Underexplored intrahost diversity of SARS-CoV-2

Nicolae Sapoval1, Medhat Mahmoud2, Michael D. Jochum3,

Yunxi Liu1, R. A. Leo Elworth1, Qi Wang4, Dreycey Albin4, Huw Ogilvie1, Michael D. Lee5,6, Sonia Villapol7, Kyle Hernandez8, Irina Maljkovic Berry9,

Jonathan Foox10, Afshin Beheshti11, Krista Ternus12, Kjersti M. Aagaard3,

David Posada13,14,15, Christopher E. Mason10, Fritz Sedlazeck2*,*†,Todd J. Treangen1*,*†∗

1Department of Computer Science, Rice University, Houston, TX, USA.

2Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX, USA.

3Baylor College of Medicine and Texas Children’s Hospital, Houston, TX, USA.

4Systems, Synthetic, and Physical Biology (SSPB) Graduate Program, Houston, TX, USA.

5Exobiology Branch, NASA Ames Research Center, Mountain View, CA, USA.

6Blue Marble Space Institute of Science, Seattle, WA, USA.

7Houston Methodist Research Institute, Houston, TX, USA.

8Department of Medicine, The University of Chicago, Chicago, IL.

9Walter Reed Army Institute of Research, Silver Spring, MD, USA.

10Department of Physiology and Biophysics, Weill Cornell Medicine, New York, New York, USA.

11KBR, Space Biosciences Division, NASA Ames Research Center, Moffett Field, CA, USA.

12Signature Science, LLC, 8329 North Mopac Expressway, Austin TX 78759.

13Biomedical Research Center (CINBIO), University of Vigo, 36310 Vigo, Spain.

14Department of Biochemistry, Genetics and Immunology

School of Biology, University of Vigo, Vigo, Spain.

15Galicia Sur Health Research Institute, 36310 Vigo, Spain.

∗To whom correspondence should be addressed; E-mail: [treangen@rice.edu.](mailto:treangen@rice.edu)

†These authors share senior authorship

**The COVID-19 pandemic has sparked an urgent need to uncover the underlying biology of this devastating coronavirus. Though RNA viruses mutate rapidly, there are a relatively small number of single nucleotide polymorphisms (SNPs) that differentiate the main SARS-CoV-2 clades that have spread throughout the world. In this study, we investigated over 7,000 SARS-CoV-2 datasets to unveil its intrahost and interhost diversity. We reveal a complex intrahost variant landscape, likely due to a combination of host mediated processes such as mRNA editing, and viral mutation. Some of these variants might reduce the power of qRT-PCR tests.**

**We also identified intrahost single nucleotide variants (iSNV) shared between COVID-19 positive patients, which might be indicative of a wide transmission bottleneck in SARS-CoV-2 In addition, we were able to identify population-wide structural variantin key genes like ORF8 and Spike. Altogether, our findings providean enhanced understanding of the SARS-CoV-2 transcriptome, inform the design of detection tests, and highlight the potential utility of iSNVs for tracking the transmission of SARS-CoV-2.**

# Introduction

Coronavirus (CoV) genomes are the largest among single strand RNA (ssRNA) viruses, ranging from 26 to 32 Kbp. While ssRNA viruses typically display very high mutation rates, coronaviruses encode an RNA polymerase with 3’-to-5’ proofreading activity that allows them to replicate their genome with high-fidelity, lowering their mutation rate [(*1*–*4*)](https://paperpile.com/c/U1nDEi/iSDa+b3GU+Gd7N+ZPaq). On March 11, 2020, the WHO determined that an outbreak of a novel coronavirus SARS-CoV-2 that began in Wuhan, China in December 2019 had reached pandemic status. Initial consensus-level GISIAD genomic data indicated that the SARS-CoV-2 mutational rate was similar to other CoVs, around 10−6*/*−7 mutations per site per replication cycle [(*5*)](https://paperpile.com/c/U1nDEi/aVN2). However, this estimate is based on consensus sequences and therefore only considers mutations in the dominant virus lineage within each infected host [(*6*)](https://paperpile.com/c/U1nDEi/ns6k). In order to properly assess the genomic diversity of any RNA virus, and specifically SARS-CoV-2, it is necessary to also consider the intrahost polymorphisms [(*6*–*9*)](https://paperpile.com/c/U1nDEi/ns6k+UxGv+Me6J+YDZg). Recent studies have claimed that host-dependent RNA editing might be a key factor for understanding the mutational landscape of SARS-CoV-2 within hosts [(*5*, *10*)](https://paperpile.com/c/U1nDEi/MRIn+aVN2). However, these studies were based on a limited number of samples (*<*20) that were sequenced on the Illumina sequencing platform. In order to explore both the intrahost and interhost mutational landscape of SARS-CoV-2, we leveraged a dataset consisting of 6,928 consensus genomes from the Global Initiative on Sharing All Influenza Data (GISAID) (*11*), 11 sequencing samples from the Baylor College of Medicine, 140 sequencing samples from the Weill Cornell College of Medicine, and one sequencing sample from the University of Melbourne (*12*).

Understanding the intrahost genomic diversity of SARS-CoV-2 is also important for different applications.Most SARS-CoV-2 detection tests rely on oligonucleotide probes and primers that must be both specific and sensitive to SARS-CoV-2. In this setting, sensitivity determines how well it can capture the diversity of all SARS-CoV-2 variants. On the other hand, specificity determines the ability to discern target sequences from an array of non-targets, which can be closely related pathogens (e.g. OC43 and 229E strains of human coronavirus that cause cold-like symptoms) or environmental and extraction kit contaminants. Moreover, recent studies on Ebolavirus and flu viruses (9, 10) highlight the importance of intrahost variation for studying viral population dynamics and transmission scenarios.

In this study, we investigate the intrahost diversity of SARS-CoV-2 by conducting a broad evaluation of (i) intrahost single nucleotide variants (iSNV), (ii) consensus-level single nucleotide polymorphisms (SNPs), and (iii) structural variants, across assembled genomes, amplicon, and metatranscriptomic datasets totaling over 7,000 samples. We performed extensive quality control checks prior to our analysis, and investigated some common potential biases in the variant data. Our analyses have implications for viral detection methods reliant on genomic data, such as qRT-PCR testing (*13*), as increased diversity can adversely affect their sensitivity. Furthermore, intrahost single nucleotide variants (iSNVs) have the potential to provide valuable insights into the transmission properties of SARS-CoV-2.

# Results

We analyzed three SARS-CoV-2 genomic datasets: GISAID public consensus sequences, sequencing reads for 11 samples collected by the Baylor College of Medicine in Houston, and sequencing reads for 140 samples collected by Weill Cornell University in New York City (NYC). The variants in GISAID represent interhost single nucleotide polymorphisms (SNPs), while the variants analyzed in the Houston and NYC datasets include both interhost SNPs and intra-host single nucleotide variants (iSNVs). The inferred phylogenetic tree of GISAID genomes with clade-defining (*14*) SNPs is shown in Figure [1B](#_gjdgxs). We note that clade-defining SNPs correspond to the geographic distribution of the sequences, with clades G and S predominantly covering North American genomes andclade V covering a portion of Asian and European genomes. We also observe that some of the clade-defining SNPs occur spuriously outside of the main phylogenetic clades. We also identified several interesting patterns of SNP and iSNV mutational patterns within the ORFs of SARS-CoV-2. Of note, SARS-CoV-2 encodes three tandem macrodomains within non-structural protein 3 (NSP3). NPS3 is essential for SARS-CoV-2 replication and represents a promising target for the development of antiviral drugs (*15*). The NSP3 protein is also one of the most diverged regions of SARS-CoV-2 compared to SARS-CoV-1 and MERS-CoV. We will now dive deep into three main results: (i) intrahost structural variant (SV) landscape, (ii) intrahost single nucleotide variant (iSNV) landscape, and (iii) exploratory analyses of shared SNPs and iSNVs within and across patients in NYC.

