

1 Genome-wide association study between SARS-CoV-2 single 2 nucleotide polymorphisms and virus copies during infections

3
4 Ke Li^{1,2,#}, Chrispin Chaguza^{1,2}, Julian Stamp³, Yi Ting Chew^{1,2}, Nicholas F.G. Chen¹, David
5 Ferguson^{4,5}, Sameer Pandya^{4,5}, Nick Kerantzas^{4,5}, Wade Schulz^{4,5}, Yale SARS-CoV-2 Genomic
6 Surveillance Initiative*, Anne M. Hahn¹, C Brandon Ogbunugafor^{2,6,7}, Virginia E. Pitzer^{1,2}, Lorin
7 Crawford^{3,8,9}, Daniel M. Weinberger^{1,2}, Nathan D. Grubaugh^{1,2,6,#}

8
9 ¹ Department of Epidemiology of Microbial Diseases, Yale School of Public Health, New Haven, CT, USA

10 ² Public Health Modeling Unit, Yale School of Public Health, New Haven, CT, USA

11 ³ Center for Computational Molecular Biology, Brown University, Providence, RI, USA

12 ⁴ Department of Laboratory Medicine, Yale School of Medicine, New Haven, CT, USA

13 ⁵ Yale School of Medicine Biorepository, Yale University, New Haven, CT, USA

14 ⁶ Department of Ecology and Evolutionary Biology, Yale University, New Haven, CT, USA

15 ⁷ Santa Fe Institute, Santa Fe, NM, USA

16 ⁸ Department of Biostatistics, Brown University, Providence, RI, USA

17 ⁹ Microsoft Research, Cambridge, MA, USA

18 * Authors listed at the end of the manuscript

19 # corresponding authors: ke.li.kl662@yale.edu; nathan.grubaugh@yale.edu

20 Abstract

21 Significant variations have been observed in viral copies generated during SARS-CoV-2
22 infections. However, the factors that impact viral copies and infection dynamics are not fully
23 understood, and may be inherently dependent upon different viral and host factors. Here, we
24 conducted virus whole genome sequencing and measured viral copies using RT-qPCR from
25 9,902 SARS-CoV-2 infections over a 2-year period to examine the impact of virus genetic
26 variation on changes in viral copies adjusted for host age and vaccination status. Using a
27 genome-wide association study (GWAS) approach, we identified multiple single-nucleotide
28 polymorphisms (SNPs) corresponding to amino acid changes in the SARS-CoV-2 genome
29 associated with variations in viral copies. We further applied a marginal epistasis test to
30 detect interactions among SNPs and identified multiple pairs of substitutions located in the
31 spike gene that have non-linear effects on viral copies. We also analyzed the temporal
32 patterns and found that SNPs associated with increased viral copies were predominantly
33 observed in Delta and Omicron BA.2/BA.4/BA.5/XBB infections, whereas those associated
34 with decreased viral copies were only observed in infections with Omicron BA.1 variants. Our
35 work showcases how GWAS can be a useful tool for probing phenotypes related to SNPs in
36 viral genomes that are worth further exploration. We argue that this approach can be used
37 more broadly across pathogens to characterize emerging variants and monitor therapeutic
38 interventions.

40 Author Summary

41 Our study explores why viral load (copies measured by RT-qPCR) varies during SARS-CoV-
42 2 infections by analyzing viral mutations and measuring viral copies in 9,902 individuals over
43 two years. We aimed to understand how genetic differences in SARS-CoV-2 influence viral
44 copies, considering host age and vaccination status. Using a genome-wide association study
45 (GWAS), we identified several single-nucleotide polymorphisms (SNPs) in the virus linked to

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

46 variations in viral levels. Notably, interactions between certain SNPs in the spike gene had
47 non-linear effects on viral copies. Our analysis revealed that SNPs associated with higher
48 viral copies were common in Delta and Omicron BA.2/BA.4/BA.5/XBB variants, while those
49 linked to lower levels were mainly found in Omicron BA.1. This research highlights GWAS as
50 a powerful tool for exploring virus genetics and suggests it can be broadly applied to monitor
51 new variants of COVID-19 and other infectious diseases.
52

53 Introduction

54 Continued SARS-CoV-2 transmission and evolution has propelled the COVID-19 pandemic.
55 Peak viral replication in the upper respiratory tract occurs during the first few days of infection
56 [1]. The viral load (or copies measured by RT-qPCR) in patient samples are valuable data to
57 understand infection dynamics, such as inferring the likelihood of disease transmission [2].
58 However, it is challenging to use viral load data, and the challenge often arises from
59 significant variations in viral load dynamics among sampled cases, which can be associated
60 with 1) host heterogeneity, e.g., age [3] and vaccination status [6–8]; 2) distinct inherent
61 properties of virus variants or sublineages [9], and 3) different sampling times [10]. For
62 example, sampling during the early stages of infection may yield higher viral loads compared
63 to later stages after viral replication has reached its peak. Nevertheless, the relative
64 importance of these factors influencing viral load has not been completely explored [11,12].
65

66 Genome-wide association studies (GWAS) have emerged as a useful tool in the field of
67 genetics, providing an approach to unraveling the complex interplay between genetic
68 variations and observable traits, including diseases and drug resistance, as reviewed in [13].
69 Several studies have employed GWAS analysis to identify and investigate the association
70 between human genetic variations across different individuals and the severity of COVID-19,
71 shedding light on genetic variations that are related to severe infections [15–17]. However,
72 few studies have utilized a GWAS method to study associations between the viral genome
73 and viral traits [19–22]. The confluence of the extensive existing research on SARS-CoV-2
74 mutations and the millions of infections that have been sequenced provides us the
75 opportunity to evaluate the application of GWAS for viral genomics. The hypothesis-free
76 approach has the potential to enhance our understanding of genetic determinants influencing
77 viral fitness and evolution and further inform effective public health strategies aimed at
78 mitigating the spread and impact of SARS-CoV-2.
79

80 In this work, we aim to investigate the impact of intrinsic viral genetic substitutions (i.e., single
81 nucleotide polymorphisms [SNPs]) on the changes in viral copies, adjusted for host age and
82 vaccination status. For this, we apply a viral GWAS analysis to SARS-CoV-2 genomic
83 sequencing and standardized RT-qPCR data collected from the Yale New Haven Hospital
84 from February 2021 to March 2023. Using whole genome sequencing data on SARS-CoV-2
85 infections, along with relevant laboratory and patient metadata, we identify associations
86 between viral SNPs and viral copies for different variants of concern (VOCs). We then examine
87 the temporal pattern of identified SNPs by constructing a phylogenetic tree, drawing upon
88 subsamples, and analyzing the time series of the fraction of SNPs occurring in the sequences.
89 This multifaceted analysis contributes to unraveling the complex dynamics of SARS-CoV-2
90 infections, providing valuable insights into the underlying viral SNPs that influence viral copies
91 in different VOCs.

92 Results

93 Viral copies vary in SARS-CoV-2 infections

94 To better understand how SARS-CoV-2 viral load varies in infected individuals, we analyzed
95 the viral copy data, along with associated host metadata (i.e., age and vaccination status),
96 and genome sequencing data from a cohort of patients tested at the Yale New Haven
97 Hospital (YNHH) located in Connecticut, US. We selected 9902 whole genome sequences
98 with available viral copy data generated from remnant SARS-CoV-2 diagnostic samples over
99 a 2-year period, from 03-Feb-2021 to 21-Mar-2023 (**Fig. 1A**). The VOCs that we identified in
100 our dataset during the sampling period included Alpha (n = 809), Delta (n = 1278), Gamma (n
101 = 36), BA.1 (n = 1818), BA.2 (n = 2432), BA.4 (n = 293), BA.5 (n = 1992), XBB (n = 698), and
102 the pre-VOC variant (named 'Other', n = 546). We conducted RT-qPCR using a standardized
103 assay targeting the nucleocapsid (CDC 'N1' primers) for each sample to allow for cross-
104 sample comparisons [23], except for a period during October 2021 when the PCR data were
105 not generated. Across all samples, the viral copies, expressed as \log_{10} (viral copies per
106 milliliter (Genome Equivalents/ml)), exhibited variations, ranging from 3.60 to 10.55, with a
107 median value of 7.26 (**Fig. 1B**). The variations in viral copies could be attributed either to the
108 introduction and/or replacement of different VOCs, each with its own epidemic curve, or to
109 the stochasticity from the sampling process. To reduce stochastic effects, we aggregated
110 the viral copies by month and still observed large variations in the viral copies across the
111 months, albeit with no consistent trend (**Fig. 1C**). Notably, we observed the lowest median
112 value of viral copies (median = 6.49) in February 2022, during which 96.3% of the sampled
113 sequences tested positive for BA.1 infections. By contrast, we observed the highest median
114 value of viral copies (median = 7.70) in June 2022, during which the sampled sequences
115 tested positive for BA.2 (64.9%), BA.4 (6%), or BA.5 (29.1%) infections. Taken together, we
116 showed a wide range of viral copies in the sampled SARS-CoV-2 infections with different
117 VOCs, utilizing data from genomic surveillance and standardized RT-qPCR tests.
118



