bioRxiv preprint doi: https://doi.org/10.1101/2020.12.09.417519; this version posted December 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Full Title

Potential role of host-mediated RNA editing in intra-host variability of SARS-CoV-2 genomes

Short Title

RNA editing induced diversity in SARS-CoV-2 genomes

Authors

Ankit K. Pathak[1], Saman Fatihi[1], Tahseen Abbas[1,2], Bharathram Uppili[1], Gyan Prakash Mishra[3], Arup Ghosh[3], Sofia Banu[4], Rahul C. Bhoyar[1], Abhinav Jain[1,2], Mohit Kumar Divakar[1,2], Mohamed Imran[1,2], Mohammed Faruq[1], Divya Tej Sowpati[4], Sunil K. Raghav[3], Lipi Thukral[1], Mitali Mukerji[1]*

Author Affiliated Institutions

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

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

[3] *Institute of Life Sciences, Bhubaneswar, Odisha, India.*

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

*Correspondence: mitali@igib.res.in

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.09.417519; this version posted December 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

ABSTRACT

Since its zoonotic transmission in the human host, the SARS-CoV-2 virus has infected millions and has diversified extensively. A hallmark feature of viral system survival is their continuous evolution and adaptation within the host. RNA editing via APOBEC and ADAR family of enzymes has been recently implicated as the major driver of intra-host variability of the SARS-CoV-2 genomes. Analysis of the intra-host single-nucleotide variations (iSNVs) in SARS-CoV-2 genomes at spatio-temporal scales can provide insights on the consequence of RNA editing on the establishment, spread and functional outcomes of the virus. In this study, using 1,347 transcriptomes of COVID-19 infected patients across various populations, we find variable prevalence of iSNVs with distinctly higher levels in Indian population. Our results also suggest that iSNVs can likely establish variants in a population. These iSNVs may also contribute to key structural and functional changes in the Spike protein that confer antibody resistance.

Keywords

SARS-CoV-2, SNVs, iSNVs, RNA Editing, APOBEC, ADAR, Hyper-editing, Spike protein, host response, antiviral response

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.09.417519; this version posted December 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

INTRODUCTION

The SARS-CoV-2 pandemic has inflicted an enormous toll on human lives and livelihoods. As of December 1, 2020, 1.5 million individuals have succumbed to COVID-19 while 64 million have been confirmed diagnosed around the globe (*1*). The outcomes of viral infection are highly variable as apparent from regional distribution of spread, susceptibility across age groups, diversity in clinical symptoms, and significant molecular differences including tissue-level expressions in hosts. At genomics scale, thousands of viral genome sequences across the globe also reveal the differential patterns of diversity during the course of the pandemic. Besides, variability in the number of cases, mortality as well as recovery rates across different ethnicities is now being widely acknowledged (*2*). The spectrum of these evolving variations in different populations is suggestive of host-dependent factors. Failure to account for this variability can confound diagnostics and therapy that are dependent on the sequence and structure of the SARS-CoV-2 genomes.

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 *quasi-species*, in a host (*3*–*8*). Host-mediated RNA editing could contribute to the high mutation rates in retroviruses (*9*, *10*) besides other factors like high-rates of replication, low fidelity of RNA-dependent RNA polymerase and genomic recombination (*11*, *12*). In a recent study, RNA editing by host deaminases has been proposed as a major driver of intra-host heterogeneity in SARS-CoV-2 genomes (*13*). The APOBEC (apolipoprotein B mRNA editing catalytic polypeptide-like) and ADAR1 (Adenosine Deaminase RNA Specific, 1) protein families are the most prominent RNA editing enzymes implicated in innate antiretroviral defence for a large family of viruses (*9*, *10*, *14*–*16*) including the coronaviruses (*17*). 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. These edits could potentially alter function through changes in RNA secondary structure, regulatory domains, as well as protein function and potentially govern host-pathogen interactions.

Although random mutations have been shown to be largely deleterious (*18*, *19*), some strains may develop enhanced survival ability by evolving mechanisms to resist antiviral responses of the host immune system or drug resistance (*20*, *21*). The fitness of the within-host viral population is dependent on the cumulative contribution of all strains, including haplotypes harbouring minor variants (*22*). Analysis of intra-host single-nucleotide variations (iSNVs), observed as heteroplasmic sites (often referred in the mitochondrial

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.09.417519; this version posted December 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

genome context) in sequence reads, could provide insights into the genomic loci that contribute to the fitness of the virus within the host. In this study, we used transcriptomic data from 1,347 samples of COVID-19 diagnosed individuals from China, Germany, Malaysia, United Kingdom, United States and different subpopulations of India to characterize iSNVs in SARS-CoV-2 genomes. We observe extensive APOBEC and ADAR editing signatures throughout the SARS-CoV-2 genome with ADAR hyperactivity in some samples. This suggests that iSNVs in the SARS-CoV-2 genomes could serve as surrogates for ongoing host-mediated RNA editing activity. Interestingly, the prevalence of iSNVs seems to differ between cohorts indicating the role of background population in SARS-CoV-2 genomic diversity. We also identify genomic sites where iSNVs likely establish as variants in a population with time. Also, the iSNVs in the Spike protein reveal a high density of alterations in functionally critical residues that could alter the antigenicity and contribute to antibody resistance. This study highlights the need for capturing iSNVs to enable more accurate models for molecular epidemiology as well as for diagnostics and vaccine design.

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.09.417519; this version posted December 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

RESULTS

Intra-host SNVs as a consequence of potential RNA editing of SARS-CoV-2 genomes

We analysed 1,347 transcriptomic samples obtained from patients diagnosed with COVID-19 from different populations (Table 1). 1,160 samples could be aligned to the SARS-CoV-2 reference genome. We identified potential editing events as iSNVs by recording heteroplasmy in the sequence reads. To ensure specificity of iSNV detection, the filtering criteria and cutoffs were established by analysing an additional set of 500 samples sequenced in replicates (see “Materials and Methods”). 958 of the 1,160 samples (~83%) harboured one or more iSNVs with frequencies ranging between 0.005 and 0.80. We recorded a total of 86,595 iSNVs with a median of ~19 iSNVs per sample (fig. S1A), 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=0.259, fig. S1C).

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 and ADAR family of enzymes respectively (Fig. 1A). ADAR mediated A-to-G/T-to-C changes were observed to be more frequent accounting for about 36% of all variant positions, though these changes had the lowest median frequencies compared to the others (fig. S1D). Although A-to-G changes were seen in disproportionate numbers in few samples, APOBEC mediated C-to-T changes were more consistently observed with a median of 5 events per sample (Fig. 1A). Both C-to-T/G-to-A and A-to-G/T-to-C combinations contributed significantly yet similarly to synonymous and missense changes, whereas only the former contributed to stop gains (fig. S1E).

iSNVs contribute to variability throughout the SARS-CoV-2 genome, both in the coding as well as in the non-coding regions. A total of 18,216 genomic positions (4,961 polymorphic) (data S3) contributed to iSNVs in one or more samples distributed evenly in all genes and protein-coding domains of the viral genome (fig. S1F). Though the frequency spectrum of iSNVs was nearly uniform with respect to the nature of inflicted amino acid sequence change (fig. S1G), a large fraction of the sites resulted in non-synonymous changes (Fig. 1B). Gain of stop codons that could alter the amount of functional proteins from the viral genomes were also observed.

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

There was an overall difference in the prevalence of iSNVs between the populations that could have arisen due to differential editing of the SARS-CoV-2 genomes within the hosts (Fig. 1C). Indian subpopulations distinctly displayed a significantly higher number of iSNVs compared to populations in Europe, China and USA. The number of samples analysed in the populations of China, USA and India were comparable, thus ruling out

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.09.417519; this version posted December 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

any bias induced due to sampling. This highlights the role of the background population in ongoing variability in the SARS-CoV-2 genome.

The top 1% frequently observed heteroplasmic positions in major cohorts that exhibit a wide range of frequency differences are shown in Fig. 1D. Each concentric circle depicts iSNV frequency in bins of 0.2 while the colour gradient in each cell signifies the percentage of samples. Thus, the innermost ring represents a cohort-wise share of samples which show a possibility of editing 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 (fig. S2) suggesting that there are preferred sites in the SARS-CoV-2 genome that could be substrates for the host editing enzymes. Interestingly, all defining variants of the current dominant A2a clade (C241T, C3037T, C14408T and A23403G) and the India specific A3i/A4 clade (C6310A, C6312A, G11083T, C13730T and C23929T) (*23*, *24*) exhibit heteroplasmy in the samples. Noticeably, all A2a clade defining variants depict a C-to-T or A-to-G nucleotide change.

Spatio-temporal dynamics of iSNVs in populations

The diversity in the prevalence of viral strains that could have potentially arisen out of these editing events is variable during different times of the pandemic as reflected from the appearance and disappearance of the genomes harbouring certain SNVs that originate from heteroplasmic sites. For instance, A2a clade that originated in March continues to be the predominant clade in the population whereas the India specific A3i/A4 clade that was the most abundant clade in India in March seems to have nearly disappeared by May (Fig. 2A).

Though most of the sites seem to convert to a fixed variant in one or all populations (Fig. 1D, fig. S2), positions like 1707, 15435, 24622, 26554 and 29029 that depict an increasing trend in the frequency of heteroplasmic variants in some cohorts suggest that these iSNVs could get fixed in the population at a later date. In a geographically restricted population, such events may contribute to the origin of new variants/haplotypes within the population over a period of time. We observe this in the East Indian cohort where the appearance of SNVs for a range of positions was seen to coincide with or after the iSNVs first surfaced in the population (Fig. 2B). This further strengthens the postulate that host-mediated editing may introduce SNVs into the travelling viral strains in a population.

Potential functional consequence of hyper-editing on Spike protein

We observed a few samples to be substrates for hyper-editing events in each cohort i.e reflected by extensive heteroplasmy at multiple sites. In order to identify these population outliers, we computed the Z-score values based on the distribution of the number of iSNVs per sample (Fig. 3A). In these samples, we observed A-to-G substitutions accounting for about one-third of all changed genomic positions implying extensive ADAR

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.09.417519; this version posted December 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

mediated editing activity (Fig. 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 (*25*). Amongst all the variants potentially induced by editing in the Spike protein, 1,068 were found to be nonsynonymous, 322 synonymous and 91 stop-coding. Fig. 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 A2a clade defining variants) is near the S1/S2 cleavage site and has already been reported in the literature to be more geographically spreading than the D614 type (*26*).

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 Fig. 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 (Fig. 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 (*27*). 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 (*28*). Besides, N439K mutation in Spike has also been shown to be involved in immune escape from a panel of

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.09.417519; this version posted December 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

neutralizing monoclonal antibodies in another study (*29*). 20 out of these 23 positions overlap with sites that are potential sites of editing in our samples (data S3).

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.09.417519; this version posted December 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

DISCUSSIONS

The outcome of SARS-CoV-2 infection in individuals appears to be panning out differently across diverse populations with respect to prevalence as well as mortality in the number of deaths per million (*1*). From the initial strain reported in Wuhan, extensive variability in clade distribution has been observed across different parts of the world (*30*). This has been attributed to the containment of strains in lockdown conditions followed by the evolution of viral genomes during local transmissions. Many uncertain aspects of variable outcome of the same virus in different hosts, reinfection and long sequelae of COVID-19 is now getting highlighted, indicating an extensive cross-talk between the virus and host genomes.

In this study, we explored the variability in the SARS-CoV-2 genomes across populations as a surrogate for ongoing editing activity of the hosts. The APOBEC and ADAR family of enzymes are extensively reported to edit different families of viral genomes (*9*, *10*, *14–16*) including the coronaviruses (*17*). Previous genome analyses of many human-associated RNA viruses have revealed footprints of past editing events with a propensity for fixation of alleles that allow them to evade the host editing machinery (*31*). 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 (Fig. 1A) which corroborates with an earlier report (*13*). Recently, ADAR1 mediated editing in SARS-CoV-2 genomes has also been shown in transcriptome data of infected human cell-lines, Vero cells and clinical samples (*32*). Though we observed ADAR hyperactivity in some individuals, APOBEC-mediated C-to-T changes were more consistently observed in our samples (Fig. 1A) which may explain the overrepresentation of C-to-T SNVs in the SARS-CoV-2 populations (*33*, *34*). Besides the APOBEC and ADAR1 editing changes, 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 (*35*). 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 (*36*).

