

Within-host genetic diversity of SARS-CoV-2 in the context of large-scale hospital-associated genomic surveillance

Alexandra A. Mushegian¹, Scott W. Long², Randall J. Olsen², Paul A. Christensen², Sishir Subedi², Matthew Chung¹, James Davis^{3,4}, James Musser², Elodie Ghedin^{1*}

1. Systems Genomics Section, Laboratory of Parasitic Diseases, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, MD USA

2. Laboratory of Molecular and Translational Human Infectious Diseases Research, Center for Infectious Diseases, Department of Pathology and Genomic Medicine, Houston Methodist Research Institute and Houston Methodist Hospital Houston, Texas, 77030

3. Division of Data Science and Learning, Argonne National Laboratory, 9700 S. Cass Ave., Lemont, Illinois, 60439

4. University of Chicago Consortium for Advanced Science and Engineering, 5801 South Ellis Avenue, Chicago, Illinois, 60637

*corresponding author: elodie.ghedin@nih.gov

1 **Abstract**

2 The COVID-19 pandemic has resulted in extensive surveillance of the genomic diversity
3 of SARS-CoV-2. Sequencing data generated as part of these efforts can also capture
4 the diversity of the SARS-CoV-2 virus populations replicating within infected individuals.
5 To assess this within-host diversity of SARS-CoV-2 we quantified low frequency (minor)
6 variants from deep sequence data of thousands of clinical samples collected by a large
7 urban hospital system over the course of a year. Using a robust analytical pipeline to
8 control for technical artefacts, we observe that at comparable viral loads, specimens
9 from patients hospitalized due to COVID-19 had a greater number of minor variants
10 than samples from outpatients. Since individuals with highly diverse viral populations
11 could be disproportionate drivers of new viral lineages in the patient population, these
12 results suggest that transmission control should pay special attention to patients with
13 severe or protracted disease to prevent the spread of novel variants.

14

15

16

17 **Introduction**

18

19 During the COVID-19 pandemic, emerging variants of SARS-CoV-2 have been globally
20 tracked due to the rapid acquisition and sharing of whole genome sequence data¹. As of
21 July 2022, close to 12 million SARS-CoV-2 consensus genome sequences have been
22 deposited to the GISAID repository (<https://www.gisaid.org>). These sequences
23 represent summaries of petabytes of raw sequencing data, which cover the 30Kb RNA
24 genome of SARS-CoV-2 at high redundancy. These deep sequence data could
25 potentially be a rich source of information about the emergence of mutations in the virus
26 population within the infected host, prior to transmission. Because an infected host is a
27 dynamic, heterogeneous environment in which viruses replicate and compete under
28 immunological pressure, it is of great interest to understand how much heterogeneity in
29 the within-host viral population lies beneath the consensus viral genome sequence.

30

31 Studies of within-host viral diversity—variously referred to in terms of minor variants,
32 quasispecies, low-frequency variants, or intrahost single nucleotide variants (iSNVs)—
33 infer the viral diversity within a specimen from the relative abundances of sequencing
34 reads supporting polymorphic sites². Such studies aim to capture de novo mutations
35 acquired by the virus over the course of within-host replication, as well as mixed
36 infections acquired through transmission of multiple lineages. In principle, this
37 information could be used to help predict the emergence of novel variants, to identify
38 sites under evolutionary selection, or to help track transmission³. It is also of interest to
39 determine if patient characteristics, behavior, or differences in clinical care strategies
40 influence the magnitude of viral diversity generated and maintained in individual patients
41 or during transmission. For example, there is evidence that SARS-CoV-2 variants with
42 multiple novel mutations have emerged in patients with protracted infections⁴.

43

44 However, the existing studies also acknowledge that such inferences must be made
45 cautiously. Within-sample read diversity can also be due to sample contamination,
46 especially with aerosolized PCR product; biases generated during reverse transcription,

47 PCR, enrichment, and library preparation steps; sequencing errors; and artefacts
48 generated during bioinformatic processing and read mapping⁵. Well-established
49 methods exist for accounting for these processes when it comes to assembling
50 consensus sequences, but it is considerably more difficult to accurately quantify within-
51 sample variation without making special efforts to counteract these sources of error.
52 Therefore, it is crucial to develop best practices for inferring minor variant diversity from
53 viral deep sequencing data, especially from opportunistic datasets generated with
54 consensus genome sequences as the primary goal and not minor variant analysis.

55

56 We explored the feasibility of extracting actionable signals about within-host viral
57 genetic diversity from the deep sequence data underlying consensus genomes by
58 focusing on a cohort from the Houston Methodist Hospital System. A network of eight
59 hospitals and an associated research institute serving a demographically diverse city of
60 7 million people, HMM began using high-output Illumina instruments to sequence all
61 SARS-CoV-2 specimens coming through the system in December 2020. End-to-end
62 processing of samples, from collection through read generation, occurred within the
63 same set of facilities and protocols with a high level of technical standardization and
64 automation. The resulting consensus sequences were deposited in GISAID and used to
65 track epidemiological trends in Houston⁶⁻⁸. The dataset is unique in that it densely
66 samples a large population over an extended period of time with rich patient metadata
67 linked to samples. However, these same advantages come with challenges from the
68 perspective of minor variant tracking: samples are processed approximately
69 sequentially, not in controlled or randomized batches, and there is limited opportunity in
70 an active high-throughput sequencing facility of this scale to sequence technical
71 replicates.

72

73 Previous studies of minor variants that carefully addressed sources of error and
74 uncertainty have emphasized different aspects of within-host viral diversification⁹⁻¹².
75 Despite methodological differences, several broad observations have been consistent
76 across studies: within-host diversity is generally low, albeit with some outlier samples

77 containing high diversity; and within-host mutations apparently independently recur
78 frequently between samples, impeding attempts to use minor variant information to infer
79 transmission. We used our large dataset to delve more deeply into unanswered
80 questions surrounding these observations. First, we used the patient metadata available
81 to explore whether there are any patient characteristics associated with high minor
82 variant richness, since such individuals might be disproportionate drivers of the
83 emergence of new consensus mutations in an analogous way to a small number of
84 patients driving superspreading events. Second, we examined a set of highly recurrent
85 minor variants to investigate how many were systematic technical artefacts vs.
86 hypermutable sites with potential phenotypic consequences. Theory and empirical data
87 show that biases in de novo generation of mutations can skew evolutionary trajectories,
88 with convergent traits often arising via pathways involving hypermutable sites.^{13,14} We
89 find that while both phenotypically important mutations and probable artefacts regularly
90 recur as minor variants, a robust association between high within-host virus diversity
91 and patient hospitalization (admission into inpatient or ICU care) could be detected. The
92 mechanism and direction of this association is unknown, but this observation supports
93 the conclusion that transmission control in healthcare settings or from severely ill
94 patients should be of particular focus in preventing the emergence of new variants.

