

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}, Yi Ting Chew^{1,2}, Nicholas F.G. Chen¹, David Ferguson^{3,4}, Sameer
5 Pandya^{3,4}, Nick Kerantzas^{3,4}, Wade Schulz^{3,4}, Yale SARS-CoV-2 Genomic Surveillance Initiative*,
6 Anne M. Hahn¹, Virginia E. Pitzer^{1,2}, Daniel M. Weinberger^{1,2}, Nathan D. Grubaugh^{1,2,5,#}

7
8 ¹ Department of Epidemiology of Microbial Diseases, Yale School of Public Health, New Haven, CT, USA

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

10 ³ Department of Laboratory Medicine, Yale School of Medicine, New Haven, CT, USA

11 ⁴ Yale Center for Genome Analysis, Yale University, New Haven, CT, USA

12 ⁵ Department of Ecology and Evolutionary Biology, Yale University, New Haven, CT, USA

13 * Authors listed at the end of the manuscript

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

15 Abstract

16 Variations in viral loads generated during SARS-CoV-2 infections can influence COVID-19
17 disease progression and the likelihood of onward transmission. However, the host and virus
18 factors that may impact viral loads and infection dynamics are not fully understood. Here, we
19 conducted virus whole genome sequencing and measured viral copies using RT-qPCR from
20 9,902 SARS-CoV-2 infections over a 2-year period to examine the relative impact of host
21 factors and virus genetic variation on changes in viral copies. We first used statistical
22 regression models to show that host age and SARS-CoV-2 variant significantly impact viral
23 copies, but vaccination status does not. Then, using a genome-wide association study
24 (GWAS) approach, we identified multiple nucleotide substitutions corresponding to amino
25 acid changes in the SARS-CoV-2 genome associated with variations in viral copies. In
26 particular, we analyzed the temporal patterns and found that SNPs associated with higher
27 viral copies were predominantly observed in Omicron BA.2/BA.4/BA.5/XBB infections,
28 whereas those associated with decreased viral copies were mostly observed in infections
29 with Delta and Omicron BA.1 variants. Our work showcases how GWAS can be a useful tool
30 for probing phenotypes related to SNPs in viral genomes, which can be used to characterize
31 emerging variants and monitor therapeutic interventions.

33 Introduction

34 Continued SARS-CoV-2 transmission and evolution has propelled the COVID-19 pandemic.
35 Peak viral replication in the upper respiratory tract occurs during the first few days of infection
36 [1]. The viral load (or copies measured by RT-qPCR) in patient samples can be used to infer
37 the likelihood of transmission, and a strong relationship between SARS-CoV-2 transmission
38 and viral load has been reported [2]. However, it is challenging to predict the transmission
39 potential of SARS-CoV-2 infections using individual viral load alone, which limits our
40 understanding of the intrinsic transmissibility of the virus at the population level [3–5]. The
41 challenge often arises from large variations in viral load dynamic in sampled cases, which can
42 be associated with 1) host heterogeneity, e.g., age [3] and vaccination status [6–8]; 2) distinct
43 inherent properties of virus variants or sublineages [9], and 3) different sampling times [10].

44 For example, sampling during the early stages of infection may yield higher viral loads
45 compared to later stages when viral replication has reached its peak. Nevertheless, the

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

46 relative importance of these factors influencing viral load has not been completely explored
47 [11,12].

48

49 Genome-wide association studies (GWAS) have emerged as a useful tool in the field of
50 genetics, providing an approach to unraveling the complex interplay between genetic
51 variations and observable traits, including diseases and drug resistance, as reviewed in [13].
52 Through testing hundreds of thousands of genetic variants across many genomes, GWAS
53 analysis enables us to identify specific mutations or genes in humans that contribute
54 significantly to observed phenotypic variations, such as HIV1-load [14]. Several studies have
55 employed GWAS analysis to identify and investigate the association between human genetic
56 variations across different individuals and the severity of COVID-19, shedding light on genetic
57 variations that are related to severe infections [15–17]. Although GWAS was originally
58 developed for humans, recently, it's been increasingly adapted and applied to study the
59 genetic basis of several microbial traits [18], especially for bacteria. Although there are studies
60 utilizing a GWAS method to validate the association between the viral genome and HIV drug
61 resistance [19], and to study the effect of human immunity on hepatitis C virus [20], the
62 application of GWAS analysis to study viral genomes and examine the association between
63 viral mutations and related phenotypes is limited. The confluence of the extensive existing
64 research on SARS-CoV-2 mutations and the millions of infections that have been sequenced
65 provides us the opportunity to evaluate the application of GWAS for viral genomics. The
66 hypothesis-free approach has the potential to enhance our understanding of genetic
67 determinants influencing viral fitness and evolution and further inform effective public health
68 strategies aimed at mitigating the spread and impact of SARS-CoV-2.

69

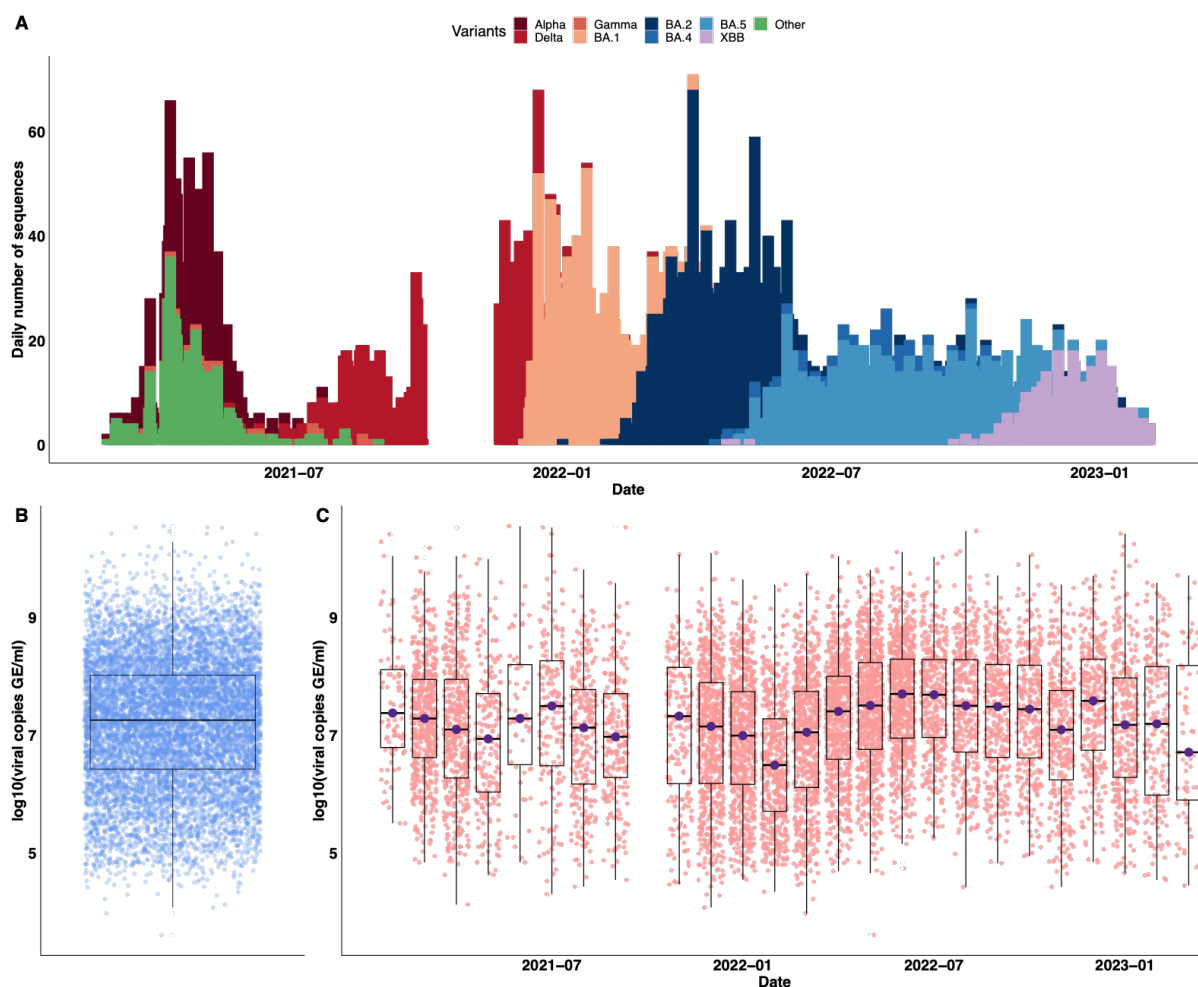
70 In this work, we aim to investigate the relative impact of host factors and intrinsic viral genetic
71 mutations (single nucleotide polymorphisms, SNPs) on the changes in viral copies. For this,
72 we apply a viral GWAS analysis to SARS-CoV-2 genomic sequencing and standardized RT-
73 qPCR data collected from the Yale New Haven Hospital from February 2021 to March 2023.
74 Further, the availability of the whole genome sequencing data on SARS-CoV-2 infections with
75 the relevant laboratory and patient metadata allows the identification of the association
76 between viral mutations and viral copies for different variants of concern (VOCs). We then
77 examine the temporal pattern of identified mutations by constructing a phylogenetic tree,
78 drawing upon subsamples, and analyzing the time series of the fraction of mutations
79 occurring in the sequences. Finally, we compare whether the temporal pattern is consistent
80 across different gene segments. This multifaceted analysis contributes to unraveling the
81 complex dynamics of SARS-CoV-2 infections, providing valuable insights into the underlying
82 viral mutations that influence viral copies in different VOCs.

83 Results

84 Viral copies vary in SARS-CoV-2 infections

85 To better understand how SARS-CoV-2 viral load varies in infected individuals, we analyzed
86 the viral copy data, along with associated host metadata (i.e., age and vaccination status),
87 and genome sequencing data from a cohort of patients tested at the Yale New Haven
88 Hospital (YNHH) located in Connecticut, US. We selected 9902 whole genome sequences
89 with available viral copy data generated from remnant SARS-CoV-2 diagnostic samples over
90 a 2-year period, from 03-Feb-2021 to 21-Mar-2023 (**Fig. 1A**). The VOCs that we identified in

91 our dataset during the sampling period included Alpha ($n = 809$), Delta ($n = 1278$), Gamma ($n = 36$), BA.1 ($n = 1818$), BA.2 ($n = 2432$), BA.4 ($n = 293$), BA.5 ($n = 1992$), XBB ($n = 698$), and
92 the pre-VOC variant (named 'Other', $n = 546$). We conducted RT-qPCR using a standardized
93 assay targeting the nucleocapsid (CDC 'N1' primers) for each sample to allow for cross-
94 sample comparisons [21], except for a period during October 2021 when the PCR data were
95 not generated. Across all samples, the viral copies, expressed as \log_{10} (viral copies per
96 milliliter (Genome Equivalents/ml)), exhibited variations, ranging from 3.60 to 10.55, with a
97 median value of 7.26 (**Fig. 1B**). The variations in viral copies could be attributed either to the
98 introduction and/or replacement of different VOCs, each with its own epidemic curve, or to
99 the stochasticity from the sampling process. To reduce stochastic effects, we aggregated
100 the viral copies by month and still observed large variations in the viral copies across the
101 months (**Fig. 1C**). Notably, we observed the lowest median value of viral copies (median =
102 6.49) in February 2022, during which 96.3% of the sampled sequences tested positive for
103 BA.1 infections. By contrast, we observed the highest median value of viral copies (median =
104 7.70) in June 2022, during which the sampled sequences tested positive for BA.2 (64.9%),
105 BA.4 (6%), or BA.5 (29.1%) infections. Taken together, we showed a wide range of viral
106 copies in the sampled SARS-CoV-2 infections with different VOCs, utilizing data from
107 genomic surveillance and standardized RT-qPCR tests.
108
109



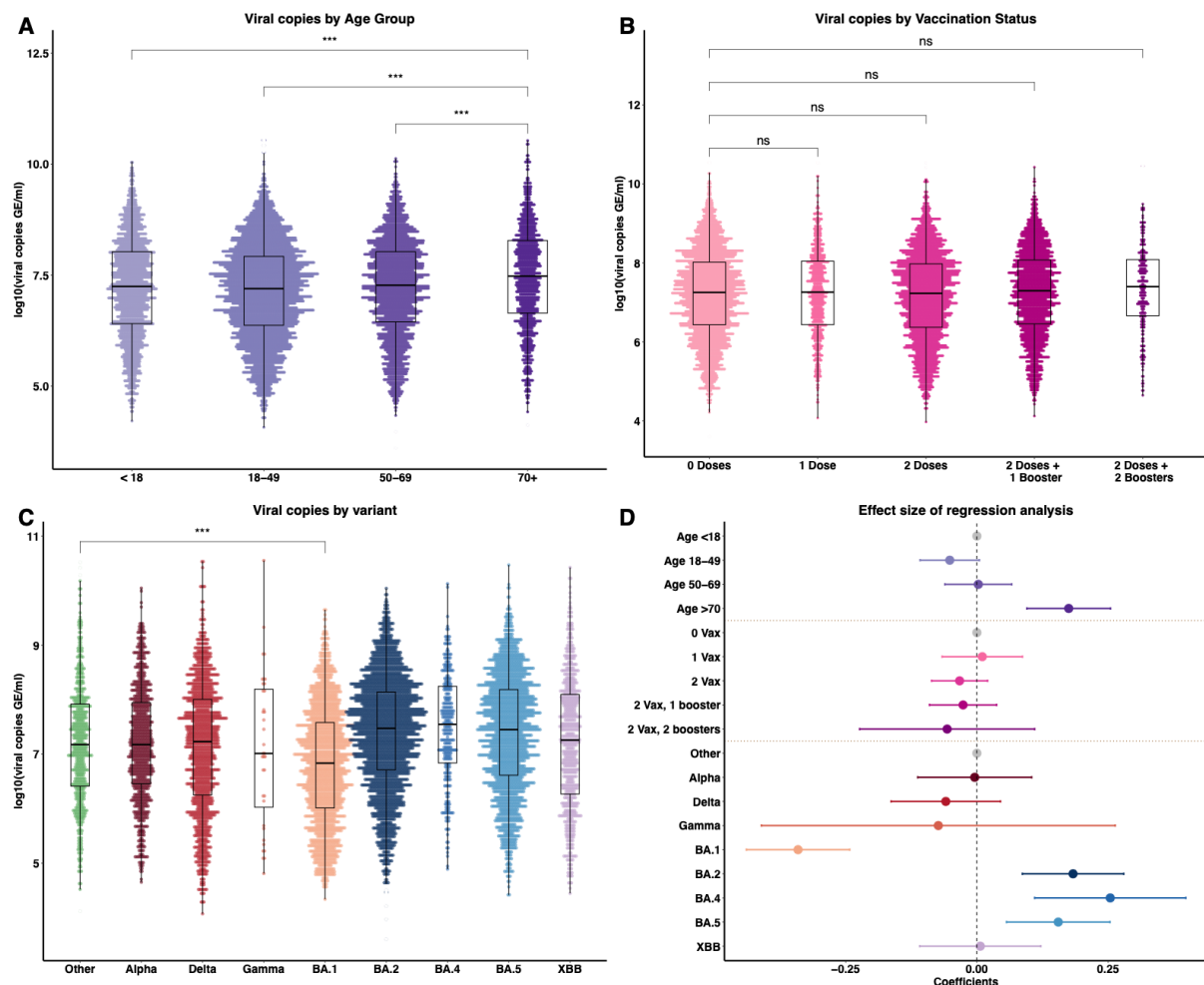
110
111 **Figure 1. Genomic sequences of SARS-CoV-2 infections and associated viral copies from cross-**
112 **sectional samples collected in Connecticut, US. (A)** The daily number of genomic sequences of
113 SARS-CoV-2 VOCs from February 2021 to March 2023. **(B)** The summary of viral copies of all samples,
114 expressed as \log_{10} (viral copies per milliliter). **(C)** The summary of viral copies aggregated by month.

