Emergence of a recurrent insertion in the N-terminal domain of the SARS-CoV-2 spike glycoprotein

Marco Gerdol[1*]

[1]University of Trieste, Department of Life Sciences, Via Giorgieri 5, 34127 Trieste, Italy

Correspondence should be addressed to: mgerdol@units.it

Abstract

Tracking the evolution of the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) through genomic surveillance programs is undoubtedly one of the key priorities in the current pandemic situation. Although the genome of SARS-CoV-2 acquires mutations at a slower rate compared with other RNA viruses, evolutionary pressures derived from the widespread circulation of SARS-CoV-2 in the human population have progressively favored the global emergence though natural selection of several variants of concern that carry multiple non-synonymous mutations in the spike glycoprotein. Such mutations are often placed in key sites within major antibody epitopes and may therefore confer resistance to neutralizing antibodies, leading to partial immune escape, or otherwise compensate minor infectivity deficits associated with other mutations. As previously shown by other authors, several emerging variants carry recurrent deletion regions (RDRs) that display a partial overlap with antibody epitopes located in the spike N-terminal domain. Comparatively, very little attention has been directed towards spike insertion mutations, which often go unnoticed due to the use of insertion-unaware bioinformatics analysis pipelines. This manuscript describe a single recurrent insertion region (RIR1) in the N-terminal domain of SARS-CoV-2 spike protein, characterized by the independent acquisition of 3-4 additional codons between Arg214 and Asp215 in different viral lineages. Even though RIR1 is unlikely to confer antibody escape, its progressive increase in frequency and its association with two distinct emerging lineages (A.2.5 and B.1.214.2) warrant further investigation concerning its effects on spike structure and viral infectivity.

Introduction

Coronaviruses generally accumulate mutations at a much lower rate than other RNA viruses, thanks to the efficient proofreading exonuclease activity exerted by nsp14, in complex the activator protein nsp10 [1,2]. As a result, the rate of molecular evolution of SARS-CoV-2 is currently estimated (as of April 10th, based on GISAID data [3]), to be close to 24 substitutions/year per genome, i.e. 8.25×10^{-4} substitutions/site/year, which is very close to previous estimates for human endemic coronaviruses [4]. Consistently with comparative genomics data obtained from other members of the Sarbecovirus subgenus, such mutations are not evenly distributed across the genome, but they are disproportionally located in the S gene, which encodes the spike glycoprotein. It is also worth noting that the S gene undergoes frequent recombination events, likely as a result of naturally occurring co-infections in the animal viral reservoirs [5], and that these events are theoretically possible also among different SARS-CoV-2 lineages [6]. The encoded transmembrane protein forms a homotrimer and plays a fundamental role in the interaction between the virus and host cells, promoting viral entry though the interaction with different membrane receptors [7], which, in the case of SARS-CoV-2 and of the closely related SARS-CoV responsible of the 2002-2004 outbreak, is represented by the angiotensin converting enzyme 2 (ACE2) [8,9].

While most of these mutations have little or no phenotypic impact at all, some may significantly influence viral transmissibility and the ability of the virus to escape host immune response. The causes underpinning such phenotypic effects may either lie in an increased viral shedding, in the alteration of the binding affinity between the spike receptor binding domain (RBD) and the host ACE2 receptor, or in the modification of key

antibody epitopes. The most striking example of a non-synonymous mutations which had a dramatic impact on the dynamics of the pandemics is most certainly represented by S:D614G. This mutation, which was not present in the ancestral lineage that caused the Wuhan outbreak, emerged in the very early phases of the pandemics, quickly becoming dominant worldwide [10], most likely due to an increased packing of functional spike protein into the virion [11]. As of April 2021, D614G is associated with more than 99% of the SARS-CoV-2 genomes sequenced daily and just a very limited number of lineages lacking this mutations are linked with local outbreaks (e.g. A.23.1 in Uganda) [12].

Even though the mutation rate of the SARS-CoV-2 genome remained relatively stable throughout 2020, growing evidence soon started to point out the presence of shared mutations across multiple independent lineages, suggesting ongoing convergent evolution and possible signatures of host adaptation [13]. While early investigations failed to identify evidence of increased transmissibility associated with such recurrent mutations [14], the nearly contemporary emergence, in different geographical locations and on different genetic backgrounds, of three variants sharing the S:N501Y non-synonymous substitution started to raise serious concerns about the possible involvement of this mutation in increasing viral infectivity. While the functional role of N501Y still remains to be fully clarified, structural modeling points towards a possible stabilization of the spike protein in the open conformation, which may increase ACE2 binding, especially in combination with other mutations targeting the RBD [15,16].

B.1.1.7, one of the three emerging lineages carrying N501Y, which originally spread in southeastern England, appears to be significantly more transmissible than wild-type genotypes [17], but is not associated with significant immune escape from the neutralizing activity of convalescent or vaccinated sera [18–21]. On the other hand, some point mutations present in the spike N-terminal domain (NTD), i.e. the deletion of a codon in position 144, lead to full escape from the activity of a few NTD-directed monoclonal antibodies [22].

The two other major variants of concerns (VOCs), B.1.351 and P.1 were linked with major outbreaks in geographical regions with very high estimated seroprevalence, i.e. in the Eastern Cape region (South Africa) [23] and in Manaus (Amazonas, Brazil) [24], respectively. Both variants are characterized by a constellation of non-synonymous mutations and accelerated rates of evolution, which suggests that their selection might have occurred in immunocompromised patients with persistent viral infection [25]. Among the many features shared by B.1.531 and P.1, the most remarkable one is the presence of two additional RBD mutations: E484K and K417N/K417T. The former one is now found in a number of other emerging lineages and it has been identified as a major player in antibody escape, due to its presence in a major epitope recognized by class II RBD-directed antibodies [26–28]. K417 on the other hand is located in an epitope recognized by class I antibodies, and is thought to provide a minor contribution to polyclonal antibody response escape [26] and to possibly stabilize, together with E484K and N501Y, the interaction between the RBD and the ACE 2 receptor [15]. The trends of new infections connected with these VOCs and other emerging variants of interest (VOIs) are being closely monitored due to the possible negative impact they might have on massive vaccination campaigns [29,30]. More recently, the status of two additional closely related lineages, first described in California, B.1.427 and B.1.429, was raised by the CDC from VOI to VOC [31]. These variants may be linked with a moderate increase in transmissibility and are associated with immune escape, due to L452R being located in a major class III RBD-directed antibody epitope [30].

Among the five aforementioned VOCs, two (B.1.1.7 and B.1.351) carry spike deletions in the NTD. Concurrently with their global spread [17,32], as well as with the emergence of large local outbreaks linked with other VOIs, such as B.1.258 [33] and B.1.526.1 [34], the overall frequency of observation of spike deletions dramatically increased over time. Such deletions were previously shown to specifically occur in four distinct NTD sites, named recurrent deletion regions (RDR) 1, 2, 3 and 4, which arose in different geographical backgrounds, in independent viral lineages, and strikingly display a significant overlap with known immune epitopes, suggesting that they may drive antibody escape [35].

Comparatively, very little attention has been directed towards spike insertions, even though such events are known to have played a fundamental role in the past evolution of SARS-CoV-2 spike protein, allowing, among the other things, the acquisition of a furin-like cleavage site. This short motif, which is thought to be a key pathogenicity determinant [36], is indeed completely absent in the closely related sarbecovirus RaTG13 [37] and only partly present in the recently described RmYN02 [38].

This work suggests that the impact of spike insertion mutations on SARS-CoV-2 genome evolution, albeit much lower than spike deletions, has been probably so far neglected. Several distinct insertion events have indeed repeatedly occurred at the very same NTD site, located between Arg214 and Asp 215, which will hereafter be referred to as Recurrent Insertion Region 1 (RIR1), whose functional implications presently remain unclear. The international spread of the A.2.5 and B.1.214.2 lineages, which both carry a three-codon insertion at RIR1 combined with multiple other non-synonymous spike mutations, strongly suggest that more attention should be put in the near future towards the acquisition of similar insertions in known VOCs and VOIs. Moreover, several bioinformatics tools presently used for SARS-CoV-2 genome assembly and variant calling often mistakenly disregard true insertion mutations, considering them as sequencing errors, which highlight that more efforts should but put towards the development and use of insertion-aware algorithms for a more efficient monitoring of emerging SARS-CoV-2 variants.

Materials and methods

All SARS-CoV-2 genome data used in this study were retrieved from GISAID (last access date April 10th 2021) [3]. In detail, all available sequenced genomes belonging to the A.2.5 and B.1.214.2 lineages and associated metadata were downloaded. While all GISAID entries were considered for reporting observation frequencies, only high quality genomes (i.e. those listed as “complete” and “high coverage”) associated a sampling date were taken into account for further analysis. Genomes containing long stretches of NNNs (i.e. > 1% of the full sequence) were discarded. Entries belonging to the sister lineages A.2.4, B.1.214, B.1.214.1, B.1.214.3 and B.1.214.4 were also retrieved and processed as mentioned above. The reference isolate Wuhan-Hu-1 was also included for tree rooting purposes. Note that several genome sequences from Panama with sampling date anterior to November 2021 were discarded due to the unreliability of associated metadata (i.e. the sampling dates appeared to be inconsistent with the very small genetic distances with recent isolates belonging to the same lineage). Overall, the A.2.5- and B.1.214.2-focused datasets included 201 and 248 sequences, respectively.

SARS-CoV-2 genome data were analyzed with the nextstrain *augur* pipeline (https://github.com/nextstrain/augur). Briefly, nucleotide sequences were aligned with MAFFT [39] and the resulting multiple sequence alignment was used as an input for a maximum likelihood phylogenetic inference analysis, carried out with IQ-TREE [40] under a generalized time reversible (GTR) model of molecular evolution. The resulting tree was further refined with *augur* using sampling date metadata, generating a timetree. Phylogenetic trees were rooted based on the oldest genotype available, which in this case was Wuhan-Hu-1, and graphically rendered using FigTree v.1.1.4.

