**BIOMI 609 Computational Genomics and Bioinformatics**

**Spring 2022**

**San Diego State University**

**Dr. Arun Sethuraman**

**Lab 4 Population Genomics**

This week, we will analyze SARS-CoV2 data to understand their population genomics and evolution, directly applying all the concepts that we have learned so far to real genomic data.

Part 1: Multiple Sequence Alignment in NCBI

1) Go to NCBI Genbank (<https://www.ncbi.nlm.nih.gov/genbank/>)

2) Search for a gene/genome that you’re interested in – say “SARS-CoV2” – this should take you to the “Reference Genome Sequence” for SARS-CoV2 (specifically the Wuhan-Hu-1 strain). We will go over the different features that are shown in the genome reference.

3) Specifically note the regions of the genome that are broken down into different protein classes. We know that this virus specifically works via attachment of the Spike protein to the ACE2 receptors on human alveolar cells (<https://www.youtube.com/watch?v=5DGwOJXSxqg>).

So let’s try and align just variants present across the Spike protein region. Scroll down the page to determine the chromosomal locations of the spike protein open reading frames.

4) Now click on “Run BLAST” on the right hand side of the website, and this will take you to the BLAST page; enter specifically just the chromosomal coordinates that you want to search for, and you can play around with percentage identity, etc. (I will guide you through this).

5) Then go ahead and hit “BLAST” – this will take a few minutes to pull up most similar sequences. I will help you navigate what the output of BLAST looks like and how to interpret that.

6) To download the obtained sequences, click on “Select All”, then “Download”->FASTA (aligned sequences). This should download a file with the sequences. Unfortunately some times, this isn’t actually a fully aligned sequence list. So we have to re-run a multiple sequence alignment, and we are going to use a software called MUSCLE to do that.

7) So now go to the EBI MUSCLE website (<https://www.ebi.ac.uk/Tools/msa/muscle/>), then upload your FASTA file that you just downloaded (by clicking on “Choose File” and picking the file you downloaded). Then set the OUTPUT FORMAT to “Pearson/FASTA”. Then go ahead and hit “Submit”. This will take a few minutes to run, and then take you to the aligned sequences. Now you can download the alignment by right-clicking on “Download Alignment File”, and clicking on “Save Link As”, and enter a name for your file. This should download the multiple sequence alignment file to your computer. Voila!

You can play around with this for different viruses – HIV, HPV, etc.

You can also play around with different SARS-CoV2 strains - delta vs omicron vs alpha, etc.

Part 2: Using MEGA to analyze viral genomes

1) Under Week 11, I’ve uploaded 2 multiple sequence alignment files here – one is a whole-genome alignment of 153 SARS-CoV2 genomes downloaded from <https://genexa.ch/sars2-bioinformatics-resources/>, the second is an alignment of 100 pol protein region from HIV1, that I have already aligned using MUSCLE. Download these.

2) Now open MEGA on your machine. Now click on Align->Edit/Build Alignment->Choose your “hiv\_pol.fasta”. This should pull up your alignment in a separate window. I will help you navigate this window, and understand what’s going on. Also, note that IF sequence 84 seems to be of very poor quality, so go ahead and click on that one and delete it. Otherwise no worries!

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application

Description automatically generated3) Now go ahead and click on “Data”->”Phylogenetic Analysis”->Click on “Yes” when prompted if this is protein-coding sequence data. Now click back on the MEGA window, and explore what the “TA” button does. I’ll help you navigate the window that opens up, where you can view the data, highlight segregating sites, translate it, etc.

4) Now click on the “Phylogeny” button, and construct a simple Neighbor Joining tree. Look at the phylogenetic tree – what can you infer about the evolutionary history of these sequences based on that tree?

5) Now close the tree, and click on “Selection”->Tajima’s Test of Neutrality. What are the statistics that are computed? Interpret these. Close this. Similarly, perform Codon-based tests of neutrality, and let’s look at these results and try to interpret them.

6) Let’s repeat these analyses with the SARS-CoV2 multiple sequence alignment file now, and compare the estimates of selection. When done, go ahead and close MEGA.

Part 3: Analyses of selection using R (pegas package):

1) Open R inside Jetstream, and install and then load the pegas library using install.packages(pegas), then library(pegas). Then change the working directory to the folder where you’ve downloaded the FASTA files.

2) Now read the fasta files:

hiv<-read.dna(“hiv\_pol.fasta”,”fasta”)

sars<-read.dna(”sarscov2.fasta”,”fasta”)

3) Estimate nucleotide diversity across both the viruses:

nuc.div(hiv)

nuc.div(sars)

What do you notice first off, about the variability in the two virus sequences across populations?

4) Now let’s perform some neutrality tests:

tajima.test(sars)

tajima.test(hiv)

What can you state about the results of these Tajima’s D tests?

Now let’s spend some time going over these two papers, and understanding why/how our findings are relevant (I’ve posted the papers on your Canvas page).

Rausch, Jason W., et al. "Low genetic diversity may be an Achilles heel of SARS-CoV-2." *Proceedings of the National Academy of Sciences* 117.40 (2020): 24614-24616.

Dearlove, Bethany, et al. "A SARS-CoV-2 vaccine candidate would likely match all currently circulating variants." *Proceedings of the National Academy of Sciences* 117.38 (2020): 23652-23662.

We will also spend some time going over the results from across millions of genomes on nextstrain.org to understand the evolution of different strains.