Original Article

Severity of COVID-19 is inversely correlated with increased number counts of non-synonymous mutations in Tokyo

Authors

Kodai Abe,[1,13] Yasuaki Kabe,[2] Susumu Uchiyama,[3,4,5] Yuka W. Iwasaki,[6] Hirotsugu Ishizu,[6] Yoshifumi Uwamino,[7,11] Toshiki Takenouchi,[8] Shunsuke Uno,[7] Makoto Ishii,[9] Mitsuru Murata,[10] Naoki Hasegawa,[7] Hideyuki Saya,[11] Yuko Kitagawa,[1] Koichi Fukunaga,[9] Masayuki Amagai,[12] Haruhiko Siomi,[6] Makoto Suematsu,[2] Kenjiro Kosaki,[13] Keio Donner Project

(KA and YK equally contributed to this work)

1 Department of Surgery, Keio University School of Medicine, Tokyo, Japan

2 Department of Biochemistry, Keio University School of Medicine, Tokyo, Japan

3 Department of Biotechnology, Graduate School of Engineering, Osaka University School of Medicine, Osaka, Japan

4 Exploratory Research Center on Life and Living Systems (ExCELLS), National Institutes of Natural Sciences, Okazaki, Japan.

5 Institute for Integrated Radiation and Nuclear Science, Kyoto University, Osaka, Japan.

6 Department of Molecular Biology, Keio University School of Medicine, Tokyo, Japan

7 Division of Infection Diseases and Infection Control, Keio University Hospital, Tokyo, Japan: Department of Infectious Diseases, Keio University School of Medicine, Tokyo, Japan

8 Department of Pediatrics, Keio University School of Medicine, Tokyo, Japan

9 Department of Internal Medicine, Keio University School of Medicine, Tokyo, Japan

10 Department of Laboratory Medicine, Keio University School of Medicine, Tokyo, Japan

11 Division of Gene Regulation, Institute for Advanced Medical Research, Keio University School of Medicine, Tokyo, Japan

12 Department of Dermatology, Keio University School of Medicine, Tokyo, Japan

13 Center for Medical Genetics, Keio University School of Medicine, Tokyo, Japan

Running title: COVID-19 severity and number of non-synonymous mutations

Corresponding author:

Kenjiro Kosaki, M.D., Ph.D.

Professor and Chair

Center for Medical Genetics, Keio University School of Medicine, Tokyo, Japan

35 Shinanomachi, Shinjuku-ku

Tokyo, 160-8582, Japan

Phone: +81-3-5363-3890

e-mail: kkosaki@keio.jp

Co-corresponding author

Makoto Suematsu, MD, PhD

Professor and Chair

Department of Biochemistry, Keio University School of Medicine, Tokyo, Japan

E-mail: gasbiology@keio.jp

Keywords: SARS-CoV-2, COVID-19, non-synonymous mutation, $3CL^{pro}$, Pro108Ser

Abstract

Background: SARS-CoV-2 genome accumulates point mutations in a constant manner. Whether the accumulation of point mutations is correlated with milder manifestations of COVID-19 remains unknown.

Methods: We performed SARS-CoV-2 genome sequencing in 90 patients with COVID-19 infection treated at a tertiary medical center in Tokyo between March and August 2020. The possible association between disease severity and viral haplotype was then assessed by counting the number of mutations in addition to performing phylogenic tree analysis, comparative amino acid sequence analysis among β-coronaviruses, and mathematical prediction of the functional relevance of amino acid substitutions.

Results: The number of non-synonymous mutations was inversely correlated with COVID-19 severity, as defined by requiring oxygen supplementation. Phylogenic tree analysis identified two predominant groups which were separated by a set of 6 single nucleotide substitutions, including four leading to non-synonymous amino acid substitutions. Among those four, Pro108Ser in 3 chymotrypsin-like protease ($3CL^{pro}$) and Pro151Leu in nucleocapsid protein occurred at conserved locations and were predicted to be deleterious. Patients with Pro108Ser in $3CL^{pro}$ and Pro151Leu in nucleocapsid protein

had a lower odds ratio for developing hypoxia requiring supplemental oxygen (odds ratio of 0.24 [95% confidence interval of 0.07-0.88, *P*-value = 0.032]) after adjustments for age and sex, compared with patients lacking this haplotype in Clade 20B.

Conclusion: Viral genome sequencing in 90 patients treated in the Tokyo Metropolitan area showed that the accumulation of point mutations, including Pro108Ser in $3CL^{pro}$ and Pro151Leu in nucleocapsid protein, was inversely correlated with COVID-19 severity. Further *in vitro* research is awaited.

Introduction

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), which is causing the coronavirus disease 2019 (COVID-19) pandemic, is spreading around the world. As an RNA virus with limited fidelity for genome replication, the SARS-CoV-2 viral genome accumulates mutations in a constant manner at an average of two nucleotides per months (GISAIDs: http://www.gisaid.org/), with the exception of large rearrangements such as a deletion spanning 382 nucleotides.[1,2] Although the 382-nucleotide deletion was associated with a milder clinical course, whether the accumulation of point mutations is associated with a better prognosis remains unknown.

Keio University Hospital, which has a catchment that includes the Tokyo Metropolitan area and surrounding prefectures, has been performing whole viral genome sequencing of SARS-CoV-2 in COVID-19 patients since March 2020 to characterize healthcare-associated infections rapidly and effectively and to prevent the spread of infection.[3] Molecular viral genome sequencing studies have indicated that the number of mutations is indeed increasing.[4,5] According to a Japanese governmental report, the daily number of newly identified COVID-19 patients in Tokyo plateaued during the time period when restrictions were imposed on foreign entry to Japan, while the number of seriously ill patients has been decreasing since May in

Tokyo, Japan (https://www.mhlw.go.jp/stf/covid-19/kokunainohasseijoukyou.html). Hence, the relative fraction of seriously ill patients has been decreasing. Actually, the mortality rates in Japan appear to be lower than those in Western countries (https://covid19.who.int/).[6] We hypothesized that the accumulation of mutations may have contributed to the decrease in clinical virulence.

Methods

Study design and participants.

