Sequence Analysis of 20,453 Severe Acute Respiratory Syndrome Coronavirus 2 Genomes from the Houston Metropolitan Area Identifies the Emergence and Widespread Distribution of Multiple Isolates of All Major Variants of Concern


From the Center for Molecular and Translational Human Infectious Diseases Research, Department of Pathology and Genomic Medicine, Houston Methodist Research Institute and Houston Methodist Hospital, Houston, Texas; the Departments of Pathology and Laboratory Medicine, and Microbiology and Immunology, Weill Cornell Medical College, New York, New York; the Consortium for Advanced Science and Engineering, 22 University of Chicago, Chicago, Illinois; the Computing, Environment and Life Sciences, Argonne National Laboratory, Lemont, Illinois; the Department of Molecular Biosciences and Institute of Molecular Biosciences, The University of Texas at Austin, Austin, Texas; and the CCDC Army Research Laboratory-South, University of Texas, Austin, Texas

Since the beginning of the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) pandemic, there has been international concern about the emergence of virus variants with mutations that increase transmissibility, enhance escape from the human immune response, or otherwise alter biologically important phenotypes. In late 2020, several variants of concern emerged globally, including the UK variant (B.1.1.7), the South Africa variant (B.1.351), Brazil variants (P.1 and P.2), and two related California variants of interest (B.1.429 and B.1.427). These variants are believed to have enhanced transmissibility. For the South Africa and Brazil variants, there is evidence that mutations in spike protein permit it to escape from some vaccines and therapeutic monoclonal antibodies. On the basis of our extensive genome sequencing program involving 20,453 coronavirus disease 2019 patient samples collected from March 2020 to February 2021, we report identification of all six of these SARS-CoV-2 variants among Houston Methodist Hospital (Houston, TX) patients residing in the greater metropolitan area. Although these variants are currently at relatively low frequency (aggregate of 1.1%) in the population, they are geographically widespread. Houston is the first city in the United States in which active circulation of all six current variants of concern has been documented by genome sequencing. As vaccine deployment accelerates, increased genomic surveillance of SARS-CoV-2 is essential to understanding the presence, frequency, and medical impact of consequential variants and their patterns and trajectory of dissemination. (Am J Pathol 2021, 191: 983–992; https://doi.org/10.1016/j.ajpath.2021.03.004)

The severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is the causative agent of coronavirus disease 2019 (COVID-19). Since first being identified in December 2019, the virus has spread globally and is responsible for massive human morbidity and mortality worldwide (https://...
Mutation (E484K) in the spike protein that decreases efficiency of infection.

These variants also are of concern because they contain a N501Y mutation in the spike protein that is associated with stronger binding to the angiotensin-converting enzyme 2 receptor, possibly contributing to increased transmissibility.

The Houston, TX, metropolitan area is the fifth largest and most ethnically diverse city in the United States, with a population of approximately 7 million. The 2400-bed Houston Methodist health system has eight hospitals and cares for a large, multietnic, and geographically and socioeconomically diverse patient population throughout greater Houston. The eight Houston Methodist hospitals have a single central molecular diagnostic laboratory, which means that all RT-PCR specimens can readily be identified, banked, and subjected to further study as needed. In addition, the Department of Pathology and Genomic Medicine has a long-standing record of integrating genome sequencing efforts into clinical care and research, especially related to microbial pathogens infecting our patients.

Before the SARS-CoV-2 virus arrived in Houston, an integrated strategy was planned to confront and mitigate this microbial threat to our patients. In addition to rapidly validating an RT-PCR test for the virus, a plan was instituted to sequence the genome of every positive specimen from patients within the Houston Methodist system, with the goal of understanding pathogen spread in our community and identifying biologically important mutant viruses. Detailed population genomics of the first three waves of SARS-CoV-2 infections in the Houston metropolitan region were previously described. Positive SARS-CoV-2 specimens continue to be sequenced with the goal of monitoring for variants of concern and genome mutations that may be associated with patient outcome or therapeutic failure.

