

1 **Title:** Temporal patterns in the evolutionary genetic distance of SARS-CoV-2 during the COVID-19
2 pandemic

3 **Running Title:** Mutation accumulation in SARS-CoV-2

4 **Keywords:** SARS-CoV-2; genetic distance; whole genome sequencing

5

6 **Authors:**

7 Jingzhi Lou^{1,+}, Shi Zhao^{1,2,+}, Lirong Cao^{1,+}, Zigui Chen³, Renee WY Chan⁴, Marc KC Chong^{1,2},
8 Benny CY Zee^{1,2}, Paul KS Chan³, and Maggie H Wang^{1,2*}

9 **Author affiliation:**

10 **1** JC School of Public Health and Primary Care, Chinese University of Hong Kong, Hong Kong,
11 China

12 **2** CUHK Shenzhen Research Institute, Shenzhen, China

13 **3** Department of Microbiology, the Chinese University of Hong Kong, Hong Kong SAR, China

14 **4** Department of Paediatric, the Chinese University of Hong Kong, Hong Kong SAR, China

15 + Joint first authors.

16 * Correspondence to: maggiew@cuhk.edu.hk (MHW)

17 **Email addresses of all authors**

18 JL: jzlou@qq.com

19 SZ: zhaoshi.cmsa@gmail.com

20 LC: caolr@link.cuhk.edu.hk

21 ZC: zigui.chen@cuhk.edu.hk

22 RWYC: reneewy@cuhk.edu.hk

23 MKCC: marc@cuhk.edu.hk

24 BCYZ: bzee@cuhk.edu.hk

25 PKSC: paulkschan@cuhk.edu.hk

26 MHW: maggiew@cuhk.edu.hk

27

28 **Abstract:**

29 **Background:** During the pandemic of coronavirus disease 2019 (COVID-19), the genetic mutations
30 occurred in severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) cumulatively or
31 sporadically. In this study, we employed a computational approach to identify and trace the emerging
32 patterns of the SARS-CoV-2 mutations, and quantify accumulative genetic distance across different
33 periods and proteins.

34 **Methods:** Full-length human SARS-CoV-2 strains in United Kingdom were collected. We
35 investigated the temporal variation in the evolutionary genetic distance defined by the Hamming
36 distance since the start of COVID-19 pandemic.

37 **Findings:** Our results showed that the SARS-CoV-2 was in the process of continuous evolution,
38 mainly involved in spike protein (S protein), the RNA-dependent RNA polymerase (RdRp) region of
39 open reading frame 1 (ORF1) and nucleocapsid protein (N protein). By contrast, mutations in other
40 proteins were sporadic and genetic distance to the initial sequenced strain did not show an increasing
41 trend.

42

43 **Introduction:**

44 The pandemic of coronavirus disease 2019 (COVID-19) caused by the severe acute respiratory
45 syndrome coronavirus 2 (SARS-CoV-2) poses severe threat to public health globally. The genetic
46 mutations in SARS-CoV-2 have been detected frequently. Although most mutations occur
47 sporadically and are purged shortly, it appears that some mutations gradually reach fixation [1].
48 Given increasing numbers of fixed mutations, the circulating SARS-CoV-2 strains may diverge from
49 the original strain in terms of an increasing accumulative genetic distance. The accumulation of
50 genetic distance reflects a steady viral evolutionary process that may affect the characteristics of
51 SARS-CoV-2, immune recognition, and effectiveness of antivirals targeting the pathogen. In this
52 study, we employed a computational approach to identify and trace the emerging patterns of the
53 SARS-CoV-2 mutations, and quantify accumulative genetic distance across different periods and
54 proteins.