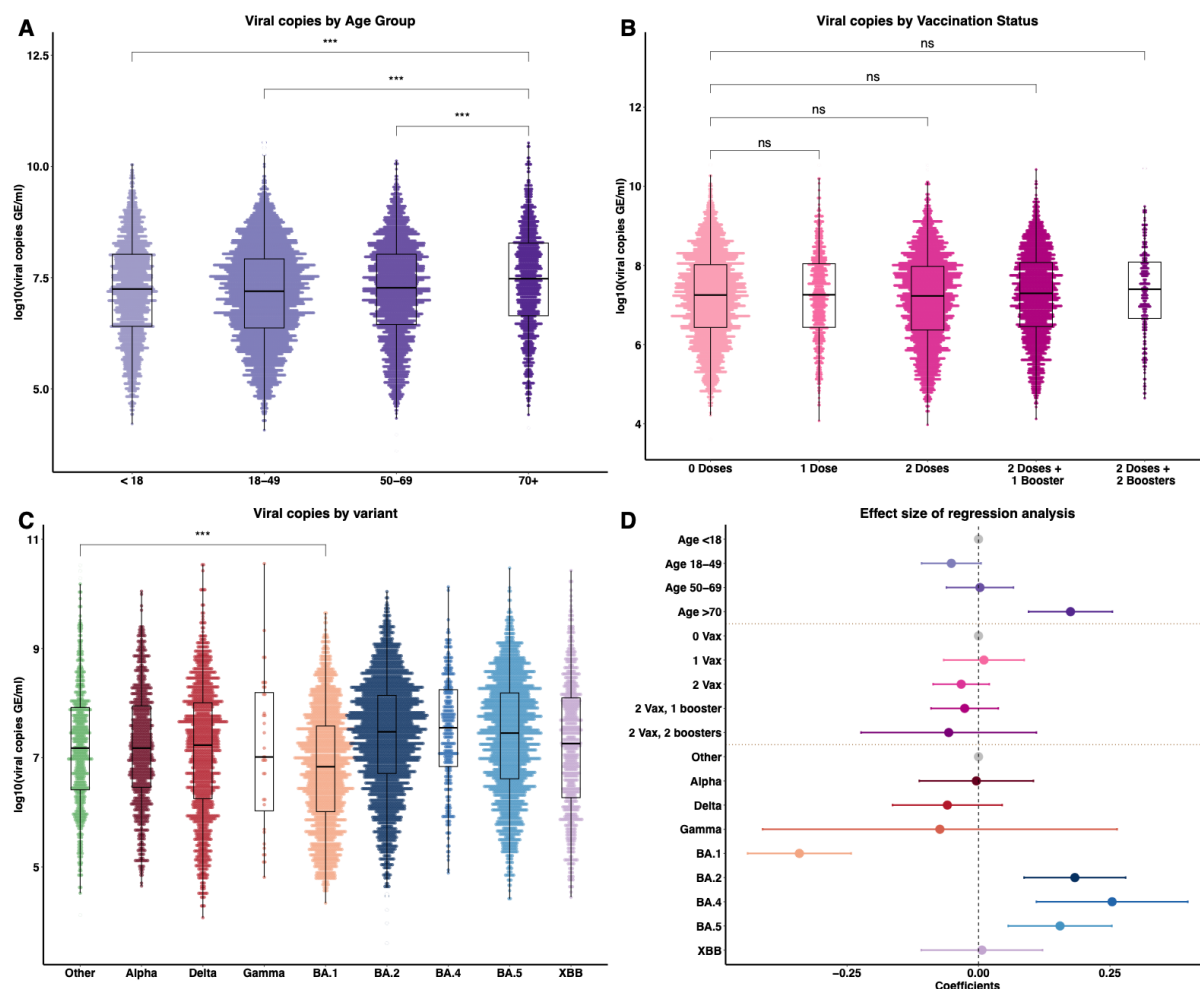
119
120 **Figure 1. Genomic sequences of SARS-CoV-2 infections and associated viral copies from cross-**
121 **sectional samples collected in Connecticut, US. (A)** The daily number of genomic sequences of
122 SARS-CoV-2 VOCs from February 2021 to March 2023. **(B)** The summary of viral copies of all samples,
123 expressed as \log_{10} (viral copies per milliliter). **(C)** The summary of viral copies aggregated by month.
124 The data gap in October 2021 is because we were unable to conduct PCR to obtain viral copies during
125 this time.

126 Viral copies correlate with age and variants, but not with vaccination 127 status

128 Having uncovered a large variability in the observed viral copies from the samples, we next
129 assessed the factors associated with these changes. To do this, we first summarized and
130 compared viral copies in various age groups (**Fig. 2A**). A positive correlation has been
131 previously reported between age and SARS-CoV-2 viral copies, showing that younger age
132 groups had lower viral copies independent of gender and/or symptom duration [24]. We
133 observed a similar result in our dataset and found that the oldest age group (i.e., >70 years
134 old) had the highest viral copies compared with other age groups (mean = 7.47, 95%
135 confidence interval (CI): [5.12, 9.49], $p < 0.001$, Wilcoxon signed-rank test). For the effect of
136 vaccination on viral copies, some studies have demonstrated that although vaccination
137 reduced the risk of infections with the Delta variant, no significant difference in peak viral
138 copies was found between fully vaccinated and unvaccinated individuals [6,7,25]. In contrast,
139 other studies have shown that vaccination reduced viral copies in BA.1 infections among
140 boosted individuals compared to unvaccinated ones [8]. These results suggest the effect of
141 vaccination on viral copies may depend on the characteristics of the infecting SARS-CoV-2
142 variant. We compared viral copies among groups with different vaccination statuses to
143 assess the impact of vaccination on viral copies (**Fig. 2B**), and no statistically significant

144 differences were detected between the groups in our data ($p > 0.05$, Wilcoxon signed-rank
145 test). Finally, we compared viral copies stratified by variant category (**Fig. 2C**). Combining
146 samples collected from all age and vaccination status groups for each variant, we found that
147 the overall mean values of viral copies were lowest for infections with BA.1 (mean = 6.83,
148 95% CI: [4.87, 8.87], $p < 0.001$, Wilcoxon signed-rank test) compared to infections with other
149 all non-BA.1 variants.

150
151 Since several factors may simultaneously impact the SARS-CoV-2 viral load, next, we sought
152 to quantify the combined impact of age, vaccination status, and VOCs on the observed viral
153 copies. To achieve this, we fitted a multivariate linear regression model, with viral copies as
154 the outcome variable and age, vaccination, and VOCs as covariates (**Fig. 2D**). We found that
155 the older age group (i.e., age >70 years old) had a positive association with viral copies (mean
156 = 0.17, 95% CI: [0.09, 0.25], $p < 0.001$) compared with the reference group (i.e., age <18
157 years old). We also found that vaccination status was not associated with viral copies (i.e.,
158 95% CIs of the vaccination coefficients span 0, $p > 0.05$). Notably, we showed that infections
159 with BA.1 were associated with reduced viral copies, with a mean effect size of -0.34 (95%
160 CI: [-0.44, -0.24], $p < 0.001$) in the same age group and vaccination status, compared to the
161 Other variant. We also showed that infections with BA.2 (mean = 0.19, 95% CI: [0.09, 0.28],
162 $p < 0.001$), BA.4, or BA.5 (mean = 0.17, 95% CI: [0.07, 0.26], $p < 0.001$) were associated with
163 increased viral copies. Among them, infections with BA.4 were associated with the largest
164 positive effect size (mean = 0.27, 95% CI: [0.12, 0.41], $p < 0.001$). Our findings demonstrated
165 that variations in viral copies were associated with infections caused by different SARS-CoV-
166 2 variants and the older age group. This implies that intrinsic factors of the viruses, such as
167 genetic mutations among distinct VOCs, are key determinants impacting viral copies.



168
 169 **Figure 2. Viral copies by category and regression analysis results.** Comparison of viral copies
 170 stratified by (A) age groups, (B) vaccination statuses, (C) variant of concerns. (D) Association of age,
 171 vaccination status, and VOCs with viral copies, expressed as \log_{10} (viral copies per milliliter (Genome
 172 Equivalents/ml)). The reference groups (in gray) are Age <18 years old, 0 doses of vaccination, and the
 173 Other variant, respectively. The positive coefficients indicate the covariate is associated with higher
 174 viral copies value compared to the reference group, and vice versa. 0 Vax, 1 Vax, 2 Vax, 2 Vax 1
 175 booster, and 2 vax 2 boosters denote vaccination statuses of 0 doses, 1 dose, 2 doses, 2 doses,
 176 and 1 booster, and 2 doses and 2 boosters, respectively, corresponding to the labels in (B). Results are
 177 shown as means with 95% confidence intervals. *** $p < 0.001$.

178 Viral GWAS reveals SARS-CoV-2 SNPs associated with viral copies

179 Having demonstrated that changes in SARS-CoV-2 viral copies are associated with infections
 180 caused by different viral variants or strains, especially Omicron BA.1/BA.2/BA.4/BA.5 variants
 181 (Fig. 2D), we then sought to identify potential genetic mutations—specifically, SNPs—that
 182 contributed to these changes in viral copies. For this, we performed a GWAS analysis using
 183 high-quality genome sequences (i.e., genome coverage > 95%). We conducted whole-
 184 genome sequencing on the 9902 SARS-CoV-2 positive specimens collected from February
 185 2021 to March 2023. Firstly, using Wuhan-Hu-1 (GenBank MN908937.3) as the reference
 186 genome, we identified 10,697 SNPs for further testing associated with viral copies as
 187 covariates. We then checked for the population structure of the 9902 genome sequences
 188 using a multidimensional scaling (MDS) method [26] (Fig. S1). We observed that Delta was
 189 an outgroup to other pre-Omicron variants (i.e., pre-VOC variant (Other), Alpha, and Gamma),
 190 and BA.1 was an outgroup to the BA.2/BA.4/BA.5/XBB cluster. In our model, we included the

191 inferred four clusters based on the MDS-computed distance to capture the viral population
192 structure. Clusters were defined using a k-means clustering method (**Fig. S1**). The host ages
193 and vaccination status were also included in the model as covariates.

