Genome-wide Mutation Landscape of SARS-CoV-2

Xiaofei Yang

1School of Computer Science and Technology, Faculty of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an, 710049 China https://orcid.org/0000-0002-5118-7755

Li Lv

3School of Electrical & Electronic Engineering, Baoji University of Arts and Sciences, Baoji, 721013, China

Kai Ye (kaiye@xjtu.edu.cn)

School of Automation Science and Engineering, Faculty of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an, 710049 China., 5Genome Institute, the First Affiliated Hospital of Xi'an Jiaotong University, Xi'an, 710061 China., 6The School of Life Science and Technology, Xi'an Jiaotong University, Xi'an, 710049.China. *To whom correspondence should be addressed, kaiye@xjtu.edu.cn.

Research article

Keywords: SARS-CoV-2, mutation, spike glycoprotein

Posted Date: August 12th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-47392/v1

Genome-wide mutation landscape of SARS-CoV-2

Xiaofei Yang[1,2], Li Lv[3], Kai Ye[2,4,5,6,*]

[1]School of Computer Science and Technology, Faculty of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an, 710049 China., [2]MOE Key Lab for Intelligent Networks & Networks Security, Faculty of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an, 710049 China., [3]School of Electrical & Electronic Engineering, Baoji University of Arts and Sciences, Baoji, 721013, China., [4]School of Automation Science and Engineering, Faculty of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an, 710049 China., [5]Genome Institute, the First Affiliated Hospital of Xi'an Jiaotong University, Xi'an, 710061 China., [6]The School of Life Science and Technology, Xi'an Jiaotong University, Xi'an, 710049.China.

*To whom correspondence should be addressed, kaiye@xjtu.edu.cn.

Abstract

Background: SARS-CoV-2 has become a pandemic and researchers have built phylogenetic trees to trace the spread of the virus. However, the accumulation rate of variations and mutational hotspots remain largely unclear.

Results: We collected more than 3,100 SARS-CoV-2 genome sequences from GISAID and profiled the landscape of whole genome variations. We detected 2,096 single nucleic variants (SNVs) and seven short deletions. 1,224 of them (58.4%) are missenses variation, altering the corresponding residues. We found the accumulation rate of SNVs in the current spreading situation is 6.36e-2/day. We found 15 missenses SNVs are extremely high frequent (existing in more than 100 genome sequences, $p < 1e\text{-}5$), effecting *ORF1ab*, *S*, *ORF3a*, *M*, *ORF8*, and *N*. Moreover, one frequent substitution at locus 23,403 changes the 614th amino acid of spike glycoprotein from D to G, potentially effecting the functions of this key protein.

Conclusion: Our study provided the genome-wide mutation landscape of SARS-CoV-2. We found the continent specific mutational patterns and 15 missenses

high frequent SNVs effecting 6 genes of the virus, may promoting the adaption of the virus during evolving.

Keywords: SARS-CoV-2; mutation; spike glycoprotein

Background

The pandemic of coronavirus disease 2019 (COVID-19), causing by Severe acute respiratory syndrome–coronavirus 2 (SARS-CoV-2), has caused more than 10 million confirmed cases and 512,000 deaths globally (as of July. 4, 2020, https://covid19.who.int/). Comparative genomic analyses indicated that the virus potentially originated from either bat [1, 2] or pangolin [3]. SARS-CoV-2 is an RNA virus, and its genome includes 29,903 nucleotides [2], 11 genes, a 5'UTR and a 3'UTR [4]. Currently, several research directions like the clinical characteristics of COVID-19 [5-7], drug target discovery [8, 9], 3D structure of spike glycoprotein [10, 11], the way of SARS-CoV-2 enter into human cells [12] are the focus of the field. In addition, researchers have identified variants occurred in the virus population, and classified virus into different types [13, 14]. the accumulation rate of variations and mutational hotspots remain largely unclear. In this work, we detected and analyzed variation from 3,160 SARS-CoV-2 strains, and constructed a graph genome to incorporate mutations in current reference genome. We found 15 high frequent (existing in more than 100 genome sequences) missenses variants, affecting *ORF1ab*, *S*, *ORF3a*, *M*, *ORF8*, and *N* proteins, which suggests the rapid evolution and sound an alarm of future outbreaks.

