TITLE

Spatio-temporal dynamics of intra-host variability in SARS-CoV-2 genomes

AUTHORS AND AFFILIATIONS

Ankit K. Pathak[1#], Gyan Prakash Mishra[2#], Bharathram Uppili[1], Safal Walia[2], Saman Fatihi[1], Tahseen Abbas[1,3], Sofia Banu[4], Arup Ghosh[2], Amol Kanampalliwar[2], Atimukta Jha[2], Sana Fatima[2], Shifu Aggarwal[2], Mahesh Shanker Dhar[5], Robin Marwal[5], Radhakrishnan V. S.[5], Kalaiarasan Ponnusamy[5], Sandhya Kabra[5], Partha Rakshit[5], Rahul C. Bhoyar[1], Abhinav Jain[1,3], Mohit Kumar Divakar[1,3], Mohamed Imran[1,3], Mohammed Faruq[1], Divya Tej Sowpati[4], Lipi Thukral[1], Sunil K. Raghav[2*], Mitali Mukerji[1,6*]

[1] CSIR - Institute of Genomics and Integrative Biology (CSIR-IGIB), Delhi, India.

[2] Institute of Life Sciences, Bhubaneswar, Odisha, India.

[3] Academy for Scientific and Innovative Research, Human Resource Development Centre Campus, Ghaziabad, Uttar Pradesh, India.

[4] CSIR - Centre for Cellular and Molecular Biology (CSIR-CCMB), Hyderabad, Telangana, India.

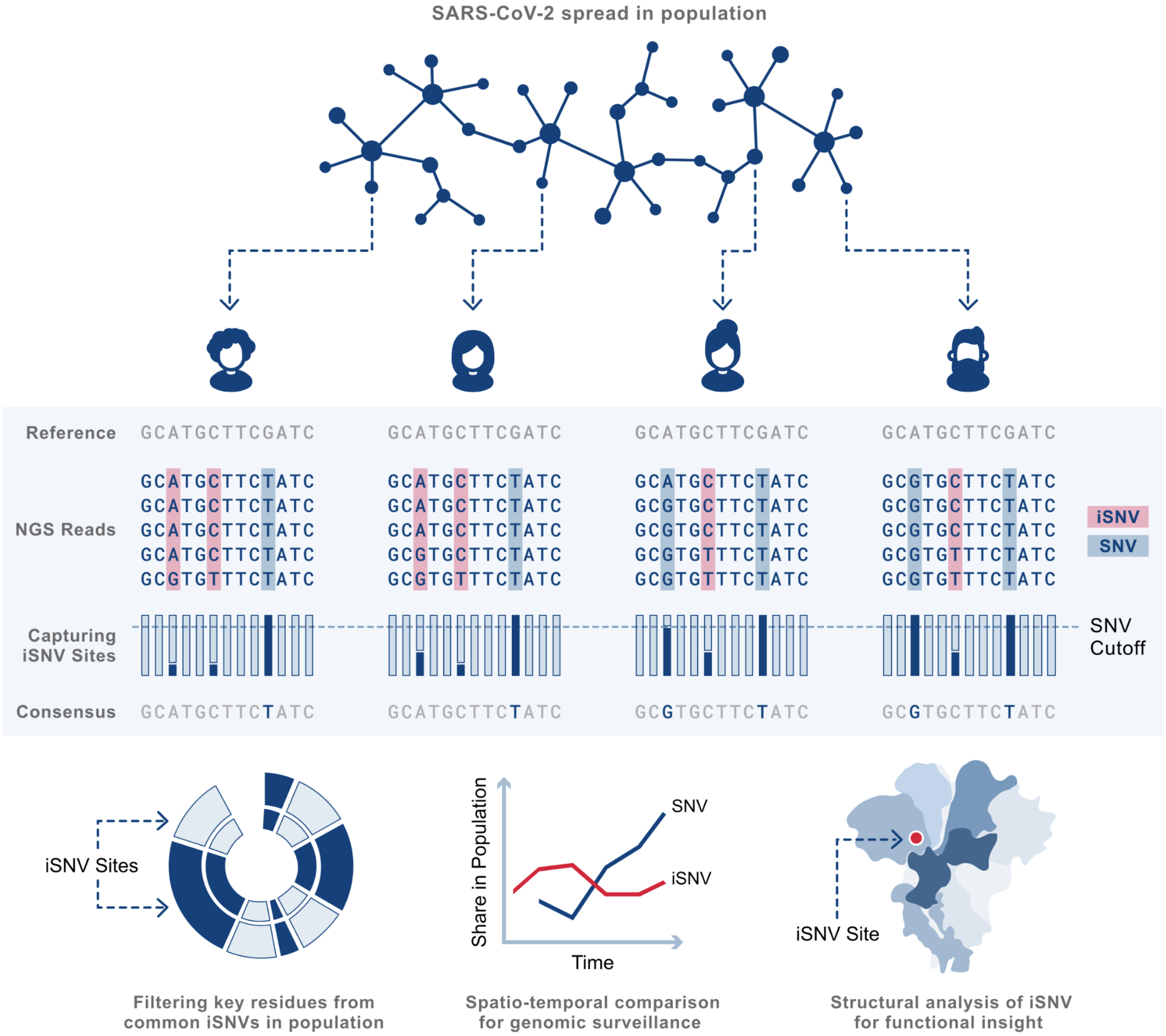
[5] Biotechnology Division, National Centre for Disease Control (NCDC), New Delhi, India.

[6] Indian Institute of Technology, Jodhpur, India.

# First author

* Correspondence: mitali@igib.res.in; sunilraghav@ils.res.in

GRAPHICAL ABSTRACT



ABSTRACT

During the course of the COVID-19 pandemic, large-scale genome sequencing of SARS-CoV-2 has been useful in tracking its spread and in identifying Variants Of Concern (VOC). Besides, viral and host factors could contribute to variability within a host that can be captured in next-generation sequencing reads as intra-host Single Nucleotide Variations (iSNVs). Analysing 1,347 samples collected till June 2020, we recorded 18,146 iSNV sites throughout the SARS-CoV-2 genome. Both, mutations in RdRp as well as APOBEC and ADAR mediated RNA editing seem to contribute to the differential prevalence of iSNVs in hosts. Noteworthy, 41% of all unique iSNVs were reported as SNVs by 30th September 2020 in samples submitted to GISAID, which increased to ~80% by 30th June 2021. Following this, analysis of another set of 1,798 samples sequenced in India between November 2020 and May 2021 revealed that majority of the Delta (B.1.617.2) and Kappa (B.1.617.1) variations appeared as iSNVs before getting fixed in the population. We also observe hyper-editing events at functionally critical residues in Spike protein that could alter the antigenicity and may contribute to immune escape. Thus, tracking and functional annotation of iSNVs in ongoing genome surveillance programs could be important for early identification of potential variants of concern and actionable interventions.

Keywords

SARS-CoV-2, SNVs, iSNVs, genomic surveillance, RNA-dependent RNA polymerase, RNA Editing, APOBEC, ADAR, Hyper-editing, Spike

INTRODUCTION

The SARS-CoV-2 pandemic has inflicted an enormous toll on human lives and livelihoods. As of June 30, 2021, 3.9 million deaths and more than 181 million confirmed cases of COVID-19 have been reported to the WHO from around the globe (1). As the virus accumulates novel mutations, probing millions of viral genome sequences has helped track the continual evolution of SARS-CoV-2. By rigorous cataloging of mutations, genomic surveillance has assisted in predicting the upcoming trajectories of infection by aiding epidemiological models with metrics of transmissibility and virulence of travelling viral strains in a population.

One of the differentiating features of RNA viruses is their exceptionally high nucleotide mutation rates that results in a set of closely related but non-identical genotypes, termed as *quasispecies*, in a host (2–7). Although random mutations have been shown to be largely deleterious (8, 9), some strains may develop enhanced survival ability by evolving mechanisms to resist antiviral responses of the host immune system or drug resistance (10, 11). The fitness of the intra-host viral population has been shown to be dependent on the cumulative contribution of all strains, including haplotypes harbouring minor variants (12). Various factors confer viral heterogeneity within a host like low fidelity of RNA-dependent RNA polymerase (RdRp) and RNA editing by host deaminases combined with the intrinsic high-rates of replication of RNA. Moreover, recent studies have shown mutations in the RdRp to affect the overall genome variability in SARS-CoV-2 (13, 14) whereas RNA editing by host deaminases has been proposed as a major driver of intra-host heterogeneity in the SARS-CoV-2 transcriptomes (15). Understanding these intra-host variations is important as they could potentially alter function through changes in RNA secondary structure, regulatory domains, as well as protein function and thereby govern host-pathogen interactions.

Genomic heterogeneity of the virus within a host can be analysed by capturing intra-host Single Nucleotide Variations (iSNVs), observed as heteroplasmic sites (often referred to in the mitochondrial genome context) in sequence reads. In this study, we accessed samples of COVID-19 diagnosed patients sequenced on an Illumina platform from two different time-periods of the pandemic. In Phase 1, we analysed 1,347 samples collected latest by June 2020 from China, Germany, Malaysia, United Kingdom, United States and different subpopulations of India to perceive a genome-wide iSNV map in SARS-CoV-2 infected populations. We observed 18,146 iSNV sites spanning the viral genome, including residues that defined the B.1 and B.6 lineages that were prominent in the listed populations before June 2020. Potential APOBEC and ADAR editing signatures were present throughout the SARS-CoV-2 genome with some samples showcasing ADAR hyperactivity. Further, mutations in RdRp and host-mediated RNA editing showed influence on the differential prevalence of iSNVs between hosts. Structural analysis of iSNVs in the Spike protein revealed a high density of alterations in functionally critical residues that could alter the antigenicity and may contribute to immune

escape. Interestingly, 41% of all unique iSNVs identified in these samples were found to be reported as an SNV by 30th September 2020 in one or more samples submitted in GISAID, increasing to ~80% by 30th June 2021. The likelihood that iSNVs can overtime manifest into SNVs in populations was further substantiated in Phase 2 by analyzing 1,798 samples sequenced in India between November 2020 and May 2021. In these samples, iSNVs could be recorded to be present in the population in most of the Delta (B.1.617.2) and Kappa (B.1.617.1) lineage-defining genomic positions prior to their fixation as SNVs by February 2021. These results highlight the importance of recording iSNVs as an extension to the genomic surveillance programs to enable more accurate models for viral epidemiology.

MATERIALS AND METHODS

Datasets

We processed Illumina next-generation sequencing samples from two different time-periods to understand the magnitude of prevalence of SARS-CoV-2 intra-host Single Nucleotide Variations (iSNVs) in infected individuals across various populations and to explore the possibility of iSNVs getting fixed in a population as a variant. For Phase 1, we accessed 1,347 publicly available samples submitted in the National Center for Biotechnology Information Sequence Read Archive (NCBI SRA) (16) till 23rd May 2020 from China, USA, Germany, Malaysia and the United Kingdom. Samples from laboratories in Bhubaneswar, Delhi and Hyderabad, that had been sequenced latest by 11th June 2020, were also included to represent SARS-CoV-2 infected East, North and South Indian populations respectively (Table 1). For Phase 2, we analyzed 1,798 samples sequenced in India between November 2020 and May 2021 for spatio-temporal review of the iSNV states of Delta (B.1.617.2) and Kappa (B.1.617.1) lineage-defining variations.

The GRCh37 human reference genome and NC_045512.2 SARS-CoV-2 Wuhan-Hu-1 reference genome were used for all analyses. High-coverage FASTA sequences submitted in GISAID (17) were accessed to get a position-wise share of SNVs in the global SARS-CoV-2 population using the 2019nCoVR database (18).

Data processing

The sequence files that were retrieved from NCBI SRA were downloaded in the SRA compressed file format and converted to FASTQ format files using NCBI's SRA-Tools (19). Adapter sequences were trimmed and low-quality reads were filtered using Trimmomatic (v=0.39) for downstream analysis (20). The trimmed reads were then aligned against the human reference genome (GRCh37/hg19) using HISAT2 (21). All paired-end reads unmapped to the human reference genome were filtered out using SAMtools (22) and mapped to the SARS-CoV-2 reference genome (NC_045512.2) using Burrows-Wheeler Aligner (BWA-MEM) (23). Since iSNVs cater to small fractions of reads, to avoid any bias induced by PCR amplification (24), all duplicate reads were removed from the SARS-CoV-2 aligned files using Picard Tools (25). BCFtools (22) was then used to generate the consensus SARS-CoV-2 FASTA sequence. Sample quality scores and alignment statistics were obtained using FastQC (26), QualiMap (27) and SAMtools (22) at each step.

We used REDItools2 (28) to call iSNVs in successfully aligned SARS-CoV-2 reads with a quality score ≥30. Additional position-level parameters in REDItools2 were used to discard some positions on the basis of site quality. The sites whose reads did not achieve a mean quality score of ≥33 were rejected and sites that were located in long homopolymeric regions of length ≥4bp (Supplementary Dataset S1) were excluded due to known sequencing errors in these regions (29). From the REDItools2 output, high-quality iSNV events were further filtered out using the following thresholds: number of minor allele reads ≥5, base coverage ≥20 and minor allele frequency ≥0.005. In order to test the pipeline's efficiency in identifying iSNVs in SARS-CoV-2 genome, we additionally downloaded and processed 500 single-end NGS samples sequenced in replicates on an Illumina platform from NCBI SRA with the Project ID PRJNA655577. We observed a high concordance between the replicates (Pearson's r=0.989, Supplementary Figure S1) in sites qualifying the pipeline thresholds. Thus, we used these cutoffs as a qualifying criterion for defining a heteroplasmic site as an iSNV in our subsequent analyses.

