

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*}

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

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

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

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

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

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

25 **ABSTRACT**

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

39

40

41

42

43

44

45

46

47

48

49

50

51 **IMPORTANCE**

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

62

63

64

65

66

67

68

69

70

71

72

73

74

75 **INTRODUCTION**

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

81 Sequencing of SARS-CoV-2 from infected patients has contributed to our knowledge of
82 the viral origin, the biology of infection, and viral transmission events as well as given insight into
83 the spread of the virus across the world. Despite efforts to prevent introductions of the virus to the
84 United States from areas of the world with active outbreaks, the first positive case of SARS-CoV-
85 2 was reported on January 19, 2020 from Washington state (6). Since this first reported
86 introduction, new outbreaks have occurred in all major US cities and areas (7). Sequencing of
87 virus from infected patients in these cities has helped to determine both the number and origin of
88 these introduction events (8, 9). Viral sequencing has also identified key amino acid changes that
89 differentiate clades of the virus in circulation (10). Identification of these clades and the associated
90 viral consensus changes aids in tracking spread of the virus. However, little has been done to
91 examine potential early detection of emerging variants before they become fixed in the population.

92 Due to the error-prone nature of viral polymerases, as well as the speed of viral replication,
93 errors are introduced into viral genomes during replication (11). These errors can range from
94 lethal (killing the virus) to beneficial, enhancing the viral lifecycle. Coronavirus polymerases are
95 unique among RNA viruses in that they possess a level of proofreading capability (12, 13). This
96 function results in a mutation rate that is significantly lower than other RNA viruses such as
97 rhinovirus or influenza A virus (14-16). Nonetheless, mutations are still introduced during viral
98 replication. Mutations can lead to changes in the consensus sequence; these specific sets of

99 mutations separate the circulating virus population into clades. Mutations in the virus genomes
100 that are not the majority within an infected host (present at lower than 50% frequency) represent
101 minority variants. Identification of these minority variants within the sequencing data can highlight
102 regions of the genome under positive selection or regions with increased mutational tolerance,
103 detect subtle virus population shifts within the infected host, and identify mutations before
104 consensus changes occur (17). These variants can also shed light on tropism, and shared
105 minority variants between samples can show patterns of viral evolution (18). The presence of
106 these variants may have long term implications for vaccine, monoclonal antibody, and drug
107 development.

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

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

125

126 RESULTS

127 Strict cutoffs are necessary for accurate identification of minority variants in SARS-CoV-2 128 sequence data

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

151 positives (**Fig. 1E**). Based on our testing, we chose to use *timo* with a coverage cutoff of 200x
152 and an allele frequency of 0.02 for the most accurate identification of minority variants within our
153 clinical samples.

154

155 *Most identified minority variants are unique to a single sample*

156 To investigate the minority variants in real SARS-CoV-2 data, we used a small cohort of
157 12 samples from the early outbreak in New York City collected and processed at NYU Langone
158 Health and NYU Grossman School of Medicine. Nasopharyngeal swabs (NS) were collected
159 between March 6, 2020 to April 9, 2020 from 11 individuals between the ages of two weeks and
160 60 years (five females, six males; one individual had samples collected at two time points).
161 Specimen collection occurred on various days post onset of illness (DPO). The samples
162 represented a variety of viral loads, ranging from 10,400 viral RNA copies/ml to 416,800 copies/ml
163 (**Supplementary Table 1**). We achieved more than 88% coverage of the genome at 5X for all 12
164 of the NS samples.

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

176 are a good representation of the genetic diversity of the virus during the time period when they
177 were collected.

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

194 To identify high confidence minority variants within this data set, we sequenced each
195 sample in duplicate, when starting material allowed (nine of 12 samples). We used a low
196 frequency threshold (0.005) to perform an initial filtering of the minority variants called by *timo* and
197 compared the minority variants across the replicate sequences. The large majority of minority
198 variants were not reproducible, indicating that they may have been introduced during the
199 amplification or sequencing processes (**Fig. 3A**). Importantly, we did not find an obvious
200 correlation between viral load and the number of reproducible minority variants in this sample set
201 ($r^2 = 0.271$) (**Fig. 3B-C**). Based on these observations, we filtered our list of variants for only those

202 that existed in both replicates in locations with coverage greater than 200X and an average allele
203 frequency above 0.02. For samples that were only sequenced once due to limited specimen
204 availability, we filtered the minority variants to include only those that were present above our
205 cutoffs and existed in another sample. We used this final list of high confidence minority variants
206 for our analyses.

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

224

225 *Transmission of minor variants between hosts.*

226 There were many instances of minority variants that were common to two or more patients
227 within our sample set (**Fig. 4C**). In order to better understand the set of shared variants, we
228 expanded our set of variants to include those present in both replicates with an allele frequency
229 greater than 0.005 and coverage greater than 200X (**Fig. 5A**). One pair of samples, NYU-VC-022
230 and NYU-VC-023 was of particular interest for these analyses given their proximity on the
231 consensus tree, and the fact that they shared the most minority variants between them (**Fig. 4C**,
232 **5A**). These variant statistics differed strongly from the remainder of the samples, signaling a
233 possible transmission event, either between these samples or across a short intermediate
234 transmission chain. To investigate this possibility, we recorded the cumulative distribution of
235 Hamming distances between samples, d , as recorded on the consensus tree, for all minority
236 variants shared between exactly two hosts (doublets). We then compared this distribution with a
237 null distribution, obtained from random pairs of variants across all of the samples. We found that
238 the majority of the doublet variants, but not those in the random pairs, were found in samples
239 where $d = 0$, suggesting that these pairs of variants are likely the result of transmission, rather
240 than of independent *de novo* mutations (**Fig. 5B**). To show that these variants were enriched
241 specifically in samples NYU-VC-022 and NYU-VC-023, we determined the fraction of doublet
242 variants compared to the sum of both the unique variants (singlets) and the doublets for all
243 samples with replicate sequencing (this includes sample 022, but not 023). NYU-VC-022 had a
244 strongly enhanced fraction of doublet variants compared with the rest of the samples in the data
245 set (**Fig. 5C**). Together, these statistics suggest a short transmission chain involving NYU-VC-
246 022 and NYU-VC-023 and indicate that transmission events contain a genetically diverse mix of
247 virus particles.

248

249 *Nonsynonymous mutations are under negative intra-host selection*

250 Upon entry into a new host, the viral population grows initially in an exponential way. As
251 such, the frequency of a mutation within the population is related to its origination time: a few early
252 mutations of larger frequency are followed by many later mutations of small frequency. This
253 feature, which is well-known in the context of Luria-Delbrück fluctuation assays, can be made
254 quantitative: a mutation originating at time t after the start of growth has an initial frequency $x =$
255 $\exp(-\lambda t)$, where λ is the growth rate of the viral population. If the mutation is nearly neutral, this
256 frequency will stay approximately constant during the subsequent growth process. These
257 dynamics generate a mutation frequency spectrum described by the Luria-Delbrück distribution
258 (28), which is characterized by a cumulative distribution function of the form $\Phi(x) = \frac{c}{x^\alpha}$. This
259 distribution gives the expected number of minority variants with frequency $> x$; the decay
260 exponent distinguishes neutral variants ($\alpha = 1$) and negatively selected variants ($\alpha > 1$). We used
261 this distribution to analyze variants with allele frequencies between 0.02 and 0.5. Mutations with
262 these frequencies are expected to arise predominantly in the first intra-cellular replication cycle,
263 which is firmly in the exponential growth phase. We analyzed the empirical cumulative frequency
264 distributions for synonymous and nonsynonymous minority variants, averaged over the samples
265 with replicate sequencing. The distribution of synonymous variants is consistent with the neutral
266 Luria-Delbrück form ($\alpha = 1$). However, non-synonymous variants showed a somewhat faster
267 decay ($\alpha \approx 1.4$), indicating weak negative selection reducing the fraction of high-frequency
268 variants (**Fig. 6A**). We note that, given the limited frequency range of reliable mutant calling,
269 substantial statistical errors of the inferred decay exponents are to be expected. Negative
270 selection on non-synonymous variants demonstrates that random mutations will often result in a
271 loss of fitness.

