The Princeton Journal of Interdisciplinary Research
ISSN 3069-8200

Exploring the Relationship Between Dominant Variant Emergence and Sequence Divergence Throughout the Sars-COV2 (COVID-19) Pandemic

Varin Nallabothula

Redmond High School, 10735 Elliston Way NE, Redmond, Washington 98053, USA

The Princeton Journal of Interdisciplinary Research
ISSN 3069-8200

Abstract

Viruses have a unique evolutionary landscape as they compete with one another and jump to different hosts. This is an understudied area of research, but due to the ongoing pandemic and the widely available sequence data there are new possibilities for understanding how viral genome evolution occurs. This study focuses on how novel variants of COVID-19 demonstrate viral sequence divergence throughout the global pandemic. The research compares and contrasts key dominant variants such as Alpha, Delta, Omicron BA.1, and Omicron XXB with the original Wuhan strain of COVID-19. The results of the comparison between the RNA nucleotides in the different variants were analyzed to determine whether some variants became dominant due to evolution, natural selection, or mutation. The results show that the variant strains all diverged and became dominant from the original Wuhan strain. However, there was no significant pattern to account for variant strain lineages becoming dominant in non-sequential order. These results indicate that natural selection is a strong driving force for viral genome evolution.

Keywords: SARS-CoV-2; COVID-19; Genomic surveillance; Viral evolution; Sequence divergence; Dominant variants; Mutation rate; Phylogenetic analysis; Genetic variability; RNA virus.

Introduction

Viruses utilize a host organism's cell biology to propagate and spread throughout populations. One such type of virus, coronaviruses, is aptly named for its morphology,

containing a "crown" of spike proteins.[1] They are RNA-based viruses and contain relatively large genomes compared to other RNA viruses.[9] Some coronaviruses are pathogenic to humans and are capable of causing many different illnesses, ranging from the common cough to the Severe Acute Respiratory Syndrome (SARS).[6] Currently, SARS-CoV-2 (Covid-19) is causing a global pandemic that poses risk-factors to people all across the world.[10] COVID-19's ability to cause Severe Acute Respiratory Syndrome when it infects a host explains the SARS portion of its name and demonstrates how harmful infection can be to human health. Covid-19 falls under the Beta genera of Coronaviruses and contains a positive-sense and single-stranded RNA (+ssRNA). Other types of coronaviruses were recently reviewed in detail.[6]

The structure of a virus is known as a virion structure, as displayed in the diagram above. This structure has the RNA in the core, protected by a circular-shaped outer membrane envelope consisting of lipids and proteins. COVID-19's virion structure consists of 4 different proteins that contribute to the structure and improve the virus's capability, along with the viral RNA. The 4 proteins are the spike protein, membrane protein, envelope protein, and nucleocapsid protein. The spike proteins are cup-shaped proteins, pointing out of the structure's exterior. This is the protein that comes into the first contact with the host's body, essentially the protein that infects the lungs through a receptor for an enzyme called Angiotensin-Converting Enzyme 2 (ACE2), which opens up the other parts of the virion structure to take control. The coronavirus is named, because of the spike proteins. The word 'Corona' means a crown, and the spike proteins are supposed to look like crown shaped proteins, hence the name Coronavirus. The membrane protein has multiple purposes. It helps protect the nucleocapsid protein from damage, helps initiate viral assembly by stabilizing nucleocapsid proteins, and facilitates a passage responsible