SARS-CoV-2 virus culture in Vero E6 cells

Virus isolation: Virus isolation from patient's swab samples was done following Harcourt. et al, 2020 (30). Vero E6 cells were grown in a 6-well plate at 80% confluency in DMEM supplemented with antibiotics (Penicillin-Streptomycin) and 10% Fetal bovine serum. Next day, Vero E6 cells were infected with 11 SARS-CoV-2 isolates at MOI of 1 in serum free media for 1 hour at 37°C with gentle rocking every 10-15 mins. One hour post infection, the inoculum was aspirated and cells were washed with PBS and supplemented with fresh complete media. The culture supernatant was aspirated 48 hours post infection, the cells were washed with 1x PBS and collected in Trizol reagent for total RNA extraction. SARS-CoV-2 virus isolation and culture was conducted in a biosafety Level-3 facility according to the biosafety guidelines issued by the Department of Biotechnology, Government of India.

qRT-PCR for viral load estimation and mRNA expression of APOBEC3B and ADARB1: RNA isolation from culture cells was performed using TRIzol Reagent (Invitrogen, cat. no. 10296010). The isolated RNA was subjected to qRT-PCR for determining the viral load. We performed one-step multiplex real-time PCR using TaqPathTM 1-Step Multiplex Master Mix (Thermo Fisher Scientific, cat. no. A28526), targeting three different SARS-CoV-2 genes with primer and probe sets specific for Spike (S), Nucleocapsid (N), and Open Reading Frame 1 (ORF1). To check the expression of APOBEC3B and ADARB1, RNA from wild-type (mock) and Vero E6 cells infected with 2 patient virus isolates was converted to cDNA using reverse transcriptase from qiaseq SARS-CoV-2 primer pool (Qiagen kit, cat no. 333896). APOBEC3b (sense: 5'-TCCACAGATCAGAAATCCGA-3'; antisense: 5'-CCATAGAGGATGGGTTCGT-3') and ADARB1 (sense: 5'-AAACATGAGTTCCAGCAGC-3'; antisense: 5'-CCCATTGGAGAGCTGAGAG-3') primer sequence were used to estimate the mRNA expression. 18S rRNA Primers (sense: 5'-CTACCACATCCAAGGAAGCA-3' and antisense: 5'-TTTTTCGTCACTACCTCCCCG-3') were used as internal control. qRT-PCR was performed using the Applied Biosystems (QuantStudio 6) machine.

Viral RNA library preparation and sequencing: Amplicon libraries for viral genome sequencing were prepared using QIAseq FX DNA Library Kit and QIAseq SARS-CoV-2 Primer Panel (Qiagen, cat. no. 180475, cat. no. 333896) as instructed by the manufacturer's manual, and the library was subsequently sequenced using Illumina NextSeq 550 platform. The adapter sequence used for each sample was compatible with Illumina sequencing instruments with 96-sample configurations (Qiaseq unique dual Y-adapter kit). The average insert length was in the 250–500 bp range. To characterize iSNVs, the *Chlorocebus sabaeus* genome (chlSab2 assembly) was used inplace of the human genome in the pipeline.

Identification of samples with potential hyper-editing events

Z-score values based on the distribution of number of iSNVs per sample were calculated for each population using the Python package SciPy (31). Samples with a Z-score value > 3 were classified as hyper-edited in each cohort (Figure 3A).

In order to categorize the specific amino acid change and the proteins containing the iSNVs, they were annotated using SnpEff version 4.5 (32). Conservation analysis of the full-length sequences of proteins was done on the basis of the six other coronaviruses (HCoV-229E, NL63, OC43, HKU1, MERS-CoV, SARS-CoV). The multiple sequence alignment of seven protein sequences was performed by clustal-omega (33). Conservation calculation of Spike amino acids was done using the ConSurf tool (34). The high-risk hyper-edited sites resulting in a single amino acid change in Spike proteins were screened and prioritized on the basis of sequence conservation and frequency i.e. number of times it is present in the total 25 hyper-edited samples. ConSurf conservation scores above 6 were considered and a total of 192 variants were filtered for mapping onto the structure to understand the effect on protein function and structure. We utilized PyMol (35) to determine interface residues from the available crystal structures. PDB 6M0J (36) and 6M17 (37) for ACE2 RBD binding and PDB 6W41 (38) and 7BZ5 (39) for antibody binding residues were used from Research Collaboratory for Structural Bioinformatics Protein Data Bank (RCSB PDB) (40).

Data handling and visualizations

Custom python scripts were used to automate the process of downloading and retrieving samples from NCBI SRA as well as to process samples through the pipeline. Python libraries such as NumPy (41), Pandas (42), Matplotlib (43) and Seaborn (44) were used for data handling and visualizations. SciPy (31) was used to compute pairwise two-sided t-tests and to calculate distribution statistics and Pearson’s correlation coefficients.

RESULTS

Intra-host Single Nucleotide Variations (iSNVs) in SARS-CoV-2 genomes

We analysed 1,347 transcriptomic samples obtained from patients diagnosed with COVID-19 by June 2020 from different populations (Table 1). 1,126 samples could be aligned to the SARS-CoV-2 reference genome. To ensure specificity of iSNV detection, the filtering criteria and cutoffs used in the sequence reads were established by analysing an additional set of 500 samples sequenced in replicates (Supplementary Figure S1). 945 of the 1,126 samples (~84%) harboured one or more iSNVs with frequencies ranging between 0.005 and 0.80 (Supplementary Dataset 3a, 3b). We recorded a total of 79,068 iSNVs with a median of ~19 iSNVs per sample (Supplementary Figure S2A), revealing extensive heteroplasmy in samples. We did not observe a significant correlation between the mean sample coverage depth and the number of iSNVs per sample suggesting that these events are unlikely to be a consequence of sequencing artefacts (Pearson's r=0.277, Supplementary Figure S2C). Further, in a sample subset, we did not find any correlation between the number of iSNVs in a sample and the Ct values for gene N (Pearson's r=-0.173, *p*=0.065), ORF1 (Pearson's r=-0.153, *p*=0.105), and S (Pearson's r=-0.132, *p*=0.162) (Supplementary Figure S2G).

Nucleotide changes due to iSNVs presented a skewed pattern, with A-to-G transition accounting for about a quarter of all changes (Figure 1A). Moreover, A-to-G changes were observed in disproportionate numbers in few samples while C-to-T changes were more consistently observed with a median of 5 events per sample. iSNVs contributed to variability throughout the SARS-CoV-2 genome, both in the coding as well as in the non-coding regions. A total of 23,392 unique iSNVs were recorded at 18,146 genomic positions (4,909 polymorphic) (Supplementary Dataset S4) in one or more samples distributed almost evenly in all genes and protein-coding domains of the genome (Supplementary Figure S3A). Though the frequency spectrum of iSNVs was nearly uniform with respect to the nature of amino acid sequence change, a large fraction of the sites lead to non-synonymous variations (Supplementary Figure S2E). Gain of stop codons that could alter the amount of functional proteins from the viral genomes were also observed.

Putative role of host and viral factor in iSNV incidence

Various factors have been shown to confer intra-host variability in many families of RNA viruses. Low-fidelity of RNA-dependent RNA polymerase (RdRp) has been established as a major contributor to this heterogeneity. Recent reports showed that mutations in the RdRp of SARS-CoV-2 may impact the number of mutations in the viral genome (13, 14). We observed significant difference in the number of iSNVs between the wild-type and C14408T mutant RdRp samples in both India ($p < 10^{-3}$) and USA ($p < 0.05$) with higher prevalence in the mutant RdRp (Figure 1B).

A bias towards the C-to-T/G-to-A and A-to-G/T-to-C nucleotide changes suggests RNA editing activity in both the strands of SARS-CoV-2 genome by the APOBEC (apolipoprotein B mRNA editing catalytic polypeptide-like) and ADAR (Adenosine Deaminase RNA Specific) family of enzymes respectively (Figure 1A). APOBECs target single-stranded nucleic acids for deamination of cytosines into uracils (C-to-U) reflecting as a C-to-T transition in the reference genome. ADARs, on the other hand, deaminate adenines into inosines (A-to-I) on double-stranded RNA (dsRNA). Inosine preferentially base pairs with cytidine which leads to incorporation of guanine in subsequent replication and an A-to-G transition. Editing could also occur in the negative strand of SARS-CoV-2 genome that would reflect as G-to-A (APOBEC) and T-to-C (ADAR) changes in the reference genome. Both C-to-T/G-to-A and A-to-G/T-to-C combinations were present throughout the genome (Supplementary Figure S3B) and contributed significantly to synonymous and missense changes, whereas only the former contributed to stop gains (Supplementary Figure S2F). Within the C14408T RdRp mutation carrying samples in India and USA, we observed differential prevalence of iSNVs ($p < 10^{-10}$) (Figure 1B). We also performed qRT-PCR analysis of APOBEC3B and ADARB1 expression in Vero cells where we clearly observed high and constitutive ADARB1 expression in both mock as well as in SARS-CoV-2 infected Vero cells (n=2), although APOBEC3B expression could not be captured (Supplementary Dataset S5). Further, we infected Vero cells with independent isolates of SARS-CoV-2 from 11 samples in the East India cohort. Comparing iSNVs at the 0th-passage and 7th-passage, we found significant accumulation of iSNVs in just 7 passages (Supplementary Dataset S6). Interestingly, the relative pattern of iSNVs with respect to the nucleotide change remained similar (Supplementary Figure S6).

Contribution of iSNVs to the spectrum of SARS-CoV-2 diversity across populations

To understand the contribution of iSNVs in the diversity of SARS-CoV-2 across various populations, we plotted a radial heatmap of the top 1% frequently observed genomic positions exhibiting a wide range of iSNV frequencies in major cohorts (Figure 1C). Each concentric ring represents an iSNV frequency range of 0.2 and the different colour gradients depict various populations with percentage of samples at a given position in each bin. Thus, the innermost ring represents a cohort-wise share of samples which show variations in a small fraction of reads (≤0.2) at a given position whereas the outermost ring represents the ones where the variant seems to have achieved near fixation (>0.8) and would be reported as an SNV in the population. Most of these positions were observed to be shared between global populations and also within Indian subpopulations (Supplementary Figure S4). Interestingly, many samples carried iSNVs at positions which defined the B.1 lineage (C241T, C3037T, C14408T and A23403G) and the India specific B.6 lineage (C6310A, C6312A, G11083T, C13730T and C23929T) (Figure 2A) (45, 46). Noteworthy, all B.1 lineage-defining variants depicted a C-to-T or an A-to-G nucleotide change.

Spatio-temporal dynamics of iSNVs

Though most of the sites seem to convert to a fixed variant in one or all populations (Figure 1C, Supplementary Figure S4), positions like 1707, 15435, 24622, 26554 and 29029 that depict an increasing trend in the iSNV frequency in few cohorts suggest that some iSNVs could get fixed in the population at a later date. We found significant overlap of iSNV sites with SNVs that have been reported in sequences submitted in GISAID. 41% of the 23,392 unique iSNVs identified in the samples were recorded as SNVs in one or more samples submitted in GISAID by 30th September 2020 with the exact change in the nucleotide sequence, increasing to ~80% by 30th June 2021 (Supplementary Dataset S4). As of June 2021, WHO had listed 4 lineages as Variants of Concern (VOC) and 7 as Variants of Interest (VOI). ~74% of all positions defining these lineages were recorded as iSNV sites in one or more samples including 75.6% of all Spike mutations (Supplementary Dataset S7). Given the origin and start of the outbreak of Delta (B.1.617.2) and Kappa (B.1.617.1) lineages in February 2021 in India (Figure 2A), we analysed 1,798 samples sequenced between November 2020 to May 2021 from different parts of the country. We found that at majority of the Delta and Kappa lineage-defining positions, SNVs appeared coinciding with or after the iSNVs first surfaced in the population (Figure 2B). This further strengthens the postulate that iSNVs may introduce SNVs into the travelling viral strains in a population which may beget newer lineages with heightened properties of transmissibility and virulence as seen with the Delta lineage.

Potential functional consequence of hyper-editing on Spike protein

We observed a few samples in each cohort to be substrates for potential hyper-editing events as reflected by an excessive number of iSNVs. In order to identify these population outliers, we computed the Z-score values based on the distribution of the number of iSNVs per sample (Figure 3A). In these samples, we observed A-to-G substitutions accounting for about one-third of all changed genomic positions implying extensive ADAR mediated editing activity (Figure 3B). Of the 11,420 annotated variants, 1,481 correspond to protein-coding variants within the Spike protein.

To understand how iSNVs could impact function through amino acids substitutions in hyper-edited samples, we examined the mutational spectra of the Spike protein in more detail, given that the initial host-viral response is triggered by the attachment of the Spike protein of SARS-CoV-2 with the host ACE2 receptor, disrupting the host cell membrane and activating viral entry (47). Amongst the 1,483 iSNVs in the Spike protein, 1,068 were found to be non-synonymous, 322 synonymous and 91 stop-coding. Figure 3C depicts the frequency of these sites on different functional protein domains, including the critical receptor-binding domain (RBD). The most frequently altered variants were D614, Y91, I105, and D428. The D614G mutation (one of the B.1 lineage defining variants) is near the S1/S2 cleavage site and has already been shown to be more geographically spreading than the D614 type (48).

We looked at other Spike mutations in more detail with respect to their functional consequences. Most striking, some amino acid sites within Spike seem to evolve into more than one type of variant. In Figure 3D, we depict these hyper-variable sites, with residue A879, N1108, S1239, D1260 evolving into either of >4 amino acid substitutions. The conservation of each residue i.e., how variable that site is amongst closely related coronavirus species was also calculated. For instance, the HR1 region and residues near the transmembrane are observed to be more conserved while NTD and RBD domains have highly variable regions, in that, especially the RBM motif (residue 437-508) seems to be less conserved (Figure 3E). Interestingly, RBM was also found to be variable in our study where specific amino acids at 442, 465, 468 conferred variations in ~11 out of 25 hyper-edited samples. The highly conserved amino acids represent a similar function of the protein across species but the substitution of critical amino acids around functional sites like RBM motif might lead to evolved or newer biological activities.

In a recent study, accelerated evolution of the virus over time leading to a repeated resurgence in an immunocompromised infected patient has been reported (49). Genomic analysis revealed that the individual had not been infected multiple times. Instead, the virus had lingered and rapidly mutated in the body leading to non-synonymous changes predominantly in the Spike region. The Spike protein and RBD harboured 57% and 38% of the total changes despite occupying only 13% and 2% of the virus genome respectively. The changes also included Q493K and F486I that corroborated with an earlier study that had recognised additional mutations (N74K, F79I, T259K, K417E, K444Q, V445A, N450D, Y453F, L455F, E484K, G485D, F490L, F490S, H655Y, R682Q, R685S, V687G, G769E, Q779K and V1128A) to be implicated in antibody resistance (50). Besides, N439K mutation in Spike has also been shown to be involved in immune escape from a panel of neutralizing monoclonal antibodies in another study (51). 20 out of these 23 positions overlap with sites that have been shown to carry iSNVs in our samples (Supplementary Dataset S8).

DISCUSSION

The course of the current pandemic has seen huge efforts in sequencing SARS-CoV-2 genomes from COVID-19 diagnosed patients and has helped catalogue the spread of mutations since the initial strain reported from Wuhan back in December 2019. Genomic surveillance has become an essential aid to epidemiological models and has facilitated prior predictions of variation(s) of concern through observations on inter-host differences in the viral genome over a period of time. In this study, we explored intra-host variability in the SARS-CoV-2 genome at spatio-temporal scales as an extended tool for genomic surveillance and examined factors contributing to its prevalence and its downstream effects.