The difference in the prevalence of iSNVs that could arise out of editing events across populations (Fig. 1C) highlights 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 (*37*, *38*). Transmission of viruses through different host populations could thus shape the mutational landscape and govern the evolution of the viral genomes. Our group has previously shown a prominent insertion-deletion polymorphism in APOBEC3b which is associated with susceptibility to *Plasmodium falciparum* (*39*). The insertion allele that retains the

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.09.417519; this version posted December 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

APOBEC3b is nearly fixed in malaria-endemic regions in India and is associated with protection from severe malaria. The prevalence of COVID-19 related mortality is nearly negligible in Indian populations that are malaria-endemic and have a high frequency of APOBEC3b (*40*). Though the correlation of APOBEC insertion with protection still needs to be tested, this nevertheless suggests the role of such family of enzymes in evolutionary outcomes of SARS-CoV-2 infection and the burden of disease.

Editing 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. Though the extent of variability differs amongst individuals and populations, there seem to be preferred sites that are shared across populations (Fig. 1D, fig. S2). In most parts of the world, the A2a clade seems to have become the most prevalent with near disappearance of other clades which were once the predominant strain in certain local regions. For example, the A3i/A4 clade which was reported to be primarily endemic to India in April 2020 (*23*, *24*) had disappeared from the entire population as of October 2020 (Fig. 2A). One of the A2a clade defining variants, D614G (A23403G), has been shown to increase the infectivity of SARS-CoV-2 (*26*). Given that all A2a clade defining positions (241, 3037, 14408 and 23403) are heteroplasmic sites with C-to-T and A-to-G changes (Fig. 1D), editing may be fueling the ongoing positive selection of the SARS-CoV-2 genomes belonging to the A2a clade.

Recent studies have also suggested that the intra-host heterogeneity of SARS-CoV-2 genome in individuals may be transferable (*41–43*) and given that a minuscule percent of population is responsible for most of the local transmission of COVID-19 infection (*44–46*), there arises a possibility of iSNV harbouring genomes to be passed through a super-spreading event. This, in an isolated population, might also lead to an altered haplotype with some unique variations (*47*). Although, with the difference in scale of the number of people infected to the number of samples analysed, recording such events remains difficult. By temporally capturing heteroplasmy and the consequential variant appearance in the population, we could determine sites in the East Indian population which possibly transitioned into a variant as a consequence of editing (Fig. 2B). This indicates that conjoint analysis of editing with such gain and loss of variants in spatio-temporal scales might provide some clues of the evolvability of sites.

Ongoing selection in the host might have contributed 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 (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

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.09.417519; this version posted December 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

C-to-T substitutions elevate the frequency of hydrophobic amino acid coding codons in SARS-CoV-2 peptides (*48*). We also mapped 192 iSNVs onto the protein structure by utilizing a full-length Spike protein model of closed conformation (Fig. 3F). Some variants are located on the surface of the receptor-binding domain (RBD) which is responsible to bind to the ACE2 receptor in the host cell. These could 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 (Fig. 3G). 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.

Since SARS-CoV-2 has been recently transmitted to humans, differential editing is likely to confound interpretations of phenotypic outcomes of infection. As mentioned, the response of the host to SARS-CoV-2 infection has been extremely variable with different sequelae during and after infection (*49*, *50*). Few cases of reinfection are being reported where the reinfected viruses have many alterations and present with different symptoms (*51*, *52*). A relook at the sequence changes that have been reported in some reinfection studies shows considerable overlap with potential editing sites (data S4). Since many of these samples were studied using nanopore sequencing technique, there is a possibility that the heteroplasmy during the first event was not captured. Interestingly, in a recent communication (in press) where the subject was resampled after two months, few heteroplasmic sites in the first infection were observed to be fixed variants in the reinfection state as shown in data S4 (NCBI SRA Project ID PRJNA674796). Although, there were also some sites in the reinfection sample where this was not observed. Our meta-analysis of limited data suggests that a fraction of the reported reinfection cases could be a consequence of persistent editing of the viral genome within hosts. Since we observe that editing can lead to a change in the amino acid sequence, it can have implications on the structure and function of the protein. Noteworthy, we observed heteroplasmy in ~87% of the sites in the Spike protein that have been recently reported to confer antibody resistance. 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 (*53*). However, there are limitations to this analysis as the reports are few, the platforms were different, and these were not followup studies at regular intervals. These observations substantiate that editing within hosts may possibly lead to an evolved immune escape ability in some strains which may seem to be a case of reinfection in a host after weeks or months of the first incidence.

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 diagnostics, design of vaccines as well as predicting the spread and infectivity of viral strains in the population. Conjoint analysis with the host variability in editing machinery should be the next step.

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.09.417519; this version posted December 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

MATERIALS AND METHODS

Datasets

We accessed 1,347 Illumina NGS samples belonging to various regions to recognise differential signatures of RNA editing in SARS-CoV-2 infected populations. Only those samples that were paired-end sequenced were used for an increased degree of accuracy in recognising novel "iSNVs" that could arise out of RNA editing events. All publicly available samples submitted in the National Center for Biotechnology Information Sequence Read Archive (NCBI SRA) (*54*) till 23rd May 2020 from China, USA, Germany, Malaysia and United Kingdom were downloaded. Samples that had been sequenced in Bhubaneswar, Delhi and Hyderabad were also included that gave us a pan-India representation of SARS-CoV-2 infected East, North and South Indian populations (Table 1).

The GRCh37 human reference genome and NC_045512.2 SARS-CoV-2 Wuhan-Hu-1 reference genome were used for all analysis. To get a position-wise share of SNVs in the global SARS-CoV-2 population, a total of 94,130 high-coverage FASTA sequences submitted in GISAID (*55*) till 30th September 2020 were downloaded. Date of collection and clade information of Indian samples for Fig. 2A was retrieved from GEAR-19 (*56*).

Data processing

The sequence files retrieved from NCBI SRA were downloaded in the SRA compressed file format and were converted to FASTQ format files using NCBI's SRA-Tools (*57*). In order to trim the adaptor sequences and filter low-quality reads from downstream analysis, we used Trimmomatic (*58*). The trimmed reads were then aligned against the human reference genome (GRCh37/hg19) using HISAT2 (*59*). All reads which couldn't be aligned in pairs to the human reference genome were filtered out from the human aligned files using SAMtools (*60*) and mapped to the SARS-CoV-2 reference genome (NC_045512.2) using Burrows-Wheeler Aligner (BWA-MEM) (*61*). Since RNA editing events deal with small fractions of reads, to avoid any bias induced by PCR amplified reads (*62*), all duplicate reads were removed from the SARS-CoV-2 aligned files using Picard Tools (*63*). BCFtools (*60*) was then used to generate the consensus SARS-CoV-2 FASTA sequence. FastQC (*64*), QualiMap (*65*) and SAMtools (*60*) were used to obtain sample quality scores and alignment statistics at each step. MAFFT (*66*) was used to create multiple sequence alignments of consensus FASTA sequences of the processed samples and the downloaded FASTA sequences from GISAID. We used the 2019nCoVR database (*67*) to annotate the sites for their codon change, amino acid sequence change and the nature of change.

We used REDItools2 (*68*) 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

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.09.417519; this version posted December 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

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 (data S5) were excluded due to known sequencing errors in these regions (66). From the iSNVs called in the REDItools2 output file, potential editing events were 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 potential editing events 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, fig. S3) in sites qualifying the pipeline thresholds. Thus, we used these cutoffs as a qualifying criterion for defining a heteroplasmic site as a potential editing site in our subsequent analyses.

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 (*69*). Samples that showed a Z-score value > 3 were classified as hyper-edited in each cohort (Fig. 3A).

In order to categorize the specific amino acid change and the proteins containing the iSNVs, they were annotated using SnpEff version 4.5 (*70*). 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 (*71*). Conservation calculation of Spike amino acids was done using the ConSurf tool (*72*). 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 (*73*) to determine interface residues from the available crystal structures. PDB 6M0J (*74*) and 6M17 (*75*) for ACE2 RBD binding and PDB 6W41 (*76*) and 7BZ5 (*77*) for antibody binding residues were used from Research Collaboratory for Structural Bioinformatics Protein Data Bank (RCSB PDB) (*78*).

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 (*79*), Pandas (*80*), Matplotlib (*81*) and Seaborn (*82*) were used for data handling and visualizations. SciPy was used to compute pairwise two-sided t-tests and to calculate distribution statistics and Pearson's correlation coefficients.

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.09.417519; this version posted December 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

REFERENCES AND NOTES

1. WHO coronavirus disease (COVID-19) dashboard. Geneva: World Health Organization, 2020. (2020), (available at https://covid19.who.int).
2. Johns Hopkins Coronavirus Resource Center. *Johns Hopkins Coronavirus Resour. Cent.* (2020), (available at https://coronavirus.jhu.edu/map.html).
3. J. C. Stack, P. R. Murcia, B. T. Grenfell, J. L. N. Wood, E. C. Holmes, Inferring the inter-host transmission of influenza A virus using patterns of intra-host genetic variation. *Proc. R. Soc. B Biol. Sci.* **280**, 20122173 (2013).
4. M. Salemi, The intra-host evolutionary and population dynamics of human immunodeficiency virus type 1: a phylogenetic perspective. *Infect. Dis. Rep.* **5**, 3 (2013).
5. C. M. Romano, M. Lauck, F. S. Salvador, C. R. Lima, L. S. Villas-Boas, E. S. A. Araújo, J. E. Levi, C. S. Pannuti, D. O'Connor, E. G. Kallas, Inter- and Intra-Host Viral Diversity in a Large Seasonal DENV2 Outbreak. *PLoS ONE*. **8**, e70318 (2013).
6. M. Ni, C. Chen, J. Qian, H.-X. Xiao, W.-F. Shi, Y. Luo, H.-Y. Wang, Z. Li, J. Wu, P.-S. Xu, S.-H. Chen, G. Wong, Y. Bi, Z.-P. Xia, W. Li, H. Lu, J. Ma, Y.-G. Tong, H. Zeng, S.-Q. Wang, G. F. Gao, X.-C. Bo, D. Liu, Intra-host dynamics of Ebola virus during 2014. *Nat. Microbiol.* **1**, 16151 (2016).
7. J. E. Muñoz-Medina, M. A. Garcia-Knight, A. Sanchez-Flores, I. E. Monroy-Muñoz, R. Grande, J. Esbjörnsson, C. E. Santacruz-Tinoco, C. R. González-Bonilla, Evolutionary analysis of the Chikungunya virus epidemic in Mexico reveals intra-host mutational hotspots in the E1 protein. *PLOS ONE*. **13**, e0209292 (2018).
8. R. Cattaneo, R. C. Donohue, A. R. Generous, C. K. Navaratnarajah, C. K. Pfaller, Stronger together: Multi-genome transmission of measles virus. *Virus Res.* **265**, 74–79 (2019).
9. M. H. Malim, APOBEC proteins and intrinsic resistance to HIV-1 infection. *Philos. Trans. R. Soc. B Biol. Sci.* **364**, 675–687 (2009).
10. R. Suspene, V. Petit, D. Puyraimond-Zemmour, M.-M. Aynaud, M. Henry, D. Guetard, C. Rusniok, S. Wain-Hobson, J.-P. Vartanian, 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 (2011).
11. W. Hu, H. Temin, Retroviral recombination and reverse transcription. *Science*. **250**, 1227–1233 (1990).
12. D. A. Steinhauer, E. Domingo, J. J. Holland, Lack of evidence for proofreading mechanisms associated with an RNA virus polymerase. *Gene*. **122**, 281–288 (1992).
13. S. Di Giorgio, F. Martignano, M. G. Torcia, G. Mattiuz, S. G. Conticello, Evidence for host-dependent RNA editing in the transcriptome of SARS-CoV-2. *Sci. Adv.* **6**, eabb5813 (2020).
14. C. Rösler, J. Köck, M. Kann, M. H. Malim, H. E. Blum, T. F. Baumert, F. von Weizsäcker, APOBEC-mediated interference with hepadnavirus production. *Hepatology*. **42**, 301–309 (2005).
15. V. V. Khrustalev, T. A. Khrustaleva, N. Sharma, R. Giri, Mutational Pressure in Zika Virus: Local ADAR-Editing Areas Associated with Pauses in Translation and Replication. *Front. Cell. Infect. Microbiol.* **7** (2017), doi:10.3389/fcimb.2017.00044.
16. K. N. Bishop, R. K. Holmes, A. M. Sheehy, N. O. Davidson, S.-J. Cho, M. H. Malim, Cytidine Deamination of Retroviral DNA by Diverse APOBEC Proteins. *Curr. Biol.* **14**, 1392–1396 (2004).
17. A. Milewska, E. Kindler, P. Vkovski, S. Zeglen, M. Ochman, V. Thiel, Z. Rajfur, K. Pyrc, APOBEC3-mediated restriction of RNA virus replication. *Sci. Rep.* **8**, 5960 (2018).
18. R. Sanjuán, A. Moya, S. F. Elena, The distribution of fitness effects caused by single-nucleotide substitutions in an RNA virus. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 8396–8401 (2004).
19. R. Sanjuán, 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 (2010).
20. S. Venkatesan, R. Rosenthal, N. Kanu, N. McGranahan, J. Bartek, S. A. Quezada, J. Hare, R. S. Harris, C. Swanton, Perspective: APOBEC mutagenesis in drug resistance and immune escape in HIV and

