

N.U.I. Maynooth, Department of Biology, 2004-2005

BI427 LABORATORY PROJECT

Whole Genome Phylogeny
Reconstruction for the
Genus *Bacillus*



NUI MAYNOOTH

Ollscoil na hÉireann Má Nuad

Bryan O' Sullivan
61714414

Supervisor: Dr. James McInerney
March 2005

This thesis is submitted in fulfillment of the
Computational Biology & Bioinformatics Degree.

DECLARATION:

I TESTIFY THAT THE WORK PRESENTED IN THIS THESIS IS ENTIRELY MY OWN AND THAT WORK WAS CARRIED OUT INDEPENDANTLY UNDER THE SUPERVISION OF DR. JAMES MCINERNEY.

BRYAN O' SULLIVAN.
11TH MARCH 2005.

ACKNOWLEDGMENTS:

Many thanks to my parents Patrick and Freda – without your guidance life would not be as I know it.

Additionally my sincerest thanks to Dr. James McInerney for his constant advice, encouragement and dedication during my undergraduate degree at N.U.I. Maynooth.

FINALLY MANY THANKS TO ALL THE MEMBERS OF THE BIOINFORMATICS & PHARMACOGENOMICS LABORATORY N.U.I. MAYNOOTH.

TABLE OF CONTENTS:

Abstract	01
Glossary of Terms	02

<u>Chapter 1 – Introduction</u>	05 - 12
1.1 The genus <i>Bacillus</i>	06
1.2 Phylogenetic Trees	07
1.3 Phylogenetic Supertrees	08
1.4 Statistical Tests	09
1.4.1 Bootstrapping Phylogenies	10
1.4.2 Bootstrapping Source Trees	10
1.4.3 Permutation Tail Probability tests	11
1.5 Horizontal Gene Transfer	
 <u>Chapter 2 – Materials and Methods</u>	 13 – 16
2.1 Bacterial Dataset	13
2.2 Phylogenetic Tree Construction	13
2.3 Supertree Construction	15
2.4 the ‘Yet Another Permutation Tail Probability’ test	15
2.5 Analysis of the agreement between the supertree and the source trees	16
2.6 Identification of genes involved in Gluconeogenesis	16
 <u>Chapter 3 – Results</u>	 17 – 21
 <u>Chapter 4 – Discussion</u>	 22 – 30
 References	 31 – 33
 Appendix A	 34 – 36

ABSTRACT:

The primary objective of this analysis was to answer the question, “What is the phylogeny of the genus *Bacillus*?” In this study, supertree methods have been used to identify this phylogeny through the identification of orthologous single gene families. A ten taxon dataset

was used for this analysis, consisting of 6 genus *Bacillus* members and 4 out-group species. Several statistical tests including bootstrap analyses and permutation tail probability tests (and variants), have been used by this analysis to infer confidence and support in the results presented here. The supertree presented here has high bootstrap support and logically classifies all genus *Bacillus* members into the same monophyletic clade. Analysis of the glycolysis pathways for all species used in this study, has aided in the identification of species-specific gene that support the relationships presented here.

GLOSSARY OF TERMS:

Branch	defines the relationship between the taxa in terms of descent and ancestry.
Congruence	Agreement between characters or trees.

Conjugation	The process of transferring a certain plasmid of DNA known as the F plasmid (or sex plasmid) from bacteria individuals who have it (known as "males") to bacteria individuals who do not already have it (known as "females") by way of direct contact between the bacteria individuals called a conjugation bridge. Once transfer is completed, the female individual becomes a male individual and both parties have a copy of the F plasmid.
Incongruence	also known as incompatibility. This refers to different characters or trees suggesting different groups of relationships. Incongruence is due to the presence of homology.
Leaves	The contemporary taxa in a dataset (as opposed to the reconstructed ancestral taxa). Also known as terminal tips.
Leaves	Represent sequences or organisms for which we have data (a.k.a. Terminal units, Operational taxonomic units).
Monophyletic group	A group of organisms with the same taxonomic title (say, genus, family, phylum etc.) that are shown phylogenetically to share a common ancestor that is exclusive to these organisms.
Node	a node represents a taxonomic unit. This can be a taxon (an existing species) or an ancestor(unknown species: represents the ancestor of 2 or more species)
Ortholog	These are homologous genes that are found in two different taxa and are performing the same function in each taxon. Together with paralogs they comprise an homologous superfamily.

Polytomy	A branch point on a tree that has more than two immediate descendents.
PTP	Permutation Tail Probability test. This is a matrix randomisation test. The columns of a matrix are randomised so that the consensus sequence and the composition for any particular column is maintained, but any signal is lost. Any measure of a tree (say tree length) should be much worse from the permuted data than from the original data <i>except</i> when the original data is no better than random in the first place.
Root	is a common ancestor of all taxa.
Systematic error (bias)	This is a facet of a dataset that will confound any tree reconstruction method. Situations such as long branch attraction and base-compositional bias are examples of systematic bias.
Transduction	The movement of genes from a bacterial donor to a bacterial recipient using a phage as the vector. A process whereby a cell can gain access to and incorporate foreign DNA brought in by a viral particle.
Transformation	The process of transferring a certain plasmid of DNA known as the F plasmid (or sex plasmid) from bacteria individuals who have it (known as "males") to bacteria individuals who do not already have it (known as "females") by way of direct contact between the bacteria individuals called a conjugation bridge. Once transfer is completed, the female individual becomes a male individual and both parties have a copy of the F plasmid.

Many thank to Dr. James McInerney for permitting of use of this glossary.

CHAPTER 01 – INTRODUCTION:

In the last decade, advances in molecular biology techniques have facilitated the sequencing of various whole bacterial genomes. To date, 10 genomes of the genus *Bacillus* have been annotated and sequenced (Bernal et al., 2001). In compiling a dataset for this analysis, 6 sequenced members of the genus *Bacillus* were chosen, together with 4 out-group bacterial species from the Phylum *Firmicutes*. A full list of the bacterial species used in the analysis can be seen in Table 1.

Supertree methods have been employed by this phylogeny reconstruction analysis. Supertree methods are more reliable when compared to other phylogeny reconstruction methods such as the partially overlapping dataset method (Daubin et al., 2001). The partially overlapping dataset method is based on the concatenation of sequences from different genes, and the construction of a phylogenetic tree based on the concatenation of such sequences. (Daubin et al., 2001). The problem of such methods is that they do not consider the different rates of evolution among genes. For example, if two genes are informative at different levels of the phylogeny, then their concatenation will attenuate this information rather than amplify it (Daubin et al., 2001). Supertrees on the other hand, reconstruct each gene tree separately and assume different models of evolution for each. Moreover, supertree methods allow the use of the largest amount of genome data with the least amount of assumptions. Partially overlapping dataset methods require assumptions regarding to missing data, whereas supertree do not make any allowances for missing data. Finally, supertree methods allow for the analysis and identification of laterally (horizontally) transferred genes.

The overall objective of this analysis has been to address the question, “Is there a phylogeny inside the genus *Bacillus*?” Previous studies have revealed that the identification of tree-like phylogenies for deep-level prokaryotic relationships are difficult to infer (Creevey

et al., 2004). Thus, this analysis aims to identify and show the phylogenetic relationships of bacteria (i.e. members of the genus *Bacillus*) at the tips of the prokaryotic phylogeny.

SECTION 1.1 THE GENUS *BACILLUS*

In 1872, Ferdinand Cohn recognised and named the bacterium *Bacillus subtilis* (Black, 2002). The organism represented a large and diverse genus of bacteria, *Bacillus*, and was placed in the family *Bacillaceae*. The family's distinguishing feature is the production of endospores, which are highly refractile resting structures formed within bacterial cells (Black, 2002). *Bacillus* is distinguished from other endospore-forming bacteria on the basis of being a strict or facultative aerobe, rod-shaped, and (usually) catalase-positive (Buchanan et al., 1974). Other endospore-forming genera in the family are: *Sporolactobacillus*, which is microaerophilic and catalase-negative; *Clostridium*, which is anaerobic and does not reduce sulphate; *Desulfotomaculum*, which is anaerobic but does reduce sulphate; *Sporosarcina*, which has a coccal morphology; and *Thermoactinomyces*, which while forming endospores, displays typical actinomycete characteristics. These genera appear to be closely related phenotypically as Gram-positive bacteria that form endospores however, genotypic analysis has revealed that these genera are more distantly related phylogenetically (Black, 2002).

