Genomic Analysis And Classification Of Novel Species *Flavobacterium gabrieli* KJJ

Presented to the Faculty of Lycoming College in partial fulfillment of the requirements for Departmental Honors in Biology

by
Kirsten Fischer
Lycoming College
April 27, 2016

Approved by:

Dr. Jeffrey Newman

Dr. David Andrew

Dr. Jeremy Ramsey

Dr. Cullen Chandler

Introduction

Microbial Systematics and Taxonomy

The diversity of bacteria is truly immense and the discovery of new species and higher taxonomic groups happens quite frequently, as evidenced by the ever expanding tree of life (Hug *et al*., 2016). The classification of prokaryotes, bacteria especially, is formally regulated by the International Committee on the Systematics of Prokaryotes and has experienced rapid change over the last fifty years. However, some feel that these rules could be even stricter for proper organization of taxonomy (Tindall *et al.*, 2010). Problems occur with the integration of newer methodologies, which creates some challenges for the researcher attempting to publish a novel species. For example, some DNA sequences that are deposited in databases are not accurate (Clarridge, 2004). Taxonomy is an artificial system that works based on the intuition of scientists rather than strict, specific standards (Konstantinidis & Tiedje, 2005). Tindall advocates that a strain shown to be a novel taxon should be characterized "as comprehensively as possible" and abide by the framework established in the Bacteriological Code (2010). The Bacteriological Code was published by the American Society for Microbiology and the present revision is from 1992 (Lapage *et al.*, 1992). This code outlines rules for classifying bacteria and naming newly identified organisms before publication in the *International Journal of Systematic Bacteriology*. Some of the guidelines given within the code include citation of authors and names, naming of new taxa, and circumstances where a name or epithet can become illegitimate and replaced (Lapage *et al.*, 1992). The Bacteriological Code, although a bit outdated since the advent of genomic analysis, anticipates the ever-changing environment of microbial identification.

The International Committee on the Systematics of Prokaryotes dictates the regulations and necessary information for publishing new bacteria species in the *International Journal of Systematic and Evolutionary Microbiology*, which was previously titled the *International Journal of Systematic Bacteriology.* These guidelines, for which the Bacteriological Code is still the key text, mandate that certain phenotypic data, the 16S rRNA gene sequence, and culture deposits in at least two publically accessible culture collections be submitted before publication (Tindall *et al.*, 2010).

The characterization of a novel species is polyphasic. There is the traditional phenotypic component, where physical traits such as colony morphology and fatty acid profiles are examined. Additionally, the level of genetic uniqueness of the organism should be established by studying the 16S rRNA gene as well as the entire genome. The preliminary characterization of a new species is achieved via comparison of the 16S rRNA gene sequence. The widespread use of 16S rRNA sequences in bacterial identification is due to the fact that it is an essential, ubiquitous gene and it evolves slowly due to its interaction with multiple proteins. The 16S rRNA gene is relatively small, but large enough for bioinformatics purposes at 1,500 base pairs (Janda & Abbott, 2007). The widespread and prominent use of the 16S rRNA gene has made it the gene of choice when making genetic comparisons between taxa using sequence similarity (Clarridge, 2004). Prior to 2006, 97% 16S rRNA sequence identity was considered the threshold below which an organism would be considered novel. The 98.5% species threshold for 16S similarity was identified by correlation with DNA-DNA Hybridization (Stackebrandt & Ebers, 2006). All organisms in this study that had less than 98.5% 16S rRNA similarity also had less than 70% DDH. This is the species threshold for DDH, which is a more traditional measure of

genetic similarity. Despite the evidence presented by Stackebrandt and Ebers in 2006, the guidelines for publication of novel species (Tindall et al., 2010) specifies that when the 16S rRNA sequence similarity is above 97%, a second measure of genomic uniqueness is required.

16S rRNA has continuously proven to be reliable for genus-level identification, but there is no set genus threshold (Janda & Abbott, 2007). Additionally, different species can have nearly identical 16S rRNA similarity, but very low DDH (Stackebrandt & Ebers, 2006). Therefore, it is highly suggested that other genes with greater resolution, particularly protein-encoding genes, be analyzed in addition to the 16S rRNA gene. Analysis of several protein coding genes is often referred to as Multi-Locus Sequence Analysis (MLSA). When compared to DDH, 16S rRNA gene sequence often lacked the same resolution (Adékambi *et al.*, 2008). Resolution of a measure is dictated by its range. 16S similarity of different species in a genus range between 95%-100%, while DDH's range is from 0%-100% giving it a broader range and better resolution.

Phylogenomics

Since the publication of the first completed genome, that of *Haemophilus influenzae* in 1995, the amount of genetic information available to researchers has exploded (Fleischmann *et al.*, 1995). Technological advances over the past decade have dramatically decreased the cost of DNA sequencing. At less than $200 to sequence a genome, it is now less expensive to sequence two genomes and use computational methods such as Genome-Genome Distance Calculator (GGDC) (Meier-Kolthoff et al., 2013) to estimate DDH than it is to use a contract lab to perform the experiment. The genome sequence can also be used to precisely determine the GC percentage, which is required for publication, as well as to identify genes present in the organism.

There are several different tools currently in use for studying genomic-level uniqueness. The phylogenomic metrics include estimated DNA-DNA Hybridization (eDDH), Average Nucleotide Identity (ANI), and Average Amino Acid Identity (AAI). DDH must be less than 70% for a novel species (Adékambi *et al.*, 2008). It is important to note that DDH is not measuring the direct sequence identity but just the efficiency of hybridization (Konstantinidis, Ramette, & Tiedje, 2006). DDH is the traditional method for genetic comparison and is the current standard in bacterial classification (Goris *et al.,* 2007). Physical measurements of DDH are typically done by heating double-stranded DNA from each organism to dissociate it and then lowering the temperature to measure the level of annealing of matching base pairs from each strand (Auch *et al.*, 2010). These experiments are time consuming, not particularly reproducible and require specialized equipment or can be contracted out with a cost of several hundred dollars (Goris *et al.,* 2007). Estimated DDH values can instead be calculated using whole genome sequences analysis (Meier-Kolthoff *et al.*, 2013).

AAI and ANI measure sequence similarity, whereas DDH measures hybridization. ANI is a pairwise genomic comparison that is becoming more prominent (Konstantinidis *et al.*, 2006). ANI was also found to correlate well with *rpoB* gene sequences (Adékambi *et al.*, 2008). However, ANI cannot detect homology below 60% identity due to a threshold level of similarity for nucleotide sequences. Therefore, distantly related species cannot be compared using ANI (Konstantinidis & Tiedje, 2005). Average Nucleotide Identity has a range from about 70%-100% identity, and so is unable to detect homology if organisms are too distantly related. Average Amino Acid Identity has a range of 30%-100% identity and can detect homology at a lower percent identity due to the fact there are more types of

amino acids than nucleotides, providing greater statistical variability to detect at lower similarities. Amino acid sequence is more highly conserved due to the degeneracy of the genetic code and selection at the level of the protein. Additionally, there are twenty possible amino acids while there are only four possible nucleotides, so this also accounts for the difference in range between AAI and ANI. Based of correlation with DDH, The threshold for AAI and ANI is <95% for a new species (Konstantinidis *et al.*, 2006). AAI and ANI describe the percent similarity of orthologous genes and proteins respectively. The percent of genes that are Bidirectional Best Hits (BBH) is another method of measuring pairwise genome sequence similarity. Genes that are bidirectional best hits are the best matches of each other when comparing two organisms' genomes. BBH can serve as a strong indication of orthologous genes (Wolf & Koonin, 2012). Orthologous genes descended from a common ancestor and retain the same function in different organisms. The Reciprocal Orthology Score Average (ROSA), a phylogenomic metric, is an unpublished tool that was developed in the Newman lab. It utilizes BBH and AAI to give an average of the percent of genes in the genome that are shared as well as the percent similarity of those shared genes. Whereas AAI and ANI only reveal the percent similarity of shared genes. No phylogenomic tool in use thus far gives as accurate a picture of global orthology and genome similarity as ROSA. Based on comparisons of bacteria with well-studied genomes, a threshold of 65 was determined through comparison to DDH and AAI values of various species.

Additionally, pairwise similarity alone is not sufficient in describing a new species, so it should be used in conjunction with other comparison methods like orthology scores as well as phenotypic data (Tindall *et al.*, 2010). In order to fully understand and appreciate

the ecological value and evolutionary niche of each bacterial species, the entire genome must be well studied (Stahl & Tiedje, 2001). Publication in the *International Journal of Systematic and Evolutionary Microbiology* only requires characterization of the 16S rRNA gene. Due to the decreasing costs and increasing availability of entire genome sequences, there is a push amongst bacterial taxonomists to change this standard (Konstantinidis & Tiedje, 2005). Nonetheless, if an organism's 16S rRNA similarity is below 98.5% to other published species, there can be confidence that it is a novel species. A cut off of 98.7% pairwise similarity was later identified as the threshold for new species classification based on a correlation study with average nucleotide identity (Kim *et al.*, 2014).

Previous work

Every December, Dr. Newman collects a water sample from the Loyalsock Creek outside of Montoursville. The water samples are spread onto Petri plates and subsequent colonies are cultured for further study. These organisms' physical traits are studied by students in the Microbiology class. Kathy Jacobs first identified *Flavobacterium gabrieli* KJJ in 2012 and Ashley Gimbel identified *Flavobacterium douthatii* ABG in 2013. The Microbiology students characterize the organisms' morphology, preferred growth conditions, various metabolic capabilities, fatty acid profiles, and antibiotic sensitivities. They also begin the initial genetic characterization of the organisms by studying the 16S rRNA gene sequence via PCR amplification and Sanger sequencing. If the 16S rRNA gene sequence is different enough (<98.7% similar) from the sequences of previously published species, then the organism is considered potentially novel. These organisms will have their entire genomes sequenced with Next Generation sequencing. NextGen genome sequencing is made accessible through the Genome Consortium for Active Teaching Using Next-Generation Sequencing (GCAT-SEEK), which provides genetic research opportunities for

undergraduates (Buonaccorsi *et al.*, 2014). The genome is initially assembled by the software NextGENe V2.3.4.2 (SoftGenetics, State College, PA). This assembled genome then undergoes manual assembly and editing to remove low quality sequences.