194

195 Using the linear regression model on viral copies for each SNP, adjusted for viral population
196 structure and host factors, we identified 31 SNPs exceeding the permuted threshold for
197 genome-wide significance ($p = 4.67 \times 10^{-6}$, dashed line, **Fig. 3A**). The threshold value was
198 calculated as 0.05 divided by 10,697 SNPs [27]. We found that the observed distribution of
199 p -values closely matches the expected distribution under the null hypothesis of no
200 association (**Fig. S2A**). To ascertain whether those SNPs have a negative or positive impact
201 on viral copies and evaluate their effect size, we extracted the coefficients (β) of the SNPs
202 with $p < 1 \times 10^{-10}$ and their standard deviations (σ) from the regression model (dashed box,
203 **Fig. 3B**). We then annotated the SNPs to identify the associated amino acids, and among
204 them, 14 SNPs were non-synonymous (i.e., changed the amino acid; **Fig. 3C**). We found that
205 a non-synonymous change N:R203M, located on the N gene, had the most significant
206 association with increased viral copies ($p = 2.68 \times 10^{-22}$, $\beta = 1.65$, $\sigma = 0.16$). By contrast, the
207 amino acid change most strongly associated with a negative effect on viral copies was
208 ORF1ab:L5086I ($p = 9.20 \times 10^{-20}$, $\beta = -1.20$, $\sigma = 0.13$). We further conducted a marginal
209 epistasis test [28–30] to detect the epistatic effects of SNPs on viral copies. We discovered
210 multiple pairs of SNPs that exhibit positive epistatic effects on viral copies, with most
211 interactions occurring in the S gene (**Fig. S3**).

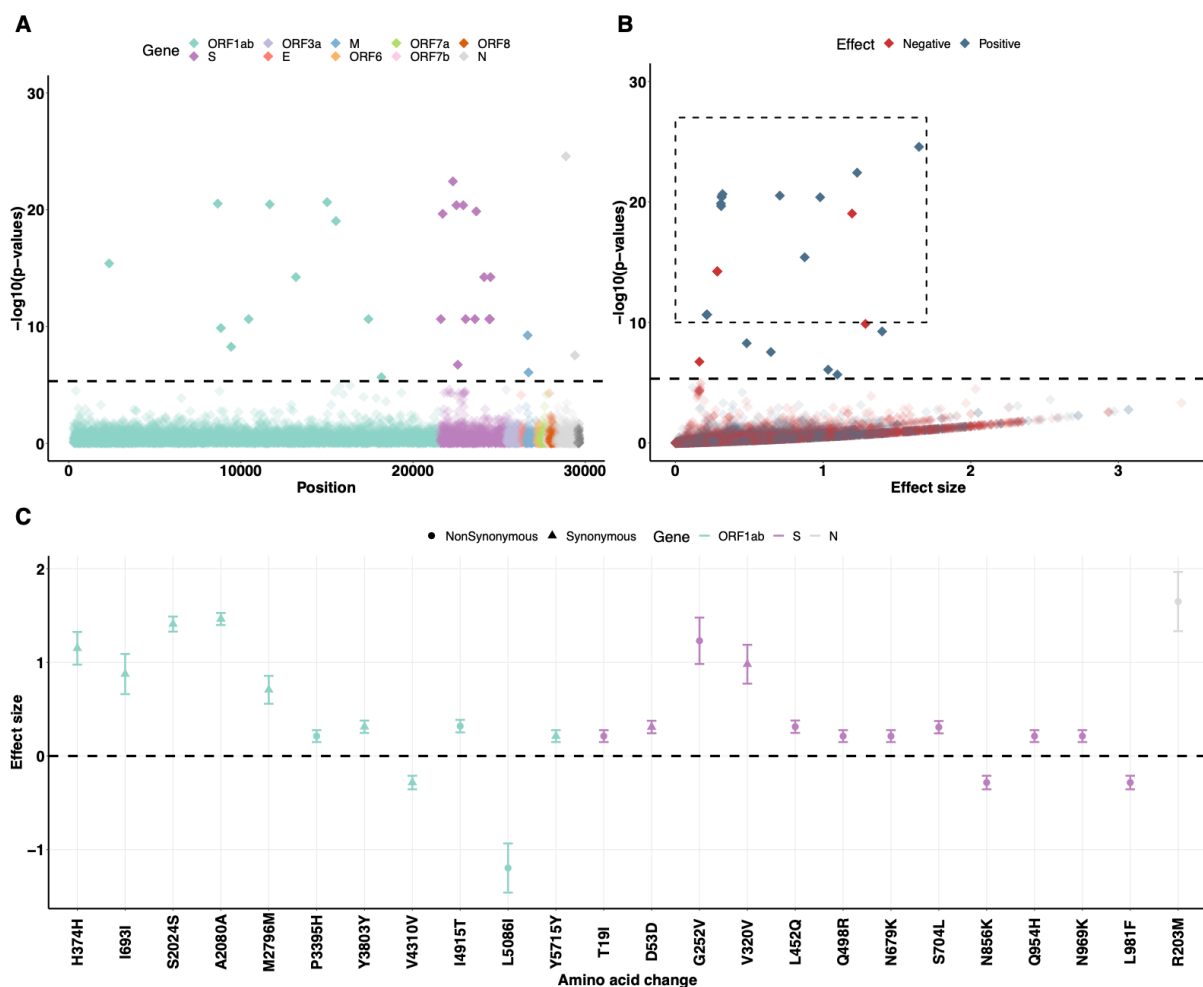
212

213 To assess the impact of adjusting for the population structure of the SARS-CoV-2 strains
214 using the MDS components on the regression results, we conducted a sensitivity analysis on
215 the genome sequences using the inferred MDS components from the pairwise SNP distance
216 matrix of SARS-CoV-2 sequences as covariates. By doing this, we identified 113 SNPs
217 exceeding the permuted threshold (**Fig. S4**). The observed distribution of p -values also
218 closely matched the expected distribution under the null hypothesis of no association (**Fig.**
219 **S2B**). The results may be more likely to reflect the SNPs that influence the viral copies
220 dependent on lineage. We also examined the association between viral copies and SNPs
221 after adjusting for the population structure based on the VOCs themselves, which broadly
222 correspond to the identified sequence clusters. We showed that only a few SNPs were found
223 (Figs. S5-8), mostly within the Omicron BA.2/BA.4/BA.5/XBB cluster (**Fig. S8**).

224

225

226

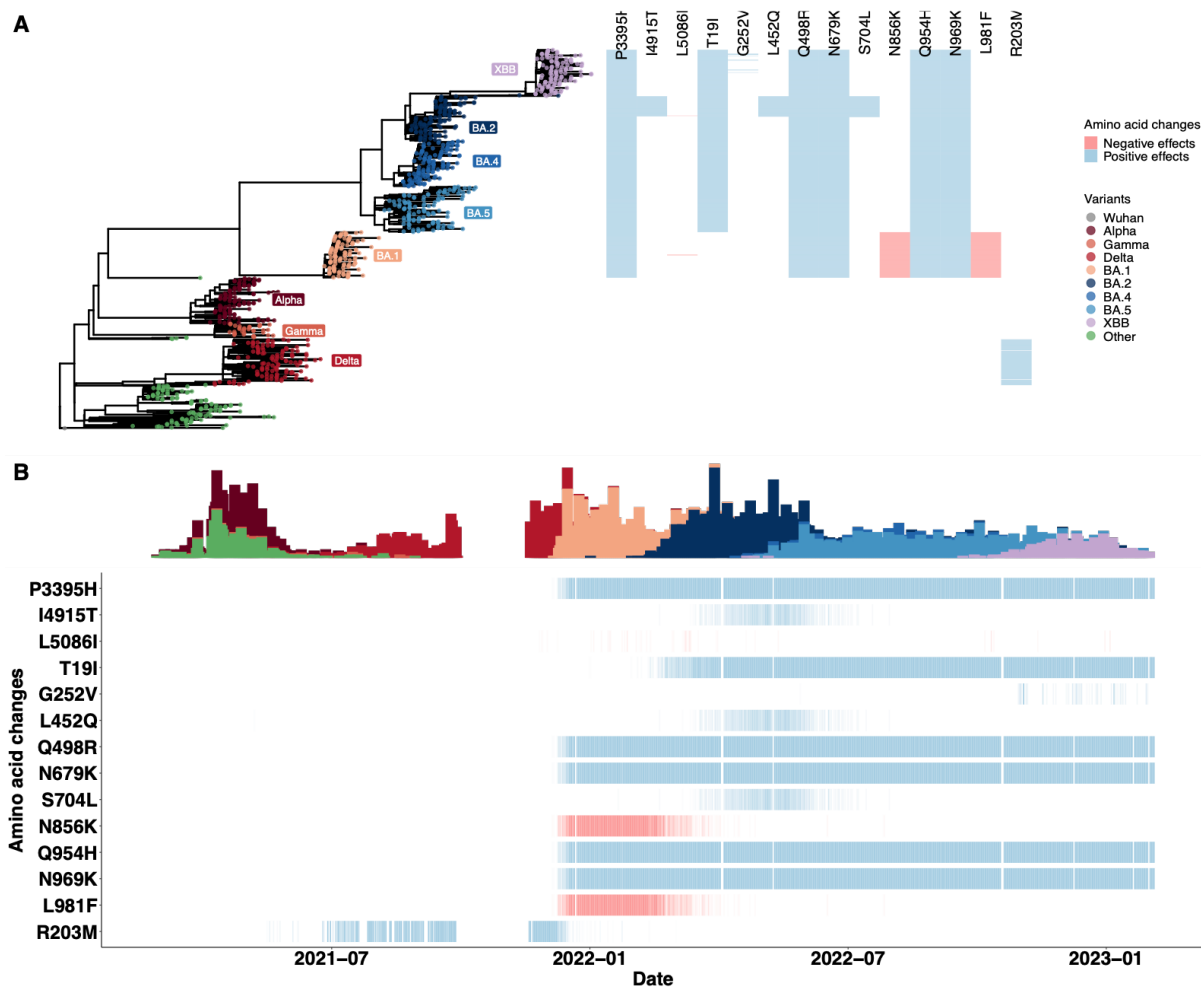


227
228
229
230
231
232
233
234
235
236
237

238 The impact of amino acid changes on viral copies is dependent on the
239 variant

240 Having identified the 14 non-synonymous SNPs with statistically significant effects on viral
241 copies in our primary analysis, we next sought to understand the temporal patterns of the
242 emergence of these amino acid changes (Fig. 4). To investigate the clustering of these SNPs,
243 we randomly sampled approximately 120 genome sequences from each VOC category (only
244 36 sequences were available for Gamma in our dataset) and generated a phylogenetic tree
245 drawing upon the subsamples (Fig. 4A). We found a clear pattern in how these mutations
246 emerged by VOC (Fig. 4A heatmap). We found that all amino acid changes associated with
247 a positive effect on viral copies were found in Delta and Omicron BA.2/BA.4/BA.5/XBB
248 infections. Often, more than one amino acid change was observed in each sampled
249 sequence, suggesting genetic linkage between these SNPs, as also shown in the epistasis

