Emergence and Evolution of a Prevalent New SARS-CoV-2 Variant in the United States

Adrian A. Pater[1], Michael S. Bosmeny[2,†], Christopher L. Barkau[2,†], Katy N. Ovington[2], Ramadevi Chilamkurthy[2], Mansi Parasrampuria[2], Seth B. Eddington[2], Abadat O. Yinusa[1], Adam A. White[2], Paige E. Metz[2], Rourke J. Sylvain[2], Madison M. Hebert[1], Scott W. Benzinger[1], Koushik Sinha[3], and Keith T. Gagnon[1,2,*]

[1]Southern Illinois University, Chemistry and Biochemistry, Carbondale, Illinois, USA, 62901.
[2]Southern Illinois University School of Medicine, Biochemistry and Molecular Biology, Carbondale, Illinois, USA, 62901.
[3]Southern Illinois University School of Computing, Carbondale, Illinois, USA, 62901.
[†]These authors contributed equally.
*Corresponding author: ktgagnon@siu.edu

Abstract

Genomic virus surveillance can lead to early identification of new variants and inform proper response during a pandemic. Using this approach, we have identified a new variant of the SARS-CoV-2 virus that emerged in the United States (U.S.) early in the coronavirus disease (COVID-19) pandemic and has become one of the most prevalent U.S variants. This new variant within the B.1.2 lineage referred to here as 20C-US, has not yet spread widely to other countries. The earliest 20C-US genomes can be traced to the southern U.S. in late May of 2020. A major early event was the rapid acquisition of five non-synonymous mutations. The changes carried by 20C-US include mutations to genes involved in virus particle maturation and release, processing of viral proteins, and RNA genome integrity and translation genes, all important for efficient and accurate virus production. In addition, 20C-US has since acquired two new non-synonymous mutations that highlight its ongoing evolution, one of which is a Q677H mutation in the spike protein adjacent to the furin cleavage site. We predict that 20C-US may already be the most dominant variant of SARS-CoV-2 in the U.S. The ongoing evolution of 20C-US, as well as other dominant region-specific variants emerging around the world, should continue to be monitored with genomic, epidemiologic, and experimental studies to understand viral evolution and predict future outcomes of the pandemic.

Introduction

In early 2020 the World Health Organization declared that coronavirus disease 2019 (COVID-19), a potentially fatal respiratory infection caused by SARS-CoV-2, was a pandemic(*1*). The high number of SARS-CoV-2 infections both globally and within the United States has presented the virus with ample space in which to acquire new mutations. It has been suggested that some mutations observed already present a fitness advantage for the virus. Notably, the D614G mutation, observed early in the pandemic, is thought to increase the transmissibility of the virus(*2, 3*). The N501Y mutation of the spike protein (S) has also been implicated in the rapid spread of two new strains in the United Kingdom and South Africa(*4-6*). Of additional concern, it is possible that some mutations may enable immune evasion and reduced vaccine efficacy(*7*).

Restrictions in human population movement during a global pandemic, as well as simultaneous acquisition of multiple mutations, could drive emergence of region-specific variants. This evolutionary paradigm might explain the rise of distinct, novel variants now being observed for SARS-CoV-2 around the world(*5, 6, 8*). Here we report the emergence of an increasingly prevalent SARS-CoV-2 strain that appears to have evolved in and remained mostly confined to the U.S. Based on existing genomic data, we predict that this variant is already the dominant variant of SARS-CoV-2 in the U.S. and may now account for the majority of U.S. COVID-19 cases.

Results

Genomic and phylogenetic and characterization of 20C-US, a prevalent new SARS-CoV-2 variant in the U.S.

In response to anticipated genetic changes occurring in the SARS-CoV-2 virus, we began sequencing viral genomes for genomic epidemiology and surveillance. With sequencing focused on the U.S. upper Midwest in the state of Illinois, we generated full genome sequences from samples taken beginning in March 2020 to present. During phylogenetic reconstruction with our Illinois genome sequences, a particular branch within the 20C clade became noticeably more pronounced (**Figure 1A**). We identified five closely co-occurring signature mutations that appeared synapomorphic to the new clade within 20C. These mutations resulted in amino acid changes of N1653D and R2613C in ORF1b, G172V in ORF3a, and P67S and P199L in the nucleocapsid (N) gene, the latter of which also introduces a stop codon mutation at position Q46 of ORF14 (**Table 1**).

To better understand this new variant at the national level, we randomly subsampled approximately 3.3% of available U.S. genomes (1905 out of 57754 genomes) from the GISAID database and constructed a phylogenetic tree using the Nexstrain pipeline(*9*) (**Figure 1B**). Analysis of each mutation within the tree and visualization of geographic distribution revealed that a substantial fraction of genomes comprised this new variant for most U.S. states (**Figure 1C**). Following the distribution of this variant over time reveals an increase in prevalence of the variant, with significant expansion since the beginning of July 2020 (**Figure 1C**). When a signature mutation for the new variant, ORF1b:N1653D, was visualized on a globally subsampled phylogenetic tree, a subclade of the 20C branch of the B.1.2 lineage corresponding to the new variant became apparent (**Figure 1D**). Geographic visualization from Nov. 1 to Dec. 31, 2020 revealed that close to 50% of all sequenced SARS-CoV-2 genomes from the U.S. are the new variant of interest (**Figure 1E**). Furthermore, this variant has thus far only been reported at very low levels in a handful of other countries, including Mexico, Australia, New Zealand, Singapore, Thailand, Taiwan, Poland and Israel. Therefore, we have termed this new variant 20C-US.

Tracing a timeline and geographic origin of the 20C-US variant

We attempted to establish a timeline and geographic origin for the emergence of 20C-US by tracing the appearance of mutations in sequenced genomes. We selected all SARS-CoV-2 genomes from the global GISAID database that possessed the five hallmark mutations of 20C-US as well as two prerequisite mutations that appear key to the 20C-US lineage based on phylogenetic reconstructions, which are ORF8:S24L and ORF1a:L3352F. The majority of genomes bearing these two mutations were from Minnesota and Louisiana during March and April, 2020. The first signature mutation of 20C-US, ORF3a:G172V, then appears in four genomes from Louisiana and one from Arizona at the beginning of April. Then a series of genome sequences are reported in late May and early June that simultaneously contain the remaining four signature mutations, suggesting either a rapid succession of mutations or an event where most or all were acquired in a single patient. Those genomes primarily originated from Texas samples, with the earliest being from the greater Houston, Texas area on May 20, 2020 (**Supplemental Spreadsheet 1**). The new 20C-US variant then becomes prevalent in SARS-CoV-2 genome sequences across the U.S. over time.

The new 20C-US variant seems to have taken root in the southern region of the U.S. in the late spring and early summer of 2020. However, the earliest genome reported to possess all signature mutations was a sample taken from a 90 years-old female patient in Spain on April 16, 2020. But this genome has an additional nucleotide mutation, C3695T, when compared to the May 20 genome from Texas, as well as the other Texas genomes that appear to branch from there. This creates a discrepancy in identifying the earliest representative genome. While reversions in mutations can occur, they are expected to be quite rare. For the early Spain and Texas genomes, the chance of the same four mutations occurring by coincidence in the same genetic background of SARS-CoV-2 is also quite low. No other genomes were found with these signature mutations outside of the U.S. until June 24, 2020 in Australia. Both the Texas and Spain genomes appear complete with no gaps. Until clarifications on submission and sampling dates or consensus genome construction and base-calling methodologies can be made, it is unclear which genome represents the earliest acquisition of the four novel mutations.

