Pages: 16
Words: 3011
Tables: 2
Figures: 2
Supplemental Figures: 3
References: 20
Contact: Ashley E. Nazario-Toole
Email: ashley.e.nazario-toole.civ@mail.mil
Guarantor: Ashley E. Nazario-Toole

Whole-genome Sequencing of SARS-CoV-2: Using Phylogeny & Structural Modeling to Contextualize Local Viral Evolution

Running Head: Contextualizing Local SARS-CoV-2 Genomes

Key Words: SARS-CoV-2, Next-generation Sequencing, Phylogenetics, Viral Evolution, Vaccine Effectiveness

Ashley E. Nazario-Toole, PhD.[1]

Hui Xia[1]

Thomas F. Gibbons, PhD.[1]

[1]59th Medical Wing, Clinical Investigations & Research Support Laboratory, Lackland Air Force Base, 1100 Wilford Hall Loop, Lackland AFB, TX 78236

Funding/COI: This study was funded by the 59MDW/STC. The authors declare no conflict of interest.

Acknowledgments: We gratefully acknowledge the authors, originating and submitting laboratories of the sequences from GISAID's EpiFlu Database on which the phylogenetic pathogen build is based. We also thank Sherry Trevino and the 59MDW Clinical Laboratory for providing the excess clinical specimens sequenced in this research.

Disclaimer: The views expressed are solely those of the authors and do not reflect the official policy or position of the US Army, US Navy, US Air Force, the Department of Defense, or the US Government.

Abstract:

The outbreak of SARS-CoV-2 has created a global pandemic resulting in over 1 million deaths worldwide. Rapid estimations of transmission and mutational patterns of virus outbreaks can be accomplished using whole genome viral sequencing. Here we report the development of a local pipeline for molecular epidemiological surveillance enabling DoD public health officials to track viral evolution and outbreaks. Sequencing of clinical specimens revealed that by June 2020, SAR-CoV-2 strains carrying the 614G mutation were the predominant cause of COVID-19 infections at JBSA/Lackland. Furthermore, we identified and mapped six additional spike protein amino acid changes, information which could potentially aid vaccine design. The sequencing and phylogenetic workflow described in this paper will enable local officials to track and better understand virus transmission events. Overall, this work could improve long-term readiness efforts by providing a mechanism for analyzing the current SARS-CoV-2 pandemic as well as future disease outbreaks.

Introduction:

In December 2019, whole genome shotgun sequencing of respiratory tract samples revealed that a novel RNA virus from the genus betacoronavirus was the causal agent of pneumonia in patients from Wuhan, China[1–3]. The virus, Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2), has since spread globally, causing approximately 30 million infections and over 950,000 deaths[4].

Prior to the discovery of SARS-CoV-2, only six coronaviruses were known to cause human disease – four which cause the common cold (HCoV-OC43, HCoV-HKU1, HCoV-NL63, and HCoV-229E) and two other strains of zoonotic origin, severe acute respiratory syndrome (SARS-CoV) and Middle East respiratory syndrome virus (MERS-CoV)[5]. SARS-CoV and MERS-CoV caused fatal respiratory disease outbreaks in 2002 and 2003 (SARS-CoV) and 2012 (MERS-CoV)[6–8]. Though phylogenetic analysis of full-length SARS-CoV-2 genomes indicates that a single zoonotic transmission event, from bats to humans, occurred mid-November 2019, the origin of the virus is still unclear[9]. SARS-CoV-2 shares 94.4% genome sequence identity with SARS-CoV but is most closely related to the bat SARSr-CoV RaTG13, sharing 96.2% sequence identity and 97% amino acid identity with the RaTG13 spike glycoprotein[10,11].

The public health response to COVID-19 has been facilitated by the unprecedented efforts of scientists sharing sequence data from clinical isolates worldwide. NCBI's GenBank and the Global Initiative on Sharing All Influenza Data (GISAID) have served as the primary repositories for SARS-CoV-2 full genome sequences. Over 100,000 sequences have been uploaded to date and these sequences have enabled the estimation of the virus mutation

