frontiers in Genetics

BRIEF RESEARCH REPORT
published: 03 February 2021
doi: 10.3389/fgene.2020.612571

SARS-CoV-2 Genomes From Oklahoma, United States

*Sai Narayanan[1,2], John C. Ritchey[1], Girish Patil[1], Teluguakula Narasaraju[1], Sunil More[2], Jerry Malayer[3], Jeremiah Saliki[1], Anil Kaul[4], Pratul K. Agarwal[3,5] and Akhilesh Ramachandran[1,2]**

[1] Oklahoma Animal Disease Diagnostic Laboratory, College of Veterinary Medicine, Oklahoma State University, Stillwater, OK, United States, [2] Department of Veterinary Pathobiology, College of Veterinary Medicine, Oklahoma State University, Stillwater, OK, United States, [3] Department of Physiological Sciences, College of Veterinary Medicine, Oklahoma State University, Stillwater, OK, United States, [4] Center for Health Sciences, Oklahoma State University, Tulsa, OK, United States, [5] High-Performance Computing Center, Oklahoma State University, Stillwater, OK, United States

Genomic sequencing has played a major role in understanding the pathogenicity of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). With the current pandemic, it is essential that SARS-CoV-2 viruses are sequenced regularly to determine mutations and genomic modifications in different geographical locations. In this study, we sequenced SARS-CoV-2 from five clinical samples obtained in Oklahoma, United States during different time points of pandemic presence in the state. One sample from the initial days of the pandemic in the state and four during the peak in Oklahoma were sequenced. Previously reported mutations including D614G in S gene, P4715L in ORF1ab, S194L, R203K, and G204R in N gene were identified in the genomes sequenced in this study. Possible novel mutations were also detected in the S gene (G1167V), ORF1ab (A6269S and P3371S), ORF7b (T28I), and ORF8 (G96R). Phylogenetic analysis of the genomes showed similarity to other SARS-CoV-2 viruses reported from across the globe. Structural characterization indicates that the mutations in S gene possibly influences conformational flexibility and motion of the spike protein, and the mutations in N gene are associated with disordered linker region within the nucleocapsid protein.

Keywords: SARS-CoV-2, nCoV-19, sequencing, mutations, Oklahoma

OPEN ACCESS

Edited by:
Cordelia Manickam, Beth Israel Deaconess Medical Center and Harvard Medical School, United States

Reviewed by:
Raja Mohan Gopalakrishnan, Beth Israel Deaconess Medical Center and Harvard Medical School, United States
Daniel Ramos Ram, Beth Israel Deaconess Medical Center and Harvard Medical School, United States

***Correspondence:**
Akhilesh Ramachandran
rakhile@okstate.edu

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: *30 October 2020*
Accepted: *31 December 2020*
Published: *03 February 2021*

Citation:
Narayanan S, Ritchey JC, Patil G, Narasaraju T, More S, Malayer J, Saliki J, Kaul A and Ramachandran A (2021) SARS-CoV-2 Genomes From Oklahoma, United States. Front. Genet. 11:612571. doi: 10.3389/fgene.2020.612571

INTRODUCTION

Toward the end of 2019, several individuals with signs of pneumonia reported to hospitals in Wuhan, the capital of Hubei Province in Central China. The etiological agent was identified to be a novel coronavirus (SARS-CoV-2/nCoV-19) on 7th January 2020 (Zheng, 2020). Human to human transmission was recorded around the same time (Nishiura et al., 2020). By the end of January 2020, WHO declared a "public health emergency of international concern" (Statement on the second meeting of the International Health Regulations 2005, 2020). As of December 22nd, 2020, a total of 78,263,502 confirmed patients and 1,722,307 deaths have been reported worldwide (Dong et al., 2020) and 2,65,620 registered cases and over 2,240 deaths in the state of Oklahoma, United States (OSDH-Covid-19 Tracker, 2020).

Numerous coronaviruses infecting different animal species including humans have been identified. SARS-CoV-2 is an enveloped positive-sense single-stranded RNA virus belonging to the

genus *Betacoronavirus* (subgenus *Sarbecovirus*) in the Coronaviridae family (order: Nidovirales). Based on genomic sequence analysis, it is reported to have originated from bats (Hu et al., 2015) and pangolins (Zhang et al., 2020a,b). Other previously identified coronaviruses known to infect humans include SARS-CoV-1, MERS-CoV, HCoV-NL63, HCoV-229E, HCoV-OC43, and HCoV-HKU1 (Graat et al., 2003; Van Elden et al., 2004; Mackay et al., 2012; Annan et al., 2013; Owusu et al., 2014; Corman et al., 2018).

Genomic sequence data is important in identifying, characterizing and understanding pathogens (Rambaut et al., 2008; Salipante et al., 2015). It can shed light on pathogenicity, virulence, drug/vaccine targets, mutation sites etc. and can also be critical in source attribution and determining microbial provenance (van Dorp et al., 2020). The first genome sequence of SARS-CoV-2 was made available on January 10th (GenBank ID: MN908947.3) (Wu et al., 2020). Multiple genomic sequences of SARS-CoV-2 from all over the world have since been deposited in public databases such as GenBank and GISAID (Global Initiative on Sharing All Influenza Data; [1]) (Elbe and Buckland-Merrett, 2017). This has facilitated extensive genomic studies leading to the identification of several mutations in the genome that can influence infectivity and virulence of the virus (Banerjee et al., 2020a; Daniloski et al., 2020; van Dorp et al., 2020).

