Virus Evolution, 2025, **11(1)**, veaf015
DOI: https://doi.org/10.1093/ve/veaf015
Advance Access Publication 11 March 2025
Research Article

Genome data artifacts and functional studies of deletion repair in the BA.1 SARS-CoV-2 spike protein

Miguel Álvarez-Herrera[1], Paula Ruiz-Rodriguez[1], Beatriz Navarro-Domínguez[1,2], Joao Zulaica[1], Brayan Grau[1], María Alma Bracho[1,3], Manuel Guerreiro[4,5], Cristóbal Aguilar-Gallardo[5], Fernando González-Candelas[1,3], Iñaki Comas[3,6], Ron Geller[1], Mireia Coscollá[1,*]

[1]Institute for Integrative Systems Biology (I[2]SysBio), University of Valencia - Spanish National Research Council (CSIC), FISABIO Joint Research Unit "Infection and Public Health", C/ Catedràtic Agustín Escardino 9, Paterna 46980, Spain
[2]Department of Genetics, Universtiy of Granada, Avenida de la Fuente Nueva, Granada 18071, Spain
[3]CIBER in Epidemiology and Public Health (CIBERESP), Av. Monforte de Lemos 3-5, Madrid 28029, Spain
[4]Department of Haematology, La Fe University and Polytechnic Hospital, Av. Fernando Abril Martorell 106, Valencia 46026, Spain
[5]La Fe Health Research Institute (IIS-La Fe), Av. Fernando Abril Martorell 106, Valencia 46026, Spain
[6]Tuberculosis Genomics Unit, Institute of Biomedicine of Valencia (IBV-CSIC), C/ Jaume Roig 11, Valencia 46010, Spain

*Corresponding author. Institute for Integrative Systems Biology (I²SysBio, University of Valencia-CSIC), C/ Catedràtic Agustín Escardino 9, Building 4, Paterna, 46980, Spain. E-mail: mireia.coscolla@csic.es

Abstract

Mutations within the N-terminal domain (NTD) of the spike (S) protein are critical for the emergence of successful severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) viral lineages. The NTD has been repeatedly impacted by deletions, often exhibiting complex and dynamic patterns, such as the recurrent emergence and disappearance of deletions in dominant variants. This study investigates the influence of repair of NTD lineage-defining deletions found in the BA.1 lineage (Omicron variant) on viral success. We performed comparative genomic analyses of >10 million SARS-CoV-2 genomes from the Global Initiative on Sharing All Influenza Data (GISAID) EpiCov database to evaluate the detection of viruses lacking S:ΔH69/V70, S:ΔV143/Y145, or both. These findings were contrasted against a screening of publicly available raw sequencing data, revealing substantial discrepancies between data repositories, suggesting that spurious deletion repair observations in GISAID may result from systematic artifacts. Specifically, deletion repair events were approximately an order of magnitude less frequent in the read-run survey. Our results suggest that deletion repair events are rare, isolated events with limited direct influence on SARS-CoV-2 evolution or transmission. Nevertheless, such events could facilitate the emergence of fitness-enhancing mutations. To explore potential drivers of NTD deletion repair patterns, we characterized the viral phenotype of such markers in a surrogate *in vitro* system. Repair of the S:ΔH69/V70 deletion reduced viral infectivity, while simultaneous repair with S:ΔV143/Y145 led to lower fusogenicity. In contrast, individual S:ΔV143/Y145 repair enhanced both fusogenicity and susceptibility to neutralization by sera from vaccinated individuals. This work underscores the complex genotype–phenotype landscape of the spike NTD in SARS-CoV-2, which impacts viral biology, transmission efficiency, and immune escape potential, offering insights with direct relevance to public health, viral surveillance, and the adaptive mechanisms driving emerging variants.

Keywords: SARS-CoV-2; spike; transmission; deletion; fusogenicity; antibody neutralization

Introduction

As we navigate the constantly shifting landscape of the coronavirus disease 2019 (COVID-19) pandemic, the remarkable potential of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) for genetic adaptation has taken center stage. Global turnover of SARS-CoV-2 lineages happened several times. In many regions, three variants of concern (VOCs) displaced the previous predominant lineages just in 2021: first, Alpha (B.1.1.7 and Q sublineages), then Delta (B.1.617.2 and AY sublineages), and, lastly, Omicron (B.1.1.529 and BA sublineages, among others) (Rambaut et al. 2020). This successive replacement of lineages throughout the pandemic suggests that the newer had a higher adaptive value than the previous ones (Chen et al. 2021, da Silva Francisco Junior et al. 2021), and thus, they are supposed to carry mutations associated with higher transmissibility and/or immune evasion.

The spike (S) protein—a type I transmembrane N-linked glycosylated protein—is a hotspot for mutations with high adaptive value. Spike proteins are located on the surface of SARS-CoV-2, and their main role is to mediate viral cell entry (Letko et al. 2020). The spike protein forms a homotrimer, which is cleaved post-transcriptionally into two subunits: S1 and S2. The S1 consists of the amino or N-terminal domain (NTD) and the receptor-binding domain (RBD), and it is responsible for binding to the host cell-surface angiotensin converting enzyme-2 (ACE2) receptors. The S2

subunit includes the trimeric core of the protein and is responsible for membrane fusion (Harvey et al. 2021, V'kovski et al. 2021). Therefore, some amino acid changes in the S protein may confer an advantage for transmission, considering its role in mediating viral cell entry. Additionally, most antibodies target sites either at the NTD or RBD, and therefore, mutations in these regions may enhance immune escape.

Specifically, deletions may play a decisive role in SARS-CoV-2 adaptive evolution, particularly in deletion-tolerant genome regions such as the S gene, as they can hardly be corrected by the proofreading activity of its RNA-dependent RNA polymerase (Minskaia et al. 2006, Denison et al. 2011, McCarthy et al. 2021). Notably, the NTD has been extensively impacted by deletions, which have emerged independently in multiple variants, often exhibiting complex and dynamic patterns, exemplified by the recurrent emergence and disappearance of the S:ΔH69/V70 deletion in dominant variants (Meng et al. 2021, McMillen et al. 2022). Acquisition of deletions in the NTD of the spike glycoprotein during long-term infections of immunocompromised patients has been reported and identified as an evolutionary pattern defined by recurrent deletion patterns that alter defined antibody epitopes (Avanzato et al. 2020, Choi et al. 2020). For example, different deletions are observed in Delta (S:ΔE156/F157-R158G), Alpha (S:ΔY144), and BA.1 (S:ΔV143/Y145) variants, all mapping to the same surface area, indicating a convergent function (Planas et al. 2021, Mishra et al. 2022). All these mutations are within the NTD site targeted by many anti-NTD neutralizing antibodies (McCallum et al. 2021a). In the Alpha variant, S:ΔH69/V70 and S:ΔY144 are directly associated with antibody escape (Kemp et al. 2021, McCallum et al. 2021a) and increased infectivity (Cantoni et al. 2022). More recent instances exist as well. Despite S:ΔY144 not being found in BA.2, its descendant lineage BJ.1 seems to have independently acquired this deletion (see cov-lineages/pango-designation issues #915 and #922 on GitHub). Then, it was passed on to XBB viruses through recombination with a BA.2-descended lineage, which was associated with increased immune escape regardless of the vaccination status or infection history (Zhang et al. 2022, Mykytyn et al. 2023, Tamura et al. 2023). Subsequently, in January 2023, the newly emerging XBB.1.5 lineage was shown to display an increased receptor-binding affinity and infectivity with respect to its parental lineage (Uriu et al. 2023, Yue et al. 2023).

The BA.1 lineage exhibited a growth advantage over previous VOCs (Puhach et al. 2022, Ward et al. 2022) and soon displaced the Delta variant. During this time, we became intrigued by the lack of overlap in NTD mutations between them. This is not the case for the Alpha variant and the BA.1 lineage, which share NTD deletions. If mutations in the NTD increased immune escape without compromising binding to ACE2, one might expect mutations in NTD to have a cumulative ("the more the better") effect. However, that would not be the case if epistatic interactions between sites prevented the fixation of particular mutations in different genomic and genetic backgrounds. Our primary objective in this study was to investigate the effect of NTD deletion repair in the BA.1 background. We examined the differences in success among three BA.1 haplotypes lacking the S:ΔH69/V70 and S:ΔV143/Y145 deletions, individually and in combination. These had been recurrently identified in immunocompromised individuals with chronic infections before widespread vaccination against COVID-19 (Avanzato et al. 2020, Choi et al. 2020, Kemp et al. 2021, McCarthy et al. 2021). Our approach integrates public data surveys with functional evidence to better understand how rare mutational events might shape viral evolution and pandemic dynamics.