From March 17 to August 31, 2020, 187 patients with a reverse transcription polymerase chain reaction (RT-PCR)-positive result who had been diagnosed as having COVID-19 at Keio University Hospital between March 17th and Augst 31th, 2020, were enrolled in the present study. Of these 187 patients, 134 COVID-19 patients underwent whole viral genome sequencing. After the exclusion of 44 patients in whom only partial genome sequences were obtained because of insufficient PCR amplification, 90 patients were included in the present study (Supplementary Figure 1). Thirty-two of these 90 patients were previously reported in an article describing their viral genome sequencing results.[3] The medical records of all the patients were reviewed, and data were obtained

regarding both the clinical characteristics of the eligible patients and the treatments that they received. PCR data obtained from samples collected from the nasopharynx, sputum or saliva were also collected. The present study protocol was approved by the ethics committee of the Keio University School of Medicine (approval number: 20200062) and was conducted in accordance with the Declaration of Helsinki.

Definitions and classification

The definitions of healthcare-associated SARS-CoV-2 transmission are shown in the Supplementary Table 1. We first divided the subjects into healthcare workers or patients and then subdivided the non-healthcare workers into those with "community onset" and those with "hospital onset" to detect nosocomial infection (Supplementary Figure 2).[7] Next, we classified the patients according to the grade of disease severity based on their symptoms or conditions as outlined in the clinical management guidelines from the World Health Organization[8] and the Ministry of Health, Labour, and Welfare in Japan (https://www.mhlw.go.jp/content/000650160.pdf). In some COVID-19 patients with mild or moderate symptoms, the presence of pneumonia could not be determined because they did not receive either chest X-ray or computed tomography examinations. Therefore, we

classified disease severity as follows: “Mild to Moderate,” patients did not require oxygen administration; “Severe,” patients required oxygen supplementation but did not require a ventilator; and “Critical,” patients who developed sepsis or acute respiratory distress syndrome and required a ventilator (Supplementary Table 2).[8]

DNA sequencing method

The whole viral genome sequences were determined as previously reported (Takenouchi et al.).[3] PCR-based amplification was performed using the Artic ncov-2019 primers, version 3 (https://github.com/artic-network/artic-ncov2019/blob/master/primer_schemes/nCoV-2019/V3/nCoV-2019.tsv) in two multiplex reactions according to the globally accepted “nCoV-2019 sequencing protocol” (https://www.protocols.io/view/ncov-2019-sequencingprotocol-bbmuik6w).[3] The sequencing library for amplicon sequencing was prepared using the Next Ultra II DNA Library Prep Kit for Illumina (New England Biolabs). Paired end sequencing was performed on the MiSeq platform (Illumina, CA). The bioinformatic pipeline used in this

study, "Mutation calling pipeline for amplicon-based sequencing of the SARS-CoV-2 viral genome," is available at https://cmg.med.keio.ac.jp/sars-cov-2/.

All point mutations including non-synonymous and synonymous mutations were annotated by the ANNOVAR software and were assessed using VarSifter (https://research.nhgri.nih.gov/software/VarSifter/), and the multiple amino acid sequence alignment of various β-coronaviruses was performed using Molecular Evolutionary Genetic Analysis software (MEGA, https://www.megasoftware.net/) (Supplementary Table 3). The effect of amino acid substitution was evaluated using the Protein Variation Effect Analyzer (PROVEAN v1.1.3, http://provean.jcvi.org/seq_submit.php). Scores less than a threshold value of -2.50 were considered as "deleterious." Finally, we used the NextStrain database 58 (https://nextstrain.org/) to visualize the SARS-CoV-2 clade identity.

Protein structure modeling and stabilization anaylsis

The three-dimensional (3D) structure was visualized using PyMol v2.4. (https://pymol.org/2/) based on publicly available SARS-CoV-2 protein structure coordinate for the dimer of $3CL^{pro}$ (protein data bank ID "6LU7") and the tetramer of

nucleocapsid protein (protein data bank ID "6VYO"). We then used the DynaMut server (http://biosig.unimelb.edu.au/dynamut/) to infer the effect of non-synonymous mutations on the $3CL^{pro}$ and nucleocapsid protein 3D structures in terms of molecular stabilities, and flexibilities.[9] We also estimated the differences of free energy change ($\Delta\Delta G$) and vibrational entropy change ($\Delta\Delta S$Vib) accompanied by the mutation using DynaMut,[10] which implements ENCoM reports on the impact on protein stability and flexibility accompanied by mutations in the wild-type structure. Structural changes, such as changes in cavity volume, packing density, and accessible surface area, are correlated with $\Delta\Delta G$, thereby $\Delta\Delta G$ can be used as an indicator of the impact of a mutation on protein stability.[11] A $\Delta\Delta G$ value of less than zero indicates that the mutation causes destabilization, while a $\Delta\Delta G$ above zero indicates protein stabilization.

Statistical analysis

The main parameter of this study was the grade of severity. Comparisons of categorical variables between two groups were assessed using the Fisher exact test. A Student *t*-test was used to compare abnormally distributed quantitative variables between two groups, and a Jonckheere-Terpstra test was used to analyze the tendency among three

groups. An exact logistic regression analysis was used to examine the relationship between the Clade 20B-T and the need for supplemental oxygen. The following covariates were considered for inclusion in the multivariate model: age group (<65 years or ≧65 years), sex, and infection group (Clade 20B-T or 20B-nonT). Statistical analyses were performed using R (version 3.6.2), and all the statistical tests were two-sided. A *P*- value <0.05 was considered significant.

Results

Viral genome sequence analysis

One hundred and eighty seven positive RT-PCR results were obtained during the study period at Keio University Hospital between March 17 and August 31, 2020, 134 samples (71.7 %) had residual samples available for viral genome sequencing (Supplementary Figure 1). The complete genome sequences of SARS-CoV-2 were determined in samples from 90 (67.2 %) of these individuals. A total of 70 viral haplotypes were observed among these 90 individuals. A mean of 11.8 mutations (Standard deviation [SD], 3.1]) separated the lineage from the founding Wuhan haplotype (the central haplotype of clade A). None of the strains had truncating mutations

(frameshift mutations, non-sense mutations). The number counts of non-synonymous mutations among the strains varied from 2 to 12 (mean, 7.5 [SD], 2.4]), compared with the Wuhan reference strain. The functional relevance of the non-synonymous mutations was predicted using the computer algorithm PROVEAN, the calculations of which are not dependent on sequence conservation among animals.

Clinical backgrounds of COVID-19 patients.

The clinical characteristics of the 90 patients whose complete viral genome sequences were obtained, are shown in the Supplementary Table 4. Nineteen patients (23.3%) required supplemental oxygen, and 8 (8.9%) developed acute respiratory distress syndrome; five of the eight patients died.

Number counts of non-synonymous mutations of SARS-CoV-2 and severity of COVID-19.