The global frequency of insertion and deletion mutations mapped on the SARS-CoV-2 S gene was retrieved from https://mendel.bii.a-star.edu.sg/ (credit to Raphael Tze Chuen Lee). Disruptive insertion and deletion mutations (i.e. those that interrupted the open reading frame of the S gene) were discarded.

Root-to-tip genetic distance analyses were performed by plotting the sampling dates against the total number of nucleotide substitutions (excluding insertions and deletions) observed in the A.2.5, B.1.214.2 and related lineages. These were calculated with MEGA X [41], compared with the reference genotype Wuhan-

Hu-1. The global average mutation rate of SARS-CoV-2 was retrieved from GISAID (as of April 10th 2021, roughly equivalent to 24 substitutions per year per genome).
The number of non-synonymous mutations, insertions and deletions affecting the spike protein of at least 50% of the genomes belonging to the A.2.5 and B.1.214.2 lineages sampled from February 1st 2021 was calculated with MEGA X [41], compared with the reference genotype Wuhan-Hu-1. The same metric was similarly computed for the variants of concern B.1.1.7, B.1.351.1, B.1.427, B.1.429 and P.1, and for the variants of interest A.23.1, A.29, B.1.525, B.1.526, B.1.526.1, B.1.526.2, B.1.617, P.2, P.3 and R.1, taking into account genomes with sampling date posterior to February 1st 2021. The widespread European lineage B.1.177 was used as a control lineage with no evidence of increased transmissibility or immune evasion properties.

Results and discussion

Presence of a recurrent insertion region (RIR1) in the N-terminal domain of SARS-CoV-2 spike protein

The analysis of the genomic data deposited in GISAID as of April 10th 2021 revealed that S gene insertions (excluding those that disrupted the open reading frame) were present in just a minor fraction of all sequenced SARS-CoV-2 genomes, i.e. roughly 0.1% of the total. The impact of insertion mutations in the S gene on viral evolution has been very limited compared with deletion mutations, which are currently found in several widespread lineages, such as B.1.1.7 and B.1.351. Overall, spike deletions have been observed with a frequency approximately 750 folds higher than insertions. As previously reported by other authors, most deletions occur in specific sites of the N-terminal domain, named Recurrent Deletion Region (RDR) 1, 2, 3 and 4 (**Figure 1**) [35], which is consistent with the higher rate of mutation observed for the S1 region in human coronaviruses compared with the more slowly evolving S2 subunit [42].
Despite their lower frequency of occurrence, insertions do not occur randomly in the S gene. Here we show that the overwhelming majority of the insertion mutations mapped so far in SARS-CoV-2 target the NTD, being in most cases (i.e. ~1500 genomes) identified at a specific site, located between Arg214 and Asp215 (**Figure 1**). Due to the convergent finding of such insertions in independent viral lineages (see below), this region will be hereafter named Recurrent Insertion Region 1 (RIR1).

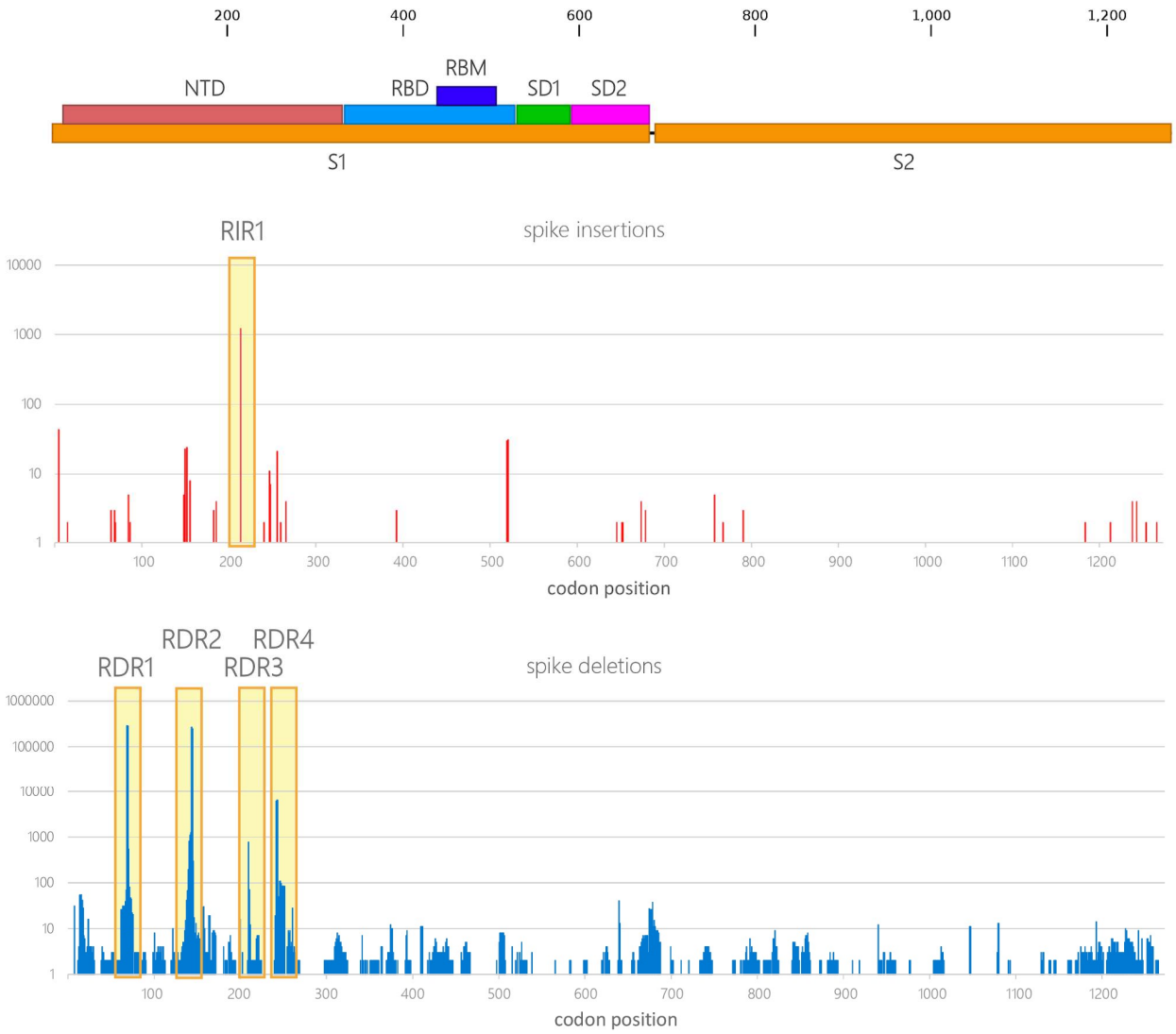

Figure 1. Schematic representation of the SARS-CoV-2 protein, with indication of the two functional S1 and S2 subunits, which are separated by a furin-like proteolytic cleavage site, the N-terminal domain (NTD), the receptor binding domain (RBD) and receptor binding motif (RBM), the SD1 and SD2 subdomains. The absolute number of observations of insertion and deletion mutations found along the S-gene are reported in a Log_{10} scale (https://mendel.bii.a-star.edu.sg/ was last accessed on April 10th 2021). The position of RDR1-RDR4 from a previous study [35], as well as the position of the newly identified RIR1 are reported.

We found that RIR1 insertions were the result of ten independent events that occurred in different branches of the SARS-CoV-2 global phylogenetic tree, which strongly suggests convergent evolution (**Table 1**). Even though the length of the insertion either spanned three or four codons (**Figure 2**), the overwhelming majority of the genomes with RIR1 insertions (99.5% of the total) only included three (**Table 1**). Out of these, only two insertion events have led to a significant community spread, whereas in the other eight cases the small number of sequenced samples and the short timespan of detection points to isolated cases that likely did not lead to a significant number of secondary infections (**Table 1**).

While the two major types of RIR1 insertions (i.e. those characterizing the A.2.5 and B.1.214.2 lineages) will be discussed in detail below, it is worth nothing that some of the other occasional events occurred in lineages that have been previously identified in VOCs. In detail, the S:ins214AQER mutation (**Figure 2**), reported in a single patient in California on January 15th 2021, occurred in a viral isolate with a B.1.429 genetic background [31]. On the other hand, S:ins214ANRN was found in four P.1 lineage genomes in in Manaus (Amazonas,

Brazil) between December 2020 and January 2021, which suggests that limited transmission in the community might have been present at that time. P.1 is currently considered one of the major VOCs and its worldwide prevalence is being constantly monitored due to the combination of three key RBD mutations (K417T, E484K and N501Y), which may jeopardize the efficiency of some vaccines by promoting antibody escape [24]. Finally, S:ins214KFH, documented in three patients in Scotland in February 2021, could be linked with the highly transmissible B.1.1.7 lineage, which is now dominant in Europe and quickly spreading worldwide [43].

Taking into account the limited efforts carried out by several countries in genomic surveillance, those reported in Table 1 may represent just a fraction of the SARS-CoV-2 variants carrying insertions at RIR1 that emerged during the course of the pandemics, either already extinct or currently part of ongoing community transmission.

