Variants of the Spike protein that had or currently have a significant presence in GISAID

Bette Korber, James Theiler, Will Fischer, Hyejin Yoon

Oct. 29, 2021

LA-UR-21-28226

I. Identification and descriptions of spike variants

Variants-Table-Oct29-2021.xlsx
This table and the accompanying alignments will be updated every 3-5 weeks.

The accompanying spreadsheet (Variants-Table-*date*.xlsx) lists of natural forms of Spike that had a significant presence in the world starting in the spring of 2021. We also include a few additional very diverse Spikes as examples of more extreme mutation patterns. Representative forms of variants of interest and variants-of-variants of interest are provided. This listing focuses on Spike to facilitate monitoring and reagent design for immune response and testing of enhanced infectivity. Representative alignments of these variants are provided, as described in the next section

Details
The listing includes variants that were identified to be of interest using tools available at https://cov.lanl.gov, and confirmed to be relatively common and increasing in frequency in least in one local geographic region, even if only transiently, during 2021.

All of the World Health Organization WHO variants of interest and concern are captured in this list, as well as many additional variants. We begin by tracking related natural forms of Spike that increased in frequency among GISAID sequences in the most recent 60-day period. We then identify the most common circulating version of the related Spikes, and subsequently determine Pango lineage that is most commonly associated with that particular form. These natural forms of Spike often, but not always, correspond to the consensus of a Pango Lineage. Exceptions arise because sometimes Pango lineages are sometimes complex and may include multiple very distinctive forms of Spike; in such cases the Pango lineage consensus and the most common form of a variant of interest frequently differ. Furthermore, many of the highly distinctive forms of Spike that we are track are found in multiple Pango lineages; this information is summarized in the Variants-Table-*date*.xlsx file.

Starting in the late summer of 2021, emergent forms of the viral Spike have tended to be new variants of the well-established Delta variant. “Delta” includes Pango lineages B.1.617.2 and AY.* sublineages, where * represents a numbered series of Delta sublineages that continues to grow.

The Spikes listed in this file are carefully curated and cross-checked against relevant sequence alignments. They are monitored to resolve regions that are difficult to automatically align near deletions, to correctly identify conserved insertions in variants, and to consider the implications of stretches of unresolved sequence in defining the dominant Spike sequence of a newly emerging lineage.

This file will be updated every 3-4 weeks. Meanwhile, automatic "XSpike", "SHIVER", and "COMMON FORMS OF SPIKE WITH A GIVEN PANGO LINEAGE DESIGNATION" updates are run several times a week at cov.lanl.gov; the automatic runs of these codes can capture newly emergent variants as they are appearing in GISAID in closer to real time.

II. GISAID Reference Alignments

VOI_RepresentativeGenomes_CodonAligned_Oct29-2021.fasta
The team at cov.lanl.gov builds a Fasta file that includes a full length representative GISAID sequence for each of the distinct variant forms of Spike included in the Variants-Table-*date*.xlsx file. To identify representative sequences, all full-length (ORF-complete) GISAID sequences that encode the exact Spike form of interest are retrieved, and the consensus of the rest of the genome among that set is determined. The earliest-sampled natural sequence in GISAID that completely matches the consensus is selected as the representative sequence of that variant. If no 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) is selected; in practice, these have differed from the consensus by only a few bases.

Oct. 29, 2021: This alignment includes all representatives of all variants with distinct Pango lineage designations in the Variants Table, as well as a small set of particularly interesting variants of Delta.

This alignment is codon aligned so the reading frame is preserved throughout the alignment of the full genome. At the top of the alignment is the GISAID reference strain, followed by a separate aligned sequence of each open reading frame for easy reference.

VOI_RepresentativeSpikes_DNA_Oct29-2021.fasta
The team at cov.lanl.gov builds an alignment that includes every Spike variant in the Variants-Table-*date*.xlsx file. To build this alignment, all of the sequences encoding the exact sequence of the representative Spike are identified in GISAID, and then most common nucleotide sequence encoding that Spike variant is determined; the sequence with the lowest EPI_ISL number is selected to represent the sequence form.

