

Independent acquisition of short insertions at the RIR1 site in the spike N-terminal domain of the SARS-CoV-2 BA.2 lineage

Samuele Greco¹, Marco Gerdol^{1*}

¹Department of Life Sciences, University of Trieste

*to whom correspondence should be addressed: mgerdol@units.it

Abstract

Although the SARS-CoV-2 variants BA.1 and BA.2 share over 30 non-synonymous substitutions in the spike glycoprotein, they show several unique mutations that were likely acquired after the split between these two major omicron lineages. One of the most intriguing mutations associated with BA.1 is the presence of the inserted tripeptide Glu-Pro-Glu within the N-terminal domain. While the functional implications of this insertion are still unclear, several other SARS-CoV-2 lineages had previously independently acquired similarly short insertions at the very same site, named RIR1. We have previously identified this site, located approximately between codon 212 and codon 216, as a hotspot of insertions, which usually involve small nucleotide sequences including three or four codons.

Here we show that similar insertion events have independently occurred at least 13 times in early 2022 within the BA.2 lineage, being occasionally associated with significant community transmission. One of these omicron sublineages, characterized by a Ser-Gly-Arg insertion in position 212, is responsible of over 2% of all SARS-CoV-2 cases recorded in Denmark, as of early April 2022. Molecular surveillance data highlight a slow but steady growth compared with the parental BA.2 lineage in all Danish regions, suggesting that the RIR1 insertion may confer a selective advantage. We report the identification of other currently circulating BA.2 sublineages showing similar insertions, whose spread should be therefore carefully monitored in the upcoming months.

Introduction

Since its first emergence in human during the Wuhan outbreak at the end of 2019, the genome of SARS-CoV-2 has been acquiring new mutations compared with the Wuhan-Hu-1 reference sequence at a rate close to 9×10^{-4} substitutions/site/year, based on GISAID data (Shu and McCauley, 2017). Nevertheless, in line with observations previously collected in other betacoronaviruses, such mutations do not display a uniform distribution, being disproportionately located within the the spike glycoprotein, which has a key role in regulating viral entry through the interaction with receptors expressed on host cell membranes (Boni et al., 2020; Guo et al., 2020). Most specifically, the overwhelming majority of S gene mutations target the S1 subunit, which includes the receptor-binding domain (RBD), responsible of ACE-2 binding, as well as the N-terminal domain (NTD), which contributes to host cell binding and stabilizes the spike trimer.

In light of its fundamental importance in receptor binding (Hoffmann et al., 2020; Li et al., 2003) and of its high antigenicity (Ju et al., 2020; Walls et al., 2020), the S1 subunit has been characterized by strong positive selection and by a very high rate of non-synonymous substitutions (i.e. nearly 2×10^{-2} amino acid/year) during the first two years of the pandemics. Although coronaviruses have a highly efficient proofreading activity thanks to the activity of the nsp10/nsp14 complex (Ma et al., 2015), this rate of molecular evolution far exceeds those observed in the HA1 subunit of seasonal influenza, which are usually in the range of $1-10 \times 10^{-3}$ substitutions/amino acid/year (based on GISAID data, (Shu and McCauley, 2017)).

The evolution of SARS-CoV-2 in the human population can be briefly summarized as follows. Although all the major lineages circulating worldwide in 2020 as the result of the initial exportation from the Hubei regions were characterized by the presence of a single spike mutation (D614G) compared with the reference strain (Korber et al., 2020), for several months none of them displayed significant fitness advantage over the others. Only at the end of 2020, with relevant community immunity building up in several countries, multiple novel variants of concern (VOCs) started to emerge nearly simultaneously. The alpha, beta and gamma variants, originally described in England, South Africa and Brazil, respectively, quickly replaced all other competing variants in these regions, displaying an apparent strong selective advantage. Curiously, these three early VOCs shared a single spike mutation (N501Y), which most likely plays a key role in stabilizing the spike trimer in an open conformation, enhancing ACE2 binding (Teruel et al., 2021; Zhu et al., 2021). N501Y was paired with several other NTD mutations and, in the case of beta and gamma, with convergent RBD mutations (i.e. K417T/N and E484K) overlapping known antibody epitopes. While the rapid spread of alpha is thought to be due to its higher intrinsic transmissibility (Davies et al., 2021), the mutational patterns of beta and gamma, together with their emergence in populations with high seroprevalence and the outcomes of *in vitro* studies, strongly suggested the acquisition of significant immune escape properties (Harvey et al., 2021; Sabino et al., 2021).

The evolutionary success of these three VOCs and of several locally emerging variant of interest (VOIs) was, however, not long-lived. Indeed, in the spring of 2021, following an initial spread in the Indian subcontinent, the new VOC delta started to quickly outcompete all the other variants, to the point that alpha, beta and gamma were brought on the verge of extinction by the end of the year, with delta accounting for over 98% of all SARS-CoV-2 sequenced genomes. Although delta displayed several S1 mutations (including L452R) that might have contributed to increased immune escape (Planas et al., 2021), the most significant contribution to its higher transmissibility was likely the presence of a single mutation occurring at the S1/S2 cleavage site, P681R, which enhanced the efficiency of proteolytic cleavage (Saito et al., 2021).

On November 25th, 2021, WHO held a press conference announcing that the fifth VOC, named omicron, was quickly spreading in South Africa, showing a concerning pattern of reinfections (Pulliam et al., 2022). Omicron replaced delta with unprecedented speed, becoming the dominant SARS-CoV-2 lineage worldwide by early 2022. Unlike previous variants, which never showed the ability to fully escape the neutralizing activity

of vaccine-elicited or convalescent sera, omicron displayed extremely significant immune evasion properties, to the point that some authors have suggested the necessity to consider this VOC as belonging to a distinct SARS-CoV-2 serotype (Simon-Loriere and Schwartz, 2022). The highly divergent antigenic properties of omicron derive from the presence of a very high number of S1 mutations, several of which have been shown to mediate ACE2 recognition and/or immune escape (Cameroni et al., 2022; Mannar et al., 2022). Curiously, many of these mutations occur at sites subjected to strong purifying selection both in previous SARS-CoV-2 variants and in bat coronaviruses, suggesting that they may cooperatively interact to mitigate slight fitness deficits associated with individual substitution, with a significant impact on the structural and biological properties of the spike trimer (Martin et al., 2022a). The highly unusual mutational pattern of omicron may, to some extent, also explain the preferential use of the TMPRSS2-independent and cathepsin-mediated endosomal cell entry route (Pia and Rowland-Jones, 2022), as well as the lower fusogenicity and altered cell tropism displayed by this variant (Meng et al., 2022). The extreme phylogenetic divergence between omicron and all the other previous SARS-CoV-2 variants raised important questions concerning its origins, which might be either sought in complex selective processes occurring in immunocompromised patients, or in an unidentified animal reservoir (Mallapaty, 2022).

Although omicron includes three lineages, only BA.1 and BA.2 have caused major outbreaks in human, whereas BA.3 has been responsible for less than one thousand documented cases worldwide. Although BA.1 and BA.2 share several S1 mutations, they display a remarkable sequence divergence, which is roughly comparable with the divergence observed among earlier VOCs. One of the most puzzling mutations only found in one out of the two omicron sublineages is a short insertion of three codons, which encode the tripeptide Glu-Pro-Glu at position 215, within the spike NTD of BA.1. Nearly all VOCs and VOIs carry small spike deletions compared with Wuhan-Hu-1, which may either confer enhanced antibody escape or act as permissive mutations, counterbalancing the fitness cost of otherwise deleterious RBD mutations (Meng et al., 2021). These small deletions independently occurred on multiple occasions during SARS-CoV-2 evolution at Recurrent Deletion Regions (RDRs) located within the NTD, marking one of the most significant signatures of convergent evolution documented during the first two years of the pandemics. On the other hand, the occurrence of spike insertions has been much rarer and consequently subjected to far less study. We have previously shown that insertions of short sequence stretches, usually encoding 3 or 4 amino acids, had been independently acquired over 50 times at the very same NTD site, named Recurrent Insertion Region 1 (RIR1). Before the emergence of BA.1, RIR1 insertions have been documented in alpha, gamma and delta, where they did not lead to significant community spread, and with two lineages (A.2.5 and B.1.214.2) that became dominant in a few Central America and Central African countries for a few months during 2021 (Gerdol et al., 2022). Although the functional role of RIR1 insertions is presently unclear, their convergent acquisition by multiple viral lineages most certainly suggests the need for increased monitoring, due to the possibility of associated fitness advantages.

While the frequency of observation of BA.1 is progressively declining as the result of the higher transmissibility of BA.2, we here show that a few BA.2 sublineages have independently acquired RIR1 insertions. Molecular surveillance data suggests that one of these sublineages in particular, hereby named BA.2+ins(L), which carries a Ser-Arg-Gly insertion at position 212 and is relatively widespread in Denmark, may display a moderate growth advantage compared with the parental BA.2 lineage.

