**Annotation Exercise – Treasure Hunt**

Learning Goals:

* Become familiar with a database that stores prokaryotic sequence data
* Become familiar with the features of a genome browser
* Be able to download data from a database
* Be able to compare gene regions on a genome browser

*Note: The purpose of this exercise is to familiarize the user with data visualization and basic data annotation.*

1. *The database that we will be using for this exercise is the Integrated Microbial Genomes database found at:* [*https://img.jgi.doe.gov*](https://img.jgi.doe.gov/)

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

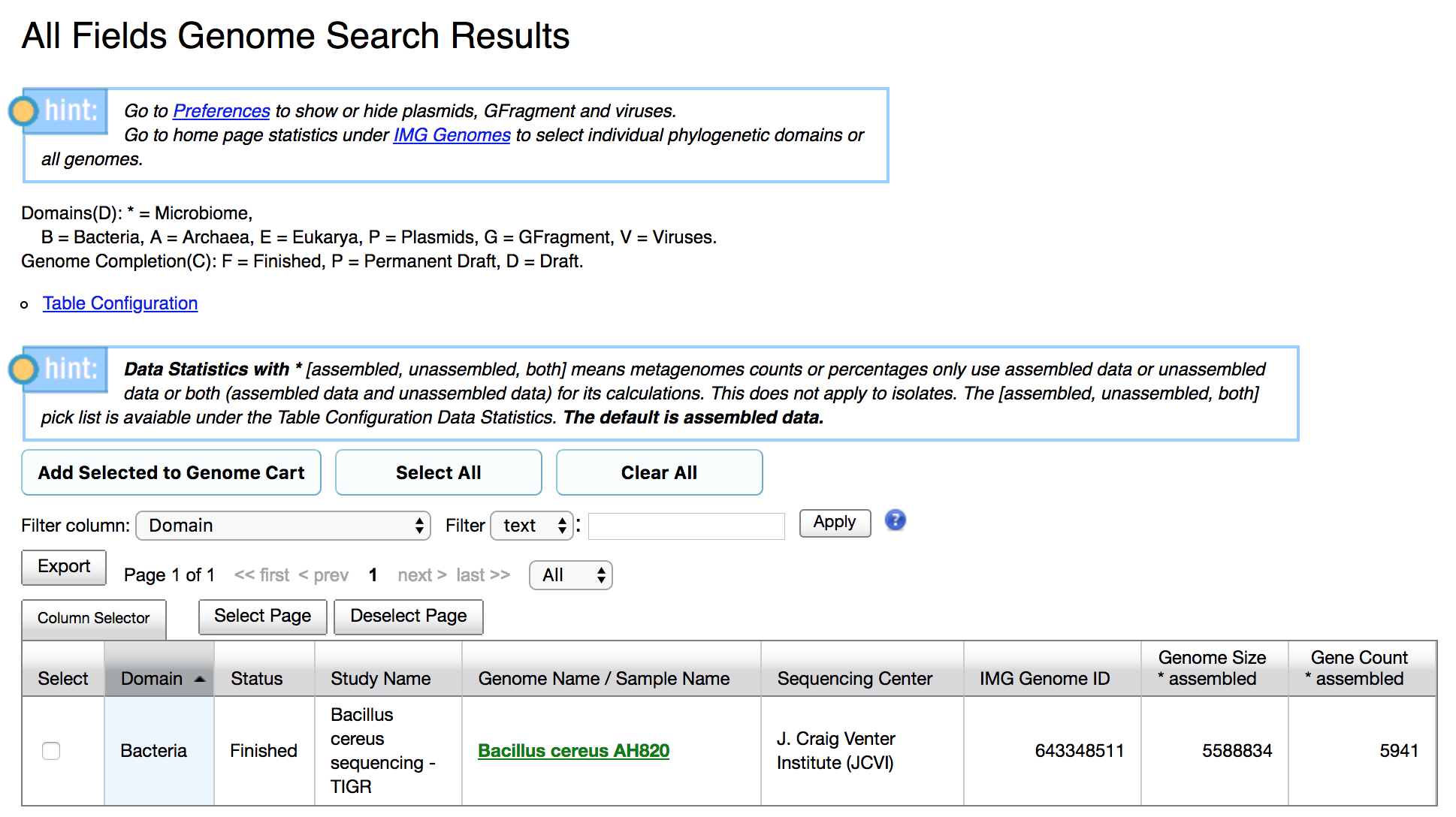
Start at the following site: <https://img.jgi.doe.gov/cgi-bin/m/main.cgi>



How many bacterial datasets are in this database?