Please contact us if you would like help retrieving a representative natural sequence for a variant we do not include in our set (seq-info@lanl.gov).

III. Key to the Variants Table

Key to Variants Table. This table contains the most common complex variant forms of Spike that were frequently sampled in GISAID (www.gisaid.org) at some point in 2021, identified using the LANL tools XSpike and SHIVER. Table counts were updated Oct. 22, 2021.

Column headers with notes, Variants Table:

A. **Classification group of Variant:** WHO VOI, VOC, Variant of Delta…
B. **Date of addition to cov.lanl.gov website listing**
C. **WHO designation:** The WHO Greek letter variant designation is as of Oct. 29, 2021 (WHO Tracking SARS-CoV-2 variants); *indicates continued monitoring.
D. **Pango lineage** most often associated with a particular Spike, as reported in GISAID as Oct. 22, 2021. Pango lineage descriptions can be found at Pango Lineages.
   - Bold letters: Variants we have identified through SHIVER, XSpike or Ember runs beginning March of 2021,
   - Non-bold letters: Variants tracked because they were listed as a WHO variant of interest, or considered of interest because they carried a very complex set of mutation.
E. **Most common Spike backbones.** The a most common natural form of a 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.
F. **Representative sequence available in the Spike fasta file**
G. **Representative sequence available in the Reference Genome fasta file**
H. **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.
I. **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.
J. **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.
K. **Number of sequences that contain this pattern, last 60 days. See above.**

L. **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 closely related, but some not obviously related. These might arise due to recombination, or mis-classification; they can change over time as Pango lineage designations can be reassigned. The current counts are based on Pango lineage assignments in the Oct. 22, 2021 GISAID data, filtered for quality control (QC) at cov.lanl.gov.
M. **All Pango Lineages associated with sequences that contain this pattern in Spike, but including Spikes that also can contain additional mutations (with counts).**
N. **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.
O. **Number of Spikes in the Pango lineage that exactly match this pattern.**
P. **Fraction of the Pango lineage that exactly match this pattern.**
Q. **Number in the Pango lineage that contain this pattern, but that also contain additional mutations.**
R. **Fraction of the Pango lineage sequences that contain this pattern.**
S. **The three countries where the exact Spike variant is most commonly sampled historically (with counts).**
T. **The three countries where Spikes that include the variant mutations are most commonly sampled historically (with counts).**
U. **The three countries where the exact Spike variant is most commonly sampled in the last 60 days (with counts).**
V. **The three countries where Spikes that include the variant mutations are most commonly sampled in the last 60 days (with counts).**
W. **An EPI_ISL accession number of a sequence that matches exactly matches the Spike mutation list.**
X. **Notes regarding insertions and deletions.** Please see the following pages for more detail
Y. **Notes.**

IV. Illustrative examples and background

Contents

Citation for these reference alignments:

Korber et al. Cell. 2020 Aug 20;182(4):812-827.e19. doi: 10.1016/j.cell.2020.06.043.

NOTES on Delta Variants-of-Variants

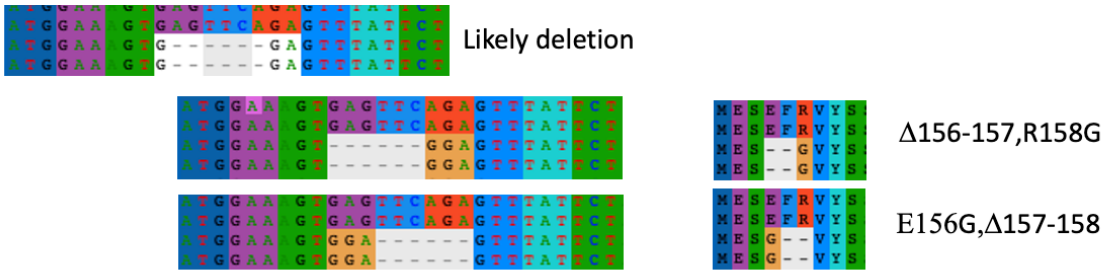
AY.* variants are distinct Pango lineage designations of Delta within the larger B.1.617.2 lineage. (See: PANGO network, new AY lineages). The interesting distinctive variant forms of Spike in Delta are often spread across many AY.* lineages. Our spreadsheet provides all of the Pango lineages that are associated with each Spike variant in the spreadsheet, including all AY.* designations, but we find it more useful for our primary purpose of tracking Spike variation to consider the actual forms of Spikes and their precise geographic locations rather than AY.* designations, and to consider Delta as a full set of all (B.1.617.2 + AY.*) Pango lineages.

