(Under materials and methods)

*Genomics data collection and processing*

SARS-CoV-2 sequencing is very rapidly conducted and made available to the public through the National Center for Biotechnology Information (NCBI) site. The reference genome used in this study is the accession: NC\_045512, one of the first complete genomes for SARS-CoV-2 sequenced from Wuhan, China in December 2019 (Wu et al. 2020). Additionally, another 435 SARS-CoV-2 genomes sequenced from various geographical locations were downloaded through the NCBI Virus database, which are the available genomes sequenced as of 1 April 2020. Using all 436 genomes, multiple alignment sequencing was performed using MAFFT (Multiple Alignment Fourier Fast Transform) software (Katoh 2013) with the NC\_045512 as the reference genome. The sequences were filtered according to its ‘quality’ and usefulness for our study by excluding sequences with excessive amount of non-base character (50% threshold). Specifically, we excluded 72 accessions that has more non-base characters than base characters. The final list of accessions is available in supplementary materials. Additionally, we also downloaded the Generic Feature Format (GFF) file from NCBI, and compiled regions of interest from UCSC genome browser for SARS-CoV-2 genome.

(Methods / Results)

*Identifying variable sites in the genome*

The data available from NCBI virus used in this study consist of the genomic sequence consensus based on their sequencing data. Therefore, unlike SNP-calling using variant calling methods that mainly uses raw sequence reads, we are able to directly align and identify parts of the genomes that are highly variable. In order to capture the variability of each site, the base frequency of each site is calculated. Then for each site, based on the variance of the base frequency, we are able to identify which sites has more variability than another, with low variance being high site variability, and high variance having low site variability.

A picture containing clock

Description automatically generated

**Figure X:** Variance calculation for each site across SARS-CoV-2 genome using the method described in Equation 1. The figure display some regions with high variance such as site 1068 and 24345 on annotated regions of notable proteins

Described in Equation 1, we calculated for each site i within the SARS-CoV-2 genome. The resulting plot display high variability for few sites, such as site 1068, 24345, located within the spike\_s2 and receptor binding domain (RBD) regions respectively, and many other sites (Figure. X).

**Combining genotypic and epidemiological finding**

In this part, we propose a method to group samples based on multiple genotypic patterns and compare them with our epidemiological findings. In the previous sections, it was shown that temperature is a correlated variable that contributes to the spreading of SARS-CoV-2, hence we attempt to combine the genotypic findings and the epidemiological findings. The first genotypic pattern to compare with temperature is grouping of samples based on its respective genotypes. The second genotypic pattern is the total genomic differences compared to the reference genome. Additional to temperature, we also display comparison compared to sequencing date to further show the possible correlation between the sequencing date and genomic changes.

*Comparing genotype groups using epidemiological data*

In the first method, the samples are grouped based on their respective genotypes. In the attempt to simplify this part, we look at specific protein regions containing high variability identified in previous section, shown in figure X. For instance, the region spike\_s2 is used in our study, and we found that within that region, there is a single site with higher variability, 24345. From this site, then we grouped our samples to the unique genotypes within the site, namely ‘C’, ‘T’, or ‘-‘, ‘-‘ indicates deletion relative to the reference genome. We further conducted principal component analysis (PCA) based on the spike\_s2 region using the dosage matrix relative to reference and found that the genotypes are quite spread out according to the first principal component (Figure X2).

A screenshot of a social media post

Description automatically generated

**Figure X2:** Principal component analysis (PCA) plot based on the spike\_s2 region dosage matrix, with samples labeled as its genotype groups. Genotype 0, 1, 2 has the base ‘-‘, ‘C’, ‘T’ respectively at site 24345.

From this genotype grouping, we can statistically compare the region temperature of each group, and such analysis can be done using analysis of variance (ANOVA), or Kruskal-Wallis test. For this specific region (spike\_s2), the P-Value for its Shapiro-Wilk test is well below 0.05, indicating that the samples are not normally distributed. Hence, we used the non-parametric test Kruskal-Wallis and the resulting p-value is 0.46, indicating that there is not enough evidence to state that the median temperature of the three genotype groups are equal.

*Correlating total genomic changes with epidemiological data*

This approach calculates the total genomic changes of a sample compared to the samples, then correlate it with specific epidemiological data information, such as region temperature or date when the samples were sequenced. In the attempt to conduct this analysis, the total genomic changes for each sample are calculated (Eq. 2).

Once the total genomic changes are calculated for each sample, conducting correlation study between the total genomic changes and the region temperature where the sample was taken is possible. Specifically, linear regression analysis is conducted on the data, where the response variable is region temperature of samples, and the predictor variable is the total genomic changes of samples.

A screenshot of a map

Description automatically generated

Figure X3: Regression between total genomic site differences vs temperature of region where samples were taken. The regression analysis yields R2 of 0.011, with p-value 0.047, slightly below 0.05.

Based on the regression study, it appears that the genomic site differences do not have significant correlation with temperature of the region where the samples are taken. Hence, from both the genotype grouping study and the total genomic site differences approach, we are unable to find a significant correlation between genotypic patterns and one of the epidemiological information, region temperature.

Reference

Olson, Nathan D., et al. "Best practices for evaluating single nucleotide variant calling methods for microbial genomics." *Frontiers in genetics* 6 (2015): 235.

Wu, F., et al. "A novel Coronavirus associated with a respiratory disease in Wuhan of Hubei province, China." *Submitted (05-JAN-2020) Shanghai Public Health Clinical Center & School of Public Health, Fudan University, Shanghai, China* (2020).