The number counts of thr non-synonymous mutations in SARS-CoV-2 found in COVID-19 patients who did and did not require supplemental oxygen were compared. The number counts of non-synonymous mutations was significantly higher among the

COVID-19 patients who did not require supplemental oxygen (Figure 1a: mean, 7.9 [SD 2.2] vs. 5.9 [SD 2.2], *P* value < 0.001). The number counts of non-synonymous mutations with deleterious PROVEAN scores was also higher among patients who did not require supplemental oxygen (mean, 1.5 [SD, 1.1] vs. 0.9 [SD 0.9], *P* value = 0.016). The number of non-synonymous mutations increased as the severity of the disease degraded (Figure 1b: JT = 404, *P* value < 0.001). The number of non-synonymous mutations with a deleterious PROVEAN score (< -2.50) among patients with COVID-19 also increased as the severity of the disease decreased (JT = 556, *P* value = 0.035).

Phylogenic tree analysis

We tested whether any of the phylogenic clade containing any of the non-synonymous mutations might contribute to a milder clinical course. The overall genetic diversity was relatively low, presumably because of effective international border restrictions and successful quarantine efforts. A divergent tree analysis of the complete viral genome sequences from the 90 patients and classification according to the internationally recommended nomenclature (GISAIDs: http://www.gisaid.org/) showed that most (i.e., 87) patients had strains derived from Clade 20B (Figure 2a).[12] The

remaining 5 patients belonged to Clade 19A, which lacks the Asp614Gly mutation in the Spike protein. Because the Asp614Gly mutation in the spike protein is known to be functionally relevant,[13] the 5 patients infected with Clade 19A viruses were excluded from further study.

Patients from Clade 20B (N = 85) were additionally divided into two subgroups by defining a subgroup as patients who had strains with no more than 5 nucleotide differences within the subgroup. The first subgroup, which we arbitrarily referred to as Clade 20B-nonT (Tokyo), had the basic haplotype of Clade 20B, which is defined by 7 mutations separating the lineage from the founding Wuhan haplotype, but had additional less than 5 single nucleotide substitutions. The second subgroup, which we arbitrarily refer to as subclade 20B-T (Tokyo), had the basic haplotype of Clade 20B but had additional 6 single nucleotide substitutions: c.4346 U>C, c.9286 C>U, c.10376 C>U, c.14708 C>U, c.28725 C>U, and c.29692 C>U (Figure 2a, yellow). Among these 6 mutations, 4 were non-synonymous mutations: c.4346 U>C (Ser543Pro), c.10376 C>U (Pro108Ser), c.14708 C>U (Ala423Val), and c.28725 C>U (Pro151Leu); the remaining 2 mutations did not affect the amino acid translation of the viral proteins. An analysis of the

cumulative total number and frequency curve showed that the relative fraction of Clade 20B-T increased during the time frame of this study (since May 2020) (Figure 2b-c).

Mapping of the suspected geographic locations where infection in each patient likely occurred indicated that patients with the Clade 20B-T strain and those with the Clade 20B-nonT strain become infected in various regions in the Tokyo metropolitan area and its neighboring prefectures (Figure 2d). This observation, together with a lack of patients with strains belonging to other clades (except for the 5 patients with Clade 19A who belonged to the same cluster) did not prove but suggested that Clade 20B and its variation Clade 20B-T were the predominant strains during the observation period.

Milder clinilcal course in Clade 20B-T patients

A comparison of the clinical characteristics between patients with the Clade 20B-T viral strain (N = 48) and those with other strains (N = 37) is shown in Table 1. Age, sex, symptoms at admission, and outcome did not differ significantly between the two main groups, but the use of oxygen supplementation was significantly lower among the patients with Clade 20B-T viral strain, compared with those with other strains (Fisher exact test: 12.5% vs. 32.4%, *P* value = 0.033). An exact logistic regression analysis showed that patients with the Clade

20B-T viral strain had a lower odds ratio for the development of hypoxia requiring supplemental oxygen, compared with those with other strains (adjusted odds ratio, 0.24 [95% CI, 0.07-0.88], *P* value = 0.032; Table 2) after adjustments for age group (< 65 years or ≧65 years) and sex (female or male).

Molecular evolutionary characterization of four non-synonymous mutations unique to Clade 20B-T.

We further aimed to decipher which of the 4 non-synonymous mutations that characterize the Clade 20B-T haplotype may contribute to a milder clinical course through a molecular evolutionary analysis. The conservation of the amino acid residues around the non-synonymous mutations observed in Clade 20B-T was evaluated using the software MEGA (https://www.megasoftware.net/) and the amino acid sequences of known β-coronaviruses. Amino acid residues at and around the Pro108Ser mutation in the $3CL^{pro}$ protein (nonstructural polyprotein 5; NSP5), and those at and around the Pro151Leu mutation in the nucleocapsid protein were highly conserved, whereas amino acid residues at and around the Ser543Pro mutation in the papain-like protease (PL^{pro}, NSP3) protein and those at and around the Ala423Val mutation in RNA-dependent RNA polymerase

(RdRp, NSP12) were only weakly conserved (Figure 3a). Furthermore, the serine in the papain-like protease at residue 543, where the Ser543Pro mutation was observed in Clade 20B-T, is substituted with proline in some of the β-coronaviruses. Similarly, the Ala at residue 423 in RdRp, where the Ala423Val mutation was observed in Clade 20B-T, is substituted with valine in some β-coronaviruses. Such evolutionary observations suggest that two substitutions, the Ser543Pro mutation in PL^{pro} (NSP3) and the Ala423Val mutation in RdRp (NSP12), are likely to be functionally neutral. In support of this notions, the PROVEAN prediction score revealed that the Ser543Pro mutation in PL^{pro} and the Ala423Val mutation in RdRp were not deleterious substitutions.

Protein structures and stabilities of Pro108Ser mutant in $3CL^{pro}$ and Pro151Leu mutant in nucleocapsid protein compared with Wuhan-strain type.

The structures of $3CL^{pro}$ and nucleocapsid protein were visualized using PyMol v2.4. (https://pymol.org/2/) and the impact of substitution was estimated by using the DynaMut server (http://biosig.unimelb.edu.au/dynamut/).

The Pro108Ser substitution in the $3CL^{pro}$ occurred at a distance from the enzymatically active sites at His41 and Cys145, which mediate amide hydrolysis (Figure

4a).[14-16] The modeling showed that the Pro108Ser substitution in $3CL^{pro}$ could not induce significant structural changes at or around residue 108 or intramolecular interactions between the residue at 108 and surrounding residues (Figure 4b). Meanwhile, importantly, the $\Delta\Delta G$ values and the $\Delta\Delta S$Vib for the protein with Ser108, compared with the protein with Pro108, were 0.362 and -0.259 kcal/mol/K, respectively (Figure 4c).