Of note, reported ancestral G142 calls within Delta spikes (instead of the G142D mutation that is characteristic of the lineage) most often result from an artifact related to the use of ARTIC 3 primers, which is corrected by the use of ARTIC 4 primers (Davis *et al.* bioRxiv https://doi.org/10.1101/2021.09.27.461949). As a consequence of the artifact, most distinct Delta variants in GISAID include both a G142 and a G142D form; we use the G142D form to represent the variants and exclude this toggling position from our counts of unique forms of Spike.

Deletions in Delta:

T19R,T95I,G142D,E156-,F157-,R158G,L452R,T478K,D614G,P681R,D950N

Variant B.1.617.2's deletion:



This 6-base deletion spans 3 amino acids, almost universally creating a G codon; this pattern is present in almost all Delta (B.1.627.2 and AY.*) Spike sequences. The above graphic shows two relevant codons for two variants without the deletion, and two with. It could be written: "156-158 EFR to G", or "E156-,F157-,R158G" or "E156G,F157-,R158-". At the nucleotide level, the translation "E156-,F157-,R158G" is slightly preferred.

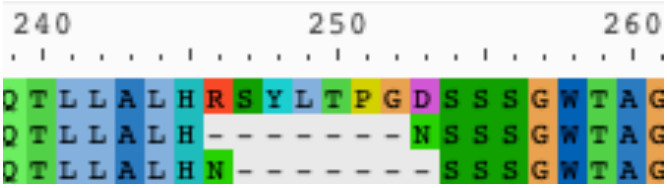
Deletions in Lambda:

G75V,T76I,R246N,S247-,Y248-,L249-,T250-,P251-,G252-,D253-,L452Q,F490S,D614G,T859N

Variant C.37's deletion

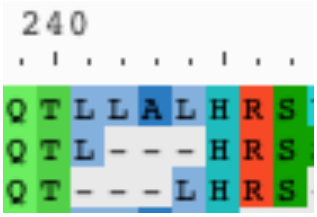
As in Delta, the deletion in Lambda also generates a proximal amino acid change that can be placed either at the beginning or end of the deletion. It could be written: "246-252 SYLTPGD to N", or "R246N,S247-,Y248-,L249-,T250-,P251-,G252-,D253-", or "R246-,S247-,Y248-,L249-,T250-,P251-,G252-,D253N".

Fig:



Deletions in Beta: B.1.351

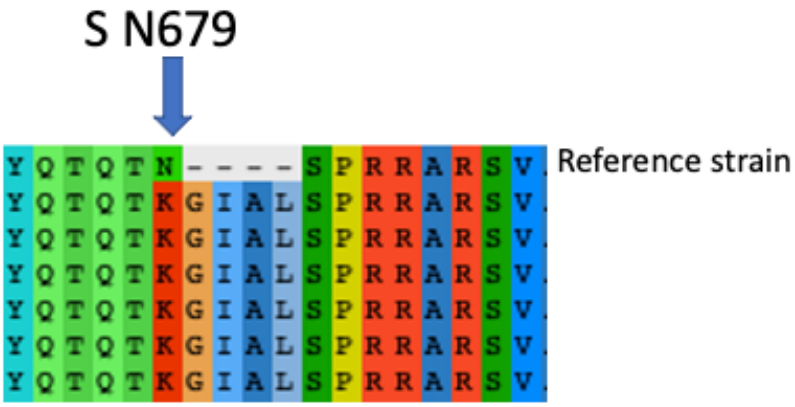
D80A,D215G,L242-,A243-,L244-,K417N,E484K,N501Y,D614G,A701V



Because of the L241 and L244 in the reference strain, this deletion can be expressed equivalently in two ways: " L241-,L242-,A243-" or "L242-,A243-,L244-".

