Hidden genomic diversity of SARS-CoV-2: implications for qRT-PCR diagnostics and transmission

Nicolae Sapoval[1], Medhat Mahmoud[2], Michael D. Jochum[3], Yunxi Liu[1], R. A. Leo Elworth[1], Qi Wang[4], Dreycey Albin[4], Huw Ogilvie[1], Michael D. Lee[5,6], Sonia Villapol[7], Kyle M. Hernandez[8,16], Irina Maljkovic Berry[9], Jonathan Foox[10], Afshin Beheshti[11], Krista Ternus[12], Kjersti M. Aagaard[3], David Posada[13,14,15], Christopher E. Mason[10], Fritz Sedlazeck[2,†], Todd J. Treangen[1,†*]

[1]Department of Computer Science, Rice University, Houston, TX, USA.
[2]Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX.
[3]Baylor College of Medicine and Texas Children's Hospital, Houston, TX.
[4]Systems, Synthetic, and Physical Biology (SSPB) Graduate Program, Houston, TX.
[5]Exobiology Branch, NASA Ames Research Center, Mountain View, CA.
[6]Blue Marble Space Institute of Science, Seattle, WA.
[7]Houston Methodist Research Institute, Houston, TX.
[8]Department of Medicine, University of Chicago, Chicago, IL.
[9]Walter Reed Army Institute of Research, Silver Spring, MD.
[10]Department of Physiology and Biophysics, Weill Cornell Medicine, New York, New York.
[11]KBR, Space Biosciences Division, NASA Ames Research Center, Moffett Field, CA.
[12]Signature Science, LLC, 8329 North Mopac Expressway, Austin TX 78759.
[13]Biomedical Research Center (CINBIO), University of Vigo, 36310 Vigo, Spain.
[14]Department of Biochemistry, Genetics and Immunology
School of Biology, University of Vigo, Vigo, Spain.
[15]Galicia Sur Health Research Institute, 36310 Vigo, Spain.
[16]Center for Translational Data Science, University of Chicago, Chicago, IL.

*To whom correspondence should be addressed; E-mail: 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 disease. Though RNA viruses mutate more rapidly than DNA viruses, 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 both intrahost and interhost diversity. Our intrahost and interhost diversity analyses yielded three major observations. First, the mutational profile of SARS-CoV-2 highlights iSNV and SNP similarity, albeit with high variability in C>T changes. Second, iSNV and SNP patterns in SARS-CoV-2 are more similar to MERS-CoV than SARS-CoV-1. Third, a significant fraction of small indels fuel the genetic diversity of SARS-CoV-2. Altogether, our findings provide insight into SARS-CoV-2 genomic diversity, inform the design of detection tests, and highlight the potential 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*). Additionally, SARS-CoV-2 contains a common 69-bp 5' leader sequence fused to the body sequence from the 3' end of the genome (*5*). 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.

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 genomic data from the Global Initiative on Sharing All Influenza Data (GISAID) (*6*) indicated that the SARS-CoV-2 mutational rate (*7*) was similar to other CoVs (*8*). 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 (*9–12*), including often overlooked structural variation. 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 (*13*, *14*). However, these studies were based on a limited number of samples (<20). 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 GISAID, 11 sequencing samples from the Baylor College of Medicine, and 140 sequencing samples from the Weill Cornell College of Medicine.

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 sensitive to SARS-CoV-2. In this setting, sensitivity determines how well it can capture the diversity of all SARS-CoV-2 variants. Lack of sensitivity leads to an increase in false positive qRT-PCR results, as few as two mismatches can result in increases in CT values and degradation in accuracy of viral load estimates (*15*, *16*). Moreover, recent studies on Ebolavirus and flu viruses (*12*, *17*) highlight the importance of intrahost variation for studying viral population dynamics and transmission scenarios. In summary, 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.

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). We evaluated structural variants across the 151 samples in both NYC and Houston; the inferred SVs are shown in Figure 1A. We also evaluated single nucleotide variants in GISAID representing single nucleotide polymorphisms (SNPs), while the variants analyzed in the Houston and NYC datasets include both SNPs and intrahost single nucleotide variants (iSNVs). The inferred phylogenetic tree of GISAID genomes with clade-defining (*18*) SNPs is shown in Figure 1B. We note that these previously reported clade-defining SNPs distinguish the geographic origin of SARS-CoV-2 genomes, with clades G and S predominantly covering North American genomes and clade V covering a portion of Asian and European genomes. We also observe that some of the clade-defining SNPs occur intermittently outside of the main phylogenetic clades. 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 identified 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 1A). Overall, since we are identifying SVs based 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. We observed a significant overlap (Kolmogorov–Smirnov test: p-value=4.95^{-5}, D=0.25) for the 98 start and 63 end breakpoints with the annotated transcription regulating sequences (TRS) (dark red Figure 1A). Subsequently, we focus on smaller SVs (<1kbp) that more likely indicate true underlying SV rather than transcription signals. We identified 247 deletions and 23 insertions across all 170 SARS-CoV-2 genomes. The imbalance of insertions and deletions is likely due to the low ability to detect insertions using short reads (*19*). Figure 1A shows the allele frequency (AF) of these SVs across all samples. We observed 8 deletions shared among 34 or more samples (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) (AF: 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 an enrichment of SVs in NSP11 and NSP12 when taking the size of the annotated regions into account (Supplementary Figure 1). 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 (*20*) 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) was present in 10 samples (AF: 6%) in ORF8, a potentially important gene for viral adaptation to humans (*21*).

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) as it is a common threshold for consensus-calling in genome assembly (*22*, *23*). Figure 2A shows the iSNV AF 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). We also note that A>G, G>A, and G>T iSNVs are common. 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). 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). 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. 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 (*24*). The NSP3 protein is also one of the most diverged regions of SARS-CoV-2 compared to SARS-CoV-1 and MERS-CoV.

We note that the mutational spectra for SNPs matches the one observed for iSNVs, namely A>G, G>A, T>C and G>T are most common (Figure 2B). 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 fraction of GISAID C>T SNPs is nearly to the fraction of Houston C>T iSNVs, clearly distinguishing GISAID SNPs and Houston iSNVs from Houston and NYC SNPs. We also note that the mutational spectra of SNPs across the genes of SARS-CoV-2 closely match the iSNV mutational spectra (Figure 2D). The mutational spectrum of NYC SNPs is significantly different from both NYC iSNVs mutational spectrum (Kolmogorov-Smirnov (KS) test: p-value $\sim 10^{-100}$) and GISAID SNPs mutational spectrum (KS test: p-value $\sim 10^{-40}$). When compared to SARS and MERS, SARS-CoV-2 has a larger proportion of G>T iSNVs (Figure 2C). 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.

