# Next generation sequencing of SARS-CoV-2 from patient specimens of Nevada reveals occurrence of specific nucleotide variants at high frequency

Paul D. Hartley[1,5*], Richard L. Tillett[8*], Yanji Xu[6], David P. AuCoin[3,4], Joel R. Sevinsky[7], Andrew Gorzalski[2], Mark Pandori[2,4], Cyprian C. Rossetto[3,4*#], Subhash C. Verma[3,4*#]

[1] Nevada Genomics Center, Research & Innovation

[2] Nevada State Public Health Lab

[3] Department of Microbiology & Immunology

[4] School of Medicine, University of Nevada, Reno

[5] University of Nevada, Reno

[6]Nevada Center for Bioinformatics, Research & Innovation

[7]Theiagen Consulting LLC

[8] Nevada Institute of Personalized Medicine, University of Nevada, Las Vegas,

* These authors contributed equally to this work and are listed in alphabetical order

[#] To whom correspondence should be addressed:

Cyprian C. Rossetto crossetto@med.unr.edu, Subhash C. Verma

scverma@med.unr.edu

**Running title:** Analysis of SARS-CoV-2 sequences in patient specimens from Nevada

**ABSTRACT:**

**Background:**

SARS-CoV-2 is the etiological agent of coronavirus disease 2019 (COVID-19), which has a wide range of disease manifestations, including relatively mild respiratory illness to severe disease requiring hospitalization. Patients with signs of COVID-19 (fever, coughing, shortness of breath) were tested for the virus by targeting specific SARS-COV-2 genomic loci via RT-PCR using RNA extracted from nasopharyngeal/nasal swabs specimens. However, this testing does not yield genetic information required for identifying viral evolution or variants. Therefore, we sequenced 200 specimens from positively identified cases in Nevada through our robust protocol for sequencing SARS-CoV-2 genomes directly from patient's nasopharyngeal/nasal swabs. This protocol enabled identification of specific nucleotide variants including those for D614G and clade defining mutations. Additionally, these sequences were used to determine the phylogenetic relationships of SARS-CoV-2 genomes of public health importance occurring in the state of Nevada.

**Results:**

Our study reports a novel finding of a variant in the nsp12 (RdRp) protein at residue 323 (314aa of orf1b) to Phenylalanine (F) from Proline (P), present in the original isolate of SARS-CoV-2 (Wuhan-1). This 323F variant is found at a very high frequency in Northern Nevada compared to other geographic areas. Structural modeling of the RdRP interface domain indicates the P323F variant did not significantly change the protein conformation.

**Conclusions:**

Our results highlight the introduction and spread of specific SARS-CoV-2 variants at a very high frequency within a distinct geographic location. This is important for clinical and public health perspectives in understanding the transmission and evolution of SARS-CoV-2 as it circulates in humans.

**Keywords:** SARS-CoV-2, COVID-19, genome enrichment, nsp12, RdRp, orf1b 314

## BACKGROUND:

Severe Acute Respiratory Syndrome coronavirus 2 (SARS-CoV-2), the cause of coronavirus disease 2019 (COVID-19), was first identified and reported in December 2019 in Wuhan, Hubei province, China [1-3]. RNA sequencing and phylogenetic analysis of a specimens taken during the initial outbreak in Wuhan determined that the virus is most closely related (89.1% nucleotide similarity) to a group of SARS-like coronaviruses (genus Betacoribavirus, subgenus Sarbecovirus) which had previously been identify in bats in China [2,4]. Coronaviruses have a recent history as emerging infections, first SARS-CoV in 2002-2003, and Middle East respiratory syndrome coronavirus (MERS-CoV) in 2012, both zoonotic infections that cause severe respiratory illness in humans [2,5-8]. Unlike SARS-CoV and MERS-CoV which displayed limited global spread, SARS-CoV-2 has spread around the world within a few months. There are specific characteristics of SARS-CoV-2 which have facilitated the transmission, including infections that result in asymptomatic or mild disease, allowing for under-characterized transmission.

SARS-CoV-2 is an enveloped, positive single-stranded RNA virus. Detection of SARS-CoV-2 in patients has primarily occurred using RT-qPCR to detect viral RNA

from respiratory specimens (primarily nasal, nasopharyngeal swabs). While RT-PCR results can be quantified through determination of a cycle threshold (Ct) value for each sample it does not yield sequence data leading to the description of genomic variants. To further a study of such variants, and to better understand the epidemiology of the virus in the state of Nevada, we developed a workflow that allowed us to sequence SARS-CoV-2 genomic RNA from patient swabs containing a broad range of viral loads. Of the sequences of SARS-CoV-2 currently submitted to common database, several were obtained after the virus had been passed in Vero cells [9,10] and others came directly from patient specimens [11,12]. Certain data has suggested a potential for lab acquired mutations following passage in cell culture [10,13]. Specifically, a report of SARS-CoV-2 passage in Vero cells which resulted in a spontaneous 9 amino acid deletion within the spike (S) protein that overlaps with furin cleavage site [13]. The loss of this site is suggested to increase the viral entry into Vero cells [14]. For both research and epidemiological purposes, sequencing of SARS-CoV-2 directly from patient specimens not only reduces the possibility of lab-adapted mutations following passage in cell culture but also reduces the time that would be spent growing the virus from the patient specimens and subsequently also reduces handling larger amounts of infectious virus. Additionally, one of the goals in developing an optimized SARS-CoV-2 NGS protocol was to be able to generate adequate depth of coverage of the viral genome while minimizing the sequencing of non-viral RNA which would allow for more specimens to be multiplexed together during sequencing.

Our workflow employs a combination of RNA amplification, conversion into Illumina-compatible sequencing libraries and enrichment of SARS-CoV-2 library

molecules prior to sequencing. Using this novel methodology, we sequenced SARS-CoV-2 from a total of 200 patient specimens collected over a 3-month period originating from Nevada. Of the 200 selected, 173 were sequenced with enough quality to be used for determining SARS-CoV-2 nucleotide variants to perform further phylogenetic analysis and study the viral epidemiology within the state of Nevada. Analysis of the data suggests a specific epidemiological course for the local epidemic within Northern Nevada. This was characterized by an initial observation of variants closely resembling isolates originating directly from China or Europe. Subsequent to government-mandated period of restrictions on business and social activity, we observed that a viral isolate not seen elsewhere in the world emerged within Northern Nevada cases (nucleotide 14,407 and 14,408). This isolate contains an amino acid change in residue P323L/F of the RdRp (nsp12). Furthermore, we found that sampled viral isolates in Southern Nevada, unlike those in Northern Nevada, closely resembled the makeup of the United States in general.

## RESULTS:

### RNA-seq workflow and assembly of SARS-CoV-2 genomes.

A total of 200 SARS-CoV-2 positive specimens collected in Nevada from March 6th to June 5th were randomly selected to have their viral genomes sequenced for variant analysis and subsequent epidemiological studies (Fig 1A and Materials & Methods). Of the sequenced specimens, 173 had >90% coverage and sufficient depth to accurately call those genomic positions with variants (Fig. 1B). An alignment of SARS-CoV-2 genomes from these specimens is presented as supplemental data (Supplemental Fig. S1) showing 173 specimens with over 90% coverage. These 173 specimens represented 133 patient specimens from Northern Nevada (including Washoe County of which Reno is the major city, the Carson-Tahoe area, and other northern, rural counties), 40 patient specimens from Southern Nevada (Clark county which encompasses Las Vegas). Nucleotide similarity and variants were determined and used to measure the phylogenetic relationships (Supplemental Fig. S2). The combined nucleotide diversity across the entire SARS-CoV-2 genome for the Nevada specimens is shown in figure 1D, along with the genomic areas that were assessed for change in frequency corresponding to amino acids D614G, P323L/F and nucleotide 379.