for entering the viral RNA into the host cell. The envelope protein protects the nucleocapsid protein with viral RNA and helps facilitate the viral RNA's entry into the host cell, like the membrane protein. Additionally, this protein's shape helps to avoid recognition by the host's immune system and lets it pass through the different semi-permeable membranes. The nucleocapsid proteins are responsible for protecting and shielding the viral RNA while also initiating the replication of the viral RNA. This protein helps transfer the viral RNA into the host cell, and after the other proteins create an entry passage into the host cell. The viral RNA contains all the instructions for the creation of proteins. Once inside a host cell, the viral RNA will send instructions to the cellular apparatus making copies or clones of the viral RNA and synthesizing thousands of other proteins in the virion structure. Many virion structures will soon spread across the body and take over even more cells in the process. In addition to these proteins, there are different ends when looking at the genomic organization of SARS-Cov2. The starting end is the 5 prime leader sequence, and the closing end is the 3 prime terminal sequence, both containing a messenger RNA (mRNA). There is also something else called open reading frames (ORFs). A typical COVID-19 genome consists of at least 6 ORFs; the first ORF represents around 67% of the entire COVID-19 viral genome. The ORFs encode for the structural proteins in the virion structure, such as the spike, envelope, membrane, and nucleocapsid proteins in the sequences of RNA. Finally, there are proteins that are coded by the viral RNA but don't contribute to the virion structure, called non-structural proteins (nsps). In SARS-Cov2, these proteins include the enzymes the viral RNA needs to replicate when infecting the host cell, such as nsp5.

Subgenomic mRNAs are responsible for gene translation, creating a 5' and 3' co-terminal in the Covid-19 viral genome. The 5' represents the leader sequence, and the 3' represents the terminal sequence. A typically 5' leader sequence and 3' terminal sequence are located with the subgenomic mRNA. The leader and terminal ends of the COVID-19 viral genome carry tiny untranslated regions, known as UTRs. There are also many nonstructural proteins (nsps), such as RNA-dependent RNA polymerase (RdRp), coronavirus main protease (3CLpro), and papain-like protease (PLpro), among others, that are encoded by the Covid-19 viral genome.

In viruses, mutations occur when the genomic sequence or the RNA codons are altered. Mutations serve as the only source of novel heritable variation upon which natural selection can act upon. COVID-19 could never have jumped from a non-human host into humans without mutations. RNA viruses have higher mutation rates than DNA viruses, and even still, COVID-19 and other +ssRNA viruses often have higher mutation rates than other RNA-based viruses.[4] This high mutation rate, in combination with high effective population sizes of viruses, allows for rapid evolution in the face of selective pressure of host immune response. The COVID-19 pandemic comes when sequencing costs are at new lows, offering a wealth of genomic data that can be used for scientific inquiry. Throughout the pandemic, Covid-19 has continued to evolve, and novel dominant variants have rapidly spread across the globe. The present study investigates the role that sequence divergence plays in the emergence of novel COVID-19 variants, providing insight into the efficacy of selection acting on the virus, patterns of novel strain emergence, and viral evolution.

Methods

Genome sequences of dominant variant strains of COVID-19 were obtained and analyzed to investigate sequence divergence. The variant strains were downloaded from the NCBI Virus SRA database.[5] The dominant COVID-19 strain varies by date and location, so the CDC data for the United States of America was used to determine this.[3] There are often hundreds or thousands of different versions prepared for one strain of interest, so to prevent any bias coming from using data from differing submitters, all variant data besides the original Wuhan strain came from the CDC Respiratory Viruses Branch, Division of Viral Diseases by Howard D. et al. under variant accession numbers: OL315334 (Alpha B.1.1.7), OM698851 (Delta B.1.617.2), ON222717 (Omicron BA.1), OP808387 (Omicron BA.2), OP651281 (Omicron BA.2.12.1), OP999924 (Omicron BA.5), OQ344804 (Omicron BQ.1.1), OQ344799 (Omicron XBB). The Wuhan strain came from the National Center for Biotechnology Information (NCBI) by Wu F. et al. under accession number NC_045512 (First available Wuhan data). MEGA11 was used to trim all data files to the same size, conduct an alignment across all data files, and then calculate sequence divergence values (# of differing sites/total # of sites). This data and the data for when each variant became the dominant strain were collected in Excel—the values of sequence diversity were inputted into RStudio for figure creation and statistical testing.

