Bio 346 Bioinformatics Lab Guide 4:

***Gene Annotation and Misc Gene Detection***

James B. Herrick, Ph.D.

*James Madison University*

I’m going to list links and very general instructions for annotating and detecting genes. The instructions won’t be as detailed as you might like… sometimes because I may have done little with a particular tool yet. You don’t have to use all of these, or explore them in any order. Go where your interests take you. This is where the fun really begins!

# Annotation of Your Entire Genome

“Annotation” is the process of where genes (open reading frames or ORFs) might be in your genome and then attempting to assign a function to each of them based upon their similarity to other genes of known function (in various databases). Many genes won’t be identified (yet!). These are based on your assemblies. The main methods for microbial genomes are *Prokka*, *RAST*, and PGAP. The latter is a proprietary annotator from NCBI so we won’t be using it until we upload our assemblies to NCBI.

## Prokka

Prokka can be accessed using [Galaxy](https://usegalaxy.org/) (not GalaxyTrakr). You’ll have to sign up separately for Galaxy itself, and upload your files just as you have in GalaxyTrakr; the interface is the same.

Look for the Prokka tool under “NGS Assembly” (or just do a search under ‘Tools’). The default parameters should be sufficient, although you might consider excluding somewhat larger contigs than the 200 bp minimum in the default settings). Prokka gives various output files, such as table and text files for searching.

## Visualization

There are a number of “genome viewers” of various flavors that will allow you to visualize your annotated contigs. [Artemis](http://sanger-pathogens.github.io/Artemis/Artemis/) is an oldie but goodie. It has an ancient interface but is still being maintained by the Sanger Wellcome Institute. Open Artemis and upload your .gff file output by Prokka.

## RAST

[RAST](https://www.patricbrc.org/app/Annotation) is a web-based annotation server found on the [PATRIC website](https://www.patricbrc.org/). PATRIC has a number of other programs you might be interested in; however, the interface takes some getting used to.

# 

# Center for Genomic Epidemiology Programs

For all of these [CGE](http://www.genomicepidemiology.org/) programs, it pays to read the instructions, the explanation of the output files and, if possible, the paper upon which the analysis is based. These are normally at the top of the ‘home’ page for each analysis.

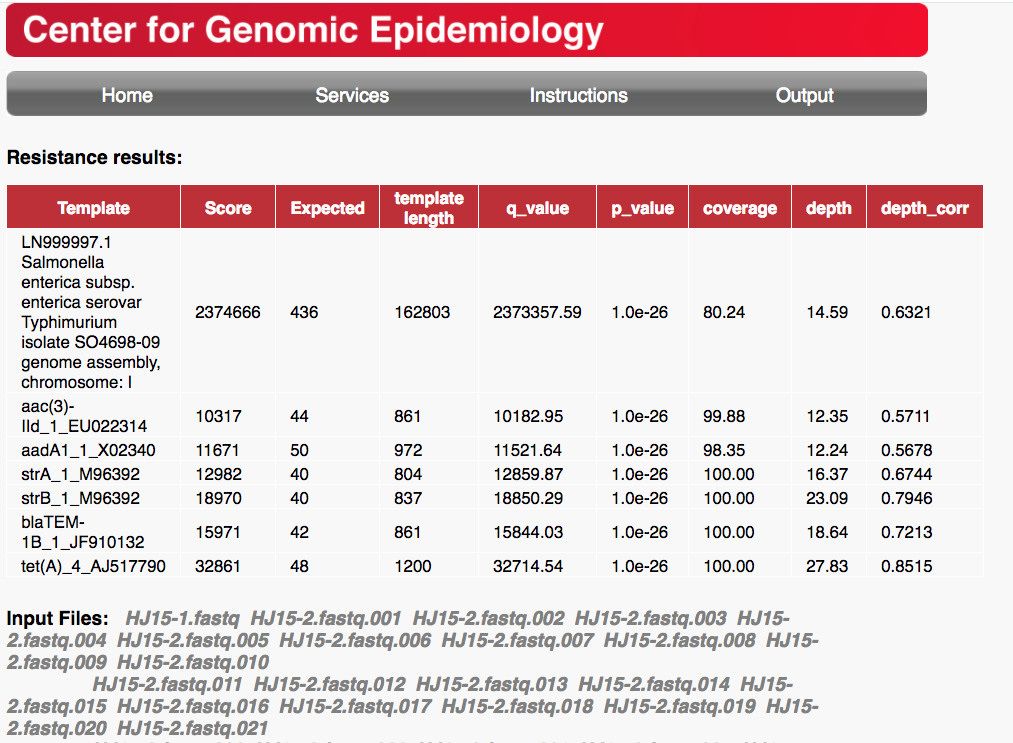
## [MLST](https://cge.cbs.dtu.dk/services/MLST-2.0/)

Although not really concerned with gene annotation oedetection *per se*, [MLST](http://www.mlst.net/) is a method to type organisms that, in the case of *Salmonella,* resolves at a somewhat finer level than serotyping. MLST is organism-specific (i.e. there is a different scheme, or set of genes used for each organism). This is the classic, 7-gene MLST.

