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

Ashley E. Nazario-Toole, PhD; Hui Xia; Thomas F. Gibbons, PhD

ABSTRACT

Introduction:

The outbreak of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has created a global pandemic resulting in over 1 million deaths worldwide. In the Department of Defense (DoD), over 129,000 personnel (civilians, dependents, and active duty) have been infected with the virus to date. Rapid estimations of transmission and mutational patterns of virus outbreaks can be accomplished using whole-genome viral sequencing. Deriving interpretable and actionable results from pathogen sequence data is accomplished by the construction of phylogenetic trees (from local and global virus sequences) and by the creation of protein maps, to visualize and predict the effects of structural protein amino acid mutations.

Materials and Methods:

We developed a sequencing and bioinformatics workflow for molecular epidemiological SARS-CoV-2 surveillance using excess clinical specimens collected under an institutional review board exempt protocol at Joint Base San Antonio, Lackland AFB. This workflow includes viral RNA isolation, viral load quantification, tiling-based next-generation sequencing, sequencing and bioinformatics analysis, and data visualization via phylogenetic trees and protein mapping.

Results:

Sequencing of 37 clinical specimens collected at JBSA/Lackland revealed that by June 2020, SAR-CoV-2 strains carrying the 614G mutation were the predominant cause of local coronavirus disease 2019 infections. We identified 109 nucleotide changes in the coding region of the SARS-CoV-2 genome (which lead to 63 unique, non-synonymous amino acid mutations), one mutation in the 5′-untranslated region (UTR), and two mutations in the 3′UTR. Furthermore, we identified and mapped six additional spike protein amino acid changes—information which could potentially aid vaccine design.

Conclusion:

The workflow presented here is designed to enable DoD public health officials to track viral evolution and conduct near real-time evaluation of future outbreaks. The generation of molecular epidemiological sequence data is critical for the development of disease intervention strategies—most notably, vaccine design. Overall, we present a streamlined sequencing and bioinformatics methodology aimed at improving long-term readiness efforts in the DoD.

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 68 million infections and over 1.5 million deaths.[4]

Before the discovery of SARS-CoV-2, only six coronaviruses were known to cause human disease—four that 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 coronavirus disease 2019 (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

59th Medical Wing, Clinical Investigations and Research Support Laboratory, Lackland Air Force Base, Lackland AFB, TX 78236, USA

The views expressed are solely those of the authors and do not reflect the official policy or position of the U.S. Army, the U.S. Navy, the U.S. Air Force, the Department of Defense, or the U.S. government.

doi:10.1093/milmed/usab031

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 of 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 oucomes.[15,16] Nevertheless, discovery of the D614G mutation illustrates 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 December 6, 2020, there have been over 129,000 confirmed SARS-CoV-2 infections in DoD personnel (military, civilian, dependents, and contractors).[17] 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 and 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.

MATERIALS AND METHODS

Study Objective and Overview

The primary objective of this work was to verify and evaluate three next-generation sequencing (NGS) methods for full-genome sequencing of SARS-CoV-2 and to develop a bioinformatics and visualization workflow for dissemination of sequence data to DoD and public health officials (Fig. S1). Three NGS techniques were evaluated with both commercially available controls and excess clinical specimens collected at JBSA/Lackland Wilford Hall Ambulatory and Surgery Center (WHASC). The three NGS methods are as follows: Amplicon-based (Tiling), hybrid-capture target enrichment (Probe), and shotgun metagenomic (Shotgun) sequencing.

Sample Collection

Nasopharyngeal swab (NPS) specimens were collected from JBSA/Lackland military members and beneficiaries from May 14, 2020 to July 28, 2020. Samples were tested for SARS-CoV-2 by the WHASC Clinical Laboratory before transfer to the Clinical Investigations and Research Support (CIRS) laboratory. Excess clinical specimens were obtained from WHASC under an institutional review board (IRB) exempt protocol (IRB reference number FWH20200103E). Upon receipt by the CIRS laboratory, samples were de-identified as follows: The accession number utilized by the clinical laboratory was coded and recorded as Sample Pin number and all other identifiers on the specimen were destroyed. Fifteen samples were obtained in June, of which only nine produced high-quality SARS-CoV-2 sequences. Thirty-six specimens were obtained in July, of which 28 yielded high-quality SARS-CoV-2 sequences. Two samples, one from June and another from July, were sequenced in duplicate. As a positive control, two sequencing libraries were prepped from Genomic RNA from SARS-Related Coronavirus 2, Isolate USA-WA1/2020 (ATCC cat. VR-1986D).