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.09.417519; this version posted December 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

cancer evolution. *Ann. Oncol. Off. J. Eur. Soc. Med. Oncol.* **29**, 563–572 (2018).
21. A. Herbert, ADAR and Immune Silencing in Cancer. *Trends Cancer*. **5**, 272–282 (2019).
22. A. V. Bordería, O. Isakov, G. Moratorio, R. Henningsson, S. Agüera-González, L. Organtini, N. F. Gnädig, H. Blanc, A. Alcover, S. Hafenstein, M. Fontes, N. Shomron, M. Vignuzzi, Group Selection and Contribution of Minority Variants during Virus Adaptation Determines Virus Fitness and Phenotype. *PLOS Pathog.* **11**, e1004838 (2015).
23. S. Banu, B. Jolly, P. Mukherjee, P. Singh, S. Khan, L. Zaveri, S. Shambhavi, N. Gaur, S. Reddy, K. Kaveri, S. Srinivasan, D. R. Gopal, A. B. Siva, K. Thangaraj, K. B. Tallapaka, R. K. Mishra, V. Scaria, D. T. Sowpati, A Distinct Phylogenetic Cluster of Indian Severe Acute Respiratory Syndrome Coronavirus 2 Isolates. *Open Forum Infect. Dis.* **7**, ofaa434 (2020).
24. P. Kumar, R. Pandey, P. Sharma, M. S. Dhar, V. A., B. Uppili, H. Vashisht, S. Wadhwa, N. Tyagi, S. Fatihi, U. Sharma, P. Singh, H. Lall, M. Datta, P. Gupta, N. Saini, A. Tewari, B. Nandi, D. Kumar, S. Bag, D. Gahlot, S. Rathore, N. Jatana, V. Jaiswal, H. Gogia, P. Madan, S. Singh, P. Singh, D. Dash, M. Bala, S. Kabra, S. Singh, M. Mukerji, L. Thukral, M. Faruq, A. Agrawal, P. Rakshit, Integrated genomic view of SARS-CoV-2 in India. *Wellcome Open Res.* **5**, 184 (2020).
25. M. Hoffmann, H. Kleine-Weber, S. Schroeder, N. Krüger, T. Herrler, S. Erichsen, T. S. Schiergens, G. Herrler, N.-H. Wu, A. Nitsche, M. A. Müller, C. Drosten, S. Pöhlmann, SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. *Cell*. **181**, 271-280.e8 (2020).
26. B. Korber, W. M. Fischer, S. Gnanakaran, H. Yoon, J. Theiler, W. Abfalterer, N. Hengartner, E. E. Giorgi, T. Bhattacharya, B. Foley, K. M. Hastie, M. D. Parker, D. G. Partridge, C. M. Evans, T. M. Freeman, T. I. de Silva, C. McDanal, L. G. Perez, H. Tang, A. Moon-Walker, S. P. Whelan, C. C. LaBranche, E. O. Saphire, D. C. Montefiori, A. Angyal, R. L. Brown, L. Carrilero, L. R. Green, D. C. Groves, K. J. Johnson, A. J. Keeley, B. B. Lindsey, P. J. Parsons, M. Raza, S. Rowland-Jones, N. Smith, R. M. Tucker, D. Wang, M. D. Wyles, Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus. *Cell*. **182**, 812-827.e19 (2020).
27. B. Choi, M. C. Choudhary, J. Regan, J. A. Sparks, R. F. Padera, X. Qiu, I. H. Solomon, H.-H. Kuo, J. Boucau, K. Bowman, U. D. Adhikari, M. L. Winkler, A. A. Mueller, T. Y.-T. Hsu, M. Desjardins, L. R. Baden, B. T. Chan, B. D. Walker, M. Lichterfeld, M. Brigl, D. S. Kwon, S. Kanjilal, E. T. Richardson, A. H. Jonsson, G. Alter, A. K. Barczak, W. P. Hanage, X. G. Yu, G. D. Gaiha, M. S. Seaman, M. Cernadas, J. Z. Li, Persistence and Evolution of SARS-CoV-2 in an Immunocompromised Host. *N. Engl. J. Med.*, NEJMc2031364 (2020).
28. A. Baum, B. O. Fulton, E. Wloga, R. Copin, K. E. Pascal, V. Russo, S. Giordano, K. Lanza, N. Negron, M. Ni, Y. Wei, G. S. Atwal, A. J. Murphy, N. Stahl, G. D. Yancopoulos, C. A. Kyratsous, Antibody cocktail to SARS-CoV-2 spike protein prevents rapid mutational escape seen with individual antibodies. *Science*, eabd0831 (2020).
29. E. C. Thomson, L. E. Rosen, J. G. Shepherd, R. Spreafico, A. da Silva Filipe, J. A. Wojcechowskyj, C. Davis, L. Piccoli, D. J. Pascall, J. Dillen, S. Lytras, N. Czudnochowski, R. Shah, M. Meury, N. Jesudason, A. De Marco, K. Li, J. Bassi, A. O'Toole, D. Pinto, R. M. Colquhoun, K. Culap, B. Jackson, F. Zatta, A. Rambaut, S. Jaconi, V. B. Sreenu, J. Nix, R. F. Jarrett, M. Beltramello, K. Nomikou, M. Pizzuto, L. Tong, E. Cameroni, N. Johnson, A. Wickenhagen, A. Ceschi, D. Mair, P. Ferrari, K. Smollett, F. Sallusto, S. Carmichael, C. Garzoni, J. Nichols, M. Galli, J. Hughes, A. Riva, A. Ho, M. G. Semple, P. J. M. Openshaw, J. K. Baillie, The ISARIC4C Investigators, the COVID-19 Genomics UK (COG-UK) consortium, S. J. Rihn, S. J. Lycett, H. W. Virgin, A. Telenti, D. Corti, D. L. Robertson, G. Snell, "The circulating SARS-CoV-2 spike variant N439K maintains fitness while evading antibody-mediated immunity" (preprint, Microbiology, 2020), , doi:10.1101/2020.11.04.355842.
30. Nextstrain: SARS-CoV-2 (2020), (available at https://nextstrain.org/ncov/global).
31. F. Poulain, N. Lejeune, K. Willemart, N. A. Gillet, Footprint of the host restriction factors APOBEC3 on the genome of human viruses. *PLOS Pathog.* **16**, e1008718 (2020).
32. E. Picardi, L. Mansi, G. Pesole, "A-to-I RNA editing in SARS-COV-2: real or artifact?" (preprint, Genomics, 2020), , doi:10.1101/2020.07.27.223172.

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.09.417519; this version posted December 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

33. P. Simmonds, 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 (2020).
34. R. Wang, Y. Hozumi, Y.-H. Zheng, C. Yin, G.-W. Wei, Host Immune Response Driving SARS-CoV-2 Evolution. *Viruses*. **12**, 1095 (2020).
35. A. Graudenzi, D. Maspero, F. Angaroni, R. Piazza, D. Ramazzotti, "Mutational Signatures and Heterogeneous Host Response Revealed Via Large-Scale Characterization of SARS-COV-2 Genomic Diversity" (preprint, Genetics, 2020), , doi:10.1101/2020.07.06.189944.
36. F. C. Camini, C. C. da Silva Caetano, L. T. Almeida, C. L. de Brito Magalhães, Implications of oxidative stress on viral pathogenesis. *Arch. Virol.* **162**, 907–917 (2017).
37. J. M. Kidd, T. L. Newman, E. Tuzun, R. Kaul, E. E. Eichler, Population Stratification of a Common APOBEC Gene Deletion Polymorphism. *PLoS Genet.* **3**, e63 (2007).
38. E. Park, J. Guo, S. Shen, L. Demirdjian, Y. N. Wu, L. Lin, Y. Xing, Population and allelic variation of A-to-I RNA editing in human transcriptomes. *Genome Biol.* **18**, 143 (2017).
39. P. Jha, S. Sinha, K. Kanchan, T. Qidwai, A. Narang, P. K. Singh, S. S. Pati, S. Mohanty, S. K. Mishra, S. K. Sharma, S. Awasthi, V. Venkatesh, S. Jain, A. Basu, S. Xu, Indian Genome Variation Consortium, M. Mukerji, S. Habib, Deletion of the APOBEC3B gene strongly impacts susceptibility to falciparum malaria. *Infect. Genet. Evol. J. Mol. Epidemiol. Evol. Genet. Infect. Dis.* **12**, 142–148 (2012).
40. H. Saman, M. Mitali, APOBEC3B and ACE1 indel polymorphisms as prima facie candidates for protection from COVID-19 (2020).
41. K. A. Lythgoe, M. Hall, L. Ferretti, M. de Cesare, G. MacIntyre-Cockett, A. Trebes, M. Andersson, N. Otecko, E. L. Wise, N. Moore, J. Lynch, S. Kidd, N. Cortes, M. Mori, A. Justice, A. Green, M. A. Ansari, L. Abeler-Dörner, C. E. Moore, T. E. A. Peto, R. Shaw, P. Simmonds, D. Buck, J. A. Todd, on behalf of OVSG Analysis Group, D. Bonsall, C. Fraser, T. Golubchik, "Shared SARS-CoV-2 diversity suggests localised transmission of minority variants" (preprint, Genomics, 2020), , doi:10.1101/2020.05.28.118992.
42. D. Wang, Y. Wang, W. Sun, L. Zhang, J. Ji, Z. Zhang, X. Cheng, Y. Li, F. Xiao, A. Zhu, B. Zhong, S. Ruan, J. Li, P. Ren, Z. Ou, M. Xiao, M. Li, Z. Deng, H. Zhong, F. Li, W. Wang, Y. Zhang, W. Chen, S. Zhu, X. Xu, X. Jin, J. Zhao, N. Zhong, W. Zhang, J. Zhao, J. Li, Y. Xu, "Population Bottlenecks and Intra-host Evolution during Human-to-Human Transmission of SARS-CoV-2" (preprint, Genetics, 2020), , doi:10.1101/2020.06.26.173203.
43. N. Sapoval, M. Mahmoud, M. D. Jochum, Y. Liu, R. A. Leo Elworth, Q. Wang, D. Albin, H. Ogilvie, M. D. Lee, S. Villapol, K. M. Hernandez, I. M. Berry, J. Foox, A. Beheshti, K. Ternus, K. M. Aagaard, D. Posada, C. E. Mason, F. Sedlazeck, T. J. Treangen, "Hidden genomic diversity of SARS-CoV-2: implications for qRT-PCR diagnostics and transmission" (preprint, Genomics, 2020), , doi:10.1101/2020.07.02.184481.
44. A. Endo, Centre for the Mathematical Modelling of Infectious Diseases COVID-19 Working Group, S. Abbott, A. J. Kucharski, S. Funk, Estimating the overdispersion in COVID-19 transmission using outbreak sizes outside China. *Wellcome Open Res.* **5**, 67 (2020).
45. D. Miller, M. A. Martin, N. Harel, O. Tirosh, T. Kustin, M. Meir, N. Sorek, S. Gefen-Halevi, S. Amit, O. Vorontsov, A. Shaag, D. Wolf, A. Peretz, Y. Shemer-Avni, D. Roif-Kaminsky, N. M. Kopelman, A. Huppert, K. Koelle, A. Stern, Full genome viral sequences inform patterns of SARS-CoV-2 spread into and within Israel. *Nat. Commun.* **11**, 5518 (2020).
46. D. C. Adam, P. Wu, J. Y. Wong, E. H. Y. Lau, T. K. Tsang, S. Cauchemez, G. M. Leung, B. J. Cowling, Clustering and superspreading potential of SARS-CoV-2 infections in Hong Kong. *Nat. Med.* **26**, 1714–1719 (2020).
47. G. K. Moreno, K. M. Braun, K. K. Riemersma, M. A. Martin, P. J. Halfmann, C. M. Crooks, T. Prall, D. Baker, J. J. Baczenas, A. S. Heffron, M. Ramuta, M. Khubbar, A. M. Weiler, M. A. Accola, W. M. Rehrauer, S. L. O'Connor, N. Safdar, C. S. Pepperell, T. Dasu, S. Bhattacharyya, Y. Kawaoka, K. Koelle, D. H. O'Connor, T. C. Friedrich, Revealing fine-scale spatiotemporal differences in SARS-CoV-2 introduction and spread. *Nat. Commun.* **11**, 5558 (2020).
48. R. Matyášek, A. Kovařík, Mutation Patterns of Human SARS-CoV-2 and Bat RaTG13 Coronavirus

