Emerging SARS-CoV-2 Mutations Circulating in the United States

Lauren Brinkac[1*], Sheila Diepold[2*], Shane Mitchell[1], Lee F. Kolakowski[2], William M. Nelson[2], Katharine Jennings[1]

[1] Noblis, Reston, VA; [2] Tetracore, Rockville, MD; *contributed equally

ABSTRACT

Coronavirus disease 2019 (COVID-19), the infectious disease caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), is a global pandemic of unprecedented scale. Despite significant increases in COVID-19 positive cases in the United States (US), the extent of genomic diversity of SARS-CoV-2 in circulation has yet to be fully elucidated. Utilizing Oxford Nanopore Technologies (ONT) complete whole genome sequences (WGS) of SARS-CoV-2 clinical isolates obtained from confirmed COVID-19 cases throughout the US, were generated and analyzed. Twenty-one isolates belong to the potential emerging 20G/8083A clade and co-harbor distinct mutations such as an additional spike protein mutation, Q913H, located within the helical fusion core of the heptad repeat 1 (HR1) region, underscoring the critical need for enhanced genomic surveillance to effectively monitor SARS-CoV-2 evolution.

INTRODUCTION

SARS-CoV-2 emerged in Wuhan, China in late 2019. On March 11, 2020 the World Health Organization (WHO) declared the outbreak a pandemic[1]. The emergence is potentially the result of a zoonotic transfer from an unknown animal species sold in a wet market in Wuhan[2]. Since the initial whole genome sequence, known as Wuhan-Hu-1, was made public[3] more than 441,773 genomes have been sequenced and deposited in GISAID (https://www.gisaid.org/) as of January 29, 2021. Genomic epidemiology has provided detailed tracking of the rise and spread of SARS-CoV-2 variants around the world. Given the fidelity of RNA viruses, variants are expected.

The variants of concern (VOC) reside in the spike (S) protein, which mediates entry of the virus into host cells and is a key target for antibody recognition[4]. These VOC and their biological implications include increased transmissibility (D614G and N501Y)[5,6], altered cell entry (P681H and Δ69/Δ70)[7], and reduced antibody neutralization (E484K)[8,9]. In this study, we analyzed the WGS of 21 SARS-CoV-2 isolates obtained from COVID-19 positive clinical samples and identified the circulation of isolates in two Midwestern US states belonging to the 20G/8083A clade[10,11]. While none of the isolates contained these VOC, they do contain potential clinically important mutations.

METHODS

Clinical nasopharyngeal and nasal samples confirmed to be polymerase chain reaction (PCR) positive for COVID-19 were de-identified and obtained from a Clinical Laboratory Improvement Amendments (CLIA) certified laboratory. The viral RNA was purified from 140 µL of the clinical sample using the QIAamp Viral RNA Mini Kit (Qiagen) following the manufacturer protocol. The cDNA was generated using random primer mix (New England BioLabs, NEB) and Superscript IV First Strand Synthesis kit (Life Technologies). Two multiplex PCR reactions, containing a total of 17 primer pairs, were used to amplify across the SARS-CoV-2 genome[12,13]. Each primer pair produces an amplicon approximately 1,900 base pairs (bp) in size with an average of 175 bp overlaps between the amplicons. The PCR was performed using the Q5 High-Fidelity DNA Polymerase (NEB). The amplicons from both primer pools were combined. The pooled amplicon was purified using Agencourt AMPure XP beads (Beckman Coulter) and quantified using the Qubit 4.0 Fluorometer and the Qubit double-stranded High Sensitivity kit (Life Technologies). The purified PCR amplicon was diluted to 4.8 ng/µL for library preparation.

A total of 60 ng of PCR amplicon was treated with Ultra II End Prep Enzyme mix (NEB). After end repair, the Native Barcodes 1-24 (ONT) were ligated using the Ultra II Ligation Module (NEB). Library preparation for sequencing on the ONT MinION used the Ligation Sequencing 1D kit, SQK-LSK-109 (ONT). Up to 24 samples, including a positive control, were pooled for sequencing on the same flow cell. Sequencing was performed on the MinION using the R9.4.1 flow cell for 4 to 12 hours, depending on the number of samples.

