GCAT-SEEK Workshop - Prokaryotic Genomics Module – Jeff Newman

A. Background

The era of genomics arguably began with the "Shot heard around the world," the publication of the first microbial genome sequence (Fleischmann *et al.*, 1995). This paper included 40 co-authors and described the whole genome shotgun sequence method using shotgun cloning, and Sanger dideoxy chain termination sequencing to create a finished 1.9 Mb genome of *Haemophilus influenzae* with 6x coverage. While template preparation, sequencing technologies and computational tools have improved dramatically over the ensuing two decades, the overall approach outlined by Fleischmann *et al.*, (Table 1) has remained surprisingly similar.

	Stage	Description
1.	Random small insert and large insert library construction	Shear genomic DNA randomly to2 kb and 15 to 20 kb, respectively
2.	Library plating	Verify random nature of library and maximize random selection of small insert and large insert clones for template production
3.	High-throughput DNA sequencing	Sequence sufficient number of sequence fragments from both ends for 6x coverage
4.	Assembly Physical gaps Sequence gaps	Assemble random sequence fragments and identify repeat regions Order all contigs (fingerprints, peptide links, X clones, PCR) and provide templates for closure
5.	Gap closure	Complete the genome sequence by primer walking
6.	Editing	Inspect the sequence visually and resolve sequence ambiguities, including frameshifts
7.	Annotation	Identify and describe all predicted coding regions (putative identifications, starts and stops, role assignments, operons, regulatory regions).

Table 1. Whole-genome sequencing strategy. (from Fleischmann et al., 1995)

B. The goals for this GCAT-SEEK workshop module are to isolate and evaluate genomic DNA from a bacterium of interest and prepare it for sequencing. A specialized sequencing facility will prepare the libraries and sequence the DNA using NextGen technologies, probably MiSeq or HiSeq, to 100x coverage.(steps 1-3 above). We will then use example data to learn how to assemble the sequences into contigs, with or without a reference, manually edit the sequence to identify more overlaps and gaps that are amenable to PCR-based closure. Participants will have a simple path that can be followed to generate and analyze a prokaryotic genome sequence chosen by the participant.

C. Vision and Change Core Competencies Addressed

These activities incorporate most/all of core concepts and competencies from the AAAS/NSF Vision and Change "Call to Action." Assembly to a reference genome and comparison of gene content and order illustrates evolutionary changes. The structure of operons and organization of genes typically reflects common biological functions. The annotation of genes and identification of instances of horizontal gene transfer is critically dependent on our understanding of **information flow**, exchange and storage. Integration of the annotated gene products into subsystems will identify pathways used by the organisms to transform energy and matter during growth. Based on knowledge of the organism's biology and phenotypic characteristics, participants will apply the process of science to make predictions about which genes/subsystems should or should not be present. Quantitative reasoning will be used to evaluate the raw sequence data based on quality scores and predict assembly metrics based on read length and number. The metabolism of the organism will be **modeled** based on the subsystems identified. Discussion of the algorithms used during assembly and the physical-chemical principles that underlie next-generation sequencing technologies will illustrate the interdisciplinary nature of Genomics. Finally, the presentation of the prokaryotic genomics methods during the day 5 portion of the workshop exploring the alternate applications of next-generation sequencing will provide practice in communication and collaboration with other disciplines.

D. GCAT-SEEK sequencing requirements.

Microbial genome sequencing can be accomplished using a variety of Next Generation Sequencing technologies, with the caveat that shorter reader lengths necessitate higher coverage levels. NextGen sequencing instruments generate massive amounts of sequence data, far more than what is needed for a single bacterial genome. Each run in the instrument also costs several thousand dollars, so the typical strategy is to organize shared runs to decrease the cost per genome. Different samples are prepared with different barcodes, which are sequences attached to each of the fragments while generating the library. This allows the sequences derived from different samples to be sorted after a "multiplexed" or combined run. Demultiplexing is often done by the sequencing facility for shared runs, however the NextGene package described below also has barcode sorting tools available on the main menu. Different instruments have advantages and disadvantages.

- Pacific Biosciences (PacBio) Single Molecule Real Time (SMRT) sequencing is the newest, least common, and most expensive but produces very long reads and is best if one needs a finished genome.
- 454 Pyrosequencing is moderately expensive but has relatively long reads to facilitate assembly.
- Ion Torrent is fastest, is inexpensive, has mid-range read lengths but a relatively high error rate and does not create paired end reads, resulting in assembly difficulty.
- The Illumina MiSeq is inexpensive, has mid-range read lengths and does create paired end reads, facilitating **de novo assembly**.
- The Illumina HiSeq is the least expensive per MB, but produces shorter paired end reads, which are better for **resequencing/alignment to a reference genome.**

100x coverage is a good target for a bacterial genome to optimize coverage or a good assembly, while still using a relatively small fraction of a run. 100x coverage of 5 Mb genome would correspond to 0.5 Gb. A single MiSeq run using the V3 2 x 300b reagent set should yield 15 Gb which is enough for about 30 genomes. A single HiSeq run can produce 500 Gb of data, which is enough for 1000 bacterial genomes! The challenge then becomes preparing and managing the DNA samples and analyzing the data.

Most Next Generation sequencers produce files or file combinations that include both the sequence,

and the "Phred" quality score for each position based on metrics read by the instrument. $Q = -10 \log_{10}P$ where P is the probability of a base-call error. (q=13 ~ p=0.05). Thus, high Q scores correspond to high quality sequence and low probability of incorrect base calls. Low quality sequences should be removed before assembly. From Wikipedia:

Phred quality sco	res are logarithmically linked to erro	or probabilities
Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10000	99.99%

E. Computer/program requirements for data analysis

Once the reads corresponding to a single sample are obtained and filtered for quality, overlaps in the sequences can be used to assemble the reads into larger contiguous sequences or "contigs" There are many algorithms for assembly, but most of the free ones run in a linux environment. The limitations in experience with and access to linux for most students and faculty teaching undergraduates presents significant problems for the assembly of raw sequence data. The options for the Windows operating system are more limited. Here we will use NextGene by Softgenetics on the Juniata GCAT-SEEK server for quality filtering and primary assembly. However, by the end of the summer, it is expected that a web-based tool (RAST2) will be available for assembly and annotation. Given the ease and minimal expense to sequence a bacterial genome, high quality web-based tools are essential for this capability to reach the masses.

The NextGene assembly will be uploaded to the Rapid Annotation with Subsystem Technology (RAST) Website (<u>http://rast.nmpdr.org/</u>) (Aziz et al., 2008, Overbeek et al., 2014) for automated annotation. The sequence-based comparison tool will compare the assembly to related genomes, which then allows the development of hypotheses regarding which contigs are adjacent and either overlapping or separated by gaps. The contigs can be manually edited and reordered using Microsoft Word, then re-uploaded to RAST for Re-Annotation

F. Time line of module

Day 2 – Tuesday June 3, 2014, Session 1a 1:00 - 3:00 – DNA Isolation (wet lab – Heim 106)

- Isolate gDNA
- Set up PCR with 16S rRNA primers

Day 2 – Tuesday June 3, 2014, Session 1b 3:00 - 5:00 – Sequence Assembly (Heim computer lab)

- Sequencer output
- Assembly How to minimally assemble, annotate, and analyze a genome sequence
 - Log into Juniata Server (192.112.102.20) using Remote Desktop Connection and the Username and Password provided.
 - o Download sequence data from Sequencing Center Server, unzip files
 - o Quality Filter reads & Primary assembly with NextGene by Softgenetics

Day 3 – Wednesday June 4, 2014 - Session 2a - 9:00-10:15 –Assembly → Annotation

- Examine Assembly with NextGene Viewer
- Download Assembly to a flash drive, examine files
- Upload to RAST for Automated Annotation
- Retrieve related genomes from GenBank, upload to RAST

Day 3 - Wednesday June 4, 2014 - Session 2b - 10:30-12:00 - Assessment of DNA Quality

- Prepare, run gel with quantitation standards, gDNA, PCR products.
- Measure DNA concentration with Qubit.

Day 3 – Wednesday June 4, 2014 - Session 3a - 1:00-2:00 – DNA QC documentation

• Examine gel, Prepare documentation to send with DNA to sequencing facility

Day 3 – Wednesday June 4, 2014 - Session 3b - 2:00-5:00 – Use of automated annotation

- Review annotation results, confirm ID of sequence
- Contig deletion, reordering, manual assembly, gap identification.
- Upload revised contigs to RAST.

Day 4 – Thursday, June 5, 2014 – Session 4 - 9:00-12:00 – Comparative Genomics

• Compare genomes of related organisms in terms of gene content (core genomes and unique genes), subsystems present, metabolic mapping, dot plots, microbial phylogenomics.

Day 4 – Thursday, June 5, 2015 – Session 5 - 1:00-5:00 – Prep for Publication

- MIGS = Minimum Information about a Genome Sequence
- How to prepare data for submission to NCBI

G. Protocols

Day 2 – Tuesday June 3, 2014, Session 1a 1:00 - 3:00 – DNA Isolation (wet lab – Heim 106)

G1a. DNA Isolation

A journey of a thousand miles begins with one step (Chinese philosopher, Lao-tzu).

The isolation of genomic DNA from most bacteria is rather straightforward, and there are several kits available from different manufacturers. We typically use the Qiagen Blood and Tissue Kit because the kit can be used with different types of samples and has consistently provided good results in the hands of even inexperienced students.

Procedure (From the Qiagen DNeasy Blood and Tissue Kit, July, 2006)

- 1. Harvest cells (maximum 2 x 10⁹ cells) from 1 mL of overnight culture in a microcentrifuge tube by centrifuging for 10 min at 5000 x g (7500 rpm). Discard supernatant.
- 2. Resuspend bacterial pellet in 180 μl enzymatic lysis buffer (20 mM Tris·Cl, pH 8.0, 2 mM sodium EDTA, 1.2% Triton_®X-100, Immediately before use, add lysozyme to 20 mg/ml.)
- Incubate for 30 min at 37°C to digest cell wall.
 After incubation, heat the heating block or water bath to 56°C if it is to be used for the incubation in step 5.
- 4. To remove proteins, add 25 μl proteinase K and 200 μl Buffer AL (without ethanol). Mix by vortexing.
 Note: Do not add proteinase K directly to Buffer AL.
 Ensure that ethanol has not been added to Buffer AL
- 5. Incubate at 56°C for 30 min.
- 6. Add 200 μl ethanol (96–100%) to the sample, and mix thoroughly by vortexing. It is important that the sample and the ethanol are mixed thoroughly to yield a homogeneous solution. A white precipitate may form on addition of ethanol. It is essential to apply all of the precipitate to the DNeasy Mini spin column. This precipitate does not interfere with the DNeasy procedure.
- 7. Pipet the mixture from above (including any precipitate) into the **DNeasy Mini spin column** placed in a 2 ml collection tube (provided). Centrifuge at 6000 x g (8000 rpm) for 1 min. Discard flow-through and collection tube.* **The DNA is now bound to the spin column membrane.**
- Place the DNeasy Mini spin column in a new 2 ml collection tube (provided), add 500 μl Buffer AW1, and centrifuge for 1 min at 6000 x g (8000 rpm). Discard flow-through and collection tube.* The DNA is still bound to the spin column membrane.
- 9. Place the DNeasy Mini spin column in a new 2 ml collection tube (provided), add 500 μl Buffer AW2, and centrifuge for 3 min at 20,000 x g (14,000 rpm) to dry the DNeasy membrane. Discard flow-through and collection tube. The DNA is still bound to the spin column membrane.