Materials and Methods

All SARS-CoV-2 genome sequence data deposited in GISAID (Shu and McCauley, 2017) up to April 11th, 2022, was screened, looking for sequences belonging to the BA.2 lineage and showing insertions between positions 210 and 220 in the S gene. Genomes displaying unidentified amino acids (i.e. X) in such insertions were flagged as the likely product of mis-assembly and discarded. Similarly, genomes assigned to the BA.2 lineage but carrying an “EPE” insertion in position 215, previously described as one of the lineage-defining mutations of the sister lineage BA.1, were also flagged as likely misclassified and were therefore discarded. All resulting genomes were grouped based on the inserted nucleotide sequence, whose exact position and phase of insertion were defined based on the alignment with the reference SARS-CoV-2 genome sequence (Wuhan-Hu-1, NCBI accession ID: NC_045512.2) and a reference BA.2 genome (EPI_ISL_8128502). Following the nomenclature scheme proposed in our previous work (Gerdol et al., 2022), each of these groups was labeled with progressive Roman numerals, following their chronological order of identification, starting from insertion L.

The molecular surveillance data from Denmark was subjected to further analysis, due to the local spread of one of the BA.2 sublineages carrying an insertion at RIR1, i.e. BA.2+ins(L). All genome data and associated metadata, deposited up to April 11th, 2022, were downloaded from GISAID. The share of BA.2 carrying insertion L was estimated, both at a national and at a regional scale (i.e. by separately taking into account Syddanmark, Sjaelland, Nordjylland, Midtjylland and Hovestaden) up to April 1st, 2022. The frequencies of observation of BA.2-ins(L) relative to all BA.2 sequenced genomes were reported the 7-day moving average, with 95% confidence intervals.

The phylogenetic tree was prepared as follows. All available SARS-CoV-2 genome sequence data deposited in GISAID up to April 5th, 2022, paired with metadata, were downloaded. Genomes with low sequencing coverage, those including stretches of undefined nucleotides (i.e. N) longer than 25 and those associated with incomplete collection date were discarded. Separate fasta files were created for viral genomes belonging to the following lineages: alpha; beta; gamma; delta; BA.1; BA.1.1; BA.2 (excluding genomes with RIR1 insertions; BA.2+ins(L) and BA.2+ins(LIII) (see **Table 1** and **Figure 1**); sequences not belonging to any VOC from the year 2020

Particular attention was placed in an additional filtering step, which involved the fasta file including the BA.2 sequences bearing RIR1 insertions, aimed at discarding the sequences that may possibly derive from low

quality assemblies. Briefly, retained sequences had to display all characterizing BA.2 mutations, defined as those present in > 80% of all sequenced genomes, as of April 5th, 2022. A few BA.2+ins(L) sequences displaying a small NSP1 deletion (codons 81-87) were also discarded due to its high frequency of observation in some BA.2 sublineages. Each fasta file was subsequently subsampled to include a representative group of sequences for the aforementioned lineages. In detail the final sequence dataset included: 50 alpha, beta, gamma, delta, BA.1 and BA.1.1 genomes; 150 genomes belonging to non-VOC lineages of 2020; 600 BA.2 genomes lacking RIR1 insertions; all available BA.2+ins(L) and BA.2+ins(LIII) passing quality filtering (i.e. 531 and 182, respectively). The Wuhan-Hu-1 strain genome (NCBI accession ID: NC_045512.2) was added as reference sequence

All the sampled sequences were merged and aligned with mafft v7.490 (Katoh and Standley, 2013), configured for closely related viral genomes as suggested by the authors. Gapped regions corresponding to insertions were not discarded. The best-fitting substitution model for the alignment was estimated with modeltest-ng v0.1.7 (Darriba et al., 2020). The Augur pipeline (Huddlestone et al., 2021) was then used to produce the sampling date-refined phylogenetic tree, with the Wuhan-Hu-1 genome set as a reference for rooting. In order to take into account the monophyletic origin of the insertions a guide tree was used in the tree building step. All scripts and command used in this analysis are available at https://github.com/54mu/SARS-CoV-2_BA.2_RIR1insertions. The tree visualization was obtained with Auspice (Hadfield et al., 2018). Extended methods are available at <https://gist.github.com/54mu/5e858e0f4aa6a7236cf5849e6ec371c7>

Results and discussion

In total, we identified 13 independent events of insertion at RIR1 associated with the BA.2 lineage. These, based on the naming scheme we have previously proposed (Gerdol et al., 2022), were chronologically ordered, from the least to the most recent one and assigned Roman numerals, from insertion L to insertion LXII (**Table 1**).

As all previously described RIR1 insertions, those associated with BA.2 involve the acquisition of either three (seven cases), four (five cases), or five (a single case) codons. However, their position was somewhat different compared with those that arose in other variants throughout 2020 and 2021 (**Figure 1A**). Indeed, four insertions (i.e. LI, LII, LIV and LX) occurred within codon 212 at phase II and four others (i.e. L, LIII, LVII and LVIII) occurred within codon 213 at phase 0. None of the 49 RIR1 insertions described in our previous work was observed in these locations (Gerdol et al., 2022). On the other hand, a single insertion (i.e. LVI) was found within codon 215, at phase II, a placement consistent with previously described RIR1 insertions. The exact location of the four remaining BA.2-associated RIR1 insertions could not be ascertained due to the possibility of multiple alternative sequence alignments with the BA.2 reference.

Due to the unusual placement of such insertions in the spike CDS, the inserted tri-, tetra- or penta-peptides in BA.2 would be usually placed between Leu212 and Val213 (albeit this residue is replaced by Gly in BA.2), and only in two cases between Arg214 and Asp215 (**Figure 1B**). Because of the out-of-frame nature of several insertions, in four cases the residue flanking the insertion at the N-terminal side (i.e. Leu212) was replaced with Phe (insertions LI, LII, LIV and LX) (**Figure 1B**). As of note, due to of the presence of the Val213Gly substitution in BA.2, the placement of most of these insertions by GISAID was often offset by a single amino acid (e.g. insertion L is reported as ins213GRG instead of ins212SGR). Consequently, a non-synonymous substitution was often incorrectly called in position 213 (e.g. Val213Ser instead of Val213Gly for insertion L). Only three out of the 49 RIR1 insertions described in our previous work were associated with a significant community spread, either locally, or globally: insertion III (lineage A.2.5), insertion IV (lineage B.1.214.2) and insertion XLI (lineage) BA.1. The first one became dominant in several Central American countries in early 2021, leading to several smaller clusters of infection abroad. The second one likely spread in central Africa in late 2020, before its importation to Belgium and Switzerland led to a few major European clusters of infection in early 2021. However, the spread of A.2.5 and B.1.214.2 was transient and neither of the two variants ever reached a global distribution. Indeed, over the following months both lineages, which predated the emergence of alpha and delta, progressively disappeared, being outcompeted by fitter variants. On the other hand BA.1, one of the two major omicron sublineages and the third major SARS-CoV-2 lineage bearing an insertion at RIR1, quickly became dominant worldwide, leading to several dozen million cases of infections and rapidly replacing delta in virtually any location.

The other 46 RIR1 insertions described in our previous work remained confined to small clusters counting no more than a few dozen sequenced cases, even though on some occasions (e.g. lineage B.1.639) undocumented community transmission likely continued for several months. Overall, these observations pointed out that the association between RIR1 insertions and the acquisition of increased fitness over previous variants could not be generalized. Due to the lack of structural overlap between RIR1 and important NTD antibody epitopes and in light of the frequent association between these insertions and non-synonymous mutations targeting the RBD, in our interpretation RIR1 insertions could be interpreted as compensatory mutations. As previously hypothesized for example for the H69/V70 deletion found in many VOCs and VOIs (Meng et al., 2021), small insertions at RIR1 may be able to mitigate slight negative fitness costs associated with the presence of non-synonymous RBD mutations affecting immune escape (Gerdol et al., 2022). From this perspective, the limited community spread of most SARS-CoV-2 variants bearing RIR1 insertions was not surprising, due to the important role played by intra-host selection in shaping spike mutational patterns, since these selective forces are not necessarily expected to lead to viral lineages with increased transmissibility. The previously documented acquisition of RIR1 insertions through *in vitro* experiments, both in response to several passages in Vero cell cultures (Shiliaev et al., 2021) and after the treatment with monoclonal antibodies (Committee for Medicinal Products for Human Use, 2021), might be

consistent with this view. Hence, the acquisition of a selective advantage by of RIR1 insertions-carrying viral lineages may be conditioned by the presence or absence of other spike non-synonymous mutations associated with negative fitness costs.

In total, 11 out of the 13 BA.2-associated RIR1 insertions were only associated with a very small number of cases and were therefore unlikely to be linked with BA.2 sublineages with enhanced fitness (**Table 1**), even though it cannot be excluded that some of these, identified very recently, may further spread in the coming weeks. Nevertheless, two exceptions were immediately evident, i.e. insertion L and insertion LIII (**Table 1**). The two BA.2 sublineages carrying such insertions will be hereafter named BA.2+ins(L) and BA.2+ins(LIII). BA.2+ins(L), first detected on January 7th, 2022, is characterized by the presence of an in-frame TCCGGCAGA insertion between codon 212 and 213, which determines the acquisition of the tripeptide Ser-Gly-Arg in position 212 (**Figure 1**). The overwhelming majority of the BA.2 genomes deposited in GISAID carrying this insertion derives from Denmark (96.5%), the country where this sublineage apparently first originated, followed by the United Kingdom (0,9%), Germany (0,7%) and Australia (0.6%). As of April 11th, 2022, a few cases have been also detected in Austria, Belgium, Germany, Israel, Italy, Norway, South Africa, Sweden and Switzerland, likely as the result of exportation from Denmark. A detailed analysis of the dynamics of the spread of the BA.2+ins(L) sublineage in this country revealed an apparent slow but steady increase in frequency of observation compared with the parental BA.2 lineage over the past two months and a half (**Figure 2**). This trend is evident at a national level, where this BA.2 sublineage accounts for approximately 2.5% of all BA.2 cases as of April 1st, 2022 (**Figure 2A**).