Results

Mutations landscape and graph genome of SARS-CoV-2

In total, we detected 2,096 single nucleic variants (SNVs) and seven short deletions (Supplementary Data). Most (1,350) of the SNVs exist in only one strain, while 22 of them exist in more than 100 reported genome sequences (Supplementary Fig. S1). For the substitution types, we found the most frequent

(807) one is C to T transitions and the second most frequent (327) is G > T transversion (Supplementary Fig. S2). We annotated 2,096 SNVs and found 1,224 are missense variant , which is significant lower than random background constructing by shuffling all SNVs in the genome 500 times (p < 7.6e-6, Supplementary Fig. S3), suggesting SARS-CoV-2 is under selection pressure during COVID-19 pandemic. To investigate the accumulation rate of SNVs, we calculated average number of SNVs for genome sequences collecting at the same date, and used a linear regression to fit the mean SNV numbers with date. We found a clear positive correlation and estimated the coefficient of date is 6.36e-2 (p = 3.12e-24, F-test) (**Fig. 1**), indicating a new SNV occurs in SARS-CoV-2 population per 16 days in current outbreak. All of these above evidences indicate the current reference genome of SARS-CoV-2 cannot reflect the genetic information of whole virus population, e.g. representing only one version of each locus and ignoring major variations in the population, which may complicate downstream analysis, such as metagenome analysis. It is necessary to construct a new SARS-CoV-2 reference genome incorporating known variants. Graph genome has been shown a good solution to address such problems [15-17]. Here, we constructed a graph genome of SARS-CoV-2 by variation graph toolkit [15] based on current reference genome and the frequent mutations (https://github.com/xjtu-omics/SARS-CoV-2_graph_genome) to facilitate future analysis in this pandemic.

SNV hotspots in SARS-CoV-2

We compared the SNV frequency with the random profiles, and detected 244 significantly high frequent SNVs (p < 1e-5, Supplementary Table S1). We found the high frequent SNVs are significantly enriched in last two 1kb bins (p < 1e-5 for bin 28,000-29,000 and 29000-29,303, Supplementary Table S2), which contain gene ORF8, N, ORF10, and 3'UTR of the virus. Of these high frequent SNVs, we further detected 22 SNVs occurred in more than 100 genome sequences (**Fig. 2**, Supplementary Table S3). All of these 22 hotspots related with six genes, including 13 hotspots locate at region of *ORF1ab*, three hotspots

locate at region of *N*, two hotspots locate at region of *ORF3a*, one hotspot locate at region of *M* and one locate at region of *ORF8* (**Fig. 2A**). We annotated 22 hotspots SNVs and found five synonymous, fifteen missense, one at 5'UTR and regulatory region of SARS-CoV-2 genome (locus 241, C > T, in 1,649 genome sequences) and one stop lost SNV (locus 20,268, A > G, in 117 genome sequences), leading to the elongation of *ORF1ab* protein (**Fig. 2A**).

To further explore the relations between SNV hotspots and geographic locations, we did an enrichment analysis and found different continents had specific enrichment patterns on SNV hotspots (fold change (FC) > 1, $p < 0.01$, **Fig. 2B,** Supplementary Fig. S4, Supplementary Table S4). Specifically, nonoverlapped SNV hotspots were significantly enriched in genome sequences from Africa, Asia, North America, and South America. Furthermore, we found five missense SNV hotspots (locus 1,059, C > T; locus 17,858, A > G; 18,060, C > T; locus 25,563, G > T, and locus 28,144, T > C) are specifically enriched in North America (FC > 1, $p < 0.01$, **Fig. 2B**, Supplementary Fig. S4, Supplementary Table S4), whose roles remain unexplored in current high percentage of global infections in America.

SNV at locus 23,403 effect the protein structure of SARS-CoV-2 spike glycoprotein

Among the 22 SNV hotspots, we highlighted one missense variant in *S* gene (locus 23,403, A > G). This substitution appears in 1,656 sequences, and significantly enriches in sequences from Europe (FC = 1.44, p = 1.80e-142) and Africa (FC = 1.67, p = 3.56e-8) (**Fig. 2B**, Supplementary Fig. S4, Supplementary Table S4). We found that this mutation firstly occurred at Jan. 28th, 2020 in a German isolate (Accession: EPI_ISL_406862) (Supplementary Table S5). After Feb. 20th, 2020, this variation gradually dominates (**Fig. 3A**).

Furthermore, this missense variant alters the 614th amino acid from Asp (D) to Gly (G) of spike glycoprotein (coded by *S* gene), the so-called S protein initiating virus entering human cells [10]. We analyzed the mutation sensitivity of each amnio acid in spike glycoprotein by Phyre2 [18], and found the mutation

sensitivity score of the 614th amino acid is 4 (**Fig. 3B,** Supplementary Table S6), indicating potential functional alternation of the spike glycoprotein. Whether this particular mutant strain has evolved to alter infection efficiency remains unclear.

