Glimmer and Other Command Line Tools

According to the JHU website, “**Glimmer** is a system for finding genes in microbial DNA, especially the genomes of bacteria, archaea, and viruses. Glimmer (Gene Locator and Interpolated Markov ModelER) uses interpolated Markov models (IMMs) to identify the coding regions and distinguish them from noncoding DNA”. (<http://ccb.jhu.edu/software/glimmer/index.shtml>)

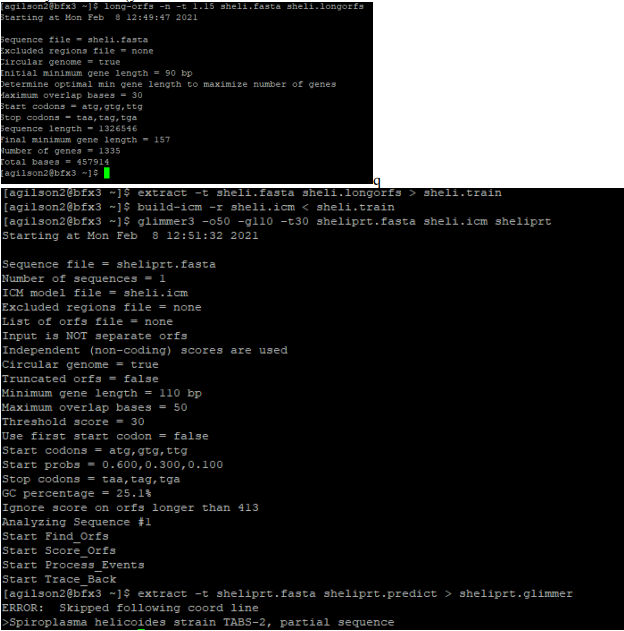
Glimmer was accessed on the JHU network’s BFX3 server, which housed not only glimmer, but other prediction programs and tools such as the SRA toolkit(also included in this pdf).

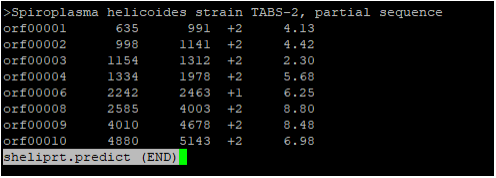
**BEDTools** was also utilized in the BFX3 server environment to analyze intersecting data points between 2 datasets. BEDtools is described as having many different functions in the aid of analysis for BED files, including intersecting, merging, counting, complementing, etc. etc. (<https://bedtools.readthedocs.io/en/latest/>)

**SRAToolkit** is a set of tools that can be used to analyze sequence read archive(SRA) data, which is normally not formatted well as it functions mainly as a repository. It acts as a means of downloading reference sequences off of the SRA database, and manipulate that data in a manner that allows for easier understanding.

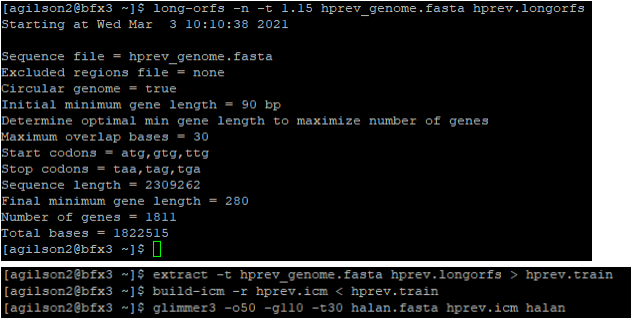
Glimmer was used in the Bioinformatics: Tools for Genome Analysis course to approximate ORF locations in 2 different examples of bacterial species. See the following two examples:

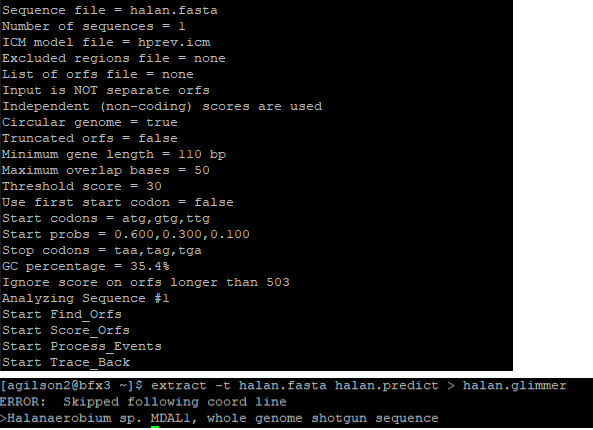
1. The following lines of code were used to analyze CDS sequences obtained from *Spiroplasma heliocoides* strain TABS-2. A training set was first created from the fasta file, which was then used in combination with Glimmer to predict the ORF’s of the partial CDS available.