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.09.417519; this version posted December 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Genomes Are Strongly Biased Towards C>U Transitions, Indicating Rapid Evolution in Their Hosts. *Genes*. **11**, 761 (2020).
49. C. del Rio, L. F. Collins, P. Malani, Long-term Health Consequences of COVID-19. *JAMA*. **324**, 1723 (2020).
50. Understanding the long-term health effects of COVID-19. *EClinicalMedicine*. **26**, 100586 (2020).
51. R. L. Tillett, J. R. Sevinsky, P. D. Hartley, H. Kerwin, N. Crawford, A. Gorzalski, C. Laverdure, S. C. Verma, C. C. Rossetto, D. Jackson, M. J. Farrell, S. Van Hooser, M. Pandori, Genomic evidence for reinfection with SARS-CoV-2: a case study. *Lancet Infect. Dis.* (2020), doi:10.1016/S1473-3099(20)30764-7.
52. K. K.-W. To, I. F.-N. Hung, J. D. Ip, A. W.-H. Chu, W.-M. Chan, A. R. Tam, C. H.-Y. Fong, S. Yuan, H.-W. Tsoi, A. C.-K. Ng, L. L.-Y. Lee, P. Wan, E. Y.-K. Tso, W.-K. To, D. N.-C. Tsang, K.-H. Chan, J.-D. Huang, K.-H. Kok, V. C.-C. Cheng, K.-Y. Yuen, Coronavirus Disease 2019 (COVID-19) Re-infection by a Phylogenetically Distinct Severe Acute Respiratory Syndrome Coronavirus 2 Strain Confirmed by Whole Genome Sequencing. *Clin. Infect. Dis.*, ciaa1275 (2020).
53. L. Wang, L. Wang, H. Zhuang, Profiling and characterization of SARS-CoV-2 mutants' infectivity and antigenicity. *Signal Transduct. Target. Ther.* **5**, 185 (2020).
54. R. Leinonen, H. Sugawara, M. Shumway, International Nucleotide Sequence Database Collaboration, The sequence read archive. *Nucleic Acids Res.* **39**, D19-21 (2011).
55. Y. Shu, J. McCauley, GISAID: Global initiative on sharing all influenza data - from vision to reality. *Euro Surveill. Bull. Eur. Sur Mal. Transm. Eur. Commun. Dis. Bull.* **22** (2017), doi:10.2807/1560-7917.ES.2017.22.13.30494.
56. A. K. Avvaru, S. Banu, P. Singh, D. T. Sowpati, Genome evolution analysis resource for covid19 (GEAR-19) (2020), (available at https://data.ccmb.res.in/gear19/).
57. SRA Toolkit Development Team, The NCBI SRA Toolkit, (available at https://ncbi.github.io/sra-tools/).
58. A. M. Bolger, M. Lohse, B. Usadel, Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinforma. Oxf. Engl.* **30**, 2114–2120 (2014).
59. D. Kim, J. M. Paggi, C. Park, C. Bennett, S. L. Salzberg, Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019).
60. H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, 1000 Genome Project Data Processing Subgroup, The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. **25**, 2078–2079 (2009).
61. H. Li, Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *ArXiv13033997 Q-Bio* (2013) (available at http://arxiv.org/abs/1303.3997).
62. V. Marx, How to deduplicate PCR. *Nat. Methods*. **14**, 473–476 (2017).
63. Broad Institute, Picard Toolkit. (2019), (available at https://github.com/broadinstitute/picard).
64. S. Andrews, FastQC: a quality control tool for high throughput sequence data (2010), (available at http://www.bioinformatics.babraham.ac.uk/projects/fastqc/).
65. K. Okonechnikov, A. Conesa, F. García-Alcalde, Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinforma. Oxf. Engl.* **32**, 292–294 (2016).
66. K. Katoh, D. M. Standley, MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
67. W.-M. Zhao, S.-H. Song, M.-L. Chen, D. Zou, L.-N. Ma, Y.-K. Ma, R.-J. Li, L.-L. Hao, C.-P. Li, D.-M. Tian, B.-X. Tang, Y.-Q. Wang, J.-W. Zhu, H.-X. Chen, Z. Zhang, Y.-B. Xue, Y.-M. Bao, The 2019 novel coronavirus resource. *Yi Chuan Hered.* **42**, 212–221 (2020).
68. E. Picardi, G. Pesole, REDItools: high-throughput RNA editing detection made easy. *Bioinformatics*. **29**, 1813–1814 (2013).
69. E. Jones, T. Oliphant, P. Peterson, SciPy: Open source scientific tools for Python. *SciPy Open Source Sci. Tools Python* (2001), (available at https://www.scipy.org/).
70. P. Cingolani, A. Platts, L. L. Wang, M. Coon, T. Nguyen, L. Wang, S. J. Land, X. Lu, D. M. Ruden, 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 (2012).

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.09.417519; this version posted December 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

71. F. Sievers, A. Wilm, D. Dineen, T. J. Gibson, K. Karplus, W. Li, R. Lopez, H. McWilliam, M. Remmert, J. Söding, J. D. Thompson, D. G. Higgins, Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7**, 539 (2011).
72. H. Ashkenazy, S. Abadi, E. Martz, O. Chay, I. Mayrose, T. Pupko, N. Ben-Tal, ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. *Nucleic Acids Res.* **44**, W344-350 (2016).
73. Schrodinger LLC, The PyMOL Molecular Graphics System, Version 1.8 (2010), (available at https://pymol.org/2/).
74. J. Lan, J. Ge, J. Yu, S. Shan, H. Zhou, S. Fan, Q. Zhang, X. Shi, Q. Wang, L. Zhang, X. Wang, Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. *Nature*. **581**, 215–220 (2020).
75. R. Yan, Y. Zhang, Y. Li, L. Xia, Y. Guo, Q. Zhou, Structural basis for the recognition of SARS-CoV-2 by full-length human ACE2. *Science*. **367**, 1444–1448 (2020).
76. M. Yuan, N. C. Wu, X. Zhu, C.-C. D. Lee, R. T. Y. So, H. Lv, C. K. P. Mok, I. A. Wilson, A highly conserved cryptic epitope in the receptor binding domains of SARS-CoV-2 and SARS-CoV. *Science*. **368**, 630–633 (2020).
77. Y. Wu, F. Wang, C. Shen, W. Peng, D. Li, C. Zhao, Z. Li, S. Li, Y. Bi, Y. Yang, Y. Gong, H. Xiao, Z. Fan, S. Tan, G. Wu, W. Tan, X. Lu, C. Fan, Q. Wang, Y. Liu, C. Zhang, J. Qi, G. F. Gao, F. Gao, L. Liu, A noncompeting pair of human neutralizing antibodies block COVID-19 virus binding to its receptor ACE2. *Science*. **368**, 1274–1278 (2020).
78. H. M. Berman, The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000).
79. C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, T. E. Oliphant, Array programming with NumPy. *Nature*. **585**, 357–362 (2020).
80. J. Reback, W. McKinney, J. V. D. Bossche, Jbrockmendel, T. Augspurger, P. Cloud, Gfyoung, Sinhrks, A. Klein, J. Tratner, C. She, M. Roeschke, Terji Petersen, W. Ayd, A. Hayden, S. Hawkins, J. Schendel, M. Garcia, V. Jancauskas, P. Battiston, Skipper Seabold, Chris-B1, H-Vetinari, S. Hoyer, W. Overmeire, Mortada Mehyar, B. Nouri, T. Kluyver, C. Whelan, K. W. Chen, *pandas-dev/pandas: v0.25.2* (Zenodo, 2019; https://zenodo.org/record/3509135).
81. J. D. Hunter, Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* **9**, 90–95 (2007).
82. M. Waskom, O. Botvinnik, J. Ostblom, S. Lukauskas, P. Hobson, MaozGelbart, D. C. Gemperline, T. Augspurger, Y. Halchenko, J. B. Cole, J. Warmenhoven, J. D. Ruiter, C. Pye, S. Hoyer, J. Vanderplas, S. Villalba, G. Kunter, E. Quintero, P. Bachant, M. Martin, K. Meyer, Corban Swain, A. Miles, T. Brunner, D. O'Kane, T. Yarkoni, M. L. Williams, C. Evans, C. Fitzgerald, Brian, *mwaskom/seaborn: v0.9.1 (January 2020)* (Zenodo, 2020; https://zenodo.org/record/3629445).

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.09.417519; this version posted December 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

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 CSIR-IGIB Delhi, ILS Bhubaneswar and CSIR-CCMB COVID-19 teams for their efforts in aggregating, processing and sharing sample 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 Khushboo Singhal for her inputs and Pragyan Acharya and Abhay Sharma for their critical reviews of the manuscript. **Funding:** This work was funded and supported by Council of Scientific and Industrial Research, India (MLP-2005). **Author contributions:** Conceptualization: MM; Pipeline Design: AKP, TA, BU; Sample Processing, Analysis and Visualization: AKP, MM; Spike Hyper-editing Analysis: SF, LT, AKP; North Indian Cohort Sample Curation: BU, MF; East Indian Cohort Sample Curation and Processing: GPM, AG, SKR; South Indian Cohort Sample Curation and Processing: SB, DTS; Replicate Sample Curation: RCB, AJ, MKD, MI; Writing (Original Draft): MM, AKP, LT, SF; Writing (Review and Editing): LT, SKR, DTS, MF, MM, AKP, TA, GPM. **Competing interests:** None declared. **Data and materials 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.

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.09.417519; this version posted December 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

