ACCESS MICROBIOLOGY
an open research platform

RESEARCH ARTICLE
Gupta *et al.*, *Access Microbiology* 2023;5:000513.v3
DOI 10.1099/acmi.0.000513.v3

MICROBIOLOGY SOCIETY
OPEN DATA OPEN ACCESS

Assessing the evolution of SARS-CoV-2 lineages and the dynamic associations between nucleotide variations

Asmita Gupta, Reelina Basu and Murali Dharan Bashyam*

Abstract

Despite seminal advances towards understanding the infection mechanism of SARS-CoV-2 (severe acute respiratory syndrome coronavirus 2), it continues to cause significant morbidity and mortality worldwide. Though mass immunization programmes have been implemented in several countries, the viral transmission cycle has shown a continuous progression in the form of multiple waves. A constant change in the frequencies of dominant viral lineages, arising from the accumulation of nucleotide variations (NVs) through favourable selection, is understandably expected to be a major determinant of disease severity and possible vaccine escape. Indeed, worldwide efforts have been initiated to identify specific virus lineage(s) and/or NVs that may cause a severe clinical presentation or facilitate vaccination breakthrough. Since host genetics is expected to play a major role in shaping virus evolution, it is imperative to study the role of genome-wide SARS-CoV-2 NVs across various populations. In the current study, we analysed the whole genome sequence of 3543 SARS-CoV-2-infected samples obtained from the state of Telangana, India (including 210 from our previous study), collected over an extended period from April 2020 to October 2021. We present a unique perspective on the evolution of prevalent virus lineages and NVs during this period. We also highlight the presence of specific NVs likely to be associated favourably with samples classified as vaccination breakthroughs. Finally, we report genome-wide intra-host variations at novel genomic positions. The results presented here provide critical insights into virus evolution over an extended period and pave the way to rigorously investigate the role of specific NVs in vaccination breakthroughs.

DATA SUMMARY

All the raw sequencing data generated and used in this study have been submitted to the Sequencing Read Archive (SRA) with project accession ID PRJNA691556. The custom R codes and associated data can be accessed from the Github repository (https://github.com/asmitagpta/nCov19-seq). The accession IDs of all the samples submitted to the GISAID database has been provided in Table S2. Supplementary material can be found in Figshare: https://doi.org/10.6084/m9.figshare.23697210.v1 [1].

INTRODUCTION

In December 2019, a local outbreak of multiple cases of acute pneumonia [later classified as coronavirus disease 2019 (COVID-19, https://www.who.int/news-room/q-a-detail/coronavirus-disease-covid-19)] was reported in Wuhan, Hubei province, China, caused by a novel coronavirus, severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). The outbreak was soon followed by a rapid worldwide transmission which led to the World Health Organization (WHO) declaring COVID-19 as a global pandemic in March 2020. The rapid transmission rate of SARS-CoV-2 has resulted in 281808270 infections and >5 million deaths worldwide (as per WHO statistics, collected up to 29 December 2021). More importantly, 2021 witnessed the emergence of multiple virus variants [2, 3], of which five were classified as Variants of Concern (VoCs) by the WHO (https://www.who.int/en/activities/tracking-SARS-CoV-2-variants/) based on their higher transmissibility [4–7] and/or enhanced

Received 24 September 2022; Accepted 20 February 2023; Published 20 July 2023
Author affiliations: [1]Laboratory of Molecular Oncology, Centre of DNA Fingerprinting and Diagnostics, Hyderabad, India.
***Correspondence:** Murali Dharan Bashyam, bashyam@cdfd.org.in
Keywords: SARS-CoV-2 genome evolution; pathogenesis; vaccination breakthrough; COVID-19; mutational co-occurrence.
Abbreviations: COVID-19, coronavirus disease 2019; iSNV, intra-host single nucleotide variation; NV, nucleotide variation; RdRp, RNA-dependent RNA polymerase; SARS-CoV-2, severe acute respiratory syndrome coronavirus 2; SNV, single nucleotide variation; VoC, variant of concern; WHO, World Health Organization.
Eight supplementary figures and six supplementary tables are available with the online version of this article.

ability to escape neutralization by antibodies [8–13]. Of the five, the B.1.617.2 and B.1.1.529 lineages (designated, and hereinafter mentioned, as 'Delta' and 'Omicron', respectively, by the WHO) have exhibited maximum transmission during 2021, making them the predominant viral forms worldwide [14, 15]. Furthermore, a constant change in the frequencies of dominant viral lineages circulating in the population, arising from the accumulation of nucleotide variations (NVs) through favourable selection, is understandably expected to be a major determinant of disease severity and possible vaccine escape [16, 17]. Since host genetics is also expected to play a major role in shaping virus evolution [18], it becomes critical to study within-host diversity arising in the form of alternate alleles and corresponding genes harbouring them. Indeed, worldwide efforts have been initiated to identify specific virus lineage(s) and/or NVs that may cause a severe clinical presentation or facilitate vaccination breakthrough. As SARS-CoV-2 continues to evolve further into distinct lineages, with potentially increased pathogenicity and/or transmission abilities, it becomes imperative to study the emerging genomic variants, especially in the context of continual reports of possible immune escape and antibody neutralization imparted by a few NVs [17, 19].

The current study was initiated to analyse the sequences generated from SARS-CoV-2-infected samples (including 210 from our previous study), collected over an extended course of 19 months (from the beginning of the pandemic to October 2021) from the state of Telangana, India. We present the major trends of dominant lineages propagating in Telangana and corresponding trends in pan-Indian regions and worldwide, across this period. We also provide a comprehensive map of all NVs in the viral genome during this period and their possible association with vaccine escape. Finally, we have analysed specific virus genomic positions displaying intra-host diversity.

METHODS

Sample collection strategy, dataset structure and features

A total of 3543 samples [1407 females and 2091 males (information unavailable for 45 samples)], representing the period 1 April 2020 to 31 October 2021, from Telangana, India, were analysed in this study (Table S1A available in the online version of this article). The sample collection strategy for the period April 2020 to February 2021 was unchanged from our previous study [20]. For the subsequent period, nasopharyngeal/oropharyngeal swabs were collected from several reverse transciptase (RT)-PCR-based testing centres as well as multi-speciality hospitals across Telangana, as per guidelines established by the Indian SARS-CoV-2 Genome Consortium (INSACOG) [21]. In addition, samples received in the Covid-19 testing laboratory in CDFD, Hyderabad, were also included in the study. The work was initiated following approvals from the Institutional Bioethics committee and Biosafety committee. Sample collection peaked during the months of June and July 2020, followed by a hiatus, and subsequently increased from March 2021 onwards, roughly coinciding with the first and second waves of the pandemic, respectively. Of the total cases, 360 represented the age group <18 years, 2674 represented the age group 18–60 years and 55 represented the age group >60 years, while age was not documented for 154 cases. The dataset also comprised two independent sets of cases belonging to local isolated transmission clusters (so-called 'super-spreader' events) (Table S1A).

Cases were classified as vaccination breakthroughs if they reported infection ≥14 days after receiving a second dose of either ChAdOx1 [22] (commercial name Covishield, based on recombinant, replication deficient chimpanzee adenovirus vector, developed at Oxford University, Oxford, UK) or BBV152 [23] (commercial name Covaxin, whole virion inactivated Vero cells developed by Bharat Biotech Ltd, India) vaccine or ≥21 days of receiving a first dose [22] [24] [25]. The dataset included a total of 313 vaccination breakthrough cases, of which 244 came from Telangana, 18 Uttar Pradesh (obtained from the Banaras Hindu University) and 51 to Chennai, Tamil Nadu (obtained from the Department of Public Health and Preventive Medicine, State Public Health Laboratory) (Table S1B, S1C). Of these, 154 were completely vaccinated and 149 had received only one vaccination dose, while the status of 10 was unavailable. The majority of cases (228/313; 72.8%) received the ChAdOx1 vaccine, while a small proportion (36/313; 11.5%) received the BBV152 [23] (Table S1 A–C). Vaccine identity was unknown for 49 samples. The vaccination dates were unavailable for 12 partially vaccinated and two completely vaccinated cases and therefore these 14 samples were excluded during NV analysis of vaccination breakthrough cases

For pan-India and worldwide analyses of widespread lineages, a dataset comprising 31546 (India) and 678438 (world) consensus genomes (*fasta* files) for the period March 2021 to October 2021 were accessed from the publicly available Global Initiative on Sharing All Influenza Data [26] (GISAID, https://www.gisaid.org/) repository; accession IDs for all the sequences submitted to GISAID from this study are listed in Table S2.

SARS-CoV-2 RNA extraction and sequencing

Total RNA was isolated in a Biosafety level 2+ (BSL-2+) environment following standard protocols using the RNA isolation kit (MagRNA-II viral RNA extraction kit, Cat. No. G2M030620; Genes2Me; molecular RNA extraction kit, Cat. No. COVEX 100PS; Q-lineBiotech; nucleic acid extraction kit, Cat. No. A200-96; Zybio Inc.) as per the manufacturers' instructions. The RNA extraction and sequencing protocol has been described in our previous study [20]. Briefly, each RNA sample was subjected to RT-PCR for Envelope (E) and RNA-dependent RNA polymerase (RdRp) genes using the nCoV-19 RT-PCR detection kit (Cat. No. NCoV-19ER100PS; Q-lineBiotech) or the ViralDetect-II multiplex real-time PCR kit for COVID-19 (which also detected

the N-gene; Cat. No. G2M020220; Genes2Me). As mentioned in our previous study, since RdRp consistently provided more robust amplification than the E-gene, we considered Ct (threshold cycle) values of RdRp alone for analysis. Samples exhibiting a Ct value of <30 (E and RdRp genes) were selected for whole genome sequencing.

The isolated RNA was reverse transcribed using random primer mix (New England Biolabs) and Superscript-IV (Thermofisher Scientifi). The synthesized cDNA was amplified using a multiplex PCR protocol, producing 98 amplicons across the SARS-CoV-2 genome (https://artic.network/, primer version V3, https://www.protocols.io/view/ncov-2019-sequencing-protocol-v3-locost-bp2l6n26rgqe/v3). The amplified products were processed for tagmentation and indexing PCR for Illumina Nextera UD Indexes Set A, B, C, D (Illumina) (384 indexes, 384 samples). All samples were processed as 96-well plate batches that consisted of one each of COVIDSeq positive control HT (CPC and, quantified (Qubit 2.0; Invitrogen) and fragment sizes were analysed in Agilent Tapestation 4200 (Agilent Technologiess). The pooled library was further normalized to 10 nM concentration and 10 µl of each normalized pool containing index adapter set A, B, C and D was combined in a new microcentrifuge tube to a final concentration of 2 nM. The pooled libraries were denatured and sequencing was performed on a Nextseq 2000 using the P2 100 Cycle kit with 1×101 bp sequencing chemistry. About 50–100 Mb of data were generated for each sample.

In silico workflow for processing genome sequencing data

The raw sequencing data in *fastq* format were subjected to quality checks including filtering low-quality reads, determination of sequencing depth and adapter trimming using Trimmomatic [27]. All reads shorter than 30 bases or with an average Phred quality score <20 were discarded. Eighteen samples were rejected because of low overall sequencing depth and poor quality. The trimmed reads were then aligned to the reference Wuhan sequence (NCBI ID NC_045512.2) using the bwa-mem [28] algorithm. Post-alignment filtering and quality assessment was performed using samtools [29]. Single nucleotide variants (SNVs) were identified using iVar [30] which works on the *mpileup* output from samtools. The variants were annotated using SnpEff [31] and further filtered to remove all problematic sites documented to be prone to accumulate sequencing errors by multiple sources as recommended earlier (https://virological.org/t/issues-with-sars-cov-2-sequencing-data/473) [32]. Sequences with high coverage, low N content and associated with complete metadata were included for further analysis. Reads were assembled to generate consensus *fasta* file using samtools mpileup and the consensus module of iVar with a base assigned as consensus if it had a minimum depth of at least 10 reads (setting ivarMinDepth=10). Sequences where the N content was >30% and sequence length <27000 were rejected. Lineage assignment was done on consensus *fasta* files using Pangolin [33] v3.1.16.

