Mar 14-15 # RESULT ANALYSIS

SARS-CoV-2 majorly impacted the world, leading to widespread economic downfall, and the effects can still be seen today.

We can takeaway the following things:

- take control of the pandemic early, this will reduce virus variability and bring neutrality even more negative, leading to more effective vaccinations in future

- developed countries should focus on helping densely populated, lers rich countries as this is where hotspots can cause increased variability and high positive selection (highly adeptable virus), and monitor outbreaks of highly positive neutrality viruses (Africa)

- SARS-CoV-2 is beginning to rapidly evolve again, the 2024 data shows a sharp jump in neutrality towards positive selection. Our current vaccines still need to be developed and people must take boosters ~~on~~ due the spike protein gene may mutate so much another pandemic occurs.

Mar 14 Statistics – T-test, P-value, Null hypo.

I wanted to test the statistical significance of the changes in SARS-CoV-2 neutrality from the start of the pandemic in 2019 the recent data from 2024.

To do this, a T-test was conducted. The test was one-tailed and heteroscedastic, because we predicted and could see in the data that the 2024 neutrality was more positive (closer to 0), and the variance of the two statistical groups were Unequal.

The T-tests were conducted with the Excel built-in functions for T-testing

P-VALUES OF EACH NEUTRALITY TEST (GLOBAL 2019/2020 VS GLOBAL 2024)

Tajima's D: 0.00222525

Fu and Li's D: 0.001142774
Fu and Li's D*: 0.000936199
Fu and Li's F: 0.0000225407
Fu and Li's F*: 0.0000135195

Fay and Wu's Normalized H: 0.000222243
Fay and Wu's Normalized E: 0.096089257

ALL VALUES SIGNIFICANTLY BELOW 0.05, MEANING I REJECTED NULL HYPOTHESIS

Hilroy

Mar 11

C program - FASTA N base Filtering

To remove sequences with unknown "N" bases, I wrote a C program that inputs the FASTA from EpiCoV and outputs a filtered FASTA free of bad data.

Overview of Code

- reads each line of FASTA until the end of an individual sequence is reached (next sequence's header)
- if sequence contained unknown bases (character "N"), skip it and do not output
- if sequence is all correct, no unknown bases, output all lines of that sequence to output FASTA

→ Ran on 2024 dataset March 11, all other aspects of this dataset same to other ones

Mar 11-13 2024 January-February dataset

Downloaded from GISAID EpiCoV March 11: Complete, High Coverage* filters

Sequences from Global (all locations), very recent collection dates of Jan 2024 and Feb 2024
3010 sequences at download, 221 after Filtering

* Due to the recency of this data, it seems that GISAID has not processed the Coverage filters fully yet. Therefore, there were still sequences in the downloaded FASTA that had sequences in the downloaded FASTA that had "N" values in their genome, meaning unknown RNA bases. For the best quality of data, these sequences were filtered out (see next page) using C code.

completed March 11, freebayes also completed March 11, Neutrality testing completed March 13

Mar

Gene_name	Coordinates	Tajima's D	FuLi's D	FuLi's D*	FuLi's F	FuLi's F*	FayWu's H	FayWu's E
gene-GU280_gp01	NC_045512.2:266:21555	-2.665858	-11.717993	-10.639525	-8.108021	-7.770627	0.309929	-2.69377
gene-GU280_gp02	NC_045512.2:21563:25384	-1.705819	-7.018726	-7.134907	-5.094482	-5.297723	0.927771	-2.3294
gene-GU280_gp03	NC_045512.2:25393:26220	-2.458876	-7.751122	-7.354333	-6.36782	-6.172238	0.290931	-2.388277
gene-GU280_gp04	NC_045512.2:26245:26472	-1.008282	-0.804932	-0.791489	-1.047627	-1.03533	0.525158	-1.167293
gene-GU280_gp05	NC_045512.2:26523:27191	-1.475966	-6.751276	-6.499666	-5.390361	-5.253959	1.008019	-2.075129
gene-GU280_gp06	NC_045512.2:27202:27387	-1.937061	-5.495366	-5.419053	-5.041753	-4.982781	0.262213	-1.710352
gene-GU280_gp07	NC_045512.2:27394:27759	-2.188509	-6.90304	-6.708154	-6.00044	-5.877491	0.246464	-2.013221
gene-GU280_gp08	NC_045512.2:27756:27887	-0.921606	-3.686068	-3.653403	-3.270014	-3.242919	0.330244	-0.936995
gene-GU280_gp09	NC_045512.2:27894:28259	-2.158882	-5.943016	-5.75451	-5.262305	-5.145258	0.268167	-2.024938
gene-GU280_gp10	NC_045512.2:28274:29533	-2.333272	-6.552621	-6.200849	-5.49873	-5.324852	0.443361	-2.413178
gene-GU280_gp11	NC_045512.2:29558:29674	-1.430663	-4.222545	-4.179669	-3.885132	-3.849931	0.253113	-1.276255

Mar 8-9 2022 Africa (Omicron) dataset

Downloaded from GISAID EpiCoV database March 8:
Complete, High Coverage filters, and Omicron filter