rate[12]. The most comprehensive mutational analysis published thus far (n > 48,800 SARS-CoV-2 complete genomes) reported an average of 7.23 mutations per sample; relative to the reference Wuhan genome NC_045512.2[13]. The global sequencing initiative also facilitated the identification of a clade newly emerged viruses carrying the D614G spike protein substitution, caused by an A to G mutation at nucleotide position 23,403[14]. Over the course of a single month, the D614G became the globally dominant form of SARS-CoV-2; especially in Europe and North America. It is suspected that mutation may confer a fitness advantage, as higher levels of viral shedding have been observed in G614 infected patients and higher *in vitro* infectious titers have been associated with G614-bearing viruses[14]. However, in patients infected with the G614 variant, viral load and clinical outcomes are not always correlated, indicating that the mutation is less important for COVID-19 disease oucomes[15,16]. Nevertheless, the discovery the D614G mutation illustrates the how SARS-CoV-2 genome sequencing informs our understanding of the biology and epidemiology of the virus. Additionally, the identification of natural polymorphisms in genes encoding the SARS-CoV-2 structural proteins, spike(S), envelope(E), membrane(M), and nucleocapsid(N), is crucial for vaccine design.

As of September 18, 2020, there have been over 40,000 confirmed SARS-CoV-2 infections in DoD personnel (military, civilian, dependents & contractors). The availability of local tools for viral genomic epidemiology surveillance could aid DoD medical and public health officials tasked with monitoring the health and readiness of military service members. Here we present a sequencing & bioinformatics workflow for stand-alone, real-time tracking of pathogen evolution at Lackland Air Force Base, TX (JBSA/Lackland). This workflow could be

adopted for other DoD installations, and could also improve long-term readiness efforts by providing a mechanism for analyzing future disease outbreaks.

Methods:

Sample Collection and Total RNA Isolation

Excess clinical specimens (nasopharyngeal swabs) were diluted 1:1 in Zymo DNA/RNA Shield (Zymo Research©, cat. R1100-250) and frozen at -80℃ until RNA isolation. 500 µL of sample was isolated using the Qiagen© RNeasy Mini Kit (Qiagen, cat. 74106), following the manufacturer's protocol.

Real-Time PCR to determine viral load pre-sequencing

A modified version of the CDC 2019-nCoV Real-Time rRT-PCR panel (CDC/DHHS) was used to detect SARS-CoV-2 RNA from July clinical specimens prior to whole genome sequencing. (Note: In an effort to focus on competency development, June specimens were directly used for sequencing library preparation.) Primers and probes designed by the CDC were commercially synthesized and comprised two 2019-nCoV-specific sets (N1, N2). Reactions were prepared using the TaqMan® RNA-to-Ct™ 1-Step Kit (ThermoFisher, cat. 4392656). A positive control, Genomic RNA from SARS-Related Coronavirus2, Isolate, USA-WA1/2020 (ATCC, cat. VR-1986D) and a non-template control, nuclease-free water, were included with each run. 20 µL RT-PCR reactions (5µL of RNA and 15µL target master mix) were performed on the ABI StepOnePlus™ Real-Time PCR system using the following conditions: 48℃ for 15 min, 95℃ for 10 min, followed by 45 cycles of 95℃ for 15 sec and 60℃ for 1 min. Quantitation of FAM-labeled

probes occurred at the end of each cycle. Amplicon-based SARS-CoV-2 genome sequencing was carried out on all samples with N1 CT < 25 (n=29 July samples).

Tiling: Amplicon-based SARS-CoV-2 genome sequencing

Paragon Genomics' CleanPlex® SARS-CoV-2 Panel (cat. 918011) for Illumina platforms was used to prepare sequencing libraries (starting concentration of 10-50 ng RNA per sample). As a positive control, sequencing libraries were also prepped for VR-1986D, Genomic RNA from SARS-Related Coronavirus 2, Isolate USA-WA1/2020. Library quality and concentration was assessed via fragment analysis using Advanced Analytics' High Sensitivity NGS Fragment Analysis Kit (cat. DNF-474-0500). For each sample, a library quality ratio score (QRS) was determined by dividing the fragment analysis trace 250-350 bp peak concentration (ng/µL) by 150-190 bp fragment peak concentration (ng/µL): excellent (QRS >10), Good (QRS 1.0 – 10), Fair (QRS <1 and >0.5), Poor (QRS < 0.5). Libraries were denatured and diluted to a final loading concentration of 1.5 pM following the Illumina NextSeq System Denature and Dilute Libraries Guide (Document # 15048776 v09), and then sequenced on the NextSeq 500 system at 2 x 151 bp using the NextSeq Mid Output v2, (300 cycle) kit (Illumina, cat. 15057939). Poor libraries (QRS < 0.5) were excluded from downstream bioinformatics analysis after sequencing.

