

Determinants of SARS-CoV-2 within-host evolutionary rates in persistently infected individuals

Authors: Mahan Ghafari^{1,2}, Steven A. Kemp^{1,2}, Matthew Hall^{1,2}, Joe Clarke², Luca Ferretti^{1,2}, Laura Thomson^{1,2}, Ruth Studley³, Emma Rourke³, COVID-19 Infection Survey Group, The COVID-19 Genomics UK (COG-UK) Consortium, Ann Sarah Walker^{4,5,6}, Tanya Golubchik^{2,7}, Katrina Lythgoe^{1,2,8}

¹Pandemic Sciences Institute, University of Oxford, Oxford, UK

²Big Data Institute, Nuffield Department of Medicine, University of Oxford, Oxford, UK

³Office for National Statistics, Newport, UK

⁴Nuffield Department of Medicine, University of Oxford, Oxford, UK

⁵The National Institute for Health Research Health Protection Research Unit in Healthcare Associated Infections and Antimicrobial Resistance at the University of Oxford, Oxford, UK.

⁶The National Institute for Health Research Oxford Biomedical Research Centre, University of Oxford, Oxford, UK

⁷The Sydney Infectious Diseases Institute (Sydney ID), School of Medical Sciences, University of Sydney, Sydney, Australia

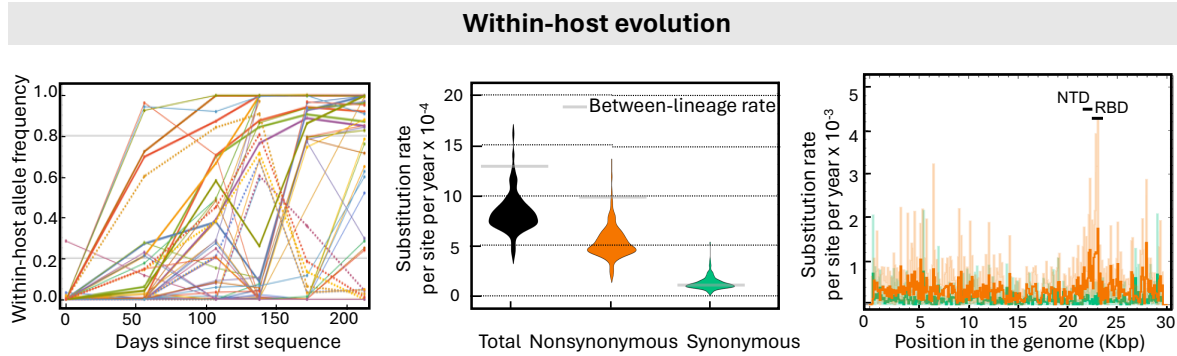
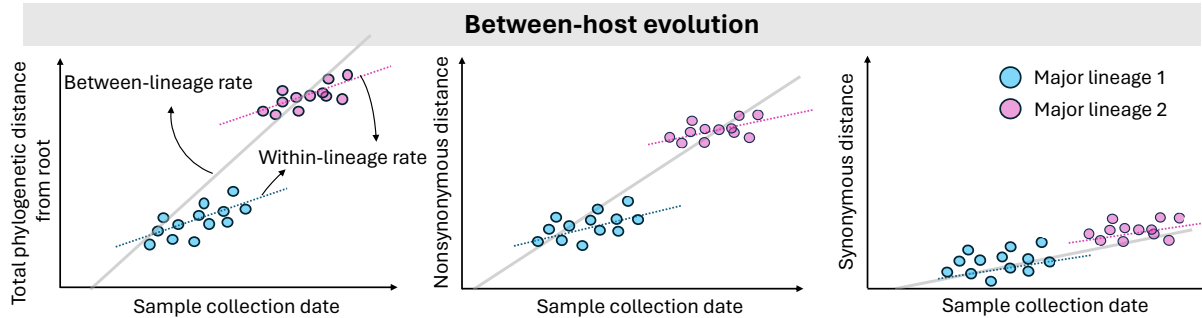
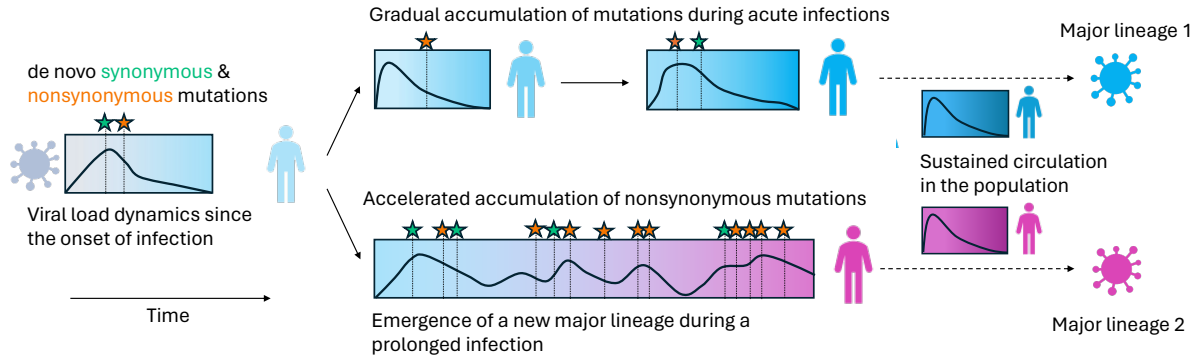
⁸Department of Biology, University of Oxford, Oxford, UK

*Correspondence: mahan.ghafari@ndm.ox.ac.uk

Abstract

Understanding the within-host evolutionary dynamics of SARS-CoV-2, particularly in relation to variant emergence, is crucial for public health. From a community surveillance study, we identified 576 persistent infections, more common among males and those over 60. Our findings show significant variation in evolutionary rates among individuals, driven by nonsynonymous mutations. Longer-lasting infections accumulated mutations faster, with no link to demographics, vaccination status, virus lineage, or prior infection. The nonsynonymous rate was particularly high within the N-terminal and receptor binding domains of *Spike*. *ORF6* was under strong purifying selection, making it a potential therapeutic target. We also identified 379 recurring mutations, with half having a negative fitness effect and very low prevalence at the between-host level, indicating some mutations are favoured during infection but disadvantageous for transmission. Our study highlights the highly heterogeneous nature of within-host evolution of SARS-CoV-2 which may in turn help inform future intervention strategies.

Keywords: SARS-CoV-2 evolution, chronic infections, within-host evolution, adaptive evolution, community surveillance, therapeutics



1 Introduction

2
3 The evolutionary dynamics of SARS-CoV-2 has been marked by the emergence of
4 highly divergent variants, including initial variants of concern (VOCs) Alpha, Beta,
5 Gamma, Delta, and Omicron, followed by second-generation variants such as BA.2.75,
6 XBB.1.5, and JN.1¹⁻³. A notable feature of these variants is that they have a large
7 number of nonsynonymous mutations compared to their closest ancestors, particularly
8 in the Spike protein's N-terminal domain (NTD) and receptor-binding domain (RBD), and
9 show signs of strong positive selection driven by increased transmissibility and antibody
10 immune escape^{4,5}. Within-host evolution of SARS-CoV-2 likely plays a key role in
11 shaping these patterns of evolutionary change over time. Many chronically infected
12 individuals also show evidence of strong viral adaptive evolution, characterised by
13 accelerated evolutionary rates that feature key lineage-defining mutations in Spike^{1,6,7}.
14 Given the likely importance of long-term (persistent) infections on the evolution of the
15 virus at the population scale, we sought to characterise the evolution of SARS-CoV-2 in
16 'typical' persistent infections.

17
18 The majority of studies on the evolutionary dynamics of persistent SARS-CoV-2
19 infections have focussed on chronic cases. These are infections with consistently high
20 viral titres, and are often found in hospitalised patients who are immunocompromised
21 and receiving treatments. However, we recently showed that persistent SARS-CoV-2
22 infections, many of which have rebounding viral loads, are also prevalent in the general
23 population⁸. There remains a major gap in our understanding of host factors
24 contributing to higher odds of experiencing persistent infections, reasons why the virus
25 undergoes accelerated adaptive evolution in certain individuals, but not in others,
26 identifying genomic regions and mutations, particularly outside of Spike, that undergo
27 adaptive evolution during persistent infections, and ultimately developing effective
28 therapeutics to clear viral infections^{9,10}. Characterisation of evolution is particularly
29 important to determine if adaptive changes during infections mirror the saltatory
30 evolution of SARS-CoV-2 observed with the emergence of new, highly divergent
31 variants. In addition, identifying mutations that present complex trade-offs, being
32 advantageous at the within-host level and detrimental at the between-host level, is
33 crucial for understanding evolutionary factors that contribute to prolonged viral
34 replication within hosts and increased odds of transmission between hosts^{6,11}.

35
36 Here, we explored the within-host evolutionary dynamics of SARS-CoV-2 in 576
37 persistently infected individuals, who participated in the Office for National Statistics
38 Covid-19 Infection Survey (ONS-CIS), and identified factors associated with rate
39 differences between individuals. Investigating the evolutionary dynamics of SARS-CoV-
40 2 within persistent infections is essential for understanding the selective pressures that
41 shape viral evolution at the within-host level, factors contributing to increased risk of
42 resistance to treatments, and also to gauge the extent to which these individuals may
43 contribute to the generation and subsequent spread of new variants¹²⁻¹⁴.

44

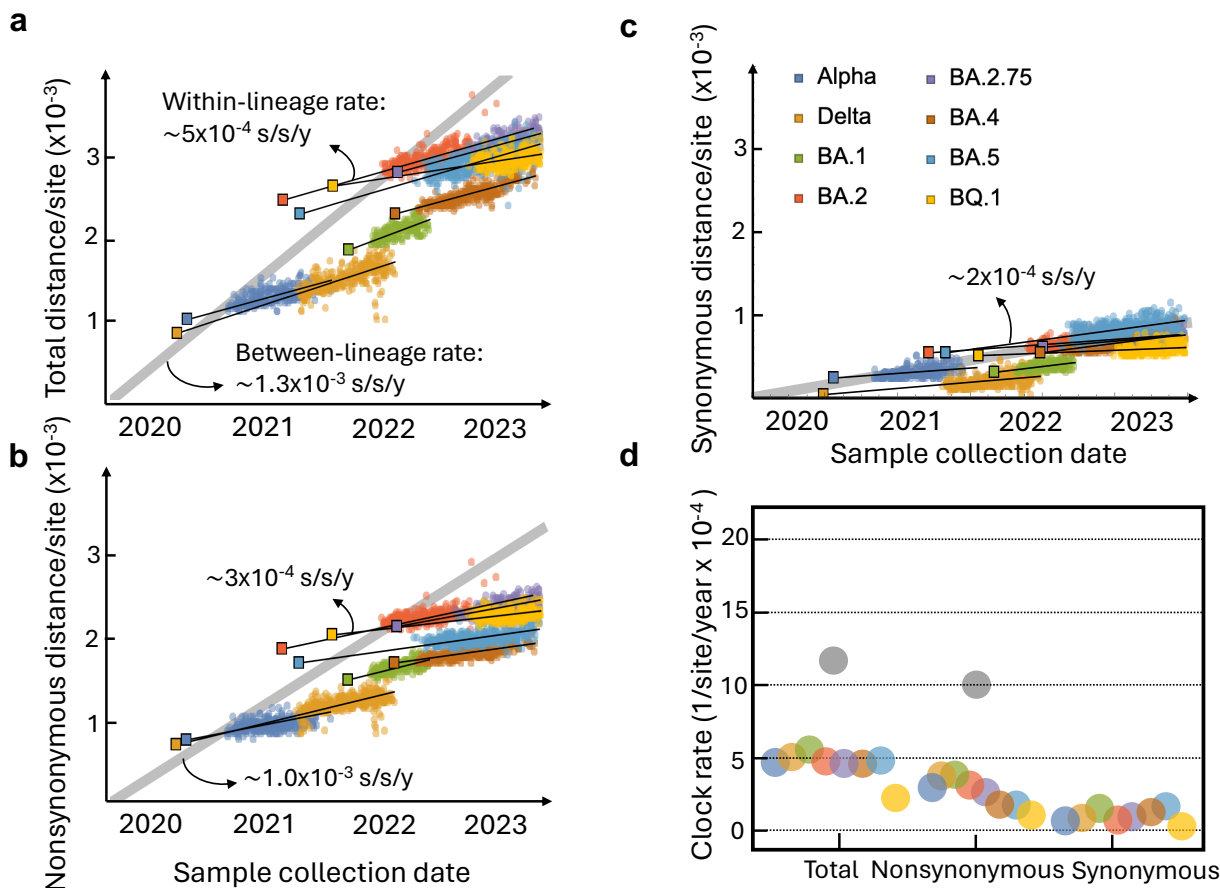
45 Results

46

47 Saltatory evolution between major lineages for nonsynonymous but not synonymous 48 mutations

49

50 Our analysis of the evolutionary dynamics of SARS-CoV-2 at the between-host level
51 identifies two distinct patterns of mutation accumulation: within-lineage and between-
52 lineage rates. Within each major viral lineage, mutations accumulate linearly over time,
53 indicating a steady evolutionary clock (**Figure 1**; see **Methods**).



54

55 **Figure 1: Evolutionary dynamics of SARS-CoV-2 at the between-host level. (a)** Mutations
56 accumulate linearly over time within each major viral lineage, punctuated by significant
57 evolutionary leaps that demarcate these lineages (between-lineage rate; grey line). **(b)** This
58 pattern is characterised by a disproportionate accumulation of nonsynonymous mutations at the
59 point of transition between major lineages, whereas **(c)** synonymous mutations accumulate at a
60 comparatively steady rate both within and across these lineages. Genetic distance within each
61 major lineage is the Hamming distance between the putative ancestral sequence (shown with
62 square markers) of that major lineage. The between-lineage distance is calculated as the
63 Hamming distance between Wuhan reference sequence (NC_045512.2) and the putative
64 ancestors of each major lineage. Lines represent the best fit from a linear regression. **(d)**
65 Substitution rate per site per year (s/s/y) for genome-wide (total), nonsynonymous, and
66 synonymous mutations, over time per major lineage. The substitution rates are $2.5\text{-}6.0 \times 10^{-4}$
67 s/s/y for genome-wide, $1.5\text{-}4.0 \times 10^{-4}$ s/s/y for nonsynonymous, and $0.5\text{-}2.5 \times 10^{-4}$ s/s/y for
68 synonymous mutations per major lineage. The between-lineage rate is highlighted with grey
69 circles.

70

71 The within-lineage rate is characterised by nonsynonymous and synonymous mutations
72 accruing at relatively similar rates. Taking synonymous mutations as a baseline for
73 neutral changes, this suggests that the within-lineage evolution is neutral or nearly
74 neutral. However, the evolutionary pattern is punctuated by significant leaps at the
75 points of transition between major lineages (see **Figure 1**). These transitions exhibit a
76 much higher rate of accumulation of nonsynonymous mutations compared to
77 synonymous ones (grey line in **Figure 1a-c**), indicating bursts of adaptive evolution that
78 distinguishes one major lineage from another ^{15,16}.

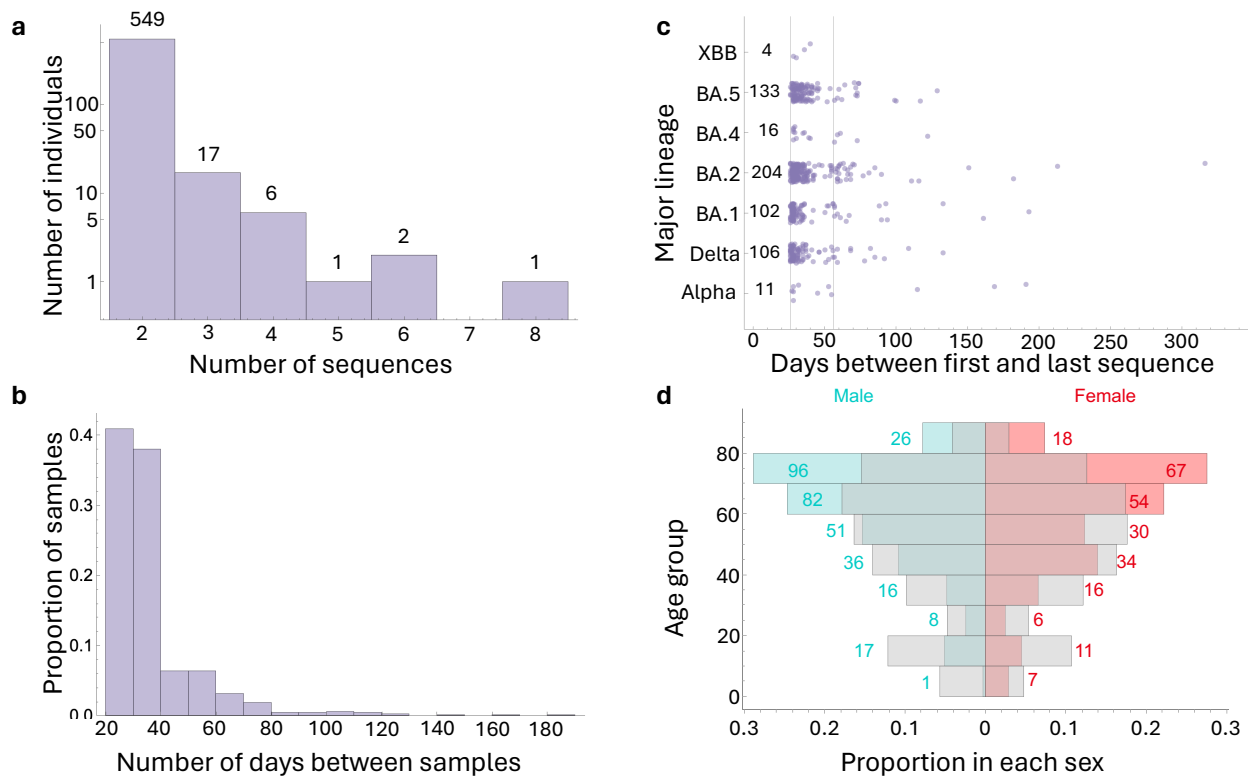
79
80 It has long been hypothesised that this saltatory pattern of SARS-CoV-2 evolution at the
81 between-host level comes from prolonged SARS-CoV-2 infections in
82 immunocompromised individuals, where the virus has extended time to adapt and
83 accumulate advantageous mutations without undergoing tight transmission bottlenecks,
84 followed by the onward transmission of the highly divergent virus to the rest of the
85 population ¹. We set out to investigate whether viral evolution in long infections is
86 consistent with this hypothesis by analysing the evolutionary dynamics of SARS-CoV-2
87 in 576 persistently infected individuals identified as part of the ONS-CIS.