Sequences from continent of Africa, collection date year of 2022, however only Omicron variants
2952 total sequences in FASTA

Note: Botswana, Africa was the origin of Omicron variant. Densely populated and low vaccinations

freebayes and CTSPLIT completed March 8
Neutrality CATE tests completed March 9

below: simplified table of result data

Gene_name	Coordinates	Tajima's D	FuLi's D	FuLi's D*	FuLi's F	FuLi's F*	FayWu's H	FayWu's E
gene-GU280_gp01	NC_045512.2:266:21555	-2.706291	-18.005758	-17.624346	-8.462254	-8.429688	0.075245	-2.459937
gene-GU280_gp02	NC_045512.2:21563:25384	-2.438433	-12.510167	-12.324957	-6.945002	-6.920228	0.245489	-2.338725
gene-GU280_gp03	NC_045512.2:25393:26220	-2.582505	-11.798325	-11.696999	-7.688078	-7.664896	0.081535	-2.318215
gene-GU280_gp04	NC_045512.2:26245:26472	-1.961859	-4.509876	-4.499717	-4.202038	-4.195815	0.218105	-1.785967
gene-GU280_gp05	NC_045512.2:26523:27191	-2.362934	-7.411109	-7.371525	-5.791923	-5.777634	0.194722	-2.18025
gene-GU280_gp06	NC_045512.2:27202:27387	-2.154502	-5.532976	-5.519094	-4.970468	-4.962539	0.207804	-1.949588
gene-GU280_gp07	NC_045512.2:27394:27759	-2.542663	-10.102402	-10.048275	-7.450586	-7.432174	0.039053	-2.221648
gene-GU280_gp08	NC_045512.2:27756:27887	-2.167176	-5.020344	-5.010191	-4.699961	-4.693524	0.065891	-1.838975
gene-GU280_gp09	NC_045512.2:27894:28259	-2.519463	-8.317784	-8.278798	-6.541666	-6.52656	0.03122	-2.18552
gene-GU280_gp10	NC_045512.2:28274:29533	-2.630637	-12.544337	-12.411344	-7.710999	-7.686024	0.085132	-2.374116
gene-GU280_gp11	NC_045512.2:29558:29674	-1.928923	-1.112684	-1.110926	-1.784006	-1.782269	0.043552	-1.588165

Mar 5-7 2021 New Zealand dataset

Mar 8

Downloaded from GISAID EpiCoV March 5:
Complete, High Coverage

Sequences from New Zealand, collection dates year 2021
total of 1723 sequences in FASTA

Note: New Zealand took control of their pandemic quite early on, leading to drastic decreases in case numbers and less impacts and more opportunities for vaccinations

freebayes and CTBPLIT completed March 6, neutrality March 7

Gene_name	Coordinates	Tajima's D	FuLi's D	FuLi's D*	FuLi's F	FuLi's F*	FayWu's H	FayWu's E
gene-GU280_gp01	NC_045512.2:266:21555	-2.790264	-21.808043	-21.221867	-10.80779	-10.74023	0.046209	-2.516184
gene-GU280_gp02	NC_045512.2:21563:25384	-2.702513	-18.038496	-17.7707	-10.84993	-10.793188	0.057738	-2.416663
gene-GU280_gp03	NC_045512.2:25393:26220	-2.56189	-8.76149	-8.687641	-6.765252	-6.737647	0.050809	-2.246391
gene-GU280_gp04	NC_045512.2:26245:26472	-2.015599	-7.689645	-7.673951	-6.719774	-6.708929	0.013151	-1.629842
gene-GU280_gp05	NC_045512.2:26523:27191	-2.306358	-6.515481	-6.481885	-5.604841	-5.587533	0.098935	-2.018513
gene-GU280_gp06	NC_045512.2:27202:27387	-1.969944	-5.562273	-5.551488	-5.125144	-5.117305	0.015686	-1.588986
gene-GU280_gp07	NC_045512.2:27394:27759	-2.309118	-9.167435	-9.134583	-7.634215	-7.615546	0.01676	-1.926189
gene-GU280_gp08	NC_045512.2:27756:27887	-1.683372	-4.669722	-4.663944	-4.386235	-4.381595	0.012698	-1.325687
gene-GU280_gp09	NC_045512.2:27894:28259	-2.3755	-8.485172	-8.449565	-7.089294	-7.070047	0.018639	-1.998452
gene-GU280_gp10	NC_045512.2:28274:29533	-2.657578	-13.68161	-13.542173	-9.372659	-9.330667	0.026554	-2.327106
gene-GU280_gp11	NC_045512.2:29558:29674	-1.748896	-2.116562	-2.113299	-2.430754	-2.427932	0.017399	-1.387827

Mar 2-3 2019 - June 2020 Italy dataset

Downloaded from GISAID EpiCoV Mar 2:
Complete, High Coverage, with Patient Status

Sequences from ~~I~~taly all regions, from start of epidemic to June 2020. Note: Italy was impacted quite majorly when virus first spread, death toll extremely high (spiked at 1.6 per 100k people daily)