Identification of recent mutations of potential interest in the 20C-US phylogenetic lineage

To further characterize 20C-US, we identified all GISAID samples with the signature mutations of ORF3a:G172V, ORF1b:N1653D, and N:P67S and that also possessed N:P199L or any mutations at position 2613 for ORF1b. We then reconstructed a phylogenetic tree with these 4681 sequences. A clear branch was observed in this new tree that was initiated by two new mutations, a synonymous mutation at the nucleotide level, C14805T, that co-occurs at the same time with a non-synonymous mutation of the ORF1a gene that changes M2606 to I2606 (**Table 1**, **Figure 2A**). The co-occurrence of these two mutations in the 20C-US lineage was first observed in late June of 2020. The first full 20C-US genome reported with this mutation and possessing the expected genotype was from Wisconsin on June 23, 2020 followed by an Illinois case on June 25, 2020. Visualizing the geographic distribution of 20C-US genomes with the ORF1a:M2606I mutation demonstrated that it has achieved high prevalence in the eastern and Midwest regions but has not yet spread widely to the western half of the U.S. (**Figure 2B**). The ORF1a:M2606I mutation currently accounts for 48% of all 20C-US genomes.

Within the new branch defined by ORF1a:M2606I, an additional smaller branch with a single mutation of potential high significance was acquired in mid-August 2020, a Q677H mutation in the spike (S) gene (**Figure 2C**). The earliest 20C-US genomes in the GISAID database having C14805T, ORF1a:M2606I, and S:Q677H originated from Minnesota and Wisconsin on August 17, 2020. Mutation of Q677 to lysine (K), arginine (R), proline (P), and histidine (H) has occurred spontaneously throughout the pandemic in many lineages, including branches of the 20A, 20B, and 20C clades. However, an S:Q677 mutation has never established and expanded until recently in the ORF1a:M2606I lineage, possibly due to epistasis(*10*). The percentage of genomes worldwide outside and inside the 20C-US lineage that have contained any S:Q677 mutation is 0.27% versus 4.77%, respectively, which represents an approximately 18-fold enrichment. Within the ORF1a:M2606I lineage, the S:Q677H mutation represents 10.2% of genomes and a map view illustrates that S:Q677H mutants remain largely localized to the upper Midwest, which includes Minnesota, Wisconsin, and Michigan (**Figure 2D**).

To better understand the emergence of new mutations, we plotted the percentage of 20C-US genomes possessing ORF1a:M2606I versus time for the U.S. and a few states with consistent genomic reporting (**Figure 2E-F**). Interestingly, we observed an apparent slowing or reduction in the percentage of 20C-US variants carrying the ORF1a:M2606I mutation. Performing a similar analysis for S:Q677H but only as a percentage of US-20C genomes that also carry the ORF1a:M2606I mutation suggests that the S:Q677H mutation is not decreasing at a rate similar to the ORF1a:M2606I mutant at this time (**Figure 2F**). It is possible that a lag in genome sequencing or reporting could account for these aberrations, especially when so few genomes containing S:Q677H in the 20C-US lineage are currently available. Indeed, S:Q677H could track more closely with ORF1a:M2606I as more genome sequences becomes available. However, it is also possible that ORF1a:M2606I results in slightly lower fitness compared to its parent 20C-US variant and over time is being out-competed. The S:Q677H may have established in the ORF1a:M2606I lineage as a compensatory mutation that rescues fitness lost in the ORF1a:M2606I genotype. Better resolution of this data through additional sequencing over time should determine the fate and impact of these two novel mutations.

Biochemical and biological implications of 20C-US mutations

Several mutations carried by the 20C-US variant suggest biochemical or functional impacts on SARS-CoV-2 biology. ORF1b:N1653D is unique and specific to the 20C-US lineage. This mutation occurs at residue 138 in the ExoN domain of nsp14, a novel RNA proofreading domain that ensures the integrity of RNA genomes and transcripts(*11*). ExoN inactivation is lethal to SARS-CoV-2(*12*). The mutation of asparagine (N) to aspartate (D) at this position represents conversion from a neutral to a negatively charged residue in a low complexity domain, possibly involved in mediating protein-protein or RNA-protein interactions(*13*) (**Figure 3A**). In addition, nsp14 plays a critical enzymatic role by installing a methyl group on the base of the 5' guanine cap for viral RNA transcripts, which is essential for translation into viral proteins(*11, 12*). Interestingly, the other essential viral factor involved in cap formation on viral transcripts is nsp16, which catalyzes 2'-*O*-methylation of the nucleotide adjacent to the terminal 7-methyl-guanine cap on viral RNA transcripts(*11, 14*). The signature 20C-US mutation ORF1b:R2613C results in a conversion of arginine (R) to cysteine (C) at residue 216 of nsp16, which would be predicted to disrupt hydrogen bonding to an adjacent glutamate and structured water molecule and possibly alter local protein stability (**Figure 3B**). These two rather unique mutations co-occur in 20C-US and could potentially alter genome integrity, mutation rates, transcript integrity, and translation efficiency of viral proteins.

The two largest SARS-CoV-2 viral RNA transcripts, ORF1a and ORF1b, are translated into polyproteins that must be further processed by proteases to release mature, functional viral nsp proteins. Two proteases within the ORF1a gene are responsible for this processing, nsp3 and nsp5(*11*). The parental ORF1a:L3352F mutation carried by 20C-US creates a mutation of leucine (L) to phenylalanine (F) at residue 89. L89 in nsp5 is involved in hydrophobic packing of a domain adjacent to the expected active site. There appears to be room to accommodate a larger phenyl ring. Packing against another juxtaposed F side group in the domain could conceivably improve hydrophobic packing and enhance protein stability(*15*) (**Figure 3C**). The more recently acquired mutation of ORF1a:M2606I results in a mutation to the rather large nsp3 protein within what is predicted to be the C-terminal 3Ecto or Y domain. This domain is implicated in anchoring the replication-transcription complex (RTC) to the endoplasmic reticulum membrane to facilitate interaction of nsp3 with other viral proteins on the cytosolic side(*16*).

ORF3a is a multifunctional protein involved in several aspects of the viral life cycle at the surface of the cell membrane and intracellular membranes(*17*). ORF3a modulates the innate immune response and apoptosis of host cells(*18*). The mutation G172V in ORF3a occurs within a conserved di-acidic Asp-X-Glu domain (171-173) involved in trafficking to the cell membrane(*19, 20*). The introduction of a valine (V) at the more variable 172 position may modulate interactions with viral or cellular factors(*21*). ORF3a plays a role in viral particle maturation and release at the cell membrane and has been proposed to co-mutate with the spike protein(*22*). Substitution of serine to leucine at position 171 of ORF3a is a common mutation of the South African SARS-CoV-2 variant 501Y.V2.

