Genetic Characteristics and Phylogeny of 969-bp S Gene Sequence of SARS-CoV-2 from Hawai‘i Reveals the Worldwide Emerging P681H Mutation

David P. Maison MS; Lauren L. Ching BS; Cecilia M. Shikuma MD; and Vivek R. Nerurkar PhD

Abstract

The COVID-19 pandemic has ravaged the world, caused over 1.8 million deaths in its first year, and severely affected the global economy. Hawai‘i has not been spared from the transmission of SARS-CoV-2 in the local population, including high infection rates in racial and ethnic minorities. Early in the pandemic, we described in this journal various technologies used for the detection of SARS-CoV-2. Herein we characterize a 969-bp SARS-CoV-2 segment of the S gene downstream of the receptor-binding domain. At the John A. Burns School of Medicine Biocontainment Facility, RNA was extracted from an oropharyngeal swab and a nasal swab from 2 patients from Hawai‘i who were infected with SARS-CoV-2 in August 2020. Following PCR, the 2 viral strains were sequenced using Sanger sequencing, and phylogenetic trees were generated using MEGAX. Phylogenetic tree results indicate that the virus has been introduced to Hawai‘i from multiple sources. Further, we decoded 13 single nucleotide polymorphisms across 13 unique SARS-CoV-2 genomes within this region of the S gene, with 1 non-synonymous mutation (P681H) found in the 2 Hawai‘i strains. The P681H mutation has unique and emerging characteristics with a significant exponential increase in worldwide frequency when compared to the plateauing of the now universal D614G mutation. The P681H mutation is also characteristic of the new SARS-CoV-2 variants from the United Kingdom and Nigeria. Additionally, several mutations resulting in cysteine residues were detected, potentially resulting in disruption of the disulfide bridges in and around the receptor-binding domain. Targeted sequence characterization is warranted to determine the origin of multiple introductions of SARS-CoV-2 circulating in Hawai‘i.

Keywords

S gene, spike protein, phylogenetic tree, genomic characterization, Hawai‘i strains, single nucleotide polymorphism, variant, proline

Abbreviations

ASGPB = Advanced Studies in Genomics, Proteomics, and Bioinformatics
A1708D = alanine to aspartic acid at amino acid 1708
A570D = alanine to aspartic acid at amino acid 570
A522S = alanine to serine at amino acid 522
A771S = alanine to serine at amino acid 771
ACE2 = angiotensin-converting enzyme 2
R577C = arginine to cysteine at amino acid 577
R52I = arginine to isoleucine at amino acid 52
N501Y = asparagine to tyrosine at amino acid 501
D614G = aspartic acid to glycine at amino acid 614
D1118H = aspartic acid to histidine at amino acid 1118
D3L = aspartic acid to leucine at amino acid 3
cDNA = complementary deoxyribonucleic acid
COVID-19 = coronavirus disease 2019
ΔG2676 = deletion of glycine amino acid 2676
ΔH69 = deletion of histidine amino acid 69
ΔF3677 = deletion of phenylalanine amino acid 3677
ΔS3675 = deletion of serine amino acid 3675
ΔY145 = deletion of tyrosine amino acid 145
ΔV70 = deletion of valine amino acid 70
DNA = deoxyribonucleic acid
EUA = emergency use authorization
GISAID = Global Initiative of Sharing All Influenza Data
E780Q = glutamic acid to glutamine at amino acid 780
Q27stop = glutamine to stop codon at amino acid 27
IBC = Institutional Biosafety Committee
IRB = Institutional Review Board
I726F = isoleucine to phenylalanine at amino acid 726
I2230T = isoleucine to threonine at amino acid 2230
I584V = isoleucine to valine at amino acid 584
MUSCLE = Multiple Sequence Comparison by Log-Expectation
NCBI = National Center for Biotechnology Information
NERVTAG = New and Emerging Respiratory Virus Threats Advisory Group
nCoV = novel coronavirus
PID = patient identification
F797C = phenylalanine to cysteine at amino acid 797
F543L = phenylalanine to leucine at amino acid 543
PCR = polymerase chain reaction
P681H = proline to histidine at amino acid 681
RBD = receptor-binding domain
RT-PCR = reverse transcriptase-polymerase chain reaction
RNA = ribonucleic acid
S982A = serine to alanine at amino acid 982
S680C = serine to cysteine at amino acid 680
S235F = serine to phenylalanine at amino acid 235
SARS-CoV-2 = severe acute respiratory syndrome coronavirus 2
SNP = single nucleotide polymorphism
T1001I = threonine to isoleucine at amino acid 1001
T716I = threonine to isoleucine at amino acid 716
TBE = tris/borate/ethylenediaminetetraacetic acid
Y73C = tyrosine to cysteine at amino acid 73
FDA = Food and Drug Administration
VOC = variant of concern
VTM = viral transport media

Introduction

The zoonotic virus responsible for the present Coronavirus disease 2019 (COVID-19) pandemic is severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) (formerly novel coronavirus [nCoV]).[1] SARS-CoV-2 emerged in Wuhan, China, at a seafood market in November 2019 and has been evolving ever since.[1-4] The COVID-19 pandemic resultant from SARS-CoV-2 has been responsible for infecting more than 107 million people worldwide and has been fatal in more than 2.3 million persons experiencing an infection.[5] The state of Hawai‘i has reported more than 26 000 cases and 400 deaths, with daily reports steady at approximately 60 new cases per day since August 2020.[6]

To understand SARS-CoV-2 emergence and the disease COVID-19, one must look at the genome as the virus evolves and emerges through genomic alterations and adaptations. SARS-CoV-2 belongs to the family of *Coronaviridae* (genus Betacoronavirus), which are viruses with 26 000–32 000 nucleotide long single-stranded positive-sense RNA genomes.[7–9] Geneticists and virologists look at the SARS-CoV-2 genome and its adaptations to analyze the nucleotide and amino acid variations. Analysis of the SARS-CoV-2 genome will allow us to track the spread through unique genomic fingerprints,[1] determine whether these adaptations alter the viral fitness, infectious capabilities,[10] and develop potential vaccines and therapeutics.[11,12]

While genes encode 20 proteins consisting of 4 structural and 16 non-structural proteins in the SARS-CoV-2 ~30 000-bp genome,[9] the gene looked to most is the S gene responsible for the spike protein. The spike protein is a 1273 amino acid long (YP_009724390.1)(NC_045512) surface protein that is the viral component accountable for interacting with the human angiotensin-converting enzyme 2 (ACE2) (UNIPROT ID Q9BYF1).[7,12-14] ACE2, genetically encoded on the human X chromosome (ENSG00000130234), is a component of the renin-angiotensin hormone system in humans and is ultimately a vasodilator.[15,16] This interaction between SARS-CoV-2 spike protein and human ACE2 via the receptor-binding domain (RBD) allows SARS-CoV-2 to enter cells and infect the human host.[12] Mutations in the spike protein can alter binding efficiency and viral fitness.[10] Indeed, some nucleotide mutations in the SARS-CoV-2 S gene change pathogenicity,[12] affect viral fitness,[10,17] reduce virulence,[1,4] and have become commonplace in tracking the spread of SARS-CoV-2.[17] These S gene and spike protein mutations can increase transmission of the virus between hosts through anatomical localization to the upper respiratory tract.[10] Therefore, to understand the SARS-CoV-2 pathogenicity, it is important to characterize the virus mutations to study viral pathogenesis and vaccine development. In this study, we report analysis of a 969-bp SARS-CoV-2 S gene from 2 patients from Hawai‘i to understand the changes in the spike protein, a target for vaccines.

