Temporal Patterns in the Evolutionary Genetic Distance of SARS-CoV-2 during the COVID-19 Pandemic

Jingzhi Lou^a Shi Zhao^{a, b} Lirong Cao^a Hong Zheng^a Zigui Chen^c
Renee W.Y. Chan^d Marc K.C. Chong^{a, b} Benny C.Y. Zee^{a, b} Paul K.S. Chan^c
Maggie H. Wang^{a, b}

^aJC School of Public Health and Primary Care, The Chinese University of Hong Kong, Hong Kong, China; ^bCUHK Shenzhen Research Institute, Shenzhen, China; ^cDepartment of Microbiology, The Chinese University of Hong Kong, Hong Kong, China; ^dDepartment of Paediatric, The Chinese University of Hong Kong, Hong Kong, China

Keywords

Severe acute respiratory syndrome coronavirus 2 · Genetic distance · Substitutions · Mutations · Evolution

Abstract

During coronavirus disease 2019 (COVID-19) pandemic, the genetic mutations of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) occurred frequently. Some mutations in the spike protein are considered to promote transmissibility of the virus, while the mutation patterns in other proteins are less studied and may also be important in understanding the characteristics of SARS-CoV-2. We used the sequencing data of SARS-CoV-2 strains in California to investigate the time-varying patterns of the evolutionary genetic distance. The accumulative genetic distances were quantified across different time periods and in different viral proteins. The increasing trends of genetic distance were observed in spike protein (S protein), the RNA-dependent RNA polymerase (RdRp) region and nonstructural protein 3 (nsp3) of open reading frame 1 (ORF1), and nucleocapsid protein (N protein). The genetic distances in ORF3a, ORF8, and nsp2 of ORF1 started to diverge from their original variants after Sep-

tember 2020. By contrast, mutations in other proteins appeared transiently, and no evident increasing trend was observed in the genetic distance to the original variants. This study presents distinct patterns of the SARS-CoV-2 mutations across multiple proteins from the aspect of genetic distance. Future investigation shall be conducted to study the effects of accumulative mutations on epidemics characteristics.

© 2022 The Author(s).

Published by S. Karger AG, Basel

Introduction

The coronavirus disease 2019 (COVID-19) pandemic caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) poses serious threats to public health globally. The pandemic has infected >90 million individuals around the world [1], and mutations in SARS-CoV-2 have been detected frequently. Although most mutations appeared for a short period of time, some amino acid substitutions gradually reached fixation in population locally or globally [2, 3]. As more mutations reach fixation, the circulating SARS-CoV-2 variants may di-

verge from the original variants in terms of genetic distance. The accumulation of genetic distance reflects a steady viral evolutionary process that may affect the characteristics of SARS-CoV-2. For instance, the D614G substitution in the S protein, which emerged in March 2020, was considered to associate with changes in COVID-19 infectivity at the population scale [3–6]. The recently reported 501Y lineage in the United Kingdom with the notable mutation N501Y in S protein was also shown to be more transmissible [7, 8].

The time-varying mutation patterns in other proteins may also be important to characterize the genetic evolution of SARS-CoV-2. In this study, we used a classical statistic to identify and trace the SARS-CoV-2-observed mutations in sequences from public databases and report the accumulated genetic distance across multiple proteins through time.

Materials and Methods

We collected the full-length human SARS-CoV-2 strains in California, USA, from Global Initiative on Sharing all Influenza Data (GISAID) [9]. All available strains amounting a total number of 9,081 were downloaded with the collection date ranging from January 22 to December 29, 2020; see online supplementary materials S1 for the details of daily sample size and summary (see www.karger.com/doi/10.1159/000520837 for all online suppl. material). The Wuhan-Hu-1 reference genome (GenBank NC_045512.2, GISAID EPI_ISL_402125), which was collected in 2019, was considered as the reference strain for sequence alignment and the initial (reference) variant for genetic distance calculation to other strains. Multiple sequences alignment was performed using MAFFT (version 7) [10]. The genetic distance to the initial variant was defined by the Hamming distance. To ensure a sufficient sample size in each time interval and to better observe the dynamic trend of genetic distance, a sliding window was applied (online suppl. materials S2). In this study, the window size W was fixed to be 15 days, and the step length was 3 days. We investigated the temporal variation of the accumulative genetic distance in the cross-sectional series.