Insertion	Lineage	GISAID entries	Other spike mutations	Earliest detection	Community transmission
AKKN	B	1	none	Mar 5th, 2020	no
AAG	A.2.5	600	L141del, G142del, V143del, D215Y, L452R, D614G	Apr 12th, 2020	yes
KLGP	B.1.177	1	E154K, A222V, D614G	Nov 13th, 2020	no
TDR	B.1.214.2	601	Q414K, N450K, D614G, T716I	Nov 22nd, 2020	yes
ANRN	P.1	4	L18F, P26S, D138Y, K417T, E484K, N501Y, D614G, D1139H, V1176F	Dec 23rd, 2020	unknown
AQER	B.1.429	1	S13I, P26S, S98F, W152C, L452R, D614G, T1027I	Jan 15th, 2021	unlikely
DLA	B.1.2	3	D614G	Jan 17th, 2021	unlikley
DRS	B.1	1	D215N, V382L, D614G, M1237I	Feb 1st, 2021	unlikley
KRI	B	3	V367F, E990A	Feb 3rd, 2021	unlikley
KFH	B.1.1.7	3	H69del, V70del, Y144del, N501Y, A570D, D614G, P681H, T716I, S982A, D1118H	Feb 12th, 2021	unlikely

Table 1. Summary of the ten independent RIR1 insertions found in the SARS-CoV-2 genome, ordered by the earliest date of detection. The assessment of presence and absence of community transmission was based on the date of the latest reported cases and on the quality of the molecular surveillance programs carried out by different countries.

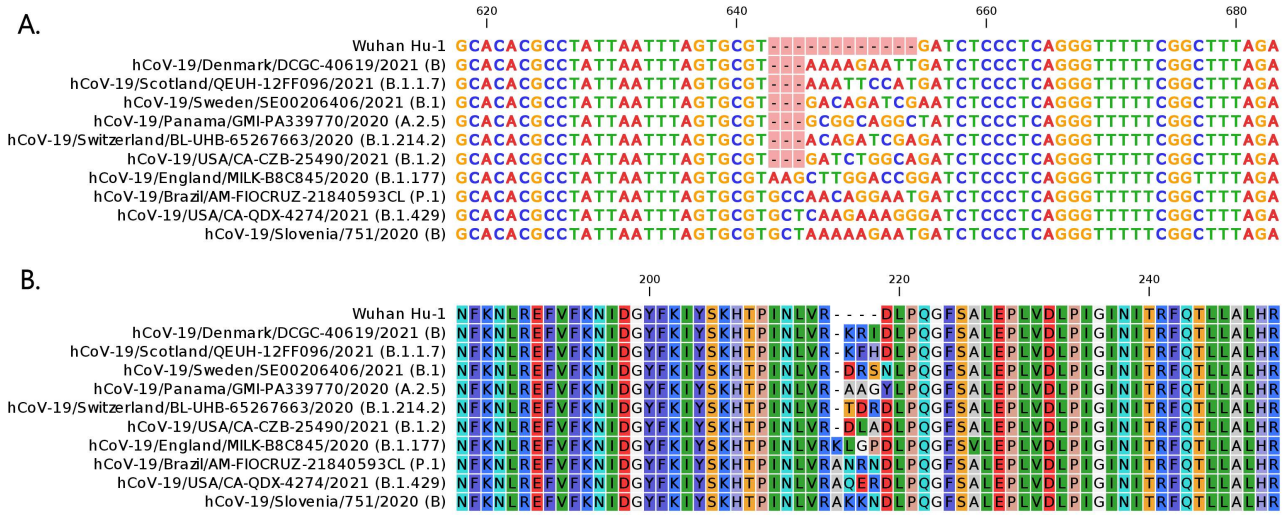

Figure 2. Multiple sequence alignment of the nucleotide (panel A) and translated protein (panel B) sequences of the SARS-CoV-2 S gene of the 10 viral lineages characterized by an insertion at RIR1, compared with the reference sequence Wuhan Hu-1. The multiple sequence alignment only displays a small portion of the S gene and of the encoded spike protein, zoomed-in and centered on RIR1.

Characterization of the lineage-defining mutations in A.2.5 and B.1.214

The only two lineages with insertions at RIR1 with solid evidence of current community transmission are A.2.5 and B.1.214.2, whose context of emergence will be discussed in detail in the following sections.

The insertion found in A.2.5 is S:ins214AAG, corresponding to the GCGGCAGGC nucleotide sequence (**Figure 2**). Remarkably, this is paired with a non-synonymous substitution of Asp215 to Tyr (as a result of a GAT->TAT codon replacement). Besides the insertion at RIR1, A.2.5 also displays the deletion of three codons (Δ141-143) in RDR2 (**Figure 1**), sometimes extending to codon 144. This region has been previously implicated in antibody escape [35] and show deletions in some relevant VOCs and VOIs, including B.1.1.7, B.1.525 and B.1.526.1. In particular, Δ144 appears to largely explain the resistance towards several NTD-directed monoclonal antibodies displayed by B.1.1.7 in vitro (Wang et al., 2021). Moreover, the insertion at RIR1 is also combined with L452R, a key mutation that confers resistance towards class III RDB-directed antibodies [26], including LY-CoV555, the basis for the commercial monoclonal antibody bamlanivimab developed by Eli Lilly (Starr et al., 2021). L452R is also found in the VOCs B.1.427 and B.1.429, which are characterized by a moderate increase in transmissibility [44], as well as in other emerging VOIs, including A.2.7 and B.1.617. Like the overwhelming majority of the variants circulating in 2021, A.2.5 also characterized by the presence of the prevalent mutations D614G. The summary of the mutations found in A.2.5 is reported in **Figure 3**.

The typical insertion found in the lineage B.1.214.2 is S:ins214TDR, which corresponds to a ACAGATCGA nucleotide sequence. This is combined with four non-synonymous mutations (**Figure 3**): besides D614G, Q414K and N450K are found in the RBD and T716I, which is shared by B.1.1.7, is found in the S2 region. As far as the two RBD mutations are concerned, Q414K is located close to K417, a key site mutated in the VOCs P.1 (K417T) and B.1.351 (K417N). This residue may be therefore part of an epitope recognized by class 1 antibodies, whose neutralizing activity is escaped by the two aforementioned VOCs [22,26]. On the other hand, N450 is also involved in the surface of interaction between the RBD and some monoclonal antibodies [45]; moreover, in a deep mutational scanning experiment, the mutation N450K determined a mild increase in ACE2 binding [28]. Currently, neither of these two RBD mutations are currently associated with other widespread SARS-CoV-2 variants. It is also worth noting that B.1.214.2 displays a large deletion (10 codons) in the N-terminal region of ORF3a, which is not found in other relevant SARS-CoV-2 lineages (**Figure 3**).

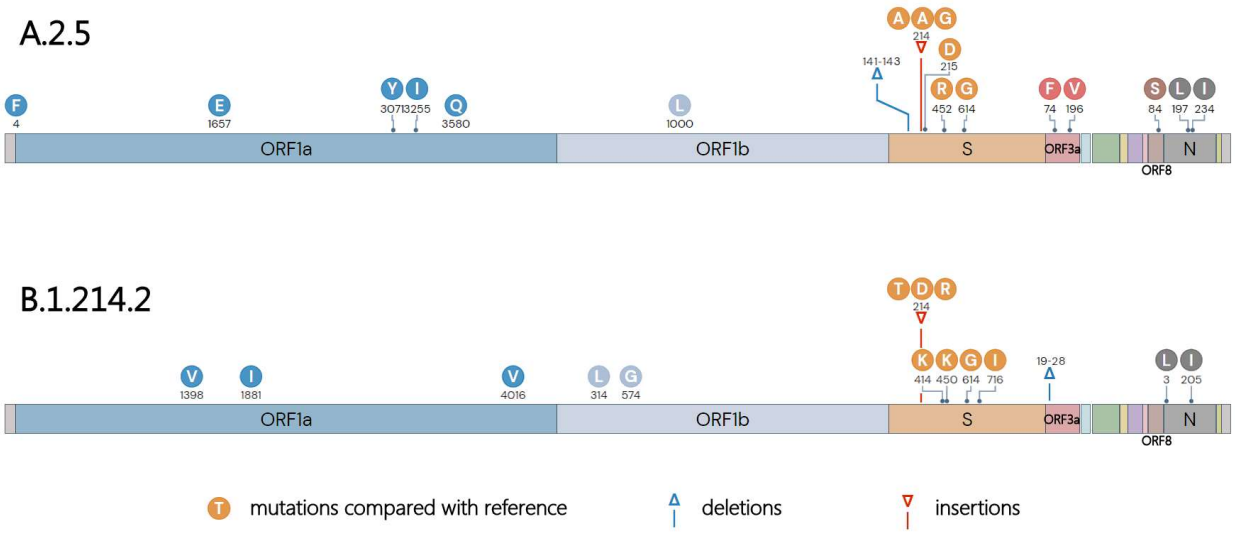

Figure 3. Key mutations associated with the A.2.5 and B.1.214.2 lineages. Genes associated with mutations (compared with the reference strain Wuhan-Hu-1) are indicated. Note the AAG and TDR insertions at RIR1. Modified from https://outbreak.info/.

Context for the emergence of A.2.5

A.2.5 stems from A.2.4, the dominant lineage in the Panama pandemics during the first half of 2020 [46]. This lineage started spreading very early in the region, with the first detections in oropharyngeal swabs sampled in late February 2020. However, A.2.4 remained largely confined to Panama, with limited evidence of international exportation, as revealed by the fact that ~98% of the genomes belonging to this lineage deposited in GISAID are from this country. While the frequency of observation of A.2.4 significantly declined over time, this lineage likely continues to be responsible of a non-negligible fraction of covid-19 cases in Panama, according to molecular surveillance data from 2021 (8 out of 198 sequenced genomes, i.e. ~4%).