By mid-August, the 20C-US variant had also acquired a Q677H mutation in the spike protein, directly adjacent to the furin cleavage site. The furin cleavage site is a novel motif not observed in SARS-CoV viruses that is proposed to significantly enhance infectivity(*23*). Furin cleavage is a critical priming step essential for efficient entry of SARS-CoV-2 viruses into cells(*23*). A mutation of interest has been the P681H in the spike protein of the novel UK variant 501Y.V1, also due to close proximity to the furin cleavage site(*5*). Q677 and P681 are mutated to histidine in 20C and 501Y.V1, respectively, suggesting a potentially important effect of histidine near the furin cleavage site. The Q677 amino acid resides in a similar region on the spike protein as D614, which is commonly mutated to a G residue(*3*) (**Figure 3D**).

Predicted dominance and spread of the 20C-US variant

The 20C-US clade accounts for a plurality of recently sequenced SARS-CoV-2 genomes in the U.S. Plotting the total percentage of 20C-US genomes in the U.S. against time up to the beginning of Dec. 2020 predicted that 20C-US would account for the majority of SARS-CoV-2 genomes from the U.S. by the end of 2020 (**Figure 4A**). Considering the delay in symptoms and lag in genome sequence reporting (available genome sequences in GISAID for December 2020 remain low), 20C-US may indeed already be the most common variant in the U.S. While widespread, the degree of representation varies. The central and Midwest U.S. regions appear to have a higher fraction of 20C-US genomes while the Northeast and West coast states have a lower fraction (**Figure 1C** and **4A**).

The rise to prevalence of the 20C-US variant, which began in late June and early July, is coincidental with the beginning of the second wave of the COVID-19 pandemic in the U.S. The 20C-US variant has continued to rapidly spread across the U.S. However, Google mobility data is consistent with no major changes in population movement patterns across the U.S. that could account for higher transmission or a perceived advantage for 20C-US over other variants (**Figure 4D**). The population mobility activities “Retail and Recreation,” “Grocery and Pharmacy,” and “Transit Station Use” exhibited little or no change. “Parks Use” declined and “Residential Stay” increased modestly in Nov. 2020. Only a small gradual increase in “Workplace Visit” was observed. Thus, 20C-US is expected to continue spreading across the U.S.

Discussion

We have characterized the emergence and rise of a prevalent SARS-CoV-2 variant that branches from the 20C clade and is highly specific to the continental U.S. It has likely surpassed 50% penetrance to become the most dominant variant in the U.S. It is unclear whether natural selection or genetic drift has driven the rise in prevalence of 20C-US. Nonetheless, its dominance has been achieved during the second and third pandemic waves in the U.S. It is interesting to note that all mutations that occurred during the establishment of the 20C-US lineage are non-synonymous. It is possible that the mutations of the 20C-US variant have conferred a fitness advantage over other variants in circulation in the U.S. population. These include mutations to proteins involved in viral replication, metabolism, and cellular exit. It has more recently also acquired additional novel mutations. Of particular interest is a Q677H mutation in the spike protein near the furin cleavage site that could potentially alter cellular entry of virus particles. Monitoring the fate of this mutation, as well as the potentially linked M2606I mutation of ORF1a, should provide unique insight into molecular evolutionary processes for SARS-CoV-2 and their impact on real-world pandemic outcomes.

The biological effects of the combined 20C-US mutations, as well as the viral characteristics of 20C-US, like fitness, transmissibility, and virulence, remain to be experimentally characterized. Nonetheless, the rise of 20C-US coincides with substantially reduced case fatality rates across the U.S. despite a sharp rise in cases. Many factors contribute to case fatality rates, including accurate case reporting(*24*). While clear evidence is lacking, it is plausible that 20C-US represents a SARS-CoV-2 variant with higher transmissibility but milder illness. Such variants could conceivably generate a fitness advantage for the virus as they are more likely to spread quietly. The flu pandemic of 1918 was extremely deadly at early stages and caused tens of millions of deaths. However, the virus eventually evolved a decreased virulence, and traces of the initial influenza virus from the 1918 pandemic can still be observed today in modern seasonal flu strains (*25*).

Although characterization of distinct novel SARS-CoV-2 variants has been limited up to this point, two driving forces that may affect their rise and prevalence are the simultaneous acquisition of multiple mutations and limited population movement between local regions during the pandemic. The UK variant 501Y.V1 acquired several mutations simultaneously. It has been proposed that this founding event occurred in an immune-compromised patient where evolutionary pressure to evade the immune system was not present (*4*). The South African variant 501Y.V2 also quickly acquired multiple mutations in the spike protein(*6*). Likewise, the 20C-US variant appears to have incorporated five mutations rapidly, including four possibly simultaneously. These events might create a jump in fitness and a temporary disequilibrium in variant competition similar to that seen earlier in the pandemic with the singular D614G mutation(*2, 3*).

The mechanism and rate of SARS-CoV-2 transmission necessitates strict measures that effectively limit population movement(*26*). Since the outbreak of the global COVID-19 pandemic, international travel has become highly restricted. Novel variants that emerge in an isolated region or country may transmit locally among that population and develop distinct genotypes and phenotypes. Thus, it would be expected that regional territories would develop their own distinct SARS-CoV-2 variants over time. When searching for novel emerging variants, focusing on local and regional data may provide an advantage. Our ability to identify the 20C-US variant can be partly attributed to our initial focus on regional, state-level data since the prevalence of the 20C-US variant was more pronounced in the U.S. Midwest.

While this manuscript was in preparation, the Nextstrain group updated their global phylogenetic analysis server for SARS-CoV-2 to begin designating emerging clades. This may enable faster identification of potential variants of interest. We found that the 20C-US variant closely tracks with the newly designated 20G clade, demonstrating that this approach will be extremely valuable for quickly searching for emerging variants. When visualized on a global geographic view, it becomes clear that variants specific to other world regions have emerged and gained prevalence, including the western coast of South America (20D), Europe (20E/EU1), Australia (20F), South Africa (20H/501Y.V2), and the United Kingdom (20I/501Y.V1) (**Figure 4F**). A detailed assessment of the emergence and rise to prevalence should also be undertaken for these variants to better understand viral evolution and inform proper global pandemic response. Unless successful vaccination efforts can be greatly accelerated, we predict the emergence of dominant novel variants in parts of the world that are relatively isolated from other global regions, possibly including Brazil, New Zealand, the African west coast, and Japan.

This study underscores the need for greater genomic surveillance of the SARS-CoV-2 virus, especially at the regional level where novel variants will first emerge. Modern genomic surveillance enables observation of evolution in real-time, prediction of major shifts in viral fitness, and assurance that vaccines are kept current.