## Output example (HJ1)

## [KmerResistance](https://cge.cbs.dtu.dk/services/KmerResistance/)

Maps the co-occurrence of k-mers between the WGS data and a database of resistance genes.

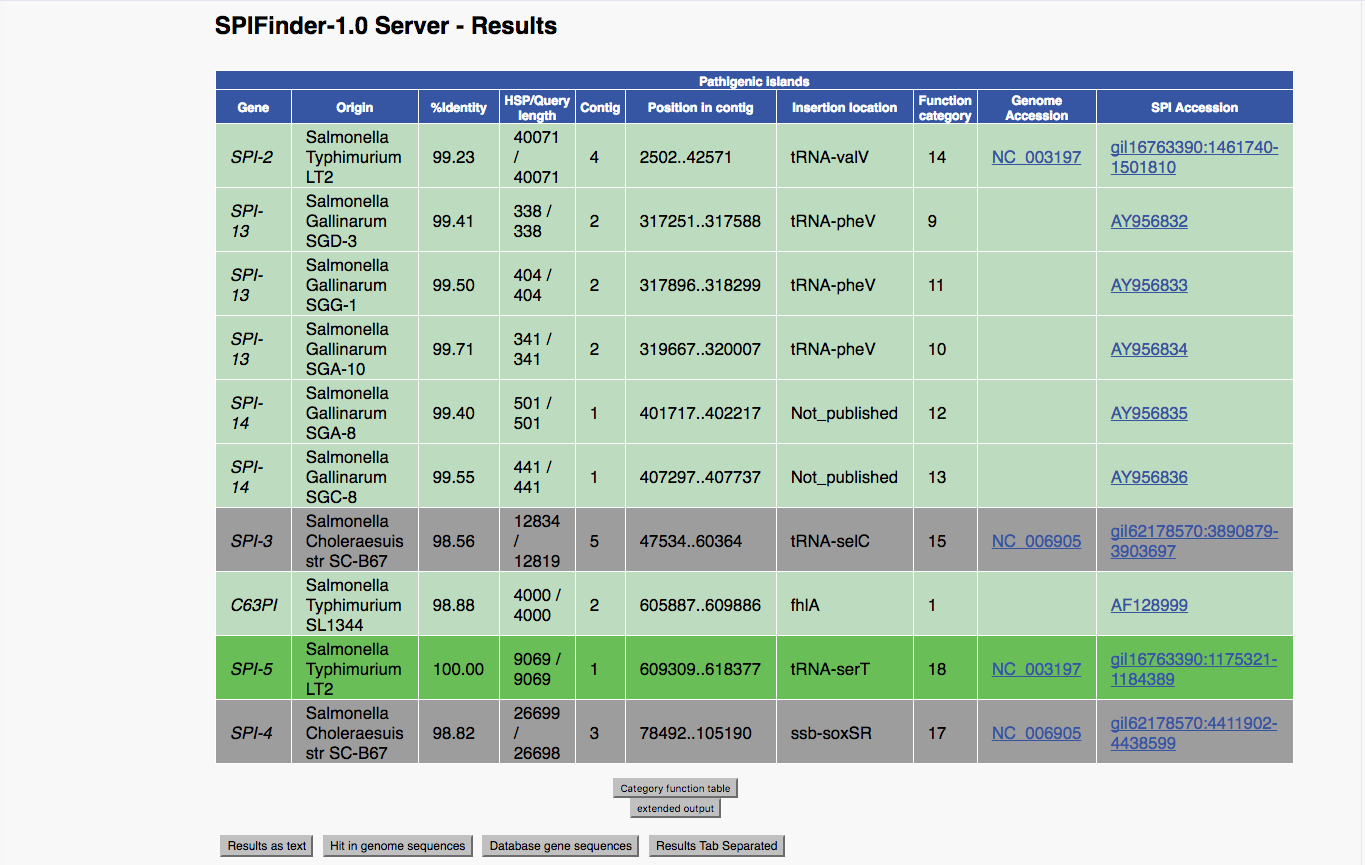


## [VirulenceFinder](https://cge.cbs.dtu.dk/services/VirulenceFinder/)

Identification of acquired virulence genes. Only available so far for *E. coli, S. aureus, Listeria,* and *Enterococcus*. Choose your assembly (fasta) as the input file.