All alleles with an allele frequency of >90% were classified as NVs. For analysis of NV cross-correlation, a methodology similar to the one reported in our previous study [20] was used. Briefly, a binary matrix was constructed for each sample with all NVs found in >5% samples as columns, indicating whether an NV of interest was present or absent in a sample. NVs exhibiting low standard deviation across samples were not included in the analysis. Pairwise Pearson correlation coefficients were estimated for this binary matrix using the *cor* function in R. *P*-values indicating the significance of association between each pairwise correlation coefficient was estimated using the *cor.test* function with *chi-square* test. This matrix was then used to visualize the NV cross-correlation maps in R with the *corrplot* function [34]. Odds ratios for estimating the association likelihoods of genomic alterations with vaccination breakthrough cases were estimated by creating contingency matrices for each NV identified in >5% of vaccinated samples and were compared with multiple random subsamples of non-vaccinated cases from March 2021 onwards.

The estimation of intra-host single nucleotide variations (iSNVs) was carried out using Lofreq [35], a variant caller with a high sensitivity to predict iSNVs with an allele frequency as low as 1% [35]. Minor alleles with allele frequencies between 2 and 50%, and minimum sequencing depth of 100×, were classified as iSNVs, and were used for further analysis.

All statistical tests and analyses were performed using custom R scripts. All structural representations were generated in PyMOL (The PyMOL Molecular Graphics System, Schrödinger, LLC).

RESULTS

B.1.617.2 ('Delta') displaced all previously circulating lineages from March 2021 onwards

We identified a clear shift in the dominant lineages present in Telangana, India, from 2020 to 2021 (Figs 1a and S1a; the complete distribution of all lineages in the dataset is provided in Table S3). The B.1.1.306 and B.1.1.326 lineages (both detected first in India and considered to have transmission links to Zambia, Somalia and Bahrain (https://cov-lineages.org/lineage_list.html) were dominant from May 2020 to September 2020. The period October 2020 to March 2021 witnessed an upsurge of B.1.36.29 (assigned as 'Indian Lineage' by Pangolin [33] and pangoLEARN [36]). December 2020 witnessed the emergence of the Kappa (B.1.617.1) lineage in the state population, and was present until April 2021. However, 'Alpha' (B.1.1.7), which was the first lineage classified by WHO as a VoC, appeared in the state population in February 2021 and was identified in samples until March 2021, consistent with other reports [37]. Spread of the Alpha lineage was higher and lasted longer in northern India, appearing as early as January 2021, and present until May 2021, while southern India witnessed a higher prevalence of the Kappa variant (B.1.617.1) (Fig. S1a,b). However, from March 2021 onwards, the Delta variant (B.1.617.2) constituted the major lineage detected in the

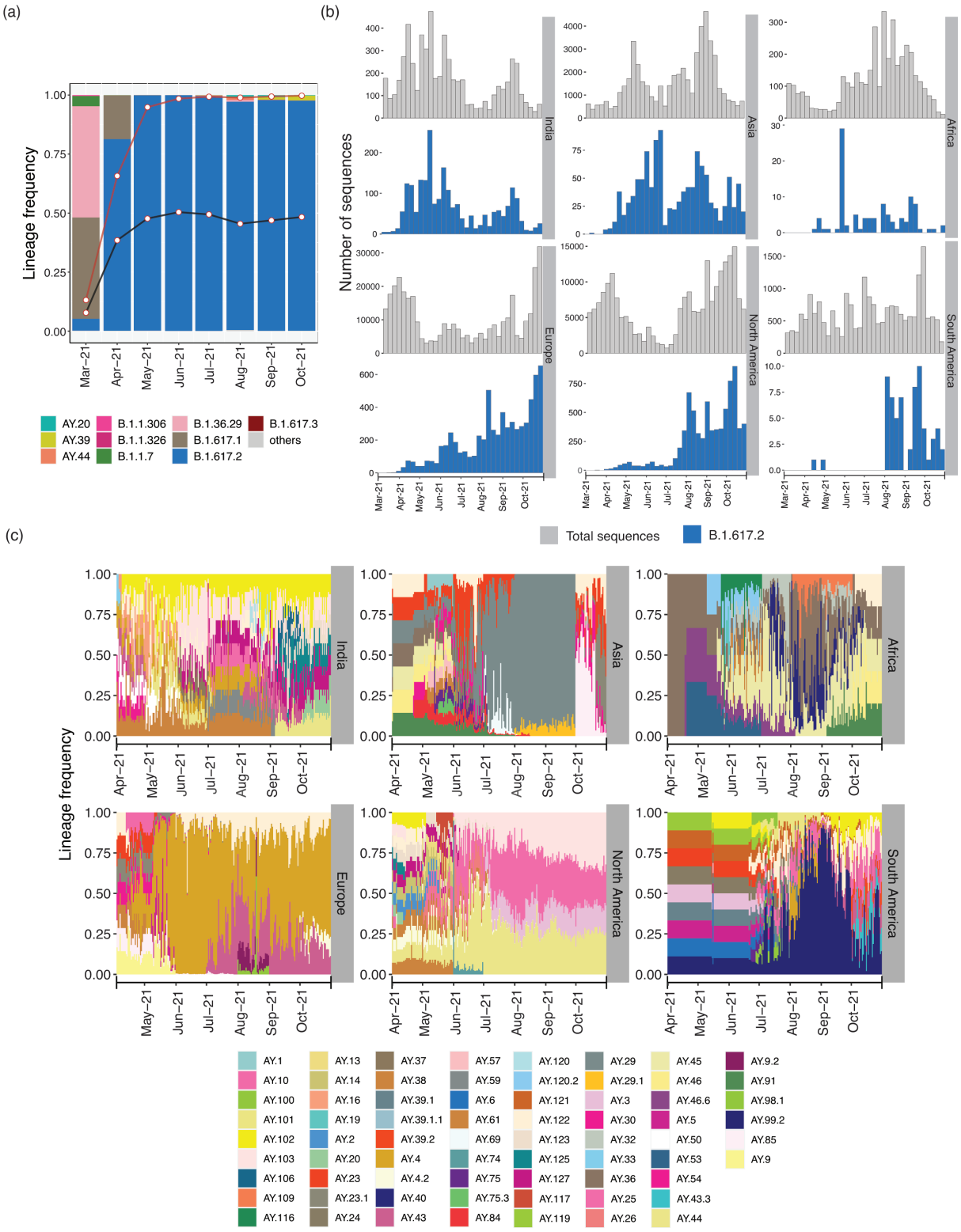

Fig. 1. Comparative timeline of major SARS-CoV-2 lineage distribution in Telangana, India. (a) Frequency distribution of the major lineages per month from March 2021 (a complete timeline starting from April 2020 is provided in Fig. S1a). The black and red lines show the frequency distribution trend of Delta and Delta plus its sub-lineages in India (excluding Telangana), respectively; white points show the corresponding monthly frequency. (b) Timeline of Delta variant distribution (blue) and total (grey) cases in India and rest of the world (all data were obtained from GISAID). (c) Timeline of changing frequency of Delta sub-lineages across the world. Only those sub-lineages present in >5% of total sequences submitted from the region were included. In (b) and (c), while estimating frequencies for the Asian countries, Indian sequences were excluded.

state, replacing all other virus lineages circulating previously, indicative of its massive spread (Figs 1a and S1a). As expected, the extent of its spread in Telangana almost paralleled its surge in the rest of the country [38] (Fig. S1b). Lineage analyses on the two sample sets representing local transmission clusters (Table S1a) did not reveal enrichment of a specific lineage compared to other samples analysed from the same period in parallel (data not shown). The spread of Delta in India pre-dated its spread in other countries (Fig. 1b). Following its gradual rise in Asia, the Delta lineage spread to Europe, North America and South America, with the highest spike in June–August 2021 (Fig. 1b). By the end of June 2021, Delta was the major lineage in all geographical regions of the world [21].

In July 2021, the Delta variant was further classified (by Pangolin) into several sub-lineages (https://cov-lineages.org/) indicative of appearance of additional NVs (https://outbreak.info). From July 2021 onwards, AY.20, AY.39 and AY.44 were the major Delta sub-lineages observed in Telangana (Fig. 1a), though the fraction of Delta sub-lineages was higher in most Indian states compared to Telangana (Fig. S1b). Moreover, the major sub-lineages in other states varied from July 2021 onwards (Fig. S1b).

The distribution of Delta sub-lineages also differed worldwide (Fig. 1c). From mid-April 2021 onwards, AY.4 was most prominent in Europe, followed by AY.43 from July 2021; by contrast, AY.103, AY.44, AY.3 and AY.25 were more common in North America from June 2021. The sub-lineage AY.29 was highly prevalent in Oceania from July 2021 (Fig. S1c), while Asian countries (excluding India) displayed a mix of various sub-lineages, with AY.29 becoming prevalent between June and September 2021 (Fig. 1c). Similar to the distribution pattern observed in India, a mix of different Delta sublineages including AY.36, AY.40, AY.45, AY.46 and AY.91 was observed in sequences from Africa during May–October 2021. Given the consistently increasing repertoire of Delta sub-lineages, constant lineage re-assignment by Pangolin and the fact that a mere assessment of the lineages per se may be insufficient to understand the complete catalogue of SARS-CoV-2 NVs, we performed a detailed analysis of NV profiles obtained from the samples.

Genome-wide analysis reveals a sudden change in the landscape of SARS-CoV-2 NVs coinciding with the second wave

Study of NVs not included in the NV signature of a lineage but that nevertheless could have arisen either due to positive selection or genetic drift may be critical from the perspective of public health surveillance. A monthly assessment revealed a significant shift in the NVs observed in Telangana during the course of the previous year (Fig. S2). Overall, a few NVs were consistently present in the population starting from April 2020, namely A23403G (D614G, S), C241T, C3037T (P924F, ORF1a: nsp3) and C14408T [P4720L (alternatively referred to as P323L), ORF1b: nsp12] (Fig. S2). Genes coding for envelope and membrane proteins (E and M genes, respectively) and the accessory protein ORF6 remained relatively free from high-frequency NVs until October 2021. The genomic landscape of the virus was marked by the presence of a few high-frequency NVs in the period from April 2020 to July 2020 (Fig. S2). The period from August 2020 to February 2021 witnessed the appearance and subsequent disappearance of many moderately frequent NVs, especially in ORF1ab. However, from March 2021 onwards, a larger number of NVs became evident at moderate to high frequencies throughout the genome (Fig. S2). Specifically, the S protein, nucleocapsid (N), ORF7a/b and nsp3 accumulated several missense NVs (Fig. 2). The S protein NVs included T19R, T95I, G142D, del156-157, A222V, L452R, T478K, D614G, P681R and D950N, along with the ubiquitous D614G (Fig. 2). With the exception of T95I, G142D and A222V, all other NVs were consistently present in >75% of samples until October 2021 and were further associated with the appearance of the Delta lineage in the population (Fig. S3). Beginning in May 2021, additional NVs appeared in the S protein, which resulted in bifurcation of Delta into various sub-lineages (Figs 2 and S3). Notably, the frequency of T95I was highest in AY.20 compared to Delta and other sub-lineages while the frequency of G142D was highest in the Kappa lineage (B.1.617.1) compared to other lineages observed thereafter. A222V was another alteration shared by both Delta and AY.44 though it was present in higher frequencies in sub-lineage AY.44. Interestingly, an observation which clearly stood out from this analysis is that sub-lineages AY.20, AY.39 and AY.44 harboured several NVs [including M153I, A243P, L244F, V1104L and V1176F (in AY.20), S221A and A1080S (in AY.39), G181V (AY.20, AY.39), and T1117I and D1260E (in AY.44)] in the N-terminal domain (NTD), receptor binding domain (RBD) and the region between the two heptapeptide repeat sequences HR1 and HR2, which were either absent or present in lower frequencies in Delta (Fig. S3).