115 The data gap in October 2021 is because we were unable to conduct PCR to obtain viral copies during
116 this time.

117 Viral copies correlate with age and variants, but not with vaccination 118 status

119 Having uncovered a large variability in the observed viral copies from the samples, we next
120 assessed whether host factors, such as age and vaccination status, or inherent differences
121 among VOCs contribute to changes in viral copies. To do this, we first summarized and
122 compared viral copies in various age groups (**Fig. 2A**). A positive correlation has been
123 previously reported between age and SARS-CoV-2 viral copies, showing that younger age
124 groups had lower viral copies independent of gender and/or symptom duration [22]. We
125 observed a similar result in our dataset and found that the older age group (i.e., >70 years
126 old) had the highest viral copies compared with other age groups (mean = 7.47, 95%
127 confidence interval (CI): [5.12, 9.49], $p < 0.001$, Wilcoxon signed-rank test). For the effect of
128 vaccination on viral copies, some studies have demonstrated that although vaccination
129 reduced the risk of infections with the Delta variant, no significant difference in peak viral
130 copies was found between fully vaccinated and unvaccinated individuals [6,7,23]. In contrast,
131 other studies have shown that vaccination reduced viral copies in BA.1 infections among
132 boosted individuals compared to unvaccinated ones [8]. These results suggest the effect of
133 vaccination on viral copies may depend on the characteristics of the infecting SARS-CoV-2
134 variant. We compared viral copies among groups with different vaccination statuses to
135 assess the impact of vaccination on viral copies (**Fig. 2B**), and no statistically significant
136 differences were detected between the groups in our data ($p > 0.05$, Wilcoxon signed-rank
137 test). Finally, we compared viral copies stratified by variant category (**Fig. 2C**). Combining
138 samples collected from all age and vaccination status groups for each variant, we found that
139 the overall mean values of viral copies were lowest for infections with BA.1 (mean = 6.83,
140 95% CI: [4.87, 8.87], $p < 0.001$, Wilcoxon signed-rank test) compared to infections with other
141 variants.

142
143 Since several factors may simultaneously impact the SARS-CoV-2 viral load, next, we sought
144 to quantify the combined impact of age, vaccination status, and VOCs on the observed viral
145 copies. To achieve this, we fitted a multivariate linear regression model, with viral copies as
146 the outcome variable and age, vaccination, and VOCs as covariates (**Fig. 2D**). We found that
147 the older age group (i.e., age >70 years old) had a positive association with viral copies (mean
148 = 0.17, 95% CI: [0.09, 0.25], $p < 0.001$) compared with the reference group (i.e., age <18
149 years old). We also found that vaccination status was not associated with viral copies (i.e.,
150 95% CIs of the vaccination coefficients span 0, $p > 0.05$). Notably, we showed that infections
151 with BA.1 were associated with reduced viral copies, with a mean effect size of -0.34 (95%
152 CI: [-0.44, -0.24], $p < 0.001$) in the same age group and vaccination status, compared to the
153 Other variant. We also showed that infections with BA.2 (mean = 0.19, 95% CI: [0.09, 0.28],
154 $p < 0.001$), BA.4, or BA.5 (mean = 0.17, 95% CI: [0.07, 0.26], $p < 0.001$) were associated with
155 increased viral copies. Among them, infections with BA.4 were associated with the largest
156 positive effect size (mean = 0.27, 95% CI: [0.12, 0.41], $p < 0.001$). Our findings demonstrated
157 that variations in viral copies were associated with infections caused by different SARS-CoV-
158 2 variants and the older age group. This implies that intrinsic factors of the viruses, such as
159 genetic mutations among distinct VOCs, are key determinants impacting viral copies.



160
 161 **Figure 2. Viral copies by category and regression analysis results.** Comparison of viral copies
 162 stratified by (A) age groups, (B) vaccination statuses, (C) variant of concerns. (D) Association of age,
 163 vaccination status, and VOCs with viral copies, expressed as \log_{10} (viral copies per milliliter (Genome
 164 Equivalents/ml)). The reference groups (in gray) are Age <18 years old, 0 doses of vaccination, and the
 165 Other variant, respectively. The positive coefficients indicate the covariate is associated with higher
 166 viral copies value compared to the reference group, and vice versa. 0 Vax, 1 Vax, 2 Vax, 2 Vax 1
 167 booster, and 2 vax 2 boosters denote vaccination statuses of 0 doses, 1 dose, 2 doses, 2 doses,
 168 and 1 booster, and 2 doses and 2 boosters, respectively, corresponding to the labels in (B). Results are
 169 shown as means with 95% confidence intervals. *** $p < 0.001$.

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

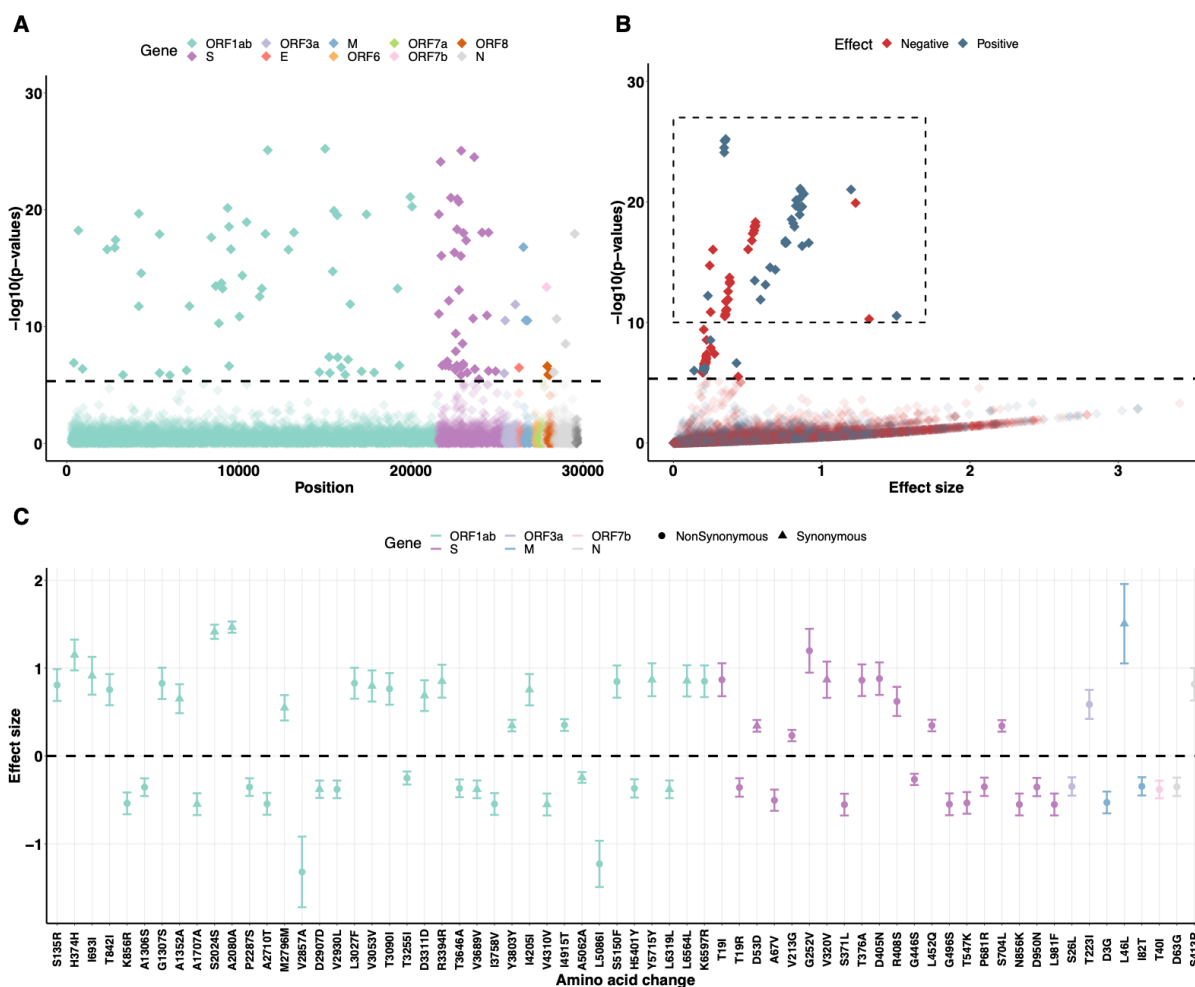
171 Having demonstrated that changes in SARS-CoV-2 viral copies are associated with infections
 172 caused by different viral variants or strains, especially Omicron BA.1/BA.2/BA.4/BA.5 variants
 173 (Fig. 2D), we then sought to identify potential genetic mutations—specifically, SNPs—that
 174 contributed to these changes in viral copies. For this, we performed a GWAS analysis using
 175 high-quality genome sequences (i.e., genome coverage > 95%). We conducted whole-
 176 genome sequencing on the 9902 SARS-CoV-2 positive specimens collected from February
 177 2021 to March 2023. Firstly, using Wuhan-Hu-1 (GenBank MN908937.3) as the reference
 178 genome, we identified 10,697 SNPs for further testing associated with viral copies as
 179 covariates. We then checked for the population structure of the 9902 genome sequences
 180 using a multidimensional scaling (MDS) method [24] (Fig. S1). We observed that Delta was
 181 an outgroup to other pre-Omicron variants (i.e., pre-VOC variant (Other), Alpha, and Gamma),
 182 and BA.1 was an outgroup to the BA.2/BA.4/BA.5/XBB cluster. We included the inferred MDS

183 components from the pairwise SNP distance matrix of SARS-CoV-2 sequences in our model
184 to capture the viral population structure. We also calculated the proportion of variance in our
185 original sequence data accounted for by the MDS results. We found that 95% of the SARS-
186 CoV-2 genetic variance could be explained by the two dimensions computed by MDS, which
187 suggested that including these two dimensions in the regression analysis could adequately
188 adjust for the population structure of the viral strains. The host factors, including age group
189 and vaccination status, were also included in our model as covariates.

190
191 Using the multivariate linear regression model on viral copies for each SNP, we identified 113
192 SNPs exceeding the permuted threshold for genome-wide significance ($p = 4.67 \times 10^{-6}$,
193 dashed line, **Fig. 3A**). The threshold value was calculated as 0.05 divided by 10,697 SNPs
194 [25]. We found that the observed distribution of p -values closely matches the expected
195 distribution under the null hypothesis of no association (**Fig. S2A**). To ascertain whether those
196 SNPs have a negative or positive impact on viral copies and evaluate their effect size, we
197 extracted the coefficients (β) of the SNPs with $p < 1 \times 10^{-10}$ and their standard errors (σ)
198 from the regression model (dashed box, **Fig. 3B**). We then annotated the SNPs to identify the
199 associated amino acids, and among them, 44 SNPs were non-synonymous (i.e., changed the
200 amino acid; **Fig. 3C**). We found that a non-synonymous change G252V, located on the S
201 gene, had the most significant association with increased viral copies ($p = 5.60 \times 10^{-22}$, $\beta =$
202 1.21 , $\sigma = 0.12$). By contrast, the amino acid change most strongly associated with a negative
203 effect on viral copies was ORF1ab:V2857A ($p = 5.66 \times 10^{-11}$, $\beta = -1.32$, $\sigma = 0.20$).

204
205 To assess the impact of adjusting for the population structure of the SARS-CoV-2 strains
206 using the MDS components on the regression results, we conducted a sensitivity analysis on
207 the genome sequences using the defined sequence clusters (a categorical variable) as
208 covariates. Clusters were defined using a k -means clustering method (**Fig. S1**). By doing this,
209 we identified only 31 SNPs exceeding the permuted threshold (**Fig. S3**), compared to 113
210 SNPs found in the analysis using MDS-computed dimensions as population control. The
211 observed distribution of p -values also closely matched the expected distribution under the
212 null hypothesis of no association (**Fig. S2B**). Of these, 25 out of the 31 identified amino acid
213 changes were consistent with those previously observed, except for ORF1ab:P3395H,
214 S:Q498R, N679K, Q954H, N969K, and N:R203M. The results may be more likely to reflect
215 the SNPs that influence the viral copies independent of lineage.

216
217 We also examined the association between viral copies and SNPs after adjusting for the
218 population structure based on the VOCs themselves, which broadly correspond to the
219 identified sequence clusters. To do this, we conducted sensitivity analyses on the genome
220 sequences within each sequence cluster defined by a k -means clustering method (**Fig. S1**).
221 We showed that only a few SNPs were found (**Figs. S4-7**), mostly within the Omicron
222 BA.2/BA.4/BA.5/XBB cluster (**Fig. S7**).