Intra-host Single Nucleotide Variations (iSNVs) can alter the fitness of the strains by influencing its virulence, infectivity or transmissibility properties which may be both beneficial or detrimental to the virus and lead to fixation or removal of alleles from the populations. Of the 23,392 unique iSNVs recorded in samples collected latest by June 2020, 41% and ~80% had been recorded as an SNV in samples submitted in GISAID (17) till 30th September 2020 and 30th June 2021 respectively (Supplementary Dataset S4). Given the global spread of the infection, even within the relatively small subset of samples analysed in this study from select population groups which were collected in the first few months of the pandemic, ~74% of all the variants listed by the WHO as VOCs or VOIs by June 2021 (52) had been recorded as an iSNV site in one or more samples (Supplementary Dataset S7). Recent studies have also suggested that the intra-host heterogeneity of SARS-CoV-2 genome in individuals may be transferable (53–55), and given that a minuscule percent of population is responsible for most of the local transmission of SARS-CoV-2 infection (56–58), there arises a possibility of iSNVs harbouring genome to be passed through a super-spreading event. This, in an isolated population, might also lead to an altered haplotype with some unique variations leading to fixation or removal of alleles from the population (59). By temporally capturing iSNVs and the consequential variant appearance in the population, we could determine iSNV sites in the East Indian population which possibly transitioned into an SNV (Supplementary Figure S5). Further analysing samples sequenced in India between November 2020 to May 2021, we found majority of Delta and Kappa lineage-defining positions to carry iSNVs before the incidence of the subsequent SNV in the population (Figure 2B). This indicates that conjoint analysis of iSNVs with such gain and loss of variants in spatio-temporal scales might provide some clues of the evolvability of sites and may prove to be an important asset in genomic surveillance. Moreover, co-occurrence of such events in distinct isolated populations may result in the independent origin of lineages with similar properties.

Although the extent of variability differs amongst individuals, there seem to be preferred iSNV sites that are shared across populations (Figure 1C, Supplementary Figure S4). By the end of 2020, the B.1 lineage had become the most prevalent in most parts of the world with near disappearance of other lineages which were once the predominant strain in certain regions. For example, the B.6 lineage which was reported to be primarily endemic to India (45, 46) in April 2020 had disappeared from the entire population by October 2020 (Figure 2A). Given that all B.1 lineage-defining positions (241, 3037, 14408 and 23403) are iSNV sites with C-to-T and A-to-G changes (Figure 1C), host-mediated RNA editing might have fueled the positive selection of the B.1 lineage. The APOBEC and ADAR family of enzymes are extensively reported to edit different families of viral genomes (60–64) including the coronaviruses (65). SARS-CoV-2 being zoonotic and of recent origin, seems to harbour a large number of sites that could be potentially targeted by the host editing machinery. The signatures of these enzymatic activities seem to be reflected in the evolving SARS-CoV-2 genomes (Figure 1A) which corroborates with an earlier report (15). The RT-qPCR expression data revealed high ADARB1 expression in SARS-CoV-2 infected Vero cells indicating its role in genome mutagenesis (Supplementary Dataset S5). Though we observed potential ADAR hyperactivity in some individuals, APOBEC-mediated C-to-T changes were more consistently observed in our samples (Figure 1A) which may explain the overrepresentation of C-to-T SNVs in the SARS-CoV-2 populations (66, 67). Besides, we also observed substantial abundance of G-to-T and C-to-A substitutions in some samples. These might be due to Reactive Oxygen Species (ROS), as hypothesized in a recent study (68). ROS activity oxidizes guanine to 7,8-dihydro-8-oxo-2'-deoxyguanosine (oxoguanine) that base pairs with adenine, leading to G-to-T transversions. This change on the negative strand would be reflected as a C-to-A transversion in the reference genome. ROS also has its role implicated in mutagenesis of many other virus families (69).

Low fidelity of RNA-dependent RNA polymerase (RdRp) is known to be a major driver of intra-host heterogeneity of RNA viruses. Correlating the number of iSNVs against mutations in RdRp revealed differential prevalence of iSNVs between the wild-type and C14408T mutant RdRp in samples from both India as well as USA, indicating positive selection of a more error-prone RdRp to maintain an active buffer of minor variants (Figure 1C). The difference in the prevalence of iSNVs could also arise out of varied RNA editing across populations (Figure 1C), highlighting the active role of the host. The editing enzymes have been shown to display genetic variability in populations, perhaps a consequence of selection based on pathogenic loads (70, 71). Transmission of viruses through different host populations could thus shape the mutational landscape and govern the evolution of the viral genomes.

Ongoing selection in the host may contribute to the fixation of some of the sites as variants in SARS-CoV-2 genomes with potential consequences on the functioning of some of the viral proteins. Structural analysis of iSNVs that occur in Spike shows that more than 40% of these variants result in a substantial change in the property of these amino acids (Supplementary Table S1). Changes in specific amino acids, especially mutations close to functional sites, may lead to drastic changes in protein structure as they can either introduce significant change in the sidechain or charge of the residue. Also, it has been shown that APOBEC influenced

C-to-T substitutions elevate the frequency of hydrophobic amino acid coding codons in SARS-CoV-2 peptides (72). We also mapped 192 iSNVs onto the protein structure by utilizing a full-length Spike protein model of closed conformation (Figure 3F). Some iSNVs are located on the surface of the receptor-binding domain (RBD) which is responsible for binding to the ACE2 receptor in the host cell. These would directly impact the binding of Spike protein to the receptor and might have a putative effect on the binding interface residues shown in surface representation (Figure 3G). Noteworthy, we observed iSNVs in ~87% of the sites in the Spike protein that have been recently reported to confer antibody resistance (Supplementary Dataset S8). These mutations can have major implications in vaccine response as they could alter the immunogenicity of the antigenic RBD peptide leading to differential antibody titers in infected individuals (73). Overall, our findings indicate a comprehensive survey for one of the SARS-CoV-2 proteins and we propose that some of these distinct patterns of sites in hyper-edited samples may directly interfere with specific functional output.

In conclusion, temporally tracking within-host variability of the virus in individuals and populations might provide important leads to the sites that are favourable or deleterious for virus survival. This information would be of enormous utility for predicting the spread and infectivity of viral strains in the population. Conjoint analysis with the intra-host variability of the SARS-CoV-2 genome should be the next step.

DATA AND CODE AVAILABILITY

The commands used in the pipeline, automated codes, detailed individual sample information and statistics are available at the following GitHub repository — https://github.com/pxthxk/iCoV19.

FUNDING

This work was funded and supported by the Council of Scientific and Industrial Research, India (MLP-2005) and ILS, Bhubaneswar intramural grant.

ACKNOWLEDGEMENTS

In memory of all who have lost their lives to COVID-19. We gratefully acknowledge the authors from the originating laboratories of the submitted samples in NCBI SRA and also the COVID-19 teams of CSIR-IGIB Delhi, ILS Bhubaneswar, CSIR-CCMB and NCDC for their efforts in aggregating, processing and sharing transcriptomic data of COVID-19 infected individuals. We also acknowledge the authors from the originating laboratories of the submitted SARS-CoV-2 genome data in GISAID. We would also like to thank Bharati Singh and Gulam Hussain Syed for providing cultured Vero E6 cells, Khushboo Singhal for her inputs and Pragyan Acharya and Abhay Sharma for their critical reviews of the manuscript.

AUTHOR CONTRIBUTIONS

Conceptualization: MM, SKR; Pipeline Design: AKP, TA, BU; Phase 1 Data Processing and Analysis: AKP, GPM; Visualization: AKP, SF[1], MM; Spike Hyper-editing Analysis: SF[1], LT, AKP; North Indian Cohort Sample Curation: BU, MF; East Indian Cohort Sample Curation and Processing: GPM, SW, AG, SKR; South Indian Cohort Sample Curation and Processing: SB, DTS; Phase 2 Sample Curation and Processing: GPM, SW, AK, AJ, SF[2], SA, BU, MSD, RM, RVS, KP, SK, PR; Cultured Vero Cell Processing and qRT-PCR: AJ, GPM; Replicate Sample Curation: RCB, AJ, MKD, MI; Writing (Original Draft): MM, AKP, LT, SF[1],GPM; Writing (Review and Editing): LT, SKR, DTS, MF, MM, AKP, GPM.

COMPETING INTERESTS

Authors declare no conflict of interest.

REFERENCES

1. WHO coronavirus disease (COVID-19) dashboard. Geneva: World Health Organization, 2020. (2021).
2. Stack,J.C., Murcia,P.R., Grenfell,B.T., Wood,J.L.N. and Holmes,E.C. (2013) Inferring the inter-host transmission of influenza A virus using patterns of intra-host genetic variation. *Proc. R. Soc. B Biol. Sci.*, **280**, 20122173.
3. Salemi,M. (2013) The intra-host evolutionary and population dynamics of human immunodeficiency virus type 1: a phylogenetic perspective. *Infect. Dis. Rep.*, **5**, 3.
4. Romano,C.M., Lauck,M., Salvador,F.S., Lima,C.R., Villas-Boas,L.S., Araújo,E.S.A., Levi,J.E., Pannuti,C.S., O'Connor,D. and Kallas,E.G. (2013) Inter- and Intra-Host Viral Diversity in a Large Seasonal DENV2 Outbreak. *PLoS ONE*, **8**, e70318.
5. Ni,M., Chen,C., Qian,J., Xiao,H.-X., Shi,W.-F., Luo,Y., Wang,H.-Y., Li,Z., Wu,J., Xu,P.-S., *et al.* (2016) Intra-host dynamics of Ebola virus during 2014. *Nat. Microbiol.*, **1**, 16151.
6. Muñoz-Medina,J.E., Garcia-Knight,M.A., Sanchez-Flores,A., Monroy-Muñoz,I.E., Grande,R., Esbjörnsson,J., Santacruz-Tinoco,C.E. and González-Bonilla,C.R. (2018) Evolutionary analysis of the Chikungunya virus epidemic in Mexico reveals intra-host mutational hotspots in the E1 protein. *PLOS ONE*, **13**, e0209292.
7. Cattaneo,R., Donohue,R.C., Generous,A.R., Navaratnarajah,C.K. and Pfaller,C.K. (2019) Stronger together: Multi-genome transmission of measles virus. *Virus Res.*, **265**, 74–79.
8. Sanjuán,R., Moya,A. and Elena,S.F. (2004) The distribution of fitness effects caused by single-nucleotide substitutions in an RNA virus. *Proc. Natl. Acad. Sci. U. S. A.*, **101**, 8396–8401.
9. Sanjuán,R. (2010) Mutational fitness effects in RNA and single-stranded DNA viruses: common patterns revealed by site-directed mutagenesis studies. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.*, **365**, 1975–1982.
10. Venkatesan,S., Rosenthal,R., Kanu,N., McGranahan,N., Bartek,J., Quezada,S.A., Hare,J., Harris,R.S. and Swanton,C. (2018) Perspective: APOBEC mutagenesis in drug resistance and immune escape in HIV and cancer evolution. *Ann. Oncol. Off. J. Eur. Soc. Med. Oncol.*, **29**, 563–572.
11. Herbert,A. (2019) ADAR and Immune Silencing in Cancer. *Trends Cancer*, **5**, 272–282.
12. Bordería,A.V., Isakov,O., Moratorio,G., Henningsson,R., Agüera-González,S., Organtini,L., Gnädig,N.F., Blanc,H., Alcover,A., Hafenstein,S., *et al.* (2015) Group Selection and Contribution of Minority Variants during Virus Adaptation Determines Virus Fitness and Phenotype. *PLOS Pathog.*, **11**, e1004838.
13. Pachetti,M., Marini,B., Benedetti,F., Giudici,F., Mauro,E., Storici,P., Masciovecchio,C., Angeletti,S., Ciccozzi,M., Gallo,R.C., *et al.* (2020) Emerging SARS-CoV-2 mutation hot spots include a novel RNA-dependent-RNA polymerase variant. *J. Transl. Med.*, **18**, 179.
14. Eskier,D., Karakülah,G., Suner,A. and Oktay,Y. (2020) RdRp mutations are associated with SARS-CoV-2 genome evolution. *PeerJ*, **8**, e9587.
15. Di Giorgio,S., Martignano,F., Torcia,M.G., Mattiuz,G. and Conticello,S.G. (2020) Evidence for host-dependent RNA editing in the transcriptome of SARS-CoV-2. *Sci. Adv.*, **6**, eabb5813.
16. Leinonen,R., Sugawara,H., Shumway,M., and International Nucleotide Sequence Database Collaboration (2011) The sequence read archive. *Nucleic Acids Res.*, **39**, D19-21.
17. Shu,Y. and McCauley,J. (2017) GISAID: Global initiative on sharing all influenza data - from vision to reality. *Euro Surveill. Bull. Eur. Sur Mal. Transm. Eur. Commun. Dis. Bull.*, **22**.
18. Zhao,W.-M., Song,S.-H., Chen,M.-L., Zou,D., Ma,L.-N., Ma,Y.-K., Li,R.-J., Hao,L.-L., Li,C.-P., Tian,D.-M., *et al.* (2020) The 2019 novel coronavirus resource. *Yi Chuan Hered.*, **42**, 212–221.
19. SRA Toolkit Development Team (2021) The NCBI SRA Toolkit.
20. Bolger,A.M., Lohse,M. and Usadel,B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinforma. Oxf. Engl.*, **30**, 2114–2120.
21. Kim,D., Paggi,J.M., Park,C., Bennett,C. and Salzberg,S.L. (2019) Graph-based genome alignment