Early attempts at the classification of *Bacillus* species were based on two characteristics: aerobic growth and endospore formation. These attempts resulted in grouping together of many bacteria possessing different physiological and genetic traits. However, since the advent of sequencing technologies, phylogenetic classification of bacterial species has been made possible. Here we use supertree methods to create a phylogenetic classification of *Bacillus* species, and further use this information to identify essential genes involved in the core of this supertree.

SECTION 1.2 PHYLOGENETIC TREES

Phylogenetics is the study of evolutionary relationships among organisms or genes. In order to illustrate the evolutionary relationships among a group of organisms, we construct phylogenetic trees. The main purpose of phylogenetic studies is to reconstruct evolutionary relationships between organisms and to estimate the time of divergence between organisms since they last shared a common ancestor. Several types of data can be used to construct phylogenetic trees. Until some 25 years ago, relationships among organisms were primarily based on morphological character data, for example presence or absence of vertebrae or beak shape. Nowadays, molecular data such as DNA sequences and protein sequences are used to construction these evolutionary relationships (Hall, 2004).

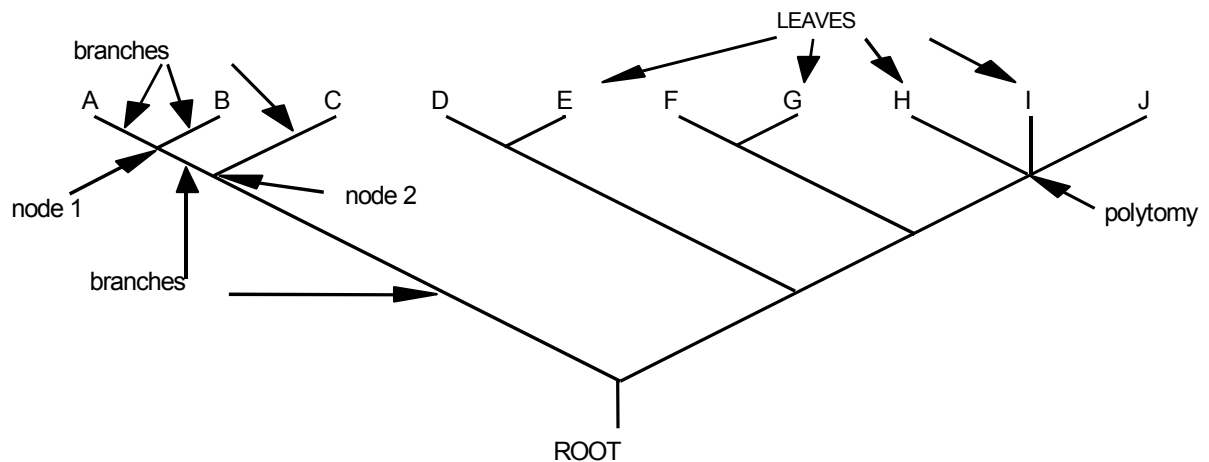


Figure 1. A hypothetical phylogenetic tree representing organisms A – J. An explanation of the terms associated with this figure (and others) can be found on page 2.

SECTION 1.3 PHYLOGENETIC SUPERTREES

Although the term ‘supertree’ was only coined and formalised in 1986 (Bininda-Emonds, 2004), the concept behind supertrees existed informally long before that time (Bininda-Emonds, 2004). A supertree aims to reconstruct the correct evolutionary relationship between groups of species. Supertree construction takes a phylogenetic approach in which many overlapping source trees, rather than the character data used to derive those trees, are combined to produce a single, larger supertree (Bininda-Emonds, 2004).

At present, several methods of constructing supertrees exist. These methods can be categorized into ‘agreement’ supertrees, i.e. those methods indicating common or uncontested groupings among the set of source trees, or ‘optimisation’ supertrees, i.e. those methods yielding the supertree(s) that has the maximum fit to the set of source trees according to some objective function (Bininda-Emonds, 2004). To date there are a minimum 16 methods and variants thereof, including semi-strict and strict methods typical of agreement supertrees, and Bayesian and Matrix representation using parsimony (MRP) methods typical of optimisation supertrees (Bininda-Emonds, 2004).

In this analysis, Clann 2.0.1 (Creevey et al., 2005) was used for the construction of our supertree. Clann uses a “Most Similar Supertree Analysis” (MSSA) method for its supertree construction, a method that can be placed in the supertree optimisation category. A most similar supertree analysis proposes a supertree containing all the leaves found in the source (gene) trees. Taking each source tree in turn, the supertree was pruned until both trees possessed the same leaf set. A simple tree-to-tree distance was calculated to evaluate similarity between the pruned supertree and the source trees. For every pair of leaves, we counted the number of nodes that separated them on each tree and calculated the absolute

difference. The sum of these pairwise distances gives the dissimilarity of the trees. To normalise for large tree bias (Purvis, 1995) the sum was divided by the total number of comparisons. In this way a proposed supertree was given a score of zero if, for all gene trees, its sub-tree on the gene-trees leaf set was identical to the gene tree. Higher scores are indicative of increasing dissimilarity. This scoring system was used as an optimality criterion for choosing between alternative supertrees. During the analysis of the dataset, heuristic searches of the tree-space were carried out to find the supertree with the minimum difference score when compared to the source trees. Sub-tree Pruning and Regrafting (SPR) as described in Paup* (Swofford, 2002) were used to traverse supertree space. (Creevey et al., 2004, Creevey et al., 2005)

SECTION 1.4 STATISTICAL TESTS

In order to evaluate the statistical significance of any particular branching pattern, statistical tests had to be performed. Statistical tests in phylogenetics permit the assessment of the degree of confidence we have in any given topology being the true topology (Felsenstein, 2004). Thus, statistical tests are responsible for the mutual development between our abilities to estimate better trees and to create more realistic models of evolution (Felsenstein, 2004). Bootstrapping (a statistical test) is one of the most popular tests, which relies heavily on resampling methods (Efron et al., 1996). Resampling methods (e.g. the bootstrap and the jackknife (not used in this analysis)) resample data or infer empirically the variability of the estimate obtained by a tree making method. Another popular statistical test used in phylogenetic analysis is the Permutation Tail Probability (PTP) test, which is designed to address whether there is any true phylogenetic signal in any given dataset (alignment) (Felsenstein, 2004).

SECTION 1.4.1 BOOTSTRAPPING PHYLOGENIES

Bootstrapping (Efron et al., 1996) is one of the most popular ways of testing the reliability of a dataset. It requires the creation of pseudoreplicate datasets by resampling. Bootstrapping allows for the assessment of whether the distribution of characters has been influenced by stochastic effects. In phylogenetic analysis, nonparametric bootstrapping is the most commonly used method (Felsenstein, 1993). The pseudoreplicate datasets are generated by randomly sampling the original character matrix to create new matrices of the same size as the original. The frequency with which a given branch is found is recorded as the bootstrap proportion. These proportions can be used as a measure of the reliability (within limitations) of individual branches in the optimal tree (Creevey et al., 2004, Felsenstein, 2004). Majority rule consensus methods assign bootstrap proportion to individual branches. Bootstrapping was applied to this analysis through the use of the SEQBOOT (Felsenstein, 1993) bootstrapping tool.

SECTION 1.4.2 BOOTSTRAPPING SOURCE TREES

In order to assess the support for internal branches on our supertree, a bootstrap analysis was performed. The same theoretical methods are applied to bootstrapping gene trees as bootstrapping alignments. Individual gene trees were resampled with replacement, until a new dataset was created with the same number of gene trees as the original (Creevey et al., 2004). A heuristic search of tree space was implemented for each pseudoreplicate and the results, reported here as bootstrap proportions (BP), were summarised using a majority rule consensus tree (Creevey et al., 2004). In this analysis, the Clann 2.0.1 supertree software was used to implement our bootstrap analysis.

SECTION 1.4.3 PERMUTATION TAIL PROBABILITY TESTS

PTP tests are based on the principle of matrix randomisation and have thus rather suitably earned themselves the name of matrix randomisation tests. The columns of the matrix are randomised resulting in the loss of any signal, while maintaining the consensus sequence and the composition for any particular column (Felsenstein, 2004). Any measure of a tree (e.g. Tree length) from the permuted should be significantly worse data than from the original data, with the exception of when the original data is no better than random.

SECTION 1.5 HORIZONTAL GENE TRANSFER

The foundation of Charles Darwin's theory of evolution is the vertical inheritance of traits from parent to offspring across successive generation (Brown, 2003). However it has recently become apparent to molecular biologists that many genes from both eukaryotes and prokaryotes have been acquired by horizontal gene transfer (HGT) (Jain et al., 1999). HGT refers to the horizontal exchange of genes between distantly related strains or even species. HGT can occur by one of three mechanisms in bacteria namely; transduction, conjugation and transformation. An explanation of these terms can be found on page 2.

