**Variants of the Spike protein that had a significant presence during 2021  
Oct. 3, 2021**

**LA-UR-21-28226**

**Summary**

The accompanying spreadsheet provides a listing of natural forms of Spike that had a significant presence in the world on at least 1 continent, starting in the spring of 2021, together with a few additional very diverse Spikes that are included as examples of more extreme mutational patterns.

**Details**

The listing includes variants that were identified to be of interest using our [SHIVER](https://cov.lanl.gov/components/sequence/COV/shiver.comp) and [XSpike](https://cov.lanl.gov/content/sequence/XSPIKE/xspike.html) codes, and confirmed to be relatively common and increasing in frequency in least in one local geographic region, even if only transiently, during 2021, using additional cov.lanl.gov tools.

All of the current [WHO variants of interest and concern](https://www.who.int/en/activities/tracking-SARS-CoV-2-variants/) are captured in this list, as well as many additional variants. We begin by tracking related natural forms of Spike, find the most common circulating version, and then subsequently determine what Pango lineage the form of Spike is associated with, rather than tracking Pango lineages. These natural forms of Spike often, but will not always, correspond to the consensus of a Pango Lineage. When a lineage is complex and includes sets of unrelated forms of Spikes the consensus and the variant of interest can differ.

THESE LISTS OF MUTATIONS REPRESENT NATURAL FORMS OF SPIKES THAT ARE MOST COMMONLY FOUND AMONG A COMMONLY EMERGING VARIANT.

Also, the Spikes listed in this file are curated and cross-checked against relevant sequence alignments. They are carefully monitored to resolve regions that are difficult to automatically align near deletions, to 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](https://cov.lanl.gov/content/sequence/XSPIKE/xspike.html)”, “[SHIVER](https://cov.lanl.gov/components/sequence/COV/shiver.comp)”, and “[COMMON FORMS OF SPIKE WITH A GIVEN PANGO LINEAGE DESIGNATION](https://cov.lanl.gov/components/sequence/COV/pangocommonforms.comp)” updates are automatically run several times a week, and the automatic runs of these codes can capture newly emergent variants as they are appearing in GISAID in closer to real time.

Please note: We build a Fasta file including a full length representative GISAID sequence for the common variant forms included in Table 1. To do this, all GISAID sequences that match the common Spike form are retrieved and the consensus of the rest of the genome among that set is determined. The oldest sampled sequence that completely matches the consensus is selected as the representative sequence of that variant; or if no sequence complete matches the consensus, the one that is closest to it. We hope to make this reference genome file available through GISAID soon. In the interim, please inquire if you need help with this.

**Table of Contents**

* **Variants Table: mutations found in distinctive lineages.**
* **Deletions in key variants.**
* **Insertions in key variants.**
* **Consequences of sequence uncertainty in newly emerging lineages**. Examples of how uncertain base calls in sequences can impact interpretation, counts, and consensus forms of emerging variants. Delta-related Pango lineages B.1.627.2, AY.1, AY.2, AY.3 are used as examples.
* **Complex variants within a Pango lineage.** Examples of the consequences of genetic complexities within Pango lineages that have given rise to situations where the complex forms of Spike that were associated within a given Pango lineage were not represented by a consensus of the lineage.

**Key to Variants Table.** The most common complex variant forms of Spike that were frequently sampled at some point in GISAID ([www.gisaid.org](https://www.gisaid.org/)) in 2021, identified using the LANL tools XSpike and SHIVER. Table count updates: Oct. 3, 2021.

**Column headers with notes, Variants Table:**

1. **Classification of Variant**
2. **Date of addition to cov.lanl.gov website listing**
3. **WHO designation:** The WHO Greek letter variant designation is as of August 9, 2021 ([WHO Tracking SARS-CoV-2 variants](https://www.who.int/en/activities/tracking-SARS-CoV-2-variants/)); \*indicates continued monitoring.
4. **Pango lineage** most often associated with a particular Spike, as reported in GISAID as of August 9, 2021. Pango lineage descriptions can be found at [Pango Lineages](https://cov-lineages.org/).

* Bold Pango letters: Variants we have identified through SHIVER, XSpike or Ember runs beginning March of 2021,
* Non-bold letters: Variant identified using XSpike or Embers, or tracked because they were listed as a WHO variant of interest, and considered of interest because they were complex or carried known resistance mutations or mutations in the furin cleavage site.

1. **Most common Spike backbones.** The a most common natural form of a Spike variant lineage.

- indicates a deletion at a site (e.g. Y144-, the Y at position 144 in the reference Spike is deleted), “+” and insert following the specified site (e.g. +143T indicates a T was inserted after position 143 in the reference Spike).

* Blue: Addition of positive charge near the furin cleavage site: 675, 677, 681 are positively charged
* 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.

1. **Number of sequences that exactly match this pattern in Spike, full data.** Note that sequences with ambiguity amino acid calls in them are not included in this tally, and this tally includes all data in our quality filtered GISAID data set starting in December 2019.
2. **Number of sequences that contain this pattern, full data.** The most common form of Spike representing a lineage are only a part of an evolving lineage. This tally counts the number of variants that contain the full specified sequence backbone, but also contain additional mutations. As lineages spread and become older, they diversify, the most common form becomes a small and smaller percentage of the total. As new variants of variants become more prevalent, we begin to identify them as a distinctive common form, and we add the most interesting of these them to the variants of variants listing.
3. **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 more contemporary GISAID global sampling, or declining and being replaced but other variants.
4. **Number of sequences that contain this pattern, last 60 days. See above.**
5. **All Pango Lineages that contain sequences that exactly match this pattern in Spike (and count).** Quite distinctive Spikes are often assigned to an array of Pango lineages, some closely related, but some not obviously related. This might arise due to recombination, or mis-classification; these may change over time as Pango lineage designations can be reassigned. These calls are based on Pango lineage assignment in our August 9, 2021 GISAID data.
6. **All Pango Lineages that contain sequences that contain match this pattern in Spike (and count).**
7. **Total count of sequences with the Pango lineage designation (based on our QC filtered set used at cov.lanl.gov).**
8. **Number of Spikes in the Pango lineage that exactly match this pattern.**
9. **Fraction of the Pango lineage that exactly match this pattern, L/K.**
10. **Number in the Pango lineage that contain this pattern.**
11. **Fraction of the Pango lineage that contain this pattern, N/K.**
12. **The three countries where the exact Spike variant is most commonly sampled.**
13. **The three countries where Spikes that include the variant mutations are most commonly sampled.**
14. **Notes regarding insertions and deletions.** Please see the following pages for more detail
15. **Notes.**