The national spread of BA.2+ins(L) was initially driven by the Nordjylland region, where this sublineage was first detected, undergoing a significant expansion relative to BA.2 from mid-January, up to now (**Figure 2B**). Based on the most recent genomic surveillance data, as of April 1st, 2022, BA.2+ins(L) may account for approximately 8.5% of all BA.2 cases recorded in the region. Midtjylland, the only Denmark region with land borders with Nordjylland, followed a similar trend: after the detection of the first cases in late-January, the share of BA.2+ins(L) genomes progressively increased over time and accounts for about 2.8% of the all BA.2 cases as of April 1st, 2022 (**Figure 2C**). Syddanmark, the southernmost continental region in Denmark, and Hovedstaden, the easternmost and capital region, displayed nearly overlapping trends. Albeit the first few cases were detected in mid-January, the frequency of observation of this sublineage relative to BA.2 started to grow in February, and the share of BA.2 genomes carrying the RIR1 insertion L in the two regions on April 1st, 2022 is approximately 0.9% and 2%, respectively (**Figure 2D and 2E**). This insertion was reported for the first time in Sjælland, the largest insular region, only on February 12th, 2022. As of April 1st, 2022, BA.2+ins(L) accounts for approximately 1% of all BA.2 cases recorded in this region (**Figure 2F**).

At a national level, the comparative analysis of the incidence of infections due to BA.2 and BA.2+ins(L) allows to draw a preliminary estimate of the weekly growth advantage displayed by the latter, which would be in the range of 20%. Although the absolute number of infections linked with this sublineage presently remains

relatively low (i.e. with the number of daily national cases recently dropping below 5,000, those linked with BA.2+ins(L) might be in the range of 50-100 per day), this moderate growth trend has been continuing for over two months, following consistent patterns in multiple Danish regions. It would be therefore important to clarify whether this apparent growth advantage over the parental BA.2 lineage is driven by an alteration of the structure and function of the spike protein or rather by other epidemiological factors, such as founder effects, as previously observed for other SARS-CoV-2 lineages in the past (Hodcroft et al., 2021).

The second relevant BA.2-related sublineage with an insertion at RIR1 was BA.2+ins(LIII). This sublineage, first detected on February 5th, 2022 in England, displays a longer insertion compared with BA.2+ins(L), i.e. ACAGTAGGAGGA, which is also found in-frame between codon 212 and 213, and encodes the tetrapeptide Thr-Val-Gly-Gly (**Figure 1**). With 290 sequenced genomes as of April 11th, 2022, insertion LIII is currently ranked the fifth most successful RIR1 insertion, after insertion III (A.2.5), IV (B.1.214.2), XLI (BA.1) and the aforementioned insertion L. During the first two months of its spread, BA.2+ins(LIII) has been almost exclusively detected in the UK (with 278 out of 281 cases being identified in England), with the exception of two single cases identified in Belgium and Israel, respectively. Although still linked with a very small absolute number of cases, this sublineage is currently displaying a very slight but constant growth over time and has been responsible for 0.2% of all cases reported in England in the second half of March 2022.

Curiously, neither these two insertions, nor the other 11 listed in **Table 1** were associated with other relevant S1 non-synonymous substitutions, questioning their previously hypothesized “compensatory” role in stabilizing otherwise slightly deleterious mutations with significant effect on immune escape (Gerdol et al., 2022). In detail, BA.2+ins(L) did not show any other associated non-synonymous mutations, neither in the S gene, nor in other genomic regions, leaving the Ser-Gly-Arg as the only sublineage-defining mutation present. The only other relevant mutation shared by all BA.2+ins(L) genomes was synonymous, i.e. C22791T (also located within the S gene), was not phylogenetically informative, as it was found in nearly one third of all BA.2 genomes. On the other hand, BA.2+ins(LIII) also displayed the presence of Thr1213Ser, a S2 non-synonymous conservative substitution rarely observed in SARS-CoV-2 prior to the emergence of BA.2 (only 62 cases had been recorded prior to December 2021), but occasionally found in BA.2. The only other two spike non-synonymous mutations linked with BA.2 RIR1 insertions were G75D in insertion LI, Y144del in insertion LX and T1136S in insertion LXII, even though several point mutations were found in other SARS-CoV-2 genes (**Table 1**).

In summary, most BA.2 associated RIR1 insertions appear to have not occurred in parallel with other mutations targeting the RBD, the polybasic site or known important epitopes recognized by neutralizing antibodies. The two major omicron lineages share a large fraction of the non-synonymous substitutions found in the S gene (i.e. 21), but each of the two also carry a dozen additional mutations not shared with the sister lineage. Some of the mutations shared by BA.1 and BA.2, if taken individually, would have been predicted to significantly decrease viral fitness. Nevertheless, they are thought to cooperatively interact by

positive epistasis, mitigating the negative fitness costs associated with their presence and possibly altering some biological properties of the spike protein (Martin et al., 2022b). Insertion XLI (i.e. ins215EPE), which is included among the mutations found in BA.1 but not shared with BA.2, undoubtedly emerged after the split between the two lineages and its presence may therefore provide a fitness advantage in combination with any of the other mutations exclusively found in BA.1. Nevertheless, we did not observe any overlap between the presence of RIR1 insertions in BA.2 and the acquisition, by convergent evolution, of any spike mutation typical of BA.1. A different interpretation would therefore be needed to explain the emergence of RIR1 insertions in BA.2, as well as the apparent slight fitness advantage associated with BA.2+ins(L). While, to date, no conclusive explanation has been provided concerning the benefits associated with the presence of ins214EPE in BA.1, it is very likely that this small insertion, in combination with other NTD non-synonymous mutations, had a significant impact on the structure of the spike protein in BA.1. These alterations to the original primary sequence of the SARS-CoV-2 spike protein might have possibly contributed to the enhanced inter-domain and inter-subunit packing, as well as to the higher accessibility for ACE2 binding, driven by its higher predisposition towards an open configuration (Ye et al., 2022). From this perspective, the acquisition of RIR1 insertions such as L, LIII and others, might be also interpreted as a trait conferring increased stability to the BA.2 spike trimer. Hence, the independent acquisition of different short sequence stretches at RIR1 in the BA.2 lineage might simply depend on the presence of pre-existing BA.2-associated mutations and be completely independent on the acquisition of additional non-synonymous mutations. Undoubtedly, further structural studies would be required to clarify the effects of the acquisition of RIR1 insertions on a BA.2 genetic background in terms of the impact on the structural conformation and stability of spike trimers.

A relevant consequence of the lack of other lineage-defining mutations in most BA.2 sublineages carrying RIR1 insertions is the inherent difficulty in their assignment through phylogenetic inference. As a matter of fact, indels are usually simply treated as missing characters and not appropriately weighted-in by most maximum likelihood approaches, including those currently used by the *agur* pipeline, i.e. IQ-TREE (Nguyen et al., 2015) and fasttree (Price et al., 2010), and this might lead to significant statistical inconsistencies, as previously noted by other authors (McTavish et al., 2015; Warnow, 2012). Here we briefly report the relationships between BA.2+ins(L) and BA.2+ins(LIII) with BA.2. Despite their clear monophyletic origin, when no guide tree was provided to the maximum likelihood phylogenetic inference analysis, BA.2 genomes carrying insertion L were scattered across the BA.2 subtree (data not shown). This inconsistent placement was possibly due to a combination between the secondary acquisition, by convergent evolution, of other synonymous (or, more rarely, non-synonymous) substitutions relatively common in other BA.2 sublineages by BA.2+ins(L), or by the presence of reversions. Only when a guide tree was provided and all sequences not showing the full mutational profile of BA.2 were discarded, all BA.2+ins(L) were correctly placed in a monophyletic clade (**Figure 3**). On the other hand, this adjustment was not required to achieve a good support of the monophyly of the BA.2+ins(LIII) sublineage, which as mentioned above also carries the spike

Thr1213Ser mutation. This highlights the fact that similar difficulties in future lineage assignments could arise for other SARS-CoV-2 lineages whose only distinctive mutations compared with their parental variant are either insertions or deletions.