223
 224 **Figure 3. GWAS analysis identifies several single nucleotide polymorphisms (SNPs) that are**
 225 **associated with the changes in viral copies. (A)** Genome-wide association results of the impact of
 226 identified SNPs on viral copies during SARS-CoV-2 infection. The dashed line indicates the permuted
 227 threshold for genome-wide significance $p = 4.67 \times 10^{-6}$. Significant SNPs are shown with solid colors.
 228 **(B)** SNPs (with $p < 1 \times 10^{-10}$) that have positive (blue) or negative (red) effects on viral copies. **(C)** The
 229 corresponding synonymous (triangles) and non-synonymous (circles) amino acid changes that
 230 associate with increased or decreased viral copies. Data is shown as means with 95% confidence
 231 intervals. The estimated effective sizes and associated standard deviations are given in **Table S1**. A
 232 Q-Q plot showing the observed distribution of p -value and the expected distribution is given in **Fig.**
 233 **S2**.
 234

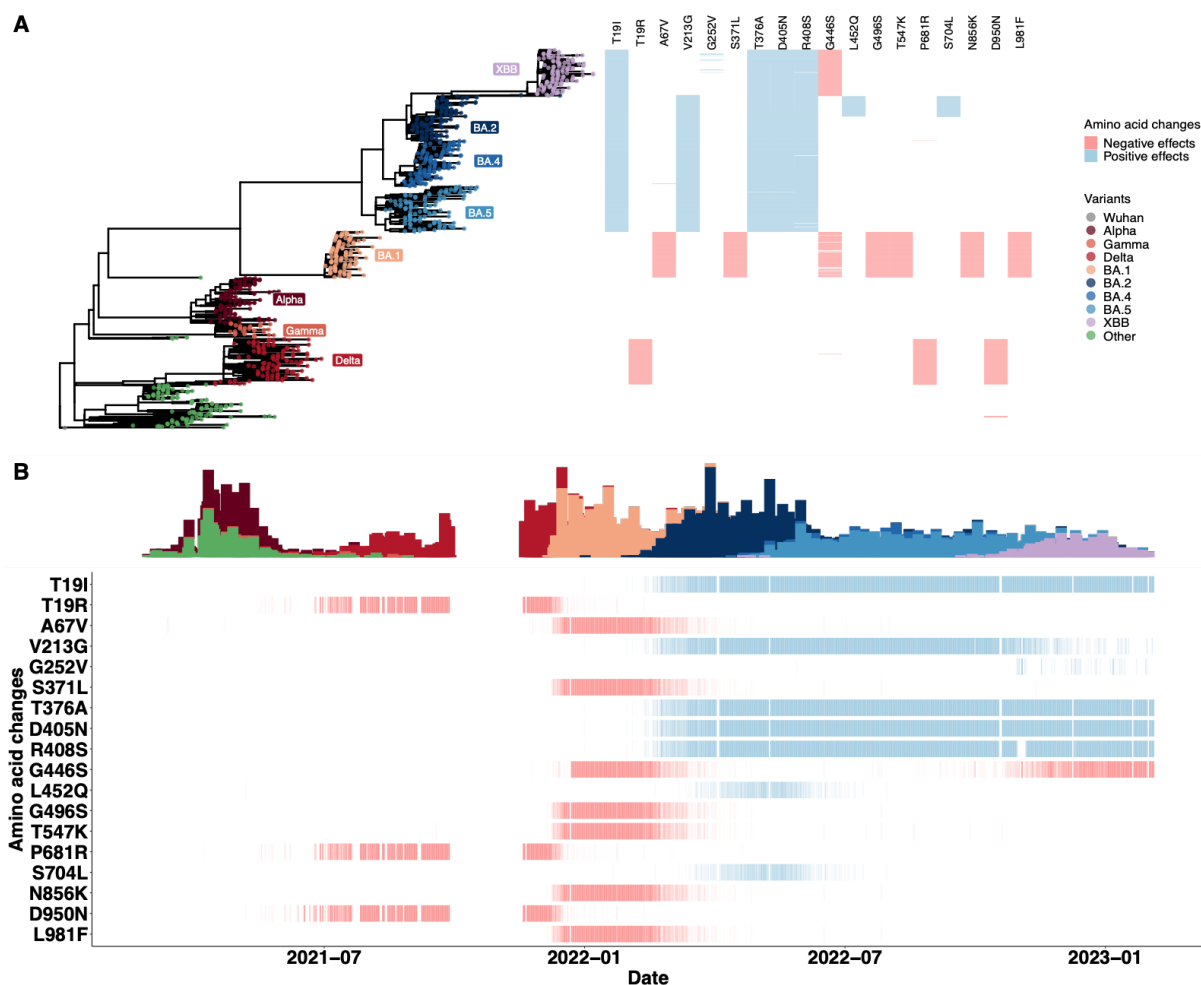
235 The impact of amino acid changes in the S gene on viral copies is
 236 dependent on the variant

237 Having identified the 44 non-synonymous SNPs with statistically significant effects on viral
 238 copies in our primary analysis, we next sought to understand the temporal patterns of the
 239 emergence of these amino acid changes (**Fig. 4**). To investigate the clustering of these SNPs,
 240 we randomly sampled approximately 120 genome sequences from each VOC category (only
 241 36 sequences were available for Gamma in our dataset) and generated a phylogenetic tree
 242 drawing upon the subsamples (**Fig. 4A**). We first focused on the 18 amino acid changes in
 243 the S gene and found a clear pattern in how these mutations emerged by VOC (**Fig. 4A**
 244 heatmap). Notably, we found that all amino acid changes associated with a positive effect on
 245 viral copies were found in BA.2/BA.4/BA.5/XBB infections. Often, more than one amino acid

246 change was observed in each sampled sequence, suggesting genetic linkage between these
247 SNPs. In particular, we identified that the amino acid changes L452Q ($p = 3.91 \times 10^{-25}$, $\beta =$
248 0.34 , $\sigma = 0.03$) and S704L ($p = 1.35 \times 10^{-24}$, $\beta = 0.34$, $\sigma = 0.03$) associated with a positive
249 effect on viral copies were typically observed in combination with BA.2 infections—
250 specifically, lineage BA.2.12.1. By contrast, a retrospective cohort study has shown that the
251 risk of death due to COVID-19 was lower for Omicron BA.1 compared with Delta (B.1.617.2)
252 infections [26]. Here, we observed that the amino acid changes with negative effects on viral
253 copies were mostly associated with BA.1 and Delta infections. Among these SNPs, we
254 observed that A67V, S371P, G496S, T547K, N856K, and L981F, exclusively associated with
255 BA.1 infections, had a comparatively larger (negative) impact on viral copies (mean = -0.54,
256 95% CI: [-0.55, -0.51]) than the others (T19R, G446S, P681R and D950N) with a mean of -
257 0.33 and a 95% CI: [-0.36, -0.27], which were also present in Delta or XBB infections.

258
259 To explore the temporal dynamics of these S gene amino acid changes, we calculated the
260 fraction of mutations occurring in the sequences for each day, thereby accounting for the
261 number of introductions to the population (**Fig. 4B**). We observed SNPs (T19I, V213G, T376A,
262 D405N, and R408S) with a positive impact on viral copies emerging in sequences sampled
263 from February 2022, when BA.2 was first detected in Connecticut. These SNPs were
264 consistently observed in almost every sequence thereafter. By contrast, we found that the
265 other two amino acid changes (L452Q and S704L) that had a positive effect on viral copies
266 were only in the samples from BA.2 infections and did not arise again in lineages of BA.4 or
267 BA.5. The G252V amino acid change was associated with higher viral copies; however, we
268 found that the SNP only appeared in a few sequences associated with XBB infections. For
269 those SNPs that had a negative association with viral copies, we observed that most of them
270 (A67V, S371P, G496S, T547K, N856K, and L981F) were present in samples associated with
271 BA.1 infections and did not persist when BA.1 was replaced by BA.2. We also noted that
272 other mutations, e.g., T19R, P681R and D950N, were only found in samples from Delta
273 infections, and that G446S ($p = 5.56 \times 10^{-16}$, $\beta = -0.26$, $\sigma = 0.03$) was first found in samples
274 from BA.1 infections but absent in BA.2/BA.4/BA.5 infections and reemerged in XBB
275 infections.

276

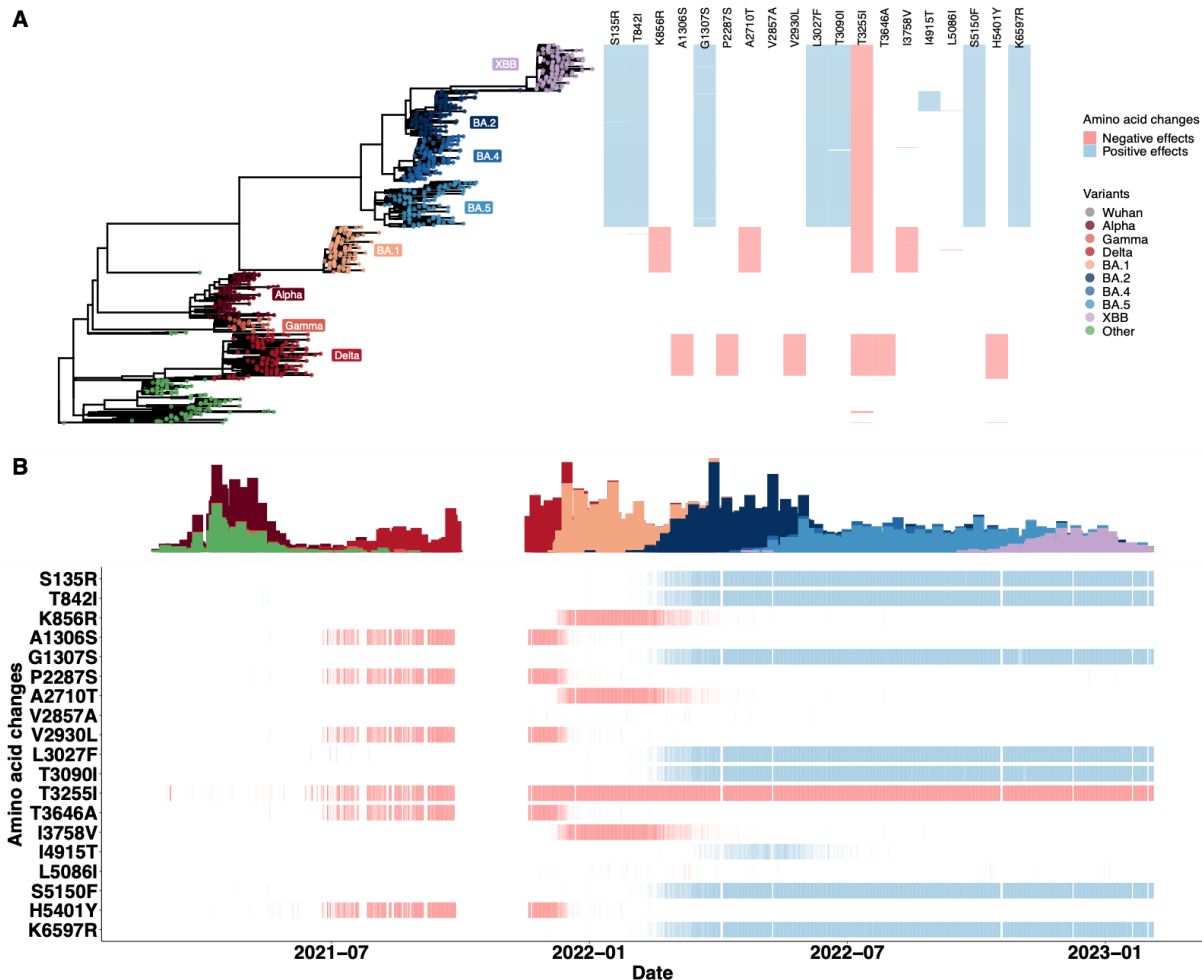


277
 278 **Figure 4. The temporal dynamics of amino acid changes in the S gene associated with changes**
 279 **in viral copies.** The results are based on the multivariate regression analysis using the two MDS
 280 components as covariates. **(A)** The phylogenetic tree estimated from a representative set of 996
 281 genome sequences showing variant assignments and the locations of amino acid changes that
 282 increase (blue) or decrease (red) viral copies. **(B)** The temporal dynamics of the SNPs from February
 283 2021 to March 2023. The transparency of the color corresponds to the mutation fraction in the daily
 284 sequence count: transparent color indicates low fractions, and opaque color indicates high fractions.
 285

286 Amino acid changes in the ORF1ab gene show similar temporal patterns
 287 to those in the S gene

288 Having observed the SNP pattern in the S gene, our next aim was to determine if the same
 289 temporal patterns persisted for the amino acid changes in the ORF1ab gene (**Fig. 5**). We
 290 found that the presence of SNPs associated with an increase in viral copies were observed
 291 in BA.2/BA.4/BA.5/XBB infections, while those associated with a decrease in viral copies
 292 were observed in Delta or BA.1 infections (**Fig. 5A**). This pattern was consistent with our
 293 observations for the S gene. Additionally, we identified similar temporal dynamics of
 294 mutations in the ORF1ab gene, mirroring the pattern seen in the S gene (**Fig. 5B**). These
 295 trends were consistently observed in the temporal dynamics of amino acid changes in the M,
 296 ORF3a, and ORF7b genes as well (**Fig. S8**). In particular, we found that ORF1ab:I4915T ($p =$
 297 2.36×10^{-25} , $\beta = 0.35$, $\sigma = 0.04$) was only associated with BA. 2 infections. We also found
 298 that, although ORF1ab:V2857A had the most significant effect in reducing viral copies ($p =$
 299 5.66×10^{-11} , $\beta = -1.32$, $\sigma = 0.20$), it only appeared in 24 samples (0.24% of the sample size)

300 with BA.1 infections. Furthermore, we identified that ORF1ab:T3255I ($p = 1.02 \times 10^{-11}$, $\beta = -$
 301 0.25, $\sigma = 0.04$) was consistently detected in Delta and other Omicron variants and was
 302 associated with decreased viral copies.
 303
 304



305
 306 **Figure 5. The temporal dynamics of amino acid changes in the ORF1ab gene associated with**
 307 **changes in viral copies.** The results are based on the multivariate regression analysis using the two
 308 MDS components as covariates. **(A)** The phylogenetic tree estimated from a representative set of 996
 309 genome sequences showing variant assignments and the locations of amino acid changes that
 310 increase (blue) or decrease (red) viral copies. **(B)** The temporal dynamics of the SNPs from February
 311 2021 to March 2023. The transparency of the color corresponds to the mutation fraction in the daily
 312 sequence count: transparent color indicates low fractions, and opaque color indicates high fractions.

313 GWAS analysis identified a subset of variant-defining amino acid
 314 substitutions

315 Experimental investigations on the specific amino acid substitutions that define emerging
 316 VOCs for SARS-CoV-2 have yielded insights into viral transmissibility and immune escape
 317 [27–33]. Here, we compared the identified SNPs using our GWAS method with all variant-
 318 defining amino acid changes (**Fig. 6**). The variant-defining amino acid changes were defined
 319 as those mutations with >75% prevalence in at least one lineage, as estimated on the
 320 outbreak.info website [33]. Our GWAS analysis identified a subset of the variant-defining
 321 amino acid substitutions in the S (**Fig. 6A**) and ORF1ab genes (**Fig. 6B**) that were associated
 322 with the changes in viral copies. More importantly, we were able to identify S L452Q and

323 S704L, which primarily emerged within the sublineage BA.2.12.1. These were not identified
 324 as variant-defining mutations on the [outbreak.info](https://www.outbreak.info) website. The results imply the applicability
 325 of GWAS of SARS-CoV-2 genome sequences to identify the associations between viral
 326 copies and amino acid changes without exclusively focusing on variant-defining SNPs.
 327



328
 329 **Figure 6. Comparison of key variant-defining amino acid changes with GWAS-identified**
 330 **substitutions.** The comparison of the key amino acid changes (dark purple) in each variant, with
 331 GWAS-identified SNPs that were associated with negative (red) or positive (blue) effects on viral copies
 332 in the (A) S gene and (B) ORF1ab gene. The results of GWAS analysis using the two dimensions
 333 computed by MDS as covariates are shown as “GWAS 1”, and the results of GWAS analysis using the
 334 categorical clusters as covariates are shown as “GWAS 2”. The effective sizes of identified SNPs using
 335 different population control methods are given in **Tables S1** and **S2**, respectively.

336 Discussion

337 We conducted a GWAS analysis on 9,902 high-quality SARS-CoV-2 genome sequences
 338 generated from two years of genomic surveillance in Connecticut, US to identify and evaluate
 339 SNPs that were associated with variations in viral copies during infections. Both viral factors
 340 and host characteristics can have impacts on viral replication at the within-host level [3,6–9].
 341 Using a GWAS approach, we were able to identify and examine virus-related factors that

342 were associated with the observed variations in viral copies. This was achieved by combining
343 data from a large cohort of individuals infected with different VOCs and employing a
344 regression model that accounted for both individual (i.e., age and vaccination status) and
345 virus-level factors (i.e., specific SNPs and genetic background) in viral copies. We identified
346 several SNPs corresponding to non-synonymous amino acid changes in the SARS-CoV-2
347 genome that were individually associated with the variations in viral copies. In particular,
348 temporal patterns of the SNPs revealed that SNPs associated with increased viral copies
349 were predominantly observed in Omicron BA.2/BA.4/BA.5/XBB infections, whereas those
350 associated with decreased viral copies were mostly observed in infections with Delta and
351 Omicron BA.1 variants.