Discussion

We analyzed more than 3,100 SARS-CoV-2 genome sequences, and detected SNV and indel mutations across whole genome. The graph genome of SARS-CoV-2 incorporates high frequent mutations, facilitating downstream analysis. We found 22 SNV hotspots existing in more than 100 genome sequences, and 15 of them are missense variants. The SNV at locus 23,403, observed in 1,656 genome sequences (52.4%), alters the 614th amino acid of spike glycoprotein from D to G, might altering its function. However, this mutation does not appear at the interface between virus S protein and human ACE2 but in the flexible hinge region of S protein, requiring further functional investigation. Besides spike glycoprotein, proteins coded by *ORF1ab*, *N*, *M*, *ORF3a*, *ORF8* are also affected by SNV hotspots (**Fig. 2A**, Supplementary Table S3) and their implications on infection and mortality rates require further investigation. Although 3' to 5' exoribonuclease of ExoN in coronavirus [19] in principle renders lower mutation rate than other RNA virus, the observed accumulation of continent specific mutational hotspots sounds an alarm that immunity gained from either vaccines or previous infection might not be sufficient to prevent future outbreaks of COVID-19.

Conclusions

In this study, we analyzed the genome-wide mutations of SARS-CoV-2 and detected 2,096 SNVs and seven short deletions. Based on the date of mutation occurring, we calculated the accumulation rate of SNVs in current spreading situation is 6.36e-2/day. Of the mutations, we found 22 SNV hotspots, affecting amino acid sequence of 6 genes, including *ORF1ab*, *N*, *M*, *ORF3a*, *ORF8*, and *S*. The SNV hotspots show a clear continent specific patterns, may resulting in different pandemic situations in different continent. The SNV hotspot at locus 23,403 (exist in 1,656 genome sequence) altering the 614th amino acid sequence

of spike glycoprotein from D to G, may resulting the structure and functional alternation and promoting the adaption of the virus during evolving.

Methods

Datasets

We downloaded SARS-CoV-2 reference genome (GenBank: MN908947) and the genome annotation file GCF_009858895.2_ASM985889v3_genomic.gff) from NCBI (www.ncbi.nlm.nih.gov/nuccore/). In addition, we obtained complete genome sequences of 3,160 SARS-CoV-2 strains from GISAID (https://www.gisaid.org/, downloaded on April 4th, 2020, the sequence Accessions were listed in Supplementary Table S7). All sequences were collected from 6 continents (Supplementary Fig. S5).

Mutation calling

We aligned 3,160 complete sequences to reference genome by BWA [20] and called mutations with SAMtools [21] and BCFtools. We used snpEff [22] to annotate the mutations.

Graph genome construction

We removed the singleton mutations, which occurred only in one strain of SARS-CoV-2 and built its graph genome with variation graph toolkit [15] based on the rest of mutations.

SNV frequency enrichment analysis

We constructed 500 random mutation profiles for each genome sequence with keeping the same mutation numbers. We calculated the mean μ_i and standard deviation σ_i of mutation frequency of position i in random mutation profiles. To determine a real SNV frequency f_i is significantly higher than random

background, we compared f_i with the maximal μ_i and calculated the *p* value by $1-pnorm\ f_i,\mu_{\max},\sigma_{\max}$, where *pnorm* was the normal distribution function in *R*.

Genomic bins enrichment of high frequent SNVs

We shuffled the high frequent SNVs across whole genome 500 times to construct the background SNV distributions. We split whole genome into 30 nonoverlapped 1kb bins, and counted SNV numbers in each bin for both real and background SNV distributions. For the *i*-th bin, we calculated the *p* value by $1-pnorm\ n_i,\mu_i,\sigma_i$, where n_i was the real SNV number in the *i*-th bin, μ_i and σ_i were the mean and standard deviation of SNV number of the *i*-th bin in 500 background SNV distributions.

SNV hotspots location enrichment analysis

For each highly frequent SNV, we calculated the FC of each continent, and did the fish exact test to calculate the *p* value by *fisher.test* function in *R*. The FC was calculated as following.

$$FC_c^i = \frac{N_c^i / N_{global}^i}{N_c / N_{global}} \quad (1)$$

where N_c^i is the number of genome sequences with SNV i from continent c, N_c is the total number of genome sequences from continent c, N_{global}^i is the total number of genome sequences with SNV i, and N_{global} is the total number of genome sequences downloaded from GISAID. Here, N_{global} = 3,159.

Statistical analysis

The statistical analysis were performed in R environment (https://www.r-project.org/).

Abbreviations

COVID-19: coronavirus disease 2019; SARS-CoV-2: severe acute respiratory syndrome–coronavirus 2; SNV: single nucleic variant; FC: fold change.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and material

The data used in our study is downloaded from GISAID (https://www.gisaid.org), the constructed graph genome is available from https://github.com/xjtu-omics/SARS-CoV-2_graph_genome, and the codes are available under request.