Table 1. Bacterial species included in this analysis

Species	Total size, Mb	Genbank(G) / TIGR(T) accession no.
<i>Bacillus cereus</i>	5.2	(T) b_cereus_10987
<i>Bacillus halodurans</i>	4.2	(G) BA000004
<i>Bacillus subtilis</i>	4.21	(G) AL009126
<i>Bacillus licheniformis</i>	4.22	(G) CP000002
<i>Bacillus thuringiensis</i>	4.7	(G) NC_005957
<i>Bacillus anthracis</i>	5.23	(T) b_anthraxis_ames
<i>Clostridium acetobutylicum</i>	4.13	(G) NC_003030
<i>Lactococcus lactis</i>	2.36	(G) NC_002662
<i>Enterococcus faecalis</i>	3.21	(G) NC_004668
<i>Mycoplasma mycoides</i>	1.21	(G) NC_005364

CHAPTER 02 - MATERIALS & METHODS:

Section 2.1 Bacterial Dataset

The dataset used in this analysis contained 10 fully sequenced bacterial genomes: *Bacillus halodurans*; *Bacillus subtilis*; *Bacillus licheniformis*; *Bacillus thuringiensis*; *Clostridium acetobutylicum*; *Lactococcus lactis*; *Enterococcus faecalis*; *Mycoplasma mycoides* (all of which were obtained from <http://www.ncbi.nlm.nih.gov/>); *Bacillus cereus* and *Bacillus anthracis* (both of which were obtained from <http://www.tigr.org/>). The sizes and Genbank/TIGR accession numbers of the bacterial genomes used in this analysis are given in Table 1.

Section 2.2 Phylogenetic Tree Construction

Using the translfas program (McInerney, 1998), our nucleotide dataset was translated into an amino acid dataset. Homologous sequences were identified by carrying out “all against all” searches of a database using the BLASTP algorithm (Altschul et al., 1997), with an E-value cut off of 10^{-7} . Only those homologous single-gene families where every member found every other member (and nothing else) were retained. The resultant single-gene families were used to construct phylogenetic gene trees that contained a minimum of 4 members. This conservative approach has been designed to exclude the inadvertent analysis of paralogues. Each of these single-gene families were then aligned using ClustalW 1.8 (Thompson et al, 1994) using the default parameter settings, and manually verified.

Using readseq (Gilbert, 1990), the verified ClustalW alignments were converted from phylip format (.phy) to nexus format (.nexus), and a paupblock was attached to each .nexus file. These files were then analysed using a Permutation Tail Probability test (PTP), which is

a function available in the Phylogenetic Analysis Using Parsimony (Swofford, 2002) software program (PAUP* 4.0b10). A PTP test is a matrix randomisation test that tests the overall phylogenetic signal in the data. It evaluates whether a candidate character state could have occurred by chance alone. For example, if data were essentially random. A PTP test generates random datasets (based on the original) and calculates of the number of times a tree with a fit as good as the original may be found by random chance. Only PTP results with a PTP score of $P = 0.01$ were accepted for further analysis. A PTP score of $P = 0.01$ gives 99% confidence that a result is better than random.

To assess the variation in the data, 100 bootstrap replicates were generated from the dataset, using the SEQBOOT program (operating with default settings) from the PHYLIP 3.6 (Felsenstein, 1993) package. Distance matrices were produced using PROTDIST (a program used to compute a distance matrix from protein sequences, where the distance for each pair of species estimates the total branch length between the two species). The resulting matrices were used as input to the program NEIGHBOR (using the neighbour-joining algorithm) which constructed neighbour-joining source trees. In each of these cases default settings were used in the execution of the programs. The resulting neighbour-joining source (gene) trees were inputted into CONSENSE to form a single majority rule consensus gene tree. These gene trees from CONSENSE were used to construct the supertree. PROTDIST, NEIGHBOR, and CONSENSE are all programs available in the PHYLIP 3.6 package.

Section 2.3 Supertree Construction

Using the default settings of Clann 2.0.1 (Creevey at al., 2005), a MSSA was carried out on the generated source (gene) trees. This supertree method takes as input a set of source trees

and seeks to identify the most similar supertree to the entire set of source trees as described by Creevey and McInerney (Creevey et al., 2004).

In order to assess the support for internal branches on a supertree, a bootstrap analysis was carried out. The individual input trees were resampled with replacement, until a new dataset was created with the same number of input trees as the original. An exhaustive search of tree space was carried out for each pseudoreplicate and the results of the bootstrap analysis were summarised using a majority-rule consensus tree of 100 replicates.

Section 2.4 the 'Yet Another Permutation Tail Probability' test

To test the null hypothesis that the phylogenetic signal in the gene trees was no better than random, a test known as, 'Yet Another Permutation Tail Probability' (YAPTP) test was performed (Creevey et al., 2004). For each gene tree, we removed the taxon names and randomly reassigned them to the leaves. This removed any congruent phylogenetic signal between gene trees, while leaving the numbers, sizes and shapes of the gene trees, the frequency with which any particular taxon was found across the gene tree, and the frequency of co-occurrence of any group of taxa within gene trees unaltered. A search of tree space was then carried out and the score of the best supertree was recorded. This was repeated 100 times. The null hypothesis, that the gene trees contain no more phylogenetic signal than would be expected by chance alone, was rejected if the score for the real gene tree was not bettered than all of the 100 sets of randomly permuted gene trees (Creevey et al., 2004, Philip et al., 2005)

Section 2.5 Analysis of the agreement between the supertree and the source trees

In order to assess the agreement between our source (gene) trees and our optimal phylogenetic supertree, Clann (Creevey et al., 2005) was used to score each source tree against the supertree. The importance of this step is to identify the 'core' source trees that

form our supertree. Results with a score of zero are in complete agreement with the supertree. Source trees with scores that tend away from zero, are less compatible with the supertree.

Section 2.6 Identification of genes involved in Gluconeogenesis

Using the Kyoto Encyclopaedia of Genes and Genomes (<http://www.genome.jp/kegg/>), with Gluconeogenesis (Glycolysis) as a reference pathway, species-specific genes were manually identified. The results of this analysis aided in the verification and validation of our supertree.

CHAPTER 03 - RESULTS:

For the 10 genomes studied, 877 single-gene families with four or more taxa were identified using the *Teaghlach.pl* perl program (Appendix A). These single-gene families were then aligned using ClustalW, and verified and validated by eye using *seaview* (Galtier et al., 1996). In order to establish the presence of true phylogenetic signal within our dataset (alignments), a Permutation Tail Probability (PTP) test was executed. Of the 877 single-gene families identified in this analysis, 742 passed the PTP test with a score of $P = 0.01$. Only those single-genes families, which passed the PTP test, were further used for analysis in this study.

A bootstrap analysis of the dataset was performed using *Seqboot*. *Seqboot* is a general bootstrapping tool. *Seqboot* reads in the dataset as its input (in this case the set of 742 single-gene families), and produces multiple pseudoreplicate datasets from the input dataset through bootstrap resampling. The output from the *Seqboot* bootstrapping tool acts as input for the *Protdist* distance matrix construction program. The output from *Protdist*, distance matrices, subsequently may act as input for other programs such as *Neighbor*. *Neighbor* implements the neighbour-joining method of Nei and Saitou (Saitou et al., 1987) and constructs a tree by successively clustering lineages, setting branch lengths as the lineages join. The neighbour-joining trees generated by *Neighbor* subsequently analysed by *Consense* to form a single consensus gene tree. In total, *Consense* formed 742 genes tree. These gene trees were used to construct a phylogenetic supertree using *Clann 2.0.1* (Figure 2).