The genome of SARS-CoV-2 is similar to many other pathogenic coronaviruses and has multiple genes that code for different proteins such as S gene (Surface glycoprotein), N gene (nucleocapsid phosphoproteins), M gene (membrane glycoprotein), E gene (envelope protein), and open reading frames, such as ORF1a, ORF1b ORF3a, ORF6, ORF7a, ORF7b, ORF8, and ORF10 (Khailany et al., 2020; van Dorp et al., 2020). The leading sequence of the viral genome is the sequence for ORF1ab, which encodes for multiple proteins including replicase polyproteins, non-structural proteins, papain-like proteinase, RNA-dependent RNA polymerase etc., which are essential for replication and survival in the host (Khailany et al., 2020). Though the exact function of ORF3a is yet to be clearly understood, it is believed to play a major role in viral release after replication in SARS-CoV-1 (Lu et al., 2009). ORF7 and ORF8 code for accessory proteins, the functions of which are yet to be clearly understood (Alam et al., 2020; Zhang et al., 2020c). Minimal roles of ORF7 and 8 in viral replication have been reported in SARS-CoV-1 alongside apoptosis stimulation of host cells (Liu et al., 2014).

The huge repository of sequence data in open-source databases such as the GenBank-NCBI and GISAID has facilitated the identification of numerous mutations and single nucleotide polymorphisms (SNPs) in the SARS-CoV-2 genome. SNPs in the genome that result in commonly reported non-synonymous mutations such as P4715L in ORF1ab, D614G in S gene, R203K, and G204R in N gene are some of the commonly reported. P4715L in ORF1ab is believed to play a major role in interaction with other proteins that regulate RNA dependent RNA polymerase (Pachetti et al., 2020). D614G (Aspartate to Glycine) mutation in the S gene has been reported to result in increased transduction into human epithelial cells (Daniloski et al., 2020). N gene mutations R203K and G204R are believed to increase viral fitness, survival and adaptation to humans (Leary et al., 2020).

In this study, we sequenced the genome of SARS-CoV-2 from 5 human clinical samples received at the Oklahoma Animal Disease Diagnostic Laboratory (OADDL) at various times during the SARS-CoV-2 pandemic in Oklahoma. Multiple mutations were detected in the genomic sequences including those already reported as well as previously unreported mutations.

MATERIALS AND METHODS

This study was approved by the Institutional Review Board (Application number: IRB-20-357) at Oklahoma State University, Stillwater, OK, 74078, United States.

Clinical Samples and Processing

Five Nasopharyngeal swabs collected from human patients received at OADDL for COVID-19 testing between March 2020 – July 2020 were used in this study. Nucleic acid extraction was performed using MagMax Viral Pathogen Nucleic acid isolation kit (Thermofisher, MA, United States) as per the manufacturer's recommended protocols. Viral presence was detected by real-time PCR using TaqPath COVID-19 Multiplex Diagnostic Solutions (Thermofisher, MA, United States). All samples had a cycle threshold value between 18 and 22.

Genomic Sequencing

Complementary DNA (cDNA) was synthesized from extracted RNA from five clinical samples that were positive for SARS-CoV-2. cDNA was then PCR amplified using ARTIC V3 primers[2] to obtain overlapping segments of the whole viral genome. DNA library repair (SQK-LSK-109, Oxford Nanopore Technologies, United Kingdom), Solid Phase Reversible Immobilization (SPRI) paramagnetic beads clean-up (AMPure XP, Beckmann Coulter, CA, United States) and adapter ligation and barcoding (NBD-001, Oxford Nanopore Technologies, United Kingdom) were done as per manufacturer recommendations. Libraries were then pooled and sequenced using MinION (Oxford Nanopore Technologies, United Kingdom) platform following manufacturer recommendations.

Genome Assembly, Alignment, and Phylogenetic Analysis

Sequences obtained were assembled *de novo* using Canu (Koren et al., 2017). To further obtain a reliable consensus genome assembly, *de novo* assemblies and sequence output files were assembled to the SARS-CoV-2 reference genome Wuhan Hu-1 (GenBank ID: MN908947.3) with minimap2 (Li, 2018) and Nanopolish (Loman et al., 2015).

To assess the uniqueness of the genomes sequenced in this study, MAFFT (Katoh et al., 2009; Katoh and Standley, 2013) was

[1] www.epicov.org

[2] https://github.com/artic-network/artic-ncov2019/tree/master/primer_schemes/nCoV-2019/V3

TABLE 1 | Genome similarity of five sequenced genomes when compared to NCBI reference genome (NC_045512.2) along with total reads generated and genome coverage obtained.

ISOLATE ID	GenBank ID	GISAID Accession number	Total number of reads	Genome coverage obtained	Genome similarity to NC_045512.2
Oklahoma-ADDL 1	MT998442	EPI_ISL_535364	264,586	4,414X	99.004
Oklahoma-ADDL 2	MW000350	EPI_ISL_535361	343,075	5,844X	99.653
Oklahoma-ADDL 3	MT998443	EPI_ISL_535362	2,181,723	37,153X	99.653
Oklahoma-ADDL 4	MT998444	EPI_ISL_535363	895,721	11,904X	99.661
Oklahoma-ADDL 5 (partial)	MW000372	EPI_ISL_487231	495,578	8,260X	86.836

used to align whole sequences of the five genomes to SARS-CoV-2 reference genome (NC_045512.2).

Gene predictions on the consensus assemblies were made using Viral Genome ORF Reader four (Wang et al., 2012) (VIGOR4) using a curated library available in the Virus Pathogen Resource (ViPR) (Pickett et al., 2012) database. Individual genes were aligned to SARS-CoV-2 Wuhan Hu-1 genome from NCBI (GenBank ID: MN908947.3) using MUSCLE aligner in MEGA-X (Kumar et al., 2018) to identify SNPs and changes in the amino acid produced by the gene.

To assess similarity to previously reported genomes, a phylogenetic analysis was made using 9072 genomes of SARS-CoV-2 from the GenBank database and the five viruses sequenced in the study. A General Time Reversible (GTR) substitution model based Unweighted Pair Group Method with Arithmetic Mean (UPGMA) alignment was constructed using MAFFT and FastTree (Price et al., 2010). Clade definitions for the sequences were identified using nine marker variants reported for classification in the GISAID database (Elbe and Buckland-Merrett, 2017).

