Algorithm for the Quantitation of Variants of Concern for Rationally Designed Vaccines Based on the Isolation of SARS-CoV-2 Hawai'i Lineage B.1.243

David P. Maison, M.S.[1,2,3],

Lauren L. Ching, B.S.[1,2,3],

Sean B. Cleveland, Ph.D.[4,5],

Alanna C. Tseng, Ph.D.[1,2,3],

Eileen Nakano, Ph.D.[1,2,3],

Cecilia M. Shikuma, M.D.[1,3,6],

Vivek R. Nerurkar, Ph.D.[1,2,3]

[1]Department of Tropical Medicine, Medical Microbiology, and Pharmacology,

[2]Pacific Center for Emerging Infectious Diseases Research,

[3]John A. Burns School of Medicine, University of Hawai'i at Mānoa, Honolulu, Hawaii 96813

[4]Hawaii Data Science Institute,

[5]Department of Cyberinfrastructure, University of Hawai'i at Mānoa, Honolulu, Hawaii 96822

[6]Hawaii Center for AIDS

Corresponding Author:

Vivek R. Nerurkar, Ph.D.

Department of Tropical Medicine, Medical Microbiology and Pharmacology

John A. Burns School of Medicine

651 Ilalo Street, BSB 320

Honolulu, Hawaii 96813

Telephone: (808) 692-1668

E-mail: nerurkar@hawaii.edu

Authors Contributions:

Clinical Studies, C.M.S.;

Conceptualization, D.P.M. and V.R.N.;

Data curation, D.P.M., L.L.C., A.C.T., E.N.;

Formal analysis, D.P.M., L.L.C., S.B.C., and V.R.N.;

Funding acquisition, V.R.N.

Investigation, D.P.M. and V.R.N.;

Project administration, V.R.N.;

Resources, D.P.M. and V.R.N.;

Software, D.P.M. and S.B.C.;

Supervision, V.R.N.;

Validation, D.P.M. and A.C.T.;

Visualization, D.P.M. and L.L.C.;

Writing - original draft preparation, D.P.M. and V.R.N;

Writing - review and editing, D.P.M., L.L.C., S.B.C., A.C.T., E.N., C.M.S., and V.R.N.;

Abstract

SARS-CoV-2 worldwide emergence and evolution has resulted in variants containing mutations resulting in immune evasive epitopes that decrease vaccine efficacy. We acquired clinical samples, analyzed SARS-CoV-2 genomes, used the most worldwide emerged spike mutations from Variants of Concern/Interest, and developed an algorithm for monitoring the SARS-CoV-2 vaccine platform. The algorithm partitions logarithmic-transformed prevalence data monthly and Pearson's correlation determines exponential emergence. The SARS-CoV-2 genome evaluation indicated 49 mutations. Nine of the ten most worldwide prevalent (>70%) spike protein changes have *r*-values >0.9. The tenth, D614G, has a prevalence >99% and *r*-value of 0.67. The resulting algorithm is based on the patterns these ten substitutions elucidated. The strong positive correlation of the emerged spike protein changes and algorithmic predictive value can be harnessed in designing vaccines with relevant immunogenic epitopes. SARS-CoV-2 is predicted to remain endemic and continues to evolve, so must SARS-CoV-2 monitoring and next-generation vaccine design.

Introduction

Since the origin of the Coronavirus Disease 2019 (COVID-19) pandemic, severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2) has rapidly evolved into seven Variants of Interest (VOI) and four Variants of Concern (VOC).[1] Further, as of July 15, 2021, over 2,559,000 SARS-CoV-2 genomic sequences have been deposited in the publicly available GenBank and the Global Initiative on Sharing Avian Influenza Data (GISAID) databases.[2] From the establishment of the now universal D614G substitution[3] to the emergence of the VOC and VOI with dozens of different mutations across their respective genomes,[1] the SARS-CoV-2 evolution, and adaptations, are apparent and constant. To give nomenclature to these evolutionary events, the Centers for Disease Control and Prevention (CDC) has classified certain lineages as VOC and VOI to denote highly adapted and immunologically evasive strains of SARS-CoV-2.[1] More recently, the World Health Organization (WHO) has further given its own classification to emerging SARS-CoV-2 lineages using letters of the Greek alphabets.[4]

Fortunately, early in the pandemic, governments and private sectors around the globe poured resources into producing efficacious vaccines. In the United States, three of these vaccines are authorized and recommended by the U.S. Food and Drug Administration (FDA).[5,6] Unfortunately, all these vaccines have reduced efficacy against all VOC,[1] an effect likely to amplify further as the virus evolves significantly from the vaccine design of the original strain. The loss of efficacy can be attributed to the alteration of immunogenic epitopes.[7] Several of these mutations are found in the spike protein, the protein used in the vaccine design, and therefore allows the virus to evade antibodies targeted to the original strain that vaccines utilize. Similar to the annual influenza virus vaccine, the evolution of SARS-CoV-2 presents the dilemma of how to redesign next-generation vaccines to keep up with the evolution of the virus.

One attempt to match the vaccine to the evolution of the virus was by Moderna. In response to the B.1.351 VOC considerably reducing the efficacy of the Novavax vaccine,[8] Moderna explored the use of the B.1.351 VOC sequence in their mRNA vaccine design.[9] Promisingly, the newly adapted vaccine increased neutralization against the B.1.351 VOC when given as a booster.[10] However, the B.1.351 was only 1.15% prevalent worldwide in April 2021. Therefore, the continuous clinical trial evaluation against emerging VOC is not practical to match the rate and diversity with which mutations and VOC emerge worldwide.

Hawai'i has been disproportionately affected by COVID-19 in terms of race, wherein 20% of the cases occur in 4% of the population of Pacific Islanders.[11,12] Understanding the SARS-CoV-2 lineage discrepancy in Hawai'i will allow for a greater understanding of the pandemic's nature worldwide. Additionally, adapting vaccines to match the nature of the viral sequence may alleviate these discrepancies. All four of the VOC recognized by the CDC are present in Hawai'i.[13]

To answer the dilemma of redesigning next-generation SARS-CoV-2 vaccines, we present and further validate our archetype quantitative analysis[14] for determining the emergence of individual mutations and variants, alike, as an algorithm. This algorithm is a platform for monitoring the virus and determining appropriate vaccine design as we proceed into the evolution and endemicity of SARS-CoV-2.[15] We utilized SARS-CoV-2 isolated in Hawai'i, and whole genome sequences (WGS) deposited in GenBank and GISAID, in combination with VOC and VOI to validate the algorithm, a prototype alpha test for rationally-designing logical next-generation vaccines. As SARS-CoV-2 evolves, so must SARS-CoV-2 monitoring and vaccine design.

Methods

Patient Samples

Human clinical samples analyzed in this report were part of the University of Hawai'i at Manoa (UHM) approved IRB - H051 study (# 2020-00367) (#NCT04360551). The samples were from two patients[14] (patient identification [PID] 498 and PID 708) collected as oropharyngeal (OPS) and nasal (NS) swabs at days 5 and 3, following symptom onset, respectively. SARS-CoV-2 positive status was confirmed using quantitative reverse-transcriptase-polymerase chain reaction (qRT-PCR). The samples were stored at -80℃ as part of the UHM IBC approved study (# 20-04-830-05).

Virus Isolation

Virus isolation was conducted using Vero E6 (ATCC CRL-1686) cells and from PID 498 OPS collected in the viral transport medium (VTM), as described previously.[16] Briefly, following 1 hour infection, cells were monitored for cytopathic effect (CPE) using a Cytosmart Microscope monitoring the same location in the flask. After observing significant CPE at 48 hours, supernatant was blind passaged three times in the Vero E6 cells. Virus isolation was confirmed with plaque assay using a double overlay, performed as previously described.[17]

Growth Kinetics

Following isolation of SARS-CoV-2, Isolate USA-HI498 2020, a growth kinetics study was conducted by seeding monolayers of Vero E6 cells in 6 well plates at a cell density of $3x10^5$/well one day prior to the assay. For the assay, cells were infected with SARS-CoV-2 USA-WA1/2020 and HI498 2020 isolates. Briefly, on the day of the assay, DMEM in 10% FBS was removed from wells with Vero cells, wells were washed twice with serum-free DMEM, and inoculated with multiplicity of infection (MOI) 0.1 and 1 virus isolates diluted in 500 µL DMEM in 2% FBS and incubated at 37ºC and 5% CO_2 for two hours. Following the two-hour adsorption, the infectious

supernatant with virus was removed, and monolayers were washed twice with DMEM in 2% FBS, and further incubated in 2 mL DMEM with 2% FBS at 37ºC and 5% CO_2 until supernatant collection at 0, 12, 24, and 48 hours.[16]

RNA Extraction, qRT-PCR, and Genomic Equivalence

For determining Genomic Equivalence, RNA extraction was conducted using the QIAamp® Viral RNA Mini Kit (Qiagen, Cat# 52906) following the manufacturer's instructions and as described previously.[14] RNA extraction was conducted on eight, ten-fold serial dilutions (10^0 to 10^{-8}) independently.[17] The primers and probes (N1 set, N2 set, and RdRp set) used are described previously.[18,19] A TaqMan® multiplexed qRT-PCR method was used for the N1 and N2 primer sets. The QuantaBio qScript® XLT One-Step qRT-PCR Tough Mix (Cat# 95132) was used to conduct the qRT-PCR on a ABI StepOnePlus™ Real-Time PCR system. A SYBR Green qRT-PCR method was used for the RdRp primer set. The QuantaBio qScript® cDNA Synthesis Kit (Cat# 95047) and QuantaBio PerfeCTa® qPCR ToughMix™, Low ROX™ (Cat# 95114) was used to perform the qRT-PCR on a ABI StepOnePlus™ Real-Time PCR system.

Standard curves for the SARS-CoV-2 isolates by N1, N2, and RdRp genes were produced by plotting cycle threshold (Ct) values against corresponding plaque forming units (PFU) per mL evaluated by plaque assay for the eight ten-fold serial dilutions of SARS-CoV-2 virus isolates.[17] The standard curve produced from the ten fold-serial dilutions of the virus was used to interpolate the results of the growth kinetics experiment. All qRT-PCR and Genomic Equivalence data were analyzed and visualized using GraphPad Prism 9 Version 9.2.0.

Whole Genome Sequencing

For WGS, RNA extraction was conducted using the third blind passage with the QIAamp® Viral RNA Mini Kit (Qiagen, Cat# 52906) following the manufacturer's instructions, as described

previously.[14] Briefly, viral RNA was eluted in 60 µL of the elution buffer. RNA extraction was confirmed using the Takara RNA LA PCR Kit (CAT #RR012A) and previously reported primer sets.[14,20] RNA was reverse transcribed into cDNA using the Takara RNA LA PCR Kit (Cat #RR012A) according to the manufacturer's protocol but with an extension time of 90 minutes. WGS was conducted by the ASGPB Core, UHM. Briefly, libraries prepared as per the manufacturer's protocol (Illumina Document #1000000025416 v09) using Illumina DNA Prep kit (Cat #20018704) and Nextera XT indexes were sequenced using the MiSeq Reagent Kit v3 (600 cycle) (Cat #MS-102-3003) and an Illumina MiSeq sequencer.

Informatics

WGS reads were compiled using the UHM MANA High-Performance Computing Cluster (HPC). Raw fastq sequence files were evaluated by the FASTQC program[21] for technical sequence error. After confirming a lack of technical errors, low-quality sequences were filtered and trimmed from each read with Trimmomatic[22] using paired-end adapter sequence NexteraPE-PE (ILLUMINACLIP:NexteraPE-PE:2:30:10) and a sliding 4 base window evaluating for quality with a PHRED score over 30. Trimmed result quality was confirmed with FASTQC. The trimmed-paired-end reads were then mapped to the NC_045512 reference genome using Bowtie2[23] and variants called with samtools mpileup[24] and transformed from VCF to FASTQ using bcftools and vcfutils[25] and finally converted to FASTA using seqtk. For comparison and validation, the fastq file was also inputted into Geneious Prime 2021.1.1 to produce FASTA files with the coronavirus assembly workflow.[26] The resultant consensus sequence was defined as Hawai'i Isolates and submitted to GenBank (SARS-CoV-2, Isolate USA-HI498 2020 (MZ664037) and SARS-CoV-2, Isolate USA-HI708 2020 (MZ664038)). The lineage of each sequence was determined with the Phylogenetic Assignment of Named Global Outbreak (PANGO) Lineage nomenclature.[27–29]

Hawai'i Sequences and Lineage Searches

All Hawai'i sequences as of July 28, 2021, from both GenBank and GISAID were downloaded and searched for the presence of potential lineages of concern using PANGO lineage as described previously.[13,27–29]

Variant Comparison

A comparison was conducted to evaluate the mutations in B.1.243 compared to 12 other VOC, VOI, and variants as of May 12, 2021. The NCBI SARS-CoV-2 resources genomic reference sequence from Wuhan was used to define the S gene (NC_045512).[30] Each of the sequences underwent pairwise alignment with NC_045512 to define S gene mutations. The Hawai'i Lineage (B.1.243) sequence selections were SARS-CoV-2 HI498 and HI708. Sequences for VOC, VOI, and other variants that garnered attention throughout this pandemic were selected with criteria of earliest complete collection dates with unambiguous S gene sequences. Lineages used were: B.1.1.7 (United Kingdom VOC, EPI_ISL_601443)[31,32], B.1.1 (Nigeria variant, EPI_ISL_729975)[33,34], B.1.351 (South Africa VOC, EPI_ISL_712081)[35], B.1.1.298 (Denmark variant, EPI_ISL_616802)[36], B.1.427 (California VOI, EPI_ISL_1531901)[37], B.1.429 (California VOI, EPI_ISL_942929)[38], P.1 (Brazil/Japan VOC, EPI_ISL_792680)[39,40], P.2 (Brazil VOI, EPI_ISL_918536)[41], B.1.617.1 (India VOI, EPI_ISL_1372093)[42], B.1.617.2 (India VOC, EPI_ISL_1663516)[43], and B.1.525 (United Kingdom/Nigeria VOI, EPI_ISL_1739895).[44]

Quantitation of Variants and Amino Acid Substitution/Deletions in Comparison to Epitope Mapping of the Spike Protein

From the aforementioned variant comparison section, the selected S gene sequences underwent pairwise alignment with NC_045512 in SnapGene, and SNPs were identified. The SNPs were inputted into the SnapGene sequence feature and Nextclade[45] to determine amino

acid substitutions (AAS). Non-synonymous substitutions were confirmed in GISAID using the metadata for each accession number.

The PANGO Server was used to identify and confirm the lineage of each of the aforementioned thirteen strains described in the Variant Comparison section.[28,29] The lineages and their collective AAS were identified and individually searched within GISAID for worldwide prevalence from March 2020 - April 2021. Each lineage was filtered separately, as were AAS. Parameters for selection were sequences that included a full month, day, and year of collection. Each month's prevalence for each lineage and AAS was logarithmically transformed and evaluated against month as an interval value with Pearson's correlation to determine an exponential increase in worldwide emergence as described previously.[14] Pearson's was calculated using RStudio version 1.3.1093 (R version 4.0.3) and plotted with the ggplot2 package. The Pearson's correlations for AAS and lineages were then compared in a corresponding pairwise heat map to evaluate if AAS emergence occurs independently of, or in tandem to, lineage emergence.

Separately, the following search parameter was used in PubMed to locate *in silico* studies predicting vaccine epitopes to SARS-CoV-2: "((B-cell) OR (B cell)) AND ((T-cell) OR (T cell)) AND (peptide) AND (vaccine epitope) AND ((SARS-CoV-2) OR (COVID-19))." From this search on January 28, 2021, the three most recent articles[46–48] and the three best matching articles[49–51] were selected for further analysis by mapping to the Spike protein. All predicted epitopes able to be searched and defined with SnapGene's "Find Protein Sequence" feature were included. Article overlaps in the systematic review were only included once.

Algorithm

The algorithm herein described was developed from the quantitation of the ten most emerged amino acid substitutions and deletions. The algorithm is as follows:

if Pearson's *r*-value ≥ 0.9:

    if Previous Month's Prevalence > 0.3:

        Emerging (Mutation of Concern) (Include in Next Generation Vaccine Design)

    else if 0.02 ≤ Previous Month's Prevalence ≤ 0.3:

        Emerging (Mutation of Interest)

    else:

        Not Emerging

else:

    if Previous Month's Prevalence ≥ 0.9:

        Emerged (Mutation of Concern) (Include in Next Generation Vaccine Design)

    else if Previous Month's Prevalence > 0.5:

        Emerged (Mutation of Interest)

    else:

        Not Emerged/Emerging

This algorithm classifies AAS and deletions into two categories based on Pearson's *r*-value, $r \geq 0.9$ and $r < 0.9$. For AAS and deletions to be called out as concerning, the *r*-value should be ≥0.9 and the previous month's worldwide prevalence of these AAS and deletions should be >30%. Further, these concerning AAS and deletions can be considered for inclusion in the next-generation vaccine design. If the $r \geq 0.9$ and the previous month's prevalence is between 2% and 30%, then the AAS or deletion is classified as interesting, and needs to be evaluated in a research setting. The same algorithm can also be applied for standardizing classifications of SARS-CoV-2 lineages as of interest or concerning.

If the r<0.9, then the focus is on previous month's prevalence of AAS and deletions, as after a mutation is established, there is no longer need to evaluate emergence. If the previous month's prevalence is ≥90%, then the mutation is established in the SARS-CoV-2 genome and should be considered as concerning and be part of the next-generation vaccine. If the previous month's prevalence is >50%, then the mutation represents the majority, and needs to be considered as interesting and evaluated in a research setting. Again, the same algorithm can also be applied for standardizing classifications of SARS-CoV-2 lineages as of interest or concerning.

B.1.243 Phylogeny and Origin Tracking

Origin tracking was accomplished as described previously.[13] The sequences from Hawai'i were obtained through both GenBank and GISAID, along with SARS-CoV-2, Isolate USA-HI498 2020 and SARS-CoV-2, Isolate USA-HI708 2020 (from this study), to accomplish the origin determination.

Results

Virus Isolation, Growth Kinetics, and Genomic Equivalence

Virus was isolated from an OPS collected five days following symptom onset from an individual with PCR confirmed SARS-CoV-2 infection and propagated in Vero E6 cells. A stock of the SARS-CoV-2, Isolate USA-HI498 2020 was produced following three blind passages in Vero E6 cells and titered to 1.28 x 10^7 PFU/mL. Minimal CPE was observed at 12 hours, moderate CPE at 24 hours, and significant CPE at 48 hours (Figure 1 A-D). Plaque assay confirmed SARS-CoV-2, Isolate USA-HI498 virus isolation at 1.28 x 10^7 PFU/mL and SARS-CoV-2 USA-WA1/2020 at 3.88 x 10^7 PFU/mL. Viral copy number analysis using N1, N2 and RdRp primers as well as microscopic observation showed no significant differences between the SARS-CoV-2, Isolate USA-HI498 2020 and SARS-CoV-2 USA-WA1/2020 (Figure 1 E-M).

Whole Genome Sequencing and Informatics

RNA extraction was confirmed using RT-PCR as described previously.[14] There were 702,978 and 792,952 reads for SARS-CoV-2, Isolate USA-HI498 2020 and HI708, respectively (Table 1). FastQC confirmed the quality of both the untrimmed and trimmed fastqc files.

Hawai'i SARS-Cov-2 Sequences and Lineage Searches

From GenBank, 317 full-genome SARS-CoV-2 sequences were obtained on July, 28, 2021. Further, an additional 2,942 sequences were obtained from GSAID. Hawai'i has 52 unique lineages in the 3,259 representative sequences (A.1, A.2.2, A.3, AY.1, AY.2, B, B.1, B.1.1, B.1.1.207, B.1.1.222, B.1.1.304, B.1.1.316, B.1.1.380, B.1.1.416, B.1.1.519, B.1.1.7, B.1.108, B.1.139, B.1.160, B.1.2, B.1.234, B.1.241, B.1.243, B.1.265, B.1.298, B.1.340, B.1.351, B.1.357, B.1.36.8, B.1.369, B.1.37, B.1.400, B.1.413, B.1.427, B.1.429, B.1.517, B.1.526, B.1.561, B.1.568, B.1.575, B.1.588, B.1.595, B.1.596, B.1.601, B.1.609, B.1.617.2, B.1.623, B.6, P.1, P.1.1, P.2, R.1). As of April 12, 2021, GISAID reported 8,809 sequences of B.1.243 lineage worldwide. The B.1.243 variant was represented by 717 of the 1,002 (72%) sequences curated from Hawai'i. This prevalence decreased to 23% by July 28, 2021. Worldwide, GISAID reported 8,809 sequences of B.1.243 lineage.

Quantitation of SARS-Cov-2 Variants and Amino Acid Substitution/Deletions in Comparison to Epitope Mapping of the Spike Protein

The output S gene alignment between the thirteen genomic sequences identified 49 SNPs. Pearson's correlation on logarithmically-transformed prevalence was calculated for the twelve SARS-CoV-2 variants in this study in order of highest to lowest *r* value as outlined in Table 2 and Figure 2.

Further, of the 49 identified SNPs in the S gene, 44 resulted in non-synonymous AAS and deletions in the protein, and 5 were synonymous as outlined in Figure 3A, 3B, 3Ci-xii and Tables 2 and 3. Pearson's correlation on logarithmically-transformed prevalence was calculated for all identified SARS-CoV-2 AAS and deletions (Tables 2, 3, Figure 4, and Supplementary Figure 1).

The PubMed search for epitope predictions returned 42 publications. In total, 393 *in silico* predicted B and T cell epitopes corresponding to the spike protein were mapped from these publications (Figure 3B). Of these, 108 epitopes involved the N-terminal domain (NTD), 102 epitopes involved the receptor binding domain (RBD), 7 epitopes involved the S1/S2 furin cleavage site, 10 epitopes involved the fusion peptide, 20 involved heptad repeat 1, 12 involved heptad repeat 2, 12 involved the transmembrane region, and 8 involved the intracellular tail domain. The remaining 112 epitopes fell outside of these domains. Further, 239 and 151 epitopes were in the S1 and S2, respectively, with at least one predicted epitope covering 97% of the spike protein.