55 **Data and methods:**

56 The full-length human SARS-CoV-2 strains in United Kingdom were obtained from Global Initiative
57 on Sharing all Influenza Data (GISAID) [2] on September 23, 2020. A total of 40,527 strains were
58 collected with the collection date ranging from January 27 to September 14, 2020. We used a
59 stratified sampling scheme to randomly selected sequences in biweekly time interval. See
60 Supplementary Materials S1 for the details of the sampling scheme and summary. As one of the first
61 reported sequences, 'China/Wuhan-Hu-1/2019', which was collected in December 31, 2019, was
62 considered as the reference strain for sequence alignment, and as the initial strain for genetic distance
63 calculation against other strains. The genetic distance to the initial strain was defined by the
64 Hamming distance. We investigated the temporal variation in the accumulative genetic distance in
65 the cross-sectional series. To better observe the dynamic trend of genetic distance, a sliding window
66 was applied (Supplementary Materials S2). Multiple sequences alignment was performed using
67 MEGA-X (version 10.1.8).

68 **Results and discussions:**

69 We observed that the genetic distance to the initial strain steadily increased with time since February
70 2020 (Figure 1). The overall distance of all proteins rose from 0 to a peaking level at 9.27 codons on

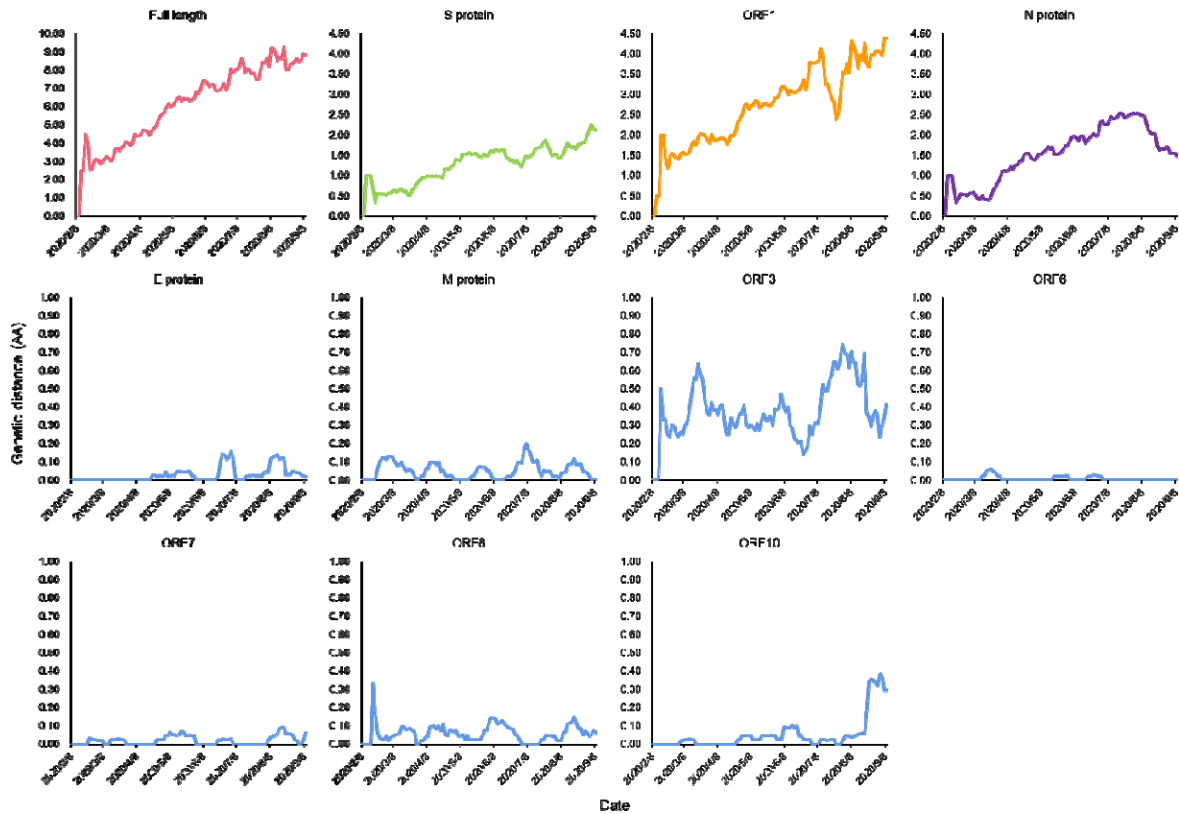
71 August 21, 2020, which implied mutations with evolutionary advantages occurred and gradually
72 accumulated. Moreover, deleterious mutants can only preserve for a short period in virus population
73 and disappeared due to functional issues or less adaption to environment, which explains the
74 fluctuations in the genetic distance curve, e.g. open reading frame 8 (ORF8) in Figure 1. By further
75 observing the distance in each protein, we found continuous increasing trends in the spike protein (S
76 protein), the ORF1 (especially in the RNA-dependent RNA polymerase, Supplementary Materials S3)
77 and the nucleocapsid protein (N protein), and no obvious trends in the membrane protein (M protein),
78 the envelope protein (E protein), the ORF3, ORF6-8 and ORF10, see Figure 1. It suggested a
79 stronger natural selection pressure on the S protein, the ORF1 and the N protein. The S protein is
80 responsible for receptor recognition and membrane fusion, and contains a receptor binding domain
81 (RBD), which is considered as the target for the SARS-CoV-2 vaccines under development[3-5].
82 The RdRp in ORF1 plays a central role in the replication and transcription cycle of SARS-CoV-2 and
83 thus is considered as a target for nucleotide analog antiviral inhibitors[6]. N protein is a
84 multifunctional and highly immunogenic determinant, whose function is mediating the packaging of
85 the viral RNA genome into the nascent virion[7]. These three proteins were critical in determining
86 the course of transmission, infection and reproduction of the virus, and thus accumulated mutations
87 in these proteins might be related to the viral adaptation to the environment both in vivo and in vitro.
88 By contrast, genetic distances of the proteins with sporadic mutation were usually below 0.5 codon
89 without demonstration of an increasing trend. To date, the mutations in these proteins have not
90 maintained, which might be due to their less competitive strength in facilitating the viral adaptation
91 to the environment.