The assembled genome was then uploaded to RAST for annotation (Aziz *et al.*, 2008). Gene annotation is a process by which the software identifies coding sequences and predicts the function of proteins based on similarity to a curated list. Proteins are grouped into pathways referred to as subsystems (Aziz *et al.*, 2008). These annotations were accessed using the SEED Viewer interface which categorizes these annotations. Annotated genomes are particularly useful in correlating the presence of genes to physical traits. Observation of a certain phenotype can sometimes be tied to genes in an organism's genome by exploring the annotated functions as well as using BBH to determine orthologous genes. The annotations are also used to create a Venn diagram, which shows how many genes are unique to each organism as well as those that are shared between the different organisms.

Flavobacteria species are frequently studied in the Newman lab due to their prevalence in freshwater. They are also a highly diverse and increasingly studied genus, particularly for their importance in freshwater biology and fish health. For example, a survey of a hard water creek in the German mountains yielded a diverse population of Flavobacteria (Brambilla *et al.,* 2007). Flavobacteria tend to be found in freshwater environments, which makes it unsurprising *Flavobacterium gabrieli* KJJ and *Flavobacterium douthatii* ABG were isolated from the Loyalsock Creek. Brambilla and her colleagues hypothesized that the bacteria are moved to the creek as a result of leaching from soil and plant roots. Additionally, the trend of finding a diverse population of Flavobacteria in fresh

water is echoed in another study surveying the Great Lakes (Loch & Faisal, 2014). They not only found a large, diverse population of *Flavobacterium* species, but 65 species that were fish-associated and some of which caused lesions. Loch and Faisal hypothesized, based on their diverse sample, that there is a growing number of Flavobacteria that are capable of infecting fish and they are becoming more heterogeneous through evolution (2014). Flavobacteria as fish pathogens have important implications not only ecologically in freshwater bodies, but in commercial fishing and hatcheries as well. Interestingly, some species of Flavobacteria normally present on a fish can also act as opportunistic pathogens after initial onset of disease (Good *et al.,* 2015). Good and her colleagues found that *Flavobacterium branchiophilum* was the causative agent of bacterial gill disease in aquaculture rainbow trout (2015). They also found high concentrations of *Flavobacterium succinicans* in diseased fish as well. The question arises as to why some species of Flavobacteria have fish pathogenicity while some do not. Additionally, what genes or proteins make one species opportunistic and the other able to initiate disease? Genes related to virulence, toxin production, and cell wall composition could all be explored for possible explanations. The correlation of physical traits to the presence of genotypes is the goal of phylogenomics.

Summary of study

In the fall semester of 2015, *Flavobacterium douthatii* ABG was studied during the Research Methods class. This study picked up where the work of Ashley Gimbel had left off. The identification of the 16S rRNA sequence and phenotypic characterization were already completed. ABG's genome was already sequenced and annotated in RAST. This study instead focused on characterizing the genome. First, comparison organisms were

chosen based on the 16S rRNA phylogenetic tree (*Flavobacterium succinicans, Flavobacterium glaciei,* and *Flavobacterium hydatis*). Genomic uniqueness was then calculated using DDH, AAI, and ROSA compared to the reference organisms. A Venn diagram was created to show genes unique and shared between ABG, *F. succinicans*, *F. glaciei, F. hydatis*, and *Flavobacterium aquatile*, which is the type species for the genus and must be in all genomic comparisons. After all of the comparisons were completed a problem with the assembly of ABG's genome was discovered. This was during the spring semester of 2016 for a biology departmental honors project. In order to save time, research efforts were instead focused on an organism with a better genome assembly. The bulk of this Honors project was spent examining KJJ. As was the case for ABG, a majority of the phenotypic work, the 16S rRNA sequence, and the assembled and annotated genome were already complete. Reference organisms were determined this time based on both the 16S rRNA phylogenetic tree as well as phylogenomic comparisons such as AAI and DDH to all available *Flavobacteria* genomes. The reference species chosen were *Flavobacterium chilense*, *Flavobacterium hibernum*, and *Flavobacterium denitrificans*. A Venn diagram was also created to determine the unique and shared genes between the reference species as well as *F. aquatile*. The phenotypic data from past work was also compared to the annotated genome to find a genetic basis to explain some differences in metabolic traits between the reference species and KJJ. Data figures such as the phylogenetic tree, a 16S matrix, the phylogenomic matrices, the Venn diagram, the list of unique genes, Fatty Acid Methyl Ester (FAME) profiles, and Biolog metabolic data were prepared in the proper format for publication in the *International Journal of Systematic and Evolutionary Microbiology.* The manuscript for publication is currently in progress for submission.

Materials and Methods

Initial Isolation and 16S rRNA Characterization

Flavobacterium douthatii ABG, *Flavobacterium gabrieli* KJJ, and their respective reference species were recovered from frozen permanents at -80^{0}C and plated on Trypticase Soy Agar (TSA) and Reasoner's 2A agar (R2A). These reference organisms were determined based upon a pairwise similarity of the 16S rRNA gene run in EZTaxon's Identify tool (Kim *et al.*, 2012). All 16S rRNA pairwise similarities were below the species threshold of 97%. The closest relatives gleaned from this analysis were characterized and had their genomes analyzed alongside ABG and KJJ. A 16S rRNA neighbor-joining phylogenetic tree was generated using Molecular Evolutionary Genetics Analysis 6 (MEGA6) software (Tamura *et al.*, 2013).

Polymerase Chain Reaction

In order to confirm the identity of the physical cultures with the 16S rRNA gene sequences deposited in GenBank, PCR and Sanger sequencing were used. Single colonies taken from these plates were inoculated into water, underwent freeze-thaw cycles to lyse the cells, and were added to a mix of *Taq* polymerase (GeneMate, BioExpress, Kaysville, UT). They were then run in the thermocycler for Polymerase Chain Reaction (PCR) to amplify the 16S rRNA gene. The program was 35 cycles of heating at 95^{0}C for three minutes to promote denaturation of DNA then after 30 seconds of additional heat, dropped to 50^{0}C for 30 seconds where the primers pair up with the genetic material, lastly each cycle ended with one minute at 72^{0}C where the DNA polymerase actually copied the DNA. The amplicon then underwent gel electrophoresis to estimate the size and concentration of the amplified PCR product. Ethidium bromide was used as the dye to stain DNA. Distance

of migration measured length: the 16S rRNA gene is about 1,500 base pairs. The brightness of the bands indicated the approximate concentration of genetic material in the PCR product. We used this estimated concentration to determine how much genetic material must be diluted to meet the requirements for the third-party lab that performed our Sanger sequencing (Beckman Coulter Genomics, Danvers, MA) which requires 20µL with a concentration of 20ng/µL.

Genomic-Based Comparisons

The genomic libraries for ABG and KJJ was prepared and sequenced on an Illumina MiSeq (V3 2 x 300 base) (Illumina, San Diego, CA) by the Indiana University Center for Genomic Studies as a part of a Genome Consortium for Active Teaching NextGen Sequencing Group (GCAT-SEEK) shared run (Buonaccorsi *et al.*, 2011, 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, State College, PA). The returned sequence was also manually edited and assembled by a Newman lab student to eliminate low quality sequences. The assembled genome was then uploaded to RAST for annotation (Aziz *et al.*, 2008). The assembled genome of ABG was 5,270,010 base pairs long with an average 90x coverage and was composed of 89 contigs. RAST determined it was made up of 4,174 protein-coding sequences that were annotated for their functions. The assembled genome of KJJ was 4,612,888 base pairs long with an average 73x coverage and was composed of 11 contigs. RAST determined it was made up of 4,016 protein-coding sequences. Subsystems were described in RAST based on the identification of possible protein coding gene sequences, the predicted proteins these

genes may be responsible for, and the proteins' ultimate function. These annotations were accessed using the SEED Viewer interface which categorizes these annotations.

RAST Tab Separated Value (*.tsv) files containing Bidirectional Best Hit comparisons for gene orthology were then exported to a program developed in the Newman lab to calculate Average Amino Acid Identity (AAI), which is considered a reliable measure of genome-level sequence identity (Konstantinidis & Tiedje, 2005). The species threshold for AAI is 95%. Average Nucleotide Identity (ANI) was calculated using the EZ Bio Cloud Average Nucleotide Identify tool. The species threshold for ANI is also 95%. DNA-DNA Hybridization is the current standard for phylogenomic comparison. eDDH values were calculated using the Genome-to-Genome Distance Calculator (Leibniz Institute DSMZ German Collection of Microorganisms and Cell Cultures, Braunschweig, Germany) (Meier-Kolthoff *et al.*, 2013). The species threshold for DDH is 70%. The Newman Lab also developed a tool called the Reciprocal Orthology Score Average (ROSA) which compares the percent of shared genes in the genome as well as the percent similarity of those shared genes.

Phenotypic Characterization

Phenotypically, streak plates provided colony morphology data. The Omnilog system (Biolog, Inc., Hayward, CA) analyzed a 96 well plate, each well containing a different metabolite. The detected the presence of dye which reacted with the organism's growth, signifying that the organism used that metabolite as a nutrient source. The Biolog software described the organism's metabolic profile and closest estimated relative.

Results

Flavobacterium douthatii ABG Fall 2015

In earlier work, the 16S rRNA sequence was assembled from Sanger sequences of the 16s rRNA gene PCR product and deposited in GenBank under accession # KF648282. The EZTaxon database was used to verify the identity of the cultured organism with the already deposited gene sequence in GenBank. This web-based service compares the 16S sequence to the bank of sequences contained within its expansive database of archaeal and prokaryotic data. Additionally, if any new species published since Ashley's study were similar to ABG they would need to be identified and included in further genetic analysis. When the 16S rRNA sequence for *Flavobacterium* sp ABG previously deposited in GenBank was compared to published species using EZTaxon, there were no pairwise similarities over 98.7%, supporting the notion that strain ABG belongs to a species that has never been studied. The closest phylogenetic relative based on the 16S rRNA sequence data was determined to be *F. succinicans* with a 98.16% similarity. A phylogenetic tree was also constructed with MEGA-6 for *Flavobacterium douthatii* based on the 16S rRNA sequence. This neighbor-joining tree showed *F. succinicans* and *F.glaciei* as the most similar to ABG (Figure 1). *F. oncorhinchi*, which had a high pairwise similarity score, clustered in a different area of the tree. These reference species were chosen because they clustered closely to ABG on the phylogenetic tree: *F. succinicans, F. hydatis, F. glaciei*. *F. aquatile*, although it did not cluster closely based on sequence, was also used for comparison because it is the type species for the genus and should be included in the comparison (Weeks, 1955).