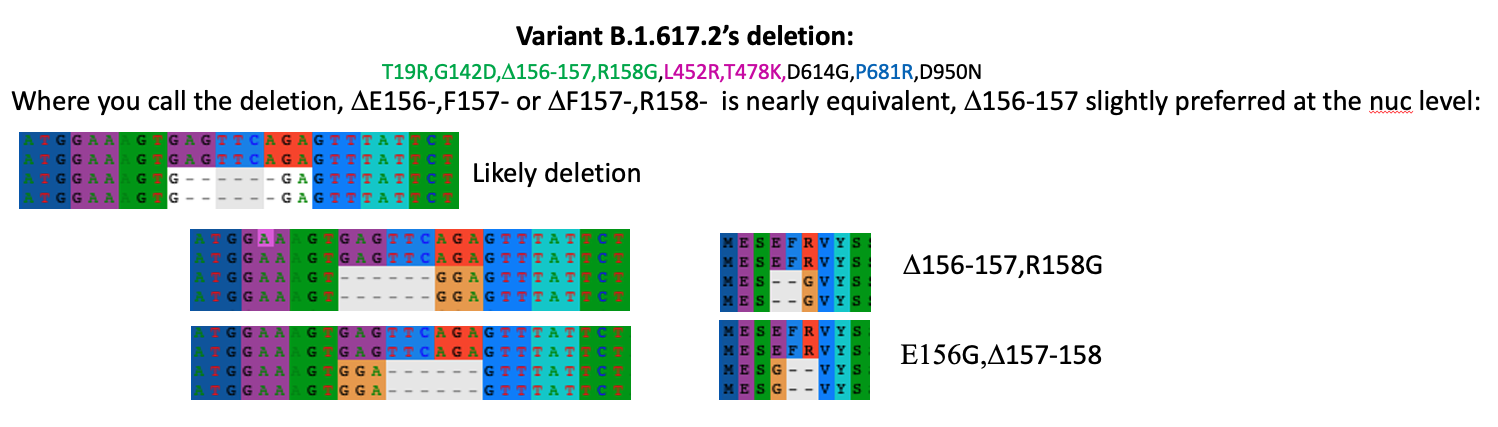
**Legend, Variant Tables.**

The variants are identified based the sampling frequency of Spike variants in [GISAID](https://www.gisaid.org/). The [Pango lineage](https://cov-lineages.org/)  most often associated with the variant form is noted; if it is in Bold, the variant was identified using SHIVER, and so it was a common form circulating in some region of the world, at least transiently, in 2021. The others included here were identified using XSpike and highlighted because they have particularly complex mutational patterns, or were listed in the WHO as an important variants, and are featured despite being rare because they are interest. If the [WHO](https://www.who.int/en/activities/tracking-SARS-CoV-2-variants/) has designated a lineage name (Greek letter) it is noted, and if they have designated it for further monitoring, it is noted by and asterisk (\*). The Counts indicate how many times the *exact* Spike form noted has been found in our cov.lanl.gov QC filtered GISAID set. We filter out about 20% of the sequences in GISAID from out baseline alignments, generally due to being incomplete.

**NOTE 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](https://www.pango.network/new-ay-lineages/)). We do not find the AY designations track well with our Delta phylogenies, but they are often mixed into different clades in our trees. Also, the interesting distinctive variant forms of Spike in Delta are often spread across many AY.\* lineages. The AY.\* linages are noted by the Pango group to show geographical clustering, but GSIAD provides the precise geographical information related to each sample. We provide all of the Pango lineages that are associated with each Spike variant in the spread sheet, including all AY.\* designations, but we find it more useful for our primary purpose of tracking Spike variation to consider the actual forms of Spike and the countries and more precise geographic locations in which the forms are sampled, and to consider Delta as a full set of all (B.1.617.2 + AY.\*) Pango lineages.

**Deletions in Key Variants**

**Deletions in Delta:**

****

This 6-base deletion spans 3 amino acids, and essentially always creates a G, and this pattern is present in almost all B.1.627.2, AY.1, AY.2, and AY.3 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 “E156G,F157-,R158-” is slightly preferred. Almost all variants of B.1.617.2, AY.1, AY.2, and AY.3 carry the 2-amino acid deletion with the G. Two currently rare forms of B.1.627.2 have been sequenced that fill in the deletions; “156-158 EFR to G”, has become either EFG, or more often ECG, in an otherwise B.1.627.2 backbone.

The most common natural forms of Spike we have found in the AY.1, .2 and .3 lineages are:

AY.1 Delta **d**  T19R,T95I,G142D,D156-157,R158G,**W258L**,**K417N**,L452R,T478K,D614G,P681R,D950N

**AY.2** Delta **d**  T19R,**V70F**,G142D,D156-157,R158G,**A222V,K417N**,L452R,T478K,D614G,P681R,D950N

AY.3\* Delta **d** T19R,G142D,D156-157,R158G,L452R,T478K,D614G,P681R,D950N

**Deletions in Lambda:**

**Variant C.37’s deletion**

G75V,T76I, D246-252,D253N,L452Q,F490S,D614G,T859N or expressed as

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

Similar to the situation in Delta, a large 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 D246-252,D253N or “E156G,F157-,R158-”. The Pango listing for Lambda was also corrected on Aug. 1, 2021.

