WS16:

Do-it-yourself Microbial Genome Sequence Analysis

Hands-On Exercises

2016

Agenda

8:15	Introduction/Overview	Michelle Giglio
8:45	CloVR background	W. Florian Fricke
9:15	Hands-on: Start CloVR	W. Florian Fricke
9:45	Break	
9:55	Hands-on: Start CloVR-Microbe	W. Florian Fricke
10:15	Hands-on: Start CloVR-Comparative	W. Florian Fricke
10:45	Genome Annotation Pipeline	Michelle Giglio
11:45	Lunch	
12:30	Hands-on: CloVR-Microbe output	W. Florian Fricke
1:00	Hands-on: Manatee	Michelle Giglio
1:30	Comparative Analysis Pipeline	Cesar Arze
2:00	Break	
2:10	Hands-on: CloVR-Comparative output	Hervé Tettelin
3:00	Future Directions	Cesar Arze
3:15	Applications with *S. pneumoniae* and *M. abscessus*	Hervé Tettelin
3:45	Applications with *E. coli*	Tracy Hazen
4:15	End of workshop	

CloVR Hands-On exercises

We have started Virtual Machines for each of you to use during the workshop. The following steps will allow you to access these VMs and use them to run the analysis pipelines.

Start CloVR -

1. Go to the DIAG website: http://diagcomputing.org/

2. login using the DIAG credentials at the beginning of this packet

NOTE: Steps #3 and #4 would be followed if you were starting a VM on your own, however, since we have already started the VMs for you, for the purpose of keeping things flowing on time for the workshop, **DO NOT** do steps 3 and 4 now. They are here for your reference for when you get back home and try this on your own.

3. Click on: 'My Account'. Then choose: 'Start CloVR'
4. This is the CloVR launch page. Click 'Launch CloVR'

5. Access the VMs we have running for you by clicking on 'My Account'. Then choose: 'Monitor VMs'.

6. There should be one VM displayed on this page with the status 'running'. Click on that row to access the VM.

Run CloVR Microbe – These steps will start the genome assembly and annotation pipelines

A. Upload data

1. On the CloVR dashboard, in the datasets panel, click on the 'Add' button, lower left side

2. A box will pop up. This box provides several options for providing CloVR with data including Browsing your computer, directing CloVR to a web location where the information is stored, or using FTP. For today we will use the url option. Paste or type this url into the box (for copying and pasting, an electronic version of this document is on the clovr.org/workshop web page): **http://cb2.igs.umaryland.edu/microbe.tgz** (Make sure that there are no trailing white spaces at the end of the url which generally happens if you cut and paste.)

3. Choose file type: 'Nucleotide FASTQ'

4. Give the file a name, make sure there are no white spaces, use something short and easy.

For example, name: 'illumina_test'

B. Configure the pipeline

1. Select the desired pipeline from the applications header panel, choose: 'Microbe'

2. Tell the pipeline which data file to use by selecting: 'Sequencing Dataset(s)'

   and then: 'illumina_test' (or whatever name you used in step 4 above)

3. Select a CloVR Microbe pipeline track. We'll do 'Assembly + Annotation' for this exercise.

4. Fill in the fields:

   - output prefix: the text you provide here will be a prefix on all output files allowing you to easily identify them. For today use 'nmen'
   - organism: type the genus and species of the organism the data comes from: 'Neisseria meningitidis'
   - DB name: choose a name for the database that will be generated to house the annotation data. For today use: 'nmen'
   - Manatee username: choose a user name for accessing the Manatee database (more on Manatee later) For today use: 'asmguest'
   - Manatee password: choose a password for accessing the Manatee database. For today use: 'stiontio65'

5. Provide a pipeline description, for example: 'asm_microbe' (again no white spaces)

6. Click 'Submit'

Run CloVR Comparative

A. Upload data

1. On the CloVR dashboard, in the datasets panel, click on the 'Add' button

2. To provide CloVR with the data to use in the pipeline, paste or type this url:
   http://cb2.igs.umaryland.edu/compare.tgz
   Again remember to make sure there are no trailing white spaces at the end
   One can also use the taxonomy browsing feature to choose genomes to include or enter genome accession numbers.

3. Choose file type: 'GenBank'.

4. Give the file a name, no white spaces. For example, name: 'neisseria_5genomes_test'

B. Configure the pipeline

1. Select the desired pipeline from the applications header panel, choose: 'Comparative'

2. Tell the pipeline which data file to use by selecting Input GenBank Tags:

   'neisseria_5genomes_test'

3. Type something in for pipeline description, for example: 'asm_comparative' (no white spaces)

4. Click 'Run'

Manatee Demonstration

1. To get started with Manatee go to **http://manatee.igs.umaryland.edu**
   Due to constraints on our resources of running 50 CloVR VMs at one time, we could not also start up Manatee VMs for this part of the workshop. We will instead use Manatee running on the IGS web server. For reference when you get back home, here is a link to instructions for accessing the Manatee VMs to view the databases resulting from the CloVR Microbe pipeline runs: http://tinyurl.com/ManateeDocs

2. Type in these credentials:
username 'asmguest'
password 'stiontio65'
database 'nmen'
This information is also on your credential sheet at the beginning of this packet.

3. You should get a 'Welcome to Manatee' page. (Figure 1) There are numerous search, display, and file download options on this page. Explore this page.

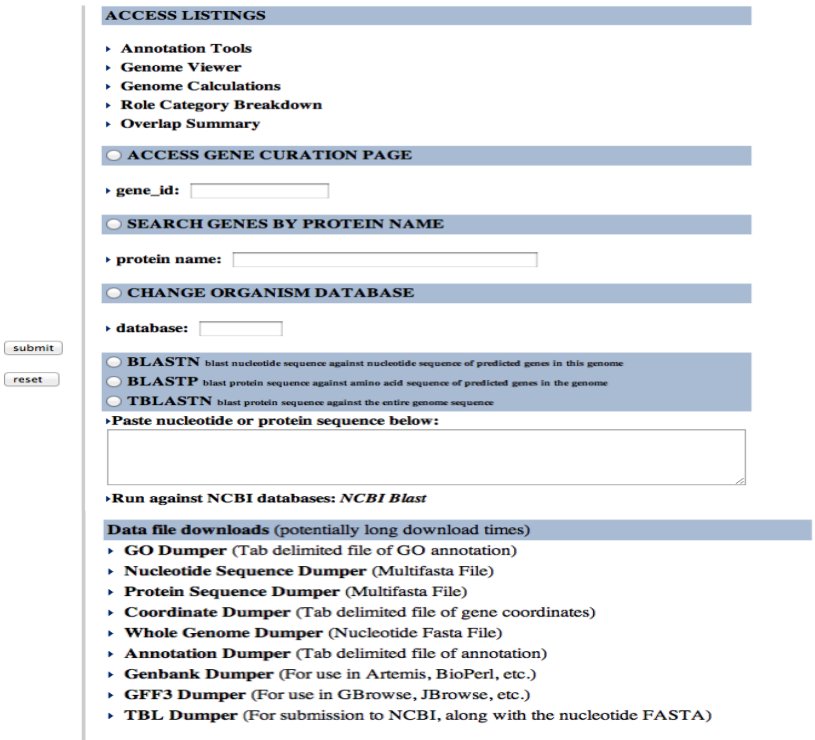


Figure 1

4. Let's click on 'Annotation Tools' at the top of the Welcome page. This gives you a page with more search options (Figure 2). Let's leave the default selection: 'all genes, ordered by role category'. Click 'Submit'.