## Intrahost Structural Variant (SV) Landscape

We identifiedn 3,311 structural variants (SVs) across 170 sequencing samples, with the majority being inversions (1,504) and tandem duplications (1,157), followed by deletions (625) and a few insertions (25) (Figure [1](#_gjdgxs)). Overall, since we are identifying SVsbased on RNA-Seq data, the majority of these SVs are likely to be highlighting variability in the SARS-CoV-2 transcriptome (*16*), which is influenced by fusion, deletions, frame-shifts, and recombination. In brief, SARS-CoV-2 contains a common 69-bp 5’ leader sequence fused to the body sequence from the 3’ end of the genome (*17*). Then, leader-to-body fusion occurs during negative-strand synthesis at short motifs called transcription-regulating sequences (TRS), which are conserved 7 bp sequences that are adjacent to the ORFs.

We observed 98 start and 63 end breakpoints overlapping with the annotated transcription regulating sequences (TRS) (dark red Figure [1A).](#_gjdgxs) This overlap is significantly different from 1000 times randomly chosen TRS (Kolmogorov–Smirnov test: p-value=4*.*95−5, D=0.25 ). Subsequently, we focus on smaller SVs (*<*1kbp) that more likely indicate true underlying SV rather than transcription signals. e identified 247 deletions and 23 insertions across all 170 SARS-CoV-2 genomes. The imbalance of insertions and deletions are likely due to the low ability to detect insertions using short reads(*18*). Figure [1A](#_gjdgxs) shows the allele frequency of these SVs across all samples. We observed 8 deletions shared among 34 or more samples (allele frequency (AF): *>*20%). a 14bp at 509bp (NSP1) (AF: 30.59%), a 9bp at 685bp (NSP1) (AF: 23.53%), a 24bp at 4532 (NSP3) (AF:25.29%) a 39bp at 21740bp (spike protein) (AF: 37.65%), a 22bp at 23558bp (spike protein) (MAF: 31.76%), a 15bp at 24014bp (spike protein) (AF: 21.18%), a 41bp at 26779bp (M protein) (AF: 34.12%) and a 14bp at 29067 (N protein) (AF: 20%) .

Next, we investigated where these SVs are mainly located with respect to the annotated regions. we identified a large cluster of SVs in NSP11 (3 del) and NSP12 (3 del) when taking the size of the annotated regions into account (see methods). In addition, it is interesting to see that a higher number of SVs are also clustering in E protein (5 del), NSP7 (5 del and 1 ins), NSP9 (7 del and 1 ins), ORF6 (6 del) and ORF7b (3 del).

We further compared our SV call set with previously reported single deletions reported by various groups. Davidson et al (*19*) reported a 24bp deletion in the subgenomic mRNA encoding the spike (S) glycoprotein that played a role in removing a proposed furin cleavage site from the S glycoprotein. We were able to identify this deletion (position: 25234bp), but only in 3 of our samples. However, in total we discovered six deletions shared among samples within the Spike protein. Three of them showed above with AF*>* 20% and the remaining at: 21984bp (9bp, AF:19.41%), 22824bp (78bp, AF: 11.76%) and at 24125bp (15bp, AF: 8.24%). We further identified five deletions, one (at 28245bp) of which was present in 10 samples (AF: 6%) in ORF8 which have been highlighted to be an important gene when studying potential viral adaptation to humans (*20*).

## Intrahost Single Nucleotide Variant (iSNV) Landscape

**We considered intrahost single nucleotide variants (iSNVs) to be those with an AF between 2% and 50% in a sample. Above 50%, all single nucleotide variants were considered to be consensus-level single nucleotide polymorphisms (SNPs).**

Figure [2A](#_30j0zll) shows the iSNV allele frequency distribution, with the peak occurring in the 2% to 5% range of the distribution.The predominant iSNVs observed are T*>*C and C*>*T (Figure [2B).](#_30j0zll) We also note that A*>*G, G*>*A, and G*>*T iSNVs are common. One potential explanation for the high fraction of G*>*T iSNVs observed can be the error introduced by Illumina sequencing, since the G*>*T transversion is the most common Illumina call error (*2*). When the distribution of iSNVs is mapped onto the SARS-CoV-2 genome, we observe that C*>*T is the dominant SNP in 10 out of 16 genes (Figure [2D).](#_30j0zll) NSP6 and NSP10 stand out as having larger fractions of T*>*C iSNVs, and NSP7 has a large fraction of A*>*C iSNVs (Figure [2D).](#_30j0zll) Additionally NSP6 and ORF3a have a high fraction of G*>*T SNPs, and ORF8 and M genes have a high fraction of T*>*C SNPs.

The SNPs extracted from the GISAID data overall match the distribution of the iSNVs. Unlike the read datasets from Houston and NYC, GISAID data is a mix of long and short read sequencing data, suggesting a potentially reduced bias in SNP calls. We note that the mutational spectra for SNPsmatches the one observed for iSNVs, namely A*>*G, G*>*A, T*>*C and G*>*T are most common (Figure [2B).](#_30j0zll) However, one striking difference is the relatively lower percentage of C*>*T changes in iSNVs from the NYC dataset (20%) compared to 40% C*>*T iSNVs for Houston samples and over 50% C*>*T in Houston and NYC SNPs. The GISAID SNP C*>*T changes are nearly identical to Houston iSNV C*>*T changes, clearly distinguishing GISAID SNPs and Houston iSNVs from Houston and NYC SNPs.

Since GISAID consensus *genomes* are based on different sequencing and assembly technologies, we note that the abundance of G*>*T mutations should not necessarily be attributed to an Illumina sequencing error, and can be a feature of the SARS-CoV-2 mutational landscape. We also note that the distribution of SNPs by type across the genes of SARS-CoV-2 closely matches the iSNV profile (Figure [2D).](#_30j0zll) he iSNV profile for NYC data is significantly different from both NYC SNPs (Kolmogorov-Smirnov (KS) test: p-value ∼ 10−100) and GISAID SNPs (KS test: p-value ∼ 10−200). NYC SNP profile has a low KS test value (0.068) when compared to GISAID SNP type distribution indicating higher similarity, but the p-value is low enough (0.0016) to consider the observed difference between the distributions significant. When compared to SARS and MERS, SARS-CoV-2 has a larger proportion of G*>*T iSNVs (Figure [2C).](#_30j0zll) The other four major iSNV types (C*>*T, T*>*C, A*>*G, and G*>*A) are well represented in all three viruses. We also note that SARS data does not have any A*>*T nor A*>*C iSNVs.

We also looked into whether clade-defining SNPs identified in a previous study (14) appear as iSNVs in our datasets. We found that a G and S clade-defining SNP appears as an iSNV position 13542 in the NSP12 gene. There are two synonymous iSNVs at this position, the more common one is a T*>*G change (seen in NYC and Houston), and a less common one is a T*>*A change. The T*>*G iSNV occurs in both NYC and Houston samples. Subsequently, we analyzed the impact of the observed SNPS and iSNVs In GISAID data we observe 1191 (36.45%) synonymous, 2021 (61.86%) missense, 40 (1.22%) stop gained and 15 (0.46%) upstream variants. In NYC data we observe 782 (31.68%) synonymous, 1549 (62.76%) missense, 73 (2.96%) stop gained, 6 (0.24%) start lost, 25 (1.01%) upstream and 31 (1.26%) downstream variants. In Houston data we observe 43 (31.16%) synonymous, 86 (62.31%) missense, 5 (3.62%) stop gained, 2 (1.45%) upstream and 2 (1.45%) downstream variants. Overall about two thirds of all observed variants are missense and about a third are synonymous.

We estimated the genetic complexity (*Sn*) (*21*)

and genetic diversity (*π*) of of SARS-CoV-2, SARS-CoV-1 and MERS (Figure [4A,B).](#_3znysh7) For both diversity and complexity all three viruses show distinct distributions of (KS test: p-value < 10−8). with a higher variance in SARS-CoV-2 We also compared the ratios of non-synonymous and synonymous diversities (*πN /πS*) for SARS-CoV-2, SARS-CoV-1 and MERS data (Figure [4C).](#_3znysh7) We note that SARSCoV-2 and SARS-CoV-1 show a trend towards purifying selection, while MERS exhibits a distinct trend towards positive selection (KS test: p-value *<* 10−7). We also observed a significant difference in the distribution of *πN /πS* ratios between iSNVs and SNPs in the NYC data (KS test, p-value 6*.*29 × 10−12). iSNVs support the purifying selection trend, while SNPs are closer to neutral mutation scenario [4D.](#_3znysh7) When we considered the *πN /πS* ratio across individual genes in SARS-CoV-2 we observed that the purifying selection trend is present in all genes (Figure [4E).](#_3znysh7) Notably, NSP2, NSP3, NSP12, NSP13 and S genes have larger variance in the rates of nonsynonymous to synonymous mutations.