Results and Discussions

Since March 2020, the average genetic distance of SARS-CoV-2 steadily increased over time compared to the initial strain (Fig. 1, top-left panel). The average genetic distance across the whole genome increased from 1.7 residues on February 25 to 12.45 residues on December 21, 2020. The increase followed a constant rate averaged 0.92 residue per month with 95% CI (0.78, 1.05) (online suppl. materials S3). Next, we analyzed the pro-

teins separately to examine which genes contributed most to the mutation accumulation. We found that the substitutions were concentrated on the S protein, the ORF1, and the N protein, with rates of increase being 0.21, 0.42, and 0.17 residues per month, respectively (Fig. 1). By further examining the ORF1 genes, we can see that its genetic distance was mostly contributed by the RdRp and the nsp2 and nsp3 regions. Each of these regions recorded 1.2 mutations on average (online suppl. material S4), among which P314L from RdRp was observed to be co-mutated with D614G on the S protein. On ORF3a and ORF8, mutation accumulation started only after September 2020, reaching 1.2 and 0.6, respectively, as of December 21, 2020. While on M, E, ORF6, ORF7, and ORF10, the average genetic distance to the Wuhan strain is nearly 0. This indicates that the membrane (M), envelope (E), and the nonstructural proteins were more conservative during the first year of SARS-CoV-2 transmission in human population using the California samples. The conservative phenomenon of these proteins was also discovered when comparing SARS-CoV-2 genomes to other SARS-related coronaviruses [11].

The accumulation of genetic distance observed in this study might be due to that the advantageous mutations usually occur in greater frequencies and will reach fixation in large populations [12]. The fluctuations of genetic distance observed in short periods might be caused by these transient deleterious mutations or by technical sequencing errors. Since some mutations that reached high frequencies locally or globally were considered to be associated with increased transmissibility [3, 4, 6, 8], the time-varying genetic distance, especially in S protein, might reflect the evolutionary process of the virus in human population.

We further computed the frequency of substitutions in the S and N protein among all the observed samples (online suppl. materials S5). For S protein, 34.4% (439 out of 1,273) of codons have ever harbored at least one residue substitution during the observation period, while 7 residue substitutions were found in >1% sequences and 6 of them were from the S1 region. For the N protein, 51.3% (215 out of 419) of codons were observed to have at least one residue switch; however, only 9 mutations were detected in >1% of sequences, 7 of which appeared in the N2a domain.

Our results showed a distinct pattern of accumulative mutations across multiple proteins of SARS-CoV-2 in terms of genetic distance against the initial strain. The continuous increasing trends were found in structural

Fig. 1. The time-varying genetic distance from the initial strain in different SARS-CoV-2 proteins. The red lines represent average genetic distance in the whole genome, S protein, ORF1, and N protein with increasing trends, while the shades represent the 95% confidence interval. The yellow dots represent genetic distance in ORF3a and ORF8 with increasing trend only after September 2020.

The blue dots represent genetic distance in M protein, E protein, ORF6, ORF7, ORF8, and ORF10 without notable trends. The red dash line indicates genetic distance at 0.5 residue. The substitution D614G emerged and reached fixation in March 2 and June 18, 2020, and N501Y emerged in December 15, 2020, in California.

protein (S and N protein) as well as some nonstructural proteins (nsp3 and RdRp). As recent studies showed that mutated functional S protein and N-terminal domain of SARS-CoV-2 conferred resistance to monoclonal antibodies [13], continuous evolution of the virus might bring considerable challenge to the development of antiviral drugs and vaccines.

Conclusion

This study presents temporal patterns of the SARS-CoV-2 mutations from the aspect of genetic distance. The continuous mutation accumulation was observed in genes encoding the S protein, N protein, and ORF3a, as well as nsp3 and RdRp in ORF1, while other proteins ap-

peared more conserved. Therefore, future investigation shall be conducted to study the effects of accumulative mutations on epidemics characteristics.

Acknowledgments

The SARS-CoV-2 sequences were collected from Global Initiative on Sharing all Influenza Data (GISAID) accessible via <https://www.gisaid.org/>. We thank the contribution of the submitting and the originating laboratories. This study was conducted using the resources of Alibaba Cloud Intelligence High Performance Cluster computing facilities, which is made free for COVID-19 research.