The reference species, along with other published *Flavobacteria*, were compared at the genome level to ABG with DDH, AAI, and ROSA. ABG appears to be equally related to both *F. chilense* (24.6%) and *F. hibernum* (24.8%) and they appear to be its closest relatives, based on estimated DDH (Figure 2).

KF648282/Flavobacterium sp. ABG Edit

ce :1416bp View query sequence

:ss :98.0% (30 – 1446)

Rank	Name	Strain	Authors	Accession	Pairwise Similarity (%)	Diff/Total nt
1	Flavobacterium succinicans	LMG 10402(T)	(Reichenbach 1989) Bernardet et al. 1996	JATV01000001	98.16	26/1415
2	Flavobacterium oncorhynchi	631-08(T)	Zamora et al. 2012	FN669776	98.02	28/1413
3	Flavobacterium glaciei	0499(T)	Zhang et al. 2006	DQ515962	97.95	29/1415
4	Flavobacterium hydatis	DSM 2063(T)	Bernardet et al. 1996	AM230487	97.69	32/1384
5	Flavobacterium chilense	LM-09-Fp(T)	Kämpfer et al. 2012	FR774915	97.68	32/1380
6	Flavobacterium aquidurense	WB-1.1.56(T)	Cousin et al. 2007	AM177392	97.60	34/1415
7	Flavobacterium granuli	Kw05(T)	Aslam et al. 2005	AB180738	97.51	35/1404
8	Flavobacterium chungangense	LMG 26729(T)	Kim et al. 2009	JASY01000008	97.46	36/1415
9	Flavobacterium hercynium	WB 4.2-33(T)	Cousin et al. 2007	AM265623	97.45	36/1413
10	Flavobacterium piscis	412R-09(T)	Zamora et al. 2014	HE612101	97.45	36/1412
11	Flavobacterium panaciterrae (Invalid name)	DCY69(T)	Jin et al. 2014	JX233806	97.34	37/1393
12	Flavobacterium pectinovorum	DSM 6368(T)	(Reichenbach 1989) Bernardet et al. 1996	AM230490	97.31	38/1415
13	Flavobacterium saccharophilum	DSM 1811(T)	(Reichenbach 1989) Bernardet et al. 1996	AM230491	97.30	38/1406
14	Flavobacterium reichenbachii	LMG 25512(T)	Ali et al. 2009	JPRL01000002	97.24	39/1415
15	Flavobacterium plurextorum	1126-1H-08(T)	Zamora et al. 2014	HE612094	97.24	39/1411
16	Flavobacterium cutihirudinis	E89(T)	Glaeser et al. 2013	JX966231	97.16	39/1373
17	Flavobacterium hibernum	DSM 12611(T)	McCammon et al. 1998	JPRK01000008	97.03	42/1415
18	Flavobacterium limicola	ST-82(T)	Tamaki et al. 2003	AB075230	97.03	42/1415

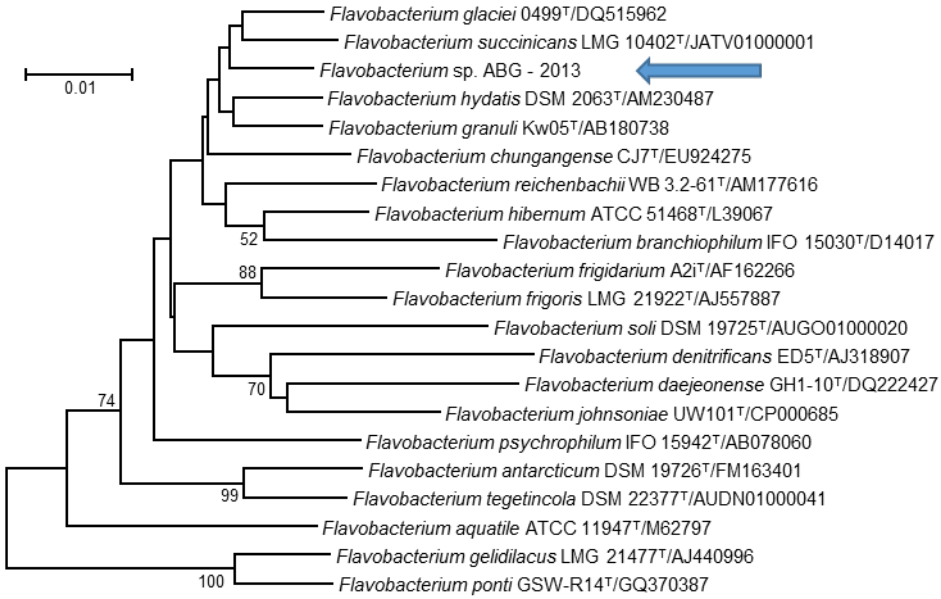

Figure 1: Upper. Best matching 16S rRNA sequences identified with EZ Taxon

Lower: Neighbor-joining phylogenetic tree showing 16S rRNA phylogeny of *Flavobaterium douthatii* ABG. The scale line denotes the distance measured to analyze relatedness along with the bootstrap values on each node out of 1000 replications.

DDH	
Flavobacterium chilense	24.6
Flavobacterium chungangense	24.3
Flavobacterium succinicans	20.5
Flavobacterium aquatile	19.9
Flavobacterium hibernum	24.8
Flavobacterium hydatis	22.9
Flavobacterium reichenbachii	24.5

Figure 2 DNA-DNA Hybridization values calculated in reference to ABG. Scores less than 70% are considered indicative of a novel species.

Additionally, *F. chilense* (81.2%) and *F. hibernum* (82.0%) were again shown to be the closest relatives to ABG on the basis of AAI (Figure 3). 16S rRNA comparisons did not show these species as the closest relative or as equidistantly related. ANI was not used because ABG and its reference species were too distantly related and below the detection limit. ROSA also showed *F. hibernum* (46.2) and *F. chilense* (46.8) as the most similar species to ABG.

Average Amino Acid Identity (AAIr)		1	2	3	4	5	6	7	8	9	10	11
Flavobacterium sp. ABG	1		67.8	81.2	80.4	81.8	75.8	82.0	78.4	80.4	81.5	71.8
Flavobacterium aquatile LMG 4008	2	67.6		67.7	67.5	67.9	70.0	67.7	67.6	67.0	67.1	68.6
Flavobacterium chilense	3	81.3	67.6		80.9	84.1	74.8	85.9	78.0	83.4	82.4	72.2
Flavobacterium chungangense LMG 26729	4	80.1	67.5	80.7		81.8	74.3	81.7	77.0	80.8	80.9	71.1
Flavobacterium denitrificans DSM 15936	5	81.9	67.8	84.0	81.9		74.7	83.6	78.0	84.8	82.7	71.8
Flavobacterium glaciei CGMCC 1.5380	6	75.5	70.1	74.7	74.4	74.5		74.7	75.9	73.7	74.0	75.2
Flavobacterium hibernum DSM 12611	7	82.1	67.5	85.8	81.7	83.7	74.8		79.7	82.6	82.4	71.6
Flavobacterium hydatis DSM 2063	8	78.2	67.5	78.0	77.1	78.0	76.0	79.8		76.7	76.0	72.6
Flavobacterium johnsoniae UW101	9	80.6	67.2	83.6	81.1	84.8	74.1	82.8	77.1		82.5	71.2
Flavobacterium reichenbachii	10	81.5	67.4	82.8	81.2	82.8	74.4	82.5	76.3	82.7		71.2
Flavobacterium succinicans LMG 10402	11	71.8	68.6	72.0	71.1	71.6	75.5	71.6	72.5	70.9	71.2	

Figure 3 Average Amino Acid Identity (AAI) calculated using the unpublished ROSA calculator developed by the Newman lab. Scores less than 95% are indicative of a new species. The values in reference to ABG are highlighted.

A Venn diagram comparing *F. succinicans, F. glaciei, F. hydatis,* and *F. aquatile* was constructed using RAST annotations (Figure 4). *Flavobacterium douthatii* had 1272 genes that were unique and all of the organisms shared 1759 genes. However, upon further examination of the annotated genome a problem was found with the edited assembly, the 16S rRNA gene was deleted. The original genome assembly was therefore investigated to determine at what point the problem occurred. In that file, it was found that the initial computer-based assembly had a repeated area (Figure 5), presumably because most organisms have multiple rRNA genes and such repeated sequences present challenges to assemblers. Unfortunately, time did not allow for this editing and so the focus shifted to an organism with a properly assembled genome that would be more time effective to study.

Further genomic comparisons should include *F. chilense* and *F. hibernum* based on this new genomic-level information indicating that they are ABG's closest published relatives.

