DETERMINING THE EVOLUTION OF THE AMINOACYL-TRNA SYNTHETASES

BY THE RATIOS OF EVOLUTIONARY DISTANCE METHOD

by

KAMYAR FARAHI

(Under the Direction of William B. Whitman)

ABSTRACT

The availability of large numbers of genomic sequences has demonstrated the importance of lateral gene transfer (LGT) in prokaryotic evolution. However, there remains considerable uncertainty concerning the frequency of LGT compared to other evolutionary processes. To examine the frequency of LGT in ancient lineages, a method was developed that utilizes the Ratios of Evolutionary Distances (or RED) to distinguish between alternative evolutionary histories. The advantages of this approach are: the variability inherent in comparing protein sequences is transparent, the direction of LGT and the relative rates of evolution are readily identified, and it is possible to detect other types of evolutionary events. RED-T, an original computer program designed to implement the RED method, was developed during our work. RED-T is a Java application capable of generating scatter plots from given distance matrixes to analyze evolutionary relationships among various levels of taxa. In addition, it is fully capable of importing new gene data for comparison with the control set we developed or allowing the user to develop a new control. The RED method was standardized using 37 genes encoding ribosomal proteins that were believed to share a vertical evolution. Using RED-T, the evolution of the genes encoding the 20 aminoacyl-tRNA synthetases was examined. Although LGTs were common in the evolution of the aminoacyl-tRNA synthetases, they were not sufficient to obscure the organismal phylogeny. Moreover, much of the apparent complexity of the gene tree was consistent with the formation of paralogs in the ancestors to the modern lineages followed by more recent loss of one paralog or the other.

INDEX WORDS: lateral gene transfer, LGT, aminoacyl-tRNA synthetases, ribosomal proteins, ratios of evolutionary distance, RED, RED-T

DETERMINING THE EVOLUTION OF THE AMINOACYL-TRNA SYNTHETASES BY THE RATIOS OF EVOLUTIONARY DISTANCE METHOD

by

KAMYAR FARAHI

B.S., The University of Georgia, 1993

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2002

DETERMINING THE THE EVOLUTION OF THE AMINOACYL-TRNA SYNTHETASES BY THE RATIOS OF EVOLUTIONARY DISTANCE METHOD

by

KAMYAR FARAHI

Major Professor: William B. Whitman

Committee: Eileen T. Kraemer
Robert J. Maier
John F. McDonald
Juergen K. W. Wiegel

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
December 2002

DEDICATION

Vincent Van Gogh yearned it, Mahatma Gandhi preached about it, Paul Erdös never knew it, and many discover it after a lifetime. Personal and emotional connectivity makes us humans, such fragile young species, existing in a world ruled by microorganisms. I have been blessed with unconditional love and infinite support from my family and friends, who have made my life meaningful. For this reason, I dedicate my work to these individuals and convey my deepest gratitude:

My grandmother who raised me from infancy; my mother's sisters who adored me as their own son; my uncle Said and Ron who served as mentors and supporters at times of change; my friends who are my brothers and sisters; my in-laws who have graciously accepted me as their son; my brother in-law, Mihir, and friend, Rahul, who portray the rare humanity in the field of medicine; the Verma family who have opened their home kindly to me and treated me as a son; my mentor Barny Whitman who is nothing less than a gentleman of science and an admirable friend; my parents, Shamsi and Djavad, who sacrificed their lives and livelihood to immigrate to a strange land for my brother and I to have a life filled with dreams and success; my brother, Ali, who is the sparkle in my eye; and my wife, Purvi, who has shown endless support and belief in me and encouraged me to pursue my dreams. She is the sunshine of my life!

ACKNOWLEDGEMENTS

I would like to thank the members of my advisory committee, Eileen T. Kramer, Rob J. Maier, Jürgen K. W. Wiegel and John F. McDonald whose assistance with this project was invaluable. In particular, I would like to thank Dr. McDonald for his mentorship and conviction in me as a researcher from the time of my undergraduate career.

My deepest gratitude to Ross Overbeek, and his colleagues – Gordon Pusch, Mark D'Souza, Natalie Maltsev, Nikos Kyrpides, at Integrated Genomics Inc., and Argonne National Laboratory for allowing access to the WIT/ERGO database of genomic sequences and metabolic pathways. This work would not be possible with out the database nor their collaborations. Furthermore, I thank the assistance of Stephanie M. Ross, Rob Onyenwoke, Dan J. Higgins, especially Rob H. Waldo and Steven McAllister in the development of RED and RED-T application.

I very much appreciate the contributions of Suchi Bhankdarkar, Russell Malmberg and Tom Hagen. Moreover, I am particularly grateful to past and present members of the Whitman lab, Warren Gardner, David Singleton, Winsten Lin, Michelle Furlong, James Henriksen, Tiffany Major, Brian Waters and Glen Dyszynski, for their valuable input and support.

Also, I would like to thank Mark Schell and Mecky Pohlschroder for opening their laboratories and providing me with bench space during the latter years of my doctorial career.

Many thanks to the office staff of the Department of Microbiology, for their kindness and assistance throughout my years as a student of department, office of graduate records for their friendly assistance in preparation of this dissertation.

I owe endless thanks to Payman Pouladdej who graciously provided all of the computer hardware and software for my work as donations or gifts. I would like to acknowledge the following faculty, Gene Michaels, William J. Payne and Richard K. Hill who are noble men that personifying the genuine beauty and value of scientific research. I am thankful for the stimulating words of Carl Woese that inspired me at times of low confidence and doubt.

My work would not have been successful, and my degree would not have been attainable without the infinite support and patience of my mentor, William B. Whitman. I thank him for being there every day, every step in the pursuit of my degree. I thank him for valuing my well being as well as my research. I thank him for believing in me and supporting me through hard times. I thank him for challenging me to bring out the best in myself, and teaching me that science is an art of communication. I thank him for exposing me to this art that is more beautiful through the angles of diversity. I thank him for teaching me the honesty and purity of science which is rare today. And, I thank him for teaching me to think and grow not only as a scientist but as a person. A wonderful thinker and buff of science, William B. Whitman is the real thing, a gentleman of science, and above all a gentleman. I am forever indebted to him for accepting me as a pupil. I will always be honored to call myself his student. And I will dedicate my life to grow as a thinker as he has shown me and to be a significant teacher as has been for me. Thank you Barny!

TABLE OF CONTENTS

Page

CHAPTER 1

LITERATURE REVIEW

Lateral Gene Transfer

Prior to the development of genomics, evidence for horizontal or lateral gene transfer in the evolution of prokaryotes was limited to a few special cases, such as antibiotic resistance, catabolic resistance and endosymbiosis. Since the recent establishment of genomics and the resulting increase of completed genome sequences, the phylogeny among organisms have been a subject of great debate. Some researchers speculate that the rate of lateral gene transfer (LGT) is frequent and distorts the phylogenetic relationships among organisms previously inferred from highly conserved genes. Other investigators argue that the rate of LGT is relatively minor compared to the rate of linear or vertical inheritance of genetic information. In order to obtain a better grasp of the evolutionary history of prokaryotes, qualitative and quantitative analyses of the rate of LGT are vital. The discovery and historical implications of LGT are discussed here, along with the implications for prokaryotic evolution.

Lateral genetic transfer among organisms is currently the subject of substantial investigation, though the evidence for genetic exchange between related and non-related organisms was first reported in the early decades of the 20th century. Prior to the discovery of LGT, evolutionary researchers believed that inheritance occurred vertically from ancestors to their direct descendents or entirely within one lineage. With respect to bacteria, reproduction by binary fission was the sole mechanism of passing on genetic

information from a parent to daughter cells. Furthermore, this vertical transmission included any changes made in the parental genome, such as mutations, which were thought to be responsible for the evolutionary changes between offspring and their ancestors.

Between the years of 1944 and 1952, three different mechanisms were reported for LGT from one bacterium to another. First was the discovery of transformation, which was described as an organism's non-specific uptake of naked DNA and incorporation into its genome. The idea of this mechanism grew out of studies of the virulence of "*Streptococcus pneumoniae*" initiated in 1928. Later, in 1944, Avery, Macleod and McCarty showed the role of deoxyribonucleic acid (DNA) when they described the ability of microorganisms to take up and integrate DNA into the genome (Avery et al. 1944). Today, reports show high occurrences of this mechanism in various species (Lorenz and Wackernagel 1994). The next mechanism for LGT identified was transduction. This mechanism of exchange was described as being mediated by a bacteriophage that was produced from another bacterial genome. With either specialized or generalized transduction, transfer of a specific portion of a genome or any random gene was possible. Conjugation was the third mechanism of genetic transfer discovered in bacteria. This mechanism is a highly specific process that requires cell-to-cell contact during which DNA is transferred from donor to recipient cells. More specifically, a plasmid, which is not essential to the bacterium, mediates the exchange of itself or even other plasmids that do not process the genes for conjugation. Many reports described bacteriophages and plasmids as vehicles of transfer (As reviewed in Hartl et al. 1984;

Moreira 2000). In addition, an alternative mechanism of transfer via transposable elements was found to facilitate such non-linear inheritance (Clewell et al. 1995).

One of the earliest reported examples of the operation of these transfer mechanisms outside the laboratory was the acquisition of antibiotic resistance from other organisms, mainly by conjugation. The marvelous discovery of penicillin by Alexander Fleming in 1928 was followed by its isolation and purification in the late 1930's by Howard Florey and Ernst Boris Chain. However, shortly after its availability for therapeutic use in 1941, microbial resistance to this miracle drug emerged (Finland 1955; Spink et al. 1945; Vogelsang 1951; Yow 1952). Although the genetic adaptation via the accumulation of resistance mutations within the microbial population was somewhat expected, the apparent transfer of these resistance genes to other bacterial populations was alarming (Falkow et al. 1971). Another documented instance of LGT is the transfer of a P transposable element, within the eukaryotic genus *Drosophila*, via a parasitic mite as the vector (Daniels et al. 1990; Houck et al. 1991). Studies showed that distantly related species, *D. melanogaster* and *D. willistoni*, possessed this element, even though the former's close relatives lacked the P transposable element, suggesting the occurrence of a LGT event in the eukaryotes.

Later, lateral transfer among distantly related organisms were discovered. For instance, it was shown that the F-plasmid of *Escherichia coli* of the domain bacteria can be transferred via a conjugal mechanism to yeast (*Saccharomyces cerevisiae*) of the domain eukarya, hence evidence for inter-domain transfer (Heinemann and Sprague 1989). Also, the well noted endosymbiosis theory presented another example of lateral transfer among distantly related organisms. This theory proposes that the eukaryotic

mitochondrion was descended from a bacterium that was acquired by a symbiotic event between a bacterium and a primitive eukaryote (Gibbons 1992; Margulis 1993). More specifically, a prokaryote, possibly an alpha-proteobacterium, invaded or was engulfed by the ancestor of the eukaryotes. With time this endosymbiont evolved into the mitochondrial organelle in eukaryotic cells (Gray 1992). Like mitochondria, chloroplasts in algae and plant cells were also thought to have been a result of lateral transfer of a cyanobacterium (Gray 1992; Kohler et al. 1997). Goff and Coleman (1984) showed a modern example of LGT that might serve as a model for the ancient events that led to the formation of the chloroplast. One species of red algae parasitizes another species and infects its host with its own nucleus, eventually leading to the replacement of the host's nucleus and essentially converting the host's cells into new parasitic cells that are capable of infecting others (Goff and Coleman 1984). These observations provided early support of lateral gene transfer and showed that such non-linear events are possible among closely as well as distantly related organisms. Yet, these observations were largely anecdotal, and most investigators believed that the primary evolutionary path for prokaryotes was vertical rather than lateral.

Recently, the field of genomics produced numerous completed genome sequences for a wide variety of organisms. Initially the genomes of human pathogens received higher priority for completion in order to better understand their individual biology. Yet, the importance of comparative genomics showed that with more diverse genomic sequences, a greater spectrum of information could be obtained. The ability to better understand the physiology, ecology, evolution and more of a single organism or groups of organisms, and the ability to predict functions of unknown coding regions warranted

the publications of the genomes of other prokaryotes and eukaryotes. Since the first completed genome of *Haemophilus influenzae* (Fleischmann et al. 1995) in 1995, numerous genomes have been completed and available for research in public and private databases, and even more genomes are enthusiastically scheduled for completion. Some completed genomes are of special interest. *Methanococcus jannaschii* was the first euryarchaeote sequenced and serves as a great model to study methanogenesis (Bult et al. 1996). *Aeropyrum pernix* (Kawarabayasi et al. 1999) was the first crenarchaeote sequenced. The hyperthermophilic bacteria *Thermotoga maritima* (Nelson et al. 1999) and *Aquifex aeolicus* (Deckert et al. 1998) show evidence of extensive LGT events with archaea. *Deinococcus radiodurans* (Lin et al. 1999; Makarova et al. 2001; White et al. 1999) is the most radiation resistant microorganism known. The first eukaryote sequenced was *S. cerevisiae* (1997). Recently, the sequence of the human genome has been completed (McPherson et al. 2001; Venter et al. 2001). With respect to evolutionary research, it is reported that comparisons of multiple genomes have disclosed substantial evidence for lateral gene transfer among distant and closely related organisms (Nelson et al. 1999).

With the aid of genomic studies, more evidence about possibly ancient endosymbiotic lateral gene transfer was reported. For instance, Margulis and colleagues believe that there is evidence that prior to mitochondrial development, the nucleolus and possibly the motility systems of the eukaryote organisms were obtained from extremely motile bacteria (Margulis 1996).

Due to its significance in the health care field, antibiotic resistance has been and continues to be extensively studied. Several nucleotide sequence analyses showed that

antibiotic resistance genes are abundant in the prokaryotic commensals of farm animals which are grown on antibiotics. The resistance genes might then be transferred to the human commensals. For instance, the *tetQ* sequence of bacteria present in the rumen and intestines of farm animals was nearly identical to the *tetQ* sequence of human gut and oral cavity flora (Nikolich et al. 1994). Additional work by Salyers and Shoemaker showed that lateral gene transfer between bacteria of various habitats had occurred. The *tetM* sequence of a soil bacterium and colonic bacterium was almost identical, suggesting that lateral gene transfer can occur between these diverse habitats (Salyers and Shoemaker 1996). Other reports showed that LGT occurred in a marine environment from humans to fish commensals and visa versa (Kruse and Sorum 1994).

Antibiotic resistance can also be grouped with other virulence factors shared by pathogenic bacteria, which are encoded by regions of the genome termed "pathogenecity islands" (Hacker et al. 1990). The genes of these islands, which also encode for virulence factors such as adhesins, secretion systems, and iron uptake systems (Hacker and Kaper 2000), were shown to correlate with genome regions that were laterally transferred (Lawrence and Ochman 1998). Furthermore, these LGT events, along with point mutations and genetic rearrangements have been shown responsible for the emergence of genetically novel organisms with altered or new pathogenic traits (Morschhauser et al. 2000). Yet, LGT is still believed to have played a primary role in the development and spreading of antibiotic resistance genes. Researchers hope that with the help of genomics many questions regarding the evolutionary development of antibiotic resistance can be answered. For instance, did antibiotic-resistant develop in a wide population of organisms

or was it specific to a group of organisms prior to lateral transfer; and how and what were the intermediates for the transfer process (Mazel and Davies 1999)?

Further findings of lateral gene transfer were reported in the field of bioremediation. Xenobiotic pollutants in water and soil are very harmful to the environment, and the discoveries of bacteria that can degrade these pollutants provide hope for eventual bioremediation. Among various lateral transfer findings in this field, it has been shown that the *tfdA* gene, which encodes for the first enzyme of 2,4-D degradation, was nearly identical in organisms from distant locations (Matheson et al. 1997). Therefore, it was proposed that LGT increased the number of bacteria with the degradation ability for this pollutant.

Genome adaptation to environmental conditions such as temperature is also speculated to have resulted from lateral gene transfer (Kandler 1994; Wiegel 1990). Data showed that some microorganisms grew over extremely broad ranges of temperature, hence allowing these organisms to adapt and cope with temperature fluctuations in their environment. It is believed that LGT may have facilitated such adaptations by allowing organisms to acquire genes from organisms adapted to other temperature ranges. For instance, genomic analyses of hyperthermophilic bacteria, *Aquifex aeolicus* and *Thermotoga maritima*, revealed that unlike other members of their domain, these two bacteria had a large section of their genome homologous to the archaea. These findings suggested a possible connection between the lifestyles of evolutionarly distant organisms and the apparent rate of LGT between them (Aravind et al. 1998).

The wide distribution of LGT evidently has significant impact on life, in particular for understanding specific gene transfers that can directly affect advancements

in arenas such as biotechnology, environment, food, and pharmaceutical products. Although the practical importance is obvious, the debatable frequency of these events has blurred the depiction of phylogenetic relationships among organisms. In the 1970's and 1980's, the universal tree of life was constructed from the 16S ribosomal RNA (rRNA), which served as a stable model for the evolutionary history of both closely and distantly related organisms. This model predicted three domains – archaea, bacteria and eukarya. More specifically, this model shows bacteria and the common ancestor of the archaea and eukaryotes as the progeny of the last "universal common ancestor". This model shed light on many taxonomic relationships within the domains (Olsen et al. 1994). Woese selected 16S rRNA as the best possible representative to depict an evolutionary history because its rate of sequence change was very slow, it was an essential component of all known organisms, and was readily isolated (Woese 1987; Woese and Fox 1977; Woese et al. 1975). Although the 16S rRNA is still widely used, in recent years some investigators have begun to question the inference of phylogenetic relationships for all life from a single gene. Taking into consideration LGT events, these skeptics believe that a gene's evolutionary history does not necessarily reflect the phylogenetic relationships among whole genomes. Furthermore, numerous reports of LGTs suggest that the utility of rRNA as a marker for inferring a universal tree of life is limited. Although not associated to LGT, other reports extended the list of limitations for this evolutionary marker. For instance, a report showed that rRNA's depiction of hyperthermophiles, such as *Aquifex* and *Thermotoga*, as closely related genera was false because their rRNA similarity was due to convergence (Galtier et al. 1999; Hirt et al. 1999). Another report estimated that the lineages leading to *E. coli* and *Salmonella enterica* split about 100 million years ago,

and since then LGT allowed both species to acquire genetically novel systems. For instance, these two species' phenotypic differences can be attributed to particular genes responsible for utilization of citrate, lactose, and propanediol, genes involved in production of indole, and genes accountable for pathogenicity. In general, these investigators estimated that 17.6% of the *E. coli* genome was acquired by LGT after the divergence of *S. enterica* (Lawrence and Ochman 1998), thus suggesting that there are genes affected by LGT and that such evolutionary events can alter existing genetic systems or develop new ones. LGT was also believed to occur frequently across domains. Examples included the presence of bacteria and archaea genes in eukaryotes (Brown and Doolittle 1997; Feng et al. 1997), metabolic genes of bacteria in archaea (Boucher and Doolittle 2000; Brown and Doolittle 1997; Doolittle and Logsdon 1998; Jain et al. 1999), and the surprising distribution of essential genes such as aminoacyl-tRNA synthetases among all three domains (Brown and Doolittle 1999; Curnow et al. 1997; Doolittle and Handy 1998; Woese et al. 2000; Wolf et al. 1999). Also, the discovery that the bacterium *Thermotoga maritima* shares about 25% of its genome with the distantly related archaeon *Pyrococcus horikoshii* and not with close relatives provided another example of LGT across domains (Nelson et al. 1999; Nesbo et al. 2001; Worning et al. 2000). These limitations and overall disagreements with the 16S rRNA led W. F. Doolittle to present a new model of phylogenetic relationships of organisms. Instead of a tree of life, a web or "net" of life would be a more suitable model to reflect the speculated high frequency of LGT events (Doolittle 1999; Doolittle 1999). A similar notion of a web-like relationship among organisms was suggested in the early 1990's by Hilario and Gogarten (1993), and even as early as the late 1970's by Reanney and Sonea (1978), who suggested that all

organisms on the planet could be viewed as a single entity, or a "global super-organism," due to the numerous genetic exchange platforms, such as bacteriophages, transposable elements and plasmids (See also Sonea and Paniset 1976).

It also important to note that with a high frequency of LGT events and the projection of a net-like evolutionary hodge-podge, the idea of a single universal ancestor is also debatable. Arguments among the research community ranged greatly. Some accept the idea of a last common ancestor (Aravind et al. 1999; Castresana and Moreira 1999; Labedan et al. 1999; Tomii and Kanehisa 1998) and also believe that the role of LGT has been greatly exaggerated (Glansdorff 2000). Some approach LGT with a doubtful attitude (Philippe and Forterre 1999), and others believe in a high LGT frequency (Doolittle 1999; Gogarten et al. 1996; Martin 1999) and do not believe it is possible to identify a universal ancestor at all. Moreover, Woese who first introduced the 16S rRNA tree suggests "the universal ancestor is not a discrete entity, rather a diverse community of cells that survives and evolves as a biological unit." He adds that this communal ancestor evolved into more complex systems that eventually served as the starting points of the lineages for the three proposed domains (Woese 1998).

Whether future findings of LGT occurrence support models of "continual gene transfer" (Jain et al. 1999) – an ongoing process of gene acquisition among organisms – or "early massive transfer" (Woese 1998) – a massive genetic exchange event which occurred long before modern species diverged and early in organisms' evolution – the importance is great. Considering the great implications on evolutionary research, the need to determine the regularity and distribution of LGT among genomes is critical. However, the older phylogenetic methods may not be best suited for this task. For many years

phylogenetic trees have been analyzed manually to infer evolutionary relationships and to discover branching inconsistencies that would imply LGT events. Yet, this approach tends to be time-consuming and difficult, especially with the growing list of completed genomes. Therefore, with the aid of novel genomic tools, several methods have been developed to examine LGT occurrence. This approach to phylogenetic analysis is referred to by some investigators as "phylogenomics" (Eisen 1998). By comparing numerous genomes, one can better understand the frequency and timeline in which LGT occurs and possibly describe genes that are more prone to LGT than others (Jain et al. 1999).

Hansmann and Martin aligned 39 gene sequences – mostly ribosomal proteins – which were thought to be well conserved across 18 completed prokaryotic genomes. Later, numerous gene trees were generated and analyzed. After analyzing this large-scale operation, they concluded that exclusion of only a few percentages of the nucleotide positions could significantly influence the apparent phylogenetic history. Essential to their objective, steps were also taken to ensure that the archaeal and bacterial domains were monophyletic in all the trees. Trees were then inspected visually for possible inter-domain LGT events. During these inspections, three separate gene trees suggested different incidences of inter-domain LGT among organisms they tested. These were deduced by visually locating genes that branched among genes from the other domains (Hansmann and Martin 2000).

In another example, Sicheritz-Pontet and Andersson (2001), developed an automated protocol that impressively generated and analyzed 8000 gene trees from seven genomes, representing all three domains. Their results suggested a high degree of

phylogenetic connections, yet the domains remained clearly distinct. Further, they reported that many of the gene trees were agreeable with the 16S rRNA evolutionary inferences.

Genome-based phylogenetic methods were developed to use gene content for deciphering phylogeny. One group developed an approach that defined orthologs by using high sequence similarity values, observed pairs of genomes, computed their fraction of shared genes, converted this fraction to distances, and then constructed a tree using neighbor-joining or least-squares methods (Snel et al. 1999). This collection was made available at a web server named SHOT (Korbel et al. 2002). Another group used FASTA scores followed by single-linkage clustering to identify orthologs, then used parsimony analysis to reconstruct trees (House and Fitz-Gibbon 2002). Another interesting method developed by Clark et al. (2002) relied on BLAST scores to compute the ratio of orthologs to the number of genes in the smaller genome, and used this to construct trees based on the least-squares method. In this method, orthologs were defined as reciprocal best match (RBM), and sequences that showed an abnormal pattern or similarity to the orthologs were removed to reduce the misidentification of orthologs. All of these methods shared several important findings. Relative phylogenetic distance measurements are reliable due to the adequate preservation of gene content. It also appeared in these different methods that separations between the domains were valid and that closely related organisms grouped as expected. Yet, a shortcoming of this genome-based approach was that distances reflecting intermediate relationships seemed to be unreliable.

Garcia-Vallve et al. (1999, 2000) reported the development of a statistical approach to predict genomic gene acquisition via LGT. Believing that genes in a genome are biased in codon usage at the species level, they used GC content and codon usage to determine the portions of the genome that may have been acquired by LGT (See also Kaplan and Fine 1998). In their study, 17 bacterial and seven archaeal genomes were used. As a result, they echoed the theory stated by Jain et al. that operational genes were more likely to be transferred than informational genes. More specifically, they were able to statistically show that "potency" of laterally transferred genes was much higher in archaeal and non-pathogenic bacterial genomes than the pathogenic bacterium – with the exception of *Mycoplasma genitalium* (Garcia-Vallve et al. 2000).

The importance of developing quantitative protocols to interpret the large volume of genomic data and to assess confidence based on statistical models is vital (Ragan 2001). Whether or not we can use quantitative methods to evaluate rates and patterns of gene loss, LGT, paralogy or other processes, still remains to be seen. However, we cannot rely only on observational analyses. Thus, it is clear that the combination of quantitative and qualitative approaches is necessary. Further, the success of the field depends on recognition of the importance of more careful annotations of the genomic data along with more reliance on laboratory experiments.

It is safe to say that lateral and vertical gene transfer mechanisms coexist in cellular evolution. Woese describes the relationship of the two mechanisms as like the "yin and yang" of evolutionary process; although opposites, their co-existence was needed for evolution as we know it (Woese 2000). Whatever the analogy, without the field of genomics this relationship may not have been appreciated, for the focus on LGT

may not have been as popular. Even though the debate is ongoing, genomics should proudly call recent LGT discoveries as an exemplary accomplishment of the field. With that said, besides LGT the importance of another evolutionary process, lineage specific gene loss, cannot be forgotten. For example, the differences in genetic content for two gama-proteobacteria, *E. coli* and *H. influenzae*, was attributed to differential gene loss along with LGT (Tatusov et al. 1996). Snel et al. (2002) use quantitative analysis of archaeal and proteobacterial genomes to show that the distribution of a great number of orthologous genes encoding prokaryotic proteins was due to lineage-specific gene loss and gene gain by LGT. In this analysis, the gene content of an ancestral genome was hypothesized as well as the evolutionary events necessary to produce modern genomes. The strength of this approach lies in realizing that the concept of taxonomic relationship must be considered along with the history of the cell's genetic makeup.

In summary, whether attempting to determine an evolutionary timetable of LGT, questioning the existence of a universal ancestor, or searching for a single gene worthy of projecting the "universal tree of life," all paths call for spending great energy and effort to determine new and better biologically based approaches to decipher and understand the evolutionary history of organisms.

Aminoacyl-tRNA Synthetases

During translation, messenger RNA (mRNA) serves as an intermediate between the cell's genetic information and the encoded proteins. The aminoacyl-tRNA synthetases (aaRSs) are enzymes that catalyze the specific aminacylation of tRNAs, specifically matching the genetic code of the tRNA as anticodons. AaRSs catalyze the 3'-esterification of tRNAs with the correct amino acid. This charging continues during the elongation phase of protein synthesis as the genetic code is translated, and the mRNA direct the synthesis of a unique sequence of amino acids, followed by the termination phase. Most cells make twenty different aaRSs, one for each type of amino acid (see below). These twenty enzymes are different, each optimized for function with its own particular amino acid and the set of tRNA molecules appropriate to that amino acid. This family of enzymes is diverse and ancient, and due to their function in the essential phase of translation, their evolution is suggested to be reflective of the early development of the genetic code (Brown and Doolittle 1995; Schimmel et al. 1993; Woese et al. 2000).

With the help of genomics the overall process of aaRSs, such as the evolutionary implications, became clearer. For the years prior to genomics, only a small group of biological systems were used to observe aaRSs, such as *E. coli*. This limited testing resulted in the misconception that 20 aaRSs would be found in nearly all organisms. In the genomic sequence of an archaeon *Methanococcus jannaschii*, homologs for only 16 of the 20 aaRSs were found (Bult et al. 1996).

The 20 aaRSs have been separated into two different groups, classes I and II (Eriani et al. 1990). Each is different in having a unique sequence motif reflecting a

domain for the active sites. The active sites are for adenylate synthesis, which is involved in the condensation of an amino acid with ATP to yield an aminoacyl adenylate, and for attaching the activated amino acid to the 3'-end of the tRNA. The active sites of class I enzymes have a Rossmann dinucleotide-binding fold made up of alternating beta-strands and alpha-helices, whereas the active sites of class II enzymes do not contain this fold but instead contain a unique antiparallel beta-fold (Arnez and Moras 1997). The difference between the classes can seen when binding ATP: class I enzymes bind the substrate in an extended conformation, while class II enzymes do so in a bent conformation (Rould et al. 1989). Rould et al. (1989) described this difference in protein conformation while observing glutaminyl-tRNA synthetases. The classes also differ in binding tRNA, where the class I enzymes approach the tRNA from the minor groove with the variable group facing the solvent, and the class II enzymes approach in an opposite manner. Exceptions exist, as one report shows that the class II enzyme, alanyl-tRNA synthetase, approaches the tRNA in a manner similar to the class I enzymes (Beuning et al. 1997). It is interesting that these two classes have nearly the same number of members. This agrees with the hypothesis of the development of paired synthetases, which requires that they gave rise to two classes with equal numbers of enzymes (Ribas de Pouplana and Schimmel 2000). Further, aaRS genes may have been involved in LGT (Doolittle and Handy 1998). An example is the class I enzyme, lysyl-tRNA synthetase, which has been proposed to have been laterally transferred to two different groups of bacteria, the spirochetes and alpha-proteobacteria (Ibba et al. 1999). In summary, these diverse, yet essential and conserved, enzymes can serve as great subjects for investigating and developing newer methods for understanding evolutionary mechanisms.

References

Anonymous. (1997) The yeast genome directory. *Nature,* 387, 5.

Aravind, L., Tatusov, R.L., Wolf, Y.I., Walker, D.R. and Koonin, E.V. (1998) Evidence for massive gene exchange between archaeal and bacterial hyperthermophiles. *Trends Genet,* 14, 442-444.

Aravind, L., Walker, D.R. and Koonin, E.V. (1999) Conserved domains in DNA repair proteins and evolution of repair systems. *Nucleic Acids Res,* 27, 1223-1242.

Arnez, J.G. and Moras, D. (1997) Structural and functional considerations of the aminoacylation reaction. *Trends Biochem Sci,* 22, 211-216.

Avery, O.T., MacLeod, C.M. and McCarty, M. (1944) Studies on the chemical nature of the substance inducing transformation of pneumococcal types. Inductions of transformation by a desoxyribonucleic acid fraction isolated from pneumococcal type III. *J. Exp. Med.,* 149, 297-326.

Beuning, P.J., Yang, F., Schimmel, P. and Musier-Forsyth, K. (1997) Specific atomic groups and RNA helix geometry in acceptor stem recognition by a tRNA synthetase. *Proc Natl Acad Sci U S A,* 94, 10150-10154.

Boucher, Y. and Doolittle, W.F. (2000) The role of lateral gene transfer in the evolution of isoprenoid biosynthesis pathways. *Mol Microbiol,* 37, 703-716.

Brown, J.R. and Doolittle, W.F. (1995) Root of the universal tree of life based on ancient aminoacyl-tRNA synthetase gene duplications. *Proc Natl Acad Sci U S A,* 92, 2441-2445.

Brown, J.R. and Doolittle, W.F. (1997) Archaea and the prokaryote-to-eukaryote transition. *Microbiol Mol Biol Rev,* 61, 456-502.

Brown, J.R. and Doolittle, W.F. (1999) Gene descent, duplication, and horizontal transfer in the evolution of glutamyl- and glutaminyl-tRNA synthetases. *J Mol Evol,* 49, 485-495.

Bult, C.J., White, O., Olsen, G.J., Zhou, L., Fleischmann, R.D., Sutton, G.G., Blake, J.A., FitzGerald, L.M., Clayton, R.A., Gocayne, J.D., Kerlavage, A.R., Dougherty, B.A., Tomb, J.F., Adams, M.D., Reich, C.I., Overbeek, R., Kirkness, E.F., Weinstock, K.G., Merrick, J.M., Glodek, A., Scott, J.L., Geoghagen, N.S. and Venter, J.C. (1996) Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. *Science,* 273, 1058-1073.

Castresana, J. and Moreira, D. (1999) Respiratory chains in the last common ancestor of living organisms. *J Mol Evol,* 49, 453-460.

Clarke, G.D., Beiko, R.G., Ragan, M.A. and Charlebois, R.L. (2002) Inferring genome trees by using a filter to eliminate phylogenetically discordant sequences and a distance matrix based on mean normalized BLASTP scores. *J Bacteriol,* 184, 2072-2080.

Clewell, D.B., Flannagan, S.E. and Jaworski, D.D. (1995) Unconstrained bacterial promiscuity: the Tn916-Tn1545 family of conjugative transposons. *Trends Microbiol,* 3, 229-236.

Curnow, A.W., Hong, K.W., Yuan, R. and Soll, D. (1997) tRNA-dependent amino acid transformations. *Nucleic Acids Symp Ser,* 36, 2-4.

Daniels, S.B., Peterson, K.R., Strausbaugh, L.D., Kidwell, M.G. and Chovnick, A. (1990) Evidence for horizontal transmission of the P transposable element between *Drosophila* species. *Genetics,* 124, 339-355.

Deckert, G., Warren, P.V., Gaasterland, T., Young, W.G., Lenox, A.L., Graham, D.E., Overbeek, R., Snead, M.A., Keller, M., Aujay, M., Huber, R., Feldman, R.A., Short, J.M., Olsen, G.J. and Swanson, R.V. (1998) The complete genome of the hyperthermophilic bacterium *Aquifex aeolicus*. *Nature,* 392, 353-358.

Doolittle, R.F. and Handy, J. (1998) Evolutionary anomalies among the aminoacyl-tRNA synthetases. *Curr Opin Genet Dev,* 8, 630-636.

Doolittle, W.F. (1999) Lateral genomics. *Trends Cell Biol,* 9, M5-8.

Doolittle, W.F. (1999) Phylogenetic classification and the universal tree. *Science,* 284, 2124-2129.

Doolittle, W.F. and Logsdon, J.M., Jr. (1998) Archaeal genomics: do archaea have a mixed heritage? *Curr Biol,* 8, R209-211.

Eisen, J.A. (1998) Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Res,* 8, 163-167.

Eriani, G., Delarue, M., Poch, O., Gangloff, J. and Moras, D. (1990) Partition of tRNA synthetases into two classes based on mutually exclusive sets of sequence motifs. *Nature,* 347, 203-206.

Falkow, S., Tompkins, L.S., Silver, R.P., Guerry, P. and Le Blanc, D.J. (1971) The problems of drug-resistant pathogenic bacteria. The replication of R-factor DNA in *Escherichia coli* K-12 following conjugation. *Ann N Y Acad Sci,* 182, 153-171.

Feng, D.F., Cho, G. and Doolittle, R.F. (1997) Determining divergence times with a protein clock: update and reevaluation. *Proc Natl Acad Sci U S A,* 94, 13028-13033.

Finland, M. (1955) Emergence of antibiotic-resistance bacteria. *New England Journal of Medicine,* 253, 909-922.

Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R., Bult, C.J., Tomb, J.F., Dougherty, B.A., Merrick, J.M. and et al. (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science,* 269, 496-512.

Galtier, N., Tourasse, N. and Gouy, M. (1999) A nonhyperthermophilic common ancestor to extant life forms. *Science,* 283, 220-221.

Garcia-Vallve, S., Palau, J. and Romeu, A. (1999) Horizontal gene transfer in glycosyl hydrolases inferred from codon usage in *Escherichia coli* and *Bacillus subtilis*. *Mol Biol Evol,* 16, 1125-1134.

Garcia-Vallve, S., Romeu, A. and Palau, J. (2000) Horizontal gene transfer in bacterial and archaeal complete genomes. *Genome Res,* 10, 1719-1725.

Garcia-Vallve, S., Romeu, A. and Palau, J. (2000) Horizontal gene transfer of glycosyl hydrolases of the rumen fungi. *Mol Biol Evol,* 17, 352-361.

Gibbons, A. (1992) Mitochondrial Eve: wounded, but not dead yet. *Science,* 257, 873-875.

Glansdorff, N. (2000) About the last common ancestor, the universal life-tree and lateral gene transfer: a reappraisal. *Mol Microbiol,* 38, 177-185.

Goff, L.J. and Coleman, A.W. (1984) Transfer of nuclei from a parasite to its host. *P Natl Acad Sci-Biol,* 81, 5420-5424.

Gogarten, J.P., Olendzenski, L., Hilario, E., Simon, C. and Holsinger, K.E. (1996) Dating the cenancester of organisms. *Science,* 274, 1750-1751; discussion 1751-1753.

Gray, M.W. (1992) The endosymbiont hypothesis revisited. *Int Rev Cytol,* 141, 233-357.

Hacker, J., Bender, L., Ott, M., Wingender, J., Lund, B., Marre, R. and Goebel, W. (1990) Deletions of chromosomal regions coding for fimbriae and hemolysins occur in vitro and in vivo in various extraintestinal *Escherichia coli* isolates. *Microb Pathog,* 8, 213-225.

Hacker, J. and Kaper, J.B. (2000) Pathogenicity islands and the evolution of microbes. *Annu Rev Microbiol,* 54, 641-679.

Hansmann, S. and Martin, W. (2000) Phylogeny of 33 ribosomal and six other proteins encoded in an ancient gene cluster that is conserved across prokaryotic genomes: influence of excluding poorly alignable sites from analysis. *Int J Syst Evol Microbiol,* 50, 1655-1663.

Hartl, D.L., Dykhuizen, D.E. and Berg, D.E. (1984) Accessory DNAs in the bacterial gene pool: playground for coevolution. *CIBA Found Symp,* 102, 233-245.

Heinemann, J.A. and Sprague, G.F., Jr. (1989) Bacterial conjugative plasmids mobilize DNA transfer between bacteria and yeast. *Nature,* 340, 205-209.

Hilario, E. and Gogarten, J.P. (1993) Horizontal transfer of ATPase genes--the tree of life becomes a net of life. *Biosystems,* 31, 111-119.

Hirt, R.P., Logsdon, J.M., Jr., Healy, B., Dorey, M.W., Doolittle, W.F. and Embley, T.M. (1999) Microsporidia are related to Fungi: evidence from the largest subunit of RNA polymerase II and other proteins. *Proc Natl Acad Sci U S A,* 96, 580-585.

Houck, M.A., Clark, J.B., Peterson, K.R. and Kidwell, M.G. (1991) Possible horizontal transfer of *Drosophila* genes by the mite *Proctolaelaps regalis*. *Science,* 253, 1125-1128.

House, C.H. and Fitz-Gibbon, S.T. (2002) Using homolog groups to create a whole-genomic tree of free-living organisms: an update. *J Mol Evol,* 54, 539-547.

Ibba, M., Losey, H.C., Kawarabayasi, Y., Kikuchi, H., Bunjun, S. and Soll, D. (1999) Substrate recognition by class I lysyl-tRNA synthetases: a molecular basis for gene displacement. *Proc Natl Acad Sci U S A,* 96, 418-423.

Jain, R., Rivera, M.C. and Lake, J.A. (1999) Horizontal gene transfer among genomes: the complexity hypothesis. *Proc Natl Acad Sci U S A,* 96, 3801-3806.

Kandler, O. (1994) The early diversification of life. 84, 152-160.

Kaplan, J.B. and Fine, D.H. (1998) Codon usage in *Actinobacillus actinomycetemcomitans*. *FEMS Microbiol Lett,* 163, 31-36.

Kawarabayasi, Y., Hino, Y., Horikawa, H., Yamazaki, S., Haikawa, Y., Jin-no, K., Takahashi, M., Sekine, M., Baba, S., Ankai, A., Kosugi, H., Hosoyama, A., Fukui, S., Nagai, Y., Nishijima, K., Nakazawa, H., Takamiya, M., Masuda, S., Funahashi, T., Tanaka, T., Kudoh, Y., Yamazaki, J., Kushida, N., Oguchi, A., Kikuchi, H. and et al. (1999) Complete genome sequence of an aerobic hyper-thermophilic crenarchaeon, *Aeropyrum pernix* K1. *DNA Res,* 6, 83-101, 145-152.

Kohler, S., Delwiche, C.F., Denny, P.W., Tilney, L.G., Webster, P., Wilson, R.J., Palmer, J.D. and Roos, D.S. (1997) A plastid of probable green algal origin in Apicomplexan parasites. *Science,* 275, 1485-1489.

Korbel, J.O., Snel, B., Huynen, M.A. and Bork, P. (2002) SHOT: a web server for the construction of genome phylogenies. *Trends Genet,* 18, 158-162.

Kruse, H. and Sorum, H. (1994) Transfer of multiple drug resistance plasmids between bacteria of diverse origins in natural microenvironments. *Appl Environ Microbiol,* 60, 4015-4021.

Labedan, B., Boyen, A., Baetens, M., Charlier, D., Chen, P., Cunin, R., Durbeco, V., Glansdorff, N., Herve, G., Legrain, C., Liang, Z., Purcarea, C., Roovers, M., Sanchez, R., Toong, T.L., Van de Casteele, M., van Vliet, F., Xu, Y. and Zhang, Y.F. (1999) The evolutionary history of carbamoyltransferases: A complex set of paralogous genes was already present in the last universal common ancestor. *J Mol Evol,* 49, 461-473.

Lawrence, J.G. and Ochman, H. (1998) Molecular archaeology of the *Escherichia coli* genome. *Proc Natl Acad Sci U S A,* 95, 9413-9417.

Lin, J., Qi, R., Aston, C., Jing, J., Anantharaman, T.S., Mishra, B., White, O., Daly, M.J., Minton, K.W., Venter, J.C. and Schwartz, D.C. (1999) Whole-genome shotgun optical mapping of *Deinococcus radiodurans*. *Science,* 285, 1558-1562.

Lorenz, M.G. and Wackernagel, W. (1994) Bacterial gene transfer by natural genetic transformation in the environment. *Microbiol Rev,* 58, 563-602.

Makarova, K.S., Aravind, L., Wolf, Y.I., Tatusov, R.L., Minton, K.W., Koonin, E.V. and Daly, M.J. (2001) Genome of the extremely radiation-resistant bacterium *Deinococcus radiodurans* viewed from the perspective of comparative genomics. *Microbiol Mol Biol Rev,* 65, 44-79.

Margulis, L. (1993) Symbiosis in cell evolution : microbial communities in the Archean and Proterozoic eons. 27, 452.

Margulis, L. (1996) Archaeal-eubacterial mergers in the origin of Eukarya: phylogenetic classification of life. *Proc Natl Acad Sci U S A,* 93, 1071-1076.

Martin, W. (1999) Mosaic bacterial chromosomes: a challenge en route to a tree of genomes. *Bioessays,* 21, 99-104.

Matheson, V.G., Forney, L.J., Suwa, Y., Nakatsu, C.H., Sewtone, A.J. and Holben, W.E. (1997) Evidence for aquisition in nature of a chromosomal 2,4-dichrolophenooxyacetic acid/ketoglutarate dioxygenase gene by different *Burkholderia* spp. *Appl Environ Microbiol,* 63, 2266-2272.

Mazel, D. and Davies, J. (1999) Antibiotic resistance in microbes. *Cell Mol Life Sci,* 56, 742-754.

McPherson, J.D., Marra, M., Hillier, L., Waterston, R.H., Chinwalla, A., Wallis, J., Sekhon, M., Wylie, K., Mardis, E.R., Wilson, R.K., Fulton, R., Kucaba, T.A., Wagner-McPherson, C., Barbazuk, W.B., Gregory, S.G., Humphray, S.J., French, L., Evans, R.S., Bethel, G., Whittaker, A., Holden, J.L., McCann, O.T., Dunham, A., Soderlund, C., Scott, C.E., Bentley, D.R., Schuler, G., Chen, H.C., Jang, W., Green, E.D., Idol, J.R., Maduro, V.V., Montgomery, K.T., Lee, E., Miller, A., Emerling, S., Kucherlapati, Gibbs, R., Scherer, S., Gorrell, J.H., Sodergren, E., Clerc-Blankenburg, K., Tabor, P., Naylor, S., Garcia, D., de Jong, P.J., Catanese, J.J., Nowak, N., Osoegawa, K., Qin, S., Rowen, L., Madan, A., Dors, M., Hood, L., Trask, B., Friedman, C., Massa, H., Cheung, V.G., Kirsch, I.R., Reid, T., Yonescu, R., Weissenbach, J., Bruls, T., Heilig, R., Branscomb, E., Olsen, A., Doggett, N., Cheng, J.F., Hawkins, T., Myers, R.M., Shang, J., Ramirez, L., Schmutz, J., Velasquez, O., Dixon, K., Stone, N.E., Cox, D.R., Haussler, D.,

Kent, W.J., Furey, T., Rogic, S., Kennedy, S., Jones, S., Rosenthal, A., Wen, G., Schilhabel, M., Gloeckner, G., Nyakatura, G., Siebert, R., Schlegelberger, B., Korenberg, J., Chen, X.N., Fujiyama, A., Hattori, M., Toyoda, A., Yada, T., Park, H.S., Sakaki, Y., Shimizu, N., Asakawa, S., et al. (2001) A physical map of the human genome. *Nature,* 409, 934-941.

Moreira, D. (2000) Multiple independent horizontal transfers of informational genes from bacteria to plasmids and phages: implications for the origin of bacterial replication machinery. *Mol Microbiol,* 35, 1-5.

Morschhauser, J., Kohler, G., Ziebuhr, W., Blum-Oehler, G., Dobrindt, U. and Hacker, J. (2000) Evolution of microbial pathogens. *Philos Trans R Soc Lond B Biol Sci,* 355, 695-704.

Nelson, K.E., Clayton, R.A., Gill, S.R., Gwinn, M.L., Dodson, R.J., Haft, D.H., Hickey, E.K., Peterson, J.D., Nelson, W.C., Ketchum, K.A., McDonald, L., Utterback, T.R., Malek, J.A., Linher, K.D., Garrett, M.M., Stewart, A.M., Cotton, M.D., Pratt, M.S., Phillips, C.A., Richardson, D., Heidelberg, J., Sutton, G.G., Fleischmann, R.D., Eisen, J.A., Fraser, C.M. and et al. (1999) Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of *Thermotoga maritima*. *Nature,* 399, 323-329.

Nesbo, C.L., L'Haridon, S., Stetter, K.O. and Doolittle, W.F. (2001) Phylogenetic Analyses of Two "Archaeal" Genes in *Thermotoga maritima* reveal multiple transfers between archaea and bacteria. *Mol Biol Evol,* 18, 362-375.

Nikolich, M.P., Hong, G., Shoemaker, N.B. and Salyers, A.A. (1994) Evidence for natural horizontal transfer of *tetQ* between bacteria that normally colonize

humans and bacteria that normally colonize livestock. *Appl Environ Microbiol,* 60, 3255-3260.

Olsen, G.J., Woese, C.R. and Overbeek, R. (1994) The winds of (evolutionary) change: breathing new life into microbiology. *J Bacteriol,* 176, 1-6.

Philippe, H. and Forterre, P. (1999) The rooting of the universal tree of life is not reliable. *J Mol Evol,* 49, 509-523.

Ragan, M.A. (2001) Detection of lateral gene transfer among microbial genomes. *Curr Opin Genet Dev,* 11, 620-626.

Reanney, D.C. (1978) Coupled evolution: adaptive interactions among the genomes of plasmids, viruses, and cells. *Int Rev Cytol Suppl,* Suppl, 1-68.

Ribas de Pouplana, L. and Schimmel, P. (2000) A view into the origin of life: aminoacyl-tRNA synthetases. *Cell Mol Life Sci,* 57, 865-870.

Rould, M.A., Perona, J.J., Soll, D. and Steitz, T.A. (1989) Structure of *E. coli* glutaminyl-tRNA synthetase complexed with tRNA(Gln) and ATP at 2.8 A resolution. *Science,* 246, 1135-1142.

Salyers, A.A. and Shoemaker, N.B. (1996) Resistance gene transfer in anaerobes: new insights, new problems. *Clin Infect Dis,* 23 Suppl 1, S36-43.

Schimmel, P., Giege, R., Moras, D. and Yokoyama, S. (1993) An operational RNA code for amino acids and possible relationship to genetic code. *Proc Natl Acad Sci U S A,* 90, 8763-8768.

Sicheritz-Ponten, T. and Andersson, S.G. (2001) A phylogenomic approach to microbial evolution. *Nucleic Acids Res,* 29, 545-552.

Snel, B., Bork, P. and Huynen, M.A. (1999) Genome phylogeny based on gene content. *Nat Genet,* 21, 108-110.

Snel, B., Bork, P. and Huynen, M.A. (2002) Genomes in flux: the evolution of archaeal and proteobacterial gene content. *Genome Res,* 12, 17-25.

Sonea, S. and Paniset, M. (1976) Towards a new bacteriology. *Rev Can Biol,* 35, 103-167.

Spink, W.W., Hall, W.H. and Ferris, V. (1945) Clinical significance of staphylococci - with natural or acquired resistance to the sulfonamides and to penicillin. *JAMA,* 128, 555-559.

Tatusov, R.L., Mushegian, A.R., Bork, P., Brown, N.P., Hayes, W.S., Borodovsky, M., Rudd, K.E. and Koonin, E.V. (1996) Metabolism and evolution of *Haemophilus influenzae* deduced from a whole- genome comparison with *Escherichia coli*. *Curr Biol,* 6, 279-291.

Tomii, K. and Kanehisa, M. (1998) A comparative analysis of ABC transporters in complete microbial genomes. *Genome Res,* 8, 1048-1059.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., Gocayne, J.D., Amanatides, P., Ballew, R.M., Huson, D.H., Wortman, J.R., Zhang, Q., Kodira, C.D., Zheng, X.H., Chen, L., Skupski, M., Subramanian, G., Thomas, P.D., Zhang, J., Gabor Miklos, G.L., Nelson, C., Broder, S., Clark, A.G., Nadeau, J., McKusick, V.A., Zinder, N., Levine, A.J., Roberts, R.J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington,

K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Di Francesco, V., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A.E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T.J., Higgins, M.E., Ji, R.R., Ke, Z., Ketchum, K.A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G.V., Milshina, N., Moore, H.M., Naik, A.K., Narayan, V.A., Neelam, B., Nusskern, D., Rusch, D.B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z., Wang, A., Wang, X., Wang, J., Wei, M., Wides, R., Xiao, C., Yan, C., et al. (2001) The sequence of the human genome. *Science,* 291, 1304-1351.

Vogelsang, T.M. (1951) The incident of penicillin-resistant pathogenic staphylococci isolated from the upper respiratory tract of young healthy persons. *Acta Pathologica et Microbiologica Scandinavia,* 29, 363-367.

White, O., Eisen, J.A., Heidelberg, J.F., Hickey, E.K., Peterson, J.D., Dodson, R.J., Haft, D.H., Gwinn, M.L., Nelson, W.C., Richardson, D.L., Moffat, K.S., Qin, H., Jiang, L., Pamphile, W., Crosby, M., Shen, M., Vamathevan, J.J., Lam, P., McDonald, L., Utterback, T., Zalewski, C., Makarova, K.S., Aravind, L., Daly, M.J., Fraser, C.M. and et al. (1999) Genome sequence of the radioresistant bacterium *Deinococcus radiodurans* R1. *Science,* 286, 1571-1577.

Wiegel, J. (1990) Temperature spans for growth: hypothesis and discussion. *FEMS Micriobiol Rev,* 75, 155-170.

Woese, C. (1998) The universal ancestor. *Proc Natl Acad Sci U S A,* 95, 6854-6859.

Woese, C.R. (1987) Bacterial evolution. *Microbiol Rev,* 51, 221-271.

Woese, C.R. (2000) Interpreting the universal phylogenetic tree. *Proc Natl Acad Sci U S A,* 97, 8392-8396.

Woese, C.R. and Fox, G.E. (1977) Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc Natl Acad Sci U S A,* 74, 5088-5090.

Woese, C.R., Fox, G.E., Zablen, L., Uchida, T., Bonen, L., Pechman, K., Lewis, B.J. and Stahl, D. (1975) Conservation of primary structure in 16S ribosomal RNA. *Nature,* 254, 83-86.

Woese, C.R., Olsen, G.J., Ibba, M. and Soll, D. (2000) Aminoacyl-tRNA synthetases, the genetic code, and the evolutionary process. *Microbiol Mol Biol Rev,* 64, 202-236.

Wolf, Y.I., Aravind, L., Grishin, N.V. and Koonin, E.V. (1999) Evolution of aminoacyl-tRNA synthetases--analysis of unique domain architectures and phylogenetic trees reveals a complex history of horizontal gene transfer events. *Genome Res,* 9, 689-710.

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

Yow, E.M. (1952) Development of proteus and pseudomonas infection during antibiotic therapy. *JAMA,* 149, 1184-1188.

CHAPTER 2

DETECTION OF LATERAL GENE TRANSFER (LGT) EVENTS IN THE EVOLUTION OF TRNA SYNTHETASES BY THE RATIOS OF EVOLUTIONARY DISTANCE METHOD[†]

[†] Farahi, K., G.D. Pusch, R. Overbeek and W.B. Whitman. To be submitted to *Journal of Molecular Evolution*.

Abstract

The availability of large numbers of genomic sequences has demonstrated the importance of lateral gene transfer (LGT) in prokaryotic evolution. However, considerable uncertainty remains concerning the frequency of LGT compared to other evolutionary processes. To examine the frequency of LGT in ancient lineages a method was developed that utilizes the Ratios of Evolutionary Distances (or RED) to distinguish between alternative evolutionary histories. The advantages of this approach are: the variability inherent in comparing protein sequences is transparent, the direction of LGT and the relative rates of evolution are readily identified, and it is possible to detect other types of evolutionary events. This method was standardized using 37 genes encoding ribosomal proteins that were believed to share a vertical evolution. Using RED-T, an original computer program designed to implement the RED method, the evolution of the genes encoding the 20 aminoacyl-tRNA synthetases was examined. Although LGTs were common in the evolution of the aminoacyl-tRNA synthetases, they were not sufficient to obscure the organismal phylogeny. Moreover, much of the apparent complexity of the gene tree was consistent with the formation of the paralogs in the ancestors to the modern lineages followed by more recent loss of one paralog or the other.

Key words: lateral gene transfer, LGT, aminoacyl-tRNA synthetases, ribosomal proteins, ratios of evolutionary distance, RED, RED-T.

Introduction

Recent advancements in genomic research and comparative genome studies have proposed a prominent role for lateral gene transfer (LGT) in the evolutionary history of prokaryotes (Ferretti, et al. 2001; Kuroda, et al. 2001; Perna, et al. 2001). The role of LGT in prokaryotic evolution has been discussed since the 1930's (Eisen 2000; Gibbons 1992; Gray 1992; Margulis 1993), and the frequency has been a subject of great interest (Doolittle 1999; Glansdorff 2000; Gogarten et al. 1996; Jain et al. 1999; Lawrence and Ochman 2002; Martin 1999). LGT may have been one of the major evolutionary mechanisms that led to the formation of the modern lineages of prokaryotes (Snel et al. 2002). For instance, Lawrence and Ochman proposed that 17.8% of the *Escherichia coli* genome may have been acquired by LGT (Lawrence and Ochman 1998). Similarly, others found that 25% of the genome of a hyperthermophilic bacterium, *Thermotoga maritima*, has high similarity with the distantly related archaeon *Pyrococcus horikoshii* and not with close relatives, thus suggesting a lateral gene acquisition across domains (Nelson et al. 1999). With the growth of genomic data and more advanced computational tools, even more LGT findings are likely to be reported, thus providing greater opportunities to study the importance of this evolutionary mechanism.

Given the limitations of existing methods to test for LGT, we sought to develop an alternative method. For instance, one common method looks for branching inconsistencies between phylogenetic trees of various genes that could be explained by LGT (Smith et al. 1992). However, potential ambiguities of phylogenetic trees may limit the usefulness of this approach. Moreover, if LGTs have actually occurred, it should be

possible to discover them by multiple methods. Therefore, an alternative method may provide additional confidence in LGTs. With this spirit, we evaluate the potential of scatter plots to identify LGTs. While this approach also relies upon phylogenetic analyses to determine evolutionary distances (E_d), it uses empirical tests rather than tree building to identify non-vertical evolution. It assumes that if the rates of evolution in two genes are constant, then plots of the E_d for one gene against another gene should be linear and the ratios of the E_d's, or RED, should be 1.0.