To further investigate patterns of difference and similarity between SNPs and iSNVs, we analyzed the functional impact of the observed variants. First, in GISAID SNPs we observe 1191 (36.45%) synonymous, 2021 (61.86%) missense, and 40 (1.22%) stop gained variants. In NYC iSNVs we observed 782 (31.68%) synonymous, 1549 (62.76%) missense, and 73 (2.96%) stop gained variants. Finally, in Houston iSNVs we observed 43 (31.16%) synonymous, 86 (62.31%) missense, and 5 (3.62%) stop gained variants. Altogether, about two thirds of all observed variants are missense and about a third are synonymous, with good agreement of these values for both SNPs and iSNVs. We also investigated the overlap between iSNV and consensus-level SNPs (Figure 3B). 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 3B). The mutational spectrum of variants that occur as both SNPs and iSNVs is similar to the general one outlined above with $\sim$65% of the changes being C>T, followed by $\sim$15% of G>T,

and ~12% of T>C.

Prior studies have found iSNVs early in virus outbreaks that later establish as SNPs (*25*, *26*). Thus, we looked into whether clade-defining SNPs identified in a previous study (*18*) co-occur with iSNVs identified in NYC and Houston datasets. We found that a G and S clade-defining SNPs co-occur with 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 both NYC and Houston), and a less common one is a T>A change occurring only in the NYC data. This indicates the emergence of an iSNV strongly correlated with the North American clade of the SARS-CoV-2.

Next, we estimated the genetic complexity (S_n) (*27*) and genetic diversity (π) of SARS-CoV-2, SARS-CoV-1 and MERS (Figure 4A,B). 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 (π_N/π_S) for iSNVs in SARS-CoV-2, SARS-CoV-1 and MERS data (Figure 4C). The genome-wide π_N/π_S values suggest that SARS-CoV-2 (median π_N/π_S: 0.554) and SARS-CoV-1 (median π_N/π_S: 0.179) might be predominantly under purifying selection, while MERS (median π_N/π_S: 1.270) seems to be overall under positive selection (KS test: p-value < 10^{-7}). We also observed a significant difference in the distribution of π_N/π_S ratios between iSNVs and SNPs in the NYC data (KS test, p-value 6.29×10^{-12}). The SARS-CoV-2 π_N/π_S values are significantly lower for iSNVs (median π_N/π_S: 0.273) than for SNPs (median π_N/π_S: 0.446, Figure 4D). The π_N/π_S ratios are consistent across ORFs/NSPs of SARS-CoV-2 (Supplementary Figure 2).

Finally, we analyzed the potential impact of iSNVs and SNPs on the probes and primers used for detection of SARS-CoV-2 (*15*, *28*). To evaluate this, we downloaded the set of probes and primers sequences available at the WHO website, as well as the Arctic primers. Among these, 263 out of 272 sequences contained at least one SNP or iSNV (Figure 5, 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 given the possibility of iSNVs establishing as SNPs (*26*).

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

Shared viral genomic variants can be indicative of transmission events and routes (*29*), and iSNVs are a critically important tool for discerning direct transmission and for bottleneck calculations (*30*). To assess our ability to identify shared iSNVs and SNPs across samples, we first compared all NYC paired samples from the same patient taken within 24 hours (Figure 6A,B). In Figure 6A, we see eight shared SNPs, one shared iSNV, and two shared iSNVs that occur as a SNP in patient 340 sample C03 and as iSNVs in patient 340 sample B03. As expected, we find multiple shared SNVs, and two of the three iSNVs in patient 340 sample B03 occur as SNPs in patient 340 sample C03. In Figure 6B, we see seven shared SNPs and four shared iSNVs. All of the iSNVs occur in both patient 639 sample D02 and patient 639 sample G01. These results highlight our ability to identify iSNVs and the feasibility of using iSNVs for identifying paired samples and potential transmission pairs.

We next calculated the number of shared iSNVs among all possible pairs of NYC samples (Figure 6C). For each pair we consider both possible assignments of donor

and recipient, narrowing down the donor alleles to only include those with AF between 0.02 and 0.5, and considering a site to be shared if the recipient also has that same variant present as either a iSNV or SNP. 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 AFs. We also observe that the number of samples with a variant at that site is unusually high (Figure 6d). In Figure 6 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 is likely due to the variant present at a high AF in the donor at site 7735 that was absent in the recipient.

Discussion

In this study, we have analyzed over 7,000 SARS-CoV-2 genomes in addition to 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 mutational spectra closely match the SNP mutational spectra inferred from the consensus 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 in both iSNVs and SNPs, when compared to SARS-CoV-1 and MERS. Additionally, we did not see A>T SNVs in SARS-CoV-1, as previously reported (*31*). 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 mutational spectra of the SNPs and iSNVs 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 (*14*). Of note, this recent study also reported that the number of observed transversions are compatible with mutation rates found in other Coronaviruses (*8*, *14*).

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 (*32*). The lower mutational complexity of NSP3 agrees with its functional implications in viral replication, and thus the need to conserve its protein structure/function (*31*, *33*). 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.

We also investigated the potential impact of iSNVs and SNPs on 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 AFs of SNP and iSNVs. We found that the SNP and iSNV profiles and AFs were concordant, indicating the potential of using shared SNPs and iSNVs and their respective AFs for 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 AFs. 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. Of note, the recent report of De Maio *et al. (34)*, many sites were examined that showed extensive homoplasy. 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 (*29*).

Despite the potential for tremendous insight, the study of intrahost variation in viruses can be confounded by multiple factors. First, the estimated AFs are impacted by variable coverage and transcription patterns. Second, low viral load (Ct values above 32) in samples can have an impact on downstream sequencing and analysis (*35*, *36*) (Supplementary Figure 3). Third, previous studies such as De Maio *et al. (34)* 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 variation across 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 downloaded 6,928 SARS-CoV-2 consensus 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.

For the other coronaviruses data we used 42 samples of SARS-CoV-1 and 53 samples of MERS viral read data (*37*) sequenced by University of Maryland School of Medicine in Baltimore, Maryland.

In total, we analyzed 7,079 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 (*38*)

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 (*39*, *40*). We have used paired-end mode for mapping reads to the SARS-CoV-2 reference genome (NC_045512).

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 (*41*).

SNV calling and annotation

We used LoFreq ver. 2.1.4 to perform variant calling on the trimmed and mapped reads (*42*). We have filtered the variants with the default LoFreq parameters: minimum coverage was set to 10, phred quality-score set to Q20 (99%), 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, and required all iSNVs to be supported by 10X minimum coverage. We annotated the SNVs found in each of the datasets with snpEff ver. 4.3 (*43*). We used SNPGenie (*44*) with the default set of parameters to estimate the genetic diversity and non-synonymous to synonymous diversity ratios in SARS-CoV-2, SARS-CoV-1 and MERS data.