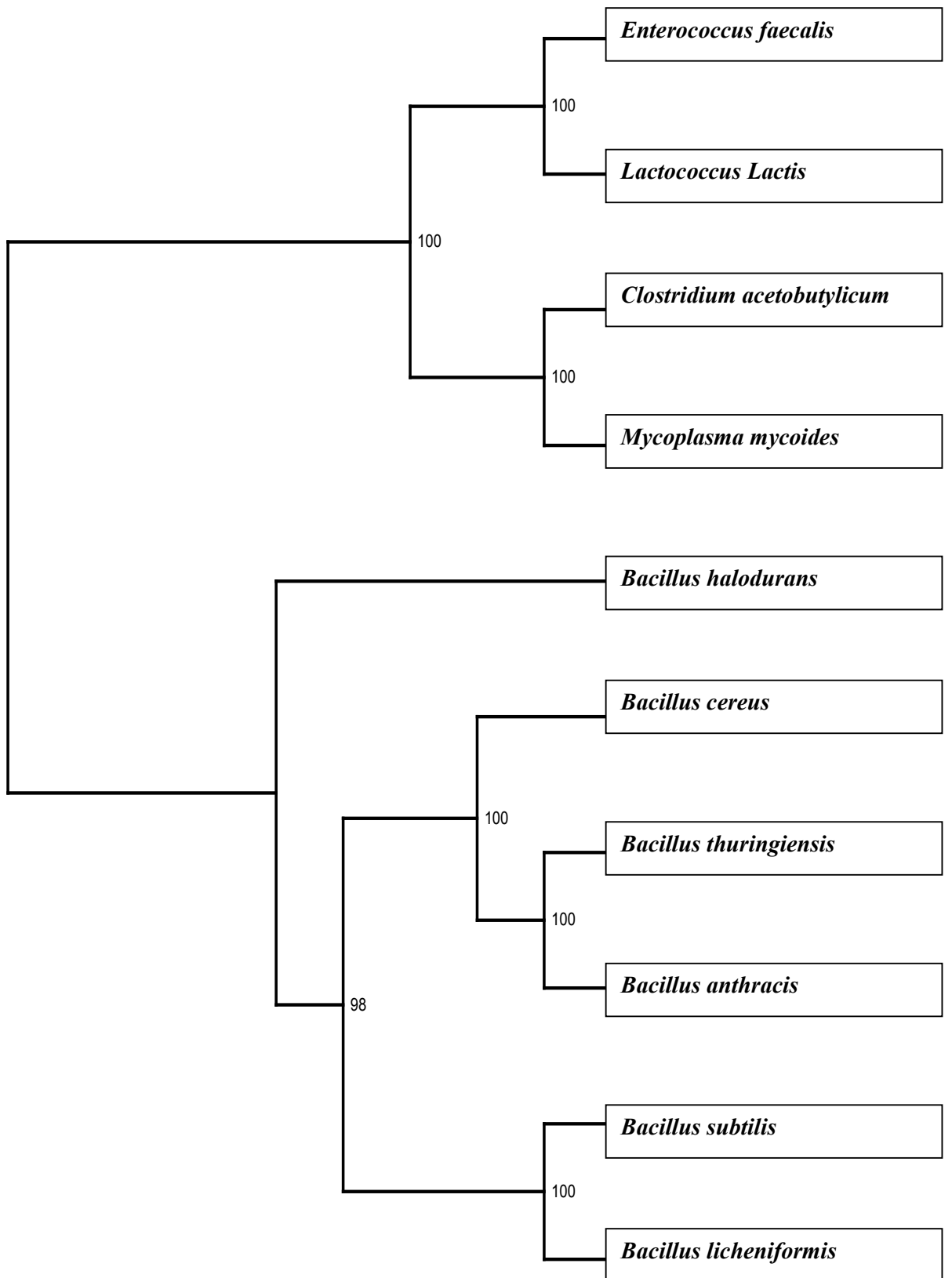


Figure 2. Optimal phylogenetic supertree from this analysis. Numbers represent the percentage BP support received by the internal branch

The unrooted phylogenetic supertree shown in figure 2 is the single optimal supertree. In order to assess the phylogenetic support and signal of the supertree, bootstrap and YAPTP test were performed respectively. As can be seen in figure 2, the numbers at the internal branches of the tree represent strong Bootstrap Proportions (BP). The high BPs represent strong bootstrap support for the branches of our supertree. The results from the YAPTP test can be seen in figure 2. The distribution of scores for the 100 best trees from the YAPTP test is centred on 670 (± 5), whereas the score obtained by our supertree (323.467224) was outside the range of scores obtained by the randomly generated trees. These results illustrate that our supertree is clearly better than randomly generated trees (Figure 3).

Results from scoring of each individual source (gene) tree against the optimal phylogenetic supertree can be seen in figure 4. As illustrated by figure 4, 38% of source trees are completely agreeable with our phylogenetic supertree, thus lending further support for the topology of our supertree.

Distribution of YAPT scores

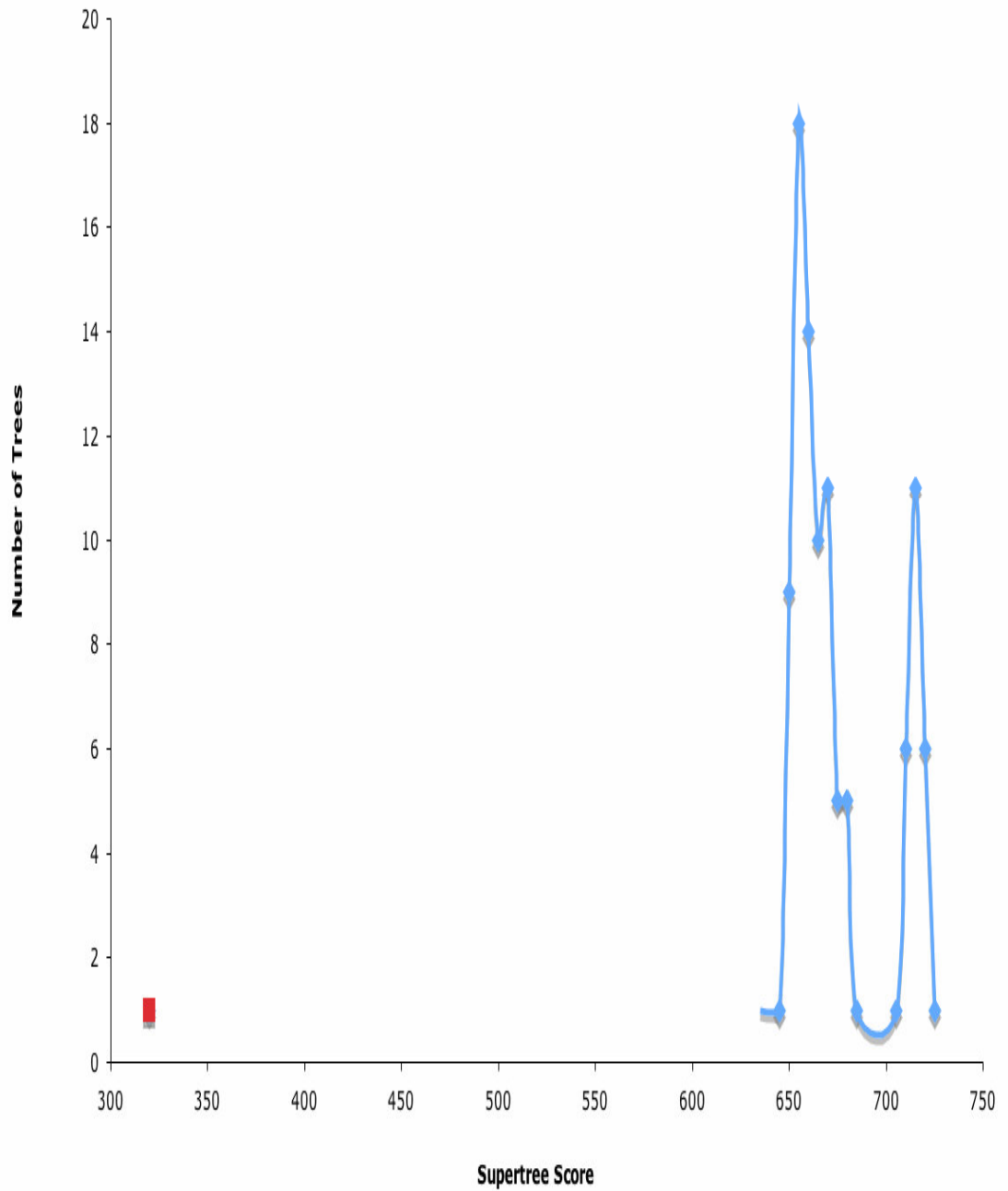


Figure 3. The results of the YAPT test, showing the score of supertree (in red) and the 100 best randomly generated trees from the YAPT analysis in blue.