For instance, in a hypothetical tree including bacteria, archaea, and eukaryotes, the evolutionary relationship among genes in the absence of LGT are indicated by the solid lines in Figure 2.1. The accompanying plot shows a correlation in the E_d for any pair of genes with the same evolutionary history (Figure 2.1B, ●). If a LGT occurred between the archaea and the bacterial sub-group B1, as imagined by the broken arrow in Figure 2.1A, the E_d between the bacterial subgroups B1 and B2 would increase to above the diagonal (Figure 2.1B, ○) The distance "i" would reflect the distance of the donating taxon from bacterial subgroup B2 (Figure 2.1B). Similarly, the inter-domain comparisons of genes from bacterial subgroup B1 with genes from the archaeal or eukaryotic domains would decrease to below the diagonal. The E_d's will be expected to be smallest to the closest relatives of the donor taxon, indicated by "ii" in Figure 2.1B. In order to evaluate this approach, the RED method was first standardized by comparisons of the ribosomal proteins, which were believed to be primarily vertically inherited (Hansmann and Martin 2000). It was then used to test for LGT events in the evolution of aminoacyl-tRNA synthetases. As key components in translation, these functionally conserved proteins recently have received great attention. On the basis of phylogenetic tree analyses, LGTs

have been proposed for the evolution of many members of this ancient family (Boucher et al. 2001; Lamour et al. 1994; Li et al. 1999, Woese et al. 2000).

Materials and Methods

Data selection. Gene sequences were extracted from the WIT/ERGO database (Overbeek et al. 2000). In the initial analysis, groups of homologous gene sequences or COGs that were well represented in the genomes were identified using automated ORF-clustering algorithms (Overbeek et al. 2000). Among these genes, 40 ribosomal proteins and 20 aminoacyl-tRNA synthetases were selected for further study, and additional homologs were obtained from other search strategies including screening annotations and BLAST searches. By the end of this process genes from a total of 44 species were found.

To screen for errors in the sequences, RED plots were generated for each gene (see Results, and Appendix C and D), and values that were outside of the diagonal, or outliers, were examined in detail. For the outliers, sequence length, functional assignment, and alignments were checked. In some cases, the length of the sequence affected E_d calculations. For example, many sequences from the eukaryotes *Arabidopsis thaliana* and *Zea mays* as well as the archaeon *Sulfolobus solfataricus* were either much shorter or longer, respectively, than the prokaryotic sequences. The short sequences probably resulted from incomplete sequences, and the longer sequences probably resulted from frame shift errors during sequencing. Because of the poor quality, sequences from

these organisms were removed from subsequent analyses. In some cases, the genes from some pairs of organisms, such as *Mycobacterium bovis* and *Mycobacterium tuberculosis*, were so closely related that the E_d's were close to zero for all of the genes analyzed. These comparisons produced outliers because dividing by values close to zero caused a high variability. For that reason, the sequences from one of the pair of organisms was removed from further analyses. This lead to the removal of sequences from *Mycobacterium bovis* and *Mycoplasma capricolum*. As a result, the maximum number of species for each gene was eventually reduced from 44 to 39 – *Aeropyrum pernix*, *Actinobacillus actinomycetemcomitans*, *Aquifex aeolicus*, *Archaeoglobus fulgidus*, *Bacillus subtilis*, *Bordetella pertussis*, *Borrelia burgdorferi*, *Campylobacter jejuni*, *Chlamydia pneumoniae*, *Chlamydia trachomatis*, *Clostridium acetobutylicum*, *Chlorobium tepidum*, *Synechocystis* sp., *Deinococcus radiodurans*, *Escherichia coli*, *Enterococcus faecalis*, *Haemophilus influenzae*, *Helicobacter pylori*, *Methanobacterium thermoautotrophicum*, *Methanococcus jannaschii*, *Mycobacterium tuberculosis*, *Mycoplasma genitalium*, *Mycoplasma pneumoniae*, *Neisseria gonorrhoeae*, *Neisseria meningitidis*, *Porphyromonas gingivalis*, *Pseudomonas aeruginosa*, *Pyrobaculum aerophilum*, *Pyrococcus abysii*, *Pyrococcus horikoshii*, *Rhodobacter capsulatus*, *Rickettsia prowazekii*, *Saccharomyces cerevisiae*, *Streptococcus mutans*, *Streptococcus pneumoniae*, *Streptococcus pyogenes*, *Thermotoga maritima*, *Treponema pallidum*, and *Yersinia pestis*.

This problem was not observed for all pairs of closely related organisms tested. For example, the the results of comparisons of the two *Neisseria* gene sequences were

only occasional outliers. Additionally, comparisons of the proteobacteria members, such as *Escherichia* to *Yersinia* genes, were never a source of high variability.

While high organism representation was beneficial, the diversity within and across the three domains and sub-domains was also important to this study. Within the 39 organisms, 31 bacterial and seven archaeal species were represented. Further, they represented 10 bacterial and two archaeal phyla. However, due to the lack of available eukaryotic genomes with reliable ORF assignments at the time that these studies were performed, only one eukaryote was used, *Saccharomyces*.

Sequences were also screened for errors specific to each gene. These errors, illustrated as outliers, included misalignments, possession of incorrect sequences, mistakes in the functional assignment of an ORF or incorrect retrieval of a sequence by the search algorithm. The ORFs producing these errors were identified, removed, and/or replaced prior to further analysis. For example, the search algorithm occasionally retrieved mitochondrial instead of the cytoplasmic sequence for yeast. In these cases, the mitochondrial sequence was replaced with the cytoplasmic homologue.

Analyses depiction. Initially, for the development and analyses of the RED method the scatter plots were generated and analyzed on Microsoft Excel 2000. For a more complete RED analysis of the genes, an interactive research tool entitled RED-T – Ratios of Evolutionary Distances for determination of alternative phylogenetic events– was developed in the Java™ programming language (*submitted*). RED-T is a graphical user interface (GUI) capable of generating scatter plots from given distance matrices to analyze evolutionary relationships among various levels of taxa. In addition to allowing

the user to further analyze any of the 60 genes discussed here, it is fully capable of importing new genes for comparison with our control, or allowing the user to develop a new control. The final phylogenetic analyses were conducted using the RED-T application. The latest version of RED-T, along with all genes reported here, help files, tutorial and source code are available at http://www.arches.uga.edu/~whitman/RED (Also see Appendix E).

All alignments were generated at the WIT/ERGO site using the ClustalW program (Thompson et al. 1994). Other tools included ClustalX 1.8, a windows interface that allowed multiple sequence alignment analysis using different protein weight matrices, and the programs of the PHYLIP package 3.x-3.6, which were used to generate phylogenetic inference data (Felsenstein 1993; Thompson et al. 1997). Evolutionary distances (E_d) were calculated by the ProtDist program using the PAM-t matrix (Felsenstein 1993). Gene trees that were used to compare with the RED plots were constructed by first using the Fitch-Margoliash method, executed by the PHYLIP's FITCH program, to generate Newick-Standard formatted data (Fitch and Margoliash 1967). This data was imported by NJ-Plot and TreeExplorer 2.12 applications to display and configure each tree (Perriere and Gouy 1996; Tamura 1993).

Quantitative variables. For each of the 60 experimental genes that were analyzed, the average sequence length in number of codons (l_s), number of comparisons (n_s), mean of *K* (*m*), and relative standard deviation of *K* (*rSD*) or the standard deviation of *K* divided by *m* was calculated for each intra-domain and inter-domain comparison (supplementary

Tables S1-4). For each pair of organisms, K was the ratio of E_d of the experimental gene divided by the mean E_d of the control genes.

To determine if m values were within the expected range, they were compared to the values m'_{AA}, m'_{AE}, m'_{BE} (supplementary Tables S1-4). Direct comparisons of the m values of the archaeal and the bacterial intra-domain comparisons were complicated by differences in the number of archaeal and bacterial genes. Because the intra-domain archaeal m_{AA} was calculated from a relatively small number of taxa, it was expected to be more sensitive to the specific taxa represented and more variable. The value m'_{AA}, was calculated for comparison of m_{AA} to m_{BB}, the mean of K for the intra-domain comparisons of the bacteria. It was calculated by taking 100 random samples of bacteria equivalent in number to the number of archaea. For each sample, m was calculated. The values m'_{AA} and SD'_{AA}, were then the mean and standard deviation of these values. When m_{AA} was not significantly different from m'_{AA}, it was also possible to calculate m'_{AE} and m'_{BE}.

Similarly, m'_{AE} was calculated to control for the effect of the n_s for the Archaea to Eukaryote comparison. In this case, the inter-domain comparisons between the archaeal and the eukaryote m_{AE} was based upon a comparison of a small number of Archaea to one Eukaryote. For the calculation of the m'_{AE} and SD'_{AE}, the archaeal genes were compared to each of the bacterial genes. Similarly, for evaluation of the inter-domain comparisons between bacteria and the eukaryotic m_{BE}, m'_{BE} and SD'_{BE} were calculated to control for the effect of n_s for the comparison of the bacteria to only one eukaryote. In this case, each archaea was compared to all of the bacteria.

Results

Designation of control proteins. Our hypothesis stated that when the evolutionary distances (E_d) for sets of genes with the same evolutionary history were plotted against one another, a linear plot would be generated. In order to test this hypothesis, 32 genes with a wide distribution among the genomic sequences were initially examined (See Appendix A). Six genes that yielded linear plots with this preliminary data sets were identified: glutamyl-tRNA synthetase, leucyl-tRNA synthetase, protein translocase SecY, and ribosomal proteins L2P, L13P, and S13P (data not shown). However, visual examination of the plots indicated a high variability and a high relative standard deviation (*rSD*) of the value *K*, which was the ratio of the E_d values for the pairs of genes. Because a stochastic error was associated with using sequence similarity to measure E_d, an attempt was made to reduce this error by averaging the E_d from more than one gene. In the first experiment, the E_d values from glutamyl-tRNA synthetase were chosen to represent the experimental gene. The control gene was represented by the average of all possible combinations for the other five genes. As shown in Figure 2.2A, the *rSD* for the glutamyl-tRNA synthetase decreased from 0.200 to about 0.130 as the number of control genes increased from one to five. Removal of ribosomal protein L13P genes to decrease the ribosomal representation, removal of all archaeal genes, replacing the glutamyl-tRNA synthetase with ribosomal protein L2P or other experimental genes had little effect on this relationship (Figure 2.2A and data not shown). In summary, as the number of genes used to calculate *K* increased, the *rSD* decreased. The largest effect was seen when the control was based on three or fewer genes. Therefore, an average of three E_d's was

chosen for subsequent studies as a compromise between the desire to lower the variability and the desire to use the lowest number of controls possible to insure that all the controls would be present in any particular genomic sequence.

The actual three genes chosen for the standard were L2P, leucyl-tRNA synthetase and SecY. Of the six original candidates, these genes were chosen in part because they represented different functional groups. Moreover, upon expansion of the data set to 39 genomes, the *K* values for these genes all possessed low *rSD*s. For further validation, phylogenetic trees for the selected 39 species for each of the three genes were very similar to each other as well as to that of the 16S rRNA genes, indicating that these genes all shared a common evolutionary history (See Appendix B). Lastly, an alternative method of combining the E_d's for the control genes was explored. The three sequences for each of the 39 organisms were concatenated, aligned and E_d's calculated (data not shown). The overall *rSD*s of *K* from the concatenated control genes were very similar to the values obtained by averaging.

Calculation of E_d. Initial studies utilized pairwise alignments because they enabled the easy addition and removal of taxa. However for some genes, pairwise alignments generated E_d values that were significantly different from those calculated from multiple alignments. In these cases, the E_d values from the pairwise alignments yielded higher *rSD* for *K* and were assumed to be incorrect. Thus, multiple alignments were used in the analyses shown here.

For most experiments, multiple alignments were generated with ClustalW, and E_d values were calculated with the PAM-t matrix. To evaluate other E_d matrices, BLOSUM,

Gonnet and PAM were used to generate multiple alignments and E_d for each control gene (Benner et al. 1994; Dayhoff et al. 1972; Dayhoff et al. 1978; Henikoff and Henikoff 1992). For comparisons with E_d values greater than 1.0, minor differences were found using any of these methods. When the E_d was less than 1.0, more variability was observed, but it was not systematic. Therefore, the method for generating multiple alignments and E_ds appeared to have little affect.

Standardization. LGT is believed to be rare in the ribosomal proteins (Matte-Tailliez et al. 2002). Therefore, these genes were used to evaluate this approach and to standardize the criteria for recognizing LGT in other genes. Within the bacteria, 39 genes for the small and large ribosomal proteins were examined, and 37 yielded linear RED plots (See Appendix C). Large ribosomal proteins L31 and L33 possessed nonlinear RED plots and high *rSD* values, 0.095 and 0.210, respectively. The large scatter apparent among these comparisons suggested that LGT may have occurred. Hansman and Martin (2000) have previously reported that L33 was poorly conserved and likely to have participated in a LGT. For these reasons, both L31 and L33 were excluded during the standardization.

The effects of number of comparisons (n_s) and sequence length (l_s) on *m* and *rSD* were analyzed. The *rSD* was not significantly correlated with n_s, and l_s was not significantly correlated with *m* (data not shown). This result suggested that these variables were independent of each other. In addition, the *rSD* was not significantly correlated with l_s for inter-domain comparisons. However, the *rSD* was correlated with l_s for intra-domain comparisons, especially for short ORF's (Figure 2.2B). To correct for this effect in subsequent analysis, the measured intra-domain *rSD* was used to calculate a

new variable, D_{rSD}. This value was the difference between the measured *rSD* and the *rSD* expected from the l_s. The expected *rSD* was calculated from the trend-line between the measured *rSD*s and l_s (Figure 2.2B). It appeared that all 37 genes possessed D_{rSD} values of ≤ 0.06.

In addition, the distribution of the *K* values was examined. If the distribution was normal, then the *rSD* would be a good predictor for the distribution. A positive kurtosis was observed for the distribution of *K* in most of the genes, which implied that the distribution was narrower than expected from the normal distribution (Snedecor and Cochran 1967). Skewness was also observed in some genes, but no obvious patterns were noted. While these results indicated that the distribution of *K* values was not normal, it did not invalidate the use of the *rSD* as a predictor.

Inter-domain comparison values were also analyzed. Of the 37 genes examined, 19 possessed archaeal homologs. These genes possessed *rSD* values of ≤ 0.135 for inter-domain comparisons between the archaea and the bacteria. A similar threshold was not set for the inter-domain comparisons between the archaea and the eukaryote and between the bacteria and the eukaryote because of the low number of comparisons. For 12 of the 19 genes, the intra-archaeal comparison m_{AA} was similar to the intra-bacterial comparisons and m'_{AA} (supplementary Table S1). This result suggested that the archaeal and bacterial domains shared similar rates of evolution.

RED Models

RED Model I: Absence of LGT – S19P. Analyses of many genes failed to detect evidence of LGT. One example was the small subunit ribosomal protein S19P (Figure 2.3). For this gene, RED plots were linear, and the *m* values were not significantly different for the intra-domain and inter-domain comparisons. First, the intra-bacterial D_{rSD} (0.029) was below the threshold for nonlinear relationships, 0.060, which suggested that none of the bacterial taxa represented were recipients of a LGT (supplementary Tables S1 and S2). The intra-archaeal comparison had an m_{AA} value of 0.571, which was not significantly different from the values of the intra-bacterial comparisons and m'_{AA} = 0.587 ± 0.047 (supplementary Table S1). Similarly, the inter-domain comparisons between archaea and eukaryotes had an m_0 value of 0.607, which was not significantly different from the m'_{AE} value of 0.568 ± 0.044. The bacteria-eukaryote inter-domain comparisons had an m_{BE} value of 0.621, which was also within the range of expected values, i.e. m'_{BE} equaled 0.566 ± 0.048. Lastly, the *rSD* measurements for all of the inter-domain comparisons were below the nonlinear threshold.

Like the S19P gene, no evidence for LGT was found for the other large and small ribosomal protein genes tested, except for L31P and L33P (see above). Since only bacterial L31P and L33P genes were examined, it was not possible to deduce inter-domain relationships. However, it was clear that D_{rSD} values, 0.095 for L31P and 0.210 for L33P, were greater than the threshold for nonlinear relationships, suggesting some of the bacteria were involved in LGT (supplementary Tables S1 and S2). Among the tRNA

synthetases, no evidence was found for LGT for the aspartyl-tRNA synthetases and the alpha and beta chains of the glycyl-tRNA synthetases (Also see Appendix D).

RED Model II: Inter-domain LGT – Valyl-tRNA synthetase. Analyses of many other genes indicated that LGT may have occurred. Two of the most common types of apparent LGTs were observed among the valyl-tRNA synthetase genes (Figure 2.4A). First, for the intra-domain comparisons, the intra-bacterial D_{rSD} value (0.122) was above the threshold for nonlinear relationships, suggesting that a bacterial taxon was the recipient of a LGT. The high *K* values resulted from comparisons of the *Rickettsia* gene with that of other bacteria, suggesting that a recent ancestor of *Rickettsia* obtained its gene from another domain. The source of the gene appeared to be the Euryarchaeota because this group had much lower *K* values, about 0.54, than comparisons between the Crenarchaeota and *Rickettsia* (0.64) or between all the Archaea and the other bacteria (0.79). This hypothesis was supported by the gene tree where *Rickettsia* appeared within the euryarchaeotic clade (Figure 2.4B) and was similar to the reports by others (Andersson et al. 1998; Woese et al. 2000). This transfer was an example of an apparent LGT from one domain to a relatively modern lineage in another domain.

A second type of common nonlinearity in RED plots resulted from comparison between domains. In this case, the mean of K values for the inter-domain comparisons were 0.774 and much lower than the values from the intra-archaeal and intra-bacterial comparisons, which were 1.20 and 1.19, respectively. Three general types of explanations seemed likely. One, the difference could be an artifact resulting from the method in which the evolutionary distances were calculated (Benner et al. 1994). Two, the rate of

evolution of this gene could have been slower prior to the formation of the domains. Three, a LGT may have occurred in the lineages ancestral to formation of the domains (Figure 2.4C at T2).

The first possibility was that the nonlinearity resulted from a systematic error in the calculation of the evolutionary distances because both moderately and distantly related genes were compared with the same algorithm. To test this point, three different BLOSUM matrices were used to generate multiple alignments. These alignments were then used to generate E_ds from different PAM or JTT matrices. All combinations were calculated, and the distances were compared. With the exception of closely related taxa, the differences in evolutionary distances were small and the variations were not systematic (data not shown). These analyses did not support the hypothesis that the nonlinearity resulted from the method of distance calculations. Because it was not possible to distinguish between the other two possibilities, this common occurrence was regarded as an ancient nonlinearity.

Although these analyses supported a LGT from the euryarchaeotes to an ancestor of *Rickettsia*, it did not support other LGTs suggested by the phylogenetic tree. For instance, as observed previously *Helicobacter* and *Campylobacter* failed to cluster with the remaining proteobacteria in the gene tree (Figure 2.4B, Woese et al. 2000). However, comparisons of these genes to those of other bacteria had *K* values of 1.19, which was similar to other intra-bacterial comparisons (m_{BB} = 1.192, supplementary Table S1). Therefore, these genes did not appear to have participated in a LGT event.

Analyses of other aminoacyl-tRNA synthetases genes provided further evidence for LGT from one domain to relatively modern lineages of other domains. The genes for

the alpha and beta chains of phenylalanyl-tRNA synthetase in the spirochetes, *Borrelia* and *Treponema*, appeared to have been received from the archaea, specifically a relative of *Pyrococcus* spp. Likewise, these spirochetes were the apparent recipients of the seryl-tRNA and methionyl-tRNA synthetase genes from the eukaryotic domain. The methionyl-tRNA synthetase genes of both *Chlorobium* and *Porphyromonas* appeared to be derived from the archaeal domain, specifically *Pyrococcus* spp. Similarly, methionyl-tRNA synthetases from the alpha proteobacterium *Rhodobacter* appeared to have been derived from spirochetes, while the same enzyme from another alpha proteobacterium *Rickettsia* was derived from *Mycobacterium* (Also see Appendix D). In addition, the ancient nonlinearity was evident for 10 of the 16 tRNA synthetase genes where homologs were identified in both domains (Table 1).

RED Model III: Ancient LGT – Isoleucyl-tRNA synthetase. Evidence for ancient LGTs between domains was also found among the isoleucyl-tRNA synthetase genes of bacteria. Support for this conclusion included nonlinear plots of the isoleucyl-tRNA synthetase genes and deviations from linearity in both the intra-domain and inter-domain comparisons (Figure 2.5A). For the intra-domain comparisons, the intra-Bacteria's D_{rSD} value (0.157) was above the threshold for nonlinearity. High *K* values resulted from comparisons between bacterial clades B1 and B2, suggesting that one of these clades obtained its gene from another domain. The *rSD* value of archaeal-bacterial inter-domain comparisons was 0.105 or below the threshold for nonlinearity (0.135), suggesting that the archaea was not the donating domain. In contrast, the high *rSD* for inter-domain comparisons between the bacteria and the eukaryote of 0.227 implicated the eukaryotes

as the donor. Because, the eukaryote and bacterial clade B2 comparisons had much lower *K* values, about 0.438, than the eukaryote and bacterial clade B1 comparisons (0.755), bacterial clade B2 was the likely recipient (Figure 2.5C, T1). The gene tree supported this assignment, where the eukaryote *Saccharomyces* appeared within the bacterial clade B2 (Figure 2.5B). In accordance, Brown et al. analyses proposed the acquisition of mupirocin-resistant isoleucyl-tRNA synthetase gene by *Staphylococcus aureus* was from eukaryotes (Brown et al. 1998). Although the *S. aureus* gene suggested by Brown et al. was not shown in Figures 2.5A and 2.5B, it clustered with other B2 genes in other analyses (data not shown). Interestingly, representatives of both the proteobacteria and the firmicutes were found in both bacterial clades. These results suggested that the LGT predated the formation of these lineages and that a common ancestor possessed both the eukaryotic and bacterial genes. The current distribution would then have resulted from the loss of one gene or the other in the modern lineages.

Analyses of other aminoacyl-tRNA synthetases also provided evidence for additional ancient LGTs. In the analysis of the asparaginyl-tRNA synthetases, the genes from some of the proteobacteria and firmicutes appeared to be derived from the Archaea. Among the proteobacteria, the recipients included *Actinobacillus* but not *Escherichia* and *Haemophilus*. It also included all the firmicutes except for *Clostridium acetobutylicum.* In a second case, a bacterial clade consisting of *Borrelia*, *Clostridium*, *Porphyromonas* and *Rhodobacter* appeared to have acquired the histidyl-tRNA synthetase gene from yeast. In a separate event, *Aeropyrum*, a member of the sub-domain Crenarchaeota, also appeared to be a recipient of the yeast gene. The latter observation was supported by the

low *K* values for the inter-domain comparisons of both *Aeropyrum* and the bacterial clade to the yeast gene (Also see Appendix D).

Lastly, the bacteria appeared to have acquired the alanyl-tRNA synthetase genes from the eukaryotic domain prior to the radiation of the modern lineages. Hence, the eukaryote form was present in all bacteria examined. This conclusion was based upon the following observations. The intra-domain values of m_{AA} and m_{BB} were the same, indicating that the gene has evolved at the same rate in the modern lineages. Similarly, the inter-domain comparison between the archaea and the eukaryote, m_{AE} was 1.43 and close to the value of m_{AA} of 1.15, which suggested that the rate of evolution was constant within these taxa. In contrast, the values m_{AB} and m_{BE} for comparison of the bacteria to the archaea or the eukaryote were much lower, 0.90 and 0.47, respectively. Because m_{BE} was lower than m_{AB}, the eukaryote appeared to be the donor to the bacteria. All of the bacteria have the same form of the gene, therefore, this transfer occurred prior to the formation of the modern bacterial lineages. The alternative that the bacteria donated the gene to the eukaryote was precluded by the high value for m_{AE} (Also see Appendix D).

RED Model IV: Multiple Inter-domain LGT – Prolyl-tRNA synthetase. Evidence for a complex series of LGT events was obtained for some genes. The RED plot of the prolyl-tRNA synthetase genes indicated deviations from linearity in both the intra-domain and inter-domain comparisons (Figure 2.6A). For the intra-domain comparisons, the intra-bacterial D_{rSD} value of 0.545 was well above the threshold for nonlinearity. Also, high *K* values were observed for comparisons between the genes of bacterial clade B1 and B2 (Figure 2.6A,B), suggesting that these clades obtained their genes from

different domains. The B1 clade consisted of *Porphyromonas* and *Deinococcus* in addition to *Borrelia*, and the B2 clade consisted of the remaining bacterial representatives. It is also interesting to note that the two spirochete genes, *Borrelia* and *Treponema*, are separated in the two bacterial clades. Thus, it was likely that a common ancestor possessed both forms of the gene. Further, the *rSD* values of inter-domain comparisons between archaea and bacteria, and between bacteria and eukaryote were 0.224 and 0.482, respectively, and also above the threshold for nonlinearity (0.135). These results suggested that all three domains played roles in separate LGT events. First, the inter-domain comparisons between the eukaryote and bacterial clade B2 had much lower *K* values, about 0.622, than comparisons between the eukaryote and bacterial clade B1 (1.662). This result indicated a LGT between the eukaryote and the bacterial clade B2 but did not indicate the direction of transfer. The direction of transfer to yeast was indicated by the high *K* values (2.253) for inter-domain comparisons between the archaea and the eukaryote. These values greatly exceeded the *K* values found within each clade of about 1.1 and were consistent with an ancestor of yeast obtaining the gene from bacterial clade B2 (Figure 2.6C, T1). In addition, yeast contained a second gene for prolyl-tRNA synthetase (Woese et al. 2000). This other gene was closely related to the archaeal genes, suggesting that it was the ancestral eukaryotic gene (data not shown). Second, the inter-domain comparisons between the archaea and bacterial clade B1 had much lower *K* values, about 0.671, than comparisons between the archaea and bacterial clade B2 (1.321). This result suggested a second inter-domain LGT event in which the archaea were the donor to bacterial clade B1 (Figure 2.6C, T2). The hypothesis was consistent with the gene tree in which the eukaryotic gene appears within the bacterial clade B2 and

the archaeal genes within the bacterial clade B1 (Figure 2.6B). However, the gene tree did not provide evidence for the directions of transfer.

Similar to the prolyl-tRNA synthetase model, analysis of threonyl-tRNA synthetase suggested that multiple LGT events have occurred (data not shown). Bacteria appear to have donated the genes to both the Crenarchaeota and the eukaryote. The intra-archaeal *rSD* value was above the nonlinear threshold, which was the result of the comparisons between the Crenarchaeota and Euryarchaeota. Further, comparisons of the Crenarchaeota with the bacteria had low *K* values, which indicated that the Crenarchaeota were recipients of this gene from the bacterial domain. Also, the low *K* values for the inter-domain comparisons between the bacteria and the eukaryote argued for a second LGT from Bacteria to the eukaryote (Also see Appendix D). These conclusions were consistent with the gene tree, which also found a close evolutionary relationship between the eukaryotic and the bacterial genes (Woese et al. 2000).

RED Model V: Rapid Expansion of a Gene Family in the Absence of Modern LGT – Cysteinyl-tRNA synthetase. These genes were an example in which phylogenetic analyses of gene trees had suggested that multiple LGTs had occurred (Li et al. 1999; Woese et al. 2000), but these transfer LGTs were not supported by RED analyses (Figure 2.7A). In addition, the gene tree made from the same data as the RED plot suggested a LGT as it grouped the yeast gene with some of the proteobacteria (Figure 2.7B). First, RED's intra-domain comparisons did not support the hypothesis that any of the archaeal or bacterial genes were derived from another domain. Although the intra-archaeal *rSD* value of 0.330 was higher than the threshold for nonlinearity, this discrepancy was due to

the low E_d of the comparisons of the two *Pyrococcus* genes. Great variability in *K* had previously been observed in comparisons of closely related taxa where the E_d values were close to zero. When this comparison was removed, the intra-archaeal *rSD* value was 0.069 and, like the intra-bacterial D_{rSD} (0.010) values, below the threshold for nonlinear relationships. In addition, the *K* values for the intra-archaeal and intra-bacterial comparisons were similar (supplementary Table S1 and S2). In contrast, the *K* values for the inter-domain comparisons (0.849, 0.479 and 0.479) were much lower and suggested either a modern origin for this gene with LGT to a number of lineages within a short period of time or a change in the rate of evolution (Figure 2.7C). If this gene was of modern origin, presumably an enzyme with the same activity existed prior to its emergence. Circumstantial evidence appears to support this latter interpretation. Because some archaea do not contain a recognizable cysteinyl-tRNA synthetase, there must be at least one other gene family with this activity (Fabrega et al. 2001; Stathopoulos et al. 2000).

RED Model VI: Intra-domain LGT – Glutamyl-tRNA synthetase. Evidence for an intra-domain LGT within the glutamyl-tRNA synthetase genes was also found by this method. Although the initial analysis with a smaller number of organisms were linear (see above), nonlinear plots were observed with the full data set in the intra-domain but not in the inter-domain comparisons (Figure 2.8A). For the intra-domain comparisons, the intra-bacterial D_{rSD} value (0.067) was above the nonlinear threshold of 0.060, suggesting that a bacterial taxon was the recipient of a LGT. The high *K* values resulted from comparisons of *Pseudomonas* with members of the delta and gamma-

proteobacteria, suggesting that *Pseudomonas* or a recent ancestor obtained its gene from another source. The *K* value between the archaea and *Pseudomonas* genes was within the range expected, and the *rSD* value of archaeal-bacterial inter-domain comparisons was 0.071 or below the nonlinear threshold. Therefore, the archaea did not appear to be the donor. In contrast, the intra-domain comparisons between *Pseudomonas* and the *Chlamydia* had a lower mean of *K* value (1.074) than comparisons to other bacteria (1.381). This result suggested that an ancestor of *Chlamydia* was the donor for the *Pseudomonas* gene (Figure 2.8C, T1).

Evidence for intra-domain LGTs was obtained for other genes. In the arginyl-tRNA synthetases genes, evidence for three different LGT events was observed. One was an intra-domain gene exchange from *Rhodobacter* to *Bacillus*. The second LGT was a transfer from a eukaryote to some members of the firmicutes, *Enterococcus* and *Streptococcus* spp. The third LGT was to the bacterium *Deinococcus* from an archaeon. The tryptophanyl-tRNA synthetase gene has also undergone both intra-domain and inter-domain LGTs. Within the bacterial domain, *Deinococcus* donated this gene to *Streptococcus* spp. A second LGT was observed from the eukaryote to an ancestor of the archaea *Pyrococcus* spp. and *Pyrobaculum*. In addition, the E_d between the beta and gamma-proteobacteria was greater than predicted. However additional evidence for a LGT was not found. The *K* values for comparisons to other bacteria suggested a normal evolutionary history and no evidence was found for a donor (Also see Appendix D). A probable explanation was that there was a duplication of this gene in an ancestor to the beta and gamma-proteobacteria, and one lineage retained one copy and the other lineage

retained the other, resulting in higher than expected divergence within this group. The published gene tree did not observe this event (Woese, et al. 2000).

RED Model VII: Ancient Divergence– Tyrosyl-tRNA synthetase. In contrast to LGT, this model proposes that an alternative, the divergence of multiple gene copies, occurred during the evolution of the tyrosyl-tRNA synthetase. The RED plot of these genes showed deviations from linearity in intra-domain comparisons, and the intra-bacterial D_{rSD} value of 0.380 was above the threshold for nonlinearity (Figure 2.9A). Also, high *K* values were observed for comparisons between the genes of bacterial clade B1 and B2 (Figure 2.9A,B). Although these results could suggest that these clades obtained their genes from different domains, the inter-domain comparisons to both bacterial clades did not support this conclusion. They had mean of *K* values similar to or slightly above that of the intra-domain comparisons of each bacterial clade to itself. These results excluded these domains as potential donors to one of the bacterial clades. Without an apparent donor, LGT was unlikely. These results were more consistent with a model in which an ancestor possessed two types of the tyrosyl-tRNA synthetase genes that diverged from a single gene (Figure 2.9C). This conclusion was supported by the observation that genes of both clades are found in the same bacterial phyla. For example, the B2-type was found in some members of the gamma-proteobacteria – *Actinobacillus* and *Haemophilus* – and beta-proteobacteria – *Bordetella*. In contrast, the B1-type was found in other members of the gamma-proteobacteria – *Escherichia*, *Pseudomonas*, and *Yersinia* – and beta-proteobacteria – *Neisseria*. Likewise, the groups of sister taxa, *Porphyromonas* and *Chlorobium* as well as *Bacillus* and *Clostridium*, also possessed both

types of this gene. In fact, both types of the tyrosyl-tRNA synthetase gene were found in *B. subtilis* and *C. acetobutylicum*, which confirmed that was it plausible for an ancestor to have both copies (data not shown, Woese et al. 2000).

Further RED plot analysis suggested that the gene type possessed by bacterial clade B1 was divergent from the bacterial clade B2, the archaea and the eukaryote genes. This difference had been previously proposed to be due to the loss of non-essential regions of the sequence (Ibba and Soll 2001). The distinction between the two gene types was illustrated by the mean of *K* values. The inter-domain comparisons (*K*) between bacterial clade B1 and the archaea or the eukaryote were slightly higher than the values found in comparisons between bacterial clade B2 and the archaea or the eukaryote.

Discussion

A major goal of the current work was to develop and evaluate a novel method for detecting LGTs. Although phylogenetically based, this method does not rely upon calculation of trees. Instead it utilizes the ratios of evolutionary distances to distinguish between alternative evolutionary histories. In this fashion, it tests whether or not the experimental gene shares the same evolutionary history as the control genes. When the evolutionary histories are different, LGT is one possible mechanism. However, any mechanism that causes changes in the evolutionary clock, such as changes in gene function or evolutionary rate, could in theory be detected. The advantages of this

approach are: the variability inherent in comparing protein sequences is transparent, the direction of LGT and the relative rates of evolution are readily identified, and it is possible to detect other types of evolutionary events.

Because of its simplicity, the distribution of the ratios of evolutionary distances can be determined empirically by examination of sets of proteins believed to share a common evolutionary history, such as the ribosomal proteins. It is then possible to set thresholds for ratios that are outside the range of values expected for genes with a common evolutionary history. The ratios method also adds an additional criterion for the recognition of LGT. The genes of the recipient taxa must be both significantly closer to the genes of the donor taxa and significantly further from the genes of their sister taxa. The application of both of these criteria facilitates the rejection of many potential LGTs suggested by discrepancies in trees. The direction of transfer is also inherent in these criteria. For instance, two deep bacterial lineages were observed in the phylogenetic tree of the prolyl-tRNA synthetase. The bacterial lineage B1 was associated with the archaeal genes and SC1, one of the two yeast genes. The other bacterial lineage B2 was associated with SC2, a second yeast gene. This tree was consistent with two possible LGT events. In the first possibility, a LGT occured from B2 to the eukaryotes as well as from the archaea to B1. In this case, SC2 would be the derived gene and SC1 would be the ancestral eukaryotic gene. In the second possibility, a LGT occured from B1 to the archaea and from the eukaryote to B2. In this example, SC2 would be the ancestral gene and SC1 would be derived. In contrast to the phylogenetic tree, the ratio method provided clear evidence in support of the first possibility. Lastly, the evolutionary rates are portrayed by

the mean of *K* values. Therefore, rate of evolution can be readily compared within and between taxonomic groups.

As early as the mid 1970's, Reanney and Sonea suggested that all organisms on the planet could be viewed as a single entity, or a "global super-organism," due to the numerous genetic exchange platforms, such as bacteriophages, transposable elements and plasmids (Hilario and Gogarten 1993; Reanney 1978; Sonea and Paniset 1976). The emergence of genomic data appeared to support this hypothesis and led to the suggestion that the evolutionary history of organisms might be likened to a "net" or web due to the high frequency of LGT (Doolittle 1999). Thus, the evolutionary history of the individual genes would vary due to LGT and there would be no consensus representing the organismal evolution. In the extreme, LGT may have erased the deep ancestral record of organismal evolution an invalidated attempts to create a universal tree of life based upon rRNA sequences (Nesbo et al. 2001; Woese 2002). Our analyses of the aminoacyl-tRNA synthetases indicated that the high frequency of LGT theory is an overestimation, and we observed a moderate frequency of LGTs that only partially obscured the organismal phylogeny.

If LGTs occurred at very high frequencies, then the apparent rate of evolution among different phylogenetic groups would differ. Our mean of *K* analyses suggested the opposite. For example, the rates of evolution or means of *K* among prolyl-tRNA synthetases of the proteobacteria and spirochetes were constant, even though both groups were involved in separate LGT events. Moreover, for many of the aminoacyl-tRNA synthetases the ancestors for the modern lineages appeared to contain multiple forms of the gene, only one of which was retained in modern organisms. This series of events

produced a complex distribution with members of the same phylogenetic groups containing different forms of the gene. Although a similar outcome might be predicted by a high rate of LGT, this possibility was eliminated by showing a constant rate of evolution within each form.

Although the current work utilized the RED method to study genes involved in translation, a highly a conserved cellular process, this method may be more generally applicable to less conserved groups of genes. It may be possible to develop new sets of controls using the RED-T application (see methods) to observe the genetic history within the proteobacteria, the unique and diverse pathways within methanogens, identify the *Pyrococcus* spp.'s distinctiveness within the archaeal domain, or set a standard for fast-clock organisms, such as the *Mycoplasma*.

References

Andersson SG, Zomorodipour A, Andersson JO, Sicheritz-Ponten T, Alsmark UC, Podowski RM, Naslund AK, Eriksson AS, Winkler HH,Kurland CG (1998) The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. Nature 396:133-140

Benner SA, Cohen MA, Gonnet GH (1994) Amino acid substitution during functionally constrained divergent evolution of protein sequences. Protein Eng 7:1323-1332

Boucher Y, Huber H, L'Haridon S, Stetter KO, Doolittle WF (2001) Bacterial origin for the isoprenoid biosynthesis enzyme HMG-CoA reductase of the archaeal orders *Thermoplasmatales* and *Archaeoglobales*. Mol Biol Evol 18:1378-1388

Brown JR, Zhang J, Hodgson JE (1998) A bacterial antibiotic resistance gene with eukaryotic origins. Curr Biol 8:R365-367

Dayhoff MO, Eck FV, Park CM (1972) In Dayhoff, M. O. National Biomedical Research Foundation, Washington, DC, 89-99

Dayhoff MO, Schwartz RM, Orcutt BC (1978) A model of evolutionary change in proteins. In Dayhoff, M. O. National Biomedical Research Foundation, Washington, DC, 345-352

Doolittle WF (1999) Lateral genomics. Trends Cell Biol 9:M5-8.

Doolittle WF (1999) Phylogenetic classification and the universal tree. Science 284:2124-2129

Eisen JA (2000) Assessing evolutionary relationships among microbes from whole-genome analysis. Curr Opin Microbiol 3:475-480

Fabrega C, Farrow MA, Mukhopadhyay B, de Crecy-Lagard V, Ortiz AR, Schimmel P (2001) An aminoacyl tRNA synthetase whose sequence fits into neither of the two known classes. Nature 411:110-114

Felsenstein J (1993) PHYLIP (Phylogeny Inference Package) version 3.5c. Distributed by the author. Deparment of Genetics, University of Washington, Seattle

Ferretti JJ, McShan WM, Ajdic D, Savic DJ, Savic G, Lyon K, Primeaux C, Sezate S, Suvorov AN, Kenton S, Lai HS, Lin SP, Qian Y, Jia HG, Najar FZ, Ren Q, Zhu H, Song L, White J, Yuan X, Clifton SW, Roe BA, McLaughlin R (2001) Complete genome sequence of an M1 strain of *Streptococcus pyogenes*. Proc Natl Acad Sci U S A 98:4658-4663

Fitch WM, Margoliash E (1967) Construction of phylogenetic trees. Science 155:279-284

Gibbons A (1992) Mitochondrial Eve: wounded, but not dead yet. Science 257:873-875

Glansdorff N (2000) About the last common ancestor, the universal life-tree and lateral gene transfer: a reappraisal. Mol Microbiol 38:177-185

Gogarten JP, Olendzenski L, Hilario E, Simon C, Holsinger KE (1996) Dating the cenancester of organisms. Science 274:1750-1751; discussion 1751-1753

Gray MW (1992) The endosymbiont hypothesis revisited. Int Rev Cytol 141:233-357

Hansmann S, Martin W (2000) Phylogeny of 33 ribosomal and six other proteins encoded in an ancient gene cluster that is conserved across prokaryotic genomes: influence of excluding poorly alignable sites from analysis. Int J Syst Evol Microbiol 50:1655-1663

Henikoff S, Henikoff JG (1992) Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A 89:10915-10919

Hilario E, Gogarten JP (1993) Horizontal transfer of ATPase genes--the tree of life becomes a net of life. Biosystems 31:111-119

Ibba M,Soll D (2001) The renaissance of aminoacyl-tRNA synthesis. EMBO Rep 2:382-387

Jain R, Rivera MC, Lake JA (1999) Horizontal gene transfer among genomes: the complexity hypothesis. Proc Natl Acad Sci U S A 96:3801-3806

Kuroda M, Ohta T, Uchiyama I, Baba T, Yuzawa H, Kobayashi I, Cui L, Oguchi A, Aoki K, Nagai Y, Lian J, Ito T, Kanamori M, Matsumaru H, Maruyama A, Murakami H, Hosoyama A, Mizutani-Ui Y, Takahashi NK, Sawano T, Inoue R, Kaito C, Sekimizu K, Hirakawa H, Kuhara S, Goto S, Yabuzaki J, Kanehisa M, Yamashita A, Oshima K, Furuya K, Yoshino C, Shiba T, Hattori M, Ogasawara N, Hayashi H, Hiramatsu K (2001) Whole genome sequencing of meticillin-resistant *Staphylococcus aureus*. Lancet 357:1225-1240

Lamour V, Quevillon S, Diriong S, N'Guyen VC, Lipinski M, Mirande M (1994) Evolution of the Glx-tRNA synthetase family: the glutaminyl enzyme as a case of horizontal gene transfer. Proc Natl Acad Sci U S A 91:8670-8674

Lawrence JG, Ochman H (1998) Molecular archaeology of the *Escherichia coli* genome. Proc Natl Acad Sci U S A 95:9413-9417

Lawrence JG, Ochman H (2002) Reconciling the many faces of lateral gene transfer. Trends Microbiol 10:1-4

Li T, Graham DE, Stathopoulos C, Haney PJ, Kim HS, Vothknecht U, Kitabatake M, Hong KW, Eggertsson G, Curnow AW, Lin W, Celic I, Whitman W, Soll D (1999) Cysteinyl-tRNA formation: the last puzzle of aminoacyl-tRNA synthesis. FEBS Lett 462:302-306

Margulis L (1993) Symbiosis in cell evolution : microbial communities in the Archean and Proterozoic eons. Second ed. Freeman, New York, xxvii, 452

Martin W (1999) Mosaic bacterial chromosomes: a challenge en route to a tree of genomes. Bioessays 21:99-104

Matte-Tailliez O, Brochier C, Forterre P, Philippe H (2002) Archaeal phylogeny based on ribosomal proteins. Mol Biol Evol 19:631-639

Nelson KE, Clayton RA, Gill SR, Gwinn ML, Dodson RJ, Haft DH, Hickey EK, Peterson JD, Nelson WC, Ketchum KA, McDonald L, Utterback TR, Malek JA, Linher KD, Garrett MM, Stewart AM, Cotton MD, Pratt MS, Phillips CA, Richardson D, Heidelberg J, Sutton GG, Fleischmann RD, Eisen JA, Fraser CM,et al. (1999) Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of *Thermotoga maritima*. Nature 399:323-329

Nesbo CL, Boucher Y, Doolittle WF (2001) Defining the core of nontransferable prokaryotic genes: the euryarchaeal core. J Mol Evol 53:340-350

Overbeek R, Larsen N, Pusch GD, D'Souza M, Selkov E, Jr., Kyrpides N, Fonstein M, Maltsev N, Selkov E (2000) WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction. Nucleic Acids Res 28:123-125

Perna NT, Plunkett G, 3rd, Burland V, Mau B, Glasner JD, Rose DJ, Mayhew GF, Evans PS, Gregor J, Kirkpatrick HA, Posfai G, Hackett J, Klink S, Boutin A, Shao Y,

Miller L, Grotbeck EJ, Davis NW, Lim A, Dimalanta ET, Potamousis KD, Apodaca J, Anantharaman TS, Lin J, Yen G, Schwartz DC, Welch RA, Blattner FR (2001) Genome sequence of enterohaemorrhagic Escherichia coli O157:H7. Nature 409:529-533

Perriere G, Gouy M (1996) WWW-query: an on-line retrieval system for biological sequence banks. Biochimie 78:364-369

Reanney DC (1978) Coupled evolution: adaptive interactions among the genomes of plasmids, viruses, and cells. Int Rev Cytol Suppl Suppl:1-68

Smith MW, Feng DF, Doolittle RF (1992) Evolution by acquisition: the case for horizontal gene transfers. Trends Biochem Sci 17:489-493

Snedecor GW, Cochran WG (1967) 6. The Iowa State University Press, Ames,

Snel B, Bork P, Huynen MA (2002) Genomes in flux: the evolution of archaeal and proteobacterial gene content. Genome Res 12:17-25

Sonea S, Paniset M (1976) Towards a new bacteriology. Rev Can Biol 35:103-167

Stathopoulos C, Li T, Longman R, Vothknecht UC, Becker HD, Ibba M, Soll D (2000) One polypeptide with two aminoacyl-tRNA synthetase activities. Science 287:479-482

Tamura K (1993) TreeExplorer (supplement of MEGA - Molecular Evolutionary Genetics Analysis). Arizona State University, Tempe, AZ, http://www.megasoftware.net

Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res 25:4876-4882

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673-4680

Woese CR, Olsen GJ, Ibba M, Soll D (2000) Aminoacyl-tRNA synthetases, the genetic code, and the evolutionary process. Microbiol Mol Biol Rev 64:202-236

Woese CR (2002) On the evolution of cells. Proc Natl Acad Sci U S A 99:8742-8747

Figure 2.1: Parental model used for interpretation of RED plots. **[A] Hypothetical tree containing representatives of the three domains.** Archaea (A), Bacteria (subgroups B1 and B2) and Eukaryote (E). Solid lines represent the phylogeny in the absence of LGT. The broken line represents a potential LGT from the Archaea to the B1 subgroup of the bacteria. Numbers represent evolutionary distances (E_d) in arbitrary units. **[B] RED plot of the hypothetical tree in A.** Both vertical (●) and lateral (○) gene transfer events are projected on this plot: y-axis represents the E_d for the experimental gene, and the x-axis represents the E_d for the control gene. The control gene is assumed to have undergone vertical evolution only. (i) reflects the E_d of the donating taxa from B2, and (ii) reflects the E_d of the recipient B1 from the modern ancestors of the donating taxa.

Figure 2.2: Factors important in standardization of RED plots. **[A] Increasing the number of control genes lowered the relative standard deviation (rSD) of *K*.** *K* is the ratio of the E_d of the experimental gene to that of the control gene(s). Glutamyl-tRNA synthetase was the experimental protein (Δ) and various combinations of leucyl-tRNA synthetase, SecY, L2P, L13P, and S13P were the control genes. The same analysis was performed without L13P (○), when all the archaeal genes were removed (●), and when the ribosomal protein L2P was the experimental gene (■). **[B] Affect of sequence length on the relative standard deviation (*rSD*) of the bacterial intra-domain comparisons.** The Power trend-line, with a R-squared value of 0.631, provides the best estimate of the *rSD* with ORF length. The equation of this trend-line was $\mathbf{y = 0.0714x^2 - 0.4122x + 0.7498}$, with **y** being the estimated *rSD* and **x** being the logarithm of the average length of the ORF (in number of codons). Above 300 codons (log value of 2.5), the sequence length did not affect the

measured *rSD*. In order to correct for ORF length below 300 codons, this equation was used to calculate an expected *rSD* and thus D_{rSD}.

Figure 2.3: RED plot of S19P. Intra-domain comparisons within the archaea (□), within the bacteria (■), inter-domain comparisons between the archaea and the eukaryote (Δ), between the archaea and the bacteria (○), and between the bacteria and the eukaryote (●).

Figure 2.4: Phylogenetic analysis of the valyl-tRNA synthetase. **[A] RED plot of valyl-tRNA synthetase.** Intra-domain comparisons within the archaea (◊), within the bacteria (except for *Rickettsia prowazekii*, RP)(■), between the bacteria and RP (Δ), inter-domain comparisons between the archaea and the bacteria (○), between the Crenarchaeota and RP (●), and between Euryarchaeota and RP (▲). **[B] Gene tree of valyl-tRNA synthetases.** The tree was calculated by the Fitch-Margoliash algorithm. The scale bar corresponded to 20 substitutions per 100 positions. Recipient genomes are in bold. **[C] Hypothetical tree illustrating the proposed evolutionary history of valyl-tRNA synthetase.** A_E and A_C represent the two archaeal sub-domains, Euryarchaeota and Crenarchaeota, respectively. B1 represents all bacteria examined except for *Rickettsia prowazekii*, E represents the Eukaryote domain, T1 represents the acquisition of the euryarchaeotic gene by RP, and T2 represents an ancient nonlinearity that could have resulted from a change in the rate of evolution or a LGT prior to formation of the domains.

Figure 2.5: Phylogenetic analysis of the isoleucyl-tRNA synthetase. **[A] RED plot of isoleucyl-tRNA synthetase.** Intra-domain comparisons within the archaea (□), within the

bacterial clade B1 or B2 (■), between bacteria clades B1 and B2 (Δ), inter-domain comparisons between the archaea and eukaryote (◊), between the archaea and the bacteria (○), between the eukaryote and bacterial clade B1 (●), and between eukaryote and bacterial clade B2 (▲). **[B] Gene tree of isoleucyl-tRNA synthetases.** The tree was calculated by the Fitch-Margoliash algorithm. The scale bar corresponded to 20 substitutions per 100 positions. Recipient genomes are in bold. **[C] Hypothetical tree illustrating the proposed evolutionary history of isoleucyl-tRNA synthetase.** A_E and A_C represent the two archaeal sub-domains Euryarchaeota and Crenarchaeota, respectively, B1 and B2 represent two clades of the bacteria, E represents the eukaryotic domain, T1 represents the acquisition of the eukaryotic gene by bacterial clade B2, and T2 represents an ancient nonlinearity that could have resulted from a change in the rate of evolution or a LGT prior to formation of the domains.

Figure 2.6: Phylogenetic analysis of the prolyl-tRNA synthetase. **[A] RED plot of prolyl-tRNA synthetase.** Intra-domain comparisons within the archaea (□), within the bacterial clade B1 or B2 (■), between bacterial clades B1 and B2 (Δ), inter-domain comparisons between the archaea and the eukaryote (♦), between the archaea and bacterial clade B1 (○), between the archaea and bacterial clade B2 (●), between the eukaryote and bacterial clade B1 (◊), and between the eukaryote and bacterial clade B2 (▲). **[B] Gene tree of prolyl-tRNA synthetases.** The tree was calculated by the Fitch-Margoliash algorithm. The scale bar corresponded to 50 substitutions per 100 positions. Recipient genomes are in bold. **[C] Hypothetical tree illustrating the proposed evolutionary history of prolyl-tRNA synthetase.** A represents the archaeal domain, B1 and B2 represent two clades of

the bacteria, E represents the eukaryotic domain, T1 represents the acquisition of bacterial clade B2 gene by the eukaryote, and T2 represents the acquisition of the archaeal gene by bacterial clade B1.

Figure 2.7: Phylogenetic analysis of the cysteinyl-tRNA synthetase. **[A] RED plot of cysteinyl-tRNA synthetase.** Intra-domain comparisons within the archaea (□), within the bacteria (■), inter-domain comparisons between the archaea and the eukaryote (Δ), between the archaea and the bacteria (○), and between the eukaryote and the bacteria (●). **[B] Gene tree of cysteinyl-tRNA synthetases.** The tree was calculated by the Fitch-Margoliash algorithm. The scale bar corresponded to 20 substitutions per 100 positions. **[C] Hypothetical tree illustrating the proposed evolutionary history of cysteinyl-tRNA synthetase.** A represents the archaea, B1 and B2 represent two clades of the bacteria, E represents the Eukaryote domain, and T either represents a modern origin for this gene with LGT to a number of lineages within a short period of time or a change in the rate of evolution.

Figure 2.8: Phylogenetic analysis of the glutamyl-tRNA synthetase. **[A] RED plot of glutamyl-tRNA synthetase.** Intra-domain comparisons within the archaea (□), within the bacteria (■), between (beta and gamma) proteobacteria and *Pseudomonas aeruginosa* PA (○), between PA and *Chlamydia pnemoniae* (CQ), *Chlamydia trachomatis* (CT) and *Deinococcus radiodurans* (DR) (Δ), and between the archaea and the bacteria (●). **[B] Gene tree of glutamyl-tRNA synthetases.** The tree was calculated by the Fitch-Margoliash algorithm. The scale bar corresponded to 20 substitutions per 100 positions. **[C] Hypothetical tree illustrating the proposed evolutionary history of glutamyl-tRNA synthetase.** B

represents the bacterial domain (excluding *Pseudomonas aeruginosa* (PA), *Chlamydia pneumoniae* (CQ) *Chlamydia trachomatis* (CT) and *Deinococcus radiodurans* (DR), A represents the archaeal domain, E represents the eukaryotic domain, T1 represents the acquisition by PA of the gene from CQ, CT and DR, and T2 represents an ancient nonlinearity.

Figure 2.9: Phylogenetic analysis of the tyrosyl-tRNA synthetase. **[A] RED plot of tyrosyl-tRNA synthetase.** Intra-domain comparisons within the archaea (□), within the bacterial clades B1 or B2 (■), between the bacterial clades B1 and B2 (Δ), inter-domain comparisons between the archaea and the eukaryote (◊), between the archaea and bacterial clade B1 (○), between the archaea and bacterial clade B2 (●), between the eukaryote and bacterial clade B1 (♦), and between the eukaryote and the bacterial clade B2 (▲). **[B] Gene tree of tyrosyl-tRNA synthetases.** The tree was calculated by the Fitch-Margoliash algorithm. The scale bar corresponded to 20 substitutions per 100 positions. **[C] Hypothetical tree illustrating the proposed evolutionary history of tyrosyl-tRNA synthetase.** A represents the archaeal domain, B1 and B2 represent two clades of the bacteria, E represents the eukaryotic domain. This model assumes the common ancestor to have possessed multiple copies of the tyrosyl-tRNA synthetase gene. The copy in B1 diverged rapidly from the genes in B2, A and E.