Competing interests

The authors declare they have no competing interests.

Funding

This work is supported by the National Science Foundation of China (61702406 and 31671372), the National Science and Technology Major Project of China (grand no. 2018ZX10302205), and the National Key R&D Program of China (2018YFC0910400 and 2017YFC0907500), the General Financial Grant from the China Postdoctoral Science Foundation (2017M623178), and the Major Project of Baoji University of Arts and Sciences (ZK16120). The funding bodies had no role in designing the experiments, collecting the data or writing the manuscript.

Authors' contributions

Both XY and KY designed this research. LL downloaded the data and remove incomplete sequence. XY detected the variation and did the analysis. XY, LL, and KY wrote this article. The authors read and approved the final manuscript.

Acknowledgements

We thank Ningxin Dang for her help to this project. In addition, we thank the sequence provider all over the world and the doctors.

References

1. Zhou P, Yang XL, Wang XG, Hu B, Zhang L, Zhang W, Si HR, Zhu Y, Li B, Huang CL *et al*: **A pneumonia outbreak associated with a new coronavirus of probable bat origin**. *Nature* 2020, **579**(7798):270-273.
2. Wu F, Zhao S, Yu B, Chen YM, Wang W, Song ZG, Hu Y, Tao ZW, Tian JH, Pei YY *et al*: **A new coronavirus associated with human respiratory disease in China**. *Nature* 2020, **579**(7798):265-269.
3. Zhang T, Wu Q, Zhang Z: **Probable Pangolin Origin of SARS-CoV-2 Associated with the COVID-19 Outbreak**. *Curr Biol* 2020, **30**(7):1346-1351 e1342.
4. Chan JF, Kok KH, Zhu Z, Chu H, To KK, Yuan S, Yuen KY: **Genomic characterization of the 2019 novel human-pathogenic coronavirus isolated from a patient with atypical pneumonia after visiting Wuhan**. *Emerg Microbes Infect* 2020, **9**(1):221-236.
5. Wang D, Hu B, Hu C, Zhu F, Liu X, Zhang J, Wang B, Xiang H, Cheng Z, Xiong Y *et al*: **Clinical Characteristics of 138 Hospitalized Patients With 2019 Novel Coronavirus-Infected Pneumonia in Wuhan, China**. *JAMA* 2020.
6. Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, Zhang L, Fan G, Xu J, Gu X *et al*: **Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China**. *Lancet* 2020, **395**(10223):497-506.
7. Guan WJ, Ni ZY, Hu Y, Liang WH, Ou CQ, He JX, Liu L, Shan H, Lei CL, Hui DSC *et al*: **Clinical Characteristics of Coronavirus Disease 2019 in China**. *N Engl J Med* 2020.
8. Liu S, Zheng Q, Wang Z: **Potential covalent drugs targeting the main protease of the SARS-CoV-2 coronavirus**. *Bioinformatics* 2020.
9. Li SR, Tang ZJ, Li ZH, Liu X: **Searching therapeutic strategy of new coronavirus pneumonia from angiotensin-converting enzyme 2: the target of COVID-19 and SARS-CoV**. *Eur J Clin Microbiol Infect Dis* 2020.

10. Walls AC, Park YJ, Tortorici MA, Wall A, McGuire AT, Veesler D: **Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein**. *Cell* 2020, **181**(2):281-292 e286.
11. Ou X, Liu Y, Lei X, Li P, Mi D, Ren L, Guo L, Guo R, Chen T, Hu J *et al*: **Characterization of spike glycoprotein of SARS-CoV-2 on virus entry and its immune cross-reactivity with SARS-CoV**. *Nat Commun* 2020, **11**(1):1620.
12. Yan R, Zhang Y, Li Y, Xia L, Guo Y, Zhou Q: **Structural basis for the recognition of SARS-CoV-2 by full-length human ACE2**. *Science* 2020, **367**(6485):1444-1448.
13. Tang X, Wu C, Li X, Song Y, Yao X, Wu X, Duan Y, Zhang H, Wang Y, Qian Z *et al*: **On the origin and continuing evolution of SARS-CoV-2**. *National Science Review* 2020.
14. Forster P, Forster L, Renfrew C, Forster M: **Phylogenetic network analysis of SARS-CoV-2 genomes**. *Proc Natl Acad Sci U S A* 2020.
15. Garrison E, Siren J, Novak AM, Hickey G, Eizenga JM, Dawson ET, Jones W, Garg S, Markello C, Lin MF *et al*: **Variation graph toolkit improves read mapping by representing genetic variation in the reference**. *Nat Biotechnol* 2018, **36**(9):875-879.
16. Rakocevic G, Semenyuk V, Lee WP, Spencer J, Browning J, Johnson IJ, Arsenijevic V, Nadj J, Ghose K, Suciu MC *et al*: **Fast and accurate genomic analyses using genome graphs**. *Nat Genet* 2019, **51**(2):354-362.
17. Yang X, Lee WP, Ye K, Lee C: **One reference genome is not enough**. *Genome Biol* 2019, **20**(1):104.
18. Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJ: **The Phyre2 web portal for protein modeling, prediction and analysis**. *Nat Protoc* 2015, **10**(6):845-858.
19. Smith EC, Blanc H, Surdel MC, Vignuzzi M, Denison MR: **Coronaviruses lacking exoribonuclease activity are susceptible to lethal**