Statement of Ethics

The study is exempt from ethics committee approval since the human SARS-CoV-2 strains were collected via public domains.

Conflict of Interest Statement

M.H.W. is a shareholder of Beth Bioinformatics Co., Ltd. B.C.Y.Z. is a shareholder of Beth Bioinformatics Co., Ltd and Health View Bioanalytics Ltd. The other authors have no conflicts of interest to declare.

References

- 1 WHO coronavirus disease (COVID-19) dashboard 2021. Available from: <https://covid19.who.int/>.
- 2 Rambaut A, Loman N, Pybus O, Barclay W, Barrett J, Carabelli A, Connor T, et al. Preliminary genomic characterisation of an emergent SARS-CoV-2 lineage in the UK defined by a novel set of spike mutations. *Virological*. 2020 Dec 21.
- 3 Korber B, Fischer WM, Gnanakaran S, Yoon H, Theiler J, Abfalterer W, et al. Tracking changes in SARS-CoV-2 spike: evidence that D614G increases infectivity of the COVID-19 virus. *Cell*. 2020;182(4):812–27.e19.
- 4 Gobeil SM, Janowska K, McDowell S, Mansouri K, Parks R, Manne K, et al. D614G mutation alters SARS-CoV-2 spike conformation and enhances protease cleavage at the S1/S2 junction. *Cell Rep*. 2020;34(2):108630.
- 5 Volz E, Hill V, McCrone JT, Price A, Jorgensen D, O'Toole Á, et al. Evaluating the effects of SARS-CoV-2 spike mutation D614G on transmissibility and pathogenicity. *Cell*. 2021;184(1):64–75.e11.
- 6 Zhao S, Lou J, Cao L, Zheng H, Chong MK, Chen Z, et al. Modelling the association between COVID-19 transmissibility and D614G substitution in SARS-CoV-2 spike protein: using the surveillance data in California as an example. *Theor Biol Med Model*. 2021 Mar 9; 18(1):10.
- 7 Leung K, Shum MH, Leung GM, Lam TT, Wu JT. Early transmissibility assessment of the N501Y mutant strains of SARS-CoV-2 in the United Kingdom, October to November 2020. *Euro Surveill*. 2021;26(1):2002106.
- 8 Zhao S, Lou J, Cao L, Zheng H, Chong MK, Chen Z, et al. Quantifying the transmission advantage associated with N501Y substitution of SARS-CoV-2 in the United Kingdom: an early data-driven analysis. *J Travel Med*. 2021;28(2):taab011.
- 9 Shu Y, McCauley J. GISAID: global initiative on sharing all influenza data – from vision to reality. *Euro Surveill*. 2017;22(13):30494.
- 10 Katoh K, Rozewicki J, Yamada KD. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Brief Bioinform*. 2019;20(4):1160–6.
- 11 Tang X, Xu C, Li X, Song Y, Yao X, Wu X, et al. On the origin and continuing evolution of SARS-CoV-2. *Natl Sci Rev*. 2020;7:1012–23.
- 12 van Dorp L, Richard D, Tan CCS, Shaw LP, Acman M, Balloux F. No evidence for increased transmissibility from recurrent mutations in SARS-CoV-2. *Nat Commun*. 2020; 11(1):5986.
- 13 Weisblum Y, Schmidt F, Zhang F, DaSilva J, Poston D, Lorenzi JC, et al. Escape from neutralizing antibodies by SARS-CoV-2 spike protein variants. *Elife*. 2020;9:e61312.

Funding Sources

This work was supported by the Food and Health Bureau, The Government of the Hong Kong Special Administrative Region (COVID190103) and (INF-CUHK-1) of Hong Kong SAR, China, and partially supported by the National Natural Science Foundation of China (NSFC) (31871340) and CUHK Direct Grant (2020.025). The funding agencies had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication.

Author Contributions

M.H.W. and J.L. conceived the study. J.L. and H.Z. collected the data. J.L. carried out the analysis and drafted the first manuscript. J.L., S.Z., L.C., and M.H.W. discussed the results and edited the manuscript. All authors critically read and revised the manuscript and gave final approval for publication.

Data Availability Statement

All data used in this work were publicly available from GISAID at <https://www.gisaid.org/>. A full list of accessions used in this study is provided in online supplementary data together with the acknowledgments of all originating and submitting laboratories.