Figure 2.1

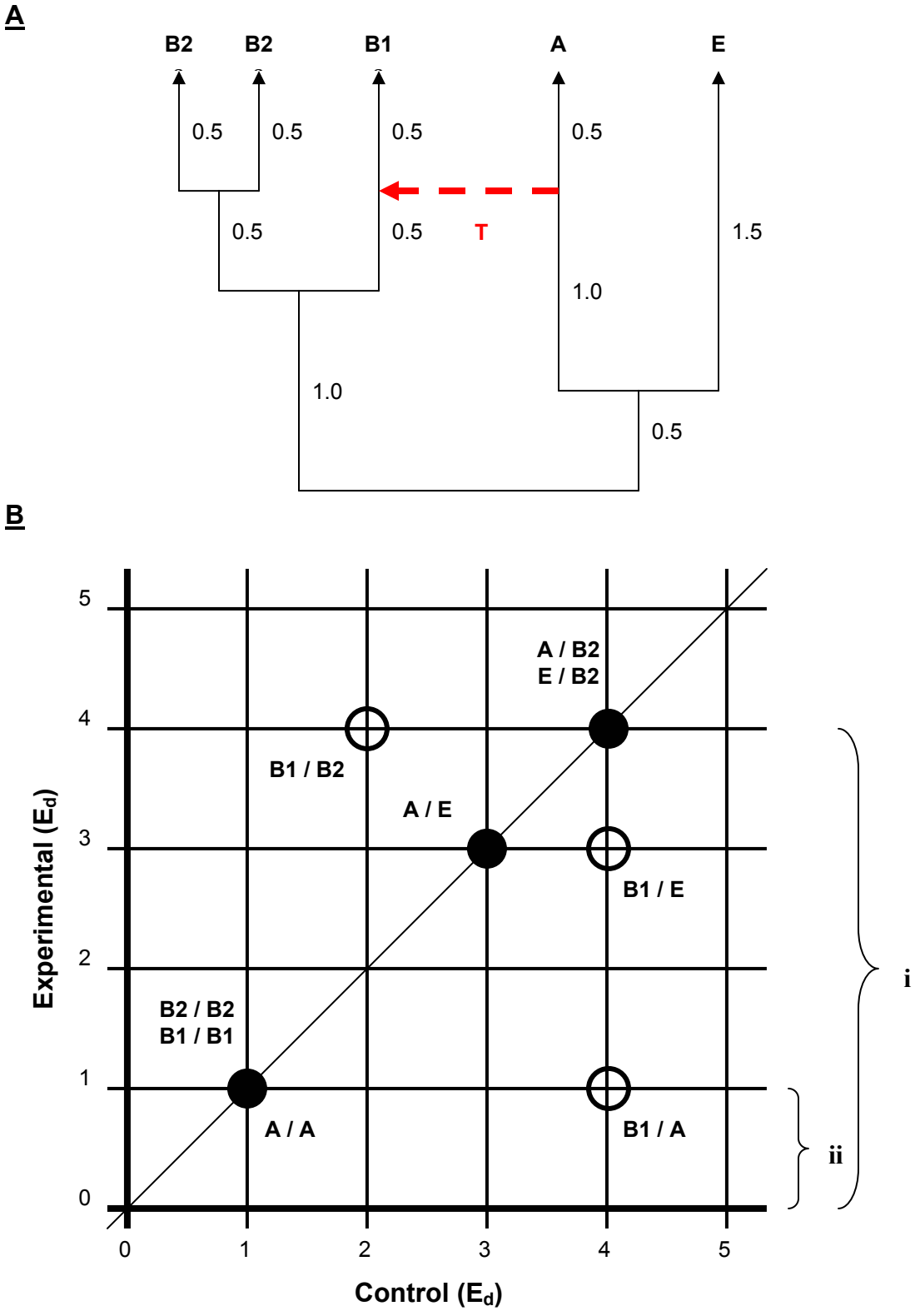
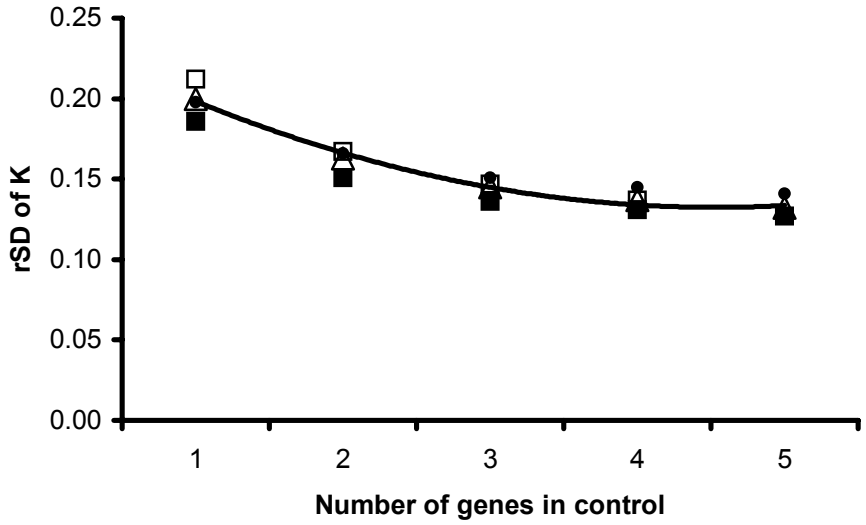

A
B2
B2
B1
A
E
0.5
0.5
0.5
0.5
0.5
0.5
T
1.5
1.0
1.0
0.5
B
5
4
3
2
1
0
Experimental (E_d)
A / B2
E / B2
B1 / B2
A / E
B1 / E
B2 / B2
B1 / B1
A / A
B1 / A
i
ii
0
1
2
3
4
5
Control (E_d)

Figure 2.2

A


0.25
0.20
0.15
0.10
0.05
0.00
rSD of K
1
2
3
4
5
Number of genes in control

B

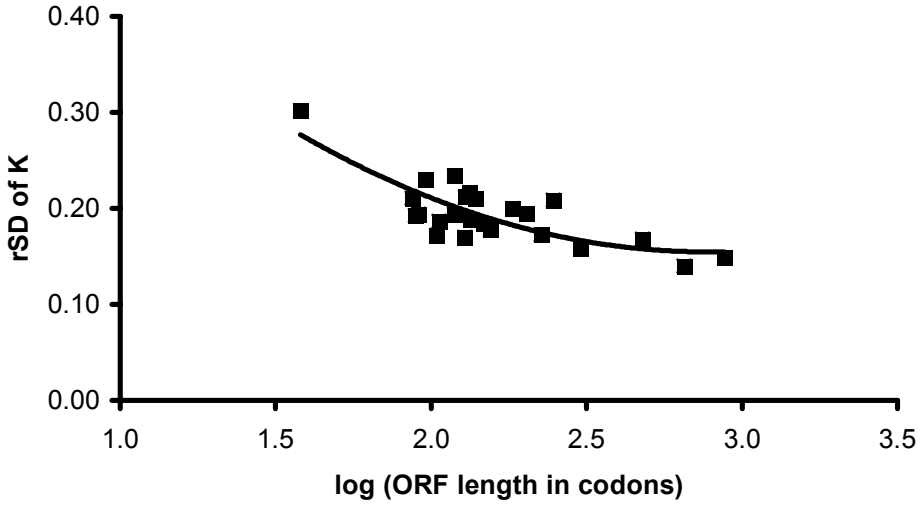

0.40
0.30
0.20
0.10
0.00
rSD of K
1.0
1.5
2.0
2.5
3.0
3.5
log (ORF length in codons)

Figure 2.3

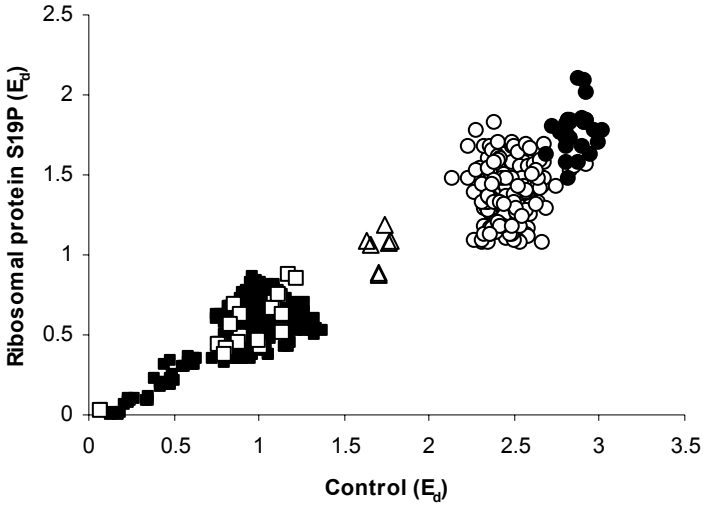


Figure 2.4

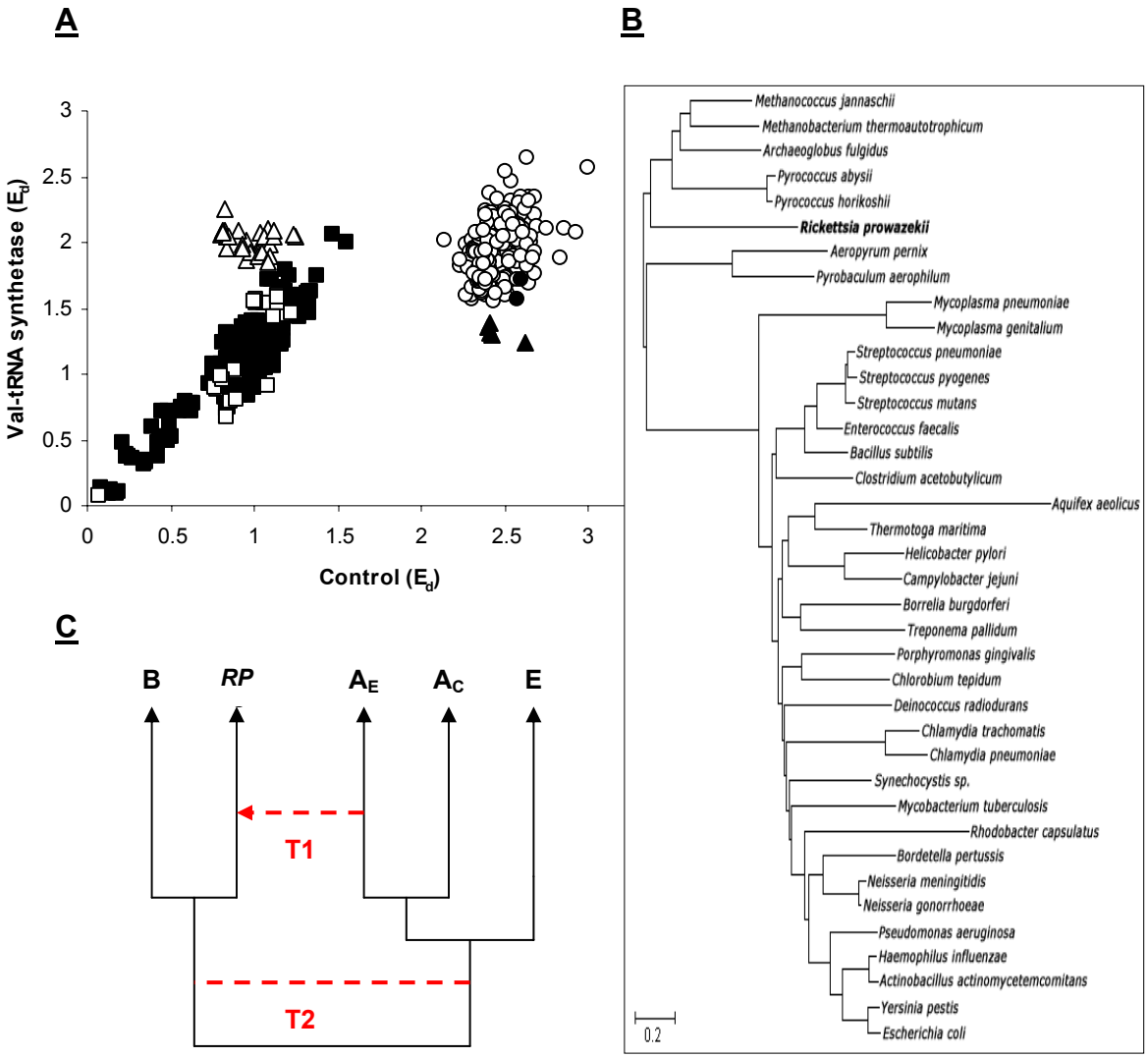

A
B
C
Val-tRNA synthetase (E_d)
Control (E_d)
0
0.5
1
1.5
2
2.5
3
B
RP
A_E
A_C
E
T1
T2
Methanococcus jannaschii
Methanobacterium thermoautotrophicum
Archaeoglobus fulgidus
Pyrococcus abysii
Pyrococcus horikoshii
Rickettsia prowazekii
Aeropyrum pernix
Pyrobaculum aerophilum
Mycoplasma pneumoniae
Mycoplasma genitalium
Streptococcus pneumoniae
Streptococcus pyogenes
Streptococcus mutans
Enterococcus faecalis
Bacillus subtilis
Clostridium acetobutylicum
Aquifex aeolicus
Thermotoga maritima
Helicobacter pylori
Campylobacter jejuni
Borrelia burgdorferi
Treponema pallidum
Porphyromonas gingivalis
Chlorobium tepidum
Deinococcus radiodurans
Chlamydia trachomatis
Chlamydia pneumoniae
Synechocystis sp.
Mycobacterium tuberculosis
Rhodobacter capsulatus
Bordetella pertussis
Neisseria meningitidis
Neisseria gonorrhoeae
Pseudomonas aeruginosa
Haemophilus influenzae
Actinobacillus actinomycetemcomitans
Yersinia pestis
Escherichia coli
0.2

Figure 2.5

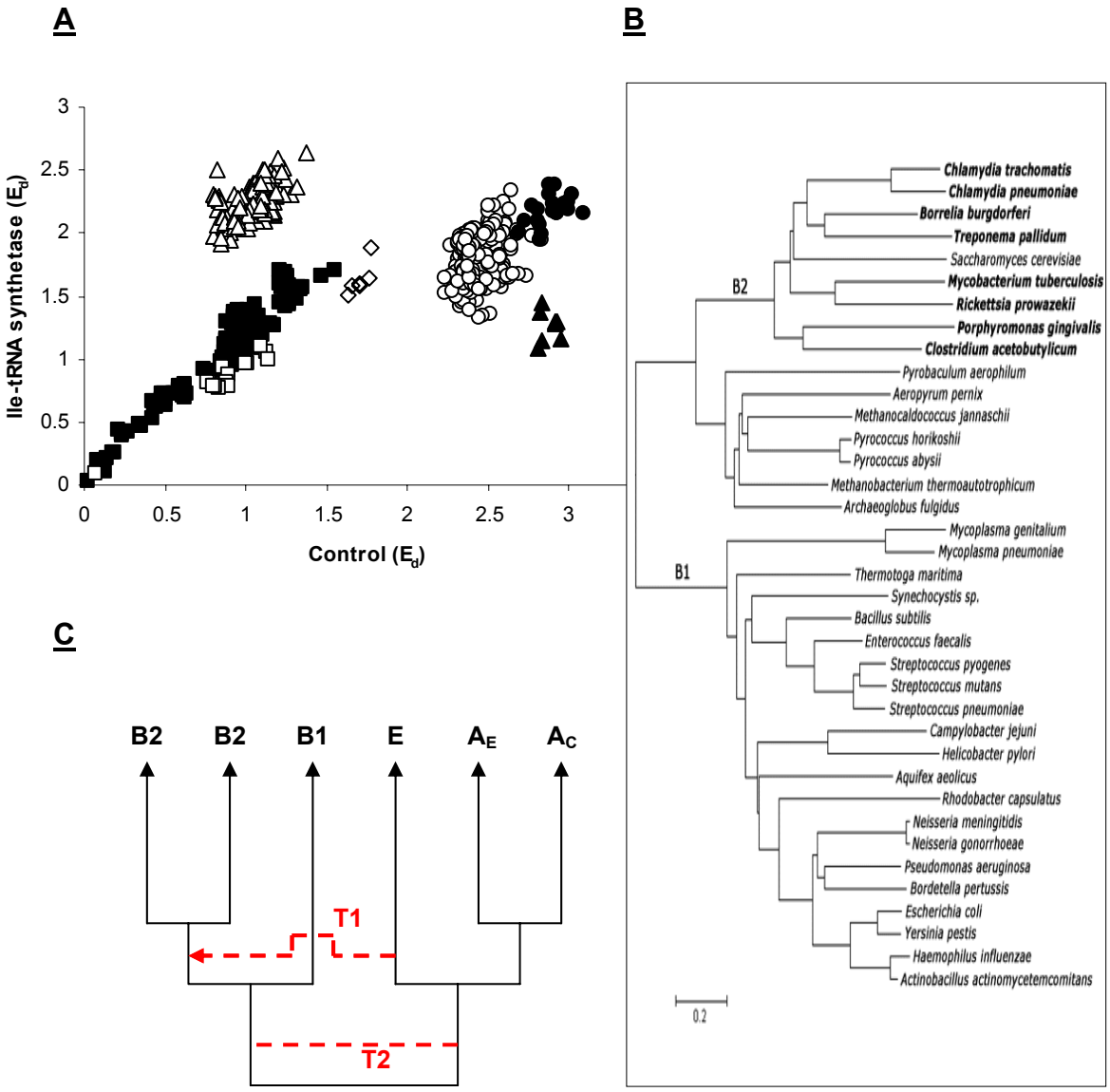

A
B
C
Ile-tRNA synthetase (Ed)
Control (Ed)
0
0.5
1
1.5
2
2.5
3
B2
B1
Chlamydia trachomatis
Chlamydia pneumoniae
Borrelia burgdorferi
Treponema pallidum
Saccharomyces cerevisiae
Mycobacterium tuberculosis
Rickettsia prowazekii
Porphyromonas gingivalis
Clostridium acetobutylicum
Pyrobaculum aerophilum
Aeropyrum pernix
Methanocaldococcus jannaschii
Pyrococcus horikoshii
Pyrococcus abysii
Methanobacterium thermoautotrophicum
Archaeoglobus fulgidus
Mycoplasma genitalium
Mycoplasma pneumoniae
Thermotoga maritima
Synechocystis sp.
Bacillus subtilis
Enterococcus faecalis
Streptococcus pyogenes
Streptococcus mutans
Streptococcus pneumoniae
Campylobacter jejuni
Helicobacter pylori
Aquifex aeolicus
Rhodobacter capsulatus
Neisseria meningitidis
Neisseria gonorrhoeae
Pseudomonas aeruginosa
Bordetella pertussis
Escherichia coli
Yersinia pestis
Haemophilus influenzae
Actinobacillus actinomycetemcomitans
0.2
B2
B2
B1
E
AE
AC
T1
T2

Figure 2.6

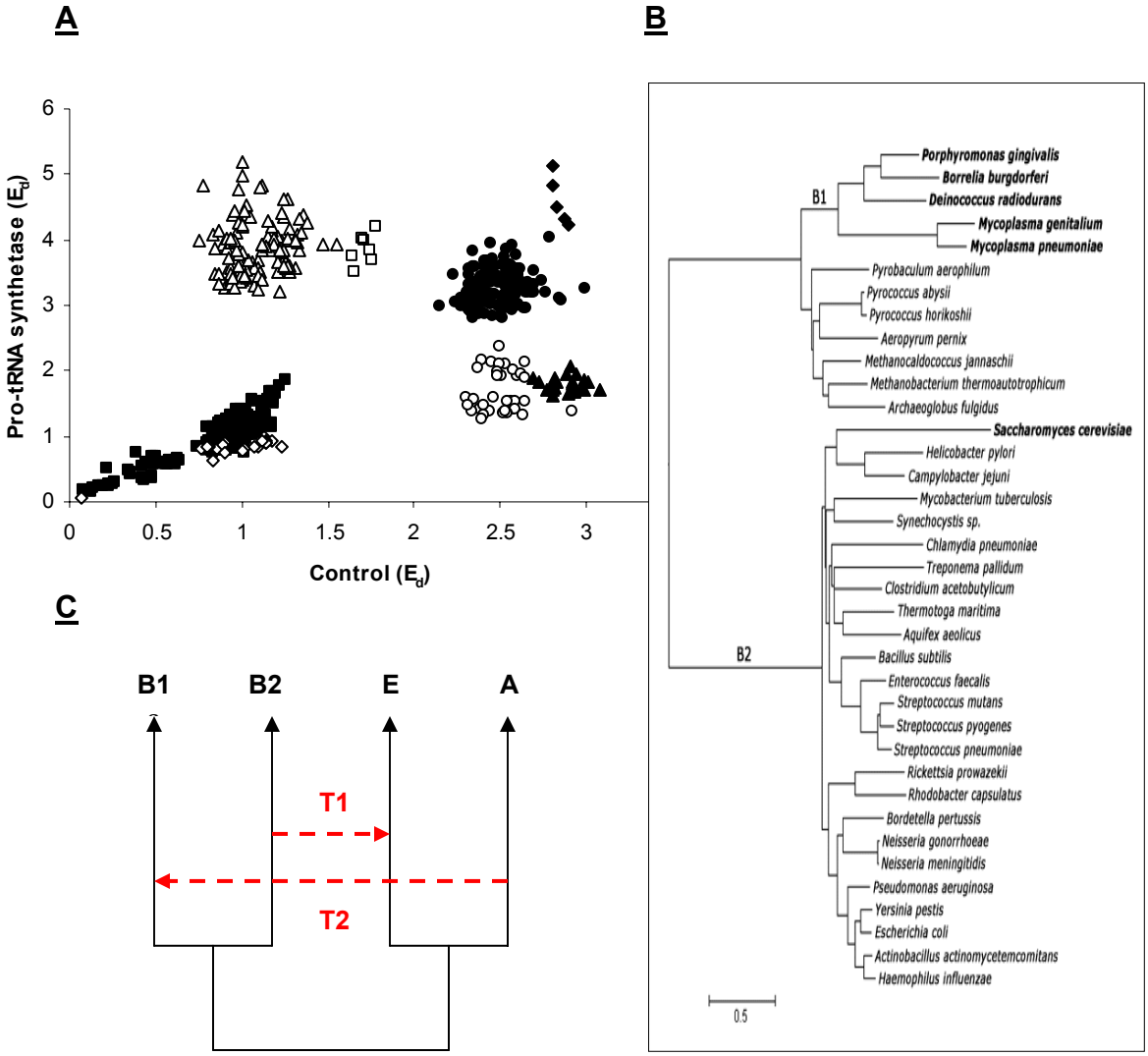

A
B
C
Pro-tRNA synthetase (E_d)
Control (E_d)
0
1
2
3
4
5
6
0
0.5
1
1.5
2
2.5
3
B1
B2
E
A
T1
T2
Porphyromonas gingivalis
Borrelia burgdorferi
Deinococcus radiodurans
Mycoplasma genitalium
Mycoplasma pneumoniae
Pyrobaculum aerophilum
Pyrococcus abysii
Pyrococcus horikoshii
Aeropyrum pernix
Methanocaldococcus jannaschii
Methanobacterium thermoautotrophicum
Archaeoglobus fulgidus
Saccharomyces cerevisiae
Helicobacter pylori
Campylobacter jejuni
Mycobacterium tuberculosis
Synechocystis sp.
Chlamydia pneumoniae
Treponema pallidum
Clostridium acetobutylicum
Thermotoga maritima
Aquifex aeolicus
Bacillus subtilis
Enterococcus faecalis
Streptococcus mutans
Streptococcus pyogenes
Streptococcus pneumoniae
Rickettsia prowazekii
Rhodobacter capsulatus
Bordetella pertussis
Neisseria gonorrhoeae
Neisseria meningitidis
Pseudomonas aeruginosa
Yersinia pestis
Escherichia coli
Actinobacillus actinomycetemcomitans
Haemophilus influenzae
0.5

Figure 2.7

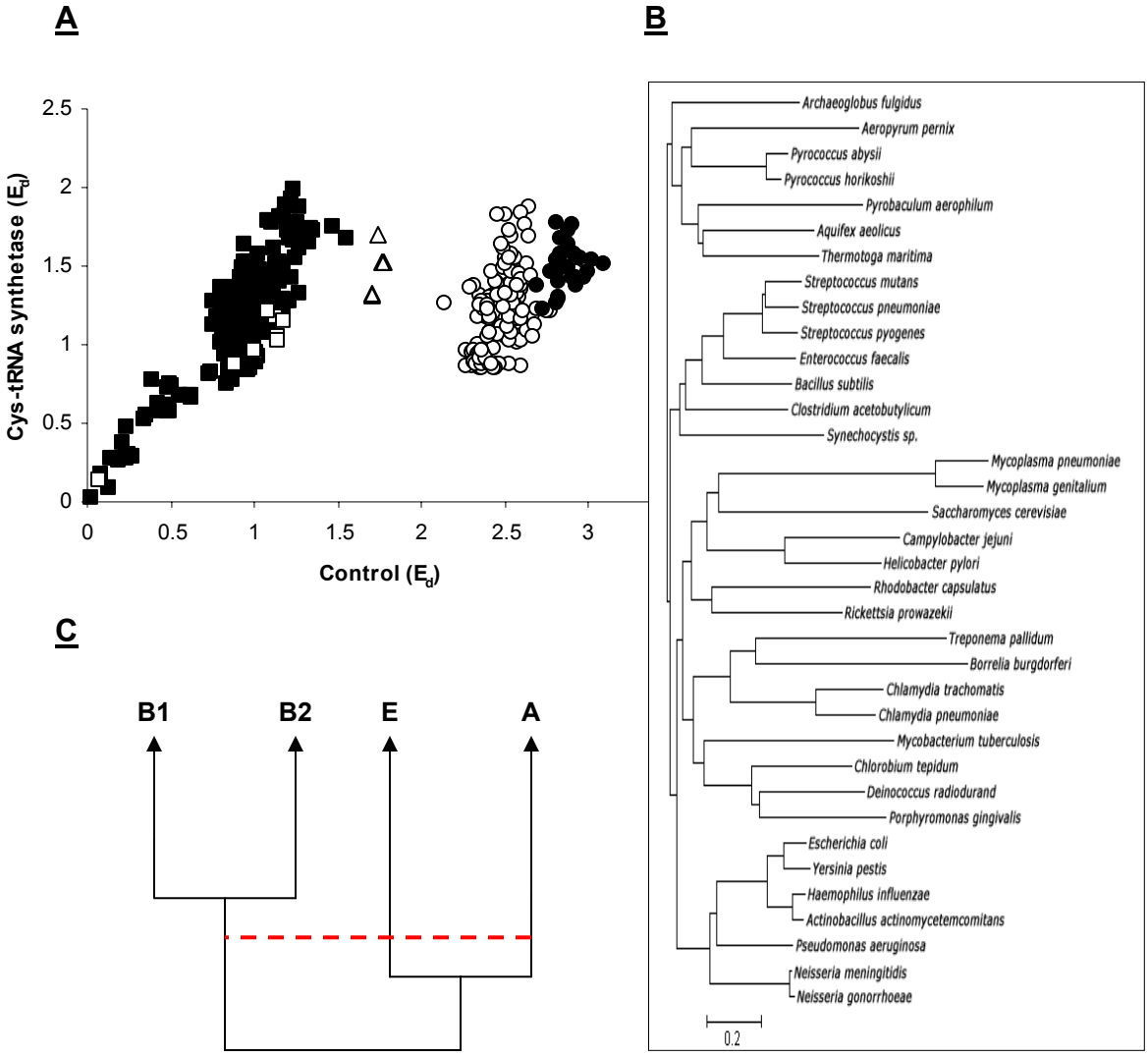

A
B
C
Cys-tRNA synthetase (E_d)
Control (E_d)
0
0.5
1
1.5
2
2.5
3
B1
B2
E
A
Archaeoglobus fulgidus
Aeropyrum pernix
Pyrococcus abysii
Pyrococcus horikoshii
Pyrobaculum aerophilum
Aquifex aeolicus
Thermotoga maritima
Streptococcus mutans
Streptococcus pneumoniae
Streptococcus pyogenes
Enterococcus faecalis
Bacillus subtilis
Clostridium acetobutylicum
Synechocystis sp.
Mycoplasma pneumoniae
Mycoplasma genitalium
Saccharomyces cerevisiae
Campylobacter jejuni
Helicobacter pylori
Rhodobacter capsulatus
Rickettsia prowazekii
Treponema pallidum
Borrelia burgdorferi
Chlamydia trachomatis
Chlamydia pneumoniae
Mycobacterium tuberculosis
Chlorobium tepidum
Deinococcus radiodurand
Porphyromonas gingivalis
Escherichia coli
Yersinia pestis
Haemophilus influenzae
Actinobacillus actinomycetemcomitans
Pseudomonas aeruginosa
Neisseria meningitidis
Neisseria gonorrhoeae
0.2

Figure 2.8

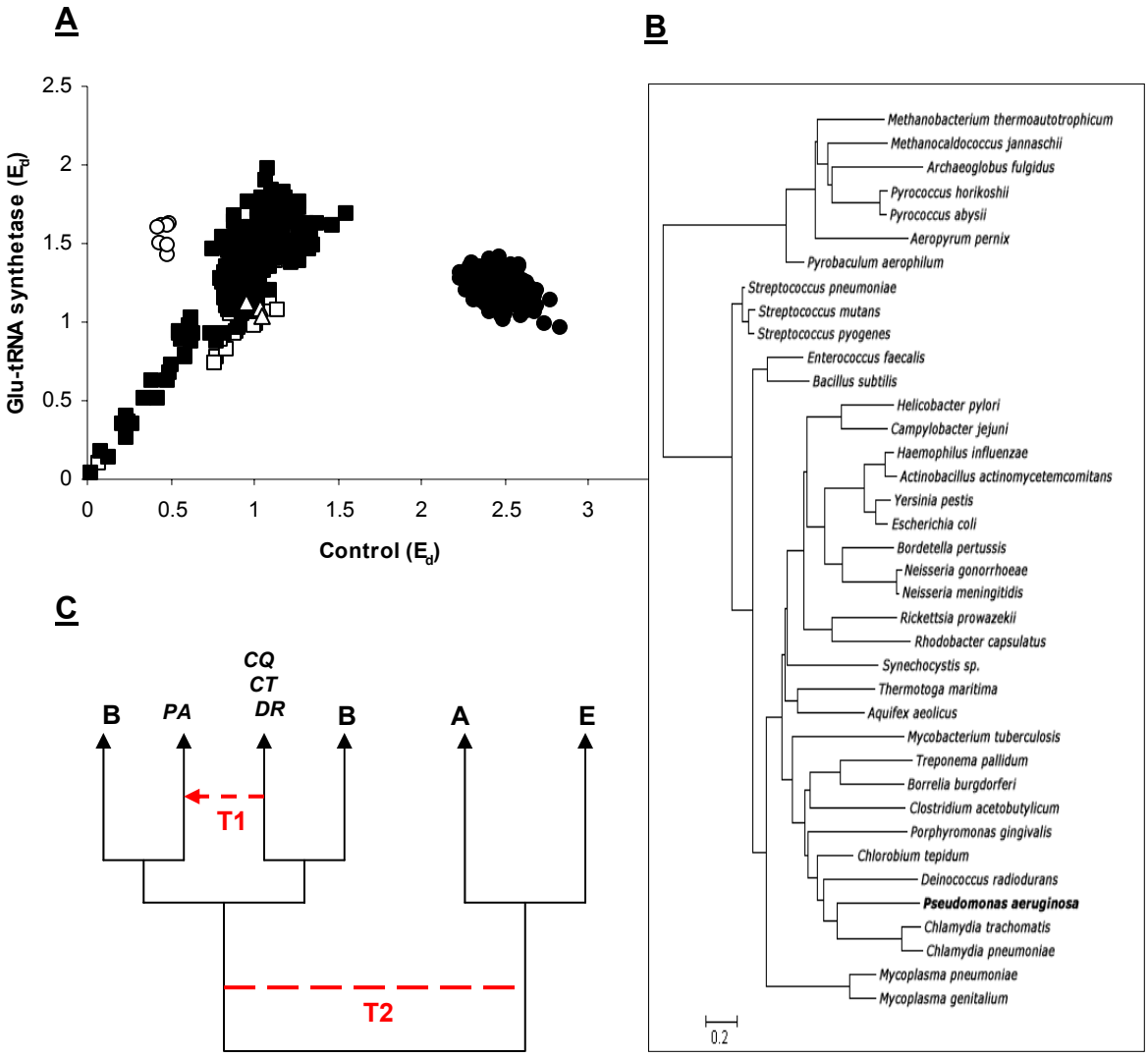

A
B
C
Glu-tRNA synthetase (E_d)
Control (E_d)
0
0.5
1
1.5
2
2.5
3
CQ
CT
DR
PA
B
B
A
E
T1
T2
Methanobacterium thermoautotrophicum
Methanocaldococcus jannaschii
Archaeoglobus fulgidus
Pyrococcus horikoshii
Pyrococcus abysii
Aeropyrum pernix
Pyrobaculum aerophilum
Streptococcus pneumoniae
Streptococcus mutans
Streptococcus pyogenes
Enterococcus faecalis
Bacillus subtilis
Helicobacter pylori
Campylobacter jejuni
Haemophilus influenzae
Actinobacillus actinomycetemcomitans
Yersinia pestis
Escherichia coli
Bordetella pertussis
Neisseria gonorrhoeae
Neisseria meningitidis
Rickettsia prowazekii
Rhodobacter capsulatus
Synechocystis sp.
Thermotoga maritima
Aquifex aeolicus
Mycobacterium tuberculosis
Treponema pallidum
Borrelia burgdorferi
Clostridium acetobutylicum
Porphyromonas gingivalis
Chlorobium tepidum
Deinococcus radiodurans
Pseudomonas aeruginosa
Chlamydia trachomatis
Chlamydia pneumoniae
Mycoplasma pneumoniae
Mycoplasma genitalium
0.2

Figure 2.9

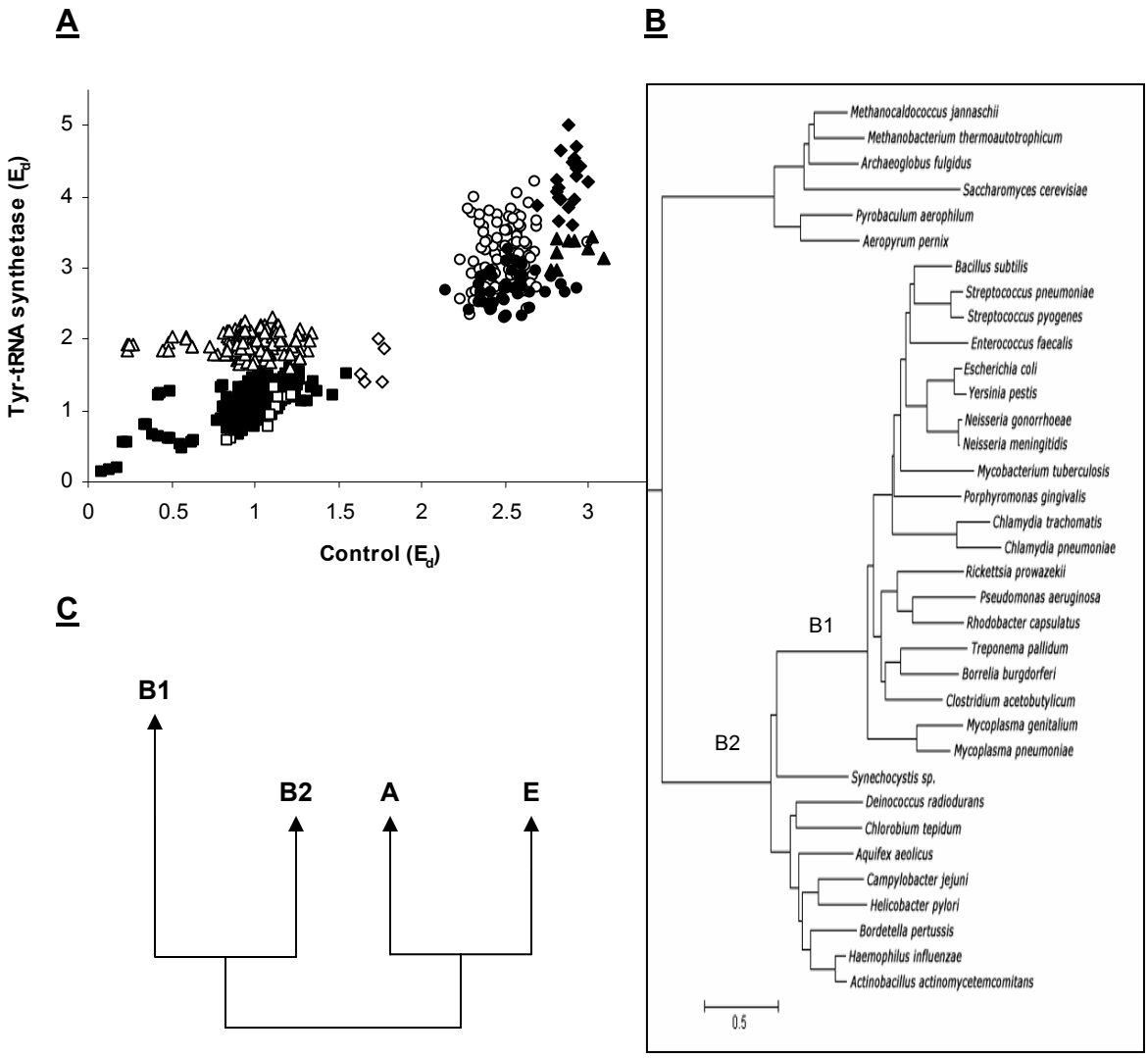

A
B
C
Tyr-tRNA synthetase (E_d)
Control (E_d)
0
1
2
3
4
5
0
0.5
1
1.5
2
2.5
3
Methanocaldococcus jannaschii
Methanobacterium thermoautotrophicum
Archaeoglobus fulgidus
Saccharomyces cerevisiae
Pyrobaculum aerophilum
Aeropyrum pernix
Bacillus subtilis
Streptococcus pneumoniae
Streptococcus pyogenes
Enterococcus faecalis
Escherichia coli
Yersinia pestis
Neisseria gonorrhoeae
Neisseria meningitidis
Mycobacterium tuberculosis
Porphyromonas gingivalis
Chlamydia trachomatis
Chlamydia pneumoniae
Rickettsia prowazekii
Pseudomonas aeruginosa
Rhodobacter capsulatus
Treponema pallidum
Borrelia burgdorferi
Clostridium acetobutylicum
Mycoplasma genitalium
Mycoplasma pneumoniae
Synechocystis sp.
Deinococcus radiodurans
Chlorobium tepidum
Aquifex aeolicus
Campylobacter jejuni
Helicobacter pylori
Bordetella pertussis
Haemophilus influenzae
Actinobacillus actinomycetemcomitans
B1
B2
0.5
B1
B2
A
E

Table 2:1: Summary of LGT events detected by RED among the aminoacyl-tRNA synthetases.

Aminoacyl-tRNA synthetases	RED model	Proposed LGT	Ancient nonlinearity[1]
Ala	III	Eukaryote **to** *Bacteria*	yes
Arg	II III VI	Archaea **to** *Deinococcus* Eukaryote **to** bacterial clade B2 *Rhodobacter* **to** *Bacillus*	yes
Asn	III	Archaea **to** bacterial clade B1	yes
Asp	I	none	no
Cys	V	none	yes
Glu	VI	*Chlamydia spp.* and *Deinococcus* **to** *Pseudomonas*	no
Gly (α, β)	I	none	na[2]
His	II III	Eukaryote **to** *Aeropyrum* Eukaryote **to** bacterial clade B1	yes
Ile	III	Eukaryote **to** bacterial clade B2	yes
Met	II	Archaea **to** *Chlorobium* and *Porphyromonas* Eukaryote **to** spirochetes	yes
Phe (α, β)	II	Archaea **to** spirochetes	yes
Pro	III IV	Archaea **to** bacterial clade B1 bacterial clade B2 **to** Eukaryote	no
Ser	II	Eukaryote **to** spirochetes	yes
Thr	IV	Bacteria **to** Eukaryote and sub-domain Crenarchaeota	no
Trp	III VI	Eukaryote **to** *Pyrococcus spp.* and *Pyrobaculum* *Deinococcus* **to** *Streptococcus pyogenes*	no
Tyr	VII	none[3]	no
Val	II	Euryarchaeota **to** *Rickettsia*	yes

[1] Nonlinearity detected for inter-domain comparisons that were consistent with either an ancient LGT or change in the rate of evolution.

[2] Only bacterial sequences were available for this gene, thus inter-domain comparisons were not analyzed.

[3] Although LGT event was not observed, RED suggested an ancient divergence (model VII) for this aminoacyl-tRNA synthetase gene.

Table 2.2: Parameters for the 40 ribosomal proteins – Intra-domain comparisons. Abbreviations are: function of each encoded gene (fx); total number of comparisons (total n_s), average sequence length in number of codons (l_s). Intra-domain includes only comparisons within the archaea or bacteria. There are two sections that are for intra-domain comparisons: archaea with other archaea (AA) and bacteria with other bacteria (BB). **AA:** number of comparisons (n_s), mean of K (m_{AA}), expected mean of K with standard deviation (m'_{AA} (SD'_{AA})), standard deviation of K (SD), and relative standard deviation of K (rSD). **BB:** number of comparisons (n_s), mean of K (m_{BB}), standard deviation of K (SD), relative standard deviation of K (rSD), relative standard deviation expected from l_s (rSD^*), and $rSD - rSD^*$ (D_{rSD}). "*na*" is not applicable, usually because of the absence of sequence for that domain.

				inTRA-domain comparisons										
				archaea vs. archaea					bacteria vs. bacteria					
	fx	total n_s	l_s	n_s	m_{AA}	m'_{AA} (SD'_{AA})	SD	rSD	n_s	m_{BB}	SD	rSD	rSD^*	D_{rSD}
Small Subunit Ribosomal Proteins (SSU)	S2P	542	252	15	0.671	0.900 (0.077)	0.173	0.258	325	0.901	0.163	0.180	0.172	0.009
	S3P	611	234	21	1.004	0.808 (0.079)	0.176	0.175	351	0.807	0.167	0.207	0.174	0.033
	S4P	441	203	15	0.813	0.952 (0.083)	0.174	0.214	276	0.958	0.187	0.195	0.179	0.016
	S5P	542	183	21	0.707	0.849 (0.072)	0.137	0.194	325	0.846	0.169	0.200	0.183	0.017
	S6P	276	118	*na*	*na*	*na*	*na*	*na*	276	1.52	0.35	0.230	0.202	0.028
	S7P	479	167	15	0.720	0.683 (0.059)	0.153	0.213	276	0.685	0.130	0.190	0.186	0.003
	S8P	479	133	15	0.657	0.921 (0.075)	0.131	0.199	276	0.904	0.196	0.217	0.196	0.020
	S9P	510	134	15	0.619	0.922 (0.067)	0.104	0.168	300	0.915	0.172	0.188	0.196	-0.008
	S10P	496	103	21	0.568	0.585 (0.058)	0.125	0.220	276	0.592	0.154	0.260	0.209	0.051
	S11P	542	129	15	0.370	0.617 (0.073)	0.061	0.165	325	0.613	0.156	0.254	0.198	0.057
	S12P	276	128	*na*	*na*	*na*	*na*	*na*	276	0.341	0.058	0.170	0.198	-0.028
	S13P	300	122	*na*	*na*	*na*	*na*	*na*	300	0.683	0.16	0.234	0.201	0.034
	S14P	325	87	*na*	*na*	*na*	*na*	*na*	325	0.973	0.205	0.211	0.219	-0.008
	S15P	378	91	*na*	*na*	*na*	*na*	*na*	378	0.87	0.169	0.194	0.216	-0.022
	S16P	300	100	*na*	*na*	*na*	*na*	*na*	231	1.076	0.272	0.253	0.211	0.042
	S17P	481	96	21	0.916	0.820 (0.070)	0.200	0.218	210	0.859	0.191	0.222	0.213	0.009
	S18P	276	83	*na*	*na*	*na*	*na*	*na*	276	1.018	0.273	0.268	0.222	0.046
	S19P	577	102	21	0.571	0.587 (0.047)	0.123	0.215	325	0.587	0.138	0.235	0.210	0.025
Large Subunit Ribosomal Proteins (LSU)	L1P	542	225	21	0.896	0.787 (0.053)	0.106	0.118	325	0.788	0.136	0.173	0.175	-0.003
	L2P[1]	682	269	21	0.707	0.700 (0.034)	0.103	0.146	406	0.697	0.093	0.133	0.170	-0.036
	L3P	481	247	21	0.875	0.986 (0.076)	0.125	0.143	253	0.981	0.205	0.209	0.172	0.036
	L4P	325	203	*na*	*na*	*na*	*na*	*na*	325	1.402	0.308	0.220	0.179	0.041
	L5P	575	181	15	0.838	0.675 (0.058)	0.143	0.171	351	0.679	0.132	0.194	0.183	0.011
	L6P	611	180	21	0.812	1.002 (0.057)	0.145	0.178	351	0.990	0.152	0.153	0.183	-0.030
	L9P	378	155	*na*	*na*	*na*	*na*	*na*	378	1.45	0.258	0.178	0.189	-0.012
	L11P	416	143	10	0.534	0.603 (0.051)	0.160	0.300	276	0.594	0.121	0.204	0.193	0.011
	L13P	412	148	15	1.030	0.891 (0.073)	0.273	0.265	253	0.899	0.165	0.184	0.192	-0.008
	L14P	609	124	15	0.564	0.505 (0.062)	0.089	0.157	378	0.505	0.121	0.240	0.200	0.041
	L15P	300	148	*na*	*na*	*na*	*na*	*na*	300	1.19	0.266	0.223	0.192	0.032
	L16P	325	139	*na*	*na*	*na*	*na*	*na*	325	0.705	0.148	0.210	0.194	0.016
	L17P	253	129	*na*	*na*	*na*	*na*	*na*	253	0.891	0.189	0.212	0.198	0.014
	L18P	351	116	*na*	*na*	*na*	*na*	*na*	351	0.94	0.243	0.259	0.203	0.055
	L19P	276	119	*na*	*na*	*na*	*na*	*na*	276	0.902	0.175	0.194	0.202	-0.008
	L20P	231	119	*na*	*na*	*na*	*na*	*na*	231	0.786	0.184	0.234	0.202	0.032
	L21P	253	106	*na*	*na*	*na*	*na*	*na*	253	1.238	0.232	0.187	0.208	-0.020
	L24P	210	104	*na*	*na*	*na*	*na*	*na*	210	1.029	0.177	0.172	0.209	-0.037
	L27P	300	89	*na*	*na*	*na*	*na*	*na*	300	0.652	0.126	0.193	0.218	-0.024
	L31P	231	76	*na*	*na*	*na*	*na*	*na*	231	1.439	0.463	0.322	0.227	0.095
	L33P	231	54	*na*	*na*	*na*	*na*	*na*	231	1.137	0.523	0.460	0.250	0.210
	L36P	153	38	*na*	*na*	*na*	*na*	*na*	153	0.478	0.144	0.302	0.277	0.025

[1] Control protein.

Table 2.3: Parameters for the 40 ribosomal proteins– Inter-domain comparisons. Abbreviations are: function of each encoded gene (fx); total number of comparisons (total n_s), average sequence length in number of codons (l_s). Inter-domain includes only comparisons across the three domains, archaea, bacteria and eukaryote. There are three sections: archaea and eukaryote (AE), archaea and bacteria (AB), and bacteria and eukaryote (BE). **AE:** number of comparisons (n_s), mean of K (m_{AE}), expected mean of K with standard deviation (m'_{AE} (SD'_{AE})), the standard deviation of K (SD), and relative standard deviation of K (rSD). **AB:** number of comparisons (n_s), mean of K (m_{AB}), standard deviation of K (SD), and relative standard deviation of K (rSD). **BE:** number of comparisons (n_s), mean of K (m_{AE}), expected mean of K with standard deviation (m'_{BE} (SD'_{BE})), standard deviation of K (SD), and relative standard deviation of K (rSD). "*na*" is not applicable, usually because of the absence of sequence for that domain.

				inTER-domain comparisons													
				archaea vs. eukaryote					bacteria vs. archaea				bacteria vs. eukaryote				
	fx	total n_s	l_s	n_s	m_{AE}	m'_{AE} (SD'_{AE})	SD	rSD	n_s	m_{AB}	SD	rSD	n_s	m_{BE}	m'_{BE} (SD'_{BE})	SD	rSD
Small Subunit Ribosomal Proteins (SSU)	S2P	542	252	6	0.818	(-[2])	0.079	0.097	168	0.785	0.088	0.112	28	0.740	(-[2])	0.072	0.097
	S3P	611	234	7	1.010	(-[2])	0.140	0.139	203	0.792	0.104	0.131	29	0.907	(-[2])	0.089	0.098
	S4P	441	203	*na*	*na*	*na*	*na*	*na*	150	1.176	0.122	0.104	*na*	*na*	*na*	*na*	*na*
	S5P	542	183	*na*	*na*	*na*	*na*	*na*	196	0.661	0.114	0.173	*na*	*na*	*na*	*na*	*na*
	S6P	276	118	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*
	S7P	479	167	6	0.652	0.688 (0.059)	0.026	0.040	156	0.686	0.088	0.128	26	0.620	0.685 (0.024)	0.049	0.079
	S8P	479	133	6	0.525	(-[2])	0.047	0.090	156	0.717	0.081	0.113	26	0.690	(-[2])	0.069	0.100
	S9P	510	134	6	0.573	(-[2])	0.046	0.080	162	0.614	0.072	0.117	27	0.572	(-[2])	0.056	0.098
	S10P	496	103	7	0.617	0.527 (0.047)	0.067	0.109	168	0.527	0.056	0.106	24	0.509	0.527 (0.038)	0.036	0.071
	S11P	542	129	6	0.342	(-[2])	0.045	0.132	168	0.407	0.054	0.133	28	0.409	(-[2])	0.047	0.114
	S12P	276	128	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*
	S13P	300	122	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*
	S14P	325	87	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*
	S15P	378	91	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*
	S16P	300	100	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*
	S17P	481	96	7	0.788	0.718 (0.081)	0.103	0.131	161	0.714	0.118	0.165	23	0.616	0.718 (0.091)	0.054	0.088
	S18P	276	83	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*
	S19P	577	102	7	0.607	0.568 (0.044)	0.066	0.108	196	0.566	0.072	0.128	28	0.621	0.566 (0.048)	0.048	0.077
Large Subunit Ribosomal Proteins (LSU)	L1P	542	225	*na*	*na*	*na*	*na*	*na*	*182*	*0.684*	*0.073*	*0.107*	*na*	*na*	*na*	*na*	*na*
	L2P[1]	682	269	7	0.546	0.632 (0.023)	0.040	0.074	217	0.632	0.055	0.086	31	0.632	0.632 (0.047)	0.039	0.061
	L3P	481	247	7	0.784	0.763 (0.091)	0.056	0.071	175	0.763	0.098	0.128	25	0.788	0.763 (0.018)	0.094	0.119
	L4P	325	203	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*
	L5P	575	181	6	0.531	(-[2])	0.048	0.090	174	0.706	0.062	0.088	29	0.646	(-[2])	0.047	0.073
	L6P	611	180	7	0.751	(-[2])	0.030	0.040	203	0.820	0.091	0.111	29	0.818	(-[2])	0.059	0.072
	L9P	378	155	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*
	L11P	416	143	*na*	*na*	*na*	*na*	*na*	130	0.504	0.047	0.094	*na*	*na*	*na*	*na*	*na*
	L13P	412	148	*na*	*na*	*na*	*na*	*na*	144	0.735	0.084	0.114	*na*	*na*	*na*	*na*	*na*
	L14P	609	124	6	0.511	0.428 (0.027)	0.033	0.064	180	0.431	0.054	0.125	30	0.487	0.431 (0.044)	0.039	0.079
	L15P	300	148	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*
	L16P	325	139	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*
	L17P	253	129	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*
	L18P	351	116	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*
	L19P	276	119	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*
	L20P	231	119	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*
	L21P	253	106	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*
	L24P	210	104	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*
	L27P	300	89	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*
	L31P	231	76	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*
	L33P	231	54	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*
	L36P	153	38	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*

[1] Control protein.

[2] Because the archaea and the bacteria do not share the same evolutionary history, m'_{AE} and m'_{BE} were not calculated.

Table 2.4: Parameters for the aminoacyl-tRNA-synthetases and SecY – Intra-domain comparisons. Abbreviations are: function of each encoded gene (fx); total number of comparisons (total n_s), average sequence length in number of codons (l_s). Intra-domain includes only comparisons within the archaea or bacteria. There are two sections that are for intra-domain comparisons: archaea with other archaea (AA) and bacteria with other bacteria (BB). **AA:** number of comparisons (n_s), mean of K (m_{AA}), expected mean of K with standard deviation (m'_{AA} (SD'_{AA})), standard deviation of K (SD), and relative standard deviation of K (rSD). **BB:** number of comparisons (n_s), mean of K (m_{BB}), standard deviation of K (SD), relative standard deviation of K (rSD), relative standard deviation expected from l_s (rSD^*), and $rSD - rSD^*$ (D_{rSD}). "*na*" is not applicable, usually because of the absence of sequence for that domain.

				inTRA-domain comparisons										
				archaea vs. archaea					bacteria vs. bacteria					
	fx	total n_s	l_s	n_s	m_{AA}	m'_{AA} (SD'_{AA})	SD	rSD	n_s	m_{BB}	SD	rSD	rSD^*	D_{rSD}
Aminacyl-tRNA Synthetases	Ala	510	877	15	1.153	1.156 (0.055)	0.226	0.196	300	1.158	0.172	0.149	0.155	-0.007
	Arg	542	571	15	1.603	B1:1.793($-^1$) B2:1.620(0.147)	0.633	0.395	325	1.995	0.608	0.305	0.156	0.149
	Asn	142	456	1	1.200	($-^1$)	*na*	($-^1$)	91	1.135	0.378	0.333	0.159	0.175
	Asp	646	552	21	0.859	0.971 (0.047)	0.156	0.181	378	0.969	0.177	0.182	0.156	0.026
	Cys	508	477	10	1.091	1.232 (0.094)	0.360	0.330	325	1.225	0.205	0.168	0.158	0.010
	Glu	561	500	15	1.099	1.381 (0.152)	0.156	0.142	378	1.429	0.321	0.225	0.157	0.067
	Gly (alpha)	190	302	*na*	*na*	*na*	*na*	*na*	190	0.671	0.107	0.159	0.167	-0.007
	Gly (beta)	210	651	*na*	*na*	*na*	*na*	*na*	210	1.596	0.223	0.140	0.155	-0.015
	His	630	421	21	1.613	B1:1.358(0.068) B2:($-^1$)	0.513	0.318	378	1.684	0.579	0.344	0.160	0.184
	Ile	542	983	15	0.993	B1:1.245(0.056) B2:1.191(0.021)	0.116	0.117	325	1.646	0.522	0.317	0.156	0.161
	Leu[2]	682	863	21	1.032	1.118 (0.056)	0.151	0.147	406	1.111	0.170	0.153	0.155	-0.002
	Met	644	645	15	1.288	B1:1.158(0.067) B2:1.195(0.060)	0.090	0.070	406	1.786	0.484	0.271	0.155	0.116
	Phe (alpha)	609	379	15	1.752	B1:1.028(0.046) B2:($-^1$)	0.467	0.267	378	1.157	0.360	0.311	0.162	0.150
	Phe (beta)	607	742	15	1.386	B1:1.731(0.072) B2:($-^1$)	0.216	0.156	406	1.960	0.672	0.343	0.155	0.188
	Pro	611	540	21	0.878	B1:1.226 (0.052) B2:1.117(0.049)	0.122	0.139	351	1.691	1.187	0.702	0.157	0.545
	Ser	445	432	6	0.861	1.123 (0.143)	0.129	0.150	300	1.118	0.247	0.221	0.159	0.062
	Thr	575	626	15	A_E: 0.973	A_E:1.097(0.066) A_C:($-^1$)	0.905	0.578	351	1.081	0.157	0.145	0.155	-0.011
	Trp	575	350	21	1.287	1.277(0.134)	0.381	0.296	351	1.429	0.523	0.366	0.163	0.203
	Tyr	508	398	10	0.894	B1:1.190(0.085) B2:1.407(0.359)	0.164	0.184	325	1.672	0.903	0.540	0.161	0.380
	Val	609	894	21	1.200	B1:1.149(0.041) RP:($-^1$)	0.210	0.175	378	1.192	0.330	0.277	0.155	0.122
	SecY[2]	682	433	21	1.261	1.189 (0.069)	0.110	0.087	406	1.192	0.168	0.141	0.159	-0.019

[1] Number of comparisons was too small to calculate due to low number of organisms.

[2] Control protein.