mutagenesis: evidence for proofreading and potential therapeutics. *PLoS Pathog* 2013, **9**(8):e1003565.

20. Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler transform**. *Bioinformatics* 2009, **25**(14):1754-1760.
21. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S: **The Sequence Alignment/Map format and SAMtools**. *Bioinformatics* 2009, **25**(16):2078-2079.
22. Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM: **A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3**. *Fly (Austin)* 2012, **6**(2):80-92.

Figure Legends

Fig. 1. Correlation between mean number of SNVs and collection date. The blue line is the fitted line and the gray region is the 95% of confidence interval.

Fig. 2. SNV hotspots. **A**. the frequency and genomic locus of 22 SNV hotspots. The genomic annotation plot is downloaded from UCSC genome browser. **B.** Continent specific enrichment of SNV hotspots. The red block means significant enrichment with $p < 0.01$, fisher-exact test. The column label is “locus, SNV type”, e.g. 28,144, T > C denotes T to C substitution at genome locus 28,144.

Fig. 3. SNV at locus 23,403 related with *S*. **A.** Relations between collection date and the proportion of genome sequences with this substitution. To avoid the bias, we removed the date with smaller than 10 genome sequences. **B.** Mutation sensitivity colored spike glycoprotein structure. The red region in the structure is the 614^{th} amino acid, which is change from D to G since this SNV. The mutation sensitivity score of the 614^{th} amino acid is 4.

Figures

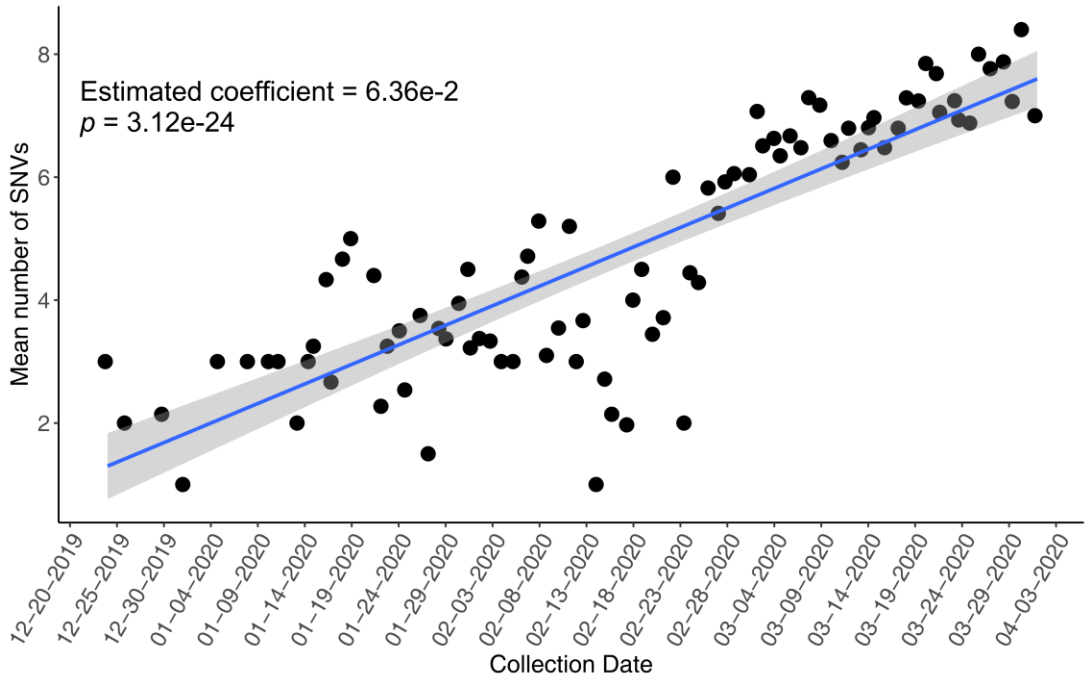

Fig. 1. Correlation between mean number of SNVs and collection date. The blue line is the fitted line and the gray region is the 95% of confidence interval.