Variant Comparison

The unique nucleotide mutations and resulting AAS and deletions for each of the twelve SARS-CoV-2 variants, in comparison to the reference sequence, are shown in Table 3. BNT162b2 (Pfizer) and mRNA-1273 (Moderna) vaccines both contain two AAS (K986P and V987P) (Figure 3C.xii).[52] Novavax and Janssen vaccines also contain two AAS, K986P and V987P. Further, additional AAS at the furin cleavage site includes, R862Q, R683Q, and R685Q (Novavax), and R682S and R685G (Janssen) (Figure 3C.xiii).[52]

B.1.243 Phylogeny and Origin Tracking

The Hawai'i SARS-CoV-2 sequences in the GenBank and GISAID were combined with all worldwide B.1.243 lineages to produce an initial MAFFT alignment of 8,820 sequences. Further,

4,273 sequences with ambiguities and 1,596 duplicate sequences were removed. The final alignment for constructing the phylogenetic tree was 2,953 unique and unambiguous B.1.243 sequences. Using this method,[13] we were able to define the origin of SARS-CoV-2, Isolate USA-HI498 2020 and HI708 (Figure 5).

Discussion

In this report, we lay the foundation for an adaptive and rational algorithm for monitoring SARS-CoV-2 evolution, quantitating variants, substitutions, and deletions, in the context of the vaccine design. Further, we describe the isolation, genetic characterization, phylogenetic analysis, and immunogenetic epitopes of the spike protein based on the SARS-CoV-2 lineage B.1.243 from Hawai'i. We employed B.1.243 to establish and validate the algorithm, as well as analyze VOC and VOI in the context of emerging spike protein amino acid changes for surveillance and future vaccine design.

Hawai'i Isolate B.1.243

Hawai'i has not been spared from this pandemic, and the Pacific Islander population here is disproportionately infected with SARS-CoV-2 as compared to other races and ethnicities.[12] Following isolation and identification of the B.1.243 lineage from the isolate SARS-CoV-2, Isolate USA-HI498 2020, and virus strain HI708, we found through curation and analysis of published sequences that the B.1.243 lineage was once the dominating lineage in Hawai'i, causing more than 40% of all cases. The Hawai'i B.1.243 lineage is not likely to escalate into a VOC, as the prevalence has decreased worldwide over the past several months. However, the disproportionate infection rates among the Pacific Islander population remains.[12]

The B.1.243 lineage that once dominated in overall prevalence in Hawai'i was introduced from Washington, California, Pennsylvania, and New Mexico, with the majority of sequences arising

from horizontal transfer within Hawai'i. SARS-CoV-2, Isolate USA-HI498 2020 was introduced in Hawai'i from New Mexico and the HI-708 SARS-CoV-2 strain originated from California. As this report is written, similar to the continental United States the SARS-CoV-2 Delta variant is rapidly spreading in Hawaii.[53]

Quantitation and Analysis of Variants of Concern

At the time of the submission, the CDC identifies four SARS-CoV-2 variants as VOC: B.1.1.7 (Alpha), B.1.351 (Beta), B.1.617.2 (Delta), and P.1 (Gamma).[1] Our quantitative data analysis supports the exponential emergence of these VOC, with the most emergent being P.1, followed by B.1.617.2, B.1.351, and B.1.1.7, with Pearson's correlation *r*-values of 0.97, 0.96, 0.94, and 0.92, respectively. The quantitative analysis described in this report gives a numerical value to each VOC emergence, predicting the likelihood that the lineage will become prevalent and spread through the population. This value then indicates which VOC genomes are likely to possess evolutionarily selective changes. Previously, using this quantitative analysis, we have demonstrated that the P681H substitution, which had a prevalence of 2% worldwide in December 2020, then emerged to 79% in April 2021.[14] Similarly, using this quantitative analysis in April 2021, we predicted the spread of the Delta variant in Hawaii and worldwide as of June 2021.[13]

In this report we demonstrate the characteristic emergence and selective evolution of VOC using the B.1.1.7 VOC as an example, with a Pearson's correlation of 0.92. The B.1.1.7 VOC, the most prevalent VOC worldwide in April 2021, has spread across the globe after emerging in the United Kingdom in December 2020.[31] As stated, the B.1.1.7 VOC represents the prototypic VOC which has evolved continuously by evading vaccine sera and becoming more transmissible.[54,55] Similarly, the Delta variant, predicted to be exponentially emerging by this

quantitative analysis with an *r*-value of 0.96 as of April 2021 at 1% prevalence, has become the most prevalent worldwide as of June 2021, representing 64% of worldwide sequences.

The virus transmissibility, prevalence, and decrease of treatment and vaccine efficacy is concerning. Each of these viral properties has been attributed to specific amino acid alterations in the spike protein. For example, the L452R substitution, prevalent in many VOC and VOI, causes a two-fold increase in viral shedding and renders multiple FDA approved monoclonal antibody treatments ineffective.[56] Thus, the quantitation of VOC leads to identification and quantitation of their respective mutations. The pairwise heat map between variants and mutations (Table 2) indicates that variants do not necessarily evolve with mutations, and the genomes may spontaneously acquire or revert to wildtype, as demonstrated in other statistical analysis for monitoring this pandemic.[57]

The Algorithm

We were able to establish the algorithm described in this report based on the ten most emerged AAS and deletions. We evaluated nine of the AAS and deletions observed among the variants in this study, as of May 12, 2021, with $r > 0.9$ and >70% prevalence in April 2021. These AAS and deletions included, P681H, ΔV70, ΔH69, N501Y, S982A, D1118H, T716I, A570D, and ΔY144. The data show that the average time for an emerging substitution ($r > 0.9$) to go from >30% monthly prevalence percentage to >50% prevalence is 2.25 months. Extrapolating these findings, the timeframe for Pfizer to identify emerging changes and manufacture them into a new version of the BNT162b2 vaccine would be 60-110 days. This timeframe is roughly equivalent to the algorithm's predictive value.[58] Additionally, the tenth AAS used in the development of the algorithm was D614G that allowed us to discern the previous month prevalence, which is an important parameter, as once an emerging mutation is established in the genome the *r* value will decrease considerably. The nine substitutions and deletions also

display an average of 4.75 months from >2% monthly prevalence to >50% monthly prevalence. Therefore, $r > 0.9$ and a prevalence of >2% is sufficient to establish a mutation as being of interest, whereas 30% prevalence escalates a mutation to the status of concern. From the evaluation of the ten total spike protein changes, the algorithm concludes that an $r > 0.9$ and a >30% prevalence percentage is an optimal time to classify a substitution as concerning, and consider the substitution for inclusion into vaccine primary structure for 60 day production time.

These mutations of interest and concern can also serve to facilitate and focus research using infectious clone[59] and pseudoviruses[60] to determine the functional characteristics. Amino acid substitutions are responsible for, i) changing epitopes at a level to evade antibodies,[1,61,62] ii) allowing the virus to localize anatomically,[63] iii) increasing viral shedding,[62] iv) increasing binding affinity,[64] v) giving the virus binding protein the ability to change conformations more efficiently,[65] and vi) causing diagnostic false negatives.[66] Therefore, of great importance in vaccine development is identifying which spike protein changes are most prevalent worldwide across all sequenced genomes. Each substitution or deletion, or combination thereof, could potentially serve as an epitope as shown in the epitope map (Figure 3C). Booster vaccines using this algorithm can therefore prepare the vaccinated for any variant they are most likely to encounter by identifying and including the emerging and emerged amino acid changes representing the majority of SARS-CoV-2 worldwide.

In silico Predicted Immunogenic Epitopes in Relation to Variants of Concern

Epitopes are found across nearly the entire spike protein. The *in silico* compilation of predicted B cell and T cell epitopes demonstrated that 53% (210/393) of all epitopes occur in the NTD and RBD. This is consistent with an *in vivo* mRNA-LNP vaccine study that found a vast majority of the CD8+ T cell response target epitopes in the N-terminal portion of the Spike protein.[67] The same study found that CD4+ T cell responses target both S1 and S2.[67] Additionally, the majority

of variant mutations and neutralizing antibody targets occur in S1.[55] As S1 is shed in the coronavirus model of fusion,[68] and S2 is responsible for fusion, the diversity of S1 in SARS-CoV-2 and the epitope targeting concentration in S1, indicate that SARS-CoV-2 vaccines will need to adapt along with the virus.

Current Vaccines, Structures, and Introduced Mutations in Relation to Variants of Concern

The Pfizer/BioNTech and Moderna/National Institute for Allergy and Infectious Diseases (NIAID) are mRNA vaccines that both use the S gene of SARS-CoV-2 S-2P. S-2P is the reference sequence with the substitution of two prolines (K986P, V987P) to stabilize the pre-fusion conformation of the spike protein.[52] The Janssen and Novavax vaccines for SARS-CoV-2 are Ad26-vectored and protein-based, respectively, and use the S-2P sequence with the addition of three AAS (R682S/Q, R683Q, R685G/Q) at the furin cleavage site.[52] These furin AAS have been shown to further stabilize the pre-fusion spike conformation, and increase neutralizing antibody titers.[52] The majority of other vaccines, including whole virus vaccines, S protein vaccines, and RBD vaccines do not utilize the S-2P method of pre-fusion stabilization.[52,68] Current post-vaccination sera has reduced neutralization against the VOC as the COVID-19 pandemic proceeds into endemicity. As the loss of neutralization is observed against the most emerging VOC and VOI herein, there is a need for a new generation of vaccines for optimal efficacy in years to come.

Future Directions:

Our findings have relevance to the future of tracking SARS-CoV-2 and of SARS-CoV-2 vaccine design. The future SARS-CoV-2 vaccines are akin to influenza virus vaccines. That seeming nature is that influenza vaccines change yearly depending on the previous years surveillance

data.[69] If established in real-time, the herein described algorithm will allow researchers to understand the evolving SARS-CoV-2 genome preemptively rather than responsively.

As the number of worldwide SARS-CoV-2 sequences swells into many millions between GenBank and GISAID, the need for an Application Programming Interface (API) between the two databases is needed now. Such an API would allow the quantitation of emerging mutations and variants to alarm when necessary, rather than at arbitrary media discretion. Real-time quantitation would then allow vaccines to evolve preemptively. As things stand, the data is self-diversifying by researchers' choice of submission to one database or the other, rendering GISAIDs mutation filters less informative on a worldwide scale. As one database headquarters in Germany (GISAID) and the other in the United States (GenBank), the world needs solidarity now more than ever as we combat this global pandemic.

Conclusions:

Here, we isolate SARS-CoV-2 in Hawai'i and evaluate the WGS of these SARS-CoV-2 isolates. We apply our archetype method for predicting exponentially emerging mutations to these isolates, then evolve the quantitative analysis into an algorithm to evaluate the VOC, VOI, and their mutations, allowing further classification of mutations as concern and interest. This algorithm can now serve as a baseline for choosing the primary structure of vaccines. Additionally, we graphically compare the S gene of the Hawai'i isolate with SARS-CoV-2 VOCs, predict the emergence of SARS-CoV-2 S gene mutations and protein substitutions and deletions, and evaluate these substitutions and deletions in the context of epitopes. In conclusion, we create a foundation for future SARS-CoV-2 monitoring and vaccine efforts as we move forward in this pandemic (Figure 6), and demonstrate the need for sequence database solidarity. These pandemic efforts cannot remain in the context of being responsive and

reactive, but need also to be preemptive and predictive. Preemptive and predictive is possible with the approach herein.

Acknowledgements

This research was supported by a grant (P30GM114737) from the Pacific Center for Emerging Infectious Diseases Research, COBRE, a grant (P20GM103466-20S1) from the INBRE, National Institute of General Medical Sciences, and a grant (U54MD007601) from Ola Hawaii, National Institute on Minority Health and Health Disparities, NIH. The H051 clinical trial is registered at ClinicalTrials.gov (#NCT04360551). Computation was supported by NSF grant #1920304 on the University of Hawai'i MANA High Performance Computing Cluster. The viral genome sequences used in this publication are publicly available from GenBank (https://www.ncbi.nlm.nih.gov/sars-cov-2/) and GISAID (https://gisaid.org). Tables of acknowledgements for the genome sequences from GISAID are available at: https://github.com/dpmaison/Algorithm-for-the-Quantitation-of-Variants-of-Concern-for-Rationally-Designed-Vaccines. Other genome sequences from GISAID are referenced in-text. We thank Dr. Jennifer Saito at the Advanced Studies in Genomics, Proteomics and Bioinformatics (ASGPB) Core, UHM for assistance with WGS.

References

1. CDC. Cases, Data, and Surveillance. Centers for Disease Control and Prevention. Published February 11, 2020. https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/variant-surveillance/variant-info.html
2. 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
3. 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
4. Tracking SARS-CoV-2 variants. World Health Organization. Published July 6, 2021. Accessed July 16, 2021. https://www.who.int/activities/tracking-SARS-CoV-2-variants
5. CDC. Different COVID-19 Vaccines. Centers for Disease Control and Prevention. Published May 27, 2021. Accessed June 14, 2021. https://www.cdc.gov/coronavirus/2019-ncov/vaccines/different-vaccines.html
6. Food and Drug Administration. COVID-19 Vaccines. *FDA*. Published online July 12, 2021. Accessed July 12, 2021. https://www.fda.gov/emergency-preparedness-and-response/coronavirus-disease-2019-covid-19/covid-19-vaccines
7. Harvey WT, Carabelli AM, Jackson B, et al. SARS-CoV-2 variants, spike mutations and immune escape. *Nat Rev Microbiol*. 2021;19(7):409-424. doi:10.1038/s41579-021-00573-0
8. Shinde V, Bhikha S, Hoosain Z, et al. Efficacy of NVX-CoV2373 Covid-19 Vaccine against the B.1.351 Variant. *N Engl J Med*. 2021;384(20):1899-1909. doi:10.1056/NEJMoa2103055
9. Moderna Announces it has Shipped Variant-Specific Vaccine Candidate, mRNA-1273.351, to NIH for Clinical Study. *Bloomberg.com*.

https://www.bloomberg.com/press-releases/2021-02-24/moderna-announces-it-has-shipped-variant-specific-vaccine-candidate-mrna-1273-351-to-nih-for-clinical-study. Published February 24, 2021. Accessed April 23, 2021.

10. Moderna Announces Positive Initial Booster Data Against SARS-CoV-2 Variants of Concern | Moderna, Inc. Accessed June 22, 2021. https://investors.modernatx.com/news-releases/news-release-details/moderna-announces-positive-initial-booster-data-against-sars-cov/
11. *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
12. Cha L, Le T, Ve'e T, Ah Soon NT, Tseng W. Pacific Islanders in the Era of COVID-19: an Overlooked Community in Need. *J Racial Ethn Health Disparities*. Published online June 24, 2021. doi:10.1007/s40615-021-01075-8
13. Maison DP, Cleveland SB, Nerurkar VR. Genomic Analysis of SARS-CoV-2 Variants of Concern Circulating in Hawai'i to Facilitate Public-Health Policies. *Res Sq*. Published online June 9, 2021. doi:DOI: 10.21203/rs.3.rs-378702/v3
14. Maison DP, Ching LL, Shikuma CM, Nerurkar VR. Genetic Characteristics and Phylogeny of 969-bp S Gene Sequence of SARS-CoV-2 from Hawaii Reveals the Worldwide Emerging P681H Mutation. *Hawaii J Health Soc Welf*. 2021;80(3):52-61. doi:https://doi.org/10.1101/2021.01.06.425497
15. Phillips N. The coronavirus is here to stay — here's what that means. *Nature*. 2021;590(7846):382-384. doi:10.1038/d41586-021-00396-2
16. Lednicky JA, Shankar SN, Elbadry MA, et al. Collection of SARS-CoV-2 Virus from the Air of a Clinic within a University Student Health Care Center and Analyses of the Viral Genomic Sequence. *Aerosol Air Qual Res*. 2020;20(6):1167-1171.

doi:10.4209/aaqr.2020.05.0202

17. Johnson BW, Russell BJ, Lanciotti RS. Serotype-Specific Detection of Dengue Viruses in a Fourplex Real-Time Reverse Transcriptase PCR Assay. *J Clin Microbiol*. 2005;43(10):4977-4983. doi:10.1128/JCM.43.10.4977-4983.2005
18. *CDC 2019-Novel Coronavirus (2019-NCoV) Real-Time RT-PCR Diagnostic Panel*. Centers for Disease Control and Prevention Accessed September 1, 2020. https://www.fda.gov/media/134922/download
19. *Protocol: Real-Time RT-PCR Assays for the Detection of SARS-CoV-2*. World Health Organization Accessed September 1, 2020. https://www.who.int/docs/default-source/coronaviruse/real-time-rt-pcr-assays-for-the-detection-of-sars-cov-2-institut-pasteur-paris.pdf?sfvrsn=3662fcb6_2
20. Yuan Y, Jun H, Lei G, et al. Molecular epidemiology of SARS-CoV-2 clusters caused by asymptomatic cases in Anhui Province, China. *BMC Infect Dis*. 2020;20(1):1-13. doi:10.1186/s12879-020-05612-4
21. Andrews S. *FastQC: A Quality Control Tool for High Throughput Sequence Data [Online]*.; 2010. https://www.bioinformatics.babraham.ac.uk/people.html
22. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. 2014;30(15):2114-2120. doi:10.1093/bioinformatics/btu170
23. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods*. 2012;9(4):357-359. doi:10.1038/nmeth.1923
24. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009;25(16):2078-2079. doi:10.1093/bioinformatics/btp352
25. Danecek P, Bonfield JK, Liddle J, et al. Twelve years of SAMtools and BCFtools. *GigaScience*. 2021;10(2). doi:10.1093/gigascience/giab008
26. Miller H. Assembly of SARS-CoV-2 genomes from tiled amplicon Illumina sequencing using Geneious Prime. Geneious. Published November 29, 2020. Accessed July 28, 2021.

https://help.geneious.com/hc/en-us/articles/360045070991-Assembly-of-SARS-CoV-2-genomes-from-tiled-amplicon-Illumina-sequencing-using-Geneious-Prime

27. Rambaut A, Holmes EC, O'Toole Á, et al. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat Microbiol*. 2020;5(11):1403-1407. doi:10.1038/s41564-020-0770-5
28. *Pangolin (Version v.3.0.5)*. Centre for Genomic Pathogen Surveillance; 2021. pangolin.cog-uk.io
29. O'Toole Á, Scher E, Underwood A, et al. pangolin: lineage assignment in an emerging pandemic as an epidemiological tool. PANGO lineages. Published 2021. Accessed March 11, 2021. github.com/cov-lineages/pangolin
30. NCBI SARS-CoV-2 Resources. National Library of Medicine National Center for Biotechnology Information. Accessed January 25, 2021. https://www.ncbi.nlm.nih.gov/sars-cov-2/
31. *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
32. Davies R. *Virus Name: HCoV-19/England/MILK-9E05B3/2020 / Accession ID: EPI_ISL_601443*. Lighthouse Lab in Milton Keynes; 2020. Accessed January 26, 2021. https://www.epicov.org/epi3/frontend#3d0e48
33. 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
34. Oluniyi PE, Ihekweazu C, Nkengasong J, Olawoye I, Happi C. Detection of SARS-CoV-2 P681H Spike Protein Variant in Nigeria. Published online December 23, 2020. https://virological.org/t/detection-of-sars-cov-2-p681h-spike-protein-variant-in-nigeria/567

35. Bhiman JN. *Virus Name: HCoV-19/South Africa/N00390/2020 / Accession ID: EPI_ISL_712081*. Port Elizabeth Provincial Hospital, National Health Laboratory Services; National Institute for Communicable Diseases of the National Health Laboratory Service; 2020. Accessed January 26, 2021. https://www.epicov.org/epi3/frontend#2d8dc1

36. Michaelsen TY. *Virus Name: HCoV-19/Denmark/DCGC-3024/2020 / Accession ID: EPI_ISL_616802*. Department of Virus and Microbiological Special Diagnostics; Albertsen lab, Department of Chemistry and Bioscience; 2020. Accessed January 28, 2021. https://www.epicov.org/epi3/frontend#4cb45e

37. Gangavarapu K. *Virus Name: HCoV-19/Mexico/BCN-ALSR-8438/2020 / Accession ID: EPI_ISL_1531902*. Centro de Diagnostico COVID-19 UABC Tijuana; 2020. https://www.epicov.org/epi3/frontend#1da1c7

38. Rodriguez AP. *Virus Name: HCoV-19/Mexico/ROO-InDRE_243/2020 / Accession ID: EPI_ISL_942929*. Instituto de Diagnostico y Referencia Epidemiologicos INDRE_RNLSP; 2020. Accessed February 14, 2021. https://www.epicov.org/epi3/frontend#31e638

39. Sekizuka T. *Virus Name: HCoV-19/Japan/IC-0561/2021 / Accession ID: EPI_ISL_792680 / P.1*. GISAID; 2021. Accessed January 28, 2021. https://www.epicov.org/epi3/frontend#524009

40. 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

41. Santos MC, Silva AM, Junior WDC, et al. *Virus Name: HCoV-19/Brazil/CE-IEC-177339/2020 / Accession ID: EPI_ISL_918536 / P.2*. LACEN - Laboratório Central de Saúde Pública do Ceara; 2021. Accessed January 28, 2021.

https://www.epicov.org/epi3/frontend#48b353

42. Raghav SK, Walia S, Ghosh A, et al. *Virus Name: HCoV-19/India/ILSGS00308/2020 / Accession ID: EPI_ISL_1372093*. Immunogenomics lab, Institute of Life Sciences, Bhubaneswar; 2020. Accessed May 11, 2021. https://www.epicov.org/epi3/frontend#54b2bf
43. Raghav SK. *Virus Name: HCoV-19/India/ILSGS00941/2020 / Accession ID: EPI_ISL_1663516*. Institute of Life Sciences - INSACOG; 2020. https://www.epicov.org/epi3/frontend#3b4b31
44. McArthur A. *Virus Name: HCoV-19/Canada/ON-SC4161/2020/ Accession ID: EPI_ISL_729975*. Ontario's COVID-19 Genomics Rapid Response Coalition; 2020. https://www.epicov.org/epi3/frontend#10db8c
45. Aksamentov I, Rubinsteyn A, Hodcroft E, et al. Nextclade Web 1.5.0. Accessed July 2, 2021. https://clades.nextstrain.org
46. Singh J, Malik D, Raina A. Immuno-informatics approach for B-cell and T-cell epitope based peptide vaccine design against novel COVID-19 virus. *Vaccine*. 2021;39(7):1087-1095. doi:10.1016/j.vaccine.2021.01.011
47. Saha R, Ghosh P, Burra VLSP. Designing a next generation multi-epitope based peptide vaccine candidate against SARS-CoV-2 using computational approaches. *3 Biotech*. 2021;11(2):47. doi:10.1007/s13205-020-02574-x
48. Pourseif MM, Parvizpour S, Jafari B, Dehghani J, Naghili B, Omidi Y. A domain-based vaccine construct against SARS-CoV-2, the causative agent of COVID-19 pandemic: development of self-amplifying mRNA and peptide vaccines. *BioImpacts*. 2020;11(1):65-84. doi:10.34172/bi.2021.11
49. Baruah V, Bose S. Immunoinformatics-aided identification of T cell and B cell epitopes in the surface glycoprotein of 2019-nCoV. *J Med Virol*. 2020;92(5):495-500. doi:10.1002/jmv.25698

50. Noorimotlagh Z, Karami C, Mirzaee SA, Kaffashian M, Mami S, Azizi M. Immune and bioinformatics identification of T cell and B cell epitopes in the protein structure of SARS-CoV-2: A systematic review. *Int Immunopharmacol*. 2020;86:106738. doi:10.1016/j.intimp.2020.106738

51. Kiyotani K, Toyoshima Y, Nemoto K, Nakamura Y. Bioinformatic prediction of potential T cell epitopes for SARS-Cov-2. *J Hum Genet*. 2020;65(7):569-575. doi:10.1038/s10038-020-0771-5

52. Dai L, Gao GF. Viral targets for vaccines against COVID-19. *Nat Rev Immunol*. 2021;21(2):73-82. doi:10.1038/s41577-020-00480-0

53. State of Hawaii D of H. Delta variant detected in all major counties. Published June 25, 2021. Accessed August 10, 2021. https://health.hawaii.gov/news/newsroom/delta-variant-detected-in-all-major-counties/

54. Davies NG, Abbott S, Barnard RC, et al. Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England. *Science*. 2021;372(6538):eabg3055. doi:10.1126/science.abg3055

55. Garcia-Beltran WF, Lam EC, St Denis K, et al. Multiple SARS-CoV-2 variants escape neutralization by vaccine-induced humoral immunity. *Cell*. 2021;184(9):2372-2383.e9. doi:10.1016/j.cell.2021.03.013

56. Moruf A. Fact Sheet For Health Care Providers Emergency Use Authorization (Eua) Of Bamlanivimab. Published online May 2021:26.