SV calling

Structural Variations were identified using Manta (version 1.6.0) (*45*). Subsequently the SV calls were merged using SURVIVOR (v1.0.7) (*46*) 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) (*47*) 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) (*48*).

Phylogenetic tree construction

We used Parsnp (ver. 1.2) (*49*) to align the GISAID genomes. We set the maximal cluster D value to 30,000, and the rest of the parameters were set to the default values. We used RAxML (*50*) to infer a phylogenetic tree from the GISAID alignment. We ran RAxML with default parameters using GTRCAT approximation model for tree scoring. We used the best-scoring maximum likelihood tree output from RAxML.

Variation in alignment of GISAID assemblies

A multiple sequence alignment of the 6,928 SARS-CoV-2 assemblies was generated with mafft v7.458 (*51*) with the *-auto* option. Variation per column in the

alignment was generated with bit v1.8.02 (*52*) which utilizes the scikit-bio (*52*, *53*) implementation for calculating Shannon uncertainty.

Probe and primer mapping

Primer and probe sequences were derived from the WHO website (*54*) and hCoV-2019/nCoV-2019 Version 3 Amplicon Set (*55*). We mapped probes and primers against the SARS-CoV-2 reference genome (NC_045512) with bowtie2 (*56*). Analysis of the primer and probe mapping regions was performed with a custom Python script and visualizations were done with R-3.6.1.

Transmission Analyses

To compute the number of shared iSNVs in each genomic pair, we utilized the variant calling results. We conducted pairwise genome comparisons and counted the number of shared variants within individual pairs. For each pair, we consider both combinations of one sample as a putative donor and one sample as a putative recipient. Shared iSNVs were then defined as iSNVs that share the same variant nucleotide between the two samples, and where the variant frequencies in the assigned donor sequences are from 0.02 to 0.5. We examined variants with frequencies $\geq$ 0.02 as the cutoff for conservative estimates to avoid including variants caused by sequencing errors. For the 140 samples from New York, given that we consider each pair twice, there are 19,460 pairs. Note, since we are looking for putative transmission events, we can only consider samples within the same geographic region, so we limited our analyses to the 140 samples that all came from New York. We 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. These positions are more prone to sequencing and mapping errors (*34*), 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 (*57*). Since the variant frequencies in recipient samples partially rely on stochastic replication processes in the early infection, we take all iSNVs (with any AFs) into account (from 0.0 to 1.0) within a shared variant for putative recipients. Furthermore iSNVs from either donors or recipients are supported by at least 10 reads to be included in the bottleneck size analysis . We use the AFs of shared iSNVs between putative donor and recipient pairs as input for the *BB bottleneck* APPROX mode (*57*). If the recipient does not have the iSNV with the same base at the same site as the donor or simply does not have any variant called at position i while mapping to the reference sequence, we assign the recipient a 0.0 AF at that position. Finally, we consider the case where the iSNV base is the same as the reference sequence base. In this case, for instance, when a variant is called at a site with 0.7 AF and no other variants are present, we take the reference base as an iSNV with 0.3 AF if there are no other reads present with an alternate allele and there are at least 10 reads mapping to the reference base.

References

1. J. W. Drake, J. J. Holland, Mutation rates among RNA viruses. *Proceedings of the National Academy of Sciences*. **96** (1999), pp. 13910–13913.

2. K. M. Peck, A. S. Lauring, Complexities of viral mutation rates. *J. Virol.* **92**, e01031–17 (2018).

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).

4. A. E. Gorbalenya, L. Enjuanes, J. Ziebuhr, E. J. Snijder, Nidovirales: evolving the largest RNA virus genome. *Virus Res.* **117**, 17–37 (2006).

5. I. Sola, F. Almazan, S. Zuniga, L. Enjuanes, Continuous and discontinuous RNA synthesis in coronaviruses. *Annual Review of Virology*. **2**, 265–288 (2015).

6. S. Elbe, G. Buckland-Merrett, Data, disease and diplomacy: GISAID's innovative

contribution to global health. *Global Challenges*. **1**, 33–46 (2017).

7. Z. Shen, Y. Xiao, L. Kang, W. Ma, L. Shi, L. Zhang, Z. Zhou, J. Yang, J. Zhong, D. Yang, L. Guo, G. Zhang, H. Li, Y. Xu, M. Chen, Z. Gao, J. Wang, L. Ren, M. Li, Genomic Diversity of Severe Acute Respiratory Syndrome–Coronavirus 2 in Patients With Coronavirus Disease 2019. *Clinical Infectious Diseases* (2020), , doi:10.1093/cid/ciaa203.

8. L. D. Eckerle, M. M. Becker, R. A. Halpin, K. Li, E. Venter, X. Lu, S. Scherbakova, R. L. Graham, R. S. Baric, T. B. Stockwell, D. J. Spiro, M. R. Denison, Infidelity of SARS-CoV Nsp14-exonuclease mutant virus replication is revealed by complete genome sequencing. *PLoS Pathog.* **6**, e1000896 (2010).

9. 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).

10. 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).

11. 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).

12. 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).

13. 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).

14. 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).

15. C. Farkas, F. Fuentes-Villalobos, J. L. Garrido, J. Haigh, M. I. Barría, Insights on early mutational events in SARS-CoV-2 virus reveal founder effects across geographical regions. *PeerJ*. **8**, e9255 (2020).

16. D. M. Whiley, T. P. Sloots, Sequence variation in primer targets affects the accuracy of viral quantitative PCR. *J. Clin. Virol.* **34**, 104–107 (2005).

17. M. D. Pauly, M. C. Procario, A. S. Lauring, A novel twelve class fluctuation test reveals higher than expected mutation rates for influenza A viruses. *Elife*. **6**, e26437 (2017).

18. D. Mercatelli, F. M. Giorgi, Geographic and Genomic Distribution of SARS-CoV-2 Mutations (2020).

19. M. Mahmoud, N. Gobet, D. I. Cruz-Dávalos, N. Mounier, C. Dessimoz, F. J. Sedlazeck, Structural variant calling: the long and the short of it. *Genome Biol.* **20**, 246 (2019).

20. A. D. Davidson, M. K. Williamson, S. Lewis, D. Shoemark, M. W. Carroll, K. Heesom, M. Zambon, J. Ellis, P. A. Lewis, J. A. Hiscox, Others, Characterisation of the transcriptome and proteome of SARS-CoV-2 using direct RNA sequencing and tandem mass spectrometry reveals evidence for a cell passage induced in-frame deletion in the spike glycoprotein that removes the furin-like cleavage site. *BioRxiv* (2020).