This report describes the identification of multiple isolates of important SARS-CoV-2 variants, including the UK B.1.1.7, South Africa B.1.351, Brazil P.1, and California B.1.429 and B.1.427 variants in Houston patient specimens collected from December 2020 through mid-February 2021. These findings represent the first detection of the South Africa and Brazil variants in Texas and the second time UK variants have been identified in Houston. Greater Houston is the first metropolis in the United States documented to have all of these important and concerning variants circulating among its residents. Our discoveries further illustrate the need for increased population genomic and epidemiologic efforts to identify and help track
dissemination of these variants, monitor development of new variants, and assess the relationship between variants and COVID-19 disease outcomes.

Materials and Methods

Patient Specimens

All specimens were obtained from individuals who were registered patients at Houston Methodist hospitals, associated facilities (eg, urgent care centers), or institutions in the greater Houston metropolitan region that use the laboratory services. Virtually all individuals had signs or symptoms consistent with COVID-19. This work was approved by the Houston Methodist Research Institute Institutional Review Board (IRB1010-0199).

SARS-CoV-2 Molecular Diagnostic Testing

Specimens obtained from symptomatic patients with a high degree of suspicion for COVID-19 were tested in the Molecular Diagnostics Laboratory at Houston Methodist Hospital using assays granted Emergency Use Authorization from the US Food and Drug Administration (https://www.fda.gov/medical-devices/emergency-situations-medical -devices/faqs-diagnostic-testing-sars-cov-2#offeringtests, last accessed February 18, 2021). Multiple molecular testing platforms were used, including the COVID-19 test or RP2.1 test with BioFire Film Array instruments (Biofire Diagnostics, Salt Lake City, UT), the Xpert Xpress SARS-CoV-2 test using Cepheid GeneXpert Infinity or Cepheid GeneXpert Xpress IV instruments (Cepheid, Sunnyvale, CA), the SARS-CoV-2 Assay using the Hologic Panther instrument, the Aptima SARS-CoV-2 Assay using the Hologic Panther Fusion system (Hologic, Marlborough, MA), and the SARS-CoV-2 assay using Abbott Alinity m instruments (Abbott Laboratories, Abbott Park, IL). All assays were performed according to the manufacturer’s instructions. Testing was performed on material obtained from nasopharyngeal, oropharyngeal, or nasal swabs immersed in universal transport media, bronchoalveolar lavage fluid, or sputum treated with dithiothreitol. To standardize specimen collection, an instructional video was generated for Houston Methodist health care workers (https://vimeo.com/396996468/2228335d56, last accessed March 9, 2021).

SARS-CoV-2 Genome Sequencing

Libraries for whole virus genome sequencing were prepared according to version 3 of the ARTIC nCoV-2019 sequencing protocol (https://artic.network/n cov-2019, last accessed March 9, 2021). Long reads were generated with the LSK-109 sequencing kit, 24 native barcodes (NBD104 and NBD114 kits), and a GridION instrument (Oxford Nanopore, Oxford, UK). Short sequence reads were generated with either a NextSeq 550 or NovaSeq 6000 instrument (Illumina, San Diego, CA).

SARS-CoV-2 Genome Sequence Analysis

Viral genomes were assembled with the BV-BRC SARS-Cov-2 assembly service (https://www.bv-brc.org/app/Comprehensive SARS2Analysis, last accessed March 9, 2021).2 The One Codex SARS-CoV-2 variant calling and consensus assembly pipeline was chosen for assembling all sequences (https://github.com/onecodex/sars-cov-2.git, last accessed February 18, 2021) using default parameters and a minimum read depth of 3. Briefly, the pipeline uses seqtk version 1.3-r116 for sequence trimming (https://github.com/lh3/seqtk.git, last accessed February 18, 2021); minimap version 2.1 for aligning reads against reference genome Wuhan-Hu-1 (NC_045512.2); samtools version 1.11 for sequence and file manipulation4; and iVar version 1.2.2 for primer trimming and variant calling.45 All COVID-19 genomes were submitted to GISAID (www.gisaid.org). Accession numbers are available in Supplemental Appendixes S1–S3.