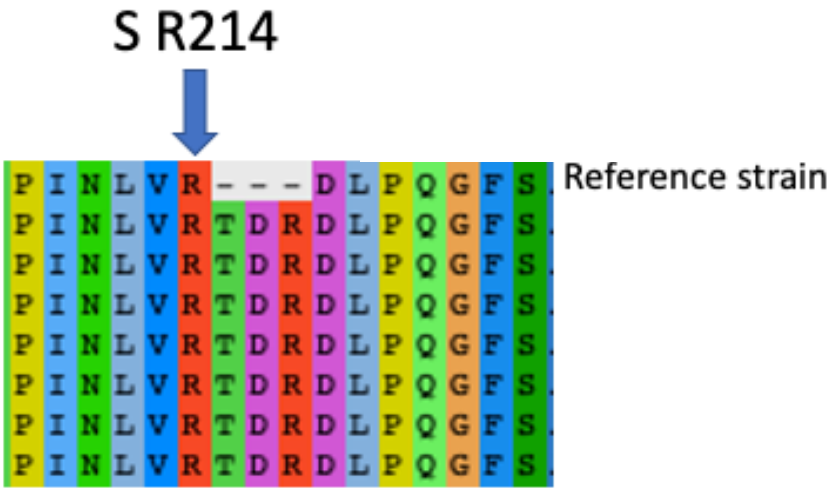
Insertion in AT.1:

P9L,C136-,N137-,D138-,P139-,F140-,L141-,G142-,V143-,Y144-,D215G,H245P,E484K,D614G,N679K,+679GIAL,E780K



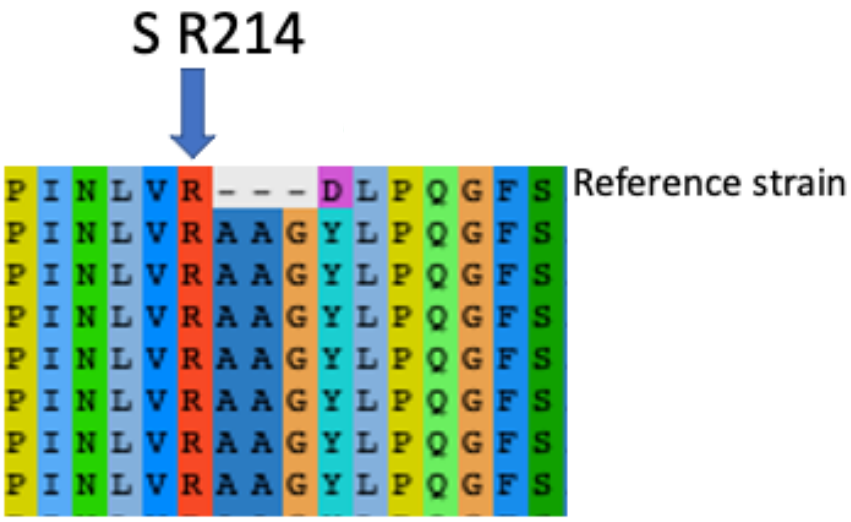
Insertion in B.1.214.2:

+214TDR,Q414K,N450K,D614G,T716I



Insertion in A.2.5.2:

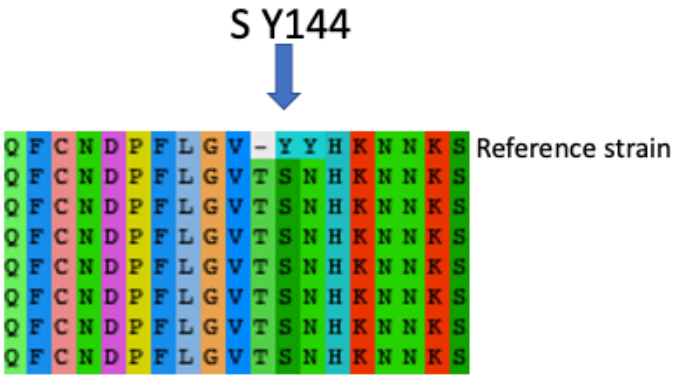
L141-,G142-,V143-,D215A,+215AGY,L452R,D614G,T739I,T859I



Insertion in B.1.621

T95I,+143T,Y144S,Y145N,R346K,E484K,N501Y,D614G,P681H,D950N

This is a very complex lineage in the region near 140-145.
The most common form has an insertion.