The precise timing of the emergence of the A.2.5 lineage, along with the acquisition of the S:ins214AAG insertion and of the other associated mutations described in the previous section, is presently unclear due to the insufficient molecular surveillance carried out in Central America during 2020. Any inference concerning the timing of the emergence of a given lineage relies on the accuracy of the metadata associated to each genome submission. With this respect, a few genomes belonging to the A.2.5 lineage, presumably sampled in Panama from April to August 2020, have been deposited in GISAID on April 8th 2021. Unfortunately, linear root-to-tip regression analysis (data not shown), as well as the high genetic relatedness with genotypes isolated in early 2021, strongly suggest that these samples might have been mistakenly assigned a wrong sampling date. Hence, the first reliable cases linked with A.2.5 in Panama can be traced back to late November 2020, all within a 100 km^2 area around the capital city Panamá. To date, less than 900 sequenced genomes out of over 350K reported covid-19 cases have been reported in this country, i.e. 0.25% of the total, far below of the threshold that would be sufficient to track emerging variants [47]. Nevertheless, A.2.5 underwent an evident expansion in Panama between December 2020 and February 2021, as revealed by the increase in estimated prevalence from ~60 to ~95%.

Not surprisingly, the remarkable spread of SARS-CoV-2 in the country (as of April 10th 2021, the incidence of infections surpassed 80,000 cases per million inhabitants) was connected with a significant number of exported cases, which have sometimes led to clusters of infection abroad. The A.2.5 lineage likely spread very early also in the neighboring Costa Rica, where multiple introductions can be inferred from the time tree, in spite of the limited amount of genomic information available (**Figure 4**). While earliest investigations

failed to identify the variant in August 2020 [48], the most recent data deposited in GISAID indicate that it may be responsible of ~30% of the covid-19 cases documented in 2021. While the spread of A.2.5 to other geographically close countries (e.g. Nicaragua, El Salvador and Honduras) seems likely, the lack of molecular surveillance in such regions prevents drawing definitive conclusions.

The first evidence of the detection of A.2.5 outside from Central America dates to December 1st 2020, in Ecuador. Reports in other Latin American countries remain sporadic, but it is worth noting that A.2.5 genomes have been so far sequenced in Suriname, Colombia, Mexico and Paraguay. A single imported cases have been also reported with similar timing in UAE (December 27th, 2020), and three were intercepted though border screening in Australia in January 2020. The first documented cases linked with A.2.5 in Europe were identified in Luxembourg, Portugal and Germany, but these are unlikely to have led to further community transmission. On the other hand, the significant number of cases reported in Campania in February-March 2021 (i.e. 98, accounting for 1% of the genotypes sequenced in this period in the region), strongly hints the occurrence of transmission in the community. Similarly, imported cases have most certainly led to local cluster of infections in different areas of the United States, with evidence of multiple independent introductions both in New Yok and Florida (**Figure 4**), even though these only represent a minor fraction (i.e. <1%) of the total number of cases recorded in the two states between February and March, 2020.

A.2.5 belongs to one of the very few surviving children lineages of the ancestral lineage A, which, after several months of limited global spread, have recently led to a few major clusters of infections. Such events, that occurred for example in Uganda (A.23.1) [12] and, most likely, also in Tanzania (A.VOI.V2) [12], are characterized, like in the case of A.2.5, by the acquisition of several spike mutations, often shared by other VOCs and VOIs.

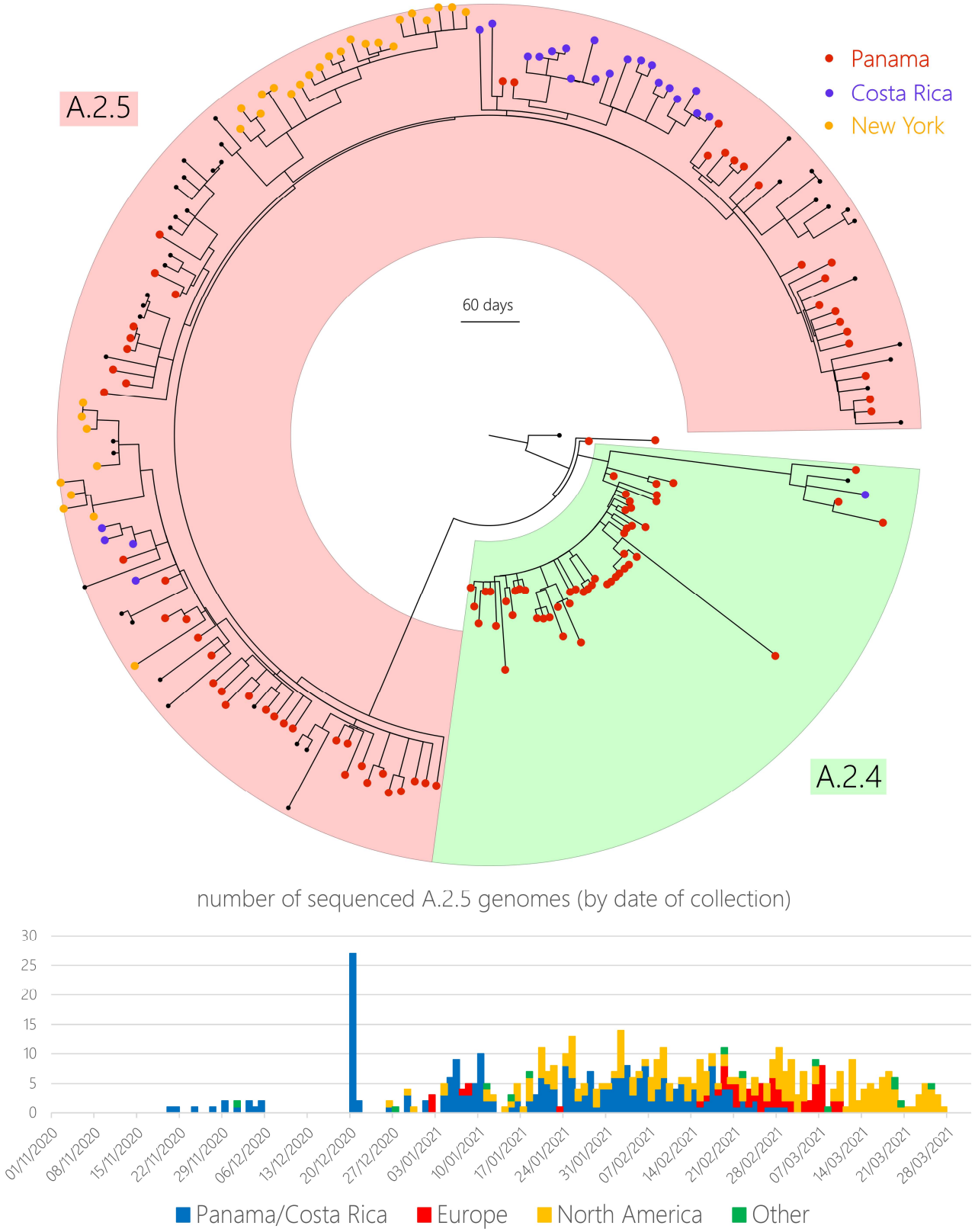

Figure 4. Upper panel: circular time tree exemplifying the phylogeny of the A.2.4 and A.2.5 lineage. Only high quality, complete genomes have been included. Red, blue and orange tip marks indicate samples from Panama, Costa Rica and New York, respectively. The Wuhan-Hu-1 strain was used to root the tree. Lower panel: number of A.2.5 SARS-CoV-2 genomes sequenced in Panama or Costa Rica, Europe, North America and other locations, shown by date of sampling, starting from November 1st 2020, shown by sampling date. Note that the numbers referred to recent dates are unreliable due to delays in data submission.

Context for the emergence of B.1.214.2

B.1.214.2 is linked with the parent lineage B.1.214, which spread in the early phases of the pandemics in the Democratic Republic of Congo and possibly in other neighboring countries where limited genomic surveillance had been carried out in the first half of 2020. As a result of exportations, four distinct sister clades later spread, with different success, in Europe between December 2020 and the first months of 2021: B.1.214.1, originally reported in Congo [49], led to a few cases in France. The few dozen genomes belonging to the B.1.214.3 lineage have been sequenced in France, Belgium, Switzerland, Spain and Mayotte. With the exception of a few cases in Portugal, B.1.214.4 spread in Denmark, where its frequency quickly declining when B.1.1.7 quickly became dominant.

Out of the four sister lineages, B.1.214.2 is, by far, the most widespread (**Figure 5**), with Belgium being the leading county in terms of genome counts. Following the first detection by molecular surveillance, dated January 3rd 2021 in Liège, this variant underwent a rapid increase in frequency, accounting for 5-6% of all the genomes sequenced in the country between week 7 and week 9, which was followed by a moderate decrease to 3-4% between week 10 and week 12. This slight drop in observed genome counts appears to be concomitant with the rise of B.1.1.7 (which, as of April 10th 2021 accounts for nearly 80% of all analyzed PCR-positive collected swabs). However, molecular phylogeny clearly reveals that the genotypes currently circulating in Belgium belong to two distinct sublineages, whose monophyly is currently poorly supported, most likely due to missing genome data concerning the earliest phases of the emergence of this lineage (**Figure 5**).

The precise geographical origins of B.1.214.2 can be only speculated, and it is not possible to establish with certainty whether its early spread in some European countries was linked with an endogenous origin, or with importation from an unknown country where the variant might have been not detected due to insufficient surveillance. The very first cases of B.1.214.2 in Europe were detected in late November 2020 in Switzerland, and more specifically in the Basel-Landschaft. This introduction led to continued community transmission, as evidenced by the large monophyletic cluster of genotypes visible in the phylogenetic tree (**Figure 5**). The circulation of this variant continues to be significant in the Basel region, with an estimated frequency of 8.5% among the genomes sequenced in March 2021.