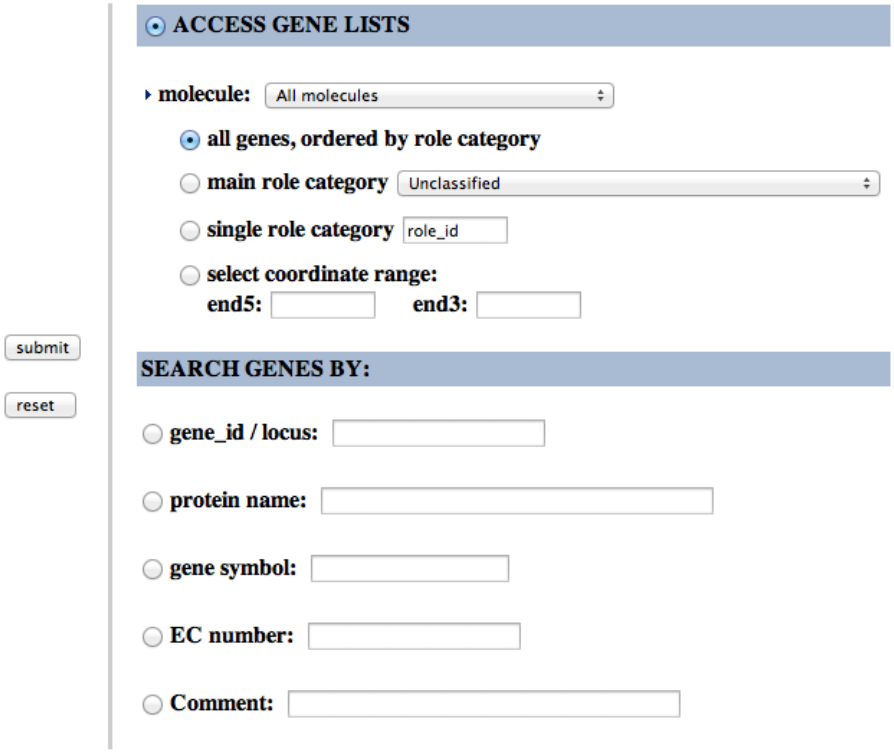


Figure 2

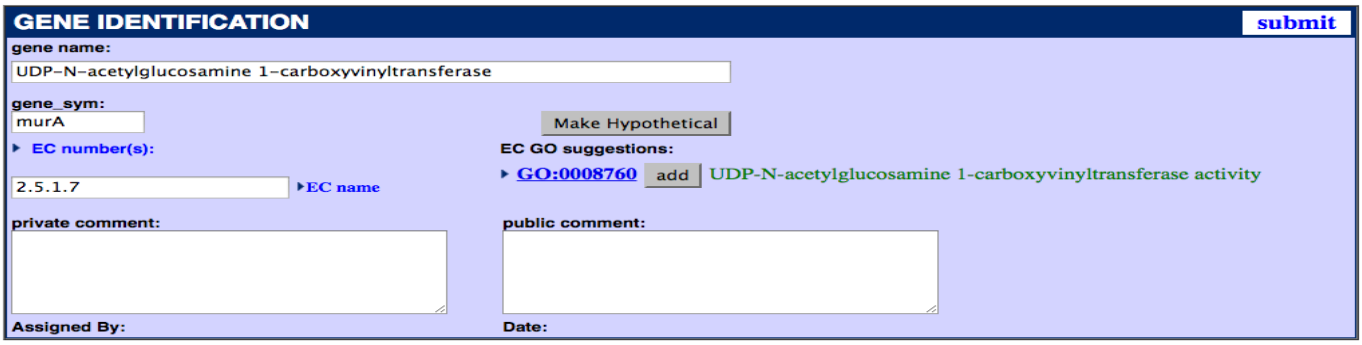
5. We then get a Gene List page (Fig 3). Click on a gene_id to get to a Gene Curation Page. (I will demo **nmen_3**)

(snippet of the list)

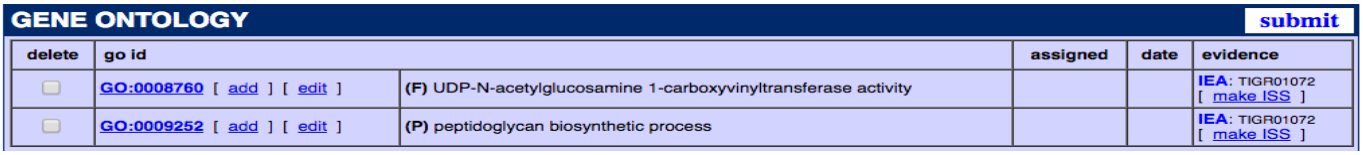
Biosynthesis and degradation of surface polysaccharides and lipopolysaccharides										Role id: 90
C	seq id	gene_id	locus	end5	end3	gene name	gene symbol	ec	other roles	start_edit
	MC58_region_for_assembly	nmen_9		10245	9322	UDP-3-O-[3-hydroxymyristoyl] N-acetylglucosamine deacetylase	lpxC	3.5.1.-		
	MC58_region_for_assembly	nmen_7		6212	4941	glycosyl transferases group 1 family protein				

Figure 3

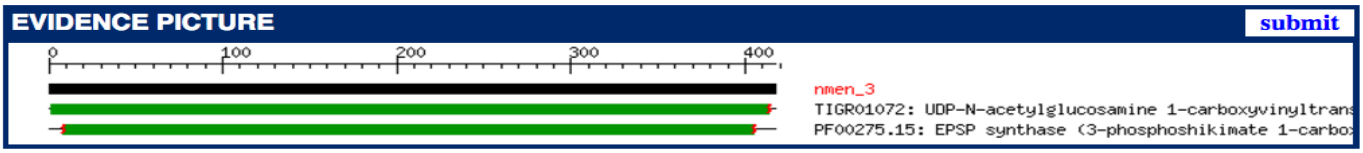
6. The Gene Curation Page – let's look at **nmen_3** (from the 'Biosynthesis of murein sacculus and peptidoglycan' section).
This page displays all of the automatically assigned annotations as well as the supporting evidence that was used to make them. (This is a long page – below I have showed a sampling of segments.) Users can change the annotations by entering information into the fields and clicking 'submit'. (Although for this workshop you have only read access to the database.) Explore the information on this page.