Structural Characterization

Crystal structures with the following codes were downloaded from protein data bank (PDB, [3]): 6VXX, 6VSB, 6M3M, and 2CJR. Computational models for the spike protein assembly were downloaded from CHARMM-GUI Archive – COVID-19 Proteins Library[4]. All measurements and analysis were performed using PyMOL opensource version 1.8.2.0[5].

RESULTS AND DISCUSSION

Five clinical rRT-PCR positive samples from different periods of the pandemic in Oklahoma, United States were chosen for the study. One of the samples (Oklahoma-ADDL-1) was received during the initial stages (April 2020) of the pandemic. An increased incidence rate of the disease was observed by the end of May 2020 (OSDH-Covid-19 Tracker, 2020). Three of the samples (Oklahoma-ADDL-2,3,4) sequenced were received during this period and the last sample (Oklahoma-ADDL-5) was received 1 month (June 2020) after this period. More than 4,000X coverage was obtained at the end of sequencing for all the samples. Following *de novo* assembly with Canu, consensus genomes were obtained after reference genome assembly with nanopolish and minimap2. The genomes sequenced have been submitted to GISAID and GenBank (**Table 1**).

The five viral genome assemblies were aligned to the reference genome (SARS-CoV-2 Wuhan-Hu-1 NC_045512.2) using MAFFT and genome similarities to the reference isolate were calculated (**Table 1**). Most of the genomes showed more than 99% similarity. For Oklahoma-ADDL-5, only a partial sequence could be generated and hence showed a lower 86.836% similarity. This could be due to reduced amplification during amplicon generation and also due to reference assembly. Other than Oklahoma-ADDL-5, Oklahoma-ADDL-1 sequenced from a clinical sample obtained during the beginning of the pandemic in Oklahoma (April 2020) showed a lower similarity to the reference genome when compared to the other four isolates.

Genes were predicted using VIGOR4 and individual genes were aligned to their respective NCBI reference genes using MUSCLE aligner with UPGMA alignment in MEGA-X to assess mutations in the genome. Using MEGA-X visualization tool, various silent and missense mutations were detected. The missense/non-synonymous mutations detected in major genes are listed in **Table 2**. While non-synonymous mutations were detected in ORF1ab, ORF1a, S, N, ORF3a, ORF7, and ORF8, none were detected in Envelope (E) gene, Membrane glycoprotein (M) gene or ORF10 gene.

TABLE 2 | Non-synonymous mutations detected and their respective amino acid changes when compared to NCBI reference genome (SARS-CoV-2, Wuhan Hu-1, NC 045512.2).

Mutation observed (nucleiotide position)	Gene	Genomes mutation detected in	Mutation type
P3371S (C10376T)	ORF1ab	Oklahoma-ADDL-5	Transition
T4412A (A13498G)	ORF1ab	Oklahoma-ADDL-4	Transition
P4715L (C14408T)	ORF1ab	Oklahoma-ADDL-1,2,3,4	Transition
A6269S (G19069T)	ORF1ab	Oklahoma-ADDL-2,3	Transversion
D614G (A23403G)	S	Oklahoma-ADDL-1,2,3,4,5	Transition
G1167V (G25062T)	S	Oklahoma-ADDL-4	Transversion
Q57H (G25563T)	ORF3a	Oklahoma-ADDL-1	Transversion
G96R (G28179C)	ORF8	Oklahoma-ADDL-2,3	Transition
T28I (C27476T)	ORF7b	Oklahoma-ADDL-5	Transversion
S194L (C28854T)	N	Oklahoma-ADDL-4	Transition
R203K (G28881A; G28882A)	N	Oklahoma-ADDL-2,3,5	Transition
G204R (G28883C)	N	Oklahoma-ADDL-2,3,5	Transversion

[3]https://www.rcsb.org/

[4]http://www.charmm-gui.org/?doc=archive&lib=covid19

[5]http://pymol.org/

Non-synonymous mutation at amino acid location 81 (C→T; Arginine → Cysteine), which codes for nsp2 (non-structural protein)was present in all the genomes sequenced in the study. The potential implication of this mutation is unknown. nsp2 along with nsp3 are known to play major roles in pathogenesis (Angeletti et al., 2020) of SARS-CoV-2 in humans. A few other possibly novel and previously reported non-synonymous mutations in ORF1a and ORF1ab were also identified. A previously reported mutation in the ORF1ab gene P4715L (Banerjee et al., 2020b) (Proline to Leucine) was recorded alongside novel mutations at various amino acid locations. P4715L mutation has been implicated to play a major role in interaction with other proteins that regulate RNA Dependent RNA polymerase activity. P3371S (Proline to Serine) mutation in ORF1ab and ORF1a was detected in Oklahoma-ADDL-5. Oklahoma-ADDL-4 carried mutation T4412A (Threonine to Alanine) in ORF1ab, while Oklahoma-ADDL-2 and 3 carried mutation A6269S (Alanine to Serine). ORF1ab has multiple functions including RNA dependent RNA polymerase activity, helicase activity, Fe–S cluster binding, Zn^{-} binding activity, methyltransferase activity (Graham et al., 2008) etc. Functional implications of these mutations are still unknown and further studies to understand functional changes caused by these mutations may aid in better understanding the pathogenesis of the viral isolates found in Oklahoma.