During the sequencing analysis we also examined the correlation between Ct values from the diagnostic RT-PCR and percentage coverage of the viral genome to determine the performance and robustness of our sequencing method in relation to available viral RNA in a specimen of a given Ct (Fig. 1C). In general, a Ct value less

than 31 resulted in at least 90% coverage to the SARS-CoV-2 genome. Importantly, our in-house developed method for viral genome enrichment and sequencing directly from the patient's specimens (nasal and nasopharyngeal swabs) was robust and yielded over 90% coverage even at Ct values ~40. This is highly significant and shows the power of our workflow in sequencing of even those specimens with inadequate amounts of specimen (due to variability in collection) and patients samples having lower viral loads in their nasal secretions.

**Prevalence of amino acid variant D614G of SARS-CoV-2 spike protein in specimens collected in Nevada.** Earlier studies have revealed the emergence, spread and potential importance of an alteration, D614G (genomic change at 23403A>G), of the spike protein [15]. This missense mutation has become a clade-distinguishing locus that differentiates viral isolates originating in Asia from those that have emerged from Europe. A total of 173 cases were analyzed to determine the number and relative proportion of the specimens which carried the D614G spike protein variant in Nevada. The cumulative frequency for D614 and G614 were plotted from March 6th to June 5th (Fig. 2A). Specimens from the beginning of March represent the earliest known cases in Nevada, and of the 14 specimens sequenced during this time period (March 6-March 15) D614 was the predominant variant. This shifted from March to June with an increasing frequency of the G614 allele. The trend for specimens originating from either Northern Nevada (N-NV) and Southern Nevada (S-NV) both show a higher frequency of G614 (Fig. 2B). We used a subsampling of sequence data from Nextstrain.org to assess the frequency of D614G in the United States and globally during the same time period (March 6th to June 5th) (Fig. 2B and 2C). The global trend by continent of

D614G is also similar, with G614 at a higher frequency, the one noted exception is in Asia, where D614 and G614 continue to exist in equal proportions (Fig. 2C).

**Frequency of SARS-CoV-2 clades in Nevada.** Worldwide, there are currently 5 main clades (19A, 19B, 20A, 20B, 20C) of SARS-CoV-2 differentiated based on specific nucleotide profiles in the Year-letter scheme of https://clades.nextstrain.org. Clade 19A and 19B are defined by C8782T and T28144C respectively. 20A is a derivative of 19A and contains mutations C3037T, C14408T and A23403G (resulting in D614G). 20B is defined by mutations G28881A, G28882A and G28883C, and 20C contains C1059T and G25563T [16].

To assess the introduction and spread of the clades in Nevada the cumulative frequency for the clades were plotted from March 6th to June 5th (Fig. 3A). The earliest sequenced specimens from Nevada were collected in the beginning of March (March 6-March 15) and are predominantly from clades 19A and 19B. Additional sequenced specimens collected from March to June revealed a shift to a higher frequency of 20C (Fig. 3A). We performed phylogenetic reconstruction of the Nevada specimens and differentiated the clades on the circular dendrogram by color (Fig. 3B). There were discordant trends in the dominant clade for specimens originating from either Northern Nevada (N-NV) and Southern Nevada (S-NV) (Fig. 3C and 3D). Specimens from Northern Nevada (Washoe County, Carson-Tahoe, and other counties) showed a prevalence of 20C, while the Southern Nevada specimens from Clark county had a larger proportion of 20A (Fig. 3C and Supplemental Fig. S3a)**.** We used a subsampling of Nextstrain.org data to assess the frequency of clades in the United States and globally by continent during the same time period (March 6th to June 5th). The

dominant clade in the United States was 20C, similar to the frequency seen in the total Nevada samples (N-NV and S-NV) (Fig. 3A and 3C). The global clade distributions were variable in areas outside of Asia while clades 19A and 19B are noted to be more prevalent in Asia. (Fig. 3D).

**Prevalence of amino acid variant P323L/F of SARS-CoV-2 nsp12 (RdRp) in Nevada.** Analysis of sequencing data revealed a novel observation for our specimens at bases 14,407 and 14,408 which results in a change at residue 323 in nsp12 (RdRp). For the Wuhan isolate at 14,407 and 14,408 there is CC for proline (P), the variants have CT for leucine (P323L) and TT for phenylalanine (P323F). To assess the introduction and spread of P323L/F in Nevada, the cumulative frequency of P323, L323 and F323 were plotted from March 6th to June 5th (Fig. 3A). Nevada specimens from the beginning of March (March 6-March 15) show P323 was the predominant variant. As additional specimens were collected and sequenced from March to June there was a shift to a higher frequency of L323 and F323 (Fig. 4A). We performed phylogenetic reconstruction of the Nevada specimens and noted the P323L/F variants on the circular dendrogram with the indicated colors (Fig. 4B). Interestingly, analysis of the Northern Nevada and Southern Nevada specimen showed very different dominant variants (Fig. 4B). In Northern Nevada the F323 was more prevalent, and in Southern Nevada L323 was more prevalent. We used a subsampling of Nextstrain.org data to assess the frequency of P323L/F in the United States and globally during the same time period (March 6th to June 5th). P323 is the predominant variant in Asia, while L323 is more prevalent in other areas of the world and F323 is only appreciably noted in North America (Fig. 4D and 4E).

## DISCUSSION:

We have developed a novel method which combines specific depletion and enrichment strategies that results in efficient SARS-CoV-2 RNA-seq with high genome coverage and depth. An advantage of this protocol is that it generates sequence data directly from swab specimens without the need to passage the virus in cell culture thereby reducing the handling of infectious material and induction of culture-acquired mutations. Another obstacle in sequencing directly from swab specimens is that most FDA-approved commercially available RNA extraction kits are specifically optimized to recover low amounts of total nucleic acids, include carrier polyA RNA that could be convertible into sequence able molecules, as has been observed previously with RNA-seq of Lassa- or Ebola-positive clinical specimens [17].

The workflow incorporates amplification of low-abundance RNA into micrograms of DNA, followed by conversion from a fraction of the DNA into Illumina-compatible sequencing libraries and enrichment of these libraries for SARS-CoV-2 sequences. In addition, during the reverse transcription step a reagent was incorporated to reduce the subsequent amplification of host ribosomal RNA. This approach is robust in that it converts low amounts of RNA into microgram quantities of DNA representative of all the RNA species (aside of rRNA) present in the specimen. This DNA can be stored indefinitely to be interrogated by multiple techniques at a later date. Additionally, RNA amplification is likely less sensitive to low viral abundance compared to RT-PCR.

Finally, the use of probes to enrich for coronavirus-specific sequencing library molecules is less sensitive to variants compared to tiling PCR amplicon approaches [18-22].

The data herein implicate that early in the pandemic, before the “stay-at-home” order on April 1st, there were multiple introductions of SARS-CoV-2 into the state of Nevada. From April 1st to the beginning of June Nevada experienced a period of semi-isolation, as the casinos and most hotels shut down, tourism and travel to the state essentially stopped. Because of the stay-at-home order and social distancing measures put in place there was less mobility of people within and between states. It is possible that these measures, compounded by potential inherent transmission variability of some viral isolates, influenced the change in the frequency of D614G, clades and P323L/F that we noted during this time period within Nevada. In addition, we also found 379C>A had a high prevalence in our study specimens compared to the subsampling of sequences from the United States and international (supplemental Fig. S4). This is a synonymous mutation in nsp1, hence the biological relevance of this nucleotide variant remains to be elucidated.