Raw nanopore signal was processed using ONT's Guppy basecaller in high accuracy mode using a single Nvidia Tesla V100 GPU. The basecalled reads were demultiplexed using ONT's Guppy barcoder to bioinformatically separate the reads into their appropriate samples. Unclassified reads were discarded. Reads were filtered to a minimum size of 1,500 base pairs (bp) and a maximum size of 3,500 bp using artic guppyplex. This process was executed according to the arctic protocol (https://artic.network/ncov-2019/ncov2019-bioinformatics-sop.html). Reads were aligned to the SARS-CoV-2 Wuhan-Hu-1 reference genome (GenBank sequence MN908947.3) using MinMap2 via ONT's medaka consensus pipeline. Variants were annotated using SnpEff v5.0[13]. Pangolin COVID-19 Lineage Assigner was used to assign SARS-CoV-2 phylogenetic lineages[14]. Complete SARS-CoV-2 genomes obtained in the US were downloaded from GISIAD (https://www.gisaid.org/) on January 31, 2021 (n=89,731). Nextstrain's NextClade software using default filtering and subsampling settings was used to perform phylogenetic analysis of US 20G clade isolates[16,17].

AKNOWLEDGEMENT

This project was funded by Noblis and Tetracore. Special thank you to Tetracore for access to COVID-19 samples.

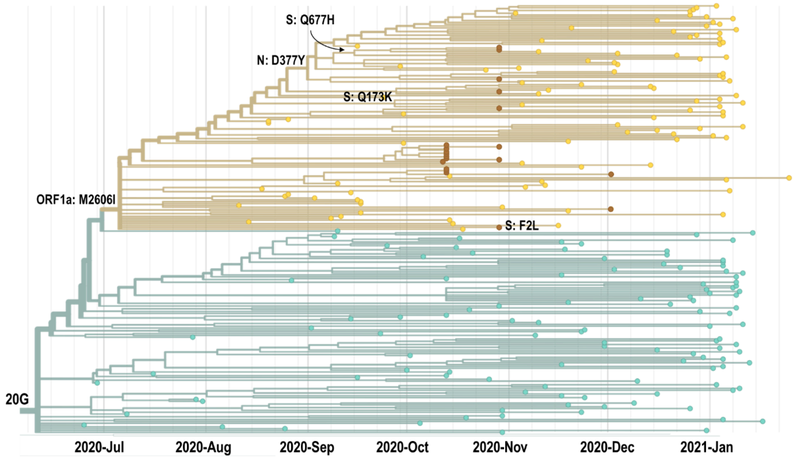

Figure 1. Phylogenetic relationship of study isolates with a US subsampling of the 20G clade. Study isolates are denoted with brown tips, grouping with isolates sharing the ORF1a:M2606I mutation (branch labelled, and nodes highlighted in yellow). Other distinguishing branch mutations are labeled accordingly.

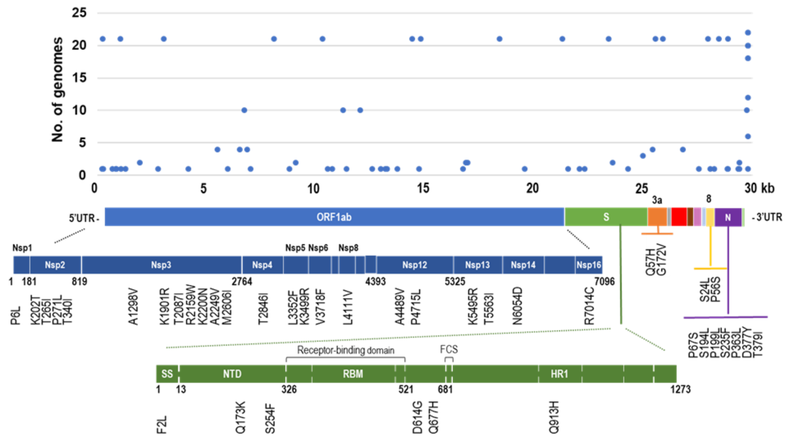

Figure 2. Frequency (top) and genome location (bottom) of mutations observed in study isolates.

Table 1. Distinguishing mutations observed in isolates sequenced in this study.