Non-synonymous mutations were also found in S, ORF3a, ORF7b, ORF8, and N gene (**Table 2**). A previously reported mutation in S gene – D614G (Laha et al., 2020) (Aspartate to Glycine), was identified in all the genomes sequenced. The D614G mutation has been reported to cause a decrease in PCR cycle thresholds, suggestive of higher upper respiratory tract viral load (Grubaugh et al., 2020; Korber et al., 2020) in the host. A previously reported deleterious variation in the protein expressed from ORF3a, Q57H (Issa et al., 2020; Laha et al., 2020), was recorded in Oklahoma-ADDL-1, the genome isolated at the beginning of the pandemic in Oklahoma. This mutation was not recorded in genomes isolated at other times in the state. The N gene also carried other previously reported mutations (**Table 2**).

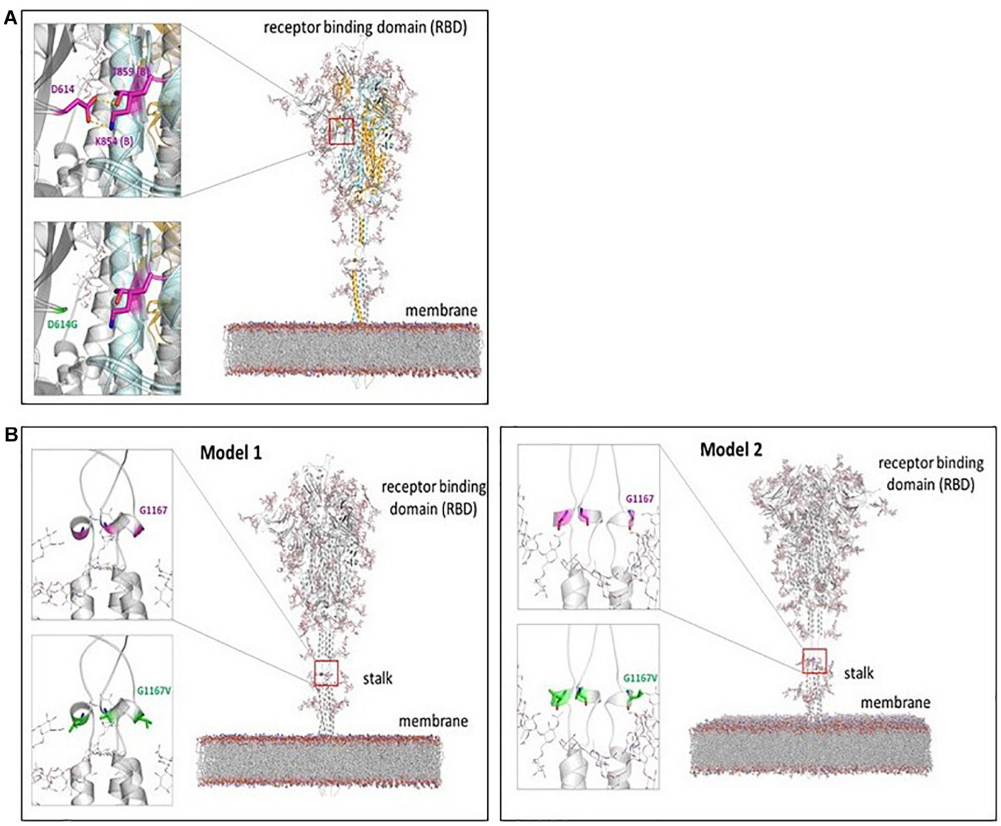

FIGURE 1 | Mutations in the Spike protein. **(A)** Change in interactions within spike protein due to D614G mutation: The mutation is located in the RBD of the spike protein. Model based on groups Croll, Seok, and Im is depicted (Choi et al., 2020; Woo et al., 2020). This model shows glycosylated protein embedded in a membrane; the spike protein is homo-trimer, with three protein chains shown in gray, cyan, and light orange. The location of Asp 614 is indicated by red square in the full model. The zoomed in view on the left shows the interacting residues. The side chain of Asp 614 from one monomer (gray chain) makes hydrophilic interactions with residues Lys 854 and Thr 859 from adjacent monomer (cyan chain indicated by **(B)** in the residue labels). Additional hydrophobic interactions with Val 860 and Leu 861 have also been suggested. The mutation D614G causes loss of all these interactions. **(B)** Location of G1167V mutation in spike protein: Two alternate computational models are shown, as the structure of the full-length spike protein has not been solved so far. These two alternate models indicate that the Gly 1167 is located in the stalk region between the RBD and the transmembrane region. The position is predicted to be a part of short alpha helical region. The mutation G1167V may affect the alpha helical region as well as the bending motion of the stalk region.

Oklahoma-ADDL-4, carried a non-synonymous mutation S194L (Banerjee et al., 2020a) while Oklahoma-ADDL 2,3,5 carried R203K and G204R (Laha et al., 2020).

A few possible novel mutations were also identified in S, ORF3a, ORF7b, and ORF8 genes (**Table 2**). A mutation in the S gene, G1167V (Glycine to Valine) was identified in Oklahoma-ADDL-4. T28I in ORF7b gene, a non-synonymous mutation (Threonine to Isoleucine) in the genome Oklahoma-ADDL-5 and in ORF8, G96R non-synonymous mutations resulting in Glycine to Arginine were noted in Oklahoma-ADDL-2 and Oklahoma-ADDL-3. These mutations in the genome indicate presence of multiple variations of the virus in Oklahoma.

A phylogenetic tree was constructed using 9,072 genomes available in the NCBI repository. The nearest neighbors to the genomes were found to be isolates reported from San Diego and Atlanta in United States and from Greece and Australia. While the Oklahoma-ADDL-1 and 5 were phylogenetically more related to isolates reported from Australia, Oklahoma-ADDL-2 and 3 were phylogenetically related to isolates reported from Greece and Atlanta, United States. Oklahoma-ADDL-4 was phylogenetically related to the isolate reported from San Diego United States. Oklahoma-ADDL-1 was identified to be in clade-GH and Oklahoma-ADDL-2,3,4,5 were identified to be in clade GR as per annotations of different marker variants in the GISAID database. The presence of multiple isolates within the Oklahoma sheds light on contagious nature of these viruses.

