Representative sequences of SARS-CoV-2 variants that are considered VOC^{TM}, VOI, or VUM by the World Health Organization.

Enabled by Data from GISAID

LANL: Bette Korber, James Theiler, Will Fischer, Hyejin Yoon
GISAID: Sebastian Maurer-Stroh

June 8, 2022

LA-UR-21-28226

I. WHO variants and additional VOIs of current or historic interest.

Designations: WHO Tracking SARS-CoV-2 variants
VOC, Variant of Concern
VOI, Variant of Interest
VUM, Variant Under Monitoring
VOC-LUM, VOC Lineages Under Monitoring

i. We are maintaining an alignment of natural sequences that correspond to the most representative forms for all of the SARS-CoV-2 variants that the WHO has noted to be of a VOC, VOI, or LUM. Key data about sampling frequency, geographic location, and Pango lineage designations is updated in the accompanying spread sheet.

ii. We also are maintaining and extended set of variants of interest. The alignments and the spreadsheet here open with the WHO representative sequence data at the top, but we also include many additional representative variants and sublineages of both current and historic interest in addition to the WHO listing of key variants.

The additional VOI's we have defined for part (ii) are based on our own selection strategies at Los Alamos based on GISAID data. These additional variants are included if they were sampled with a significantly increasing frequencies in multiple geographic locations over a distinct, generally transient, period of time during the pandemic. Our analyses for selection of these forms are based on actual Spike sequences, not Pango lineages designations, as some very distinctive forms of Spike are repeated in many different Pango lineages, and conversely, some Pango lineages contain very distinctive forms of Spike. Also, Pango lineage designations often change over time, while forms of Spike stay stable. Current information about the designated Pango lineages most associated with each Spike variant form is provided, however, to enable the variants to be recognized in the context of the current designations in the literature, and to appropriately reflect the World Health Organization SARS-CoV-2 variant nomenclature.

Spike Variant Data Table and Alignments

The accompanying spreadsheet lists a natural form of a SARS-CoV-2 sequence that was found to be most representative of each of the WHO designated VOC, VOI, or VUM (See "WHO Tracking SARS-CoV-2 variants": WHO Tracking SARS-CoV-2 variants), as well as additional VOIs. The spreadsheet provides details of mutational patterns and of sampling frequencies of these common forms both 60 days prior to this report and historically within entire collection at GISAID.

The accompanying fasta files include (i) a full-length genome codon aligned set of the each of the representative forms, and (ii) an alignment of Spike reference forms. We emphasize Spike to facilitate monitoring and reagent design for immune response and testing of enhanced infectivity. The most commonly sampled natural version of each variant listed as a “baseline form”; this form generally corresponds to an early consensus form of the lineage, and to the sequence that was of founder of the lineage. Reference strains with representative mutations in other regions of the genome that are most typical of particular forms of Spike are included in the full-length genome alignments.

Variants are continuously evolving and so the most representative sequence of a given form of Spike may change over time.

II. Reference Alignments for WHO GISAID Sequences

WHO-GISAID_RepresentativeSpikeSequences_2022-06-08.fasta

The team at Los Alamos National Lab (cov.lanl.gov) builds an alignment of all intact SARS-CoV-2 Spike sequences available in GISAID, and the most common form of the spike protein for each WHO variant is resolved. We then identify the most common nucleotide sequence encoding that spike variant, and the natural sequence with the lowest EPI_ISL number that perfectly matches that gene sequence is selected to represent that particular variant form of Spike.

WHO-GISAID_RepresentativeGenomes_CodonAligned_2022-06-08.fasta

We next build an alignment based on all full-length (ORF-complete) GISAID sequences that encodes the exact representative Spike form of interest. The consensus of the rest of the genome among that set is determined. The earliest-sampled intact natural sequence in GISAID EpiCoV that completely matches the full genome consensus is selected to be the representative sequence of that variant. If no intact natural sequence completely matching the consensus is available, the natural sequence that is closest to the consensus is selected (i.e. with the lowest Hamming distance); in practice, these have differed from the consensus by only a few bases.

We the build a codon-aligned genome alignment of these most representative sequences spanning all open reading frames, so the reading frame is preserved throughout the alignment of the full genome. At the top of the full genome alignment is the GISAID SARS-CoV-2 reference strain, immediately followed by a separate aligned sequence specifying each open reading frame/gene in isolation; this is intended as a guide to enable one to easily resolve where mutations are occurring throughout the genome.

III. Key to the Variants Table

Key to Variants Table. This spreadsheet/table contains the most common forms of Spike among WHO designated VOI/VOC/VUMs sampled in GISAID EpiCoV (www.gisaid.org), followed by other interesting variants.