A striking observation in the N protein was the disappearance of RG203KR (triplet 28881-3GGG>AAC) from January 2021 onwards (Fig. 2). We observed the simultaneous appearance of R203M in Alpha as well as in Delta and its sub-lineages. Furthermore, the Delta lineage was marked additionally by N protein NVs D63G, G215C and D377Y. More importantly, several N protein NVs were restricted to Delta sub-lineages (being absent in Delta itself) including L161F, K361Q and K369T (AY.39), and A252S (AY.20) (Fig. S3).

Among the non-structural proteins that constitute ORF1a/b, nsp3 is the largest, and exhibited an increase in frequency of several NVs (I880V, A1306S, P1640L, P2046L, P2287S) from March, 2021, most of them being associated with Delta and its sub-lineages (Figs 2 and S3). The accessory proteins 7a and b, relatively free of NVs until December 2020, accumulated distinct high-frequency NVs including V82A (7a), T120I (7a) and T40I (7b) from January 2021 (Figs 2 and S3). Similarly, the NV landscape of ORF3a was significantly altered post-December 2020 and ORF1b:nsp12 (RdRp) accumulated the G5068S NV in addition to the ubiquitously

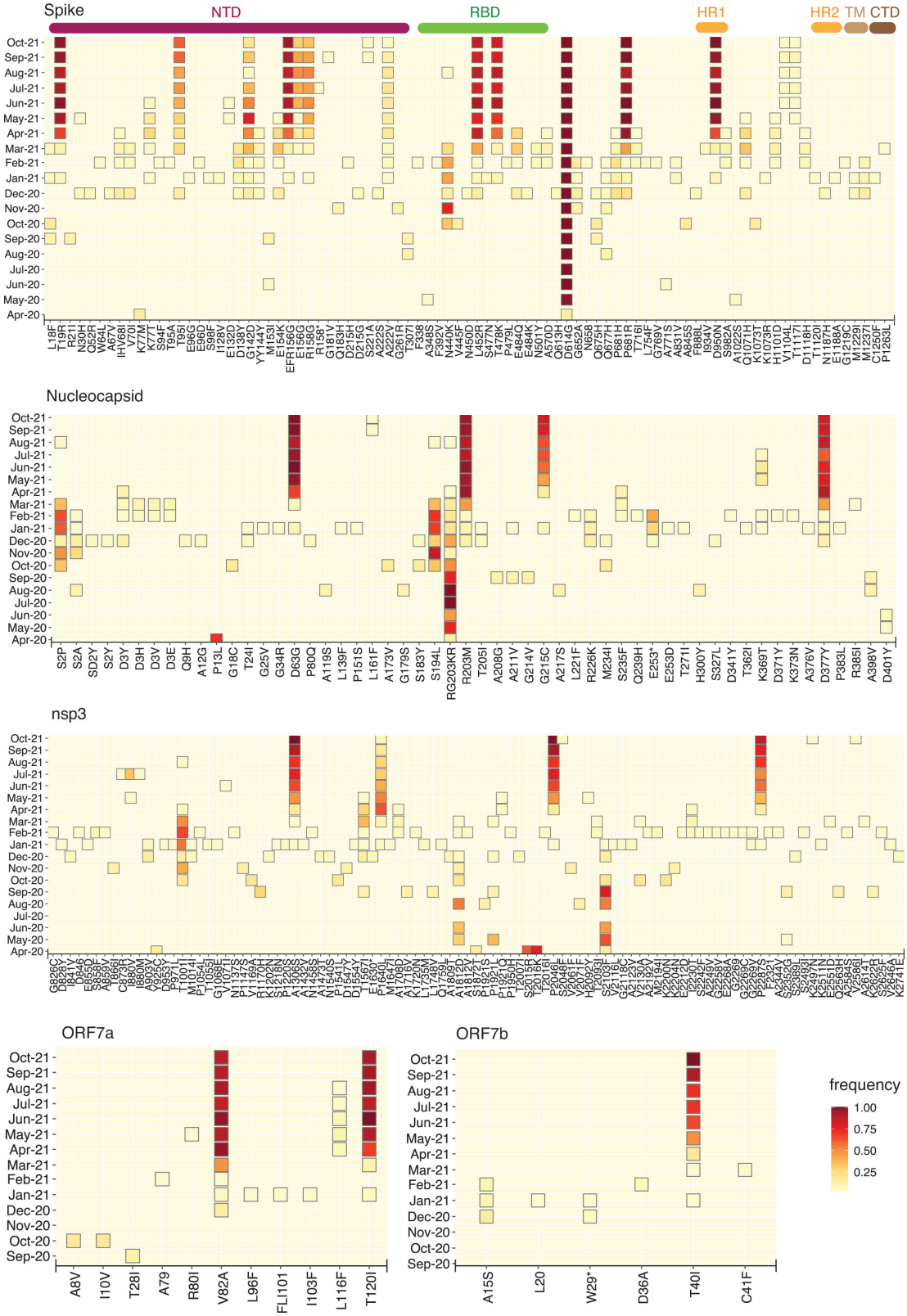

Fig. 2. Monthly timeline of SARS-CoV-2 NV frequencies; variants present in <3% of samples were excluded. Key domains in the S protein (first panel) are indicated at the top: NTD, N-terminal domain (brown); RBD, receptor binding domain (green); S2 subunit containing SD1 (subdomain 1), SD2 (subdomain 2) and S1/S2 cleavage sites (orange); CTD, C-terminal domain (black). NV frequencies for N protein (second panel), nsp3 (third panel), and ORF7a and ORF7b (bottom panels) are also shown.

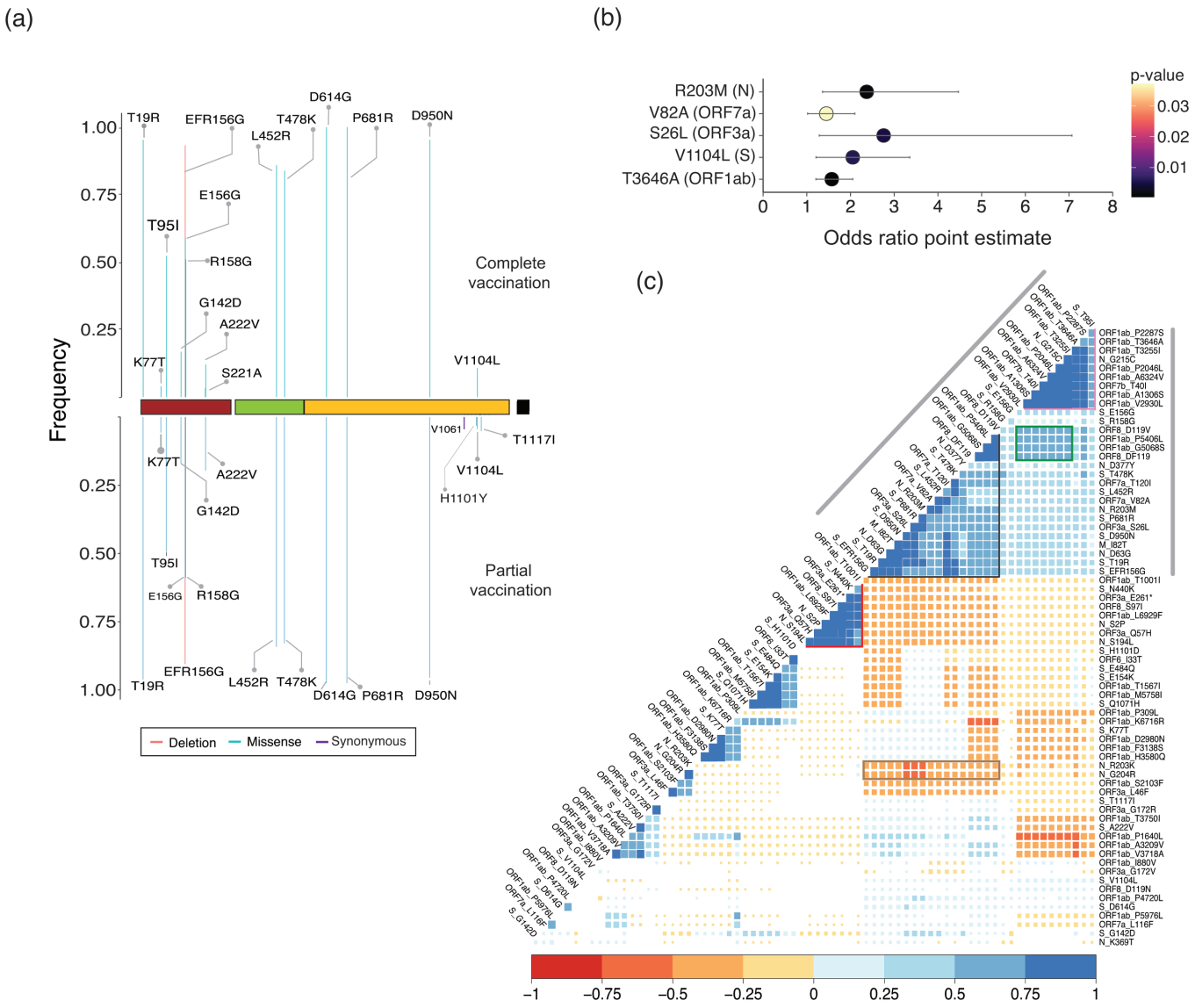

Fig. 3. NV frequency in vaccinated cases and pairwise cross-correlation analysis. (a) Frequency of S protein amino acid alterations present in >3% of completely (top, calculated for a total of 152 samples) and partially (bottom, total of 137 samples) vaccinated cases. (b) Odds ratio indicating the extent of association of specific variants with vaccination breakthrough cases. (c) Pairwise cross-correlation plot between all non-synonymous missense NVs present in >3% of all the SARS-CoV-2 samples identified from Telangana, India, during April 2020 to October, 2021; the size of all coloured 'squares' is inversely proportional to the corresponding *P*-value of the correlation. Positions with *P*>0.05 appear as blank (or white). The colour key for positive (blue) and negative (red) correlation is given below the plot.

present P4720L (Fig. S4). The frequencies of all these NVs in ORF7a/b, ORF3a and nsp12 were consistently high in samples classified as either Delta or its sub-lineages AY.20, AY.39 and AY.44, potentially reflecting an extended genomic variation footprint not observed in previous lineages (Figs S3 and S4). Furthermore (and similar to the observations made with respect to lineages), an inspection of the NVs in the samples collected from two local 'super-spreader' events (Table S1A) did not reveal enrichment of any specific NVs.

In a nutshell, the Delta sub-lineages were marked by the presence of several NVs, over and above those that defined the Delta lineage. Second, AY.20 and AY.39 had relatively higher numbers of NVs compared to other sub-lineages observed in the state. Since the emergence of vaccine breakthrough cases coincided roughly with the emergence of Delta and its sub-lineages, we further investigated whether the lineages or the NVs associated with them showed a higher likelihood of occurrence in vaccination breakthrough cases.