Total sequences 1568

freebayes and CIPSPLIT completed March 2, neutrality March ~~2~~3

Gene_name	Coordinates	Tajima's D	FuLi's D	FuLi's D*	FuLi's F	FuLi's F*	FayWu's H	FayWu's E
gene-GU280_gp01	NC_045512.2:266:21555	-2.721789	-18.589195	-18.077164	-9.695315	-9.628904	0.089809	-2.488476
gene-GU280_gp02	NC_045512.2:21563:25384	-2.716259	-15.628966	-15.358515	-9.5528	-9.496064	0.06177	-2.437155
gene-GU280_gp03	NC_045512.2:25393:26220	-2.578916	-11.623758	-11.522084	-8.55892	-8.521221	0.044834	-2.252256
gene-GU280_gp04	NC_045512.2:26245:26472	-1.559861	-2.379486	-2.375872	-2.547041	-2.543978	0.088824	-1.280964
gene-GU280_gp05	NC_045512.2:26523:27191	-2.339838	-8.612319	-8.573009	-7.170735	-7.149339	0.04519	-1.985686
gene-GU280_gp06	NC_045512.2:27202:27387	-2.01773	-5.990295	-5.976383	-5.446341	-5.43648	0.02079	-1.636578
gene-GU280_gp07	NC_045512.2:27394:27759	-2.249273	-8.144005	-8.099416	-6.673223	-6.650795	0.24666	-2.075798
gene-GU280_gp08	NC_045512.2:27756:27887	-1.302691	-2.23834	-2.236135	-2.331758	-2.329826	0.110803	-1.073405
gene-GU280_gp09	NC_045512.2:27894:28259	-2.158586	-4.187885	-4.173446	-4.112427	-4.102558	0.057024	-1.812216
gene-GU280_gp10	NC_045512.2:28274:29533	-2.527249	-9.8577	-9.756529	-7.278723	-7.244008	0.097442	-2.260702
gene-GU280_gp11	NC_045512.2:29558:29674	-2.016159	-6.804019	-6.788334	-6.052705	-6.041747	0.021664	-1.635935

Feb 26-29 2019-May 2020 GLOBAL dataset

Mar 2-3

- ~~test~~ now that pipeline has been finalized the first dataset for formal analysis was obtained:

Downloaded from GISAID EpiCoV Feb 26:
Complete, High Coverage, with Patient Status

Sequences from all global locations, start of pandemic in December 2019 to May 2020;
9935 sequences in total in FASTA

Ran through pipeline, simplified result below

freebayes and split-VCF (CTSPLIT) completed Feb 27
neutrality testing completed Feb 29

Gene_name	Coordinates	Tajima's D	FuLi's D	FuLi's D*	FuLi's F	FuLi's F*	FayWu's H	FayWu's E
gene-GU280_gp01	NC_045512.2:266:21555	-2.687141	-32.730705	-32.431465	-12.009747	-11.995693	0.02025	-2.391884
gene-GU280_gp02	NC_045512.2:21563:25384	-2.665126	-26.913971	-26.740633	-11.512069	-11.499301	0.025459	-2.369421
gene-GU280_gp03	NC_045512.2:25393:26220	-2.602483	-18.790056	-18.714504	-9.891827	-9.881908	0.027996	-2.303241
gene-GU280_gp04	NC_045512.2:26245:26472	-2.387288	-11.131731	-11.119412	-8.452331	-8.447262	0.008024	-2.036911
gene-GU280_gp05	NC_045512.2:26523:27191	-2.502661	-10.559076	-10.536423	-7.275305	-7.26935	0.035705	-2.195493
gene-GU280_gp06	NC_045512.2:27202:27387	-2.337372	-8.538905	-8.530309	-6.889363	-6.885457	0.01863	-1.994837
gene-GU280_gp07	NC_045512.2:27394:27759	-2.532944	-15.788511	-15.757904	-10.269608	-10.261526	0.006472	-2.195935
gene-GU280_gp08	NC_045512.2:27756:27887	-2.194906	-8.154276	-8.148948	-6.889233	-6.88625	0.005852	-1.833866
gene-GU280_gp09	NC_045512.2:27894:28259	-2.512329	-13.727953	-13.700127	-9.055339	-9.048083	0.023415	-2.1924
gene-GU280_gp10	NC_045512.2:28274:29533	-2.635528	-18.984985	-18.89645	-9.581333	-9.571378	0.020572	-2.331517
gene-GU280_gp11	NC_045512.2:29558:29674	-2.292399	-8.99019	-8.982628	-7.31498	-7.31125	0.008307	-1.936601

Feb 25 First Run (NY2022) RESULTS

Bioinformatics Pipeline is now complete

CATE outputted a [population name]_genelist.nt file which can be locally downloaded and opened in Excel

↳ includes π (average pairwise differences), number of segregating sites, and other relavent variables, however below table has been simplified for space (see full chart on trifold and Github)

