Chryseobacterium haifense : A Genomic Report

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

by

Tom Sontag

Lycoming College

September, 2013

Approved by:

Jeffrey D. Newman

Jeffrey D. Newman (Signature)

Edward G. Gabriel

(Signature)

Eileen M. Peluso

(Signature)

(Signature)

Mary E. Morrison

Abstract:

The phylogenetic position of a number of bacteria within the family Flavobacteriaceae has been questioned. To address the question, the whole genomes of several organisms were sequenced, and this project is focused on *Chryseobacterium haifense*. The advances in next generation sequencing (NGS) technologies have caused a decrease in cost for whole genome sequencing. This decreased cost has led to more genomes being sequenced and in the process has caused a large demand for bioinformatics tools to handle the genomic data. To analyze the genomic data, the 930,000 reads were assembled in several steps, using several different software packages to refine the assembly to fewer than 700 contiguous sequences. Automated annotation using the Rapid Annotation using Subsystem Technologies (RAST) server identified the organism's genes and known pathways which were compared to its phenotypes. The Reciprocal Orthology Score Average (ROSA) genomic similarity calculator showed that *Chryseobacterium haifense* is as different from "true" Chryseobacteria as other separate genera are which has led to the conclusion that *Chryseobacterium haifense* does not belong within the *Chryseobacterium* genus.

Introduction:

The *Chryseobacterium* genus was established from the genus *Flavobacterium* by Vandamme *et al.* (1994). Chryseobacteria were described as gram-negative, nonmotile, non-spore-forming rods; and Vandamme designated *Chryseobacterium gleum* as the type species for the genus (Vandamme *et al.* 1994). The original description of *C. gleum* was as *Flavobacterium gleum* and a color change was reported after treatment with 20% KOH (Holmes *et al.* 1984), which was described by Reichenbach in 1980 to detect the presence of flexirubin pigments (Bernardet 2002). The description of the genus *Chryseobacterium* included the presence of flexirubin (Vandamme *et al.* 1994).

Chryseobacterium haifense was originally isolated from raw cow's milk. The cells are aerobic, Gram-negative, non-motile rods, occurring singly, in pairs or in short chains. They grow from 4-41°C, at 0-2.5% NaCl and at a pH range from 6.5 to pH 10.5. The most abundant fatty acids are: 15:0 iso (41.6%), 15:0 anteiso (16.6%) and 17:0 iso 3-OH (10.3%). This organism was classified as a *Chryseobacterium* because of its phenotypic similarity and its 96.7% 16S rRNA similiarity to *C. hispanicum* (Hantsis-Zacharov and Halpern 2007).

A very weak color change was reported when *C. haifense* was treated with 20% KOH (Hantsis-Zacharov and Halpern 2007). The weak color change in *C. haifense* was not reproducible by the Newman Lab (Data not published). The colony color of *C. haifense* was reported to be yellow when grown in presence of light but a cream color in the absence of light (Hantsis-Zacharov and Halpern 2007). One goal of this Genomic analysis is to provide insight into the production of flexirubin pigments.

Genome sequencing has become a more common practice due to decreased cost. Sequencing technology relying on highly-parallel optical sensing of DNA synthesis reactions has advanced significantly within the last decade (Bragg *et al.* 2013). The advances in the next-generation sequencing (NGS) have shifted whole genome sequencing from larger facilities to small research labs (Kisand and Lettieri 2013). Not all sequencing technology has relied on intense optical sensing of polymerization reactions. Ion Torrent Technology (Life Technologies) has sought to reduce the dependence on expensive photon sensors (Rothberg *et al.* 2011). Ion Torrent Technology sequencing relies on sensors designed to detect hydrogen ions released by DNA polymerase during DNA synthesis (Rothberg *et al.* 2011). Rothberg *et al.* reported that the sequencing accuracy of the ion based approach was similar to the photon based results (2011). With advances in both optical and proton sensors the cost of sequencing is likely to continue decreasing. The National Human Genome Research Institute has calculated the decreased cost per megabase from 2001 to 2013 (http://www.genome.gov/images/content/cost_per_megabase.jpg). The rate of decrease in cost per megabase is larger than Moore's Law, which is based on observations of advances in computing hardware over the history of computers.

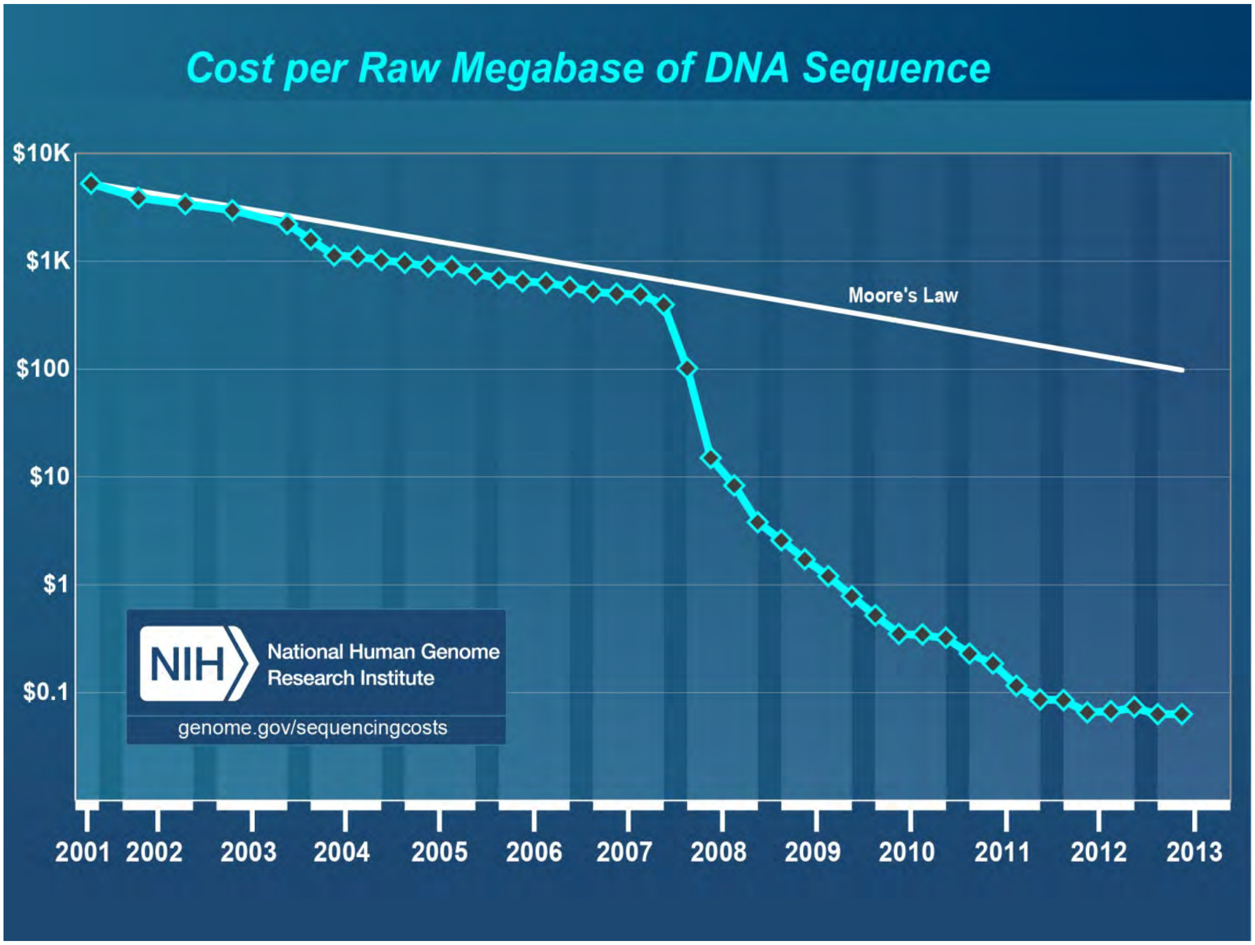


Figure 1: The Cost of Genome Sequencing as reported by NHGR I

These advances in NGS have made it possible for smaller research institutions to sequence whole genomes (Salzberg *et al.* 2008). Pagani *et al.* indicated almost a twofold increase in total number of genomic sequences in the Genome Online Database (GOLD) since the 2009 review, which was attributed to advances in NGS and decrease in cost (2011). Markowitz *et al.* confirmed that the dramatic decrease in sequencing cost led to a considerable increase in new genome data sets (2012). Grigoriev *et al.* predicted that the total number of sequenced genomes would drastically increase with the decreased cost (2012). Such advances have allowed Lycoming College to sequence multiple organisms' genomes.

Lycoming College was able to obtain the whole genome sequencing of *Chryseobacterium haifense* through the Genome Consortium for Active Teaching (GCAT) SEEKquence, or GCAT-SEEK. GCAT-SEEK was founded to provide students the ability to learn the cutting edge of biology. Due to decreased cost in NGS more graduate level research labs and corporate labs have begun using NGS. In order for students to be competitive for jobs and graduate programs it is important for them to have experience with NGS. GCAT-SEEK provides a means for faculty to offer low cost NGS projects to students (Buonaccorsi *et al*. 2011). The GCAT-SEEK program has earned funding from the Howard Hughes Medical Institute (http://www.hhmi.org/news/hhmicolleges20120524.html) and from the National Science Foundation (**# DBI-1248096** and **# DBI-1061893**).

Advances in NGS can be attributed to increasing interest in genomics as a method of biomedical research. Following the publication of the human genome in 2001, Julian Davies argued that sequencing the human microbiota, or the collection of bacteria that inhabit the various niches within humans, would further benefit biomedical research (NIH HMP Working Group *et al.* 2009). The Human Microbiome Project (HMP) was a five year, National Institutes of Health (NIH) initiative that sought to advance biomedical research through sequencing samples from various regions of the human body to identify organisms and gene functions present (NIH HMP Working Group *et al.* 2009). The HMP predicted the addition of 900 bacterial genome sequences to the public database (NIH HMP Working Group *et al.* 2009). As of July 2009, more than 500 bacterial genomes were being sequenced by various facilities (NIH HMP Working Group *et al.* 2009). Human microbiome groups from around the world launched an International Human Microbiome Consortium (IHMC), which predicted that more than 1,000 genomes will be sequenced among the groups (Human Microbiome Jumpstart Reference Strains

Consortium *et al.* 2010). Since the foundation of the HMP about 5,000 bacterial strains have been isolated from the human body and been submitted for whole genome sequencing (Fodor *et al.* 2012). One of the roadblocks faced by the HMP is determining which group of taxa should be prioritized for genomic sequencing (Fodor *et al.* 2012). Fodor *et al.* report that 97% 16S rRNA similarity generally indicates “same” species, but it does not mean that the genomic variation confirms (2012). The high priority taxa were determined by selecting organisms that had less than 90% identity to GOLD-Human or HMP strains (Fodor *et al.* 2012). Genomic sequencing of clinical microbiological specimens has increased the ability to study both cultivatable and uncultivatable bacteria (Conlan *et al.* 2012). Direct sequencing of microbial communities, metagenomics, provides information for uncultivatable bacteria (Conlan *et al.* 2012). Genomic sequencing is valuable for detecting multiple organisms within a single sample, which is not easily done in culture (Conlan *et al.* 2012). It was originally predicted that humans would contain a large core microbiome at the species level, but after sequencing samples from many individuals that does not seem to be the case. However, there do seem to be similarities at higher-order taxonomic levels (Hamady and Knight 2009). Many bacterial species with medical implications have been sequenced due to the HMP. The large quantity of bacteria sequenced due to the HMP led to increased technologies and tools. The HMP is also responsible for depositing a large number of bacterial genomes into genomic databases, which have had metabolic gene pathways mapped. These mapped pathways are sometimes analogous, especially essential pathways, to pathways observed in other bacteria, such as *Chryseobacterium haifense. De novo* metabolic pathways did not need to be mapped for *C. haifense* because many of the genes were related to known mapped genes.