57. Zhao LP, Lybrand TP, Gilbert PB, et al. Tracking SARS-CoV-2 Spike Protein Mutations in the United States (2020/01 - 2021/03) Using a Statistical Learning Strategy. *BioRxiv Prepr Serv Biol*. Published online June 15, 2021:2021.06.15.448495. doi:10.1101/2021.06.15.448495

58. Weise E, Weintraub K. A COVID-19 vaccine life cycle: from DNA to doses. USA Today News. Published February 7, 2021. Accessed June 28, 2021.

https://www.usatoday.com/in-depth/news/health/2021/02/07/how-covid-vaccine-made-step-step-journey-pfizer-dose/4371693001/

59. Xie X, Muruato A, Lokugamage KG, et al. An Infectious cDNA Clone of SARS-CoV-2. *Cell Host Microbe*. 2020;27(5):841-848.e3. doi:10.1016/j.chom.2020.04.004

60. Tsai W-Y, Ching LL, Hsieh S-C, Melish ME, Nerurkar VR, Wang W-K. A real-time and high-throughput neutralization test based on SARS-CoV-2 pseudovirus containing monomeric infrared fluorescent protein as reporter. *Emerg Microbes Infect*. Published online April 30, 2021:1-38. doi:10.1080/22221751.2021.1925163

61. Jangra S, Ye C, Rathnasinghe R, et al. *The E484K Mutation in the SARS-CoV-2 Spike Protein Reduces but Does Not Abolish Neutralizing Activity of Human Convalescent and Post-Vaccination Sera*. Infectious Diseases (except HIV/AIDS); 2021. doi:10.1101/2021.01.26.21250543

62. Deng X, Garcia-Knight MA, Khalid MM, et al. Transmission, infectivity, and neutralization of a spike L452R SARS-CoV-2 variant. *Cell*. Published online April 20, 2021:S0092-8674(21)00505-5. doi:10.1016/j.cell.2021.04.025

63. 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

64. Ali F, Kasry A, Amin M. The new SARS-CoV-2 strain shows a stronger binding affinity to ACE2 due to N501Y mutant. *Med Drug Discov*. Published online March 2, 2021. doi:10.1016/j.medidd.2021.100086

65. Mansbach RA, Chakraborty S, Nguyen K, Montefiori DC, Korber B, Gnanakaran S. The SARS-CoV-2 Spike variant D614G favors an open conformational state. *Sci Adv*. 2021;7(16):eabf3671. doi:10.1126/sciadv.abf3671

66. Solutions for surveillance of the S gene mutation in the B.1.1.7 (501Y.V1) SARS-CoV-2 strain lineage. Behind the Bench. Published December 31, 2020. Accessed March 14, 2021.

https://www.thermofisher.com/blog/behindthebench/solutions-for-surveillance-of-the-s-gene-mutation-in-the-b117-501yv1-sars-cov-2-strain-lineage/

67. Laczkó D, Hogan MJ, Toulmin SA, et al. A Single Immunization with Nucleoside-Modified mRNA Vaccines Elicits Strong Cellular and Humoral Immune Responses against SARS-CoV-2 in Mice. *Immunity*. 2020;53(4):724-732.e7. doi:10.1016/j.immuni.2020.07.019

68. Pallesen J, Wang N, Corbett KS, et al. Immunogenicity and structures of a rationally designed prefusion MERS-CoV spike antigen. *Proc Natl Acad Sci*. 2017;114(35):E7348-E7357. doi:10.1073/pnas.1707304114

69. CDC. Selecting Viruses for the Seasonal Flu Vaccine. Centers for Disease Control and Prevention. Published October 26, 2020. Accessed April 22, 2021. https://www.cdc.gov/flu/prevent/vaccine-selection.htm

Figures and Tables

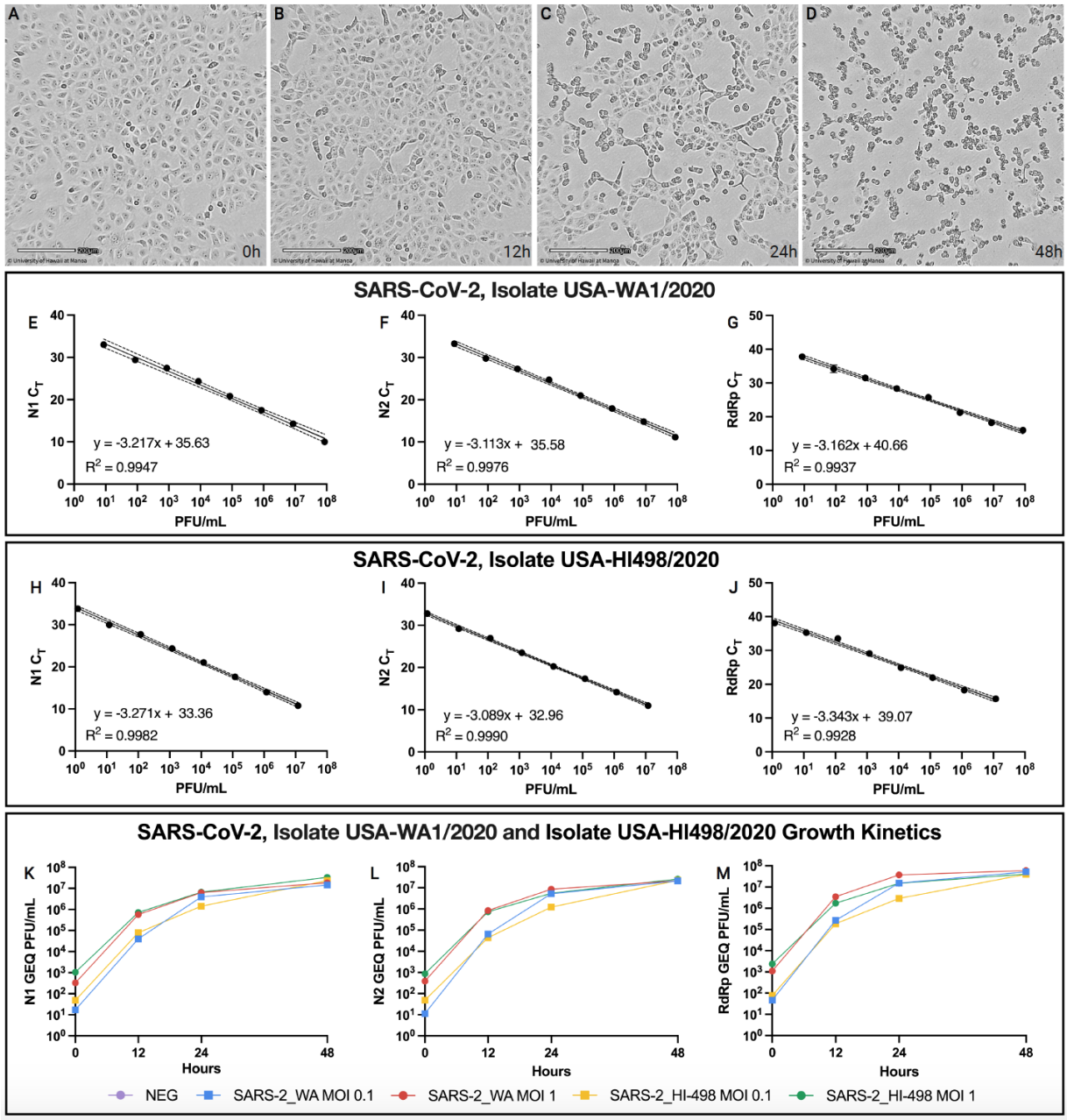


Figure 1. Cytopathic Effect and Growth Kinetics of SARS-CoV-2, Isolate USA-HI498 2020

The figure shows time-lapse cell images of VeroE6 cells infected with SARS-CoV-2 isolate USA-HI498 2020 at different time points, demonstrating the cytopathic effect (CPE) induced by the virus at multiplicity of infection (MOI) 1. A) 0 hr., B) 12 hr., C) 24 hr., and D) 48 hr. Scale bar equals 200 µm. E) Genomic equivalent (GEQ) comparison between SARS-CoV-2 USA-HI498

2020 isolate at MOI 0.1 (yellow) and 1 (green), with the SARS-CoV-2 Washington (WA) isolate at MOI 0.1 (blue) and 1 (red), using N1 (E,H,K), N2 (F,I,L) and RdRp (G,J,M) primers.

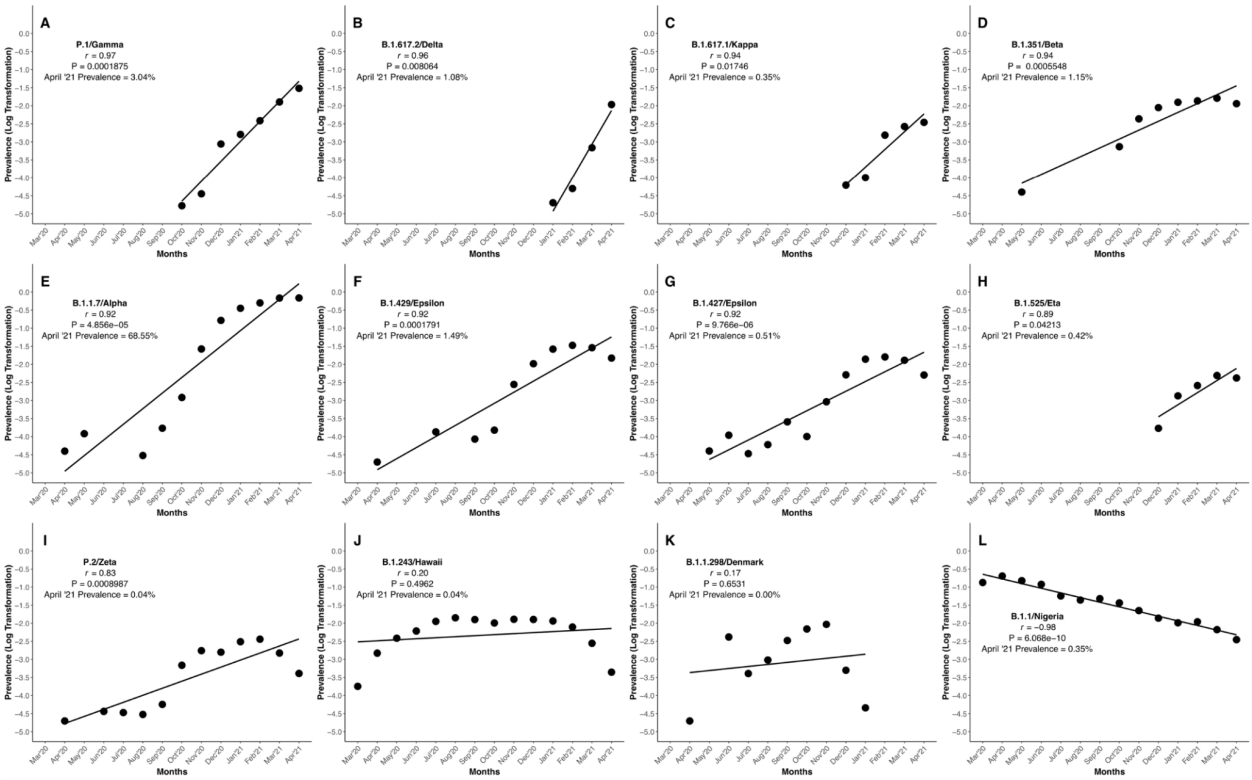


Figure 2. Pearson's Correlation on Logarithmically-Transformed Prevalence Ratios of SARS-CoV-2 Variant of Concern, Variants of Interest, and other Lineages

This figure demonstrates the quantitation of SARS-CoV-2 variants of concern, variants of interest, and other lineages. The emergence and disappearance of variants/lineages of SARS-CoV-2 is evaluated by Pearson's correlation of logarithmic transformation prevalence data. Variants are displayed in order of decreasing *r* value (A) P.1, B) B.1.617.2, C) B.1.617.1, D) B.1.351, E) B.1.1.7, F) B.1.429, G) B.1.427, H) B.1.525, I) P.2, J) B.1.243, K) B.1.1.298, and L) B.1.1).

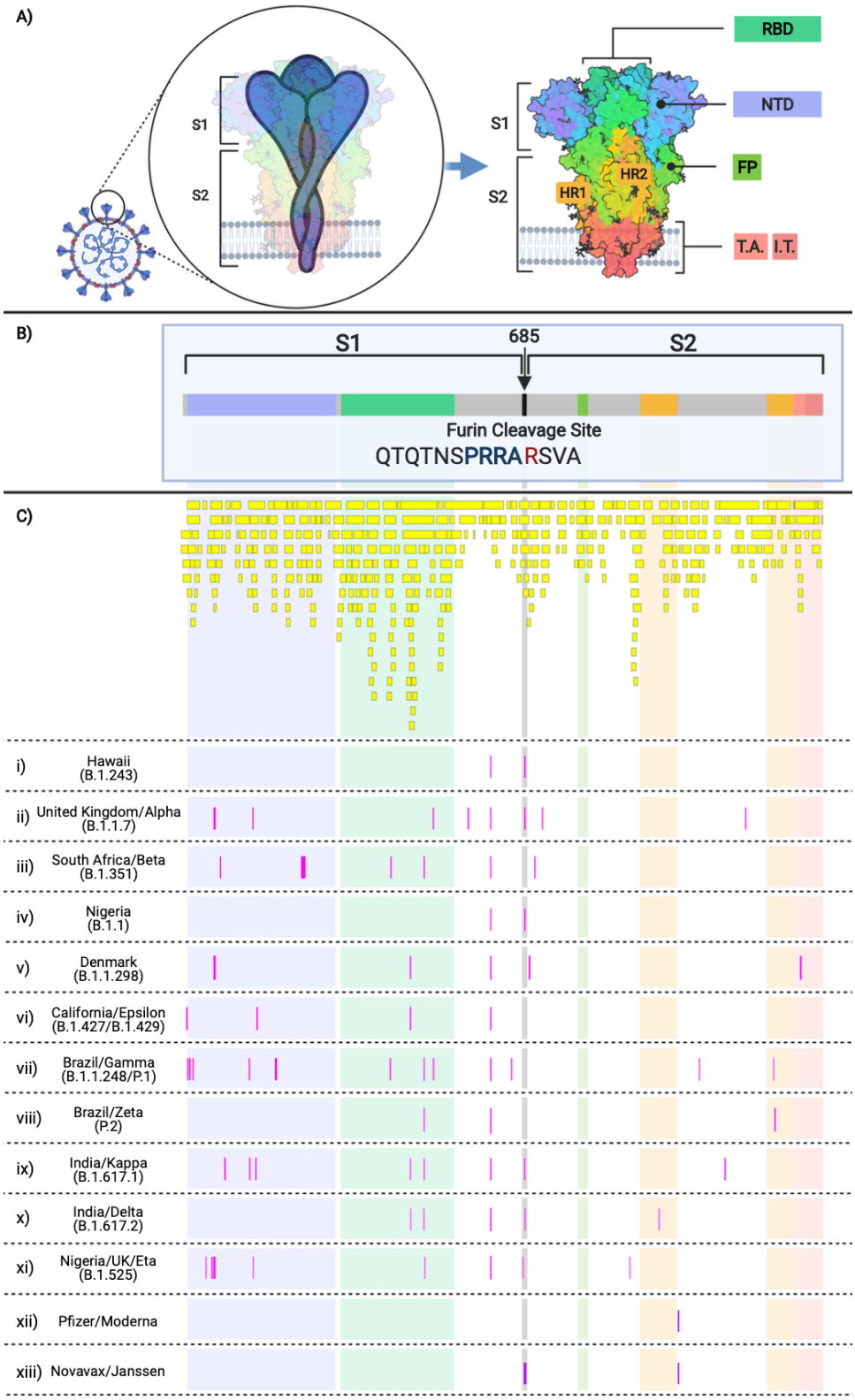


Figure 3. SARS-CoV-2 Spike Protein Domains and Relation to B and T cell Epitopes, Variant Amino Acid Substitutions, and Vaccine Amino Acid Substitutions

This figure demonstrates the evolution of the SARS-CoV-2 variants by depicting the location of the variants substitutions and deletions in the context of spike domains and epitopes. A) Cartoon rendering of SARS-CoV-2 and the 1,273 amino acid long spike protein overlay onto the color-coded crystallographic structure determined by electron microscopy (PBD ID: 6VXX-PDB). The individual protein domains are color-coded: N-terminal domain (NTD) (light purple) (residues 14-305), receptor-binding domain (RBD) (teal green) (residues 319-541), furin (F) (residues 682-685), fusion protein (FP) (green) (residues 788-806), heptad repeat 1 (HR1) (orange) (residues 912-984), heptad repeat 2 (HR2) (orange) (residues 1163-1213), transmembrane anchor (TM) (light pink) (1213-1237), and intracellular tail domain (IT) (dark pink) (1237-1273). B) Two-dimensional layout of the spike protein and domains with the addition of the S1/S2 furin cleavage site (RRA/R) (682-685) (black). C) *In silico* predicted B and T cell epitope loci revealing 393 *in silico* B and T cell epitopes mapped here individually as a yellow boxes i)-xiii) Amino acid substitutions present in the corresponding variant shown in pink boxes in comparison to the reference sequence NC_045512. i) B.1.243 Hawaii; ii) B.1.1.7 United Kingdom; iii) B.1.351 South Africa; iv) B.1.1 Nigeria; v) B.1.1.298 Denmark; vi) B.1.427 and B.1.429 California; vii) P.1 Brazil; viii) P.2 Brazil; ix) B.1.617.1 India; x) B.1.617.2 India; xi) B.1.525 United Kingdom/Nigeria; xii) Pfizer and Moderna mRNA sequences with artificially added substitutions K986P and V987P; xiii) Novavax and Janssen mRNA sequences with artificially added substitutions R682S/Q, R683Q, R685G/Q, K986P, and V987P.