88
89 Persistent infections are more frequent in older individuals and males

90
91 We defined persistent infections as those with at least two RT-PCR positive samples
92 with a high viral RNA titre (cycle threshold values ≤ 30), collected at time intervals of at
93 least 26 days apart, and representing the same infection. We previously identified 381
94 persistent infections within the ONS-CIS using samples collected between 2 Nov 2020
95 to 15 August 2022 ⁸. For the current analysis, we extended this dataset to samples
96 collected up to 21 March 2023, and so covering the entire duration of ONS-CIS before it
97 was paused ², and thereby identifying an additional 195 persistent infections (see
98 **Methods**).

99
100 In total, our dataset comprised 576 cases of persistent SARS-CoV-2 infections,
101 including 11 infections with B.1.1.7 (referred to as Alpha), 106 B.1.617.2 (referred to as
102 Delta), 102 with BA.1, 204 with BA.2, 16 with BA.4, 133 with BA.5, and 4 with XBB major
103 lineages. All persistent infections had viral sequencing data from at least two time
104 points; 27 had sequencing data from three or more time points, typically collected at 20-
105 to 40-day intervals, and the longest-lasting persistent infection spanned nearly a year
106 with eight sequenced time points (**Figure 2a-c**).

107



108 **Figure 2: Baseline characteristics of persistent SARS-CoV-2 infections.** (a) Number of
 109 sequences per persistent infection. Numbers on each bar show the number of persistent
 110 infections per category. (b) Distribution of numbers of elapsed days between consecutive
 111 sequences collected per persistent infection. In cases where a persistent infection has multiple
 112 samples, each pair of consecutive samples is considered. (c) Number of days between the
 113 earliest and latest genomic samples for each persistent infection, with each point representing a
 114 persistent infection. Solid vertical lines are drawn at the 26- and 56-day marks to denote the
 115 thresholds for persistent infections lasting at least one month and two months, respectively.
 116 Numbers on the side of each bar shows the total number of persistent infections per major
 117 lineage. (d) Proportion of persistent infections in each sex and per age-group. Numbers on each
 118 bar show the raw number of persistent infections in each age-group. Grey bars on either side
 119 show the relative proportion of infections with a single positive PCR within the ONS COVID
 120 Infection Survey per sex and age group.

122
 123 Compared to individuals with a single positive PCR test within the ONS-CIS (hereafter
 124 referred to as non-persistent infections), persistently infected individuals were more
 125 prevalent in the above 60 age groups (X-squared = 8.98, df = 1, p-value = 0.00273; see
 126 also **Figure 2d**). We also found a significant association between sex and type of
 127 infection (X-squared = 21.28, df = 1, p-value = 3.97×10^{-6}), with males representing
 128 57.8% of persistently infected cases compared to 48.1% of non-persistently infected
 129 cases. Although we lacked specific information about the underlying health conditions of
 130 participants, the age and sex profile of individuals with persistent infections closely
 131 mirrors the demographic characteristics of individuals diagnosed with Type 1 and Type
 132 2 diabetes in England ¹⁷.

133
 134
 135
 136

137 Nucleotide diversity increases during infection

138

139 We began investigating the within-host evolutionary dynamics of the virus in these 576
140 individuals by first identifying intra-host single nucleotide variants (iSNVs) for each
141 sample collected during infection, and measuring nucleotide diversity, π , over time (see
142 **Methods**). An iSNV was called at a given genomic position if there was a minimum read
143 depth of 10 at that position and a minor allele present at frequency of 20-50%. Positions
144 where the majority of reads were gaps, and those where observed iSNVs are unlikely to
145 represent genuine within-host diversity, were excluded from the analysis (see
146 **Methods**).

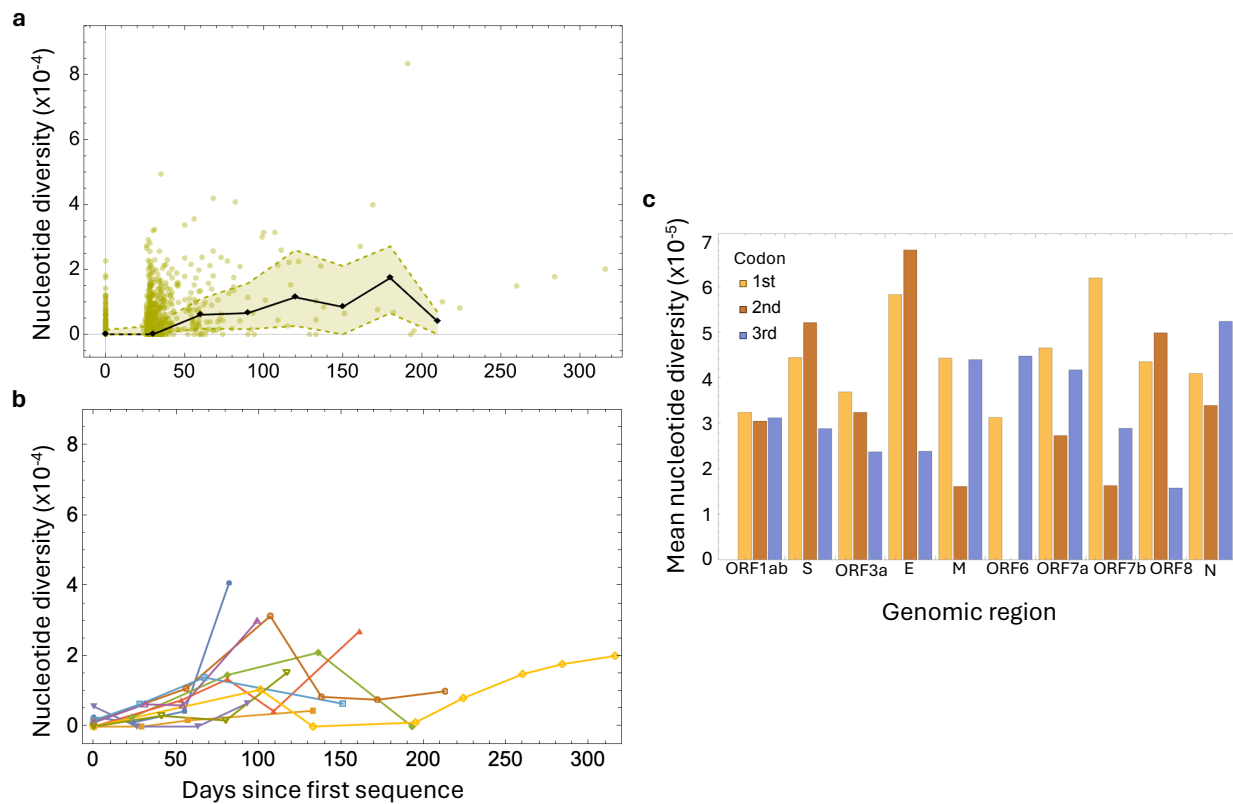
147

148 In the great majority of cases, nucleotide diversity at the earliest time point for each
149 persistent infection was very low, with more than 61% (355/576) of infections displaying
150 no detectable diversity (**Figure 3a**). This suggests that the first sample in most
151 persistent infections was collected near the onset of infection, and with infection initiated
152 by a single, or very closely related, variants^{18,19}. The average within-host diversity (π) of
153 all sampling time points was approximately 4×10^{-5} per nucleotide which is more than an
154 order of magnitude smaller than the between-host diversity at approximately 5×10^{-4} per
155 nucleotide². As might be expected, this indicates that samples collected from the same
156 infection have much lower diversity than samples collected independently from different
157 individuals²⁰. Despite significant variation in diversity over time across different
158 infections (**Figure 3b**), genetic diversity tended to increase until approximately 100 days
159 after the first time point, at which point it either declined or began to plateau in most
160 cases. This pattern suggests that iSNVs appearing late in the infection do not
161 significantly contribute to the overall nucleotide diversity. This could be because they
162 reach mutation-selection balance, remain at low frequency due to their deleterious
163 fitness effects, or rapidly increase in frequency and become fixed. A similar pattern has
164 also been observed during the within-host evolution of HIV²¹.

165

166 We also measured nucleotide diversity by codon position. The first and second codon
167 positions typically induce nonsynonymous changes, while most mutations in third
168 position result in synonymous changes²². Looking at the first and second position
169 across different genomic regions within our samples from persistent infections, the
170 lowest nucleotide diversity was in open reading frame 6 (*ORF6*), with no diversity at the
171 second position, indicating this genomic region is highly conserved and likely subject to
172 strong purifying selection. Conversely, the *Envelope* (*E*) gene exhibited the highest
173 diversity at the first two codon positions, followed by *Spike* (*S*) and *ORF8* (**Figure 3c**).
174 Some of the other genomic regions such as *ORF1ab* had a more uniform diversity
175 across all three codon positions while *ORF6* and *Nucleocapsid* (*N*) had higher
176 synonymous diversity compared to nonsynonymous diversity across all genomic
177 regions.

178



179
180
181
182
183
184
185
186
187
188

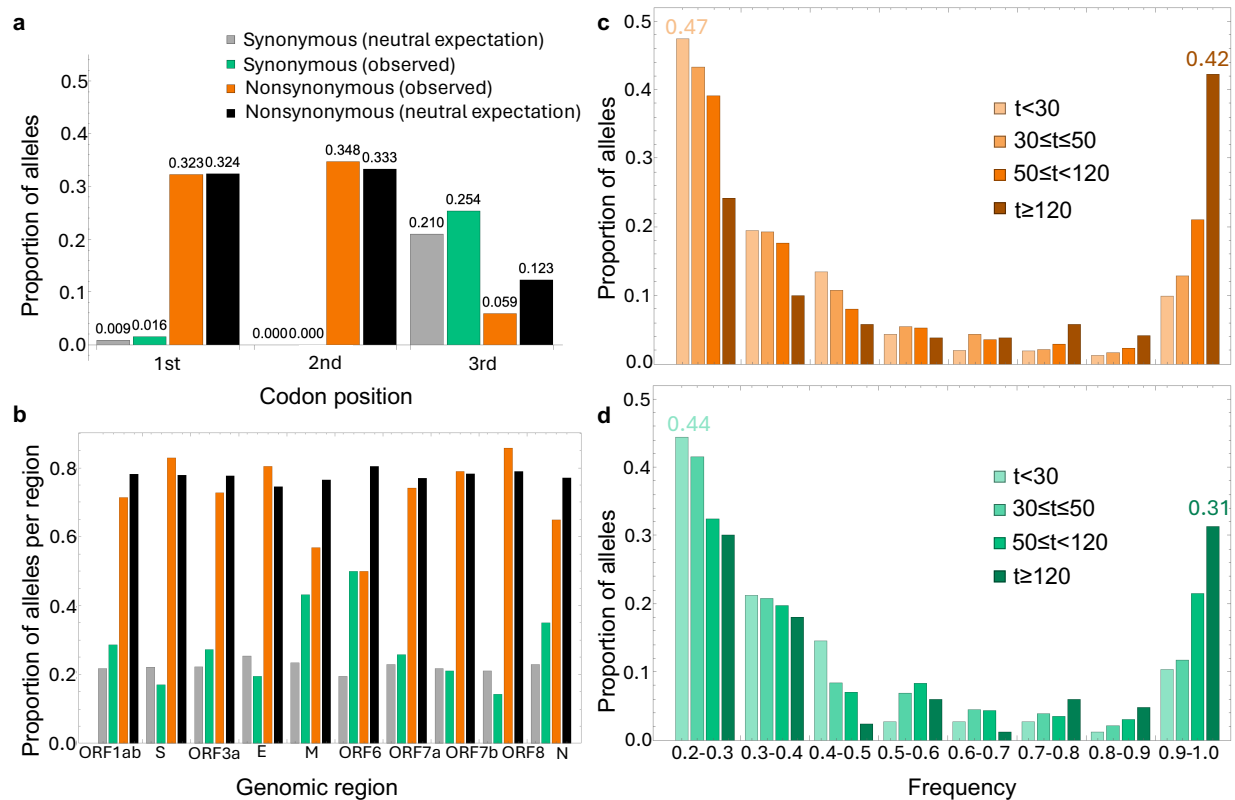
Figure 3: Within-host nucleotide diversity. (a) Aggregate nucleotide diversity (π) over time across all persistent infections. Each data point represents the diversity of a sample from a persistent infection at a given time since the first sequenced sample in that infection ($t=0$). The black line shows the median nucleotide diversity in 30-day intervals and the shaded area covers the interquartile range. (b) Nucleotide diversity over time for persistent infections with three or more samples. (c) Mean nucleotide diversity per codon position in each genomic region including the Open Reading Frames (ORFs), *Spike* (S), *Envelope* (E), *Membrane* (M), and *Nucleocapsid* (N).

189 Higher prevalence of nonsynonymous mutations later in infection

190

191 Next, we identified synonymous and nonsynonymous mutations present at 20%
192 frequency or above at any time point over the course of infection, taking the majority
193 allele at the first time point as reference (see **Methods**). Nearly 67% of all mutant alleles
194 and 73% of those within the coding region were nonsynonymous, with less than 2%
195 synonymous at the first and second codon positions (**Figure 4a**). *ORF6*, *Membrane* (M),
196 and *N* had the highest proportion of synonymous compared to nonsynonymous
197 mutations, and *ORF8* the lowest (**Figure 4b**). Comparing the allele frequency of
198 mutations at different points during infections, towards the start of infections (less than
199 120 days since the first sampled time point), both nonsynonymous and synonymous
200 alleles were typically at comparable frequencies, predominantly below 50% (**Figures**
201 **4c,d**). However, later on nonsynonymous alleles tended to be at higher frequencies,
202 likely indicative of positive selection (see **Figure 4c,d**).

203



204
 205 **Figure 4: Basic characteristics of mutant alleles.** (a) Proportion of synonymous (green) and
 206 nonsynonymous (orange) mutant alleles per codon position observed in samples from
 207 persistent infections, taking the majority allele at the first time point as reference, compared to
 208 expectations under neutrality, taking NC_045512.2 as reference. (b) Proportion of alleles per
 209 mutation type for each genomic region including the Open Reading Frames (ORFs), *Spike* (S),
 210 *Envelope* (E), *Membrane* (M), and *Nucleocapsid* (N). (c) Proportion of synonymous and (d)
 211 nonsynonymous alleles over time across different frequency bands. The proportions of alleles
 212 within the smallest and largest frequency bands are highlighted for both early ($t < 30$) and late
 213 ($t \geq 120$) stages of infection.

214
 215 Nonsynonymous alleles were two to three times more prevalent than synonymous ones
 216 across all frequency bands (see **Supplementary Figure 1**), with about 73% of mutants
 217 in the coding region that exceeded 50% frequency being nonsynonymous. This ratio is
 218 close to the expectation under neutrality, with 78% of all possible mutations across the
 219 genome expected to be nonsynonymous²² (see **Figure 4a,b**). Given it has previously
 220 been found that half of the mutations causing nonsynonymous changes are purged both
 221 at the between-host level and during acute infections ($dN/dS \approx 0.5$)^{18,23}, observing a
 222 ratio of nonsynonymous mutations that is similar to the neutral expectation in
 223 persistently infected individuals suggests that at least some genomic regions are under
 224 positive selection.

225
 226 Variation in evolutionary rates among infections is driven by nonsynonymous changes

227
 228 To determine the within-host evolutionary rates for each infection, we used changes in
 229 allele frequency relative to first sequenced time point (hereafter referred to as the
 230 baseline) as a proxy for measuring evolutionary distance over time (see **Methods**).

231 Within this framework, a full sweep of a mutant allele (a frequency change of 100%)
232 contributes 1 unit of distance and a partial sweep with a frequency change of 40%
233 contributes 0.4 units. This definition of evolutionary distance does not invoke any
234 assumptions about the founder population, which might differ from the population at
235 baseline, as it relies on absolute changes in allele frequencies to measure evolutionary
236 distance ²⁴.

237
238 Allele frequencies change over the course of infection both as a result of sampling noise
239 and actual evolution. To assess the impact of sampling noise on the variation in allele
240 frequencies over time, we required that at least one allele be present at a frequency of
241 $\geq 20\%$ at at least one time point per persistent infection. In other words, if no allele
242 meets this threshold in any sample from a persistently infected individual, we will not
243 (incorrectly) assume there is zero noise (or evolution) due to insufficient data. We also
244 limited our analysis to samples with sufficient sequencing coverage to ensure unbiased
245 estimates of genetic distance per site and, therefore, excluded samples where the
246 number of overlapping base pairs between the consensus sequence at the baseline and
247 the consensus sequence of the sample was less than half the length of the genome.
248 Approximately 14% (82/576) of persistent infections did not meet these criteria and
249 were excluded. We categorised the genetic distances as either synonymous or
250 nonsynonymous, depending on whether the mutant alleles induced a synonymous or
251 nonsynonymous change to the consensus sequence at the baseline for each persistent
252 infection.

253
254 To determine within-host evolutionary rates we used linear regression models, with the
255 slope of the regression line representing the rate of evolution and the y-intercept the
256 level of background noise in the data. The non-zero y-intercept could be attributed to
257 sampling noise and/or residual population structure at baseline ²⁴. To determine the
258 most appropriate model for measuring within-host evolutionary rates, we compared
259 several linear regression models with varying levels of complexity based on their
260 Bayesian Information Criterion (BIC) values. This comparison included a null model
261 which assumed a single fixed slope and y-intercept for all persistent infections (see
262 **Methods**).

