# Validation of Reciprocal Orthology Score Average (ROSA) Through Comparison With Other Genomic Metrics and Traditional Phenotypic Methods.

Andrew Gale

4/24/15

**Table of Contents**

# I. Abstract:

The study of bacterial and archaeal taxonomy is hindered by the lack of a robust, widely accepted taxonomic metric that is applicable from the strain level to the genus level and beyond. As a solution to this the Reciprocal Orthology Score Average (ROSA) tool was designed and validated. The ROSA tool is superior to other metrics due to its ability to compare conserved and non-conserved regions of an organism's genome and provide a whole genome measure of similarity. Unlike other metrics, the ROSA tool is also able to account for overall genome size differences between two organisms. When outliers were identified with the tool it was found that other genomic tools and phenotypic results supported the ROSA taxonomic placements. In this study 9 separate organisms were found to be members of the same species, one novel species combination was validated, 11 novel genera were suggested, 9 novel families were suggested, and 1 organism was validated as belonging to a separate phylum. An Average Amino Acid Identity genus threshold was proposed at 72- 64. Based on the agreement with other genomic and phenotypic metrics, ROSA is shown to be a valid tool for bacterial differentiation at all levels of phylogeny.

## **II.** Introduction

## A. History of Microbial Taxonomy

Taxonomy is the science of defining the relationship between groups of organisms based on their characteristics and assigning names to them. Within microbial taxonomy a key concept is that of the type specimen. A type specimen is the official representative of a group, and it can be either for a species or up to a family. To be placed in a group the organism must be compared to the type specimen.

The species concept for non-sexually reproducing organisms has been one of significant change and debate. The first classification attempt for bacteria was done by Antonie van Leeuwenhoek in 1676 using his single lens microscope. He called these organisms animalcules. Later Otto Friedrich Muller was the first to denote a specific group from these animalcules as members of the later defined *Vibrio* genus based on their morphology. The term bacterium was later introduced by Christian Gottfied Ehrenberg in 1838. The currently accepted kingdoms of Bacteria and Achaea were not suggested until Woese and his group defined the Archaea domain in 1990 (Woese et al., 1990).

The first true attempt at bacterial phylogeny came from Ferdinand Cohn in 1872; he showed that the organisms could be classified in the pattern described by Carl von Linné. In this early stage, species were mostly characterized by their relevance to medical microbiology. Properties that were used to identify and characterize species in this time period included requirements for growth, morphology, chemical reactions, and

potential for pathogenicity. Many of these properties are still used today but are supplemented with some genetic component. However, many have denounced this haphazard system (Stackenbrandt 2006). As time progressed, bacterial morphology, supplemented by physiology, became the primary means of identification. From the period of 1950 to 1980 a number of methods were introduced to microbial taxonomy, the most important of which were DNA-DNA Hybridization, rRNA sequencing (particularly 16s rRNA), and the concept of chemotaxonomy. An interesting note is that in 1980 a group of microbiologists showed that out of 40,000 known names only 2500 were validly published, meaning that they were described in a journal with type strains deposited for safe keeping (Stackenbrandt 2006). With the increase in whole genomic sequences the possibility of a universal metric is becoming closer to reality, as even previously unculturable organisms can be studied with such methods.

## B. Impairments of Taxonomy

Beyond determining and describing a standard, well-accepted metric across all taxonomy, one of the largest problems for microbial taxonomy, is the issue of pathology and medical relevance. The beginnings of microbial taxonomy were highly influenced and dictated by the potential pathogenicity of the organism. Due to the large role that bacteria play in human health, many organisms are still defined based on their potential impact on human health. Beyond pathogenicity, misclassifications of other organisms may inhibit the correct reclassification of others as well. The challenge for reclassification of these organisms could be the overly large size of the correct genera,

due to misclassifications of others, and lack of sufficient data to fully argue reclassification.

A classic example of pathogenicity versus phylogeny is the case of the *Escherichia coli* species. It is important to note that the original strain of *E. coli* is thought to be lost and as such a neotype strain, or replacement type strain, was designated for the organism (Meier-Kolthoff et al., 2013). The genus *Shigella*, consisting of *S. dysenteriae*, *S. flexerni*, *S. boydii*, and *S. sonnei*, are, from an evolutionary viewpoint, strains of the *E. coli* species (Lan and Reevs 2002). However due to small genetic differences, which alone would normally not be enough for reclassification or divergence, there is a large difference in the pathology of the strains. Due to these pathological differences the organisms were originally classified as different species. Despite these pathogenic differences, the organisms are genetically members of the same species (Lan and Reevs 2002). However, reclassification has been avoided due to the medical significance of these organisms, in order to reduce confusion for treating doctors and patients.

To further compound the difficulty in *E. coli* classification, study done in 2010 by Lukjancenko et al., found that in 61 sequenced *E. coli* genomes the core genome only encompassed 20% of the total coding sequences, far below what would be expected for members of the same species (Figure 1) (Lukjancenko et al., 2010). A core genome is the set of genes shared by any group of organisms while the pan genome is the total genes between the organisms. This metabolic diversity does not have a significant impact on the DDH values of the organisms, with all comparisons still retaining >70% hybridization.

This study showcases that phenotypic diversity may not always lead to a similar level of genotypic diversity.

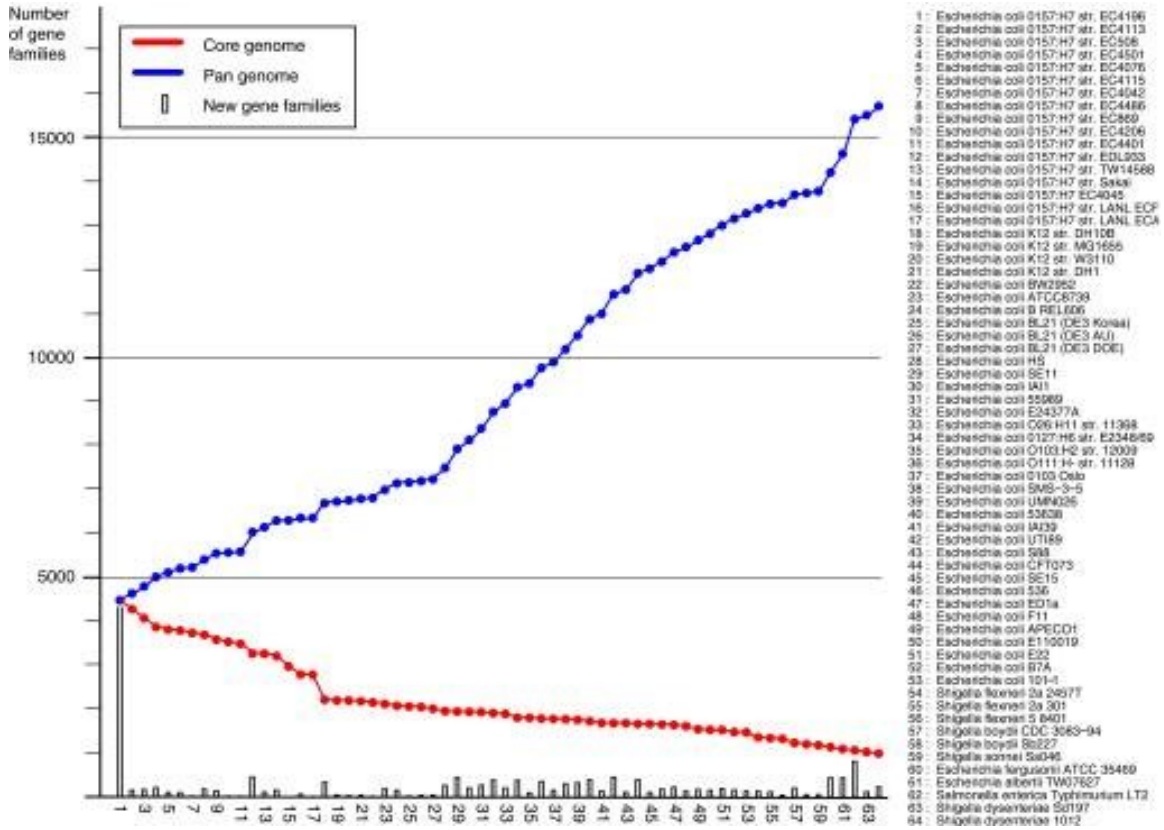


Figure 1: Pan- and core genome plot of the analyzed genomes. The *blue* pan-genome curve connects the cumulative number of genes present in the analyzed genomes. The *red* core genome curve connects the conserved number of gene families. The *gray bars* show the numbers of novel gene families identified in each genome. (Lukjancenko et al., 2010).

A second group with medical relevance and outdated taxonomy is the Bacillus genus. In particular the *B. cereus* group, consisting of *B. anthracis*, *B. cereus*, *B. thuringiensis*, *B. mycoides*, *B. pseudomycoides*, *B. weihenstephanensis* and *B. medusa*, show significant similarity at the genotype level, in particular having 99-100% 16s rRNA, ANI (Average Nucleotide Identity), and AAI (Average Amino Acid Identity) similarity, each greater than 96%. Despite the high level of genome similarity, with DDH values <

70%, they are still classified as different species due to their different pathogenicity. The species differ from each other in the severity and host of the disease. For example *B. thuringiensis* produces and pesticide while *B. anthracis* is the causative agent of anthrax. Resistance to reclassification comes on the grounds that their radically different phenotypes warrant enough medical significance that reclassification would be detrimental to public health (Bavykin et al., 2004).

*Yersinia pestis*, the causative agent of the black plague, was found to be a strain of *Yersinia pseudotuberculosis* that arose approximately 15000 years ago. The *pestis* strain has a segment of DNA called a pathogenicity island that causes it to have a radically different pathology when compared to its ancestor. The cause of this pathogenicity is thought to be the acquisition of a virulence plasmid (Achtman et al., 1999). In light of this pathogenicity reclassification of *Yersinia pestis* as a strain of *Yersinia pseudotuberculosis* has been denied due to the health implications of the phenotype.

Genera may also be nested within another genus, but may resist reclassification due to insufficient genomic data. Such is the case of *Paenibacillus* and *Brevibacillus* in the genus *Bacillus* (Xu and Cote 2003). The *Agrobacterium* genus is nested within the *Rhizobium* genus but reclassification has been resisted due to its characteristic plant pathogenicity (Farrand et al., 2003). A final barrier for correct classification is the potential size of the new genus. The genus *Azotobacter* is nested in the *Pseudomonas* genus but was initially classified due to their nitrogen fixation traits. Reclassification, although suggested, is delayed due to the large size of the *Pseudomonas* genus, 218

species, many are hesitant to add to it until many of the organisms within the genus are correctly reclassified (Palleroni 2010).

## C. Genomic Metrics

A universal, objective, and standardized method to delineate taxa at all ranks within *Archaea* and *Bacteria* would be the "holy grail" of microbial taxonomists. Most taxonomic metrics have focused primarily on species (Tindall et al., 2010), some have focused on subspecies and genus-level classification with the least attention on suprageneric ranks. No formal definitions exist for ranks of family through domain. A whole genome approach which classifies all taxa levels with clear thresholds and replaces DNA-DNA hybridization (DDH) has potential to revolutionize microbial systematics. One of the reasons many are keen to replace DDH has the standard definition is the difficulties associated with the metric. DDH is inherently an error prone method with small variations in the concentration of regents vastly changing the results. Furthermore it is impossible to build a database for ease of DDH values, as each comparison must be done separately. Furthermore beyond the 70% species threshold values have no correlation with greater levels of taxonomy.

Since the pre-genomic era, assignment of a microbial entity to a particular species has been officially based on DNA-DNA Hybridization (DDH). The first clear and standardized "species-concept" was based on strains having at least approximately 70% DDH and with a melting temperature ($T_m$) difference of less than $5^oC$ (Wayne 1987). Thus far, this definition is the only defined and accepted taxonomic metric to date (Wayne 1987). While DDH advanced microbial taxonomy, it does have many

disadvantages. If this metric was applied to eukaryotes the entire primate order would become the same species. Many microbial taxonomists have focused on finding an appropriate wet-lab or computational replacement that has the same or greater taxonomic resolution (Goris et al., 2007; Schleifer, 2009, Stackebrandt et al., 2002). Stackebrandt and Ebers showed that 98.5% 16s rRNA similarity was related to the original 70% DDH (Stackebrandt and Ebers, 2006). 16s rRNA similarity revolutionized microbial taxonomy, but it does not always provide enough resolution to properly distinguish microbial species (Schleifer, 2009). Consequently, groups looked to single-protein coding sequences for greater evolutionary resolution. Later they looked to artificially concatenated genes, and followed by Multi-locus sequence typing or MLST, of protein-coding genes such as gyrB (Richert et al., 2005), rpoB (Adekambi et al., 2009), recN (Glazunova et al., 2009), groEL (Teng et al., 2002), recA (Eisen, 1995), DnaJ (Bustard and Gupta, 1997) and sodA (Poyart et al., 1998) to only name a few. When applicable, such approaches provide greater taxonomic resolution for only some taxa. For example, analysis of the groEL has been shown to be more powerful than gyrB, rpoB, and sodA (Glazunova et al., 2009). While such single-gene and multi-locus approaches aid in classification in addition to DDH and %16S rRNA sequencing, the selection of genes to use may be subjective and difficult, imposing the definition of universal genes between genome pairs (Tindall et al., 2010). Different genes are used to show species delimitation in different taxa, but using different genes may blur the species concept (Schleifer, 2009). In light of the latter problem and due to the exponentially growing

number of microbial genomes being sequenced each year, computational approaches have been designed to utilize whole genome sequences in classification.

Utilizing the growing number of genome sequences, a few very different whole genome computational approaches have been published. Average nucleotide identity (ANI) is the average percent identity of pairwise 1 kilobase sequences with BLAST matches with a minimum of 60% sequence identity and an alignable region of at least 70% similarity. A threshold of about 94% ANI was shown to correlate to the 70% DDH cutoff (Konstantindinis & Tiedje, 2005). A later comparison of ANI and DDH concluded that 95% ANI and 69% conserved DNA corresponded to a 70% DDH (Goris et al., 2007). Currently, there are multiple versions of ANI: $ANI_b$ (blast, as explained above) and $ANI_m$ (based as the Mummer algorithm), which differ in how they process the original genomic sequence (Richter and Rosselló-Móra, 2009). Average amino acid identity (AAI) is another metric like ANI but with a cutoff of 30% identity and an alignable region of at least 70% based on BLAST matches at the amino acid level (Konstantindinis and Tiedje, 2005).

Oligonucleotide frequencies are being considered as species-specific indicators (Takahashi et al., 2009). The tetranucleotide frequency correlation coefficients (TETRA) are most helpful when complementing $ANI_m$ analysis, but may not be universally applicable (Richter and Rosella-Mora, 2009). A few available software packages include calculations of $ANI_b$, $ANI_m$, and TETRA such as Jspecies (Richter and Rosselló-Móra, 2009). In contrast to analyses that utilize only a few shared CDSs, such as MLSA, techniques that incorporate all shared genes, such as AAI and ANI, offer better

taxonomic resolution than techniques that group microbes based on marker genes (Konstantindinis & Tiedje, 2005). The Genome-Genome-Distance-Calculator is an *in silico* version of DDH. While offering a genome-era relevant mechanism to acquire DDH for organisms, it still suffers from the same previously mentioned flaws as standard DDH (Auch et al., 2010). Other limitations of the method are the black box calculations and high scoring segment pairs.

The Rapid Annotation Using Subsystems Technology (RAST) website is a genomic analyzer tool developed by Aziz et al., in 2008. The RAST site uses subsystems technology to annotate genomes at the SEED quality. The site allows comparisons of genomes through a sequence or KEGG, Kyoto Encyclopedia of Genes and Genomes, based comparison. Furthermore the site allows metabolic reconstructions and the mapping of genes to subsystems. RAST annotation output can be used to help construct draft genomes and make predictions as to the phenotype of the organism (Aziz et al., 2008).

This Honor's Thesis presents the Reciprocal Orthology Score Average (ROSA), a novel genome similarity metric that integrates the similarity of orthologs and the percentage of the genome composed of orthologs. It has been designed to conserve the species concept based on the 70% DDH cut-off but without the disadvantages of DDH. ROSA is a metric that encompasses shared protein-coding genes based in part on genome similarity, genome conservation, and percentage of the genome that is orthologous, all thought to be related to DDH (Goris et al., 2007). ROSA is the first computational substitute to DDH that has ample resolution to differentiate all levels of

taxa. We have defined thresholds for all taxonomic levels and show the potential of ROSA to be the basis for a genome-based prokaryotic classification. Further discussed in this paper are the reconciliation of taxonomic outliers and general complications with current taxonomy.

## III. Methods

Genome sequence:

Genomic DNA was isolated from *Flavobacterium aquatile* LMG 4008$^T$ using the Qiagen DNeasy Blood and Tissue kit according to the manufacturer's instructions. Libraries were prepared and then sequenced on an Illumia MiSeq (V3 2 x 300 base) by the Indiana University Center for Genome Studies as part of a genome Consortium for Active Teaching NextGen Sequencing Group (Gcat-Seek) shared run (Buonaccorsi et al., 2014). Sequencing reads were filtered (median phred score >20), trimmed (phred score >16) and assembled using the paired-end de novo assembly option in NextGene V2.3.4.2 (SoftGenetics). The assembled contigs were uploaded to RAST (Aziz et al., 2008) for analysis. Contigs with a length of less then 200 bp and low coverage, <10x, were discarded due to unreliability. Smaller contigs (under 5000bp) were BLASTed against the uploaded genome to discard those that were repeats of a larger contig. A total of 2,056,329 reads were assembled into 7 contigs with a total genome length of 3,491,115 base pairs and 3,222 coding sequences.

Genomic DNA was isolated from *Flavobacterium sp. Nov AED* using the Qiagen DNeasy Blood and Tissue kit according to the manufacturers instructions. Libraries were prepared and then sequenced on an Illumia MiSeq (V3 2 x 300 base) by the Indiana University Center for Genome Studies as part of a genome Consortium for Active Teaching NextGen Sequencing Group (Gcat-Seek) shared run (Buonaccorsi et al., 2014). Sequencing reads were filtered (median phred score >20), trimmed (phred score >16) and assembled using the paired-end de novo assembly option in NextGene V2.3.4.2

(SoftGenetics). The assembled contigs were uploaded to RAST (Aziz et al., 2008) for analysis. The assembled contigs were uploaded to RAST (Aziz et al., 2008) for analysis. Contigs with a length of less then 200 bp and low coverage, <10x, were discarded due to unreliability. Smaller contigs (under 5000bp) were blasted against the uploaded genome to discard those that were repeats of a larger contig. A total of 1,682,774 reads were assembled into 15 contigs with a total genome length of 3,925,153 base pairs and 3,462 coding sequences.

16s rRNA tree and % differences:

In order to compare 16s rRNA sequences between different species of the same genus, EzTaxon's CompGen feature was used in conjunction with 16s rRNA sequences of the type strain for the type species as deposited in NCBI databases. Coding sequences recovered from the CompGen were aligned by the ClustalW program in MEGA6. Aligned sequences were used to form a neighbor-joining tree using the Kimura two-parameter method from the available sequences. Bootstrap values (expressed as percentages of 1000 replications) are given at the nodes (http://eztaxon-e.ezbiocloud.net, Kim *et al.,* 2012; Tamura *et al.,* 2013). 16s rRNA % differences were calculated using the pairwise similarity method in MEGA6. Sequences aligned by ClustalW were compared using the bootstrap method with n=1000.

RAST Annotation:

In order to facilitate this study, many genomes were uploaded into the RAST webserver for annotation and usage of the sequence based comparison, which formed the core for many tests. Organisms whose sequences were not within the RAST

database were uploaded using their standard recommended procedures (Aziz et al., 2008). Genomic sequences were taken from their respective deposits within the NCBI genome portal; WGS sequences were used with preference due to their more complete and accurate nature.

Biolog:

Phenotypic characterizations generated in this study were done using the Biolog Omnilog system. Strains were grown on BUG+Blood agar and transferred to Biolog GenIII plates at concentrations recommended by standard protocol's. Strains were incubated for 36 hours at $30^0$C in the Biolog incubator. Metabolic results were analyzed using the Omnilog scanner system (bioMerieux).

Genomic comparisons:

To simulate DDH, strains were compared using the Genome to Genome Distance Calculator software. Version 2 was used as per publisher recommendations. All strain comparisons were type strains unless otherwise indicated (http://ggdc.dsmz.de) (Auch et al., 2010).

Average Nucleotide Identity values were calculated using the using the algorithm described by Goris et al., (2007). Software support and utilization of this metric was done using the ExGenome web service (http://eztaxon-e.ezbiocloud.net/ezgenome/ani) (Kim et al., 2012).

Average Amino Acid Identity was calculated via the Newman Lab ROSA calculator using sequence based comparison output files from RAST (http://lycofs01.lycoming.edu/~newman/ROSA/index.htm).

%Bi-directional Best Hit values were calculated via the Newman Lab ROSA calculator using sequence based comparison output files from RAST (http://lycofs01.lycoming.edu/~newman/ROSA/index.htm).

ROSA values were calculated via the Newman Lab ROSA calculator using sequence based comparison output files from RAST (http://lycofs01.lycoming.edu/~newman/ROSA/index.htm).

Relative gene counts were constructed in two ways. Both methods made use of the output from the sequence based comparisons from RAST. Strains were compared in groups of five, with each organism serving as the reference and the other four as comparisons. Output TSV files were analyzed in two ways, a manual and an automated method. TSV files analyzed manually were transferred to Excel and sorted by hit type with bidirectional hits being favored. After sorting gene counts were manually determined using the spreadsheet cell number and the counts for different combinations were noted. Total counts were averaged to better represent overall relatedness in gene count. The second method was done through an automated Excel program. The program was designed, tested, and validated using a manual sorted comparison cluster. Gene clusters of particular importance were pulled from the comparisons and analyzed based on their annotated function.

Literature search:

Phenotypic characteristics and genomic comparison values were, where noted, taken from literature when available.

# IV. Results:

## A. The Design of ROSA:

To design a genome-based taxonomic metric that could simulate DDH at the species level, we sought to incorporate two terms, percent conserved DNA and ortholog similarity, that both contribute to the ability of DNA from two different organisms to hybridize (Goris et al., 2007). Therefore, we reasoned that a novel microbial genomic metric, the orthology score (OS), could be determined by multiplying ortholog similarity by genome conservation.

$$OS = ortholog\ similarity * genome\ conservation$$

Formula 1: OS formula

We sought to determine genome conservation based on protein-coding genes. % BBH was designed based on the total amino acids encoded by BBH-called genes of the compared genome over the total amino acids encoded by all protein-coding genes of the reference genome (Formula 2).

$$\%\ BBH = \frac{\Sigma length_{bbh}}{length_{ref}}$$

Formula 2: %BBH formula.

Ortholog similarity could've be based on two previously published metrics, ANI (specifically ANIb) or AAI (Konstantinidis and Tiedje, 2005), which have the same 95% thresholds at the species level. The degeneracy of the genetic code causes the nucleotide sequence similarity of orthologous genes to be lower than the amino acid

sequence similarity of orthologous proteins. The fact that 70% DDH correlated to both 95% ANI and 96% AAI may result from the fact that the AAI calculation includes a greater number of orthologs, many of which are below the threshold level to be included in the ANI calculation. Thus amino acid sequence comparisons (BLASTP-based BBH) are a more sensitive way to detect orthologs.

A new tool to calculate AAI based on the RAST server was designed (http://lycofs01.lycoming.edu/~newman/ROSA/index.htm). This AAI (AAIr) is calculated for all BBH-called genes without a sequence similarity threshold. Moreover, rather than averaging the AAI of all genes, we normalize the AAI contribution of a gene by its length; all of the shared amino acids per BBH gene were summed and divided by the total number of amino acids of orthologs (Formula 3).

$$AAIr = \left[\frac{\Sigma(percent\ identity * lengthBBH)}{\Sigma lengthBBH}\right] * 100$$

Formula 3: AAIr formula

We compared our AAIr to the previously published AAI from the Konstantadinis group (AAIk) for 120 genome-pair comparisons (15 pairs per each taxonomic comparison level). AAIk results were identical whether no cut-off or 20% cut-off was used. With no cut-off, AAIk and AAIr yield similar results with only a minor percent difference. < 2%, presumably due to the normalization of AAIper gene to compute AAIr. With a 30% similarity threshold, AAIk was anywhere from 12.6 to 22% different compared to AAIr for interdomain comparisons; this difference decreased as the strains

become more taxonomically similar. As all BBH-called genes are orthologous within RAST, it was not necessary to impose a threshold to ensure inclusion of other orthologs.

Once we had decided to use AAI instead of ANI for ortholog similarity, we explored the possibility of squaring it. Squaring AAI simulates two strands of DNA coming together during hybridization, with a term for each strand. An OS value based on either AAI and $AAIr^2$ were very similar (Figs 1A and 1B), but the correlation of the $AAIr^2$ based OS to averaged reciprocal DDH (rDDH) values was higher for all relationships by both linear correlation and Kendall's non-parametric analysis, a more robust statistical methodology for datasets that are do not belong to any particular distribution. Therefore, the genome similarity term, OS is based on $AAIr^2$ (Formula 4).

$$OS = \left(\frac{AAIr}{100}\right)^2 \%BBH$$

Formula 4: Final OS formula

Next, we reasoned that averaging the OS values between two pairwise comparisons would minimize effects of different genome sizes. For example, a 1 Mbp genome compared to a 10Mbp genome can have 100% orthology, yet a 10Mbp compared to a 1Mbp can have at maximum 10% orthology. By averaging the reciprocal comparisons the formula allows a better comparison at the whole genomic level between two strains that is not limited by their genomic sizes. Likewise, the Reciprocal Orthology Score Average (ROSA) is based on the average of each OS value for a pair of reciprocal comparisons (Formula 5).

$$ROSA = \frac{(OS_{A\to B} + OS_{B\to A})}{2}$$

Formula 5: ROSA formula

A user-friendly tool was designed in conjunction with the RAST annotation server to enable wide use of ROSA for genome-based microbial taxonomy analysis and related purposes.

| Level | Same | Different | expected range | min | max | mean | n= | below range | above range |
|---|---|---|---|---|---|---|---|---|---|
| 8 | species | strain | >65 | 49 | 99.59 | 85.98 | 312 | 3 | N/A |
| 7 | genus | species | 35-65 | 6.85 | 95.8 | 36.63 | 521 | 294 | 38 |
| 6 | family | genus | 15-35 | 5.8 | 54.99 | 18.86 | 524 | 125 | 19 |
| 5 | order | family | 10-15 | 4.75 | 23.19 | 11.58 | 314 | 79 | 22 |
| 4 | class | order | 8-10 | 4.15 | 14.99 | 8.22 | 286 | 121 | 58 |
| 3 | phylum | class | 6-8 | 4.3 | 11.4 | 6.62 | 123 | 42 | 14 |
| 2 | domain | phylum | 3-6 | 1.85 | 7.66 | 4.45 | 211 | 10 | 6 |
| 1 | | domain | <3 | 1.36 | 4.44 | 2.42 | 47 | 0 | 9 |

Table 1: ROSA taxonomic ranges with comparison count and number of outliers per taxonomic range.

Ranges were developed for ROSA values and subsequently compared to current taxonomy. When applicable type strains were used for comparisons since they are the

official representative of their species. For strains that are the same species it is expected that ROSA values are greater than 65. For strains that are different species but part of the same genus ROSA values will range from 35-65. For strains that are in the same family but different genera ROSA values will range from 15-35. For strains that are in different family but the same order ROSA values will range from 10-15. For strains that are in different orders but the same class ROSA values will range from 8-10. For strains that are in different classes but the same phylum expected ROSA values will range from 6-8. For strains that are in different phyla but the same domain expected ROSA values will range from 3-6. For organisms that are in different domains expected ROSA values will be less than 3 (Table 1).

## B. Example ROSA Clusters

| Staphylococcus aureus A | ROSA | 1241616 | 426430 | 93062 | 93061 | 367830 | 418127 | 158879 | 359787 | 282459 | 273036 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Staphylococcus aureus DSM 20231 TTT | 1241616 | | | | | | | | | | |
| Staphylococcus aureus str. Newman | 426430 | 94.9 | | | | | | | | | |
| Staphylococcus aureus COL | 93062 | 94.8 | 92.7 | | | | | | | | |
| Staphylococcus aureus NCTC 8325 | 93061 | 94.8 | 93.3 | 94.0 | | | | | | | |
| Staphylococcus aureus USA300 | 367830 | 93.5 | 93.1 | 94.8 | 92.7 | | | | | | |
| Staphylococcus aureus Mu3 | 418127 | 90.9 | 90.9 | 89.1 | 89.1 | 89.4 | | | | | |
| Staphylococcus aureus N315 | 158879 | 90.7 | 89.2 | 89.6 | 88.8 | 89.7 | 94.9 | | | | |
| Staphylococcus aureus JH1 | 359787 | 90.5 | 89.1 | 89.0 | 89.2 | 89.4 | 93.0 | 93.3 | | | |
| Staphylococcus aureus MSSA476 | 282459 | 90.4 | 88.4 | 90.4 | 90.9 | 90.1 | 87.9 | 89.4 | 88.3 | | |
| Staphylococcus aureus RF122 | 273036 | 88.6 | 86.4 | 86.8 | 86.9 | 85.6 | 86.8 | 86.7 | 85.4 | 86.0 | |
| Staphylococcus aureus MRSA252 | 282458 | 85.7 | 85.1 | 85.9 | 85.2 | 86.6 | 85.3 | 86.9 | 85.5 | 87.8 | 83.7 |

| Streptococcus pyogenes B | ROSA | 1123316 | 293653 | 160490 | 198466 | 193567 | 160491 | 186103 | 370551 | 370554 | 319701 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Streptococcus pyogenes DSM 20565 T | 1123316 | | | | | | | | | | |
| Streptococcus pyogenes MGAS5005 | 293653 | 93.7 | | | | | | | | | |
| Streptococcus pyogenes M1GAS | 160490 | 92.0 | 91.4 | | | | | | | | |
| Streptococcus pyogenes MGAS315 | 198466 | 88.6 | 87.4 | 85.6 | | | | | | | |
| Streptococcus pyogenes SSI-1 | 193567 | 87.7 | 86.4 | 84.4 | 95.4 | | | | | | |
| Streptococcus pyogenes str. Manfredo | 160491 | 89.0 | 88.7 | 87.4 | 87.7 | 87.2 | | | | | |
| Streptococcus pyogenes MGAS8232 | 186103 | 88.5 | 84.8 | 87.1 | 87.2 | 85.9 | 89.1 | | | | |
| Streptococcus pyogenes MGAS9429 | 370551 | 87.5 | 89.9 | 88.0 | 85.8 | 83.4 | 88.9 | 86.0 | | | |
| Streptococcus pyogenes MGAS10750 | 370554 | 88.6 | 88.3 | 86.4 | 85.3 | 83.2 | 86.8 | 85.7 | 86.9 | | |
| Streptococcus pyogenes MGAS6180 | 319701 | 87.8 | 87.9 | 88.6 | 84.8 | 82.1 | 86.7 | 85.7 | 89.4 | 85.8 | |
| Streptococcus pyogenes MGAS2096 | 370553 | 85.7 | 88.4 | 86.4 | 83.5 | 81.5 | 86.5 | 83.0 | 93.1 | 83.5 | 87.322 |

| Escherichia coli C | ROSA | 866789 | 83333 | 316407 | 511145 | 316385 | 481805 | 155864 | 364106 | 83334 | 405955 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Escherichia coli DSM 30083 TTTT = JCM 1649 TTTT | 866789 | | | | | | | | | | |
| Escherichia coli K12 | 83333 | 81.4 | | | | | | | | | |
| Escherichia coli W3110 | 316407 | 81.7 | 98.1 | | | | | | | | |
| Escherichia coli str. K-12 substr. MG1655 | 511145 | 82.1 | 97.7 | 97.7 | | | | | | | |
| Escherichia coli str. K-12 substr. DH10B | 316385 | 79.5 | 94.1 | 94.3 | 95.3 | | | | | | |
| Escherichia coli ATCC 8739 | 481805 | 80.0 | 89.4 | 89.8 | 90.2 | 88.0 | | | | | |
| Escherichia coli O157:H7 EDL933 | 155864 | 74.6 | 80.8 | 80.9 | 80.5 | 78.3 | 79.9 | | | | |
| Escherichia coli UTI89 | 364106 | 92.0 | 79.6 | 80.0 | 80.5 | 78.0 | 78.6 | 73.6 | | | |
| Escherichia coli O157:H7 | 83334 | 73.2 | 79.2 | 79.4 | 79.0 | 76.9 | 78.3 | 92.2 | 72.5 | | |
| Escherichia coli APEC O1 | 405955 | 90.6 | 77.3 | 77.6 | 77.7 | 75.5 | 76.6 | 71.4 | 88.3 | 70.2 | |
| Escherichia coli CFT073 | 199310 | 85.8 | 77.1 | 77.4 | 77.5 | 75.4 | 77.6 | 71.6 | 85.7 | 70.2 | 82.0 |

Table 2: ROSA values for intraspecies strains. Table 2A. ROSA values for strains of *Staphylococcus aureus*. Table 2B. ROSA values for strains of *Streptococcus pyogenes*. Table 2C. ROSA values for strains of *Escherichia coli*.

ROSA values for strains of *Staphylococcus aureus* were acquired using full genome sequences. ROSA values from within the cluster ranged from ~ 82 to ~98. Strains within the species all had ROSA values greater than the 65 threshold for intraspecies relatedness (Table 2A). ROSA values for strains of *Streptococcus pyogenes* were acquired using full genome sequences. ROSA values from within the cluster ranged from ~ 81 to ~96. Strains within the species all had ROSA values greater than the 65

threshold for intraspecies relatedness (Table 2B). ROSA values for strains of *Escherichia coli* were acquired using full genome sequences. ROSA values from within the cluster ranged from ~ 70 to ~98. Strains within the species all had ROSA values greater than the 65 threshold for intraspecies relatedness (Figure 2C).

| **Staphylococcus** | **ROSA** | **93062** | **367830** | **93061** | **282459** | **158878** | **273036** | **282458** | **176279** | **176280** | **279808** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Staphylococcus aureus subsp. aureus COL | **93062** | | | | | | | | | | |
| Staphylococcus aureus subsp. aureus USA300 | **367830** | 94.8 | | | | | | | | | |
| Staphylococcus aureus subsp. aureus NCTC 8325 | **93061** | 93.9 | 92.7 | | | | | | | | |
| Staphylococcus aureus subsp. aureus MSSA476 | **282459** | 90.3 | 90.1 | 90.8 | | | | | | | |
| Staphylococcus aureus subsp. aureus Mu50 | **158878** | 88.0 | 88.2 | 88.1 | 87.5 | | | | | | |
| Staphylococcus aureus RF122 | **273036** | 86.7 | 85.6 | 86.8 | 86.0 | 86.1 | | | | | |
| Staphylococcus aureus subsp. aureus MRSA252 | **282458** | 85.9 | 86.5 | 85.2 | 87.8 | 84.8 | 83.6 | | | | |
| Staphylococcus epidermidis RP62A | **176279** | 43.4 | 43.3 | 42.5 | 43.5 | 43.3 | 43.5 | 43.9 | | | |
| Staphylococcus epidermidis ATCC 12228 | **176280** | 43.2 | 43.2 | 42.6 | 43.7 | 42.6 | 43.7 | 43.2 | 86.4 | | |
| Staphylococcus haemolyticus JCSC1435 | **279808** | 41 | 41.2 | 40.7 | 41.1 | 40.6 | 41.3 | 41.2 | 44.9 | 44.9 | |
| Staphylococcus saprophyticus ATCC 15305$^{T}$ | **342451** | 38.3 | 38.3 | 48.0 | 38.6 | 37.6 | 38.6 | 38.3 | 39. | 39.7 | 40.1 |

ROSA values for organisms within the *Staphylococcus* genus ranged from ~37 to ~95 (Table 3). Strains of *S. aureus* had ROSA values ranging from 83 to 95 when compared to other strains of the organism, and ROSA values ranging from 37 to ~ 44 when compared to organisms within the *Staphylococcus* genus, values consistent with expected ROSA ranges for Intraspecies and Intragenus comparisons (Table 3). Strains of *S. epidermidis* had a ROSA value of 86.44 when compared to strains within the same species and ROSA values ranging from 43 to 44 when compared to other members of the genus (Table 3). Strains of *S. haemolyticus* showed ROSA values ranging from 40 to 45 when compared to other members of the *Staphylococcus* genus (Table 3). Strains of

Table 3: ROSA values intraspecies and intragenus comparisons within the Staphylococcus genus. Intraspecies comparisons were done between strains of S. aureus and S. epidermis. Intragenus comparisons were done between strains of S. aureus, S. epidermis, S. haemolyticus, and S. saprophyticus.

*S. saprophyticus* had ROSA values ranging from 37 to 40 when compared to other members of the Staphylococcus genus (Table 3).

## C. Brucella

| Brucella | ROSA | 224914 | 546272 | 262698 | 430066 | 204722 | 470137 | 483179 |
|---|---|---|---|---|---|---|---|---|
| Brucella melitensis 16M$^T$ | 224914 | | | | | | | |
| Brucella melitensis ATCC 23457 biovar 2 | 546272 | 95.3 | | | | | | |
| Brucella abortus biovar 1 str. 9-941 | 262698 | 94.5 | 94.5 | | | | | |
| Brucella abortus S19 | 430066 | 94.0 | 95.3 | 97.4 | | | | |
| Brucella suis 1330$^T$ | 204722 | 94.1 | 94.7 | 96.1 | 95.0 | | | |
| Brucella suis ATCC 23445 | 470137 | 90.1 | 90.3 | 92.0 | 92.3 | 93.5 | | |
| Brucella canis ATCC 23365$^T$ | 483179 | 90.6 | 91.3 | 92.3 | 93. | 95.1 | 96.0 | |
| Brucella ovis ATCC 25840$^T$ | 444178 | 90.9 | 93.5 | 92.0 | 92.4 | 92.4 | 88.0 | 89.0 |

Table 4: ROSA values for intraspecies and intragenus for members of the *Brucella* genus. Intraspecies comparisons were done with strains of the species *B. melitensis, B. abortus*, and *B. suis*. Intragenus comparisons were done between *B. melitensis, B. abortus, B. suis, B. canis,* and *B. ovis*.

ROSA values for organisms within the *Brucella* genus ranged from 88 to 97 (Table 4). Intraspecies comparisons for *B. melitensis, B. abortus,* and *B. suis* all had values greater than 93 (Table 4). Intragenus comparisons for *B. melitensis, B. abortus, B. suis, B. canis*, and *B. ovis* had ROSA values greater than 88 (Table 4). ROSA values for the intraspecies comparisons fell within the expected range for strains belonging to the same species. ROSA values for intragenus comparisons fell above the expected range for intragenus comparisons and within the intraspecies comparison range, suggesting the organisms are all members of the same species (Table 4). GGDC values for reciprocal comparisons of members within the genus showed high levels of similarity. All members had values greater than 95, values greater than the standard 70% for species threshold.

No organisms had values less than 70 that would indicate separateness of the species (Table 5).

| | | DDH (Formula 2) | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | NC_003317 NC_003318 | NC_012441 NC_012442 | NC_006932 NC_006933 | NC_010740 NC_010742 | AE014291 AE014292 | CP000911 CP000912 | NC_010103 NC_010104 | NC_009504 NC_009505 |
| Brucella melitensis 16M$^T$ | NC_003317 NC_003318 | | 99.3 | 98 | 98.1 | 97.7 | 97.4 | 97.5 | 96.6 |
| Brucella melitensis ATCC 23457 biovar 2 | NC_012441 NC_012442 | 99.3 | | 98.2 | 98.2 | 97.9 | 97.5 | 97.7 | 96.8 |
| Brucella abortus biovar 1 str. 9-941 | NC_006932 NC_006933 | 98 | 98.2 | | 100 | 97.8 | 97.5 | 97.7 | 96.8 |
| Brucella abortus S19 | NC_010740 NC_010742 | 98.1 | 98.2 | 100 | | 97.8 | 97.5 | 97.7 | 96.8 |
| Brucella suis 1330$^T$ | AE014291 AE014292 | 97.7 | 97.9 | 97.8 | 97.8 | | 98.7 | 100 | 97.3 |
| Brucella suis ATCC 23445 | CP000911 CP000912 | 97.4 | 97.5 | 97.5 | 97.5 | 98.7 | | 98.5 | 96.9 |
| Brucella canis ATCC 23365$^T$ | NC_010103 NC_010104 | 97.5 | 97.7 | 97.7 | 97.7 | 100 | 98.5 | | 97.1 |
| Brucella ovis ATCC 25840$^T$ | NC_009504 NC_009505 | 96.6 | 96.8 | 96.8 | 96.8 | 97.3 | 96.9 | 97.1 | |

Table 5: GGDC values for members of the *Brucella* genus. Simulated DDH output for all reciprocal comparisons showed values greater then 96.

## D. Francisella and Thiomicrospira

| Francisella | ROSA | 393011 | 458234 | 119857 | 418136 | 393115 | 441952 | 177416 | 401614 | 484022 | 317025 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Francisella tularensis subsp. holarctica OSU18 | 393011 | | | | | | | | | | |
| Francisella tularensis subsp. holarctica FTA | 458234 | 93.6 | | | | | | | | | |
| Francisella tularensis subsp. holarctica LVS | 119857 | 92.7 | 92.5 | | | | | | | | |
| Francisella tularensis subsp. tularensis WY96-3418 | 418136 | 88.9 | 88.8 | 88.3 | | | | | | | |
| Francisella tularensis subsp. tularensis FSC198 | 393115 | 88.7 | 88.6 | 88.2 | 91.9 | | | | | | |
| Francisella tularensis subsp. mediasiatica FSC147 | 441952 | 87.4 | 87.7 | 87.8 | 90.1 | 90.0 | | | | | |
| Francisella tularensis subsp. tularensis Schu 4 | 177416 | 85.3 | 85.3 | 84.9 | 87.8 | 89.4 | 86.5 | | | | |
| Francisella tularensis subsp. novicida U112 | 401614 | 84.4 | 84.3 | 83.8 | 85.8 | 85.5 | 84.6 | 81.9 | | | |
| Francisella philomiragia subsp. philomiragia ATCC 25017 | 484022 | 55.8 | 55. | 55.4 | 56.5 | 56.3 | 56.2 | 54.5 | 58.5 | | |
| Thiomicrospira crunogena XCL-2 | 317025 | 8.0 | 8.1 | 8.0 | 8.0 | 8.1 | 8.0 | 8.2 | 8.1 | 8.0 | |
| Thiomicrospira denitrificans ATCC 33889$^T$ | 326298 | 5.0 | 5.0 | 5.03 | 5.0 | 5.0 | 4.9 | 5.1 | 5.1 | 5.2 | 7.0 |

Table 6: ROSA values for members of the *Francisella* genus and the *Thiomicrospira* genus. Intraspecies comparisons were done between strains of *F. tularensis*. Intragenus comparisons were done between *F. tularensis* and *F. philomiragia*. Intragenus comparisons were done between *T. crunogen* and *F. denitrificans*. Intraclass comparisons were done between *F. tularensis, F. philomiragia, T. crunogen*, and *F. denitrificans*.

ROSA values for strains belonging to *F. tularensis* had values ranging from 81 to ~94, values within the intraspecies range (Table 6). ROSA value for intragenus

comparison between *T. crunogen* and *F. denitrificans* was ~7, within the intraphyla range. Comparisons between *Francisella tularensis* and strains of the *Thiomicrospira* genus ranged from ~5 to ~8 spanning the intraphyla and intraclass ranges (Table 6).

## E. Flavobacterium:

Complete 16s rRNA sequence data when compared yielded a phylogenetic tree that encompassed the entirety of the *Flavobacterium* genus. The tree showed deep branching among the genus with several clusters being significantly separated from the type species. *F. aquatile* had 16s rRNA % similarities ranging from 96.2% to 93.9% within cluster 1 (Table 7A). *F. aquatile* had 16s rRNA % similarities ranging from 94.7% to 92.8% within cluster 2 (Table 7B).

| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|---|
| Flavobacterium_denitrificans_ED5T | 1 | | | | | | | |
| Flavobacterium_johnsoniae_UW101T | 2 | 95.5 | | | | | | |
| Flavobacterium_chilense_LM-09-FpT | 3 | 95.8 | 96.7 | | | | | |
| Flavobacterium_chungangense_CJ7T | 4 | 96.0 | 97.0 | 98.5 | | | | |
| Flavobacterium_hibernum_ATCC_51468T | 5 | 94.0 | 94.8 | 96.8 | 96.6 | | | |
| Flavobacterium_hydatis_DSM_2063T | 6 | 94.4 | 95.8 | 97.2 | 97.1 | 97.2 | | |
| Flavobacterium_reichenbachii_WB_3.2-61T | 7 | 95.1 | 95.9 | 97.0 | 96.6 | 97.0 | 96.9 | |
| Flavobacterium_aquatile_ATCC_11947T/ | 8 | 93.3 | 93.4 | 95.1 | 95.2 | 94.4 | 94.9 | 94.1 |

| | | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| Flavobacterium_limnosediminis_JC2902T | 1 | | | | | |
| Flavobacterium_cauense_R2A-7T | 2 | 98.6 | | | | |
| Flavobacterium_saliperosum_S13T | 3 | 98.1 | 98.3 | | | |
| Flavobacterium_sasangense_YC6274T | 4 | 92.7 | 93.1 | 92.5 | | |
| Flavobacterium_suncheonense_GH29-5T | 5 | 95.7 | 96.1 | 94.4 | 93.1 | |
| Flavobacterium_aquatile_ATCC_11947T | 6 | 94.7 | 94.1 | 93.4 | 94.6 | 92.8 |

Table 7: **16%rRNA similarities between members of the *Flavobacterium* genus.** A. 16s rRNA % differences for organisms in cluster 1 compared to *F. aquatile* and within the cluster. B. 16s rRNA % differences for organisms in cluster 2 compared to *F. aquatile* and within the cluster.

Based on the tree, clusters with a significant number of organisms whose genomes were sequenced had Average amino acid values calculated and compared to others in their cluster as well as compared to *F. aquatile* (Figure8 ). Cluster 1 had less then 70% Average amino acid similarity to *F. aquatile* while retaining 77-84% similarity within the cluster (Table 8A). Cluster 2 had less then 70% Average Amino Acid similarity to *F. aquatile* while retaining ~80% and greater similarity within its cluster (Table 8B). Indicating that the flavobacterium genus needs to be split.

A

| Average Amino Acid Identity | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| F. succinicans LMG 10402 | 1 | | 68.6 | 71.5 | 71.1 | 71.0 | 71.4 | 71.5 | 72.5 | 70.6 | 70.9 | 67.5 |
| F. aquatile LMG 4008 | 2 | 68.6 | | 67.7 | 67.5 | 67.9 | 67.9 | 67.7 | 67.6 | 67.0 | 67.1 | 69.1 |
| F. chilense LM-09-FpT | 3 | 71.8 | 67.6 | | 81.0 | 71.3 | 84.2 | 86.1 | 78.2 | 83.1 | 82.3 | 67.7 |
| F. chungangense CJ7T | 4 | 71.1 | 67.5 | 80.7 | | 71.8 | 81.8 | 81.7 | 76.8 | 80.0 | 80.9 | 67.6 |
| F. daejeonense DSM 17708 | 5 | 71.1 | 67.9 | 71.1 | 71.9 | | 71.4 | 70.7 | 71.4 | 71.2 | 71.3 | 68.0 |
| F. denitrificans ED5T | 6 | 71.6 | 67.9 | 84.0 | 81.9 | 71.6 | | 83.6 | 77.9 | 84.8 | 82.6 | 68.4 |
| F. hibernum ATCC 51468T | 7 | 71.6 | 67.5 | 85.9 | 81.7 | 70.6 | 83.6 | | 79.7 | 82.6 | 82.3 | 67.9 |
| F. hydatis DSM 2063T | 8 | 72.5 | 67.5 | 78.2 | 76.8 | 71.4 | 77.8 | 79.8 | | 76.7 | 76.1 | 68.0 |
| F. johnsoniae UW101T | 9 | 71.0 | 67.2 | 83.1 | 80.2 | 71.6 | 84.8 | 82.8 | 77.1 | | 82.5 | 67.6 |
| F. reichenbacii WB 3.2-61T | 1 | 71.1 | 67.4 | 82.7 | 81.1 | 71.3 | 82.8 | 82.5 | 76.3 | 82.6 | | 67.2 |
| F. soli DSM 19725 | 1 | 68 | 69 | 67 | 68 | 68 | 68 | 68 | 68 | 67 | 67 | |

B

| Average Amino Acid Identity | | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|
| F. suncheonense DSM 17707 | 1 | | 68.2 | 80.9 | 79.1 | 79.2 | 80.5 |
| F. aquatile LMG 4008 | 2 | 68.1 | | 69.0 | 68.2 | 68.3 | 69.1 |
| F. cauense R2A-7 | 3 | 80.9 | 69.1 | | 82.5 | 84.6 | 88.7 |
| F. enshiense DK69 | 4 | 79.1 | 68.3 | 82.5 | | 85.4 | 82.3 |
| F. limnosediminis JC2902 | 5 | 79.3 | 68.5 | 84.6 | 85.4 | | 84.8 |
| F. saliperosum S13 | 6 | 80.8 | 69.1 | 88.8 | 82.6 | 84.9 | |

Table 8. **AAI values for select members of the *Flavobacterium* genus.** A. AAI values for organisms in cluster 1 compared to *F. aquatile* and within the cluster.1B. AAI values for organisms in cluster 2 compared to *F. aquatile* and within the cluster.