Methods

Patient Samples and Viral RNA Extraction

The 2 patients (patient identification [PID] 00498 and PID 00708) analyzed in this report were part of the University of Hawai‘i at Mānoa Institutional Review Board (IRB)-approved H051 study (IRB# 2020-00367). Patient PID 00498 and patient PID 00708 are both males, 1 identifies as white, and the other identifies as Japanese, Okinawan, and Filipino; their mean age is 29.5 years. These patients were previously identified as having SARS-CoV-2, and oropharyngeal swab (OS - PID 00498) and nasal swab (NS - PID 00708) were collected in August 2020, 3 days after first polymerase chain reaction (PCR) positive diagnosis. Samples were stored at -80°C. Neither of the patients had traveled outside of Honolulu in the week before their first PCR positive SARS-CoV-2 diagnosis, however, both identified potential sources of exposure in Honolulu.

Swabs stored in viral transport media (VTM) at -80°C were thawed in the biosafety cabinet at the John A. Burns School of Medicine high containment laboratory as part of the University of Hawai‘i at Mānoa Institutional Biosafety Committee (IBC)-approved study (IBC#20-04-830-05). VTM was centrifuged to separate the supernatant from debris, aliquoted, and 140 µL of the VTM was used for viral RNA purification using the QIAamp® Viral RNA Mini Kit (Cat# 52906) following the manufacturer›s instructions. The samples were eluted in 30 µL of the elution buffer.

Reverse Transcriptase Polymerase Chain Reaction and Sequencing

As per the manufacturer's instructions, purified viral RNA was transcribed into cDNA using the Takara RNA LA PCR Kit (Cat #RR012A) with random 9 mers and an extension time of 90 minutes. Primer sets were designed based on published sequences and were procured from Integrated DNA Technologies (Coralville, IA). A 1127-bp segment of the S gene was amplified using the Takara RNA LA PCR Kit (Cat #RR012A) and primers CF and CR (Figure 1).[18] PCR was conducted according to the manufacturer's instructions and cycled on the Applied Biosystems GeneAmp® PCR System 9600. PCR products were then electrophoresed on 1.5% agarose 1x TBE gels at 50V, and the amplicons of interest were purified using the Qiagen QIAquick Gel Extraction Kit (Cat# 28704).

Sanger sequencing was conducted on the amplicons using 4 primers (CF, CR,[18] CR2, and CR3) at the Advanced Studies in Genomics, Proteomics, and Bioinformatics (ASGPB) core facility at the University of Hawai‘i at Mānoa (Figure 1). The resulting sequences were input into and verified using both MEGAX[19,20] and SnapGene software (Insightful Science, www.snapgene.com) and aligned using Multiple Sequence Comparison by Log-Expectation (MUSCLE) program[21] to define the contiguous sequence. The resulting 969-bp consensus sequences were uploaded to the National Center for Biotechnology Information (NCBI) database. The S gene 969-bp region encompasses nucleotides 23 042 to 24 010 and corresponds to amino acids 494 to 816, which involves the 3' and C-terminal of the RBD that ends at nucleotide 23 185 and amino acid 541.[7]

Single Nucleotide Polymorphism (SNP) Analysis

Sixty-eight coronavirus strains representing alpha and beta lineages were selected from the NCBI database representing 25 distinct geographical locations spanning the pandemic duration and at least 1 SARS-CoV-2 strain per month from December 2019 to September 2020. Of these 68 coronavirus strains, 55 were SARS-CoV-2 strains. All SARS-CoV-2 sequences, including previously published Hawai‘i sequences, were first aligned, and redundant sequences were removed from further analysis.

Coronavirus sequences were aligned with the 969-bp S gene region with SnapGene using MUSCLE, and the corresponding region was used for future analysis.[21] The non-SARS-CoV-2 strains were removed if the 969-bp S gene of SARS-CoV-2 sequence did not align with the S gene of the non-SARS-CoV-2 strains. Based on the alignment, SNPs were identified and annotated into SnapGene to analyze the amino acid substitutions.

Upon finding the P681H mutation among the 2 Hawai'i strains in this study, the Global Initiative of Sharing All Influenza Data (GISAID) database[22,23] was used to filter worldwide SARS-CoV-2 sequences by the P681H mutation to create a ratio of sequences containing the P681H mutation to all sequences reported in the GISAID database for a given month. Inclusion criteria were for sequences providing a full month, day, and year. The D614G mutation underwent assessment in the same manner for comparison. All prevalence data converted into ratio underwent a logarithmic transformation. Pearson's correlation tests between P681H frequency versus month, D614G frequency versus month, and P681H frequency versus D614G frequency were conducted and verified using GraphPad Prism version 9.0.0 for Mac (GraphPad Software, San Diego, California USA, www.graphpad.com), JASP version 0.14,[24] and RStudio version 1.3.1093 (R version 4.0.3).[25]

Phylogenetic Tree

After the SNP analysis, incomplete sequences were removed before the construction of the phylogenetic tree. The phylogenetic tree was constructed using MEGAX.[19] The alignment was first done using the program MUSCLE.[21] The phylogenetic tree was then generated with Maximum Likelihood parameters with 1,000 bootstraps in MEGAX[19,20] using the University of Hawai'i MANA High Performance Computing Cluster. The output tree from MEGAX was rooted using FigTree version 1.4.4 based on alpha coronavirus human 299E (KF514433).[26]

Results

Gene Amplification and Sequence Analysis

SARS-CoV-2 genomic sequences were detected by reverse transcriptase-polymerase chain reaction (RT-PCR) in both the patients, PID 00498 and PID 00708, using various primers spanning the S gene (Figure 1). A 1127-bp segment was amplified and sequenced, and the entire sequence of the amplicon was aligned with at least 1 forward and 1 reverse sequence (translated into reverse-complement) to span the whole 1127-bp region. For final sequence analysis, a 969-bp sequence verified by sequencing the 5' and 3' ends was used. The 2 Hawai'i sequences were deposited in the GenBank, accession numbers MW237663 for PID 00498 and MW237664 for PID 00708.

SNP Analysis

Of the 55 original non-Hawai'i SARS-CoV-2 strains, 47 were redundant in the 969-bp segment of the S gene. Of the 12 Hawai'i strains deposited in the GenBank, 9 were redundant in the 969-bp sequence region. Thus, we analyzed 8 non-Hawai'i SARS-CoV-2 strains and 3 SARS-CoV-2 strains from Hawai'i. With the addition of the 2 SARS-CoV-2 strain sequences from this study, a total of 13 SARS-CoV-2 sequences were compared, and SNPs encompassing the 969-bp region of the S gene were analyzed (Table 1). The alignment containing the 13 sequences revealed 13 SNPs (Table 1) (Figure 1). Eleven of the 13 mutations resulted in non-synonymous mutations (A522S, F543L, R577C, I584V, D614G, S680C, P681H, I726F, A771S, E780Q, and F797C) (Table 1 and Figure 1). Two of the 13 mutations resulted in a synonymous mutation (amino acid 541 and 790) (Table 1 and Figure 1). The P681H mutation is unique to the Hawai'i strains from this study (MW237663 and MW237664).