21. D. Muth, V. M. Corman, H. Roth, T. Binger, R. Dijkman, L. T. Gottula, F. Gloza-Rausch, A. Balboni, M. Battilani, D. Rihtarič, Others, Attenuation of replication by a 29 nucleotide deletion in SARS-coronavirus acquired during the early stages of human-to-human transmission. *Sci. Rep.* **8**, 1–11 (2018).

22. C. F. Wright, M. J. Morelli, G. Thébaud, N. J. Knowles, P. Herzyk, D. J. Paton, D. T. Haydon, D. P. King, Beyond the consensus: dissecting within-host viral population diversity of foot-and-mouth disease virus by using next-generation genome sequencing. *J. Virol.* **85**, 2266–2275 (2011).

23. J. Quick, N. D. Grubaugh, S. T. Pullan, I. M. Claro, A. D. Smith, K. Gangavarapu, G. Oliveira, R. Robles-Sikisaka, T. F. Rogers, N. A. Beutler, D. R. Burton, L. L. Lewis-Ximenez, J. G. de Jesus, M. Giovanetti, S. C. Hill, A. Black, T. Bedford, M. W. Carroll, M. Nunes, L. C. Alcantara Jr, E. C. Sabino, S. A. Baylis, N. R. Faria, M. Loose, J. T. Simpson, O. G. Pybus, K. G. Andersen, N. J. Loman, Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples. *Nat. Protoc.* **12**, 1261–1276 (2017).

24. M.-H. Lin, D. C. Moses, C.-H. Hsieh, S.-C. Cheng, Y.-H. Chen, C.-Y. Sun, C.-Y. Chou, Disulfiram can inhibit MERS and SARS coronavirus papain-like proteases via different modes. *Antiviral Res.* **150**, 155–163 (2018).

25. R. Rodriguez-Roche, H. Blanc, A. V. Bordería, G. Díaz, R. Henningsson, D. Gonzalez, E. Santana, M. Alvarez, O. Castro, M. Fontes, M. Vignuzzi, M. G. Guzman, Increasing Clinical Severity during a Dengue Virus Type 3 Cuban Epidemic: Deep Sequencing of Evolving Viral Populations. *J. Virol.* **90**, 4320–4333 (2016).

26. P. Parameswaran, P. Charlebois, Y. Tellez, A. Nunez, E. M. Ryan, C. M. Malboeuf, J. Z. Levin, N. J. Lennon, A. Balmaseda, E. Harris, M. R. Henn, Genome-wide patterns of intrahuman dengue virus diversity reveal associations with viral phylogenetic clade and interhost diversity. *J. Virol.* **86**, 8546–8558 (2012).

27. J. Gregori, M. Salicru, E. Domingo, A. Sanchez, J. I. Esteban, F. Rodríguez-Frías, J. Quer, Inference with viral quasispecies diversity indices: clonal and NGS approaches. *Bioinformatics*. **30**, 1104–1111 (2014).

28. K. A. Khan, P. Cheung, Presence of mismatches between diagnostic PCR assays and coronavirus SARS-CoV-2 genome. *Royal Society Open Science*. **7**, 200636 (2020).

29. C. J. Worby, M. Lipsitch, W. P. Hanage, Shared Genomic Variants: Identification of Transmission Routes Using Pathogen Deep-Sequence Data. *Am. J. Epidemiol.* **186**, 1209–1216 (2017).

30. M. P. Zwart, S. F. Elena, Matters of Size: Genetic Bottlenecks in Virus Infection and Their Potential Impact on Evolution. *Annu Rev Virol.* **2**, 161–179 (2015).

31. G. M. Pavlović-Lažetić, N. S. Mitić, M. V. Beljanski, Bioinformatics analysis of SARS coronavirus genome polymorphism. *BMC Bioinformatics*. **5**, 65 (2004).

32. Y. M. O. Alhammad, M. M. Kashipathy, A. Roy, J.-P. Gagne, L. Nonfoux, P. McDonald, P. Gao, K. P. Battaile, D. K. Johnson, G. G. Poirier, Others, The SARS-CoV-2 conserved macrodomain is a highly efficient ADP-ribosylhydrolase. *bioRxiv* (2020).

33. Y. M. Báez-Santos, A. M. Mielech, X. Deng, S. Baker, A. D. Mesecar, Catalytic function and substrate specificity of the papain-like protease domain of nsp3 from the Middle East respiratory syndrome coronavirus. *J. Virol.* **88**, 12511–12527 (2014).

34. N. De Maio, C. Walker, R. Borges, L. Weilguny, G. Slodkowicz, N. Goldman, Issues with SARS-CoV-2 sequencing data. *Virological*. **6**, 80–92 (2012).

35. F. Thorburn, S. Bennett, S. Modha, D. Murdoch, R. Gunson, P. R. Murcia, The use of next generation sequencing in the diagnosis and typing of respiratory infections. *J. Clin. Virol.* **69**, 96–100 (2015).

36. B. Huang, A. Jennison, D. Whiley, J. McMahon, G. Hewitson, R. Graham, A. De Jong, D. Warrilow, Illumina sequencing of clinical samples for virus detection in a public health laboratory. *Scientific Reports*. **9** (2019), , doi:10.1038/s41598-019-41830-w.

37. C. M. Coleman, M. B. Frieman, Emergence of the Middle East respiratory syndrome coronavirus. *PLoS Pathog.* **9** (2013).

38. A. M. Bolger, M. Lohse, B. Usadel, Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics.* **30**, 2114–2120 (2014).

39. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows--Wheeler transform. *Bioinformatics*. **25**, 1754–1760 (2009).

40. H. Li, Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:1303. 3997* (2013).

41. H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, The sequence alignment/map format and SAMtools. *Bioinformatics.* **25**, 2078–2079 (2009).

42. A. Wilm, P. P. K. Aw, D. Bertrand, G. H. T. Yeo, S. H. Ong, C. H. Wong, C. C. Khor, R. Petric, M. L. Hibberd, N. Nagarajan, LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. *Nucleic Acids Res.* **40**, 11189–11201 (2012).

43. 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* . **6**, 80–92 (2012).

44. C. W. Nelson, L. H. Moncla, A. L. Hughes, SNPGenie: estimating evolutionary parameters to detect natural selection using pooled next-generation sequencing data. *Bioinformatics*. **31**, 3709–3711 (2015).

45. X. Chen, O. Schulz-Trieglaff, R. Shaw, B. Barnes, F. Schlesinger, M. Källberg, A. J. Cox, S. Kruglyak, C. T. Saunders, Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics*. **32**, 1220–1222 (2016).

46. D. C. Jeffares, C. Jolly, M. Hoti, D. Speed, L. Shaw, C. Rallis, F. Balloux, C. Dessimoz, J. Bähler, F. J. Sedlazeck, Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nat. Commun.* **8**, 1–11 (2017).