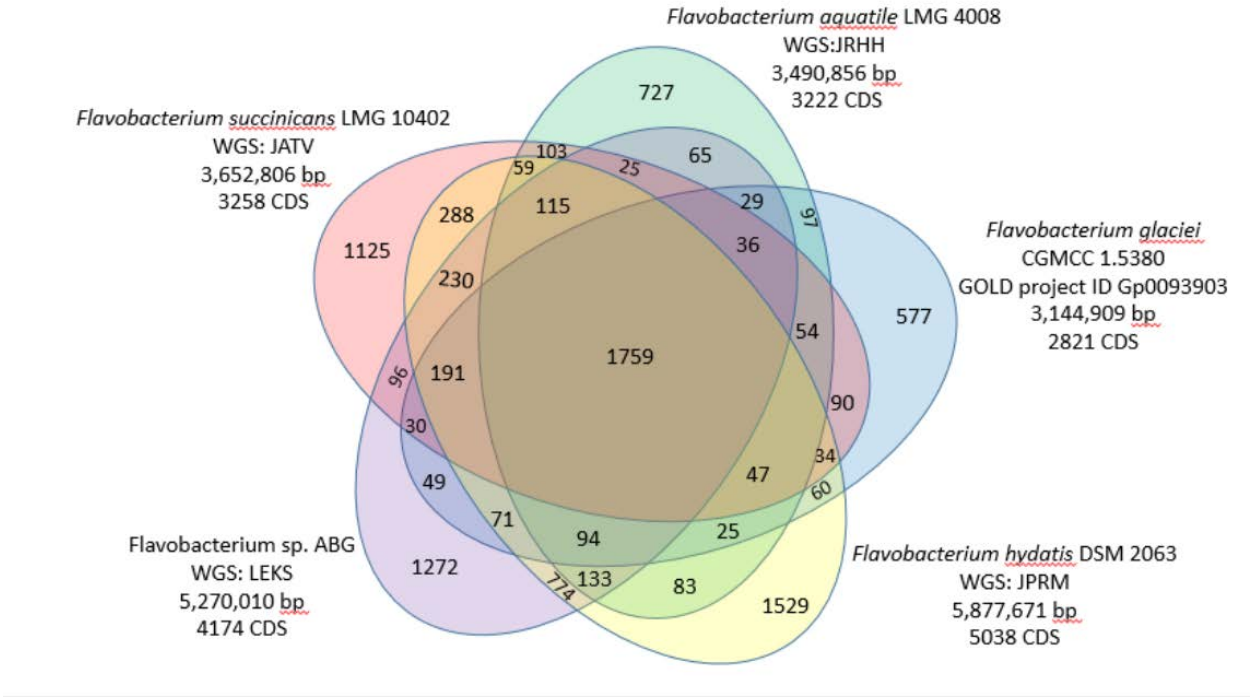

Figure 4 Venn diagram comparing unique and shared genes of ABG and its reference species. ABG was found to have 1272 annotated unique genes and all of the organisms shared 1759 genes.

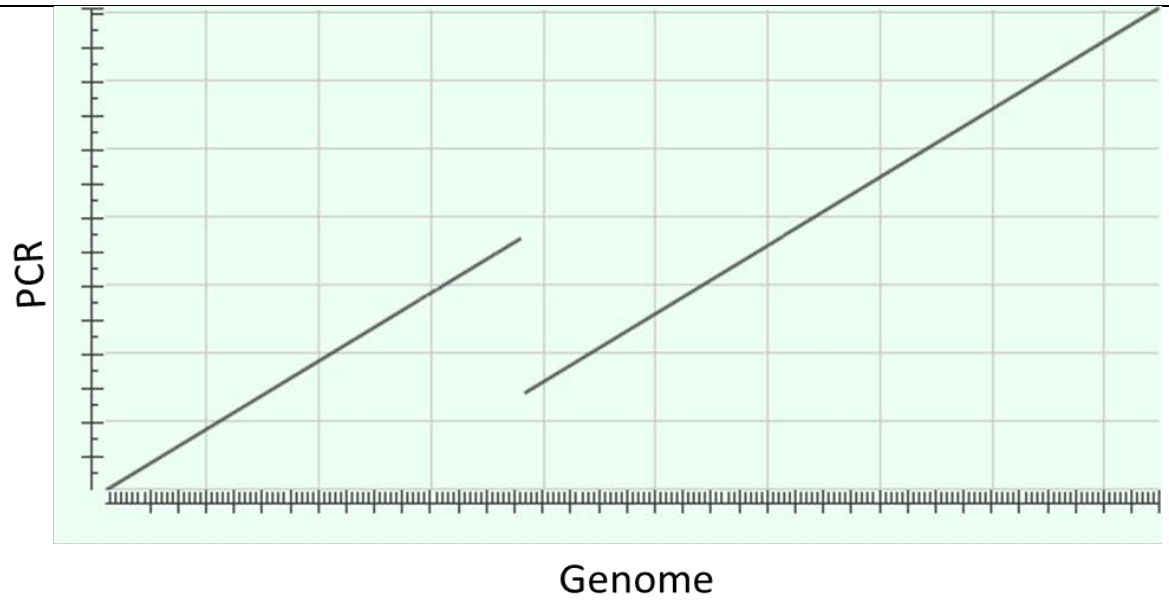

Figure 5. Dot Plot comparison of ABG 16S rRNA sequence obtained from genome (X axis) vs sequence produced by sequencing 16S rRNA PCR product illustrating that part of the 16S gene is represented twice in the genome derived sequence.

Flavobacterium gabrieli KJJ Spring 2016

Flavobacterium gabrieli KJJ was originally identified by Kathy Jacobs in 2012 and its 16S rRNA and genome sequences were already annotated and assembled by other students. The assembly of the genome was first examined for possible problems by checking RAST annotation and using the Basic Local Alignment Tool (BLAST) to compare it to the deposited genome in GenBank. Once this checked out, the 16S rRNA sequence was used to verify the identity of the cultured organism using PCR amplification and Sanger sequencing. The 16S sequence was compared to the EZTaxon database to make sure no new organisms had been published since Kathy's study that were more similar or the same species. EZTaxon's most similar result was *F. chungangense* (98.03%). A neighbor-joining phylogenetic tree was constructed based on the 16S rRNA sequence. *F. chilense* and *F. chungangense* were the species that clustered closest to KJJ (Figure 6). A 16S rRNA matrix was constructed from MEGA6 to quantify the 16S similarity of KJJ and some of its closest relatives. *F. chungangense* was the most similar with a 16S rRNA similarity of 98.23%. This is below the 98.7% threshold for a new species.

of kjj Edit

nce :1509bp View query sequence

ess :100% (1 – 1510)

Rank	Name	Strain	Authors	Accession	Pairwise Similarity (%)	Diff/Total nt
1	Flavobacterium chungangense	LMG 26729(T)	Kim et al. 2009	JASY01000008	98.03	29/1472
2	Flavobacterium chilense	LM-09-Fp(T)	Kämpfer et al. 2012	FR774915	97.84	30/1388
3	Flavobacterium saccharophilum	DSM 1811(T)	(Reichenbach 1989) Bernardet et al. 1996	AM230491	97.73	33/1453
4	Flavobacterium glaciei	0499(T)	Zhang et al. 2006	DQ515962	97.68	33/1422
5	Flavobacterium aquidurense	WB-1.1.56(T)	Cousin et al. 2007	AM177392	97.67	34/1460
6	Flavobacterium tructae	435-08(T)	Zamora et al. 2014	HE612100	97.67	34/1458
7	Flavobacterium spartansii	T16(T)	Loch and Faisal 2014	JX287799	97.54	34/1383
8	Flavobacterium oncorhynchi	631-08(T)	Zamora et al. 2012	FN669776	97.53	35/1417
9	Flavobacterium succinicans	LMG 10402(T)	(Reichenbach 1989) Bernardet et al. 1996	JATV01000001	97.42	38/1472
10	Flavobacterium reichenbachii	LMG 25512(T)	Ali et al. 2009	JPRL01000002	97.28	40/1472
11	Flavobacterium pectinovorum	DSM 6368(T)	(Reichenbach 1989) Bernardet et al. 1996	AM230490	97.27	40/1467
12	Flavobacterium panaciterrae (Invalid name)	DCY69(T)	Jin et al. 2014	JX233806	97.27	38/1393
13	Flavobacterium hydatis	DSM 2063(T)	Bernardet et al. 1996	AM230487	97.26	39/1423
14	Flavobacterium piscis	412R-09(T)	Zamora et al. 2014	HE612101	97.26	40/1459
15	Flavobacterium hercynium	WB 4.2-33(T)	Cousin et al. 2007	AM265623	97.21	41/1467
16	Flavobacterium cutihirudinis	E89(T)	Glaeser et al. 2013	JX966231	97.09	40/1373
17	Flavobacterium chungbukense	CS100(T)	Lim et al. 2011	HM627539	97.01	44/1470
18	Flavobacterium plurextorum	1126-1H-08(T)	Zamora et al. 2014	HE612094	96.98	44/1457
19	Flavobacterium psychrolimnae	LMG 22018(T)	Van Trappen et al. 2005	AJ585428	96.97	44/1452
20	Flavobacterium hibernum	DSM 12611(T)	McCammon et al. 1998	JPRK01000008	96.94	45/1472

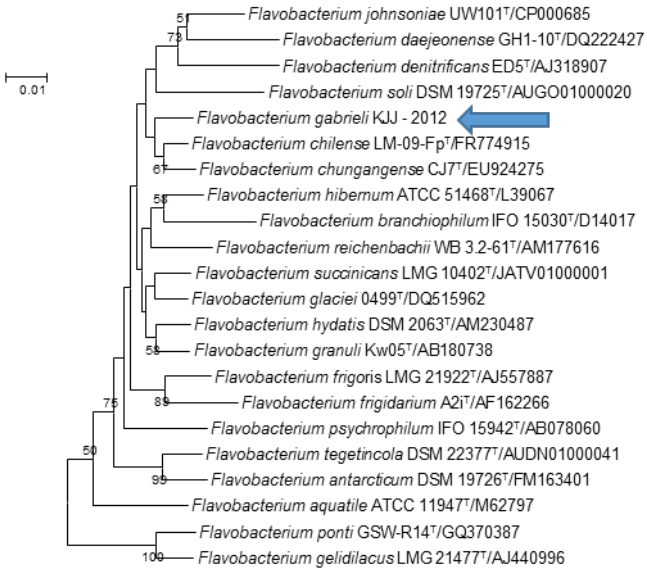

Figure 6 Upper. Best matching 16S rRNA sequences identified with EZ Taxon

Lower. Neighbor-joining phylogenetic tree showing 16S rRNA phylogeny of

Flavobacterium gabrieli KJJ.