GISAID reported the first P681H mutation on March 12, 2020 (EPI_ISL_430887).[27] Further, from March 01, 2020, through January 31, 2021, GISAID reports a total of 65 959 strains that have the P681H mutation. During that same time, GISAID has reported approximately 480 822 SARS-CoV-2 strains (Table 2). Pearson's correlation between time in months versus prevalence of P681H (Figure 2A) and D614G (Figure 2B) of logarithmically transformed data indicates an increase in the number of strains having the P681H mutation (r=0.97, P <.0001) (Figure 2A) and plateauing of the D614G mutation (r=0.77, P=.005) (Figure 2B). P681H mutations were not reported in May 2020. Further, Pearson's correlation indicates a positive correlation (r=0.69, P=.03) between the worldwide prevalence of P681H and D614G (Figure 2C).

Phylogenetic Tree

Of the 13 SARS-CoV-2 sequences used for SNP identification, 2 were incomplete due to unidentified nucleotides and were removed from the phylogenetic analysis. Similarly, of the 13 non-SARS-CoV-2 sequences, 4 did not align to the 969-bp S gene due to large insertions or deletions and were removed from further analysis. Therefore, the final phylogenetic tree was constructed using 20 coronavirus sequences, 11 SARS-CoV-2, and 9 non-SARS-CoV-2 sequences (Figure 3). Based on the phylogenetic tree constructed using the Maximum Likelihood method, the alpha and beta coronaviruses segregated as expected. Further, the beta coronaviruses lineages A, B, C, and D segregated with a bootstrap value of >70. Within the beta coronavirus lineage B, the SARS-CoV-1 and the bat coronaviruses were distinctly segregated from the SARS-CoV-2 with a bootstrap value of 92.

Within the SARS-CoV-2 branch of beta coronavirus lineage B, the D614G mutation was the defining node for branch separation. The 3 sequences lacking the D614G mutation (NC_045512, MT344949, and MT093571) are separated from sequences with the D614G mutation with a bootstrap value of 100, except for the MW066483 Hawai'i sequence, which also does not have the D614G mutation. Wuhan (NC_045512)

and Hawai‘i (MT344949) strains are identical and contain no mutations, as NC_045512 is the reference genome for SARS-CoV-2. The Sweden strain (MT093571) has a F797C mutation. The Hawai‘i strain MW064483 is the next closest cluster to the D614G defining node and additionally contains the synonymous tyrosine mutation at amino acid 790. All the remaining sequences have the D614G mutation. Clustering near the Hawai‘i strain MW066483 are 2 strains from the state of New York (MW035565 and MW035511), both containing the E780Q mutation and MW035565 also containing the A522S mutation and a synonymous phenylalanine mutation at amino acid 541. Branching from the New York strains cluster is a strain from the state of Washington (MT994395), exclusively having the I584V mutation. The 2 Hawai‘i strains from this study (MW237663 and MW237664) have the emerging P681H mutation and cluster closely with previously published SARS-CoV-2 sequences from Hawai‘i (MT627421) and China (MT407659). MT627421 strain from Hawai‘i and MT407659 strain from China are identical and are the only sequences to contain the D614G mutation exclusively.

In summary, SARS-CoV-2 strains from Hawai‘i deposited in the GenBank in March 2020 clustered with sequences from Wuhan, Sweden, China, and the state of New York. The SARS-CoV-2 strains in this study cluster from the state of Washington and with sequences from China and Hawai‘i. Five of the 13 SARS-CoV-2 sequences used in the phylogenetic tree are sequences from Hawai‘i, marked with an asterisk. Coronavirus lineage determinations are based on phylogenetic trees constructed by Chan and colleagues[28] and Su and colleagues.[29]

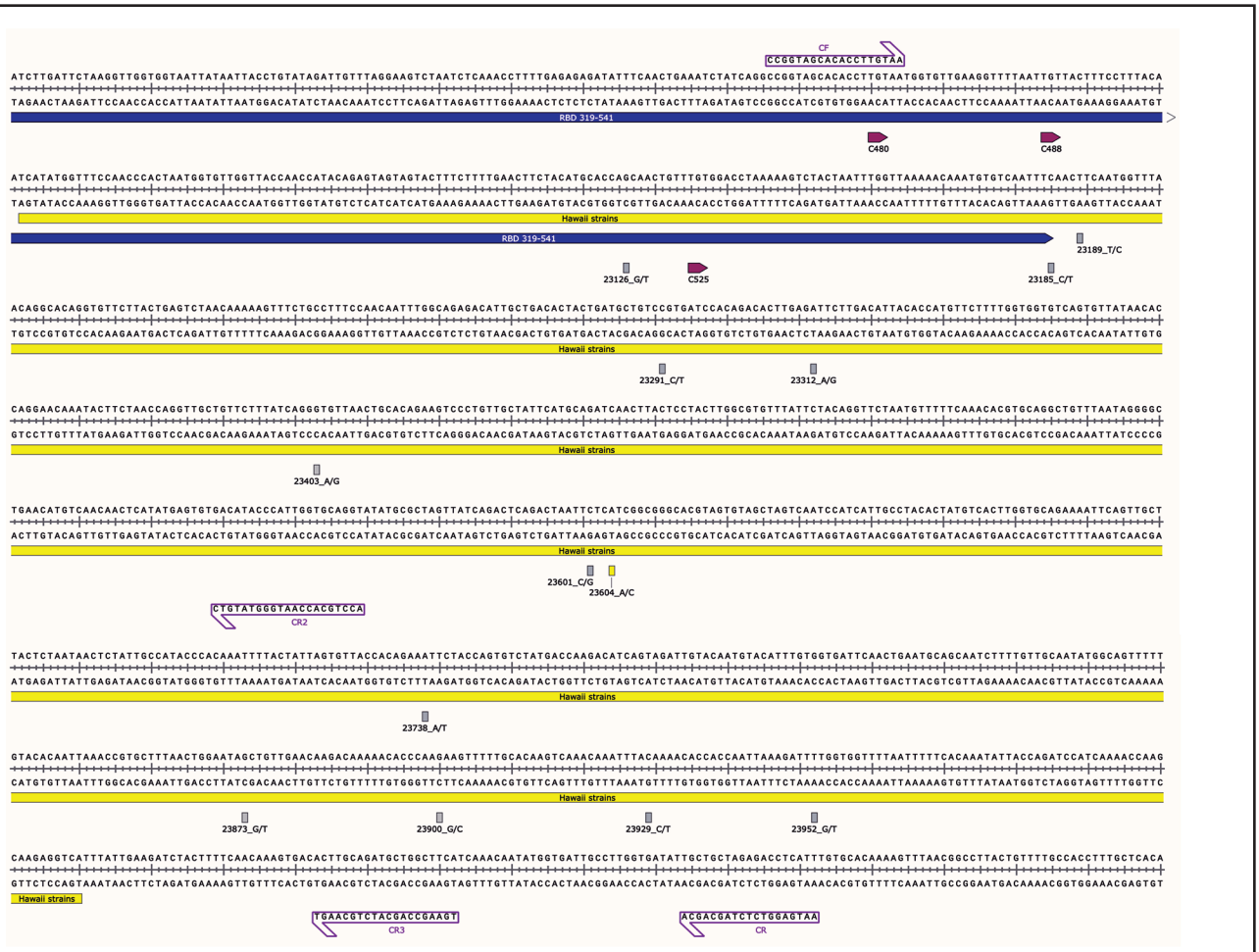

Figure 1. SARS-CoV-2 S Gene Region used in this Study Along with Annotated Primers, Mutations, and Cysteine Residues of the Receptor Binding Domain

This figure represents the Hawai‘i strain MW237663 and MW237664 sequences. The primer pair, CF/CR, was used to amplify the 1127-bp S gene fragment, and primers CF, CR, CR2, and CR3 depicted with purple boxes were used for Sanger sequencing. The yellow box indicates the start and end of the 969-bp sequence. The blue line indicates the 3’ end of the S gene receptor-binding domain (RBD). RBD cysteine residues are shown in depicted boxes. All mutations found in this study are in their respective loci with nucleotide numbers, and rectangular boxes correlating to the sense strand as indicated after the nucleotide number and underscore in the figure (nucleotide/protein mutations: G23126T/A522S, C23185T/F541F, T23189C/F543L, C23291T/R577C, A23312G/I584V, A23403G/D614G, C23601G/S680C, C23604A/P681H, A23738T/I726F, G23873T/A771S, G23900C/E780Q, C23929T/Y790Y, and T23952G/F797C). All boxes are grey except for the P681H mutation seen in the Hawai‘i strains from this study, shown with a yellow rectangle. Image was generated with the SnapGene software (from Insightful Science; available at www.snapgene.com) and was created with BioRender.com.