47. A. R. Quinlan, I. M. Hall, BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. **26**, 841–842 (2010).

48. M. Krzywinski, J. Schein, I. Birol, J. Connors, R. Gascoyne, D. Horsman, S. J. Jones, M. A. Marra, Circos: an information aesthetic for comparative genomics. *Genome Res.* **19**, 1639–1645 (2009).

49. T. J. Treangen, B. D. Ondov, S. Koren, A. M. Phillippy, The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. *Genome Biol.* **15**, 524 (2014).

50. A. Stamatakis, RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*. **30**, 1312–1313 (2014).

51. K. Katoh, D. M. Standley, MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).

52. M. D. Lee, *Bioinformatics Tools (bit)* (2018; https://github.com/AstrobioMike/bioinf_tools#citation-info).

53. *Scikit-bio Development Team. Scikit-bio: A Bioinformatics Library for Data Scientists, Students, and Developers* (2020; http://scikit-bio.org).

54. WHO in-house qRT-PCR assays, (available at https://www.who.int/docs/default-source/coronaviruse/whoinhouseassays.pdf?sfvrsn=de3a76aa_2).

55. Artic Network Sequencing Primers. *GitHub*, (available at https://github.com/artic-network/artic-ncov2019).

56. B. Langmead, S. L. Salzberg, Fast gapped-read alignment with Bowtie 2. *Nat. Methods*. **9**, 357–359 (2012).

57. A. S. Leonard, D. B. Weissman, B. Greenbaum, E. Ghedin, K. Koelle, Transmission bottleneck size estimation from pathogen deep-sequencing data, with an application to human influenza A virus. *J. Virol*. **91**, e00171–17 (2017).

Acknowledgments

The authors would like to acknowledge feedback and discussion contributions on the effects of variants on the qRT-PCR detection methods provided by Jamie Purcell. The authors would also like to thank Luay Nakhleh for suggestions specific to comparative genomic analyses of SARS-CoV-1 and MERS-COV. Finally, the authors would also like to thank all members of the COVID-19 International Research Team (www.cov-irt.org) for their helpful feedback during weekly meetings.

Disclaimer

This material has been reviewed by the Walter Reed Army Institute of Research. There is no objection to its presentation and/or publication. The views and conclusions expressed in this article are those of the authors and do not necessarily reflect the official policy or position of the Department of the Army, Department of the Navy, Department of Defense, ODNI, IARPA, ARO, or US Government.

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.

Author contributions

N.S. led the iSNV and SNP analyses, interpreted the results, generated the figures, and wrote the manuscript. M.J. analyzed the impact of polymorphisms on probes and primers, and generated figures.Y.L. analyzed single nucleotide variant data and generated the figures. D.A. analyzed phylogenetic data and generated the figures. M.D.L. analyzed genomic data and generated 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. S.V. edited the manuscript, provided exchange of ideas, and generated figures. C.M. provided the RNA-seq data and contributed to the manuscript. T.J.T. supervised the analyses, 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. A.B. edited the manuscript and provided 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. All co-authors read and edited the manuscript and provided constructive feedback.

Competing interests

Authors declare no competing interests.

Data availability

Variant calling files, raw data and supplementary figures and files are available at 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. Scripts used for probe and primer analysis and visualization are available at: https://github.com/COV-IRT/microbial/tree/master/manuscript_reference

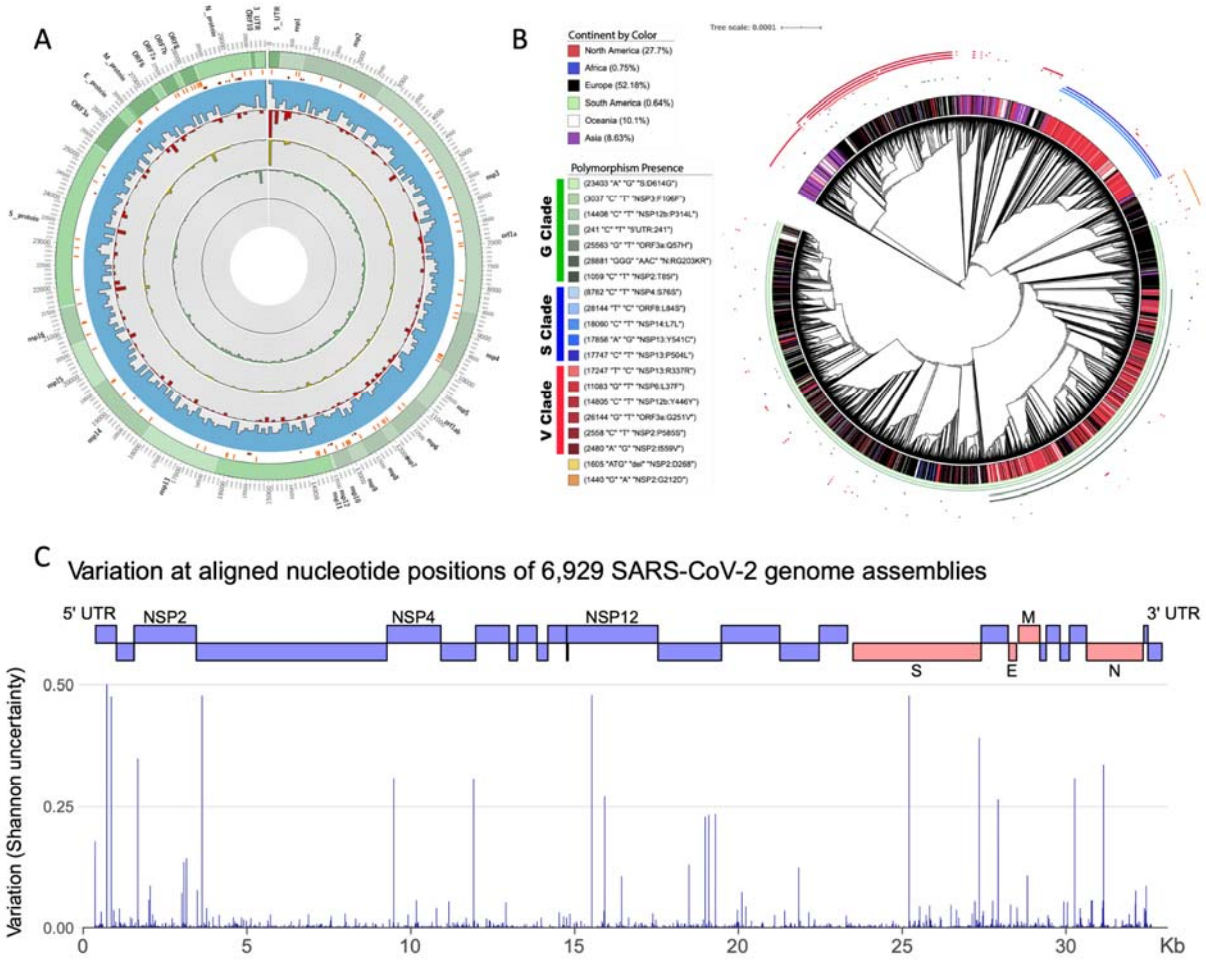

Figure 1: Overview of general diversity of SARS-CoV-2. A. From outer to inner layers: Annotation of SARS-CoV-2 genome (green), transcription-regulating sequences (TRS) (orange), PCR primer designs (dark red), intrahost variant density including iSNVs (blue), deletions start sites (red), duplication start 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 SNPs as described in (*18*). The G clade SNPs 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. **C.** This figure shows the variability of positions in SARS-CoV-2 overlaid with the protein coding regions in the genome.