Location	Mutation	No. of isolates	Date (2020)
Iowa	ORF1ab: T5563I; **S: Q677H**, **Q913H**; N: D377Y	1	29-Oct
	ORF1ab: T5563I; **S: Q677H**; ORF8: 28254del; N: D377Y	1	
	N: T379I	1	
Nebraska	ORF1ab: A2249V; **S: Q173K**; N: P364L	1	
	ORF1ab: T340I, A1298V, K1901R, T2846I; **S: F2L**; ORF8: P56S	1	
	ORF1ab: K202T, K5495R; **S: S254F**	1	
	ORF1ab: R2159W	10	12-Oct
	ORF1ab: T2087I, K2200N	3	13-Oct
	ORF1ab: P271L, T2087I, K2200N	1	2-Dec
	ORF1ab: P6L, L3495F, V3718F, L4111F, A4489V; N: S194L, F363I	1	

RESULTS

Twenty-one SARS-CoV-2 isolates obtained from COVID-19 positive nasopharyngeal and nasal swab samples were collected in Iowa (n=3) and Nebraska (n=18) between October 12, 2020 and December 2, 2020, and sequenced. All genomes covered > 99.5% of the Wuhan-Hu-1 reference genome with a mean coverage of 3888x. Genomes were represented by the globally abundant B.1 lineage, B.1.2 sublineage, and all 21 isolates grouped within the potential emerging 20G/8083A clade, corresponding to the ORF1a:M2606I mutation, noted for its increasing US prevalence[10,11] (**Figure 1**).

Sixty-three distinct mutations were identified in coding regions of the 21 isolates. Of these over half (n=40, 65%) are nonsynonymous occurring in ORF1ab (11 out of the 16 nonstructural proteins (NSP) integral for viral replication and transcription), S and nucleoside (N) structural proteins, and ORF3a and ORF8 accessory proteins (**Figure 2**). The most prevalent mutations genome-wide include ORF1ab:C3037T, nsp12:P323L and S:D614G characteristic of B.1 lineage SARS-CoV-2 isolates, as well as conserved mutations consisting of nsp2:T85I, nsp3:M1788I, nsp5:L89F, nsp14:N129D, nsp16:R216C, ORF3a:Q57H, ORF3a:G172V, ORF8:S24L, N:P67S, and N:P199L, and ORF1ab:C14805T as similarly reported in 20G/8083A Midwest isolates[10,11].

Of the mutations involving the S protein, two isolates obtained from samples collected in Iowa, also contain the recently acquired S:Q677H mutation as reported in other 20G/8083A isolates[10,11] (**Table 1**). This mutation (red) is of interest due to its position in the QTQ**T**N consensus sequence where the bolded threonine is an ACE2 receptor interacting residue located three amino acids upstream of the furin cleavage site (FCS)[18]. Novel to this clade and unique among SARS-CoV-2 genomes is the co-occurrence in a single isolate of an additional S protein mutation, Q913H. It is also interesting that both glutamine mutations were substituted with histidine, as noted elsewhere[2] and in B.1.1.7/501Y.V1 with P681H[10]. Located within the helical fusion core of the HR1 region, the effect of Q913H, if any, on HR1-HR2 interactions and stability of the six-helical bundle (6-HB) essential for viral fusion is not known[19]. The second S:Q677H containing isolate co-harbors a ORF8:28254del involving the loss of an adenine, one of several globally occurring deletions in this gene[20,21]. Furthermore, co-occurring in both S:Q677H isolates are nsp13:T239I and N:D377Y. Additional S protein mutations of interest were identified in two isolates collected in Nebraska. These isolates contain either a Q173K or S254F substitution located in the N-terminal domain (NTD) of the S protein and are of interest as the NTD has been shown to act as a non-RBD antibody-targeting domain[22,23]. The substitution of a charged lysine at position 173, in place of the neutral glutamine, may alter antibody interaction[10]. Last, a single isolate also obtained in Nebraska, contains a F2L mutation located in the signal signature (SS) domain of the S protein. While leucine is less sterically hindered than phenylalanine, the effect of this conservative replacement is unknown. Furthermore, there were 16 additional isolates sequenced in this study lacking any mutations in the spike protein.

CONCLUSION

Through whole genome sequencing this study identified 21 SARS-CoV-2 isolates sharing similarly conserved mutations with other recently reported US Midwest isolates of the 20G/8083A clade, while also containing some unique differences. Most notably, the S protein mutations consisting of F2L located in the SS domain, Q173K and D377Y in the NTD, and Q913H adjacent to the FCS. The functional significance of these and other mutations identified as part of this study need to be confirmed experimentally. Furthermore, increases in regional and national genomic sequencing will help elucidate the extent to which these mutations are circulating in the population. These findings underscore the critical need for enhanced frontline and national genomic surveillance to effectively monitor SARS-CoV-2 evolution. As certain variants demonstrate increased transmissibility and aspects of immune evasion, there is an urgent need to monitor the potential for impacting therapeutic and vaccine efficacy.