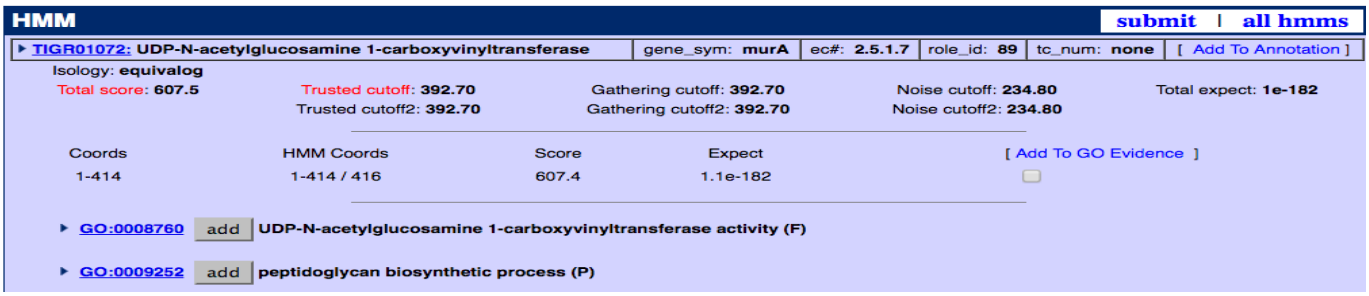
Protein name and other annotations.



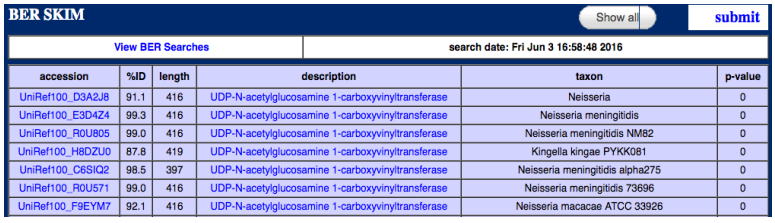
Gene Ontology annotations



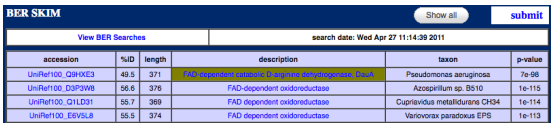
Evidence Summary graphical display



HMM evidence

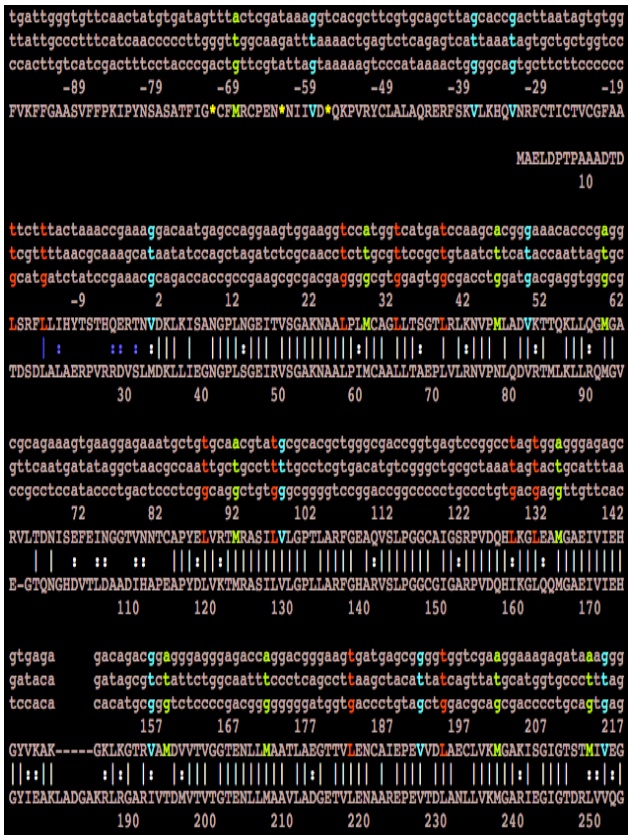


BER evidence



BER evidence with characterized protein

7. Explore the links from this page
Page links include BER alignment files (by clicking on match protein names in the BER skim section), sequence pages (by clicking on the 'View Sequences' button near the top left of the page), TIGRFAM and Pfam links (by clicking on the HMM accession numbers), etc.



Example alignment.

CDS

```
>nmen_3
GTGGACAAACTGAAAATCTCCGCAAACGGCCCGCTCAACGGGGAAATAACGGTCTCGGGC
GCGAAAAACGCGGCATTGCCGCTGATGTGCGCGGGTTTGCTGACATCGGGTACGTTGCGC
CTGAAAAACGTCCCTATGCTGGCAGATGTGAAAACCACGCAAAAGCTGCTTCAGGGGATG
GGCGCGCGCGTCCTGACCGACAATATCAGCGAATTTGAAATCAACGGCGGTACGGTAAAC
AATACCTGCGCCCCTTACGAGTTGGTCCGAACGATGCGCGCTTCGATTTTGGTGCTGGGC
CCGACGCTGGCGCGTTTCGGCGAGGCGCAAGTCAGCCTGCCGGGCGGCTGCGCCATCGGT
TCGCGCCCCGTCGATCAGCATTTGAAAGGCTTGGAAGCGATGGGTGCTGAGATTGTTATC
GAACACGGTTACGTCAAAGCCAAAGGCAAACTCAAAGGTACGCGCGTGGCGATGGATGTC
GTTACCGTCGGCGGCACGGAAAACCTGCTGATGGCGGCGACGCTGGCGGAAGGTACGACG
GTTTTGGAAAACTGCGCCATTGAGCCTGAAGTGGTCGATTTGGCGGAATGCCTGGTCAAA
ATGGGCGCGAAAATCAGCGGCATCGGTACGTCCACAATGATTGTGGAAGGGGTGGACGAG
CTGCAAGGCTGCGAACACAGCGTCGTCCCCGACCGGATCGAGGCGGGGACGTTCCTGTGC
GCGGTGGCGATAACCGGTGGCAGGGTGGTTTTGCGGAATGCCGCGCCGAAAACGATGGAA
GTGGTGTTGGACAAACTGGTTGAGGCAGGTGCGGTGATTGAGGCGGGCGACGATTGGATC
GCCATCGATATGCGGCAGCGTCCGAAGGCGGTGGACATCCGCACGGTCGTCCACCCCGGC
TTCCCCACGGATATGCAGGCGCAGTTTATGGCATTGAATGCCGTGGCGGAGGGAAGCTGC
CGCGTGGTGGAAACGATTTTTGAAAACCGCTTTATGCACGTCCCCGAGTTGAACCGGATG
GGGGCGAACATCACAACCGAGGGCAATACGGCATTTGTGCAGGGTGTGGAACAGCTTTCC
GGCGCAGTCGTCAAGGCGACGGATTTGCGTGCGTCCGCCAGCCTCGTTATCGCCGGTTTG
GCGGCGCGAGGCGAAACCGTGGTCGAACAGATTTACCACTTGGATCGCGGTTATGAAAAT
ATTGAAAAAAAACTCGGCAGCGTCGGCGCGAAAATCGAGCGCGTATCCGGCTGA
```