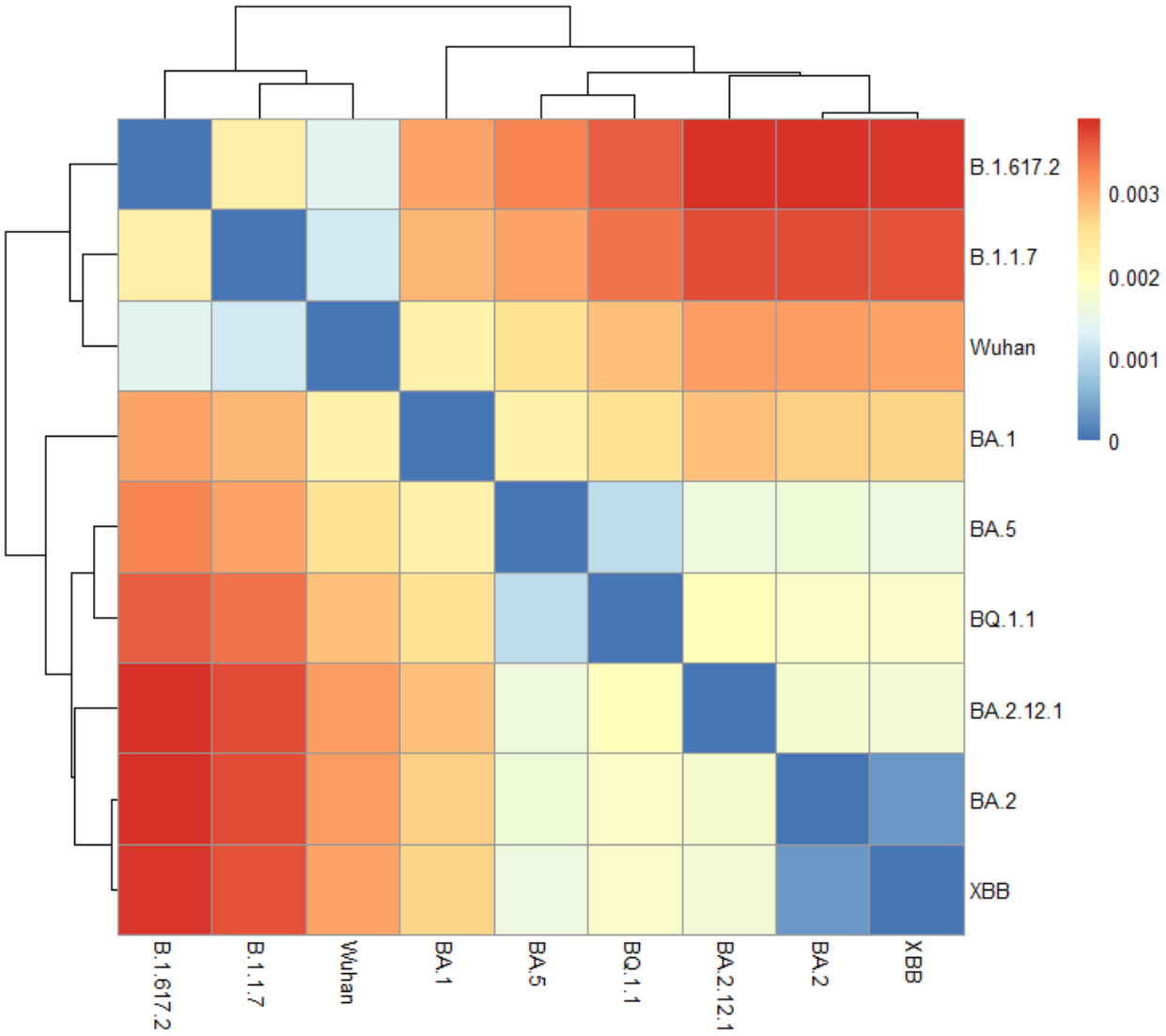
Results and Discussion

A heatmap of sequence divergence values between all strains of interest shows some interesting trends in the data (Figure 1). Notably, the Wuhan strain is an intermediate in divergence values between Alpha and Delta and the Omicron strains. Since both Alpha/Delta and the Omicron strains originate from the Wuhan variant, it makes sense that the greatest