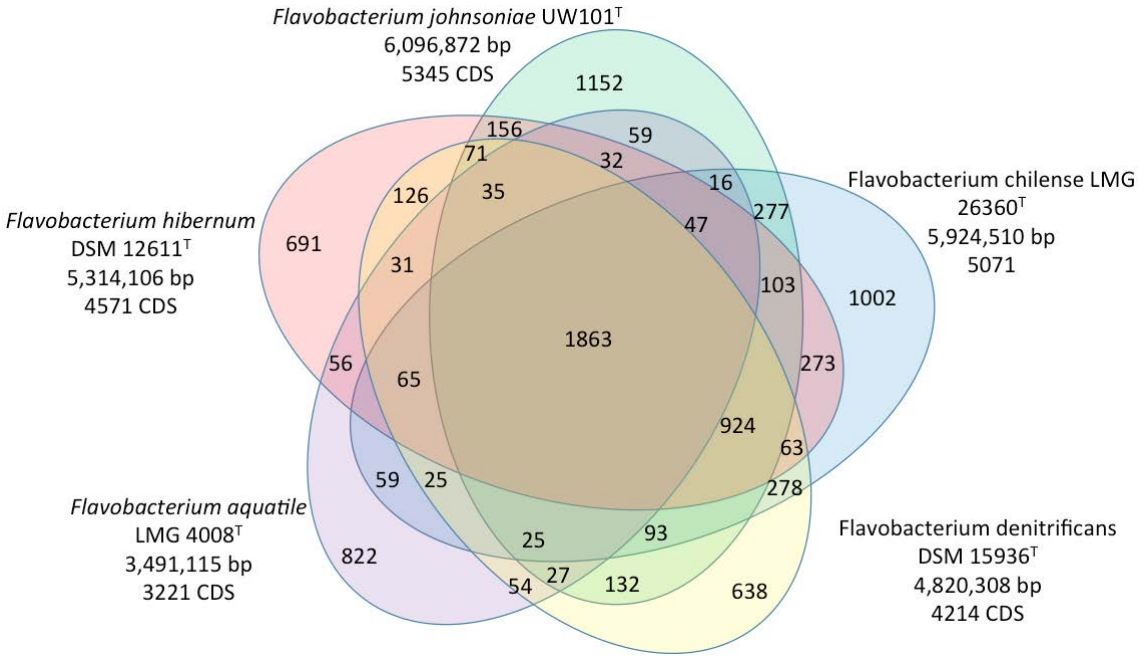


Figure 2. **Genomic similarity at the gene level for *F. hibernum*, *F. johnsoniae*, *F. chilense*, *F. denitrificans*, and *F. aquatile*.** Venn-diagram shows the number of orthologous genes shared within each possible comparison combination.

The core genome of *F. hibernum, F. johnsoniae, F. chilense, F. denitrificans*, and *F. aquatile* consisted of 1,863 genes. *F. hibernum*, *F. johnsoniae*, *F. chilense*, and *F. denitrificans* shared 924 genes (Figure 2). *F. hibernum* contained 691 unique genes with a genome size of 5,314,106 bp and 4,571 CDS. *F. johnsoniae* contained 1,152 unique genes with a genome size of 6,096,872 bp and 5,345 CDS. *F. chilense* contained 1002 unique genes with a genome size of 5,924,510 bp and 5,071 CDS. *F. denitrificans* contained 638 unique genes with a genome size of 4,820,308 bp and 4,214 CDS. *F. aquatile* contained 822 unique genes with a genome size of 3,491,115 bp and 3,221 CDS (Figure 2).

| | *F. reichenbachii* | *F. hibernum* | *F. chungangense* | *F. chiliense* | *F. hydatis* | *F. aquatile* | | *F. reichenbachii* | *F. hibernum* | *F. chungangense* | *F. chiliense* | *F. hydatis* | *F. aquatile* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| neg control | 21 | 23 | 24 | 20 | 15 | 32 | gelatin | 96 | 98 | 93 | 99 | 52 | 100 |
| dextrin | 98 | 98 | 99 | 99 | 82 | 100 | glycyl-L-proline | 90 | 93 | 95 | 96 | 38 | 49 |
| D-maltose | 98 | 98 | 99 | 98 | 37 | 100 | L-alanine | 21 | 54 | 7 | 91 | 39 | 7 |
| D-trehalose | 20 | 69 | 10 | 98 | 48 | 15 | L-arginine | 53 | 75 | 22 | 89 | 42 | 25 |
| D-cellobiose | 99 | 72 | 98 | 99 | 8 | 11 | L-aspartic acid | 84 | 95 | 91 | 96 | 58 | 26 |
| gentiobiose | 99 | 98 | 99 | 98 | 42 | 16 | L-glutamic acid | 98 | 96 | 97 | 96 | 81 | 80 |
| sucrose | 15 | 70 | 97 | 7 | 57 | 100 | L-histidine | 23 | 43 | 23 | 71 | 18 | 7 |
| D-turanose | 39 | 11 | 20 | 9 | 12 | 10 | L-pyroglutamic acid | 31 | 12 | 13 | 11 | 10 | 11 |
| stachyose | 23 | 26 | 11 | 10 | 13 | 14 | L-serine | 37 | 80 | 74 | 90 | 39 | 9 |
| pos control | 98 | 98 | 98 | 97 | 98 | 100 | lincomycin | 18 | 37 | 10 | 12 | 40 | 9 |
| pH 6 | 97 | 97 | 96 | 97 | 98 | 34 | guanidine HCl | 12 | 51 | 9 | 13 | 37 | 12 |
| pH 5 | 12 | 15 | 12 | 94 | 16 | 11 | niaproof 4 | 15 | 13 | 14 | 17 | 17 | 11 |
| D-raffinose | 20 | 70 | 14 | 12 | 26 | 21 | pectin | 54 | 53 | 97 | 91 | 54 | 86 |
| α-D-lactose | 40 | 19 | 19 | 21 | 24 | 19 | D-galacturonic acid | 98 | 96 | 98 | 97 | 38 | 18 |
| D-melibiose | 41 | 13 | 14 | 11 | 11 | 11 | L-galacturonic acid lactone | 9 | 38 | 9 | 11 | 9 | 7 |
| β-methyl-D-glucoside | 24 | 69 | 98 | 17 | 14 | 7 | D-gluconic acid | 23 | 9 | 27 | 14 | 13 | 13 |
| D-salicin | 25 | 69 | 98 | 98 | 13 | 8 | D-glucuronic acid | 85 | 44 | 97 | 54 | 17 | 14 |
| N-acetyl-D-glucosamine | 99 | 96 | 6 | 97 | 72 | 7 | glucuronamide | 38 | 20 | 33 | 30 | 19 | 7 |
| N-acetyl-β-D-mannosamine | 32 | 13 | 10 | 14 | 12 | 7 | mucic acid | 27 | 8 | 15 | 12 | 11 | 7 |
| N-acetyl-D-galactosamine | 27 | 95 | 36 | 97 | 45 | 9 | quinic acid | 20 | 14 | 21 | 14 | 12 | 15 |
| N-acetyl neuraminic acid | 11 | 66 | 7 | 97 | 6 | 13 | D-saccharic acid | 22 | 9 | 14 | 14 | 10 | 16 |
| 1% NaCl | 68 | 72 | 94 | 91 | 47 | 13 | vancomycin | 24 | 94 | 95 | 93 | 98 | 10 |
| 4% NaCl | 13 | 11 | 13 | 15 | 12 | 8 | tetrazolium violet | 62 | 79 | 45 | 98 | 64 | 55 |
| 8% NaCl | 18 | 11 | 16 | 17 | 17 | 15 | tetrazolium blue | 99 | 97 | 95 | 99 | 100 | 38 |
| α-D-glucose | 98 | 97 | 98 | 98 | 56 | 100 | p-hydroxy-phenylacetic acid | 9 | 11 | 9 | 14 | 9 | 7 |
| D-mannose | 98 | 97 | 98 | 98 | 61 | 88 | methyl pyruvate | 38 | 62 | 12 | 84 | 11 | 8 |
| D-fructose | 97 | 93 | 34 | 88 | 72 | 12 | D-lactic acid methyl ester | 39 | 15 | 27 | 19 | 13 | 21 |
| D-galactose | 99 | 97 | 51 | 98 | 72 | 51 | L-lactic acid | 36 | 10 | 19 | 17 | 10 | 8 |
| 3-methyl glucose | 10 | 9 | 13 | 9 | 12 | 7 | citric acid | 31 | 14 | 21 | 16 | 12 | 7 |
| D-fucose | 14 | 9 | 6 | 17 | 9 | 10 | α-keto-glutaric acid | 34 | 10 | 18 | 13 | 12 | 14 |
| L-fucose | 97 | 66 | 68 | 20 | 13 | 9 | D-malic acid | 30 | 12 | 16 | 14 | 10 | 16 |
| L-rhamnose | 96 | 67 | 98 | 97 | 11 | 7 | L-malic acid | 25 | 42 | 15 | 15 | 57 | 18 |
| inosine | 13 | 9 | 7 | 8 | 8 | 12 | bromo-succinic acid | 24 | 10 | 6 | 9 | 9 | 7 |
| 1% Na-lactate | 95 | 92 | 96 | 94 | 28 | 6 | nalidixic acid | 21 | 14 | 12 | 13 | 23 | 98 |
| fusidic acid | 15 | 9 | 9 | 10 | 16 | 10 | LiCl | 13 | 10 | 9 | 12 | 14 | 20 |
| D-serine | 16 | 14 | 10 | 13 | 16 | 16 | K-tellurite | 20 | 19 | 16 | 20 | 23 | 27 |
| D-sorbitol | 18 | 14 | 13 | 12 | 15 | 18 | tween-40 | 52 | 89 | 96 | 96 | 42 | 73 |
| D-mannitol | 23 | 10 | 18 | 11 | 15 | 15 | γ-amino-butyric acid | 9 | 15 | 12 | 16 | 12 | 8 |
| D-arabitol | 22 | 8 | 11 | 11 | 9 | 13 | α-hydroxy-butyric acid | 17 | 12 | 13 | 15 | 10 | 7 |
| myo-inositol | 21 | 8 | 15 | 8 | 10 | 13 | β-hydroxy-D,L-butyric acid | 25 | 12 | 19 | 15 | 11 | 13 |
| glycerol | 20 | 37 | 8 | 9 | 37 | 9 | α-keto-butyric acid | 13 | 7 | 7 | 9 | 5 | 7 |
| D-glucose-6-PO4 | 54 | 48 | 27 | 15 | 23 | 21 | acetoacetic acid | 9 | 50 | 41 | 37 | 16 | 10 |
| D-fructose-6-PO4 | 54 | 44 | 90 | 21 | 30 | 13 | propionic acid | 17 | 9 | 6 | 10 | 9 | 49 |
| D-aspartic acid | 6 | 7 | 6 | 7 | 5 | 7 | acetic acid | 8 | 72 | 89 | 95 | 38 | 97 |
| D-serine | 6 | 6 | 6 | 7 | 6 | 7 | formic acid | 25 | 10 | 8 | 15 | 7 | 15 |
| troleandomycin | 15 | 9 | 9 | 10 | 15 | 6 | aztreonam | 97 | 95 | 98 | 97 | 96 | 100 |
| rifamycin SV | 95 | 95 | 96 | 95 | 96 | 99 | Na-butyrate | 20 | 12 | 15 | 17 | 17 | 26 |
| minocycline | 18 | 13 | 12 | 15 | 18 | 19 | Na bromate | 11 | 11 | 16 | 16 | 15 | 26 |

Figure 3. Biolog GenIII phenotypic tests for *F. reichenbachii*, *F. hibernum F. chungangense*, *F. chilense*, *F. hydatis*, and *F. aquatile.*

Biolog Gen III plates were used to analyze strains of *F. reichenbachii, F. hibernum F. chungangense, F. chilense, F. hydatis,* and *F. aquatile*. Significant differences were noted in a number of growth conditions, carbon source utilization, and inhibitory complexes between *F. aquatile* and the other species. *F. reichenbachii, F. hibernum F. chungangense*, and *F. chilense* showed growth at 1% NaCl while *F. aquatile* exhibited no growth. *F. reichenbachii, F. hibernum F. chungangense*, and *F. chilense* all utilized D-fructose, L-aspartic acid, L-serine, D-galacturonic acid, and tetrazolium blue, while *F. aquatile* was unable to. *F. aquatile* was able to survive in the presence of nalidixic acid while *F. reichenbachii, F. hibernum F. chungangense*, and *F. chilense* were inhibited (Figure 3).

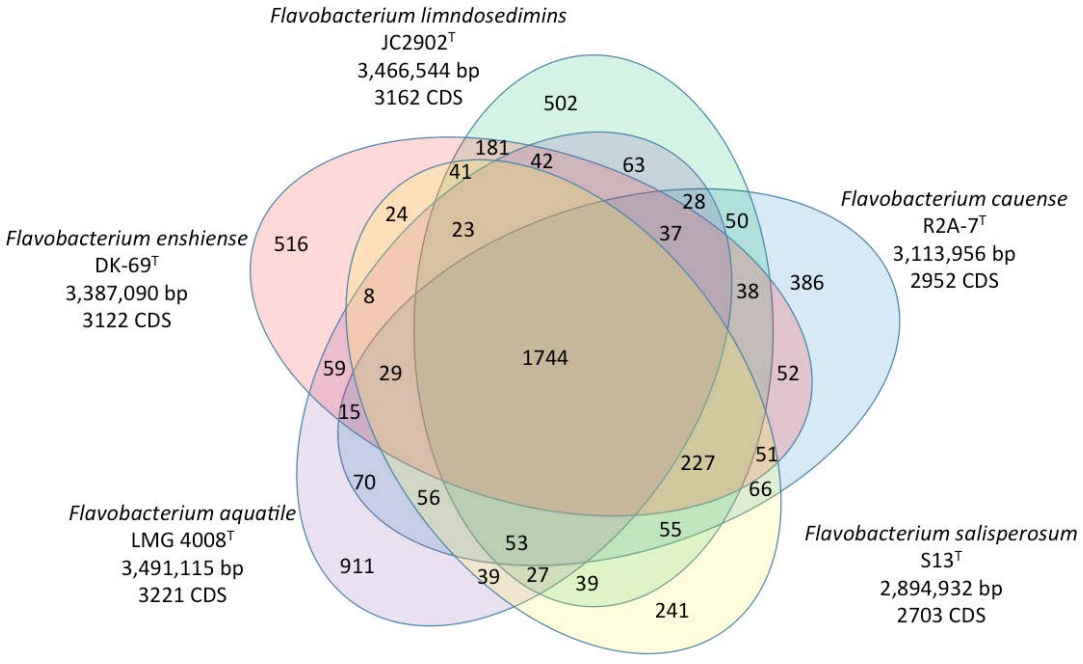


Figure 4. **Genomic similarity at the gene level for F. *enshiense*, *F. limnosediminis*, *F. cauense*, *F. salisperosum*, and *F. aquatile*.** Venn-diagram shows the number of orthologous genes shared within each possible comparison combination.

The core genome of *F. enshiense, F. limnosediminis, F. cauense, F. salisperosum*, and *F. aquatile* consisted of 1744 genes. *F. enshiense, F. limnosediminis, F. cauense*, and *F. salisperosum* shared 277 genes (Figure 4). *F. enshiense* contained 516 unique genes with a genome size of 3,387,090 bp and 3,122 CDS. *F. cauense* contained 386 unique genes with a genome size of 3,113,956 bp and 2,952 CDS. *F. salisperosum* contained 241 unique genes with a genome size of 2,894,932 bp and 2,703 CDS. *F. limnosediminis* contained 502 unique genes with a genome size of 3,466,544 bp and 3,162 CDS. *F. aquatile* contained 911 unique genes with a genome size of 3,491,115 bp and 3221 CDS (Figure 4).

A

| | ROSA sorted | 37752 | 946677 | 1121888 | 376686 | 362418 | 1453505 | 991 | 1450525 | 1121887 | 1121897 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Flavobacterium hibernum DSM 12611 | 37752 | | | | | | | | | | |
| Flavobacterium chilense LMG 26360 | 946677 | 54.5 | | | | | | | | | |
| Flavobacterium denitrificans DSM 15936 | 1121888 | 52.0 | 49.6 | | | | | | | | |
| Flavobacterium johnsoniae UW101 | 376686 | 49.4 | 51.4 | 51.9 | | | | | | | |
| Flavobacterium reichenbachii LMG 25512 | 362418 | 48.0 | 48.1 | 49.8 | 50.0 | | | | | | |
| Flavobacterium chungangense LMG 26729 | 1453505 | 45.3 | 43.3 | 46.4 | 44.1 | 46.3 | | | | | |
| Flavobacterium hydatis | 991 | 42.3 | 41.8 | 39.3 | 38.7 | 36.8 | 35.5 | | | | |
| Flavobacterium succinicans LMG 10402 | 1450525 | 33.1 | 32.1 | 33.5 | 30.6 | 30.3 | 31.0 | 32.6 | | | |
| Flavobacterium daejeonense DSM 17708 | 1121887 | 30.2 | 28.5 | 31.5 | 29.4 | 29.3 | 32.0 | 26.9 | 30.5 | | |
| Flavobacterium soli DSM 19725 | 1121897 | 27.8 | 26.7 | 28.8 | 26.5 | 26.9 | 26.9 | 26.4 | 28.8 | 25.8 | |
| Flavobacterium aquatile LMG 4008 | 1453498 | 27.0 | 25. | 27.5 | 25.4 | 26.0 | 25.6 | 26.2 | 30.5 | 25.7 | 32.2 |

B

| | ROSA sorted | 1341155 | 1341154 | 1341181 | 1121899 | 1107311 | 1453498 |
|---|---|---|---|---|---|---|---|
| Flavobacterium saliperosum S13 | 1341155 | | | | | | |
| Flavobacterium cauense R2A-7 | 1341154 | 66.4 | | | | | |
| Flavobacterium limnosediminis JC2902 | 1341181 | 56.7 | 55.3 | | | | |
| Flavobacterium suncheonense DSM 17707 | 1121899 | 53.6 | 52.7 | 47.7 | | | |
| Flavobacterium enshiense DK69 | 1107311 | 53.0 | 52.1 | 57.6 | 47.2 | | |
| Flavobacterium aquatile LMG 4008 | 1453498 | 34.0 | 33.5 | 31.6 | 32.4 | 30.5 | |

ROSA values for members of cluster 1, as well as closely related organisms by 16s rRNA and aquatile were acquired using full genome sequences. ROSA values from within the cluster ranged from 35.542 and 54.551. Compared to *F. aquatile*, ROSA values compared to members of the cluster ranged from 25.487 to 32.218 (Table 9A). ROSA

values from members within cluster 2 ranged from 47.206 to 66.422, while when compared to *F. aquatile* values ranged from 34.025 and 30.578 (Table 9B). ROSA values indicate that both cluster 1 and cluster 2 are forming separate genera and as such should be split from the *Flavobacterium* genus.

## F. Bacillus

| **Bacillus** | **ROSA** | **280477** | **198094** | **261594** | **412694** | **222523** | **226900** | **315749** | **279010** | **326423** | **692420** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Bacillus anthracis str. Australia 94 | **280477** | | | | | | | | | | |
| Bacillus anthracis str. Ames | **198094** | 96.0 | | | | | | | | | |
| Bacillus anthracis str. 'AmesAncestor' | **261594** | 95.8 | 95.5 | | | | | | | | |
| Bacillus thuringiensis str. AlHakam | **412694** | 82.9 | 84.8 | 81.8 | | | | | | | |
| Bacillus cereus ATCC 10987 | **222523** | 72.9 | 73.6 | 72.4 | 73.7 | | | | | | |
| Bacillus cereus ATCC 14579$^{T}$ | **226900** | 72.1 | 74.2 | 71.6 | 73.6 | 69.2 | | | | | |
| Bacillus cereus subsp. Cytotoxis NVH 391-98 | **315749** | 49.7 | 50.8 | 49.7 | 51.0 | 49.0 | 50.2 | | | | |
| Bacillus licheniformis ATCC 14580$^{T}$ | **279010** | 18.3 | 18.6 | 18.2 | 18.7 | 17.5 | 18.3 | 19.7 | | | |
| Bacillus amyloliquefaciens FZB42 | **326423** | 18.0 | 18.3 | 17.9 | 18.5 | 17.5 | 18.1 | 19.2 | 37.2 | | |
| Bacillus amyloliquefaciens DSM 7$^{T}$ | **692420** | 17.9 | 18.2 | 17.8 | 18.4 | 17.3 | 17.9 | 19.2 | 37.1 | 80.9 | |
| Bacillus subtilis subsp. subtilis str. 168$^{TTTTT}$ | **224308** | 18.0 | 18.4 | 17.9 | 18.5 | 17.3 | 18.0 | 19.4 | 40.0 | 53.3 | 52.4 |

Table 10: ROSA values for members of the *Bacillus* genus. Intraspecies comparisons were done between strains of *B. anthracis*, *B. cereus*, and B. *amyloliquefaciens*. Intragenus comparisons were done between *B. anthracis, B. thuringiensis, B. cereus, B. licheniformis, B. amyloquefaciens,* and *B. subtilis*.

ROSA values for strains belonging to *B. anthracis* had values ranging from 95 to 96 (Table 10). ROSA values for strains belonging to *B. cereus* had values ranging from 49 to 75 (Table 10). ROSA values for strains belonging to *B. amyloliquefaciens* had a score of approximately 81 (Table 10). ROSA values for strains comparing *B. anthracis, B. thuringiensis*, and *B. cereus* had ranges from 49 to 85 (Table 10). Rosa values for strains comparing *B. licheniformis, B. amyloliquefaciens*, and *B. subtilis* ranged between 37 to 52 (Table 10). Comparisons between different members of the genus showed extensive

variation in ROSA values, ranging from ~ 17, below the expected range for members of the same genus, and greater than 65, above the expected range for members of the same genus (Table 10). These ROSA values indicate that *B. cereus, B. thuringiensis, and B. anthracis* belong to the same species with *B. cereus* subsp. cytotoxis being a new species and the cluster a separate genus from the *B. subtilis* cluster.

| | Average Amino Acid Identity (AAIr) | 692420 | 326423 | 261594 | 198094 | 280477 | 222523 | 226900 | 315749 | 279010 | 224308 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Bacillus amyloliquefaciens DSM 7$^{T}$ | 692420 | | 95.6 | 57.7 | 58 | 57.5 | 57.7 | 57.4 | 57.9 | 71.2 | 80.3 |
| Bacillus amyloliquefaciens FZB42 | 326423 | 95.6 | | 58.2 | 58 | 57.9 | 57.6 | 57.7 | 58.4 | 71.6 | 80.3 |
| Bacillus anthracis str. 'AmesAncestor' | 261594 | 57.5 | 58.1 | | 99.9 | 99.9 | 95.0 | 93.3 | 83.0 | 57.6 | 58.1 |
| Bacillus anthracis str. Ames | 198094 | 57.5 | 58.0 | 99.9 | | 99.9 | 95.1 | 93.3 | 82.9 | 57.5 | 58.0 |
| Bacillus anthracis str. Australia 94 | 280477 | 57.3 | 57.9 | 99.9 | 99.9 | | 94.9 | 93.2 | 82.8 | 57.5 | 57.9 |
| Bacillus cereus ATCC 10987 | 222523 | 57.7 | 57.7 | 95.2 | 95 | 95.1 | | 92.9 | 83.3 | 57.5 | 57.7 |
| Bacillus cereus ATCC 14579$^{T}$ | 226900 | 57.4 | 57.5 | 93.4 | 93 | 93.3 | 93.0 | | 83.1 | 57.4 | 57.6 |
| Bacillus cereus subsp. Cytotoxis NVH 391-98 | 315749 | 57.8 | 58.4 | 83.0 | 83 | 82.9 | 83.2 | 83.1 | | 57.8 | 57.9 |
| Bacillus licheniformis ATCC 14580$^{T}$ | 279010 | 71.3 | 71.7 | 57.8 | 58 | 57.8 | 57.5 | 57.4 | 58.1 | | 72.8 |
| Bacillus subtilis subsp. subtilis str. 168$^{T}$ | 224308 | 80.3 | 80.4 | 58.3 | 58 | 58.2 | 57.8 | 57.9 | 58.2 | 72.9 | |
| Bacillus thuringiensis str. AlHakam | 412694 | 57.9 | 58.1 | 97.4 | 98 | 97.5 | 95.2 | 93.4 | 83.3 | 57.5 | 58.0 |

Table 11: AAI values for members in the Bacillus genus. Reciprocal comparisons were done for *B. amyloliquefaciens, B. anthracis, B. cereus, B. licheniformis, B. subtilis*, and *B. thuringiensis*.

AAI values for members of the *Bacillus* genus were calculated compared to other members of the cluster. Distinct clustering was found that was consistent with ROSA value clustering. Members of the same species had AAI values of greater than 95 for all but the different strains of *B. cereus*, with the Cytotoxis strain falling outside expected ranges. *B. anthracis, B. cereus*, and *B. thuringiensis* had high levels of AAI similarity, in some cases above the species threshold (Table 11). B. *amyloquiefaciens, B. licheniformis*, and *B. subtilis* had AAI values similar to organisms found in the same

genus, yet falling outside of the typical genus values compared to other members in the *Bacillus* genus (Table 11). These AAI values indicate that *B. cereus, B. thuringiensis, and B. anthracis* belong to the same species with *B. cereus* subsp. cytotoxis being a new species and the cluster a separate genus from the *B. subtilis* cluster.

| | %BBH | 692420 | 326423 | 261594 | 198094 | 280477 | 222523 | 226900 | 315749 | 279010 | 224308 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Bacillus amyloliquefaciens DSM 7$^{T}$ | 692420 | | 89.1 | 47.4 | 48.0 | 46.5 | 43.5 | 46.8 | 58.6 | 71.6 | 79.2 |
| Bacillus amyloliquefaciens FZB42 | 326423 | 87.9 | | 46.4 | 46.8 | 45.8 | 43.6 | 46.4 | 57.4 | 70.6 | 80.0 |
| Bacillus anthracis str. 'AmesAncestor' | 261594 | 59.9 | 59.4 | | 95.5 | 94.1 | 76.3 | 80.6 | 82.2 | 60.1 | 57.8 |
| Bacillus anthracis str. Ames | 198094 | 61.7 | 61.7 | 95.6 | | 94.5 | 78.0 | 84.3 | 84.6 | 62.1 | 59.8 |
| Bacillus anthracis str. Australia 94 | 280477 | 61.9 | 61.5 | 97.9 | 97.9 | | 79.2 | 83.4 | 84.5 | 62.0 | 59.9 |
| Bacillus cereus ATCC 10987 | 222523 | 60.5 | 61.4 | 83.8 | 84.2 | 82.3 | | 82.4 | 83.8 | 60.3 | 59.1 |
| Bacillus cereus ATCC 14579$^{T}$ | 226900 | 61.7 | 62.8 | 83.8 | 86.0 | 82.2 | 77.8 | | 84.4 | 62.1 | 60.4 |
| Bacillus cereus subsp. Cytotoxis NVH 391-98 | 315749 | 56 | 55 | 62.0 | 62.8 | 60.6 | 57.7 | 61.0 | | 56.3 | 54.8 |
| Bacillus licheniformis ATCC 14580$^{T}$ | 279010 | 75 | 75 | 49.6 | 49.9 | 48.2 | 45.2 | 48.8 | 61.0 | | 75.3 |
| Bacillus subtilis subsp. subtilis str. 168$^{T}$ | 224308 | 83 | 85 | 48.0 | 48.8 | 46.8 | 44.4 | 47.7 | 60.2 | 75.4 | |
| Bacillus thuringiensis str. AlHakam | 412694 | 60 | 61 | 84.7 | 86.8 | 83.5 | 75.9 | 81.2 | 82.4 | 61.0 | 59.1 |

Table 12: %BBH values for members in the *Bacillus* genus. Reciprocal comparisons were done for *B. amyloliquefaciens, B. anthracis, B. cereus, B. licheniformis, B. subtilis*, and *B. thuringiensis*.

%BBH values for members of the *Bacillus* genus showed distinct and highly variable values beyond what would be expected for members of the same genus (Table 12). Strains of the same species had high levels of orthology, >85%, except members of *B. cereus*, who had values in some cases less then 60% orthology (Table 12). *B. anthracis, B. cereus*, and *B. thurgensis* had high levels of ortholgoy, in some cases approaching and even exceeding the suspected species threshold (Table 12). *B. amyloquiefaciens, B. licheniformis,* and *B. subtilis* had ortholgoy values similar to organisms found in the same genus, yet falling outside of the typical genus values compared to other members in the *Bacillus* genus (Table 12).

| | | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|
| Bacillus_anthracis_ATCC_14578T | 1 | | | | | | |
| Bacillus_anthracis_Ames | 2 | 99.9 | | | | | |
| Bacillus_cereus_ATCC_14579T | 3 | 99.9 | 99.8 | | | | |
| Bacillus_thuringiensis_ATCC_10792T | 4 | 99.7 | 99.6 | 99.8 | | | |
| Bacillus_amyloliquefaciens_subsp._amyloliquefaciens_DSM_7T | 5 | 93.8 | 93.9 | 93.7 | 93.6 | | |
| Bacillus_licheniformis_ATCC_14580T | 6 | 94 | 94.1 | 93.9 | 93.9 | 98.2 | |
| Bacillus_subtilis_subsp._subtilis_NCIB_3610T | 7 | 93.9 | 94 | 93.9 | 93.8 | 99.5 | 98.3 |

Table 13: Reciprocal %16s rRNA differences between members of the *Bacillus* genus.

Members of the Bacillus genus showed vastly differing levels of 16s rRNA % similarity in 16s rRNA similarity ranged from .001, or 99.9% similarity, to 0.064, or 94.6% similarity (Table 13). *B. anthracis, B. cereus*, and *B. thuringiensis* had high levels of 16s rRNA similarity, in some cases approaching and even exceeding the suspected species threshold (Table 13). *B. amyloliquefaciens*, *B. licheniformis*, and *B. subtilis* had 16s rRNA similarity values similar to organisms found in the same genus, yet falling outside of the typical genus values compared to other members in the *Bacillus* genus (Table 13).

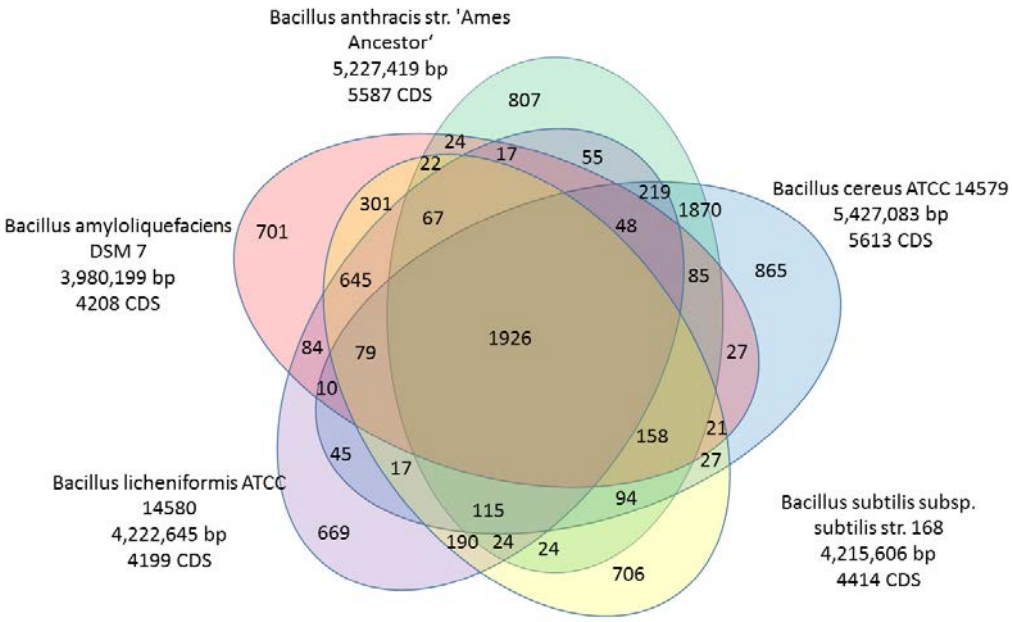


Figure 5: **Genomic similarity at the gene level for *B. licheniformis*, *B. amyloliquefaciens*, *B. anthracis*, *B. cereus*, and *B. subtilis*.** Venn-diagram shows the number of orthologous genes shared within each possible comparison combination.

The core genome of *B. licheniformis, B. amyloliquefaciens, B. anthracis, B. cereus*, and *B. subtilis* consisted of 1,926 genes. *B. licheniformis* had 669 unique genes with a genome size of 4,222,645 bp and 4,199 CDS. *B. amyloliquefaciens* had 701 unique genes with a genome size of 3,980,199 bp and 4,208 CDS. *B. anthracis* had 807 unique genes with a genome size of 5,227,419 bp and 5,587 CDS. *B. cereus* had 865 unique genes with a genome size of 5,427,083 bp and 5,613 CDS. *B. subtilis* had 706 unique genes with a genome size of 4,215,606 bp and 4,414 CDS (Figure 5). *B. licheniformis, B. amyloliquefaciens,* and *B. subtilis* shared a further 645 genes (Figure 5). *B. cereus* and *B. anthracis* shared a further 1,870 genes between them (Figure 5).

**B. anthracis and B. cereus only**

A

652 Spore germination protein GerLA
653 Spore germination protein GerLB
654 Spore germination protein GerLC
655 hypothetical protein
656 Ferrous iron transport protein B
657 Ferrous iron transport protein B
658 Ferrous iron transport protein A
711 Phosphonate ABC transporter phosphate-binding periplasmic component (TC 3.A.1.9.1)
712 Spore germination protein GerYB
713 Spore germination protein GerYC
714 Spore germination protein GerYA
715 Glycerate kinase (EC 2.7.1.31)
948 Chemotaxis protein methyltransferase CheR (EC 2.1.1.80)
949 sensor histidine kinase/response regulator
950 FIG011741: hypothetical protein
951 hypothetical protein
952 FIG005030: hypothetical protein
953 FIG017032: hypothetical protein
954 PlcB, ORFX, ORFP, ORFB, ORFA, ldh gene
955 FIG01226432: hypothetical protein
956 FIG016527: hypothetical protein
1278 Ferric iron ABC transporter, iron-binding protein
1279 Ferric iron ABC transporter, ATP-binding protein
1280 Ferric iron ABC transporter, permease protein
1281 Phosphonoacetaldehyde hydrolase (EC 3.11.1.1)
1285 hypothetical protein
1286 internalin, putative
1287 Methylthioribulose-1-phosphate dehydratase related protein
1288 Methylthioribulose-1-phosphate dehydratase related protein
1289 CBS domain containing protein
1290 Possible divergent polysaccharide deacetylase
1647 HigA protein (antitoxin to HigB)
1648 permease, putative
1649 Phosphonate ABC transporter phosphate-binding periplasmic component (TC 3.A.1.9.1)
1650 Chromosome initiation inhibitor
1651 Ferric iron ABC transporter, ATP-binding protein
1652 Ferric iron ABC transporter, iron-binding protein
2999 FIG01227996: hypothetical protein
3000 FIG01231776: hypothetical protein
3001 FIG01227706: hypothetical protein
3002 Malate:quinone oxidoreductase (EC 1.1.5.4)
3003 Autoinducer 2 (AI-2) ABC transport system, periplasmic AI-2 binding protein LsrB
3004 Autoinducer 2 (AI-2) ABC transport system, membrane channel protein LsrD
3005 Autoinducer 2 (AI-2) ABC transport system, membrane channel protein LsrC
3006 Autoinducer 2 (AI-2) ABC transport system, fused AI2 transporter subunits and ATP-binding component
3374 Phosphonate ABC transporter phosphate-binding periplasmic component (TC 3.A.1.9.1)
33753-oxoacyl-[acyl-carrier protein] reductase paralog (EC 1.1.1.100)
3376 ABC transporter permease protein
3377 ABC transporter, ATP-binding protein
3378 FIG01226010: hypothetical protein
3379 FIG01225310: hypothetical protein
3381 Glutamate synthase [NADPH] large chain (EC 1.4.1.13)
33825'-nucleotidase YjjG (EC 3.1.3.5)
3383 Possible protein-tyrosine-phosphatase (EC 3.1.3.48)
3385 Protein export cytoplasm protein SecA ATPase RNA helicase (TC 3.A.5.1.1)
3386 FIG01225608: hypothetical protein
3387 Phosphohydrolase, MutT/Nudix family
3388 Tryptophanyl-tRNA synthetase (EC 6.1.1.2)
33891-deoxy-D-xylulose 5-phosphate reductoisomerase (EC 1.1.1.267)
3391 Integral membrane protein
3392 acetyltransferase, GNAT family
3393 hypothetical protein
3396 Ubiquinone/menaquinone biosynthesis methyltransferase UBIE (EC 2.1.1.-)
3397 FIG01227072: hypothetical protein
3398 membrane protein, putative
3399 TPR domain protein
3400 Phosphonate ABC transporter phosphate-binding periplasmic component (TC 3.A.1.9.1)

**B. subtilis, B. lichenformis and B. amyloliquefaciens only**

B

378 Hydroxyaromatic non-oxidative decarboxylase protein B (EC 4.1.1.-)
379 Hydroxyaromatic non-oxidative decarboxylase protein C (EC 4.1.1.-)
380 Hydroxyaromatic non-oxidative decarboxylase protein D (EC 4.1.1.-)
381 Hydroxyaromatic non-oxidative decarboxylase protein D (EC 4.1.1.-)
384 FIG01230034: hypothetical protein
385 Uncharacterized protein yczF precursor
386 Spore germination protein GerKA
387 Spore germination protein GerKC
388 Spore germination protein GerKB
389 ABC transporter ATP-binding protein YvcR
390 ABC transporter, permease protein
416 PTS system, mannitol-specific IIC component (EC 2.7.1.69)
417 PTS system, mannitol-specific IIA component (EC 2.7.1.69)
418 Mannitol-1-phosphate 5-dehydrogenase (EC 1.1.1.17)
419 Tartrate dehydrogenase (EC 1.1.1.93) /
1349 phage-like element PBSX protein XkdA
1352 Phage-like element PBSX protein xkdB
1353 DNA replication protein DnaC
1354 FIG01227758: hypothetical protein
1355 XkdD
1356 Phage-like element PBSX protein xtrA
1358 Phage terminase small subunit
1359 Phage terminase, large subunit [SA bacteriophages 11, Mu50B]
1360 Phage-like element PBSX protein xkdE
1361 FIG01230357: hypothetical protein
1362 Phage-like element PBSX protein xkdG
1365 Lin1275 protein
1366 Phage-like element PBSX protein xkdJ
1368 Phage-like element PBSX protein xkdK
1369 Phage tail fibers
1370 Phage-like element PBSX protein xkdN
1373 Phage-like element PBSX protein xkdP
1374 Phage-like element PBSX protein xkdQ
1375 FIG01234021: hypothetical protein
1757 Flagellar hook-basal body complex protein FliE
1760 Flagellar assembly protein FliH
1762 Flagellar protein FliJ
1763 Flagellar protein FlbB
1764 Flagellar hook-length control protein FliK
1767 Flagellar protein FlbD
1768 Flagellar biosynthesis protein FliL
1772 Flagellar biosynthesis protein FliZ
1779 Flagellar synthesis regulator FleN
1780 Chemotaxis response regulator protein-glutamate methylesterase CheB (EC 3.1.1.61)
1782 Positive regulator of CheA protein activity (CheW)
1783 Chemotaxis protein CheC -- inhibitor of MCP methylation
1784 Chemotaxis protein CheD
1785 RNA polymerase sigma factor for flagellar operon
1786 FIG01231449: hypothetical protein
3390 Na(+) H(+) antiporter subunit B
3391 Na(+) H(+) antiporter subunit C
3393 Na(+) H(+) antiporter subunit E
3394 Na(+) H(+) antiporter subunit F
3395 Na(+) H(+) antiporter subunit G

Table 14: A. Selected unique genes shared between *B. anthracis* and *B. cereus*. Selected unique genes shared between *B. subtilis*, *B. licheniformis*, and *B. amyloliquefaciens.*

Some unique shared genes and operons found between *B. anthracis* and *B. cereus* are different sporeulation proteins, different iron transporters, chemotaxing proteins, malate catabolism operons, some specialized membrane proteins, and some antibiotic resistance proteins (Table 14A). Some unique shared genes and operons

found between *B. subtilis, B. licheniformis*, an *B. amyloliquefaciens* are different non-oxidative carboxylase proteins, sporulation proteins, mannitol catabolism proteins, unique phage genomic DNA, flagellar proteins, and multiple Na/H transport proteins (Table 14B). *B. anthracis* and *B. cereus* had similar types of unique genes. Genes that were not shared between the two consisted primarily of captured phage genes and a few specialized pathogenicity islands and secretory compounds (Data not shown).

## G. *Chlamydiacae*

| | ROSA | 115713 | 138677 | 182082 | 115711 | 218497 | 264202 | 227941 | 243161 | 471473 | 315277 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Chlamydophila pneumoniae CWL029 | **115713** | | | | | | | | | | |
| Chlamydophila pneumoniae J138 | **138677** | 99.0 | | | | | | | | | |
| Chlamydophila pneumoniae TW-183 | **182082** | 98.9 | 98.7 | | | | | | | | |
| Chlamydophila pneumoniae AR39 | **115711** | 97.8 | 98.0 | 97.4 | | | | | | | |
| Chlamydophila abortus S26/3 | **218497** | 40.6 | 40.6 | 40.5 | 40.4 | | | | | | |
| Chlamydophila felis Fe/C-56 | **264202** | 40.0 | 39.9 | 39.9 | 39.8 | 66.7 | | | | | |
| Chlamydophila caviae GPIC | **227941** | 39.8 | 39.8 | 39.7 | 39.6 | 67.5 | 68.9 | | | | |
| Chlamydia muridarum Nigg | **243161** | 32.6 | 32.6 | 32.5 | 32.5 | 36.7 | 36.6 | 36.5 | | | |
| Chlamydia trachomatis L2b/UCH-1/proctitis | **471473** | 33.1 | 33.1 | 33.0 | 32.9 | 37.5 | 36.8 | 36.7 | 69.1 | | |
| Chlamydia trachomatis A/HAR-13 | **315277** | 32.9 | 32.8 | 32.8 | 32.7 | 37.3 | 36.9 | 36.8 | 69.2 | 95.4 | |
| Chlamydia trachomatis D/UW-3/CX | **272561** | 23.8 | 23.7 | 23.7 | 23.7 | 27.4 | 26.7 | 26.6 | 55.6 | 95.1 | 96.9 |

Table 15: ROSA values for the *Chlamydiacae* family. Intraspecies comparisons were done for *Chlamydia trachomatis* and *Chlamydophila pneumoniae*. Intragenus comparisons were done between *Chlamydia trachomatis* and *Chlamydia muridarum*. Intragenus comparisons were done between *Chlamydophila pneumoniae, Chlamydophila abortus, Chlamydophila felis*, and *Chlamydophila caviae*. Intrafamily comparisons were done between *Chlamydia trachomatis, Chlamydia muridarum*, *Chlamydophila pneumoniae, Chlamydophila abortus, Chlamydophila felis*, and *Chlamydophila caviae*.

Intraspecies comparisons between strains of *Chlamydophila pneumoniae* yielded ROSA values between 97.4 and 99.0 (Table 15). Intraspecies comparisons between strains of *Chlamydia trachomatis* yielded ROSA values between 69.1 and 96.9 (Table 15). Intragenus comparisons between *Chlamydophila pneumoniae*, *Chlamydophila abortus,*

*Chlamydophila felis*, and *Chlamydophila caviae* yielded ROSA values ranging from 39.7 to 40.6 (Table 15). Intragenus comparisons between *Chlamydia trachomatis* and *Chlamydia muridarum* yielded ROSA values ranging from 26.6 and 36.8 (Table 15). Intrafamily comparisons were done between *Chlamydia trachomatis, Chlamydia muridarum, Chlamydophila pneumoniae, Chlamydophila abortus, Chlamydophila felis,* and *Chlamydophila caviae* yielded ROSA values ranging from 26.6 and 36.6 (Table 15). ROSA values suggest a third possible genus consisting of *C. abortus*, *C. felis*, and *C. caviae*, as well as supporting the *Chlamydophila* and *Chlamydia* split.

| | | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| Chlamydophila_pneumoniae_TW-183T | 1 | | | | | |
| Chlamydophila_abortus_S26 | 2 | 95.8 | | | | |
| Chlamydophila_felis_Fe/C-56 | 3 | 95.2 | 98.3 | | | |
| Chlamydophila_caviae_GPICT | 4 | 95.5 | 99.3 | 98.5 | | |
| Chlamydia_muridarum_Nigg | 5 | 94.4 | 95.6 | 95.7 | 95.9 | |
| Chlamydia_trachomatis_A | 6 | 93.8 | 95.0 | 94.9 | 95.2 | 98.4 |

Table 16: 16s rRNA % similarity for members of the *Chlamydiacae* family.

Members of the *Chlamydiacae* family showed distinctive clustering according to 16s rRNA (Table 16). Members of the *Chlamydophila* genus clustered highly with one another, 95.5 to 99.3, with *C. abortus, C. felis*, and *C. caviae* having 16s rRNA similarities nearing the cutoff for same species, 98.3 to 99.3 (Table 16). Members of the *Chlamydia* genus clustered with each other by 16s rRNA similarity and separately from members of the *Chlamydophila* genus, 98.4 and 93.8 to 95.9 respectively (Table 16).

| | Average Amino Acid Identity (AAIr) | 243161 | 315277 | 272561 | 471473 | 218497 | 227941 | 264202 | 115711 | 115713 | 138677 | 182082 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Chlamydia muridarum Nigg | **243161** | | 85.0 | 65.8 | 85.3 | 64.2 | 64.4 | 64.3 | 62.3 | 62.3 | 62.1 | 62.3 |
| Chlamydia trachomatis A/HAR-13 | **315277** | 84.9 | | 98.6 | 98.7 | 64.1 | 64.1 | 64.0 | 61.8 | 61.8 | 61.8 | 61.9 |
| Chlamydia trachomatis D/UW-3/CX | **272561** | 84.9 | 99.3 | | 98.7 | 64.1 | 64.1 | 64.0 | 61.7 | 62.0 | 61.7 | 61.9 |
| Chlamydia trachomatis L2b/UCH-1/proctitis | **471473** | 85.2 | 98.6 | 97.2 | | 64.3 | 64.3 | 64.3 | 61.9 | 61.9 | 61.9 | 61.9 |
| Chlamydophila abortus S26/3 | **218497** | 64.3 | 64.2 | 64.2 | 64.4 | | 84.2 | 84.0 | 67.7 | 67.8 | 67.5 | 68.0 |
| Chlamydophila caviae GPIC | **227941** | 64.5 | 64.2 | 64.3 | 64.4 | 84.2 | | 84.9 | 67.7 | 68.1 | 67.7 | 68.2 |
| Chlamydophila felis Fe/C-56 | **264202** | 64.4 | 64.2 | 64.3 | 64.4 | 83.9 | 84.7 | | 67.6 | 67.9 | 67.6 | 67.9 |
| Chlamydophila pneumoniae AR39 | **115711** | 62.1 | 61.8 | 61.7 | 61.8 | 67.4 | 67.6 | 67.5 | | 99.9 | 99.9 | 99.8 |
| Chlamydophila pneumoniae CWL029 | **115713** | 62.1 | 61.8 | 61.9 | 61.8 | 67.3 | 67.7 | 67.5 | 99.9 | | 99.9 | 99.9 |
| Chlamydophila pneumoniae J138 | **138677** | 62.1 | 61.7 | 61.7 | 61.8 | 67.3 | 67.5 | 67.5 | 99.9 | 99.9 | | 99.8 |
| Chlamydophila pneumoniae TW-183 | **182082** | 62.1 | 61.7 | 61.8 | 61.8 | 67.5 | 67.9 | 67.6 | 99.8 | 99.9 | 99.8 | |

Table 17: AAI values for members of the *Chlamydiacae* family.