92 Our results showed that the SARS-CoV-2 was in the process of continuous evolution, mainly
93 involved in the S protein, the RdRp region of ORF1 and the N protein. In August 24, 2020, the first
94 cases of COVID-19 re-infection was officially identified and reported that the viruses corresponding
95 to the first and second infections carried 13 amino acid differences [8]. Another study showed that
96 functional S protein and N-terminal domain of SARS-CoV-2 with mutations conferred resistance to
97 monoclonal antibodies [9]. Continuous evolution of the virus might bring considerable challenge to
98 the development of antiviral drugs and vaccines.

99 **Conclusion:**

100 This study presents the evolutionary process of SARS-CoV-2 virus from the aspect of genetic
101 distance. The continuous mutation accumulation was observed in genes encoding the S protein, the
102 RdRp region and the N protein, but not be observed in genes encoding other proteins to date.
103 Therefore, future investigation is warranted to study the characteristics and the effects associated
104 with the accumulative mutations in SARS-CoV-2, as well as the treatment or control strategies.

105



106

107 **Figure 1. The time-varying genetic distance from initial strain in different proteins.**

108 The green, orange and purple lines represent genetic distance in S protein, ORF1 and N protein with
109 increasing trends, respectively. The blue lines represent genetic distance in M protein, E protein,
110 ORF3, ORF6, ORF7, ORF8 and ORF10 without notable trends.

111 **Declarations**

112 **Authors' contributions**

113 MHW conceived the study. JL collected the data and carried out the analysis. JL, SZ, LC and MHW
114 discussed the results. JL drafted the first manuscript. JL, SZ, LC and MHW reviewed and edited the
115 manuscript. All authors critically read and revised the manuscript and gave final approval for
116 publication.

117 **Funding**

118 This work is supported by the Health and Medical Research Fund (HMRF) Commissioned Research
119 on COVID-19 [COVID190103] and [INF-CUHK-1] of Hong Kong SAR, China, and partially
120 supported by the National Natural Science Foundation of China (NSFC) [31871340] and CUHK
121 Direct Grant [4054524].

122 **Disclaimer**

123 The funding agencies had no role in the design and conduct of the study; collection, management,
124 analysis, and interpretation of the data; preparation, review, or approval of the manuscript; or
125 decision to submit the manuscript for publication.

126 **Acknowledgements**

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

132 **Conflict of interests**

133 MHW is a shareholder of Beth Bioinformatics Co., Ltd. BCYZ is a shareholder of Beth
134 Bioinformatics Co., Ltd and Health View Bioanalytics Ltd. Other authors declared no competing
135 interests.