Association of specific genomic NVs with vaccination breakthrough cases

A large fraction (70%) of vaccination breakthrough cases belonged to the Delta variant, followed by its sub-lineages AY.44 (5.2%), AY.20, AY.43 and AY.39 (5% each) (Fig. S1d). Interestingly, 57, 39, 29 and 22% of all samples classified as AY.32, AY.43, AY.35 and AY.20 respectively, belonged to vaccination breakthrough cases (Fig. S1d). We next analysed all S protein NVs present in >3% of vaccinated cases. T19R, deletion EFR156G (156-157del), L452R, T478K, D614G, P681R and D950N, present in high frequency in the total dataset, were also present with highest frequencies (>80%) in vaccinated cases (Fig. S5), as expected. In addition to these, T95I, which was present in slightly higher frequencies in samples belonging to AY.20 compared to Delta, was found in 50.8% of all vaccinated samples. Another S protein alteration, V1104L (located in the S2 subunit), was present in ~7.5% of all vaccinated cases. Intriguingly, the frequency of this NV was higher in AY.20-associated samples than in Delta itself (Fig. S3). Separate analyses of partial and completely vaccinated cases identified S221A (2.3% cases) as an exclusive event in completely vaccinated cases (Fig. 3a). Similarly, T1117I (~3.9% cases) and H1101Y (~1.9% cases) were identified exclusively in partially vaccinated cases (Fig. 3a). We computed odds ratios to estimate the significance of association of genome-wide NVs with vaccination breakthrough cases. V1104L (S), S26L (ORF3a), V82A (ORF7a), R203M (N) and T3646A (ORF1ab) exhibited odds ratios of 1.5–3.0 [P<0.05, at 95% confidence interval (Fig. 3b); details of upper and lower confidence interval limits are provided in Table S4], thereby providing further support towards favourable occurrence of these NVs in vaccination breakthrough cases

NV cross-correlation maps reveal presence of extended NV signatures associated with different lineages

An NV (missense only) cross-correlation map arranged by hierarchical clustering revealed several clusters highlighting the frequency of co-occurring NVs within samples. The first major cluster immediately apparent from the cross-correlation map (Fig. 3c, grey lines at the sides) was formed by an extensive set of co-occurring NVs that highlight the Delta lineage (https://outbreak.info/situation-reports/delta). Within this large cluster, however, we identified the presence of two smaller (sub) clusters. One of these (black outline; Fig. 3c) encompassed S protein NVs P681R, L452R, T478K, D950N and T19R which were observed in high frequency in samples belonging to the Delta lineage (Fig. S3). This sub-cluster additionally displayed enriched co-occurrence of NVs in other genes, namely T120I, V82A (ORF7a), S26L (ORF3a), del 119/120 (ORF8), I82T (M) and D63G and D377Y (N), indicative of a potential extended genomic signature of the Delta lineage. Interestingly, this sub-cluster was mutually exclusive with another set of NVs [N440K (S), S2P, S194L (N), Q57H, stop gained E261* (ORF3a)] that were part of the B.1.36.29 lineage (Fig. 3c, the lineage defining NVs highlighted by a red outline) which was abundant before the emergence of Delta as mentioned above (Figs. 2 and S3). The Delta-defining cluster was also negatively correlated to the double amino acid changes in N protein R203K and G204R (brown rectangle; Fig. 3c) that was prevalent before the emergence of Delta, indicating a completely unique signature formed by NVs in Delta, mutually exclusive to that of previously circulating viral lineages in the population.

The second sub-cluster (pink outline; Fig. 3c) included T95I (S), G215C (N), T40I (ORF7b) and several NVs located in ORF1ab including A1306S, P2046L, P2287S (nsp3), T3255I (nsp4), T3646A [nsp6; also favourably associated with vaccination breakthrough cases (Fig. 3b)] and A6324V (nsp14). These NVs were found to be present in a large proportion (>90%) of samples associated with Delta sub-lineages AY.20, AY.39 and AY.44 (Fig. S3). Though these NVs also exhibited positive correlation with the NVs in the Delta sub-cluster (Fig. 3c; black outline) a stronger correlation among the NVs within this new cluster points to a divergent evolution of these sub-lineages in the population. Furthermore, several NVs associated within this Delta sub-lineage sub-cluster displayed positive associations with a set of NVs formed by ORF8 (D119V, del 119/120) and ORF1ab (P5406L, G5068S) (green rectangle; Fig. 3c), reflecting an additional set of co-occurring common NVs occurring in Delta sub-lineages (AY.20, AY.29, AY.44) (Fig. S4).

We extended the analysis to vaccine breakthrough cases and observed a positive cross-correlation between S26L (ORF3a), P4720L (ORF1ab) and D614G (S) (correlation coefficient r^2>0.75) which was absent in non-vaccinated cases (Fig. S6). Interestingly, S26L (ORF3a) also exhibited highly significant association with vaccination breakthrough cases (Fig. 3b). Another positive cross-correlation was observed among V82A (ORF7a) and S protein NVs, L452R, T478K (r^2>0.75, as opp and, T95I (r^2 >0.5; <0.5 in non-vaccinated cases). Finally, another instance of increased positive correlation in vaccination breakthrough cases was formed by R203M (N), V82A and T120I (ORF7a) compared to non-vaccinated cases (Fig. S6). This aligns well with our previous observation, where both R203M and V82A displayed favourable odds of occurrence in vaccination breakthrough cases (Fig. 3b). Overall, the findings suggest that the genomic alterations in the S protein (L452R, T478K, P681R) which have previously been reported to increase escape from neutralizing antibodies and potentially associate with vaccination breakthroughs [8, 11] might arise in tandem with genomic changes located in non-S genes which are involved in modulating the downstream processes in viral life cycle (ORF1ab, N) and regulating the host immune response (ORF7a/b, ORF3a), thus increasing the probability of virus survival and causing breakthroughs.

SARS-CoV-2 exhibits genomic plasticity as multiple sites contribute to generation of intra-host variants

Since virus genomic diversity arises within a host, it becomes important to uncover minor alleles originating in the form of iSNVs. A total of 545 samples (15.4% of the total dataset) exhibited iSNVs, of which a majority [234; 43% (6.6% of the total dataset)] harboured minor allelic variants at a single genomic position and 165 samples [30% (4.6% of the total dataset)] exhibited iSNVs

at two positions (Fig. 4a). Interestingly, 11 samples contained more than nine sites with minor alleles. However, the Ct values in seven of these 11 cases were >22 (Table S5) indicating possible false calls due to sequencing error(s), as reported earlier [39–41]. The presence ofmore than nine iSNV sites in the remaining three samples could probably reflect mixed infections, as reported earlier [39]. N, nsp11, nsp9 and ORF3a exhibited a higher frequency of iSNVs (when normalized to gene size) compared to other proteins (Table S6).

We next endeavoured to map the co-occurrence of NVs and iSNVs in the genome. Across the 545 samples, a total of 229 genomic loci were involved in iSNVs, of which 113 loci were found to also share NVs (Fig. 4b). This observation stands in contrast to the total number of genomic loci which formed NVs across the dataset (5224), of which only 229 sites shared iSNVs. To investigate whether any iSNV site(s) which shared NVs coincided with mutation 'hotspot' region(s), we mapped all genomic positions which shared iSNV and NVs and estimated the sample frequencies of major and alternative alleles at these loci (Fig. S7). The analysis revealed that a very few [9053 (nsp4), 11201, 11332, 11418 (nsp6), 21618, 23403, 23604 (S), 26767 (N) and 28881 (N)] shared sites exhibited NVs (major alleles) which were widespread (sample frequency >45%) in the dataset, suggesting that these loci could be mutation hotspots.

Another critical and arguably more important insight that iSNV analysis provides is by enabling effective temporal tracking of a novel minor allele from the time of its emergence. This can potentially identify an important NV before it becomes widespread in the population, and hence being of great benefit for public health surveillance. To this end, we evaluated changes in sample and allele frequencies of minor alleles identified in iSNV sites (Fig. 4c). The results revealed few iSNVs in ORF1ab (683T, 5584G, 12946C and 19684T), ORF3a (25489T, 25506T, 25506G, 25677T, 25855C and 25907T) and multiple positions in N (28290T, 28725T, 29000T, 29032T, 29034T, 29140T and 29179T) where the sample and allele frequencies consistently increased from the time of their emergence (Fig. 4c). Allele frequencies of minor alleles at all these iSNV positions (with a few exceptions discussed below) reached 80% over time, attesting to a transition from iSNV to NV. Furthermore, most of these iSNVs emerged in January 2021 (with a few emerging in December 2020 or February 2021), indicating a possible escape from a transmission bottleneck. A few notable examples include 25907T (leading to 172V in ORF3a, present in 53% of the total samples) and 29034T in N (leading to 254V, present in 24% of all samples).

In addition, a few iSNVs exhibited a significant increase in sample frequency across a long time period, despite exhibiting an allele frequency of <50% (Fig. 4c). A notable example was all alternative alleles at position 5270 (nsp3), which exhibited an allele frequency <16%, but were nonetheless detected in a significant fraction of samples (across the 3543 samples, 5270A in 33% cases, 5270C in 20%, 5270G in 5%) (Fig. 4c). Another such instance was observed at position 29033 (N; 29033T in 6%, 29033A in 1.2%) (Fig. 4c).

We next focused on S protein iSNVs, expected to be important in shaping virus transmission and immune escape. In addition, S protein iSNVs are reported to be rare events due to evolutionary constraints with many becoming lost, attesting to a narrow transmission bottleneck [42]. A few minor alleles, 19I (C21618), 30H (A21650), 879S (G24197), 1101Y (C24863), 1118Y (C24914) and 1252C (C25317), which were first observed in independent samples in January 2021, were consistently present in samples until October 2021 as well (sample distribution and allele frequency shown in Fig. 4c). Among these, the S protein iSNV-associated minor alleles that exhibited an increased sample frequency from the time of their emergence were 30H, 1101Y and 1252C (the distribution of their allelic frequencies in shown in Fig. 4c). However, the allelic frequencies of 30H and 1252C, in all the samples in which they were detected, was <50% (with the exception of one sample which harboured the 1252C allele at an allele frequency >75%). By contrast, from June to September 2021, the sample frequency of 1101Y increased from 0.6 to 3%, with a concordant increase in corresponding allele frequency (Fig. 4c). Interestingly, the H1101Y NV was also present in >5% of all partially vaccinated samples. Subsequent analysis on the potential functional impact of these iSNVs is currently underway.