Materials and methods

GISAID survey data retrieval and processing

We performed two surveys to track the absence of NTD deletions S:ΔH69/V70 and S:ΔV143/Y145 in sequences assigned to the Omicron lineage BA.1 or any of its sublineages (hereafter referred to collectively as "BA.1"). The first survey included 11 334 504 available SARS-CoV-2 sequences fetched from the GISAID (Khare et al. 2021) EpiCoV database on 15 June 2022. The GISAID EPI_SET is available at DOI:10.55876/gis8.230801ex. We ran Nextalign CLI v1.9.0 (Hadfield et al. 2018) with default scoring settings to obtain aligned genomes and peptides corresponding to each gene. Genomes from nonhuman hosts were discarded. Then, we filtered out genomes containing >5% ambiguous bases and 1000 gaps, or at least one indeterminate position among the lineage-defining sites of the spike protein. A total of 10 353 158 genomes passed these filters. Then, BA.1 genomes were selected and classified into one of three haplotypes, depending on their lack of S:ΔH69/V70 (RepΔ69/70), S:ΔV143/Y145 (RepΔ143/145), or both (RepBoth). After running Nextclade CLI v3.8.2 (Aksamentov et al. 2021), we applied a BA.1 lineage filter alongside specific mutation and motif inclusion and exclusion criteria. Manual curation followed. For RepΔ69/70, sequences were selected if they carried S:ΔV143, S:ΔY144, either S:ΔG142 or S:ΔY145, and a lack of S:ΔH69 and S:ΔV70. Sequences were required to match the S:69-70:HV motif. For RepΔ143/145, sequences were selected if they carried S:ΔH69, S:ΔV70, either S:G142D or S:Y145D, and a lack of S:ΔV143, S:ΔY144, and either S:ΔG142 or S:ΔY145. The required motif was S:142-145:DVYY. For RepBoth, sequences were selected if they carried S:G142D and a lack of S:ΔH69, S:ΔV70, S:ΔG142, S:ΔV143, S:ΔY144, and S:ΔY145. Sequences were required to match both the S:69-70:HV and S:142-145:DVYY motifs. We introduced an additional quality control (QC) step based on known control mutations: because BA.1 viruses typically exhibit the ORF1a:ΔS3675/F3677 (NSP6) deletion, sequences lacking this deletion were discarded. We used R v4.3.3 along with tidyverse v2.0.0 (Wickham et al. 2019) to conduct, manage, and visualize these analyses.

To assess potential artifacts in the quality of GISAID data, we performed an initial exploratory analysis of raw sequencing reads obtained via the European Nucleotide Archive (ENA). First, we identified GISAID genomes with corresponding raw reads deposited in the ENA by cross-referencing GISAID identifiers (virus name and accession number). Raw reads were downloaded and reprocessed to test the congruence between the GISAID consensus sequences and the raw read mappings. The mappings were manually inspected to confirm genotypes, and we calculated two metrics for each haplotype: the ratio of fixed deletion repairs (allele frequency >0.95) and the ratio of cases with significant deletion repair read support (allele frequency >0.10). Both the exploratory analysis and full ENA survey workflow described hereunder employed the same workflow outline and software.

ENA survey data retrieval and processing

Following this exploratory comparison, we extended the analysis to independently contrast the GISAID survey. We performed a second survey of combinations of NTD-repaired deletions using SARS-CoV-2 "read run" records via the ENA Portal API. This survey tracked the absence of deletions S:ΔH69/V70 and S:ΔV143/Y145 (separately and combined) in sequences assigned to lineage BA.1 (Omicron variant). We searched for runs from samples collected from 1 November 2021 to 1 August 2022, filtering by the National Center for Biotechnology Information taxonomy code

(2697049 for SARS-CoV-2) and host scientific name (*Homo sapiens*). Instrument platforms matching DNBseq, Element, and capillary sequencing were discarded. We also discarded runs with an RNAseq library strategy, as well as transcriptomic, metagenomic, and metatranscriptomic library sources. The survey brought out 1 541 892 matching records as of 10 December 2024 (Supplementary Table 1). After filtering for data URL availability, 1 250 257 records remained. The date of submission of these records ranged between 10 November 2021 and 1 December 2024. We did a nearly three-fold reduction of the dataset size, keeping 464 859 records selected at random for further analysis (see also Supplementary Table 1). Then, we ran an in-house pipeline implemented as a Snakemake (Köster and Rahmann 2012) workflow to detect NTD-repaired deletions. We ran Snakemake v8.25.3 with Python v3.12.7.

For each run, the following steps were carried out. FASTQ files were downloaded using the FTP URLs found in the ENA metadata. Then, read preprocessing and QC were performed using fastp v0.23.4 (Chen 2023) with default settings. Filtered reads were mapped to a complete genome sequence from a BA.1 SARS-CoV-2 isolate (GenBank accession number: OZ070629.1) with minimap2 v2.28 (Li and Birol 2018, Li and Alkan 2021), using the recommended presets according to the software documentation ("sr" for Illumina and Ion Torrent reads, "map-ont" for Oxford Nanopore reads, and "map-hifi" for PacBio reads). After the mapping step, pileups were calculated using SAMtools v1.20 (Danecek et al. 2021), and consensus genomes were obtained with iVar v1.4.3 (Grubaugh et al. 2019) with a minimum insertion frequency of 0.9, a minimum depth of 40, and otherwise default settings. These sequences were fed to pangolin v4.3 (O'Toole et al. 2021) to perform lineage assignment. The output was used to further filter the dataset, keeping records with passing pangolin QC and an assignment of BA.1 lineage or Omicron (BA.1-like) constellation. Then, variant calling was performed on the remaining runs with iVar v1.4.3 as well, with a minimum frequency threshold of 0.05, a minimum depth of 40, and otherwise default settings. Results were further annotated using SnpEff v5.2 and filtered with SnpSift v5.2 (Cingolani et al. 2012). We used a custom SnpEff database based on the mapping reference. Finally, matching records were classified into one of the deletion repair haplotypes—RepΔ69/70, RepΔ143/145, or RepBoth—while records with no evidence of repaired deletions were discarded. We established a minimum allele frequency threshold of 0.75 for each repaired deletion allele. For the two single repair haplotypes, we set a maximum frequency threshold of 0.25 nonrepaired deletion.

Additionally, we generated graphic visualizations of the genome region between sites 21694 and 22010 for visual inspection using R v4.3.3 with ape v5.8 (Paradis et al. 2019), Rsamtools v2.18.0 (Morgan et al. 2023), tidyverse v2.0.0 (Wickham et al. 2019), and ggpubr v0.6.0 (Kassambara 2023). The repository at https://github.com/PathoGenOmics-Lab/ena-spike-ntd-repdel-analysis (v1.0.0) contains the Snakemake workflow and the code to perform the initial screening.

Phylogenetic evaluation of event occurrences and clustering

We assessed the number of emergences in a phylogeny and whether these were ancestral to subsequent events. We placed genomes matching deletion repair haplotypes on a global-scale mutation-annotated tree (MAT) of public sequences (McBroome et al. 2021), under a maximum parsimony criterion, using UShER v0.6.2 (Turakhia et al. 2021). For GISAID data, genomes were placed on a tree dated to the survey date (15 June 2022), while for ENA data, consensus sequences inferred through our pipeline were placed on a tree dated to the last input record (11 March 2022).

To quantify the phylogenetic clustering of each haplotype, we developed tcfinder v1.0.0, a lightweight and efficient tool that performs an exhaustive breadth-first search of the target tips on a tree. This tree must be provided using the tabular "phylo4" structure, as defined in phylobase v0.8.10 (https://github.com/fmichonneau/phylobase). The tool is available in Bioconda (Grüning et al. 2018). We selected clusters containing at least two members and at least 90% target tips. This requirement was reduced from 100% to account for potential sequencing errors and ambiguous placements on the phylogenetic tree.

The number of emergences in the phylogeny for each haplotype was calculated by adding up the number of clusters to the number of unclustered target tips. To adequately compare the phylogeny-derived metrics between haplotypes, we calculated a normalized clustering index as the ratio between the size of each cluster and the number of sequences that were collected between the first and last cases of the cluster. A 7-day padding was added to both ends of the time window to account for missing data and short-lived clusters. This approach effectively standardizes the size of each cluster, correcting differences in surveillance efforts and viral prevalence at different geographic regions and time periods. In essence, it provides a specific rate of emergence and the subsequent cluster size relative to the broader population background. Transmission, on the other hand, is usually considered as a proxy of viral fitness because it is related to its basic reproductive number, reflecting the ability of the virus to replicate, persist, and spread within hosts and in the population (Domingo 2016). For observations backed by raw reads data, we interpreted this metric as an estimation of transmission fitness. Consequently, for clusters exhibiting cross-border clusters (i.e. involving samples from different countries), we further divided the cluster time window into country-specific subwindows. Then, the denominator of the normalized clustering index estimate was calculated as the sum of the number of records in the data repository for each country-specific time window. For GISAID data, the full survey was used as background. For ENA data, the 464 859 analyzed records were used. Regardless, there were no significant differences between the distribution of fitness estimates within each haplotype, whether using country-specific subwindows ("slicing" method) or not ("adding" method) (Supplementary Fig. 1). Results in the main text use the first method. Differences in the distribution of the normalized clustering index between haplotypes and host age were evaluated using Wilcoxon rank-sum tests. To determine age fold changes, the ratio of group average values was calculated. Statistical analyses were performed and visualized using R v4.1.2 along with tidyverse v2.0.0 (Wickham et al. 2019) and ggpubr v0.4.0 (Kassambara 2023).