Of the 14,885 complete SARS-CoV-2 genomes currently available (as of August 14, 2020) in NCBI there are only 6 genomes that have the P323F variant (accession number: MT706208, LR860619, MT345877, MT627429, MT810889, MT811171). In this study 62 of the 133 specimens from Northern Nevada contain P323F. That is 46% of specimens from Northern Nevada contain P323F compared to 0.04% of NCBI deposited SARS-CoV-2 isolates. We performed structural modeling of the P323L/F variation and did not find any significant change to the nsp12 confirmation with either

P323L or P323F (Supplemental Fig. S5), therefore this variant is most likely a neutral mutation and does not confer either a fitness advantage or disadvantage to transmission or pathogenicity of SARS-CoV-2.

We find that the overall trend for D614G in Nevada during this time period was similar to what was observed in other states and internationally, with the exception of within Asia where the D614 allele had originated. We noted that there were differences between Northern Nevada and Southern Nevada. In Northern Nevada clade 20C and F323 were more frequent, while during this same time period in Southern Nevada clade 20A and L323 were more prevalent. These data indicate that there were distinct genomic profiles of the SARS-CoV-2 viruses that were circulating in these populations during the initial months of the pandemic while stay-at-home order were in place to help prevent transmission of the virus.

**CONCLUSIONS:**

SARS-COV-2 is currently a global pandemic on a scale that has not been experienced since the 1918 influenza pandemic. Viral genomes accumulate nucleotide variations while passing in the human population and these mutations may confer phenotypic differences including immune response and anti-viral drug resistance. We have developed a protocol to enrich SARS-CoV-2 RNA directly from patient nasal/nasopharyngeal specimens that does not rely on amplicon based sequencing strategies nor the need of passing the virus into tissue culture and reducing the possible introduction of lab-adapted mutations. These specimens were subjected to next generation sequencing (NGS) to determine the phylogenetic relationship of SARS-CoV-

2 sequences within Nevada during the early months (March to June) of the pandemic. Our study identified a very rare mutation in the RdRp protein (P323F) found to occur at a very high frequency among patients of Nevada. This may be important for clinical and public health perspectives in understanding the transmission and evolution of SARS-CoV-2 as it circulates in humans.

## METHODS:

**SARS-CoV-2 specimen and library preparation.** Nasal and Nasopharyngeal swab specimens were received at the Nevada State Public Health Lab (NSPHL) or Southern Nevada Public Health Lab (SNPHL) and RNA extraction was completed using either a QIAamp Viral RNA Mini Kit (QIAGEN) or Mag-Bind Viral DNA/RNA kit (Omega Biotek). Specimens were tested for the presence of coronaviral RNA using FDA-approved kits that employed RT-PCR to detect SARS-COV-2 RNA.

A set of 200 coronavirus positive specimens were selected for genome sequencing. Specimens were treated with DNase I (QIAGEN) for 30 minutes at room temperature and concentrated using RNeasy Minelute spin columns (QIAGEN) based on the manufacturer supplied protocol. These concentrated samples were converted into Illumina-compatible sequencing libraries with a QIAseq FX Single Cell RNA Library kit (QIAGEN). RNA samples were annealed to a 1:12.5 dilution of QIAseq FastSelect -HMR probes (QIAGEN) to reduce subsequent amplification of human ribosomal RNA. After treatment to remove trace DNA from the samples, a reverse transcription reaction was carried out using random hexamers. The synthesized DNA was ligated to one another, followed by isothermal linear amplification. Amplified DNA (1 μg) was

enzymatically sheared to an average insert size of 300 bp, and Illumina-compatible dual-indexed sequencing adapters were ligated to the ends. Next, about 300 ng of adapter-ligated sample was amplified with 6 cycles of PCR with KAPA HiFi HotStart polymerase (Roche Sequencing Solutions). Enrichment of library molecules containing SARS-CoV-2 sequence was conducted with a myBaits kit and coronavirus-specific biotinylated probes (Arbor Biosciences). Each enrichment used 500 ng of PCR-amplified DNA, was carried out based on manufacturer instructions at a hybridization temperature of 65° for 16 hours, and was completed with 8-16 cycles of PCR using KAPA HiFi HotStart polymerase. Samples were sequenced using an Illumina Next-seq mid-output (2 x 75). The generated FASTQ files from the sequencing reaction were analyzed as described below. The data files are available at GISAID, NCBI under the https://www.ncbi.nlm.nih.gov/bioproject/657893

**Computational analysis.** Sequence pair libraries were trimmed using Trimmomatic, version 0.39 and adapter-clipping setting “2:30:10:2:keepBothReads” [23]. Read pairs were aligned against the Wuhan reference genome (NC_045512.2) by Bowtie 2, version 2.3.5, local alignment [24]. PCR optical duplicates were removed via Picard MarkDuplicates [25].

Variants were called using Freebayes, version 1.0.2, with ploidy set to 1, minimum allele frequency 0.75, and minimum depth of 4 [26]. No variants were called in the first 200 bp and final 63 bp of the COVID-19 genome. High-quality variant sites were selected where site “QUAL > 20” using *vcffilter*, VCFlib version 1.0.0_rc2 [27]. Individual genomes were reconstructed by their filter-passing variants using *bcftools consensus*

and only where aligned coverage depth ≥ 4; bases with coverage below four are reported as unknown (Ns) [28].

A set of 3,644 complete, high-coverage SARS-CoV-2 genomes reported in the July 15, 2020 Nextstrain.org global subsample and metadata were obtained from GISAID and combined with our own samples to determine their phylogenetic placement [29,30]. Four of the global samples were set aside after screening for unexpected FASTA characters. The combined sets of global and Nevadan samples were aligned together, with metadata, by the *augur* phylodynamic pipelines of the *ncov* build of the *nextstrain* command-line tool, version 2.0.0.post1 [31].

Cumulative frequency of D614G, clades (19A, 19B, 20A, 20B, 20C), and P323L/F were calculated at each time point based on the total number of specimens up to the indicated date. Plots and pie charts were generated using GraphPad Prism (version 8).

**nsp12 protein modeling.** Sequence of nsp12 (RdRp) protein for SARS-CoV-2 (YP_009725307.0) was retrieved from NCBI protein database and 3D model was structured based on a previously published report (PDB ID: 6XEZ [32]). In addition to nsp12 (chain A), the model also contains nsp7 (chain C), nsp8 (chain B and D), nsp13 (chain E and F), ligands (ZN, Mg) and RNA template and product strands. Mutational changes to residue 323 within nsp12 were performed using PyMol Molecular Graphics System (version 2.0, Schrödinger LLC). The original proline (P) was mutated to either leucine (L) or phenylalanine (F) as indicated, these residues along with residues containing side chains within 5 Å of P323L/F are shown as sticks. The rotamers for

each P323L/F were assessed and those with least rotational strain and steric hindrance were used to generate the final image. To determine any NCBI deposited sequences which contain the P323F variant, standard protein BLAST from the BLASTp suite was used to find nsp12 protein sequences which contained FSTVFP**F**TSFGP (P323F is bold and underline) from full length SARS-CoV-2 genomes. The P323F amino acid changes were confirmed with the NCBI deposited nucleotide sequences.