Illumina adaptor sequences were trimmed using the BaseSpace Onsite FASTQ Toolkit v1.0.0. Primer sequences were removed using the fgbio toolkit, installed in a Linux environment (http://bioconda.github.io/recipes/fgbio/README.html) and a tab delimited file with primer genomic coordinates provided by Paragon Genomics. Adapter/primer trimmed FASTQ files were aligned to the SARS-CoV-2 reference genome (NC_045512.2) using Illumina DRAGEN Bio-

IT Platform. Genome coverage uniformity and mapping was visualized in IGV (BAM and VCF files) and consensus FASTA files were created using the fgbio toolkit.

Probe: Hybrid-Capture Target Enrichment Sequencing

First and second strand cDNA synthesis was carried with Maxima H Minus dsDNA Synthesis Kit (Thermo Scientific, cat. K2561) from 13 µL RNA. Synthesized cDNAs were cleaned with Qiagen's QIAquick PCR Purification Kit (Qiagen, cat #28104) and the 260/280 ratio and concentration were determined using NanoDrop. Enriched NGS libraries were constructed using Illumina's Nextera DNA Flex Pre-Enrichment kit (cat. 20025524), the IDT for Illumina UD Indexes Set A (cat. 20027213), and the Respiratory Virus Oligos Panel (cat. 20042472), which contains ~7800 probes to detect respiratory viruses, including SARS-CoV-2.

Briefly, 300-500 ug of cDNA was underwent tagmentation, clean-up, and pre-enrichment amplification. After amplification, individual libraries underwent probe hybridization, probe capture, enrichment amplification, and quantification. Libraries were pooled, denatured, diluted to a concentration of 1.5 pM and sequenced at 2 x 151 bp using the NextSeq Mid Output v2, (300 cycle) kit (Illumina, cat. 15057939).

Adaptor sequences were trimmed using BaseSpace Onsite and trimmed, concatenated FASTQ files were aligned to the SARS-CoV-2 reference genome (NC_045512.2) using Illumina DRAGEN Bio-IT Platform. Genome coverage uniformity and mapping was visualized in IGV (BAM and VCF files) and consensus FASTA files were created using the fgbio toolkit.

Shotgun metagenomics sequencing

cDNA was reverse transcribed (without DNase treatment) from 5 µL of RNA per sample using the Ovation® RNA-Seq System kit (Tecan, cat. 7102). The cDNA concentration was determined by fragment analysis and the maximum volume of input DNA was used to prepare shotgun libraries using the Illumina Nextera DNA Flex Library Prep kit (cat. 20018704) and Nextera DNA CD Indexes (cat. 20018707). Fragment analysis (Adv. Analytical, cat. DNF-474-0500) was used to determine library size and concentration and the libraries were normalized, denatured, and then diluted to a loading concentration of 1.8 pM. A NextSeq 500/550 High Output v2, 150 cycles kit (Illumina, cat. 15057931) was used to sequence at 2 x 76 bp.

Adaptor sequences were trimmed as previously described and the BaseSpace Onsite Kraken Metagenomics app was used for Shotgun sequencing analysis. The SNAP aligner was used to filter human sequences by aligning to RefSeq hg 19. Alignment to the SARS-CoV-2 genome and generation of consensus sequences FASTA files was carried out as described above.

Phylogenetic tree construction

The Nexstrain conda environment, Auger (bioinformatics tooling) and auspice (the Nextstrain visualization app), was downloaded and installed locally[17]. The Nextstrain SARS-CoV-2 tutorial, including snakemake file with auger commands, was downloaded (https://nextstrain.github.io/ncov) and utilized as the foundation of the CIRS SARS-CoV-2 pathogen build. An additional 48 global SARS-CoV-2 FASTA sequences (from 5/3/2020 – 8/14/2020) were downloaded from GSAID and added to the pathogen build. Next, the consensus genome sequences (FASTA) for CIRS clinical specimens (n = 39, with two duplicate samples) and FASTA's for the positive control samples (n = 2) were added. Metadata for the

CIRS specimens includes the following: Sample ID (PIN), date collected, location (North America, TX, 59MDW), and date uploaded. 59MDW viral genomes (FASTA sequences) were uploaded to GSAID (post-Public Affairs clearance).

To build phylogenetic trees in Nextstrain, the Snakemake command was run, referencing CIRS sequences and the Nextstrain CIRS SARS-CoV-2 pathogen build, and auspice was used to visual the output .json file in the Nextstrain environment.