FIGURES AND TABLES

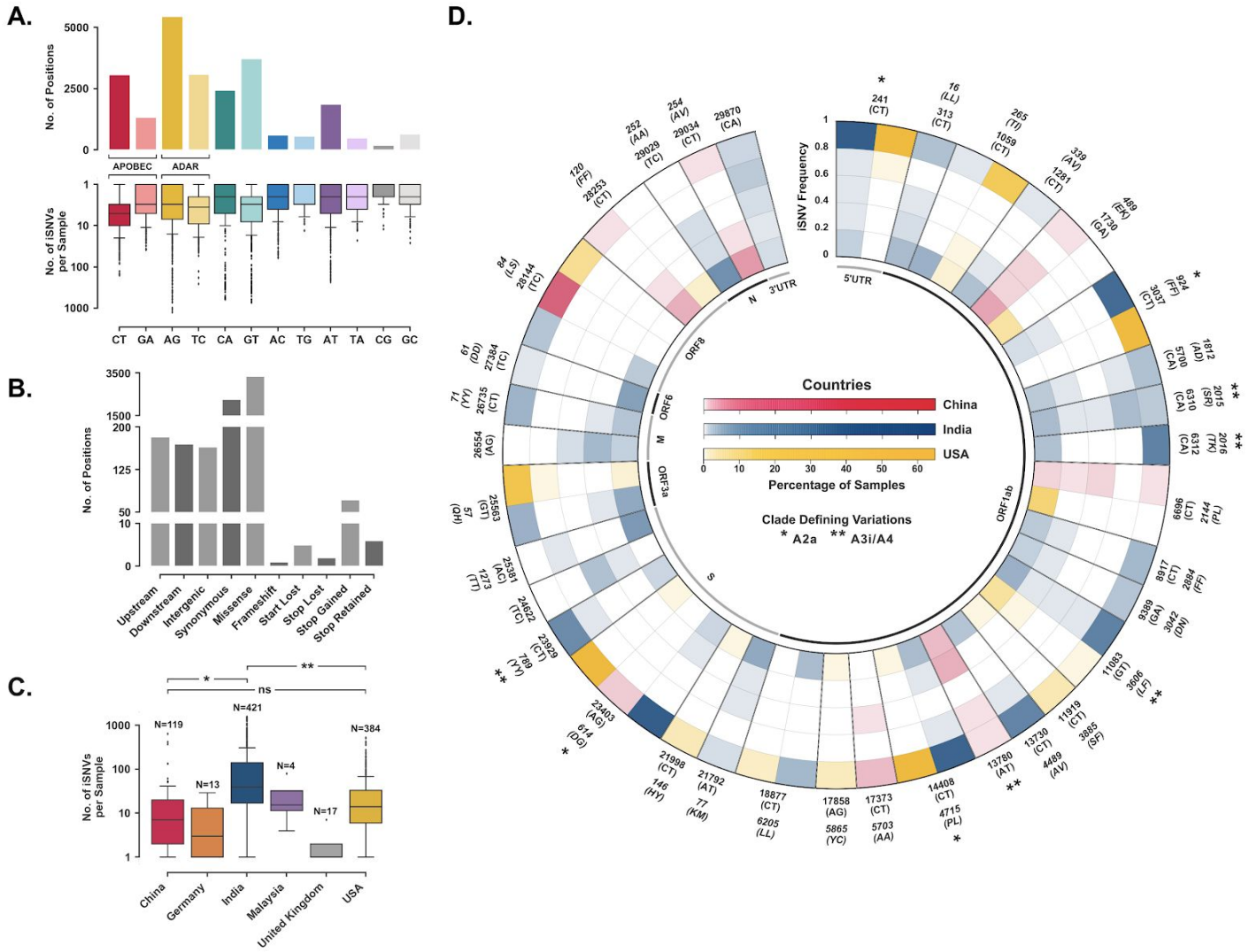


Fig. 1. Spectrum of iSNVs in samples across populations. (**A**) Split plot depicting the distribution of iSNV sites (n=23516) with respect to nucleotide changes in the SARS-CoV-2 genome and across samples (**B**) Potential impact of iSNVs (n=6251) vis-a-vis nature of amino acid sequence change in the SARS-CoV-2 genome (**C**) Distribution of iSNVs in samples of different population cohorts. The significance of the pairwise two-sided t-tests is indicated on top (*$p < 10^{-6}$, **$p < 10^{-16}$, ns = non-significant) (**D**) Radial plot with concentric rings representing the extent of position-wise heterogeneity in samples in the global populations. iSNV frequency distribution in samples is shown for select heteroplasmic positions in frequency bins of 0.2. The colour gradient in each cell represents the percentage of samples. The outer labels denote the nucleotide change and amino acid change (italicized) with the position of change. Variations that are represented in the A2a and A3i/A4 clades have been marked (*) and (**) respectively.

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.09.417519; this version posted December 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

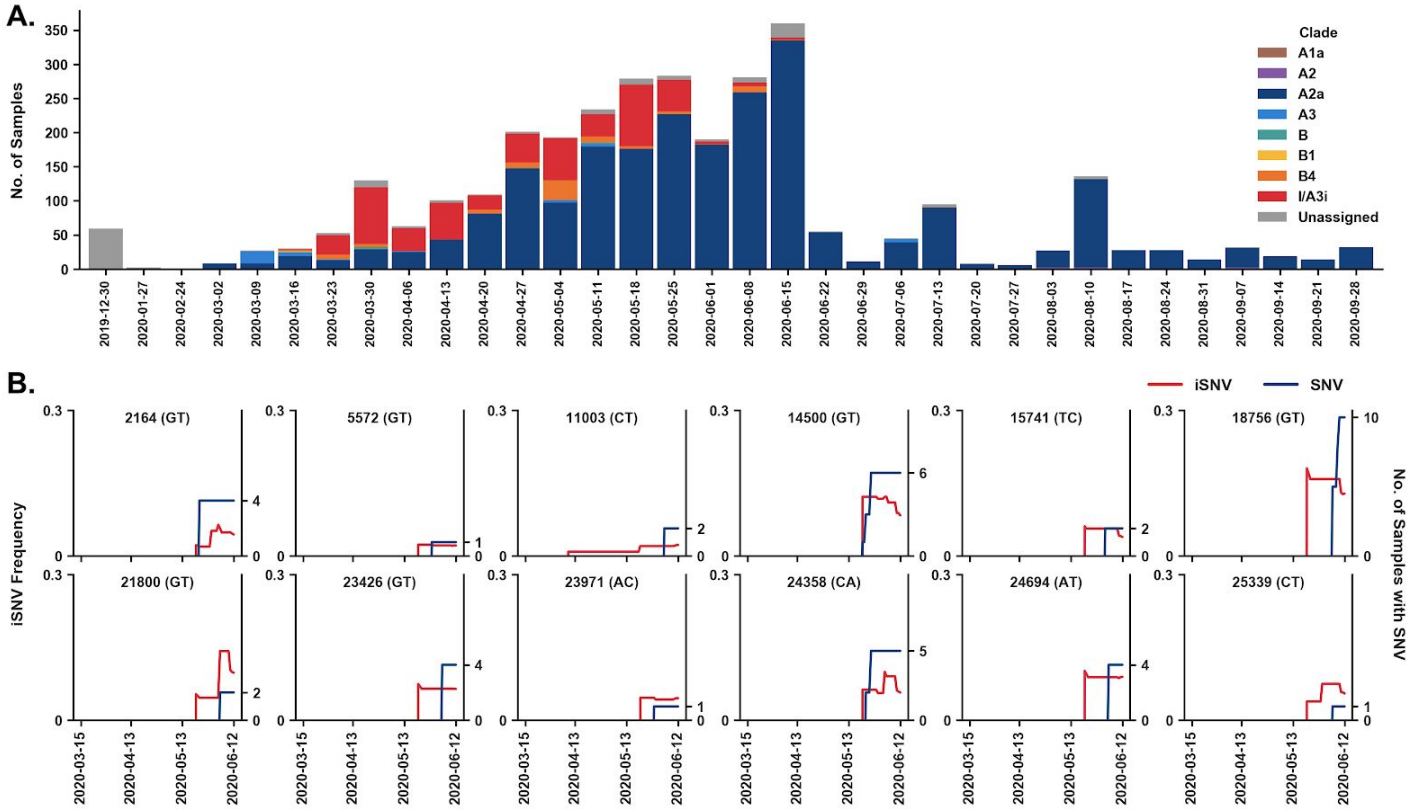


Fig. 2. Spatio-temporal dynamics of iSNVs. (**A**) Temporal (weekly) distribution of clades in SARS-CoV-2 samples submitted pan-India. The proportions of A2a and A3i/A4 clades are represented in blue and red colours respectively (**B**) Temporal trends of iSNV frequency and incidence of SNV for select heteroplasmic positions in the East Indian cohort. The left y-axis denotes the average frequency of iSNVs and the right y-axis denotes the number of samples with SNVs. Red and blue lines illustrate the iSNV frequency and the number of samples with SNVs at the site in the cohort respectively.

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.09.417519; this version posted December 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

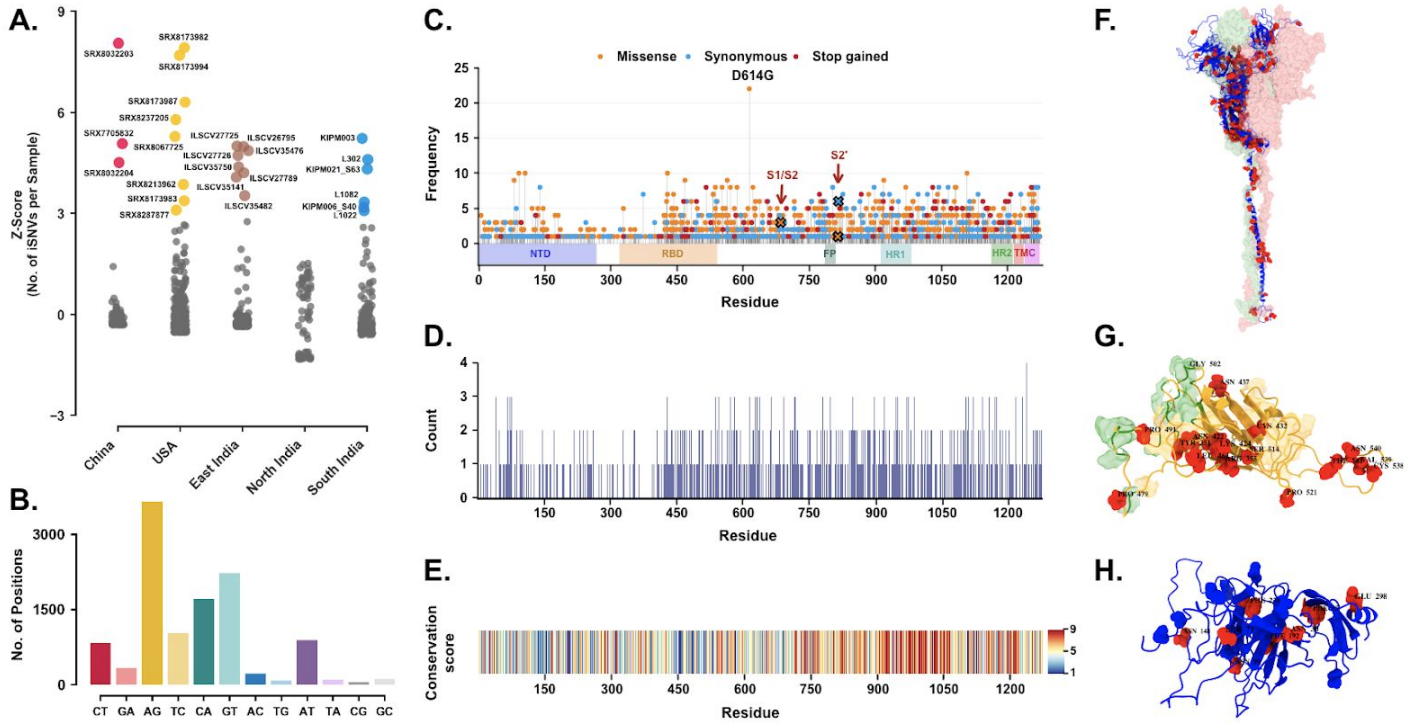


Fig. 3. Functional impact of hyper-editing on Spike protein. (**A**) Dot plot illustrating cohort-wise outliers in Z-score values based on the distribution of number of iSNVs per sample (**B**) Distribution of iSNV sites (n=11392) with respect to nucleotide changes in the SARS-CoV-2 genome in hyper-edited 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.

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.09.417519; this version posted December 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

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**	**463**	**384**
India	**489**	**486**	**421**
East India	246	246	197
North India	77	74	62
South India	166	166	162
TOTAL	**1347**	**1160**	**958**

Table 1. Cohort-wise share of samples. 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).

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.09.417519; this version posted December 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

SUPPLEMENTARY MATERIALS

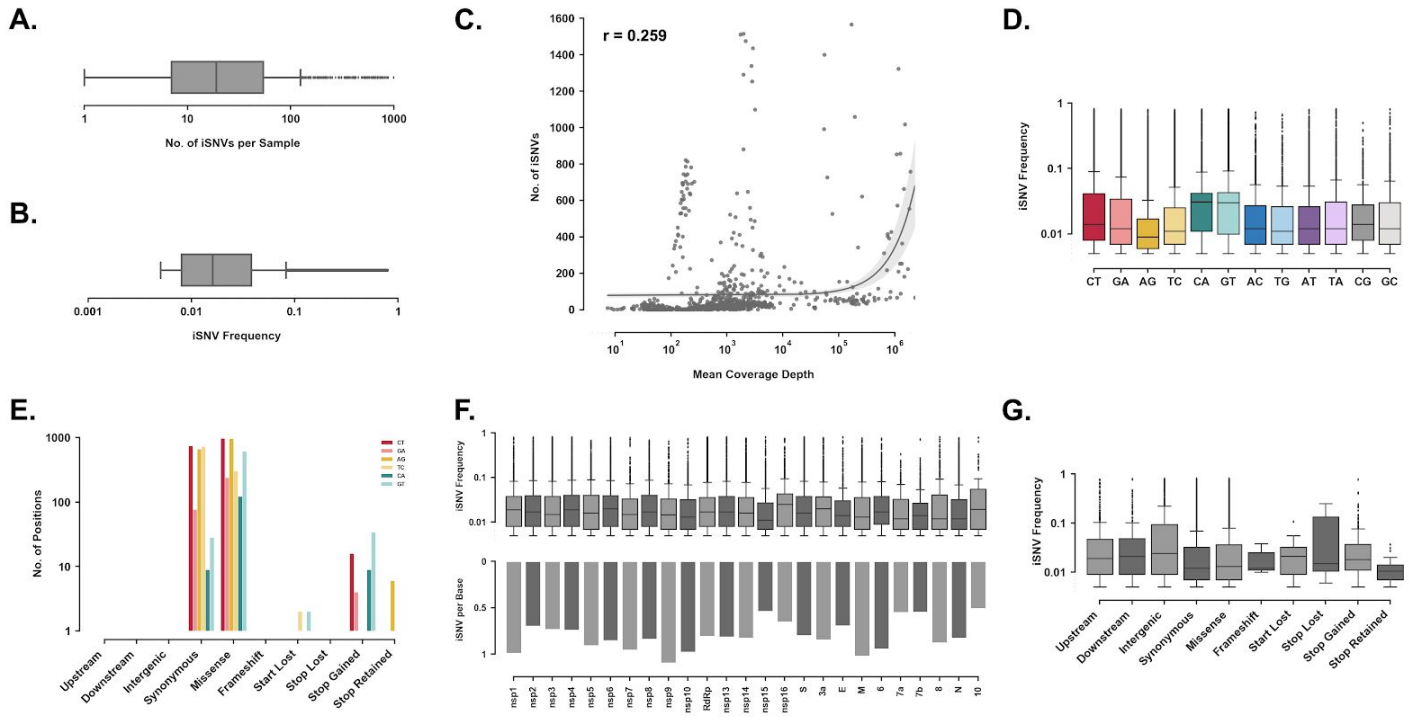


Fig. S1. Distribution of iSNVs in samples. (**A**) Distribution of the number of iSNVs per sample (n=958) (**B**) Distribution of the frequencies of all captured iSNVs (n=86595) (**C**) Scatter plot to assess the correlation between number of iSNVs and mean coverage depth in samples (n=958) (**D**) Distribution of iSNV frequencies with respect to the nucleotide change (**E**) Distribution of nucleotide change mediated by iSNVs with respect to nature of amino acid sequence change (**F**) Split plot depicting the distribution of iSNV frequencies and number of iSNVs per base in each protein-coding domain (**G**) Distribution of iSNV frequencies (n=20547) with respect to nature of amino acid sequence change.

