**Introduction to Genome Annotation**

**by Selene Nikaido**

Adapted from **Annotation Exercise – Treasure Hunt by Anne Rosenwald, Gaurav Arora, and Vinayak Mathur**

*Note: The purpose of this exercise is to familiarize the user with data visualization and basic data annotation. The database that we will be using for this exercise is the Integrated Microbial Genomes database found at:* https://img.jgi.doe.gov/cgi-bin/edu/main.cgi

**Learning Goals**

* To navigate a database that stores prokaryotic sequence data including data from the Human Microbiome Project (HMP)
* To search the features of a genome browser
* To download data from a database
* To compare gene regions on a genome browser

**Background Requirements**

This exercise assumes that a student understands:

* The structure and function of DNA
* Biological classification and scientific names of organisms
* Ideas of inheritance
* The relationship between DNA and appearance (phenotype)

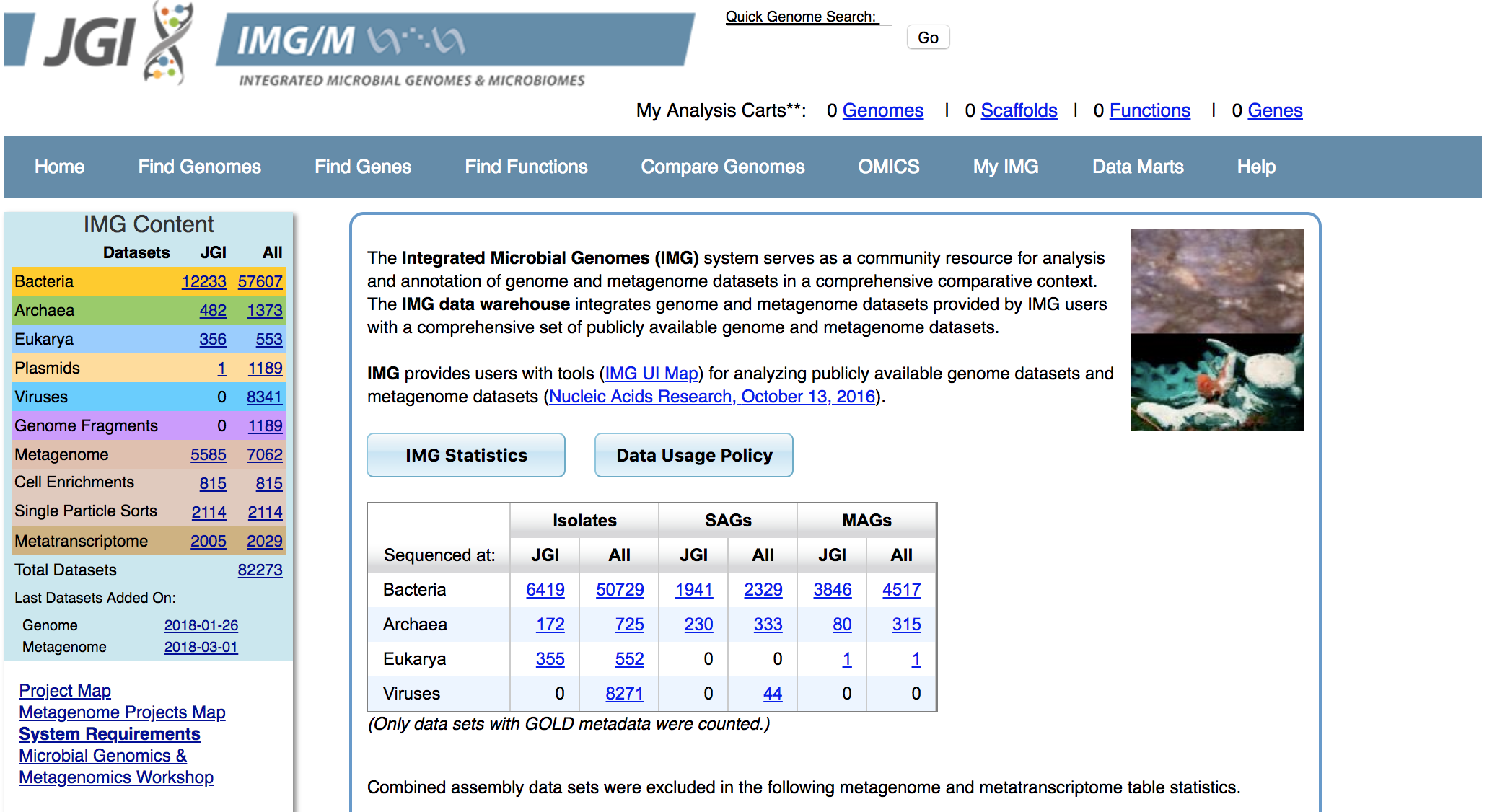
**Pre-Exercise Activity: Genome Annotation Slide Deck**