RNA Isolation

The NPS samples were diluted 1:1 in Zymo DNA/RNA Shield (Zymo Research, cat. R1100-250) and frozen at −80°C until RNA isolation. 500 μL of sample were isolated using the Qiagen RNeasy Mini Kit (Qiagen, cat. 74106), following the manufacturer's protocol.

Real-Time PCR

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 before whole-genome sequencing (Table S1). Note that, in an effort to focus on competency development, all June specimens were directly used for sequencing library preparation and only a subset was tested using qPCR. Primers and probes designed by the CDC were commercially synthesized and comprised two 2019-nCoV-specific sets (N1 and N2). Reactions were prepared using the TaqMan RNA-to-Ct 1-Step Kit (ThermoFisher, cat. 4392656). A positive control, Genomic RNA from SARS-Related Coronavirus 2, Isolate USA-WA1/2020 (ATCC, cat. VR-1986D) and a non-template control, nuclease-free water, were included with each run. 20-μL RT-PCRs (5 μL of RNA and 15 μL of target master mix) were performed on the ABI StepOnePlus Real-Time PCR system using the following conditions: 48°C for 15 min, 95°C for 10 min, and followed by 45 cycles of 95°C for 15 sec and 60°C 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. 9180 11) 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) (Table S1). For each sample, a library