Methods

Sequencing of SARS-CoV-2 Samples

RNA extraction was performed using Mag-Bind Viral RNA XPress Kit (Omega Bio-Tek) on samples received from Illinois Department of Public Health (IDPH) in inactivation buffer consisting of 200 µL of sample in Viral Transport Media (VTM) and 240 µL of lysis buffer (239 µL of TNA lysis buffer and 1 µL of carrier RNA per sample). cDNA synthesis from the extracted RNA was performed using ABI High-Capacity cDNA Reverse Transcription Kit, following manufacturer's recommended protocol. SARS-CoV-2 was detected using N2 primers (Integrated DNA Technologies, IDT) that target the N2 region of the nucleocapsid gene. PrimeTime Gene Expression Master Mix (IDT) and 2 µL of cDNA was used to determine the C_t value of each sample. Sequencing was performed with

Oxford Nanopore Technology's MinION platform using ARTIC Network protocol (https://artic.network/ncov-2019) with slight modifications to the protocol. Briefly, 25 µL reactions were performed for each pool using 5 µL of cDNA template and 4 µL of 10 µM primer pool in respective reactions. Finally, 0.5 µL of 10 mM Deoxynucleotide (dNTP) solution mix (New England Biolabs) was added to each reaction and nuclease -free water was used to make up the remaining reaction volume. The PCR was run by combining the annealing and extension steps at 63°C. The thermocycler was set at 98°C for initial denaturation for 30 seconds, followed by 35 cycles of denaturation at 98°C for 15 seconds, annealing and extension at 63°C for 5 min and holding at 4°C indefinitely. From each Multiplex PCR, 5 µL of several reactions from each pool from the batch was used to run on 1% agarose gel. The remaining 20 µL of each pool were pooled and clean-up was performed using an equal volume of AMPure XP beads (Beckman Coulter) and quantified using 1 µL of sample with the Qubit dsDNA HS Assay Kit on a Qubit 2.0 Fluorometer. Following quantification, 60 ng of each sample was end-prepped using 0.75 µL of Ultra II End-Prep enzyme and 1.75 µL of Ultra II End-Prep buffer with a total volume of 15 µL. The samples were then barcoded using the 96 native barcoding kit (Oxford Nanopore Technology) and further processed using the ARTIC Network protocol (https://artic.network/ncov-2019).

Basecalling on completed sequencing runs was performed using Guppy high accuracy model and further demultiplexed using Guppy barcoder using strict parameters requiring barcodes to be present at both ends. Reads were filtered based on length quality (400-700 bp) and mapped by aligning to the MN908947.3 reference genome using minimap2. Medaka was used to create the consensus sequence and call variants for each of the samples. The variants identified were fed into longshot to produce a set of high-confidence variants.

Dataset

SARS-CoV-2 genomes sequences used in this paper that were not generated by us were acquired from the GISAID Initiative (https://www.gisaid.org/). This dataset was updated on 2021-01-04. Individual sequences were compiled by GISAID from contributions from individual labs across the world. Their information is compiled in the supplementary section. Sequences generated by our laboratory, currently in the process of being submitted to GISAID, were added to this dataset. Our samples were all within the state of Illinois.

All sequences were evaluated through the command-line version of Nextstrain's NextClade software (https://www.npmjs.com/package/@neherlab/nextclade)(*9*) to derive a list of all nucleotide and amino-acid substitutions. 20C-US samples were then filtered from the larger population using a criterion of necessary amino-acid mutations. Because not all sequences are complete (contain gaps) there is the possibility of a sequence being within the 20C-US population but not reporting one of the required mutations. The formula for filtering therefore used a flexible criterion.

Phylogenetic Trees and Maps

Once appropriately filtered from the greater genomic sample pool, sample FASTA files were used for phylogenetic inference. Briefly, the Augur pipeline utilized by Nextstrain filters sequences for metadata values and N content before aligning them using MAFFT(*27*). As a reference, the Nextstrain nCoV toolkit aligns SARS-CoV-2 sequences to the SARS-CoV-2 isolate Wuhan-Hu-1 complete genome (GenBank: NC_045512.2) and by default removes any nucleotide insertions relative to this sequence. The pipeline conducts a maximum likelihood phylogenetic analysis using the general time reversible model allowing for invariant sites and a gamma distribution (GTR+I+G) in IQ-TREE(*28*). Augur further refines and annotates the tree in various ways. The resulting files are combined and used by Auspice to visualize phylogenetic relationships and geographic distributions of SARS-CoV-2 across time. For the 20C-US subset, all samples were used. For the model of all SARS-CoV-2 variants within the United States of America, a subset of the data was generated based on Nextstrain's standard filtering criteria. Briefly, lower-quality samples (those with large numbers of gaps or lacking sufficient metadata) are filtered out, then the dataset is sorted based on the month it was acquired as well as the U.S. state it was acquired in. From each of these subsets, Nextstrain attempts to randomly pick an equal number of samples. All these samples are then recombined and processed together.

Acknowledgements

We thank the Illinois Department of Public Health for access to patient samples. This work was funded by discretionary funds from the Southern Illinois University (SIU) School of Medicine Dean's Office and the Office of the Vice Chancellor for Research at SIU Carbondale. It was also funded by the Walder Foundation through a collaborative agreement with the Open Commons Consortium as part of the Chicago CAN initiative.

Figure Legends

Figure 1. Phylogenetic reconstruction and geographic visualization of SARS-CoV-2 variant 20C-US in Illinois, the United States, and globally. Phylogenetic reconstruction of SARS-CoV2 using (**A**) 352 genomes sequenced from Illinois from March 2020 through December 2020 and (**B**) genome sequences randomly subsampled from the U.S. at ~3.3% of all U.S. genomes in the GISAID database (as of Jan. 4, 2021). (**C**) Geographic visualization of the fraction of 20C-US variant genomes in the U.S. during the 2-month intervals of May 1 to Jun. 30, Jul. 1 to Aug. 31, Sep. 1 to Oct. 31, and Nov. 1 to Dec. 31, 2020. (**D**) Phylogenetic tree reconstruction of SARS-CoV-2 during the 2-month interval of Nov. 1 to Dec. 31, 2020 using randomly subsampled global genomes (3819) from GISAID (as of Jan. 4, 2021). (**E**) Global map view of the fraction of 20C-US variant genomes from around the world. For all trees and maps, the signature ORF1b:N1653D was used to highlight the 20C-US variant compared to all other SARS-CoV-2 variants.

Figure 2. Characterization of recent mutations of the SARS-CoV-2 variant 20C-US. (**A**-**B**) Phylogenetic reconstruction and geographic visualization (during the 2-month interval of Nov. 1 to Dec. 31, 2020) of all SARS-CoV-2 variant 20C-US genomes (4683) in the GISAID database (as of Jan. 4, 2021). The ORF1a:M2606I mutant genotype is colored to distinguish it from all other genetic variants within the 20C-US tree. (**C**-**D**) Phylogenetic reconstruction and geographic visualization (during the 2-month interval of Nov. 1 to Dec. 31, 2020) for all SARS-CoV-2 variant 20C-US genomes (4683) in the GISAID database (as of Jan. 4, 2021). The S:Q677H mutant genotype is distinguished from all genetic other variants within the 20C-US tree. (**E**) Plot depicting the rise in percentage of 20C-US, 20C-US possessing ORF1a:M2606I, and 20C-US possessing S:Q677H genomes for all U.S. SARS-CoV-2 genomes in the GISAID database during the indicated months (as of Jan. 4, 2021). (**F**) Percentage of 20C-US genomes that possess the ORF1a:M2606I mutation or the percentage of ORF1a:M2606I mutants that also possess the S:Q677H mutation versus time.