272

273 *Transmission droplets are likely heterogeneous*

274 To expand upon the hypothesis that minority variants could be shared between samples
275 by transmission, we used only reproducible variants (those observed in both sequencing
276 replicates) present at an allele frequency above 0.005 and coverage above 200X (**Fig 3A, 5A**)
277 and computed the probability that the transmitted viral population is heterogeneous in sequence.
278 To do this, we evaluated the cumulative mutant weight $Y(x)$, which is defined as the sum of the
279 expected frequencies of mutant clades with frequency $> x$ that arise on the host-specific ancestral
280 (wild-type) background. By construction, this weight discounts all double mutants that arise on
281 the background of an earlier mutant. In **Fig. 6B**, we show the empirical mutant weight functions
282 for synonymous, nonsynonymous, and all mutations, which are computed from the frequency
283 counts using a random-genealogy assumption and averaged over all samples with replicate
284 sequencing (Methods). From this, we infer a substantial weight of all minor variants even if we
285 restrict the frequency range to above the cutoff for variant calling, $Y(x_0) = 0.30$ for
286 $x_0 = 0.5 \times 10^{-2}$. The complement of this weight is an upper bound for the frequency of the
287 ancestral genotype in the evolved viral population, $X_{wt} < 1 - Y(x_0) = 0.70$. Similarly, the weight
288 of non-synonymous mutations, $Y_n(x_0) = 0.18$, determines an upper bound for the frequency of
289 the ancestral amino acid sequence in the evolved viral population, $X_a < 1 - Y_n(x_0) = 0.82$. These
290 ancestral frequencies, in turn, determine the probabilities that a transmission event of n virions is
291 monomorphic in the wild type nucleotide sequence, $p(n) = X_{wt}^n$, or in amino acid sequence,
292 $p_a(n) = X_a^n$. From the inferred weights, we can conclude that even transmission events of
293 moderate virion count ($n \sim 10$) are likely to be polymorphic in nucleotide sequence ($p(10) < 0.03$)
294 and in amino acid sequence ($p_a(10) < 0.14$) (**Fig. 6B**). These analyses show that most
295 transmission droplets of this size would transport minor variants between hosts.

296 DISCUSSION

297 It has long been understood that intra-host viral populations are heterogeneous in nature
298 (29-32). We were particularly interested in the level of viral diversity early during the SARS-CoV-
299 2 pandemic and whether or not minority variant analysis could be used to inform transmission
300 events. Here we performed an in-depth analysis of minority variants within a small set of SARS-
301 CoV-2 samples from the early virus outbreak in New York City. Confident identification of minority
302 variants is complicated by errors introduced during amplification and sequencing and therefore
303 we first determined the best approach for stringent calling of minority variants. We tested six
304 minority variant callers using simulated SARS-CoV-2 deep sequence data and our results
305 highlight the need for stringent coverage and allele frequency cutoffs in minority variant analyses.
306 Using our determined cutoffs, we found a number of shared minority variants between samples
307 and provide evidence that some variants may be passed during transmission events. Together,
308 our results lay the groundwork for future studies of minority variants in SARS-CoV-2 infections.

309 Viral replication is inherently error-prone, and these replication errors result in a diverse
310 population of viruses within a single host (29, 33, 34). Viral sequencing easily allows for the
311 determination of the consensus sequence of the majority population within a host. However,
312 capturing and accurately identifying the other viral mutations that do not constitute the majority is
313 more difficult. The variant callers that we tested take diverse approaches, such as haplotype-
314 based methods (freebayes and haplotype caller), or alignment to the reference (VarScan, iVar
315 and timo). We found that in each category there were tools that performed well and tools that
316 performed more poorly. Freebayes was the least precise and called the highest number of false
317 positives. The false positive rate of both iVar and mutect2 was 0, however these callers were
318 relatively conservative, missing a number of true positives. Many studies use iVar for variant
319 calling in viral genomes and our data suggest that this may result in true variants being
320 overlooked. Haplotype caller, *timo*, and VarScan all performed nearly perfectly, missing only
321 variants that existed at very low frequencies (< 0.01) or at low coverage (< 100X). It is clear from

322 these data that variant callers and cutoffs should be carefully selected in order to increase
323 confidence in the identified variants in studies such as this one. However, even with the best
324 variant calling software, preparation and processing steps necessary for sequencing viral
325 genomes generate layers of error that can make low frequency minority variants virtually
326 indistinguishable from processing errors. In replicate sequencing, with a relatively low frequency
327 cutoff of 0.005, we found that consistently less than 10% of identified minority variants were
328 reproducible, regardless of viral load. Our data suggest that these process errors greatly interfere
329 with confident minority variant prediction and replicate sequencing as well as very stringent cutoffs
330 are thus essential for the identification of variants.

331 The majority of changes we identified in our data, both at the consensus level and in minority
332 variants, were C to U transitions, consistent with published reports (35-38). However, we found
333 few unique consensus changes (only six of the 20 identified mutations) while the majority of
334 minority variants were unique to a single sample (44 of the 54 minority variants). Similarly, the
335 number of consensus changes across samples was relatively consistent, but the number of
336 minority variants differed more significantly. We found no correlation between the number of
337 minority variants and the viral load within our data set, despite studies that suggest that low viral
338 load increases false positive minority variants, likely due to replicated sequencing of our samples
339 (39).

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

348 One frequently debated topic is the possibility that minority variants could be passed between
349 individuals during a transmission event. This possibility depends on a number of factors including
350 the frequency of the minority variant and the size of the transmission bottleneck. For SARS-CoV-
351 2, the size of the bottleneck has been reported to be as few as one to as many as one thousand
352 (40-43). It is likely that transmission between individuals involves multiple transmission events
353 over the course of an interaction, rather than just one. Multiple transmission events would
354 increase the number of viral particles passed between individuals. Our analyses suggest that
355 transmission events are unlikely to be homogeneous and that most virions in the host differ by
356 acquired mutations from the founder genome that was transmitted. This notion is supported by
357 studies that have shown evidence of mixed SARS-CoV-2 infections (18, 44). Moreover, we find
358 that intra-host selection shapes the distribution of minor variants in the high-frequency regime,
359 which includes the variants relevant for transmission. Our current analysis covers broad negative
360 selection on non-synonymous mutations. Future, more densely sampled data may also permit the
361 identification of positively selected minor variants.

362 Further supporting transmission of minority variants, we identified several instances of shared
363 minority variants within our sample set. Some variants were shared between two individuals, while
364 other variants were widely shared between many individuals. Two of our samples, NYU-VC-022
365 and NYU-VC-023, contained many more uniquely shared variants (doublets) than any other set
366 of samples, and these samples were also the closest on the consensus tree. To rule out the
367 possibility of contamination, we re-extracted, amplified and sequenced these samples many
368 months after initial sequencing, and confirmed the presence of the high confidence variants.
369 These data contribute to an argument for transmission of minority variants; however, these
370 conclusions are limited by the sample size and by the lack of metadata supporting the potential
371 for transmission and we would caution against using minority variants alone to determine
372 transmission between individuals. Future studies with large data sets and more in-depth metadata
373 from contact tracing would help to further these conclusions. We also found variants that are

374 shared between many samples in our data set. These variants are shared between samples that
375 are dispersed across the consensus tree and therefore are unlikely to be shared through
376 transmission events. Instead, these variants are likely the product of *de novo* mutations, perhaps
377 in regions of the genome that have an increased tolerance for mutation. In our analyses, we did
378 not find a significant relationship between sites with widely shared minority variants and frequently
379 mutated positions on the tree, though a large sample set would be necessary to explore this
380 further. This phenomenon has been previously suggested for widely shared variants in SARS-
381 CoV-2 infection, and the proposal of mutational hotspots within RNA virus genomes is also
382 substantiated (36, 39, 45).