Start with the slide deck, Genome Solver: Genome Annotation. Before continuing with the exercise, answer the following questions based on the slide deck.

1. What is an annotation? When annotating a map, what does one define? In data figures, we call this a figure legend.
2. A genome is the complete genetic information of an organism. What are two broad categories of genome annotation? For genomes, you would preface a question of annotation with “where in the genetic data (information) is…?”
3. What is the difference between structural and functional annotation? Both attempt to define the landscape of a genome, but from two different points of view.
4. Reading frames attempt to interpret DNA sequence into protein sequence. If a segment of DNA (mRNA) is translated from the 5’ end to the 3’ end, why are there six reading frames? There are two distinct answers to this question.
5. Technically, what is an **open** reading frame? Identifying the reading frame that has an open reading frame will locate possible protein-coding genes.
6. The slide deck says that a region between two stop codons of the same reading frame is an open reading frame. What else must be found between the two stop codons? It is necessary to start translation. [Note: An open reading frame (ORF) is found in the same reading frame in prokaryotes. In eukaryotes, an ORF may span more than one reading frame because eukaryotic genes have introns.]
7. What is the importance of a Shine-Dalgarno sequence for bacterial genome annotation?
8. Computer programs are used to assemble DNA sequences into a genome sequence. To annotate genome sequences, computer programs look for patterns in the DNA sequence. What is the significance of finding an AT rich region? GC rich region? Shine-Dalgarno sequence? AG rich region?
9. Besides *ab initio* methods described above (#8), what is another way one can use a computer program to identify genes? [*Ab initio* means “from the beginning,” and it generally refers to starting from theory or based on theory – e.g., consensus sequences.] Hint: This second method requires prior knowledge that is based on direct information and ideas of evolution.
10. Functional annotation of genes relies on laboratory-derived information from related organisms. A database of previously identified genes is used to search for similarities to unannotated sequences. [For more information about hidden Markov models see <https://www.nature.com/articles/nbt1004-1315>.] The “known” proteins and their genes are characterized using laboratory and computation methods. Gene ontology describes the concept of categorizing or indexing proteins in a database. What are three categories of indexing proteins in gene ontology?
11. Direct laboratory experiments, human curation of genomes, and simple sequence comparisons, which of these provides the strongest evidence to characterize a “gene”? Which of these is the easiest to perform?
12. Let’s say you just discovered a new bacterium. Its genome is relatively small, and it is now possible to sequence the entire genome in your own lab. Although the sequencing of the DNA and the computational work to put together the sequenced pieces into a contig (contiguous sequence) can be quite labor-intensive, the more laborious work comes next – genome annotation. Why do you want to identify and remove from your analyses tRNA and rRNA genes FIRST?
13. The next step in genome annotation is to use computer programs to identify possible protein-coding genes by performing structural and functional annotations. Genomic data are stored in searchable databases. This exercise will introduce you to these databases. Let’s begin.