95

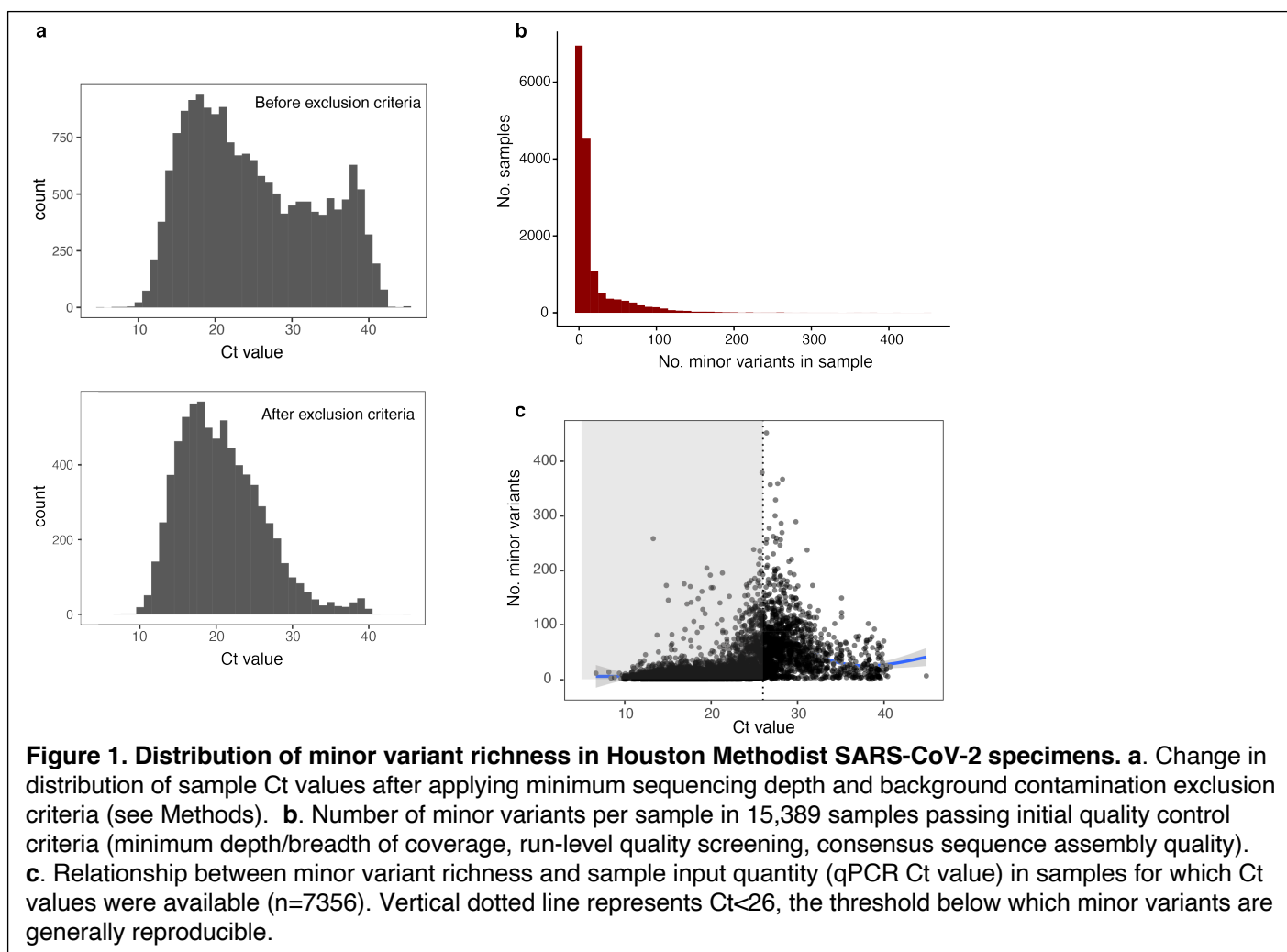
96 **Results**

97

98 *Sample inclusion criteria, minor variant detection and reproducibility*

99 Between the beginning of December 2020 and the end of November 2021, a total
100 39,880 samples were collected at Houston Methodist, encompassing a wide range of
101 symptomatic and asymptomatic patients and healthcare workers. These were
102 sequenced across 70 NovaSeq sequencing runs. We narrowed down the dataset
103 considerably according to various criteria such as the sequencing depth and the input
104 quantity of viral RNA, approximated by the quantitative PCR cycle threshold value, Ct
105 (see *Methods*). Since no-template negative controls were not sequenced, we used
106 Ct \geq 40 samples as pseudo-negative controls to assess the level of background PCR

107 amplicon contamination in each run, reasoning that background contamination of deeply
108 sequenced samples does not necessarily impact consensus sequence calling but can
109 affect the appearance of within-host diversity. We excluded runs containing at least
110 three $Ct \geq 40$ samples in which the coverage breadth and depth were not statistically
111 different from the $Ct < 40$ samples in the run ($t\text{-test} \geq 0.01$). This conservative criterion
112 resulted in the exclusion of 22 sequencing runs. We also limited all our analyses to
113 samples with at least 100x coverage over at least 98% of the genome, and excluded
114 samples where the consensus sequence was flagged as poor quality by Nextclade¹⁵ or
115 that did not have a lineage assigned by Pango¹⁶, and used the earliest available sample
116 from patients from whom multiple samples were collected. These initial filtering criteria
117 narrowed the dataset to 15,389 samples with a lower average Ct value (19.96) than the
118 total population (23.47) (**Fig. 1a**).



119

120 Minor variants were identified using a variant calling pipeline (*timo*), which was
121 previously demonstrated to have high precision and recall of minor variants, given a
122 background rate of sequencing error⁵. The output of this pipeline also differs from many
123 minor variant callers in that it identifies minor variants that are reversions to the
124 ancestral reference allele at sites where the consensus allele differs from the reference.
125 We considered minor variants at sites with a total depth of coverage of at least 100
126 reads, where the minority variant made up at least 1% of reads at the site at a minimum
127 depth of 50 reads (thus requiring a more stringent standard of minimum minor variant
128 frequency at sites with lower coverage). We excluded any minor variants at PCR primer
129 binding sites of either of the primer sets used over the course of the study, and
130 excluded any sites called as gaps or Ns in the consensus or minority fraction. For 54 of
131 the samples, technical replicates re-sequenced from the original RNA were available,
132 which we used to assess how many of the minor variants were reproducible. We found
133 that both the presence/absence and within-host frequency of minor variants were highly
134 reproducible in samples with Ct values <26 (**SuppFig.1 a,b**). This was consistent with
135 the range of input quantities at which minor variants were reproducible in several
136 previous studies^{11,17}. At higher Ct values, many more minor variants failed to be
137 detected in the second replicate of sequencing or were detected but at substantially
138 different frequencies. Sequencing depth either across the whole genome or at individual
139 sites was not clearly associated with reproducibility (**SuppFig.1 c,d**). We thus
140 concluded that minor variants were more likely to be spurious in lower-input samples
141 but were reliably detected in a single replicate of sequencing in higher-input samples,
142 and that sample input amount rather than sequencing depth was a more reliable
143 indicator of sample quality for this purpose.

144

145 *Within-host minor variant diversity*

146 In the 15,389 samples passing the initial quality control criteria, 9,771 (63%) contained
147 minor variants at fewer than 10 nucleotide positions, with a long tail of samples
148 containing much higher diversity (**Fig.1b**). Consistently with previous studies⁹, there was

149 a strong correlation with Ct value, with the most diversity concentrated in moderately
150 low-input samples (Ct~28-30) (**Fig.1c**). There could be technical or biological reasons
151 for higher minor variant richness in lower viral load samples: they are inherently more
152 variable due to stochastic sampling and thus are more sensitive to contamination and
153 technical artefacts, but they also could have been collected early or late in the infection,
154 reflecting different points in the mutation/selection trajectory of the viral population.
155 Given the previous findings about reproducibility, we limited further analysis to samples
156 with Ct<26, acknowledging that although minor variant data from these samples is likely
157 more accurate, this stringent criterion likely affects the composition of the patient cohort
158 studied, since it excludes patients with low viral loads. Because diagnostic PCR was
159 carried out on different instruments across the healthcare system, Ct values were not
160 available for all samples; thus the final dataset contained 6,140 samples. The median
161 number of minor variants in samples in this dataset was 5, with a maximum of 379. A
162 slight positive association with Ct value remained, which is likely due to genuine
163 biological factors in this range of input values (**Fig.1b**, *grey region*).

164
165 The final dataset of samples spanned 47 sequencing runs encompassing several
166 distinct stages of the epidemic. In December 2020 and early January 2021, the
167 dominant consensus sequences were from variant B.1.2 and an assortment of smaller
168 lineages. Starting in late January, a period of declining cases was strongly dominated
169 by the Alpha variant (B.1.1.7), which was replaced in July by the Delta variant
170 (B.1.617.2) in a late summer/autumn surge (**SuppFig.2**). Sample collection dates were
171 roughly, but not completely, chronologically associated with runs (**SuppTable 1**). Even
172 after stringent filtering criteria, there remained a run-level effect on the detection of
173 minor variants, which appeared to be related to the run-level average of sequencing
174 depths rather than individual sample sequencing depths (**SuppFig.3**). Sequencing
175 batch effects are therefore important to consider when assessing minor variant diversity.