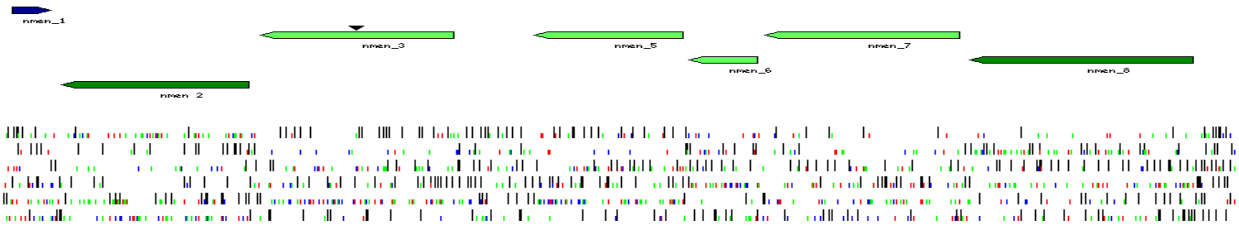
Protein

```
>nmen_3
VDKLKISANGPLNGEITVSGAKNAALPLMCAGLLTSGTLRLKNVPMLADVKTTQKLLQGM
GARVLTDNISEFEINGGTVNNTCAPYELVRTMRASILVLGPTLARFGEAQVSLPGGCAIG
SRPVDQHLKGLEAMGAEIVIEHGYVKAKGKLKGTRVAMDVVTVGGTENLLMAATLAEGTT
VLENCAIEPEVVDLAECLVKMGAKISGIGTSTMIVEGVDELQGCEHSVVPDRIEAGTFLC
AVAITGGRVVLRNAAPKTMEVVLDKLVEAGAVIEAGDDWIAIDMRQRPKAVDIRTVVHPG
FPTDMQAQFMALNAVAEGSCRVVETIFENRFMHVPELNRMGANITTEGNTAFVQGVEQLS
GAVVKATDLRASASLVIAGLAARGETVVEQIYHLDRGYENIEKKLGSVGAKIERVSG
```

Example Sequences

8. You can get back to the Welcome page at any time by clicking 'Home' in the upper right section of all Manatee pages.

9. Let's explore Genome Viewer. Click on the Genome Viewer link at the top of the Gene Curation Page, or go back to the Home page and click on Genome Viewer there.



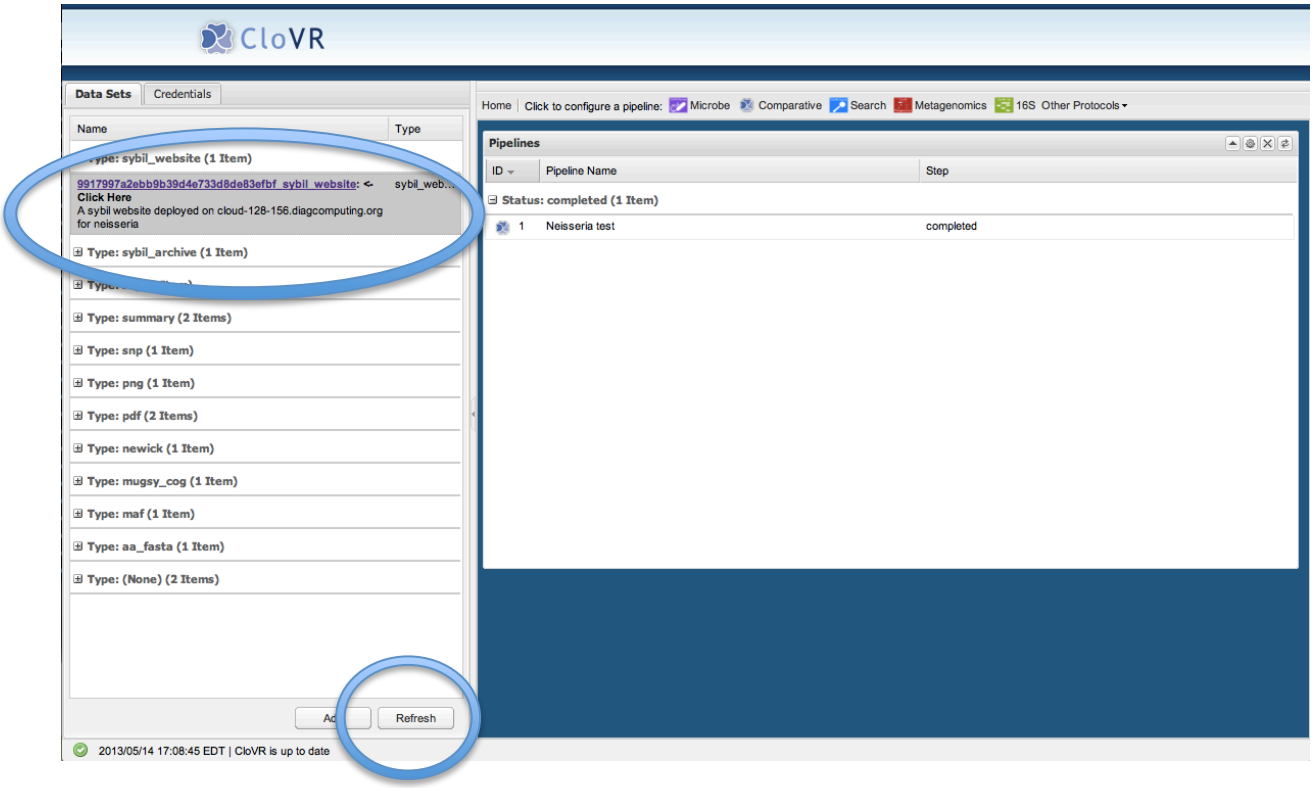
10. A standing Manatee demo is always available at manatee.igs.umaryland.edu.
The database is 'cgspu', the username is 'training1', the password is 'training1'. This user account has read, but not write, privileges to the database.