The complete pipeline (including the phylogenetic placement, phylogenetic clustering quantification, fitness estimation, and data visualizations) is available as a Snakemake workflow at https://github.com/PathoGenOmics-Lab/transcluster (v2.1.0). We also studied the location and time span of the clusters using the associated sample metadata from GISAID.

Biological characterization of BA.1 deletion repairs

Deletion repairs of S:ΔH69/V70 and S:ΔV143/ΔY145 were introduced individually and in combination into a pCG1 plasmid encoding a codon-optimized BA.1 spike protein (Giménez et al. 2022) by site-directed mutagenesis. All the constructs were verified by Sanger sequencing. Pseudotyped vesicular stomatitis virus

(VSV) encoding a green fluorescent protein (GFP) reporter gene and carrying the different spike proteins was produced as previously reported (Gozalbo-Rovira et al. 2020). To assess the effects on virus production, pseudotyped VSV carrying each construct was produced independently three times. The resulting viruses were then titrated by infecting Vero E6 cells (kindly provided by Dr Luis Enjuanes; CNB-Spanish National Research Council (CSIC), Spain) or Vero E6-TMPRSS2 cells (JCRB Cell Bank catalogue code: JCRB1819) for 16 h, followed by quantification of GFP-expressing infected cells using a live cell microscope (Incucyte SX5; Sartorius) to obtain the number of focus-forming units (FFUs) per milliliter.

To assess thermal stability, 500 FFUs of these pseudotyped viruses were incubated for 15 min at a range of temperatures in a thermal cycler (30.4, 31.4, 33.0, 35.2, 38.2, 44.8, 47.0, 48.6, and 49.6°C; Biometra TOne Gradient, Analytik Jena), and the surviving virus was used to infect VeroE6-TMPRSS2 cells for 16 h. The GFP signal in each well was then determined using a live-cell microscope (Incucyte SX5, Sartorius). The average GFP signal observed in mock-infected wells was subtracted from all infected wells, followed by standardization of the GFP signal to the average GFP signal from wells incubated at 30.4°C. Finally, we fitted a three-parameter log-logistic function to the data using the drc v3.0-1 R package ("LL.3" function) and calculated the temperature resulting in a 50% reduction in virus infection ("ED" function).

To assess the effects on neutralization by polyclonal sera, we used six sera from convalescent patients from the first COVID-19 wave in Spain and six sera from individuals that had been administered two doses of the BioNTech-Pfizer Comirnaty COVID-19 vaccine. Sera samples were obtained from the La Fe University and Polytechnic Hospital of Valencia and were collected after informed written consent had been obtained, with approval by the ethical committee and institutional review board (registration number 2020-123-1). The neutralization capacity of the sera (NT50) was obtained as previously described on VeroE6-TMPRSS2 cells (Giménez et al. 2022). We used a previously described flow cytometry assay based on the use of polyclonal sera to examine surface expression (Grzelak et al. 2020). Briefly, HEK293T cells were transfected with the different S mutants using the calcium chloride method. After 24 h, cells were detached using phosphate-buffered saline (PBS) with 1 mM ethylenediaminetetraacetic acid (EDTA), washed, and incubated in ice with different polyclonal sera (three from convalescent patients from the first COVID-19 wave in Spain and one from individuals that had been administered two doses of the BioNTech-Pfizer Comirnaty COVID-19 vaccine) at a 1:300 dilution in PBS containing 0.5% BSA and 2 mM EDTA for 30 min. Next, cells were washed three times with PBS, stained with anti-IgG Alexa Fluor 647 (Thermo Fisher Scientific) at a 1:400 dilution and analyzed by flow cytometry similarly treated un-transfected controls to set the threshold for positive cells.

For cell–cell fusion assays, we used a split Venus fluorescent protein system (García-Murria et al. 2019). HEK293T cells were grown overnight in 24 well plates (1.5 × 105 cells/well) using Dulbecco's modified Eagle medium supplemented with 10% fetal bovine serum. After 24 h, cells were transfected using Lipofectamine 2000 (Invitrogen) with 0.5 μg of either a 1:1 mixture of the S plasmids and a Jun-Nt Venus fragment (Addgene 22 012) plasmid or a mixture of hACE2 plasmid (kindly provided by Dr Markus Hoffman; German Primate Center, Goettingen, Germany) (Hoffmann et al. 2020) and the Fos-Ct Venus fragment (Addgene 22013). After 24 h, cells were counted, and the S-transfected cells were mixed at a 1:1 ratio with ACE2-transfected cells and seeded in 96 well plates (3×10^4 cells/well) in 100 μl of media. Cells transfected with the Wuhan-Hu-1 served as a positive control, while cells transfected with hACE2 and Jun-Nt Venus were used as a negative control. We obtained the GFP Integrated Intensity (GCU·μm^2/image) in each condition using a live-cell imaging platform (Incucyte SX5, Sartorius) at 24 h post-mixing and standardized to the signal obtained from the positive control (Wuhan-Hu-1 spike protein). All experiments were performed at least three times in triplicates.

Statistical analyses were performed and visualized using R v4.1.2, along with tidyverse v2.0.0 (Wickham et al. 2019) and ggpubr v0.4.0 (Kassambara 2023) to facilitate the analysis and enhance the visualization of these results. Comparisons were conducted utilizing *t*-tests. We used unpaired tests for all assays except neutralization (log-transformed NT50 values) and surface expression, for which we used paired tests, as we used the same sets of polyclonal sera in each experiment. To determine fold change values, the ratio of group average values was calculated.

Results

Detection and clustering dynamics of spike NTD deletion repair in the BA.1 lineage using GISAID genomes

We first examined the differences in success within the Omicron BA.1 lineage depending on the presence of ΔH69/V70 and ΔV143/Y145 in the NTD of the spike protein through a global survey of viral genome data. This domain is characterized by recurrent deletions occurring in distinct independent lineages (McCarthy et al. 2021, Meng et al. 2021, Planas et al. 2021, McCallum et al. 2021a, 2021b, Tamura et al. 2023). We investigated potential patterns of deletion repair events and the factors behind their detection through a survey of 11.3 million consensus genomes from the GISAID's EpiCov platform (Fig. 1a). We identified 5499 sequences that carried repaired S:ΔH69/V70, S:ΔV143/Y145, or both (Fig. 2a, Table 1; see full details in Supplementary Table 2). The most frequently observed group consisted of sequences exhibiting the repaired S:ΔH69/V70 in the BA.1 background, while the remaining repair patterns were less frequently observed.

We also identified clusters of each haplotype through a phylogenetic estimation using worldwide SARS-CoV-2 genomes. We measured an average within-cluster collection date window of 24 ± 27 days, with 70% of clusters including only within-border observations (Supplementary Fig. 2). The haplotype bearing the S:ΔH69/V70 single repair in GISAID consensus sequences had the highest median normalized clustering index, followed by the double repair and, finally, by the S:ΔV143/Y145 single repair (Fig. 2b). We then investigated the association of host age. No significant differences in host age were observed between deletion repair haplotypes (Supplementary Fig. 3). However, clusters of observations with repaired S:ΔH69/V70 in GISAID consensus sequences were associated with higher host age compared to nonclustered genomes ($P = .0010$; Supplementary Fig. 4), an effect likely independent of sampling location (Supplementary Figs 5 and 6). These findings should be interpreted cautiously due to substantial missing patient metadata (Supplementary Fig. 7) and the potential for sequence artifacts in the genome region of interest in SARS-CoV-2 sequences from GISAID.