It is important to dry the membrane of the DNeasy Mini spin column, since residual ethanol may interfere with subsequent reactions. This centrifugation step ensures that no residual ethanol will be carried over during the following elution. Following the centrifugation step, remove the DNeasy Mini spin column carefully so that the column does not come into contact with the flow-through, since this will result in carryover of ethanol. If carryover of ethanol occurs, empty the collection tube, then reuse it in another centrifugation for 1 min at 20,000 x g (14,000 rpm).

- 10. Place the DNeasy Mini spin column in a clean 1.5 ml or 2 ml microcentrifuge tube (not provided), and pipet 200 μl Buffer AE directly onto the DNeasy membrane. Incubate at room temperature for 1 min, and then centrifuge for 1 min at 6000 x g (8000 rpm) to elute. Elution with 100 μl (instead of 200 μl) increases the final DNA concentration in the eluate, but also decreases the overall DNA yield. The DNA is now in the eluate (liquid) that came through the column.
- 11. Recommended: For maximum DNA yield, repeat elution with 100 μ l as described in step 10. This step leads to increased overall DNA yield. A new microcentrifuge tube can be used for the second elution step to prevent dilution of the first eluate.

G1b. PCR amplification of rRNA gene fragment

The purpose of this specific PCR is to ensure that there are no inhibitors contaminating the DNA sample, and ideally to sequence the PCR product via the Sanger method to confirm that the DNA is from the expected organism. We don't need to spend \$200 to sequence *E.coli* again!

Design of oligonucleotide primers to amplify and sequence ribosomal RNA genes.

The 16S rRNA gene is present in all Bacteria and Archaea. Certain sequences within the gene have not changed much in billions of years due to their essential nature for the function of the 16S rRNA gene product. These conserved sequences can be used as primer annealing sites to amplify the 16S rRNA gene by the Polymerase Chain Reaction (PCR). Many researchers around the world use the same common set of "Universal" oligonucleotide primers that we will use today. (Lane, 1991)

27f - 5' - AGAGTTTGATCMTGGCTCAG 1492r - 5' - TACGGYTACCTTGTTACGACTT

The 16S rRNA gene is a little larger than 1500 bp, so these primers will amplify nearly the full length gene. Notice that there are some non-standard letters (M,Y) in the primer sequences. These correspond to "degenerate" positions, i.e. positions that are less highly conserved, so that more than one base must be included to be "Universal". Standard nucleotide naming conventions are listed below

IUPAC Nucleic acid codes

A = Adenine	C = Cytosine
G = Guanine	T = Thymine
	U = Uracil
R = Purine (A or G)	Y = Pyrimidine (C, T, or U)
M = C or A	K = T, U, or G

W = T, U, or A	S = C or G
B = C, T, U, or G (not A)	D = A, T, U, or G (not C)
H = A, T, U, or C (not G)	V = A, C, or G (not T, not U)
N = Any base (A, C, G, T, or U)	

Thus, half of the 27f primers have a C at position 12, and half have an A. Likewise, half of the 1492r primers have a C at position 6 and half have a T. During oligonucleotide synthesis, this is accomplished by adding a mixture of the desired nucleotides when adding the nucleotide to the specified position.

During the Polymerase Chain Reaction (PCR), heating of the double stranded template DNA to 94°C separates the two strands. Upon cooling to 55°C, the primers will hybridize (base pair) with their complementary sequences on the template DNA. Heating to 72°C allows the thermal stable Taq DNA polymerase to add new nucleotides to end of the primer to produce double stranded DNA. This process is continued in a thermal cycler to produce in excess of 10⁹ copies of the DNA fragmemt defined by the two primers.

Procedure:

- 1. Obtain and label a 0.2 mL thin wall PCR tube for each sample and an extra as a control.
- 2. Prepare a master mix containing the following for each PCR (plus an extra half for good luck/pipetting errors)
 - 12.5 μL 2x Taq Premix (contains enzyme, buffer, dNTPs)
 4 μL Primer 27f (5 μM) 5' -AGAGTTTGATCMTGGCTCAG 3'
 4 μL Primer 1492r (5 μM) 5' -TACGGYTACCTTGTTACGACTT 3'
 3.5 μL dH₂O
- 3. Pipette 24 μL master mix into each PCR tube, **add 1 μL of the appropriate DNA sample** or sterile water (negative control).
- 4. Close tubes, load in thermal cycler, initiate thermal cycling program.

<u>Program = rRNA.fl</u>		
Phase 1 (initial denaturation) - 1 cycle	Initial denaturation	2 min . @ 94°C
Phase 2 (standard cycle) 35 cycles	standard denaturation Primer annealing Primer extension ²	30 sec. @ 94°C 30 sec. @ 50°C 1.5 min @ 72°C
<u>Phase 3 (extra extension) - 1 cycle</u>	Primer extension	9 min. @ 72°C

Day 2, Tuesday June 3, 2014 Session 1b - 3:00-5:00 – Download, Filter & Assemble Data

G1c. Primary Assembly

- 1. Login to lab computer with userid: Guest2 pw: BiolOgyDept
- Use Remote Desktop Connection to log into the Juniata GCAT-SEEK Server (192.112.102.20). Use the username and password provided to you (in computer lab, click "Use another account"). This server has 64 GB RAM, sufficient for a reasonably rapid assembly of prokaryotic genomes.
- 3. Use Google Chrome to visit the sequencing center download site <u>https://lims.cgb.indiana.edu/gs454/</u> <u>JeffNewman_Lycoming/</u> and login with username and password provided to you.
- Right click on the desired file, choose "save file as" and specify an appropriate download location (your folder on the data drive).
- 5. On the Start menu, choose 7-Zip File Manager, then browse to your files, select them, click the extract button, then OK. Close the 7-Zip File Manager.
- 6. Using Windows File Manager, move the uncompressed

R2 file to the R1 folder, delete the R2 folder, and simplify the R1 folder name.



- Double click on NextGene to launch the program.
 Select: Illumina, de novo assembly, sequence assembly and click next.
- 8. Click the format conversion button, then click add, then select the two fastq files. Remove low quality data using the settings shown at right, and click OK.
- 9. After conversion has been completed, use the file manager and review the conversion log text files to note the percentage of reads converted. After beginning the assembly process, return to these documents to **discuss the meaning and significance of each line in the conversion log**
- 10. On the subsequent page, click load, select the two successfully converted *.fasta files and click Next.
- 11. Assemble using the default settings shown at right.
- 12. Click Finish, then click, Run NextGene. Depending on the server load and number of sequences, the assembly may take from 30 min to several hours to complete. Allow the assembly to run overnight.

ut:	
Data\2014 Worksh	op-Prok\jeff test\GSF634-31-Flavo-sp-R30-53_S23\GSF634-31-Flavobacteri. Add
	Remove
	Remove All
ile Format Type	
Illumina	FASTQ
C SOLID	CSFASTA->FASTA
C Roche	FNA/QUAL
C Ion Torrent	FASTQ
C Other	SAGE Library
ttings	
Htings Median Score 1 Called Base Nu Trim or Reject Paired Reads D	Immunol Immunol <t< td=""></t<>
ttings Median Score 1 Called Base Nu Trim or Reject Paired Reads D Remove 5'	Import of Each Read >= 25 mbor of Each Read >= 25 Read when >= 3 Base(s) with Score <=
ttings Called Base Nu Called Base Nu Trim or Reject Paired Reads D Remove 5'	Investod 20 IV Nax # of Uncalled Bases <= 3 mber of Each Read >= 25 3 Base(s) web Score <= 16
ttings ✓ Median Score 11 ✓ Called Base Nu ✓ Trim or Reject ✓ Pared Reads D □ Remove 5' 0 □ Keep Only Bas □ Trim by Sequel	hveshold >= 20 IV Max # of Uncalled Bases <=
ttings V Median Score 1 Called Base Nu V Trim or Reject Parred Reads D Remove 5' Keep Only Bas Trim by Sequer Trim by Sequer	http://discupression.com/discupression/discupresdindiscupression/discupression/discupression/discupressio

	Snow Project Log:
ep	Assembly
	Save the Original Sequences with Assembled Ones
	Assembly Method
Application	C De Bruijn (PE Assembly
	PE Assembly Options
Load Data	I Pared Read Data
	✓ Long Library Size >1000 Bases) Section Size 400
	Word Length 21 Minimum Scaffold Length 150
ondensation	High Coverage Limited: May Coverage = 350
	Thigh corologo childen that corologo - 1000
	Final Contig Merging Reduce Memory Usage
Assembly	Final Contig Merging
Assembly	🏹 Final Contig Merging 🛛 🗂 Reduce Memory Usage
Assembly	Final Conkig Merging Final Conkig Merging
Assembly	F Final Contig Merging F Reduce Memory Usage
Assembly	F Final Contig Merging F Reduce Memory Usage
Assembly Alignment	Final Contig Merging T Reduce Memory Usage
Assembly Alignment Post Processing	Final Contig Merging Fi Reduce Memory Usage
Assembly Alignment Post Processing	F Final Contig Merging F Reduce Memory Usage
Assembly Alignment Post Processing	Final Contig Merging Reduce Memory Usage
Assembly Alignment Post Processing	Final Contig Merging Ti Reduce Memory Usage
Assembly Alignment Protessing	F Final Contig Merging F Reduce Memory Usage
Assembly Alignment Post Processing	Final Contig Merging Reduce Memory Usage Default Settings Save Settings Load Settings



Day 3 – Wednesday June 4, 2014 - Session 2a - 9:00-10:15 –Assembly → Annotation

G2a1 – Examine & Download Assembly

- 1. Login to a lab computer, and use Remote Desktop Connection to login to the GCAT-SEEK Windows server at Juniata.
- 2. After the assembly was completed, it should have been opened in the NextGene Viewer shown above. From the image, one can get a sense of the quality of the assembly. For example, red lines are used to separate the contigs, and the grey lines indicate the coverage of the genome. It is apparent that about half of the sequence data (4000 kb) has a little over 50x coverage and is assembled into a few large contigs, while another half has less than 10x coverage in many small contigs. This is a characteristic pattern of a contaminated DNA sample. One can, however, use the difference in coverage to delete the contaminant sequences, focusing just on the large, high coverage contigs.
- 3. Use the file manager to review the assembly files. **Copy** the two convert log text files into the output folder, then select all of the files smaller than 10 Mb, right click and send to a compressed folder for downloading. Right click on the compressed folder, choose copy. Minimize remote desktop connection, and on your local machine, paste the file onto your flashdrive, and unzip the compressed folder.