and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.*, **37**, 907–915.
22. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G., Durbin,R., and 1000 Genome Project Data Processing Subgroup (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
23. Li,H. (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *ArXiv13033997 Q-Bio*.
24. Marx,V. (2017) How to deduplicate PCR. *Nat. Methods*, **14**, 473–476.
25. Broad Institute (2019) Picard Toolkit.
26. Andrews,S. (2010) FastQC: a quality control tool for high throughput sequence data.
27. Okonechnikov,K., Conesa,A. and García-Alcalde,F. (2016) Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinforma. Oxf. Engl.*, **32**, 292–294.
28. Picardi,E. and Pesole,G. (2013) REDItools: high-throughput RNA editing detection made easy. *Bioinformatics*, **29**, 1813–1814.
29. Ivády,G., Madar,L., Dzsudzsák,E., Koczok,K., Kappelmayer,J., Krulisova,V., Macek,M., Horváth,A. and Balogh,I. (2018) Analytical parameters and validation of homopolymer detection in a pyrosequencing-based next generation sequencing system. *BMC Genomics*, **19**, 158.
30. Harcourt,J., Tamin,A., Lu,X., Kamili,S., Sakthivel,S.Kumar., Murray,J., Queen,K., Tao,Y., Paden,C.R., Zhang,J., *et al.* (2020) Isolation and characterization of SARS-CoV-2 from the first US COVID-19 patient. biorXiv doi: https://dx.doi.org/10.1101/2020.03.02.972935, 07 March 2020, pre-print: not peer-reviewed.
31. Jones,E., Oliphant,T. and Peterson,P. (2001) SciPy: Open source scientific tools for Python. *SciPy Open Source Sci. Tools Python*.
32. Cingolani,P., Platts,A., Wang,L.L., Coon,M., Nguyen,T., Wang,L., Land,S.J., Lu,X. and Ruden,D.M. (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)*, **6**, 80–92.
33. Sievers,F., Wilm,A., Dineen,D., Gibson,T.J., Karplus,K., Li,W., Lopez,R., McWilliam,H., Remmert,M., Söding,J., *et al.* (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.*, **7**, 539.
34. Ashkenazy,H., Abadi,S., Martz,E., Chay,O., Mayrose,I., Pupko,T. and Ben-Tal,N. (2016) ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. *Nucleic Acids Res.*, **44**, W344-350.
35. Schrodinger LLC (2010) The PyMOL Molecular Graphics System, Version 1.8.
36. Lan,J., Ge,J., Yu,J., Shan,S., Zhou,H., Fan,S., Zhang,Q., Shi,X., Wang,Q., Zhang,L., *et al.* (2020) Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. *Nature*, **581**, 215–220.
37. Yan,R., Zhang,Y., Li,Y., Xia,L., Guo,Y. and Zhou,Q. (2020) Structural basis for the recognition of SARS-CoV-2 by full-length human ACE2. *Science*, **367**, 1444–1448.
38. Yuan,M., Wu,N.C., Zhu,X., Lee,C.-C.D., So,R.T.Y., Lv,H., Mok,C.K.P. and Wilson,I.A. (2020) A highly conserved cryptic epitope in the receptor binding domains of SARS-CoV-2 and SARS-CoV. *Science*, **368**, 630–633.
39. Wu,Y., Wang,F., Shen,C., Peng,W., Li,D., Zhao,C., Li,Z., Li,S., Bi,Y., Yang,Y., *et al.* (2020) A noncompeting pair of human neutralizing antibodies block COVID-19 virus binding to its receptor ACE2. *Science*, **368**, 1274–1278.
40. Berman,H.M. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
41. Harris,C.R., Millman,K.J., van der Walt,S.J., Gommers,R., Virtanen,P., Cournapeau,D., Wieser,E., Taylor,J., Berg,S., Smith,N.J., *et al.* (2020) Array programming with NumPy. *Nature*, **585**, 357–362.
42. Reback,J., McKinney,W., Bossche,J.V.D., Jbrockmendel, Augspurger,T., Cloud,P., Gfyoung, Sinhrks, Klein,A., Tratner,J., *et al.* (2019) pandas-dev/pandas: v0.25.2 Zenodo.
43. Hunter,J.D. (2007) Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.*, **9**, 90–95.
44. Waskom,M., Botvinnik,O., Ostblom,J., Lukauskas,S., Hobson,P., MaozGelbart, Gemperline,D.C.,

Augspurger,T., Halchenko,Y., Cole,J.B., *et al.* (2020) mwaskom/seaborn: v0.9.1 (January 2020) Zenodo.

45. Banu,S., Jolly,B., Mukherjee,P., Singh,P., Khan,S., Zaveri,L., Shambhavi,S., Gaur,N., Reddy,S., Kaveri,K., *et al.* (2020) A Distinct Phylogenetic Cluster of Indian Severe Acute Respiratory Syndrome Coronavirus 2 Isolates. *Open Forum Infect. Dis.*, **7**, ofaa434.
46. Kumar,P., Pandey,R., Sharma,P., Dhar,M.S., A.,V., Uppili,B., Vashisht,H., Wadhwa,S., Tyagi,N., Fatihi,S., *et al.* (2020) Integrated genomic view of SARS-CoV-2 in India. *Wellcome Open Res.*, **5**, 184.
47. Hoffmann,M., Kleine-Weber,H., Schroeder,S., Krüger,N., Herrler,T., Erichsen,S., Schiergens,T.S., Herrler,G., Wu,N.-H., Nitsche,A., *et al.* (2020) SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. *Cell*, **181**, 271-280.e8.
48. Korber,B., Fischer,W.M., Gnanakaran,S., Yoon,H., Theiler,J., Abfalterer,W., Hengartner,N., Giorgi,E.E., Bhattacharya,T., Foley,B., *et al.* (2020) Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus. *Cell*, **182**, 812-827.e19.
49. Choi,B., Choudhary,M.C., Regan,J., Sparks,J.A., Padera,R.F., Qiu,X., Solomon,I.H., Kuo,H.-H., Boucau,J., Bowman,K., *et al.* (2020) Persistence and Evolution of SARS-CoV-2 in an Immunocompromised Host. *N. Engl. J. Med.*, 10.1056/NEJMc2031364.
50. Baum,A., Fulton,B.O., Wloga,E., Copin,R., Pascal,K.E., Russo,V., Giordano,S., Lanza,K., Negron,N., Ni,M., *et al.* (2020) Antibody cocktail to SARS-CoV-2 spike protein prevents rapid mutational escape seen with individual antibodies. *Science*, 10.1126/science.abd0831.
51. Thomson,E.C., Rosen,L.E., Shepherd,J.G., Spreafico,R., da Silva Filipe,A., Wojcechowskyj,J.A., Davis,C., Piccoli,L., Pascall,D.J., Dillen,J., *et al.* (2021) Circulating SARS-CoV-2 spike N439K variants maintain fitness while evading antibody-mediated immunity. *Cell*, **184**, 1171-1187.e20.
52. World Health Organization (2021) Tracking SARS-CoV-2 variants.
53. Lythgoe,K.A., Hall,M., Ferretti,L., de Cesare,M., MacIntyre-Cockett,G., Trebes,A., Andersson,M., Otecko,N., Wise,E.L., Moore,N., *et al.* (2020) Within-host genomics of SARS-CoV-2. biorXiv doi: https://doi.org/10.1101/2020.05.28.118992, 28 May 2020, pre-print: not peer-reviewed.
54. Wang,D., Wang,Y., Sun,W., Zhang,L., Ji,J., Zhang,Z., Cheng,X., Li,Y., Xiao,F., Zhu,A., *et al.* (2021) Population Bottlenecks and Intra-host Evolution During Human-to-Human Transmission of SARS-CoV-2. *Front. Med.*, **8**, 585358.
55. Sapoval,N., Mahmoud,M., Jochum,M.D., Liu,Y., Leo Elworth,R.A., Wang,Q., Albin,D., Ogilvie,H., Lee,M.D., Villapol,S., *et al.* (2020) Hidden genomic diversity of SARS-CoV-2: implications for qRT-PCR diagnostics and transmission. biorXiv doi: https://doi.org/10.1101/2020.07.02.184481, 2 July 2020, pre-print: not peer-reviewed.
56. Endo,A., Centre for the Mathematical Modelling of Infectious Diseases COVID-19 Working Group, Abbott,S., Kucharski,A.J. and Funk,S. (2020) Estimating the overdispersion in COVID-19 transmission using outbreak sizes outside China. *Wellcome Open Res.*, **5**, 67.
57. Miller,D., Martin,M.A., Harel,N., Tirosh,O., Kustin,T., Meir,M., Sorek,N., Gefen-Halevi,S., Amit,S., Vorontsov,O., *et al.* (2020) Full genome viral sequences inform patterns of SARS-CoV-2 spread into and within Israel. *Nat. Commun.*, **11**, 5518.
58. Adam,D.C., Wu,P., Wong,J.Y., Lau,E.H.Y., Tsang,T.K., Cauchemez,S., Leung,G.M. and Cowling,B.J. (2020) Clustering and superspreading potential of SARS-CoV-2 infections in Hong Kong. *Nat. Med.*, **26**, 1714–1719.
59. Moreno,G.K., Braun,K.M., Riemersma,K.K., Martin,M.A., Halfmann,P.J., Crooks,C.M., Prall,T., Baker,D., Baczenas,J.J., Heffron,A.S., *et al.* (2020) Revealing fine-scale spatiotemporal differences in SARS-CoV-2 introduction and spread. *Nat. Commun.*, **11**, 5558.
60. Malim,M.H. (2009) APOBEC proteins and intrinsic resistance to HIV-1 infection. *Philos. Trans. R. Soc. B Biol. Sci.*, **364**, 675–687.
61. Suspene,R., Petit,V., Puyraimond-Zemmour,D., Aynaud,M.-M., Henry,M., Guetard,D., Rusniok,C., Wain-Hobson,S. and Vartanian,J.-P. (2011) Double-Stranded RNA Adenosine Deaminase ADAR-1-Induced Hypermutated Genomes among Inactivated Seasonal Influenza and Live

Attenuated Measles Virus Vaccines. *J. Virol.*, **85**, 2458–2462.

62. Rösler,C., Köck,J., Kann,M., Malim,M.H., Blum,H.E., Baumert,T.F. and von Weizsäcker,F. (2005) APOBEC-mediated interference with hepadnavirus production. *Hepatology*, **42**, 301–309.
63. Khrustalev,V.V., Khrustaleva,T.A., Sharma,N. and Giri,R. (2017) Mutational Pressure in Zika Virus: Local ADAR-Editing Areas Associated with Pauses in Translation and Replication. *Front. Cell. Infect. Microbiol.*, **7**.
64. Bishop,K.N., Holmes,R.K., Sheehy,A.M., Davidson,N.O., Cho,S.-J. and Malim,M.H. (2004) Cytidine Deamination of Retroviral DNA by Diverse APOBEC Proteins. *Curr. Biol.*, **14**, 1392–1396.
65. Milewska,A., Kindler,E., Vkovski,P., Zeglen,S., Ochman,M., Thiel,V., Rajfur,Z. and Pyrc,K. (2018) APOBEC3-mediated restriction of RNA virus replication. *Sci. Rep.*, **8**, 5960.
66. Simmonds,P. (2020) Rampant C→U Hypermutation in the Genomes of SARS-CoV-2 and Other Coronaviruses: Causes and Consequences for Their Short- and Long-Term Evolutionary Trajectories. *mSphere*, **5**, e00408-20, /msphere/5/3/mSphere408-20.atom.
67. Wang,R., Hozumi,Y., Zheng,Y.-H., Yin,C. and Wei,G.-W. (2020) Host Immune Response Driving SARS-CoV-2 Evolution. *Viruses*, **12**, 1095.
68. Graudenzi,A., Maspero,D., Angaroni,F., Piazza,R. and Ramazzotti,D. (2021) Mutational signatures and heterogeneous host response revealed via large-scale characterization of SARS-CoV-2 genomic diversity. *iScience*, **24**, 102116.
69. Camini,F.C., da Silva Caetano,C.C., Almeida,L.T. and de Brito Magalhães,C.L. (2017) Implications of oxidative stress on viral pathogenesis. *Arch. Virol.*, **162**, 907–917.
70. Kidd,J.M., Newman,T.L., Tuzun,E., Kaul,R. and Eichler,E.E. (2007) Population Stratification of a Common APOBEC Gene Deletion Polymorphism. *PLoS Genet.*, **3**, e63.
71. Park,E., Guo,J., Shen,S., Demirdjian,L., Wu,Y.N., Lin,L. and Xing,Y. (2017) Population and allelic variation of A-to-I RNA editing in human transcriptomes. *Genome Biol.*, **18**, 143.
72. Matyášek,R. and Kovařík,A. (2020) Mutation Patterns of Human SARS-CoV-2 and Bat RaTG13 Coronavirus Genomes Are Strongly Biased Towards C>U Transitions, Indicating Rapid Evolution in Their Hosts. *Genes*, **11**, 761.
73. Wang,L., Wang,L. and Zhuang,H. (2020) Profiling and characterization of SARS-CoV-2 mutants' infectivity and antigenicity. *Signal Transduct. Target. Ther.*, **5**, 185.

TABLES AND FIGURES

Cohort	No. of samples processed	No. of samples with SARS-CoV-2 reads	No. of samples with iSNVs
China	**136**	**128**	**119**
Germany	**18**	**17**	**13**
Malaysia	**4**	**4**	**4**
United Kingdom	**62**	**62**	**17**
USA	**638**	**443**	**384**
India	**489**	**472**	**408**
East India	246	245	197
North India	77	61	49
South India	166	166	162
TOTAL	**1347**	**1126**	**945**

Table 1. Cohort-wise share of samples for Phase 1. The number of processed samples for each population curated from samples submitted in NCBI SRA till 23rd May 2020 (China, Germany, Malaysia, United Kingdom, USA) and samples sequenced in laboratories from Bhubaneswar, Delhi, and Hyderabad (India) latest by 11th June 2020.

