SARS-CoV-2: Tracking Variants of Concern

color

Version 2.0 – Updated 3.30.21

Summary

The Problem: Variants of concern (VOC) that may increase transmissibility or decrease efficacy of treatments or vaccines are arising in SARS-CoV-2 strains throughout the world. Tracking these VOC is crucial for public health, but the current method of designating viral lineages is inadequate.

B.1.1.7 Classification: The B.1.1.7 viral lineage, first identified in the UK, has been reported to increase transmissibility. Therefore, tracking the spread of this variant worldwide has been a priority. However, current tracking of B.1.1.7 faces the pitfall that lineage definition is not based on important VOC, and strains can be designated as B.1.1.7 without the relevant VOC. Additionally, using S-gene dropout to track the B.1.1.7 lineage both misses strains with the crucial VOC N501Y and identifies many samples that do not contain N501Y.

E484K Tracking: The E484K variant has arisen as a particularly concerning variant, as it has been reported to reduce the efficacy of some vaccines. It was first noted in the B.1.351 lineage in South Africa. Importantly, our analysis has identified that this VOC is arising in several additional lineages around the world and should be watched independent of a specific lineage.

Discussion: While lineage designation can be a useful tool during the initial identification of VOC, current methodologies face pitfalls that may be hampering public health tracking and scientific communication. By not tracking and reporting on VOC independent of lineages, scientific understanding of the effects of these VOC may be impeded.

Background

As the SARS-CoV-2 virus continues to accumulate novel variants, scientists have been identifying and characterizing variants of concern (VOC) or variants of interest (VOI) as they emerge.[1,2] These VOCs are primarily arising in the spike protein of the virus, a key viral component that allows the virus to attach to and invade human cells.[3] Some variants have been reported to make the virus more transmissible, increase severity of COVID-19 (the disease caused by SARS-CoV-2), or reduce natural immunity or vaccine efficacy, but how they do so is still under investigation.[2] One region of the spike protein known as the receptor binding domain (RBD) appears to play a key role.[1]

The nomenclature used to describe the viral lineages that contain these variants may be muddling tracking efforts.[4] Currently these genetic variations are tracked and discussed from an evolutionary perspective as viral lineages. These lineages are defined by Pangolin,[5] a software program that takes the genomic sequence of a particular sample and algorithmically assigns it a lineage designation. Lineages that are currently considered to be of particular interest, and the key VOC and VOI they contain, are listed in Table 1. Key lineages that are circulating in the United States, and the known attributes of these lineages, are being tracked by the US Centers for Disease Control.[6]

In this document, we refer to specific genomic changes as **variants**, and an evolutionarily connected collection of variants as a **lineage**. A **strain** is a lineage that has acquired new properties. Tracking lineages, as is common practice, has the advantage of following groups of mutations that may combine to affect the viral phenotype, in a phenomenon known as epistasis.[7,8] However, the methodology used to define these lineages is crucial, and strict monitoring of these lineages without following the individual VOC can miss signals important for public health monitoring.

Lineage	del69_70	K417N	K417T	N439K	L452R	S477N	T478K	E484K	N501Y	P681H	Count	1st large cluster	Alternate nomenclature
B.1.1.7	+								+	+	127,922	United Kingdom	501Y.V1
B.1.351		+						+	+		3,061	South Africa	501Y.V2
P.1			+					+	+		823	Brazil	501Y.V3
P.2								+			1,206	Brazil	
B.1.526								+			1,532	US/New York	
B.1.526						+					439	US/New York	
B.1.526.1					+						202		
B.1.526.2						+					163		
B.1.427					+						2,817	US/West Coast	CAL.20C
B.1.429					+						6,785	US/West Coast	CAL.20C
B.1.160						+					10,173	Western Europe	
B.1.1.519							+			+	1,751		
B.1.243										+	2,898		
B.1.258	+			+							6,174		

Table 1. SARS-CoV-2 lineages and key variants in the spike protein. SARS-CoV-2 lineages deposited in GISAID were analyzed for the presence of the main variants of concern or interest in the spike protein. Variants were identified using an internally developed genotyping approach. For each lineage, defined by Pangolin, only the combination of variants with the highest count has been included; for B.1.526 two sub-lineages with different VOC are being tracked;[9] the count of B.1.258 includes minor lineages (B.1.258.*) with the same subset of defining variants in the spike protein as shown here. Counts are of the number of complete, high coverage strains deposited in GISAID between 1-Nov-2020 and 29-Mar-2021, excluding strains collected before 1-Nov-2020 (total =342,815).