To evaluate the reliability of GISAID-derived data, we compared its consensus sequences to corresponding raw reads from the ENA. We cross-referenced the available ENA entries with all GISAID identifiers of genomes with deletion repairs to identify matching records (Supplementary Table 3). Among 18 genomes with matching ENA runs, <25% confirmed fixed deletion repairs as indicated by the GISAID consensus, and fewer

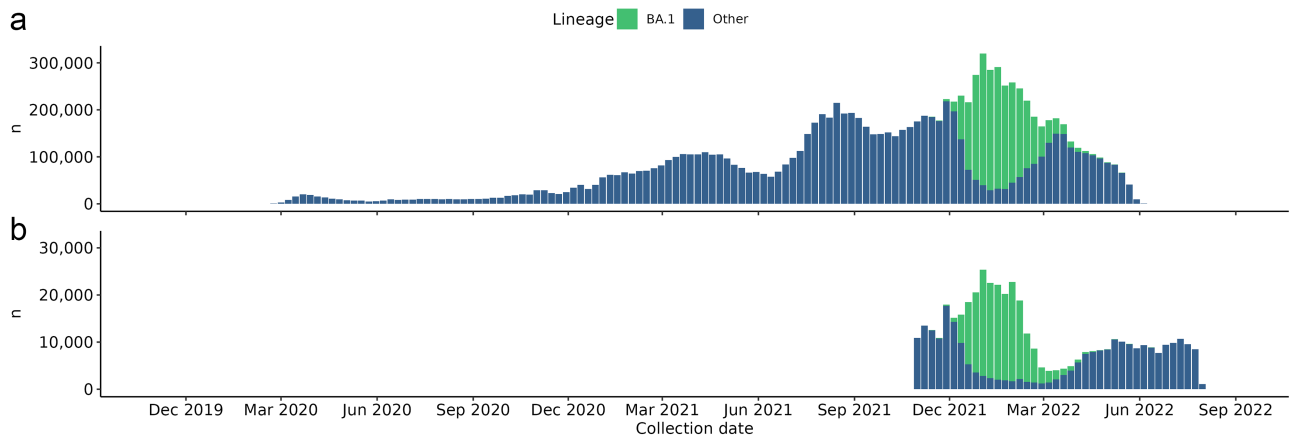

Figure 1. Distribution of collection dates of the two survey datasets. Bars show the weekly number of records in each survey dataset, excluding those with missing collection dates. Records identified as deletion repair-bearing were extracted from the BA.1 data partition in both surveys, which is highlighted in a lighter shade. a) GISAID dataset, with lineages extracted from metadata. b) ENA dataset (3-fold random subset), with lineages assigned running pangolin v4.3 on consensus sequences reconstructed from read mappings (see Materials and methods), as no lineage metadata is provided in database records. Although lineage filtering prior to data processing was not possible for the ENA dataset, the chosen time frame ensured adequate representation of BA.1 records.

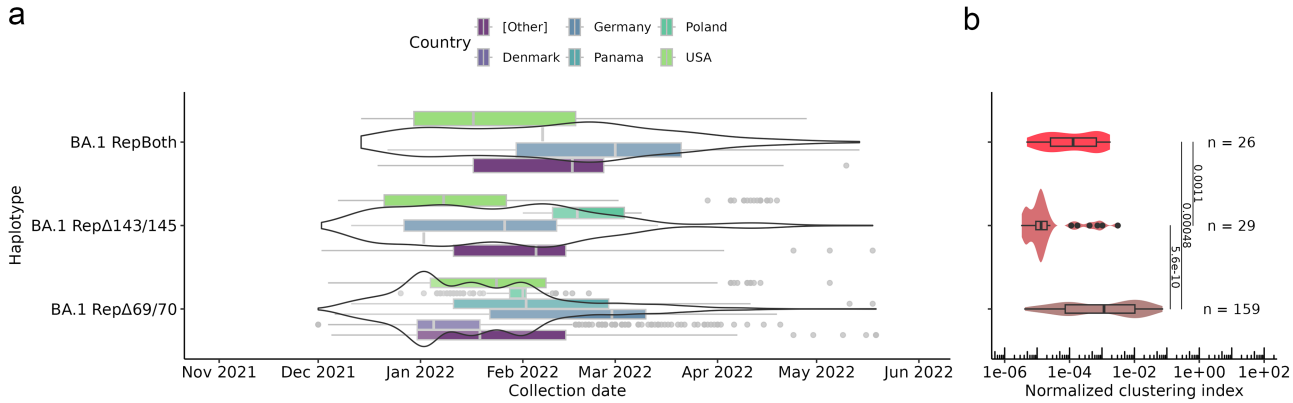

Figure 2. Summary of GISAID survey results for BA.1 genomes with repaired S:ΔH69/V70, S:ΔV143/Y145 or both. a) Sample collection timelines broken down by sampling location. The five countries with the highest number of observations globally are displayed. b) Differences in the normalized clustering index between deletion repair haplotypes. This metric corrects the size of each tree cluster regarding local spatial and temporal sequencing efforts and lineage abundances (see Materials and methods). The differences were assessed using Wilcoxon rank-sum tests, and P-values are displayed above each line connecting a pair of haplotypes. Each data point corresponds to a transmission cluster. Group sizes (number of clusters, n) are displayed next to the P-values.

Table 1. Description of Omicron BA.1 haplotypes with repaired NTD deletions detected through the GISAID survey.

Haplotype	Observations	Ratio to BA.1	Clusters in phylogeny	Emergences in phylogeny
BA.1 RepΔ69/70	4643	$1.92 \cdot 10^{-3}$	159	3578
BA.1 RepΔ143/145	557	$2.30 \cdot 10^{-4}$	29	499
BA.1 RepBoth	299	$1.24 \cdot 10^{-4}$	26	215

The ratio to BA.1 is calculated as the fraction of observations over the number of BA.1 observations found in the survey ($n = 2\,420\,866$).

than 60% had significant read support of the deletion repair genotype. Specifically, for repaired S:ΔH69/V70, seven genomes matched ENA runs, with only one exhibiting a fixed repaired deletion. For repaired S:ΔV143/Y145, one genome matched ENA runs, with none showing fixed repairs. For genomes with both repaired deletions, 10 matched ENA runs, with 3 displaying fixed repairs. These findings suggest that most GISAID sequences may not reliably represent deletion repair genotypes, highlighting the need for further validation of such analyses.

Detection and transmission dynamics of spike NTD deletion repair in the BA.1 lineage using raw sequencing data

In our survey of viral genome data from GISAID, we were able to identify repairs to deletion mutations, which are theoretically rare events. However, their validity may be confounded by sequencing artifacts. Public consensus genome sequences often suffer from low coverage, poor sequencing quality, and imputed data, hindering the detection of genuine deletion repair events. To address potential artifacts, we analyzed 464 859 read run records from

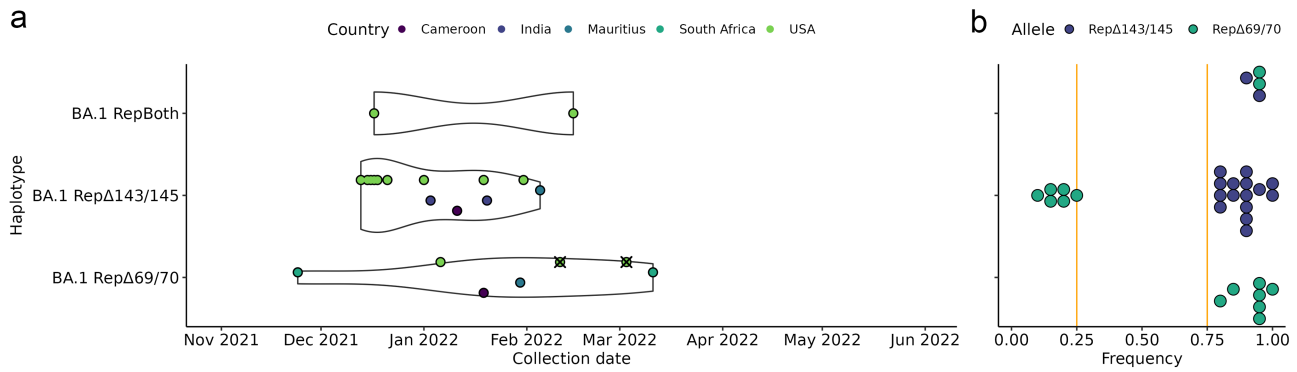

Figure 3. Summary of ENA survey results for BA.1 records with repaired S:ΔH69/V70, S:ΔV143/Y145 or both. a) Sample collection timelines broken down by sampling location. The five countries with at least one observation are displayed. Crossed markers indicate observations belonging to the single detected transmission cluster. The timeframe aligns with Figure 2A for consistency. b) Allele frequency of deletion repairs in each run record. For each allele, each point represents its binned relative frequency in a different record. Point bins have a width of 0.05. Note that, for haplotype RepΔV143/Y145, some runs bear repaired S:ΔH69/V70 with a frequency under the 0.25 threshold (see Materials and methods). No differences in the estimated transmission fitness between deletion repair haplotypes are reported, as we did not detect transmission for all haplotypes (see Table 2).

Table 2. Description of Omicron BA.1 haplotypes with repaired NTD deletions detected through the ENA survey.