The data provided by NGS is not contained in one continuous DNA sequence but rather in many tiny reads which can be assembled into fewer, larger contiguous sequences (contigs) based on overlaps. Bacterial DNA is contained in one circular chromosome, but NGS provides millions of short reads. The first step in analyzing the data is assembling the reads into contigs or ideally the continuous bacterial chromosome. There are various software packages available to assemble the NGS reads (Zhang *et al.* 2011). Most of these packages run on the Linux operating system, but our experience with and access to Linux has been a limiting factor. One Windows-based package, NextGENe (SoftGenetics, State College, PA), is available to us on the GCAT-SEEK server. Because it is running on a powerful server with a large amount (64 Gb) of RAM, the millions of reads from NGS can be assembled into fewer, larger contigs. The small number of contigs from NextGENe (SoftGenetics) can be further assembled into larger contigs (supercontigs) through the software Geneious Pro (Biomatters) running on a standard personal computer. Artemis (Carver *et al.* 2012) is a genomic tool that aids in the assembly process by allowing reordering and manipulation of contigs and supercontigs based on comparison to a reference organism. These assemblies can be deposited into genomic databases.

As more genome sequences become available, genomic databases are required to store large quantities of sequence data. Several databases have been developed, such as the Integrated Microbial Genome database (IMG), the NCBI genome database, the Genomes Online Database (GOLD), and the Comprehensive Microbial Resource (Langille *et al.* 2012), and each database provides access to different sets of tools. The Department of Energy (DOE) created the Joint Genome Institute (JGI) system to study bioenergy and environmental applications using genomics (Grigoriev *et al.* 2011). Various databases and analytical systems have been developed by JGI. One such system is IMG, which allows the genomic study of microbes such as annotation, analysis and microbial gene distribution (Grigoriev *et al.* 2011; Markowitz *et al.* 2012). The IMG system has integrated public draft and complete microbial genomes from the three domains of life, including a large number of plasmids and viruses (Markowitz *et al.* 2012). GOLD allows continuous monitoring of metagenome and genome sequences, along with their metadata (Pagani *et al.* 2011). The J. Craig Venter Institute, a non-profit genomics research institute, offers a genomic database, Comprehensive Microbial Resourse (CMR), The CMR is a free website dedicated to providing information for all publicly available, complete prokaryotic genomes (http://cmr.jcvi.org/tigr-scripts/CMR/CmrHomePage.cgi).

Genomic databases allow for more than just storage of genomic sequences. Genome sequence information such as: organization into chromosomal replicons (finished genomes) or scaffolds and contigs (draft genomes); predicted protein-coding sequences (CDSs); some RNA-coding genes; and protein product names, is recorded in IMG (Markowitz *et al.* 2012). Genomic and metagenomic data along with their metadata greatly increases value and can lead to more accurate comparative analyses and biological interpretations of the sequence data, for this reason the GOLD metadata is incorporated into IMG (Pagani *et al.* 2011). The metadata from GOLD is associated with each IMG genome through a data integration pipeline (Markowitz *et al.* 2012). IMG also includes information such as: CRISPR repeats, signal peptides, transmembrane helices, and RNAs (Markowitz *et al.* 2012). Functional categories and clusters of orthologous genese (COGs) , or an attempt to phylogenetically classify hypothetical and known proteins encoded by genomes based on evolutionary relations (Koonin 2003), are used to generate annotations of protein family and domain characterizations through IMG. (Markowitz *et al.* 2012). IMG provides lists of potential paralogs and orthologs for each gene based on NCBI BLASTp sequence similarity (Markowitz *et al.* 2012). Observing closely related genes is crucial in the comparison of closely related organisms. IMG also offers an “Expert Review” version, which permits scientists to curate and review functional annotations of microbial genomes (Markowitz *et al.* 2012).

After depositing assembled genomic sequences into a database, the process of identifying gene locations within the genome can begin. The Rapid Annotation using Subsystems Technology (RAST) server is a fully automated service for annotating bacterial and archaeal genomes (Aziz *et al.* 2008) within a period of a few hours to a few days, depending on the server load. The RAST server annotates bacterial and archaeal genomes by providing initial gene calls, gene functions and metabolic reconstructions (Aziz *et al.* 2008). A subsystem is defined as a set of functional roles and the subsystem is expanded by connecting the functional roles to specific genes in specific genomes (Aziz *et al.* 2008). As of 2008 there were over 600 reported subsystems in which functions for greater than 500,000 protein-encoding genes in over 500 bacterial and archaeal genomes (Aziz *et al.* 2008)., while 4,873 COGs included only 136,711 proteins from 50 bacterial genomes and 13 archaeal genomes (Koonin 2003). COGs are generated based on the notion that at least three proteins from distant genomes that are more similar to each other than any proteins from the same genome belong to an orthologous cluster (Tatusov *et al.* 2000). Genes that perform similar roles but are more distantly related, or paralogs, are not accounted for in COGS, but in subsystems both orthologs and paralogs are incorporated because the genes are organized based on their function rather than their evolutionary relationship.

Annotated genomes can provide insight into taxonomic relationships. In 2002, an ad hoc committee re-evaluated the species definition for bacteria (Stackebrandt *et al.* 2002). DNA-DNA re-association, also known as DNA-DNA hybridization (DDH), and 16S rDNA analyses were considered to methods of great promise (Stackebrandt *et al.* 2002). Stackebrandt *et al.* also agreed that sequencing of housekeeping genes and DNA profiling also offer promise to defining bacterial species (2002). The human genome had only been published a year earlier and whole genome sequencing costs were still expensive. The ad hoc committee has not met since 2002, therefore the methods used to determine the bacterial species definition is outdated.

A definition for bacterial species in the genomic era is desired but not yet fully obtained. Insights into intra-species diversity and a new bacterial species definition is possible due to genomic sequencing (Konstantinidis *et al.* 2005). The DDH standard (70%), used in the traditional definition of bacterial species, is universally applicable; however, it is criticized for being difficult to implement (Konstantinidis *et al.* 2005; Auch *et al.* 2010; Richter and Rosselló-Móra 2009). The scientific community finds the species definition based on 70% DDH standard lacking (Konstantinidis *et al.* 2005). DDH is time-consuming and labor intensive, which has led to its use by only a select few specialized laboratories (Auch *et al.* 2010). Reliable diagnosis of infectious disease agents and intellectual property rights are impacted by poor standards (Konstantinidis *et al.* 2005). DDH measures the efficiency of the hybridization of the DNA, not the sequence identity (Konstantinidis *et al.* 2005). Understanding genetic differences among closely related bacteria is critical in redefining the bacterial species definition in a genomic era (Konstantinidis *et al.* 2005). The genomic era definition requires high resolution characterization of many bacterial groups (Konstantinidis *et al.* 2005). ANI is the average nucleotide identity of the total genomic sequence shared between two bacterial strains (as measured by a Basic Local Alignment Search Tool or BLAST search result above a certain threshold) and is considered a sensitive method for identifying evolutionary-relatedness between closely related bacteria (Konstantinidis *et al.* 2005). It has been reported that 70% DDH corresponds to 95% ANI (Konstantinidis and Tiedje 2007). Another method used to attempt sequence based comparisons of closely related bacteria is the genome-to-genome distances (GGD) (Auch *et al.* 2010). Auch *et al.* believes that average nucleotide identity along with GGD could be used to recreate DDH *in silico* (2010). ANI values are based on pairwise alignment of genome stretches (Richter and Rosselló-Móra 2009). Tang reported that clear-cut genetic boundaries do exist between bacterial lineages which can be detected by genomic analysis (Tang *et al*. 2013).

Another whole genome-based metric called Average Amino Acid Identity (AAI) is based on amino acid sequences rather than nucleotides (Konstantinidis *et al.* 2005). This AAI value can be used to compare whole genome-level similarity between organisms.

There are two questions to be answered when comparing genome similarity: how similar are the shared genes and what percentage of genes is shared among the genomes. ANI assesses similarities between the nucleotides (ATCG) of the sequences by using BLASTN, which uses high-scoring pairs, to identify similar nucleotide sequences. The pitfall, to using high-scoring pairs to identify nucleotide sequence similarities between the organisms, is that protein families that are distantly related are not included. Proteins are made of amino acids. Amino acids are encoded by triplet codes of nucleotides (codons), which is specified by the Genetic Code. There are 64 codons of nucleotides: but only 20 amino acids, which means that multiple codes may be used to specify the same amino acid. For example AAA and AAG both encode the amino acid Phenylalanine. This is important because it allows nucleotides to mutate without changing the amino acid. Amino acids are also classified based on their behaviors in water and their charges. If a mutation causes a positive amino acid to be changed to another positively charged amino acid then the mutation is said to be conservative. The function may change slightly but it will still perform a similar job. As these conservative mutations accumulate the nucleotide sequences tend to mutate rather rapidly compared to the protein functions. Amino acid sequences change slower than nucleotide sequences, because changes in nucleotides do not always cause changes in the amino acid sequences due to the Wobble Hypothesis which states that the third base in the triplet code can be altered but still result in the same amino acid.

ANI is accurate for species level comparisons, but is not adequate for higher level taxonomic levels. The AAI calculation is based on the protein-length –weighted pairwise identity of orthologous proteins, as determined by bidirectional best hits (BBH) between a reference genome and up to ten comparison genomes. Genes that encode proteins that perform the same function in different organisms, such as human and chimp α-globin (related to hemoglobin), are said to be orthologous. A bidirectional best hit occurs when a certain gene in the reference genome is matched to a gene in a comparison genome and vice versa.

ANI and AAI measure the similarity of shared genes and proteins respectively. And %BBH addresses the percentage of genes that are shared. The Orthology Score (OS) is a calculation based on AAI and %BBH. The AAI value was found to be most important and thus is squared, in order to simulate DDH. The %BBH varies in reciprocal comparisons if the genomes being compared have significantly different sizes. Since %BBH is a factor in calculating OS, OS will also vary. Reciprocal comparisons can be averaged to calculate the Reciprocal Orthology Score Average (ROSA). The goal of ROSA is to provide a single metric that calculates genome similarity based on AAI & %BBH that yields values similar to DDH. The development of the ROSA metric is outside the scope of this project. ROSA does, however, provide a means to determine the phylogenetic relationship of *Chryseobacterium haifense.*

Methods:

Chryseobacterium haifense was obtained from the DSMZ.

<u>Genomic DNA Isolation and Whole Genome Sequencing</u>

The genomic DNA was extracted from *Chryseobacterium haifense* using Qiagen Blood and Tissue kit in the 2011 Molecular Biology class. The genomic DNA was then sequenced, using Ion Torrent Technology, by Penn State’s Genomics Core.

<u>Initial Assembly of Ion Torrent Reads</u>

NextGENe v2.17 (Softgenetics) was used for the primary assembly, using the following options: application type: de novo; instrument: Roche 454; Assembly method type: Greedy.

<u>Geneious Assembly Optimization</u>

The contigs provided by the NextGENe (SoftGenetics) package were imported in FASTA format to Geneious 5.6 (Biomatters) for secondary assembly. A secondary analysis was completed to assemble the NextGENe contigs into supercontigs, or an assemblage of overlapping contigs. The purpose was to assemble as many contigs into as few supercontigs as possible to provide information for an accurate analysis. A *de novo* assembly of the contigs, without base pair trimming, was completed following the Geneious 5.6 (Biomatters) suggestions. Because many discrepancies were noted near the ends of contigs, assemblies were repeated in which different numbers of bases were trimmed from each end.

Editing Ambiguous Bases

Ambiguities in contigs were located where reads overlapped. Each contig was individually selected and viewed. Ambiguities were identified using the Geneious command: find next disagreement. Ambiguities were corrected based on position, which correlates to sequence accuracy.

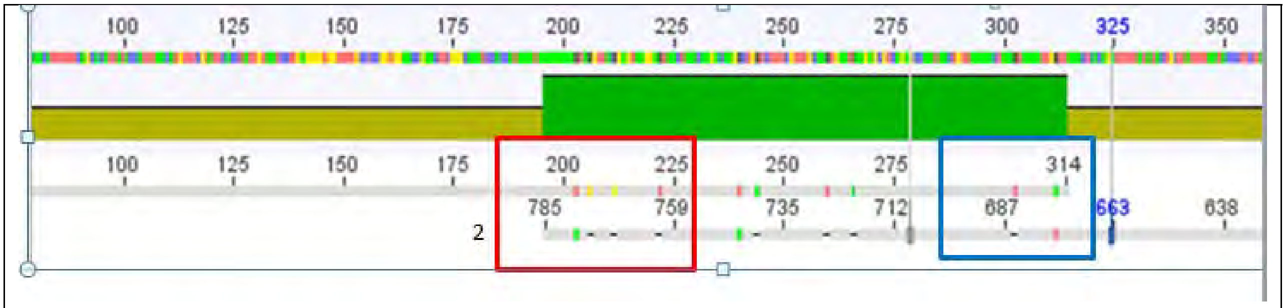


Figure 2 Screenshot From Geneious Two contigs are aligned for overlaps with ambiguities present.