AAI values for the *Chlamydiacae* family showed distinct clustering at both the genus and the species level. Members of the *Chlamydia* genus had AAI values between 65.8 and 85.3 (Table 17). Strains *of C. trachomatis* had AAI values ranging from 98.6 to 99.3 (Table 17). Members of the *Chlamydophila* genus had AAI values ranging from 67.6 to 85 (Table 17). Strains of *C. pneumoniae* had AAI values of 99.8 to 99.9 (Table 17). *C. abortus, C. caviae*, and *C. felis* had distinct clustering to each other with values ranging from 83.9 to 84.9 (Table 17). Comparisons of members of the *Chlamydia* genus to members of the *Chlamydophila* genus had values ranging from 61.7 to 64.4 (Table 17).

| | Percent Bidirectional Best Hit (% BBH) | 243161 | 315277 | 272561 | 471473 | 218497 | 227941 | 264202 | 115711 | 115713 | 138677 | 182082 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Chlamydia muridarum Nigg | **243161** | | 96.5 | 98.6 | 96.3 | 87.4 | 83.7 | 84.4 | 78.5 | 78.5 | 78.8 | 78.0 |
| Chlamydia trachomatis A/HAR-13 | **315277** | 95.5 | | 99.9 | 98.5 | 88.5 | 84.8 | 85.5 | 79.7 | 79.7 | 79.7 | 79.0 |
| Chlamydia trachomatis D/UW-3/CX | **272561** | 94.8 | 97.9 | | 98.7 | 88.5 | 84.1 | 85.0 | 80.0 | 79.3 | 79.9 | 79.0 |
| Chlamydia trachomatis L2b/UCH-1/proctitis | **471473** | 93.7 | 97.5 | 99.6 | | 88.0 | 83.1 | 83.9 | 79.4 | 79.4 | 79.5 | 79.1 |
| Chlamydophila abortus S26/3 | **218497** | 90.0 | 92.7 | 44.6 | 93.4 | | 91.8 | 91.7 | 84.3 | 83.9 | 84.6 | 83.1 |
| Chlamydophila caviae GPIC | **227941** | 92.1 | 94.0 | 45.0 | 94.3 | 98.4 | | 95.8 | 85.2 | 84.3 | 85.4 | 83.6 |
| Chlamydophila felis Fe/C-56 | **264202** | 92.0 | 93.7 | 44.9 | 94.1 | 97.6 | 95.8 | | 85.5 | 84.7 | 85.5 | 84.0 |
| Chlamydophila pneumoniae AR39 | **115711** | 89.5 | 91.7 | 44.3 | 92.9 | 93.1 | 87.9 | 88.9 | | 97.5 | 97.9 | 97.1 |
| Chlamydophila pneumoniae CWL029 | **115713** | 90.3 | 92.6 | 44.6 | 93.8 | 94.2 | 88.4 | 89.8 | 98.7 | | 99.5 | 98.9 |
| Chlamydophila pneumoniae J138 | **138677** | 90.2 | 92.5 | 44.7 | 93.8 | 94.1 | 88.9 | 89.6 | 98.5 | 99.1 | | 98.7 |
| Chlamydophila pneumoniae TW-183 | **182082** | 90.3 | 92.7 | 44.5 | 93.7 | 93.7 | 87.8 | 89.7 | 98.4 | 99.1 | 99.3 | |

Table 18: %BBH for members of the *Chlamydiacae* family.

Members of the *Chlamydiacae* family showed distinct clustering in orthology. Strains of *C. pneumoniae* shared over 98% of their genome (Table 18). Members of the *Chlamydophila* genus shared between 83.1% to 98.4% of their genome. In particular *C. abortus, C. caviae*, and *C. felis* shared over 90% of their genome when compared to each other (Table 18). Strains of *C. trachomatis* shared over 98% of their genome and over 93% with *C. muridarum* (Table 18). Percentage of genome conserved between members of the two genera ranged from 44.5% to over 90%; strain UW-3 of *C. trachomatis* had particularly low conservation with members of *Chlamydophila* with its orthology being below 45% (Table 18).

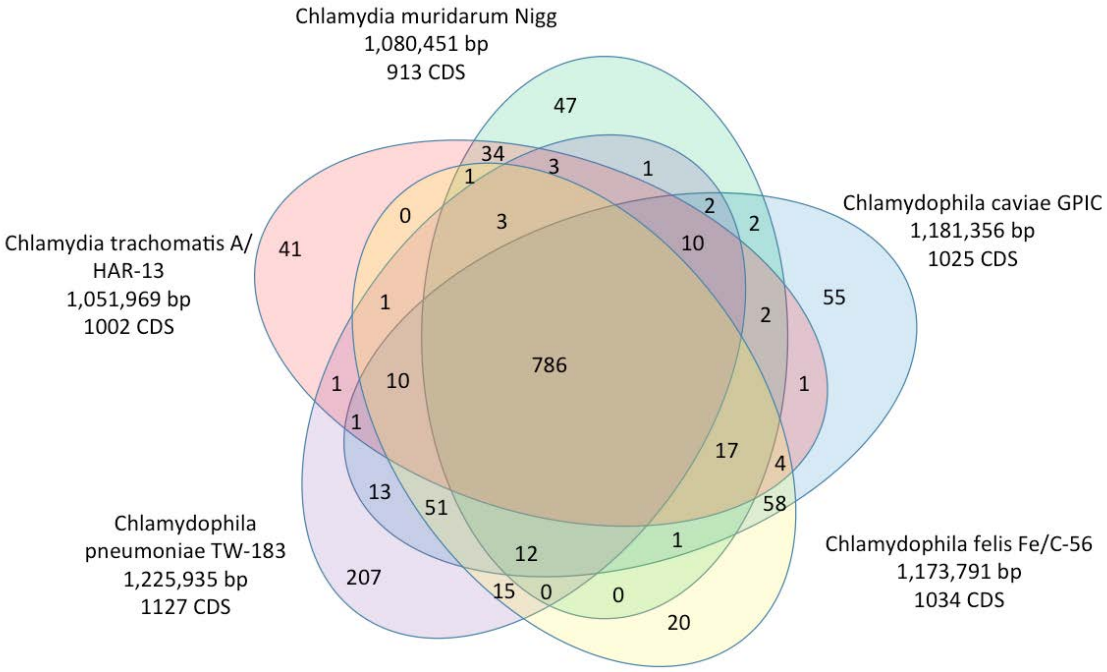


Figure 6: **Genomic similarity at the gene level for *C. pneumoniae, C. trachomatis, C. muridarum, C. caviae*, and *C. felis*.** Venn-diagram shows the number of orthologous genes shared within each possible comparison combination.

The core genome of *C. pneumoniae, C. trachomatis, C. muridarum, C. caviae*, *C. felis* consisted of 786 genes. *C. pneumoniae* had 206 unique genes with a genome size of 1,225,935 bp and 1127 CDS. *C. trachomatis* had 41 unique genes with a genome size of 1,051,969 bp and 1002 CDS. *C. muridarum* had 47 unique genes with a genome size of 1,090,451 bp and 913 CDS. *C. caviae* had 55 unique genes with a genome size 1,181,356 bp and 1025 CDS. *C. felis* had 20 unique genes with a genome size of 1,173,791 bp and 1034 CDS (Figure 6). *C. trachomatis* and *C. muridarum* shared a further 34 genes uniquely to each other (Figure 6). These genes encoded for hypothetical proteins and specialized membrane proteins. *C. caviae*, and *C. felis* shared a further 58 genes uniquely to their genus (Figure 6). These genes encoded for a few specialized transport and adhesion proteins.

## H. *Helicobacter*

| ROSA (sorted) | | 102618 | 563041 | 570508 | 85962 | 512562 | 85963 | 382638 | 936155 | 217 | 235279 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Helicobacter pylori NCTC 11637 | **102618** | | | | | | | | | | |
| Helicobacter pylori G27 | **563041** | 86.8 | | | | | | | | | |
| Helicobacter pylori P12 | **570508** | 83.1 | 84.5 | | | | | | | | |
| Helicobacter pylori 26695 | **85962** | 83.6 | 84.4 | 84.9 | | | | | | | |
| Helicobacter pylori Shi470 | **512562** | 82.3 | 83.7 | 81.5 | 81.9 | | | | | | |
| Helicobacter pylori J99 | **85963** | 81.6 | 82.5 | 82.4 | 82.1 | 81.7 | | | | | |
| Helicobacter acinonychis str. Sheeba | **382638** | 70.0 | 70.5 | 69.6 | 70.1 | 71.0 | 70.3 | | | | |
| Helicobacter felis CS1, ATCC 49179 | **936155** | 23.9 | 24.2 | 24.1 | 24.3 | 24.7 | 24.5 | 25.1 | | | |
| Helicobacter mustelae 43772 | **217** | 19.9 | 20.2 | 20.1 | 20.2 | 20.5 | 20.4 | 20.8 | 19.5 | | |
| Helicobacter hepaticus ATCC 51449 | **235279** | 17.3 | 17.4 | 17.3 | 17.4 | 17.6 | 17.5 | 17.8 | 17.0 | 20.7 | |
| Helicobacter canadensis MIT 98-5491 | **537970** | 16.2 | 16.2 | 16.1 | 16.2 | 16.5 | 16.3 | 16.6 | 15.5 | 19.1 | 21.4 |

Table 19: ROSA values for members of the *Helicobacter* family. Intraspecies comparisons were done for strains of *H. pylori*. Intragenus comparisons were done between *H. pylori, H. acinonychis, H. felis, H. mustelae, H. hepaticus*, and *H. canadensis*.

Intraspecies comparisons for strains of *H. pylori* had ROSA values between 81.6 and 86.8 (Table 19). Intragenus comparisons between members of the *Helicobacter* genus had ROSA values ranging from 16.2 to 70.5 (Table 19). Members of *H. pylori* and *H. acinonychis* formed a distinct cluster from other members of the *Helicobacter* genus. Values between the two organisms ranged between 69.6 and 70.5 (Table 19). Other members of the genus had values ranging from 16.2 and 25.1 and formed no distinct clustering with other members (Table 19). ROSA values suggest that *H. pylori* and *H. acinonychis* are members of the same species. *H. felis*, *H. mustelae*, H. hepatics, and *H. canadensis* are all forming separate genera from both each other and from *H. pylori* and *H. acinonychis*. ROSA values suggest that the *Helicobacter* genus be split into 5 genera.

| | | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| Helicobacter_pylori_NCTC_11638T | 1 | | | | | |
| Helicobacter_acinonychis_Sheeba | 2 | 98.2 | | | | |
| Helicobacter_felis_ATCC_49179T | 3 | 95.5 | 95.0 | | | |
| Helicobacter_mustelae_ATCC_43772T | 4 | 94.1 | 93.6 | 93.4 | | |
| Helicobacter_hepaticus_Hh-2T | 5 | 93.7 | 92.8 | 93.2 | 96.3 | |
| Helicobacter_canadensis_MIT_98-5491T | 6 | 94.3 | 93.7 | 93.4 | 95.9 | 97.4 |

Table 20: 16s rRNA % similarities for members for the *Helicobacter* genus.

16s rRNA values showed on strong clustering of organisms and second possible cluster (Table 20). *H. pylori* and *H. acinonychis* formed a cluster with a 16s rRNA % similarity of 98.2% (Table 20). A possible second cluster was formed between *H. mustelae, H hepaticus*, and *H. canadensis* with 16s rRNA values ranging from 95.9 and 97.4 (Table 20). *H. felis* showed little clustering with other organisms with its 16s rRNA values ranging from 92.8 to 95.5 to 93.2 (Table 20).

| | Average Amino Acid Identity (AAIr) | 512562 | 537970 | 936155 | 102618 | 382638 | 235279 | 217 | 85962 | 563041 | 85963 | 570508 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Helicobacter pylori Shi470 | **512562** | | 51.0 | 58.5 | 94.9 | 90.3 | 52.5 | 54.6 | 94.5 | 94.9 | 93.4 | 94.4 |
| Helicobacter canadensis MIT 98-5491 | **537970** | 50.9 | | 50.5 | 51.0 | 51.0 | 54.7 | 51.9 | 50.9 | 50.8 | 50.8 | 50.9 |
| Helicobacter felis CS1, ATCC 49179 | **936155** | 58.3 | 50.6 | | 57.8 | 58.2 | 51.6 | 54.0 | 58.1 | 58.1 | 57.9 | 57.9 |
| Helicobacter pylori NCTC 11637 | **102618** | 95.0 | 51.6 | 58.3 | | 90.2 | 52.6 | 54.5 | 95.8 | 96.6 | 94.5 | 95.9 |
| Helicobacter acinonychis str. Sheeba | **382638** | 89.7 | 51.0 | 58.0 | 89.7 | | 52.6 | 54.1 | 89.8 | 90.0 | 89.4 | 89.8 |
| Helicobacter hepaticus ATCC 51449 | **235279** | 52.6 | 54.8 | 51.5 | 52.2 | 52.7 | | 54.7 | 52.6 | 52.5 | 52.4 | 52.6 |
| Helicobacter mustelae 43772 | **217** | 54.8 | 52.4 | 54.1 | 54.3 | 54.0 | 54.8 | | 54.4 | 54.5 | 54.4 | 54.8 |
| Helicobacter pylori 26695 | **85962** | 94.5 | 51.1 | 58.1 | 95.8 | 90.3 | 52.5 | 54.1 | | 95.6 | 94.4 | 95.7 |
| Helicobacter pylori G27 | **563041** | 94.9 | 51.2 | 58.5 | 96.5 | 90.4 | 52.6 | 54.1 | 95.5 | | 94.1 | 95.5 |
| Helicobacter pylori J99 | **85963** | 93.4 | 51.1 | 58.1 | 94.6 | 90.2 | 52.6 | 54.5 | 94.3 | 94.2 | | 94.5 |
| Helicobacter pylori P12 | **570508** | 94.3 | 51.1 | 58.2 | 95.8 | 90.3 | 52.7 | 54.4 | 95.7 | 95.4 | 94.4 | |

Table 21: AAI values for members of the *Helicobacter* genus.

AAI values for members of the *Helicobacter* genus showed great variation with only a single distinct cluster emerging (Table 21). Strains of *H. pylori* had AAI values

ranging from 94.3 and 95.6 (Table 21). In addition *H. acinonychis* had AAI values ranging from 89.7 and 90 to members of *H. pylori* (Table 21). Other members of the genus had no visible clustering between them and AAI values ranging from 50.5 and 58.5 when compared to other members of the genus (Table 21).

| | Percent Bidirectional Best Hit (% BBH) | 512562 | 537970 | 936155 | 102618 | 382638 | 235279 | 217 | 85962 | 563041 | 85963 | 570508 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Helicobacter pylori Shi470 | **512562** | | 62.3 | 70.1 | 92.2 | 88.3 | 60.1 | 68.6 | 90.2 | 90.9 | 93.1 | 89.1 |
| Helicobacter canadensis MIT 98-5491 | **537970** | 64.5 | | 60.1 | 63.1 | 66.1 | 68.5 | 71.4 | 62.5 | 62.3 | 63.8 | 61.8 |
| Helicobacter felis CS1, ATCC 49179 | **936155** | 74.3 | 61.3 | | 73.8 | 77.9 | 62.0 | 68.8 | 72.9 | 72.4 | 74.8 | 71.8 |
| Helicobacter pylori NCTC 11637 | **102618** | 90.4 | 59.9 | 68.3 | | 87.0 | 58.8 | 66.5 | 89.2 | 90.8 | 90.5 | 87.6 |
| Helicobacter acinonychis str. Sheeba | **382638** | 86.8 | 61.7 | 70.8 | 86.2 | | 59.4 | 69.7 | 84.2 | 84.4 | 86.5 | 83.3 |
| Helicobacter hepaticus ATCC 51449 | **235279** | 67.4 | 74.4 | 65.9 | 67.4 | 69.2 | | 73.1 | 65.3 | 65.4 | 66.9 | 64.8 |
| Helicobacter mustelae 43772 | **217** | 68.4 | 68.5 | 64.8 | 67.8 | 72.5 | 65.0 | | 67.2 | 66.9 | 68.6 | 65.8 |
| Helicobacter pylori 26695 | **85962** | 93.1 | 62.4 | 70.9 | 93.0 | 88.7 | 60.6 | 70.3 | | 92.2 | 93.5 | 92.0 |
| Helicobacter pylori G27 | **563041** | 94.8 | 62.0 | 69.9 | 95.5 | 88.9 | 60.4 | 69.8 | 92.9 | | 94.6 | 92.0 |
| Helicobacter pylori J99 | **85963** | 94.1 | 62.0 | 70.8 | 92.2 | 88.0 | 60.1 | 69.0 | 90.9 | 91.6 | | 90.3 |
| Helicobacter pylori P12 | **570508** | 94.0 | 62.2 | 70.8 | 93.1 | 88.3 | 60.0 | 69.0 | 93.4 | 93.5 | 94.5 | |

Table 22: %BBH values for members of the *Helicobacter* genus.

Members of the *Helicobacter* genus had high and low levels of orthology in their genomes with only one distinct cluster forming (Table 22). Strains of *H. pylori* shared 90.4% to 94.5% of their genome with *H. acinonychis* sharing 83.3% to 86.8% of its genome to members of *H. pylori* (Table 22). Other members of the genus had low levels of orthology when compared to others of the genus, values ranged from 58.8% to 74.4% of the genomes shared. *H. canadensis* and *H. hepaticus* showed particularly low levels of orthology compared to other members of the genus (Table 22).

| | | DDH (Formula 2) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| Helicobacter pyloriNCTC 11637 =CCUG 17874$^{T}$ | 1 | | 66.3 | 62.8 | 61.9 | 55.9 | 54.8 | 37.1 | 19.5 | 27.1 | 26 |
| Helicobacter pylori G27 | 2 | 66.3 | | 62 | 61.7 | 55.9 | 54.5 | 37.2 | 19 | 20.4 | 21.3 |
| Helicobacter pyloriP12 | 3 | 62.8 | 62 | | 61.5 | 55.4 | 54.3 | 36.7 | 23.2 | 20.4 | 20.6 |
| Helicobacter pylori 26695 | 4 | 61.9 | 61.7 | 61.5 | | 56.4 | 54 | 37 | 19.7 | 19.8 | 20.8 |
| Helicobacter pyloriShi470 | 5 | 55.9 | 55.9 | 55.4 | 56.4 | | 50.1 | 37.1 | 19.7 | 20.2 | 21.1 |
| Helicobacter pyloriJ99 | 6 | 54.8 | 54.5 | 54.3 | 54 | 50.1 | | 36.9 | 21.2 | 20.1 | 20.8 |
| Helicobacter acinonychis str. Sheeba | 7 | 37.1 | 37.2 | 36.7 | 37 | 37.1 | 36.9 | | 18.8 | 27.9 | 21.5 |
| Helicobacter felis CS1, ATCC 49179$^{T}$ | 8 | 19.5 | 19 | 23.2 | 19.7 | 19.7 | 21.2 | 18.8 | | 23.4 | 33.6 |
| Helicobacter hepaticus ATCC 51449 | 10 | 27.1 | 20.4 | 20.4 | 19.8 | 20.2 | 20.1 | 27.9 | 23.4 | | 21.4 |
| Helicobacter canadensis MIT 98-5491, ATCC 700968$^{T}$ | 11 | 26 | 21.3 | 20.6 | 20.8 | 21.1 | 20.8 | 21.5 | 33.6 | 21.4 | |

Table 23: GGDC values for members of the *Helicobacter* genus.

Members of the *Helicobacter* genus formed a single distinct cluster consisting of members of *H. pylori* and possibly *H. acinonychis*. Strains of *H. pylori* had GGDC values greater than 87 when compared to each other and *H. acinonychis* had values ranging from 67.6 and 69.8 to strains of *H. pylori* (Table 23). Other organisms had values in the negatives when compared to other members of the genus (Table 23).

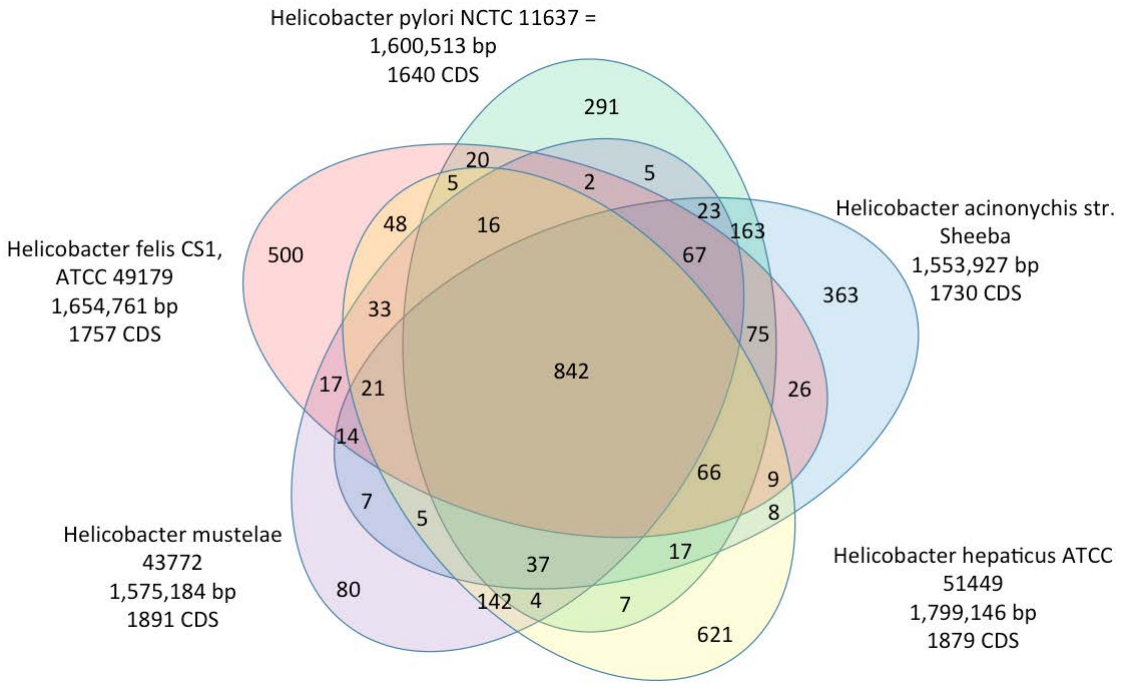


Figure 7: **Genomic similarity at the gene level for *H. mustelae, H. felis, H. hepaticus, H. pylori,* and *H. acinonychis*.** Venn-diagram shows the number of orthologous genes shared within each possible comparison combination.

The core genome of *H. mustelae*, *H. felis, H. hepaticus*, *H. pylori*, and *H. acinonychis* consisted of 842 genes. *H. mustelae* had 80 unique genes with a genome size of 1,575,184 bp and 1891 CDS *H. felis* had 500 unique genes with a genome size of 1,654,761 bp and 1757 CDS. *H. hepaticus* had 621 unique genes with a genome size of 1,799,146 bp and 1879 CDS. *H. pylori* had 291 unique genes with a genome size 1,600,513 bp and 1640 CDS. *H. acinonychis* had 363 unique genes with a genome size of 1,533,927 bp and 1730 CDS (Figure 7). *H. hepaticus* and *H. mustelae* shared a further 142 unique genes (Figure 7). H acinonychis and *H. pylori* shared a further 163 genes unique genes (Figure 7). *H. pylori* and *H. acinonychis* shared operons mostly consisting of metal regulatory proteins, osmotic proteins, and haloacid proteins. A large number of

genes were found to be unique to *H. acinonychis* and *H. pylori*. When compared it was found that these genes consisted mainly of two different types of operons, the first was captured phage genomes, the second consisted of pathogenicity islands or vacuolating cytotoxins (Table 24).

A

| | H. pylori only |
|---|---|
| 21 | Superfamily I DNA and RNA helicase and helicase subunits-like protein |
| 22 | cag island protein, CYTOTOXICITY ASSOCIATED IMMUNODOMINANT ANTIGEN |
| 23 | cag pathogenicity island protein B |
| 24 | cag pathogenicity island protein C |
| 25 | cag island protein |
| 26 | CAG pathogenicity island protein 23 (Protein picB) |
| 27 | cag pathogenicity island protein |
| 28 | cag island protein |
| 29 | cag island protein |
| 30 | cag island protein |
| 31 | cag pathogenicity island protein (cag18) |
| 32 | cag pathogenicity island protein (cag17) |
| 33 | cag island protein |
| 34 | hypothetical protein |
| 35 | cag pathogenicity island protein |
| | |
| 59 | FIG00711344: hypothetical protein |
| 60 | FIG00711868: hypothetical protein |
| 61 | Exonuclease SbcC |
| 62 | hypothetical protein |
| 63 | MobC-like protein |
| 64 | hypothetical protein |
| 65 | hypothetical protein |
| 66 | YafQ toxin protein |
| 67 | Rep |
| 68 | hypothetical protein |
| | |
| 288 | DNA-cytosine methyltransferase (EC 2.1.1.37) |
| 289 | FIG00710380: hypothetical protein |
| 290 | FIG00710380: hypothetical protein |
| 291 | DNA-cytosine methyltransferase (EC 2.1.1.37) |
| 292 | FIG00710779: hypothetical protein |
| | |
| 382 | hypothetical protein |
| 383 | cag pathogenicity island protein (cag1) |
| 384 | cag pathogenicity island protein (cag3) |
| 385 | cag island protein |
| 386 | Type IV secretion system protein VirD4 |
| 388 | cag pathogenicity island protein Z |
| 389 | cag pathogenicity island protein (cag7) |
| 390 | hypothetical protein |
| 391 | FIG00711566: hypothetical protein |
| 392 | FIG00711488: hypothetical protein |
| 393 | FIG00710817: hypothetical protein |
| 394 | hypothetical protein |
| | |
| 569 | hypothetical protein |
| 570 | phage/colicin/tellurite resistance cluster terY protein |
| 572 | FIG00712352: hypothetical protein |
| 573 | FIG00712090: hypothetical protein |
| 574 | hypothetical protein |
| 575 | hypothetical protein |
| 576 | FIG00710954: hypothetical protein |
| 577 | hypothetical protein |
| 578 | FIG00712400: hypothetical protein |
| 579 | FIG00711265: hypothetical protein |
| 580 | FIG00710476: hypothetical protein |
| 581 | hypothetical protein |
| 583 | hypothetical protein |
| 584 | hypothetical protein |
| 585 | FIG00711106: hypothetical protein |
| | |
| 1119 | hypothetical protein |
| 1120 | cag pathogenicity island protein |
| 1121 | cag pathogenicity island protein (cag12) |
| 1122 | cag pathogenicity island protein |
| 1123 | Inner membrane protein forms channel for type IV secretion of T-DNA complex (VirB8) |
| 1124 | cag pathogenicity island protein (cag9) |
| 1125 | cag island protein |

B

| | H. acinonychis only |
|---|---|
| 176 | putative |
| 177 | type III restriction enzyme R protein |
| 178 | type III restriction enzyme R protein |
| 179 | type III restriction enzyme R protein |
| 180 | type III restriction enzyme R protein |
| 181 | hypothetical protein |
| 182 | |
| 183 | type IIS restriction enzyme M1 protein (mod) |
| 184 | adenine specific DNA methyltransferase (dpnA) |
| 185 | type IIS restriction enzyme R protein (MBOIIR) |
| 186 | hypothetical protein |
| | |
| 487 | hypothetical protein |
| 489 | hypothetical protein |
| 490 | vacuolating cytotoxin |
| 491 | vacuolating cytotoxin |
| 492 | vacuolating cytotoxin |
| 493 | vacuolating cytotoxin |
| 494 | vacuolating cytotoxin |
| 495 | vacuolating cytotoxin |
| 496 | vacuolating cytotoxin |
| 497 | vacuolating cytotoxin |
| 498 | hypothetical protein |
| 499 | vacuolating cytotoxin |
| 500 | vacuolating cytotoxin |
| 501 | vacuolating cytotoxin |
| | |
| | |
| 728 | ISCco1, transposase orfA |
| 729 | ISCco1, transposase orfB |
| 731 | hypothetical protein |
| 732 | hypothetical protein |
| 733 | hypothetical protein |
| 734 | hypothetical protein |
| 735 | ISCco1, transposase orfA |
| 736 | hypothetical protein |
| 737 | hypothetical protein |
| 738 | tetracycline resistance protein tetA(P), putative |
| | |
| 1216 | vacuolating cytotoxin |
| 1217 | vacuolating cytotoxin |
| 1218 | vacuolating cytotoxin |
| 1219 | hypothetical protein |
| 1220 | vacuolating cytotoxin |
| 1222 | vacuolating cytotoxin |
| 1223 | vacuolating cytotoxin |
| 1224 | vacuolating cytotoxin |
| 1225 | vacuolating cytotoxin |
| 1226 | vacuolating cytotoxin |
| 1227 | vacuolating cytotoxin |
| | |
| 1291 | tetracycline resistance protein tetA(P), putative |
| 1292 | tetracycline resistance protein tetA(P), putative |
| 1293 | Integrase |
| 1294 | adenine specific DNA methyltransferase |

Table 24: Unique genes belonging to *H. pylori* (A) and *H. acinonychis* (B).

# I. Flavobacteriacae

| | ROSA (sorted) | 525257 | 1121870 | 1216967 | 1218108 | 376686 | 1122226 | 1121909 | 1121912 | 865937 | 398720 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Chryseobacterium gleum F93, ATCC 35910 | **525257** | | | | | | | | | | |
| Epilithonimonas tenax DSM 16811 | **1121870** | 30.0 | | | | | | | | | |
| Elizabethkingia meningoseptica ATCC 13253 = NBRC 12535 | **1216967** | 27.7 | 25.5 | | | | | | | | |
| Empedobacter brevis ATCC 43319 | **1218108** | 16.7 | 16.7 | 16.3 | | | | | | | |
| Flavobacterium johnsonia johnsoniae UW101 | **376686** | 13.1 | 13.2 | 12.9 | 13.0 | | | | | | |
| Mesonia mobilis DSM 19841 | **1122226** | 11.5 | 12.9 | 12.0 | 13.4 | 16.8 | | | | | |
| Gaetbulibacter saemankumensis DSM 17032 | **1121909** | 10.9 | 12.3 | 11.8 | 12.5 | 18.1 | 23.6 | | | | |
| Gelidibacter mesophilus DSM 14095 | **1121912** | 10.6 | 11.8 | 11.5 | 12.0 | 16.4 | 22.7 | 28.1 | | | |
| Gillisia limnaea R-8282, DSM 15749 | **865937** | 10.2 | 11.8 | 11.0 | 11.5 | 15.5 | 26.1 | 21.9 | 22.6 | | |
| Leeuwenhoekiella blandensis MED217 | **398720** | 10.2 | 11.3 | 10.8 | 11.3 | 15.6 | 26.0 | 21.7 | 21.1 | 23.2 | |
| Polaribacter franzmannii ATCC 700399 | **1248440** | 7.9 | 9.3 | 8.9 | 9.4 | 11.1 | 15.6 | 18.8 | 16.0 | 15.3 | 14.7 |

Table 25: Intrafamily ROSA values for *Chryseobacterium gleum*, *Epilithonimonas tenax*, *Elizabethkingia meningoseptica*, *Empedobacter brevis*, *Flavobacterium johnsoniae*, *Mesonia mobilis*, *Gaetbulibacter saemankumensis*, *Gelidibacter mesophilus*, *Gillisia limnaea*, *Leeuwenhoekiella blandensis*, and *Polaribacter franzmannii*.

Intrafamily ROSA values for the Flavobacteriacae family yielded two distinct clusters. The first cluster consisted of *Chryseobacterium gleum*, *Epilithonimonas tenax*, and *Elizabethkingia meningoseptica* with ROSA values ranging from 16.3 to 30 (Table 25). The second cluster consisted of *Flavobacterium johnsoniae*, *Mesonia mobilis*, *Gaetbulibacter saemankumensis*, *Gelidibacter mesophilus*, *Gillisia limnaea*, *Leeuwenhoekiella blandensis*, and *Polaribacter franzmannii* with ROSA values ranging from 11.1 to 26.1 (Table 25). ROSA values between the two clusters ranged from7.9 to 13.4 (Table 25). ROSA values suggest two distinct families forming the fiurst containing of *Chryseobacterium gleum*, *Epilithonimonas tenax*, and *Elizabethkingia meningoseptica* and the second containing *Flavobacterium johnsoniae*, *Mesonia mobilis*, *Gaetbulibacter saemankumensis*, *Gelidibacter mesophilus*, *Gillisia limnaea*, *Leeuwenhoekiella blandensis*, and *Polaribacter franzmannii.*

| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Flavobacterium_johnsoniae_DSM_2064 | 1 | | | | | | | | | | |
| Chryseobacterium_gleum_CCUG_14555 | 2 | 83.4 | | | | | | | | | |
| Chryseobacterium_meningosepticum_strain_13253T | 3 | 83.9 | 94.5 | | | | | | | | |
| Empedobacter_brevis_LMG_4011T | 4 | 39.6 | 39.4 | 39.3 | | | | | | | |
| Epilithonimonas_tenax_EP105 | 5 | 82.9 | 94.3 | 95.0 | 39.3 | | | | | | |
| Gaetbulibacter_saemankumensis_SMK-12_16S | 6 | 90.0 | 81.6 | 83.1 | 38.9 | 82.1 | | | | | |
| Gelidibacter_mesophilus_2SM29T | 7 | 88.8 | 82.3 | 84.0 | 39.4 | 83.0 | 92.9 | | | | |
| Gillisia_limnaea_R-8282 | 8 | 89.1 | 81.9 | 82.7 | 40.0 | 81.1 | 90.2 | 89.9 | | | |
| Leeuwenhoekiella_blandensis_MED217 | 9 | 87.8 | 82.5 | 84.1 | 39.2 | 82.2 | 90.8 | 89.9 | 89.3 | | |
| Mesonia_mobilis_KMM_6059 | 10 | 88.3 | 82.5 | 84.1 | 39.9 | 82.2 | 90.2 | 89.7 | 91.7 | 90.6 | |
| Polaribacter_franzmannii_ATCC_700399 | 11 | 87.5 | 82.3 | 83.3 | 39.6 | 82.6 | 88.7 | 88.1 | 87.7 | 87.8 | 89.0 |

Table 26: 16s rRNA values for members of the Flavobacteriacae.

16s rRNA values for members of the Flavobacteriacae revealed two distinct clusters. The first cluster consisted of *Chryseobacterium gleum* and *Epilithonimonas tenax* with a16s rRNA % similarity of 94.5% (Table 26). The second cluster contained *Flavobacterium johnsoniae*, *Mesonia mobilis*, *Gaetbulibacter saemankumensis*, *Gelidibacter mesophilus*, *Gillisia limnaea*, *Leeuwenhoekiella blandensis*, and *Polaribacter franzmannii* with 16s rRNA % similarity ranging from 87.5% to 92.9% (Table 25). *Empedobacter brevis* had low levels of 16s rRNA similarity with values ranging from 39.3% to 40% (Table 26).

| | Average Amino Acid Identity (AAIr) | 376686 | 525257 | 1216967 | 1218108 | 1121870 | 1121909 | 1121912 | 865937 | 398720 | 1122226 | 1248440 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Flavobacterium johnsonia johnsoniae UW101 | **376686** | | 50.5 | 49.8 | 50.7 | 50.3 | 57.6 | 55.9 | 55.5 | 54.7 | 55.6 | 53.7 |
| Chryseobacterium gleum F93, ATCC 35910 | **525257** | 50.2 | | 67.1 | 54.2 | 69.8 | 47.5 | 47.2 | 47.1 | 46.7 | 48.1 | 46.8 |
| Elizabethkingia meningoseptica ATCC 13253 = NBRC 12535 | **1216967** | 49.6 | 67.1 | | 52.9 | 65.5 | 48.5 | 47.7 | 47.7 | 47.5 | 48.5 | 47.1 |
| Empedobacter brevis ATCC 43319 | **1218108** | 50.2 | 54.2 | 52.9 | | 52.9 | 49.7 | 48.8 | 48.4 | 48.0 | 49.6 | 47.9 |
| Epilithonimonas tenax DSM 16811 | **1121870** | 50.1 | 69.8 | 65.3 | 53.0 | | 47.8 | 47.8 | 47.5 | 47.3 | 47.7 | 47.2 |
| Gaetbulibacter saemankumensis DSM 17032 | **1121909** | 57.2 | 47.7 | 48.4 | 49.8 | 47.8 | | 66.0 | 60.6 | 60.0 | 60.9 | 59.8 |
| Gelidibacter mesophilus DSM 14095 | **1121912** | 55.5 | 47.4 | 47.7 | 48.8 | 47.8 | 66.0 | | 60.3 | 59.3 | 59.8 | 57.8 |
| Gillisia limnaea R-8282, DSM 15749 | **865937** | 55.1 | 47.1 | 47.7 | 48.7 | 47.6 | 60.6 | 60.4 | | 60.9 | 63.4 | 56.6 |
| Leeuwenhoekiella blandensis MED217 | **398720** | 54.6 | 46.9 | 47.6 | 48.0 | 47.3 | 60.2 | 59.4 | 61.0 | | 62.4 | 55.6 |
| Mesonia mobilis DSM 19841 | **1122226** | 55.5 | 48.3 | 48.6 | 49.6 | 47.8 | 60.9 | 59.7 | 63.4 | 62.1 | | 55.9 |
| Polaribacter franzmannii ATCC 700399 | **1248440** | 53.4 | 46.8 | 47.1 | 48.0 | 47.2 | 59.7 | 57.9 | 56.5 | 55.7 | 55.9 | |

Table 27: AAI values for members of the Flavobacteriacae family.

AAI values of the Flavobacteriacae family yielded two distinct clusters. The first cluster consisted or *Chryseobacterium gleum*, Epilithonimonas tenax and Elizabethkingia

meningoseptica with AAI values ranging from 65.3 to 69.8 (Table 27). The second cluster *Flavobacterium johnsoniae*, *Mesonia mobilis*, *Gaetbulibacter saemankumensis*, *Gelidibacter mesophilus*, *Gillisia limnaea*, *Leeuwenhoekiella blandensis*, and *Polaribacter franzmannii* with AAI values ranging from 55.6 to 66.0 (Table 27). *Empedobacter brevis* had no clustering and AAI values ranging from 54.2 to 47.9 (Table 27).

| | Percent Bidirectional Best Hit (% BBH) | 376686 | 525257 | 1216967 | 1218108 | 1121870 | 1121909 | 1121912 | 865937 | 398720 | 1122226 | 1248440 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Flavobacterium johnsonia johnsoniae UW101 | **376686** | | 53.4 | 63.4 | 61.7 | 64.9 | 71.4 | 60.9 | 61.0 | 60.6 | 70.3 | 36.4 |
| Chryseobacterium gleum F93, ATCC 35910 | **525257** | 50.3 | | 72.6 | 67.3 | 74.7 | 61.8 | 53.1 | 54.0 | 52.6 | 62.9 | 32.8 |
| Elizabethkingia meningoseptica ATCC 13253 = NBRC 12535 | **1216967** | 41.0 | 50.4 | | 58.5 | 61.4 | 55.0 | 47.3 | 47.9 | 45.3 | 55.4 | 29.5 |
| Empedobacter brevis ATCC 43319 | **1218108** | 40.0 | 46.3 | 58.3 | | 61.4 | 55.8 | 47.3 | 48.7 | 46.4 | 59.0 | 29.9 |
| Epilithonimonas tenax DSM 16811 | **1121870** | 39.5 | 48.5 | 57.8 | 57.5 | | 57.6 | 47.1 | 50.3 | 46.0 | 59.5 | 29.5 |
| Gaetbulibacter saemankumensis DSM 17032 | **1121909** | 38.2 | 34.8 | 45.2 | 45.7 | 50.4 | | 54.3 | 53.6 | 50.8 | 62.9 | 33.8 |
| Gelidibacter mesophilus DSM 14095 | **1121912** | 45.1 | 41.3 | 53.5 | 53.8 | 56.5 | 74.6 | | 65.7 | 60.4 | 73.0 | 37.9 |
| Gillisia limnaea R-8282, DSM 15749 | **865937** | 40.2 | 37.7 | 48.3 | 48.8 | 54.1 | 65.6 | 58.4 | | 59.3 | 71.0 | 35.2 |
| Leeuwenhoekiella blandensis MED217 | **398720** | 44.0 | 40.3 | 50.6 | 52.1 | 55.1 | 69.1 | 59.7 | 65.7 | | 76.2 | 37.2 |
| Mesonia mobilis DSM 19841 | **1122226** | 38.2 | 36.3 | 46.4 | 49.6 | 53.3 | 64.1 | 54.3 | 58.9 | 57.8 | | 32.6 |
| Polaribacter franzmannii ATCC 700399 | **1248440** | 41.2 | 39.1 | 50.8 | 51.6 | 54.3 | 71.6 | 57.9 | 60.7 | 57.3 | 67.4 | |

Table 28: %BBH values for members of the Flavobacteriacae family.

Orthology between members of the Flavobacteriacae family yielded no distinct clustering. A few members stood out with exceptionally high or low %BBH scores. *Flavobacterium johnsoniae* had both some of the highest and lowest genome conservation between members of its family. Values ranged from 53.4% to 71.4% when the reference organism but ranged from 38.2% to 50.3% when used as the comparison organism (Table 28). *Mesonia mobilis*, *Gaetbulibacter saemankumensis*, *Gelidibacter mesophilus*, *Gillisia limnaea*, and *Leeuwenhoekiella blandensis* had high levels of genome conservation ranging from 60.4% to 76.2% (Table 28). *Polaribacter franzmannii* had the lowest genome conservation with other members of the family with values ranging from 29.5 to 37.2% of the genome conserved (Table 28).

# V. Discussion:

## A. ROSA Formula

ROSA is the first genome-based microbial classification system with described thresholds given for each taxonomic level. Previous metrics have only had defined species thresholds (DDH, GGDC, ANI, AAIr, TETRA) or species and tentative genus thresholds (% 16S rRNA similarity) (Stackebrandt and Ebers, 2006; Konstantindinis & Tiedje, 2005; Konstantindinis and Tiedje, 2005). Its increased taxonomic resolution enables more accurate delimitation of taxa than with previous metrics. As ROSA is based on protein-coding genes, its scores also incorporate the ideas of polyphasic and chemotaxonomic approaches, and metabolic characteristics of compared isolates. Built on the premises of the gain, loss and divergence of protein-coding genes, ROSA is the first whole-genome metric to be built on concise and clear biological concepts. In order to fully grasp the depth of ROSA 2338 reciprocal comparisons were done across all levels of taxonomy (Table 1).

A great number of outliers were encountered when the taxonomic thresholds established for ROSA were paired against the current taxonomic placement of each organism for a total of 834 outliers or 35.2% of all data points. Groups with the largest percentages of outliers were those at the species-genus level and the class-order level, with approximately half of their total comparison being outliers. Species-strain, domain, and order-family levels comparisons had the fewest number of outliers. A reason for the species-strain level having the fewest percentage of outliers is that taxonomists have a

clear definition already in place. In order to more fully understand the ROSA tool, outliers were chosen for a more in depth study of their genomic characteristics in order to validate ROSA's placement of the organism or find any flaws in the ROSA classification. Organisms at the species-strain level, species-genus level, and the genus-family level were chosen due to the more robust understanding of similarities necessary for those levels of taxonomy. Currently there only exists a defined threshold for the strain-species level, yet there is a general consensus on some metrics for the species-genus level. Genus-family level thresholds are more poorly understood but outliers at this stage were thought to possibly yield more decisive data than outliers at other points.

## B. Example Clusters:

Strain designation is in general a reliable taxonomic placement of organisms. Except a few cases, such as medical pathology mentioned in II B, organisms classified within the same species have, in general a strong taxonomic certainty. As such, strain level comparisons were the first to be done in order to examine ROSA's classification strength. When 11 strains of *Staphylococcus aureus* were compared, ROSA values were all well above the 65 threshold proposed for differentiation, having ROSA values ranging from ~ 82 and ~98 (Table 2A). When 11 strains of *Streptococcus pyogenes* were compared ROSA values were all well above the 65 threshold proposed for differentiation, having ROSA values ranging from ~ 81 and ~96 (Table 2B). When 11 strains of *Escherichia coli* were compared ROSA values were above the 65 threshold proposed for differentiation, having ROSA values ranging from ~ 70 and ~98 (Table 2C).

It is worth noting that although it was mentioned earlier some of the uncertainties in *E. coli* classification, the strains chosen for this level of comparison had all been extensively studied at the genomic level and as such their current classification was well supported.

Strains of Staphylococcus aureus had ROSA values ranging from 83.6 to 94.8, well within the 65 threshold for members of the same species and as such are correctly identified as the same species (Table 3). The two strains of Staphylococcus epidermis had a ROSA value of 86.4 when compared to each other; this value is greater than the 65 threshold suggested for species level differentiation and as such, they are correctly identified as the same species (Table 3).

When comparing different species of the same genus it is hypothesized that ROSA values ranging from 35 – 65 would indicate intragenus classification between two organisms (Table 1). A case example of this is the intragenus comparisons of strains of S. aureus, S. epidermis, S. haemolyticus, and S. saprophyticus. These ROSA values ranged from 37.6 to 43.9 (Table 3). These values fall within the proposed range of 35 to 65 for intragenus comparisons and as such these strains are thought to be correctly identified as different species of the same genus (Table 1).

## C. Brucella:

The Brucella genus was the first to have ROSA values that suggested a taxonomy different from the currently accepted taxonomic ranking. Although the ROSA values for intraspecies comparisons, ranging from 93.5 to 97.4, inducated same species taxonomic ranking, current taxonomy for *B. melitensis*, *B. abortus*, *B. suis*, *B. canis*, and *B. ovis* have

them classified as different species of the same genus. ROSA values for the genus range from 88 to 95.0 (Table 4). These values suggest that the genus consists of different strains of the same species. As mentioned earlier the official definition for a species is >70% DDH, and GGDC can be used as a computational substitute for DDH. When GGDC comparisons were run for the organisms all had reciprocal GGDC values greater than the 70% species threshold. GGDC values ranged from 96.6 to 100 (Table 5). These values coupled with the ROSA scores indicate that current taxonomy is wrong for the Brucella genus and the species need to be reclassified as all members of the same species. Verger et al. in 1985 showed that all 6 members of this genus showed sufficient clustering through physical DDH to classify them as members of the same species. Verger and his colleagues found that species had DDH values ranging from 89% to 110% hybridization to the immobilized strain (Verger et al. 1985). (Note that some values were greater than 100%, due to inherent flaws within the DDH metric that cause deviation in the measured values). In the paper the authors suggested that the organisms be reclassified as one species with different relevant Biovars for each of the "species". Ultimately this taxonomic change was noted but disregarded due to the confusion it would cause within the medical community. It is interesting to note that the DDH, GGDC, and ROSA values all had little deviation in their magnitude when compared to each other (Table 4. 5, and Verger et al. 1985). Based on the large amount of evidence in both ROSA values and the official standard of DDH it is concluded that the ROSA values and assignment of taxonomic position for the Brucella genus as a monospecific genus is correct. However as mentioned in the introduction, taxonomy will

resist correction due to the medical relevance of the organisms and the fears that it will cause confusion for medical practitioners and patients.

## D. Francisella and Thiomicrospira:

ROSA values were generated for both intraspecies and intragenus comparisons for the Francisella genus. *F. tularensis* had 8 strains belonging to 4 subspecies compared to each other. ROSA values ranged from 81.9 to 93.6 within the expected range for members of the same species (Table 1, Table 6). These values indicate that the current taxonomy for the species is correct and serve to validate ROSA on the intraspecies level. For intragenus comparisons of the strains of *F. tularensis* and *F. philomiragia* ROSA values ranged from 55 to 58.5 (Table 6). These values agree with the suggested ROSA range for organisms that are in the same genus but different species, 35 – 65 (Table 1). Based on the ROSA values it is suggested that current taxonomic ranking is correct and serve to validate ROSA in differentiating members of the same genus that are different species.

Two species of the Thiomicrosira genus had ROSA values calculated between each other and the strains of *F. tularensis* and F. philomiraga. Comparisons between *F. crunogens* and strains of *F. tularensis* and *F. philomiragia* yielded ROSA values ranging from 8.0-8.1, within the expected intraclass range of ROSA (Table 6). Comparisons between *F. denitrificans* and strains of *F. tularensis* and *F. philomiragia* yielded ROSA values ranging from 4.9 to 5.2, these values were in the range of intraphyla instead of the expected intraclass range (Table 6). Furthermore the ROSA score between T. crunogena and *F. denitrificans* was 7.0, in the intraclass range (Table 6). A literature

search reveals that *F. denitrificans* was recently reclassified by Takai et al. in 2006 as a member of the Sulfurimonas genus (Takai et al, 2006). They found that S. denitrificans was originally classified based on its 16s rRNA sequence and incorrectly identified electron donors and acceptors; during a more detailed approach it was found that the S. denitrificans had physiological traits more closely related to members of Sulfurimonas and as such suggested its reclassification. In regards to the new classification the ROSA scores of the organisms make more sense. Thiomicrospira and Francisella are both members of the gammaproteobacteria class while Sulfurimonas is a member of the Epsilonproteobacteria class. S. denitrificans had intraphyla ROSA levels when compared to the members of the Francisella genus, this updated taxonomy fits the ROSA predictions. Comparisons between T. crunogena and S. denitrificans yielded a ROSA value of 7, not within the expected intraphyla range. A possible explanation for this is that at higher taxa levels found in extreme environments, genomes may be more highly conserved, which would give these organisms an inflated ROSA score. In the paper Takai et al. also built on a theme in extremophilic organisms, that "metabolic versatility of deep-sea epsilonproteobacteria is not relevant to their 16S rRNA gene phylogeny" (Yakai et al 2006). This theme is important to remember as many extremophilic and pathogenic organisms are classified almost solely based on their 16s rRNA, as are many unculturable organisms. If the 16s rRNA gene can not solely be used for extremophilic organisms' classifications then is it possible that additional metabolic information must also be used for pathogenic organisms?

## **E. Flavobacterium**:

The Flavobacterium genus was established in 1923 with the type species aquatile having been described in 1889 (Holmes and Owen 1979; Bergey *et al.* 1923). After the emended description of the genus that it is only to include organisms that are pigmented, non-motile, strictly aerobic, and have guanine-plus-cytosine (G+ C) contents ranging from 30 to 42 mol%. A number of groups have been moved out of the genus and into new or existing genera (Holmes et al 1984). These have included Sphingobacterium, Cytophaga, and Chryseobacterium (Vandamme et al. 1984; Reichenbach 1989; Yabuuchi et al. 1983). These species were moved out on the basis of updated phenotypic data that was required for the emended genus description. A that helped lead tothe emended genus description was the polyphasic study done by Bernardet et al. that restricted the genus to ten species on the *aquatile* rRNA branch with more specific phenotypic characteristics. The emended genus based on the *F. aquatile* branch, has the following features: gram-negative rods that are motile by gliding, are chemoorganotrophic and aerobic, produce cream to yellow colonies on agar, and decompose several polysaccharides. The habitats of these organisms are widely distributed in both soil and freshwater, with some species being pathogenic to fish. The G + C contents of the proposed emended genus by Bernardet have G+C contents ranging from 32 to 37 mol% (Bernardet et al. 1996).