250 test (**Fig. S3**), such as S:Q954H and N969K. In particular, we identified that the amino acid
 251 changes S:L452Q ($p = 3.91 \times 10^{-25}$, $\beta = 0.34$, $\sigma = 0.03$) and S704L ($p = 1.35 \times 10^{-24}$, $\beta =$
 252 0.34 , $\sigma = 0.03$) associated with a positive effect on viral copies were typically observed in
 253 combination with BA.2 infections—specifically, lineage BA.2.12.1. We also observed that the
 254 amino acid changes with negative effects on viral copies (ORF1ab:L5086I, S:N856K and
 255 L981F) were only associated with BA.1 infections.
 256



257
 258 **Figure 4. The temporal dynamics of non-synonymous amino acid changes in the ORF1ab gene**
 259 **(P3395H, I4915T and L5086I), S gene (T19I, G252V, L452Q, Q498R, N679K, S704L, N856K, Q954H,**
 260 **N969K and L981F), and N gene (R203M) associated with changes in viral copies.** The results are
 261 based on the multivariate regression analysis using the sequence clusters (i.e., a categorical variable)
 262 inferred from the MDS components. (A) The phylogenetic tree estimated from a representative set of
 263 996 genome sequences showing variant assignments and the locations of amino acid changes that
 264 increase (blue) or decrease (red) viral copies. (B) The temporal dynamics of the SNPs from February
 265 2021 to March 2023. The transparency of the color corresponds to the mutation fraction in the daily
 266 sequence count: transparent color indicates low fractions, and opaque color indicates high fractions.
 267 The temporal dynamics of the SNPs, using MDS-inferred distance as a population control, are shown
 268 in **Fig. S9-11**.
 269

270 To explore the temporal dynamics of these amino acid changes, we calculated the fraction
 271 of SNPs occurring in the sequences for each day, thereby accounting for the number of
 272 introductions to the population (**Fig. 4B**). We observed most SNPs with a positive impact on
 273 viral copies emerging in sequences sampled from February 2022, when BA.2 was first
 274 detected in Connecticut. These SNPs were consistently observed in almost every sequence

275 thereafter. By contrast, we found that the other two amino acid changes (S:L452Q and S704L)
276 that had a positive effect on viral copies were only in the samples from BA.2 infections and
277 did not arise again in sublineages of BA.4 or BA.5. S:G252V was associated with higher viral
278 copies; however, we found that the SNP only appeared in a few sequences associated with
279 XBB infections. Notably, the N:R203M mutation was only associated with Delta infections.
280 For the SNPs (ORF1ab:L5086I, S:N856K and L981F) that had a negative association with viral
281 copies, we observed that they were present in samples associated with BA.1 infections and
282 did not persist when BA.1 was replaced by BA.2.
283

284 Discussion

285 We conducted a GWAS analysis on 9,902 high-quality SARS-CoV-2 genome sequences
286 generated from two years of genomic surveillance in Connecticut, US to identify and evaluate
287 SNPs that were associated with variations in viral copies during infections. Using a GWAS
288 approach, we were able to identify and examine virus-related factors that were associated
289 with the observed variations in viral copies independent of host factors. This was achieved
290 by combining data from a large cohort of individuals infected with different VOCs and
291 employing a regression model for viral copies that accounted for virus-level factors (i.e.,
292 specific SNPs and genetic background), adjusted for individual factors (i.e., age and
293 vaccination status). We identified several SNPs corresponding to non-synonymous amino
294 acid changes in the SARS-CoV-2 genome that were individually or jointly associated with the
295 variations in viral copies. In particular, temporal patterns of the SNPs revealed that SNPs
296 associated with increased viral copies were predominantly observed in Delta and Omicron
297 BA.2/BA.4/BA.5/XBB infections, whereas those associated with decreased viral copies were
298 mostly observed in infections with Omicron BA.1 variants.
299

300 Using a GWAS approach, we successfully identified a subset of variant-defining amino acid
301 changes in Delta and Omicron variants (**Fig. S12**). Note that we did not detect any
302 substitutions in the Alpha and Gamma variants (likely due to the low sample size for Gamma).
303 We also identified SNPs that did not define any major variant category, including S:L452Q
304 and S704L that were specifically associated with BA.2.12.1, a sublineage of BA.2 that briefly
305 dominated during the pandemic (i.e., dominated mainly in the US between March and May
306 2022). This highlights the application of GWAS for identifying SNPs associated with important
307 phenotypic effects without requiring a set of lineage-defining mutations to be defined a priori.
308 Nevertheless, there are several reasons why we only detected a subset of the SNPs that
309 defined different VOCs. Firstly, SNPs with small effect sizes may not be detected due to the
310 stringent statistical significance thresholds applied in GWAS. Secondly, lineage-defining
311 SNPs that are in low linkage disequilibrium with the causal mutations may not be detected
312 [39], even if they may be functionally relevant. Our results showcase how GWAS can help to
313 narrow the focus of SNPs associated with specific phenotypes, generating hypotheses for
314 further investigation.
315

316 A key result from our analysis is that SNPs associated with viral copies did not exhibit the
317 same temporal dynamics, even though they could have similar (either positive or negative)
318 effects on viral copies, suggesting they may have independent effects on viral copies. Some
319 amino acid changes, for example, ORF1ab:I4915T (positive effects), were only present in
320 samples with BA.2 infections and disappeared when new Omicron variants emerged. Other
321 SNPs (e.g., S:T19R), while also associated with higher viral copies, were observed and

322 persisted in all BA.2/BA.4/BA.5/XBB infections. The distinct temporal pattern of SNPs,
323 dependent on VOCs, may help explain the different fitness levels (e.g., intrinsic
324 transmissibility or immune escape) of each variant [40,41]. Notably, we found three SNPs,
325 ORF1ab:L5086I, S:N856K and L981F, were associated with decreased viral copies in BA.1
326 infections. The negative impact on viral copies should be interpreted with caution. Although
327 the possibility that these SNPs have a direct impact on reducing viral copies cannot be ruled
328 out, it is also likely that the estimated negative effects are due to a synthetic association with
329 other SNPs. Further study may be required to disentangle the direct effects of these SNPs
330 from the confounding influences of other genetic variations and to confirm their functional
331 impact on viral copies.

332
333 In this study, we employed a series of single SNP regression models to identify the underlying
334 SNPs associated with the changes in viral copies without accounting for potential interactions
335 between SNPs. We noted that several synonymous SNPs located in the ORF1ab gene were
336 identified to have an impact on viral copies. The synonymous SNPs were likely linked to non-
337 synonymous SNPs that were under positive selection. In such cases, the synonymous SNPs
338 can be carried along with the non-synonymous SNPs, resulting in their significance in the
339 GWAS analysis, as shown in the subsequent epistasis test (**Fig. S3**). Nevertheless, the
340 method provided an initial set of SNPs that are worth further exploration, pinpointing
341 important mutations associated with viral copies and providing valuable insights into the
342 overall genetic landscape of the viral population. The method, thus, represented an important
343 first step towards understanding detailed epistatic effects among these mutations on viral
344 copies. A paired or higher-order SNP regression study could be conducted as a subsequent
345 step to test potential interactions or joint effects among different SNPs.

346
347 There are limitations to our study. First, we assumed that the distribution of times between
348 infection and sample collection was similar through time and across variants as these data
349 were not available. Given our samples were taken frequently over a 2-year period, we do not
350 anticipate that this assumption will qualitatively impact our results. Second, our study
351 primarily focuses on the genetic variants in VOCs, neglecting other factors such as host
352 immune responses or environmental influences, partially captured by the host-associated
353 covariates, including age and vaccination status in this study, that may also contribute to the
354 changes in viral copies. Further study will be needed to address the impact of these factors
355 on viral copies, for example, genome-to-genome analysis to reveal the impact of host-viral
356 genetic interactions in SARS-CoV-2 infections [20,56]. Third, our data were obtained from a
357 specific geographic region, whose population diversity may not necessarily be similar to other
358 settings; therefore, extrapolating these findings to a broader population may require caution.
359 Additionally, focusing solely on consensus genomic changes in the analysis could overlook
360 the genetic diversity within the sample, which may also influence variations in viral load.
361 Despite these constraints, our study highlights the importance of sustained genomic
362 surveillance and the need for comprehensive analyses to understand the nuanced impact of
363 specific genetic variations on viral copies at the within-host level, and its implications for viral
364 transmissibility and immune escape at the population level. Further work and collaborative
365 efforts are essential to elucidate the complex interplay between viral genetics, host factors,
366 and the dynamics of transmission associated with emerging variants. Such studies could
367 inform predictive early warning public health systems regarding the emergence of potentially
368 highly transmissible viral strains based on their constellation of mutations.

369
370 Recently, Duesterwald et al. [12] used genome sequence data and a machine-learning
371 approach to predict cycle threshold (Ct) values of SARS-CoV-2 infections based on the k-

372 mers. Similar to our findings, they suggested that S:L452 and P681 were hallmarks of VOCs,
373 implying impacts on the observed Ct values in clinical samples. Although the machine-
374 learning approach may capture broader patterns and interactions within the genome on Ct
375 values, they lack interpretability compared to regression models. For example, regression-
376 based models could offer insights into the direct association between specific genetic
377 variants and viral copies. In addition, regression-based models may perform well even with
378 limited sample sizes [19], provided that the assumptions of the model are met and the
379 predictors are informative, whereas using machine-learning methods with small sample sizes
380 can be challenging. However, the viral GWAS method may not be appropriate in situations
381 where there is insufficient genetic diversity in the viral population under study, as this can
382 limit the power to detect meaningful associations between mutations and viral traits.
383 Additionally, it may not be suitable when the phenotypic traits of interest are not well-defined
384 or accurately measured.