B.1.1.7 challenges

Lineage assignment does not always correlate with VOC

Identifying and tracking the emergence and spread of these SARS-CoV-2 variants is crucial for public health. However, the current methods have several shortcomings that could be hampering such efforts. For example, the so-called B.1.1.7 lineage, first identified in the UK, is typically characterized by the del69_70, N501Y, and P681H variants. However, our review of strains annotated to be of the B.1.1.7 lineage deposited in GISAID[10,11] shows a wider range of variants associated with this lineage (Table 2). Fundamentally, lineages are assigned by the Pangolin software based on a minimum number of matching defining variants, without prioritizing variants in critical functional domains (e.g. RBD), let alone specific VOC. B.1.1.7 has a total of 17 non-synonymous variants across the viral genome, but only 8 variants within the spike protein. However, the underlying algorithm assigns B.1.1.7 to sequences with at least 5 of the 17 defining variants. Consequently, sequences have been assigned to B.1.1.7 even in the absence of the N501Y variant, which has been one of the defining VOC for this strain.

Conversely, by focusing on lineages and not on the underlying variants, VOC that arise in novel or "misclassified" lineages can be missed. The variant N501Y has not only been observed in B.1.1.7, B.1.351 and P.1, but also in 100 other lineages (Figure 1A).

Lineage	L18F	del69_70	del144	A222V	E484K	N501Y	P681H	Count
B.1.1.7		+	+			+	+	125,484
B.1.1.7		+				+	+	1,817
B.1.1.7			+			+	+	1,636
B.1.1.7						+	+	1,583
B.1.1.7		+	+				+	470
B.1.1.7	+	+	+			+	+	463
B.1.1.7		+	+			+		313
B.1.1.7		+	+	+		+	+	165
B.1.1.7		+	+		+	+	+	63

Table 2. Variant diversity in strains annotated as B.1.1.7. SARS-CoV-2 lineages annotated as B.1.1.7 in GISAID were analyzed for the presence of variants of concern in the spike protein. Multiple subgroups exist that lack key variants such as N501Y and del69_70; these specimens were assigned to B.1.1.7 as a result of characterizing variants in other parts of the viral genome. Note that this distribution is biased due to the large proportion of UK specimens deposited in GISAID: ~63%of B.1.1.7 specimens were sequenced in the UK.

S-gene dropout is a flawed proxy for B.1.1.7

While large-scale strain sequencing efforts across the globe have been invaluable to identify and track VOC, alternate methods that use proxy signals rather than whole genome sequencing have also been employed. Most notably, so-called "S-gene dropout" has been used to track the spread of the B.1.1.7 lineage, and implicitly the VOC N501Y. This method assumes that samples that test positive for SARS-CoV-2, but in which the S-gene (which encodes for the spike protein) is not detected are an indication of the B.1.1.7 lineage due to the del69_70 variant causing reduced amplification of the S-gene probes. While this pattern can indeed be used to triage suspected cases for confirmatory sequencing, it cannot be used to uniquely identify this lineage. Specifically, the deletion of amino acids 69-70 has been observed in 105 lineages, of which only 34 lineages carry the relevant N501Y variant. For example lineage B.1.258, which has frequently been observed in Denmark (n=1,041) and Slovenia (n=1,062), carries del69_70 but not N501Y (Table 1, Figure 1B). Within the US, we observed that S-gene dropout was strongly associated with N501Y in some states (e.g. California), but not in others (e.g. Massachusetts and Rhode Island), which could confuse narratives about the spread of B.1.1.7.