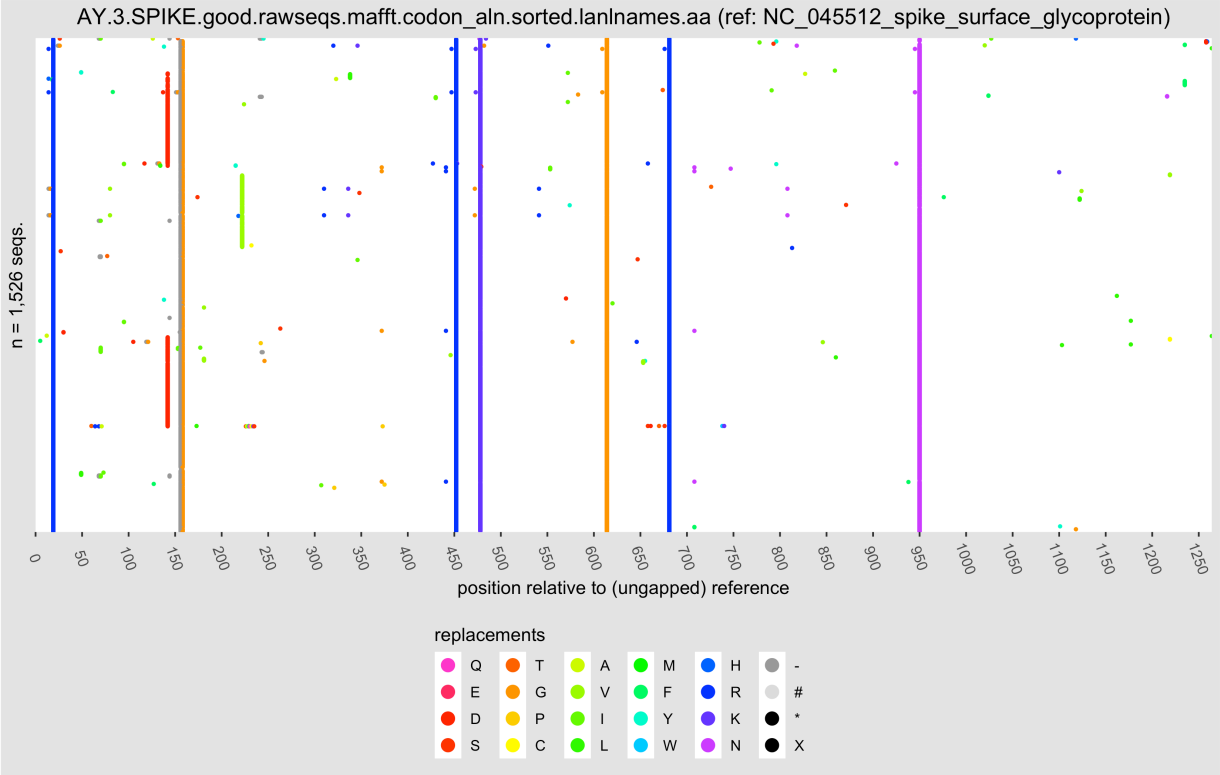
Consequences of sequence uncertainty in newly emerging lineages

As new lineages emerge, there are often many sequences with strings of Ns, or ambiguous base calls. The cov.lanl.gov pipeline filters out such sequences, and as a consequence we undercount the number of sequences in GISAID that have been specified as members of related Pango lineages. Our strategy, however, enables us to use strictly high quality sequences to identify and track the common Spike mutations in an emerging lineage. If this is not done carefully and missing data treated appropriately, mutations that are highly conserved and typify a newly emerging lineage can be missed.