Based on the 16s rRNA neighbor joining tree, two clusters were identified that had enough members for potential novel genera status due to their branching within

the genus. The johnsoniae cluster consisted of *F. hibernum*, *F. chiliense*, *F. denitrificans*, *F. johnsoniae*, *F. reichenbachii*, F. chungagense, and *F. hydatis*, while the salisperosum cluster contained *F. salisperosum*, F. cuaense, *F. limnosediminis*, F. enshiense, and *F. suncheonense*. 16s rRNA% similarities within the clusters when compared to *F. aquatile* showed significant differences from the type species while retaining a level of similarity within their cluster (Table 7A). Species within the cluster consisting of *F. denitrificans*, *F. johnsoniae*, *F. chiliense*, *F. chungangense*, F*F. hibernum*, *F. hydatis*, and *F. reichenbachii*, shared 94% to 99% 16s rRNA similarity, while *F. aquatile* shared less then 95% 16s rRNA similarity. Organisms in the salisperosum cluster consisting of *F. salisperosum*, F. cuaense, *F. limnosediminis*, *F. suncheonense*, and F. enshiense had similarities between 95.4% and 99.1% (Table 7B). This 16s rRNA grouping, in both the neighbor joining tree and % difference, indicates that organisms in both johnsoniae cluster and the salisperosum cluster differentiated from *F. aquatile* significantly since their divergence from a common ancestor.

AAI values calculated for both clusters showed significant differences between organisms within each cluster when compared to organisms outside of their cluster, specifically *F. aquatile* (Table 8A). For both clusters *F. aquatile* showed values between 67% and 69% AAI when compared to organisms within the cluster (Table 8A, 8B). Within the *johnsoniae* cluster AAI values ranged from 78.762% an 86.109%, within the salisperosum cluster AAI values ranged from 79.126% and 88.822% (Table 8A, 8B 2). These elevated AAI values within each cluster as well as the lower values exhibited by each species when compared to *F. aquatile* indicate significant amino acid sequence

differences between the organisms and *F. aquatile.* This indicates that the cluster's amino acid sequences differentiated from *F. aquatile* significantly since their divergence from a common ancestor. The AAI values seen within the clusters fall within the suspected range of organisms that belong in the same genus, yet still remain different species. The ranges compared to *F. aquatile* fall in the range of organisms within the same family but different genera (Figure 2).

When the genomes of *F. hibernum*, *F. johnsoniae*, *F. chilense*, *F. denitrificans*, and *F. aquatile* were compared at the gene level it was evident that there was a large amount of similarity between *F. hibernum*, *F. johnsoniae*, *F. chilense*, and *F. denitrificans*; they shared 924 genes that were not present in *F. aquatile.* The core genome consisted of 1863 genes. Each organism had a large number of unique coding sequences *F. hibernum* contained 691 unique genes, *F. johnsoniae* contained 1152 unique genes, *F. chilense* contained 1002 unique genes, *F. denitrificans* contained 628 unique genes, *F. aquatile* contained 822 unique genes (Figure 3). Unique genes within each organism correlated positively with the genome size. Organisms with the largest genomes, *F. johnsoniae* and *F. chilense* contained the most unique genes. *F. hibernum* and *F. denitrificans*, contained the smaller number of unique genes despite being the middle two genomes in size. *F. aquatile* had the smallest genome yet the 3$^{rd}$ most unique coding sequences (Figure 3). This suggests that *F. aquatile* belongs to a different genus when compared to the organisms in johnsoniae cluster. Additionally, the organisms in johnsoniae cluster had genomes ranging from 4.8 to 6.1 Mbp and 4214 to 5345 CDS, *F. aquatile* had a genome size of 3.5 Mbp and 3221 CDS. This difference in

genome size and coding sequences indicates a distinct taxonomic difference between the two groups of organisms beyond that of belonging to the same genus.

Phenotypic results from Biolog GenIII plates revealed phenotypic differences between organisms in johnsoniae cluster and *F. aquatile*. *F. reichenbachii*, *F. hibernum F. chungangense*, and *F. chilense* showed growth at 1% NaCl and were able to utilize D-fructose, L-aspartic acid, L-serine, D-galacturonic acid, and tetrazolium blue while *F. aquatile* was unable to. However *F. aquatile* was able to proliferate in naladixic acid while *F. reichenbachii*, *F. hibernum F. chungangense*, and *F. chilense* were unable to (Figure 4). These distinct phenotypes that separated the cluster from *F. aquatile* can be located as operons within the 924 shared genes that *F. aquatile* does not contain. Within the shared genes there are 5 major operons encoding over 90 genes for carbohydrate metabolism. Furthermore the organisms contain an operon for nitrite metabolism, various drug and metal resistance proteins, iron and sulfur utilization enzymes, a DNA metabolism operon, and 2 operons for flexirubin synthesis.

When the genomes of *F. cauense*, *F. aquatile*, F. *enshiense*, F. *limndosedimins,* and *F. salisperosum* were compared at the gene level it was evident that there was a large amount of similarity between F. *enshiense*, *F. limnosediminis*, F. *cauense*, and *F. salisperosum,* they shared 227 genes that were not present in *F. aquatile.* The core genome consisted of 1744 genes. Each organism had a significant number of unique coding sequences; *F. cauense* contained 386 unique genes, F. *enshiense* contained 516 unique genes, F. *limndosedimins* contained 502 unique genes, *F. salisperosum* contained

241 unique genes, while *F. aquatile* contained 911 unique genes (Figure 4). Unique genes within each organism correlated positively with the genome size and number of CDS. Organisms within the cluster with the largest genomes, F. *enshiense* and *F. limnosediminis* contained the most unique genes. *F. cauense* and *F. salisperosum*, contained the smallest number of unique genes. *F. aquatile* had a similar sized genome to F. enshiense and *F. limnosediminis* yet contained the most unique coding sequences (Figure 4). However the number of unique genes it had was far greater than the size of its genome would lead one to believe. Furthermore the large number of genes it did not share with the other organisms, 227, and the low number it shares with the others, fewer than 100 per organism, supports its classification as a separate genus. Some genes that are shared by all of the organisms in the salisperosum cluster but not aquatile are genes encoding for various metabolic proteins, some antibiotic resistance proteins, metalloresistance proteins, a large number of proteins with unknown functions, and a large number of transcriptional regulators.

ROSA values for each of the clusters were calculated and sorted for further analysis. ROSA values for organisms in johnsoniae cluster ranged from 35.527 to 54.222 within the cluster and ranged from 25.485 to 32.215 when compared to *F. aquatile (*Figure 5A). ROSA values for organisms in the salisperosum cluster ranged from 52.119 to 66.422 within the cluster and ranged from 30.578 to 34.025 when compared to *F. aquatile* (Figure 5B). The ROSA values suggest that the organisms in Johnsoniae cluster and the salisperosum cluster should be classified into new genera other than flavobacterium in the Flavobacteriacae family.

The phenotypic and genomic tests conducted on a cluster consisting of *F. hibernum*, *F. chiliense*, *F. denitrificans*, *F. johnsoniae*, *F. reichenbachii*, F. *chungagense*, and *F. hydatis*, as well as a cluster consisting of *F. salisperosum*, F. *cuaense*, *F. limnosediminis*, F. *enshiense*, and *F. suncheonense*, suggest that each group forms a separate genera within the Flavobacteriacae family. It is suggested that the cluster consisting of *F. salisperosum*, F. *cuaense*, *F. limnosediminis*, F. *enshiense*, and *F. suncheonense* be reclassified with the genus name Gregorabacterium, and the cluster containing the organisms *F. hibernum*, *F. chiliense*, *F. denitrificans*, *F. johnsoniae*, *F. reichenbachii*, F. *chungagense*, and *F. hydatis* be reclassified with the genus name Lycobacterium. Both novel genera will be within the Flavobacteriacae family.

## F. Bacillus:

The Bacillus genus was established in 1872 with Bacillus subtilis being described in 1835 and designated the type species in 1872. There are currently 299 validly published species (Parte 2014). Organisms in this genus occupy a large range of habitats, ranging from deserts, hot springs, polar regions, and as pathogens. General characteristics of members are that they are aerobic, may be facultative anaerobic, spore forming, alkaliphilic, halotolerent, and gram positive. Members are generally mesophilic although some can grow at higher temperatures. Most prefer a neutral pH (Zhu et al., 2014).

ROSA values for members of the Bacillus genus showcased two distinct clusters. The Cereus cluster consisted of *B. anthracis*, *B. thuringiensis*, and *B. cereus*. Intraspecies comparisons for *B. anthracis* yielded values between 81.8 and 96 (Table 10). These values are consistent with the expected range for intraspecies comparisons. Intraspecies values for strains of *B. cereus* was 69.2 for two of the strains, within expected range, and 49.2 to 50.2, for the cytotoxis subsp.; this value was below what was expected for intraspecies comparisons (Table 10). Rosa values between *B. cereus* strains, B thuringiensis, and *B. cereus* strains ranged from 69.2 to 84.8, for all but *B. cereus* subsp cytotoxis. For the cytotoxis subsp. ROSA values ranged from 49.7 to 50.8, within the expected values for intragenus comparisons (Table 10). The second cluster consisted of *B. licheniformis*, *B. amyloliquefaciens*, and *B. subtilis*. Intraspecies value for the two strains of *B. amyloliquefaciens* was 80.9, within the expected range (Table 10). Intragenus comparisons between *B. licheniformis*, *B. amyloliquefaciens*, and *B. subtilis* ranged from 37.1 to 52.4, within the expected range (Table 10). Values between the two clusters ranged from 17.3 to 19.7, outside the range expected for members of the same genus but expected for members of the same family (Table 10).

AAI values support the original clustering formed by the ROSA matrix. Members of the Cereus cluster had AAI ranging from 93 to 99.9 (Table 11). The standard cutoff for AAI is generally defined as 94-95. While some of them, mainly *B. cereus* ATCC 14579, fall outside of this boundary, the amount they do so is extremely small and would not warrant reclassification. The only member of this group that truly falls out of the cluster is *B. cereus* subsp. cytotoxis. The strain has AAI values ranging from 82.8 to 83.3 for all

members of the Cereus cluster (Table 11). Although this puts the organisms at odds with classical classification, the values showcase again how related the other organisms are to each other. Within the second cluster strains of B. amyloquiefaciens had AAI value of 95.6, within the expected range for members of the same species. Within the cluster different organisms had AAI values ranging from 71.3 to 80.4 (Table 11). When members of both clusters were compared, AAI values ranged from 57.3 to 58.4 between the members of each cluster, providing a strict genomic boundary between the two clusters (Table 11).

When observing the orthology of the strains, the same distinct clustering pattern appears. Members of the *B. cereus* cluster show a high level of orthology between themselves. Except for *B. cereus* subsp cytotoxis, which had %BBH values from 57.7 to 62.8, the members of the *B. cereus* cluster had %BBH scores ranging from 77 to 97.9 (Table 12). For the *B. subtilis* cluster, the two strains of *B. amyloliquefaciens* had %BBH scores of 89.1 and 87.9 (Table 12). Strains from within Cluster 2 had intracluster values ranging from 75 to 85 (Table 12). When organisms were compared between clusters, BBH values ranged from 46.5 to 61.9 (Table 12). These values showcase a strict shared genome barrier with the members of each cluster having higher levels of conservation between members of their own cluster, and less compared to the other cluster. It should be noted that the organisms in the first cluster showcase higher intragenus values than the organisms of cluster 2, differing by over 20% in some cases (Table 12).

The 16s rRNA clustering further shows this relationship between the organisms. The organisms deemed as the *B. cereus* cluster showed 99.6 to 99.9 16s rRNA % similarity when compared to each other (Table 13). Organisms in the *B. subtilis* cluster showcased this high level of similarity as well, with all of them sharing a value greater than 98.2% 16s rRNA similarity. Members of the  all had values greater than what would be expected for organisms classified as different species while only *B. licheniformis* and *B. subtilis* showed increased levels of similarity for the *B. subtilis* cluster(Table 13). When the two clusters were compared to each other a clear separation was noticed between the two clusters. 16s rRNA values ranged from 93.6 to 94.1 between the two clusters, showing a clear separation through 16s (Table 13). Overall 16s values showcase two clusters that are highly related within each cluster but show little relation to each other.

When the genomes of *B. licheniformis*, *B. amyloliquefaciens*, *B. anthracis*, *B. cereus*, and *B. subtilis* were compared at the gene level it was evident that there was a large amount of similarity between *B. anthracis* and *B. cereus;* they shared 1870 genes that were not present in *B. licheniformis*, *B. amyloliquefaciens*, and *B. subtilis,* who shared 645 genes not present in *B. anthracis* and *B. cereus (Figure 5).* The core genome consisted of 1926 genes. Each organism had a significant number of unique coding sequences; *B. licheniformis* contained 669 unique genes, *B. amyloliquefaciens* contained 701 unique genes, *B. anthracis* contained 807 unique genes, *B. cereus* contained 865 unique genes, while *B. subtilis* contained 706 unique genes (Figure 5). The genes shared between *B. licheniformis*, *B. amyloliquefaciens*, and *B. subtilis* encode different non-oxidative carboxylase proteins, sporulation proteins, mannitol catabolism proteins,

unique phage genomic DNA, flagellar proteins, and multiple Na/H transport proteins (Table 14B). Genes shared by *B. anthracis* and *B. cereus* encode for sporulation proteins, different iron transport proteins, chemo-taxis proteins, malate catabolism operons, some specialized membrane proteins, and some antibiotic resistance proteins (Table 14A).

Based on the data a number of conclusions can be drawn about the Bacillus genus. The *B. cereus* cluster consisting of *B. anthracis*, *B. thuringiensis*, and *B. cereus* should all be classified as the same species within one genus. The only strain not within that species would be the *B. cereus* cytotoxis subsp. A recent literature search has shown that the cytotoxis subspecies has recently been moved to its own species, *B. cytotoxicus* through the work done by Guinebretière et al. in 2013. They showed that the organism is different from its parent type through MLST, 16s rRNA, and DDH (Guinebretière et al. 2013). In light of this reclassification the ROSA values perfectly predict the placement of the organisms. It is common knowledge that *B. anthracis*, B. *thuringiensis*, and *B. cereus* are the same species (Helgason et al., 2000). Further evidence of this comes from the AAI, %BBH, and gene count comparisons (Table 11, 12 and Figure 5). For each of these metrics, members of this cluster showcase levels of similarity that would be expected of organisms from the same species. Guinebretière et al. showed that the DDH values between the three organisms were all greater than the 70% species threshold, indicating same species taxonomic ranking (Guinebretière et al. 2013). The gene counts of *B. cereus* and *B. anthracis* showed a large amount of similarity, however there was still a significant number of unique genes. Genome

analysis yielded that these unique genes were a few membrane proteins, a large number of captured phage genomes, over 70% of the genes, and two pathogenicity islands found in *B. anthracis*. When these values are placed in context of their pathology it showcases their unique phylogeny. Both are pathogens that only differed in the severity of the disease that they cause, with anthracis causing a much more serious disease and containing two extra pathogenicity islands.

The *B. subtilis* cluster containing *B. licheniformis*, *B. amyloliquefaciens*, and *B. subtilis* showed similarities between each other that far outstripped the similarities they showed to the other cluster. Based on the 16s rRNA data, the AAI values, %BBH values, and the gene counts, it appears that the ROSA classification of the organisms was correct. The three organisms shared an AAI score over 71 between each other yet had AAI values lower than 60 when compared to the other members of the cluster. Furthermore, the new *B. cytotoxicus* species had AAI values greater than 80 when compared to its cluster, yet organisms from the *B. subtilis* cluster fell far below that when compared to *B. cereus* cluster organisms, with values ranging from 57.3 to 58.4 (Table 11). These values form a strong separation between the two clusters. %BBH shows the same clustering, members within the second cluster share over 70% of their genes with each other while sharing far less than 60% with members of the other cluster (Table 12). 16s rRNA also forms a barrier between the two groups at less than 94.1 % similarity, below the 95% recommendation for members of the same genus (Table 13). When gene counts were observed the second cluster shared 645 genes uniquely between them that were not found in thec*Cereus* cluster, with the *cereus* cluster

holding 1870 genes uniquely to them (Figure 5). These genes encoded a large number of catabolic proteins that could provide a distinct phenotype for the organisms. Additional well-known evidence for this separation is the fact that none of the organisms in cluster 2 cause a disease sharing any of the pathology similar to the organisms in the *B. cereus* cluster . These stark differences between the two clusters, as well as the high levels of similarities unique to both clusters, indicate that the two should be split into different genera. This suggestion is further supported by the ROSA values splitting the genus into two separate genera within the same family.

## G. *Chlamydia*ceae:

The genus *Chlamydia* was established in 1945 with *Chlamydia* trachomatis being characterized in 1935 and moved to the type species in 1945. There are currently 6 validly published members. The genus *Chlamydophila* was split from the *Chlamydia* genus in 1999 with *Chlamydophila* pisttaci being the type for it. This has led to a large amount of controversy against the split (Schachter et al., 2001). The split occurred due to differences in glycogen staining, genome size, and ribosomal operons with the 16s rRNA difference straddling the guideline for new genus establishment. Members of both genera are generally obligate intracellular parasites. They are given their own order due to their unique developmental cycle.

ROSA values for the *Chlamydiacae* family revealed two semi-separate clusters connected by two organisms. The *Chlamydophila* cluster consists of the members of the

*Chlamydophila* genus, and the second containing members of the *Chlamydia* genus (Table 15). Intraspecies comparisons between members of the *Chlamydophila* genus ranged from 97.4 to 99.0. Intragenus comparisons between members of the *Chlamydophila* genus ranged from 39.7 to 68.9 (Table 15). Intraspecies comparisons between members of the *Chlamydia* genus ranged from 69.2 to 96.9. Intragenus comparisons between members of the *Chlamydia* genus ranged from 27.4 to 69.1 (Table 15). The two genera are connected by the ROSA values of *C. felis*, *C. caviae*, and *C. muridarum* which had ROSA values ranging form 32.6 to 37.3 when compared between clusters (Table 15). Interestingly the organisms of the *Chlamydophila* cluster had higher than expected ROSA values when comparing *C. abortus*, *C. felis*, and *C. caviae*. The values ranged from 66.7 to 68.9, which are in the borderline range expected for strains of the same species (Table 15). This suggests that *C. abortus*, *C. felis*, and *C. caviae* may be all strains of the same species. In the *Chlamydia* cluster a similar trend occurred with *C. muridarum* having a ROSA score of 69.1 compared to two strains of *C. trachomatis* and a score of 55.6 when compared to the third. However since it had the higher than expected ROSA values compared to the type strain, ROSA values indicate that *C. muridarum* is a strain of the *C. trachomatis* species (Table 15).

16s rRNA % similarities also support a mixing of the two clusters. 16s rRNA % similarities within organisms of the *Chlamydophila* cluster ranged from 95.2 to 99.3 (Table 16). In particular *C. abortus*. *C. felis*, and *C. caviae* had higher than expected 16s rRNA % similarities, with values ranging from 98.3% to 99.3% (Table 16). These values suggest that *C. abortus*. *C. felis*, and *C. caviae* are highly related organisms. Organisms in

the *Chlamydia* cluster had a 16s rRNA % similarity of 98.4% (Table 16). This value is just below the species threshold of 98.5% suggesting that the organisms, if they are separate species, may have diverged very recently. Intercluster comparisons ranged from 93.8% to 95.9% similarity (Table 16). The organisms with values greater than the suggested 94%-95% species threshold were *C. felis*, *C. caviae*, and *C. muridarum* (Table 16). These values suggest that the two genera are either members of the same genus that will soon diverge fully, or members of different genera recently diverged.

AAI values show a slightly different cluster pattern than 16s rRNA and ROSA values. AAI values for intraspecies comparisons for members of the *Chlamydophila* cluster ranged from 99.8 to 99.9 while intragenus ranged from 67.5 to 84.9 (Table 17). Values within *C. abortus*, *C. caviae*, and *C. felis* ranged from 83.9 to 84.7 (Table 17). These values indicate that the three species diverged significantly from each other as the AAI species threshold is 95. Based on the AAI values the organisms are well classified as different species. Members of the *Chlamydia* cluster showed intraspecies AAI values ranging from 97.2 to 99.3 and intragenus values ranging from 84.9 to 85.2 (Table 17). These values suggest that the organisms are separated into two distinct taxonomic species, once again citing the 95 threshold for AAI. Intercluster comparisons showed values ranging from 61.7 to 67.9 (Table 17). The values for *C. pneumoniae* and members of the *Chlamydia* cluster are significantly lower, by 5 or more, than the values by *C. abortus*, *C. felis*, and *C. caviae*. This indicates that there are significant differences between *C. pneumoniae* and members of the *Chlamydophila* cluster. It is interesting that the values for the *C. abortus*, *C. felis*, and *C. caviae* when compared to those of the

*Chlamydophila* cluster nearly matched the values when compared to strains of *C. pneumoniae* (Table 17). This could indicate that *C. abortus*, *C. felis*, and *C. caviae* are in the process of diverging from both sets of organisms

%BBH values show a high level of similarity between the organisms. Intraspecies comparisons between members of the *Chlamydophila* cluster ranged from 97.1% to 99.3% of genome conserved with intragenus comparisons ranging from 83.9% to 98.4% of the genomes conserved (Table 18). Within this group two distinct subclusters could be observed, one consisting of *C. abortus*, *C. caviae*, and *C. felis* who had %BBH ranging from 91.7% to 98.4% while having values ranging from 83.9% to 93.7% when compared to the rest of the cluster (Table 18). This subclustering suggests a distinct difference between strains of *C. pneumoniae* and *C. abortus*, *C. felis*, and *C. caviae*. Furthermore the high genome similarity between *C. abortus*, *C. felis*, and *C. caviae* suggests that the organisms belong to the same species instead of the separate species they are now. Members of the *Chlamydia* cluster had intraspecies values ranging from 97.5% to 99.9% of the genomes conserved with *C. muridarum* having 93.7% to 98.6% of its genome conserved with members of its cluster (Table 18). Intercluster values ranged from 78% to 94.3% of the genomes conserved, with *C. trachomatis* strain D having extremely low values ranging from 44.3% to 45% (Table 18). *C. abortus*, *C. felis*, and *C. caviae* had higher %BBH when compared to members of the *Chlamydia* cluster than when compared to members of the *Chlamydophila* cluster, the opposite trend from AAI. This once again provides evidence for *C. abortus*, *C. felis*, and *C. caviae* acting as an intermediately diverging group separating the divergence of two other genera. Since the

values of those three species are intermediate between the other members of each cluster and show intense clustering together it provides evidence that the organisms may be diverging into their own genus relative to the other two clusters.

When the genomes of *C. pneumoniae*, *C. trachomatis*, *C. muridarum*, *C. caviae*, and *C. felis* were compared at the gene level it was evident that there was a large amount of similarity between the organisms with a core genome that consisted of 786 genes. Only *C. pneumoniae* had a significant number of unique genes 207 unique genes, *C. trachomatis* contained 41 unique genes, *C. muridarum* contained 47 unique genes, *C. caviae* contained 55 unique genes, while *C. felis* contained 20 unique genes (Figure 6). *C. pneumoniae*, *C. caviae*, and *C. felis* all shared a further 51 genes (Figure 6). *C. trachomatis* and *C. muridarum* shared a further 34 genes that encode for hypothetical proteins and specialized membrane proteins (Figure 6). *C. caviae*, and *C. felis* shared a further 58 genes unique to their genus that encode for a few specialized transport and adhesion proteins (Figure 6). Based on the number of genes shared between each combination it is easy to see that *C. pneumoniae* has a large number of genes different from the other members of the genus and as such probably diverged earlier from the others. *C. trachomatis* and *C. muridarum* shared enough genes to put them in their own cluster relative to other organisms as did *C. caviae* and *C. felis* (Figure 6). These gene count clusterings indicate that there are three distinct diverging clusters within the *Chlamydia*ceae family.

Based on the data it is likely that there are three separate genera emerging from a single parent genus. ROSA values indicate heavy clustering within the genus, with the entire genus have ROSA values that put organisms within the same species, to within different genera (Table 15). It was in the ROSA values that the first evidence arose for three separate divergences. Since members of the *Chlamydophila* cluster and the *Chlamydia* cluster were linked by ROSA scores from *C. felis*, *C. abortus*, and *C. caviae* it showed that the three clusters have only recently split form each other (Table 15). 16s rRNA scores supported this clustering. All organisms, except for two reciprocal comparisons, showed 16s similarity that was greater than the suggested threshold for different genera. Within the 16s three separate clusters of organisms emerged. The first cluster contained *C. pneumoniae*, the second contained *C. muridarum* and *C. trachomatis*, and the third contained *C. felis*, *C. abortus*, and *C. caviae* (Table 16). When these subclusters were compared to each other 16s values were just above those suggested for a genus threshold, suggesting that these organisms are in the process of fully diverging. AAI values for the organisms showed even more distinct clustering between the groups. Each subcluster had a high AAI to members within, but relatively low AAI values to other members of the genus (Table 17). %BBH values did not fully support this story due to the overall high levels of orthology between the organisms. Yet through examination, each of the subclusters show increased relatedness within themselves (Table 18). Gene counts show that the organisms are all highly similar to each other in protein functions yet show the beginnings of enough variation to warrant clustering. It is suggested that these subclusters are separated into different genera, the

first containing *C. pneumoniae*, the second consisting of *C. trachomatis* and *C. muridarum*, and a third consisting of *C. felis*, *C. abortus*, and *C. caviae*. It seems that the *Chlamydia* cluster diverged from *C. pneumoniae* first with the third cluster currently undergoing this process. AAI values between the clusters show that the first and the *Chlamydia* cluster are less similar than the third and first as do the %BBH values (Table 17, 18). Gene counts show this divergence pattern as well. *C. pneumoniae*, *C. caviae*, and *C. felis* share more genes between them than *C. trachomatis*, *C. pneumoniae* and *C. muridarum* share between them. Furthermore the number of genes shared uniquely between members of each cluster supports this classification. This cluster showcases ROSA's ability to analyze clusters that have not fully diverged and provide a pattern of evolution between clusters. Evidence for this diverging relationship is supported within the all species tree of life. The 16s tree for the *Chlamydia*ceae family shows branch splitting consistent with that suggested by ROSA (Yarza et al., 2008).

## H. Helicobacter:

The helicobacter genus was established in 1989 with the type species Helicobacter pylori being first described in 1985. It currently has 35 validly published members. Members of this genus are typically fastidious and slow growing with specific nutritional and environmental requirements. These organisms are found as pathogens, in a variety of species. They are gram negative and posses a helical shape. These organisms are able to survive in low pH, are normally highly susceptible to antibiotics, are highly motile with flagella, and are microaerophilic.

ROSA values for the Helicobacter genus revealed one distinct cluster and 4 organisms that did not cluster with any other organisms. The cluster consisted of the strains of *H. pylori* and *H. acinonychis*; the organisms that did not cluster well were *H. felis*, *H. mustelae*, *H. hepaticus*, and *H. canadensis*. Intraspecies comparisons in cluster 1 ranged from 81.6 to 86.8 with intragenus comparisons in cluster 1 ranged from 70 to 71 (Table 19). Organisms that did not cluster well had ROSA values ranging from 16.1 to 25.1, below the expected range for intragenus comparisons and in the range of intrafamily comparisons (Table 19). Based on the ROSA values the Helicobacter genus should be split into 5 different genera, one containing the members of the first cluster, and a genus for each of the other organisms. *H. pylori* and *H. acinonychis* are being classified as the same species by ROSA values.

16s rRNA % similarity revealed different clustering as compared to the ROSA values. 16s data revealed two separate clusters, one consisting of *H. pylori*, *H. acinonychis*, and *H. felis* with 16s rRNA values ranging from 95.0% to 98.2% (Table 20). The second cluster consisted of *H. mustelae*, *H. hepaticus*, and *H. canadensis* with 16s rRNA values ranging from 95.9% to 97.4% (Table 20). Intercluster values ranged from 92.8% to 94.3% (Table 20). The 16s values for cluster one were at the very borderline for the suggested 94-95% threshold for 16s rRNA genus delimitation. For *H. pylori* and *H. acinonychis* the 16s value was just below the 98.5% threshold for same species differentiation. For cluster 2 all values were well within the expected range for members of the same genus. Based on these values it seems that the Helicobacter genus has two major clusters. One of these clusters, cluster 2, has strong intragenus values while the

other, cluster 1, has borderline species and genus threshold values with the numbers falling just above or below the threshold.

AAI values for members of the helicobacter genus had either low levels of similarity between each other or high levels of similarity. Within the AAI values a single distinct cluster emerges containing strains of *H. pylori* and *H. acinonychis*. AAI values between the strains of H pylori range from 92.4 to 95.8 (Table 21). When *H. acinonychis* was compared to strains of *H. pylori* AAI values ranged from 89.7 to 90 (Table 21). Other members of the genus when compared to *H. pylori* strains, *H. acinonychis*, and each other had AAI values were well below 60 in all cases. (Table 21). The low values show that there is a low amount of relatedness between the organisms. *H. pylori* strains met the 94-95 AAI value threshold for members of the same species and therefore are correctly classified. *H. acinonychis* had AAI values that were under the range for members of the same species but still had high levels of similarities between itself and strains of *H. pylori*. These values predict a high level of similarity between the two organisms as shown by the ROSA and AAI values.

%BBH values revealed distinct clustering between the organisms. *H. pylori* strains had %BBH ranging from 90.4% to 94.6% of their genomes conserved. *H. pylori* stains and *H. acinonychis* had 83.3% to 86.8% of their genomes conserved (Table 22). Other members of the genus showed relatively low levels of orthology compared to each other, *H. pylori* strains, and *H. acinonychis*, had %BBH values ranged from 58.8% to 74.4% of the genomes conserved. This distinct clustering of *H. acinonychis* and *H. pylori*

indicates a high level of relatedness between the two that is not shared by other members of the genus.

GGDC values for members of the Helicobacter genus revealed distinct clustering. *H. canadensis*, *H. hepaticus*, *H. mustelae*, and *H. felis* all had GGDC values <70 (Table 23). This shows that the organisms are not within the same species of *H. pylori*, *H. acinonychis*, nor any other member in the genus. When *H. pylori* strains were examined all had GGDC values above the 70% DDH threshold for members of the same species (Table 23). When *H. acinonychis* was compared to strains of *H. pylori* values just below the threshold were obtained, ranging from 68.3% to 69.8% (Table 23). Although the values were not within the threshold it is important to remember that DDH, and its insilico counterpart have a minimum standard deviation of +or-3 (Guinebretière et al. in 2013). With this deviation applied the values between the organisms are truly borderline, furthermore the species definition is arbitrary. As such, the classification could go either way; it all depends on the characteristics.

When the genomes of *H. mustelae*, *H. felis*, *H. hepaticus*, *H. pylori*, and *H. acinonychis* were compared at the gene level the core genome consisted of 842 genes. The number of unique coding sequences each organisms contained followed no patterns *H. mustelae* contained 80 unique genes, *H. felis* contained 500 unique genes, *H. hepaticus* contained 621 unique genes, *H. pylori* contained 291 unique genes, while *H. acinonychis* contained 363 unique genes (Figure 7). *H. mustelae* contained an extremely low number of unique genes, below what would be expected for an organism falling so

strongly outside the rest of the organisms. *H. mustelae* and *H. hepaticus* shared a further 142 genes that consisted mostly of hypothetical proteins (Figure 7). *H. pylori* and *H. acinonychis* shared a further 163 unique genes (Figure 7). These genes encoded primarily metal regulating proteins, haloacid proteins expected due to their site of infection, and osmotic regulatory proteins. When unique genes between the *H. pylori* and *H. acinonychis* were compared it was found that the differences were composed of two different types of operons. The first was captured phage genomes, the second consisted of pathogenicity islands or vacuolating cytotoxins respectively (Table 24).

Based on the data a number of conclusions can be drawn about the Helicobacter genus. The first and most significant is that it should be divided into 5 different genera: the first containing *H. pylori* and *H. acinonychis*, the second containing *H. felis*, the third containing *H. mustelae*, the fourth containing *H. hepaticus*, and the fifth containing *H. canadensis*. ROSA values, DDH values, AAI values, BBH values, and gene levels all support the split. Although 16s rRNA similarities do not support the split this method is looking only at a single gene in an organism. Furthermore since the gene is so highly conserved it is likely that the high values comparing these organisms may be due to the environment they are found in, similar to the Sulfurimonas denitrificans misclassification (Takai et al., 2006). The second conclusion that can be drawn is that *H. pylori* and *H. acinonychis* are members of the same species that are close to diverging into separate species. ROSA values indicate that they are members of the same species while 16s rRNA values have them being borderline same species (Table 19 and 20). AAI and %BBH values show *H. acinonychis* clustering tightly with strains of *H. pylori*, not

enough to be a full member but close enough to the threshold to be close (Table 21 and 32). rDDH values showed high levels of similarity between the organisms classifying them as newly diverged or close to divergence, with standard deviation taken into consideration (Table 23). Most significantly the organisms share a large number of genes unique to them that other members of the genus do not contain. The differences in their genes primarily concern captured phage genomes and different toxins that they produce and secrete (Figure 7). Eppinger and his colleagues found that *H. acinonychis* was derived from a large cat consuming a human infected with *H. pylori* at most 200,000 and possibly as short as 100,000 years ago. Changes between the organisms are small and consist primarily of a 5 membrane proteins that allows it to evade the immune surveillance of the big cat host. Interestingly enough some strains of *H. acinonychis* are more similar to African strains of *H. pylori* than some African *H. pylori* strains are to European strains of *H. pylori* (Eppinger et al. 2006). It is most likely that the two species became fully isolated from each other only recently, as big cats and humans regularly hunted each other until recently, and as such are only now fully undergoing the speciation process. Based on this it is recommended that despite the values that classify them as the same species the two remain separate species as divergence is happening, especially with the quick decline in cheetah populations creating a bottleneck.

## I. Flavobacteriacae:

ROSA values for the Flavobacteriacae family showed two distinct clusters emerging from it. The Chryseobacterium cluster consisted of Chryseobacterium gleum, Epilithonimonas tenax, Elizabethkingia meningoseptica, and Empedobacter brevis with ROSA values ranging from 16.3 to 30.0, the expected range for members of the same family (Table 25). The Flavobacterium cluster consisted of Flavobacterium johnsoniae, Mesonia mobilis, Gaetbulibacter saemankumensis, Gelidibacter mesophilus, Gillisia limnaea, Leeuwenhoekiella blandensis, and Polaribacter franzmannii and had ROSA values ranging from 15.3 to 28.1, values within the expected range for members of the same family (Table 25). When the two clusters were compared to each other ROSA values ranged from 7.9 to 13.4 values expected for different families in the same order (Table 25). Based on these values it is expected that the two clusters are separate families within the same order.

16s rRNA similarity values almost mirror this relationship. Chryseobacterium gleum, Chryseobacterium meningoseptica, and Epilithonimonas tenax clustered with each other with 16s rRNA % similarities ranging from 94.3% to 95%, within the possible same family range (Table 26). The second cluster consisted of Flavobacterium johnsoniae, Mesonia mobilis, Gaetbulibacter saemankumensis, Gelidibacter mesophilus, Gillisia limnaea, Leeuwenhoekiella blandensis, and Polaribacter franzmannii which had 16s rRNA % similarity ranging from 87.5% to 92.9%, within the suspected same family range (Table 26). Empedobacter brevis had 16s rRNA similarity with values ranging from 39.3% to 40%, this is thought to be caused by a bad sequence (Table 26). 16s rRNA %

similarity values ranged between the clusters ranged from 81.6% to 84.0%, showing that the two clusters were well separated by 16s rRNA (Table 26).

AAI values formed a similar clustering pattern as the ROSA values. The first cluster consisted of Chryseobacterium gleum, Elizabethkingia meningoseptica and Epilithonimonas tenax with AAI values ranging from 65.3 to 69.8, although no known threshold exists for family level AAI values it can be concluded that these organisms do share some level of similarity (Table 27). The second cluster contained Flavobacterium johnsoniae, Mesonia mobilis, Gaetbulibacter saemankumensis, Gelidibacter mesophilus, Gillisia limnaea, Leeuwenhoekiella blandensis, and Polaribacter franzmannii with AAI values ranging from 55.6 to 66.0, showing similar values to the first cluster (Table 27). Empedobacter brevis had no clustering and AAI values ranging from 54.2 to 47.9 (Table 27). When the two clusters were compared between each other AAI values ranged from 47.1 to 49.8, showing a clear divergence between the two clusters (Table 27).

%BBH values showed similar clustering as AAI. Mesonia mobilis, Gaetbulibacter saemankumensis, Gelidibacter mesophilus, Gillisia limnaea, and Leeuwenhoekiella blandensis had high levels of genome conservation ranging from 60.4% to 76.2% (Table 28). These values indicate a high level of similarity between the genera. Chryseobacterium gleum, Elizabethkingia meningoseptica and Epilithonimonas tenax had high levels of similarity as well, ranging from 71.4% to 74.7% (Table 28). When the clusters were compared against each other %BBH values ranged from 34.8% to 53.4%, providing a distinct separation between the genera. Polaribacter franzmannii had the lowest genome conservation with other members of the family with values ranging from

29.5 to 37.2% of the genome conserved (Table 28). This low level of orthology is most likely due to genome size between the organisms.

Overall it is suggested that the Flavobacteriacae family be split into two separate families under the same order. ROSA values showcase a distinct split between the family with values suggesting the above mentioned taxonomic strain. AAI, %BBH, and 16s rRNA similarities did not provide a distinct threshold as none currently exist at the family level for the respective methods. However they provide distinct clustering between the two proposed families and low levels of similarity between the two. Furthermore work done in the Newman lab has found that the Chryseobacterium genus is sufficiently different from the Flavobacterium to warrant its own family. Other genera that show this same difference to Flavobacterium and relatedness to Chryseobacterium include the Elizabethkingia and Epilithonimonas genera, which match the ROSA suggested taxonomy (Data unpublished).

## J. Overall Conclusions:

A number of overall conclusions can be drawn from this study. The most significant one is the validity and resolution of the ROSA tool. All other current taxonomic metrics have only a well accepted species threshold. A few such, as AAI and 16s rRNA, have proposed genus level boundaries, but none are well accepted. ROSA is the only tool that provides thresholds at all levels of taxonomy, making it superior to any other system currently in use. Furthermore it is also one of the easiest methods to use. A number of methods require difficult wet lab procedures that are highly error

prone, such as DDH; have unclear calculations, such as GGDC; or require a level of opinion that renders its validity suspect, such as MLSA. ROSA relies on clear, easy to understand metrics, and operates on methodologies based on biological concepts. Another advantage of the ROSA metric is its ability to be a whole genome comparison metric. Many methods, such as ANI and AAI, only focus on the relatedness of the conserved sections of the genome. As such they are unable to fully account for the true relatedness of two organisms at a whole genomic level. By comparing the similarities of the orthologous sections and factoring in how much of the genome is orthologous, ROSA provides much greater resolution at all levels of taxonomy. One final advantage of the ROSA metric is that it is a reciprocal comparison and as such has a built in correction against bias by genome size. A final conclusion is a tentative AAI genus level threshold and a tentative %BBH species level threshold based on the values produced in this study. It is proposed that a genus level AAI threshold be at the low 70's, approximately 72-74. A tentative species threshold for %BBH is proposed for 77% to 80%.

A possible weakness of ROSA is the classification of extremophiles. In extreme environments the selection acts quicker and allows less deviation. Since AAI forms such a crucial part in the ROSA calculation an inability to change amino acids over evolutionary time will raise the ROSA values above what would be expected. This can be seen when comparing *Thiomicrospira crunogens* to *Sulfurimonas denitrificans*. The ROSA values were not what would be expected, being one level of taxonomy higher than what would be predicted (Table 4). This weakness s shared with all other genomic characterization tools.

This is only a select sample of the possible number of species that require classification. Currently only a small percentage of known, validly published bacterial species have full genomic sequences. Of these, only a small percentage are type strain sequences of the organism. As such there is still a large amount of work to be done in order to correct microbial taxonomy. A list of additional analyzed clusters can be found in the Appendix.

## V. References:


Achtman, M., Zurth, K., Morelli, G., Torrea, G., Guiyoule, A. & Carniel, E. (1999). Yersinia pestis, the cause of plague, is a recently emerged clone of Yersinia pseudotuberculosis. *Proc Natl Acad Sci USA*96, 14043–8.

Adékambi, T., Shinnick, T. M., Raoult, D. & Drancourt, M. (2008). Complete rpoB gene sequencing as a suitable supplement to DNA-DNA hybridization for bacterial species and genus delineation. *Int J Syst Evol Microbiol* 58, 1807–14.

Auch, A. F., Jan, M. von, Klenk, H.-P. P. & Göker, M. (2010). Digital DNA-DNA hybridization for microbial species delineation by means of genome-to-genome sequence comparison. *Stand Genomic Sci* 2, 117–34.

Aziz, R.K., Bartels, D., Best, A.A., DeJongh, M., Disz, T., Edwards, R.A., Formsma, K., Gerdes, S., Glass, E.M. & other authors (2008). The RAST Server: rapid annotations using subsystems technology. *BMC Genomics* 9, 75.

Bavykin, S. G., Lysov, Y. P., Zakhariev, V., Kelly, J. J., Jackman, J., Stahl, D. A. & Cherni, A. (2004).Use of 16S rRNA, 23S rRNA, and gyrB gene sequence analysis to determine phylogenetic relationships of Bacillus cereus group microorganisms. *J Clin Microbiol*42, 3711–30.

Bergey, D. H., F. C. Harrison, R. S. Breed, B. W. Hammer, and F. M. Huntoon (ed.). 1923. Bergey's manual of determinative bacteriology, p. 116. The Wil- liams & Wilkins Co., Baltimore.

Buonaccorsi, Vincent, Mark Peterson, Gina Lamendella, Jeff Newman, Nancy Trun, Tammy Tobin, Andres Aguilar, Arthur Hunt, Craig Praul, Deborah Grove, Jim Roney, and Wade Roberts. "Vision and Change through the Genome Consortium for Active Teaching Using Next-Generation Sequencing (GCAT-SEEK)." *CBE Life Sci Educ* 13.1 (2014): 1-2. Web.

Bustard, K. & Gupta, R. S. (1997). The sequences of heat shock protein 40 (DnaJ) homologs provide evidence for a close evolutionary relationship between the Deinococcus-thermus group and cyanobacteria. *J Mol Evol* 45, 193–205.

Ceccarelli, D. & Colwell, R. R. (2014). *Vibrio* ecology, pathogenesis, and evolution. *Front Microbiol* 5, 256.

Choi, J.-H. H., Lee, K. M., Lee, M.-K. K., Cha, C.-J. J. & Kim, G.-B. B. (2014). Bifidobacterium faecale sp. nov., isolated from human faeces. *Int J Syst Evol Microbiol* 64, 3134–9

Cole, S. T., Brosch, R., Parkhill, J., Garnier, T., Churcher, C., Harris, D., Gordon, S. V., Eiglmeier, K., Gas, S. & other authors. (1998). Deciphering the biology of *Mycobacterium* tuberculosis from the complete genome sequence. *Nature* 393, 537–44.

Devulder, G., Pérouse de Montclos, M. & Flandrois, J. P. (2005). A multigene approach to phylogenetic analysis using the genus *Mycobacterium* as a model. *Int J Syst Evol Microbiol* 55, 293–302.

Eisen, J. A. (1995). The RecA protein as a model molecule for molecular systematic studies of bacteria: comparison of trees of RecAs and 16S rRNAs from the same species. *J Mol Evol* 41, 1105–23.

Eppinger, M., Baar, C., Linz, B., Raddatz, G., Lanz, C., Keller, H., Morelli, G., Gressmann, H., Achtman, M. & Schuster, S. C. (2006). Who ate whom? Adaptive Helicobacter genomic changes that accompanied a host jump from early humans to large felines. *PLoS Genet* 2, e120

Everett, K. D., Bush, R. M. & Andersen, A. A.(1999). Emended description of the order *Chlamydia*les, proposal of Para*Chlamydia*ceae fam. nov. and Simkaniaceae fam. nov., each containing one monotypic genus, revised taxonomy of the family *Chlamydia*ceae, including a new genus and five new species, and standards for the identification of organisms. *Int J Syst Bacteriol* 49 Pt 2, 415–40.

Facklam, R. (2002). What happened to the streptococci: overview of taxonomic and nomenclature changes. *Clin Microbiol Rev* 15, 613–30.

Farrand, S. K., Berkum, P. B. Van & Oger, P.(2003). Agrobacterium is a definable genus of the family Rhizobiaceae. *Int J Syst Evol Microbiol* 53, 1681–7.

Gilbert, M. J., Kik, M., Miller, W. G., Duim, B. & Wagenaar, J. A. (2015). *Campylobacter* iguaniorum sp. nov., isolated from reptiles. *Int J Syst Evol Microbiol*65, 975–82.

Glazunova, O. O., Raoult, D. & Roux, V. (2009).Partial sequence comparison of the rpoB, sodA, groEL and gyrB genes within the genus *Streptococcus*. *Int J Syst Evol Microbiol* 59, 2317–22.

Goodwin, C. S., J. A. Armstrong, T. Chilvers, M. Peters, M. D. Collins, L. Sly, W. McConnell, and W. E. S. Harper. (1989). Transfer of *Campylobacter pylori* and *Campylobacter mustelae* to *Helicobacter* gen. nov. as *Helicobacter pylori* comb. nov. and *Helicobacter mustelae* comb. nov., respectively. Int. J. Syst. Bacteriol. 39:397–405.

Goris, J., Konstantinidis, K. T., Klappenbach, J. A., Coenye, T., Vandamme, P. & Tiedje, J. M. (2007).DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. *Int J Syst Evol Microbiol* 57, 81–91.

Goudenège, D., Labreuche, Y., Krin, E., Ansquer, D., Mangenot, S., Calteau, A., Médigue, C., Mazel, D., Polz, M. F. & Roux, F. Le. (2013). Comparative genomics of pathogenic lineages of *Vibrio* nigripulchritudo identifies virulence-associated traits.*ISME J* 7, 1985–96

Guinebretière, M.-H. H., Auger, S., Galleron, N., Contzen, M., Sarrau, B. De, Buyser, M.-L. L. De, Lamberet, G., Fagerlund, A., Granum, P. E. (2013). Bacillus cytotoxicus sp. nov. is a novel thermotolerant species of the Bacillus cereus Group occasionally associated with food poisoning. *Int J Syst Evol Microbiol* 63, 31–40.

Helgason, E., Okstad, O. A., Caugant, D. A., Johansen, H. A., Fouet, A., Mock, M., Hegna, I. & Kolstø, A. B. (2000). Bacillus anthracis, Bacillus cereus, and Bacillus thuringiensis--one species on the basis of genetic evidence. *Appl Environ Microbiol* 66, 2627–30.

Holmes, B., R. J. Owen, and T. A. McMeekin. (1984). Genus *Flavobactenum* Bergey, Harrison, Breed, Hammer and Huntoon 1923, 97AL,p. 353-361. *In* N. R. Krieg and J. G. Holt (ed.), Bergey's manual of systematic bacteriology, vol. 1. The Williams and Wilkins Co., Baltimore.

Kim, O.S., Cho, Y.J., Lee, K., Yoon, S.H., Kim, M., Na, H., Park, S.C., Jeon, Y.S., Lee, J.H., Yi, H., Won, S., Chun, J. (2012). Introducing EzTaxon: a prokaryotic 16S rRNA Gene sequence database with phylotypes that represent uncultured species. *Int J Syst Evol Microbiol* 62, 716–721.

Konstantinidis, K.T. and Tiedje, J.M. (2005). Genomic insights that advance the species definition for prokaryotes. *Proc Natl Acad Sci USA* 102, 2567-2572

Konstantinidis, K. T. & Tiedje, J. M. (2005).Towards a genome-based taxonomy for prokaryotes. *J Bacteriol* 187, 6258–64.

Lan, R. & Reeves, P. R. (2002). Escherichia coli in disguise: molecular origins of Shigella. *Microbes Infect* 4, 1125–32.

Lukjancenko, O., Wassenaar, T. M. & Ussery, D. W. (2010). Comparison of 61 sequenced Escherichia coli genomes. *Microb Ecol* 60, 708–20

Mattarelli, P., Bonaparte, C., Pot, B. & Biavati, B.(2008). Proposal to reclassify the three biotypes of Bifidobacterium longum as three subspecies: Bifidobacterium longum subsp. longum subsp. nov., Bifidobacterium longum subsp. infantis comb. nov. and Bifidobacterium longum subsp. suis comb. nov. *Int J Syst Evol Microbiol* 58, 767–72.

Meier-Kolthoff, J. P., Hahnke, R. L., Petersen, J., Scheuner, C., Michael, V., Fiebig, A., Rohde, C., Rohde, M., Fartmann, B. & other authors. (2014).Complete genome sequence of DSM 30083(T), the type strain (U5/41(T)) of Escherichia coli, and a proposal for delineating subspecies in microbial taxonomy. *Stand Genomic Sci* 9, 2.