Accessing the Sybil website on your VM

Find the Sybil website dataset in your dataset panel.

This could require clicking 'Refresh' at the bottom.

Click the link that should look like 'blablabla_sybil_website' where blablabla is a big identifier.

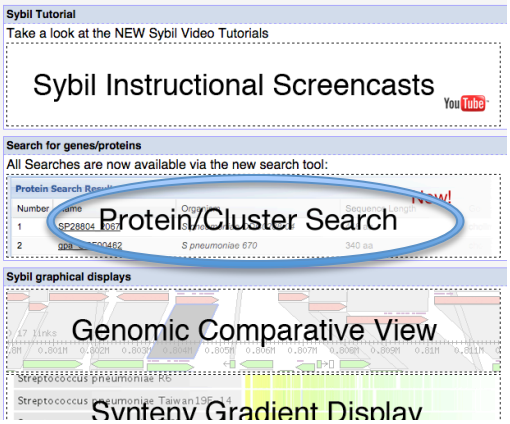


Sybil home page

This brings you to the Sybil home page. At the bottom right you will see the genomes included in this particular Sybil instance. We are going to compare five complete gap-free genomes of the Gram-negative bacterium *Neisseria meningitidis*. Each genome has only one chromosome of ~2.2Mb. Please note that CloVR and Sybil also work with genomes containing multiple replicons as well as draft genomes, in which case replicon/contig information would be reported in searches and displays.

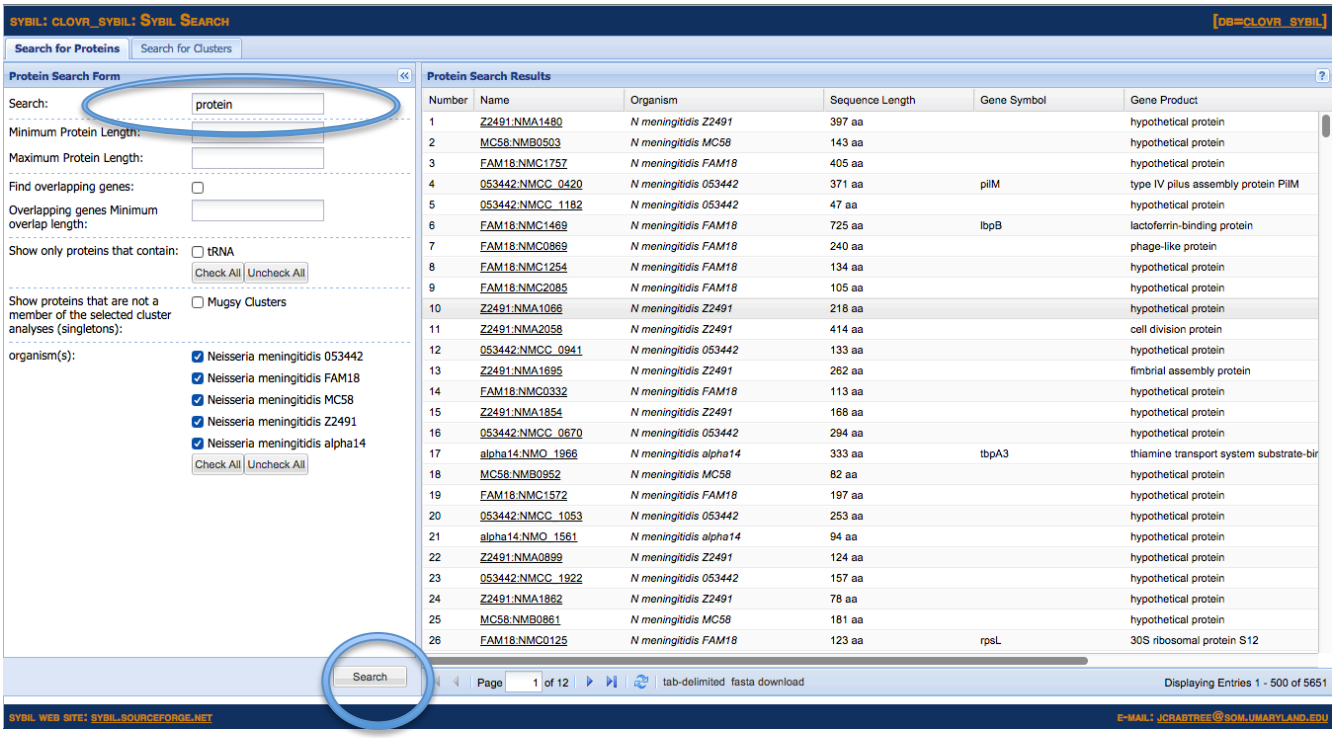
Protein search

1. From the Sybil homepage click on the Protein/Cluster Search button.



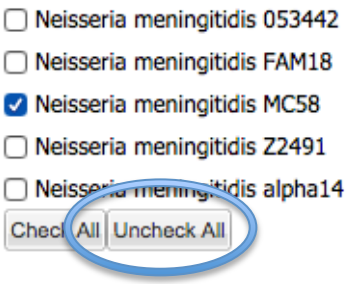
This will take you to the protein search page.

2. First we'll search for a gene by keyword. Enter the word 'protein' in the Search box and click 'Search' at the bottom.



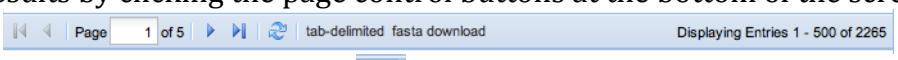
This will bring up genes that contain the word 'protein' in their gene product annotation.

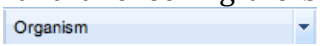
3. Click 'Uncheck All' then select only '*Neisseria meningitidis* MC58' and click 'Search' to restrict the search to only that specific genome/strain.



Search form tips

- Browse the results by clicking the page control buttons at the bottom of the screen.