Table 1. Comparison of Single Nucleotide Polymorphisms and Resultant Amino Acid Substitutions in a 969-bp S gene of SARS-CoV-2 Strains from Hawai'i

ACCESSION AND IDENTIFIER	SNP																									
	23126 NT	522 AA	23185 NT	541 AA	23189 NT	543 AA	23291 NT	577 AA	23312 NT	584 AA	23403 NT	614 AA	23601 NT	680 AA	23604 NT	681 AA	23738 NT	726 AA	23873 NT	771 AA	23900 NT	780 AA	23929 NT	790 AA	23952 NT	797 AA
MW237663_HI	G	Ala	C	Phe	T	Phe	C	Arg	A	Ile	**G**	**Gly**	C	Ser	**A**	**His**	A	Ile	G	Ala	G	Glu	C	Tyr	T	Phe
MW237664_HI	G	Ala	C	Phe	T	Phe	C	Arg	A	Ile	**G**	**Gly**	C	Ser	**A**	**His**	A	Ile	G	Ala	G	Glu	C	Tyr	T	Phe
MW064483.1_HI	G	Ala	C	Phe	T	Phe	C	Arg	A	Ile	A	Asp	C	Ser	C	Pro	A	Ile	G	Ala	G	Glu	**T**	**Tyr**	T	Phe
MT627421_HI	G	Ala	C	Phe	T	Phe	C	Arg	A	Ile	**G**	**Gly**	C	Ser	C	Pro	A	Ile	G	Ala	G	Glu	C	Tyr	T	Phe
MT344949_HI	G	Ala	C	Phe	T	Phe	C	Arg	A	Ile	A	Asp	C	Ser	C	Pro	A	Ile	G	Ala	G	Glu	C	Tyr	T	Phe
NC_045512.2_CHN	G	Ala	C	Phe	T	Phe	C	Arg	A	Ile	A	Asp	C	Ser	C	Pro	A	Ile	G	Ala	G	Glu	C	Tyr	T	Phe
MT407659.1_CHN	G	Ala	C	Phe	T	Phe	C	Arg	A	Ile	**G**	**Gly**	C	Ser	C	Pro	A	Ile	G	Ala	G	Glu	C	Tyr	T	Phe
MW035565_NY	**T**	**Ser**	**T**	**Phe**	T	Phe	C	Arg	A	Ile	**G**	**Gly**	C	Ser	C	Pro	A	Ile	G	Ala	**C**	**Gln**	C	Tyr	T	Phe
MW035511_NY	G	Ala	C	Phe	T	Phe	C	Arg	A	Ile	**G**	**Gly**	C	Ser	C	Pro	A	Ile	G	Ala	**C**	**Gln**	C	Tyr	T	Phe
MT994395_WA	G	Ala	C	Phe	T	Phe	C	Arg	**G**	**Val**	**G**	**Gly**	C	Ser	C	Pro	A	Ile	G	Ala	G	Glu	C	Tyr	T	Phe
MT093571.1_SWE	G	Ala	C	Phe	T	Phe	C	Arg	A	Ile	A	Asp	C	Ser	C	Pro	A	Ile	G	Ala	G	Glu	C	Tyr	**G**	**Cys**
MT451798.1_AUS	N	-	C	Phe	**Y**	**Leu/ Phe**	**Y**	**Arg/ Cys**	A	Ile	A	Asp	**S**	**Ser/ Cys**	C	Pro	**W**	**Ile/ Phe**	**K**	**Ala/ Ser**	G	Glu	C	Tyr	T	Phe
MT394864.1_DEU	N	-	C	Phe	T	Phe	C	Arg	A	Ile	**G**	**Gly**	C	Ser	C	Pro	A	Ile	G	Ala	G	Glu	C	Tyr	T	Phe

Abbreviations: HI, Hawai'i; CHN, China; NY, New York; WA, Washington; SWE, Sweden; AUS, Australia; DEU, Germany;
SNP, Single Nucleotide Polymorphism; NT, nucleotide; A, Adenine; C, Cytosine; G, Guanine; T, Thymine; K, Guanine or Thymine; S, Guanine or Cytosine; W, Adenine or Thymine; Y, Cytosine or Thymine;
AA, Amino Acid; Ala, Alanine; Arg, Arginine; Asp, Aspartic Acid; Cys, Cysteine; Gln, Glutamine; Glu, Glutamic Acid; Gly, Glycine; His, Histidine; Ile, Isoleucine; Leu, Leucine; Phe, Phenylalanine; Pro, Proline; Ser, Serine; Tyr, Tyrosine; Val, Valine

Table 2. The Distribution and Frequency of P681H and D614G Mutations Among All SARS-CoV-2 Sequences by Month Reported in the GISAID Database in Year 2020 and 2021

Month and Year	n	P681H n (%)	D614G n (%)
March 2020	47 120	9 (0.02%)	34 375 (73.0%)
April 2020	44 386	8 (0.02%)	37 782 (85.1%)
May 2020	22 230	0 (0%)	20 412 (91.8%)
June 2020	23 975	14 (0.06%)	23 126 (96.5%)
July 2020	22 917	53 (0.2%)	22 285 (97.2%)
August 2020	25 439	229 (0.9%)	25 094 (98.6%)
September 2020	29 831	223 (0.8%)	29,674 (99.5%)
October 2020	50 910	407 (0.8%)	50 672 (99.5%)
November 2020	59 650	2701 (4.5%)	59 495 (99.7%)
December 2020	73 405	18 347 (25.0%)	72 868 (99.3%)
January 2021	80 959	43 968 (54.3%)	80 516 (99.4%)