The Pro151Leu substitution in nucleocapsid protein was in close proximity to the surface of the tetramer of the critical RNA-binding N-terminal domain (NTD) (Figure 4d).[17] The Pro151Leu substitution in nucleocapsid protein could not induce significant changes in structure at or around the residue at 151 or the intramolecular interactions of the residue at 151 with the surrounding residues (Figure 4e). However, the $\Delta\Delta G$ values and the $\Delta\Delta S$Vib for the protein with Leu151, compared with the protein with P151, were 0.771 kcal/mol and -0.140 kcal/mol/K, respectively (Figure 4f). These modeling and calculation suggest the stabilization of the protein structure with decreased overall molecular flexibility by Pro108Ser substitution in $3CL^{pro}$ and Pro151Leu substitution in nucleocapsid protein, respectively.

Discussion

The number count of non-synonymous mutations was inversely correlated with the severity of COVID-19 disease in a cohort of 90 patients whose viral genome sequences had been completely sequenced. Patients with a viral haplotype containing both the $3CL^{pro}$ Pro108Ser mutation and the nucleocapsid protein Pro151Leu mutation tended to have a milder disease course than those with a viral haplotype lacking these two mutations.

Our observation that a mutation of $3CL^{pro}$, Pro108Ser, was associated with disease severity strongly supports the notion that inhibitors for $3CL^{pro}$, the main protease that cleaves viral polyproteins into functional proteins,[18] could be a promising antiviral agent that may be effective against SARS-CoV,[19,20] as well as SARS-Cov-2.[21] Further enzymatic analysis of the Pro108Ser mutation in $3CL^{pro}$ is needed.

The other observation that the Pro151Leu mutation in nucleocapsid protein, which enters the host cells along with the viral RNA, may be associated with disease severity is also intriguing in that nucleocapsid protein is responsible for promoting viral replication and processing the assembly and release of viral particles.[22]

The computer-based protein structure predicted that the mutant $3CL^{pro}$ and the mutant nucleocapsid proteins are less flexible, with no significant changes in protein structure or stabilities. However, their interactions and dominance are not known.

In general, RNA viruses continue to survive and proliferate by constantly changing their forms and adapting to various environments, but quasispecies with a high replicability seem to be unfavorable in terms of long-term survival because of a high sensitivity to environmental influences (i.e., survival of the fittest).[23] On the other hand, quasispecies with a low replicability may have reduced infectivity and pathogenicity but actually have long-term survival advantages because they are less likely to be subjected to natural selection (i.e., survival of the flattest).[24,25] Coronaviruses, including SARS-CoV or SARS-CoV-2, are well known to have a low fidelity in replicating their genome.[26,27] Of the proteins involved in the replication of the viral genome, proteases, including $3CL^{pro}$, are critical for viral replication.[28] Therefore, the Pro108Ser mutation in $3CL^{pro}$ may lead to a situation in which the virus is less susceptible to natural selection and long-term survival is more favorable, reducing viral infectivity and virulence.

The present study had several limitations. First, the decrease in the fraction of critically ill patients might be accounted for by the seasonality of vulnerability to viruses, in general.[29] One report has suggested that SARS-CoV-2 is more stable and has a higher persistence at the same temperature than SARS-CoV.[30] However, the lack of such seasonal trends in the SARS-CoV-2 pandemic observed in other countries suggests that a

seasonal bias is unlikely. Second, drawing a decisive conclusion based on a relatively small number of subjects at a single center could be premature in terms of the relative virulence of the two subclades in Tokyo. However, the near absence of haplotype groups other than Clade 20B-T or 20B-nonT in our cohort of 87 patients, whose suspected locations of infection were scattered across the city (Figure 2d), strongly suggest that these two groups represent the predominant or almost exclusive strains in Tokyo. Currently, the number of whole viral genome sequences determined in Japan and deposited in international databases is not sufficient to conclude that viral strains with a Clade 20B-nonT haplotype carrying the $3CL^{Pro}$ Pro108Ser mutation and the nucleocapsid protein Pro151Leu mutation are predominant in Japan. Nevertheless, recent data deposited from a tertiary medical center in Nagoya, Japan's third largest city located 500 km away from Tokyo, support our notion. Among the 27 viral strains collected in Nagoya between March and August of 2020, 19 strains belonged to the Clade 20B-T haplotype and 8 belonged to the Clade 20B-nonT haplotype (Supplementary Figure 3). This distribution essentially recapitulates our observations in Tokyo.

In conclusion, viral genome sequencing in 90 patients living in the Tokyo Metropolitan area showed that the accumulation of point mutations, including Pro108Ser

in $3CL^{pro}$ and Pro151Leu in nucleocapsid protein, was inversely correlated with COVID-19 severity. Further *in vitro* research is awaited.

Author contributions

KA contributed to writing of the report and data analysis. YK, SU, TM and MN contributed to review and editing of the report and data analysis. YI, HI, TT and HS contributed to sequencing and analysis. YU, SU and NH contributed to public health intelligence and case identification. MI and KF contributed to clinical data and clinical care. HS, YK and MA contributed to writing and editing of the report. MM contributed to diagnostics and laboratory management. MS and KK had the idea for the study and contributed to diagnostics, formal analysis, and writing and editing of the report.

Declaration of interests

Authors have no conflicts of interests.

Funding sources

This work was supported by Keio Gijuku Academic Development Funds and by AMED under Grant Number JP20he0622043.

Data sharing.

We downloaded the full nucleotide sequences of the SARS-CoV-2 genomes from the GISAID database (https://www.gisaid.org/). A table of the contributors is available in the Acknowledgements Table. We have uploaded the full nucleotide sequences of our cohort to the GISAID database.

Acknowledgements

We thank all the patients and healthcare workers who have fought against COVID-19.

This work was supported by the Keio Donner Project and is devoted to the late Professor Shibasaburo Kitasato, the founder of Keio University School of Medicine. We also thank the members of Center for Medical Genetics in Keio University School of Medicine, and SUNTORY Co., Ltd.

References

1. Su YC, Anderson DE, Young BE, et al. Discovery and Genomic Characterization of a 382-Nucleotide Deletion in ORF7b and ORF8 during the Early Evolution of SARS-CoV-2. *mBio* 2020; 11(4): e01610-20. https://doi.org/10.1128/mBio.01610-20.