Gene_name	Coordinates	Tajima's D	FuLi's D	FuLi's D*	FuLi's F	FuLi's F*	FayWu's H	FayWu's E
gene-GU280_gp01	NC_045512.2:266:21555	-2.11934	-3.40766	-2.44115	-3.69896	-2.64921	1.126316	-3.28987
gene-GU280_gp02	NC_045512.2:21563:25384	-1.67382	-2.12573	-1.78243	-2.42077	-1.97467	1.301753	-2.73915
gene-GU280_gp03	NC_045512.2:25393:26220	-1.99828	-3.04393	-2.38176	-3.30832	-2.56932	0.877667	-2.41751
gene-GU280_gp04	NC_045512.2:26245:26472	NA	NA	NA	NA	NA	NA	NA
gene-GU280_gp05	NC_045512.2:26523:27191	-1.43714	-1.60634	-1.35316	-1.84461	-1.55085	0.940813	-1.779
gene-GU280_gp06	NC_045512.2:27202:27387	-1.46801	-1.95589	-1.7764	-2.14048	-1.91712	0.422305	-1.22915
gene-GU280_gp07	NC_045512.2:27394:27759	-1.6388	-1.98733	-1.57814	-2.26284	-1.79671	1.132855	-2.27059
gene-GU280_gp08	NC_045512.2:27756:27887	-0.27429	0.689539	0.73235	0.53135	0.543074	1.27117	-1.21187
gene-GU280_gp09	NC_045512.2:27894:28259	-1.09932	-0.66752	-0.50933	-0.92231	-0.74598	0.708664	-1.35094
gene-GU280_gp10	NC_045512.2:28274:29533	-1.81172	-2.42414	-1.81856	-2.72129	-2.04615	1.162088	-2.66215
gene-GU280_gp11	NC_045512.2:29558:29674	NA	NA	NA	NA	NA	NA	NA

Feb 22 cont. # CATE Neutrality Tests

After the vcf CTSPLIT finished running, The following was completed:

- check split_VCF folder, .out and .error files from computing cluster (method will vary depending on cluster used) to make sure CATE ran successfully - should see a folder within split_VCF, within it another titled the population name used, then the processed VCF (split_VCF/Name/pop name/.vcf)

IMPORTANT: due to some programmed features of CATE, it creates an extra folder that the neutrality test does not recognize. (usually the name of the input .vcf) Therefore, processed VCF (the innermost folder in tree) must be moved to directly inside split_VCF, tree as follows:
working directory/split_VCF/population_name/processed.VCF containing the

ex. mine was Project/split_VCF/NY22/processed.VCF

then the parameters.json file was modified as follows:

input path = split_VCF (output from CTSPLIT)
output path = results

Universal gene list = "genelist.txt"
↳ upload R genelist by SFTP client

Other hardware-dependent parameters can remain the same from CTSPLIT function parameters

Feb 22

R Bioconductor: CATE Genelist

Before conducting Neutrality testing on CATE, Genelist must first be created. Requirements as follows:

- Tab-delimited text file
Column 1: Gene Name (ex. gene-GU280_gp01)
Column 2: formatted gene details:
[Accession ID]:[start position]:[stop position]
positions are genomic locuses

R Code Overview

- loads necessary libraries, updates if needed
- connects to BiomaRt Ensembl database
↳ downloads necessary genes using Accession ID NC_045512.2 (Wuhan-Hu-1)
- formats into genelist

Alternative Method

By downloading the SARS-CoV-2 GFF feature file from Ensembl, we can also use CATE's built-in GFF → Gene function

Feb 21

CATE vcf SPLIT (CTSPLIT)

Feb 22

To begin analysis with CATE, we must first "split" our input VCF. CATE will create its proprietary file structure and process the VCF.

→ connect to Compute Canada Narval nodes (or chosen computing cluster)
(I used MobaXTerm as SSH client)

- cloned CATE Github
  ↳ copy parameter file and CATE executable from cloned to my working directory as well

- in working directory, created folder named vcf_Full and uploaded VCF (from SARS-CoV-2-freeboyes and after C code processing) using Filezilla (SFTP)

- also uploaded completed Population file (not in vcf_Full, in main working directory)

→ modified base CATE parameter file as follows:

input path: vcf_Full output path: split_VCF
intermediate: intermediate

Prometheus: No (not needed)
CPU cores / SNPs per time / etc. = used DEFAULTS
hardware dependent, Compute Canada used

VCF sample details - Ploidy: 1 (VIRUS = HAPLOID)

VCF Split mode: CTSPLIT
Population file path: full/path/to/populationfile
Column numbers: we used 1 and 2 (see pop file)

Feb 17 C code: construct Population File

population.c

CATE requires a population file with the names of each individual sequence (SAMPLE columns) from the input VCF file. The population file also includes a column for the population name (user-defined)

Requirements for CATE CTSPLIT
→ Tab-delimited .txt (.tsv = tab-delimited)

Column 1: Sample ID (all sequences in VCF)
Column 2: Population ID (ex. NEWYORK22)