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

389

390

391

392

393

394

395

396

397

398

399

400 **MATERIALS AND METHODS**

401 RNA extraction and SARS-CoV-2 quantification

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

412

413 Reverse transcription and generation of amplicons

414 Amplification of the viral genome was performed using a modified version of the ARTIC
415 consortium protocol for nCoV-2019 sequencing (<https://artic.network/ncov-2019>) and the
416 methods described in Gonzalez-Reiche *et al.* (8). Briefly, RNA extracted from patient samples
417 was reverse transcribed and subsequently amplified using the Superscript III one-step RT-PCR
418 system with Platinum Taq DNA Polymerase (Thermo-Fisher, 12574018) using nested cycling
419 conditions. Cycling conditions were as follows: 45 $^{\circ}$ C for 60' for RT, 94 $^{\circ}$ C for 2', followed by 12
420 cycles of 94 $^{\circ}$ C for 15 s, 55 $^{\circ}$ C for 30 s and 68 $^{\circ}$ C for 8 min; followed by 35 cycles of 94 $^{\circ}$ C for 15s,
421 55 $^{\circ}$ C for 30 s and 68 $^{\circ}$ C for 2 min 30 s; 68 $^{\circ}$ C for 5' and an 8 $^{\circ}$ C hold. Each sample was processed
422 with two separate pools of primers, pool A and pool B, resulting in alternating and overlapping
423 amplicons that cover the SARS-CoV-2 genome (**Table 2**). Gel electrophoresis was used to
424 confirm amplification of a 2 kb product.

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

Primer Set A		
	Forward Primer (5'-3')	Reverse Primer (5'-3')
A1	CCAGGTAACAAACCAACCACTTT	GCCACTGCGAAGTCAACTGAACA
A2	TGGAACCTTACACCAGTTGTTTCAGAC	AGCATCTTGTAGAGCAGGTGGA
A3	AAACCGTGTGGTACTAATTATATGCCTT	TCACGAGTGACACCACCATCAA
A4	ACGGTCTTTGGCTTGATGACGT	TTTGACCGTGATGCAGCCATGC
A5	GCTAAATTCCTAAAACTAATTGTTGTCGC	GCGGACATACTTATCGGCAATTTTGTTA
A6	TGTTGGTGATTATTTTGTGCTGACAT	CGCTTAACAAAGCACTCGTGGA
A7	ACCCAGGAGTCAAATGGAAATTGA	CCTGAGGGAGATCACGCACTAA
A8	ACCCATTGGTGCAGGTATATGC	TGCAGTAGCGCGAACAAAATCT
A9	TGTGGCTCAGCTACTTCATTGC	GGCCAGTTCCTAGGTAGTAGAAAT

Primer Set B		
	Forward Primer (5'-3')	Reverse Primer (5'-3')
B1	CTGGAATATTGGTGAACA	GCCGACAACATGAAGACAGTGT
B2	GGTCCAACCTTATTTGGATGGAGCTGAT	AAAACACNTAAAGCAGCGGTTGA
B3	GTCACAACATTGCTTTGATATGGAACG	TGGGCCTCATAGCACATTGGTA
B4	ATTGTGGGCTCAATGTGTCCAG	AGCATAGACGAGGTCTGCCATT
B5	CCTAAATGTGATAGAGCCATGCCT	TGCGAGCAGAAGGGTAGTAGAG
B6	CTGAGCGCACCTGTTGTCTATG	TGAACCTGTTTGCGCATCTGTT
B7	TTCGAAGACCCAGTCCCTACTT	AGTGACACTTGCAGATGCTGGCT
B8	GCTGTAGTTGTCTCAAGGGCTGTTGTT	GCTCCCAATTTGTAATAAGAAAGC
B9	ACTTGTACGCCTAAACGAACA	TAGGCAGCTCTCCCTAGCATTG

426

427 Library Preparation and Sequencing platforms

428 All libraries were prepared using the Nextera XT library preparation kit (Nextera), scaled
429 down to 0.25x of the manufacturer's instructions. Briefly, PCR products were normalized to
430 0.2ng/uL. DNA was then fragmented, tagged, amplified and barcoded (Illumina Nextera DNA dual
431 indexes), cleaned with a 0.9x bead cleanup and pooled at equal molarity. A 0.7x bead cleanup
432 was performed on the final pool and libraries were sequenced on either the Illumina MiSeq or the
433 Illumina NextSeq using either the 2x150 bp or 2x300 bp paired end protocol.

434

435 Generation of simulated data and testing of minority variant callers

436 Reads were simulated using the NEAT v2.0 next generation read simulator (25). First, a
437 mutation model was built using `genMutModel.py` (NEAT) providing the VCF obtained from
438 <https://bigd.big.ac.cn/ncov/variation/statistics?lang=en> (downloaded April 2020) and the NCBI
439 SARS-CoV-2 reference genome (NC_045512.2) as input. An error model was built using
440 `genSeqErrorModel.py` (NEAT) providing paired end reads from a high coverage library within our
441 data set. GC and fragment length models were built using `computeGC.py` and `computeFraglen.py`
442 respectively (NEAT) using the NYU-VC-003 bam file as input and SARS-CoV-2 reference for
443 `computeGC.py`. These four models were then provided to NEAT `genReads.py` along with the
444 reference fasta and a mutation rate of 0.0045 (0.45%) to produce a “golden VCF” file containing
445 ~160 SNPs. Several copies of this golden VCF were made, each with the same variants but with
446 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
447 one VCF was made with random allele frequencies where any given variant was 3x more likely
448 to have an AF < 0.5 than > 0.5. Each VCF was then provided as input to NEAT `genReads.py`
449 along with the reference, error model, fragment length model, GC model, and the following
450 params: ploidy = 100, read length = 150, coverage = 100,000, and mutation rate = 0 in order to
451 use only variants in the VCF to simulate paired end fastq libraries with SNPs from the original
452 golden VCF file inserted at varying allele frequencies as encoded in the individual VCF files.

453 Each set of simulated paired end fastq libraries was then down-sampled at the following
454 fractions: 0.1 (~10000X), 0.01 (~1000X), 0.001 (~100X), 0.0001 (~10X), 0.00001 (~1X), 0.002
455 ~(200X), 0.003 (~300X), 0.005 (~500X) using `seqtk v1.2-r94` (<https://github.com/lh3/seqtk>) and a
456 different seed for each down-sampling process to create different fastq files with varying levels of
457 coverage from the original data.

458 Each pair of downsampled fastq files, along with the original, was quality and adapter
459 trimmed using `trimmomatic v0.36` with the following parameters:
460 `ILLUMINACLIP:adapters.fa:2:30:10:8:true LEADING:20 TRAILING:20 SLIDINGWINDOW:4:20`
461 `MINLEN:20` (47). The trimmed reads were aligned to the Wuhan-Hu-1 SARS-CoV-2 reference

462 genome (NC_045512.2) using BWA mem v0.7.17 with the -K parameter set to 100000000 for
463 reproducibility and -Y to use soft clipping for supplementary alignments (48). Duplicates were
464 marked using GATK MarkDuplicatesSpark v4.1.7.0 (24).

465 Variants were called using six separate methods:

- 466 1. GATK Mutect2 v4.1.7.0 with default parameters. Variants were then filtered using
467 GATK FilterMutectCalls v4.1.7.0 (19).
- 468 2. Freebayes v1.1.0-54-g49413aa with ploidy set to 1 and a minimum allele frequency (-
469 F) set to 0.01 (note: freebayes default ploidy is 2, -F is 0.2) (20).
- 470 3. Our in-house pipeline, *timo*, with the minor variant frequency cutoff (-c) option set to
471 0.001, and the coverage cutoff (-C) option set to 1.
- 472 4. VarScan v2.4.2 with -min-coverage set to 1 and -min-var-freq set to 0.01. The input
473 for VarScan was piped from the output of samtools mpileup using the default
474 parameters. VarScan generates a .snp file, which we parse into a VCF file (21).
- 475 5. iVAR v1.2.3 using the default parameters and the minimum frequency (-t) option set
476 to 0.001. The input to ivar is also piped from the output of mpileup using the options
477 -aa -A -d 0 -B -Q 0. The 'PASS' field in the output of iVar was ignored in generation of
478 the vcf files (22).
- 479 6. GATK HaplotypeCaller v4.1.7.0 with the -ploidy option set to 100. This generates a vcf
480 with both snps and indels. GATK selectVariants was used to extract just the snps from
481 these files (23, 24).