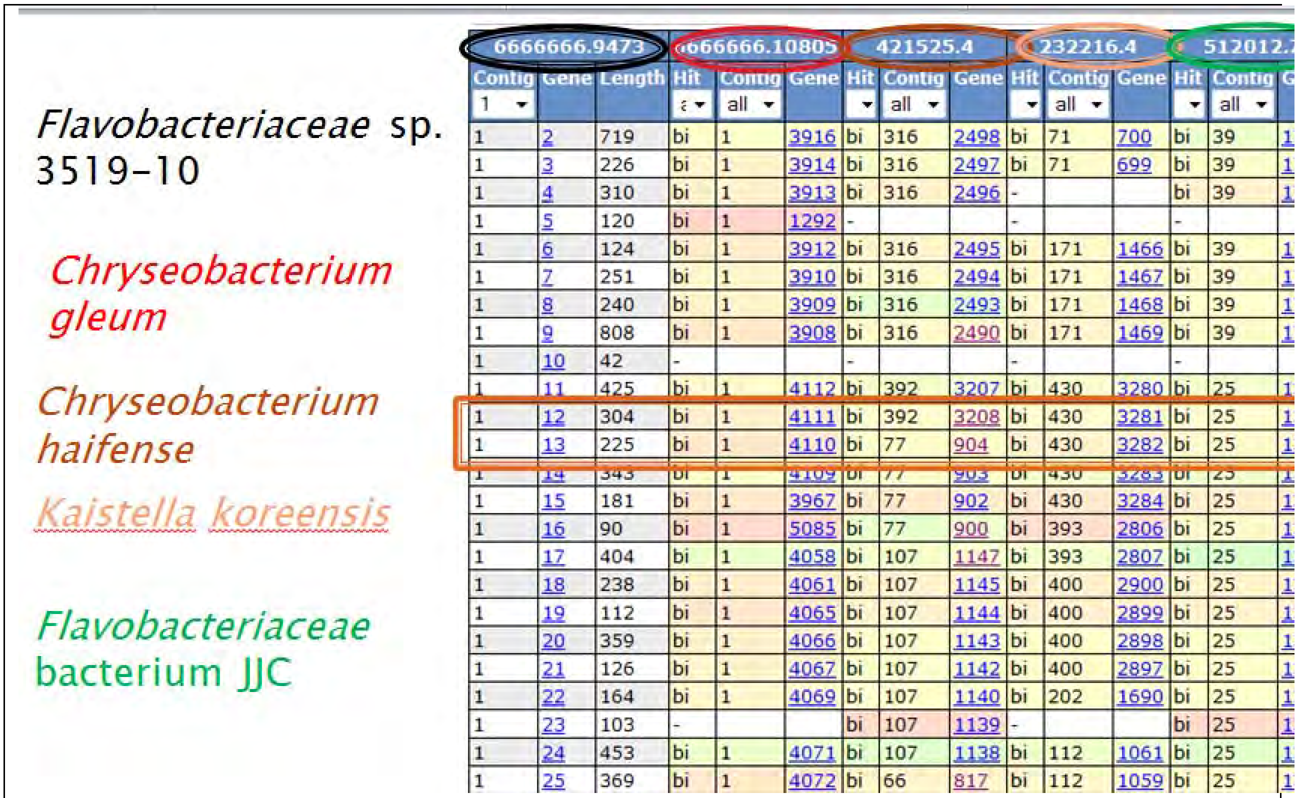
More Central bases in a contig were selected over less central bases on the joining contig. A red and a blue box were added to a Geneious screenshot (Fig. 2) to provide a visual aid. The ambiguous bases, presented in the red box, were corrected to the top contig because the bases in question were more centrally located in the contig. The ambiguous bases, presented in the blue box, were corrected based on the lower contig because the bases were more central than the bases on the top contig. This method was continued for all of the supercontigs created in the Geneious assembly.

RAST Annotation

The sequence data, after being edited for ambiguity, was uploaded to the Rapid Annotation using Subsystem Technologies for automated annotation (http://rast.nmpdr.org/) (Aziz *et al.* 2008).

Contig Reordering through Artemis and RAST

The ambiguity-free secondary assembly was uploaded to RAST and Artemis. RAST was used to create a sequence based comparison between *C. haifense*, *C. gleum, C. koreense, Flavobacteriaceae* sp. JJC and *Flavobacteriaceae* sp. 3519-10. *Flavobacteriaceae* sp. 35-1910 was used as the reference because it had a finished genome. The sequence based comparison was used to identify contigs that may be adjacent to each other for potential gap closure.

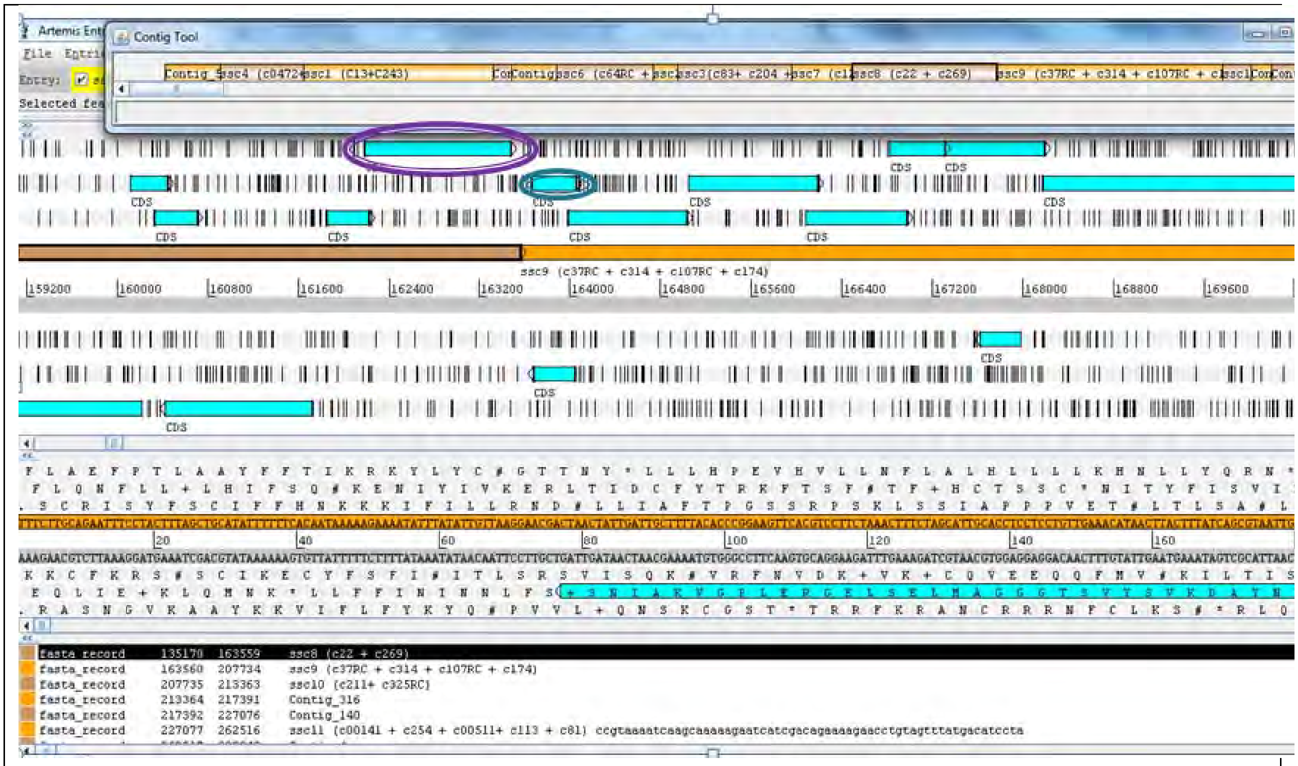

Flavobacteriaceae sp. 3519-10
Chryseobacterium gleum
Chryseobacterium haifense
Kaistella koreensis
Flavobacteriaceae bacterium JJC

6666666.9473			666666.10805			421525.4			232216.4			512012.		
Contig	Gene	Length	Hit	Contig	Gene	Hit	Contig	Gene	Hit	Contig	Gene	Hit	Contig	G
1 ▾			▾	all ▾		▾	all ▾		▾	all ▾		▾	all ▾	
1	2	719	bi	1	3916	bi	316	2498	bi	71	700	bi	39	1
1	3	226	bi	1	3914	bi	316	2497	bi	71	699	bi	39	1
1	4	310	bi	1	3913	bi	316	2496	-			bi	39	1
1	5	120	bi	1	1292	-			-			-		
1	6	124	bi	1	3912	bi	316	2495	bi	171	1466	bi	39	1
1	7	251	bi	1	3910	bi	316	2494	bi	171	1467	bi	39	1
1	8	240	bi	1	3909	bi	316	2493	bi	171	1468	bi	39	1
1	9	808	bi	1	3908	bi	316	2490	bi	171	1469	bi	39	1
1	10	42	-			-			-			-		
1	11	425	bi	1	4112	bi	392	3207	bi	430	3280	bi	25	1
1	12	304	bi	1	4111	bi	392	3208	bi	430	3281	bi	25	1
1	13	225	bi	1	4110	bi	77	904	bi	430	3282	bi	25	1
1	14	343	bi	1	4109	bi	77	903	bi	430	3283	bi	25	1
1	15	181	bi	1	3967	bi	77	902	bi	430	3284	bi	25	1
1	16	90	bi	1	5085	bi	77	900	bi	393	2806	bi	25	1
1	17	404	bi	1	4058	bi	107	1147	bi	393	2807	bi	25	1
1	18	238	bi	1	4061	bi	107	1145	bi	400	2900	bi	25	1
1	19	112	bi	1	4065	bi	107	1144	bi	400	2899	bi	25	1
1	20	359	bi	1	4066	bi	107	1143	bi	400	2898	bi	25	1
1	21	126	bi	1	4067	bi	107	1142	bi	400	2897	bi	25	1
1	22	164	bi	1	4069	bi	107	1140	bi	202	1690	bi	25	1
1	23	103	-			bi	107	1139	-			bi	25	1
1	24	453	bi	1	4071	bi	107	1138	bi	112	1061	bi	25	1
1	25	369	bi	1	4072	bi	66	817	bi	112	1059	bi	25	1

Figure 3: RAST Sequence based Comparison

The gene locations were predicted based on gene order in closely related organisms (Fig. 3). If two genes were located adjacent to one another in the reference genome and the other organisms, then it was hypothesized that the genes were adjacent to one another in the *C. haifense* genome. For example, Genes 12 and 13 (Orange box) were adjacent in the reference organism and the corresponding genes were located side by side in the other organisms, therefore, it was hypothesized that genes 3208 and gene 904 would be adjacent to one another and thus contig 392 and 77 would belong adjacent to one another within the *C. haifense* chromosome (Orange box).

After hypotheses were formed about gene location, within the chromosome, the contigs, which had been ordered by length, were reordered using the Artemis contig reordering tool. The tertiary assembly was re-uploaded to RAST.


Artemis Ent
Contig Tool
File Entri
Entry:
Selected fea
Contig_5ssc4 (c0472+ssc1 (C13+C243)
CoContig ssc6 (c64RC + ssc ssc3(c83+ c204 +ssc7 (c1 ssc8 (c22 + c269)
ssc9 (c37RC + c314 + c107RC + c ssc1 Con Con
CDS
ssc9 (c37RC + c314 + c107RC + c174)
159200
160000
160800
161600
162400
163200
164000
164800
165600
166400
167200
168000
168800
169600
20
40
60
80
100
120
140
160

```
fasta_record    135170  163559   ssc8 (c22 + c269)
fasta_record    163560  207734   ssc9 (c37RC + c314 + c107RC + c174)
fasta_record    207735  213363   ssc10 (c211+ c325RC)
fasta_record    213364  217391   Contig_316
fasta_record    217392  227076   Contig_140
fasta_record    227077  262516   ssc11 (c00141 + c254 + c00511+ c113 + c81) ccgtaaaatcaagcaaaaagaatcatcgacagaaaagaacctgtagtttatgacatccta
```

Figure 4: Artemis Contig Reordering and Identified ORFs

Artemis was used to identify Open Reading Frames (ORFs). The ORFs were used to confirm or refute the hypotheses. The ORFs were identified using BLAST. If the ORFs on the two adjoining contigs encoded different regions of the same protein the hypothesis was confirmed, if not the hypothesis was refuted. A visual was provided as an example (Fig. 4). The ORFs on the edge of two adjoining contigs (Fig. 4 circles) were selected for BLAST searches because the orientation was correct. The BLAST searches would identify the genes encoded by the ORF, which was then used to determine whether there were gaps between the contigs that were missing.