Tables and Figures

Designation	Insertion	Insertion type	GISAID entries	Other mutations not shared with BA.2	Earliest detection	Latest detection*
L	212:SGR	in frame (codon 213 phase 0)	1447	/	Jan 7th, 2022	Apr 5th, 2022
LI	212:NNR	out-of-frame (codon 212 phase 2)	2	Spike_G75D, Spike_L214F, NS3_L140F, NSP1_G82del, NSP1_H83del, NSP1_M85V, NSP1_V84del, NSP4_A146V	Feb 2nd, 2022	Feb 4th, 2022
LII	212:ASG	out-of-frame (codon 212 phase 2)	1	Spike_L214F, NSP4_T28I	Feb 3rd, 2022	Feb 3rd, 2022
LIII	212:TVGG	in frame (codon 213 phase 0)	290	Spike_L1231S	Feb 5th, 2022	Apr 4nd, 2022
LIV	212:GG	out-of-frame (codon 212 phase 2)	3	Spike_L214F, NSP13_S36P	Feb 8th, 2022	Feb 21st, 2022
LV	211:LTPT/212:TPTL	ambiguous; in frame (codon 212 phase 0/codon 213 phase 0)	31	/	Feb 19th, 2022	Apr 2nd, 2022
LVI	212:MAEL	out-of-frame (codon 212 phase 2)	9	NSP1_P80S, NSP2_T573I, NS7a_P99L, NS3_L140F, NSP4_A146V, Spike_Y144del, Spike_L214F	Mar 4th, 2022	Mar 31st, 2022
LVII	214:EIGT	out-of-frame (codon 215 phase 2)	1	NS3_L108F, NSP14_L157F	Mar 5th, 2022	Mar 5th, 2022
LVIII	212:DRG	in frame (codon 213 phase 0)	1	NSP3_A1561V, NSP3_T1482I, NSP12_M633T, NSP12_A585S	Mar 10th, 2022	Mar 10th, 2022
LIX	212:TGNTL	in frame (codon 213 phase 0)	11	E_R61C	Mar 14th, 2022	Mar 23rd, 2022
LX	214:DLE	ambiguous; out-of-frame (codon 215 phase 2/codon 216 phase 1)	4	N_G30del, N_G34del, N_N29del, NSP3_L1266I, NSP3_S1265del	Mar 15th, 2022	Mar 15th, 2022
LXI	211:NL/212:NLL	ambiguous; in frame (codon 212 phase 0/codon 213 phase 0)	1	M_R146C	Mar 22nd, 2022	Mar 22nd, 2022
LXII	212:GHVN	ambiguous (codon 213 phase 0/phase I/phase II)	1	Spike_T1136S, NSP1_M85del, NSP1_V86del	unknown	unknown

Table 1: Summary of the 13 independent RIR1 insertions identified in the SARS-CoV-2 BA.2 lineage. Latest detection, as of April 11th, 2022.

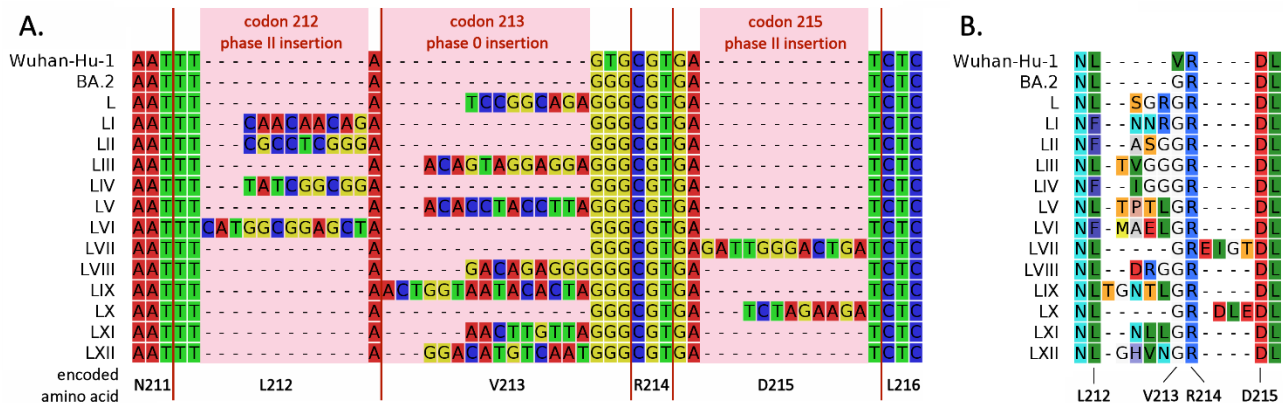


Figure 1: Panel A: Multiple sequence alignment of the nucleotide sequences of the S gene of viral genomes belonging to the BA.2 lineage bearing insertions at RIR1, compared with the reference sequences of Wuhan-Hu-1 and BA.2. The multiple sequence alignment only shown a small region of the S gene (i.e. codon 211-codon 216). Red vertical bars highlight codon boundaries. Panel B: amino acid sequences encoded by the nucleotide sequences displayed in panel A. Codon/residue numbering refers to the Wuhan-Hu-1 reference sequence.

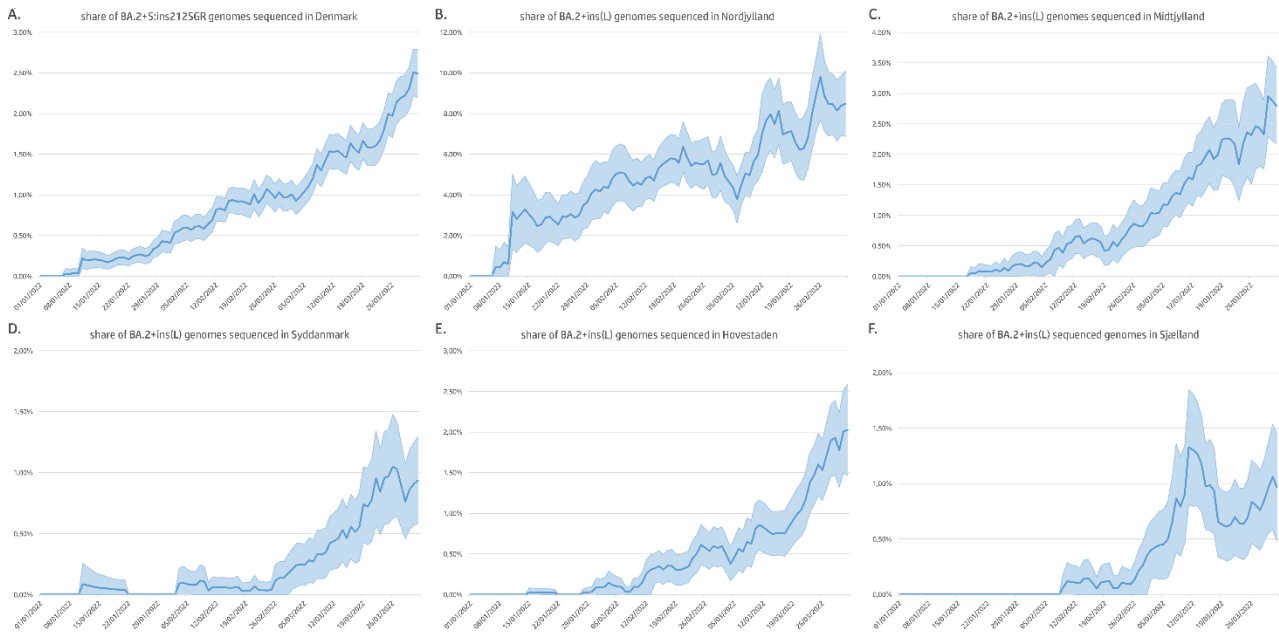


Figure 2: Share of BA.2 genomes bearing the L insertion at RIR1 in Denmark (panel A), with a regional breakdown for Nordjylland (panel B), Midtjylland (panel C), Syddanmark (panel D), Sjælland (panel E) and Hovestaden (panel F). The graphs report the 7-day moving average frequencies of observations of BA.2+ins(L) genomes, relative to all BA.2 genomes, with 95% confidence intervals.

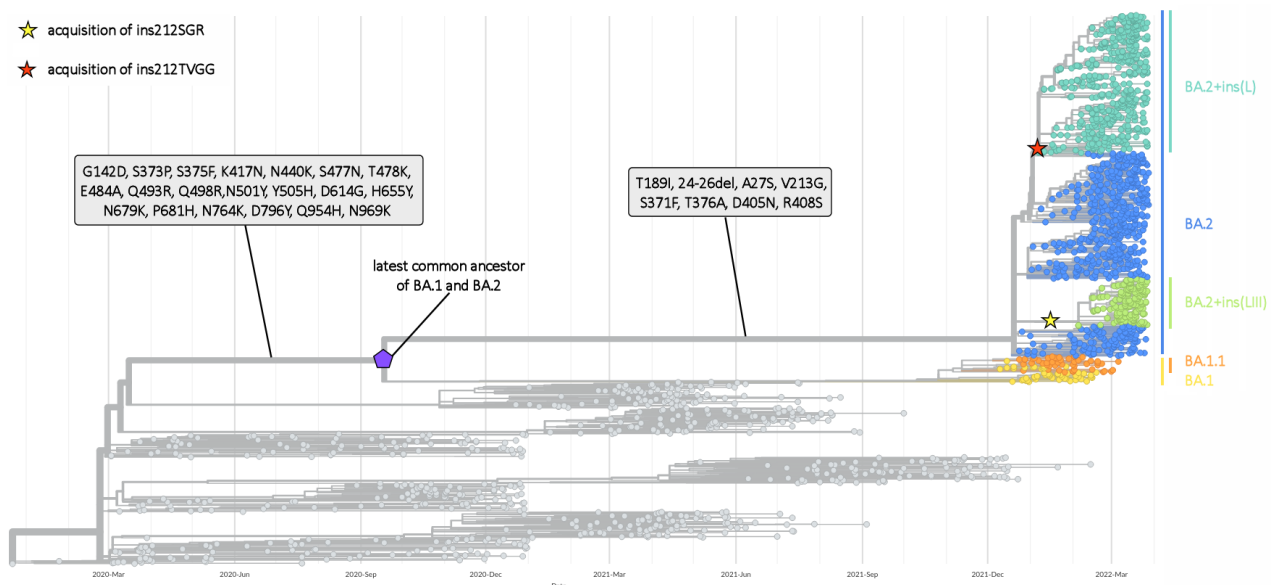


Figure 3: Time tree summarizing SARS-CoV-2 phylogeny and highlighting the placement of BA2+ins(L) and BA2+ins(L)III relative with BA.1 and the other BA.2 sublineages. The acquisition of lineage-defining spike mutations in reported in grey boxes. The acquisition of insertions at RIR1 are marked by stars. The interactive tree can be visualized at https://nextstrain.org/community/54mu/SARS-CoV-2_BA.2_RIR1insertions/treeViz?c=BA_group.

