Diversity and selection of SARS-CoV-2 minority variants in the early New York City outbreak

Short title: Analysis of minority variants in SARS-CoV-2 sequence data

Roder, AE.[1], Khalfan, M.[2], Johnson, KEE.[2], Ruchnewitz, D.[3], Knoll, M.[2], Banakis, S.[1], Wang, W.[1], Samanovic, MI.[4], Mulligan, MJ.[4], Gresham, D.[2], Lässig, M.[3], Łuksza, M.[5], Ghedin, E.[1,2*]

[1]Systems Genomics Section, Laboratory of Parasitic Diseases, NIAID, NIH, Bethesda, MD 20894, USA

[2]Center for Genomics and Systems Biology, Department of Biology, New York University, New York, NY 10003, USA

[3]Institute for Biological Physics, University of Cologne, Cologne, Germany

[4]New York University Vaccine Center, Department of Medicine, New York, NY 10016, USA

[5]Department of Oncological Sciences, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA

*Corresponding author. Email: elodie.ghedin@nih.gov

ABSTRACT

High error rates of viral RNA-dependent RNA polymerases lead to diverse intra-host viral populations during infection. Errors made during replication that are not strongly deleterious to the virus can lead to the generation of minority variants. Here we analyzed minority variants within the SARS-CoV-2 data in 12 samples from the early outbreak in New York City, using replicate sequencing for reliable identification. While most minority variants were unique to a single sample, we found several instances of shared variants. We provide evidence that some higher-frequency minority variants may be transmitted between patients or across short transmission chains, while other lower-frequency, more widely shared variants arise independently. Further, our data indicate that even with a small transmission bottleneck, the heterogeneity of intra-host viral populations is enhanced by minority variants present in transmission samples. Our data suggest that analysis of shared minority variants could help identify regions of the SARS-CoV-2 genome that are under increased selective pressure, as well as inform transmission chains and give insight into variant strain emergence.

IMPORTANCE

When viruses replicate inside a host, the virus replication machinery makes mistakes. Over time, these mistakes create mutations that result in a diverse population of viruses inside the host. Mutations that are neither lethal to the virus, nor strongly beneficial, can lead to minority variants. In this study, we analyzed the minority variants in SARS-CoV-2 patient samples from New York City during the early outbreak. We found common minority variants between samples that were closely related and showed that these minority variants may be transmitted from one patient to another. We show that in general, transmission events between individuals likely contain genetically diverse viral particles, and we find signatures of selection governing intra-host evolution. We conclude that the analysis of shared minority variants can help to identify transmission events and give insight into emergence of new viral variants.

INTRODUCTION

The circulation of a novel coronavirus was reported in late 2019 out of Wuhan Province, China (1-3). Originally named nCoV-2019, the virus was officially named SARS-CoV-2 in early February 2020 (4). The World Health Organization declared SARS-CoV-2 a global pandemic in March 2020 and as of April 14, 2021, the virus had infected close to 142 million people and caused more than 3 million deaths worldwide (5).

Sequencing of SARS-CoV-2 from infected patients has contributed to our knowledge of the viral origin, the biology of infection, and viral transmission events as well as given insight into the spread of the virus across the world. Despite efforts to prevent introductions of the virus to the United States from areas of the world with active outbreaks, the first positive case of SARS-CoV-2 was reported on January 19, 2020 from Washington state (6). Since this first reported introduction, new outbreaks have occurred in all major US cities and areas (7). Sequencing of virus from infected patients in these cities has helped to determine both the number and origin of these introduction events (8, 9). Viral sequencing has also identified key amino acid changes that differentiate clades of the virus in circulation (10). Identification of these clades and the associated viral consensus changes aids in tracking spread of the virus. However, little has been done to examine potential early detection of emerging variants before they become fixed in the population.

Due to the error-prone nature of viral polymerases, as well as the speed of viral replication, errors are introduced into viral genomes during replication (11). These errors can range from lethal (killing the virus) to beneficial, enhancing the viral lifecycle. Coronavirus polymerases are unique among RNA viruses in that they possess a level of proofreading capability (12, 13). This function results in a mutation rate that is significantly lower than other RNA viruses such as rhinovirus or influenza A virus (14-16). Nonetheless, mutations are still introduced during viral replication. Mutations can lead to changes in the consensus sequence; these specific sets of mutations separate the circulating virus population into clades. Mutations in the virus genomes that are not the majority within an infected host (present at lower than 50% frequency) represent minority variants. Identification of these minority variants within the sequencing data can highlight regions of the genome under positive selection or regions with increased mutational tolerance, detect subtle virus population shifts within the infected host, and identify mutations before consensus changes occur (17). These variants can also shed light on tropism, and shared minority variants between samples can show patterns of viral evolution (18). The presence of these variants may have long term implications for vaccine, monoclonal antibody, and drug development.

Confident prediction of minority variants requires significant sequence read coverage and the frequency at which identified variants are considered valid is debated. Numerous software packages exist to identify single nucleotide variants (SNVs) within sequence data, but both the approaches and results can differ significantly.

With the goal of identifying and understanding the scope of minority variants during SARS-CoV-2 infection, we used a small cohort of 12 samples from 11 individuals that were infected with SARS-CoV-2 early in the pandemic during the New York City outbreak. We first used simulated SARS-CoV-2 data to test the ability of different variant-calling software packages to accurately identify minority variants in SARS-CoV-2 sequence data. We then used these methods to analyze the minority variants present in our cohort. We found a number of variants in common between closely related samples that suggest the possibility of variant sharing through short transmission chains. Analyzing the frequency distributions within hosts suggests that even with a small transmission bottleneck, transmitted populations are likely heterogeneous. Furthermore, we find signatures of selection even within the high-frequency variants relevant for transmission. This highlights the importance of accurately identifying minority variants in SARS-CoV-2 sequence data as a tool for uncovering areas of selection within the genome and for tracking spread and emergence of novel variants.

RESULTS

Strict cutoffs are necessary for accurate identification of minority variants in SARS-CoV-2 sequence data

Accurate identification of minority variants, even with stringent coverage and frequency thresholds, is complicated by the fact that both PCR amplification of the genome and sequencing can introduce errors. Minority variants can be difficult to separate from these errors. Many methods exist for identifying minority variants within deep sequence data, however, they vary in both their bioinformatic and statistical approaches. With this in mind, we tested the ability of five popular variant-calling software packages (iVar, VarScan, HaplotypeCaller, Mutect2, and freebayes) and one in-house pipeline (*timo*) to accurately identify minority variants at both set and random allele frequencies and across a range of down-sampled coverages (19-24). We used the NEAT software package to simulate SNVs in the SARS-CoV-2 data, incorporating variants through both a mutation model based on publicly available SARS-CoV-2 sequence data, as well as a sequencing error model based on reads specific to the sequencing platform used. We initially simulated data at a coverage of 100,000x and accounted for variable read depths through random down-sampling (25). We then aligned reads and called variants using the six tools. At approximately 200X coverage, iVar and mutect2 were accurate, but too conservative in their calls, sacrificing recall for precision. All tools outperformed freebayes in calling true positives, which identified the most variants, but this included a high number of false positives. VarScan, HaplotypeCaller and *timo* all performed well, though VarScan had slightly lower precision than the other two tools (**Fig. 1A-B**). Looking at performance across coverages at a set allele frequency of 0.02, we determined that both HaplotypeCaller and our in-house caller, *timo*, performed well for capturing low frequency alleles at relatively low read depths (>0.02, 200X) (**Fig. 1C-D**). Using simulated data with SNVs at random allele frequencies, we found that at approximately 200X, *timo* accurately identified all variants above a frequency cutoff of 2% without calling any false positives (**Fig. 1E**). Based on our testing, we chose to use *timo* with a coverage cutoff of 200x and an allele frequency of 0.02 for the most accurate identification of minority variants within our clinical samples.

Most identified minority variants are unique to a single sample

To investigate the minority variants in real SARS-CoV-2 data, we used a small cohort of 12 samples from the early outbreak in New York City collected and processed at NYU Langone Health and NYU Grossman School of Medicine. Nasopharyngeal swabs (NS) were collected between March 6, 2020 to April 9, 2020 from 11 individuals between the ages of two weeks and 60 years (five females, six males; one individual had samples collected at two time points). Specimen collection occurred on various days post onset of illness (DPO). The samples represented a variety of viral loads, ranging from 10,400 viral RNA copies/ml to 416,800 copies/ml (**Supplementary Table 1**). We achieved more than 88% coverage of the genome at 5X for all 12 of the NS samples.

To determine the major clades represented within our samples, we mapped them against a global tree using 10,932 global isolates. We characterized the main genetic clades by identifying non-synonymous amino acid mutations that originate in prevalent viral population subtrees and used the Wuhan/Hu-1/2019 strain to root the tree. The New York isolates mapped to two major clades. Ten of the sequences belonged to clade 20C, defined by mutations S:D614G, ORF1b:P314L, ORF3a:Q57H, and ORF1a:T265I, while two sequences, from the two samples from the same patient (NYU-VC-009), mapped to clade 20B, defined by the mutations S:D614G, ORF1b:P314L, N:R203K, N:G204R, and ORF14:G50N (**Fig. 2A-B**). These two clades were circulating in New York City during the time period when the samples were collected. The first clade was the dominant clade in March and April, constituting 80-90% of the viral population. The second clade was circulating at a frequency of 5-10% at that time, showing that our data samples are a good representation of the genetic diversity of the virus during the time period when they were collected.

We analyzed the full set of mutations in our isolates and identified 20 unique consensus changes across the 12 samples, including changes in six of the 10 coding regions, in the 5' UTR and in one intergenic region. Samples had between five and 10 consensus changes, an average of approximately eight per sample as compared to the Wuhan/Hu-1 reference strain. As expected, due to the length of the gene, ORF1a contained the most changes with seven unique changes. There were three consensus changes found in all 12 samples, including 5'UTR:C241U, ORF1a:C3037U, and S:A23403G (**Fig. 2C**). The S:A23403G (aa S:D614G) mutation is a defining mutation associated with European derived strains of the virus and found to be associated with increased transmission (26, 27). Of the 20 unique consensus changes, 13 of them represented non-synonymous changes while seven were synonymous or in non-coding regions. The non-synonymous changes were also found more frequently in multiple samples, representing 62 of the 95 total changes in the data. Of these 95 total changes, the overwhelming majority were transitions with very few transversions. C to U transitions were the most frequent, followed by G to A and A to G changes (**Fig. 2D**). As expected, none of the identified consensus changes were unique to our samples and can be found in many publicly available sequences within the USA East Coast clade.