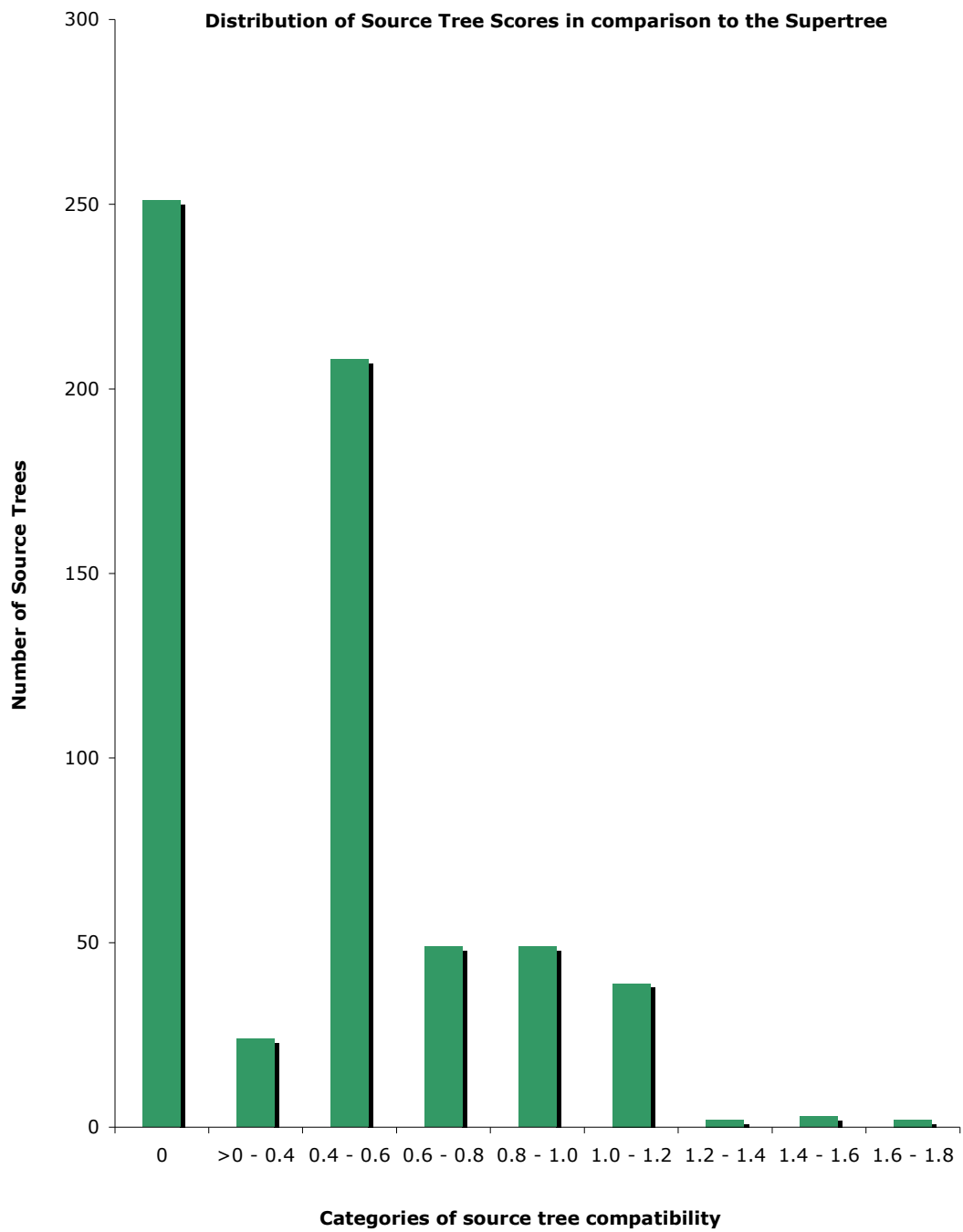


Figure 4. Scores resulting from scoring our source (gene) trees against the optimal phylogenetic supertree. 34% of source trees are completely compatible with the supertree.

DISCUSSION:

For this analysis, computational identification and phylogenetic reconstruction of single-gene families from ten bacterial genomes, resulted in the construction of a phylogenetic supertree (Figure 2). The supertree presented in this analysis logically classifies the six members of the genus *Bacillus* as being distinctly related, and also classifies the four out-group genomes as being more closely related to each other than to the genus *Bacillus* members. A large volume of morphological, biochemical, statistical and phylogenetic data exists, which support the supertree that is presented in this analysis.

For this analysis, scores from the YAPTP test indicate that the optimal supertree is very dissimilar to the randomised genes trees (generated by the YAPTP test). In addition to this, the scores obtained when the gene trees were compared to the optimal supertree, illustrate that there is strong agreement between the gene trees and the optimal supertree. These results indicate that the gene trees show a high level of congruence and extremely consistent signals. Furthermore, almost all of the relationships in this study (for example between *Bacillus thuringiensis* and *Bacillus anthracis*) receive 100% BP support. These strong BP supports demonstrate high levels of support for the internal branches on the supertree, and thus provide additional data in support of our supertree (Figure 2).

In addition to the strong statistical support, useful morphological and phylogenetic data also exists in agreement with the phylogenetic supertree. Logically, all *Bacillus* species are grouped together into a monophyletic clade on the supertree. According to the supertree presented here, *Bacillus subtilis* and *Bacillus lichenformis* are each others closest relative and thought to have evolved from a common ancestral bacterium. Genome comparisons between

B. licheniformis and *B. subtilis* have revealed that 66% of *B. licheniformis* genes have orthologs in *B. subtilis* (Rey et al., 2004), confirming the accuracy of the relationship presented in the supertree. In addition to this genomic data, *B. licheniformis* and *B. subtilis* possess many of the same characteristics and features, in that they are both rod shaped, gram-positive, catalase positive, spore forming, motile bacteria. These common characteristics lend even further support to the accuracy of this relationship. Despite the remarkable similarities between the *B. licheniformis* and *B. subtilis* genomes, there are notable differences in the number and location of prophages, transposable elements and a number of extracellular enzymes and secondary metabolic pathway operons that distinguish these species (Rey et al., 2004)

The next clade identified in the supertree includes *Bacillus cereus*, *Bacillus thuringiensis*, and *Bacillus anthracis*. As illustrated on the supertree presented in figure 2, *Bacillus thuringiensis* and *Bacillus anthracis* are viewed as being more closely related to each other (indicating a recent evolutionary origin) than they are to *Bacillus cereus*, consistent with other studies concerning these bacterial species (Ivanova et al., 2003, Parkhill et al., 2003, Helgason et al., 2000). Both *Bacillus thuringiensis* and *Bacillus anthracis* (animal and human pathogens respectively) share characteristics common to the genus *Bacillus*, and it is known that the genomes of both species are highly conserved and that homologous genes can be found in both. For example *Bacillus anthracis* is known to contain two homologues of the *Bacillus thuringiensis* immune inhibitor A metalloprotease (BA0672 and BA1295), that enhances virulence in insects through cleavage of bacteriocidal lectin (Read et al., 2003). *Bacillus cereus* is also known to share genome similarity with both *B. anthracis* and *B. thuringiensis* and numerous homologous genes have been identified in all three genomes including the channel forming type III haemolysins (BA5701, BA2241) and a complex of three non-haemolytic enterotoxins (BA1887-1889) (Read et al., 2003). However

despite sharing many common characteristics *B. cereus* is distinguished from *B. anthracis* and *B. thuringiensis* by the presence of plasmid-borne specific toxins (*B. anthracis* and *B. thuringiensis*) and a poly-D-glutamyly capsule, which mediates the invasive stage of the infection (*B. anthracis*) (Ivanova et al., 2003).

The final *Bacillus* species used in this analysis, *Bacillus halodurans*, occupies a position outside of each of the above discussed clades and appears to be one of the earliest evolving member of the genus *Bacillus* to be represented on this tree. *B. halodurans* is classified as an alkaliphilic bacterium meaning that it cannot grow or grows poorly under neutral pH conditions, but grows well at pH >9.5 and requires Na⁺ for growth. Analysis of the genome of *Bacillus halodurans* has revealed 4,066 protein coding sequences, 2,141 of which have functional assignments, 1,182 are conserved protein coding sequences with unknown function and 743 have no match to any protein in any protein database (Takami et al., 2000). Sequence comparisons between *B. halodurans* and *B. subtilis* have revealed that among the total protein coding sequences of *B. halodurans*, at total of 8.8% match sequences of proteins found only in *Bacillus subtilis* (Takami et al., 2000). Furthermore, the *B. halodurans* chromosome is known to contain large regions that are collinear with the genomes of *B. licheniformis* and *B. subtilis* (Takami et al., 2000). The unique protein sequence similarity between *B. halodurans* and *B. subtilis* and the matching regions of chromosomal collinearity between *B. halodurans* and *B. licheniformis* and *B. subtilis*, helps to explain in part why *B. licheniformis* and *B. subtilis* are seen to have diverged from *B. halodurans*. These observations correlate well with and give further confidence and support to our supertree.