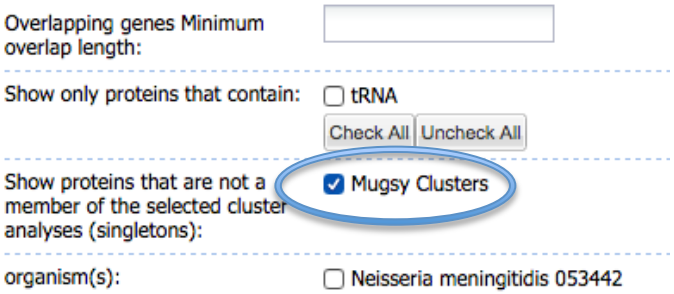
- Hide the search window by clicking the  button at the top of the search form.
- Hide columns by clicking the down arrow in the column header and unchecking the box. The arrow will appear when you hover over the header with your mouse. 

- Reorder columns by clicking and dragging the column header.
- Sort by the 'Name' column by clicking on the column header.

3. Next, let's search for genes that are unique to the '*Neisseria meningitidis* MC58' genome.

The CloVR Comparative pipeline has grouped genes predicted in all genomes into Mugsy-based clusters of orthologs. This method of clustering makes use of the local synteny obtained from the Mugsy whole genome multiple alignment (see this reference for details: Angiuoli SV, Dunning Hotopp JC, Salzberg SL, Tettelin H. (2011) Improving pan-genome annotation using whole genome multiple alignment. BMC Bioinformatics 12: 272). This approach was briefly described in Sonia Agrawal's presentation.
There are many other methods for generating clusters of orthologs including Jaccard-based clusters of orthologs (JOCs)(Crabtree J, Angiuoli SV, Wortman JR, White OR (2007) Sybil: methods and software for multiple genome comparison and visualization. Methods Mol. Biol. 408: 93-108) and Blast Score Ratio-based clusters (see Tracy Hazen's presentation later today). Multiple clustering methods can be loaded into the same database in which case Sybil will allow selection of which one to use for a given search.

- Check the box next to 'Mugsy Clusters'. This will limit our search results to those genes that are singletons in the Mugsy protein clustering. This means that they are not members of a Mugsy protein cluster.
- Click 'Search'.



This will bring up all MC58 genes that contain the word 'protein' somewhere in their annotation but that are not contained in a Mugsy gene cluster. This means that these genes are specific to a particular genome, i.e. singletons (note that some of these genes might still have a BLAST hit to another genome, for example if they are located in repeated regions).

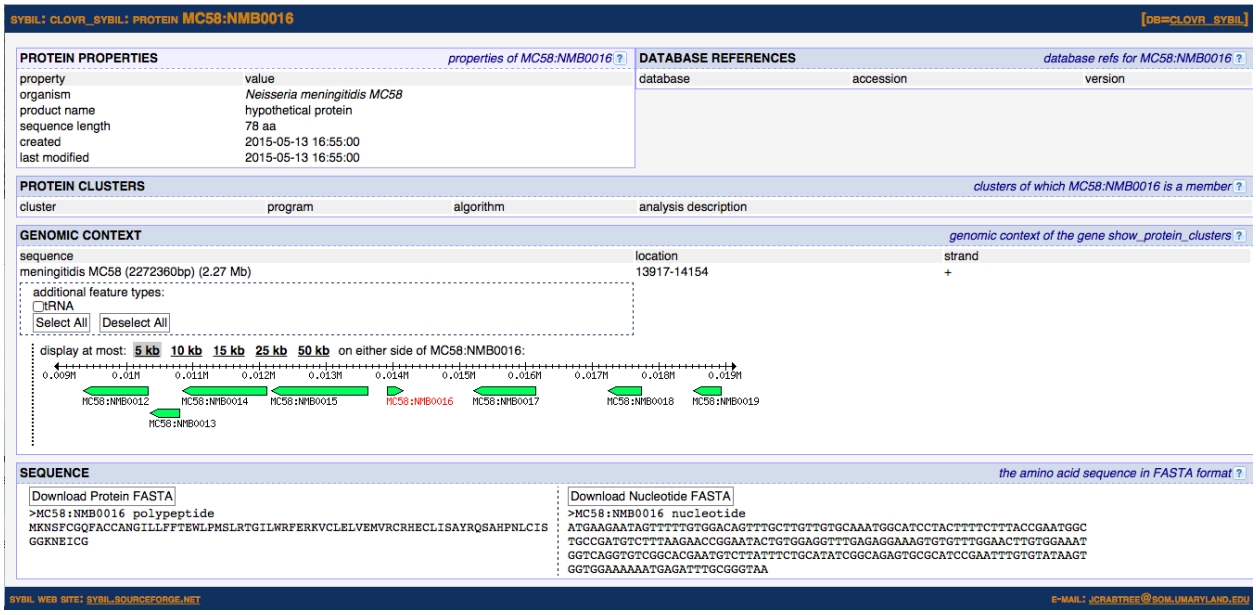
5. Click on the name of the gene 'MC58:NMB0016' which is annotated as a 'hypothetical protein' to go to the protein report page.

Protein Search Results

Number	Name ▲	Organism	Sequence Length	Gene Product
123	MC58:NMB0016	*N meningitidis MC58*	78 aa	hypothetical protein
139	MC58:NMB0046	*N meningitidis MC58*	74 aa	hypothetical protein
57	MC58:NMB0067	*N meningitidis MC58*	225 aa	polysialic acid capsule biosynthesis protein SiaD
156	MC58:NMB0093	*N meningitidis MC58*	28 aa	hypothetical protein
112	MC58:NMB0099	*N meningitidis MC58*	47 aa	hypothetical protein

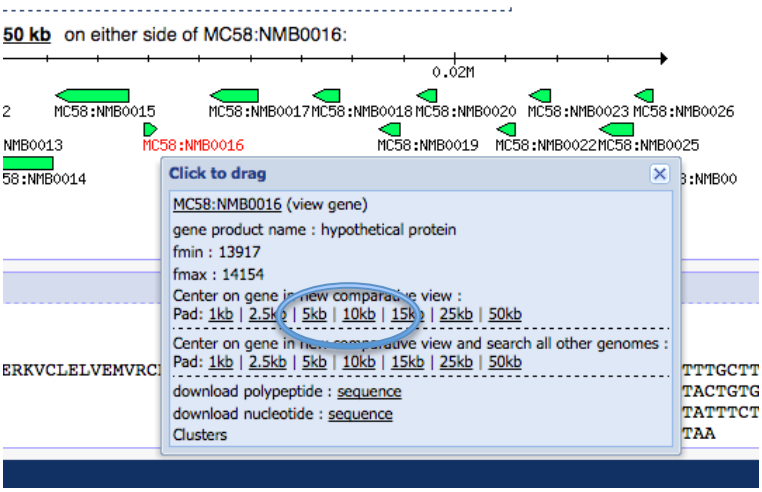
Protein report

The protein report page provides summary information about a gene in addition to providing graphical representations of the genomic context and the BLAST results.