divergence values would be seen between Alpha/Delta and the newest emerged Omicron variant (XBB). Divergence values range from 0.000313 (XBB: BA.2) to 0.00389 (BA.2.12.1:B.1.617.2), roughly an order of magnitude difference. The hierarchical clustering of the divergence data also shows notable features (Figure 2). The Omicron strains clustered separately from Wuhan, Alpha (B.1.1.7), and Delta (B.1.617.2). Some Omicron strains are more closely clustered together than others, and Wuhan and Alpha clustered more closely than Wuhan and Delta. Investigating whether the amount of time between strains explains levels of divergence is shown in two ways (Figure 3 & Figure 4). First, divergence values between Wuhan and each subsequence dominant variant were plotted against the number of days (since Wuhan's emergence) it took for each variant to become the dominant strain (Figure 3). A linear regression was performed with a p-value of 0.005945. Second, divergence values were obtained from one dominant variant to the next (in sequential order), and this was plotted against the number of days between each variant becoming dominant (Figure 4). A linear regression gave a p-value of 0.473.

Figure 1

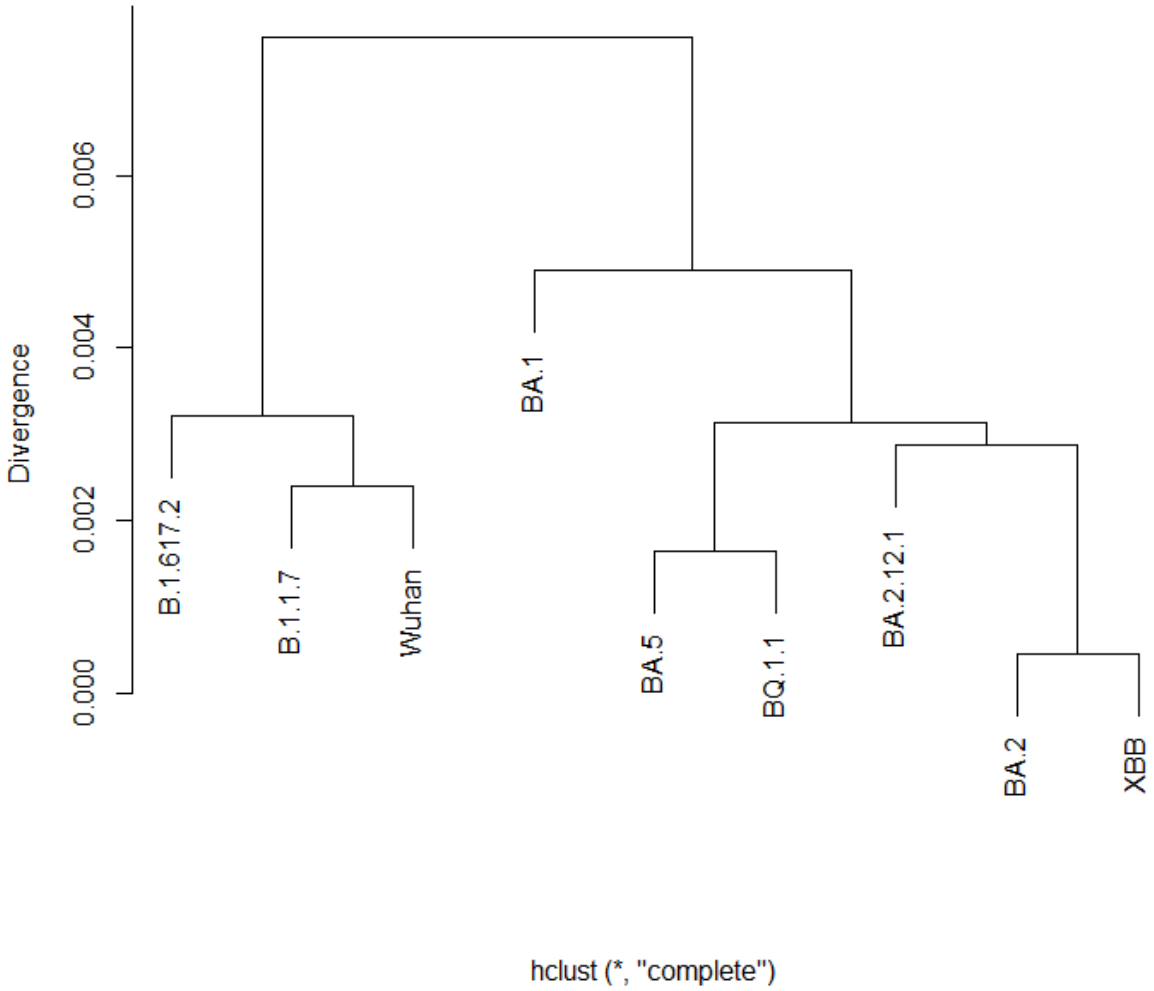
The Sequence Divergence Values Comparing Each of The Eight Variants and The Wuhan Strain Visualized Through a Heatmap.