482 Intersections between the workflow VCF files (produced by Mutect2, Freebayes, timo, VarScan,
483 iVar and haplotype caller) and the golden VCF file were generated using bcftools isec v1.9 (48).

484 The output from bcftools isec was then analyzed and compared against the respective AF-
485 specific golden VCF to compare allele frequencies using a custom script.

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

487

488 Assembly of genomes and consensus sequences

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

504

505 Identification of minority variants

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

513

514 Generation of phylogenetic trees

515 Isolates of human SARS-CoV-2 were retrieved from the GISAID EpiCov database as of 2020-10-
516 15 (52). The of 3' and 5' regions of sequences were truncated, and sequences containing more
517 than 1% ambiguous sites or those which had an incomplete collection date annotation were
518 removed. From the remaining set we sampled randomly up to 1000 isolates per month leaving
519 10932 isolates as representatives of the global population (**Supplementary Table 2**). The
520 sequences were aligned with MAFFT v7.467 (53) to a reference isolate from GenBank (54)
521 (Accession: MN908947, Wuhan-Hu-1, isolate collected on December 19th 2019 in Wuhan,
522 China). This alignment of the selected 10932 isolates, including the consensus sequences from
523 the NYU Langone samples, was used to infer the maximum likelihood phylogeny under the
524 nucleotide substitution model GTR+G in IQTree (55). The tree topology was assessed using the
525 ultrafast bootstrap function with 1000 replicates (56). To root the tree, we specified the reference
526 isolate hCoV-19/Wuhan/Hu-1/2019 (GISAID-Accession: EPI_ISL_402125), which is identical in
527 sequence to the GenBank isolate used in the alignment step. We inferred the sequences of
528 internal nodes, the optimized timing of internal nodes and resolved polytomies on the final ML-
529 Tree with TreeTime (57). We used a fixed clock rate of 8×10^4 (stdev = 4×10^4) mutations/
530 (bp day) under a skyline coalescent tree prior and we rooted the tree using the same reference
531 isolate as with the IQTree step of topology reconstruction (GISAID-Accession: EPI_ISL_402125)
532 (58). The clock rate was computed as the total number of mutations on the tree, divided by the
533 total length of branches of the timed tree. This rate was optimized by iterative runs of TreeTime
534 until convergence. The time of the root of the tree is estimated to December 19, 2019.

535

536 Identification of circulating clades

537 We characterize the main genetic clades by identifying non-synonymous amino-acid mutations
538 that originate prevalent viral population subtrees. We computed global population clade frequency

539 as follows: (1) Individual isolates, which we index with i , are assigned a smoothed multiplicity
540 factor, $n_i(t) = \exp[-(t - t_i)^4 / (2\sigma^4)]$, where t_i is the collection date of the isolate, and the
541 squared Gaussian kernel is $\sigma = 3$ days. Sample frequencies of isolates are computed as $x_i^s(t) =$
542 $N_i(t) / N(t)$, where t_i is the sampling time and $N(t) = \sum_i N_i(t)$. (2) To correct for regional sequence
543 sampling bias, we computed reweighted frequencies by calibration with the daily incidence data
544 from JHU(59), $x_i(t) = m_{c(i)} x_i^s(t)$, where $c(i)$ is the continent of isolate i . The reweighting factors
545 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
546 c , which is obtained from the JHU data on a given date t (59). We use the following broad
547 geographical regions: USA East Coast, USA West Coast, North America remainder, Europe,
548 Asia, China, South America, Africa, and Oceania. (3) From the corrected isolate frequencies, we
549 obtained global clade frequencies $X_\alpha(t) = \sum_{i \in \alpha} x_i(t)$. We kept all clades that have reached a
550 threshold frequency 5% on any day since the start of the epidemic.

551

552 Distance statistics of doublet variants

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

$$556 \quad P_0(d) = \frac{\sum_{k < k'} n_k n_{k'} H(-d_{kk'} + d)}{\sum_{k < k'} n_k n_{k'}}$$

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

560

561 Mutant weight functions

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

$$565 \quad Y_m = Y_{m-1} + x_m(1 - Y_{m-1}) \quad (3)$$

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

571

572 Data Availability

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

575

576 Ethics Statement

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

581

582

583

584

585

586

587 **ACKNOWLEDGEMENTS**

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

595

596

597

598

599

600

601

602

603

604

605

606

607

608

609

610

611

612

613 **FIGURE LEGENDS**

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

625

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

635

636 **Figure 3. Reproducibility of minority variants across sequencing replicates. (A)** UpsetR
637 plots show the shared and unique minority variants identified by *timo* at an allele frequency of >
638 0.5% in replicate amplification/sequencing runs from clinical NS samples. Red numbers below

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

641

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

655

656 **Figure 5. Uniquely shared variants are enriched at close distances on the consensus tree.**

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

665

666 **Figure 6. SARS-CoV-2 transmission droplets are heterogeneous. (A)** Data distributions $\Phi_s(x)$
667 (synonymous mutations, blue), $\Phi_n(x)$ (nonsynonymous mutations, orange), and $\Phi(x) = \Phi_s(x) +$
668 $\Phi_n(x)$ (all mutations, green) are plotted together with fit functions of the form (1) (dashed lines).
669 **(B)** Empirical mutant weight functions $Y_s(x)$ (synonymous mutations, blue), $Y_n(x)$
670 (nonsynonymous mutations, orange), and $Y(x)$ (all mutations, green); see Methods.

671

672

673

674

675

676

677

678

679

680

681

682

683

684

685

686

687

688

689

690 **REFERENCES**

- 691 1. Zhou P, Yang XL, Wang XG, Hu B, Zhang L, Zhang W, Si HR, Zhu Y, Li B, Huang CL, Chen HD,
692 Chen J, Luo Y, Guo H, Jiang RD, Liu MQ, Chen Y, Shen XR, Wang X, Zheng XS, Zhao K, Chen
693 QJ, Deng F, Liu LL, Yan B, Zhan FX, Wang YY, Xiao GF, Shi ZL. 2020. A pneumonia outbreak
694 associated with a new coronavirus of probable bat origin. *Nature* 579:270-273.
- 695 2. Wu F, Zhao S, Yu B, Chen YM, Wang W, Song ZG, Hu Y, Tao ZW, Tian JH, Pei YY, Yuan ML,
696 Zhang YL, Dai FH, Liu Y, Wang QM, Zheng JJ, Xu L, Holmes EC, Zhang YZ. 2020. A new
697 coronavirus associated with human respiratory disease in China. *Nature* 579:265-269.
- 698 3. Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, Zhao X, Huang B, Shi W, Lu R, Niu P, Zhan
699 F, Ma X, Wang D, Xu W, Wu G, Gao GF, Tan W, China Novel Coronavirus I, Research T.
700 2020. A Novel Coronavirus from Patients with Pneumonia in China, 2019. *N Engl J Med*
701 382:727-733.
- 702 4. Coronaviridae Study Group of the International Committee on Taxonomy of V. 2020. The
703 species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV
704 and naming it SARS-CoV-2. *Nat Microbiol* 5:536-544.
- 705 5. Organization WH. 10/02/2020 2020. WHO Coronavirus Disease (COVID-19) Dashboard.
706 <https://covid19.who.int/>. Accessed 10/01/2020.
- 707 6. Bedford T, Greninger AL, Roychoudhury P, Starita LM, Famulare M, Huang ML, Nalla A,
708 Pepper G, Reinhardt A, Xie H, Shrestha L, Nguyen TN, Adler A, Brandstetter E, Cho S,
709 Giroux D, Han PD, Fay K, Frazar CD, Ilcisin M, Lacombe K, Lee J, Kiavand A, Richardson M,
710 Sibley TR, Truong M, Wolf CR, Nickerson DA, Rieder MJ, Englund JA, Hadfield J, Hodcroft
711 EB, Huddleston J, Moncla LH, Muller NF, Neher RA, Deng X, Gu W, Federman S, Chiu C,