Fig:



**Deletions in Beta: B.1.351**

D80A,D215G,D242-244,K417N,E484K,N501Y,D614G,A701V

This deletion can be expressed equivalently in two ways: D241-243 == D242-244

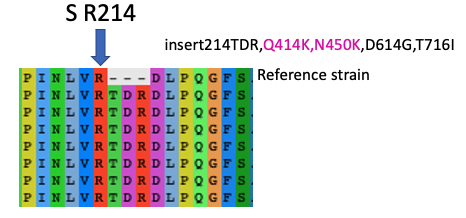


**Insertions in Key Variants**

**Insertion in AT.1:**



**Insertion in B.1.214.2:**



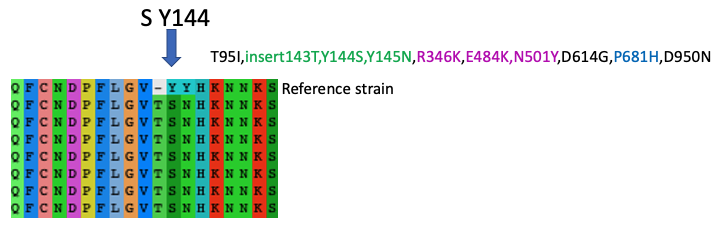
**Insertion in A.2.5.2:**



**Insertion in B.1.621**

This is a very complex lineage in the region near 140-145.

The most common form has an insertion.



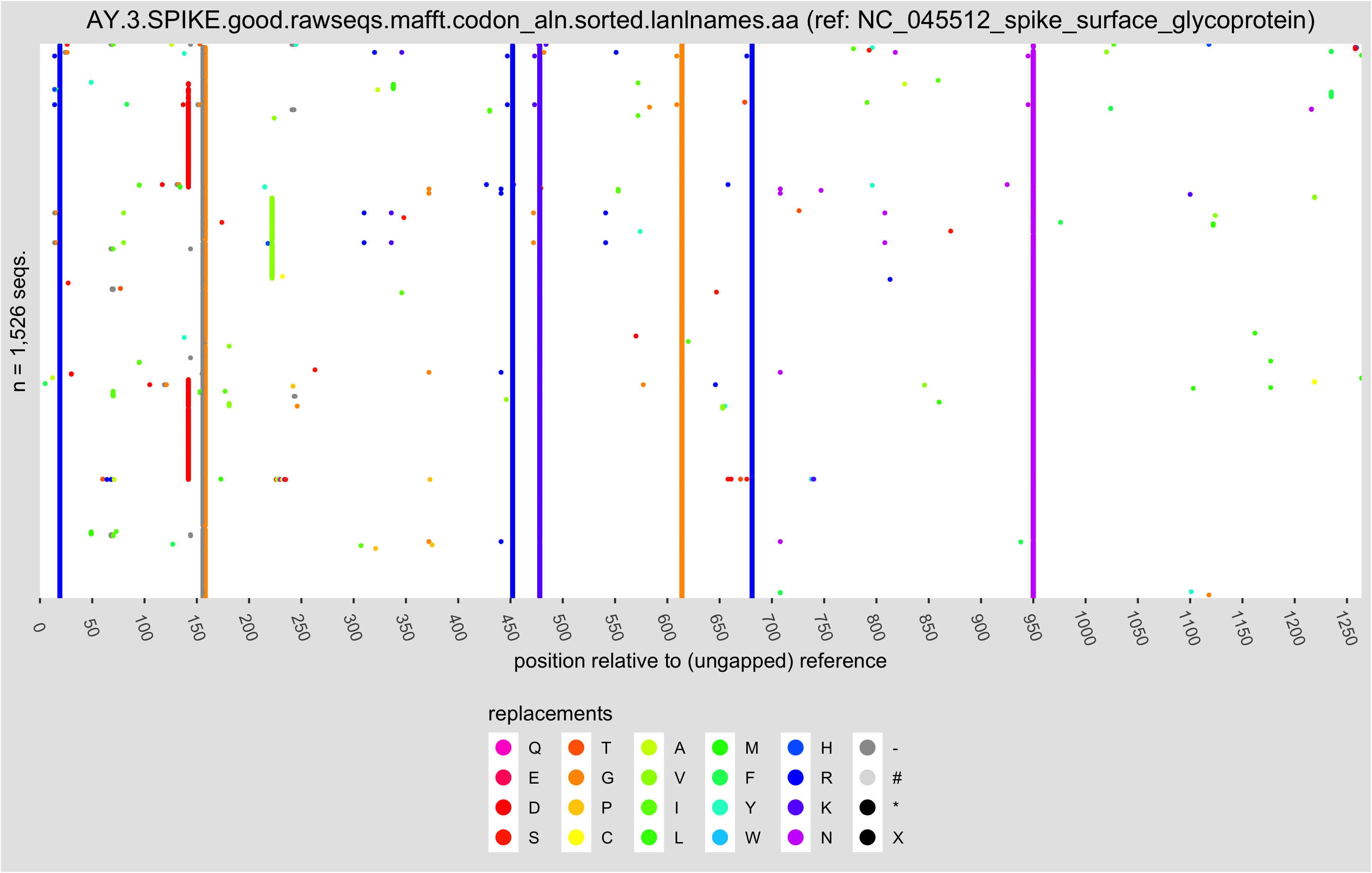
**Consequences of sequence uncertainty in newly emerging lineages**

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 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. Below we show an alignment of 1,526 sequences from AY.3, sampled 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. On the next page are shown similar plots for a subsample of B.1.617.2, AY.1 and AY.2 and AY.3. Data shown here is from unfiltered data from an example from GISAID, July 16,2021.