Abbreviations: GISAID, Global Initiative of Sharing All Influenza Data

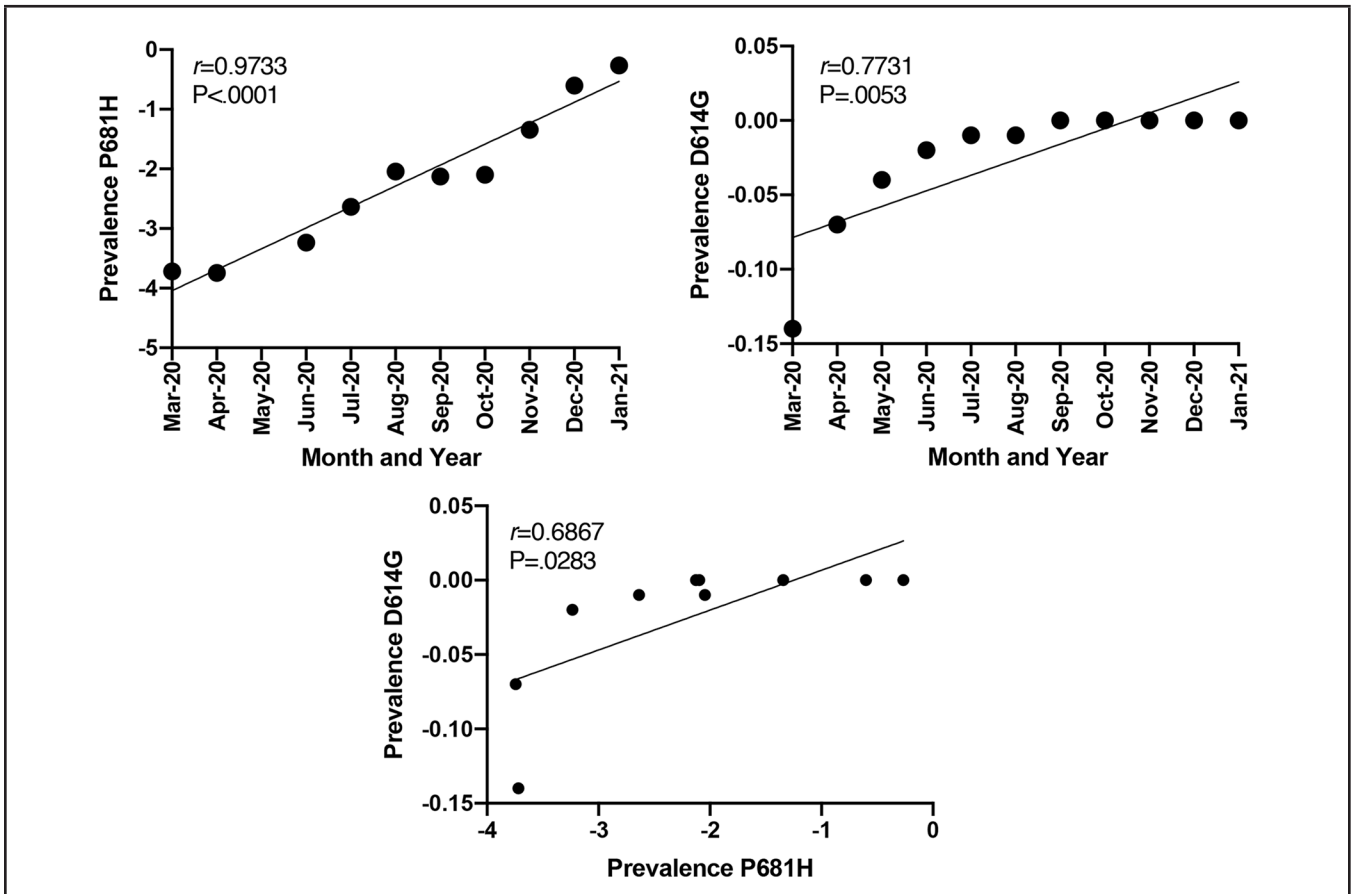

Figure 2. Pearson's Correlation of Logarithmically Transformed Data Showing Positive Correlation for Time in Months Versus Both P681H and D614G Mutations

(A) Graphical representation of the logarithmic transformed ratio of P681H mutation among all reported Global Initiative of Sharing All Influenza Data (GISAID) strains on the y-axis and month on the x-axis. Linear regression line shown along with Pearson's correlation, r=0.97, P=2.157e-06. (B) Graphical representation of the logarithmic transformed ratio of D614G mutation among all reported GISAID strains on the y-axis and month on the x-axis. Linear regression line shown along with Pearson's correlation, r=0.77, P=.005. (C) Graphical representation of the logarithmic transformed ratio of D614G mutation among all reported GISAID strains on the y-axis and the logarithmic transformed ratio of P681H mutations among all reported GISAID strains on the x-axis. Linear regression line shown along with Pearson's correlation, r=0.69, P=.03. Graphs were created with GraphPad Prism version 9.0.0 for Mac (GraphPad Software, San Diego, CA; www.graphpad.com).

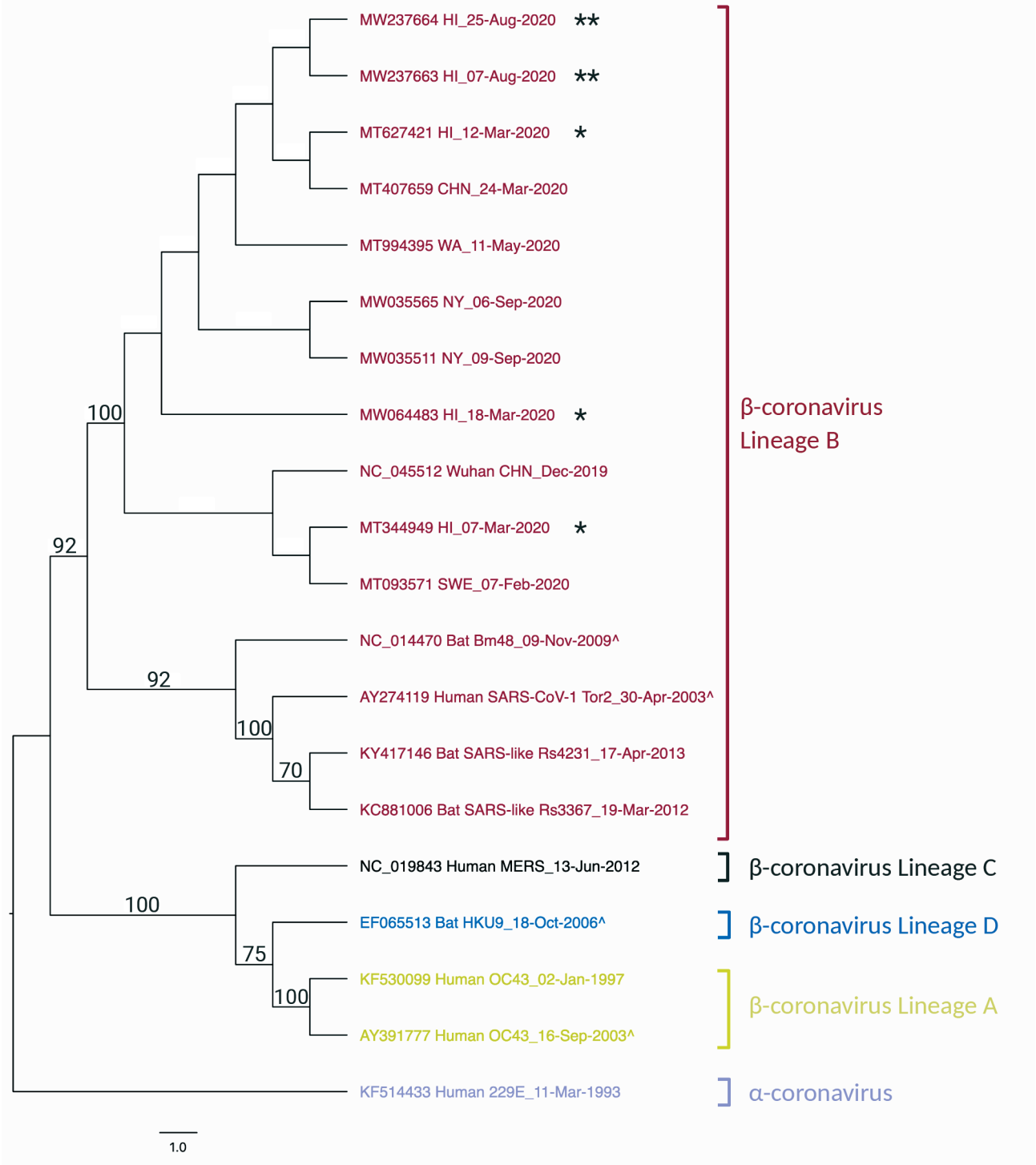

Figure 3. Maximum Likelihood Phylogenetic Tree Constructed Using a 969-bp S gene Region of SARS-CoV-2