4. Examine the downloaded files. In particular, take note of the StatInfo.txt, the

AssembledSequences.fasta, the ContigMerge and ScaffoldContigs files.

In the *StatInfo.txt document to review the assembly process and resulting statistics

- Total Reads Number: 2034788
- Matched Reads Number: 1983986
- Unmatched Reads Number: 50802
- Assembled Sequences Number: 61
- Average Sequence Length: 57497
- Minimum Sequence Length: 158
- Maximum Sequence Length: 641985
- N50 Length: 366076

[Final Contig Merge Results Statistics Report]

- Final Contig Merge Sequences Number: 13
- Final Contig Merge Average Sequence Length: 269063
- Final Contig Merge Minimum Sequence Length: 173
- Final Contig Merge Maximum Sequence Length: 856388
- Final Contig Merge N50 Length: 586767

[Alignment Statistics Information]

- Matched Reads Count: 1977550
- Unmatched Reads Count: 0
- Number of Matched Bases: 562514128
- Number of Unmatched Bases That are Recorded as Mutations: 605431
- Number of Unmatched Bases That are NOT Recorded as Mutations: 2353746
- Average Read Length: 285
- Average Coverage: 161
- Reference Length: 3507364
- Number of Covered Bases: 3507355

What does each statistic mean, what is the significance of each?

G2a2. Upload sequences to RAST for initial annotation.

At this stage, there may still be more than one hundred contigs. If the genome is from a novel species, or a species for which there is no reference sequence, one can still use the most closely related genome sequence available (preferably within the same genus) to help determine the proper order and orientation of contigs. One method to do this involves an automated annotation using the Rapid Annotation with Subsystems Technology (RAST) website (http://rast.nmpdr.org/) (Aziz et al., 2008). This annotation will identify genes within the sequence, and when compared to one or more related, annotated, and preferably finished genomes, can suggest which contigs are adjacent to each other and possibly overlapping.

Procedure

Use Firefox to login to the RAST website (<u>http://rast.nmpdr.org/</u>). Click "Your Jobs" → "Upload New Job". On the subsequent page, Browse to the Assembly2.fasta file from the previous section, then click "Use the data and go to step 2

	RAST Server - Jobs Overview +
€)@	rast. nmpdr.org /rast.cgi?page=Jobs
Most	Visited 🥹 Getting Started 🗌 Suggested Sites 🗍 Web Slice Gallery
For mo	Fe information about The SEED please visit <u>theSEED.org.</u>
Home	Your Jobs
	Jobs Overview
As	(Upload New Job 1 16:18:02 2013, there are 100 jobs in
Job	Private Organism Preferences
The ov	erview below list all genomes currently processed and the progress on the annotation. To get a more deta
The ov	erview below list all genomes currently processed and the progress on the annotation. To get a more deta 9 of questions or problems using this service, please contact: <u>rast@mcs.anl.gov</u> .
The ov In case Progre	erview below list all genomes currently processed and the progress on the annotation. To get a more deta e of questions or problems using this service, please contact: <u>rast@mcs.anl.gov</u> . ess bar color key:

2 Open a new tab in your browser go to	🗲 👁 rast.nmpdr.org/rast.cgi
	🖉 Most Visited 🥹 Getting Started 🗍 Suggested Sites 🗍 Web Slice Gallery
the NCBI website	Upload a Genome
(<u>http://www.ncbi.nlm.nih.gov/</u>)	Review genome data
and perform a Taxonomy search for	We have analyzed your upload and have computed the following information.
the organism. Copy the NCBI	Contin statistics
taxonomy ID into the appropriate	
hav on the BAST name and click the	Statistic As uploaded After splitting into scatfolds Sequence size 3279168 3277347
box on the RAST page and there the	Number of contigs 901 1253
lookup button. Click "Use this data	GC content (%) 40.0 40.0
and go to step 3"	Median sequence size 1703 1178
	Mean sequence size 3639.5 2615.6
	Longest contig size 43/0/ 39383
	Please enter or verify the following information about this organism:
	Required information:
	(leave blank if NCBL Taxonomy ID unknown)
	Find the taxonomy id for your organism by searching for its name in the NCBI taxonomy brow
	Taxonomy string: Bacteria; Bacteroidetes/Chlorobi group; Bacteroidetes; Flavobacteriia;
	Flavobacteriales; Flavobacteriaceae; Chryseobacterium; Chryseobacterium k
	Domain: 💿 Bacteria 💿 Archaea 💿 Virus
	Genus: Chryseobacterium
	Species: koreense
	Strain: CCUG 49689
	Constin Coder
3. Enter the requested information, change	Home Your Jobs
the Figram version to the highest	Upload a Genome
number, check "build metabolic	Complete Upload
model", then click "Finish the	By answering the following questions you will help us improve our ability to track problems in processing your genome:
unload"	Optional information:
upioau	Sequencing Method Sanger Mix of Sanger and Pyrosequencing Pyrosequencing other
	Number of contigs 501-1000 -
	Number of contigs 501-1000 • Average Read Length 260
	Number of contigs 501-1000 • Average Read Length 260 (leave blank if unknown)
	Number of contigs 501-1000 • Average Read Length 260 (leave blank if unknown) Please consider the following options for the RAST annotation pipeline: RAST Annotation Settings:
	Number of contigs 501-1000 • Average Read Length 260 (leave blank if unknown) Please consider the following options for the RAST annotation pipeline: RAST Annotation Settings: Select gene caller RAST Please select which type of gene calling you would like RAST to perform. Note the backfilling of gaps.
	Number of contigs 501-1000 • Average Read Length 260 (leave blank if unknown) Please consider the following options for the RAST annotation pipeline: RAST Annotation Settings: Select gene caller RAST Select FIGFam version for Release59 • Choose the version of FIGfams to be used to process this genome.
	Number of contigs 501-1000 • Average Read Length 260 (leave blank if unknown) Please consider the following options for the RAST annotation pipeline: RAST Annotation Settings: Select gene caller RAST Please select which type of gene calling you would like RAST to perform. Note the backfilling of gaps. Select FIGFam version for Release59 • Choose the version of FIGfams to be used to process this genome. Hthis run Automatically fix errors? Ves
	Number of contigs 501-1000 • Average Read Length 260 (leave blank if unknown) Please consider the following options for the RAST annotation pipeline: RAST Annotation Settings: Select gene caller RAST • Please select which type of gene calling you would like RAST to perform. Note the backfilling of gaps. Select FIGfam version for Release59 • Choose the version of FIGfams to be used to process this genome. this run Automatically fix errors? Automatically fix errors? Yes Fix frameshifts? If you wish for the pipeline to fix frameshifts, check this option. Otherwise fram
	Number of contigs 501-1000 • Average Read Length 260 (leave blank if unknown) Please consider the following options for the RAST annotation pipeline: RAST Annotation Settings: Select gene caller RAST Please select which type of gene calling you would like RAST to perform. Note the backfilling of gaps. Select FIGfam version for Release59 • Choose the version of FIGfams to be used to process this genome. this run Automatically fix errors? Automatically fix errors? Yes Fix frameshifts? Yes If you wish for the pipeline to fix frameshifts, check this option. Otherwise fram Build metabolic model? Yes If you wish AST to build a metabolic model for this genome, check this option.
	Number of contigs 501-1000 • Average Read Length 260 (leave blank if unknown) Please consider the following options for the RAST annotation pipeline: RAST Annotation Settings: RAST Annotation Settings: Please select which type of gene calling you would like RAST to perform. Note the backfilling of gaps. Select FIGfam version for Release59 • Choose the version of FIGfams to be used to process this genome. this run Automatically fix errors? Yes Fix frameshifts? Yes If you wish for the pipeline to fix frameshifts, check this option. Otherwise fram Build metabolic model? Yes Backfill gaps? Yes If you wish for the pipeline to bast large gaps rm issing genes, check this option. Backfill gaps? Yes If you wish for the pipeline to bast large gaps rm issing genes, check this option.
	Number of contigs Sol1:1000 • Average Read Length 260 (leave blank if unknown) Please consider the following options for the RAST annotation pipeline: RAST Annotation Settings: RAST Annotation Settings: Please select which type of gene calling you would like RAST to perform. Note the backfilling of gaps. Select FIGfam version for Release59 • Choose the version of FIGfams to be used to process this genome. this run Automatically fix errors? Yes Fix frameshifts? Yes If you wish for the pipeline to fix frameshifts, check this option. Otherwise fram Build metabolic model? Yes Backfill gaps? Yes If you wish for the pipeline to bast large gaps for missing genes, check this option. Backfill gaps? Yes If you wish for the pipeline to bast large gaps for missing genes, check this option. Set verbose level 0 Set this to the verbosity level of choice for error messages.
	Number of contigs Sol1:1000 • Average Read Length 260 (leave blank if unknown) Please consider the following options for the RAST annotation pipeline: RAST Annotation Settings: RAST Annotation Settings: Please select which type of gene calling you would like RAST to perform. Note the backfilling of gaps. Select FIGFam version for Release59 • Choose the version of FIGFams to be used to process this genome. this run Automatically fix errors? Yes The automatic annotation process may run into problems, such as gene candidates), private rools the pipeline to fix frameshifts? Build metabolic model? Yes If you wish for the pipeline to bast large gaps for missing genes, check this option. Backfill gaps? Yes If you wish for the pipeline to bast large gaps for missing genes, check this option. Build metabolic model? Yes If you wish for the pipeline to bast large gaps for missing genes, check this option. Backfill gaps? Yes If you wish for the vipeline to bast large gaps for missing genes, check this option. Set verbose level 0 Set this to the verbosity level of choice for error messages. Disable replication Yes Even if this job is identical to a previous job, run it from scratch.
	Number of contigs Sol1:1000 • Average Read Length 260 (leave blank if unknown) Please consider the following options for the RAST annotation pipeline: RAST Annotation Settings: Select gene caller RAST Please select which type of gene calling you would like RAST to perform. Note the backfilling of gaps. Select FIGFam version for Release59 Choose the version of FIGFams to be used to process this genome. this run Automatically fix errors? Automatically fix errors? Yes The automatic annotation process may run into problems, such as gene candidates), r Fix frameshifts? Yes Build metabolic model? Yes If you wish for the pipeline to fix frameshifts, check this option. Backfill gaps? Yes Turn on debug? Yes Turn on debug? Yes Disable replication Yes Ves Even if this to the verbosity level of choice for error messages. Disable replication Yes