To identify high confidence minority variants within this data set, we sequenced each sample in duplicate, when starting material allowed (nine of 12 samples). We used a low frequency threshold (0.005) to perform an initial filtering of the minority variants called by *timo* and compared the minority variants across the replicate sequences. The large majority of minority variants were not reproducible, indicating that they may have been introduced during the amplification or sequencing processes (**Fig. 3A**). Importantly, we did not find an obvious correlation between viral load and the number of reproducible minority variants in this sample set ($r^2 = 0.271$) (**Fig. 3B-C**). Based on these observations, we filtered our list of variants for only those that existed in both replicates in locations with coverage greater than 200X and an average allele frequency above 0.02. For samples that were only sequenced once due to limited specimen availability, we filtered the minority variants to include only those that were present above our cutoffs and existed in another sample. We used this final list of high confidence minority variants for our analyses.

Using these cutoffs, we identified 54 minority variants across the 12 NS samples, 29 of which were unique to the samples in which they were detected. High confidence minority variants were detected in eight of the 10 gene coding regions, as well as in the 5' UTR. The highest number of variants were in ORF1a (**Fig. 4A**). As with the identified consensus changes, there were more transitions than transversions with C to U transitions accounting for the overwhelming majority of the changes (**Fig. 4B**). In contrast to the consensus changes, the number of variants was more variable between samples, ranging from as few as one to as many as 13 in one sample **(Fig. 4C)**. Of the 38 different variants identified across the samples, approximately 20% were found in more than one sample. Close to 50% of the shared variants were present in pairs of samples while the others were shared between 3-5 samples. Samples 022 and 023 shared the highest number of variants (**Fig. 4C**). Thirty-five of the minority variants led to nonsynonymous changes, compared to 12 synonymous changes; both synonymous and nonsynonymous changes were represented within the shared variants (**Fig. 4D**). There was only one instance of a minority variant that was present at the same location as a consensus change within our data, in ORF1a at amino acid position 1429 (**Fig. 4E**). Ultimately, we found that most minority variants were unique to a single sample, reinforcing the randomness of errors made by the viral RdRp which result in minority variants.

Transmission of minor variants between hosts.

There were many instances of minority variants that were common to two or more patients within our sample set (**Fig. 4C**). In order to better understand the set of shared variants, we expanded our set of variants to include those present in both replicates with an allele frequency greater than 0.005 and coverage greater than 200X (**Fig. 5A**). One pair of samples, NYU-VC-022 and NYU-VC-023 was of particular interest for these analyses given their proximity on the consensus tree, and the fact that they shared the most minority variants between them (**Fig. 4C, 5A**). These variant statistics differed strongly from the remainder of the samples, signaling a possible transmission event, either between these samples or across a short intermediate transmission chain. To investigate this possibility, we recorded the cumulative distribution of Hamming distances between samples, d, as recorded on the consensus tree, for all minority variants shared between exactly two hosts (doublets). We then compared this distribution with a null distribution, obtained from random pairs of variants across all of the samples. We found that the majority of the doublet variants, but not those in the random pairs, were found in samples where $d = 0$, suggesting that these pairs of variants are likely the result of transmission, rather than of independent *de novo* mutations (**Fig. 5B**). To show that these variants were enriched specifically in samples NYU-VC-022 and NYU-VC-023, we determined the fraction of doublet variants compared to the sum of both the unique variants (singlets) and the doublets for all samples with replicate sequencing (this includes sample 022, but not 023). NYU-VC-022 had a strongly enhanced fraction of doublet variants compared with the rest of the samples in the data set (**Fig. 5C**). Together, these statistics suggest a short transmission chain involving NYU-VC-022 and NYU-VC-023 and indicate that transmission events contain a genetically diverse mix of virus particles.

<u>Nonsynonymous mutations are under negative intra-host selection</u>

Upon entry into a new host, the viral population grows initially in an exponential way. As such, the frequency of a mutation within the population is related to its origination time: a few early mutations of larger frequency are followed by many later mutations of small frequency. This feature, which is well-known in the context of Luria-Delbrück fluctuation assays, can be made quantitative: a mutation originating at time t after the start of growth has an initial frequency $x = \exp(-\lambda t)$, where λ is the growth rate of the viral population. If the mutation is nearly neutral, this frequency will stay approximately constant during the subsequent growth process. These dynamics generate a mutation frequency spectrum described by the Luria-Delbrück distribution (28), which is characterized by a cumulative distribution function of the form $\Phi(x) = \frac{c}{x^{\alpha}}$. This distribution gives the expected number of minority variants with frequency $> x$; the decay exponent distinguishes neutral variants ($\alpha = 1$) and negatively selected variants ($\alpha > 1$). We used this distribution to analyze variants with allele frequencies between 0.02 and 0.5. Mutations with these frequencies are expected to arise predominantly in the first intra-cellular replication cycle, which is firmly in the exponential growth phase. We analyzed the empirical cumulative frequency distributions for synonymous and nonsynonymous minority variants, averaged over the samples with replicate sequencing. The distribution of synonymous variants is consistent with the neutral Luria-Delbrück form ($\alpha = 1$). However, non-synonymous variants showed a somewhat faster decay ($\alpha \approx 1.4$), indicating weak negative selection reducing the fraction of high-frequency variants (**Fig. 6A**). We note that, given the limited frequency range of reliable mutant calling, substantial statistical errors of the inferred decay exponents are to be expected. Negative selection on non-synonymous variants demonstrates that random mutations will often result in a loss of fitness.

Transmission droplets are likely heterogeneous

To expand upon the hypothesis that minority variants could be shared between samples by transmission, we used only reproducible variants (those observed in both sequencing replicates) present at an allele frequency above 0.005 and coverage above 200X (**Fig 3A, 5A**) and computed the probability that the transmitted viral population is heterogeneous in sequence. To do this, we evaluated the cumulative mutant weight $Y(x)$, which is defined as the sum of the expected frequencies of mutant clades with frequency $> x$ that arise on the host-specific ancestral (wild-type) background. By construction, this weight discounts all double mutants that arise on the background of an earlier mutant. In **Fig. 6B**, we show the empirical mutant weight functions for synonymous, nonsynonymous, and all mutations, which are computed from the frequency counts using a random-genealogy assumption and averaged over all samples with replicate sequencing (Methods). From this, we infer a substantial weight of all minor variants even if we restrict the frequency range to above the cutoff for variant calling, $Y(x_0) = 0.30$ for $x_0 = 0.5 \times 10^{-2}$. The complement of this weight is an upper bound for the frequency of the ancestral genotype in the evolved viral population, $X_{\rm wt} < 1 - Y(x_0) = 0.70$. Similarly, the weight of non-synonymous mutations, $Y_n(x_0) = 0.18$, determines an upper bound for the frequency of the ancestral amino acid sequence in the evolved viral population, $X_a < 1 - Y_n(x_0) = 0.82$. These ancestral frequencies, in turn, determine the probabilities that a transmission event of n virions is monomorphic in the wild type nucleotide sequence, $p(n) = X_{\rm wt}^n$, or in amino acid sequence, $p_a(n) = X_a^n$. From the inferred weights, we can conclude that even transmission events of moderate virion count ($n \sim 10$) are likely to be polymorphic in nucleotide sequence ($p(10) < 0.03$) and in amino acid sequence ($p_a(10) < 0.14$) (**Fig. 6B**). These analyses show that most transmission droplets of this size would transport minor variants between hosts.

DISCUSSION

It has long been understood that intra-host viral populations are heterogeneous in nature (29-32). We were particularly interested in the level of viral diversity early during the SARS-CoV-2 pandemic and whether or not minority variant analysis could be used to inform transmission events. Here we performed an in-depth analysis of minority variants within a small set of SARS-CoV-2 samples from the early virus outbreak in New York City. Confident identification of minority variants is complicated by errors introduced during amplification and sequencing and therefore we first determined the best approach for stringent calling of minority variants. We tested six minority variant callers using simulated SARS-CoV-2 deep sequence data and our results highlight the need for stringent coverage and allele frequency cutoffs in minority variant analyses. Using our determined cutoffs, we found a number of shared minority variants between samples and provide evidence that some variants may be passed during transmission events. Together, our results lay the groundwork for future studies of minority variants in SARS-CoV-2 infections.

Viral replication is inherently error-prone, and these replication errors result in a diverse population of viruses within a single host (29, 33, 34). Viral sequencing easily allows for the determination of the consensus sequence of the majority population within a host. However, capturing and accurately identifying the other viral mutations that do not constitute the majority is more difficult. The variant callers that we tested take diverse approaches, such as haplotype-based methods (freebayes and haplotype caller), or alignment to the reference (VarScan, iVar and timo). We found that in each category there were tools that performed well and tools that performed more poorly. Freebayes was the least precise and called the highest number of false positives. The false positive rate of both iVar and mutect2 was 0, however these callers were relatively conservative, missing a number of true positives. Many studies use iVar for variant calling in viral genomes and our data suggest that this may result in true variants being overlooked. Haplotype caller, *timo*, and VarScan all performed nearly perfectly, missing only variants that existed at very low frequencies (< 0.01) or at low coverage (< 100X). It is clear from these data that variant callers and cutoffs should be carefully selected in order to increase confidence in the identified variants in studies such as this one. However, even with the best variant calling software, preparation and processing steps necessary for sequencing viral genomes generate layers of error that can make low frequency minority variants virtually indistinguishable from processing errors. In replicate sequencing, with a relatively low frequency cutoff of 0.005, we found that consistently less than 10% of identified minority variants were reproducible, regardless of viral load. Our data suggest that these process errors greatly interfere with confident minority variant prediction and replicate sequencing as well as very stringent cutoffs are thus essential for the identification of variants.

The majority of changes we identified in our data, both at the consensus level and in minority variants, were C to U transitions, consistent with published reports (35-38). However, we found few unique consensus changes (only six of the 20 identified mutations) while the majority of minority variants were unique to a single sample (44 of the 54 minority variants). Similarly, the number of consensus changes across samples was relatively consistent, but the number of minority variants differed more significantly. We found no correlation between the number of minority variants and the viral load within our data set, despite studies that suggest that low viral load increases false positive minority variants, likely due to replicated sequencing of our samples (39).

We also saw only one instance of a minority variant in the same genomic location as a consensus change in our data set — in ORF1a at aa position 1429. We initially expected to see this pattern more frequently as all mutations in the consensus tree must have been a minority in an intra-host viral population at some point. The fact that within our data, we see this pattern infrequently could suggest that selected mutations move from minority to majority very quickly and therefore capturing them as minority variants is less likely; or could suggest the opposite, that it takes a very long time for this change to occur and thus, capturing it within a small data set would be rare. This will be an interesting avenue to explore in future studies.

One frequently debated topic is the possibility that minority variants could be passed between individuals during a transmission event. This possibility depends on a number of factors including the frequency of the minority variant and the size of the transmission bottleneck. For SARS-CoV-2, the size of the bottleneck has been reported to be as few as one to as many as one thousand (40-43). It is likely that transmission between individuals involves multiple transmission events over the course of an interaction, rather than just one. Multiple transmission events would increase the number of viral particles passed between individuals. Our analyses suggest that transmission events are unlikely to be homogeneous and that most virions in the host differ by acquired mutations from the founder genome that was transmitted. This notion is supported by studies that have shown evidence of mixed SARS-CoV-2 infections (18, 44). Moreover, we find that intra-host selection shapes the distribution of minor variants in the high-frequency regime, which includes the variants relevant for transmission. Our current analysis covers broad negative selection on non-synonymous mutations. Future, more densely sampled data may also permit the identification of positively selected minor variants.