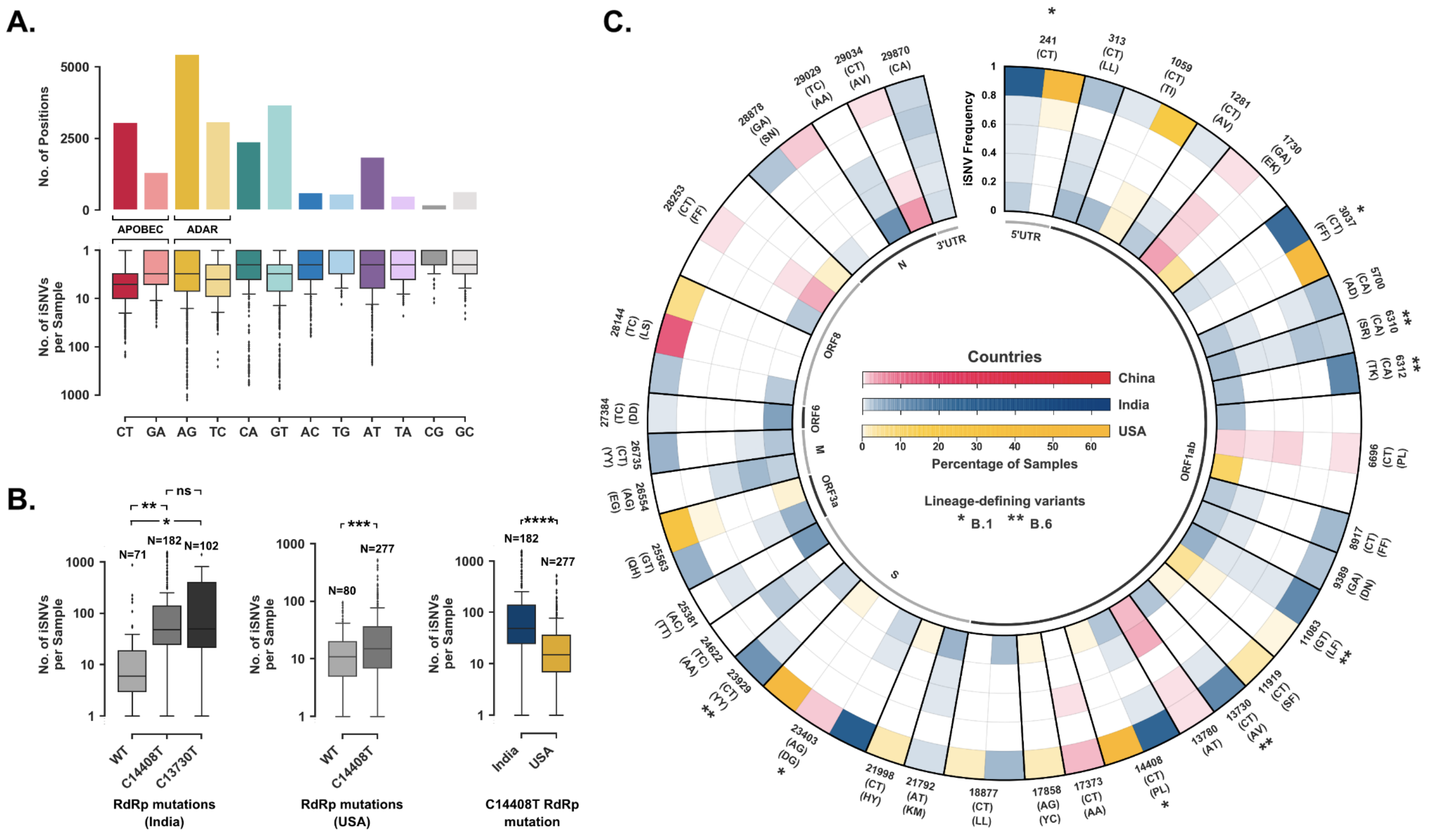


Figure 1. Spectrum of iSNVs in Phase 1 samples (**A**) Split plot depicting the distribution of unique iSNV sites (n=23392) with respect to the nucleotide change in the SARS-CoV-2 genome and across samples (**B**) Box plots showing comparison of iSNV prevalence between wild-type and RdRp mutant samples in India and USA (* $p < 10^{-6}$, ** $p < 10^{-3}$, *** $p < 0.05$, **** $p < 10^{-10}$). (**C**) Radial plot showcasing frequency distribution for selected iSNV sites in China, India and USA denoted in Red, Blue and Yellow respectively. Each concentric ring depicts an iSNV frequency range of 0.2 and the different colour gradients depict different populations with percentage of samples at a given position in each bin. The outer labels denote the position of change, nucleotide change and amino acid change. Variations that define the B.1 and B.6 lineages have been marked (*) and (**) respectively.

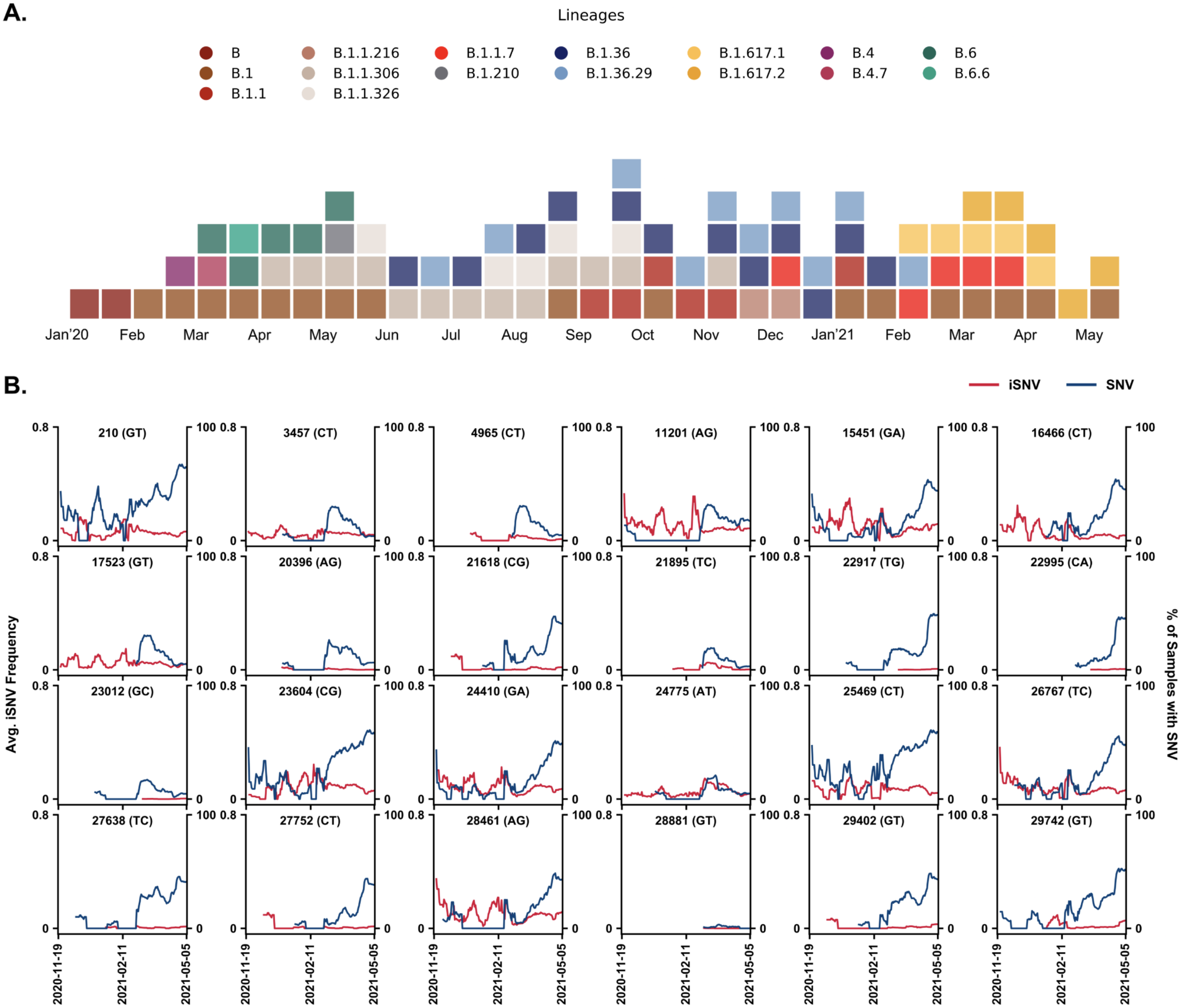


Figure 2. Spatio-temporal dynamics of iSNVs. (**A**) Temporal (fortnightly) distribution of prominent lineages identified in SARS-CoV-2 samples submitted pan-India till May 2021. Only lineages with at least 10% representation in the total samples submitted in a given fortnight period were selected. (**B**) Line plot showing temporal trends of iSNV frequency and incidence of SNV for Delta and Kappa lineage-defining positions in India between November 2020 to May 2021. The left y-axis denotes the average iSNV frequency and the right y-axis denotes the percentage of samples with SNVs on a 14-day rolling basis. Red and blue lines illustrate the average iSNV frequency and the percentage of samples with SNVs at the site respectively.

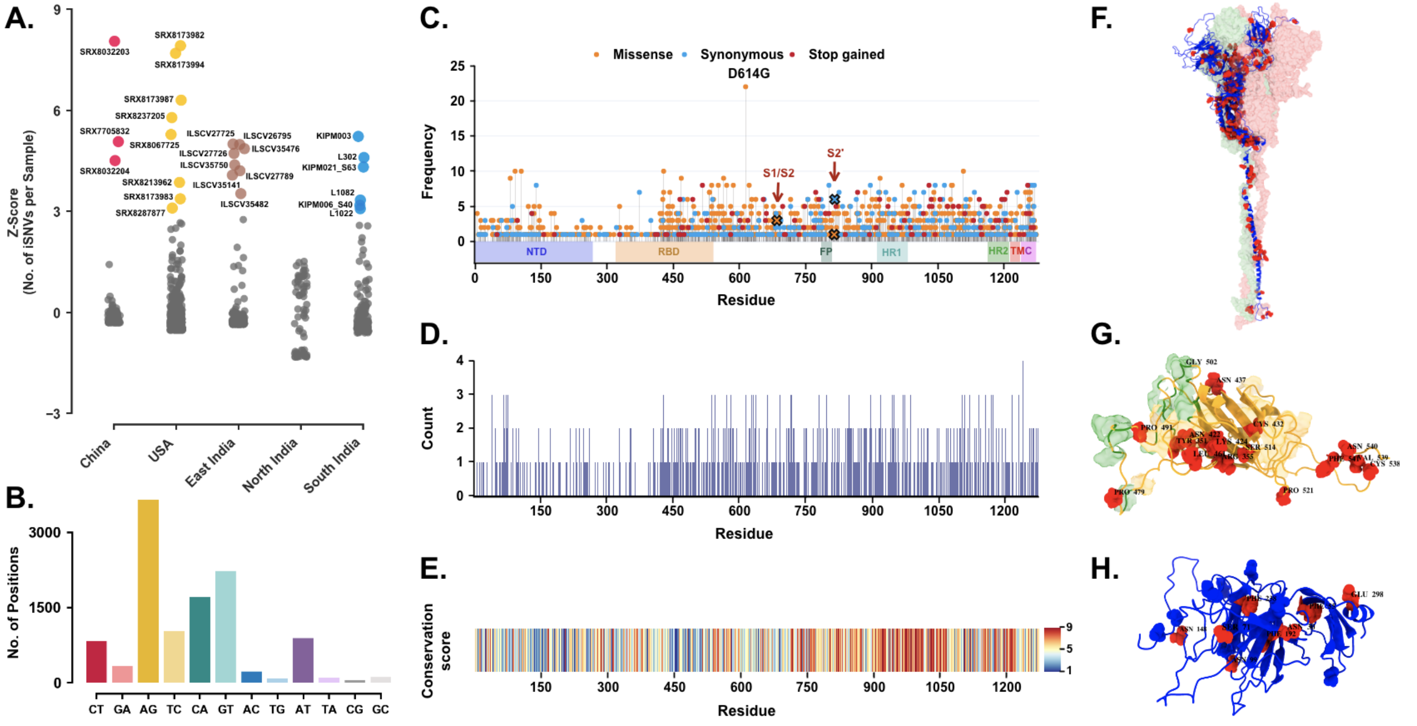
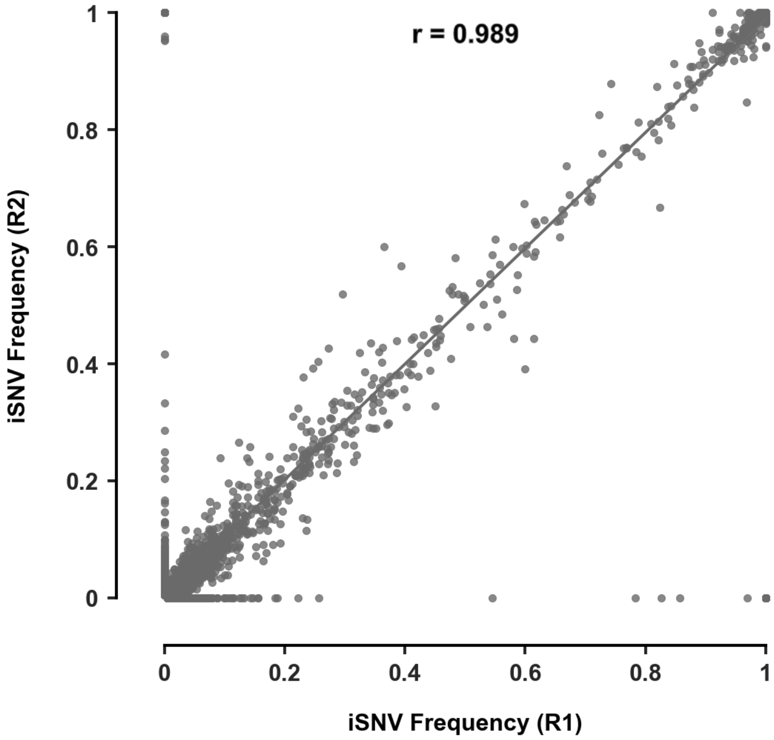


Figure 3. Functional impact of hyper-editing on Spike protein. (**A**) Dot plot illustrating cohort-wise outliers in Z-score values based on the distribution of number of iSNVs per sample (**B**) Bar plot representing distribution of iSNV sites (n=11392) with respect to nucleotide changes in the SARS-CoV-2 genome in hyper-edited samples (**C**) Needle plot depicting the distribution of protein-coding changes due to iSNVs in Spike protein. Non-synonymous, synonymous, and stop variants are shown in orange, blue, and red respectively. The length of the needle depicts the occurrence of altered residues out of the total of 25 hyper-edited samples. Protein domain architecture is indicated as horizontal boxes (**D**) Counts of amino acid substitutions at each residue location in the Spike protein (**E**) Conservation score of residues in the Spike protein. A conservation score of 9 denotes a highly conserved while a score of 1 denotes a highly variable position (**F**) Selected iSNVs with missense changes mapped onto the Spike trimer structure where red dots represent altered sites. The other two chains of the trimer are shown in surface representation (**G**) A close-up view of the RBD domain in surface representation harbouring the altered sites (iSNVs) in hyper-edited samples. Green highlights the ACE2 binding surface and orange highlights the antibody binding surface (**H**) A close-up view of the NTD domain with altered residues shown in red and N-glycosylation sites in blue. Altered residues are labelled in black text: amino acid name and position.