## DECLARATIONS:

### Ethics Approval:

Deidentified human specimens (nasal and nasopharyngeal swabs) were used for the extraction of viral RNA all the experiments were done in accordance with guidelines of the University of Nevada, Reno. The University of Nevada, Reno Institutional Review Board (IRB) reviewed this project and determined this study to be EXEMPT FROM IRB REVIEW according to federal regulations and University policy.

The Environmental and Biological Safety committee of the University of Nevada, Reno, approved methods and techniques used in this study.

### Competing Interests:

The authors declare that they have no competing interests with the contents of this article.

## DATA AVAILABILITY:

All reported data are deposited and available at GISAID with the following accession numbers: EPI_ISL_515409, EPI_ISL_515407, EPI_ISL_515408, EPI_ISL_515375,

370 EPI_ISL_515376, EPI_ISL_515373, EPI_ISL_515374, EPI_ISL_515371,
371 EPI_ISL_515372, EPI_ISL_515370, EPI_ISL_515416, EPI_ISL_515417,
372 EPI_ISL_515414, EPI_ISL_515415, EPI_ISL_515379, EPI_ISL_515412,
373 EPI_ISL_515413, EPI_ISL_515377, EPI_ISL_515410, EPI_ISL_515378,
374 EPI_ISL_515411, EPI_ISL_515364, EPI_ISL_515365, EPI_ISL_515362,
375 EPI_ISL_515363, EPI_ISL_515360, EPI_ISL_515361, EPI_ISL_515405,
376 EPI_ISL_515406, EPI_ISL_515403, EPI_ISL_515404, EPI_ISL_515368,
377 EPI_ISL_515401, EPI_ISL_515369, EPI_ISL_515402, EPI_ISL_515366,
378 EPI_ISL_515367, EPI_ISL_515400, EPI_ISL_515353, EPI_ISL_515354,
379 EPI_ISL_515351, EPI_ISL_515352, EPI_ISL_515350, EPI_ISL_515359,
380 EPI_ISL_515357, EPI_ISL_515358, EPI_ISL_515355, EPI_ISL_515356,
381 EPI_ISL_515342, EPI_ISL_515343, EPI_ISL_515340, EPI_ISL_515461,
382 EPI_ISL_515341, EPI_ISL_515462, EPI_ISL_515460, EPI_ISL_515348,
383 EPI_ISL_515349, EPI_ISL_515346, EPI_ISL_515347, EPI_ISL_515344,
384 EPI_ISL_515345, EPI_ISL_515331, EPI_ISL_515298, EPI_ISL_515452,
385 EPI_ISL_515332, EPI_ISL_515299, EPI_ISL_515453, EPI_ISL_515296,
386 EPI_ISL_515450, EPI_ISL_515330, EPI_ISL_515297, EPI_ISL_515451,
387 EPI_ISL_515294, EPI_ISL_515295, EPI_ISL_515293, EPI_ISL_515339,
388 EPI_ISL_515337, EPI_ISL_515458, EPI_ISL_515338, EPI_ISL_515459,
389 EPI_ISL_515335, EPI_ISL_515456, EPI_ISL_515336, EPI_ISL_515457,
390 EPI_ISL_515333, EPI_ISL_515454, EPI_ISL_515334, EPI_ISL_515455,
391 EPI_ISL_515319, EPI_ISL_515320, EPI_ISL_515441, EPI_ISL_515321,
392 EPI_ISL_515442, EPI_ISL_515440, EPI_ISL_515328, EPI_ISL_515449,

EPI_ISL_515329, EPI_ISL_515326, EPI_ISL_515447, EPI_ISL_515327, EPI_ISL_515448, EPI_ISL_515324, EPI_ISL_515445, EPI_ISL_515325, EPI_ISL_515446, EPI_ISL_515322, EPI_ISL_515443, EPI_ISL_515323, EPI_ISL_515444, EPI_ISL_515308, EPI_ISL_515429, EPI_ISL_515309, EPI_ISL_515390, EPI_ISL_515397, EPI_ISL_515430, EPI_ISL_515310, EPI_ISL_515398, EPI_ISL_515431, EPI_ISL_515395, EPI_ISL_515396, EPI_ISL_515393, EPI_ISL_515394, EPI_ISL_515391, EPI_ISL_515392, EPI_ISL_515317, EPI_ISL_515438, EPI_ISL_515318, EPI_ISL_515439, EPI_ISL_515315, EPI_ISL_515436, EPI_ISL_515316, EPI_ISL_515437, EPI_ISL_515313, EPI_ISL_515434, EPI_ISL_515314, EPI_ISL_515435, EPI_ISL_515311, EPI_ISL_515399, EPI_ISL_515432, EPI_ISL_515312, EPI_ISL_515433, EPI_ISL_515418, EPI_ISL_515419, EPI_ISL_515386, EPI_ISL_515387, EPI_ISL_515420, EPI_ISL_515384, EPI_ISL_515385, EPI_ISL_515382, EPI_ISL_515383, EPI_ISL_515380, EPI_ISL_515381, EPI_ISL_515306, EPI_ISL_515427, EPI_ISL_515307, EPI_ISL_515428, EPI_ISL_515304, EPI_ISL_515425, EPI_ISL_515305, EPI_ISL_515426, EPI_ISL_515302, EPI_ISL_515423, EPI_ISL_515303, EPI_ISL_515424, EPI_ISL_515300, EPI_ISL_515388, EPI_ISL_515421, EPI_ISL_515301, EPI_ISL_515389, EPI_ISL_515422

## AUTHOR CONTRIBUTION:

PDH: conceptualization, formal analysis, methodology, writing – original draft preparation, writing – review and editing

RLT: conceptualization, formal analysis, methodology, writing – original draft preparation, writing – review and editing

XY: formal analysis, writing – review and editing

DPA: conceptualization, funding

JRS: formal analysis, writing – review and editing

AG: methodology

MP: conceptualization, formal analysis, methodology, project administration, funding, writing – review and editing

CCR: conceptualization, formal analysis, methodology, project administration, funding, writing – original draft preparation, writing – review and editing

SCV: conceptualization, formal analysis, methodology, project administration, funding, writing – review and editing

**ACKNOWLEDGEMENTS:**

We thank the staff at the Nevada State Public Health Lab (NSPHL). This work was supported by NSPHL, University of Nevada, Reno Vice President for Research and Innovation (VPRI), Department of Microbiology & Immunology, UNR School of Medicine, Nevada IDeA Network of Biomedical Research Excellence (INBRE) from the National Institute of General Medical Sciences (GM 103440 and GM 104944) from the National Institutes of Health (NIH). The authors wish to acknowledge the support of Research & Innovation and the Office of Information Technology at the University of Nevada, Reno for computing time on the Pronghorn High-Performance Computing Cluster.

## REFERENCES:

1. Zhou P, Yang X-L, Wang X-G, Hu B, Zhang L, Zhang W, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature. 2020;579: 270–273.
2. Petrosillo N, Viceconte G, Ergonul O, Ippolito G, Petersen E. COVID-19, SARS and MERS: are they closely related? Clin Microbiol Infect. 2020;26: 729–734.
3. Viruses CSG of TIC on T of, Coronaviridae Study Group of the International Committee on Taxonomy of Viruses. The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. Nature Microbiology. 2020. pp. 536–544. doi:10.1038/s41564-020-0695-z
4. Hu D, Zhu C, Ai L, He T, Wang Y, Ye F, et al. Genomic characterization and infectivity of a novel SARS-like coronavirus in Chinese bats. Emerg Microbes Infect. 2018;7: 154.
5. de Wit E, van Doremalen N, Falzarano D, Munster VJ. SARS and MERS: recent insights into emerging coronaviruses. Nat Rev Microbiol. 2016;14: 523–534.
6. Peiris JSM, Lai ST, Poon LLM, Guan Y, Yam LYC, Lim W, et al. Coronavirus as a possible cause of severe acute respiratory syndrome. Lancet. 2003;361: 1319–1325.
7. Zhong NS, Zheng BJ, Li YM, Poon, Xie ZH, Chan KH, et al. Epidemiology and cause of severe acute respiratory syndrome (SARS) in Guangdong, People's Republic of China, in February, 2003. Lancet. 2003;362: 1353–1358.
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: 1814–1820.