Figure 3. Location of select 20C-US mutations on the three-dimensional structure of their respective proteins. (**A**) Structure of nsp14 is shown in orange cartoon and the N138 amino acid in cyan spheres (PDB 5C8S). (**B**) Structure of nsp16 is shown in green cartoon and the R216 amino acid in purple spheres (PDB 6YZ1). The neighboring glutamate residue and the structured water molecule that it is predicted to hydrogen bond with are shown as green sticks and a red dot sphere, respectively. An S-adenosyl-methionine analog is shown as a space filling model in the nsp16 active site. (**C**) The structure of nsp5 is shown in yellow cartoon and the L89 amino acid in blue spheres (PDB 7KHP). A nearby phenylalanine available for increased hydrophobic packing is shown as orange sticks. (**D**) Structure of the trimeric spike protein with one monomer shown as a blue cartoon and the other two shown as gray cartoons (PDB 7JJI). The position of the Q77 amino acid is shown in spheres. Additional important mutations of the spike protein described previously are also shown: the 69-70

amino acid deletion in cyan spheres, the N501 amino acid in red spheres, the E484 amino acid in orange spheres, and the D614 amino acid in magenta spheres.

Figure 4. Timeline for rise to prevalence of the SARS-CoV-2 variant 20C-US, case fatality rates, and google mobility and 20C-US viral load approximations. (**A**) Plot depicting the rise in percentage of 20C-US genomes for all U.S. SARS-CoV-2 genomes in the GISAID database during the indicated states and indicated months up to Nov. 30. A curve fit to the average of all U.S. states projecting the continued rise of 20C-US prevalence in the genome database. (**B**) Google mobility plots for diverse activities for the entire U.S. Data is presented as percent change from pre-pandemic baseline. (**C**) Global map view of using Nextstrain's designated SARS-CoV-2 clades from Aug. 1 to Dec. 31, 2020 using a phylogenetic tree reconstructed from 3819 randomly subsampled genomes in the GISAID database.

References

1. D. Cucinotta, M. Vanelli, WHO declares COVID-19 a pandemic. *Acta Bio Medica: Atenei Parmensis* **91**, 157 (2020).
2. J. A. Plante *et al.*, Spike mutation D614G alters SARS-CoV-2 fitness. *Nature*, 1 (2020).
3. B. Korber *et al.*, Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus. *Cell* **182**, 812 (Aug 20, 2020).
4. K. Leung, M. H. Shum, G. M. Leung, T. T. Lam, J. T. Wu, Early transmissibility assessment of the N501Y mutant strains of SARS-CoV-2 in the United Kingdom, October to November 2020. *Eurosurveillance* **26**, 2002106 (2021).
5. N. G. Davies *et al.*, Estimated transmissibility and severity of novel SARS-CoV-2 Variant of Concern 202012/01 in England. *medRxiv*, 2020.12.24.20248822 (2020).
6. H. Tegally *et al.*, Emergence and rapid spread of a new severe acute respiratory syndrome-related coronavirus 2 (SARS-CoV-2) lineage with multiple spike mutations in South Africa. *medRxiv*, 2020.12.21.20248640 (2020).
7. E. Dumonteil, C. Herrera, Polymorphism and selection pressure of SARS-CoV-2 vaccine and diagnostic antigens: implications for immune evasion and serologic diagnostic performance. *Pathogens* **9**, 584 (2020).
8. D. Mercatelli, F. M. Giorgi, Geographic and Genomic Distribution of SARS-CoV-2 Mutations. (2020).
9. J. Hadfield *et al.*, Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* **34**, 4121 (2018).
10. R. Sanjuan, J. M. Cuevas, A. Moya, S. F. Elena, Epistasis and the adaptability of an RNA virus. *Genetics* **170**, 1001 (Jul, 2005).
11. P. V'Kovski, A. Kratzel, S. Steiner, H. Stalder, V. Thiel, Coronavirus biology and replication: implications for SARS-CoV-2. *Nature reviews. Microbiology*, (Oct 28, 2020).
12. N. S. Ogando *et al.*, The Enzymatic Activity of the nsp14 Exoribonuclease Is Critical for Replication of MERS-CoV and SARS-CoV-2. *Journal of virology* **94**, (Nov 9, 2020).
13. Y. Ma *et al.*, Structural basis and functional analysis of the SARS coronavirus nsp14-nsp10 complex. *Proceedings of the National Academy of Sciences of the United States of America* **112**, 9436 (Jul 28, 2015).
14. E. Decroly *et al.*, Coronavirus nonstructural protein 16 is a cap-0 binding enzyme possessing (nucleoside-2'O)-methyltransferase activity. *Journal of virology* **82**, 8071 (Aug, 2008).
15. J. Lee *et al.*, Crystallographic structure of wild-type SARS-CoV-2 main protease acyl-enzyme intermediate with physiological C-terminal autoprocessing site. *Nature communications* **11**, 5877 (Nov 18, 2020).
16. J. Lei, Y. Kusov, R. Hilgenfeld, Nsp3 of coronaviruses: Structures and functions of a large multi-domain protein. *Antiviral research* **149**, 58 (Jan, 2018).

17. R. Minakshi *et al.*, The SARS Coronavirus 3a protein causes endoplasmic reticulum stress and induces ligand-independent downregulation of the type 1 interferon receptor. *PloS one* **4**, e8342 (Dec 17, 2009).
18. C. Diemer *et al.*, Cell type-specific cleavage of nucleocapsid protein by effector caspases during SARS coronavirus infection. *Journal of molecular biology* **376**, 23 (Feb 8, 2008).
19. N. Nishimura, W. E. Balch, A di-acidic signal required for selective export from the endoplasmic reticulum. *Science* **277**, 556 (Jul 25, 1997).
20. E. Issa, G. Merhi, B. Panossian, T. Salloum, S. Tokajian, SARS-CoV-2 and ORF3a: Nonsynonymous Mutations, Functional Domains, and Viral Pathogenesis. *mSystems* **5**, (May 5, 2020).
21. R. McBride, B. C. Fielding, The role of severe acute respiratory syndrome (SARS)-coronavirus accessory proteins in virus pathogenesis. *Viruses* **4**, 2902 (Nov 7, 2012).
22. S. S. Hassan, P. P. Choudhury, P. Basu, S. S. Jana, Molecular conservation and differential mutation on ORF3a gene in Indian SARS-CoV2 genomes. *Genomics* **112**, 3226 (Sep, 2020).
23. M. Hoffmann, H. Kleine-Weber, S. Pohlmann, A Multibasic Cleavage Site in the Spike Protein of SARS-CoV-2 Is Essential for Infection of Human Lung Cells. *Molecular cell* **78**, 779 (May 21, 2020).
24. F. J. Angulo, L. Finelli, D. L. Swerdlow, Estimation of US SARS-CoV-2 Infections, Symptomatic Infections, Hospitalizations, and Deaths Using Seroprevalence Surveys. *JAMA network open* **4**, e2033706 (Jan 4, 2021).
25. J. K. Taubenberger, D. M. Morens, 1918 Influenza: the mother of all pandemics. *Revista Biomedica* **17**, 69 (2006).
26. J. A. Lewnard, N. C. Lo, Scientific and ethical basis for social-distancing interventions against COVID-19. *The Lancet. Infectious diseases* **20**, 631 (2020).
27. K. Katoh, K.-i. Kuma, H. Toh, T. Miyata, MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic acids research* **33**, 511 (2005).
28. L.-T. Nguyen, H. A. Schmidt, A. Von Haeseler, B. Q. Minh, IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular biology and evolution* **32**, 268 (2015).