Table 2.4: Parameters for the aminoacyl-tRNA-synthetases and SecY – Inter-domain comparisons. Abbreviations are: function of each encoded gene (fx); total number of comparisons (total n_s), average sequence length in number of codons (l_s). Inter-domain includes only comparisons across the three domains, archaea, bacteria and eukaryote. There are three sections: archaea and eukaryote (AE), archaea and bacteria (AB), and bacteria and eukaryote (BE). **AE:** number of comparisons (n_s), mean of K (m_{AE}), expected mean of K with standard deviation (m'_{AE} (SD'_{AE})) – when two clades were present the expectation was calculated for each one individually, the standard deviation of K (SD), and relative standard deviation of K (rSD). **AB:** number of comparisons (n_s), mean of K (m_{AB}), standard deviation of K (SD), and relative standard deviation of K (rSD). **BE:** number of comparisons (n_s), mean of K (m_{AE}), expected mean of K with standard deviation (m'_{BE} (SD'_{BE})) – when two clades were present the expectation was calculated for each one individually, standard deviation of K (SD), and relative standard deviation of K (rSD). "*na*" is not applicable, usually because of the absence of sequence for that domain.

				inTER-domain comparisons													
				archaea vs. eukaryote					bacteria vs. archaea				bacteria vs. eukaryote				
	fx	total n_s	l_s	n_s	m_{AE}	m'_{AE} (SD'_{AE})	SD	rSD	n_s	m_{AB}	SD	rSD	n_s	m_{BE}	m'_{BE} (SD'_{BE})	SD	rSD
Aminoacyl-tRNA Synthetases	Ala	510	877	6	1.428	0.899 (0.068)	0.054	0.038	162	0.899	0.077	0.086	27	0.472	0.899 (0.027)	0.060	0.127
	Arg	542	571	6	1.144	B1:0.854(0.074) B2:0.800(0.071)	0.103	0.090	168	0.812	0.107	0.132	28	0.677	B1:0.854(0.085) B2:0.800(0.042)	0.147	0.217
	Asn	142	456	2	0.850	B1:0.679(0.026) B2:0.605(0.028)	0.018	0.021	32	0.565	0.066	0.115	16	0.586	B1:0.479(0.006) B2:0.605(0.010)	0.049	0.083
	Asp	646	552	7	0.796	0.699 (0.053)	0.033	0.042	210	0.699	0.057	0.082	30	0.720	0.699 (0.039)	0.041	0.056
	Cys	508	477	5	0.849	0.479 (0.083)	0.083	0.098	140	0.479	0.093	0.194	28	0.519	0.479 (0.039)	0.049	0.094
	Glu	561	500	*na*	*na*	*na*	*na*	*na*	168	1.219	0.087	0.071	*na*	*na*	*na*	*na*	*na*
	Gly (alpha)	190	302	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*
	Gly (beta)	210	651	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*	*na*
	His	549	421	7	1.333	*na*	0.212	0.159	196	0.809	0.106	0.131	28	0.885	*na*	0.127	0.144
	Ile	542	983	6	0.957	B1:0.764(0.048) B2:0.649(0.036)	0.055	0.057	168	0.731	0.077	0.105	28	0.664	B1:0.764(0.037) B2:0.649(0.018)	0.151	0.227
	Leu[2]	682	863	7	1.018	1.344 (0.045)	0.079	0.077	217	1.344	0.070	0.052	31	1.270	1.344 (0.037)	0.054	0.042
	Met	644	645	6	0.914	B1:0.624(0.062) B2:0.804(0.060)	0.092	0.101	186	0.717	0.125	0.174	31	0.711	B1:0.624(0.036) B2:0.804(0.064)	0.158	0.222
	Phe (alpha)	609	379	6	1.098	B1:0.767(0.048) B2:0.668(0.040)	0.079	0.072	180	0.761	0.099	0.131	30	0.655	B1:0.767(0.089) B2:0.668(0.040)	0.037	0.057
	Phe (beta)	607	742	*na*	*na*	*na*	*na*	*na*	186	1.218	0.198	0.162	*na*	*na*	*na*	*na*	*na*
	Pro	611	540	7	2.253	B1:0.671(0.131) B2:1.321(0.071)	0.111	0.049	203	1.209	0.271	0.224	29	0.794	B1:0.671(0.052) B2:1.321(0.071)	0.391	0.492
	Ser	445	432	4	0.761	0.546 (0.059)	0.063	0.083	108	0.554	0.074	0.134	27	0.542	0.550 (0.046)	0.048	0.089
	Thr	575	626	6	1.602	EA:1.043(0.063) CA:0.668(0.053)	0.297	0.186	174	0.981	0.157	0.160	29	0.458	EA:1.043(0.021) CA:0.669 (-[1])	0.045	0.099
	Trp	575	350	*na*	*na*	*na*	*na*	*na*	203	1.015	0.148	0.145	*na*	*na*	*na*	*na*	*na*
	Tyr	508	398	5	0.953	B1:1.374(0.094) B2:1.119(0.108)	0.147	0.155	140	1.283	0.197	0.154	28	1.357	B1:1.374(0.130) B2:1.119(0.063)	0.195	0.144
	Val	609	894	na	na	na	na	na	210	0.774	0.089	0.115	na	na	na	na	na
	SecY[2]	682	433	7	1.436	1.024 (0.050)	0.075	0.053	217	1.025	0.083	0.081	31	1.098	1.024 (0.053)	0.067	0.061

[1] Number of comparisons was too small to calculate due to low number of organisms.

[2] Control protein.

CHAPTER 3

RED-T: AN APPLICATION UTILIZING THE RATIOS OF EVOLUTIONARY DISTANCES FOR DETERMINATION OF ALTERNATIVE PHYLOGENETIC EVENTS[†]

[†] Farahi, K., W.B. Whitman, E.T. Kraemer. Submitted to *Bioinformatics*. 12/11/ 2002.

Summary and Availability

RED-T is a Java application for phylogenetic analysis that is based on a unique method, RED, that utilizes the ratios of evolutionary distances to distinguish between alternative evolutionary histories. RED-T allows the user to examine if any given experimental gene shares the same evolutionary history as the designated control gene(s). Moreover, the tool detects any differences in evolutionary history, and allows the user to examine comparisons of E_d for a likely explanation. Lateral gene transfer (LGT), which may have a significant influence in organismal evolution especially in prokaryote evolution, is one mechanism that could explain the findings of these RED-T analyses.

Availability: The application is available online at http://www.arches.uga.edu/~whitman/RED.

Contact: Kamyar Farahi, 541 Biological Sciences Building, Department of Microbiology, The University of Georgia, Athens, GA 30602, USA; *telephone*: 706.542.4692, *fax*: 706.542.2674, *email*: kfarahi@arches.uga.edu. Eileen Kraemer, 415 GSRC, Computer Science Department, The University of Georgia, Athens GA, 30602, USA, *telephone*: 706.542.5799, *fax*: 706.542.2966, *email*: eileen@cs.uga.edu.

Introduction and Application Description

Deciphering organismal evolution and detecting phylogenetic events have captured biologists' attention for generations. With the growth of genomic data and development of advanced methods of analysis, evolutionary biology has taken a great leap forward in understanding the evolutionary history of prokaryotes. An important discovery has been the realization that lateral gene transfer (LGT) occurs at high rates in many prokaryotic lineages (Lawrence and Ochman 1998; Doolittle 1999; Nelson, Clayton et al. 1999; Lawrence and Ochman 2002). Furthermore, LGT may have been one of the major evolutionary mechanisms that led to the formation of the modern lineages of prokaryotes (Snel, Bork et al. 2002).

Methods to detect LGT depend upon searching for incongruities in phylogenetic trees or differences in the DNA composition of LGT genes (Lawrence and Ochman 2002). Given its importance, we sought to develop an alternative strategy to detect LGTs. Similar to phylogenetic trees, this method relies upon phylogenetic analyses to determine evolutionary distances (E_d). It assumes that if the rates of evolution in two genes are constant, then plots of the E_d for one gene against another gene should be linear. Therefore, it is possible to use empirical tests rather than tree building to identify non-vertical evolution. For the application of this method, RED, an interactive research tool entitled RED-T – Ratios of Evolutionary Distances for determination of alternative phylogenetic events was developed. RED-T is a Java application capable of generating scatter plots from given distance matrixes to analyze evolutionary relationships among various levels of taxa. In addition, it is fully capable of importing new gene data for

comparison with the control set we developed (RED control) or allowing the user to develop a new control.

With the assistance of the application wizard, the user can import Phylip formatted distance matrices (Felsenstein 1993), or columnar formats (see RED help files). RED-T allows the user to rename and map each sequence ID that is imported to the appropriate organism and taxonomy provided by the tools' catalogue. This resource is in accordance with EMBL classifications (Stoesser, Baker et al. 2002) and fully adjustable by advanced users. In addition, imported sequence IDs from the WIT/ERGO database are automatically mapped and can then be edited by the user (Overbeek, Larsen et al. 2000). One other advantage of the mapping step is to allow the user to analyze multiple paralogous genes of an organism or examine unknown sequences.

The imported data can be used as the experimental gene for comparison with the tool's default control, which was developed from a set of genes with a constant evolutionary history. Alternative to the RED control, the user can develop an original control based on RED's protocol, which is simulated by the tool's wizard. This allows the analysis of different evolutionary histories, such as rapid evolution of the *Mycoplasma* gene family or even transposon phylogeny.

The RED-T provides the user with extensive analysis tools to evaluate evolutionary relationships of a gene, ranging from intra-domain and inter-domain comparisons (e.g. archaeal domain versus groups of proteobacteria) to specific relationships of two organisms. In addition, the tool calculates quantitative variables reflecting values for the selected comparisons, provides highlight mode and zoom settings to evaluate the user's selection with other comparisons, and a journal interface

for analysis notes (Figure 3.1). Any content of the analyses can be printed for further study. Also, phylogenetic trees can be imported for the user to compare and contrast with the RED-T plot. This feature also allows the user to import other graphical references, such as tables, graphs or diagrams to be saved under the same analyses folder for further efforts.

In order to evaluate the RED method, RED-T was applied to a data set of ribosomal proteins believed to have been vertically inherited (Hansmann and Martin 2000). Next, 20 aminoacyl-tRNA synthetases were chosen to check for alternative phylogenetic events, such as LGT. Although these synthetases are members of a functionally conserved family, LGT has been proposed for the evolution of many members of this ancient family (Woese, Olsen et al. 2000; Boucher, Huber et al. 2001). These and other genes are available as samples in the RED-T package, version 2.1. When the plots of the E_d's are nonlinear, LGT is one possible mechanism. However, any perturbation of the evolutionary clock, such as changes in gene function or evolutionary rate, could in theory be detected. The advantages of this approach are: the variability inherent in comparing protein sequences is transparent, the direction of LGT and the relative rates of evolution are readily identified, and it is possible to detect other types of evolutionary events. For more understanding of the theory and development of the RED method and further biological applications of RED-T, see Farahi, et al. (to be submitted).

RED-T application is available as a Java™ Archive (JAR) file. This compression format allows the user to launch the full application on any operating environment providing the system has JRE (Java™ Runtime Environment) version 1.4 or above. In addition, the complete source code and documentation are available for downloading by

registered users. For additional information regarding current and upcoming features, please visit http://www.arches.uga.edu/~whitman/RED.

References

Boucher,Y., Huber, H., L'Haridon,S., Stetter,K.O., Doolittle,W.F. (2001). Bacterial origin for the isoprenoid biosynthesis enzyme HMG-CoA reductase of the archaeal orders Thermoplasmatales and Archaeoglobales. *Mol Biol Evol*. **18**, 1378-88.

Doolittle,W.F. (1999). Phylogenetic classification and the universal tree. *Science*. **284**, 2124-9.

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

Hansmann,S. and Martin,W. (2000). Phylogeny of 33 ribosomal and six other proteins encoded in an ancient gene cluster that is conserved across prokaryotic genomes: influence of excluding poorly alignable sites from analysis. *Int J Syst Evol Microbiol*. **50**, 1655-63.

Lawrence,J.G. and Ochman,H. (1998). Molecular archaeology of the *Escherichia coli* genome. *Proc Natl Acad Sci U S A*. **95**, 9413-7.

Lawrence,J.G. and Ochman,H. (2002). Reconciling the many faces of lateral gene transfer. *Trends Microbiol*. **10**, 1-4.

Nelson,K.E., Clayton,R.A., Gill,S.R., Gwinn,M.L., Dodson,R.J., Haft,D.H., Hickey,E.K., Peterson,J.D., Nelson,W.C., Ketchum,K.A., McDonald,L., Utterback,T.R., Malek,J.A., Linher,K.D., Garrett,M.M., Stewart,A.M., Cotton,M.D., Pratt,M.S., Phillips,C.A., Richardson,D., Heidelberg,J., Sutton,G.G., Fleischmann,R.D., Eisen,J.A., Fraser,C.M.,et al. (1999). Evidence for lateral gene transfer between

Archaea and bacteria from genome sequence of *Thermotoga maritima*. *Nature*. **399**, 323-9.

Overbeek,R., Larsen,N., Pusch,G.D., D'Souza,M., Selkov,E.,Jr., Kyrpides,N., Fonstein,M., Maltsev,N., Selkov,E. (2000). WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction. *Nucleic Acids Res*. **28**, 123-5.

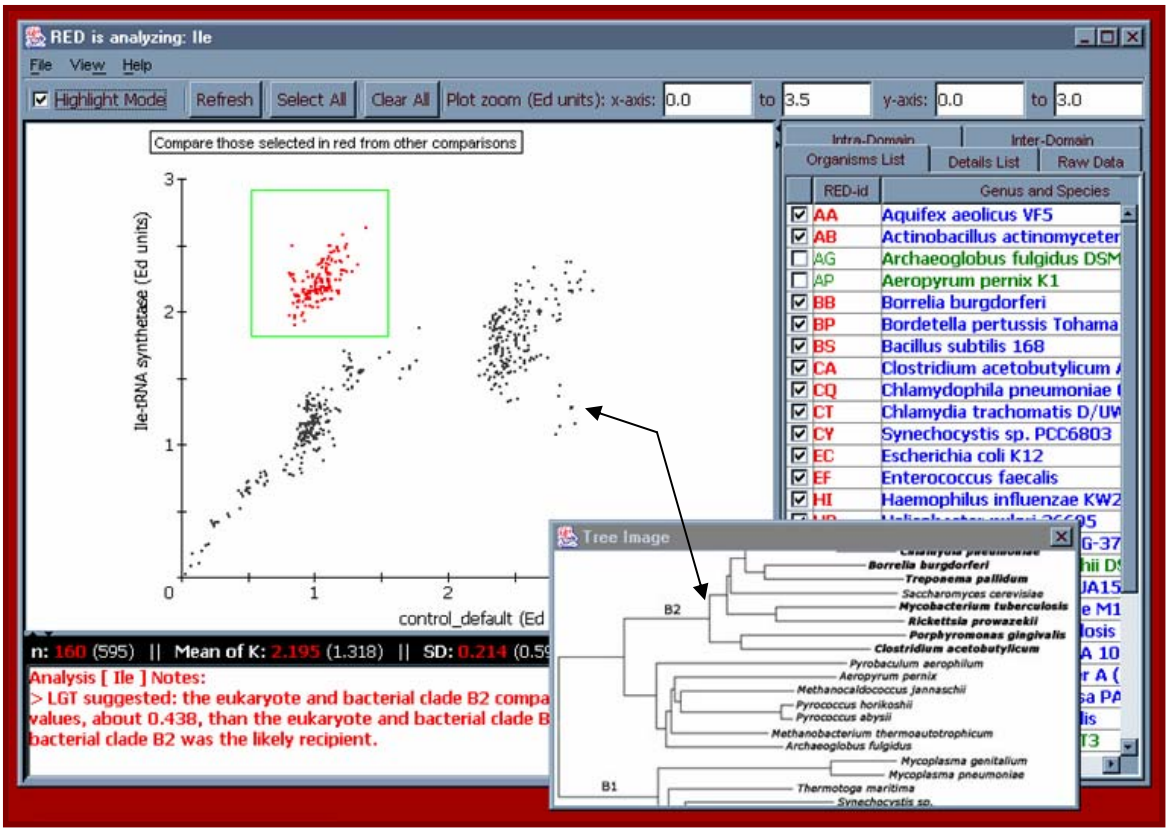
Snel,B., Bork,P., Huynen,M.A. (2002). Genomes in flux: the evolution of archaeal and proteobacterial gene content. *Genome Res*. **12**, 17-25.

Stoesser,G., Baker,W., van den Broek,A., Camon,E., Garcia-Pastor,M., Kanz,C., Kulikova,T., Leinonen,R., Lin,Q., Lombard,V., Lopez,R., Redaschi,N., Stoehr,P., Tuli,M.A., Tzouvara,K., Vaughan,R. (2002). The EMBL Nucleotide Sequence Database. *Nucleic Acids Res*. **30**, 21-6.

Woese,C.R., Olsen,G.J., Ibba,M., Soll,D. (2000). Aminoacyl-tRNA synthetases, the genetic code, and the evolutionary process. *Microbiol Mol Biol Rev*. **64**, 202-36.

Figure 3.1: **Screen shot of RED-T application.** The main window is separated into three resizable partitions: The main partition contains the plot generator as illustrated here in highlight mode and a bar displaying quantitative variables that reflect values for the selected portion of the plot. The second partition contains a text box for taking notes of analysis. The last partition contains five tabbed windows to provide different levels of taxonomic selection. In addition, the floating window displays an imported phylogenetic tree. Illustrated here is the analysis of the isoleucyl-tRNA synthetase gene, in highlight mode. High *K* values resulted from comparisons between bacterial clades B1 and B2 (outlined box), suggesting that one of these clades obtained its gene from another domain. It is clear that the eukaryote and bacterial clade B2 (arrows in plot and tree window) comparisons had much lower *K* values than the eukaryote and bacterial clade B1 comparisons, thus bacterial clade B2 was the likely recipient of the eukaryotic isoleucyl-tRNA synthetase gene.

Figure 3.1


RED is analyzing: Ile
File View Help
Highlight Mode
Refresh
Select All
Clear All
Plot zoom (Ed units): x-axis: 0.0 to 3.5 y-axis: 0.0 to 3.0
Compare those selected in red from other comparisons
Ile-tRNA synthetase (Ed units)
control_default (Ed
n: 160 (595) || Mean of K: 2.195 (1.318) || SD: 0.214
Analysis [Ile] Notes:
> LGT suggested: the eukaryote and bacterial clade B2 compa
values, about 0.438, than the eukaryote and bacterial clade B
bacterial clade B2 was the likely recipient.
Intra-Domain
Inter-Domain
Organisms List
Details List
Raw Data
RED-id
Genus and Species
AA Aquifex aeolicus VF5
AB Actinobacillus actinomyceter
AG Archaeoglobus fulgidus DSM
AP Aeropyrum pernix K1
BB Borrelia burgdorferi
BP Bordetella pertussis Tohama
BS Bacillus subtilis 168
CA Clostridium acetobutylicum
CQ Chlamydophila pneumoniae
CT Chlamydia trachomatis D/UW
CY Synechocystis sp. PCC6803
EC Escherichia coli K12
EF Enterococcus faecalis
HI Haemophilus influenzae KW2
Tree Image
Borrelia burgdorferi
Treponema pallidum
Saccharomyces cerevisiae
Mycobacterium tuberculosis
Rickettsia prowazekii
Porphyromonas gingivalis
Clostridium acetobutylicum
B2
Pyrobaculum aerophilum
Aeropyrum pernix
Methanocaldococcus jannaschii
Pyrococcus horikoshii
Pyrococcus abysii
Methanobacterium thermoautotrophicum
Archaeoglobus fulgidus
Mycoplasma genitalium
Mycoplasma pneumoniae
B1
Thermotoga maritima
Synechocystis sp.

CHAPTER 4

CONCLUSIONS

The availability of large numbers of genomic sequences has demonstrated the importance of lateral gene transfer (LGT) in prokaryotic evolution. However, there remains considerable uncertainty concerning the frequency of LGT compared to other evolutionary processes. We successfully set out to examine the frequency of LGT in ancient lineages and develop an alternative method to distinguish between possible evolutionary histories. This method utilized the Ratios of Evolutionary Distances (or RED). The advantages of this approach were demonstrated: the variability inherent in comparing protein sequences was transparent, the direction of LGT and the relative rates of evolution were readily identified, and it was possible to detect other types of evolutionary events. In accordance with our hypothesis, RED was standardized using genes that were believed to share a vertical evolution. A number of quantitative measurements, such as the distribution of E_d, were set as empirical thresholds to further distinguish various different evolutionary models. Following standardization, the evolutionary histories of aminoacyl-tRNA synthetases were examined. RED-T, an original computer program designed to implement the RED method, was developed during this work.

While examining the aminoacyl-tRNA synthetases, evidence for different models were observed using the RED method. One model represented evolution in the absence of

a LGT event and solely a vertical evolution. In addition to most ribosomal proteins, the evolution of the aspartyl-tRNA synthetase and alpha and beta chains of the glycyl-tRNA synthetases were consistent with this model. Next, models that hypothesized various types of LGT events were most common. The analyses of four of the remaining tRNA synthetases – arginyl-, histidyl-, prolyl-, tryptophanyl-, were consistent with combinations of various types of LGTs categorized by rank of donor and recipient. One type was shared by the evaluation of the arginyl-, histidyl-, methionyl-, seryl-, and valyl-tRNA synthetase which proposes LGT from one domain to a recent lineage, such as a specific bacterial genus of another domain. The next type proposed an ancient LGT event between two domains, such as to a bacterial group. Evidence for this type was the most common among the aminoacyl-tRNA synthetases, with five of the aminoacyl-tRNA synthetases, alanyl-, arginyl-, asparaginyl-, histidyl-, isoleucyl- and tryptophanyl-tRNA synthetase possessing this pattern of evolution. Prolyl- and threonyl-tRNA synthetases represented another type of LGT, which proposed a different direction of transfer from the previous types. Lateral transfer was proposed from a bacterial group to the archaea or the eukaryote domain. Another LGT type proposed an early LGT event. It was defined as a relatively modern origin for the gene with LGT to a number of lineages within a short period of time. However, a similar pattern might be caused by a recent increase in the rate of evolution. Cysteinyl-tRNA synthetase and 10 other aminoacyl-tRNA synthetases, illustrated an ancient nonlinearity which was either due to the increase in rate of evolution of this gene which may have been slower prior to the formation of the domains, or a LGT that occurred in the lineages ancestral to formation of the domains. The last proposal of by LGT was portrayed by LGT type that showed LGT events occurred within

a domain. Arginyl-, glutamyl- and tryptophanyl-tRNA synthetase were examples of this type. Finally, an alternative model was observed by RED, of gene evolution that supports events resulting from divergence of multiple copies of a gene among various taxa. Tyrosyl-tRNA synthetase was an example for this model.

Our results showed that although LGTs were common in the evolution of the aminoacyl-tRNA synthetases, they were not sufficient to obscure the organismal phylogeny. Moreover, much of the apparent complexity of the gene tree was consistent with the formation of the paralogs in the ancestors to the modern lineages followed by more recent loss of one paralog or the other in modern lineages.

The current work utilized the RED method to study genes involved in translation, a highly conserved cellular process, but this method may be more generally applicable to less conserved groups of genes. For instance, it may be possible to develop new sets of controls using the RED-T application to observe the genetic history within the proteobacteria, the unique and diverse pathways within methanogens, identify the *Pyrococcus* spp.'s distinctiveness within the archaeal domain, or set a standard for fast-clock organisms such as the *Mycoplasma*.

Overall, it is important to mention that genomics has a lot more to say about the influence of LGT on cellular evolution. In addition, considering the common occurrences of this evolutionary mechanism, LGT evidently has significant impact on evolution. In particular, gene transfers may have direct affects on genes important for biotechnology, environmental studies, food sciences, and pharmaceutical products. Therefore, it is vital to develop new methods to analyze upcoming data, both quantitatively and qualitatively. Currently, phylogenetic methods that detect a gene's irregular distribution among

organisms appear to be the most informative to identify LGT. Although, humans still will prove to outperform much software, more automated methods might be developed to handle the great volume of genomic data. The large-scale data analysis methods that rely greatly on statistical models need to be improved. This has lead to numerous misinterpretations because the basic logic of biology has been difficult to mimic. Therefore, this field must recognize the importance of traditional research and commit itself to examining genomic data from the biological point of view and not rely solely on data processing. Although improvements have been encouraging, the gap between these two opposite point of views must lessen. One idea would be to present the computational methods with more user-friendly interfaces, or interpret the genomics data using visual methods. A true balance must be established, and with time and patience the methods can be developed to use genomics in revolutionary ways.

APPENDIX A

INITIAL RIBOSOMAL PROTEIN L2P CONTROL STUDIES

Additional data that was mentioned but not presented in the "designation of control proteins" section of Chapter 2 is presented here. Plots of L2P used as a standard with other genes are shown in Figure A1-16. This standard was compared to the other five genes which also yielded linear plots. In addition, L2P was compared to 11 other genes selected from the initial 32 genes. Overall, the 16 genes represented a diverse range that had produced both linear and non-linear plots in the initial analyses. As expected, the L2P compared with the earlier five genes (Figure A4-6, 11 and 13) yielded obvious linear plots. Some of the other plots were roughly linear, and appeared to contain linear as well as nonlinear components. Some contained a linear component near the origin but appeared to flatten out at higher values (Figure A2, 7-8, 10, 12, 14 and 16). Others had two or more clusters with both linear and non-linear patterns. These plots visually suggested a high variability, which was confirmed by calculating the relative standard deviation (*rSD*) of the value *K*, or the ratio of E_d(experimental) to E_d(L2P). For all 16 plots, the *rSD* of *K* ranged from 0.153 to 0.506 (Table A1). The linear plots had a lower *rSD* than the nonlinear (heavily scattered) plots.

<u>Figure A1-16</u>: Plots of L2P used as a standard to compare with 16 other genes.

These plots were created in the early stages of RED method development (see Chapter 2). Microsoft Excel software was used to generate these E_d comparison plots. E_d for each gene was calculated according to the Methods in Chapter 2. Each of the genes analyzed here contained a different set of organismal sample.

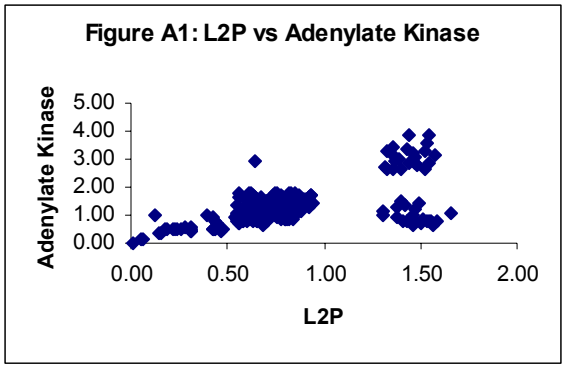


Figure A1: L2P vs Adenylate Kinase
Adenylate Kinase
5.00
4.00
3.00
2.00
1.00
0.00
0.00
0.50
1.00
1.50
2.00
L2P

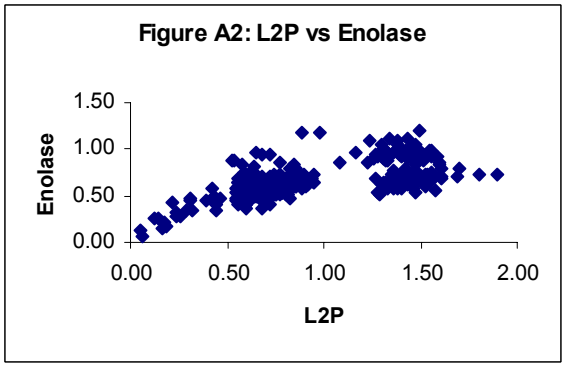


Figure A2: L2P vs Enolase
Enolase
1.50
1.00
0.50
0.00
0.00
0.50
1.00
1.50
2.00
L2P

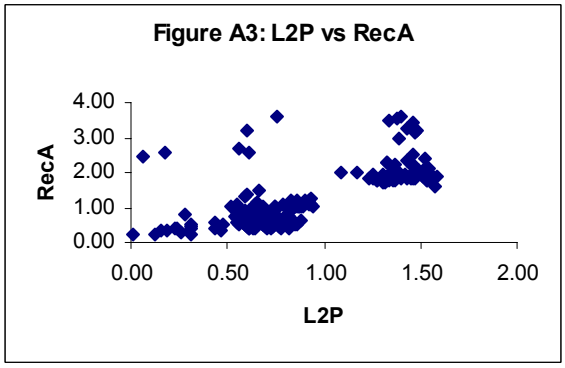


Figure A3: L2P vs RecA
RecA
4.00
3.00
2.00
1.00
0.00
0.00
0.50
1.00
1.50
2.00
L2P

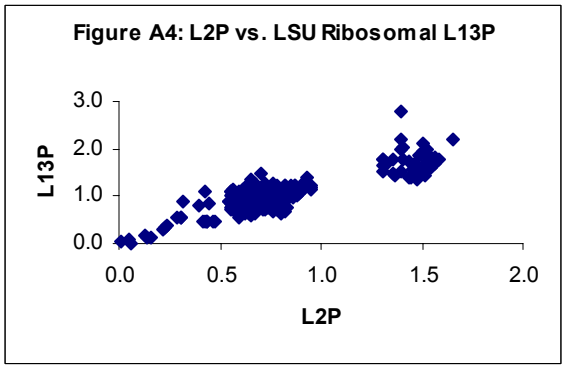


Figure A4: L2P vs. LSU Ribosomal L13P
L13P
3.0
2.0
1.0
0.0
0.0
0.5
1.0
1.5
2.0
L2P

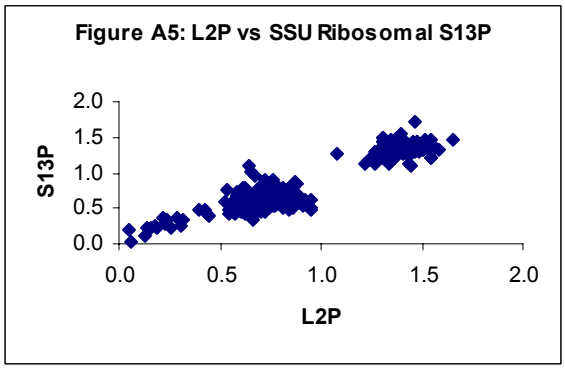


Figure A5: L2P vs SSU Ribosomal S13P
S13P
2.0
1.5
1.0
0.5
0.0
0.0
0.5
1.0
1.5
2.0
L2P

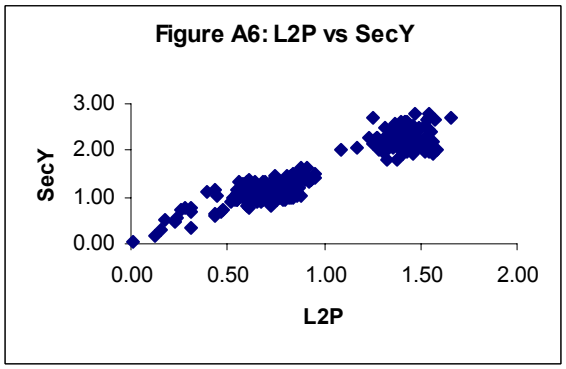


Figure A6: L2P vs SecY
SecY
3.00
2.00
1.00
0.00
0.00
0.50
1.00
1.50
2.00
L2P

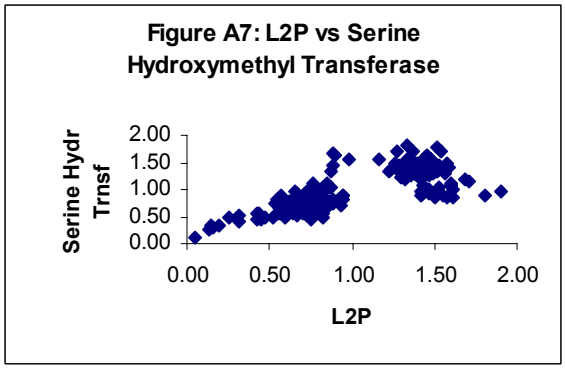


Figure A7: L2P vs Serine Hydroxymethyl Transferase
Serine Hydr Trnsf
2.00
1.50
1.00
0.50
0.00
0.00
0.50
1.00
1.50
2.00
L2P

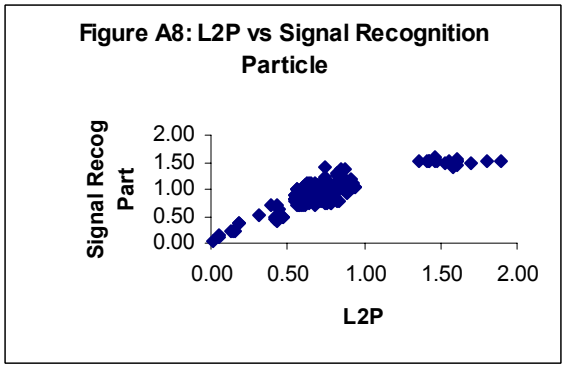


Figure A8: L2P vs Signal Recognition Particle
Signal Recog Part
2.00
1.50
1.00
0.50
0.00
0.00
0.50
1.00
1.50
2.00
L2P

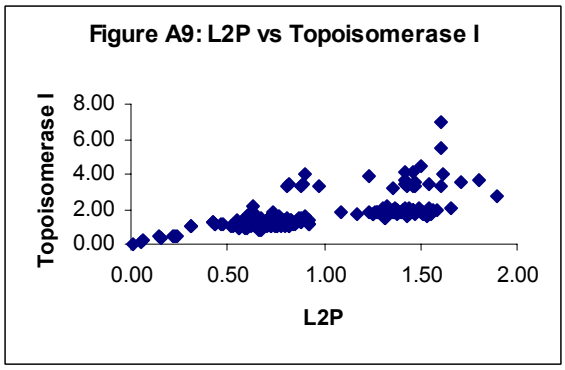


Figure A9: L2P vs Topoisomerase I
Topoisomerase I
8.00
6.00
4.00
2.00
0.00
0.00
0.50
1.00
1.50
2.00
L2P

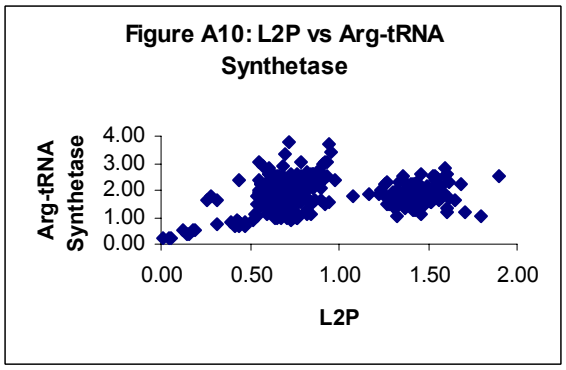


Figure A10: L2P vs Arg-tRNA Synthetase
Arg-tRNA Synthetase
4.00
3.00
2.00
1.00
0.00
0.00
0.50
1.00
1.50
2.00
L2P

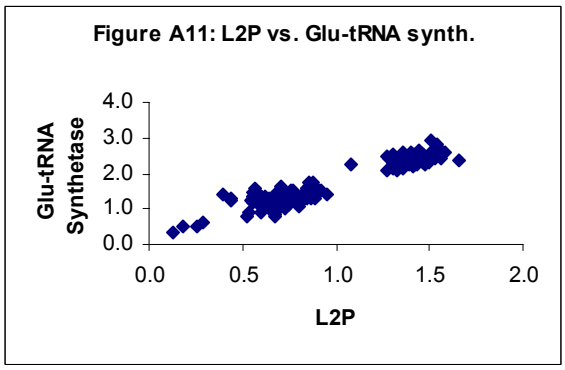


Figure A11: L2P vs. Glu-tRNA synth.
Glu-tRNA Synthetase
4.0
3.0
2.0
1.0
0.0
0.0
0.5
1.0
1.5
2.0
L2P

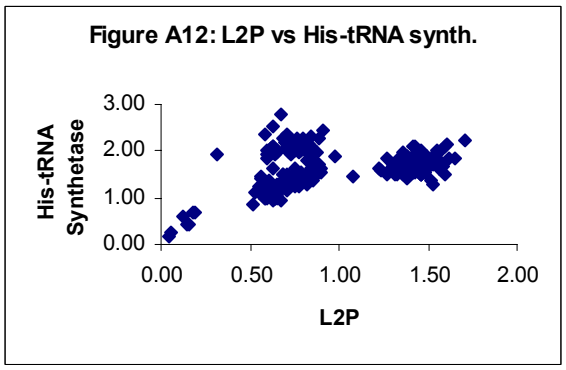


Figure A12: L2P vs His-tRNA synth.
His-tRNA Synthetase
3.00
2.00
1.00
0.00
0.00
0.50
1.00
1.50
2.00
L2P

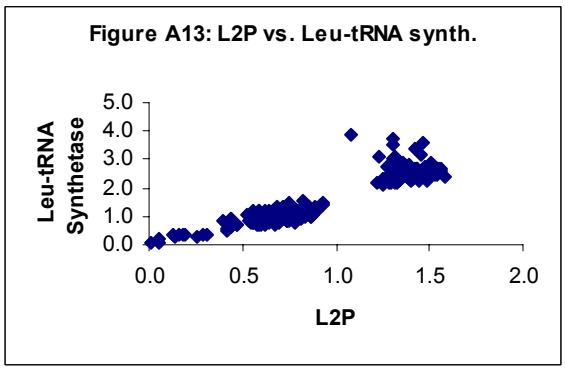


Figure A13: L2P vs. Leu-tRNA synth.
Leu-tRNA Synthetase
5.0
4.0
3.0
2.0
1.0
0.0
0.0
0.5
1.0
1.5
2.0
L2P

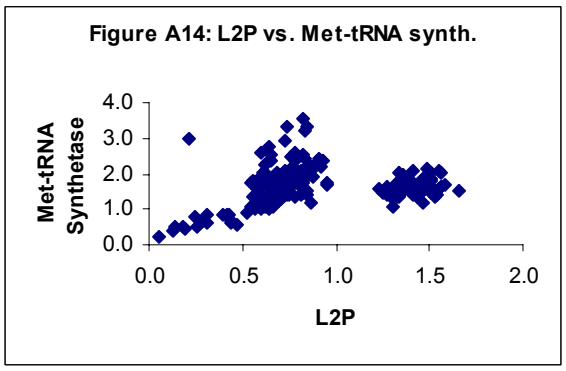


Figure A14: L2P vs. Met-tRNA synth.
Met-tRNA Synthetase
4.0
3.0
2.0
1.0
0.0
0.0
0.5
1.0
1.5
2.0
L2P

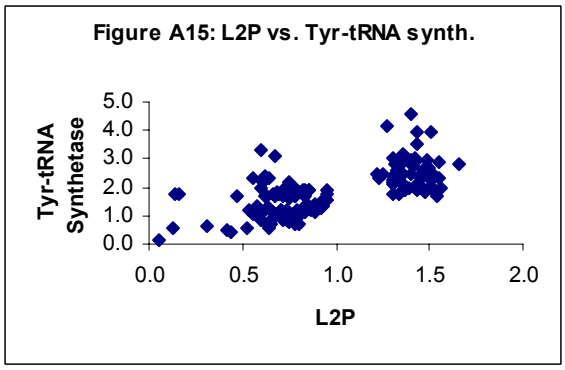


Figure A15: L2P vs. Tyr-tRNA synth.
Tyr-tRNA Synthetase
5.0
4.0
3.0
2.0
1.0
0.0
0.0
0.5
1.0
1.5
2.0
L2P

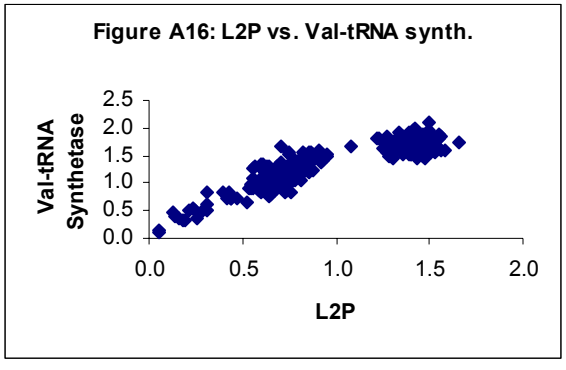


Figure A16: L2P vs. Val-tRNA synth.
Val-tRNA Synthetase
2.5
2.0
1.5
1.0
0.5
0.0
0.0
0.5
1.0
1.5
2.0
L2P

Table A1: Statistics for gene plots in Figure A1-16. *K* was the ratio of E_d (experimental) divided by the E_d (L2P) for each pair of genes compared; *m* was the mean of *K*; *SD* was the standard deviation of *K*; relative *SD* (*rSD*) was the *SD* divided by *m*.

Related plots of Figure A	Experimental proteins	*m*	*SD*	*rSD*
1	Adenylate Kinase	0.668	0.338	0.506
2	Enolase	1.391	0.463	0.333
3	RecA	0.938	0.353	0.376
4	Ribosomal Protein L13P (LSU)	0.802	0.220	0.274
5	Ribosomal Protein S13P (SSU)	1.113	0.229	0.205
6	SecY	0.588	0.095	0.161
7	Serine Hydroxymethyl Transferase	1.030	0.244	0.237
8	Signal Recognition Particle	0.784	0.156	0.199
9	Histidyl-tRNA Synthetase	0.604	0.222	0.367
10	Arginly-tRNA Synthetase	0.523	0.245	0.468
11	Glutamyl-tRNA Synthetase	0.559	0.086	0.153
12	DNA Topoisomerase I	0.592	0.169	0.285
13	Leucyl-tRNA Synthetase	0.622	0.123	0.198
14	Methionyl-tRNA Synthetase	0.546	0.239	0.439
15	Tyrosyl-tRNA Synthetase	0.560	0.191	0.342
16	Valyl-tRNA Synthetase	0.676	0.135	0.200

APPENDIX B

ANALYSIS OF 16S rRNA WITH THE RED CONTROL

The RED method was standardized using a control of 39 organisms as mentioned in methods of Chapter 2.Within the 39 organisms, 31 bacterial and seven archaeal species were represented. Further, they represented 10 bacterial and two archaeal phyla. However, due to the lack of available eukaryotic genomes with reliable ORF assignments at the time that these studies were performed, only one eukaryote was used, *Saccharomyces*. Figure B1 shows a 16S rRNA gene tree for these organisms and their proposed evolutionary relationships.

To determine if the control genes shared the same evolution history as the 16S rRNA genes, the E_d's for the 16S rRNAs of all 39 organisms were calculated and plotted against the control. The resulting plot was not linear (Figure B2), suggesting either that these genes did not share the same evolutionary history as the 16S rRNA or that the rate of evolution within the 16S rRNA genes was not constant. Although the intra-domain comparisons and the comparisons between the bacteria and eukaryotes were in a linear relationship, the other two inter-domain comparisons were not linear. Other common algorithms used to calculate E_d, such as Maximum Likelihood (ML) and Kimura 2, gave similar results. Also, changing the number of nucleotide residues of the 16S rRNA sequences from 850 bp to 1350 bp when calculating evolutionary distances did not affect

the outcome, and further confirmed this non-linear relationship. All of the archaea genes used in these comparisons were thermophilic, and it is possible that this property may have affected the rate of evolution in this lineage. A lower slope might be expected if rRNAs from thermophiles evolved under more structural constraints than the rRNAs from mesophiles. In support of this hypothesis, although the relationship within the archaea was linear, the inter-domain comparisons of the archaea with the thermophilic bacteria (*Aquifex aeolicus* and *Thermotoga maritima*) slope (0.162) was lower than found within the mesophilic bacteria (0.212). In Figure B2, this difference in slopes is illustrated as a dotted oval and distinguishes the comparisons of the archaea with the theromphilic bacteria from comparisons of archaea with the mesophilic bacteria. In addition, as illustrated in Figure B2 with a solid circle, the inter-domain comparisons of archaea with eukaryote has a higher slope (0.441) than both intra-bacterial (0.295) and intra-archaeal (0.180) best-fit lines. One could assume that this discrepancy about the best-fit lines is due to possible rapid evolution. Also, the only organism representing the eukaryote domain, *Sacchromyces cerevisiae*, appeared to have a very high slope when compared with the archaea. Other yeast and eukaryotic sequences were retrieved and checked to see if this discrepancy was due to a miscalculation of a mis-annotated or fragmented yeast sequence. The results showed that other eukaryotic comparison with the archaea appeared normal, and on the best-fit line as shown by the solid line in Figure B2 (data not shown).

Figure B1: Phylogenetic relationship of all 39 organisms, using 16S rRNA as the evolutionary marker. Distances between nucleotide sequences were computed using the Jukes-Cantor formula of the PHYLIP package. These distances were then used in distance matrix program Fitch-Margoliash. The scale bar corresponded to 10 substitutions per 100 positions.

Figure B2: Evolutionary distance (E_d) of 16S rRNA of all 39 control organisms compared to the control. Figure key: intra-bacterial best fit line (solid); intra-archaeal best fit line (dotted); inter-AE is the comparisons between the archaea and the eukaryote (solid circle); inter-BE (smaller rectangle) is the comparisons between the bacteria and the eukaryote; inter-AB (larger rectangle) is the comparisons between the archaea and the bacteria; inter-domain comparisons of the archaea with the thermophilic bacteria (dotted circle).

Figure B1, Appendix B

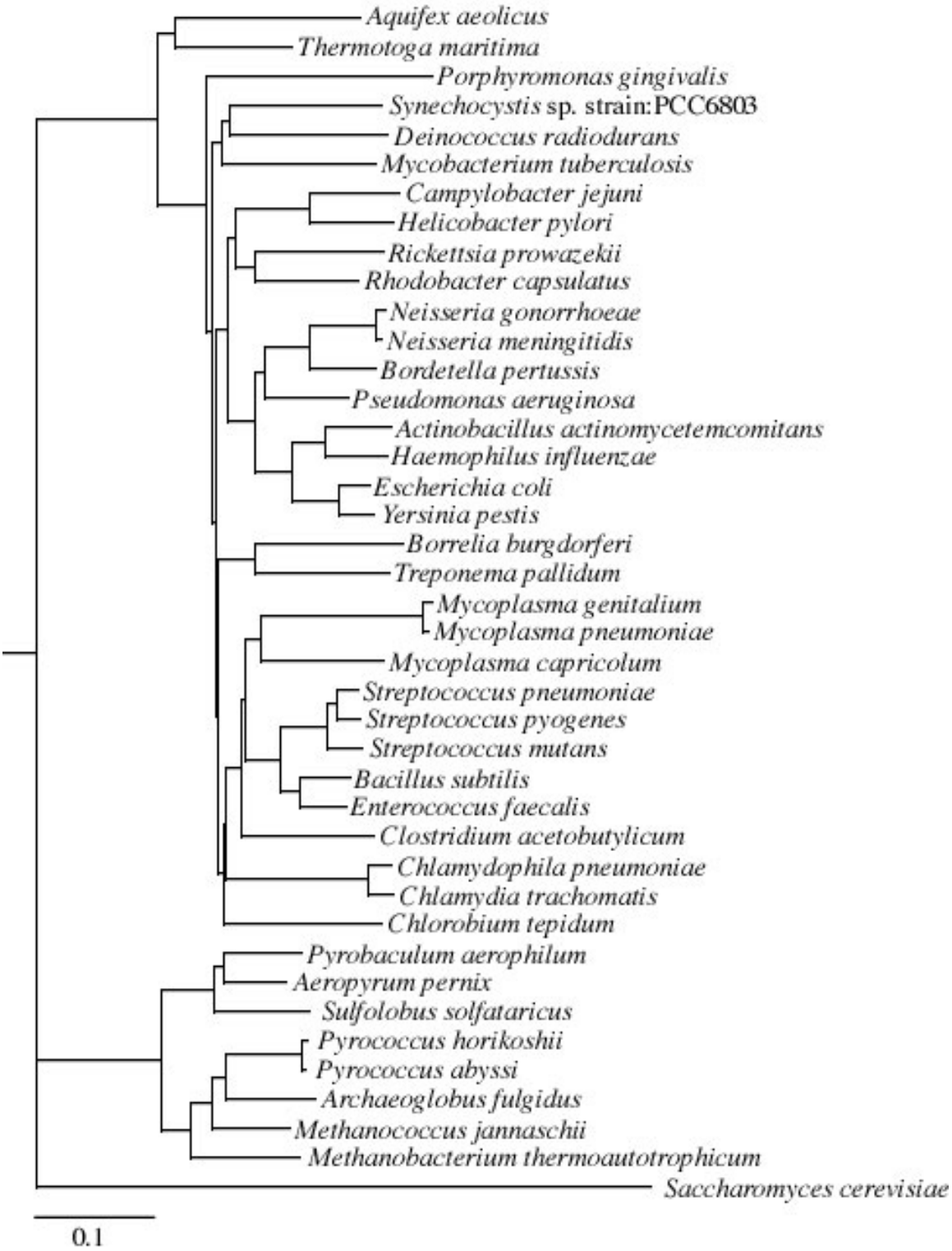
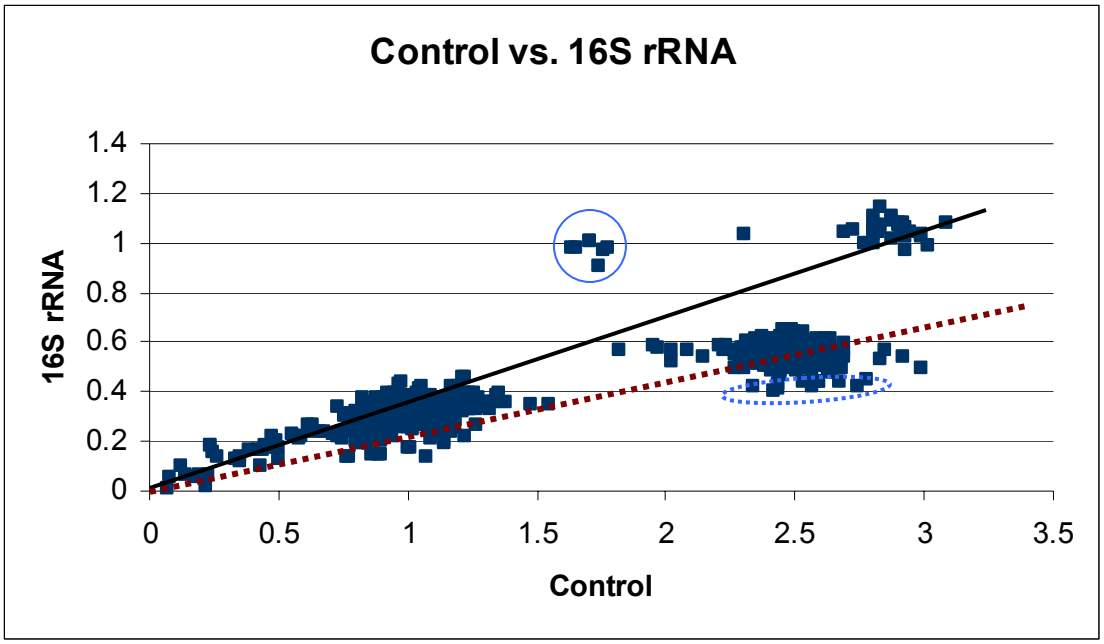


Figure B2, Appendix B



APPENDIX C

RED ANLYSES OF THE RIBOSOMAL PROTEINS

The RED method was standardized by using ribosomal genes. As mentioned in the "standardization" section of Chapter 2, within the bacteria, 39 genes for the small and large ribosomal proteins were examined, and all but two yielded linear RED plots. These two genes, L31 and L33, and other ribosomal gene analysis using the RED plots are described here. Here, the RED analyses of 17 of 18 small ribosomal subunit (SSU) proteins – S2, S3, S4, S5, S6, S7, S8, S9, S10, S11, S12, S13, S14, S15, S16, S17, and S18 - and 21 of 22 large ribosomal subunit (LSU) proteins – L1, L3, L4, L5, L6, L9, L11, L13, L14, L15, L16, L17, L18, L19, L20, L21, L24, L27, L31, L33, and L36 – were described. Not included are S19 and a control protein L2P. For a detailed analysis of ribosomal protein S19P refer to Chapter 2's Results, Model I section.

SSU Ribosomal protein S2

Analysis of the small subunit ribosomal protein S2 genes yielded a nonlinear relationship between the intra-domain and inter-domain groups supported both visually and quantitatively. Figure C1 illustrated that the intra-domain (oval I) and inter-domain (oval II, III and IV) comparisons did not have linear relationships. In addition, quantitative values also supported this nonlinear relationship. Since three domains were represented, there were three groups of values for the inter-domain comparisons of each gene: the comparison between of the archaea and the eukaryote (oval II), the comparison between the archaea and the bacteria (oval III) and the comparison between bacteria and the eukaryote (oval IV). By comparing the mean of *K* values (*m*) of the intra-archaeal and intra-bacterial comparisons, a best-fit line representing the intra-domain comparisons can be estimated to establish any linear relationships with the inter-domain comparisons. The intra-archaeal comparison had a m_{AA} value of 0.671, which was less than the intra-bacterial m_{BB} value of 0.901. Adjusting for the discrepancy in n_s values of the intra-archaeal and intra-bacteria comparisons, m'_{AA} and its standard deviation (SD'_{AA}) was analyzed. The m'_{AA} was 0.900 ± 0.077 suggested that a value of 0.077 (SD'_{AA}) above or below the m'_{AA} was acceptable. The intra-archaea m_{AA} value did not fall within this range, thus the difference between intra-archaeal and intra-bacterial *m* values was significant. Further, this may suggest that the archaea may have a slower rate of evolution than the bacteria. The intra-bacterial comparisons D_{rSD} (0.009) clearly fell below the threshold for nonlinearity, and also did not suggest any significant distribution of E_d's.

Similarly, the inter-domain comparisons were questioned for linearity and distribution of the E_d's by observing the *m* values of the three inter-domain components

along with m'_{AE} and m'_{BE}. But since m_{AA} was not significantly different from m'_{AA}, it was not possible to calculate m'_{AE} and m'_{BE}. The *rSD* for the values in ovals II, III and IV comparisons were 0.097, 0.112, and 0.097, respectively, which were below the nonlinear threshold. This suggested a tight distribution of the inter-domain comparisons.

SSU Ribosomal protein S3

Analysis of the small subunit ribosomal protein S3 genes yielded a nonlinear relationship between the intra-domain and inter-domain groups supported both visually and quantitatively. Figure C2 illustrated that the intra-domain (oval I) and inter-domain (oval II, III and IV) comparisons did not have a linear relationships. In addition, quantitative values also supported this nonlinear relationship. Since three domains were represented, there were three groups of values for the inter-domain comparisons of each gene: the comparison between the archaea and the eukaryote (oval II), the comparison between the archaea and the bacteria (oval III) and the comparison between the bacteria and the eukaryote (oval IV). By comparing the mean of *K* values (*m*) of the intra-archaeal and intra-bacterial comparisons, a best-fit line representing the intra-domain comparisons can be estimated to establish any linear relationships with the inter-domain comparisons. The intra-archaeal had a m_{AA} value of 1.004, which was greater than the of the intra-bacterial m_{BB} value of 0.807. Adjusting for the discrepancy in n_s values of the intra-archaeal and intra-bacterial m'_{AA} and its standard deviation (SD'_{AA}) was analyzed. The m'_{AA} was 0.808 ± 0.079 suggested that a value of 0.079 (SD'_{AA}) above or below the m'_{AA}

was acceptable. The intra-archaeal m_{AA} value did not fall within this range, thus the difference between intra-archaeal and intra-bacteria *m* values was significant. Further, this may have suggested that the archaea may have a faster rate of evolution than bacteria. The intra-bacterial D_{rSD} (0.033) clearly fell below the threshold for nonlinearity and also suggested no significant distribution of E_d's.

Similarly, the inter-domain comparisons were questioned for linearity and distribution of the E_d's by observing the *m* values of the three inter-domain components along with m'_{AE} and m'_{BE}. But since m_{AA} was not significantly different from m'_{AA}, it was not possible to calculate m'_{AE} and m'_{BE}. The *rSD* for the values in ovals II, III and IV comparisons were 0.140, 0.131, and 0.098, respectively. For the latter two, the value fell below the nonlinear threshold and suggested a tight distribution. For the inter-domain comparison between the archaea and the eukaryote, the values were slightly above the nonlinearity threshold. This suggested a slight scatter that can be attributed to the low number of comparisons.

SSU Ribosomal protein S4

Analysis of the small subunit ribosomal protein S4 genes yielded a linear relationship between the intra-domain and inter-domain groups supported both visually and quantitatively. Figure C3 illustrated that the intra-domain (oval I) and inter-domain (oval II) comparisons had linear relationships. In addition, quantitative values (See Chapter 2: Tables S1-2) also supported this linear relationship. Since two domains were represented, there was only one group of values for the inter-domain comparisons of each

gene: the comparison between the archaea and the bacteria (oval II). By comparing the mean of *K* values (*m*) of the intra-archaeal and intra-bacterial comparisons, a best-fit line representing the intra-domain comparisons can be estimated to establish any linear relationships with the inter-domain comparisons. The intra-archaeal m_{AA} value of 0.813, was less than the intra-bacterial m_{BB} of 0.958. Adjusting for the discrepancy in n_s values of the intra-archaeal and intra-bacterial comparisons, *m*' and its standard deviation (*SD*') was analyzed. The m'_{AA} was 0.952 ± 0.083 suggested that a value of 0.083 (SD'_{AA}) above or below the m'_{AA} was acceptable. The intra-archaeal m_{AA} value fell within this range, thus the difference between intra-archaeal and intra-bacterial *m* values was not significant, and an acceptable deviation for a linear relationship. In addition to the linear agreement indicated by the *m* values, the intra-bacterial D_{rSD} (0.016) fell below the threshold for nonlinearity. These quantitative analyses clearly supported both of the intra-domain comparison groups as linear to each other, which indicated that the archaea and bacteria share the same evolutionary rate.

Similarly, the inter-domain comparisons were questioned for linearity and distribution of the E_d's by observing the *rSD* value of inter-domain comparisons between the archaea and the bacteria was 0.107 which fell below the nonlinearity threshold. Further, the collective mean of *K* values for comparisons of inter-domain comparisons agreed with the best-fit line of the intra-domain comparisons. These observations suggested a linear relationship, and overall did not indicate LGT for the S4P gene.