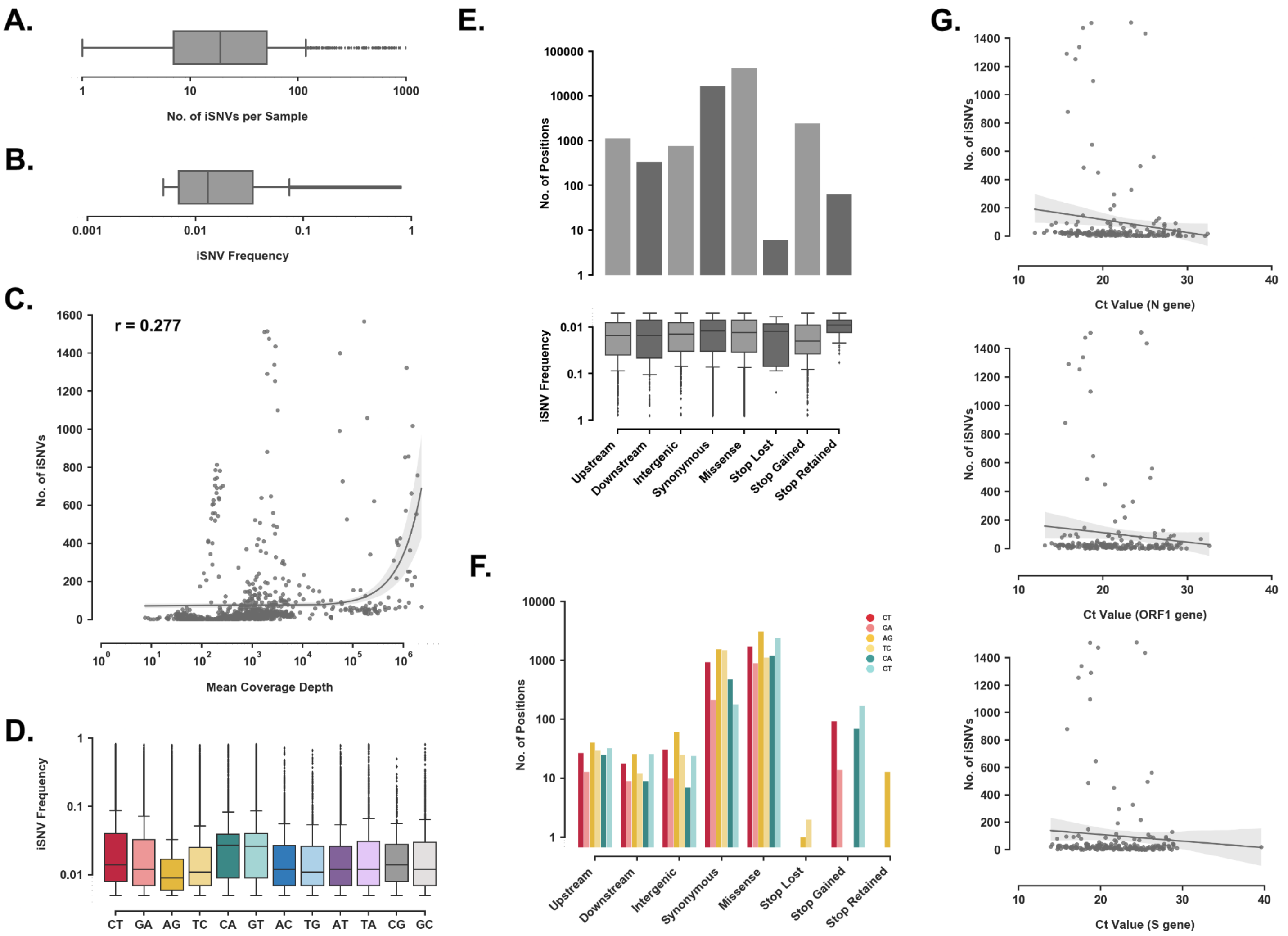
SUPPLEMENTARY DATA

	Hydrophobic	Polar	Positive	Negative
Hydrophobic	45	31	3	4
Polar	14	19	21	6
Positive	11	2	7	2
Negative	19	6	0	2

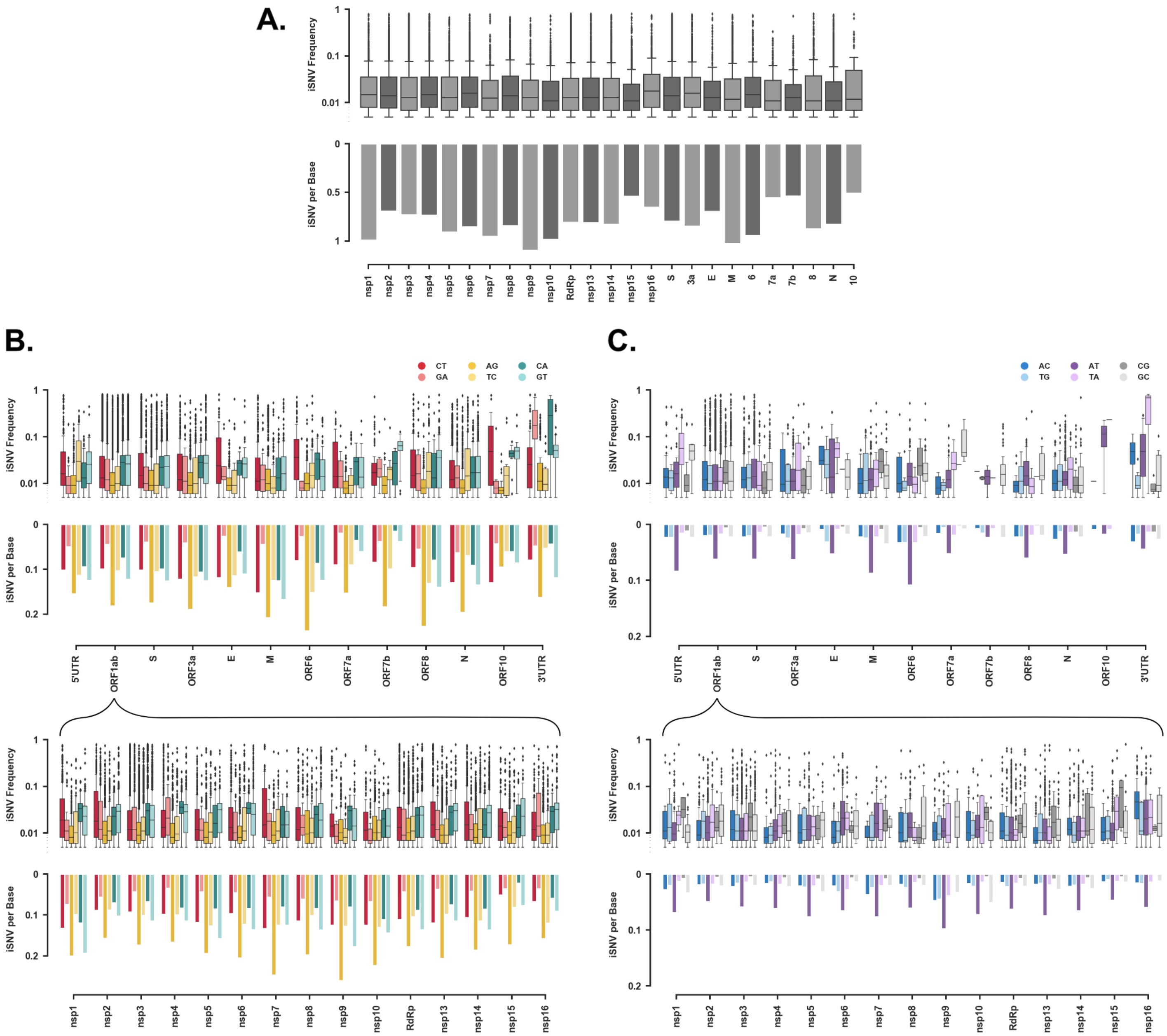
Supplementary Table S1: Type of amino acid change at Spike iSNV sites in hyper-edited samples. The iSNV associated Spike variants in the hyper-edited samples characterised according to hydrophobic, polar and charged (positive and negative) nature of the residue. The rows indicate the amino acid property of the original residue and the column indicates the property of the modified residue. The number of iSNV associated variants falling in each of the pair categories are listed.



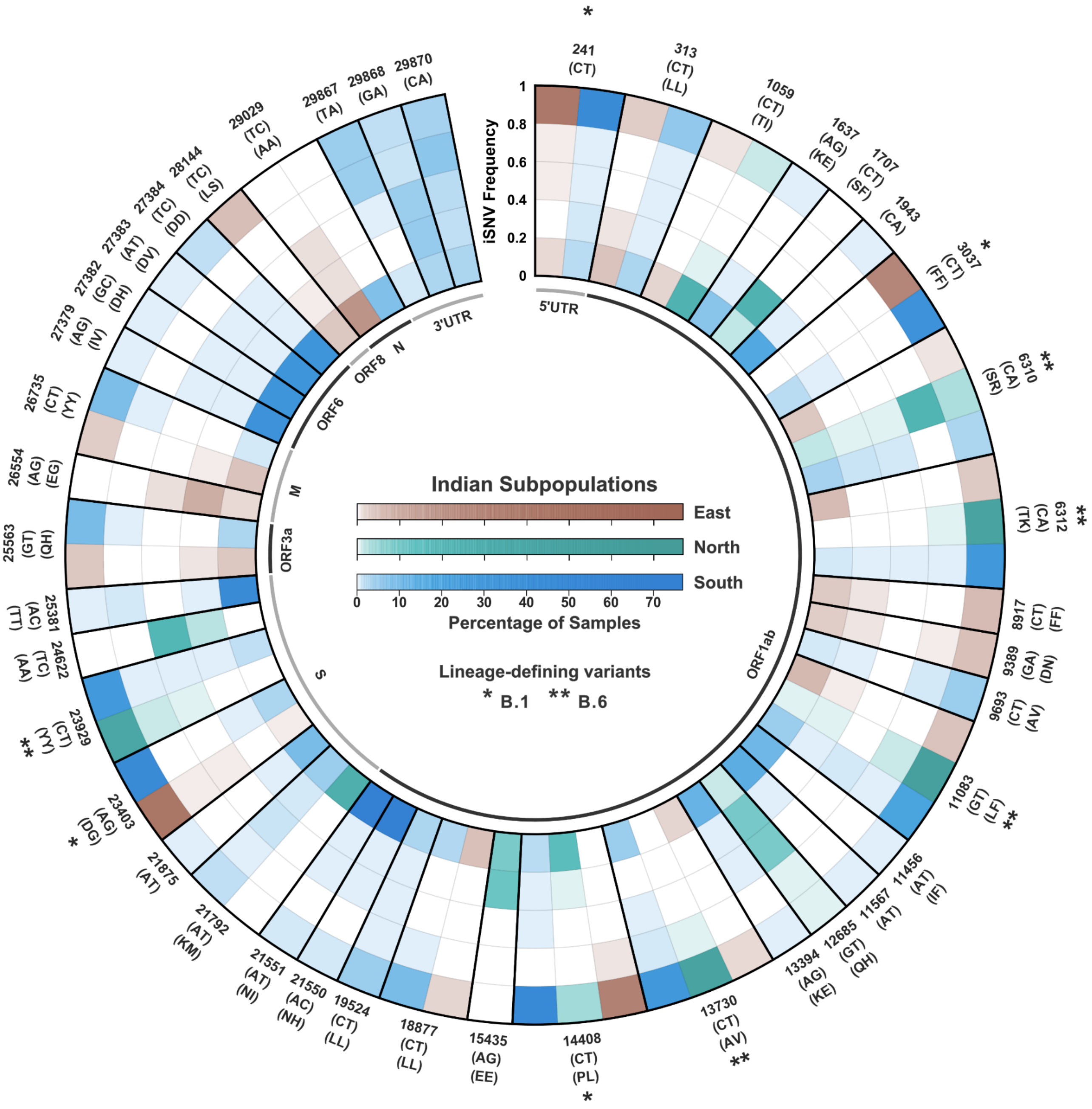
Supplementary Figure S1. iSNV concordance between replicates. Correlation plot illustrating the concordance in iSNV frequencies between replicates (n=500).



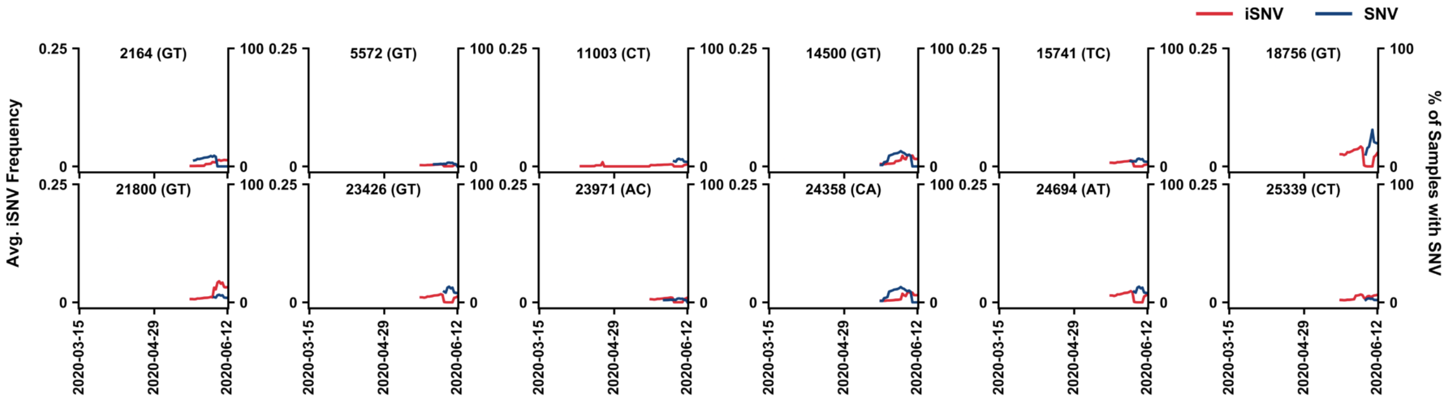
Supplementary Figure S2. Distribution of iSNVs in Phase 1 samples. (**A**) Distribution of the number of iSNVs per sample (n=945) (**B**) Distribution of the frequencies of all captured iSNVs (n=79068) (**C**) Scatter plot to assess the correlation between number of iSNVs and mean coverage depth in samples (n=945) (**D**) Distribution of iSNV frequencies with respect to the nucleotide change (**E**) Split plot showing distribution of number of iSNVs and its potential impact vis-a-vis nature of amino acid substitution (n=62971) (**F**) Distribution of nucleotide change mediated by iSNVs with respect to nature of amino acid sequence change (**G**) Correlation plot to compare the concordance between the number of iSNVs in samples and Ct values of N, ORF1 and S gene respectively.