2. Young BE, Fong SW, Chan YH, et al. Effects of a major deletion in the SARS-CoV-2 genome on the severity of infection and the inflammatory response: an observational cohort study. *Lancet* 2020; 396: 603–11. https://doi.org/10.1016/S0140-6736(20)31757-8.

3. Takenouchi T, Iwasaki Y, Harada S, et al. Clinical Utility of SARS-CoV-2 Whole Genome Sequencing in Deciphering Source of Infection. *J Hosp Infect* 2020. https://doi.org/10.1016/j.jhin.2020.10.014.

4. Koyama T, Platt D, Parida L. Variant analysis of SARS-CoV-2 genomes. *Bull World Health Organ* 2020;98:495–504. doi: http://dx.doi.org/10.2471/BLT.20.253591

5. Chen J, Wang R, Wang M, Wei GW. Mutations Strengthened SARS-CoV-2 Infectivity. *J Mol Biol* 2020; 432: 5212-26. https://doi.org/10.1016/j.jmb.2020.07.009

6. Kumar M, Taki K, Gahlot R, Sharma A, Dhangar K. A chronicle of SARS-CoV-2: Part-I - Epidemiology, diagnosis, prognosis, transmission and treatment. *Sci Total Environ* 2020; 734: 139278. https://doi.org/10.1016/j.scitotenv.2020.139278

7. Meredith LW, Hamilton WL, Warne B, et al. Rapid implementation of SARS-CoV-2 sequencing to investigate cases of health-care associated COVID-19: a prospective genomic surveillance study. *Lancet Infect Dis* 2020; https://doi.org/10.1016/

8. World Health Organization. Clinical Management of COVID-19 Interim guidance 27 May 2020. WHO/2019-nCoV/clinical/2020.5

9. Rodrigues CH, Pires DE, Ascher DB. DynaMut: predicting the impact of mutations on protein conformation, flexibility and stability. *Nucleic Acids Res* 2018; 46:W350-5.

10. Frappier V, Najmanovich RJ. A Coarse-Grained Elastic Network Atom Contact Model and Its Use in the Simulation of Protein Dynamics and the Prediction of the Effect of Mutations. *PLoS Comput Biol* 2014; 10(4): e1003569. doi:10.1371/journal.pcbi.1003569.

11. Eriksson AE, Baase WA, Zhang XJ, et al. Response of a protein structure to cavity-creating mutations and its relation to the hydrophobic effect. *Science* 1992; 255(5041): 178-83.

12. Rambaut A, Holmes EC, O'Toole A, et al. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat Microbiol* 2020; 5: 1403-7. https://doi.org/10.1038/s41564-020-0770-5.

13. Korber B, Fischer WM, Gnanakaran S, et al. Tracking changes in SARS-CoV-2 spike: evidence that Asp614Gly mutation increases infectivity of the COVID-19 virus. *Cell* 2020; 182(4): 812-27.e19. doi: 10.1016/j.cell.2020.06.043.

14. Jin Z,, Du X, Xu Y, et al. Structure of Mpro from SARS-CoV-2 and discovery of its inhibitors. *Nature* 2020; 582: 289-93.

15. Chen S, Chen L, Luo H, et al. Enzymatic activity characterization of SARS coronavirus 3C-like protease by fluorescence resonance energy transfer technique1. *Acta Pharmacologica Sinica* 2005; 26(1): 99-106.

16. Sun H, Luo H, Yu C, et al. Molecular cloning, expression, purification, and mass spectrometric characterization of 3C-like protease of SARS coronavirus. *Protein Expr Purif* 2003; 32: 302–8.

17. Kang S, Yang M, Hong Z, et al. Crystal structure of SARS-CoV-2 nucleocapsid protein RNA binding domain reveals potential unique drug targeting sites. *Acta Pharmaceutica Sinica B* 2020; 10(7): 1228-38.

18. Yoshimoto FK. The Proteins of Severe Acute Respiratory Syndrome Coronavirus‐2 (SARS CoV‐2 or n‐COV19), the Cause of COVID‐19. *Protein J* 2020; 39: 198–216.

19. Pillaiyar T, Manickam M, Namasivayam V, Hayashi Y, Jung SH. An Overview of Severe Acute Respiratory Syndrome−Coronavirus (SARS-CoV) 3CL Protease Inhibitors: Peptidomimetics and Small Molecule Chemotherapy. *J Med Chem* 2016, 59, 6595−628.

20. Li Q, Kang CB. Progress in Developing Inhibitors of SARS-CoV-2 3C-Like Protease. *Microorganisms* 2020; 8(1250); doi:10.3390/microorganisms8081250.

21. Hung HC, Ke YY, Huang SY, et al. Discovery of M Protease Inhibitors Encoded by SARS-CoV-2. *Antimicrob Agents Chemother* 2020; 64: e00872-20. https://doi.org/10.1128/AAC.00872-20.

22. Rahman MS, Islam MR, Alam AR, et al. Evolutionary dynamics of SARS-CoV-2 nucleocapsid (N) protein and its consequences. *J Med Virol* 2020. https://doi.org/10.1002/jmv.26626

23. Domingo E, Perales C. Viral quasispecies. *PLOS Genetics* 2019; https://doi.org/10.1371/journal.pgen.1008271.

24. Bingham RJ, Dykeman EC, Twarock R. RNA Virus Evolution via a Quasispecies-Based Model Reveals a Drug Target with a High Barrier to Resistance. *Viruses* 2017, 9, 347; doi:10.3390/v9110347.

25. Tejero H, Marin A, Montero F. The relationship between the error catastrophe, survival of the flattest, and natural selection. *BMC Evol Biol* 2011; 11: 2. http://www.biomedcentral.com/1471-2148/11/2.

26. Novella IS, Presloid JB, Taylor RT. RNA replication errors and the evolution of virus pathogenicity and virulence. *Curr Opin Virol* 2014; 9: 143-7.

27. Smith EC, Denison MR. Implications of altered replication fidelity on the evolution and pathogenesis of coronaviruses. *Curr Opin Virol* 2012; 2: 519-24.

28. Babe LM, Craik CS. Viral Proteases: Evolution of Diverse Structural Motifs to Optimize Function. *Cell* 1997; 91: 427–30.

29. Li Y, Reeves RM, Wang X, et al. Global patterns in monthly activity of influenza virus, respiratory syncytial virus, parainfluenza virus, and metapneumovirus: a systematic analysis. *Lancet Glob Health* 2019; 7: e1031–45.

30. He J, Tao H, Yan Y, Huang SY, Xiao Y. Molecular Mechanism of Evolution and Human Infection with SARS-CoV-2. *Viruses* 2020; 12: 428; doi:10.3390/v12040428.