- 712 Duchin J, Gautom R, Melly G, Hiatt B, Dykema P, Lindquist S, Queen K, Tao Y, Uehara A,
713 Tong S, et al. 2020. Cryptic transmission of SARS-CoV-2 in Washington State. medRxiv
714 doi:10.1101/2020.04.02.20051417.
- 715 7. Prevention CfDca. 10/01/2020 2020. Coronavirus Disease 2019 (COVID-19).
716 covid.cdc.gov/covid-data-tracker. Accessed 10/1/2020.
- 717 8. Gonzalez-Reiche AS, Hernandez MM, Sullivan MJ, Ciferri B, Alshammary H, Obla A, Fabre
718 S, Kleiner G, Polanco J, Khan Z, Albuquerque B, van de Guchte A, Dutta J, Francoeur N,
719 Melo BS, Oussenko I, Deikus G, Soto J, Sridhar SH, Wang YC, Twyman K, Kasarskis A,
720 Altman DR, Smith M, Sebra R, Aberg J, Krammer F, Garcia-Sastre A, Luksza M, Patel G,
721 Paniz-Mondolfi A, Gitman M, Sordillo EM, Simon V, van Bakel H. 2020. Introductions and
722 early spread of SARS-CoV-2 in the New York City area. *Science* 369:297-301.
- 723 9. Maurano MT, Ramaswami S, Westby G, Zappile P, Dimartino D, Shen G, Feng X, Ribeiro-
724 Dos-Santos AM, Vulpescu NA, Black M, Hogan M, Marier C, Meyn P, Zhang Y, Cadley J,
725 Ordonez R, Luther R, Huang E, Guzman E, Serrano A, Belovarac B, Gindin T, Lytle A, Pinnell
726 J, Vougiouklakis T, Boytard L, Chen J, Lin LH, Rapkiewicz A, Raabe V, Samanovic-Golden
727 MI, Jour G, Osman I, Aguerro-Rosenfeld M, Mulligan MJ, Cotzia P, Snuderl M, Heguy A.
728 2020. Sequencing identifies multiple, early introductions of SARS-CoV2 to New York City
729 Region. medRxiv doi:10.1101/2020.04.15.20064931.
- 730 10. Koyama T, Platt D, Parida L. 2020. Variant analysis of SARS-CoV-2 genomes. *Bull World*
731 *Health Organ* 98:495-504.
- 732 11. Peck KM, Lauring AS. 2018. Complexities of Viral Mutation Rates. *J Virol* 92.

- 733 12. Denison MR, Graham RL, Donaldson EF, Eckerle LD, Baric RS. 2011. Coronaviruses: an RNA
734 proofreading machine regulates replication fidelity and diversity. *RNA Biol* 8:270-9.
- 735 13. Smith EC, Blanc H, Surdel MC, Vignuzzi M, Denison MR. 2013. Coronaviruses lacking
736 exoribonuclease activity are susceptible to lethal mutagenesis: evidence for proofreading
737 and potential therapeutics. *PLoS Pathog* 9:e1003565.
- 738 14. Lewis-Rogers N, Seger J, Adler FR. 2017. Human Rhinovirus Diversity and Evolution: How
739 Strange the Change from Major to Minor. *J Virol* 91.
- 740 15. Sanjuan R, Nebot MR, Chirico N, Mansky LM, Belshaw R. 2010. Viral mutation rates. *J Virol*
741 84:9733-48.
- 742 16. Zhao Z, Li H, Wu X, Zhong Y, Zhang K, Zhang YP, Boerwinkle E, Fu YX. 2004. Moderate
743 mutation rate in the SARS coronavirus genome and its implications. *BMC Evol Biol* 4:21.
- 744 17. Capobianchi MR, Rueca M, Messina F, Giombini E, Carletti F, Colavita F, Castilletti C, Lalle
745 E, Bordi L, Vairo F, Nicastrì E, Ippolito G, Gruber CEM, Bartolini B. 2020. Molecular
746 characterization of SARS-CoV-2 from the first case of COVID-19 in Italy. *Clin Microbiol*
747 *Infect* 26:954-956.
- 748 18. Lythgoe K.A. HM, Ferretti L, et al. 2020. Shared SARS-CoV-2 diversity suggests localised
749 transmission of minority variants. bioRxiv
750 doi:<https://doi.org/10.1101/2020.05.28.118992>.
- 751 19. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, Gabriel S, Meyerson
752 M, Lander ES, Getz G. 2013. Sensitive detection of somatic point mutations in impure and
753 heterogeneous cancer samples. *Nat Biotechnol* 31:213-9.

- 754 20. Garrison EM, Gabor. 2012. Haplotype-based variant detection from short-read
755 sequencing. ArXiv 1207.3907v2.
- 756 21. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, Miller CA, Mardis ER, Ding
757 L, Wilson RK. 2012. VarScan 2: somatic mutation and copy number alteration discovery in
758 cancer by exome sequencing. *Genome Res* 22:568-76.
- 759 22. Grubaugh ND, Gangavarapu K, Quick J, Matteson NL, De Jesus JG, Main BJ, Tan AL, Paul
760 LM, Brackney DE, Grewal S, Gurfield N, Van Rompay KKA, Isern S, Michael SF, Coffey LL,
761 Loman NJ, Andersen KG. 2019. An amplicon-based sequencing framework for accurately
762 measuring intrahost virus diversity using PrimalSeq and iVar. *Genome Biol* 20:8.
- 763 23. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del
764 Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM, Sivachenko AY,
765 Cibulskis K, Gabriel SB, Altshuler D, Daly MJ. 2011. A framework for variation discovery
766 and genotyping using next-generation DNA sequencing data. *Nat Genet* 43:491-8.
- 767 24. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A,
768 Jordan T, Shakir K, Roazen D, Thibault J, Banks E, Garimella KV, Altshuler D, Gabriel S,
769 DePristo MA. 2013. From FastQ data to high confidence variant calls: the Genome Analysis
770 Toolkit best practices pipeline. *Curr Protoc Bioinformatics* 43:11 10 1-11 10 33.
- 771 25. Stephens ZD, Hudson ME, Mainzer LS, Taschuk M, Weber MR, Iyer RK. 2016. Simulating
772 Next-Generation Sequencing Datasets from Empirical Mutation and Sequencing Models.
773 *PLoS One* 11:e0167047.

- 774 26. B K, WM F, al. e. 2020. Spike mutation pipeline reveals the emergence of a more
775 transmissible form of SARS-CoV-2. bioRxiv
776 doi:<https://doi.org/10.1101/2020.04.29.069054>.
- 777 27. Júnior IJM, Polveiro RC, Souza GM, Bortolin DI, Sasaki FT, Lima ATM. 2020. The global
778 population of SARS-CoV-2 is composed of six major subtypes. bioRxiv.
- 779 28. Luria SE, Delbruck M. 1943. Mutations of Bacteria from Virus Sensitivity to Virus
780 Resistance. *Genetics* 28:491-511.
- 781 29. Holland J, Spindler K, Horodyski F, Grabau E, Nichol S, VandePol S. 1982. Rapid evolution
782 of RNA genomes. *Science* 215:1577-85.
- 783 30. Eigen M. 1993. Viral quasispecies. *Sci Am* 269:42-9.
- 784 31. Domingo E, Martin V, Perales C, Grande-Perez A, Garcia-Arriaza J, Arias A. 2006. Viruses
785 as quasispecies: biological implications. *Curr Top Microbiol Immunol* 299:51-82.
- 786 32. Luring AS, Andino R. 2010. Quasispecies theory and the behavior of RNA viruses. *PLoS*
787 *Pathog* 6:e1001005.
- 788 33. Novella IS, Presloid JB, Taylor RT. 2014. RNA replication errors and the evolution of virus
789 pathogenicity and virulence. *Curr Opin Virol* 9:143-7.
- 790 34. Fitzsimmons WJ, Woods RJ, McCrone JT, Woodman A, Arnold JJ, Yennawar M, Evans R,
791 Cameron CE, Luring AS. 2018. A speed-fidelity trade-off determines the mutation rate
792 and virulence of an RNA virus. *PLoS Biol* 16:e2006459.
- 793 35. Sapoval N, Mahmoud M, Jochum MD, Liu Y, Elworth RAL, Wang Q, Albin D, Ogilvie H, Lee
794 MD, Villapol S, Hernandez KM, Berry IM, Foox J, Beheshti A, Ternus K, Aagaard KM, Posada
795 D, Mason CE, Sedlazeck F, Treangen TJ. 2020. Hidden genomic diversity of SARS-CoV-2:

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

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

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

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

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

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

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

- 818 M, Zoufaly A, Maritschnik S, Borkovec M, Hufnagl P, Nairz M, Weiss G, Wolfinger MT, von
819 Laer D, Superti-Furga G, Lopez-Bigas N, Puchhammer-Stockl E, Allerberger F, Michor F,
820 Bock C, Bergthaler A. 2020. Genomic epidemiology of superspreading events in Austria
821 reveals mutational dynamics and transmission properties of SARS-CoV-2. *Sci Transl Med*
822 12.
- 823 42. Wang DX, Wang YQ, Sun WY, Zhang L, Ji JK, Zhang ZY, Cheng XY, Li YM, Xiao F, Zhu AR,
824 Zhong B, Ruan SC, Li JD, Ren PD, Ou ZH, Xiao MF, Li M, Deng ZQ, Zhong HZ, Li FQ, Wang
825 WJ, Zhang YW, Chen WJ, Zhu SD, Xu X, Jin X, Zhao JX, Zhong NS, Zhang WW, Zhao JC, Li
826 JH, Xu YH. 2021. Population Bottlenecks and Intra-host Evolution During Human-to-
827 Human Transmission of SARS-CoV-2. *Frontiers in Medicine* 8.
- 828 43. Lythgoe KA, Hall M, Ferretti L, de Cesare M, MacIntyre-Cockett G, Trebes A, Andersson M,
829 Otecko N, Wise EL, Moore N, Lynch J, Kidd S, Cortes N, Mori M, Williams R, Vernet G,
830 Justice A, Green A, Nicholls SM, Ansari MA, Abeler-Dorner L, Moore CE, Peto TEA, Eyre
831 DW, Shaw R, Simmonds P, Buck D, Todd JA, Oxford Virus Sequencing Analysis G, Connor
832 TR, Ashraf S, da Silva Filipe A, Shepherd J, Thomson EC, Consortium C-GU, Bonsall D, Fraser
833 C, Golubchik T. 2021. SARS-CoV-2 within-host diversity and transmission. *Science*
834 doi:10.1126/science.abg0821.
- 835 44. Wang Y, Wang D, Zhang L, Sun W, Zhang Z, Chen W, Zhu A, Huang Y, Xiao F, Yao J, Gan M,
836 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
837 Z, Zhong H, Xu X, Song T, Mok CKP, Peiris M, Zhong N, Zhao J, Li Y, Li J, Zhao J. 2021. Intra-
838 host variation and evolutionary dynamics of SARS-CoV-2 populations in COVID-19
839 patients. *Genome Med* 13:30.

- 840 45. Cheung PP, Rogozin IB, Choy KT, Ng HY, Peiris JS, Yen HL. 2015. Comparative mutational
841 analyses of influenza A viruses. *RNA* 21:36-47.
- 842 46. CDC. 2020. CDC 2019-Novel Coronavirus (2019-nCoV) Real-Time RT-PCR Diagnostic Panel.
- 843 47. Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina
844 sequence data. *Bioinformatics* 30:2114-20.
- 845 48. Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler
846 transform. *Bioinformatics* 25:1754-60.
- 847 49. Anonymous. 2018. Picard toolkit, *on* Broad Institute.
848 <http://broadinstitute.github.io/picard/>. Accessed 2020-10-02.
- 849 50. Galanti LS, Dennis; Gunsalus, Kristin C. . 2017. Pheniqs: Fast and flexible quality-aware
850 sequence demultiplexing. bioRxiv doi:<https://doi.org/10.1101/128512>.
- 851 51. Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM.
852 2012. A program for annotating and predicting the effects of single nucleotide
853 polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118;
854 iso-2; iso-3. *Fly (Austin)* 6:80-92.
- 855 52. Elbe S, Buckland-Merrett G. 2017. Data, disease and diplomacy: GISAID's innovative
856 contribution to global health. *Glob Chall* 1:33-46.
- 857 53. Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7:
858 improvements in performance and usability. *Mol Biol Evol* 30:772-80.
- 859 54. Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. 2017.
860 GenBank. *Nucleic Acids Res* 45:D37-D42.

- 861 55. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A,
862 Lanfear R. 2020. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference
863 in the Genomic Era. *Mol Biol Evol* 37:1530-1534.
- 864 56. Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. 2018. UFBoot2: Improving
865 the Ultrafast Bootstrap Approximation. *Mol Biol Evol* 35:518-522.
- 866 57. Sagulenko P, Puller V, Neher RA. 2018. TreeTime: Maximum-likelihood phylodynamic
867 analysis. *Virus Evol* 4:vex042.
- 868 58. Kingman JFC. 1982. The coalescent. *Stochastic Processes and their Applications* 13:235-
869 248.
- 870 59. Dong E, Du H, Gardner L. 2020. An interactive web-based dashboard to track COVID-19 in
871 real time. *Lancet Infect Dis* 20:533-534.
- 872