4. In a new browser tab, Go to <u>http://www.ncbi.nlm.nih.gov/genome/browse/</u> and search for your organism's genus.

C f	lm.nih.gov/genome/	/browse/								5
NCBI Resources 🕑 How To 🖸	2									Sign in to NCE
Genome	e •							Search		
enome Information by or	ganism									
lavobacterium (taxid:237)			Search by o	rganism	Clear			Down	load Repo	rts from FTP site
Duoniow [24] Eukonyotoo [0]	Brokonyotoc [20]	inusos (0) Plasmida (4	1							
Diverview [24] Luxaryotes [0]	Flokalyotes [50]		1							
First Previous	Shown: 1 - 24 out (of 24 items	Next L	ast				Dov	vnload sele	ected records
Organiem/Namo	Kingdom	Group			SubGroup	Size (Mb)	Chr	Organallas	Disemide	BioDrojecte
Organisminanie	All	All		▼ All		▼ 3126 (MD)	CIII	organeties	Flashilus	DioFrojecta
Flavobacterium	Bacteria	Bacteroidetes/Chl	orobi group		Bacteroidetes	5.34	-	-	1	9
Flavobacterium antarcticum	Bacteria	Bacteroidetes/Chl	orobi group		Bacteroidetes	3.08	-	-	-	1
Flavobacterium branchiophilum	Bacteria	Bacteroidetes/Chl	orobi group		Bacteroidetes	3.56	1	-	1	1
Flavobacterium cauense	Bacteria	Bacteroidetes/Chl	orobi group		Bacteroidetes	3.11	-		-	1
Flavobacterium columnare	Bacteria	Bacteroidetes/Chl	orobi group		Bacteroidetes	3.16	1	-	-	1
Flavobacterium daejeonense	Bacteria	Bacteroidetes/Chl	orobi group		Bacteroidetes	4.24	-		-	1
Flavobacterium denitrificans	Bacteria	Bacteroidetes/Chl	orobi group		Bacteroidetes	4.82	-	-	-	1
Flavobacterium enshiense	Bacteria	Bacteroidetes/Chl	orobi group		Bacteroidetes	3.39	-		-	1
Flavobacterium filum	Bacteria	Bacteroidetes/Chl	orobi group		Bacteroidetes	3.19	-	-	-	1
Flavobacterium frigidarium	Bacteria	Bacteroidetes/Chl	orobi group		Bacteroidetes	3.63	-	-	-	1
Flavobacterium frigoris	Bacteria	Bacteroidetes/Chl	orobi group		Bacteroidetes	3.93	-	-	-	1
Flavobacterium gelidilacus	Bacteria	Bacteroidetes/Chl	orobi group		Bacteroidetes	3.44	-	-	-	1
Flavobacterium indicum	Bacteria	Bacteroidetes/Chl	orobi group		Bacteroidetes	2.99	1	-	-	1
Flavobacterium johnsoniae	Bacteria	Bacteroidetes/Chl	orobi group		Bacteroidetes	6.1	1	-	-	1
Flavobacterium limnosediminis	Bacteria	Bacteroidetes/Chl	orobi group		Bacteroidetes	3.47	-	-	-	1
Flavobacterium psychrophilum	Bacteria	Bacteroidetes/Chl	orobi group		Bacteroidetes	2.86	1	-	1	1

5. Click on the link for an organism of interest (closest relatives), then the INSDC link or the RefSeq link followed by the wgs link. Download the GenBank file, unzip (twice), and upload to RAST for annotation.

STREDI K	Resources 🖄 Ho	w To 🗹									Sign in to NCBI
Senome	Ge	enome 🔻								Search	
		Limits	Advanced								Help
isplay Setti	ings: 🕑 Overview							Sei	nd to: ☑		
										Tools	
Organism (Overview ; Geno	me Assembly and Annota	tion report [1]						ID: 23235	BLAST Genome	
Flavok	bacterium	ı limnosedimiı	nis								
Flavobacteri	ium limnosedimin	is overview								Polated information	
Lineage: Bac	cteria[4255]: Bacter	oidetes[377]: Flavobacteriia	1291: Flavobacteriales[126	61: Flavobacteriac	eae[118]: Fi	avobacterium	[24]: Flavo	acterium		Related information	
										DIOFIOJECI	
limnosedimi	iinis[1]									Componente	
Iimnosedimi	iinis[1] hacterium limno	sediminis IC2002								Components	
• E Flavol Submitter:	inis[1] <i>bacterium limno</i> : Seoul National U	sediminis JC2902 Iniversity								Components Protein	
• E Flavol Submitter: Genome ser	binis[1] bacterium limno Seoul National U equencing of the ty	sediminis JC2902 Iniversity ype strain of Flavobacteriur	n limnosediminis sp. no	v. for taxonomic	studies.					Components Protein Taxonomy	
• E Flavol Submitter: Genome set Status: Con Morpholog	inis[1] bacterium limno Seoul National U equencing of the ty ntig IV: Gram:Negative	sediminis JC2902 Iniversity ype strain of Flavobacteriur	n limnosediminis sp. no	v. for taxonomic	studies.					Components Protein Taxonomy	
Imnosedimi ■ Flavol Submitter: Genome set Status: Con Morpholog Environme	inis[1] bacterium limno : Seoul National U :quencing of the ty ntig gy: Gram:Negative ent: Salinity:NonH	sediminis JC2902 Iniversity /pe strain of Flavobacteriur alophilic, OxygenReq:Aero	n limnosediminis sp. no bic, OptimumTemperatu	v. for taxonomic ure:30C, Temper	studies. atureRango	e:Mesophilic	, Habitat:/	Aquatic		Components Protein Taxonomy Recent activity	
Imnosedimi • E Flavolt Submitter: Genome set Status: Com Morpholog Environme Phenotype	iniis[1] bacterium limno : Seoul National U quencing of the ty ntig yy: Gram:Negative ent: Salinity:NonH e: BioticRelations]	sediminis JC2902 niversity pe strain of Flavobacteriur alophilic, OxygenReq.Aerc nip.FreeLiving	n limnosediminis sp. no bic, OptimumTemperatu	v. for taxonomic ure:30C, Temper	studies. atureRangi	e:Mesophilic	, Habitat:/	Aquatic		Components Protein Taxonomy Recent activity	Turn Off Clear
Imnosedimi ■ Flavol Submitter: Genome se Status: Con Morpholog Environme Phenotype Assembly: BioProject:	initis[1] bacterium limno : Seoul National U quencing of the ty ntig yy: Gram:Negative ent: Salinity:NonH :: BioticRelationsl GCA_00049855 g: PRJNA229861	sediminis JC2902 niversity ype strain of Flavobacteriur alalophilic. OxygenReq:Aerc nip:FreeLiving 1 Fiim1.0 scaffolds: 56 cc PRINA20619	n limnosediminis sp. no bic, OptimumTemperatu ontigs: 56 N50: 311,990	v. for taxonomic ure:30C, Temper) L 50: 4	studies. atureRang	e:Mesophilic	, Habitat:/	quatic		Components Protein Taxonomy Recent activity Flavobacterium limnosedimi	T <u>urn Off</u> <u>Clear</u> nis
Imnosedimi ■ Flavol Submitter: Genome se Status: Com Morpholog Environme Phenotype Assembly: BioProject: Type	initis[1] bacterium limno : Seoul National U quencing of the ty- ntig yy: Gram:Negative ent: Salinity:NonH :: BioticRelationsl GCA_000498535 Is: PRJNA229861, Name	sediminis JC2902 niversity ype strain of Flavobacteriur alophilic, OxygenReq.Aero in p: FreeLiving .1 Flim1.0 scaffolds: 56 co PRJNA206419 RefSeq	n limnosediminis sp. no bic, OptimumTemperatu ontigs: 56 N50: 311,990 INSDC	v. for taxonomic ure:30C, Temper) L50: 4 Size (Mb)	studies. atureRange GC%	e:Mesophilic Protein	, Habitat:A	tRNA	Gene	Components Protein Taxonomy Recent activity Flavobacterium limnosedimi	 <u>Turn Off Clear</u> nis Genome
Imnosedimi • ☐ Flavot Submitter: Genome se Status: Con Morpholog Environme Phenotype Assembly: BioProject: Type	initis[1] bacterium limno Seoul National U equencing of the ty ent; gram:Negative ent; Salinity:NonH :: BioticRelationsl GCA_000498535 is: PRJNA229861; Name master WGS	sediminis JC2902 niversity ype strain of Flavobacteriur alaophilic, OxygenReq:Aeron nip:FreeLiving , 1 Film1.0 scaffolds: 56 cr PRJNA206419 RefSeq NZ_AVG60000000 1	n limnosediminis sp. no bic, OptimumTemperatu ontigs: 56 N50: 311,990 INSDC AVGG0000000.1	v. for taxonomic ure:30C, Temper 0 L50: 4 Size (Mb) 3.47	studies. atureRange GC% 38.5	e:Mesophilic Protein 3,117	, Habitat: A	tRNA 48	Gene 3,167	Components Protein Taxonomy Recent activity Flavobacterium limnosedimi Flavobacterium sp. R30-53	T <u>urn Off</u> <u>Clear</u> nis Genome
Imnosedimi • ☐ Flavot Submitter: Genome se Status: Con Morpholog Environme Phenotype Assembly: BioProject: Type	inis[1] bacterium limno Seoul National U squencing of the ty ntig ny: Gram:Negative ent: Salinity:NonH : BioticRelationsl GCA_000498535 is: PRJNA229861 Name master W3S	sediminis JC2902 niversity pe strain of Flavobacteriur alophilic, OxygenReq:Aero ip:FreeLiving 1. Film1.0 scaffolds: 56 cc .PRJNA205419 RefSeq NZ_AVGG00000000.1	n limnosediminis sp. no bic, OptimumTemperatu ontigs: 56 N50: 311,990 INSDC Avesgooocooco.1	v. for taxonomic ure:30C, Temper) L50: 4 Size (Mb) 3.47	studies. atureRange GC% 38.5	e:Mesophilic Protein 3,117	, Habitat: A rRNA 2	tRNA 48	Gene 3,167	Components Protein Taxonomy Recent activity Flavobacterium limnosedimi Flavobacterium sp. R30-53	Turn Off Clear nis Genome taxonomy
Imnosedimi ■ Flavot Submitter: Genome se Status: Com Morpholog Environme Phenotype Assembly: BioProject Type	inis[1] bacterium limno sevel National U quencing of the ty mitig yy: Gram.Negative prit. Salinity: NonH s: BioticRelations) is: PRJNA229861 Name master WGS	sediminis JC2902 niversity pe strain of Flavobacteriur alophilic, OxygenReq:Aero ip:FreeLiving 1. Film1.0 scaffolds: 56 cc .PRJNA205419 RefSeq NZ_AVGG00000000 1	n limnosediminis sp. no bic, OptimumTemperatu ontigs: 56 N50: 311,990 INSDC Avescoocococo.1	v. for taxonomic ure:30C, Temper) L50: 4 Size (Mb) 3.47	studies. atureRange GC% 38.5	e:Mesophilic Protein 3,117	, Habitat:A rRNA 2	tRNA 48	Gene 3,187	Components Protein Taxonomy Recent activity Flavobacterium limnosedimi Flavobacterium sp. R30-53 Flavobacterium sp. R30-53	Turn Off Clear nis taxonomy (1) Taxonomy
Imnosedimi • Elavol Submitter: Genome ses Status: Con Morpholog Environme Phenotype Assembly: BioProject: Type	inis[1] bacterium limno : Seoul National U quencing of the ty intig y: Gram:Negative mit: Salinity:NonH : BioticRelations] is: PRJNA229861 Name master W35	sediminis JC2902 niversity pe strain of Flavobacteriur alophilic, OxygenReq:Aero ip:FreeLiving 1. Flim1.0 scaffolds: 56 cc .PRJNA206419 .RefSeq .HZ_AV360000000.1	n limnosediminis sp. no bic, OptimumTemperatu ontigs: 56 N50: 311,990 INSDC AVGG0000000.1	v. for taxonomic ure:30C, Temper) L50: 4 Size (Mb) 3.47	studies. atureRange GC% 38.5	Protein 3,117	, Habitat:A rRNA 2	tRNA 48	Gene 3,167	Components Protein Taxonomy Recent activity Flavobacterium limnosedimi Flavobacterium sp. R30-53 Flavobacterium sp. R30-53 Flavobacterium sp. R30-53 R30-53 (0)	Turn Off Clear nis Genome taxonomy (1) Taxonomy
Imnosedimi • Elavol Submitter: Genome se Status: Con Morpholog Environme Phenotype Assembly: BioProject Type	inis[1] bacterium limmo Seoul National U quencing of the ty ntig gy: Gram:Negative mt: Salinity:NonH e: BioticRelations is: PRJNA229861 Name master WGS	sediminis JC2902 niversity ype strain of Flavobacteriur alophilic, OxygenReq:Aerc ip:FreeLiving .1 Flim1.0 scaffolds: 56 cd PRJNA206419 RefSeq NZ_AVGG0000000.1	n limnosediminis sp. no bic, OptimumTemperati ontigs: 56 N50: 311,990 INSDC Avegooocooo.1	v. for taxonomic ure:30C, Temper 0 L50: 4 Size (Mb) 3.47	studies. atureRange GC% 38.5	Protein 3.117	, Habitat:A rRNA 2	tRNA 48	Gene 3,107	Components Protein Taxonomy Recent activity Flavobacterium limnosedimi Flavobacterium sp. R30-53 Flavobacterium sp. R30-53 Ravobacterium sp. Ravobacterium sp. Ravobacterium sp. Ravobacterium sp. R	T <u>urn Off</u> Clear nis Genome taxonomy (1) Taxonomy Taxonomy

