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.

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

	Stage	Description
1.	Random small insert and large insert library construction	Shear genomic DNA randomly to --2 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).

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 **evolution**ary 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 sub**systems** 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 scores are logarithmically linked to error 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 (http://rast.nmpdr.org/) (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.
    - Download sequence data from Sequencing Center Server, unzip files
    - 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 10^9 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.)

3. 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.**

8. 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 µl 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^{o}C separates the two strands. Upon cooling to 55^{o}C, the primers will hybridize (base pair) with their complementary sequences on the template DNA. Heating to 72^{o}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 10^9 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 dH_2O

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**.

Program = rRNA.fl

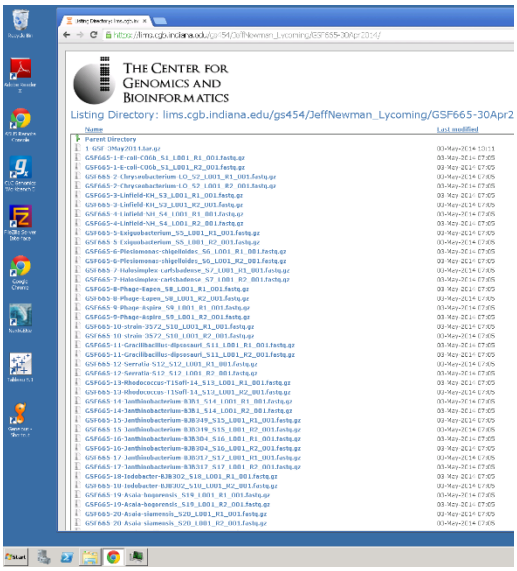
Phase 1 (initial denaturation) - 1 cycle	Initial denaturation	2 **min**. @ 94^{o}C
Phase 2 (standard cycle) 35 cycles	standard denaturation	30 sec. @ 94^{o}C
	Primer annealing	30 sec. @ 50^{o}C
	Primer extension[2]	1.5 min @ 72^{o}C
Phase 3 (extra extension) - 1 cycle	Primer extension	**9 min.** @ 72^{o}C

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: Biol0gyDept*

2. 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 https://lims.cgb.indiana.edu/gs454/JeffNewman_Lycoming/ and login with username and password provided to you.

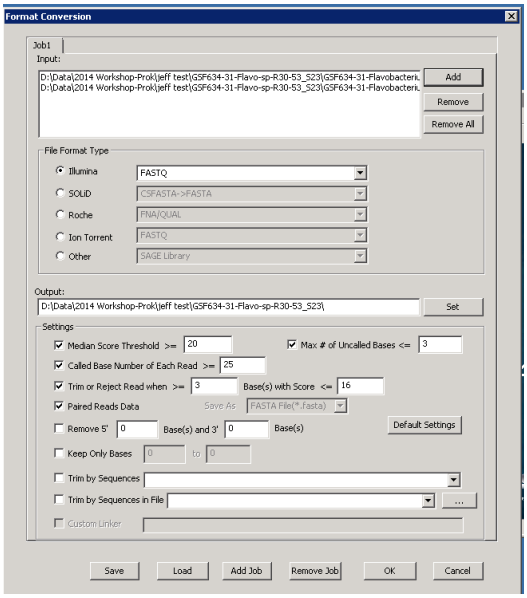
4. Right click on the desired file, choose "save file as" and specify an appropriate download location (your folder on the data drive).

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.

7. 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.

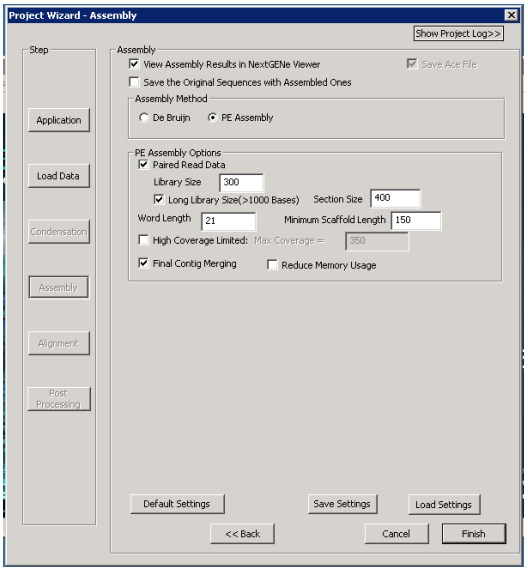


9. After conversion has been completed, use the file manager and review the conversion log text files to note the percentage of reads converted. After beginning the assembly process, return to these documents to **discuss the meaning and significance of each line in the conversion log**

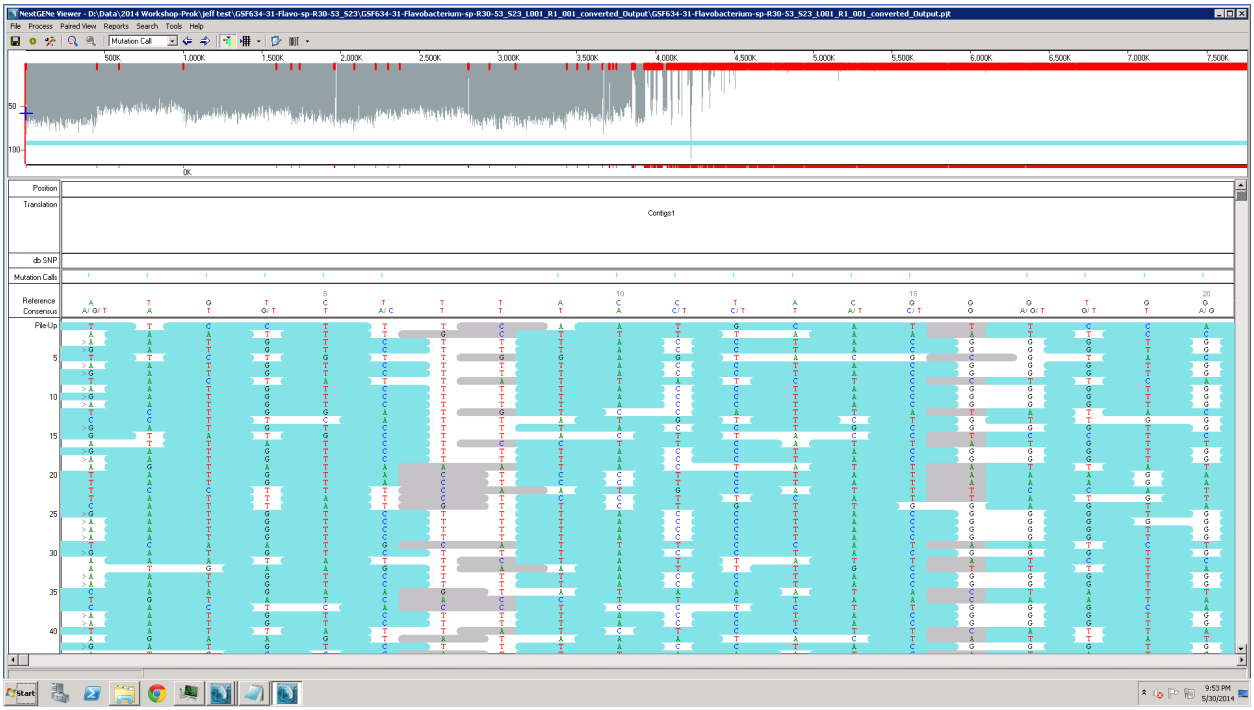
10. On the subsequent page, click load, select the two successfully converted *.fasta files and click Next.

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.