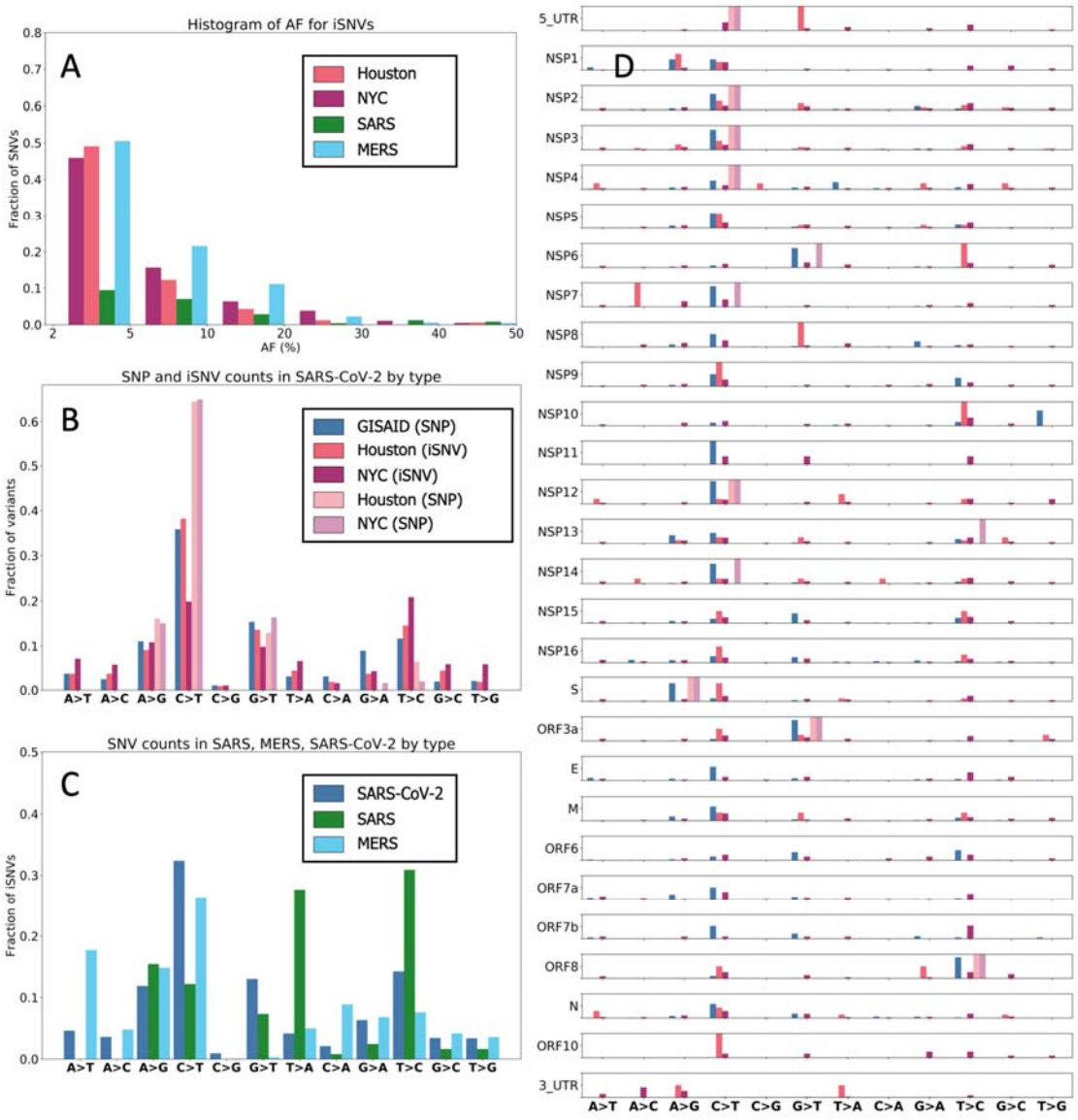

Figure 2: Mutational frequencies of iSNV and SNPs. A. *Distribution of iSNV AF.* We note that the distribution of AF is strictly less than 50% as iSNVs are below consensus-level by definition. **B.** *Mutational spectrum of SARS-CoV-2.* **C.** *Mutational spectra of SARS-CoV-1, SARS-CoV-2, and MERS.* **D.** *Mutational spectrum of SARS-CoV-2 by ORF/NSP.*