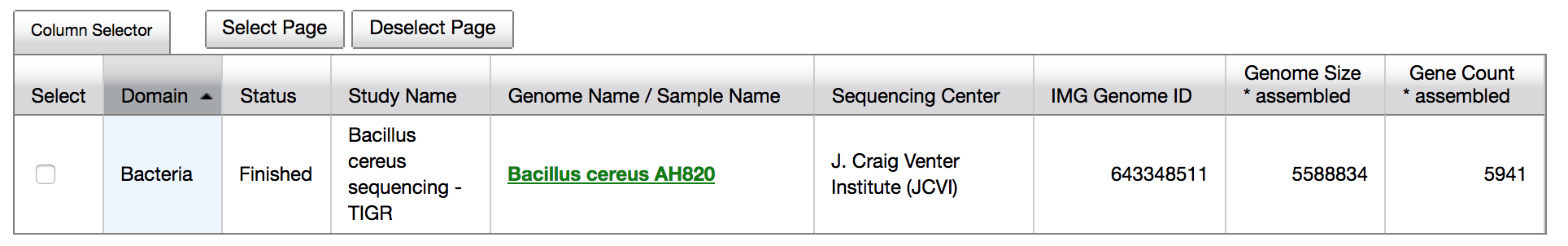
16753 exclusively in the JGI database. 83178 in total (As of August 2020)

2. Under the ‘Find Genomes’ tab, click on ‘Genome Search’. On the next page in the search bar type *Bacillus cereus* AH820 and press Search. It will open a page with search results – Click on the link under ‘Hits’



On the next page you will see three different kind of completion statuses of the genome. What are they? See red box above : Finished, Permanent Draft and Draft

3. What is the finishing level of the *Bacillus cereus* AH820 genome?



4. Click on *Bacillus cereus* AH820 Genome:

Look under ‘Overview’ for the following questions:

Where was the genome of this bacterium sequenced? (*Hint: look at the row titled ‘Sequencing Center’)*

J. Craig Venter Institute (JCVI)

When and where was the bacterium isolated? (*Hint: look at the row titled ‘Geographic Location’)*

|  |  |
| --- | --- |
| **Geographic Location** | October 1995 in Akershus Norway, from the periodontal pocket of a 76 year old female patient with marginal periodontitis |

What is the Phenotype of this organism? (*Hint: look at the row under ‘Phenotype’)*

|  |  |
| --- | --- |
| **Phenotype** | Pathogen |

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

Graphical user interface, application

Description automatically generated

How long is the genome in number of bps? (*Hint: look under the column titled ‘DNA, total number of bases’)*