Haplotype	Observations	Ratio to BA.1	Average repair allele frequency	Transmission clusters	Minimum independent emergences
BA.1 RepΔ69/70	7	$3.76 \cdot 10^{-5}$	0.925 ± 0.073	1	6
BA.1 RepΔ143/145	15	$8.05 \cdot 10^{-5}$	0.884 ± 0.064	0	15
BA.1 RepBoth	2	$1.07 \cdot 10^{-5}$	0.945 ± 0.021	0	2

The ratio to BA.1 is calculated as the fraction of observations over the number of BA.1 read runs found in the survey analysis ($n = 186\,387$). The average and standard deviation of the frequency of the deletion repair allele(s) associated with each haplotype are reported as well.

the ENA (Fig. 1b), randomly selected out of 1.54 million collected between 1 November 2021 and 1 August 2022. This allowed us to reassess haplotype assignments with greater reliability.

We observed substantial discrepancies compared to the initial GISAID genome survey. We identified 24 samples that carried repaired S:ΔH69/V70, S:ΔV143/Y145, or both (Fig. 3a, Table 2; see full details in Supplementary Table 4) with an average haplotype-defining allele frequency of 0.918 (Fig. 3b). The ratio of repaired deletions relative to BA.1 observations was ∼3- to 51-fold lower (depending on the haplotype) than in the GISAID dataset, underscoring a significant drop in detection. This suggests that the earlier survey was influenced by artifacts or biases in sequence processing.

We then performed a phylogenetic estimation of transmission clusters. A single cluster was detected (Table 2), involving two records with repaired S:ΔH69/V70 (accession numbers SRR19566289 and SRR20496144) collected 20 days apart in California and New Jersey, USA. The allele frequency of the deletion repair declined 12% between infections. Notably, the assigned haplotype exhibited the highest normalized clustering index—interpreted as an estimation of transmission fitness for observations supported by reads data—in the earlier GISAID survey (Fig. 2b). In short, generally, each observation was detected as an independent emergence with limited epidemiological impact, suggesting that these NTD-repaired deletion events might be rare, isolated, and unlikely to drive SARS-CoV-2 evolution or transmission dynamics by themselves.

NTD deletion repair in the Omicron BA.1 background has a nonaccumulative effect on viral phenotype

NTD deletion repair events could have still contributed to viral diversity in a broader evolutionary context. Even isolated events might provide new opportunities for fitness-enhancing mutations to emerge, particularly in high-transmission scenarios. To investigate whether certain viral characteristics act as drivers of certain patterns of deletion repair compared to others, we analyzed the effect of these events on the Omicron BA.1 spike using a pseudotyped VSV system. We generated VSV pseudotyped with the BA.1 spike proteins with repaired S:ΔH69/V70, repaired S:ΔV143/Y145, or both deletions repaired. We then used these pseudotyped viruses to examine possible phenotypic effects on spike function, including the efficiency of virus production, thermal stability, surface expression, susceptibility to antibody neutralization, and fusogenicity.

To assess possible effects on virus production, which encompasses both viral egress and infection of the next cell, VSV was pseudotyped with all spike constructs under identical conditions at the same time, and the amount of virus produced titrated on Vero E6 cells (Fig. 4a). Virus production was not significantly affected by S:ΔV143/Y145 repair (0.77-fold change; $P = .060$) but was significantly reduced upon repairing S:ΔH69/V70 (0.28-fold; $P = 6.6 \cdot 10^{-4}$) or both deletions (0.34-fold; $P = 1.6 \cdot 10^{-4}$). A similar effect was observed when the titer of the virus was evaluated in cells expressing the TMPRSS2 coreceptor, indicating a substantial negative effect of S:ΔH69/V70 repair by itself on virus production (0.26-fold; $P = .017$). Reduced viral production can stem from lower stability of the spike proteins or lower expression on the cell surface. To test the former, we obtained the temperature resulting in a 50% reduction of virus titer for each spike variant (i.e. 50% inactivation temperature; Fig. 4b). No significant differences were observed, suggesting that spike protein stability was not the driver of these differences. We therefore examined whether the different constructs had altered cell expression of the different constructs by flow cytometry, using polyclonal sera from four individuals (Fig. 4c). We did not detect differences in the median

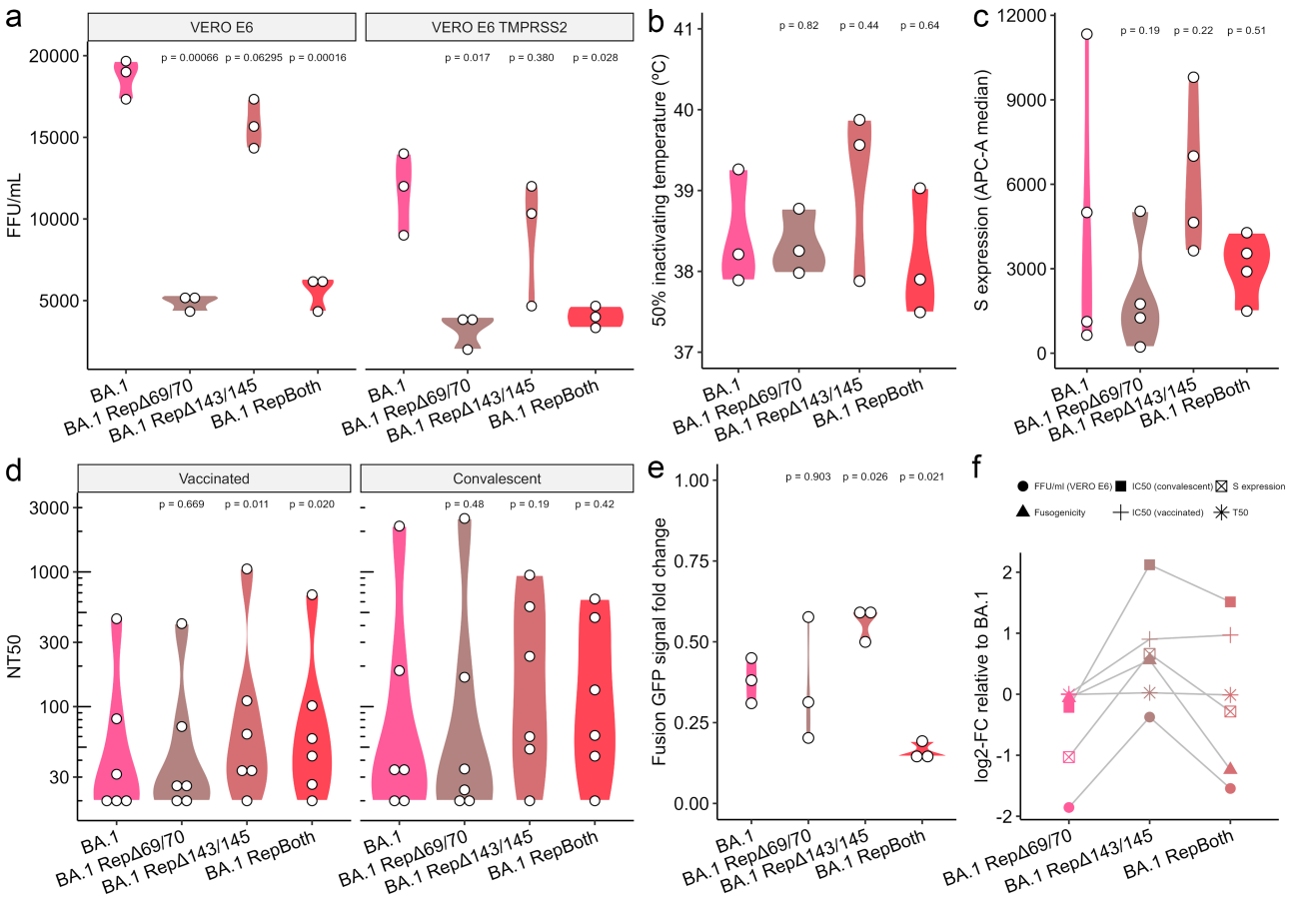

Figure 4. Experimental assessment of the effect of repairing NTD deletions in Omicron BA.1 background. a) Comparison of viral titres obtained for each spike variant using pseudotyped VSV on the indicated cell line. b) Comparison of the temperature resulting in 50% inactivation of pseudotyped VSV carrying the indicated S protein. c) Surface expression of the spike protein in transfected HEK293T cells quantified using flow cytometry. d) Reciprocal 50% neutralization titres (NT50) of sera from convalescent and vaccinated individuals. e) Relative ability of each spike variant to drive cell-cell fusion relative to the Wuhan-Hu-1 spike. f) Summary of the log2-transformed fold change for each spike variant relative to that of BA.1 in terms of average virus production in VERO E6 cells, average fusogenicity, median NT50, average surface expression, and average 50% inactivating temperature. Subfigures A-E represent the median and interquartile range of at least three replicates. Differences were assessed using Wilcoxon rank-sum tests.

cell surface expression of the spike protein between the canonical BA.1 protein and any deletion repair haplotype. However, the repair of S:ΔV143/Y145 resulted in higher expression than S:ΔH69/V70 (3.2-fold; $P = .0030$) and the double repair (1.9-fold; $P = .032$), resembling the effect observed with virus production.