The bluer one of the boxes is, the higher the similarity is between the two Covid-19 strains. The redder one of the boxes is, the greater the difference is between the two Covid-19 strains.

Figure 2

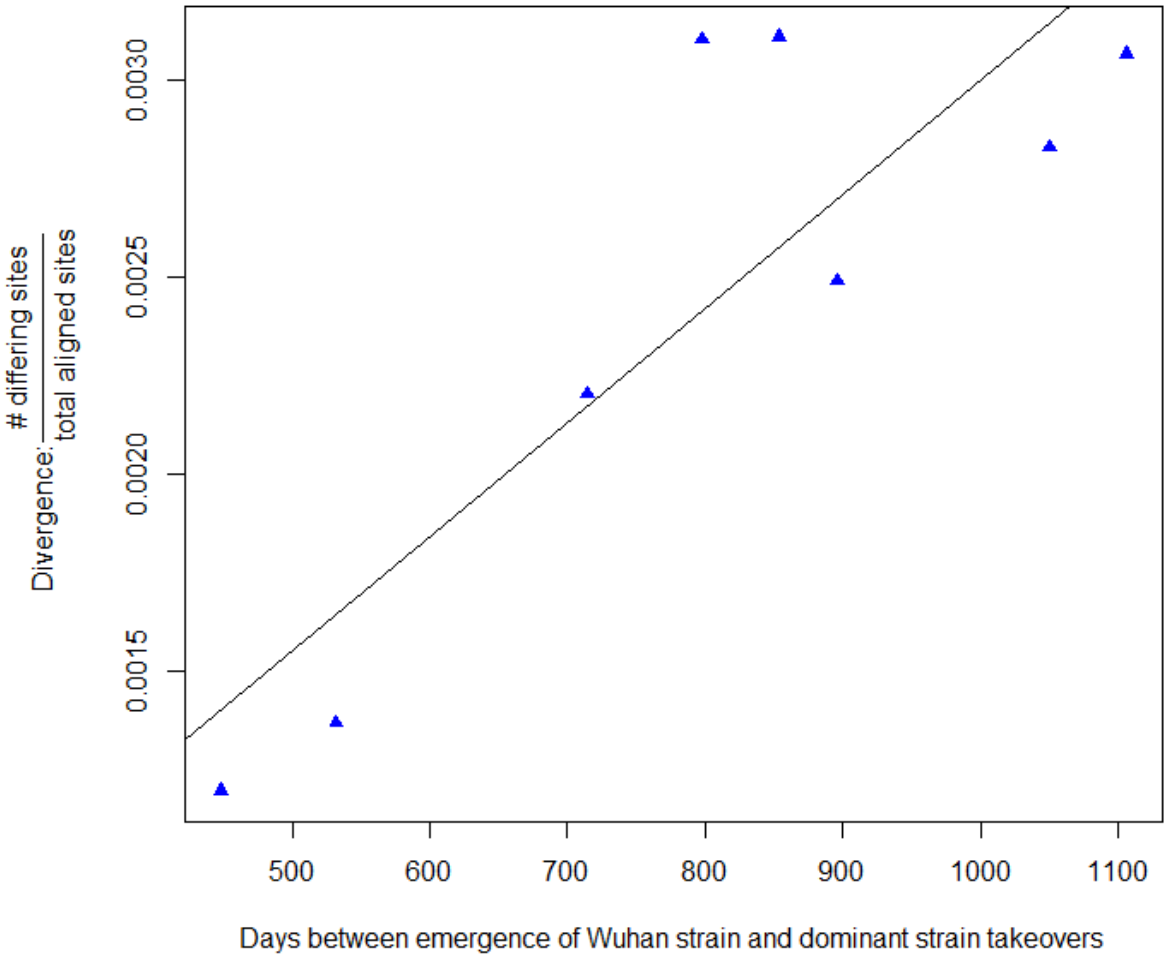
This Dendrogram Shows Relationships Between Variants Based on Their Sequence Divergence.