Declaration of Competing Interest

The authors declare they have no competing interests.

Acknowledgements

The authors would like to acknowledge the tremendous efforts made by the clinicians, researchers and public health authorities that allowed the collection of SARS-CoV-2 genome data and made sequence data available in a timely manner through GISAID, as well as the great efforts made by the developers of nextstrain to assist researchers in SARS-CoV-2 evolution studies.

Funding Statement

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

- Boni, M.F., Lemey, P., Jiang, X., Lam, T.T.-Y., Perry, B.W., Castoe, T.A., Rambaut, A., Robertson, D.L., 2020. Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic. *Nature Microbiology* 5, 1408–1417. <https://doi.org/10.1038/s41564-020-0771-4>
- Cameroni, E., Bowen, J.E., Rosen, L.E., Saliba, C., Zepeda, S.K., Culap, K., Pinto, D., VanBlargan, L.A., De Marco, A., di Iulio, J., Zatta, F., Kaiser, H., Noack, J., Farhat, N., Czudnochowski, N., Havenar-Daughton, C., Sprouse, K.R., Dillen, J.R., Powell, A.E., Chen, A., Maher, C., Yin, L., Sun, D., Soriaga, L., Bassi, J., Silacci-Fregni, C., Gustafsson, C., Franko, N.M., Logue, J., Iqbal, N.T., Mazzitelli, I., Geffner, J., Grifantini, R., Chu, H., Gori, A., Riva, A., Giannini, O., Ceschi, A., Ferrari, P., Cippà, P.E., Franzetti-Pellanda, A., Garzoni, C., Halfmann, P.J., Kawaoka, Y., Hebner, C., Purcell, L.A., Piccoli, L., Pizzuto, M.S., Walls, A.C., Diamond, M.S., Telenti, A., Virgin, H.W., Lanzavecchia, A., Snell, G., Vesler, D., Corti, D., 2022. Broadly neutralizing antibodies overcome SARS-CoV-2 Omicron antigenic shift. *Nature* 602, 664–670. <https://doi.org/10.1038/s41586-021-04386-2>
- Committee for Medicinal Products for Human Use, 2021. GlaxoSmithKline use of sotrovimab (VIR-7831/GSK4182136) for the treatment of COVID-19 (No. EMA/304600/2021). European Medicines Agency.
- Darriba, D., Posada, D., Kozlov, A.M., Stamatakis, A., Morel, B., Flouri, T., 2020. ModelTest-NG: A New and Scalable Tool for the Selection of DNA and Protein Evolutionary Models. *Mol Biol Evol* 37, 291–294. <https://doi.org/10.1093/molbev/msz189>
- 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., Wenseleers, T., Gimma, A., Waites, W., Wong, K.L.M., van Zandvoort, K., Silverman, J.D., CMMID COVID-19 Working Group, COVID-19 Genomics UK (COG-UK) Consortium, Diaz-Ordaz, K., Keogh, R., Eggo, R.M., Funk, S., Jit, M., Atkins, K.E., Edmunds, W.J., 2021. Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England. *Science* 372, eabg3055. <https://doi.org/10.1126/science.abg3055>
- Gerdol, M., Dishnica, K., Giorgetti, A., 2022. Emergence of a recurrent insertion in the N-terminal domain of the SARS-CoV-2 spike glycoprotein. *Virus Research* 310, 198674. <https://doi.org/10.1016/j.virusres.2022.198674>
- Guo, H., Hu, B.-J., Yang, X.-L., Zeng, L.-P., Li, B., Ouyang, S., Shi, Z.-L., 2020. Evolutionary Arms Race between Virus and Host Drives Genetic Diversity in Bat Severe Acute Respiratory Syndrome-Related Coronavirus Spike Genes. *J Virol* 94, e00902-20. <https://doi.org/10.1128/JVI.00902-20>

- Hadfield, J., Megill, C., Bell, S.M., Huddleston, J., Potter, B., Callender, C., Sagulenko, P., Bedford, T., Neher, R.A., 2018. Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* 34, 4121–4123. <https://doi.org/10.1093/bioinformatics/bty407>
- Harvey, W.T., Carabelli, A.M., Jackson, B., Gupta, R.K., Thomson, E.C., Harrison, E.M., Ludden, C., Reeve, R., Rambaut, A., Peacock, S.J., Robertson, D.L., 2021. SARS-CoV-2 variants, spike mutations and immune escape. *Nat Rev Microbiol* 19, 409–424. <https://doi.org/10.1038/s41579-021-00573-0>
- Hodcroft, E.B., Zuber, M., Nadeau, S., Vaughan, T.G., Crawford, K.H.D., Althaus, C.L., Reichmuth, M.L., Bowen, J.E., Walls, A.C., Corti, D., Bloom, J.D., Velesler, D., Mateo, D., Hernando, A., Comas, I., González-Candelas, F., Stadler, T., Neher, R.A., 2021. Spread of a SARS-CoV-2 variant through Europe in the summer of 2020. *Nature* 595, 707–712. <https://doi.org/10.1038/s41586-021-03677-y>
- Hoffmann, M., Kleine-Weber, H., Schroeder, S., Krüger, N., Herrler, T., Erichsen, S., Schiergens, T.S., Herrler, G., Wu, N.-H., Nitsche, A., Müller, M.A., Drosten, C., Pöhlmann, S., 2020. SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. *Cell* 181, 271-280.e8. <https://doi.org/10.1016/j.cell.2020.02.052>
- Huddleston, J., Hadfield, J., Sibley, T.R., Lee, J., Fay, K., Ilcisin, M., Harkins, E., Bedford, T., Neher, R.A., Hodcroft, E.B., 2021. Augur: a bioinformatics toolkit for phylogenetic analyses of human pathogens. *Journal of Open Source Software* 6, 2906. <https://doi.org/10.21105/joss.02906>
- Ju, B., Zhang, Q., Ge, J., Wang, R., Sun, J., Ge, X., Yu, Jiazhen, Shan, S., Zhou, B., Song, S., Tang, X., Yu, Jinfang, Lan, J., Yuan, J., Wang, H., Zhao, Juanjuan, Zhang, S., Wang, Y., Shi, X., Liu, L., Zhao, Jincun, Wang, X., Zhang, Z., Zhang, L., 2020. Human neutralizing antibodies elicited by SARS-CoV-2 infection. *Nature* 584, 115–119. <https://doi.org/10.1038/s41586-020-2380-z>
- Katoh, K., Standley, D.M., 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* 30, 772–780. <https://doi.org/10.1093/molbev/mst010>
- Korber, B., Fischer, W.M., Gnanakaran, S., Yoon, H., Theiler, J., Abfalterer, W., Hengartner, N., Giorgi, E.E., Bhattacharya, T., Foley, B., Hastie, K.M., Parker, M.D., Partridge, D.G., Evans, C.M., Freeman, T.M., de Silva, T.I., Angyal, A., Brown, R.L., Carrilero, L., Green, L.R., Groves, D.C., Johnson, K.J., Keeley, A.J., Lindsey, B.B., Parsons, P.J., Raza, M., Rowland-Jones, S., Smith, N., Tucker, R.M., Wang, D., Wyles, M.D., McDanal, C., Perez, L.G., Tang, H., Moon-Walker, A., Whelan, S.P., LaBranche, C.C., Saphire, E.O., Montefiori, D.C., 2020. Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus. *Cell* 182, 812-827.e19. <https://doi.org/10.1016/j.cell.2020.06.043>
- Li, W., Moore, M.J., Vasilieva, N., Sui, J., Wong, S.K., Berne, M.A., Somasundaran, M., Sullivan, J.L., Luzuriaga, K., Greenough, T.C., Choe, H., Farzan, M., 2003. Angiotensin-converting enzyme 2 is a functional receptor for the SARS coronavirus. *Nature* 426, 450–454. <https://doi.org/10.1038/nature02145>
- Ma, Y., Wu, L., Shaw, N., Gao, Y., Wang, J., Sun, Y., Lou, Z., Yan, L., Zhang, R., Rao, Z., 2015. Structural basis and functional analysis of the SARS coronavirus nsp14–nsp10 complex. *PNAS* 112, 9436–9441. <https://doi.org/10.1073/pnas.1508686112>
- Mallapaty, S., 2022. Where did Omicron come from? Three key theories. *Nature* 602, 26–28. <https://doi.org/10.1038/d41586-022-00215-2>
- Mannar, D., Saville, J.W., Zhu, X., Srivastava, S.S., Berezuk, A.M., Tuttle, K.S., Marquez, A.C., Sekirov, I., Subramaniam, S., 2022. SARS-CoV-2 Omicron variant: Antibody evasion and cryo-EM structure of spike protein-ACE2 complex. *Science* 375, 760–764. <https://doi.org/10.1126/science.abn7760>
- Martin, D.P., Lytras, S., Lucaci, A.G., Maier, W., Grüning, B., Shank, S.D., Weaver, S., MacLean, O.A., Orton, R.J., Lemey, P., Boni, M.F., Tegally, H., Harkins, G., Scheepers, C., Bhiman, J.N., Everatt, J., Amoako, D.G., San, J.E., Giandhari, J., Sigal, A., NGS-SA, Williamson, C., Hsiao, N.-Y., von Gottberg, A., De Klerk, A., Shafer, R.W., Robertson, D.L., Wilkinson, R.J., Sewell, B.T., Lessells, R., Nekrutenko, A., Greaney, A.J., Starr, T.N., Bloom, J.D., Murrell, B., Wilkinson, E., Gupta, R.K., de Oliveira, T., Kosakovsky Pond, S.L., 2022a. Selection analysis identifies unusual clustered mutational changes in Omicron lineage BA.1 that likely impact Spike function. *bioRxiv* 2022.01.14.476382. <https://doi.org/10.1101/2022.01.14.476382>
- Martin, D.P., Lytras, S., Lucaci, A.G., Maier, W., Grüning, B., Shank, S.D., Weaver, S., MacLean, O.A., Orton, R.J., Lemey, P., Boni, M.F., Tegally, H., Harkins, G.W., Scheepers, C., Bhiman, J.N., Everatt, J., Amoako, D.G., San, J.E., Giandhari, J., Sigal, A., NGS-SA, Williamson, C., Hsiao, N., von Gottberg, A., De Klerk, A., Shafer, R.W., Robertson, D.L., Wilkinson, R.J., Sewell, B.T., Lessells, R., Nekrutenko, A., Greaney,