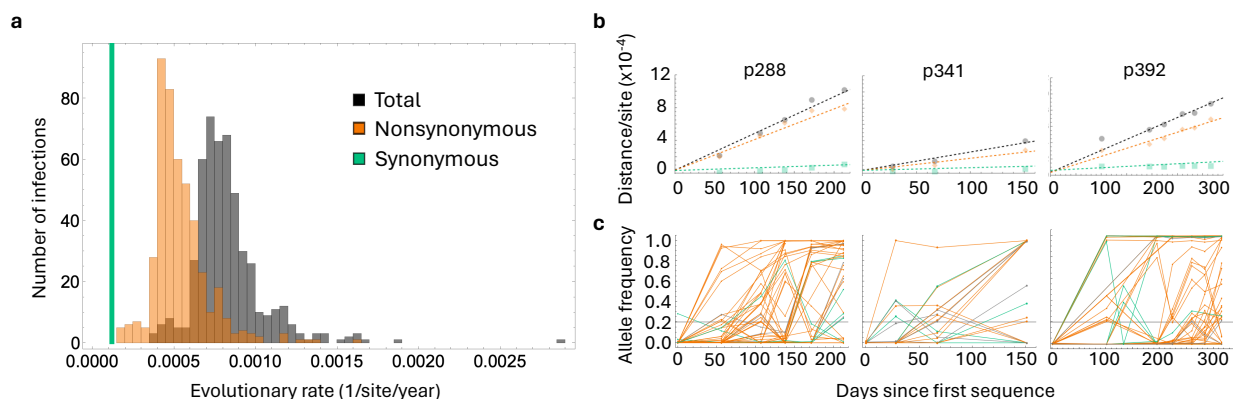
263
264 For genome-wide and nonsynonymous genetic distances, a linear mixed-effect model
265 which assigned a unique evolutionary rate to each persistent infection, but a fixed y-
266 intercept for all infections, gave the best fit (**Supplementary Table 1**). For synonymous
267 distances, a model with a single rate for all infections, but a random y-intercept for each
268 infection, was the most appropriate model. This suggests there was considerable
269 variation in the rate of evolution among individuals, predominantly influenced by
270 nonsynonymous changes, and no strong evidence supporting variation in rate of
271 synonymous evolution across individuals. We also confirmed that the level of noise in
272 allele frequencies is not associated with different sequencing centres (see **Methods** and
273 **Supplementary Table 1**).

274

275 The median genome-wide evolutionary rate was 7.9×10^{-4} substitutions per site per year
276 (s/s/y) with an interquartile range (IQR) of 7.0 - 9.0×10^{-4} s/s/y (**Figure 5a**). Almost 95%
277 (469/494) of persistent infections exhibited an evolutionary rate exceeding 5.5×10^{-4}
278 s/s/y, indicating that the vast majority of individuals experienced a rate surpassing the
279 between-host within-lineage evolutionary rate of SARS-CoV-2 which typically ranges
280 from 2.5 to 5.0×10^{-4} s/s/y for the Alpha, Delta, and Omicron sublineages (see **Figure 1**).
281 Furthermore, 23% (114/494) of the infections had an evolutionary rate higher than the
282 between-lineage rate of 1×10^{-3} s/s/y. The rate of nonsynonymous evolution was 5.0×10^{-4}
283 (IQR: 4.4 - 6.1×10^{-4}) s/s/y, which was about four times higher than the synonymous rate
284 of 1.2×10^{-4} s/s/y across most persistent infections (see **Figure 5**).

285
286 The considerably higher rate of nonsynonymous evolution indicates at least some
287 nonsynonymous mutations are subject to positive selection, and moreover that this
288 selective pressure differs among individuals. In contrast, the preference for a regression
289 model with a single rate for synonymous mutations implies that these mutations are
290 evolutionarily neutral or nearly neutral, evolving at approximately the same rate across
291 all individuals.

292



293
294 **Figure 5: Rates of genome-wide, nonsynonymous, and synonymous evolution in**
295 **persistently infected individuals. (a)** Distribution of inferred evolutionary rates per individual,
296 based on analyses using a linear mixed-effects model optimised for the best fit to the data (as
297 indicated by the lowest BIC value). The model differentiates between unique genome-wide
298 (black) and nonsynonymous (orange) rates for each individual, while applying a single
299 synonymous rate (green) across all individuals. **(b)** Illustrates the evolutionary distance over
300 time for three selected persistently infected individuals – see Supplementary Figure 7 for all 576
301 persistent infections. Points on the graphs represent the total genetic distance from the
302 consensus sequence at the initial time point, calculated based on allele frequency changes over
303 time. Dashed lines indicate the regression lines that best fit these data. **(c)** Shows the mutant
304 allele frequency trajectories for the three persistent infections examined, categorised into
305 synonymous, nonsynonymous, and non-coding (grey) mutations – see Supplementary Figure 2
306 for trajectories in all individuals with measurable evolution in at least 3 time points. Each
307 mutation that reached a minimum frequency of 20% at least at one time point is shown. We can
308 see partial and full sweeps of de novo mutations over the course of persistent infections. A
309 horizontal grey line across the graphs marks the 20% allele frequency threshold.

310

311 To assess how well our model choices fit the data, we further examined the 13
312 persistent infections with three or more sequenced samples, and that included at least

313 one measurement of genetic distance for both synonymous and nonsynonymous
314 mutations (**Supplementary Figure 2**). The best fit regression lines captured most of the
315 changes in the genetic distances over time, with nonsynonymous mutations occurring
316 more frequently, and reaching higher frequencies, than synonymous ones
317 (**Supplementary Figures 2a,b**; see also **Figures 5b,c**). We typically observed two
318 distinct patterns in allele frequencies across different persistent infections. In some
319 cases, transient alleles emerged together at one time point, before disappearing at the
320 later time points, suggesting we are capturing distinct subpopulations within infections.
321 In other cases, we observed the near complete sweep of mutations from low to high
322 frequencies. Other cases were largely a combination of both patterns with some
323 mutations appearing and disappearing in groups while others were present throughout
324 most of the infection (see **Supplementary Figure 3**).

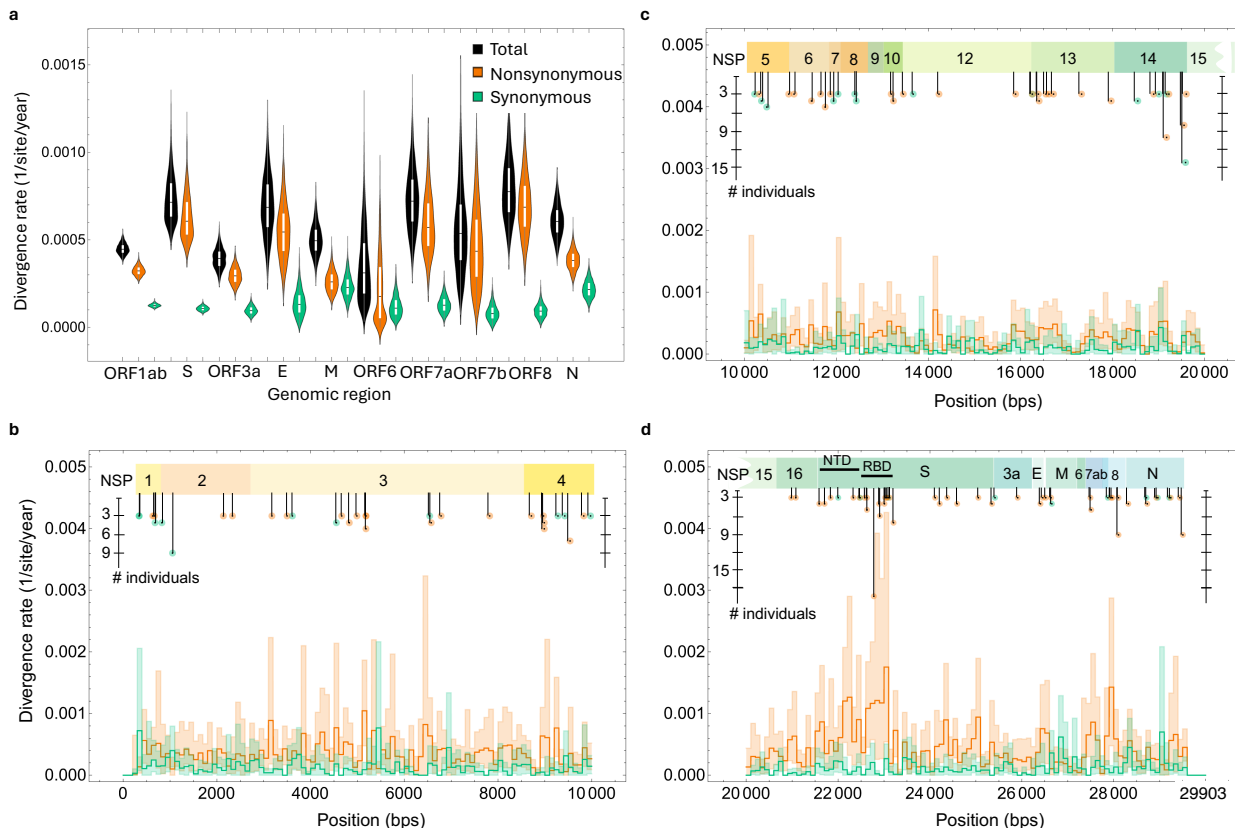
325 The nonsynonymous divergence rate is highest in the receptor binding domain

327
328 To explore evolutionary rate variation across the genome, we next assumed the
329 consensus sequence at baseline represents the founder virus, and that the start of
330 infection occurred at the midpoint between the last negative PCR test and the first
331 sampled time point of the persistent infection. For the majority of infections, the last
332 negative PCR test was taken between 20 to 40 days before the baseline (see
333 **Supplementary Figure 4**). Using the estimated infection start dates, we calculated an
334 evolutionary rate for each region of the genome, aggregating across all individuals (see
335 **Methods**). We called this the divergence rate to distinguish it from the approach we
336 took to measure evolutionary rates per individual, because most infections had only a
337 limited number of mutations, which precluded a calculation of a per-individual rate per
338 gene or gene segment. This commonly used approach to measuring within-host
339 divergence rates comes with two key disadvantages compared to the intra-infection
340 evolutionary rates we measured in the previous section. First, it requires estimating the
341 time elapsed since the start of the infection rather than using only known sample
342 collection dates. Second, this method has a tendency to ascribe any changes in allele
343 frequencies, or their absence, to substitution rates rather than to sampling noise.

344
345 We observed considerable variability in the rate of divergence across the genome
346 (**Figure 6**). The bulk of this rate variation among different genomic regions came from
347 nonsynonymous changes, with the rate of synonymous divergence remaining relatively
348 uniform across most regions, except for the *M* and *N* genes which had a synonymous
349 rate nearly double that of the other regions (**Figure 6a**). *ORF8* and *S* had the highest
350 rates of nonsynonymous divergence, nearly five times greater than the rates of
351 synonymous divergence, whereas *ORF6* showed the lowest rate of nonsynonymous
352 divergence, further indicating it is likely under strong purifying selection.

353
354 Looking at divergence rates across non-overlapping gene segments of 100 base pairs
355 in length, most segments in *ORF1ab* and *S*, which together make up approximately
356 85% of the SARS-CoV-2 genome, displayed low levels of variation in synonymous

357 divergence rates, while nonsynonymous rates varied up to 5 times in some segments of
358 *ORF1ab*, and 10 times in *S* (**Figure 6b-d**). The end tail of the RBD in *S* (22990 to
359 23090) had the highest rates of nonsynonymous divergence, suggesting that it is under
360 strongest positive selection (**Figure 6d**). Accelerated nonsynonymous evolution in the
361 NTD and RBD during persistent infections supports the idea that these infections are
362 the main source behind the emergence of highly divergent variants at the population
363 level. New major lineages that successfully spread also exhibit an overabundance of
364 nonsynonymous mutations in the same genomic regions compared to other circulating
365 lineages at the time of their emergence in the population.
366



367
368 **Figure 6: Virus divergence rates across the genome.** (a) This panel presents the estimated
369 divergence rates from the putative founder, showcasing genome-wide (black), nonsynonymous
370 (orange), and synonymous (green) substitution rates across different regions. The distributions
371 represent the bootstrap estimates derived from 576 persistent infections. (b-d) Display the
372 estimated divergence rate per 100 (nonoverlapping) base pair segments of the genome for non-
373 structural proteins (NSPs): NSP 1 to 4 in (b), NSP 5 to 15 in (c), and NSP 15 and 16, along with
374 other structural non-structural proteins and accessory factors in (d). Shaded area represents the
375 95% confidence intervals from bootstrapping. Recurrent mutations identified in three or more
376 persistent infections are highlighted.
377

378 Recurrent within-host mutations with transient fitness advantage

379
380 We found 379 (262 nonsynonymous and 117 synonymous) mutations found in at least
381 two individuals among the 576 persistent infections (**Source file**; see also **Figure 5b-d**).
382 The highest concentration of these recurrent mutations that were nonsynonymous were
383 in *ORF8* (24 mutations), *E* (14 mutations), and *S* (210 mutations), whereas the highest

384 concentration of recurrent synonymous mutations was in *ORF7b* (3 mutations) and *M*
385 (14 mutations).

386
387 The per-lineage fitness effect of recurrent mutations was measured at the between-host
388 level using a globally representative SARS-CoV-2 phylogeny²⁵. When fitness effects
389 were examined within the same major lineage as the virus from persistent infections,
390 54% of these mutations showed a positive fitness effect (**Supplementary Figure 5a-c**).
391 Most recurrent mutations also had very low population-prevalence with nearly 47%
392 being present in less than 0.01% of all samples within ONS-CIS sequences from the
393 same major lineage as the virus from persistent infections (**Supplementary Figure 5d**).
394 This suggests that almost half of the recurrent mutations have a fitness advantage at
395 the within-host level but a fitness disadvantage and low prevalence at the between-host
396 level.

397
398 The most recurrent mutations were S:N405D (with corresponding nucleotide
399 substitution A22775G in 8 infections), NSP14: T516T (T19587A, in 13 infections), and
400 NSP14:C382G (T19183G, in 10 infections), all of which were found in persistent
401 Omicron infections, BA.2, BA.4, and BA.5. The highly recurrent *Spike* mutations that
402 were found in at least three persistent infections and had very high between-host fitness
403 effects were S:L452R, S:K356T, and S:T547K all of which are lineage-defining
404 mutations (see **Supplementary Figure 5e**). In particular, S:K356T is lineage-defining
405 for BA.2.86 and was found in multiple BA.2 and BA.5 persistent infections. On the other
406 hand, most of the highly recurrent mutations with strong negative between-host fitness
407 effects were concentrated in various non-structural proteins of *ORF1ab* (see
408 **Supplementary Figure 5e**).

409
410 We also investigated potential associations between host characteristics and recurrent
411 mutations in SARS-CoV-2 persistent infections. Specifically, we examined whether
412 there is an association between the age group of the persistently infected individual and
413 the number of times a mutation recurs (**Supplementary Figure 5f**), the between-host
414 fitness effect of recurrent mutations and the age group of the individual in which they
415 appeared (**Supplementary Figure 5g**), and the fitness effect of the recurrent mutations
416 with respect to the duration of persistent infections (**Supplementary Figure 5h**).
417 However, we found no strong associations between these factors.

418
419 Infection duration is correlated with evolutionary rates

420
421 We found no significant associations ($\Delta\text{BIC} < 0$) of age, sex, vaccination status, prior
422 infection, or virus lineage with within-host evolution rates. This evaluation was based on
423 comparing the BIC values of the best-fit regression model for determining within-host
424 rates with models that included each of these parameters as an additional fixed effect
425 (see **Supplementary Table 2**). Notably, our observation that the within-host
426 evolutionary rates do not significantly differ between vaccinated and unvaccinated
427 individuals suggests that vaccination does not lead to accelerated evolutionary rates.

428 We were also interested in investigating whether experiencing a viral rebound had an
429 impact on evolutionary rates. To do this, we categorised persistent infections into either
430 persistent-chronic (consistently positive PCR tests throughout the infection) or
431 persistent-rebounding (at least one negative PCR test during the infection); see also ref
432 ⁸ for more about these two categories. We found weak evidence ($\Delta\text{BIC}=1$) in support of
433 a positive association between experiencing a rebounding viral load and an elevated
434 nonsynonymous evolutionary rate. After controlling for duration of infection, since it is
435 more likely to identify persistent-rebounding infections when the infections are longer
436 (i.e. more time to pick up a negative PCR test during a prolonged infection), by only
437 examining a subset of infection where the duration of infection is longer than at least 56
438 days, we found no association between viral rebound and higher evolutionary rates
439 ($\Delta\text{BIC} < 0$). However, we did identify a positive association ($\Delta\text{BIC} > 2$) between the
440 evolutionary rates and the duration of infection, indicating that longer infections exhibit
441 higher rates of nonsynonymous evolution. To determine if this association was biased
442 by the lower genetic diversity typically seen in shorter infections, which could result in
443 lower evolutionary rate estimates, we also examined longer infections lasting at least 56
444 days. Our analysis confirmed statistical support ($\Delta\text{BIC} > 2$) for the positive relationship
445 between infection duration and evolutionary rates, even within these subsets of
446 infections (see **Supplementary Table 2**).

447

448 **Discussion**

449

450 We characterised viral genomic diversity and within-host evolutionary rates in 576
451 individuals with persistent SARS-CoV-2 infections, identified through large-scale
452 community surveillance, and including samples collected between November 2020 to
453 March 2023. Central to our investigation was the hypothesis that persistent infections
454 could serve as the primary source for the saltatory evolution of the virus at the between-
455 host level, mirroring the same evolutionary changes we see with the emergence of
456 highly-divergent variants. This premise led us to identify host characteristics associated
457 with prolonged infections and to characterise viral evolutionary patterns across the
458 genome and between individuals.

459

460 We observed significant variability in within-host viral evolutionary rates between
461 infections. This variability was predominantly attributed to the different rates at which
462 individuals accumulated nonsynonymous mutations, with the rate of synonymous
463 mutations being similar among all individuals and typically more than four-fold slower
464 than the rate of nonsynonymous mutations. This variability among individuals explains
465 previous findings of limited consensus change mutations in some individuals and over-
466 abundance of mutations in others ^{8,26}. We also observed considerable variability in
467 nonsynonymous evolutionary rates across most of the genome, but not synonymous
468 rates, with the receptor binding domain of the Spike protein having the highest rate of
469 nonsynonymous evolution relative to all other genomic regions. We also found elevated
470 synonymous rates in *M* and *N* genes which suggest they could have functional benefit

471 for mRNA stability and translation efficacy, particularly on phosphorylation sites that are
472 abundant in *N*⁹.