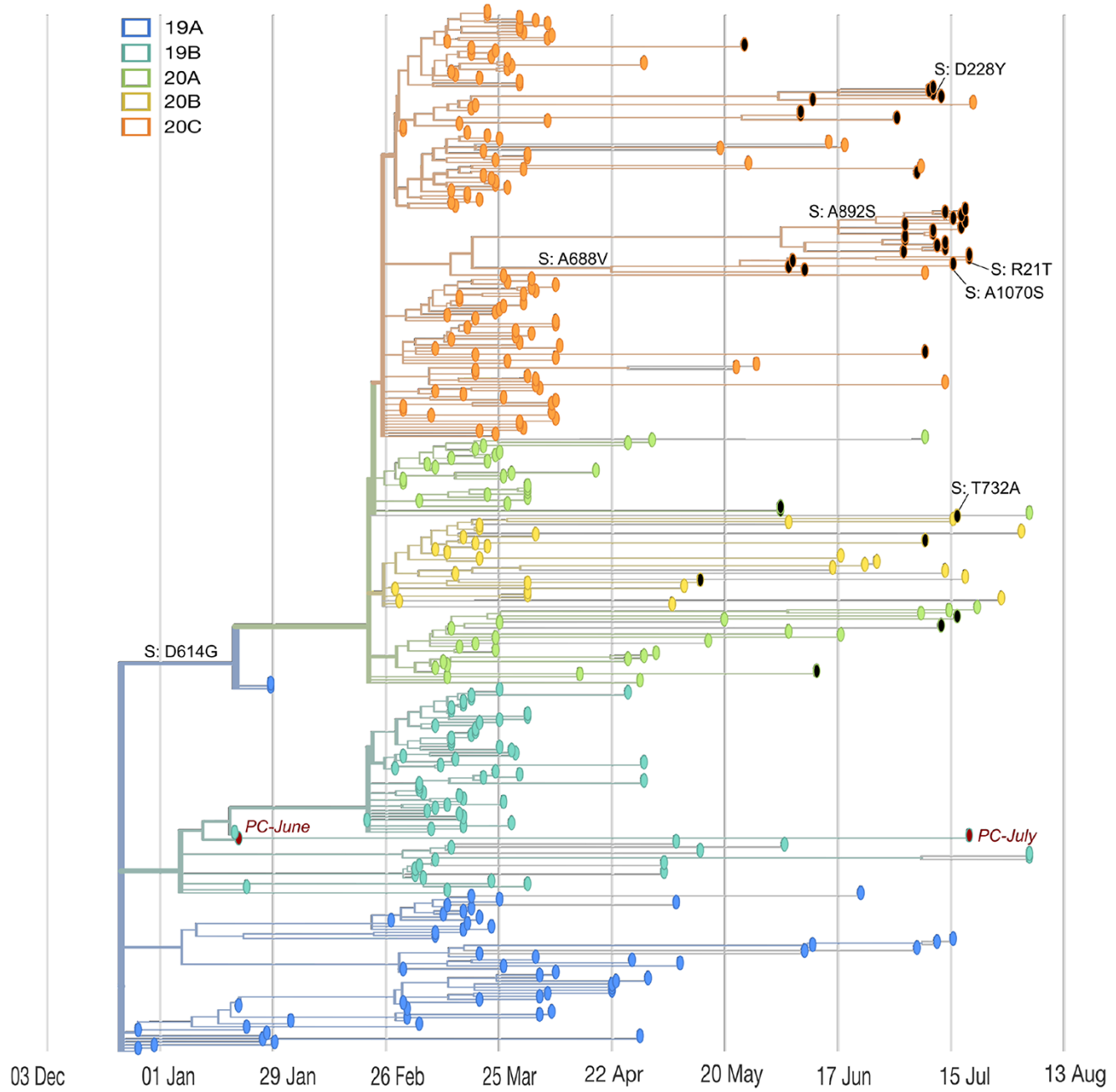

FIGURE 1. Genomic and epidemiological diversity of 59MDW SARS-CoV-2 sequences. The SARS-CoV-2 phylogeny was constructed in a local version of Nextstrain (Auspice.us) and shows the evolutionary relationship of 59MDW SARS-CoV-2 genomes (black dots) and a subsample of publicly available SARSCoV-2 genomes collected worldwide ($n = 301$, colored by viral clade [see legend]). This phylogeny contextualizes mutation rates and transmission patterns of SARS-CoV-2 strains collected at JBSA/Lackland and can be viewed locally using Auspice.us and a provided .json file. The tree is a time-resolved display rooted relative to early samples collected in Wuhan, China, and site numbering and genome structure uses Wuhan-Hu-1/2019 as reference. Positive control samples are shown in red and mutations in the spike protein are indicated in 59MDW samples.

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), and 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 number 15048776 v09), and then sequenced on the NextSeq 500 system at 2×151 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 out with Maxima H Minus dsDNA Synthesis Kit (Thermo Scientific, cat. K2561) from 13 μL of RNA. Synthesized cDNAs were cleaned with Qiagen's QIAquick PCR Purification Kit (Qiagen, cat. number 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 ~7,800 probes to detect respiratory viruses, including SARS-CoV-2.

Briefly, 300-500 μg of cDNA underwent tagmentation, cleanup, 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 2×151 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.

TABLE I. Summary of Nucleotide and Non-synonymous Amino Acid Mutations in 59MDW SARS-CoV-2 Sequences

Gene	Number of nucleotide mutations	Nucleotide mutation frequency	Number of non-synonymous amino acid changes
ORF1a	44	0.404	23
ORF1b	21	0.193	13
S	14	0.128	7
ORF3a	11	0.101	9
E	1	0.009	1
M	0	0.000	0
ORF6	1	0.009	0
ORF7a	2	0.018	1
ORF7b	5	0.046	4
N	9	0.083	4
ORF10	1	0.009	1

109 nucleotide mutations were identified in coding regions of SARS-CoV-2 genomes sequenced from 59MDW samples. The mutation frequency per gene was calculated by dividing the number of mutations per gene by number of genome-wide mutations. Column 4 lists the number of non-synonymous amino acid changes per gene.

Shotgun Metagenomics Sequencing

The 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 2×76 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 Nextstrain conda environment, Auger (bioinformatics tooling) and Auspice (the Nextstrain open-source visualization tool (https://auspice.us), was downloaded and installed locally.[18] 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 May 3, 2020 to August 14, 2020) were downloaded from GISAID and added to the pathogen build. Next, the consensus genome sequences (FASTA) for CIRS clinical specimens ($n = 39$, with two duplicate samples) and FASTAs for the positive control samples ($n = 2$) were added. Metadata for the CIRS

specimens include the following: Sample ID (PIN), date collected, location (North America, TX, 59MDW), and date uploaded. 59MDW viral genomes (FASTA sequences) were uploaded to GISAID.