Geospatial Analysis

The patient home address zip codes were used to visualize the geospatial distribution of spread for each variant of concern. Figures were generated using Tableau version 2020.3.4 (Tableau Software, Mountain View, CA).

Results

So far, 20,453 complete genomes of SARS-CoV-2 from clinical specimens collected from patients in the Houston metropolitan area during the period March 2020 to February 2021 have been sequenced. Variants of concern were first detected among specimens collected in December 2020, and now 23 UK variants (B.1.1.7), 2 South African variants (B.1.351), and 4 Brazilian variants (P.1) have been identified. One hundred and sixty two patients infected with the California variants (B.1.429, N = 143; B.1.427, N = 19) and 39 patients infected with Brazil P.2 variants were also identified (Table 1).

UK Variant of Concern (B.1.1.7)

The UK variant B.1.1.7 was first identified in September 2020 in the United Kingdom and was designated as a variant of concern in South London on December 14, 2020. It was strongly associated with a resurgence of SARS-CoV-2 infections in that region and rapidly became the dominant lineage. More importantly, the United Kingdom has the most extensive SARS-CoV-2 genome sequencing program in the world, making it particularly well-situated to rapidly identify new variants. Of the approximately 500,000 SARS-CoV-2 genome sequences submitted to GISAID from global sources, approximately
The South Africa B.1.351 variant of concern was first identified in a COVID-19 epidemic wave occurring in Nelson Mandela Bay in October 2020. This variant was concerning because of its large number of spike protein mutations (including K417N, E484K, and N501Y) and apparent increased transmissibility. These three mutations are located in the receptor binding domain of spike protein (Figure 1) and has spread from California to 34 states since October 2020 (B.1.429 and B.1.427 isolation report, https://outbreak.info/situation-reports?pango=B.1.429, last accessed February 19, 2021). The California variants are noteworthy primarily for their emergence and rapid spread in Los Angeles County and identification elsewhere in the United States. However, as of February 17, 2021, they have not been designated as variants of concern by the CDC.

Geospatial Distribution of Variants

Table 1: Variants of Concern or Variant of Interest Identified in the Houston, TX, Metropolitan Area

<table>
<thead>
<tr>
<th>Variant</th>
<th>No. of isolates</th>
</tr>
</thead>
<tbody>
<tr>
<td>B.1.1.7</td>
<td>23</td>
</tr>
<tr>
<td>B.1.351</td>
<td>1</td>
</tr>
<tr>
<td>P.1</td>
<td>4</td>
</tr>
<tr>
<td>P.2</td>
<td>39</td>
</tr>
<tr>
<td>B.1.429</td>
<td>143</td>
</tr>
<tr>
<td>B.1.427</td>
<td>19</td>
</tr>
</tbody>
</table>

The emergence of what is now known as the California variant, originally known as CAL.20C and later designated as lineages B.1.429 and B.1.427, was first identified in Los Angeles County in July 2020 as a single isolate (B.1.429 Lineage Report, outbreak.info 2021, https://outbreak.info/situation-reports?pango=B.1.429, last accessed February 18, 2021). This variant re-emerged in October 2020 and was associated with an increasing number of cases during a wave of SARS-CoV-2 infections in the region. Variant B.1.429 accounted for 36% of isolates collected from late November to late December 2020 in Los Angeles County. Since November 2020, this variant has been detected in 42 states in the United States, and was first found in Houston Methodist Hospital patients in specimens obtained in late December 2020. So far, 143 and 19 patients with the B.1.429 and B.1.427 isolates, respectively, have been identified (Table 1). The B.1.427 variant is closely related to B.1.429 (Figure 1) and has spread from California to 34 states since October 2020 (B.1.427 Lineage Report, https://outbreak.info/situation-reports?pango=B.1.427, last accessed February 19, 2021). The California variants are noteworthy primarily for their emergence and rapid spread in Los Angeles County and identification elsewhere in the United States. However, as of February 17, 2021, they have not been designated as variants of concern by the CDC.
Herein we report detection of the UK (B.1.1.7), South Africa (B.1.351), and Brazil (P.1) SARS-CoV-2 variants of concern from patients in the Houston metropolitan region. Geographically widespread dissemination of the Cal.20C California (B.1.429 and B.1.427) variants of interest were identified. These SARS-CoV-2 variants are distributed