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.09.417519; this version posted December 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

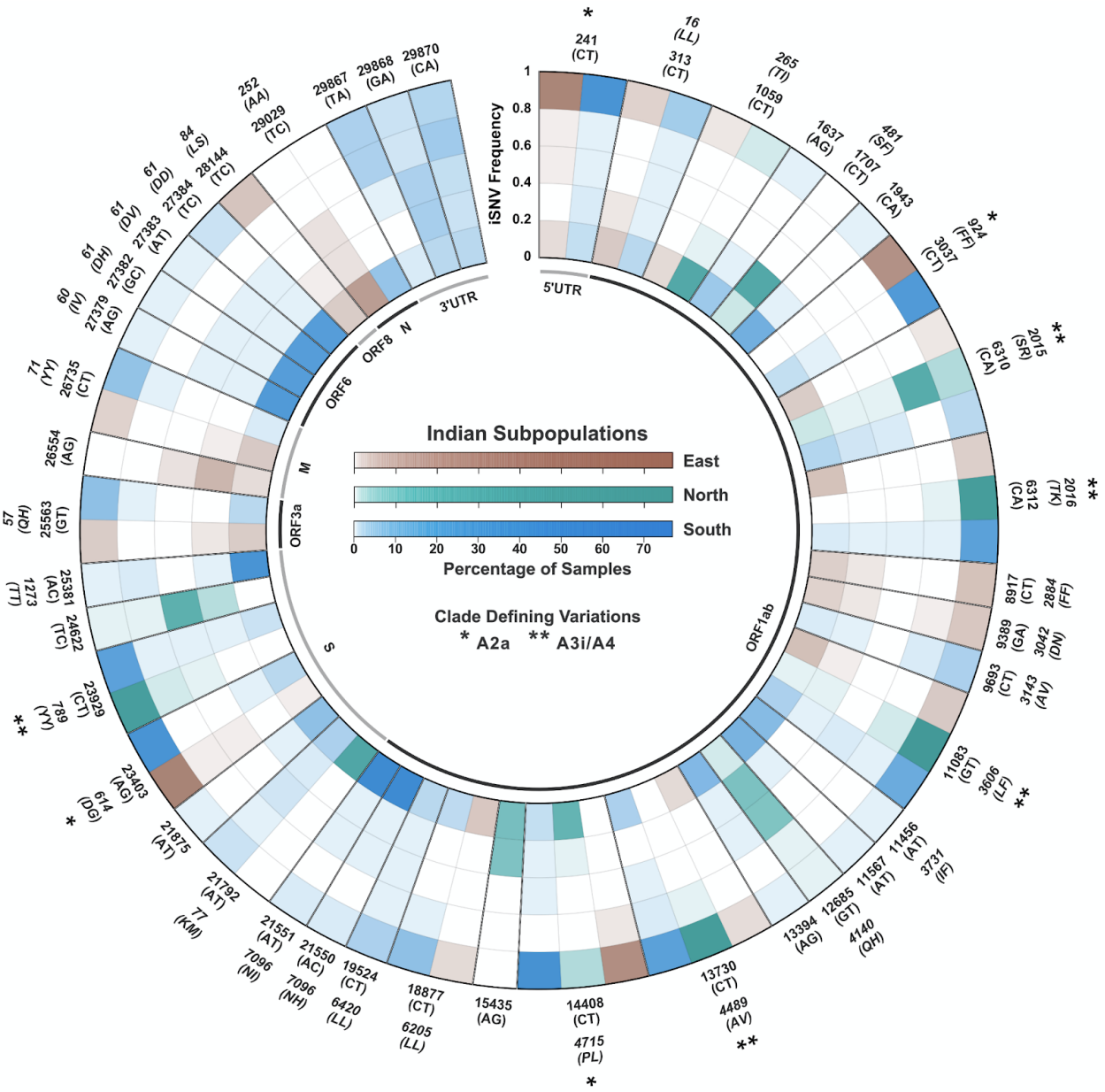


Fig. S2. Spectrum of iSNVs in Indian subpopulations. Radial plot with concentric rings representing the extent of position-wise heterogeneity in samples in the Indian subpopulations. iSNV frequency distribution in samples is shown for select heteroplasmic positions in frequency bins of 0.2. The colour gradient in each cell represents the percentage of samples. The outer labels denote the nucleotide change and amino acid change (italicized) with the position of change. Variations that are represented in the A2a and A3i/A4 clades have been marked (*) and (**) respectively.

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.09.417519; this version posted December 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

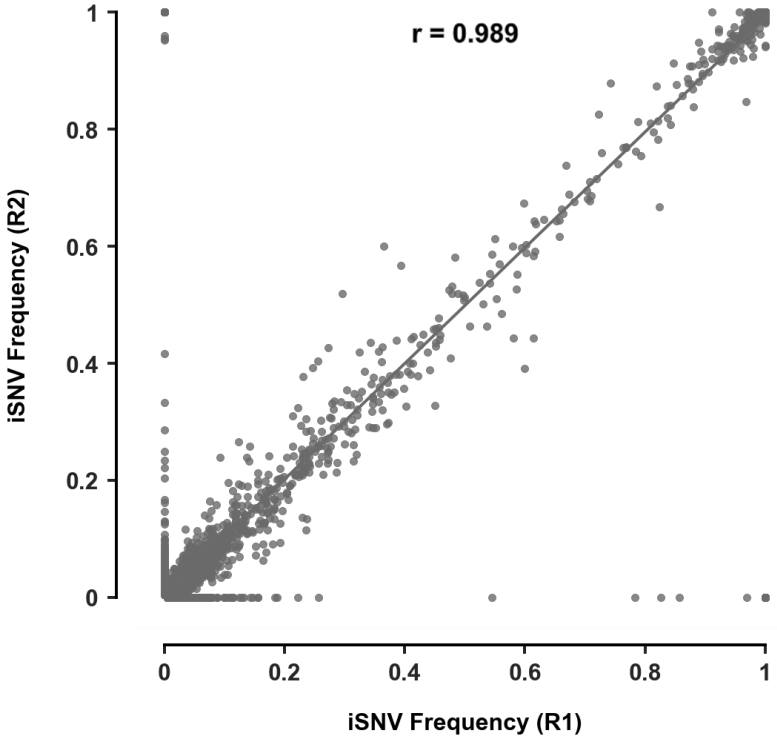


Fig. S3. iSNV concordance between replicates. Correlation plot illustrating the concordance in iSNV frequencies between replicates (n=500).

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.09.417519; this version posted December 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

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

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 column indicates the property of the modified residue. The number of iSNV associated variants falling in each of the pair categories are listed.

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.09.417519; this version posted December 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Supplementary Materials Captions

Data File S1: Data summary of pipeline analyses and metadata for each sample

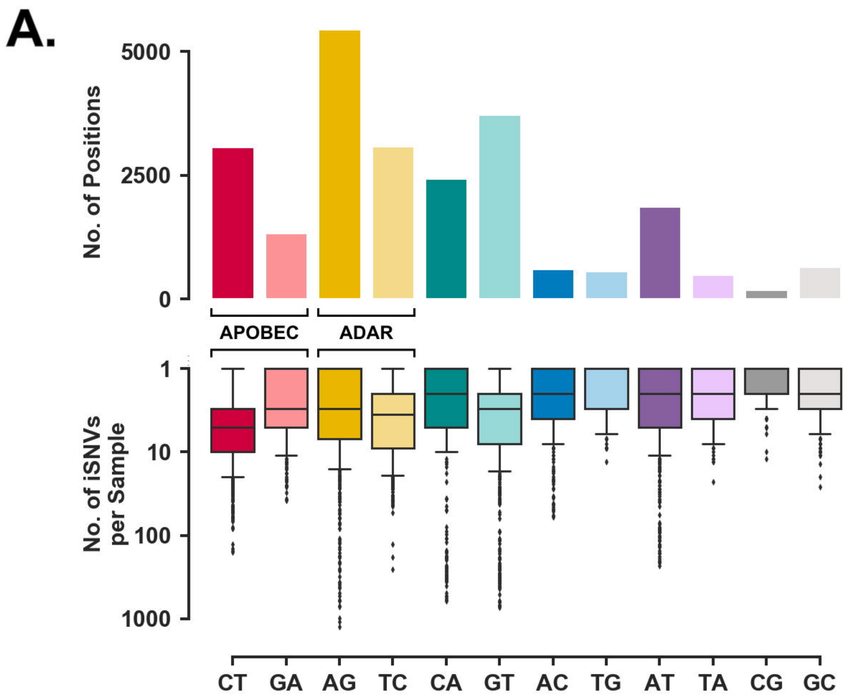
Data File S2a: Sample-wise frequencies of all recorded iSNVs in global populations

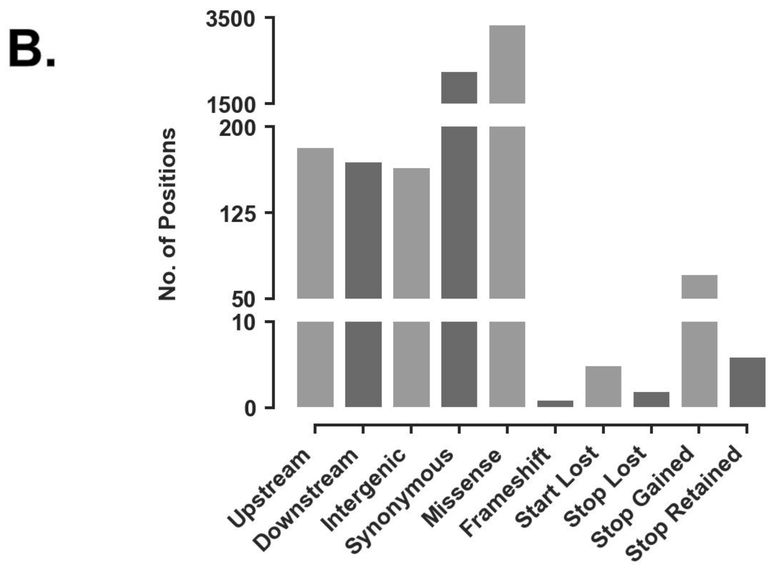
Data File S2b: Sample-wise frequencies of all recorded iSNVs in Indian subpopulations

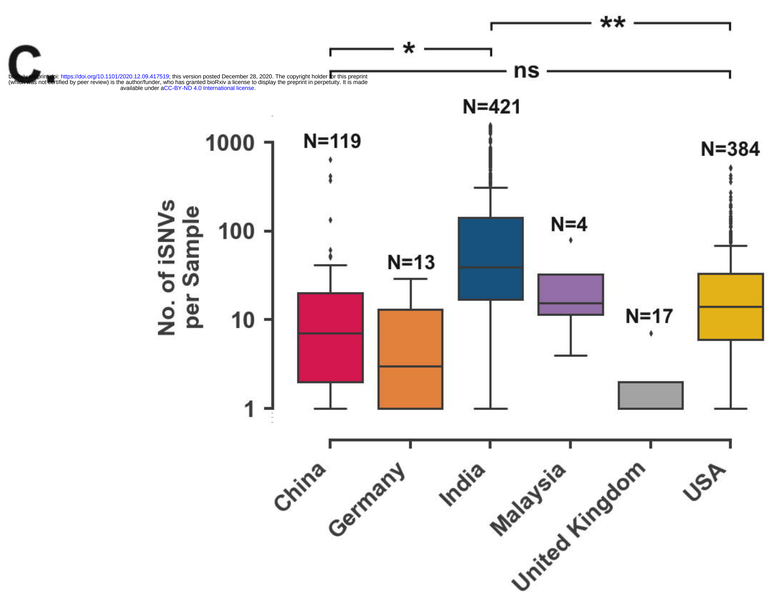
Data File S3: Summary of nucleotide change, codon change, amino acid sequence change, and nature of change of all recorded iSNV sites