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 was reviewed and the PDB structure 6VYB was downloaded.[19] 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 per Wrapp et al.[19,20]

RESULTS

In June 2020, we received 15 NPS specimens from the WHASC Clinical Laboratory under an IRB exempt protocol. Fourteen of the specimens tested positive for SARS-CoV-2 at the WHASC laboratory, while one was negative. Upon arrival at the CIRS laboratory, 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 NGS 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 (Fig. S1). We then used fragment analysis to calculate a library QRS, defined by the concentration (ng/μL) of fragments of expected library amplicon

TABLE II. Non-synonymous Mutations in SARS-CoV-2 Structural Proteins

Gene	Nucleotide mutation	Amino acid mutation	Biochemical property	59MDW sample(s)	Frequency
S	21,642 G > C	R21T	Positively charged to polar but neutral	3829	2.70%
S	22,245 G > T	D228Y	Negatively charged to polar but neutral	3680	2.70%
S	23,403 A > G	D614G	Negatively charged to nonpolar	All	100.00%
S	23,625 C > T	A688V	Both nonpolar	3829, 3831, 3784, 2940, 2913, 3014	16.22%
S	2,375 A > G	T732A	Polar but neutral to nonpolar	3,786	2.70%
S	24,236 G > T	A892S	Nonpolar to polar but neutral	3807, 3723, 3804, 3776, 3808	13.51%
S	24,770 G > T	A1070S	Nonpolar to polar but neutral	3784	2.70%
E	26,455 C > T	P71S	Nonpolar to polar but neutral	2973, 3488, 2956	8.11%
N	28,362 C > G	G30A	Both nonpolar	3654, 3681, 3096, 3709, 3680	13.51%
N	28,845 C > T	R191L	Positively charged to nonpolar	3654, 3681, 3096, 3709, 3680	13.51%
N	28,775 C > T	P168S	Nonpolar to polar but neutral	3488	2.70%
N	28,821 C > A	S183Y	Both polar but neutral	3014, 2913, 3784, 2940, 3829, 3831, 3717, 3535, 3687, 3724, 3540, 3537, 3802, 3542, 3667, 3808, 3776, 3804, 3273, 3807	54.05%
N	28,881 G > A	R203K	Both positively charged	2576, 3786, 3634	8.11%
N	28,882 G > A	G204R	Nonpolar to positively charged	2576, 3786, 3634	8.11%

SARS-CoV-2 whole-genome sequencing revealed 14 non-synonymous mutations in genes encoding structural proteins (S = spike; E = envelope; N = nucleocapsid) in 59MDW specimens. The biochemical properties of each non-synonymous change was characterized using the NCBI structure program, and the mutation frequency was calculated by dividing the number of specimens with a mutation by the total number of sequenced clinical specimens ($N = 37$).

size (250-350 bp) normalized to the concentration of fragments representing primer dimers and nonspecific products (150-190 bp) (Table S1). 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 2×151 bp long reads on a NextSeq Mid Output flow cell. Once the CleanPlex primer sequences were removed from the reads, only libraries with good and excellent QRS scores were mapped to the SARS-CoV-2 reference genome