Geneious 6.0 Assembly

An updated version of Geneious was released during the assembly and alignment project. The Geneious 5.6 assembled sequence data was imported into Geneious 6.0. A *de novo* assembly, with 50bp trimmed on the 3' and 5' regions, was completed following the manufacturer's instructions.

Phylogenetic Tree

A neighbor joining tree including *C. haifense* and closely related species was completed using MEGA5 with a bootstrap value of 1000 replications (Fig. 4).

ROSA

The Newman Lab has developed a new method to calculate AAI which has been implemented as a JavaScript program by Dr Eileen Peluso RAST was used to create 11 different sequence based comparisons with each organism as the reference. The comparisons were uploaded to the ROSA calculator (http://lycofs01.lycoming.edu/~newman/rosa/) and the results were recorded in Microsoft Excel.

Results:

Geneious Assembly Optimization:

The trimming of 50bp on the 3' and 5' provided the fewest contigs derived from the 2761 NextGene contigs (Table 1). The number of contigs produced and the number of unused contigs decreased with each bp trimming increase until the 50bp trimming with the exception of the 45 bp trimming method. The 45bp trimming method did not agree with the observed correlation. The number of supercontigs produced along with the number of unused reads increased significantly between the 50 and 60bp methods. The N50 Length for the 50bp trim assembly was 11,431bp. The max length was 41,158bp.

Table 1: Geneious 5.6 Assembly Report Data

Bp Trimmed	Supercontigs produced	Unused Contigs	Total Contigs Imported
0	504	1,062	2,761
15	436	701	2,761
25	415	644	2,761
35	389	609	2,761
45	528	739	2,761
50	385	576	2,761
60	571	942	2,761

Geneious 6.0 Assembly

The Geneious 6.0 assembly combined more reads and contigs from the Geneious 5.6 assembly to create fewer contigs (Table 2). The max length was 73,060bp and the N50 length was 18,906.

Table 2: Geneious 6.0 Assembly Report Data

BP Trimmed	Supercontigs Produced	Unused Contigs	Total Contigs Imported
50	152	408	935

Phylogenetic Tree

Four discrete clusters were observed based on 16S rRNA similarity (Fig. 5). The red cluster contains members of the Epilithonimonas genus. The orange cluster contains the type species for the Chryseobacteria genus, *Chryseobacterium gleum*, along with other members of the Chryseobacterium genus. *Chryseobacterium haifense* was not clustered with *C. gleum*, but rather with a few members of the Chryseobacteria genus along with the unnamed species, *Flavobacteriaceae* sp. JJC and *Flavobacteriaceae* sp. 3519-10 (Fig 5. Blue braces). The green cluster includes members of the *Flavobacterium* genus.

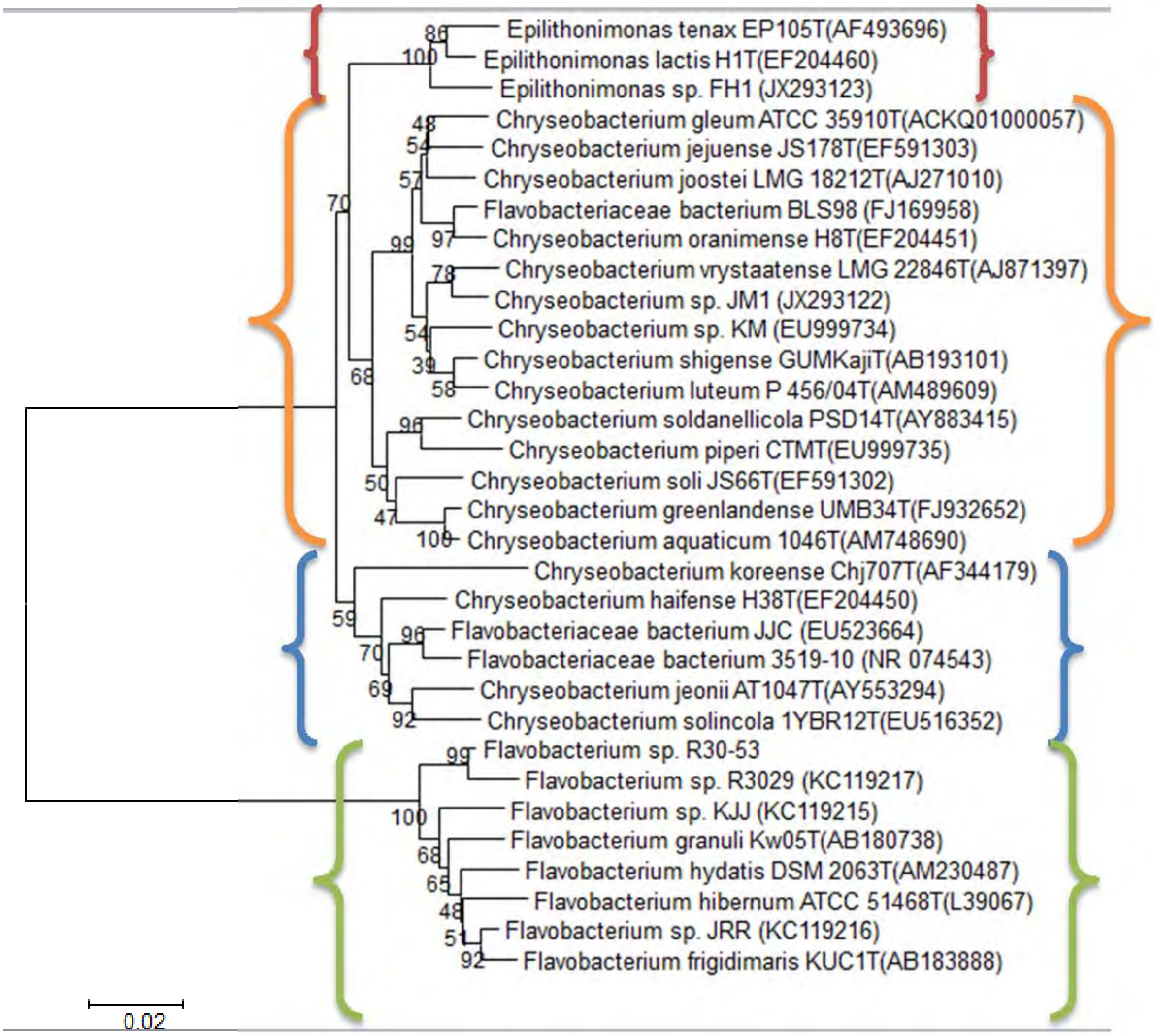

86
100
Epilithonimonas tenax EP105T(AF493696)
Epilithonimonas lactis H1T(EF204460)
Epilithonimonas sp. FH1 (JX293123)
48
54
57
70
99
97
Chryseobacterium gleum ATCC 35910T(ACKQ01000057)
Chryseobacterium jejuense JS178T(EF591303)
Chryseobacterium joostei LMG 18212T(AJ271010)
Flavobacteriaceae bacterium BLS98 (FJ169958)
Chryseobacterium oranimense H8T(EF204451)
78
Chryseobacterium vrystaatense LMG 22846T(AJ871397)
Chryseobacterium sp. JM1 (JX293122)
54
Chryseobacterium sp. KM (EU999734)
39
68
58
Chryseobacterium shigense GUMKajiT(AB193101)
Chryseobacterium luteum P 456/04T(AM489609)
96
Chryseobacterium soldanellicola PSD14T(AY883415)
Chryseobacterium piperi CTMT(EU999735)
50
Chryseobacterium soli JS66T(EF591302)
47
Chryseobacterium greenlandense UMB34T(FJ932652)
100
Chryseobacterium aquaticum 1046T(AM748690)
Chryseobacterium koreense Chj707T(AF344179)
59
Chryseobacterium haifense H38T(EF204450)
96
Flavobacteriaceae bacterium JJC (EU523664)
70
Flavobacteriaceae bacterium 3519-10 (NR 074543)
69
Chryseobacterium jeonii AT1047T(AY553294)
92
Chryseobacterium solincola 1YBR12T(EU516352)
99
Flavobacterium sp. R30-53
Flavobacterium sp. R3029 (KC119217)
100
Flavobacterium sp. KJJ (KC119215)
68
Flavobacterium granuli Kw05T(AB180738)
65
Flavobacterium hydatis DSM 2063T(AM230487)
48
Flavobacterium hibernum ATCC 51468T(L39067)
51
Flavobacterium sp. JRR (KC119216)
92
Flavobacterium frigidimaris KUC1T(AB183888)
0.02

Figure 5: Evolutionary relationships of taxa. The evolutionary history was inferred using the Neighbor-Joining method [1]. The optimal tree with the sum of branch length = 0.46788618 is shown. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (1000 replicates) are shown next to the branches [2]. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances were computed using the Maximum Composite Likelihood method [3] and are in the units of the number of base substitutions per site. The analysis involved 30 nucleotide sequences. All positions containing gaps and missing data were eliminated. There were a total of 1259 positions in the final dataset. Evolutionary analyses were conducted in MEGA5 [4]

Subsystems analysis

Genes with known functions were grouped within functional-based subsystems by the Rapid Annotation with Subsystems Technology Website (RAST) (Fig. 6). Some subsystems contained hundreds of genes, while other subsystems had only a few. Several subsystems did not contain any genes, suggesting that the specific function is not present in the organism.

Category
Subcategory Subsystem (number of genes)
Cofactors, Vitamins, Prosthetic Groups, Pigments (131)
Biotin (1) Biotin biosynthesis Experimental (1) Cofactors, Vitamins, Prosthetic Groups, Pigments - no subcategory (0) Quinone cofactors (8) Menaquinone and Phylloquinone Biosynthesis (8) Tetrapyrroles (15) Heme and Siroheme Biosynthesis (15) Riboflavin, FMN, FAD (31) Riboflavin, FMN and FAD metabolism in plants (12) riboflavin to FAD (5) Flavodoxin (4) Riboflavin, FMN and FAD metabolism (10) Fe-S clusters (0) Mycofactocin (0) Pyridoxine (10) Pyridoxin (Vitamin B6) Biosynthesis (10) NAD and NADP (10) NAD and NADP cofactor biosynthesis global (10) Coenzyme B (0) Folate and pterines (40) Folate biosynthesis cluster (8) Folate Biosynthesis (15) 5-FCL-like protein (17) Lipoic acid (1) Lipoic acid metabolism (1) Coenzyme F420 (0) Coenzyme M (0) Coenzyme A (15) Coenzyme A Biosynthesis cluster (5) Coenzyme A Biosynthesis (10)
Cell Wall and Capsule (77)
Capsular and extracellular polysacchrides (19) dTDP-rhamnose synthesis (7) Rhamnose containing glycans (12) Gram-Negative cell wall components (16) KDO2-Lipid A biosynthesis (16) Cell Wall and Capsule - no subcategory (41) Murein Hydrolases (5) Peptidoglycan Biosynthesis (20) UDP-N-acetylmuramate from Fructose-6-phosphate Biosynthesis (4) Recycling of Peptidoglycan Amino Sugars (1) Recycling of Peptidoglycan Amino Acids (5) Peptidoglycan biosynthesis--gjo (6)