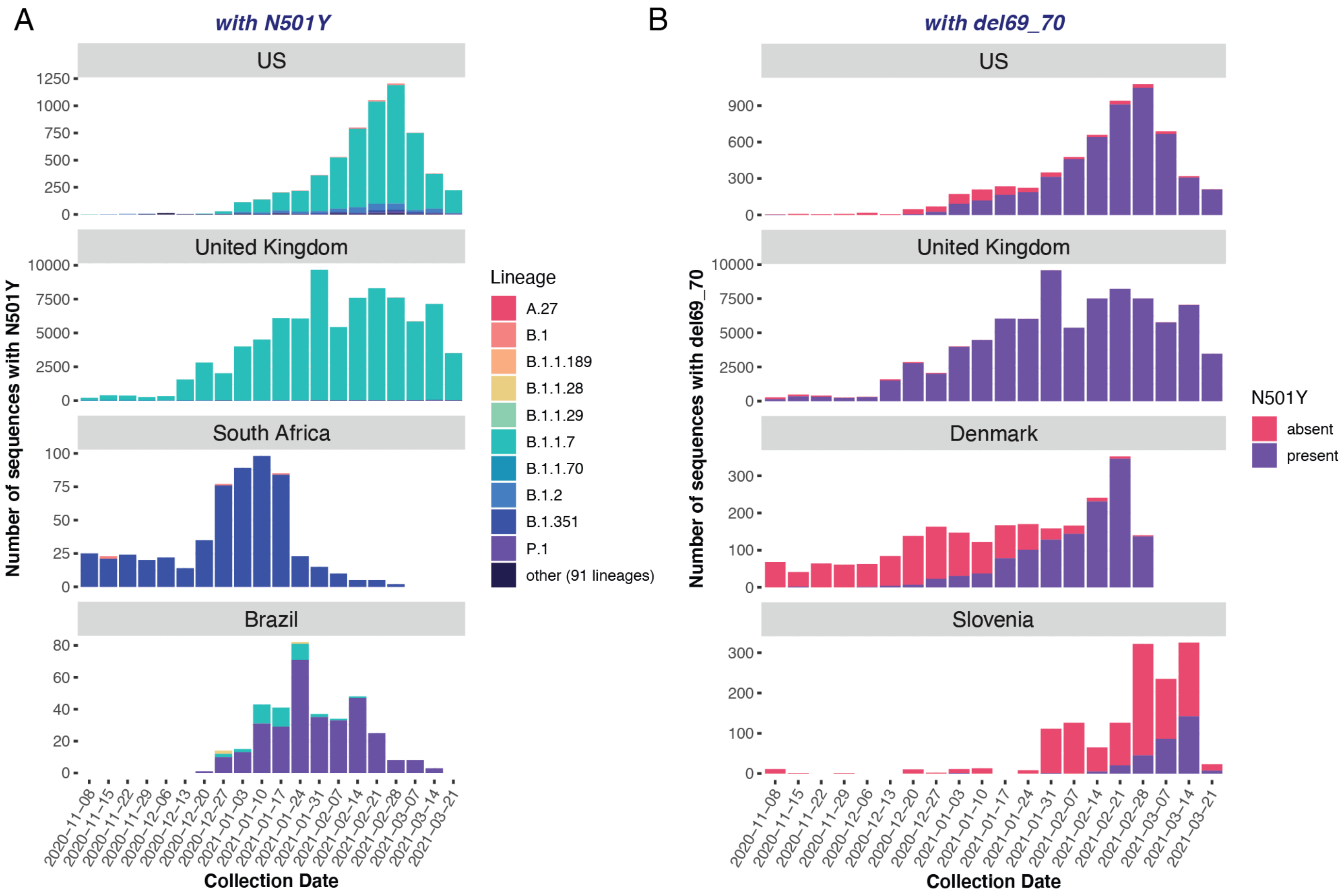

Figure 1. Characteristic variants of the B.1.1.7 lineage. (A) Specimens containing N501Y, one of the key variants of the B.1.1.7 lineage, are associated with a total of 103 lineages worldwide; lineages with ≤50 observations were lumped into the "other" bucket. **(B)** Strains with the del69_70 variant that causes S-gene dropout vary in their association with N501Y across different countries.

E484K can be missed by lineage-based classifications

In recent months, the world's attention has been drawn to the E484K variant. It was first identified in the B.1.351 lineage in South Africa, where it has been reported to reduce the efficacy of vaccines (especially the Oxford/AstraZeneca vaccine) and antibody treatments. However, we and others have observed that E484K is found in diverse lineages beyond B.1.351. In our observations it occurred not only in B.1.351 (n=3,176), P.2 (1,212) and P.1 (861), but also a total of 3,279 times (38.4% of observations) across 102 different lineages (Figure 2A). Assuming many of these lineages have emerged independently, this supports the hypothesis that E484K confers a competitive advantage in some backgrounds.

E484K has been observed at least 2,894 times in the USA, conservatively estimated based on complete sequences collected between 1-Nov-2020 and 21-Mar-2021. Only 625 of these observations (21.6%) were annotated as the "South Africa" and "Brazil" lineages (B.1.351, P.1 & P.2) (Figure 2B). E484K has emerged both in the New York cluster (B.1.526) and in other lineages across 49 states in the US. The rapidly increasing frequency of variants at the E484 position, including E484K and E484Q, amongst all strains in the US supports the urgent tracking of this variant.