9. Licastro D, Rajasekharan S, Dal Monego S, Segat L, D'Agaro P, Marcello A. Isolation and Full-Length Genome Characterization of SARS-CoV-2 from COVID-19 Cases in Northern Italy. J Virol. 2020;94. doi:10.1128/JVI.00543-20

10. Kim D, Lee J-Y, Yang J-S, Kim JW, Kim VN, Chang H. The Architecture of SARS-CoV-2 Transcriptome. Cell. 2020;181: 914–921.e10.

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

12. Holland LA, Kaelin EA, Maqsood R, Estifanos B, Wu LI, Varsani A, et al. An 81 nucleotide deletion in SARS-CoV-2 ORF7a identified from sentinel surveillance in Arizona (Jan-Mar 2020). J Virol. 2020. Available: https://jvi.asm.org/content/jvi/early/2020/04/30/JVI.00711-20.full.pdf?mod=article_inline

13. Davidson AD, Williamson MK, Lewis S, Shoemark D, Carroll MW, Heesom KJ, et al. Characterisation of the transcriptome and proteome of SARS-CoV-2 reveals a cell passage induced in-frame deletion of the furin-like cleavage site from the spike glycoprotein. Genome Medicine. 2020. doi:10.1186/s13073-020-00763-0

14. Walls AC, Park Y-J, Tortorici MA, Wall A, McGuire AT, Veesler D. Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. Cell. 2020;181: 281–292.e6.

15. Korber B, Fischer WM, Gnanakaran S, Yoon H, Theiler J, Abfalterer W, et al. Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus. Cell. 2020. doi:10.1016/j.cell.2020.06.043

16. Nextclade. [cited 17 Aug 2020]. Available: https://clades.nextstrain.org

17. Matranga CB, Andersen KG, Winnicki S, Busby M, Gladden AD, Tewhey R, et al. Enhanced methods for unbiased deep sequencing of Lassa and Ebola RNA viruses from clinical and biological samples. Genome Biol. 2014;15: 519.

18. O'Flaherty BM, Li Y, Tao Y, Paden CR, Queen K, Zhang J, et al. Comprehensive viral enrichment enables sensitive respiratory virus genomic identification and analysis by next generation sequencing. Genome Res. 2018;28: 869–877.

19. Briese T, Kapoor A, Mishra N, Jain K, Kumar A, Jabado OJ, et al. Virome Capture Sequencing Enables Sensitive Viral Diagnosis and Comprehensive Virome Analysis. mBio. 2015. doi:10.1128/mbio.01491-15

20. Paskey AC, Frey KG, Schroth G, Gross S, Hamilton T, Bishop-Lilly KA. Enrichment post-library preparation enhances the sensitivity of high-throughput sequencing-based detection and characterization of viruses from complex samples. BMC Genomics. 2019;20: 155.

21. Paden CR, Tao Y, Queen K, Zhang J, Li Y, Uehara A, et al. Rapid, Sensitive, Full-Genome Sequencing of Severe Acute Respiratory Syndrome Coronavirus 2. Emerg Infect Dis. 2020;26. doi:10.3201/eid2610.201800

22. Xiao M, Liu X, Ji J, Li M, Li J, Yang L, et al. Multiple approaches for massively parallel sequencing of SARS-CoV-2 genomes directly from clinical samples. Genome Med. 2020;12: 57.

23. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30: 2114–2120.

24. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nature Methods. 2012. pp. 357–359. doi:10.1038/nmeth.1923

509 25. Picard Toolkit. In: Github [Internet]. 2019 [cited 17 Aug 2020]. Available:
510 http://broadinstitute.github.io/picard
511 26. Garrison E, Marth G. Haplotype-based variant detection from short-read
512 sequencing. arXiv [q-bio.GN]. 2012. Available: http://arxiv.org/abs/1207.3907
513 27. Garrison E. Vcflib: A C++ library for parsing and manipulating VCF files. 2019.
514 Available: https://github.com/vcflib/vcflib
515 28. Li H. A statistical framework for SNP calling, mutation discovery, association
516 mapping and population genetical parameter estimation from sequencing data.
517 Bioinformatics. 2011;27: 2987–2993.
518 29. Elbe S, Buckland-Merrett G. Data, disease and diplomacy: GISAID's innovative
519 contribution to global health. Global Challenges. 2017. pp. 33–46.
520 doi:10.1002/gch2.1018
521 30. Nextstrain. Genomic epidemiology of novel coronavirus - Global subsampling. In:
522 Nextstrain.org [Internet]. 15 Jul 2020 [cited 15 Jul 2020]. Available:
523 https://nextstrain.org/ncov/global/2020-07-15?d=tree&l=clock&legend=closed
524 31. Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, Callender C, et al.
525 Nextstrain: real-time tracking of pathogen evolution. Bioinformatics. 2018;34: 4121–
526 4123.
527 32. Chen J, Malone B, Llewellyn E, Grasso M, Shelton PMM, Olinares PDB, et al.
528 Structural Basis for Helicase-Polymerase Coupling in the SARS-CoV-2 Replication-
529 Transcription Complex. Cell. 2020. doi:10.1016/j.cell.2020.07.033

530
531

532

533

## FIGURE LEGENDS:

**Figure 1. Workflow of SARS-CoV-2 genome sequencing and analysis from nasopharyngeal patient specimens in Nevada.** (A) RNA was extracted from Nasal or Nasopharygeal (NP) swabs taken from patients in Nevada and first used to determine the presence of SARS-CoV-2 genomes by RT-qPCR. Next generation sequencing (NGS) libraries were prepared from positive specimens, this included steps for ribosomal RNA depletion and SARS-CoV-2 enrichment. Subsequent libraries were pooled and used for whole genome sequencing at the Nevada Genomics Center on the Illumina NextSeq 500 instrument. FASTQ files were aligned to the reference genome, and analyzed to determine nucleotide variation and phylogenetic relationship. (B) A total of 200 specimens were sequenced, of which 174 had over 99% coverage of the SARS-CoV-2 genome. This included 133 patient specimens from Northern Nevada, 40 from Southern Nevada and 1 specimen that was re-sequenced. (C) Correlation between RT-qPCR Ct value and the percentage of coverage in the whole genome sequencing after trimming and alignment. (D) Nucleotide variants across the SARS-CoV-2 genome in the 173 specimens from Nevada from March 6th to June 5th.

**Figure 2. Distribution of D614G in Nevada and comparison with the United States and global proportion.** (A) Cumulative frequency of D614G in 173 patient specimens from Nevada from March 6th to June 5th, 2020 (D614 is indicated by teal, G614 is indicated by yellow). Pie charts depict the cumulative proportion up to the indicated time point (March 15, April 1, May 1, June 5). The total number of specimens included at each time point is specified below each pie chart. Effective dates of emergency orders

and regulatory responses to SARS-CoV-2 spread in Nevada are indicated on the frequency graph time axis. (B) Proportion of D614G in the United States from March 6th to June 5th, specimens from Nevada are divided in the geographic area that they originated from, Northern Nevada (N-NV) includes 133 specimens from Washoe county, Carson-Tahoe, and other northern counties, and Southern Nevada (S-NV) includes 40 specimens from Clark county. (C) Global proportion of D614G in the shown regions during the same time period from a subsampling of sequences deposited in Nextstrain.org. The size of the pie chart corresponds to the relative specimen number for each region.