Data File S4: iSNVs in samples of three reinfection studies mapped to iSNV sites

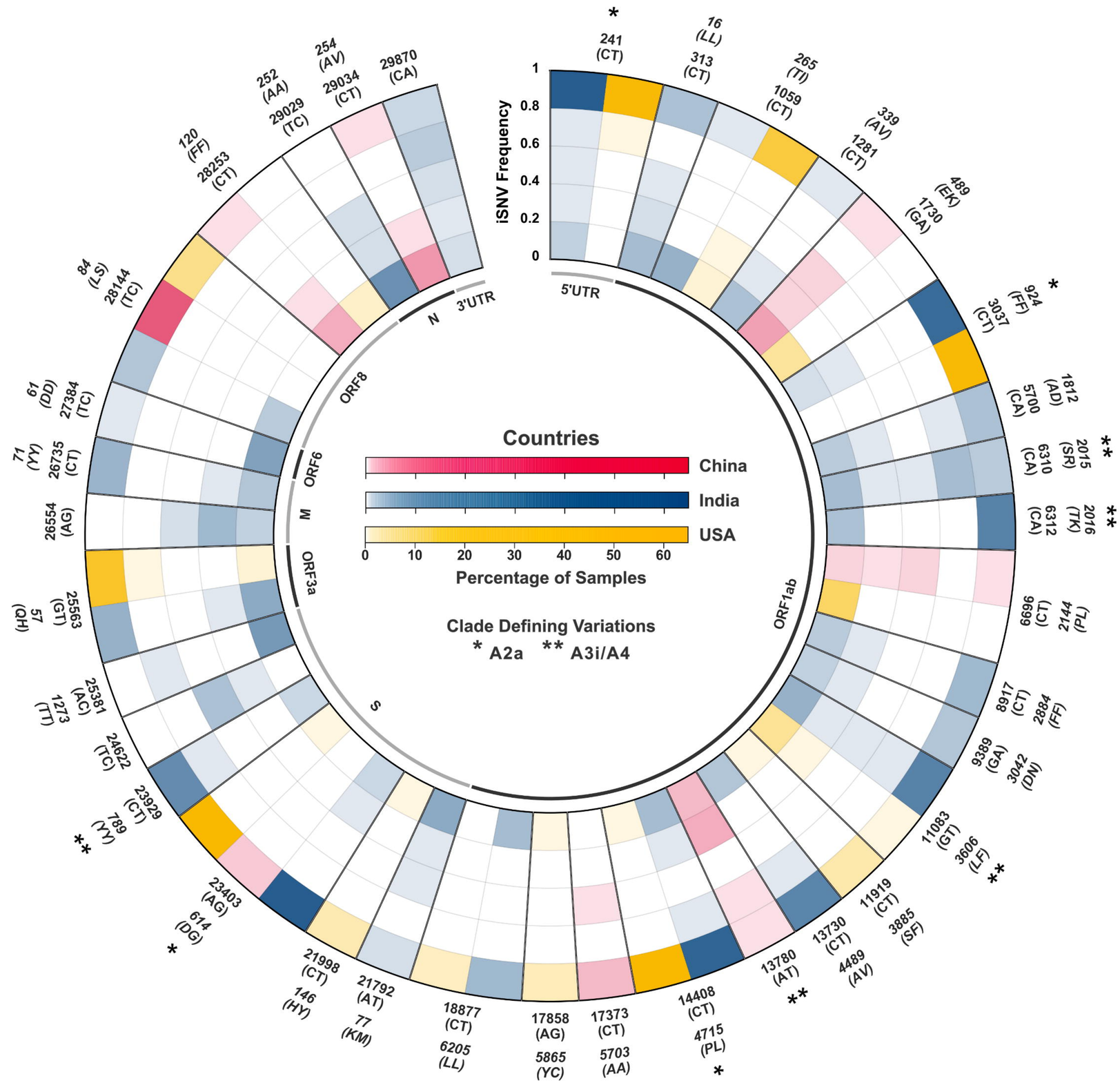
Data File S5: Coordinates of homopolymeric regions in the SARS-CoV-2 genome

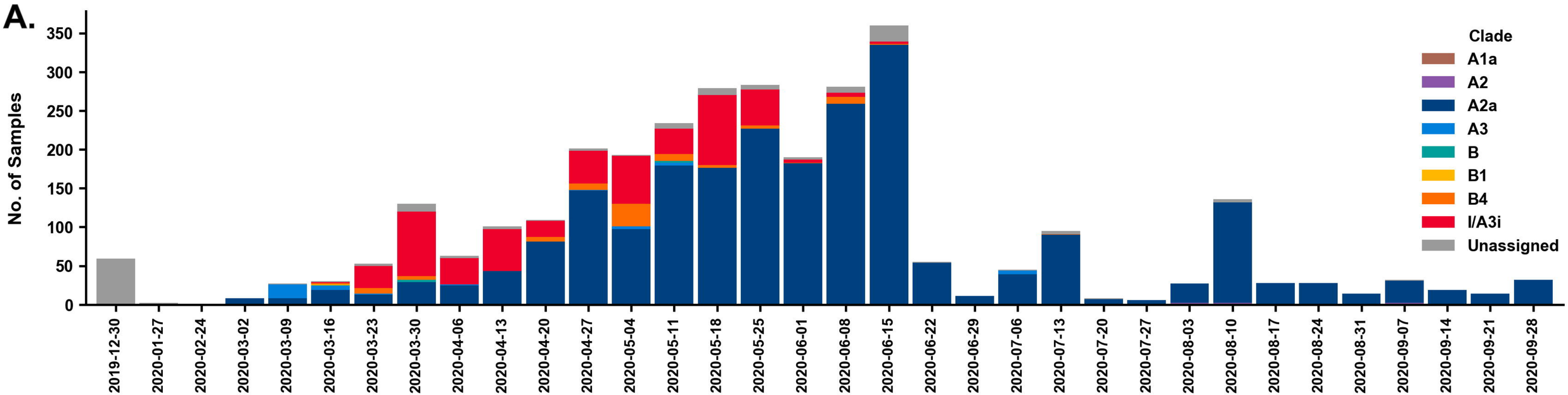


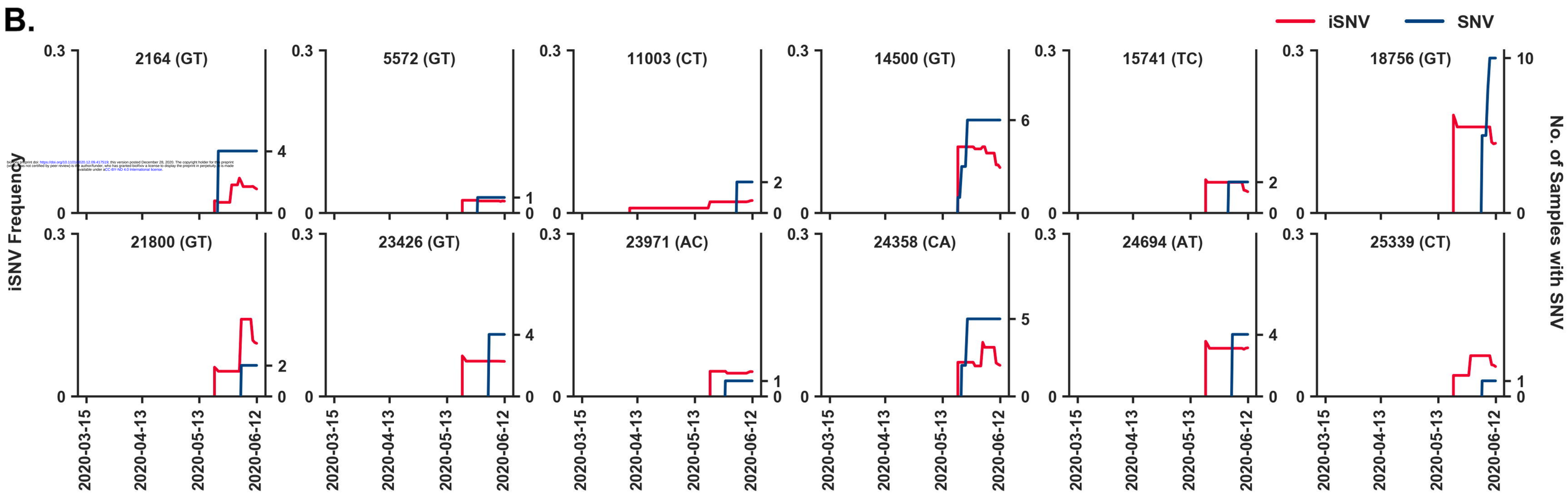




https://doi.org/10.1101/2020.12.09.417519; this version posted December 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