385
386 With the availability of high-quality whole-genome sequences for SARS-CoV-2, we
387 demonstrated that GWAS analysis of the viral genome can identify SNPs that associate with
388 positive or negative impacts on viral copies in VOCs, revealing important biological insights
389 and enhancing our understanding of within-host viral dynamics. We argue that the application
390 of GWAS analyses to study viral genomes provides a particularly tractable tool to identify
391 potential SNPs of interest for further evaluation across different viral pathogens. It is
392 particularly useful to understand the genetic basis of viral virulence, transmission, resistance
393 to antiviral treatments, and host-virus interactions for several reasons. First, the small genome
394 size of viruses and high evolutionary rates make it easier to perform comprehensive genome-
395 wide scans for SNPs and to experimentally test the impacts of SNPs on specific traits.
396 Second, significant phenotypic variations (e.g., viral loads and antibody responses) are often
397 observed in viral infections, despite limited changes in the viral genome. GWAS can help to
398 identify SNPs that correlate with these phenotypic variations, providing insights into the
399 genetic basis of these traits. Third, the increasing accessibility to sequence viral genomes
400 makes it possible to perform GWAS on rich datasets, enabling in-depth analysis of the
401 temporal dynamics of viral evolution. Together, the applicability of GWAS analyses to study
402 viral genomes can provide a new approach for exploring the intricate interplay between
403 genetic mutations and phenotypes, informing strategies for managing and mitigating the
404 impact of emerging viral variants, and contributing to the development of potential
405 therapeutic interventions.

406 Materials & Methods

407 Ethics

408 The Institutional Review Board from the Yale University Human Research Protection Program
409 determined that the RT-qPCR testing and sequencing of de-identified remnant COVID-19
410 clinical samples obtained from clinical partners conducted in this study is not research
411 involving human subjects (IRB Protocol ID: 2000028599).

412 Clinical sample collection and measurement of viral copies by RT-qPCR
413 SARS-CoV-2 positive samples (nasal swabs in viral transport media) were collected through
414 the Yale New Haven Hospital (YNHH) System as a part of routine inpatient and outpatient

415 testing and sent to the Yale SARS-CoV-2 Genomic Surveillance Initiative. Using the MagMAX
416 viral/pathogen nucleic acid isolation kit, nucleic acid was extracted from 300µl of each clinical
417 sample and eluted into 75µl of elution buffer. Extracted nucleic acid was then used as
418 template for a “research use only” (RUO) RT-qPCR assay [23] to test for presence of SARS-
419 CoV-2 RNA. Ct values from the nucleocapsid target (CDC-N1 primer-probe set [57]) were
420 used to derive viral copy numbers using a previously determined standard curve for this
421 primer set [58]. A positive RNA control with defined viral copy number (1000/µl) was used to
422 standardize results across individual runs.

423 Whole genome sequencing

424 Libraries were prepared for sequencing using the Illumina COVIDSeq Test (RUO version) and
425 quantified using the Qubit High Sensitivity dsDNA kit. Negative controls were included for
426 RNA extraction, cDNA synthesis, and amplicon generation. Prepared libraries were
427 sequenced at the Yale Center for Genomic Analysis on the Illumina NovaSeq with a 2x150
428 approach and at least 1 million reads per sample.

429
430 Reads were then aligned to the Wuhan-Hu-1 reference genome (GenBank MN908937.3)
431 using BWA-MEM v.0.7.15 [59]. Adaptor sequences were then trimmed, primer sequences
432 masked, and consensus genomes called (simple majority >60% frequency) using iVar
433 v1.3.133 [60] and SAMtools v1.11 [61]. When <20 reads were present at a site an ambiguous
434 “N” was used, with negative controls consisting of ≥99% Ns. The Pangolin lineage assignment
435 tool [62] was used for assigning viral lineages.

436 Clinical metadata

437 We obtained patient metadata and vaccination records from the YNHH system and the
438 Center for Outcomes Research and Evaluation (CORE) and matched these records to
439 sequencing data through unique sample identifiers. Duplicate patient records or those with
440 missing or inconsistent metadata and vaccination date were removed from the GWAS
441 analysis. We also removed patient records with persistent infections (>28 days since first
442 positive test).

443 We determined vaccination status at time of infection by comparing the sample collection
444 date to the patient’s vaccination record dates. We categorized vaccine statuses based on
445 the number of vaccine doses received at least 14 days before the collection date. Patient
446 vaccination statuses at the time of infection were categorized as follows: non-vaccine, one-
447 dose vaccine, two-dose vaccine, two-dose vaccine with one booster, or two-dose vaccine
448 with two boosters. We calculated the age of each patient as the difference between the
449 date of birth and the sampling date.

450 Single nucleotide polymorphisms

451 To identify single nucleotide polymorphisms (SNPs), we first aligned the 9902 genome
452 sequences using *nextalign* (v3.2.1) [63] with the reference genome of the Wuhan-Hu-1
453 genome (GenBank accession: MN908937.3). Then, SNPs were identified using *snp-sites*
454 (v2.4.1) [64], with the reference genome of the Wuhan-Hu-1 genome (GenBank accession:
455 MN908937.3). We also normalized the SNPs in the generated VCF file, such that multiallelic
456 SNPs were separated into different rows. Normalizing the SNPs ensured that each SNP was
457 one-hot encoded and analyzed separately. Note that we did not include ambiguous SNPs,

458 deletions and insertions in our GWAS analysis. We used *vcf-annotator* (v0.7) to annotate
459 SNPs to corresponding amino acid changes.

460 Multidimensional scaling and population control

461 To reveal the underlying structure of the 9902 genome sequences. We first used *snp-dists*
462 (v0.7.0) [65] to convert the aligned sequences (a FASTA alignment) to a SNP distance matrix.
463 We then applied a multidimensional scaling (MDS) method [26] to transform the SNP distance
464 matrix into a geometric configuration while preserving the original pairwise relationships. The
465 scaling was conducted using *cmdscale* function in an R package *stats* (v3.6.2). We set the
466 maximal dimensional parameter $k = 2$.

467
468 To measure the goodness of the transformation, we calculated the distance between the
469 original genome sequencing data and compared it with the new distances determined by
470 MDS. This involved arranging the two matrices of distances into two columns and computing
471 the correlation coefficient (i.e., r) between them. Finally, we used r^2 to measure the
472 proportion of variance in the original distance matrix explained by the new computed distance
473 matrix.

474
475 To determine the clusters (i.e., categorical variables) from MDS, we applied the k-means
476 clustering method using the *kmeans* function implemented in R statistical software (v4.0.2).
477 We set the number of centroids $k = 4$.
478

479 Testing for associations between viral copies and SNPs

480 In this work, we conducted a series of single SNP regression analyses to test for associations
481 between viral copies and SNPs, adjusted for host ages, vaccination status and viral
482 population. The linear regression model is written as follows:

$$483 \quad Y \sim \alpha W + \beta_i SNP_i + e,$$

484 where Y is a vector of normalized \log_{10} -transformed viral copies, W is a matrix of covariates,
485 including age (a categorical variable with four age groups of “<18”, “18-49”, “50-69”, and
486 “>70” years old), vaccination status (a categorical variable with vaccination statuses of “0
487 doses”, “1 dose”, “2 doses”, “2 doses and 1 booster”, “2 doses and 2 boosters”), a
488 population control variable for different viral variants (a categorical variable with cluster
489 numbers of “1”, “2”, “3” and “4”, see **Fig. S1** for detailed clusters), and an intercept, and α is
490 a vector that corresponds to coefficients of the covariates. In particular, SNP_i is a vector of
491 genotype values for all samples at each SNP, i . It is a binary variable: 0 represents the SNP
492 is not present in the genome sequence, whereas 1 represents its presence. β_i is the effective
493 size of each identified SNP, i . We also conducted a sensitivity analysis including two terms
494 $\xi_1 d_1, \xi_2 d_2$ as covariates in the model for population control. The vectors d_1, d_2 represent the
495 two dimensions computed by MDS, and ξ_1, ξ_2 are the coefficients of the dimension
496 covariates. The random effect of residual errors is presented here by e , which is assumed to
497 follow a normal distribution with a mean of 0 and a standard deviation of σ_e .

498 Marginal epistasis test

499 We applied a marginal epistasis test method to explore the interactions between SNPs on
500 viral copies, using an R package *mvMAPIT* (v.2.0.3) [28–30]. This method maps SNPs with
501 non-zero marginal epistatic effects—the combined pairwise interaction effects between a

502 given SNP and all other SNPs—identifying candidate variants involved in epistasis without
503 needing to identify the exact partners with which the variants interact.
504

505 Phylogenetic tree construction and comparison to variant-defining 506 substitutions

507 We employed *iq-tree* (v2.2.2.6) [66] of a representative set using 996 of our 9902 genome
508 sequences for tree construction, using Wuhan-Hu-1 (GenBank MN908937.3) as the reference
509 genome. We specified the HKY substitution model and set the number of bootstrap replicates
510 to 1,000. To visualize the phylogenetic tree, we used the *ggtree* (v1.4.11) implemented in the
511 R statistical software (v4.0.2). The variant-defining amino acid changes were defined as those
512 mutations with >75% prevalence in at least one lineage, as estimated on outbreak.info
513 website [38]. Note that we did not include deletions in variant-defining substitutions.

514 Data and code availability

515 We used the R statistical software (v4.0.2) for all statistical analyses and visualization. Data
516 and code used in this study are publicly available on Github:
517 https://github.com/grubaughlab/2024_paper_GWAS. All genome sequences used for the
518 GWAS analysis and a subset of the associated metadata (accession number, virus name,
519 collection date, originating lab and submitting lab, and the list of authors) in this dataset are
520 published in GISAID's EpiCoV database: <https://doi.org/10.55876/gis8.240219fh>. The de-
521 identified and coded clinical metadata associated with the sequenced samples are available
522 upon request with IRB approval.