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

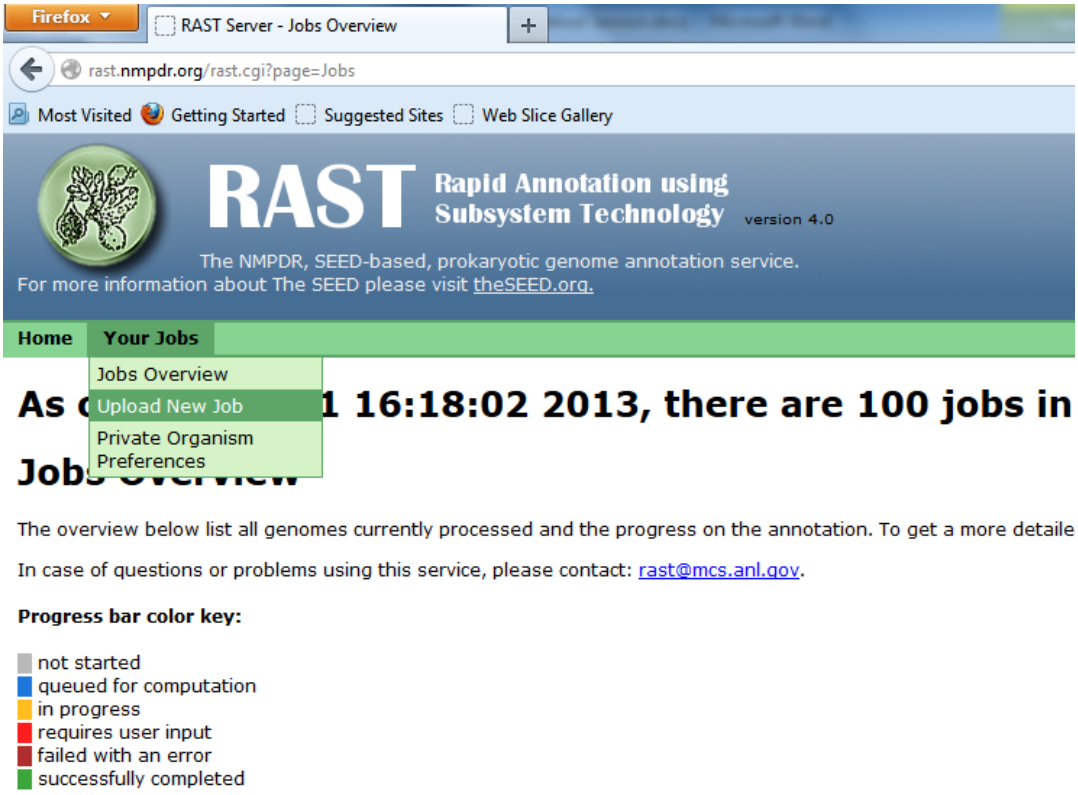
What does each statistic mean, what is the significance of each?

G2a2. Upload sequences to RAST for initial annotation.

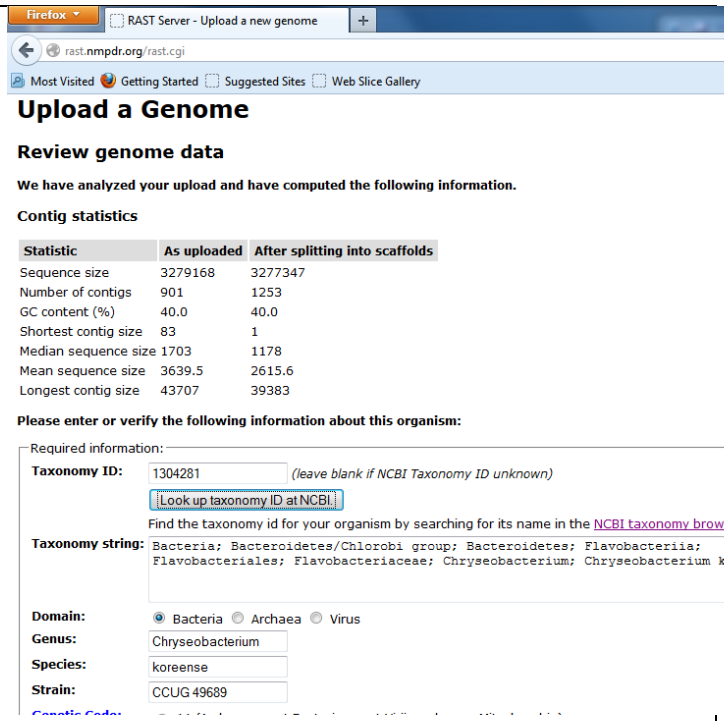
At this stage, there may still be more than one hundred contigs. If the genome is from a novel species, or a species for which there is no reference sequence, one can still use the most closely related genome sequence available (preferably within the same genus) to help determine the proper order and 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

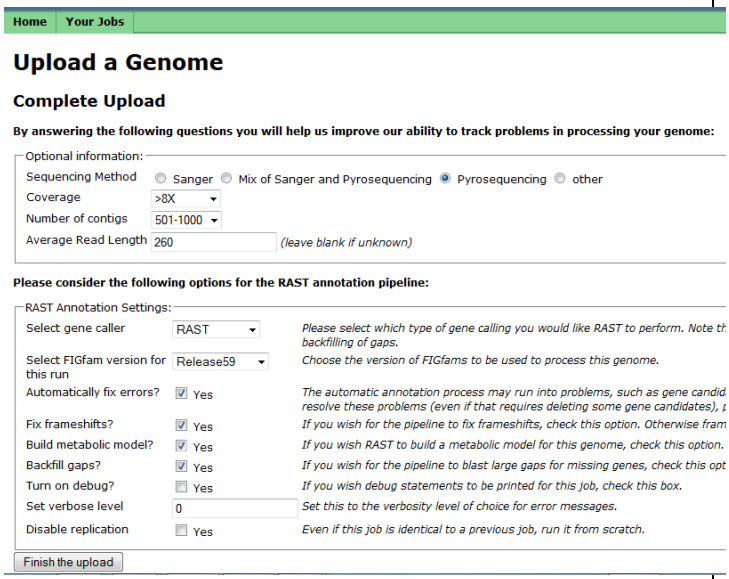
1. Use **Firefox** to login to the RAST website (http://rast.nmpdr.org/). Click "Your Jobs" → "Upload New Job". On the subsequent page, Browse to the Assembly2.fasta file from the previous section, then click "Use the data and go to step 2



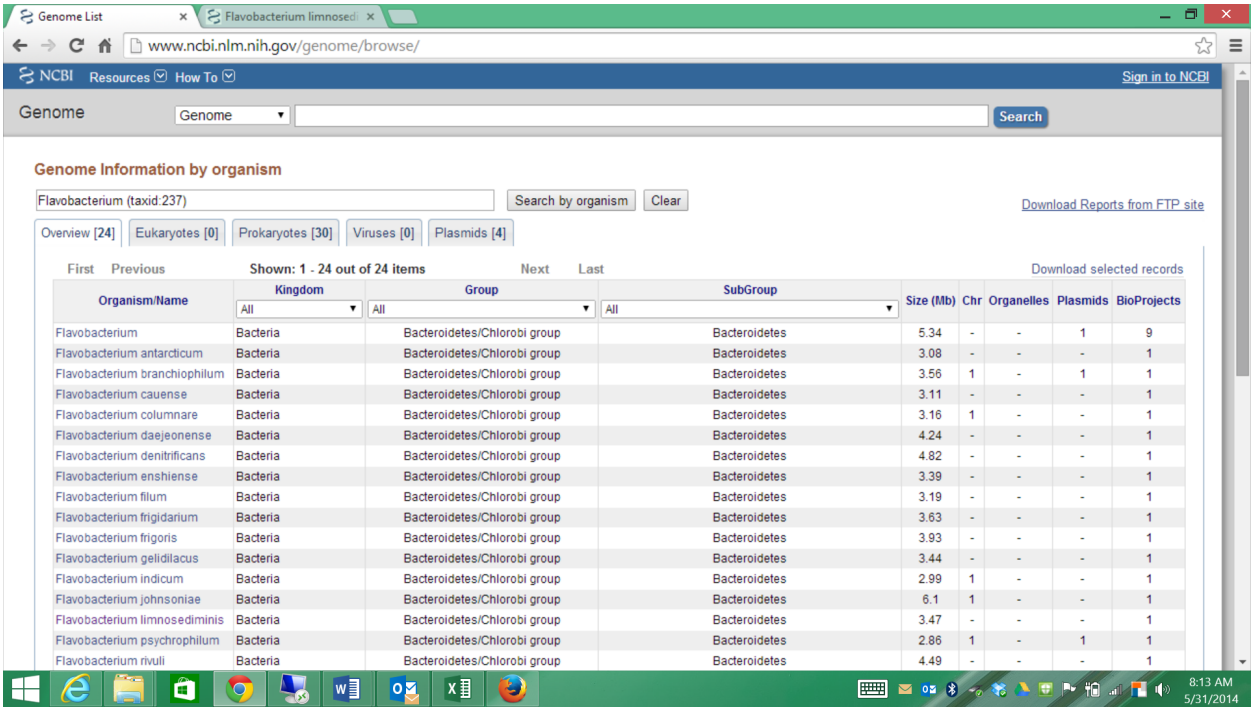
2. Open a new tab in your browser, go to the NCBI website (http://www.ncbi.nlm.nih.gov/) and perform a Taxonomy search for the organism. Copy the NCBI taxonomy ID into the appropriate box on the RAST page and click the lookup button. Click "Use this data and go to step 3"