Below we show an alignment of 1,526 sequences from AY.3, sampled very soon after the AY.3 designation was made by the Pango group. The mutations each sequence carries carry relative to the ancestral Wuhan form are highlighted. Data shown here is from unfiltered data from an example from GISAID, July 16,2021. In the top figure, ambiguous amino acids due to missing data are not shown.

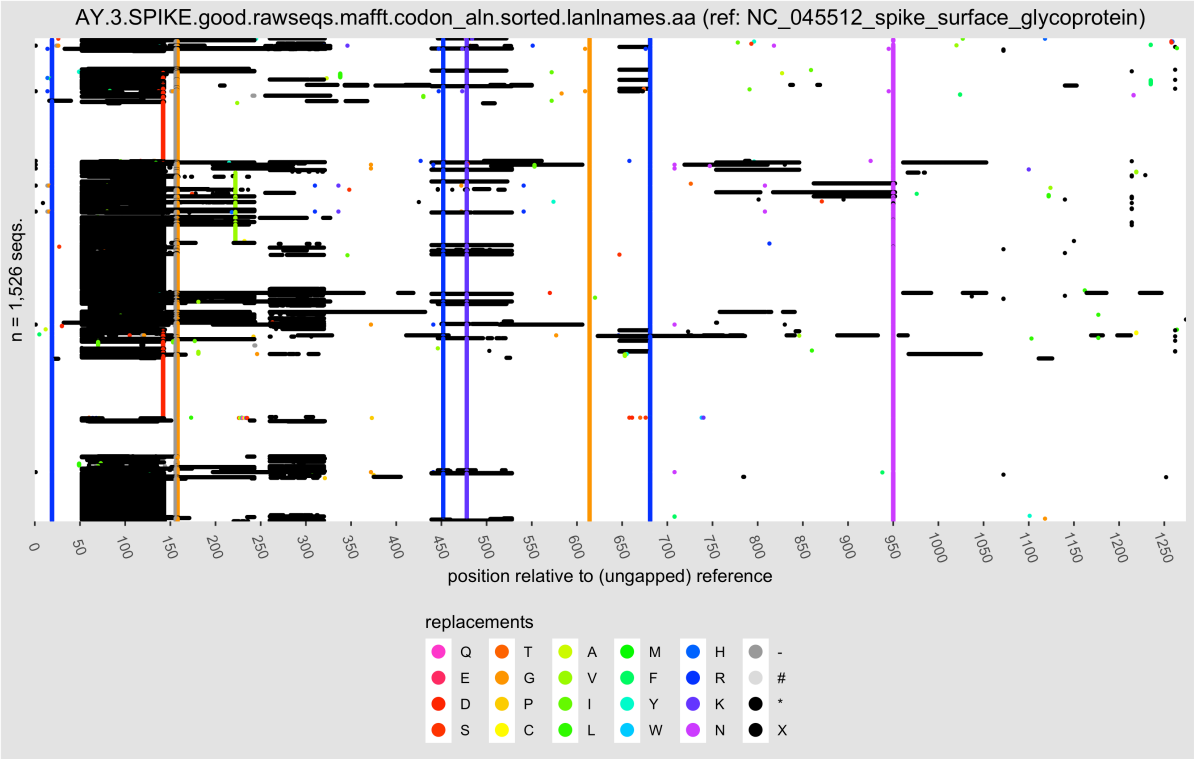
Example: AY.3

Mutations: T19R,G142D,E156-,F157-,R158G,L452R,T478K,D614G,P681R,D950N



Ambiguous base calls (Ns) given rise to uncertain amino acids (X for unknown), are shown in black and added in, in the following figure. Treating the “X” (unknown) characters as ancestral results in an

AY.3 consensus Spike that lacks the highly characteristic Delta pattern in Spike “E156-,F157-,R158G”, although all AY.3 complete sequences preserved this pattern. We treat ambiguous base calls correctly at cov.lanl.gov, but some groups treat N as an ancestral base, which can result in missing important mutations in emerging lineages in consensus sequence reconstructions.



Complex variants within Pango lineages

Example: B.1.526 Iota (example taken from Aug. 9, 2021).