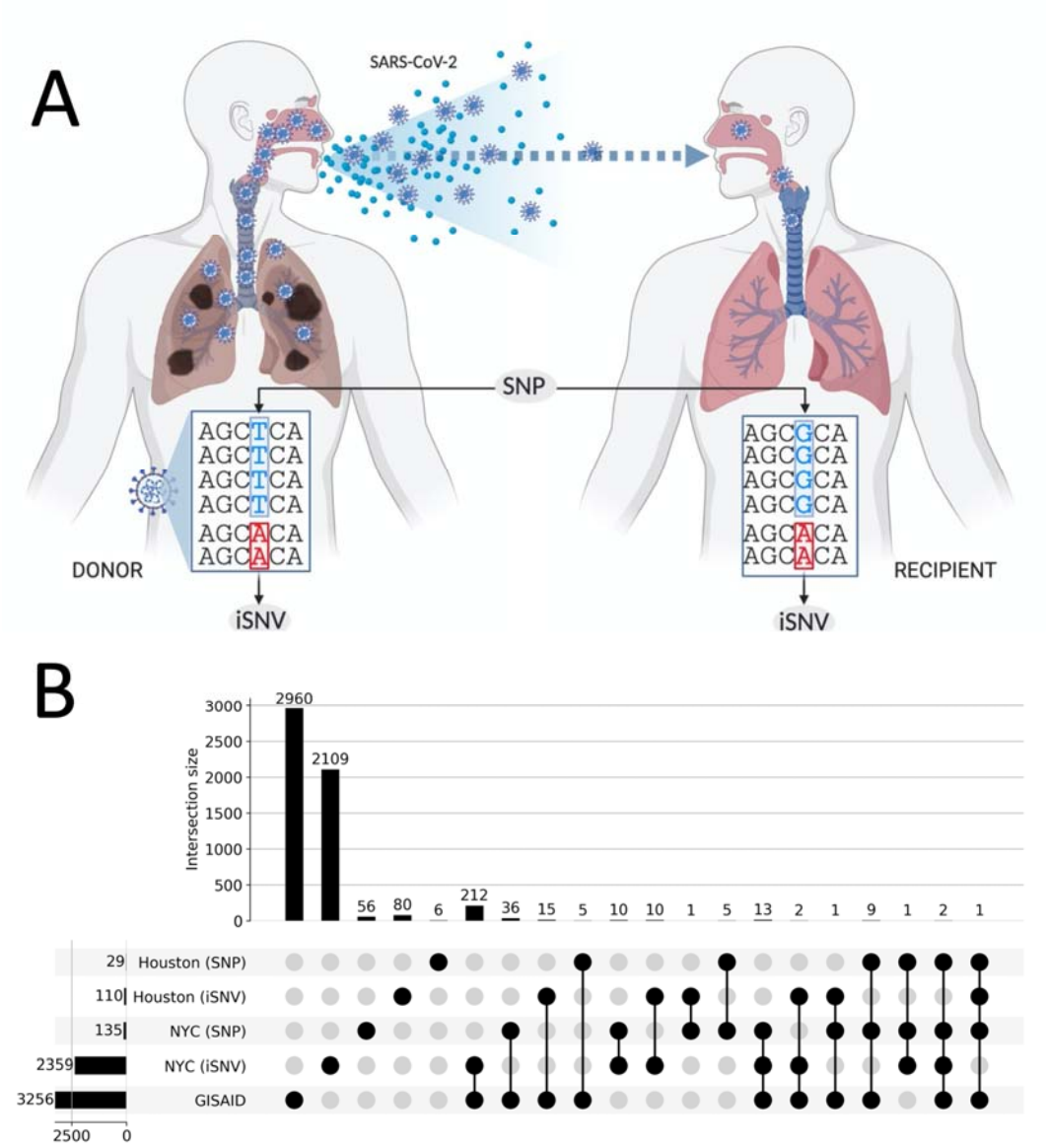

Figure 3: **Shared SNPs and SNVs across datasets. A.** Illustration differentiating what we define as an intrahost SNV (iSNV) and an interhost consensus-level SNP. **B**. This UpSet plot captures the shared single nucleotide variants between iSNVs and consensus-level SNPs. The horizontal bars on the left show the total number of variants in the given category. Vertical bars indicate the size of the intersection between highlighted (with black circles) sets. Every variant contributes to exactly one intersection size to avoid double counting.

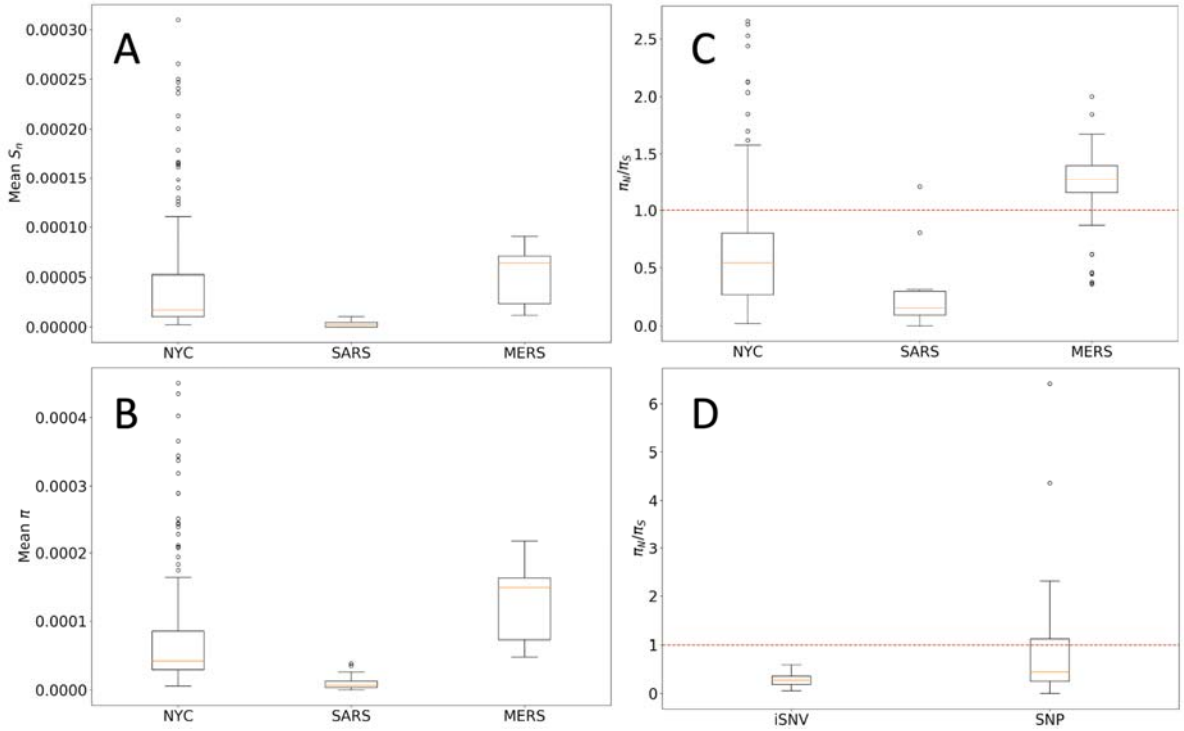

Figure 4: **Complexity and diversity in Coronaviruses. A.** Intrahost complexity of Coronavirus samples. This plot shows the mean S_n complexity of samples for SARS-CoV-2, SARS-CoV-1 and MERS. **B.** Diversity of Coronavirus samples. This plot shows the mean π diversity of samples. **C.** Synonymous vs non-synonymous diversity ratios. **D.** *Syn. vs non-syn. diversity ratios for iSNVs and SNPs in NYC data.*