I decided to write another C program to automatically extract the sequence names from the SAMPLE columns in the VCF (formatted from the earlier VCF format/processing C code in the Pipeline)

Code Overview

- parses input VCF line-by-line until headers are found (begins with "#")

- tokenizes header string by tabs (tab-delimited)
  ↳ ignores all columns other than the SAMPLE ones (after FORMAT)

- for each SAMPLE column name, output to column 1 of a tab-delimited text file, output user-defined population name in column 2

RESULT = formatted CATE population file
↳ completed run on First Run VCF

Feb 16 cont. C VCF processing

github
filter_multiple.c
filter_multiple_cleanup.c

Issue Encountered

CATE software only analyzes SNPs for neutrality testing (Single Nucleotide Polymorphisms) meaning data with Indels must be removed

Solution: C program to filter out non-SNP variation in VCF generated by SARS-CoV-2-freebayes
↳ my code does this by two passes

PASS ONE

- tokenizes each line of VCF file (tab-delimited)
- prints each column of each row from the input VCF file to an intermediate VCF file as it is read, however if REF or ALT column contains more than one nucleotide base variation, SKIP it
  ↳ CHROM, ID, POS columns of offending line still printed to output file

PASS TWO

= removes all lines that are shorter than 50 characters from the intermediate VCF, stores the final cleaned-up results in an output VCF file = all offending lines removed

RESULT = formatted VCF file ready for analysis by CATE
↳ completed run of code on the First Run data (New York 2021 Oct-2022Apr)

Feb 16 C code: VCF processing

Overview of VCF format (Variant Call File)

= stores gene sequence variations, (SNP, indels, etc.) Columns Include:

CHROM / POS / ID columns

the sequence location on chromosome, postion (coordinates) sequence of variation, identifier name of the variation (unique or reference to database)

REF

the nucleotide base from the reference sequence at this location of variation (for comparison)

ALT

the alternative allele nucleotide base (variation) (ones that vary from the reference genome)

INFO

provides information (such as variation type)
↳ SARS-CoV-2-freebayes provides Allele Frequency data

FORMAT

outlines what information will be included in the following SAMPLE columns (SARS-CoV-2-freebayes only provides GT (genotype) data, see below)

SAMPLE (all columns)

contain genotype information for each individual sequence (each has its own column for each variant row)

↳ "1" in column = sequence contains ALT allele
↳ "0" in column = sequence contains REF allele

Feb 13 Neutrality Tests (Background)

Feb

Popgen Text

Tajima's D (1989)

- compares two measures of genetic diversity:
  - → Segregating Sites: positions in genome where variation is observed (ex. SNP)
  - → Pairwise Differences: average number of nucleotide differences between pairs of DNA sequences in sample (= genetic diversity of population)

Fu and Li's D, F, D*, F* (1993)

- extension of Tajima's D, but focuses on the distribution of singleton (rare) mutations in the population = only in a few individuals, giving an additional level of granularity
- compares number of singleton mutations with number of mutations (D = total number, F = average)
- star (*) variations correct for specific biases

Fay and Wu's Normalized H and E

- H focuses on Positive selections

→ adaptive evolution

- E more similar to previous tests with minor improvements

Feb 11 Analysis of VCF files: CATE

CATE = CUDA Accelerated Testing of Evolution → COMPUTING CLUSTER
↳ Nvidia technology for GPU computing

conducts evolutionary testing (ex. neutrality tests) on large-scale VCF data very efficiently

written with C, C++, CUDA, already tested with GISAID SARS-CoV-2 data (reliable)

Utility Functions (some not included for brevity)

VCF Splitter: required to use CATE, creates proprietary file structure and indexes the VCF input

FASTA Spitter/Merger: seperates/merges FASTA files

Haplotype Extractor: finds unique haplotypes for different regions, reconstructs FASTA

EVOLUTIONARY TESTS (research background later)

Tajima's D: Tajima's D statistic (1989)
Fu and Li's: Fu and Li's D, D*, F, F* statistics (1993)
Fay and Wu's: Fay and Wu's normalized H and E (2006)
→ can also calculate above three all at once

Mcdonald-Kreitman neutrality index: (1991)

Fixation Index: (1965) (Fst)

Extended Haplotype Homozygosity (EHH): (2002)

Feb 11 Analysis of VCF files: CATE

CATE = CUDA Accelerated Testing of Evolution
↳ Nvidia technology for GPU computing
→ COMPUTING CLUSTER

conducts evolutionary testing (ex. neutrality tests) on large-scale VCF data very efficiently

written with C, C++, CUDA, already tested with GISAID SARS-CoV-2 data (reliable)

Utility Functions (some not included for brevity)

VCF Splitter: required to use CATE, creates proprietary file structure and indexes the VCF input

FASTA Splitter/Merger: seperates/merges FASTA files

Haplotype Extractor: finds unique haplotypes for different regions, reconstructs FASTA

EVOLUTIONARY TESTS (research background later)

Tajima's D: Tajima's D statistic (1989)
Fu and Li's: Fu and Li's D, D*, F, F* statistics (1993)
Fay and Wu's: Fay and Wu's normalized H and E (2006)
→ can also calculate above three all at once