1. Click on the '10kb' button just above the image. This will redraw the image with 10kb on either side of the gene.
2. Click on 'tRNA' just above the 10kb button to show any tRNA present in this region. One appears as a small empty box between NMB0011 and NMB0012.
3. Click on the tRNA box itself to bring up its popup window with coordinates.
4. Click on some genes in the display to pull up additional information in popup windows.
5. Click on the gene labeled in red – MC58:NMB0016 – to bring up its popup window.
6. Notice that no clusters are listed for this gene since it was selected as a singleton.

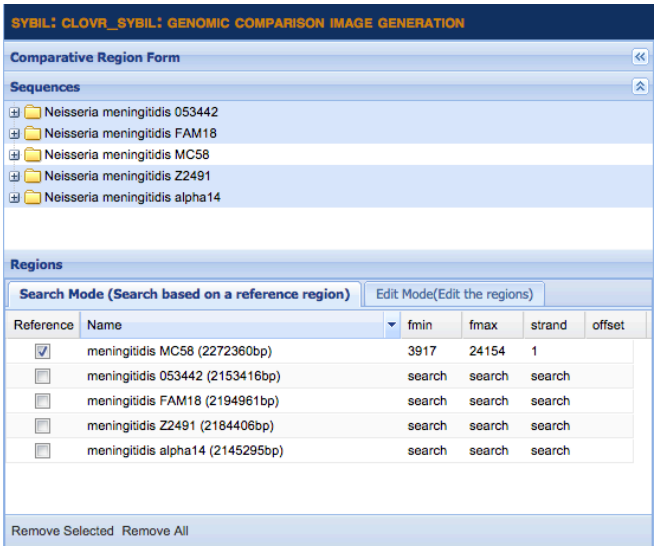
7. Click on the '10kb' link just below 'Center on gene in new comparative view'.



Clicking this link will take you to the 'Genomic Comparative View' and will center your view on this gene's coordinates.

Genomic comparative view

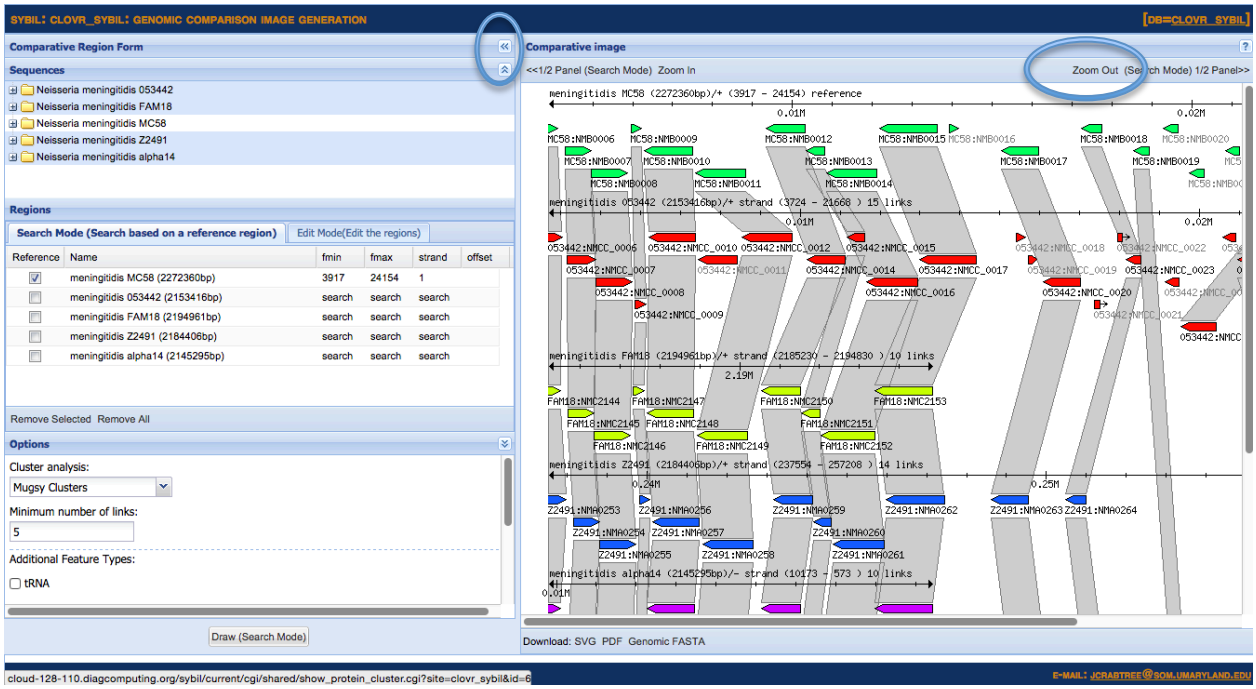
The genomic comparative view provides a graphical representation of the genomic context of multiple genomes.



1. The first step is pulling in some genomes to search. This is done by selecting sequences from the 'Sequences' table and dragging those sequences into the table below – labeled 'Search Mode'. You can select multiple sequences using the shift or command/ctrl key. For this exercise, select the remaining 4 genomes – 053442, FAM18, Z2491 and alpha14 – and drag them into the table.

2. Click 'Draw (Search Mode)' at the bottom of the form.

This will draw the reference region we selected – our gene of interest with 10kb on either side – with matching genomes below. Once the image has loaded, click the  button to hide the search form, leaving just the picture on the screen.



3. Click the box labeled 'Zoom Out' at the top right hand side of the screen to extend the left and right flanks. Notice that now a second segment of the FAM18 genome is displayed on the right side. It only appeared now because enough flanking genes with Mugsy cluster links are available to display it. The default number of links is 5 but it can be changed in the search form.
4. The gene names in grey font are singletons (not members of a Mugsy cluster).
5. Notice the breaks and coordinates in the FAM18 and alpha14 genomes.