Figure 1

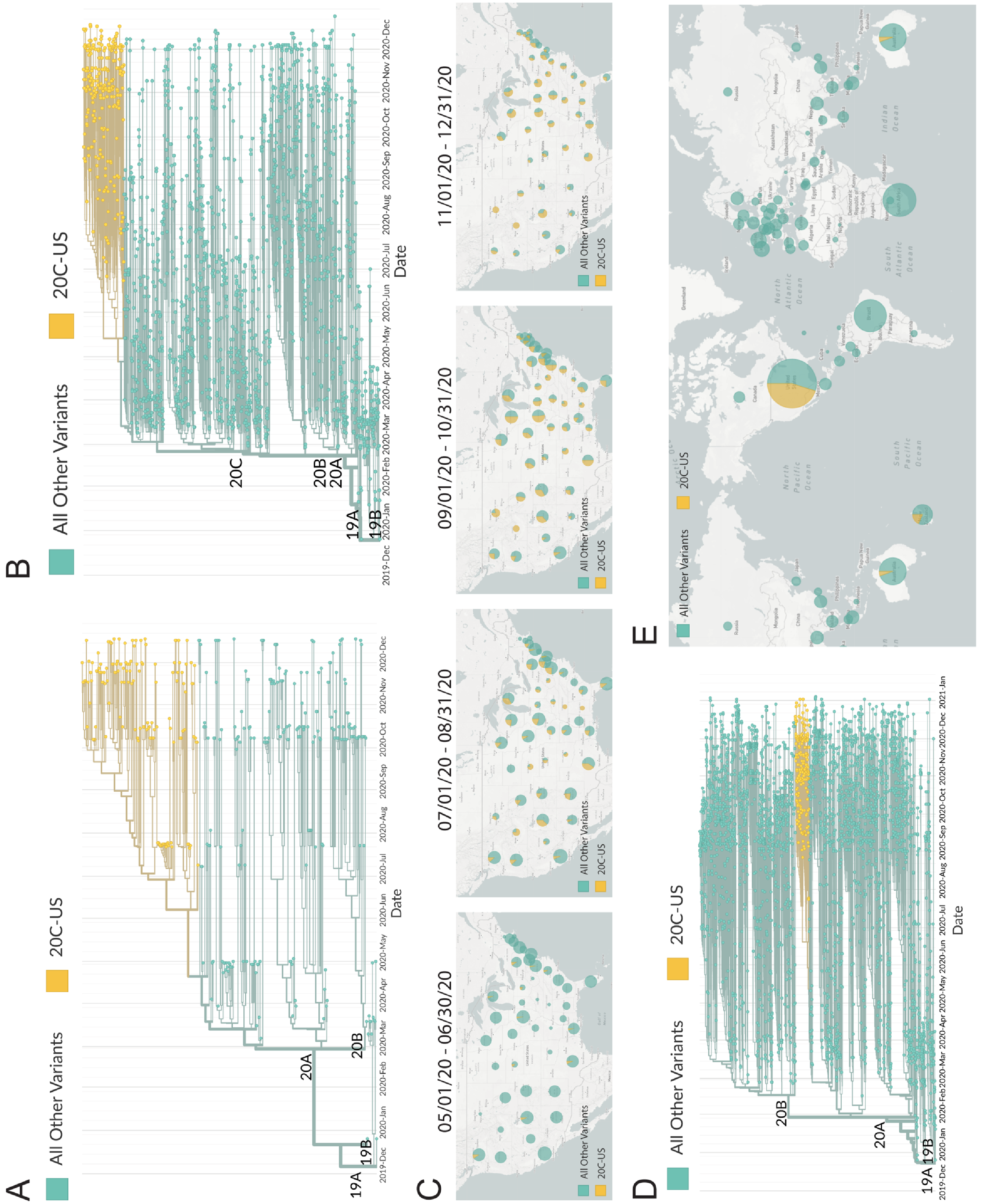
A
All Other Variants
20C-US
19A
19B
20A
20B
2019-Dec
2020-Jan
2020-Feb
2020-Mar
2020-Apr
2020-May
2020-Jun
2020-Jul
2020-Aug
2020-Sep
2020-Oct
2020-Nov
2020-Dec
Date
B
All Other Variants
20C-US
19A
19B
20A
20B
20C
2019-Dec
2020-Jan
2020-Feb
2020-Mar
2020-Apr
2020-May
2020-Jun
2020-Jul
2020-Aug
2020-Sep
2020-Oct
2020-Nov
2020-Dec
Date
C
05/01/20 - 06/30/20
07/01/20 - 08/31/20
09/01/20 - 10/31/20
11/01/20 - 12/31/20
All Other Variants
20C-US
D
All Other Variants
20C-US
19A
19B
20A
20B
2019-Dec
2020-Jan
2020-Feb
2020-Mar
2020-Apr
2020-May
2020-Jun
2020-Jul
2020-Aug
2020-Sep
2020-Oct
2020-Nov
2020-Dec
2021-Jan
Date
E
All Other Variants
20C-US
North Pacific Ocean
South Pacific Ocean
North Atlantic Ocean
South Atlantic Ocean
Indian Ocean
Greenland
Canada
United States
Brazil
Russia
China
Mongolia
India
Australia
South Africa