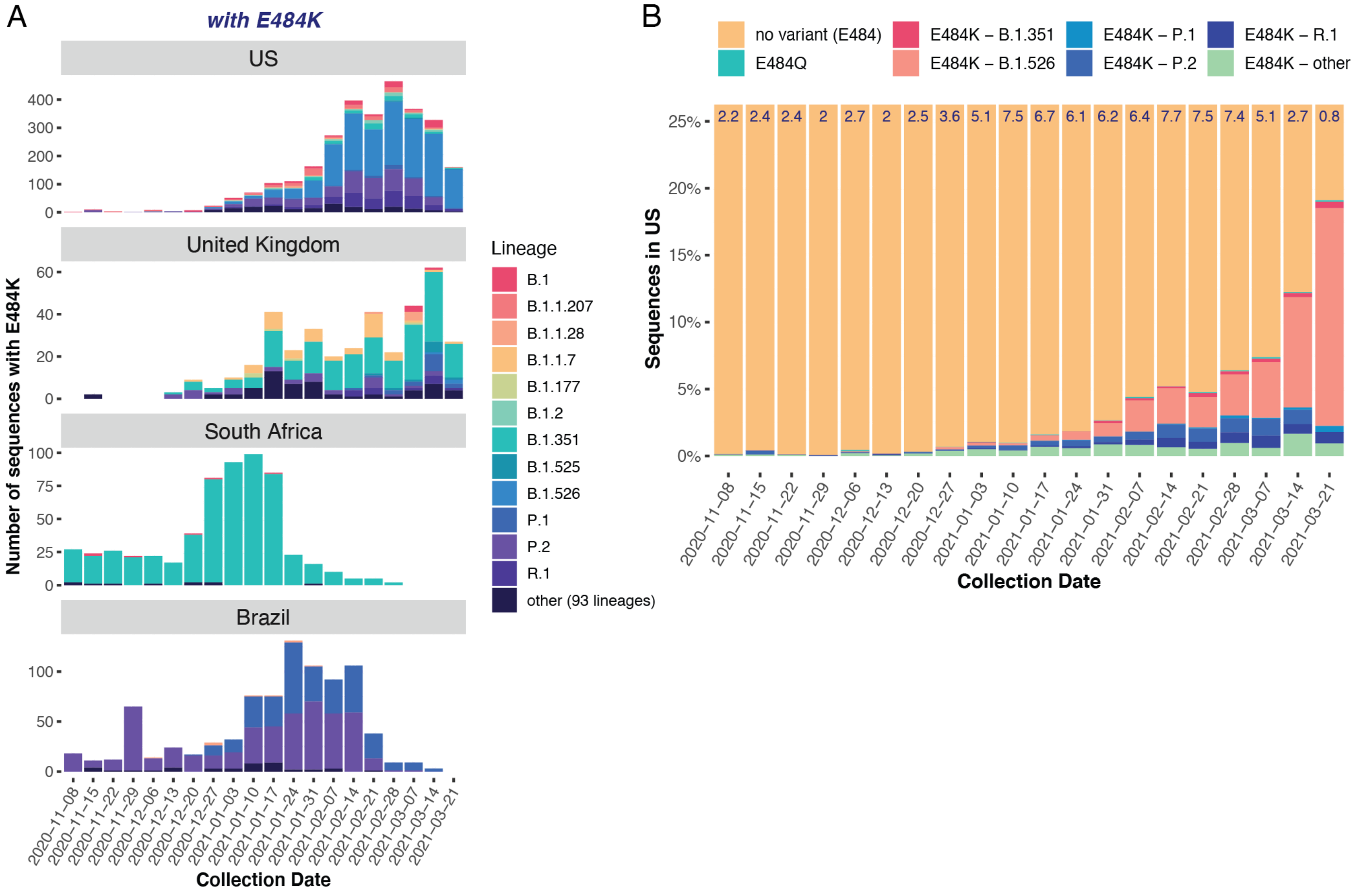

Figure 2. E484 variants are found across diverse lineages. (A) Specimens containing E484K have been assigned to a total of 105 lineages world-wide, including the widely tracked B.1.351 and P.1/P.2; lineages with ≤50 observations were lumped into the "other" bucket. **(B)** Normalizing the frequency of a VOC against the number of specimens sequenced in the US shows that E484K is being observed at an increasing frequency; the number at the top of each bar denotes the weekly total (*1,000) of sequenced specimens (complete and high coverage sequences). Note that large-scale sequencing of local outbreaks (e.g. B.1.526 in New York) will introduce bias in the relative contribution of these lineages.