The major European clusters of community transmission clearly led to the exportation of cases abroad, either in neighboring regions (such as Rhineland-Palatinate in Germany or the Netherlands). As of April 10th 2021, B.1.214.2 have been reported in France, Germany, Netherlands, Ireland, Portugal and United Kingdom. While the cumulative number of known cases is still low, the genetic relatedness among the isolates from Germany and France suggest the presence of early community transmission.

Evidence of spread of this variant in other continents has been, so far, limited. A single case is known from the African continent (Senegal) and two cases have been reported in Asia (Philippines on January 3rd 2021, and Indonesia on March 1st 2021). Several B.1.215.2 genomes have been recently sequenced across the United States, mostly in Maryland and Virginia, indicating multiple introductions and possible early signs of ongoing community transmission.

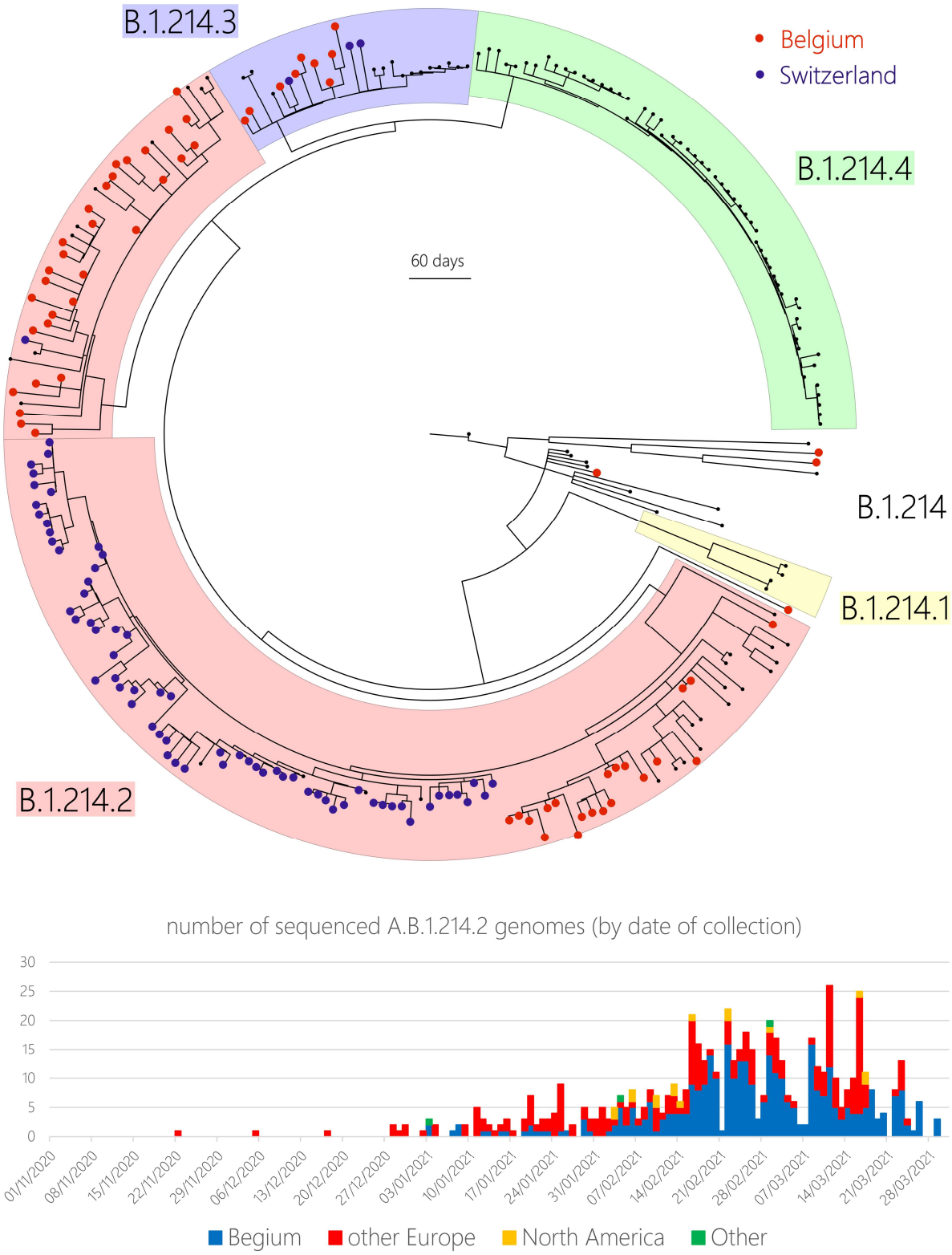

Figure 5. Upper panel: circular time tree exemplifying the phylogeny of the B.1.214.2 lineage, of the parental lineage B.1.214, and of the sister lineages B.1.214.1, B.1.214.3 and B.1.214.4. Only high quality, complete genomes have been included. Red and blue tip marks indicate samples from Belgium and Switzerland, respectively. The Wuhan-Hu-1 strain was used to root the tree. Lower panel: number of B.1.214.2 SARS-CoV-2 genomes sequenced in Belgium, other European countries, North America and other locations, shown by date of sampling, starting from November 1st 2020. Note that the numbers referred to recent dates are unreliable due to delays in data submission.

Further evolutionary considerations

RDR1 is located in a loop which connects the NTD β strands 15 and 16, a region which, unlike RDR1, 2, 3 and 4, does not show any overlap with known major NTD antigenic sites (Cerutti et al., 2021; McCallum et al., 2021). Hence, the involvement of the insertions reported in this manuscript in antibody escape is unlikely, even though the possibility that this modification may lead to paired structural alterations at distantly related sites, leading to a reduced surface accessibility of canonical antibody epitopes cannot be ruled out. Moreover, this region does not show the presence of recurrent indels in the pairwise comparisons with other closely related bat coronaviruses (data not shown), suggesting that RIR1 has not been prone to structural alterations during the radiation of bat coronaviruses. Nevertheless, the spread of the A.2.5 and B.1.214.2 lineages in different geographical contexts suggests that RIR1 insertions are unlikely to have a detrimental impact on the three-dimensional structure of the spike protein or to dramatically reduce the infectivity of these variants. At the same time, the well-defined length of the insertions (either 3 or 4 codons) suggests that some critical structural constraints, that may prevent the selection of shorter insertions or limit their associated evolutionary benefits, might exist.

Detailed structural modeling studies should be carried out to predict which consequences, if any, RIR1 insertions have on the structural organization of the spike protein, on the stability of the interaction between the RBD and the ACE2 receptor, on the accessibility of distant sites to antibody recognition and on the efficiency of the proteolytic cleavage. For instance, the NTD His69/Val70 deletion, which, like RIR1, is found in multiple independent lineages, does not determine a significant antibody escape *in vitro* [22]. However, it is thought to have an important impact on the structure of the spike protein, by conferring increased cleavage at the S1/S2 site, thereby allowing higher incorporation in the virion (Kemp et al., 2021). In light of these observations, some NTD indels apparently not related with immune escape may act as permissive mutations, by compensating small infectivity deficits associated with other RBD mutations (i.e. L452R in A.2.5 and Q414K and N450K in B.1.214.2). The emergence of RIR1 insertions in multiple independent lineages is most certainly suggestive of convergent evolution, and their recent occurrence in some VOCs (B.1.1.7, B.1.429 and P.1) indicates that similar insertion events are very likely to be documented in other emerging VOIs the next few months.

Root-to-tip genetic distance regression analyses indicate that A.2.5 accumulated a number of nucleotide substitutions slightly higher that the majority of the other circulating SARS-CoV-2 strains, as evidenced by the deviation from average rate of genome-wide molecular evolution, estimated to be 8.25×10^{-4} substitutions/site/year (based on GISAID data, data retrieved on April 10th 2021). Unfortunately, the lack of genomic data from the second half of 2020 currently prevents to estimate whether this occurred in a progressive manner or abruptly, as previously evidenced for B.1.1.7 and P.1 [24] (**Figure 6A**). On the other hand, the overall rate of genomic evolution of B.1.214.2 was in line with the global evolutionary trends (**Figure 6B**).

Nevertheless, as previously mentioned, both lineages have disproportionately accumulated a relatively large number of non-synonymous mutations in the S gene, which include, besides the insertions at RIR1, also a four-codon deletion in the case of A.2.5 (**Figure 3**). As a matter of fact, compared with the reference Wuhan Hu-1 genotype, both major RIR1 variants display a total number of substituted, deleted or inserted amino acids in the spike protein, which is in line and often superior to that of several other VOCs and VOIs. In detail, the average number of substitutions/insertions/deletions calculated for A.2.5 (i.e. 9, **Figure 6**, detailed in **Table 1**) is only second to the VOCs B.1.1.7, B.1.351 and P.1, and slightly lower than the emerging lineage P.3. Curiously, this rate of molecular evolution appears to be much higher than the VOCs B.1.427 and B.1.429 [44], and superior to that of other recently described emerging lineage A variants, such as A.23.1 and A.27 [12]. The total number of substitutions/insertions/deletions found in B.1.214.2, albeit slightly inferior to A.2.5 (i.e. ~7, see **Table 1**), was still relatively high, in line with most other emerging VOIs and higher than other

widespread lineages which are not linked with increased transmissibility or immune evasion properties (exemplified by B.1.177 in **Figure 6**).