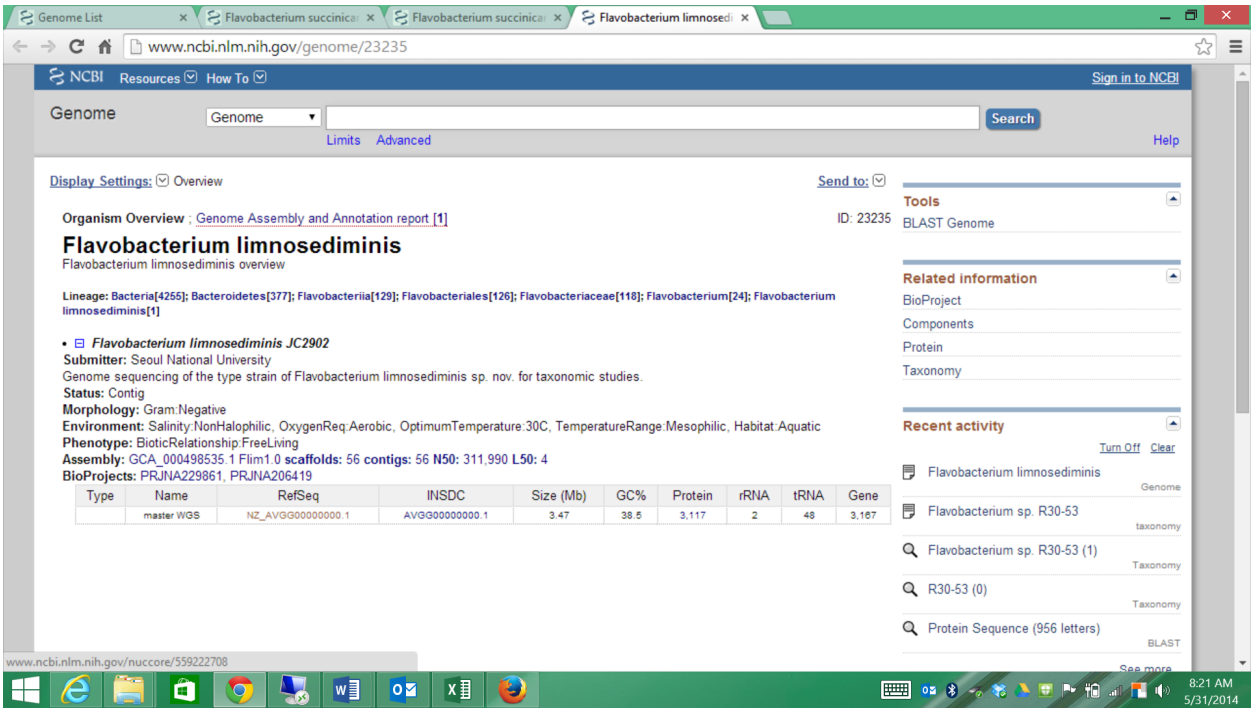
3. Enter the requested information, change the FIGfam version to the highest number, check "build metabolic model", then click "Finish the upload"



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



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.



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.

3. 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.

4. 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 198µl of working buffer to the appropriate tube, then add 2µl of DNA.
5. For each of the two standards, add 190µl of working buffer to the appropriate tube, then add 10µl 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

1. 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 μL 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):

   Lane 1 – 3 μL uncut λ DNA (20 ng/μL)

   Lane 2 – 3 μL uncut λ DNA (50 ng/μL)

   Lane 3 – 3 μL uncut λ DNA (80 ng/μL)

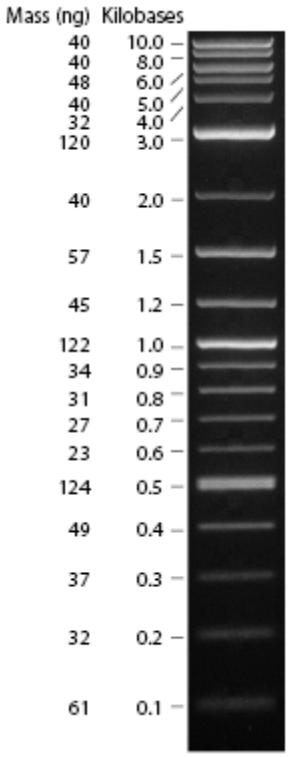
   Lane 4 – 3 μL gDNA Sample 1st elution

   Lane 5 – 3 μL gDNA Sample 2nd elution

   Lane 6 - 5 μL 2 log ladder (500 ng total)

   Lane 7 – 5 μL PCR negative control

   Lane 8 – 5 μL PCR



4. Run gel at 50v during lunch. After fastest migrating blue dye (bromophenol blue) has migrated 2/3 the length of the gel, turn off power, carefully remove gel from chamber, drain, slide onto piece of plastic wrap. Photograph the gel under UV light, print on a color printer

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)

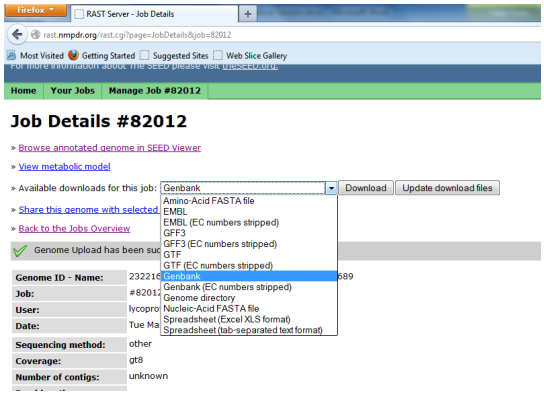
Method	1st elution	2nd elution
Gel electrophoresis		

Qubit		

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

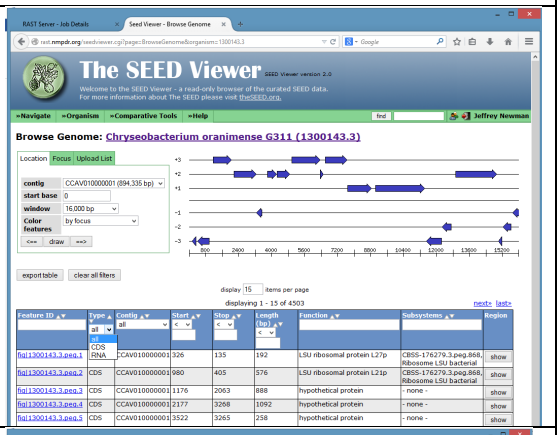
G3a - Review annotation results

1. Use **Firefox** to login to the RAST website (http://rast.nmpdr.org/). 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"



2. In the Organism Tab near the top of the page, choose Genome Browser, then in the second column choose RNA.

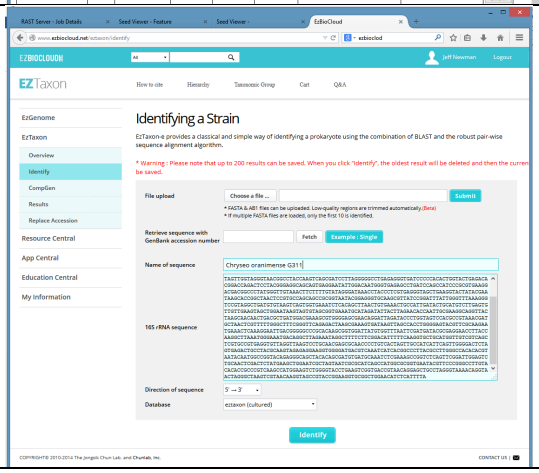
3. Click the next link until you find the small subunit rRNA gene, then click the feature ID link, then the sequence link. Select and copy the sequence.