Parte, A. C. (2014). LPSN--list of prokaryotic names with standing in nomenclature. *Nucleic Acids Res* 42, D613–6.

Palleroni, N. J. (2010). The Pseudomonas story. *Environ Microbiol* 12, 1377–83.

Poyart, C., Quesne, G., Coulon, S., Berche, P. & Trieu-Cuot, P. (1998). Identification of streptococci to species level by sequencing the gene encoding the manganese-dependent superoxide dismutase. *J Clin Microbiol* 36, 41–7.

Reichenbach, H. (1989). Order I. *Cytophagales* Leadbetter 1974, 99*'-, p. 2011-2013. *In* J. T. Staley, M. P. Bryant, N. Pfennig, and J. G. Holt (ed.), Bergey's manual of systematic bacteriology, vol. 3. The Williams and Wilkins Co., Baltimore.

Richer, L., Golubic, S., Guédès, R. L., Ratiskol, J., Payri, C. & Guezennec, J. (2005). Characterization of exopolysaccharides produced by cyanobacteria isolated from Polynesian microbial mats. *Curr Microbiol*51, 379–84.

Richter, M. & Rosselló-Móra, R. (2009). Shifting the genomic gold standard for the prokaryotic species definition. *Proc Natl Acad Sci USA* 106, 19126-19131.

Sawabe, T., Ogura, Y., Matsumura, Y., Feng, G., Amin, A. R., Mino, S., Nakagawa, S., Sawabe, T., Kumar, R. & other authors. (2013). Updating the *Vibrio* clades defined by multilocus sequence phylogeny: proposal of eight new clades, and the description of *Vibrio* tritonius sp. nov. *Front Microbiol* 4, 414.

Schachter, J., Stephens, R. S., Timms, P., Kuo, C., Bavoil, P. M., Birkelund, S., Boman, J., Caldwell, H., Campbell, L. A. & other authors. (2001). Radical changes to *Chlamydia*l taxonomy are not necessary just yet. *Int J Syst Evol Microbiol* 51, 249; author reply 251–253

Schleifer, K. H. (2009). Classification of Bacteria and Archaea: past, present and future. *Syst Appl Microbiol*32, 533–42.

Stackebrandt, E., Frederiksen, W., Garrity, G. M., Grimont, P. A., Kämpfer, P., Maiden, M. C., Nesme, X., Rosselló-Mora, R., Swings, J. & other authors. (2002). Report of the ad hoc committee for the re-evaluation of the species definition in bacteriology. *Int J Syst Evol Microbiol* 52, 1043–7.

Stackebrandt, E. & Ebers, J. (2006). Taxonomic parameters revisited: tarnished gold standards. Microbiol Today 33, 152–155.

Stackebrandt,E. ( 2006 ). The history of microbial species definitions, conventions:. In: Reconciling Microbial Systematics & Genomics :3-3.

Stahl, D. A. & Urbance, J. W. (1990). The division between fast- and slow-growing species corresponds to natural relationships among the mycobacteria. *J Bacteriol*172, 116–24.

Takahashi, M., Kryukov, K. & Saitou, N. (2009).Estimation of bacterial species phylogeny through oligonucleotide frequency distances. *Genomics* 93, 525–33.

Takai, K., Suzuki, M., Nakagawa, S., Miyazaki, M., Suzuki, Y., Inagaki, F. & Horikoshi, K. (2006).Sulfurimonas paralvinellae sp. nov., a novel mesophilic, hydrogen- and sulfur-oxidizing chemolithoautotroph within the Epsilonproteobacteria isolated from a deep-sea hydrothermal vent polychaete nest, reclassification of Thiomicrospira denitrificans as Sulfurimonas denitrificans comb. nov. and emended description of the genus Sulfurimonas. *Int J Syst Evol Microbiol* 56, 1725–33.

Tamura K., Stecher G., Peterson D., Filipski A., and Kumar S. (2013). MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0. *Molecular Biology and Evolution* 30: 2725-2729.

Tsuchida, S., Kitahara, M., Nguema, P. P., Norimitsu, S., Fujita, S., Yamagiwa, J., Ngomanda, A., Ohkuma, M. & Ushida, K. (2014).*Lactobacillus* gorillae sp. nov., isolated from the faeces of captive and wild western lowland gorillas (Gorilla gorilla gorilla). *Int J Syst Evol Microbiol* 64, 4001–6

Teng, L.-J. J., Hsueh, P.-R. R., Tsai, J.-C. C., Chen, P.-W. W., Hsu, J.-C. C., Lai, H.-C. C., Lee, C.-N. N. & Ho, S.-W. W. (2002). groESL sequence determination, phylogenetic analysis, and species differentiation for viridans group streptococci. *J Clin Microbiol* 40, 3172–8.

Tindall, B. J., Rosselló-Móra, R., Busse, H.-J. J., Ludwig, W. & Kämpfer, P. (2010). Notes on the characterization of prokaryote strains for taxonomic purposes. *Int J Syst Evol Microbiol* 60, 249–66.

Vandamme,P., J.-F. Bernardet, P Segers,K. Kersters,and B. Holmes. (1994). New perspectives in the classification of the flavobacteria: description of *Chryseobacterium* gen. nov., *Beigeyella* gen. nov., and *Ernpedobacter*nom. rev. Int. J. Syst. Bacteriol. 44827-831.

Verger, J.-M., F. Grimont, P. A. D. Grimont, and M. Grayon. (1985). "Brucella, a Monospecific Genus as Shown by Deoxyribonucleic Acid Hybridization." *International Journal of Systematic Bacteriology* 35.3:292-95.

Wayne, L. G. (1988). International Committee on Systematic Bacteriology: announcement of the report of the ad hoc Committee on Reconciliation of Approaches to Bacterial Systematics. *International Journal of Systematic Bacteriology*. 268, 433–4.

Woese, C. R., Kandler, O. & Wheelis, M. L. (1990).Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proc Natl Acad Sci USA* 87, 4576–9.

Xu, D. & Côté, J.-C. C. (2003). Phylogenetic relationships between Bacillus species and related genera inferred from comparison of 3’ end 16S rDNA and 5' end 16S-23S ITS nucleotide sequences. *Int J Syst Evol Microbiol* 53, 695–704

Yarza, P., M. Richter, J. Peplies, J. Euzeby, R. Amann, K. H. Schleifer, W. Ludwig, F. O. Glöckner, and R. Rossello-Mora. 2008. The All-Species Living Tree project: A 16S rRNA-based phylogenetic tree of all sequenced type strains. Syst Appl Microbiol 31:241-250

Yabuuchi, E., T. Kaneko, I. Yano, C. W. Moss, and N. Miyoshi. (1983). *Sphingobacterium* gen. nov., *Sphingobacterium spiritivomm* comb. nov., *Sphingobacterium multivorum* comb. nov., *Sphingobacteriumrnizutae* sp. nov., and *Flavobacterium indologenes* sp. nov.: glucose-nonfermenting gram-neg- ative rods in CDC groups IIk-2 and IIb. Int. J. Syst. Bacteriol. 33580-598

Zhu, D., Tanabe, S.-H. H., Xie, C., Honda, D., Sun, J. & Ai, L. (2014). Bacillus ligniniphilus sp. nov., an alkaliphilic and halotolerant bacterium isolated from sediments of the South China Sea. *Int J Syst Evol Microbiol* 64, 1712–7.

# VII Appendix:

## A. Bifidobacterium

| Bifidobacterium | ROSA | 580050 | 555970 | 442563 | 302912 | 367928 | 566552 | 518635 | 473819 | 398513 | 206672 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Bifidobacterium animalis lactis Bl-04, ATCC SD5219 | **580050** | | | | | | | | | | |
| Bifidobacterium animalis lactis DSM 10140$^{Tsubsp}$ | **555970** | 99.6 | | | | | | | | | |
| Bifidobacterium animalis subsp. lactis AD011 | **442563** | 98.5 | 98.3 | | | | | | | | |
| Bifidobacterium animalis animalis ATCC 25527$^{T}$ | **302912** | 85.1 | 85.0 | 84.8 | | | | | | | |
| Bifidobacterium adolescentis ATCC 15703$^{T}$ | **367928** | 35.0 | 35.0 | 35.1 | 35.1 | | | | | | |
| Bifidobacterium catenulatum DSM 16992 | **566552** | 34.6 | 34.7 | 34.8 | 34.5 | 60.4 | | | | | |
| Bifidobacterium angulatum DSM 20098$^{T}$ | **518635** | 33.3 | 33.3 | 33.4 | 33.5 | 45.8 | 45.1 | | | | |
| Bifidobacterium dentium ATCC 27678 | **473819** | 32.9 | 32.9 | 33.0 | 32.8 | 52.0 | 51.4 | 40.4 | | | |
| Bifidobacterium bifidum NCIMB 41171 | **398513** | 31.1 | 31.1 | 31.2 | 31.3 | 37.2 | 37.7 | 36.3 | 33.4 | | |
| Bifidobacterium longum NCC2705 | **206672** | 32.2 | 32.2 | 32.3 | 32.3 | 39.9 | 40.5 | 38.7 | 36.4 | 40.3 | |
| Bifidobacterium longum subsp. infantis ATCC 15697$^{Tsubsp}$ | **391904** | 29.3 | 29.3 | 29.3 | 29.3 | 36.7 | 36.0 | 35.7 | 32.1 | 37.4 | 63.7 |

Table 29: ROSA values for members of the *Bifidobacterium* genus. Intraspecies comparisons were done between *B. animalis* and strains of *B. longum*. Intragenus comparisons were done between strains of *B. animalis, B. adolescentis, B. catenulatum, B. angulatum, B. dentium, B. bifidum*, and *B. longum*.

ROSA values for members of the *Bifidobacterium* genus ranged from 29.3 to 99.6 (Table 29). Intraspecies comparisons between strains of *B. animalis* ranged between 84.8 and 99.6 while values between strains of *B. longum* was 40.3 (Table 29). Two distinct clusters were observed in the genus. One consisting of *B. animalis* and *B. adolescentis* had ROSA values ranging form 35 to 99.6. The second consisting of *B. catenulatum, B. angulatum, B. dentium, B. bifidum*, and *B. longum* had ROSA values ranging from 32.1 to 60.4 (Table 29). ROSA values between the two clusters ranged from 29.3 to 34.6 (Table 29). ROSA values suggest splitting the genus into two separate genera consisting of first cluster and the second cluster.

| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|---|
| Bifidobacterium_animalis_subsp._animalis_ATCC_25527T | 1 | | | | | | | |
| Bifidobacterium_animalis_subsp._lactis_DSM_10140T | 2 | 99.0 | | | | | | |
| Bifidobacterium_catenulatum_DSM_16992T | 3 | 93.8 | 94.4 | | | | | |
| Bifidobacterium_adolescentis_ATCC_15703T | 4 | 93.1 | 93.1 | 96.1 | | | | |
| Bifidobacterium_angulatum_DSM_20098T | 5 | 93.7 | 93.9 | 97.3 | 96.6 | | | |
| Bifidobacterium_dentium_ATCC_27534T | 6 | 93.1 | 93.5 | 96.9 | 96.7 | 95.7 | | |
| Bifidobacterium_bifidum_ATCC_29521T | 7 | 93.0 | 93.0 | 95.3 | 94.7 | 95.0 | 95.4 | |
| Bifidobacterium_longum_subsp._infantis_ATCC_15697T | 8 | 93.5 | 93.4 | 95.2 | 95.2 | 95.7 | 95.4 | 95.6 |

Table 30: 16s rRNA % similarity values for members of the *Bifidobacterium* genus.

Members of the *Bifidobacterium* genus showed differing levels of 16s rRNA similarity and formed two distinct clusters with one cluster consisting of a single species. 16s rRNA values within the genus varied from 99% similarity to as low as 93% similarity. Organisms that clustered together, *B. catenulatum, B. adolescentis, B. angulatum, B. dentium, B. bifidum*, and *B. longum*, had 16s rRNA similarities ranging form 94.7% to 97.3% similarity (Table 30).

| | Average Amino Acid Identity (AAIr) | 367928 | 518635 | 302912 | 580050 | 555970 | 398513 | 566552 | 473819 | 442563 | 206672 | 391904 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Bifidobacterium adolescentis ATCC 15703$^T$ | 367928 | | 75.2 | 67.3 | 67.5 | 68 | 72 | 85 | 82 | 67.7 | 72.5 | 72.3 |
| Bifidobacterium angulatum DSM 20098$^T$ | 518635 | 75.3 | | 66.0 | 66.2 | 66 | 71 | 75 | 74 | 66.4 | 72.2 | 72.5 |
| Bifidobacterium animalis animalis ATCC 25527$^T$ | 302912 | 67.5 | 65.9 | | 95.6 | 96 | 65 | 68 | 67 | 95.8 | 66.6 | 66.1 |
| Bifidobacterium animalis lactis BI-04, | 580050 | 67.6 | 66.1 | 96.0 | | 100 | 65 | 67 | 68 | 99.8 | 66.5 | 66.2 |
| Bifidobacterium animalis lactis DSM 10140T | 555970 | 67.7 | 66.1 | 96.0 | 99.9 | | 65.5 | 67.5 | 67.6 | 99.8 | 66.5 | 66.2 |
| Bifidobacterium bifidum NCIMB 41171 | 398513 | 71.9 | 70.9 | 65.5 | 65.9 | 65.9 | | 71.9 | 71.1 | 66.1 | 73.3 | 73.6 |
| Bifidobacterium catenulatum DSM 16992 | 566552 | 85.4 | 75.3 | 67.5 | 67.4 | 67.5 | 71.9 | | 82.3 | 67.7 | 72.2 | 71.6 |
| Bifidobacterium dentium ATCC 27678 | 473819 | 82.3 | 74.0 | 67.2 | 67.4 | 67.4 | 71.0 | 82.4 | | 67.6 | 71.5 | 71.2 |
| Bifidobacterium animalis subsp. lactis AD011 | 442563 | 67.7 | 66.2 | 96.0 | 99.8 | 99.8 | 65.9 | 67.7 | 67.8 | | 66.5 | 66.4 |
| Bifidobacterium longum NCC2705 | 206672 | 72.6 | 72.4 | 66.7 | 66.4 | 66.5 | 73.2 | 72.2 | 71.9 | 66.5 | | 93.5 |
| Bifidobacterium longum subsp. infantis ATCC 15697 | 391904 | 72.2 | 72.5 | 66.1 | 65.9 | 66.0 | 73.0 | 71.6 | 71.2 | 66.1 | 93.1 | |

Table 31: AAI values for members of the *Bifidobacterium* genus.

AAI values for members of the *Bifidobacterium* genus showed large levels of variation between different organisms. Strains of the same organism always had AAI over 93% with most having over 95% (Table 31). Comparisons between different species had values that ranged from 85.2 to 65.5 with most species falling in the mid-70's when compared to each other. Distinct differences were shown between strains of *B. animalis* and other members of the genus. *B. animalis* had the lowest intragenus comparison values of all organisms. All other organisms showed values greater than 70% AAI when compared to each other (Table 31).

| | Percent Bidirectional Best Hit (% BBH) | 367928 | 518635 | 302912 | 580050 | 555970 | 398513 | 566552 | 473819 | 442563 | 206672 | 391904 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Bifidobacterium adolescentis ATCC 15703$^T$ | **367928** | | 81.8 | 80.2 | 79.3 | 79.2 | 70.6 | 82.8 | 67.4 | 79.2 | 71.9 | 60.6 |
| Bifidobacterium angulatum DSM 20098$^T$ | **518635** | 80.0 | | 79.1 | 77.7 | 77.6 | 70.5 | 78.5 | 63.8 | 77.5 | 69.6 | 58.0 |
| Bifidobacterium animalis animalis ATCC 25527$^T$ | **302912** | 74.2 | 74.8 | | 92.4 | 92.3 | 69.2 | 72.4 | 60.7 | 91.9 | 66.1 | 55.7 |
| Bifidobacterium animalis lactis BI-04 | **580050** | 74.1 | 74.6 | 93.0 | | 99.6 | 68.5 | 73.0 | 60.7 | 98.8 | 66.5 | 55.8 |
| Bifidobacterium animalis lactis DSM 10140$^T$ | **555970** | 74.1 | 74.6 | 93.0 | 99.6 | | 68.5 | 73.0 | 60.7 | 98.7 | 66.5 | 55.8 |
| Bifidobacterium bifidum NCIMB 41171 | **398513** | 73.5 | 74.3 | 77.0 | 75.4 | 75.4 | | 74.1 | 59.1 | 75.3 | 72.9 | 61.3 |
| Bifidobacterium catenulatum DSM 16992 | **566552** | 83.2 | 80.8 | 79.1 | 79.3 | 79.2 | 71.8 | | 66.8 | 79.0 | 74.2 | 61.3 |
| Bifidobacterium dentium ATCC 27678 | **473819** | 86.5 | 84.0 | 84.3 | 83.8 | 83.8 | 73.2 | 84.8 | | 83.6 | 76.5 | 62.9 |
| Bifidobacterium animalis subsp. lactis AD011 | **442563** | 74.0 | 74.6 | 92.6 | 98.8 | 98.6 | 68.0 | 72.7 | 60.4 | | 66.6 | 55.4 |
| Bifidobacterium longum NCC2705 | **206672** | 79.7 | 78.5 | 79.5 | 79.3 | 79.2 | 77.3 | 81.2 | 65.1 | 79.4 | | 66.6 |
| Bifidobacterium longum subsp. infantis ATCC 15697 | **391904** | 79.7 | 77.7 | 78.4 | 78.6 | 78.554 | 77.9 | 79.0 | 63.7 | 78.2 | 79.8 | |

Table 32: %BBH values for members of the *Bifidobacterium* genus.

Orthology for members of the *Bifidobacterium* genus showed great levels of variation between strains of the same species as well as members of the genus. *B. animalis* strains showed high levels of orthology, with all strains sharing well over 92.6 % of their genome (Table 32). Strains of *B. longum* showed low levels of orthology, sharing ~less than 70% of their genome. Members of the genus showed great variety in their orthology to each other. *B. longum* strains shared on average less than 65% of their genome to other members of the genus, with some strains sharing as little as 55%. *B.*

*bifidum* and *B. dentium* also had low levels of orthology to other members of their genus, sharing as little as 68% and 60% of their genomes respectively (Table 32). Other members of the genus shared well over 70% of their genome to each other, excluding the aforementioned cases (Table 32).

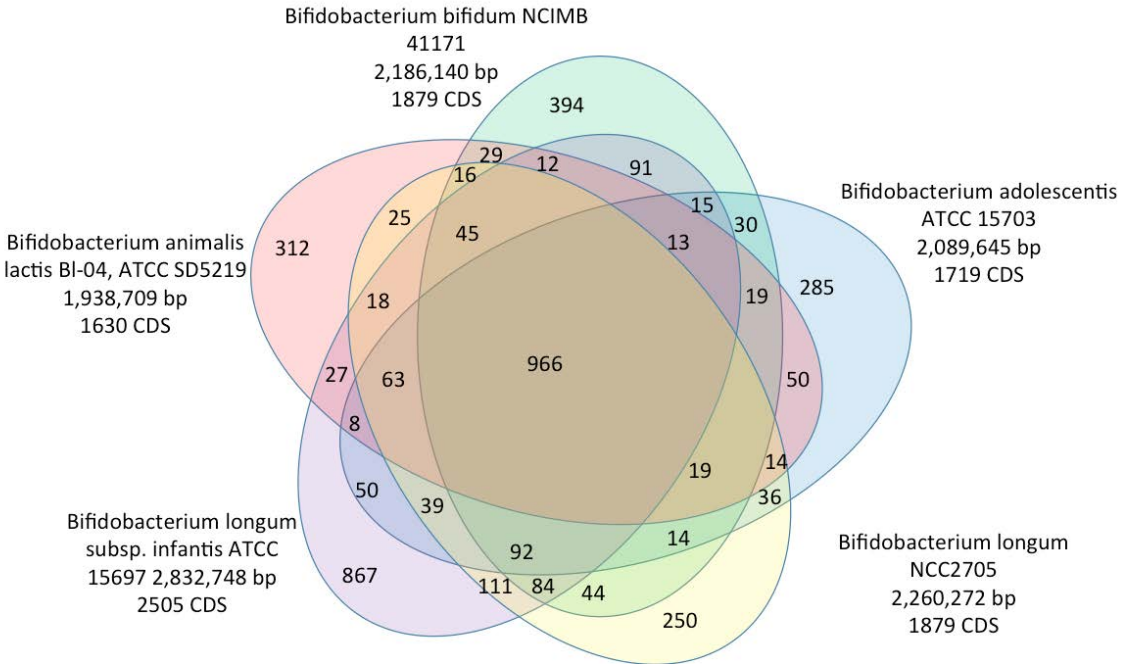


Figure 8: **Genomic similarity at the gene level for *B. longum subsp.infantis*, *B. animalis*, *B. longum T*, *B. bifidum*, and *B. adolescentis*.** Venn-diagram shows the number of orthologous genes shared within each possible comparison combination.

The core genome of *B. longum subsp.infantis*, *B. animalis, B. longum T, B. bifidum*, and *B. adolescentis* consisted of 966 genes. *B. longum subsp.infantis* had 867 unique genes with a genome size of 2,832,748 bp and 2,505 CDS. *B. animalis* had 312 unique genes with a genome size of 1,938,709 bp and 1,630 CDS. *B. bifidum* had 394 unique genes with a genome size of 2,186,140 bp and 1,879 CDS. *B. adolescentis* had

285 unique genes with a genome size 2,089,645 bp and 1,719 CDS. *B. longum* (type) had 250 unique genes with a genome size of 2,260,272 bp and 1,879 CDS (Figure 8). The most significant amount of shared genes occurred between the two strains of *B. longum*, with them sharing a further 111 genes uniquely between each other (Figure 8). Genes that were found in all of the above organisms but *B. animalis* included genes primarily for polysaccharide metabolic capabilities as well as a few unique antimicrobial peptide proteins (Table 33).

| | Genes found in all but B. animalis |
|---|---|
| 273 | ABC transporter ATP-binding protein |
| 274 | ABC-type antimicrobial peptide transport system, permease component |
| 275 | ABC-type antimicrobial peptide transport system, permease component |
| 423 | Phosphoenolpyruvate-protein phosphotransferase of PTS system (EC 2.7.3.9) |
| 424 | Phosphocarrier protein of PTS system |
| 445 | DNA-damage-inducible protein F |
| 461 | hypothetical protein |
| 462 | ATP binding protein of ABC transporter |
| 463 | possible permease protein of ABC transporter system |
| 464 | possible permease protein of ABC transporter system |
| 465 | Integral membrane protein |
| 466 | 34 kDa membrane antigen precursor |
| 467 | Putative high-affinity iron permease |
| 545 | Sucrose transporter ScrT, MFS family |
| 586 | Permease of the drug/metabolite transporter (DMT) superfamily |
| 587 | probable Hsp20-family heat shock chaperone |
| 700 | Dehydrogenases with different specificities (related to short-chain alcohol dehydrogenases) |
| 909 | ABC-type antimicrobial peptide transport system, permease component |
| 1081 | thiamine biosynthesis protein |
| 1082 | Thiazole biosynthesis protein ThiG |
| 1083 | Sulfur carrier protein ThiS |
| 1483 | dentin sialophosphoprotein preproprotein |
| 1484 | Permease of the drug/metabolite transporter (DMT) superfamily |

Table 33: Selected genes shared by the organisms referenced in Figure 8 except *B. animalis*.

The core genome of *B. dentium, B. animalis, B. adolescentis, B. angulatum*, and *B. catenulatium* consisted of 1,024 genes. *B. dentium* had 586 unique genes with a genome size of 2,642,081 bp and 2,211 CDS. *B. animalis* had 314 unique genes with a

genome size of 1,938,483 bp and 1,632 CDS. *B. angulatum* had 252 unique genes with a genome size of 2,000,615 bp and 1,659 CDS. *B. adolescentis* had 231 unique genes with a genome size 2,089,645 bp and 1,719 CDS. B. catenulatium had 268 unique genes with a genome size of 2,058,429 bp and 1,779 CDS (Figure 9). *B. dentium, B. adolescentis, B. angulatum*, and *B. catenulatum* shared a further 112 genes unique to them that were not found in *B. amimalis*. These unique genes consisted primarily of sugar metabolism genes, a collagen adhesion gene, as well as a few sulfur and iron utilizing genes (Table 34).

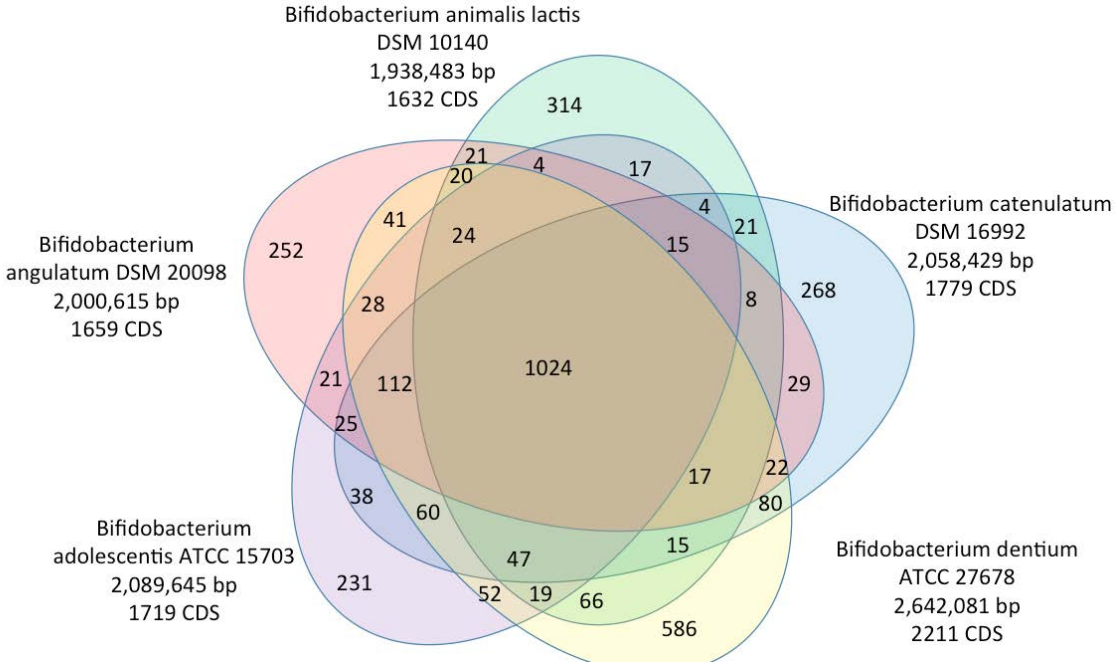


Figure 9: **Genomic similarity at the gene level for *B. dentium*, *B. animalis*, *B. adolescentis T*, *B. angulatum*, and *B. catenulatum*.** Venn-diagram shows the number of orthologous genes shared within each possible comparison combination.

Genes found in B. dentium, B. adolescentis , B. angulatum, and B. catenulatum only

| | |
|---|---|
| 87 | Sodium/dicarboxylate symporter |
| 118 | MSM (multiple sugar metabolism) operon regulatory protein |
| 190 | Dihydropteroate synthase (EC 2.5.1.15) |
| 191 | GTP cyclohydrolase I (EC 3.5.4.16) type 1 |
| 202 | Lactaldehyde reductase (EC 1.1.1.77) |
| 458 | Dihydroorotate dehydrogenase (EC 1.3.3.1) |
| 477 | Phosphoenolpyruvate-protein phosphotransferase of PTS system (EC 2.7.3.9) |
| 478 | Phosphotransferase system, phosphocarrier protein HPr |
| 479 | Ribose operon repressor |
| 492 | Alpha-L-arabinofuranosidase II precursor (EC 3.2.1.55) |
| 493 | Transcriptional regulators |
| 494 | Alpha-L-arabinofuranosidase II precursor (EC 3.2.1.55) |
| 563 | FIG00424242: hypothetical protein |
| 564 | ABC-type antimicrobial peptide transport system, ATPase component |
| 565 | possible permease protein of ABC transporter system |
| 566 | ABC transporter permease protein |
| 567 | Ferrous iron transport permease EfeU |
| 568 | Periplasmic protein p19 involved in high-affinity Fe2+ transport |
| 569 | Putative high-affinity iron permease |
| 583 | Sulfur carrier protein ThiS |
| 584 | Sulfur carrier protein adenylyltransferase ThiF |
| 585 | Thiazole biosynthesis protein ThiG |
| 590 | Sortase A, LPXTG specific |
| 646 | Permease of the drug/metabolite transporter (DMT) superfamily |
| 775 | collagen adhesin precursor |
| 782 | Maltodextrin glucosidase (EC 3.2.1.20) |
| 1373 | Hydroxymethylpyrimidine phosphate synthase ThiC / Thiamin-phosphate pyrophosphorylase (EC 2.5.1.3) |
| 2068 | Sporulation regulatory protein WhiB |
| 2092 | NAD(P)H oxidoreductase YRKL (EC 1.6.99.-) @ Putative NADPH-quinone reductase (modulator of drug activity B) @ Flavodoxin 2 |

Table 34: Selected shared genes between ***B. dentium*, *B. adolescentis, B. angulatum*, and *B. catenulatum.***

## Bifidobacterium Discussion

The Bifidobacterium genus was established in 1924 with Bifidobacterium bifidum being described in 1900 as a member of the bacillus group and subsequently separated as the type in 1924 (Parte 2014). Currently there are 51 validly published members of the genus. Members of the Bifidobacterium genus are gram positive, anaerobic, non-spore-forming, non-motile, lactate and acetate forming, and having high G+C content.

They are normally located within the gastrointestinal tracts of mammals and some bees. They have also been located in human waste and in dairy products. Members of this genus are typically symbiotic with their host and play critical roles in maintaining the health of their host (Choi et al., 2014).

ROSA values for members of the Bifidobacterium genus revealed two distinct clusters. The *B. animalis* cluster contained *B. animalis* and *B. adolescentis*. The *B. bifidum* cluster contained *B. adolescentis*, *B. catenulatum*, *B. angulatum*, *B. dentium*, *B. bifidum*, and *B. longum*. Due to the much higher ROSA values within the cluster C. adolescentis was placed within the *B. bifidum* cluster. Intraspecies comparisons between different strains of *B. animalis* yielded ROSA values between 84.8 and 99.6 (Table 29). These values indicate that all strains are part of the same species. ROSA values between members of the *B. bifidum* cluster ranged from 32.1 to 60.4 (Table 29). Although some of these values fall out of the expected range for intragenus comparisons, the majority for each species fall within the expected values, showcasing their relatedness to each other as members of the same genus. ROSA values between members of the two clusters ranged from 29.3 to 34.8 (Table 29). These values indicate that members in this comparison may have split very recently as the values for intercluster comparisons are just below the threshold for different genera. Based on the ROSA scores the Bifidobacterium genus should be split into two separate genera.

16s rRNA % similarities between members of the genus further shows two distinct clusters previously described above. According to 16s rRNA values only *B. animalis* strains were able to cluster with other *B. animalis* strains with % similarities at

99% to each other and below 94.4 when compared to other members of the genus (Table 30). Members of the *B. bifidum* cluster had 16s rRNA similarity values ranging from 94.7 to 97.3 (Table 30). Intragenus comparisons between the clusters yielded a clear distinction between the two clusters with 16s rRNA values ranging from 93.0 to 94.4, below the suggested genus threshold (Table 30).

AAI values for the Bifidobacterium genus provided further evidence for two separate clusters within the genus. Strains of *B. animalis* have AAI values ranging from 96.0 to 100 (Table 31). Members of the *B. bifidum* cluster had AAI values ranging from 71.6 to 85 while the intraspecies comparison for the *B. bifidum* cluster had an AAI value of 93.3 (Table 31). The AAI value between the two strains of *B. longum* fall below the expected AAI value for members of the same species. For comparisons between strains of the different clusters AAI values ranged from 65.5 to 68 (Table 31). These consistently low values show a distinct genomic difference between the two clusters of the genus, suggesting that they have greater differences than the suggested same genus standard in taxonomy.

%BBH values for the Bifidobacterium provide further evidence for two separate distinct clusters within the Bifidobacterium genus. Strains of *B. animalis* share at least 92.4 % of its genome and at most 99.6% of its genome (Table 32). Members of The *B. bifidum* cluster shared over 71.9% to 84.8% of their genomes to each other within the cluster (Table 32). Between the two clusters BBH values ranged from 60.4% to 74.6% of the genomes shared between the two clusters (Table 32). *B. longum* Subsp infantis and *B. dentium* showed extremely low levels of orthology within their cluster and outside of

it. Percentages of shared genome range from 55.4% to 60.6% and 59.1% to 67.4% (Table 32). One explanation for this vast difference in genome sizes, is that the two organisms have genome sizes significantly larger than other members, which will lower the BBH score.

When the genomes of *B. longum ATCC*, *B. animalis*, *B. bifidum*, *B. adolescentis*, and *B. longum NCC* were compared at the gene level the core genome consisted of 966 genes. Each organism had a significant number of unique coding sequences *B. longum ATCC* contained 867 unique genes, *B. animalis* contained 312 unique genes, *B. bifidum* contained 394 unique genes, *B. adolescentis* contained 285 unique genes, while *B. longum NCC* contained 250 unique genes (Figure 8). A further 111 genes were shared between the two strains of *B. longum* (Figure 8). A total of 92 genes were shared by all organisms except for *B. animalis* (Figure 8). No other significant gene counts were observed, however a pattern can be deciphered from the number of genes shared in smaller combinations. The number of genes shared between organisms and *B. animalis* were all extremely small, often numbering below 20, while when another organism was being compared gene counts were in general much higher, showing that *B. animalis* is the least similar of all the genomes (Figure 8). The genes shared by all organisms except for *B. animalis* encode genes primarily for polysaccharide metabolic capabilities as well as a few unique antimicrobial peptide proteins (Table 33).

When the genomes of *B. dentium*, *B. animalis*, *B. adolescentis T*, *B. angulatum*, and *B. catenulatum* were compared at the gene level the core genome consisted of 1024 genes. Each organism had a significant number of unique coding sequences *B.*

*dentium* contained 586 unique genes, *B. animalis* contained 314 unique genes, *B. adolescentis T* contained 231 unique genes, *B. angulatum* contained 252 unique genes, while *B. catenulatum* contained 268 unique genes (Figure 8). A further 112 genes were shared between the by all organisms except for *B. animalis* (Figure 8). No other significant gene counts were observed, however a pattern can be deciphered from the number of genes shared in smaller combinations. The number of genes shared between organisms and *B. animalis* were all extremely small, often numbering below 20, while when another organism was being compared gene counts were in general much higher, showing that *B. animalis* is the least similar of all the genomes (Figure 8). The genes shared by all organisms except for *B. animalis* encode primarily of sugar metabolism genes, a collagen adhesion gene, as well as a few sulfur and iron utilizing genes (Table 34).

Based on the data it is suggested that the Bifidobacterium genus be split into two separate genera. 16s rRNA values show the splitting of the genus into two separate clusters, one that contains only *B. animalis* strains, and one that contains *B. adolescentis*, *B. catenulatum*, *B. angulatum*, *B. dentium*, *B. bifidum*, and *B. longum* (Table 30). Intracluster values were far lower than what would be expected for members of the same genus, below the 95% that is being suggested for members of the same species (Table 30). Conversely members within each cluster showcase a large amount of 16s rRNA similarity, falling above the 95% threshold (Table 30). AAI values showed the same clustering but was the first to suggest that *B. longum* subsp infantis may be a separate species from the parent organism. AAI values within the first cluster

96.0 to 100 while members of the *B. bifidum* cluster had AAI values ranging from 71.6 to 85 while the intraspecies comparison for the *B. bifidum* cluster has an AAI value of 93.3 (Table 31). AAI values between the clusters ranged from 65.5 to 68, showing a strong separation between the two clusters (Table 31). %BBH values repeated this trend in defining two clusters as well as supporting *B. longum* subsp infantis as its own species. BBH values within the first cluster indicated that 92.4 % of its genome was shared while members of the *B. bifidum* cluster had over 71.9% to 84.8% of their genomes shared with each other within the cluster while the intraspecies comparison for the *B. bifidum* cluster has aBBH value of 66.6% (Table 31).

BBH values between the clusters ranged from 60.4% to 74.6 &, showing a strong separation between the two clusters (Table 31). The other organism that had low levels of orthology was *B. dentium* but this was due to the size of its genome. When gene numbers were compared it was seen that *B. animalis* had the fewest genes shared between the other members of the genus, often having 100 or more genes shared by all other members that *B. animalis* did not contain (Figure 8 and 8). These genes encoded for a number of polysaccharide utilization proteins as well as specific resistance and different metal utilization genes (Table 33 and 20). Based on these values and the ROSA values it is suggested that the Bifidobacterium genus be split into two, with one section containing the strains of *B. animalis* and the other genus containing *B. adolescentis*, *B. catenulatum*, *B. angulatum*, *B. dentium*, *B. bifidum*, and *B. longum*. It is further recommended that based on ROSA, AAI, BBH, 16s, genome size, number of CDS, and number of genes shared, that the two strains of *B. longum* be separated into separate

species. Mattereli and his colleagues found that the *B. longum* and *B. longum* subsp infantis had DDH values below the 70% threshold, however they were hesitant to reclassify them as a new species due to phenotypic similarities (Matterli et al., 2008).

## B. *Campylobacter*:

| | ROSA (sorted) | 192222 | 360112 | 195099 | 360109 | 306254 | 306264 | 306263 | 360106 | 360105 | 360104 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Campylobacter jejuni subsp. jejuni NCTC 11168 | **192222** | | | | | | | | | | |
| Campylobacter jejuni subsp. jejuni HB93-13 | **360112** | 86.0 | | | | | | | | | |
| Campylobacter jejuni RM1221 | **195099** | 85.5 | 82.0 | | | | | | | | |
| Campylobacter jejuni subsp. doyleiI 269.97 | **360109** | 77.7 | 76.2 | 75.4 | | | | | | | |
| Campylobacter coli RM2228 | **306254** | 60.4 | 60.1 | 60.7 | 55.4 | | | | | | |
| Campylobacter upsaliensis RM3195 | **306264** | 42.5 | 42.6 | 41.3 | 42.2 | 39.8 | | | | | |
| Campylobacter lari RM2100 | **306263** | 36.7 | 37.4 | 36.0 | 34.4 | 34.2 | 31.1 | | | | |
| Campylobacter fetus subsp. fetus 82-40 | **360106** | 23.3 | 22.9 | 22.6 | 21.9 | 21.4 | 20.7 | 22.0 | | | |
| Campylobacter curvus 525.92 | **360105** | 21.3 | 21.3 | 20.8 | 20.1 | 19.4 | 19.0 | 20.4 | 26.4 | | |
| Campylobacter concisus 13826 | **360104** | 20.4 | 20.3 | 20.0 | 19.4 | 18.7 | 18.3 | 19.7 | 24.7 | 42.8 | |
| Campylobacter hominis ATCC BAA-381 | **360107** | 19.4 | 19.7 | 19.1 | 18.5 | 18.3 | 17.9 | 19.1 | 23.8 | 22.4 | 21.6 |

Table 35: ROSA values for members of the *Campylobacter* genus. Intraspecies comparisons were done betweene *C. jejuni*. Intragenus comparisons were done between *C. jejuni, C. coli, C. upsaliensis, C. lari, C. fetus, C. curvus, C. concisus,* and *C. hominis.*

ROSA values between strains of *C. jejuni* ranged from 75.4 to 86 (Table 35). Intragenus comparisons had ROSA values from 18.3 to 60.7. Intragenus ROSA values showed two clusters of organisms. The first cluster consisted of *C. jejuni, C. coli, C. upsaliensis*, and *C. lari*, ROSA values ranged from 21.1 to 60.7 (Table 35). The ROSA value for the second cluster consisting of *C. curvis* and *C. concisus* was 42.8 (Table 35). *C. fetus* and *C. hominis* did not cluster with any other organisms and had ROSA values between 18 and 23 (Table 35). The ROSA values indicate that the *Campylobacter* genus should be split into 4 separate genera, with cluster 1 and cluster 2 in their own genera as well as C. curbis and *C. concisus* having their own genera.

| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|---|
| Campylobacter_jejuni_subsp._doylei_ATCC_49349T | 1 | | | | | | | | |
| Campylobacter_jejuni_subsp._jejuni_ATCC_33560T | 2 | 99.8 | | | | | | | |
| Campylobacter_coli_LMG_9860T | 3 | 98.3 | 98.4 | | | | | | |
| Campylobacter_lari_subsp._lari_ATCC_35221 | 4 | 98.2 | 98.2 | 97.1 | | | | | |
| Campylobacter_upsaliensis_strain_CCUG_14913 | 5 | 95.8 | 95.8 | 94.7 | 94.9 | | | | |
| Campylobacter_fetus_subsp._fetus_ATCC_27374T | 6 | 93.8 | 93.8 | 94.9 | 93.7 | 92.5 | | | |
| Campylobacter_curvus_ATCC_35224T | 7 | 92.0 | 92.0 | 92.8 | 91.6 | 91.5 | 94.2 | | |
| Campylobacter_concisus_ATCC_33237T | 8 | 92.6 | 92.6 | 93.3 | 93.1 | 91.7 | 95.0 | 96.8 | |
| Campylobacter_hominis_ATCC_BAA-381T | 9 | 90.2 | 90.2 | 90.2 | 89.9 | 88.8 | 91.5 | 92.7 | 93.0 |

Table 36: 16s rRNA %similarity for members of the *Campylobacter* genus.

Members of the *Campylobacter* genus showed two distinct clusters through 16s rRNA % similarity (Table 36). The first cluster consisted of *C. jejuni, C. coli, C. lari*, and *C. upsaliensis* whom had 16s rRNA % similarities ranging from 99.8 and 94.7 (Table 36). The score cluster consists of *C. concius* and *C. curvus* with a 16s rRNA % similarity of 96.8 (Table 36).

| | **Average Amino Acid Identity (AAIr)** | **306264** | **306254** | **360104** | **360105** | **360106** | **360107** | **195099** | **360109** | **360112** | **192222** | **306263** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Campylobacter upsaliensis RM3195 | **306264** | | 73.3 | 55.8 | 55.6 | 55.5 | 54.7 | 73.4 | 73.3 | 73.5 | 73.9 | 65.6 |
| Campylobacter coli RM2228 | **306254** | 73.5 | | 55.6 | 55.8 | 56.2 | 54.6 | 85.9 | 84.6 | 85.3 | 85.5 | 68.0 |
| Campylobacter concisus 13826 | **360104** | 55.9 | 55.8 | | 73.9 | 59.6 | 57.8 | 56.4 | 56.3 | 56.0 | 56.4 | 55.8 |
| Campylobacter curvus 525.92 | **360105** | 55.7 | 56.0 | 74.1 | | 60.7 | 58.1 | 56.1 | 56.1 | 56.6 | 56.3 | 56.4 |
| Campylobacter fetus subsp. fetus 82-40 | **360106** | 55.7 | 56.0 | 59.4 | 60.6 | | 59.2 | 56.0 | 56.0 | 55.8 | 55.8 | 56.5 |
| Campylobacter hominis ATCC BAA-381 | **360107** | 54.8 | 54.9 | 58.2 | 58.5 | 59.5 | | 54.7 | 54.9 | 54.9 | 54.9 | 55.5 |
| Campylobacter jejuni RM1221 | **195099** | 73.8 | 85.9 | 56.4 | 56.1 | 56.0 | 54.1 | | 95.2 | 97.2 | 97.8 | 68.1 |
| Campylobacter jejuni subsp. doylei 269.97 | **360109** | 73.4 | 84.3 | 56.3 | 56.1 | 56.0 | 54.9 | 94.7 | | 95.5 | 95.7 | 68.1 |
| Campylobacter jejuni subsp. jejuni HB93-13 | **360112** | 73.5 | 85.3 | 56.0 | 56.4 | 55.8 | 54.7 | 97.0 | 95.5 | | 97.3 | 68.2 |
| Campylobacter jejuni subsp. jejuni NCTC 11168 | **192222** | 74.2 | 85.6 | 56.3 | 56.4 | 55.8 | 54.5 | 97.5 | 96.2 | 97.4 | | 68.4 |
| Campylobacter lari RM2100 | **306263** | 65.6 | 67.8 | 56.0 | 56.5 | 56.2 | 55.3 | 68.0 | 68.2 | 68.1 | 68.4 | |

Table 37: AAI values for members of the *Campylobacter* genus.

AAI values for members of the *Campylobacter* genus showed significant variation amongst each other. Strains of the same organism had AAI values ranging from 94.7 to 97.4 (Table 37). From AAI values two clusters emerged, the first consisting of *C. curvus* and *C. concisus* which had AAI values of ~ 74 when compared to each other. The second

cluster consisted of *C. jejuni, C. upsaliensis, C. coli*, and potentially *C. lari*. AAI values within this cluster ranged from 73.2 to 85.5 (Table 37). The values of the partial outlier, *C. lari* were consistently between 65 and 68 (Table 37). AAI values for other organisms were below 60 and were consistently in the mid 50's (Table 37).

| | **Percent Bidirectional Best Hit (% BBH)** | 306264 | 306254 | 360104 | 360105 | 360106 | 360107 | 195099 | 360109 | 360112 | 192222 | 306263 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Campylobacter upsaliensis RM3195 | **306264** | | 72.7 | 54.8 | 58.9 | 67.7 | 62.7 | 77.8 | 76.7 | 80.9 | 81.2 | 76.5 |
| Campylobacter coli RM2228 | **306254** | 74.9 | | 57.6 | 61.0 | 69.8 | 65.6 | 84.8 | 77.3 | 86.2 | 87.6 | 80.2 |
| Campylobacter concisus 13826 | **360104** | 62.8 | 63.0 | | 80.3 | 75.0 | 72.2 | 68.0 | 64.4 | 70.8 | 71.4 | 71.3 |
| Campylobacter curvus 525.92 | **360105** | 63.7 | 63.3 | 75.9 | | 75.3 | 72.3 | 69.7 | 65.3 | 71.1 | 72.8 | 70.5 |
| Campylobacter fetus subsp. fetus 82-40 | **360106** | 66.0 | 66.2 | 64.6 | 68.5 | | 70.7 | 72.6 | 67.8 | 75.1 | 77.3 | 72.4 |
| Campylobacter hominis ATCC BAA-381 | **360107** | 56.9 | 56.3 | 56.4 | 59.7 | 64.4 | | 61.8 | 57.4 | 63.8 | 64.0 | 62.7 |
| Campylobacter jejuni RM1221 | **195099** | 74.7 | 79.7 | 57.7 | 62.4 | 71.4 | 67.5 | | 80.3 | 88.1 | 92.0 | 81.2 |
| Campylobacter jejuni subsp. doylei 269.97 | **360109** | 79.9 | 78.0 | 58.2 | 62.7 | 71.7 | 65.4 | 86.9 | | 88.0 | 90.6 | 79.9 |
| Campylobacter jejuni subsp. jejuni HB93-13 | **360112** | 76.8 | 79.0 | 58.5 | 62.3 | 72.0 | 67.2 | 86.0 | 79.2 | | 92.3 | 83.4 |
| Campylobacter jejuni subsp. jejuni NCTC 11168 | **192222** | 73.7 | 77.3 | 57.2 | 61.3 | 72.2 | 65.9 | 87.2 | 78.4 | 89.1 | | 79.8 |
| Campylobacter lari RM2100 | **306263** | 68.21 | 68.3 | 55.0 | 57.7 | 65.9 | 61.9 | 74.3 | 68.2 | 77.6 | 77.4 | |

Table 38: %BBH values for members of the *Campylobacter* genus.

Orthology for members of the *Campylobacter* genus showed a large amount of variation. Members of the same species had on average over 90% orthology. A distinct clustering was formed between of *C. jejuni, C. upsaliensis, C. coli*, and potentially *C. lari* with the species sharing between 70% and 90% of their genomes (Table 38). A second grouping was found between *C. concisus* and *C. curvus* who shared between 76% and 80% of their genome (Table 38). Other members of the genus had %BBH lower than 60% when compared to other members of the genus (Table 38).

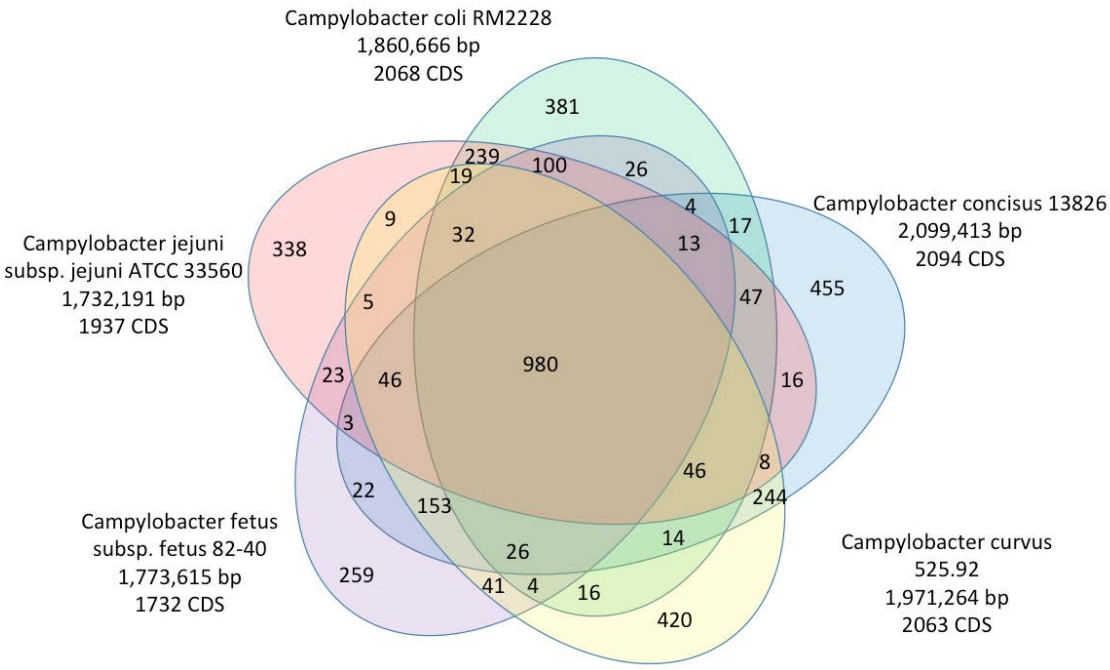


Figure 10: **Genomic similarity at the gene level for *C. fetus, C. jejuni, C. coli, C. conisus,* and *C. curvus*.** Venn-diagram shows the number of orthologous genes shared within each possible comparison combination.

