Comparing the direction and degree of selection on mutations between SARS-CoV-2 phylogenetic clades

Katherine Feldmann

Introduction

The discontinuous evolution of SARS-CoV-2 led to large jumps in genetic differentiation and the emergence of unique variants that can be characterized by one or more phylogenetic clades (Faria et al., 2021, Hodcroft et al., 2021, Naveca et al., 2021, Tegally et al., 2021, Viana et al., 2022, Volz et al., 2021). Mutations unique to certain phylogenetic clades can experience different directions and degrees of selection which may influence SARS-CoV-2 evolution (Neher 2022). The goal of this project is to generate two-clade comparisons for SARS-CoV-2 phylogenetic clades and identify significantly enriched mutations.

Methods

SARS-CoV-2 mutation data

The dataset used to compare nucleotide and amino acid mutations between SARS-CoV-2 phylogenetic clades was developed using genome sequences from the National Institutes of Health, the COVID-19 Genomics UK Consortium, and the China National Center for Bioinformation.

To identify mutations exhibiting different directions and degrees of selection between clades, mutations from the fourteen phylogenetic clades in the dataset were compared to all other clades, resulting in 91 two-clade comparisons.

Calculate q-values for two-clade comparisons

Within each two-clade comparison, the observed and expected counts of each mutation were compared using a Fisher's exact test. The expected counts were calculated using four-fold synonymous mutations. To account for the large number of hypothesis tests generated for each two-clade comparison, p-values were adjusted with a false-discovery rate correction, generating q-values.

Visualize q-values using volcano plots

Volcano plots were used to determine what mutations are significantly enriched for certain clades in a two-clade comparison. Mutations are considered significantly enriched if they have a large magnitude of difference (x-axis: log2 fold-change) and that difference is significant (y-axis: -log10 q-value). Fold-change was calculated as the ratio of observed to expected counts for clade 1 divided by the ratio for clade 2. Therefore, mutations with a positive fold-change are enriched in clade 1, and clade 2 for negative values.

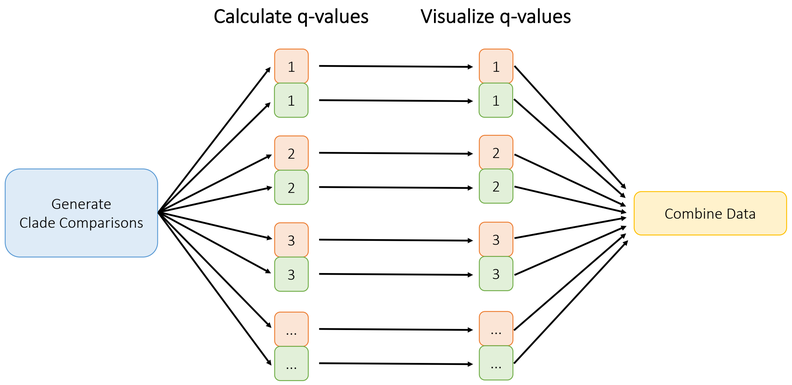

Figure 1: A Snakemake pipeline was used to parallelize (1) calculating q-values using a Fisher's exact test and false-discovery rate correction and (2) visualizing q-values using volcano plots across the 91 two-clade comparisons. Orange and green boxes indicate that nucleotide and amino acid comparisons for each two-clade comparison were also generated in parallel.

Results

The computational pipeline designed to calculate and visualize q-values in parallel for fourteen SARS-CoV-2 phylogenetic clades compared nearly 8 million nucleotide mutations and over 6 million amino acid mutations across 91 two-clade comparisons. To visualize the calculated q-values, the pipeline generated 91 nucleotide and 91 amino acid volcano plots. After filtering for expected counts greater or equal to five, 25 and 20 two-clade comparisons did not plot any significantly enriched nucleotide and amino acid mutations, respectively.

Synonymous amino acid mutations in the spike protein are near deletions

To compare significantly enriched amino acid mutations between all clades in the dataset, the dataset was filtered for the minimum q-value for each mutation enriched in each clade. An important point to note is that mutations that appear highly significant across many clades could be due to comparisons with one clade. After filtering these minimum q-values for synonymous mutations in the spike protein, many of the significantly enriched synonymous amino acid mutations are near known deletions.

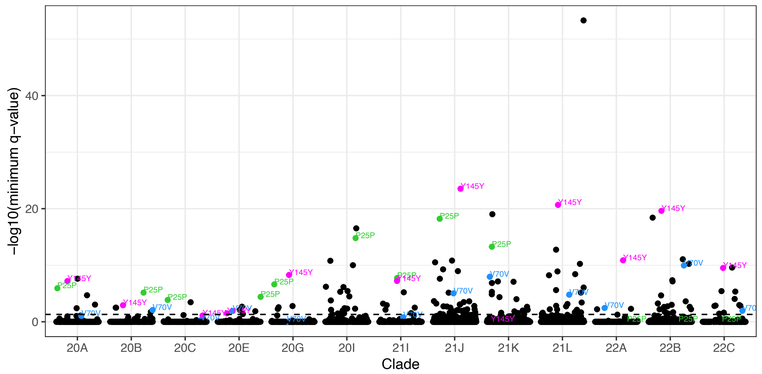

Figure 2: Minimum q-values for synonymous amino acid mutations in the spike protein. Mutations not near notable deletions are black. P25P is near L24-, P25- and P26-. V70V is near H69- and V70-. Y145Y is near Y144-. For more information visit https://covariants.org/shared-mutations.