## [SPIFinder](https://cge.cbs.dtu.dk/services/SPIFinder/)

Identifies “Salmonella Pathogenicity Islands”. Choose your assembly (fasta) as the input file.



## Pathogenicity Island Functional Categories



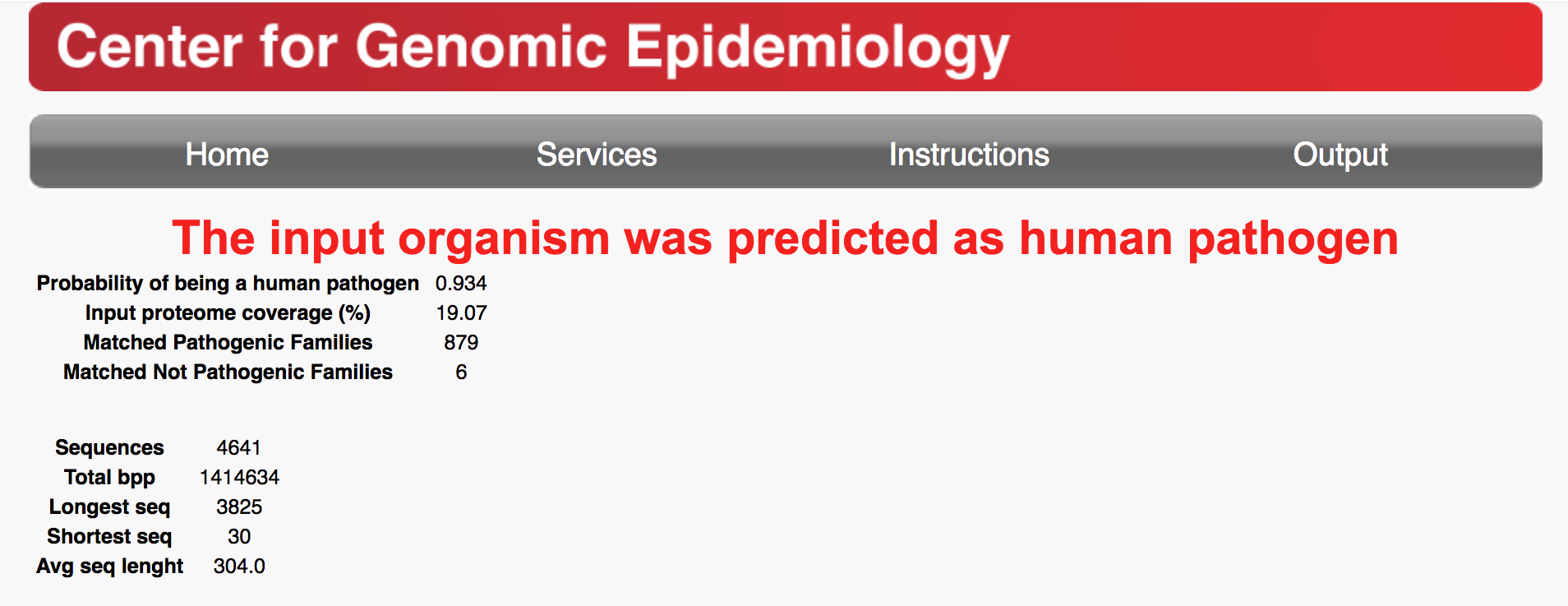
## 

## [SeroTypeFinder](https://cge.cbs.dtu.dk/services/SerotypeFinder/)

Prediction of serotypes in total or partial sequenced isolates of *E. coli*. (Not an annotation tool but listed here because it’s potentially useful for typing *E. coli*.). Again, use your assembly (fasta) as the input file.