Day 3 - Wednesday June 4, 2014 - Session 2b - 10:30-12:00 - DNA Quality Control

G2b1 - Prepare 0.8% agarose gel

- 1. Attach dams to the ends of a gel tray and align comb in tray parallel with and 1-2 cm from the end of the tray.
- 2. Add 0.32 g of agarose to 40 mL water in a 125 mL erlenmeyer flask, heat mixture in microwave on high setting until mixture begins to boil (~ 1min). Do not let the solution boil over.
- Using a folded paper towel to hold the neck of the erlenmeyer flask, swirl the gel mixture well, and return to microwave. Heat for an additional 30 - 45 sec, or until mixture begins to boil. Bring to a boil a third time to get all of the agarose dissolved.
- Add 0.8 mL 50x TAE buffer, 10 μL 2 mg/mL ethidium bromide (final conc = 0.5 μg/mL), swirl to mix, pour into gel tray, allow to stand at room temp for 20 - 30 min to solidify.

G2b2 – Measure DNA concentration with Qubit Fluoometer

1. Prepare working buffer: (extra 3 samples allow for 2 standards and for pipetting error)

Qubit dsDNA Buffer: [Number of samples+3]*199µl = _____

Qubit reagent (fluorophore): [Number of samples+3]*1µl = _____

- 2. Vortex the working buffer to mix
- 3. Label Qubit Assay tubes on cap with sample ID, or S1 or S2 for the standards
- 4. For each sample, add 198 μ l of working buffer to the appropriate tube, then add 2 μ l of DNA.
- 5. For each of the two standards, add 190 μ l of working buffer to the appropriate tube, then add 10 μ l of standard.
- 6. Vortex each sample for 2-3 seconds to mix
- 7. Incubate for 2 minutes at room temperature
- 8. On the Qubit fluorometer, press DNA, then dsDNA Broad Range, then YES.
- 9. When directed, insert standard 1, close the lid, and press Read
- 10. Repeat step 9 for standard 2. This produces your two-point standard calibration.
- 11. Read each sample by inserting the tube into the fluorometer, closing the lid, and pressing **Read Next Sample**

G2b3 - Gel electrophoresis

- Fill gel chamber with 1x TAE buffer such that the level of liquid just covers center platform. Remove comb from gel, place gel tray in chamber with the wells near the negative electrode (black), add sufficient 1x TAE to just cover the gel.
- 2. Cut a small piece of parafilm, place on bench near gel, "spot" a 1-2 μ L aliquot of loading dye onto parafilm for each sample to be loaded on gel.
- 3. Load gel as outlined below by drawing sample into pipette tip and pipetting up and down onto a spot of loading dye to mix, then loading sample into well of gel. Be careful not to poke pipette tip through bottom of well. Samples should be loaded in the following order (from left right):
 Mass (ng) Kilobases

Lane 1 – 3 μL uncut λ DNA (20 ng/μL)	40 40 48	10.0 - 8.0 - 6.0 <	
Lane 2 – 3 μL uncut λ DNA (50 ng/μL)	40 32 120	5.0 4.0 3.0 -	
Lane 3 – 3 μL uncut λ DNA (80 ng/μL)	40	2.0 -	
Lane 4 – 3 μL gDNA Sample 1 st elution	10	2.00	
Lane 5 – 3 μ L gDNA Sample 2 nd elution	57	1.5 -	
Lang 6 - 5 ul 2 log ladder (500 ng total)	45	1.2 -	
Late 0 - 5 μ L 2 log ladder (500 lig total)	122	1.0 - 0.9 -	
Lane 7 – 5 μL PCR negative control	31	0.8 -	
	27	0.7 -	
Lane 8 – 5 µL PCR	23	0.6 -	
	124	0.5 -	-
4. Run gel at 50v during lunch. After fastest migrating blue dye	49	0.4 –	
(bromophenol blue) has migrated 2/3 the length of the gel, turn off	37	0.3 -	
power, carefully remove gel from chamber, drain, slide onto piece of	32	0.2 -	
plastic wrap. Photograph the gel under OV light, print on a color			
printer	61	0.1 -	

- 5. Compare the intensity of the bands in the samples you prepared to the intensity of the bands with known amounts/concentrations to estimate the concentration of DNA as best you can, and enter the data below.
- 6. Confirm that 16SrRNA fragment was successfully amplified. If possible, it is always a good idea to confirm source of DNA by Sanger sequencing of PCR product

Estimates of genomic DNA Concentration (ng/uL)

Method	1 st elution	2 nd elution
Gel electrophoresis		

Qubit

Т

Day 3 – Wednesday June 4, 2014 - Session 3b - 2:00-5:00 – Use of automated annotation

G3a - Review annotation results

Г

 Use Firefox to login to the RAST website (<u>http://rast.nmpdr.org/</u>). Click "Your Jobs" → "Jobs Overview". On the subsequent page, click "View details", then note the available downloads for the genome. Click on "Browse annotated genome in SEED viewer" 	BAST Server- Jab Details Image: Comparing Server- Jab Details
 2. In the Organism Tab near the top of the page, choose Genome Browser, then in the second column choose RNA. 3. Click the next link until you find the small subunit rRNA gene, then click the feature ID link, then the sequence link. Select and copy the sequence. 	Matthew viel basis Sections Image: Section of Sectio
 4. Login to EzTaxon at http://www.ezbiocloud.net/eztaxon . This is the best website to search for matching 16srRNA sequences because it includes only Type strains. 5. Click Identify, then paste your sequence into the box, enter a name for the sequence and click the identify button at the bottom of the page. Is the best match what you expected? 	Bit Horizeric Habitali Best Horizeric Habitali Best Horizeric Habitalia Best Horizeri Habitalia Best Horizeri Habit



7. Click on a spe	subsystem of in cific subsystem	nteres genes	t, then presen	click t in tl	on t ne o	he s rgar	ubs [,] nism	/ste of i	m sj nte	orea rest	adsh	neet	tab	to i	ider	ntify	the
	private organisms ∡▼ Pattern ∡∓	Taxonom	У _▲▼	di	splayir	ng 1 - 1	14 of 1	4									_
	Organism _★ ▼	Domain Bacter -	Variant [?]	active yes	SusA	SusB	SusC	SusD	SusG	SusE	SusF	SusR	GBE	COG	DocI	CelJ	
	<u>Chryseobacterium</u> koreense CCUG 49689 (232216.5)	Bacteria	1.x	yes	846	843	839	840					1225, 1226, 175, 50				
	Flavobacterium sp. MED217 (313593.3)	Bacteria	1.x	yes	1107		1111, 1521, 1754	<u>1110</u> , <u>1522</u>					1588				
	HTCC2559 (216432.3) Bacteroides vulgatus ATCC 8482 (435590.6)	Bacteria	1.x 1.x	yes	<u>1038</u>	<u>1303</u>	<u>1033</u> , <u>1302</u> , <u>3436</u>	<u>1301</u>				1305	<u>2978</u>		<u>888</u>		
	Parabacteroides distasonis ATCC 8503 (435591.10)	Bacteria	1.x	yes			<u>1479</u> , <u>3473</u>					<u>1612</u>					
	Flavobacteria sp. BBFL7 (156586.3) Robiginitalea biformata	Bacteria Bacteria	1.x 1.x	yes yes			<u>1452</u> , <u>2685</u> <u>1765</u>	<u>2684</u> 1763					<u>2692</u> 2603,				
	Bacteroides thetaiotaomicron VPI-5482 (226186.1)	Bacteria	1.0	yes	3702	<mark>3701</mark> , <u>4579</u>	1119, 2951, 3089, 3309, 3700,	<mark>1118</mark> , 2950, 3699, 4668, 484	3696	3698	3697	3090, 3703	003				