473
474 Although older individuals were more likely to experience persistent infections, we found
475 no evidence to suggest that host factors such as age, sex, vaccination status, virus
476 lineage, previous infection, or dynamics of viral RNA titres significantly affected
477 evolutionary rates. However, we did observe a positive association between
478 evolutionary rates and the duration of infection, with longer-lasting infections exhibiting
479 higher rates of nonsynonymous evolution. We speculate that individuals with longer
480 infections may have more impaired immune responses, and/or be undergoing
481 treatment, which may result in faster rates of adaptive evolution. Our examination of
482 recurrent within-host mutations which are rare in the general population and have
483 negative between-host fitness effects further illustrates the complex evolutionary
484 dynamics at play within persistent infections. These mutations likely confer a selective
485 advantage within hosts due to enhanced replication rates and/or immune evasion.
486 However, they may prove detrimental at the between-host level, for example if they
487 result in reduced transmissibility of the virus between individuals^{11,27,28}.

488
489 We found that *ORF6* had the lowest levels of nonsynonymous diversity and divergence
490 rate compared to the other genomic regions, indicating it is functionally conserved
491 during persistent infections. Strikingly, we found no diversity in the second codon
492 position of *ORF6*; all mutations at this position would be nonsynonymous. These
493 observations are consistent with several studies that have highlighted the crucial role of
494 *ORF6* in viral replication and disease progression^{29–31}. These results suggest that
495 *ORF6* could be a promising candidate for the development of therapeutic drugs for
496 treating individuals with persistent infections³².

497
498 Many of the recurrent mutations identified in our study have been found to have
499 functional importance for SARS-CoV-2. For example, the mutation S:G446V is linked to
500 treatment resistance³³. The mutation NSP3:T820I frequently occurs in patients treated
501 with Nirmatrelvir and Ritonavir³⁴, while NSP7:L3935L is commonly found in cancer
502 patients and those undergoing immunosuppressive or steroid therapies³⁵. Another
503 mutation, S:D1153Y, is known for its antibody escape properties³⁶. The mutation
504 M:N117K may play a role in the glycosylation of the virus³⁷. Also, recurrent mutations
505 S:L216F, S:S98F, and N:P151S have previously been identified as being under
506 multilevel selection, beneficial at the within-host level but deleterious at the between-
507 host level¹¹.

508
509 Our findings shed light on the complex interplay between persistent SARS-CoV-2
510 infections, the demographic characteristics of those infected, and the evolutionary
511 mechanisms driving the virus evolution within these individuals. This study also
512 underscores how persistent infections may contribute to the emergence of highly
513 divergent variants, with factors such as the duration of infection and accelerated rate of

514 evolution at nonsynonymous sites, particularly in the RBD of Spike protein, influencing
515 their evolutionary rates.

516

517 **Methods**

518

519 ONS COVID-19 Infection Survey

520

521 This work contains statistical data from ONS which is Crown Copyright. The use of the
522 ONS statistical data in this work does not imply the endorsement of the ONS in relation
523 to the interpretation or analysis of the statistical data. This work uses research datasets
524 which may not exactly reproduce National Statistics aggregates.

525

526 The Office for National Statistics Covid-19 Infection Survey (ONS-CIS) is a UK
527 household-based surveillance study, which began in the UK from April 2020³⁸ and was
528 first paused in March 2023². Our analysis here covered the period from 2 Nov 2020 to
529 21 March 2023. Households from nationwide address lists were invited to participate
530 (every household member aged two years and above), ensuring as representative a
531 cross-section of the population as possible. Participants gave written informed consent
532 to contribute swab samples (self-collected or by a parent/carer for those under 12
533 years), irrespective of symptoms, and completed a questionnaire for each assessment.

534

535 Most of the participants in the survey consented to routine PCR sampling at weekly
536 intervals for the first month of enrollment and monthly thereafter for the duration of the
537 study^{2,8}. From December 2020, all cases where a participant tested positive with a high
538 viral load (Ct \leq 30), their sample was further sent for sequencing.

539

540 Sequencing

541

542 Samples were sequenced at one of five sequencing centres, University of Oxford
543 (OXON), Northumbria University and associated NHS foundation trusts (NORT),
544 National Infection Service Public Health England (PHEC), Quadram Institute
545 Bioscience, Norwich (NORW), and Wellcome Sanger Institute (Sanger). The great
546 majority of samples were sequenced on Illumina Novaseq, with the rest using Oxford
547 Nanopore GridION or MINION. The standard consensus FASTA sequences for all
548 ONS-CIS samples were generated using the ARTIC Nextflow processing pipeline (v1)
549³⁹, or veSeq, an RNA sequencing protocol based on a quantitative targeted enrichment
550 strategy^{2,40} with consensus sequences produced using Shiver (v1.5.8)⁴¹. For additional
551 information about the survey, sequencing protocol, and FASTA consensus sequence
552 protocol see^{2,8}.

553

554

555

556

557 Identification of persistent infections

558
559 We used the consensus sequences generated using ARTIC Nextflow or Shiver to
560 determine whether two or more sequences from the same individual were from the
561 same infection, using the method outlined in ⁸. Briefly, if two sequences from the same
562 individual were collected at least 26 days apart, were of the same major lineage, and
563 shared a rare single nucleotide polymorphism (SNP) compared to the population-level
564 consensus, the individual was determined to be persistently infected. Our analysis
565 covered infections with the Alpha, Delta, Omicron BA.1, BA.2, BA.4, BA.5, and XBB
566 major lineages, and a SNP was deemed to be rare if found in <400 samples of that
567 lineage (see **Supplementary Figure 6**). Due to possible misclassification of some BA.2
568 sequences as BA.5 and vice versa using the Pango lineage nomenclature ⁴², we
569 considered the possibility that some BA.5 sequences could belong to a BA.2 infection.
570 This approach identified 3 cases of BA.2 persistent infections, which included at least
571 one sequence misclassified as a BA.5 lineage. Without requiring any additional
572 adjustment to separate second-generation BA.2 (e.g. BA.2.75) and BA.5 (e.g. BQ.1)
573 major lineages from their closest ancestors, our method reliably recovered subsets of
574 infections within BA.2 and BA.5 that were attributable to second-generation variants.
575 Specifically, we found 21 BA.2.75 and 25 BQ.1 persistent infections.

576 577 Identifying intra-host single nucleotide variants

578
579 We called an intra-host single nucleotide variant (iSNV) at a given position in the
580 genome if there were 10 or more bases called at that position, including gaps, and if the
581 most common minor allele was present at 20% or more but less than 50%. The small
582 number of bases required to call an iSNV was chosen because many samples had low
583 viral titre, whilst the 20% threshold was to avoid biases introduced by differing amounts
584 of sequencing noise across all the samples.

585
586 We also identified mutations, which we defined as iSNVs or major alleles that differed
587 from the majority allele at the first sampling time point, and reached at least 20%
588 frequency at baseline or any of the subsequent time points. Whereas iSNVs are always
589 less than 50% frequency by definition, a mutation can be at any frequency above 20%
590 (including 100%). To ensure consistency of methods across our analyses, we also
591 defined the majority-rule consensus at each sampling as the majority allele, with a
592 minimum of 10 bases to call a consensus at any given position. Unless stated
593 otherwise, when we refer to the consensus we mean the majority-rule consensus, not
594 the consensus generated using ARTIC Nextflow or Shiver.

595
596 Some positions in the genome are prone to having low frequency iSNVs in a high
597 proportion of samples, and are often sequencing centre specific. Although we do not
598 know what causes these low frequency iSNVs, they are unlikely maintained through
599 descent and we therefore label them 'artefactual iSNVs'. For each sequencing centre in

600 our study, we masked genomic positions where an iSNV was present at $\geq 2\%$ frequency
601 in more than 1% of samples from that sequencing centre.

602

603 Nucleotide diversity

604

605 Nucleotide diversity was calculated using the π statistic, which is the common measure
606 of diversity least affected by the number of sequences used in the analysis⁴³. For each
607 persistent infection, nucleotide diversity at a given time point is given by:

608

$$\pi = \frac{1}{L} \sum_{\ell=1}^L D_{\ell}$$

609

610

611 where L represents the number of nucleotide positions being examined, and D_{ℓ} the
612 genetic diversity at locus ℓ with an iSNV present at a frequency $\geq 20\%$. This is
613 calculated as:

614

$$D_{\ell} = \frac{2}{N(N-1)} \sum_{i \neq j} n_i n_j$$

615

616

617 where n_i represents the number of nucleotides $i = A, C, G$ or T (not including gaps), and
618 N the total number of reads at that locus.

619

620 Estimating within-host genetic distance

621

622 We used differences in mutant allele frequencies between two sequences from the
623 same infection to calculate the genetic distance between the sequences. This is similar
624 to an approach that has been used to measure within-host evolutionary rates of
625 influenza A in a chronically infected individual²⁴. We calculated changes in allele
626 frequency relative to the first sequenced time point in each persistent infection.
627 Synonymous and nonsynonymous distance was determined by whether the mutant
628 allele would result in the same (synonymous) or a different (nonsynonymous) amino
629 acid being coded for compared to the first time point in the infection.

630

631 Following this definition of evolutionary distance, a mutant allele i , present at frequency
632 $f_i(t_0)$ at the first time point and $f_i(t_k)$ at the k th time point contributes $|f_i(t_k) - f_i(t_0)|$ to the
633 pairwise distance between the two sequences. More generally, if the pair of sequences
634 has M mutant alleles, the total genetic distance between them is

635

$$636 \quad d(t_k, t_0) = \sum_{i \in M} |f_i(t_k) - f_i(t_0)| ,$$

637

638 where $| \cdot |$ represents the absolute change of allele frequency. We excluded pairs of
639 samples where the total number of overlapping base pairs between the two consensus
640 sequences is smaller than 50% of genome length as these can give rise to deflated or
641 inflated measures of genetic distance per site.

642

643 Estimating within-host evolutionary rate

644

645 We quantified within-host evolutionary rates by assuming a linear relationship between
646 the genetic distance and the time elapsed since the first sequence was collected from
647 each individual.

648

649 A linear regression model represented the changes in genetic distance relative to first
650 sequence over time within each persistent infection as

651

$$652 \quad d(t_k, t_0) \approx r |t_k - t_0| + e ,$$

653

654 where r is the evolutionary rate and e is the y-intercept, which represents the expected
655 amount of noise when measuring genetic distance. The noise could arise from either
656 sequencing error or undiagnosed population structure²⁴. If a persistent infection does
657 not have a detectable mutant allele that reaches frequency $\geq 20\%$, we exclude that
658 individual from evolutionary rate analysis as we cannot quantify the contribution of noise
659 in frequency change of alleles.

660

661 Our analysis encompassed five different regression models with varied levels of
662 complexity (see **Supplementary Table 1**) to estimate genome-wide, synonymous, and
663 nonsynonymous within-host evolutionary rates. We used the Bayesian Information
664 Criterion (BIC) value for model selection, balancing model complexity against fit quality.

665

666 The y-intercept can be interpreted as the baseline level of noise in changes of allele
667 frequencies. With a fixed nonsynonymous y-intercept at 3.4×10^{-5} substitutions per site
668 and an average of 4.5 nonsynonymous mutations per infection, we can estimate that
669 roughly 23% of the variations in nonsynonymous allele frequencies may be attributed to
670 noise. Conversely, for a typical synonymous mutation characterised by a y-intercept of
671 2.2×10^{-5} substitutions per site and an average of 1.6 synonymous mutations per
672 infection, about 40% of changes in allele frequencies are driven by noise. While we
673 expect the contribution of sampling noise to be the same for both synonymous and
674 nonsynonymous mutations, biological factors such as selection and functional
675 constraints may not be uniform across different mutation types. More specifically, given
676 that synonymous mutations are more likely to be neutral or nearly neutral, their baseline
677 noise can be more reflective of sampling noise and the stochastic nature of viral
678 replication and mutation.

679

680 We examined the following linear regression models for measuring evolutionary rates
681 and baseline noise:

682

683 (i) Complete pooling: $d_i(t) = r_0 t + e_0 + \varepsilon_i(t)$

684

685 This model assumes a single (fixed) underlying rate, denoted as r_0 , and intercept, e_0 ,
686 which describes a common evolutionary rate and noise contribution across all
687 individuals. The error term $\varepsilon_i(t)$ represents the residual unexplained variability in
688 distance, $d_i(t)$ for persistent infection i .

689

690 Models (ii) to (v) all incorporate partial pooling with varying degrees of complexity.

691

692 (ii) Random intercept: $d_i(t) = r_0 t + e_i + e_0 + \varepsilon_i(t)$

693

694 A linear mixed effect model which assumes a shared rate, r_0 , and error, e_0 , across all
695 infections (fixed effects) with each infection i also having a unique intercept e_i , indicative
696 of individual-level noise variation (random effect).

697

698 (iii) Random slope with one fixed intercept: $d_i(t) = (r_0 + r_i) t + e_0 + \varepsilon_i(t)$

699

700 A linear mixed effect model which assumes a single (fixed) underlying rate, r_0 , and error,
701 e_0 , shared by all individuals in addition to a unique underlying rate, r_i , for each persistent
702 infection, i (random effect).

703

704 (iv) Random slope with multiple fixed intercepts: $d_i(t) = (r_0 + r_i) t + \sum_j e_j + \varepsilon_i(t)$

705

706 Considering potential sequencing centre-specific noise, we categorised y-intercepts into
707 nine groups, based on where the sequences were sampled. For instance, if the initial
708 sample from a persistently infected individual was sequenced in Sanger Institute
709 ("Sanger") and a subsequent sample in the University of Oxford ("OXON"), the y-
710 intercept corresponding to this persistent infection belong to the j =("Sanger", "OXON")
711 category. There are a total of nine such y-intercept categories, represented as
712 $j \in \{(\text{NORT}, \text{PHEC}), (\text{NORT}, \text{NORW}), (\text{NORT}, \text{Sanger}), (\text{OXON}, \text{PHEC}), (\text{Sanger},$
713 $\text{OXON}), (\text{NORT}), (\text{PHEC}), (\text{OXON}), (\text{Sanger})\}$. There are 9 pairs of samples that are
714 (NORT, PHEC), 4 (NORT, NORW), 90 (NORT, Sanger), 14 (OXON, PHEC), 1 (Sanger,
715 OXON), 147 (NORT), 16 (PHEC), 10 (OXON), and 331 (Sanger). We assessed these
716 categories for their impact on baseline noise in the data, assuming their influence is
717 constant over time. This model therefore introduces nine fixed effects e_j to account for
718 variations in y-intercepts due to sequencing noise levels.

719

720 (v) No pooling: $d_i(t) = r_i t + e_i$

721

722 Each persistent infection, denoted as i , has a unique rate and error term. In practice,
723 this model cannot be applied to our dataset because the number of measurements is
724 smaller than the number of random effects, as persistent infections with only two
725 samples yield a single measurement for genetic distance.

726
727 Our analysis showed, based on the lowest BIC value, that the random slope with one
728 fixed intercept regression model (iii) best explains genome-wide and nonsynonymous
729 evolutionary rates while the random intercept regression model (ii) best explains
730 synonymous rate for persistent infections. The lines of best fit for all the persistent
731 infections with measurable evolution is shown in **Supplementary Figure 7**.

732

733 Estimating within- and between-lineage rates at the between-host level

734

735 To assess the saltatory evolution of SARS-CoV-2 at the between-host level, we used a
736 previously identified representative sample from the ONS-CIS dataset ². This dataset
737 covered sequences from the Alpha, Delta, Omicron BA.1, BA.2 (excluding BA.2.75),
738 BA.2.75, BA.4, BA.5 (excluding BQ.1), and BQ.1 lineages. We then constructed the
739 ancestral sequence for each major lineage using TreeTime ⁴⁴ and calculated total,
740 nonsynonymous, and synonymous Hamming distances between samples from each
741 major lineage relative to the ancestral sequence of the same major lineage. Finally, to
742 estimate the between-lineage rate, we calculated the total, nonsynonymous, and
743 synonymous Hamming distances between the Wuhan reference sequence
744 (NC_045512.2) and the ancestral sequence for each major lineage.

745

746 Divergence rate from putative founder

747

748 Since persistent infections on average have 5 mutations across the genome (IQR: 2, 8),
749 estimating an evolutionary rate for different segments of the genome at an individual
750 level is not practical. We therefore used the majority-rule consensus sequence at the
751 first time point of each persistent infection as a proxy for the founding virus. We then
752 estimated the start time of infection as the midpoint between the last negative PCR test
753 and the first sequence from the persistent infection. We measured the typical
754 evolutionary rate (rather than mean) from the putative founder across all individuals for
755 each segment of the genome.

756

757 While this method is frequently used for calculating within-host divergence rates for
758 viruses like HIV ⁴⁵, it will miss early fixation events that might have shifted the
759 consensus sequence away from the true founding virus by the time the first sample was
760 collected; assumes the founding viral population was genetically homogeneous ⁴⁶; does
761 not control for noise which could bias estimates of the divergence rate. Nonetheless,
762 aggregating across a large number of individuals should help mitigate these effects.

763

764 This approach involved treating each measurement of divergence from the putative
765 founder at any given time point, t , as an independent observation, regardless of its
766 associated persistent infection. The divergence from the founder for each genomic
767 segment at any time point, including baseline, was defined as the cumulative frequency
768 of all mutant alleles within that segment at time t . For example, if there were no mutant
769 alleles within a genomic segment at a given time point, we recorded a divergence of
770 zero. Subsequently, we used a linear regression with a zero y -intercept at the start time
771 of infection to calculate the divergence rate from the putative founder for each genomic
772 segment. This can be expressed as $d^{(n)}_i(t) = r_i t + \varepsilon^{(n)}_i(t)$, where r_i is the divergence rate
773 for genomic segment i , and $d^{(n)}_i(t)$ is calculated as the genetic divergence of sample n
774 from its putative founder within segment i at time t . Each sample, n , from a persistent
775 infection represents one measurement of $d^{(n)}_i(t)$. If a sample is collected at time $t=t^*$ and
776 has no mutant alleles within segment i , then $d^{(n)}_i(t^*)=0$. For each sample, the estimated
777 start of infection is taken as $t=0$. Each sample from an individual acts as an independent
778 observation of genetic distance for segment i . The error term $\varepsilon^{(n)}_i(t)$ represents the
779 residual unexplained variability in distance, $d^{(n)}_i(t)$, for sample n .