136 **Ethics approval and consent to participate**

137 The human SARS-CoV-2 strains were collected via public domains, and thus neither ethical
138 approval nor individual consent was not applicable.

139 **Availability of materials**

140 All data used in this work were publicly available.

141 **References:**

- 142 1. Benvenuto D, Demir AB, Giovanetti M, Bianchi M, Angeletti S, Pascarella S, et al. Evidence for mutations in
143 SARS-CoV-2 Italian isolates potentially affecting virus transmission. *Journal of medical virology*. 2020. Epub
144 2020/06/04. doi: 10.1002/jmv.26104. PubMed PMID: 32492183; PubMed Central PMCID: PMC7300971.
- 145 2. Shu Y, McCauley J. GISAID: Global initiative on sharing all influenza data - from vision to reality. *Euro*
146 *surveillance : bulletin European sur les maladies transmissibles = European communicable disease bulletin*. 2017;22(13).
147 Epub 2017/04/07. doi: 10.2807/1560-7917.Es.2017.22.13.30494. PubMed PMID: 28382917; PubMed Central PMCID:
148 PMC5388101.
- 149 3. Jackson LA, Anderson EJ, Roupael NG, Roberts PC, Makhene M, Coler RN, et al. An mRNA Vaccine against
150 SARS-CoV-2 - Preliminary Report. *The New England journal of medicine*. 2020. Epub 2020/07/15. doi:
151 10.1056/NEJMoa2022483. PubMed PMID: 32663912; PubMed Central PMCID: PMC7377258.
- 152 4. Mercado NB, Zahn R, Wegmann F, Loos C, Chandrashekar A, Yu J, et al. Single-shot Ad26 vaccine protects
153 against SARS-CoV-2 in rhesus macaques. *Nature*. 2020. Epub 2020/07/31. doi: 10.1038/s41586-020-2607-z. PubMed
154 PMID: 32731257.
- 155 5. Wang Q, Zhang Y, Wu L, Niu S, Song C, Zhang Z, et al. Structural and Functional Basis of SARS-CoV-2 Entry by
156 Using Human ACE2. *Cell*. 2020;181(4):894-904.e9. Epub 2020/04/11. doi: 10.1016/j.cell.2020.03.045. PubMed PMID:
157 32275855; PubMed Central PMCID: PMC7144619.
- 158 6. Gao Y, Yan L, Huang Y, Liu F, Zhao Y, Cao L, et al. Structure of the RNA-dependent RNA polymerase from
159 COVID-19 virus. *Science (New York, NY)*. 2020;368(6492):779-82. Epub 2020/04/12. doi: 10.1126/science.abb7498.
160 PubMed PMID: 32277040; PubMed Central PMCID: PMC7164392.
- 161 7. Zinzula L, Basquin J, Bohn S, Beck F, Klumpe S, Pfeifer G, et al. High-resolution structure and biophysical
162 characterization of the nucleocapsid phosphoprotein dimerization domain from the Covid-19 severe acute respiratory
163 syndrome coronavirus 2. *Biochemical and Biophysical Research Communications*. 2020. Epub 2020/10/12. doi:
164 10.1016/j.bbrc.2020.09.131. PubMed PMID: 33039147; PubMed Central PMCID: PMC7532810.
- 165 8. To KK, Hung IF, Ip JD, Chu AW, Chan WM, Tam AR, et al. COVID-19 re-infection by a phylogenetically distinct
166 SARS-coronavirus-2 strain confirmed by whole genome sequencing. *Clinical infectious diseases : an official publication*
167 *of the Infectious Diseases Society of America*. 2020. Epub 2020/08/26. doi: 10.1093/cid/ciaa1275. PubMed PMID:
168 32840608; PubMed Central PMCID: PMC7499500.
- 169 9. Weisblum Y, Schmidt F, Zhang F, DaSilva J, Poston D, Lorenzi JCC, et al. Escape from neutralizing antibodies by
170 SARS-CoV-2 spike protein variants. 2020:2020.07.21.214759. doi: 10.1101/2020.07.21.214759 %J bioRxiv.

171