Г

Click the b area fam orga How	back button to re a, select the feat iliar with the ava nism's phenoty a does this comp	eturn to the Org ures in subsyste iilable informati pes? are to publishee	anism overview page, then in the subsy ms tab and browse the different catego on. What predictions would you make a d descriptions of the organism?	stems information ries to become about the
Suboyat	m Information			
Subsyste	em Information			
Subsystem s	tatistics Features in Subsystems			
export to fil	e clear all filters			
			display 15 items per page	and had
			displaying 1 - 15 of 52	next» last»
Category 🛓	▼ Subcategory ★▼	Subsystem 🛓 🔻	Role 🔭	Features
Respiration	• all			
Respiration	ATP synthases	FOF1-type ATP synthase	ATP synthase C chain (EC 3.6.3.14)	tiq 232216.5.peq.3583
Respiration	ATP synthases	E0E1 type ATP synthase	ATP synthase delta chain (EC 3.6.3.14)	fig 222216.5.peg.2254
Respiration	ATP Synciases	TOT POPE ATP SYNChase	Air synchase b chain (EC 5.0.5.14)	fig 232216.5.peg.2256
Respiration	ATP synthases	F0F1-type ATP synthase	ATP synthase beta chain (EC 3.6.3.14)	fig 232216.5.peg.3208
Respiration	ATP synthases	F0F1-type ATP synthase	ATP synthase gamma chain (EC 3.6.3.14)	fig 232216.5.peg.2129
Respiration	ATP synthases	F0F1-type ATP synthase	ATP synthase alpha chain (EC 3.6.3.14)	fiq 232216.5.peq.2253
Respiration	ATP synthases	F0F1-type ATP synthase	ATP synthase A chain (EC 3.6.3.14)	fiq 232216.5.peq.3584
Respiration	ATP synthases	F0F1-type ATP synthase	ATP synthase epsilon chain (EC 3.6.3.14)	fig 232216.5.peg.3210
Respiration	Electron accepting reactions	Terminal cytochrome C oxidases	Cytochrome c oxidase subunit CcoP (EC 1.9.3.1)	fiq 232216.5.peq.3306
Respiration	Electron accepting reactions	Terminal cytochrome C oxidases	Cytochrome c oxidase subunit CcoO (EC 1.9.3.1)	fiq 232216.5.peq.1159, fiq 232216.5.peq.3308
Respiration	Electron accepting reactions	Terminal cytochrome C oxidases	Type cbb3 cytochrome oxidase biogenesis protein CcoS, involved in heme b insertion	fig 232216.5.peg.3309
Respiration	Electron accepting reactions	Terminal cytochrome C oxidases	Cytochrome c oxidase subunit CcoN (EC 1.9.3.1)	fig 232216.5.peg.777, fig 232216.5.peg.1158.
				fig 232216.5.peg.3308
Respiration	Electron accepting reactions	Terminal cytochrome C oxidases	Type cbb3 cytochrome oxidase biogenesis protein CcoG, involved in Cu oxidation	fiq1232216.5.peq.3308 fiq1232216.5.peq.3305, fiq1232216.5.peq.3362
Respiration Respiration	Electron accepting reactions Electron accepting reactions	Terminal cytochrome C oxidases Anaerobic respiratory reductases	Type cbb3 cytochrome oxidase biogenesis protein CcoG, involved in Cu oxidation Arsenate reductase (EC 1.20.4.1)	fiq1232216.5.peq.3308 fiq1232216.5.peq.3305, fiq1232216.5.peq.3362 fiq1232216.5.peq.3146, fiq1232216.5.peq.2118

G3b – Improvement of the assembly, contig reordering gap identification

1. On the links above the organism overview, choose comparative tools, then **sequence-based comparison**. Select a reference genome that is most closely related to your organism of interest and is preferably finished. Select your organism and several other closely related organisms as comparison genomes and click "compute".



2. When the new page appears, change the display to 300 items per page and click first. This **table displays the orthologous genes** in the same order as in in the reference genome. Examine the column headings and discuss the significance of the information in the column. What conclusions can be drawn, and what new hypothesis could be developed from this information?



numbers in the fasta file. However, one can **"mouse over" a cell to obtain information** on the exact name of the contig. For example, the pop-up box in the left panel below for Protein Encoding Gene (peg) 1113 from our genome in the 2nd set of columns shows that it is present on the scaffold "15+23+18+7+12+43" (RAST contig 3). The protein is about the same size as in the reference genome (632 aa's). Note that the proteins (Chaperone HtpG) are 93.17% identical. Note also that the gene begins at 499,442, very close to the end of the contig (499,962).

The pop-up box in the right panel below for peg 598 from our genome in the 2nd set of columns shows that it is present on the scaffold "1+14" (RAST contig 1). The protein is about the same size as in the reference genome (334 aa's). Note that the proteins (RecA) are 97.59% identical. Note also that the gene ends at 635,830, very close to the end of the contig (636,548).

The orthologs (bidirectional best hits) of these genes are # 2166 and 2169 in the reference genome, and 3935 & 3937 in the 3rd organism; 2907 and 2904 in the 4th organism; 844 and 845 in the 5th organism; and 2774 and 2773 in the last organism. What hypotheses would be reasonable about organism #2?

10451	Server - Job	Seed Vie	wer	×s	eed V	iewer - Bl	last	Seed	Viewer - B	Blast	Seed	Viewer - B	last	÷			RAST S	erver - J	ob 🤇	Seed Viewer	×	Seed \	/iewer - Bla	ist Se	ed Viewer - E	last	See	d Viewer -	Blast	+		
•	rast.nmpdr.or	g/seedview	er.cgi						⊤ C ⁱ	8 - Ips	n		٩	☆	Ê.	∔ îr	= €)@r	est. nm p	dr.org/se	edviewer.cgi					∠ G _i	8 - Ip	sn		٩	☆	Ê.	F 1
21	2140 58	7 bi	3	1055	hi	43	3018	bi	5	2028	hi	21	822	bi	11	2700	21	2140	587	bi 3	105	5 bi	43	3918	pi 5	2928	bi	21	822	bi	11	2799
21	2141 31	1 bi	3	1056	hi	43	3010	bi	5	2027	hi	21	823	bi	11	2708	21	2141	314	bi 3	105	<u>6</u> bi	43	<u>3919</u>	pi 5	<u>2927</u>	bi	21	823	bi	11	2798
21	2142 77	5 bi	3	1057	hi	43	3920	bi	5	2026	hi	21	824	bi	11	2797	21	2142	776	bi 3	105	Z bi	43	<u>3920</u>	oi 5	2926	bi	21	824	bi	11	2797
21	2142 77	76 bi	3	1057	hi	43	3920	bi	5	2920	hi	21	825	bi	11	2796	21	2143	1676	bi 3	105	<u>B</u> bi	43	<u>3921</u>	oi 5	2925	bi	21	825	bi	11	2796
21	2143 10	bi	3	1050	hi	43	3921	bi	5	2923	bi	21	923 927	bi	11	2790	21	2144	296	bi 3	1059	9 bi	43	<u>3922</u>	pi 5	2923	bi	21	827	bi	11	2795
21	2145 76) Di	3	1059	bi	43	2022	bi	5	2022	bi	21	020	bi	11	2722	21	2145	760	bi 3	106	Q bi	43	<u>3923</u>	oi 5	2922	bi	21	828	bi	11	2794
21	2145 700		2	1000	DI bi	40	2923	5	5	2021	DI bi	21	020	5	11	2794	21	2146	98	bi 3	106	1 bi	43	3925	oi 5	2921	bi	21	829	bi	11	2793
21	2140 98	-	3	1001	Uni	43	120	bi	5	1212	UI	40	2790	Uni	7	2/93	21	2147	965	-		uni	2	129	pi 2	1213	un	ii 48	2789	uni	7	2036
21	2147 90		12	4154	uni	2	2756	Uni	2	2044	uni	90	2/02	uni	7	2030	21	2148	205	uni 12	4154	4 uni	25	2756	uni 8	3944	un	ii 34	2001	uni	7	2035
1	2140 20	- uni	12	542	uni	25	2/30	uni bi	0	424	uni bi	26	2001	uni	17	2035	21	2149	275	uni 1	543	uni	37	3435	oi 1	434	bi	36	2111	uni	17	3745
21	2149 27		1	242	uni Li	37	2026	DI bi	1	434	DI I	30	2111	unii Lii	1/	2742	21	2150	473	bi 3	106	2 bi	43	3926	oi 5	2920	bi	21	830	bi	11	2790
21	2150 47.	S DI	3	1002	DI	43	3920	DI	5	2920	DI	15	501	DI	11	2790	21	2151	211	-		-					bi	15	591	bi	11	2792
21	2151 21	-	-	2160	-	46	4145	-	1	400	UI	15	1000	DI bi	2	152	21	2152	475	uni 5	2169	9 uni	46	4145	uni 1	498	ur	ii 31	1832	bi	2	153
	2152 47	uni	-	1004	uni uni	10	1010	uni	1	424	dill	54	2054	UI .	47	153	21	2153	434	uni 5	188	4 uni	15	1818	uni 1	434	bi	54	2964	uni	17	3745
1	2153 434	+ uni	5	1884	uni	15	1818	uni	1	934	DI	54	2904	uni	1/	3/45	21	2154	44	-		-	-				1-	-		-	-	
1	2154 44	-		-	-		-	-	-	-	-			-			21	2155	159	-	1	-					-	+	+	-		1
1	2155 159			1201	-	40	4117	1	2	0.20	-	50	2107		17	2745	21	2156	357	bi 4	138	4 uni	46	4117	ini 2	920	bi	59	3197	uni	17	374
1	2156 35	bi	4	1384	uni	46	4117	uni	2	920	DI	59	3197	uni	17	3/45	21	2157	341	bi 3	110	5 bi	43	3927	ni 5	2917	hi	21	834	bi	11	278
1	2157 34	bi	3	1105	DI	43	3927	DI	5	2917	DI	21	834	DI	11	2787	21	2150	131	bi 3	110	6 bi	43	3028	ni 5	2016	hi	21	835	bi	11	2784
1	<u>2158</u> 13	L DI	3	1106	bı	43	3928	bi	5	2916	bi	21	835	bi	11	2786	21	2150	220	01 5	110		45	3920		2910	bi	21	035	bi	11	2700
1	2159 23) -		-	-			-			bi	21	<u>836</u>	bi	11	2783	21	2139	230			-				-		21	030	bi	11	2703
1	2160 38	-			-			-			-			bi	11	2785	21	2100	30				42	2020		2014	+	+	+	DI I		2703
1	2161 12	7 bi	3	1107	bi	43	3929	bi	5	2914	-			bi	11	2784	21	2101	127	DI 3	110		43	3929	5	2914	-		007	DI	11	2784
1	2162 32	5 bi	3	1108	bi	43	3931	bi	5	2913	bi	21	<u>837</u>	bi	11	2782	21	2162	325	DI 3	1100		43	3931	5	2913	DI	21	837	DI	11	2782
21	2163 49	bi	3	1109	bi	43	3932	bi	5	2912	bi	21	838	bi	11	2781	21	2163	49	DI 3	110	2 01	43	3932	5	2912	DI	21	838	DI	11	278
21	2164 99	bi	3	1110	bi	43	3933	bi	5	2911	bi	21	<u>839</u>	bi	11	<u>2780</u>	21	2164	99	bi 3	1110	Q DI	43	3933	5	2911	DI	21	839	bi	11	2780
21	2165 13	L bi	3	1111	bi	43	<u>3934</u>	bi	5	2909	bi	21	<u>840</u>	bi	11	2779	21	2165	131	bi 3	111	<u>1</u> bi	43	<u>3934</u>	DI 5	2909	bi	21	840	bi	11	2779
21	2166 633	2 bi	3	<u>1113</u>	bi	43	<u>3935</u>	bi	5	<u>2907</u>	bi	21	<u>844</u>	bi	11	<u>2774</u>	21	2166	632	bi 3	111	<u>3</u> bi	43	<u>3935</u>	pi 5	2907	bi	21	844	bi	11	2774
1	2167 963	2 <mark>uni</mark>	5	<u>1872</u>	in a		1000	44.7	12	11007	i	45	~~~ <u>`</u>	uni	33	4424	21	2167	962	uni 5	187	2 uni	3	<u>232</u> (uni 3	1837	un	ji 45	2662	uni	33	4424
1	2168 478	3 -			ngi	558151.4	4.peg.1	113						-			21	2168	478	-		-			·		-		_	-		
1	2169 334	ŧ bi	1	<u>598</u>	loca	tion: 15+	+23+18+	+7+1	2+43_len	gth:499	962 4	99442 49	7550	bi	11	2773	21	2169	334	bi 1	<u>598</u>	bi	43	<u>3937</u>	pi 5	<u>2904</u>	bi	21	845	bi	11	2773
1	2170 39	5 <mark>bi</mark>	1	<u>597</u>	iden	tity: 0.93	317							bi	11	2772	21	2170	395	bi 1	597	fial5	58151 4 n	og 598				21	846	bi	11	2772
1	2171 26	5 bi	1	<u>596</u>	func	tion: Cha	aperone	prot	ein HtpG					bi	11	2771	21	2171	266	bi 1	<u>596</u>	inglos		Landba ^C	00040-00400	ic coco	20	21	<u>847</u>	bi	11	2771
21	2172 579) -			-		i –	ŕ.	<u> </u>	1	-			-			21	2172	579	-		length	on: 1+14_ 0.341	iengun.o.	00040 00400	00000	130			-		
21	2173 33	5 bi	1	<u>594</u>	bi	43	3942	bi	5	2899	bi	21	<u>848</u>	bi	11	2769	21	2173	335	bi 1	594	identit	y: 0.9759					21	848	bi	11	2769
1	2174 329) bi	1	<u>593</u>	bi	43	3943	bi	5	2898	bi	21	<u>849</u>	bi	11	2768	21	2174	329	bi 1	<u>593</u>	function	on: RecA	protein				21	849	bi	11	2768
1	2175 293	B bi	1	<u>592</u>	-			-			-			bi	11	2767	21	2175	293	bi 1	592	-					-			bi	11	2767
1	2176 114) bi	1	<u>591</u>	uni	3	220	uni	2	923	uni	7	239	bi	11	2766	21	2176	114	bi 1	591	uni	3	220	uni 2	923	un	ii 7	239	bi	11	2766
21	2177 26	7 bi	1	590	bi	27	2940	-			-			bi	11	2765	21	2177	267	bi 1	590	bi	27	2940			-			bi	11	2765
1	2178 253	3 bi	1	589	bi	43	3945	bi	5	2897	bi	21	850	bi	11	2764	21	2178	253	bi 1	589	bi	43	3945	bi 5	2897	bi	21	850	bi	11	2764
1	2179 420) bi	1	588	bi	43	3946	bi	5	2896	bi	21	852	bi	11	2763	21	2179	420	bi 1	588	bi	43	3946	oi 5	2896	bi	21	852	bi	11	2763
1	2180 443	3 bi	1	587	bi	43	3947	bi	5	2895	bi	21	853	bi	11	2762	21	2180	443	bi 1	587	bi	43	3947	oi 5	2895	bi	21	853	bi	11	2762
1	2181 56	bi	1	586	bi	43	3948	-			bi	21	854	bi	11	2761	21	2181	56	bi 1	586	bi	43	3948			bi	21	854	bi	11	2761
1	2182 35	7 bi	1	585	bi	43	3949	bi	5	2893	-			bi	11	2760	21	2182	357	bi 1	585	bi	43	3949	oi 5	2893	-	-		bi	11	2760
21	2183 129) bi	1	584	bi	43	3950	bi	5	2892	bi	21	855	bi	11	2759	21	2183	129	bi 1	584	bi	43	3950	ni 5	2892	hi	21	855	bi	11	2759
1	2184 39	7 bi	1	578	bi	43	3952	bi	5	2890	bi	21	860	bi	11	2756	21	2184	397	bi 1	578	bi	43	3952	ni 5	2890	hi	21	860	bi	11	2756
21	2185 74	bi	5	2094	bi	43	3953	bi	5	2889	-			bi	1	15	21	2185	74	bi 5	200	4 bi	43	3953	ni 5	2880	-			hi	1	15
1	2186 44	-			-						-			-	-	-	21	2100	44	5	209	- 01	15	2222		2009	+	+	+	-		12
21	2187 50	hi	1	576	uni	43	3953	-	1	1	-			-			21	2180	50	bi 1	570	-	42	2052		-	÷	+	+	÷	-	+
	2188 20	bi bi	1	575	hi	43	3954	bi	5	2889	hi	21	861	hi	11	2752	21	2187	50	0 1	5/6	uni	43	2054		2000	-	21	061	-	11	2752
11		- 11/1		1000	100		2221	101	-	2000		~ ~	224		~~	2.22	121	2188	1530	0 1	3/5	101	45	3954	כן וכ	2088	101	21	001	DI	11	2/53