Figure 1

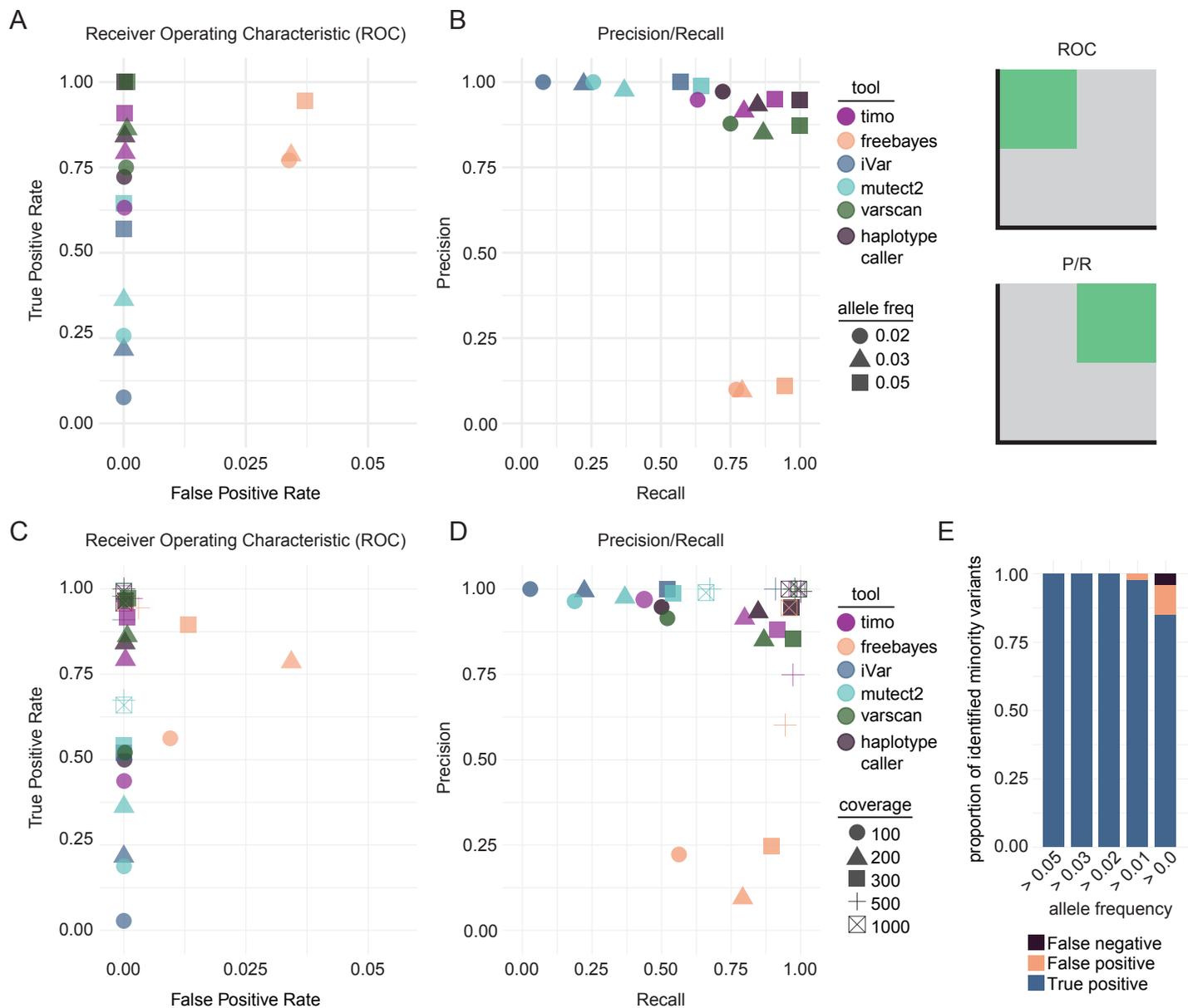


Figure 2

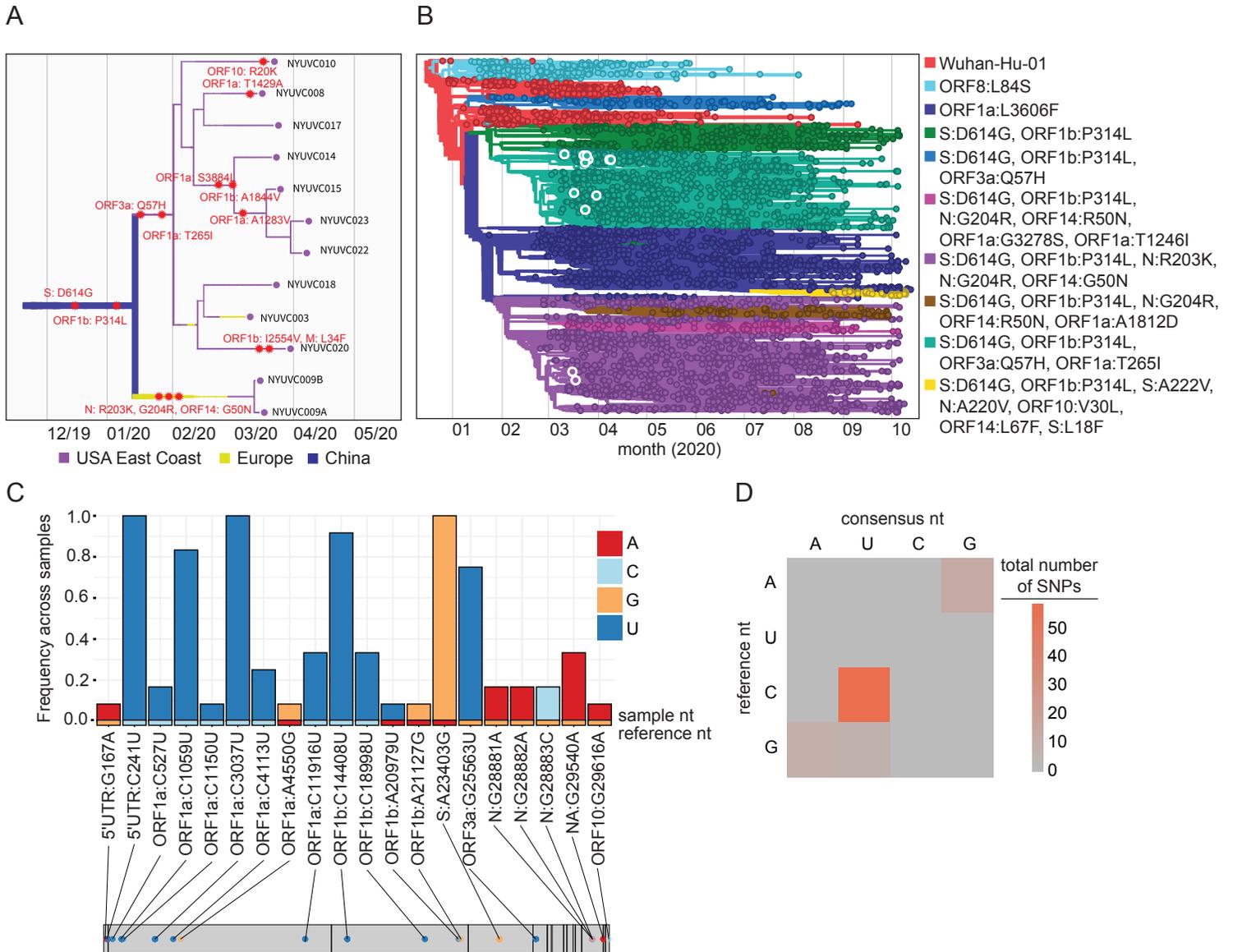


Figure 3

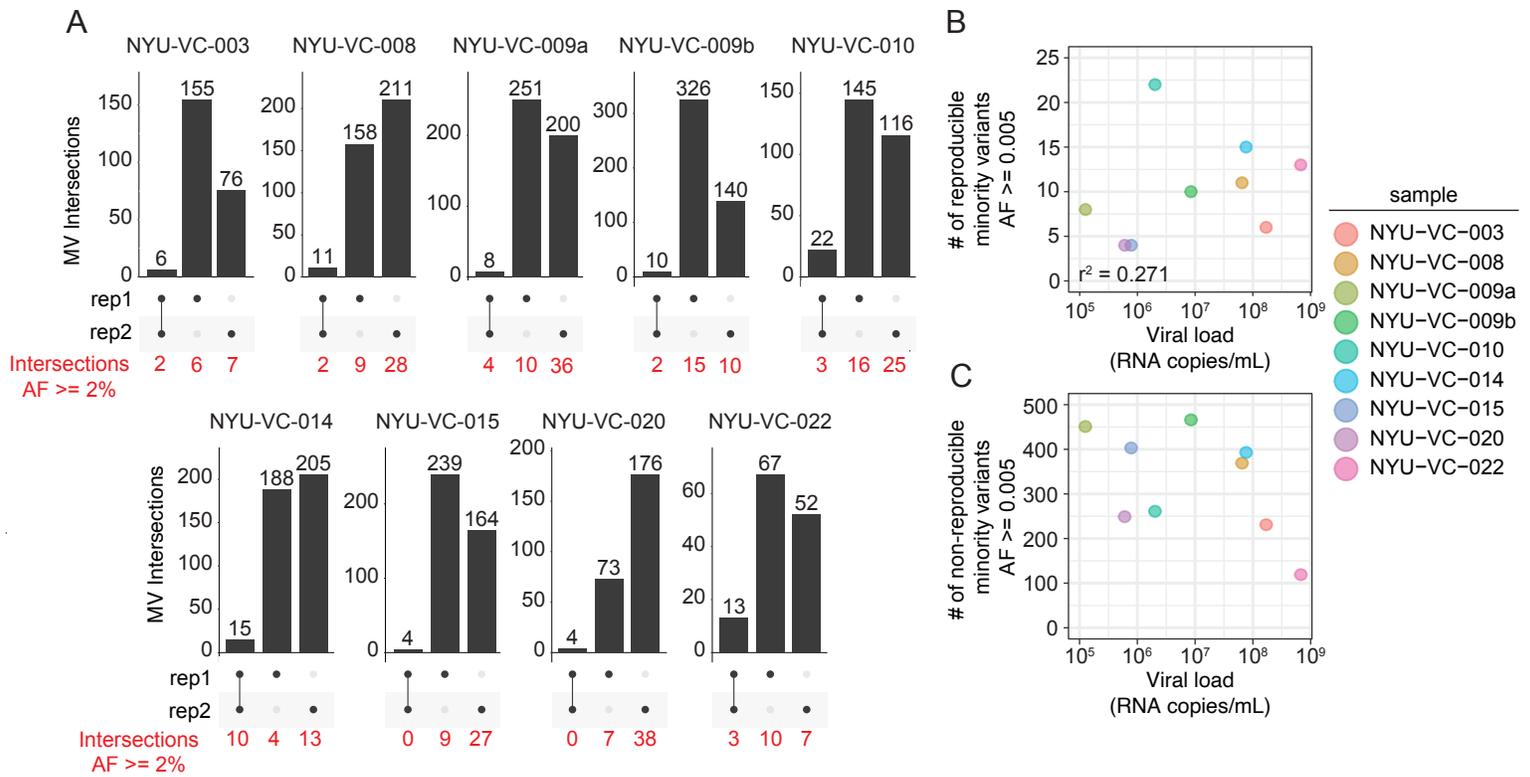