Validate analyses with synonymous mutations

To validate the analyses generated by the computational pipeline, the proportion of significant synonymous amino acid mutations was compared to proportions generated randomly from the dataset. Because synonymous mutations do not result in a change to the amino acid sequence, these mutations are experiencing neutral selection and therefore should not be significantly enriched. Therefore, the proportion of significant synonymous mutations should be less than those randomly generated.

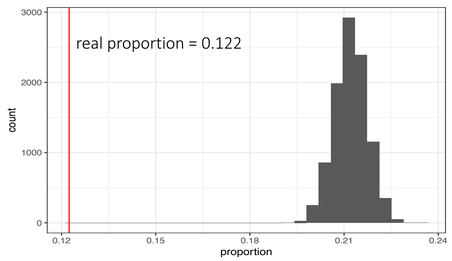

Figure 3: After 10,000 randomly generated significant synonymous proportions, the actual proportion of significant synonymous mutations in the dataset (red line = 0.122) is less than all values within the null distribution.

Conclusions

Although mutations experiencing different directions and degrees of selection can be identified using experimental approaches (Starr et al., 2022), using a computational pipeline to identify significantly enriched mutations is an effective way to analyze the entire SARS-CoV-2 genome rather than a specific region of a protein (e.g., receptor-binding domain of the spike protein). Based on the volcano plots, distantly-related clades have more significantly enriched mutations than closely-related clades, and the two proteins that contribute the largest proportion of significantly enriched mutations are ORF8 and the spike protein. Although the computationally-derived results span the entire SARS-CoV-2 genome, these results, in addition to prior knowledge, indicate that analyses should be focused on the spike protein due to its importance in virus selection. This project also determined that significantly enriched synonymous amino acid mutations in the spike protein are near known deletions. This result represents a potential caveat with the computationally-derived results – the synonymous mutations may be due to incorrect sequence alignment at the deletion site. Finally, for mutations not proximate to deletions, the proportion of significantly enriched synonymous amino acid mutations being less than the randomly generated proportions validates the results of this project.

Literature Cited

Faria NR, Mellan TA, Whittaker C, Claro IM, Candido D da S, Mishra S, Crispim MAE, Sales FCS, Hawryluk I, McCrone JT, et al. 2021. Genomics and epidemiology of the P.1 SARS-CoV-2 lineage in Manaus, Brazil. Science. 372(6544):815–821.

Hodcroft EB, Zuber M, Nadeau S, Vaughan TG, Crawford KHD, Althaus CL, Reichmuth ML, Bowen JE, Walls AC, Corti D, et al. 2021. Spread of a SARS-CoV-2 variant through Europe in the summer of 2020. Nature. 595(7869):707–712.

Naveca FG, Nascimento V, de Souza VC, Corado A de L, Nascimento F, Silva G, Costa Á, Duarte D, Pessoa K, Mejía M, et al. 2021. COVID-19 in Amazonas, Brazil, was driven by the persistence of endemic lineages and P.1 emergence. Nat Med. 27(7):1230–1238.

Neher RA. 2022. Contributions of adaptation and purifying selection to SARS-CoV-2 evolution.

Starr TN, Greaney AJ, Hannon WW, Loes AN, Hauser K, Dillen JR, Ferri E, Farrell AG, Dadonaite B, McCallum M, et al. 2022. Shifting mutational constraints in the SARS-CoV-2 receptor-binding domain during viral evolution. Science. 377(6604):420–424.

Tegally H, Moir M, Everatt J, Giovanetti M, Scheepers C, Wilkinson E, Subramoney K, Makatini Z, Moyo S, Amoako DG, et al. 2022. Emergence of SARS-CoV-2 Omicron lineages BA.4 and BA.5 in South Africa. Nat Med. 28(9):1785–1790.

Viana R, Moyo S, Amoako DG, Tegally H, Scheepers C, Althaus CL, Anyaneji UJ, Bester PA, Boni MF, Chand M, et al. 2022. Rapid epidemic expansion of the SARS-CoV-2 Omicron variant in southern Africa. Nature. 603(7902):679–686.

Volz E, Mishra S, Chand M, Barrett JC, Johnson R, Geidelberg L, Hinsley WR, Laydon DJ, Dabrera G, O'Toole Á, et al. 2021. Assessing transmissibility of SARS-CoV-2 lineage B.1.1.7 in England. Nature. 593(7858):266–269.

Acknowledgments

The SARS-CoV-2 mutation dataset used in this project was developed by the Bloom Lab at the Fred Hutchinson Cancer Center and can be accessed at *https://github.com/jbloomlab/SARS2-mut-fitness/*. Special thanks to Dr. Jesse Bloom and Will Hannon for their assistance on this project.

Fred Hutch Cancer Center

Molecular & Cellular Biology in Seattle
AN INTERDISCIPLINARY Ph.D. PROGRAM *offered through the*
UNIVERSITY OF WASHINGTON *and* FRED HUTCH