Figure Legends

Figure 1: Number of non-synonymous mutations of SARS-CoV-2 is inversely correlated with COVID-19 disease severity. a. The number of non-synonymous mutations (vertical axis) was significantly lower in COVID-19 patients who required supplemental oxygen than in those who did not (Student *t*-test: mean, 7.9 [SD, 2.2] vs. 5.9 [SD, 2.2], **P* value <0.001). b. The number of non-synonymous mutations (vertical axis) tended to decrease as the COVID-19 severity increased (Jonckheere-Terpstra trend analysis: JT = 404, * *P* value < 0·001). SARS-CoV-2, severe acute respiratory syndrome coronavirus 2; COVID-19, coronavirus disease 2019.

Figure 2: Phylogenic tree analysis, temporal trends, and spatial distribution around Keio University Hospital (purple dot in Fig. 2d) showed consistent increase of a strain with a unique haplotype. a. Phylogenic tree analysis of whole viral genome sequences recovered at Keio University Hospital consisted of clades 19A and 20B defined by GISAID (https://www.gisaid.org/). We further defined the predominant Clade "20B-T" which had the six additional mutations compared with the remaining strains "Clade 20B-nonT" of Clade 20B. b. Temporal trends of the number of patients of Clade 20B-T and Clade 20B-nonT. Clade 20B-T became predominant over Clade 20B-nonT. c. The cumulative frequency of Clade 20B-T increased steadily to exceed 50%. d. The suspected location of infection of individuals from Clade 20B-T and those from Clade 20B-nonT scattered over Tokyo Metropolitan area and its neighboring prefecture. COVID-19, coronavirus disease 2019. NSP, Non-structural polyprotein; PL^{pro}, papain-like proteinase; $3CL^{pro}$, 3 chymotrypsin-like protease; RdRp, RNA dependent RNA polymerase; ORF, open reading frame.

Figure 3: Multiple amino acid sequence alignments of various betacorona viruses and locations of mutated amino acid residues in Clade 20B-T. a. The structure of the genomic region that encodes nonstructural polyproteins of SARS-CoV-2. Multiple sequence alignments homologous proteins of 7 β-coronaviruses at and around 3 non-synonymous mutations: Ser543Pro in the NSP3 protein (PL^{pro}), Pro108Ser in the NSP5 ($3CL^{pro}$) protein, and Ala423Val in the NSP12 (RdRp) protein. b. The structure of the genomic region that encodes nucleocapsid protein of SARS-CoV-2. Multiple sequence alignments homologous proteins of 7 β-coronaviruses at and around the non-synonymous substitution Pro151Leu in the nucleocapsid protein. SARS-CoV-2, severe acute respiratory syndrome coronavirus 2. ORF, open reading frame; NSP, nonstructural protein; PL^{pro}, papain-like protease; $3CL^{pro}$, 3 chymotrypsin-like protease; RdRp, RNA-dependent RNA polymerase; SARS, severe acute respiratory syndrome; MERS, middle east respiratory syndrome; NTD, N-terminal domain; CTD, C-terminal domain; COVID-19, coronavirus disease 2019; MEGA, Molecular Evolutionary Genetics Analysis.

Figure 4: The modeled structure and calculated stabilities of Pro108Ser in 3CL^{pro} and Pro151Leu in nucleocapsid protein. a. The structure of the dimer of 3CLpro in SARS-CoV-2 (PDB ID ‘‘6LU7’’) and modeled structure of the single mutated 3CL^{pro}. Pro108Ser mutation is distant from the active sites at His41 and Cys145 of 3CL^{pro}. b. Predicted interactomic interactions and relationship with the surrounding residues between Wuhan-strain type and Pro108Ser mutant in 3CL^{pro} were described. c. The impact of Pro108Ser mutation in 3CL^{pro} on free energy change were estimated based on 3D structure in the DynaMut. d. The structure of the tetramer of nucleocapsid protein in SARS-CoV-2 was indicated (PDB ID ‘‘6VYO’’). Pro151Leu mutation is in close proximity to the surface of the tetramer of the critical RNA-binding N-terminal domain. e. Predicted interactomic interactions and relationship with the surrounding residues between Wuhan-strain type and Pro151Leu mutant in nucleocapsid protein. f. The impact of Pro151Leu mutation on free energy change in nucleocapsid protein were estimated using DynaMut. 3CL^{pro}, 3 chymotrypsin-like protease; SARS-CoV-2, severe acute respiratory syndrome coronavirus 2; PDB, protein data bank; 3D, three-dimensional; $\Delta\Delta G$, difference of the free energy change; $\Delta\Delta S$Vib, difference of the vibrational entropy change between Wuhan-strain type and mutants.

Table 1: Comparison of clinical features between Clade 20B-T and Clade 20B-nonT

N = 85	Clade 20B-T (N = 48)	Clade 20B-nonT (N = 37)	*P* value
Mean Age (years old)	41·0 [SD 15·9]	43·7 [SD 19·7]	0.504
Sex (male / female)	34 / 14	21 / 16	0.252
Symptoms at admission			
Cough	22 (45·8 %)	14 (37·8 %)	0.512
Dysosmia	6 (12·5 %)	5 (13·5 %)	1.000
Dysgeusia	4 (8·3 %)	5 (13·5 %)	0.494
Fever (≧ 37·5℃)	22 (45·8 %)	16 (43·2 %)	0.830
Sepsis	0	0	-
ARDS	2 (4·2 %)	4 (10·8 %)	0.396
Treatment			
Oxygen supplementation	6 (12·5 %)	12 (32·4 %)	0.033
Methylprednisolone treatment	2 (4·2 %)	5 (13·5 %)	0.232
Ventilator usage	1 (2·1 %)	3 (8·1 %)	0.313
ICU admission	1 (2·1 %)	3 (8·1 %)	0.313
Death	1 (2·1 %)	1 (2·7 %)	1.000

PCR, polymerase chain reaction; ARDS, acute respiratory distress syndrome; ICU, intensive care unit

Table 2: Logistic regression analysis of candidate predictors for requiring supplemental oxygen

	Univariate model		Multivariate model	
	OR (95% CI)	*P* value	Adjusted OR (95% CI)	*P* value
Age, years				
< 65	1 (ref)	-	1 (ref)	-
≧ 65	12.02 (2.66-65.58)	<0.001	14.67 (3.32-64.79)	<0.001
Sex				
Female	1 (ref)	-	-	-
Male	2.20 (0.60-10.19)	0.269	-	-
Infection				
Clade 20B-nonT	1 (ref)	-	1 (ref)	-
Clade 20B-T	0.30 (0.08-1.00)	0.033	0.24 (0.07-0.88)	0.032

OR, odds ratio; CI, cofidence interval

Figure 1: Number of non-synonymous mutations of SARS-CoV-2 is inversely correlated with COVID-19 disease severity.