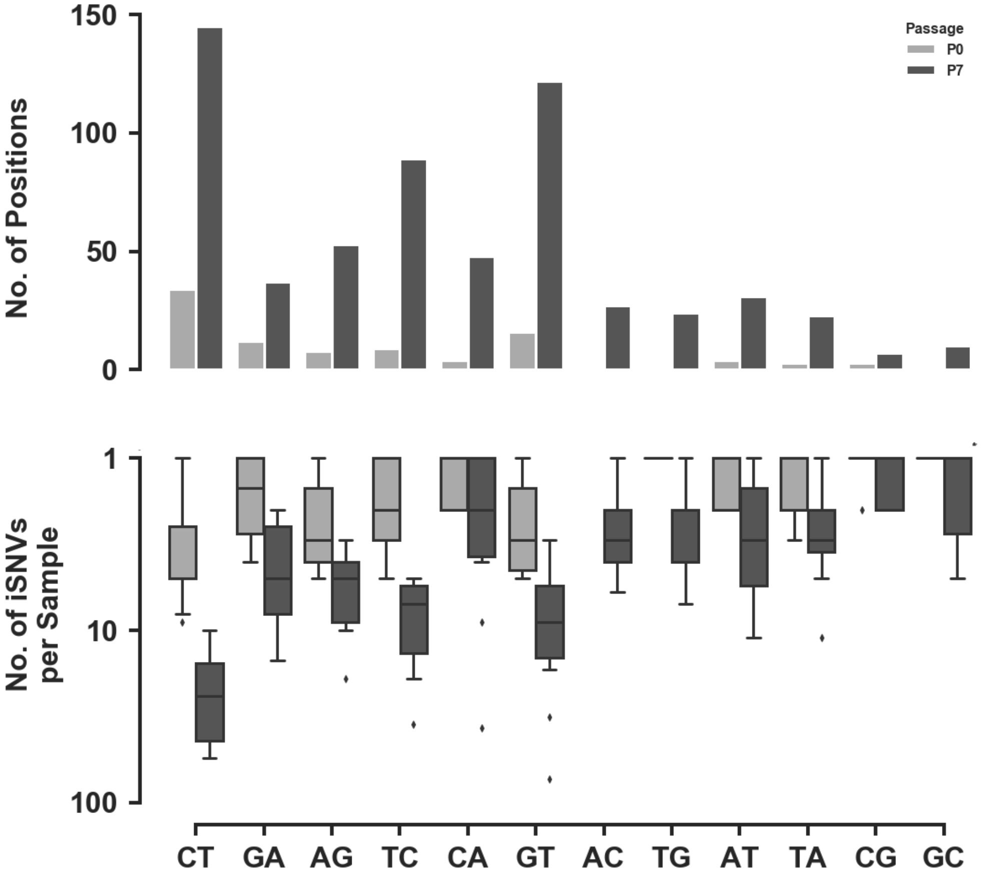
Supplementary Figure S3. Distribution of iSNVs across the SARS-CoV-2 genome. (**A**) Split plot depicting the distribution of iSNV frequencies and number of iSNVs per base in each protein-coding domain. (**B**) Split plot depicting the distribution of iSNV frequencies and number of iSNVs per base in each protein-coding domain with respect to potential APOBEC (CT/GA), ADAR (AG/TC) and ROS (CA/GT) inflicted nucleotide changes. (**C**) Split plot depicting the distribution of iSNV frequencies and number of iSNVs per base in each protein-coding domain with respect to other nucleotide changes.



Supplementary Figure S4. Spectrum of iSNVs in Indian subpopulations. Radial plot showcasing frequency distribution for select iSNV sites in East India, North India and South India denoted in Brown, Green and Blue respectively. Each concentric ring depicts an iSNV frequency range of 0.2 and the different colour gradients depict different populations with percentage of samples at a given position in each bin. The outer labels denote the position of change, nucleotide change and amino acid change. Variations that define the B.1 and B.6 lineages have been marked (*) and (**) respectively.



Supplementary Figure S5. Spatio-temporal dynamics of iSNVs in East India. Line plot showing temporal trends of iSNV frequency and incidence of SNV in the East India cohort. The left y-axis denotes the average iSNV frequency and the right y-axis denotes the percentage of samples with SNVs on a 14-day rolling basis. Red and blue lines illustrate the average iSNV frequency and the percentage of samples with SNVs at the site respectively.



Supplementary Figure S6. Accumulation of iSNVs in Vero cells. Comparison of iSNV prevalence at the 0th and the 7th passage of Vero cells infected with independent isolates of SARS-CoV-2 from 11 samples in the East India cohort.

Supplementary Dataset Legends

Supplementary Dataset S1: Coordinates of homopolymeric regions in the SARS-CoV-2 genome

Supplementary Dataset S2: Data summary of pipeline analyses and metadata for each sample

Supplementary Dataset S3a: Phase 1 sample-wise frequencies of all recorded iSNVs in global populations

Supplementary Dataset S3b: Phase 1 sample-wise frequencies of all recorded iSNVs in Indian subpopulations

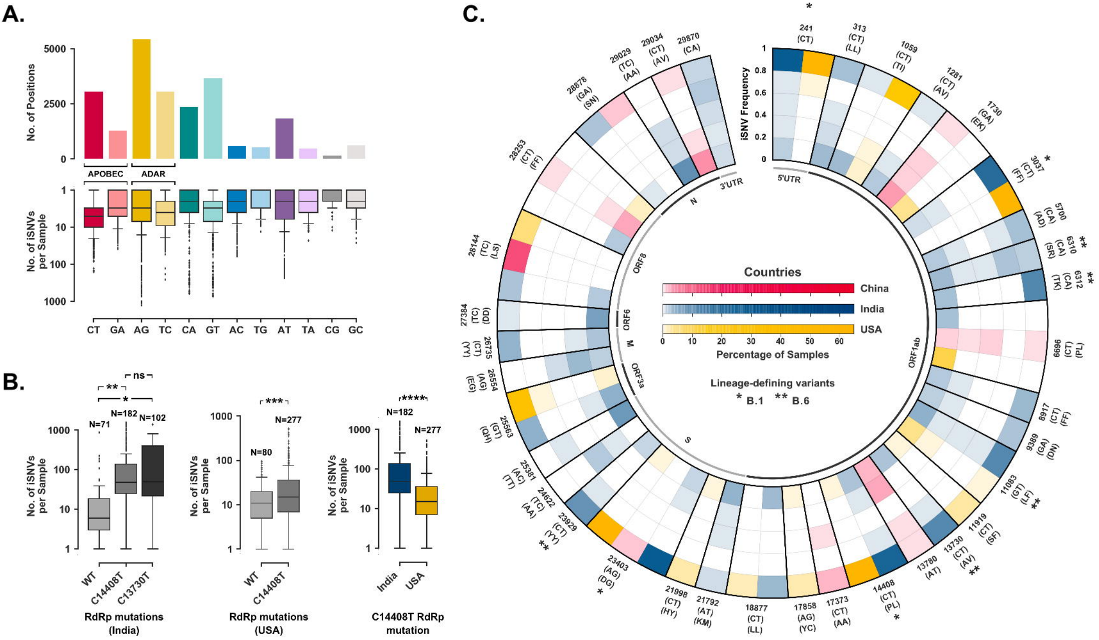
Supplementary Dataset S4: Summary of nucleotide change, codon change and amino acid sequence change of all recorded iSNV sites in Phase 1 samples and the number of samples with respective SNVs in GISAID

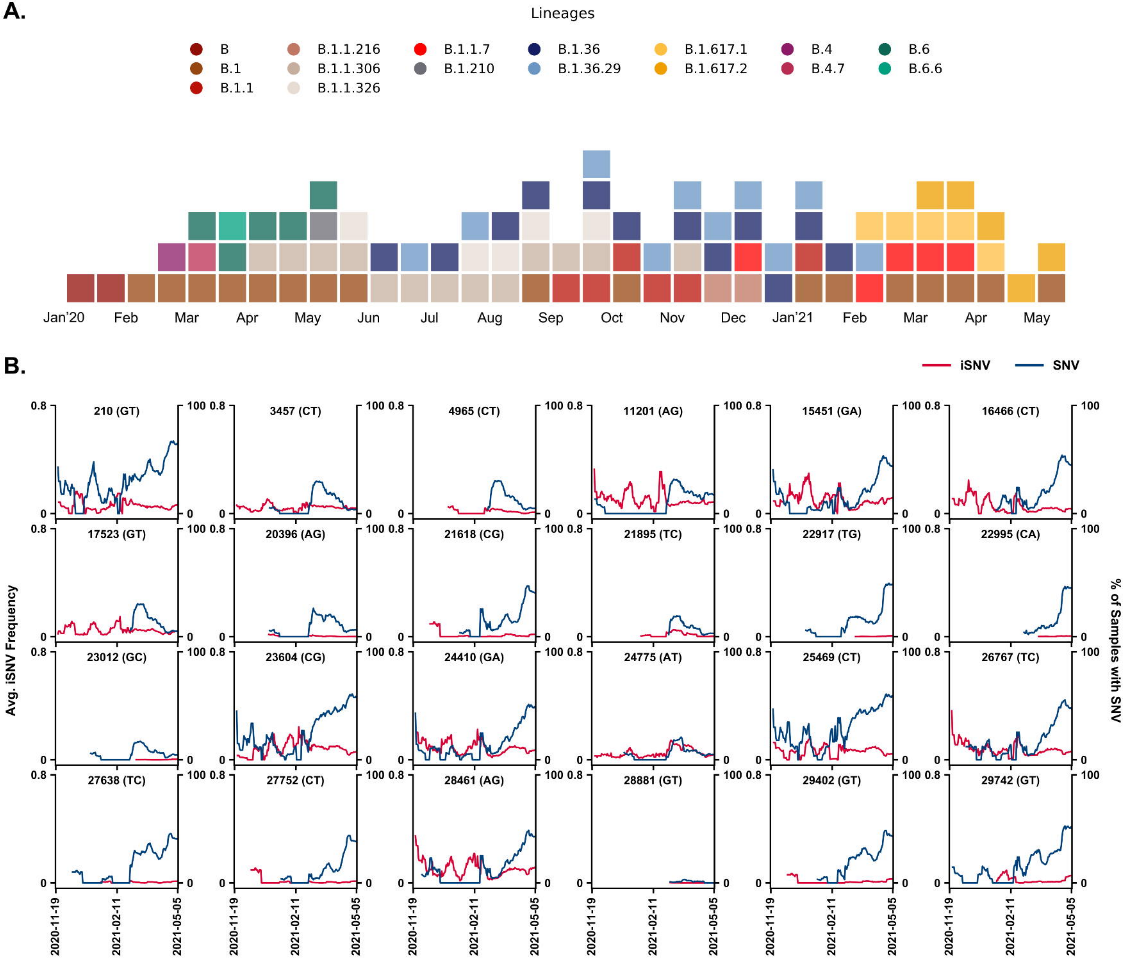
Supplementary Dataset S5: qRT-PCR results for mRNA expression of APOBEC3B and ADARB1 genes

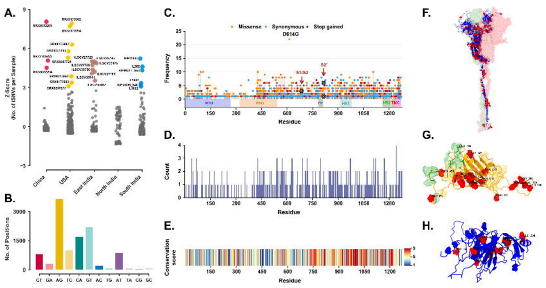
Supplementary Dataset S6: Sample-wise frequencies of all recorded iSNVs in SARS-CoV-2 infected cultures of Vero E6 cells

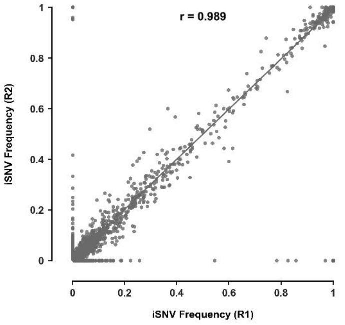
Supplementary Dataset S7: iSNVs mapped to lineage-defining variants of VOCs and VOIs listed by the WHO by 30th June 2021

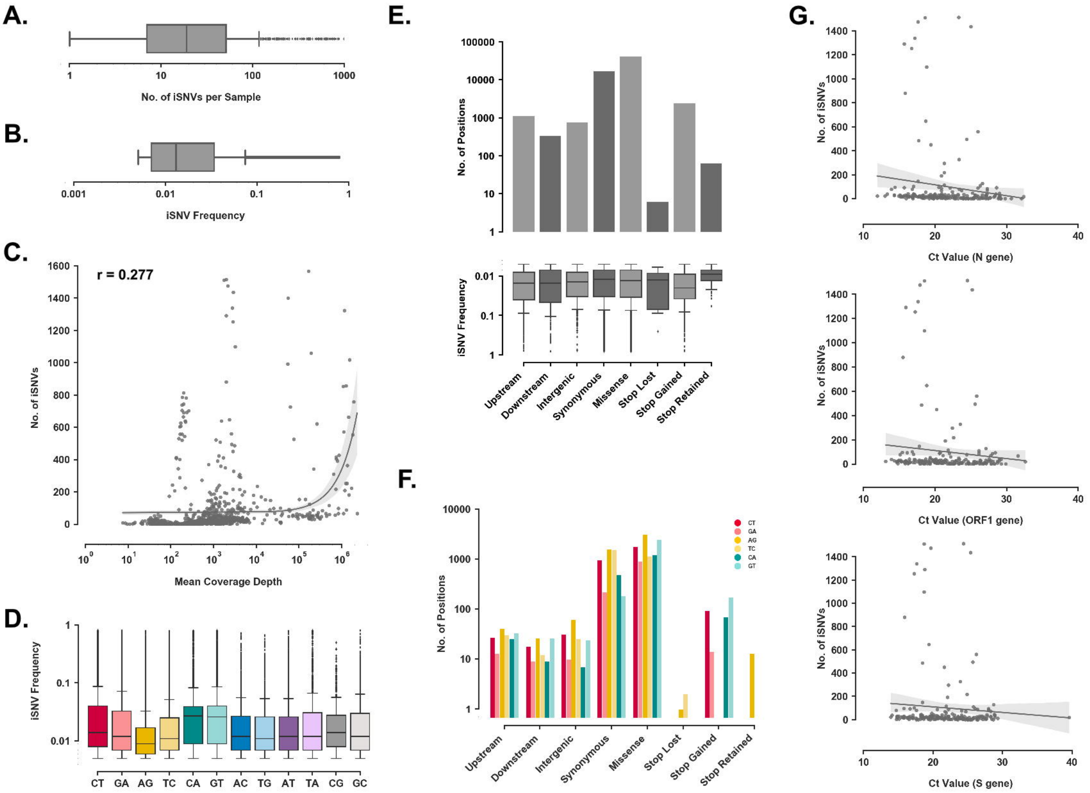
Supplementary Dataset S8: iSNVs mapped to potential immune-escape variants