Protein structural analysis

Cryo-EM three-dimensional structure of SARS-CoV-2 spike glycoprotein available in RCSB Protein Data Bank were reviewed and the PDB structure 6VYB was downloaded[18]. SWISS-MODEL (https://swissmodel.expasy.org) was utilized with homology modeling online server with Spike glycoprotein NCBI Reference Sequence: YP_009724390.1 to fill in missing residues. The resulting file was utilized in PyMOL to create renderings of the spike protein with predicted epitope regions per Grifoni *et al.* colored black, and specific areas of the trimer colored as in Wrapp *et al*[19,18].

Results:

Under an IRB exempt protocol, we received 15 nasopharyngeal swab (NP) specimens from the Wilford Hall Clinical Laboratory in June 2020. Fourteen of the specimens tested positive for SARS-CoV-2 at the WHMC laboratory, while one was negative. Upon arrival at the CIRS lab, samples were assigned a 4-digit PIN number and all other identifying information was removed. Total RNA was extracted using the Qiagen RNeasy Mini kit. Paragon Genomics' CleanPlex® SARS-CoV-2 Panel, an amplicon-based next-generation sequencing library

preparation method, was used to prepare sequencing libraries from all 15 clinical specimens (Sample 2875 in duplicate) and from a positive control, Genomic RNA from SARS-Related Coronavirus 2, Isolate USA-WA1/2020 (Figure S1). We then used fragment analysis to calculate a library quality ratio score (QRS) where the concentration (ng/uL) of fragments of the expected library amplicon size (250-350bp) was normalized the by the concentration of fragments which represented primer dimers and non-specific products (150-190bp). Of the 16 specimen libraries, five (2913, 2965, 3124, 2875A and 2875B) had excellent QRS scores (QRS>10), 2973 and 3014 were good (QRS 1.0 – 10), 3095 was fair (QRS <1 and >0.5), and the remaining nine were poor (QRS < 0.5). The QRS score for the positive control library (PC-June) was excellent. All samples were sequenced at 2x151 bp long reads on a NextSeq Mid Output flow cell. Once the CleanPlex primer sequences were removed from the reads, only libraries with Good & Excellent QRS scores mapped to the SARS-CoV-2 reference genome (NC_045512.2) at over 99.95% genome coverage. Thus, all poor tiling libraries from the June samples were excluded from further downstream analysis. We then re-sequenced a subset of the June specimens using shotgun-metagenomic and probe (hybrid-capture) sequencing techniques in order to cross validate our SARS-CoV-2 whole genome sequencing and determine if these methods were more or less sensitive to low viral loads than tiling preps (**Figure S1**). SARS-CoV-2 was successfully sequenced from only one of the “poor” tiling samples, 2576, with both shotgun and probe methods (97.81% and 100% 20X genome coverage for shotgun and probe, respectively). Another “poor” sample, 2699, sequenced at 20X depth of 73.38% for shotgun and at 100% for probe. We were able to use the probe sequence reads for 2576 and 2699 to generate variant call files (VCF) and SARS-CoV-2 consensus genome FASTA sequences. Based on these results, we

concluded that probe-based sequencing is a useful alternative to tiling and shotgun sequencing when working with samples with low viral loads. However, despite the increased sensitivity, the preparation of probe-based sequencing libraries is 3X longer than tiling library prep - 3 days to prepare a single probe library versus 1 day per tiling library. In the event that actionable sequencing data is required, tiling library prep is the most efficient and effective method, especially for specimens with high viral loads (N1 C_T > 25).

We received an additional 38 exempt clinical specimens in July. Prior to sequencing, we used a modified version of the CDC real-time PCR assay to determine the C_T values for the N1 and N2 targets. Specimens with N1 C_T > 25 (n=29) and the USA-WA1/2020 genomic RNA (PC-July) were used to prepare tiling sequence libraries. As before, a single specimen (3667) was sequenced in duplicate. All 31 samples yielded either good or excellent QRS scores, and consensus SARS-CoV-2 genome sequences were generated for all libraries.

To visualize the phylogeny of the 59MDW samples, we created a pathogen build for a locally installed version of Nextstrain, an open-source visualization tool from the GASAID (gisaid.org) (Hadfield). First, we modified the Nextstrain SARS-CoV-2 tutorial, which contained 419 global SARS-CoV-2 genomes, by adding an additional 48 specimens sequenced between May and August 2020, bringing the total number of global SARS-CoV-2 sequences in the build to 467. These global strains were collected in different regions and classified into one of five distinct clades identified in Nextstrain – 19A, 19B, 20A, 20B, and 20C. Next, the SARS-CoV-2 sequences and metadata from 59MDW (n=41) were added to the build. Finally, the completed build was used to generate a .json file containing 342 viral genomes. This file was then run in

auspice, the program to render Nextstrain visualizations, creating time and divergence phylogenies, diversity plots, and map-based visualizations of the data (**Figure 1**).