We have further investigated the overlap between iSNV and consensus-level SNPs (Figure [3).](#_1fob9te) We note that there are 15 mutations that have been found in GISAID data, NYC data, and Houston data independently. We also observed that 230 SNVs occur both as an iSNV in at least one sample and as SNPs in the GISAID data. Finally, there are 2 iSNVs that also occur as SNPs (Figure [3).](#_1fob9te) The mutational spectrume variants that occur as both SNPs and iSNVs is similar to the general one outlined above with ∼65% of the changes being C*>*T, followed by ∼15% of G*>*T, and ∼12% of T*>*C.

Finally, we analyzed the potential impact of iSNVs and SNPs on the probes and primers used for detection of SARS-CoV-2(*13, 22*). To evaluate this, we downloaded the set of probes and primers sequences available at the [WHO website,](https://www.who.int/docs/default-source/coronaviruse/whoinhouseassays.pdf) as well as the [Arctic primers.](https://github.com/artic-network/artic-ncov2019/tree/master/primer_schemes/nCoV-2019/V3) Among these, 263 out of 272 sequences contained at least one SNP or iSNV (Figure [5,](#_2et92p0) Table S2). On average, each probe/primer sequence contained 2.529 iSNV and/or 2.477 SNPs . These results suggest the potential for a drop in the sensitivity of the affected probes and primers. We also note that since the iSNV and SNP mutational profiles mimic each other for specific mutations, the potential impact of iSNVs on primer and probe binding should not be overlooked.

## Exploratory Transmission Analysis of Shared SNPs and iSNVs within and across patients

Shared viral genomic variants can be indicative of transmission events and routes (REF). . As proof of concept, we compared paired samples from the same patient taken within 24 hours to assess our ability to detect shared SNV(Figure [6A](#_tyjcwt) and [6B)](#_tyjcwt). In Figure [6A,](#_tyjcwt) we see eight shared SNPs, one shared iSNV, and two shared variants that occur as a SNP in COVHA-20200315-P10-C03-P and as iSNVs in COVHA-20200315-P10-B03-P. As

expected, we find multiple shared SNVs, and two of the three iSNVs in COVHA-20200315P10-B03-P increase to SNPs in the same patient. In Figure [6B,](#_tyjcwt) we see seven shared SNPs and four shared iSNVs. All of the iSNVs occur in both COVHA-20200316-P12-D02-P and COVHA-20200316-P12-G01-P. These results show that our approach is valid for detection of iSNVs?

We next calculated the number of shared iSNVs among all possible pairs of samples (Figure [6C](#_tyjcwt)) For each pair we consider both possible assignments of donor and recipient, narrowing down the donor alleles to only include those with AF between 0.03 and 0.5, and considering a site to be shared if the recipient also has that same variant present at any frequency. We show these results on the raw data from the iSNV calls, as well as on the same data but after applying masking to sites near the ends of the genome. For the raw data before masking, most pairs have 0 to 3 shared variants, with about 150 pairs having 4 or more shared SNVs (Figure 6c). After masking sites near the genome ends, these numbers drop substantially by reducing likely noise from the variant calls, and we see most pairs sharing 0 to 2 variants. When examining each possible pair, one immediately noticeable trend is that site 29871 yields strong signals for shared SNVs between samples with large and similar allele frequencies. We also observe that the number of samples with a variant at that site is unusually high (Figure 6d). In the recent report of De Maio *et al.* (*23*), many sites were examined that showed extensive homoplasy.

In Figure [6](#_tyjcwt) panels E and F, we see two examples of pairs of samples that not only share multiple iSNVs but also at a similar AF. In these pairs, we find many instances of large estimated bottleneck sizes. The lower estimate of 3 for the pair in Figure [6E](#_tyjcwt) is likely due to the variant present at a high AFy in the donor at site 7735 that was absent in the recipient. Out of all samples, we found that the majority of pairs show no signal for an inferred large bottleneck. This is to be expected given that the majority of pairs in a large batch of sequenced SARS-CoV-2 samples are not expected to have been direct or indirect transmissions.

# Discussion

In this study, we have analyzed RNA-seq datasets from 151 COVID-19 positive patients in depth to describe the intrahost variation in SARS-CoV-2. Our analyses yielded four major observations. First, the iSNV profile closely matches the SNP profile inferred from the assembled genomes. In particular, the SARS-CoV-2 genome is enriched with C*>*T changes overall, both for iSNVs and SNPs. Genes NSP6 and NSP10 are particularly enriched for T*>*C mutations, while NSP7 has an enrichment of A*>*C SNVs. Second, the mutational profile of SARS-CoV-2 largely matches that of other Coronaviruses, but with some key differences. SARS-CoV-2 has a significantly larger proportion of G*>*T changes both intra and interhost, when compared to SARS-CoV-1 and MERS. Additionally, we did not see A*>*T SNVs in SARS-CoV-1, as previously reported (*25*). Third, while the SV spectra is likely reflecting the transcriptome landscape of SARS-CoV-2, we detected a significant fraction of small indels that fuel the genetic diversity of SARS-CoV-2. Fourth, the biases present in the SNPs and iiSNVs indicate that there is a complex interplay between endogenous SARS-CoV-2 mutational processes and host-dependent RNA editing. This observation is in line with several recent studies that propose APOBEC and ADAR deaminase activity as a likely driver of the C*>*T changes in the SARS-CoV-2 genomes (*5*).

We also investigated the potential impact of iSNVs and SNPson probes and primers commonly used in RT-PCR based detection and amplicon sequencing of SARS-CoV-2. Most probes we analyzed contain both SNPs and iSNVs. While many platforms can tolerate a few single nucleotide mismatches without the loss of target hybridization, the overall diversity exhibited by SARS-CoV-2 presents potential challenges for probe and primer development. Since we observed a close connection between the SNPs and iSNVs, for future probe and primer designs it could be useful to track the iSNVs to potentially predict and avoid variable regions of the genome. With the integration of these data into design processes at early stages, greater sensitivity could be achieved for hybridization primers and probes even as the virus evolves.

We analyzed paired samples taken from the same COVID-19 positive patient within 24 hours of one another to analyze allele frequencies of SNP and iSNVs. We found that the SNP and iSNV results highlighted strong signal, and also offered a potential hint into the ability of using shared SNPs and iSNVs and their respective allele frequencies for characterizing the elapsed time between samples by tracking intrahost SARS-CoV-2 population dynamics. We also scanned all of the NYC COVID-19 positive samples for putative transmission pairs; we highlighted two examples of potential direct or indirect pairs given shared iSNVs at strikingly similar, high allele frequencies. While these analyses cannot confirm sample pairs as having been involved in direct transmissions without additional confirmatory metadata, this exploratory analysis suggests the possible presence of such transmission pairs (insert ref PMID: 29149252). We also reported high sequence conservation within the NSP3 region, a region that is one of the most diverged from SARS-CoV-1 and MERS-CoV. A number of convergent findings suggest de-mono-ADP-ribosylation of STAT1 by the SARS-CoV-2 NSP3 as a putative cause of the cytokine storm observed in the most severe cases of COVID-19 (*26*). The lower mutational complexity of NSP3 agrees with its functional implications in viral replication (27), and thus the need to conserve its protein structure/function (*25*). Thus, NSP3 may be a good target for drug development since it is well conserved and is essential for viral replication. Follow up studies will be required to solidify functional implications of these observations.

Despite the potential for tremendous insight, the study of intrahost variation in viruses can be confounded by multiple factors. First, the estimated allele frequencies are impacted by variable coverage and transcription patterns. Second, the potentially low viral load in samples can have an impact on downstream sequencing and analysis. Third, previous studies such as De Maio *et al.* (*23*) highlight SARS-CoV-2 sites marked as prone to high homoplasy and need to be taken into consideration for transmission analyses. Lastly, lack of additional metadata imposes a barrier to an in depth study of transmission events. These factors should be addressed in the future studies of iSNVs in SARS-CoV-2.