|  |  |
| --- | --- |
| **DNA, total number of bases** | 5588834 |

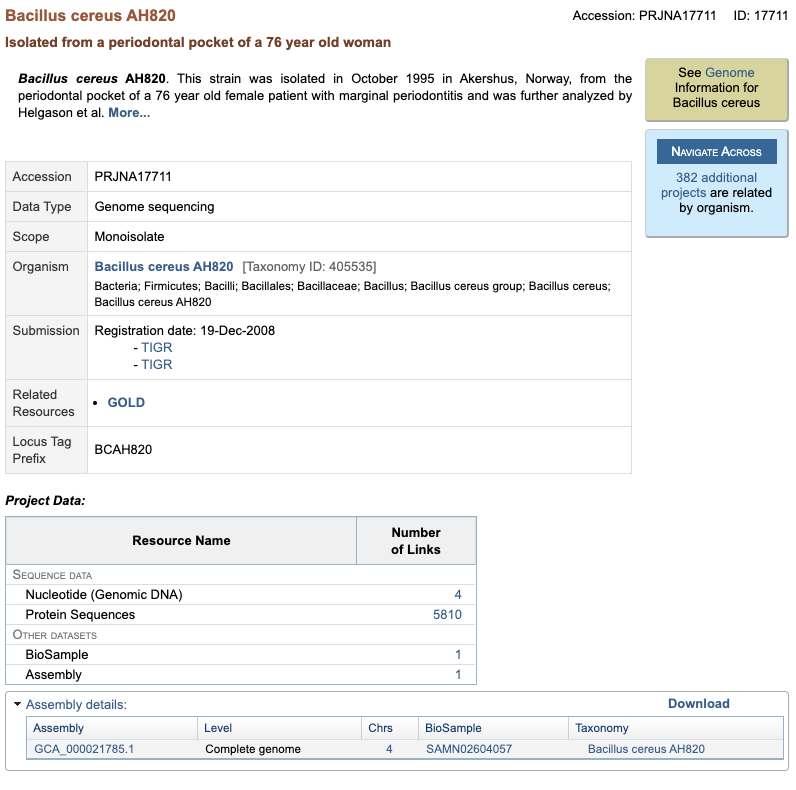
What is the GC% associated with this organism? (*Hint: look under the column titled ‘DNA, total number of bases’)*

|  |  |
| --- | --- |
| DNA G+C number of bases | 1973369 |

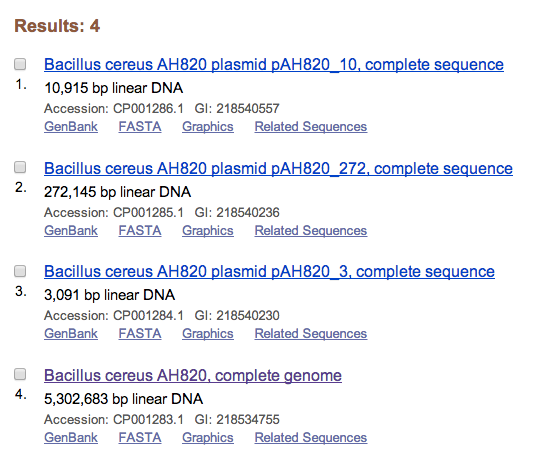
How many genes have been annotated as protein coding? (*Hint: look under the column titled ‘Gene, total number’)*

|  |  |
| --- | --- |
| **Genes total number** | 5941 |

5. Under ‘Overview’, look for the ‘Bioproject Accession’ and copy the accession number. Open a new tab and go to <http://www.ncbi.nlm.nih.gov/bioproject/> and paste the accession number in the search bar. You will be directed to a new page on the NCBI website.



Click on the number next to the ‘Nucleotide (Genomic DNA)’ column and you will be directed to a page all the sequences associated with this organism can be accessed.



Which scaffold is the longest? (*Give accession number)*

Accession # CP001286.1

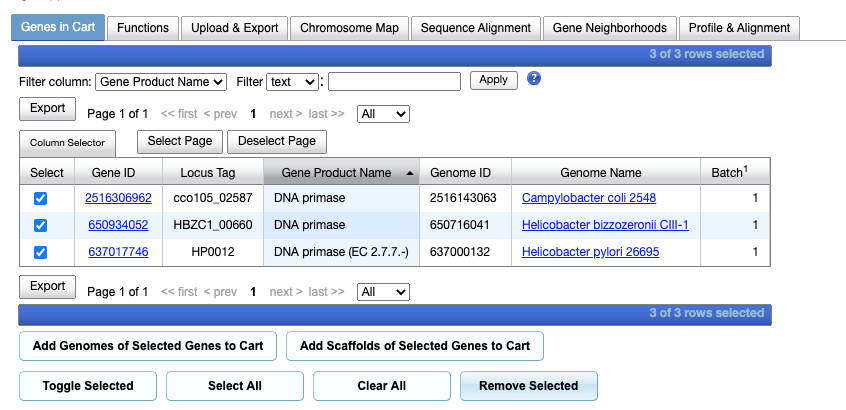
If you wanted to save the genome sequence for research or a classroom exercise you could click on the FASTA link and save the FASTA sequence of this genome as a .txt file