780

781 To ensure an equal representation of each persistent infection in the divergence rate
782 assessment for a genomic segment, we limited our analysis to two divergence
783 measurements per individual—one at the baseline and another selected randomly from
784 later in the infection. We then performed bootstrapping across all individuals and every
785 possible pair of divergence measurements per individual to create a distribution of
786 divergence rate estimates for each genomic segment.

787

788 **Data availability**

789

790 All raw consensus sequences have been made publicly available as part of the COG-
791 UK Consortium

792 (<https://webarchive.nationalarchives.gov.uk/ukgwa/20230505214946/https://www.cogco>
793 [nsortium.uk/priority-areas/data-linkage-analysis/](https://www.cogco.nsortium.uk/priority-areas/data-linkage-analysis/)) and are available from the European
794 Nucleotide Archive at EMBL-EBI under accession number [PRJEB37886](https://www.ebi.ac.uk/ena/record/PRJEB37886).

References

1. Markov, P.V., Ghafari, M., Beer, M., Lythgoe, K., Simmonds, P., Stilianakis, N.I., and Katzourakis, A. (2023). The evolution of SARS-CoV-2. *Nat. Rev. Microbiol.* 21, 361–379.
2. Lythgoe, K.A., Golubchik, T., Hall, M., House, T., Cahuantzi, R., MacIntyre-Cockett, G., Fryer, H., Thomson, L., Nurtay, A., Ghafari, M., et al. (2023). Lineage replacement and evolution captured by 3 years of the United Kingdom Coronavirus (COVID-19) Infection Survey. *Proc. Biol. Sci.* 290. 10.1098/rspb.2023.1284.
3. Roemer, C., Sheward, D.J., Hisner, R., Gueli, F., Sakaguchi, H., Frohberg, N., Schoenmakers, J., Sato, K., O'Toole, Á., Rambaut, A., et al. (2023). SARS-CoV-2 evolution in the Omicron era. *Nat. Microbiol.* 8, 1952–1959.
4. Carabelli, A.M., Peacock, T.P., Thorne, L.G., Harvey, W.T., Hughes, J., de Silva, T.I., Peacock, S.J., Barclay, W.S., de Silva, T.I., Towers, G.J., et al. (2023). SARS-CoV-2 variant biology: immune escape, transmission and fitness. *Nat. Rev. Microbiol.* 21, 162–177.
5. Kistler, K.E., Huddleston, J., and Bedford, T. (2022). Rapid and parallel adaptive mutations in spike S1 drive clade success in SARS-CoV-2. *Cell Host Microbe* 30, 545-555.e4.
6. Harari, S., Tahor, M., Rutsinsky, N., Meijer, S., Miller, D., Henig, O., Halutz, O., Levytskyi, K., Ben-Ami, R., Adler, A., et al. (2022). Drivers of adaptive evolution during chronic SARS-CoV-2 infections. *Nat. Med.* 28, 1501–1508.
7. Wilkinson, S.A.J., Richter, A., Casey, A., Osman, H., Mirza, J.D., Stockton, J., Quick, J., Ratcliffe, L., Sparks, N., Cumley, N., et al. (2022). Recurrent SARS-CoV-2 mutations in immunodeficient patients. *Virus Evol* 8, veac050.
8. Ghafari, M., Hall, M., Golubchik, T., Ayoubkhani, D., House, T., MacIntyre-Cockett, G., Fryer, H.R., Thomson, L., Nurtay, A., Kemp, S.A., et al. (2024). Prevalence of persistent SARS-CoV-2 in a large community surveillance study. *Nature* 626, 1094–1101.
9. Bouhaddou, M., Reuschl, A.-K., Polacco, B.J., Thorne, L.G., Ummadi, M.R., Ye, C., Rosales, R., Pelin, A., Batra, J., Jang, G.M., et al. (2023). SARS-CoV-2 variants evolve convergent strategies to remodel the host response. *Cell* 186, 4597-4614.e26.
10. Chaguza, C., Hahn, A.M., Petrone, M.E., Zhou, S., Ferguson, D., Breban, M.I., Pham, K., Peña-Hernández, M.A., Castaldi, C., Hill, V., et al. (2023). Accelerated SARS-CoV-2 intrahost evolution leading to distinct genotypes during chronic infection. *Cell Rep. Med.* 4, 100943.
11. Bonetti Franceschi, V., and Volz, E. (2024). Phylogenetic signatures reveal multilevel selection and fitness costs in SARS-CoV-2. *Wellcome Open Res.* 9, 85.
12. Li, Y., Choudhary, M.C., Regan, J., Boucau, J., Nathan, A., Speidel, T., Liew, M.Y., Edelstein, G.E., Kawano, Y., Uddin, R., et al. (2024). SARS-CoV-2 viral clearance and evolution varies by type and severity of immunodeficiency. *Sci. Transl. Med.* 16. 10.1126/scitranslmed.adk1599.
13. Gonzalez-Reiche, A.S., Alshammary, H., Schaefer, S., Patel, G., Polanco, J., Carreño, J.M., Amoako, A.A., Rooker, A., Cognigni, C., Floda, D., et al. (2023). Sequential intrahost evolution and onward transmission of SARS-CoV-2 variants. *Nat. Commun.* 14, 1–13.

14. Ghafari, M., Liu, Q., Dhillon, A., Katzourakis, A., and Weissman, D.B. (2022). Investigating the evolutionary origins of the first three SARS-CoV-2 variants of concern. *Front. Virol.* 2. 10.3389/fviro.2022.942555.
15. Tay, J.H., Porter, A.F., Wirth, W., and Duchene, S. (2022). The Emergence of SARS-CoV-2 Variants of Concern Is Driven by Acceleration of the Substitution Rate. *Mol. Biol. Evol.* 39. 10.1093/molbev/msac013.
16. Neher, R.A. (2022). Contributions of adaptation and purifying selection to SARS-CoV-2 evolution. *Virus Evol.* 8, veac113.
17. Public Health England (2016). Diabetes Prevalence Model. <https://assets.publishing.service.gov.uk/media/5a82c07340f0b6230269c82d/Diabetesprevalencemodelbriefing>.
18. Lythgoe, K.A., Hall, M., Ferretti, L., de Cesare, M., MacIntyre-Cockett, G., Trebes, A., Andersson, M., Otecko, N., Wise, E.L., Moore, N., et al. (2021). SARS-CoV-2 within-host diversity and transmission. *Science* 372. 10.1126/science.abg0821.
19. Shi, Y.T., Harris, J.D., Martin, M.A., and Koelle, K. (2024). Transmission bottleneck size estimation from DE Novo viral genetic variation. *Mol. Biol. Evol.* 41, msad286.
20. Álvarez-Herrera, M., Sevilla, J., Ruiz-Rodríguez, P., Vergara, A., Vila, J., Cano-Jiménez, P., González-Candelas, F., Comas, I., and Coscollá, M. (2024). VIPERA: Viral Intra-Patient Evolution Reporting and analysis. *Virus Evol.* 10, veae018.
21. Zanini, F., Brodin, J., Thebo, L., Lanz, C., Bratt, G., Albert, J., and Neher, R.A. (2015). Population genomics of inpatient HIV-1 evolution. *Elife* 4, e11282.
22. Otto, S.P., Day, T., Arino, J., Colijn, C., Dushoff, J., Li, M., Mechai, S., Van Domselaar, G., Wu, J., Earn, D.J.D., et al. (2021). The origins and potential future of SARS-CoV-2 variants of concern in the evolving COVID-19 pandemic. *Curr. Biol.* 31, R918–R929.
23. Wang, H., Pipes, L., and Nielsen, R. (2021). Synonymous mutations and the molecular evolution of SARS-CoV-2 origins. *Virus Evol.* 7, veaa098.
24. Lumby, C.K., Zhao, L., Breuer, J., and Illingworth, C.J.R. (2020). A large effective population size for established within-host influenza virus infection. *Elife* 9, e56915.
25. Bloom, J.D., and Neher, R.A. (2024). Fitness effects of mutations to SARS-CoV-2 proteins. *Virus Evol.* 9, vead055.
26. Översti, S., Gaul, E., Jensen, B.-E.O., and Kühnert, D. (2023). Phylogenetic meta-analysis of chronic SARS-CoV-2 infections in immunocompromised patients shows no evidence of elevated evolutionary rates. *bioRxiv*, 2023.11.01.565087. 10.1101/2023.11.01.565087.
27. Harari, S., Miller, D., Fleishon, S., Burstein, D., and Stern, A. (2024). Using big sequencing data to identify chronic SARS-Coronavirus-2 infections. *Nat. Commun.* 15, 1–12.
28. Lythgoe, K.A., Gardner, A., Pybus, O.G., and Grove, J. (2017). Short-Sighted Virus Evolution and a Germline Hypothesis for Chronic Viral Infections. *Trends Microbiol.* 25, 336–348.
29. Kehrer, T., Cupic, A., Ye, C., Yildiz, S., Bouhaddou, M., Crossland, N.A., Barrall, E.A., Cohen, P., Tseng, A., Çağatay, T., et al. (2023). Impact of SARS-CoV-2 ORF6 and its variant polymorphisms on host responses and viral pathogenesis. *Cell Host Microbe* 31, 1668-1684.e12.

30. Miyamoto, Y., Itoh, Y., Suzuki, T., Tanaka, T., Sakai, Y., Koido, M., Hata, C., Wang, C.-X., Otani, M., Moriishi, K., et al. (2022). SARS-CoV-2 ORF6 disrupts nucleocytoplasmic trafficking to advance viral replication. *Commun. Biol.* 5, 1–15.
31. Reuschl, A.-K., Thorne, L.G., Whelan, M.V.X., Ragazzini, R., Furnon, W., Cowton, V.M., De Lorenzo, G., Mesner, D., Turner, J.L.E., Dowgier, G., et al. (2024). Evolution of enhanced innate immune suppression by SARS-CoV-2 Omicron subvariants. *Nat. Microbiol.* 9, 451–463.
32. Li, G., Hilgenfeld, R., Whitley, R., and De Clercq, E. (2023). Therapeutic strategies for COVID-19: progress and lessons learned. *Nat. Rev. Drug Discov.* 22, 449–475.
33. Huygens, S., GeurtsvanKessel, C., Gharbharan, A., Bogers, S., Worp, N., Boter, M., Bax, H.I., Kampschreur, L.M., Hassing, R.-J., Fiets, R.B., et al. (2024). Clinical and virological outcome of monoclonal antibody therapies across severe acute respiratory syndrome Coronavirus 2 variants in 245 immunocompromised patients: A multicenter prospective cohort study. *Clin. Infect. Dis.*, ciae026.
34. Carlin, A.F., Clark, A.E., Chaillon, A., Garretson, A.F., Bray, W., Porrachia, M., Santos, A.T., Rana, T.M., and Smith, D.M. (2023). Virologic and immunologic characterization of Coronavirus disease 2019 recrudescence after nirmatrelvir/ritonavir treatment. *Clin. Infect. Dis.* 76, e530–e532.
35. Hosaka, Y., Yan, Y., Naito, T., Oyama, R., Tsuchiya, K., Yamamoto, N., Nojiri, S., Hori, S., Takahashi, K., and Tabe, Y. (2023). SARS-CoV-2 evolution among patients with immunosuppression in a nosocomial cluster of a Japanese medical center during the Delta (AY.29 sublineage) surge. *Front. Microbiol.* 14. 10.3389/fmicb.2023.944369.
36. Dadonaite, B., Crawford, K.H.D., Radford, C.E., Farrell, A.G., Yu, T.C., Hannon, W.W., Zhou, P., Andrabi, R., Burton, D.R., Liu, L., et al. (2023). A pseudovirus system enables deep mutational scanning of the full SARS-CoV-2 spike. *Cell* 186, 1263-1278.e20.
37. Shajahan, A., Pepi, L.E., Rouhani, D.S., Heiss, C., and Azadi, P. (2021). Glycosylation of SARS-CoV-2: structural and functional insights. *Anal. Bioanal. Chem.* 413, 7179–7193.
38. Pouwels, K.B., House, T., Pritchard, E., Robotham, J.V., Birrell, P.J., Gelman, A., Vihta, K.-D., Bowers, N., Boreham, I., Thomas, H., et al. (2021). Community prevalence of SARS-CoV-2 in England from April to November, 2020: results from the ONS Coronavirus Infection Survey. *Lancet Public Health* 6, e30–e38.
39. ncov2019-artic-nf: A Nextflow pipeline for running the ARTIC network’s fieldbioinformatics tools (<https://github.com/artic-network/fieldbioinformatics>), with a focus on ncov2019 (Github).
40. Bonsall, D., Golubchik, T., de Cesare, M., Limbada, M., Kosloff, B., MacIntyre-Cockett, G., Hall, M., Wymant, C., Ansari, M.A., Abeler-Dörner, L., et al. (2020). A comprehensive genomics solution for HIV surveillance and clinical monitoring in low-income settings. *J. Clin. Microbiol.* 58. 10.1128/jcm.00382-20.
41. Wymant, C., Blanquart, F., Golubchik, T., Gall, A., Bakker, M., Bezemer, D., Croucher, N.J., Hall, M., Hillebregt, M., Ong, S.H., et al. (2018). Easy and accurate reconstruction of whole HIV genomes from short-read sequence data with shiver. *Virus Evol.* 4, vey007.
42. Rambaut, A., Holmes, E.C., O’Toole, Á., Hill, V., McCrone, J.T., Ruis, C., du Plessis, L., and Pybus, O.G. (2020). A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat. Microbiol.* 5, 1403–1407.

43. Zhao, L., and Illingworth, C.J.R. (2019). Measurements of intrahost viral diversity require an unbiased diversity metric. *Virus Evol.* 5, vey041.
44. Sagulenko, P., Puller, V., and Neher, R.A. (2018). TreeTime: Maximum-likelihood phylodynamic analysis. *Virus Evol.* 4, vex042.
45. Druelle, V., and Neher, R.A. (2023). Reversions to consensus are positively selected in HIV-1 and bias substitution rate estimates. *Virus Evol.* 9, veac118.
46. Raghwani, J., Redd, A.D., Longosz, A.F., Wu, C.-H., Serwadda, D., Martens, C., Kagaayi, J., Sewankambo, N., Porcella, S.F., Grabowski, M.K., et al. (2018). Evolution of HIV-1 within untreated individuals and at the population scale in Uganda. *PLoS Pathog.* 14, e1007167.

Acknowledgement

The CIS was funded by the Department of Health and Social Care and the UK Health Security Agency, with in-kind support from the Welsh Government, the Department of Health on behalf of the Northern Ireland Government and the Scottish Government. The COVID-19 Infection Survey Group of the COVID-19 Genomics UK (COG-UK) Consortium was supported by funding from the Medical Research Council part of UK Research & Innovation, the National Institute of Health Research (NIHR) (grant code: MC_PC_19027) and Genome Research Limited, operating as the Wellcome Sanger Institute. We acknowledge use of data generated through the COVID-19 Genomics Programme funded by the Department of Health and Social Care. A.S.W. is supported by the NIHR Health Protection Research Unit in Healthcare Associated Infections and Antimicrobial Resistance at the University of Oxford in partnership with the UK Health Security Agency (NIHR200915) and the NIHR Oxford Biomedical Research Centre, and is an NIHR Senior Investigator. K.L. is supported by the Royal Society and the Wellcome Trust (107652/Z/15/Z) and by the Li Ka Shing Foundation. The research was supported by the Wellcome Trust Core Award grant number 203141/Z/16/Z, with funding from the NIHR Oxford BRC. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR, the Department of Health, the Department of Health and Social Care or the UK Health Security Agency.

Supplementary Tables

Supplementary Table 1: Model comparison for estimating within-host evolutionary rates.

Comparison of regression models for estimating genome-wide (GW), nonsynonymous (NS), and synonymous (S) evolutionary rates. Each model is presented with its corresponding equation and Bayesian Information Criterion (BIC) value, which assesses model fit to the data. Parameters e_0 and r_0 represent fixed effects for y-intercept at time $t=0$ (corresponding to the day when the first sample from a persistent infection was collected) and rate across all persistent infections, respectively; $d_i(t)$ represents distance at time t for persistent infection i (dependent variable); r_i and e_i represent random effects for evolutionary rate and intercept per persistent infection, respectively; $\varepsilon_i(t)$ is the error term which represents the unexplained variability in the dependent variable; the index j corresponds to nine categories for y-intercept labelled based on sequencing centre(s) that genetic samples are collected from. Models with lowest BIC values are highlighted with an underline.

Regression model	Equation	BIC (GW)	BIC (NS)	BIC (S)
Complete pooling	$d_i(t) = r_0 t + e_0 + \varepsilon_i(t)$	-8059	-7755	-6792
Random intercept	$d_i(t) = r_0 t + e_i + e_0 + \varepsilon_i(t)$	-8131	-7822	<u>-6880</u>
Random slope with one fixed intercept	$d_i(t) = (r_0+r_i) t + e_0 + \varepsilon_i(t)$	<u>-8146</u>	<u>-7866</u>	-6863
Random slope with multiple fixed intercepts	$d_i(t) = (r_0+r_i) t + \sum_j e_j + \varepsilon_i(t)$	-8142	-7860	-6830
No pooling	$d_i(t) = r_i t + e_i + \varepsilon_i(t)$	*	*	*

*Number of observations is smaller than the number of random effects.