The core genome of *C. fetus, C. jejuni, C. coli, C. conisus,* and *C. curvus* consisted of 980 genes. *C. fetus* had 259 unique genes with a genome size of 1,773,615 bp and 1732 CDS. *C. jejuni* had 338 unique genes with a genome size of 1,732,191 bp and 1937 CDS. *C. coli* had 381 unique genes with a genome size of 1,860,666 bp and 2068 CDS. *C. concisus* had 455 unique genes with a genome size 2,099,413 bp and 2094 CDS. *C. curvus* had 420 unique genes with a genome size of 1,971,264 bp and 2063 CDS (Figure 10). *C. fetus, C. jejuni,* and *C. coli* shared an additional 100 genes (Figure 10). *C. jejuni* and *C. coli* shared an additional 239 genes(Figure 10). *C. concisus* and *C. curvis* shared an additional 244 genes (Figure 10). Some of the shared genes between *C. curvus* and *C. concisus* included iron utilization, carbon metabolism, sulfur using proteins, CRISPR

proteins, secreted proteins, metal resistance proteins, and a tartate pathway (Table 39). Unique genes shared between *C. jeuni* and *C. coli* include an erythrocyte associated operon, multidrug resistance, export proteins, and enterocholine uptake proteins (Table 40).

C. concisus and C. curvus only

44Carbon monoxide dehydrogenase CooS subunit (EC 1.2.99.2)
118Flavocytochrome c flavin subunit
193Putative two-component response regulator
194Flavocytochrome c flavin subunit
205Ferrichrome-iron receptor
206Methyltransferase
207Vitamin B12 ABC transporter, permease component BtuC
208Iron ABC transporter, ATP-binding protein
238Putative arylsulfate sulfotransferase (EC 2.8.2.22)
251UPF0141 membrane protein YijP possibly required for phosphoethanolamine modification of lipopolysaccharide
268Benzoyl-CoA reductase subunit BadG (EC 1.3.99.15)
338TonB-dependent receptor; Outer membrane receptor for ferrienterochelin and colicins
339Uncharacterized iron-regulated membrane protein; Iron-uptake factor PiuB
367diguanylate cyclase/phosphodiesterase (GGDEF & EAL domains) with PAS/PAC sensor(s)
501Z4855 protein
502Acyl-coenzyme A synthetases/AMP-(fatty) acid ligases; putative surfactin synthetase homolog
503Acyltransferase (EC 2.3.1.-) # surfactin synthetase cluster
505Putative transmembrane protein
506membrane protein, inferred for ABFAE pathway
542sulfate/thiosulfate import ATP-binding protein CysA (sulfate-transporting ATPase) [EC:3.6.3.25]
710Methyl-accepting chemotaxis protein
724Iron(III) ABC transporter, ATP-binding protein
725Vitamin B12 ABC transporter, permease component BtuC
726Esterase/lipase
728Iron(III) ABC transporter, solute-binding protein
823Lead, cadmium, zinc and mercury transporting ATPase (EC 3.6.3.3)
860Ferric uptake regulation protein FUR
916Outer membrane protein of RND family multidrug efflux pump / Multidrug efflux system CmeDEF
917Putative acetate efflux pump, MadN
930CRISPR-associated protein, TM1800 family
931CRISPR-associated protein, TM1801 family
932CRISPR-associated protein, TM1802 family
933hypothetical protein
936Phage integrase
1388TETRATRICOPEPTIDE REPEAT FAMILY PROTEIN
1391conserved hypothetical secreted protein
1392TETRATRICOPEPTIDE REPEAT FAMILY PROTEIN
1495polysulfide reductase, subunit C
1496polysulfide reductase, subunit B
1497polysulfide reductase, subunit A
1630Transcriptional regulator
1631UPF0272 protein CA_C0774
1632ATP-UTILIZING ENZYME OF THE PP-LOOP SUPERFAMILY
1633NCAIR mutase (PurE)-related protein
1634L(+)-tartrate dehydratase alpha subunit (EC 4.2.1.32)
1635L(+)-tartrate dehydratase beta subunit (EC 4.2.1.32)

Table 39: Selected shared genes of *C. concisus* and *C. curvus.*

## C. jejuni and C. coli only

| | |
|---|---|
| 17 | IcmF-related protein |
| 21 | IncF plasmid conjugative transfer protein TraG |
| 27 | FIG00471674: hypothetical protein |
| 29 | hypothetical protein |
| 30 | hypothetical protein |
| 32 | conserved hypothetical protein |
| 33 | conserved hypothetical protein |
| 36 | transcriptional regulator, Cro/CI family |
| 37 | FIG00470782: hypothetical protein |
| 39 | hypothetical protein |
| 42 | FIG00469806: hypothetical protein |
| 43 | hypothetical protein |
| 44 | hypothetical protein |
| 45 | FIG00469781: hypothetical protein |
| 47 | Erythrocyte membrane-associated antigen |
| 48 | Integrase-recombinase protein XERCD family |
| 172 | Putative acyl carrier protein |
| 173 | 3-oxoacyl-[acyl-carrier-protein] synthase, KASIII (EC 2.3.1.41) |
| 174 | HAD-superfamily phosphatase, subfamily IIIC |
| 221 | Multidrug-efflux transporter, major facilitator superfamily (MFS) (TC 2.A.1) |
| 371 | Multiple antibiotic resistance protein marC |
| 504 | Highly acidic protein |
| 505 | L-Proline/Glycine betaine transporter ProP |
| 759 | CAAX amino terminal protease family protein |
| 760 | Histone acetyltransferase HPA2 and related acetyltransferases |
| 1024 | Nicotinamidase (EC 3.5.1.19) |
| 1042 | Chromosome (plasmid) partitioning protein ParB / Stage 0 sporulation protein J |
| 1155 | Arsenate reductase (EC 1.20.4.1) |
| 1264 | Nonheme iron-containing ferritin |
| 1285 | DNA/RNA non-specific endonuclease |
| 1311 | FIG00710232: hypothetical protein |
| 1312 | FIG00469622: hypothetical protein |
| 1313 | FIG00469658: hypothetical protein |
| 1314 | GTP-binding protein |
| 1647 | Putative bacterial haemoglobin |
| 1648 | Predicted D-lactate dehydrogenase, Fe-S protein, FAD/FMN-containing |
| 1697 | Motility accessory factor |
| 1710 | Cinnamyl alcohol dehydrogenase/reductase (EC 1.1.1.195) @ Alcohol dehydrogenase (EC 1.1.1.1) |
| 1711 | Homolog of BLC protein |
| 1715 | Allophanate hydrolase 2 subunit 2 (EC 3.5.1.54) |
| 1717 | Allophanate hydrolase 2 subunit 1 (EC 3.5.1.54) |
| 1718 | Lactam utilization protein LamB |
| 1830 | Capsular polysaccharide export system inner membrane protein KpsE |
| 1831 | Capsular polysaccharide export system periplasmic protein KpsD |
| 1856 | Capsular polysaccharide export system protein KpsC |
| 1857 | Capsular polysaccharide export system protein KpsS |
| 1858 | Probable integral membrane protein Cj1412c |
| 1859 | Cytochrome P450 family protein |
| 1860 | membrane protein |
| 1871 | hydrogenase, (NiFe)/(NiFeSe) small subunit family |
| 1919 | Putative integral membrane protein |
| 1920 | Iron compound ABC uptake transporter substrate-binding protein |
| 1921 | Enterochelin uptake ATP-binding protein |
| 1922 | Enterochelin uptake permease CeuC |
| 1923 | Enterochelin uptake permease CeuB |

Table 40: Selected shared genes of *C. jejuni* and *C. coli*.

Campylobacter Discussion

The *Campylobacter* genus was established in 1963 with *Campylobacter fetus* being described in 1919 then designated the type species in 1963. There are currently approximately 33 members in the genus (Parte 2014). Members of the genus have been isolated from a diverse range of vertebrates, mammals, reptiles and birds. As such members that are found in reptiles are adapted to a large range of temperatures (Gilbert et al., 2015). Strains from the genus are gram negative, microaerophilic, oxidase positive, non-fermentive, have flagella and are spiral shaped. Many members are pathogenic in their hosts (Gilbert et al., 2015).

ROSA values for members of the genus *Campylobacter* showed one larger cluster, one cluster with two members, and two that do not cluster at all. The larger cluster, cluster 1, consists of *C. jejuni, C. coli, C. upsaliensis*, and *C. lari.* The smaller cluster, cluster 2, consists of C. curbus, *C. concisus*, and the two that do not cluster are *C. fetus*, and *C. hominis* (Table 35). Intraspecies comparisons for strains of *C. jejuni* had ROSA values ranging from 75.4 to 86.0, within the expected range for intraspecies (Table 35). Intragenus comparisons within the cluster had scores ranging from 31.1 to 60.4; these values straddled the intragenus comparison range but only with one comparison to the non-type (Table 35). The second cluster had an intragenus ROSA value of 42.8. The two clusters when compared had intercluster values ranging from 18.3 to 22.9 (Table 35). These values suggest that they are in different genera of the same family. The two organisms that did not cluster efficiently, *C. hominis* and *C. fetus* had ROSA values ranging from 18.3 to 22.4 and 19 to 23.3 respectively (Table 35). These values

suggest that the two outlying organisms do not belong within the *Campylobacter* genus and that the genus should be split into the two clusters.

16s rRNA similarity values further supports this clustering. Members of Cluster 1 had intraspecies 16s rRNA % similarity of 99.8% (Table 36). Intragenus comparisons for cluster 1 had 16s rRNA % similarities ranging from 94.7 %to 98.3%, these values are either borderline or above the genus level threshold suggested for 16s rRNA % similarity (Table 36). Cluster 2 had 16s rRNA % similarity of 96.8%, above the level for genus separation (Table 36). Intercluster comparisons had 16s rRNA % similarities ranging form 91.5% to 93.3%, well below the genus threshold for 16s rRNA (Table 36). These intercluster values suggest that the two clusters are seperate genera from each other. *C. hominis* and *C. fetus* had 16s rRNA % similarities ranging from 88.8% to 93% and 92.5% to 94.9% respectively; these values indicate that the organisms belong in their own genus each.

AAI values support the clustering of the genus mentioned above. Members of cluster 1 had intraspecies AAI ranging from 94.7 to 97.8, values that are borderline below or well above the species threshold for AAI (Table 37). Intragenus comparisons for members of cluster 1 had AAI values of 68.1 to 85.3 (Table 37). AAI values for intergenus comparisons within cluster 2 ranged from 73.9 to 74.1 (Table 37). Intercluster comparisons between members of both clusters range from 55.8 to 60.8, showing a distinct separation between the two clusters (Table 37). AAI values of the two outliers *C. hominis* and *C. fetus* ranged from 55.7 to 60.7 and 54.1 to 59.5 respectively,

values that are far off the clustering of other organisms. It is important to note that these outlier values are in the range as the intercluster comparison values.

%BBh values provided further evidence for the clustering seen in the ROSA values. Intraspecies comparisons in cluster 1 shared 78.4% to 92.3% of their genomes (Table 38). Intragenus comparisons within cluster 1 ranged from 72.7% to 87.65 of the genome conserved (Table 38) For cluster 2 organisms had 75.9% to 80.3% of their genomes conserved (Table 38). Intercluster comparisons had BBH values ranging from 55% to 68.2% (Table 38). These values indicate that the two clusters share a relationship greater than that of a genus. The outliers *C. hominis* and *C. fetus* have %BBH values of 56.4% to 64% and 64.6% to75.3% between other members of the genus (Table 38). These values are similar to those of the intercluster comparisons.

When the genomes of *C. fetus, C. jejuni, C. coli, C. conisus, and C. curvus* were compared at the gene level, the core genome consisted of 980 genes. Each organism had a significant number of unique coding sequences. *C. fetus* contained 259 unique genes, *C. jejuni* contained 338 unique genes, *C. coli* contained 381 unique genes*, C. conisus* contained 455 unique genes, while *C. curvus* contained 420 unique genes (Figure 10). *C. coli* and *C. jejuni* shared a further 239 genes indicating a significant relationship between them (Figure 10). These genes encode for an erythrocyte-associated operon, multidrug resistance, export proteins, and enterocholine uptake proteins (Table 40). *C. fetus, C. coli* and *C. jejuni* shared an additional 100 genes while *C. coli, C. fetus, C. conisus*, and *C. curvis* shared 153 genes and *C. curvus* and *C. concisus* shared an additional 244 genes between them (Figure 10). The genes shared by *C. curvus* and *C.*

*concisus* encoded for iron utilization carbon metabolism, sulfur using proteins, CRISPR proteins, secreted proteins, metal resistance proteins, and a tartate pathway (Table 39).

Based on the data presented it is recommended that the *Campylobacter* genus be split into two different genera, with a possible 3rd and 4th using the outliers. The first genus would consist of organisms in cluster 1 and the second would consist of the organisms in cluster 2. The two outliers based on their scores should be their own separate 3rd and 4th. 16s rRNA separate the organisms into two clusters as previously mentioned with the two outliers mentioned above. AAI values and %BBH values support this classification for the organisms with each one forming identical groupings within the cluster. Gene count comparisons showcased this relationship as well. *C. fetus* had the lowest level of orthology with the other organisms and shared equal numbers of genes with each subcluster (Figure 10).

*C. jejuni* and *C. coli,* both members of cluster 1, shared a significant amount of additional genes between them that encoded for a number of metabolic processes unique to them (Figure 10). *C. curvus* and *C. concisus* shared an additional 244 genes unique to them that encoded for a variety of unique metal related genes as well as CRISPR proteins and metabolic pathways (Figure 10). These differences in genes provide further evidence that the two clusters belong in separate genera. ROSA values predict the creation of 2 larger clusters and 2 outliers whom all belong to different genera within the same family (Table 35). The 16s rRNA, BBH, AAI, and gene comparisons support the separation of the two main clusters and the two outliers. Based on the

information it is recommended that cluster 1, cluster 2, *C. hominis*, and *C. fetus* be separated into separate genera within the same family.

## C. *Lactobacillus*

| | ROSA (sorted) | 321956 | 390333 | 405566 | 272621 | 324831 | 257314 | 557433 | 334390 | 321967 | 543734 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Lactobacillus delbrueckii subsp. bulgaricus ATCC BAA-365 | **321956** | | | | | | | | | | |
| Lactobacillus delbrueckii subsp. bulgaricus ATCC 11842 | **390333** | 86.9 | | | | | | | | | |
| Lactobacillus helveticus DPC 4571 | **405566** | 29.5 | 28.5 | | | | | | | | |
| Lactobacillus acidophilus NCFM | **272621** | 28.2 | 27.7 | 50.2 | | | | | | | |
| Lactobacillus gasseri ATCC 33323 | **324831** | 25.8 | 25.3 | 31.5 | 34.3 | | | | | | |
| Lactobacillus johnsonii NCC 533 | **257314** | 25.0 | 24.5 | 30.9 | 35.3 | 63.9 | | | | | |
| Lactobacillus reuteri JCM 1112 | **557433** | 14.8 | 14.5 | 15.4 | 15.8 | 15.8 | 15.5 | | | | |
| Lactobacillus fermentum IFO 3956 | **334390** | 14.8 | 14.8 | 14.4 | 14.2 | 14.2 | 14.0 | 30.3 | | | |
| Lactobacillus casei ATCC 334 | **321967** | 14.5 | 14.2 | 13.9 | 14.1 | 14.3 | 14.0 | 14.5 | 14.8 | | |
| Lactobacillus casei BL23 | **543734** | 14.0 | 13.7 | 13.4 | 13.9 | 14.4 | 13.9 | 14.3 | 14.4 | 84.4 | |
| Lactobacillus brevis ATCC 367 | **387344** | 13.5 | 13.1 | 13.1 | 13.3 | 13.8 | 13.2 | 19.4 | 19.1 | 16.4 | 16.0 |

Table 41: ROSA values for members of the *Lactobacillus* genus. Intraspecies comparisons were done between strains of *L. delbrueckii* and *L. casei*. Intragenus comparisons were done between *L. delbrueckii, L. helveticus, L. acidophilus, L. gasseri, L. johnsonii, L. fermentum, L. casei*, and *L. brevis*.

ROSA values for intraspecies comparisons between strains of *L. delbrueckii* and *L. casei* had values of 86.9 and 84.4 respectively (Table 41). Intragenus ROSA values L delbrueckii, *L. helveticus*, *L. acidophilus*, *L. gasseri*, *L. johnsonii*, *L. casei*, and *L. brevis* ranged from 13.1 to 63.9 (Table 41). Two potential small clusters formed. *L. helveticus* and *L. acidophilus* had a ROSA value of 50.2 (Table 41). *L. acidophilus*, *L. gasseri*, and *L. johnsonii* formed a small cluster with ROSA values ranging from 34.3 and 63.9 (Table 41). Other members of *Lactobacillus* showed no distinct clustering with ROSA values ranging from 13.1 to 30.3 (Table 41). ROSA values suggest that *L. delbrueckii, L. reuteri, L. fermentum*, and *L. brevis* are all split from the genus to form single member genera. *L.*

*gasseri* and *L. johnsonii* along with *L. helveticus* and *L. acidophilus* should form two species genera each.

| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| Lactobacillus_delbrueckii_ATCC_11842T | 1 | | | | | | | | | |
| Lactobacillus_helveticus_DSM_20075T | 2 | 93.3 | | | | | | | | |
| Lactobacillus_acidophilus_CIP_76.13T | 3 | 93.8 | 98.4 | | | | | | | |
| Lactobacillus_johnsonii_ATCC_33200T | 4 | 91.9 | 93.1 | 92.9 | | | | | | |
| Lactobacillus_gasseri_ATCC_33323T | 5 | 91.9 | 93.3 | 92.9 | 99.5 | | | | | |
| Lactobacillus_reuteri_JCM_1112T | 6 | 88.5 | 88.8 | 89.0 | 90.3 | 90.1 | | | | |
| Lactobacillus_fermentum_CECT_562T | 7 | 88.4 | 88.0 | 88.5 | 90.0 | 90.1 | 93.9 | | | |
| Lactobacillus_fermentum_NBRC_3956 | 8 | 88.6 | 88.0 | 88.6 | 90.1 | 90.1 | 94.0 | 99.7 | | |
| Lactobacillus_casei_BL23T | 9 | 88.1 | 87.9 | 87.8 | 89.3 | 89.0 | 91.5 | 90.1 | 90.2 | |
| Lactobacillus_brevis_ATCC_14869T | 10 | 88.5 | 87.6 | 87.8 | 90.3 | 90.0 | 91.5 | 91.3 | 91.5 | 91.7 |

Table 42: 16s rRNA %similarity values for members of the *Lactobacillus* genus.

16s rRNA values for the *Lactobacillus* genus showed a large amount of variation with a few clusters forming from the values (Table 42). *L. helveticus* and *L. acidophilus* had 98.4% 16s rRNA similarity, *L. gasseri* and *L. johnsonii* had 99.5% 16s rRNA similarity, the two strains of *L. fermentum* had 99.7% 16s rRNA similarity (Table 42). Two large clusterings appeared in the table. The first consisted of *L. delbrueckii*, *L. helveticus*, *L. acidophilus*, *L. gasseri*, and *L. johnsonii* had 16s rRNA % similarity values ranging from 91.9 to 99.5 (Table 42). The second consisted of *L. fermentum*, *L. reuteri*, *L. casei*, and *L. brevis* had 16s rRNA % similarity values ranging from 90.1 to 94.0 (Table 42).

| | Average Amino Acid Identity (AAIr) | 272621 | 387344 | 321967 | 543734 | 390333 | 321956 | 334390 | 324831 | 405566 | 257314 | 557433 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Lactobacillus acidophilus NCFM | **272621** | | 50.1 | 51.0 | 51.1 | 65.2 | 65.2 | 51.8 | 68.1 | 84.3 | 69.6 | 53.3 |
| Lactobacillus brevis ATCC 367 | **387344** | 50.1 | | 54.4 | 54.5 | 50.3 | 50.2 | 57.1 | 50.3 | 51.1 | 50.4 | 56.8 |
| Lactobacillus casei ATCC 334 | **321967** | 51.2 | 54.5 | | 98.6 | 51.5 | 51.4 | 52.6 | 51.2 | 52.2 | 51.0 | 52.5 |
| Lactobacillus casei BL23 | **543734** | 51.1 | 54.4 | 98.7 | | 51.3 | 51.1 | 52.1 | 51.3 | 52.3 | 50.8 | 51.9 |
| Lactobacillus delbrueckii subsp. bulgaricus ATCC 11842 | **390333** | 65.0 | 50.2 | 51.6 | 51.4 | | 99.2 | 52.0 | 62.4 | 66.4 | 62.6 | 51.5 |
| Lactobacillus delbrueckii subsp. bulgaricus ATCC BAA-365 | **321956** | 65.1 | 50.1 | 51.4 | 51.1 | 99.2 | | 51.0 | 62.4 | 66.8 | 62.6 | 51.3 |
| Lactobacillus fermentum IFO 3956 | **334390** | 51.6 | 57.2 | 52.7 | 52.2 | 51.7 | 51.1 | | 51.5 | 51.9 | 51.3 | 65.6 |
| Lactobacillus gasseri ATCC 33323 | **324831** | 68.2 | 50.4 | 51.3 | 51.4 | 62.8 | 62.7 | 51.7 | | 69.3 | 88.2 | 52.8 |
| Lactobacillus helveticus DPC 4571 | **405566** | 83.7 | 51.0 | 52.0 | 52.0 | 66.3 | 66.7 | 51.9 | 68.7 | | 69.1 | 53.5 |
| Lactobacillus johnsonii NCC 533 | **257314** | 69.8 | 50.1 | 51.2 | 51.4 | 62.9 | 62.8 | 51.5 | 88.8 | 69.6 | | 52.7 |
| Lactobacillus reuteri JCM 1112 | **557433** | 53.5 | 56.7 | 52.3 | 52.0 | 51.4 | 51.3 | 65.6 | 52.6 | 53.4 | 52.8 | |

Table 43: AAI values for members of the *Lactobacillus* genus.

Members of the *Lactobacillus* genus had a large amount of variation in their AAI values. Intraspecies comparisons between strains of *L. casei* and *L. delbrueckii* had AAI values of 98.7 and 99.2 respectively (Table 43). Only two organisms had significant AAI values, *L. gasseri* and *L. johnsonii*, with a value of 88.5 (Table 43). Intragenus AAI values formed no distinct clustering between different organisms of the genus, AAI values between different members of the genus ranged from 50.1 to 88.2 (Table 43).

| | Percent Bidirectional Best Hit (% BBH) | 272621 | 387344 | 321967 | 543734 | 390333 | 321956 | 334390 | 324831 | 405566 | 257314 | 557433 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Lactobacillus acidophilus NCFM | **272621** | | 49.8 | 45.7 | 43.5 | 68.0 | 73.1 | 52.6 | 76.2 | 68.4 | 72.7 | 55.6 |
| Lactobacillus brevis ATCC 367 | **387344** | 56.4 | | 50.0 | 47.2 | 57.3 | 62.0 | 61.9 | 59.2 | 51.9 | 55.2 | 63.7 |
| Lactobacillus casei ATCC 334 | **321967** | 62.3 | 60.8 | | 84.5 | 63.6 | 69.3 | 61.6 | 64.9 | 57.7 | 62.0 | 61.3 |
| Lactobacillus casei BL23 | **543734** | 62.6 | 60.9 | 89.0 | | 63.5 | 69.3 | 62.7 | 66.4 | 57.0 | 63.6 | 62.6 |
| Lactobacillus delbrueckii subsp. bulgaricus ATCC 11842 | **390333** | 62.4 | 46.6 | 43.0 | 40.4 | | 92.4 | 52.9 | 64.3 | 60.8 | 59.5 | 52.6 |
| Lactobacillus delbrueckii subsp. bulgaricus ATCC BAA-365 | **321956** | 59.8 | 44.9 | 40.8 | 38.3 | 84.2 | | 50.8 | 61.5 | 58.3 | 56.9 | 50.5 |
| Lactobacillus fermentum IFO 3956 | **334390** | 53.3 | 54.8 | 45.0 | 43.3 | 56.9 | 62.4 | | 55.8 | 52.4 | 52.8 | 70.5 |
| Lactobacillus gasseri ATCC 33323 | **324831** | 71.4 | 49.2 | 43.9 | 42.6 | 65.1 | 70.5 | 51.2 | | 61.7 | 79.0 | 54.9 |
| Lactobacillus helveticus DPC 4571 | **405566** | 73.9 | 48.6 | 44.2 | 41.8 | 68.9 | 74.2 | 54.7 | 70.7 | | 66.4 | 55.8 |
| Lactobacillus johnsonii NCC 533 | **257314** | 72.7 | 49.7 | 45.2 | 43.0 | 65.0 | 70.2 | 52.8 | 84.1 | 62.2 | | 55.9 |
| Lactobacillus reuteri JCM 1112 | **557433** | 55.5 | 56.4 | 44.4 | 43.0 | 56.9 | 61.7 | 70.3 | 58.8 | 51.9 | 55.9 | |

Table 44: %BBH values for members of the *Lactobacillus* genus.

%BBH values for members of the *Lactobacillus* genus showed a large amount of variation. Strains of *L. casei* and *L. delbrueckii* shared 87 and 90% of their genome respectively (Table 44). Intragenus comparisons between members of *Lactobacillus* showed only one possible cluster consisting of *L. gasseri*, *L. helveticus*, and *L. johnsonii* who shared on average 70% of their genome and had similar levels of orthology when compared to other organisms (Table 44).

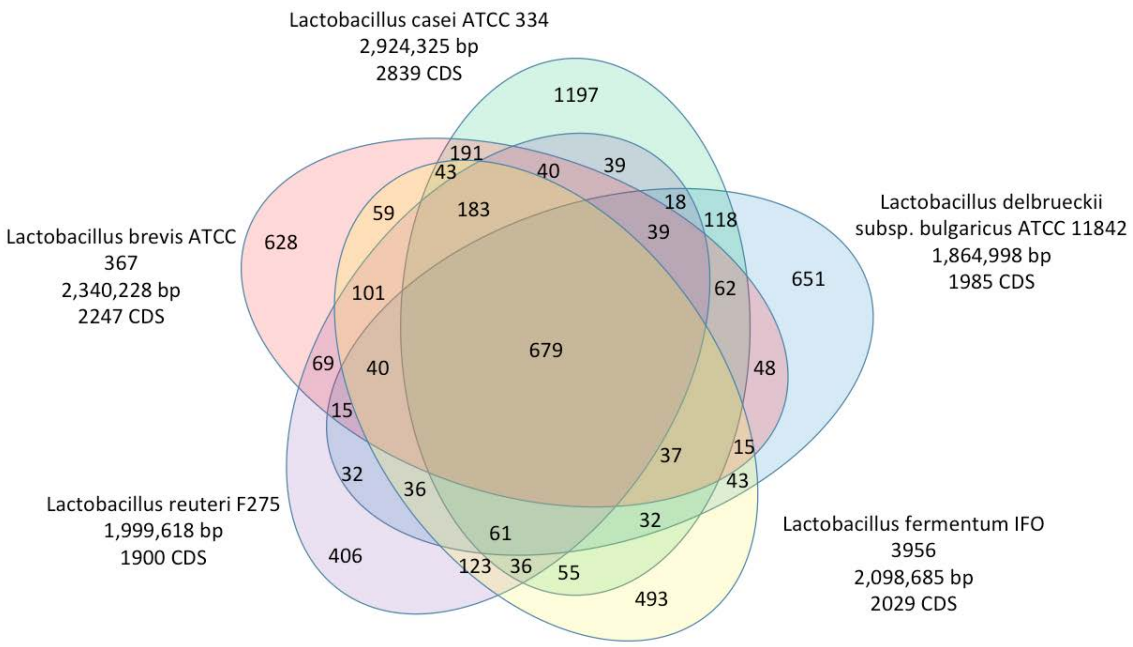


Figure 11: **Genomic similarity at the gene level for *L. reuteri*, *L. brevis*, *L. casei*, *L. delbrueckii*, and *L. fermentum*.** Venn-diagram shows the number of orthologous genes shared within each possible comparison combination.

The core genome of for *L. reuteri*, *L. brevis*, *L. casei*, *L. delbrueckii*, and *L. fermentum* consisted of 679 genes. *L. reuteri* had 406 unique genes with a genome size of 1,999,618 bp and 1900 CDS. *L. brevis* had 628 unique genes with a genome size of 2,340,228 bp and 2247 CDS. *L. casei* had 1197 unique genes with a genome size of 2,924,325 bp and 2839 CDS. *L. delbrueckii* had 651 unique genes with a genome size 1,864,998 bp and 1985 CDS. *L. fermentum* had 493 unique genes with a genome size of 2,098,685 bp and 2029 CDS (Figure 11). *L. brevis* and *L. casei* shared a further 191 genes between them, *L. casei* further shared 118 genes with *L. delbrueckii* (Figure 11). *L. fermentum* and *L. reuteri* shared a further 123 genes between them (Figure 11). *L. reuteri*, *L. brevis*, *L. casei*, and *L. fermentum* shared a further 183 genes between them

that *L. delbrueckii* did not contain (Figure 11). These shared genes were primarily scattered transcriptional regulatory proteins and hypothetical proteins.

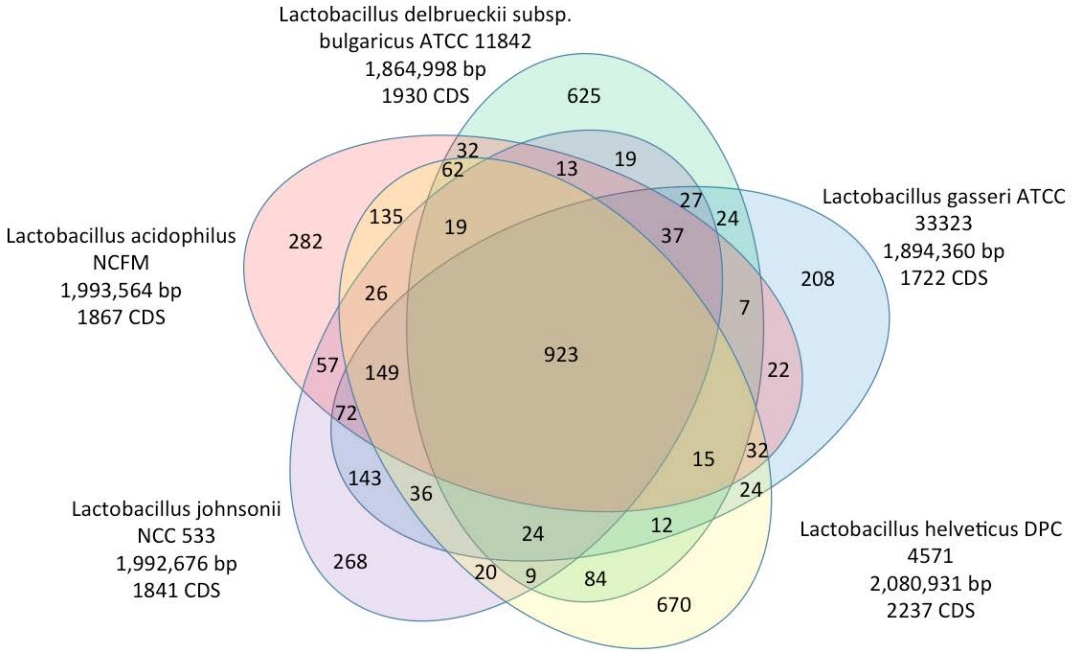


Figure 12: **Genomic similarity at the gene level for *L. johnsonii*, *L. acidophilus*, *L. delbrueckii*, *L. gasseri*, and *L. helveticus*.** Venn-diagram shows the number of orthologous genes shared within each possible comparison combination.

The core genome of for *L. johnsonii*, *L. acidophilus*, *L. delbrueckii*, *L. gasseri*, and *L. helveticus* consisted of 923 genes. *L. johnsonii* had 268 unique genes with a genome size of 1,992,676 bp and 1841 CDS. *L. acidophilus* had 285 unique genes with a genome size of 1,993,564 bp and 1867 CDS. *L. delbrueckii* had 625 unique genes with a genome size of 1,864,998 bp and 1930 CDS. *L. gasseri* had 208 unique genes with a genome size 1,894,360 bp and 1722 CDS. *L. helveticus* had 670 unique genes with a genome size of 2,080,931 bp and 2237 CDS (Figure 12). *L. johnsonii*, *L. acidophilus*, *L. gasseri*, and *L. helveticus* shared 149 genes that *L. delbrueckii* did not (Figure 12). *L. johnsonii* and *L. gasseri* shared a further 143 genes (Figure 12). These genes primarily deal with sugar

uptake and transport as well as a few resistance proteins (Table 45). *L. helveticus* and *L. acidophilus* shared a further 135 genes uniquely between them (Figure 12). These genes include sugar transport, citrate utilization, resistance proteins, and an operon of Imidazolonepropionase related amidohydrolase that may indicate an operon for metabolism of compounds containing C-N bonds (Table 46).

| | L. gasseri and L. johnsonii only. |
|---|---|
| 487 | PTS system, mannose-specific IIB component (EC 2.7.1.69) |
| 488 | phosphoenolpyruvate-dependent sugar phosphotransferase system EIIC, probable sorbose specific |
| 489 | phosphoenolpyruvate-dependent sugar phosphotransferase system EIID, probable fructose specific |
| 490 | PTS system, mannose-specific IIA component (EC 2.7.1.69) |
| 491 | Oligo-1,6-glucosidase (EC 3.2.1.10) |
| | |
| 556 | Antirepressor [SA bacteriophages 11, Mu50B] |
| 557 | hypothetical protein |
| 563 | Phage related protein |
| 564 | DNA modification methylase |
| 566 | hypothetical protein |
| 569 | ORF009 |
| 570 | Phage portal protein, SPP1 Gp6-like |
| 571 | Phage Mu protein F like protein |
| 573 | Lactobacillus delbrueckii phage mv4 main capsid protein Gp34 homolog lin2390 |
| 806 | contains gram positive anchor domain |
| | |
| 1402 | Lantibiotic transport permease protein |
| 1431 | ABC-type multidrug transport system, ATPase and permease components |
| 1616 | Anthranilate synthase, amidotransferase component like (EC 4.1.3.27) |
| | |
| 1657 | putative membrane protein |
| 1658 | hypothetical protein |
| 1660 | phosphoenolpyruvate-dependent sugar phosphotransferase system EIIC, probable galactitol specific |
| 1661 | Integral membrane protein |
| 1665 | Cytochrome d ubiquinol oxidase subunit I (EC 1.10.3.-) |
| 1666 | Cytochrome d ubiquinol oxidase subunit II (EC 1.10.3.-) |
| 1668 | Transport ATP-binding protein cydC |
| 1669 | Putative prenyltransferase, contains 1,4-dihydroxy-2-naphthoate octaprenyltransferase domain |

Table 45: Selected genes unique to *L. gasseri* and *L. johnsonii*.

| | L. helveticus and L. acidophilus only. |
|---|---|
| 61 | Surface layer protein |
| 72 | penicillin-binding protein |
| 192 | S-layer protein precursor |
| 218 | putative fibronectin domain |
| 240 | Cation transport ATPase |
| 440 | PTS system sucrose-specific IIABC component |
| 516 | Acetate kinase (EC 2.7.2.1) |
| 945 | 2,3,4,5-tetrahydropyridine-2,6-dicarboxylate N-acetyltransferase (EC 2.3.1.89) |
| 946 | N-acetyl-L,L-diaminopimelate deacetylase (EC 3.5.1.47) |
| 947 | Dihydrodipicolinate synthase (EC 4.2.1.52) |
| 948 | Dihydrodipicolinate reductase (EC 1.3.1.26) |
| 949 | N-acetyl-L,L-diaminopimelate aminotransferase (EC 2.6.1.-) |
| 954 | polyferredoxin |
| 1014 | putative mucus binding precursor |
| | |
| 1029 | [Citrate [pro-3S]-lyase] ligase (EC 6.2.1.22) |
| 1030 | Citrate lyase gamma chain, acyl carrier protein (EC 4.1.3.6) |
| 1031 | Citrate lyase beta chain (EC 4.1.3.6) |
| 1032 | Citrate lyase alpha chain (EC 4.1.3.6) |
| | |
| 1145 | Potassium uptake protein ktrB |
| 1148 | Flavodoxin |
| 1151 | penicillin-binding protein |
| | |
| 1583 | Imidazolonepropionase related amidohydrolase |
| 1584 | Imidazolonepropionase related amidohydrolase |
| 1585 | Oligopeptide ABC transporter, periplasmic oligopeptide-binding protein oppA (TC 3.A.1.5.1) |
| | |
| 1801 | Stress-responsive transcriptional regulator PspC |
| 1815 | thermostable pullulanase( EC:3.2.1.41 ) |

Table 46: Selected genes unique to *L. helveticus* and *L. acidophilus*.

Lactobacillus Discussion

The *Lactobacillus* genus was established in 1901 with *Lactobacillus delbrueckii* being described in 1896 as a member of the *Bacillus* genus and subsequently separated as the type in 1901. Currently there are approximately 214 validly published species (Parte 2014). Members of this genus are known as the main representative of the lactic acid bacteria that occur in high polysaccharide environments. Such environments may be the digestive tracts of mammals and insects, plant-derived materials, products of the dairy industry, and the reproductive tract of women. The genus is crucial to the health of the host's gastrointestinal tract. As derived from their name they are known for their lactic acid production from lactose and other sugars. Each species of animal hosts different species of the genus (Tsuchida et al., 2014).

ROSA values for the *Lactobacillus* genus showed 4 distinct genera, two families and one order that encloses them all. The first cluster consisted of two strains of *L. delbrueckii*, the second cluster consisted of *L. helveticus* and *L. acidophilus*, the third cluster contained *L. gasseri* and *L. johnsonii*; the fourth cluster contained both strains of *L. brevis*. The first family contained *L. delbrueckii*, *L. helveticus*, *L. acidophilus*, *L. gasseri*, and *L. johnsonii*, while the second family contained *L. casei*, *L. reuteri*, *L. fermentum*, and *L. brevis* (Table 41). For intraspecies comparisons of strains of *L. delbrueckii* and *L. casei* ROSA values were 86.9 and 84.4 respectively (Table 41). Cluster 2 had intragenus values of 50.2, cluster 3 had intragenus values of 63.9 (Table 41). Values within the two family clusters ranged from 15.4 to 29.5 and 16.0 to 19.4, within the expected range for

intrafamily comparisons (Table 41). Comparisons between the two families yielded ROSA values between 13.1 and 14.8, within the expected range for interfamily comparisons (Table 41). These ROSA values indicate that the *Lactobacillus* genus has a much higher taxonomic ranking than currently proposed.

16s rRNA values showed the *Lactobacillus* has two distinct clusters. The first consists of *L. delbrueckii*, *L. helveticus*, *L. acidophilus*, *L. johnsonii*, and L . gasseri. The second cluster consisted of *L. reuteri*, *L. fermentum*, *L. casei*, and *L. brevis*. Members in cluster 1 had two spots of high similarity, between *L. acidophilus* and *L. helveticus* and between *L. johnsonii* and *L. gasseri*; these organisms had 98.4% and 99.5% similarity respectively (Table 42). Within the cluster, members shared between 91.9% and 93.8% 16s rRNA similarity (Table 42). These values within the cluster fall below the proposed 94% threshold for genus delimitation, supporting the ROSA classification. The similarities between *L. acidophilus* and *L. helveticus* were within the expected range for different species of the same genus, supporting the ROSA predictions. The values between *L. johnsonii* and *L. gasseri* were higher than what was expected for organisms within the same genus but are different species. This is another case in which relying on only a single highly conserved gene can lead to false positives. Within cluster two, intraspecies comparisons yielded a value of 99.7% similarity, within the expected range (Table 42). Intragenus comparisons yielded values below the 94% genus differentiation threshold, supporting the ROSA suggested classification. Values between the two clusters ranged well below 90% similarity, showing a great difference between the two

groups (Table 42). However since there is no accepted or proposed family level threshold no well-supported analysis could be made of these values.

AAI values within the genus supported the ROSA classification. Intraspecies comparisons between strains of *L. casei* and *L. delbrueckii* yielded AAI values of 98.7 and 99.2 respectively (Table 43). Proposed genera from cluster 1 including *L. helveticus* and *L. acidophilus* had an AAI value of 84.3, the proposed genus of *L. gasseri* and *L. johnsonii* yielded an AAI value of 88.8 (Table 43). These values are well within the estimated range for members of the same genus that are different species. Members of the proposed family 1, consisting of *L. delbrueckii*, *L. helveticus*, *L. acidophilus*, *L. johnsonii*, and L . gasseri, had AAI values ranging from 62.4 to 69.8 (Table 43). Members of the proposed family 2, consisting of *L. reuteri*, *L. fermentum*, *L. casei*, and *L. brevis*, had AAI values ranging from 52.3 to 65.6 (Table 43). AAI values between the two proposed families ranged from 50.1 to 51.9 (Table 43). Although there is no official threshold for AAI beyond the level of species it is clear that there is a difference in the relatedness between the two proposed families. The low levels of AAI between the proposed members showcase the differences between the organisms as well as highlighting the higher level of similarity between members of the cluster.

%BBH values within the genus supported the ROSA classification. Intraspecies comparisons between strains of *L. casei* and *L. delbrueckii* yielded AAI values of 84.5% to 89% and 84.2% to 92.4% of the genomes conserved respectively (Table 44). Proposed genera from cluster 1 including *L. helveticus* and *L. acidophilus* shared 72.7% of their genome, the proposed genus of *L. gasseri* and *L. johnsonii* shared 79% to 84% of their

genomes (Table 44). Members of the proposed family 1, consisting of *L. delbrueckii*, *L. helveticus*, *L. acidophilus*, *L. johnsonii*, and *L. gasseri*, had %BBH values ranging from 59.8% to 66.4% of genome conserved (Table 43). Members of the proposed family 2, consisting of *L. reuteri*, *L. fermentum*, *L. casei*, and *L. brevis*, had %BBH values ranging from 54.2% to 70.5% of their genomes conserved (Table 43). %BBH values between the two proposed families ranged from 38.3% to 56.4% of their genomes conserved (Table 43). Although there is no official threshold for BBH it is clear that there is a difference in the relatedness between the two proposed families. The low levels of conservation between the proposed members showcase the differences between the organisms as well as highlighting the higher level of similarity between members of the cluster.

When the genomes of *L. reuteri*, *L. brevis*, *L. casei*, *L. delbrueckii*, and *L. fermentum* were compared at the gene level the core genome consisted of only 679 genes. Each organism had a significant number of unique coding sequences *L. reuteri* contained 406 unique genes, *L. brevis* contained 628 unique genes, *L. casei* contained 1197 unique genes, *L. delbrueckii* contained 651 unique genes, while *L. fermentum* contained 493 unique genes (Figure 11). L brevis and *L. casei* shared a further 191 genes between them, *L. casei* further shared 118 genes with *L. delbrueckii* (Figure 11). *L. fermentum* and *L. reuteri* shared a further 123 genes between them (Figure 11). *L. reuteri*, *L. brevis*, *L. casei*, and *L. fermentum* shared a further 183 genes between them that *L. delbrueckii* did not contain (Figure 11). These shared genes were primarily scattered transcriptional regulatory proteins and hypothetical proteins. These organisms seem to be clustering in their own unique families within the cluster. This is supported

by the gene counts as no significant number of genes was shared between the organisms beyond what those shared between the four of them. This number is significant as they are all clustering in their own families outside of the one that includes *L. delbrueckii*.

When the genomes *L. johnsonii*, *L. acidophilus*, *L. delbrueckii*, *L. gasseri*, and *L. helveticus* were compared at the gene level, the core genome consisted of 923 genes. Each organism had a significant number of unique coding sequences. *L. johnsonii* contained 268 unique genes, *L. acidophilus* contained 282 unique genes, *L. delbrueckii* contained 625 unique genes, *L. gasseri* contained 208 unique genes, while *L. helveticus* contained 670 unique genes (Figure 12). *L. johnsonii*, *L. acidophilus*, *L. gasseri*, and *L. helveticus* shared 149 genes that *L. delbrueckii* did not (Figure 12). *L. johnsonii* and *L. gasseri* shared a further 143 genes (Figure 12). These genes primarily deal with sugar uptake and transport as well as a few resistance proteins (Table 44). *L. helveticus* and *L. acidophilus* shared a further 135 genes uniquely between them (Figure 12). These genes include sugar transport, citrate utilization, resistance proteins, and an operon of Imidazolonepropionase related amidohydrolase that may indicate an operon for metabolism of compounds containing C-N bonds (Table 42). When comparing the gene counts between the organisms it is easy to see the formation of genera between them. L acidophilus and *L. helveticus* shared 135 genes between them that the others did not contain; they are forming their own separate genus by the ROSA and 16s data. The second genus is formed by *L. johnsonii* and *L. gasseri*; they share an additional 143 genes

uniquely between them that the others do not, following the genus pattern suggested by ROSA.

Based on the data above it is recommended that the *Lactobacillus* genus be split into 2 different families, each containing multiple genera. The first family would consist of *L. delbrueckii*, *L. helveticus*, *L. acidophilus*, *L. gasseri*, and *L. johnsonii*. *L. helveticus* and *L. acidophilus* would form their own genus within the family as well as *L. gasseri* and *L. johnsonii* forming their own genus. All other members of the family would be single species genera. The second family would contain *L. casei*, *L. reuteri*, *L. fermentum*, and *L. brevis*, each of which would be single species genera. These families would both belong to a single order. The rationale behind this decision is that it is supported by all levels of evidence except for 16s rRNA. The 16s rRNA has a few outliers to this classification scheme, mainly calling *L. johnsonii* and *L. gasseri* as the same species. This is easily rectified as the 16s rRNA gene being too highly conserved between these two organisms. Furthermore based on ROSA, AAI, and %BBH those two organisms diverged not too long ago, which would explain why the 16s rRNA is still so highly conserved. One characteristic of these organisms that supports the reclassification is that each animal species hosts primarily one of the members of this genus. This provides evidence for long-term coevolution between symbiotic and host. Coevolution between host and symbiotic would provide sufficient isolation and selection to facilitate the diversification between strains of a single species and members of a genus.

## D. *Mycobacterium*

| | ROSA (sorted) | 83332 | 419947 | 336982 | 83331 | 410289 | 233413 | 272631 | 262316 | 243243 | 350058 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Mycobacterium tuberculosis H37Rv$^{TTT}$ | **83332** | | | | | | | | | | |
| Mycobacterium tuberculosis H37Ra | **419947** | 96.5 | | | | | | | | | |
| Mycobacterium tuberculosis F11 | **336982** | 95.9 | 96.0 | | | | | | | | |
| Mycobacterium tuberculosisCDC1551 | **83331** | 95.9 | 95.1 | 95.6 | | | | | | | |
| Mycobacterium bovis BCG str. Pasteur 1173P2 | **410289** | 93.4 | 95.0 | 93.5 | 93.3 | | | | | | |
| Mycobacterium bovisAF2122/97 | **233413** | 95.8 | 94.1 | 94.6 | 95.2 | 95.8 | | | | | |
| Mycobacterium leprae TN | **272631** | 43.8 | 43.6 | 43.8 | 43.6 | 43.6 | 44.0 | | | | |
| Mycobacterium avium subsp. paratuberculosis str. k10 | **262316** | 40.6 | 40.6 | 40.5 | 40.3 | 40.5 | 40.6 | 39.5 | | | |
| Mycobacterium avium 104 | **243243** | 39.2 | 39.3 | 39.2 | 38.8 | 39.2 | 39.2 | 38.6 | 85.6 | | |
| Mycobacterium vanbaalenii PYR-1$^{T}$ | **350058** | 27.3 | 27.3 | 27.3 | 27.0 | 27.2 | 27.2 | 28.7 | 30.1 | 28.9 | |
| Mycobacterium smegmatis str.MC2 155 | **246196** | 25.9 | 26.0 | 25.9 | 25.6 | 25.8 | 25.8 | 27.9 | 29.0 | 27.8 | 35.1 |

Table 47: ROSA values for members of the *Mycobacterium* genus. Intraspecies comparisons were done between strains of *M. tuberculosis*, *M. bovis*, and *M. avium*. Intragenus comparisons were done between *M. tuberculosis*, *M. bovis*, *M. avium*, M leprae, *M. vanbaalenii*, and *M. smegmatis*.