523 Acknowledgements

524 We would like to thank Verity Hill, Seth Redmond, Jiye Kwon, Rafael Lopes, Sophie Taylor,
525 and Philip Jack for their helpful conversations and feedback on this work. This project is
526 supported by the Centers for Disease Control and Prevention (CDC) Broad Agency
527 Announcement Contract 75D30122C14697 (NDG). This work does not necessarily represent
528 the views of the CDC.
529

530 Competing Interest Statement

531 NDG is a paid consultant for BioNTech, DMW has received consulting fees from Pfizer,
532 Merck, and GSK, unrelated to this manuscript, and has been PI on research grants from
533 Pfizer and Merck to Yale, unrelated to this manuscript.

534 Yale SARS-CoV-2 Genomic Surveillance Initiative Authors

535 Tara Alpert, Kaya Bilguvar, Kendall Billig, Mallery Breban, Anderson Brito, Christopher
536 Castaldi, Rebecca Earnest, Bony De Kumar, Joseph Fauver, Chaney Kalinich, Tobias Koch,
537 Marie Landry, Shrikant Mane, Isabel Ott, David Peaper, Mary Petrone, Kien Pham, Jessica
538 Rothman, Irina Tikhonova, Chantal Vogels, Anne Watkins
539

540 References

- 541 1. Killingley B, Mann AJ, Kalinova M, Boyers A, Goonawardane N, Zhou J, et al. Safety, tolerability
542 and viral kinetics during SARS-CoV-2 human challenge in young adults. *Nat Med.* 2022;28:
543 1031–1041. doi:10.1038/s41591-022-01780-9
- 544 2. Marks M, Millat-Martinez P, Ouchi D, Roberts CH, Alemany A, Corbacho-Monné M, et al.
545 Transmission of COVID-19 in 282 clusters in Catalonia, Spain: a cohort study. *Lancet Infect Dis.*
546 2021;21: 629–636. doi:10.1016/S1473-3099(20)30985-3
- 547 3. Jones TC, Biele G, Mühlemann B, Veith T, Schneider J, Beheim-Schwarzbach J, et al.
548 Estimating infectiousness throughout SARS-CoV-2 infection course. *Science.* 2021;373.
549 doi:10.1126/science.abi5273
- 550 4. Cevik M, Tate M, Lloyd O, Maraolo AE, Schafers J, Ho A. SARS-CoV-2, SARS-CoV, and MERS-
551 CoV viral load dynamics, duration of viral shedding, and infectiousness: a systematic review and
552 meta-analysis. *Lancet Microbe.* 2021;2: e13–e22. doi:10.1016/S2666-5247(20)30172-5
- 553 5. Singanayagam A, Patel M, Charlett A, Lopez Bernal J, Saliba V, Ellis J, et al. Duration of
554 infectiousness and correlation with RT-PCR cycle threshold values in cases of COVID-19,
555 England, January to May 2020. *Euro Surveill.* 2020;25. doi:10.2807/1560-
556 7917.ES.2020.25.32.2001483
- 557 6. Singanayagam A, Hakki S, Dunning J, Madon KJ, Crone MA, Koycheva A, et al. Community
558 transmission and viral load kinetics of the SARS-CoV-2 delta (B.1.617.2) variant in vaccinated
559 and unvaccinated individuals in the UK: a prospective, longitudinal, cohort study. *Lancet Infect*
560 *Dis.* 2022;22: 183–195. doi:10.1016/S1473-3099(21)00648-4
- 561 7. Kissler SM, Fauver JR, Mack C, Tai CG, Breban MI, Watkins AE, et al. Viral Dynamics of SARS-
562 CoV-2 Variants in Vaccinated and Unvaccinated Persons. *N Engl J Med.* 2021;385: 2489–2491.
563 doi:10.1056/NEJMc2102507
- 564 8. Puhach O, Adea K, Hulo N, Sattoune P, Genecand C, Iten A, et al. Infectious viral load in
565 unvaccinated and vaccinated individuals infected with ancestral, Delta or Omicron SARS-CoV-2.
566 *Nat Med.* 2022;28: 1491–1500. doi:10.1038/s41591-022-01816-0
- 567 9. Boucau J, Marino C, Regan J, Uddin R, Choudhary MC, Flynn JP, et al. Duration of Shedding of
568 Culturable Virus in SARS-CoV-2 Omicron (BA.1) Infection. *N Engl J Med.* 2022;387: 275–277.
569 doi:10.1056/NEJMc2202092
- 570 10. Hay JA, Kennedy-Shaffer L, Kanjilal S, Lennon NJ, Gabriel SB, Lipsitch M, et al. Estimating
571 epidemiologic dynamics from cross-sectional viral load distributions. *Science.* 2021;373.
572 doi:10.1126/science.abh0635
- 573 11. Fryer HR, Golubchik T, Hall M, Fraser C, Hinch R, Ferretti L, et al. Viral burden is associated with
574 age, vaccination, and viral variant in a population-representative study of SARS-CoV-2 that
575 accounts for time-since-infection-related sampling bias. *PLoS Pathog.* 2023;19: e1011461.
576 doi:10.1371/journal.ppat.1011461
- 577 12. Duesterwald L, Nguyen M, Christensen P, Wesley Long S, Olsen RJ, Musser JM, et al. Using
578 Genome Sequence Data to Predict SARS-CoV-2 Detection Cycle Threshold Values.
579 doi:10.1101/2022.11.14.22282297
- 580 13. Uffelmann E, Huang QQ, Munung NS, de Vries J, Okada Y, Martin AR, et al. Genome-wide
581 association studies. *Nature Reviews Methods Primers.* 2021;1: 1–21. doi:10.1038/s43586-021-
582 00056-9
- 583 14. McLaren PJ, Porreca I, Iaconis G, Mok HP, Mukhopadhyay S, Karakoc E, et al. Africa-specific

- 584 human genetic variation near CHD1L associates with HIV-1 load. *Nature*. 2023;620: 1025–1030.
585 doi:10.1038/s41586-023-06370-4
- 586 15. Karim M, Dunham I, Ghoussaini M. Mining a GWAS of Severe Covid-19. *The New England*
587 *journal of medicine*. 2020. pp. 2588–2589. doi:10.1056/NEJMc2025747
- 588 16. Roberts GHL, Partha R, Rhead B, Knight SC, Park DS, Coignet MV, et al. Expanded COVID-19
589 phenotype definitions reveal distinct patterns of genetic association and protective effects. *Nat*
590 *Genet*. 2022;54: 374–381. doi:10.1038/s41588-022-01042-x
- 591 17. Genomewide Association Study of Severe Covid-19 with Respiratory Failure. *N Engl J Med*.
592 2020;383: 1522–1534. doi:10.1056/NEJMoa2020283
- 593 18. Power RA, Parkhill J, de Oliveira T. Microbial genome-wide association studies: lessons from
594 human GWAS. *Nat Rev Genet*. 2017;18: 41–50. doi:10.1038/nrg.2016.132
- 595 19. Power RA, Davaniah S, Derache A, Wilkinson E, Tanser F, Gupta RK, et al. Genome-Wide
596 Association Study of HIV Whole Genome Sequences Validated using Drug Resistance. *PLoS*
597 *One*. 2016;11: e0163746. doi:10.1371/journal.pone.0163746
- 598 20. Ansari MA, Pedergrana V, L C Ip C, Magri A, Von Delft A, Bonsall D, et al. Genome-to-genome
599 analysis highlights the effect of the human innate and adaptive immune systems on the hepatitis
600 C virus. *Nat Genet*. 2017;49: 666–673. doi:10.1038/ng.3835
- 601 21. Ansari MA, Aranday-Cortes E, Ip CL, da Silva Filipe A, Lau SH, Bamford C, et al. Interferon
602 lambda 4 impacts the genetic diversity of hepatitis C virus. *Elife*. 2019;8.
603 doi:10.7554/eLife.42463
- 604 22. Hahn G, Wu CM, Lee S, Lutz SM, Khurana S, Baden LR, et al. Genome-wide association
605 analysis of COVID-19 mortality risk in SARS-CoV-2 genomes identifies mutation in the SARS-
606 CoV-2 spike protein that colocalizes with P.1 of the Brazilian strain. *Genet Epidemiol*. 2021;45:
607 685–693. doi:10.1002/gepi.22421
- 608 23. Vogels CBF, Breban MI, Ott IM, Alpert T, Petrone ME, Watkins AE, et al. Multiplex qPCR
609 discriminates variants of concern to enhance global surveillance of SARS-CoV-2. *PLoS Biol*.
610 2021;19: e3001236. doi:10.1371/journal.pbio.3001236
- 611 24. Zhou C, Zhang T, Ren H, Sun S, Yu X, Sheng J, et al. Impact of age on duration of viral RNA
612 shedding in patients with COVID-19. *Aging*. 2020;12: 22399–22404. doi:10.18632/aging.104114
- 613 25. Acharya CB, Schrom J, Mitchell AM, Coil DA, Marquez C, Rojas S, et al. Viral Load Among
614 Vaccinated and Unvaccinated, Asymptomatic and Symptomatic Persons Infected With the
615 SARS-CoV-2 Delta Variant. *Open Forum Infect Dis*. 2022;9: ofac135. doi:10.1093/ofid/ofac135
- 616 26. Torgerson WS. Multidimensional scaling: I. Theory and method. *Psychometrika*. 1952;17: 401–
617 419. doi:10.1007/bf02288916
- 618 27. VanderWeele TJ, Mathur MB. SOME DESIRABLE PROPERTIES OF THE BONFERRONI
619 CORRECTION: IS THE BONFERRONI CORRECTION REALLY SO BAD? *Am J Epidemiol*.
620 2018;188: 617–618. doi:10.1093/aje/kwy250
- 621 28. Crawford L, Zeng P, Mukherjee S, Zhou X. Detecting epistasis with the marginal epistasis test in
622 genetic mapping studies of quantitative traits. *PLoS Genet*. 2017;13: e1006869.
623 doi:10.1371/journal.pgen.1006869
- 624 29. Stamp J, DenAdel A, Weinreich D, Crawford L. Leveraging the genetic correlation between traits
625 improves the detection of epistasis in genome-wide association studies. *G3*. 2023;13.
626 doi:10.1093/g3journal/jkad118
- 627 30. Zhou X. A UNIFIED FRAMEWORK FOR VARIANCE COMPONENT ESTIMATION WITH
628 SUMMARY STATISTICS IN GENOME-WIDE ASSOCIATION STUDIES. *Ann Appl Stat*. 2017;11:

- 629 2027–2051. doi:10.1214/17-AOAS1052
- 630 31. Ward IL, Bermingham C, Ayoubkhani D, Gethings OJ, Pouwels KB, Yates T, et al. Risk of covid-
631 19 related deaths for SARS-CoV-2 omicron (B.1.1.529) compared with delta (B.1.617.2):
632 retrospective cohort study. *BMJ*. 2022;378: e070695. doi:10.1136/bmj-2022-070695
- 633 32. Volz E, Hill V, McCrone JT, Price A, Jorgensen D, O’Toole Á, et al. Evaluating the Effects of
634 SARS-CoV-2 Spike Mutation D614G on Transmissibility and Pathogenicity. *Cell*. 2021;184: 64–
635 75.e11. doi:10.1016/j.cell.2020.11.020
- 636 33. Tian D, Sun Y, Zhou J, Ye Q. The Global Epidemic of the SARS-CoV-2 Delta Variant, Key Spike
637 Mutations and Immune Escape. *Front Immunol*. 2021;12: 751778.
638 doi:10.3389/fimmu.2021.751778
- 639 34. Hodcroft EB, Domman DB, Snyder DJ, Oguntuyo KY, Van Diest M, Densmore KH, et al.
640 Emergence in late 2020 of multiple lineages of SARS-CoV-2 Spike protein variants affecting
641 amino acid position 677. *medRxiv*. 2021. doi:10.1101/2021.02.12.21251658
- 642 35. Cosar B, Karagulleoglu ZY, Unal S, Ince AT, Uncuoglu DB, Tuncer G, et al. SARS-CoV-2
643 Mutations and their Viral Variants. *Cytokine Growth Factor Rev*. 2022;63: 10–22.
644 doi:10.1016/j.cytogfr.2021.06.001
- 645 36. Chen J, Wang R, Wang M, Wei G-W. Mutations Strengthened SARS-CoV-2 Infectivity. *J Mol*
646 *Biol*. 2020;432: 5212–5226. doi:10.1016/j.jmb.2020.07.009
- 647 37. Harvey WT, Carabelli AM, Jackson B, Gupta RK, Thomson EC, Harrison EM, et al. SARS-CoV-2
648 variants, spike mutations and immune escape. *Nat Rev Microbiol*. 2021;19: 409–424.
649 doi:10.1038/s41579-021-00573-0
- 650 38. Gangavarapu K, Latif AA, Mullen JL, Alkuzweny M, Hufbauer E, Tsueng G, et al. Outbreak.info
651 genomic reports: scalable and dynamic surveillance of SARS-CoV-2 variants and mutations. *Nat*
652 *Methods*. 2023;20: 512–522. doi:10.1038/s41592-023-01769-3
- 653 39. Tang X, Wu C, Li X, Song Y, Yao X, Wu X, et al. On the origin and continuing evolution of SARS-
654 CoV-2. *Natl Sci Rev*. 2020;7: 1012–1023. doi:10.1093/nsr/nwaa036
- 655 40. Carabelli AM, Peacock TP, Thorne LG, Harvey WT, Hughes J, COVID-19 Genomics UK
656 Consortium, et al. SARS-CoV-2 variant biology: immune escape, transmission and fitness. *Nat*
657 *Rev Microbiol*. 2023;21: 162–177. doi:10.1038/s41579-022-00841-7
- 658 41. Souza PFN, Mesquita FP, Amaral JL, Landim PGC, Lima KRP, Costa MB, et al. The spike
659 glycoprotein of SARS-CoV-2: A review of how mutations of spike glycoproteins have driven the
660 emergence of variants with high transmissibility and immune escape. *Int J Biol Macromol*.
661 2022;208: 105–125. doi:10.1016/j.ijbiomac.2022.03.058
- 662 42. Kimura I, Kosugi Y, Wu J, Zahradnik J, Yamasoba D, Butlertanaka EP, et al. The SARS-CoV-2
663 Lambda variant exhibits enhanced infectivity and immune resistance. *Cell Rep*. 2022;38:
664 110218. doi:10.1016/j.celrep.2021.110218
- 665 43. Yadav PD, Bergeron É, Flora MS. Emerging SARS-COV-2 Variants: Genomic Variations,
666 Transmission, Pathogenesis, Clinical Impact and Interventions. *Frontiers Media SA*; 2023.
667 Available: <https://play.google.com/store/books/details?id=AHi6EAAAQBAJ>
- 668 44. BMvd V, Dingemans J, Bank LE, von Wintersdorff CJ. Viral load dynamics in healthcare workers
669 with COVID-19 during Delta and Omicron era. [cited 10 Feb 2024]. Available:
670 <https://europepmc.org/article/ppr/ppr485029>
- 671 45. Lentini A, Pereira A, Winqvist O, Reinius B. Monitoring of the SARS-CoV-2 Omicron BA.1/BA.2
672 variant transition in the Swedish population reveals higher viral quantity in BA.2 cases. *bioRxiv*.
673 2022. doi:10.1101/2022.03.26.22272984

- 674 46. Russell TW, Townsley H, Abbott S, Hellewell J, Carr EJ, Chapman LAC, et al. Combined
675 analyses of within-host SARS-CoV-2 viral kinetics and information on past exposures to the
676 virus in a human cohort identifies intrinsic differences of Omicron and Delta variants. *PLoS Biol.*
677 2024;22: e3002463. doi:10.1371/journal.pbio.3002463
- 678 47. Kopsidas I, Karagiannidou S, Kostaki EG, Kousi D, Douka E, Sfikakis PP, et al. Global
679 Distribution, Dispersal Patterns, and Trend of Several Omicron Subvariants of SARS-CoV-2
680 across the Globe. *Trop Med Infect Dis.* 2022;7. doi:10.3390/tropicalmed7110373
- 681 48. Motozono C, Toyoda M, Tan TS, Hamana H, Goto Y, Aritsu Y, et al. The SARS-CoV-2 Omicron
682 BA.1 spike G446S mutation potentiates antiviral T-cell recognition. *Nat Commun.* 2022;13:
683 5440. doi:10.1038/s41467-022-33068-4
- 684 49. Lin X, Sha Z, Trimpert J, Kunec D, Jiang C, Xiong Y, et al. The NSP4 T492I mutation increases
685 SARS-CoV-2 infectivity by altering non-structural protein cleavage. *Cell Host Microbe.* 2023;31:
686 1170–1184.e7. doi:10.1016/j.chom.2023.06.002
- 687 50. Luo CH, Morris CP, Sachithanandham J, Amadi A, Gaston D, Li M, et al. Infection with the
688 SARS-CoV-2 Delta Variant is Associated with Higher Infectious Virus Loads Compared to the
689 Alpha Variant in both Unvaccinated and Vaccinated Individuals. *medRxiv.* 2021.
690 doi:10.1101/2021.08.15.21262077
- 691 51. von Wintersdorff CJH, Dingemans J, van Alphen LB, Wolffs PFG, van der Veer BMJW, Hoebe
692 CJPA, et al. Infections with the SARS-CoV-2 Delta variant exhibit fourfold increased viral loads
693 in the upper airways compared to Alpha or non-variants of concern. *Sci Rep.* 2022;12: 13922.
694 doi:10.1038/s41598-022-18279-5
- 695 52. Li B, Deng A, Li K, Hu Y, Li Z, Shi Y, et al. Viral infection and transmission in a large, well-traced
696 outbreak caused by the SARS-CoV-2 Delta variant. *Nat Commun.* 2022;13: 460.
697 doi:10.1038/s41467-022-28089-y
- 698 53. Fall A, Eldesouki RE, Sachithanandham J, Morris CP, Norton JM, Gaston DC, et al. The
699 displacement of the SARS-CoV-2 variant Delta with Omicron: An investigation of hospital
700 admissions and upper respiratory viral loads. *EBioMedicine.* 2022;79: 104008.
701 doi:10.1016/j.ebiom.2022.104008
- 702 54. Hay JA, Kissler SM, Fauver JR, Mack C, Tai CG, Samant RM, et al. Quantifying the impact of
703 immune history and variant on SARS-CoV-2 viral kinetics and infection rebound: A retrospective
704 cohort study. *Elife.* 2022;11. doi:10.7554/eLife.81849
- 705 55. Jones RP, Ponomarenko A. COVID-19-Related Age Profiles for SARS-CoV-2 Variants in England
706 and Wales and States of the USA (2020 to 2022): Impact on All-Cause Mortality. *Infect Dis Rep.*
707 2023;15: 600–634. doi:10.3390/idr15050058
- 708 56. Bartha I, Carlson JM, Brumme CJ, McLaren PJ, Brumme ZL, John M, et al. A genome-to-
709 genome analysis of associations between human genetic variation, HIV-1 sequence diversity,
710 and viral control. *Elife.* 2013;2: e01123. doi:10.7554/eLife.01123
- 711 57. Lu X, Wang L, Sakthivel SK, Whitaker B, Murray J, Kamili S, et al. US CDC Real-Time Reverse
712 Transcription PCR Panel for Detection of Severe Acute Respiratory Syndrome Coronavirus 2.
713 *Emerg Infect Dis.* 2020;26: 1654–1665. doi:10.3201/eid2608.201246
- 714 58. Vogels CBF, Brito AF, Wyllie AL, Fauver JR, Ott IM, Kalinich CC, et al. Analytical sensitivity and
715 efficiency comparisons of SARS-CoV-2 RT-qPCR primer-probe sets. *Nature Microbiology.*
716 2020;5: 1299–1305. doi:10.1038/s41564-020-0761-6
- 717 59. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv*
718 [q-bio.GN]. 2013. Available: <http://arxiv.org/abs/1303.3997>
- 719 60. Grubaugh ND, Gangavarapu K, Quick J, Matteson NL, De Jesus JG, Main BJ, et al. An