SSU Ribosomal protein S5

Analysis of the small ribosomal subunit protein S5 genes yielded an apparent nonlinear relationship between the intra-domain and inter-domain groups supported both visually and quantitatively. Figure C4 illustrated that the intra-domain (oval I) and inter-domain (oval II) comparisons did not have a linear relationship. In addition, quantitative values (See Chapter 2: Tables S1-2) also supported this nonlinearity. Since two domains were represented, there was only one group of values for the inter-domain comparisons of each gene: the comparison between the archaea and the bacteria (oval II). By comparing the mean of *K* values (*m*) of the intra-archaeal and intra-bacterial comparisons, a best-fit line representing the intra-domain comparisons can be estimated to establish any linear relationships with the inter-domain comparisons. The intra-archaeal had a m_{AA} value of 0.707, which was lower than the intra-bacterial m_{BB} of 0.846. Adjusting for the discrepancy in n_s values of the intra-archaeal and intra-bacterial comparisons, m'_{AA} and its standard deviation (SD'_{AA}) was analyzed. The m'_{AA} was 0.849 ± 0.072, suggested that a value of 0.072 (SD'_{AA}) above or below the m'_{AA} was acceptable. The intra-archaeal m_{AA} value fell within this range, thus the difference between intra-archaeal and intra-bacterial *m* values was not significant. In addition to the linear agreement indicated by the *m* values, the intra-bacterial D_{rSD} (0.017) fell below the threshold for nonlinearity. These quantitative analyses clearly supported both of the intra-domain comparison groups as linear to each other, indicating that the archaea and bacteria share the same evolutionary rate.

Similarly, the inter-domain comparisons were questioned for linearity and distribution of the E_d's by observing the *rSD* value of inter-domain comparisons between the archaea and the bacteria was 0.173 which fell above the nonlinearity threshold. This suggested significant scatter among the inter-domain comparisons, yet there was lack of any clustering among these two domain comparisons. Thus, this alone is a weak argument since there were no other indications. Further, the collective mean *K* values for comparisons of inter-domain comparisons fell below the best-fit line of the intra-domain comparisons. This argued for a possible early LGT of this gene (S5P gene) to or from the bacteria prior to the separation of the archaeal and eukaryotic lineages.

SSU Ribosomal protein S6

For the plot analysis of this gene, see the 'SSU lone intra-domain comparisons' section below, and Figure C5.

SSU Ribosomal protein S7

Analysis of the small subunit ribosomal protein S7 genes yielded a linear relationship between the intra-domain and inter-domain groups supported both visually and quantitatively. Figure C6 illustrated that the intra-domain (oval I) and inter-domain (oval II, III and IV) comparisons had a linear relationships. In addition, quantitative

values also supported this linear relationship. Since three domains were represented, there were three groups of values for the inter-domain comparisons of each gene: the comparison between the archaea and the eukaryote (oval II), the comparison between the archaea and the bacteria (oval III) and the comparison between the bacteria and the eukaryote (oval IV). By comparing the mean of *K* values (*m*) of the intra-archaeal and intra-bacterial comparisons, a best-fit line representing the intra-domain comparisons can be estimated to establish any linear relationships with the inter-domain comparisons. The intra-archaeal had a m_{AA} value of 0.720, which was greater than the of the intra-bacterial m_{BB} value of 0.685. Adjusting for the discrepancy in n_s values of the intra-archaeal and intra-bacterial m'_{AA} and its standard deviation (SD'_{AA}) was analyzed. The m'_{AA} was 0.685 ± 0.059 suggested that a value of 0.059 (SD'_{AA}) above or below the m'_{AA} was acceptable. The intra-archaeal m_{AA} value did fall within this range, thus the difference between intra-archaeal and intra-bacteria *m* values was not significant. Further, this may have suggested that the archaea shares the same rate of evolution as the bacteria. The intra-bacterial D_{rSD} (0.003) clearly fell below the threshold for nonlinearity and also suggested no significant distribution of E_d's.

Similarly, the inter-domain comparisons were questioned for linearity and distribution of the E_d's by observing the *m* values of the three inter-domain components along with m'_{AE} and m'_{BE} . First, inter-domain comparisons between the archaea and the eukaryote, or oval II in Figure C6, had a m_{AE} value (0.652) that fell within the m'_{AE} value of 0.688 ± 0.059 (SD'_{AE}). Further, this inter-domain comparison appeared to be in linear agreement with the inter-domain comparisons of the archaea and the bacteria, or oval III, which had a m_{AB} value of 0.686. The last group, the inter-domain comparisons between

the bacteria and the eukaryote, or oval IV, had a m_{BE} value of 0.620 that fells below the m'_{AE} value of 0.685 ± 0.024 (SD'_{BE}), thus not linear with the inter-domain comparisons of the archaea and the bacteria. The *rSD* measurements indicated a linear relationship among the inter-domain comparisons. The *rSD* for the values in ovals II, III and IV were 0.040, 0.128, and 0.079, respectively, which fell below the nonlinearity threshold, thus there did not appear to be any scatter.

SSU Ribosomal protein S8

Analysis of the small subunit ribosomal protein S8 genes yielded a nonlinear relationship between the intra-domain and inter-domain groups supported both visually and quantitatively. Figure C7 illustrated that the intra-domain (oval I) and inter-domain (oval II, III and IV) comparisons did not have a linear relationships. In addition, quantitative values also supported this nonlinear relationship. Since three domains were represented, there were three groups of values for the inter-domain comparisons of each gene: the comparison between the archaea and the eukaryote (oval II), the comparison between the archaea and the bacteria (oval III) and the comparison between the bacteria and the eukaryote (oval IV). By comparing the mean of *K* values (*m*) of the intra-archaeal and intra-bacterial comparisons, a best-fit line representing the intra-domain comparisons can be estimated to establish any linear relationships with the inter-domain comparisons. The intra-archaeal had a m_{AA} value of 0.657, which was less than the of the intra-bacterial m_{BB} value of 0.904. Adjusting for the discrepancy in n_s values of the intra-archaeal and

intra-bacterial m'_{AA} and its standard deviation (SD'_{AA}) was analyzed. The m'_{AA} was 0.921 ± 0.075 suggested that a value of 0.075 (SD'_{AA}) above or below the m'_{AA} was acceptable. The intra-archaeal m_{AA} value did not fall within this range, thus the difference between intra-archaeal and intra-bacteria *m* values was significant. Further, this may have suggested that the archaea may have a slower rate of evolution than bacteria. The intra-bacterial D_{rSD} (0.020) fell below the threshold for nonlinearity and also suggested no significant distribution of E_d's.

Similarly, the inter-domain comparisons were questioned for linearity and distribution of the E_d's by observing the *m* values of the three inter-domain components along with m'_{AE} and m'_{BE}. But since m_{AA} was not significantly different from m'_{AA}, it was not possible to calculate m'_{AE} and m'_{BE}. The *rSD* for the values in ovals II, III and IV comparisons were 0.090, 0.113, and 0.100, respectively. All fell below the nonlinear threshold and suggested a tight distribution.

In summary, the collective mean of *K* values for comparisons of inter-domain comparisons was in agreement with a best-fit line of the intra-domain comparisons, measured by averaging of the two different intra-domain comparison rates. Thus, the S8P gene did not appear to have been involved in an event other than vertical evolution.

SSU Ribosomal protein S9

Analysis of the small subunit ribosomal protein S9 genes yielded a nonlinear relationship between the intra-domain and inter-domain groups supported both visually

and quantitatively. Figure C8 illustrated that the intra-domain (oval I) and inter-domain (oval II, III and IV) comparisons did not have a linear relationships. In addition, quantitative values also supported this nonlinear relationship. Since three domains were represented, there were three groups of values for the inter-domain comparisons of each gene: the comparison between the archaea and the eukaryote (oval II), the comparison between the archaea and the bacteria (oval III) and the comparison between the bacteria and the eukaryote (oval IV). By comparing the mean of *K* values (*m*) of the intra-archaeal and intra-bacterial comparisons, a best-fit line representing the intra-domain comparisons can be estimated to establish any linear relationships with the inter-domain comparisons. The intra-archaeal had a m_{AA} value of 0.619, which was less than the of the intra-bacterial m_{BB} value of 0.915. Adjusting for the discrepancy in n_s values of the intra-archaeal and intra-bacterial m'_{AA} and its standard deviation (SD'_{AA}) was analyzed. The m'_{AA} was 0.922 ± 0.067 suggested that a value of 0.067 (SD'_{AA}) above or below the m'_{AA} was acceptable. The intra-archaeal m_{AA} value did not fall within this range, thus the difference between intra-archaeal and intra-bacteria *m* values was significant. Further, this may have suggested that the archaea may have a slower rate of evolution than bacteria. The intra-bacterial D_{rSD} (-0.008) clearly fell below the threshold for nonlinearity and also suggested no significant distribution of E_d's.

Similarly, the inter-domain comparisons were questioned for linearity and distribution of the E_d's by observing the *m* values of the three inter-domain components along with m'_{AE} and m'_{BE}. But since m_{AA} was not significantly different from m'_{AA}, it was not possible to calculate m'_{AE} and m'_{BE}. The *rSD* for the values in ovals II, III and IV

comparisons were 0.080, 0.117, and 0.098, respectively. All fell below the nonlinear threshold and suggested a tight distribution.

SSU Ribosomal protein S10

Analysis of the small subunit ribosomal protein S10 genes yielded a linear relationship between the intra-domain and inter-domain groups supported both visually and quantitatively. Figure C9 illustrated that the intra-domain (oval I) and inter-domain (oval II, III and IV) comparisons had a linear relationships. In addition, quantitative values also supported this linear relationship. Since three domains were represented, there were three groups of values for the inter-domain comparisons of each gene: the comparison between the archaea and the eukaryote (oval II), the comparison between the archaea and the bacteria (oval III) and the comparison between the bacteria and the eukaryote (oval IV). By comparing the mean of *K* values (*m*) of the intra-archaeal and intra-bacterial comparisons, a best-fit line representing the intra-domain comparisons can be estimated to establish any linear relationships with the inter-domain comparisons. The intra-archaeal had a m_{AA} value of 0.568, which was greater than the of the intra-bacterial m_{BB} value of 0.592. Adjusting for the discrepancy in n_s values of the intra-archaeal and intra-bacterial m'_{AA} and its standard deviation (SD'_{AA}) was analyzed. The m'_{AA} was 0.585 ± 0.058 suggested that a value of 0.058 (SD'_{AA}) above or below the m'_{AA} was acceptable. The intra-archaeal m_{AA} value did fall within this range, thus the difference between intra-archaeal and intra-bacteria *m* values was not significant. Further, this may

have suggested that the archaea shares the same rate of evolution as the bacteria. The intra-bacterial D_{rSD} (0.051) fell below the threshold for nonlinearity and also suggested no significant distribution of E_d's.

Similarly, the inter-domain comparisons were questioned for linearity and distribution of the E_d's by observing the *m* values of the three inter-domain components along with m'_{AE} and m'_{BE} . First, inter-domain comparisons between the archaea and the eukaryote, or oval II in Figure C9, had a m_{AE} value (0.617) that fell within the m'_{AE} value of 0.527 ± 0.047 (SD'_{AE}). Further, this inter-domain comparison appeared to be in linear agreement with the inter-domain comparisons of the archaea and the bacteria, or oval III, which had a m_{AB} value of 0.527. The last group, the inter-domain comparisons between the bacteria and the eukaryote, or oval IV, had a m_{BE} value of 0.509 that agreed with the m'_{AE} value of 0.527 ± 0.038 (SD'_{BE}), thus linear with the inter-domain comparisons of the archaea and the bacteria. The *rSD* measurements indicated a linear relationship among the inter-domain comparisons. The *rSD* for the values in ovals II, III and IV were 0.109, 0.106, and 0.071, respectively, which fell below the nonlinearity threshold, thus there did not appear to be any scatter.

The *m* values for this gene supported a linear relationship between the intra-domain and inter-domain comparisons. In summary, this plot represented the case where the experimental gene (S10P) had the same evolutionary history as the control genes due to a linear relationship of the plotted E_d's. Moreover, there is no evidence of LGT for the S10P gene.

SSU Ribosomal protein S11

Analysis of the small subunit ribosomal protein S11 genes yielded a nonlinear relationship between the intra-domain and inter-domain groups supported both visually and quantitatively. Figure C10 illustrated that the intra-domain (oval I) and inter-domain (oval II, III and IV) comparisons did not have linear relationships. In addition, quantitative values also supported this nonlinear relationship. Since three domains were represented, there were three groups of values for the inter-domain comparisons of each gene: the comparison between of the archaea and the eukaryote (oval II), the comparison between the archaea and the bacteria (oval III) and the comparison between bacteria and the eukaryote (oval IV). By comparing the mean of *K* values (*m*) of the intra-archaeal and intra-bacterial comparisons, a best-fit line representing the intra-domain comparisons can be estimated to establish any linear relationships with the inter-domain comparisons. The intra-archaeal comparison had a m_{AA} value of 0.370, which was less than the intra-bacterial m_{BB} value of 0.613. Adjusting for the discrepancy in n_s values of the intra-archaeal and intra-bacteria comparisons, m'_{AA} and its standard deviation (SD'_{AA}) was analyzed. The m'_{AA} was 0.617 ± 0.073 suggested that a value of 0.073 (SD'_{AA}) above or below the m'_{AA} was acceptable. The intra-archaea m_{AA} value did not fall within this range, thus the difference between intra-archaeal and intra-bacterial *m* values was significant. Further, this may suggest that the archaea may have a slower rate of evolution than the bacteria. The intra-bacterial comparisons D_{rSD} (0.057) fell below the threshold for nonlinearity, and also did not suggest any significant distribution of E_d's.

Similarly, the inter-domain comparisons were questioned for linearity and distribution of the E_d's by observing the *m* values of the three inter-domain components

along with m'_{AE} and m'_{BE}. But since m_{AA} was not significantly different from m'_{AA}, it was not possible to calculate m'_{AE} and m'_{BE}. The *rSD* for the values in ovals II, III and IV comparisons were 0.132, 0.133, and 0.114, respectively, which were below the nonlinear threshold. This suggested a tight distribution of the inter-domain comparisons.

In summary, the collective mean *K* values for comparisons of inter-domain comparisons fell below the best-fit line of the intra-domain comparisons. This argued for an early LGT this S11P gene to or from the bacterial domain prior to the separation of the archaea and the eukaryote lineages.

SSU Ribosomal proteins S12, S13, S14, S15 and S16

For the plot analysis of these genes, see the 'SSU lone intra-domain comparisons' section below. Also see Figures C11-15.

SSU Ribosomal protein S17

Analysis of the small subunit ribosomal protein S17 genes yielded a linear relationship between the intra-domain and inter-domain groups supported both visually and quantitatively. Figure C16 illustrated that the intra-domain (oval I) and inter-domain (oval II, III and IV) comparisons did not have a linear relationships. In addition, quantitative values also supported this nonlinear relationship. Since three domains were

represented, there were three groups of values for the inter-domain comparisons of each gene: the comparison between the archaea and the eukaryote (oval II), the comparison between the archaea and the bacteria (oval III) and the comparison between the bacteria and the eukaryote (oval IV). By comparing the mean of *K* values (*m*) of the intra-archaeal and intra-bacterial comparisons, a best-fit line representing the intra-domain comparisons can be estimated to establish any linear relationships with the inter-domain comparisons. The intra-archaeal had a m_{AA} value of 0.916, which was greater than the of the intra-bacterial m_{BB} value of 0.859. Adjusting for the discrepancy in n_s values of the intra-archaeal and intra-bacterial m'_{AA} and its standard deviation (SD'_{AA}) was analyzed. The m'_{AA} was 0.820 ± 0.070 suggested that a value of 0.070 (SD'_{AA}) above or below the m'_{AA} was acceptable. The intra-archaeal m_{AA} value did fall within this range, thus the difference between intra-archaeal and intra-bacteria *m* values was not significant. Further, this may have suggested that the archaea shares the same rate of evolution as the bacteria. The intra-bacterial D_{rSD} (0.009) clearly fell below the threshold for nonlinearity and also suggested no significant distribution of E_d's.

Similarly, the inter-domain comparisons were questioned for linearity and distribution of the E_d's by observing the *m* values of the three inter-domain components along with m'_{AE} and m'_{BE} . First, inter-domain comparisons between the archaea and the eukaryote, or oval II in Figure C16, had a m_{AE} value (0.788) that fell within the m'_{AE} value of 0.718 ± 0.081 (SD'_{AE}). Further, this inter-domain comparison appeared to be in linear agreement with the inter-domain comparisons of the archaea and the bacteria, or oval III, which had a m_{AB} value of 0.714. The last group, the inter-domain comparisons between the bacteria and the eukaryote, or oval IV, had a m_{BE} value of 0.616 that agreed

with the m'_{AE} value of 0.718 ± 0.091 (SD'_{BE}), thus linear with the inter-domain comparisons of the archaea and the bacteria. The *rSD* measurements indicated a linear relationship among the inter-domain comparisons. The *rSD* for the values in ovals II and IV were 0.131 and 0.088, respectively, which fell below the nonlinearity threshold, thus there did not appear to be any scatter. But, the *rSD* value of the inter-domain comparisons between the archaea and the bacteria was 0.165, which fell above the nonlinearity threshold. Although this indicated a scatter, observations of any distinct comparisons group were observed.

The *m* values for this gene supported a linear relationship between the intra-domain and inter-domain comparisons. In summary, this plot represented the case where the experimental gene (S17P) had the same evolutionary history as the control genes due to a linear relationship of the plotted E_d's. Moreover, there is no evidence of LGT for the S17P gene.

SSU Ribosomal protein S18 and other lone intra-domain comparisons of ribosomal proteins S6, S12, S13, S14, S15 and S16

Detailed plot generation and analysis of the six additional small subunit ribosomal proteins was done using the RED-T. All plots contained only the bacterial gene, hence the lone intra-domain comparisons (oval I). All Figures – Figure C17 for S18 and Figures C5, C11-15 for the others – illustrated linear relationships of these genes to the control, but with different bacterial rates of evolution, or mean of *K* values. S6, S12, S13, S14, S15, S16 and S18 had mean of *K* values of 1.520, 0.340, 0.683, 0.973, 0.870, 1.076 and

1.018, respectively. In addition, all of their respective D_{rSD} have values 0.028, -0.028, 0.034, -0.008, -0.022, 0.042 and 0.046 which fell below the nonlinearity threshold for the intra-bacterial comparisons. Overall, observations suggested LGT had not occurred for these genes among the bacteria examined.

LSU Ribosomal protein L1

Analysis of the large subunit ribosomal protein L1 genes yielded a nonlinear relationship between the intra-domain and inter-domain groups supported both visually and quantitatively. Figure C18 illustrated that the intra-domain (oval I) and inter-domain (oval II) comparisons had linear relationships. In addition, quantitative values also supported this linear relationship. Since two domains were represented, there was only one group of values for the inter-domain comparisons of each gene: the comparison between the archaea and the bacteria (oval II). By comparing the mean of *K* values (*m*) of the intra-archaeal and intra-bacterial comparisons, a best-fit line representing the intra-domain comparisons can be estimated to establish any linear relationships with the inter-domain comparisons. The intra-archaeal comparison had a m_{AA} value of 0.896, which was greater than the intra-bacterial m_{BB} value of 0.788. Adjusting for the discrepancy in n_s values of the intra-archaeal and intra-bacteria comparisons, m'_{AA} and its standard deviation (SD'_{AA}) was analyzed. The m'_{AA} was 0.787 ± 0.053 suggested that a value of 0.053 (SD'_{AA}) above or below the m'_{AA} was acceptable. The intra-archaea m_{AA} value did not fall within this range, thus the difference between intra-archaeal and intra-bacterial *m* values was significant. Further, this may suggest that the archaea may have a faster rate of evolution than the bacteria. The intra-bacterial comparisons D_{rSD} (-0.003) clearly fell below the threshold for nonlinearity, and also did not suggest any significant distribution of E_d's.

Similarly, the inter-domain comparisons were questioned for linearity and distribution of the E_d's by observing the *m* value of the sole inter-domain component. The *rSD* for the value for the inter-domain comparisons between the archaea and the bacteria,

or oval II, was 0.107, which was below the nonlinear threshold. This suggested a tight distribution of the inter-domain comparisons.

LSU Ribosomal protein L3

Analysis of the large subunit ribosomal protein L3 genes yielded a linear relationship between the intra-domain and inter-domain groups supported both visually and quantitatively. Figure C19 illustrated that the intra-domain (oval I) and inter-domain (oval II, III and IV) comparisons had a linear relationships. In addition, quantitative values also supported this linear relationship. Since three domains were represented, there were three groups of values for the inter-domain comparisons of each gene: the comparison between the archaea and the eukaryote (oval II), the comparison between the archaea and the bacteria (oval III) and the comparison between the bacteria and the eukaryote (oval IV). By comparing the mean of *K* values (*m*) of the intra-archaeal and intra-bacterial comparisons, a best-fit line representing the intra-domain comparisons can be estimated to establish any linear relationships with the inter-domain comparisons. The intra-archaeal had a m_{AA} value of 0.875, which was less than the of the intra-bacterial m_{BB} value of 0.981. Adjusting for the discrepancy in n_s values of the intra-archaeal and intra-bacterial m'_{AA} and its standard deviation (SD'_{AA}) was analyzed. The m'_{AA} was 0.986 ± 0.076 suggested that a value of 0.076 (SD'_{AA}) above or below the m'_{AA} was acceptable. The intra-archaeal m_{AA} value fell within this range, thus the difference between intra-archaeal and intra-bacteria *m* values was not significant. Further, this may have suggested that the archaea shares the same rate of evolution as the bacteria. The

intra-bacterial D_{rSD} (0.036) fell below the threshold for nonlinearity and also suggested no significant distribution of E_d's.

Similarly, the inter-domain comparisons were questioned for linearity and distribution of the E_d's by observing the *m* values of the three inter-domain components along with m'_{AE} and m'_{BE} . First, inter-domain comparisons between the archaea and the eukaryote, or oval II in Figure C19, had a m_{AE} value (0.784) that fell within the m'_{AE} value of 0.763 ± 0.091 (SD'_{AE}). Further, this inter-domain comparison appeared to be in linear agreement with the inter-domain comparisons of the archaea and the bacteria, or oval III, which had a m_{AB} value of 0.763. The last group, the inter-domain comparisons between the bacteria and the eukaryote, or oval IV, had a m_{BE} value of 0.788 that agreed with the m'_{AE} value of 0.763 ± 0.018 (SD'_{BE}), thus linear with the inter-domain comparisons of the archaea and the bacteria. The *rSD* measurements indicated a linear relationship among the inter-domain comparisons. The *rSD* for the values in ovals II, III and IV were 0.071, 0.128 and 0.119, respectively, which fell below the nonlinearity threshold, thus there did not appear to be any scatter.

The *m* values for this gene supported a linear relationship between the intra-domain and inter-domain comparisons. In summary, this plot represented the case where the experimental gene (L3P) had the same evolutionary history as the control genes due to a linear relationship of the plotted E_d's. Moreover, there is no evidence of LGT for the L3P gene.

LSU Ribosomal protein L4

For the plot analysis of this gene, see the 'LSU lone intra-domain comparisons' section below. Also see Figure C20.

LSU Ribosomal protein L5

Analysis of the large subunit ribosomal protein L5 genes yielded a nonlinear relationship between the intra-domain and inter-domain groups supported both visually and quantitatively. Figure C21 illustrated that the intra-domain (oval I) and inter-domain (oval II, III and IV) comparisons did not have a linear relationships. In addition, quantitative values also supported this nonlinear relationship. Since three domains were represented, there were three groups of values for the inter-domain comparisons of each gene: the comparison between the archaea and the eukaryote (oval II), the comparison between the archaea and the bacteria (oval III) and the comparison between the bacteria and the eukaryote (oval IV). By comparing the mean of *K* values (*m*) of the intra-archaeal and intra-bacterial comparisons, a best-fit line representing the intra-domain comparisons can be estimated to establish any linear relationships with the inter-domain comparisons. The intra-archaeal had a m_{AA} value of 0.838, which was greater than the of the intra-bacterial m_{BB} value of 0.679. Adjusting for the discrepancy in n_s values of the intra-archaeal and intra-bacterial m'_{AA} and its standard deviation (SD'_{AA}) was analyzed. The m'_{AA} was 0.675 ± 0.058 suggested that a value of 0.058 (SD'_{AA}) above or below the m'_{AA}

was acceptable. The intra-archaeal m_{AA} value did not fall within this range, thus the difference between intra-archaeal and intra-bacteria *m* values was significant. Further, this may have suggested that the archaea may have a faster rate of evolution than bacteria. The intra-bacterial D_{rSD} (0.011) fell below the threshold for nonlinearity and also suggested no significant distribution of E_d's.

Similarly, the inter-domain comparisons were questioned for linearity and distribution of the E_d's by observing the *m* values of the three inter-domain components along with m'_{AE} and m'_{BE}. But since m_{AA} was not significantly different from m'_{AA}, it was not possible to calculate m'_{AE} and m'_{BE}. The *rSD* for the values in ovals II, III and IV comparisons were 0.090, 0.088, and 0.073, respectively. All fell below the nonlinear threshold and suggested a tight distribution.

LSU Ribosomal protein L6

Analysis of the large subunit ribosomal protein L6 genes yielded a nonlinear relationship between the intra-domain and inter-domain groups supported both visually and quantitatively. Figure C22 illustrated that the intra-domain (oval I) and inter-domain (oval II, III and IV) comparisons did not have a linear relationships. In addition, quantitative values also supported this nonlinear relationship. Since three domains were represented, there were three groups of values for the inter-domain comparisons of each gene: the comparison between the archaea and the eukaryote (oval II), the comparison between the archaea and the bacteria (oval III) and the comparison between the bacteria

and the eukaryote (oval IV). By comparing the mean of *K* values (*m*) of the intra-archaeal and intra-bacterial comparisons, a best-fit line representing the intra-domain comparisons can be estimated to establish any linear relationships with the inter-domain comparisons. The intra-archaeal had a m_{AA} value of 0.812, which was less than the of the intra-bacterial m_{BB} value of 0.990. Adjusting for the discrepancy in n_s values of the intra-archaeal and intra-bacterial m'_{AA} and its standard deviation (SD'_{AA}) was analyzed. The m'_{AA} was 1.002 ± 0.057 suggested that a value of 0.057 (SD'_{AA}) above or below the m'_{AA} was acceptable. The intra-archaeal m_{AA} value did not fall within this range, thus the difference between intra-archaeal and intra-bacteria *m* values was significant. Further, this may have suggested that the archaea may have a slower rate of evolution than bacteria. The intra-bacterial D_{rSD} (-0.013) clearly fell below the threshold for nonlinearity and also suggested no significant distribution of E_d's.

Similarly, the inter-domain comparisons were questioned for linearity and distribution of the E_d's by observing the *m* values of the three inter-domain components along with m'_{AE} and m'_{BE}. But since m_{AA} was not significantly different from m'_{AA}, it was not possible to calculate m'_{AE} and m'_{BE}. The *rSD* for the values in ovals II, III and IV comparisons were 0.040, 0.111, and 0.072, respectively. All fell below the nonlinear threshold and suggested a tight distribution.

LSU Ribosomal protein L9

For the plot analysis of this gene, see the 'LSU lone intra-domain comparisons' section below. Also see Figure 23.

LSU Ribosomal protein L11

Analysis of the large subunit ribosomal protein L11 genes yielded a linear relationship between the intra-domain and inter-domain groups supported both visually and quantitatively. Figure C24 illustrated that the intra-domain (oval I) and inter-domain (oval II) comparisons had linear relationships. In addition, quantitative values also supported this linear relationship. Since two domains were represented, there was only one group of values for the inter-domain comparisons of each gene: the comparison between the archaea and the bacteria (oval II). By comparing the mean of *K* values (*m*) of the intra-archaeal and intra-bacterial comparisons, a best-fit line representing the intra-domain comparisons can be estimated to establish any linear relationships with the inter-domain comparisons. The intra-archaeal comparison had a m_{AA} value of 0.534, which was less than the intra-bacterial m_{BB} value of 0.594. Adjusting for the discrepancy in n_s values of the intra-archaeal and intra-bacteria comparisons, m'_{AA} and its standard deviation (SD'_{AA}) was analyzed. The m'_{AA} was 0.603 ± 0.051 suggested that a value of 0.051 (SD'_{AA}) above or below the m'_{AA} was acceptable. The intra-archaea m_{AA} value fell within this range, thus the difference between intra-archaeal and intra-bacterial *m* values was not significant. Further, this may suggest that the archaea and bacteria share similar rates of evolution than the bacteria. The intra-bacterial comparisons D_{rSD} (0.011) fell below the threshold for nonlinearity, and also did not suggest any significant distribution of E_d's.

Similarly, the inter-domain comparisons were questioned for linearity and distribution of the E_d's by observing the *m* value of the sole inter-domain component. The

rSD for the value for the inter-domain comparisons between the archaea and the bacteria, or oval II, was 0.094, which was below the nonlinear threshold. This suggested a tight distribution of the inter-domain comparisons.

LSU ribosomal proteins – L13P

Analysis of the large subunit ribosomal protein L13 genes yielded a linear relationship between the intra-domain and inter-domain groups supported both visually and quantitatively. Figure C25 illustrated that the intra-domain (oval I) and inter-domain (oval II) comparisons had linear relationships. In addition, quantitative values also supported this linear relationship. Since two domains were represented, there was only one group of values for the inter-domain comparisons of each gene: the comparison between the archaea and the bacteria (oval II). By comparing the mean of *K* values (*m*) of the intra-archaeal and intra-bacterial comparisons, a best-fit line representing the intra-domain comparisons can be estimated to establish any linear relationships with the inter-domain comparisons. The intra-archaeal comparison had a m_{AA} value of 1.030, which was greater than the intra-bacterial m_{BB} value of 0.899. Adjusting for the discrepancy in n_s values of the intra-archaeal and intra-bacteria comparisons, m'_{AA} and its standard deviation (SD'_{AA}) was analyzed. The m'_{AA} was 0.891 ± 0.073 suggested that a value of 0.073 (SD'_{AA}) above or below the m'_{AA} was acceptable. The intra-archaea m_{AA} value fell within this range, thus the difference between intra-archaeal and intra-bacterial *m* values was not significant. Further, this may suggest that the archaea and bacteria share similar rates of evolution than the bacteria. The intra-bacterial comparisons D_{rSD} (-0.008) clearly

fell below the threshold for nonlinearity, and also did not suggest any significant distribution of E_d's.

Similarly, the inter-domain comparisons were questioned for linearity and distribution of the E_d's by observing the *m* value of the sole inter-domain component. The *rSD* for the value for the inter-domain comparisons between the archaea and the bacteria, or oval II, was 0.114, which was below the nonlinear threshold. This suggested a tight distribution of the inter-domain comparisons.

LSU Ribosomal protein L14

Analysis of the large subunit ribosomal protein L14 genes yielded a linear relationship between the intra-domain and inter-domain groups supported both visually and quantitatively. Figure C26 illustrated that the intra-domain (oval I) and inter-domain (oval II, III and IV) comparisons had a linear relationships. In addition, quantitative values also supported this linear relationship. Since three domains were represented, there were three groups of values for the inter-domain comparisons of each gene: the comparison between the archaea and the eukaryote (oval II), the comparison between the archaea and the bacteria (oval III) and the comparison between the bacteria and the eukaryote (oval IV). By comparing the mean of *K* values (*m*) of the intra-archaeal and intra-bacterial comparisons, a best-fit line representing the intra-domain comparisons can be estimated to establish any linear relationships with the inter-domain comparisons. The intra-archaeal had a m_{AA} value of 0.564, which was greater than the of the intra-bacterial

m_{BB} value of 0.505. Adjusting for the discrepancy in n_s values of the intra-archaeal and intra-bacterial m'_{AA} and its standard deviation (SD'_{AA}) was analyzed. The m'_{AA} was 0.505 ± 0.062 suggested that a value of 0.062 (SD'_{AA}) above or below the m'_{AA} was acceptable. The intra-archaeal m_{AA} value fell within this range, thus the difference between intra-archaeal and intra-bacteria *m* values was not significant. Further, this may have suggested that the archaea shares the same rate of evolution as the bacteria. The intra-bacterial D_{rSD} (0.041) fell below the threshold for nonlinearity and also suggested no significant distribution of E_d's.

Similarly, the inter-domain comparisons were questioned for linearity and distribution of the E_d's by observing the *m* values of the three inter-domain components along with m'_{AE} and m'_{BE} . First, inter-domain comparisons between the archaea and the eukaryote, or oval II in Figure C26, had a m_{AE} value (0.511) that did not fall within the m'_{AE} value of 0.428 ± 0.027 (SD'_{AE}). Further, this inter-domain comparison appeared not to be in linear agreement with the inter-domain comparisons of the archaea and the bacteria, or oval III, which had a m_{AB} value of 0.431. The last group, the inter-domain comparisons between the bacteria and the eukaryote, or oval IV, had a m_{BE} value of 0.487 that agreed with the m'_{AE} value of 0.431 ± 0.044 (SD'_{BE}), thus linear with the inter-domain comparisons of the archaea and the bacteria. The *rSD* measurements indicated a linear relationship among only two of the inter-domain comparisons, oval III and IV. The *rSD* for the values in ovals II, III and IV were 0.064, 0.125 and 0.079, respectively, which fell below the nonlinearity threshold, thus there did not appear to be any scatter.

LSU Ribosomal proteins L15, L16, L17, L18, L19, L20, L21, L24 and L27

For the plot analysis of these genes, see the 'LSU lone intra-domain comparisons' section below. Also see Figures C27-35.

LSU Ribosomal proteins L31

Analysis of the large ribosomal subunit protein L31 genes portrayed an evolutionary history that possibly involved LGT. This plot contained only the bacterial gene, hence the lone intra-domain comparisons. With only the bacteria analyzed in this gene plot, a nonlinear relationship within the intra-domain comparisons of the bacteria was supported both visually and quantitatively. Figures C36 illustrated that the intra-bacterial comparisons was separated into two distinct plot clusters, oval Ia and Ib. Further, the quantitative values showed support for these clusters. The intra-Bacteria D_{rSD} (0.095) fell above the non-linear threshold, thus projecting a wide distribution among bacterial comparisons. Detail analysis showed that ovals Ia and Ib consisted of comparisons between two distinct bacteria. Oval Ia contained mainly genes of *Synechocystis* sp. and *Rhodobacter capsulatus* as compared with other bacteria. Oval Ib contained comparisons of all other bacteria, such as the proteobacteria. Mean of *K* analysis illustrated these intra-bacterial comparisons' distinction. Oval Ib (2.158) had higher slopes than oval Ia (1.279). Furthermore, according to our hypothesis, this observation leads to consider that one of these two bacteria could have been the recipient

of this gene from another domain. This question remains unanswered here due to lack of organisms representing these two domains.

LSU Ribosomal protein L33

Analysis of the large ribosomal subunit protein L33P genes portrayed an evolutionary history that possibly involved LGT. This plot contained only the bacterial gene, hence the lone intra-domain comparisons. With only bacteria analyzed for this gene plot, a nonlinear relationship within the intra-bacterial comparisons was supported both visually and quantitatively. Figures C37 illustrated that the intra-bacterial comparisons was separated into three distinct plot clusters, oval Ia, Ib and Ic. Further, the quantitative values showed support for these groupings. The intra-bacterial D_{rSD} (0.210) fell above the nonlinear threshold, thus projected a wide distribution among bacterial comparisons. Detail analysis showed that ovals Ia and Ib consisted of comparisons between two distinct bacteria, *Streptococcus mutans* and *Rickettsia prowazekii*. Mean of *K* analysis illustrated these distinctions. Both oval Ib (2.697) and Ic (1.810) had higher slopes than oval Ia (1.124). Furthermore, according to our hypothesis, this observation leads to consider that one of these two bacteria could have been the recipient of this gene from another domain, if not from the bacterial domain.

Beginning with *Rickettsia* , it appeared that the mean of *K* of *Rickettsia* vs. other proteobacteria (1.082) agreed with the best-fit line of oval Ia. On the other hand, the mean of *K* of oval Ic, containing comparisons of *Rickettsia* vs. the bacteria (1.871), such

as the Gram-type positives, was clearly higher than oval Ia. This suggested that either *Rickettsia* or the Gram-type positives may have been involved in LGT with either the archaea or the eukaryote domain. This question remains unanswered here due to lack of organisms representing these two domains.

Next, we analyzed oval Ib and Ic, which contained comparisons between *Streptococcus* and low GC Gram-type positives. Mean of *K* analysis showed that *Streptococcus* vs. proteobacteria had a low mean of *K* value of 1.214, similar to the mean of *K* of oval Ia, whereas *Streptococcus* vs. low GC Gram-type bacteria had a much higher mean of *K* (oval Ic=1.745) than oval Ia. This suggested that *Streptococcus* received the gene encoding L33P from the proteobacteria.

LSU Ribosomal protein L36 and other lone intra-domain comparisons of ribosomal proteins L4, L9, L15, L16, L17, L18, L19, L20, L21, L24 and L27

Detailed plot generation and analysis of the 12 additional large subunit ribosomal proteins was done using the RED-T. All plots contained only the bacterial gene, hence the lone intra-domain comparisons (oval I). All Figures – Figure C38 for L36 and Figures 20, 23, 27-35 for the others – illustrated linear relationships of these genes to the control, but with different bacterial rates of evolution, or mean of *K* values. L4, L9, L15, L16, L17, L18, L19, L20, L21, L24, L27, and L36 had mean of *K* values of 1.402, 1.450, 1.190, 0.705, 0.891, 0.940, 0.902, 0.786, 1.238, 1.029, 0.652, and 0.478, respectively. In addition, all of their respective D_{rSD} have values 0.041, -0.012, 0.032, 0.016, 0.014,

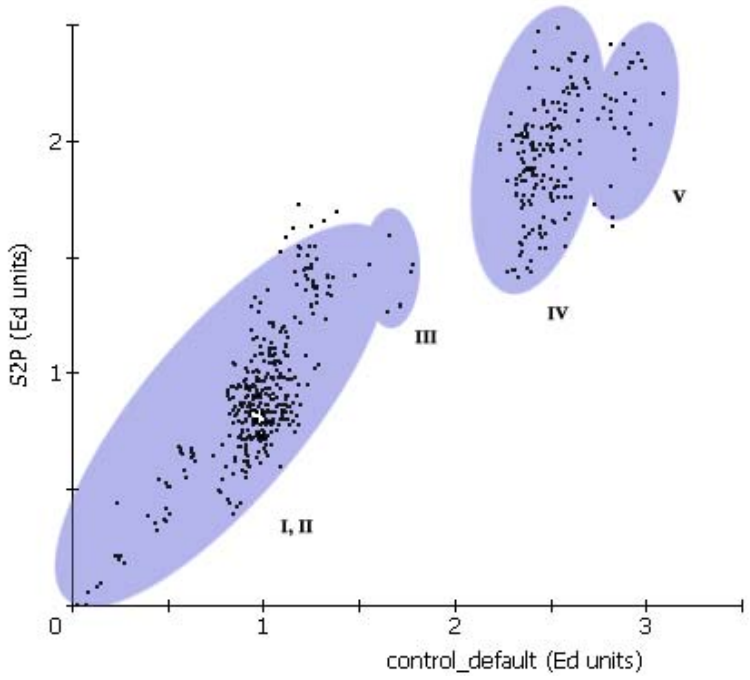
0.055, -0.008, 0.032, -0.020, -0.037, -0.024, and 0.025, which fell below the nonlinearity threshold for the intra-bacterial comparisons. Overall, observations suggested LGT had not occurred for these genes among the bacteria examined.

Figures C1-2, C6-10 and C16: RED plots of SSU ribosomal proteins in Appendix C. Intra-domain comparisons within the archaea (oval I), within the bacteria (oval II), inter-domain comparisons between the archaea and the eukaryote (oval III), between the archaea and the bacteria (oval IV), and between the bacteria and the eukaryote (oval V).

Figures C3-4: RED plots of SSU ribosomal proteins in Appendix C. Intra-domain comparisons within the archaea (oval I), within the bacteria (oval II), inter-domain comparisons between the archaea and the bacteria (oval III).

Figures C5, C11-15 and C17: RED plots of SSU ribosomal proteins in Appendix C. Intra-domain comparisons within the bacteria (oval I). No other domain analyzed for these genes.

Figure C1 (top) and C2 (bottom), Appendix C



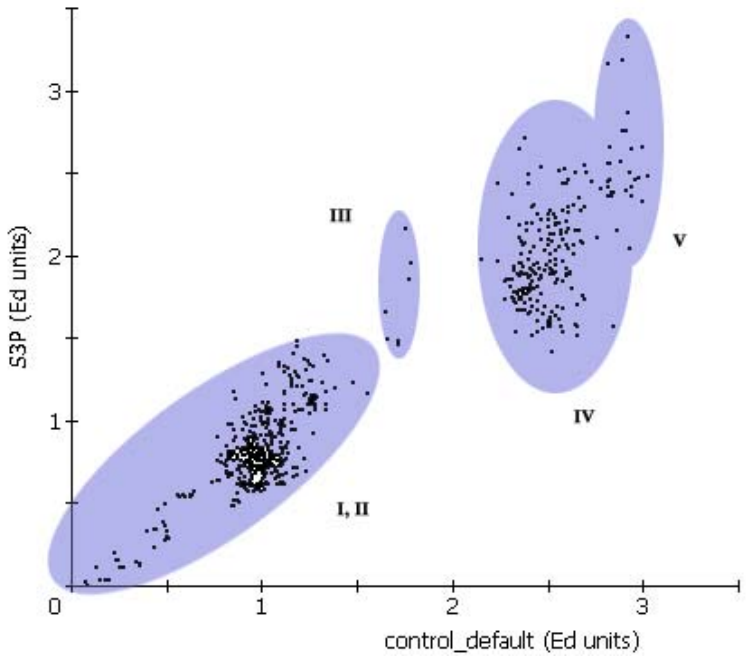
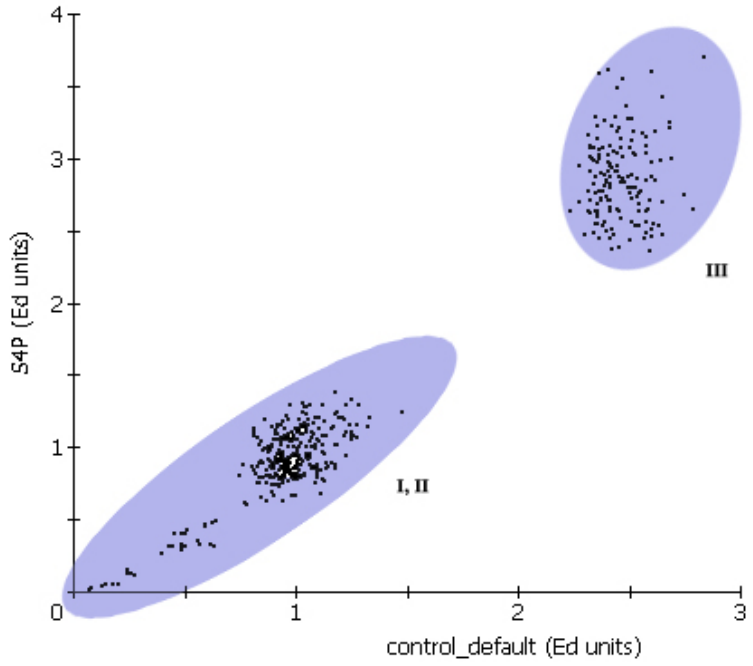


Figure C3 (top) and C4 (bottom), Appendix C



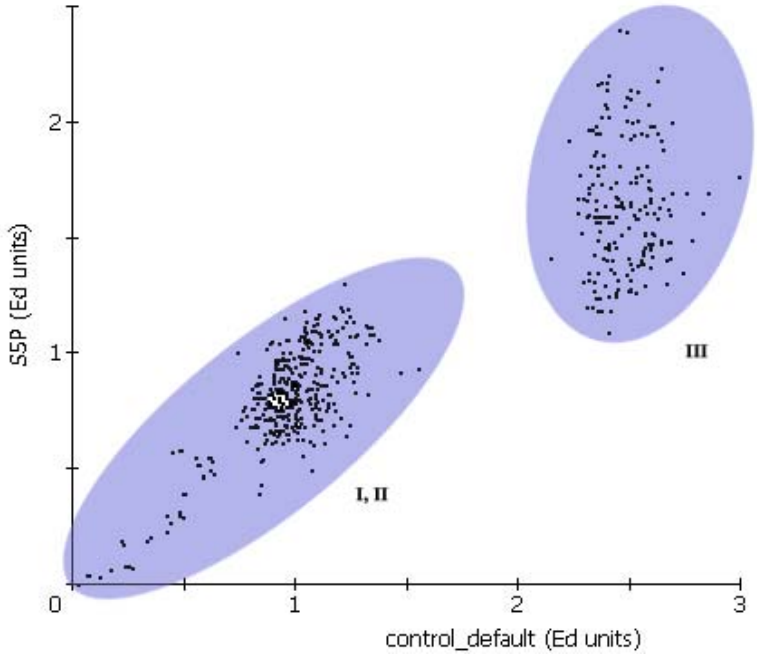
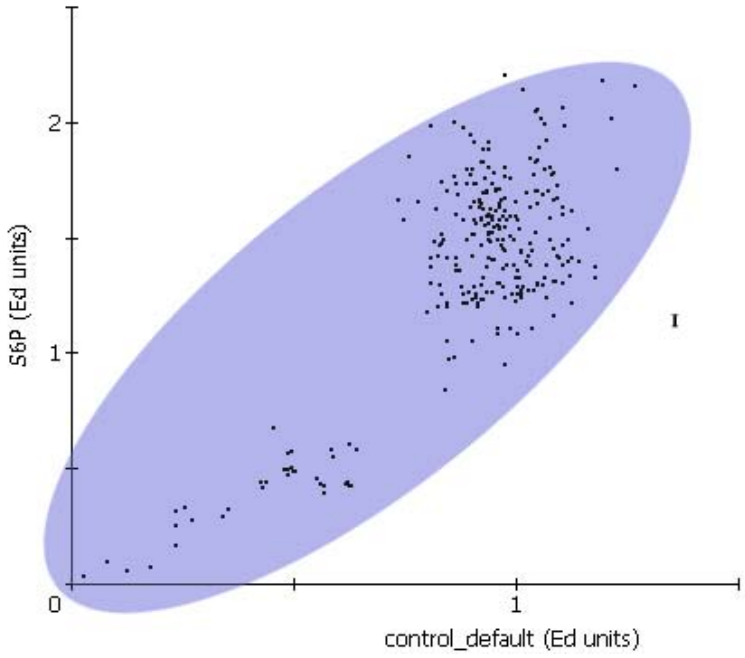


Figure C5 (top) and C6 (bottom), Appendix C



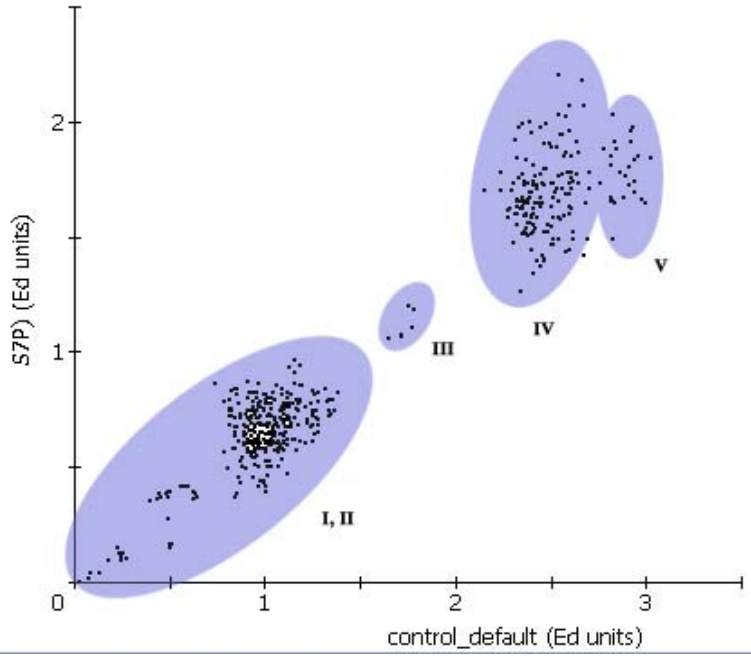
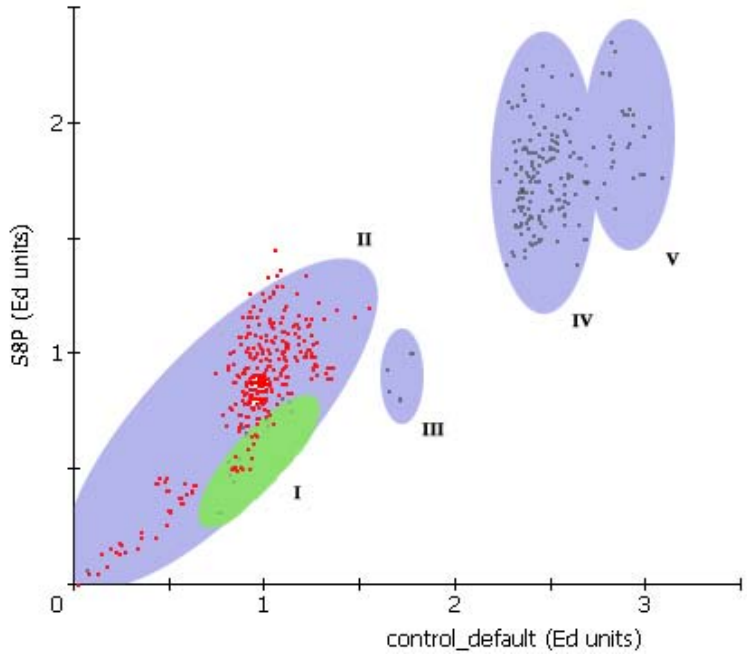


Figure C7 (top) and C8 (bottom), Appendix C



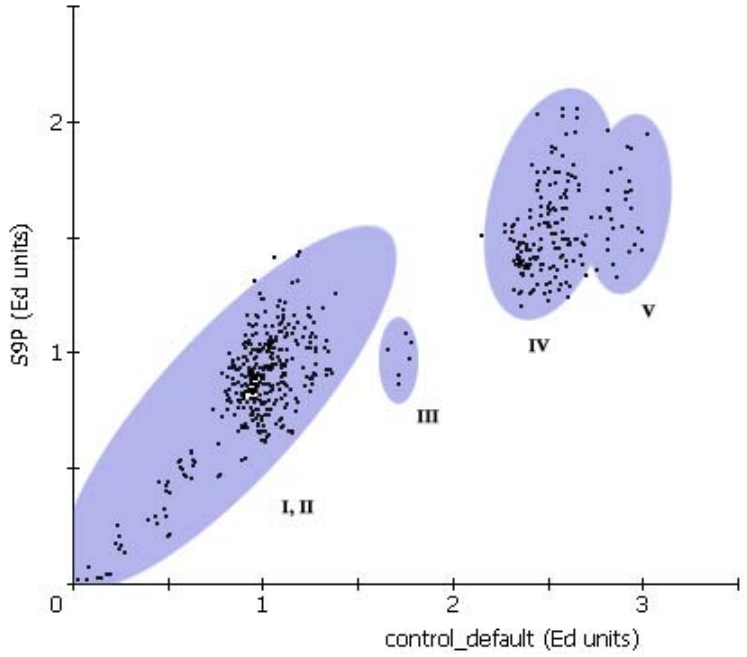
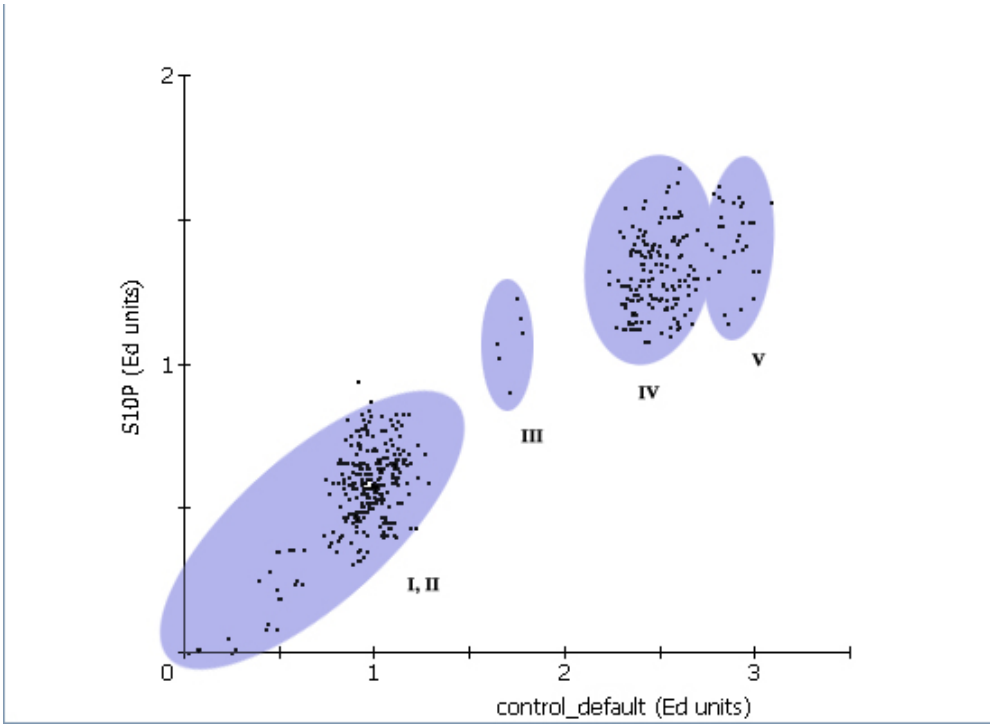


Figure C9 (top) and C10 (bottom), Appendix C



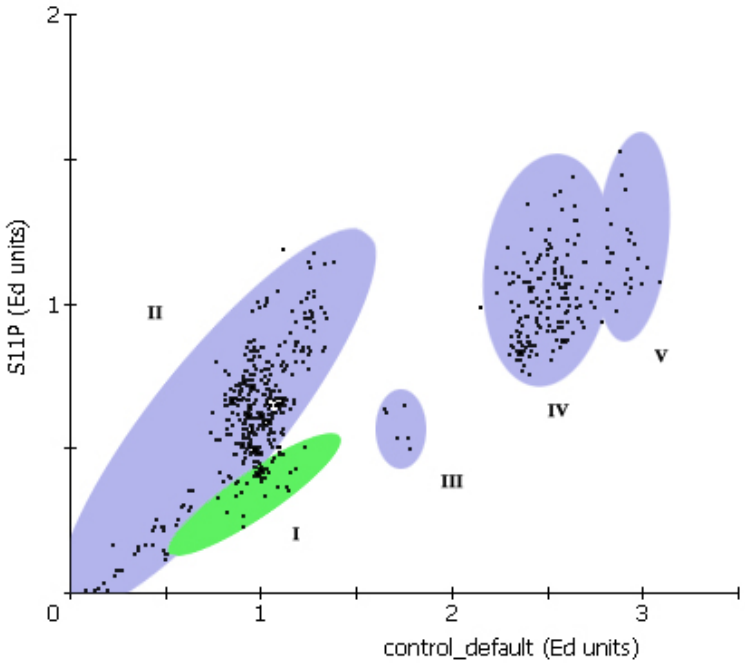
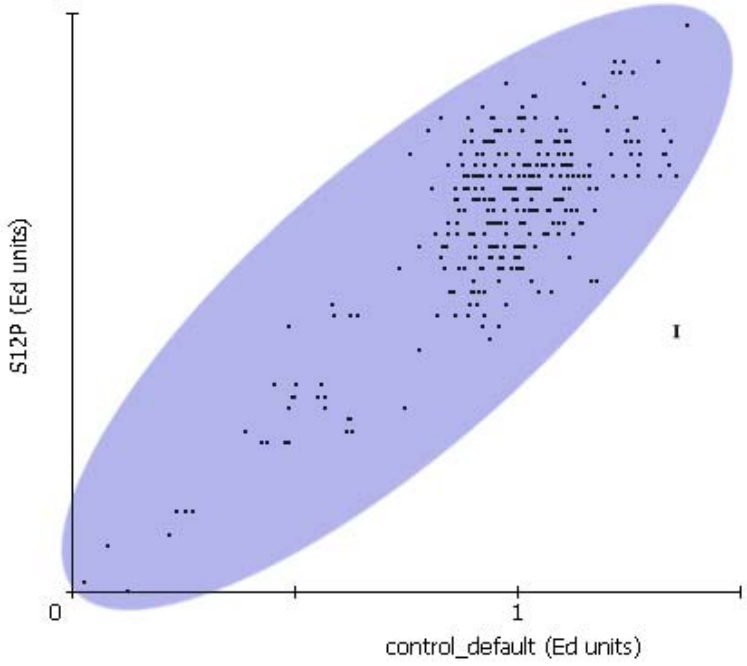