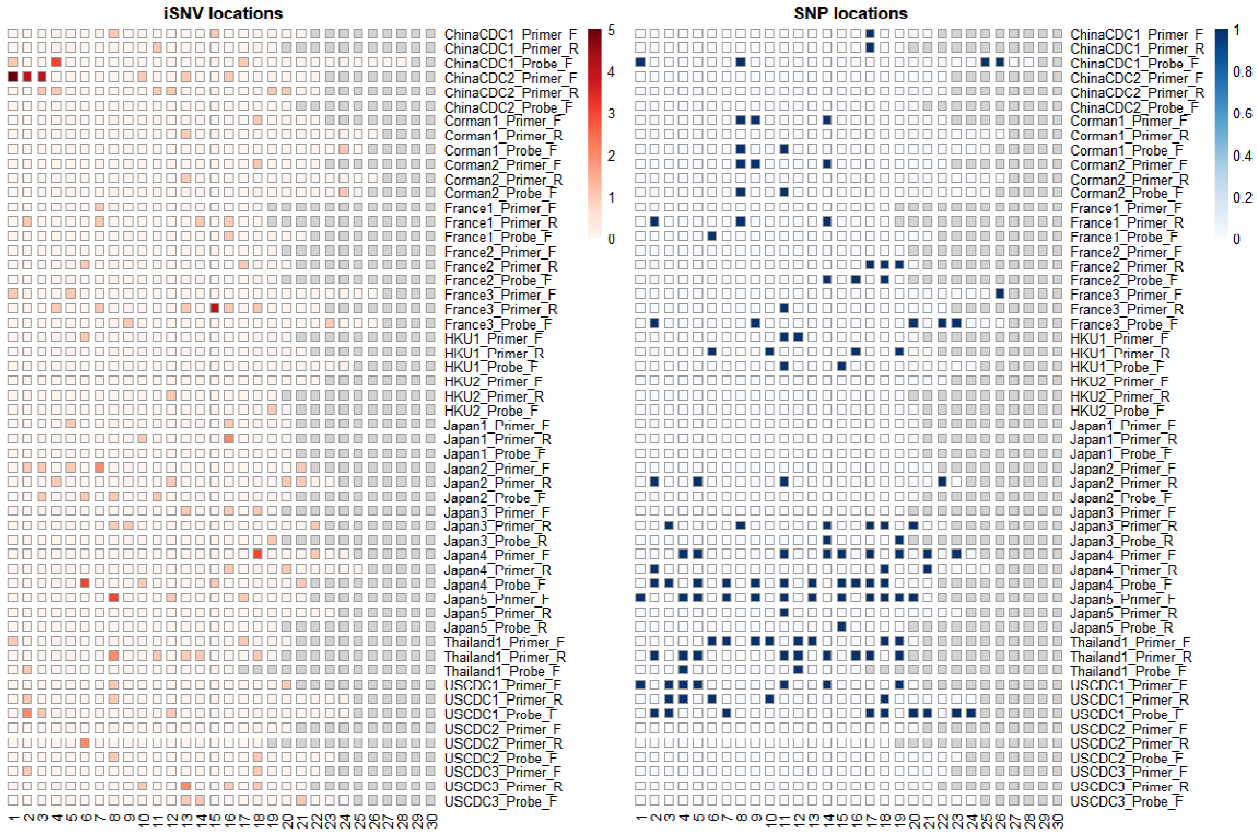

Figure 5: **iSNV and SNP presence on widely-used primers and probes.** This figure shows the locations on WHO probes and primers that contain iSNVs (left) and SNPs (right). Columns correspond to base pair positions within the probe, and the sequences are 3' aligned. Rows corresponding to the oligonucleotide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

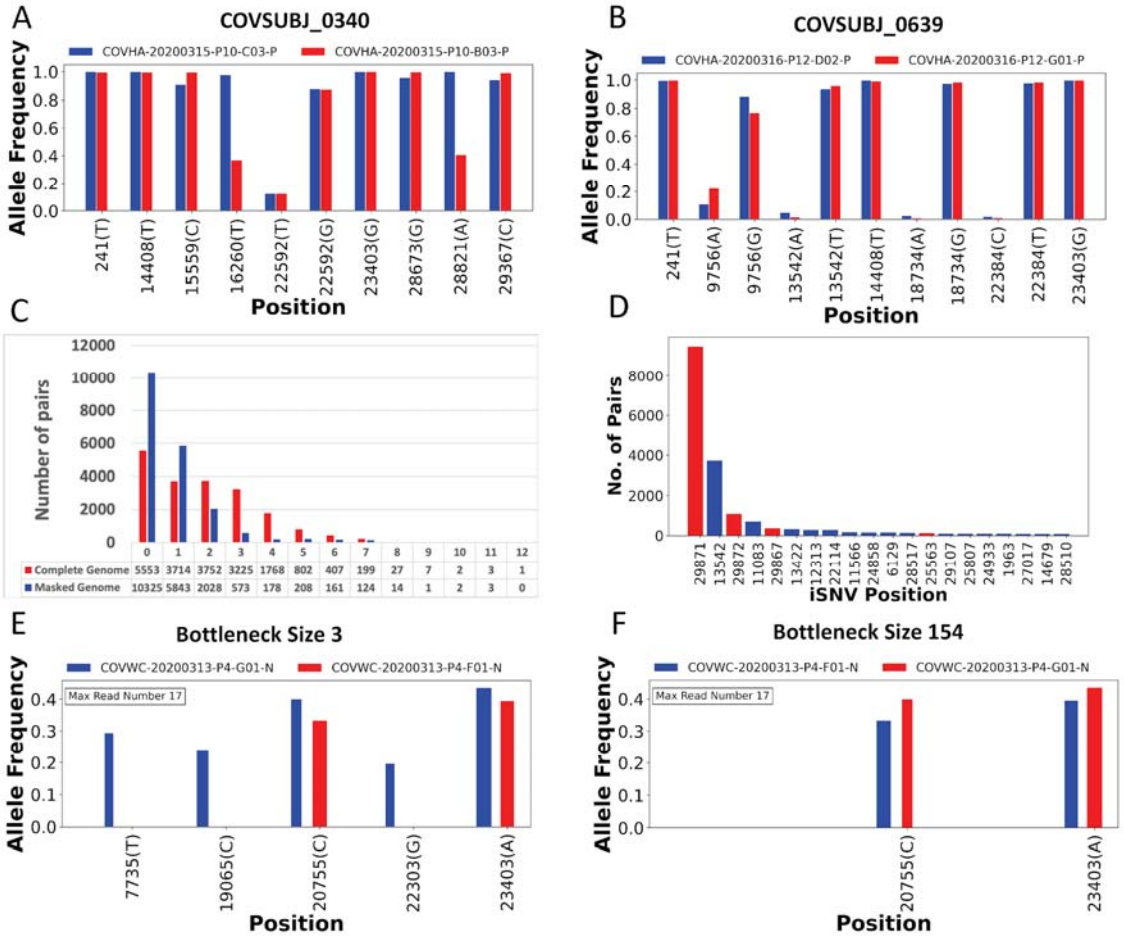

Figure 6: In-depth analysis of shared iSNVs. A. Paired samples from patient COVSUBJ 0340 in NYC. **B.** Paired samples from patient COVSUBJ 0639 in NYC. **C.** The distribution of the number of genomic pairs and their shared iSNVs. **D.** The number of samples with iSNVs at given nucleotide positions. Red color represents positions that are highly homplasic and masked in the bottleneck analysis. **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**.