352
353 Using GWAS analysis, we identified a subset of variant-defining amino acid changes in Delta
354 and Omicron variants. Note that we did not detect any substitutions in the Alpha and Gamma
355 variants (likely due to the low sample size for Gamma). We also identified SNPs that did not
356 define any major variant category, including S:L452Q and S704L that were specifically
357 associated with BA.2.12.1, a sublineage of BA.2 that briefly dominated during the pandemic
358 (i.e., dominated mainly in the US between March and May 2022). This highlights the
359 application of GWAS for identifying SNPs associated with important phenotypic effects
360 without requiring a set of lineage-defining mutations to be defined a priori. Nevertheless,
361 there are several reasons why we only detected a subset of the SNPs that defined different
362 VOCs. Firstly, SNPs with small effect sizes may not be detected due to the stringent statistical
363 significance thresholds applied in GWAS. Secondly, lineage-defining SNPs that are in low
364 linkage disequilibrium with the causal mutations may not be detected [34], even if they may
365 be functionally relevant. Our results showcase how GWAS can help to narrow the focus of
366 SNPs associated with specific phenotypes, generating hypotheses for further investigation.

367
368 A key result from our analysis is that SNPs associated with viral copies did not exhibit the
369 same temporal dynamics, even though they could have similar (either positive or negative)
370 effects on viral copies. Some amino acid changes, for example, ORF1ab:I4915T (positive
371 effects), were only present in samples with BA.2 infections and disappeared when new
372 Omicron variants emerged. Other SNPs (e.g., S:T19R and ORF1ab:S135R), while also
373 associated with higher viral copies, were observed and persisted in all BA.2/BA.4/BA.5/XBB
374 infections. The distinct temporal pattern of SNPs, dependent on VOCs, may help explain the
375 different fitness levels (e.g., intrinsic transmissibility or immune escape) of each variant
376 [35,36]. For example, S:L452Q has been suggested to increase viral infectivity and affinity to
377 angiotensin-converting enzyme 2, thereby enhancing the ability to circumvent vaccine-
378 induced immune responses [37]. Different VOCs have been suggested to exhibit variations in
379 cell tropisms, with BA.1 showing a preference for the upper respiratory tract [1,38]. This
380 preference could be associated with specific mutations which may have causal effects on
381 intracellular viral replication dynamics in different host cell types. Recent studies indicated
382 that BA.2 infections were characterized by a higher peak viral load, as shown in [39–41]. This
383 is consistent with our analysis, where we found that SNPs associated with an increase in viral
384 copies were linked to BA.2 infections. The shift from negative to positive effect SNPs on viral
385 copies from BA.1 to BA.2 may explain the observed rapid replacement of BA.1 by BA.2 and
386 the subsequent fast spread of BA.2 infections [42].

387
388 In addition, we showed that S:G446S was associated with decreased viral copies in primarily
389 BA.1 and XBB infections. Motozono et al. [43] have previously demonstrated that the
390 mutation could affect antigen presentation and potentiate antiviral activity by vaccine-
391 induced T cells, leading to enhanced T cell recognition. We also identified a previously

392 studied amino acid change, T3255I, in the ORF1ab gene. This SNP is located at site 492 in
393 non-structural protein 4 (NSP4) and carried by both the Delta and Omicron variants. The
394 mutation has been implied to increase the replication capacity of the virus and improve its
395 ability to evade host immune responses [44]. Using phylogenomic analysis, we identified that
396 the mutation was present in both the Delta and Omicron variants, with a negative impact on
397 viral copies. These two mutations, S:G446S and ORF1ab:T3255I were found in different
398 genetic backgrounds, which may provide evidence that they may influence the viral load
399 independent of the background/lineage.

400
401 Several studies indicate that the Delta variant is linked to higher viral loads and increased
402 transmissibility compared to the Alpha variant [45–47]. In contrast, another study found no
403 significant difference in the mean peak viral load (indicated by a cycle threshold value)
404 between Delta infections and the pre-VOC variant [7,48]. Here, we identified mostly SNPs
405 associated with decreased viral copies in Delta infections, except for one N:R203M (**Fig. S9**).
406 The discrepancy of these results suggests that the relationship between Delta variant
407 infections, viral copies, and associated factors is multifaceted, requiring careful consideration
408 of several factors and large datasets to investigate such patterns. Differences in population
409 demographics, the dynamic nature of viral evolution, or study methodologies (e.g., diagnostic
410 assays) could all influence the observed viral copies. Previous studies inferring viral load
411 differences in infections by VOCs normally analyzed viral copy data from large cohorts of
412 individuals without explicitly accounting for the influence of host factors, such as age, host
413 immunity, symptomatic conditions, and infection history. These factors could potentially
414 contribute to variations in viral load. Not only does host immunity and variants shape within-
415 host viral dynamics of SARS-CoV-2 infections, but individual heterogeneity also plays a more
416 important role in determining the viral kinetics [49]. For instance, an increased viral load was
417 observed primarily in the younger population with Delta infections [46]. The association of
418 higher viral loads with a younger age group raises intriguing questions about age-related
419 susceptibility and immune responses, both for specific variants and for SARS-CoV-2 in
420 general, as discussed in [50]. Alternatively, inconsistent results in various studies may arise
421 from using data on different sub-lineages of SARS-CoV-2, for example those belonging to
422 the Delta variant (e.g., B.1.617.2 and AY). Another possibility for not detecting any positive-
423 effect SNPs in Delta infections is that the increased viral copies in Delta infections may result
424 from the concurrent accumulation of multiple mutations, which may introduce non-additive
425 epistatic effects that can be difficult to unpick [28]. In this study, however, we employed a
426 series of single SNP regression models to identify the underlying SNPs associated with the
427 changes in viral copies. This was achieved by performing individual regression analyses for
428 each of the 10,697 SNPs one at a time, without accounting for potential interactions between
429 SNPs. A multiple SNP regression study, which allows for testing potential interactions or joint
430 effects among multiple SNPs, albeit challenging due to multicollinearity, will be left for future
431 study.

432
433 There are limitations to our study. First, the common cooccurrences of identified SNPs make
434 it impossible to evaluate their individual effects on viral copies, which is an overall limitation
435 of GWAS. However, our findings provide valuable insights into the overall genetic landscape
436 of the viral population and identify potential genetic variants that could be further explored.
437 Second, we assumed that the distribution of times between infection and sample collection
438 was similar through time and across variants as these data were not available. Given our
439 samples were taken frequently over a 2-year period, we do not anticipate that this assumption
440 will qualitatively impact our results. Third, our study primarily focuses on the genetic variants
441 in VOCs, neglecting other factors such as host immune responses or environmental

442 influences, partially captured by the host-associated covariates, including age and
443 vaccination status in this study, that may also contribute to the changes in viral copies.
444 Further study will be needed to address the impact of these factors on viral copies, for
445 example, genome-to-genome analysis to reveal the impact of host-viral genetic interactions
446 in SARS-CoV-2 infections [20,51]. Fourth, our data were obtained from a specific geographic
447 region, whose population diversity may not necessarily be similar to other settings; therefore,
448 extrapolating these findings to a broader population may require caution. Additionally,
449 focusing solely on consensus genomic changes in the analysis could overlook the genetic
450 diversity within the sample, which may also influence variations in viral load. Despite these
451 constraints, our study highlights the importance of sustained genomic surveillance and the
452 need for comprehensive analyses to understand the nuanced impact of specific genetic
453 variations on viral copies at the within-host level, and its implications for viral transmissibility
454 and immune escape at the population level. Further work and collaborative efforts are
455 essential to elucidate the complex interplay between viral genetics, host factors, and the
456 dynamics of transmission associated with emerging variants. Such studies could inform
457 predictive early warning public health systems regarding the emergence of potentially highly
458 transmissible viral strains based on their constellation of mutations.

459
460 Recently, Duesterwald et al. [12] used genome sequence data and a machine-learning
461 approach to predict cycle threshold (Ct) values of SARS-CoV-2 infections based on the *k*-
462 mers. Similar to our findings, they suggested that S:L452 and P681 were hallmarks of VOCs,
463 implying impacts on the observed Ct values in clinical samples. Although the machine-
464 learning approach may capture broader patterns and interactions within the genome on Ct
465 values, they lack interpretability compared to regression models. For example, regression-
466 based models could offer insights into the direct association between specific genetic
467 variants and viral copies. In addition, regression-based models may perform well even with
468 limited sample sizes [19], provided that the assumptions of the model are met and the
469 predictors are informative, whereas using machine-learning methods with small sample sizes
470 can be challenging.

471
472 With the availability of high-quality whole-genome sequences for SARS-CoV-2, we
473 demonstrated that GWAS analysis of the viral genome can identify SNPs that associate with
474 positive or negative impacts on viral copies in VOCs, revealing important biological insights
475 and enhancing our understanding of within-host viral dynamics. We argue that the application
476 of GWAS analyses to study viral genomes provides a particularly tractable tool to identify
477 potential SNPs of interest for further evaluation using genetic engineering for several reasons.
478 First, the small genome size of viruses and high evolutionary rates make it easier to perform
479 comprehensive genome-wide scans for SNPs and to experimentally test the impacts of SNPs
480 on specific traits. Second, significant phenotypic variations (e.g., viral loads and antibody
481 responses) are often observed in viral infections, despite limited changes in the viral genome.
482 GWAS can help to identify SNPs that correlate with these phenotypic variations, providing
483 insights into the genetic basis of these traits. Third, the increasing accessibility to sequence
484 viral genomes makes it possible to perform GWAS on rich datasets, enabling in-depth
485 analysis of the temporal dynamics of viral evolution. Together, the applicability of GWAS
486 analyses to study viral genomes can provide a new approach for exploring the intricate
487 interplay between genetic mutations and phenotypes, informing strategies for managing and
488 mitigating the impact of emerging viral variants, and contributing to the development of
489 potential therapeutic interventions.

490

491 Materials & Methods

492 Ethics

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

497 Clinical sample collection and measurement of viral copies by RT-qPCR

498 SARS-CoV-2 positive samples (nasal swabs in viral transport media) were collected through
499 the Yale New Haven Hospital (YNHH) System as a part of routine inpatient and outpatient
500 testing and sent to the Yale SARS-CoV-2 Genomic Surveillance Initiative. Using the MagMAX
501 viral/pathogen nucleic acid isolation kit, nucleic acid was extracted from 300µl of each clinical
502 sample and eluted into 75µl of elution buffer. Extracted nucleic acid was then used as
503 template for a “research use only” (RUO) RT-qPCR assay [21] to test for presence of SARS-
504 CoV-2 RNA. Ct values from the nucleocapsid target (CDC-N1 primer-probe set [52]) were
505 used to derive viral copy numbers using a previously determined standard curve for this
506 primer set [53]. A positive RNA control with defined viral copy number (1000/µl) was used to
507 standardize results across individual runs.

508 Whole genome sequencing

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

514

515 Reads were then aligned to the Wuhan-Hu-1 reference genome (GenBank MN908937.3)
516 using BWA-MEM v.0.7.15 [54]. Adaptor sequences were then trimmed, primer sequences
517 masked, and consensus genomes called (simple majority >60% frequency) using iVar
518 v1.3.133 [55] and SAMtools v1.11 [56]. When <20 reads were present at a site an ambiguous
519 “N” was used, with negative controls consisting of ≥99% Ns. The Pangolin lineage assignment
520 tool [57] was used for assigning viral lineages.

521 Clinical metadata

522 We obtained patient metadata and vaccination records from the YNHH system and the
523 Center for Outcomes Research and Evaluation (CORE) and matched these records to
524 sequencing data through unique sample identifiers. Duplicate patient records or those with
525 missing or inconsistent metadata and vaccination date were removed from the GWAS
526 analysis. We also removed patient records with persistent infections.

527 We determined vaccination status at time of infection by comparing the sample collection
528 date to the patient’s vaccination record dates. We categorized vaccine statuses based on
529 the number of vaccine doses received at least 14 days before the collection date. Patient

530 vaccination statuses at the time of infection were categorized as follows: non-vaccine, one-
531 dose vaccine, two-dose vaccine, two-dose vaccine with one booster, or two-dose vaccine
532 with two boosters. We calculated the age of each patient as the difference between the
533 date of birth and the sampling date.

534 Single nucleotide polymorphisms

535 To identify single nucleotide polymorphisms (SNPs), we first aligned the 9902 genome
536 sequences using *nextalign* (v3.2.1) [58] with the reference genome of the Wuhan-Hu-1
537 genome (GenBank accession: MN908937.3). Then, SNPs were identified using *snp-sites*
538 (v2.4.1) [59], with the reference genome of the Wuhan-Hu-1 genome (GenBank accession:
539 MN908937.3). We also normalized the SNPs in the generated VCF file, such that multiallelic
540 SNPs were separated into different rows. Normalizing the SNPs ensured that each SNP was
541 one-hot encoded and analyzed separately. Note that we did not include ambiguous SNPs,
542 deletions and insertions in our GWAS analysis. We used *vcf-annotator* (v0.7) to annotate
543 SNPs to corresponding amino acid changes.

544 Multidimensional scaling and population control

545 To reveal the underlying structure of the 9902 genome sequences. We first used *snp-dists*
546 (v0.7.0) [60] to convert the aligned sequences (a FASTA alignment) to a SNP distance matrix.
547 We then applied a multidimensional scaling (MDS) method [24] to transform the SNP distance
548 matrix into a geometric configuration while preserving the original pairwise relationships. The
549 scaling was conducted using *cmdscale* function in an R package *stats* (v3.6.2). We set the
550 maximal dimensional parameter $k = 2$.

551
552 To measure the goodness of the transformation, we calculated the distance between the
553 original genome sequencing data and compared it with the new distances determined by
554 MDS. This involved arranging the two matrices of distances into two columns and computing
555 the correlation coefficient (i.e., r) between them. Finally, we used r^2 to measure the
556 proportion of variance in the original distance matrix explained by the new computed distance
557 matrix.

558 Testing for associations between viral copies and SNPs

559 In this work, we conducted a series of single SNP regression analyses to test for associations
560 between viral copies and SNPs. The linear regression model is written as follows:

$$561 \quad Y \sim \alpha W + \beta_i SNP_i + \xi_1 d_1 + \xi_2 d_2 + e,$$

562 where Y is a vector of normalized \log_{10} -transformed viral copies, W is a matrix of covariates,
563 including age (a categorical variable with four age groups of “<18”, “18-49”, “50-69”, and
564 “>70” years old), vaccination status (a categorical variable with vaccination statuses of “0
565 doses”, “1 dose”, “2 doses”, “2 doses and 1 booster”, “2 doses and 2 boosters”) and an
566 intercept, and α is a vector that corresponds to coefficients of the covariates. In particular,
567 SNP_i is a vector of genotype values for all samples at each SNP, i . It is a binary variable: 0
568 represents the SNP is not present in the genome sequence, whereas 1 represents its
569 presence. β_i is the effective size of each identified SNP, i . The vectors d_1, d_2 represent the
570 two dimensions computed by MDS, and ξ_1, ξ_2 are the coefficients of the dimension
571 covariates. The random effect of residual errors is presented by e , which is assumed to follow
572 a normal distribution with a mean of 0 and a standard deviation of σ_e .