Supplementary Table 2: Evaluation of associations between various host factors and within-host evolutionary rates. (a)

This table examines the impact of integrating individual host factors—age, sex, vaccination status, prior infection, virus lineage, duration of infection, and RNA viral load dynamics—into the best-fit regression model as fixed effect parameters and comparing best fits using the Bayesian Information Criterion (BIC) values. The baseline model is a linear mixed-effects regression, identified as the optimal fit for genome-wide (GW), nonsynonymous (NS), and synonymous (S) distances over time (see Supplementary Table 1). Each of the seven factors is added as a fixed effect to this baseline model, with categorical variables including age (aged 60 and above: 295; aged below 60: 199), sex (male: 293; female: 201), vaccination status (received at least one dose: 470; no vaccination: 24), prior infection (none: 478; at least one: 16), viral lineage (10 Alpha, 95 Delta, 87 BA.1, 173 BA.2, 14 BA.4, 111 BA.5, and 4 XBB with measurable evolution), and viral load dynamics (experienced viral rebound: 32; no rebound detected: 462). Duration of infection is classed as a continuous variable ranging from 26 to 316 days per infection. (b) Comparing the BIC values for a subset of infections with durations lasting longer than 36 days (198 infections) and 56 days (110 infections) between the null model and a model that includes duration of infection as an additional fixed effect parameter.

(a)

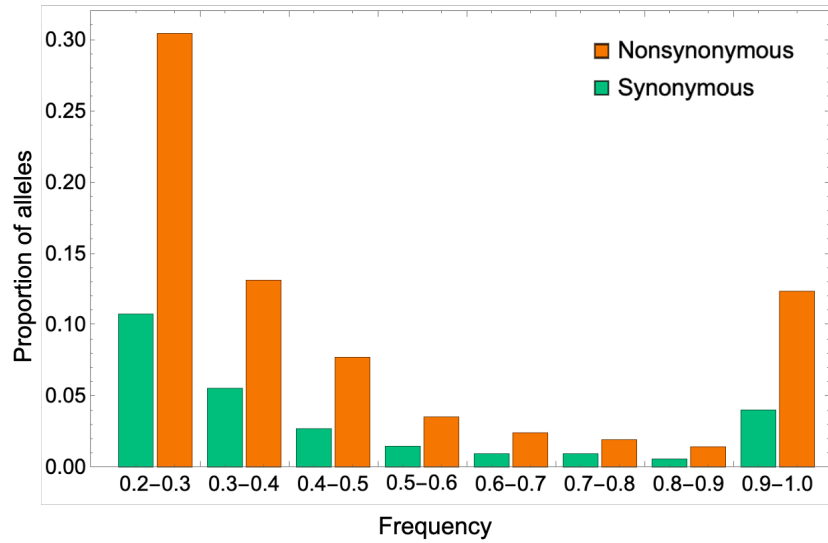
Fixed effects	BIC (GW)	BIC (NS)	BIC (S)
Null model	-8146	-7866	-6880
Virus lineage	-8120	-7838	-6847
Prior infection	-8141	-7860	-6874
Vaccination status	-8141	-7861	-6874
Sex	-8142	-7862	-6874
Age	-8143	-7861	-6874
Viral load dynamics	-8145	-7867	-6877
Duration of infection	-8150*	-7868*	-6881

(b)

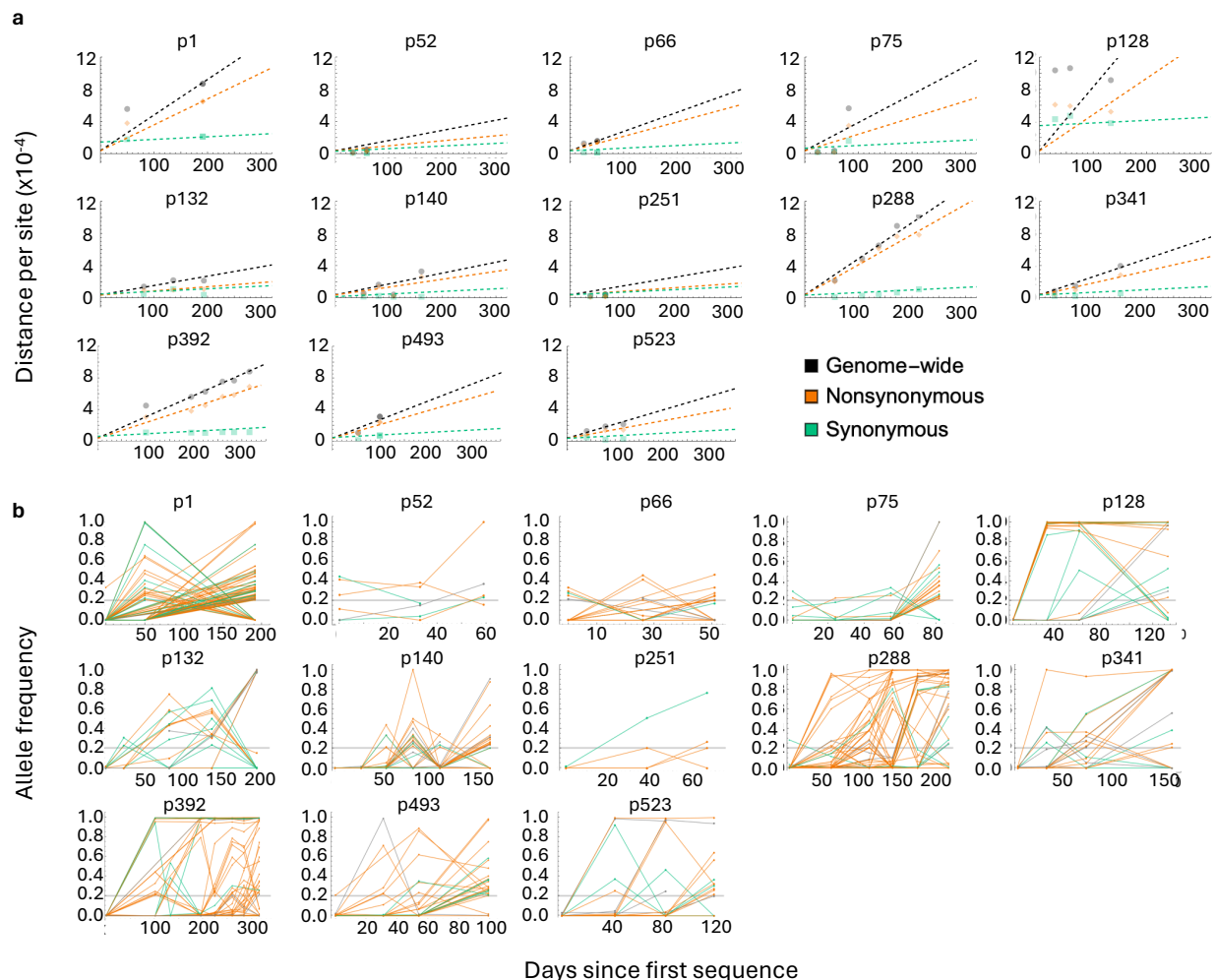
Fixed effects	BIC (GW)	BIC (NS)	BIC (S)
Null model (t>36)	-2907	-2925	-2661
Duration of infection (t>36)	-2910*	-2928*	-2659
Null model (t>56)	-1559	-1558	-1605
Duration of infection (t>56)	-1560	-1560*	-1602

*Indicates $\Delta BIC = BIC_{Null} - BIC_{Alternative} > 2$.

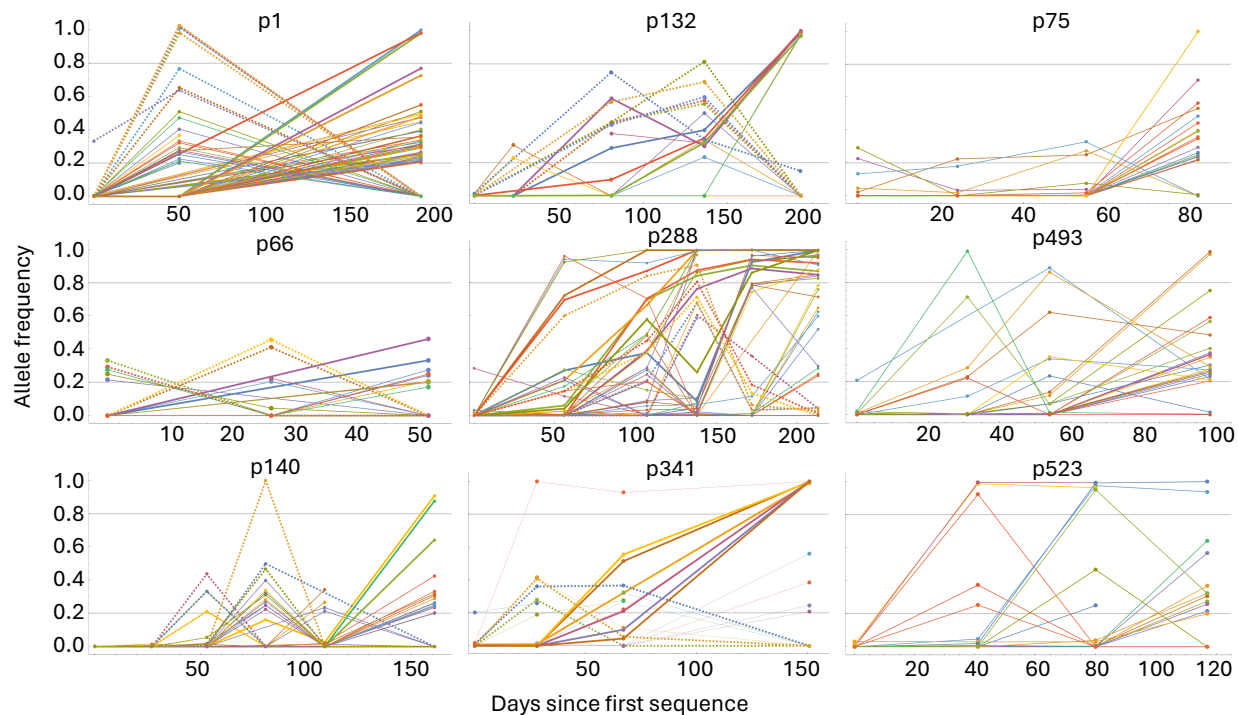
Supplementary Figures



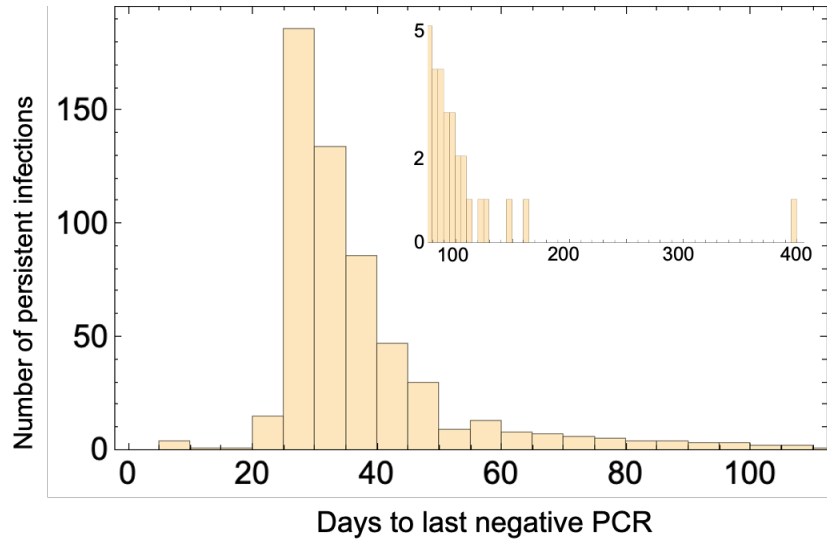
Supplementary Figure 1: Site frequency spectrum. Proportion of synonymous (green) and nonsynonymous (orange) mutations in persistent infections across all frequency bands.



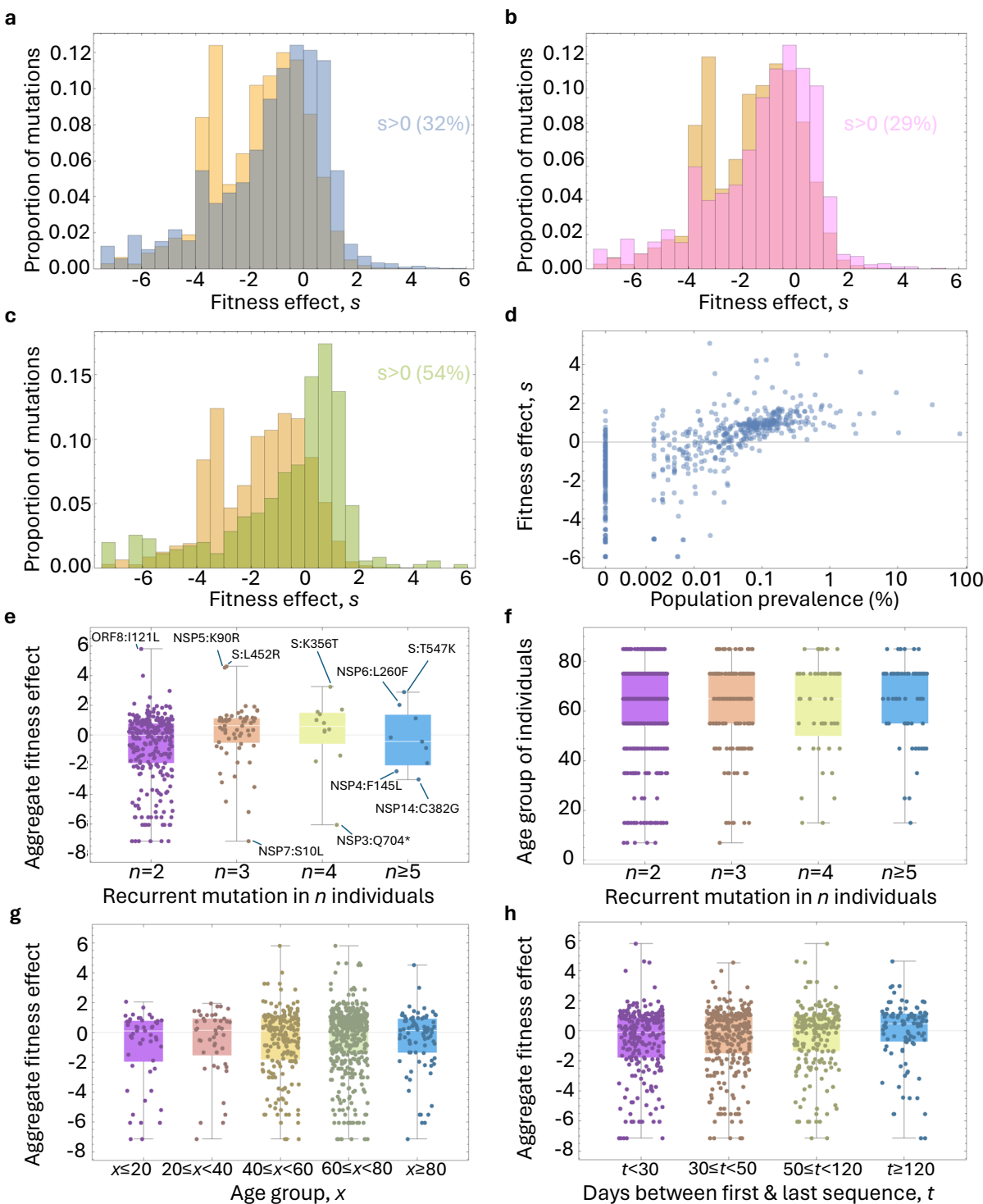
Supplementary Figure 2: Rates of genome-wide, nonsynonymous, and synonymous evolution in 13 persistently infected individuals. (a) Illustrates the evolutionary distance over time for a subset of 13 persistently infected individuals, each characterised by a minimum of three temporal data points and the presence of at least one synonymous and one nonsynonymous mutant allele. Points on the graph represent the total genetic distance from the consensus sequence at the initial time point, calculated based on allele frequency changes over time. Dashed lines indicate the regression lines that best fit these data. (b) Shows the allele frequency trajectories for the 13 persistent infections examined, categorised into synonymous, nonsynonymous, and non-coding (grey) mutations. Each mutation that reached a minimum frequency of 20% at least at one time point is shown. A horizontal grey line across the graphs marks the 20% allele frequency threshold.



Supplementary Figure 3: Temporal allele frequency dynamics in nine persistent infections. The figure illustrates two distinct patterns of allele dynamics over time. In the left column (infections p1, p66, and p140), we observe transient allele groups that emerge at one time point, with some reaching high frequencies before vanishing in subsequent time points (dashed lines). Consensus sequence samples from p1, p66, and p140 (as well as the other 6 infections shown here) form a monophyletic clade on a representative phylogeny of non-persistently infected individuals⁸. Additionally, certain alleles that were not present at the early stages of infection surge to high frequencies towards the end of infection (bold solid lines). Conversely, the middle column (infections p132, p288, and p341) showcases alleles that experience a sweep from low to high frequencies, with some ultimately disappearing (dashed lines) and others reaching fixation (bold solid lines). The right column (p75, p288, and p523) show allele frequency dynamics that is a mix of the two patterns with some alleles appearing and disappearing in groups while other are present in the population in at least two time points, with some reaching fixation without disappearing at later time points.