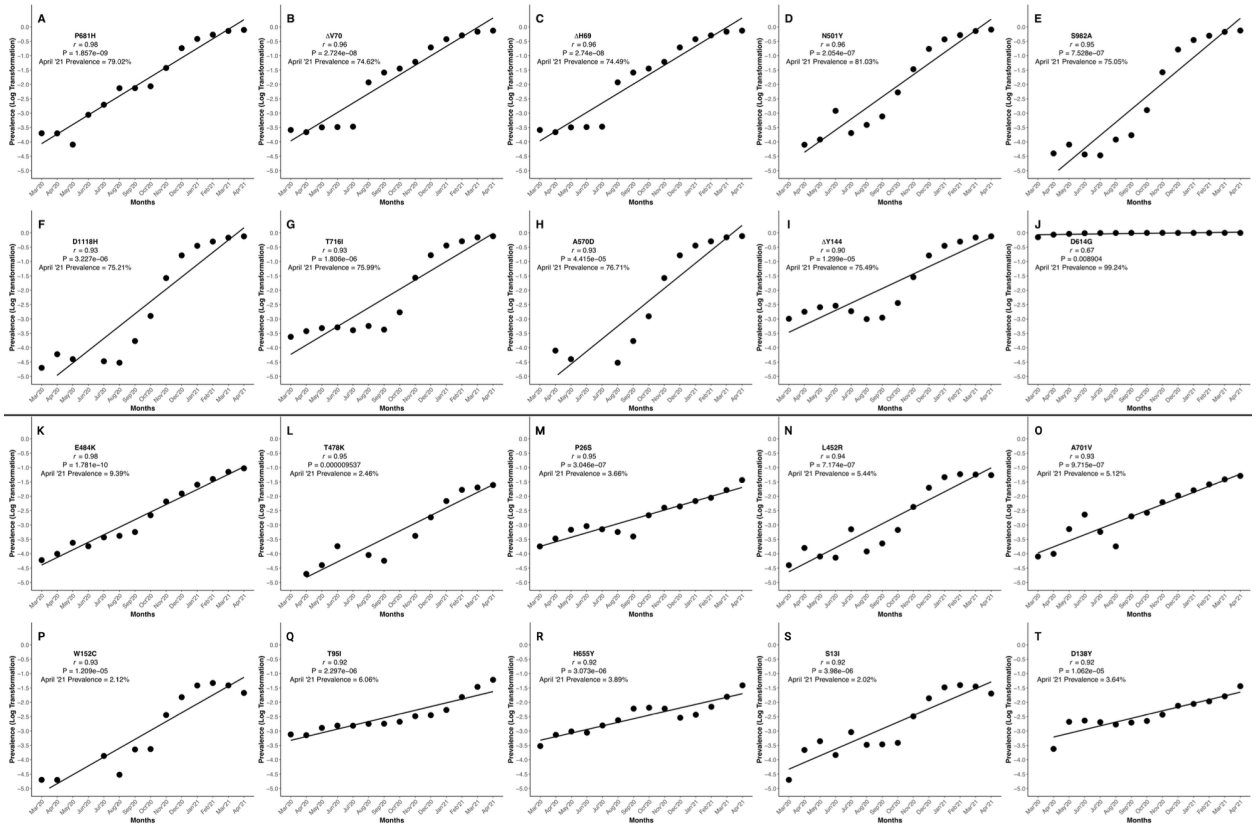


Figure 4. Pearson's Correlation of Logarithmically-Transformed Prevalence Ratios of the Most Emergent SARS-CoV-2 Mutations of Concern and Interest Selected via the Algorithm

This figure shows the graphical representation of the logarithmically-transformed prevalence data used to calculate the Pearson's correlation of each of the twenty most emerged (of concern) and emergent (of interest) spike protein substitutions and deletions. The substitutions and deletions of concern here are in order of decreasing *r* value, and each has a unique alphabet identifier A) P681H, B) ΔV70, C) ΔH69, D) N501Y, E) S982A, F) D1118H, G) T716, H) A570D, I) ΔY144, and J) D614G. The algorithm uses the monthly prevalence data from these ten spike protein substitutions and deletions, and they are the most concerning of all spike changes. The substitutions and deletions of interest here are in order of decreasing *r* value and each unique substitution or deletion is denoted by a letter of the English alphabet, K) E484K, L)

T478K, M) P26S, N) L452R, O) A701V, P) W152C, Q) T95I, R) H655Y, S) S13I, and T) D138Y. Graphs were generated using RStudio version 1.3.1093 (R version 4.0.3) and the ggplot2 package. Graphs were compiled and the final figure generated using Biorender.com.

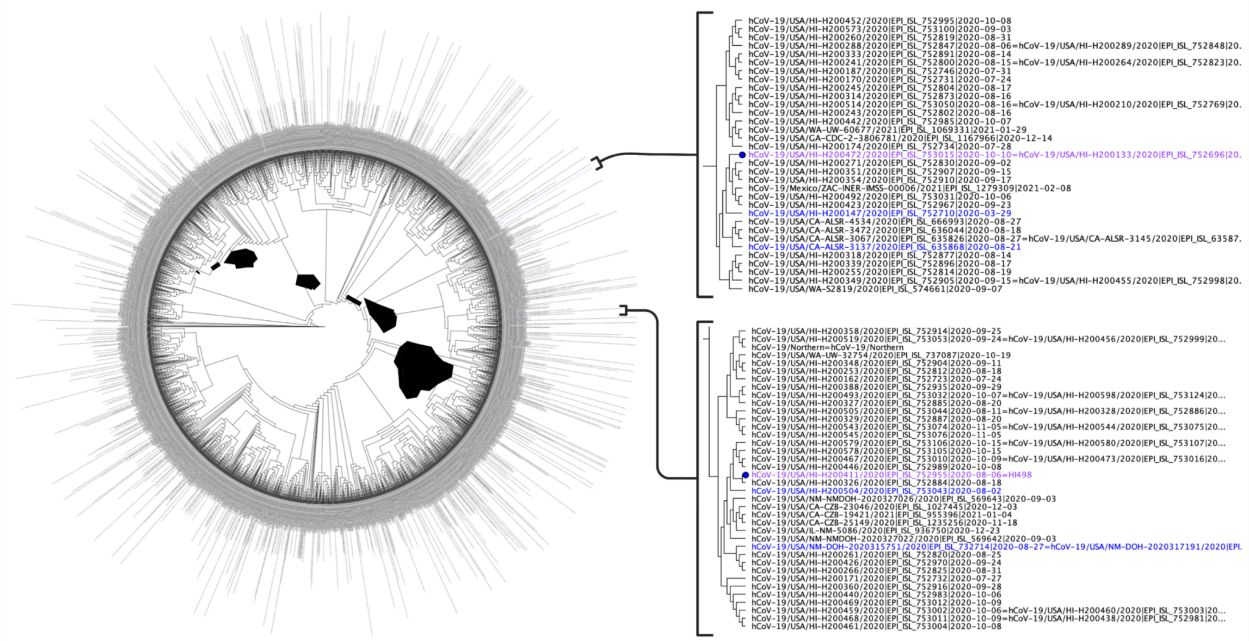


Figure 5. Phylogenetic Tree of all B.1.243 Lineage Sequences Worldwide

This figure displays the phylogenetic tree used to determine the origin of the B.1.243 sequences used in this study. We use 8,822 SARS-CoV-2 B.1.243 whole-genome sequences published in the Global Initiative on Sharing Avian Influenza Data (GISAID) and GenBank as of April 12, 2021 to define the origin. From the 8,822, 4,273 had ambiguous nucleotides between the 5' and 3' untranslated regions as determined using multiple alignment using fast Fourier transform (MAFFT). Further, 1,588 were duplicate sequences and eight had duplicate identifications as determined by the sRNA Toolbox. Therefore, the final tree was constructed using FastTree in Geneious Prime 2021.1.1 (http://www.geneious.com) from 2,953 unique and unambiguous SARS-CoV-2 whole-genome sequences. The HI498 (purple text) origin is defined as New Mexico (blue text) and the HI-708 (purple text) origin is defined as California (blue text).

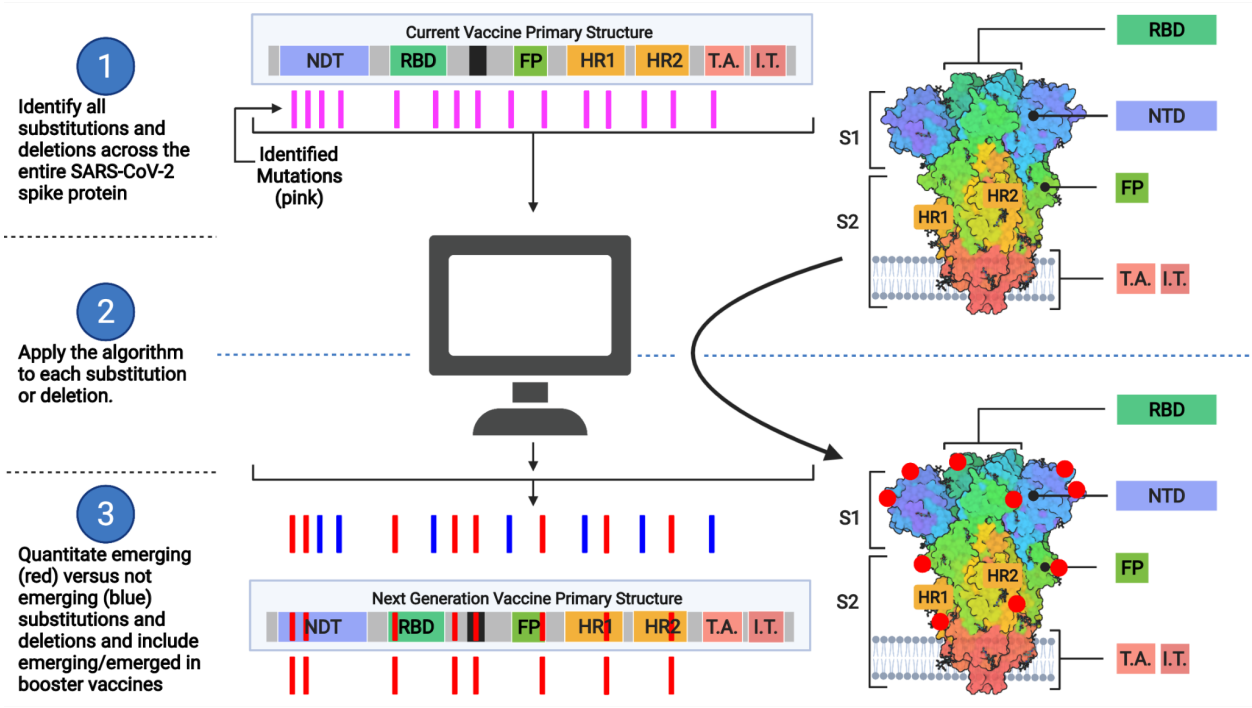
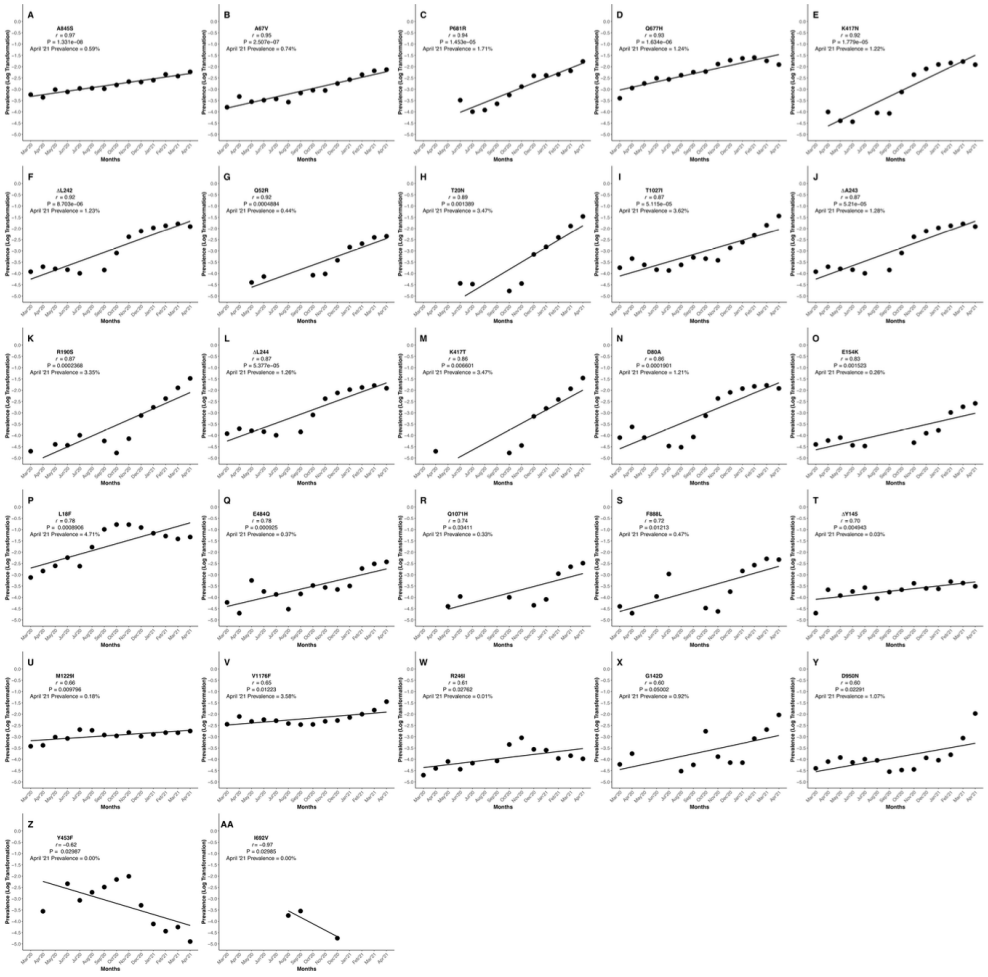


Figure 6. Applying the Algorithm to Vaccine Design

The model displays how the algorithm would lead to the design of next-generation vaccines based on the S gene. Part one of the model identifies all SARS-CoV-2 genomic mutations and spike amino acid substitutions and deletions (represented as pink lines) based on the worldwide sequence databases GISAID and GenBank. The current vaccine design as it translates into proteins is depicted in the top center and top right. Part two of the model will apply the quantitative analysis and algorithm described in this report to each of the protein changes identified in part one. The quantitative analysis determines emergence via logarithmic transformation of prevalence and Pearson's correlation. The algorithm then applies criteria to the quantitative analysis and previous months prevalence for determining which changes are likely to be in the majority of SARS-CoV-2 for incorporation in the next-generation vaccine. Part three of the model determines which substitutions and deletions are exponentially emerging or emerged (red lines) and which are not (blue lines). From part three, the mRNA sequence of vaccines can then incorporate the emerging and emerged mutations so that the folded protein

(bottom right) will contain the protein changes (red dots) most prevalent worldwide by the time the next-generation vaccine is manufactured and administered. As a result, these substitutions and deletions will present the most appropriate epitopes of the SARS-CoV-2 spike protein to vaccine recipients.



Supplementary Figure 1. Pearson's Correlation on Logarithmically-Transformed Prevalence Ratios of the Remaining SARS-CoV-2 Variant Amino Acid Substitutions and Deletions Not Currently Selected via the Algorithm

This figure shows the graphical representation of the SARS-CoV-2 spike protein amino acid substitutions and deletions not currently concerning due to low previous month prevalence, low *r* value, or insignificant P value. Though not yet of concern as of April '21, those substitution and deletions represented by high *r* values should be cause for close monitoring. Each graph denoted by an alphabetical character or characters represents a unique amino acid substitution or deletion in the spike protein of SARS-CoV-2 (A) A845S, B) A67V, C) P681R, D) Q677H, E) K417N, F) ΔL242, G) Q52R, H) T20N, I) T1027I, J) ΔA243, K) R190S, L) ΔL244, M) K417T, N)

D80A, O) E154K, P) L18F, Q) E484Q, R) Q1071H, S) F888L, T) ΔY145, U) M1229I, V) V1176F, W) R246I, X) G142D, Y) D950N, Z) Y453F, and AA) I692V). Graphs were generated using RStudio version 1.3.1093 (R version 4.0.3) and the ggplot2 package. Graphs were compiled and the final figure generated using Biorender.com.

Table 1. Genetic Characteristics of the Hawaii SARS-CoV-2 Variant B.1.243 Isolates, USA-HI498 2020* and USA-HI708 2020**

Gene	Nucleotide			Amino Acid		
	Loci	Wild Type	Mutant	Loci	Wild Type	Mutant
5' UTR	241	C	T	-	-	-
ORF1ab	3,037	C	T	924	Phe (F)	Phe (F)
ORF1ab	10,741	C	T	3492	Asp (D)	Asp (D)
ORF1ab	12,076***	C	T	3937	Asn (N)	Asn (N)
ORF1ab	14,408	C	T	4715	Pro (P)	Leu (L)
ORF1ab	20,268	A	G	6668	Leu (L)	Leu (L)
S	23,403	A	G	614	Asp (D)	Gly (G)
S	23,604	C	A	681	Pro (P)	His (H)
S	24,076	T	C	838	Gly (G)	Gly (G)
N	28,854	C	T	194	Ser (S)	Leu (L)
N	29,266***	G	A	331	Leu (L)	Leu (L)
3' UTR	29,710	T	C	-	-	-

GenBank accession *MZ664037 and **MZ664038, ***exclusive to SARS-CoV-2, Isolate USA-HI498 2020

Table 2. Pearson's Correlation of Logarithmic Transformed Prevalence Ratios of Amino Acid Substitutions (AAS) in the Spike Protein and Variants from March 2020 - April 2021 and Pairwise Heatmap of Absolute Value Difference of *r*-value between Corresponding AAS and Variants

				Variant	P.1 Gamma	B.1.617.2 Delta	B.1.617.1 Kappa	B.1.351 Beta	B.1.1.7 Alpha	B.1.429 Epsilon	B.1.427 Epsilon	B.1.525 Eta	P.2 Zeta	B.1.243 -	B.1.1.298 -	B.1.1 -
				Origin	Brazil	India	India	South Africa	United Kingdom	California	California	United Kingdom	Brazil	Hawai'i	Denmark	Nigeria
				n	12,485	2,758	2,037	14,798	580,502	26,779	12,305	3,411	2,350	9,726	1,533	41,833
				April '21 Prevalence (%)	3.04	1.08	0.35	1.15	68.55	1.49	0.51	0.42	0.04	0.04	0.00	0.35
				P Value	0.00	0.01	0.02	0.00	0.00	0.00	0.00	0.04	0.00	0.50	0.65	0.00
Substitution	**n**	**April '21 Prevalence (%)**	**P value**	**r**	0.97	0.96	0.94	0.94	0.92	0.92	0.92	0.89	0.83	0.20	0.17	-0.98
E484K	60,990	9.39	0.000	0.98	0.010			0.046				0.093	0.157		0.810	
P681H	641,501	79.02	0.000	0.98					0.053					0.779		1.959
A845S	4,830	0.59	0.000	0.97												1.950
ΔV70	614,413	74.62	0.000	0.96					0.041			0.073			0.790	
ΔH69	613,401	74.49	0.000	0.96					0.040			0.073			0.790	
N501Y	636,491	81.03	0.000	0.96	0.015			0.021	0.036				0.133		0.785	
S982A	594,828	75.05	0.000	0.95					0.025							
A67V	5,784	0.74	0.000	0.95								0.057				
T478K	17,655	2.46	0.000	0.95		0.017										
P26S	18,361	3.66	0.000	0.95	0.029								0.119			
P681R	8,571	1.71	0.000	0.94		0.022	0.002									
L452R	56,039	5.44	0.000	0.94		0.026	0.002			0.020	0.020					
A701V	35,639	5.12	0.000	0.93				0.004								
D1118H	595,048	75.21	0.000	0.93					0.009							
W152C	37,487	2.12	0.000	0.93						0.012	0.012					
Q677H	23,205	1.24	0.000	0.93								0.037				
T716I	608,252	75.99	0.000	0.93					0.004							
A570D	609,134	76.71	0.000	0.93					0.002							
T95I	31,010	6.06	0.000	0.92			0.016									
K417N	15,367	1.22	0.000	0.92				0.015								
H655Y	18,155	3.89	0.000	0.92	0.055								0.093		0.746	
ΔL242	14,464	1.23	0.000	0.92				0.019								
Q52R	3,138	0.44	0.000	0.92								0.026				
S13I	33,379	2.02	0.000	0.92						0.001	0.001					
D138Y	19,518	3.64	0.000	0.92	0.058								0.089		0.742	
ΔY144	600,274	75.49	0.000	0.90					0.026			0.006				
T20N	13,656	3.47	0.001	0.89	0.087								0.061			
T1027I	14,941	3.62	0.000	0.87	0.104								0.043			
ΔA243	15,132	1.28	0.000	0.87				0.068								
R190S	13,441	3.35	0.000	0.87	0.105								0.042		0.695	
ΔL244	14,735	1.26	0.000	0.87				0.069								
K417T	13,226	3.47	0.007	0.86	0.118								0.029		0.682	
D80A	15,126	1.21	0.000	0.86				0.083								
E154K	1,505	0.26	0.002	0.83			0.109									
L18F	91,123	4.71	0.001	0.78	0.190								0.043			
E484Q	2,447	0.37	0.001	0.78			0.157									
Q1071H	1,783	0.33	0.034	0.74			0.196									
F888L	3,698	0.47	0.012	0.72								0.170				
ΔY145	482	0.03	0.005	0.70					0.220			0.188				
D614G	1,446,840	99.24	0.009	0.67	0.306	0.295	0.271	0.270	0.255	0.249	0.249	0.223	0.158	0.470	0.494	1.650
M1229I	2,058	0.18	0.010	0.66											0.488	
V1176F	19,419	3.58	0.012	0.65	0.327								0.179			
R246I	290	0.01	0.028	0.61				0.331								
G142D	3,190	0.92	0.050	0.60			0.338									
D950N	2,895	1.07	0.023	0.60		0.363										
K150N	180	0.04	0.040	0.60												
V308L	1,212	0.08	0.060	0.51												
Y453F	1,689	0.00	0.030	-0.62											0.799	
I692V	19	0.00	0.030	-0.97											1.145	

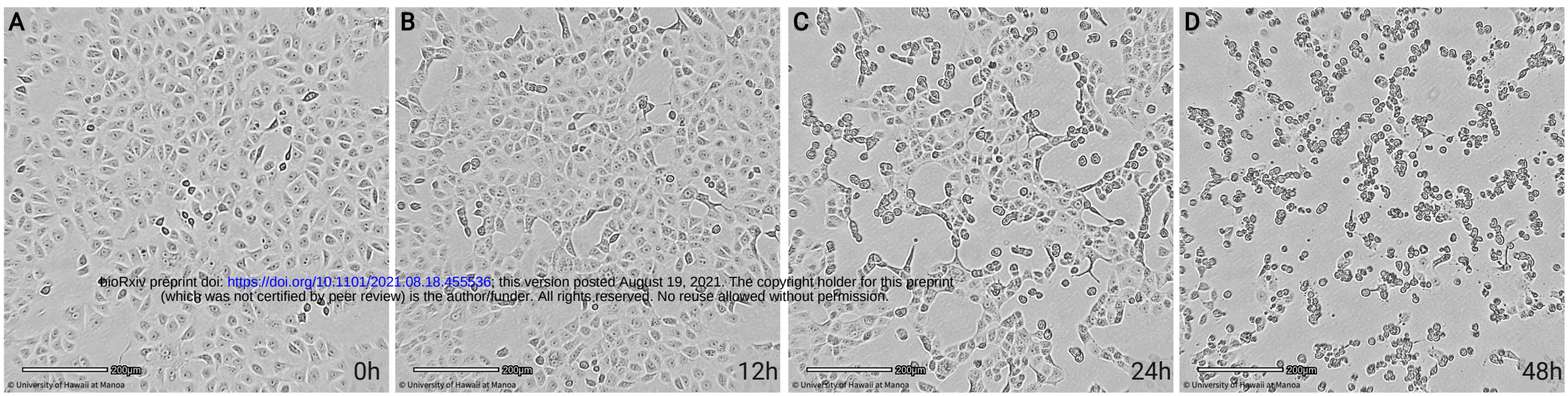
Table 3. Comparison of Single Nucleotide Polymorphisms (SNP) and Resultant Amino Acid Substitutions (AAS) Among SARS-CoV-2 Variants