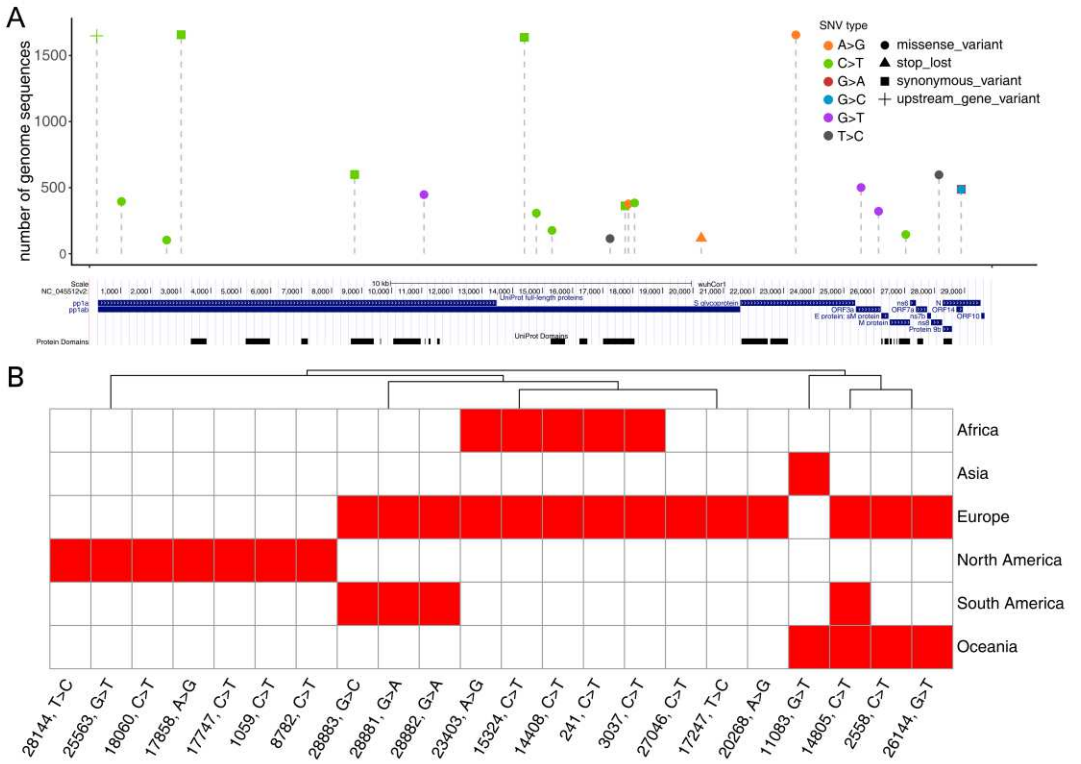

Fig. 2. SNV hotspots. **A**. the frequency and genomic locus of 22 SNV hotspots. The genomic annotation plot is downloaded from UCSC genome browser. **B.** Continent specific enrichment of SNV hotspots. The red block means significant enrichment with $p < 0.01$, fisher-exact test. The column label is "locus, SNV type", e.g. 28,144, T > C denotes T to C substitution at genome locus 28,144.