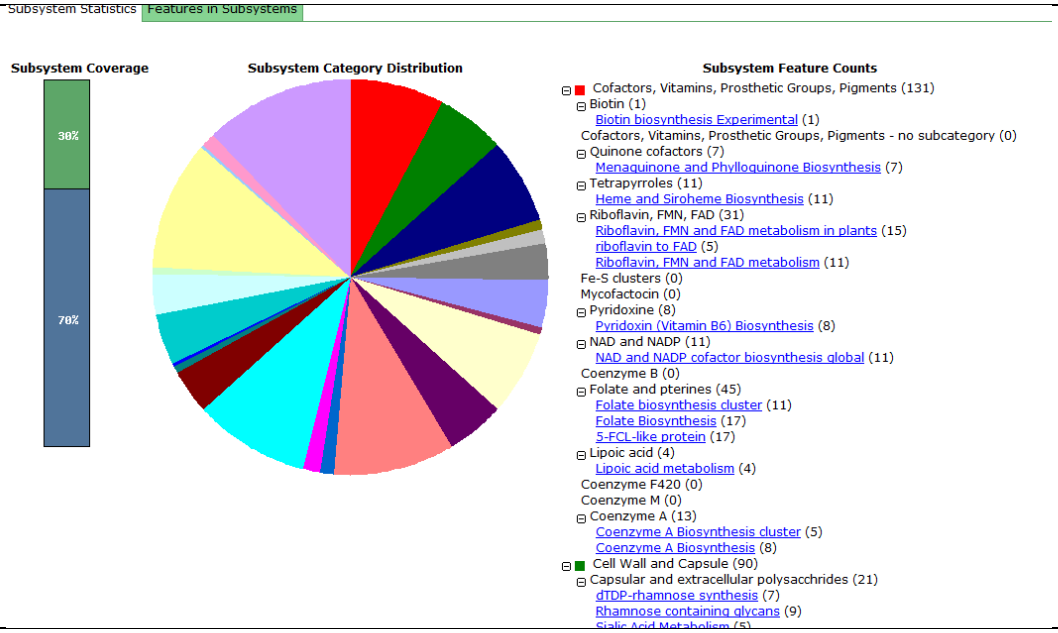
4. Login to EzTaxon at http://www.ezbiocloud.net/eztaxon . This is the best website to search for matching 16srRNA sequences because it includes only Type strains.

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.

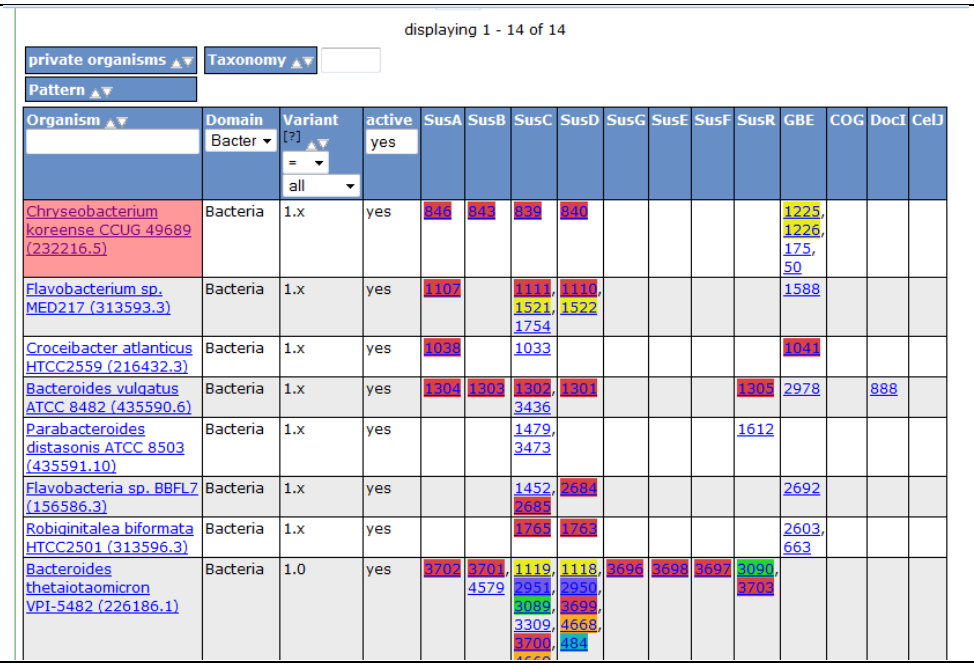
Is the best match what you expected?



6. On the first Seed Viewer page, click the back arrow to returnto the organism overview, then click the plus signs to expand the list of subsystem feature counts. After all of the subsystems have been expanded, copy and paste the list to a Microsoft Excel document. Save the document with a name such as “Chryseobacterium sp. Subsystems”



7. Click on a subsystem of interest, then click on the subsystem spreadsheet tab to identify the specific subsystem genes present in the organism of interest.



8. Click the back button to return to the Organism overview page, then in the subsystems information area, select the features in subsystems tab and browse the different categories to become familiar with the available information. What predictions would you make about the organism's phenotypes?
How does this compare to published descriptions of the organism?

Subsystem Information

Subsystem Statistics | Features in Subsystems

export to file | clear all filters

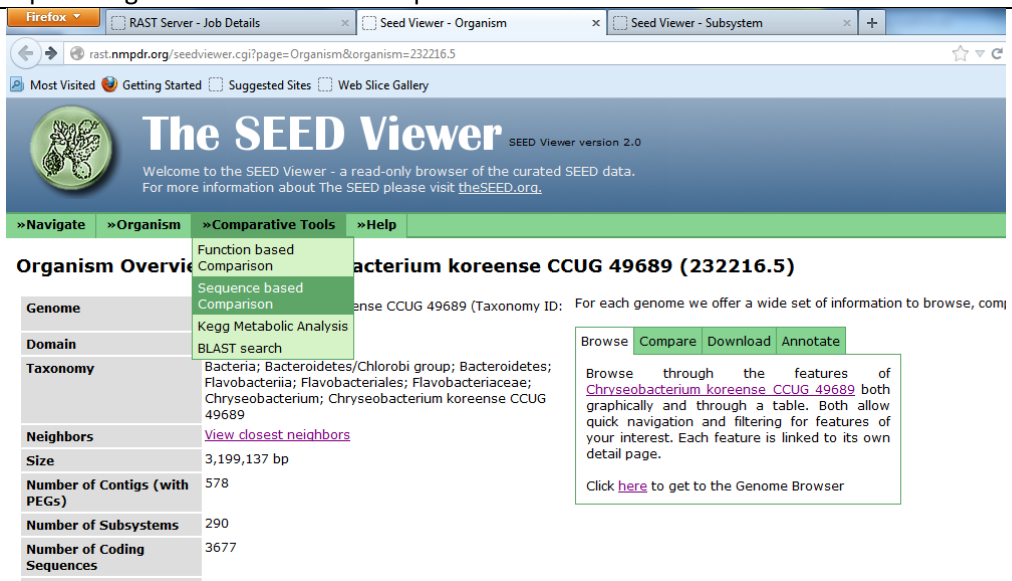
display 15 items per page

displaying 1 - 15 of 52 next» last»