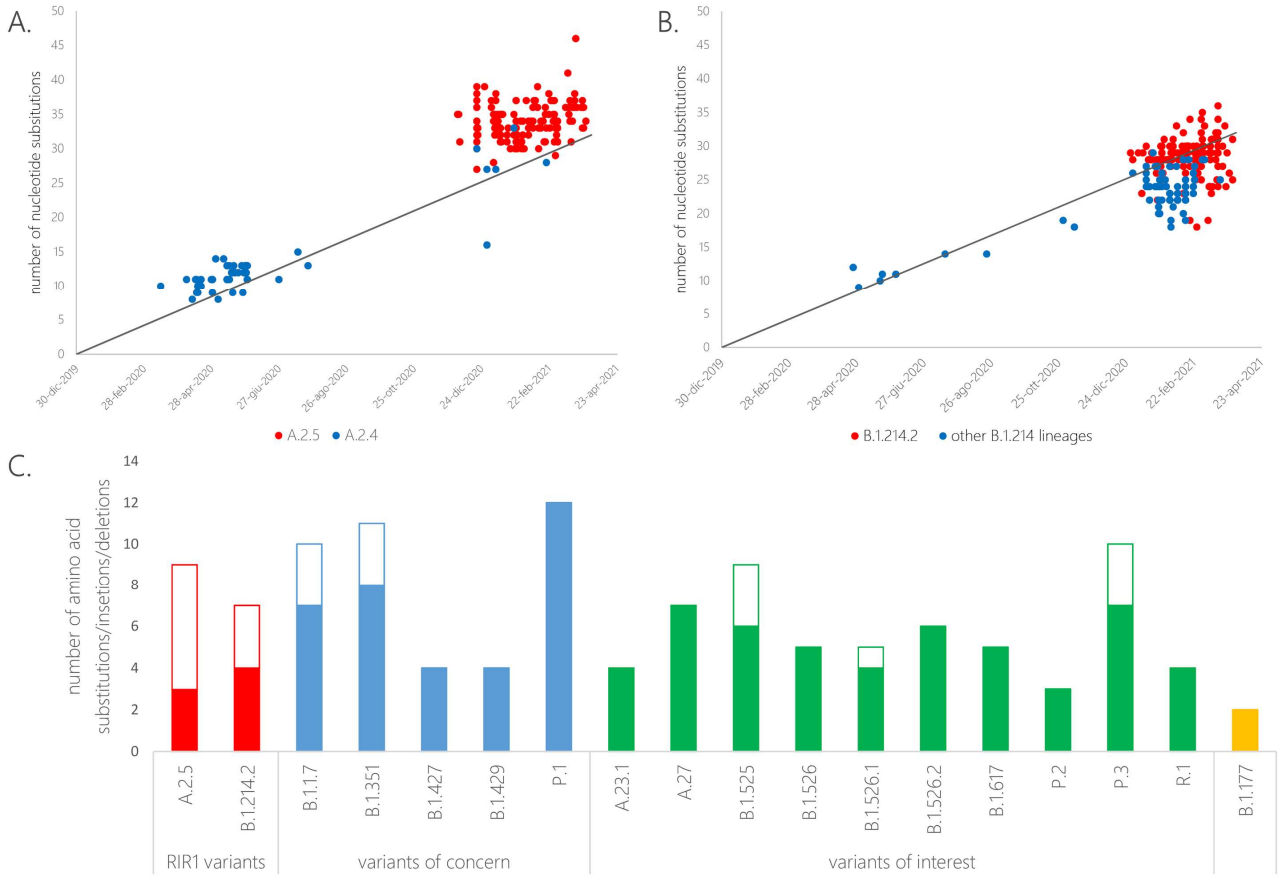

Figure 6. Panels A and B: root-to-tip genetic distance (number of nucleotide substitutions) of the genomes belonging to the A.2.5 and B.1.214.2 lineages, respectively, compared with parent and sister lineages. The line represents the average rate of mutation of all SARS-CoV-2 sequenced genomes, according to GISAID (i.e. 24 substitutions per genome per year, as of April 10th 2021). Note that insertions and deletions were excluded from this calculation. Panel C: mean number of amino acid substitutions (full bars), insertions and deletions (empty bars) observed in the spike protein of the two major variants with insertions at RIR1, the five VOCs and a number of VOIs. The lineage B.1.177 has been included as a non-VOC/VOI reference.

Technical issues in the detection of insertions in SARS-CoV-2 genomes

Several different bioinformatics tools are routinely used for variant calling and SARS-CoV-2 genome assembly. The mostly depend on the sequencing method used and with the preference of different research groups with consolidated pipelines of analysis, whose performance is usually reliable over a broad range of sequencing outputs. However, most of these methods are based on the mapping of sequencing reads to a reference genome, and rely on the analysis of mapping files for variant calling [53]. The management of gaps (i.e. insertions and deletions) may vary from algorithm to algorithm, and the management of insertions, in particular may be problematic, leading to laboratory-specific biases. As a matter of fact, the raw sequencing reads generated with the Oxford Nanopore technology (i.e. MinION, GridION or PromethION platforms) are expected to present a relatively high error rate, in particular for what concerns the presence of artificial indels in homopolymeric sequence stretches. Usually, the presence of such sequencing errors can be overcome thanks to the high sequencing coverage obtained, that allows a consensus-based variant calling [54]. Hence,

some of the algorithms specifically designed and used to carry out such tasks are not insertion-aware, i.e. they explicitly disregard any insertion detected in the alignment between the reads and the reference genome as a probable sequencing error.

This issue was particularly evident for both A.2.5 and B.1.214.2, since of relevant number of GISAID entries lacked the expected lineage-defining insertions at RIR1, despite the shared ancestry of all genotypes (**Figures 4 and 5**). In detail, as of April 10th 2021, only 61% of the A.2.5 genomes carry S:ins214AAG, and only 73% of the B.1.214.2 genomes carry S:ins214TDR. These fraction of deposited genomes lacking insertions at RIR1 remains very high even if we only take into account complete, high quality genomes (i.e. 16% and 43% for A.2.5 and B.1.214.2, respectively), indicating that these artefacts are not linked with low sequencing coverage. As a striking example, just 17 out of the 131 B.214.2 genomes sequenced in Switzerland correctly report the insertion at RIR1. These have been submitted to GISAID by the university hospitals of Basel and Geneva, which most likely use in their routine genome analyses insertion-aware tools. On the other hand, all the B.214.2 genomes sequenced in Switzerland that lack the TDR insertion were deposited by the same institution, i.e. ETH Zürich, which uses V-pipe [55], a tool that disregards the possibility of insertions compared with the reference genome sequence.

Clearly, similar issues may affect other SARS-CoV-2 lineages carrying insertion mutations, not just in the S gene, but potentially also in any genomic regions. Hence, taking into consideration the still limited, but global spread of A.2.5 and B.1.214.2, as well as the insertions occurring at RIR1 here documented in other relevant widespread viral lineages, the use of insertion-aware genome analysis pipelines should be encouraged to allow an improved monitoring of novel insertion mutants that may otherwise go unnoticed. Similarly, existing tools that currently do not include insertion-aware variant calling, should be implemented by the developers to allow the discrimination between *bona fide* insertions and those linked with sequencing errors. While insertions have so far just occurred in a very limited subset of circulating genotypes, they may be linked with phenotypic alterations, both in terms of increase immune escape or virulence, as previously demonstrated for deletions [22,52]. In conclusion, it is the opinion of the author that more attention should be directed towards a close monitoring of the emergence of these still rare but puzzling mutations, especially whenever they are associated with lineages that have already been previously identified as VOCs or VOIs.

References

1. Denison, M.R.; Graham, R.L.; Donaldson, E.F.; Eckerle, L.D.; Baric, R.S. Coronaviruses: An RNA Proofreading Machine Regulates Replication Fidelity and Diversity. *RNA Biol.* **2011**, *8*, 270–279, doi:10.4161/rna.8.2.15013.
2. Ma, Y.; Wu, L.; Shaw, N.; Gao, Y.; Wang, J.; Sun, Y.; Lou, Z.; Yan, L.; Zhang, R.; Rao, Z. Structural Basis and Functional Analysis of the SARS Coronavirus Nsp14–Nsp10 Complex. *Proc. Natl. Acad. Sci.* **2015**, *112*, 9436–9441, doi:10.1073/pnas.1508686112.
3. Shu, Y.; McCauley, J. GISAID: Global Initiative on Sharing All Influenza Data – from Vision to Reality. *Eurosurveillance* **2017**, *22*, doi:10.2807/1560-7917.ES.2017.22.13.30494.
4. Ren, L.; Zhang, Y.; Li, J.; Xiao, Y.; Zhang, J.; Wang, Y.; Chen, L.; Paranhos-Baccalà, G.; Wang, J. Genetic Drift of Human Coronavirus OC43 Spike Gene during Adaptive Evolution. *Sci. Rep.* **2015**, *5*, doi:10.1038/srep11451.
5. Boni, M.F.; Lemey, P.; Jiang, X.; Lam, T.T.-Y.; Perry, B.W.; Castoe, T.A.; Rambaut, A.; Robertson, D.L. Evolutionary Origins of the SARS-CoV-2 Sarbecovirus Lineage Responsible for the COVID-19 Pandemic. *Nat. Microbiol.* **2020**, *5*, 1408–1417, doi:10.1038/s41564-020-0771-4.
6. Ignatieva, A.; Hein, J.; Jenkins, P.A. Evidence of Ongoing Recombination in SARS-CoV-2 through Genealogical Reconstruction. *bioRxiv* **2021**, 2021.01.21.427579, doi:10.1101/2021.01.21.427579.
7. Millet, J.K.; Jaimes, J.A.; Whittaker, G.R. Molecular Diversity of Coronavirus Host Cell Entry Receptors. *FEMS Microbiol. Rev.* **2020**, doi:10.1093/femsre/fuaa057.