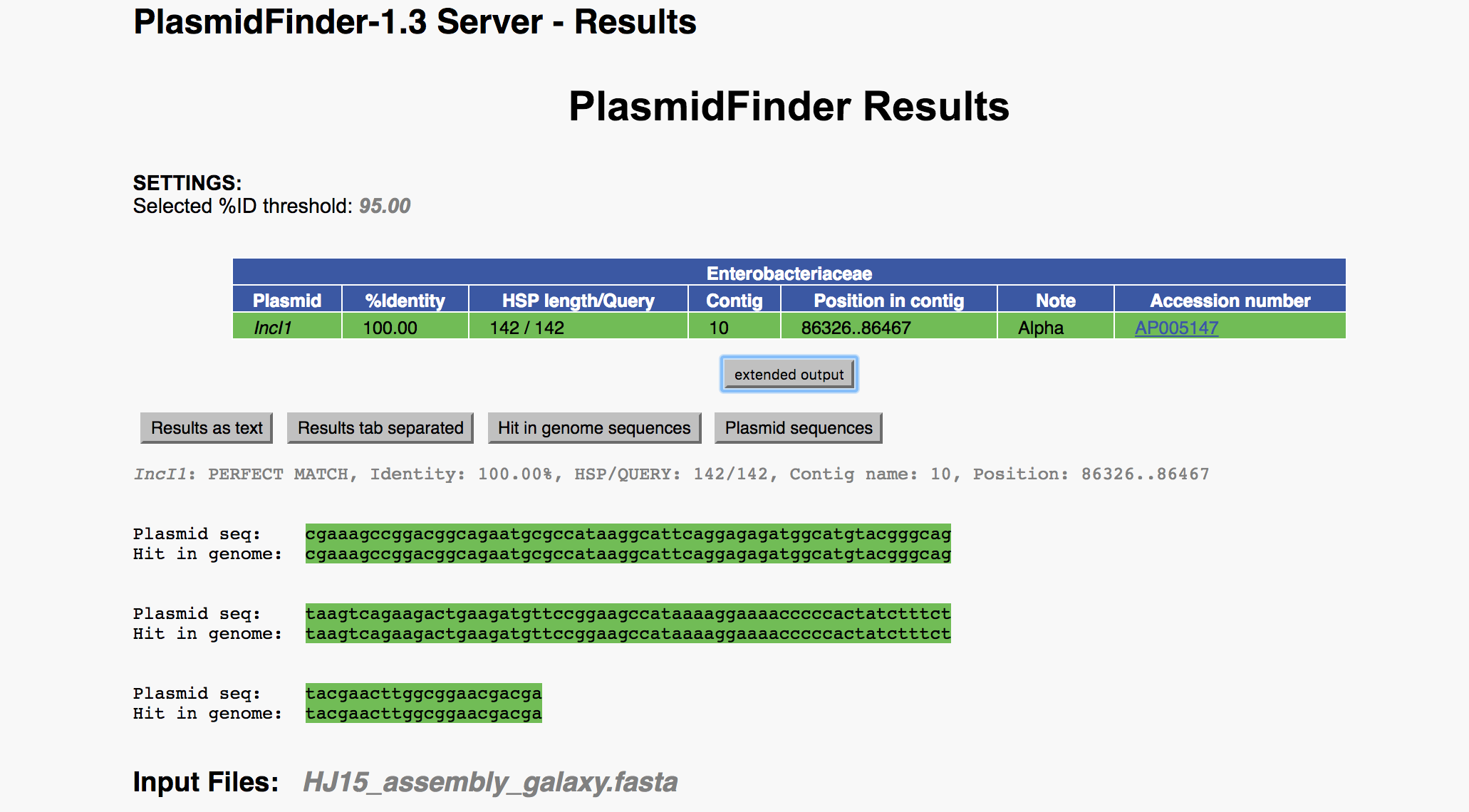
## [PathogenFinder](https://cge.cbs.dtu.dk/services/PathogenFinder/) (HJ1)

Prediction of a bacterium’s pathogenicity towards human hosts. Choose *γ-proteobacteria* under ‘phylum or class’ and *assembled genome* under ‘Sequencing Platform’. Then use your assembly (fasta) as the input file.



## [PlasmidFinder](https://cge.cbs.dtu.dk/services/PlasmidFinder/)

Identifies plasmids in total or partial sequenced isolates of bacteria. It is very accurate, with few false positives (although it does suffer from false negatives). Select the *Enterobacteriace* database. You can experiment with the thresholds. Use your assembly as the input file. “The filename must not contain spaces.