Column headers with notes, Variants Table:

A. **WHO Classification group of Variant**
B. **Dates of WHO classifications, or if not a specific WHO classification, dates of addition to these GISAID tables.**
C. **WHO designation:** The WHO Greek letter variant designation from WHO Tracking SARS-CoV-2 variants; *indicates continued monitoring.
D. **Pango lineage(s)** most often associated with a particular WHO variant. Pango lineage descriptions can be found at Pango Lineages.
E. **Mutation list relative to baseline variant sequence.** WHO is monitoring the following Omicron mutations: R346K, L452X, and F486V. *Add* means add a mutation relative to the baseline from, *revert* means the Wuhan ancestral amino acid is restored at a given position.
F. **Most common Spike backbones.** Mutation list relative to the Wuhan reference strain WIV04/2019|EPI_ISL_402124 for the most common natural form of each Spike variant lineage.
   Note that "-" indicates a deletion at a site (e.g. Y144-, the Y at position 144 in the reference Spike is deleted), "+" indicates an insertion following the specified site (e.g. +143T indicates a T was inserted after position 143 in the reference Spike), and "_" indicates the ancestral value (*e.g.* D614_ indicates the ancestral D at site 614; it is equivalent to D614D).
   Colors highlight mutation in regions of interest:
   Blue: Addition of positive charge near the furin cleavage site: 675, 677, 681 are positively charged, or the H655Y substitution
   Green: NTD supersite: 13-20, 140-158, 242-264
   Magenta: RBD: 330-521
   Red: D614_, The ancestral Spike D614 amino acid is the dominant form in this lineage, the underscore indicates ancestral.
   Turquoise: The Heptad Repeat 1 region, HR1: 908-985
G. **Is the representative sequence available in the current Spike fasta file? (X == yes)**

H. **Is the representative sequence available in the Reference Genome fasta file? (X== yes)**

I. **Number of sequences that exactly match this pattern in Spike, full data.** This tally includes all data in our quality filtered GISAID data set starting in December 2019.

J. **Number of sequences that contain this pattern, full data.** The most common form of Spike representing a lineage is always part of an evolving lineage. This tally represents the number of variants that contain the full specified set of mutations (the "sequence backbone"), but that may also contain additional mutations. As lineages spread over time, they diversify, and the most common form becomes a smaller percentage of the total. As new variants of variants become more prevalent, we identify them as distinctive common forms, and add the most interesting of these to the “variants-of-variants” listing.

K. **Number of sequences that exactly match this pattern, last 60 days.** This tally is a rough indication of whether a particular form of Spike is still present in a contemporary GISAID global sampling, is declining and being replaced by other variants, or is no longer sampled.

L. **Number of sequences that contain this pattern, last 60 days. See above.**

M. **All Pango Lineages that contain sequences that exactly match this pattern in Spike (with counts).** Quite distinctive Spikes are often assigned to an array of Pango lineages. Some are closely related, but some not obviously phylogenetically related; such exceptions can arise due to recombination, or mis-classification; they also can change over time as Pango lineage designations can be reassigned. The current counts are based on Pango lineage assignments in the current data, after being filtered for quality control (QC) at cov.lanl.gov.

N. **All Pango Lineages that are associated with sequences that contain this pattern in Spike, including Spikes that also contain additional mutations (with counts).**

O. **Total count of sequences with the Pango lineage designation in which the Spike variant is most commonly found.** These counts are based on our QC filtered set used at cov.lanl.gov.

P. **Number of Spikes in the Pango lineage that exactly match this pattern.**

Q. **Fraction of the Pango lineage that exactly match this pattern (column R/Q).**

R. **Number in the Pango lineage that contain this pattern, but that also contain additional mutations.**

S. **Fraction of the Pango lineage sequences that contain this pattern (column T/Q).**

T. **The three countries where the exact Spike variant is most commonly sampled historically (with counts).**

U. **The three countries where Spikes that include the variant mutations are most commonly sampled historically (with counts).**

V. **The three countries where the exact Spike variant is most commonly sampled in the last 60 days (with counts).**

W. **The three countries where Spikes that include the variant mutations are most commonly sampled in the last 60 days (with counts).**

X. **The full sequence name used in the Spike fasta files**.

Y. **The lowest (first) EPI_ISL accession number among the set of sequences that exactly matches the pattern of Spike mutations listed in column F, used to create the Spike fasta file.**

Z. **The full sequence name used in the Spike full length genome fasta files**.

AA. **Brief notes regarding the lineage.**

Citation for these reference alignments:

GISAID

The tools we use for tracking SARS-CoV-2 variants were first described in Korber et al. Cell. 2020 Aug 20;182(4):812-827.e19. doi: 10.1016/j.cell.2020.06.043.

Further information can be found at cov.lanl.gov