Further supporting transmission of minority variants, we identified several instances of shared minority variants within our sample set. Some variants were shared between two individuals, while other variants were widely shared between many individuals. Two of our samples, NYU-VC-022 and NYU-VC-023, contained many more uniquely shared variants (doublets) than any other set of samples, and these samples were also the closest on the consensus tree. To rule out the possibility of contamination, we re-extracted, amplified and sequenced these samples many months after initial sequencing, and confirmed the presence of the high confidence variants. These data contribute to an argument for transmission of minority variants; however, these conclusions are limited by the sample size and by the lack of metadata supporting the potential for transmission and we would caution against using minority variants alone to determine transmission between individuals. Future studies with large data sets and more in-depth metadata from contact tracing would help to further these conclusions. We also found variants that are shared between many samples in our data set. These variants are shared between samples that are dispersed across the consensus tree and therefore are unlikely to be shared through transmission events. Instead, these variants are likely the product of *de novo* mutations, perhaps in regions of the genome that have an increased tolerance for mutation. In our analyses, we did not find a significant relationship between sites with widely shared minority variants and frequently mutated positions on the tree, though a large sample set would be necessary to explore this further. This phenomenon has been previously suggested for widely shared variants in SARS-CoV-2 infection, and the proposal of mutational hotspots within RNA virus genomes is also substantiated (36, 39, 45).

Taken together, our findings establish a framework for the study of minority variants within SARS-CoV-2 sequence data and provide evidence for heterogeneous transmission of SARS-CoV-2 that likely contributes to the sharing of minority variants. These findings have long term implications for vaccine and drug development and set groundwork for the exciting potential of detection of minority variants within the population before their emergence as consensus nucleotides.

MATERIALS AND METHODS

RNA extraction and SARS-CoV-2 quantification

Total RNA was extracted from 300µL of nasopharyngeal swab (NS) or plasma samples collected at the NYU Langone Health between March 6, 2020 and April 9, 2020 (**Supplementary Table 1**). Samples were collected and stored in viral transport media (BD, 220220) and RNA was extracted using the QIAamp® Viral RNA Mini Kit (Qiagen, 52904) according to the manufacturer's instructions. Quantitative real-time PCR was performed according to the "CDC Real-Time RT-PCR Panel for Detection 2019-Novel Coronavirus" protocol with three SARS-CoV-2 virus-specific primers/probe sets (N1, N2, N3, (Integrated DNA Technologies, cat. 10006606)) to test for the presence of SARS-CoV-2 (46). A standard curve was generated using the CDC Positive Template Control (PTC) RNA and was used to calculate viral copies/mL. In total, 12 NS samples were used for genomic analysis.

Reverse transcription and generation of amplicons

Amplification of the viral genome was performed using a modified version of the ARTIC consortium protocol for nCoV-2019 sequencing (https://artic.network/ncov-2019) and the methods described in Gonzalez-Reiche *et al.* (8). Briefly, RNA extracted from patient samples was reverse transcribed and subsequently amplified using the Superscript III one-step RT-PCR system with Platinum Taq DNA Polymerase (Thermo-Fisher, 12574018) using nested cycling conditions. Cycling conditions were as follows: 45°C for 60' for RT, 94°C for 2', followed by 12 cycles of 94°C for 15 s, 55°C for 30 s and 68°C for 8 min; followed by 35 cycles of 94°C for 15s, 55°C for 30 s and 68°C for 2 min 30 s; 68°C for 5' and an 8°C hold. Each sample was processed with two separate pools of primers, pool A and pool B, resulting in alternating and overlapping amplicons that cover the SARS-CoV-2 genome (**Table 2**). Gel electrophoresis was used to confirm amplification of a 2 kb product.

Table 2: Oligonucleotides used for SARS-CoV-2 genome amplification

	Primer Set A	
	Forward Primer (5'-3')	**Reverse Primer (5'-3')**
A1	CCAGGTAACAAACCAACCAACTTT	GCCACTGCGAAGTCAACTGAACA
A2	TGGAACTTACACCAGTTGTTCAGAC	AGCATCTTGTAGAGCAGGTGGA
A3	AAACCGTGTTTGTACTAATTATATGCCTT	TCACGAGTGACACCACCATCAA
A4	ACGGTCTTTGGCTTGATGACGT	TTTGACCGTGATGCAGCCATGC
A5	GCTAAATTCCTAAAAACTAATTGTTGTCGC	GCGGACATACTTATCGGCAATTTTGTTA
A6	TGTTGGTGATTATTTTGTGCTGACAT	CGCTTAACAAAGCACTCGTGGA
A7	ACCCAGGAGTCAAATGGAAATTGA	CCTGAGGGAGATCACGCACTAA
A8	ACCCATTGGTGCAGGTATATGC	TGCAGTAGCGCGAACAAAATCT
A9	TGTGGCTCAGCTACTTCATTGC	GGCCCAGTTCCTAGGTAGTAGAAAT
	Primer Set B	
	Forward Primer (5'-3')	**Reverse Primer (5'-3')**
B1	CTGGAATATTGGTGAACA	GCCGACAACATGAAGACAGTGT
B2	GGTCCAACTTATTTGGATGGAGCTGAT	AAAACACNTAAAGCAGCGGTTGA
B3	GTCACAACATTGCTTTGATATGGAACG	TGGGCCTCATAGCACATTGGTA
B4	ATTGTGGGCTCAATGTGTCCAG	AGCATAGACGAGGTCTGCCATT
B5	CCTAAATGTGATAGAGCCATGCCT	TGCGAGCAGAAGGGTAGTAGAG
B6	CTGAGCGCACCTGTTGTCTATG	TGAACCTGTTTGCGCATCTGTT
B7	TTCGAAGACCCAGTCCCTACTT	AGTGACACTTGCAGATGCTGGCT
B8	GCTGTAGTTGTCTCAAGGGCTGTTGTT	GCTCCCAATTTGTAATAAGAAAGC
B9	ACTTGTCACGCCTAAACGAACA	TAGGCAGCTCTCCCTAGCATTG

Library Preparation and Sequencing platforms

All libraries were prepared using the Nextera XT library preparation kit (Nextera), scaled down to 0.25x of the manufacturer's instructions. Briefly, PCR products were normalized to 0.2ng/uL. DNA was then fragmented, tagged, amplified and barcoded (Illumina Nextera DNA dual indexes), cleaned with a 0.9x bead cleanup and pooled at equal molarity. A 0.7x bead cleanup was performed on the final pool and libraries were sequenced on either the Illumina MiSeq or the Illumnia NextSeq using either the 2x150 bp or 2x300 bp paired end protocol.

Generation of simulated data and testing of minority variant callers

Reads were simulated using the NEAT v2.0 next generation read simulator (25). First, a mutation model was built using genMutModel.py (NEAT) providing the VCF obtained from https://bigd.big.ac.cn/ncov/variation/statistics?lang=en (downloaded April 2020) and the NCBI SARS-CoV-2 reference genome (NC_045512.2) as input. An error model was built using genSeqErrorModel.py (NEAT) providing paired end reads from a high coverage library within our data set. GC and fragment length models were built using computeGC.py and computeFraglen.py respectively (NEAT) using the NYU-VC-003 bam file as input and SARS-CoV-2 reference for computeGC.py. These four models were then provided to NEAT genReads.py along with the reference fasta and a mutation rate of 0.0045 (0.45%) to produce a "golden VCF" file containing ~160 SNPs. Several copies of this golden VCF were made, each with the same variants but with differing allele frequencies: 0.01, 0.02, 0.03, 0.05, 0.1, 0.2, 0.25, 0.5 (one fixed AF per file), and one VCF was made with random allele frequencies where any given variant was 3x more likely to have an AF < 0.5 than > 0.5. Each VCF was then provided as input to NEAT genReads.py along with the reference, error model, fragment length model, GC model, and the following params: ploidy = 100, read length = 150, coverage = 100,000, and mutation rate = 0 in order to use only variants in the VCF to simulate paired end fastq libraries with SNPs from the original golden VCF file inserted at varying allele frequencies as encoded in the individual VCF files.

Each set of simulated paired end fastq libraries was then down-sampled at the following fractions: 0.1 (~10000X), 0.01 (~1000X), 0.001 (~100X), 0.0001 (~10X), 0.00001 (~1X), 0.002 ~(200X), 0.003 (~300X), 0.005 (~500X) using seqtk v1.2-r94 (https://github.com/lh3/seqtk) and a different seed for each down-sampling process to create different fastq files with varying levels of coverage from the original data.

Each pair of downsampled fastq files, along with the original, was quality and adapter trimmed using trimmomatic v0.36 with the following parameters: ILLUMINACLIP:adapters.fa:2:30:10:8:true LEADING:20 TRAILING:20 SLIDINGWINDOW:4:20 MINLEN:20 (47). The trimmed reads were aligned to the Wuhan-Hu-1 SARS-CoV-2 reference genome (NC_045512.2) using BWA mem v0.7.17 with the -K parameter set to 100000000 for reproducibility and -Y to use soft clipping for supplementary alignments (48). Duplicates were marked using GATK MarkDuplicatesSpark v4.1.7.0 (24).

Variants were called using six separate methods:

1. GATK Mutect2 v4.1.7.0 with default parameters. Variants were then filtered using GATK FilterMutectCalls v4.1.7.0 (19).
2. Freebayes v1.1.0-54-g49413aa with ploidy set to 1 and a minimum allele frequency (-F) set to 0.01 (note: freebayes default ploidy is 2, -F is 0.2) (20).
3. Our in-house pipeline, *timo*, with the minor variant frequency cutoff (-c) option set to 0.001, and the coverage cutoff (-C) option set to 1.
4. VarScan v2.4.2 with –min-coverage set to 1 and –min-var-freq set to 0.01. The input for VarScan was piped from the output of samtools mpileup using the default parameters. VarScan generates a .snp file, which we parse into a VCF file (21).
5. iVAR v1.2.3 using the default parameters and the minimum frequency (-t) option set to 0.001. The input to ivar is also piped from the output of mpileup using the options -aa -A -d 0 -B -Q 0. The ‘PASS’ field in the output of iVar was ignored in generation of the vcf files (22).
6. GATK HaplotypeCaller v4.1.7.0 with the -ploidy option set to 100. This generates a vcf with both snps and indels. GATK selectVariants was used to extract just the snps from these files (23, 24).

Intersections between the workflow VCF files (produced by Mutect2, Freebayes, timo, VarScan, iVar and haplotype caller) and the golden VCF file were generated using bcftools isec v1.9 (48). The output from bcftools isec was then analyzed and compared against the respective AF-specific golden VCF to compare allele frequencies using a custom script.