Then, KJJ's genome was compared on a broad scale using AAI and DDH to a variety of *Flavobacteria* genomes publically available in the GenBank database as well as genomes from organisms in closely related genera such as *Mesoflavibacter zeaxanthinifaciens*. This genome-level comparison ensured that we were tailoring our phylogenomic metrics to include the most similar organisms, not just those based on 16S rRNA sequence as was the case with ABG. In both KJJ and ABG, genomic characterization identified different closest relatives than the 16S rRNA sequence did. Estimated DDH showed *F. hibernum* (26.7%) and *F. chilense* (26.1%) were the most similar species, while *F. chungangense* was not nearly as similar (24.8%) as it was in the 16S comparison. AAI showed that *F. chilense* and *F. denitrificans* were both the closest species and equally similar to KJJ (85.8%), while *F chungangense* was only 82.5% similar. ANI found that *F. hibernum* (82.6%) and *F. chilense* (82.4%) were the most similar, again *F. chungangense* was only 80.5%. The slight differences between the metrics speak to their differing levels of accuracy and mechanism. AAI, with the best resolution and less variability than ANI, was the metric relied on most heavily in this study. Nonetheless, all of the phylogenomic metric scores were below the 70% threshold for DDH and 95% threshold for AAI and ANI, supporting *Flavobacterium gabrieli* KJJ's distinction as an uncharacterized species. They were all grouped into one convenient table, which is a new formatting change for simplicity (Figure 7).

					Estimated DNA-DNA Hybridization (eDDH)										
Accession		16S	ANI		1	2	3	4	5	6	7	8	9	10	11
JARP00000000	*Flavobacterium gabrieli* KJJ			1		26.1	25.2	26.7	25.3	25.0	24.8	22.5	20.9	19.6	19.1
LSYT00000000	*Flavobacterium chilense* LMG 26360^T	98.23	82.2	2	85.8		24.8	27.3	24.6	25.7	24.7	22.3	20.7	19.8	18.9
AUDL00000000	*Flavobacterium denitrificans* DSM 15936^T	96.55	82.4	3	85.8	84.1		24.6	24.3	25.9	24.1	22.0	20.5	20.0	19.3
JPRK00000000	*Flavobacterium hibernum* DSM 12611^T	97.14	82.6	4	85.2	85.9	83.6		25.3	24.9	24.7	23.3	21.1	20.1	18.6
JPRL00000000	*Flavobacterium reichenbachii* LMG 25512^T	97.05	82.0	5	84.0	82.4	82.7	82.4		24.9	24.0	22.2	20.9	19.4	18.9
CP000685	*Flavobacterium johnsoniae* UW101^T	96.30	81.7	6	83.5	83.4	84.8	82.6	82.7		24.2	22.6	20.8	20.3	19.8
JASY00000000	*Flavobacterium chungangense* LMG 26729^T	98.32	80.5	7	82.5	80.9	81.9	81.7	81.2	81.1		22.1	21.0	19.4	18.7
JPRM00000000	*Flavobacterium hydatis* DSM 2063^T	97.56	77.6	8	78.3	78.0	78.0	79.7	76.3	77.1	77.0		21.4	20.1	18.4
FQ859183	*Flavobacterium glaciei* CGMCC 1.5380^T	97.73	75.8	9	75.3	74.8	74.7	74.8	74.4	74.1	74.3	76.0		19.9	19.1
JRHH00000000	*Flavobacterium aquatile* LMG 4008^T	95.54	72.3	10	68.0	67.6	67.8	67.5	67.4	67.2	67.5	67.5	70.1		19.4
AULQ00000000	*Mesoflavibacter zeaxanthinifaciens* DSM 18436^T	89.35	68.7	11	58.8	58.0	58.5	58.2	58.1	58.3	58.4	58.4	59.8	58.3	
					Average Amino Acid Identity (AAI)										

Figure 7 Phylogenomic matrix including values for 16S comparison, ANI, AAI, and DDH. The closest scores are in green and the least similar scores are in red.

A Venn diagram was constructed using RAST annotation to compare KJJ with *F. denitrificans, F. hibernum, F. chilense,* and *F. aquatile* which were chosen based on the phylogenomic matrices (Figure 8). *Flavobacterium gabrieli* KJJ had 505 unique genes (Figure 10) and all of the organisms shared 1861 genes.

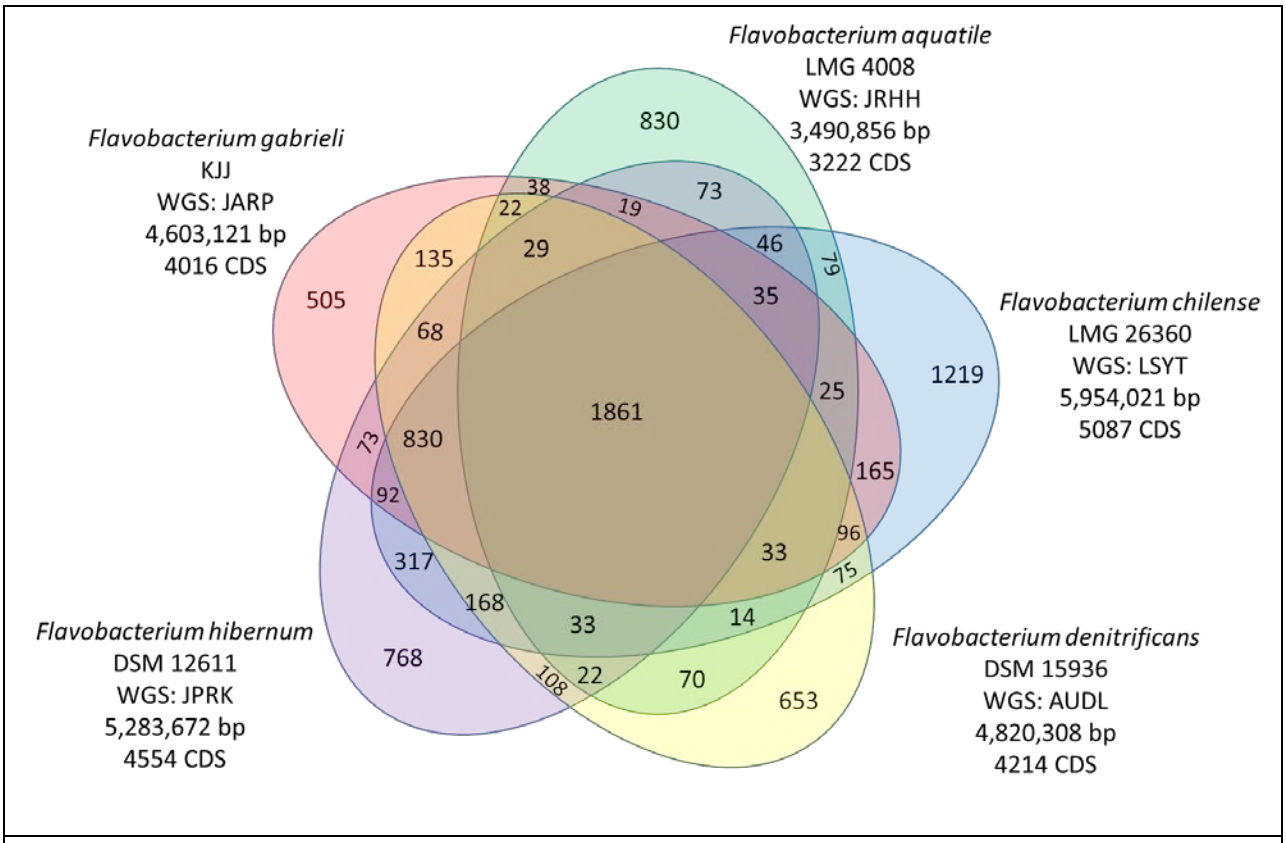

Figure 8 Venn diagram comparing unique and shared genes of KJJ and its reference species. KJJ was found to have 505 annotated unique genes and all of the organisms shared 1861 genes.

Biolog metabolic profiles for *F. chilense, F. chungangense,* and *F. hibernum* were compared to KJJ for differences in metabolic capabilities (Figure 9). The annotated genome was then explored in RAST for a possible genetic basis for the differences in phenotypic traits. *Flavobacterium gabrieli* KJJ was not able to use trehalose, while *F. chilense* and *F. hibernum* were able to use it quite well. Trehalose is a disaccharide composed of two glucose molecules linked by an α,α-1,1 glucoside bond (Jain & Roy, 2009). The trehalose pathway genes in *F. chilense* and KJJ were compared side by side in RAST. The trehalose permease gene, the enzyme responsible for uptake of trehalose, was found in *F. chilense* but not in KJJ. KJJ was also missing the gene for trehalase, which is responsible for breaking the

glycosidic linkage for trehalose utilization. These two missing genes explain the lack of trehalose utilization in KJJ whereas its close phylogenomic relative *F. chilense* was able to effectively use trehalose. Another metabolic difference that was examined was hexosamine utilization. *F. chilense* and KJJ were both able to use N-acetyl glucosamine. However, KJJ was uniquely able to use N-acetyl mannosamine while *F. chilense* uniquely used N-acetyl galactosamine and N-acetyl neuraminic acid.