176

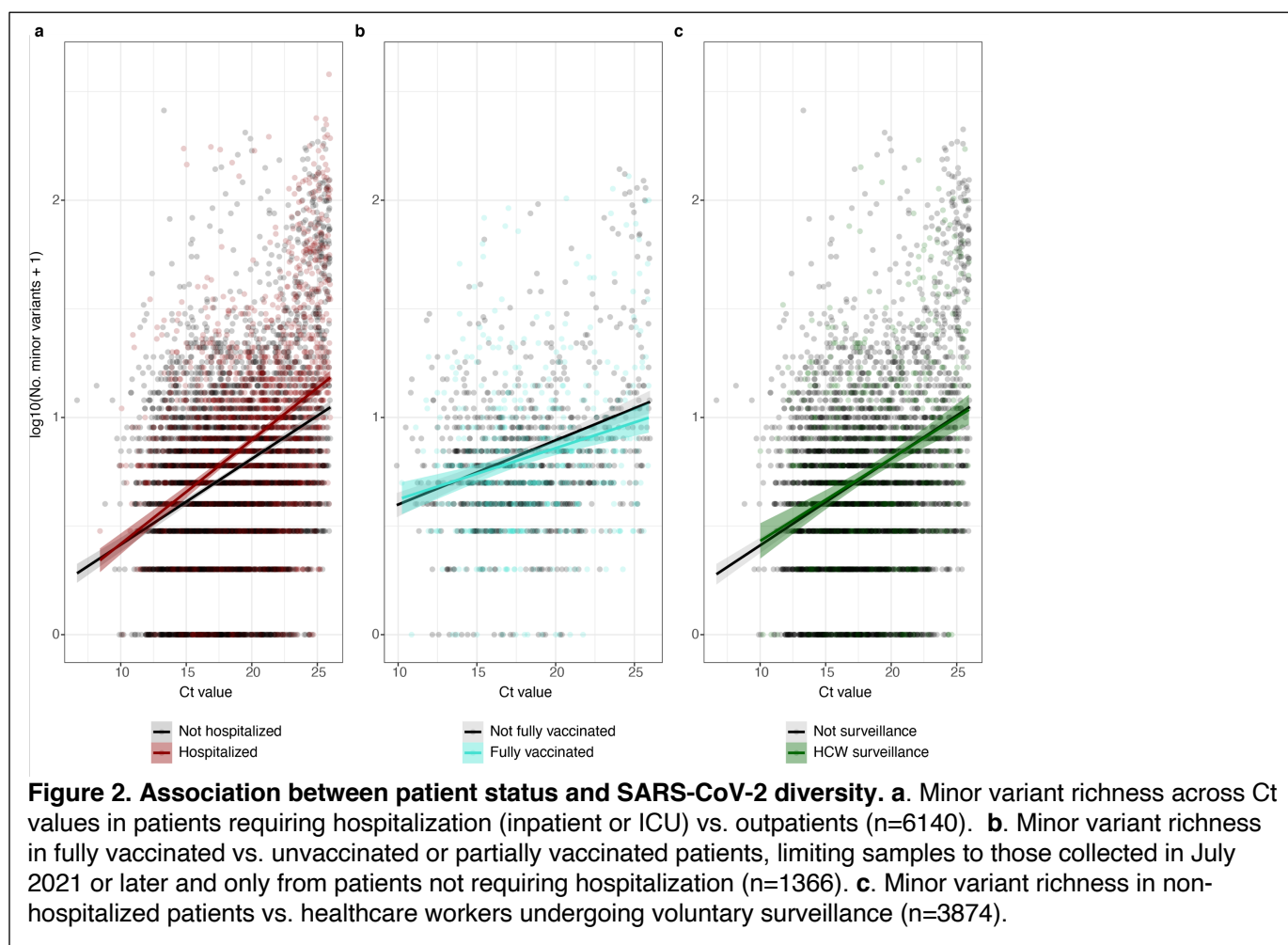
177 *Clinical correlates of within-host diversity*

178 Having controlled as much as possible for artefacts and systematic errors, we next
179 determined if unusually high minor variant richness was associated with any clinical
180 features. From available medical records, we obtained deidentified data on patient
181 demographics (age group, sex, and ethnicity), comorbidities (chronic heart, lung, liver
182 and kidney disease; hypertension; diabetes; obesity; HIV status; previous organ
183 transplant status; cancer), and aspects of clinical treatment (whether the patient was
184 hospitalized and/or was treated with plasma or monoclonal antibodies). Because many
185 of these factors are highly correlated with each other, we used a random forest
186 classification model to query the relative importance of these factors in grouping
187 samples as “high diversity” or “low diversity.” We considered “high diversity” samples to
188 be those with more than 5 minor variants, which was the median number. While the
189 overall performance of the random forest model was poor (ROC AUC=0.58), suggesting
190 that the clinical features included were insufficient to classify high vs. low diversity
191 samples without additional information, the most important variable in the trained model
192 was hospital admission. A chi-squared test confirmed that high-variant samples were
193 overrepresented among hospitalized patients—treated as inpatients or admitted to
194 intensive care—compared to outpatients ($X^2 = 131.58$, $df = 1$, $p < 2.2e-16$). The odds
195 ratio of the association of hospitalization with high minor variant diversity was 1.84 (95%
196 CI 1.66-2.05). There was a higher proportion of hospitalized patients, as compared to
197 non-hospitalized individuals, with at least 1 minor variant, more than 5, or more than 10
198 minor variants in the sample (**SuppFig.4**).

199
200 To take sample viral input amount and sequencing run batch into account, we
201 constructed a linear mixed effects model with hospital admission and sample Ct value
202 as fixed effects, sequencing run as a random effect, and the log-transformed number of
203 minor variants in the sample as the response variable. Hospital admission and Ct value
204 were both significantly associated with minor variant diversity (**Table 1a, Fig.2a**).

205 One plausible reason samples from patients requiring hospitalization could have higher
206 minor variant richness is if they were on average collected later in the infection than
207 samples from non-hospitalized patients. We did not have information on the number of

208 days post infection for any of the samples. As another way to probe the relationship
209 between disease severity and minor variant richness, that is less likely to be related to
210 infection duration, we examined minor variant diversity in samples from July 2021
211 onwards when an appreciable number of vaccine breakthrough cases occurred.
212 Assuming that vaccination was associated with less severe disease even among the
213 population not requiring hospitalization, we compared minor variant richness in the
214 infections of fully vaccinated and unvaccinated or partially vaccinated non-hospitalized
215 individuals. At comparable Ct values, minor variant diversity was significantly higher (p
216 <0.05) in the unvaccinated than in the fully vaccinated cases (**Table 1b, Fig.2b**). Finally,
217 we compared minor variant diversity in samples from healthcare workers undergoing



218 voluntary surveillance testing with samples from non-hospitalized patients. In this case,

219 minor variant richness was not significantly different between these two groups (**Table**
220 **1c, Fig.2c**).

221
222 As a complementary approach to evaluating the association between patient factors,
223 sample characteristics, and minor variant richness, we constructed a LASSO regression
224 model containing the comorbidities, treatments, and demographic factors, as well as Ct
225 values, median sequencing depth, collection month and run. In the best model
226 ($\lambda=0.0013$) the deviance ratio was 0.226, meaning that the combination of
227 variables we included explained approximately 22.6% of the variation in the log-
228 transformed number of variants. We also constructed a version of this model which
229 excluded all factors that had a strong temporal bias (vaccination status, consensus
230 variant, and collection month), because temporal trends in this study design were
231 impossible to separate from sequencing batch (run) effects. In this model (deviance
232 ratio 0.17), hospital admission was clearly associated with the highest increase in minor
233 variant richness (**SuppFig.5**).

234
235 Taken together, these complementary modeling strategies suggest there is substantial
236 unexplained variation in within-host minor variant richness. They also highlight that
237 severity of disease—as exemplified here by hospitalization or lack of vaccination—
238 warrants further study as a correlate of within-host diversity independently of diversity
239 associated with viral load or viral-load-related technical artefacts.

240
241 *Robustness of clinical associations to analytical thresholds*

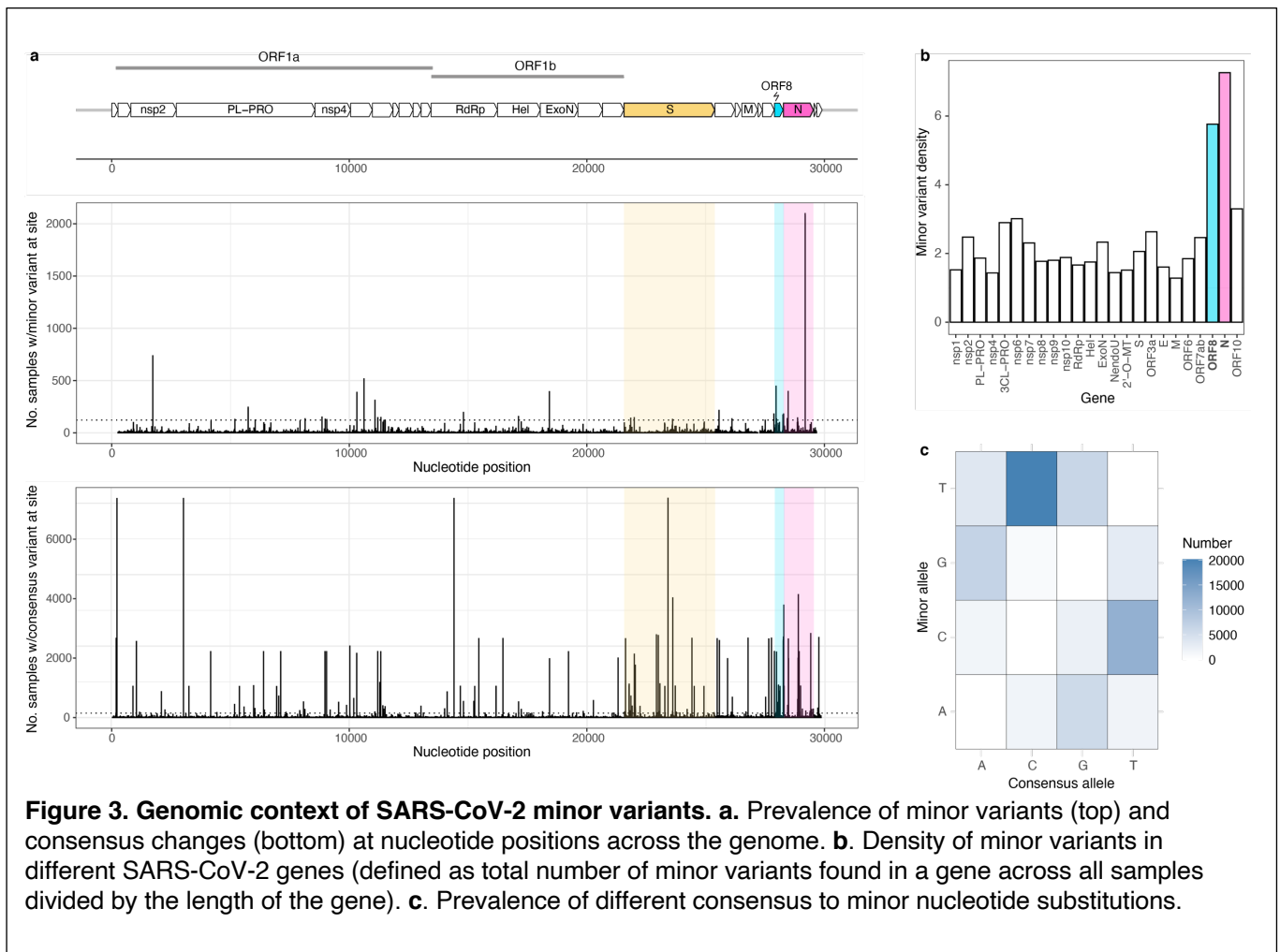
242 To evaluate how sensitive the associations discussed above were to the criteria used to
243 identify minor variants, we generated three additional datasets with different levels of
244 stringency across criteria. In Alternate Dataset 1, we used samples with 200x coverage
245 or more over at least 98% of the genome. We required minor variants to have a
246 minimum of 2% allele frequency (MAF) at sites with at least 200x coverage and be
247 supported by a minimum of 100 reads. In Alternate Dataset 2, we used samples with
248 500x coverage or more over at least 98% of the genome, and required minor variants to