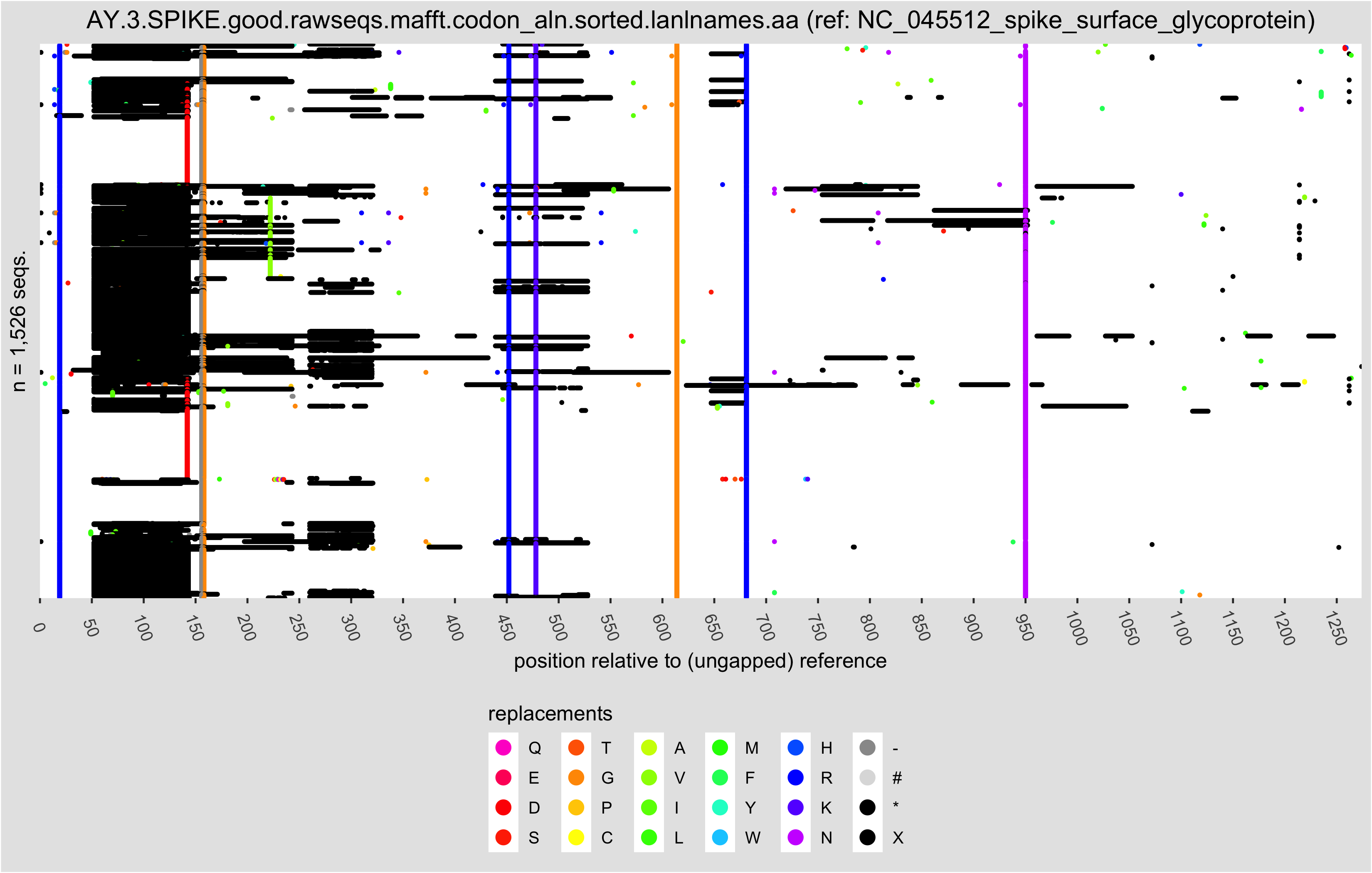
**Example: AY.3**

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

Omitting codons with ambiguous base calls:



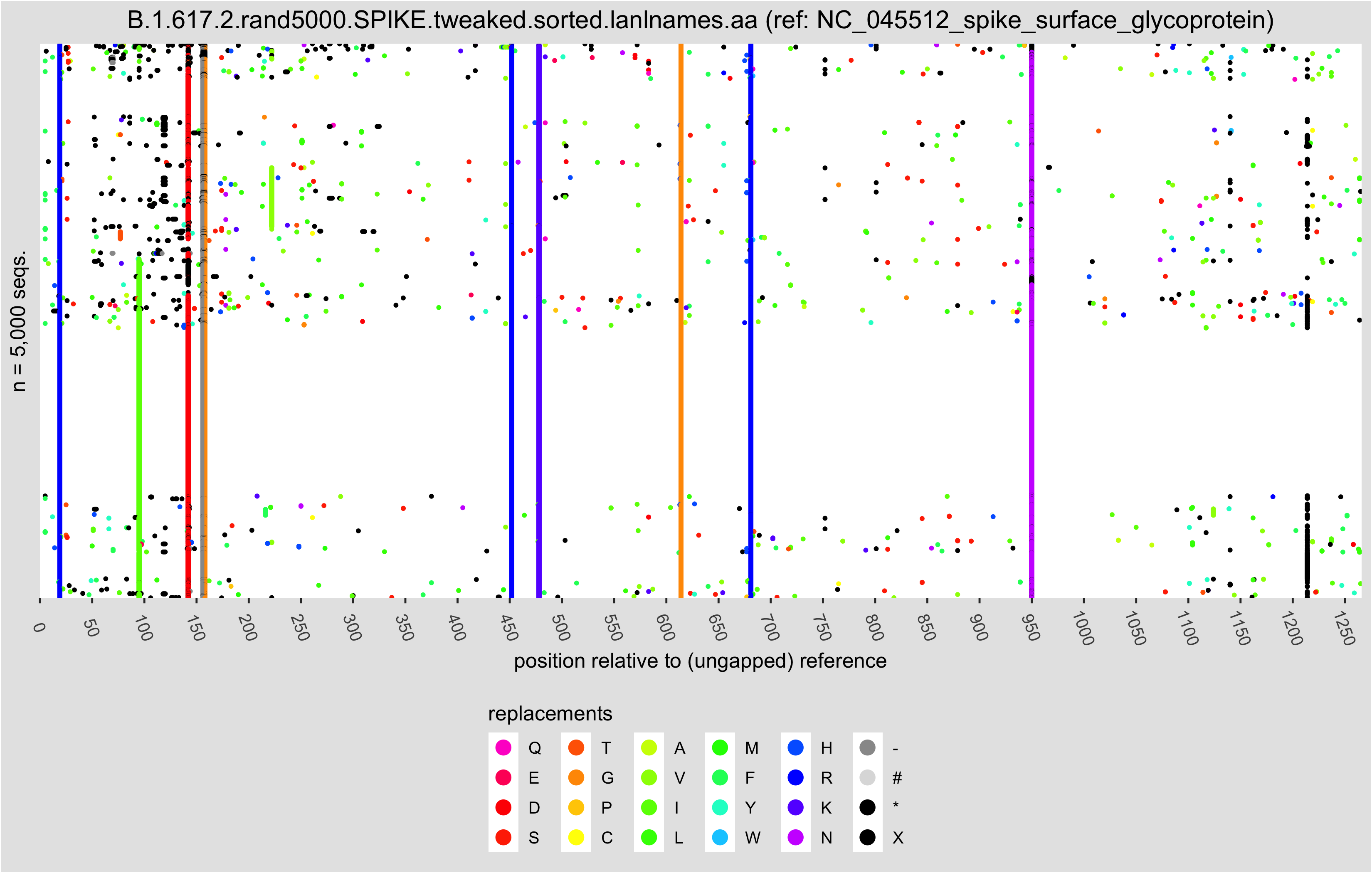
Ambiguous base calls given rise to uncertain amino acids (X for unknown) in black:



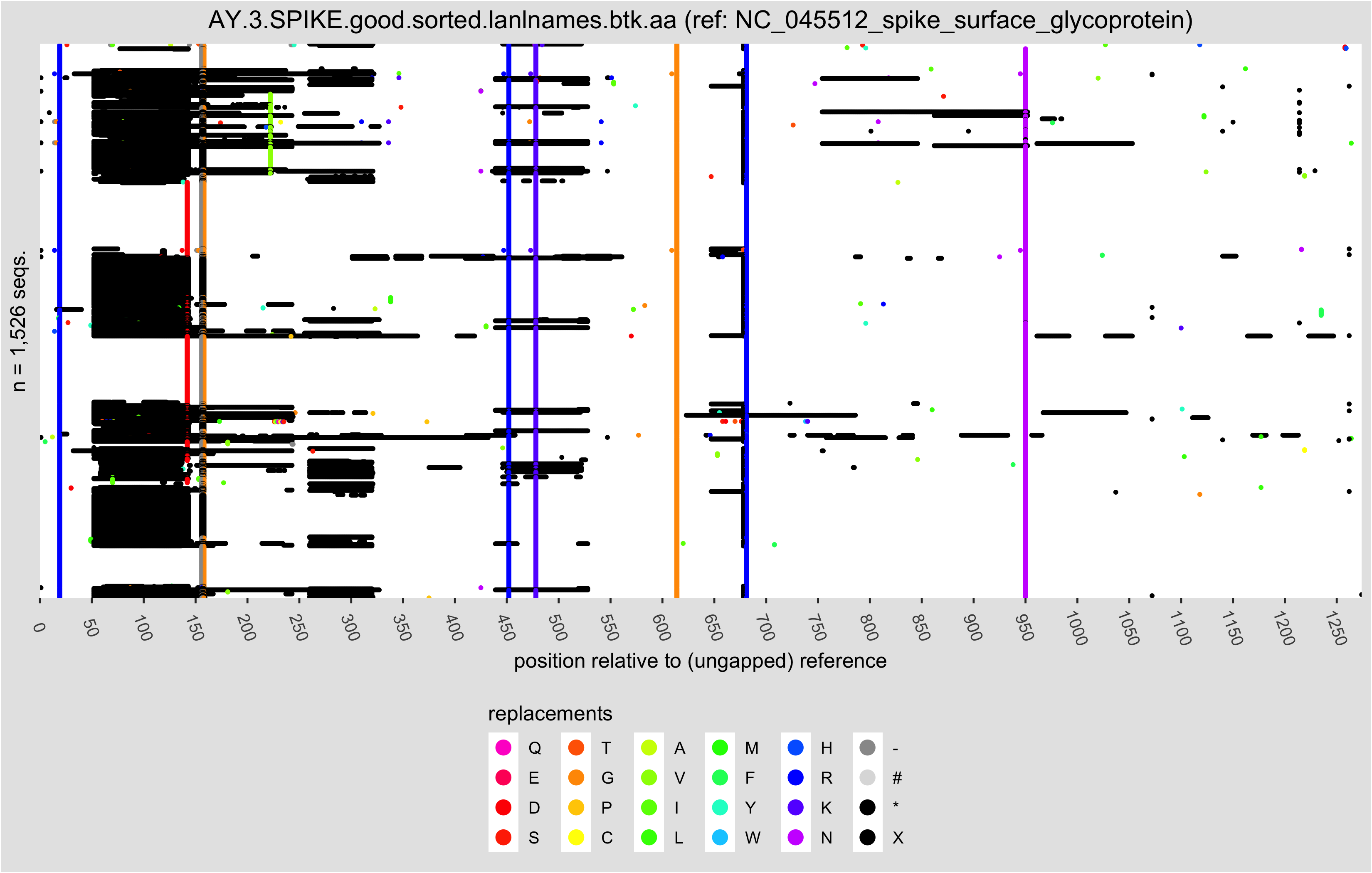
**Example: B.1.617.2**

Sequences for this lineage are currently generally well resolved. There are still many uncertainties in AY.1, AY.2, and AY.3