As neutralizing antibodies can be key drivers of viral evolution, we questioned whether deletion repair affected neutralization by polyclonal sera from convalescent donors from the first epidemic wave in Spain and those dually vaccinated with the Comirnaty mRNA vaccine ($n = 6$ each; Fig. 4d). Significant increases in susceptibility to neutralization in sera from vaccinated individuals against viruses with repaired S:ΔV143/Y145 (2.1-fold; $P = .011$) or both S:ΔH69/V70 and S:ΔV143/Y145 repaired (1.5-fold; $P = .020$) were observed. A similar trend was observed with convalescent sera, but statistical significance was not reached ($P = .19$ and $P = .42$, respectively) possibly due to a higher intra-sample variability. Thus, these results indicate that increased neutralization is driven by the repair of S:ΔV143/Y145 in the BA.1 background.

Cell-to-cell spread via fusion of the plasma membrane could potentially reduce exposure to neutralizing antibodies, compensating for increased susceptibility to neutralization. Hence, we also examined whether deletion repair could alter the ability of the spike protein to fuse cells (Fig. 4e). Interestingly, the repair of S:ΔV143/Y145 increased cell fusion relative to the BA.1 spike protein (1.5-fold; $P = .026$), while the repair of both S:ΔH69/V70 and S:ΔV143/Y145 led to a decrease of >50% in the average fusogenicity (0.42-fold; $P = .021$). The presence of S:ΔH69/V70 by itself did not seem to play a role in cell–cell fusion in a BA.1 background.

Discussion

Deletions in the SARS-CoV-2 genome have a significant impact on viral adaptation and fitness, often surpassing the effects of single nucleotide variants (McCarthy et al. 2021, Meng et al. 2021, Cantoni et al. 2022, Mishra et al. 2022, Venkatakrishnan et al. 2023). In fact, deletions in the NTD region of the spike protein are fixed in many prominent viral variants. Our understanding of the repair patterns of deletions is crucial for elucidating the genetic and phenotypic characteristics of the VOCs. In this work, we demonstrate that repairing these deletions can alter viral characteristics and potentially influence the success, emergence, and transmission dynamics of specific viral haplotypes.

The *in vitro* assessment of the viral characteristics driving different patterns of deletion repair in BA.1 showed differences in spike virus production and infectivity, fusogenicity, and sera neutralization, but no differences in thermal stability. Repair of S:ΔH69/V70 led to reduced virus titers individually and in

combination with S:ΔV143/Y145 repair. We did not observe differences in thermal stability with the BA.1 protein, suggesting that the effect of S:ΔH69/V70 and S:ΔV143/Y145 repair did not affect full-protein stability.

Cell–cell fusion and subsequent formation of syncytia can be mediated by viral membrane glycoproteins (Li et al. 2003, Tseng et al. 2005) such as the spike protein found in SARS-CoV-2, and it is a key feature of SARS-CoV-2 infection (Buchrieser et al. 2020, Bussani et al. 2020, Hoffmann et al. 2020). In line with prior research, our results point to a lower fusogenicity of the unaltered BA.1 spike protein compared to that of Wuhan-Hu-1, which has been attributed to its inefficient spike cleavage and unfavored TMPRSS2-mediated cell entry (Meng et al. 2021). Interestingly, deletion repair patterns exerted a nonaccumulative influence on the fusogenicity of the BA.1 spike protein: repair of S:ΔH69/V70 alone did not have a significant impact and repair of S:ΔV143/Y145 promoted higher fusogenicity, while repair of both deletions led to a significant decrease in fusogenicity.

Deletion repair did also affect sera neutralization. In the BA.1 background, repaired S:ΔV143/Y145 was associated with a significant increase in sensitivity to neutralization by polyclonal sera obtained from vaccinated individuals. Conversely, we did not detect a significant association of repaired S:ΔV143/Y145 with sensitivity to sera from convalescent, "first-wave" patients. Repair of S:ΔH69/V70 did not affect neutralization with either serum group. Although our surrogate system might not fully reflect the complex interactions of the spike protein with host cells and the immune system *in vivo*, and other regions of the spike protein or the viral genome might also contribute to SARS-CoV-2 fitness and adaptation, our results point to the variant-independent influence of NTD variability on immune effects. These findings point to deletion repair events being rare and isolated, as their context-dependent and often deleterious effects generally suggest limited adaptive advantage. This may also indicate epistatic incompatibilities that constrain their contribution to viral fitness and transmission. In agreement with the *in vitro* results, sequence surveys suggest that deletion repair events were rare and likely isolated, with limited direct impact on SARS-CoV-2 evolution or transmission dynamics. This aligns with the hypothesis that recurrent deletion and apparent repair events in early VOC may have been driven by persistent viral reservoirs, where prolonged infections or animal reservoirs allow for cycles of mutation accumulation and reversion before reintroduction. In contrast, such dynamics appear less relevant in second-generation dominant lineages (e.g. BA.2 sublineages), as reviewed in Roemer et al. (2023).

All survey efforts are expected to be biased by geographic and temporal differences in sequencing efforts and lineage prevalence. Sequence analyses could be affected by sampling bias and technical artifacts. However, despite our attempts to account for demographic differences between variants, our analysis could be affected by biases in genome sequencing and publication practices.

In our study, we observed instances of apparent genetic reversions in GISAID consensus sequences that could, in fact, be attributable to missing information in poorly sequenced regions or polymorphic sites. The ENA survey revealed a significantly lower frequency of deletion repair events compared to GISAID, and the majority of one-to-one comparisons of GISAID genomes against raw reads data did not confirm the deletion repair genotype. This suggests potential artifacts in GISAID genomes, likely arising from a replacement of problematic genomic regions with the homologous ancestral sequence. Such undocumented adjustments can create reference bias, distorting the interpretation of genetic events and complicating genomic analyses, particularly given the QC mechanisms employed by repositories like GISAID. In fact, key global resources such as the daily-updated MATs we used for clustering and transmission estimation (McBroome et al. 2021) could well be impacted by these issues. We highlight the need for greater scrutiny of sequence deposition and QC mechanisms. Ongoing efforts such as the Viridian initiative (Hunt et al. 2024) have already pointed out these issues and are working toward reducing their impact. Incidentally, the limitations of the FASTA format itself also contribute to this problem, since it restricts the amount of metadata and quality information that can directly accompany sequences. This choice, while historically convenient, may no longer align with the need for richer data representations. Accordingly, submission of raw sequencing data from genome studies to public repositories should be prioritized.

We also acknowledge that our study does not specifically address the mutational mechanism that led to deletion repair. We speculate that recombination between different infecting variants could be behind the occurrence of deletion-repairing haplotypes, at least in some cases. We further speculate that template switching at wrong loci during replication ("imperfect homologous recombination," as hypothesized in Chrisman et al. 2021) could explain the emergence of rare indel events such as deletion repairs in the absence of coinfections. Recombination might have played a role in the evolution of SARS-CoV-2 from the ancestral coronavirus before diversification in humans (Boni et al. 2020, Andersen et al. 2020, Kirtipal et al. 2020, MacLean et al. 2021, Wells et al. 2021), and studies using focused datasets have also pointed to ongoing recombination and proposed insights into its dynamics (Jackson et al. 2021, VanInsberghe et al. 2021, Ignatieva et al. 2022, Perez-Florido et al. 2023). Still, the emergence of the first demonstrably recombinant lineages of epidemiological relevance was not reported until years into the pandemic (Tamura et al. 2023, Yue et al. 2023). Due to the poor phylogenetic resolution of the datasets, our efforts ultimately led to an inconclusive result, rendering us unable to articulate a systematic verification of the recombination hypothesis.

In summary, our study describes the occurrence of rare deletion repair events in the SARS-CoV-2 spike NTD and investigates their impact on viral fitness and clinical characteristics depending on the genetic context. In doing so, we have provided novel insights into the repair patterns of NTD deletions in the Omicron BA.1 background and their phenotypic consequences. While significant progress has been recently made in estimating the fitness effects of individual mutations across the SARS-CoV-2 genome (Bloom and Neher 2023), these approaches do not focus on how mutations interact within different genomic contexts. There have also been efforts to specifically characterize the variability of the S protein NTD through full swap assays, comparing the ancestral (Wuhan-Hu-1) spike protein with that of Alpha and Omicron BA.1 (Cantoni et al. 2022), Delta and Omicron BA.1 and BA.2 (Meng et al. 2021), and even with more distant viruses like SARS-CoV (Qing et al. 2021). However, to our best knowledge, no previous study has undertaken a combinatorial approach to characterize each separate marker. Our work reveals that the repair of specific deletions might be driven by distinct phenotypic traits, such as changes in viral characteristics, including fusogenicity, infectivity, and susceptibility to neutralization. Understanding these genotype–phenotype relationships can inform the evolutionary dynamics and adaptation of SARS-CoV-2 variants. Future studies building upon these findings will contribute to the development of effective strategies to monitor and mitigate the impact of NTD deletions in emerging SARS-CoV-2 variants.