In summary, our analysis of intrahost variationacross 151 samples from COVID-19 positive patients revealed a complex landscape of within-host diversity that will likely shed additional light on the elusive mechanisms driving the rapid dissemination of SARS-CoV-2. Metatranscriptomic analysis is a powerful tool for interrogating the genomic and transcriptomic landscape of RNA viruses, as it provides a simultaneous peek into viral, bacterial, and host gene expression. Future studies able to integrate all three of these perspectives may hold the key to novel therapies and treatments of this devastating pandemic.

# Materials and methods

## Datasets

We downloaed 6,928 *de novo* assembled SARS-CoV-2 genomes from the GISAID database, available on April, 18th, 2020. We only selected high quality, complete (*>*29 Kbp) genomes We used read data from 11 patient samples collected by Baylor College of Medicine in Houston, Texas. We have also used read data from 140 patient samples collected by Weill Cornell College of Medicine in New York City, New York. Both datasets consist of Illumina NovaSeq 6000 paired-end reads. Host and bacterial genetic material has been removed from the datasets, and we performed all analyses on the viral read data.

We have also additionally used one SARS-CoV-2 sample obtained from direct RNA sequencing with Oxford Nanopore (*12*). For the other coronaviruses data we used 42 samples of SARS-CoV-1 and 53 samples of MERS viral read data (*28*) sequenced by University of Maryland School of Medicine in Baltimore, Maryland.

In total, we analyzed 7,080 SARS-CoV-2, 42 SARS-CoV-1, and 53 MERS samples.

## Read QC and mapping

We processed the Illumina paired-end reads using Trimmomatic ver. 0.39 (*29*) to remove adapter sequences and trim low quality base pairs. We used a universal set of Illumina adapters as a reference for the adapter removal. We set the maximum mismatch count to 2, palindrome clip threshold to 30 and simple clip threshold to 10. We also trimmed leading and trailing low quality (quality value below 3) and ambiguous (N) base pairs. Finally, we applied sliding window trimming cutting the read if the quality score of 4 contiguous bases made the average score drop below 15. After trimming in the final read set we included the reads above the length of 36 with both reads from a pair passing quality control.



We aligned the trimmed reads to the reference genome using Burrows-Wheeler Alignment tool (BWA) ver. 0.7.17 (*30, 31*). We have used paired-end mode for mapping reads to the reference.

We used SAMtools ver. 1.9 to convert the output of *BWA* from SAM to BAM format, and to sort and generate indices for the BAM files (*32*).



## SNV calling and annotation

We used LoFreq ver. 2.1.4 to perform variant calling on the trimmed and mapped reads (*33*). We have filtered the variants with the default LoFreq parameters: minimum coverage was set to 10 and strand-bias FDR correction p-value is greater than 0.001. We have also filtered out the variants occurring below 0.02 AF threshold for the subsequent analyses.

We have annotated the SNVs found in each of the datasets with snpEff ver. 4.3 (*34*).

## SV calling

Structural Variations were identified using Manta (version 1.6.0) (*35*). Subsequently the SV calls were merged using SURVIVOR (v1.0.7) (*36*) using a 100 bp maximum distance between the breakpoints and requiring that the SV types are in agreement in order to merge two SV across the samples. We annotated the SV using a simple 1bp overlap method using bedtools (v2.27.1) (*37*) intersect using the annotations. The same method was used to establish if the start or stop breakpoints of an SV are overlapping with the TRS sites. To test the significance of the overlap we used a permutation test where we randomized the TRS sites (using bedtools random) to generate random TRS with length of 5bp, 1000 times and calculated per TRS the number of start/stop breakpoints of the SV catalog. Subsequently we used this together with the observed overlap using a Kolmogorov–Smirnov (ks.test) with an alternative set to ”two.sided” in R (v 3.2.2).

To generate SV and SNV densities we computed the number of variations per type within a 100bp window. For each variant we counted 1/AF where AF is the frequency of that variant across the samples. This was done based on a custom script available on request. The plot was generated using Circos (v 0.69-8) (*38*)

## Multiple sequence alignment

We have used Parsnp to align the GISAID genomes.

## Phylogenetic tree construction

We have used RAxML (*39*) to infer a phylogenetic tree from the GISAID alignment.

## Transmission Analyses

To compute the number of shared iSNVs in each genomic pair, we utilized the variant calling results and conducted pairwise genome comparisons to count the number of shared variants within individual pairs. Shared iSNVs are those with the same variant nucleotide, and the variant frequencies in the assigned donor sequences are from 0.02 to 0.5. we examined variants with frequencie ≥ 0.02as the cutoff for conservative estimates and to avoid including variants caused by sequence errors. Each genome in a pair is assigned as a donor once. Thus, for a total of 140 samples from New York, there should be 19,460 pairs. Note, since we are looking for putative transmission events, we only analyze 140 samples from New York and exclude all the Houston samples in the transmission analysis. To find pairs with close iSNV allele frequencies we first count the number of genome-wide variants shared between any two samples. Since the variant frequencies in recipient samples partially rely on stochastic replication process in the early infection, we take all iSNVs (with any AFs) into account. If a recipient has a iSNV at position *i* which has the same nucleotide as its donor’s iSNV does at the same position, we say the donor and the recipient share a iSNV at site *i* and record their corresponding AFs as input for the *BB bottleneck* APPROX mode. If the recipient does not have the same base as the donor does or the recipient does not have any iSNVs at position *i* while mapping to the reference sequence, we call this iSNV in the donor sequence as a unique donor iSNV and assign the recipient a 0*.*0 AF at that position.

We have masked the iSNVs that occur between positions 1-55 and 29804-29903 in the genome. Additionally, we masked 25 nucleotide positions between 56-29804 that are highly homoplasic with little phylogenetic signal. These positions are more prone to sequencing and mapping errors (*23*), and therefore were not used in the transmission analyses. We applied the BB bottleneck software to approximate SARS-CoV-2 bottleneck sizes, that is, the founding viral population size in the recipient host (*24*). To

# References

* + 1. J. W. Drake, J. J. Holland, *Proceedings of the National Academy of Sciences* **96**, 13910 (1999).
    2. K. M. Peck, A. S. Lauring, *Journal of Virology* **92**, e01031 (2018).
    3. M. R. Denison, R. L. Graham, E. F. Donaldson, L. D. Eckerle, R. S. Baric, *RNA Biology* **8**, 270 (2011).
    4. A. E. Gorbalenya, L. Enjuanes, J. Ziebuhr, E. J. Snijder, *Virus Research* **117**, 17 (2006).
    5. S. Di Giorgio, F. Martignano, M. G. Torcia, G. Mattiuz, S. G. Conticello, *Science Advances*

p. eabb5813 (2020).

* + 1. L. L. Poon, *et al.*, *Nature Genetics* **48**, 195 (2016).
    2. C. Barbezange, *et al.*, *Frontiers in Microbiology* **9**, 2596 (2018).
    3. M. K. Borucki, *et al.*, *PloS One* **14** (2019).

9. D. J. Park, *et al.*, *Cell* **161**, 1516 (2015).

1. D. Ramazzotti, *et al.*, *bioRxiv* (2020).
2. S. Elbe, G. Buckland-Merrett, *Global Challenges* **1**, 33 (2017).
3. G. Taiaroa, *et al.*, *bioRxiv* (2020).
4. C. Farkas, F. Fuentes-Villalobos, J. L. Garrido, J. Haigh, M. I. Barr´ıa, *PeerJ* **8**, e9255 (2020).
5. D. Mercatelli, F. M. Giorgi (2020).
6. W. Rut, *et al.*, *bioRxiv* (2020).
7. D. Kim, *et al.*, *Cell* (2020).
8. I. Sola, F. Almazan, S. Zuniga, L. Enjuanes, *Annual Review of Virology* **2**, 265 (2015).
9. M. Mahmoud, *et al.*, *Genome Biology* **20**, 246 (2019).
10. A. D. Davidson, *et al.*, *BioRxiv* (2020).
11. D. Muth, *et al.*, *Scientific Reports* **8**, 1 (2018).
12. J. Gregori, *et al.*, *Bioinformatics* **30**, 1104 (2014).
13. K. A. Khan, P. Cheung, *Royal Society Open Science* **7**, 200636 (2020).
14. N. De Maio, *et al.*, *Virological* **6**, 80 (2012).
15. A. S. Leonard, D. B. Weissman, B. Greenbaum, E. Ghedin, K. Koelle, *Journal of Virology*