Mcdonald-Kreitman neutrality index: (1991)

Fixation Index: (1965) (Fst)

Extended Haplotype Homozygosity (EHH): (2002)

Feb 8

SARS-CoV-2-freebayes: First Run

21-22 New York

- used Windows Linux Subsystem (WSL)
  ↳ allows for use of Bioconda library
- downloaded Conda software (Python package manager)
  ↳ and prerequisite packages for freebayes
  = Bioconda, Conda-Forge, Samtools, Vcftools
- followed SARS-CoV-2-freebayes Installation guide
  Github

Note: some problems regarding admin commands occured on Compute Canada / ARC (cloud computing clusters) therefore a local machine execution was deemed the best course of action. However, some library incompatibilities were found to cause errors. To solve this, the Conda libraries VCFlib and SnpEff, along with the base freebayes library, were downgraded to 1.0.3, 5.0, and 1.3.6, respectively (versions from approximately 2023 worked best)

- indexed Reference Sequence using samtools faidx
- increase stack size / file limit (ulimit) = if needed
- Execute SARS-CoV-2-freebayes Pipeline
  ↳ this study used nolimit binary versions

= VCF results of First Run sequences

(New York 2021 Oct - 2022 Apr)

Feb 11

Feb 7 SARS-CoV-2-freebayes

FINDING VARIATIONS: FASTA to VCF

Freebayes - Github
= Bioinformatics Library for variant detection
↳ Bayesian genetic variant detector, finds SNPs, InDels, MNPs, complex events

= haplotype-based (differs from other models)
↳ calls variants based on the literal sequences of reads aligned to a particular Target, not precise alignment)

→ Uses short-read alignments (BAM files) for individuals in the population and a reference genome to report combinations of genotypes in VCF format

Actual Library Used: SARS-CoV-2-freebayes

↳ conda environment, specializes in SARS-CoV-2 (provides necessary BAM, etc.)

Feb 3 cont.

DATA FROM GISAID EPICOV cont.

EPI_ISL ID: EPI_ISL_12357962
virus Name: hCoV-19/USA/NY-WMC2022-1969/2021
Date/Loc: 2021-12-31, New York USA
YYYY/MM/DD

EPI_ISL_12357971
hCoV-19/USA/NY-WMC2022-1979/2022
2022-01-01, New York USA

EPI_ISL_12358086
hCoV-19/USA/NY-WMC2022-2228/2021
2021-12-31, New York, USA

EPI_ISL_12358087
hCoV-19/USA/NY-WMC2022-2229/2021
2021-12-31, New York USA

EPI_ISL_12358100
hCoV-19/USA/NY-WMC2022-2246/2021
2021-12-31, New York USA

12 FASTA Datasets in total,
2021 October - 2022 April, New York USA

Feb 3

FIRST ANALYSIS FOR PIPELINE

DATA FROM GISAID.EPICOV

EPI_ISL_12154784 ← EPI_ISL Accession ID

virus Name: hCoV-19/USA/NY-COV-4066/2021
Date/Loc: 2021-12-27, New York, USA
YYYY/MM/DD

EPI_ISL_12154785
hCoV-19/USA/NY-COV-4067/2021
2021-12-29, New York, USA

EPI_ISL_12154787
hCoV-19/USA/NY-COV-4070/2022
2022-04-19, New York, USA

EPI_ISL_12155757
hCoV-19/USA/NY-3303/2021
2021-10-22, New York USA

EPI_ISL_12155759
hCoV-19/USA/NY-3321/2021
2021-11-22, New York USA

EPI_ISL_12156129
hCoV-19/USA/NY-3405/2021
2021-12-03, New York USA

EPI_ISL_12175204
hCoV-19/USA/NY-MSK-184/2021
2021-01-24, New York USA

Jan 29

Feb 3

TYPES OF GENETIC VARIATION

Single-Nucleotide Polymorphism (SNP)

= substitution of a single nucleotide at a specific position in the genome
↳ present in a large fraction of population (>1%)

ex. ATCTAACGTAC VS ATCTAGCGTAC
97% population 3% population

Insertion/Deletion (Indels)

= specific nucleotide sequence is added (insertion) or not present (deletion) or single
↳ commonly causes frameshift mutation

ex. cystic fibrosis = deletion causing missing AA

Multi-Nucleotide Polymorphism (MNP)

= includes double, triple, more nucleotide substitutions at a genome position

virus Nam
Date/Loc
YYYY/MM/

POPULATION GENETICS TEXTBOOK USED FOR STUDY =

An Introduction to Population Genetics:
Theory and Applications

by Rasmus Nielsen, Montgomery Slatkin

Bio
Econ
English Engl

Jan 28

HARDY-WEINBERG EQUILIBRIUM

allele and genotype frequencies from generation to generation in a population will remain CONSTANT in the absence of other evolutionary influences