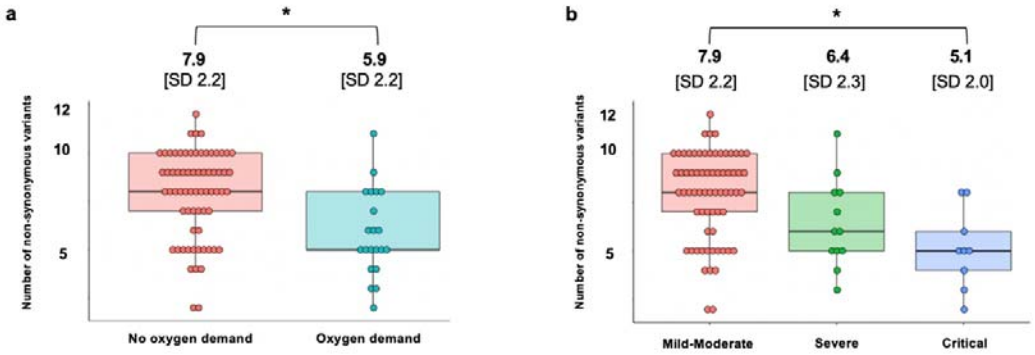


Figure 2: Phylogenic tree analysis, temporal trends, and spatial distribution around Keio University Hospital (purple dot in Fig. 2d) showed consistent increase of a strain with a unique haplotype.

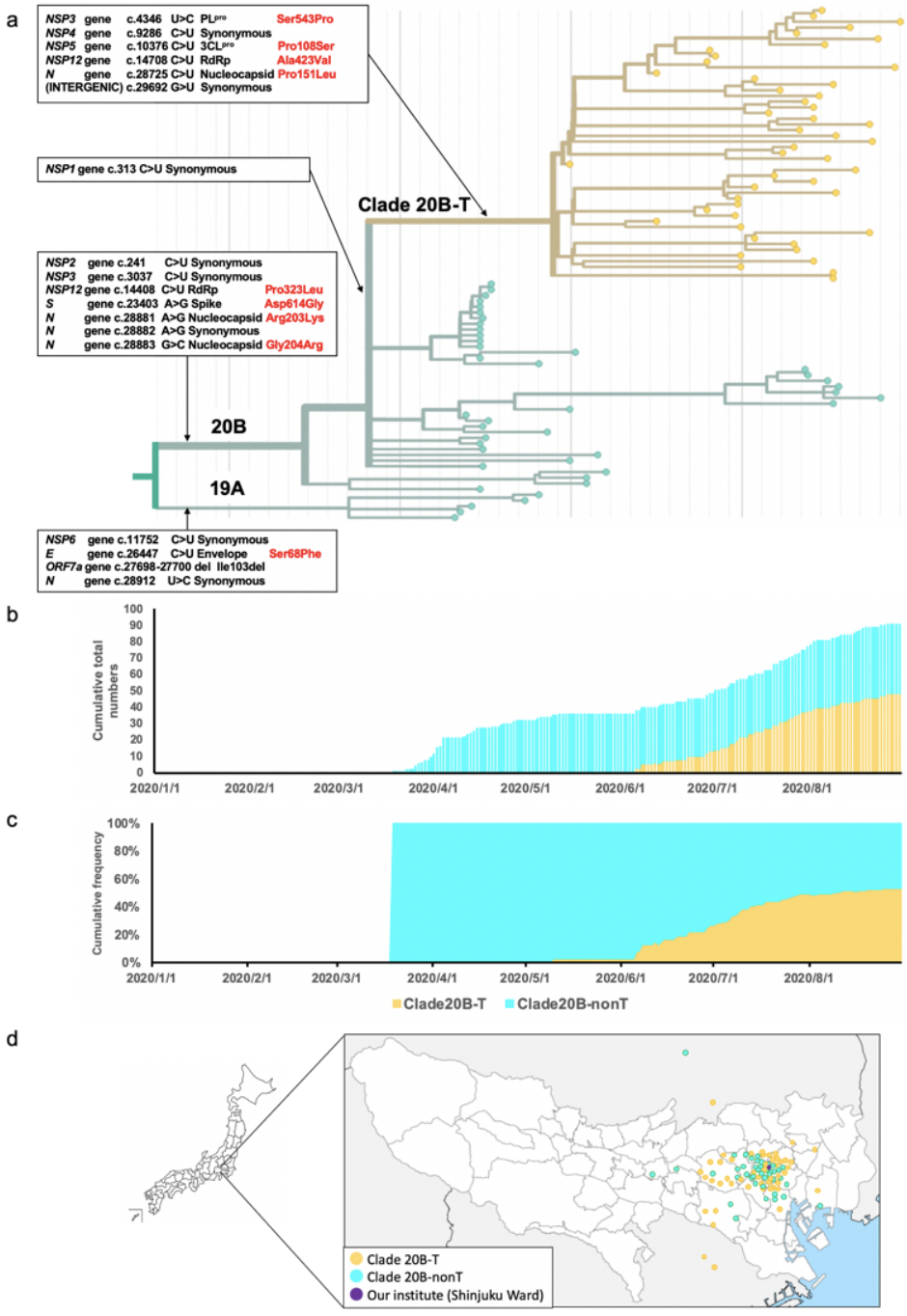


Figure 3: Multiple amino acid sequence alignments of various betacorona viruses and locations of mutated amino acid residues in Clade 20B-T.