ACCESSION AND IDENTIFIER	SNP and AAS																																							
	S13I		L18F		T20N		P26S		Q52R		A67V		ΔH69		ΔV70		D80A		T95I		D138Y		G142D		Y144		ΔY145		W152C		E154K		R190S								
	21,600 NT	13 AA	21,614 NT	18 AA	21,621 NT	20 AA	21,638 NT	26 AA	21,717 NT	52 AA	21,762 NT	67 AA	21,766-8 NT	69 AA	21,769-71 NT	70 AA	21,801 NT	80 AA	21,846 NT	95 AA	21,974 NT	138 AA	21,987 NT	142 AA	21,994 NT	144 AA	21,995-7 NT	145 AA	22,018 NT	152 AA	22,032 NT	154 AA	22,132 NT	190 AA							
NC_045512_Reference Genome_Wuhan	G	Ser (S)	C	Leu (L)	C	Thr (T)	C	Pro (P)	A	Gln (Q)	C	Ala (A)	ACA	His (H)	TGT	Val (V)	A	Asp (D)	C	Thr (T)	G	Asp (D)	G	Gly (G)	T	Tyr (Y)	TAC	Tyr (Y)	G	Trp (W)	G	Glu (E)	G	Arg (R)							
EPI_ISL_601443_UK_B.1.1.7_Alpha	G	Ser (S)	C	Leu (L)	C	Thr (T)	C	Pro (P)	A	Gln (Q)	C	Ala (A)	---	Δ	---	Δ	A	Asp (D)	C	Thr (T)	G	Asp (D)	G	Gly (G)	C	Tyr (Y)	---	Δ	G	Trp (W)	G	Glu (E)	G	Arg (R)							
EPI_ISL_712081_South Africa_B.1.351_Beta	G	Ser (S)	C	Leu (L)	C	Thr (T)	C	Pro (P)	A	Gln (Q)	C	Ala (A)	ACA	His (H)	TGT	Val (V)	C	Ala (A)	C	Thr (T)	G	Asp (D)	G	Gly (G)	T	Tyr (Y)	TAC	Tyr (Y)	G	Trp (W)	G	Glu (E)	G	Arg (R)							
EPI_ISL_729975_Nigeria_B.1.1	G	Ser (S)	C	Leu (L)	C	Thr (T)	C	Pro (P)	A	Gln (Q)	C	Ala (A)	ACA	His (H)	TGT	Val (V)	A	Asp (D)	C	Thr (T)	G	Asp (D)	G	Gly (G)	T	Tyr (Y)	TAC	Tyr (Y)	G	Trp (W)	G	Glu (E)	G	Arg (R)							
EPI_ISL_616802_Denmark_B.1.1.298	G	Ser (S)	C	Leu (L)	C	Thr (T)	C	Pro (P)	A	Gln (Q)	C	Ala (A)	---	Δ	---	Δ	A	Asp (D)	C	Thr (T)	G	Asp (D)	G	Gly (G)	T	Tyr (Y)	TAC	Tyr (Y)	G	Trp (W)	G	Glu (E)	G	Arg (R)							
EPI_ISL_942929_L452R_B.1.427/429_Epsilon	T	Ile (I)	C	Leu (L)	C	Thr (T)	C	Pro (P)	A	Gln (Q)	C	Ala (A)	ACA	His (H)	TGT	Val (V)	A	Asp (D)	C	Thr (T)	G	Asp (D)	G	Gly (G)	T	Tyr (Y)	TAC	Tyr (Y)	T	Cys (C)	G	Glu (E)	G	Arg (R)							
EPI_ISL_792680_Brazil_B.1.1.248/P.1_Gamma	G	Ser (S)	T	Phe (F)	A	Asn (N)	T	Ser (S)	A	Gln (Q)	C	Ala (A)	ACA	His (H)	TGT	Val (V)	A	Asp (D)	C	Thr (T)	T	Tyr (Y)	G	Gly (G)	T	Tyr (Y)	TAC	Tyr (Y)	G	Trp (W)	G	Glu (E)	T	Ser (S)							
EPI_ISL_918536_Brazil_P.2_Zeta	G	Ser (S)	C	Leu (L)	C	Thr (T)	C	Pro (P)	A	Gln (Q)	C	Ala (A)	ACA	His (H)	TGT	Val (V)	A	Asp (D)	C	Thr (T)	G	Asp (D)	G	Gly (G)	T	Tyr (Y)	TAC	Tyr (Y)	G	Trp (W)	G	Glu (E)	G	Arg (R)							
EPI_ISL_1739895_B.1.525_Eta	G	Ser (S)	C	Leu (L)	C	Thr (T)	C	Pro (P)	G	Arg (R)	T	Val (V)	---	Δ	---	Δ	A	Asp (D)	C	Thr (T)	G	Asp (D)	G	Gly (G)	C	Tyr (Y)	---	Δ	G	Trp (W)	G	Glu (E)	G	Arg (R)							
EPI_ISL_1372093_India_B.1.617.1_Kappa	G	Ser (S)	C	Leu (L)	C	Thr (T)	C	Pro (P)	A	Gln (Q)	C	Ala (A)	ACA	His (H)	TGT	Val (V)	A	Asp (D)	T	Ile (I)	G	Asp (D)	A	Asp (D)	T	Tyr (Y)	TAC	Tyr (Y)	G	Trp (W)	A	Lys (K)	G	Arg (R)							
EPI_ISL_1663516_India_B.1.617.2_Delta	G	Ser (S)	C	Leu (L)	C	Thr (T)	C	Pro (P)	A	Gln (Q)	C	Ala (A)	ACA	His (H)	TGT	Val (V)	A	Asp (D)	C	Thr (T)	G	Asp (D)	G	Gly (G)	T	Tyr (Y)	TAC	Tyr (Y)	G	Trp (W)	G	Glu (E)	G	Arg (R)							
SARS-CoV-2, Isolate USA-HI498/2020_B.1.243	G	Ser (S)	C	Leu (L)	C	Thr (T)	C	Pro (P)	A	Gln (Q)	C	Ala (A)	ACA	His (H)	TGT	Val (V)	A	Asp (D)	C	Thr (T)	G	Asp (D)	G	Gly (G)	T	Tyr (Y)	TAC	Tyr (Y)	G	Trp (W)	G	Glu (E)	G	Arg (R)							

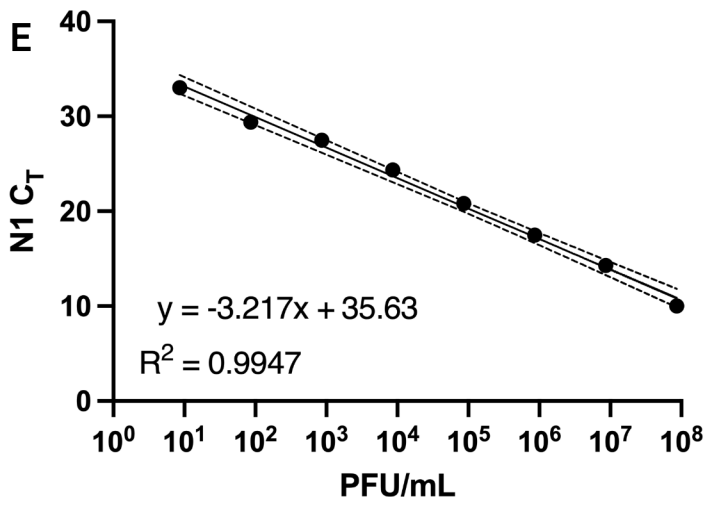
ACCESSION AND IDENTIFIER	SNP and AAS																																							
	L241		ΔL242		ΔA243		ΔL244		R246I		K417T		K417N		L452R		Y453F		T478K		E484K		E484Q		N501Y		A570D		D614G		H655Y		Q677H								
	22,284-5 NT	241 AA	22,296-8 NT	242 AA	22,289-91 NT	243 AA	22,292 NT	244 AA	22,299 NT	246 AA	22,812 NT	417 AA	22,813 NT	417 AA	22,917 NT	452 AA	22,920 NT	453 AA	22,995 NT	478 AA	23,012 NT	484 AA	23,012 NT	484 AA	23,063 NT	501 AA	23,271 NT	570 AA	23,403 NT	614 AA	23,525 NT	655 AA	23,593 NT	677 AA							
NC_045512_Reference Genome_Wuhan	TA	Leu (L)	CTT	Leu (L)	GCT	Ala (A)	T	Leu (L)	G	Arg (R)	A	Lys (K)	G	Lys (K)	T	Leu (L)	A	Tyr (Y)	C	Thr (T)	G	Glu (E)	G	Glu (E)	A	Asn (N)	C	Ala (A)	A	Asp (D)	C	His (H)	G	Gln (Q)							
EPI_ISL_601443_UK_B.1.1.7_Alpha	TA	Leu (L)	CTT	Leu (L)	GCT	Ala (A)	T	Leu (L)	G	Arg (R)	A	Lys (K)	G	Lys (K)	T	Leu (L)	A	Tyr (Y)	C	Thr (T)	G	Glu (E)	G	Glu (E)	T	Tyr (Y)	A	Asp (D)	G	Gly (G)	C	His (H)	G	Gln (Q)							
EPI_ISL_712081_South Africa_B.1.351_Beta	--	Leu (L)*	---	Δ	---	Δ	-	Leu (L)*	T	Ile (I)	A	Lys (K)	T	Asn (N)	T	Leu (L)	A	Tyr (Y)	C	Thr (T)	A	Lys (K)	A	Lys (K)	T	Tyr (Y)	C	Ala (A)	G	Gly (G)	C	His (H)	G	Gln (Q)							
EPI_ISL_729975_Nigeria_B.1.1	TA	Leu (L)	CTT	Leu (L)	GCT	Ala (A)	T	Leu (L)	G	Arg (R)	A	Lys (K)	G	Lys (K)	T	Leu (L)	A	Tyr (Y)	C	Thr (T)	G	Glu (E)	G	Glu (E)	A	Asn (N)	C	Ala (A)	G	Gly (G)	C	His (H)	G	Gln (Q)							
EPI_ISL_616802_Denmark_B.1.1.298	TA	Leu (L)	CTT	Leu (L)	GCT	Ala (A)	T	Leu (L)	G	Arg (R)	A	Lys (K)	G	Lys (K)	T	Leu (L)	T	Phe (F)	C	Thr (T)	G	Glu (E)	G	Glu (E)	A	Asn (N)	C	Ala (A)	G	Gly (G)	C	His (H)	G	Gln (Q)							
EPI_ISL_942929_L452R_B.1.427/429_Epsilon	TA	Leu (L)	CTT	Leu (L)	GCT	Ala (A)	T	Leu (L)	G	Arg (R)	A	Lys (K)	G	Lys (K)	G	Arg (R)	A	Tyr (Y)	C	Thr (T)	G	Glu (E)	G	Glu (E)	A	Asn (N)	C	Ala (A)	G	Gly (G)	C	His (H)	G	Gln (Q)							
EPI_ISL_792680_Brazil_B.1.1.248/P.1_Gamma	TA	Leu (L)	CTT	Leu (L)	GCT	Ala (A)	T	Leu (L)	G	Arg (R)	C	Thr (T)	G	Lys (K)	T	Leu (L)	A	Tyr (Y)	C	Thr (T)	A	Lys (K)	A	Lys (K)	T	Tyr (Y)	C	Ala (A)	G	Gly (G)	T	Tyr (Y)	G	Gln (Q)							
EPI_ISL_918536_Brazil_P.2_Zeta	TA	Leu (L)	CTT	Leu (L)	GCT	Ala (A)	T	Leu (L)	G	Arg (R)	A	Lys (K)	G	Lys (K)	T	Leu (L)	A	Tyr (Y)	C	Thr (T)	A	Lys (K)	A	Lys (K)	A	Asn (N)	C	Ala (A)	G	Gly (G)	C	His (H)	G	Gln (Q)							
EPI_ISL_1739895_B.1.525_Eta	TA	Leu (L)	CTT	Leu (L)	GCT	Ala (A)	T	Leu (L)	G	Arg (R)	A	Lys (K)	G	Lys (K)	T	Leu (L)	A	Tyr (Y)	C	Thr (T)	A	Lys (K)	A	Lys (K)	A	Asn (N)	C	Ala (A)	G	Gly (G)	C	His (H)	C	His (H)							
EPI_ISL_1372093_India_B.1.617.1_Kappa	TA	Leu (L)	CTT	Leu (L)	GCT	Ala (A)	T	Leu (L)	G	Arg (R)	A	Lys (K)	G	Lys (K)	G	Arg (R)	A	Tyr (Y)	C	Thr (T)	C	Gln (Q)	C	Gln (Q)	A	Asn (N)	C	Ala (A)	G	Gly (G)	C	His (H)	G	Gln (Q)							
EPI_ISL_1663516_India_B.1.617.2_Delta	TA	Leu (L)	CTT	Leu (L)	GCT	Ala (A)	T	Leu (L)	G	Arg (R)	A	Lys (K)	G	Lys (K)	G	Arg (R)	A	Tyr (Y)	A	Lys (K)	C	Gln (Q)	G	Glu (E)	A	Asn (N)	C	Ala (A)	G	Gly (G)	C	His (H)	G	Gln (Q)							
SARS-CoV-2, Isolate USA-HI498/2020_B.1.243	TA	Leu (L)	CTT	Leu (L)	GCT	Ala (A)	T	Leu (L)	G	Arg (R)	A	Lys (K)	G	Lys (K)	T	Leu (L)	A	Tyr (Y)	C	Thr (T)	G	Glu (E)	G	Glu (E)	A	Asn (N)	C	Ala (A)	G	Gly (G)	C	His (H)	G	Gln (Q)							

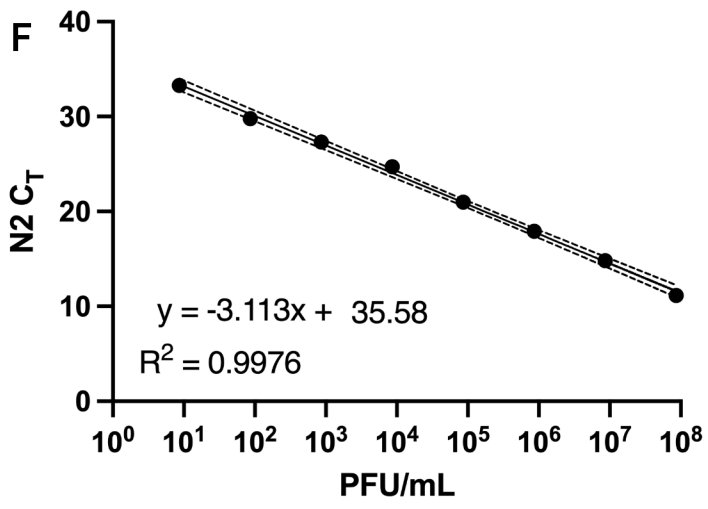
ACCESSION AND IDENTIFIER	SNP and AAS																																							
	P681H		P681R		I692V		A701V		T716I		G838		A845S		F888L		S929		D950N		S982A		T1027I		Q1071H		D1118H		D1146		V1176F		M1229I								
	23,604 NT	681 AA	23,604 NT	681 AA	23,636 NT	692 AA	23,644 NT	701 AA	23,709 NT	716 AA	24,076 NT	838 AA	24,095 NT	845 AA	24,224 NT	888 AA	24,349 NT	929 AA	24,410 NT	950 AA	24,506 NT	982 AA	24,642 NT	1027 AA	24,775 NT	1071 AA	24,914 NT	1118 AA	25,000 NT	1146 AA	25,088 NT	1176 AA	25,249 NT	1229 AA							
NC_045512_Reference Genome_Wuhan	C	Pro (P)	C	Pro (P)	A	Ile (I)	C	Ala (A)	C	Thr (T)	T	Gly (G)	G	Ala (A)	T	Phe (F)	T	Ser (S)	G	Asp (D)	T	Ser (S)	C	Thr (T)	A	Gln (Q)	G	Asp (D)	C	Asp (D)	G	Val (V)	G	Met (M)							
EPI_ISL_601443_UK_B.1.1.7_Alpha	A	His (H)	A	His (H)	A	Ile (I)	C	Ala (A)	T	Ile (I)	T	Gly (G)	G	Ala (A)	T	Phe (F)	T	Ser (S)	G	Asp (D)	G	Ala (A)	C	Thr (T)	A	Gln (Q)	C	His (H)	C	Asp (D)	G	Val (V)	G	Met (M)							
EPI_ISL_712081_South Africa_B.1.351_Beta	C	Pro (P)	C	Pro (P)	A	Ile (I)	T	Val (V)	C	Thr (T)	T	Gly (G)	G	Ala (A)	T	Phe (F)	T	Ser (S)	G	Asp (D)	T	Ser (S)	C	Thr (T)	A	Gln (Q)	G	Asp (D)	C	Asp (D)	G	Val (V)	G	Met (M)							
EPI_ISL_729975_Nigeria_B.1.1	A	His (H)	A	His (H)	A	Ile (I)	C	Ala (A)	C	Thr (T)	T	Gly (G)	T	Ser (S)	T	Phe (F)	T	Ser (S)	G	Asp (D)	T	Ser (S)	C	Thr (T)	A	Gln (Q)	G	Asp (D)	C	Asp (D)	G	Val (V)	G	Met (M)							
EPI_ISL_616802_Denmark_B.1.1.298	C	Pro (P)	C	Pro (P)	G	Val (V)	C	Ala (A)	C	Thr (T)	T	Gly (G)	G	Ala (A)	T	Phe (F)	T	Ser (S)	G	Asp (D)	T	Ser (S)	C	Thr (T)	A	Gln (Q)	G	Asp (D)	T	Asp (D)	G	Val (V)	T	Ile (I)							
EPI_ISL_942929_L452R_B.1.427/429_Epsilon	C	Pro (P)	C	Pro (P)	A	Ile (I)	C	Ala (A)	C	Thr (T)	T	Gly (G)	G	Ala (A)	T	Phe (F)	C	Ser (S)	G	Asp (D)	T	Ser (S)	C	Thr (T)	A	Gln (Q)	G	Asp (D)	C	Asp (D)	G	Val (V)	G	Met (M)							
EPI_ISL_792680_Brazil_B.1.1.248/P.1_Gamma	C	Pro (P)	C	Pro (P)	A	Ile (I)	C	Ala (A)	C	Thr (T)	T	Gly (G)	G	Ala (A)	T	Phe (F)	T	Ser (S)	G	Asp (D)	T	Ser (S)	T	Ile (I)	A	Gln (Q)	G	Asp (D)	C	Asp (D)	T	Phe (F)	G	Met (M)							
EPI_ISL_918536_Brazil_P.2_Zeta	C	Pro (P)	C	Pro (P)	A	Ile (I)	C	Ala (A)	C	Thr (T)	T	Gly (G)	G	Ala (A)	T	Phe (F)	T	Ser (S)	G	Asp (D)	T	Ser (S)	C	Thr (T)	A	Gln (Q)	G	Asp (D)	C	Asp (D)	T	Phe (F)	G	Met (M)							
EPI_ISL_1739895_B.1.525_Eta	C	Pro (P)	C	Pro (P)	A	Ile (I)	C	Ala (A)	C	Thr (T)	T	Gly (G)	G	Ala (A)	C	Leu (L)	T	Ser (S)	G	Asp (D)	T	Ser (S)	C	Thr (T)	A	Gln (Q)	G	Asp (D)	C	Asp (D)	G	Val (V)	G	Met (M)							
EPI_ISL_1372093_India_B.1.617.1_Kappa	G	Arg (R)	G	Arg (R)	A	Ile (I)	C	Ala (A)	C	Thr (T)	T	Gly (G)	G	Ala (A)	T	Phe (F)	T	Ser (S)	G	Asp (D)	T	Ser (S)	C	Thr (T)	T	His (H)	G	Asp (D)	C	Asp (D)	G	Val (V)	G	Met (M)							
EPI_ISL_1663516_India_B.1.617.2_Delta	G	Arg (R)	G	Arg (R)	A	Ile (I)	C	Ala (A)	C	Thr (T)	T	Gly (G)	G	Ala (A)	T	Phe (F)	T	Ser (S)	A	Asn (N)	T	Ser (S)	C	Thr (T)	A	Gln (Q)	G	Asp (D)	C	Asp (D)	G	Val (V)	G	Met (M)							
SARS-CoV-2, Isolate USA-HI498/2020_B.1.243	A	His (H)	A	His (H)	A	Ile (I)	C	Ala (A)	C	Thr (T)	C	Gly (G)	G	Ala (A)	T	Phe (F)	T	Ser (S)	G	Asp (D)	T	Ser (S)	C	Thr (T)	A	Gln (Q)	G	Asp (D)	C	Asp (D)	G	Val (V)	G	Met (M)							