Allele Frequency (phenotype)

$$p + q = 1$$

- p — dominant
- q — recessive

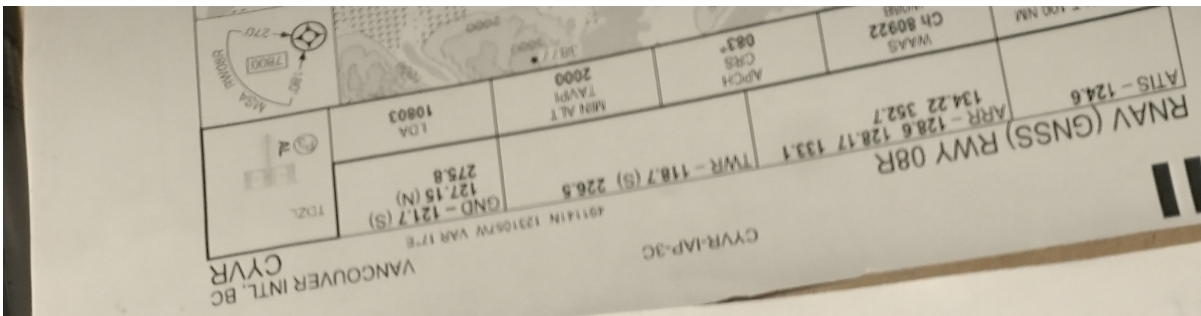
Genotype Frequency

$$p^2 + 2pq + q^2 = 1$$

- p^2 — homozygous dominant
- $2pq$ — hetero-zygous
- q^2 — homozygous recessive

REQUIRES:

- no natural selection
- no mutations
- no genetic migration
- large population (reduce impacts of genetic drift)
  - founder effect
  - bottleneck effect
- Random Mating (no sexual selection)



Jan 19

Data of SARS-CoV-2 Genome: GISAID EPICOV
↳ FASTA files (nucleotide sequence)

cloud computing network used for analysis: Compute Canada

First Analysis to develop pipeline:
- assorted variants
- Location = New York, USA
- 2021-2023 timeframe

* complete applications for services by next week

Jan 20

(Gene Migration)
Gene Flow = changes in alleles due to mixing with foreign populations
(Immigration/Emigration)

ENSEMBL - Genome Database Project
= centralized resource by European Bioinformatics Institute for access to annotated genomes of many species

annotated = relationships between genes and locations

BIOMART = R Bioinformatics Library
= allows automated retrival and processing of gene data, can be used with Ensembl

Math Math
Bio Bio
Econ Chem
English English

Jan 16

Wright-Fisher model (marbles) → Binomial distribution

Genetic Drift

= changes in allele frequencies (variations) caused purely by chance = random events unrelated to natural selection

- can cause **loss** of alleles
- affects small populations greatly

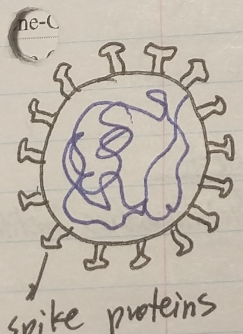
bottleneck effect = population sharply reduced by natural disaster

founder effect = small group splits off from main population

→ or can cause **fixation** (100%) of one allele

STRUCTURE OF THE SARS-CoV-2 GENOME

↳ coronaviruses are haploid
= contain one copy of genome, no chromosomes



stored in single strand of RNA inside virus
≈ 30,000 nucleotides in length

- open reading frames, coding regions (genes)

Jan 16

Jan 12 cont. ~~bioinformatics programs~~

R = main language, learn basics before Feb

Python also used, C++, etc.

recommend cloud computing cluster

↳ Linux shell command-line

Jan 13 Evolution

- natural selection (survival of the fittest)