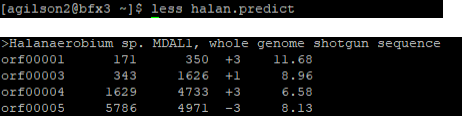
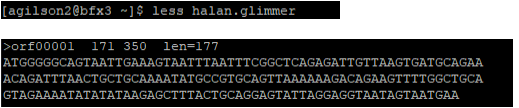


And the following ORF predictions were created. Confirmation of these predictions was made using FGENESB, see the “Prediction Software” section.

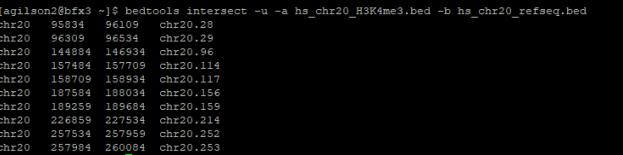
1. A similar method was used to examine contigs from an unknown strain of *Halanaerobium*, with a training set being created from the whole genome of *Halanaerobium praevalens*. This method was used once again to determine ORF sites from the FASTA file, and to also obtain the DNA sequence of those ORFs. These were once again confirmed using FGENESB.





The ORF predictions were placed into a .predict file in the same manner as the first example.  

BEDtools was utilized in the Bioinformatics: Tools for Genome Analysis course to analyze potential areas in which two BED files, one being the reference genome and the other containing H3K4me3 methylation state information, were intersected to determine methylation states in relation to coding exons.

1. 

10 regions were found to intersect a coding exon, which could imply that these exons are in fact methylated and in turn could be transcriptionally less active than those which are not methylated.



245 regions of the H3K4me3 file were found to NOT intersect with coding exons in the reference.

The SRAToolkit was utilized during my course, Next Generation Sequencing and Data Analysis, in which the toolkit was used to analyze SRR6956031 to perform a FASTQC quality score check, *de novo* assembly of the sequence reads, and blastn that assembly to identify the most closely related genome. The following code produced the result that SRR6956031 was most closely related to *Stenotrophomonas pavanii*. Many of the functions were available to me in the “data” folder available in JHU’s BFX3 server.

[agilson2@bfx3 ~]$ prefetch SRR6956031

[agilson2@bfx3 ~]$ fastq-dump -F --split-files --gzip SRR6956031

[agilson2@bfx3 ~]$ cd 410.666

[agilson2@bfx3 410.666]$ cd exam

[agilson2@bfx3 exam]$ vdb-config –interactive

[x]

[agilson2@bfx3 exam]$ prefetch SRR6956031

[agilson2@bfx3 exam]$ fastq-dump -F –split-files –gzip SRR6956031

\*Generated SRR6956031\_1.fastq.gz & SRR6956031\_2.fastq.gz, and

A file folder for SRR6956031

Read 1005748 spots for SRR6956031

Written 1005748 spots for SRR6956031

[agilson2@bfx3 exam]$ md5sum SRR6956031\_1.fastq.gz

b17a9131e0c19074785bdae353b9c82f SRR6956031\_1.fastq.gz

[agilson2@bfx3 exam]$ md5sum SRR6956031\_2.fastq.gz

b7fc7702a4213cbc7dc352a9c04cd527 SRR6956031\_2.fastq.gz

[agilson2@bfx3 exam]$ fastqc SRR6956031\_1.fastq.gz

[agilson2@bfx3 exam]$ fastqc SRR6956031\_2.fastq.gz

[agilson2@bfx3 exam]$ mkdir -p ~/public\_html

[agilson2@bfx3 exam]$ chmod -R 755 ~/public\_html

[agilson2@bfx3 exam]$ cp SRR6956031\_1\_fastqc.html ~/public\_html/

[agilson2@bfx3 exam]$ cp SRR6956031\_2\_fastqc.html ~/public\_html/

#1: <http://bfx3.aap.jhu.edu/~agilson2/SRR6956031_1_fastqc.html>

#2: <http://bfx3.aap.jhu.edu/~agilson2/SRR6956031_2_fastqc.html>

[agilson2@bfx3 exam]$ trimmomatic PE -phred33 SRR6956031\_1.fastq.gz SRR6956031\_2.fastq.gz SRR6956031\_1trim.fastq.gz SRR6956031\_1unpaired.fastq.gz SRR6956031\_2trim.fastq.gz SRR6956031\_2unpaired.gz LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:100