Figure 4

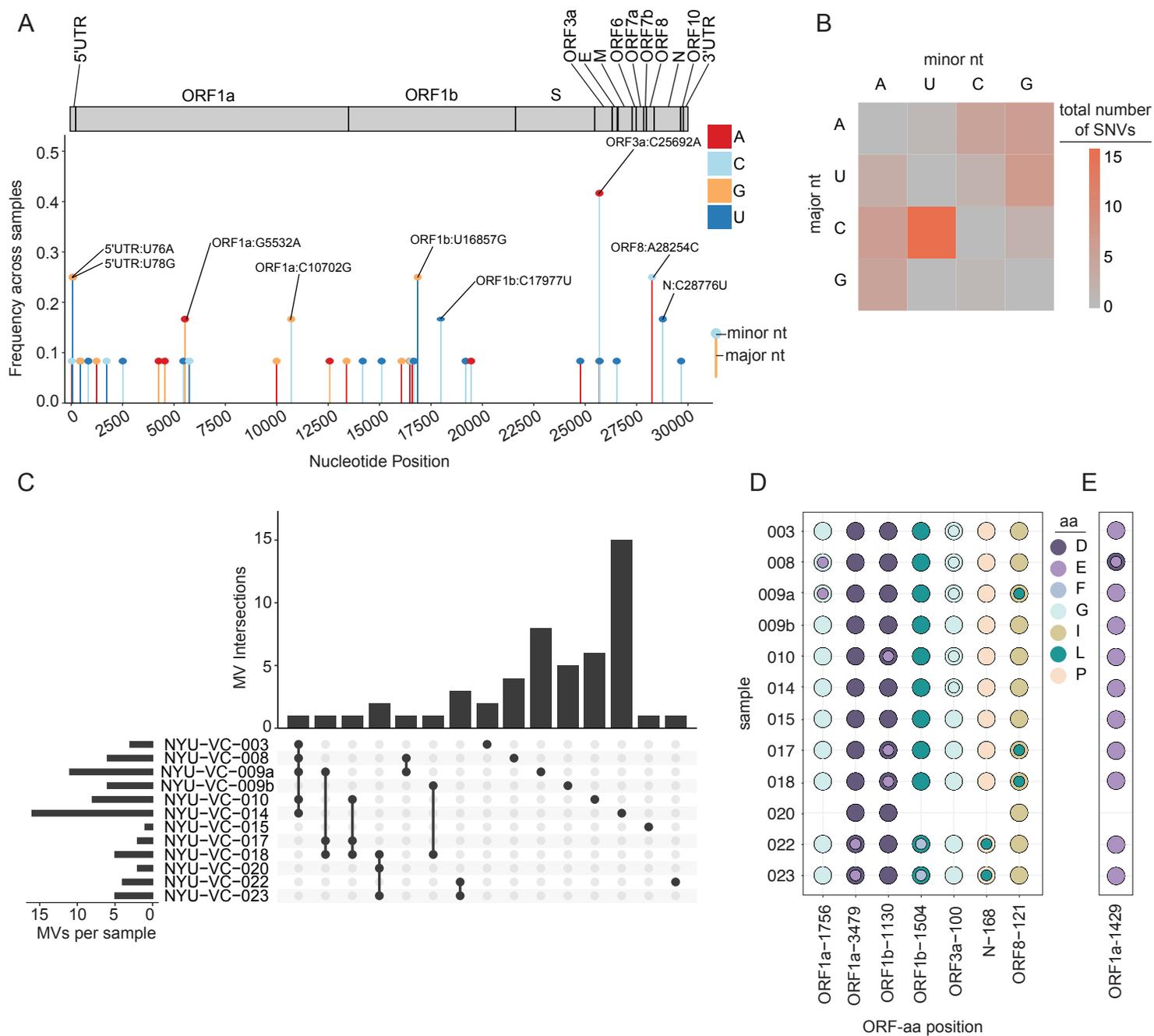


Figure 5

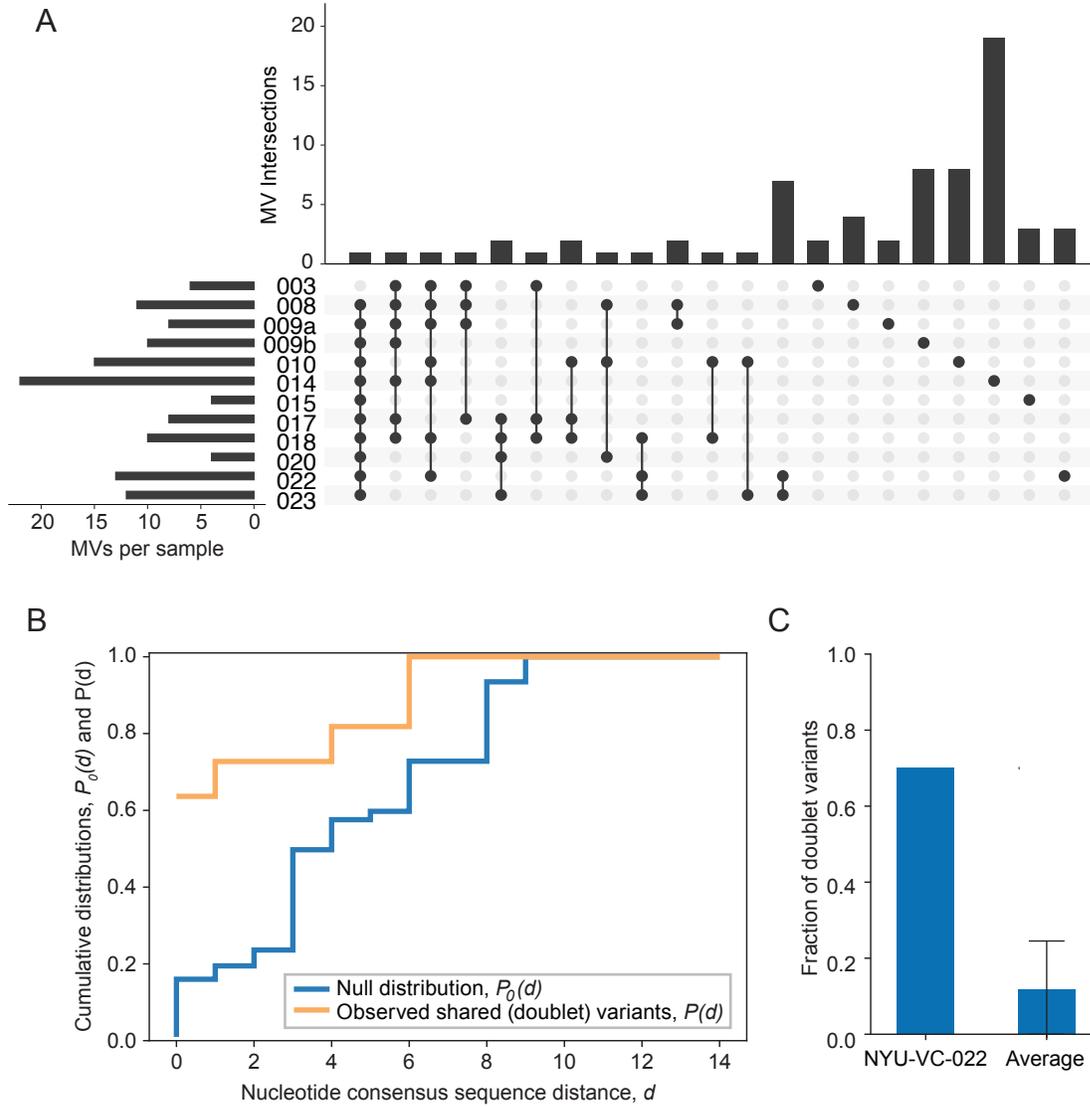


Figure 6