DISCUSSION

SARS-CoV-2 continues to cause significant morbidity and mortality worldwide, making it important to perform regular genomic surveillance to detect emerging virus NVs. The estimated mutation rate of SARS-CoV-2 is about 1.1×10^{-3} substitutions per site per year [43]. The constant virus evolution leading to emergence of new variants is shaped by several factors including host genetics and immune response [44, 45], strong selection pressure created due to neutralizing antibodies [46], etc. The infection rate reduced significantly across India after peaking in August 2020. However, India witnessed a ferocious 'second wave' of infection during February to June 2021, with the number of deaths several fold higher than that observed in the first wave [47]. This massive spread has mostly been attributed to the Delta variant, which has been shown to be associated with higher transmissibility rates than previously circulating Alpha and Kappa lineages [48, 49]. The Delta variant rapidly displaced all previously circulating viral lineages owing perhaps to its increased fitness [R_0 (R_e) 60–70%] with viraemia 1000-fold higher than most previous lineages [49]. The increased adaptability of a pathogen virus under 'waning' immune pressure or partial immunization has been reported in other studies [50]. While other widely transmitted lineages such as Alpha emerged in other countries before vaccination programmes were implemented [6], the countrywide spread of Delta probably

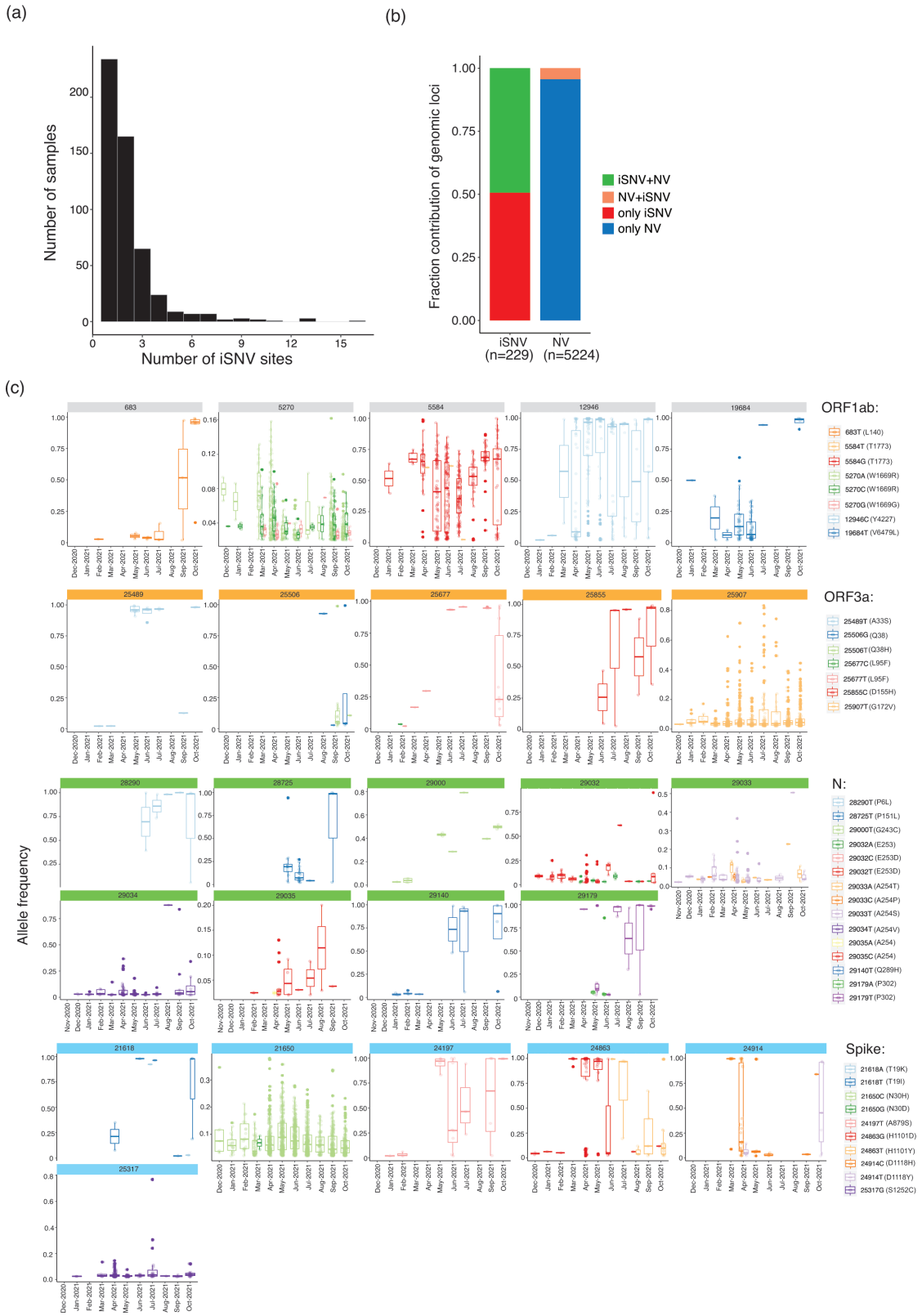

Fig. 4. SARS-CoV-2 iSNVs identified from Telangana, India. (a) Distribution of number of iSNVs in samples. (b) Distribution of genomic loci with shared and unique iSNVs and NVs. (c) Timeline of allele frequency changes in the minor alleles in ORF1ab, ORF3a, N and S proteins. Box plots indicate allelic frequency distribution while points represent samples in which the allele was identified. Only those alternate alleles whose allele frequencies were either consistent or increased during the indicated timeline are shown.

occurred during the implementation of vaccination in India. This suggests that higher mutation accumulation observed in this variant could be a result of increased selection pressure under a modified host microenvironment to which the virus was exposed. This suggestion assumes significance given the Delta variant's ultrafast replication speeds which makes it detectable within 4 days after exposure [49]. Despite the preponderance of Delta during the second wave, the spread of specific virus lineage(s)/sub-lineages did exhibit differences in various geographical regions within India. Under these circumstances, it becomes imperative to study the evolution of the virus within a specific demography over an extended period.

From June 2021 onwards, the Delta variant itself was divided into several sub-lineages (based on Pangolin classification) but we have presented results mainly for the more frequently observed ones, namely AY.20, AY.39 and AY.44. The preferential occurrence of certain sub-lineages over others in different states within India and in different countries indicates a potential role of population-specific host genetic factors which might govern the favourable spread of one sub-lineage over another. However, we have not evaluated the association between host genetics and viral sub-lineages in this study. Although there are multiple reports of the association of a few S protein NVs (especially T478K, P681R and L452R) with viral transmissibility and immune escape [1151], we have performed extensive analyses on the entire landscape of SARS-CoV-2 genomic NVs in this study. Although these NVs have been reported to occur in samples belonging to the Delta lineage (as documented in GISAID, and https://outbreak.info/situation-reports/delta), their association with vaccination breakthrough events was not reported, highighting the importance of our analysis. However, an absence of information on neutralization antibody levels in vaccinated individuals did make it difficult for us to establish a strong association with vaccination breakthrough cases. Nation-wide sero-surveys have suggested that sero-positivity or presence of neutralization IgG antibodies can exist even in unvaccinated individuals [52] (probably arising from natural infection). To the best of our knowledge, a comparison between sero-positive unvaccinated and vaccinated individuals for specific NV prevalence has not been performed. Also, keeping in mind the interpretation pitfalls that may be created due to the small size of vaccination breakthrough cases, it is important to validate the results presented here on a larger sample set, which we are currently pursuing. This work gains further significance in the context of recent reports that discuss the causal relationship between specific genomic variations and their potential to cause enhanced immune escape, even in individuals who have received vaccinations that are currently recognized by WHO [53–55].

NV cross-correlation analysis is a powerful tool to establish moderately or tightly linked genome-wide signatures. Our analyses revealed the complete footprint of genomic alterations affiliated with specific viral lineages. The cross-correlation analysis not only described the entire set of alterations associated with Delta and its sub-lineages, but also revealed several NVs that were completely mutually exclusive with those associated with previously circulating lineages. Another significant observation made possible through the cross-correlation analysis was the preferential co-occurrence of specific NVs in Delta sub-lineages but absent from Delta itself, which could not be deduced from linear analysis of NV timelines. The analyses also suggested greater enrichment of certain co-occurring NVs in vaccination breakthrough (compared to other) cases.

Previous studies have provided evidence on how iSNVs impart genomic plasticity [56] and direct viral genome evolution through inter-host transmission cycles [57, 58]. Due to lack of data on donor–recipient pairs and primary contacts (including family members) of infected individuals, we were unable to perform iSNV bottleneck estimation in this study. Moreover, despite having a substantial dataset size, the information output is hampered by lack of clinical information such as infection symptoms, hospitalization status, etc. Nevertheless, a few observations are worth highlighting. First, we did not detect a significant correlation between samples showing iSNVs and their vaccination status or their age or with a specific lineage (data not shown). Second, we identified specific iSNVs such as 25907T in ORF3a that exhibited increased sample and allele frequencies with time. Interestingly, the 172V alteration generated from 25907T has been shown to improve protein stability, owing to increased local hydrophobic interactions, in recent studies [59]. The stability of ORF3a plays a crucial role in its functionality as an apoptosis-inducing protein leading to cell death [60] and membrane rearrangement during SARS-CoV-2 infection [61].

Our iSNV analysis has potentially revealed an important S protein allele, namely 1101Y, where the 'Y' allele frequency showed an upward trend from April 2021 and was labelled as an NV as its frequency becamereached >50%. Further, this NV was also preferentially associated with partially vaccinated samples. Interestingly, H1101Y, V1104L and T1117I are located between the two heptapeptide repeat sequences HR1 and HR2 within the S2 subunit of the S protein (Fig. S8). Earlier reports have suggested that both V1104L and H1101Y could increase local stability and alter the surface character of the S protein, thereby aiding favourable evolution of the virus [62, 63]. We therefore recommend including H1101Y and V1104L under active surveillance.

This study provides fresh insights into how the virus genome landscape has evolved over the duration of 19 months. Multiple factors related to host genetics, including co-morbidities and innate immune response, have been shown to play an important role in determining the evolution of virus variants [18]. However, since the current study is focused on samples obtained from the state of Telangana, we expect less host genetic variation among the patients analysed when compared to a pan-India study. The novelty of the study stems from its all all-inclusive approach and the identification of

missense NVs in the context of cross-correlation and intra-host diversity analysis. We suggest mutation cross-correlation and iSNV analyses as two important tools for future studies targeting other geographical regions as well as a more recent timeline (including the emergence and rapid spread of Omicron, which nevertheless resulted in a much 'milder' [64] and subdued third wave across India; https://ourworldindata.org/covid-cases). More importantly, we laid greater emphasis on individual NVs rather than the lineages per se, given the recent and frequent 're-classifications' of SARS-CoV-2 lineages by Pangolin [36]. Our study has facilitated a better understanding of how different aspects of virus genome dynamics are inter-linked. Future functional studies on important NVs identified in this study may reveal their possible role(s) in virus transmission and vaccine escape.

Funding Information
The study was supported by the 'INSACOG' (RAD-22017/28/2020-KGD-DBT) and National Genomics Core (BT/INF/22/SP28169/2019,07/03/2019) grants from the Department of Biotechnology, Ministry of Science and Technology, Government of India. A.G. acknowledges support from the Science and Engineering Research Board, Department of Science and Technology (DST-SERB) in the form of a National-Postdoctoral Fellowship (NPDF, PDF/2019/002427).

Acknowledgements
We are grateful to Dr Nagamani, Telangana state nodal officer, for coordinating sample collection from across the state. We thank all Telangana state sentinel sites assigned for identifying samples for genome sequencing, namely Gandhi Medical College and Hospital, Secunderabad; Institute of Preventive Medicine, Hyderabad, Osmania Medical College, Hyderabad, Nizam's Institute of Medical Sciences (NIMS), Hyderabad, Fever Hospital, Hyderabad, Government Medical College (GMC), Siddipet, Government Medical College, Warangal, Government General Hospital and Medical College, Suryapet, Government Medical College (GMC), Mahbubnagar, and Government Medical College, Nizamabad. We also acknowledge Dr S. Raju (State Public Health Laboratory (SPHL, Chennai, TN) and Dr Gyaneshwar Chaubey, Banaras Hindu University (BHU, UP), for access to vaccination breakthrough samples. We thank the Telangana State Government, the Indian Council of Medical Research, Government of India, and the University of Hyderabad, Hyderabad, for procurement of consumables and equipment to perform screening of patient samples. We are grateful to Dr K. Thangaraj, Dr Ashwin B. Dalal, Dr Rashna Bhandari and Dr R. Harinarayanan, CDFD, Hyderabad, for co-ordinating the activities of the COVID-19 testing laboratory at CDFD. We are grateful to Mr Mandla Vasanth Kumar for his kind assistance in computational analysis. All volunteers and 'COVID warriors' from CDFD, Hyderabad, are gratefully acknowledge for their significant contribution in screening of samples. We also acknowledge the National Genomics Core (NGC) – CDFD for performing SARS-CoV-2 genome sequencing.