The Tamura-Nei model and the Maximum Likelihood method were used to infer evolutionary history using 1000 bootstraps. The tree with the highest log-likelihood is shown in the figure. Next to the branches is shown the percentage of trees in which the associated taxa clustered together; only values greater than 70 are displayed. Neighbor-Join and BioNJ algorithms were applied to a matrix of pairwise distances estimated using the Tamura-Nei model to automatically obtain the initial tree for the heuristic search and then selecting the topology with superior log likelihood value. This analysis involves nucleotide sequences from 20 taxa. There were a total of 1366 final dataset positions. Evolutionary analyses were conducted in MEGAX19,20 using the University of Hawai'i MANA High-Performance Computing Cluster. The tree was rooted using the alphacoronavirus human 229E (KF514433) taxa in FigTree version 1.4.4. Yellow text is used for betacoronavirus lineage A, red text denotes betacoronavirus lineage B, black text denotes betacoronavirus lineage C, blue text shows betacoronavirus lineage D, and purple text is for alphacoronavirus.

* represents strains from Hawai'i. ** represents strains from this study. ^ indicates sequence submission date rather than collection date. Created with BioRender.com.

Discussion

This report focuses on SARS-CoV-2 strains from 2 patients in Hawai‘i based on a 969-bp S gene. The sequence and phylogenetic analysis indicate and support that the S gene is continuously mutating as previously reported,[30] and Hawai‘i may be harboring a unique strain with an emerging mutation in an altered spike protein. Analysis of the SNPs found in the 969-bp S gene region indicates the loss of a proline residue and the gain of cysteine residues. These mutations potentially alter the spike protein monomeric and trimeric structures.

The P681H represents the loss of a proline residue and the gain instead of an imidazole-containing histidine residue. According to the GISAID database, the P681H mutation is found worldwide in 65 959 strains reported as of January 31, 2021. The Pearson's correlation test of the logarithmic transformed P681H prevalence of the mutation versus time indicates that the P681H mutation is exponentially increasing worldwide, and the sequences encompassing the P681H mutation are dominating significantly when compared to other SARS-CoV-2 strains. This significant finding suggests that there is a selective pressure in favor of this mutation.

A study looking at the effect of the loss of a proline residue in the spike protein of the mouse hepatitis virus, a coronavirus, observed altered pathology, fusion kinetics, and enhanced infectivity.[31] The same study also suggests that prolines in regions adjacent to RBDs may not be essential for fusion but may significantly change structure and function of the spike protein.[31] Recent SARS-CoV-2 studies indicate that the P681H mutation is immediately juxtaposed to the amino acid 682–685, furin cleavage site, identified at the S1/S2 linkage site, which predicted enhance systemic infection,[12,32] and increased membrane fusion.[11] Additionally, the proline in the P681H mutation is within the epitope found to be the highest-ranking B and T cell epitope based on the in silico long-term population-scale epitope prediction for vaccine development study.[11] Therefore, the sequence surrounding this P681H mutation is predictably the loci where the immune response is targeted, meaning this mutation could be the first identified SARS-CoV-2 mutation of antigenic evolution.[1] Further studies are warranted to analyze the pathogenicity and virulence of the newly identified P681H mutation seen in the Hawai‘i strains and whether this is a viral evasion mechanism to deter antibody recognition or another increase in fitness. Now that SARS-CoV-2 vaccines are available in the United States under the Food and Drug Administration-Emergency Use Authorization (FDA-EUA), it is critical to evaluate the epitope-altering mutations. These mutations could change the effectiveness of FDA-EUA SARS-CoV-2 vaccines that rely on the structure of the spike protein.[33,34]

Twelve full length sequences have been published from Hawai‘i (MT627420.1, MT627421.1, MW064481.1, MW064482.1, MW064483.1, MW064596.1, MW064825.1, MW064826.1, MW065225.1, MW190887.1, MT344948.1, MT344949.1). All of these original strains introduced to Hawai‘i were collected in March 2020, and 66.6% (8 of 12) have the D614G mutation. The D614G mutation has become universal throughout the SARS-CoV-2 strains.[10] The D614G mutation is known to enhance infectivity and replication and localize the virus to the upper respiratory tract to increase transmission.[10,17] Interestingly, several mutations (nucleotide position 241 C→T, position 3037 C→T, and position 14 408 C→T, etc) exist alongside the D614G mutation[10,17] Similarly, in both Hawai‘i SARS-CoV-2 strains, the P681H mutation is also present alongside D614G.[10,13,17] This observation indicates that similar to the D614G mutation, the P681H mutation is becoming globally prevalent among SARS-CoV-2 sequences.

The R577C, S680C, and F797C mutations depicted in Table 1 are also very prominent mutations in that they present possible new disulfide bridges forming within and around the RBD. The RBD possesses 8 cysteine residues, with disulfide bridges formed between amino acids 336:361, 379:432, 480:488, and 391:525.[7,14] The aforementioned cysteine mutations may interact with these known bridges or create new bridges. The F797C mutation is seen alone in the Sweden strain (MT093571.1). The R577C and S680C mutations are present together in the strain from Australia (MT451798.1). Similar to the P681H, the S680C mutation is also within the epitope region of the B and T cell epitope in silico prediction model for vaccine development.[11] Further studies are warranted to evaluate the disulfide bridge configurations and whether an odd number of cysteines in this region can result in a dynamic bridge or the addition of several cysteines can alter the spike protein structure. Such studies would help to understand if these mutations are evolutionary mechanisms that alter virulence, or perhaps influence fusion kinetics. The other non-synonymous SNPs found in this study (A522S, F543L, I584V, I726F, A771S, E780Q) are not as apparent in presenting drastic evolutionary change, but they too deserve further analysis.

Recently, the New and Emerging Respiratory Virus Threats Advisory Group (NERVTAG) group based out of London, England, has reported on a new SARS-CoV-2 variant (variant of concern [VOC] 202012/01).[35,36] NERVTAG reports that the variant has increased transmissibility, and further studies are underway to confirm their report.[37] This variant includes amino acid mutations in ORF1ab, spike, Orf8, and N.[35,37,38] The 6 ORF1ab mutations are T1001I, A1708D, I2230T, and ΔS3675, ΔG2676, and ΔF3677.[35,37,38] The 10 spike mutations are ΔH69, ΔV70, ΔY145, N501Y, A570D, D614G, P681H, T716I, S982A, and D1118.[35,37,38] The Orf8 mutations are Q27stop, R52I, and Y73C.[37,38] The N protein mutations are D3L and S235F.[37,38] When comparing the SNPs encompassing the 969-bp of 2 strains from this study to the reference genome for VOC202012/01, EPI_ISL_601443,[39] we found 2 similar mutations, D614G and P681H.[37] Further, EPI_ISL_601443 shows the N501Y, A570D, D614G, P681H, and T716 mutations in the 969-bp region, while

the 2 Hawai‘i strains, MW237663 and MW237664, display the D614G and P681H mutations. Additionally, a new variant in Nigeria (B.1.1.207)(EPI_ISL_729975)[40] has been defined by the P681H mutation found in the 2 Hawai‘i strains.