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

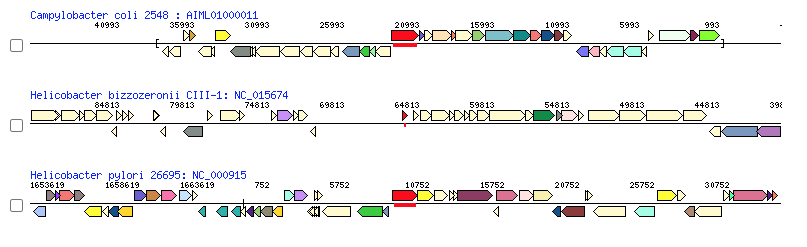
1. Start at the following site:

<https://img.jgi.doe.gov/cgi-bin/m/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, 2516306962 (each separated by commas) in the ‘Keyword’ box. On the next page click on the tab under ‘Gene Hits’



On the resulting page select the boxes associated with all the genes and click on ‘ Add selected Gene to Cart’ tab. On the following page click on the ‘Gene Neighborhoods’ tab and then select the forward direction (5’ – 3’ direction). You will see a graphical interface associated with your genes. Based on this interface answer the following questions:



What are the names of the organisms that are being compared in the graphical output?

*Helicobacter pylori, Helicobacter bizzozeronnii*  and *Campylobacter coli*

1. The gene that you are interested in is colored in red. How does the length of this gene in *H. pylori* compare to other organisms?

The gene associated with DNA primase is truncated in *H. bizzozeronnii*

1. Scroll over the gene in *H. pylori* and click on the region. 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:

What are some of the gene ontology terms associated with this gene (Hint: see the GO terms?

[GO:0003677](http://www.ebi.ac.uk/ego/DisplayGoTerm?id=GO:0003677) - DNA binding [evidence=[IEA](http://www.geneontology.org/GO.evidence.shtml)]  
[GO:0003896](http://www.ebi.ac.uk/ego/DisplayGoTerm?id=GO:0003896) - DNA primase activity [evidence=[IEA](http://www.geneontology.org/GO.evidence.shtml)]  
[GO:0003899](http://www.ebi.ac.uk/ego/DisplayGoTerm?id=GO:0003899) - DNA-directed RNA polymerase activity [evidence=[IEA](http://www.geneontology.org/GO.evidence.shtml)]  
[GO:0046872](http://www.ebi.ac.uk/ego/DisplayGoTerm?id=GO:0046872) - metal ion binding [evidence=[IEA](http://www.geneontology.org/GO.evidence.shtml)]  
[GO:0016779](http://www.ebi.ac.uk/ego/DisplayGoTerm?id=GO:0016779) - nucleotidyltransferase activity [evidence=[IEA](http://www.geneontology.org/GO.evidence.shtml)]  
[GO:0016740](http://www.ebi.ac.uk/ego/DisplayGoTerm?id=GO:0016740) - transferase activity [evidence=[IEA](http://www.geneontology.org/GO.evidence.shtml)]  
[GO:0008270](http://www.ebi.ac.uk/ego/DisplayGoTerm?id=GO:0008270) - zinc ion binding [evidence=[IEA](http://www.geneontology.org/GO.evidence.shtml)]

What are some of the protein families associated with this gene? (*Hint: see the information associated with Pfam)*

[fam01807](http://pfam.xfam.org/family/PF01807)- zf-CHC2   
[pfam08275](http://pfam.xfam.org/family/PF08275)- Toprim\_N   
[pfam13662](http://pfam.xfam.org/family/PF13662)- Toprim\_4   
[pfam16730](http://pfam.xfam.org/family/PF16730)- DnaGprimase\_HBD

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

What protein families are associated with this gene?

No information available

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 this genome in comparison with *H. pylori?*

*H. pylori*  seems well annotated. *H. bizzozeronii*  needs more evidence for functional domains (potential for research project perhaps?)