well	carbon source or condition	Flavobacterium gabrieli KJJ	Flavobacterium chiliense	Flavobacterium chungangense	Flavobacterium hibernum	well	carbon source or condition	Flavobacterium gabrieli KJJ	Flavobacterium chiliense	Flavobacterium chungangense	Flavobacterium hibernum
A01	neg control	21	24	24	23	E01	gelatin	92	99	93	98
A02	dextrin	99	98	99	98	E02	glycyl-L-proline	53	95	95	91
A03	D-maltose	90	98	99	98	E03	L-alanine	44	83	7	33
A04	D-trehalose	12	98	10	97	E04	L-arginine	44	87	22	64
A05	D-cellobiose	57	98	98	99	E05	L-aspartic acid	44	97	91	93
A06	gentiobiose	99	98	99	97	E06	L-glutamic acid	70	97	97	95
A07	sucrose	14	11	97	98	E07	L-histidine	30	71	23	23
A08	D-turanose	14	12	20	9	E08	L-pyroglutamic acid	10	9	13	14
A09	stachyose	14	13	11	32	E09	L-serine	44	89	74	73
A10	pos control	98	97	98	98	E10	lincomycin	44	11	10	10
A11	pH 6	97	96	96	98	E11	guanidine HCl	40	18	9	39
A12	pH 5	27	67	12	16	E12	niaproof 4	16	14	14	12
B01	D-raffinose	17	14	14	96	F01	pectin	28	45	97	71
B02	α-D-lactose	16	21	19	22	F02	D-galacturonic acid	56	98	98	96
B03	D-melibiose	17	13	14	10	F03	L-galacturonic acid lactone	12	8	9	53
B04	β-methyl-D-glucoside	14	17	98	97	F04	D-gluconic acid	10	14	27	8
B05	D-salicin	10	98	98	98	F05	D-glucuronic acid	18	47	97	51
B06	N-acetyl-D-glucosamine	99	97	6	96	F06	glucuronamide	17	21	33	18
B07	N-acetyl-β-D-mannosamine	77	16	10	12	F07	mucic acid	10	10	15	7
B08	N-acetyl-D-galactosamine	48	97	36	95	F08	quinic acid	10	12	21	16
B09	N-acetyl neuraminic acid	11	97	7	95	F09	D-saccharic acid	10	10	14	8
B10	1% NaCl	55	92	94	63	F10	vancomycin	63	94	95	94
B11	4% NaCl	14	13	13	12	F11	tetrazolium violet	90	99	45	70
B12	8% NaCl	19	16	16	11	F12	tetrazolium blue	99	99	95	96
C01	α-D-glucose	99	98	98	97	G01	p-hydroxy-phenylacetic acid	15	12	9	11
C02	D-mannose	99	97	98	97	G02	methyl pyruvate	14	71	12	85
C03	D-fructose	44	89	34	94	G03	D-lactic acid methyl ester	13	14	27	17
C04	D-galactose	99	97	51	97	G04	L-lactic acid	10	13	19	11
C05	3-methyl glucose	10	13	13	8	G05	citric acid	16	14	21	15
C06	D-fucose	12	16	6	8	G06	α-keto-glutaric acid	11	10	18	10
C07	L-fucose	16	20	68	87	G07	D-malic acid	13	11	16	12
C08	L-rhamnose	46	96	98	93	G08	L-malic acid	46	13	15	15
C09	inosine	11	8	7	8	G09	bromo-succinic acid	10	8	6	7
C10	1% Na-lactate	91	93	96	94	G10	nalidixic acid	16	13	12	13
C11	fusidic acid	11	10	9	9	G11	LiCl	11	11	9	11
C12	D-serine	16	15	10	12	G12	K-tellurite	21	18	16	18
D01	D-sorbitol	17	15	13	11	H01	tween-40	50	97	96	85
D02	D-mannitol	12	11	18	9	H02	γ-amino-butyric acid	16	12	12	16
D03	D-arabitol	10	12	11	7	H03	α-hydroxy-butyric acid	16	11	13	12
D04	myo-inositol	11	10	15	7	H04	β-hydroxy-D,L-butyric acid	13	11	19	13
D05	glycerol	44	8	8	8	H05	α-keto-butyric acid	12	8	7	7
D06	D-glucose-6-PO4	36	16	27	28	H06	acetoacetic acid	30	31	41	47
D07	D-fructose-6-PO4	27	21	90	34	H07	propionic acid	12	10	6	7
D08	D-aspartic acid	7	7	6	8	H08	acetic acid	51	91	89	60
D09	D-serine	7	7	6	6	H09	formic acid	16	11	8	9
D10	troleandomycin	12	10	9	9	H10	aztreonam	96	96	98	97
D11	rifamycin SV	96	94	96	96	H11	Na-butyrate	16	14	15	13
D12	minocycline	16	13	12	13	H12	Na bromate	15	14	16	12

Figure 9 Biolog metabolic profiles of *F. chilense, F. chungangense,. F. hibernum,* and KJJ.

Characteristics that distinguish KJJ from its relatives are highlighted.

The genes for hexosamine utilization were compared side by side in RAST for *F. chilense* and KJJ. For the NagX N-acetyl glucosamine related transporter, gene 2698 in *F. chilense* was a unidirectional best hit for 3010 in KJJ, meaning *F. chilense* may have a duplication of this gene or that KJJ lost a copy. Since these genes are not bidirectional best hits they are not orthologs of each other, or not the best match. They were still paralogs, however, meaning they are similar genes but may have different functions. A mutation in this transporter gene may have affected the specificity, allowing it to transport different hexosamines in *F. chilense* and KJJ. Another unidirectional best hit in a beta-hexosaminidase, which may have influenced the specificity for different glucose derivatives. Other beta-hexosaminidase genes were bidirectional best hits or orthologs. These genes were doing exactly the same function in both organisms. This may explain why both organisms were capable of using N-acetyl glucosamine. A third metabolic difference that was examined in genetic detail was β-methyl glucoside utlilization. *F. hibernum* and *F. chungangense* were both able to utilize β-methyl glucoside while KJJ and *F. chilense* were not. Gene maps in RAST were examined to determine which genes may be responsible for this difference (Figure 11). These gene sequences were then searched in the BLAST database to find organisms with highly similar, orthologous gene sequences. Organisms that appeared to have an ortholog were compared to the Biolog data to confirm the influence of the ortholog on metabolism. Interestingly, *F. hibernum* and *F. chungangense* both shared orthologs of TonB dependent receptors, two beta-glucosidase genes, and an esterase gene that were not found to be orthologs in *F. chilense* and KJJ. This

may explain why *F. chungangense F. hibernum* were able to use β-methyl glucoside while their close phylogenomic relatives could not.