We observed a single nucleotide change in both positive control specimens (PC-June, PC-July) when compared to the reference sequence for USA-WA1/2020 (**Figure 1**). As the change was observed in PC samples sequenced using all three methods (tiling, probe, and shotgun), we suspect that the change could be due to the fact that the USA-WA1/2020 virus was purchased from ATCC (cat # VR-1986D), and may have undergone multiple passages through cell culture prior to RNA extraction. Further, analysis of the phylogenetic tree and an alignment of the variant call files (IGV VCF alignment) verified that genomes of both specimens sequenced in duplicate (2875 and 3667) were identical **(Figure S2**). Additionally, the analysis revealed that several individuals carried the same viral strains (**Figure S3**). Finally, we did an in-depth analysis of a cluster of infections in the 20C clade by zooming in on the cluster in the Nextstrain divergence phylogeny and using IGV VCF alignment IGV (**Figure S2**). In this single cluster of 15 specimens, 3 individuals shared the same virus. Furthermore, we found two other sets of individuals who had the same virus, indicative of multiple virus transmission events within the JBSA population. Overall, we identified 109 nucleotide changes in the coding region of the SARS-CoV-2 genome (which led to 63 unique, non-synonymous amino acid mutations), one mutation in the 5'UTR, and two mutations in the 3'UTR (**Table 1**).

Next, we focused our analysis on nucleotide mutations which cause non-synonymous amino acid changes the viral structural proteins (**Table 2**). Seven non-synonymous changes in the spike protein, one in the envelope, and 6 in the nucleocapsid were identified. Interestingly, no mutations in the membrane were identified. All 37 viruses contained the G614 variant

(nucleotide 23,403 A>G), which is expected based on global sequence data and the dominance of the D614G in North America. The second most frequent spike mutation (6/37 samples) was found in nucleotide 23,625 C>T, which caused a change in amino acid 688 from Alanine (A) to Valine (V). Of note, 688 is the final amino acid the novel furin cleavage site of the SARS-CoV-2 spike glycoprotein[20]. Five viruses contained a mutation in amino acid 892 from Alaline (A) to Serine (S). Additional spike mutations were: R21T, D228Y, T732A, and A1070S, each identified in single viruses. To visualize the effect of these mutations on the viral spike protein, we utilized PyMOL to overlay the mutations (shown in grey) on the prefusion CryoEM rendering of the trimeric spike glycoprotein, with predicted linear B cell epitopes shown in black[19]. Interestingly, the virus isolated from sample 3784 carried three spike protein mutations (D614G, A688V, and A1070S) (**Figure 2A**), more than was observed for any other specimen. The model for viruses carrying mutations at D614G and A892S shows that amino acid 892 lies on the surface of the S2 domain, but not within a predicted immune epitope (**Figure 2B**). The remaining non-synonymous spike mutations were low frequency (found in only a single virus). Nevertheless, plotting these changes on the spike protein model reveals that mutation amino acid 732 is embedded within the S2 domain of the protein, and may thus be inaccessible to the host immune response (**Figure 2C**). Amino acid 288 lies within the N-terminal domain, but in close proximity to the receptor binding domain (**Figure 2D**).

Discussion:

The present study describes a local pipeline for molecular epidemiological surveillance to track viral evolution and outbreaks. The workflow described in this paper: viral RNA isolation, viral load quantification, tiling-based next-generation sequencing, sequencing & bioinformatics

analysis, and data visualization can be accomplished in less than a week using tools available at the JBSA/Lackland CIRS laboratory. Tracking viral mutations is essential for effective vaccine design and sequencing may also identify super-spreading transmission events.

Ultimately, this work will enable public health and infectious disease officials to utilize secure, password protected, phylogenetic and protein modeling data for rapid assessment of a variety of questions with respect to SARS-CoV-2. The fast turn-around time could be useful for local officials seeking real-time information on the overall on the health and readiness of military service members, dependents, civilians, and contractors.