Intraspecies ROSA values for strains of *M. tuberculosis* ranged from 95.1 to 96.5, ROSA values for strains of *M. bovis* were 95.8, ROSA values for strains of *M. avium* were 85.6 (Table 47). Intragenus values for members of *Mycobacterium* ranged from 25.8 to 95.8 (Table 47). Two distinct clusters formed on the basis of ROSA values. The first cluster consisted of *M. tuberculosis*, *M. bovis*, *M. leprae*, and *M. avium*. Members of this cluster had ROSA values ranging from 39.2 to 95.8 (Table 47). Strains of *M. tuberculosis* and *M. bovis* had ROSA values over 94.1 (Table 47). Strains of *M. leprae* and *M. avium* had ROSA values ranging from 38.8 to 44.0 (Table 47). *M. vanbaalenii* and *M. smegmatis* had a ROSA value of 35.1 (Table 47). ROSA values suggest splitting the *Mycobacterium* genus into two genera, one containing *M. tuberculosis*, *M. bovis*, *M. leprae*, and *M. avium,* and the second consisting of *M. vanbaalenii* and *M. smegmatis.* ROSA values suggest that *M. tuberculosis* and *M. bovis* are members of the same species.

| | | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|
| Mycobacterium_tuberculosis_H37RvT | 1 | | | | | | |
| Mycobacterium_bovis_BCG_Pasteur_1173P2 | 2 | 100 | | | | | |
| Mycobacterium_leprae_TN | 3 | 98.3 | 98.3 | | | | |
| Mycobacterium_avium_subsp._avium_ATCC_25291T | 4 | 98.5 | 98.5 | 98.1 | | | |
| Mycobacterium_avium_subsp._paratuberculosis_ATCC_19698T | 5 | 98.5 | 98.5 | 98.1 | 100.0 | | |
| Mycobacterium_vanbaalenii_PYR-1T | 6 | 96.0 | 96.0 | 95.3 | 96.0 | 96.0 | |
| Mycobacterium_smegmatis_ATCC_19420T | 7 | 96.3 | 96.3 | 95.5 | 96.2 | 96.2 | 97.6 |

Table 48: 16s rRNA values for members of the *Mycobacterium* genus.

16s rRNA values for members of the *Mycobacterium* genus yielded two distinct clusters. The first cluster consisted of *M. tuberculosis*, *M. bovis*, *M. avium*, and *M. leprae*, with % similarities ranging from 98.1% to 100% (Table 48). Within this cluster *M. bovis* had 100% 16s rRNA similarity to *M. tuberculosis* (Table 48). The second cluster consisted on *M. vanbaalenii* and *M. smegmatis* with a 16s rRNA % similarity of 97.6% 16s rRNA values between these two clusters ranged from 95.3% to 96.2% similarity (Table 48).

| | Average Amino Acid Identity (AAIr) | 243243 | 262316 | 233413 | 410289 | 272631 | 246196 | 83331 | 336982 | 419947 | 83332 | 350058 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Mycobacterium avium 104 | **243243** | | 98.5 | 78.8 | 78.8 | 79.8 | 70.3 | 78.7 | 78.9 | 78.8 | 78.7 | 70.0 |
| Mycobacterium avium subsp. paratuberculosis str. k10 | **262316** | 98.5 | | 78.8 | 78.7 | 80.0 | 70.3 | 78.7 | 78.7 | 78.6 | 78.6 | 70.0 |
| Mycobacterium bovisAF2122/97 | **233413** | 78.8 | 78.8 | | 99.8 | 81.1 | 69.7 | 99.5 | 99.6 | 99.6 | 99.6 | 69.3 |
| Mycobacterium bovis BCG str. Pasteur 1173P2 | **410289** | 78.8 | 78.8 | 99.8 | | 81.2 | 69.9 | 99.6 | 99.7 | 99.6 | 99.6 | 69.3 |
| Mycobacterium leprae TN | **272631** | 79.3 | 79.6 | 81.0 | 81.0 | | 71.2 | 81.0 | 80.9 | 80.9 | 81.0 | 70.4 |
| Mycobacterium smegmatis str.MC2 155 | **246196** | 70.2 | 70.2 | 69.7 | 69.6 | 71.3 | | 69.6 | 69.7 | 69.6 | 69.6 | 74.0 |
| Mycobacterium tuberculosisCDC1551 | **83331** | 78.7 | 78.7 | 99.5 | 99.5 | 81.1 | 69.4 | | 99.8 | 99.7 | 99.7 | 69.2 |
| Mycobacterium tuberculosis F11 | **336982** | 78.9 | 78.6 | 99.6 | 99.6 | 81.1 | 69.8 | 99.8 | | 99.8 | 99.8 | 69.2 |
| Mycobacterium tuberculosis H37Ra | **419947** | 78.8 | 78.5 | 99.6 | 99.6 | 81.1 | 69.8 | 99.7 | 99.8 | | 99.9 | 69.2 |
| Mycobacterium tuberculosis H37Rv$^{TTT}$ | **83332** | 78.8 | 78.7 | 99.6 | 99.6 | 81.2 | 69.5 | 99.8 | 99.8 | 99.9 | | 69.1 |
| Mycobacterium vanbaalenii PYR-1$^{T}$ | **350058** | 70.0 | 70.1 | 69.3 | 69.4 | 70.7 | 74.0 | 69.3 | 69.2 | 69.3 | 69.2 | |

Table 49: AAI values for members of the *Mycobacterium* genus.

AAI values within the *Mycobacterium* genus had a great degree of variability and range. Strains of the same species, *M. avium*, *M. bovis*, and *M. tuberculosis* had AAI values of 98.5, 99.6, and 99.8 respectively (Table 49). Intragenus comparisons had AAI values ranging from 69.2 to 99.7 (Table 49). Strains of *M. bovis* when compared to

strains of *M. tuberculosis* had AAI values ranging from 99.6 to 99.7 (Table 49). Strains of *M. avium* and *M. leprae* when compared to *M. bovis* and *M. tuberculosis* had AAI values ranging 78.7 to 81.2 (Table 49). *M. smegmatis* and *M. vanbaalenii* had a reciprocal AAI value of 74 but AAI values ~70 when compared to other members of the genus (Table 49).

| | Percent Bidirectional Best Hit (% BBH) | 243243 | 262316 | 233413 | 410289 | 272631 | 246196 | 83331 | 336982 | 419947 | 83332 | 350058 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Mycobacterium avium 104 | **243243** | | 92.5 | 70.2 | 70.7 | 91.4 | 48.5 | 68.5 | 70.1 | 70.8 | 70.2 | 54.1 |
| Mycobacterium avium subsp. paratuberculosis str. k10 | **262316** | 83.9 | | 69.3 | 69.8 | 91.0 | 48.0 | 68.1 | 69.6 | 70.3 | 69.6 | 53.0 |
| Mycobacterium bovisAF2122/97 | **233413** | 56.3 | 61.5 | | 96.9 | 95.1 | 40.4 | 95.1 | 95.7 | 95.8 | 96.5 | 45.9 |
| Mycobacterium bovis BCG str. Pasteur 1173P2 | **410289** | 55.6 | 60.8 | 95.3 | | 93.7 | 39.6 | 92.1 | 93.9 | 96.0 | 93.5 | 45.2 |
| Mycobacterium leprae TN | **272631** | 30.1 | 33.1 | 38.9 | 38.8 | | 21.8 | 37.5 | 38.7 | 38.8 | 38.2 | 24.9 |
| Mycobacterium smegmatis str.MC2 155 | **246196** | 64.2 | 69.7 | 65.9 | 66.8 | 88.2 | | 64.7 | 66.4 | 67.1 | 66.2 | 67.4 |
| MycobacteriumtuberculosisCDC1551 | **83331** | 56.6 | 62.1 | 97.1 | 96.3 | 95.1 | 41.0 | | 97.6 | 97.8 | 97.8 | 46.2 |
| Mycobacterium tuberculosis F11 | **336982** | 55.8 | 61.3 | 95.0 | 94.6 | 94.7 | 40.0 | 94.3 | | 97.0 | 96.1 | 45.9 |
| Mycobacterium tuberculosis H37Ra | **419947** | 55.8 | 61.2 | 94.0 | 95.6 | 93.9 | 40.0 | 93.5 | 95.6 | | 96.0 | 45.6 |
| Mycobacterium tuberculosis H37Rv$^{TTT}$ | **83332** | 56.3 | 61.7 | 96.6 | 94.9 | 94.8 | 40.7 | 95.1 | 96.4 | 97.4 | | 46.2 |
| Mycobacterium vanbaalenii PYR-1$^{T}$ | **350058** | 64.0 | 69.5 | 67.6 | 67.7 | 90.2 | 60.7 | 66.1 | 67.9 | 68.3 | 67.9 | |

Table 50: %BBH values for members of the *Mycobacterium* genus.

Orthology between members of the *Mycobacterium* genus exhibited great variation and produced multiple clusters. Strains of the same species, *M. avium*, *M. bovis*, and *M. tuberculosis* had ~87%, ~96%, and ~93% to ~98% of their genomes conserved respectively (Table 50). Strains of *M. bovis* when compared to strains of *M. tuberculosis* shared~95% of their genomes, in some cases exhibiting greater conservation than some strains of *M. tuberculosis* exhibited towards each other (Table 50). Strains of *M. avium* and *M. leprae* when compared to *M. bovis* and *M. tuberculosis* shared ~56% to ~70% of their genomes (Table 50). In particular *M. leprae* had high levels of orthology with the rest of the genus, exhibiting greater than 88% genome conservation (Table 50). *M. smegmatis* and *M. vanbaalenii* shared ~64% of their genomes, *M. vanbaalenii* shared ~64% of its genome to other members of the genus

while *M. smegmatis* shared ~21% to ~39% of its genome to other members of its genus (Table 50).

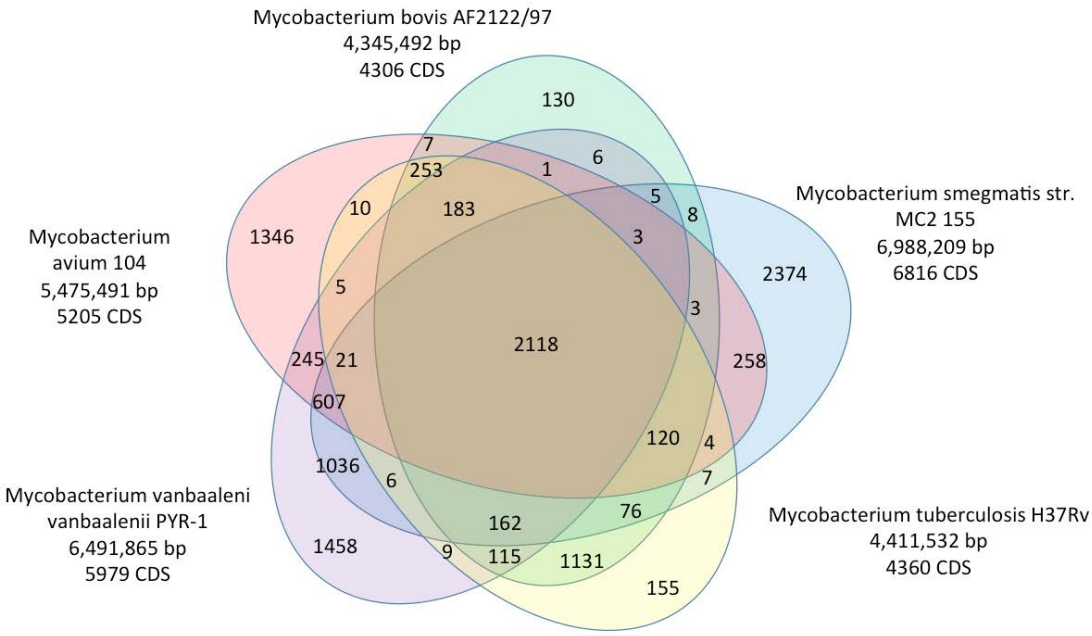


Figure 12: **Genomic similarity at the gene level for *M. vanbaalenii*, *M. avium*, *M. bovis*, *M. smegmatis*, and *M. tuberculosis*.** Venn-diagram shows the number of orthologous genes shared within each possible comparison combination.

The core genome of *M. vanbaalenii*, *M. avium*, *M. bovis*, *M. smegmatis*, and *M. tuberculosis* consisted of 2118 genes. *M. vanbaalenii* had 1458 unique genes with a genome size of 6,491,865 bp and 5679 CDS. *M. avium* had 1346 unique genes with a genome size of 5,475,491 bp and 5205 CDS. *M. bovis* had 130 unique genes with a genome size of 4,345,492 bp and 4306 CDS *M. smegmatis* had 2374 unique genes with a genome size 6,988,209 bp and 6816 CDS. *M. tuberculosis* had 155 unique genes with a genome size of 4,411,532 bp and 4360 CDS (Figure 12). *M. vanbaalenii*, and *M.*

*smegmatis* shared a further 1036 genes (Figure 12). These genes encoded for a variety of nitrogen utilization compounds, hydro peroxide resistance proteins, a unique sodium pump, unique metal resistance and utilization proteins, antibiotic resistance proteins, putrescence transporters, xylose utilization proteins, urea metabolism, and other carbohydrate metabolism proteins (Table 51). *M. bovis* and *M. tuberculosis* shared a further 1131 genes (Figure 12). *M. vanbaalenii* and *M. avium* shared a further 245 genes (Figure 12). *M. avium* and *M. smegmatis* shared a further 258 genes (Figure 12). *M. avium*, *M. smegmatis*, and *M. vanbaalenii* shared a further 607 genes (Figure 12). *M. avium*, *M. tuberculosis* and *M. bovis* shared a further 253 genes (Figure 12). *M. bovis*, *M. tuberculosis*, and *M. smegmatis* shared a further 120 genes (Figure 12).

| M. vanbaaleni and M. smegmatis only. | | | |
|---|---|---|---|
| 41 | putative regulatory protein | 2320 | [NiFe] hydrogenase nickel incorporation protein HypA |
| 42 | Urea carboxylase-related amino acid permease | 2321 | [NiFe] hydrogenase metallocenter assembly protein HypF |
| 43 | acetoacetyl CoA reductase [imported] (relatedtoshort-) | 2322 | [NiFe] hydrogenase metallocenter assembly protein HypC |
| 44 | 2-hydroxycyclohexanecarboxyl-CoA dehydrogenase (EC 1.1.1.-) | 2323 | [NiFe] hydrogenase metallocenter assembly protein HypD |
| 45 | Agmatinase (EC 3.5.3.11) | 2324 | [NiFe] hydrogenase metallocenter assembly protein HypE |
| 46 | polysaccharide deacetylase family protein | 2325 | benABC operon transcriptional activator BenR |
| | | 2327 | Uptake hydrogenase small subunit precursor (EC 1.12.99.6) |
| | | 2328 | Uptake hydrogenase large subunit (EC 1.12.99.6) |
| 92 | Phosphocarrier protein of PTS system | 2329 | Hydrogenase maturation protease (EC 3.4.24.-) |
| 93 | PTS system, fructose-specific IIA component (EC 2.7.1.69) | 2331 | NifU-like nitrogen fixation protein |
| 94 | 1-phosphofructokinase (EC 2.7.1.56) | 2332 | Transcriptional regulator |
| 95 | Transcriptional repressor of the fructose operon, DeoR family | 2333 | hypothetical protein |
| | | | |
| 337 | Isovaleryl-CoA dehydrogenase (EC 1.3.99.10) | 2369 | [NiFe] hydrogenase metallocenter assembly protein HypF |
| 338 | hypothetical protein | 2370 | hydrogenase assembly chaperone hypC-hupF (hypC) |
| 339 | hypothetical protein | 2371 | Hydrogenase maturation protease (EC 3.4.24.-) |
| 340 | Enoyl-CoA hydratase (EC 4.2.1.17) | 2378 | [NiFe] hydrogenase nickel incorporation-associated protein HypB |
| 341 | Methylcrotonyl-CoA carboxylase biotin-containing subunit (EC 6.4.1.4) | 2379 | [NiFe] hydrogenase nickel incorporation protein HypA |
| 342 | Methylcrotonyl-CoA carboxylase carboxyl transferase subunit (EC 6.4.1.4) | | |
| 343 | hypothetical protein | 2548 | Urea ABC transporter, ATPase protein UrtE |
| 344 | Two-component system, regulatory protein | 2549 | Urea ABC transporter, ATPase protein UrtD |
| 345 | Transcriptional regulator VpsT | 2550 | Urea ABC transporter, permease protein UrtC |
| 346 | Nitrilotriacetate monooxygenase component B (EC 1.14.13.-) | 2551 | Urea ABC transporter, permease protein UrtB |
| 347 | hypothetical protein | 2552 | Urea ABC transporter, urea binding protein |
| 348 | Aldehyde dehydrogenase (EC 1.2.1.3) | | |
| 349 | Long-chain-fatty-acid--CoA ligase (EC 6.2.1.3) | 3332 | Spermidine Putrescine transport ATP-binding protein potA (TC_3.A.1.11.1) |
| 350 | hypothetical protein | 3333 | Putrescine ABC transporter putrescine-binding protein potF (TC 3.A.1.11.2) |
| 351 | hypothetical protein | 3334 | Putrescine transport system permease protein potH (TC 3.A.1.11.2) |
| 352 | Acyl-CoA dehydrogenase, short-chain specific (EC 1.3.99.2) | 3335 | Putrescine transport system permease protein potI (TC 3.A.1.11.2) |
| | | 3336 | hypothetical protein |
| 430 | Oligopeptide transport ATP-binding protein oppF (TC 3.A.1.5.1) | 3337 | Integral membrane protein |
| 431 | Oligopeptide transport ATP-binding protein oppD (TC 3.A.1.5.1) | | |
| 432 | Dipeptide transport system permease protein dppB (TC 3.A.1.5.2) | 3396 | Nitrite-sensitive transcriptional repressor NsrR |
| 433 | Dipeptide transport system permease protein dppC (TC 3.A.1.5.2) | | |
| 434 | ABC transporter dipeptide binding protein | 3879 | Prolyl oligopeptidase family protein |
| | | 3880 | Putative ammonium transporter MJ0058 |
| 592 | Amino acid permease family protein | 3881 | hypothetical protein |
| 593 | Lactate-responsive regulator LldR in Actinobacteria, GntR family | 3882 | Mlr7324 protein |
| 594 | Enoyl-[acyl-carrier-protein] reductase [NADPH] (EC 1.3.1.10) | 3883 | Flavodoxin reductases (ferredoxin-NADPH reductases) family 1; (EC 1.14.13.-) |
| 596 | 2,3-dihydro-2,3-dihydroxybenzoate dehydrogenase (EC 1.3.1.28) | 3884 | hypothetical protein |
| 597 | hypothetical protein | | |
| 598 | glutamine synthetase family protein | 3891 | Transcriptional regulator protein |
| 599 | glutamine synthetase family protein | 3892 | Protein involved in biosynthesis of mitomycin antibiotics/polyketide fumonisin |
| 600 | Lipase/esterase | 3893 | Myo-inositol 2-dehydrogenase (EC 1.1.1.18) |
| 601 | 3-oxoacyl-[acyl-carrier protein] reductase (EC 1.1.1.100) | 3894 | Inosose dehydratase (EC 4.2.1.44) |
| 602 | Cyclohexanone monooxygenase (EC 1.14.13.22) | | |
| | | 4482 | Autolysis histidine kinase LytS |
| 725 | Na(+) H(+) antiporter subunit G | 4483 | alginate biosynthesis regulatory protein AlgR (lytT) |
| 726 | Na(+) H(+) antiporter subunit F | 4484 | putative integral membrane protein |
| 727 | Na(+) H(+) antiporter subunit E | 4485 | putative transmembrane transport protein |
| 728 | Na(+) H(+) antiporter subunit D | 4486 | INTEGRAL MEMBRANE PROTEIN (Rhomboid family) |
| 729 | Na(+) H(+) antiporter subunit C | 4487 | sodium:solute symporter protein |
| 730 | Na(+) H(+) antiporter subunit A; Na(+) H(+) antiporter subunit B | | |
| | | 5210 | Xylose ABC transporter, permease protein xylH |
| 917 | Transcriptional regulator, GntR family | 5211 | D-xylose transport ATP-binding protein xylG |
| 918 | Hydantoin permease | 5212 | Xylose ABC transporter, periplasmic xylose-binding protein xylF |
| 919 | Hydantoin racemase (EC 5.1.99.-) | 5213 | Xylose isomerase (EC 5.3.1.5) |
| 920 | polysaccharide deacetylase family protein | 5214 | Xylose repressor XylR (ROK family) |
| 921 | Allantoicase (EC 3.5.3.4) | 5215 | Xylulose kinase (EC 2.7.1.17) |
| 922 | Phosphohydrolase | 5217 | Acetoacetyl-CoA synthetase (EC 6.2.1.16) |
| | | 5218 | D-beta-hydroxybutyrate dehydrogenase (EC 1.1.1.30) |
| 1001 | hypothetical protein | 5220 | Transcriptional regulator, LysR family |
| 1002 | D-3-phosphoglycerate dehydrogenase (EC 1.1.1.95) | 5221 | Monoglyceride lipase (EC 3.1.1.23) |
| 1003 | Glucarate dehydratase (EC 4.2.1.40) | | |
| 1004 | Transcriptional regulator, IclR family | 5422 | Ammonium transporter |
| 1005 | Protein of unknown function DUF81 | 5423 | Glutamine synthetase type I (EC 6.3.1.2) |
| 1007 | Oligopeptide transport ATP-binding protein oppD (TC 3.A.1.5.1) | 5424 | Glucosamine--fructose-6-phosphate aminotransferase [isomerizing] (EC 2.6.1.16) |
| 1008 | Dipeptide transport system permease protein dppC (TC 3.A.1.5.2) | 5425 | Glutamate synthase [NADPH] large chain (EC 1.4.1.13) |
| 1009 | Probable ABC transporter permease | 5426 | Glutamate synthase [NADPH] large chain (EC 1.4.1.13) |
| 1010 | ABC transporter, substrate binding protein | | |
| 1012 | Phosphodiesterase/alkaline phosphatase D | 5474 | Propanediol dehydratase reactivation factor large subunit |
| 1015 | Glucarate dehydratase (EC 4.2.1.40) | 5475 | Propanediol dehydratase small subunit (EC 4.2.1.28); |
| 1016 | 5-dehydro-4-deoxyglucarate dehydratase (EC 4.2.1.41) | 5476 | Glycerol dehydratase large subunit (EC 4.2.1.30) |
| 1017 | Ketoglutarate semialdehyde dehydrogenase (EC 1.2.1.26) # hydroxy-L-proline-inducible | 5477 | amino acid permease family protein |
| 1019 | TRANSMEMBRANE SERINE | 5478 | Putrescine aminotransferase (EC 2.6.1.82) |
| 1020 | TolA protein | 5479 | aldehyde-alcohol dehydrogenase (sucD) |
| 1021 | TolA protein | 5480 | hypothetical protein |
| 1022 | hypothetical protein | 5481 | hypothetical protein |
| | | 5482 | carbon dioxide concentrating mechanism protein CcmL, putative |
| 1200 | Magnesium and cobalt transport protein CorA | 5483 | microcompartment protein |
| 1201 | Glycosyl transferase, group 1 | 5485 | Homoserine kinase (EC 2.7.1.39) |
| 1202 | Aliphatic amidase amiE (EC 3.5.1.4) | 5486 | 3-oxoacyl-[acyl-carrier protein] reductase (EC 1.1.1.100) |
| 1203 | Selenophosphate synthetase | 5487 | transcriptional regulator, GntR family, putative |
| 1204 | Biotin synthase-related enzyme | 5488 | glutamyl-Q-tRNA synthetase |
| 1205 | monooxygenase, putative | 5489 | Amino acid ABC transporter |
| 1206 | NAD/NADP transhydrogenase alpha subunit | 5490 | amino acid ABC transporter, ATP-binding protein (glnQ) |
| 1207 | Aliphatic amidase amiE (EC 3.5.1.4) | | |
| 1208 | SII0783 protein | 5737 | COG0523: Putative GTPases (G3E family) |
| | | 5738 | Nitrile hydratase subunit alpha (EC 4.2.1.84) (Nitrilase) (NHase) |
| 1504 | Organic hydroperoxide resistance transcriptional regulator | 5739 | PROBABLE NITRILE HYDRATASE SUBUNIT BETA PROTEIN (EC 4.2.1.84) |
| 1505 | Organic hydroperoxide resistance protein | | |
| | | | |
| 1520 | Flavohemoprotein (Nitric oxide dioxygenase) (EC 1.14.12.17) | 2176 | ABC transporter, permease protein, putative |
| | | 2177 | Pyoverdine efflux carrier and ATP binding protein |
| 1814 | Probable amino acid permease | 2178 | ABC transporter, ATP-binding protein homolog |
| 1815 | Regulatory protein | 2179 | amino acid permease family protein, putative |
| 1816 | hypothetical protein | 2180 | Copper amine oxidase precursor (EC 1.4.3.6) |
| 1817 | putative cytochrome P450 hydroxylase | 2181 | hypothetical protein |
| 1818 | Ferredoxin | | |
| 1819 | Glutamine synthetase type I (EC 6.3.1.2) | 2120 | Iron-dependent peroxidase |
| 1820 | Glutamine ABC transporter, periplasmic glutamine-binding protein (TC 3.A.1.3.2) | 2121 | Copper metallochaperone, bacterial analog of Cox17 protein |
| 1821 | Gamma-glutamyl-GABA hydrolase | 2122 | Uncharacterized iron-regulated membrane protein; Iron-uptake factor PiuB |
| 1822 | Gamma-glutamyl-GABA hydrolase | | |
| 1823 | putative regulatory protein | | |
| 1825 | Aldehyde dehydrogenase (EC 1.2.1.3) | | |
| 1826 | Glutaryl-CoA dehydrogenase (EC 1.3.99.7) | | |
| 1827 | 4-aminobutyrate aminotransferase (EC 2.6.1.19) | | |
| | | | |
| 2253 | Core component NikM of nickel ECF transporter / Additional core component NikN of nickel ECF transporter | | |
| 2254 | Core component NikM of nickel ECF transporter / Additional core component NikN of nickel ECF transporter | | |
| 2255 | Transmembrane component NikQ of energizing module of nickel ECF transporter | | |
| 2256 | ATPase component NikO of energizing module of nickel ECF transporter | | |

Table 51: Selected unique genes shared by *M. smegmatis* and *M. vanbaalenii*.

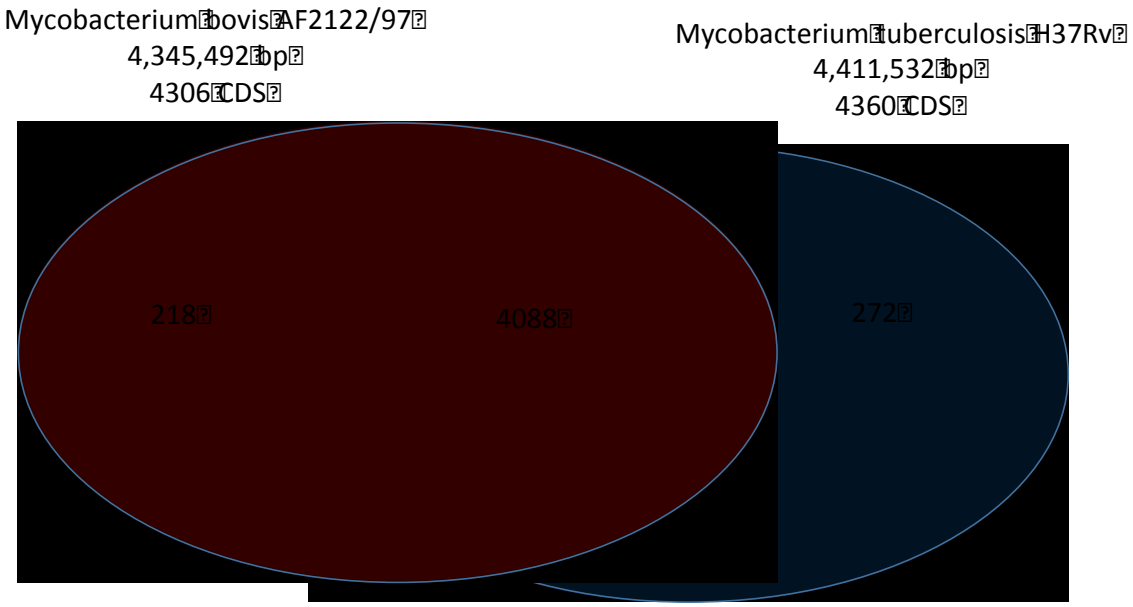


Figure 13: **Genomic similarity at the gene level for *M. bovis*, and *M. tuberculosis*.** Venn-diagram shows the number of orthologous genes shared within each possible comparison combination.

*M. bovis* and *M. tuberculosis* were chosen for a more detailed orthology examination. *M. bovis* and *M. tuberculosis* had a core genome consisting of 4088 genes. The organisms also had 218 and 272 unique genes respectively (Figure 13). When these genes were examined it was found that they consisted primarily of hypothetical proteins along with a few inner and outer membrane proteins.

## Mycobacterium Discussion

The *Mycobacterium* genus was established in 1896 with *Mycobacterium* tuberculosis being described in 1883 and designated as the type species in 1896. There are currently 170 validly published species. Organisms in the *Mycobacterium* genus can be pathogenic or non-pathogenic, fast or slow growing, and are found mostly in aquatic environments. Some of the main characteristic features of this genus are their yellow, scotochromatic pigmentation, and their unique cell wall, which is neither truly gram negative or gram positive. Members of the genus are typically aerobic and non-motile, acid-alcohol-fast, non-spore forming, and containing a thick hydrophobic cell wall rich in mycolic acids. Many can grow on simple substrates using ammonia and amino acids as a source for glycerol, nitrogen, and carbon (Devulder et al., 2005).

ROSA values for the *Mycobacterium* genus led to the formation of two different clusters. The *M. tuberculosis* cluster consisted of strains of *M. tuberculosis*, strains of *M. bovis*, strains of *M. avium*, and *M. leprae*. Intraspecies comparisons within this genus ranged from 93.4 to 96.5 for strains of *M. tuberculosis*, 95.8, and 85.6 for strains of *M. bovis* and *M. avium* respectively (Table 51). Intragenus comparisons within the cluster ranged from 38.6 to 44.0 for non-bovis strains (Table 51). *M. bovis* had ROSA scores <93 when compared to strains of *M. tuberculosis* (Table 51). These values indicate that *M. bovis* and *M. tuberculosis* are members of the same species, which disagrees with current taxonomy. Intragenus comparisons support *M. leprae* and *M. avium* as members of the genus with *M. bovis* and *M. tuberculosis*. *M. smegmatis* cluster consisted of *M. vanbaalenii* and *M. smegmatis*, with a ROSA score of 35.1 (Table 51).

The ROSA score for these two organisms is at the very border of the threshold for same genus classification between two organisms; they may belong to different genera in the same family. When members of the *M. tuberculosis* cluster and *M. smegmatis* cluster were compared ROSA values ranged from 25.9 to 30.1 (Table 51). These values suggest that the two clusters belong to different genera within the same family of organisms.

16s rRNA % similarity provides more evidence for the clustering as described by ROSA. Members of the *M. tuberculosis* cluster had 16s rRNA % similarities ranging from 98.1% to 100% between members (Table 47). Most members had values at or below the level necessary for species differentiation within the same genus. *M. bovis* and *M. tuberculosis* had 100% 16s rRNA % similarity between each other (Table 47). *M. vanbaalenii* and *M. smegmatis* had 97.6% 16s rRNA % similarity, a value that promotes species differentiation (Table 47). 16s rRNA values between the two clusters ranged from 95.3% to 96.3% similarity (Table 47). These values do not cross over the suggested threshold for 16s rRNA genus differentiation but do serve to showcase that there is a genetic gap between the two strains.

AAI values also mirrored the ROSA suggested placement. Strains of *M. tuberculosis*, *M. bovis*, and *M. avium* had AAI values ranging from 99.7 to 99.8, 99.5 to 99.6 and 98.5 respectively (Table 48). Strains of *M. bovis* and *M. tuberculosis* had AAI values ranging from 99.5 to 99.8 (Table 48). These values indicate that *M. bovis* is a member of the *M. tuberculosis* species due to its high degree of AAI similarity being well above the threshold for the metric. Intergenus comparisons within the cluster ranged from 78.7 to 81.2 (Table 48). Although there are no thresholds for AAI genus level

comparisons these values show a steady similarity between the organisms. Members of *M. smegmatis* cluster had intergenus AAI values of 74.0 (Table 48). These values are similar to the intergenus comparisons of the *M. tuberculosis* cluster. AAI values between the two clusters ranged from 69.1 to 71.3 indicating a clear separation between the two clusters (Table 48).

%BBH values matched the ROSA taxonomic predication except in the case of *M. leprae*. Strains of *M. tuberculosis*, *M. bovis*, and *M. avium* shared 93.5% to 97.8%, 95.3% to 96.9%, and 83.9% to 92.5% of their genomes respectively (Table 49). *M. bovis* and *M. tuberculosis* shared 92.1% to 97.1% of their genomes suggesting a higher level of relatedness than different species (Table 49). Intergenus comparisons within the cluster, excluding *M. leprae*, ranged from 54.6% to 70.8% of their genomes conserved (Table 49). *M. leprae* had at most 38.9% of its genome conserved to any member of the genus and 21.8% of its genome conserved at worst. However with *M. leprae* as a reference %BBH scores were inflated far beyond what would be expected (Table 49). Members of *M. smegmatis* cluster shared 67.4% of their genomes (Table 49). This value is within the range of the *M. tuberculosis* cluster showcasing similar levels of orthology. Intergenus comparisons between the two clusters, excluding *M. leprae*, ranged from sharing 39.6% to 67.9% of their genomes (Table 49). The inflated values for *M. vanbaalenii* and *M. smegmatis* as comparisons are due to their larger than average genome size, which will raise the %BBH value. Based on the %BBH values it is proposed that the two clusters described by the ROSA tool are valid taxonomic relationships.

When the genomes of *M. vanbaalenii*, *M. avium*, *M. bovis*, *M. smegmatis*, and *M. tuberculosis* were compared at the gene level, the core genome consisted of 2118 genes. Most organisms had a significant number of unique coding sequences. *M. vanbaalenii* contained 1458 unique genes, *M. avium* contained 1346 unique genes, *M. bovis* contained 130 unique genes, *M. smegmatis* contained 2364 unique genes, while *M. tuberculosis* contained 706 unique genes (Figure 13). *M. bovis* and *M. tuberculosis* shared a further 1131 genes (Figure 13). This high level of gene level similarity and having such low levels of gene uniqueness indicates that these two organisms are more closely related than species of the same genus. To further support this close relatedness in all comparisons between the other 3 organisms *M. bovis* and *M. tuberculosis* showed the highest amount of similarity when compared together. When they were compared separately the number of genes conserved between organisms usually numbered under 10 (Figure 13).

*M. smegmatis* and *M. vanbaalenii* shared 1036 genes between them, yet each retained a large amount of unique genes, 2374 and 1459 respectively (Figure 13). *M. avium* showed high levels of similarity to *M. bovis*, *M. tuberculosis*, *M. smegmatis* and *M. vanbaalenii*. When compared to *M. vanbaalenii* alone, *M. smegmatis* alone, and the two together they shared 245, 258, and 607 genes respectively (Figure 13). When compared to *M. bovis* and *M. tuberculosis* it shared 253 genes with them (Table 14). Although this seems like a low amount of gene level similarity with members of its genus, it must also be noted the effect the genome size will have on this. *M. bovis* and *M. tuberculosis* share over 80% of their genome just in the core genome and the genes

they share uniquely to themselves. In that context the amount of genes the organisms share is well within the amount expected for members of the same genus. Genome size also explains the high number of genes *M. avium* shares with *M. smegmatis* and *M. vanbaalenii*. Both have much larger genome sizes than the other organisms and as such will more than likely have more room to house genes that can be shared between the organisms. When *M. bovis* and *M. tuberculosis* were compared by themselves it was found that they shared 4088 genes with 218 and 272 unique to each respectively (Figure 15). These genes encoded primarily for hypothetical proteins along with a few inner and outer membrane proteins, consistent for diseases with slightly different pathologies and tropism.

Based on the data presented for the cluster a number of conclusions can be drawn. Based on the ROSA values it is suggested that the *Mycobacterium* genus be split into two separate genera within the same family. The first genus will consist of *M. tuberculosis*, *M. bovis*, *M. leprae*, and *M. avium*, the second will consist of *M. vanbaalenii* and *M. smegmatis* (Table 51). Within the ROSA values it was suggested that *M. tuberculosis* and *M. bovis* be reclassified as the same species. 16s rRNA similarities for members of the *M. tuberculosis* cluster showed this trend as well as supporting the combining of *M. bovis* and *M. tuberculosis*. *M. smegmatis* cluster had 16s rRNA similarity values within the range expected of organisms in the same genus. Intercluster comparisons showed that the two clusters have a definitive split under the 94% suggested genus level threshold (Table 47). AAI values further support the clustering of the organisms suggested by ROSA.

The AAI scores for the *M. tuberculosis* cluster supported classification of *M. bovis* and *M. tuberculosis* as the same species. Values within the *M. tuberculosis* cluster ranged from 78.7 to 81.2. *M. smegmatis* cluster had an internal AAI value of 74. Intercluster comparisons had AAI values of 69.1 to 71.3 (Table 48). Although there is no accepted genus level threshold for AAI the values exhibited showed distinct differences between the clusters. %BBH provides further evidence for splitting the two clusters, when excluding *M. leprae* for reasons explained later on. Values also support the combining of *M. bovis* and *M. tuberculosis*, %BBh values within the *M. tuberculosis* cluster ranged from 54.6% to 97.1%. Within *M. smegmatis* cluster BBH values were 67.4%. Intercluster comparisons ranged between 39.6% to 67.9% (Table 49). The difference in intercluster value vs. intracluster values provides evidence for the split in the genera. *M. leprae* was an outlier of the group with %BBH values ranging from 21.8% to 38.9% (Table 49). It must be noted that even with these low values a pattern of similarity could be noted. *M. leprae* shared over 30% of its genome with members of the *M. tuberculosis* cluster while it only shared at most 24.8% of its genome with members of *M. smegmatis* cluster. This showcases that although the values are lower than expected there is still a clear level of phylogeny between *M. leprae* and members of the *M. tuberculosis* cluster. When the similarity was compared on the gene level it was found that the organisms all share 2118 genes (Figure 13). *M. bovis* and *M. tuberculosis* shared a further 1131 genes uniquely between them, while having 130 and 155 genes unique to them respectively. This high level of gene conservation as well as low level of gene uniqueness suggests that the two species should be reclassified as members of the

same species. *M. smegmatis* and *M. vanbaalenii* shared an additional 1036 genes between each other, indicating a large level of relatedness between them. Although they shared a large number of genes with *M. avium*, more than would be expected based on the ROSA values, this is somewhat expected due to the larger genome sizes of *M. smegmatis* and *M. vanbaalenii*.

Based on the above analysis it is recommended that the *Mycobacterium* genus be split into two separate genera within the same family. The first genus consisting of *M. tuberculosis*, *M. bovis*, *M. avium*, and *M. leprae* with *M. bovis* and *M. tuberculosis* being combined into one species. *M. leprae* is an interesting case that shows the strength of the ROSA metric. If one was to only look at the AAI or %BBH of the organism it would be easy to misinterpret the values and miss the true classification of the organism. Extensive study has been done concerning the genome of *M. leprae*. It was found that the organisms originally had a genome as large as *M. tuberculosis*, however a few thousand years ago it began to undergo extensive reductive evolution. During this process a large number of previously essential genes were deleted or mutated into a form that was no longer functional (Cole et al. 2001). In regards to this massive size reduction it is estimated that if the genes could be recognized and compared back to the genomes within its cluster, a large number relate to possible orthologs and raise the %BBH score to a level expected for that level of comparison. The second genus would consist of *M. smegmatis* and *M. vanbaalenii*. Both of the proposed genera would remain within the family the organisms already belong to. It is interesting to note that the classification proposed by ROSA separates the organisms by two phenotypic

characteristics as well. The first is that members of the first proposed genus are all commonly associated with disease while those in the second proposed genus are rarely or never associated with disease. The second phenotype is that members in the first proposed genus are all considered slow growers while those in the second proposed genus are fast growers (Stahl and Urbance, 1990).

## E. *Streptococcus*:

| | Reciprocal Orthology Score Average (ROSA) | 373153 | 453364 | 216600 | 246201 | 467705 | 210007 | 208435 | 198466 | 160490 | 552526 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Streptococcus pneumonia pneumoniae D39 | **373153** | | | | | | | | | | |
| Streptococcus pneumoniae CDC0288-04 | **453364** | 88.6 | | | | | | | | | |
| Streptococcus pneumoniae 23F | **216600** | 84.7 | 78.6 | | | | | | | | |
| Streptococcus mitis NCTC 12261$^{T}$ | **246201** | 60.7 | 56.1 | 58.4 | | | | | | | |
| Streptococcus gordonii str. Challis substr. CH1 | **467705** | 38.6 | 35.3 | 37.7 | 35.8 | | | | | | |
| Streptococcus mutans UA159 | **210007** | 28.1 | 25.9 | 26.9 | 25.6 | 30.0 | | | | | |
| Streptococcus agalactiae 2603V/R | **208435** | 26.8 | 24.8 | 27.4 | 24.6 | 26.1 | 30.6 | | | | |
| Streptococcus pyogenes MGAS315 | **198466** | 25.9 | 24.1 | 25.6 | 22.6 | 25.8 | 28.1 | 33.4 | | | |
| Streptococcus pyogenes M1 GAS | **160490** | 27.5 | 25.7 | 27.3 | 24.5 | 27.5 | 30.3 | 35.3 | 85.6 | | |
| Streptococcus equi subsp. zooepidemicus MGCS10565 | **552526** | 27.8 | 25.8 | 27.2 | 24.9 | 26.5 | 30.4 | 33.3 | 43.7 | 46.5 | |
| Streptococcus equi subsp. Equi | **148942** | 26.2 | 24.1 | 25.9 | 23.3 | 25.3 | 28.7 | 31.4 | 44.4 | 45.6 | 79.4 |

Table 52: ROSA values for members of the *Streptococcus* genus. Intraspecies comparisons were done for strains of *S. pneumoniae*, *S. pyogenes*, and *S. equi*. Intragenus comparisons were done between *S. pneumoniae*, *S. mitis*, *S. gordonii*, *S. mutans*, *S. agalactiae*, *S. pyogenes*, and *S. equi*.

Intraspecies ROSA values for strains of *S. pneumoniae*, *S. pyogenes*, and *S. equi* had ROSA values ranging 84.7 and 88.6, 85.6 and 79.4 respectively (Table 52). Intragenus comparisons between *S. pneumoniae*, *S. mitis*, *S. gordonii*, *S. mutans*, *S. agalactiae*, *S. pyogenes*, and *S. equi* had ROSA values ranging from 23.2 to 60.7 (Table 52). Two distinct clusters arose from ROSA values. The first consisted of *S. pneumoniae*, *S. mitis*, and *S. gordonii*, whom had ROSA value ranging from 35.3 to 60.7 (Table 52). The second cluster consisting of *S. pyogenes* and *S. agalactiae* had ROSA values ranging from 31.4 to 46.5 (Table 52). Intragenus comparisons between other organisms had ROSA

values ranging from 20.1 to 33.6 (Table 52). ROSA values suggest splitting the *Streptococcus* genus into 4 genera. The fist containing *S. pneumoniae*, *S. mitis*, and *S. gordonii,* the second *S. mutans*, the third *S. agalactiae,* and the fourth *S. pyogenes*, and *S. equi.*

| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|---|
| Streptococcus_pneumoniae_ATCC_33400T | 1 | | | | | | | | |
| Streptococcus_pseudopneumoniae_ATCC_BAA-960T | 2 | 99.6 | | | | | | | |
| Streptococcus_mitis_NCTC_12261T | 3 | 99.4 | 99.4 | | | | | | |
| Streptococcus_gordonii_ATCC_10558T | 4 | 96.6 | 96.7 | 97.1 | | | | | |
| Streptococcus_mutans_ATCC_25175T | 5 | 92.8 | 93.0 | 93.2 | 94.6 | | | | |
| Streptococcus_agalactiae_ATCC_13813T/ | 6 | 94.0 | 94.0 | 94.4 | 94.4 | 92.3 | | | |
| Streptococcus_pyogenes_JCM_5674T | 7 | 94.0 | 94.0 | 94.2 | 95.0 | 92.5 | 96.3 | | |
| Streptococcus_equi_subsp._equi_ATCC_33398T | 8 | 93.3 | 93.3 | 93.3 | 94.0 | 93.3 | 94.4 | 95.3 | |
| Streptococcus_equi_subsp._zooepidemicus_ATCC_43079T | 9 | 92.6 | 92.7 | 92.7 | 93.3 | 92.6 | 93.7 | 94.7 | 99.0 |

Table 53: 16s rRNA % similarity values for members of the *Streptococcus* genus.

Intraspecies for *S. pneumoniae* and *S. equi* were 99.6% and 99.0% respectively (Table 53). 16s rRNA values between *S. pneumoniae*, *S. mitis*, and *S. gordonii* ranged from 96.6% to 99.4% (Table 53). 16s rRNA values between *S. pyogenes* and *S. equi* were 96.3% (Table 53). Other intragenus comparisons ranged from 92.3% to 95% 16s rRNA similarity (Table 53).

| | Average Amino Acid Identity (AAIr) | 208435 | 148942 | 552526 | 467705 | 246201 | 210007 | 373153 | 216600 | 453364 | 160490 | 198466 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Streptococcus agalactiae 2603V/R | **208435** | | 69.6 | 70.0 | 64.0 | 64.4 | 69.4 | 64.3 | 64.2 | 65.0 | 71.9 | 71.6 |
| Streptococcus equi subsp. Equi | **148942** | 69.5 | | 97.6 | 64.6 | 64.5 | 68.9 | 64.7 | 64.1 | 64.9 | 79.0 | 79.7 |
| Streptococcus equi subsp. zooepidemicus MGCS10565 | **552526** | 70.1 | 97.7 | | 64.5 | 65.6 | 69.3 | 65.2 | 65.0 | 65.6 | 79.1 | 79.4 |
| Streptococcus gordonii str. Challis substr. CH1 | **467705** | 64.5 | 65.1 | 64.7 | | 74.4 | 66.6 | 74.6 | 74.4 | 74.5 | 65.6 | 65.8 |
| Streptococcus mitis NCTC 12261$^{T}$ | **246201** | 65.9 | 66.3 | 66.4 | 74.4 | | 66.9 | 92.9 | 91.8 | 93.4 | 66.7 | 66.4 |
| Streptococcus mutans UA159 | **210007** | 69.7 | 69.2 | 69.4 | 66.3 | 66.2 | | 66.1 | 65.0 | 65.6 | 68.8 | 69.1 |
| Streptococcus pneumonia pneumoniae D39 | **373153** | 64.8 | 65.1 | 65.5 | 74.3 | 92.8 | 66.1 | | 98.2 | 98.6 | 65.2 | 65.3 |
| Streptococcus pneumoniae 23F | **216600** | 64.8 | 64.5 | 65.3 | 74.1 | 92.4 | 65.6 | 98.1 | | 98.7 | 65.3 | 65.1 |
| Streptococcus pneumoniae CDC0288-04 | **453364** | 64.8 | 64.9 | 65.7 | 74.4 | 92.8 | 65.8 | 98.6 | 98.7 | | 65.6 | 65.2 |
| Streptococcus pyogenes M1 GAS | **160490** | 71.6 | 79.1 | 79.1 | 65.4 | 65.8 | 69.0 | 65.3 | 65.1 | 66.1 | | 98.1 |
| Streptococcus pyogenes MGAS315 | **198466** | 71.7 | 79.8 | 79.4 | 65.4 | 65.3 | 69.0 | 65.3 | 65.0 | 65.8 | 98.0 | |

Table 54: AAI values for members of the *Streptococcus* genus.