**91**, e00171 (2017).

1. G. M. Pavlovic´-Lazˇetic´, N. S. Mitic´, M. V. Beljanski, *BMC bioinformatics* **5**, 65 (2004).
2. Y. M. Alhammad, *et al.*, *bioRxiv* (2020).
3. Y. M. Ba´ez-Santos, A. M. Mielech, X. Deng, S. Baker, A. D. Mesecar, *Journal of virology*

**88**, 12511 (2014).

1. C. M. Coleman, M. B. Frieman, *PLoS Pathogens* **9** (2013).
2. A. M. Bolger, M. Lohse, B. Usadel, *Bioinformatics* **30**, 2114 (2014).
3. H. Li, R. Durbin, *Bioinformatics* **25**, 1754 (2009).
4. H. Li, *arXiv preprint arXiv:1303.3997* (2013).
5. H. Li, *et al.*, *Bioinformatics* **25**, 2078 (2009).
6. A. Wilm, *et al.*, *Nucleic Acids Research* **40**, 11189 (2012).
7. P. Cingolani, *et al.*, *Fly* **6**, 80 (2012).
8. X. Chen, *et al.*, *Bioinformatics* **32**, 1220 (2016).
9. D. C. Jeffares, *et al.*, *Nature Communications* **8**, 1 (2017).
10. A. R. Quinlan, I. M. Hall, *Bioinformatics* **26**, 841 (2010).
11. M. Krzywinski, *et al.*, *Genome Research* **19**, 1639 (2009).
12. A. Stamatakis, *Bioinformatics* **30**, 1312 (2014).

# Acknowledgments

## Funding information

N.S. and Y.L. are supported by the Department of Computer Science, Rice University. Q.W., D.A., T.J.T, and R.A.L.E. are supported by startup funds from Rice University. M.J. is supported under NIH award No. R01HD091731 from the NICHD. F.J.S. acknowledges funding and part of the data was produced by Baylor College of Medicine under NIAID (U19AI144297-01). A.B. is supported by supplemental funds for COVID-19 research from Translational Research Institute through NASA Cooperative Agreement NNX16AO69A (T-0404) and further funding was provided by KBR, Inc. D.P. is supported by the European Research Council (ERC-617457-PHYLOCANCER), Spanish Ministry of Economy and Competitiveness, and Xunta de Galicia. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the ODNI, IARPA, ARO, or the US Government. [ADD REMAINING FUNDING INFO]

## Author contributions

N.S. analyzed and interpreted the data, generated the figures, and wrote the manuscript. Y.L. analyzed single nucleotide variant data and generated the figures. D.A. analyzed phylogenetic data and generated the figures. Q.W. analyzed and interpreted viral transmission data, generated figures, and wrote the manuscript. R.A.L.E. interpreted the viral transmission and phylogenetic data, and wrote and edited the manuscript. T.J.T. interpreted the data, edited and wrote the manuscript. M.M. and F.J.S. lead the SV analysis, interpretation of the data and edited and wrote the manuscript. M.J. analyzed the impact of polymorphisms on probes and primers, and generated figures. A.B. edited the manuscript and exchange of ideas. K.T. reviewed the SNV commands and called variants in public COVID-19 metatranscriptomes for comparison. D.P. proposed some of the analyses, helped with their interpretation, and contributed to manuscript writing. [ADD REMAINING CONTRIBUTION INFO]

## Competing interests

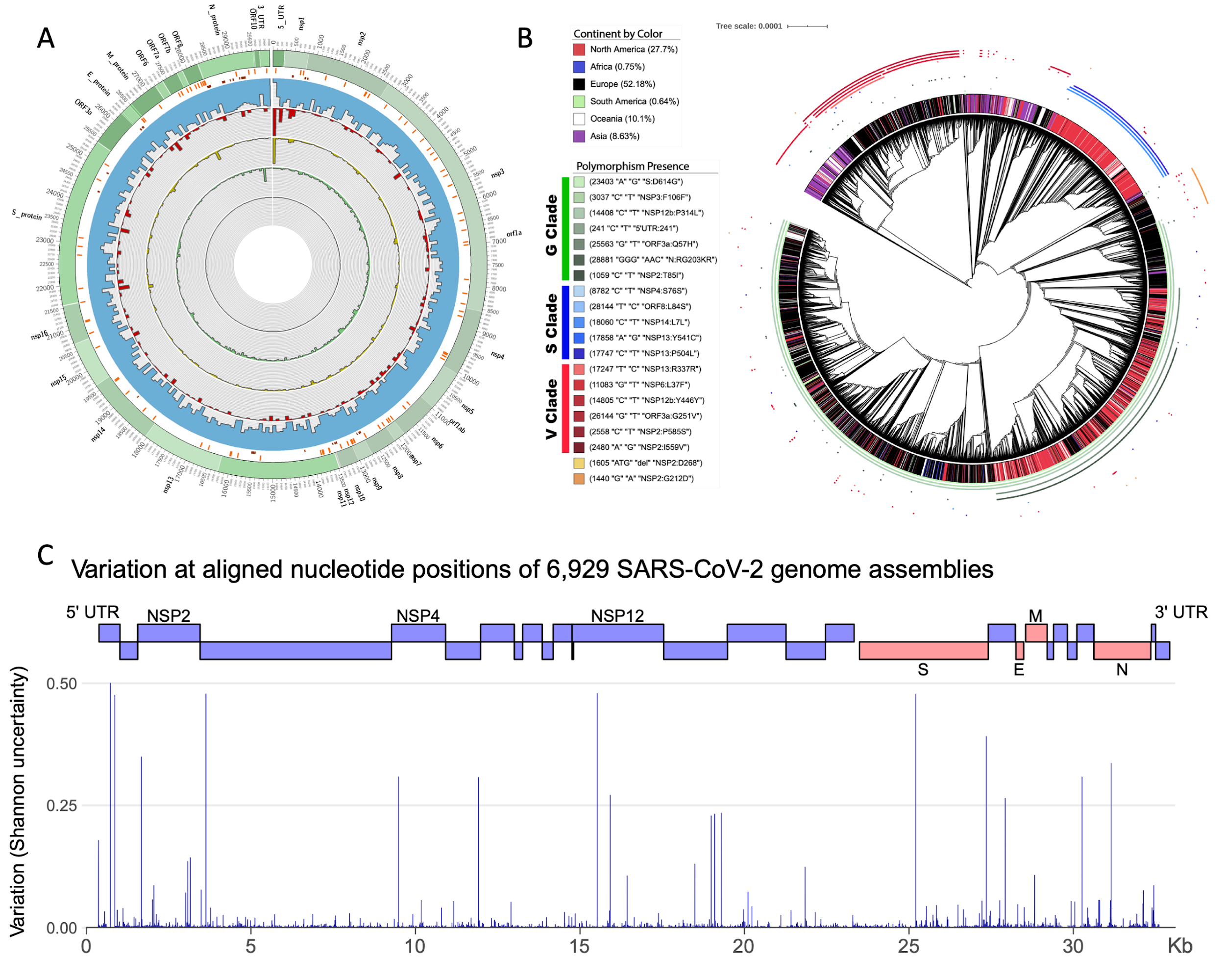
Authors declare no competing interests.