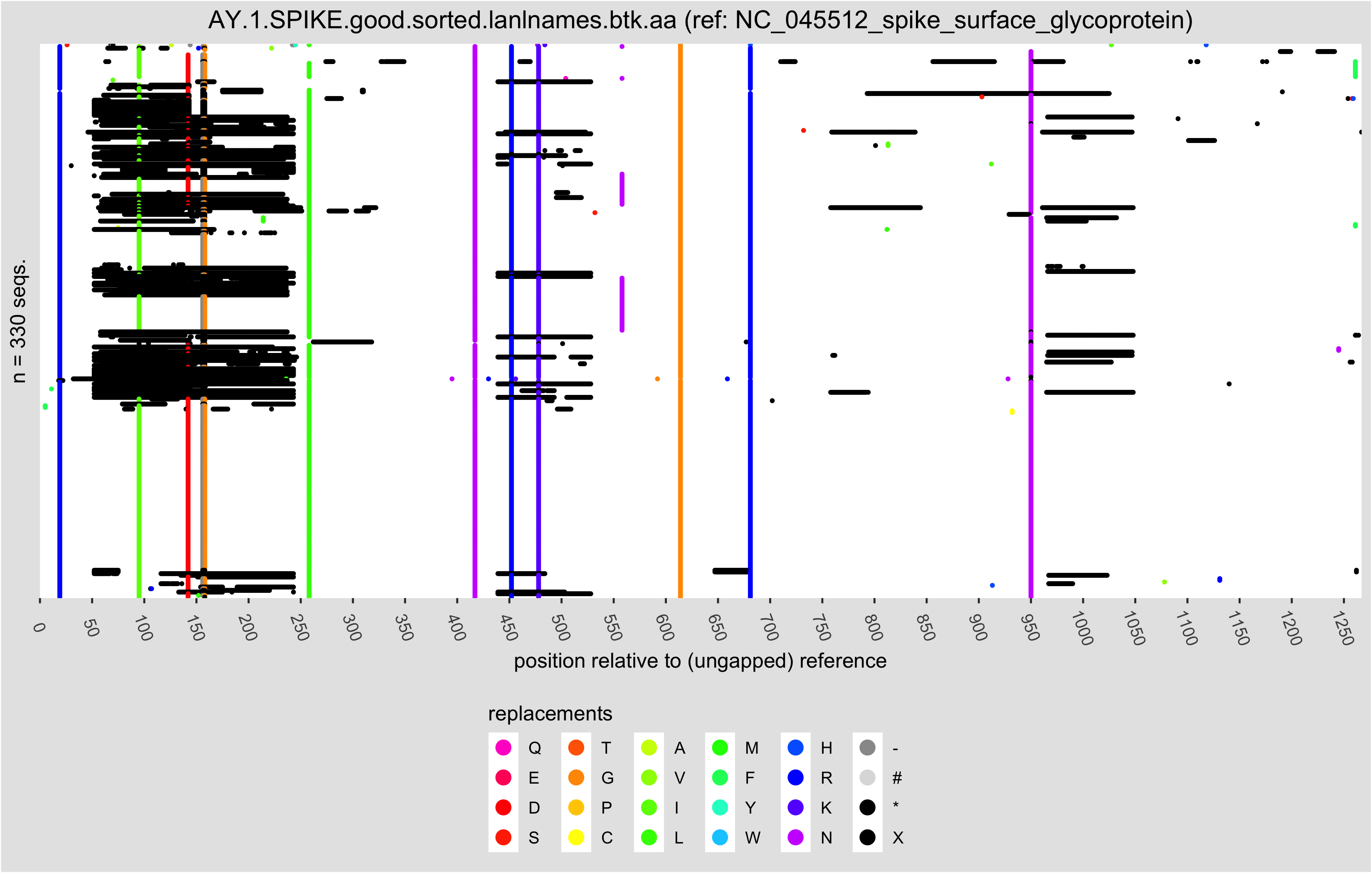
B.1.617.2: subsample of 5000 sequences



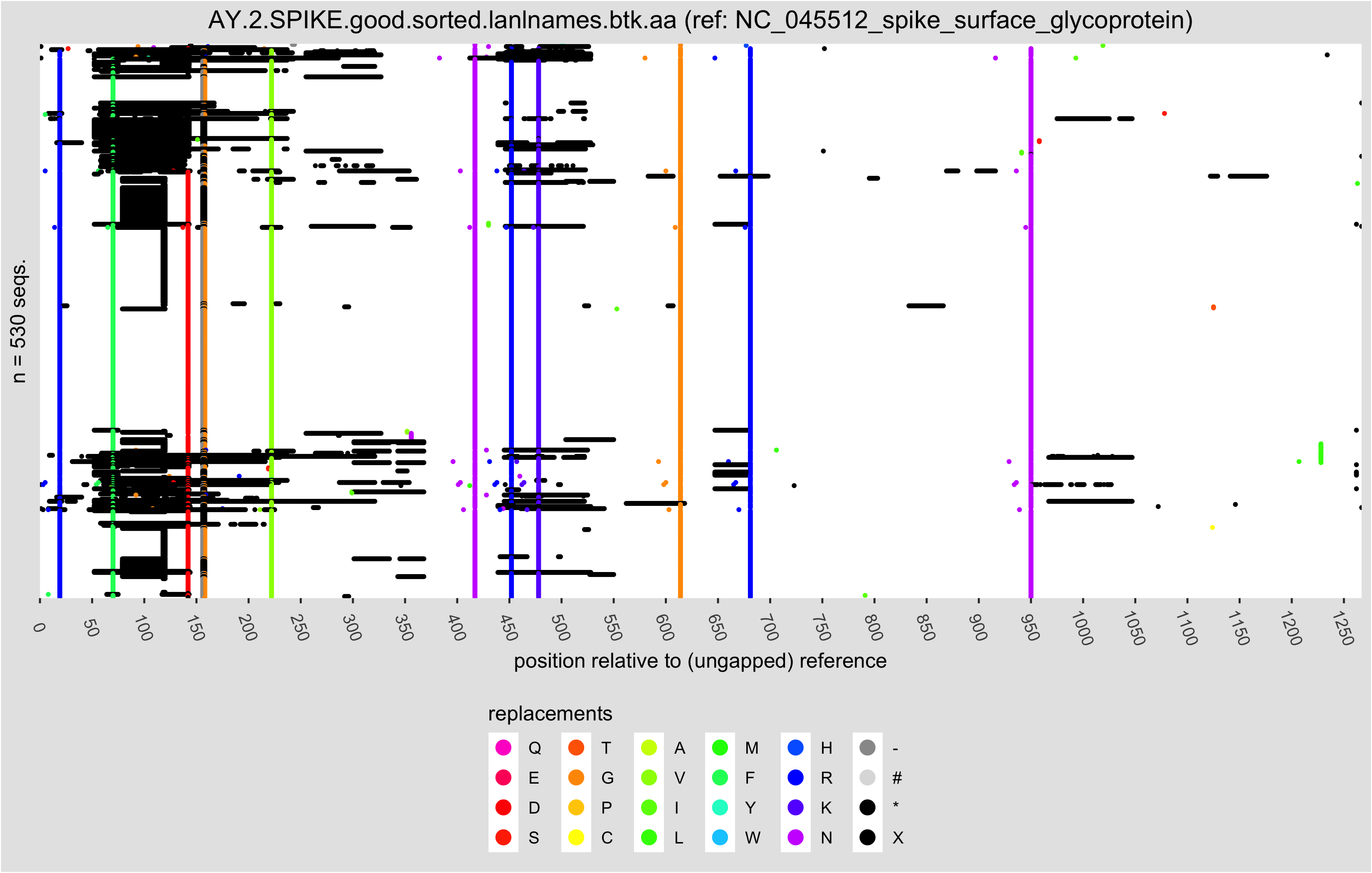
AY.3: Sample of 1,526 sequences



AY.1: Sample of 330 sequences



AY.2: Sample of 530 sequences



**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. Thus we retain it here from illustrative purposes. But this particular issue has been better resolved by the Pango group, and the very distinctive variant in B.1.526 now renamed B.1.637.**

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

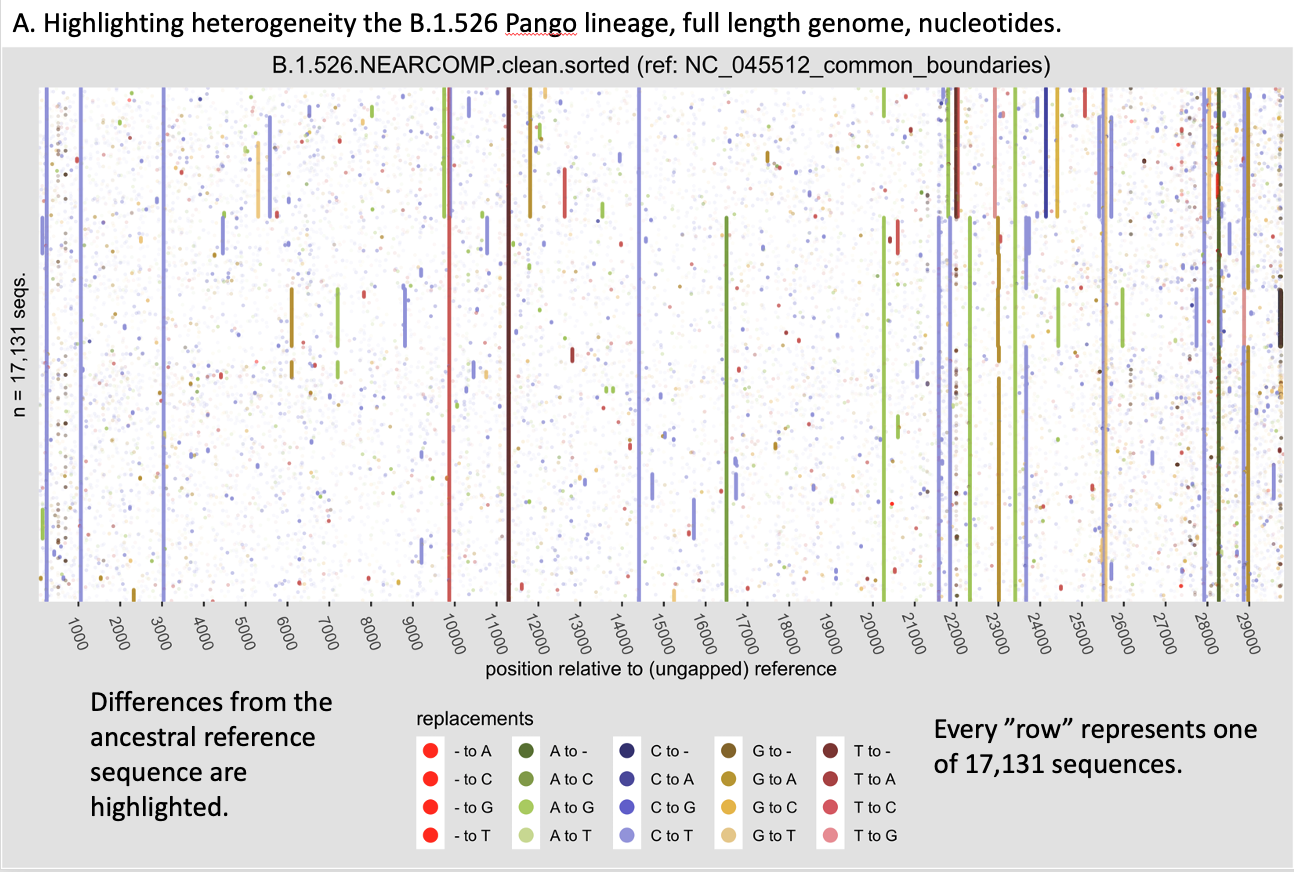
There are as of this writing 3 distinctive lineages within B.1.526. Only one amino acid in Spike is shared among all three, G614D. A consensus of B.1.526 from this data sample would suggest that B.1.526 has no mutation in the Spike RBD that has reached a 50% level, yet every single Spike sequence within this lineage carries one of three RBD mutations, either E484K, S477N, or L452R. The Pango group have 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, and went back to B.1.526. Currently we consider the three different forms of Spike within this lineage as forms of Spike that have been common enough to merit tracking. The third from is very distinct from the first two throughout the genome.