As expected, the Wuhan strain is most diverged from the most recently emerging Omicron XBB. Additionally, the omicron strains cluster separately from the Wuhan, Alpha (B.1.17), and Delta strains (B.1.617.2), indicating the relationship of the evolutionary timeline between these groups of strains.

Figure 3

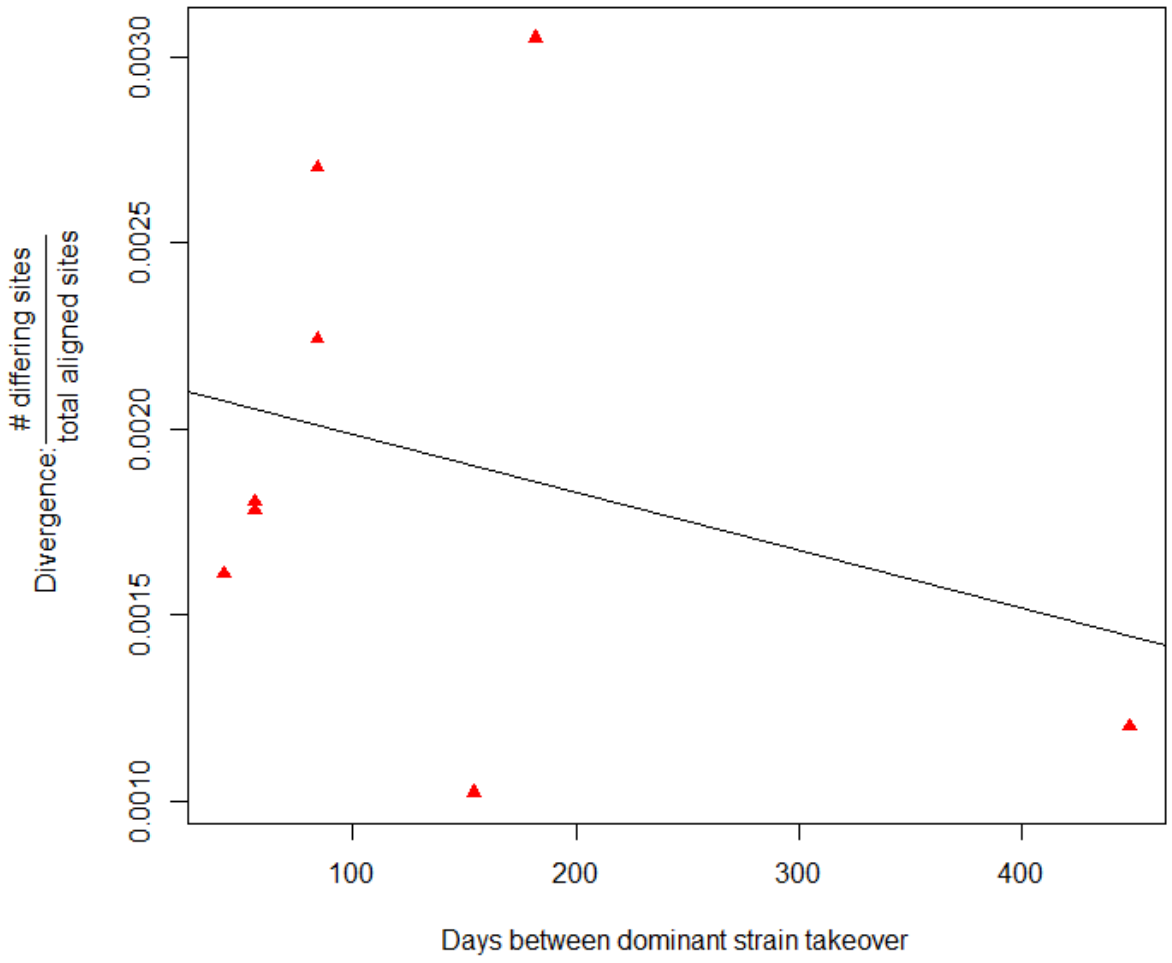
Days Between Emergence of Wuhan Strain and Dominant Strain Takeovers