## Data availability

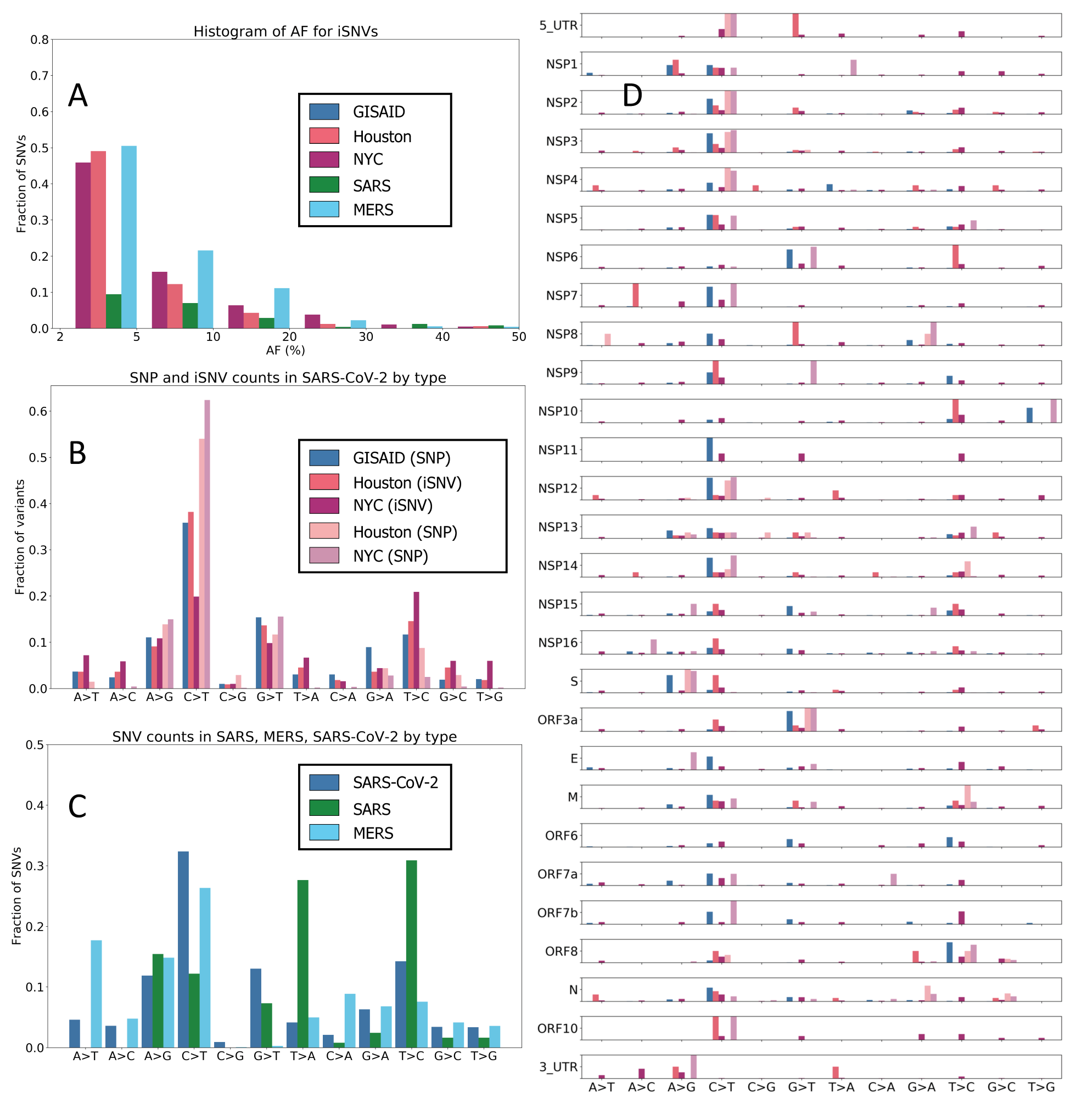
Variant calling files and other raw data are available at [https://rice.box.com/v/SARS-COV-2-SNV-data.](https://rice.box.com/v/SARS-COV-2-SNV-data) Assembled genomes for SARS-CoV-2 used in the analysis are available at GISAID. SARS-CoV-1 and MERS read data were obtained from the study PRJNA233943. Scripts used for data analysis are available at [https://gitlab.com/treangenlab/covirt\_scripts.](https://gitlab.com/treangenlab/covirt_scripts)

# Supplementary materials

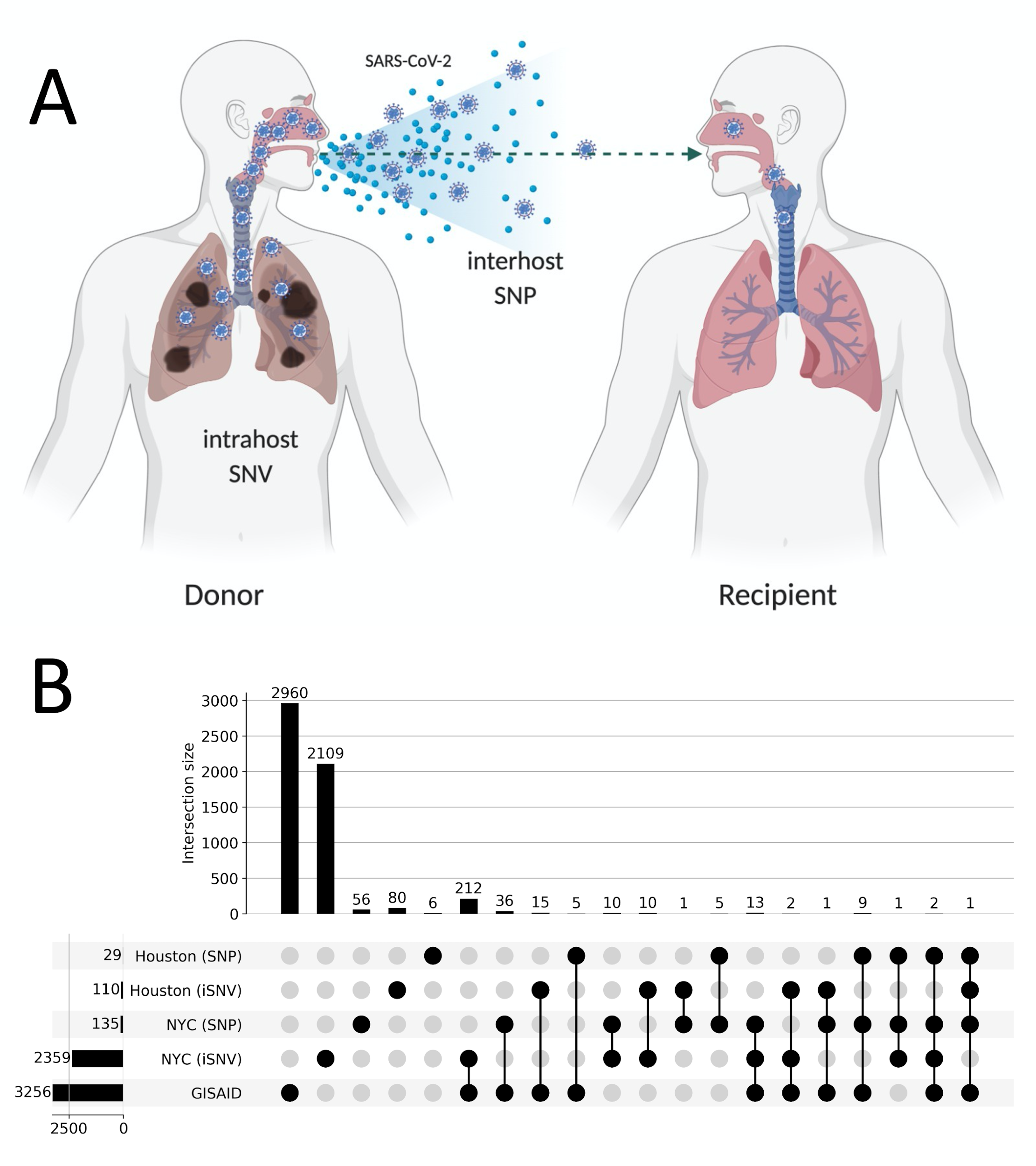
Supplementary Text Table S1, S2



**Figure 1: Overview of general diversity of SARS-CoV-2. A.** From outer to inner layers: Annotation of SARS-CoV-2 genome (green), transcription-regulating sequences (TRS) (orange), PCR primer designs (dark red), intrahost variant density including iSNVs (blue), deletions start sites (red), duplication start sites (yellow), inversion start sites (green) and insertions (dark green) along the entire genome. For SNPs + iSNVs + SVs we plotted the density scaled by their allele frequency across the population over 100bp windows. **B.** Directly outside of the tree branches is the continuous annotation ring for the continents corresponding to each GISAID sample. The set of smaller non-continuous rings, surrounding the continent annotation ring, are the clade-specific interhost polymorphisms as described in (*14*). The G clade polymorphisms are colored as different shades of green, the S clade ones are colored different shades of blue, and the V clade ones are different shades of red.