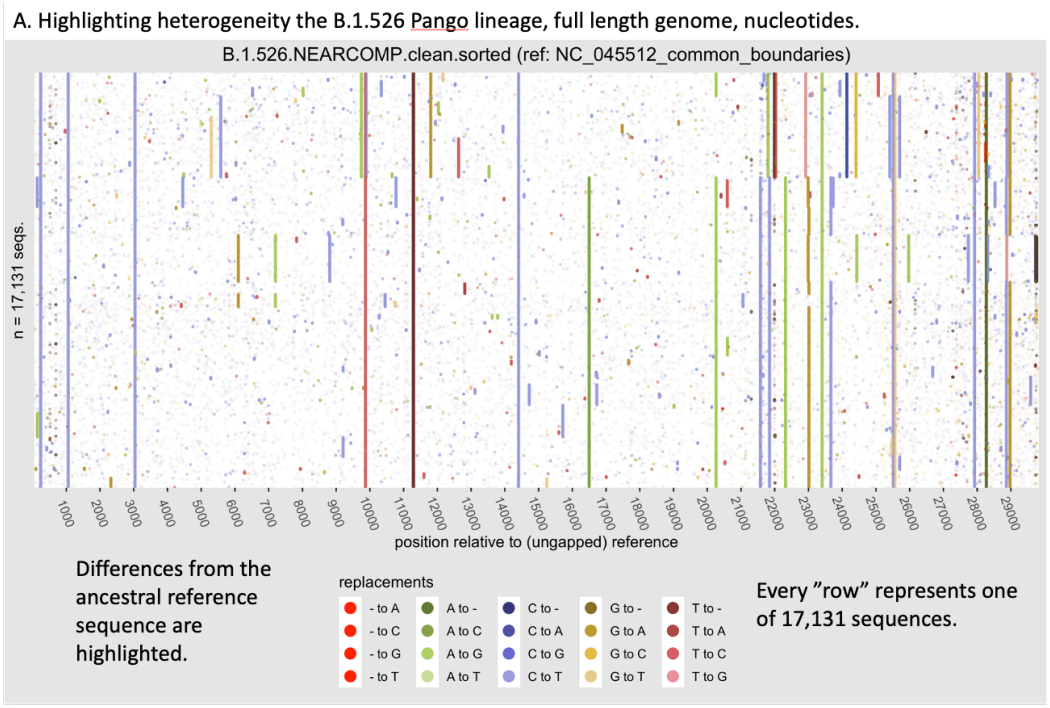
Note: the problem B.1.526 described below persisted for some months, and is was good illustration of the kind of complexity that can arise in a Pango lineages, and why Pango lineage consensus sequences may sometimes be not be representative of the lineage. We retain it here from illustrative purposes, although this particular issue has been better resolved by the Pango group, and the very distinctive variant in B.1.526 has been renamed B.1.637.

A complex lineage that illustrates how a Pango lineage *consensus* can give a misleading impression about the forms of Spike within a designated lineage.