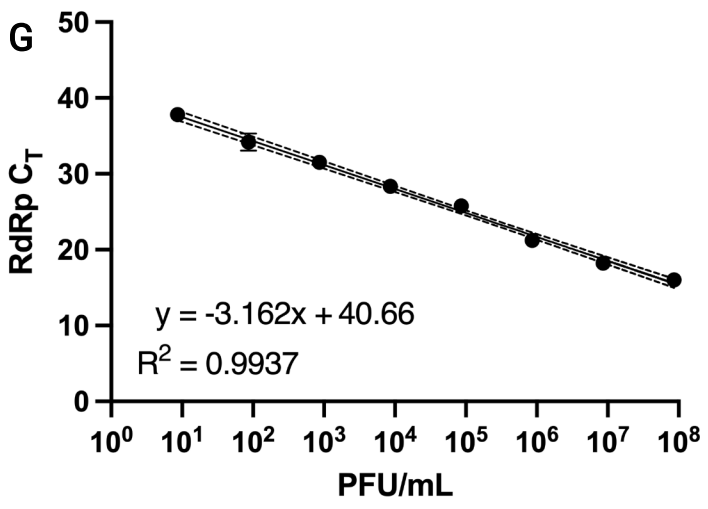
*combine to code for one Leucine



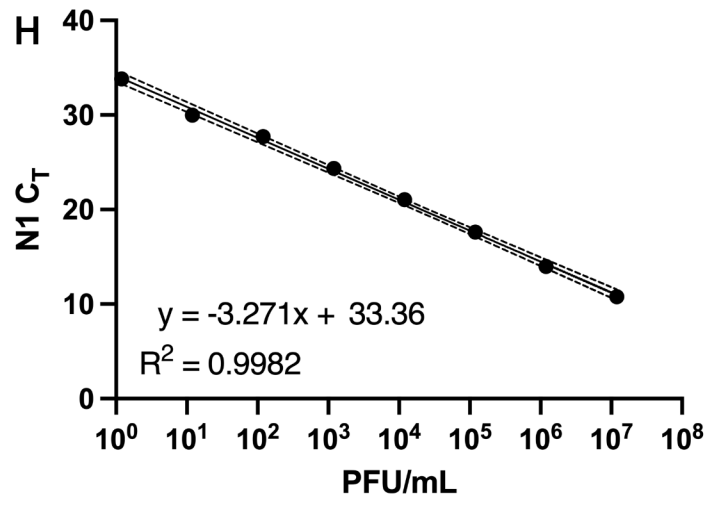
SARS-CoV-2, Isolate USA-WA1/2020

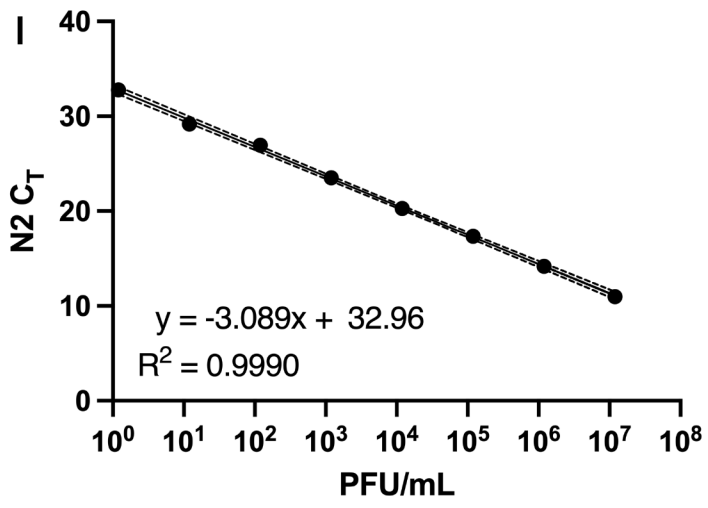


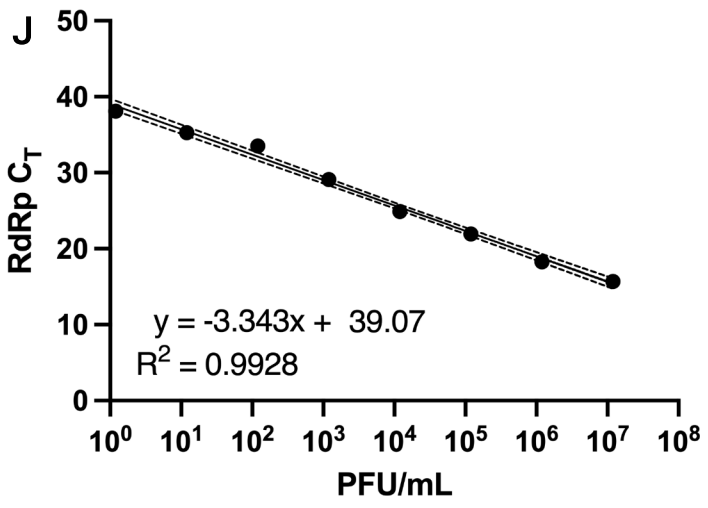




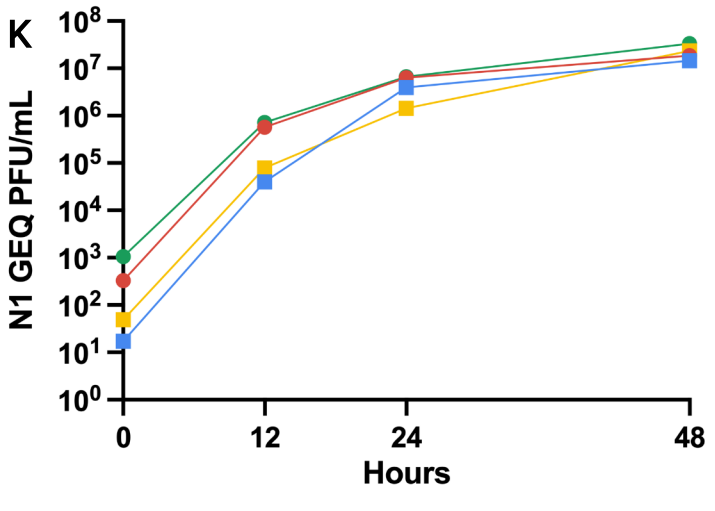
SARS-CoV-2, Isolate USA-HI498/2020

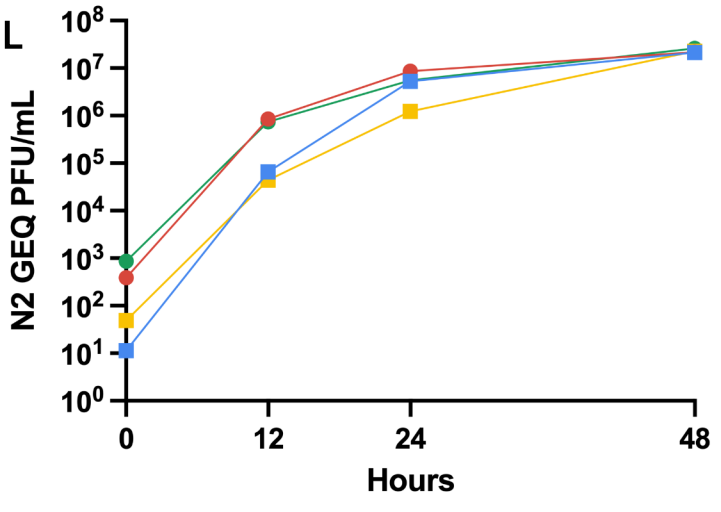


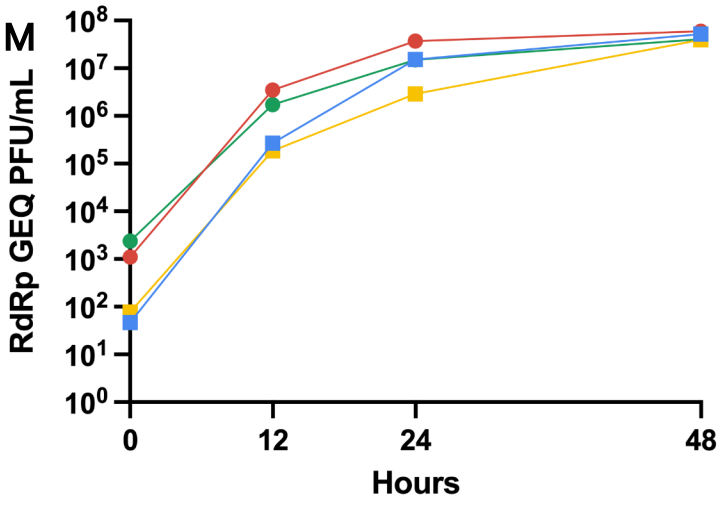




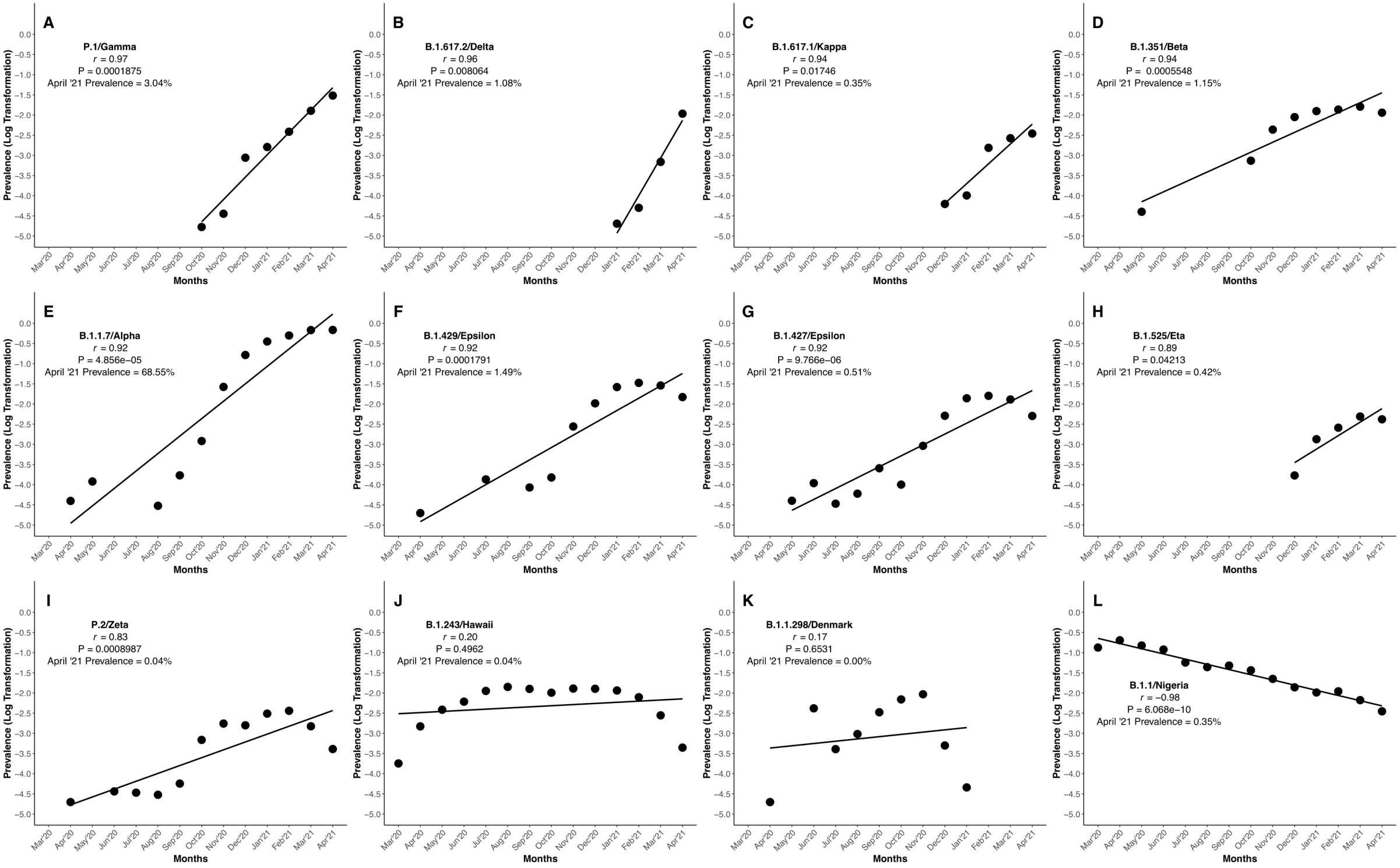
SARS-CoV-2, Isolate USA-WA1/2020 and Isolate USA-HI498/2020 Growth Kinetics



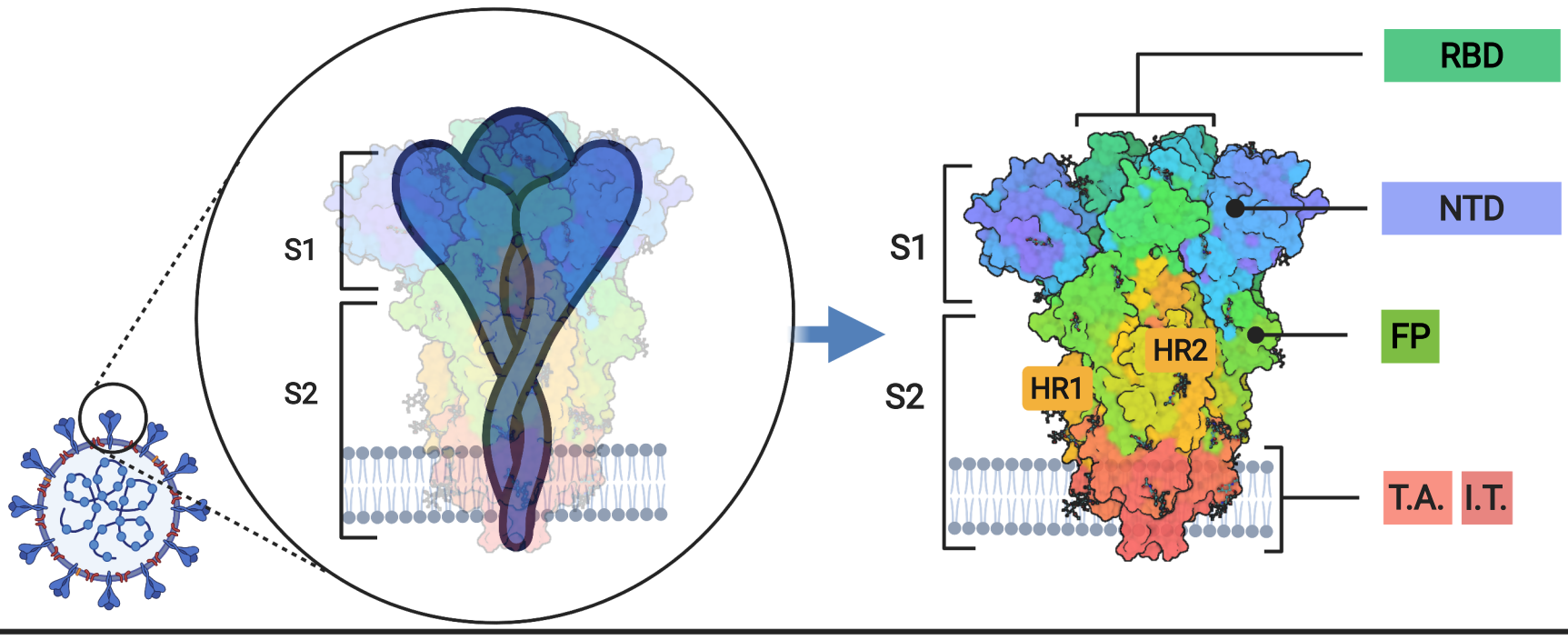




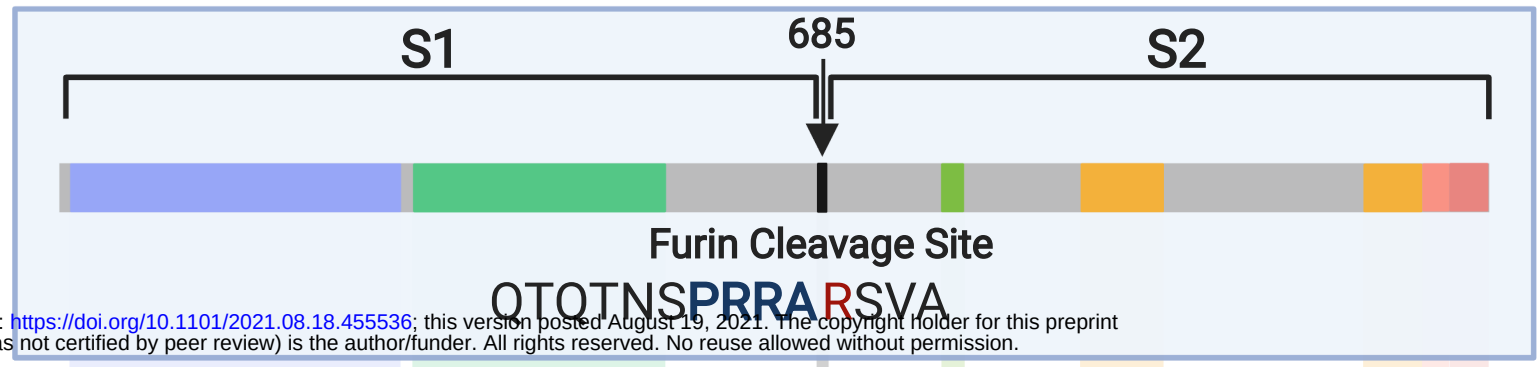
NEG SARS-2_WA MOI 0.1 SARS-2_WA MOI 1 SARS-2_HI-498 MOI 0.1 SARS-2_HI-498 MOI 1


A
P.1/Gamma
r = 0.97
P = 0.0001875
April '21 Prevalence = 3.04%
B
B.1.617.2/Delta
r = 0.96
P = 0.008064
April '21 Prevalence = 1.08%
C
B.1.617.1/Kappa
r = 0.94
P = 0.01746
April '21 Prevalence = 0.35%
D
B.1.351/Beta
r = 0.94
P = 0.0005548
April '21 Prevalence = 1.15%
E
B.1.1.7/Alpha
r = 0.92
P = 4.856e−05
April '21 Prevalence = 68.55%
F
B.1.429/Epsilon
r = 0.92
P = 0.0001791
April '21 Prevalence = 1.49%
G
B.1.427/Epsilon
r = 0.92
P = 9.766e−06
April '21 Prevalence = 0.51%
H
B.1.525/Eta
r = 0.89
P = 0.04213
April '21 Prevalence = 0.42%
I
P.2/Zeta
r = 0.83
P = 0.0008987
April '21 Prevalence = 0.04%
J
B.1.243/Hawaii
r = 0.20
P = 0.4962
April '21 Prevalence = 0.04%
K
B.1.1.298/Denmark
r = 0.17
P = 0.6531
April '21 Prevalence = 0.00%
L
B.1.1/Nigeria
r = −0.98
P = 6.068e−10
April '21 Prevalence = 0.35%
Prevalence (Log Transformation)
Months
Mar'20
Apr'20
May'20
Jun'20
Jul'20
Aug'20
Sep'20
Oct'20
Nov'20
Dec'20
Jan'21
Feb'21
Mar'21
Apr'21

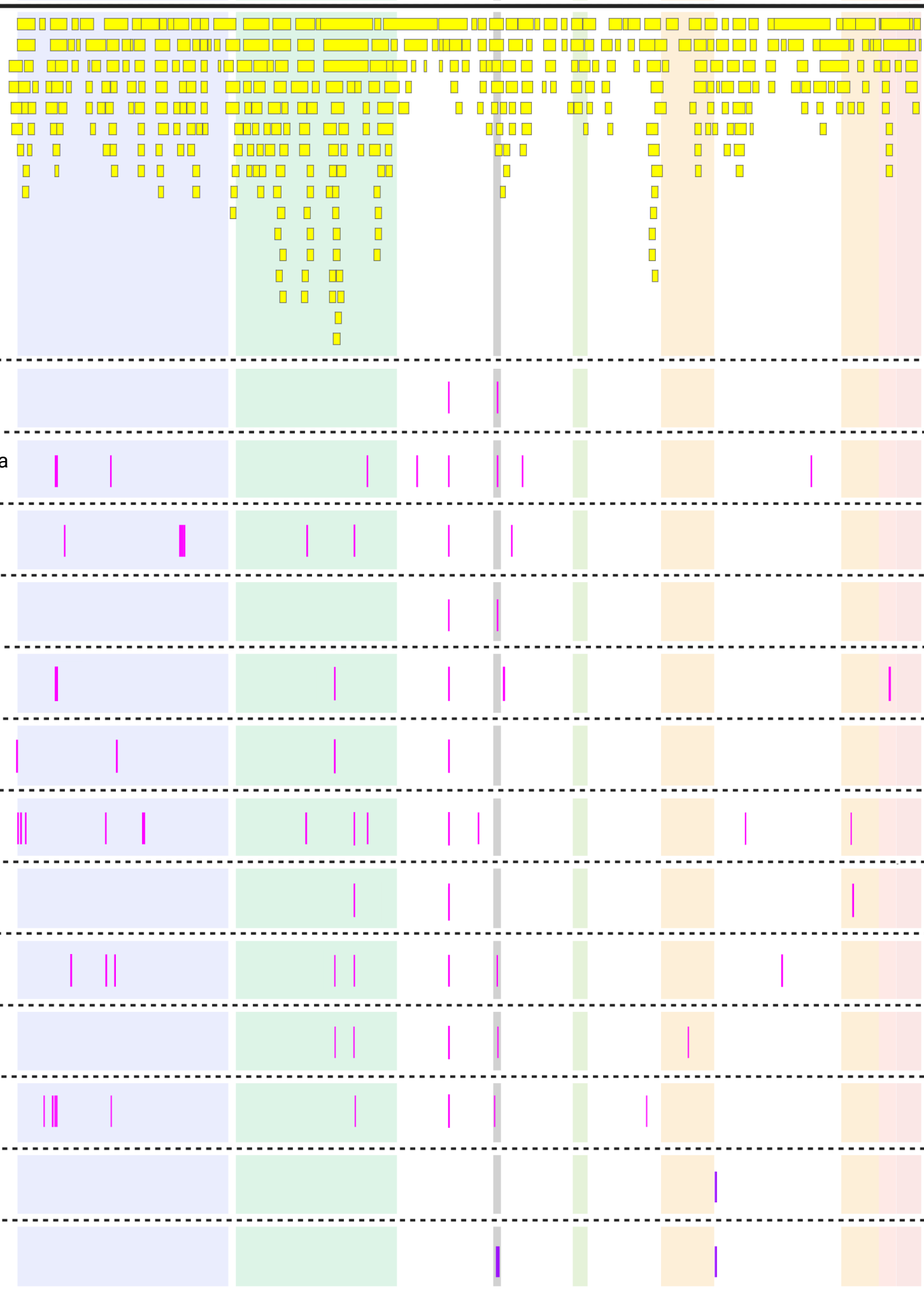
A)