All the *Bacillus* species were logically placed together in the same clade on our supertree. Similarly, the out-group species were identified to be more closely related to each other than

to any of the *Bacillus* species in this analysis. As can be seen from figure 2. *Enterococcus faecalis* and *Lactococcus lactis* are shown to be more closely related to each other than they are to *Clostridium acetobutylicum* and *Mycoplasma mycoides*, and *vice versa*. The relationship between *E. faecalis* and *L. lactis* is an obvious and accurate relationship. Both species are member of the order *Lactobacillales*, although from different families. *E. faecalis* and *L. lactis* are both non-motile, gram-positive cocci, have a fermentative metabolism and are catalase negative. They both inhabit similar habitats in animals, namely the gastrointestinal tract (Paulsen et al., 2003, Bolotin et al., 2001). These bacteria can thus withstand low pH environments and also have resistance to bile acids. Homologous genes have been identified in both species, for example the *ycdB* gene which is present as a single copy in *L. lactis* and as a double copy in *E. faecalis* (Bolotin et al., 2001).

The relationship between *Mycoplasma mycoides* and *Clostridium acetobutylicum* as described by our supertree is less obvious. Both bacteria are members of the Phylum *Firmicutes* (low-GC gram-positive bacteria), however they do not belong to the same genus. The underlying basis for this relationship is most probably because these two species are the two most dissimilar genomes to every other genome used in this analysis, but because *M. mycoides* and *C. acetobutylicum* are both members of the Phylum *Firmicutes*, their genomes will thus contain genes that are common to all members of this phylum.

Finally, definitive biochemical data also exists in support of the supertree. Through the analysis of the Kyoto Encyclopaedia of Genes and Genomes (<http://www.genome.jp/kegg/>), we have identified several species-specific genes involved in the Glycolysis pathway that lend additional support to our supertree (Figure 5, Figure 6). The grouping of all *Bacillus* species together on our supertree, has fundamental biochemical support. Analysing the glycolysis pathway for all ten bacterial species has revealed that the genes encoding Acetate

CoA ligase and CoA independent aldehyde dehydrogenase are a unique feature of the genus *Bacillus* (based on this data). Based on the wide diversity of species used, it appears that these genes arose at the same time as the formation of the genus *Bacillus* (based on the data presented in this analysis) (Figure 5).

Additionally, there is biochemical support for the sub-clades within the Bacillus clade. For example figure 6 illustrates the relationship between *Bacillus subtilis* and *Bacillus licheniformis*. This relationship is distinctive in that these two species are the only species in this analysis that encode alcohol dehydrogenase (NADP+). This information thus provides added strength and accuracy to the relationship presented between *B. subtilis* and *B. thuringiensis* (Figure 6).

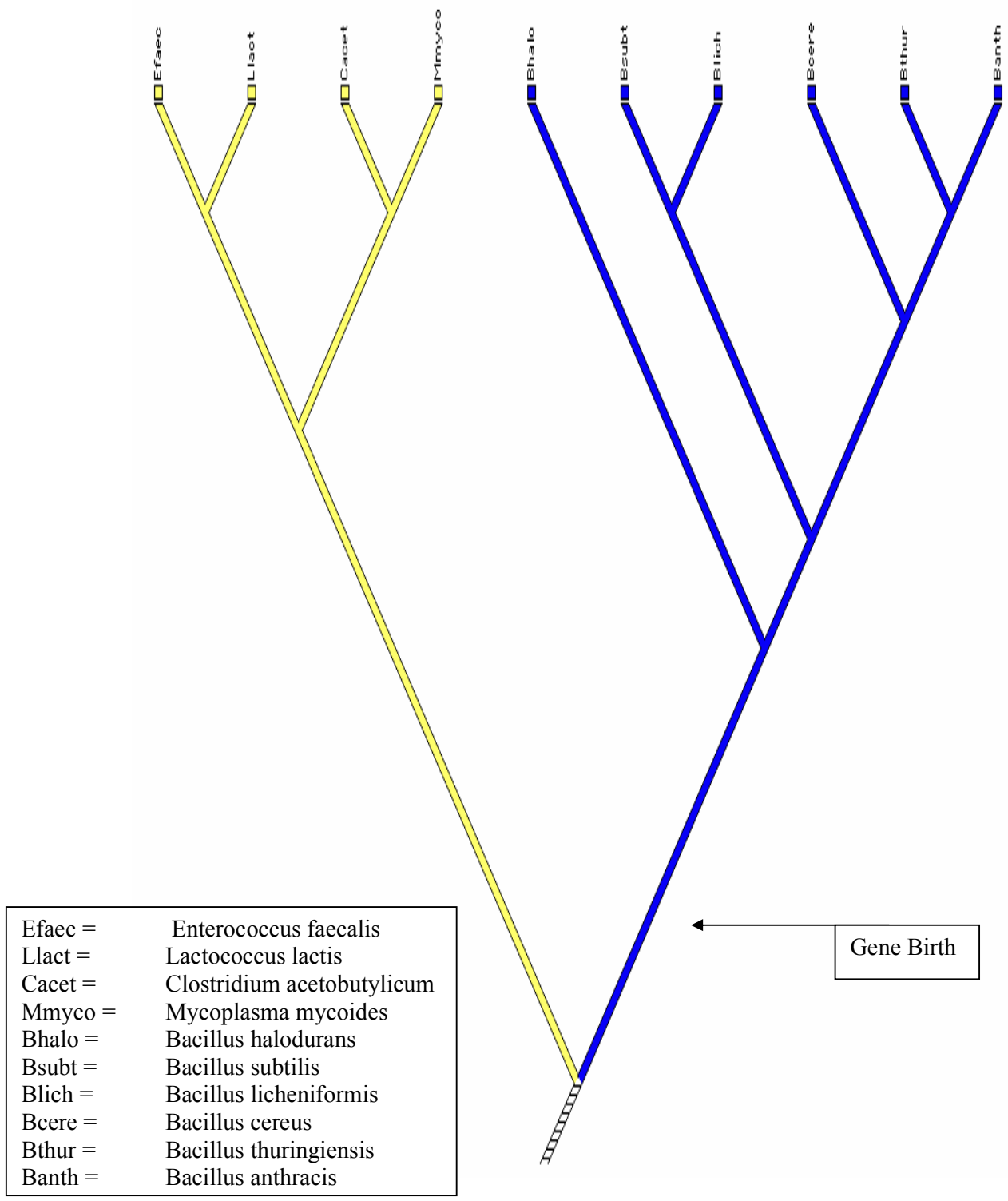


Figure 5. A phylogenetic tree representing the presence (Blue branches) and absence (Yellow branches) of the genes encoding Acetate CoA ligase and CoA independent aldehyde dehydrogenase in the bacterial species used in this analysis.