When finished editing, save the file in plain text format and re-upload to RAST. Be sure to take note of the GC composition, as this is a frequently reported piece of information, particularly in novel species papers. Reordering and manually combining contigs is an arduous, but ultimately rewarding process as the number of contigs decreases. There is much that can be learned from draft genomes, and indeed, most genome sequencing projects are not finished due to the high cost with relatively little benefit.

Day 4 – Thursday, June 5, 2014 – Session 4 - 9:00-12:00 – Comparative Genomics

While there are many tools available for analysis of genome sequences, few, if any have the capabilities and accessibility of RAST and the SEED Viewer. What do you use with your students?

Question 1. How does my organism's genome align with its relatives?

- 1. After rearranging and deleting some contigs yesterday, we re-uploaded the genome to RAST. Go to RAST as before, login, click view details, then browse annotated genome in SEED Viewer.
- 2. On the Comparative Tools tab, choose sequence-based comparison. Choose your genome as a reference, and two or three closely related genomes as comparison genomes and click compute.
- 3. After the genomes have been compared, click on BLASTDOTPLOT for a pair of genomes. Note in the example on the left side that the *C. angstadtii* contig at around 4Mb corresponds to the end of the *C. gleum* genome (the type species for the genus). That contig can be selected and moved down to the end. Clicking redraw yields the center dotplot. The new contig at 4.02 Mbp looks like it should be second to last, so it can be moved. It takes some time, but this tool can facilitate contig reordering relative to a reference... to yield the dotplot at right.



Question 2. Does my organism have any unique gene clusters?

- On the circular map comparing genomes on sequence-based comparison results page, note the red dot, which corresponds to the area of the genome shown in the table to the left. Click outside an area on the map where there is a gap indicating the presence of genes in reference that are not present in comparison genomes.
- 2. Increase the number of items to display to 100, and click to move the red dot so that the gap is noticeable on the table. Mouse over the genes in the reference to see what the unique genes are. The example shown below was not very informative because most of the genes encoded hypothetical proteins. What are hypothetical proteins? How might one develop hypotheses regarding the function of these proteins?
- 3. Click a link for one of the genes to see the context. Note that homologous genes are color coded and that the aqua colored gene from Chitinophaga pinensis is annotated as a Phage tail fiber protein. The reference organism apparently has a prophage!



Question 3. What genes are shared among all of the organisms? Which are unique to my organism?



2. Open the .tsy file from the sequence-based	
21 Open the lost he hom the bequence buccu	A A E C O E P G H I I E C L M N O P Q protect protect protec
comparison in Excel. Click Ctrl+F to	Correg Gene Legels Gene 16 Parcten He Corteg Gene Gene 16 2033.7 Anotion He Corteg Gene Gene 16 2033.7 Anotion He Corteg Gene Gene 16 2033.7 Anotion L Corteg Gene Gene 16 2033.7 Anotion Corteg Corteg Gene Gene 16 2033.7 Anotion L Corteg Gene Gene 16 2033.7 Anotion Corteg Corteg Gene Gene 16 2033.7 Anotion L Corteg Gene Gene 16 2033.7 Anotion Corteg Corteg Gene Gene 16 2033.7 Anotion Corteg Corteg Corteg Gene Gene 16 2033.7 Anotion Corteg Corte
activate the find (realized function and	4 1 5 369 (§) (\$300.4;eg) hypothetical partiesis b) 27 277 (§) (\$300.4;eg) hypothetical partiesis b) 21 207 (§) (\$300.4;eg) hypothetical partiesis h) 1 31 31 109 (§) (\$300.4;eg) h) h) 107 (\$400.4;eg) h)
activate the ind/replace function and	i i
replace all dashes with nothing	1 2 Configuration deep 2 Antoniones designed and points and a configuration of the configuration designed and points desi
replace all address with hotning.	1 13 29 (s)(30)(34)(46)(3) Instrumentation preference in the second presence in the second presecond presence in the second presence
Find and Replace	11 16 10 6 6 6 10 11 16 10 16 10
Find Replace	1 D Description D Description Description <thdescription< th=""> <thdescription< th=""></thdescription<></thdescription<>
Find what:	1 13 2016[03033.406]3 hypothetics/protein 16 27.101 [02033.166]3 16.01 hypothetics/protein 16 2.11 [02033.166]3 15.11 hypothetics/protein 1 14 2016[03033.406]3 hypothetics/protein 16 27.101 [02033.166]3 15.11 hypothetics/protein 1 14 2016[03033.406]3 hypothetics/protein 16.11 hypothetics/protein 16.11 hypothetics/protein 1 15 2016[03033.406]3 16.01 hypothetics/protein 16.11 hypothetics/protein 17.11 hypothetics/protein 1 16 2016[03033.406]3 16.11 hypothetics/protein 16.11 hypothetics/protein 16.11 hypothetics/protein
Replace with:	1 30 110
Cotions >> 10	1 1 20 46 (5)(3)(3)(4),40(2) 1
	11 12 200 (g)(033.4 apg.2) Psynthetic prefers H 27.331 (g)(233.1 apg.2) 51.40 (p)(233.4 apg.2) 51.40 (p)(23.4 a
Replace All Replace Find All Find Next Close	b 1 D
	1 0 37 (rg)(3011.4 (rg)/2) 0 contractive data protocol protocol 27 70 (rg)(3021.4 (rg)/2) 50 (rg)/2 (
	(d) 1 42 ShellStates4 headeted order (d) ShellStates4 headeted order
	🟥 😧 👔 militar 👔 Salarystafens. 👔 Salarystafens. 👔 Salarystafens. 👔 Salarystafens. 👔 Salarystafens. 👔 Salarystafens.
2. Click Ctrl+A to select all, then on the Data	
menu, choose sort, and set a sort level	Sort
for each comparison organism using the	🛞 Add Level 🗶 Delete Level 🗈 Conv Level 🔿 💌 Ontions 🕅 My data bas bear
"Hit" column to sort. Click OK. Zoom out	Column Sort On Order
"Hit" column to sort. Click OK. Zoom out to see the full width, and scroll down to	Column Sort On Order Sort by Hit v Values A to Z v Then by Hit v Values A to Z v
"Hit" column to sort. Click OK. Zoom out to see the full width, and scroll down to identify the last gene in which all	Column Sort On Order Sort by Hit v Values v A to Z v Then by Hit v Values A to Z v Then by Hit v Values A to Z v
"Hit" column to sort. Click OK. Zoom out to see the full width, and scroll down to identify the last gene in which all	Column Sort On Order Sort by Hit v Values v A to Z v Then by Hit v Values A to Z v Then by Hit v Values A to Z v
"Hit" column to sort. Click OK. Zoom out to see the full width, and scroll down to identify the last gene in which all comparison organisms have a "bi" -	Column Sort On Order Sort by Hit Values A to Z Then by Hit Values A to Z
"Hit" column to sort. Click OK. Zoom out to see the full width, and scroll down to identify the last gene in which all comparison organisms have a "bi" - directional best hit. This is the core	Column Sort On Order Sort by Hit v Values v A to Z v Then by Hit v Values v A to Z v Then by Hit v Values v A to Z v Then by Hit v Values v A to Z v Then by Hit v Values v A to Z v Then by Hit v Values v A to Z v Then by Hit v Values v A to Z v Then by Hit v Values v A to Z v Then by Hit v Values v A to Z v Then by Hit v Values v A to Z v
"Hit" column to sort. Click OK. Zoom out to see the full width, and scroll down to identify the last gene in which all comparison organisms have a " bi "- directional best hit. This is the core genome. Select these rows, determine	Column Sort On Order Sort by Hit v Values A to Z v Then by Hit v Values A to Z v
"Hit" column to sort. Click OK. Zoom out to see the full width, and scroll down to identify the last gene in which all comparison organisms have a "bi" - directional best hit. This is the core genome. Select these rows, determine their number and copy and paste to a	Column Sort On Order Sort by Ht v Values v A to Z v Then by Ht v Values V A to Z v Then by Ht v Values V A to Z v
"Hit" column to sort. Click OK. Zoom out to see the full width, and scroll down to identify the last gene in which all comparison organisms have a " bi "- directional best hit. This is the core genome. Select these rows, determine their number and copy and paste to a	Column Sort On Order Sort by Ht values v A to Z v Then by Ht values A to Z v