The pipeline used to analyze the data is available at https://github.com/gencorefacility/MAD.

Assembly of genomes and consensus sequences

Reads were base-called with Picard Tools IlluminaBasecallsToFastq v2.17.11 and demultiplexed using Pheniqs allowing for 1 mismatch in sample index sequences (49, 50). Illumina sequencing adapters and primer sequences were trimmed with Trimmomatic v0.36 (47). The trimmed reads were aligned to the Wuhan-Hu-1 SARS-CoV-2 reference genome (NC_045512.2) using BWA mem v0.7.17 with the -K parameter set to 100000000 for reproducibility and -Y to use soft clipping for supplementary alignments (48). The two primer pool libraries for each biological sample were merged into one alignment file using Picard Tools MergeSamFiles v2.17.11. Duplicates were marked using GATK MarkDuplicatesSpark v4.1.3.0 (https://gatk.broadinstitute.org/hc/en-us/articles/360037224932-MarkDuplicatesSpark). Variants were called using GATK HaplotypeCaller v4.1.3.0 with -ploidy set to 1 and filtered for single nucleotide variants with Quality Depth > 2, Fisher Strand < 60, Mapping Quality > 40, and Symmetric Odds Ratio > 4.0. Viral consensus sequences were generated from VCF files based on the NC_045512.2 reference using GATK FastaAlternateReferenceMaker v4.1.3.0; regions below 5x were masked with Ns. Predicted SNV effects were called using SnpEff v4.3i (51). The pipeline used to analyze the data is available at https://github.com/gencorefacility/covid19.

Identification of minority variants

Minority variants were identified using our in-house python script, *timo,* that iterates through merged alignment files (https://github.com/GhedinLab/timo). Minority variants were initially called if present at, or above, a .1% frequency at a position with at least 1x coverage, identified in both forward and reverse reads, and had a Phred score of at least 25. Of the 12 samples included in these analyses, nine were sequenced in duplicate. Only minority variants present in both outputs at an allele frequency greater than 0.02, at a coverage of at least 200X were considered for follow up analysis.

Generation of phylogenetic trees

Isolates of human SARS-CoV-2 were retrieved from the GISAID EpiCov database as of 2020-10-15 (52). The of 3' and 5' regions of sequences were truncated, and sequences containing more than 1% ambiguous sites or those which had an incomplete collection date annotation were removed. From the remaining set we sampled randomly up to 1000 isolates per month leaving 10932 isolates as representatives of the global population (**Supplementary Table 2**). The sequences were aligned with MAFFTv7.467 (53) to a reference isolate from GenBank (54) (Accession: MN908947, Wuhan-Hu-1, isolate collected on December 19th 2019 in Wuhan, China). This alignment of the selected 10932 isolates, including the consensus sequences from the NYU Langone samples, was used to infer the maximum likelihood phylogeny under the nucleotide substitution model GTR+G in IQTree (55). The tree topology was assessed using the ultrafast bootstrap function with 1000 replicates (56). To root the tree, we specified the reference isolate hCoV-19/Wuhan/Hu-1/2019 (GISAID-Accession: EPI ISL 402125), which is identical in sequence to the GenBank isolate used in the alignment step. We inferred the sequences of internal nodes, the optimized timing of internal nodes and resolved polytomies on the final ML-Tree with TreeTime (57). We used a fixed clock rate of 8×10^4 (stdev = 4×10^4) mutations/ (bp day) under a skyline coalescent tree prior and we rooted the tree using the same reference isolate as with the IQTree step of topology reconstruction (GISAID-Accession: EPI ISL 402125) (58). The clock rate was computed as the total number of mutations on the tree, divided by the total length of branches of the timed tree. This rate was optimized by iterative runs of TreeTime until convergence. The time of the root of the tree is estimated to December 19, 2019.

Identification of circulating clades

We characterize the main genetic clades by identifying non-synonymous amino-acid mutations that originate prevalent viral population subtrees. We computed global population clade frequency as follows: (1) Individual isolates, which we index with i, are assigned a smoothened multiplicity factor, $n_i(t) = \exp[-(t-t_i)^4/(2\sigma^4)]$, where t_i is the collection date of the isolate, and the squared Gaussian kernel is $\sigma = 3$ days. Sample frequencies of isolates are computed as $x_i^s(t) = N_i(t)/N(t)$, where t_i is the sampling time and $N(t) = \sum_i N_i(t)$. (2) To correct for regional sequence sampling bias, we computed reweighed frequencies by calibration with the daily incidence data from JHU(59), $x_i(t) = m_{c(i)} x_i^s(t)$, where $c(i)$ is the continent of isolate i. The reweighing factors are defined by $m_c(t) = y_c(t) \,/\, \sum_{i\in c} x_i^s(t)$. Here $y_c(t)$ denotes the fraction of incidence in continent c, which is obtained from the JHU data on a given date t (59). We use the following broad geographical regions: USA East Coast, USA West Coast, North America remainder, Europe, Asia, China, South America, Africa, and Oceania. (3) From the corrected isolate frequencies, we obtained global clade frequencies $X_\alpha(t) = \sum_{i\in\alpha} x_i(t)$. We kept all clades that have reached a threshold frequency 5% on any day since the start of the epidemic.

Distance statistics of doublet variants

The cumulative distance distribution for doublet pairs, $P(d)$ was compared with the corresponding null distribution for random pairs of variants across different hosts in our sample set. The cumulative null distribution is given by

$$P_0(d) = \frac{\sum_{k<k'} n_k n_{k'} \, H(-d_{kk'} + d)}{\sum_{k<k'} n_k n_{k'}}$$

where n_k is the number of singlet variants in host k and H is the Heaviside step function (Fig 5A). In these analyses, higher multiplets were excluded because they are a priori unlikely to occur under transmission.

Mutant weight functions

Empirical mutant weight distributions $Y(x)$ of intra-host variants were constructed from a list of mutants ordered by decreasing frequency, $(x_1, x_2, \dots)$, i.e., by increasing origination time. We recursively computed

$$Y_m = Y_{m-1} + x_m(1 - Y_{m-1}) \qquad (3)$$

and plotted Y_m vs. x_m. This recursion used a random-genealogy assumption: the mth mutation appears with probability $(1 - Y_{m-1})$ on the ancestral background and with probability Y_{m-1} on the background of a previous mutation. This recursion was evaluated independently for all mutations, synonymous mutations, and non-synonymous mutations, giving the weight functions $Y(x)$, $Y_s(x)$, and $Y_a(x)$ reported in Fig 6B

Data Availability

Data is available in NCBI GenBank and SRA. All accession IDs can be found in Supplementary Table 1, and data in SRA can be found under BioProject ID PRJNA721724.

Ethics Statement

The participants in this research study provided written informed consent in advance of any study activities. The informed consent form was reviewed and approved by the NYU Langone Institutional Review Board (IRB). The study protocol (study number i18-02035) was reviewed and approved by the NYU Langone IRB.

ACKNOWLEDGEMENTS

We would like to thank Robert Sebra and Harm van Bakel of the Icahn School of Medicine at Mt Sinai for their suggestions on primer design and protocol development early on, as well as the participants who joined our study and the NYU Vaccine Center staff. This work was supported in part by R01AI140766 (DG), the Division of Intramural Research (DIR) of the NIAID/NIH (EG), AI148574 (MJM), and the Deutsche Forschungsgemeinschaft grant SFB 1310 (ML). MŁ is a Pew Biomedical Scholar and was partially supported by the NIAID Centers of Excellence for Influenza Research and Surveillance (contract HHSN272201400008C).

FIGURE LEGENDS

Figure 1. Analysis of variant callers across several allele frequencies on simulated SARS-CoV-2 data. (**A**) Receiver Operating Characteristic (ROC) of tested variant callers across a range of allele frequencies (AF). ROC is a function of the true positive rate (true positive/condition positive) and the false positive rate (false positive/condition negative). (**B**) Precision/Recall (PR) curves of variant callers across a range of allele frequencies. PR graphs precision (true positive/true positive + false positive) against recall, also known as the true positive rate (true positive/condition positive. Green boxes show area of the graph which indicate superior performance based on these metrics. (**C**) ROC of tested variant callers across a range of down-sampled coverages at a set AF of 0.02 (**D**) PR of tested variant callers across a range of down-sampled coverages at a set AF of 0.02. (**E**) Variant calling performance of *timo* at ranges of AFs in simulated data where minority variants were placed at random allele frequencies.

Figure 2. Phylogeny of New York City SARS-CoV-2 samples. (**A**, **B**) Maximum-likelihood timed strain tree reconstructed from 10932 sequences from GISAID (Methods). The tree is colored by major genetic clades, the isolates from this study are shown in detail on the left panel and highlighted in the right panel. **(C)** Consensus changes found with the 12 samples plotted across the SARS-CoV-2 genome. Y axis represents the frequency of a given consensus change within our cohort, where 1.0 indicates the change is found in all 12 samples. Bars are colored according to the nucleotide and the reference nucleotide (Wuhan-Hu-1) is shown along the bottom of the graph. (**D**) Heatmap showing the frequency of transitions and transversions represented in the identified consensus changes.

Figure 3. Reproducibility of minority variants across sequencing replicates. (**A**) UpsetR plots show the shared and unique minority variants identified by *timo* at an allele frequency of > 0.5% in replicate amplification/sequencing runs from clinical NS samples. Red numbers below

indicate intersections at an allele frequency of 0.02. (**B, C**) Correlation between viral load and both reproducible and non-reproducible minority variants shown in panel (A).

Figure 4. Minority variants in SARS-CoV-2 sequence data. (**A**) High confidence minority variants (identified in replicate amplification/sequencing runs) graphed across the SARS-CoV-2 genome. Height of bar indicates frequency of the variant across the cohort samples. Shared variants (present in > 1 sample) are labeled with the gene ID and the nucleotide change. (**B**) Heatmap showing the frequency of transitions and transversions represented in the identified minority variants. (**C**) UpsetR plot showing sharing of minority variants between samples in the cohort. Vertical bars indicate the size of the shared set of variants while dots and connecting lines show which samples share a given set of variants. Horizontal bars show total numbers of variants identified in each sample. (**D**) Circle plot showing shared, non-synonymous minority variants across the 12 samples. Outer circle represents the major amino acid at the indicated position which inner circle represents the amino acid coded by the minority variant. Circles are not shown for samples/regions where coverage at that position was not >= 200X. (**E**) Circle plot of ORF1a, aa position 1429.

Figure 5. Uniquely shared variants are enriched at close distances on the consensus tree. (**A**) UpsetR plot showing sharing of minority variants between samples in the cohort at an allele frequency of 0.005. Vertical bars indicate the size of the shared set of variants while dots and connecting lines show which samples share a given set of variants. Horizontal bars show total numbers of variants identified in each sample. (**B**) Cumulative distribution of Hamming distances between samples for doublet minor variants, $P(d)$, and for random pairs of variants across all samples, $P_0(d)$ (Methods). (**C**) The fraction of doublet variants in sample NYU-VC-022 is significantly enriched as compared to the remaining samples, due to the 6 variants shared with sample NYU-VC-023.