Figure 2

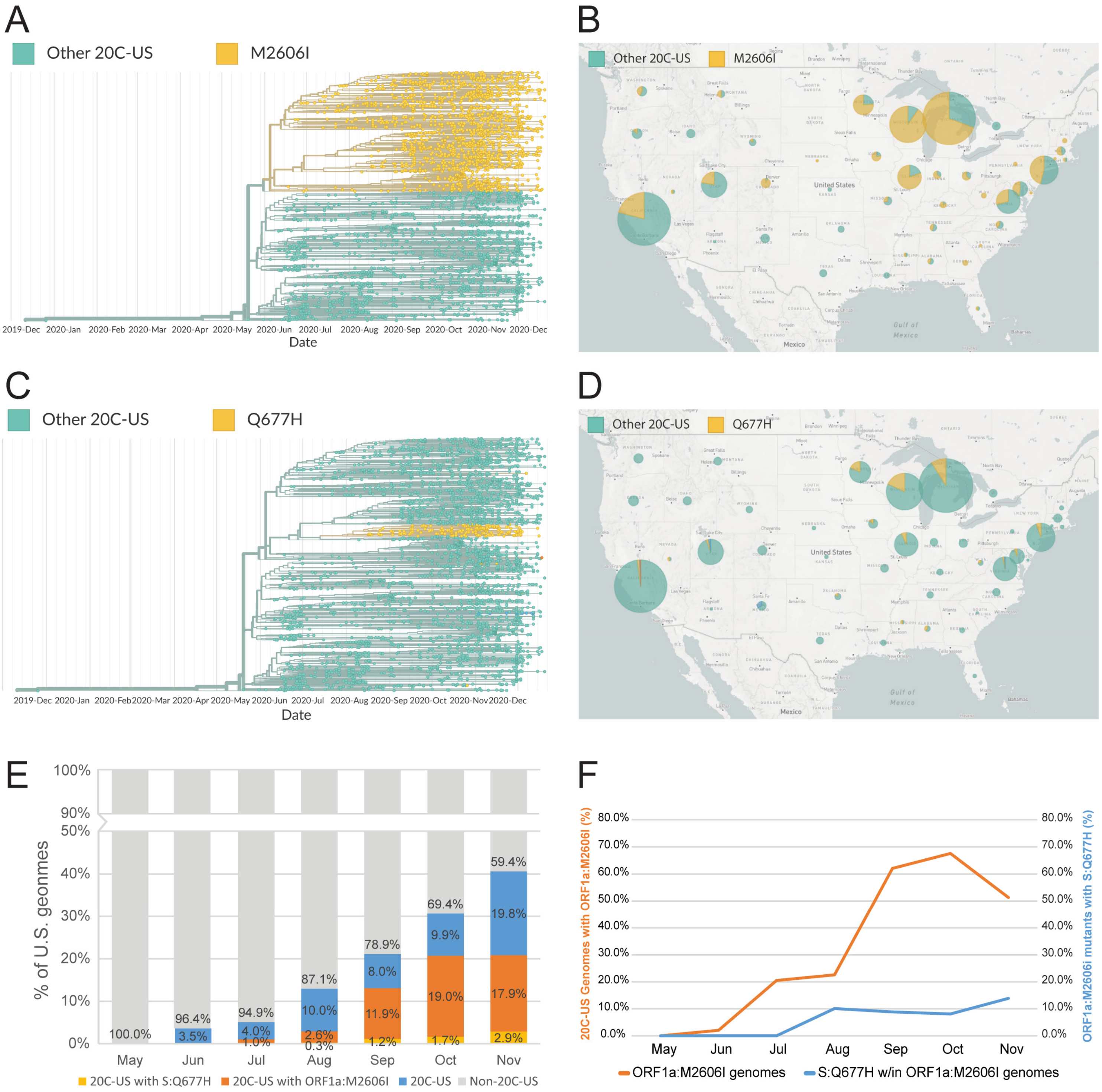
A
🟩 Other 20C-US
🟨 M2606I
2019-Dec
2020-Jan
2020-Feb
2020-Mar
2020-Apr
2020-May
2020-Jun
2020-Jul
2020-Aug
2020-Sep
2020-Oct
2020-Nov
2020-Dec
Date
B
🟩 Other 20C-US
🟨 M2606I
United States
Mexico
Gulf of Mexico
C
🟩 Other 20C-US
🟨 Q677H
2019-Dec
2020-Jan
2020-Feb
2020-Mar
2020-Apr
2020-May
2020-Jun
2020-Jul
2020-Aug
2020-Sep
2020-Oct
2020-Nov
2020-Dec
Date
D
🟩 Other 20C-US
🟨 Q677H
United States
Mexico
Gulf of Mexico
E
% of U.S. geonmes
100%
90%
50%
40%
30%
20%
10%
0%
100.0%
96.4%
3.5%
94.9%
4.0%
1.0%
87.1%
10.0%
2.6%
0.3%
78.9%
8.0%
11.9%
1.2%
69.4%
9.9%
19.0%
1.7%
59.4%
19.8%
17.9%
2.9%
May
Jun
Jul
Aug
Sep
Oct
Nov
20C-US with S:Q677H
20C-US with ORF1a:M2606I
20C-US
Non-20C-US
F
20C-US Genomes with ORF1a:M2606I (%)
ORF1a:M2606i mutants with S:Q677H (%)
80.0%
70.0%
60.0%
50.0%
40.0%
30.0%
20.0%
10.0%
0.0%
May
Jun
Jul
Aug
Sep
Oct
Nov
ORF1a:M2606I genomes
S:Q677H w/in ORF1a:M2606I genomes