This plot shows divergence values between Wuhan and all dominant variants plotted against the number of days since the emergence of the Wuhan strain it took for each variant to become dominant. Linear regression showed a significant correlation (p-value = 0.005945). This indicates that as new dominant strains emerge, they will generally be more divergent than the ancestral Wuhan variant. This also supports the general idea that virus divergence happens steadily.

Figure 4

Days Between Dominant Strain Takeover



This plot shows divergence values between each dominant strain in sequential order (Wuhan: Alpha, Alpha: Delta, Delta: Omicron BA.1, etc.) plotted against the number of days between variants becoming dominant. A linear regression did not show a significant linear correlation (p-value = 0.473). It could be argued that this provides evidence that the amount of

time between dominant variants does not explain how much divergence will occur but fails to account for different evolutionary lineages becoming dominant in non-sequential order.

The results determined that the Omicron XBB and Omicron BA.2 strains were the most closely related to each other. This finding is somewhat surprising since XBB is the most recently emerged dominant omicron strain, and BA.2 is older. This fact could imply that XBB evolved from BA.2, especially since this divergence value (0.000313) is not clustered with the others. On the other hand, it was determined that the Omicron BA.2.12.1 and B.1.617.2 strains were the furthest related strains. This result is also somewhat non-intuitive since the furthest diverged relationship does not include the Wuhan strain. However, upon further examination, the pieces add up since both the Omicron cluster of variants and the Alpha/Delta cluster have diverged away from the Wuhan strain and thus would be further from one another than from the Wuhan strain. Based on these implications, it is evident that COVID-19 is getting more divergent from the original Wuhan strain over time. This makes sense since mutations accumulate over time, and some will become fixed, especially if they are adaptive. Additionally, there is no evidence of a relationship between the amount of divergence occurring and how long it takes for a new dominant variant to arise. This result provides evidence for selection acting on Covid-19. If random mutations were accumulating in the population due to genetic drift, it would be expected that the longer it takes for a new variant to emerge, the more divergent the variant would be from the last. While variants are becoming more diverged from the Wuhan strain over time, selection purges deleterious mutations. Selection would be expected to act on a virus with such a large effective population size. According to a study conducted by researchers, the estimated effective population size of Omicron B.1.1.7 is roughly 300.[8] Based on the results, it can be inferred that

each of the Covid-19 variants analyzed is undergoing random mutations. These results show that the random mutations and periods it took for each variant to become dominant could be used to implement things such as public policy or virus prevention.

One limiting factor of this study is the number of Covid-19 variants analyzed. Although a constant stream of new dominant variants is emerging, a sample size of nine variants is ultimately not statistically rigorous. As Covid-19 continues to evolve and new variants emerge, divergence patterns may also emerge. In addition to the number of variants analyzed, more information may be identified if additional sequences were studied for each variant. Here, one sequence submission per variant was used, but there is likely some level of intra-variant sequence variation. Taking this variation into account may provide a more nuanced statistical analysis, but it was out of this project's scope.