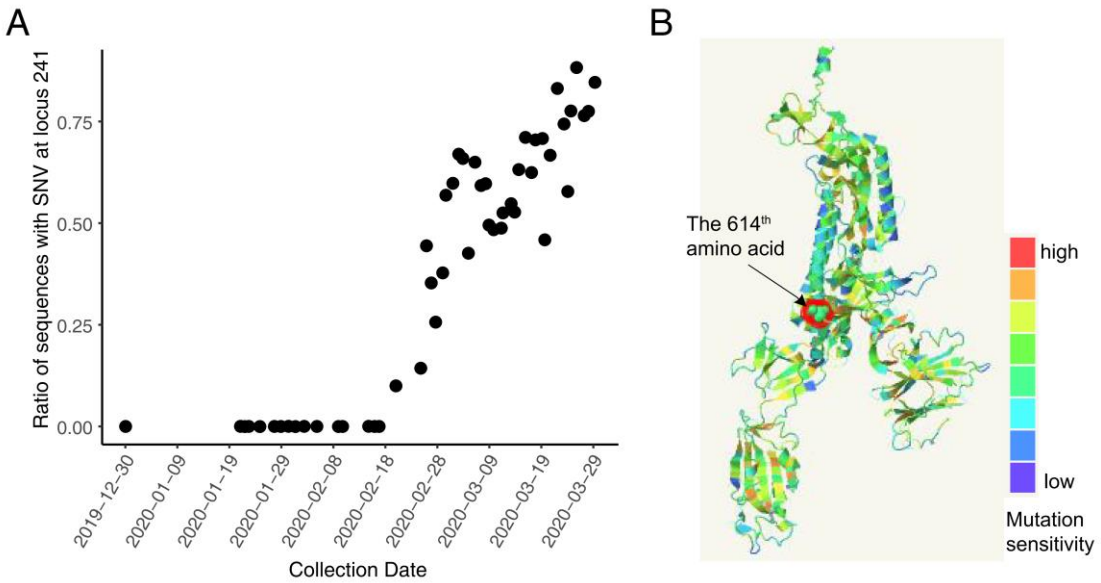

Fig. 3. SNV at locus 23,403 related with *S*. **A.** Relations between collection date and the proportion of genome sequences with this substitution. To avoid the bias, we removed the date with smaller than 10 genome sequences. **B.** Mutation sensitivity colored spike glycoprotein structure. The red region in the structure is the 614^{th} amino acid, which is change from D to G since this SNV. The mutation sensitivity score of the 614^{th} amino acid is 4.

Figures

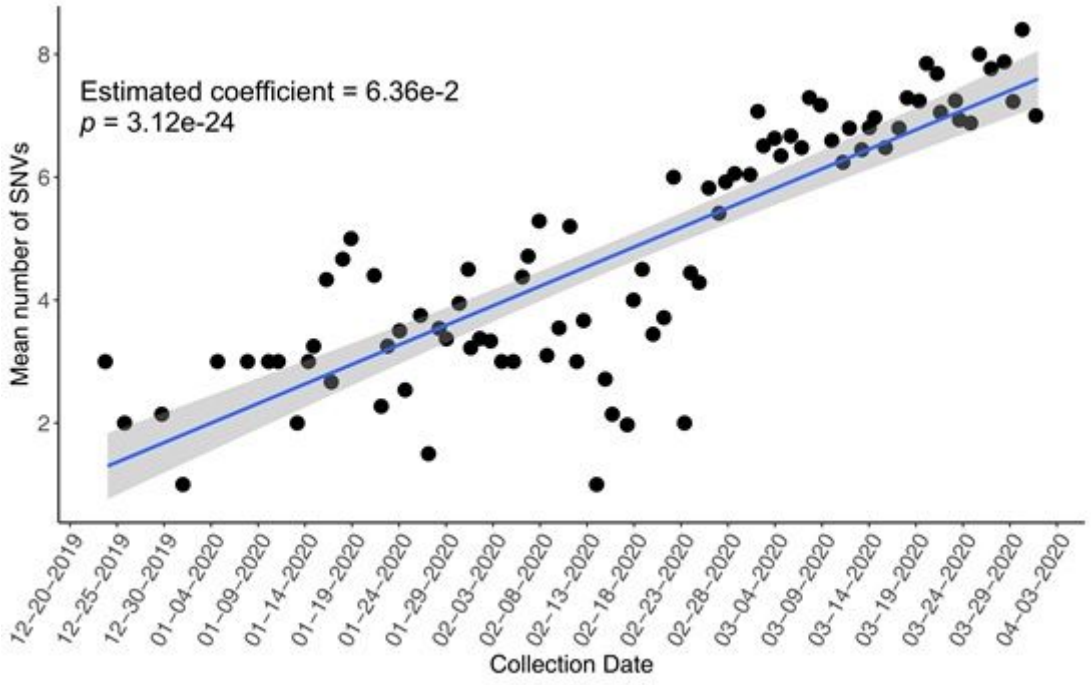

Figure 1

Correlation between mean number of SNVs and collection date. The blue line is the fitted line and the gray region is the 95% of confidence interval.