Functional Significance of S Gene Mutations

D614G mutation in S gene, which corresponds to changes in the spike protein, has already been reported widely in the literature

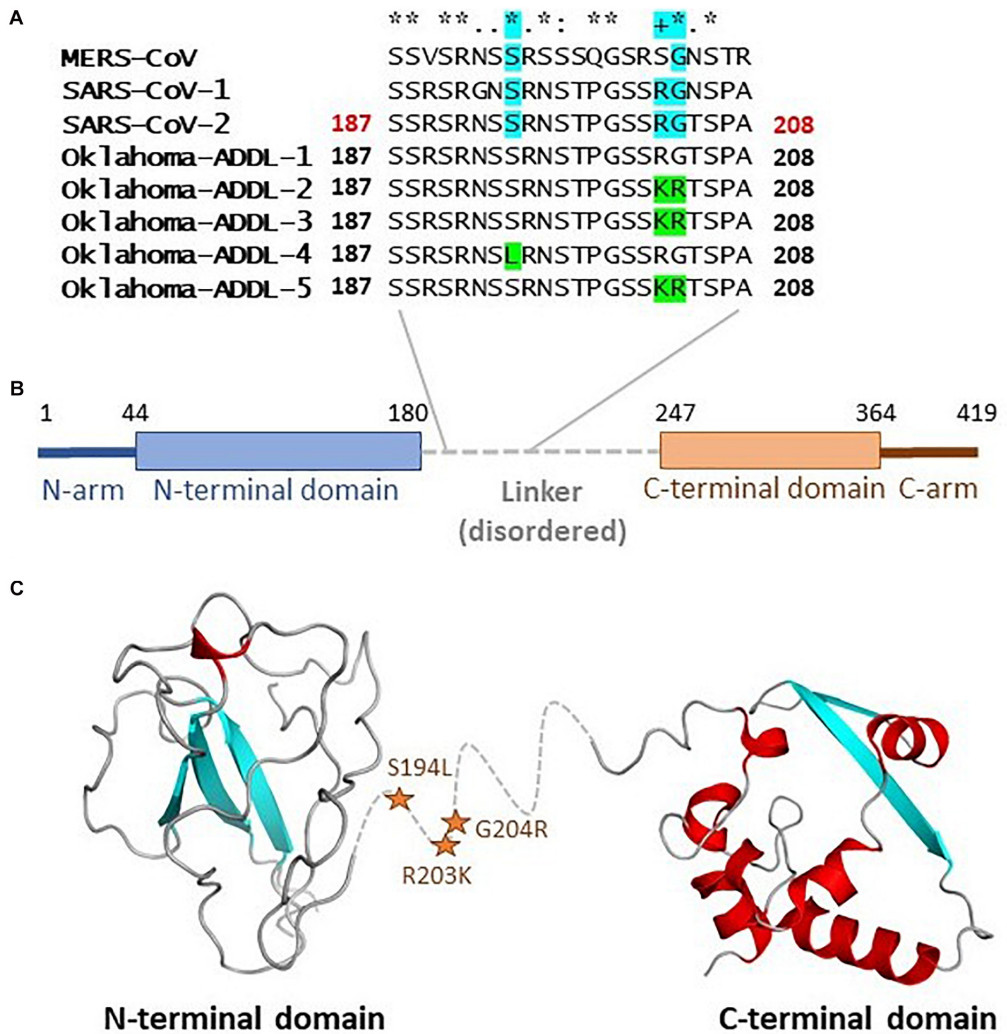

FIGURE 2 | Mutations in the nucleocapsid protein: The N gene product plays a role in RNA packaging and viral release. **(A)** The mutations S194L, R203K, and G204R are located in the highly conserved linker region as indicated in panel. **(B)** Various regions in the protein. **(C)** The structure of full-length protein is not available, structure of NTD (based on 6M3M chain A) and CTD (based on 6WJI chain A) are depicted with linker region indicated with the dashed line in the middle. The location of the mutations is indicated by stars.

(Grubaugh et al., 2020; Korber et al., 2020; Laha et al., 2020). As shown in **Figure 1A**, crystal structure and the molecular models of spike protein assembly developed by several groups indicate that the residue Asp 614 makes contact with several other residues in the trimeric spike protein assembly (Choi et al., 2020; Walls et al., 2020; Woo et al., 2020; Wrapp et al., 2020). Most noticeably, the sidechain of Asp 614 makes hydrophilic contacts with residues Lys854 and Thr859 (of the adjacent monomer of spike protein). Additional hydrophobic interactions with residues Val 860 and Leu 861 have also been reported (Isabel et al., 2020). The mutation to Gly 614 removes all these side chain interactions. The functional relevance of this mutation is currently being explored in a number of ongoing investigations (Choi et al., 2020; Gobeil et al., 2020), particularly in relation to conformational dynamics of the receptor binding domain (RBD).

The mutation G1167V in S gene is proposed to be located in the stalk region of spike protein. A structure of full-length spike protein is currently not available. Models prepared by computational groups (Choi et al., 2020; Woo et al., 2020) (**Figure 1B**) have indicated that this mutation is located in the heptad repeat (HR) linker region between RBD of the trimeric spike protein and the transmembrane (TM) region. Two alternate computational models developed by Woo et al., 2020 and Choi et al., 2020 (Choi et al., 2020; Woo et al., 2020) (shown in **Figure 1B**), indicate that Gly 1167 is possibly part of a short alpha helical region. The mutation to Val 1167 introduces a relatively bulkier side chain, and its effect on the secondary structure is currently unknown. However, this mutation is located in the stalk region, which is proposed to affect the bending motions of the stalk causing large movements of the RBD. The conformational flexibility of spike protein and movements of the RBD have already been suggested to be linked to its function of binding to ACE2 (Gobeil et al., 2020).