![Diagram of SARS-CoV-2 variants]

**Figure 1**  
A: Schematic showing structural changes present in the spike protein of the major severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) variants identified in the study. B: Mapping of important changes onto the cryoEM structure of spike protein. The color scheme matches that used in A. Blue, amino-terminal domain (NTD); purple, receptor-binding domain (RBD); orange, S1 domain (S1); and yellow, S2 domain (S2). Aggregate mutations present in variants of concern are colored in red when amino acid residues are present in the resolved structure. *Left panel*: Side view of SARS-CoV-2 prefusion-stabilized spike. *Right panel, top view*: Structure of PDB 6vsb was used as reference.

**Discussion**

Herein we report detection of the UK (B.1.1.7), South Africa (B.1.351), and Brazil (P.1) SARS-CoV-2 variants of concern from patients in the Houston metropolitan region. Geographically widespread dissemination of the Cal.20C California (B.1.429 and B.1.427) variants of interest were identified. These SARS-CoV-2 variants are distributed...
across a large geospatial region in the metropolitan region (Figure 2), indicating successful patient-to-patient transmission among Houstonians. None of the affected patients was from a common household or reported recent international travel, suggesting that every infection was independently acquired locally or during domestic travel. Given that Houston is a culturally and ethnically diverse population center with two international airports, a major shipping center, and a global energy sector, the discovery of patients infected with each of the four concerning SARS-CoV-2 variants is not unexpected but it is disquieting. With this report, Houston now becomes the first US city to document patients infected with each of the four SARS-CoV-2 variants of concern, testament to the extensive sequencing of COVID-19 patient samples.

The P.2 variant gained recent attention in the scientific and lay press because it has been reported to cause SARS-CoV-2 reinfections (https://virological.org/t/spike-e484k-mutation-in-the-first-sars-cov-2-reinfection-case-confirmed-in-brazil-2020/584, last accessed February 19, 2021). Thirty nine P.2 infections were identified among Houston patients. Although it is currently a numerically minor cause of all Houston-area infections, P.2 is now the most common SARS-CoV-2 variant of concern in this population.

The E484K amino acid replacement in spike protein is characteristic of P.1, P.2, and B.1.351 strains (Figure 1). It has independently arisen in many different SARS-CoV-2 genomic backgrounds, including some B.1.1.7 strains. This amino acid replacement has caused substantial public health concern because of its potentially detrimental effects on neutralizing activity of therapeutic monoclonal antibodies, sera obtained from naturally infected individuals, and post-vaccination sera. That is, the E484K amino acid change may facilitate vaccine escape. Among the Houston SARS-CoV-2 genomes, E484K was detected 84 times (0.4% of the total genomes sequenced). It was first detected in a respiratory specimen collected in July 2020, near the peak of our second massive wave of infections, and has been identified in many diverse genomic backgrounds thereafter. Because of this strong signal of convergent evolution, all Houston SARS-CoV-2 genomes will continue to be closely monitored for the E484K amino acid change.

Recently, the Q677H amino acid change in spike protein has been identified in SARS-CoV-2 patient samples collected in multiple US states and other global locations. Q677H has arisen in at least six distinct genomic backgrounds. A Q667P amino acid change has also been identified. Among the Houston genomes, Q677H occurred 288 times (1.4%) and is encoded by two different nucleotide changes. Two other amino acid changes, 677P (in 330 genomes; 1.6%) and Q677K (in 2 genomes; <0.1%) were identified in Houston. Taken together, these data suggest selection for a yet to be determined biologic phenotype associated with amino acid replacements at position 677.