Intraspecies clustering between strains of *S. pyogenes*, *S. pneumoniae*, and *S. equi* had AAI values of 98.1, 98.4, and 97.7 respectively (Table 54). AAI values for members of the *Streptococcus* genus showed two distinct clusters within the genus. The first cluster consisted of *S. pyogenes* and *S. equi* with AAI values ranging from 79.1 to 79.9 (Table 55). The second cluster consisted of *S. pneumoniae*, *S. mitis*, and *S. gordonii* with AAI values ranging from 74.4 to 93.4 (Table 54). Other members of the genus had AAI values ranging from 64.0 to 70 when compared to members of the two clusters (Table 54).

| | Percent Bidirectional Best Hit (% BBH) | 208435 | 148942 | 552526 | 467705 | 246201 | 210007 | 373153 | 216600 | 453364 | 160490 | 198466 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Streptococcus agalactiae 2603V/R | **208435** | | 65.9 | 71.6 | 63.3 | 57.3 | 66.4 | 66.8 | 67.6 | 57.8 | 75.2 | 70.0 |
| Streptococcus equi subsp. Equi | **148942** | 64.1 | | 87.1 | 59.6 | 53.7 | 62.3 | 64.1 | 63.5 | 55.1 | 79.4 | 74.3 |
| Streptococcus equi subsp. zooepidemicus MGCS10565 | **552526** | 64.4 | 79.6 | | 60.0 | 53.5 | 62.9 | 64.0 | 62.1 | 55.4 | 78.0 | 71.1 |
| Streptococcus gordonii str. Challis substr. CH1 | **467705** | 63.2 | 61.0 | 67.1 | | 63.2 | 71.3 | 72.0 | 69.8 | 62.6 | 70.8 | 64.6 |
| Streptococcus mitis NCTC 12261$^{T}$ | **246201** | 58.4 | 55.3 | 60.9 | 66.1 | | 61.5 | 73.5 | 71.3 | 63.7 | 62.4 | 57.0 |
| Streptococcus mutans UA159 | **210007** | 60.2 | 57.9 | 63.3 | 64.8 | 54.0 | | 63.4 | 61.7 | 56.4 | 67.6 | 60.7 |
| Streptococcus pneumonia pneumoniae D39 | **373153** | 62.1 | 60.6 | 66.0 | 67.2 | 67.3 | 65.4 | | 86.7 | 92.4 | 69.3 | 63.3 |
| Streptococcus pneumoniae 23F | **216600** | 64.1 | 62.0 | 65.9 | 66.8 | 66.6 | 64.6 | 89.1 | | 77.4 | 69.4 | 64.0 |
| Streptococcus pneumoniae CDC0288-04 | **453364** | 60.1 | 59.5 | 64.1 | 64.9 | 65.9 | 63.4 | 90.0 | 84.1 | | 67.0 | 62.0 |
| Streptococcus pyogenes M1 GAS | **160490** | 61.9 | 66.7 | 70.6 | 57.2 | 49.2 | 60.0 | 59.9 | 58.8 | 51.6 | | 86.9 |
| Streptococcus pyogenes MGAS315 | **198466** | 60.3 | 65.4 | 67.6 | 55.2 | 47.2 | 57.1 | 58.0 | 57.0 | 50.4 | 91.3 | |

Table 55: %BBH values for members of the *Streptococcus* genus.

Orthology between members of the *Streptococcus* genus revealed two distinct clusters of organisms. The first cluster consisted of *S. equi* and *S. pyogenes* with orthology ranging from 65.4% to 79.4% of the genomes conserved (Table 55). Intraspecies comparisons of strains within the cluster revealed 84% of genome conservation for strains of *S. equi* and 90% of genome conservation for strains of *S. pyogenes* (Table 55). The second cluster consisted of *S. pneumoniae*, *S. gordonii*, and *S. mitis* whom shared 69.8% to 73.5% of their genomes (Table 55). Strains of *S.*

*pneumoniae* shared 77% to 90% of their genomes (Table 55). *S. mutans* and *S. mitis* showed little orthology with the rest of the genus, sharing 47.2% to 71.3% of their genome (Table 55).

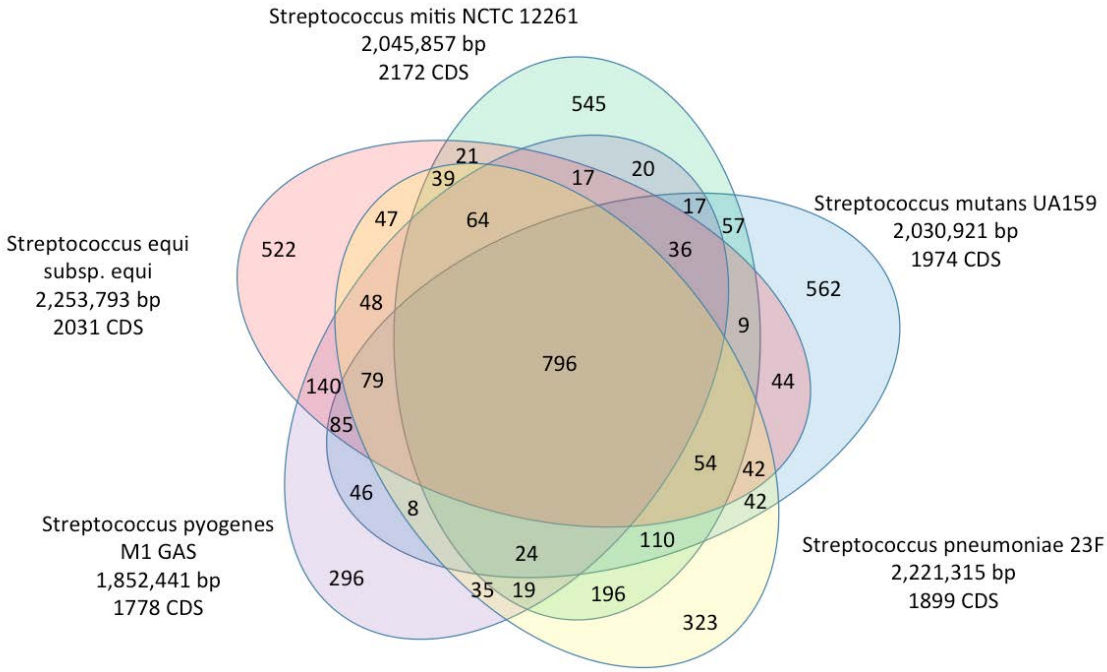


Figure 14: **Genomic similarity at the gene level for *S. pyogenes*, *S. equi*, *S. mitis*, *S. mutans*, and *S. pneumoniae*.** Venn-diagram shows the number of orthologous genes shared within each possible comparison combination.

The core genome of *S. pyogenes*, *S. equi*, *S. mitis*, *S. mutans*, and *S. pneumoniae* consisted of 796 genes. *S. pyogenes* had 296 unique genes with a genome size of 1,852,441 bp and 1778 CDS. *S. equi* had 522 unique genes with a genome size of 2,253,491 bp and 2031 CDS. *S. mitis* had 545 unique genes with a genome size of 2,045,857 bp and 2172 CDS *S. mutans* had 562 unique genes with a genome size 2,030,921 bp and 1874 CDS. *S. pneumoniae* had 323 unique genes with a genome size of 2,221,315 bp and 1899 CDS (Figure 14). *S. mitis*, *S. mutans*, and *S. pneumoniae* shared a

further 110 genes (Figure 14). *S. mitis* and *S. pneumoniae* shared a further 196 genes (Figure 14). These genes encoded metal-related proteins and carbohydrate metabolism (Table 56). *S. pyogenes* and *S. equi* shared a further 140 genes (Figure 14). These genes encode for the synthesis of Streptolysin S, cell surface proteins, drug resistance, a captured phage genome, and a specialized ribonucleotide reductase (Table 56).

A

| | S. equi and S. pyogenes only. |
|---|---|
| 85 | Putative Nudix hydrolase YfcD (EC 3.6.-.-) |
| 121 | Pathogenicity island SaPIn2 |
| 122 | Phage transcriptional repressor |
| 251 | Sensory transduction protein kinase (EC 2.7.3.-) |
| 252 | Sensory transduction protein kinase (EC 2.7.3.-) |
| 253 | Response regulator FasA or ComE or BlpR |
| | |
| 388 | Streptococcal cell surface hemoprotein receptor Shr |
| 390 | Cell surface protein Shp, transfers heme from hemoglobin to apo-SiaA/HtsA |
| 391 | Heme ABC transporter (Streptococcus), heme and hemoglobin-binding protein SiaA/HtsA |
| 394 | Putative ABC transporter (ATP-binding protein), spy1791 homolog |
| 395 | Putative ABC transporter ATP-binding protein, spy1790 homolog |
| 486 | Streptolysin S biosynthesis protein B (SagB) |
| 487 | Streptolysin S biosynthesis protein C (SagC) |
| 488 | Streptolysin S biosynthesis protein D (SagD) |
| 489 | Streptolysin S self-immunity protein (SagE) |
| 490 | Streptolysin S biosynthesis protein (SagF) |
| | |
| 509 | Ribonucleotide reductase of class Ib (aerobic), beta subunit (EC 1.17.4.1) |
| 510 | Ribonucleotide reduction protein NrdI |
| 511 | Ribonucleotide reductase of class Ib (aerobic), alpha subunit (EC 1.17.4.1) |
| | |
| 1452 | Fibronectin-binding protein |
| 1453 | Fibronectin-binding protein |
| 1470 | D-beta-hydroxybutyrate permease |
| 1471 | D-beta-hydroxybutyrate dehydrogenase (EC 1.1.1.30) |
| 1472 | Acetate CoA-transferase beta subunit (EC 2.8.3.8) |
| 1473 | Acetate CoA-transferase alpha subunit (EC 2.8.3.8) |
| 1474 | 3-ketoacyl-CoA thiolase (EC 2.3.1.16); Acetyl-CoA acetyltransferase (EC 2.3.1.9) |
| 1475 | Transcriptional regulators, LysR family |
| | |
| 1578 | Phage endopeptidase |
| 1579 | Tail protein |
| 1580 | Minor tail protein GP26 |
| 1581 | conserved hypothetical protein - phage associated |
| 1582 | conserved hypothetical protein - phage associated |
| 1583 | Major tail protein |
| 1585 | Structural protein |
| 1586 | hypothetical phage protein |
| 1587 | Phage protein |
| 1588 | phi Mu50B-like protein |
| 1591 | Phi Mu50B-like protein |
| 1950 | Multidrug resistance protein B |

B

| | S. mitis and S. pneumonia only. |
|---|---|
| 136 | Ferric iron ABC transporter, permease protein # pitB, pitC |
| 137 | Ferric iron ABC transporter, ATP-binding protein # pitD |
| | |
| 193 | ABC-type polysaccharide transport system, permease component |
| 194 | ABC transporter, permease protein |
| 195 | ABC transporter, substrate-binding protein |
| | |
| 216 | Exopolysaccharide biosynthesis transcriptional activator EpsA |
| 217 | Manganese-dependent protein-tyrosine phosphatase (EC 3.1.3.48) |
| 218 | Tyrosine-protein kinase transmembrane modulator EpsC |
| 219 | Tyrosine-protein kinase EpsD (EC 2.7.1.112) |
| | |
| 391 | Conserved thiamin-related protein TenA |
| 392 | Hydroxyethylthiazole kinase (EC 2.7.1.50) |
| 393 | Thiamin-phosphate pyrophosphorylase (EC 2.5.1.3) |
| 394 | Core component YkoE of thiamin-regulated ECF transporter for HydroxyMethylPyrimidine |
| 396 | Transmembrane component YkoC of energizing module of thiamin-regulated ECF transporter for HydroxyMethylPyrimidine |
| 397 | Conserved thiamin-related protein TenA |
| 398 | Predicted thiazole transporter ThiW |
| 399 | Hydroxyethylthiazole kinase (EC 2.7.1.50) |
| 400 | Thiamin-phosphate pyrophosphorylase (EC 2.5.1.3) |
| 401 | Phosphomethylpyrimidine kinase (EC 2.7.4.7) |
| | |
| 404 | Copper-exporting ATPase (EC 3.6.3.4) |
| | |
| 414 | metallo-beta-lactamase family protein |
| | |
| 831 | Deblocking aminopeptidase (EC 3.4.11.-) |
| 832 | Branched-chain amino acid transport system carrier protein |
| 834 | Uracil-DNA glycosylase |
| 835 | Xaa-His dipeptidase (EC 3.4.13.3) |
| 838 | Serine/threonine protein phosphatase (EC 3.1.3.16) |
| | |
| 926 | cadmium resistance transporter, putative |
| | |
| 1377 | Multi antimicrobial extrusion protein (Na(+)/drug antiporter), MATE family of MDR efflux pumps |
| 1468 | Spore coat polysaccharide biosynthesis protein spsC |
| 1792 | Copper-transporting P-type ATPase (EC 3.6.3.4) |
| 1813 | Magnesium and cobalt transport protein corA |
| | |
| | |
| | |
| | |

Table 56. A. Selected unique genes shared between *S. equi* and *S. pyogenes*. B. Selected unique genes shared between *S. mitis* and *S. pneumoniae*.

## Streptococcus Discussion

The *Streptococcus* genus was established in 1884 with *Streptococcus* pyogenes being characterized as the type species in the same year. There are currently 110 validly published members. General characteristics of this genus are that they are gram positive, round, divide in bending chains, and have colony-like morphology. They inhabit a diverse range of organisms and are generally pathogenic. Due to the large diversity of the group there are few characteristics that span the entire genus (Parte 2014).

ROSA values for members of the *Streptococcus* genus revealed two distinct clusters and 2 outliers. The *S. pneumoniae* cluster consisted of *S. pneumoniae*, *S. mitis*, and *S. gordonii*. The *S. pyogenes* cluster consisted of *S. pyogenes* and *S. equi*. The two outliers were *S. mutans* and *S. agalactiae*. ROSA values for intraspecies comparisons ranged from 78.6 to 88.6, 85.6, and 79.4 for strains of S. pneumonia, *S. pyogenes*, and *S. equi* respectively, values within the expected ROSA range (Table 52). Intergenus comparisons for members of the *S. pneumoniae* cluster ranged from 35.3 to 60.7, intergenus comparisons for the *S. pyogenes* cluster ranged from 43.7 to 46.5, all within the expected ROSA range for members of the same genus (Table 52), the two outliers had ROSA values ranging from 24.6 to 30.6, within the range expected for genera of the same family. Intercluster comparisons ranged from 24.5 to 28.1, ranges that are expected for different genera that are part of the same family (Table 52). The ROSA

values predict that the two clusters are going to cluster into separate genera within the same species with the two outliers forming their own genera as well.

16s rRNA similarities between strains of *S. pneumoniae* and *S. equi* were 99.6% and 99.0% respectively (Table 53). These values indicate that the organisms are indeed members of the same species. Intergenus comparisons within the *S. pneumoniae* cluster ranged from 96.6% to 99.4% (Table 53). These values are in some cases above what we would expect from members of the same genus, but as mentioned earlier 16s rRNA is sometimes too highly conserved between two taxonomically distinct species. Within the *S. pyogenes* cluster intergenus comparisons had a value of 96.3%, this value is well within what would be expected for organisms of the same genus but different species (Table 53). When 16s rRNA sequences between the two outliers were compared between clusters, values ranged from 92.3% to 95% 16s rRNA similarity (Table 53). These values fit within the proposed genus threshold of 94%-95% similarity, thus supporting the ROSA clustering.

AAI values within the *Streptococcus* genus further supported the ROSA clustering. Intraspecies clustering between strains of *S. pyogenes*, S. pneumonia, and *S. equi* had AAI values of 98.1, 98.4, and 97.7 respectively (Table 54). AAI values between members of the *S. pneumoniae* cluster ranged from 74.4 to 93.4 (Table 54). Although no hard threshold exists for the genus level, the tight clustering of the organisms shows a large amount of similarity between them. Members of the *S. pyogenes* cluster had intragenus AAI values ranging from 79.1 to 89.9, showing a similar level of relatedness as the organisms in the *S. pneumoniae* cluster. When the outliers were compared to

both clusters, and the clusters were compared to each other, AAI values ranged from 64.0 to 70 (Table 54). These AAI values show that there is a significant difference in the relatedness between the two clusters and the two outliers. This provides evidence for a separation consistent with the ROSA values and the 16s rRNA values.

%BBH values also showed two distinct clusters and two outliers. Intraspecies comparisons for the *S. pneumoniae* cluster yielded %BBH values of 77.4% to 90% conserved (Table 55). These values indicate that the strains are most probably members of the same species. Intragenus comparisons within the *S. pneumoniae* cluster had %BBH values ranging from 69.8% to 73.5% of their genomes shared (Table 55). These values show a high level of similarity between the organisms. Strains of *S. equi* and *S. pyogenes* had %BBH values ranging from 79.6% to 87.1% and 86.9% to 91.3% (Table 55). These values indicates that the strains are members of the same species. Intragenus comparisons between the organisms within the cluster revealed %BBH values ranging from 65.4% to 79.4% of the genomes conserved (Table 56). When the outliers were compared to both clusters, and the clusters were compared to each other, %BBH values ranged from 47.2% to 64.3% of their genomes conserved (Table 56). These values indicates that there is a distinct taxonomic separation between the two outliers and the two clusters as they do not share a significant portion of their genomes.

When the genomes of *S. pyogenes*, *S. equi*, *S. mitis*, *S. mutans*, and S. pneumonia were compared at the gene level the core genome consisted of 796 genes. Each organism had a significant number of unique coding sequences *S. pyogenes* contained 296 unique genes, *S. equi* contained 522 unique genes, *S. mitis* contained 545 unique

genes, *S. mutans* contained 562 unique genes, while B S. pneumonia contained 323 unique genes (Figure 14). Within the gene counts the same clustering pattern can be observed as that which was suggested by the ROSA values. *S. pyogenes* and *S. equi* shared an additional 140 genes with each other and an additional 85 genes with *S. mutans* (Figure 14). They shared a significantly fewer number of genes with *S. pneumoniae* and *S. mitis*, under 3- for each (Figure 14). These lower levels of orthology between the members of the two clusters indicate a separation at the genomic level. *S. mitis* and *S. pneumoniae* shared an additional 196 genes between each other and an additional 110 genes with *S. mutans* (Figure 14). These values indicate that the two species are highly related to each other. The fact that *S. mutans* had similar amounts of orthology with members of both clusters indicates that the organism is not related to either cluster at the genus level, but may be a member of the same family. The genes shared between *S. mitis* and *S. pneumoniae* primarily encoded for metal-related proteins and carbohydrate metabolism (Table 57). The genes shared between *S. pyogenes* and *S. equi* encode for synthesis of Streptolysin S, cell surface proteins, drug resistance, a captured phage genome, and a specialized ribonucleotide reductase (Table 55).

Based on the values it can be concluded that the *Streptococcus* genus consists of two distinct genera with two outliers that may each form their own genera. The first new genus would contain the organisms in the *S. pneumoniae* cluster, S. pneumonia, *S. mitis*, and *S. gordonii*, while the *S. pyogenes* cluster would contain *S. pyogenes* and *S. equi* (Table 52). This clustering pattern was repeated in all other manners of genomic

similarity. The 16s rRNA values within each cluster was above the proposed 94%-95% similarity between organisms. When the two clusters were compared between each other the 16s rRNA values were borderline of this suggested threshold ranging from 92.3% to 95% (Table 53). The values that do not directly fit into this threshold, the 95% value between *S. pyogenes* and *S. gordonii* could be due to the highly conserved nature of 16s rRNA sequences. Furthermore the threshold is a general value and may not be an exact definition as the proposed threshold is still evolving to reflect taxonomy. Similar values were obtained when the outliers were compared to other members of the genus.

AAI values supported the ROSA clustering as well. Organisms within the *S. pneumoniae* cluster showed AAI values ranging 74.4 to 93.4 (Table 54). Organisms in the *S. pyogenes* cluster showed AAI values ranging from 79.1 to 89.9 (Table 54). These two ranges showed considerable overlap in their values. Yet when compared to each other the two clusters had AAI values ranging from 64.0 to 70 (Table 54). These values show that the two clusters had a large amount of similarity within their cluster but had a significant drop in similarity when compared to each other. This supports the ROSA taxonomical classification. %BBH values showed distinct clustering as well. Members of the *S. pneumoniae* cluster had %BBH values ranging from 69.8% to 73.5% when compared within cluster, while members of the *S. pyogenes* cluster had %BBH values ranging from 65.4% to 79.4% when compared within the cluster. When the two clusters were compared to each other values ranged from 47.2% to 64.3 (Table 55). These values indicate that the two clusters had a large amount of similarity within their cluster but had a significant drop in similarity when compared to each other.

When the genomes of the organisms were compared at the gene level strains of the *S. pneumoniae* cluster shared more genes with other members of the *S. pneumoniae* cluster, while members of the *S. pyogenes* cluster shared more genes with other members of the *S. pyogenes* cluster. The outlier organisms shared a near equal amount of genes with members of both clusters (Figure 14). This gene pattern indicates that the organisms are forming their own unique distinct clusters that support the ROSA format. As such, based on the evidence provided it is recommended that the organisms be reclassified based on their ROSA values. Disease causing phenotype provides support for this classification as the *S. pyogenes* cluster contains the beta-hemolytic streptococci and the *S. pneumoniae* cluster contains the non beta-hemolytic streptococci (Facklam, 2002).

## F. *Vibrio*

| ROSA (sorted) | | 1219071 | 1224742 | 1224743 | 1219076 | 1219067 | 945543 | 1219077 | 1219061 | 675814 | 243277 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Vibrio harveyi NBRC 15634 = ATCC 14126 | **1219071** | | | | | | | | | | |
| Vibrio campbellii CAIM 519 = NBRC 15631 | **1224742** | 70.3 | | | | | | | | | |
| Vibrio rotiferianus CAIM 577 = LMG 21460 | **1224743** | 68.1 | 63.8 | | | | | | | | |
| Vibrio alginolyticus NBRC 15630 = ATCC 17749 | **1219076** | 61.8 | 60.1 | 58.2 | | | | | | | |
| Vibrio natriegens NBRC 15636 = ATCC 14048 = DSM 759 | **1219067** | 49.6 | 50.4 | 49.8 | 55.1 | | | | | | |
| Vibrio brasiliensis LMG 20546 | **945543** | 42.5 | 41.9 | 42.9 | 41.9 | 39.9 | | | | | |
| Vibrio azureus NBRC 104587 | **1219077** | 42.3 | 45.9 | 41.3 | 41.6 | 38.6 | 34.9 | | | | |
| Vibrio vulnificus NBRC 15645 | **1219061** | 41.4 | 39.8 | 41.6 | 40.4 | 36.5 | 38.3 | 31.4 | | | |
| Vibrio coralliilyticus ATCC BAA-450 | **675814** | 39.6 | 40.0 | 38.9 | 39.4 | 35.9 | 49.0 | 32.7 | 34.1 | | |
| Vibrio cholerae O1 biovar El Tor str. N16961 | **243277** | 37.0 | 38.2 | 37.8 | 38.2 | 36.2 | 39.5 | 33.1 | 37.4 | 37.0 | |
| Vibrio nigripulchritudo ATCC 27043 | **1051649** | 30.8 | 31.4 | 31.1 | 30.9 | 29.7 | 33.0 | 27.3 | 28.6 | 30.6 | 31.5 |

Table 57: ROSA values for members of the *Vibrio* genus. Intragenus comparisons were done between *V. harveyi*, *V. campbelli, V. rotiferanus, V. alginolyticus*, *V. brasiliensis*, *V. azureus, V. vulnificus*, *V. coralliilyticus*, *V. cholerae*, and *V. nigripulchritudo*.

Intragenus ROSA values for *V. harveyi*, *V. campbelli, V. rotiferianus*, *V. alginolyticus*, *V. brasiliensis*, *V. azureus, V. vulnificus*, *V. coralliilyticus*, *V. natriegens*, *V. cholerae*, and *V. nigripulchritudo*, ranged from 29.7 to 70.3 (Table 57). Within the ROSA

values one cluster and one outlier emerged. The outlier, *V. nigripulchritudo*, had ROSA values ranging from 27.3 to 31.1 when compared to other members of the genus (Table 57). A cluster containing *V. harveyi*, *V. campbelli,* V. rotiferanus, *V. alginolyticus*, *V. brasiliensis*, *V. azureus, V. vulnificus*, *V. coralliilyticus*, and *V. cholerae* had ROSA values ranging from 31.4 to 70.3 (Table 57). *V. harveyi*, *V. campbelli,* and *V. rotiferianus* formed a subcluster with ROSA values ranging from 63.8 to 70.3 (Table 57). ROSA values suggest *V. nigripulchuritudo* be split from the rest of the *Vibrio* genus to form its own genus.

| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Vibrio_harveyi_ATCC_14126T | 1 | | | | | | | | | | |
| Vibrio_campbellii_ATCC_BAA-1116T | 2 | 99.2 | | | | | | | | | |
| Vibrio_rotiferianus_LMG_21460T | 3 | 99.4 | 99.8 | | | | | | | | |
| Vibrio_alginolyticus_NBRC_15630T | 4 | 98.9 | 99.4 | 99.3 | | | | | | | |
| Vibrio_natriegens_ATCC_14048T | 5 | 94.4 | 94.9 | 94.8 | 95.3 | | | | | | |
| Vibrio_brasiliensis_LMG_20546T | 6 | 97.4 | 97.6 | 97.6 | 97.5 | 95.1 | | | | | |
| Vibrio_azureus_NBRC_104587T | 7 | 99.0 | 99.4 | 99.4 | 99.6 | 95.0 | 97.2 | | | | |
| Vibrio_vulnificus_NBRC_15645T | 8 | 96.0 | 96.4 | 96.3 | 96.4 | 96.1 | 96.0 | 96.6 | | | |
| Vibrio_coralliilyticus_ATCC_BAA-450T | 9 | 96.0 | 96.2 | 96.3 | 96.6 | 94.2 | 97.4 | 96.2 | 96.2 | | |
| Vibrio_cholerae_CECT_514T | 10 | 92.6 | 93.0 | 93.0 | 93.2 | 94.1 | 92.8 | 93.2 | 94.8 | 92.8 | |
| Vibrio_nigripulchritudo_ATCC_27043T | 11 | 97.0 | 97.3 | 97.3 | 97.1 | 95.3 | 97.6 | 97.2 | 96.0 | 96.2 | 93.0 |

Table 60: 16s rRNA % similarity values for members of the *Vibrio* genus.

Members of the *Vibrio* genus had a large amount of variation in their 16s rRNA % similarities. From the values a cluster could be determined that included *V. harveyi*, *V. campbelli, V. rotiferianus*, *V. alginolyticus*, *V. brasiliensis*, *V. azureus, V. vulnificus*, *V. coralliilyticus*, *V. natrigens*, and *V. nigripulchritudo* with % similarities ranging from 94.2% up to 99.8% similarity (Table 60). Within this cluster a subcluster formed between *V. harveyi*, *V. campbelli, V. alginolyticus*, and *V. azureus* whom displayed 16s rRNA similarities ranging from 98.9% to 99.8% (Table 60). *V. cholerae* had particularly low levels of 16s rRNA similarity when compared to other members of the genus, 16s rRNA similarity ranged from 92.6% to 94.8% similarity (Table 60).

| | Average Amino Acid Identity (AAIr) | 1219076 | 1219077 | 945543 | 1224742 | 243277 | 675814 | 1219071 | 1219067 | 1051649 | 1224743 | 1219061 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Vibrio alginolyticus NBRC 15630 = ATCC 17749 | **1219076** | | 79.5 | 75.6 | 87.2 | 73.3 | 74.5 | 87.3 | 86.1 | 70.1 | 86.5 | 76.2 |
| Vibrio azureus NBRC 104587 | **1219077** | 79.7 | | 74.3 | 80.5 | 71.7 | 72.2 | 80.5 | 79.4 | 69.9 | 80.4 | 72.4 |
| Vibrio brasiliensis LMG 20546 | **945543** | 75.5 | 74.1 | | 76.4 | 74.0 | 80.1 | 75.6 | 75.7 | 70.8 | 75.8 | 73.7 |
| Vibrio campbellii CAIM 519 = NBRC 15631 | **1224742** | 87.1 | 80.4 | 76.4 | | 73.6 | 74.9 | 93.5 | 84.8 | 70.3 | 90.3 | 76.1 |
| Vibrio cholerae O1 biovar El Tor str. N16961 | **243277** | 73.2 | 71.7 | 74.2 | 73.8 | | 73.1 | 72.8 | 73.3 | 70.3 | 73.5 | 73.0 |
| Vibrio coralliilyticus ATCC BAA-450 | **675814** | 74.7 | 72.4 | 80.5 | 75.1 | 73.2 | | 74.7 | 74.0 | 69.7 | 75.1 | 72.4 |
| Vibrio harveyi NBRC 15634 = ATCC 14126 | **1219071** | 87.2 | 80.8 | 75.9 | 93.6 | 72.9 | 74.9 | | 84.1 | 70.1 | 90.7 | 76.6 |
| Vibrio natriegens NBRC 15636 = ATCC 14048 = DSM 759 | **1219067** | 86.0 | 79.5 | 75.8 | 84.9 | 73.3 | 74.1 | 84.2 | | 70.1 | 83.7 | 75.6 |
| Vibrio nigripulchritudo ATCC 27043 | **1051649** | 69.9 | 69.7 | 70.8 | 70.4 | 70.3 | 69.4 | 70.1 | 69.8 | | 70.3 | 68.9 |
| Vibrio rotiferianus CAIM 577 = LMG 21460 | **1224743** | 86.7 | 80.2 | 76.0 | 90.4 | 73.6 | 74.8 | 90.8 | 83.5 | 70.4 | | 76.8 |
| Vibrio vulnificus NBRC 15645 | **1219061** | 76.1 | 72.9 | 73.9 | 76.4 | 73.1 | 72.4 | 76.7 | 75.6 | 69.1 | 76.8 | |

Table 61: AAI values for members of the *Vibrio* genus.

AAI values for members of the *Vibrio* genus showed great variation and the formation of a possible cluster. *V. harveyi*, *V. campbelli, V. alginolyticus, V. rotiferianus*, *and V. natrigens* with AAI vales ranging from 86.0 to 93.6 (Table 57). *V. azureus*, *V. brasiliensis*, *V. cholerae*, V. *coralliilyticus*, and *V. vulnificus* had AAI values ranging from 71.7 to 76.2 when compared to other members of the genus (Table 57). *V. nigripulchritudo* had significantly lower levels of AAI when compared to other members of its genus, with values ranging from 69.7 to 70.4 (Table 57).

| | Percent Bidirectional Best Hit (% BBH) | 1219076 | 1219077 | 945543 | 1224742 | 243277 | 675814 | 1219071 | 1219067 | 1051649 | 1224743 | 1219061 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Vibrio alginolyticus NBRC 15630 = ATCC 17749 | **1219076** | | 70.4 | 75.7 | 80.2 | 79.2 | 66.8 | 77.8 | 74.5 | 57.0 | 77.0 | 75.4 |
| Vibrio azureus NBRC 104587 | **1219077** | 61.1 | | 61.1 | 67.3 | 67.3 | 54.7 | 58.0 | 57.0 | 47.0 | 59.0 | 60.9 |
| Vibrio brasiliensis LMG 20546 | **945543** | 71.1 | 65.9 | | 70.7 | 77.9 | 69.4 | 68.7 | 67.4 | 57.7 | 71.4 | 74.1 |
| Vibrio campbellii CAIM 519 = NBRC 15631 | **1224742** | 78.0 | 74.7 | 73.1 | | 77.3 | 65.9 | 76.2 | 69.0 | 56.7 | 76.4 | 73.5 |
| Vibrio cholerae O1 biovar El Tor str. N16961 | **243277** | 63.2 | 61.3 | 66.0 | 63.2 | | 57.3 | 59.0 | 59.8 | 50.8 | 61.5 | 68.1 |
| Vibrio coralliilyticus ATCC BAA-450 | **675814** | 74.9 | 70.5 | 82.7 | 76.2 | 80.8 | | 72.2 | 69.4 | 60.9 | 72.5 | 74.3 |
| Vibrio harveyi NBRC 15634 = ATCC 14126 | **1219071** | 84.5 | 72.1 | 79.1 | 84.5 | 80.5 | 69.3 | | 73.0 | 59.4 | 85.3 | 79.4 |
| Vibrio natriegens NBRC 15636 = ATCC 14048 = DSM 759 | **1219067** | 74.5 | 65.3 | 71.8 | 71.0 | 75.0 | 61.4 | 67.1 | | 54.9 | 70.5 | 69.2 |
| Vibrio nigripulchritudo ATCC 27043 | **1051649** | 68.9 | 65.3 | 74.1 | 70.0 | 76.7 | 65.5 | 65.7 | 66.6 | | 68.0 | 70.9 |
| Vibrio rotiferianus CAIM 577 = LMG 21460 | **1224743** | 78.4 | 69.2 | 77.5 | 80.0 | 78.5 | 66.0 | 80.0 | 72.2 | 57.5 | | 77.2 |
| Vibrio vulnificus NBRC 15645 | **1219061** | 63.8 | 58.2 | 66.7 | 63.5 | 72.2 | 55.9 | 61.7 | 58.5 | 49.3 | 63.8 | |

Table 60: %BBH values for members of the *Vibrio* Genus.

Orthology between members of the *Vibrio* genus showed a large amount of variation. A single cluster formed between *V. harveyi, V. alginolyticus*, *V. rotiferanus*, and *V. campbellii*, which shared 75.7% to 85.3% of their genome (Table 60). *V. azureus*, *V. brasiliensis*, *V. choleraee*, V. *coralliilyticus, V. natrigens*, and *V. vulnificus* shared between 55.9% to 76.4 % of their genomes with other members of the genus (Table 60). *V. nigripulchritudo* had low levels of orthology with other members of its genus with values ranging from 47% to 57.7% of the genomes conserved (Table 60).

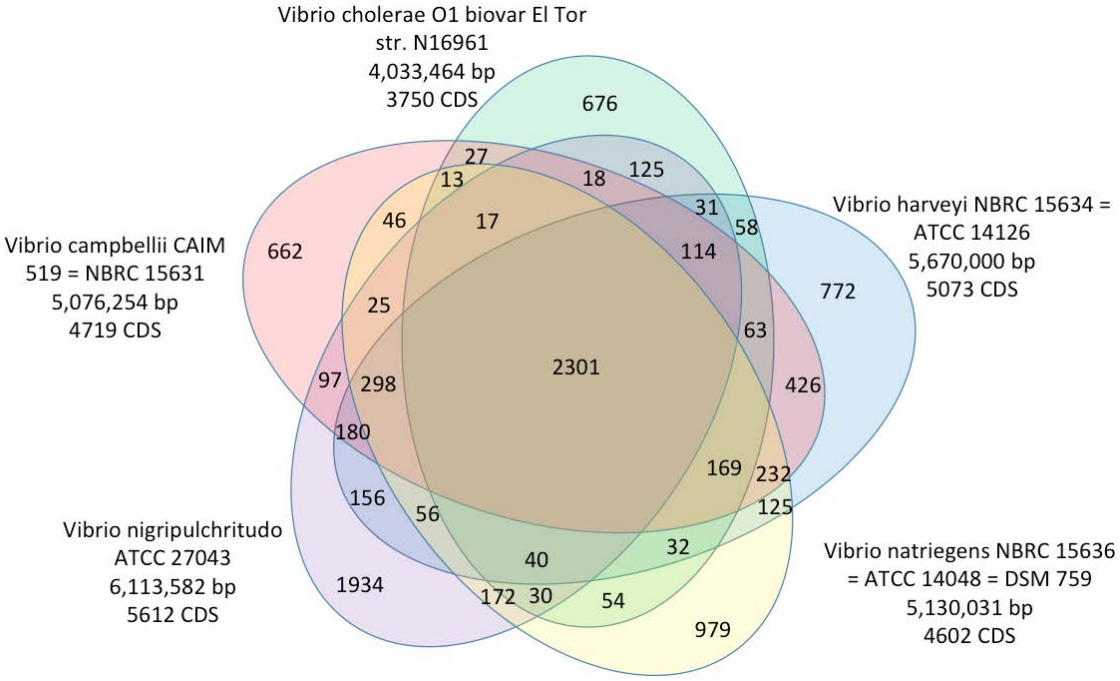


Figure 14: **Genomic similarity at the gene level for *V. nigripulchritudo, V. campbellii*, *V. choleraee*, *V. harveyi*, *and V. natrigens*.** Venn-diagram shows the number of orthologous genes shared within each possible comparison combination.

The core genome of *V. nigripulchritudo, V. campbellii*, *V. choleraee*, *V. harveyi*, *and V. natrigens* consisted of 2301 genes. *V. nigripulchritudo* had 1934 unique genes with a genome size of 6,113,582 bp and 5612 CDS. *V. campbellii* had 662 unique genes with a genome size of 5,076,254 bp and 4719 CDS. *V. harveyi* had 772 unique genes with

a genome size of 5,670,000 bp and 5073 CDS S. cholerae had 676 unique genes with a genome size 4,033,46 bp and 3750 CDS. V. natrigens had 979 unique genes with a genome size of 5,130,031 bp and 4602 CDS (Table 14). *V. campbellii* and *V. harveyi* shared a further 426 genes (Table 14). *V. nigripulchritudo, V. campbellii*, *V. harveyi*, *and V. natrigens* shared a further 298 genes (Table 14). *V. nigripulchritudo, V. campbellii*, *V. choleraee*, and *V. harveyi* shared a further 114 genes (Table 14). *V. natrigens* and *V. nigripulchritudo* shared a further 172 genes (Table 14). *V. nigripulchritudo* had a large number of genes unique to it. These genes encoded for a large amount of polysaccharide metabolism proteins, a secreted proteases, metal utilization and resistance proteins, phosphonate metabolism, resistance proteins, chemotaxis proteins, transcriptional regulators and possible anaerobic genes (Table 61).

| | Only in V. nigripulchritudo. | | |
|---|---|---|---|
| | | | |
| 445 | Methyl-accepting chemotaxis protein | | |
| 446 | Methyl-accepting chemotaxis protein | 2957 | RNA polymerase sigma-70 factor |
| 678 | Acriflavin resistance protein | 2958 | hypothetical protein |
| 679 | Acriflavin resistance protein | 2959 | Beta-lactamase class C and other penicillin binding proteins |
| | | 2960 | hypothetical protein |
| 1256 | Phosphonate ABC transporter ATP-binding protein (TC 3.A.1.9.1) | 2961 | Prolyl endopeptidase (EC 3.4.21.26) |
| 1257 | Phosphonate ABC transporter phosphate-binding periplasmic component (TC 3.A.1.9.1) | 2962 | 4-Hydroxy-2-oxoglutarate aldolase (EC 4.1.3.16) |
| 1258 | Phosphonate ABC transporter permease protein phnE (TC 3.A.1.9.1) | 2963 | Phosphogluconate repressor HexR, RpiR family |
| 1259 | Transcriptional regulator PhnF | 2964 | hypothetical protein |
| 1260 | PhnG protein | 2965 | D-3-phosphoglycerate dehydrogenase (EC 1.1.1.95) |
| 1261 | PhnH protein | 2966 | FIG01074520: hypothetical protein |
| 1262 | PhnI protein | 2967 | Glucose-6-phosphate isomerase, archaeal (EC 5.3.1.9) |
| 1263 | PhnJ protein | 2968 | Dihydroxyacetone ABC transport system, permease protein |
| 1264 | Phosphonates transport ATP-binding protein PhnK | 2969 | ABC transport system, sugar-binding protein |
| 1265 | Phosphonates transport ATP-binding protein PhnL | 2970 | glycosyl transferase, family 2 |
| 1266 | Metal-dependent hydrolase involved in phosphonate metabolism | 2971 | Peptidyl-tRNA hydrolase, archaeal type (EC 3.1.1.29) |
| 1267 | ATP-binding protein PhnN; Guanylate kinase (EC 2.7.4.8) | 2972 | Beta-lactamase class C and other penicillin binding proteins |
| 1268 | Metal-dependent hydrolases of the beta-lactamase superfamily I; PhnP protein | 2973 | L-ribulose-5-phosphate 4-epimerase (EC 5.1.3.4) |
| | | 2974 | L-xylulose 5-phosphate 3-epimerase (EC 5.1.3.-) homolog |
| 1998 | 2-deoxy-D-gluconate 3-dehydrogenase (EC 1.1.1.125) | 2975 | L-xylulose/3-keto-L-gulonate kinase (EC 2.7.1.-) |
| 1999 | 4-deoxy-L-threo-5-hexosulose-uronate ketol-isomerase (EC 5.3.1.17) | 2976 | 6-phosphogluconolactonase (EC 3.1.1.31) |
| 2000 | TRAP-type C4-dicarboxylate transport system, large permease component | 2977 | putative sugar transport system permease ABC transporter protein |
| 2001 | TRAP-type C4-dicarboxylate transport system, small permease component | 2978 | Various polyols ABC transporter, periplasmic substrate-binding protein |
| 2002 | TRAP-type transport system, periplasmic component, predicted N-acetylneuraminate-binding protein | 2979 | Glycerol-3-phosphate ABC transporter, ATP-binding protein UgpC (TC 3.A.1.1.3) |
| 2003 | Predicted D-glucarate or D-galactorate regulator, GntR family | 2980 | Transcriptional regulator, GntR family |
| 2004 | hypothetical protein | | |
| 2005 | Rhamnogalacturonides degradation protein RhiN | 3345 | Secreted trypsin-like serine protease |
| | | 3346 | Secreted trypsin-like serine protease |
| 2656 | Glutathione S-transferase | 3347 | Secreted trypsin-like serine protease |
| 2657 | 3-oxoacyl-[acyl-carrier protein] reductase (EC 1.1.1.100) | | |
| 2658 | FIG01075954: hypothetical protein | 3875 | Phosphonate ABC transporter ATP-binding protein (TC 3.A.1.9.1) |
| 2659 | Maltose/maltodextrin transport ATP-binding protein MalK (EC 3.6.3.19) | 3876 | Phosphonate ABC transporter phosphate-binding periplasmic component (TC 3.A.1.9.1) |
| 2660 | Maltose/maltodextrin ABC transporter, permease protein MalG | 3877 | Phosphonate ABC transporter permease protein phnE (TC 3.A.1.9.1) |
| 2661 | ABC-type sugar transport systems, permease components | 3878 | Transcriptional regulator PhnF |
| 2662 | ABC transporter, periplasmic substrate-binding protein | 3879 | PhnG protein |
| 2663 | hypothetical protein | 3880 | PhnH protein |
| 2664 | Aldo-keto reductase | 3881 | PhnI protein |
| 2665 | Sorbitol dehydrogenase (EC 1.1.1.14) | 3882 | PhnJ protein |
| 2666 | Choline dehydrogenase (EC 1.1.99.1) | 3883 | Phosphonates transport ATP-binding protein PhnK |
| 2667 | Sugar phosphate isomerases/epimerase | 3884 | Phosphonates transport ATP-binding protein PhnL |
| 2668 | hypothetical protein | 3885 | Metal-dependent hydrolase involved in phosphonate metabolism |
| 2669 | hypothetical protein | 3886 | ATP-binding protein PhnN; Guanylate kinase (EC 2.7.4.8) |
| 2670 | Rrf2 family transcriptional regulator, group III | 3887 | Metal-dependent hydrolases of the beta-lactamase superfamily I; PhnP protein |

Table 61: Selected portion of genes unique to *V. nigripulchritudo*.

Vibrio Discussion

The *Vibrio* genus was established in 1854 with the type species *Vibrio* cholerae being described in the same year (Parte 2014). Currently there are approximately 111 validly published species grouped within 19 clades based on MLSA with 8 housekeeping genes (Sawabe et al., 2013). Species from this genus are typically gram negative motile rods with polar flagellum that are enclosed within a sheath, halophilic, mesophilic, chemo-organotrophic, and are facultatively fermentive, able to ferment D-Glucose, D-fructose, maltose and glycerol (Baumann et al., 1984, Gomez-Gil et al., 2014). Organisms can typically reduce nitrate to nitrite and require Na+ . Many organisms within the genus are isolated from aquatic environments both as pathogens, as is the case for V. cholerae, and as free living organisms. With the rapid increase in known members of the *vibrio* genus, organisms that lack one or more of the textbook properties are being described (Gomez-Gil et al., 2014).

ROSA values for the *Vibrio* genus revealed two distinct clusters. The *V. harveyi* cluster contained *V. harveyi*, *V. campbelli, V. rotiferianus, V. alginolyticus*, V. brasiliensis, *V. azureus,* V. vulnificus, V. coralliilyticus, *and V. natrigens*, while the second consisted of only a single organism, V. nigripulchritudo. ROSA values within the *V. harveyi* cluster ranged from 31.4 to 70.3 (Table 57). These values indicate that the members of the cluster all belong to the same genus as different species, except for *V. harveyi*, *V. campbelli, and V. rotiferianus* who formed their own subcluster with values ranging from 63.8 to 70.3, just above and below the species threshold for ROSA (Table 57). This

indicates that the three organisms may be members of the same species, however the values are just at the threshold and need more evidence for either direction. Values between the cluster and V. nigripulchuritudo ranged from 27.3 to 31.1, firmly in the same family different genus range.

16s rRNA values supported a slightly different clustering. The values formed a number of unique clusters within the genus. First *V. harveyi*, *V. campbelli,* V. rotiferanus*, V. alginolyticus*, and *V. azureus* all clustered with 16s rRNA values <98.5%, the species threshold, indicating that they are all the same species by 16s rRNA*. V. natrigens*, V. vulnificus, V. coralliilyticus, and V. nigripulchuritudo had 16s values ranging from 94.4 to 97.4% indicating that they are members of the same genus as *V. harveyi*, *V. campbelli,* V. rotiferanus*, V. alginolyticus*, and *V. azureus* . V. cholera clustered by itself with 16s rRNA values ranging form 92.6% to 94.1% (Table 60). These values indicate that of all the organisms, the type, V. cholerae was the only one not in the genus by 16s rRNA similarity. This cluster showcases one of the weaknesses of 16s rRNA. The method only looks at one highly conserved gene, and as such may fail to fully capture the relatedness of two genomes.

AAI values support the original clustering given by the ROSA values. Members of the *V. harveyi* cluster had AAI values ranging from 71.7 to 93.6 (Table 57). When *V. harveyi*, *V. campbelli, and V. rotiferianus* were compared to each other AAI values ranged from 90.7 to 93.6 indicating that the organisms share a higher level of similarity than other members of the *Vibrio* genus but not quite enough to be above the species threshold for AAI (Table 57). When V. nigripulchuritudo was compared to other

members of the genus it showed a great amount of difference with AAI values ranging form 69.7 to 70.4 (Table 57). These values indicate that the rest of the genus belongs to a single genus that V. nigripulchuritudo does not belong to, supporting the original ROSA value clustering.

%BBH values showed a different clustering pattern as AAI values. A cluster formed between *V. harveyi, V. alginolyticus*, V. rotiferanus, *and V. campbellii*, which shared 75.7% to 85.3% of their genome (Table 60). These values indicate that these organisms show a high level of similarity between each other*. V. azureus*, V. brasiliensis, V. cholerae, V. coralliilyticus*, V. natrigens*, and V. vulnificus shared between 55.9% to 76.4 % of their genomes with other members of the genus (Table 60). Although these values are relatively low the lower values primarily come from *V. azureus* . V. nigripulchuritudo had low levels of orthology with other members of its genus with values ranging from 47% to 57.7% of the genomes conserved (Table 60). These low values indicate that V. nigripulchuritudo is extremely different from other members of the cluster and genus.

When the genomes of V. nigripulchuritudo*, V. campbellii*, V. cholerae, *V. harveyi*, *and V. natrigens* were compared at the gene level the core genome consisted of 2301 genes. Each organism had a significant number of unique coding sequences. V. nigripulchuritudo contained 1934 unique genes, V. cholerae contained 671 unique genes, *V. harveyi* contained 772 unique genes*, V. natrigens* contained 979 unique genes, while*V. campbellii* contained 662 unique genes (Figure 17)*. V. campbellii* and *V. harveyi* shared a further 426 genes (Figure 17). This indicates that the organisms are related. V.

nigripulchuritudo had over 1934 genes unique to it (Figure 17). These genes encoded for a large amount of polysaccharide metabolism proteins, a secreted proteases, metal utilization and resistance proteins, phosphonate metabolism, resistance proteins, chemotaxis proteins, transcriptional regulators and possible anaerobic genes (Table 60). This large amount of unique genes and relatively low amount of shared genes indicate that V. nigripulchuritudo is highly diverged from the rest of the genus.

Based on the data above a reclassification is deemed necessary for the *Vibrio* genus. It is suggested that the genus be split into two, one genus containing *V. harveyi*, *V. campbelli, V. rotiferianus, V. alginolyticus*, V. brasiliensis, *V. azureus,* V. vulnificus, V. coralliilyticus, *and V. natrigens*, and the other containing V. nigripulchuritudo. This is done on the basis of ROSA values, AAI values, %BBH values, 16s rRNA values, and gene level comparisons. ROSA values for the group showed distinct clustering at the genus level for all members of the genus except V. nigripulchuritudo, who clustered at the family level. For 16s rRNA values clustering was observed between all members except V. cholerae, however as mentioned 16s rRNA is looking at only 1 highly conserved genes and therefore can not be used by itself to resolve taxonomy. AAI values showed intense clustering between the organisms, particularly *V. harveyi, V. campbellii*, and V. rotiferanus. All members of the cluster had AAI values exceeding 71 while V. nigripulchuritudo had AAI values below 71, showing a distinct genomic difference between the two groups. Similarly the %BBH values of *V. harveyi, V. campbellii*, and V. rotiferanus showed higher levels of orthology compared to the rest of the cluster while the cluster overall ranged from 55.9% to 85.3% of the genomes conserved. When

compared to V. nigripulchuritudo %BBH ranged from 47% to 57.7%, showing a distinct difference when compared to the other cluster. Gene counts show the same story with members of the cluster sharing a strong percentage of their genes with distinct metabolic pathways. V. nigripulchuritudo shared a small percentage of its genome and retained 1934 genes unique to it with diverse metabolic capabilities. It is based on this evidence that it is recommended that the *Vibrio* genus be split into two new genera, the first containing solely V. nigripulchuritudo. The second genus will contain *V. harveyi*, *V. campbelli, V. rotiferianus, V. alginolyticus*, V. brasiliensis, *V. azureus,* V. vulnificus, V. coralliilyticus, *and V. natrigens*. Within this genus it is recommended that *V. harveyi*, *V. campbelli, and V. rotiferianus* are combined into a single species. A recent paper by Goudenege et al in 2013 found that V. nigripulchuritudo strains had little diversity among their genomes and were only distantly related to other *Vibrio* species (Goudenege et al. 2013). This provides literature support for the ROSA classification for V. nigripulchuritudo.