249 have at least 3% MAF supported by a minimum of 20 reads. In Alternate Dataset 3, we
250 used samples with 1000x coverage or more over at least 98% of the genome and minor
251 variants with at least 1% MAF. Because the sample size was significantly reduced, we
252 relaxed the Ct criterion from $Ct < 26$ to $Ct < 35$ for the regression analyses on vaccination
253 status and healthcare worker surveillance, on the assumption that random errors in
254 higher Ct samples would not differ between the groups. Despite the varying sample
255 sizes and minor variant detection limits, all three datasets showed significantly higher
256 minor variant richness in hospitalized patients (**SuppFig.6**). This factor was the most
257 important in all three random forest models, it was in the top three highest coefficients in
258 all three LASSO regression models in which all factors were included, and it had the
259 highest coefficient in all three LASSO models in which temporally-biased factors were
260 excluded. The other two factors of interest – vaccination status and healthcare worker
261 status – differed significantly in the extent of the association depending on the dataset
262 used. Minor variant richness was significantly lower in vaccinated patients and in
263 healthcare workers in Alternate Dataset 1, but only the healthcare worker factor was
264 significant in Alternate Dataset 2, and neither factor was significant in alternate dataset
265 3. The ROC AUC values for the random forest classification model were similar for all
266 three datasets (0.58-0.60), while the fraction of variation explained by the LASSO
267 regression model was slightly higher for alternate dataset 3 (28% for the model
268 including all factors, 16% for the model excluding temporally biased factors). Aside from
269 the hospitalization variable, the coefficients of many factors changed substantially
270 between alternate datasets, even changing sign (for example, plasma treatment and
271 monoclonal antibody treatment were associated with increased or decreased minor
272 variant richness depending on the dataset). We concluded that the thresholds and
273 criteria used to identify minor variants could significantly affect the strength of observed
274 associations, but that the higher richness of minor variants in samples from patients
275 requiring hospitalization was robust.

276

277 *Mutational patterns of highly prevalent minor variants*

278 Having determined that minor variant richness was robustly associated with
 279 hospitalization, we set out to analyze the mutational patterns in the genome. Across all
 280 samples, minor variants were found mostly concentrated in the Orf8 and N genes, an
 281 observation consistent with previous characterization of these genes as
 282 hypermutable^{18,19} (**Fig.3a,b**); this enrichment pattern was similar in samples from
 283 hospitalized and non-hospitalized patients (**SuppFig. 7**). The most common within-host
 284 mutation observed was C>T, consistent with previous studies and with the hypothesis
 285 that nucleic acid editing by host enzymes contributes to the mutational spectrum¹¹
 286 (**Fig.3c**). Surprisingly, C>T mutations were also the most prevalent among non-
 287 reproducible minor variants, despite the fact that C>T mutations are not known to be
 288 common sequencing errors²⁰ (**SuppFig. 8**).



290 We were particularly interested in sites containing minor variants in a high proportion of
291 the samples. Mutational hotspots are of interest due to their potentially important role in
292 convergent evolution. Therefore, a plausible purpose for monitoring minor variants in
293 deep sequencing data is to identify sites with increased probabilities of mutation as a
294 special focus for targeted mutational analysis. Doing so would require the ability to
295 distinguish genuine mutational hotspots from recurrent artefacts.

296
297 We focused on 34 positions in the genome where minor variants were present in at
298 least 2% of samples (**Fig.4a**). Minor variants at most of these positions were found at a
299 range of frequencies from 1% to 50%, with several exceptions (**Fig.4b**). A majority of
300 these sites included samples in which the minor variant was a reversion to the ancestral
301 allele, suggesting repeated mutation at these sites. The gene containing the highest
302 number (8) of highly recurrently mutated sites was N, but another 7 of these sites were
303 found in the proteases PLpro or 3CLpro. These essential enzymes are involved in viral
304 replication and immune modulation and thus high-profile targets for antiviral drug
305 development^{21,22}. This further justifies special attention to the mutational properties of
306 these sites, since responsible development of antivirals ought to consider likely paths to
307 the evolution of resistance, including loci with higher than average standing genetic
308 variation within hosts. We cross-referenced the list of the 34 highly shared positions with
309 highly shared sites from previous studies, with global patterns of consensus SNPs
310 (queried from GISAID using cov-spectrum.org²³), and with known phenotypically
311 important convergent mutations²⁴.

312
313 Several of the highly shared minor variant sites were highly polymorphic on the
314 consensus level both within this dataset and in the US-wide GISAID data. Such sites
315 pose challenges for interpretation because this pattern could be indicative of genuinely
316 hypermutable sites but is also difficult to distinguish from cross-contamination because
317 multiple consensus variants are often present in the same run. Indeed, in many sites
318 that were polymorphic on the consensus level, minor variants only appeared in runs in
319 which multiple consensus variants were present (**SuppFig.9**). One exception was

320

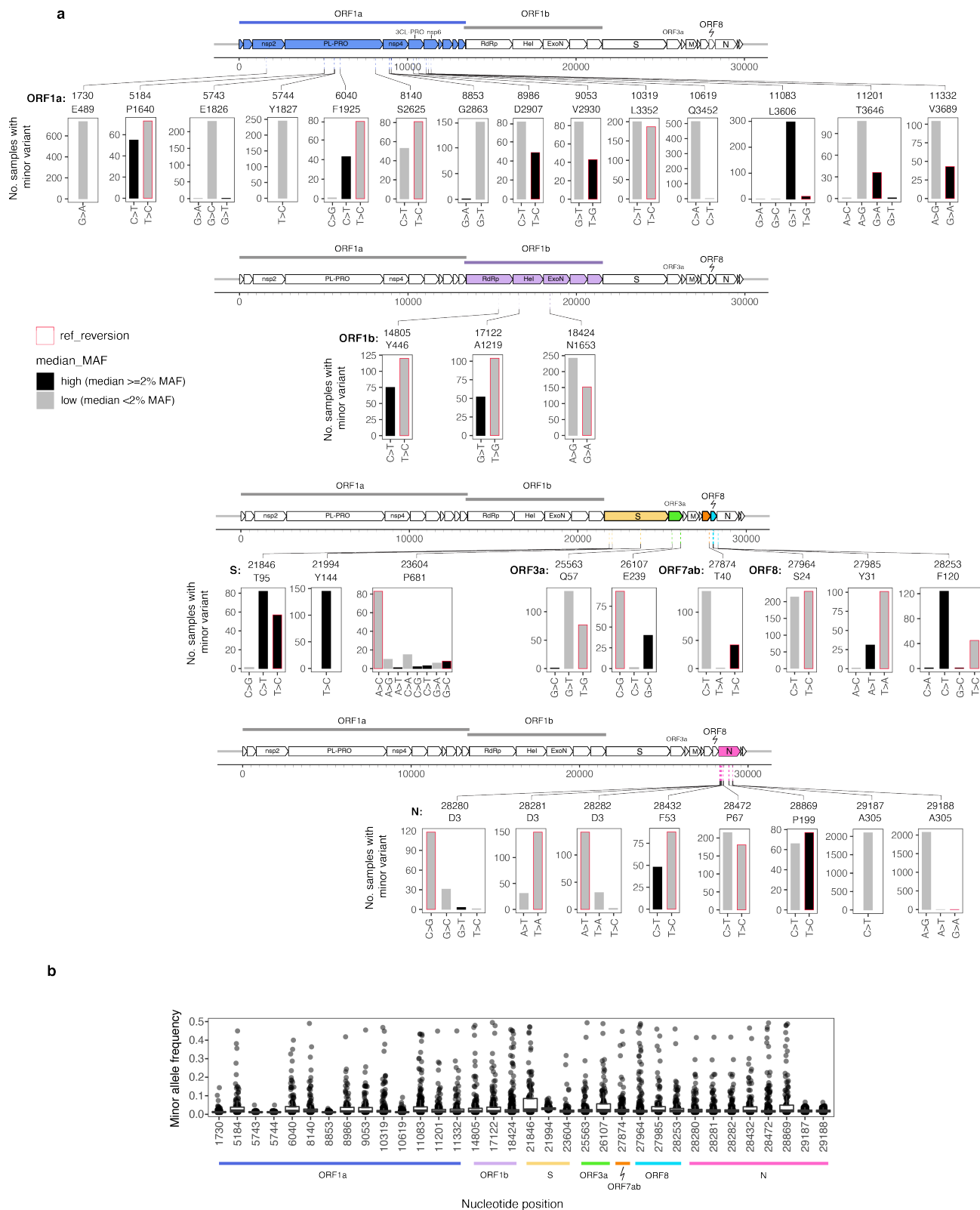


Figure 4. Nucleotide substitutions and minor allele frequencies at highly mutable sites. a. Minor variant changes at 34 nucleotide positions containing minor variants in at least 2% of samples. **b.** Minor allele frequencies of variants at these sites.