Figure 6. SARS-CoV-2 transmission droplets are heterogeneous. (A) Data distributions $\Phi_s(x)$ (synonymous mutations, blue), $\Phi_n(x)$ (nonsynonymous mutations, orange), and $\Phi(x) = \Phi_s(x) + \Phi_n(x)$ (all mutations, green) are plotted together with fit functions of the form (1) (dashed lines). **(B)** Empirical mutant weight functions $Y_s(x)$ (synonymous mutations, blue), $Y_n(x)$ (nonsynonymous mutations, orange), and $Y(x)$ (all mutations, green); see Methods.

REFERENCES

1. Zhou P, Yang XL, Wang XG, Hu B, Zhang L, Zhang W, Si HR, Zhu Y, Li B, Huang CL, Chen HD, Chen J, Luo Y, Guo H, Jiang RD, Liu MQ, Chen Y, Shen XR, Wang X, Zheng XS, Zhao K, Chen QJ, Deng F, Liu LL, Yan B, Zhan FX, Wang YY, Xiao GF, Shi ZL. 2020. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature 579:270-273.

2. Wu F, Zhao S, Yu B, Chen YM, Wang W, Song ZG, Hu Y, Tao ZW, Tian JH, Pei YY, Yuan ML, Zhang YL, Dai FH, Liu Y, Wang QM, Zheng JJ, Xu L, Holmes EC, Zhang YZ. 2020. A new coronavirus associated with human respiratory disease in China. Nature 579:265-269.

3. Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, Zhao X, Huang B, Shi W, Lu R, Niu P, Zhan F, Ma X, Wang D, Xu W, Wu G, Gao GF, Tan W, China Novel Coronavirus I, Research T. 2020. A Novel Coronavirus from Patients with Pneumonia in China, 2019. N Engl J Med 382:727-733.

4. Coronaviridae Study Group of the International Committee on Taxonomy of V. 2020. The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. Nat Microbiol 5:536-544.

5. Organization WH. 10/02/2020 2020. WHO Coronavirus Disease (COVID-19) Dashboard. https://covid19.who.int/. Accessed 10/01/2020.

6. Bedford T, Greninger AL, Roychoudhury P, Starita LM, Famulare M, Huang ML, Nalla A, Pepper G, Reinhardt A, Xie H, Shrestha L, Nguyen TN, Adler A, Brandstetter E, Cho S, Giroux D, Han PD, Fay K, Frazar CD, Ilcisin M, Lacombe K, Lee J, Kiavand A, Richardson M, Sibley TR, Truong M, Wolf CR, Nickerson DA, Rieder MJ, Englund JA, Hadfield J, Hodcroft EB, Huddleston J, Moncla LH, Muller NF, Neher RA, Deng X, Gu W, Federman S, Chiu C,

Duchin J, Gautom R, Melly G, Hiatt B, Dykema P, Lindquist S, Queen K, Tao Y, Uehara A, Tong S, et al. 2020. Cryptic transmission of SARS-CoV-2 in Washington State. medRxiv doi:10.1101/2020.04.02.20051417.

7. Prevention CfDCa. 10/01/2020 2020. Coronavirus Disease 2019 (COVID-19). covid.cdc.gov/covid-data-tracker. Accessed 10/1/2020.

8. Gonzalez-Reiche AS, Hernandez MM, Sullivan MJ, Ciferri B, Alshammary H, Obla A, Fabre S, Kleiner G, Polanco J, Khan Z, Alburquerque B, van de Guchte A, Dutta J, Francoeur N, Melo BS, Oussenko I, Deikus G, Soto J, Sridhar SH, Wang YC, Twyman K, Kasarskis A, Altman DR, Smith M, Sebra R, Aberg J, Krammer F, Garcia-Sastre A, Luksza M, Patel G, Paniz-Mondolfi A, Gitman M, Sordillo EM, Simon V, van Bakel H. 2020. Introductions and early spread of SARS-CoV-2 in the New York City area. Science 369:297-301.

9. Maurano MT, Ramaswami S, Westby G, Zappile P, Dimartino D, Shen G, Feng X, Ribeiro-Dos-Santos AM, Vulpescu NA, Black M, Hogan M, Marier C, Meyn P, Zhang Y, Cadley J, Ordonez R, Luther R, Huang E, Guzman E, Serrano A, Belovarac B, Gindin T, Lytle A, Pinnell J, Vougiouklakis T, Boytard L, Chen J, Lin LH, Rapkiewicz A, Raabe V, Samanovic-Golden MI, Jour G, Osman I, Aguero-Rosenfeld M, Mulligan MJ, Cotzia P, Snuderl M, Heguy A. 2020. Sequencing identifies multiple, early introductions of SARS-CoV2 to New York City Region. medRxiv doi:10.1101/2020.04.15.20064931.

10. Koyama T, Platt D, Parida L. 2020. Variant analysis of SARS-CoV-2 genomes. Bull World Health Organ 98:495-504.

11. Peck KM, Lauring AS. 2018. Complexities of Viral Mutation Rates. J Virol 92.

12. Denison MR, Graham RL, Donaldson EF, Eckerle LD, Baric RS. 2011. Coronaviruses: an RNA proofreading machine regulates replication fidelity and diversity. RNA Biol 8:270-9.

13. Smith EC, Blanc H, Surdel MC, Vignuzzi M, Denison MR. 2013. Coronaviruses lacking exoribonuclease activity are susceptible to lethal mutagenesis: evidence for proofreading and potential therapeutics. PLoS Pathog 9:e1003565.

14. Lewis-Rogers N, Seger J, Adler FR. 2017. Human Rhinovirus Diversity and Evolution: How Strange the Change from Major to Minor. J Virol 91.

15. Sanjuan R, Nebot MR, Chirico N, Mansky LM, Belshaw R. 2010. Viral mutation rates. J Virol 84:9733-48.

16. Zhao Z, Li H, Wu X, Zhong Y, Zhang K, Zhang YP, Boerwinkle E, Fu YX. 2004. Moderate mutation rate in the SARS coronavirus genome and its implications. BMC Evol Biol 4:21.

17. Capobianchi MR, Rueca M, Messina F, Giombini E, Carletti F, Colavita F, Castilletti C, Lalle E, Bordi L, Vairo F, Nicastri E, Ippolito G, Gruber CEM, Bartolini B. 2020. Molecular characterization of SARS-CoV-2 from the first case of COVID-19 in Italy. Clin Microbiol Infect 26:954-956.

18. Lythgoe K.A. HM, Ferretti L, et al. 2020. Shared SARS-CoV-2 diversity suggests localised transmission of minority variants. bioRxiv doi:https://doi.org/10.1101/2020.05.28.118992.

19. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, Gabriel S, Meyerson M, Lander ES, Getz G. 2013. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol 31:213-9.

20. Garrison EM, Gabor. 2012. Haplotype-based variant detection from short-read sequencing. ArXiv 1207.3907v2.

21. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, Miller CA, Mardis ER, Ding L, Wilson RK. 2012. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res 22:568-76.

22. Grubaugh ND, Gangavarapu K, Quick J, Matteson NL, De Jesus JG, Main BJ, Tan AL, Paul LM, Brackney DE, Grewal S, Gurfield N, Van Rompay KKA, Isern S, Michael SF, Coffey LL, Loman NJ, Andersen KG. 2019. An amplicon-based sequencing framework for accurately measuring intrahost virus diversity using PrimalSeq and iVar. Genome Biol 20:8.

23. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43:491-8.

24. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, Banks E, Garimella KV, Altshuler D, Gabriel S, DePristo MA. 2013. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinformatics 43:11 10 1-11 10 33.

25. Stephens ZD, Hudson ME, Mainzer LS, Taschuk M, Weber MR, Iyer RK. 2016. Simulating Next-Generation Sequencing Datasets from Empirical Mutation and Sequencing Models. PLoS One 11:e0167047.

26. B K, WM F, al. e. 2020. Spike mutation pipeline reveals the emergence of a more transmissible form of SARS-CoV-2. bioRxiv doi:https://doi.org/10.1101/2020.04.29.069054.

27. Júnior IJM, Polveiro RC, Souza GM, Bortolin DI, Sassaki FT, Lima ATM. 2020. The global population of SARS-CoV-2 is composed of six major subtypes. bioRxiv.

28. Luria SE, Delbruck M. 1943. Mutations of Bacteria from Virus Sensitivity to Virus Resistance. Genetics 28:491-511.

29. Holland J, Spindler K, Horodyski F, Grabau E, Nichol S, VandePol S. 1982. Rapid evolution of RNA genomes. Science 215:1577-85.

30. Eigen M. 1993. Viral quasispecies. Sci Am 269:42-9.

31. Domingo E, Martin V, Perales C, Grande-Perez A, Garcia-Arriaza J, Arias A. 2006. Viruses as quasispecies: biological implications. Curr Top Microbiol Immunol 299:51-82.

32. Lauring AS, Andino R. 2010. Quasispecies theory and the behavior of RNA viruses. PLoS Pathog 6:e1001005.

33. Novella IS, Presloid JB, Taylor RT. 2014. RNA replication errors and the evolution of virus pathogenicity and virulence. Curr Opin Virol 9:143-7.

34. Fitzsimmons WJ, Woods RJ, McCrone JT, Woodman A, Arnold JJ, Yennawar M, Evans R, Cameron CE, Lauring AS. 2018. A speed-fidelity trade-off determines the mutation rate and virulence of an RNA virus. PLoS Biol 16:e2006459.

35. Sapoval N, Mahmoud M, Jochum MD, Liu Y, Elworth RAL, Wang Q, Albin D, Ogilvie H, Lee MD, Villapol S, Hernandez KM, Berry IM, Foox J, Beheshti A, Ternus K, Aagaard KM, Posada D, Mason CE, Sedlazeck F, Treangen TJ. 2020. Hidden genomic diversity of SARS-CoV-2:

implications for qRT-PCR diagnostics and transmission. bioRxiv doi:10.1101/2020.07.02.184481.

36. Goswami P, Bartas M, Lexa M, Bohalova N, Volna A, Cerven J, Cervenova V, Pecinka P, Spunda V, Fojta M, Brazda V. 2020. SARS-CoV-2 hot-spot mutations are significantly enriched within inverted repeats and CpG island loci. Brief Bioinform doi:10.1093/bib/bbaa385.

37. Simmonds P. 2020. Rampant C-->U Hypermutation in the Genomes of SARS-CoV-2 and Other Coronaviruses: Causes and Consequences for Their Short- and Long-Term Evolutionary Trajectories. mSphere 5.