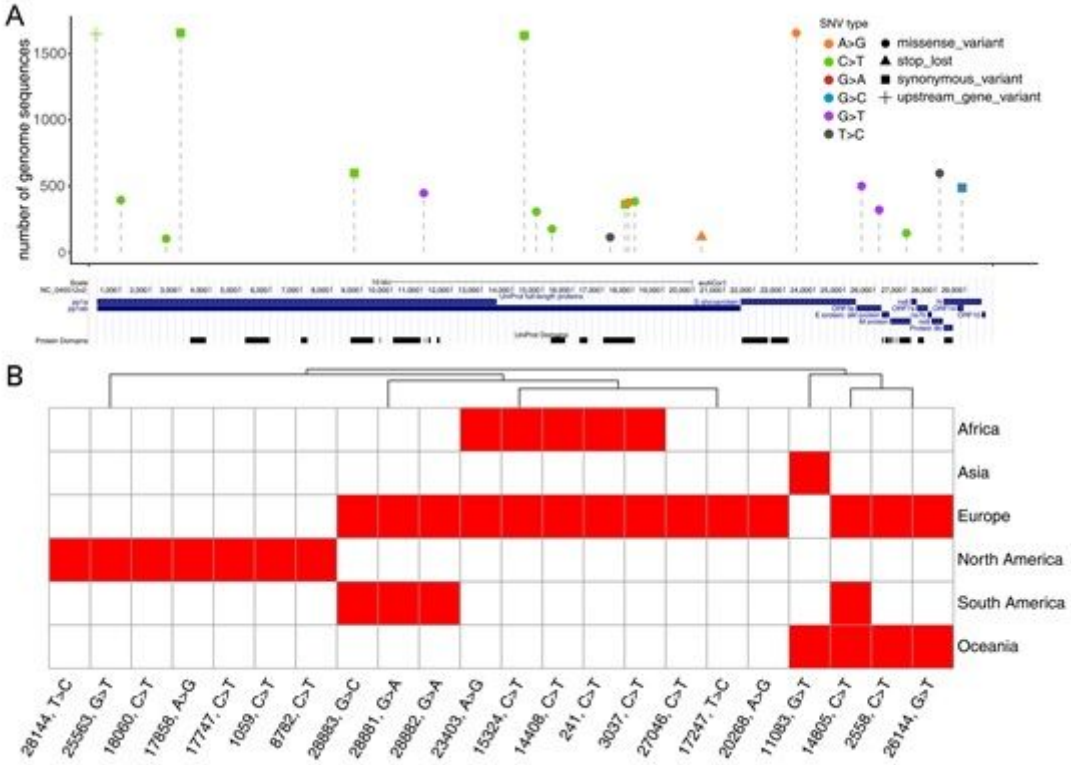

Figure 2

SNV hotspots. A. the frequency and genomic locus of 22 SNV hotspots. The genomic annotation plot is downloaded from UCSC genome browser. B. Continent specific enrichment of SNV hotspots. The red block means significant enrichment with p < 0.01, fisher-exact test. The column label is “locus, SNV type”, e.g. 28,144, T > C denotes T to C substitution at genome locus 28,144.

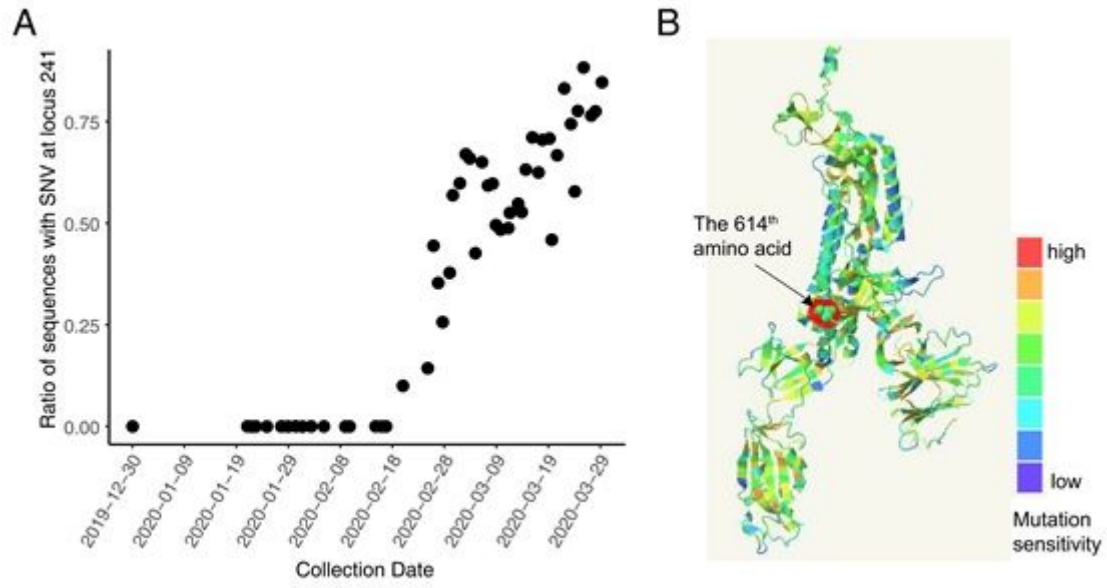

Figure 3

SNV at locus 23,403 related with S. A. Relations between collection date and the proportion of genome sequences with this substitution. To avoid the bias, we removed the date with smaller than 10 genome sequences. B. Mutation sensitivity colored spike glycoprotein structure. The red region in the structure is the 614th amino acid, which is change from D to G since this SNV. The mutation sensitivity score of the 614th amino acid is 4.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- themutationswithannotation.vcf.txt
- SupplementaryTables.xlsx
- SupplementaryMaterials.docx