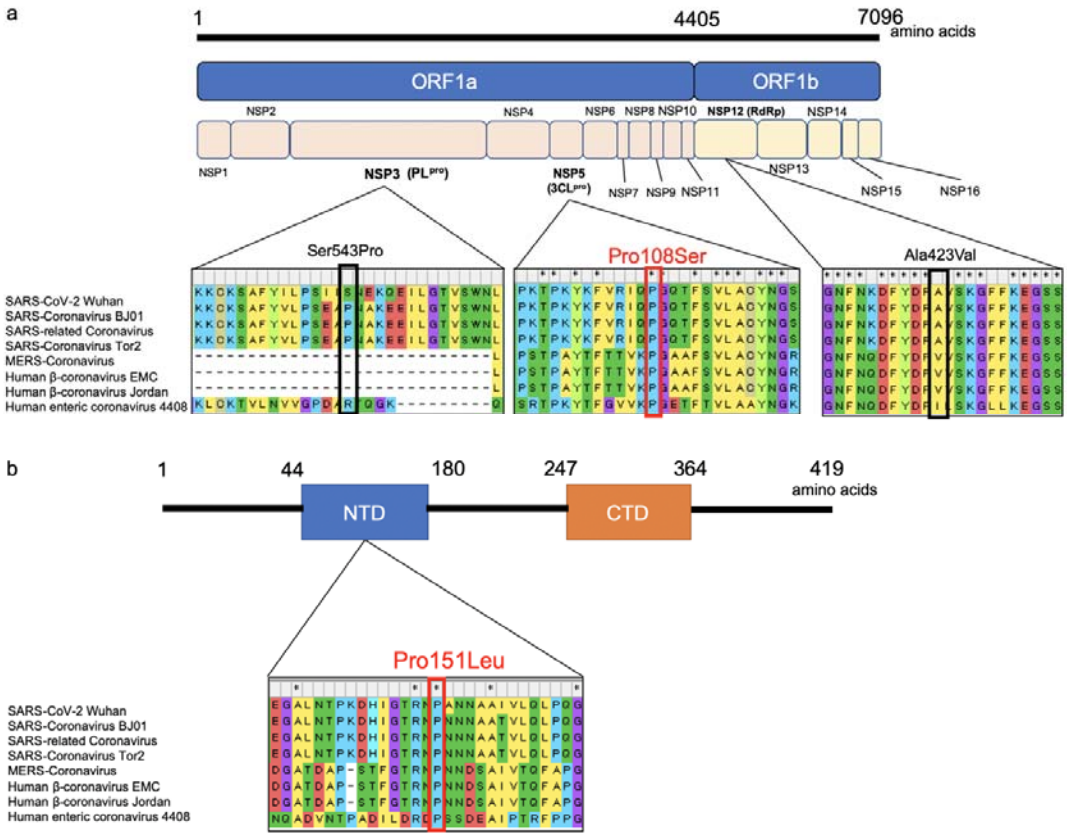
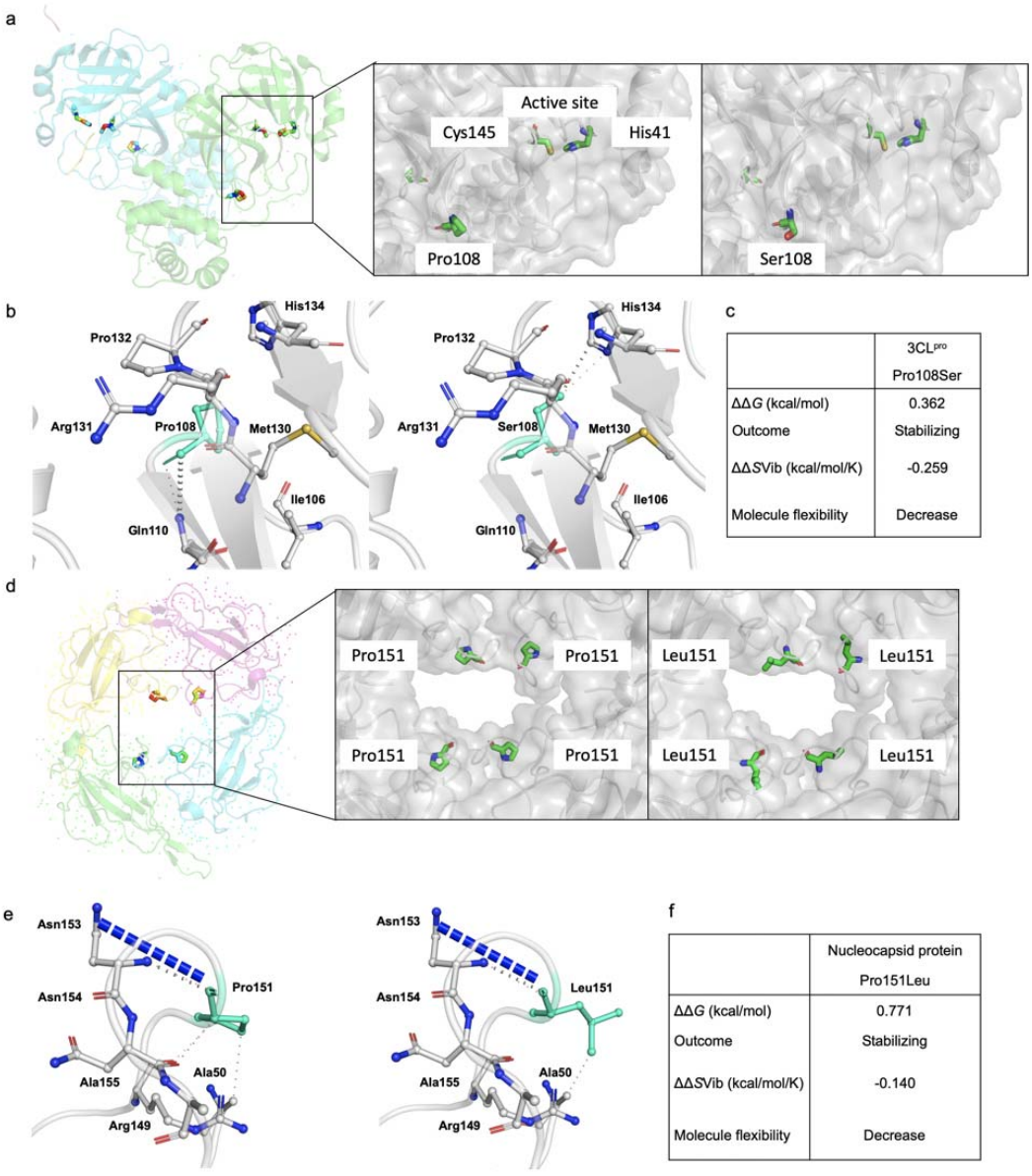


Figure 4: The modeled structure and calculated stabilities of Pro108Ser in 3CL^{pro} and Pro151Leu in nucleocapsid protein.



	3CL^{pro} Pro108Ser
ΔΔ*G* (kcal/mol)	0.362
Outcome	Stabilizing
ΔΔ*S*Vib (kcal/mol/K)	-0.259
Molecule flexibility	Decrease

	Nucleocapsid protein Pro151Leu
ΔΔ*G* (kcal/mol)	0.771
Outcome	Stabilizing
ΔΔ*S*Vib (kcal/mol/K)	-0.140
Molecule flexibility	Decrease