38. Di Gioacchino A, Šulc P, Komarova AV, Greenbaum BD, Monasson R, Cocco S. 2021. The Heterogeneous Landscape and Early Evolution of Pathogen-Associated CpG Dinucleotides in SARS-CoV-2. Molecular Biology and Evolution doi:10.1093/molbev/msab036.

39. Valesano ALR, Kalee E; Dimcheff, Derek E; Blair, Christopher N; Fitzsimmons, William J; Petrie, Joshua G; Martin, Emily T; Lauring, Adam S. 2021. Temporal dynamics of SARS-CoV-2 mutation accumulation within and across infected hosts. bioRxiv doi:https://doi.org/10.1101/2021.01.19.427330.

40. Martin MAK, Katia. 2021. Reanalysis of deep-sequencing data from Austria points towards a small SARS-COV-2 transmission bottleneck on the order of one to three virions. bioRxiv doi:https://doi.org/10.1101/2021.02.22.432096.

41. Popa A, Genger JW, Nicholson MD, Penz T, Schmid D, Aberle SW, Agerer B, Lercher A, Endler L, Colaco H, Smyth M, Schuster M, Grau ML, Martinez-Jimenez F, Pich O, Borena W, Pawelka E, Keszei Z, Senekowitsch M, Laine J, Aberle JH, Redlberger-Fritz M, Karolyi

M, Zoufaly A, Maritschnik S, Borkovec M, Hufnagl P, Nairz M, Weiss G, Wolfinger MT, von Laer D, Superti-Furga G, Lopez-Bigas N, Puchhammer-Stockl E, Allerberger F, Michor F, Bock C, Bergthaler A. 2020. Genomic epidemiology of superspreading events in Austria reveals mutational dynamics and transmission properties of SARS-CoV-2. Sci Transl Med 12.

42. Wang DX, Wang YQ, Sun WY, Zhang L, Ji JK, Zhang ZY, Cheng XY, Li YM, Xiao F, Zhu AR, Zhong B, Ruan SC, Li JD, Ren PD, Ou ZH, Xiao MF, Li M, Deng ZQ, Zhong HZ, Li FQ, Wang WJ, Zhang YW, Chen WJ, Zhu SD, Xu X, Jin X, Zhao JX, Zhong NS, Zhang WW, Zhao JC, Li JH, Xu YH. 2021. Population Bottlenecks and Intra-host Evolution During Human-to-Human Transmission of SARS-CoV-2. Frontiers in Medicine 8.

43. Lythgoe KA, Hall M, Ferretti L, de Cesare M, MacIntyre-Cockett G, Trebes A, Andersson M, Otecko N, Wise EL, Moore N, Lynch J, Kidd S, Cortes N, Mori M, Williams R, Vernet G, Justice A, Green A, Nicholls SM, Ansari MA, Abeler-Dorner L, Moore CE, Peto TEA, Eyre DW, Shaw R, Simmonds P, Buck D, Todd JA, Oxford Virus Sequencing Analysis G, Connor TR, Ashraf S, da Silva Filipe A, Shepherd J, Thomson EC, Consortium C-GU, Bonsall D, Fraser C, Golubchik T. 2021. SARS-CoV-2 within-host diversity and transmission. Science doi:10.1126/science.abg0821.

44. Wang Y, Wang D, Zhang L, Sun W, Zhang Z, Chen W, Zhu A, Huang Y, Xiao F, Yao J, Gan M, Li F, Luo L, Huang X, Zhang Y, Wong SS, Cheng X, Ji J, Ou Z, Xiao M, Li M, Li J, Ren P, Deng Z, Zhong H, Xu X, Song T, Mok CKP, Peiris M, Zhong N, Zhao J, Li Y, Li J, Zhao J. 2021. Intra-host variation and evolutionary dynamics of SARS-CoV-2 populations in COVID-19 patients. Genome Med 13:30.

45. Cheung PP, Rogozin IB, Choy KT, Ng HY, Peiris JS, Yen HL. 2015. Comparative mutational analyses of influenza A viruses. RNA 21:36-47.

46. CDC. 2020. CDC 2019-Novel Coronavirus (2019-nCoV) Real-Time RT-PCR Diagnostic Panel.

47. Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114-20.

48. Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754-60.

49. Anonymous. 2018. Picard toolkit, *on* Broad Institute. http://broadinstitute.github.io/picard/. Accessed 2020-10-02.

50. Galanti LS, Dennis; Gunsalus, Kristin C. . 2017. Pheniqs: Fast and flexible quality-aware sequence demultiplexing. bioRxiv doi:https://doi.org/10.1101/128512.

51. Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin) 6:80-92.

52. Elbe S, Buckland-Merrett G. 2017. Data, disease and diplomacy: GISAID's innovative contribution to global health. Glob Chall 1:33-46.

53. Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30:772-80.

54. Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. 2017. GenBank. Nucleic Acids Res 45:D37-D42.

55. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, Lanfear R. 2020. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Mol Biol Evol 37:1530-1534.

56. Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. 2018. UFBoot2: Improving the Ultrafast Bootstrap Approximation. Mol Biol Evol 35:518-522.

57. Sagulenko P, Puller V, Neher RA. 2018. TreeTime: Maximum-likelihood phylodynamic analysis. Virus Evol 4:vex042.

58. Kingman JFC. 1982. The coalescent. Stochastic Processes and their Applications 13:235-248.

59. Dong E, Du H, Gardner L. 2020. An interactive web-based dashboard to track COVID-19 in real time. Lancet Infect Dis 20:533-534.

Figure 1

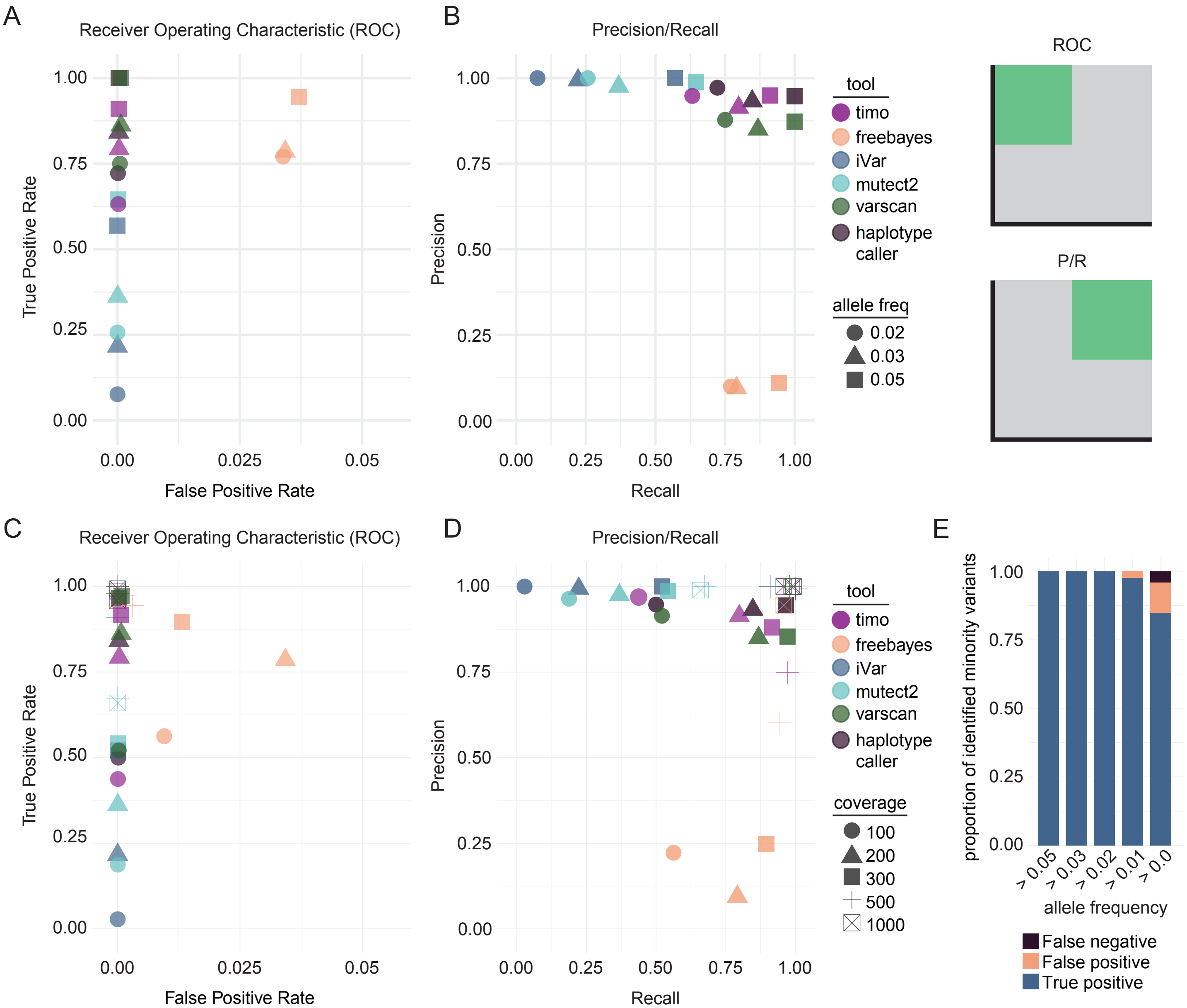

A
Receiver Operating Characteristic (ROC)
True Positive Rate
1.00
0.75
0.50
0.25
0.00
0.00
0.025
0.05
False Positive Rate
B
Precision/Recall
Precision
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
Recall
tool
timo
freebayes
iVar
mutect2
varscan
haplotype caller
allele freq
0.02
0.03
0.05
ROC
P/R
C
Receiver Operating Characteristic (ROC)
True Positive Rate
1.00
0.75
0.50
0.25
0.00
0.00
0.025
0.05
False Positive Rate
D
Precision/Recall
Precision
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
Recall
tool
timo
freebayes
iVar
mutect2
varscan
haplotype caller
coverage
100
200
300
500
1000
E
proportion of identified minority variants
1.00
0.75
0.50
0.25
0.00
> 0.05
> 0.03
> 0.02
> 0.01
> 0.0
allele frequency
False negative
False positive
True positive