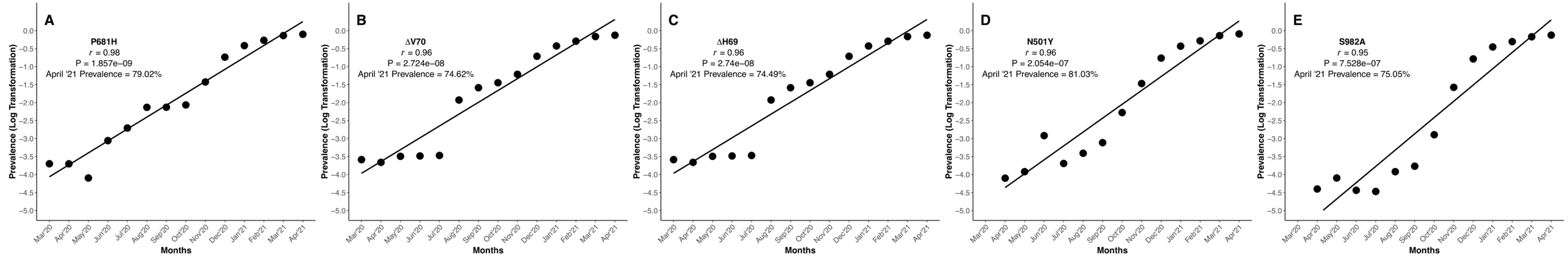


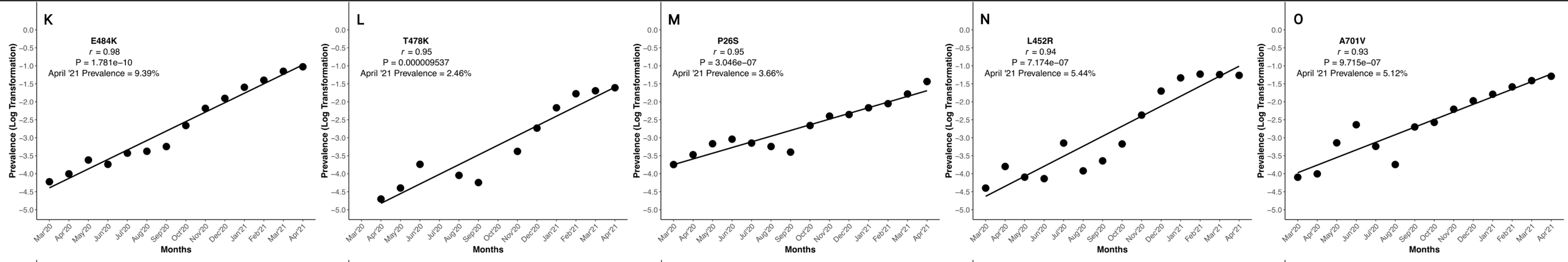
B)

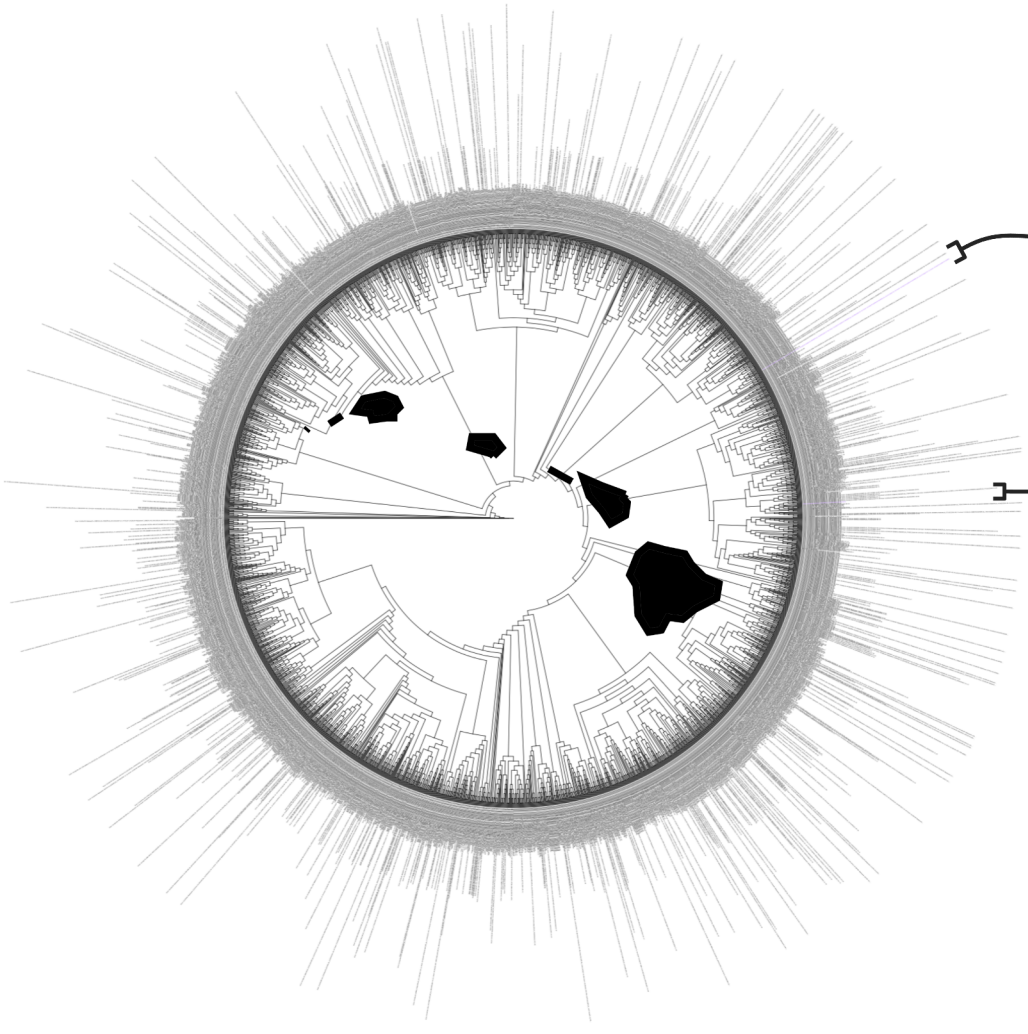


C)




A
P681H
r = 0.98
P = 1.857e−09
April '21 Prevalence = 79.02%
B
ΔV70
r = 0.96
P = 2.724e−08
April '21 Prevalence = 74.62%
C
ΔH69
r = 0.96
P = 2.74e−08
April '21 Prevalence = 74.49%
D
N501Y
r = 0.96
P = 2.054e−07
April '21 Prevalence = 81.03%
E
S982A
r = 0.95
P = 7.528e−07
April '21 Prevalence = 75.05%
F
D1118H
r = 0.93
P = 3.227e−06
April '21 Prevalence = 75.21%
G
T716I
r = 0.93
P = 1.806e−06
April '21 Prevalence = 75.99%
H
A570D
r = 0.93
P = 4.415e−05
April '21 Prevalence = 76.71%
I
ΔY144
r = 0.90
P = 1.299e−05
April '21 Prevalence = 75.49%
J
D614G
r = 0.67
P = 0.008904
April '21 Prevalence = 99.24%
Prevalence (Log Transformation)
Months
Mar'20
Apr'20
May'20
Jun'20
Jul'20
Aug'20
Sep'20
Oct'20
Nov'20
Dec'20
Jan'21
Feb'21
Mar'21
Apr'21


K
E484K
r = 0.98
P = 1.781e−10
April '21 Prevalence = 9.39%
L
T478K
r = 0.95
P = 0.000009537
April '21 Prevalence = 2.46%
M
P26S
r = 0.95
P = 3.046e−07
April '21 Prevalence = 3.66%
N
L452R
r = 0.94
P = 7.174e−07
April '21 Prevalence = 5.44%
O
A701V
r = 0.93
P = 9.715e−07
April '21 Prevalence = 5.12%
P
W152C
r = 0.93
P = 1.209e−05
April '21 Prevalence = 2.12%
Q
T95I
r = 0.92
P = 2.297e−06
April '21 Prevalence = 6.06%
R
H655Y
r = 0.92
P = 3.073e−06
April '21 Prevalence = 3.89%
S
S13I
r = 0.92
P = 3.98e−06
April '21 Prevalence = 2.02%
T
D138Y
r = 0.92
P = 1.062e−05
April '21 Prevalence = 3.64%
Prevalence (Log Transformation)
Months
Mar'20
Apr'20
May'20
Jun'20
Jul'20
Aug'20
Sep'20
Oct'20
Nov'20
Dec'20
Jan'21
Feb'21
Mar'21
Apr'21



hCoV-19/USA/HI-H200452/2020|EPI_ISL_752995|2020-10-08
hCoV-19/USA/HI-H200573/2020|EPI_ISL_753100|2020-09-03
hCoV-19/USA/HI-H200260/2020|EPI_ISL_752819|2020-08-31
hCoV-19/USA/HI-H200288/2020|EPI_ISL_752847|2020-08-06=hCoV-19/USA/HI-H200289/2020|EPI_ISL_752848|20.
hCoV-19/USA/HI-H200333/2020|EPI_ISL_752891|2020-08-14
hCoV-19/USA/HI-H200241/2020|EPI_ISL_752800|2020-08-15=hCoV-19/USA/HI-H200264/2020|EPI_ISL_752823|20.
hCoV-19/USA/HI-H200187/2020|EPI_ISL_752746|2020-07-31
hCoV-19/USA/HI-H200170/2020|EPI_ISL_752731|2020-07-24
hCoV-19/USA/HI-H200245/2020|EPI_ISL_752804|2020-08-17
hCoV-19/USA/HI-H200314/2020|EPI_ISL_752873|2020-08-16
hCoV-19/USA/HI-H200514/2020|EPI_ISL_753050|2020-08-16=hCoV-19/USA/HI-H200210/2020|EPI_ISL_752769|20.
hCoV-19/USA/HI-H200243/2020|EPI_ISL_752802|2020-08-16
hCoV-19/USA/HI-H200442/2020|EPI_ISL_752985|2020-10-07
hCoV-19/USA/WA-UW-60677/2021|EPI_ISL_1069331|2021-01-29
hCoV-19/USA/GA-CDC-2-3806781/2020|EPI_ISL_1167966|2020-12-14
hCoV-19/USA/HI-H200174/2020|EPI_ISL_752734|2020-07-28
hCoV-19/USA/HI-H200472/2020|EPI_ISL_753015|2020-10-10=hCoV-19/USA/HI-H200133/2020|EPI_ISL_752696|20.
hCoV-19/USA/HI-H200271/2020|EPI_ISL_752830|2020-09-02
hCoV-19/USA/HI-H200351/2020|EPI_ISL_752907|2020-09-15
hCoV-19/USA/HI-H200354/2020|EPI_ISL_752910|2020-09-17
hCoV-19/Mexico/ZAC-INER-IMSS-00006/2021|EPI_ISL_1279309|2021-02-08
hCoV-19/USA/HI-H200492/2020|EPI_ISL_753031|2020-10-06
hCoV-19/USA/HI-H200423/2020|EPI_ISL_752967|2020-09-23
hCoV-19/USA/HI-H200147/2020|EPI_ISL_752710|2020-03-29
hCoV-19/USA/CA-ALSR-4534/2020|EPI_ISL_666993|2020-08-27
hCoV-19/USA/CA-ALSR-3472/2020|EPI_ISL_636044|2020-08-18
hCoV-19/USA/CA-ALSR-3067/2020|EPI_ISL_635826|2020-08-27=hCoV-19/USA/CA-ALSR-3145/2020|EPI_ISL_63587.
hCoV-19/USA/CA-ALSR-3137/2020|EPI_ISL_635868|2020-08-21
hCoV-19/USA/HI-H200318/2020|EPI_ISL_752877|2020-08-14
hCoV-19/USA/HI-H200339/2020|EPI_ISL_752896|2020-08-17
hCoV-19/USA/HI-H200255/2020|EPI_ISL_752814|2020-08-19
hCoV-19/USA/HI-H200349/2020|EPI_ISL_752905|2020-09-15=hCoV-19/USA/HI-H200455/2020|EPI_ISL_752998|20.
hCoV-19/USA/WA-S2819/2020|EPI_ISL_574661|2020-09-07

hCoV-19/USA/HI-H200358/2020|EPI_ISL_752914|2020-09-25
hCoV-19/USA/HI-H200519/2020|EPI_ISL_753053|2020-09-24=hCoV-19/USA/HI-H200456/2020|EPI_ISL_752999|20...
hCoV-19/Northern=hCoV-19/Northern
hCoV-19/USA/WA-UW-32754/2020|EPI_ISL_737087|2020-10-19
hCoV-19/USA/HI-H200348/2020|EPI_ISL_752904|2020-09-11
hCoV-19/USA/HI-H200253/2020|EPI_ISL_752812|2020-08-18
hCoV-19/USA/HI-H200162/2020|EPI_ISL_752723|2020-07-24
hCoV-19/USA/HI-H200388/2020|EPI_ISL_752935|2020-09-29
hCoV-19/USA/HI-H200493/2020|EPI_ISL_753032|2020-10-07=hCoV-19/USA/HI-H200598/2020|EPI_ISL_753124|20...
hCoV-19/USA/HI-H200327/2020|EPI_ISL_752885|2020-08-20
hCoV-19/USA/HI-H200505/2020|EPI_ISL_753044|2020-08-11=hCoV-19/USA/HI-H200328/2020|EPI_ISL_752886|20...
hCoV-19/USA/HI-H200329/2020|EPI_ISL_752887|2020-08-20
hCoV-19/USA/HI-H200543/2020|EPI_ISL_753074|2020-11-05=hCoV-19/USA/HI-H200544/2020|EPI_ISL_753075|20...
hCoV-19/USA/HI-H200545/2020|EPI_ISL_753076|2020-11-05
hCoV-19/USA/HI-H200579/2020|EPI_ISL_753106|2020-10-15=hCoV-19/USA/HI-H200580/2020|EPI_ISL_753107|20...
hCoV-19/USA/HI-H200578/2020|EPI_ISL_753105|2020-10-15
hCoV-19/USA/HI-H200467/2020|EPI_ISL_753010|2020-10-09=hCoV-19/USA/HI-H200473/2020|EPI_ISL_753016|20...
hCoV-19/USA/HI-H200446/2020|EPI_ISL_752989|2020-10-08
hCoV-19/USA/HI-H200411/2020|EPI_ISL_752955|2020-08-06=HI498
hCoV-19/USA/HI-H200326/2020|EPI_ISL_752884|2020-08-18
hCoV-19/USA/HI-H200504/2020|EPI_ISL_753043|2020-08-02
hCoV-19/USA/NM-NMDOH-2020327026/2020|EPI_ISL_569643|2020-09-03
hCoV-19/USA/CA-CZB-23046/2020|EPI_ISL_1027445|2020-12-03
hCoV-19/USA/CA-CZB-19421/2021|EPI_ISL_955396|2021-01-04
hCoV-19/USA/CA-CZB-25149/2020|EPI_ISL_1235256|2020-11-18
hCoV-19/USA/IL-NM-5086/2020|EPI_ISL_936750|2020-12-23
hCoV-19/USA/NM-NMDOH-2020327022/2020|EPI_ISL_569642|2020-09-03
hCoV-19/USA/NM-DOH-2020315751/2020|EPI_ISL_732714|2020-08-27=hCoV-19/USA/NM-DOH-2020317191/2020|EPI.
hCoV-19/USA/HI-H200261/2020|EPI_ISL_752820|2020-08-25
hCoV-19/USA/HI-H200426/2020|EPI_ISL_752970|2020-09-24
hCoV-19/USA/HI-H200266/2020|EPI_ISL_752825|2020-08-31
hCoV-19/USA/HI-H200171/2020|EPI_ISL_752732|2020-07-27
hCoV-19/USA/HI-H200360/2020|EPI_ISL_752916|2020-09-28
hCoV-19/USA/HI-H200440/2020|EPI_ISL_752983|2020-10-06
hCoV-19/USA/HI-H200469/2020|EPI_ISL_753012|2020-10-09
hCoV-19/USA/HI-H200459/2020|EPI_ISL_753002|2020-10-06=hCoV-19/USA/HI-H200460/2020|EPI_ISL_753003|20...
hCoV-19/USA/HI-H200468/2020|EPI_ISL_753011|2020-10-09=hCoV-19/USA/HI-H200438/2020|EPI_ISL_752981|20...
hCoV-19/USA/HI-H200461/2020|EPI_ISL_753004|2020-10-08

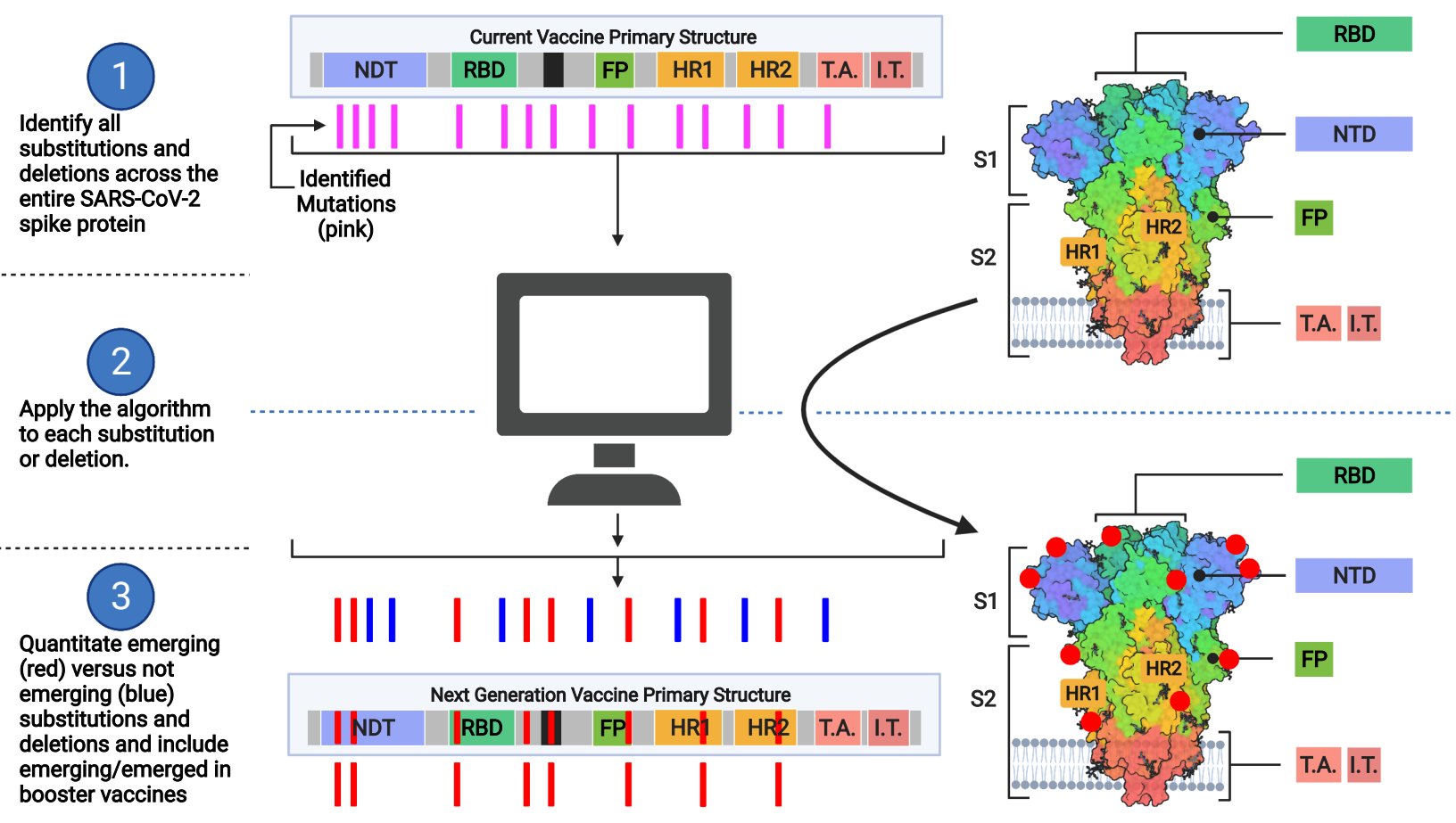

1
Identify all substitutions and deletions across the entire SARS-CoV-2 spike protein
2
Apply the algorithm to each substitution or deletion.
3
Quantitate emerging (red) versus not emerging (blue) substitutions and deletions and include emerging/emerged in booster vaccines
Current Vaccine Primary Structure
NDT
RBD
FP
HR1
HR2
T.A.
I.T.
Identified Mutations (pink)
Next Generation Vaccine Primary Structure
NDT
RBD
FP
HR1
HR2
T.A.
I.T.
RBD
NTD
FP
HR2
HR1
S1
S2
T.A.
I.T.
RBD
NTD
FP
HR2
HR1
S1
S2
T.A.
I.T.