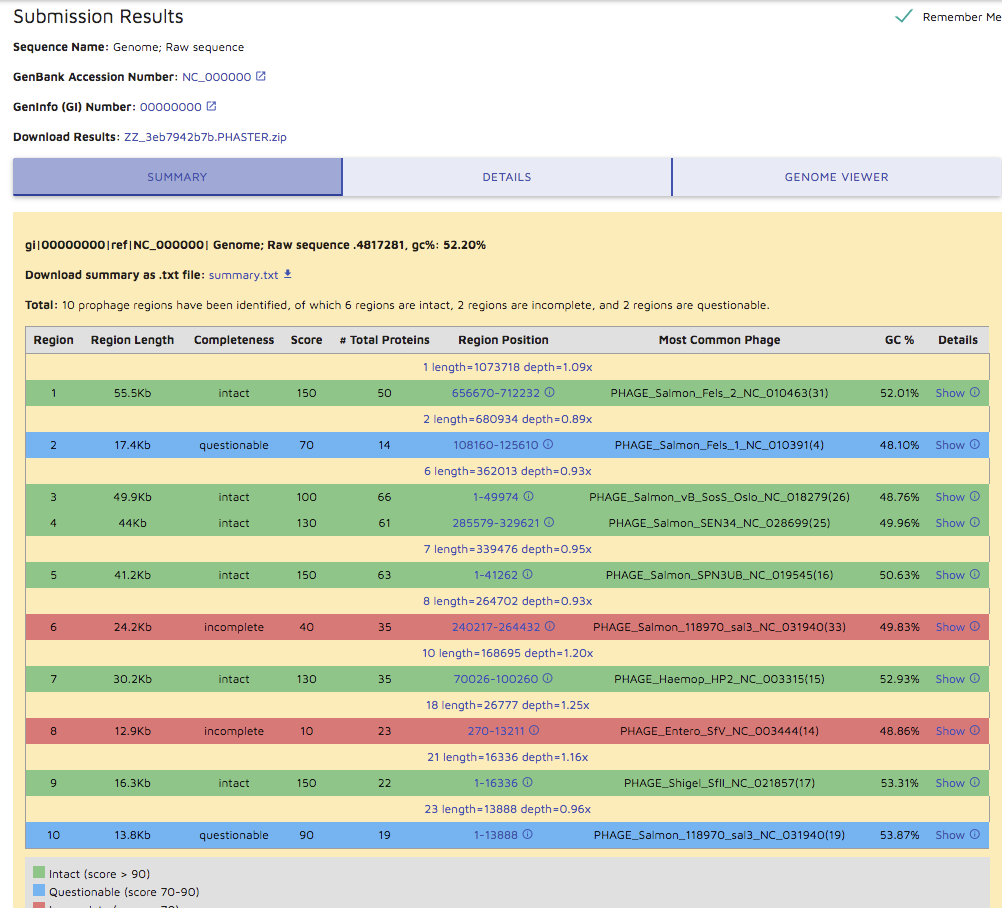


# Other Programs

## [PHASTER](http://phaster.ca/submissions/ZZ_ef8dd57185?batch_id=BB_9e9c4ae92f)

Rapid identification and annotation of **prophage** sequences within bacterial genomes and plasmids

* There is a short tutorial video on the home page that I recommend you watch. The *Help* tab at the top of the page also has very good instructions on using the site.
* Upload your fasta assembly file. Check “My FASTA file consists of metagenomic contigs”
* This will take a little while (minutes to hours), depending upon how many are ahead of you in the queue.
* (Be sure and check out the ‘Genome Viewer’ output.)



## [INTEGRALL](http://integrall.bio.ua.pt/?intro)

identification of **integrons**, which are are genetic systems that allow bacteria to capture and express antibiotic resistance gene cassettes.

* For this site, you actually have to cut and paste your assembled fasta file into the [“BLAST” page](http://integrall.bio.ua.pt/?search#). Open your file in a text viewer (such as TextEdit on the Mac, or the free viewer SublimeText), if possible. Converting it to text in Word might work. Select the entire file and copy it.
* Click on ‘BLAST/Search’ on the INTEGRALL site and paste your file. Then hit ‘search’.

**Note:** cutting and pasting these huge files will tax your computer’s clipboard and CPU. Be patient. I found that Safari worked better than Chrome on my Mac.

* The results will appear in the same window. If nothing appears, apprently no integrons were found. (I know, it’s kind of silly to me, too.)

## [Virulence Factors of Pathogenic Bacteria](http://www.mgc.ac.cn/cgi-bin/VFs/jsif/main.cgi) (website)

This is a database, with useful information on the virulence factors -- including pathogenicity islands -- of *Salmonella* and other bacterial pathogens. It will help you understand some of your results obtained above. (Note: last time I checked, this site wouldn’t load for me. It might be down or even non-functional now.)