Functional Significance of N Gene Mutations

The N gene product is reported to be the nucleocapsid protein which plays a role in RNA packaging and viral particle release (Zeng et al., 2020). The crystal structure of the full length nucleocapsid protein has not been solved so far, however, it is proposed to consist of several regions (**Figure 2**) including N-arm, N terminal domain (NTD), linker region, C-terminal domain (CTD) and the C-tail. The structures of NTD and CTD from SARS-CoV-2 has been solved. N gene mutations idenfied in this study (S194L, R203K, and G204R) are located in the region 180–247, which is suggested to be a flexible linker region that lacks organized structure. Based on small angle X-ray scattering (SAXS) studies, it is proposed that this region is extended and may contain some residual secondary structure (Zeng et al., 2020).

The site of two of these mutations (S194 and G204) are fully conserved between MERS-CoV, SARS-CoV-1, and the reference SARS-CoV-2 sequence, while the site of third mutation (R203) is conserved in the SARS-CoV-1 and the reference SARS-CoV-2 sequence. The conservation of these residues between different viruses may be an indication of their functional role. It is widely discussed that protein regions that lack secondary structure become structured in presence of proper binding partners or may be involved in signaling mediated by flexibility and conformational sampling. Future studies would be important in characterizing the functional relevance of these mutations.

CONCLUSION

We sequenced five SARS-CoV-2 genomes from clinical samples collected from Oklahoma, United States between March and July 2020. Genome assembly and annotation studies identified several new mutations as well as previously reported mutations. Notably, presence of D614G mutation in the S gene was found in all the isolates. Detection of multiple mutations in the viral genomes collected from a narrow geographic region within a few months of the pandemic underscores the ability of SARS-CoV-2 to undergo rapid genomic alterations. Further studies are needed to better understand if these mutations can potentially influence host susceptibility, pathogenicity and virulence. Phylogenetic analysis of the viral genomes revealed high similarities with isolates reported from Australia, Greece and United States (Atlanta and San Diego), indicating possible multiple introductions to the state. Preliminary characterization based on available structural information of the SARS-CoV-2 proteins indicates that the mutations in S gene possibly influences conformational flexibility and motion of the spike protein, and the mutations in N gene are associated with disordered linker region within the nucleocapsid protein. In the future, mass sequencing of clinical isolates is needed to comprehensively identify genomic variations of SARS-CoV-2 in specific geographic locations.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://www.ncbi.nlm.nih.gov/genbank/, MT 998442MW 000350 MT 998443 MT 998444 MW000372 https://www.gisaid.org/, EPI_ISL_535364, EPI_ISL_535361, EPI_ISL_535362, EPI_ISL_535363, and EPI_ISL_487231.

ETHICS STATEMENT

This study was approved by the Institutional Review Board (Application Number: IRB-20-357) at Oklahoma State University, Stillwater OK 74078, United States.

AUTHOR CONTRIBUTIONS

SN and AR performed the sequencing. SN and JR performed the bioinformatics analysis. All the authors helped with sample collection. AR, AK, SM, JM, and JS helped with preparing and reviewing the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This study was partially funded by FDA-CVM-VetLIRN (Grant No. 5U18FD006671-02). PA acknowledges a grant from NIGMS under award number GM105978. This manuscript has been released as pre-print at medRxiv (Narayanan et al., 2020).

ACKNOWLEDGMENTS

We thank all the laboratory personnel involved in COVID-19 testing at the Oklahoma Animal Disease Diagnostic Laboratory. We also thank Dr. Ip Hon (National Wildlife Health center, WI, United States) for his advice on genome sequencing.

REFERENCES

Alam, I., Kamau, A., Kulmanov, M., Arold, S. T., Pain, A., Gojobori, T., et al. (2020). Functional pangenome analysis provides insights into the origin, function and pathways to therapy of SARS-CoV-2 coronavirus. *bioRxiv* [Preprint]. doi: 10.1101/2020.02.17.952895

Angeletti, S., Benvenuto, D., Bianchi, M., Giovanetti, M., Pascarella, S., and Ciccozzi, M. (2020). COVID-2019: the role of the nsp2 and nsp3 in its pathogenesis. *J. Med. Virol.* 92, 584–588. doi: 10.1002/jmv.25719

Annan, A., Baldwin, H. J., Corman, V. M., Klose, S. M., Owusu, M., Nkrumah, E. E., et al. (2013). Human betacoronavirus 2c EMC/2012–related viruses in bats, Ghana and Europe. *Emerg. Infect. Dis.* 19:456.

Banerjee, A., Sarkar, R., Mitra, S., Lo, M., Dutta, S., and Chawla-Sarkar, M. (2020). The novel coronavirus enigma: phylogeny and mutation analyses of SARS-CoV-2 viruses circulating in India during early 2020. *bioRxiv* [Preprint]. doi: 10.1101/2020.05.25.114199

Banerjee, S., Seal, S., Dey, R., Mondal, K. K., and Bhattacharjee, P. (2020). Mutational spectra of SARS-CoV-2 orf1ab polyprotein and Signature mutations in the United States of America. *bioRxiv* [Preprint]. doi: 10.1101/2020.05.01.071654

Choi, Y. K., Cao, Y., Frank, M., Woo, H., Park, S.-J., Yeom, M. S., et al. (2020). Structure, dynamics, receptor binding, and antibody binding of fully-glycosylated Full-length SARS-CoV-2 spike protein in a viral membrane. *bioRxiv* [Preprint]. doi: 10.1101/2020.10.18.343715

Corman, V. M., Muth, D., Niemeyer, D., and Drosten, C. (2018). Hosts and sources of endemic human coronaviruses. *Adv. Virus Res.* 100, 163–188. doi: 10.1016/bs.aivir.2018.01.001