Author contributions
M.D.B. and A.G. conceptualized the study. A.G., M.D.B. and R.B. developed the methodology. A.G. carried out the formal analysis. Writing (Original draft preparation) was carried out by A.G. M.D.B. and A.G. worked on reviewing and edited the draft, R.B. provided suggestions on editing the draft.

Conflicts of interest
The author(s) declare that there are no conflicts of interest.

Ethical statement
The work presented here was carried out following approvals from the CDFD Institutional Bioethics and Biosafety committee.

References

1. Gupta A, Basu R, Dharan Bashyam M. Supplement to Assessing the evolution of SARS-CoV-2 lineages and the dynamic associations between nucleotide variations. *Figshare*. 2023.
2. Frampton D, Rampling T, Cross A, Bailey H, Heaney J, *et al*. Genomic characteristics and clinical effect of the emergent SARS-CoV-2 B.1.1.7 lineage in London, UK: a whole-genome sequencing and hospital-based cohort study. *Lancet Infect Dis* 2021;21:1246–1256.
3. Cherian S, Potdar V, Jadhav S, Yadav P, Gupta N. Convergent evolution of SARS-CoV-2 spike mutations, L452R, E484Q and P681R, in the second wave of COVID-19 in Maharashtra, India. *Molecular Biology* 2021.
4. Davies NG, Abbott S, Barnard RC, Jarvis CI, Kucharski AJ, *et al*. Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England. *Science* 2021;372:eabg3055.
5. Kumar V, Singh J, Hasnain SE, Sundar D. Possible link between higher transmissibility of B.1.617 and B.1.1.7 variants of SARS-CoV-2 and increased structural stability of its spike protein and hACE2 affinity. *Biophysics* 2021.
6. Volz E, Mishra S, Chand M, Barrett JC, Johnson R, *et al*. Assessing transmissibility of SARS-CoV-2 lineage B.1.1.7 in England. *Nature* 2021;593:266–269.
7. Torjesen I. Covid-19: Delta variant is now UK's most dominant strain and spreading through schools. *BMJ* 2021;373:1445.
8. Weisblum Y, Schmidt F, Zhang F, DaSilva J, Poston D, *et al*. Escape from neutralizing antibodies by SARS-CoV-2 spike protein variants. *Elife* 2020;9:e61312.
9. Xie X, Zou J, Fontes-Garfias CR, Xia H, Swanson KA, *et al*. Neutralization of N501Y mutant SARS-CoV-2 by BNT162b2 vaccine-elicited sera. *bioRxiv* 2021:2021.01.07.425740.
10. Zhou D, Dejnirattisai W, Supasa P, Liu C, Mentzer AJ, *et al*. Evidence of escape of SARS-CoV-2 variant B.1.351 from natural and vaccine-induced sera. *Cell* 2021;184:2348–2361.
11. Planas D, Veyer D, Baidaliuk A, Staropoli I, Guivel-Benhassine F, *et al*. Reduced sensitivity of SARS-CoV-2 variant Delta to antibody neutralization. *Nature* 2021;596:276–280.
12. Edara V-V, Pinsky BA, Suthar MS, Lai L, Davis-Gardner ME, *et al*. Infection and Vaccine-Induced Neutralizing-Antibody Responses to the SARS-CoV-2 B.1.617 Variants. *N Engl J Med* 2021;385:664–666.
13. Cele S, Jackson L, Khoury DS, Khan K, Moyo-Gwete T, *et al*. SARS-CoV-2 Omicron has extensive but incomplete escape of Pfizer BNT162b2 elicited neutralization and requires ACE2 for infection. *medRxiv* 2021:2021.12.08.21267417.
14. Kirola L. Genetic emergence of B.1.617.2 in COVID-19. *New Microbes New Infect* 2021;43:100929.
15. Wolter N, Jassat W, Walaza S, Welch R, Moultrie H, *et al*. Early assessment of the clinical severity of the SARS-CoV-2 omicron variant in South Africa: a data linkage study. *Lancet* 2022;399:437–446.
16. Garcia-Beltran WF, Lam EC, St Denis K, Nitido AD, Garcia ZH, *et al*. Multiple SARS-CoV-2 variants escape neutralization by vaccine-induced humoral immunity. *Cell* 2021;184:2372–2383.
17. Willett BJ, Grove J, MacLean OA, Wilkie C, De Lorenzo G, *et al*. SARS-CoV-2 Omicron is an immune escape variant with an altered cell entry pathway. *Nat Microbiol* 2022;7:1161–1179.

18. Carter-Timofte ME, Jørgensen SE, Freytag MR, Thomsen MM, Brinck Andersen N-S, *et al*. Deciphering the role of host genetics in susceptibility to severe COVID-19. *Front Immunol* 2020;11:1606.
19. Harvey WT, Carabelli AM, Jackson B, Gupta RK, Thomson EC, *et al*. SARS-CoV-2 variants, spike mutations and immune escape. *Nat Rev Microbiol* 2021;19:409–424.
20. Gupta A, Sabarinathan R, Bala P, Donipadi V, Vashisht D, *et al*. A comprehensive profile of genomic variations in the SARS-CoV-2 isolates from the state of Telangana, India. *J Gen Virol* 2021;102:001562.
21. Mlcochova P, Kemp S, Dhar MS. The Indian SARS-cov-2 genomics consortium (INSACOG). *Nature* 2021.
22. Pouwels KB, Pritchard E, Matthews PC, Stoesser N, Eyre DW, *et al*. Effect of Delta variant on viral burden and vaccine effectiveness against new SARS-CoV-2 infections in the UK. *Nat Med* 2021;27:2127–2135.
23. Ella R, Vadrevu KM, Jogdand H, Prasad S, Reddy S, *et al*. Safety and immunogenicity of an inactivated SARS-CoV-2 vaccine, BBV152: a double-blind, randomised, phase 1 trial. *Lancet Infect Dis* 2021;21:637–646.
24. Nordström P, Ballin M, Nordström A. Effectiveness of heterologous ChAdOx1 nCoV-19 and mRNA prime-boost vaccination against symptomatic Covid-19 infection in Sweden: A nationwide cohort study. *Lancet Reg Health Eur* 2021;11:100249.
25. Voysey M, Clemens SAC, Madhi SA, Weckx LY, Folegatti PM, *et al*. Safety and efficacy of the ChAdOx1 nCoV-19 vaccine (AZD1222) against SARS-CoV-2: an interim analysis of four randomised controlled trials in Brazil, South Africa, and the UK. *Lancet* 2021;397:99–111.
26. Elbe S, Buckland-Merrett G. Data, disease and diplomacy: GISAID's innovative contribution to global health. *Glob Chall* 2017;1: 33–46.
27. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 2014;30:2114–2120.
28. Li H, Durbin R. Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics* 2009;25:1754–1760.
29. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, *et al*. The sequence alignment/map format and SAMtools. *Bioinformatics* 2009;25:2078–2079.
30. Grubaugh ND, Gangavarapu K, Quick J, Matteson NL, De Jesus JG, *et al*. An amplicon-based sequencing framework for accurately measuring intrahost virus diversity using PrimalSeq and iVar. *Genome Biol* 2019;20:8.
31. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, *et al*. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly* 2012;6:80–92.
32. Turakhia Y, De Maio N, Thornlow B, Gozashti L, Lanfear R, *et al*. Stability of SARS-CoV-2 phylogenies. *PLoS Genet* 2020;16: e1009175.
33. Rambaut A, Holmes EC, O'Toole Á, Hill V, McCrone JT, *et al*. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat Microbiol* 2020;5:1403–1407.
34. Friendly M. Corrgrams: exploratory displays for correlation matrices. *Am Stat* 2002;56:316–324.
35. Wilm A, Aw PPK, Bertrand D, Yeo GHT, Ong SH, *et al*. LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. *Nucleic Acids Res* 2012;40:11189–11201.
36. O'Toole Á, Scher E, Underwood A, Jackson B, Hill V, *et al*. Assignment of epidemiological lineages in an emerging pandemic using the pangolin tool. *Virus Evol* 2021;7:veab064.
37. Singh J, Rahman SA, Ehtesham NZ, Hira S, Hasnain SE. SARS-CoV-2 variants of concern are emerging in India. *Nat Med* 2021;27:1131–1133.
38. Jha N, Hall D, Kanakan A, Mehta P, Maurya R, *et al*. Geographical landscape and transmission dynamics of SARS-CoV-2 variants across India: a longitudinal perspective. *Front Genet* 2021;12:753648.
39. Tonkin-Hill G, Martincorena I, Amato R, Lawson ARJ, Gerstung M, *et al*. Patterns of within-host genetic diversity in SARS-CoV-2. *Elife* 2021;10:e66857.
40. Jackson B, Boni MF, Bull MJ, Colleran A, Colquhoun RM, *et al*. Generation and transmission of interlineage recombinants in the SARS-CoV-2 pandemic. *Cell* 2021;184:5179–5188.
41. Valesano AL, Rumfelt KE, Dimcheff DE, Blair CN, Fitzsimmons WJ, *et al*. Temporal dynamics of SARS-CoV-2 mutation accumulation within and across infected hosts. *PLoS Pathog* 2021;17:e1009499.
42. Lythgoe KA, Hall M, Ferretti L, de Cesare M, MacIntyre-Cockett G, *et al*. Within-host genomics of SARS-CoV-2. *Genomics* 2020.
43. Martin MA, VanInsberghe D, Koelle K. Insights from SARS-CoV-2 sequences. *Science* 2021;371:466–467.
44. Ovsyannikova IG, Haralambieva IH, Crooke SN, Poland GA, Kennedy RB. The role of host genetics in the immune response to SARS-CoV-2 and COVID-19 susceptibility and severity. *Immunol Rev* 2020;296:205–219.
45. Di Maria E, Latini A, Borgiani P, Novelli G. Genetic variants of the human host influencing the coronavirus-associated phenotypes (SARS, MERS and COVID-19): rapid systematic review and field synopsis. *Hum Genomics* 2020;14:30.
46. Andreano E, Piccini G, Licastro D, Casalino L, Johnson NV, *et al*. SARS-CoV-2 escape from a highly neutralizing COVID-19 convalescent plasma. *Proc Natl Acad Sci U S A* 2021;118:e2103154118.
47. Dhar MS, Marwal R, Vs R, Ponnusamy K, Jolly B, *et al*. Genomic characterization and epidemiology of an emerging SARS-CoV-2 variant in Delhi, India. *Science* 2021;374:995–999.
48. Singanayagam A, Hakki S, Dunning J, Madon KJ, Crone MA, *et al*. Community transmission and viral load kinetics of the SARS-CoV-2 delta (B.1.617.2) variant in vaccinated and unvaccinated individuals in the UK: a prospective, longitudinal, cohort study. *Lancet Infect Dis* 2022;22:183–195.
49. Li B, Deng A, Li K, Hu Y, Li Z, *et al*. Viral infection and transmission in a large, well-traced outbreak caused by the SARS-CoV-2 Delta variant. Epidemiology 2021. DOI: 10.1101/2021.07.07.21260122.
50. Grenfell BT, Pybus OG, Gog JR, Wood JLN, Daly JM, *et al*. Unifying the epidemiological and evolutionary dynamics of pathogens. *Science* 2004;303:327–332.
51. McCallum M, Walls AC, Sprouse KR, Bowen JE, Rosen LE, *et al*. Molecular basis of immune evasion by the Delta and Kappa SARS-CoV-2 variants. *Science* 2021;374:1621–1626.
52. Murhekar MV, Bhatnagar T, Thangaraj JWV, Saravanakumar V, Santhosh Kumar M, *et al*. Seroprevalence of IgG antibodies against SARS-CoV-2 among the general population and healthcare workers in India, June-July 2021: A population-based cross-sectional study. *PLoS Med* 2021;18:e1003877.
53. Ju B, Zheng Q, Guo H, Fan Q, Li T, *et al*. Immune escape by SARS-CoV-2 Omicron variant and structural basis of its effective neutralization by a broad neutralizing human antibody VacW-209. *Cell Res* 2022;32:491–494.
54. Ju B, Fan Q, Wang M, Liao X, Guo H, *et al*. Antigenic sin of wild-type SARS-CoV-2 vaccine shapes poor cross-neutralization of BA.4/5/2.75 subvariants in BA.2 breakthrough infections. *Nat Commun* 2022;13:7120.
55. Kurhade C, Zou J, Xia H, Liu M, Chang HC, *et al*. Low neutralization of SARS-CoV-2 Omicron BA.2.75.2, BQ.1.1 and XBB.1 by parental mRNA vaccine or a BA.5 bivalent booster. *Nat Med* 2023;29:344–347.
56. Karamitros T, Papadopoulou G, Bousali M, Mexias A, Tsiodras S, *et al*. SARS-CoV-2 exhibits intra-host genomic plasticity and low-frequency polymorphic quasispecies. *J Clin Virol* 2020;131:104585.
57. Wang Y, Wang D, Zhang L, Sun W, Zhang Z, *et al*. Intra-host variation and evolutionary dynamics of SARS-CoV-2 populations in COVID-19 patients. *Genome Med* 2021;13:30.
58. Lythgoe KA, Hall M, Ferretti L, Cesare M, MacIntyre-Cockett G. On behalf of the oxford virus sequencing analysis group. *Science* 2021;372:eabg0821.