**Figure 3. Distribution of SARS-CoV-2 clades in Nevada.** (A) Cumulative frequency of SARS-CoV-2 clades in 173 patient specimens from Nevada during March 6th to June 5th. The five clades are colored 19A (blue), 19B (teal), 20A (green), 20B (yellow) and 20C (orange). Pie charts depict the cumulative proportion up to the indicated time point (March 15, April 1, May 1, June 5). The total number of specimens included at each time point is specified below each pie chart. Dates of emergency orders and regulations meant to slow the spread of SARS-CoV-2 in Nevada are indicated on the time scale of the frequency graph. (B) Circular dendrogram depicting clades from Nevada specimens. (C) Pie chart of the clades from northern Nevada (N-NV), southern Nevada (S-NV) and the United States. (D) Pie charts show the proportion of clades from global regions during the same time period from a subsampling of sequences deposited in Nextstrain.org. The size of the pie chart corresponds to the relative specimen number for each region.

**Figure 4. Distribution of P323L/F (nsp12, RdRp) in Nevada.** (A) Cumulative frequency of P323L/F (nsp12, RdRp) in 173 patient specimens from Nevada during March 6th to June 5th. The amino acid at position 323 is indicated by teal for proline (P), yellow for leucine (L) and blue for phenylalanine (F). Pie charts depict the cumulative proportion up to the indicated time point (March 15, April 1, May 1, June 5). The total number of specimens included at each time point is specified below each pie chart. Dates of emergency orders and regulations meant to slow the spread of SARS-CoV-2 in Nevada are indicated on the time scale of the frequency graph. (B) Circular dendrogram representing the distribution of amino acid change at residue 323 of nsp12 from a global subsampling of sequences deposited in Nextstrain.org from March 6th to June 5th, the larger dots indicate specimens from Nevada. (C) Pie chart indicating the ratio of P/L/F in Northern NV and Southern NV specimens from this study. (D) Proportion of P323L/F from a subsampling of sequences deposited in Nextstrain.org for the United States and (E) global regions from March 6th to June 5th. The size of the pie chart corresponds to the relative specimen number for each region.

## SUPPLEMENTARY FIGURES

**Figure S1. Alignment of SARS-CoV-2 sequences from NV patient's specimens.** SARS-CoV-2 sequences from patient specimen were aligned together with the original COVID-19 sequence from Wuhan, NC_045512.2, using multiple sequences alignment tool MUSCLE 3.8.31. The aligned sequences were sorted based on number of Ns in each sequence from the smallest to the largest. The sorted alignment file in .aln format

603 was uploaded to the web portal of NCBI Multiple Sequence Alignment Viewer, Version
604 1.15.0, for visualization.

605
606 **Figure S2. Dendrogram of Nevada specimens in context of other sequenced**
607 **specimens.** (a) Nucleotide mutation clock from Nextstrain.org with the SARS-CoV-2
608 genome isolated from Washington (USA/WA 1/2020) on January 24th and first
609 specimen from Nevada (A0004) on March 5th indicated in red. (b) Circular dendrogram
610 of Nevada specimens from March 6th to June 5th positioned within a subsample of
611 global sequences from Nextstrain.org during the same time period, the larger dots
612 indicate specimens from Nevada. The five clades are colored 19A (blue), 19B (teal),
613 20A (green), 20B (yellow) and 20C (orange).

614

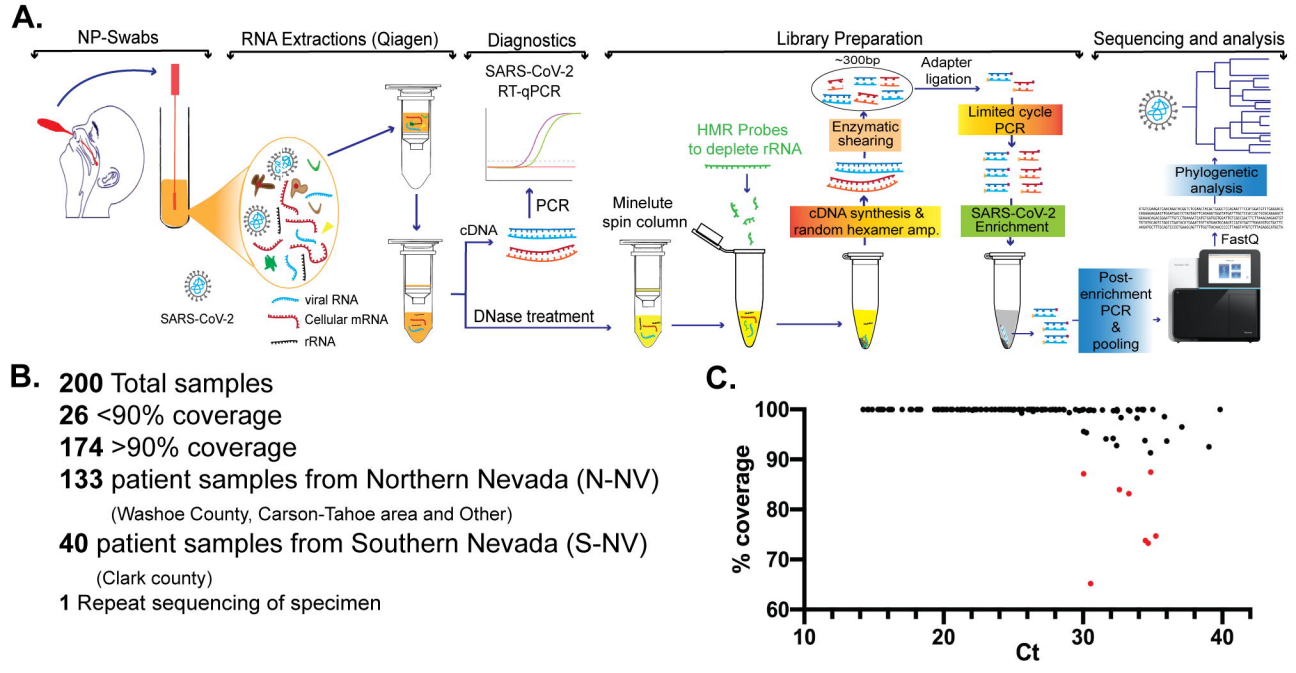
615 **Figure S3. Global distribution of clades from March 6th to June 5th.** Pie charts
616 depict the proportion of the clades within (a) four main areas in Nevada, these include
617 91 specimens from Washoe county (upper left), 23 specimens from Carson-Tahoe
618 (middle left), 40 specimens from Clark county (bottom right), and 19 from rural Nevada
619 (middle). The five clades are colored 19A (blue), 19B (teal), 20A (green), 20B (yellow)
620 and 20C (orange). (b) Pie chart of the clades in each indicated country from March 6th
621 to June 5th generated from a subsampling of sequences deposited in Nextstrain.org.
622 The size of the pie chart corresponds to the relative specimen number for each region.