Daniloski, Z., Guo, X., and Sanjana, N. E. (2020). The D614G mutation in SARS-CoV-2 Spike increases transduction of multiple human cell types. *bioRxiv* [Preprint]. doi: 10.1101/2020.06.14.151357

Dong, E., Du, H., and Gardner, L. (2020). An interactive web-based dashboard to track COVID-19 in real time. *Lancet Infect. Dis.* 20, 533–534. doi: 10.1016/S1473-3099(20)30120-1

Elbe, S., and Buckland-Merrett, G. (2017). Data, disease and diplomacy: GISAID's innovative contribution to global health. *Glob. Challeng.* 1, 33–46. doi: 10.1002/gch2.1018

Gobeil, S. M.-C., Janowska, K., McDowell, S., Mansouri, K., Parks, R., Manne, K., et al. (2020). D614G mutation alters SARS-CoV-2 spike conformational dynamics and protease cleavage susceptibility at the S1/S2 junction. *bioRxiv* [Preprint]. doi: 10.1101/2020.10.11.335299

Graat, J. M., Schouten, E. G., Heijnen, M.-L. A., Kok, F. J., Pallast, E. G. M., de Greeff, S. C., et al. (2003). A prospective, community-based study on virologic assessment among elderly people with and without symptoms of acute respiratory infection. *J. Clin. Epidemiol.* 56, 1218–1223.

Graham, R. L., Sparks, J. S., Eckerle, L. D., Sims, A. C., and Denison, M. R. (2008). SARS coronavirus replicase proteins in pathogenesis. *Virus Res.* 133, 88–100. doi: 10.1016/j.virusres.2007.02.017

Grubaugh, N. D., Hanage, W. P., and Rasmussen, A. L. (2020). Making sense of mutation: what D614G means for the COVID-19 pandemic remains unclear. *Cell* 182, 794–795. doi: 10.1016/j.cell.2020.06.040

Hu, B., Ge, X., Wang, L.-F., and Shi, Z. (2015). Bat origin of human coronaviruses. *Virol. J.* 12:221. doi: 10.1186/s12985-015-0422-1

Isabel, S., Graña-Miraglia, L., Gutierrez, J. M., Bundalovic-Torma, C., Groves, H. E., Isabel, M. R., et al. (2020). Evolutionary and structural analyses of SARS-CoV-2 D614G spike protein mutation now documented worldwide. *Sci. Rep.* 10:14031. doi: 10.1038/s41598-020-70827-z

Issa, E., Merhi, G., Panossian, B., Salloum, T., and Tokajian, S. (2020). SARS-CoV-2 and ORF3a: non-synonymous mutations and polyproline regions. *bioRxiv* [Preprint]. doi: 10.1101/2020.03.27.012013

Katoh, K., Asimenos, G., and Toh, H. (2009). Multiple alignment of DNA sequences with MAFFT. *Methods Mol. Biol.* 537, 39–64. doi: 10.1007/978-1-59745-251-9_3

Katoh, K., and Standley, D. M. (2013). Multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* 30, 772–780. doi: 10.1093/molbev/mst010

Khailany, R. A., Safdar, M., and Ozaslan, M. (2020). Genomic characterization of a novel SARS-CoV-2. *Gene Rep.* 19:100682. doi: 10.1016/j.genrep.2020.100682

Korber, B., Fischer, W. M., Gnanakaran, S., Yoon, H., Theiler, J., Abfalterer, W., et al. (2020). Tracking changes in SARS-CoV-2 Spike: evidence that D614G increases infectivity of the COVID-19 virus. *Cell* 182, 812.e19–827.e19. doi: 10.1016/j.cell.2020.06.043

Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., and Phillippy, A. M. (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* 27, 722–736. doi: 10.1101/gr.215087.116

Kumar, S., Stecher, G., Li, M., Knyaz, C., and Tamura, K. M. (2018). Molecular evolutionary genetics analysis across computing platforms. *Mol. Biol. Evol.* 35, 1547–1549. doi: 10.1093/molbev/msy096

Laha, S., Chakraborty, J., Das, S., Manna, S. K., Biswas, S., and Chatterjee, R. (2020). Characterizations of SARS-CoV-2 mutational profile, spike protein stability and viral transmission. *bioRxiv* [Preprint]. doi: 10.1101/2020.05.03.066266

Leary, S., Gaudieri, S., Chopra, A., Pakala, S., Alves, E., John, M., et al. (2020). Three adjacent nucleotide changes spanning two residues in SARS-CoV-2 nucleoprotein: possible homologous recombination from the transcription-regulating sequence. *bioRxiv* [Preprint]. doi: 10.1101/2020.04.10.029454

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 34, 3094–3100. doi: 10.1093/bioinformatics/bty191

Liu, D. X., Fung, T. S., Chong, K. K.-L., Shukla, A., and Hilgenfeld, R. (2014). Accessory proteins of SARS-CoV and other coronaviruses. *Antiviral. Res.* 109, 97–109. doi: 10.1016/j.antiviral.2014.06.013

Loman, N. J., Quick, J., and Simpson, J. T. (2015). A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nat. Methods* 12, 733–735. doi: 10.1038/nmeth.3444

Lu, W., Xu, K., and Sun, B. (2009). SARS accessory proteins ORF3a and 9b and their functional analysis. *Mol. Biol. SARS Coronavirus* 2009, 167–175.

Mackay, I. M., Arden, K. E., Speicher, D. J., O'Neil, N. T., McErlean, P. K., Greer, R. M., et al. (2012). Co-circulation of four human coronaviruses (HCoVs) in Queensland children with acute respiratory tract illnesses in 2004. *Viruses* 4, 637–653.