8. Walls, A.C.; Park, Y.-J.; Tortorici, M.A.; Wall, A.; McGuire, A.T.; Veesler, D. Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. *Cell* **2020**, *181*, 281-292.e6, doi:10.1016/j.cell.2020.02.058.
9. Ren, W.; Qu, X.; Li, W.; Han, Z.; Yu, M.; Zhou, P.; Zhang, S.-Y.; Wang, L.-F.; Deng, H.; Shi, Z. Difference in Receptor Usage between Severe Acute Respiratory Syndrome (SARS) Coronavirus and SARS-Like Coronavirus of Bat Origin. *J. Virol.* **2008**, *82*, 1899–1907, doi:10.1128/JVI.01085-07.
10. Korber, B.; Fischer, W.M.; Gnanakaran, S.; Yoon, H.; Theiler, J.; Abfalterer, W.; Hengartner, N.; Giorgi, E.E.; Bhattacharya, T.; Foley, B.; et al. Tracking Changes in SARS-CoV-2 Spike: Evidence That D614G Increases Infectivity of the COVID-19 Virus. *Cell* **2020**, *182*, 812-827.e19, doi:10.1016/j.cell.2020.06.043.
11. Zhang, L.; Jackson, C.B.; Mou, H.; Ojha, A.; Peng, H.; Quinlan, B.D.; Rangarajan, E.S.; Pan, A.; Vanderheiden, A.; Suthar, M.S.; et al. SARS-CoV-2 Spike-Protein D614G Mutation Increases Virion Spike Density and Infectivity. *Nat. Commun.* **2020**, *11*, 6013, doi:10.1038/s41467-020-19808-4.
12. Bugembe, D.L.; Phan, M.V.T.; Ssewanyana, I.; Semanda, P.; Nansumba, H.; Dhaala, B.; Nabadda, S.; O'Toole, Á.N.; Rambaut, A.; Kaleebu, P.; et al. A SARS-CoV-2 Lineage A Variant (A.23.1) with Altered Spike Has Emerged and Is Dominating the Current Uganda Epidemic. *medRxiv* **2021**, 2021.02.08.21251393, doi:10.1101/2021.02.08.21251393.
13. van Dorp, L.; Acman, M.; Richard, D.; Shaw, L.P.; Ford, C.E.; Ormond, L.; Owen, C.J.; Pang, J.; Tan, C.C.S.; Boshier, F.A.T.; et al. Emergence of Genomic Diversity and Recurrent Mutations in SARS-CoV-2. *Infect. Genet. Evol.* **2020**, *83*, 104351, doi:10.1016/j.meegid.2020.104351.
14. van Dorp, L.; Richard, D.; Tan, C.C.S.; Shaw, L.P.; Acman, M.; Balloux, F. No Evidence for Increased Transmissibility from Recurrent Mutations in SARS-CoV-2. *Nat. Commun.* **2020**, *11*, 1–8, doi:10.1038/s41467-020-19818-2.
15. Nelson, G.; Buzko, O.; Spilman, P.; Niazi, K.; Rabizadeh, S.; Soon-Shiong, P. Molecular Dynamic Simulation Reveals E484K Mutation Enhances Spike RBD-ACE2 Affinity and the Combination of E484K, K417N and N501Y Mutations (501Y.V2 Variant) Induces Conformational Change Greater than N501Y Mutant Alone, Potentially Resulting in an Escape Mutant. *bioRxiv* **2021**, 2021.01.13.426558, doi:10.1101/2021.01.13.426558.
16. Teruel, N.; Mailhot, O.; Najmanovich, R.J. Modelling Conformational State Dynamics and Its Role on Infection for SARS-CoV-2 Spike Protein Variants. *bioRxiv* **2020**, 2020.12.16.423118, doi:10.1101/2020.12.16.423118.
17. Davies, N.G.; Abbott, S.; Barnard, R.C.; Jarvis, C.I.; Kucharski, A.J.; Munday, J.D.; Pearson, C.A.B.; Russell, T.W.; Tully, D.C.; Washburne, A.D.; et al. Estimated Transmissibility and Impact of SARS-CoV-2 Lineage B.1.1.7 in England. *Science* **2021**, doi:10.1126/science.abg3055.
18. Lustig, Y.; Nemet, I.; Kliker, L.; Zuckerman, N.; Yishai, R.; Alroy-Preis, S.; Mendelson, E.; Mandelboim, M. Neutralizing Response against Variants after SARS-CoV-2 Infection and One Dose of BNT162b2. *N. Engl. J. Med.* **2021**, *0*, null, doi:10.1056/NEJMc2104036.
19. Wang, G.-L.; Wang, Z.-Y.; Duan, L.-J.; Meng, Q.-C.; Jiang, M.-D.; Cao, J.; Yao, L.; Zhu, K.-L.; Cao, W.-C.; Ma, M.-J. Susceptibility of Circulating SARS-CoV-2 Variants to Neutralization. *N. Engl. J. Med.* **2021**, *0*, null, doi:10.1056/NEJMc2103022.
20. Wang, Z.; Schmidt, F.; Weisblum, Y.; Muecksch, F.; Barnes, C.O.; Finkin, S.; Schaefer-Babajew, D.; Cipolla, M.; Gaebler, C.; Lieberman, J.A.; et al. MRNA Vaccine-Elicited Antibodies to SARS-CoV-2 and Circulating Variants. *bioRxiv* **2021**, 2021.01.15.426911, doi:10.1101/2021.01.15.426911.
21. Xie, X.; Zou, J.; Fontes-Garfias, C.R.; Xia, H.; Swanson, K.A.; Cutler, M.; Cooper, D.; Menachery, V.D.; Weaver, S.; Dormitzer, P.R.; et al. Neutralization of N501Y Mutant SARS-CoV-2 by BNT162b2 Vaccine-Elicited Sera. *bioRxiv* **2021**, 2021.01.07.425740, doi:10.1101/2021.01.07.425740.
22. Wang, P.; Nair, M.S.; Liu, L.; Iketani, S.; Luo, Y.; Guo, Y.; Wang, M.; Yu, J.; Zhang, B.; Kwong, P.D.; et al. Antibody Resistance of SARS-CoV-2 Variants B.1.351 and B.1.1.7. *Nature* **2021**, 1–6, doi:10.1038/s41586-021-03398-2.
23. Sykes, W.; Mhlanga, L.; Swanevelder, R.; Glatt, T.N.; Grebe, E.; Coleman, C.; Pieterson, N.; Cable, R.; Welte, A.; Berg, K. van den; et al. *Prevalence of Anti-SARS-CoV-2 Antibodies among Blood Donors in Northern Cape, KwaZulu-Natal, Eastern Cape, and Free State Provinces of South Africa in January 2021.*; In Review, 2021;

24. Sabino, E.C.; Buss, L.F.; Carvalho, M.P.S.; Prete, C.A.; Crispim, M.A.E.; Fraiji, N.A.; Pereira, R.H.M.; Parag, K.V.; Peixoto, P. da S.; Kraemer, M.U.G.; et al. Resurgence of COVID-19 in Manaus, Brazil, despite High Seroprevalence. *The Lancet* **2021**, *397*, 452–455, doi:10.1016/S0140-6736(21)00183-5.
25. Choi, B.; Choudhary, M.C.; Regan, J.; Sparks, J.A.; Padera, R.F.; Qiu, X.; Solomon, I.H.; Kuo, H.-H.; Boucau, J.; Bowman, K.; et al. Persistence and Evolution of SARS-CoV-2 in an Immunocompromised Host. *N. Engl. J. Med.* **2020**, *383*, 2291–2293, doi:10.1056/NEJMc2031364.
26. Greaney, A.J.; Starr, T.N.; Barnes, C.O.; Weisblum, Y.; Schmidt, F.; Caskey, M.; Gaebler, C.; Cho, A.; Agudelo, M.; Finkin, S.; et al. Mutational Escape from the Polyclonal Antibody Response to SARS-CoV-2 Infection Is Largely Shaped by a Single Class of Antibodies. *bioRxiv* **2021**, 2021.03.17.435863, doi:10.1101/2021.03.17.435863.
27. Starr, T.N.; Greaney, A.J.; Dingens, A.S.; Bloom, J.D. Complete Map of SARS-CoV-2 RBD Mutations That Escape the Monoclonal Antibody LY-CoV555 and Its Cocktail with LY-CoV016. *bioRxiv* **2021**, 2021.02.17.431683, doi:10.1101/2021.02.17.431683.
28. Starr, T.N.; Greaney, A.J.; Hilton, S.K.; Ellis, D.; Crawford, K.H.D.; Dingens, A.S.; Navarro, M.J.; Bowen, J.E.; Tortorici, M.A.; Walls, A.C.; et al. Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding. *Cell* **2020**, *182*, 1295-1310.e20, doi:10.1016/j.cell.2020.08.012.
29. Madhi, S.A.; Baillie, V.; Cutland, C.L.; Voysey, M.; Koen, A.L.; Fairlie, L.; Padayachee, S.D.; Dheda, K.; Barnabas, S.L.; Bhorat, Q.E.; et al. Efficacy of the ChAdOx1 NCoV-19 Covid-19 Vaccine against the B.1.351 Variant. *N. Engl. J. Med.* **2021**, *0*, null, doi:10.1056/NEJMoa2102214.
30. Shen, X.; Tang, H.; Pajon, R.; Smith, G.; Glenn, G.M.; Shi, W.; Korber, B.; Montefiori, D.C. Neutralization of SARS-CoV-2 Variants B.1.429 and B.1.351. *N. Engl. J. Med.* **2021**, *0*, null, doi:10.1056/NEJMc2103740.
31. Zhang, W.; Davis, B.D.; Chen, S.S.; Martinez, J.M.S.; Plummer, J.T.; Vail, E. Emergence of a Novel SARS-CoV-2 Strain in Southern California, USA. *medRxiv* **2021**, 2021.01.18.21249786, doi:10.1101/2021.01.18.21249786.
32. Tegally, H.; Wilkinson, E.; Giovanetti, M.; Iranzadeh, A.; Fonseca, V.; Giandhari, J.; Doolabh, D.; Pillay, S.; San, E.J.; Msomi, N.; et al. Emergence and Rapid Spread of a New Severe Acute Respiratory Syndrome-Related Coronavirus 2 (SARS-CoV-2) Lineage with Multiple Spike Mutations in South Africa. *medRxiv* **2020**, 2020.12.21.20248640, doi:10.1101/2020.12.21.20248640.
33. Brejova, B.; Hodorová, V.; Boršová, K.; Čabanová, V.; Reizigová, L.; Paul, E.D.; Čekan, P.; Klempa, B.; Nosek, J.; Vinař, T. B.1.258Δ, a SARS-CoV-2 Variant with ΔH69/ΔV70 in the Spike Protein Circulating in the Czech Republic and Slovakia - SARS-CoV-2 Coronavirus / NCoV-2019 Genomic Epidemiology Available online: https://virological.org/t/b-1-258-a-sars-cov-2-variant-with-h69-v70-in-the-spike-protein-circulating-in-the-czech-republic-and-slovakia/613 (accessed on 5 April 2021).
34. Lasek-Nesselquist, E.; Lapierre, P.; Schneider, E.; George, K.S.; Pata, J. The Localized Rise of a B.1.526 SARS-CoV-2 Variant Containing an E484K Mutation in New York State. *medRxiv* **2021**, 2021.02.26.21251868, doi:10.1101/2021.02.26.21251868.
35. McCarthy, K.R.; Rennick, L.J.; Nambulli, S.; Robinson-McCarthy, L.R.; Bain, W.G.; Haidar, G.; Duprex, W.P. Recurrent Deletions in the SARS-CoV-2 Spike Glycoprotein Drive Antibody Escape. *Science* **2021**, *371*, 1139–1142, doi:10.1126/science.abf6950.
36. Johnson, B.A.; Xie, X.; Bailey, A.L.; Kalveram, B.; Lokugamage, K.G.; Muruato, A.; Zou, J.; Zhang, X.; Juelich, T.; Smith, J.K.; et al. Loss of Furin Cleavage Site Attenuates SARS-CoV-2 Pathogenesis. *Nature* **2021**, *591*, 293–299, doi:10.1038/s41586-021-03237-4.
37. Ge, X.-Y.; Wang, N.; Zhang, W.; Hu, B.; Li, B.; Zhang, Y.-Z.; Zhou, J.-H.; Luo, C.-M.; Yang, X.-L.; Wu, L.-J.; et al. Coexistence of Multiple Coronaviruses in Several Bat Colonies in an Abandoned Mineshaft. *Virol. Sin.* **2016**, *31*, 31–40, doi:10.1007/s12250-016-3713-9.
38. Zhou, H.; Chen, X.; Hu, T.; Li, J.; Song, H.; Liu, Y.; Wang, P.; Liu, D.; Yang, J.; Holmes, E.C.; et al. A Novel Bat Coronavirus Closely Related to SARS-CoV-2 Contains Natural Insertions at the S1/S2 Cleavage Site of the Spike Protein. *Curr. Biol.* **2020**, *30*, 2196-2203.e3, doi:10.1016/j.cub.2020.05.023.
39. Katoh, K.; Misawa, K.; Kuma, K.; Miyata, T. MAFFT: A Novel Method for Rapid Multiple Sequence Alignment Based on Fast Fourier Transform. *Nucleic Acids Res.* **2002**, *30*, 3059–3066, doi:10.1093/nar/gkf436.