321 had all combinations of minor and consensus nucleotides present throughout the study
322 period; and sites 28280-28282 (N: D3), which had two alleles at various frequencies
323 present at each position in the codon throughout the study period. S: P681 has
324 undergone different substitutions in multiple variants of concern; N: D3 contains a whole
325 codon substitution in the Alpha lineage. These patterns are consistent with the
326 explanation that these are highly mutable loci in which mutations have a high likelihood
327 of becoming fixed in a lineage with a fitness advantage, like the variants of concern.

328
329 A somewhat puzzling pattern was observed for the highly prevalent minor variants at
330 sites 1730 (orf1a: E489), 8853 (orf1a: G2863), 10619 (orf1a: Q3452), and 29187-29188
331 (N:A305) (**Fig.4a**). These sites were highly conserved at the consensus level but
332 contained minor variants at low frequencies in a substantial fraction of specimens
333 throughout the study period with a notable increase in the fraction of samples with minor
334 variants in the later part of the study period (starting approximately with run 57/July
335 2021) (**SuppFig.9**). This period corresponds to when the vast majority of sequenced
336 specimens were from the Delta lineages.

337
338 Some of the other sites had evidence suggesting technical or bioinformatic artefacts.
339 For example, nucleotide position 11083 was identified as a highly recurrent minor
340 variant in Lythgoe et al and Tonkin-Hill et al^{10,11}, as well as a highly homoplastic site
341 across the SARS-CoV-2 phylogeny²⁵. It is immediately adjacent to a long T-
342 homopolymer site, a well-known cause of sequencing error. Nucleotide position 21994
343 (S:Y144) is adjacent to a characteristic deletion in the Alpha (B.1.1.7) lineage,
344 suggesting that mis-alignment of reads at the deletion site might contribute to the
345 appearance of minor variants at that position. A particularly anomalous pair of minor
346 variant sites were at positions 29187 and 29188, in the N gene, which were present in
347 more than 30% of the samples studied but never at higher than 7% minor allele
348 frequency. The amino acid mutation that these minor variants corresponded to, N:
349 A305V, was extremely rare on the consensus level among sequences deposited to

350 GISAID (**SuppFig.10**), with only 72 sequences containing this mutation found in the
351 United States, of which 59 were from Houston Methodist Hospital.

352
353 Finally, as another approach to examining whether convergent evolution could be linked
354 to mutational patterns observed in minor variant data, we examined the frequency in the
355 minor variant dataset of mutations repeatedly associated with increases in SARS-CoV-2
356 fitness. Obermeyer et al²⁴ showed that single nucleotide mutations associated with
357 increased transmissibility evolved independently multiple times in different lineages.
358 Several mutations in the Spike protein, namely E484K, N501Y, K417N, and L18F, also
359 independently evolved in several lineages of concern²⁶. We examined the prevalence of
360 these 22 nucleotide substitutions as minor variants in our data. Eight of these were not
361 found as minor variants in any samples, whereas 14 (13 of them in the Spike protein)
362 were found in at least one sample (**SuppFig.11**). The most prevalent of these minor
363 variant mutations was S:T95I, which was at nucleotide position 21846, one of the
364 identified highly shared sites. This mutation, in the N-terminal domain of the Spike
365 protein, has arisen independently in at least 30 consensus lineages and is associated
366 with significant increases in viral fitness, but its phenotypic effect has not, to our
367 knowledge, been experimentally characterized. Given the tendency of this mutation to
368 frequently arise both within hosts and be successfully transmitted between hosts, further
369 characterization of the effects of this change may be warranted to inform design of
370 drugs and antibodies.

371 372 **Discussion**

373 We explored the feasibility of characterizing within-host diversity by extracting minor
374 variants data from clinical genomic surveillance samples of a large, densely sampled
375 population to supplement consensus-level understanding of viral variants. The clearest
376 finding of this study is that although within-host diversity is generally low, higher within-
377 host diversity is associated with patients requiring hospitalization. Previous studies of
378 minor variants in SARS-CoV-2 have consistently identified outlier samples with high
379 numbers of within-host minor variants even after stringent quality control but were

380 unable to examine the implications of such specimens due to their rarity. Exactly such
381 rare events (i.e. patients with highly diverse viral populations) may be disproportionate
382 drivers of viral recombination and transmission of standing genetic variation on which
383 host-mediated natural selection can act. In this dataset, the absolute number of these
384 outliers was high enough that we could test whether any of the host factors were
385 associated with higher viral diversity.

386
387 Every novel consensus variant must at some point arise as a within-host variant, so it is
388 crucial to understand what contexts may be likely to incubate viral population diversity.
389 We found that, despite the noise introduced by variation in sequencing runs and sample
390 Ct values, the signal was strong enough to observe a clear correlation between higher
391 minor variant richness in more severely ill patients (those admitted to inpatient care or
392 the ICU) and lower richness in vaccinated patients, in whom the number of replication
393 cycles is assumed to be constrained. Previously, similar observations had been hinted
394 at in smaller studies comparing cancer patients with healthcare workers, comparing
395 mildly and severely ill patients, and examining minor variants in samples from patients
396 of different ages^{27–29}. Studies in which patients were longitudinally sampled have
397 shown fluctuating numbers of minor variants over time, with little directional trend^{9,30}.
398 Combined with our data, this suggests that within-host diversity is temporally dynamic,
399 but in the aggregate is more likely to be high in more severely ill patients. This has
400 implications for infection control strategies, for example bolstering the case that
401 transmission control in healthcare settings and among symptomatically infected patients
402 should be a critically high priority for preventing the emergence of new viral variants.

403
404 The direction of causality for the association between severe disease and high within-
405 host diversity is unclear. It is plausible that more severe disease is the result of a more
406 prolonged or quickly-replicating infection, during which more mutations can
407 accumulate³¹; but, conversely, it is possible that more diverse infections drive more
408 severe disease³². It is also possible that the immune responses during severe disease
409 are distinctive in ways that affect selection on the viral population, or that there are

410 mechanistic links between the comorbidities associated with risk of hospitalization and
411 the dynamics of immune selection. For example, obesity is associated with more
412 negative outcomes in influenza, and a study of influenza virus diversity in mice found
413 that influenza A virus replicated faster and accumulated more diversity in obese than in
414 lean mice, an effect that appeared to be mediated by the differential robustness of the
415 interferon response³³. Future studies should conduct case-control comparisons of
416 technically replicated longitudinal samples of patients with different disease time-
417 courses to understand how virus population diversity changes over time in different
418 types of hosts, under different immunological and clinical conditions. Such dynamics are
419 well understood in viruses such as HIV, but may be more complex in SARS-CoV-2,
420 which appears to elicit highly heterogeneous immune responses in different patients³⁴.

421
422 In multiple previous studies of SARS-CoV-2 within-host diversity, mutational hotspots
423 have hampered attempts to use minor variants to track transmission of closely related
424 lineages^{11,35}, since it is difficult to distinguish hypermutable sites, recurrent artefacts,
425 and sites that are genuinely co-transmitted. Notably, in Tonkin-Hill et al, there was no
426 correlation between the probability of transmission between two patients and the
427 number of minor variants they shared; samples that were epidemiologically distant from
428 each other often had more than 10 minor variants in common. This was true even after
429 excluding minor variants that were generally highly prevalent in the dataset. Other
430 authors have pointed out that automatically excluding minor variants that are present in
431 many samples is not always warranted in epidemiological inference, for example when
432 examining a superspreading event in which a large group of people were interacting
433 and transmitting virus for a substantial length of time³⁵. In general, it appears that
434 convergent within-host *de novo* mutation is common enough to significantly complicate
435 inferences of transmission of within-host diversity. For these reasons, it was warranted
436 to play closer attention to the characteristics of sites containing minor variants in many
437 samples. We found evidence both for highly recurrent artefacts and for phenotypically
438 important recurrent mutations, the latter of which may be a high priority for targeted
439 mutational studies.