Acknowledgements

We gratefully acknowledge all data contributors, i.e. the authors and their originating laboratories responsible for obtaining the specimens, and their submitting laboratories for generating the genetic sequence and metadata and sharing via the GISAID Initiative, on which this research is based. We also extend our acknowledgment to all contributors and submitters of raw reads hosted by the ENA and other members of the International Nucleotide Sequence Database Collaboration, which are also integral to this study. We thank Dr Óscar González Recio (INIA-CSIC, Spain) for providing sequencing data from a BA.1 sample with S:142G and repaired NTD deletions (GISAID accession code: EPI_ISL_9805648) for an initial exploratory analysis related to S gene target failure, Dr Luis Enjuanes (CNB-CSIC, Spain) for providing Vero E6 cells, and Dr Markus Hoffman (German Primate Center, Germany) for providing the hACE2 plasmid. We thank Dr Irving Cancino Muñoz for his insightful feedback and recommendations on sequence data processing. We thank Francisco José Martínez Martínez for comments on phylogeny visualization and transmission cluster analyses. We thank Jimena Solana González for taking the time to review the manuscript and offering thoughtful feedback. Finally, we thank the Associate Editor, Dr Richard Neher, for his detailed and prompt feedback during the review process. The computations were performed on the HPC cluster Garnatxa at the Institute for Integrative Systems Biology (I^2SysBio, Spain), a joint collaborative research institute involving the University of Valencia and the CSIC.

Author contributions

M.Á.-H., B.N.-D., P.R.-R., and M.C. conceived the theoretical framework. M.Á.-H. and B.N.-D. devised the initial idea. M.Á.-H. and M.C. conceived the final study design. M.Á.-H. implemented and performed the data retrieval, data curation, and computations. J.Z., B.G., and R.G. designed the experiments. J.Z. and B.G. performed the experiments. M.Á.-H. carried out the statistical analyses of the experiments. M.G. and C.A.-G. contributed biological samples. M.A.B. did project management. M.Á.-H. drafted the manuscript with support from B.N.-D., P.R.-R., R.G., and M.C. F.G.-C. aided in interpreting the results. M.Á.-H., P.R.-R., B.N.-D., R.G., F.G.-C., I.C., and M.C. discussed the results and commented on the manuscript. M.C. supervised the project.

Supplementary data

Supplementary data is available at *VEVOLU Journal* online.

Conflict of interest: None declared.

Funding

This research work was funded by the European Commission—NextGenerationEU/PRTR (Regulation EU 2020/2094), through CSIC's Global Health Platform (PTI+ Salud Global) to M.C., R.G., I.C., and F.G.-C. M.Á.H. is supported by the Generalitat Valenciana and the European Social Fund "ESF Investing in your future" through grant CIACIF/2022/333. This work was also a part of projects CNS2022-135116 (M.C.) and CNS2022-135100 (R.G.) funded by MCIN/AEI/10.13039/501100011033 and the European Union NextGenerationEU/PRTR.

Data availability

The GISAID EPI_SET underlying this article is available via https://dx.doi.org/10.55876/gis8.230801ex. Raw reads data are publicly available via the ENA under the accession numbers provided in Supplementary Table 1. The software we developed and used to find clusters in a global phylogeny, tcfinder v1.0.0, is a command line tool released as free software under the GNU GPLv3 license. Its source code is available in GitHub (https://github.com/PathoGenOmics-Lab/tcfinder). It is also available as a package in the Bioconda channel for the Conda package manager (https://anaconda.org/bioconda/tcfinder). The pipeline used for the phylogenetic analysis, transcluster v2.1.0, is a Snakemake workflow also released as free software under the GNU GPLv3 license, and its source code is available in GitHub (https://github.com/PathoGenOmics-Lab/transcluster). The processed data, code, configuration files, results of the ENA survey, and both phylogenetic analyses are available via https://dx.doi.org/10.20350/digitalCSIC/17032. The lists of identifiers assigned to each haplotype under study are also incorporated into the article and its Supplementary material. Code for performing the ENA survey and the pipeline for processing the resulting raw sequencing data are available as free software under the GNU GPLv3 license, and its source code is available in GitHub (https://github.com/PathoGenOmics-Lab/ena-spike-ntd-repdel-analysis, v1.0.0).

References

Aksamentov I, Roemer C, Hodcroft EB *et al*. Nextclade: clade assignment, mutation calling and quality control for viral genomes. *J Open Source Softw* 2021;**6**:3773.

Andersen KG, Rambaut A, Lipkin WI *et al*. The proximal origin of SARS-CoV-2. *Nat Med* 2020;**26**:450–452.

Avanzato VA, Matson MJ, Seifert SN *et al*. Case study: prolonged infectious SARS-CoV-2 shedding from an asymptomatic immunocompromised individual with cancer. *Cell* 2020;**183**:1901–1912.e9.

Bloom JD, Neher RA. Fitness effects of mutations to SARS-CoV-2 proteins. *Virus Evolut* 2023;**9**:vead055.

Boni MF, Lemey P, Jiang X *et al*. Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic. *Nat Microbiol* 2020;**5**:1408–17.

Buchrieser J, Dufloo J, Hubert M *et al*. Syncytia formation by SARS-CoV-2-infected cells. *EMBO J* 2020;**39**:e106267.

Bussani R, Schneider E, Zentilin L *et al*. Persistence of viral RNA, pneumocyte syncytia and thrombosis are hallmarks of advanced COVID-19 pathology. *EBioMedicine* 2020;**61**:103104.

Cantoni D, Murray MJ, Kalemera MD *et al*. Evolutionary remodelling of N-terminal domain loops fine-tunes SARS-CoV-2 spike. *EMBO Rep* 2022;**23**:e54322.

Chen S. Ultrafast one-pass FASTQ data preprocessing, quality control, and deduplication using fastp. *iMeta* 2023;**2**:e107.

Chen Z, Chong KC, Wong MCS *et al*. A global analysis of replacement of genetic variants of SARS-CoV-2 in association with containment capacity and changes in disease severity. *Clin Microbiol Infect* 2021;**27**:750–57.

Choi B, Choudhary MC, Regan J *et al*. Persistence and evolution of SARS-CoV-2 in an immunocompromised host. *N Engl J Med* 2020;**383**:2291–93.

Chrisman BS, Paskov K, Stockham N *et al*. Indels in SARS-CoV-2 occur at template-switching hotspots. *BioData Mining* 2021;**14**:20.

Cingolani P, Patel VM, Coon M *et al*. Using *Drosophila melanogaster* as a model for genotoxic chemical mutational studies with a new program, SnpSift. *Front Genet* 2012;**3**:35.

Danecek P, Bonfield JK, Liddle J *et al*. Twelve years of SAMtools and BCFtools. *GigaScience* 2021;**10**:giab008.

da Silva Francisco Junior R, Lamarca AP, de Almeida LGP *et al*. Turnover of SARS-CoV-2 lineages shaped the pandemic and

enabled the emergence of new variants in the state of Rio de Janeiro, Brazil. *Viruses* 2021;**13**:2013.

Denison MR, Graham RL, Donaldson EF *et al*. Coronaviruses: an RNA proofreading machine regulates replication fidelity and diversity. *RNA Biol* 2011;**8**:270–79.

Domingo E. Long-term virus evolution in nature. In: Domingo E (ed.), *Virus as Populations: composition, complexity, dynamics, and biological implications*. Boston: Academic Press, 2016, 227–62.

García-Murria MJ, Expósito-Domínguez N, Duart G *et al*. A bimolecular multicellular complementation system for the detection of syncytium formation: a new methodology for the identification of Nipah virus entry inhibitors. *Viruses* 2019;**11**:229.

Giménez E, Albert E, Zulaica J *et al*. Severe acute respiratory syndrome coronavirus 2 adaptive immunity in nursing home residents following a third dose of the Comirnaty coronavirus disease 2019 vaccine. *Clinl Infect Dis* 2022;**75**:e865–8.

Gozalbo-Rovira R, Gimenez E, Latorre V *et al*. SARS-CoV-2 antibodies, serum inflammatory biomarkers and clinical severity of hospitalized COVID-19 patients. *J Clin Virol* 2020;**131**:104611.

Grubaugh ND, Gangavarapu K, Quick J *et al*. An amplicon-based sequencing framework for accurately measuring intrahost virus diversity using PrimalSeq and iVar. *Genome Biol* 2019;**20**:8.

Grüning B, Dale R, Sjödin A *et al*. Bioconda: sustainable and comprehensive software distribution for the life sciences. *Nat Methods* 2018;**15**:475–76.

Grzelak L, Temmam S, Planchais C *et al*. A comparison of four serological assays for detecting anti–SARS-CoV-2 antibodies in human serum samples from different populations. *Sci Trans Med* 2020;**12**:eabc3103.