Discussion

We found that annotation with a specific lineage (e.g. the "UK" strain B.1.1.7) does not guarantee that sequence to carry the VOC (e.g. N501Y). Tracking VOC using viral lineages provides an incomplete view, especially as the current techniques for assigning strains to lineages do not give weight to key variants.

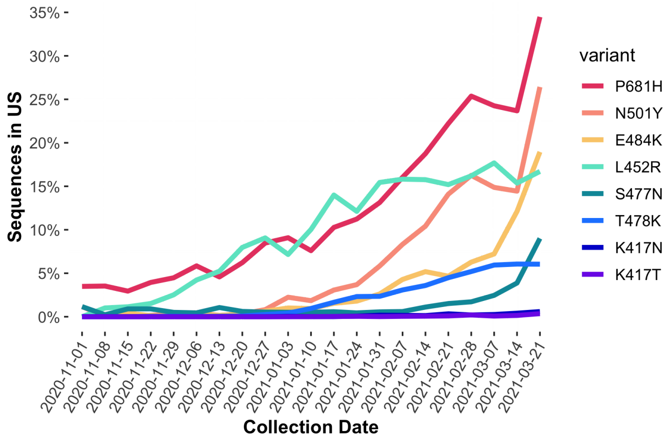

Figure 3. VOC prevalence in the United States. Tracking individual variants provides a robust approach to become and/or remain aware of variants of concern such as E484K and N501Y.

Instead of focusing tracking efforts and scientific communication on lineages, an alternative approach could be to focus on individual or specific combinations of VOC.[4] For example, Figure 3 tracks VOC in the United States during the winter months. This reveals the gradual rise of cases with N501Y, beyond those cases annotated as B.1.1.7, and E484K. Similarly, the steady increase of cases with the variants of concern L452R (associated with the recent USA/West Coast outbreak) and P681H can be directly observed.

However, one limitation with this approach is that the effects of these variants may be impacted by other variants that are present in these lineages, a phenomenon known as epistasis. Therefore it may be necessary to track these variants both within the context of lineages that are defined by key variants, as well as track VOC on their own.

References

1. Wang R, Chen J, Gao K, Hozumi Y, Yin C, Wei G-W. Analysis of SARS-CoV-2 mutations in the United States suggests presence of four substrains and novel variants. *Commun Biol*. 2021;4(1):228.
2. Lauring AS, Hodcroft EB. Genetic Variants of SARS-CoV-2-What Do They Mean? *JAMA*. 2021;325(6):529-531.
3. Greaney AJ, Loes AN, Crawford KHD, et al. Comprehensive mapping of mutations to the SARS-CoV-2 receptor-binding domain that affect recognition by polyclonal human serum antibodies. *Cold Spring Harbor Laboratory*. Published online January 4, 2021:2020.12.31.425021.
4. Callaway E. "A bloody mess": Confusion reigns over naming of new COVID variants. *Nature*. 2021;589(7842):339.
5. 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.
6. CDC. SARS-CoV-2 Variant Classifications and Definitions.
7. Wu K, Werner AP, Moliva JI, et al. mRNA-1273 vaccine induces neutralizing antibodies against spike mutants from global SARS-CoV-2 variants. *Cold Spring Harbor Laboratory*. Published online January 25, 2021:2021.01.25.427948.
8. Liu Y, Liu J, Xia H, et al. Neutralizing Activity of BNT162b2-Elicited Serum. *N Engl J Med*. Published online March 8, 2021.
9. Zhou H, Dcosta BM, Samanovic MI, Mulligan MJ, Landau NR, Tada T. B.1.526 SARS-CoV-2 variants identified in New York City are neutralized by vaccine-elicited and therapeutic monoclonal antibodies. *bioRxiv*. Published online March 24, 2021:2021.03.24.436620.
10. Elbe S, Buckland-Merrett G. Data, disease and diplomacy: GISAID's innovative contribution to global health. *Glob Chall*. 2017;1(1):33-46.
11. Shu Y, McCauley J. GISAID: Global initiative on sharing all influenza data - from vision to reality. *Euro Surveill*. 2017;22(13).