59. Bianchi M, Borsetti A, Ciccozzi M, Pascarella S. SARS-cov-2 orf3a: mutability and function. *Int J Biol Macromol* 2021;170:820–826.
60. Issa E, Merhi G, Panossian B, Salloum T, Tokajian S. SARS-CoV-2 and ORF3a: nonsynonymous mutations, functional domains, and viral pathogenesis. *mSystems* 2020;5:e00266-20.
61. Ren Y, Shu T, Wu D, Mu J, Wang C, *et al*. The ORF3a protein of SARS-CoV-2 induces apoptosis in cells. *Cell Mol Immunol* 2020;17:881–883.
62. Bascos NA, Mirano-Bascos D, Abesamis KI, Bagoyo CA, Mallapre OT, *et al*. Structural analysis of spike protein mutations in the SARS-CoV-2 theta (P.3) variant biophysics. *Philipp J Sci* 2021;150. DOI: 10.56899/150.05.31.
63. Chand GB, Banerjee A, Azad GK. Identification of twenty-five mutations in surface glycoprotein (Spike) of SARS-CoV-2 among Indian isolates and their impact on protein dynamics. *Gene Rep* 2020;21:100891.
64. Lewnard JA, Kahn MM, Lipsitch M R, Tartof SY. Clinical among patients infected with Omicron (B.1.1.529) SARS-cov-2 variant in southern California. *Epidemiology* 2022.

Peer review history

VERSION 2

Editor recommendation and comments

https://doi.org/10.1099/acmi.0.000513.v2.3

Andrew Bosworth; Public Health England, UNITED KINGDOM

Date report received: 20 February 2023
Recommendation: Accept

Comments: The work presented is clear and the arguments well formed.

SciScore report

https://doi.org/10.1099/acmi.0.000513.v2.1

iThenticate report

https://doi.org/10.1099/acmi.0.000513.v2.2

Author response to reviewers to Version 1

RESPONSE TO COMMENTS FROM RECEIVING EDITOR AND REVIEWER'S

At the outset, we thank the receiving editor and the reviewers for the critical comments on the manuscript. I am giving below point-by-point response to the comments/queries raised by the Editor and the reviewers.

Editor comments:

Remarks: The reviewers have provided detailed commentary on your manuscript and have highlighted minor concerns with the work presented, which I encourage you to address in full. In particular there is a need to elaborate further on some of the methodologies used in your study, and improve some of the displayed figures. This is a study that would be of interest to the field and community.

Response: We thank the editor for positive comments on the manuscript.

Comment#1: Please deposit the data underlying the work in the Society's data repository Figshare account here: https://microbiology.figshare.com/submit. Please also cite this data in the Data Summary of the main manuscript and list it as a unique reference in the References section. When you resubmit your article, the Editorial staff will post this data publicly on Figshare and add the DOI to the Data Summary section where you have cited it. This data will be viewable on the Figshare website with a link to the preprint and vice versa, allowing for greater discovery of your work, and the unique DOI of the data means it can be cited independently.

Response: The data has been deposited in the Microbiology Society's figshare account (https://figshare.com/articles/dataset/Supplementary_tables/21346110/3). The same has now been cited in the revised manuscript in the Data Summary section and a unique reference has also been given (**Reference no 64 in Data Summary**).

Comment#2: Provide more detail in the Methods section and ensure that software is consistently cited and its version and parameters included.

Response: As per the Editor's instructions (and additional suggestions from the reviewers), the methods section has been elaborated further in the revised manuscript. In addition, all data has been uploaded in Github and figshare accounts for reproducing the figures Specifically, the methods section under the sub-heading '**SARS-CoV-2 RNA extraction and sequencing**' lines 125-128, 131-133, 135-140 include additional details. We also confirm that all software are appropriately cited and version and parameters are included in the manuscript.

Reviewer #1

Comment #1: I have some concerns about some of the methodology presented in this paper. Firstly, line 119 specifies that extraction of total RNA occurred at BSL2, were there any considerations taken into place for handling a level 3 pathogen?

Response: The RT-PCR based COVID19 diagnostics laboratory was set up in CDFD as per the guidelines established by Indian Council of Medical Research (ICMR, Government of India), following the regulatory approvals from Department of Biotechnology (DBT, Government of India). This was actually a BSL2+ laboratory and the same is now mentioned in the revised manuscript (line 120). The virus handling area was completely secluded within this laboratory and only inactivated virus was brought out into the RNA isolation area which itself was also secluded within a cabin. The staff working in both these sections were also independent with no overlap.

Comment #2: Line 130: You mention "The synthesized cDNA", however, there isn't any report of how the cDNA was made - I assume this was part of the kits mentioned in the prior paragraph. Was cDNA generated with oligodT primers and random hexamers?

Response: We thank the reviewer for raising this point. The random hexamer pool was used for cDNA synthesis and the protocol was exactly as described in our previous study (Gupta et al., 2021, reference no 19). We have now added these details in the revised manuscript (lines 124-138).

Comment #3: For the *in-silico*methodology, there is no mention of how primer sequences from the ARTIC V3 scheme were removed from sequencing reads in the pipeline. Please specify how you did this. If this hasn't been done this can contribute to false variant discoveries. I have checked the code on Github and this part is hashed out in the script where you have commented mask primers using iVar. Please confirm how you accounted for this in your methodology.

Response: We thank the reviewer for these comments. We did remove/trim the ARTIC V3 primer sequences from the reads before calling the variants. This line was commented out in the GitHub code as we were also testing a few alternative pipelines for variant calling. We confirm that the results presented in the manuscript do include primer trimming. We have now 'uncommented' this line in the Github code to remove the discrepancy.

Comment #4: Line 158: you provide a link to virological.org - please can you add this to your reference list and specify which parts of the reference you have followed.

Response: We have now added an appropriate reference for this link in the methods section (sub-heading – "*In silico*workflow for processing genome sequencing data", reference no 31) of the revised manuscript.

To further clarify, we have specifically used the VCF file provided in the above reference, which lists all sites to be used with caution, when calling variants. This VCF file can also be found here –

https://github.com/W-L/ProblematicSites_SARS-CoV2.

Comment #5: Further comments regarding methodology:

What method was used to translate the nucleotide sequences to amino acid sequences for you to infer phenotype changes?

Response: We used SNPEff software (cited in the methods section, reference no. 30, of the revised manuscript) to annotate the variant call files (VCFs), which also performs the nucleotide to amino acid conversion based on reference genome and GFF (genome feature format) file; both of which were downloaded from NCBI GenBank database (https://www.ncbi.nlm.nih.gov/sars-cov-2/).

Comment #6:2. Presentation of results. Figure 1: Please check for colorblind accessibility.

The figure Panel 1C is quite difficult to interpret with the amount of different colours, make sure key points are in the figure legend/manuscript.

Response: We thank the reviewer for this insightful observation. However, in order to depict the lineage evolution accurately and given the large number of variants/lineages, we had limited discrete colour options. Additionally, for figures like these, it is not possible to use a sequential colour palette which conventionally is more colour blind friendly. However, we have elaborated on the key findings from this figure sufficiently in the main manuscript text (**Results section, line numbers – 234-245**). We would also like to bring to the reviewer's kind attention, the following publications, which follow a similarly discrete colour theme for depicting data similar to ours–

1.https://www.microbiologyresearch.org/content/journal/mgen/10.1099/mgen.0.000589

2. https://www.microbiologyresearch.org/content/journal/mgen/10.1099/mgen.0.000684

3. https://www.microbiologyresearch.org/content/journal/mgen/10.1099/mgen.0.000838

Comment #7: Figure 2: There is no mention of N, Orf71/7b or nsp5 in the figure legend. However, this is a good visualisation of the variants increasing in frequency over time.

Response: We are grateful to the reviewer for this suggestion and we have now appropriately modified the figure legend in the revised manuscript.

Comment #8: Figure 3: If possible, could you supplement the titles for complete and partial vaccination on panel a? I found myself having to flick between pages to understand this figure, and I think these labels will help the reader to digest the information. This might also be a good opportunity to remind us how many samples were used in the analysis with N=x in the figure legend or supplemented on the figure. How does the time frame differ between the completely or partially vaccinated groups? I.e. were samples collected across the same time period? Is there any variation in this? Does this explain any of the differences or are you confident it is vaccination status?

Response: Figure 3 and the corresponding legend has now been modified as per the reviewer's suggestion. The sample collection timeline for completely and partially vaccinated samples was unbiased and was spread across an overlapping time period (shown in the figure below) which can also be inferred from dates mentioned in Tables S1A, B, and C. Since, there is no variation between the two sets of cases i.e. partial and completely vaccinated, hence, the observed differences between mutational status were not confounded by the time period of sample collection.

Figure showing a parallel sample collection timeline of partial and completely vaccinated cases

Comment #9: Figure 4: Presents the iSNVs observed over time at selected positions. If appropriate, it would be useful to show the amino acid changes that are associated with the nucleotide changes you are reporting.

Response: Figure 4 has now been modified in the revised manuscript to incorporate the above suggestion.

Comment #10: The introduction is concise and sets the tone for the study whilst highlighting the importance of genomic data. One key focus of the study is to consider vaccine escape variants. Perhaps some more references and examples of this in the introduction would be useful.

The results go on to describe the rise of the delta variant fromMarch 2021, and how sub-variants of delta were being described from July 2021. They highlight the importance of having a closer look at the NVs in the next section. In the results section, I noticed the report of C14408T being described as P4720L when actually this should be P323L, which is the hitchhiker mutation for S:D614G that emerged early on in the pandemic.