X 🖌	17 - (*	- (-		and the second second		Me	sorhizobiun	n loti MAF	F303099.tsv	- Microso	ft Excel									×
File	Hom	ie Inse	rt Page Layout	Formulas Data Revie	w View	Acroba	ıt												ہ 🕜 ۵	- 67
From Access	From Web	From From Text So	n Other arces 7	Refresh All + @ Edit Links	Ž↓ <u>Z Z</u> Z↓ Sort	Filter	Clear Reapply Advanced	Text to Column	Remove s Duplicates	Data Validation	Consolio	ate What- Analysis	If Grou	up Ungroup	Subtotal	9클 Show E 크릴 Hide D	Detail etail			
		Get Externa	Data	Connections	50	on or Filter				Data Tor	DIS				Jutime		la			
	E6376		Jx AB	C transporter binding prot	ein															
	A	В	C D	E	F	G	H	1.1	J	K	L	м	N	0	Р	Q	R	S	Т	U
6375	1	4849	73 tig 26683 5	5 methyltetrahydrofolatehomo	cysteine me				0						0					
6376	1	5786	110 fig 26683/	ABC transporter binding prote	in .				0						0					
6377	1	699	276 tig 26683 A	ABCtype multidrug transport s	ystem, ATPa				0						0					
6378	1	5494	324 Tig 26683 A	Acetyi xyian esterase (EC 3.1.1	.41)				0						0					
6379	1	3164	293 tig 26683 a	acio prosphatase					0						0					
6380	1	1305	345 Tigi 26683 A	ADPHIDOSYIGIYCONYOROlase (EC	3.2)				0						0					
6283	1	4245	210 fig/20085 d	ammomeuriyin ansierase Annorohis dimothul sulfovido	roductoco cl				0						0					
6383	1	3867	180 fig 20083 #	autotransporter protein	reductase ci				0						0					
6384	1	920	180 fig 20083 a	autotransporter protein					0						0					
6385	1	4477	97 fig126683 /	Ava CO101 and related protei	nc				0						0					
6386	1	269	325 fig126683 F	B burgdorferi predicted codir	g region BB(0						0					
6387	1	4066	341 fig 26683 F	Biotin carboxylase (EC 6.3.4.1	4)				0						0					
6388	1	2150	447 fig 26683 0	Carbamoviphosphate synthas	e large subu				0						0					
6389	1	4377	384 fig 26683 (Ceramide glucosyltransferase	(EC 2.4.1.80				0						0					
6390	3	7076	235 fig 26683 0	Chemotaxis signal transduction	on protein				0						0					
6391	1	2500	230 fig 26683 0	Chromosome segregation ATP	ases				0						0					
6392	1	1690	330 fig 26683 0	COG family: dihydrodipicolini	ate reductas				0						0					
6393	1	3650	377 fig 26683 c	contains similarity to Dalanir	e:Dlactate li				0						0					
6394	1	3108	197 fig 26683 c	contains weak similarity to Tr	bB (188) and				0						0					
6395	1	3632	109 fig 26683 c	contains weak similarity to xy	lose isomer				0						0					
6396	1	5577	346 fig 26683 [Diadenosine tetraphosphatas	e and relate				0						0					
6397	3	7123	85 fig 26683 [DNA ligase homolog					0						0					
6398	2	7007	177 fig 26683 [DNA methylase					0						0					
6399	2	7002	325 fig 26683 [DNA methylation					0						0					
6400	1	362	155 fig 26683 [DNA packaging protein gp3					0						0					
6401	1	1608	153 fig 26683 [DNAbinding protein					0						0					
6402	1	3599	931 fig 26683 [DNAdirected DNA polymerase					0						0					

J. References

- Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25, 3389-3402.
- Aziz, R.K., Bartels, D., Best, A.A., DeJongh, M., Disz, T., Edwards, R.A., Formsma, K., Gerdes, S., Glass, E.M. & other authors (2008). The RAST Server: rapid annotations using subsystems technology. *BMC Genomics* 9, 75.
- Brenner, D.J. (1973). Deoxyribonucleic acid reassociation in the taxonomy of enteric bacteria. *Int J Syst Bacteriol* 23, 298-307.
- Chan, J.Z.-M., Halachev, M.R., Loman, N.J., Constantinidou, C. & Pallen, M.J. (2012). Defining bacterial species in the genomic era: insights from the genus *Acinetobacter*. *BMC Microbiol* 12, 302.
- Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM, et al. (1995) Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science. 1995 Jul 28;269(5223):496-512.
- Goris, J., Konstantinidis, K.T., Klappenbach, J.A., Coenye, T., Vandamme, P. & Tiedje, J.M. (2007). DNA–DNA hybridization values and their relationship to whole-genome sequence similarities. *Int J Syst Evol Microbiol* 57, 81–91.
- Konstantinidis, K.T. & Tiedje, J.M. (2005). Genomic insights that advance the species definition for prokaryotes. *Proc Natl Acad Sci USA* **102**, 2567-2572.
- Kurtz, S., Phillippy, A., Delcher, A.L., Smoot, M., Shumway, M., Antonescu, C. & Salzberg, S.L. (2004). Versatile and open software for comparing large genomes. *Genome Biol* 5, r12.
- Lane, D. J. (1991). 16S/23S rRNA sequencing. Nucleic acid techniques in bacterial systematics. E. Stackebrandt and M. Goodfellow, eds. New York, NY, John Wiley and Sons: 115-175.
- Markowitz, VM; Chen I-MA, Palaniappan K, Chu K, Szeto E, Grechkin Y, Ratner A, Jacob B, Huang J, Williams P, Huntemann M, Anderson I, Mavromatis K, Ivanova NN, Kyrpides NC (2012). <u>"IMG: the integrated microbial genomes database and comparative analysis system"</u>. Nucleic Acids Res. (England) 40 (1): D115-22. <u>DOI:10.1093/nar/gkr1044</u>. <u>PMC 3245086</u>. <u>PMID</u> 22194640.
- Overbeek, R., Begley, T., Butler, R.M., Choudhuri, J.V., Chuang, H.Y., Cohoon, M., de Crécy-Lagard, V., Diaz, N., Disz, T. & other authors (2005). The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res* 33, 5691-5702.
- Overbeek, R., Robert Olson, Gordon D Pusch, Gary J Olsen, James J Davis, Terry Disz, Robert A Edwards, Svetlana Gerdes, Bruce Parrello, Maulik Shukla, Veronika Vonstein, Alice R Wattam, Fangfang Xia, Rick Stevens. (2014) The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). Nucleic Acids Research.

- Raphael, B.H., Lautenschlager, M., Kalb, S.R., de Jong, L.I., Frace, M., Lúquez,, C., Barr, J.R., Fernández, R.A. & Maslanka, S.E. (2012). Analysis of a unique *Clostridium botulinum* strain from the Southern hemisphere producing a novel type E botulinum neurotoxin subtype. *BMC Microbiol* 12, 245.
- Richter, M. & Rosselló-Móra, R. (2009). Shifting the genomic gold standard for the prokaryotic species definition. *Proc Natl Acad Sci USA* **106**, 19126-19131.
- Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA and Barrell B (2000) Artemis: sequence visualization and annotation. *Bioinformatics* **16**;10;944-5
- Stackebrandt, E. & Ebers, J. (2006). *Taxonomic parameters revisited: tarnished gold standards. Microbiol Today* 33, 152–155.
- Stackebrandt, E. & Goebel, B.M. (1994). *Taxonomic note: a place for DNA-DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology. Int J Syst Evol Microbiol* 44, 846-849.
- Stackebrandt, E., Frederiksen, W., Garrity, G.M., Grimont, P.A., Kämpfer, P., Maiden, M.C., Nesme, X., Rosselló-Mora, R., Swings, J. & other authors (2002). *Int J Syst Evol Microbiol* 52, 1043–1047.
- Thompson CC, Emmel VE, Fonseca EL et al. (2013) Streptococcal taxonomy based on genome sequence analyses. [v1; ref status: indexed, http://f1000r.es/o1] F1000Research 2013, 2:67 (doi: 10.12688/f1000research.2-67.v1)
- Tindall, B.J., Roselló-Móra, R., Busse, H.-J., Ludwig, W. & Kämpfer, P. (2010). Notes on the characterization of prokaryote strains for taxonomic purposes. *Int J Syst Evol Microbiol* 60, 249-266.
- Wu, D., Hugenholtz, P., Mavromatis, K., Pukall, R., Dalin, E., Ivanova, N.N., Kunin, V., Goodwin, L., Wu, M. & other authors (2009). A phylogeny-driven genomic encyclopedia of Bacteria and Archaea. *Nature* 462, 1056-1060.
- Zhang, Y.M., Tian, C.F., Sui, X.H., Chen, W.F. & Chen, W.X. (2012). Robust Markers Reflecting Phylogeny and Taxonomy of Rhizobia. *PLoS One* 7, e44936