**BIOMI 609 Computational Genomics and Bioinformatics**

**Spring 2022**

**San Diego State University**

**Dr. Arun Sethuraman**

**Comprehensive Final Exam**

**Due via Canvas on May 12th, 2022 (Thursday) at 11:59 PM**

1. Here is an SRA link to a whole genome sequencing run of the bacterium *Treponema pallidum subsp. pallidum* which causes endemic syphilis - <https://sra-download.ncbi.nlm.nih.gov/traces/sra59/SRR/017897/SRR18326765>

You will download this SRA link, (1) convert it into FASTQ using SRA toolkit, (2) assess quality of reads using FASTQC, and trim reads (if needed) (3) then assemble the reads using SPAdes, (4) assess assembly quality using QUAST against the CDC’s *T. pallidum subsp. pallidum* reference genome (<https://www.ncbi.nlm.nih.gov/data-hub/assembly/GCF_001655275.1/>) (5) perform *ab initio* annotation of the genome using AUGUSTUS using *E. coli* as a training model, (6) display the annotated genome as a genome browser using JBrowse.

Submit your code for 1-4, along with the output of FASTQC, final assembly (as a FASTA file), QUAST output (as a PDF or HTML), the GFF3/GTF output of AUGUSTUS, and the HTML file from JBrowse.

(10+10+10+10+10 = 50 points)

2. From the annotated genome, extract one coding sequence. Now use BLAST to obtain homologous “hits” against this query (make sure to exclude *Treponema pallidum* - otherwise all your hits will be from the same species), obtain 100 of these sequences, perform a multiple sequence alignment of these genes, and construct a maximum likelihood phylogeny using RAxML. Visualize this phylogeny using FigTree, and interpret the tree - how did Treponema as a genus evolve? What are its closest neighboring species? Are they also pathogenic? Explain your results. (25 points).

3. Read Beale et al. (2021): [10.1038/s41564-021-01000-z](https://dx.doi.org/10.1038%2Fs41564-021-01000-z) on the global evolution of Treponema pallidum. You will now utilize the variant file produced from this study (I’ve uploaded this to your Canvas page - tpa.vcf). Use ADMIXTURE to estimate population structure from this dataset. Determine the optimal number of subpopulations using the cross validation error method, and plot your results. I have also uploaded a file called “indivs”, which contains the ID’s of all the individual *T. pallidum* strains used in this study, along with an Excel file containing information about the geographic origins of these strains. Thereon, use R to determine the ancestries (for example, to determine which individuals belong predominantly to the first subpopulation, say you’ve read the ancestry file into R as: x<-read.table(“tp.3.Q”, header=FALSE), then you can use y<-which(x$V1>=0.9), which will give you the columns of individuals with predominantly ancestry from the first subpopulation, say you’ve read the indivs file into a variable called z, then you can just use z[y] which will give you all the individual strains in the first subpopulation. Now refer to the Excel file provided to interpret the subpopulation structure within global strains of T. pallidum. Explain your answer with respect to the findings from the paper (25 points).