References:

1. Zhu N, Zhang D, Wang W, et al. A Novel Coronavirus from Patients with Pneumonia in China, 2019. *N Engl J Med*. 2020;382(8):727-733. doi:10.1056/NEJMoa2001017

2. Chan JF-W, Yuan S, Kok K-H, et al. A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. *The Lancet*. 2020;395(10223):514-523. doi:10.1016/S0140-6736(20)30154-9

3. Huang C, Wang Y, Li X, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet*. 2020;395(10223):497-506. doi:10.1016/S0140-6736(20)30183-5

4. *WHO Coronavirus Disease (COVID-19) Dashboard. Geneva: World Health Organization, 2020. Available Online: Https://Covid19.Who.Int/ (Last Cited: 9/20/2020).*

5. Su S, Wong G, Shi W, et al. Epidemiology, Genetic Recombination, and Pathogenesis of Coronaviruses. *Trends Microbiol*. 2016;24(6):490-502. doi:10.1016/j.tim.2016.03.003

6. Cui J, Li F, Shi Z-L. Origin and evolution of pathogenic coronaviruses. *Nat Rev Microbiol*. 2019;17(3):181-192. doi:10.1038/s41579-018-0118-9

7. Drosten C, Günther S, Preiser W, et al. Identification of a Novel Coronavirus in Patients with Severe Acute Respiratory Syndrome. *N Engl J Med*. 2003;348(20):1967-1976. doi:10.1056/NEJMoa030747

8. Zaki AM, van Boheemen S, Bestebroer TM, Osterhaus ADME, Fouchier RAM. Isolation of a Novel Coronavirus from a Man with Pneumonia in Saudi Arabia. *N Engl J Med*. 2012;367(19):1814-1820. doi:10.1056/NEJMoa1211721

9. Rambaut A. Phylodynamic analysis of SARS-CoV-2 genomes, 6 Mar-2020. *Virological*. March 2020. https://virological.org/t/phylodynamic-analysis-176-genomes-6-mar-2020/356; ; accessed September 20, 2020.

10. Zhou P, Yang X-L, Wang X-G, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature*. 2020;579(7798):270-273. doi:10.1038/s41586-020-2012-7

11. Wan Y, Shang J, Graham R, Baric RS, Li F. Receptor Recognition by the Novel Coronavirus from Wuhan: an Analysis Based on Decade-Long Structural Studies of SARS Coronavirus. Gallagher T, ed. *J Virol*. 2020;94(7):e00127-20. doi:10.1128/JVI.00127-20

12. Koyama T, Platt D, Parida L. Variant analysis of SARS-CoV-2 genomes. *Bulletin World Health Organization*. 2020;(98):495-504. doi:http://dx.doi.org/10.2471/BLT.20.253591

13. Mercatelli D, Giorgi FM. Geographic and Genomic Distribution of SARS-CoV-2 Mutations. *Frontiers in Microbiology*. 2020;11:1800. doi:10.3389/fmicb.2020.01800

14. Korber B, Fischer WM, Gnanakaran S, et al. Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus. *Cell*. 2020;182(4):812-827.e19. doi:10.1016/j.cell.2020.06.043

15. Becerra-Flores M, Cardozo T. SARS-CoV-2 viral spike G614 mutation exhibits higher case fatality rate. *Int J Clin Pract*. May 2020:e13525-e13525. doi:10.1111/ijcp.13525

16. Isabel S, Graña-Miraglia L, Gutierrez JM, et al. Evolutionary and structural analyses of SARS-CoV-2 D614G spike protein mutation now documented worldwide. *Scientific Reports*. 2020;10(1):14031. doi:10.1038/s41598-020-70827-z

17. Hadfield J, Megill C, Bell SM, et al. Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics*. 2018;34(23):4121-4123. doi:10.1093/bioinformatics/bty407

18. Wrapp D, Wang N, Corbett KS, et al. Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science*. 2020;367(6483):1260-1263. doi:10.1126/science.abb2507

19. Grifoni A, Sidney J, Zhang Y, Scheuermann RH, Peters B, Sette A. A Sequence Homology and Bioinformatic Approach Can Predict Candidate Targets for Immune Responses to SARS-CoV-2. *Cell Host Microbe*. 2020;27(4):671-680.e2. doi:10.1016/j.chom.2020.03.002

20. Xing Y, Li X, Gao X, Dong Q. Natural Polymorphisms Are Present in the Furin Cleavage Site of the SARS-CoV-2 Spike Glycoprotein. *Front Genet*. 2020;11:783-783. doi:10.3389/fgene.2020.00783