440

441 Our observations suggest that identifying genuine mutational hotspots requires both
442 understanding the genomic context (noting adjacent deletions, homopolymers or other
443 structural features that might affect spurious minor variant calling) and also the wider
444 sequencing context. For example, the increase in the prevalence of several mutations in
445 the same later runs is difficult to explain. Although it is plausible that different lineages
446 would have different within-host mutation rates, it is more difficult to imagine a
447 mechanism by which certain lineages would have elevated rates of mutation only at
448 specific sites. It is also difficult to rule out that some unknown technical change in
449 sequencing conditions also contributed to changes in relative prevalence at these sites.
450 Similarly, we suspect that the high prevalence of minor variants at sites 29187-29188
451 may be an artefact of the specific combination of methods used at this sequencing
452 facility, because consensus variants at this position were very rare in consensus
453 sequences from GISAID and primarily came from Houston Methodist, in samples with
454 very different genetic backgrounds and collection dates. The existence of consensus
455 mutations specific to particular sequencing labs was noted early in the pandemic²⁵, so
456 caution when detecting unusual mutations on the minor variant level is particularly
457 warranted.

458

459 One of the purposes of this study was to examine the general feasibility of extracting
460 minor variant data from samples not collected for this purpose. Despite the exceptional
461 level of quality control involved in the generation of our sequences in the clinical
462 context, we found that unavoidable technical artefacts, in particular batch effects and
463 the clear effect of RNA input quantity on minor variant calling—even when limiting
464 samples to those with high coverage across the genome—hampered the ability to draw
465 definitive conclusions in the absence of technical replicates and required us to limit our
466 analyses to high input samples. Our results demonstrate that caution is required when,
467 for example, analyzing minor variants from sequence data repositories^{36,37}. Rates of
468 different types of error may meaningfully differ between batches of samples such as
469 sequencing runs, between laboratory protocols and even due to factors such as

470 whether individual tubes or plates were used in library preparation ^{38,39}. If the
471 idiosyncratic physical conditions under which library preparation occurs differentially
472 affect error rates, then developing universally applicable error models may be extremely
473 difficult. At the same time, the composition of batches may be non-random in
474 biologically meaningful ways (e.g. samples from an outbreak in a particular location or
475 population are likely to be sequenced in the same batch), making it difficult to
476 disentangle biological and technical causes of patterns of minor variant prevalence.
477 These results show that, without the development of more mature methods for
478 correcting for numerous different sources of technical noise, deep sequence data
479 cannot be used for routine monitoring of within-host viral diversity in the same way that
480 consensus sequences are used for genomic surveillance.

481
482 Targeted studies of within-host diversity that take these technical issues into account
483 can, however, lead to a greater understanding of the mutational biases of the virus and
484 characteristics of the within-patient environment that affect viral diversification. In our
485 exploratory study, the clearest emergent signal is that infections with high virus diversity
486 are enriched among hospitalized patients. This has clear implications for prioritizing
487 transmission control in healthcare settings and Further dissection of within-host viral
488 dynamics is required to determine whether knowledge of a patient's viral population
489 diversity can better inform clinical care.

490

491 **Methods**

492

493 *Patient population and ethics*

494 The work was approved by the Houston Methodist Research Institute Institutional
495 Review Board (IRB1010-0199). Specimens from patients were obtained primarily from
496 symptomatic patients with a suspicion for COVID-19 disease from outpatient,
497 emergency, labor and delivery, and other types of clinics. Specimens from healthcare
498 workers were collected as part of a non-mandatory workplace surveillance program.
499 Specimens were tested in the Molecular Diagnostics Laboratory at Houston Methodist

500 Hospital using assays granted Emergency Use Authorization (EUA) from the FDA
501 ([https://www.fda.gov/medical-devices/emergency-situations-medical-devices/faqs-](https://www.fda.gov/medical-devices/emergency-situations-medical-devices/faqs-diagnostic-testing-sars-cov-2#offeringtests)
502 [diagnostic-testing-sars-cov-2#offeringtests](https://www.fda.gov/medical-devices/emergency-situations-medical-devices/faqs-diagnostic-testing-sars-cov-2#offeringtests), last accessed June 7, 2021). Standardized
503 specimen collection methods were used (<https://vimeo.com/396996468/2228335d56>).
504 Multiple molecular testing platforms were used, including the COVID-19 test or RP2.1
505 test with BioFire Film Array instruments, the Xpert Xpress SARS-CoV-2 test using
506 Cepheid GeneXpert Infinity or Cepheid GeneXpert Xpress IV instruments, the Cobas
507 SARS-CoV-2 & Influenza A/B Assay using the Roche Liat system, the SARS-CoV-2
508 Assay using the Hologic Panther instrument, the Aptima SARS-CoV-2 Assay using the
509 Hologic Panther Fusion system, the Cobas SARS-CoV-2 test using the Roche 6800
510 system, and the SARS-CoV-2 assay using Abbott Alinity instruments.

511

512 *Library preparation and sequencing*

513 Libraries for whole SARS-CoV-2 genome sequencing were prepared according to
514 version 3 (<https://community.artic.network/t/sars-cov-2-version-4-scheme-release/312>,
515 last accessed August 19, 2021) of the ARTIC nCoV-2019 sequencing protocol. We
516 used a semi-automated workflow described previously^{6,7} that employed BioMek i7 liquid
517 handling workstations (Beckman Coulter Life Sciences) and MANTIS automated liquid
518 handlers (FORMULATRIX). Short sequence reads were generated with a NovaSeq
519 6000 instrument (Illumina).

520

521 *Sample selection*

522 The initial dataset was comprised of 39,880 samples from 70 Novaseq runs. These runs
523 also often contained samples from other institutions or collection time periods, which we
524 took into account when assessing the possibility of within-run cross-contamination but
525 did not otherwise analyze. Median sequence depth of samples was broadly but not
526 perfectly correlated with sample input quantity. We noted that very low input samples,
527 i.e. with Ct values ≥ 40 , generally had very low coverage, but there was a small subset
528 with high coverage comparable to high input samples (**SuppFig.12**). We excluded from
529 the analysis any runs containing at least three Ct ≥ 40 samples in which the coverage

530 breadth and depth were not statistically different from the Ct<40 samples in the run (t-
531 test>0.01 for median coverage or for fraction of genome with at least 1000x coverage).
532 We further limited all our analyses to samples with at least 100x coverage over at least
533 98% of the genome excluding the 5' and 3' UTRs. We also excluded samples where the
534 consensus sequence was flagged as poor quality under the default quality control
535 criteria of Nextclade¹⁵ (QC score >100) or that did not have a lineage assigned by
536 Pango¹⁶. In cases where multiple samples were collected longitudinally from the same
537 patient, we chose the earliest sample.

538

539 *Consensus sequence assembly and minor variant calling*

540 Adapter sequences were trimmed from reads using trimmomatic v0.39⁴⁰ with the
541 following options: ILLUMINACLIP:\${params.adapters}:2:30:10:8:true LEADING:20
542 TRAILING:20 SLIDINGWINDOW:4:20 MINLEN:20. Reads were aligned to the
543 Wuhan/Hu-1 SARS-CoV2 genome (RefSeq: NC_045512.2) using minimap2 v2.17 with
544 the preset genomic short-read mapping option⁴¹. ARTIC v3 primer sequences⁴² were
545 removed using iVar v.1.3.1 with a minimum quality threshold of 0 and including all reads
546 with no primer sequences found⁴³. Consensus sequences and minor variants were
547 called using an in-house variant calling pipeline, timo, available at
548 <https://github.com/GhediniLab/timo>.

549

550 *Analysis*

551 Calculations, visualizations and statistical analyses were carried out in R v4.0.3 (R
552 Foundation for Statistical Computing). Packages used for analysis included tidyverse
553 1.3.1⁴⁴, glmnet 4.1.2⁴⁵, nlme 3.1.149⁴⁶, randomForest 4.6.14⁴⁷, pROC 1.17.0.1.
554 Consensus sequence quality control was carried out in Nextclade 1.4.1¹⁵.

555

556 *Additional Data Files*

557 Inclusion_table.csv includes IDs of samples with information about which samples were
558 included in which analyses, and will include SRA accession numbers for each sample

559 when data deposition is complete. Files used in minor variant analysis are available in
560 Github repository <https://github.com/GhedinSGS/HMH-SARS-CoV2-minorvariants>.

561

562 *Sequence and code availability*

563 Raw sequence data are available under Bioproject PRJNA767338. Pipeline used for
564 minor variant calling is available at <https://github.com/GhedinLab/timo>, and data files
565 and code used for analyses are available at [https://github.com/GhedinSGS/HMH-](https://github.com/GhedinSGS/HMH-SARS-CoV2-minorvariants)
566 [SARS-CoV2-minorvariants](https://github.com/GhedinSGS/HMH-SARS-CoV2-minorvariants).