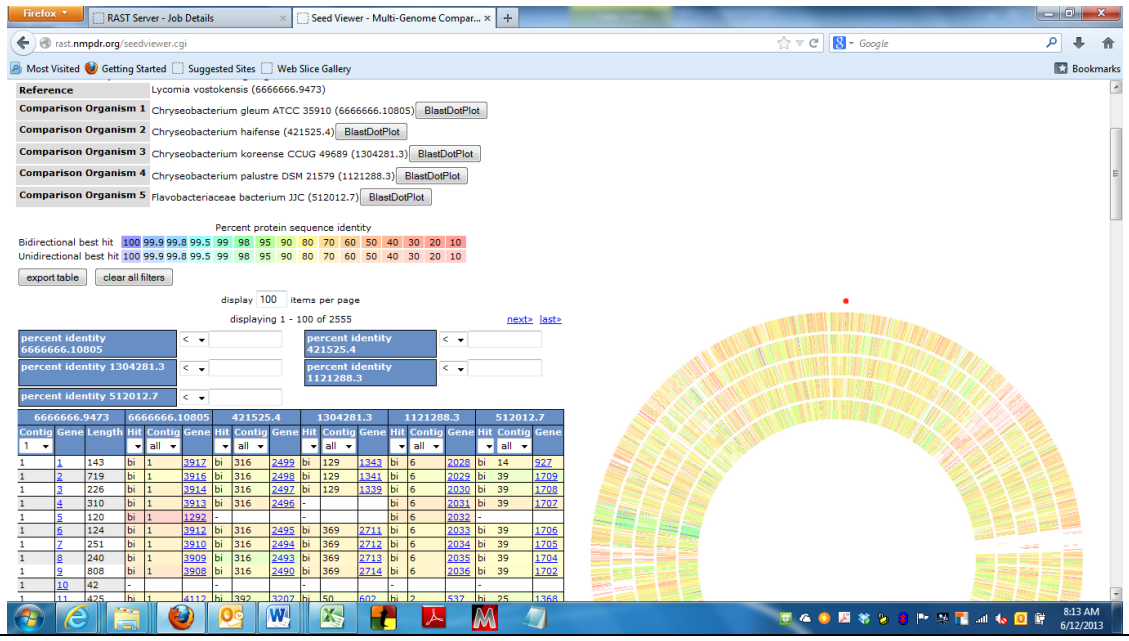
Category	Subcategory	Subsystem	Role	Features
Respiration	all			
Respiration	ATP synthases	F0F1-type ATP synthase	ATP synthase C chain (EC 3.6.3.14)	fig|232216.5.peg.3583
Respiration	ATP synthases	F0F1-type ATP synthase	ATP synthase delta chain (EC 3.6.3.14)	fig|232216.5.peg.2254
Respiration	ATP synthases	F0F1-type ATP synthase	ATP synthase B chain (EC 3.6.3.14)	fig|232216.5.peg.2255, 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)	fig|232216.5.peg.2253
Respiration	ATP synthases	F0F1-type ATP synthase	ATP synthase A chain (EC 3.6.3.14)	fig|232216.5.peg.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)	fig|232216.5.peg.3306
Respiration	Electron accepting reactions	Terminal cytochrome C oxidases	Cytochrome c oxidase subunit CcoO (EC 1.9.3.1)	fig|232216.5.peg.1159, fig|232216.5.peg.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	fig|232216.5.peg.3305, fig|232216.5.peg.3362
Respiration	Electron accepting reactions	Anaerobic respiratory reductases	Arsenate reductase (EC 1.20.4.1)	fig|232216.5.peg.1146, fig|232216.5.peg.2118
Respiration	Electron donating reactions	Respiratory Complex I	NADH-ubiquinone oxidoreductase chain K (EC 1.6.5.3)	fig|232216.5.peg.3011

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?

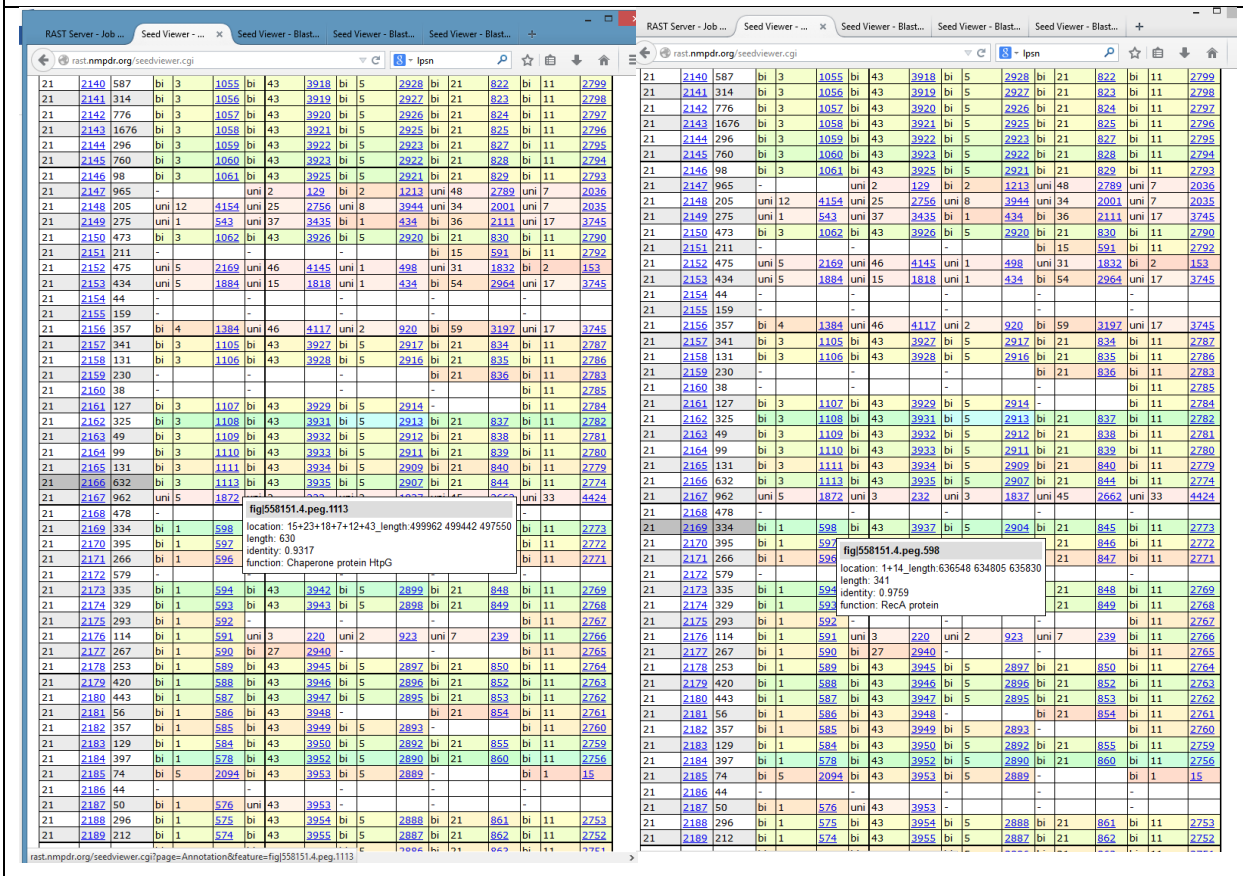


3. RAST renumbers the contigs, so the contig numbers in the table will not correspond to the contig

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?