Table 1. Genetic Characteristics of the Hawaii SARS-CoV-2 Variant B.1.243 Isolates, USA-HI498 2020* and USA-HI708 2020**

Gene	Nucleotide			Amino Acid		
	Loci	Wild Type	Mutant	Loci	Wild Type	Mutant
5' UTR	241	C	T	-	-	-
ORF1ab	3,037	C	T	924	Phe (F)	Phe (F)
ORF1ab	10,741	C	T	3492	Asp (D)	Asp (D)
ORF1ab	12,076***	C	T	3937	Asn (N)	Asn (N)
ORF1ab	14,408	C	T	4715	Pro (P)	Leu (L)
ORF1ab	20,268	A	G	6668	Leu (L)	Leu (L)
S	23,403	A	G	614	Asp (D)	Gly (G)
S	23,604	C	A	681	Pro (P)	His (H)
S	24,076	T	C	838	Gly (G)	Gly (G)
N	28,854	C	T	194	Ser (S)	Leu (L)
N	29,266***	G	A	331	Leu (L)	Leu (L)
3' UTR	29,710	T	C	-	-	-

GenBank accession *MZ664037 and **MZ664038, ***exclusive to SARS-CoV-2, Isolate USA-HI498 2020

Table 2. Pearson's Correlation of Logarithmic Transformed Prevalence Ratios of Amino Acid Substitutions (AAS) in the Spike Protein and Variants from March 2020 - April 2021 and Pairwise Heatmap of Absolute Value Difference of r-value Between Corresponding AAS and Variants

				Variant	P.1 Gamma	B.1.617.2 Delta	B.1.617.1 Kappa	B.1.351 Beta	B.1.1.7 Alpha	B.1.429 Epsilon	B.1.427 Epsilon	B.1.525 Eta	P.2 Zeta	B.1.243 -	B.1.1.298 -	B.1.1 -
				Origin	Brazil	India	India	South Africa	United Kingdom	California	California	United Kingdom	Brazil	Hawai'i	Denmark	Nigeria
				n	12,485	2,758	2,037	14,798	580,502	26,779	12,305	3,411	2,350	9,726	1,533	41,833
				April '21 Prevalence (%)	3.04	1.08	0.35	1.15	68.55	1.49	0.51	0.42	0.04	0.04	0.00	0.35
				P Value	0.00	0.01	0.02	0.00	0.00	0.00	0.00	0.04	0.00	0.50	0.65	0.00
Substitution	**n**	**April '21 Prevalence (%)**	**P value**	**r**	0.97	0.96	0.94	0.94	0.92	0.92	0.92	0.89	0.83	0.20	0.17	-0.98
E484K	60,990	9.39	0.000	0.98	0.010			0.046				0.093	0.157		0.810	
P681H	641,501	79.02	0.000	0.98					0.053					0.779		1.959
A845S	4,830	0.59	0.000	0.97												1.950
ΔV70	614,413	74.62	0.000	0.96					0.041			0.073			0.790	
ΔH69	613,401	74.49	0.000	0.96					0.040			0.073			0.790	
N501Y	636,491	81.03	0.000	0.96	0.015			0.021	0.036				0.133		0.785	
S982A	594,828	75.05	0.000	0.95					0.025							
A67V	5,784	0.74	0.000	0.95								0.057				
T478K	17,655	2.46	0.000	0.95		0.017										
P26S	18,361	3.66	0.000	0.95	0.029								0.119			
P681R	8,571	1.71	0.000	0.94		0.022	0.002									
L452R	56,039	5.44	0.000	0.94		0.026	0.002			0.020	0.020					
A701V	35,639	5.12	0.000	0.93				0.004								
D1118H	595,048	75.21	0.000	0.93					0.009							
W152C	37,487	2.12	0.000	0.93						0.012	0.012					
Q677H	23,205	1.24	0.000	0.93								0.037				
T716I	608,252	75.99	0.000	0.93					0.004							
A570D	609,134	76.71	0.000	0.93					0.002							
T95I	31,010	6.06	0.000	0.92			0.016									
K417N	15,367	1.22	0.000	0.92				0.015								
H655Y	18,155	3.89	0.000	0.92	0.055								0.093		0.746	
ΔL242	14,464	1.23	0.000	0.92				0.019								
Q52R	3,138	0.44	0.000	0.92								0.026				
S13I	33,379	2.02	0.000	0.92						0.001	0.001					
D138Y	19,518	3.64	0.000	0.92	0.058								0.089		0.742	
ΔY144	600,274	75.49	0.000	0.90					0.026			0.006				
T20N	13,656	3.47	0.001	0.89	0.087								0.061			
T1027I	14,941	3.62	0.000	0.87	0.104								0.043			
ΔA243	15,132	1.28	0.000	0.87				0.068								
R190S	13,441	3.35	0.000	0.87	0.105								0.042		0.695	
ΔL244	14,735	1.26	0.000	0.87				0.069								
K417T	13,226	3.47	0.007	0.86	0.118								0.029		0.682	
D80A	15,126	1.21	0.000	0.86				0.083								
E154K	1,505	0.26	0.002	0.83			0.109									
L18F	91,123	4.71	0.001	0.78	0.190								0.043			
E484Q	2,447	0.37	0.001	0.78			0.157									
Q1071H	1,783	0.33	0.034	0.74			0.196									
F888L	3,698	0.47	0.012	0.72								0.170				
ΔY145	482	0.03	0.005	0.70					0.220			0.188				
D614G	1,446,840	99.24	0.009	0.67	0.306	0.295	0.271	0.270	0.255	0.249	0.249	0.223	0.158	0.470	0.494	1.650
M1229I	2,058	0.18	0.010	0.66											0.488	
V1176F	19,419	3.58	0.012	0.65	0.327								0.179			
R246I	290	0.01	0.028	0.61				0.331								
G142D	3,190	0.92	0.050	0.60			0.338									
D950N	2,895	1.07	0.023	0.60		0.363										
K150N	180	0.04	0.040	0.60												
V308L	1,212	0.08	0.060	0.51												
Y453F	1,689	0.00	0.030	-0.62											0.799	
I692V	19	0.00	0.030	-0.97											1.145	

Table 3. Comparison of Single Nucleotide Polymorphisms (SNP) and Resultant Amino Acid Substitutions (AAS) or Deletions in the Spike Gene Among SARS-CoV-2 Variants

ACCESSION AND IDENTIFIER	SNP and AAS																															
	S13I		L18F		T20N		P26S		Q52R		A67V		ΔH69		ΔV70		D80A		T95I		D138Y		G142D		Y144		ΔY145		W152C		E154K		R190S	
	21,600 NT	13 AA	21,614 NT	18 AA	21,621 NT	20 AA	21,638 NT	26 AA	21,717 NT	52 AA	21,762 NT	67 AA	21,766-8 NT	69 AA	21,769-71 NT	70 AA	21,801 NT	80 AA	21,846 NT	95 AA	21,974 NT	138 AA	21,987 NT	142 AA	21,994 NT	144 AA	21,995-7 NT	145 AA	22,018 NT	152 AA	22,022 NT	154 AA	22,132 NT	190 AA
NC_045512_Reference Genome_Wuhan	G	Ser (S)	C	Leu (L)	C	Thr (T)	C	Pro (P)	A	Gln (Q)	C	Ala (A)	ACA	His (H)	TGT	Val (V)	A	Asp (D)	C	Thr (T)	G	Asp (D)	G	Gly (G)	T	Tyr (Y)	TAC	Tyr (Y)	G	Trp (W)	G	Glu (E)	G	Arg (R)
EPI_ISL_601443_UK_B.1.1.7_Alpha	G	Ser (S)	C	Leu (L)	C	Thr (T)	C	Pro (P)	A	Gln (Q)	C	Ala (A)	---	Δ	---	Δ	A	Asp (D)	C	Thr (T)	G	Asp (D)	G	Gly (G)	C	Tyr (Y)	---	Δ	G	Trp (W)	G	Glu (E)	G	Arg (R)
EPI_ISL_712081_South Africa_B.1.351_Beta	G	Ser (S)	C	Leu (L)	C	Thr (T)	C	Pro (P)	A	Gln (Q)	C	Ala (A)	ACA	His (H)	TGT	Val (V)	C	Ala (A)	C	Thr (T)	G	Asp (D)	G	Gly (G)	T	Tyr (Y)	TAC	Tyr (Y)	G	Trp (W)	G	Glu (E)	G	Arg (R)
EPI_ISL_729975_Nigeria_B.1.1	G	Ser (S)	C	Leu (L)	C	Thr (T)	C	Pro (P)	A	Gln (Q)	C	Ala (A)	ACA	His (H)	TGT	Val (V)	A	Asp (D)	C	Thr (T)	G	Asp (D)	G	Gly (G)	T	Tyr (Y)	TAC	Tyr (Y)	G	Trp (W)	G	Glu (E)	G	Arg (R)
EPI_ISL_616802_Denmark_B.1.1.298	G	Ser (S)	C	Leu (L)	C	Thr (T)	C	Pro (P)	A	Gln (Q)	C	Ala (A)	---	Δ	---	Δ	A	Asp (D)	C	Thr (T)	G	Asp (D)	G	Gly (G)	T	Tyr (Y)	TAC	Tyr (Y)	G	Trp (W)	G	Glu (E)	G	Arg (R)
EPI_ISL_942929_L452R_B.1.427/429_Epsilon	T	Ile (I)	C	Leu (L)	C	Thr (T)	C	Pro (P)	A	Gln (Q)	C	Ala (A)	ACA	His (H)	TGT	Val (V)	A	Asp (D)	C	Thr (T)	G	Asp (D)	G	Gly (G)	T	Tyr (Y)	TAC	Tyr (Y)	T	Cys (C)	G	Glu (E)	G	Arg (R)
EPI_ISL_792680_Brazil_B.1.1.248/P.1_Gamma	G	Ser (S)	T	Phe (F)	A	Asn (N)	T	Ser (S)	A	Gln (Q)	C	Ala (A)	ACA	His (H)	TGT	Val (V)	A	Asp (D)	C	Thr (T)	T	Tyr (Y)	G	Gly (G)	T	Tyr (Y)	TAC	Tyr (Y)	G	Trp (W)	G	Glu (E)	T	Ser (S)
EPI_ISL_918536_Brazil_P.2_Zeta	G	Ser (S)	C	Leu (L)	C	Thr (T)	C	Pro (P)	A	Gln (Q)	C	Ala (A)	ACA	His (H)	TGT	Val (V)	A	Asp (D)	C	Thr (T)	G	Asp (D)	G	Gly (G)	T	Tyr (Y)	TAC	Tyr (Y)	G	Trp (W)	G	Glu (E)	G	Arg (R)
EPI_ISL_1739895_B.1.525_Eta	G	Ser (S)	C	Leu (L)	C	Thr (T)	C	Pro (P)	G	Arg (R)	T	Val (V)	---	Δ	---	Δ	A	Asp (D)	C	Thr (T)	G	Asp (D)	G	Gly (G)	C	Tyr (Y)	---	Δ	G	Trp (W)	G	Glu (E)	G	Arg (R)
EPI_ISL_1372093_India_B.1.617.1_Kappa	G	Ser (S)	C	Leu (L)	C	Thr (T)	C	Pro (P)	A	Gln (Q)	C	Ala (A)	ACA	His (H)	TGT	Val (V)	A	Asp (D)	T	Ile (I)	G	Asp (D)	A	Asp (D)	T	Tyr (Y)	TAC	Tyr (Y)	G	Trp (W)	A	Lys (K)	G	Arg (R)
EPI_ISL_1663516_India_B.1.617.2_Delta	G	Ser (S)	C	Leu (L)	C	Thr (T)	C	Pro (P)	A	Gln (Q)	C	Ala (A)	ACA	His (H)	TGT	Val (V)	A	Asp (D)	C	Thr (T)	G	Asp (D)	G	Gly (G)	T	Tyr (Y)	TAC	Tyr (Y)	G	Trp (W)	G	Glu (E)	G	Arg (R)
SARS-CoV-2, Isolate USA-HI498/2020_B.1.243	G	Ser (S)	C	Leu (L)	C	Thr (T)	C	Pro (P)	A	Gln (Q)	C	Ala (A)	ACA	His (H)	TGT	Val (V)	A	Asp (D)	C	Thr (T)	G	Asp (D)	G	Gly (G)	T	Tyr (Y)	TAC	Tyr (Y)	G	Trp (W)	G	Glu (E)	G	Arg (R)

ACCESSION AND IDENTIFIER	SNP and AAS																															
	L241		ΔL242		ΔA243		ΔL244		R246I		K417T		K417N		L452R		Y453F		T478K		E484K		E484Q		N501Y		A570D		D614G		H655Y		Q677H	
	22,284-5 NT	241 AA	22,286-8 NT	242 AA	22,289-91 NT	243 AA	22,292 NT	244 AA	22,299 NT	246 AA	22,812 NT	417 AA	22,813 NT	417 AA	22,917 NT	452 AA	22,920 NT	453 AA	22,995 NT	478 AA	23,012 NT	484 AA	23,012 NT	484 AA	23,063 NT	501 AA	23,271 NT	570 AA	23,403 NT	614 AA	23,525 NT	655 AA	23,593 NT	677 AA
NC_045512_Reference Genome_Wuhan	TA	Leu (L)	CTT	Leu (L)	GCT	Ala (A)	T	Leu (L)	G	Arg (R)	A	Lys (K)	G	Lys (K)	T	Leu (L)	A	Tyr (Y)	C	Thr (T)	G	Glu (E)	G	Glu (E)	A	Asn (N)	C	Ala (A)	A	Asp (D)	C	His (H)	G	Gln (Q)
EPI_ISL_601443_UK_B.1.1.7_Alpha	TA	Leu (L)	CTT	Leu (L)	GCT	Ala (A)	T	Leu (L)	G	Arg (R)	A	Lys (K)	G	Lys (K)	T	Leu (L)	A	Tyr (Y)	C	Thr (T)	G	Glu (E)	G	Glu (E)	T	Tyr (Y)	A	Asp (D)	G	Gly (G)	C	His (H)	G	Gln (Q)
EPI_ISL_712081_South Africa_B.1.351_Beta	--	Leu (L)*	---	Δ	---	Δ	-	Leu (L)*	T	Ile (I)	A	Lys (K)	T	Asn (N)	T	Leu (L)	A	Tyr (Y)	C	Thr (T)	A	Lys (K)	A	Lys (K)	T	Tyr (Y)	C	Ala (A)	G	Gly (G)	C	His (H)	G	Gln (Q)
EPI_ISL_729975_Nigeria_B.1.1	TA	Leu (L)	CTT	Leu (L)	GCT	Ala (A)	T	Leu (L)	G	Arg (R)	A	Lys (K)	G	Lys (K)	T	Leu (L)	A	Tyr (Y)	C	Thr (T)	G	Glu (E)	G	Glu (E)	A	Asn (N)	C	Ala (A)	G	Gly (G)	C	His (H)	G	Gln (Q)
EPI_ISL_616802_Denmark_B.1.1.298	TA	Leu (L)	CTT	Leu (L)	GCT	Ala (A)	T	Leu (L)	G	Arg (R)	A	Lys (K)	G	Lys (K)	T	Leu (L)	T	Phe (F)	C	Thr (T)	G	Glu (E)	G	Glu (E)	A	Asn (N)	C	Ala (A)	G	Gly (G)	C	His (H)	G	Gln (Q)
EPI_ISL_942929_L452R_B.1.427/429_Epsilon	TA	Leu (L)	CTT	Leu (L)	GCT	Ala (A)	T	Leu (L)	G	Arg (R)	A	Lys (K)	G	Lys (K)	G	Arg (R)	A	Tyr (Y)	C	Thr (T)	G	Glu (E)	G	Glu (E)	A	Asn (N)	C	Ala (A)	G	Gly (G)	C	His (H)	G	Gln (Q)
EPI_ISL_792680_Brazil_B.1.1.248/P.1_Gamma	TA	Leu (L)	CTT	Leu (L)	GCT	Ala (A)	T	Leu (L)	G	Arg (R)	C	Thr (T)	G	Lys (K)	T	Leu (L)	A	Tyr (Y)	C	Thr (T)	A	Lys (K)	A	Lys (K)	T	Tyr (Y)	C	Ala (A)	G	Gly (G)	T	Tyr (Y)	G	Gln (Q)
EPI_ISL_918536_Brazil_P.2_Zeta	TA	Leu (L)	CTT	Leu (L)	GCT	Ala (A)	T	Leu (L)	G	Arg (R)	A	Lys (K)	G	Lys (K)	T	Leu (L)	A	Tyr (Y)	C	Thr (T)	A	Lys (K)	A	Lys (K)	A	Asn (N)	C	Ala (A)	G	Gly (G)	C	His (H)	G	Gln (Q)
EPI_ISL_1739895_B.1.525_Eta	TA	Leu (L)	CTT	Leu (L)	GCT	Ala (A)	T	Leu (L)	G	Arg (R)	A	Lys (K)	G	Lys (K)	T	Leu (L)	A	Tyr (Y)	C	Thr (T)	A	Lys (K)	A	Lys (K)	A	Asn (N)	C	Ala (A)	G	Gly (G)	C	His (H)	C	His (H)
EPI_ISL_1372093_India_B.1.617.1_Kappa	TA	Leu (L)	CTT	Leu (L)	GCT	Ala (A)	T	Leu (L)	G	Arg (R)	A	Lys (K)	G	Lys (K)	G	Arg (R)	A	Tyr (Y)	C	Thr (T)	C	Gln (Q)	C	Gln (Q)	A	Asn (N)	C	Ala (A)	G	Gly (G)	C	His (H)	G	Gln (Q)
EPI_ISL_1663516_India_B.1.617.2_Delta	TA	Leu (L)	CTT	Leu (L)	GCT	Ala (A)	T	Leu (L)	G	Arg (R)	A	Lys (K)	G	Lys (K)	G	Arg (R)	A	Tyr (Y)	A	Lys (K)	C	Gln (Q)	G	Glu (E)	A	Asn (N)	C	Ala (A)	G	Gly (G)	C	His (H)	G	Gln (Q)
SARS-CoV-2, Isolate USA-HI498/2020_B.1.243	TA	Leu (L)	CTT	Leu (L)	GCT	Ala (A)	T	Leu (L)	G	Arg (R)	A	Lys (K)	G	Lys (K)	T	Leu (L)	A	Tyr (Y)	C	Thr (T)	G	Glu (E)	G	Glu (E)	A	Asn (N)	C	Ala (A)	G	Gly (G)	C	His (H)	G	Gln (Q)

ACCESSION AND IDENTIFIER	SNP and AAS																															
	P681H		P681R		I692V		A701V		T716I		G838		A845S		F888L		S929		D950N		S982A		T1027I		Q1071H		D1118H		D1146		V1176F		M1229I	
	23,604 NT	681 AA	23,604 NT	681 AA	23,636 NT	692 AA	23,644 NT	701 AA	23,709 NT	716 AA	24,076 NT	838 AA	24,095 NT	845 AA	24,224 NT	888 AA	24,349 NT	929 AA	24,410 NT	950 AA	24,506 NT	982 AA	24,642 NT	1027 AA	24,775 NT	1071 AA	24,914 NT	1118 AA	25,000 NT	1146 AA	25,088 NT	1176 AA	25,249 NT	1229 AA
NC_045512_Reference Genome_Wuhan	C	Pro (P)	C	Pro (P)	A	Ile (I)	C	Ala (A)	C	Thr (T)	T	Gly (G)	G	Ala (A)	T	Phe (F)	T	Ser (S)	G	Asp (D)	T	Ser (S)	C	Thr (T)	A	Gln (Q)	G	Asp (D)	C	Asp (D)	G	Val (V)	G	Met (M)
EPI_ISL_601443_UK_B.1.1.7_Alpha	A	His (H)	A	His (H)	A	Ile (I)	C	Ala (A)	T	Ile (I)	T	Gly (G)	G	Ala (A)	T	Phe (F)	T	Ser (S)	G	Asp (D)	G	Ala (A)	C	Thr (T)	A	Gln (Q)	C	His (H)	C	Asp (D)	G	Val (V)	G	Met (M)
EPI_ISL_712081_South Africa_B.1.351_Beta	C	Pro (P)	C	Pro (P)	A	Ile (I)	T	Val (V)	C	Thr (T)	T	Gly (G)	G	Ala (A)	T	Phe (F)	T	Ser (S)	G	Asp (D)	T	Ser (S)	C	Thr (T)	A	Gln (Q)	G	Asp (D)	C	Asp (D)	G	Val (V)	G	Met (M)
EPI_ISL_729975_Nigeria_B.1.1	A	His (H)	A	His (H)	A	Ile (I)	C	Ala (A)	C	Thr (T)	T	Gly (G)	T	Ser (S)	T	Phe (F)	T	Ser (S)	G	Asp (D)	T	Ser (S)	C	Thr (T)	A	Gln (Q)	G	Asp (D)	C	Asp (D)	G	Val (V)	G	Met (M)
EPI_ISL_616802_Denmark_B.1.1.298	C	Pro (P)	C	Pro (P)	G	Val (V)	C	Ala (A)	C	Thr (T)	T	Gly (G)	G	Ala (A)	T	Phe (F)	T	Ser (S)	G	Asp (D)	T	Ser (S)	C	Thr (T)	A	Gln (Q)	G	Asp (D)	T	Asp (D)	G	Val (V)	T	Ile (I)
EPI_ISL_942929_L452R_B.1.427/429_Epsilon	C	Pro (P)	C	Pro (P)	A	Ile (I)	C	Ala (A)	C	Thr (T)	T	Gly (G)	G	Ala (A)	T	Phe (F)	C	Ser (S)	G	Asp (D)	T	Ser (S)	C	Thr (T)	A	Gln (Q)	G	Asp (D)	C	Asp (D)	G	Val (V)	G	Met (M)
EPI_ISL_792680_Brazil_B.1.1.248/P.1_Gamma	C	Pro (P)	C	Pro (P)	A	Ile (I)	C	Ala (A)	C	Thr (T)	T	Gly (G)	G	Ala (A)	T	Phe (F)	T	Ser (S)	G	Asp (D)	T	Ser (S)	T	Ile (I)	A	Gln (Q)	G	Asp (D)	C	Asp (D)	T	Phe (F)	G	Met (M)
EPI_ISL_918536_Brazil_P.2_Zeta	C	Pro (P)	C	Pro (P)	A	Ile (I)	C	Ala (A)	C	Thr (T)	T	Gly (G)	G	Ala (A)	T	Phe (F)	T	Ser (S)	G	Asp (D)	T	Ser (S)	C	Thr (T)	A	Gln (Q)	G	Asp (D)	C	Asp (D)	T	Phe (F)	G	Met (M)
EPI_ISL_1739895_B.1.525_Eta	C	Pro (P)	C	Pro (P)	A	Ile (I)	C	Ala (A)	C	Thr (T)	T	Gly (G)	G	Ala (A)	C	Leu (L)	T	Ser (S)	G	Asp (D)	T	Ser (S)	C	Thr (T)	A	Gln (Q)	G	Asp (D)	C	Asp (D)	G	Val (V)	G	Met (M)
EPI_ISL_1372093_India_B.1.617.1_Kappa	G	Arg (R)	G	Arg (R)	A	Ile (I)	C	Ala (A)	C	Thr (T)	T	Gly (G)	G	Ala (A)	T	Phe (F)	T	Ser (S)	G	Asp (D)	T	Ser (S)	C	Thr (T)	T	His (H)	G	Asp (D)	C	Asp (D)	G	Val (V)	G	Met (M)
EPI_ISL_1663516_India_B.1.617.2_Delta	G	Arg (R)	G	Arg (R)	A	Ile (I)	C	Ala (A)	C	Thr (T)	T	Gly (G)	G	Ala (A)	T	Phe (F)	T	Ser (S)	A	Asn (N)	T	Ser (S)	C	Thr (T)	A	Gln (Q)	G	Asp (D)	C	Asp (D)	G	Val (V)	G	Met (M)
SARS-CoV-2, Isolate USA-HI498/2020_B.1.243	A	His (H)	A	His (H)	A	Ile (I)	C	Ala (A)	C	Thr (T)	C	Gly (G)	G	Ala (A)	T	Phe (F)	T	Ser (S)	G	Asp (D)	T	Ser (S)	C	Thr (T)	A	Gln (Q)	G	Asp (D)	C	Asp (D)	G	Val (V)	G	Met (M)

*combine to code for one Leucine