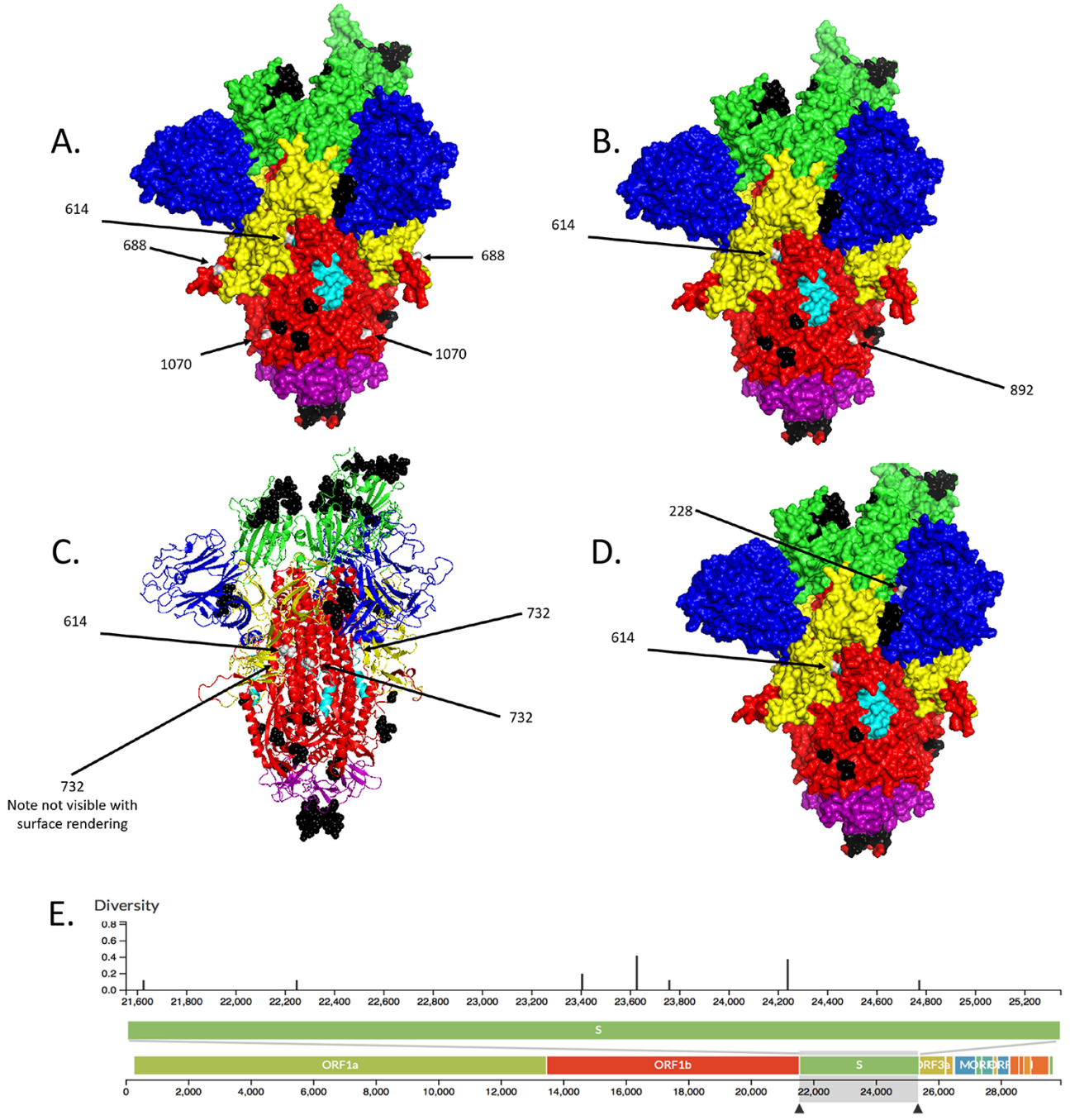

FIGURE 2. 59MDW amino acid mutations overlaid on spike structure in prefusion conformation. (A-D) Spike protein amino acid mutations were mapped onto the prefusion SARS-CoV-2 spike glycoprotein structure (PDB: 6VSB; Wrapp et al.[19]) using PyMOL. Predicted B-cell epitopes (Grifoni et al.[20]) are shown in black, and 59MDW amino acid changes are shown in gray. Spike protein domains: Yellow = S1 (Note, blue = N-terminal domain [NTD] and green = receptor-binding domain [RBD] are also part of S1); cyan = fusion peptide (FP); red = S2 domain; purple = connector domain; (CD). (E) Diversity of the S protein of SARS-CoV-2 sequenced from 59MDW specimens.

(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 (Fig. 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% 20× genome coverage for shotgun and probe, respectively). Another "poor" sample, 2699, sequenced at 20× 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 3× 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 are 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. Before 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 GISAID (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 2020 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 (Fig. 1).

We observed a single nucleotide change in both positive control specimens (PC-June and PC-July) when compared to the reference sequence for USA-WA1/2020 (Fig. 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. number VR-1986D) and may have undergone multiple passages through cell culture before 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 (Fig. S2). Additionally, the analysis revealed that several individuals carried the same viral strains (Fig. 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 in IGV (Fig. S2). In this single cluster of 15 specimens, three individuals shared the same viral sequence. Furthermore, we found two other sets of individuals who had the same viral sequence, 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′-untranslated region (5′UTR), and two mutations in the 3′UTR (Table I).