REFERENCES

1. WHO Director-General's opening remarks at the media briefing on COVID-19 – 11 March 2020. WHO. 11 March 2020.
2. Garry RF. Mutations arising in SARS-CoV-2 spike on sustained human-to-human transmission and human-to-animal passage, Virological. 2 January 2021.
3. Wu F, et al. A new coronavirus associated with human respiratory disease in China. Nature. 2020 Mar;579(7798):265-269.
4. Galloway SE, et al. Emergence of SARS-CoV-2 B.1.1.7 lineage - United States, December 29, 2020–January 12, 2021. CDC. 15 January 2021.
5. Korber B, et al. Tracking changes in SARS-CoV-2 spike: evidence that D614G increases infectivity of the COVID-19 virus. Cell. 2020 Aug 20;182(4):812-827.e19.
6. Leung K, et al. Early transmissibility assessment of the N501Y mutant strains of SARS-CoV-2 in the United Kingdom, October to November 2020. Euro Surveill. 2021 Jan;26(1):2002106.
7. Hoffmann M, et al. A multibasic cleavage site in the spike protein of SARS-CoV-2 is essential for infection of human lung vells. Mol Cell. 2020 May 21;78(4):779-784.e5.
8. Voloch C, et al. Genomic characterization of a novel SARS-CoV-2 lineage from Rio de Janeiro, Brazil. medRxiv 2020.12.23.20248598
9. Faria NR, et. al. Genomic characterisation of an emergent SARS-CoV-2 lineage in Manaus: preliminary findings. Virological. 12 January 2021.
10. Pater A, et al. Emergence and evolution of a prevalent new SARS-CoV-2 variant in the United States. bioRxiv 2021.01.11.426287.
11. Tu H, et al. Distinct patterns of emergence of SARS-CoV-2 spike variants including N501Y in clinical samples in Columbus Ohio. bioRxiv 2021.01.12.426407.
12. Resende PC, et al. SARS-CoV-2 genomes recovered by long amplicon tiling multiplex approach using nanopore sequencing and applicable to other sequencing platforms. bioRxiv 2020.04.30.069039.
13. Resende, PC, et al. Genomic surveillance of SARS-CoV-2 reveals community transmission of a major lineage during the early pandemic phase in Brazil. bioRxiv 2020.06.17.158006.
14. Cingolani P, et al. 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 Apr-Jun;6(2):80-92.
15. Rambaut A, et al. Preliminary genomic characterisation of an emergent SARS-CoV-2 lineage in the UK defined by a novel set of spike mutations. Virological. 8 December 2020.
16. Hadfield J, et al. Nextstrain: real-time tracking of pathogen evolution. Bioinformatics. 2018 Dec 1;34(23):4121-4123.
17. Sagulenko P, et al. TreeTime: Maximum-likelihood phylodynamic analysis. Virus Evol. 2018 Jan 8;4(1):vex042.
18. Andersen KG, et al. The proximal origin of SARS-CoV-2. Nat Med. 2020 Apr;26(4):450-452.
19. Huang Y, et al. Structural and functional properties of SARS-CoV-2 spike protein: potential antivirus drug development for COVID-19. Acta Pharmacol Sin. 2020 Sep;41(9):1141-1149.
20. Pereira F. Evolutionary dynamics of the SARS-CoV-2 ORF8 accessory gene. Infect Genet Evol. 2020 Nov;85:104525.
21. Gaurav S, et al. Identification of unique mutations in SARS-CoV-2 strains isolated from India suggests its attenuated pathotype. bioRxiv 2020.06.06.137604;
22. Suryadevara N, et al. Neutralizing and protective human monoclonal antibodies recognizing the N-terminal domain of the SARS-CoV-2 spike protein. bioRxiv 2021.01.19.427324.
23. Chi X, et al. A neutralizing human antibody binds to the N-terminal domain of the Spike protein of SARS-CoV-2. Science. 2020 Aug 7;369(6504):650-655.