- A.J., Starr, T.N., Bloom, J.D., Murrell, B., Wilkinson, E., Gupta, R.K., de Oliveira, T., Kosakovsky Pond, S.L., 2022b. Selection analysis identifies clusters of unusual mutational changes in Omicron lineage BA.1 that likely impact Spike function. *Mol Biol Evol* msac061. <https://doi.org/10.1093/molbev/msac061>
- McTavish, E.J., Steel, M., Holder, M.T., 2015. Twisted trees and inconsistency of tree estimation when gaps are treated as missing data – The impact of model mis-specification in distance corrections. *Molecular Phylogenetics and Evolution* 93, 289–295. <https://doi.org/10.1016/j.ympev.2015.07.027>
- Meng, B., Abdullahi, A., Ferreira, I.A.T.M., Goonawardane, N., Saito, A., Kimura, I., Yamasoba, D., Gerber, P.P., Fathi, S., Rathore, S., Zepeda, S.K., Papa, G., Kemp, S.A., Ikeda, T., Toyoda, M., Tan, T.S., Kuramochi, J., Mitsunaga, S., Ueno, T., Shirakawa, K., Takaori-Kondo, A., Brevini, T., Mallery, D.L., Charles, O.J., CITIID-NIHR BioResource COVID-19 Collaboration, Genotype to Phenotype Japan (G2P-Japan) Consortium, Ecuador-COVID19 Consortium, Bowen, J.E., Joshi, A., Walls, A.C., Jackson, L., Martin, D., Smith, K.G.C., Bradley, J., Briggs, J.A.G., Choi, J., Madisson, E., Meyer, K.B., Mlcochova, P., Ceron-Gutierrez, L., Doffinger, R., Teichmann, S.A., Fisher, A.J., Pizzuto, M.S., de Marco, A., Corti, D., Hosmillo, M., Lee, J.H., James, L.C., Thukral, L., Veessler, D., Sigal, A., Sampaziotis, F., Goodfellow, I.G., Matheson, N.J., Sato, K., Gupta, R.K., 2022. Altered TMPRSS2 usage by SARS-CoV-2 Omicron impacts infectivity and fusogenicity. *Nature* 603, 706–714. <https://doi.org/10.1038/s41586-022-04474-x>
- Meng, B., Kemp, S.A., Papa, G., Datir, R., Ferreira, I.A.T.M., Marelli, S., Harvey, W.T., Lytras, S., Mohamed, A., Gallo, G., Thakur, N., Collier, D.A., Mlcochova, P., Robson, S.C., Loman, N.J., Connor, T.R., Golubchik, T., Martinez Nunez, R.T., Ludden, C., Corden, S., Johnston, I., Bonsall, D., Smith, C.P., Awan, A.R., Bucca, G., Torok, M.E., Saeed, K., Prieto, J.A., Jackson, D.K., Hamilton, W.L., Snell, L.B., Moore, Catherine, Harrison, E.M., Goncalves, S., Fairley, D.J., Loose, M.W., Watkins, J., Livett, R., Moses, S., Amato, R., Nicholls, S., Bull, M., Smith, D.L., Barrett, Jeff, Aanensen, D.M., Curran, M.D., Parmar, S., Aggarwal, D., Shepherd, J.G., Parker, M.D., Glaysher, S., Bashton, M., Underwood, A.P., Pacchiarini, N., Loveson, K.F., Templeton, K.E., Langford, C.F., Sillitoe, J., de Silva, T.I., Wang, D., Kwiatkowski, D., Rambaut, A., O’Grady, J., Cottrell, S., Holden, M.T.G., Thomson, E.C., Osman, H., Andersson, M., Chauhan, A.J., Hassan-Ibrahim, M.O., Lawniczak, M., Alderton, A., Chand, M., Constantinidou, C., Unnikrishnan, M., Darby, A.C., Hiscox, J.A., Paterson, S., Martincorena, I., Volz, E.M., Page, A.J., Pybus, O.G., Bassett, A.R., Ariani, C.V., Chapman, M.H.S., Li, K.K., Shah, R.N., Jesudason, N.G., Taha, Y., McHugh, M.P., Dewar, R., Jahun, A.S., McMurray, C., Pandey, S., McKenna, J.P., Nelson, A., Young, G.R., McCann, C.M., Elliott, S., Lowe, H., Temperton, B., Roy, S., Price, A., Rey, S., Wyles, M., Rooke, S., Shaaban, S., de Cesare, M., Letchford, L., Silveira, S., Pelosi, E., Wilson-Davies, E., Hosmillo, M., O’Toole, Á., Hesketh, A.R., Stark, R., du Plessis, L., Ruis, C., Adams, H., Bourgeois, Y., Michell, S.L., Gramatopoulos, D., Edgeworth, J., Breuer, J., Todd, J.A., Fraser, C., Buck, D., John, M., Kay, G.L., Palmer, Steve, Peacock, S.J., Heyburn, D., Weldon, D., Robinson, E., McNally, A., Muir, P., Vipond, I.B., Boyes, J., Sivaprakasam, V., Salluja, T., Dervisevic, S., Meader, E.J., Park, N.R., Oliver, K., Jeffries, A.R., Ott, S., da Silva Filipe, A., Simpson, D.A., Williams, C., Masoli, J.A.H., Knight, B.A., Jones, C.R., Koshy, C., Ash, A., Casey, A., Bosworth, A., Ratcliffe, L., Xu-McCrae, L., Pymont, H.M., Hutchings, S., Berry, Lisa, Jones, K., Halstead, F., Davis, T., Holmes, C., Iturriza-Gomara, M., Lucaci, A.O., Randell, P.A., Cox, A., Madona, P., Harris, K.A., Brown, J.R., Mahungu, T.W., Irish-Tavares, D., Haque, T., Hart, J., Witele, E., Fenton, M.L., Liggett, S., Graham, C., Swindells, E., Collins, J., Eltringham, G., Campbell, S., McClure, P.C., Clark, G., Sloan, T.J., Jones, C., Lynch, J., Warne, B., Leonard, S., Durham, J., Williams, T., Haldenby, S.T., Storey, N., Alikhan, N.-F., Holmes, N., Moore, Christopher, Carlile, M., Perry, M., Craine, N., Lyons, R.A., Beckett, A.H., Goudarzi, S., Fearn, C., Cook, K., Dent, H., Paul, H., Davies, R., Blane, B., Girgis, S.T., Beale, M.A., Bellis, K.L., Dorman, M.J., Drury, E., Kane, L., Kay, S., McGuigan, S., Nelson, R., Prestwood, L., Rajatileka, S., Batra, R., Williams, R.J., Kristiansen, M., Green, A., Justice, A., Mahanama, A.I.K., Samaraweera, B., Hadjirin, N.F., Quick, J., Poplawski, R., Kermack, L.M., Reynolds, N., Hall, G., Chaudhry, Y., Pinckert, M.L., Georgana, I., Moll, R.J., Thornton, A., Myers, R., Stockton, J., Williams, C.A., Yew, W.C., Trotter, A.J., Trebes, A., MacIntyre-Cockett, G., Birchley, A., Adams, A., Plimmer, A., Gatica-Wilcox, B., McKerr, C., Hilvers, E., Jones, H., Asad, H., Coombes, J., Evans, J.M., Fina, L., Gilbert, L., Graham, L., Cronin, M., Kumziene-Summerhayes, S., Taylor, Sarah, Jones, S., Groves, D.C., Zhang, P., Gallis, M., Louka, S.F., Starinskij, I., Jackson, C., Gourtovaia, M., Tonkin-Hill, G., Lewis, K., Tovar-Corona, J.M., James, K., Baxter, L., Alam, M.T., Orton, R.J., Hughes, J.,