Next, we focused our analysis on nucleotide mutations that cause non-synonymous amino acid changes in the viral structural proteins (Table II). Seven non-synonymous changes in the spike protein, one in the envelope, and six 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 in the novel furin cleavage site of the SARS-CoV-2 spike glycoprotein.[21] Five viruses contained a mutation in amino acid 892 from Alanine (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 gray) on the prefusion CryoEM rendering of the trimeric spike glycoprotein, with predicted linear B cell epitopes shown in black[20] (Fig. 2). Interestingly, the virus isolated from sample 3784 carried three spike protein mutations (D614G, A688V, and A1070S) (Fig. 2A), higher than the mutations 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 (Fig. 2B). The remaining non-synonymous spike mutations were of low frequency (found in only a single virus). Nevertheless, plotting these changes on the spike protein model reveals that mutation at amino acid 732 is embedded within the S2 domain of the protein and may thus be inaccessible to the host immune

response (Fig. 2C). Amino acid 228 lies within the N-terminal domain, but in close proximity to the receptor-binding domain (Fig. 2D).

CONCLUSIONS

The present study describes a methodology for molecular epidemiological surveillance to track viral evolution and outbreaks. The workflow described in this article—viral RNA isolation, viral load quantification, tiling-based NGS, sequencing and 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.

One limitation of the present study is that descriptive data on patients (i.e., age, sex, symptoms, disease severity, and patient outcomes) was inaccessible to the authors. However, examining how novel SARS-CoV-2 mutations impact on patient disease outcomes is key to enhancing our understanding of COVID-19.[22] Inclusion of patient clinical information in future studies will strengthen the sequencing and bioinformatics workflow presented here by enabling clinicians to link disease symptomology to genomic data.

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 turnaround time could be useful for DoD command officials seeking real-time information on the overall health and readiness of military service members, dependents, civilians, and contractors.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the contributors, originating and submitting laboratories, of the GISAID's EpiFlu Database SARS-CoV-2 sequences 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.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Military Medicine* online.

FUNDING

This work was funded by the 59th Medical Wing Clinical Investigations Program.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

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): 727-33.
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. Lancet 2020; 395(10223): 514-23.
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.
4. World Health Organization: WHO coronavirus disease (COVID-19) dashboard. Geneva. Available at https://covid19.who.int/, 2020; accessed October 12, 2020.
5. Su S, Wong G, Shi W, et al: Epidemiology, genetic recombination, and pathogenesis of coronaviruses. Trends Microbiol 2016; 24(6): 490-502.
6. Cui J, Li F, Shi Z-L: Origin and evolution of pathogenic coronaviruses. Nat Rev Microbiol 2019; 17(3): 181-92.
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): 1967-76.
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): 1814-20.
9. Rambaut A: Phylodynamic analysis of SARS-CoV-2 genomes. Virological. Available at https://virological.org/t/phylodynamic-analysis-176-genomes-6-mar-2020/356, March 6, 2020; 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): 270-3.
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. J Virol 2020; 94(7): e00127-20.
12. Koyama T, Platt D, Parida L: Variant analysis of SARS-CoV-2 genomes. Bull World Health Organ 2020; 98: 495-504. http://dx.doi.org/10.2471/BLT.20.253591.
13. Mercatelli D, Giorgi FM: Geographic and genomic distribution of SARS-CoV-2 mutations. Front Microbiol 2020; 11: 1800. Published July 22, 2020.
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): 812-27.e19.
15. Becerra-Flores M, Cardozo T: SARS-CoV-2 viral spike G614 mutation exhibits higher case fatality rate. Int J Clin Pract 2020; 74(8): e13525.
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. Sci Rep 2020; 10(1): 14031.
17. US Department of Defense: DOD COVID-19 cumulative totals. Available at https://www.defense.gov/Explore/Spotlight/Coronavirus/; accessed October 12, 2020.
18. Hadfield J, Megill C, Bell SM, et al: Nextstrain: real-time tracking of pathogen evolution. Bioinformatics 2018; 34(23): 4121-3.
19. Wrapp D, Wang N, Corbett KS, et al: Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. Science 2020; 367(6483): 1260-3.
20. 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): 671-80.e2.
21. 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. Published July 17, 2020.
22. Voss JD, Skarzynski M, McAuley EM, et al: Variants in SARS-CoV-2 associated with mild or severe outcome. medRxiv 2020. https://doi.org/10.1101/2020.12.01.20242149.