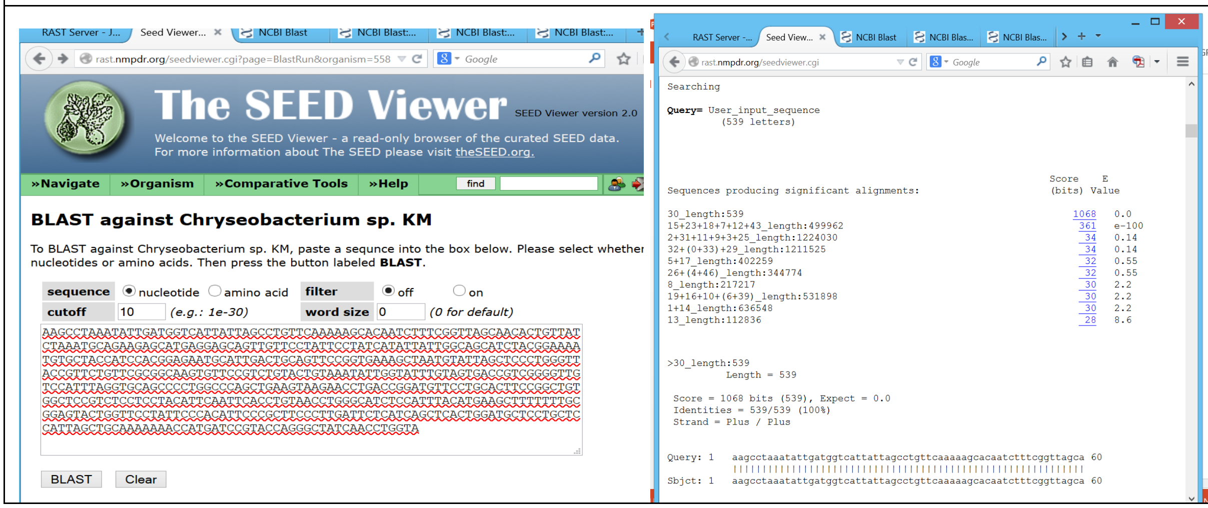


4. Because the 3' end if the "1+14" contig appears to be located after the 3 ' end of the "15+23+18+7+12+43" contig, the "1+14" contig must be flipped using a reverse complement tool such as at http://reverse-complement.com/ . In some cases, as shown below, two contigs can be assembled manually by using the find function in Word to identify overlaps. In other cases, there are no detectable overlaps, but reordering the contigs still has value to recognize synteny in dot plots.
What can cause problems with assembly?
How can these be identified?
How can gaps in the sequence be closed?

Spend a little time now to connect some contigs, and identify some gaps that can be closed as indicated above. After each edit, **save the file as plain text** with an incrementally increasing number so that it is possible to backtrack if an error is made.

When working with your own data, use Word or another program to **move the appropriate contigs into the correct order and orientation**. Keep in mind that if your reference genome is also composed of multiple contigs, it may be possible to reorder them and improve that assembly as well.

Overlapping sequences assembled into one scaffold

5. It is common for low quality, or contaminating reads to not assemble properly and to produce short contigs. These can be identified by performing BLAST searches against the assembled genome to determine if a sequence is already represented in a larger contig, or performing a BLAST search against GenBank to determine if the sequence is derived from another organism. Contigs 48 and 49 below can obviously be deleted. Other contigs like #30 can be BLASTed against the genome. On any SEED Viewer page, click the comparative tools tab, then BLAST search. 6. Copy the sequence of any contigs in question into the box, click the nucleotide radio button, then click BLAST. If there is a larger sequence with significant hits, the short contig can probably be safely deleted.
>30_length:539 AAGCCTAAATATTGATGGTCATTATTAGCCTGTTCAAAAAGCACAATCTTTCGGTTAGCAACACTGTTATCTAAATGCAGAAGAGC ATGAGGAGCAGTTGTTCCTATTCCTATCATATTATTGGCAGCATCTACGGAAAATGTGCTACCATCCACGGAGAATGCATTGACTG CAGTTCCGGTGAAAGCTAATGTATTAGCTCCCTGGGTTACCGTTCTGTTCGCGGCAAGTGTTCCGTCTGTACTGTAAATATTGGTA TTTGTAGTGACCGTCGGGGTTGTCCATTTAGGTGCAGCCCCTGGCCCAGCTGAAGTAAGAACCTGACCGGATGTTCCTGCACTTCC GGCTGTGGCTCCGTCTCCTCCTACATTCAATTCACCTGTAACCTGGGCATCTCCATTTACATGAAGCTTTTTTTGCGGAGTACTGG TTCCTATTCCCACATTCCCGCTTCCCTTGATTCTCATCAGCTCACTGGATGCTCCTGCTCCATTAGCTGCAAAAAAACCATGATCC GTACCAGGGCTATCAACCTGGTA >48_length:173 AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAC A >49_length:172 ACCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC


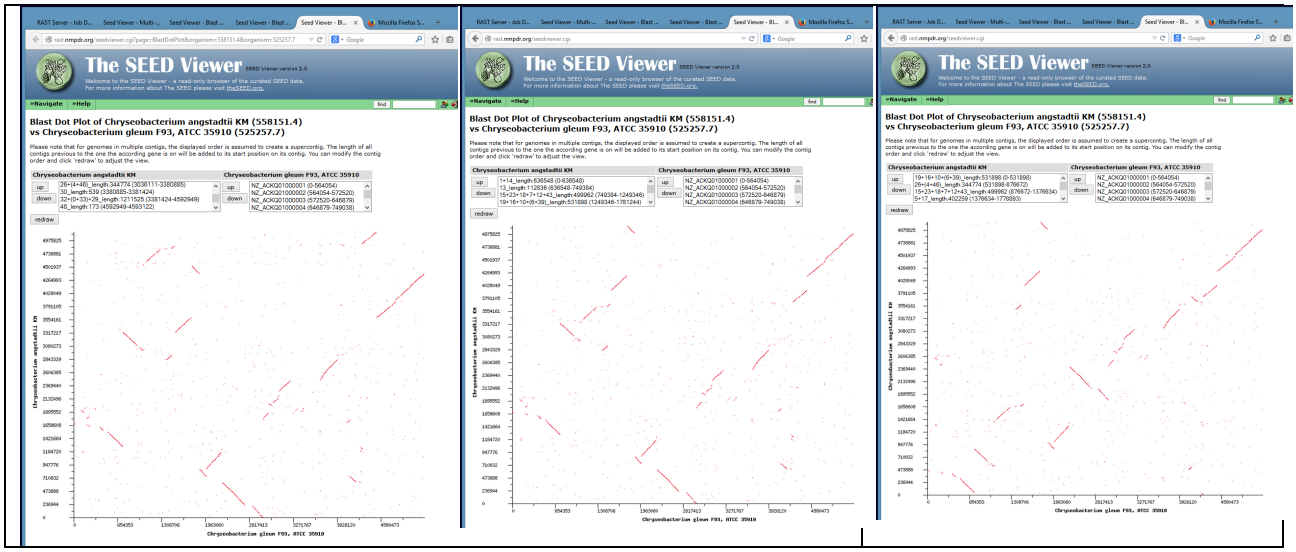
When finished editing, save the file in plain text format and re-upload to RAST. Be sure to take note of the GC composition, as this is a frequently reported piece of information, particularly in novel species papers. Reordering and manually combining contigs is an arduous, but ultimately rewarding process 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.