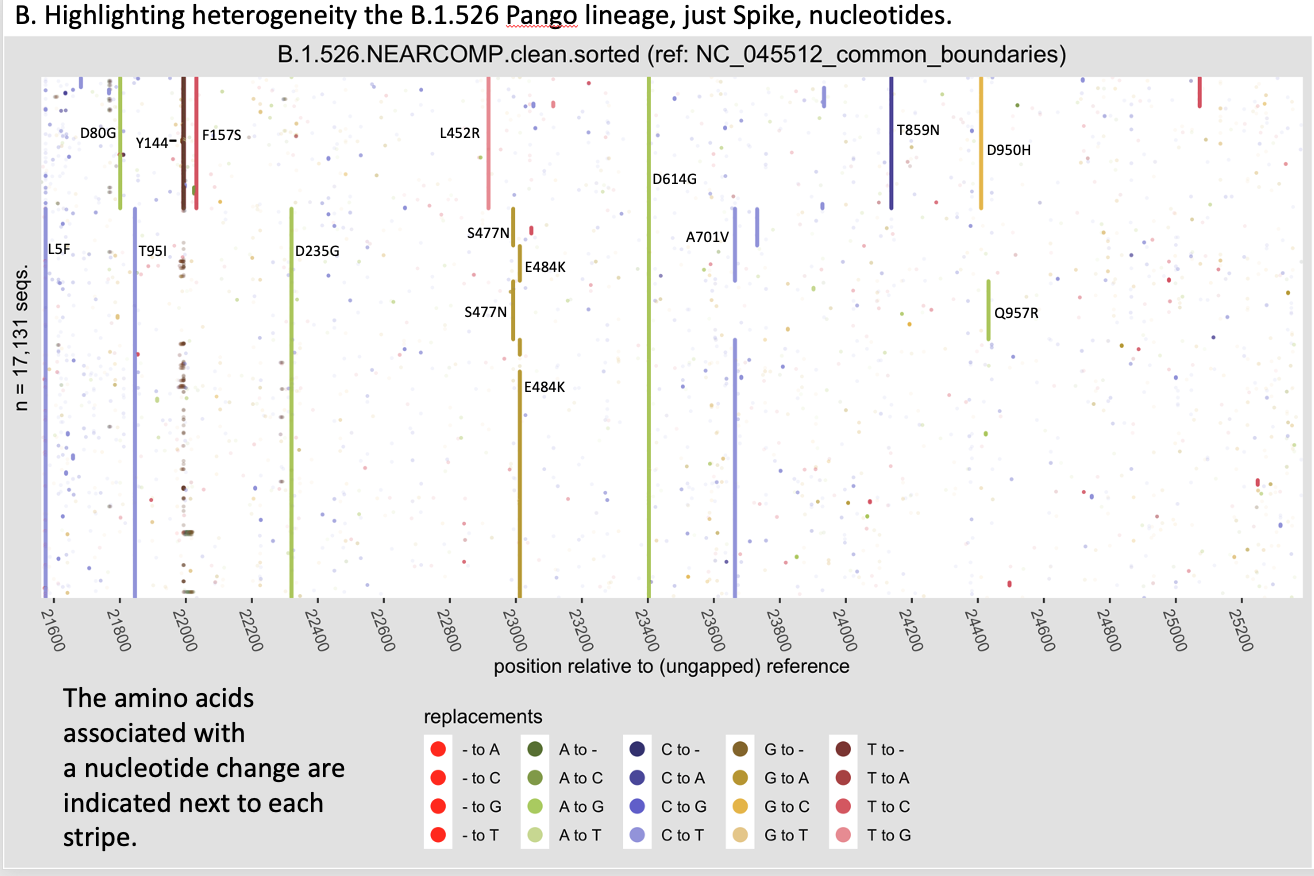
**B.1.526**  L5F,T95I,D253G,E484K,D614G,A701V

**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.





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

Occasionally a sublineage surfaces with a Pango linage that is rare within the lineage but still of potential interest for continued monitoring. One could request a new Pango lineage number specifically for the sublineage, eventually, if they remain of interest.

The consensus form of B.1.234 just carries D614G in Spike. The baseline forms sub-lineages that were transiently increasing in Texas and California are highlighted in blue.

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] |