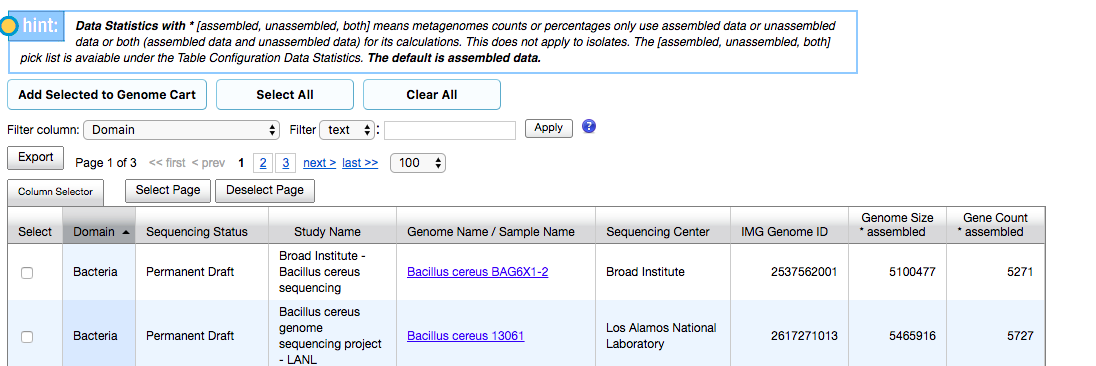
**Part 1: Exploring and downloading data from a human microbiome database:**

Using a browser, visit the following site: <https://img.jgi.doe.gov/cgi-bin/edu/main.cgi>. You will see a homepage similar to the one shown on the next page. From this homepage, find answers to the following questions.

1. How many bacterial datasets are in this database?
2. What is JGI?
3. What is IMG/M?
4. The projects map (see menu on the left) shows where bacteria were collected for genome annotation. List some places where bacteria were collected. Were they always on land? To get back to the homepage, click on the IMG/M tab in the menu bar across the top.
5. Craig Venter, once CEO of a private company called Celera, used shot-gun sequencing to compete with the public Human Genome Project to finish a draft of the human genome. He started a non-profit organization for a new adventure. What is it and what does it do? Visit these sites to find out. <http://www.genomenewsnetwork.org/articles/2004/03/04/sargasso.php>; <https://www.jcvi.org/gos>

Many bacteria cannot be cultured in the laboratory. So, Venter’s shot-gun sequencing and computer programs to assemble the sequences are methods used to identify new species.

Under the ‘Find Genomes’ tab, click on ‘Genome Search’. On the next page in the search bar type *Bacillus cereus*. On the results page that appears, click on the number next to “Genome Name/Sample Name.” A table like the one below appears on the final page.



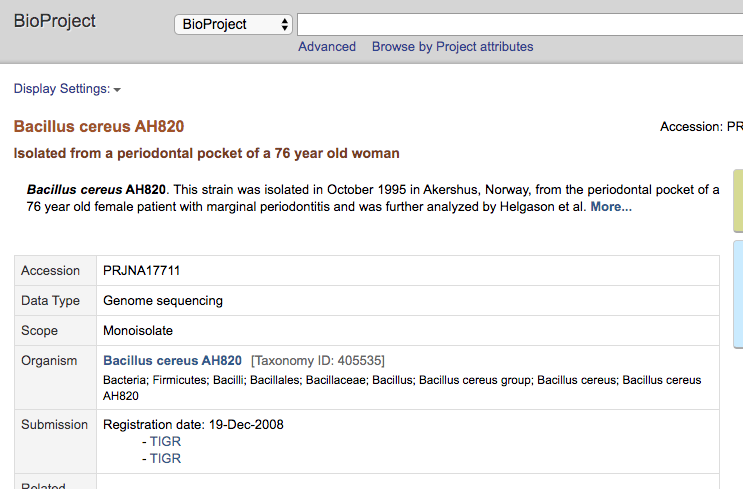
1. What are the different states of sequencing status?
2. Search (use browser Find function) for AH820. You may have to scroll to the next page. What is the sequencing status of the *Bacillus cereus* AH820 genome? Does this mean that the annotation is complete for this strain of *B. cereus?*

Click on *Bacillus cereus* AH820 Genome Name. Look under ‘Overview’ to answer the following questions:

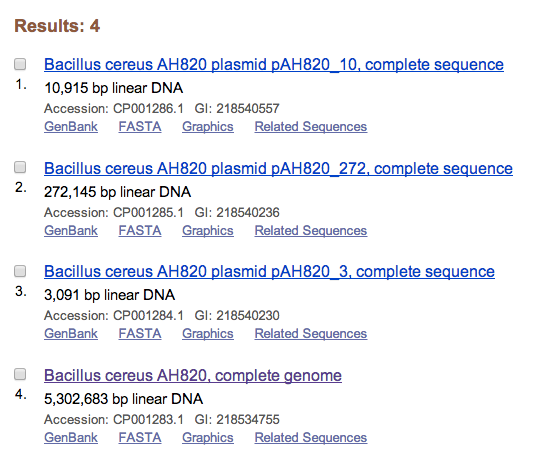
1. Where was the genome of this bacterium sequenced? (*Hint: look at the row titled ‘Sequencing Center’)*
2. When and where was the bacterium isolated? (*Hint: look at the row titled ‘Geographic Location’)*
3. Name three phenotypes associated with this organism? (*Hint: look at the rows under ‘Phenotypes/Metabolism from Pathway Assertion’)*

Look at the information under ‘Genome Statistics’ to answer the following questions:

1. What is the size of the genome in number of bps? (*Hint: look under the column titled ‘DNA, total number of bases’)*
2. What is the GC% associated with this organism? (*Hint: look under the column titled ‘DNA, total number of bases’ for G+C)*
3. How many genes have been annotated as protein coding? (*Hint: look under the column titled ‘Gene, total number’)*

Under ‘Overview’, look for the ‘NCBI Bioproject Accession’ and click on the accession number. You will be directed to a new page (similar to the one on the right) on the NCBI site, <https://www.ncbi.nlm.nih.gov/>.

Near the bottom of the page, find the ‘Project Data’ table. Click on the number next to the ‘Nucleotide’ in the ‘Resource Name’ column. You will be directed to a page where all sequences associated with this organism can be accessed. Below is an image of the results page.



1. Which scaffold (genome assembly) has the most base pairs (bp)? (*Give accession number)*

If you want to save the genome sequence for research or a classroom exercise, check mark the box next to the genome sequence you would like to save, pull down the ‘Summary’ tab near the top, check mark ‘FASTA (text)’. and save the FASTA sequence of this genome as a \*.txt file.

**Part 2: Comparing gene regions on a genome browser (you will need to be logged into the IMG database for this part):**

Return to the following site: https://img.jgi.doe.gov/cgi-bin/edu/main.cgi

Open the drop-down menu titled ‘Find Genes’. Click on the ‘Gene Search’ tab and on the next page enter accession IDs 650934052, 637017746, and 2516306962 in the ‘Keyword’ box. [Separate the accession IDs with commas as the search function is comma delimited. Do not include the word, ‘and.’] Click on the number next to the ‘IMG Gene ID’ in the ‘Quick Search Filter’ column.

On the resulting page, select the boxes associated with each gene and click on the ‘Add Selected to Gene Cart’ tab. On the following page, click on the ‘Gene Neighborhoods’ tab near the top, and then select the forward direction (5’ – 3’ direction of each selected gene is left to right). Click on ‘Show Neighborhoods’ button. You will see a graphical interface associated with your genes. Based on this interface answer the following questions:

1. What are the scientific names of the organisms that are being compared in the graphical output?
2. The gene that you are interested in is colored in red. How does the size of this gene in *H. pylori* compare to other organisms?

Scroll over the gene in *H. pylori* and click on it. You will be directed to a Gene information page (**make sure that you open this page in a NEW tab - Windows users: right click and select ‘Open link in New Tab’. Mac users: Press Ctrl and click on your mouse pad, and then select ‘Open link in New Tab’**). Based on the information on this page answer the following questions:

1. What are some of the gene ontology terms associated with this gene (*Hint: see the GO terms*).
2. What are some of the protein families associated with this gene? (*Hint: see the information associated with Pfam).*

Go back to the graphical interface with all the genes from your analysis cart represented. Click on the same gene region (in red) for *H. bizzozeronii.*

1. What protein families are associated with this gene?
2. How does this compare to the information for the gene in *H. pylori?* Based on this information, what can you infer about the annotation of the *H. bizzozeronii* genome in comparison with *H. pylori*. Scroll over the outlined “genes” of *H. bizzozeronii* to confirm your inference.

**Reflection**. Write a paragraph about what you learned about genomes and annotation of genomes from this exercise.