623

624 **Figure S4. Distribution of nucleotide variant 379C>A.** (a) Green line at the far left
625 end of the genome denoted nucleotide position 379 of nsp1. (b) Circular dendrogram of
626 global subsample of sequences from Nextstrain.org with NV specimens indicated by

larger dots. (c) Pie chart indicating the proportion of sequences with either the cytidine (C) or adenosine (A) at position 379 from the Nevada specimens. Subsample of sequences from Nextstrain.org were used to generate the proportion of 379C>A in (d) the indicated states within the U.S. and (e) internationally. The size of the pie chart corresponds to the relative specimen number for each region.

**Figure S5. Structure of SARS-CoV-2 nsp12 (RdRp) P323L/F.** (a) Diagram depicting ORF1b genomic location and encoded proteins. Below the linear protein schematic of nsp12 specific nucleotide variants at position 14,407 and 14,408 and the resulting amino acid changes are indicated. (b) SARS-CoV-2 replicase complex modeled from 6XEZ template. This model includes nsp12, nsp7, nsp8, nsp13, ligands (ZN, Mg), the template and product strand of RNA. P323L/F within the interface domain of nsp12 is located at the left side of the complex. (c) Model nsp12 showing P323L/F within the interface domain. Residue 323 is shown with either P, L or F and amino acids with side chains within 5 Å of residue 323 are depicted as sticks in the cartoon model.




**B.**

**200** Total samples

**26** <90% coverage

**174** >90% coverage

**133** patient samples from Northern Nevada (N-NV)

(Washoe County, Carson-Tahoe area and Other)

**40** patient samples from Southern Nevada (S-NV)

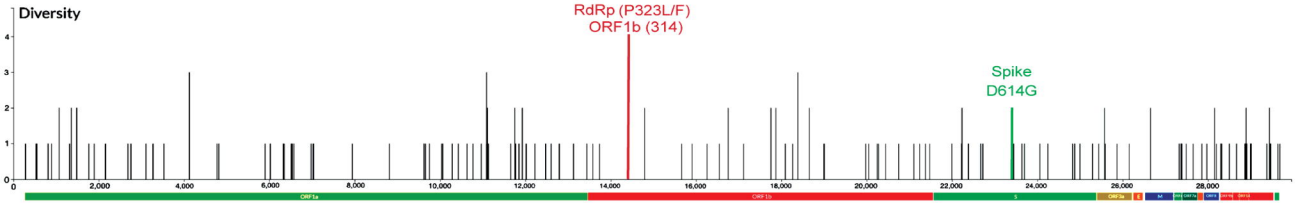
(Clark county)