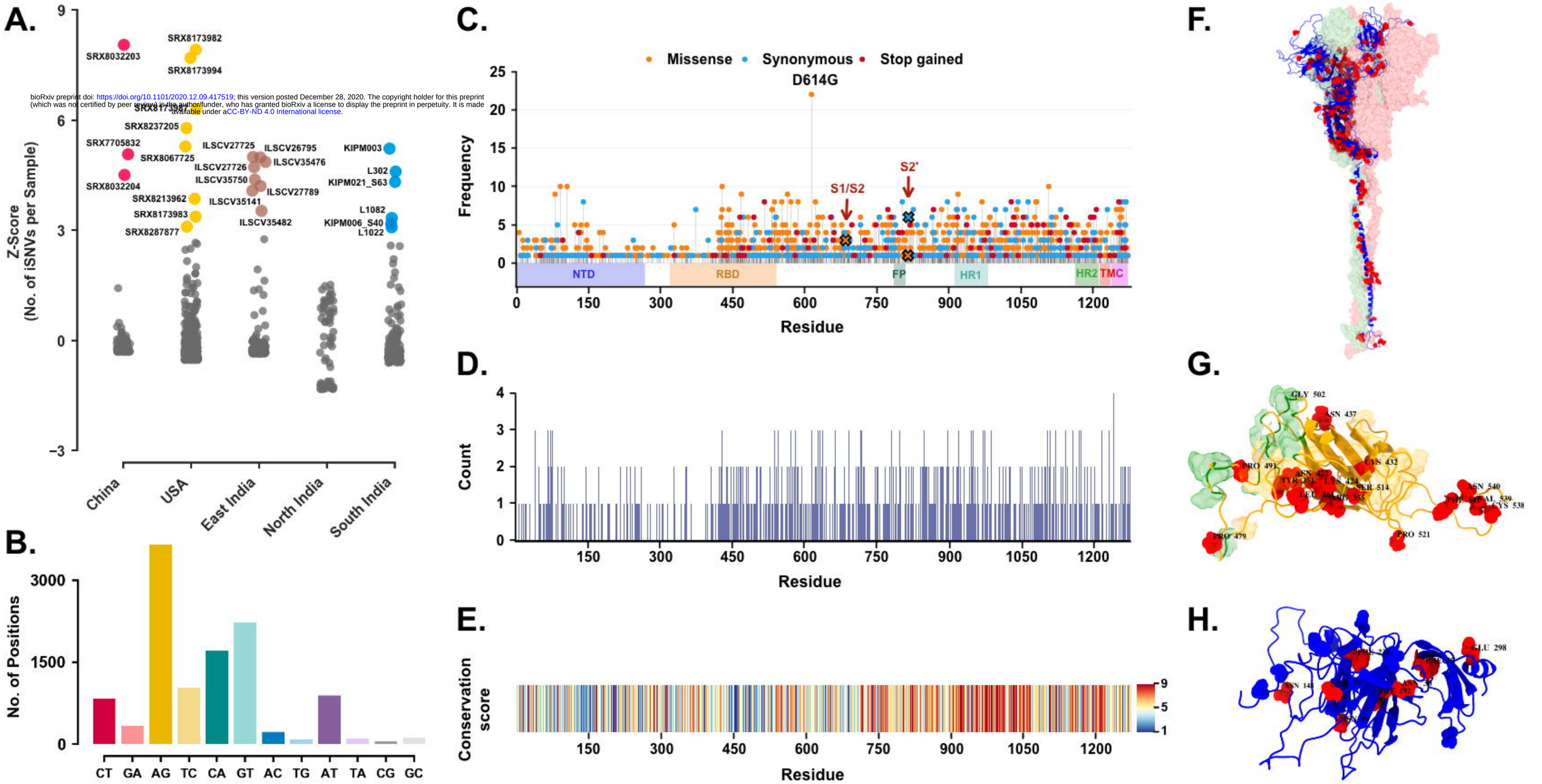


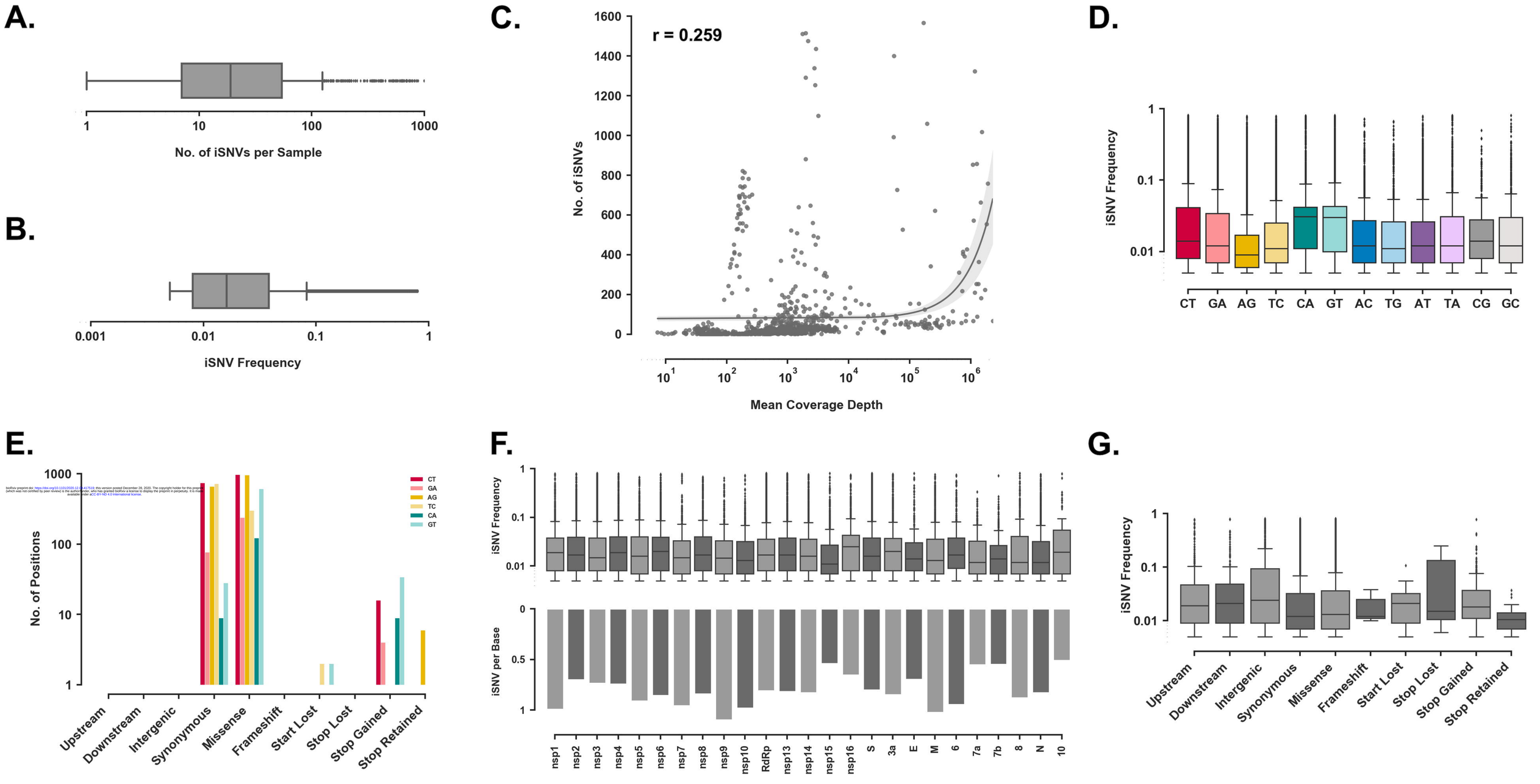




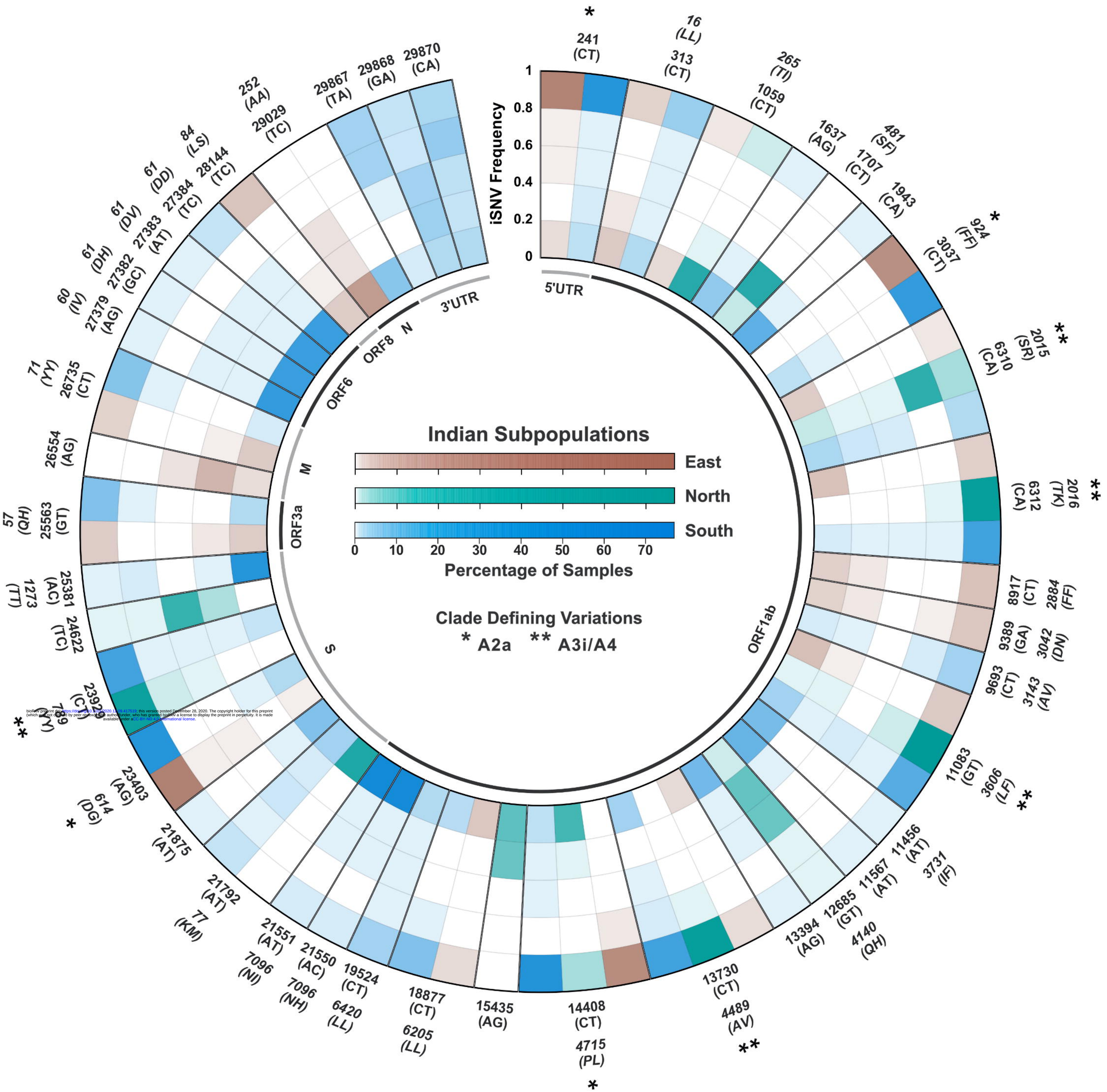
bioRxiv preprint doi: https://doi.org/10.1101/2020.12.09.417519; this version posted December 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.09.417519; this version posted December 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

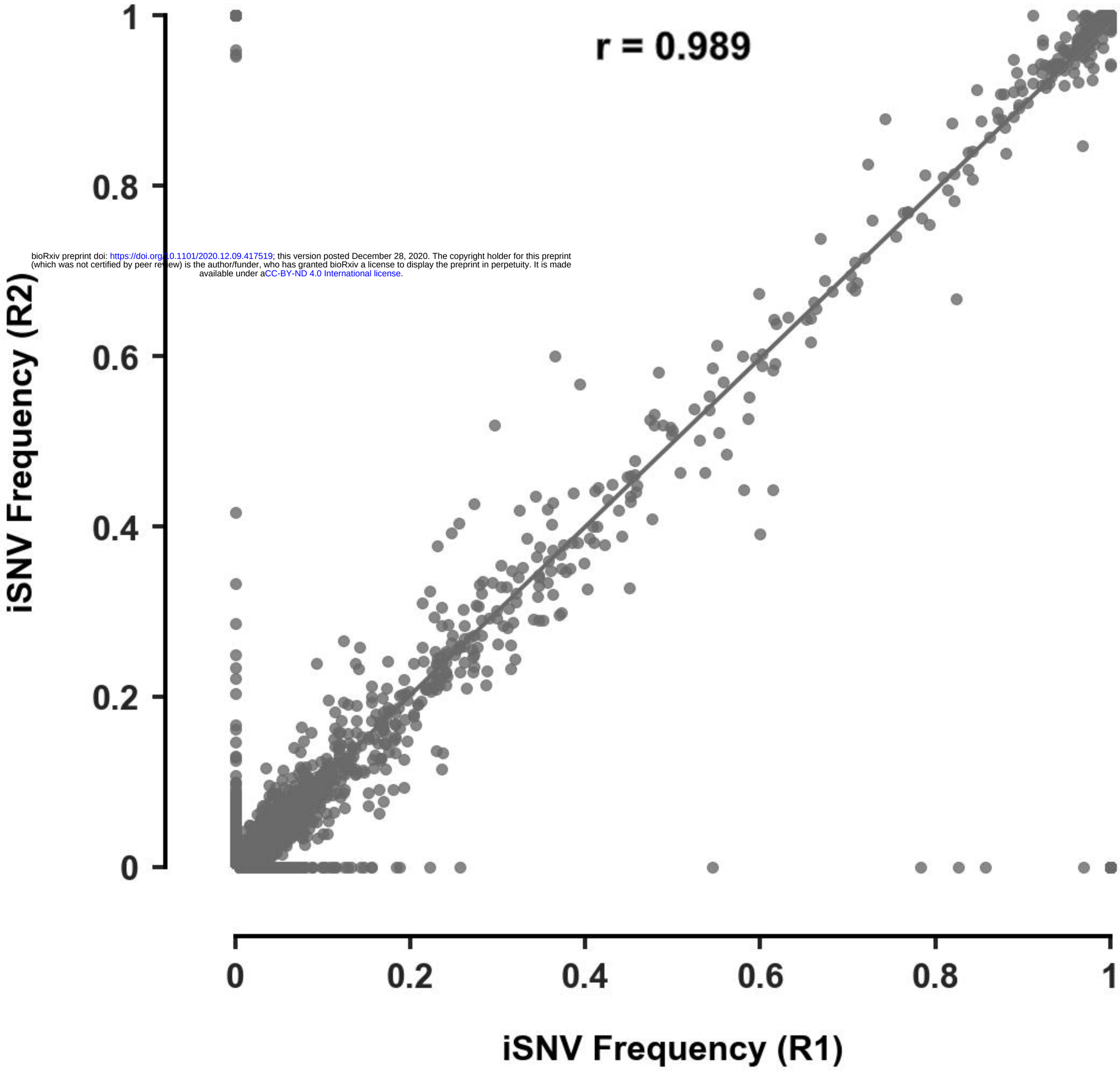




bioRxiv preprint doi: https://doi.org/10.1101/2020.12.28.417519; this version posted December 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.



bioRxiv preprint doi: https://doi.org/10.1101/2020.12.29.417519; this version posted December 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.



bioRxiv preprint doi: https://doi.org/10.1101/2020.12.09.417519; this version posted December 28, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.