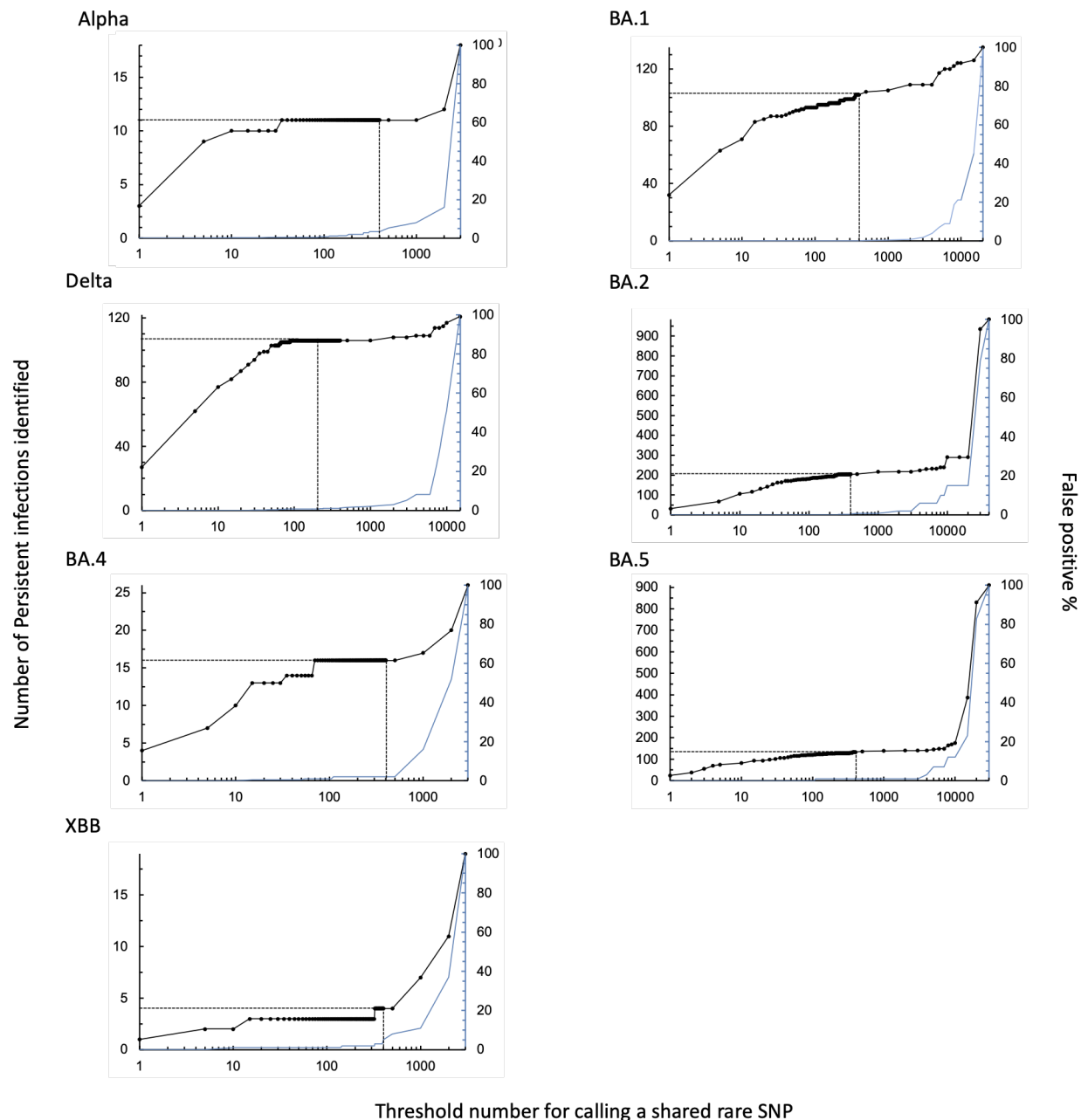


Supplementary Figure 4: Number of days elapsed since the last time a persistently infected individual had a negative PCR test. The histogram plot includes all 576 identified persistent infections.

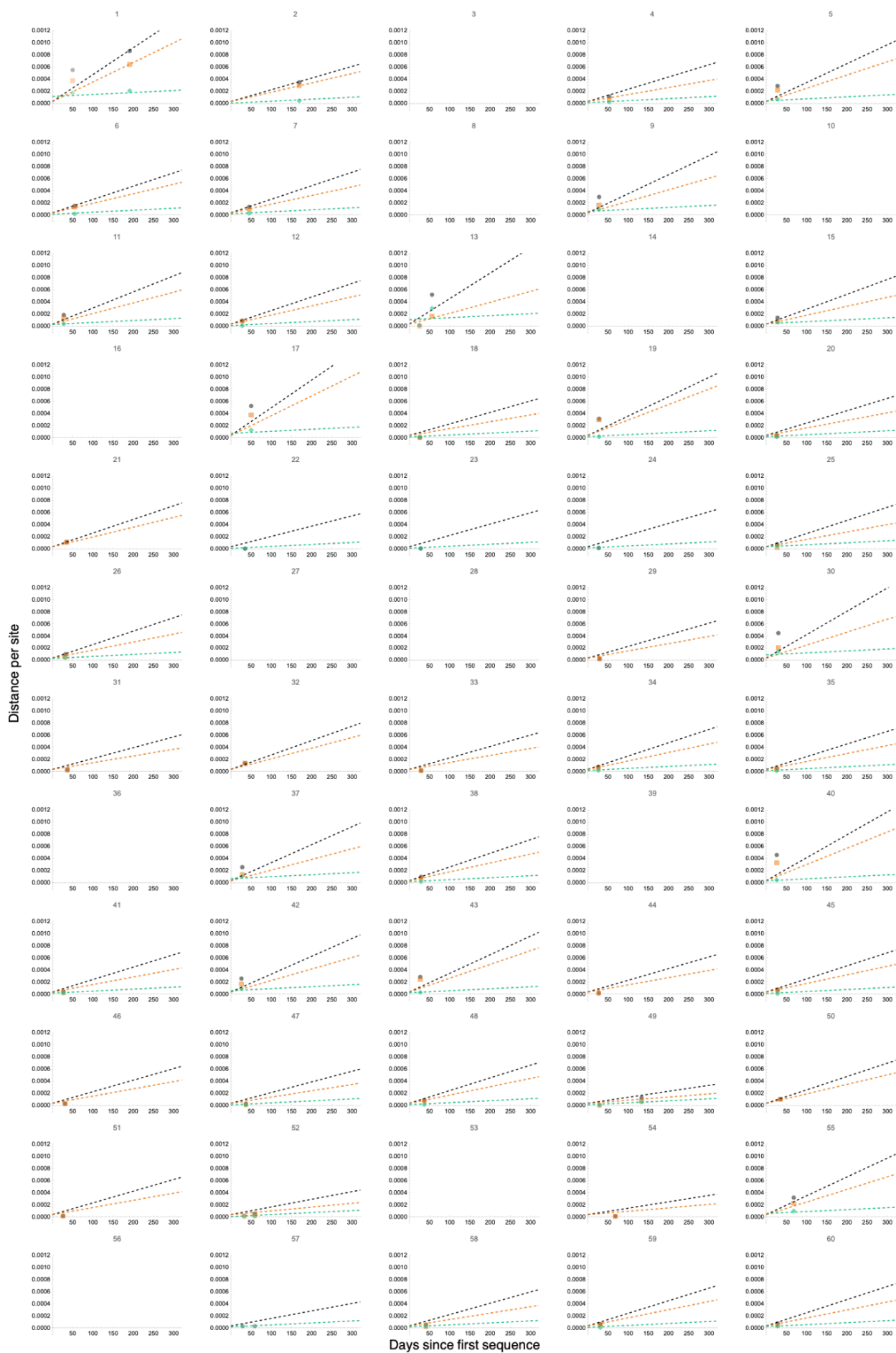


Supplementary Figure 5: Between-host fitness effect and prevalence of recurrent mutations identified in persistently-infected individuals. (a) Distribution of between-host fitness effects of all SARS-CoV-2 mutations on a global phylogeny (orange), between-host fitness of all mutations found in persistently infected individuals (blue), (b) for those found only in a single persistent infection (magenta), and (c) for those found in two or more persistent infections (green). The percentage of mutations in persistent infections with a positive between-host fitness effect (s) is highlighted on each graph in (a)-(c). The between-host fitness effect of mutations in persistent infections corresponds to the fitness effect of that mutation on a global phylogeny within the same major viral lineage that was found to be in the persistently infected individual. For example, if a recurrent mutation is found in two persistently infected individuals with BA.2 and BA.5 infections, the between-host fitness effect of that mutation in both the BA.2

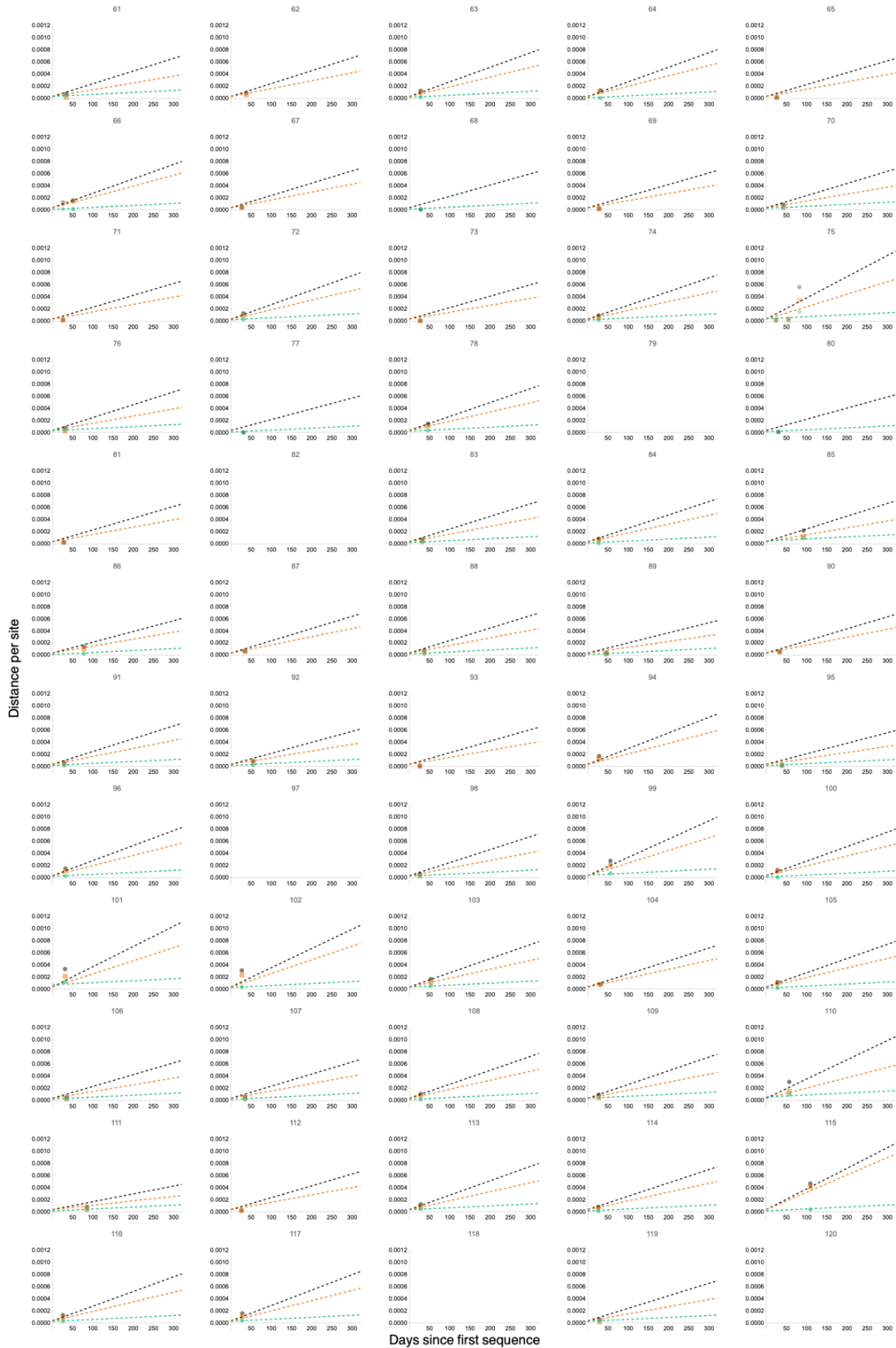
and BA.5 major lineages is recorded. **(d)** The between-host fitness effect of recurrent mutations found in persistent infections and their corresponding prevalence across all ONS-CIS sequences of the same major lineage as the persistent infection. **(e)** The aggregate between-host fitness effect (averaged across all major lineages of SARS-CoV-2 on a global phylogeny) of recurrent mutations found in n persistent infections. Some of the mutations with extremely high and low fitness effects are highlighted. **(f)** Age-group of all individuals which share n recurrent mutations. **(g)** Aggregate fitness effect of recurrent mutations per age group. **(h)** Aggregate fitness effect of recurrent mutations based on the duration of the persistent infection (as measured based on number of days between first and last sequence from a persistent infection) in which they emerged. Fitness effect of mutations are taken from https://github.com/jbloomlab/SARS2-mut-fitness/blob/main/results_public_2024-04-19/nt_fitness/ntmut_fitness_by_clade.csv²⁵.

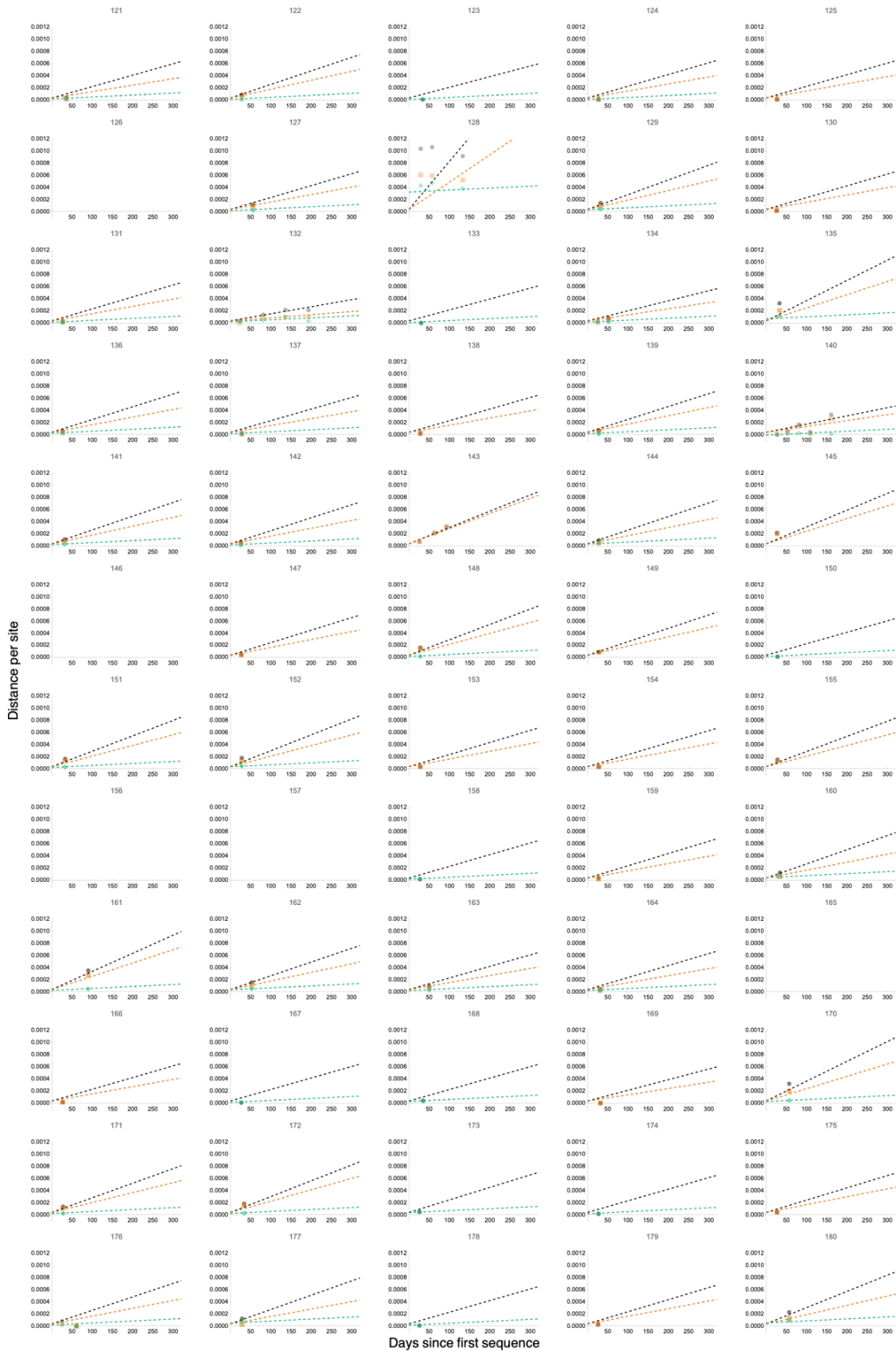


Supplementary Figure 6: Number of persistent infections identified with a shared rare SNP as a function of the threshold number of cases for calling a rare SNP. A threshold value of 1 for a rare SNP means the rare SNP is only found in one sequence of that lineage in the ONS-CIS dataset, excluding sequences from any persistently infected individuals. The number of persistent infections identified gives the number of persistent infections lasting at least 26 days we would identify as persistent in the ONS-CIS using the given threshold (black). The false positive percentage gives the percentage of times two random samples of the same major lineage taken from the ONS-CIS would be falsely identified as belonging to the same persistent infection (blue; 1,000 pairs of samples were considered). As the threshold value for calling a rare SNP increases, the number of persistent infections identified (black) increases, but so does the false positive rate. Similar to the approach we took in our previous study⁸, we chose a threshold number of 400 (vertical dashed line) in this study for identifying persistent infections, since for this threshold the percentage of false positives were 0-3% for all major lineages, but the number of persistent infections identified has begun to plateau. We allowed for possible misclassification of some BA.2 and BA.5 major lineages by allowing for potential identification of persistent infections with a mix of BA.2 and BA.5 samples.