Gram-Positive cell wall components (1) Teichuronic acid biosynthesis (1) Cell wall of Mycobacteria (0)
Virulence, Disease and Defense (85)
Adhesion (0) Toxins and superantigens (0) Bacteriocins, ribosomally synthesized antibacterial peptides (0) Resistance to antibiotics and toxic compounds (72) Copper homeostasis (16) Cobalt-zinc-cadmium resistance (40) Resistance to fluoroquinolones (4) Arsenic resistance (4) Beta-lactamase (8) Virulence, Disease and Defense - no subcategory (0) Detection (0) Invasion and intracellular resistance (13) Mycobacterium virulence operon involved in protein synthesis (SSU ribosomal proteins) (4) Mycobacterium virulence operon involved in DNA transcription (2) Mycobacterium virulence operon possibly involved in quinolinate biosynthesis (4) Mycobacterium virulence operon involved in protein synthesis (LSU ribosomal proteins) (3)
Potassium metabolism (8)
Potassium metabolism - no subcategory (8) Potassium homeostasis (8)
Photosynthesis (0)
Light-harvesting complexes (0) Photosynthesis - no subcategory (0) Electron transport and photophosphorylation (0)
Miscellaneous (24)
Conversion of Succinyl-CoA to Propionyl-CoA (0) Plant-Prokaryote DOE project (15) Iron-sulfur cluster assembly (15) Miscellaneous - no subcategory (9) Phosphoglycerate mutase protein family (1) DedA family of inner membrane proteins (1) Muconate lactonizing enzyme family (4) Broadly distributed proteins not in subsystems (3)
Phages, Prophages, Transposable elements, Plasmids (42)
Phage family-specific subsystems (0) Transposable elements (40) Conjugative transposon, Bacteroidales (40) Phages, Prophages (1) Phage introns (1) Phages, Prophages, Transposable elements, Plasmids - no subcategory (1) Integrons (1) Pathogenicity islands (0)

- Gene Transfer Agent (GTA) (0)
- Plasmid related functions (0)

Membrane Transport (60)

- Protein secretion system, Type II (0)
- ABC transporters (0)
- Protein secretion system, Type VII (Chaperone/Usher pathway, CU) (0)
- Protein translocation across cytoplasmic membrane (3)
  - Bacterial signal recognition particle (SRP) (2)
  - Twin-arginine translocation system (1)
- Protein secretion system, Type V (0)
- Protein secretion system, Type I (0)
- Cation transporters (16)
  - Magnesium transport (2)
  - Copper Transport System (14)
- Protein secretion system, Type III (0)
- Uni- Sym- and Antiporters (6)
  - Proton-dependent Peptide Transporters (5)
  - NhaA, NhaD and Sodium-dependent phosphate transporters (1)
- Membrane Transport - no subcategory (35)
  - Ton and Tol transport systems (35)
- TRAP transporters (0)
- Sugar Phosphotransferase Systems, PTS (0)
- Protein secretion system, Type VI (0)
- Protein secretion system, Type VIII (Extracellular nucleation/precipitation pathway, ENP) (0)
- Protein and nucleoprotein secretion system, Type IV (0)

Iron acquisition and metabolism (4)

- Siderophores (0)
- Iron acquisition and metabolism - no subcategory (4)
  - Hemin transport system (4)
- Iron transpot (0)

RNA Metabolism (102)

- RNA processing and modification (83)
  - RNA pseudouridine syntheses (5)
  - tRNA nucleotidyltransferase (1)
  - Methylthiotransferases (2)
  - Ribonucleases in Bacillus (2)
  - RNA processing and degradation, bacterial (5)
  - RNA methylation (8)
  - ATP-dependent RNA helicases, bacterial (2)
  - 16S rRNA modification within P site of ribosome (6)
  - tRNA modification Bacteria (30)
  - mnm5U34 biosynthesis bacteria (6)
  - Queuosine-Archaeosine Biosynthesis (8)
  - Ribonuclease H (2)
  - tRNA processing (6)
- Transcription (18)
  - Transcription initiation, bacterial sigma factors (3)

- RNA polymerase bacterial (3)
- Transcription factors bacterial (10)
- Rrf2 family transcriptional regulators (2)

RNA Metabolism - no subcategory (1)

- Group II intron-associated genes (1)

Nucleosides and Nucleotides (79)

Pyrimidines (27)

- pyrimidine conversions (17)
- De Novo Pyrimidine Synthesis (10)

Purines (38)

- De Novo Purine Biosynthesis (18)
- Purine conversions (20)

Nucleosides and Nucleotides - no subcategory (7)

- Ribonucleotide reduction (2)
- Adenosyl nucleosidases (5)

Detoxification (7)

- Nucleoside triphosphate pyrophosphohydrolase MazG (1)
- Nudix proteins (nucleoside triphosphate hydrolases) (2)
- Housecleaning nucleoside triphosphate pyrophosphatases (4)

Protein Metabolism (143)

Protein folding (11)

- GroEL GroES (2)
- Protein chaperones (6)
- Periplasmic disulfide interchange (1)
- Peptidyl-prolyl cis-trans isomerase (2)

Selenoproteins (1)

- Selenoprotein O (1)

Protein biosynthesis (83)

- tRNA aminoacylation, Val (1)
- tRNA aminoacylation, Met (1)
- tRNA aminoacylation, Ile (2)
- tRNA aminoacylation, Arg (1)
- Translation initiation factors bacterial (6)
- tRNA aminoacylation, Gly (1)
- Ribosome activity modulation (1)
- tRNA aminoacylation, Ala (1)
- tRNA aminoacylation, Trp (1)
- Ribosome LSU bacterial (32)
- Programmed frameshift (2)
- tRNA aminoacylation, Cys (1)
- Translation termination factors bacterial (10)
- tRNA aminoacylation, His (1)
- tRNA aminoacylation, Asp and Asn (2)
- Translation elongation factors bacterial (7)
- tRNA aminoacylation, Lys (1)
- tRNA aminoacylation, Thr (1)
- Translation elongation factor G family (3)
- tRNA aminoacylation, Glu and Gln (3)

tRNA aminoacylation, Ser (1)
tRNA aminoacylation, Tyr (1)
tRNA aminoacylation, Leu (1)
tRNA aminoacylation, Phe (2)
Protein processing and modification (28)
Protein-L-isoaspartate O-methyltransferase (1)
Ribosomal protein S12p Asp methylthiotransferase (2)
Peptide methionine sulfoxide reductase (2)
N-linked Glycosylation in Bacteria (6)
Lipoprotein Biosynthesis (4)
Modification of eukaryotic initiation factor 5A (2)
Signal peptidase (3)
G3E family of P-loop GTPases (metallocenter biosynthesis) (8)
Protein degradation (20)
Aminopeptidases (EC 3.4.11.-) (2)
Protein degradation (6)
Metallocarboxypeptidases (EC 3.4.17.-) (2)
Dipeptidases (EC 3.4.13.-) (2)
Serine endopeptidase (EC 3.4.21.-) (1)
Proteolysis in bacteria, ATP-dependent (6)
Omega peptidases (EC 3.4.19.-) (1)
Cell Division and Cell Cycle (16)
Checkpoint control (0)
Heterocyst formation (0)
Cell Division and Cell Cycle - no subcategory (16)
Bacterial Cytoskeleton (16)
Motility and Chemotaxis (0)
Magnetotaxis (0)
Motility and Chemotaxis - no subcategory (0)
Flagellar motility in Prokaryota (0)
Social motility and nonflagellar swimming in bacteria (0)
Regulation and Cell signaling (17)
Regulation and Cell signaling - no subcategory (12)
cAMP signaling in bacteria (8)
LysR-family proteins in Salmonella enterica Typhimurium (1)
LysR-family proteins in Escherichia coli (1)
Stringent Response, (p)ppGpp metabolism (2)
Signal transduction in Eukaryotes (0)
Quorum sensing and biofilm formation (0)
Proteolytic pathway (0)
Regulation of virulence (0)
Programmed Cell Death and Toxin-antitoxin Systems (5)
Toxin-antitoxin replicon stabilization systems (5)
Secondary Metabolism (5)
Secondary Metabolism - no subcategory (0)
Lipid-derived mediators (0)
Plant Octadecanoids (0)

- Bacterial cytostatics, differentiation factors and antibiotics (0)
- Biosynthesis of phenylpropanoids (0)
- Hydrocarbons (0)
- Aromatic amino acids and derivatives (0)
- UV-absorbing secondary metabolites (0)
- Plant Alkaloids (1)
  - Alkaloid biosynthesis from L-lysine (1)
- Biologically active compounds in metazoan cell defence and differentiation (0)
- Plant Hormones (4)
  - Auxin biosynthesis (4)

DNA Metabolism (83)

- DNA repair (49)
  - Uracil-DNA glycosylase (2)
  - DNA repair, bacterial MutL-MutS system (4)
  - DNA repair, UvrABC system (5)
  - DNA repair, bacterial photolyase (1)
  - DNA repair system including RecA, MutS and a hypothetical protein (2)
  - DNA repair, bacterial (15)
  - DNA repair, bacterial RecFOR pathway (9)
  - DNA Repair Base Excision (7)
  - DNA repair, bacterial UvrD and related helicases (4)
- CRISPs (5)
  - CRISPRs (5)
- DNA Metabolism - no subcategory (14)
  - Type I Restriction-Modification (4)
  - Restriction-Modification System (4)
  - DNA ligases (1)
  - YcfH (3)
  - DNA structural proteins, bacterial (2)
- DNA replication (12)
  - DNA topoisomerases, Type I, ATP-independent (7)
  - DNA replication strays (1)
  - DNA topoisomerases, Type II, ATP-dependent (4)
- DNA recombination (3)
  - RuvABC plus a hypothetical (3)
- DNA uptake, competence (0)

Regulons (0)

Atomic Regulons (0)

Fatty Acids, Lipids, and Isoprenoids (73)

- Phospholipids (12)
  - Glycerolipid and Glycerophospholipid Metabolism in Bacteria (12)
- Triacylglycerols (4)
  - Triacylglycerol metabolism (4)
- Fatty acids (29)
  - Fatty Acid Biosynthesis FASII (17)
  - Fatty acid metabolism cluster (12)
- Fatty Acids, Lipids, and Isoprenoids - no subcategory (20)

    - Polyhydroxybutyrate metabolism (20)
- Isoprenoids (8)
    - Myxoxanthophyll biosynthesis in Cyanobacteria (1)
    - Mevalonate Branch of Isoprenoid Biosynthesis (7)

Nitrogen Metabolism (14)

- Nitrogen Metabolism - no subcategory (7)
    - Nitrosative stress (4)
    - Ammonia assimilation (3)
- Denitrification (7)
    - Denitrifying reductase gene clusters (3)
    - Denitrification (4)

Dormancy and Sporulation (5)

- Spore DNA protection (0)
- Dormancy and Sporulation - no subcategory (5)
    - Spore Core Dehydration (1)
    - Persister Cells (3)
    - Sporulation-associated proteins with broader functions (1)

Respiration (55)

- Biotin (1)
- ATP synthases (0)
- Electron accepting reactions (8)
    - Terminal cytochrome C oxidases (5)
    - Anaerobic respiratory reductases (3)
- Electron donating reactions (35)
    - Respiratory Complex I (14)
    - Respiratory dehydrogenases 1 (1)
    - Succinate dehydrogenase (6)
    - NADH ubiquinone oxidoreductase (14)
- Reverse electron transport (0)
- Respiration - no subcategory (12)
    - Biogenesis of cbb3-type cytochrome c oxidases (5)
    - Biogenesis of c-type cytochromes (2)
    - Soluble cytochromes and functionally related electron carriers (5)
- Sodium Ion-Coupled Energetics (0)

Stress Response (50)

- Osmotic stress (3)
    - Osmoregulation (3)
- Dessication stress (0)
- Acid stress (0)
- Oxidative stress (23)
    - Protection from Reactive Oxygen Species (4)
    - Oxidative stress (17)
    - Glutathione: Non-redox reactions (1)
    - Glutathione: Redox cycle (1)
- Cold shock (1)
    - Cold shock, CspA family of proteins (1)
- Heat shock (14)

- Heat shock dnaK gene cluster extended (14)

Detoxification (9)

- D-tyrosyl-tRNA(Tyr) deacylase (1)
- Uptake of selenate and selenite (1)

Stress Response - no subcategory (3)

- Dimethylarginine metabolism (2)
- Hfl operon (1)

Periplasmic Stress (4)

- Periplasmic Stress Response (4)

Metabolism of Aromatic Compounds (10)

Peripheral pathways for catabolism of aromatic compounds (1)

- Quinate degradation (1)

Anaerobic degradation of aromatic compounds (0)