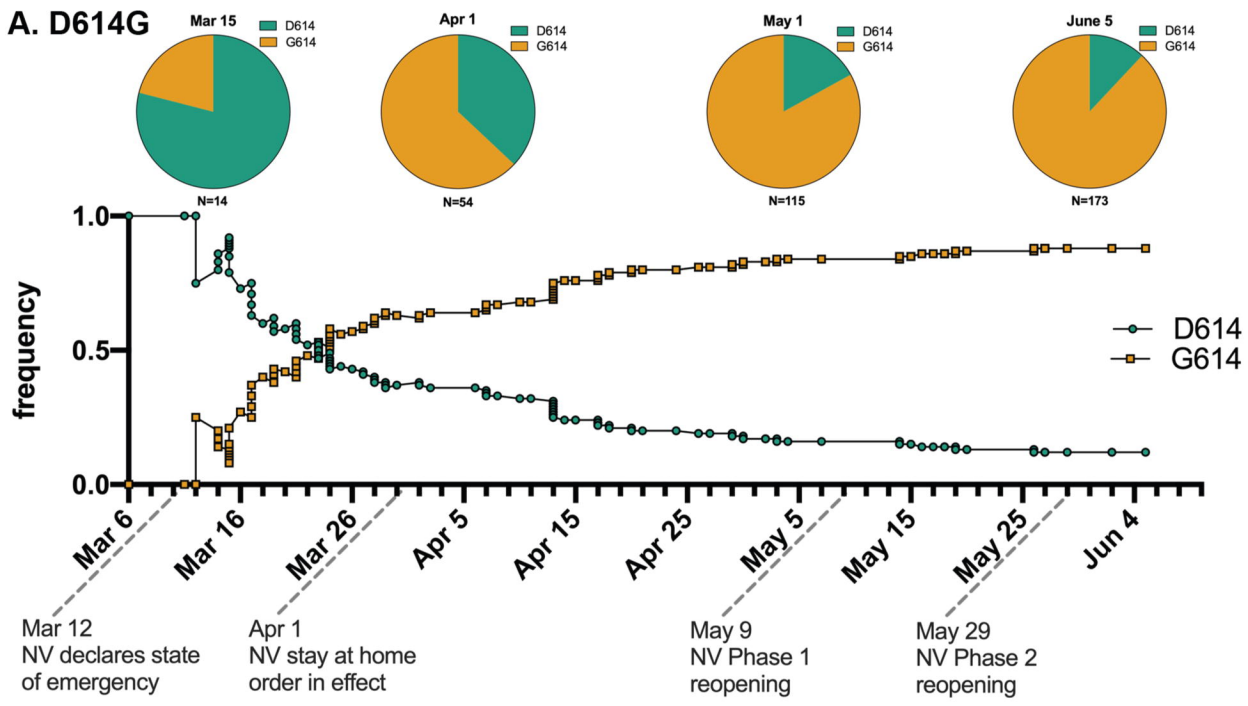
**1** Repeat sequencing of specimen

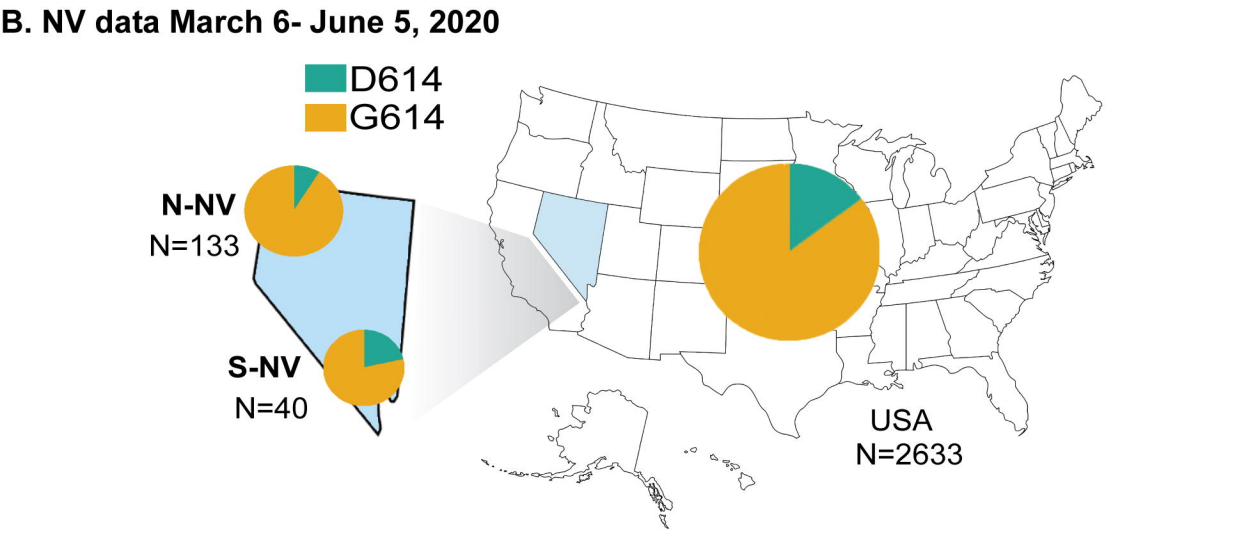
**C.**

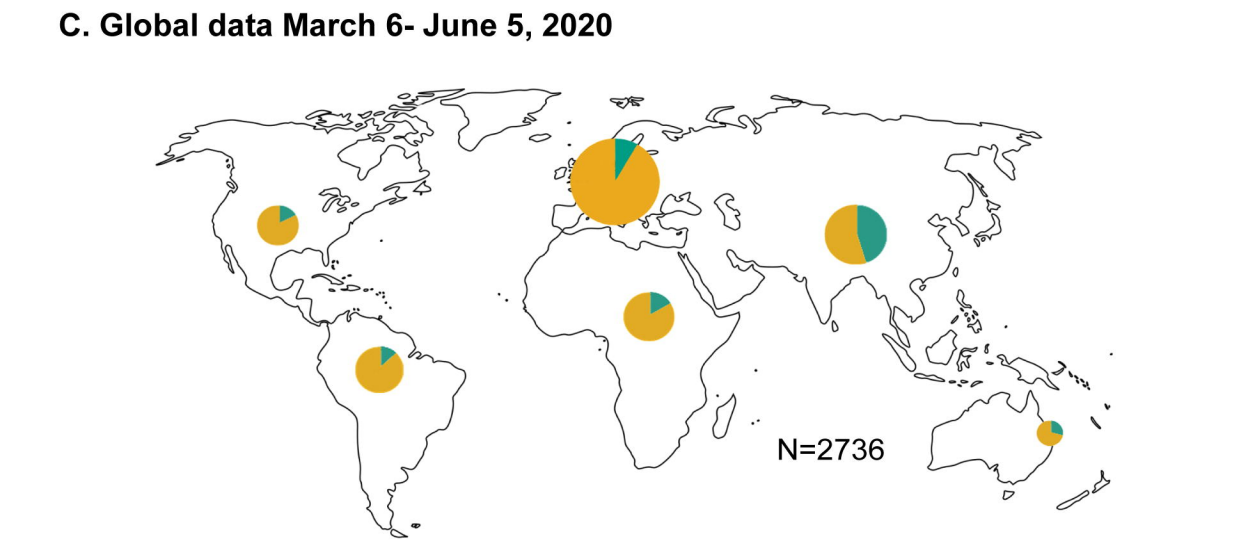


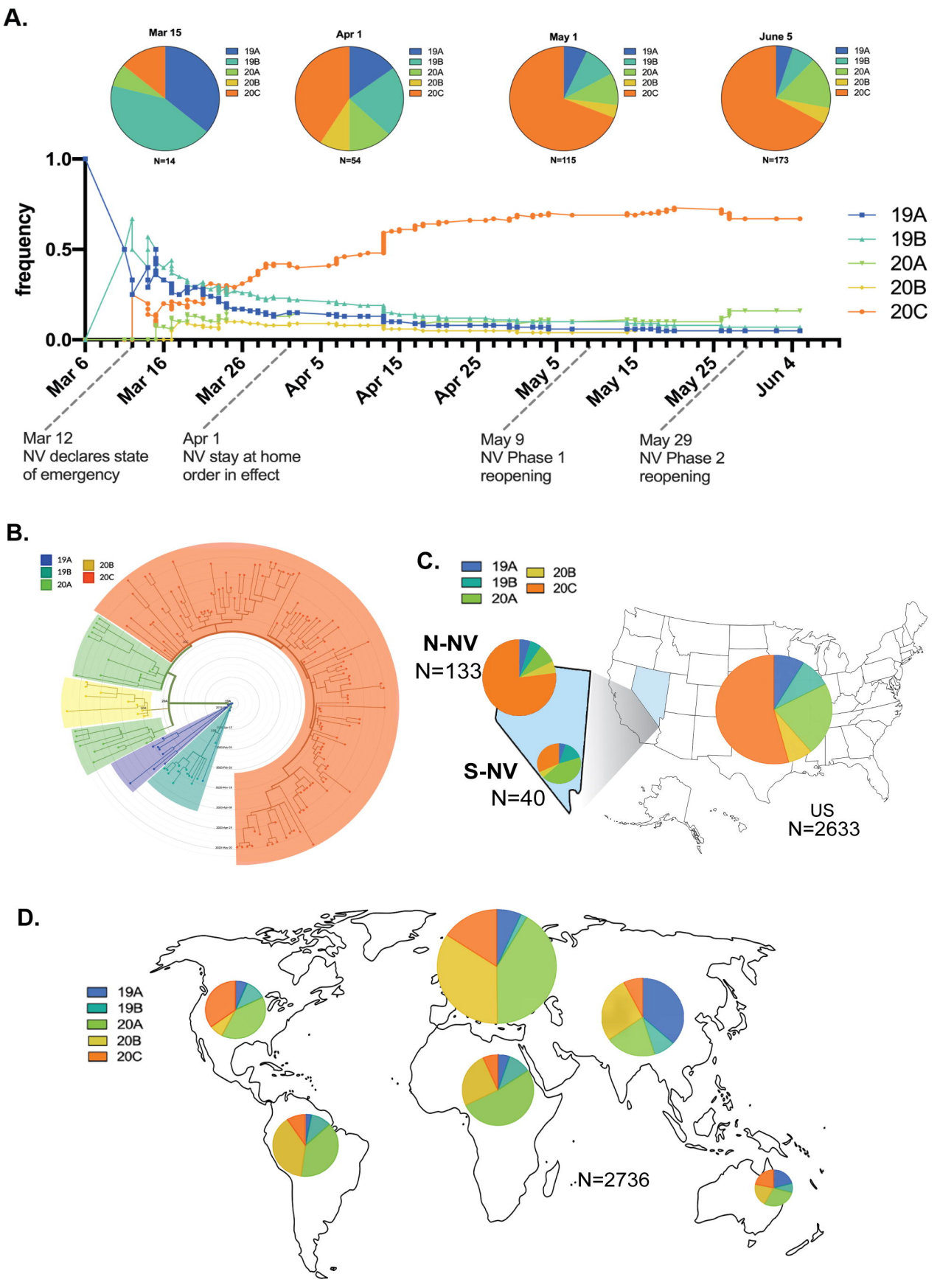

**D.**




A. D614G
Mar 15
Apr 1
May 1
June 5
D614
G614
N=14
N=54
N=115
N=173
frequency
1.0
0.5
0.0
D614
G614
Mar 6
Mar 16
Mar 26
Apr 5
Apr 15
Apr 25
May 5
May 15
May 25
Jun 4
Mar 12
NV declares state of emergency
Apr 1
NV stay at home order in effect
May 9
NV Phase 1 reopening
May 29
NV Phase 2 reopening


B. NV data March 6- June 5, 2020
D614
G614
N-NV
N=133
S-NV
N=40
USA
N=2633


C. Global data March 6- June 5, 2020
N=2736


A.
Mar 15
Apr 1
May 1
June 5
19A
19B
20A
20B
20C
N=14
N=54
N=115
N=173
1.0
0.5
0.0
frequency
Mar 6
Mar 16
Mar 26
Apr 5
Apr 15
Apr 25
May 5
May 15
May 25
Jun 4
Mar 12
NV declares state
of emergency
Apr 1
NV stay at home
order in effect
May 9
NV Phase 1
reopening
May 29
NV Phase 2
reopening
B.
C.
N-NV
N=133
S-NV
N=40
US
N=2633
D.
N=2736