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.

```
[agilson2@bfx3 ~]$ long-orfs -n -t 1.15 sheli.fasta sheli.longorfs
Starting at Mon Feb  8 12:49:47 2021

Sequence file = sheli.fasta
Excluded regions file = none
Circular genome = true
Initial minimum gene length = 90 bp
Determine optimal min gene length to maximize number of genes
Maximum overlap bases = 30
Start codons = atg,gtg,ttg
Stop codons = taa,tag,tga
Sequence length = 1326546
Final minimum gene length = 157
Number of genes = 1335
Total bases = 457914
[agilson2@bfx3 ~]$
```

```
[agilson2@bfx3 ~]$ extract -t sheli.fasta sheli.longorfs > sheli.train
[agilson2@bfx3 ~]$ build-icm -r sheli.icm < sheli.train
[agilson2@bfx3 ~]$ glimmer3 -o50 -g110 -t30 sheliprt.fasta sheli.icm sheliprt
Starting at Mon Feb  8 12:51:32 2021

Sequence file = sheliprt.fasta
Number of sequences = 1
ICM model file = sheli.icm
Excluded regions file = none
List of orfs file = none
Input is NOT separate orfs
Independent (non-coding) scores are used
Circular genome = true
Truncated orfs = false
Minimum gene length = 110 bp
Maximum overlap bases = 50
Threshold score = 30
Use first start codon = false
Start codons = atg,gtg,ttg
Start probs = 0.600,0.300,0.100
Stop codons = taa,tag,tga
GC percentage = 25.1%
Ignore score on orfs longer than 413
Analyzing Sequence #1
Start Find_Orfs
Start Score_Orfs
Start Process_Events
Start Trace_Back
[agilson2@bfx3 ~]$ extract -t sheliprt.fasta sheliprt.predict > sheliprt.glimmer
ERROR:  Skipped following coord line
>Spiroplasma helicoides strain TABS-2, partial sequence
```

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

```
>Spiroplasma helicoides strain TABS-2, partial sequence
orf00001      635      991  +2     4.13
orf00002      998     1141  +2     4.42
orf00003     1154     1312  +2     2.30
orf00004     1334     1978  +2     5.68
orf00006     2242     2463  +1     6.25
orf00008     2585     4003  +2     8.80
orf00009     4010     4678  +2     8.48
orf00010     4880     5143  +2     6.98
sheliprt.predict (END)
```

2. 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.

```
[agilson2@bfx3 ~]$ long-orfs -n -t 1.15 hprev_genome.fasta hprev.longorfs
Starting at Wed Mar  3 10:10:38 2021

Sequence file = hprev_genome.fasta
Excluded regions file = none
Circular genome = true
Initial minimum gene length = 90 bp
Determine optimal min gene length to maximize number of genes
Maximum overlap bases = 30
Start codons = atg,gtg,ttg
Stop codons = taa,tag,tga
Sequence length = 2309262
Final minimum gene length = 280
Number of genes = 1811
Total bases = 1822515
[agilson2@bfx3 ~]$
```

```
[agilson2@bfx3 ~]$ extract -t hprev_genome.fasta hprev.longorfs > hprev.train
[agilson2@bfx3 ~]$ build-icm -r hprev.icm < hprev.train
[agilson2@bfx3 ~]$ glimmer3 -o50 -g110 -t30 halan.fasta hprev.icm halan
```

```
Sequence file = halan.fasta
Number of sequences = 1
ICM model file = hprev.icm
Excluded regions file = none
List of orfs file = none
Input is NOT separate orfs
Independent (non-coding) scores are used
Circular genome = true
Truncated orfs = false
Minimum gene length = 110 bp
Maximum overlap bases = 50
Threshold score = 30
Use first start codon = false
Start codons = atg,gtg,ttg
Start probs = 0.600,0.300,0.100
Stop codons = taa,tag,tga
GC percentage = 35.4%
Ignore score on orfs longer than 503
Analyzing Sequence #1
Start Find_Orfs
Start Score_Orfs
Start Process_Events
Start Trace_Back
```

```
[agilson2@bfx3 ~]$ extract -t halan.fasta halan.predict > halan.glimmer
ERROR:  Skipped following coord line
>Halanaerobium sp. MDAL1, whole genome shotgun sequence
```

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

```
[agilson2@bfx3 ~]$ less halan.predict

>Halanaerobium sp. MDAL1, whole genome shotgun sequence
orf00001      171      350  +3    11.68
orf00003      343     1626  +1     8.96
orf00004     1629     4733  +3     6.58
orf00005     5786     4971  -3     8.13
```

```
[agilson2@bfx3 ~]$ less halan.glimmer

>orf00001  171 350  len=177
ATGGGGGCAGTAATTGAAAGTAATTTAATTTCGGCTCAGAGATTGTTAAGTGATGCAGAA
ACAGATTTAACTGCTGCAAAATATGCCGTGCAGTTAAAAAAGACAGAAGTTTTGGCTGCA
GTAGAAAATATATATAAGAGCTTTACTGCAGGAGTATTAGGAGGTAATAGTAATGAA
```

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.