Figure C11 (top) and C12 (bottom), Appendix C



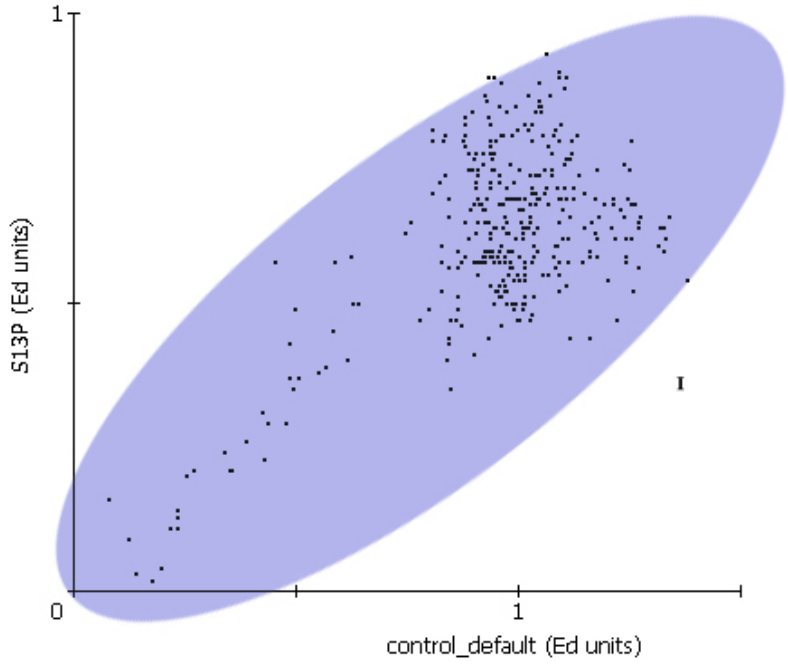
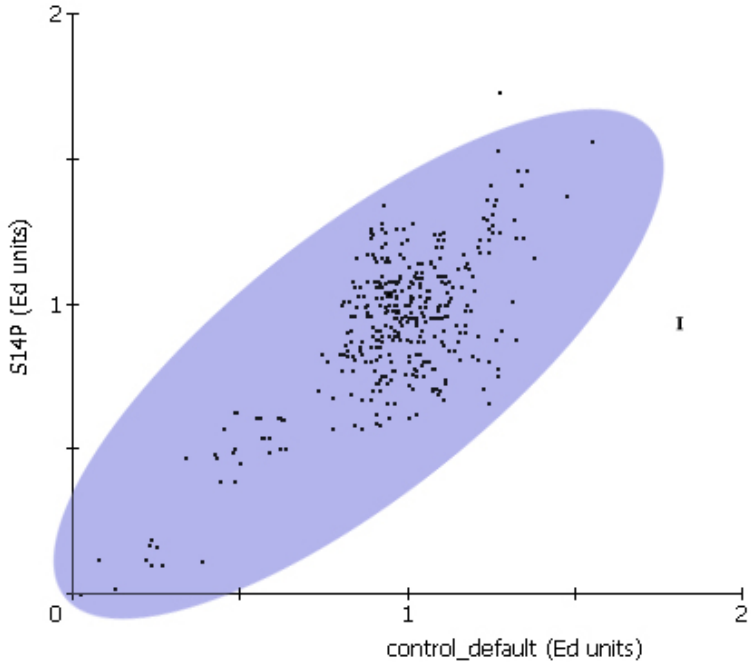


Figure C13 (top) and C14 (bottom), Appendix C



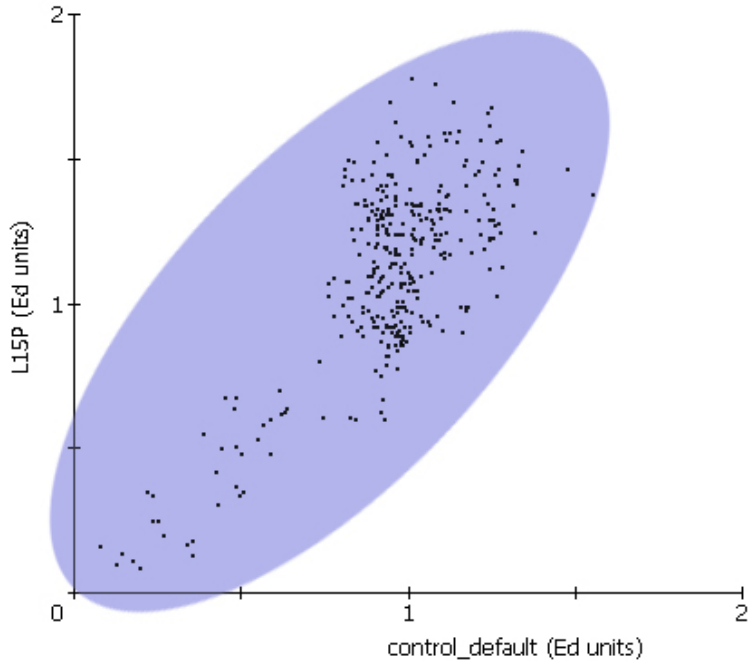
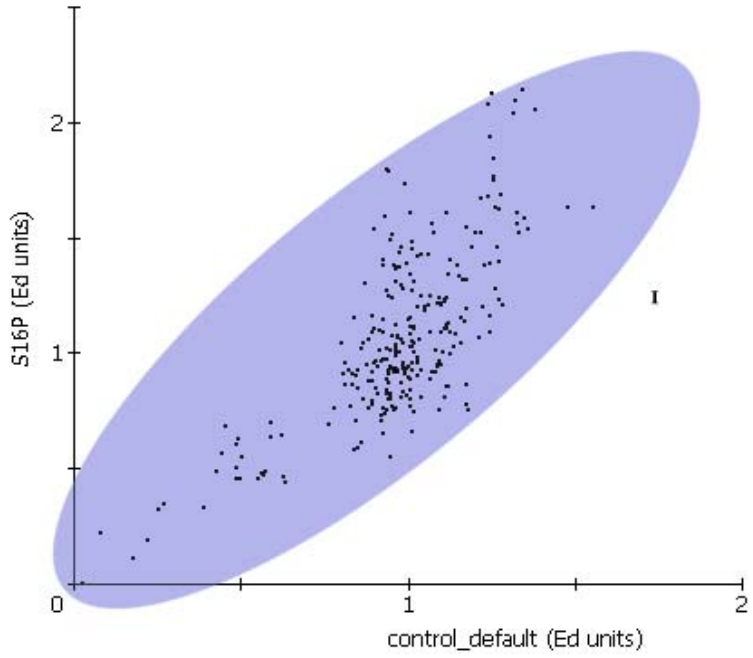


Figure C15 (top) and C16 (bottom), Appendix C



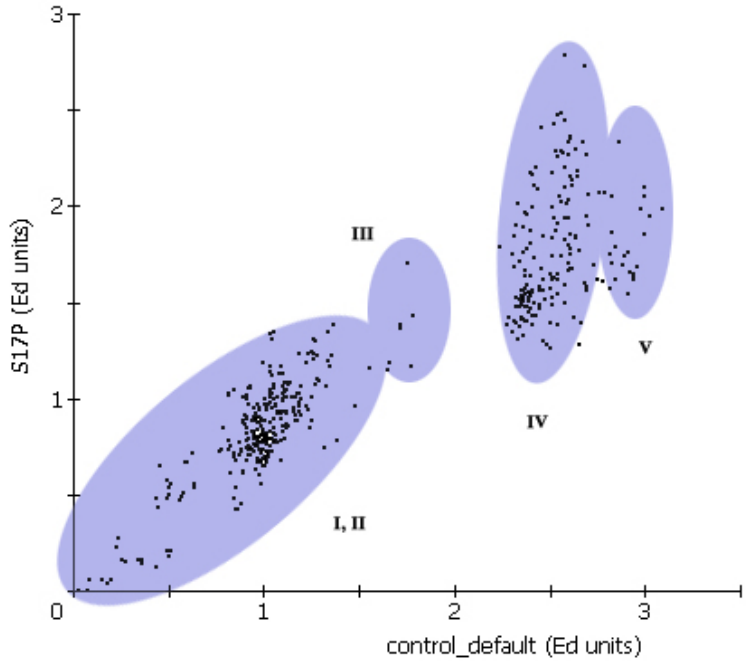
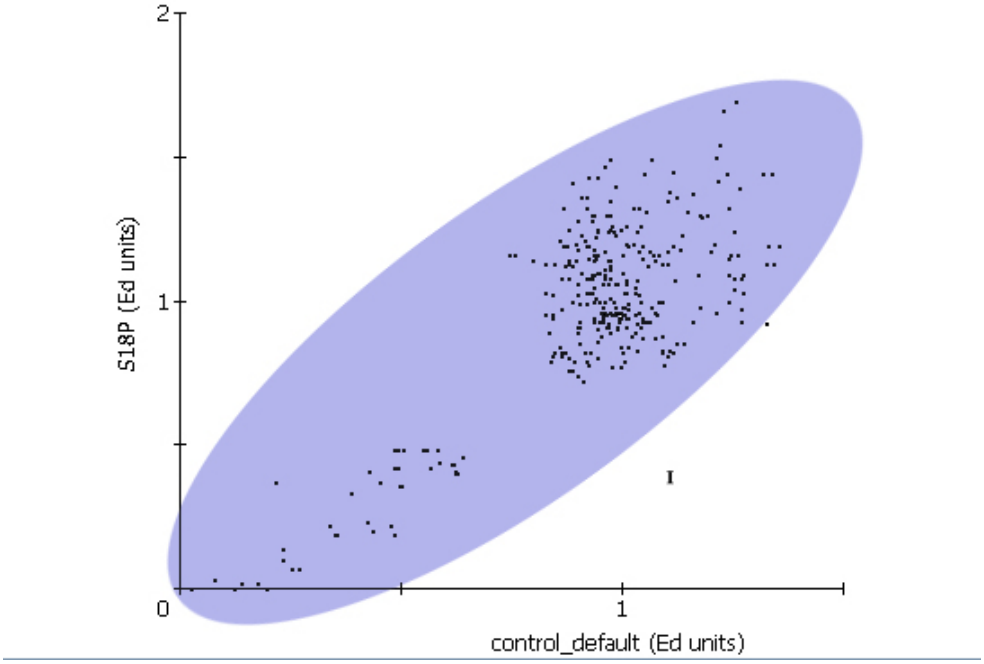


Figure C17, Appendix C



Figures C19, C21-22 and C26: RED plots of LSU ribosomal proteins in Appendix C. Intra-domain comparisons within the archaea (oval I), within the bacteria (oval II), inter-domain comparisons between the archaea and the eukaryote (oval III), between the archaea and the bacteria (oval IV), and between the bacteria and the eukaryote (oval V).

Figures C18 and C24-25: RED plots of SSU ribosomal proteins in Appendix C. Intra-domain comparisons within the archaea (oval I), within the bacteria (oval II), inter-domain comparisons between the archaea and the bacteria (oval III).

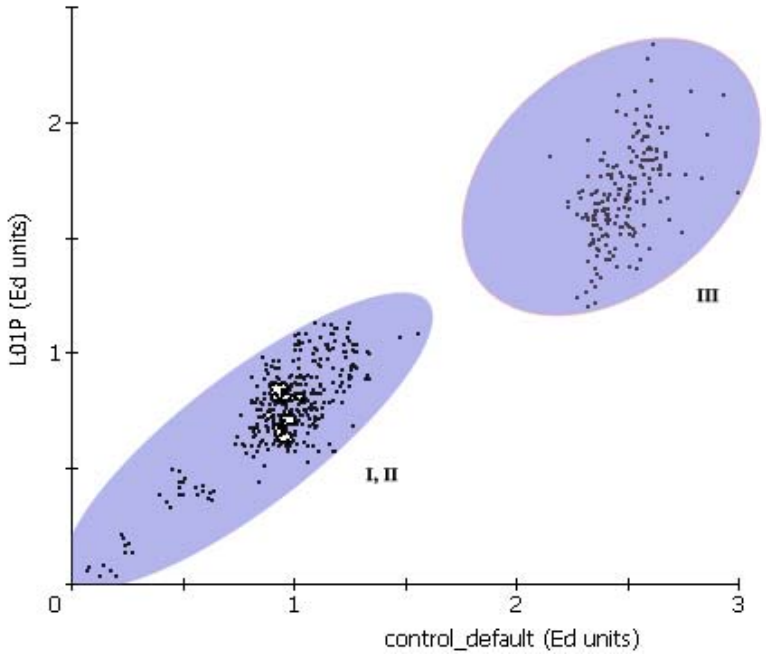
Figures C20, C23, C27-35 and C38: RED plots of LSU ribosomal proteins in Appendix C. Intra-domain comparisons within the bacteria (oval I). No other domain analyzed for these genes.

Figures C36: RED plots of LSU ribosomal proteins L31 in Appendix C. Intra-domain comparisons within the bacteria (oval Ia, Ib). Oval Ia contained mainly genes of *Synechocystis* sp. and *Rhodobacter capsulatus* as compared with other bacteria. Oval Ib contained comparisons of all other bacteria, such as the proteobacteria. No other domain analyzed for these genes.

Figures C37: RED plots of LSU ribosomal proteins L33 in Appendix C. Intra-domain comparisons within the bacteria (oval Ia, Ib, Ic). Ovals Ia and Ib contained comparisons between two distinct bacteria, *Streptococcus mutans* and *Rickettsia*

prowazekii. Ovals Ib and Ic contained comparisons between *Streptococcus* and low GC Gram-type positives. No other domain analyzed for these genes.

Figure C18 (top) and C19 (bottom), Appendix C



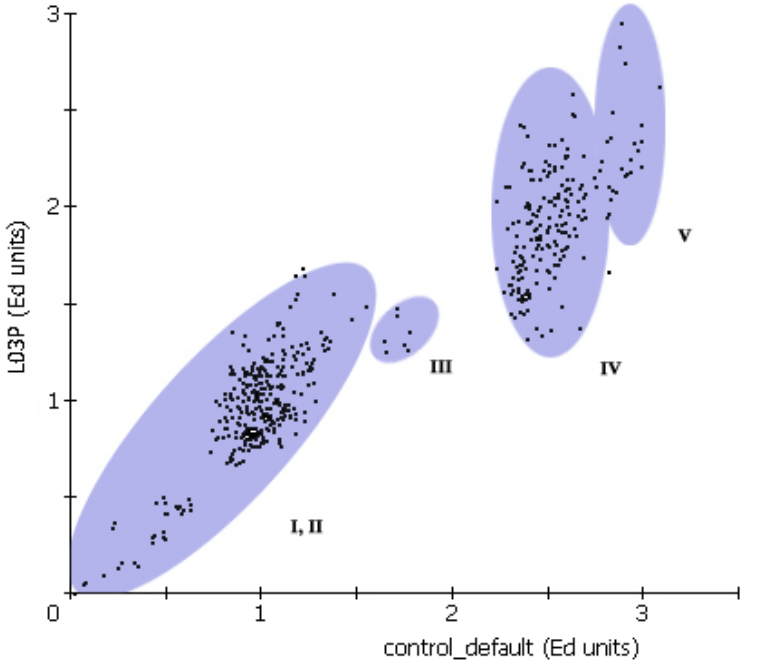
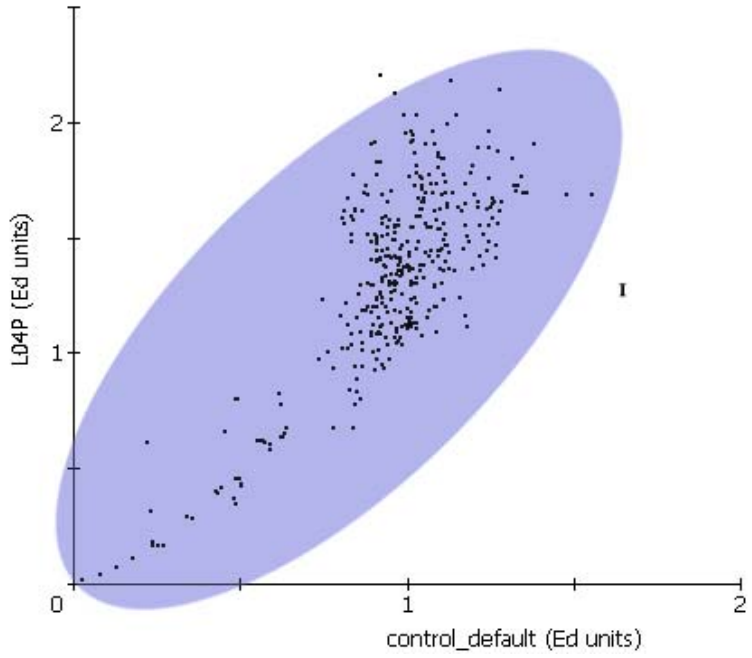


Figure C20 (top) and C21 (bottom), Appendix C



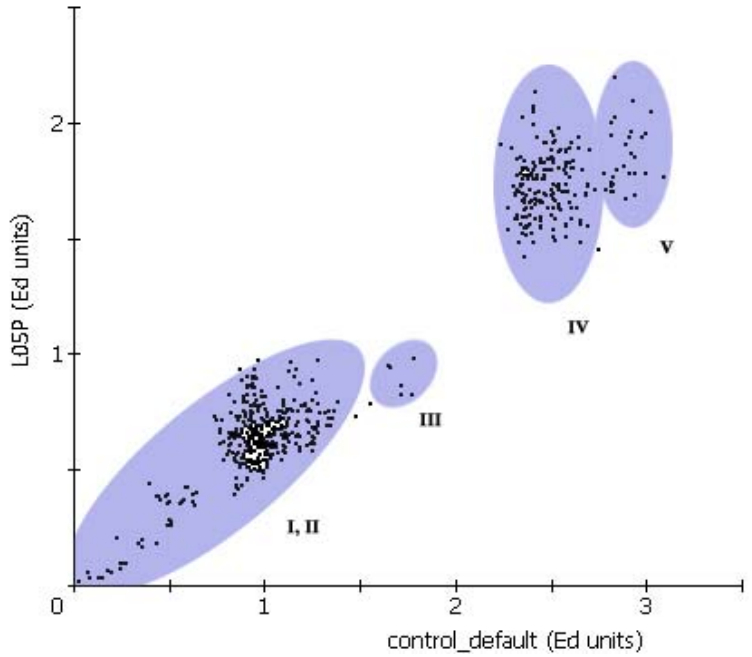
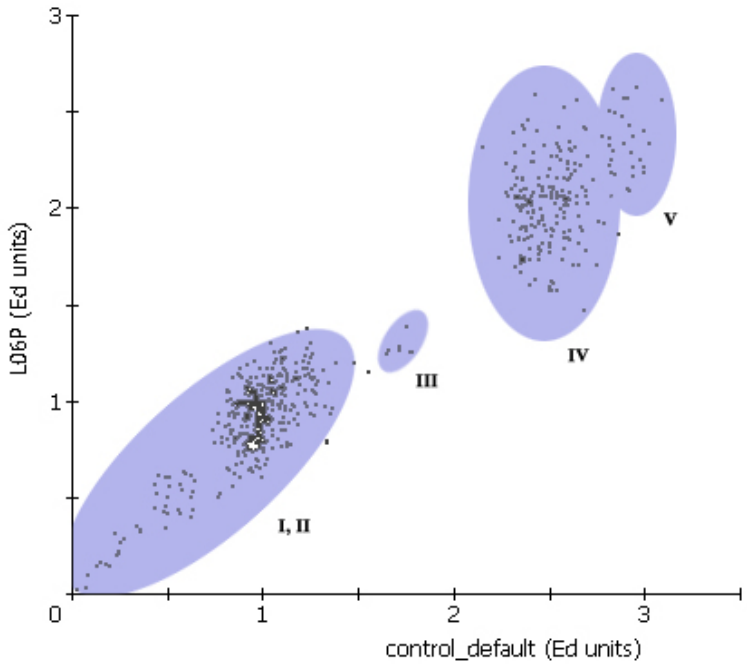


Figure C22 (top) and C23 (bottom), Appendix C



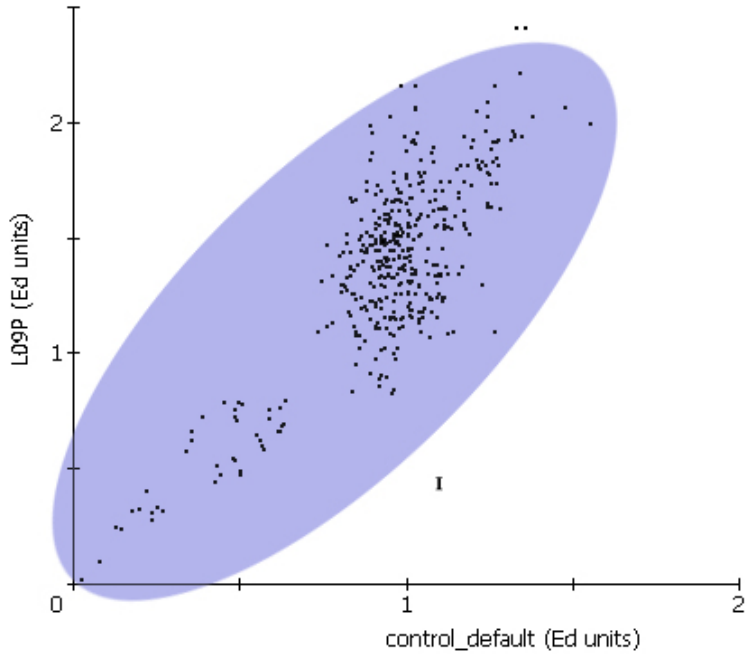
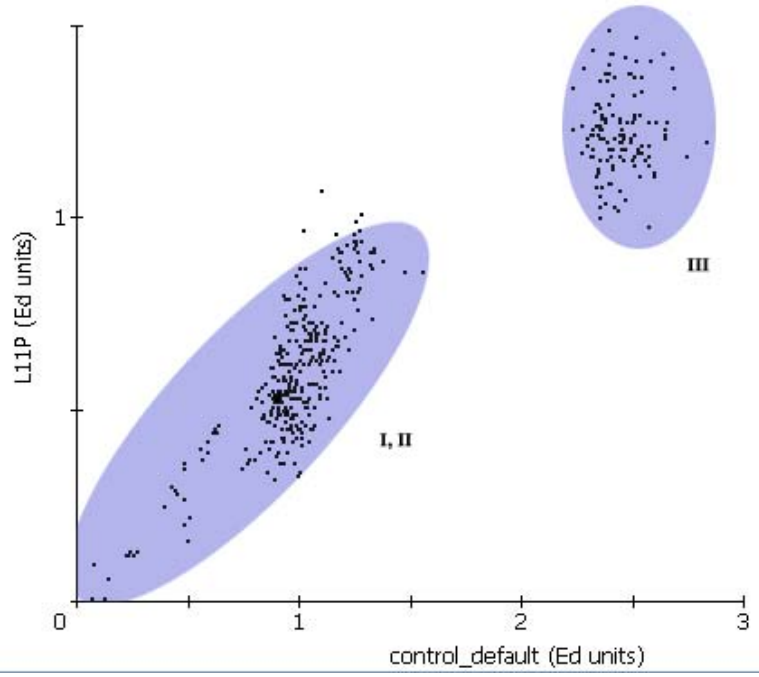


Figure C24 (top) and C25 (bottom), Appendix C



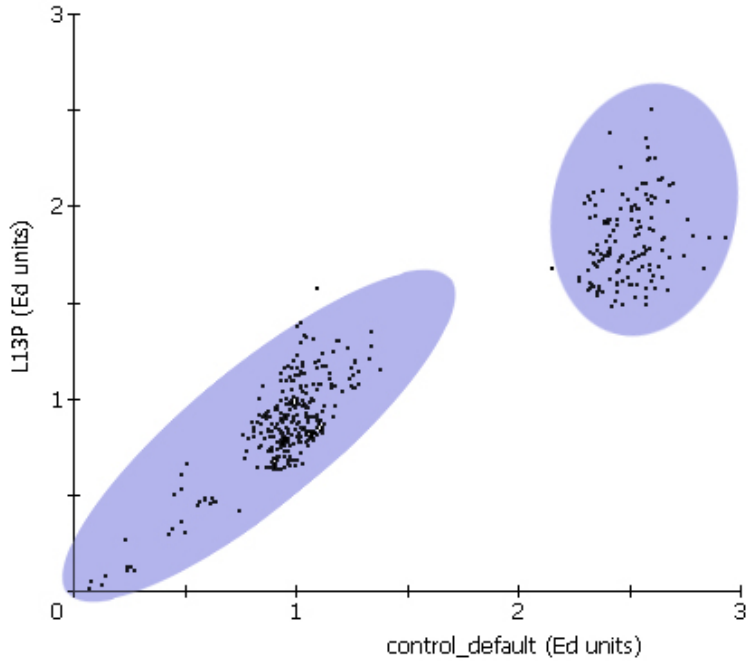
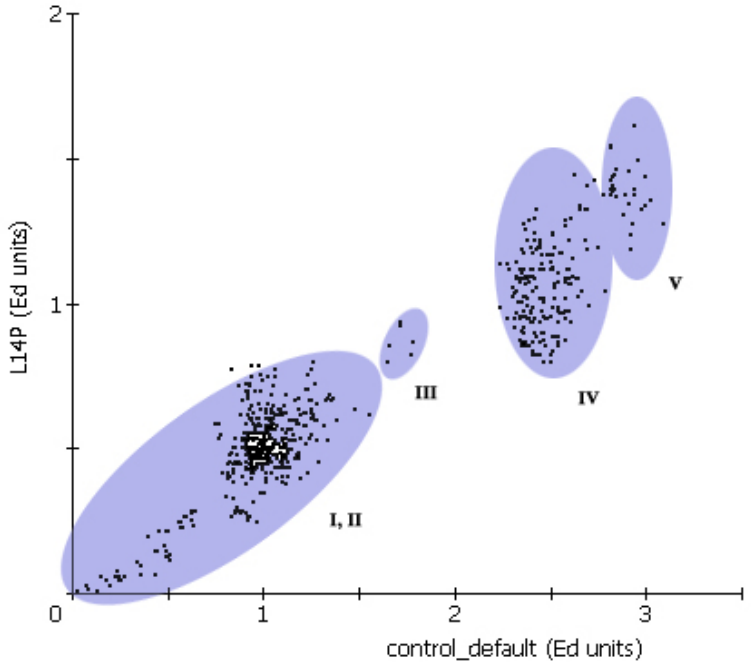


Figure C26 (top) and C27 (bottom), Appendix C



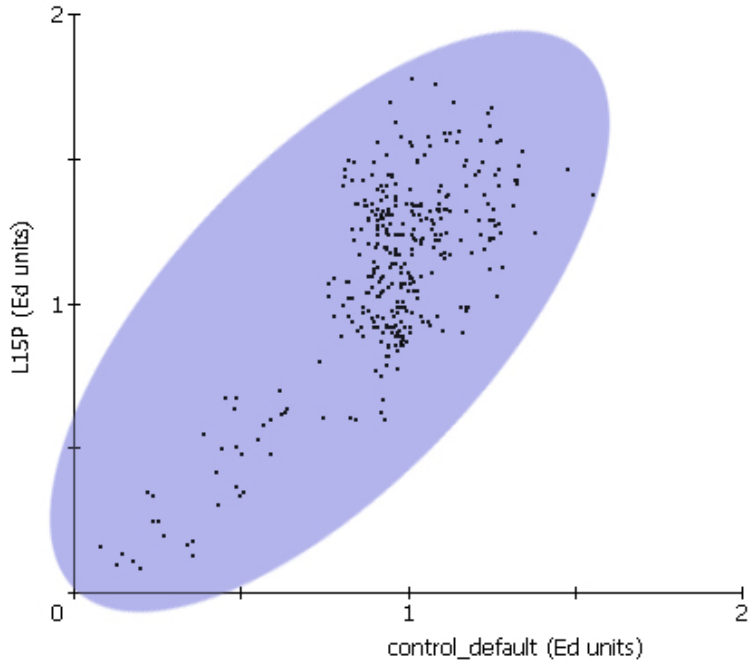
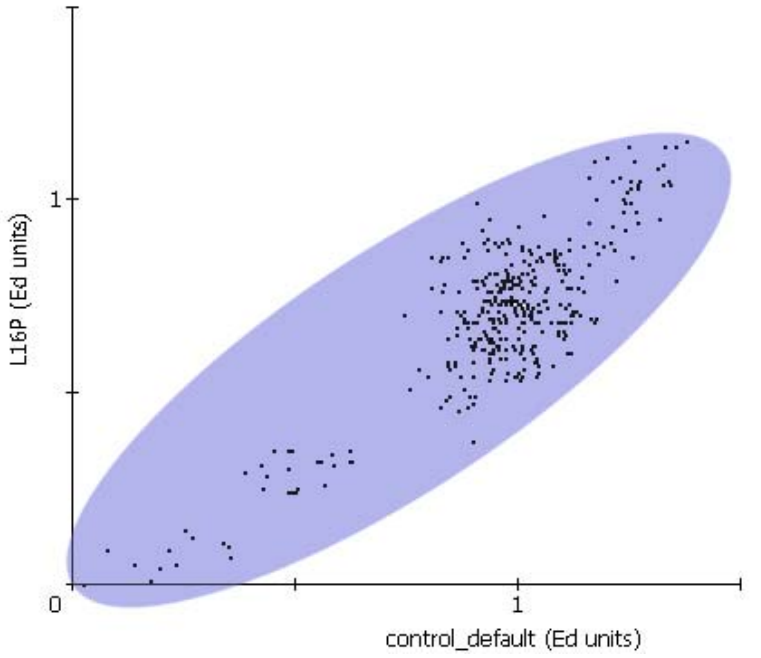


Figure C28 (top) and C29 (bottom), Appendix C



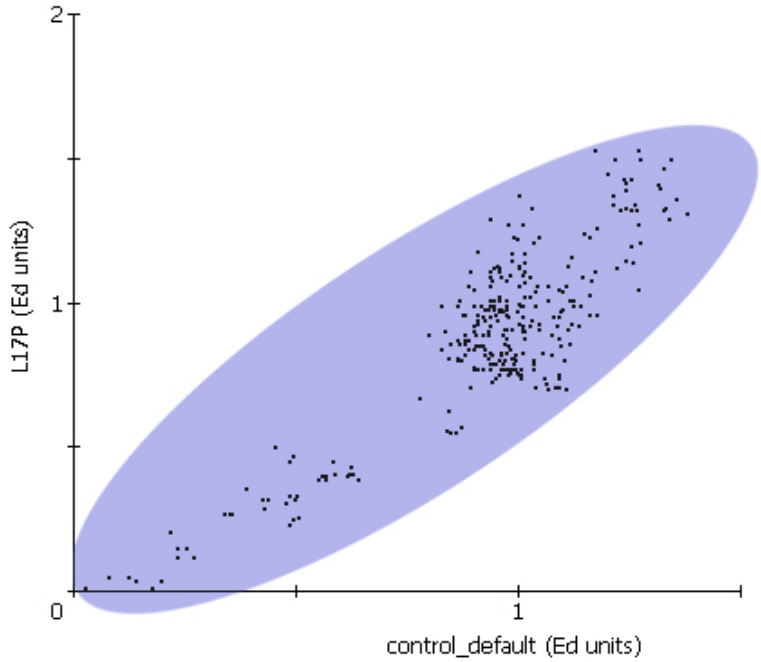
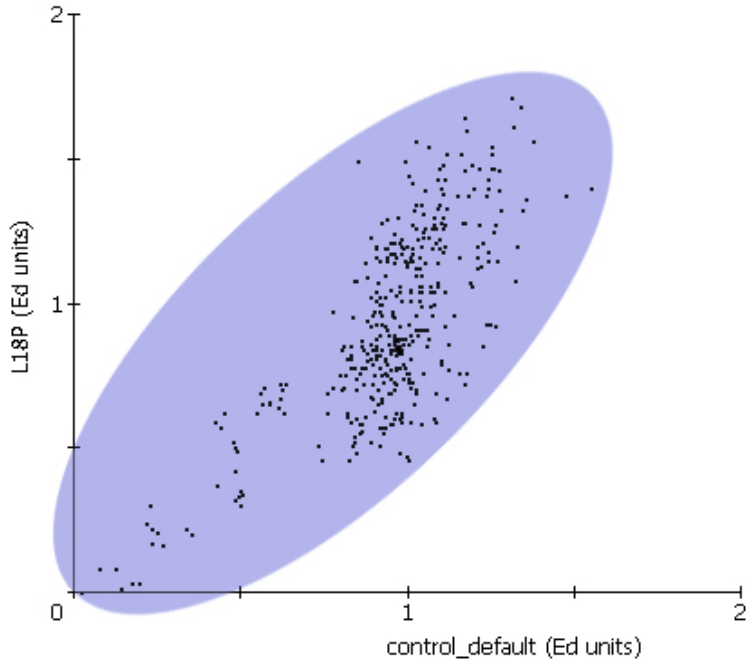


Figure C30 (top) and C31 (bottom), Appendix C



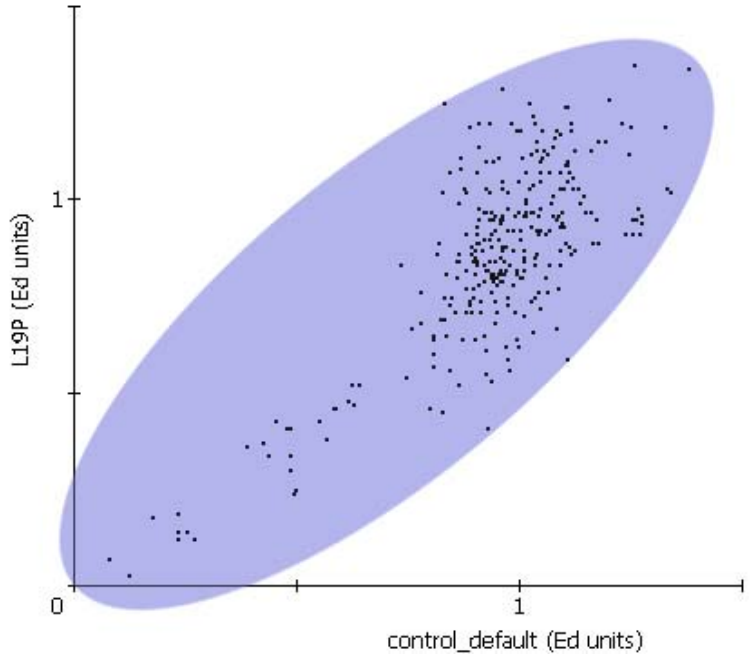
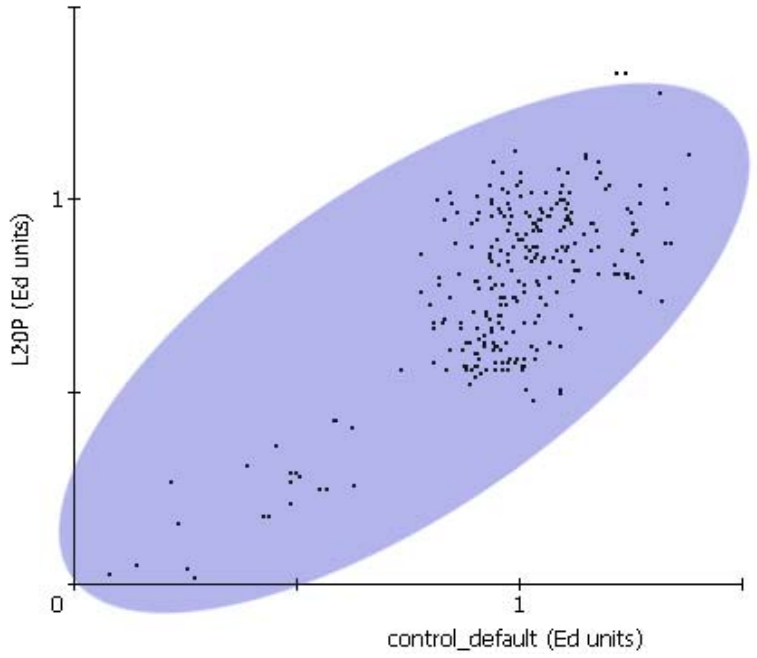


Figure C32 (top) and C33 (bottom), Appendix C



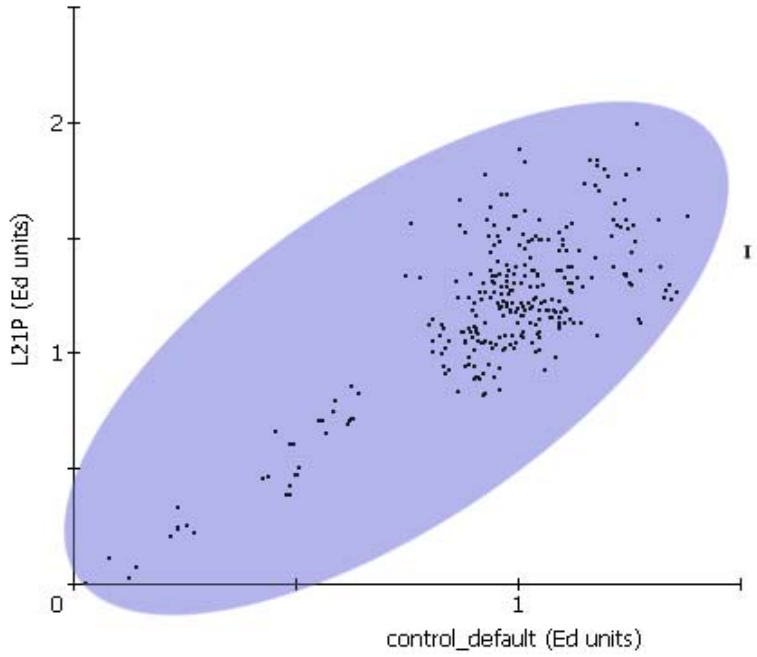
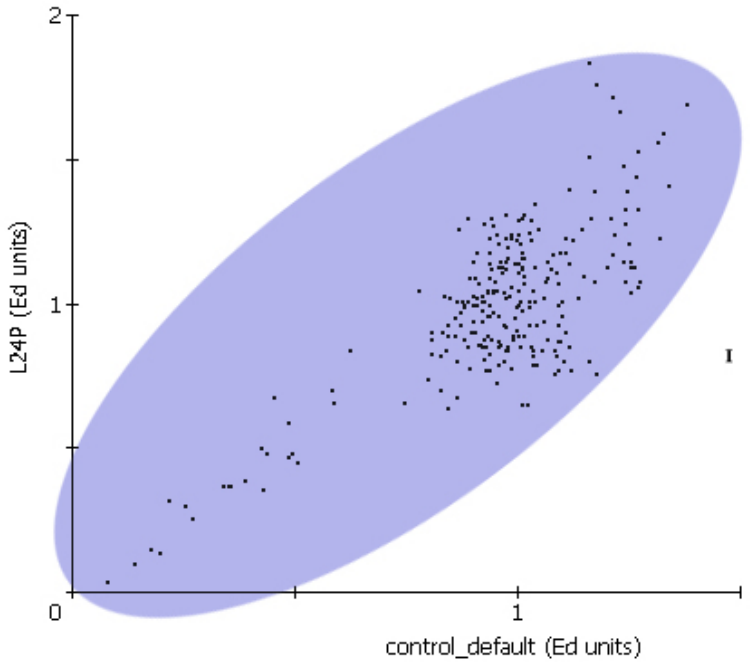


Figure C34 (top) and C35 (bottom), Appendix C



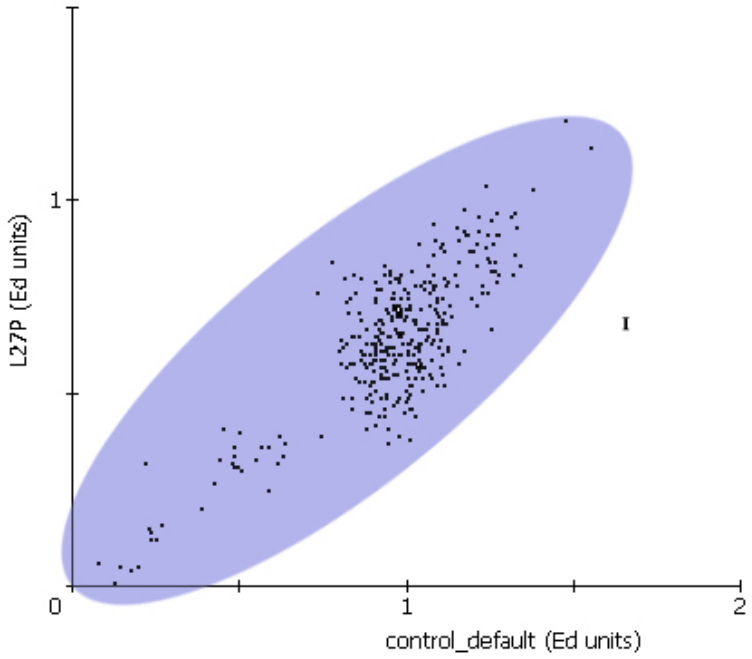
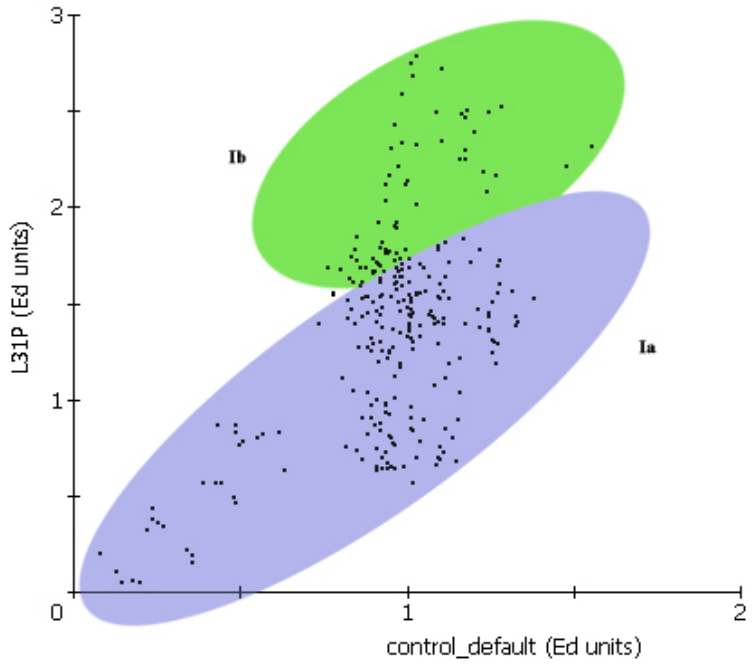


Figure C36 (top) and C37 (bottom), Appendix C



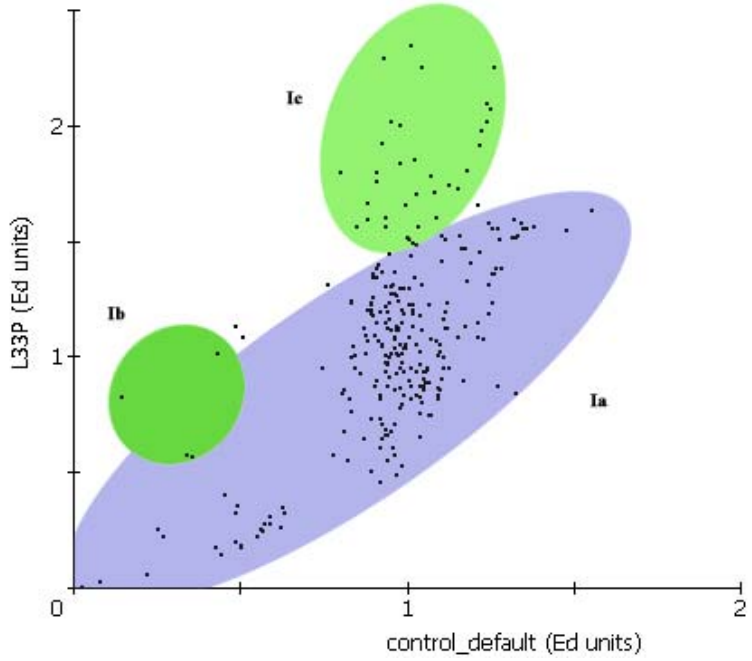
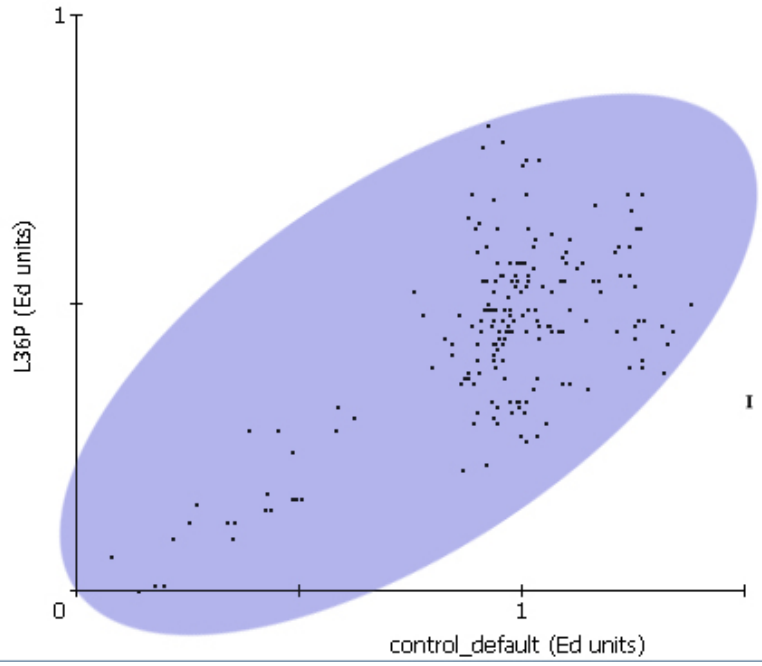


Figure C38, Appendix C



APPENDIX D

RED ANLYSES OF THE AMINOACYL-TRNA SYNTHETASES

The RED method was standardized by using ribosomal genes, and then used to test for LGT events in the evolution of aminoacyl-tRNA synthetases. Chapter 2 provided detail analyses for the 20 aminoacyl-tRNA synthetases (aaRSs). The 20 aaRSs were categorized into seven different RED models. One aaRS was selected to represent each model, and for these aaRSs, detail analyses including plots and hypothetical evolutionary histories diagrams were provide. In addition, phylogenetic trees published by Woese et al. (2000) accompany each aaRS plot (see Chapter 2 references). Not including the control leucyl-tRNA synthetase, this appendix includes the plots and hypothetical evolutionary history diagrams for the remaining 12 aaRS genes examined in Chapter 2. The aaRS genes included are alanyl-, arginyl, asparaginyl-, aspartyl-, glycyl-, histidyl-, methionyl-, alpha and beta chain phenylalanyl-, seryl-, threonyl-, and tryptophanyl-tRNA synthetase.

Figure D1: RED plot of alanyl-tRNA synthetase. Intra-domain comparisons within the archaea (oval I), within the bacteria (oval II), inter-domain comparisons between the archaea and eukaryote (oval III), between the archaea and the bacteria (IV), between the eukaryote and bacteria (oval V).

Figure D2: Hypothetical tree illustrating the proposed evolutionary history of alanyl -tRNA synthetase. A_E and A_C represent the two archaeal sub-domains Euryarchaeota and Crenarchaeota, respectively, B1 and B2 represent two clades of the bacteria, E represents the eukaryotic domain, T represents the acquisition of the eukaryotic gene by bacterial ancestor.

Figure D3: Gene tree of alanyl-tRNA synthetases. Phylogenetic tree of AlaRS sequences, Figure 10 of Woese et al. 2000 (see Chapter 2 references). Organisms in red font belong to the bacterial domain, blue font belong to the archaeal domain, and orange font belong to the eukaryotic domain.

Figure D4: RED plot of arginyl-tRNA synthetase. Intra-domain comparisons within the archaea (oval I), within the bacteria (oval II), between the bacterium *Bacillus* and normal sister taxa, *Enterococcus* and *Streptococcus* (oval III), between the bacterial clade B1 and B2 (oval IV), inter-domain comparisons between the archaea and eukaryote (oval V), between the archaea and the bacteria (VI), between archaea and the bacterium *Deinococcus* (oval VII), between the eukaryote and the bacterial clade B1 (oval VIII), and between the eukaryote and the bacterial clade B2 (oval IX).

Figure D5: Hypothetical tree illustrating the proposed evolutionary history of arginyl -tRNA synthetase. B1 and B2 represent two clades of the bacteria, E represents the eukaryotic domain, and T1 represents the acquisition of the bacterium *Rhodobacter* (RC) gene by another bacterium *Baciulls* (BS).

Figure D6: Hypothetical tree illustrating the proposed evolutionary history of arginyl -tRNA synthetase. A represent the archaeal domain, B1 and B2 represent two clades of the bacteria, E represents the eukaryotic domain, T2 represents the aquisision of the archaeal gene by the bacterium *Deinococcus,* T3 represents the acquisition of the eukaryotic gene by the bacterial clade B2, and T4 represents an ancient nonlinearity LGT.

Figure D7: Gene tree of arginyl -tRNA synthetases. Phylogenetic tree of ArgRS sequences, Figure 20 of Woese et al. 2000 (see Chapter 2 references). Organisms in red font belong to the bacterial domain, blue font belong to the archaeal domain, and orange font belong to the eukaryotic domain.

Figure D8: RED plot of asparaginyl -tRNA synthetase. Intra-domain comparisons within the archaea (oval I), within the bacteria (oval II), between the bacterial clades B1 and B2 (oval III), inter-domain comparisons between the archaea and eukaryote (oval IV), between the archaea and the bacterial calde B2 (oval V), between the archaea and the bacterial calde B1 (oval VI), and between the eukaryote and the bacteria (oval VII).

Figure D9: Hypothetical tree illustrating the proposed evolutionary history of asparaginyl -tRNA synthetase. A represents the archaeal domain, B1 and B2 represent two clades of the bacteria, E represents the eukaryotic domain, T1 represents an ancient nonlinearity LGT, and T2 represents the acquisition of the archaeal gene by the bacterial clade B1.

Figure D10: Gene tree of asparaginyl-tRNA synthetases. Phylogenetic tree of AsnRS sequences, Figure 14 of Woese et al. 2000 (see Chapter 2 references). The plot in Figure D8 only analyzed N1 type of the yeast gene. Organisms in red font belong to the bacterial domain, blue font belong to the archaeal domain, and orange font belong to the eukaryotic domain.

Figure D11: RED plot of aspartyl -tRNA synthetase. Intra-domain comparisons within the archaea (oval I), within the bacteria (oval II), inter-domain comparisons between the archaea and eukaryote (oval III), between the archaea and the bacteria (oval IV), and between the eukaryote and the bacteria (oval V).

Figure D12: Gene tree of aspartyl-tRNA synthetases. Phylogenetic tree of AspRS sequences, Figure 16 of Woese et al. 2000 (see Chapter 2 references). The plot in Figure D11 only analyzed D types of this gene. Organisms in red font belong to the bacterial domain, blue font belong to the archaeal domain, and orange font belong to the eukaryotic domain.

Figure D13: RED plot of glycyl-tRNA synthetase alpha chain. Intra-domain comparisons within the bacteria (oval I). No other domain analyzed for these genes.

Figure D14: RED plot of glycyl-tRNA synthetase beta chain. Intra-domain comparisons within the bacteria (oval I). No other domain analyzed for these genes.

Figure D15: Gene tree of glycyl-tRNA synthetases alpha and beta chain. Phylogenetic tree of GlyRS sequences, Figure 21A of Woese et al. 2000 (see Chapter 2 references). Organisms in red font belong to the bacterial domain, blue font belong to the archaeal domain, and orange font belong to the eukaryotic domain.

Figure D16: RED plot of histidyl -tRNA synthetase. Intra-domain comparisons within the archaea (oval I), within the bacteria (oval II), between the archaeal and the crenarchaeota *Aeropyrum* (oval III), between the bacterial clade B1 and B2 (oval IV), inter-domain comparisons between the eukaryote and the archaea not including *Aeropyrum* (oval V), between the eukaryote and *Aeropyrum* (oval VI), between the bacteria and the archaea not including *Aeropyrum* (oval VII), between *Aeropyrum* and clade B1 (oval VIII), between the eukaryote and the bacterial clade B2 (oval IX), and between the eukaryote and the bacterial clade B1 (oval X).

Figure D17: Hypothetical tree illustrating the proposed evolutionary history of histidyl -tRNA synthetase. A_E and A_C represent the two archaeal sub-domains

Euryarchaeota and Crenarchaeota, respectively, B1 and B2 represent two clades of the bacteria, E represents the eukaryotic domain, T1 represents the acquisition of the eukaryotic gene by bacterial clade B1, T2 represents the acquisition of the eukaryotic gene by the crenarchaeota *Aeropyrum* (AP), and T3 represents an ancient nonlinearity LGT.

Figure D18: Gene tree of histidyl-tRNA synthetases. Phylogenetic tree of HisRS sequences, Figure 12 of Woese et al. 2000 (see Chapter 2 references). Organisms in red font belong to the bacterial domain, blue font belong to the archaeal domain, and orange font belong to the eukaryotic domain.

Figure D19: RED plot of methionyl-tRNA synthetase. Intra-domain comparisons within the archaea (oval I), within the bacteria (oval II), between the bacterial clade B1 and B2 (oval III), inter-domain comparisons between the eukaryote and the archaea (oval IV), between the archaea and the bacteria not including *Chlorobium* and *Porphyromonas* (oval V), between archaea and *Chlorobium* and *Porphyromonas* of bacterial clade B1 (oval VI), between the eukaryote and the bacteria not including the spirochetes (oval VII), and between the eukaryote and the spirochetes of bacterial clade B1 (oval VIII).

Figure D20: Hypothetical tree illustrating the proposed evolutionary history of methionyl-tRNA synthetase. A represent the archaeal domains, B1 and B2 represent two clades of the bacteria, E represents the eukaryotic domain, T1 represents the acquisition of the eukaryotic gene by the spirochetes – *Borrelia* (BB) and *Treponema*

(TP), T2 represents the acquisition of the archaeal gene by the *Chlorobium* (CL) and *Porphyromonas* (PG), and T3 represents an ancient nonlinearity LGT.

Figure D21: Gene tree of methionyl-tRNA synthetases. Phylogenetic tree of MetRS sequences, Figure 5 of Woese et al. 2000 (see Chapter 2 references). Organisms in red font belong to the bacterial domain, blue font belong to the archaeal domain, and orange font belong to the eukaryotic domain.

Figure D22: RED plot of phenylalanyl-tRNA synthetase alpha chain. Intra-domain comparisons within the archaea (oval I), within the bacteria (oval II), between the spirochetes and the other bacteria (oval III), inter-domain comparisons between the eukaryote and the archaea (oval IV), between the archaea and the bacteria not including the spirochetes (oval V), between archaea and the spirochetes (oval VI), between the eukaryote and the bacteria (oval VII).

Figure D23: RED plot of phenylalanyl-tRNA synthetase beta chain. Intra-domain comparisons within the archaea (oval I), within the bacteria (oval II), between the spirochetes and the other bacteria (oval III), inter-domain comparisons between the archaea and the bacteria not including the spirochetes (oval IV), between archaea and the spirochetes (oval V).

Figure D24: Hypothetical tree illustrating the proposed evolutionary history of phenylalanyl-tRNA synthetase alpha and beta chains. A represent the archaeal

domains, B represents the bacterial domain, E represents the eukaryotic domain, T1 represents the acquisition of the archaeal gene by the spirochetes – *Borrelia* (BB) and *Treponema* (TP), T2 represents an ancient nonlinearity LGT.

Figure D25: Gene tree of phenylalanyl-tRNA synthetases alpha and beta chains. Phylogenetic tree of PheRS sequences, Figure 2 of Woese et al. 2000 (see Chapter 2 references). Organisms in red font belong to the bacterial domain, blue font belong to the archaeal domain, and orange font belong to the eukaryotic domain.