Vattipally, S., Ragonnet-Cronin, M., Nascimento, F.F., Jorgensen, D., Boyd, O., Geidelberg, L., Zarebski, A.E., Raghwan, J., Kraemer, M.U.G., Southgate, J., Lindsey, B.B., Freeman, T.M., Keatley, J.-P., Singer, J.B., de Oliveira Martins, L., Yeats, C.A., Abudahab, K., Taylor, B.E.W., Menegazzo, M., Danesh, J., Hogsden, W., Eldirdiri, S., Kenyon, A., Mason, J., Robinson, T.I., Holmes, A., Price, J., Hartley, J.A., Curran, T., Mather, A.E., Shankar, G., Jones, R., Howe, R., Morgan, S., Wastenge, E., Chapman, M.R., Mookerjee, S., Stanley, Rachael, Smith, W., Peto, T., Eyre, D., Crook, D., Vernet, G., Kitchen, C., Gulliver, H., Merrick, I., Guest, M., Munn, R., Bradley, D.T., Wyatt, T., Beaver, C., Foulser, L., Palmer, Sophie, Churcher, C.M., Brooks, E., Smith, K.S., Galai, K., McManus, G.M., Bolt, F., Coll, F., Meadows, L., Attwood, S.W., Davies, A., De Lacy, E., Downing, F., Edwards, S., Scarlett, G.P., Jeremiah, S., Smith, N., Leek, D., Sridhar, S., Forrest, S., Cormie, C., Gill, H.K., Dias, J., Higginson, E.E., Maes, M., Young, J., Wantoch, M., Jamrozy, D., Lo, S., Patel, M., Hill, V., Bewshea, C.M., Ellard, S., Auckland, C., Harrison, I., Bishop, C., Chalker, V., Richter, A., Beggs, A., Best, A., Percival, B., Mirza, J., Megram, O., Mayhew, M., Crawford, L., Ashcroft, F., Moles-Garcia, E., Cumley, N., Hopes, R., Asamaphan, P., Niebel, M.O., Gunson, R.N., Bradley, A., Maclean, A., Mollett, G., Blacow, R., Bird, P., Helmer, T., Fallon, K., Tang, J., Hale, A.D., Macfarlane-Smith, L.R., Harper, K.L., Carden, H., Machin, N.W., Jackson, K.A., Ahmad, S.S.Y., George, R.P., Turtle, L., O'Toole, E., Watts, J., Breen, C., Cowell, A., Alcolea-Medina, A., Charalampous, T., Patel, A., Levett, L.J., Heaney, J., Rowan, A., Taylor, G.P., Shah, D., Atkinson, L., Lee, J.C.D., Westhorpe, A.P., Jannoo, R., Lowe, H.L., Karamani, A., Ensell, L., Chatterton, W., Pusok, M., Dadrah, A., Symmonds, A., Sluga, G., Molnar, Z., Baker, P., Bonner, S., Essex, S., Barton, E., Padgett, D., Scott, G., Greenaway, J., Payne, B.A.I., Burton-Fanning, S., Waugh, S., Raviprakash, V., Sheriff, N., Blakey, V., Williams, L.-A., Moore, J., Stonehouse, S., Smith, L., Davidson, R.K., Bedford, L., Coupland, L., Wright, V., Chappell, J.G., Tsoleridis, T., Ball, J., Khakh, M., Fleming, V.M., Lister, M.M., Howson-Wells, H.C., Berry, Louise, Boswell, T., Joseph, A., Willingham, I., Duckworth, N., Walsh, S., Wise, E., Moore, N., Mori, M., Cortes, N., Kidd, S., Williams, R., Gifford, L., Bicknell, K., Wyllie, S., Lloyd, A., Impey, R., Malone, C.S., Cogger, B.J., Levene, N., Monaghan, L., Keeley, A.J., Partridge, D.G., Raza, M., Evans, C., Johnson, K., Betteridge, E., Farr, B.W., Goodwin, S., Quail, M.A., Scott, C., Shirley, L., Thurston, S.A.J., Rajan, D., Bronner, I.F., Aigrain, L., Redshaw, N.M., Lensing, S.V., McCarthy, S., Makunin, Alex, Balcazar, C.E., Gallagher, M.D., Williamson, K.A., Stanton, T.D., Michelsen, M.L., Warwick-Dugdale, J., Manley, R., Farbos, A., Harrison, J.W., Sambles, C.M., Studholme, D.J., Lackenby, A., Mbisa, T., Platt, S., Miah, S., Bibby, D., Manso, C., Hubb, J., Dabrera, G., Ramsay, M., Bradshaw, D., Schaefer, U., Groves, N., Gallagher, E., Lee, D., Williams, D., Ellaby, N., Hartman, H., Manesis, N., Patel, V., Ledesma, J., Twohig, K.A., Allara, E., Pearson, C., Cheng, J.K.J., Bridgewater, H.E., Frost, L.R., Taylor-Joyce, G., Brown, P.E., Tong, L., Broos, A., Mair, D., Nichols, J., Carmichael, S.N., Smollett, K.L., Nomikou, K., Aranday-Cortes, E., Johnson, N., Nickbakhsh, S., Vamos, E.E., Hughes, M., Rainbow, L., Eccles, R., Nelson, C., Whitehead, M., Gregory, R., Gemmell, M., Wierzbicki, C., Webster, H.J., Fisher, C.L., Signell, A.W., Betancor, G., Wilson, H.D., Nebbia, G., Flaviani, F., Cerda, A.C., Merrill, T.V., Wilson, R.E., Cotic, M., Bayzid, N., Thompson, T., Acheson, E., Rushton, S., O'Brien, S., Baker, D.J., Rudder, S., Aydin, A., Sang, F., Debebe, J., Francois, S., Vasylyeva, T.I., Zamudio, M.E., Gutierrez, B., Marchbank, A., Maksimovic, J., Spellman, K., McCluggage, K., Morgan, M., Beer, R., Afifi, S., Workman, T., Fuller, W., Bresner, C., Angyal, A., Green, L.R., Parsons, P.J., Tucker, R.M., Brown, R., Whiteley, M., Bonfield, J., Puethe, C., Whitwham, A., Liddle, Jennifer, Rowe, W., Siveroni, I., Le-Viet, T., Gaskin, A., Johnson, R., Abnizova, I., Ali, M., Allen, L., Anderson, R., Ariani, C., Austin-Guest, S., Bala, S., Barrett, Jeffrey, Bassett, A., Battleday, K., Beal, J., Beale, M., Bellany, S., Bellerby, T., Bellis, K., Berger, D., Berriman, M., Bevan, P., Binley, S., Bishop, J., Blackburn, K., Boughton, N., Bowker, S., Brendler-Spaeth, T., Bronner, I., Brooklyn, T., Buddenborg, S.K., Bush, R., Caetano, C., Cagan, A., Carter, N., Cartwright, J., Monteiro, T.C., Chapman, L., Chillingworth, T.-J., Clapham, P., Clark, R., Clarke, A., Clarke, C., Cole, D., Cook, E., Coppola, M., Cornell, L., Cornwell, C., Corton, C., Crackett, A., Cranage, A., Craven, H., Craw, S., Crawford, M., Cutts, T., Dabrowska, M., Davies, M., Dawson, J., Day, C., Densem, A., Dibling, T., Dockree, C., Dodd, D., Dogga, S., Dorman, M., Dougan, G., Dougherty, M., Dove, A., Drummond, L., Dudek, M., Durrant, L., Easthope, E., Eckert, S., Ellis, P., Farr, B., Fenton, M., Ferrero, M., Flack, N., Fordham, H., Forsythe, G., Francis, M., Fraser, A., Freeman, A., Galvin, A., Garcia-Casado, M., Gedny, A., Girgis, S., Glover, J., Gould, O., Gray, A., Gray, E., Griffiths, C., Gu, Y., Guerin, F., Hamilton, W., Hanks, H., Harrison, E., Harrott, A., Harry, E., Harvison, J., Heath, P., Hernandez-Koutoucheva, A.,