3.
```
[agilson2@bfx3 ~]$ bedtools intersect -u -a hs_chr20_H3K4me3.bed -b hs_chr20_refseq.bed
chr20   95834   96109   chr20.28
chr20   96309   96534   chr20.29
chr20   144884  146934  chr20.96
chr20   157484  157709  chr20.114
chr20   158709  158934  chr20.117
chr20   187584  188034  chr20.156
chr20   189259  189684  chr20.159
chr20   226859  227534  chr20.214
chr20   257534  257959  chr20.252
chr20   257984  260084  chr20.253
```

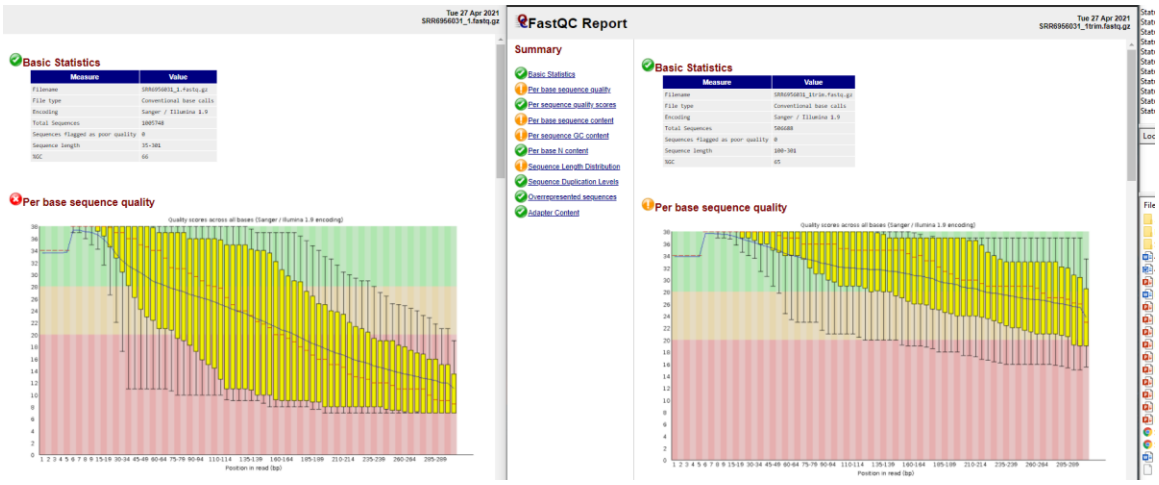
10 regions were found to intersect a coding exon, which could imply that these exons are in fact methylated and in turn could be transcriptionally less active than those which are not methylated.

```
[agilson2@bfx3 ~]$ bedtools intersect -v  -a hs_chr20_H3K4me3.bed -b hs_chr20_refseq.bed | wc -l
245
```

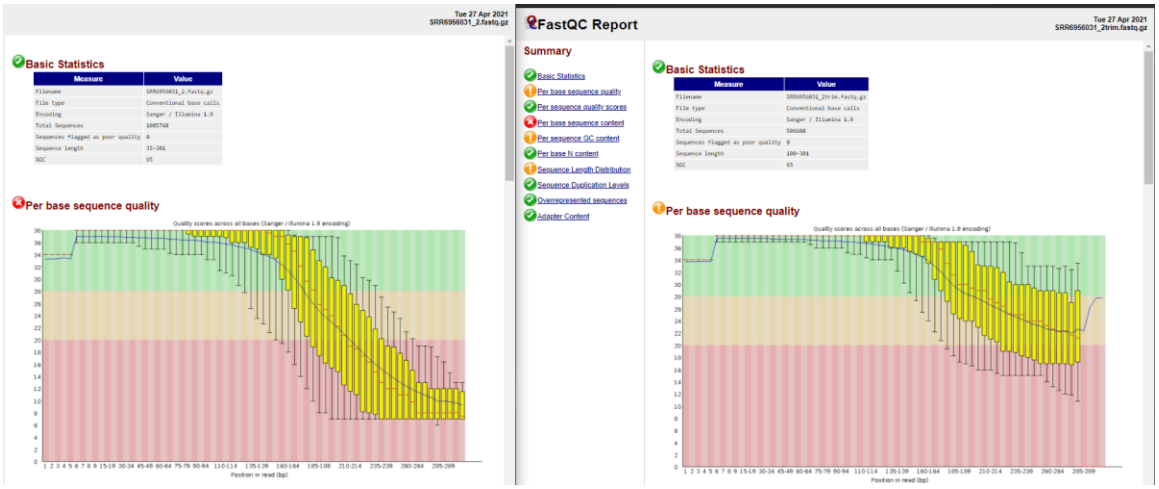
245 regions of the H3K4me3 file were found to NOT intersect with coding exons in the reference.

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]$ cat abyss-fac-SRR6956031.txt
n       n:1000  n:N50   min     median mean    N50     max     sum
20      20      4       1251    175347 225702 402501 787650  4514048 draft_asmSRR6956031_1K.fasta
[agilson2@bfx3 exam]$
```

[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***