567

568 *Acknowledgements*

569 We thank Drs. Marc Boom and Dirk Sostman for their ongoing support. We thank
570 Matthew Ojeda Saavedra, Kristina Reppond, Madison N Shyer, Jessica Cambric,
571 Ryan Gadd, Rashi M Thakur, Akanksha Batajoo, Regan Mangham, Sindy Pena and
572 Trina Trinh for technical support. The research was supported by the Houston Methodist
573 Academic Institute Infectious Diseases Fund and many generous Houston
574 philanthropists. This work utilized the computational resources of the NIH High
575 Performance Computing (HPC) Biowulf cluster (<http://hpc.nih.gov>).

576

577 *Funding Statement*

578 This project was supported in part by the Houston Methodist Academic Institute
579 Infectious Diseases Fund, and federal funds from the National Institute of Allergy and
580 Infectious Diseases, National Institutes of Health, Department of Health and Human
581 Services, under Contract No. 75N93019C00076 (J.J.D.). This work was also supported
582 in part by the Division of Intramural Research (DIR) of the NIAID/NIH.

583

584

585 **TABLE 1. Linear mixed-effects models for association of hospitalization,**
586 **vaccination and healthcare worker surveillance samples on minor variant**
587 **richness.**

588 a. Effect of sample Ct value and patient hospitalization status on log10 transformed
589 minor variant richness. Sequencing run is included as a random effect.

	numDF	denDF	F-value	p-value
(Intercept)	1	6090	602.4491	<.0001
Ct	1	6090	1370.695	<.0001
admitted_hospital	1	6090	77.6004	<.0001
Ct:admitted_hospital	1	6090	11.5969	7.00E-04

590

591 b. Effect of sample Ct value and patient vaccination status on log10 transformed
592 minor variant richness. Only samples from non-hospitalized patients collected in
593 July 2021 or later are included. Sequencing run is included as a random effect.

	numDF	denDF	F-value	p-value
(Intercept)	1	1346	297.22177	<.0001
Ct	1	1346	152.42258	<.0001
vaccine_status	1	1346	4.52737	0.0335
Ct:vaccine_status	1	1346	4.31531	0.0380

594

595 c. Effect of sample Ct value and healthcare worker surveillance sample status on
596 log10 transformed minor variant richness. Only samples from non-hospitalized
597 individuals are included. Sequencing run is included as a random effect.

	numDF	denDF	F-value	p-value
(Intercept)	1	3824	455.9225	<.0001
Ct	1	3824	669.4242	<.0001
surveillance_sample	1	3824	0.1950	0.6588
Ct:surveillance_sample	1	3824	0.0207	0.8857

598

599

600 **FIGURE LEGENDS**

601 **Figure 1. Distribution of minor variant richness in Houston Methodist SARS-CoV-**

602 **2 specimens. a.** Change in distribution of sample Ct values after applying minimum

603 sequencing depth and background contamination exclusion criteria (see Methods).

604 **b.** Number of minor variants per sample in 15,389 samples passing initial quality control

605 criteria (minimum depth/breadth of coverage, run-level quality screening, consensus

606 sequence assembly quality). **c.** Relationship between minor variant richness and

607 sample input quantity (qPCR Ct value) in samples for which Ct values were available

608 (n=7356). Vertical dotted line represents Ct<26, the threshold below which minor

609 variants are generally reproducible.

610

611 **Figure 2. Association between patient status and SARS-CoV-2 diversity. a.** Minor

612 variant richness across Ct values in patients requiring hospitalization (inpatient or ICU)

613 vs. outpatients (n=6140). **b.** Minor variant richness in fully vaccinated vs. unvaccinated

614 or partially vaccinated patients, limiting samples to those collected in July 2021 or later

615 and only from patients not requiring hospitalization (n=1366). **c.** Minor variant richness

616 in non-hospitalized patients vs. healthcare workers undergoing voluntary surveillance

617 (n=3874).

618

619 **Figure 3. Genomic context of SARS-CoV-2 minor variants. a.** Prevalence of minor

620 variants (top) and consensus changes (bottom) at nucleotide positions across the

621 genome. **b.** Density of minor variants in different SARS-CoV-2 genes (defined as total

622 number of minor variants found in a gene across all samples divided by the length of the

623 gene). **c.** Prevalence of different consensus to minor nucleotide substitutions.

624

625 **Figure 4. Nucleotide substitutions and minor allele frequencies at highly mutable**

626 **sites. a.** Minor variant changes at 34 nucleotide positions containing minor variants in at

627 least 2% of samples. **b.** Minor allele frequencies of variants at these sites.

628

629

630 **Supplemental Figures**

631

632 **Supplementary Figure 1. Reproducibility of minor variants in 54 samples with**
633 **technical replicates. a,b.** Reproducibility of minor variant detection (black vs. red) and
634 minor allele frequency in samples with different input quantities, by category and by
635 individual samples. “Unknown” Ct values represent samples diagnosed on the Hologic
636 Aptima instrument, which gives results in relative light units (values >100 in individual
637 sample panels) or on the Biofire Diagnostics instrument, which does not quantify viral
638 load (“NA” in individual sample panels). **c.** Minor variant reproducibility in samples with
639 different ranges of median sequencing depth. **d.** Sequencing depth at site and minor
640 allele frequency of minor variants that were subsequently detected in the second
641 technical replicate (“yes”) vs. not detected (“no”).

642

643 **Supplementary Figure 2. Collection dates and consensus virus lineage of final**
644 **sample set.**

645

646 **Supplementary Figure 3. Run-level effects on minor variant richness in filtered**
647 **sample set. a.** Minor variant richness in high coverage and Ct<26 samples in each
648 sequencing run. **b.** Relationship between sample’s median sequencing coverage and
649 number of minor variants. **c.** Median number of minor variants for samples in each run
650 represented as a function of run–level median of median coverage.

651

652 **Supplementary Figure 4. Categories of minor variant diversity among hospitalized**
653 **and non-hospitalized patients.**

654

655 **Supplementary Figure 5. LASSO regression coefficients for association of patient**
656 **and sample factors of interest with minor variant richness. a.** Results of model in
657 which all factors of interest are included. **b.** Results of model in which factors with a
658 strong temporal dimension were excluded (collection month, consensus lineage,
659 vaccination status).

660

661 **Supplementary Figure 6. Associations between patient status and SARS-CoV-2**
662 **diversity in three datasets using different thresholds for sample coverage and**
663 **minor variant detection.**

664

665 **Supplementary Figure 7. Density of minor variants in different SARS-CoV-2 genes**
666 **in samples from hospitalized and non-hospitalized patients.** Minor variant density is
667 defined as total number of minor variants found in a gene across all samples divided by
668 the length of the gene.

669

670 **Supplementary Figure 8.** Prevalence of different consensus > minor nucleotide
671 substitutions at minor variant sites that were detected vs. not detected in a second
672 sequencing replicate.

673

674 **Supplementary Figure 9. Prevalence of minor alleles and consensus alleles at the**
675 **34 most frequent minor variant sites, by run.** For consensus allele prevalence, all
676 samples from the run are included, regardless of whether they were analyzed in the
677 minor variant study, on the assumption that they may contribute to cross-contamination
678 in other samples.

679

680 **Supplementary Figure 10. Minor variant prevalence in the Houston Methodist**
681 **dataset vs. prevalence of consensus mutations at these sites in US-wide SARS-**
682 **CoV-2 sequences from GISAID.** GISAID sequences were queried on June 13, 2022
683 covering the entire length of the pandemic in the U.S. to identify the number of
684 sequences that had any nucleotide substitution (A,C,T,G) at the 34 nucleotide positions
685 that most frequently had minor variants in the Houston dataset.

686

687 **Supplementary Figure 11. Prevalence across time of minor variants**
688 **corresponding to recurrent amino acid changes associated with increased SARS-**
689 **CoV-2 fitness, as identified in Obermeyer et al²⁴.**

690

691 **Supplementary Figure 12. Sequencing depth and Ct values of Houston Methodist**
692 **SARS-CoV-2 samples prior to filtering.**

693

694

695

696

697 **References**

698

699 1. du Plessis, L. *et al.* Establishment and lineage dynamics of the SARS-CoV-2

700 epidemic in the UK. *Science* eabf2946 (2021) doi:10.1126/science.abf2946.

701 2. Grubaugh, N. D. *et al.* An amplicon-based sequencing framework for accurately

702 measuring intrahost virus diversity using PrimalSeq and iVar. *Genome Biol.* **20**, 8

703 (2019).

704 3. Worby, C. J., Lipsitch, M. & Hanage, W. P. Shared Genomic Variants: Identification

705 of Transmission Routes Using Pathogen Deep-Sequence Data. *Am. J. Epidemiol.*

706 **186**, 1209–1216 (2017).

707 4. The CITIID-NIHR BioResource COVID-19 Collaboration *et al.* SARS-CoV-2

708 evolution during treatment of chronic infection. *Nature* **592**, 277–282 (2021).

709 5. Roder, Ae. *et al.* Diversity and selection of SARS-CoV-2 minority variants in the

710 early New York City outbreak.