5	Two-component response regulator	2350	Alpha-L-Rha alpha-1,3-L-rhamnosyltransferase (EC 2.4.1.-)
30	Non-specific DNA-binding protein Dps / Iron-binding ferritin-like antioxidant protein / Ferroxidase (EC 1.16.3.1)	2352	UDP-glucose 4-epimerase (EC 5.1.3.2)
39	Adenylate cyclase (EC 4.6.1.1)	2364	UDP-4-amino-4-deoxy-L-arabinose--oxoglutarate aminotransferase (EC 2.6.1.-)
43	Serine esterase (EC 3.1.1.1)	2369	Putative CDP-glycerol:glycerophosphate glycerophosphotransferase (EC 2.7.8.-)
48	putative reductase	2370	Glycosyltransferase (EC 2.4.1.-)
51	Glucanase C	2373	Streptomycin biosynthesis StrF domain protein
52	Polyphosphate kinase (EC 2.7.4.1)	2374	Glycosyl transferase, group 2 family protein
54	Prolidase (EC 3.4.13.9)	2376	putative glycosyltransferase
59	Diguanylate cyclase (GGDEF domain) with GAF sensor	2387	Asparagine synthetase [glutamine-hydrolyzing] (EC 6.3.5.4)
61	Two-component sensor histidine kinase	2473	gnl|WGS:AAAB|ebiP470|gb|EAA02729
66	EstC	2688	DNA-damage-inducible protein D
67	salicylate esterase	2690	Phosphoenolpyruvate synthase (EC 2.7.9.2)
68	salicylate esterase	2693	Small heat shock protein
69	Hydrolase, alpha/beta fold family functionally coupled to Phosphoribulokinase	2697	Polyribonucleotide nucleotidyltransferase (EC 2.7.7.8)
102	LemA protein	2698	Osmotically inducible protein Y precursor
103	possible Galanin	2704	Osmotically inducible protein Y precursor
106	Putative bacteriophage protein	2705	Osmotically inducible protein Y precursor
109	glucose-fructose oxidoreductase	2720	Rhs family protein
253	SanA protein	2723	Mll6838 protein
256	Bll1930 protein	2729	Acyl-CoA dehydrogenase; probable dibenzothiophene desulfurization enzyme
364	Two-component system response regulator	2732	ThiJ/Pfpl family protein
408	FIG145533: Methyltransferase (EC 2.1.1.-)	2757	Glyceraldehyde-3-phosphate dehydrogenase, putative
420	Lipoprotein spr precursor	2760	Xylose isomerase (EC 5.3.1.5)
477	Glycosyltransferase (EC 2.4.1.-)	2768	possible sensor for regulator EvgA(EC:2.7.3.-)
572	Putative inner membrane protein	2769	Anti-sigma B factor RsbT
604	N-acetyltransferase	2770	Anti-sigma B factor RsbT
703	Nucleotidyltransferase (EC 2.7.7.-)	2771	RsbS, negative regulator of sigma-B
704	Nucleotidyltransferase (EC 2.7.7.-)	2772	RsbR, positive regulator of sigma-B
729	TsaC protein (YrdC domain) required for threonylcarbamoyladenosine t(6)A37 modification in tRNA	2775	Potassium efflux system KefA protein / Small-conductance mechanosensitive channel
730	cAMP-binding proteins - catabolite gene activator and regulatory subunit of cAMP-dependent protein kinases	2779	Lipoprotein
737	putative membrane protein	2780	protein of unknown function DUF892
782	cAMP-binding proteins - catabolite gene activator and regulatory subunit of cAMP-dependent protein kinases	2781	Glycosyltransferase
797	Two-component response regulator	2785	putative membrane protein
798	histidine kinase sensor protein	2798	transcriptional regulator, AraC family
799	Circadian oscillation regulator KaiB	2802	Transcriptional regulator, AraC family
800	Circadian oscillation regulator KaiB	2828	MutT-like protein
801	Circadian clock protein KaiC	2900	Outer membrane lipoprotein omp16 precursor
807	Transcriptional regulator, ArsR family	2907	Hep_Hag
846	Transcriptional regulator, AraC family	2985	Succinyl-CoA ligase [ADP-forming] beta chain (EC 6.2.1.5)
847	putative 6-aminohexanoate-dimer hydrolase	3091	S23 ribosomal
851	Phytochrome, two-component sensor histidine kinase (EC 2.7.3.-); Cyanobacterial phytochrome B	3143	Lacl family transcriptional regulator
869	diguanylate cyclase/phosphodiesterase (GGDEF & EAL domains) with PAS/PAC sensor(s)	3166	Phage protein
870	Two-component system response regulator	3169	Chemotaxis protein methyltransferase CheR (EC 2.1.1.80)
873	Conserved protein	3179	Phosphate regulon transcriptional regulatory protein PhoB (SphR)
874	FAD dependent oxidoreductase	3180	Signal transduction histidine kinase
882	putative helicase	3186	6-aminohexanoate-dimer hydrolase (EC 3.5.1.46)
897	Modulator of drug activity B	3188	Transcriptional regulator, AraC family
898	Aldo-keto reductase	3205	no hits
902	Aldo-keto reductase	3233	ATPase involved in DNA repair
904	Flavodoxin	3273	Cytochrome c oxidase polypeptide I (EC 1.9.3.1) # putative
905	Amidohydrolase domain protein	3360	ThiJ/Pfpl family protein
906	Transcriptional regulator, TetR family	3373	L-asparaginase (EC 3.5.1.1)
907	UDP-glucose 4-epimerase (EC 5.1.3.2)	3384	TonB-dependent siderophore receptor
910	Oxygen-insensitive NAD(P)H nitroreductase (EC 1.-.-.-) / Dihydropteridine reductase (EC 1.5.1.34)	3407	Isochorismatase (EC 3.3.2.1)
935	Phosphatidylserine decarboxylase (EC 4.1.1.65)	3593	Transcriptional regulator
1036	Chemotaxis protein methyltransferase CheR (EC 2.1.1.80)	3595	Protease
1037	Two-component response regulator CheY subfamily	3596	Lanthionine biosynthesis protein LanB
1038	transcriptional regulator, HxlR family	3597	Lanthionine biosynthesis protein LanB
1040	3-oxoacyl-[acyl-carrier protein] reductase (EC 1.1.1.100)	3607	DNA-methyltransferase
1042	Glyoxalase family protein	3608	orf98
1163	RDD domain containing protein	3614	helix-turn-helix- domain containing protein, AraC type
1265	Type I restriction-modification system, specificity subunit S (EC 3.1.21.3)	3615	Short chain dehydrogenase
1353	Transcriptional regulator, AraC family	3653	Phophatidylinositol-4-phosphate 5-kinase (EC 2.7.1.68)
1354	Short chain dehydrogenase	3657	internalin, putative
1387	Endoglucanase D precursor (EC 3.2.1.4)	3658	Cell well associated RhsD protein
1407	internalin, putative	3662	Outer membrane protein H precursor
1441	Tsr1131 protein	3668	BatD
1453	Isochorismatase (EC 3.3.2.1)	3763	cAMP-dependent Kef-type K+ transport system
1457	Mll4938 protein	3768	Exodeoxyribonuclease III (EC 3.1.11.2)
1526	Ribosomal-protein-L7p-serine acetyltransferase	3821	ThiJ/Pfpl family protein
1664	Putative DNA-binding protein in cluster with Type I restriction-modification system	3825	Probable glutathione S-transferase-related transmembrane protein (EC 2.5.1.18)
1675	Phytoene desaturase, neurosporene or lycopene producing (EC 1.3.-.-)	3876	Glutathione-regulated potassium-efflux system protein KefB
1686	Beta-galactosidase (EC 3.2.1.23)	3879	probable membrane protein STY1534
1851	Transcriptional regulator	3884	NG,NG-dimethylarginine dimethylaminohydrolase 1 (EC 3.5.3.18)
2333	Glucose-1-phosphate cytidylyltransferase (EC 2.7.7.33)	3885	Cyanophycinase (EC 3.4.15.6)
2334	Similar to CDP-glucose 4,6-dehydratase (EC 4.2.1.45)	3886	Arginine/ornithine antiporter ArcD
2335	dTDP-glucose 4,6-dehydratase (EC 4.2.1.46)	3889	TonB family protein / TonB-dependent receptor
2336	2,4-dihydroxyhept-2-ene-1,7-dioic acid aldolase (EC 4.1.2.-)	3890	Cyanophycinase (EC 3.4.15.6)
2337	Acetolactate synthase large subunit (EC 2.2.1.6)	3892	Transcriptional regulator
2338	Acetaldehyde dehydrogenase, acetylating, (EC 1.2.1.10) in gene cluster for degradation of phenols, cresols, catechol	3893	Glucosamine-6-phosphate deaminase (EC 3.5.99.6)
2339	4-hydroxy-2-oxovalerate aldolase (EC 4.1.3.39)	3907	Peptide methionine sulfoxide reductase MsrA (EC 1.8.4.11)
2340	dTDP-glucose 4,6-dehydratase (EC 4.2.1.46)	3956	Biotin synthesis protein BioC
2341	CDP-4-dehydro-6-deoxy-D-glucose 3-dehydratase (EC 4.2.1.-)	3957	Transcriptional regulator, AraC family
2343	glycosyl transferase, family 2	3965	Glucosamine-6-phosphate deaminase (EC 3.5.99.6)
2344	WzxE protein	3967	Major facilitator superfamily (MFS) transport protein
2346	UDP-glucose 4-epimerase (EC 5.1.3.2)	3972	Alpha-1,2-mannosidase
2347	Glycosyl transferase, group 1 family protein	3977	TonB family protein / TonB-dependent receptor
2348	dTDP-rhamnosyl transferase RfbF (EC 2.-.-.-)	4014	Phage major capsid protein

Figure 10 List of annotated genes unique to *Flavobacterium gabrieli* KJJ.

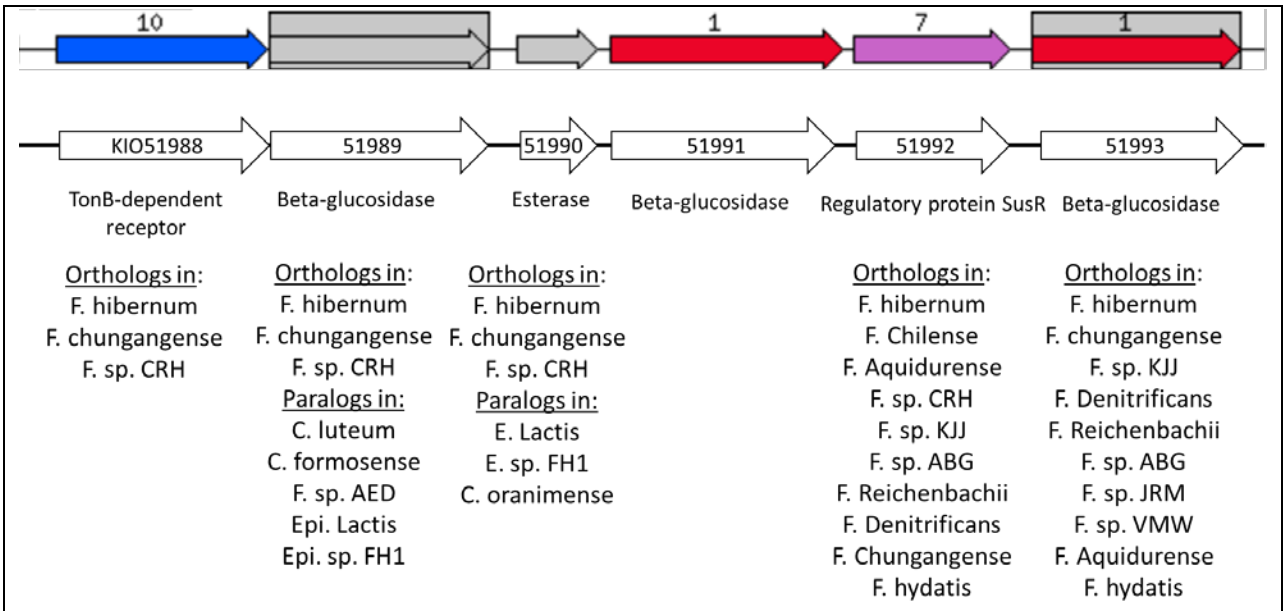

Figure 11 Gene map of β-glucosidase genes in *F. hibernum*. BLAST results are indicated below with organisms that contained orthologs or paralogs based on percent sequence similarity. GenBank accessions are listed in the white arrows for each gene.

Discussion

The first study characterizing *Flavobacterium douthatii* ABG offered valuable lessons in the process by which the lab characterizes novel organisms, despite fatal assembly errors that prevent it from being publishable. Instead of choosing reference species based first on the 16S rRNA phylogenetic tree, we have implemented a new procedure. The reference species are selected based on a broad genomic-level comparison using estimated DDH, AAI, and BBH to all genomes deposited in GenBank within that genus as well as closely related genera. Because 16S rRNA does not always give the most accurate picture of phylogeny, the whole genome must be compared. This, too, is imperfect because not all species have had their whole genomes sequenced yet. Past studies selected reference species based on 16S rRNA alone, but later research using the genome showed that entirely different species should have been used. The emphasis on identifying phenotypic differences and relating them to the genome is not standard in bacteria classification. Identification of these genetic differences can support how unique the new species is, as well as verify that the phenotypic data are accurate. Although many labs do not explicitly compare genotype to phenotype, doing so can provide a more accurate depiction of the organism's phylogeny and homology to their closest ancestors. In this study, examining the genome provided an explanation for the differences in hexosamine utilization between KJJ and its reference species. However, RAST annotates some genes as merely 'hypothetical' without predicting a function for that protein. The curated list of protein functions will be updated with new understanding of gene to protein relationships (Aziz *et al.*, 2008). Annotation of these hypothetical genes will improve and function predictions will become more comprehensive. The study of ABG also highlighted the importance of verifying the

quality of the genome assembly and ensuring that no genes of importance have been deleted. An examination using RAST as well as BLAST to compare to the genome deposited in GenBank is important. This should be done before the preliminary genome comparisons. I was not able to perform proper genome assembly due to time constraints and the advanced skills necessary. In order to publish ABG, the raw genome sequence will need to be reassembled. This is an example of the challenges of many people collaborating on a project over time. The process of the scientific method is self-correcting, problems are subsequently identified and remedied.