573 Phylogenetic tree construction and comparison to variant-defining 574 substitutions

575 We employed *iq-tree* (v2.2.2.6) [61] of a representative set using 996 of our 9902 genome
576 sequences for tree construction, using Wuhan-Hu-1 (GenBank MN908937.3) as the reference
577 genome. The variant-defining amino acid changes were defined as those mutations
578 with >75% prevalence in at least one lineage, as estimated on outbreak.info website [33].
579 Note that we did not include deletions in variant-defining substitutions.

580 Data and code availability

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

589 Acknowledgements

590 We would like to thank Verity Hill, Seth Redmond, Jiye Kwon, Rafael Lopes, Sophie Taylor,
591 and Philip Jack for their helpful conversations and feedback on this work. This project is
592 supported by the Centers for Disease Control and Prevention (CDC) Broad Agency
593 Announcement Contract 75D30122C14697 (NDG). This work does not necessarily represent
594 the views of the CDC.
595

596 Competing Interest Statement

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

600 Yale SARS-CoV-2 Genomic Surveillance Initiative Authors

601 Tara Alpert, Kaya Bilguvar, Kendall Billig, Mallery Breban, Anderson Brito, Christopher
602 Castaldi, Rebecca Earnest, Bony De Kumar, Joseph Fauver, Chaney Kalinich, Tobias Koch,
603 Marie Landry, Shrikant Mane, Isabel Ott, David Peaper, Mary Petrone, Kien Pham, Jessica
604 Rothman, Irina Tikhonova, Chantal Vogels, Anne Watkins
605
606
607
608
609
610
611

612 References

- 613 1. Killingley B, Mann AJ, Kalinova M, Boyers A, Goonawardane N, Zhou J, et al. Safety, tolerability
614 and viral kinetics during SARS-CoV-2 human challenge in young adults. *Nat Med.* 2022;28:
615 1031–1041. doi:10.1038/s41591-022-01780-9
- 616 2. Marks M, Millat-Martinez P, Ouchi D, Roberts CH, Alemany A, Corbacho-Monné M, et al.
617 Transmission of COVID-19 in 282 clusters in Catalonia, Spain: a cohort study. *Lancet Infect Dis.*
618 2021;21: 629–636. doi:10.1016/S1473-3099(20)30985-3
- 619 3. Jones TC, Biele G, Mühlemann B, Veith T, Schneider J, Beheim-Schwarzbach J, et al.
620 Estimating infectiousness throughout SARS-CoV-2 infection course. *Science.* 2021;373.
621 doi:10.1126/science.abi5273
- 622 4. Cevik M, Tate M, Lloyd O, Maraolo AE, Schafers J, Ho A. SARS-CoV-2, SARS-CoV, and MERS-
623 CoV viral load dynamics, duration of viral shedding, and infectiousness: a systematic review and
624 meta-analysis. *Lancet Microbe.* 2021;2: e13–e22. doi:10.1016/S2666-5247(20)30172-5
- 625 5. Singanayagam A, Patel M, Charlett A, Lopez Bernal J, Saliba V, Ellis J, et al. Duration of
626 infectiousness and correlation with RT-PCR cycle threshold values in cases of COVID-19,
627 England, January to May 2020. *Euro Surveill.* 2020;25. doi:10.2807/1560-
628 7917.ES.2020.25.32.2001483
- 629 6. Singanayagam A, Hakki S, Dunning J, Madon KJ, Crone MA, Koycheva A, et al. Community
630 transmission and viral load kinetics of the SARS-CoV-2 delta (B.1.617.2) variant in vaccinated
631 and unvaccinated individuals in the UK: a prospective, longitudinal, cohort study. *Lancet Infect*
632 *Dis.* 2022;22: 183–195. doi:10.1016/S1473-3099(21)00648-4
- 633 7. Kissler SM, Fauver JR, Mack C, Tai CG, Breban MI, Watkins AE, et al. Viral Dynamics of SARS-
634 CoV-2 Variants in Vaccinated and Unvaccinated Persons. *N Engl J Med.* 2021;385: 2489–2491.
635 doi:10.1056/NEJMc2102507
- 636 8. Puhach O, Adea K, Hulo N, Sattoune P, Genecand C, Iten A, et al. Infectious viral load in
637 unvaccinated and vaccinated individuals infected with ancestral, Delta or Omicron SARS-CoV-2.
638 *Nat Med.* 2022;28: 1491–1500. doi:10.1038/s41591-022-01816-0
- 639 9. Boucau J, Marino C, Regan J, Uddin R, Choudhary MC, Flynn JP, et al. Duration of Shedding of
640 Culturable Virus in SARS-CoV-2 Omicron (BA.1) Infection. *N Engl J Med.* 2022;387: 275–277.
641 doi:10.1056/NEJMc2202092
- 642 10. Hay JA, Kennedy-Shaffer L, Kanjilal S, Lennon NJ, Gabriel SB, Lipsitch M, et al. Estimating
643 epidemiologic dynamics from cross-sectional viral load distributions. *Science.* 2021;373.
644 doi:10.1126/science.abh0635
- 645 11. Fryer HR, Golubchik T, Hall M, Fraser C, Hinch R, Ferretti L, et al. Viral burden is associated with
646 age, vaccination, and viral variant in a population-representative study of SARS-CoV-2 that
647 accounts for time-since-infection-related sampling bias. *PLoS Pathog.* 2023;19: e1011461.
648 doi:10.1371/journal.ppat.1011461
- 649 12. Duesterwald L, Nguyen M, Christensen P, Wesley Long S, Olsen RJ, Musser JM, et al. Using
650 Genome Sequence Data to Predict SARS-CoV-2 Detection Cycle Threshold Values.
651 doi:10.1101/2022.11.14.22282297
- 652 13. Uffelmann E, Huang QQ, Munung NS, de Vries J, Okada Y, Martin AR, et al. Genome-wide
653 association studies. *Nature Reviews Methods Primers.* 2021;1: 1–21. doi:10.1038/s43586-021-
654 00056-9
- 655 14. McLaren PJ, Porreca I, Iaconis G, Mok HP, Mukhopadhyay S, Karakoc E, et al. Africa-specific

- 656 human genetic variation near CHD1L associates with HIV-1 load. *Nature*. 2023;620: 1025–1030.
657 doi:10.1038/s41586-023-06370-4
- 658 15. Karim M, Dunham I, Ghoussaini M. Mining a GWAS of Severe Covid-19. *The New England*
659 *journal of medicine*. 2020. pp. 2588–2589. doi:10.1056/NEJMc2025747
- 660 16. Roberts GHL, Partha R, Rhead B, Knight SC, Park DS, Coignet MV, et al. Expanded COVID-19
661 phenotype definitions reveal distinct patterns of genetic association and protective effects. *Nat*
662 *Genet*. 2022;54: 374–381. doi:10.1038/s41588-022-01042-x
- 663 17. Genomewide Association Study of Severe Covid-19 with Respiratory Failure. *N Engl J Med*.
664 2020;383: 1522–1534. doi:10.1056/NEJMoa2020283
- 665 18. Power RA, Parkhill J, de Oliveira T. Microbial genome-wide association studies: lessons from
666 human GWAS. *Nat Rev Genet*. 2017;18: 41–50. doi:10.1038/nrg.2016.132
- 667 19. Power RA, Davaniah S, Derache A, Wilkinson E, Tanser F, Gupta RK, et al. Genome-Wide
668 Association Study of HIV Whole Genome Sequences Validated using Drug Resistance. *PLoS*
669 *One*. 2016;11: e0163746. doi:10.1371/journal.pone.0163746
- 670 20. Ansari MA, Pedergrana V, L C Ip C, Magri A, Von Delft A, Bonsall D, et al. Genome-to-genome
671 analysis highlights the effect of the human innate and adaptive immune systems on the hepatitis
672 C virus. *Nat Genet*. 2017;49: 666–673. doi:10.1038/ng.3835
- 673 21. Vogels CBF, Breban MI, Ott IM, Alpert T, Petrone ME, Watkins AE, et al. Multiplex qPCR
674 discriminates variants of concern to enhance global surveillance of SARS-CoV-2. *PLoS Biol*.
675 2021;19: e3001236. doi:10.1371/journal.pbio.3001236
- 676 22. Zhou C, Zhang T, Ren H, Sun S, Yu X, Sheng J, et al. Impact of age on duration of viral RNA
677 shedding in patients with COVID-19. *Aging*. 2020;12: 22399–22404. doi:10.18632/aging.104114
- 678 23. Acharya CB, Schrom J, Mitchell AM, Coil DA, Marquez C, Rojas S, et al. Viral Load Among
679 Vaccinated and Unvaccinated, Asymptomatic and Symptomatic Persons Infected With the
680 SARS-CoV-2 Delta Variant. *Open Forum Infect Dis*. 2022;9: ofac135. doi:10.1093/ofid/ofac135
- 681 24. Torgerson WS. Multidimensional scaling: I. Theory and method. *Psychometrika*. 1952;17: 401–
682 419. doi:10.1007/bf02288916
- 683 25. VanderWeele TJ, Mathur MB. SOME DESIRABLE PROPERTIES OF THE BONFERRONI
684 CORRECTION: IS THE BONFERRONI CORRECTION REALLY SO BAD? *Am J Epidemiol*.
685 2018;188: 617–618. doi:10.1093/aje/kwy250
- 686 26. Ward IL, Bermingham C, Ayoubkhani D, Gethings OJ, Pouwels KB, Yates T, et al. Risk of covid-
687 19 related deaths for SARS-CoV-2 omicron (B.1.1.529) compared with delta (B.1.617.2):
688 retrospective cohort study. *BMJ*. 2022;378: e070695. doi:10.1136/bmj-2022-070695
- 689 27. Volz E, Hill V, McCrone JT, Price A, Jorgensen D, O’Toole Á, et al. Evaluating the Effects of
690 SARS-CoV-2 Spike Mutation D614G on Transmissibility and Pathogenicity. *Cell*. 2021;184: 64–
691 75.e11. doi:10.1016/j.cell.2020.11.020
- 692 28. Tian D, Sun Y, Zhou J, Ye Q. The Global Epidemic of the SARS-CoV-2 Delta Variant, Key Spike
693 Mutations and Immune Escape. *Front Immunol*. 2021;12: 751778.
694 doi:10.3389/fimmu.2021.751778
- 695 29. Hodcroft EB, Domman DB, Snyder DJ, Oguntuyo KY, Van Diest M, Densmore KH, et al.
696 Emergence in late 2020 of multiple lineages of SARS-CoV-2 Spike protein variants affecting
697 amino acid position 677. *medRxiv*. 2021. doi:10.1101/2021.02.12.21251658
- 698 30. Cosar B, Karagulleoglu ZY, Unal S, Ince AT, Uncuoglu DB, Tuncer G, et al. SARS-CoV-2
699 Mutations and their Viral Variants. *Cytokine Growth Factor Rev*. 2022;63: 10–22.
700 doi:10.1016/j.cytogfr.2021.06.001

- 701 31. Chen J, Wang R, Wang M, Wei G-W. Mutations Strengthened SARS-CoV-2 Infectivity. *J Mol*
702 *Biol.* 2020;432: 5212–5226. doi:10.1016/j.jmb.2020.07.009
- 703 32. Harvey WT, Carabelli AM, Jackson B, Gupta RK, Thomson EC, Harrison EM, et al. SARS-CoV-2
704 variants, spike mutations and immune escape. *Nat Rev Microbiol.* 2021;19: 409–424.
705 doi:10.1038/s41579-021-00573-0
- 706 33. Gangavarapu K, Latif AA, Mullen JL, Alkuzweny M, Hufbauer E, Tsueng G, et al. Outbreak.info
707 genomic reports: scalable and dynamic surveillance of SARS-CoV-2 variants and mutations. *Nat*
708 *Methods.* 2023;20: 512–522. doi:10.1038/s41592-023-01769-3
- 709 34. Tang X, Wu C, Li X, Song Y, Yao X, Wu X, et al. On the origin and continuing evolution of SARS-
710 CoV-2. *Natl Sci Rev.* 2020;7: 1012–1023. doi:10.1093/nsr/nwaa036
- 711 35. Carabelli AM, Peacock TP, Thorne LG, Harvey WT, Hughes J, COVID-19 Genomics UK
712 Consortium, et al. SARS-CoV-2 variant biology: immune escape, transmission and fitness. *Nat*
713 *Rev Microbiol.* 2023;21: 162–177. doi:10.1038/s41579-022-00841-7
- 714 36. Souza PFN, Mesquita FP, Amaral JL, Landim PGC, Lima KRP, Costa MB, et al. The spike
715 glycoprotein of SARS-CoV-2: A review of how mutations of spike glycoproteins have driven the
716 emergence of variants with high transmissibility and immune escape. *Int J Biol Macromol.*
717 2022;208: 105–125. doi:10.1016/j.ijbiomac.2022.03.058
- 718 37. Kimura I, Kosugi Y, Wu J, Zahradnik J, Yamasoba D, Butlertanaka EP, et al. The SARS-CoV-2
719 Lambda variant exhibits enhanced infectivity and immune resistance. *Cell Rep.* 2022;38:
720 110218. doi:10.1016/j.celrep.2021.110218
- 721 38. Yadav PD, Bergeron É, Flora MS. Emerging SARS-COV-2 Variants: Genomic Variations,
722 Transmission, Pathogenesis, Clinical Impact and Interventions. *Frontiers Media SA;* 2023.
723 Available: <https://play.google.com/store/books/details?id=AHi6EAAAQBAJ>
- 724 39. BMvd V, Dingemans J, Bank LE, von Wintersdorff CJ. Viral load dynamics in healthcare workers
725 with COVID-19 during Delta and Omicron era. [cited 10 Feb 2024]. Available:
726 <https://europepmc.org/article/ppr/ppr485029>
- 727 40. Lentini A, Pereira A, Winqvist O, Reinius B. Monitoring of the SARS-CoV-2 Omicron BA.1/BA.2
728 variant transition in the Swedish population reveals higher viral quantity in BA.2 cases. *bioRxiv.*
729 2022. doi:10.1101/2022.03.26.22272984
- 730 41. Russell TW, Townsley H, Abbott S, Hellewell J, Carr EJ, Chapman LAC, et al. Combined
731 analyses of within-host SARS-CoV-2 viral kinetics and information on past exposures to the
732 virus in a human cohort identifies intrinsic differences of Omicron and Delta variants. *PLoS Biol.*
733 2024;22: e3002463. doi:10.1371/journal.pbio.3002463
- 734 42. Kopsidas I, Karagiannidou S, Kostaki EG, Kousi D, Douka E, Sfrikakis PP, et al. Global
735 Distribution, Dispersal Patterns, and Trend of Several Omicron Subvariants of SARS-CoV-2
736 across the Globe. *Trop Med Infect Dis.* 2022;7. doi:10.3390/tropicalmed7110373
- 737 43. Motozono C, Toyoda M, Tan TS, Hamana H, Goto Y, Aritsu Y, et al. The SARS-CoV-2 Omicron
738 BA.1 spike G446S mutation potentiates antiviral T-cell recognition. *Nat Commun.* 2022;13:
739 5440. doi:10.1038/s41467-022-33068-4
- 740 44. Lin X, Sha Z, Trimpert J, Kunec D, Jiang C, Xiong Y, et al. The NSP4 T492I mutation increases
741 SARS-CoV-2 infectivity by altering non-structural protein cleavage. *Cell Host Microbe.* 2023;31:
742 1170–1184.e7. doi:10.1016/j.chom.2023.06.002
- 743 45. Luo CH, Morris CP, Sachithanandham J, Amadi A, Gaston D, Li M, et al. Infection with the
744 SARS-CoV-2 Delta Variant is Associated with Higher Infectious Virus Loads Compared to the
745 Alpha Variant in both Unvaccinated and Vaccinated Individuals. *medRxiv.* 2021.
746 doi:10.1101/2021.08.15.21262077