Figure 2

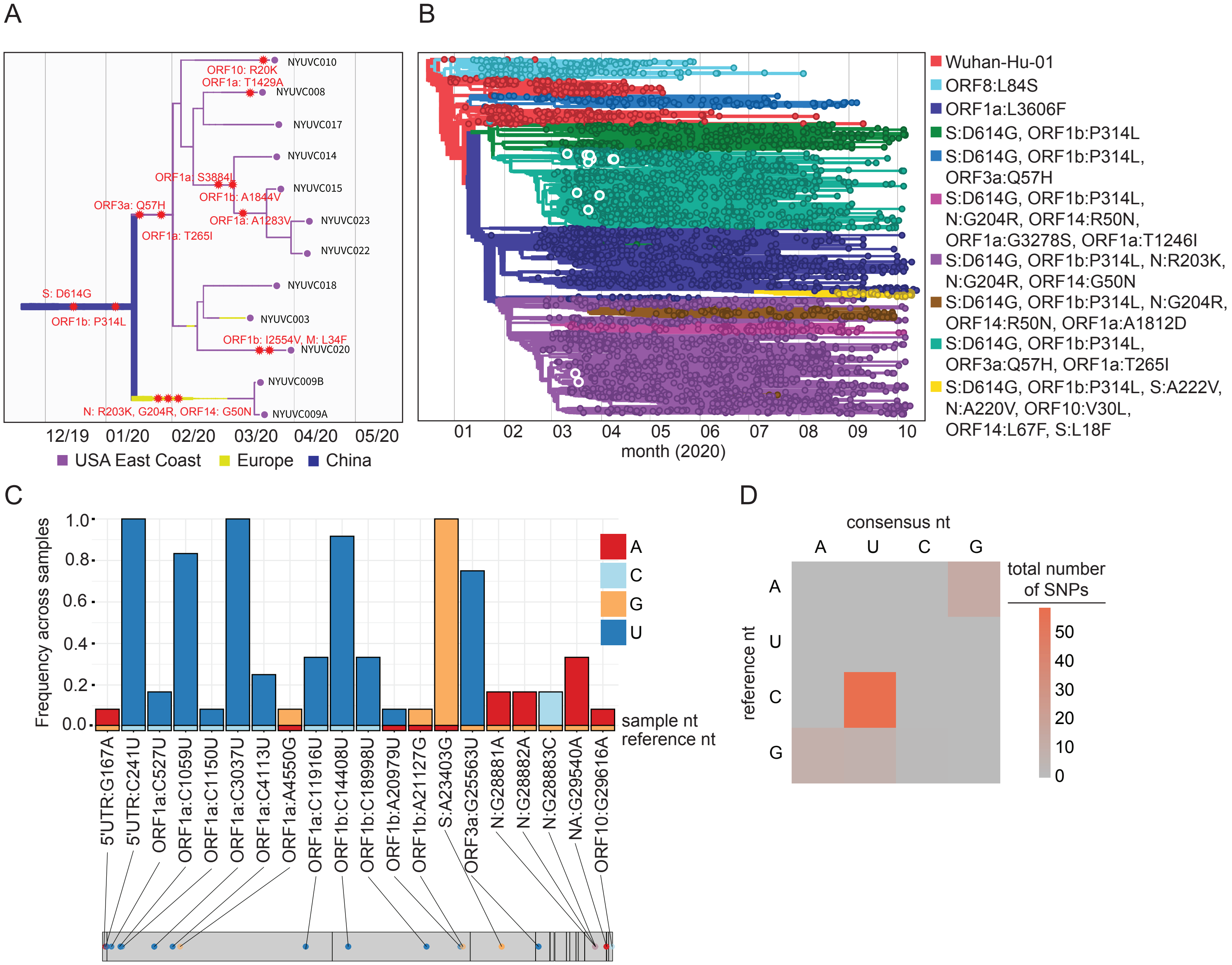

A
ORF10: R20K
ORF1a: T1429A
NYUVC010
NYUVC008
NYUVC017
NYUVC014
ORF1a: S3884L
NYUVC015
ORF1b: A1844V
ORF3a: Q57H
ORF1a: A1283V
NYUVC023
ORF1a: T265I
NYUVC022
NYUVC018
S: D614G
ORF1b: P314L
NYUVC003
ORF1b: I2554V, M: L34F
NYUVC020
NYUVC009B
N: R203K, G204R, ORF14: G50N
NYUVC009A
12/19
01/20
02/20
03/20
04/20
05/20
USA East Coast
Europe
China
B
01
02
03
04
05
06
07
08
09
10
month (2020)
Wuhan-Hu-01
ORF8:L84S
ORF1a:L3606F
S:D614G, ORF1b:P314L
S:D614G, ORF1b:P314L, ORF3a:Q57H
S:D614G, ORF1b:P314L, N:G204R, ORF14:R50N, ORF1a:G3278S, ORF1a:T1246I
S:D614G, ORF1b:P314L, N:R203K, N:G204R, ORF14:G50N
S:D614G, ORF1b:P314L, N:G204R, ORF14:R50N, ORF1a:A1812D
S:D614G, ORF1b:P314L, ORF3a:Q57H, ORF1a:T265I
S:D614G, ORF1b:P314L, S:A222V, N:A220V, ORF10:V30L, ORF14:L67F, S:L18F
C
Frequency across samples
1.0
0.8
0.6
0.4
0.2
0.0
A
C
G
U
sample nt
reference nt
5'UTR:G167A
5'UTR:C241U
ORF1a:C527U
ORF1a:C1059U
ORF1a:C1150U
ORF1a:C3037U
ORF1a:C4113U
ORF1a:A4550G
ORF1a:C11916U
ORF1b:C14408U
ORF1b:C18998U
ORF1b:A20979U
ORF1b:A21127G
S:A23403G
ORF3a:G25563U
N:G28881A
N:G28882A
N:G28883C
NA:G29540A
ORF10:G29616A
D
consensus nt
A
U
C
G
reference nt
A
U
C
G
total number of SNPs
50
40
30
20
10
0

Figure 3

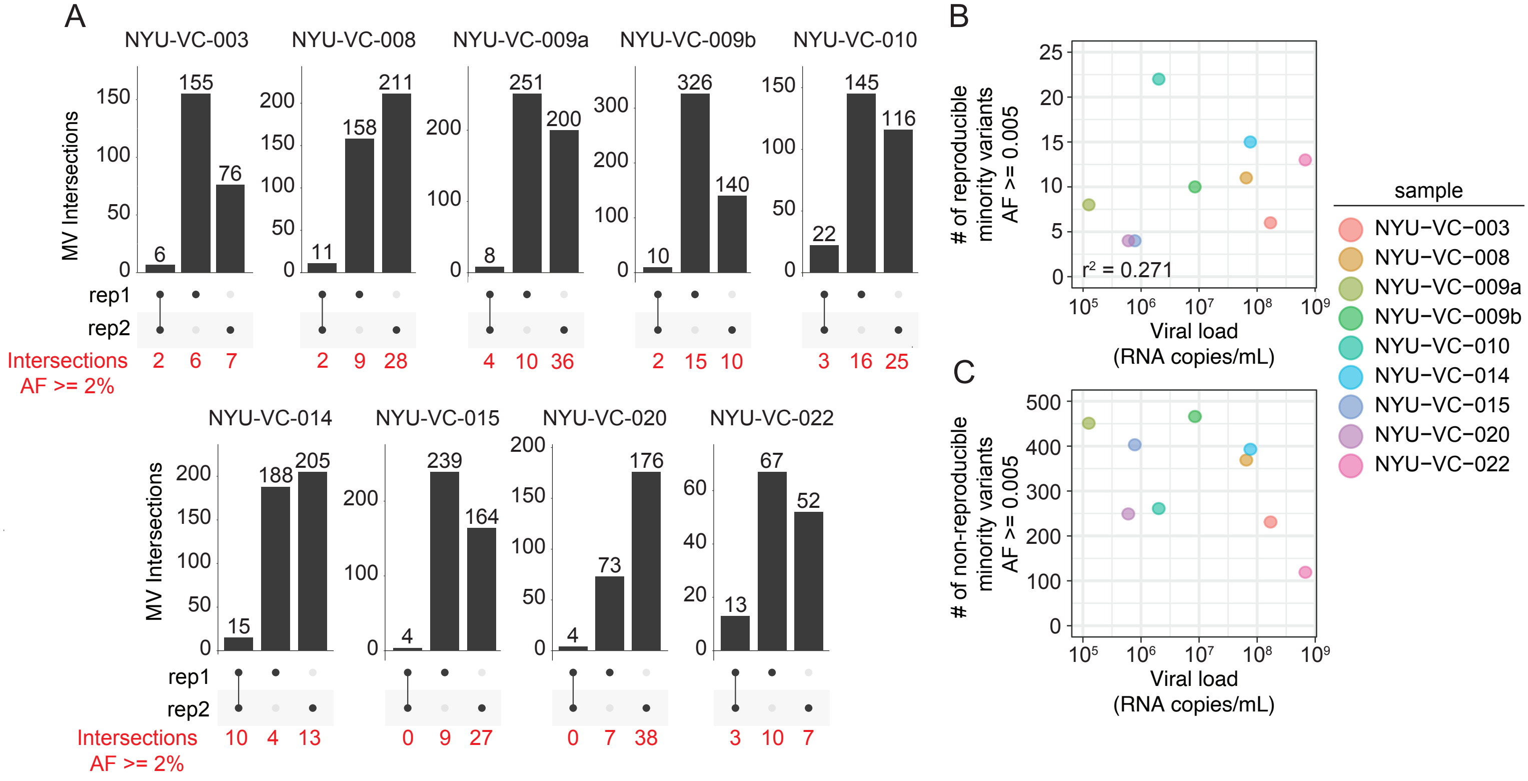

A
NYU-VC-003
NYU-VC-008
NYU-VC-009a
NYU-VC-009b
NYU-VC-010
MV Intersections
6 155 76
11 158 211
8 251 200
10 326 140
22 145 116
rep1
rep2
Intersections AF >= 2%
2 6 7
2 9 28
4 10 36
2 15 10
3 16 25
NYU-VC-014
NYU-VC-015
NYU-VC-020
NYU-VC-022
15 188 205
4 239 164
4 73 176
13 67 52
10 4 13
0 9 27
0 7 38
3 10 7
B
# of reproducible minority variants AF >= 0.005
r² = 0.271
Viral load (RNA copies/mL)
10^5 10^6 10^7 10^8 10^9
C
# of non-reproducible minority variants AF >= 0.005
sample
NYU−VC−003
NYU−VC−008
NYU−VC−009a
NYU−VC−009b
NYU−VC−010
NYU−VC−014
NYU−VC−015
NYU−VC−020
NYU−VC−022