In studying *Flavobacterium gabrieli* KJJ I learned the importance of verifying the physical culture with 16S rRNA PCR and Sanger sequencing for identification verification. It took several weeks of culturing organisms from different permanent vials to recover a strain that was later confirmed to match the deposited sequence for KJJ. The identification of the physical organism must be validated so that it can be deposited into culture collections. It is important to ensure the effective communication and organization of information and data within the lab. A lot of data is generated and saved from previous students' work. These data must be organized in a manner that makes it easy to access and compile for publication. I am running into a few difficulties finding some data that I need because of ineffective organization or loss of old data. Some of this older data may need to also be replicated due to discrepancies with literature or consistency. The Newman lab has been making several efforts to effectively organize the vast amount of information in the lab network space such as typed lab journals, the use of spreadsheets, and file naming systems. In order to publish in the *International Journal of Systematic and Evolutionary Microbiology* the proper figures must be formatted for publication quality. These figures

include a matrix including 16S, DDH, ANI, and AAI in one table. The Venn diagram as well as tables for the Biolog and fatty acids will also be published. Additionally, a manuscript describing the new species in the format of the journal must be written.

In September 2011, the Subcommittee on the taxonomy of *Flavobacterium* and *Cytophaga*-like bacteria, which is a part of the greater International Committee on Systematics of Prokaryotes met in Sapporo, Japan (Bowman & Bernardet, 2013). This committee sets the guidelines for the *International Journal of Systematic and Evolutionary Microbiology*, They discussed recent taxonomic developments within the *Flavobacteriaceae* family. First, they discussed the fact that the number of species classified into the genera *Flavobacterium* and *Chryseobacterium* has increased considerably in recent years due to the description of many novel species belonging to these genera. This was partially attributed to young scientists wishing to publish papers rapidly and sometimes classifying organisms to existing genera rather than undertaking the additional effort of describing a new genus. Time and funding become an issue for the types of studies needed to characterize a new genus. At the meeting, J. Chun, a member of the subcommittee, also discussed the recent retraction of a species originally published under Flavobacteria. He felt that PhD and post-doc students are under publication and time pressures. His discussion highlights the common problem of the "race against time" to publish a novel organism. This time constraint on research can lead to misclassification.

The theory that many Flavobacteria have been misclassified, whether due to outdated techniques or rushed science, is echoed in the reclassification of [*Flavobacterium*] *salegens* to a new genus *Salegentibacter salegens* gen. nov., comb. nov. (McCammon & Bowman, 2000). McCammon and Bowman even cite Bernardet's sentiment in their

introduction: “Until recently the nomenclature status of the genus was heterogeneous and confused (Holmes *et al.,* 1984; Bernardet *et al.*, 1996).” It was on the basis of the 16S rRNA sequence, which showed *[F.] salegens* appearing on a distinct branch from its previous genus *Flavobacterium*. This allowed McCammon and Bowman to make the novel genus assertion. As genome data and new technology become more widely available, it will be easier for scientists to create appropriate taxa for misclassified species.

In conclusion, new methodologies, particularly the wide accessibility of genomic data as well as the ability to compare genomic orthology accurately have made bacterial taxonomy more accessible to researchers. These new methods, however, should also be used in conjunction with the more “traditional” phenotype-centered methods to paint an accurate picture of the identity of a new species (Tindall *et al.*, 2010). Additionally, more attention must be paid to taxa above species rank to avoid the problem currently experienced with Flavobacteria being too heterogeneous for ANI comparison and the potential of being paraphyletic. There currently is no criterion for a level of divergence that would lead to creation of new genera or families. More dedicated senior scientists are needed in the effort to strictly standardize bacterial taxonomy and establish accurate genera to classify the wide variety of microbes waiting to be discovered. There is also the need reclassify those which are misplaced with new technology or the increased availability of genomic data (McCammon & Bowman, 2000).

Works Cited

Adékambi, T., Shinnick, T., 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,** 1807-1814.

Ali, Z., Cousin, S., Frühling, A., Brambilla, E., Schumann, P., Yang, Y., & Stackebrandt, E. (2009). *Flavobacterium rivuli* sp. nov., *Flavobacterium subsaxonicum* sp. nov., *Flavobacterium swingsii* sp. nov. and *Flavobacterium reichenbachii* sp. nov., isolated from a hard water rivulet. *Int J Syst Evol Microbiol* **59**, 2610-2617.

Auch, A.F., von Jan, M., Klenk, H.P., & Göker, M. (2010). Digital DNA-DNA Hybridization for microbial species delineation by means of genome-to-genome sequence comparison. *Standards in Genomic Sciences* **2**, 117-134.

Aziz, R.K., Bartels, D., Best, A.A., DeJongh, M., Disz, T., Edwards, R.A., Formsma, K., Gerdes, S., Glass, E.M., Kubal, M. & other others (2008). The RAST server: rapid annotations using subsystems

Bernardet, J.F., & Bowman, J.P. (1994). Genus I. Flavobacterium. In *Bergey's Manual of Determinative Bacteriology,* pp. 112-153. Edited by J.G. Holt. Philadelphia, PA: Lippincott, Williams, & Wilkins.

Bernardet, J.F., & Bowman, J.P. (2013). International Committee on Systematics of Prokaryotes, Subcommitee on the taxonomy of *Flavobacterium* and *Cytophaga*-like bacteria minutes of the meetings, 7 September 2011, Sapporo, Japan. *Int J Syst Evol Microbiol* **63**, 2752-2754.

Bierbaum, G., Sahl, H.G. (2009). Lantibiotics: mode of action, biosynthesis and bioengineering. *Curr Pharm Biotechnol* **10**, 2-18.

Buonaccorsi, V.P., Boyle, M.D., Grove, D., Praul, C., Sakk, E., Stuart, A., Tobin, T., Hosier, J., Carney, S.L., & others. (2011). GCAT-SEEKquence: genome consortium for active teaching of undergraduates through increased faculty access to next-generation sequencing data. *CBE Life Sci Educ* **10**, 342-345.

Buonaccorsi, V.P., Peterson, M., Lamendella, G., Newman, J.D., Trun, N., Tobin, T., Aguilar, A., Hunt, A., Praul, C. & others. (2014). Vision and change through the genome consortium for active teaching using next-generation sequencing (GCAT-SEEK). *CBE Life Sci Educ* **13**, 1-2.

Clarridge, J.E. (2004). Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases. *Clin. Microbiol. Rev.* **17**, 840-862.

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

Good, C., Davidson, J., Wiens, G.D., Welch, T.J., & Summerfelt, S. (2015). Flavobacterium branchiophilum and F. succinicans associated with bacterial gill disease in rainbow trout Oncohynchus mykiss (Walbaum) in water recirculation aquaculture systems. *J Fish Dis* **38**, 409-413.

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 Sys Evol Microbiol* **57**, 81-91.

Hug, L.A., Baker, B.J., Anantharaman, K., Brown, C.T., Probst, A.J., Castelle, C.J., Butterfield, C.N., Hernsdorf, A.W., Amano, Y., Ise, K., Suzuki, Y., Dudek, N., Relman, D.A., Finstad, K.M., Amundson, R., Thomas, B.C., & Banfield, J.F. (2016). A new view of the tree of life. *Nature Microbiol* **48**, 1-6.

Jain, N.K., & Roy, I. (2009). Effect of trehalose on protein structure. *Protein Sci* **18**, 24-36.

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.

Klenk, H.P., & Göker, M. (2010). En route to a genome-based classification of *Archaea* and *Bacteria*? *Syst Appl Microbiol* **33**, 175-182.

Konstantinidis, K.T. & Tiedje, J.M. (2005). Towards a genome-based taxonomy for prokaryotes. *J Bacteriol* **187**, 6258-6264.

Konstantinidis, K.T., Ramette, A., & Tiedje, J.M. (2006). The bacterial species definition in the genomic era. *Philo Trans R Soc Lond B Biol Sci* **361**, 1929-1940.

Lapage, S., P., Sneath, P., A., Lessel, E., F., Skerman, V., D., Seeliger, H., R., & Clark, W., A. (1992). *International Code of Nomenclature of Bacteria,* 4th revision. Washington, D.C.: ASM Press.

Loch, T.P., & Faisal, M. (2014). Deciphering the biodiversity of fish-pathogenic *Flavobacterium* spp. Recovered from the Great Lakes basin. *Dis Aquat Organ* **112**, 45-57.

McCammon, S.A., & Bowman, J.P. (2000). Taxonomy of Antarctic *Flavobacterium* species: description of *Flavobacterium gillisiae* sp. nov., *Flavobacterium tegetincola*

sp. nov. and *Flavobacterium xanthum* sp. nov., nom. rev. and reclassification of [*Flavobacterium*] *salegens* as *Salegentibacter salegens* gen. nov., comb. nov. *Int J Sys Evol Microbiol* **50**, 1055-1063.

Meier-Kolthoff, J.P., Auch, A.F., Klenk, H.P., & Goker, M. (2013). Genome sequence-based species delimitation with confidence intervals and improved distance functions. *BMC Bioinformatics* **14**, 60.

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

Stahl, D.A., & Tiedje, J. (2001). Microbial ecology and genomics: A crossroads of opportunity. The American Academy of Microbiology colloquium, 23-25 February 2001, Singer Island, FL.

Tamura, K., Stecher, G., Peterson, D., Filipski, A., & Kumar, S. (2013). MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Mol Biol Evol* **30**, 2725-2729.

Tindall, B.J., Rosselló-Móra, R., Busse, H.J., Ludwig, W., & Kämpfer. (2010). Notes on the characterization of prokaryote strains for taxonomic purposes. *Int J Sys Evol Microbiol* **60**, 249-266.

Tindall, B.J. (2015). Updating Rule 15 of the international code of nomenclature of bacteria. *Int J Sys Evol Microbiol* **65**, 2766-2768.

Weeks, O.B. (1955). *Flavobacterium aquatile* (Frankland and Frankland) Bergey *et al.*, type species of the genus *Flavobacterium*. *J Bacteriol* **69**, 649-658.

Wheat, P., F. (2001). Development of antimicrobial susceptibility testing methodology. *J Antimicrob Chemother* **48**, 1-4.

Wolf, Y.I., & Koonin, E.V. (2012). A tight link between orthologs and bidirectional best hits in bacterial and archaeal genomes. *Genome Biol Evol* **4**, 1286-1294.

Zhang, D.C., Wang, H.X., Liu, H.C., Dong X.Z., & Zhou, P.J. (2006). Flavobacterium glaciei sp. nov., a novel psychrophilic bacterium isolated from the China No.1 glacier. *Int J Sys Evol Microbiol* **56**, 2921-2925.