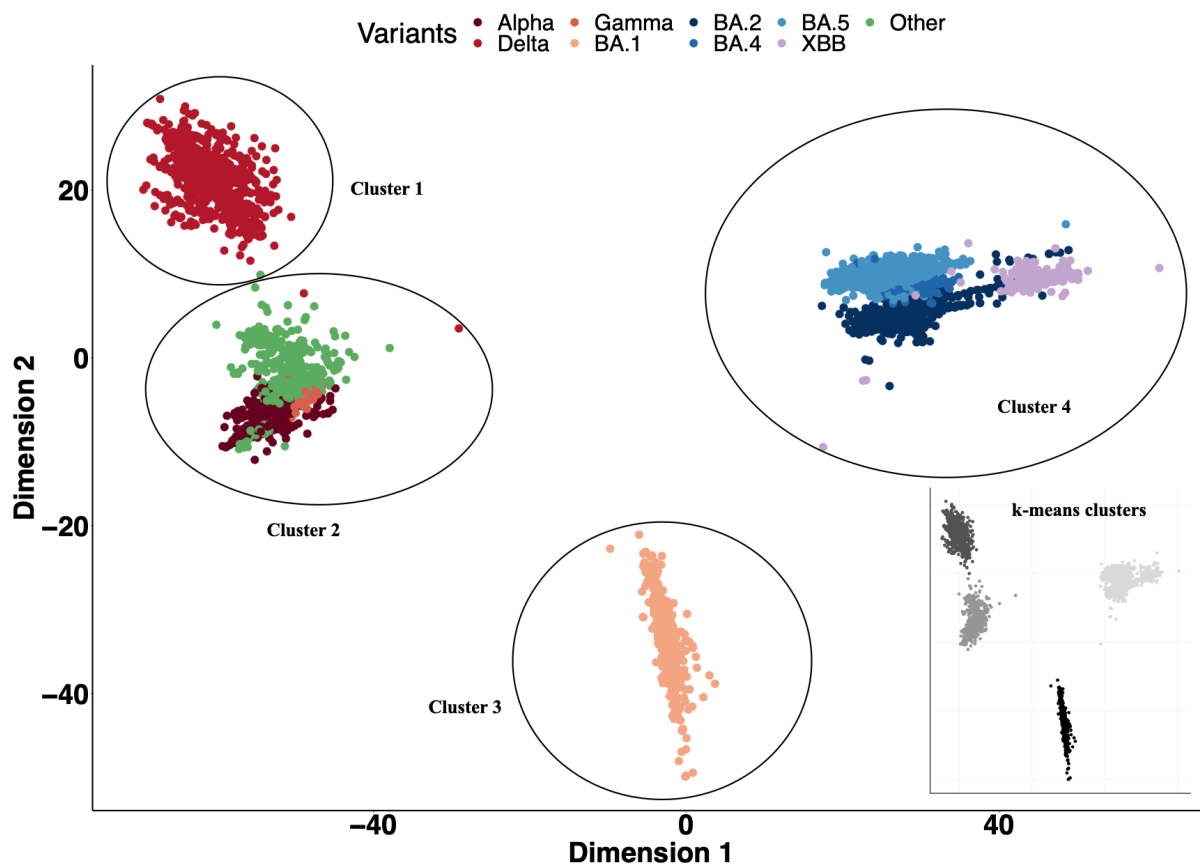
- 747 46. von Wintersdorff CJH, Dingemans J, van Alphen LB, Wolffs PFG, van der Veer BMJW, Hoebe
748 CJPA, et al. Infections with the SARS-CoV-2 Delta variant exhibit fourfold increased viral loads
749 in the upper airways compared to Alpha or non-variants of concern. *Sci Rep.* 2022;12: 13922.
750 doi:10.1038/s41598-022-18279-5
- 751 47. Li B, Deng A, Li K, Hu Y, Li Z, Shi Y, et al. Viral infection and transmission in a large, well-traced
752 outbreak caused by the SARS-CoV-2 Delta variant. *Nat Commun.* 2022;13: 460.
753 doi:10.1038/s41467-022-28089-y
- 754 48. Fall A, Eldesouki RE, Sachithanandham J, Morris CP, Norton JM, Gaston DC, et al. The
755 displacement of the SARS-CoV-2 variant Delta with Omicron: An investigation of hospital
756 admissions and upper respiratory viral loads. *EBioMedicine.* 2022;79: 104008.
757 doi:10.1016/j.ebiom.2022.104008
- 758 49. Hay JA, Kissler SM, Fauver JR, Mack C, Tai CG, Samant RM, et al. Quantifying the impact of
759 immune history and variant on SARS-CoV-2 viral kinetics and infection rebound: A retrospective
760 cohort study. *Elife.* 2022;11. doi:10.7554/eLife.81849
- 761 50. Jones RP, Ponomarenko A. COVID-19-Related Age Profiles for SARS-CoV-2 Variants in England
762 and Wales and States of the USA (2020 to 2022): Impact on All-Cause Mortality. *Infect Dis Rep.*
763 2023;15: 600–634. doi:10.3390/idr15050058
- 764 51. Bartha I, Carlson JM, Brumme CJ, McLaren PJ, Brumme ZL, John M, et al. A genome-to-
765 genome analysis of associations between human genetic variation, HIV-1 sequence diversity,
766 and viral control. *Elife.* 2013;2: e01123. doi:10.7554/eLife.01123
- 767 52. Lu X, Wang L, Sakthivel SK, Whitaker B, Murray J, Kamili S, et al. US CDC Real-Time Reverse
768 Transcription PCR Panel for Detection of Severe Acute Respiratory Syndrome Coronavirus 2.
769 *Emerg Infect Dis.* 2020;26: 1654–1665. doi:10.3201/eid2608.201246
- 770 53. Vogels CBF, Brito AF, Wyllie AL, Fauver JR, Ott IM, Kalinich CC, et al. Analytical sensitivity and
771 efficiency comparisons of SARS-CoV-2 RT-qPCR primer-probe sets. *Nature Microbiology.*
772 2020;5: 1299–1305. doi:10.1038/s41564-020-0761-6
- 773 54. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv*
774 [q-bio.GN]. 2013. Available: <http://arxiv.org/abs/1303.3997>
- 775 55. Grubaugh ND, Gangavarapu K, Quick J, Matteson NL, De Jesus JG, Main BJ, et al. An
776 amplicon-based sequencing framework for accurately measuring intrahost virus diversity using
777 PrimalSeq and iVar. *Genome Biol.* 2019;20: 8. doi:10.1186/s13059-018-1618-7
- 778 56. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of
779 SAMtools and BCFtools. *Gigascience.* 2021;10. doi:10.1093/gigascience/giab008
- 780 57. O'Toole Á, Scher E, Underwood A, Jackson B, Hill V, McCrone JT, et al. Assignment of
781 epidemiological lineages in an emerging pandemic using the pangolin tool. *Virus Evol.* 2021;7:
782 veab064. doi:10.1093/ve/veab064
- 783 58. Aksamentov I, Roemer C, Hodcroft E, Neher R. Nextclade: clade assignment, mutation calling
784 and quality control for viral genomes. *J Open Source Softw.* 2021;6: 3773.
785 doi:10.21105/joss.03773
- 786 59. Page AJ, Taylor B, Delaney AJ, Soares J, Seemann T, Keane JA, et al. SNP-sites: rapid efficient
787 extraction of SNPs from multi-FASTA alignments. *Microb Genom.* 2016;2: e000056.
788 doi:10.1099/mgen.0.000056
- 789 60. Creators Seemann, Torsten1 Show affiliations 1. The University of Melbourne. Source code for
790 snp-dists software. doi:10.5281/zenodo.1411986
- 791 61. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, et al.

792 Corrigendum to: IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in
793 the Genomic Era. *Mol Biol Evol.* 2020;37: 2461. doi:10.1093/molbev/msaa131