**Figure 2:** **Mutational frequencies of iSNV and SNPs. A.** *Distribution of iSNV AF.* We note that the distribution of AF is strictly less than 50% as iSNVs are below consensus-level by definition. We see a spike in iSNVs from 2% to 5%, with a rapid decrease in higher allele frequencies (AFs). **B. i***SNV frequency in SARS-CoV-2 by type.* We see that the largest fraction of both iSNVs and SNPs corresponds to a C*>*T change. The A*>*G, G*>*A, T*>*C, and G*>*T changes are also common. **C.** *Comparative iSNV frequencies in SARS-CoV-1, SARS-CoV-2, and MERS.* We note that SARS-CoV-1 data lacks A*>*T changes, and the G*>*T change is more common in the SARS-CoV-2 data. **D. i***SNV frequency by ORF/NSP.* We observe that NSP6 and NSP10, have an enrichment of T*>*C changes, and NSP7 has an enrichment of A*>*C change

**Figure 3**: **Shared SNPs and SNVs across datasets. A.** Illustration differentiating what we define as an intrahost SNV (iSNV) and as an interhost consensus-level SNP. **B**. This UpSet plot captures the shared single nucleotide variants between iSNVs and consensus-level SNPs. There are 15 polymorphisms that occur both as iSNVs and SNPs, and are present in all three datasets independently. There are a total of 230 iSNVs that also appear as SNPs in the GISAID data.

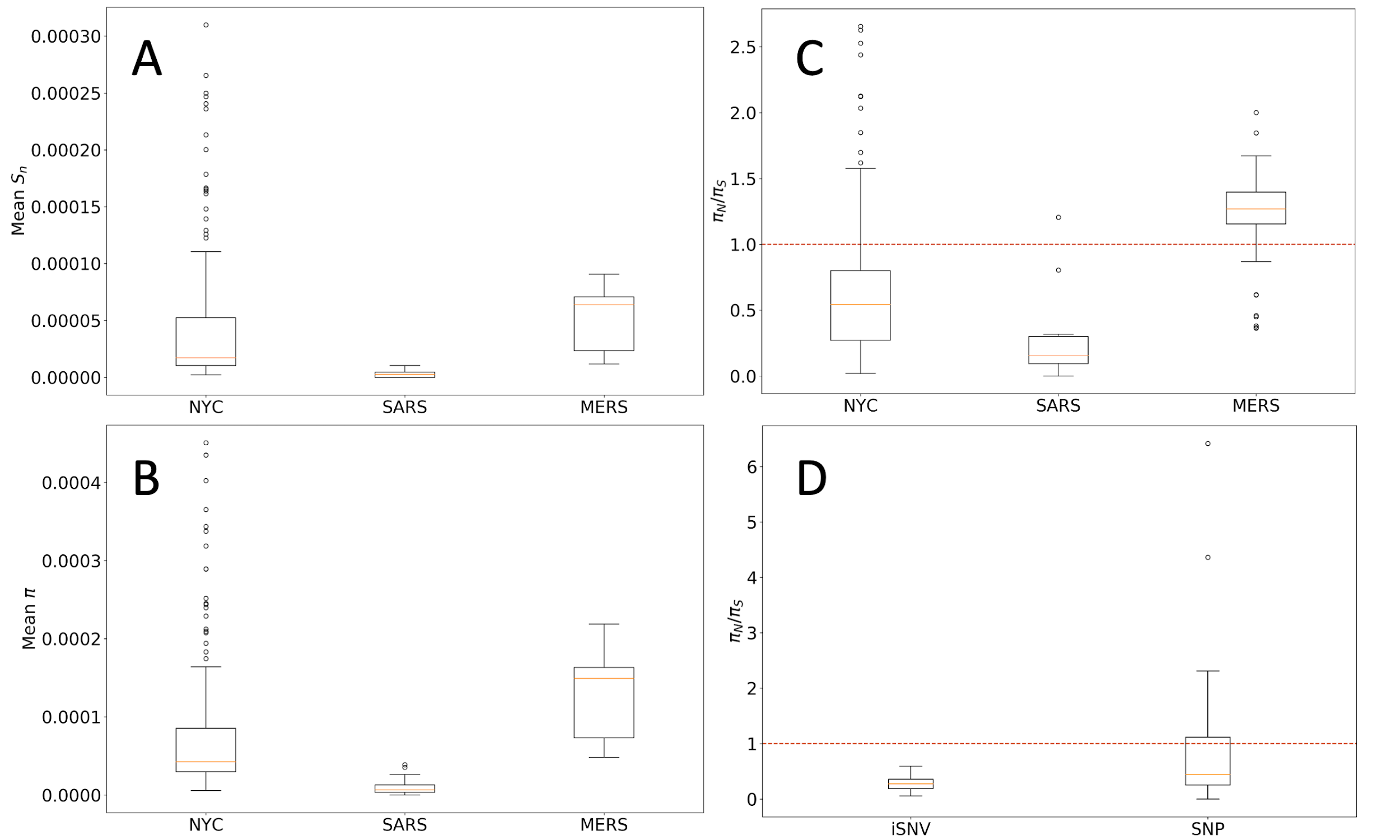


Figure 4: **Complexity and diversity in Coronaviruses. A.** *Intrahost complexity of Coronavirus samples.* This plot shows the mean *Sn* complexity of samples for SARS-CoV-2, SARS-CoV-1 and MERS. **B.** *Diversity of Coronavirus samples.* This plot shows the mean *π* diversity of samples. **C.** *Synonymous vs non-synonymous mutation ratios.* We note that the data from MERS samples indicate positive selection, while SARS-CoV-1 and SARS-CoV-2 samples indicate purifying selection. **D.** *Syn. vs non-syn. mutation ratios for iSNVs and SNPs in NYC data.*