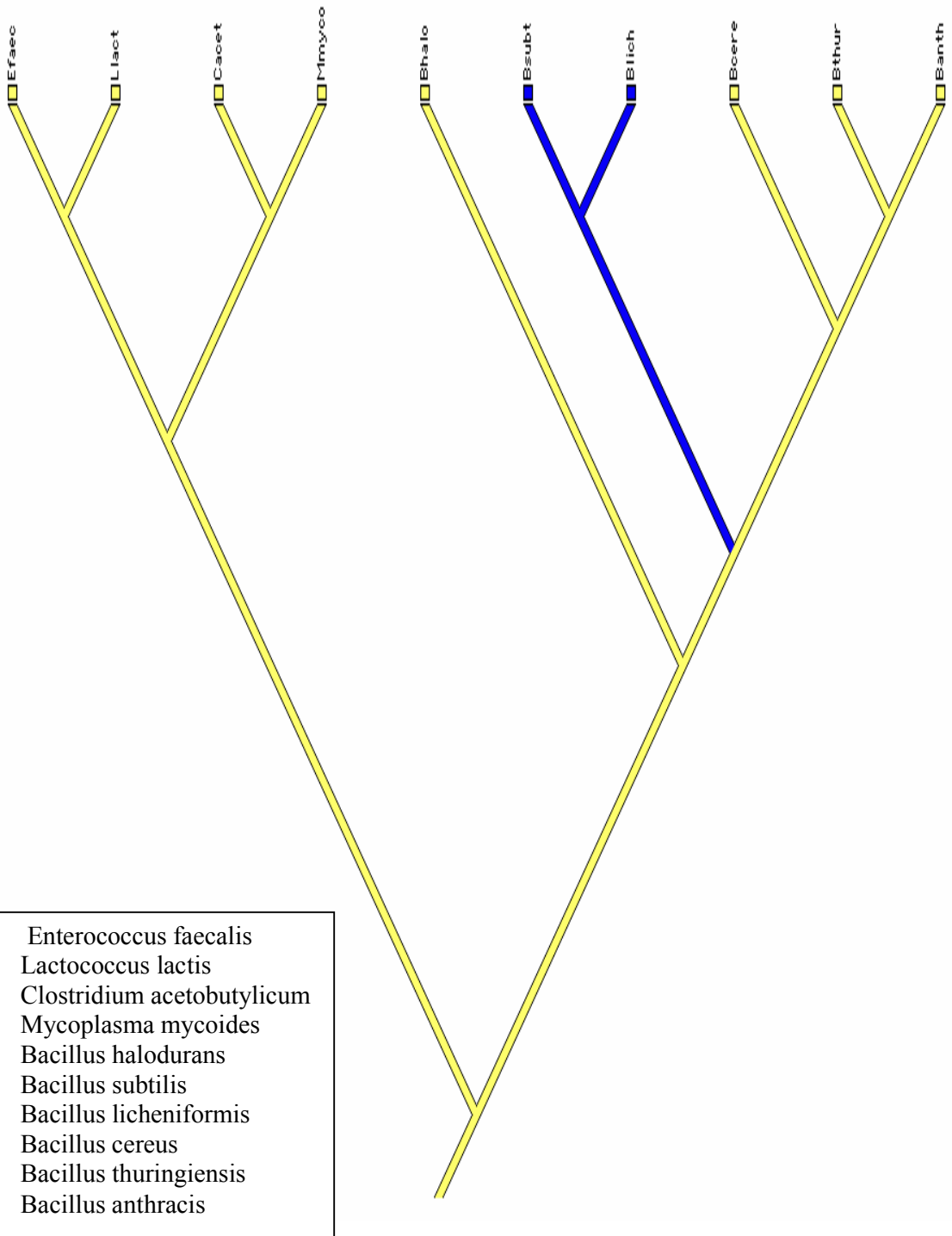


Figure 6. A phylogenetic tree representing the presence (Blue branches) and absence (Yellow branches) of the gene that encodes alcohol dehydrogenase (NADP⁺). As can be seen, alcohol dehydrogenase (NADP⁺) is unique to *B. subtilis* and *B. thuringiensis*

An area of potential research resulting from this study would be the analysis of HGT in the genus *Bacillus*. As can be seen from figure 4, the scores for 53% of the source trees from this analysis lie within the range of greater than 0 and 1.2 in terms of source tree compatibility. Statistical analyses of the supertree presented in this study have revealed high correlations in terms of congruence and confidence. However, 53% of source trees are not in complete agreement with the supertree, therefore it is logical to infer that these source trees may have a different evolutionary history. It is reasonable to ask, “How could these gene trees have such a different evolutionary history?” There are 3 possible hypotheses to this question: Hypothesis one is based on the idea that the source tree and underlying genes may be candidates for horizontal gene transfer; Hypothesis two is founded on the idea of systematic bias. This hypothesis deduces that incongruent source trees, with putative differences in evolutionary history, may be due to a base composition bias; Finally hypothesis three correlates incompatible source tree scores with poor phylogenetic reconstruction (e.g. choosing the wrong maximum likelihood model) (Felsenstein, 2004). Identification and examination of incongruent source trees in relation to the above hypotheses could be the focus of a putative study.

Various forms of statistical, biochemical and phylogenetic data have been presented in this study, which give confidence to and supports the supertree presented in this analysis. As more and more bacterial genomes become publicly available, the importance for evolutionary studies such as this one, will increase. The results from such analyses will have many practical uses, for example, taxonomic classification, protein function prediction, genetic engineering, drug design and etc. In order for the true benefit of phylogenetic analysis to become a reality, more efficient software needs to be developed. This software needs to

produce more accurate results quickly and also needs to be made more user friendly and accessible.

In conclusion, this analysis has focused on the classification of bacterial species based on the identification of single-gene families. These single-gene families have been used to construct a phylogenetic supertree representing all of the bacteria used in this study. In order to infer confidence in our results and assess the reliability of our data, statistical tests such as PTP tests, YAPTP tests and bootstrap analyses were performed. Various lines of evidence, including statistical, biochemical and phylogenetic, support the accuracy of our supertree. To the best of our ability the topology of the supertree, and phylogeny of the genus *Bacillus*, identified by this study, are correct.

REFERENCES:

- Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*. 25: 3389 – 3402.
- Bernal, A., Ear, U. and Kyrpides N. 2001. Genomes OnLine Database (GOLD): a monitor of genome projects world-wide. *Nucleic Acid Research*. 29: 126 – 127.
- Bininda-Emonds, O.R.P. 2004. The evolution of Supertrees. *Trends in Ecology and Evolution*. 19: 315 – 322.
- Black, J. G. 2002. *Microbiology: principles and exporations*. Wiley, New York, USA.
- Bolotin, A., Wincker, P., Mauger., S., Jaillon, O., Malarne, K., Wissenbach, J., Ehrlich, S.D. and Sorokin, A. 2001. The complete genome sequence of the lactic acid bacterium *Lactococcus lactis ssp. lactis* IL1403. *Genome Research*. 11:731 – 753.
- Brown, J.R. 2003. Ancient Horizontal Gene Transfer. *Nature*. 4: 121 – 132.
- Buchanan, R.E. and Gibbons, N.E. 1974. *Bergey's Manual of Determinative Bacteriology* 8th Edition. The Williams & Wilkins Company, Baltimore, USA.
- Creevey C.J. and McInerney J.O. 2005. Clann: investigating phylogenetic information through supertree analyses. *Bioinformatics*. 21: 390 – 392.
- Creevey, C.J., Fitzpatrick, D.A., Philip, G.K., Kinsella, R.J., O'Connell, M.J., Pentony, M.M., Travers, S.A., Wilkinson, M. and McInerney, J.O. 2004. Does a tree-like phylogeny only exist at the tips in the prokaryotes?. *Proceedings of the royal society London B: Biological Sciences*. 271: 2551-2558.
- Daubin, V., Gouy, M. and Perriere, G. 2001. Bacterial molecular phylogeny using supertree approach. *Genome Research*. 12: 155 – 164.
- Efron, B., Halloran, E. and Holmes, S. 1996. Bootstrap confidence levels for phylogenetic trees. *Proceeding of the national academy of sciences of the United States of America*. 23: 13429 – 13434.
- Felsenstein, J. 1993. *PHYLIP: Phylogeny Inference Package*. Seattle, Washington. Distributed by author.
- Felsenstein, J. 2004. *Inferring Phylogenies*. Sinauer Associates, Sunderland, Massachusetts.
- Galtier, N., Gouy, M. and Gautier, C. 1996. SeaView and Phylo_win, two graphic tools for sequence alignment and molecular phylogeny. *Computer Applications to Biosciences* . 12: 543 - 548.