Metabolism of central aromatic intermediates (8)

- Catechol branch of beta-ketoadipate pathway (3)
- Salicylate and gentisate catabolism (2)
- Homogentisate pathway of aromatic compound degradation (3)

Metabolism of Aromatic Compounds - no subcategory (1)

- Gentisare degradation (1)

Amino Acids and Derivatives (268)

Glutamine, glutamate, aspartate, asparagine; ammonia assimilation (16)

- Glutamine, Glutamate, Aspartate and Asparagine Biosynthesis (12)
- Glutamate dehydrogenases (1)
- Glutamine synthetases (1)
- Glutamate and Aspartate uptake in Bacteria (2)

Histidine Metabolism (4)

- Histidine Degradation (4)

Arginine; urea cycle, polyamines (27)

- Polyamine Metabolism (3)
- Arginine and Ornithine Degradation (7)
- Arginine Biosynthesis extended (7)
- Arginine Biosynthesis -- gjo (7)
- Cyanophycin Metabolism (3)

Lysine, threonine, methionine, and cysteine (70)

- Methionine Degradation (11)
- Methionine Biosynthesis (24)
- Threonine and Homoserine Biosynthesis (12)
- Threonine degradation (3)
- Lysine Biosynthesis DAP Pathway, GJO scratch (8)
- Cysteine Biosynthesis (12)

Amino Acids and Derivatives - no subcategory (0)

Branched-chain amino acids (60)

- Isoleucine degradation (23)
- Leucine Degradation and HMG-CoA Metabolism (18)
- Valine degradation (19)

Polyamines (0)

Aromatic amino acids and derivatives (43)

- Common Pathway For Synthesis of Aromatic Compounds (DAHP synthase to chorismate) (7)

- Chorismate Synthesis (10)
  - Chorismate: Intermediate for synthesis of Tryptophan, PAPA antibiotics, PABA, 3-hydroxyanthranilate and more. (12)
  - Phenylalanine and Tyrosine Branches from Chorismate (3)
  - Tryptophan synthesis (11)
- Proline and 4-hydroxyproline (6)
  - Proline Synthesis (2)
  - A Hypothetical Protein Related to Proline Metabolism (2)
  - Proline, 4-hydroxyproline uptake and utilization (2)
- Alanine, serine, and glycine (42)
  - Glycine Biosynthesis (5)
  - Alanine biosynthesis (7)
  - Serine Biosynthesis (7)
  - Glycine cleavage system (4)
  - Glycine and Serine Utilization (19)

Sulfur Metabolism (10)

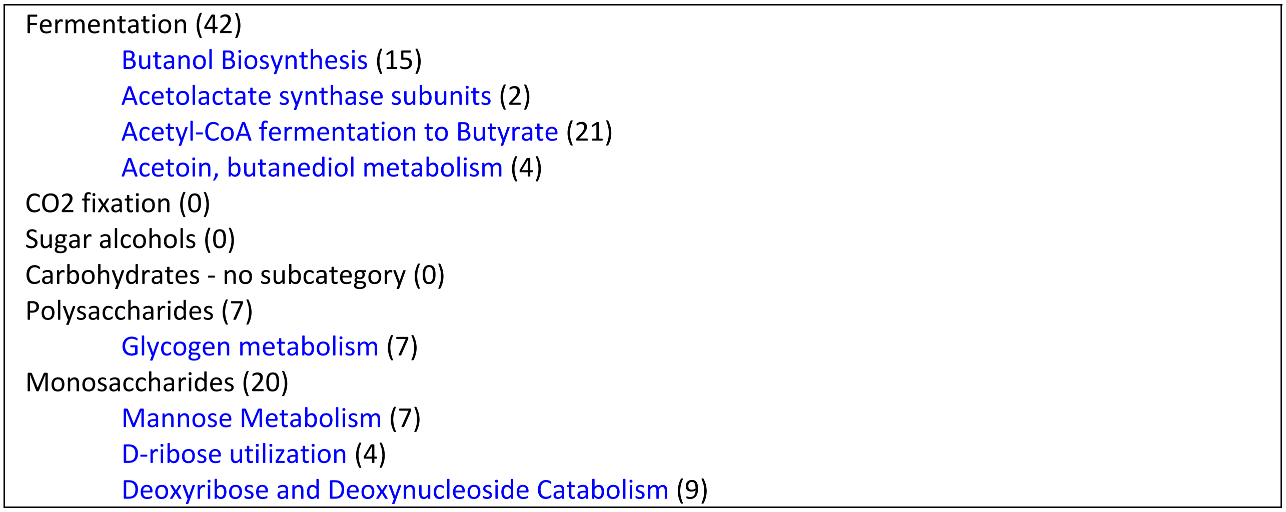
- Inorganic sulfur assimilation (0)
- Sulfur Metabolism - no subcategory (10)
  - Thioredoxin-disulfide reductase (7)
  - Galactosylceramide and Sulfatide metabolism (3)
- Organic sulfur assimilation (0)

Phosphorus Metabolism (14)

- Phosphorus Metabolism - no subcategory (14)
  - Phosphate metabolism (12)
  - Polyphosphate (2)

Carbohydrates (209)

- Central carbohydrate metabolism (84)
  - Methylglyoxal Metabolism (5)
  - Pyruvate metabolism II: acetyl-CoA, acetogenesis from pyruvate (7)
  - Pyruvate Alanine Serine Interconversions (8)
  - Glyoxylate bypass (5)
  - Glycolysis and Gluconeogenesis (13)
  - Dehydrogenase complexes (15)
  - TCA Cycle (18)
  - Pentose phosphate pathway (6)
  - Pyruvate metabolism I: anaplerotic reactions, PEP (7)
- Aminosugars (0)
- Di- and oligosaccharides (15)
  - Sucrose utilization (3)
  - Maltose and Maltodextrin Utilization (9)
  - Lactose utilization (3)
- Glycoside hydrolases (0)
- One-carbon Metabolism (39)
  - Serine-glyoxylate cycle (34)
  - One-carbon metabolism by tetrahydropterines (5)
- Organic acids (2)
  - Lactate utilization (2)



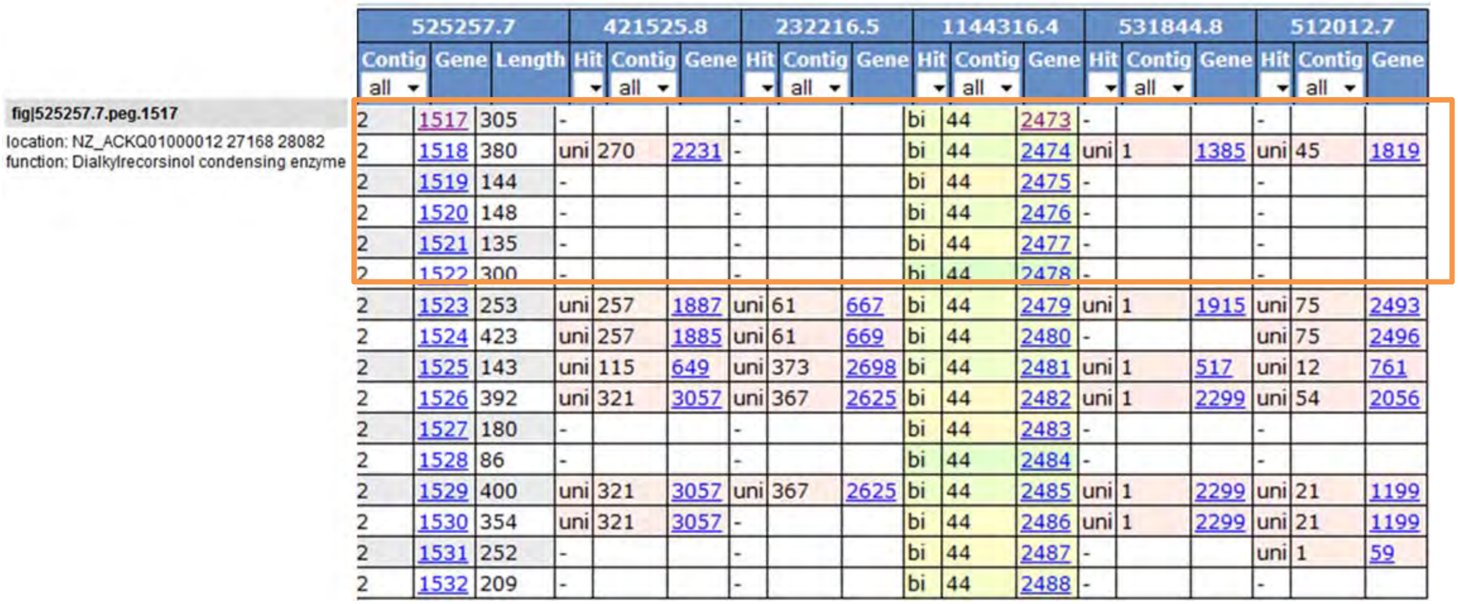
- Fermentation (42)
  - Butanol Biosynthesis (15)
  - Acetolactate synthase subunits (2)
  - Acetyl-CoA fermentation to Butyrate (21)
  - Acetoin, butanediol metabolism (4)
- CO2 fixation (0)
- Sugar alcohols (0)
- Carbohydrates - no subcategory (0)
- Polysaccharides (7)
  - Glycogen metabolism (7)
- Monosaccharides (20)
  - Mannose Metabolism (7)
  - D-ribose utilization (4)
  - Deoxyribose and Deoxynucleoside Catabolism (9)

Figure 6: Subsystem Feature Counts from RAST

Sequence Based Comparison

It was observed that *Chryseobacterium gleum* and *Chryseobacterium* sp. CF314 both had genes encoding the dialkylrecorsinol condensing enzyme, but the other comparison organisms, including *Chryseobacterium haifense*, did not have the dialkylrecorsinol condensing enzyme, a flexirubin biosynthesis gene. There are six other genes (1518- 1522, Orange box, Fig. 7), which are thought to be related to Flexirubin biosynthesis, that are present in *C. gleum* and *C.* sp. CF314, but not the others, including ***Chryseobacterium haifense***.

C. gleum *C. haifense* *C. koreense* *C.* sp. CF314 *F.* sp. 3519-10 *F.* sp. JJC



fig|525257.7.peg.1517
location: NZ_ACKQ01000012 27168 28082
function: Dialkylrecorsinol condensing enzyme

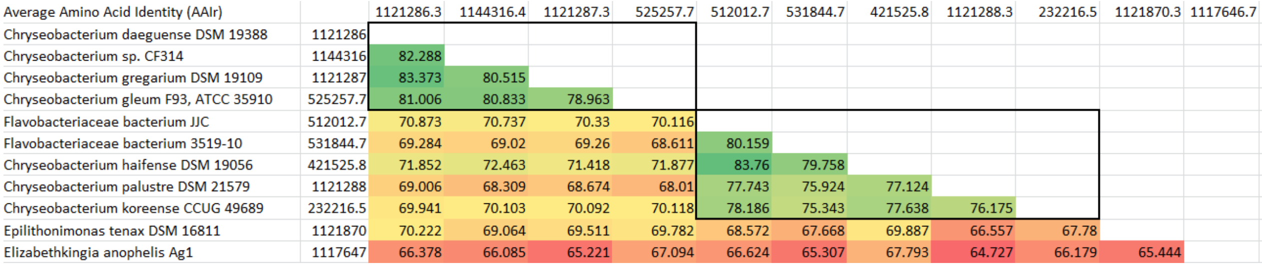
525257.7			421525.8			232216.5			1144316.4			531844.8			512012.7		
Contig	Gene	Length	Hit	Contig	Gene	Hit	Contig	Gene	Hit	Contig	Gene	Hit	Contig	Gene	Hit	Contig	Gene
all ▾			▾	all ▾		▾	all ▾		▾	all ▾		▾	all ▾		▾	all ▾	
2	1517	305	-			-			bi	44	2473	-			-		
2	1518	380	uni	270	2231	-			bi	44	2474	uni	1	1385	uni	45	1819
2	1519	144	-			-			bi	44	2475	-			-		
2	1520	148	-			-			bi	44	2476	-			-		
2	1521	135	-			-			bi	44	2477	-			-		
2	1522	300	-			-			bi	44	2478	-			-		
2	1523	253	uni	257	1887	uni	61	667	bi	44	2479	uni	1	1915	uni	75	2493
2	1524	423	uni	257	1885	uni	61	669	bi	44	2480	-			uni	75	2496
2	1525	143	uni	115	649	uni	373	2698	bi	44	2481	uni	1	517	uni	12	761
2	1526	392	uni	321	3057	uni	367	2625	bi	44	2482	uni	1	2299	uni	54	2056
2	1527	180	-			-			bi	44	2483	-			-		
2	1528	86	-			-			bi	44	2484	-			-		
2	1529	400	uni	321	3057	uni	367	2625	bi	44	2485	uni	1	2299	uni	21	1199
2	1530	354	uni	321	3057	-			bi	44	2486	uni	1	2299	uni	21	1199
2	1531	252	-			-			bi	44	2487	-			uni	1	59
2	1532	209	-			-			bi	44	2488	-			-		

Figure 7: RAST Sequence Based Comparison *Chryseobacterium gleum* was used as the reference for a sequence based comparison. A region containing flexirubin biosynthesis genes (orange box) was identified.