Narayanan, S., Ritchey, J. C., Patil, G., Teluguakula, N., More, S., Malayer, J., et al. (2020). SARS-CoV-2 genomes from Oklahoma, USA. *medRxiv* [Preprint]. doi: 10.1101/2020.09.15.20195420

Nishiura, H., Linton, N. M., and Akhmetzhanov, A. R. (2020). Initial cluster of novel coronavirus (2019-nCoV) infections in Wuhan, China is consistent with substantial human-to-human transmission. *J. Clin. Med.* 9:488.

OSDH-Covid-19 Tracker (2020). Available at: https://coronavirus.health.ok.gov/sites/g/files/gmc786/f/2020.07.10_weekly_epi_report.pdf (accessed December 20, 2020).

Owusu, M., Annan, A., Corman, V. M., Larbi, R., Anti, P., Drexler, J. F., et al. (2014). Human coronaviruses associated with upper respiratory tract infections in three rural areas of Ghana. *PLoS One* 9:e99782. doi: 10.1371/journal.pone.0099782

Pachetti, M., Marini, B., Benedetti, F., Giudici, F., Mauro, E., Storici, P., et al. (2020). Emerging SARS-CoV-2 mutation hot spots include a novel RNA-dependent-RNA polymerase variant. *J. Transl. Med.* 18, 1–9.

Pickett, B. E., Sadat, E. L., Zhang, Y., Noronha, J. M., Squires, R. B., Hunt, V., et al. (2012). ViPR: an open bioinformatics database and analysis resource for virology research. *Nucleic Acids Res.* 40, D593–D598. doi: 10.1093/nar/gkr859

Price, M. N., Dehal, P. S., and Arkin, A. P. (2010). FastTree 2 – approximately maximum-likelihood trees for large alignments. *PLoS One* 5:e9490. doi: 10.1371/journal.pone.0009490

Rambaut, A., Pybus, O. G., Nelson, M. I., Viboud, C., Taubenberger, J. K., and Holmes, E. C. (2008). The genomic and epidemiological dynamics of human influenza A virus. *Nature* 453, 615–619. doi: 10.1038/nature06945

Salipante, S. J., Roach, D. J., Kitzman, J. O., Snyder, M. W., Stackhouse, B., Butler-Wu, S. M., et al. (2015). Large-scale genomic sequencing of extraintestinal pathogenic *Escherichia coli* strains. *Genome Res.* 25, 119–128.

Statement on the second meeting of the International Health Regulations 2005 (2020). *Emergency Committee regarding the outbreak of novel coronavirus (2019-nCoV)*. Available online at: https://www.who.int/news-room/detail/30-01-2020-statement-on-the-second-meeting-of-the-international-health-regulations-(2005)-emergency-committee-regarding-the-outbreak-of-novel-coronavirus-(2019-ncov) (accessed November 7, 2020).

van Dorp, L., Acman, M., Richard, D., Shaw, L. P., Ford, C. E., Ormond, L., et al. (2020). Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. *Infect. Genet. Evol.* 83:104351. doi: 10.1016/j.meegid.2020.104351

Van Elden, L. J. R., Anton, M. A. M., Van Alphen, F., Hendriksen, K. A. W., Hoepelman, A. I. M., Van Kraaij, M. G. J., et al. (2004). Frequent detection of human coronaviruses in clinical specimens from patients with respiratory tract infection by use of a novel real-time reverse-transcriptase polymerase chain reaction. *J. Infect. Dis.* 189, 652–657.

Walls, A. C., Park, Y.-J., Tortorici, M. A., Wall, A., McGuire, A. T., and Veesler, D. (2020). Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein. *Cell* 181, 281.e6–292.e6. doi: 10.1016/j.cell.2020.02.058

Wang, S., Sundaram, J. P., and Stockwell, T. B. (2012). VIGOR extended to annotate genomes for additional 12 different viruses. *Nucleic Acids Res.* 40, W186–W192. doi: 10.1093/nar/gks528

Woo, H., Park, S.-J., Choi, Y. K., Park, T., Tanveer, M., Cao, Y., et al. (2020). Modeling and simulation of a fully-glycosylated full-length SARS-CoV-2 spike protein in a viral membrane. *bioRxiv* [Preprint]. doi: 10.1101/2020.05.20.103325

Wrapp, D., Wang, N., Corbett, K. S., Goldsmith, J. A., Hsieh, C.-L., Abiona, O., et al. (2020). Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science* 367, 1260–1263. doi: 10.1126/science.abb2507

Wu, F., Zhao, S., Yu, B., Chen, Y.-M., Wang, W., Song, Z.-G., et al. (2020). A new coronavirus associated with human respiratory disease in China. *Nature* 579, 265–269.

Zeng, W., Liu, G., Ma, H., Zhao, D., Yang, Y., Liu, M., et al. (2020). Biochemical characterization of SARS-CoV-2 nucleocapsid protein. *Biochem. Biophys. Res. Commun.* 527, 618–623. doi: 10.1016/j.bbrc.2020.04.136

Zhang, T., Wu, Q., and Zhang, Z. (2020a). Pangolin homology associated with 2019-nCoV. *bioRxiv* [Preprint]. doi: 10.1101/2020.02.19.950253

Zhang, T., Wu, Q., and Zhang, Z. (2020b). Probable pangolin origin of SARS-CoV-2 associated with the COVID-19 outbreak. *Curr. Biol.* 30, 1346.e7–1351.e7. doi: 10.1016/j.cub.2020.03.022

Zhang, Y., Zhang, J., Chen, Y., Luo, B., Yuan, Y., Huang, F., et al. (2020c). The ORF8 protein of SARS-CoV-2 mediates immune evasion through potently downregulating MHC-I. *bioRxiv* [Preprint]. doi: 10.1101/2020.05.24.111823

Zheng, J. (2020). SARS-CoV-2: an emerging coronavirus that causes a global threat. *Int. J. Biol. Sci.* 16, 1678–1685. doi: 10.7150/ijbs.45053

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.