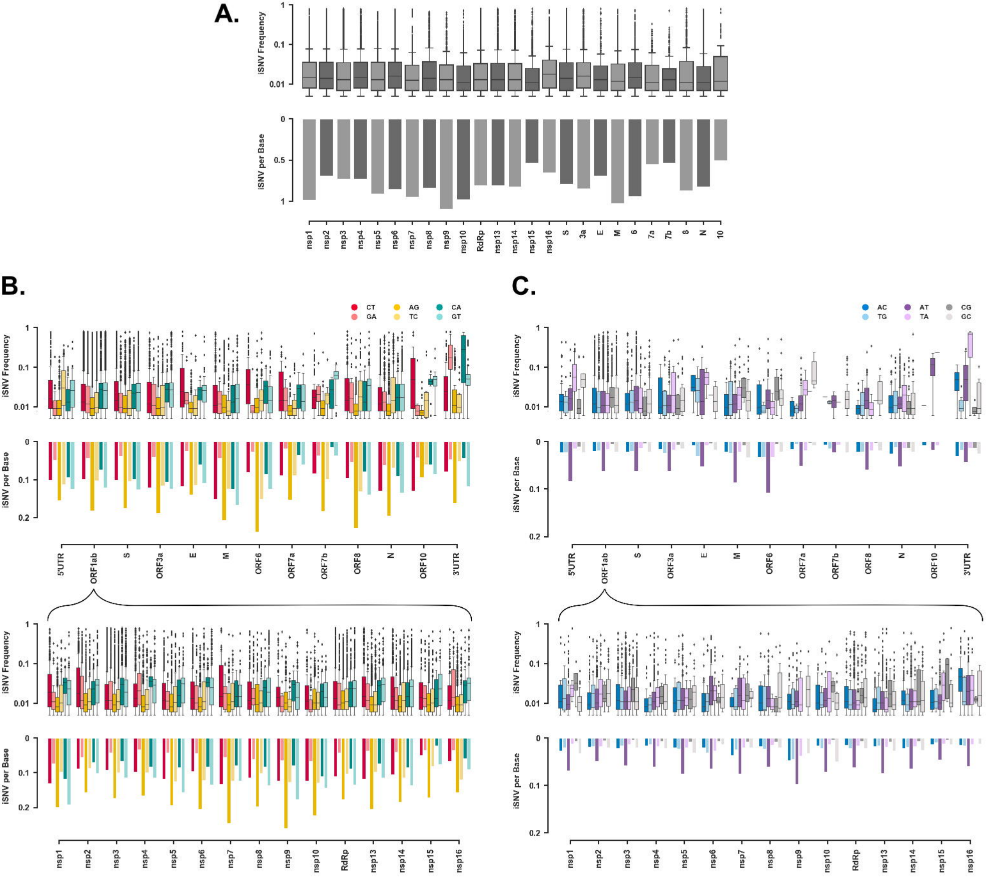
A.
No. of Positions
5000
2500
0
APOBEC
ADAR
No. of iSNVs per Sample
1
10
100
1000
CT
GA
AG
TC
CA
GT
AC
TG
AT
TA
CG
GC
B.
ns
**
*
N=71
N=182
N=102
No. of iSNVs per Sample
1000
100
10
1
WT
C14408T
C13730T
RdRp mutations (India)
***
N=80
N=277
No. of iSNVs per Sample
1000
100
10
1
WT
C14408T
RdRp mutations (USA)
****
N=182
N=277
No. of iSNVs per Sample
1000
100
10
1
India
USA
C14408T RdRp mutation
C.
iSNV Frequency
1
0.8
0.6
0.4
0.2
0
5'UTR
ORF1ab
S
ORF3a
M
ORF6
ORF8
N
3'UTR
Countries
China
India
USA
0
10
20
30
40
50
60
Percentage of Samples
Lineage-defining variants
* B.1 ** B.6
*
241 (CT)
313 (CT) (LL)
1059 (CT) (TI)
1281 (CT) (AV)
1730 (GA) (EK)
3037 (CT) (FF) *
5700 (CA) (AD)
6310 (CA) (SR) **
6312 (CA) (TK) **
6696 (CT) (PL)
8917 (CT) (FF)
9389 (GA) (DN)
11083 (GT) (LF) **
11919 (CT) (SF)
13730 (CT) (AV) **
13780 (AT)
14408 (CT) (PL) *
17373 (CT) (AA)
17858 (AG) (YC)
18877 (CT) (LL)
21792 (AT) (KM)
21998 (CT) (HY)
23403 (AG) (DG) *
23929 (CT) (YY) **
24622 (TC) (AA)
25381 (AC) (TT)
25563 (GT) (QH)
26554 (AG) (EG)
26735 (CT) (YY)
27384 (TC) (DD)
28144 (TC) (LS)
28253 (CT) (FF)
28878 (GA) (SN)
29029 (TC) (AA)
29034 (CT) (AV)
29870 (CA)

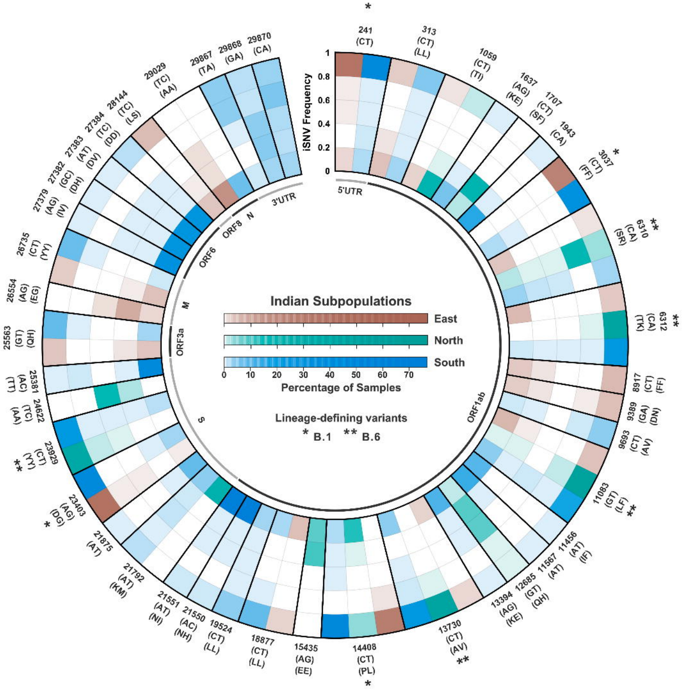

A.
Lineages
B
B.1
B.1.1
B.1.1.216
B.1.1.306
B.1.1.326
B.1.1.7
B.1.210
B.1.36
B.1.36.29
B.1.617.1
B.1.617.2
B.4
B.4.7
B.6
B.6.6
Jan'20
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Jan'21
Feb
Mar
Apr
May
B.
iSNV
SNV
Avg. iSNV Frequency
% of Samples with SNV
210 (GT)
3457 (CT)
4965 (CT)
11201 (AG)
15451 (GA)
16466 (CT)
17523 (GT)
20396 (AG)
21618 (CG)
21895 (TC)
22917 (TG)
22995 (CA)
23012 (GC)
23604 (CG)
24410 (GA)
24775 (AT)
25469 (CT)
26767 (TC)
27638 (TC)
27752 (CT)
28461 (AG)
28881 (GT)
29402 (GT)
29742 (GT)
0.8
0
100
0
2020-11-19
2021-02-11
2021-05-05

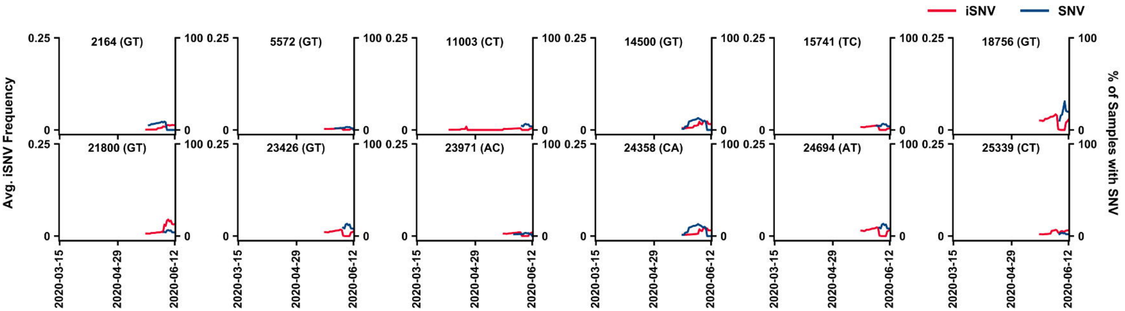

A.
Z-Score
(No. of ISNVs per Sample)
9
6
3
0
−3
SRX8032203
SRX8173982
SRX8173994
SRX8173867
ILSCV27725
ILSCV27728
ILSCV35750
ILSCV35141
ILSCV35476
ILSCV27783
SRX8067725
SRX8172883
L382
L3622
China
USA
East India
North India
South India
B.
No. of Positions
3000
1500
0
CT
GA
AG
TC
CA
GT
AC
TG
AT
TA
CG
GC
C.
Missense
Synonymous
Stop gained
D614G
S1/S2
S2'
Frequency
25
20
15
10
5
0
NTD
RBD
FP
HR1
HR2
TMC
0
150
300
450
600
750
900
1050
1200
Residue
D.
Count
4
3
2
1
0
150
300
450
600
750
900
1050
1200
Residue
E.
Conservation score
9
5
1
150
300
450
600
750
900
1050
1200
Residue
F.
G.
H.

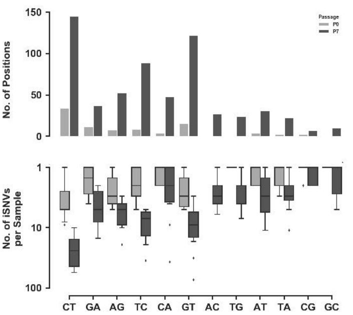

r = 0.989
iSNV Frequency (R2)
1
0.8
0.6
0.4
0.2
0
0
0.2
0.4
0.6
0.8
1
iSNV Frequency (R1)


A.
1
10
100
1000
No. of iSNVs per Sample
B.
0.001
0.01
0.1
1
iSNV Frequency
C.
r = 0.277
1600
1400
1200
1000
800
600
400
200
0
No. of iSNVs
10^0
10^1
10^2
10^3
10^4
10^5
10^6
Mean Coverage Depth
D.
1
0.1
0.01
iSNV Frequency
CT
GA
AG
TC
CA
GT
AC
TG
AT
TA
CG
GC
E.
100000
10000
1000
100
10
1
No. of Positions
0.01
0.1
1
iSNV Frequency
Upstream
Downstream
Intergenic
Synonymous
Missense
Stop Lost
Stop Gained
Stop Retained
F.
10000
1000
100
10
1
No. of Positions
● CT
● GA
● AG
● TC
● CA
● GT
Upstream
Downstream
Intergenic
Synonymous
Missense
Stop Lost
Stop Gained
Stop Retained
G.
1400
1200
1000
800
600
400
200
0
No. of iSNVs
10
20
30
40
Ct Value (N gene)
1400
1200
1000
800
600
400
200
0
No. of iSNVs
10
20
30
40
Ct Value (ORF1 gene)
1400
1200
1000
800
600
400
200
0
No. of iSNVs
10
20
30
40
Ct Value (S gene)


A.
iSNV Frequency
1
0.1
0.01
iSNV per Base
0
0.5
1
nsp1
nsp2
nsp3
nsp4
nsp5
nsp6
nsp7
nsp8
nsp9
nsp10
RdRp
nsp13
nsp14
nsp15
nsp16
S
3a
E
M
6
7a
7b
8
N
10
B.
CT
GA
AG
TC
CA
GT
iSNV Frequency
1
0.1
0.01
iSNV per Base
0
0.1
0.2
5'UTR
ORF1ab
S
ORF3a
E
M
ORF6
ORF7a
ORF7b
ORF8
N
ORF10
3'UTR
iSNV Frequency
1
0.1
0.01
iSNV per Base
0
0.1
0.2
nsp1
nsp2
nsp3
nsp4
nsp5
nsp6
nsp7
nsp8
nsp9
nsp10
RdRp
nsp13
nsp14
nsp15
nsp16
C.
AC
TG
AT
TA
CG
GC
iSNV Frequency
1
0.1
0.01
iSNV per Base
0
0.1
0.2
5'UTR
ORF1ab
S
ORF3a
E
M
ORF6
ORF7a
ORF7b
ORF8
N
ORF10
3'UTR
iSNV Frequency
1
0.1
0.01
iSNV per Base
0
0.1
0.2
nsp1
nsp2
nsp3
nsp4
nsp5
nsp6
nsp7
nsp8
nsp9
nsp10
RdRp
nsp13
nsp14
nsp15
nsp16


Indian Subpopulations
East
North
South
0
10
20
30
40
50
60
70
Percentage of Samples
Lineage-defining variants
* B.1 ** B.6
iSNV Frequency
1
0.8
0.6
0.4
0.2
0
5'UTR
ORF1ab
S
ORF3a
M
ORF6
ORF8
N
3'UTR
*
241 (CT)
313 (CT) (LL)
1059 (CT) (TI)
1637 (AG) (KE)
1707 (CT) (SF)
1943 (CA)
*
3037 (CT) (FF)
**
6310 (CA) (SR)
**
6312 (CA) (TK)
8917 (CT) (FF)
9389 (GA) (DN)
9693 (CT) (AV)
11083 (GT) (LF)
**
11456 (AT) (IF)
11567 (AT)
12685 (GT) (QH)
13394 (AG) (KE)
13730 (CT) (AV)
**
14408 (CT) (PL)
*
15435 (AG) (EE)
18877 (CT) (LL)
19524 (CT) (LL)
21550 (AC) (NH)
21551 (AT) (NI)
21792 (AT) (KM)
21875 (AT)
23403 (AG) (DG)
*
23929 (CT) (YY)
**
24622 (TC) (AA)
25381 (AC) (TT)
25563 (GT) (QH)
26554 (AG) (EG)
26735 (CT) (YY)
27379 (AG) (IV)
27382 (GC) (DH)
27383 (AT) (DV)
27384 (TC) (DD)
28144 (TC) (LS)
29029 (TC) (AA)
29867 (TA)
29868 (GA)
29870 (CA)


iSNV
SNV
Avg. iSNV Frequency
% of Samples with SNV
0.25
0
100
0
2164 (GT)
5572 (GT)
11003 (CT)
14500 (GT)
15741 (TC)
18756 (GT)
21800 (GT)
23426 (GT)
23971 (AC)
24358 (CA)
24694 (AT)
25339 (CT)
2020-03-15
2020-04-29
2020-06-12


150
100
50
0
No. of Positions
Passage
P0
P7
1
10
100
No. of iSNVs per Sample
CT
GA
AG
TC
CA
GT
AC
TG
AT
TA
CG
GC