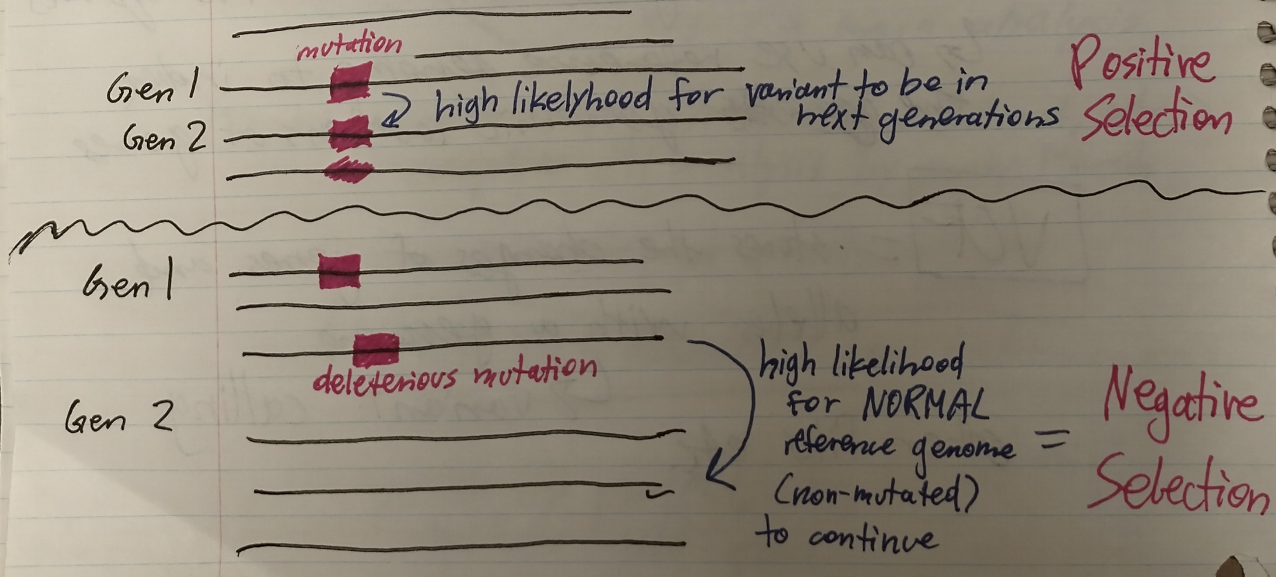
↳ mutations

Types

** different from advantageous and deleterious

Haplotype (gene combination)

Normal Gene Region (no change)



* neutral selection = doesn't really matter if mutated or not, equal amounts in next gen

Jan 7

- bioinformatics research
  ↳ study of large scale genomic data using computer analysis

SARS-COV-2 variants

Alpha, Beta, **Delta**, Gamma, **Omicron**

Jan 12 bioinformatics files and structures

FASTA = stores nucleotide sequences
(ex. ATGTCATCGGATC)
↳ however, has additions of other relavent information

BAM/SAM = alignment data of nucleotide sequences
↳ can use reference genome to index the ~~genome~~ genomic locations of genes

VCF = stores the changes of genes and alleles with a genome
↳ variant calling

other files etc.

Jan 5

Timelines

Jan 7

January: general planning, background research, finding data, networking [redacted]

February:

Early Feb { finish background research, (read papers relavent to topic), plan what tools to use, and learn the necessary programming concepts needed, learn some Bioinformatics concepts, population genetics

Jan 12

Mid Feb { begin working on project (download ~~necessary~~ software and libraries), finalize data used

Late Feb { complete first analysis (ex. tests from NFC, 2022 Jan-Mar)
↳ use learned lessons to expedite main analysis

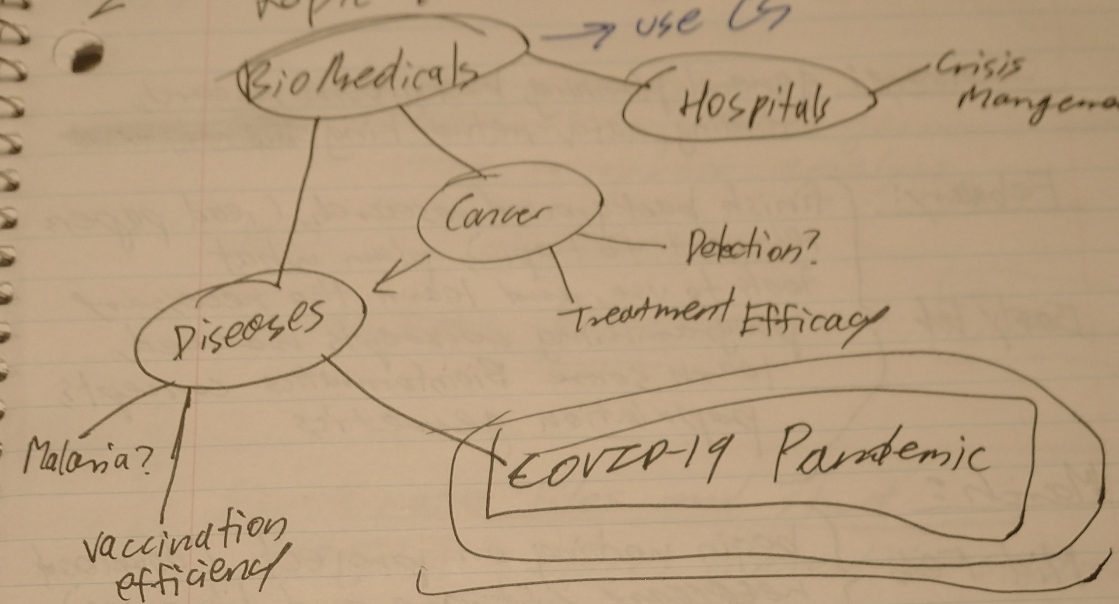
March: begin writing for trifold (abstract, etc.)
complete all analysis by early March
↳ interpret results

MARCH 15 = deadline for online

April: polish up project, construct trifold and glue writing

Jan 2

Topic brainstorming → use CS



Math
Bio
Ecor
Engl

COVID-19 Pandemic - SARS-CoV-2

- using CS = bioinformatics

- world taken by suprise
- EVOLVED QUICKLY = new variants
- must prevent similar global pandemics in future

→ How? What were differences in variants? etc.

↳ Changes in Genome

TOPIC = Analysis of Genetic Changes in SARS-CoV-2 and Variants