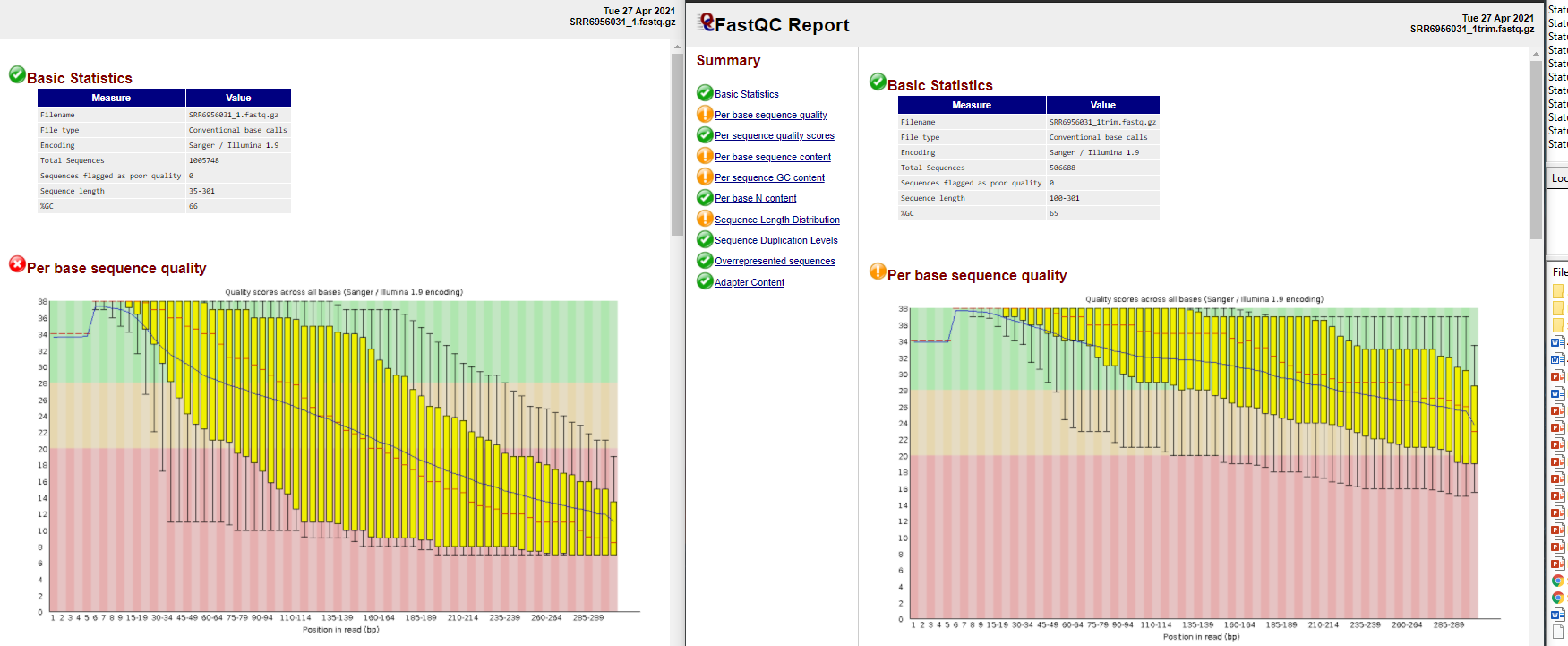
[agilson2@bfx3 exam]$ fastqc SRR6956031\_1trim.fastq.gz