- 720 amplicon-based sequencing framework for accurately measuring intrahost virus diversity using
721 PrimalSeq and iVar. *Genome Biol.* 2019;20: 8. doi:10.1186/s13059-018-1618-7
- 722 61. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of
723 SAMtools and BCFtools. *Gigascience.* 2021;10. doi:10.1093/gigascience/giab008
- 724 62. O'Toole Á, Scher E, Underwood A, Jackson B, Hill V, McCrone JT, et al. Assignment of
725 epidemiological lineages in an emerging pandemic using the pangolin tool. *Virus Evol.* 2021;7:
726 veab064. doi:10.1093/ve/veab064
- 727 63. Aksamentov I, Roemer C, Hodcroft E, Neher R. Nextclade: clade assignment, mutation calling
728 and quality control for viral genomes. *J Open Source Softw.* 2021;6: 3773.
729 doi:10.21105/joss.03773
- 730 64. Page AJ, Taylor B, Delaney AJ, Soares J, Seemann T, Keane JA, et al. SNP-sites: rapid efficient
731 extraction of SNPs from multi-FASTA alignments. *Microb Genom.* 2016;2: e000056.
732 doi:10.1099/mgen.0.000056
- 733 65. Creators Seemann, Torsten1 Show affiliations 1. The University of Melbourne. Source code for
734 snp-dists software. doi:10.5281/zenodo.1411986
- 735 66. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, et al.
736 Corrigendum to: IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in
737 the Genomic Era. *Mol Biol Evol.* 2020;37: 2461. doi:10.1093/molbev/msaa131
- 738
- 739

740 Supplementary Figures & Tables

741

742 **Supplemental Figure 1. Results of multidimensional scaling.** The population structure of the 9902
743 genome sequences using a multidimensional scaling (MDS) method. Clusters are defined using a *k*-
744 means clustering method, as demonstrated on the bottom right corner.

745

746 **Supplemental Figure 2. Q-Q plots of GWAS p-values.** Q-Q plots (quantile-quantile plots) showing
747 the p-values from GWAS analysis using (A) two MDS-computed components, or (B) MDS-inferred four
748 clusters as covariates in the regression model.

749

750 **Supplemental Figure 3. Marginal epistasis tests identify single nucleotide polymorphisms (SNPs)
751 that have epistatic interactions with others and are associated with the changes in viral copies.**

752 (A) Marginal epistasis test results of the SNPs (annotated as amino acid changes) that have marginal
753 epistatic effects on viral copies. The dashed line indicates the permuted threshold for genome-wide
754 significance $p = 0.05/171 = 2.74 \times 10^{-4}$. Significant mutations are shown with solid colors. (B) The *p*-
755 values and (C) the effect size of pairwise interaction tests among the significant mutations.

756

757 **Supplemental Figure 4. GWAS analysis identifies several single nucleotide polymorphisms
758 (SNPs) that are associated with the changes in viral copies.** (A) Genome-wide association results

759 of the impact of identified SNPs on viral copies during SARS-CoV-2 infection. The dashed line
760 indicates the permuted threshold for genome-wide significance $p = 4.67 \times 10^{-6}$. Significant SNPs are
761 shown with solid colors. (B) SNPs (with $p < 1 \times 10^{-10}$) that have positive (blue) or negative (red) effects
762 on viral copies. (C) The corresponding synonymous (triangles) and non-synonymous (circles) amino
763 acid changes that associate with increased or decreased viral copies. Data is shown as means with
764 95% confidence intervals. The estimated effective sizes and associated standard deviations are given
765 in Table S1. A Q-Q plot showing the observed distribution of p-value and the expected distribution is
766 given in Fig. S2.

767

768 **Supplemental Figure 5. GWAS analysis using only Cluster 1 data (shown in Fig. S1).** (A) Genome-

769 wide association results of the impact of identified SNPs on viral copies during SARS-CoV-2 infection.
770 The dashed line indicates the permuted threshold for genome-wide significance $p = 4.03 \times 10^{-5}$
771 (0.05/1242 SNPs). Significant SNPs are shown with solid colors. (B) SNPs (with $p < 1 \times 10^{-10}$) that have
772 positive (blue) or negative (red) effects on viral copies. (C) The corresponding synonymous (triangles)
773 amino acid changes that associate with increased or decreased viral copies. Data shown as means
774 with 95% confidence intervals.

775

776 **Supplemental Figure 6. GWAS analysis using Cluster 2 data (shown in Fig. S1).** (A) Genome-wide

777 association results of the impact of identified SNPs on viral copies during SARS-CoV-2 infection. The
778 dashed line indicates the permuted threshold for genome-wide significance $p = 3.68 \times 10^{-5}$ (0.05/1357
779 SNPs). Significant SNPs are shown with solid colors. (B) SNPs (with $p < 1 \times 10^{-10}$) that have positive
780 (blue) or negative (red) effects on viral copies. (C) The corresponding synonymous (triangles) amino
781 acid changes that associate with increased or decreased viral copies. Data shown as means with 95%
782 confidence intervals.

783

784 **Supplemental Figure 7. GWAS analysis using Cluster 3 data (shown in Fig. S1).** (A) Genome-

785 wide association results of the impact of identified SNPs on viral copies during SARS-CoV-2
786 infection. The dashed line indicates the permuted threshold for genome-wide significance $p =$
787 2.80×10^{-5} (0.05/1784 SNPs). Significant SNPs are shown with solid colors. (B) SNPs (with $p <$
788 1×10^{-10}) that have positive (blue) or negative (red) effects on viral copies. (C) The corresponding
789 non-synonymous (circles) amino acid changes that associate with increased or decreased viral
790 copies. Data shown as means with 95% confidence intervals.

791

792 **Supplemental Figure 8. GWAS analysis using Cluster 4 data (shown in Fig. S1).** (A) Genome-wide
793 association results of the impact of identified SNPs on viral copies during SARS-CoV-2 infection. The
794 dashed line indicates the permuted threshold for genome-wide significance $p = 7.91 \times 10^{-6}$ (0.05/6314
795 SNPs). Significant SNPs are shown with solid colors. (B) SNPs (with $p < 1 \times 10^{-10}$) that have positive
796 (blue) or negative (red) effects on viral copies. (C) The corresponding synonymous (triangles) and non-
797 synonymous (circles) amino acid changes that associate with increased or decreased viral copies.
798 Data shown as means with 95% confidence intervals.
799

800 **Supplemental Figure 9. The temporal dynamics of amino acid changes in the S gene**
801 **associated with changes in viral copies.** The results are based on the multivariate regression
802 analysis using the two MDS components as covariates. (A) The phylogenetic tree estimated from a
803 representative set of 996 genome sequences showing variant assignments and the locations of
804 amino acid changes that increase (blue) or decrease (red) viral copies. (B) The temporal dynamics of
805 the SNPs from February 2021 to March 2023. The transparency of the color corresponds to the
806 mutation fraction in the daily sequence count: transparent color indicates low fractions, and opaque
807 color indicates high fractions.
808

809 **Supplemental Figure 10. The temporal dynamics of amino acid changes in the ORF1ab gene**
810 **associated with changes in viral copies.** The results are based on the multivariate regression
811 analysis using the two MDS components as covariates. (A) The phylogenetic tree estimated from a
812 representative set of 996 genome sequences showing variant assignments and the locations of amino
813 acid changes that increase (blue) or decrease (red) viral copies. (B) The temporal dynamics of the SNPs
814 from February 2021 to March 2023. The transparency of the color corresponds to the mutation fraction
815 in the daily sequence count: transparent color indicates low fractions, and opaque color indicates high
816 fractions.
817

818 **Supplemental Figure 11. The temporal dynamics of amino acid changes in the ORF3a gene**
819 **(S26L and T223I), M gene (D3G and I82T), ORF7b gene (T40I) and N gene (D63G and S413R)**
820 **associated with changes in viral copies.** The results are based on the multivariate regression
821 analysis using the two MDS components as covariates. (A) The phylogenetic tree estimated from a
822 representative set of 996 genome sequences showing variant assignments and the locations of amino
823 acid changes that increase (blue) or decrease (red) viral copies. (B) The temporal dynamics of the SNPs
824 from February 2021 to March 2023. The transparency of the color corresponds to the mutation fraction
825 in the daily sequence count: transparent color indicates low fractions, and opaque color indicates high
826 fractions.
827

828 **Supplemental Figure 12. Comparison of key variant-defining amino acid changes with GWAS-**
829 **identified substitutions.** The comparison of the key amino acid changes (dark purple) in each variant,
830 with GWAS-identified SNPs that were associated with negative (red) or positive (blue) effects on viral
831 copies in the (A) S gene and (B) ORF1ab gene. The results of GWAS analysis using the two dimensions
832 computed by MDS as covariates are shown as “GWAS 1”, and the results of the analysis using the
833 categorical clusters as covariates are shown as “GWAS 2”. The effective sizes of identified SNPs using
834 different population control methods are given in Tables S1 and S2.
835

836 **Supplemental Table 1. The identified amino acid changes associated estimated effective sizes**
837 **and standard deviations using the multivariate linear regression model with categorical clusters**
838 **as covariates.**
839

840
841 **Supplemental Table 2. The identified amino acid changes associated estimated effective sizes**
842 **and standard deviations using the multivariate linear regression model with MDS-computed**
843 **dimensions as covariates.**
844