Conclusion

A notable extension from these results involves investigating the mutations that foster inter-variant divergence. Patterns may exist at certain loci of the Covid-19 genome that explain how selection fosters divergence between variants. This may help understand the distribution of fitness effects (DFE) of mutations in COVID-19—viruses needing host cell machinery to replicate experience a nested selection hierarchy. For instance, a single mutation could benefit competition between viral particles within a host and deleterious transmission between hosts. Multilevel selection also suggests that other factors favor the most pathogenic COVID-19 strains and predict more severe disease in populations with high crowd interactions.[2] Selection acting at multiple levels may create a unique DFE of mutations that cannot be extrapolated to other organisms. Previous analyses have shown that RNA-based viruses have a low tolerance to

accumulate mutations and stay functional, making the rate of deadly mutations very common.[7] The large amount of available COVID-19 sequence data may be a great system to provide novel insight into the DFE of viruses.

In short, here, it was shown that COVID-19 is diverging over time, yet there is no consistent trend in divergence from one dominant variant to the next, implying that selection acts as a strong driving force for its evolution.

References

Almeida, J. D., & Tyrrell, D. A. (1967). The morphology of three previously uncharacterized human respiratory viruses that grow in organ culture. Journal of General Virology, 1(2), 175-178.

Blackstone, N. W., Blackstone, S. R., & Berg, A. T. (2020). Variation and multilevel selection of SARS-CoV-2. Evolution; international journal of organic evolution, 74(10), 2429–2434. https://doi.org/10.1111/evo.14080.

Centers for Disease Control and Prevention. (2023). COVID Data Tracker: Variant Proportions. Retrieved March 5, 2023, from https://covid.cdc.gov/covid-data-tracker/#variant-proportions

Duffy, S., Shackelton, L. A., & Holmes, E. C. (2008). Rates of evolutionary change in viruses: patterns and determinants. Nature Reviews Genetics, 9(4), 267-276.

Hatcher, E. L., Zhdanov, S. A., Bao, Y., Blinkova, O., Nawrocki, E. P., Ostapchuck, Y., Schaffer, A. A., & Brister, J. R. (2017). Virus Variation Resource - improved response to emergent viral outbreaks. Nucleic Acids Research, 45(D1), D482-D490. https://doi.org/10.1093/nar/gkw1065.

Kaur, N., Singh, R., Dar, Z., Bijarnia, R. K., Dhingra, N., & Kaur, T. (2021). Genetic comparison among various coronavirus strains for the identification of potential vaccine targets of SARS-CoV2. Infection, Genetics and Evolution, 89, 104490.

Sanjuán, R., Moya, A., & Elena, S. F. (2004). The distribution of fitness effects caused by single-nucleotide substitutions in an RNA virus. Proceedings of the National Academy of

Sciences of the United States of America, 101(22), 8396–8401. https://doi.org/10.1073/pnas.0400146101.

Volz, E., Mishra, S., Chand, M., Barrett, J. C., Johnson, R., Geidelberg, L., Hinsley, W. R., Laydon, D. J., Dabrera, G., O'Toole, Á., Amato, R., Ragonnet-Cronin, M., Harrison, I., Jackson, B., Ariani, C. V., Boyd, O., Loman, N. J., McCrone, J. T., Gonçalves, S., Jorgensen, D., … Ferguson, N. M. (2021). Assessing transmissibility of SARS-CoV-2 lineage B.1.1.7 in England. Nature, 593(7858), 266–269. https://doi.org/10.1038/s41586-021-03470-x.

Woo, P. C., Wang, M., Lau, S. K., Xu, H., Poon, R. W., Guo, R., ... & Yuen, K. Y. (2007). Comparative analysis of twelve genomes of three novel group 2c and group 2d coronaviruses reveals unique group and subgroup features. Journal of virology, 81(4), 1574-1585.

Zheng, J. (2020). SARS-CoV-2: an emerging coronavirus that causes a global threat. International journal of biological sciences, 16(10), 1678

The Princeton Journal of Interdisciplinary Research
ISSN 3069-8200