Figure D26: RED plot of seryl-tRNA synthetase. Intra-domain comparisons within the archaea (oval I), within the bacteria (oval II), between the spirochetes and the other bacteria (oval III), inter-domain comparisons between the archaea and the eukaryote (oval IV), between the archaea and the bacteria not including the spirochetes (oval V), between archaea and the spirochetes (oval VI), and between the bacteria and the eukaryote (oval VII).

Figure D27: Hypothetical tree illustrating the proposed evolutionary history of seryl-tRNA synthetase. A represent the archaeal domains, B represents the bacterial domain, E represents the eukaryotic domain, T1 represents the acquisition of the eukaryotic gene by the spirochetes – *Borrelia* (BB) and *Treponema* (TP), T2 represents an ancient nonlinearity LGT.

Figure D28: Gene tree of seryl-tRNA synthetases. Phylogenetic tree of SerRS sequences, Figure 7 of Woese et al. 2000 (see Chapter 2 references). The plot in Figure D26 only analyzed S1 type of the yeast gene.Organisms in red font belong to the bacterial domain, blue font belong to the archaeal domain, and orange font belong to the eukaryotic domain.

Figure D29: RED plot of threonyl-tRNA synthetase. Intra-domain comparisons within the archaea not including *Aeropyrum* (oval I), within the bacteria (oval II), between the archaea *Aeropyrum* and the other archaea (oval III), inter-domain comparisons between the eukaryote and the archaeal not including *Aeropyrum* (oval IV), between the eukaryote and *Aeropyrum* (oval V), between the bacteria and the archaea not including the *Aeropyrum* (oval VI), between the bacteria and *Aeropyrum* (oval VII), and between the bacteria and the eukaryote (oval VIII).

Figure D30: Hypothetical tree illustrating the proposed evolutionary history of threonyl-tRNA synthetase. A_E and A_C represent the two archaeal sub-domains Euryarchaeota and Crenarchaeota, respectively, B represents the bacterial domain, E represents the eukaryotic domain, T1 represents the acquisition of the bacterial gene by archaeal sub-domain Crenarchaeota, T2 represents the acquisition of the bacterial gene by the eukaryote.

Figure D31: Gene tree of threonyl-tRNA synthetases. Phylogenetic tree of ThrRS sequences, Figure 9 of Woese et al. 2000 (see Chapter 2 references). Organisms in red

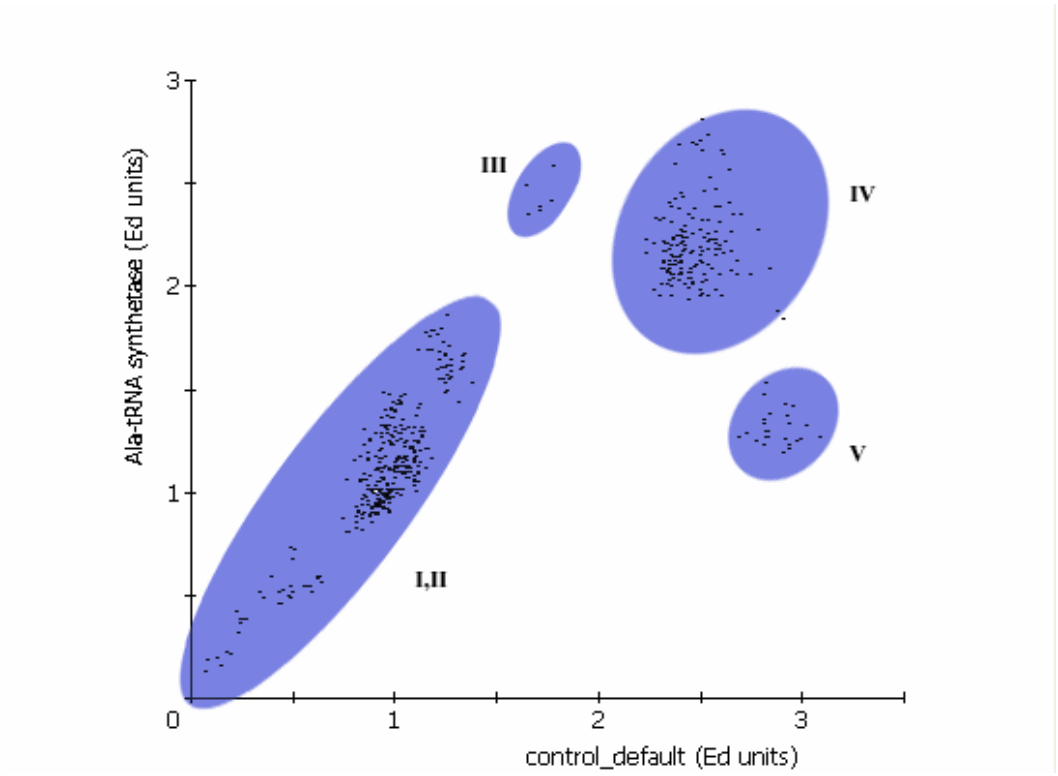
font belong to the bacterial domain, blue font belong to the archaeal domain, and orange font belong to the eukaryotic domain.

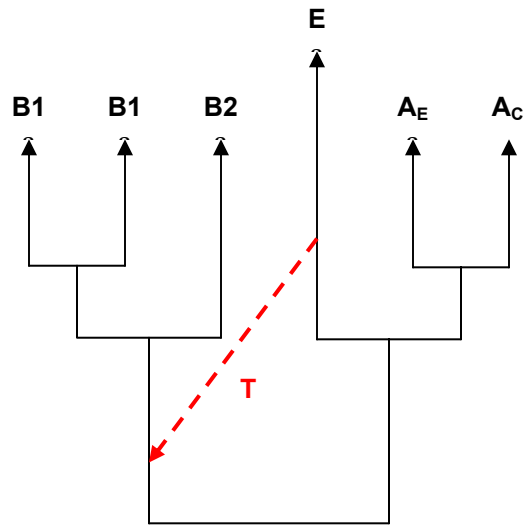
Figure D32: RED plot of tryptophanyl-tRNA synthetase. Intra-domain comparisons within the archaea (oval I), within the bacteria not including oval III comparisons (oval II), comparisons of *Bacillus* and *Enterococcus* with *Streptococcus* (oval III), comparisons of *Neisseria* and *Pseudomonas* with other gamma-proteobacteria (oval III), comparisons between *Deincoccus* and *Streptococcus* (oval IV), inter-domain comparisons between the eukaryote and the archaea (oval V).

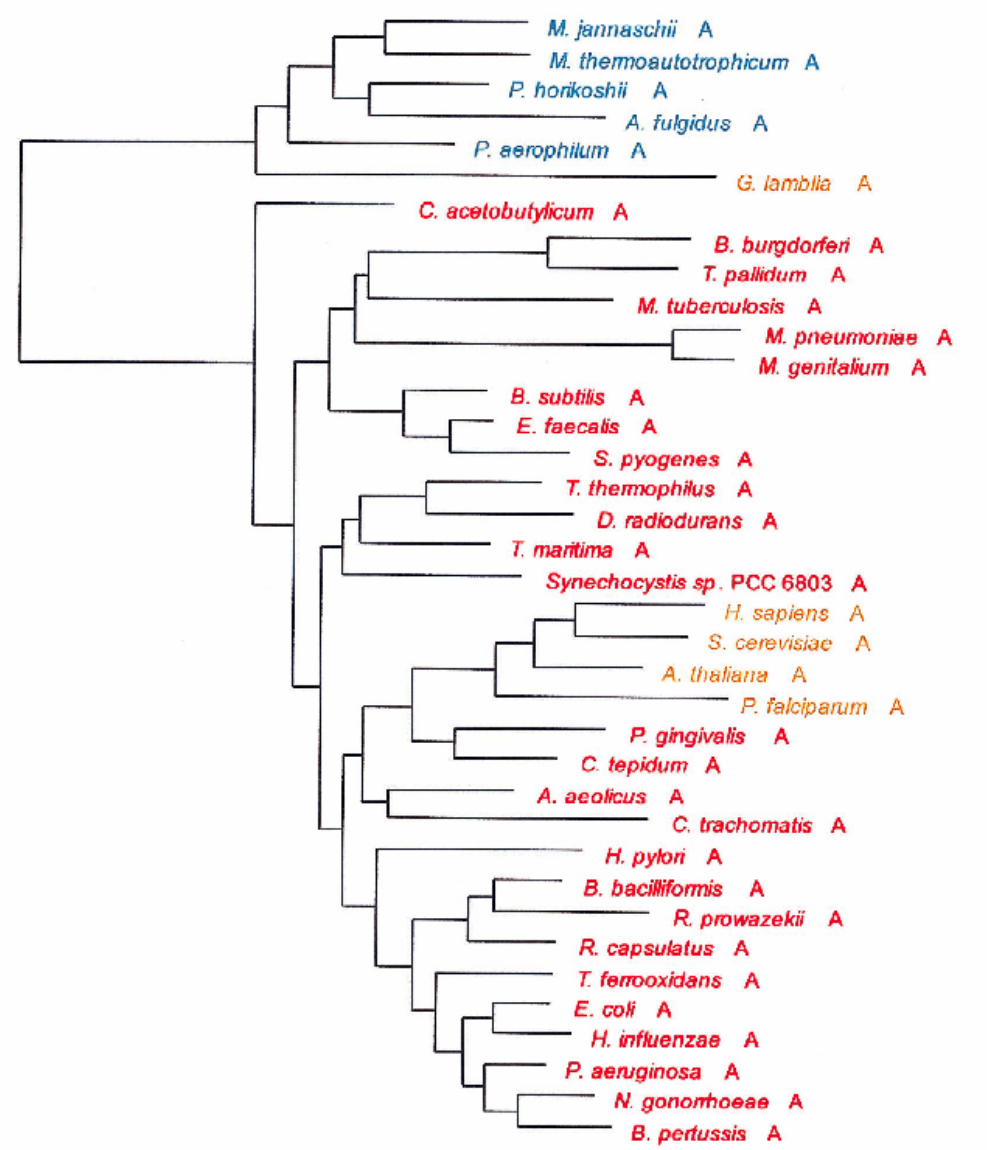
Figure D33: Hypothetical tree illustrating the proposed evolutionary history of tryptophanyl-tRNA synthetase. A1 and A2 represent two archaeal clades, B1 and B2 represents two bacterial clades, E represents the eukaryotic domain, T1 represents the acquisition of *Deinococcus* gene by the *Streptococcus* genes, T2 represents the acquisition of the eukaryotic gene by the archaea *Pyrococcus* and *Pyrobaculum*.

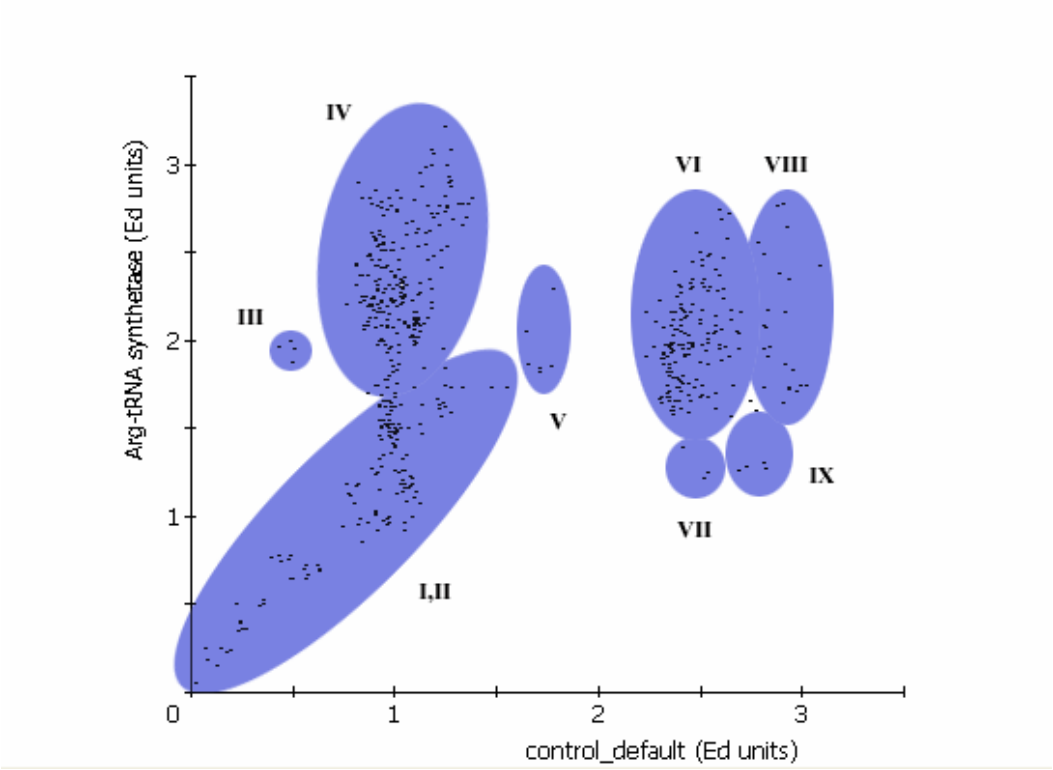
Figure D34: Gene tree of tryptophanyl-tRNA synthetases. Phylogenetic tree of TrpRS sequences, Figure 19 of Woese et al. 2000 (see Chapter 2 references). The plot in Figure D32 only analyzed W2 type of the *Deinococcus* gene, and W1 of the *Enterococcus* gene.Organisms in red font belong to the bacterial domain, blue font belong to the archaeal domain, and orange font belong to the eukaryotic domain.

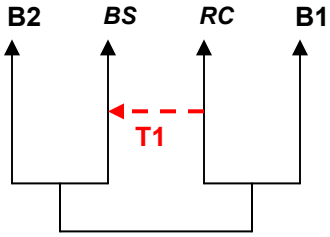
Figures D1 (top) and D2 (bottom), Appendix D

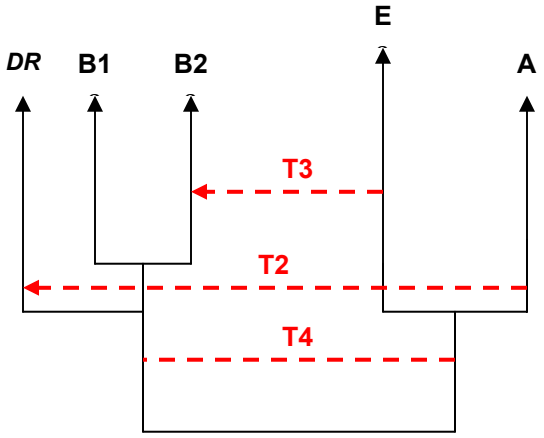


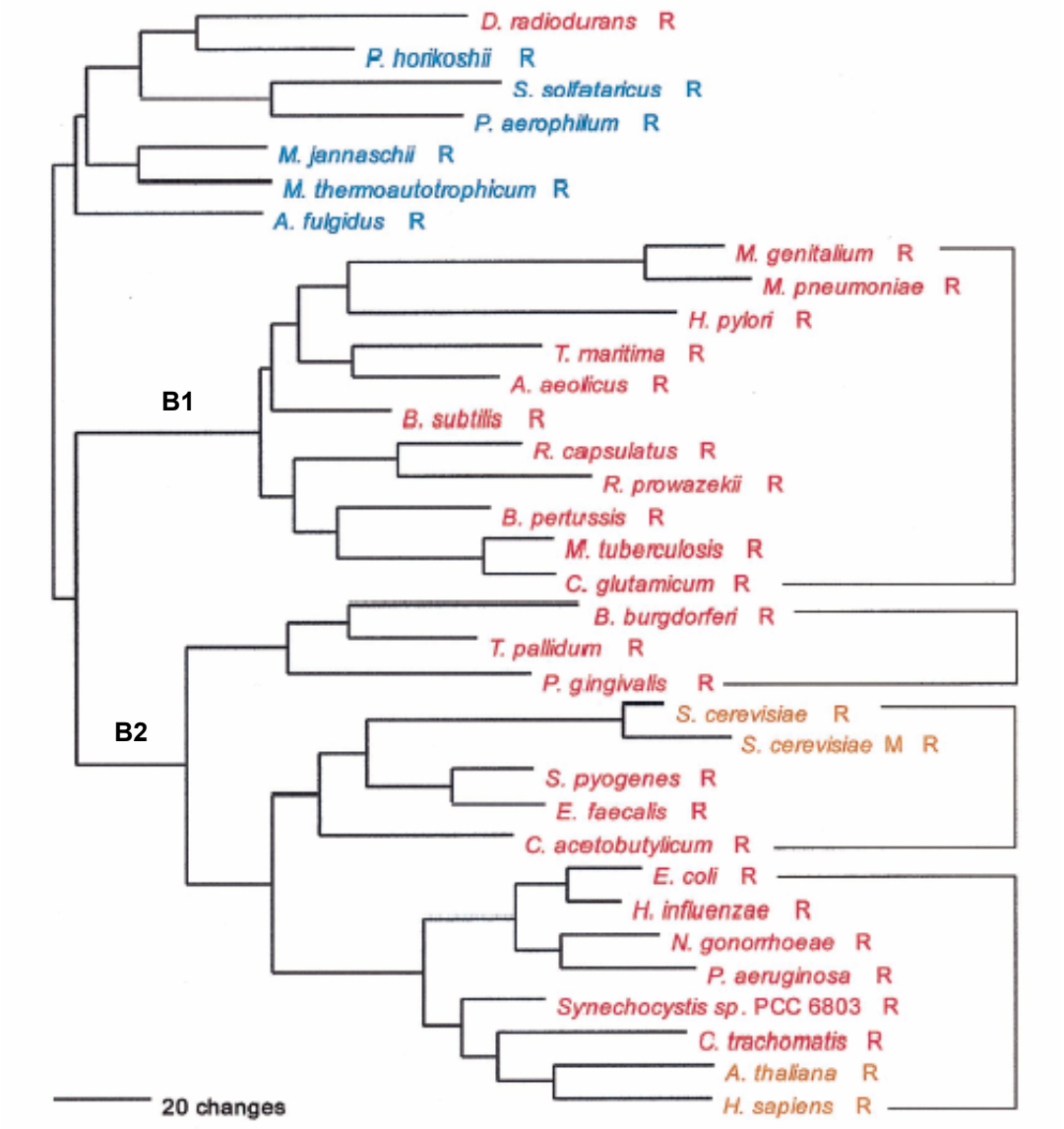



M. jannaschii A
M. thermoautotrophicum A
P. horikoshii A
A. fulgidus A
P. aerophilum A
G. lamblia A
C. acetobutylicum A
B. burgdorferi A
T. pallidum A
M. tuberculosis A
M. pneumoniae A
M. genitalium A
B. subtilis A
E. faecalis A
S. pyogenes A
T. thermophilus A
D. radiodurans A
T. maritima A
Synechocystis sp. PCC 6803 A
H. sapiens A
S. cerevisiae A
A. thaliana A
P. falciparum A
P. gingivalis A
C. tepidum A
A. aeolicus A
C. trachomatis A
H. pylori A
B. bacilliformis A
R. prowazekii A
R. capsulatus A
T. ferrooxidans A
E. coli A
H. influenzae A
P. aeruginosa A
N. gonorrhoeae A
B. pertussis A

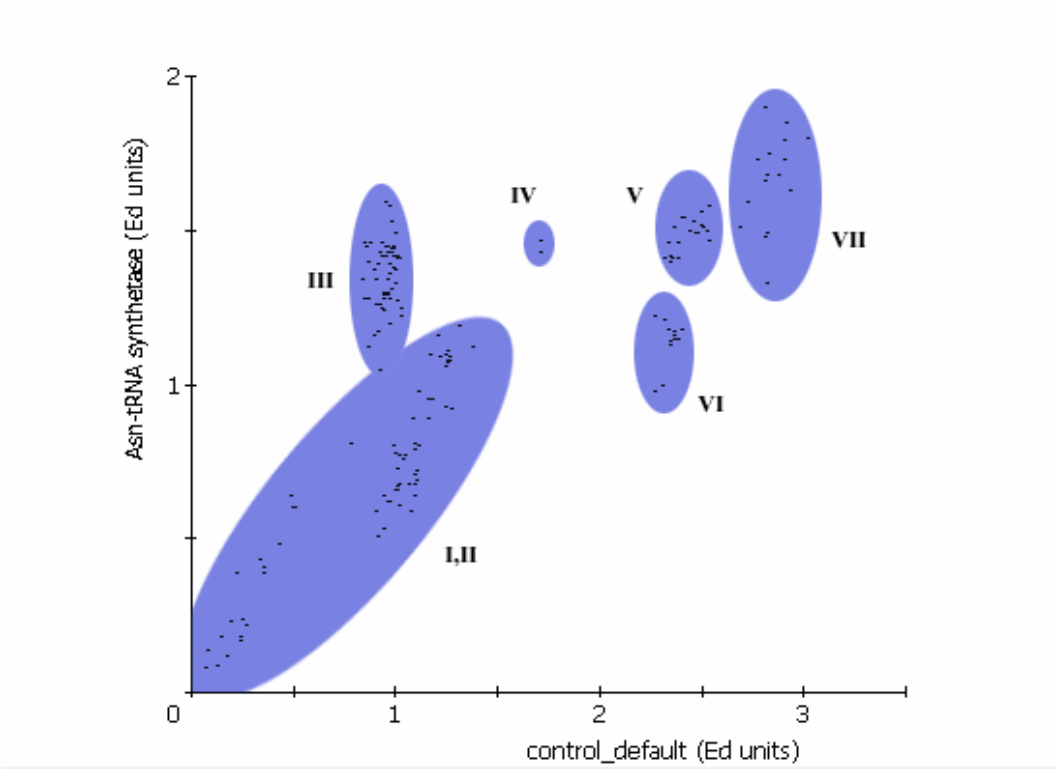

IV
VI
VIII
III
V
IX
VII
I,II
Arg-tRNA synthetase (Ed units)
control_default (Ed units)
0
1
2
3

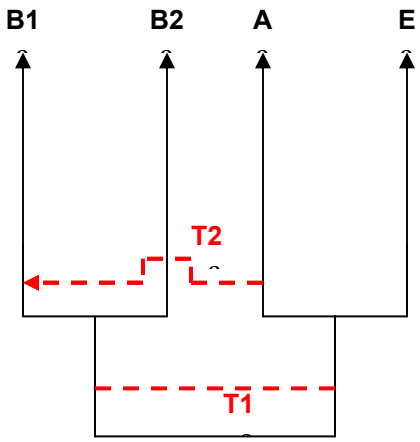

B2
BS
RC
B1
T1

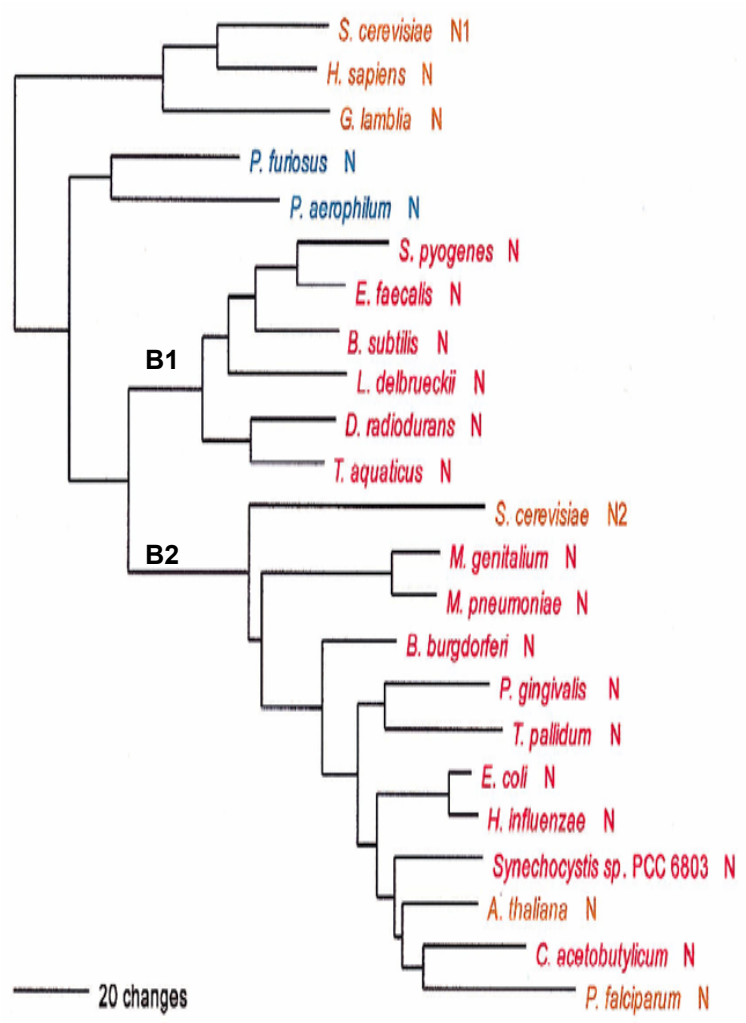

E
DR
B1
B2
A
T3
T2
T4

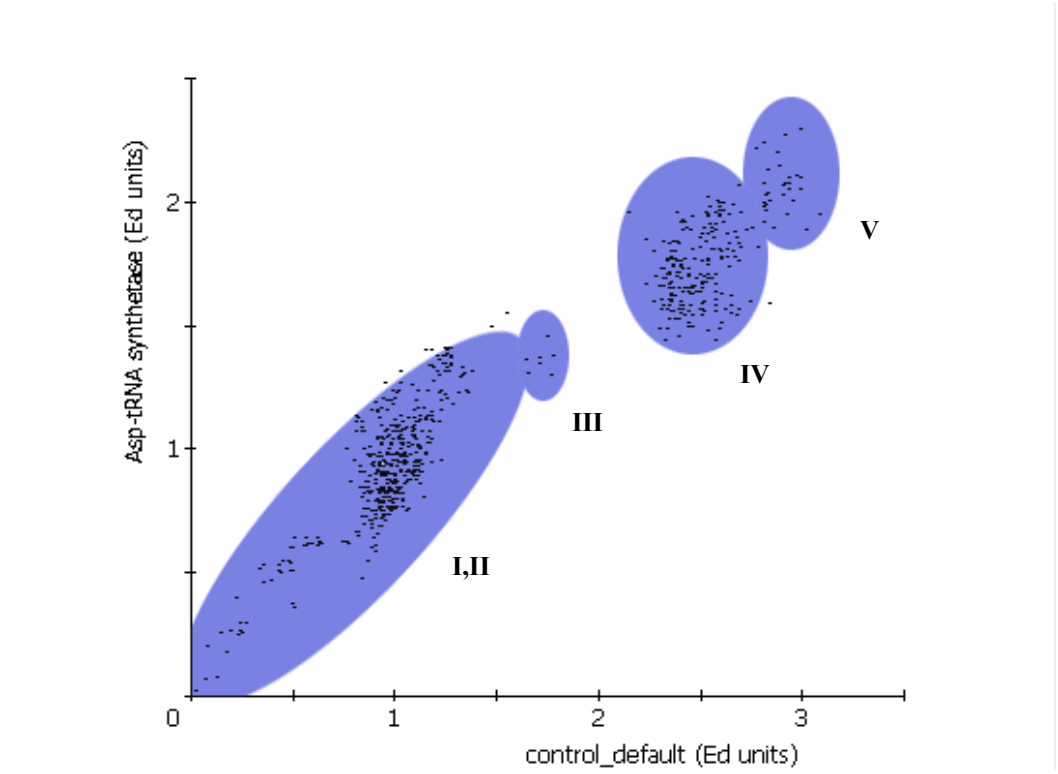

D. radiodurans R
P. horikoshii R
S. solfataricus R
P. aerophilum R
M. jannaschii R
M. thermoautotrophicum R
A. fulgidus R
M. genitalium R
M. pneumoniae R
H. pylori R
T. maritima R
A. aeolicus R
B1
B. subtilis R
R. capsulatus R
R. prowazekii R
B. pertussis R
M. tuberculosis R
C. glutamicum R
B. burgdorferi R
T. pallidum R
P. gingivalis R
S. cerevisiae R
B2
S. cerevisiae M R
S. pyogenes R
E. faecalis R
C. acetobutylicum R
E. coli R
H. influenzae R
N. gonorrhoeae R
P. aeruginosa R
Synechocystis sp. PCC 6803 R
C. trachomatis R
A. thaliana R
H. sapiens R
20 changes

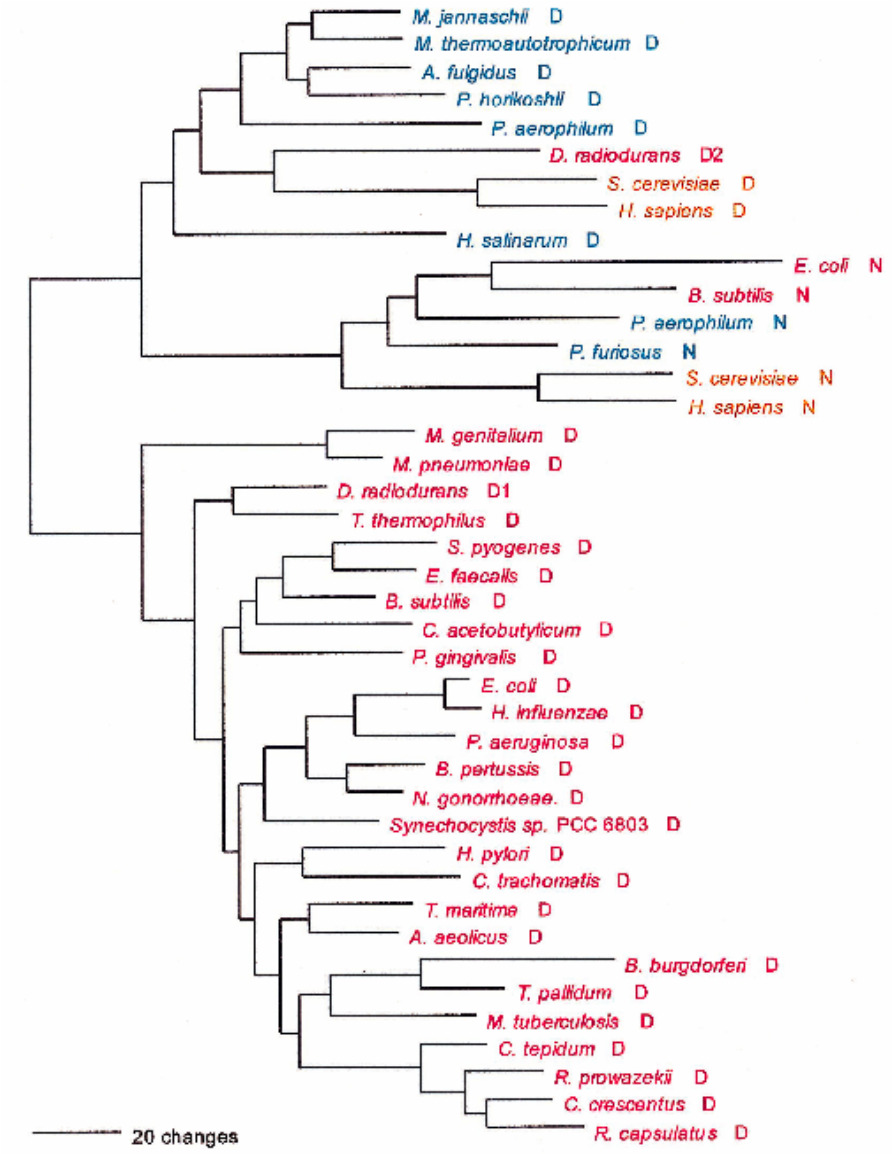
Figures D8 (top) and D9 (bottom), Appendix D



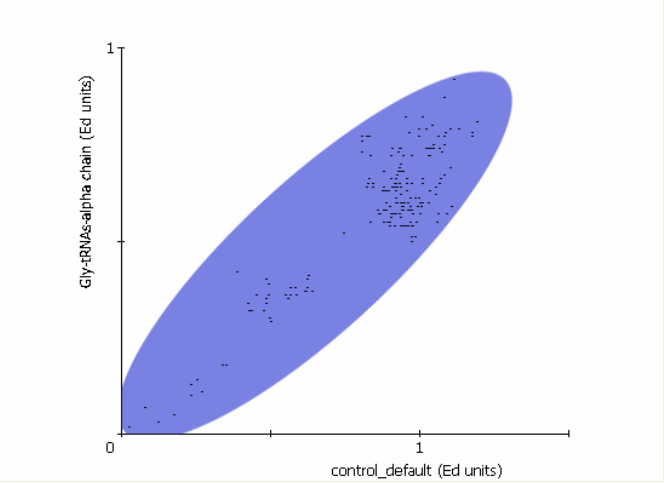


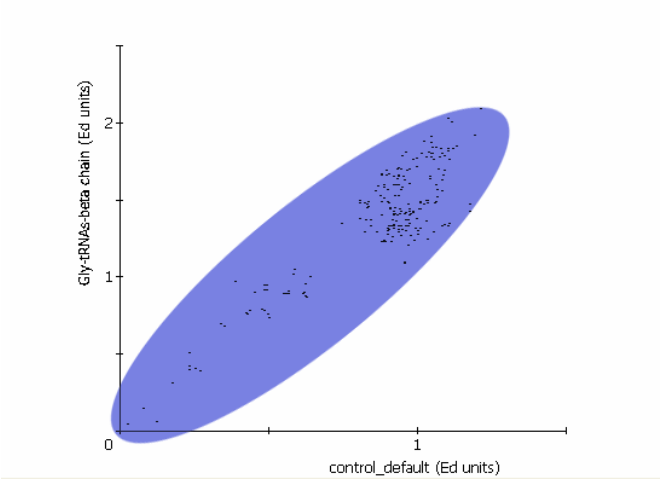

S. cerevisiae N1
H. sapiens N
G. lamblia N
P. furiosus N
P. aerophilum N
S. pyogenes N
E. faecalis N
B. subtilis N
B1
L. delbrueckii N
D. radiodurans N
T. aquaticus N
S. cerevisiae N2
B2
M. genitalium N
M. pneumoniae N
B. burgdorferi N
P. gingivalis N
T. pallidum N
E. coli N
H. influenzae N
Synechocystis sp. PCC 6803 N
A. thaliana N
C. acetobutylicum N
P. falciparum N
20 changes

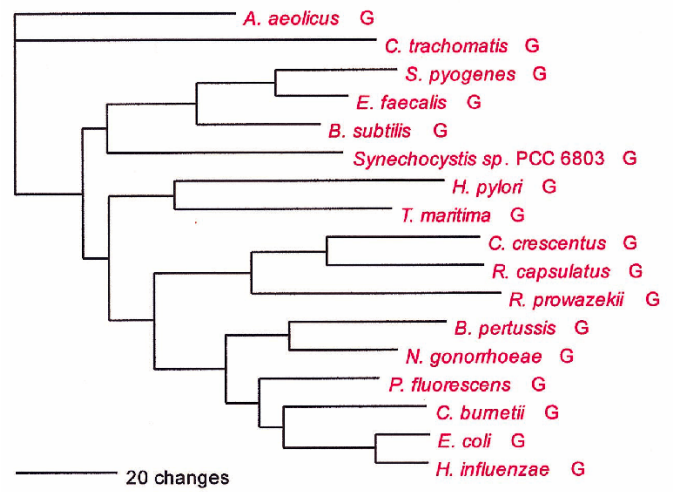

Asp-tRNA synthetase (Ed units)
control_default (Ed units)
0
1
2
3
I,II
III
IV
V


M. jannaschii D
M. thermoautotrophicum D
A. fulgidus D
P. horikoshii D
P. aerophilum D
D. radiodurans D2
S. cerevisiae D
H. sapiens D
H. salinarum D
E. coli N
B. subtilis N
P. aerophilum N
P. furiosus N
S. cerevisiae N
H. sapiens N
M. genitalium D
M. pneumoniae D
D. radiodurans D1
T. thermophilus D
S. pyogenes D
E. faecalis D
B. subtilis D
C. acetobutylicum D
P. gingivalis D
E. coli D
H. influenzae D
P. aeruginosa D
B. pertussis D
N. gonorrhoeae. D
Synechocystis sp. PCC 6803 D
H. pylori D
C. trachomatis D
T. maritima D
A. aeolicus D
B. burgdorferi D
T. pallidum D
M. tuberculosis D
C. tepidum D
R. prowazekii D
C. crescentus D
R. capsulatus D
20 changes

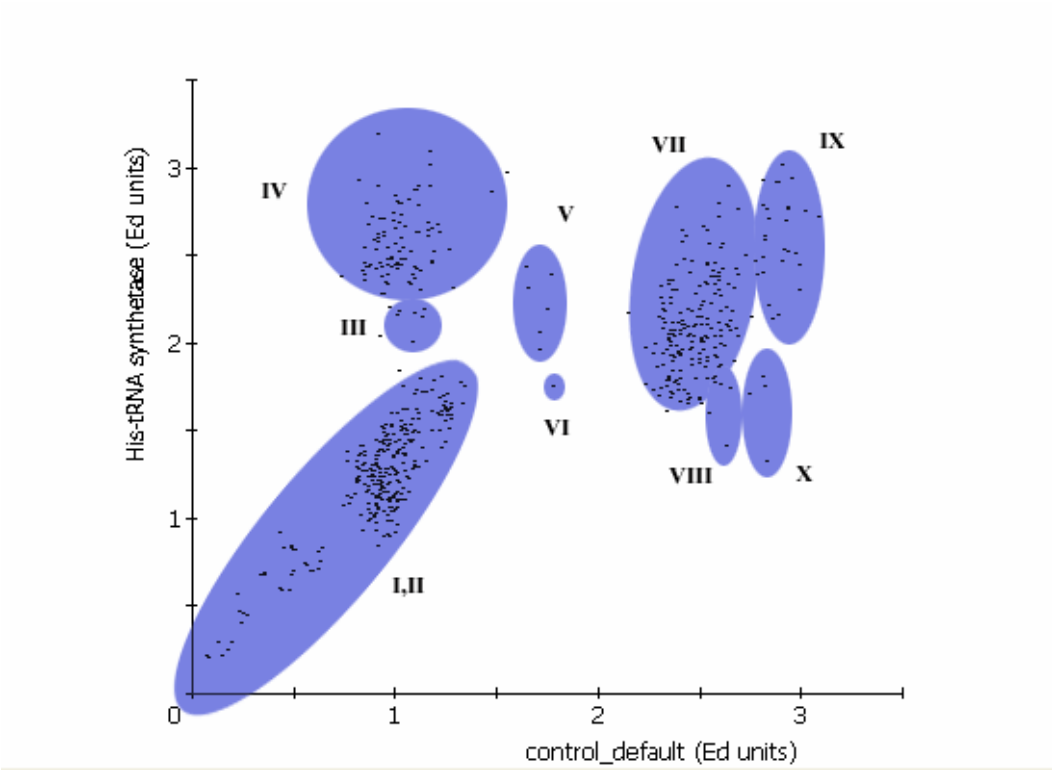
Figures D13 (top), D14 (middle) and D15 (bottom), Appendix D

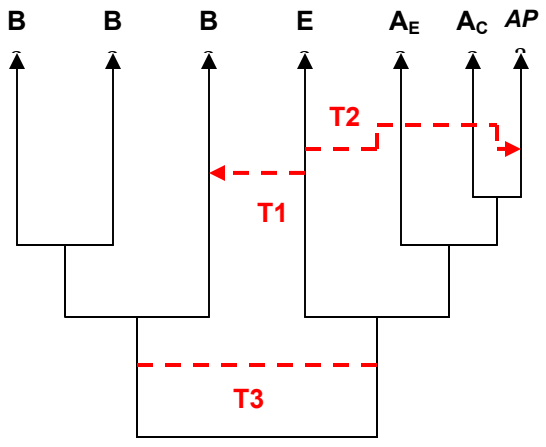


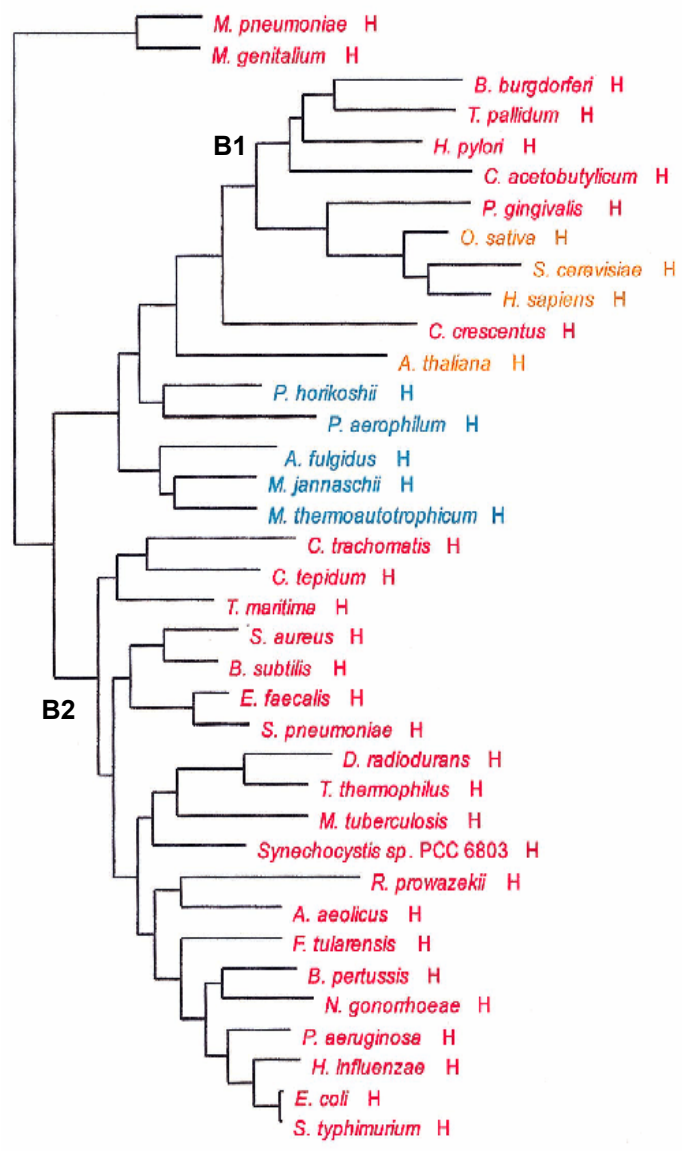




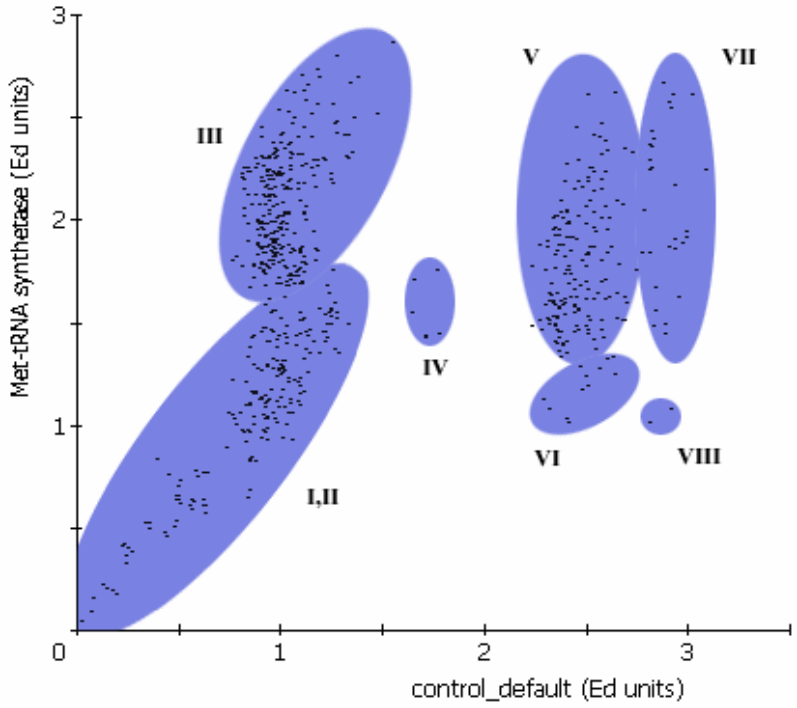
Figures D16 (top) and D17 (bottom), Appendix D

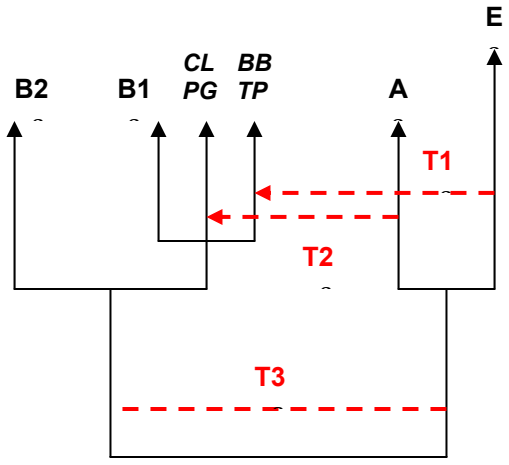


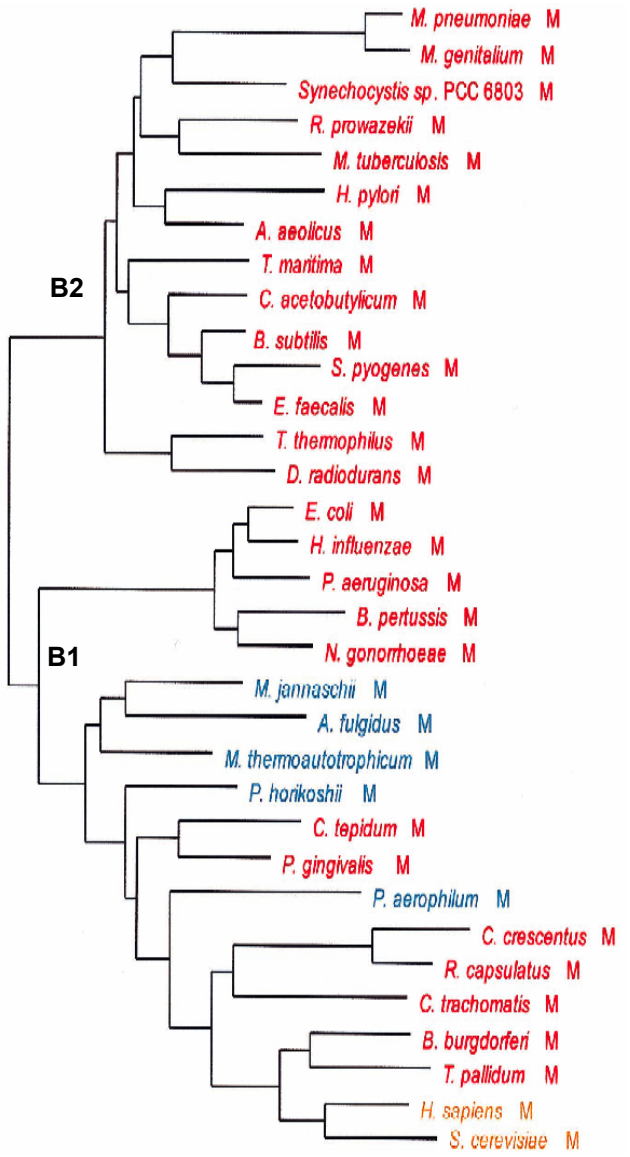



M. pneumoniae H
M. genitalium H
B. burgdorferi H
T. pallidum H
B1
H. pylori H
C. acetobutylicum H
P. gingivalis H
O. sativa H
S. cerevisiae H
H. sapiens H
C. crescentus H
A. thaliana H
P. horikoshii H
P. aerophilum H
A. fulgidus H
M. jannaschii H
M. thermoautotrophicum H
C. trachomatis H
C. tepidum H
T. maritima H
S. aureus H
B. subtilis H
B2
E. faecalis H
S. pneumoniae H
D. radiodurans H
T. thermophilus H
M. tuberculosis H
Synechocystis sp. PCC 6803 H
R. prowazekii H
A. aeolicus H
F. tularensis H
B. pertussis H
N. gonorrhoeae H
P. aeruginosa H
H. influenzae H
E. coli H
S. typhimurium H

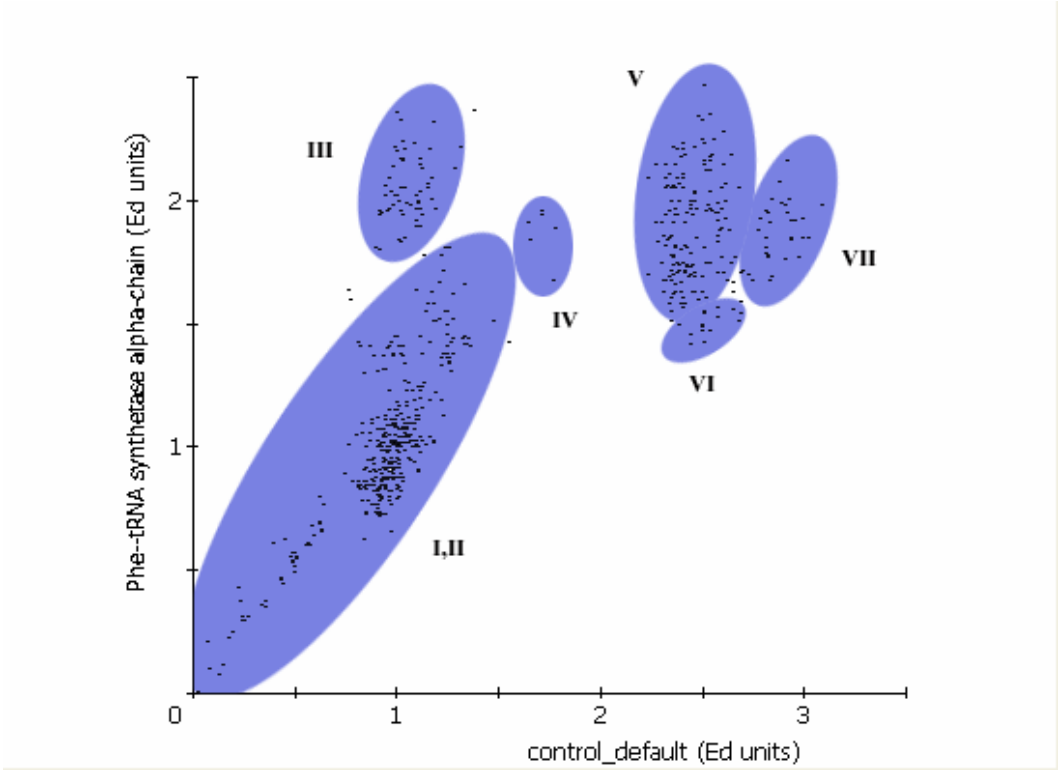
Figures D19 (top) and D20 (bottom), Appendix D

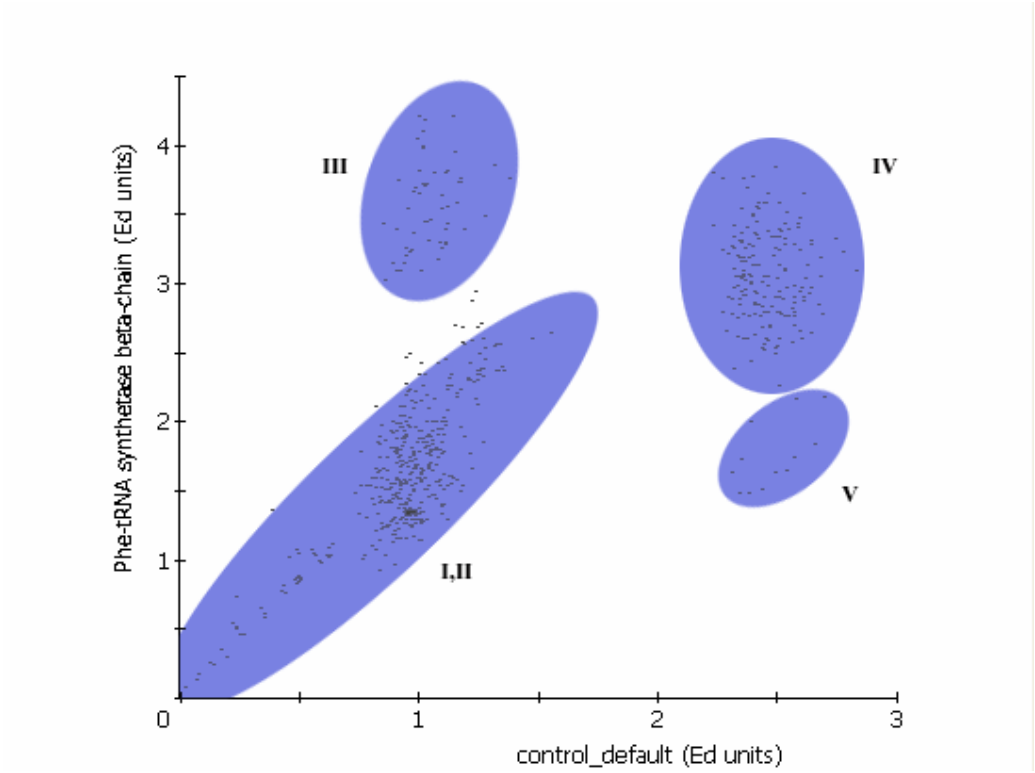




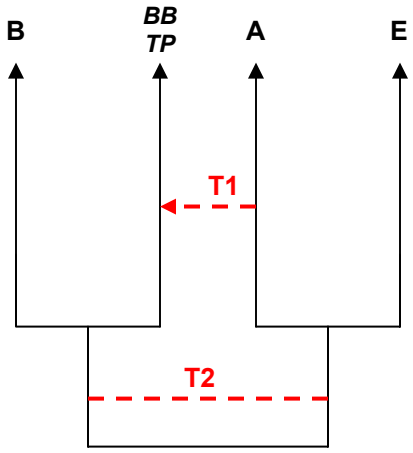

M. pneumoniae M
M. genitalium M
Synechocystis sp. PCC 6803 M
R. prowazekii M
M. tuberculosis M
H. pylori M
A. aeolicus M
T. maritima M
B2
C. acetobutylicum M
B. subtilis M
S. pyogenes M
E. faecalis M
T. thermophilus M
D. radiodurans M
E. coli M
H. influenzae M
P. aeruginosa M
B. pertussis M
N. gonorrhoeae M
B1
M. jannaschii M
A. fulgidus M
M. thermoautotrophicum M
P. horikoshii M
C. tepidum M
P. gingivalis M
P. aerophilum M
C. crescentus M
R. capsulatus M
C. trachomatis M
B. burgdorferi M
T. pallidum M
H. sapiens M
S. cerevisiae M

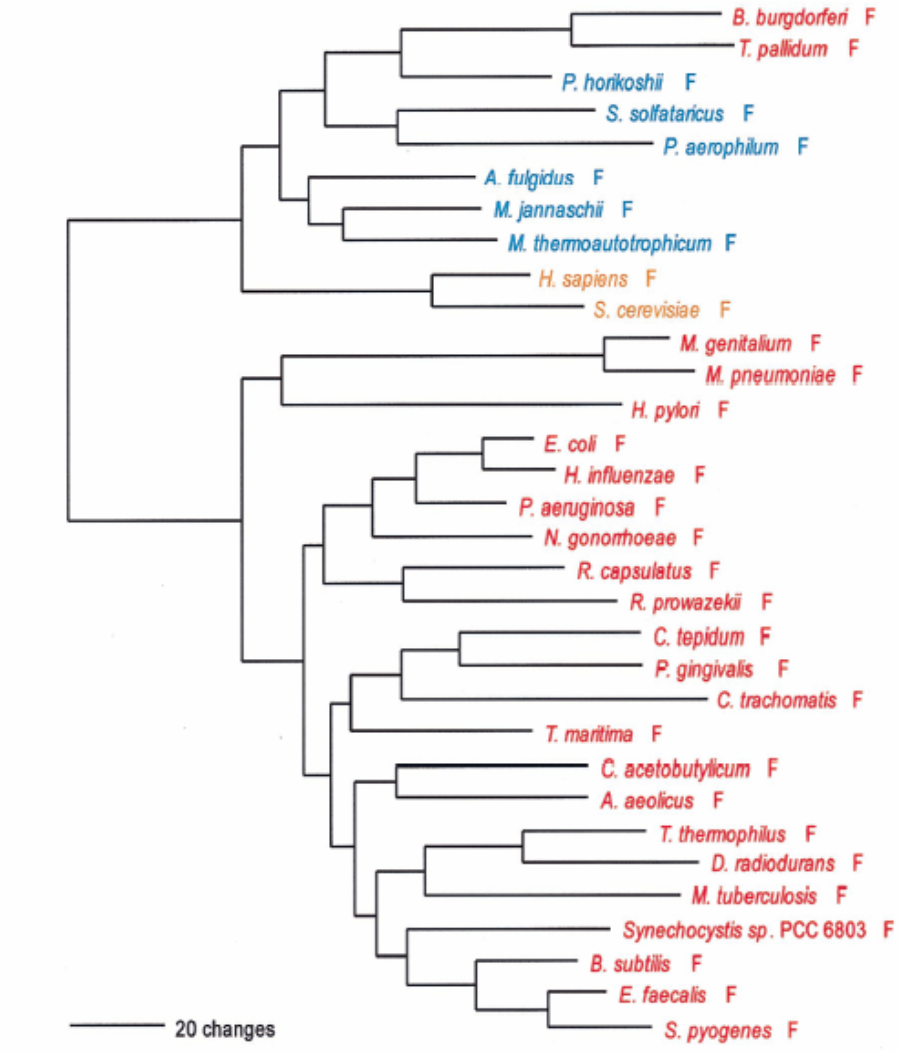
Figures D22 (top) and D23 (bottom), Appendix D



