**BIOMI 609 Computational Genomics and Bioinformatics**

**Spring 2022**

**Dr. Arun Sethuraman**

**Final Exam Review - Lab**

**Part 1: Using Unix-based scripts**

Recall that Unix commands that are predominantly used include:

ls - list contents of a directory

cd - change directory

mv - cut + paste

cp - copy + paste

mkdir - make a new directory

rm - delete a file/directory

wget - download a file from the web

grep - search for the occurrence of a string

wc - word count

vim - text editor inside Unix

export PATH=$PATH:/path/to/set - sets the path to a directory

clear - clears screen

pwd - path to current working directory

chmod +x script.sh - sets executable permission to script.sh

cat - print the contents of a file to the screen

head - display the first few lines of a file

tail - display the last few lines of a file

**Part 2: File types**

FASTQ - contains reads with their PHRED quality scores; paired reads are usually indicated with filename\_1.fastq and filename\_2.fastq

FASTA - contains either a single sequence or multiple sequence, including aligned sequences

SAM/BAM - alignment to a reference genome

VCF - variant call format

BED - binary variant call format

NEWICK - phylogenetic tree format

Other common formats:

TXT - text files

CSV - comma separated values - often used for tables

TSV - tab separated values - often used for tables

**Part 3: Quality Control, trimming reads as needed**

Make a new folder, call it FinalReview, then cd into that folder.

In this folder, go ahead and download these two FASTQ files - these are from E. coli, in case you’d like to run steps with respect to a reference genome (e.g. QUAST, reference-guided annotation using AUGUSTUS).

curl -O -J -L https://osf.io/shqpv/download

curl -O -J -L https://osf.io/9m3ch/download

Go ahead and download these two read FASTQ files (alternately, you can download them from Canvas under Final Review), then run fastqc using:

fastqc SRR957824\_500K\_R1.fastq.gz SRR957824\_500K\_R2.fastq.gz

Now observe the HTML files output by FASTQC; what do you observe? How would you trim these files? Thereon, go ahead and use trimmomatic to trim them based on your observation.

trimmomatic PE -threads 6 SRR957824\_500K\_R1.fastq.gz SRR957824\_500K\_R2.fastq.gz SRR957824\_500K\_R1\_trimmed.fastq.gz SRR957824\_500K\_R2\_trimmed.fastq.gz SRR957824\_500K\_R1\_untrimmed.fastq.gz SRR957824\_500K\_R2\_untrimmed.fastq.gz TRAILING:30

Alternately, if you’re using the jar version of trimmomatic, your command will look like this:

java -jar trimmomatic-0.39.jar PE SRR957824\_500K\_R1.fastq.gz SRR957824\_500K\_R2.fastq.gz SRR957824\_500K\_R1\_trimmed.fastq.gz SRR957824\_500K\_R2\_trimmed.fastq.gz SRR957824\_500K\_R1\_untrimmed.fastq.gz SRR957824\_500K\_R2\_untrimmed.fastq.gz TRAILING:30

**Part 4: Assembling reads, Quality assessment using QUAST**

Using Velvet:

First run velveth runk <kmer size> -shortPaired -separate -fastq SRR957824\_500K\_R1.fastq.gz SRR957824\_500K\_R2.fastq.gz

Thereon velvetg runk

Using SPAdes:

spades.py -k <kmersize> --pe1-1 SRR957824\_500K\_R1.fastq.gz --pe1-2 SRR957824\_500K\_R2.fastq.gz -o spadesrun

Using Abyss:

abyss-pe name=example k=<kmersize> in=” SRR957824\_500K\_R1.fastq.gz SRR957824\_500K\_R2.fastq.gz”

To run QUAST:

quast.py assembly1.fasta assembly2.fasta

Then observe the results files produced by QUAST.

**Part 5: Annotation using AUGUSTUS, creating a JBrowse instance**

These are really easy to do - so I’ll refer these to the last lab manual where I have detailed instructions on running AUGUSTUS and JBrowse using The Galaxy Project.

**Part 6: Multiple Sequence Alignment and Phylogenetic reconstruction**

Now copy a part of the assembled FASTA file, go to BLAST, search for less similar sequences using BLAST, download a FASTA file of the top few hits. Thereon, use MAFFT to align these sequences:

mafft hits.fasta > hits\_aligned.fasta

Thereon, you can build a ML phylogeny using:

raxmlHPC -s hits\_aligned.fasta -m GTRGAMMA -p 12345 -n hits\_ecoli

Visualize this using FigTree or any other tool e.g. <http://etetoolkit.org/treeview/>

**Part 7: Working with VCF files, population genomics**

Please refer to the manual for these - you’ve done plenty of examples of these recently (e.g. estimating pi, plotting it, doing ADMIXTURE runs, plotting ancestry proportions, etc).