794

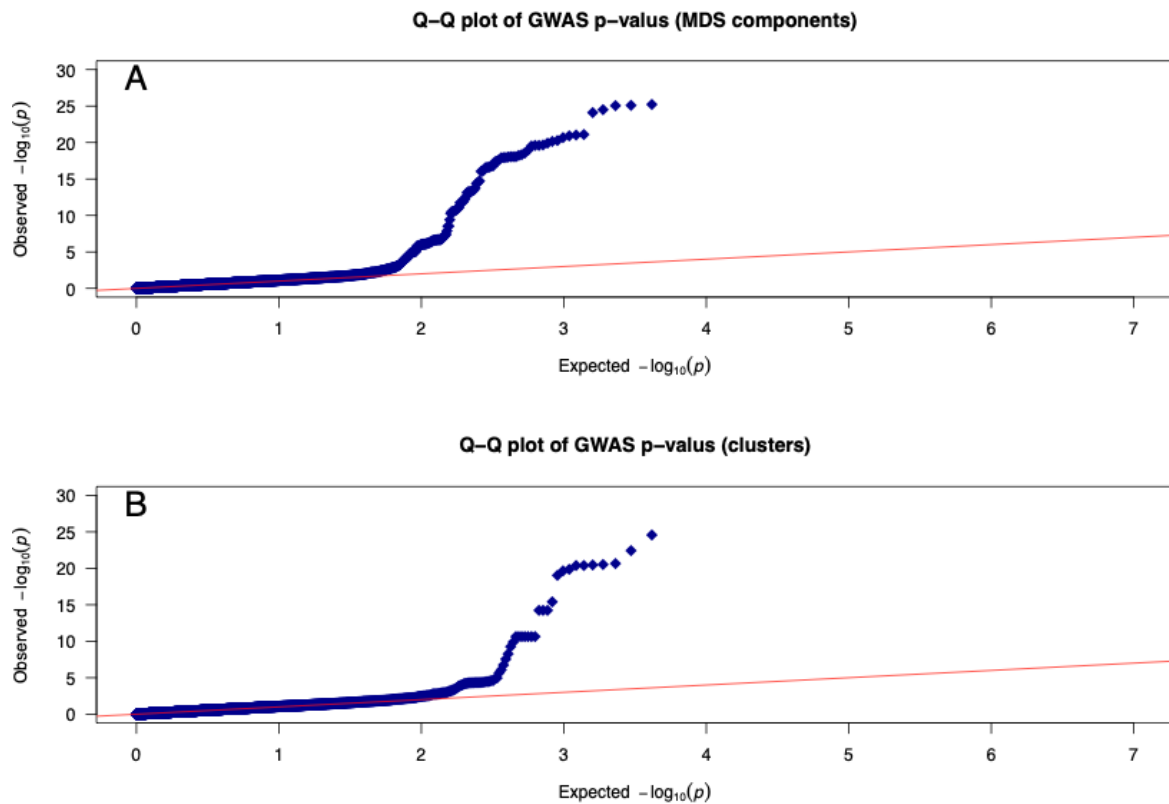
795

796 Supplementary Figures & Tables



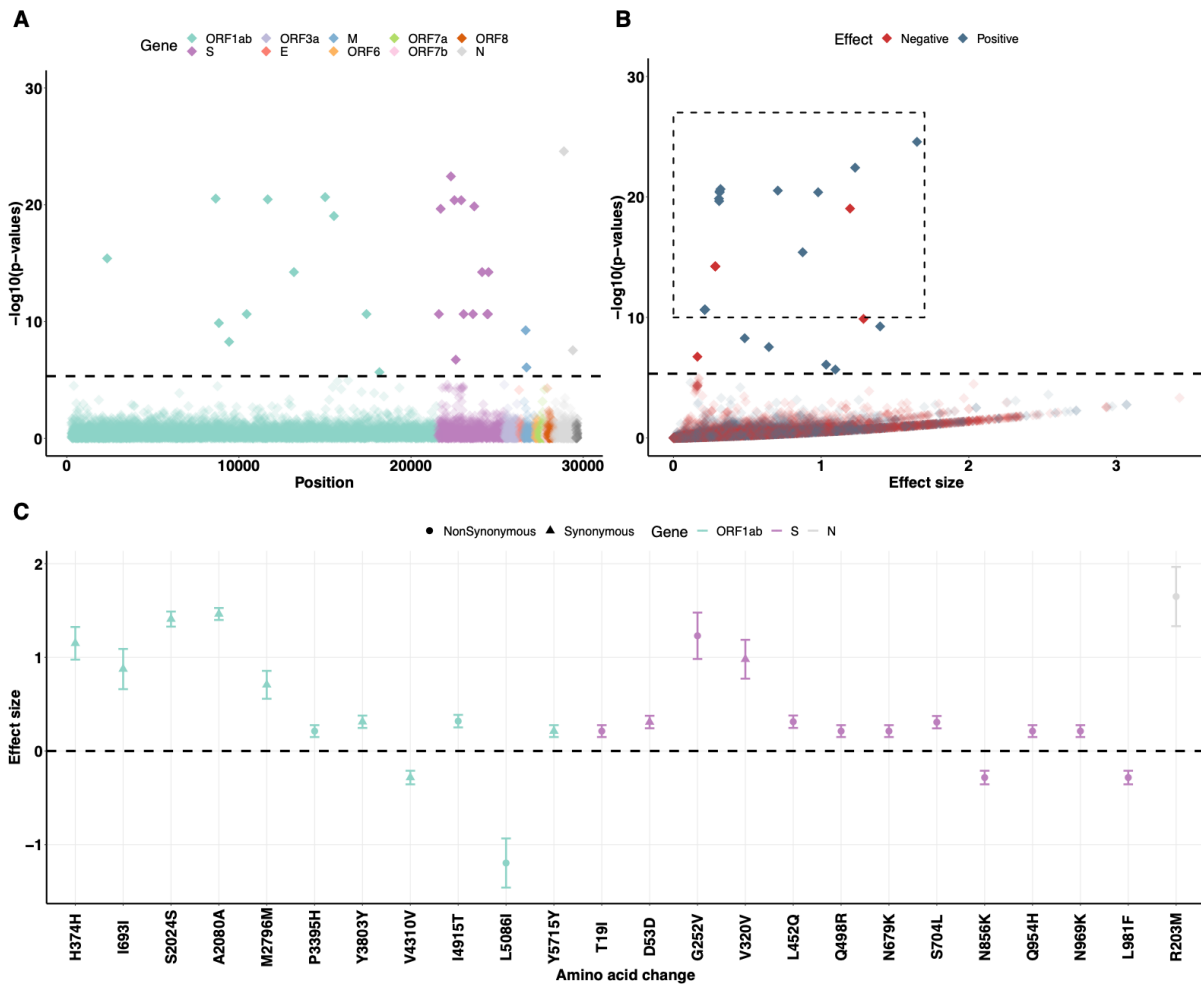
797
798
799
800
801
802

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



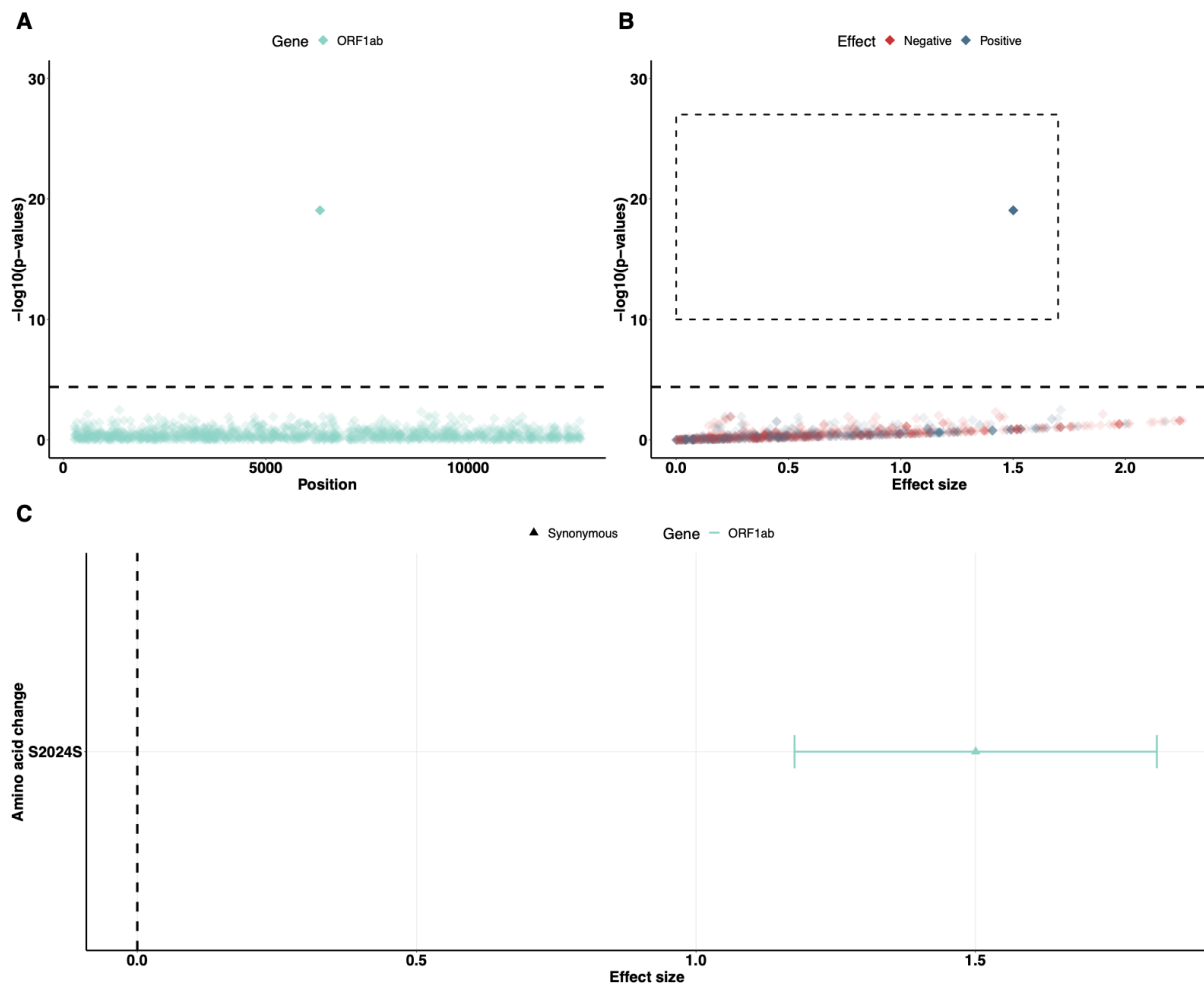
803
804
805
806
807
808
809
810
811
812
813
814
815
816
817
818
819

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



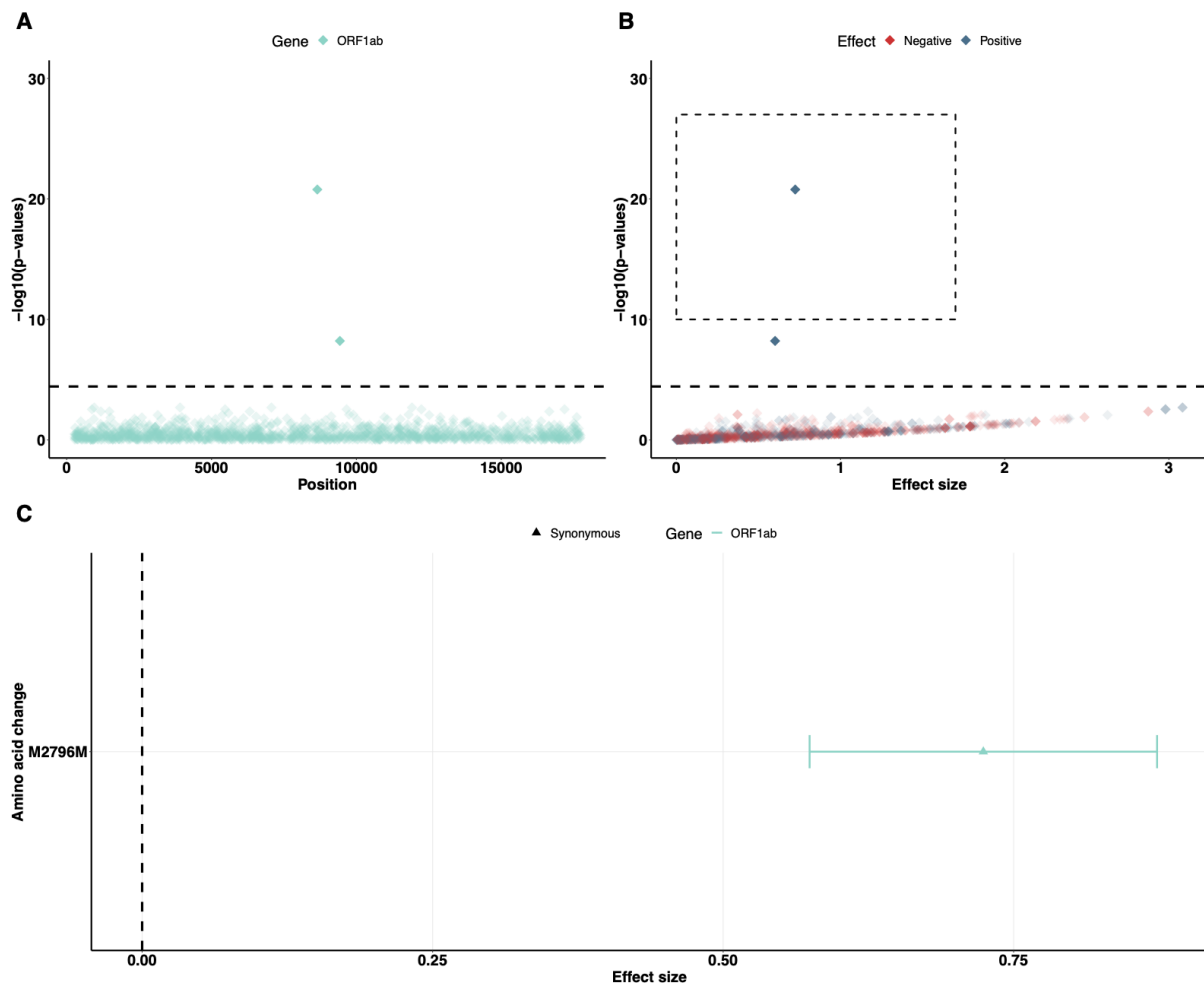
820
821
822
823
824
825
826
827
828
829
830

Supplemental Figure 3. GWAS analysis using the four clusters (shown in Fig. S1) as covariates for population control. (A) Genome-wide association results of the impact of identified SNPs on viral copies during SARS-CoV-2 infection. The dashed line indicates the permuted threshold for genome-wide significance $p = 4.67 \times 10^{-6}$ ($0.05/10697$ SNPs). Significant SNPs are shown with solid colors. **(B)** SNPs (with $p < 1 \times 10^{-10}$) that have positive (blue) or negative (red) effects on viral copies. **(C)** The corresponding synonymous (triangles) and non-synonymous (circles) amino acid changes that associate with increased or decreased viral copies. Data shown as means with 95% confidence intervals. The estimated effective sizes and associated standard deviations are given **Table S2**.



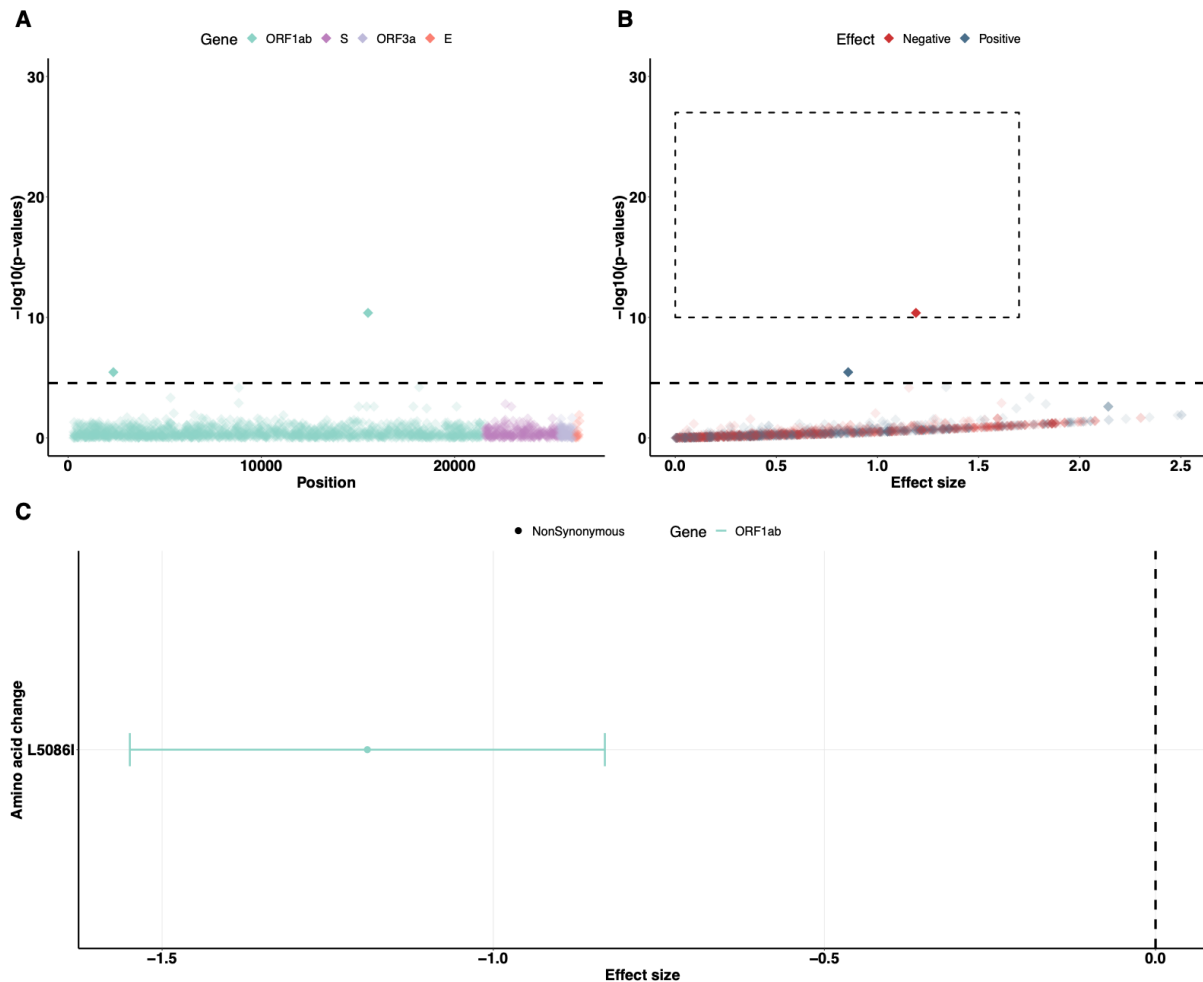
831
832
833
834
835
836
837
838
839

Supplemental Figure 4. GWAS analysis using only Cluster 1 data (shown in Fig. S1). (A) Genome-wide association results of the impact of identified SNPs on viral copies during SARS-CoV-2 infection. The dashed line indicates the permuted threshold for genome-wide significance $p = 4.03 \times 10^{-5}$ (0.05/1242 SNPs). Significant SNPs are shown with solid colors. (B) SNPs (with $p < 1 \times 10^{-10}$) that have positive (blue) or negative (red) effects on viral copies. (C) The corresponding synonymous (triangles) amino acid changes that associate with increased or decreased viral copies. Data shown as means with 95% confidence intervals.



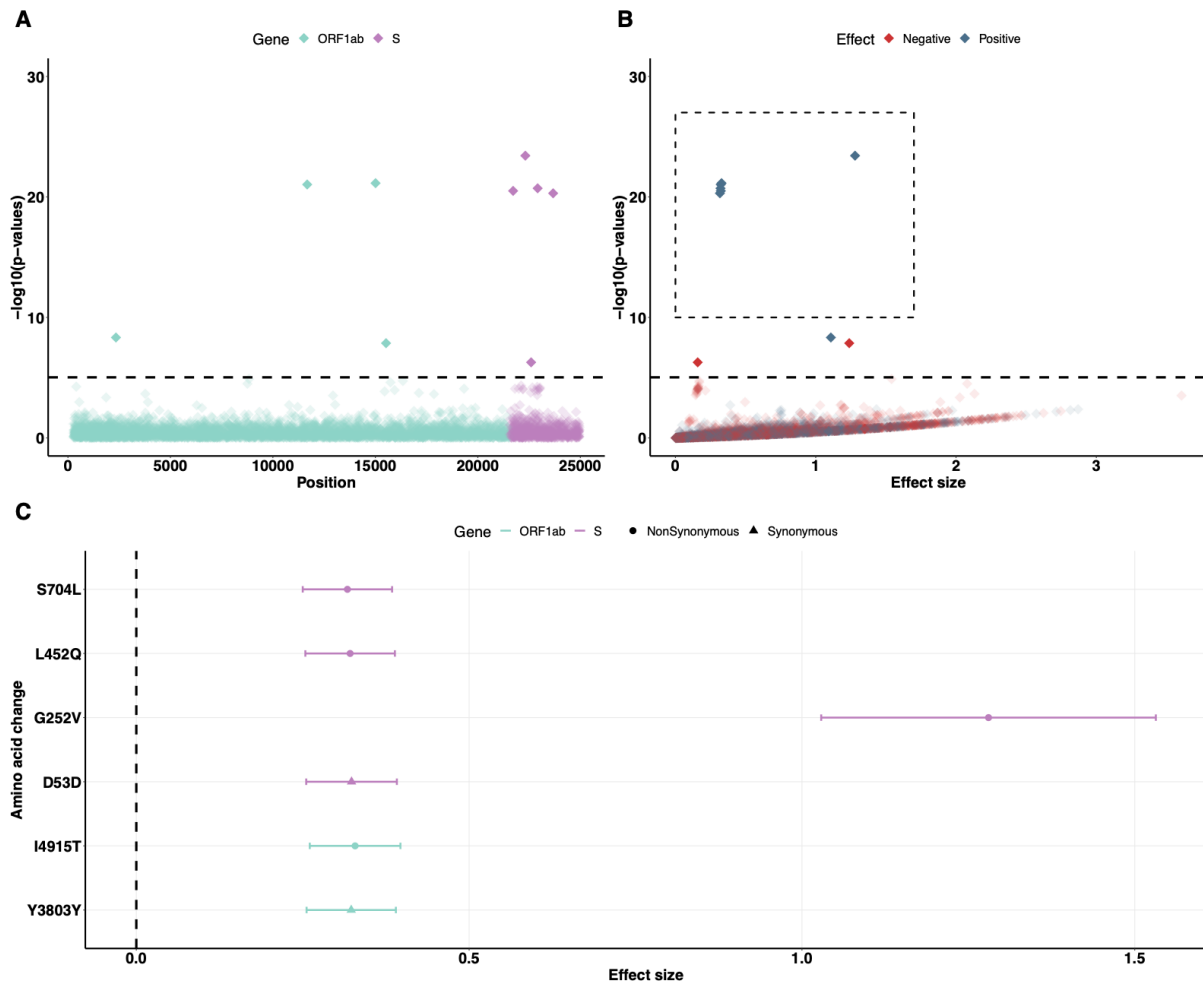
840
841
842
843
844
845
846
847
848
849
850

Supplemental Figure 5. GWAS analysis using Cluster 2 data (shown in Fig. S1). (A) Genome-wide association results of the impact of identified SNPs on viral copies during SARS-CoV-2 infection. The dashed line indicates the permuted threshold for genome-wide significance $p = 3.68 \times 10^{-5}$ (0.05/1357 SNPs). Significant SNPs are shown with solid colors. (B) SNPs (with $p < 1 \times 10^{-10}$) that have positive (blue) or negative (red) effects on viral copies. (C) The corresponding synonymous (triangles) amino acid changes that associate with increased or decreased viral copies. Data shown as means with 95% confidence intervals.