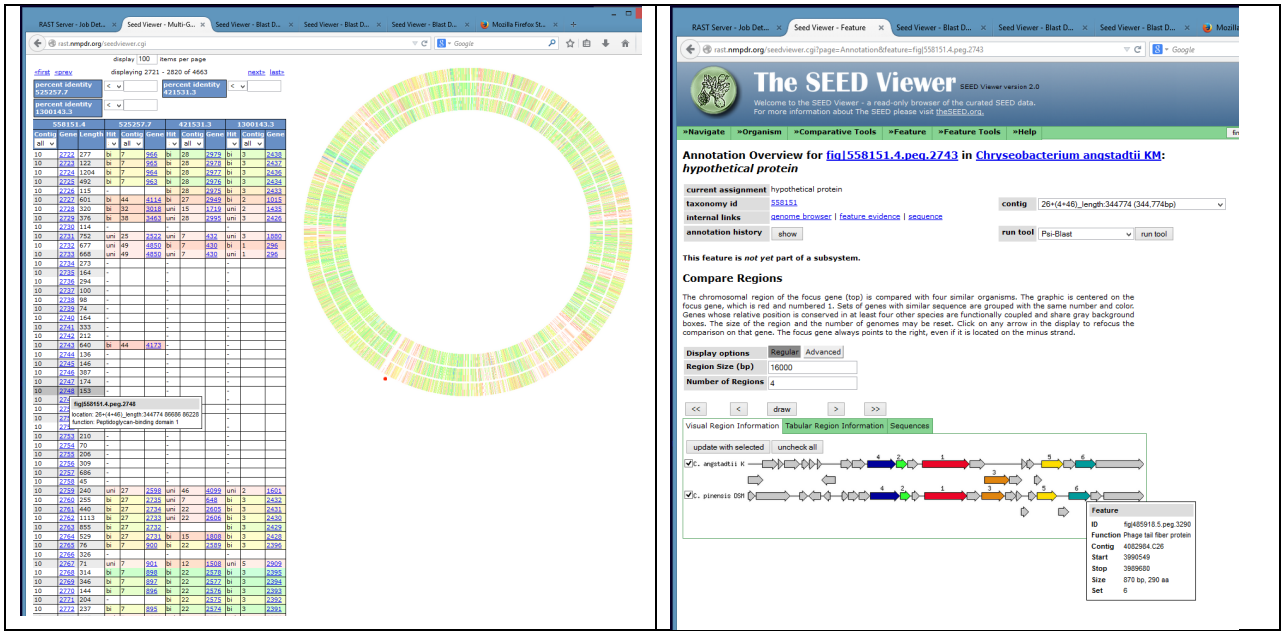


Question 2. Does my organism have any unique gene clusters?

1. 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.

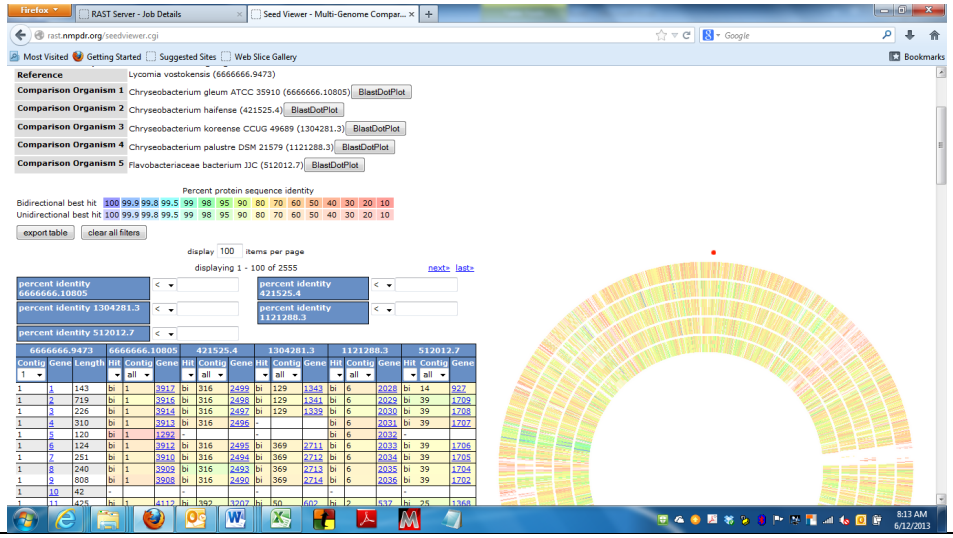
2. Increase the number of items to display to 100, and click to move the red dot so that the gap is noticeable on the table. Mouse over the genes in the reference to see what the unique genes are. The example shown below was not very informative because most of the genes encoded hypothetical proteins. 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!

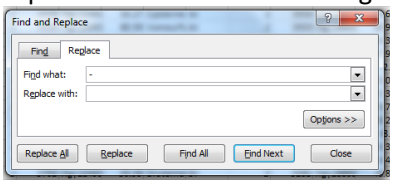


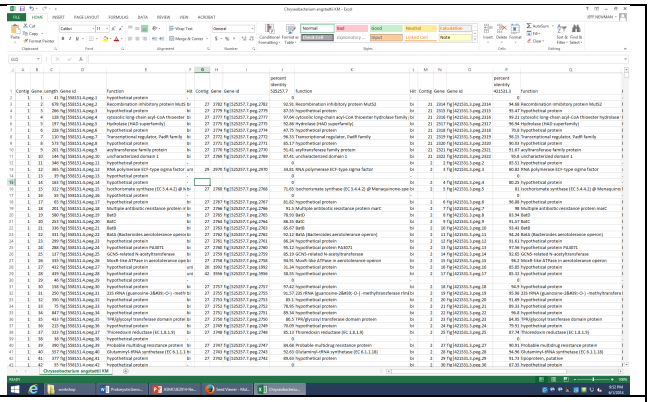
Question 3. What genes are shared among all of the organisms? Which are unique to my organism?

1. On the sequence-based comparison page, copy and paste the list of organisms to an excel spreadsheet. Copy the name of the reference stain to the clipboard, then click the “export table button”. Save the .tsv file on your flash drive with the name of the reference strain followed.

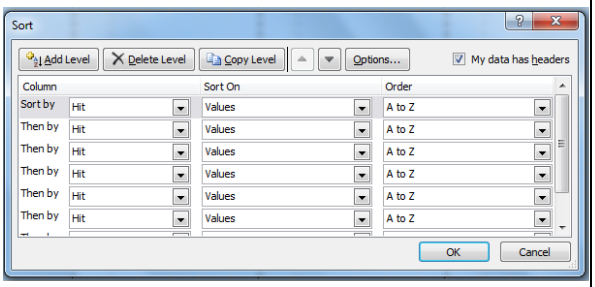


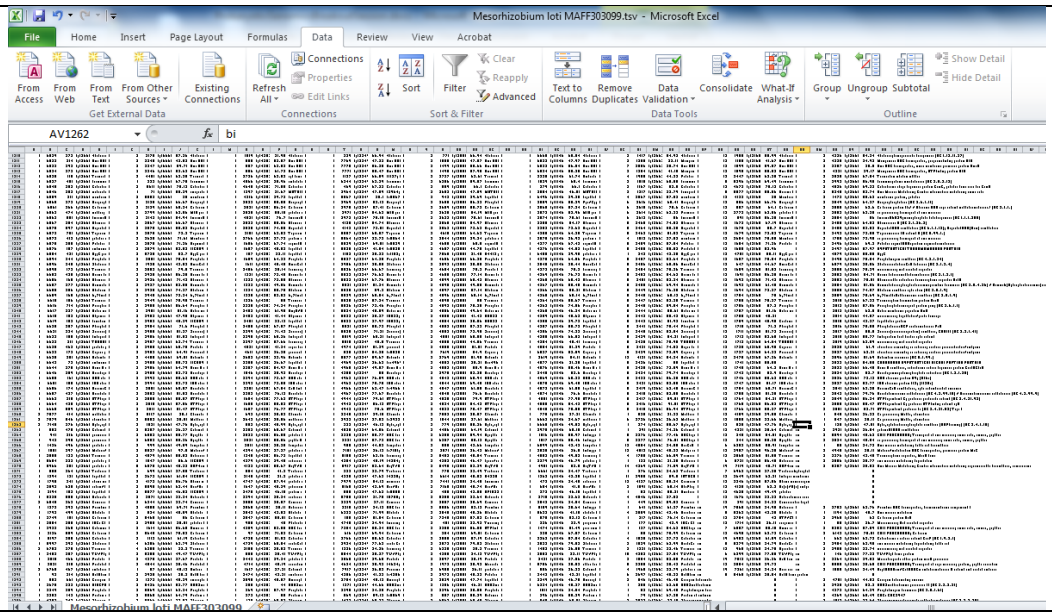
2. Open the .tsv file from the sequence-based comparison in Excel. Click Ctrl+F to activate the find/replace function and replace all dashes with nothing.