711 <http://biorxiv.org/lookup/doi/10.1101/2021.05.05.442873> (2021)

712 doi:10.1101/2021.05.05.442873.

713 6. Long, S. W. *et al.* Sequence Analysis of 20,453 Severe Acute Respiratory Syndrome

714 Coronavirus 2 Genomes from the Houston Metropolitan Area Identifies the

715 Emergence and Widespread Distribution of Multiple Isolates of All Major Variants of

716 Concern. *Am. J. Pathol.* **191**, 983–992 (2021).

717 7. Olsen, R. J. *et al.* Trajectory of Growth of Severe Acute Respiratory Syndrome

718 Coronavirus 2 (SARS-CoV-2) Variants in Houston, Texas, January through May

- 719 2021, Based on 12,476 Genome Sequences. *Am. J. Pathol.* **191**, 1754–1773
720 (2021).
- 721 8. Christensen, P. A. *et al.* Signals of Significantly Increased Vaccine Breakthrough,
722 Decreased Hospitalization Rates, and Less Severe Disease in Patients with
723 Coronavirus Disease 2019 Caused by the Omicron Variant of Severe Acute
724 Respiratory Syndrome Coronavirus 2 in Houston, Texas. *Am. J. Pathol.*
725 S000294402200044X (2022) doi:10.1016/j.ajpath.2022.01.007.
- 726 9. Valesano, A. L. *et al.* Temporal dynamics of SARS-CoV-2 mutation accumulation
727 within and across infected hosts. *PLOS Pathog.* **17**, e1009499 (2021).
- 728 10. Lythgoe, K. A. *et al.* SARS-CoV-2 within-host diversity and transmission. *Science*
729 eabg0821 (2021) doi:10.1126/science.abg0821.
- 730 11. Tonkin-Hill, G. *et al.* Patterns of within-host genetic diversity in SARS-CoV-2. *eLife*
731 **10**, e66857 (2021).
- 732 12. Braun, K. M. *et al.* Acute SARS-CoV-2 infections harbor limited within-host diversity
733 and transmit via tight transmission bottlenecks. *PLOS Pathog.* **17**, e1009849 (2021).
- 734 13. Storz, J. F. *et al.* The role of mutation bias in adaptive molecular evolution: insights
735 from convergent changes in protein function. 1.
- 736 14. Stoltzfus, A. & Yampolsky, L. Y. Climbing Mount Probable: Mutation as a Cause of
737 Nonrandomness in Evolution. *J. Hered.* **100**, 637–647 (2009).
- 738 15. Hadfield, J. *et al.* Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics*
739 **34**, 4121–4123 (2018).

- 740 16. O'Toole, Á. *et al.* Assignment of Epidemiological Lineages in an Emerging Pandemic
741 Using the Pangolin Tool. *Virus Evol.* veab064 (2021) doi:10.1093/ve/veab064.
- 742 17. Ortiz, A. T. *et al.* *Within-host diversity improves phylogenetic and transmission*
743 *reconstruction of SARS-CoV-2 outbreaks.*
744 <http://biorxiv.org/lookup/doi/10.1101/2022.06.07.495142> (2022)
745 doi:10.1101/2022.06.07.495142.
- 746 18. Zinzula, L. Lost in deletion: The enigmatic ORF8 protein of SARS-CoV-2. *Biochem.*
747 *Biophys. Res. Commun.* **538**, 116–124 (2021).
- 748 19. Lin, M. J. *et al.* Host–pathogen dynamics in longitudinal clinical specimens from
749 patients with COVID-19. *Sci. Rep.* **12**, 5856 (2022).
- 750 20. Stoler, N. & Nekrutenko, A. Sequencing error profiles of Illumina sequencing
751 instruments. *NAR Genomics Bioinforma.* **3**, lqab019 (2021).
- 752 21. Rut, W. *et al.* Activity profiling and crystal structures of inhibitor-bound SARS-CoV-2
753 papain-like protease: A framework for anti-COVID-19 drug design. *Sci. Adv.* **6**,
754 eabd4596 (2020).
- 755 22. Shin, D. *et al.* Papain-like protease regulates SARS-CoV-2 viral spread and innate
756 immunity. *Nature* **587**, 657–662 (2020).
- 757 23. Chen, C. *et al.* CoV-Spectrum: analysis of globally shared SARS-CoV-2 data to
758 identify and characterize new variants. *Bioinformatics* **38**, 1735–1737 (2022).
- 759 24. Obermeyer, F. *et al.* Analysis of 6.4 million SARS-CoV-2 genomes identifies
760 mutations associated with fitness. *Science* abm1208 (2022)
761 doi:10.1126/science.abm1208.

- 762 25. de Maio, N. *et al.* Issues with SARS-CoV-2 sequencing data.
- 763 26. Martin, D. P. *et al.* The emergence and ongoing convergent evolution of the SARS-
764 CoV-2 N501Y lineages. *Cell* S0092867421010503 (2021)
765 doi:10.1016/j.cell.2021.09.003.
- 766 27. Siqueira, J. D. *et al.* SARS-CoV-2 genomic and quasispecies analyses in cancer
767 patients reveal relaxed intrahost virus evolution.
768 <http://biorxiv.org/lookup/doi/10.1101/2020.08.26.267831> (2020)
769 doi:10.1101/2020.08.26.267831.
- 770 28. Al Khatib, H. A. *et al.* Within-Host Diversity of SARS-CoV-2 in COVID-19 Patients
771 With Variable Disease Severities. *Front. Cell. Infect. Microbiol.* **10**, 575613 (2020).
- 772 29. Kuipers, J. *et al.* Within-patient genetic diversity of SARS-CoV-2.
773 <http://biorxiv.org/lookup/doi/10.1101/2020.10.12.335919> (2020)
774 doi:10.1101/2020.10.12.335919.
- 775 30. Simons, L. M. *et al.* Assessment of Virological Contributions to COVID-19 Outcomes
776 in a Longitudinal Cohort of Hospitalized Adults. *Open Forum Infect. Dis.* **9**, ofac027
777 (2022).
- 778 31. Li, J. *et al.* Two-step fitness selection for intra-host variations in SARS-CoV-2. *Cell*
779 *Rep.* 110205 (2021) doi:10.1016/j.celrep.2021.110205.
- 780 32. Töpfer, A. *et al.* Sequencing approach to analyze the role of quasispecies for
781 classical swine fever. *Virology* **438**, 14–19 (2013).
- 782 33. Honce, R. *et al.* Obesity-Related Microenvironment Promotes Emergence of Virulent
783 Influenza Virus Strains. *mBio* **11**, (2020).

- 784 34. Mathew, D. *et al.* Deep immune profiling of COVID-19 patients reveals distinct
785 immunotypes with therapeutic implications. *Science* **369**, eabc8511 (2020).
- 786 35. Nicholson, M. D. *et al.* Response to comment on “Genomic epidemiology of
787 superspreading events in Austria reveals mutational dynamics and transmission
788 properties of SARS-CoV-2”. *Sci. Transl. Med.* **13**, eabj3222 (2021).
- 789 36. Pathak, A. K. *et al.* Spatio-temporal dynamics of intra-host variability in SARS-CoV-2
790 genomes. 11.
- 791 37. Armero, A., Berthet, N. & Avarre, J.-C. Intra-Host Diversity of SARS-Cov-2 Should
792 Not Be Neglected: Case of the State of Victoria, Australia. *Viruses* **13**, 133 (2021).
- 793 38. Walker, A. W. A Lot on Your Plate? Well-to-Well Contamination as an Additional
794 Confounder in Microbiome Sequence Analyses. *mSystems* **4**, (2019).
- 795 39. Lam, C. *et al.* Sars-CoV-2 Genome Sequencing Methods Differ In Their Ability To
796 Detect Variants From Low Viral Load Samples. *J. Clin. Microbiol.* (2021)
797 doi:10.1128/JCM.01046-21.
- 798 40. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina
799 sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
- 800 41. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**,
801 3094–3100 (2018).
- 802 42. Tyson, J. R. *et al.* Improvements to the ARTIC multiplex PCR method for SARS-
803 CoV-2 genome sequencing using nanopore.
804 <http://biorxiv.org/lookup/doi/10.1101/2020.09.04.283077> (2020)
805 doi:10.1101/2020.09.04.283077.

- 806 43. Castellano, S. *et al.* iVar, an Interpretation-Oriented Tool to Manage the Update and
807 Revision of Variant Annotation and Classification. *Genes* **12**, 384 (2021).
- 808 44. Wickham, H. *et al.* Welcome to the Tidyverse. *J. Open Source Softw.* **4**, 1686
809 (2019).
- 810 45. Friedman, J., Hastie, T. & Tibshirani, R. Regularization Paths for Generalized Linear
811 Models via Coordinate Descent. *J. Stat. Softw.* **33**, (2010).
- 812 46. nlme: Linear and Nonlinear Mixed Effects Models.
- 813 47. Liaw, A. & Wiener, M. Classification and Regression by randomForest. **2**, 5 (2002).
- 814