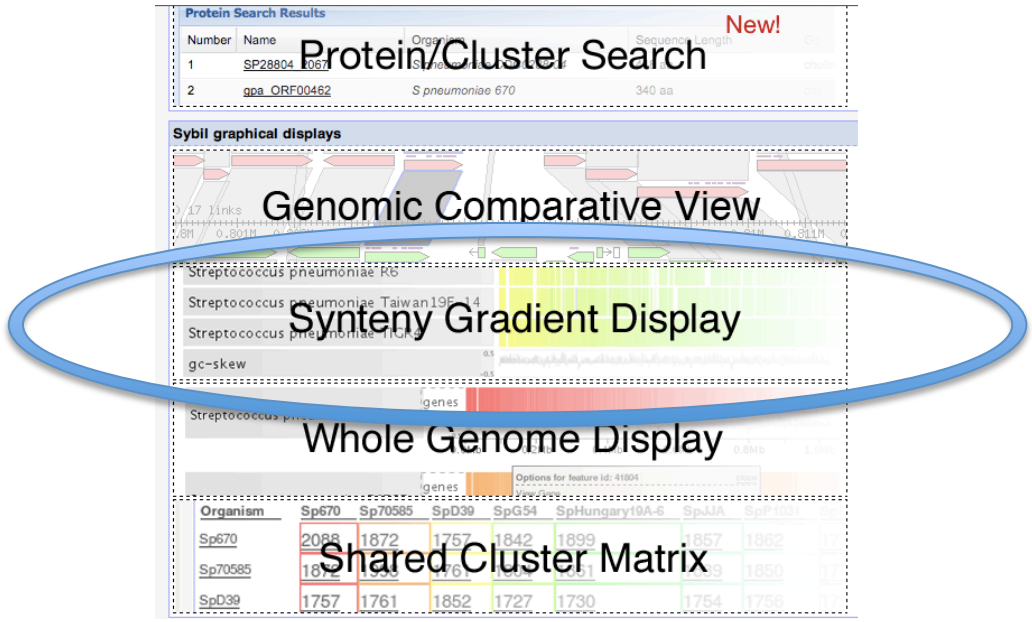
Genomic comparative view tips

- Click on genes to see more information about them.
- Click on cluster links (grey polygons) to see more information about them, including a list of all the cluster members whether they are shown on the current display on not.
- Zoom in/out and pan controls are available in the sub-title pane at the top. Note that in 'Search Mode' this will zoom/pan the reference and re-search the other genomes for matching regions.
- Export the image you are seeing to SVG or PDF by clicking the export buttons at the bottom.

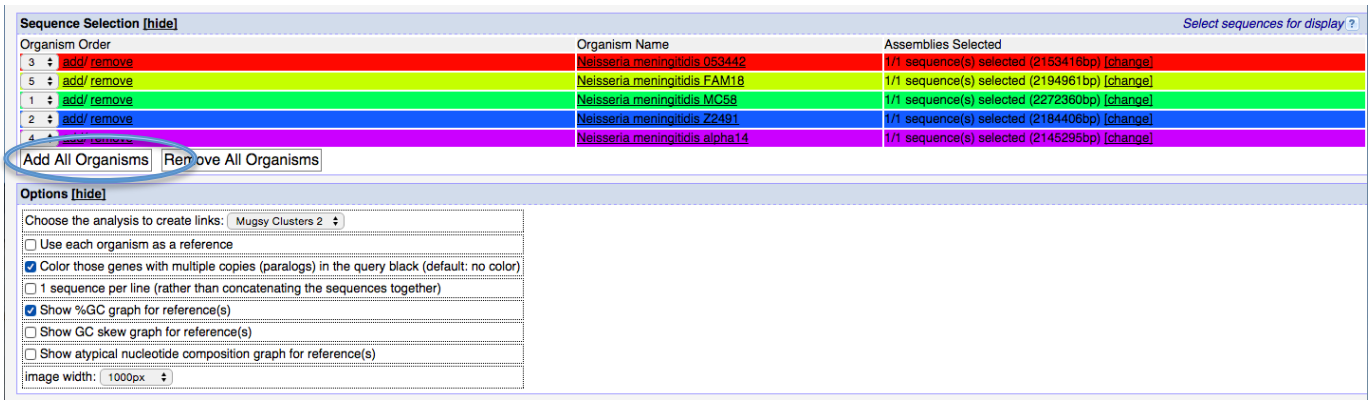
Gradient display

Lastly we'll look at the whole-genome gradient display.

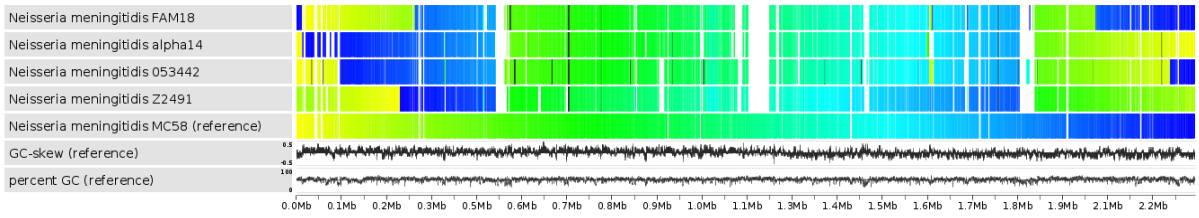
1. Click on the link at the top right hand side of your browser window labeled 'DB=neisseria'. This link is available on all views and will take you back to the Sybil homepage.
2. Click on the button labeled 'Synteny Gradient Display'



3. Click on the 'Add All Organisms' button at the end of the list of organisms in the 'Sequence Selection' box. The organism in position one will be the default reference.
4. Check the box next to 'Color those genes with multiple copies (paralogs) in the query black'.
5. Check the box next to 'Show %GC graph for reference(s)'.
6. Click 'Draw'.



This will draw a gradient display comparing the reference genome to the other genomes selected.



This particular display shows MC58 as the reference genome. Genes from the remaining 4 genomes are drawn above the MC58 gene but are colored based on their position in their native genome.

Gradient display tips

- Clicking on genes in the reference brings up popups about those genes and you can link to protein report/cluster report pages.
- Export the images using the 'PNG', 'SVG', 'PDF' and 'JPEG' buttons.

View the Sybil screencast tutorials (linked from the front page)
http://www.youtube.com/user/SybilScreencasts

Sybil sourceforge website
http://sybil.sourceforge.net

Sybil publications

Riley DR, Angiuoli SV, Crabtree J, Dunning Hotopp JC, Tettelin H (2012) Using Sybil for interactive comparative genomics of microbes on the web. Bioinformatics 28: 160-166.

Crabtree J, Angiuoli SV, Wortman JR, White OR (2007) Sybil: methods and software for multiple genome comparison and visualization. Methods Mol. Biol. 408: 93-108.