40. Minh, B.Q.; Schmidt, H.A.; Chernomor, O.; Schrempf, D.; Woodhams, M.D.; von Haeseler, A.; Lanfear, R. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol. Biol. Evol.* **2020**, *37*, 1530–1534, doi:10.1093/molbev/msaa015.
41. Kumar, S.; Stecher, G.; Li, M.; Knyaz, C.; Tamura, K. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Mol. Biol. Evol.* **2018**, *35*, 1547–1549, doi:10.1093/molbev/msy096.
42. Kistler, K.E.; Bedford, T. Evidence for Adaptive Evolution in the Receptor-Binding Domain of Seasonal Coronaviruses OC43 and 229e. *eLife* **2021**, *10*, e64509, doi:10.7554/eLife.64509.
43. Volz, E.; Mishra, S.; Chand, M.; Barrett, J.C.; Johnson, R.; Geidelberg, L.; Hinsley, W.R.; Laydon, D.J.; Dabrera, G.; O'Toole, Á.; et al. Assessing Transmissibility of SARS-CoV-2 Lineage B.1.1.7 in England. *Nature* **2021**, 1–17, doi:10.1038/s41586-021-03470-x.
44. Tchesnokova, V.; Kulakesara, H.; Larson, L.; Bowers, V.; Rechkina, E.; Kisiela, D.; Sledneva, Y.; Choudhury, D.; Maslova, I.; Deng, K.; et al. Acquisition of the L452R Mutation in the ACE2-Binding Interface of Spike Protein Triggers Recent Massive Expansion of SARS-Cov-2 Variants. *bioRxiv* **2021**, 2021.02.22.432189, doi:10.1101/2021.02.22.432189.
45. Yin, R.; Guest, J.D.; Taherzadeh, G.; Gowthaman, R.; Mittra, I.; Quackenbush, J.; Pierce, B.G. Structural and Energetic Profiling of SARS-CoV-2 Antibody Recognition and the Impact of Circulating Variants. *bioRxiv* **2021**, 2021.03.21.436311, doi:10.1101/2021.03.21.436311.
46. Franco, D.; Gonzalez, C.; Abrego, L.E.; Carrera, J.-P.; Diaz, Y.; Caicedo, Y.; Moreno, A.; Chavarria, O.; Gondola, J.; Castillo, M.; et al. Early Transmission Dynamics, Spread, and Genomic Characterization of SARS-CoV-2 in Panama. *Emerg. Infect. Dis.* **2020**, *27*, 612–615, doi:10.3201/eid2702.203767.
47. Vavrek, D.; Speroni, L.; Curnow, K.J.; Oberholzer, M.; Moeder, V.; Febbo, P.G. Genomic Surveillance at Scale Is Required to Detect Newly Emerging Strains at an Early Timepoint. *medRxiv* **2021**, 2021.01.12.21249613, doi:10.1101/2021.01.12.21249613.
48. Molina-Mora, J.A.; Cordero-Laurent, E.; Godínez, A.; Calderón, M.; Brenes, H.; Soto-Garita, C.; Pérez-Corrales, C.; Rica, C.-C.C.I. de E.G. del S.-C.-2 C.; Drexler, J.F.; Moreira-Soto, A.; et al. SARS-CoV-2 Genomic Surveillance in Costa Rica: Evidence of a Divergent Population and an Increased Detection of a Spike T1117I Mutation. *bioRxiv* **2020**, 2020.12.21.423850, doi:10.1101/2020.12.21.423850.
49. Ntoumi, F.; Mfoutou Mapanguy, C.C.; Tomazatos, A.; Pallerla, S.R.; Linh, L.T.K.; Casadei, N.; Angelov, A.; Sonnabend, M.; Peter, S.; Kremsner, P.G.; et al. Genomic Surveillance of SARS-CoV-2 in the Republic of Congo. *Int. J. Infect. Dis.* **2021**, *105*, 735–738, doi:10.1016/j.ijid.2021.03.036.
50. Cerutti, G.; Guo, Y.; Zhou, T.; Gorman, J.; Lee, M.; Rapp, M.; Reddem, E.R.; Yu, J.; Bahna, F.; Bimela, J.; et al. Potent SARS-CoV-2 Neutralizing Antibodies Directed against Spike N-Terminal Domain Target a Single Supersite. *Cell Host Microbe* **2021**, doi:10.1016/j.chom.2021.03.005.
51. McCallum, M.; Marco, A.D.; Lempp, F.; Tortorici, M.A.; Pinto, D.; Walls, A.C.; Beltramello, M.; Chen, A.; Liu, Z.; Zatta, F.; et al. N-Terminal Domain Antigenic Mapping Reveals a Site of Vulnerability for SARS-CoV-2. *bioRxiv* **2021**, 2021.01.14.426475, doi:10.1101/2021.01.14.426475.
52. Kemp, S.A.; Meng, B.; Ferriera, I.A.; Datir, R.; Harvey, W.T.; Papa, G.; Lytras, S.; Collier, D.A.; Mohamed, A.; Gallo, G.; et al. Recurrent Emergence and Transmission of a SARS-CoV-2 Spike Deletion H69/V70. *bioRxiv* **2021**, 2020.12.14.422555, doi:10.1101/2020.12.14.422555.
53. Kubik, S.; Marques, A.C.; Xing, X.; Silvery, J.; Bertelli, C.; Maio, F.D.; Pournaras, S.; Burr, T.; Duffourd, Y.; Siemens, H.; et al. Guidelines for Accurate Genotyping of SARS-CoV-2 Using Amplicon-Based Sequencing of Clinical Samples. *bioRxiv* **2020**, 2020.12.01.405738, doi:10.1101/2020.12.01.405738.
54. Bull, R.A.; Adikari, T.N.; Ferguson, J.M.; Hammond, J.M.; Stevanovski, I.; Beukers, A.G.; Naing, Z.; Yeang, M.; Verich, A.; Gamaarachchi, H.; et al. Analytical Validity of Nanopore Sequencing for Rapid SARS-CoV-2 Genome Analysis. *Nat. Commun.* **2020**, *11*, 6272, doi:10.1038/s41467-020-20075-6.
55. Posada-Céspedes, S.; Seifert, D.; Topolsky, I.; Jablonski, K.P.; Metzner, K.J.; Beerenwinkel, N. V-Pipe: A Computational Pipeline for Assessing Viral Genetic Diversity from High-Throughput Data. *Bioinformatics* **2021**, doi:10.1093/bioinformatics/btab015.