ROSA Analysis

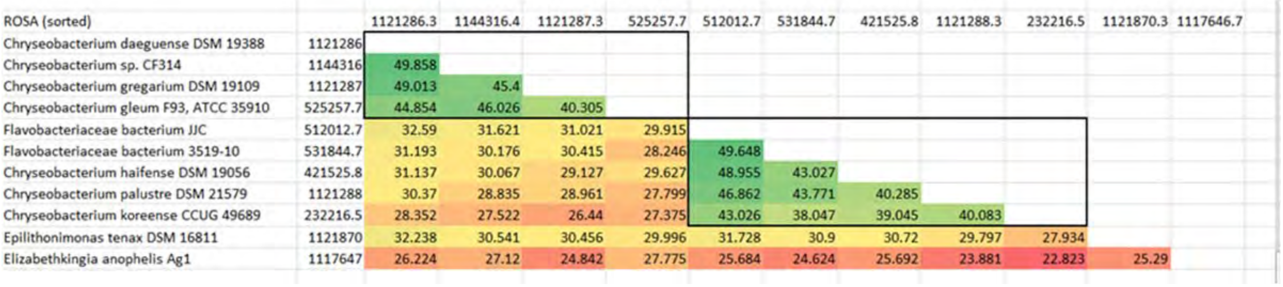
OS Values, %BBH values and AAI values for the individual comparisons along with the matrices can be found in the supplementary Microsoft Excel sheet. The AAI (Fig. 8A) and ROSA (Fig. 8B) indicated the presence of related clusters (black boxes).

A



Average Amino Acid Identity (AAIr)		1121286.3	1144316.4	1121287.3	525257.7	512012.7	531844.7	421525.8	1121288.3	232216.5	1121870.3	1117646.7
Chryseobacterium daeguense DSM 19388	1121286											
Chryseobacterium sp. CF314	1144316	82.288										
Chryseobacterium gregarium DSM 19109	1121287	83.373	80.515									
Chryseobacterium gleum F93, ATCC 35910	525257.7	81.006	80.833	78.963								
Flavobacteriaceae bacterium JJC	512012.7	70.873	70.737	70.33	70.116							
Flavobacteriaceae bacterium 3519-10	531844.7	69.284	69.02	69.26	68.611	80.159						
Chryseobacterium haifense DSM 19056	421525.8	71.852	72.463	71.418	71.877	83.76	79.758					
Chryseobacterium palustre DSM 21579	1121288	69.006	68.309	68.674	68.01	77.743	75.924	77.124				
Chryseobacterium koreense CCUG 49689	232216.5	69.941	70.103	70.092	70.118	78.186	75.343	77.638	76.175			
Epilithonimonas tenax DSM 16811	1121870	70.222	69.064	69.511	69.782	68.572	67.668	69.887	66.557	67.78		
Elizabethkingia anophelis Ag1	1117647	66.378	66.085	65.221	67.094	66.624	65.307	67.793	64.727	66.179	65.444	

B.



ROSA (sorted)		1121286.3	1144316.4	1121287.3	525257.7	512012.7	531844.7	421525.8	1121288.3	232216.5	1121870.3	1117646.7
Chryseobacterium daeguense DSM 19388	1121286											
Chryseobacterium sp. CF314	1144316	49.858										
Chryseobacterium gregarium DSM 19109	1121287	49.013	45.4									
Chryseobacterium gleum F93, ATCC 35910	525257.7	44.854	46.026	40.305								
Flavobacteriaceae bacterium JJC	512012.7	32.59	31.621	31.021	29.915							
Flavobacteriaceae bacterium 3519-10	531844.7	31.193	30.176	30.415	28.246	49.648						
Chryseobacterium haifense DSM 19056	421525.8	31.137	30.067	29.127	29.627	48.955	43.027					
Chryseobacterium palustre DSM 21579	1121288	30.37	28.835	28.961	27.799	46.862	43.771	40.285				
Chryseobacterium koreense CCUG 49689	232216.5	28.352	27.522	26.44	27.375	43.026	38.047	39.045	40.083			
Epilithonimonas tenax DSM 16811	1121870	32.238	30.541	30.456	29.996	31.728	30.9	30.72	29.797	27.934		
Elizabethkingia anophelis Ag1	1117647	26.224	27.12	24.842	27.775	25.684	24.624	25.692	23.881	22.823	25.29	

Figure 8: AAI and ROSA Data Matrices. The Average Amino Acid Identity (AAI) values for the comparison right as of *Chryseobacterium haifense* and the 10 other closely related organisms are oriented the same left to top to bottom (A). The Reciprocal Orthology Score Average (ROSA) values for the comparison of *Chryseobacterium haifense* and the 10 other closely related organisms are oriented the same left to right as top to bottom (B). Boxes (A and B) are used to denote clusters of high scoring comparisons.

Discussion:

Geneious 5.6 Assembly and Geneious 6.0 Assembly

The Geneious assembly with 50bp trim was determined to be the optimal assembly, which was determined after observing the number of contigs used to make supercontigs, number of unused contigs and the number of newly made supercontigs. The assembly with the most used reads into the fewest contigs along with the fewest unused reads was considered to be the most optimum assembly.

The correlation between the number of bp trimmed and the number of contigs generated is likely due to the incorrect bases on the edges of the contigs due to inaccurate flow-calls which introduce insertion/deletion errors at a raw rate of 2.84%, generally resulting in over-called short-homopolymers and under-called long-homopolymers (Bragg *et al.* 2013). As can be seen in Table 1, the number of contigs generated decreased from the non-trimmed assembly until the 35bp-trimmed assembly (Table 1). This result likely occurred due to the trimming of error containing regions near the edges of the contig. As more of the errors were removed more of the contigs were recognized as overlapping. The 45bp-trimmed did not fit the correlation (Table 1), and should be further investigated. If 35bp are trimmed from the edge of the contig and there were only around 35bp error on the edge of the contigs, then when trimming 45bp valuable, accurate regions of the DNA could be lost. With valuable, accurate regions of DNA being removed, it could be difficult for the Geneious Assembly software to combine contigs based on similarity. However, this trend was continued by the 50bp-trim method (Table 1), the 50bp method removed sequencing errors without affecting accurate DNA sequences. The number of contigs generated increased with the 60bp-trim method (Table 1). The explanation responsible for the results observed with the 60bp-trim method is that as more of the DNA is being removed, accurate nucleotides in the sequence were removed and the Geneious Assembler could not assemble the contigs properly due to deletion of the overlapping sequences.

The 50bp-trim followed the correlation observed for the assemblies from the non-trim to the 35bp-trim; however, the 45bp-trim did not fit this correlation. The exact process for the trimming and assembly through Geneious is not known, because it is protected by copyrights. The value selected in the trimming option indicated the smallest number of bp to be deleted. This indicated that the number of bp trimmed from the 3’ and 5’ ends were at least the values selected, but could potentially lead to a higher amount of bp to be trimmed (Eg. 35bp trim could lead to 40bp trim). This being said, it is hypothesized that there were actually more bp trimmed in the 45bp-trim method than in the 50bp-trim method. This hypothesis was not tested, but could be a potential point for future research. If the hypothesis is true than the correlation observed still stands. This correlation suggested that the amount of incorrect bp trimmed from the ends of contigs improve the assembly; however, if the trimming exceeds the error region of the contig and begins to disrupt correct sequences, then the amount of contigs generated will increase, because the Geneious Assembler will not be able to detect the similarities of the contigs.

Phylogenetic Tree

The orange and blue braces (Fig. 5) represent members of the genus *Chryseobacterium*. The red braces (Fig. 5) indicate members of the *Epilithonimonas* genus. And the green braces (Fig. 5) represent the *Flavobacterium* genus. The blue braces include the organisms: *C. haifense, C. koreense,C. jeonii* AT1047T, *C. solincola* 1YBR12T, *F.* sp 3519-10, and *F.* sp JJC, but this branch is clearly separated from the orange branch which includes fellow Chryseobacteria. The *Epilithonimonas* branch is more closely related to the true Chryseobacteria, orange braces because it contains the type species for the genus, than the branch containing *C. haifense. Epilithonamonas* and *Flavobacterium* genera were selected to represent organisms accepted as non-Chryseobacteria. The organisms within the blue braces are more distantly related to the accepted Chryseobacteria genus, branch with *C. gleum*, than *Epilithonamonas*, which is accepted as a separate genus. It can be concluded that the 16S rRNA indicates a need to reclassify: *C. haifense, C. koreense,C. jeonii* AT1047T, *C. solincola* 1YBR12T, *F.* sp 3519-10, and *F.* sp JJC. Previous reclassifications, such as the *Kaistella koreensis* to the genus *Chryseobacterium*, have relied on 16S rRNA sequencing, but 16S rRNA should not be enough to reclassify organisms. 16S rRNA similarity should be used along with whole genome analysis metrics such as ROSA and phenotypic testing in order to reclassify organisms.

Subsystem Analysis and Sequence Based Comparison

Chryseobacterium haifense had genes necessary for all living organisms such as: genes for aminoacyl-tRNA synthetases, which attach the proper amino to tRNA for protein biosynthesis; DNA helicase genes, which are used to separate the DNA strand; and genes coding for bacterial cytoskeleton, which adds support and stability to the cell and is also used in cell division.

Chryseobacterium haifense contains 8 genes associated with cAMP signaling in bacteria. The cAMP signal pathway is used to regulate the lactose and other catabolite utilization pathways. *Chryseobacterium haifense* also contains genes associated with the lactose utilization pathway. It is no surprise that *C. haifense* contains genes for lactose utilization and the regulation of the lactose pathway because *C. haifense* was isolated from raw cow's milk where there would be an abundance of lactose.

Chryseobacterium haifense was described as a non-motile bacterium, and it is no surprise that there are not any genes in the motility subsystem. *C. haifense* does not contain genes associated with photosynthesis.

Chyrseobacterium haifense contains genes necessary to carry out the Glycolysis/gluconeogenesis pathway, the Tricarboxylic Acid Cycle, Pentose phosphate pathway, glyoxolate, and genes for pyruvate metabolism such as the anaplerotic reactions. The presence of these genes confirms that *C. haifense* is truly an aerobic bacterium as originally described.

Chryseobacterium haifense contains 42 genes responsible for the fermentation process. *C. haifense* was reported to produce acid from glucose, fructose, lactose and maltose, which is not surprising because it contains the necessary genes to perform fermentation and to utilize each of the sugars.

Chryseobacterium haifense does not contain flexirubin biosynthesis genes. The dialkylrecorsinol condensing enzyme is a flexirubin biosynthesis gene, found in *Chryseobacterium gleum* but not in *C. haifense* (Fig. 7). The genes in the orange box (Fig. 7) are flexirubin biosynthesis genes. These genes are present in the "true" Chryseobacteria such as *C. gleum* and *Chryseobacterium* CF314 but are not present in the wrongly classified Chryseobacteria such as *C. haifense* and *C. koreense.* This further confirms that *Chryseobacterium haifense* should be reclassified as a separate genus.