Through an odds ratio analysis, the authors have found mutations that could be associated with breakthrough cases.

The order of the results and the associated figures do take you through a logical story for the data that is being discussed.

Response: We once again thank the reviewer for positive comments. As suggested, we have now added a few references in the introduction section of the revised manuscript in the context of vaccine escape variants (**Line 55,Reference nos. 15-16; lines 62-64, reference no 16,18**). We have also added a phrase **(Line 253)**further clarifying the nomenclature of P4270L mutation in nsp12, also called as P323L. Please note that both refer to the same mutation.

Comment #11: The introduction could have more literature to highlight other work that assesses the impact of variants on vaccine breakthrough - I am sure I have seen studies from the lab that talks about reduced neutralising effect of new variants etc. So even if there is no epidemiological data - it would be good to discuss work that has been conducted experimentally. Likewise in the discussion. The authors highlight that to their knowledge, this hasn't been done before. However, I believe there is certainly literature that could be brought in here to discuss the importance of their findings. The discussion highlights the pitfalls of the study and reinforces the need to assess this on larger sample sizes.

Response: In response to the reviewer's comments, we have made necessary changes in the revised manuscript and have added references which discuss the immune escaping viral variants (**Introduction - references 15-16, 18, lines 55, lines 62-64; Discussion – lines 508-512, references – 52-54**)

Comment #12: In the abstract, line one you use "its" - I think it will be more clear to refer to this as SARS-CoV-2 first.

In the introduction the authors say that the WHO declared the pandemic in January 2020, this is incorrect and will need updating to March 2020. (Line 41)

Response: The modifications suggested above has been implemented in the revised manuscript.

Reviewer #2

Comment #1: Materials and Methods

Some clarification on how many samples provided sequence information would be helpful, perhaps in the Abstract where sample numbers are mentioned. The authors sate that the whole genome sequence from 3543 samples were analysed. However, the Materials and Methods section states that samples with rt-PCR Ct values <30 were sequenced and Table S1A shoes 1955 samples with appropriate Ct values. Also, lines 128-129 in Materials and Methods states that Ct values of <30 for both the RdRp and E genes were used as a determinant, but table S1A only shows Ct values for the RdPd gene.

Response: Regarding the absent Ct value information for samples in Table S1A, we didn't receive the corresponding data from the source site or centre which carried out the RT-PCR analysis. However, each site did confirm that all samples have a Ct value of <30, though the exact values were unavailable.

The Ct values of only RdRp were considered for choosing samples for downstream processing. This point was mentioned in our previous study, which has now been cited in the methods section of the revised manuscript (line – 131, reference no. 19).

Comment #2: Lines 95-97: the first mention of the vaccines occurs here, and it seems more logical to include the manufacturing details here rather than later in lines 104-107.

Response: The corrections as suggested above have now been included in the revised manuscript (lines 97-101).

Comment #3: Lines 124=127: Rather than write multiple viral genes, just refer to the three genes the two assays actually target.

Response: The corrections as suggested above have now been included in the revised manuscript.

Comment #4: Line 130: A little more detail on how the cDNA was synthesised would be helpful. Checking the Artic website link provided did not clarify exactly how this was achieved.

Response: As mentioned above in response to comment #2 raised by the first reviewer, appropriate details are now included in the revised manuscript.

Comment #5: Lines 207-208: replace the existing text with "the Delta variant (B.1.617.2} constituted the major linage detected in the state," etc.

Line 244: replace "last one year" with "previous year".

Response: The corrections as suggested above have now been included in the revised manuscript (line nos. 214, 250).

Comment #6: Some clarification on nomenclature might be helpful, e.g., lines 266-267 where "higher frequencies in AY.44 compared to Delta", might be better written as "though it was present in higher frequencies in sub-lineage AY.44" and delete the unnecessary reference to Delta. This fits in with the authors later descriptions of sub-lineages elsewhere (e.g., lines 290-291).

Response: The corrections as suggested above have now been included in the revised manuscript (line 273).

Comment #7: Lines 304-304:…. Cases belonged to the Delta variant, followed by its sub-lineages AY.44 et.

Line329: A rather the an AN NV.

Line 361: "reflecting an additional"

Line 461: "The Delta variant

Line 463: fold rather than folds

Line 519: "Nevertheless, a few observations etc.

Response: The corrections as suggested above have now been included in the revised manuscript. However, we would like to point out that 'an NV' (and not a NV) is the correct and so we have not made this change (line nos. 311, 368, 468, 470, 530 respectively).

Comment #8: References

The references do not appear to be in the journal style, which seems to be the first five author names, followed by et al.

Response: We thank the reviewer for this suggestion, however as per the ACMI guidelines, the referencing style will be modified by the journal editorial office after acceptance of the manuscript. Therefore, we have not made any corrections in the reference style in the revised manuscript.

Based on the above, we hope that the Editor shall find the revised manuscript suitable for publication in the Access Microbiology.

VERSION 1

Editor recommendation and comments

https://doi.org/10.1099/acmi.0.000513.v1.5

Andrew Bosworth; Public Health England, UNITED KINGDOM

Date report received: 14 December 2022
Recommendation: Minor Amendment

Comments: This is a study that would be of interest to the field and community. The reviewers have highlighted minor concerns with the work presented. Please ensure that you address their comments. Please deposit the data underlying the work in the Society's data repository Figshare account here: https://microbiology.figshare.com/submit. Please also cite this data in the Data Summary of the main manuscript and list it as a unique reference in the References section. When you resubmit your article, the Editorial staff will post this data publicly on Figshare and add the DOI to the Data Summary section where you have cited it. This data will be viewable on the Figshare website with a link to the preprint and vice versa, allowing for greater discovery of your work, and the unique DOI of the data means it can be cited independently. Please provide more detail in the Methods section and ensure that software is consistently cited and its version and parameters included. The reviewers have provided detailed commentary on your manuscript, which I encourage you to address in full. In particular there is a need to elaborate further on some of the methodologies used in your study, and improve some of the displayed figures. I look forward to receiving your revised manuscript. Best wishes Dr Andrew Bosworth

Reviewer 2 recommendation and comments

https://doi.org/10.1099/acmi.0.000513.v1.4

Anonymous.

Date report received: 07 December 2022
Recommendation: Minor Amendment

Comments: This paper builds on work published in 2021 on a smaller set of samples and is a comprehensive, well-written and well-conceived, study. The appropriate techniques were used, and ethical approvals obtained. The results were discussed with reference to the relevant literature in a clear and concise manner. I have only a few, very minor, comments on clarifying some of the methodological issues. Materials and Methods Some clarification on how many samples provided sequence information would be helpful, perhaps in the Abstract where sample numbers are mentioned. The authors sate that the whole genome sequence from 3543 samples were analysed. However, the Materials and Methods section states that samples with rt-PCR Ct values

Please rate the manuscript for methodological rigour
Good

Please rate the quality of the presentation and structure of the manuscript
Very good

To what extent are the conclusions supported by the data?
Strongly support

Do you have any concerns of possible image manipulation, plagiarism or any other unethical practices?
No

Is there a potential financial or other conflict of interest between yourself and the author(s)?
No

If this manuscript involves human and/or animal work, have the subjects been treated in an ethical manner and the authors complied with the appropriate guidelines?
Yes

Reviewer 1 recommendation and comments

https://doi.org/10.1099/acmi.0.000513.v1.3

Rebekah Penrice-Randal; University of Liverpool, Infection and Microbiomes, 146 Brownlow Hill, Ic2 building, Liverpool, UNITED KINGDOM
https://orcid.org/0000-0002-0653-2097

Date report received: 23 November 2022
Recommendation: Major Revision

Comments: 1. Methodological rigour, reproducibility and availability of underlying data I have some concerns about some of the methodology presented in this paper. Firstly, line 119 specifies that extraction of total RNA occured at BSL2, were there any considerations taken into place for handling a level 3 pathogen? Line 130: You mention "The synthesized cDNA", however, there isn't any report of how the cDNA was made - I assume this was part of the kits mentioned in the prior paragraph. Was cDNA generated with oligodT primers and random hexamers? For the in silico methodology, there is no mention of how primer sequences from the ARTIC V3 scheme were removed from sequencing reads in the pipeline. Please specify how you did this. If this hasn't been done this can contribute to false variant discoveries. I have checked the code on github and this part is hashed out in the script where you have commented mask primers using iVar. Please confirm how you accounted for this in your methodology. Line 158: you provide a link to virological.org - please can you add this to your reference list and specify which parts of the reference you have followed. Further comments regarding methodology: What method was used to translate the nucleotide sequences to amino acid sequences for you to infer phenotype changes? 2. Presentation of results Figure 1: Please check for colourblind accessibility. Panel 1C is quite difficult to interpret with the amount of different colours, make sure key points are in the figure legend/manuscript. Figure 2: There is no mention of N, Orf71/7b or nsp5 in the figure legend. However, this is a good visualisation of the variants increasing in frequency over time. Figure 3: If possible, could you supplement the titles for complete and partial vaccination on panel a? I found myself having to flick between pages to understand this figure, and I think these labels will help the reader to digest the information. This might also be a good opportunity to remind us how many samples were used in the analysis with N=x in the figure legend or supplemented on the figure. How does the time frame differ between the completely or partially vaccinated groups? I.e. were samples collected accross the same time period? Is there any variation in this? Does this explain any of the differences or are you confident it is vaccination status? Figure 4: Presents the iSNVs observed over time at selected positions. If appropriate, it would be useful to show the amino acid changes that are associated with the nucleotide changes you are reporting. 3. How the style and organization of the paper communicates and represents key findings The introduction is concise and sets the tone for the study whilst highlighting the importance of genomic data. One key focus of the study is to consider vaccine escape variants. Perhaps some more references and examples of this in the introduction would be useful. The results go on to describe the rise of the delta variant from March 2021, and how sub-variants of delta were being described from July 2021. They highlight the importance of having a closer look at the NVs in the next section. In the results section, I noticed the report of C14408T being described as P4720L when actually this should be P323L, which is the hitchhiker mutation for S:D614G that emerged early on in the pandemic. Through an odds ratio analysis, the authors have found mutations that could be associated with breakthrough cases. The order of the results and the associated figures do take you through a logical story for the data that is being discussed. 4. Literature analysis or discussion The introduction could have more literature to highlight other work that assesses the impact of variants on vaccine breakthrough - I am sure I have seen studies from the lab that talks about reduced neutralising effect of new variants etc. So even if there is no epidemiological data - it would be good to discuss work that has been conducted experimentally. Likewise in the discussion. The authors highlight that to their knowledge, this hasn't been done before. However, I believe there is certainly literature that could be brought in here to discuss the importance of their findings. The discussion highlights the pitfalls of the study and reinforces the need to assess this on larger sample sizes. 5. Any other relevant comments In the abstract, line one you use "its" - I think it will be more clear to refer to this as SARS-CoV-2 first. In the introduction the authors say that the WHO declared the pandemic in January 2020, this is incorrect and will need updating to March 2020. (Line 41)

Please rate the manuscript for methodological rigour
Satisfactory

Please rate the quality of the presentation and structure of the manuscript
Satisfactory

To what extent are the conclusions supported by the data?
Partially support

Do you have any concerns of possible image manipulation, plagiarism or any other unethical practices?
No

Is there a potential financial or other conflict of interest between yourself and the author(s)?
No

If this manuscript involves human and/or animal work, have the subjects been treated in an ethical manner and the authors complied with the appropriate guidelines?
Yes

SciScore report

https://doi.org/10.1099/acmi.0.000513.v1.1

iThenticate report

https://doi.org/10.1099/acmi.0.000513.v1.2