851
852 **Supplemental Figure 6. GWAS analysis using Cluster 3 data (shown in Fig. S1).** (A) Genome-wide
853 association results of the impact of identified SNPs on viral copies during SARS-CoV-2 infection. The
854 dashed line indicates the permuted threshold for genome-wide significance $p = 2.80 \times 10^{-5}$ (0.05/1784
855 SNPs). Significant SNPs are shown with solid colors. (B) SNPs (with $p < 1 \times 10^{-10}$) that have positive
856 (blue) or negative (red) effects on viral copies. (C) The corresponding non-synonymous (circles) amino
857 acid changes that associate with increased or decreased viral copies. Data shown as means with 95%
858 confidence intervals.

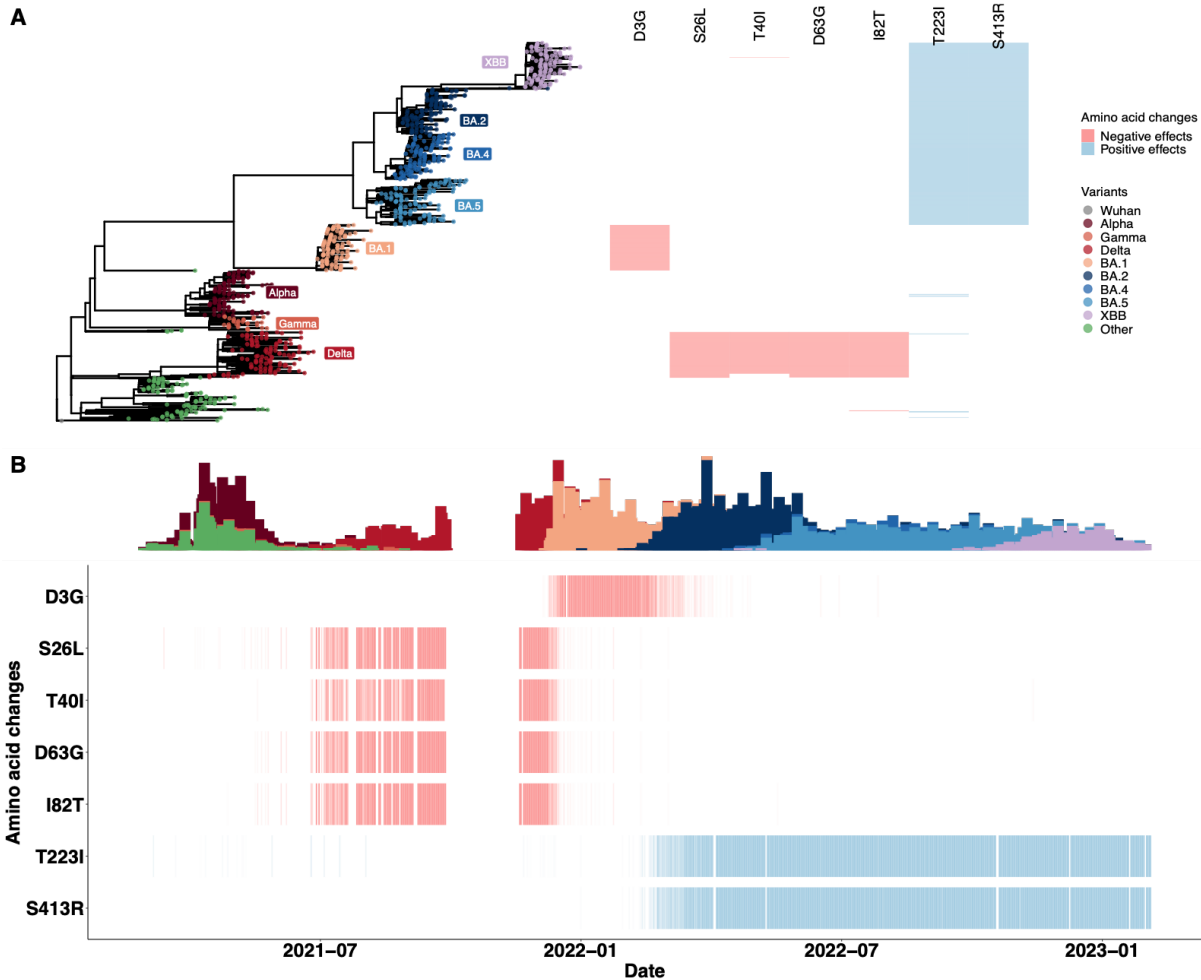
859
860
861
862
863



864
865
866
867
868
869
870
871
872
873
874
875
876

Supplemental Figure 7. GWAS analysis using Cluster 4 data (shown in Fig. S1). (A) Genome-wide association results of the impact of identified SNPs on viral copies during SARS-CoV-2 infection. The dashed line indicates the permuted threshold for genome-wide significance $p = 7.91 \times 10^{-6}$ (0.05/6314 SNPs). Significant SNPs are shown with solid colors. (B) SNPs (with $p < 1 \times 10^{-10}$) that have positive (blue) or negative (red) effects on viral copies. (C) The corresponding synonymous (triangles) and non-synonymous (circles) amino acid changes that associate with increased or decreased viral copies. Data shown as means with 95% confidence intervals.

877



878

879

880

881

882

883

884

885

886

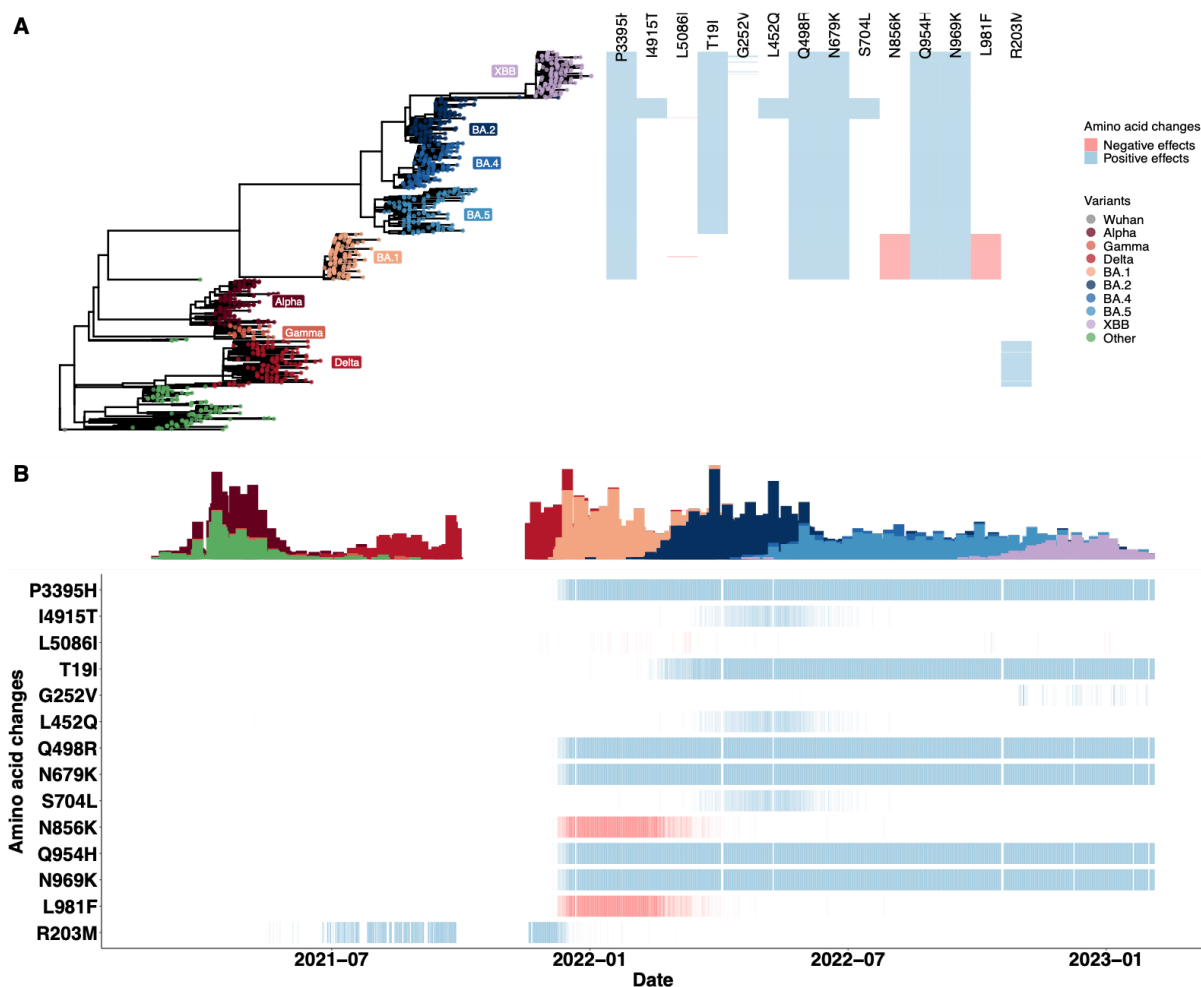
887

888

889

890

Supplemental Figure 8. The temporal dynamics of amino acid changes in the ORF3a gene (S26L and T223I), M gene (D3G and I82T), ORF7b gene (T40I) and N gene (D63G and S413R) associated with changes in viral copies. The results are based on the multivariate regression analysis using the two MDS components as covariates. **(A)** The phylogenetic tree estimated from a representative set of 996 genome sequences showing variant assignments and the locations of amino acid changes that increase (blue) or decrease (red) viral copies. **(B)** The temporal dynamics of the SNPs from February 2021 to March 2023. The transparency of the color corresponds to the mutation fraction in the daily sequence count: transparent color indicates low fractions, and opaque color indicates high fractions.



891
892 **Supplemental Figure 9. The temporal dynamics of amino acid changes in the ORF1ab gene**
893 **(P3395H, I4915T and L5086I), S gene (T19I, G252V, L452Q, Q498R, N679K, S704L, N856K, Q954H,**
894 **N969K and L981F), and N gene (R203M) associated with changes in viral copies.** The results are
895 based on the multivariate regression analysis using the sequence clusters (i.e., a categorical variable)
896 inferred from the MDS components, such that $Y \sim \alpha W + \beta_i SNP_i + \eta Cluster + e$. **(A)** The phylogenetic
897 tree estimated from a representative set of 996 genome sequences showing variant assignments and
898 the locations of amino acid changes that increase (blue) or decrease (red) viral copies. **(B)** The temporal
899 dynamics of the SNPs from February 2021 to March 2023. The transparency of the color corresponds
900 to the mutation fraction in the daily sequence count: transparent color indicates low fractions, and
901 opaque color indicates high fractions.

902
903
904
905

AminoAcidChange	coefficients	Standard deviation
D3G	-0.5289821	0.06197068
T19R	-0.3583990	0.05236202
T19I	0.8675393	0.09368981
S26L	-0.3470353	0.05215769
T40I	-0.3810594	0.05035466
D63G	-0.3508300	0.05236169
A67V	-0.5034788	0.06039024
I82T	-0.3456858	0.05196310
S135R	0.8074352	0.09059436
V213G	0.2323215	0.03223704
T223I	0.5872878	0.08264916
G252V	1.1976598	0.12468634
S371L	-0.5534597	0.06193220
T376A	0.8620845	0.08999643
D405N	0.8797794	0.09241162
R408S	0.6207915	0.08287705
S413R	0.8157651	0.09230739
G446S	-0.2664795	0.03198609
L452Q	0.3466409	0.03292778
G496S	-0.5499511	0.06210286
T547K	-0.5341148	0.06147332
P681R	-0.3506782	0.05225140
S704L	0.3418159	0.03284982
T842I	0.7544330	0.08850513
N856K	-0.5521347	0.06227279
K856R	-0.5393076	0.06198051
D950N	-0.3534498	0.05196289
L981F	-0.5521347	0.06227279
A1306S	-0.3553043	0.05035843
G1307S	0.8259626	0.08905459

906			
907	P2287S	-0.3541119	0.05015436
908	A2710T	-0.5450373	0.06223708
909			
910	V2857A	-1.3198257	0.20083279
911	V2930L	-0.3796890	0.05042589
912			
913	L3027F	0.8277260	0.08812009
914	T3090I	0.7634935	0.08992720
915			
916	T3255I	-0.2507549	0.03705032
917			
918	T3646A	-0.3687172	0.05037977
919	I3758V	-0.5466932	0.06187469
920			
921	I4915T	0.3524137	0.03336264
922	L5086I	-1.2284741	0.13159986
923			
924	S5150F	0.8465102	0.09162000
925			
926	H5401Y	-0.3689352	0.05186462
927	K6597R	0.8501061	0.09021831
928			
929			

930 **Supplemental Table 1. The identified amino acid changes associated estimated effective sizes**
931 **and standard deviations using the multivariate linear regression model with MDS-computed**
932 **dimensions as covariates.**

933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951

	AminoAcidChange	coefficients	Standard deviation
952			
953			
954			
955	T19I	0.2123842	0.03172457
956			
957	R203M	1.6492504	0.15825678
958			
959	G252V	1.2301542	0.12381144
960			
961	L452Q	0.3127277	0.03308999
962			
963	Q498R	0.2123842	0.03172457
964			
965	N679K	0.2123842	0.03172457
966			
967	S704L	0.3077803	0.03301503
968			
969	N856K	-0.2832514	0.03621949
970			
971	Q954H	0.2123842	0.03172457
972			
973	N969K	0.2123842	0.03172457
974			
975	L981F	-0.2832514	0.03621949
976			
977	P3395H	0.2123842	0.03172457
978			
979	I4915T	0.3189265	0.03351622
	L5086I	-1.1958547	0.13116867

980 **Supplemental Table 2. The identified amino acid changes associated estimated effective sizes**
981 **and standard deviations using the multivariate linear regression model with categorical clusters**
982 **as covariates.**
983
984