Figure 4

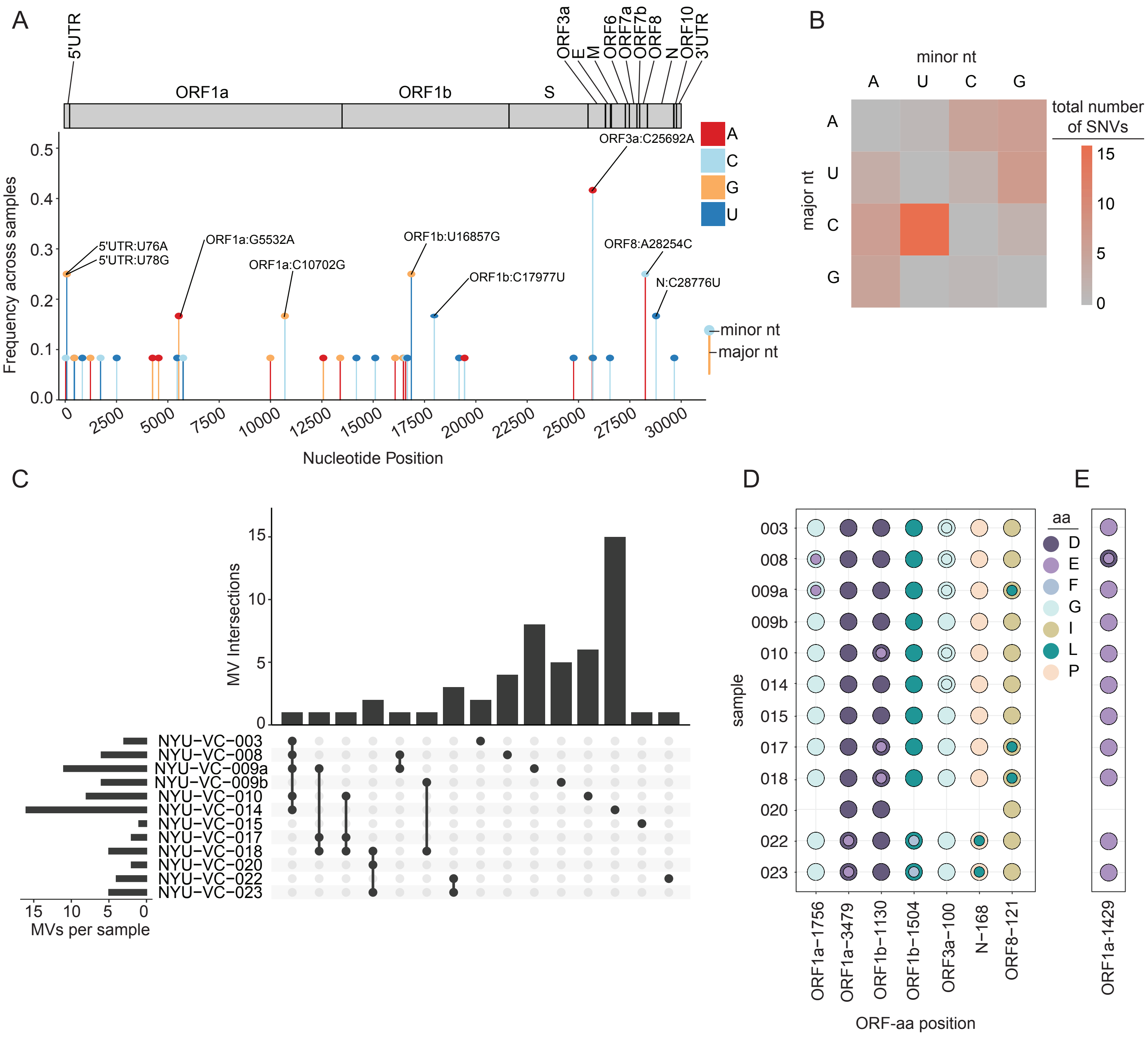

A
5'UTR
ORF1a
ORF1b
S
ORF3a
E
M
ORF6
ORF7a
ORF7b
ORF8
N
ORF10
3'UTR
A
C
G
U
Frequency across samples
0.0
0.1
0.2
0.3
0.4
0.5
5'UTR:U76A
5'UTR:U78G
ORF1a:G5532A
ORF1a:C10702G
ORF1b:U16857G
ORF1b:C17977U
ORF3a:C25692A
ORF8:A28254C
N:C28776U
minor nt
major nt
0
2500
5000
7500
10000
12500
15000
17500
20000
22500
25000
27500
30000
Nucleotide Position
B
minor nt
A
U
C
G
major nt
A
U
C
G
total number of SNVs
15
10
5
0
C
MV Intersections
0
5
10
15
NYU−VC−003
NYU−VC−008
NYU−VC−009a
NYU−VC−009b
NYU−VC−010
NYU−VC−014
NYU−VC−015
NYU−VC−017
NYU−VC−018
NYU−VC−020
NYU−VC−022
NYU−VC−023
15
10
5
0
MVs per sample
D
sample
003
008
009a
009b
010
014
015
017
018
020
022
023
ORF1a−1756
ORF1a−3479
ORF1b−1130
ORF1b−1504
ORF3a−100
N−168
ORF8−121
ORF-aa position
aa
D
E
F
G
I
L
P
E
ORF1a−1429

Figure 5

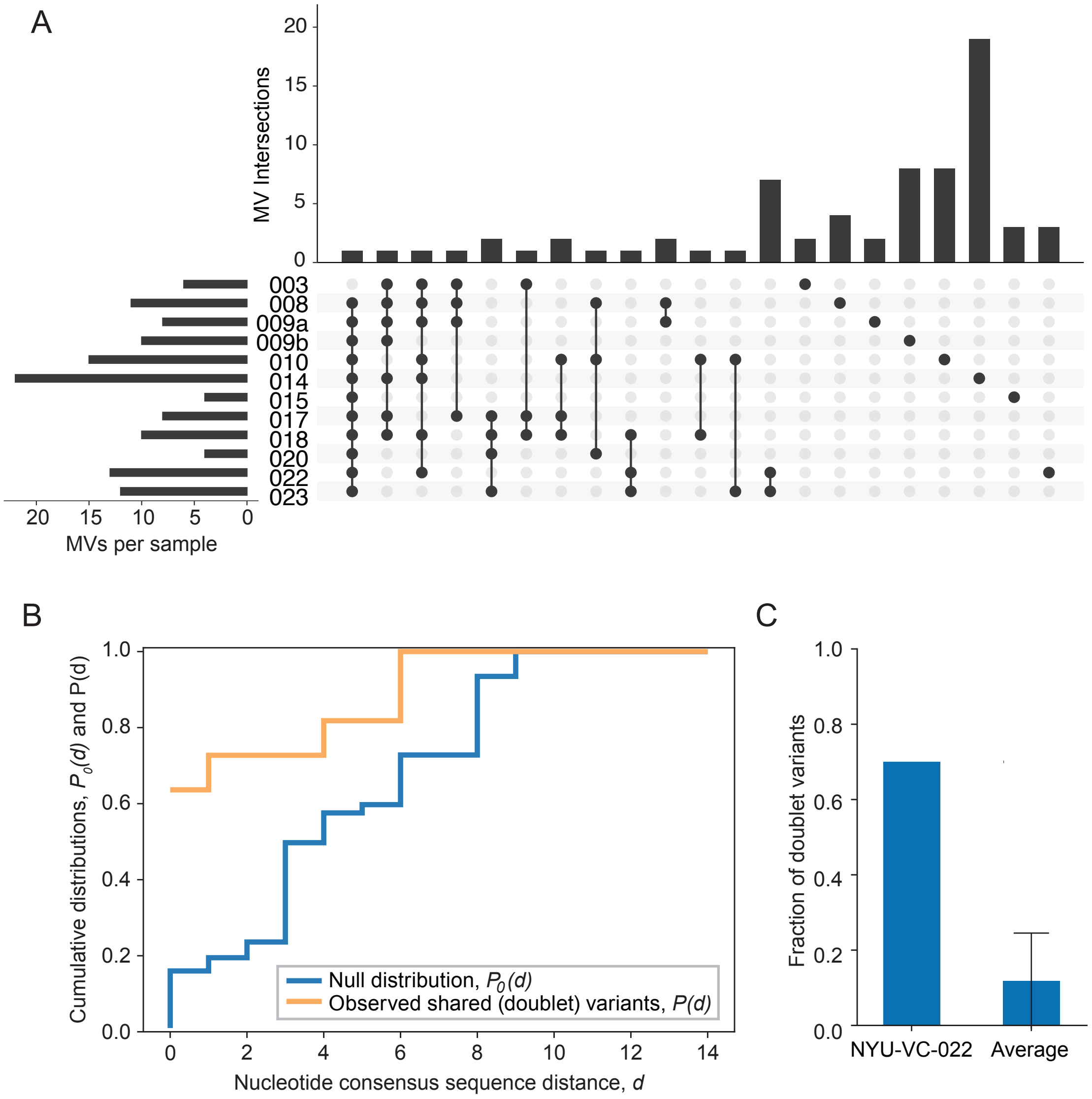
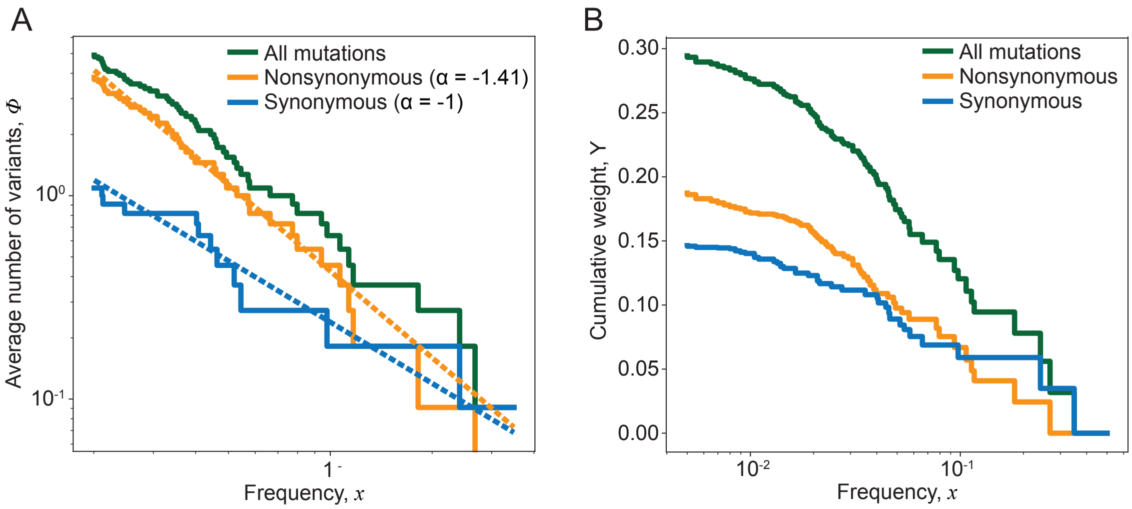

A
MV Intersections
20
15
10
5
0
003
008
009a
009b
010
014
015
017
018
020
022
023
20
15
10
5
0
MVs per sample
B
Cumulative distributions, $P_0(d)$ and P(d)
1.0
0.8
0.6
0.4
0.2
0.0
0
2
4
6
8
10
12
14
Nucleotide consensus sequence distance, *d*
Null distribution, $P_0(d)$
Observed shared (doublet) variants, *P(d)*
C
Fraction of doublet variants
1.0
0.8
0.6
0.4
0.2
0.0
NYU-VC-022
Average

Figure 6


A
All mutations
Nonsynonymous (α = -1.41)
Synonymous (α = -1)
Average number of variants, Φ
10^0
10^{-1}
1 -
Frequency, x
B
All mutations
Nonsynonymous
Synonymous
Cumulative weight, Y
0.30
0.25
0.20
0.15
0.10
0.05
0.00
10^{-2}
10^{-1}
Frequency, x