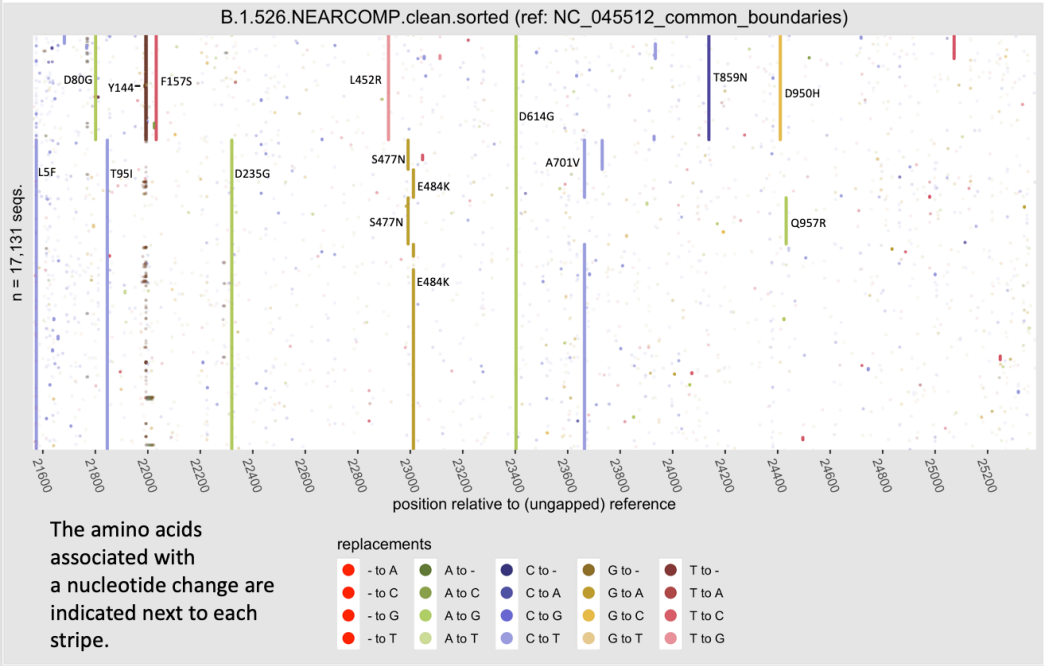
In the original Pango designation B.1.526 there had been 3 very distinctive lineages. Only one amino acid in Spike was shared among all three, G614D. A consensus of B.1.526 from this data taken during this period would suggest that B.1.526 has no mutation in the Spike RBD, as none reached a 50% level, yet every single Spike sequence within this lineage carries one of three RBD mutations, either E484K, S477N, or L452R. The Pango group transitioned from calling this set "B.1.526" to "B.1526.1, B.1526.2 or B.1526.3", and then rescinded the three B.1.526 sublineage names, reverting back to B.1.526. Eventually they renamed the most distinctive variant lineage B.1.637. As we considered different forms of Spike for tracking, we found all three forms common enough to merit following, and we did. A consensus of B.1.526 would not have matched the RBD any of the three variants. The third from is very distinct from the first two throughout the full genome (data not shown).

B.1.526 L5F,T95I,D253G,E484K,D614G,A701V
B.1.526 L5F,T95I,D253G,**S477N**,D614G,Q957R
B.1.526 **D80G,Y144-,F157S,L452R**,D614G,**T859N,D950H** **-> now renamed B.1.637**

The consensus B.1.526 Spike, "L5F,T95I,D253G,D614G" doesn't match any form of Spike in this lineage.



B. Highlighting heterogeneity the B.1.526 Pango lineage, just Spike, nucleotides.



Example: B.1.234 [G142S,E180V,D614G,Q677H]

Occasionally a sublineage surfaces within a Pango linage that is not the most common form of the lineage but still of potential interest for continued monitoring. One could request a new Pango lineage number be considered specifically for the sublineage, eventually, if they remain of interest.

The consensus form of B.1.234 just carries D614G in Spike. There were sub-lineages that were transiently increasing in Texas and California that are highlighted in blue, with the pattern: G142S,E180V,D614G,Q677H.

This is based on Aug. 9, 2021 data:

Pango Lineage	Lineage Count	Form Count	%	Mutation string
B.1.234	5715	3416	59.8%	[D614G] (consensus)
		362	6.3%	[D614G,P812L]
		310	5.4%	[G142S,E180V,D614G,Q677H]
		150	2.6%	[D614G,N679K]
		132	2.3%	[G142S,E180V,D614G,Q677H,S940F]

Example: B.1.1.284 [M153T,G184S,D614G,Q677H]

The most common form of B.1.1.284 just carries D614G in Spike. The baseline form of the sub-lineage that was more highly mutated and transiently increasing in Japan is highlighted in blue.

Pango Lineage	Lineage Count	Form Count	%	Mutation string
B.1.1.284	8841	3306	37.4%	[M153T,D614G] (consensus)
		3491	39.5%	[D614G]
		480	5.4%	[M153T,G184S,D614G,Q677H]
		183	2.1%	[M153T,D614G,P793S]
		109	1.2%	[M153T,S255Y,D614G]