ROSA Analysis

The ROSA Sorted matrix (Fig. 10) appears to indicate 3 separate phylogenetic clusters, but the two bottom clusters are analogous. The top cluster includes *Chryseobacterium gleum* and other "true" Chryseobacteria. The bottom cluster includes *Chryseobacterium haifense* and the other Chryseobacteria that need to be reclassified. This is further supported when observing the *Chryseobacterium gleum* and *Epilithonomonas tenax* comparison and the *C. gleum* and *C. haifense* comparison. The ROSA value, for the *Chryseobacterium gleum* and *Epilithonomonas tenax*, is 29.996, which indicates a genus level separation. The ROSA value between *Chryseobacterium gleum* and *Chryseobacterium haifense* is 29.627, which is lower than the genus level separation seen for *Chryseobacterium gleum* and *Epilithonomonas* *tenax.* The other organisms closely related to *C. haifense* such as *Flavobacteriaceae* sp. 3519-10, *Flavobacteriaceae* sp. JJC*,* Chryseobacterium *koreense* and *Chryseobacterium palustre* all have a ROSA value below 30 when compared to *C. gleum*, which indicates a genus level separation.

Conclusion:

Based on the ROSA analysis and the 16S rRNA phylogeny, the comparison organisms including *Chryseobacterium haifense* belong to a different genus than *Chryseobacterium gleum.* This genus will likely be the re-instated *Kaistella* genus. *Kaistella koreensis* was the sole representative, before being wrongly reclassified as *Chryseobacterium koreense. Chryseobacterium haifense, Chryseobacterium koreense,* and the other "false" Chryseobacteria belong in the genus *Kaistella.* The *Chryseobacterium haifense* genome is currently a draft genome ready for publication. The draft genome of *Zavarzinella formosa* DSM 19928 was published in the Journal of Bacteriology (Guo *et al.* 2012) with 594 contigs (Aziz *et al.* 2008). The draft genome of *C. haifense* has 676 contigs, which is slightly higher than *Zavarzinella formosa* DSM 19928 but should still be within an acceptable range for publication.

Works Cited

Auch, A.F., von Jan, M., Klenk, H., and Göker, M. (2010). Digital DNA-DNA hybridization for microbial species delineation by means of genome-to-genome sequence comparison. SIGS 2:117-134.

Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, Formsma K, Gerdes S, Glass EM, Kubal M, Meyer F, Olsen GJ, Olson R, Osterman AL, Overbeek RA, McNeil LK, Paarmann D, Paczian T, Parrello B, Pusch GD, Reich C, Stevens R, Vassieva O, Vonstein V, Wilke A, and Zagnitko O. (2008). The RAST Server: rapid annotations using subsystems technology. BMC 9:75.

Bernardet JF, Nakagawa Y, Holmes B; Subcommittee on the taxonomy of Flavobacterium and Cytophaga-like bacteria of the International Committee on Systematics of Prokaryotes. (2002). Proposed minimal standards for describing new taxa of the family Flavobacteriaceae and emended description of the family. IJSEM 53(Pt 3): 1049-70.

Bragg LM, Stone G, Butler MK, Hugenholtz P, Tyson GW. (2013). Shining a light on dark sequencing: characterising errors in Ion Torrent PGM data. PLoS Comput Biol 9(4): e1003031.

Buonaccorsi VP, Boyle MD, Grove D, Praul C, Sakk E, Stuart A, Tobin T, Hosler J, Carney SL, Engle MJ, Overton BE, Newman JD, Pizzorno M, Powell JR, and Trun N. (2011). GCAT-SEEKquence: genome consortium for active teaching of undergraduates through increased faculty access to next-generation sequencing data. CBE Life Sci Educ 10(4): 342-5.

Carver T, Berriman M, Tivey A, Patel C, Böhme U, Barrell BG, Parkhill J and Rajandream MA. (2008). Artemis and ACT: viewing, annotating and comparing sequences stored in a relational database. Bioinformatics 24(23):2672-6

Conlan S, Kong HH, and Segre JA. (2012). Species-level analysis of DNA sequence data from the NIH Human Microbiome Project. PLoS One 7(10): e47075.

Fodor AA, DeSantis TZ, Wylie KM, Badger JH, Ye Y, Hepburn T, Hu P, Sodergren E, Liolios K, Huot-Creasy H, Birren BW, Earl AM. (2012). The "most wanted" taxa from the human microbiome for whole genome sequencing. PLoS One 7(7): e41294.

Goris, J., Konstantinidis, K.T., Klappenbach, J.A., Coenye, T., Vandamme, P., and Tiedje, J.M. (2007). DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. IJSEM 57:81-91.

Grigoriev, I.V., Nordberg, H., Shabalov, I., Aerts, A., Cantor, M., Goodstein, D., Kuo, A., Minovitsky, S., Nikitin, R., Ohm, R.A., Otillar, R., Poliakov, A., Ratnere, I., Riley, R., Smirnova, T., Rokhsar, D., and Dubchak, I. (2012). The Genome Portal of the Department of Energy Joint Genome Institute. Nucleic Acids Research 40: D26-D32.

Guo M, Han X, Jin T, Zhou L, Yang J, Li Z, Chen J, Geng B, Zou Y, Wan D, Li D, Dai W, Wang H, Chen Y, Ni P, Fang C, and Yang R. (2012). Genome sequences of three species in the family Planctomycetaceae. J. Bacteriology 194(14): 3740-1.

Hamady, M., and Knight, R. (2009). Microbial community profiling for human microbiome projects: Tools, techniques, and challenges. Genome Res. 19(7): 1141-52.

Hantsis-Zacharov E, and Halpern M. (2007). Chryseobacterium haifense sp. nov., a psychrotolerant bacterium isolated from raw milk. IJSEM 57(Pt 10): 2344-8.

Holmes, B., Owen., R.J., Steigerwalt., A.G., and Brenner, D.J. (1984). *Flavobacterium* gleum, as New Species found in Human Clinical Specimens. IJSB 34(1): 21-25.

Human Microbiome Jumpstart Reference Strains Consortium, Nelson KE, Weinstock GM, Highlander SK, Worley KC, Creasy HH, Wortman JR, Rusch DB, Mitreva M, Sodergren E, Chinwalla AT, Feldgarden M, Gevers D, Haas BJ, Madupu R, Ward DV, Birren BW, Gibbs RA, Methe B, Petrosino JF, Strausberg RL, Sutton GG, White OR, Wilson RK, Durkin S, Giglio MG, Gujja S, Howarth C, Kodira CD, Kyrpides N, Mehta T, Muzny DM, Pearson M, Pepin K, Pati A, Qin X, Yandava C, Zeng Q, Zhang L, Berlin AM, Chen L, Hepburn TA, Johnson J, McCorrison J, Miller J, Minx P, Nusbaum C, Russ C, Sykes SM, Tomlinson CM, Young S, Warren WC, Badger J, Crabtree J, Markowitz VM, Orvis J, Cree A, Ferriera S, Fulton LL, Fulton RS, Gillis M, Hemphill LD, Joshi V, Kovar C, Torralba M, Wetterstrand KA, Abouellleil A, Wollam AM, Buhay CJ, Ding Y, Dugan S, FitzGerald MG, Holder M, Hostetler J, Clifton SW, Allen-Vercoe E, Earl AM, Farmer CN, Liolios K, Surette MG, Xu Q, Pohl C, Wilczek-Boney K, Zhu D. (2010). A catalog of reference genomes from the human microbiome. Science 328(5981): 994-9.

Kisand V, and Lettieri T. (2013). Genome sequencing of bacteria: sequencing, de novo assembly and rapid analysis using open source tools. BMC Genomics 14:211.

Konstantantinidis, K.T., and Tiedje, J.M. (2007). Prokaryotic taxonomy and phylogeny in the genomic era: advancements and challenges ahead. Current Opinion in Microbiology 10:504-509.

Konstantantinidis, K.T., Ramette, A., and Tiedje,J.M. (2005). The bacterial species definition in the genomic era. Phil. Trans. R. Soc. B. 361:1929-1940.

Koonin EV. The Clusters of Orthologous Groups (COGs) Database: Phylogenetic Classification of Proteins from Complete Genomes. 2002 Oct 9 [Updated 2003 Aug 13]. In: McEntyre J, Ostell J, editors. The NCBI Handbook [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2002-. Chapter 22. Available from: http://www.ncbi.nlm.nih.gov/books/NBK21090/

Langille, M.G.I., Laird, M.R., Hsiao, W.W.L., Chiu, T.A., Eisen, J.A., and Brinkman, F.S.L. (2012). MicrobeDB: a locally maintainable database of microbial genomic sequences. Bioinformatics 28(14): 1947-1948.

Lee, H.C., Lai, K., Lorenc, M.T., Imelfort, M., Duran, C., and Edwards, D. (2011). Bioinformatics tools and databases for analysis of next-generation sequence data. Briefings in Functional Genomics 2(1): 12-24.

Markowitz, V.M., Chen, I.A., Palaniappan, K., Chu,K., Szeto, E., Grechkin, Y., Ratner, A., Jacob,B., Huang, J., Williams, P., Huntemann, M., Anderson, I., Mavromatis, K., Ivanova, N.N, and Kyrpides, N.C. (2012). IMG: the integrated microbial genomes databaseand comparative analysis system. Nucleic Acids Research 40:D115-D122.

NIH HMP Working Group, Peterson J, Garges S, Giovanni M, McInnes P, Wang L, Schloss JA, Bonazzi V, McEwen JE, Wetterstrand KA, Deal C, Baker CC, Di Francesco V, Howcroft TK, Karp RW, Lunsford RD, Wellington CR, Belachew T, Wright M, Giblin C, David H, Mills M, Salomon R, Mullins C, Akolkar B, Begg L, Davis C, Grandison L, Humble M, Khalsa J, Little AR, Peavy H, Pontzer C, Portnoy M, Sayre MH, Starke-Reed P, Zakhari S, Read J, Watson B, Guyer M. The NIH Human Microbiome Project. Genome Res. 19(12): 2317-23.

Pagani, I., Liolios, K., Jansson, J., Chen, I.A., Smirnova,T., Nosrat,B., Markowitz, V.M., and Kyrpides, N.C. (2012). The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata. Nucleic Acids Research 40:D571-D579.

Richter, M., and Roselló-Móra, R. (2009). Shifting the genomic gold standard for the prokaryotic species definition. PNAS 106(45): 19126-19131.

Rothberg JM, Hinz W, Rearick TM, Schultz J, Mileski W, Davey M, Leamon JH, Johnson K, Milgrew MJ, Edwards M, Hoon J, Simons JF, Marran D, Myers JW, Davidson JF, Branting A, Nobile JR, Puc BP, Light D, Clark TA, Huber M, Branciforte JT, Stoner IB, Cawley SE, Lyons M, Fu Y, Homer N, Sedova

M, Miao X, Reed B, Sabina J, Feierstein E, Schorn M, Alanjary M, Dimalanta E, Dressman D, Kasinskas R, Sokolsky T, Fidanza JA, Namsaraev E, McKernan KJ, Williams A, Roth GT, Bustillo J. (2011). An integrated semiconductor device enabling non-optical genome sequencing. Nature 475(7356): 348-52.

Salzberg SL, Sommer DD, Puiu D, and Lee VT.(2008). Gene-boosted assembly of a novel bacterial genome from very short reads. PLoS Comput Biol 4(9): e1000186.

Stackebrandt, E., Frederiksen, W., Garrity, G.M., Grimont, P.A.D., Kämpfer, P., Maiden, M.C.J., Nesme, X., Roselló-Mora, R., Swings, J., Trüper, H.G., Vauterin, L., Ward, A.C., and Whitman, W.B (2002). Report of the ad hoc committee for the reevaluation of the species definition in bacteriology. IJSEM 52:1043-1047.

Tang, L., Li, Y., Deng, X., Johnston, R.N., Liu, G., and Liu, S. (2013). Defining natural species of bacteria: clear-cut genomic boundaries revealed by a turning point in nucleotide sequence divergence.

Tatusov, R.L., Galperin, M.Y., Natale, D.A., and Koonin, E.V. (2000) The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Research 28(1):33-36.

Vandamme, P., Bernardet, J.F., Segers, P., Kersters, K., and Holmes, B. (1994). New Perspectives in the Classification of the Flavobacteria: Description of *Chryseobacterium* gen. nov., *Bergeyella* gen. nov., and *Empedobacter* norn. rev. IJSB 44(4): 827-831.

Zhang J, Chiodini R, Badr A, and Zhang G. (2011). The impact of next-generation sequencing on genomics. J Genet Genomics 38(3): 95-109.