The 2 Hawai‘i strains analyzed in this study cluster together predictably due to the emerging P681H mutation. These 2 strains also cluster closely with a strain from China and a previously published Hawai‘i strain. Other previously published Hawai‘i strains cluster with SARS-CoV-2 strains from New York, Wuhan, Sweden, and China. These analyses and resultant phylogenetic tree indicate that the virus has likely been introduced to Hawai‘i through several sources.

Over the past year, SARS-CoV-2 worldwide has evolved and will continue to do so. As of this report's publication, 6 new SARS-CoV-2 variants have been reported from the United Kingdom (B.1.1.7/VOC 202012/01),[41] South Africa (B.1.351/501Y.V2),[41] Denmark (B.1.1.298/Mink Cluster V),[42,43] Nigeria (B.1.1.207),[41] Brazil (B.1.1.248/P.1),[44] and California (B.1.429/L452R).[45] This fast pace of evolutionary changes will affect pathogenicity of SARS-CoV-2[12,46] and warrants further in silico, in vitro, and in vivo studies.

In summary, COVID-19 in Hawai‘i and the pandemic originating in Wuhan in the 2019–2020 winter is still ongoing. The virus continues to mutate, and the effects and outcomes of several of these mutations have yet to be elucidated. This study demonstrates a partial sequence from the first SARS-CoV-2 strain possessing the P681H non-synonymous mutation. In Hawai‘i, Native Hawaiians and Pacific Islanders have a significantly high prevalence of SARS-CoV-2 compared to other ethnic minorities and whites.[47] Therefore, characterizing viral sequences from these minority groups is important to understand virus transmission and pathogenicity better.

Conflicts of Interest

None of the authors identify a conflict of interest.

Acknowledgments

This research was supported by a grant (P30GM114737) from the Pacific Center for Emerging Infectious Diseases Research, Centers of Biomedical Research Excellence, National Institute of General Medical Sciences, National Institutes of Health (NIH); by a grant (U54MD007601) from Ola Hawai‘i, National Institute on Minority Health and Health Disparities, NIH; and by a contract (CT-MAY-2000282) from the City and County of Honolulu. We thank Dr. Sean Cleveland of Information Technology Services at the University of Hawai‘i for assistance with the phylogenetic tree, the University of Hawai‘i MANA High Performance Computing Cluster for use of the facility, and Dr. Vedbar Khadka for assistance with MEGAX. We thank the nurses and staff of the Hawai‘i Center for AIDS for assisting with the H051 study and the patients for participating in this study.

Authors' Affiliations:
- Department of Tropical Medicine, Medical Microbiology, and Pharmacology, John A. Burns School of Medicine, University of Hawai‘i at Mānoa, Honolulu, HI (DPM, LLC, CMS, VRN)
- Pacific Center for Emerging Infectious Diseases Research, John A. Burns School of Medicine, University of Hawai‘i at Mānoa, Honolulu, HI (DPM, LLC, VRN)
- Hawai‘i Center for AIDS, John A. Burns School of Medicine, University of Hawai‘i at Mānoa, Honolulu, HI (CMS)

Correspondence to:
Vivek R. Nerurkar PhD; Department of Tropical Medicine, Medical Microbiology and Pharmacology, John A. Burns School of Medicine, 651 Ilalo Street, BSB 320, Honolulu, HI 96813; E-mail: nerurkar@hawaii.edu

References

1. Day T, Gandon S, Lion S, Otto SP. On the evolutionary epidemiology of SARS-CoV-2. *Curr Biol*. 2020;30(15):R849–R857. doi:10.1016/j.cub.2020.06.031
2. Wang C, Liu Z, Chen Z, et al. The establishment of reference sequence for SARS-CoV-2 and variation analysis. *J Med Virol*. 2020;92(6):667–674. doi:10.1002/jmv.25762
3. Zhou P, Yang X-L, Wang X-G, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature*. 2020;579(7798):270–273. doi:10.1038/s41586-020-2012-7
4. Gaurav S, Pandey S, Puvar A, et al. Identification of unique mutations in SARS-CoV-2 strains isolated from India suggests its attenuated pathotype. *Microbiology*. 2020. doi:10.1101/2020.06.06.137604
5. Johns Hopkins Coronavirus Resource Center. Accessed February 10, 2021. https://coronavirus.jhu.edu/
6. Johns Hopkins Coronavirus Resource Center. Hawaii: COVID-19 Overview. Accessed February 10, 2021. https://coronavirus.jhu.edu/region/hawaii
7. Lan J, Ge J, Yu J, et al. Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. *Nature*. 2020;581(7807):215–220. doi:10.1038/s41586-020-2180-5
8. Yadav P, Potdar V, Choudhary M, et al. Full-genome sequences of the first two SARS-CoV-2 viruses from India. *Indian J Med Res*. 2020;0(0):0. doi:10.4103/ijmr.IJMR_663_20
9. Ching L, Chang SP, Nerurkar VR. COVID-19 special column: principles behind the technology for detecting SARS-CoV-2, the cause of COVID-19. *Soc Welf*. 2020;79(5):7.
10. Plante JA, Liu Y, Liu J, et al. Spike mutation D614G alters SARS-CoV-2 fitness. *Nature*. Published online October 26, 2020. doi:10.1038/s41586-020-2895-3
11. Yarmarkovich M, Warrington JM, Farrel A, Maris JM. Identification of SARS-CoV-2 vaccine epitopes predicted to induce long-term population-scale immunity. *Cell Rep Med*. 2020;1(3):100036. doi:10.1016/j.xcrm.2020.100036
12. Huang Y, Yang C, Xu X, Xu W, Liu S. Structural and functional properties of SARS-CoV-2 spike protein: potential antivirus drug development for COVID-19. *Acta Pharmacol Sin*. 2020;41(9):1141–1149. doi:10.1038/s41401-020-0485-4
13. Callaway E. Making sense of coronavirus mutations. Different SARS-CoV-2 strains haven't yet had a major impact on the course of the pandemic - but they might in future. *Nature*. 2020;585:174–177.
14. Hati S, Bhattacharyya S. Impact of thiol–disulfide balance on the binding of Covid-19 spike protein with angiotensin-converting enzyme 2 receptor. *ACS Omega*. 2020;5(26):16292–16298. doi:10.1021/acsomega.0c02125
15. Donoghue M, Hsieh F, Baronas E, et al. A novel angiotensin-converting enzyme–related carboxypeptidase (ACE2) converts angiotensin i to angiotensin 1-9. *Circ Res*. 2000;87(5). doi:10.1161/01.RES.87.5.e1
16. Clarke NE, Turner AJ. Angiotensin-converting enzyme 2: the first decade. *Int J Hypertens*. 2012;2012:1–12. doi:10.1155/2012/307315
17. Korber B, Fischer WM, Gnanakaran S, et al. Tracking changes in SARS-CoV-2 spike: evidence that D614G increases infectivity of the COVID-19 virus. *Cell*. 2020;182(4):812–827.e19. doi:10.1016/j.cell.2020.06.043
18. Yuan Y, He J, Gong L, et al. Molecular epidemiology of SARS-CoV-2 clusters caused by asymptomatic cases in Anhui Province, China. *BMC Infect Dis*. 2020;20(1). doi:10.1186/s12879-020-05612-4
19. Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA X: Molecular evolutionary genetics analysis across computing platforms. Battistuzzi FU, ed. *Mol Biol Evol*. 2018;35(6):1547–1549. doi:10.1093/molbev/msy096
20. Stecher G, Tamura K, Kumar S. Molecular evolutionary genetics analysis (MEGA) for macOS. Russo C, ed. *Mol Biol Evol*. 2020;37(4):1237–1239. doi:10.1093/molbev/msz312
21. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. 2004;32(5):1792–1797. doi:10.1093/nar/gkh340
22. Shu Y, McCauley J. GISAID: Global initiative on sharing all influenza data – from vision to reality. *Eurosurveillance*. 2017;22(13). doi:10.2807/1560-7917.ES.2017.22.13.30494
23. Elbe S, Buckland-Merrett G. Data, disease and diplomacy: GISAID's innovative contribution to global health: Data, Disease and Diplomacy. *Glob Chall*. 2017;1(1):33–46. doi:10.1002/gch2.1018
24. Team J. *JASP (Version 0.14)[Computer Software]*. 2020. https://jasp-stats.org/
25. Team Rs. *RStudio: Integrated Development for R. RStudio*. 2020. http://www.rstudio.com/.
26. Rambaut A. *FigTree*. 2018. http://tree.bio.ed.ac.uk/software/figtree/