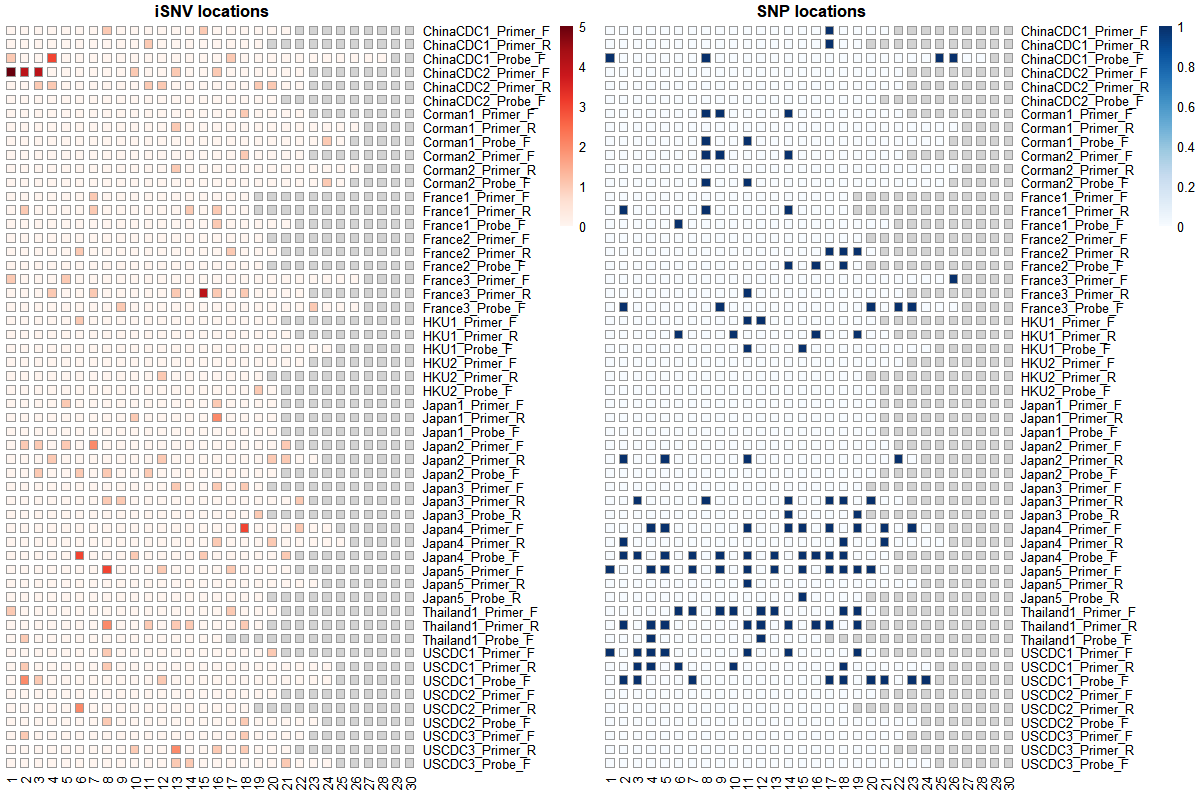
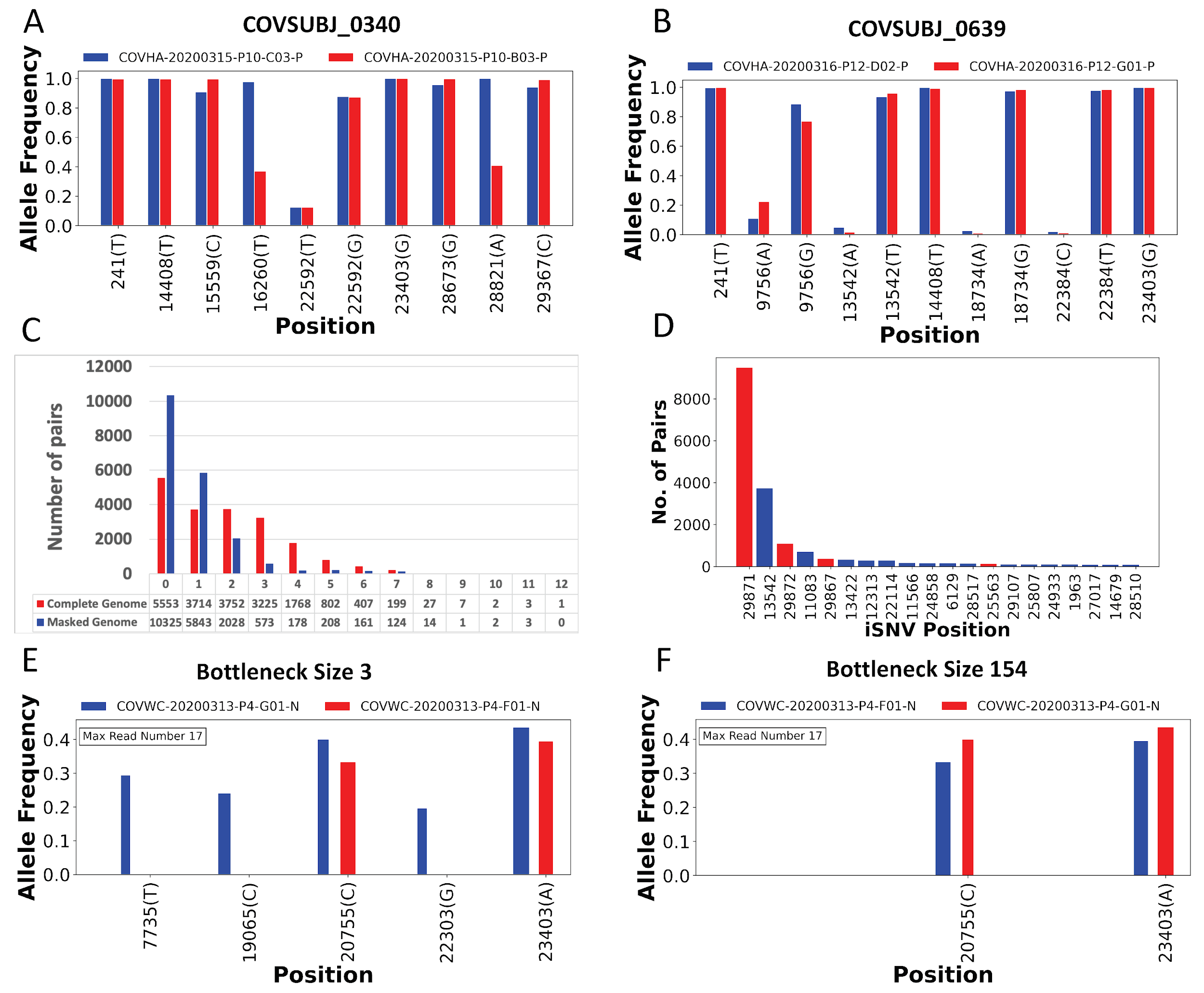


Figure 5: **iSNV and SNP presence on widely-used primers and probes.** This figure shows the locations on WHO probes and primers that contain iSNVs (left) and SNPs (right). Columns correspond to base pair positions within the probe, and the sequences are 3’ aligned. Rows corresponding to the oligonucleotide sequences and highlighted squares indicate that the position is affected by a SNV in one or more samples.



**Figure 6: In-depth analysis of shared iSNVs. A.** Paired samples from patient COVSUBJ 0340 in NYC. Ten total shared SNVs, including three iSNVs and seven SNPs. **B.** Paired samples from patient COVSUBJ 0639 in NYC. Eleven total shared SNVs, including four iSNVs and seven SNPs. **C.** The distribution of the number of genomic pairs and their shared iSNVs. **D.** The number of samples with iSNVs at given nucleotide positions. **E, F.** Allele frequencies and presence of shared iSNVs between two unpaired samples. Blue color represents donor and red color represents recipient. The bar width is proportional to the number of reads supporting the variants. The minimum bar width represents 10 reads. Bottleneck size was estimated to be 3 for **E** and 154 for **F**.

1. [J. W. Drake, J. J. Holland, Mutation rates among RNA viruses. *Proceedings of the National Academy of Sciences*. **96** (1999), pp. 13910–13913.](http://paperpile.com/b/U1nDEi/iSDa)

2. [K. M. Peck, A. S. Lauring, Complexities of viral mutation rates. *J. Virol.* **92**, e01031–17 (2018).](http://paperpile.com/b/U1nDEi/b3GU)

3. [M. R. Denison, R. L. Graham, E. F. Donaldson, L. D. Eckerle, R. S. Baric, Coronaviruses: an RNA proofreading machine regulates replication fidelity and diversity. *RNA Biol.* **8**, 270–279 (2011).](http://paperpile.com/b/U1nDEi/Gd7N)

4. [A. E. Gorbalenya, L. Enjuanes, J. Ziebuhr, E. J. Snijder, Nidovirales: evolving the largest RNA virus genome. *Virus Res.* **117**, 17–37 (2006).](http://paperpile.com/b/U1nDEi/ZPaq)

5. [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. *Science Advances*, eabb5813 (2020).](http://paperpile.com/b/U1nDEi/aVN2)

6. [L. L. M. Poon, T. Song, R. Rosenfeld, X. Lin, M. B. Rogers, B. Zhou, R. Sebra, R. A. Halpin, Y. Guan, A. Twaddle, Others, Quantifying influenza virus diversity and transmission in humans. *Nat. Genet.* **48**, 195 (2016).](http://paperpile.com/b/U1nDEi/ns6k)

7. [C. Barbezange, L. Jones, H. Blanc, O. Isakov, G. Celniker, V. Enouf, N. Shomron, M. Vignuzzi, S. van der Werf, Seasonal genetic drift of human influenza A virus quasispecies revealed by deep sequencing. *Front. Microbiol.* **9**, 2596 (2018).](http://paperpile.com/b/U1nDEi/UxGv)

8. [M. K. Borucki, N. M. Collette, L. L. Coffey, K. K. A. Van Rompay, M. H. Hwang, J. B. Thissen, J. E. Allen, A. T. Zemla, Multiscale analysis for patterns of Zika virus genotype emergence, spread, and consequence. *PLoS One*. **14** (2019).](http://paperpile.com/b/U1nDEi/Me6J)

9. [D. J. Park, G. Dudas, S. Wohl, A. Goba, S. L. M. Whitmer, K. G. Andersen, R. S. Sealfon, J. T. Ladner, J. R. Kugelman, C. B. Matranga, Others, Ebola virus epidemiology, transmission, and evolution during seven months in Sierra Leone. *Cell*. **161**, 1516–1526 (2015).](http://paperpile.com/b/U1nDEi/YDZg)

10. [D. Ramazzotti, F. Angaroni, D. Maspero, C. Gambacorti-Passerini, M. Antoniotti, A. Graudenzi, R. Piazza, Characterization of intra-host SARS-CoV-2 variants improves phylogenomic reconstruction and may reveal functionally convergent mutations. *bioRxiv* (2020).](http://paperpile.com/b/U1nDEi/MRIn)