Hadfield J, Megill C, Bell SM *et al*. Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* 2018;**34**:4121–23.

Harvey WT, Carabelli AM, Jackson B *et al*. SARS-CoV-2 variants, spike mutations and immune escape. *Nat Rev Microbiol* 2021;**19**:409–24.

Hoffmann M, Kleine-Weber H, Schroeder S *et al*. SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor. *Cell* 2020;**181**:271–280.e8.

Hunt M, Hinrichs AS, Anderson D *et al*. Addressing pandemic-wide systematic errors in the SARS-CoV-2 phylogeny. bioRxiv, 2024

Ignatieva A, Hein J, Jenkins PA. Ongoing recombination in SARS-CoV-2 revealed through genealogical reconstruction. *Mol Biol Evolut* 2022;**39**:msac028.

Jackson B, Boni MF, Bull MJ *et al*. Generation and transmission of interlineage recombinants in the SARS-CoV-2 pandemic. *Cell* 2021;**184**:5179–5188.e8.

Kassambara A. *ggpubr: "ggplot2" Based Publication Ready Plots*. https://rpkgs.datanovia.com/ggpubr/ (R package version 0.6.0). CRAN, 2023

Kemp SA, Collier DA, Datir RP *et al*. SARS-CoV-2 evolution during treatment of chronic infection. *Nature* 2021;**592**:277–82.

Khare S, Gurry C, Freitas L *et al*. GISAID's role in pandemic response. *China CDC Weekly* 2021;**3**:1049–51.

Kirtipal N, Bharadwaj S, Kang SG. From SARS to SARS-CoV-2, insights on structure, pathogenicity and immunity aspects of pandemic human coronaviruses. *Infect Genet Evol* 2020;**85**:104502.

Köster J, Rahmann S. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics* 2012;**28**:2520–22.

Letko M, Marzi A, Munster V. Functional assessment of cell entry and receptor usage for SARS-CoV-2 and other lineage B betacoronaviruses. *Nat Microbiol* 2020;**5**:562–69.

Li H, Alkan C. New strategies to improve minimap2 alignment accuracy. *Bioinformatics* 2021;**37**:4572–74.

Li H, Birol I. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 2018;**34**:3094–100.

Li W, Moore MJ, Vasilieva N *et al*. Angiotensin-converting enzyme 2 is a functional receptor for the SARS coronavirus. *Nature* 2003;**426**:450–54.

MacLean OA, Lytras S, Weaver S *et al*. Natural selection in the evolution of SARS-CoV-2 in bats created a generalist virus and highly capable human pathogen. *PLoS Biol* 2021;**19**:e3001115.

McBroome J, Thornlow B, Hinrichs AS *et al*. A daily-updated database and tools for comprehensive SARS-CoV-2 mutation-annotated trees. *Mol Biol Evolut* 2021;**38**:5819–24.

McCallum M, De Marco A, Lempp FA *et al*. N-terminal domain antigenic mapping reveals a site of vulnerability for SARS-CoV-2. *Cell* 2021a;**184**:2332–2347.e16.

McCallum M, Walls AC, Sprouse KR *et al*. Molecular basis of immune evasion by the Delta and Kappa SARS-CoV-2 variants. *Science (New York, N Y)* 2021b;**374**:1621–26.

McCarthy KR, Rennick LJ, Nambulli S *et al*. Recurrent deletions in the SARS-CoV-2 spike glycoprotein drive antibody escape. *Science (New York, N Y)* 2021;**371**:1139–42.

McMillen T, Jani K, Robilotti EV *et al*. The spike gene target failure (SGTF) genomic signature is highly accurate for the identification of Alpha and Omicron SARS-CoV-2 variants. *Sci Rep* 2022;**12**:18968.

Meng B, Kemp SA, Papa G *et al*. Recurrent emergence of SARS-CoV-2 spike deletion H69/V70 and its role in the Alpha variant B.1.1.7. *Cell Rep* 2021;**35**:109292.

Minskaia E, Hertzig T, Gorbalenya AE *et al*. Discovery of an RNA virus 3′->5′ exoribonuclease that is critically involved in coronavirus RNA synthesis. *Proc Natl Acad Sci USA* 2006;**103**:5108–13.

Mishra T, Dalavi R, Joshi G *et al*. SARS-CoV-2 spike E156G/Δ157-158 mutations contribute to increased infectivity and immune escape. *Life Sci Alliance* 2022;**5**:e202201415.

Morgan M, Pagès H, Obenchain V *et al*. Rsamtools. Bioconductor. 2023.

Mykytyn AZ, Rosu ME, Kok A *et al*. Antigenic mapping of emerging SARS-CoV-2 omicron variants BM.1.1.1, BQ.1.1, and XBB.1. *Lancet Microbe* 2023;**4**:e294–5.

O'Toole Á, Scher E, Underwood A *et al*. Assignment of epidemiological lineages in an emerging pandemic using the pangolin tool. *Virus Evolut* 2021;**7**:veab064.

Paradis E, Schliep K, Schwartz R. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* 2019;**35**:526–28.

Perez-Florido J, Casimiro-Soriguer CS, Ortuño F *et al*. Detection of high level of co-infection and the emergence of novel SARS CoV-2 delta-omicron and omicron-omicron recombinants in the Epidemiological Surveillance of Andalusia. *Int J Mol Sci* 2023;**24**:2419.

Planas D, Veyer D, Baidaliuk A *et al*. Reduced sensitivity of SARS-CoV-2 variant Delta to antibody neutralization. *Nature* 2021;**596**:276–80.

Puhach O, Adea K, Hulo N *et al*. Infectious viral load in unvaccinated and vaccinated individuals infected with ancestral, Delta or Omicron SARS-CoV-2. *Nat Med* 2022;**28**:1491–500.

Qing E, Kicmal T, Kumar B *et al*. Dynamics of SARS-CoV-2 spike proteins in cell entry: control elements in the amino-terminal domains. *mBio* 2021;**12**.

Rambaut A, Holmes EC, O'Toole Á *et al*. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat Microbiol* 2020;**5**:1403–07.

Roemer C, Sheward DJ, Hisner R *et al*. SARS-CoV-2 evolution in the Omicron era. *Nat Microbiol* 2023;**8**:1952–59.

Tamura T, Ito J, Uriu K *et al*. Virological characteristics of the SARS-CoV-2 XBB variant derived from recombination of two Omicron subvariants. *Nat Commun* 2023;**14**:2800.

Tseng C-TK, Tseng J, Perrone L *et al*. Apical entry and release of severe acute respiratory syndrome-associated coronavirus in polarized Calu-3 lung epithelial cells. *J Virol* 2005;**79**:9470–79.

Turakhia Y, Thornlow B, Hinrichs AS *et al*. Ultrafast sample placement on existing tRees (UShER) enables real-time phylogenetics for the SARS-CoV-2 pandemic. *Nat Genet* 2021;**53**:809–16.

Uriu K, Ito J, Zahradnik J *et al*. Enhanced transmissibility, infectivity, and immune resistance of the SARS-CoV-2 omicron XBB.1.5 variant. *Lancet Infect Dis* 2023;**23**:280–81.

VanInsberghe D, Neish AS, Lowen AC *et al*. Recombinant SARS-CoV-2 genomes circulated at low levels over the first year of the pandemic. *Virus Evolut* 2021;**7**:veab059.

Venkatakrishnan AJ, Anand P, Lenehan PJ *et al*. Expanding repertoire of SARS-CoV-2 deletion mutations contributes to evolution of highly transmissible variants. *Sci Rep* 2023;**13**:257.

V'kovski P, Kratzel A, Steiner S *et al*. Coronavirus biology and replication: implications for SARS-CoV-2. *Nat Rev Microbiol* 2021;**19**:155–70.

Ward T, Glaser A, Overton CE *et al*. Replacement dynamics and the pathogenesis of the Alpha, Delta and Omicron variants of SARS-CoV-2. *Epidemiol Infect* 2022;**151**:e32.

Wells HL, Letko M, Lasso G *et al*. The evolutionary history of ACE2 usage within the coronavirus subgenus *Sarbecovirus*. *Virus Evolut* 2021;**7**:veab007.

Wickham H, Averick M, Bryan J *et al*. Welcome to the tidyverse. *J Open Source Softw* 2019;**4**:1686.

Yue C, Song W, Wang L *et al*. ACE2 binding and antibody evasion in enhanced transmissibility of XBB.1.5. *Lancet Infect Dis* 2023;**23**:278–80.

Zhang X, Chen -L-L, Ip JD *et al*. Omicron sublineage recombinant XBB evades neutralising antibodies in recipients of BNT162b2 or CoronaVac vaccines. *Lancet Microbe* 2022;**4**:e131.

Virus Evolution, 2025, 11(1), veaf015 , DOI: https://doi.org/10.1093/ve/veaf015, Advance Access Publication 11 March 2025, Research Article