Many population genomic studies performed in various global locations have clearly demonstrated that SARS-CoV-2
variants with biologically relevant phenotypes have evolved. Emergence of new variants underlines the need for ongoing extensive genomic sequencing efforts for early identification and public health warning. In support of these efforts, our laboratory has devoted substantial resources to SARS-CoV-2 genomics, resulting in sequence analysis of more genomes than any other state in the United States. Since March 2020, approximately 36,500 SARS-CoV-2—positive patients have received care in our Houston Methodist health system, and 20,453 virus genomes have been sequenced. In total, this data set represents 56% of our Houston Methodist COVID-19 patients. Inasmuch as almost 500,000 COVID-19 infections have been reported in the Houston metropolitan area (Texas Medical Center COVID-19 Dashboard 2021, https://www.tmc.edu/coronavirus-updates/covid-19-positive-cumulative-cases, last accessed February 19, 2021), the genome of 4.1% of all cases reported in our area has been sequenced. On the basis of modeling, this sample depth may be sufficient to identify all variants occurring at a biologically relevant frequency. Because of the wide geographic catchment of the eight-hospital system that serves a diverse patient population, the data presented herein likely reflect a reasonably detailed overview of SARS-CoV-2 genomic diversity throughout our metropolis. This comparatively deep sampling of the Houston metropolitan SARS-CoV-2 population was used to identify patients infected with variants of concern, and provide information regarding the timeframe of initial presence and frequency of each variant. The strategy was modeled on the aggressive genome sequencing being conducted in the United Kingdom, a global leader in SARS-CoV-2 genome sequencing.

The large SARS-CoV-2 genome data set and comprehensive infrastructure of the Houston Methodist Hospital are critical resources. By linking the SARS-CoV-2 whole genome sequence data to patient metadata present in the electronic medical record, analytic tools such as high-performance compute clusters and machine learning were used to investigate the relationship between genomic diversity and phenotypic traits, such as strain virulence or patient outcomes. For example, recent reports of increased mortality caused by B.1.1.7 variant strains are concerning and worthy of further investigation. The goal is to sequence the SARS-CoV-2 genome of every infected patient in the health care system in near real time, and expand outward to other patients in our community. Consistent with these goals, the American Rescue Plan, announced by the Biden administration, proposes to substantially fund sequencing capacity in the United States (White House Briefing Room, https://www.whitehouse.gov/briefing-room/legislation/2021/01/20/president-biden-announces-american-rescue-plan, last accessed February 19, 2021). These results from a major metropolitan region in the United States underline the necessity of greatly increased genome surveillance to rapidly identify and track the emergence and introduction of SARS-CoV-2 variants in the United States and local areas.

Acknowledgments

We thank Drs. Jessica Thomas and Zejuan Li, Erika Walker, the many talented and dedicated molecular technologists, and many labor pool volunteers in the Molecular Diagnostics Laboratory and Methodist Research Institute for dedicated efforts; Drs. Marc Boom and Dirk Sostman for support, many generous Houston philanthropists for tremendous support, and the Houston Methodist Academic Institute Infectious Diseases Fund, which have made this ongoing project possible; Jessica W. Podnar and personnel in the University of Texas Genome Sequencing and Analysis Facility for sequencing some of the genomes in this study; Hung-Chen Kuo and G. Nguyen for genome sequencing support; Maulik Shukla and Marcus Nguyen at PATRIC Bioinformatics Resource Center for assistance with sequence analysis; the originating and submitting laboratories of the severe acute respiratory syndrome coronavirus 2 genome sequences from GISAID’s EpiFlu Database, used in some of the work presented herein; many colleagues for critical reading of the manuscript and suggesting improvements; and Dr. Kathryn Stockbauer, Sasha Pejerrey, Adrienne Winston, and Heather McConnell for help with figures, tables, and editorial contributions.

Author Contributions

J.M.M. conceptualized and designed the project; S.W.L., R.J.O., P.A.C., S.S., R.O., J.J.D., M.S., P.Y., L.P., K.R., M.N.S, J.C., I.J.F, and J.G. performed research. All authors wrote the manuscript.

Supplemental Data

Supplemental material for this article can be found at http://doi.org/10.1016/j.ajpath.2021.03.004.

References