27. Roychoudhury P, Hong X, Jerome K, Greninger A. *Virus Name: HCoV-19/USA/WA-UW-555/2020 / Accession ID: EPI_ISL_430887*. UW Virology Lab; 2020. https://www.epicov.org/epi3/frontend#3ba4e8
28. Chan JF-W, Yuan S, Kok K-H, et al. A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. *The Lancet*. 2020;395(10223):514–523. doi:10.1016/S0140-6736(20)30154-9
29. Su S, Wong G, Shi W, et al. Epidemiology, genetic recombination, and pathogenesis of coronaviruses. *Trends Microbiol*. 2016;24(6):490–502. doi:10.1016/j.tim.2016.03.003
30. Li Q, Wu J, Nie J, Li X, Huang W, Wang Y. The impact of mutations in SARS-CoV-2 spike on viral infectivity and antigenicity. *Cell*. 2020;182:1284–1294. doi:10.1016/j.cell.2020.07.012
31. Singh M, Kishore A, Maity D, et al. A proline insertion-deletion in the spike glycoprotein fusion peptide of mouse hepatitis virus strongly alters neuropathology. *J Biol Chem*. 2019;294(20):8064–8087. doi:10.1074/jbc.RA118.004418
32. Wang Q, Qiu Y, Li J-Y, Zhou Z-J, Liao C-H, Ge X-Y. A unique protease cleavage site predicted in the spike protein of the novel pneumonia coronavirus (2019-nCoV) potentially related to viral transmissibility. *Virol Sin*. 2020;35(3):337–339. doi:10.1007/s12250-020-00212-7
33. Chung JY, Thone MN, Kwon YJ. COVID-19 vaccines: The status and perspectives in delivery points of view. *Adv Drug Deliv Rev*. Published online December 2020:S0169409X20302829. doi:10.1016/j.addr.2020.12.011
34. Pfizer. *Pfizer and BioNTech Choose Lead mRNA Vaccine Candidate Against COVID-19 and Commence Pivotal Phase 2/3 Global Study*. Accessed January 2, 2021. https://www.pfizer.com/news/press-release/press-release-detail/pfizer-and-biontech-choose-lead-mrna-vaccine-candidate-0
35. GISAID. *UK Reports New Variant, Termed VUI 202012/01*. GISAID; 2020. https://www.gisaid.org/references/gisaid-in-the-news/uk-reports-new-variant-termed-vui-20201201/
36. NERVTAG, Horby P, Barclay W, et al. *NERVTAG Meeting on SARS-CoV-2 Variant under Investigation VUI-202012/01*. New and Emerging Respiratory Virus Threats Advisory Group; 2020. https://khub.net/documents/135939561/338928724/SARS-CoV-2+variant+under+investigation%2C+meeting+minutes.pdf/962e866b-161f-2fd5-1030-32b6ab467896?t=1608470511452
37. *Investigation of Novel SARS-CoV-2 Variant - Variant of Concern 202012/01*. Public Health England https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/947048/Technical_Briefing_VOC_SH_NJL2_SH2.pdf
38. Rambaut A, Loman N, Pybus O, et al. *Preliminary Genomic Characterisation of an Emergent SARS-CoV-2 Lineage in the UK Defined by a Novel Set of Spike Mutations*. COVID-19 Genomics Consortium UK; 2020. https://virological.org/t/preliminary-genomic-characterisation-of-an-emergent-sars-cov-2-lineage-in-the-uk-defined-by-a-novel-set-of-spike-mutations/563
39. Alderton A, Amato R, Goncalves S, et al. *Virus Name: HCoV-19/England/MILK-9E05B3/2020/ Accession ID: EPI_ISL_601443*. Lighthouse Lab in Milton Keynes; 2020. Accessed December 28, 2020. https://www.epicov.org/epi3/frontend#57fc35
40. Oluniyi PE. *Virus Name: HCoV-19/Nigeria/OS-CV296/2020/Accession ID: EPI_ISL_729975*. Nigeria Centre for Disease Control (NCDC); 2020. https://www.epicov.org/epi3/frontend#376f5
41. CDC. Coronavirus Disease 2019 (COVID-19): Emerging SARS-CoV-2 Variants. Centers for Disease Control and Prevention. Published December 30, 2020. Accessed December 30, 2020. https://www.cdc.gov/coronavirus/2019-ncov/more/science-and-research/scientific-brief-emerging-variants.html
42. Oude Munnink BB, Sikkema RS, Nieuwenhuijse DF, et al. Transmission of SARS-CoV-2 on mink farms between humans and mink and back to humans. *Science*. Published online November 10, 2020:eabe5901. doi:10.1126/science.abe5901
43. WHO | SARS-CoV-2 mink-associated variant strain – Denmark. WHO. Accessed January 2, 2021. http://www.who.int/csr/don/03-december-2020-mink-associated-sars-cov2-denmark/en/
44. Faria NR, Claro IM, Candido D, et al. Genomic characterisation of an emergent SARS-CoV-2 lineage in Manaus: preliminary findings - SARS-CoV-2 coronavirus / nCoV-2019 Genomic Epidemiology. Virological. Published January 12, 2021. Accessed January 28, 2021. https://virological.org/t/genomic-characterisation-of-an-emergent-sars-cov-2-lineage-in-manaus-preliminary-findings/586
45. Gangavarapu K, Basler T, Shephard J, Austin B. Virus Name: HCoV-19/USA/CA-AL-SR-4704/2020/Accession ID: EPI_ISL_730092. San Diego County Public Health Laboratory; Andersen lab at Scripps Research; 2020. Accessed January 28, 2021. https://www.epicov.org/epi3/frontend#6dc3b
46. Gussow AB, Auslander N, Faure G, Wolf YI, Zhang F, Koonin EV. Genomic determinants of pathogenicity in SARS-CoV-2 and other human coronaviruses. *Proc Natl Acad Sci*. 2020;117(26):15193–15199. doi:10.1073/pnas.2008176117
47. Hawaii COVID-19 Data: Which Racial and Ethnic Groups Have Been Most Affected? State of Hawai'i - Department of Health Accessed March 4, 2021. https://health.hawaii.gov/coronavirusdisease2019/what-you-should-know/current-situation-in-hawaii/#race