- Hobbs, R., Holland, D., Holmes, S., Hornett, G., Hough, N., Huckle, L., Hughes-Hallet, L., Hunter, A., Inglis, S., Iqbal, S., Jackson, A., Jackson, D., Verdejo, C.J., Jones, M., Kallepally, K., Kay, K., Keatley, J., Keith, A., King, A., Kitchin, L., Kleanthous, M., Klimekova, M., Korlevic, P., Krasheninikova, K., Lane, G., Langford, C., Laverack, A., Law, K., Lensing, S., Lewis-Wade, A., Liddle, Jennifer, Lin, Q., Lindsay, S., Linsdell, S., Long, R., Lovell, Jamie, Lovell, Jon, Mack, J., Maddison, M., Makunin, Aleksei, Mamun, I., Mansfield, J., Marriott, N., Martin, M., Mayho, M., McClintock, J., McHugh, S., MapcMinn, L., Meadows, C., Mobley, E., Moll, R., Morra, M., Morrow, L., Murie, K., Nash, S., Nathwani, C., Naydenova, P., Neaverson, A., Nerou, E., Nicholson, J., Nimz, T., Noell, G.G., O'Meara, S., Ohan, V., Olney, C., Ormond, D., Oszlanczi, A., Pang, Y.F., Pardubska, B., Park, N., Parmar, A., Patel, G., Payne, M., Peacock, S., Petersen, A., Plowman, D., Preston, T., Quail, M., Rance, R., Rawlings, S., Redshaw, N., Reynolds, J., Reynolds, M., Rice, S., Richardson, M., Roberts, C., Robinson, K., Robinson, M., Robinson, D., Rogers, H., Rojo, E.M., Roopra, D., Rose, M., Rudd, L., Sadri, R., Salmon, N., Saul, D., Schwach, F., Seekings, P., Simms, A., Sinnott, M., Sivadasan, S., Siwek, B., Sizer, D., Skeldon, K., Skelton, J., Slater-Tunstall, J., Sloper, L., Smerdon, N., Smith, Chris, Smith, Christen, Smith, J., Smith, K., Smith, M., Smith, S., Smith, T., Sneade, L., Soria, C.D., Sousa, C., Souster, E., Sparkes, A., Spencer-Chapman, M., Squares, J., Stanley, Robert, Steed, C., Stickland, T., Still, I., Stratton, M., Strickland, M., Swann, A., Swiatkowska, A., Sycamore, N., Swift, E., Symons, E., Szluha, S., Taluy, E., Tao, N., Taylor, K., Taylor, Sam, Thompson, S., Thompson, M., Thomson, M., Thomson, N., Thurston, S., Toombs, D., Topping, B., Tovar-Corona, J., Ungureanu, D., Uphill, J., Urbanova, J., Van, P.J., Vancollie, V., Voak, P., Walker, D., Walker, M., Waller, M., Ward, G., Weatherhogg, C., Webb, N., Wells, A., Wells, E., Westwood, L., Whipp, T., Whiteley, T., Whitton, G., Widaa, S., Williams, M., Wilson, M., Wright, S., Duncan, L.M., Carabelli, A.M., Kenyon, J.C., Lever, A.M., De Marco, A., Saliba, C., Culap, K., Cameroni, E., Matheson, N.J., Piccoli, L., Corti, D., James, L.C., Robertson, D.L., Bailey, D., Gupta, R.K., 2021. Recurrent emergence of SARS-CoV-2 spike deletion H69/V70 and its role in the Alpha variant B.1.1.7. *Cell Reports* 35, 109292. <https://doi.org/10.1016/j.celrep.2021.109292>
- Nguyen, L.-T., Schmidt, H.A., von Haeseler, A., Minh, B.Q., 2015. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Molecular Biology and Evolution* 32, 268–274. <https://doi.org/10.1093/molbev/msu300>
- Pia, L., Rowland-Jones, S., 2022. Omicron entry route. *Nat Rev Immunol* 22, 144–144. <https://doi.org/10.1038/s41577-022-00681-9>
- Planas, D., Veyer, D., Baidaliuk, A., Staropoli, I., Guivel-Benhassine, F., Rajah, M.M., Planchais, C., Porrot, F., Robillard, N., Puech, J., Prot, M., Gallais, F., Gantner, P., Velay, A., Le Guen, J., Kassis-Chikhani, N., Edriss, D., Belec, L., Seve, A., Courtellemont, L., Péré, H., Hocqueloux, L., Fafi-Kremer, S., Prazuck, T., Mouquet, H., Bruel, T., Simon-Lorière, E., Rey, F.A., Schwartz, O., 2021. Reduced sensitivity of SARS-CoV-2 variant Delta to antibody neutralization. *Nature* 596, 276–280. <https://doi.org/10.1038/s41586-021-03777-9>
- Price, M.N., Dehal, P.S., Arkin, A.P., 2010. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLOS ONE* 5, e9490. <https://doi.org/10.1371/journal.pone.0009490>
- Pulliam, J.R.C., van Schalkwyk, C., Govender, N., von Gottberg, A., Cohen, C., Groome, M.J., Dushoff, J., Mlisana, K., Moultrie, H., 2022. Increased risk of SARS-CoV-2 reinfection associated with emergence of Omicron in South Africa. *Science* 0, eabn4947. <https://doi.org/10.1126/science.abn4947>
- 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., Oikawa, M.K., Salomon, T., Cucunuba, Z.M., Castro, M.C., Santos, A.A. de S., Nascimento, V.H., Pereira, H.S., Ferguson, N.M., Pybus, O.G., Kucharski, A., Busch, M.P., Dye, C., Faria, N.R., 2021. Resurgence of COVID-19 in Manaus, Brazil, despite high seroprevalence. *The Lancet* 397, 452–455. [https://doi.org/10.1016/S0140-6736\(21\)00183-5](https://doi.org/10.1016/S0140-6736(21)00183-5)
- Saito, A., Irie, T., Suzuki, R., Maemura, T., Nasser, H., Uriu, K., Kosugi, Y., Shirakawa, K., Sadamasu, K., Kimura, I., Ito, J., Wu, J., Iwatsuki-Horimoto, K., Ito, M., Yamayoshi, S., Loeber, S., Tsuda, M., Wang, L., Ozono, S., Butlertanaka, E.P., Tanaka, Y.L., Shimizu, R., Shimizu, K., Yoshimatsu, K., Kawabata, R., Sakaguchi, T., Tokunaga, K., Yoshida, I., Asakura, H., Nagashima, M., Kazuma, Y., Nomura, R., Horisawa, Y., Yoshimura, K., Takaori-Kondo, A., Imai, M., Tanaka, S., Nakagawa, S., Ikeda, T., Fukuhara, T., Kawaoka, Y., Sato, K., 2021. Enhanced fusogenicity and pathogenicity of SARS-CoV-2 Delta P681R mutation. *Nature* 1–10. <https://doi.org/10.1038/s41586-021-04266-9>

- Shiliaev, N., Lukash, T., Palchevska, O., Crossman, D.K., Green, T.J., Crowley, M.R., Frolova, E.I., Frolov, I., 2021. Natural and Recombinant SARS-CoV-2 Isolates Rapidly Evolve In Vitro to Higher Infectivity through More Efficient Binding to Heparan Sulfate and Reduced S1/S2 Cleavage. *Journal of Virology* 95, e01357-21. <https://doi.org/10.1128/JVI.01357-21>
- Shu, Y., McCauley, J., 2017. GISAID: Global initiative on sharing all influenza data – from vision to reality. *Euro Surveill* 22. <https://doi.org/10.2807/1560-7917.ES.2017.22.13.30494>
- Simon-Loriere, E., Schwartz, O., 2022. Towards SARS-CoV-2 serotypes? *Nat Rev Microbiol* 20, 187–188. <https://doi.org/10.1038/s41579-022-00708-x>
- Teruel, N., Mailhot, O., Najmanovich, R.J., 2021. Modelling conformational state dynamics and its role on infection for SARS-CoV-2 Spike protein variants. *PLOS Computational Biology* 17, e1009286. <https://doi.org/10.1371/journal.pcbi.1009286>
- Walls, A.C., Park, Y.-J., Tortorici, M.A., Wall, A., McGuire, A.T., Velesler, D., 2020. Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. *Cell* 181, 281-292.e6. <https://doi.org/10.1016/j.cell.2020.02.058>
- Warnow, T., 2012. Standard maximum likelihood analyses of alignments with gaps can be statistically inconsistent. *PLoS Curr* 4, RRN1308. <https://doi.org/10.1371/currents.RRN1308>
- Ye, G., Liu, B., Li, F., 2022. Cryo-EM structure of a SARS-CoV-2 omicron spike protein ectodomain. *Nat Commun* 13, 1214. <https://doi.org/10.1038/s41467-022-28882-9>
- Zhu, X., Mannar, D., Srivastava, S.S., Berezuk, A.M., Demers, J.-P., Saville, J.W., Leopold, K., Li, W., Dimitrov, D.S., Tuttle, K.S., Zhou, S., Chittori, S., Subramaniam, S., 2021. Cryo-electron microscopy structures of the N501Y SARS-CoV-2 spike protein in complex with ACE2 and 2 potent neutralizing antibodies. *PLOS Biology* 19, e3001237. <https://doi.org/10.1371/journal.pbio.3001237>