[agilson2@bfx3 exam]$ fastqc SRR6956031\_2trim.fastq.gz

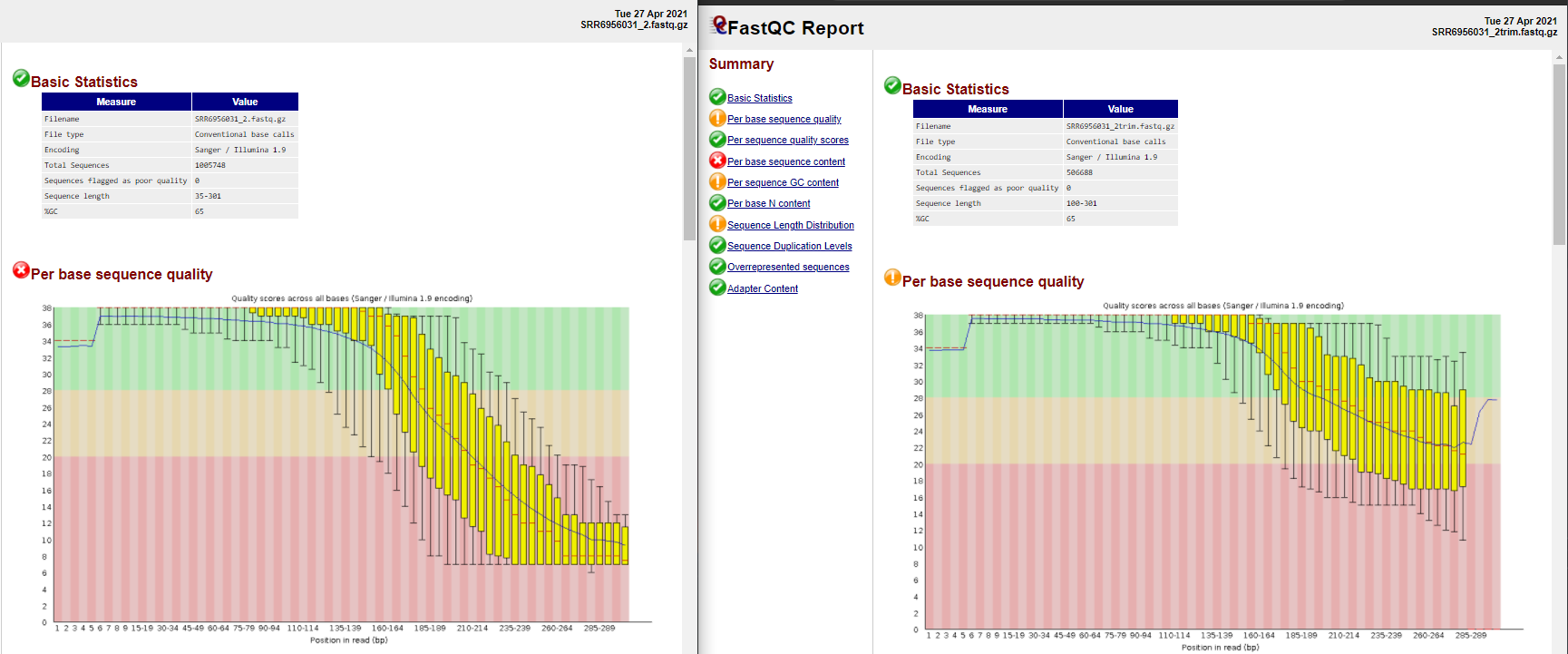
[agilson2@bfx3 exam]$ cp SRR6956031\_1trim\_fastqc.html ~/public\_html/

[agilson2@bfx3 exam]$ cp SRR6956031\_2trim\_fastqc.html ~/public\_html/

#1: <http://bfx3.aap.jhu.edu/~agilson2/SRR6956031_1trim_fastqc.html>

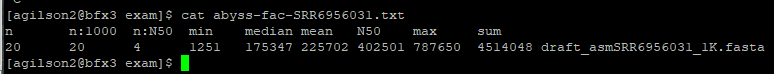


#2: <http://bfx3.aap.jhu.edu/~agilson2/SRR6956031_2trim_fastqc.html>



[agilson2@bfx3 exam]$ --careful -m 20 -t 8 -k 21,33,55,77 -o asmSRR6956031\_spades

[agilson2@bfx3 exam]$ /opt/410.666/data/scripts/sort\_contigs.pl -b -m 1000 -p -z asmSRR6956031\_spades/scaffolds.fasta draft\_asmSRR6956031\_1K.fasta

[agilson2@bfx3 exam]$ /opt/410.666/data/scripts/abyss-fac-all -t 1000 draft\_asmSRR6956031\_1K.fasta > abyss-fac-SRR6956031.txt [agilson2@bfx3 exam]$ makeblastdb -dbtype nucl -parse\_seqids -in draft\_asmSRR6956031\_1K.fasta

[agilson2@bfx3 exam]$ head -n 1 draft\_asmSRR6956031\_1K.fasta

[agilson2@bfx3 exam]$ blastdbcmd -entry NODE\_1\_length\_787650\_cov\_22.222449 -db draft\_asmSRR6956031\_1K.fasta -out NODE\_1\_length\_787650\_cov\_22.222449.fasta

[agilson2@bfx3 exam]$ blastn -db /opt/410.666/data/blastdb/ref\_prok\_rep\_genomes -query NODE\_1\_length\_787650\_cov\_22.222449.fasta -out exam\_part\_5.out

Top BLASTN hit: ***Stenotrophomonas pavanii***