- Hall, B.G. 2004. *Phylogenetics Trees Made Easy: A How-To Manual for Molecular Biologists* 2nd Edition. Sinauer Associates, Sunderland, Massachusetts, USA.
- Ivanova, N., Sorokin, A., Anderson, I., Galleron, N., Candelon, B., Kapatral, V., Bhattacharyya, A., Reznik, G., Mikhailova, N., Lapidus, A., Chu, L., Mazur, M., Goltsman, E., Larsen, N., D'Souza, M., Walunas, T., Grechkin, Y., Pusch, G., Haselkorn, R., Fonstein, M., Ehrlich, S.D., Overbeek, R. and Kyrpides, N. 2003. Genome sequence of *Bacillus cereus* and comparative analysis with *Bacillus anthracis*. *Nature*. 423: 87 – 91.
- Jain, R., Rivera, M.C. and Lake, J.A. 1999. Horizontal gene transfer among genomes: the complexity hypothesis. *Proceeding of the national academy of sciences of the United States of America*. 96: 3801 – 3806.
- McInerney, J.O. 1998. GCUA: general codon usage analysis. *Bioinformatics*. 14: 372 – 373.
- Parkhill, J. and Berry, C. 2003. Relative pathogenic values. *Nature*. 423: 23 – 24.
- Paulsen, I.T., Banerjee, L., Myers, G.S.A., Nelson, K.E., Seshadri, R., Read, T.D., Fouts, D.E., Eisen, J.A., Gill, S.R., Heidelberg, J.F., Tettelin, H., Dodson, R.J., Umayam, L., Brinkac, L., Beanan, M., Daugherty, S., DeBoy, R.T., Durkin, S., Kolonay, J., Madupu, R., Nelson, W., Vamatheman, J., Tran, B., Upton, J., Hansen, T., Shetty, J., Khouri, H., Utterback, T., Radune, D., Ketchum, K.A., Dougherty, B.A. and Fraser, C.M. 2003. Role of Mobile DNA in the evolution of vancomycin-resistant *Enterococcus faecalis*. *Science*. 299: 2071-2074.
- Philip, G.K., Creevey, C.J. and McInerney, J.O. 2005. The Opisthokonta and the Ecdysozoa may not be Clades: Stronger Support for the Grouping of Plant and Animal than for Animal and Fungi and Stronger Support for the Coelomata than Ecdysozoa. *Molecular Biology and Evolution*. In press - Feb 9 2005.
- Read, T.D., Peterson, S.N., Tourasse, N., Baillie, L.W., Paulsen, I.T., Nelson, K.E., Tettelin, H., Fouts, D.E., Eisen, J.A., Gill, S.R., Holtzapple, E.K., Okstad, O.A., Helgason, E., Rilstone, J., Wu, M., Kolonay, J.F., Beanan, M.J., Dodson, R.J., Brinkac, L.M., Gwinn, M., DeBoy, R.T., Madpu, R., Daugherty, S.C., Durkin, A.S., Haft, D.H., Nelson, W.C., Peterson, J.D., Pop, M., Khouri, H.M., Radune, D., Benton, J.L., Mahamoud, Y., Jiang, K.J., Hance, I.R., Wiedman, J.F., Berry, K.J., Plaut, R.D., Wolf, A.M., Watkins, K.L., Nierman, W.C., Hazen, A., Cline, R., Redmond, C., Thwaite, J.E., White, O., Salzberg, S.L., Thomason, B., Friedlander, A.M., Koehler, T.M., Hanna, P.C., Kolsto, A.B. and Fraser, C.M. 2003. The genome sequence of *Bacillus anthracis* Ames and comparison to closely related bacteria. *Nature*. 423: 81 – 86.
- Rey, M.W., Ramaiya, P., Nelson, B.A., Brody-Karpin, S.D., Zaretsky, E.J., Tang, M., Lopez de Leon, A., Xiang, H., Gusti, V., Groth Clausen, I., Olsen, P.B., Rasmussen, M.D., Andersen, J.T., Jorgensen, P.L., Larsen, T.S., Sorokin, A., Bolotin, A., Lapidus, A., Galleron, N., Ehrlich, S.D. and Berka, R.M. 2004. Complete genome sequences of the industrial bacterium *Bacillus licheniformis* and comparisons with closely related *Bacillus* species. *Genome Biology*. 5(10):R77

- Saitou, N. and Nei, M. 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*. 4: 406-425
- Swofford, D.L. 2002. PAUP* : Phylogenetic Analysis Using Parsimony (* and other methods), v.4. Sinauer Associates, Sunderland, Massachusetts, USA.
- Takami, H., Nakasone, K., Takaki, Y., Maeno, G., Sasaki, R., Masui, N., Fuji, F., Hirama, C., Nakamura, Y., Ogasawara, N., Kuhara, S. and Horikoshi, K. 2000. Complete genome sequence of the alkiphilic bacterium *Bacillus halodurans* and genomic sequence comparisons with *Bacillus subtilis*. *Nucleic Acids Research*. 28: 4317-4331.
- Tettelin, H., Fouts, D.E., Eisen, J.A., Gill, S.R., Holtzapple, E.K., Okstad, O.A., Helgason, E., Rilstone, J., Wu, M., Kolonay, J.F., Beanan, M.J., Dodson, R.J., Brinkac, L.M., Gwinn, M., DeBoy, R.T., Madpu, R., Daugherty, S.C., Scott-Durkin, A., Haft, D.H., Nelson, W.C., Peterson, J.D., Pop, M., Khouri, H.M., Radune, D., Benton, J.L., Mahamoud, Y., Jiang, L., Hance, I.R., Weidman, J.F., Berry, K.J., Plaut, R.D., Wolf, A.M., Watkins, K.L., Nierman, W.C., Hazen, Cline, R., Redmond, C., Thwaite, J.E., White, O., Salzberg, S.L., Thomason, B., Friedlander, A.M., Koehler, T.M., Hanna, P.C., Kolsto, A.B. and Fraser, C.M. 2003. The genome sequence of *Bacillus anthracis* Ames and comparison to closely related bacteria. *Nature*. 423: 81 – 86.
- Thompson, J.D., Higgins, D.G. and Gibson, T. J. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignments through sequence weighting, position specific gap penalties and weight matrix choice. *Nucleic Acid Research*. 22: 4673 - 4680.

APPENDIX A:

```
#!/usr/bin/perl
```

```
use strict;  
use warnings;
```

Teaghlach.pl – Used to identify single- and multi-gene families

```
my $fas = "";
```

```
#this is for use with RandomBlast or Systematic blast  
print "ENTER THE NAME OF THE FILE YOU WISH TO WORK WITH :\n";  
$fas = <STDIN>; #takes in command line parameter  
chomp $fas; #removes new lines from the end of STDIN
```

```
my $max = 0;  
print "ENTER THE NUMBER OF FILES YOU WISH TO WORK WITH :\n";  
$max = <STDIN>;  
chomp $max;
```

```
my $i = 0; #will be used in the for loop
```

```
for ($i = 0; $i <= $max; $i++)
```

```
{  
    my $current_file_name = "$i$fas"; #this interprets the value of the variable - ie 0fas.txt  
    1fas.txt
```

```
    if (open (INPUT, $current_file_name))  
        #if statement is only executed if the file is present - ie when non sequential files are present  
        #INPUT is used to access the data in whatever $current_file_name is  
        {  
            my @array = <INPUT>;  
            #takes the contents of the current file and puts it into an array - each new line is new entry in  
            the array  
            close INPUT;
```

```
my $B_ce = 0; #Bacillus cereus
my $B_ha = 0; #Bacillus halodurans
my $B_su = 0; #Bacillus subtilis
my $B_li = 0; #Bacillus licheniformis
my $B_th = 0; #Bacillus thuringiensis
my $B_an = 0; #Bacillus anthracis
my $C_ac = 0; #Clostridium acetobutylicum
my $L_la = 0; #Lactococcus lactis
my $E_fa = 0; #Enterococcus faecalis
my $M_my = 0; #Mycoplasma mycoides
```

```
my $FileString = ""; #holds the entire contents of the file as a string
```

```
foreach (@array)
#for each element in the array, using $_ (default variable) to store the current element
of the array
{
    if (/B_anth/)
    {
        $B_an++; #searches $_ for the expression - increment counter if found
    }

    elsif (/B_cere/)
    {
        $B_ce++;
    }

    elsif (/C_acet/)
    {
        $C_ac++;
    }

    elsif (/L_lact/)
    {
        $L_la++;
    }

    elsif (/M_myco/)
    {
        $M_my++;
    }

    elsif (/E_faec/)
    {
        $E_fa++;
    }

    elsif (/B_thur/)
```

```

    {
        $B_th++;
    }

    elseif (/B_halo/)
    {
        $B_ha++;
    }

    elseif (/B_lich/)
    {
        $B_li++;
    }

    elseif (/B_subt/)
    {
        $B_su++;
    }

    $FileString .= $_;
    # .= acts like append - appends the contents of each element of the array onto the
end.
    #Whole file will be held in this variable when the foreach loop is finished.

} #closes the foreach loop

my $single = $current_file_name.".single";
my $sum = $B_su + $B_li + $B_ha + $B_ce + $C_ac + $L_la + $M_my + $E_fa +
$B_th + $B_an;

if ($B_su<2 && $B_li<2 && $B_ha<2 && $B_ce<2 && $C_ac<2 && $L_la<2 &&
$M_my<2 && $E_fa<2 && $B_th<2 && $B_an<2 && $sum > 3)
{ #single gene family

    open (OUTPUT, ">$single"); #assign a name to your output file
    print OUTPUT $FileString; #prints the of $FileString to the output file and not the
screen
    close OUTPUT;
}

my $multi = $current_file_name.".multi";

if ($B_su>1 || $B_li >1 || $B_ha >1 || $B_ce >1 || $C_ac >1 || $L_la >1 || $M_my >1
|| $E_fa >1 || $B_th >1 || $B_an >1 && $sum >3)
{
    open (OUTPUT, ">$multi");
    print OUTPUT $FileString;
}

```

```
        close OUTPUT;
    }
} #closes if statement
} #closes for loop

exit;
```