2. Click Ctrl+A to select all, then on the Data menu, choose sort, and set a sort level for each comparison organism using the “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 new worksheet.



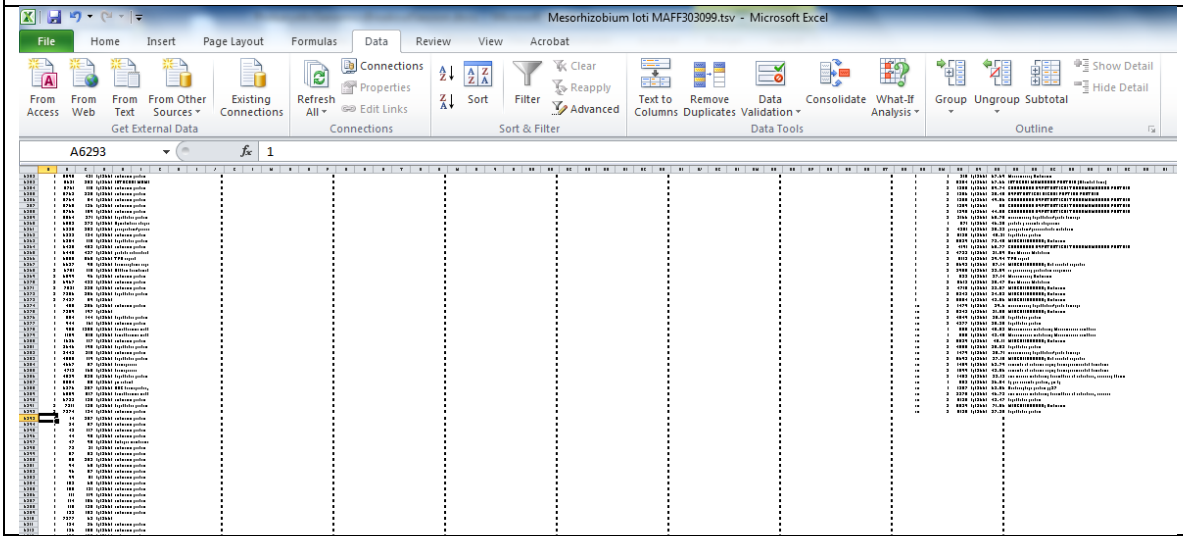


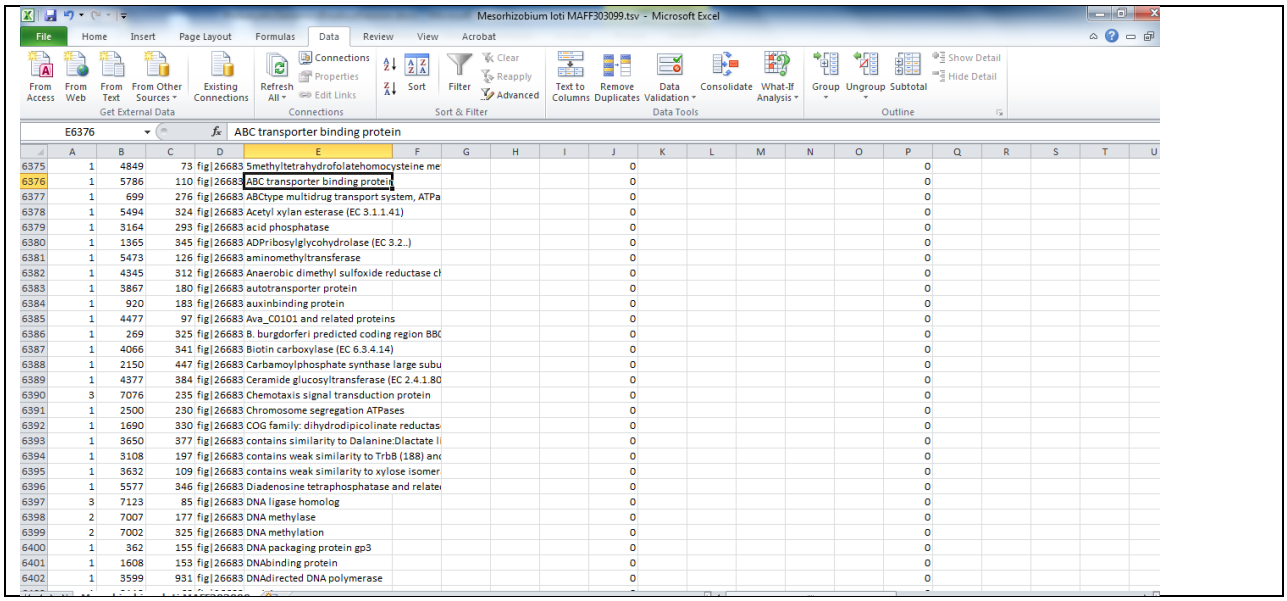
3. Scroll down to determine the number of genes that are "bi" for all of the organisms except the last (e.g. 1,2,3,4 but not 5) and copy that list to a new worksheet and label appropriately. Repeat using one less organism each time (present in 1,2,3, not 4,5). Ultimately determine the number of unique genes and copy them to a new worksheet. Sort these by function (column E) to group those not listed as hypothetical protein.

   What proteins are encoded by these genes?

   What unique phenotypes do you think the organism should have?

4. Repeat this process with different sort orders (e.g. 1,2,4,5, not 3) and different reference genomes (you will need to create separate tsv files) to determine the genes belonging to different sets. Why must one use the different .tsv files? i.e. why can't a single .tsv file just be sorted differently?





	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
6375	1	4849	73	fig|26683	5methyltetrahydrofolatehomocysteine me					0						0					
6376	1	5786	110	fig|26683	ABC transporter binding protein					0						0					
6377	1	699	276	fig|26683	ABCtype multidrug transport system, ATPa					0						0					
6378	1	5494	324	fig|26683	Acetyl xylan esterase (EC 3.1.1.41)					0						0					
6379	1	3164	293	fig|26683	acid phosphatase					0						0					
6380	1	1365	345	fig|26683	ADPribosylglycohydrolase (EC 3.2..)					0						0					
6381	1	5473	126	fig|26683	aminomethyltransferase					0						0					
6382	1	4345	312	fig|26683	Anaerobic dimethyl sulfoxide reductase ch					0						0					
6383	1	3867	180	fig|26683	autotransporter protein					0						0					
6384	1	920	183	fig|26683	auxinbinding protein					0						0					
6385	1	4477	97	fig|26683	Ava_C0101 and related proteins					0						0					
6386	1	269	325	fig|26683	B. burgdorferi predicted coding region BB(					0						0					
6387	1	4066	341	fig|26683	Biotin carboxylase (EC 6.3.4.14)					0						0					
6388	1	2150	447	fig|26683	Carbamoylphosphate synthase 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	Chemotaxis signal transduction protein					0						0					
6391	1	2500	230	fig|26683	Chromosome segregation ATPases					0						0					
6392	1	1690	330	fig|26683	COG family: dihydrodipicolinate reductas					0						0					
6393	1	3650	377	fig|26683	contains similarity to Dalanine:Dlactate li					0						0					
6394	1	3108	197	fig|26683	contains weak similarity to TrbB (188) and					0						0					
6395	1	3632	109	fig|26683	contains weak similarity to xylose isomer					0						0					
6396	1	5577	346	fig|26683	Diadenosine tetraphosphatase 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)**. "IMG: the integrated microbial genomes database and comparative analysis system" . Nucleic Acids Res. (England) **40** (1): D115-22. DOI:10.1093/nar/gkr1044. PMC 3245086. PMID 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