Figure 3

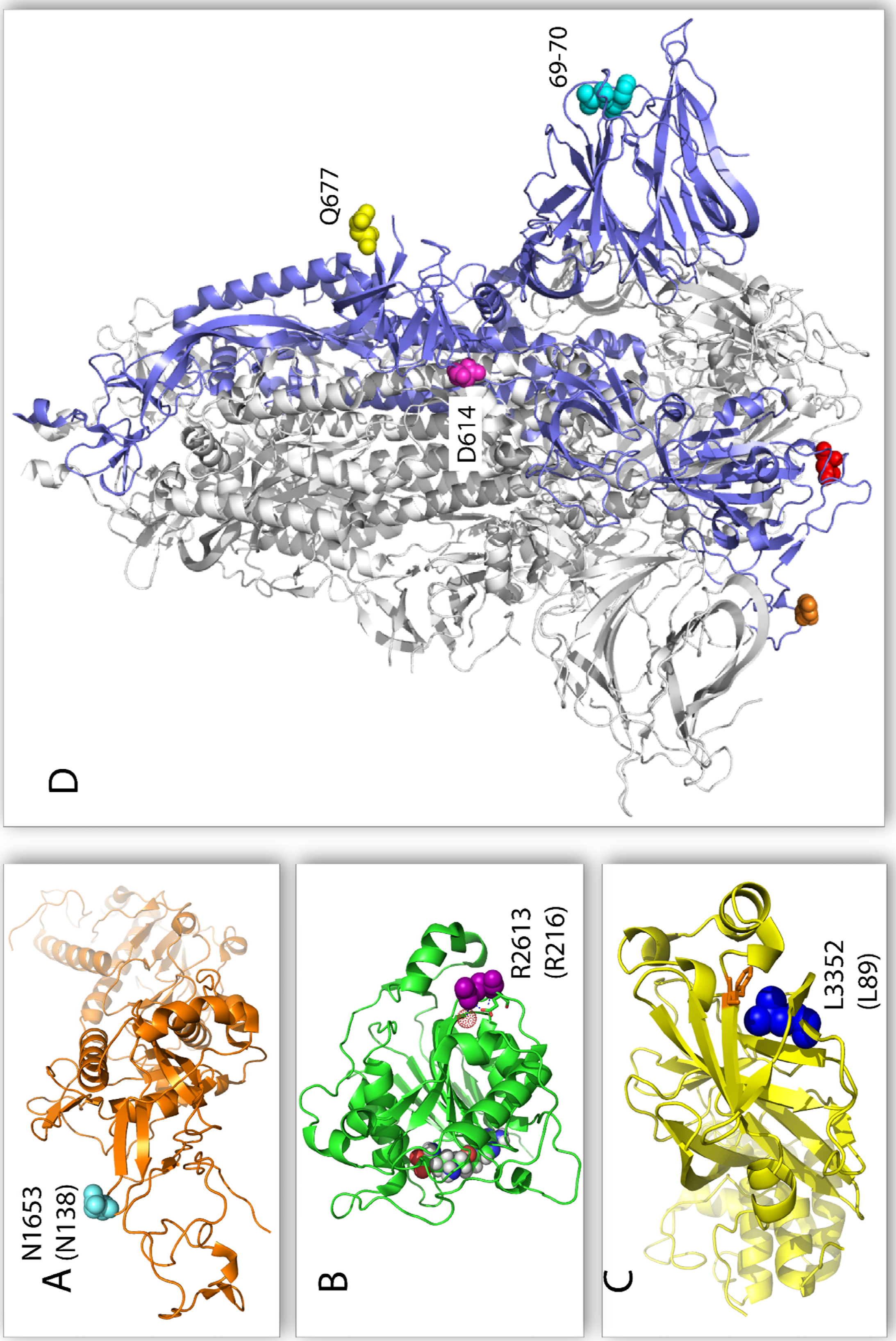

Figure 4

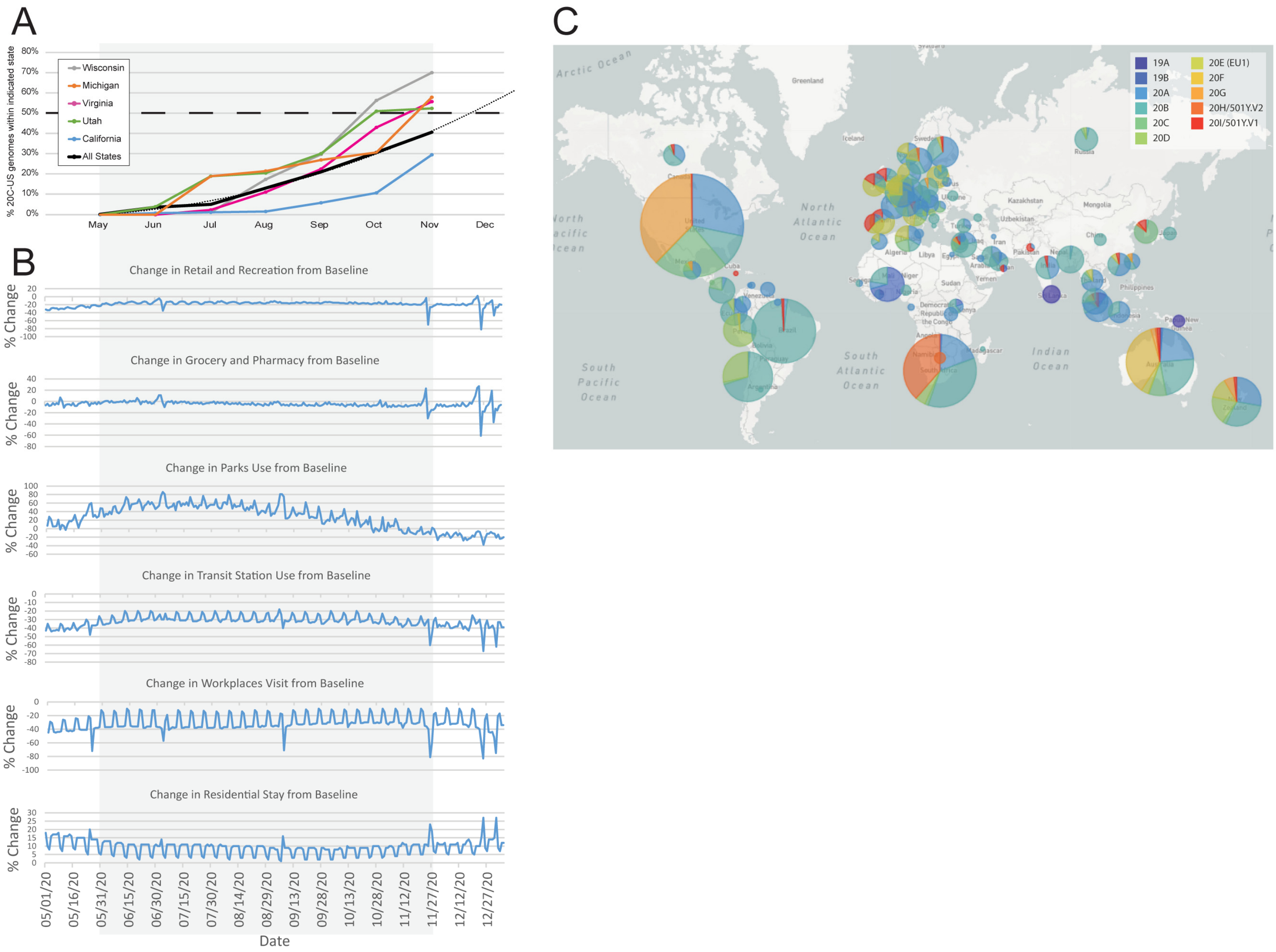
A
% 20C-US genomes within indicated state
80%
70%
60%
50%
40%
30%
20%
10%
0%
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Wisconsin
Michigan
Virginia
Utah
California
All States
B
Change in Retail and Recreation from Baseline
Change in Grocery and Pharmacy from Baseline
Change in Parks Use from Baseline
Change in Transit Station Use from Baseline
Change in Workplaces Visit from Baseline
Change in Residential Stay from Baseline
% Change
05/01/20
05/16/20
05/31/20
06/15/20
06/30/20
07/15/20
07/30/20
08/14/20
08/29/20
09/13/20
09/28/20
10/13/20
10/28/20
11/12/20
11/27/20
12/12/20
12/27/20
Date
C
19A
19B
20A
20B
20C
20D
20E (EU1)
20F
20G
20H/501Y.V2
20I/501Y.V1
Arctic Ocean
North Atlantic Ocean
South Atlantic Ocean
South Pacific Ocean
Indian Ocean
North Pacific Ocean

Table 1: Key 20C-US Mutations.

Signature Mutation	Mutation Name	Nucleotide Change	Amino Acid Change	Protein Name	Protein Function
Yes	ORF1b:N1653D	A18424G	N1653D	ORF1b, nsp14	RNA proofreading, RNA capping, methyltransferase
Yes	ORF1b:R2613C	C21304T	R2613C	ORF1b, nsp16	RNA capping, methyltransferase
Yes	ORF3a:G172V	G25907T	G172V	ORF3a	cellular immune modulation, virus maturation and exit
Yes	N:P67S	C28472T	P67S	nucleocapsid	viral particle structure, RNA binding
Yes	N:P199L	C28869T	P199L	nucleocapsid	viral particle structure, RNA binding
	ORF14:Q64*	C28869T	STOP	ORF14	unknown
	ORF1a:M2606I	G8083A	M2606I	ORF1a, nsp3	protease, viral polyprotein maturation
	S:Q677H	G23593T	Q677H	spike	cell membrane binding, viral entry