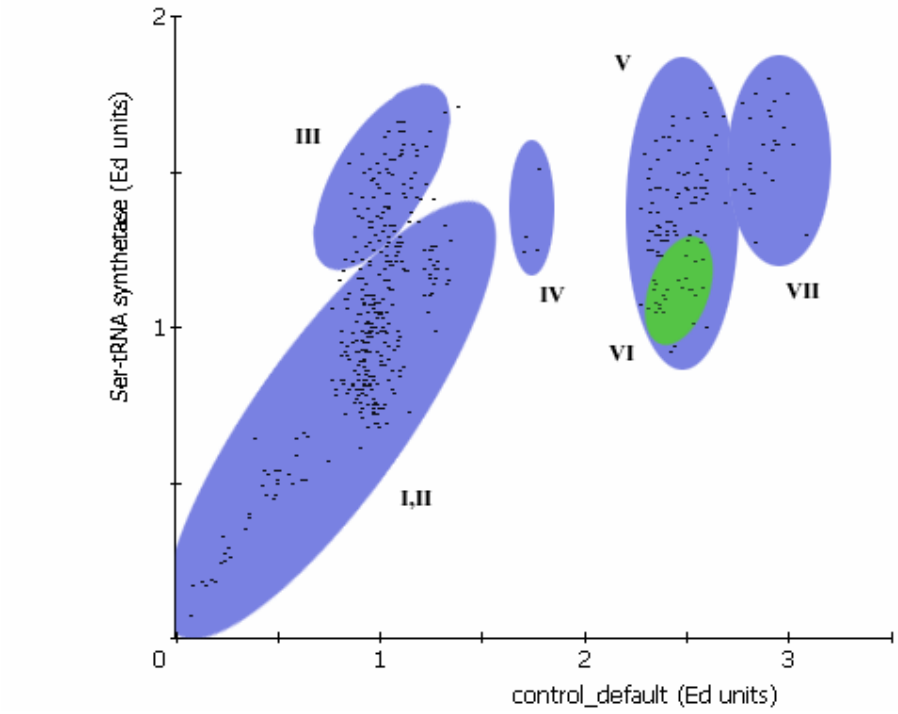


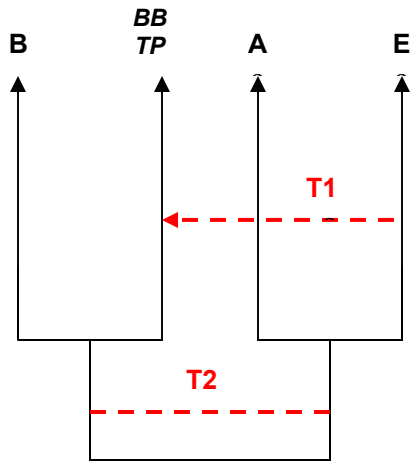
Figures D24 (top) and D25 (bottom), Appendix D

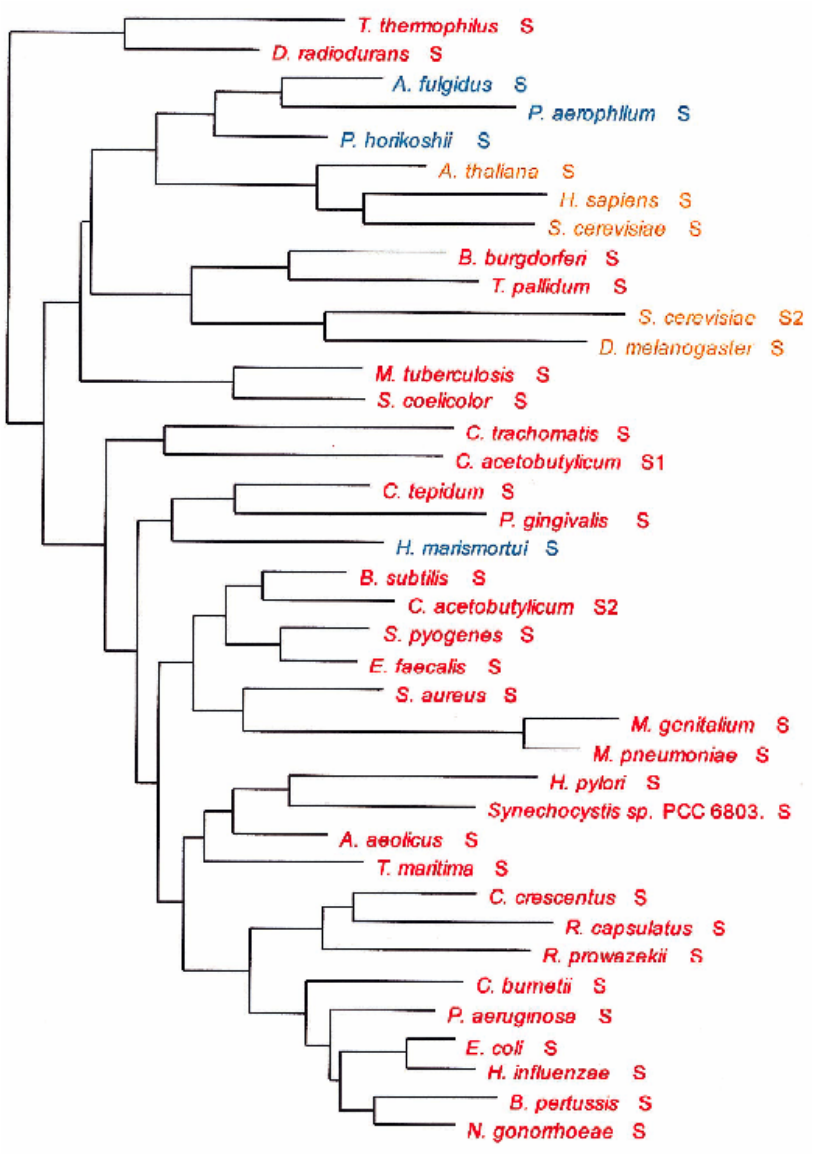




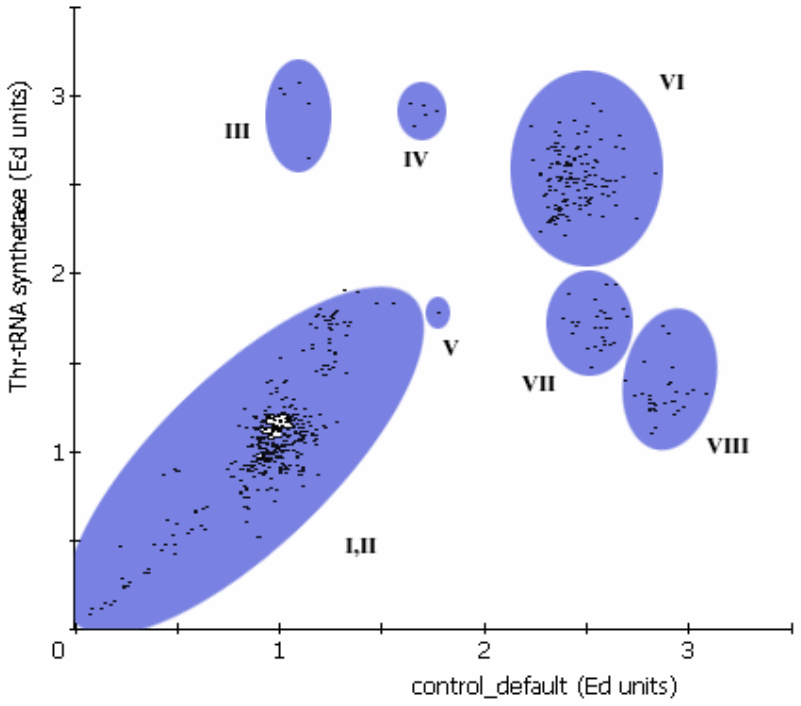
Figures D26 (top) and D27 (bottom), Appendix D

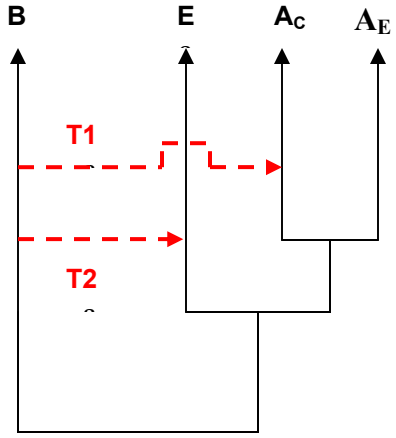


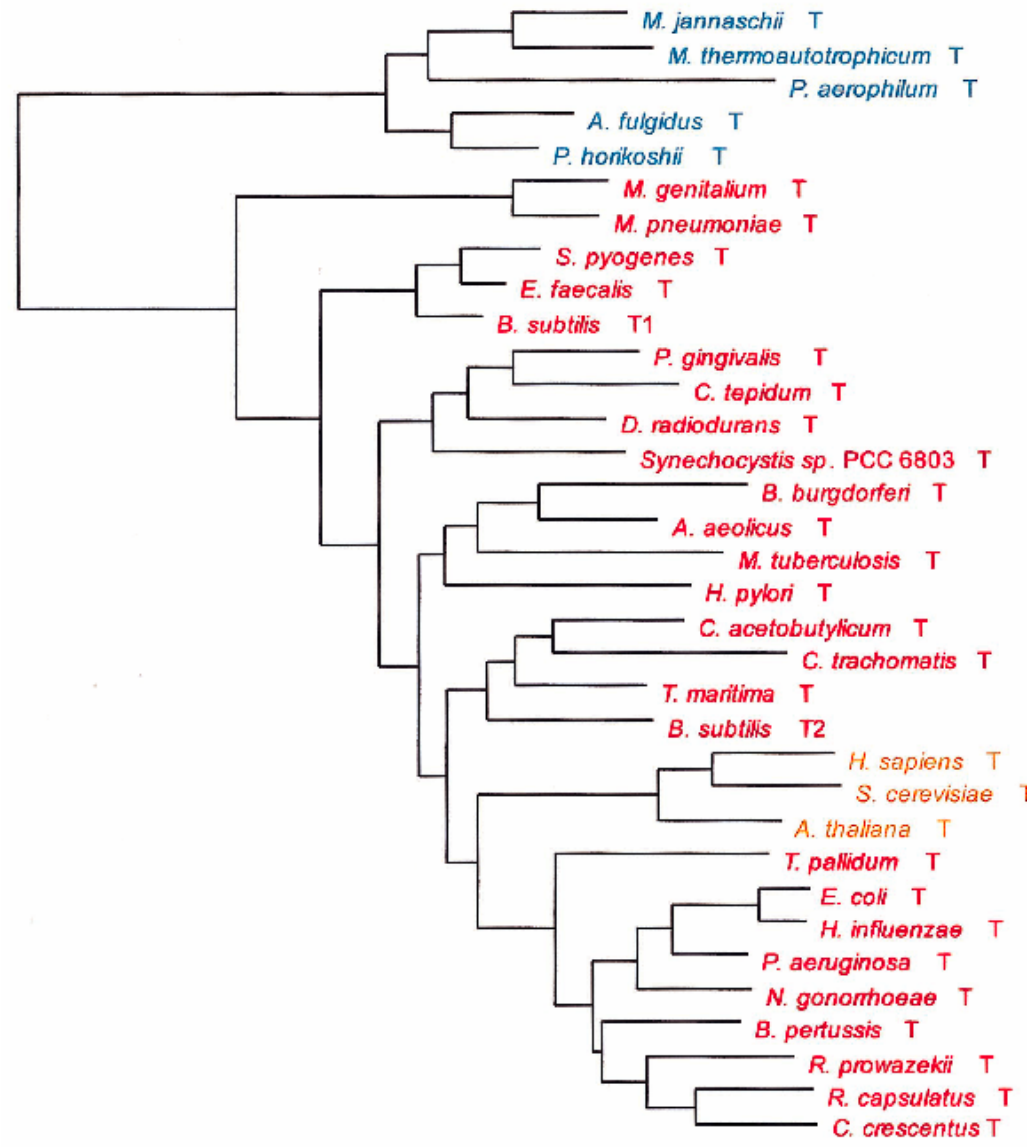



T. thermophilus S
D. radiodurans S
A. fulgidus S
P. aerophilum S
P. horikoshii S
A. thaliana S
H. sapiens S
S. cerevisiae S
B. burgdorferi S
T. pallidum S
S. cerevisiae S2
D. melanogaster S
M. tuberculosis S
S. coelicolor S
C. trachomatis S
C. acetobutylicum S1
C. tepidum S
P. gingivalis S
H. marismortui S
B. subtilis S
C. acetobutylicum S2
S. pyogenes S
E. faecalis S
S. aureus S
M. genitalium S
M. pneumoniae S
H. pylori S
Synechocystis sp. PCC 6803. S
A. aeolicus S
T. maritima S
C. crescentus S
R. capsulatus S
R. prowazekii S
C. burnetii S
P. aeruginosa S
E. coli S
H. influenzae S
B. pertussis S
N. gonorrhoeae S

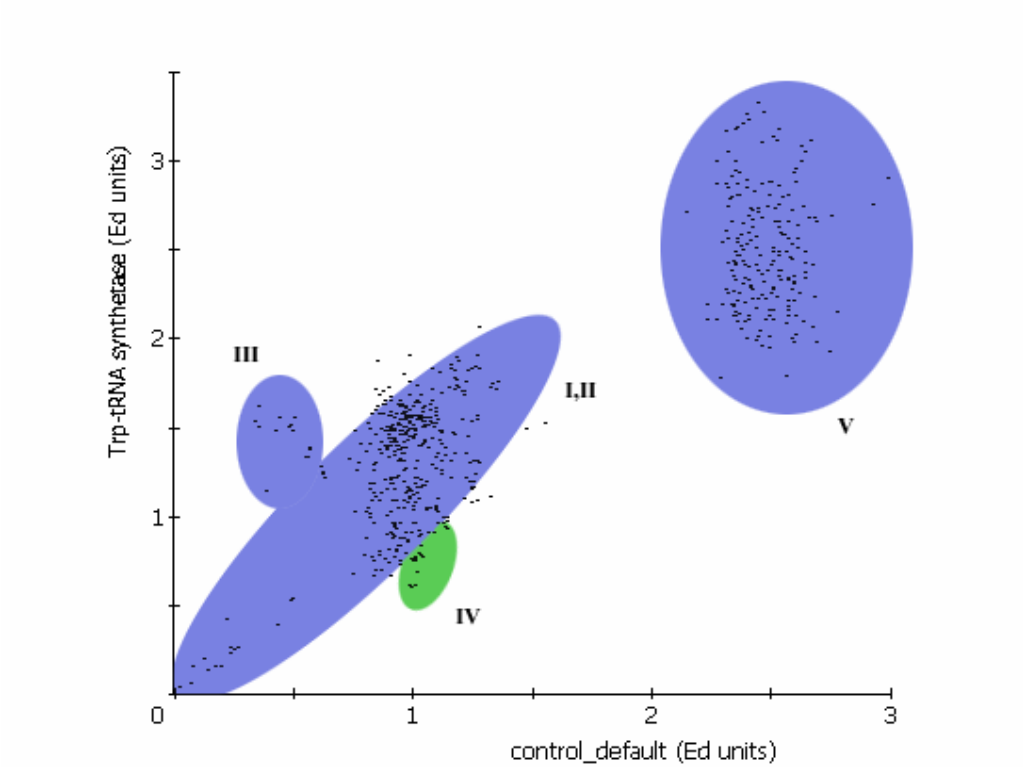
Figures D29 (top) and D30 (bottom), Appendix D

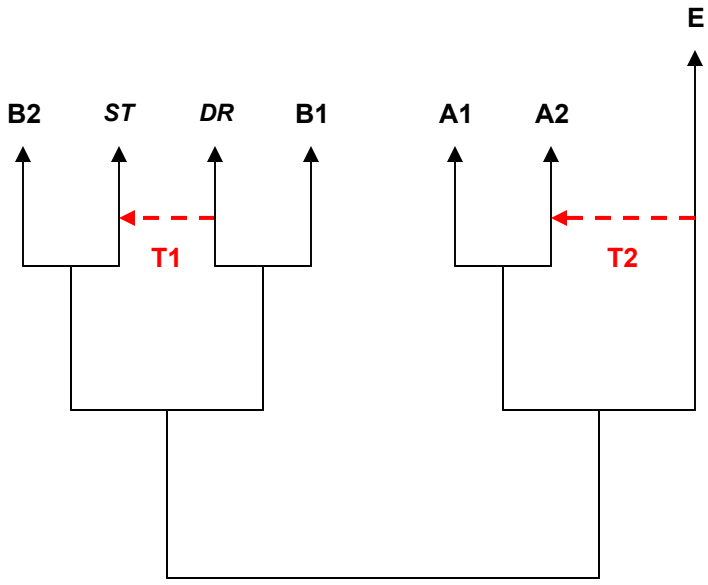


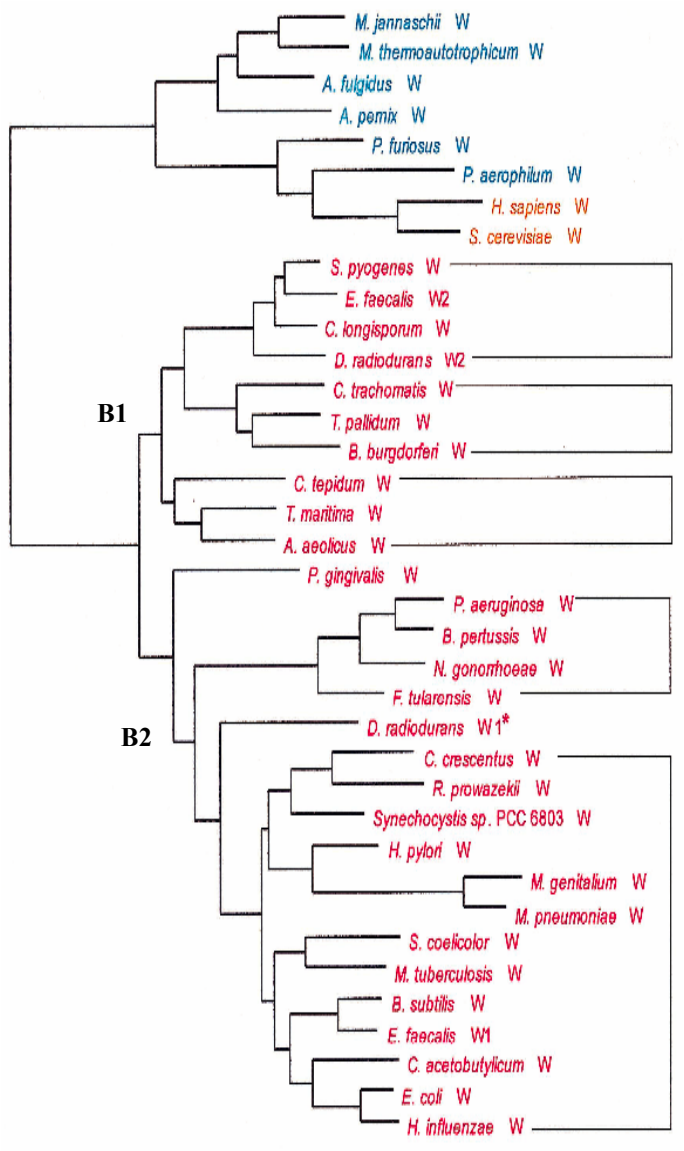



M. jannaschii T
M. thermoautotrophicum T
P. aerophilum T
A. fulgidus T
P. horikoshii T
M. genitalium T
M. pneumoniae T
S. pyogenes T
E. faecalis T
B. subtilis T1
P. gingivalis T
C. tepidum T
D. radiodurans T
Synechocystis sp. PCC 6803 T
B. burgdorferi T
A. aeolicus T
M. tuberculosis T
H. pylori T
C. acetobutylicum T
C. trachomatis T
T. maritima T
B. subtilis T2
H. sapiens T
S. cerevisiae T
A. thaliana T
T. pallidum T
E. coli T
H. influenzae T
P. aeruginosa T
N. gonorrhoeae T
B. pertussis T
R. prowazekii T
R. capsulatus T
C. crescentus T

Figures D32 (top) and D33 (bottom), Appendix D






M. jannaschii W
M. thermoautotrophicum W
A. fulgidus W
A. pernix W
P. furiosus W
P. aerophilum W
H. sapiens W
S. cerevisiae W
S. pyogenes W
E. faecalis W2
C. longisporum W
D. radiodurans W2
C. trachomatis W
T. pallidum W
B. burgdorferi W
C. tepidum W
T. maritima W
A. aeolicus W
P. gingivalis W
P. aeruginosa W
B. pertussis W
N. gonorrhoeae W
F. tularensis W
D. radiodurans W1*
C. crescentus W
R. prowazekii W
Synechocystis sp. PCC 6803 W
H. pylori W
M. genitalium W
M. pneumoniae W
S. coelicolor W
M. tuberculosis W
B. subtilis W
E. faecalis W1
C. acetobutylicum W
E. coli W
H. influenzae W
B1
B2

APPENDIX E

RED-T APPLICATION DOCUMENTATION

This section contains supplementary digital material for RED-T application, as described in Chapter 3. The following subsections include the help files, which includes application feature descriptions and frequently asked questions, is supplied with the application and accessed on-line. Accessibility of RED-T tools and the feedback form are described in the next two sections of this appendix. The underlined text generally represents the active hyperlinks. The home page for these files is http://www.arches.uga.edu/~whitman.

Help for RED-T application (help.html)

Welcome to RED-T (version 2.1.x) Help Page

This help page is to provide information for general use and explanation of various features of the RED-T application.

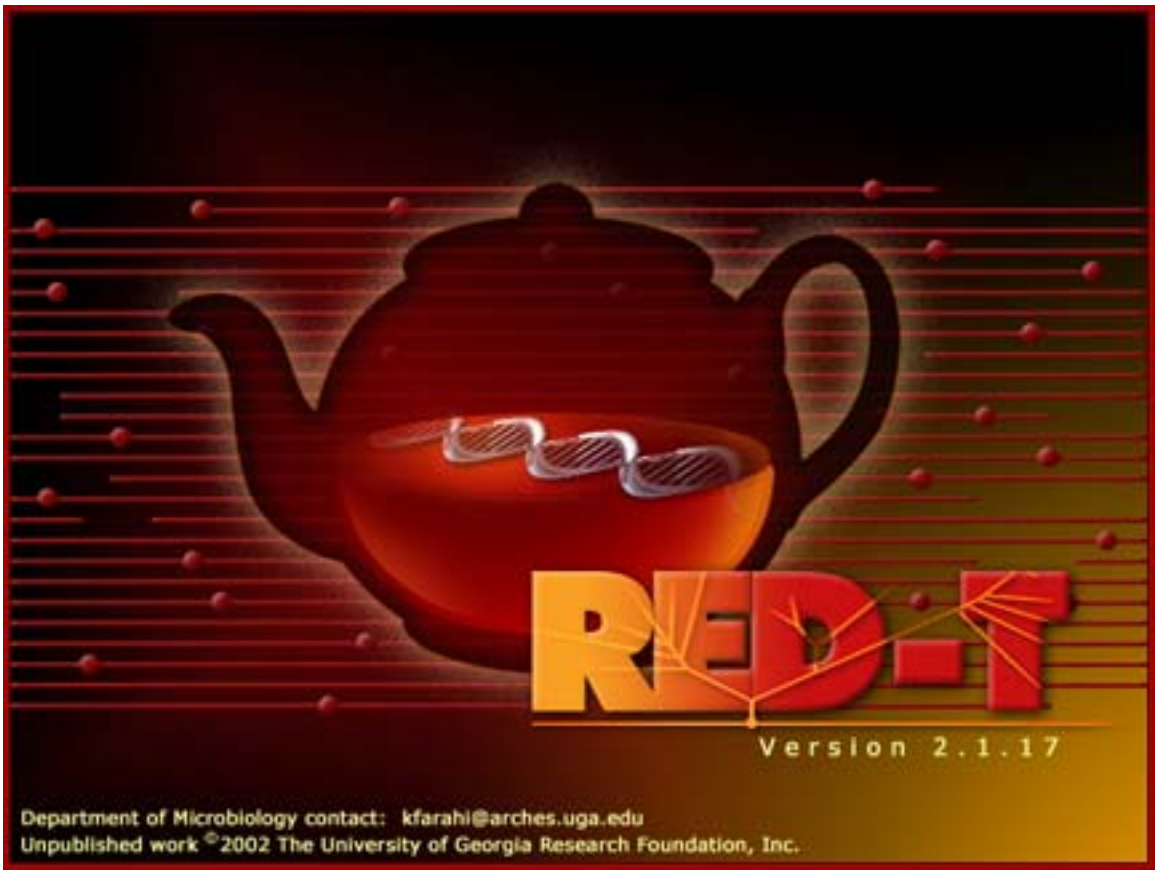


TABLE OF CONTENTS

- Why was RED developed?
- Where can I get more information on this application?
- Why do I need to register before I use these tools?
- How do I cite RED in literature or other works?
- What are the licensing rights for RED?
- How can I find out which version of this application I am using?
- How can I obtain RED source code?
- I can download the archive files (*.exe), but cannot extract (unpack) or open the archive files.
- How can I find out what operating system or type and version of browser I am using?
- After downloading, when I launch (start) RED.jar, the splash screen comes up only.
- The box-selection (neon-green square) does not select plots I want.
- Why are the plot axes different for different plots?
- I cannot import an image of a tree using the import feature.
- I get an error every time I am importing a columnar formatted data.
- How do I add a new organisms to RED-T's organisms list?
- How do I report a bug?
- What are some features planned for implementation in future versions?
- How do I suggest a feature to be implemented in the future versions?

- What features are planned for the next RED release, v.2.2?

5. **Contact information**

WHAT IS RED, AND HOW IS IT DIFFERENT FROM RED-T?

RED-T is a Java application for phylogenetic analysis that is based on a unique method, RED, that utilizes the ratios of evolutionary distances to distinguish between alternative evolutionary histories. RED-T, Ratios of Evolutionary Distances for determination of alternative phylogenetic events, allows the user to examine if any given experimental gene shares the same evolutionary history as the designated control gene(s). Moreover, the tool detects any differences in evolutionary history, and allows the user to examine comparisons of E_d for a likely explanation. Lateral gene transfer (LGT), which may have a significant influence in organismal evolution especially in prokaryotes evolution, is one mechanism that could explain the findings of these RED-T analyses.

The figure below is a screen shot of the RED-T application. The main window is separated into three resizable partitions: The main partition contains the plot generator as illustrated here in highlight mode and a bar displaying quantitative variables that reflect values for the selected portion of the plot. The second partition contains a text box for taking notes of analysis. The last partition contains five tabbed windows to provide different levels of taxonomic selection. In addition, the floating window displays an

imported phylogenetic tree. Illustrated here is the analysis of the isoleucyl-tRNA synthetase gene, in highlight mode. High *K* values resulted from comparisons between bacterial clades B1 and B2 (outlined box), suggesting that one of these clades obtained its gene from another domain. It is clear that the eukaryote and bacterial clade B2 (arrows in plot and tree window) comparisons had much lower *K* values than the eukaryote and bacterial clade B1 comparisons, thus bacterial clade B2 was the likely recipient of the eukaryotic isoleucyl-tRNA synthetase gene.

Back to Top

GETTING STARTED

- **Software requirements:**

  RED-T is not distributed as an executable (*.exe), but as a JAR file (RED.jar). Since JAR is operating system independent (can be used with Windows, Apple, Linux), all that is need to run this application is JAVA Runtime Environment (JRE) 1.4 or above. If your machine does not have this version of JRE, it will attempt to retrieve and install the latest JRE. If you would like to download and install it yourself, find the latest JRE along with installation instructions at the Sun Microsystems web site, java.sun.com.

If you don't know which operating system or platform is running on your machine, you can click the below buttons to retrieve this information, along with information about your browser type/version and monitor.

Since the tool is currently distributed as a archive file (*.exe), a software to decompress this file, such as WinZip or WinRAR is necessary. For more information go to the following link.

Back to Top

- **Installation instructions:**

There is no installation required.

If you are downloading the tool (packaged as an archive file), then follow these easy steps:

1. If prompted to 'OPEN' or 'SAVE' the file, choose to 'SAVE'.
2. After downloading the latest RED-T package, unpack it. To unpack it, double click on the *.exe file and choose a directory, and/or a folder such as "RED" to download into. This maintains the file structure that is essential for operation of the application.
3. To launch (start) RED-T, double click on the **RED.jar** file...
4. Choose from different packages, with one including the "ANALYSES" folder which contains the 60 gene samples analyzed during our investigation (in "Samples" sub-folder).

a. **To download ANALYSES folder, link here and save the "RED-T_Analyses.exe" file.**
b. Then unpack like above insructions.
c. Then place (save) the ANALYSES folder in the same location as the RED.jar file (next to the VisualNumerics folder).

Remember that all files and folders are required to be in the same directory as the RED.jar for the application to launch.

If you are launching the tool from your browser, then don't worry about a thing. Your browser will do everything for you. Just click on the active link for the application, and it will launch if your browser has JRE 1.4 or above.

Back to Top

FEATURES

- **File Menu items:**
    - **New** - this item starts the RED-T wizard that provides the user with three options: import new data to compare with the default control, import new data with the current control, or generate a new control.

- **Open** - this item allows the user to open a previous plot saved under the 'Analyses' folder. In order to open a analysis saved by a previous version of RED-T (i.e. applet), use the 'New' menu item to import using the wizard.

- **Save** and **Save As...** - these items allow the user to save the current analysis in a project folder under the 'Analyses' folder. The control used, the imported experimental data, the merged comparisons (Raw data), taxonomical mapping information identified during control/experimental development, and the notes taken are saved under this folder. 'Save as...' allows the user to specify a name for this project folder. 'Save' item automatically saves the project as the previous name, or uses the name specified under wizard protocol.

- **Import tree** - this feature allows the user to import a graphic of a tree (or any other graphic such as a table, screenshot, plots, or pathways) to assist in the phylogenetic analyses. The images are imported into the floating Tree Image window. This window is turned on by the View menu item. Note that all graphics have to be in GIF or JPEG (JPG) formats.

- **Print** - This item allows the user print different parts of the analyses. The user is given a list of items that can be printed - Plot & Calculations (checked as default), Notes, Organisms (content of the Organisms tab), and Raw data.

Back to Top

- **View Menu items:** (these items can be turned on/off during analysis)

    - **Tree Image** - this item displays the imported tree image, which is either imported by the File Menu's Import Tree option, or an image already stored in the project folder. Remember that this image can be a table, plot, or any other image that the investigator would like to have during analysis.

    - **m-line** - allows for further interpretation of the mean of *K* values. Two lines are drawn through the origin with the slope equal to the mean of *K* (*m*). One (red colored) is immobile and reflects the *m* value of the displayed plots, which equals the quantitative variable displayed. The other line (gray colored) is able to pivot about the origin to make comparisons to the immobile line (red colored). This allows the investor to visually see the differences between various data (other clusters or individual plots).

Back to Top

- **Toolbar items:**

    - **Highlight mode** - when this feature is chosen (check box under the main menu bar) the user can highlight data from specific taxonomic groups. This feature is very useful for identifying outlying clusters.

- **Refresh/Select All/Clear All buttons** - These button are found in the top left corner above the main partition. The 'Refresh' button allows the user to update the plot with any changes, such as selecting a new set of data or clearing the previous view. 'Select All' button allows the user to select all possible comparisons, and 'Clear All' button clears the plot completely.
- **Plot Zoom** - this feature allows the user to zoom in and out of the plot by setting the range of the x and y-axes in evolutionary distance (Ed) units.
- **Boxed-selection feature** - this feature allows a user to obtain the identification of all data points within a boxed area. This feature does not need to be turned. To use it, select the part of the plot for identification by holding the (left) mouse button and dragging the mouse from left to right.

Back to Top

- **Main partition** - this window displays the current plot, including the labeled axes and data selected.

- **Quantitative variables bar:** this bar is found immediately below the main partition and includes the sample size or number of values (*n*), mean of *K* (*m*), standard deviation of *K* (*SD*), and relative standard deviation (*rSD*), which is the *SD* of *K* divided by the m of *K*. *K* is the ratio of E_d of the experimental gene (y-axis) divided by the mean E_d of the control genes (x-axis) for each pair of

organisms. The values for the selected data are shown in **red** font, while the values for all of the data are in parentheses.

Back to Top

- **Taxonomic selection partition:**

    - **Intra-group and Inter-group tabs:** in order to evaluate evolutionary relationships between taxonomic groups, the intra-group and inter-group tabs are available. For instance using these tabs, it is possible to compare genes from the archaeal domain to genes from the proteobacteria.

      The intra-group comparisons tab, allows the user to compare within a group or taxonomic level. For example, if you would like to see all comparisons within Crenarchaeota, then expand the Archaea domain label by one level (click on the "+" next to Archaea) and click or highlight the sub-domain Crenarchaeota.

      Similarly the inter-group comparisons tab contains two panels which allows the user to compare between groups. For example, if you would like to see all comparisons between Crenarchaeota and Bacteria, then click or highlight the Crenarchaeota in one of the panel and click or highlight the Bacteria in the second panel.

The levels of taxonomy is modeled after EMBL and ERGO databases. The tree-file structure can by expanded by clicking on the "+", and collapsed by clicking on the "-" next to each label.

Back to Top

- **Organisms tab** - this window identifies the data currently being displayed or analyzed. Every organism for which data is available is listed. When the data is selected, the box in the first column is checked. The second column contains the abbreviation (RED-id's) for each organism used during importing of data. When the data is selected, the RED-id turns into a bold **red** font. The third column contains the genus and species identification for each organism. The fourth column contains the taxonomic hierarchy for each organism. The last column is the original identification code or accession number every sequence from the original imported data. The third, fourth and fifth column is color coded for each of the three organismal domains - green for Archaea, blue for Bacteria, and purple for Eukarya.

- **n-Comparisons tab** - this window allows the user to select specific data points for analysis. Each row includes a box for selection of the data point, the RED-id's of the two organisms being compared, the x and y coordinates of the data point, and the y/x ratio.

- **Raw data tab** - is just that, raw data! This is the data as read by RED-T as it loads a plot. So if you like to cut and paste from this window and save it

for your reports, this is the place to do it. This is also a good place to see the list of selected data because only the selected data is displayed. Each row contains the RED-id's of the two organisms being compared, the x and y coordinates of the data point, and the y/x ratio.

Back to Top

- **Analysis notes** - this separate partition is found below the main partition and contains a text box for taking notes. These notes are saved with each plot. Examples of notes are included for the genes we analyzed in the ANALYSES folder.

- **Wizard for creating your own control genes:**
  - number of controls
  - importing data formats
  - mapping
  - confirmations and naming

- **Wizard for importing your own experimental gene:**
  - importing data formats
  - mapping
  - confirmations and naming

Back to Top

TUTORIAL (coming soon!)

Opening an existing Analyses

Importing new experimental data to compare with RED-T' default control

Importing new experimental data to be compared with a developed control

Developing a new control

Back to Top

FREQUENTLY ASKED QUESTIONS (FAQ)

Why do I need to register before I use these tools?

If you are a first time user of RED-T and plan to download any of the RED-T tools, we encourage you to register by completing the feedback form. This is not mandatory! This is only for record keeping and notification of any updates. We also welcome your comments and suggestions via this form. Tell us what you think about the RED-T tools, Help files, our web sites, etc.

Back to Top

Where can I get more information on this application?

For more information on RED-T or the RED method, please contact us, for both manuscripts have been submitted and being reviewed by respective publications.

Back to Top

The box-selection (neon-green square) does not select plots I want.

Make sure the highlight mode is not on while choosing the area to select. Since the highlight mode displays all data, including the data not currently being analyzed, it does not see the data that is not active (black in highlight mode). Basically you are selecting empty space. Turn off the highlight mode by un-checking the box below the menu bar.

Back to Top

Why are the plot axes different for different plots?

The plot adjusts to the range of the data provided. It will add one E_d unit to the max value of every axis. To change these initial settings, adjust the axis ranges by using the Zoom feature below the Menu bar.

Back to Top

How can I find out what version of this application I am using?

From the main menu, choose **About...** from the Help Menu item. The application's splash screen will have the version assignment in the bottom left, under "RED-T".

Back to Top

How can I obtain RED source code?

All of the *.class files can be downloaded sepratly from the RED-T package downloads. In the same download menu, click on the link for RED-T_2.1.x_source_code.exe and download this archive file to your machine. Please read the licensing rights before making any changes to the code.

Back to Top

I can download the archive (*.exe) file, but cannot extract (unpack) or open the archive file.

You probably do not have an extracting software, like WinZip or WinRAR, installed on your machine. Go to download.com and download WinZip, WinRAR or another software that can unpack archive files.

If this is not the problem, then you might have downloaded a corrupt RED-T file. Please attempt another download before contacting us via email or the feedback form.

Back to Top

How can I find out what operating system or type and version of browser I am using?

Use the below diagnostic buttons to retrieve information on your browser type/version, operating system or platform, and resolution information of your monitor.

Back to Top

After downloading, when I launch RED.jar, the splash screen comes up only.

First, make sure that the file structure is as instructed by this page. If the same thing happens, this indicates that you do not have JRE version 1.4 or above installed on your machine. Please follow the instructions in software requirements to install the latest JRE on your machine.

Back to Top

I cannot import an image of a tree using the import feature.

If the image is not in GIF or JPEG (JPG) format you will not be able to import the file. In order to transform your image into one of these acceptable formats, use a graphic editor to save as either of these formats.

Back to Top

I get an error every time I am importing a columnar formatted data.

This is a formatting error that you will encounter if the file you are importing is not recognized by the wizard. There are two common errors:

1. Make sure that the first line of the columnar format file begins with "**....RED**". If importing data from older versions, such as the applet, you must add this manually. The "....RED" is followed by the x and y axis labels in this format:

   ....RED(*x-axis label,y-axis label*)

   ....RED(control_default,Pro-tRNA synthetase)

2. The columnar data must be in correct format. Each line should contain information exactly like this:

   organism1,organism2 *tab* x-axis coordinate *tab* y-axis coordinate tab ratio *return-carriage*

   RAA00970,RAB00502 0.906 0.99 1.092

Back to Top

How do I add a new organisms to RED-T' organisms list?

Hopefully you would not have to since we will update the list to reflect publicly available genomes. The only way to do this is to update the *org_list.txt* file. This will be in the same location as the RED.jar file. Follow the delimited format of this text file, beginning with an capitalized abbreviation, genus and species name, and so on.

Back to Top

What are the licensing rights for RED?

RED-T was developed during research at The University of Georgia, Department of Microbiology. All rights reserved. Currently, RED-T is

Unpublished work ©2002 The University of Georgia Research Foundation, Inc.

For more details or updates on the licensing of REd-T, please contact us.

Back to Top

Citing RED-T in literature or other works:

If you wish to cite the use of RED-T in your publications, we suggest the following citation.

- **When referring to RED-T in the main text of your publication, you may choose text similar this format:**

  "...phylogenetic analyses were conducted using RED-T version 2.x (Farahi et al. 2002)."

- **When including a RED-T citation in the Literature Cited/Bibliography section, you may use the following:**

  Farahi K, Whitman WB, and Kraemer ET. **RED-T: application for utilizing the Ratios of Evolutionary Distances for**

determination of alternative phylogenetic events.

Bioinformatics Applications Note. 2002. (*in preparation*).

- **While this manuscript has been submitted and being reviewed,you may cite the following:**

  Kamyar Farahi, William B. Whitman, and Eileen T. Kraemer (2002) **RED-T: application for utilizing the Ratios of Evolutionary Distances for determination of alternative phylogenetic events.** (http://www.arches.uga.edu/~whitman/RED). University of Georgia, Athens, GA, USA. 2002.

Back to Top

How do I report a bug?

Any feedback would be appreciated. Please describe the problem you have experienced in as much detail as possible. Provide as many operational steps as possible that led to the error. Fill out the feedback form, or if you are more comfortable, just email us. In your email, please include the application version. Please enter "Bug report" in the subject field of this email. We will update you with the fix as soon as possible.

Back to Top

What are some features planned for implementation in future versions?

Future versions will include more features to improve the phylogenetic data analysis. These include:

- **Tree generator:** allows the user to generate phylogenetic trees and interactively compare them with the RED plots. This feature allows a user to import a Newick-standard formatted data and generate a tree. This tree will be interactive with the plot. The user can use this feature to compare a phylogenetic history of a gene with the RED plot.
- **Symbolizer:** organismal shape designation for clear identification of displayed data.
- **Mouse over identification mode:** allows individual plot identification to improve identification of the organismal distribution.

Back to Top

How do I suggest a feature for implementation in the future versions?

Any feedback would be appreciated. Please describe the feature you are suggesting and its range of application. Fill out the feedback form, or if you are more comfortable, just email us. In your email, please include the application version you are using. Please enter "Feature suggestion" in the subject field of this email. We will follow up with you as we attempt to implement your idea.

Back to Top

What features are planned for the next RED-T release, v.2.2?

One feature that will be implemented in the next version (v2.2) is to have a choice of symbols for every organism in the plot. This feature will be added to the Organisms List window, next to each RED-id. The user will have a variety symbols to choose, such as open and closed circles, squares and other geometric shapes. This will assist in the analyses of the plot by visually distinguishing the clades or clusters.

Back to Top

CONTACT INFORMATION

Kamyar Farahi - Department of Microbiology, University of Georgia

email: kfarahi@arches.uga.edu

telephone: 706.542.4692

Eileen T. Kraemer - Department of Computer Science, University of Georgia

email: eileen@cs.uga.edu

telephone: 706.542.5799

William B. Whitman - Department of Microbiology, University of Georgia

email: whitman@arches.uga.edu

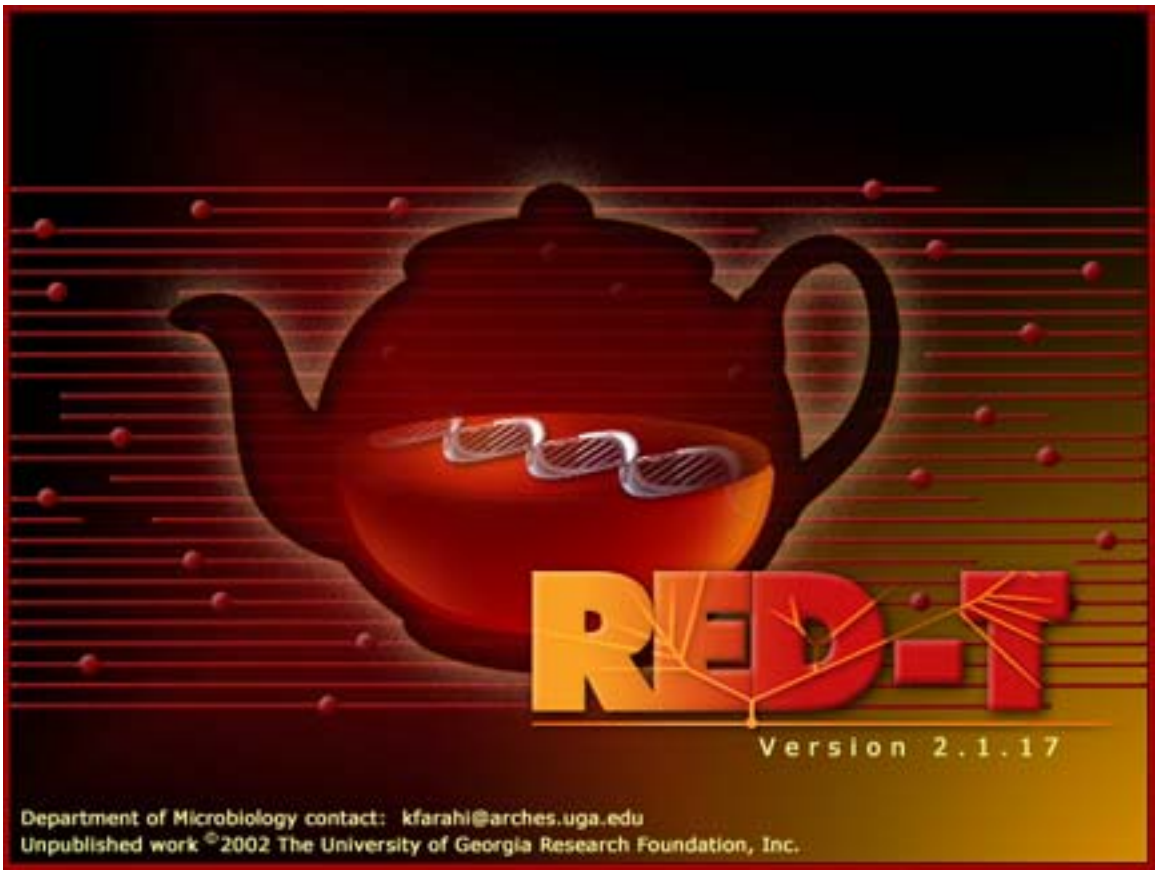
telephone: 706.542.4219

Back to Top

Accessibility of RED-T (phd.html)

RED-T: an application utilizing the Ratios of Evolutionary Distances for determination of alternative phylogenetic events

Welcome to my project web page! We have developed a method, termed RED, to analyze the evolutionary history of individual genes and the potential of lateral gene transfer (LGT). A preliminary description of this method was presented briefly at our poster (ASM 2001).



RED-T is a Java application for phylogenetic analysis that is based on a unique method, RED, that utilizes the ratios of evolutionary distances to distinguish between alternative evolutionary histories. RED-T, Ratios of Evolutionary Distances for determination of alternative phylogenetic events, allows the user to examine if any given experimental gene shares the same evolutionary history as the designated control gene(s). Moreover, the tool detects any differences in evolutionary history, and allows the user to examine comparisons of E_d for a likely explanation. Lateral gene transfer (LGT), which may have a significant influence in organismal evolution especially in prokaryotes evolution, is one mechanism that could explain the findings of these RED-T analyses. RED-T can be downloaded below. In addition to the application, complete interactive access to 60 genes is provided. These genes include many aminoacyl t-RNA synthetases and ribosomal proteins.

Three Java tools are available below. These tools are:

RED-T application package - RED-T application package includes the most recent software and should be sufficient for most investigators.Supplemental ANALYSES package, which contains the 60 gene samples analyzed during our investigation, is also available.

RED-T web-based applet - original version of the tool available for long-term users

The Matrix Integrator - This applet converts matrix formatted data into the columnar format recognized by RED-T (organism1,organism2 *tab* value). This feature is also implemented in the RED-T application.

If you are a first time user, we would appreciate it if you would register before downloading/using any of these tools. This registration is largely for record keeping and notification of any updates, but we also welcome your comments and suggestions.

RED-T application:

Contents of the files	Downloadable files	Approximate size of file
RED-T **application** only	**RED-T_2.1.17.exe**	~609 KB
ANALYSES folder only	**RED-T_Analyses.exe**	~1.825 MB
RED-T application **bundle** including the ANALYSES folder	**RED-T_2.1.17_full.exe**	~2.328 MB
RED-T **source codes** (*.class)	**RED-T_2.1.17_source_code.exe**	~208 KB

This application can be accessed in two different ways. First, the application can be downloaded to your machine. Second, the application can be launched from this web site. If you use this method, only the analyses and not the tool will be saved to your machine. Lastly, the source code is also available for downloading.

This application is created and distributed as a JAR file. It requires JRE 1.4 or later. Visit java.sun.com to obtain the latest JRE.

The complete RED-T application package requires downloading both the RED-T tool and the optional Analyses folder.

Help documentation which includes FAQ, feature descriptions, and more... is available as part of the application. Also, it can be accessed from this web site: click here.

- **Download RED-T application package**

  Download RED-T tool.

  If prompted to 'OPEN' or 'SAVE' the file, choose to 'SAVE'.

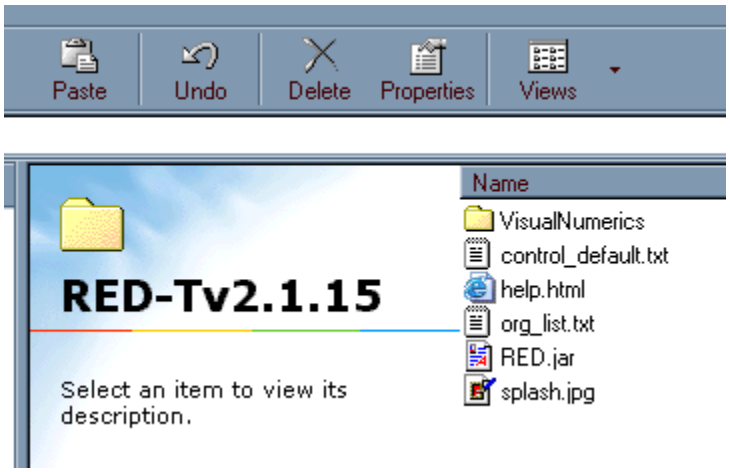
  This is distributed as an self-extracting file (RED-T_2.1.*x*.exe).

  after downloading the latest RED-T package, you must extract it.

  run the *.exe file (Windows: double click on it),

choose a folder on your computer to extract the contents of the package.

this operation maintains the file structure that is essential for



operation of the application, as seen in the example folder content below:

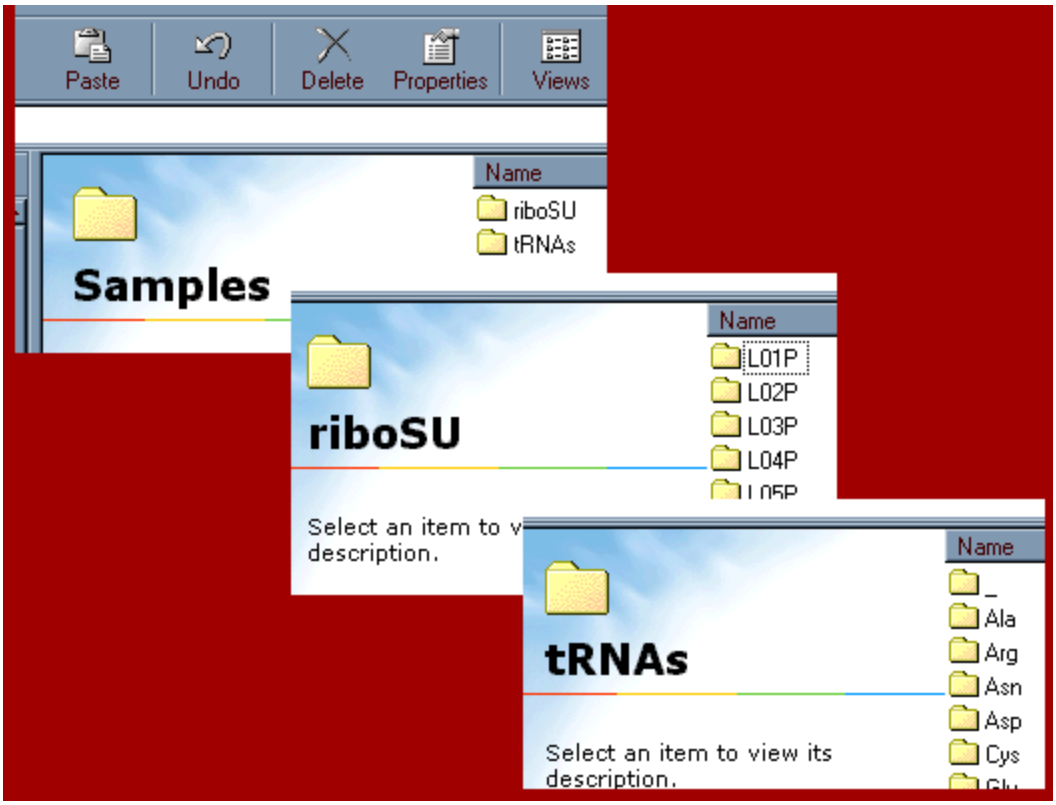
To launch RED-T, double click on the RED.jar file within the folder.

...this package includes the current help files, but **not** the "ANALYSES" folder which contains the 60 genes analyzed during our investigation (in "Samples" folder).

To download ANALYSES folder, link here and save the "Analyses.exe" file.

To extract the ANALYSES folder, run the *.exe file and choose the folder where the RED.jar is for extracting. in the same location as the **RED.jar** file and next to the **VisualNumerics** folder.

Below is the content and file structure of the ANALYSES folder:



Remember that all other files and folders, except the "ANALYSES" folder, which is <u>not essential</u> for the operation of the application, are required in the same directory as the RED.jar for the application to launch.

Also, a larger file, which contains the ANALYSES folder and the RED-T application can be downloaded here. Extract the contents to a folder on your machine and you are ready to run RED-T.

- **Launch it from here! (*to be implemented*)**

  If prompted to 'OPEN' or 'SAVE' the JAR file, choose to 'OPEN'.

  Remember if you use this method, only the analyses and not the tool will be saved to your machine.

- **Download the source code for the latest RED-T version**

  this includes all *.class files.

  extract by running the *.exe file and choose a location on your computer to extract it to.

RED-T applet (web-based form predecessor of the RED-T application)

- **Applet**

  launch the applet from this link only if the location (path) of the data file is known. Otherwise, to use the applet with our data files select one of the links below.

Data files available for use with the applet includes these 69 genes:

- **40 ribosomal proteins**
- **20 aminoacyl-tRNA synthetase**
- **9 aromatic amino acid biosynthetic genes**

The same ribosomal protein and aminoacyl tRNA synthetase genes are also available within the ANALYSES folder of the latest version of the RED-T application, which can be accessed above.

Tips on using this applet

RED's Matrix Integrator applet

This **applet** converts matrix formatted data into the columnar format recognized by RED-T (organism1,organism2 *tab* value). This feature is also implemented in the RED-T application.

CONTACT INFORMATION: [back to top]

Kamyar Farahi

Department of Microbiology, University of Georgia

William B. Whitman

Department of Microbiology, University of Georgia

mailing address: Department of Microbiology, University of Georgia, Athens, GA 30602-2605 (USA)

telephone: 706.542.4692

e-mail: kfarahi@arches.uga.edu

/ or use our feedback form /

Feedback form (feedback.html)

RED-T' FEEDBACK & REGISTRATION

Please *register* here if you are a first time user of RED-T and plan to download any of the RED-T tools. This is only for record keeping and notification of any updates.

We also welcome your comments and suggestions. Tell us what you think about the RED-T tools, Help files, our web sites, etc.

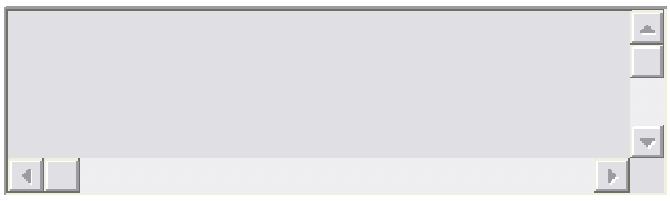
Please select one of the following that best pertains to your interest:

- ◉ Registration (please include what version of the application used)
- ○ Reporting a bug (please include the version number. Instruction on how to obtain the version number, please see the Help files.)
- ○ Make a suggestion (please give reason why this would be useful)
- ○ Other...

...regarding:

Select one... Other:

Enter your comments in the space provided below:



Tell us something about you: (fields with '*' are required to complete this form)

Name*

Organization/affiliation*

Position*

E-mail*

Telephone

Fax

Where did you hear about this work?

Tell us a little about your work. How do you intend to use RED-T?

Please contact me as soon as possible.

Please never contact me!

Submit This | Clear Form

Kamyar Farahi. email kfarahi@arches.uga.edu, telephone: 706.542.4692.

Department of Microbiology, The University of Georgia, Athens, GA, 30602, USA.