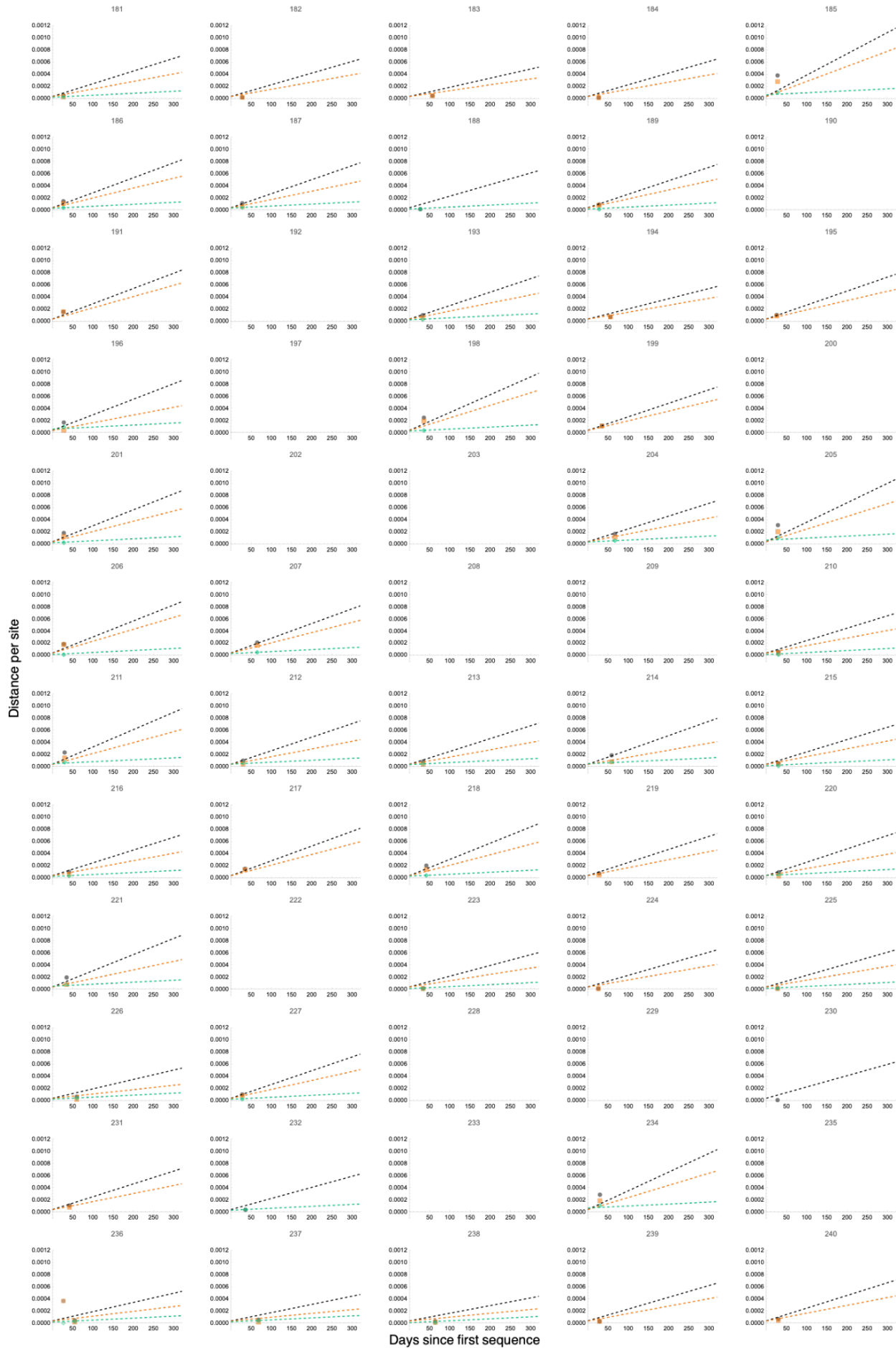


Days since first sequence

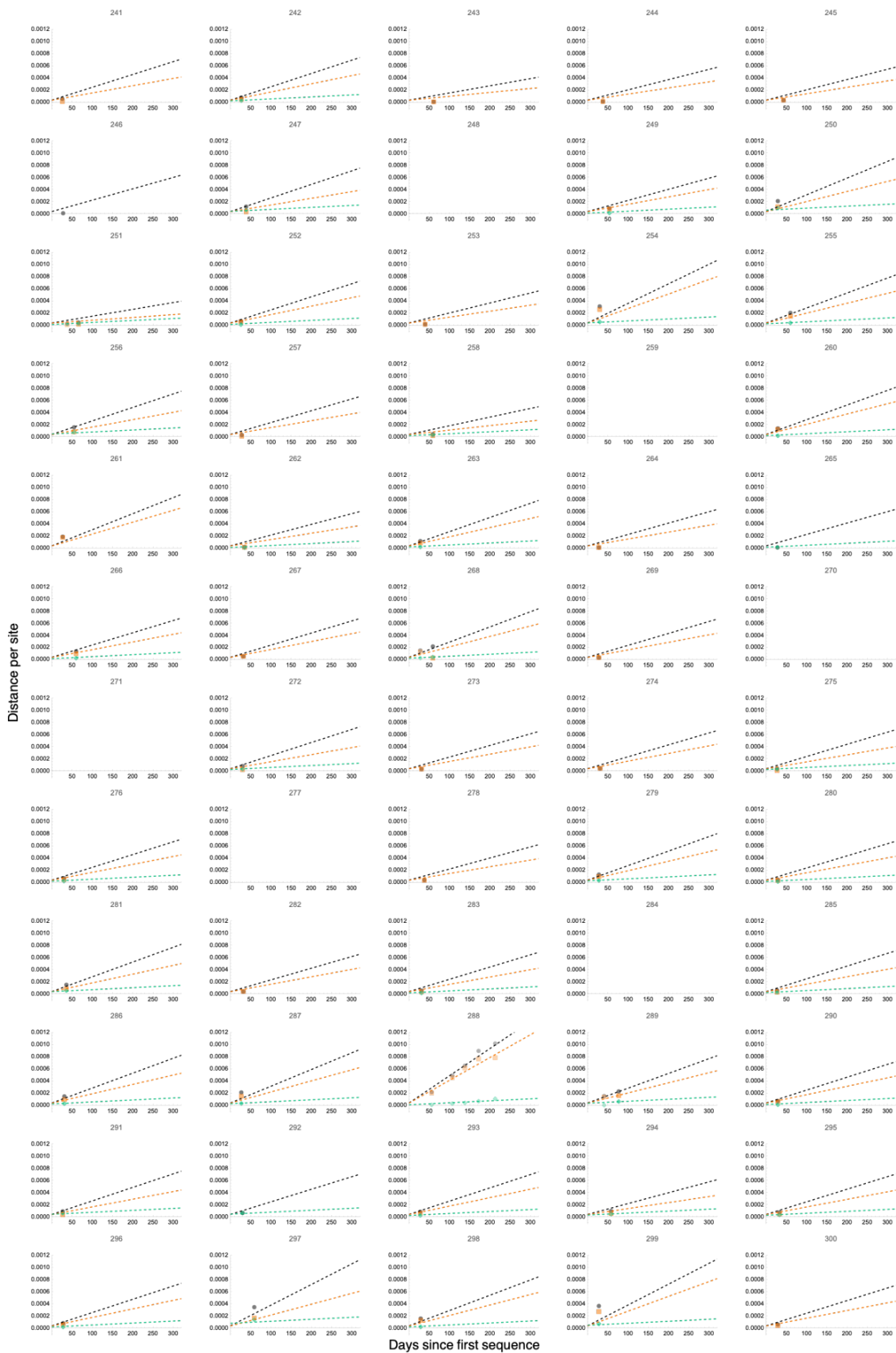


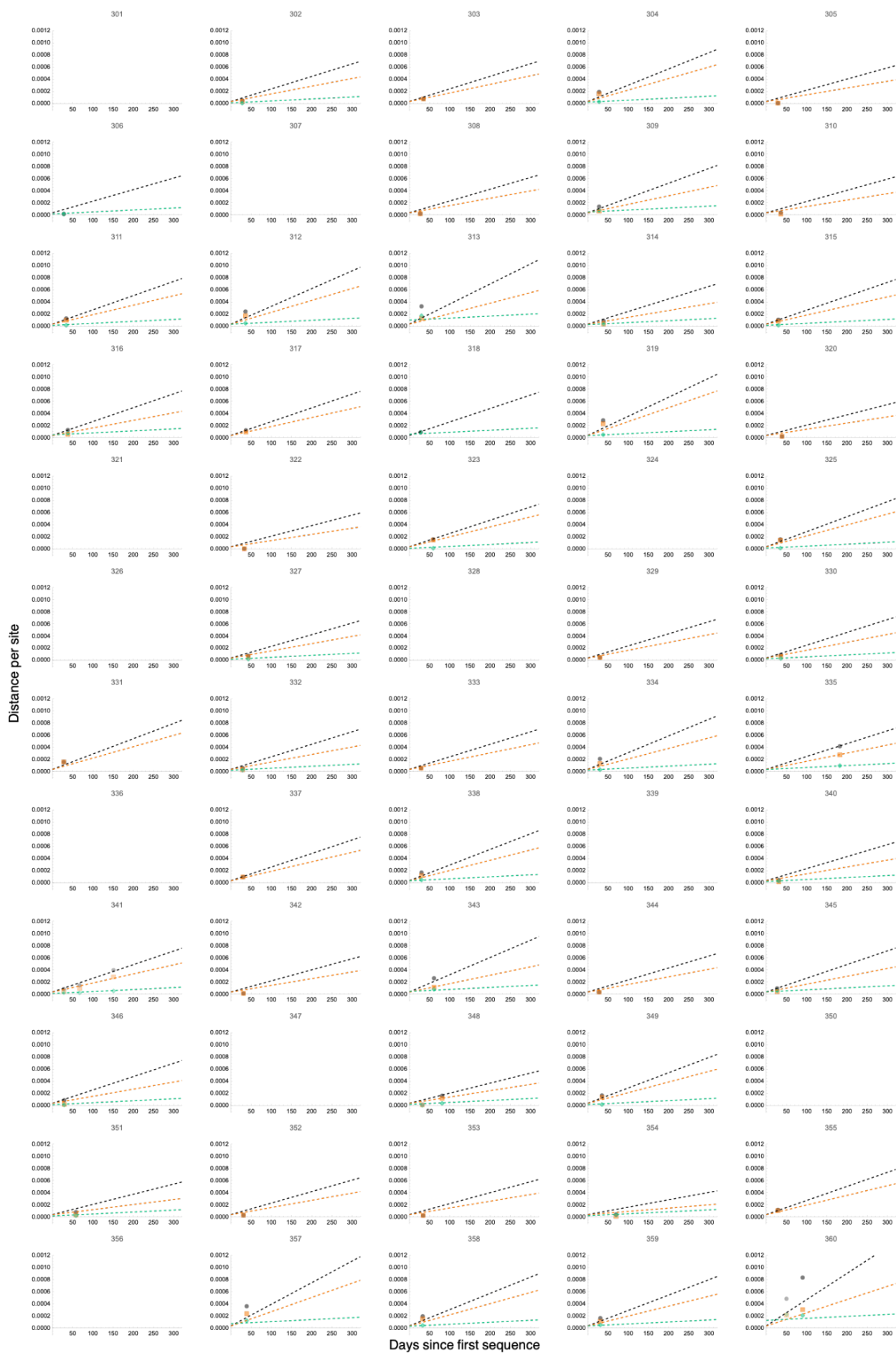


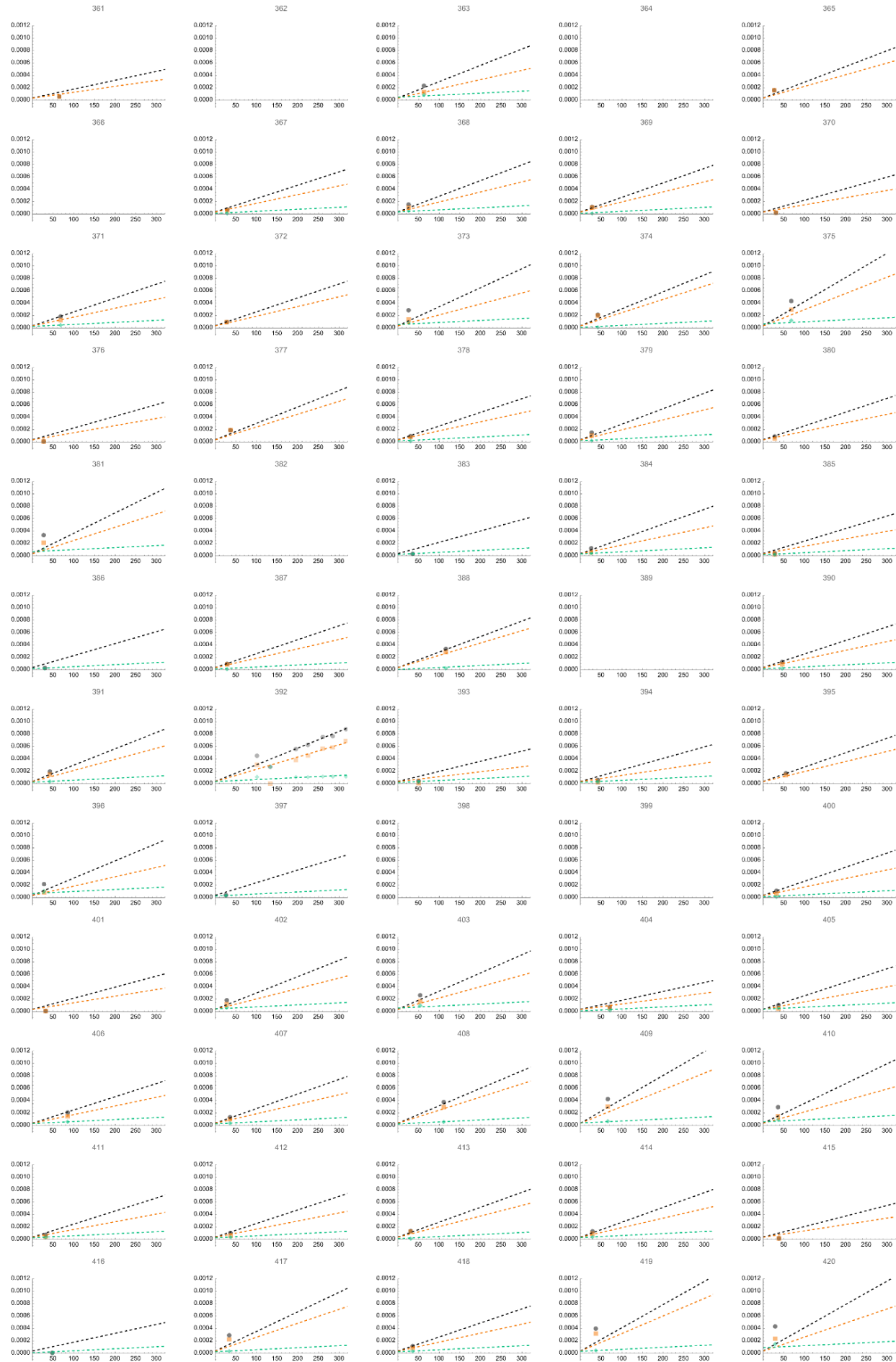
Days since first sequence

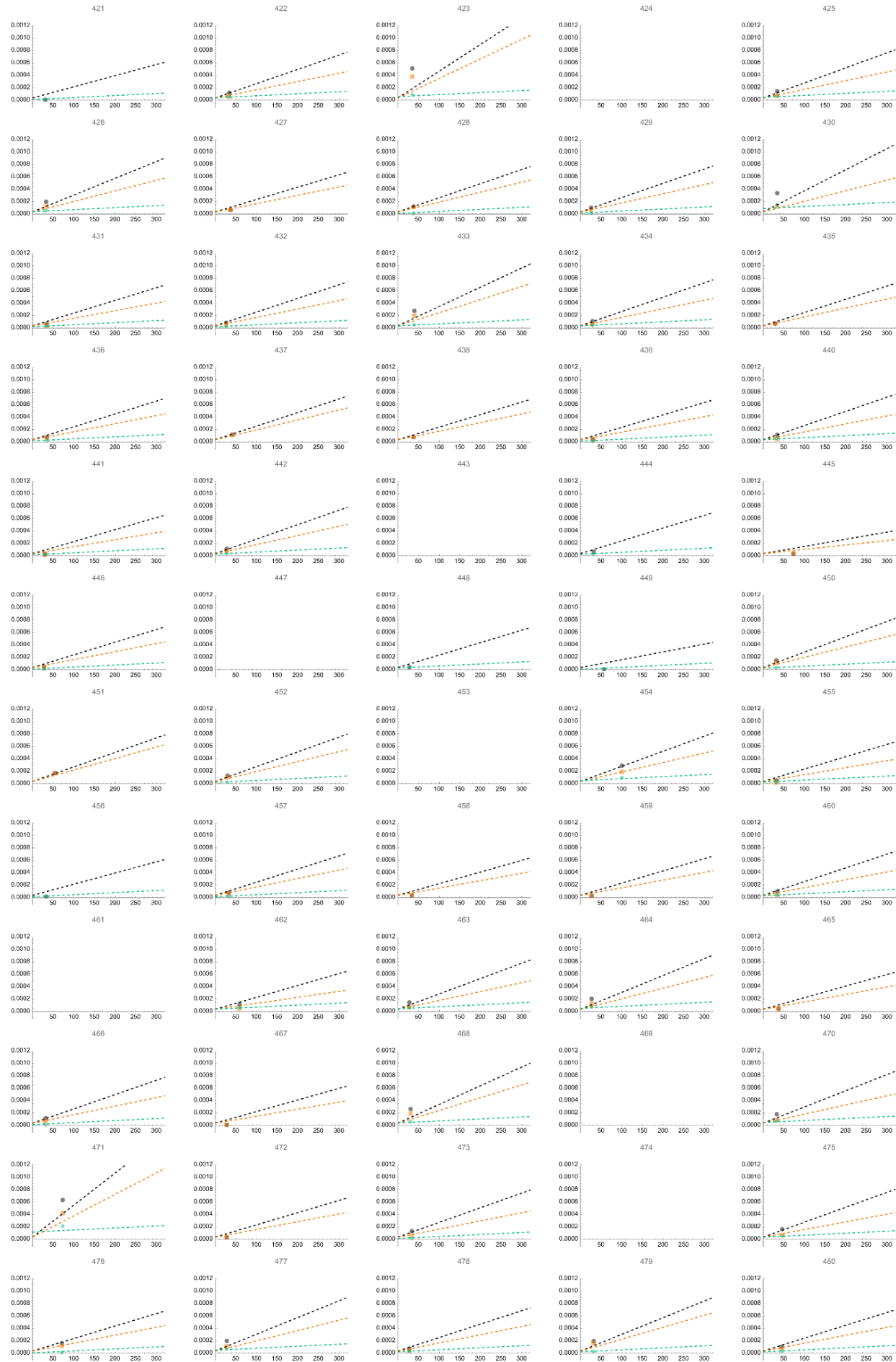


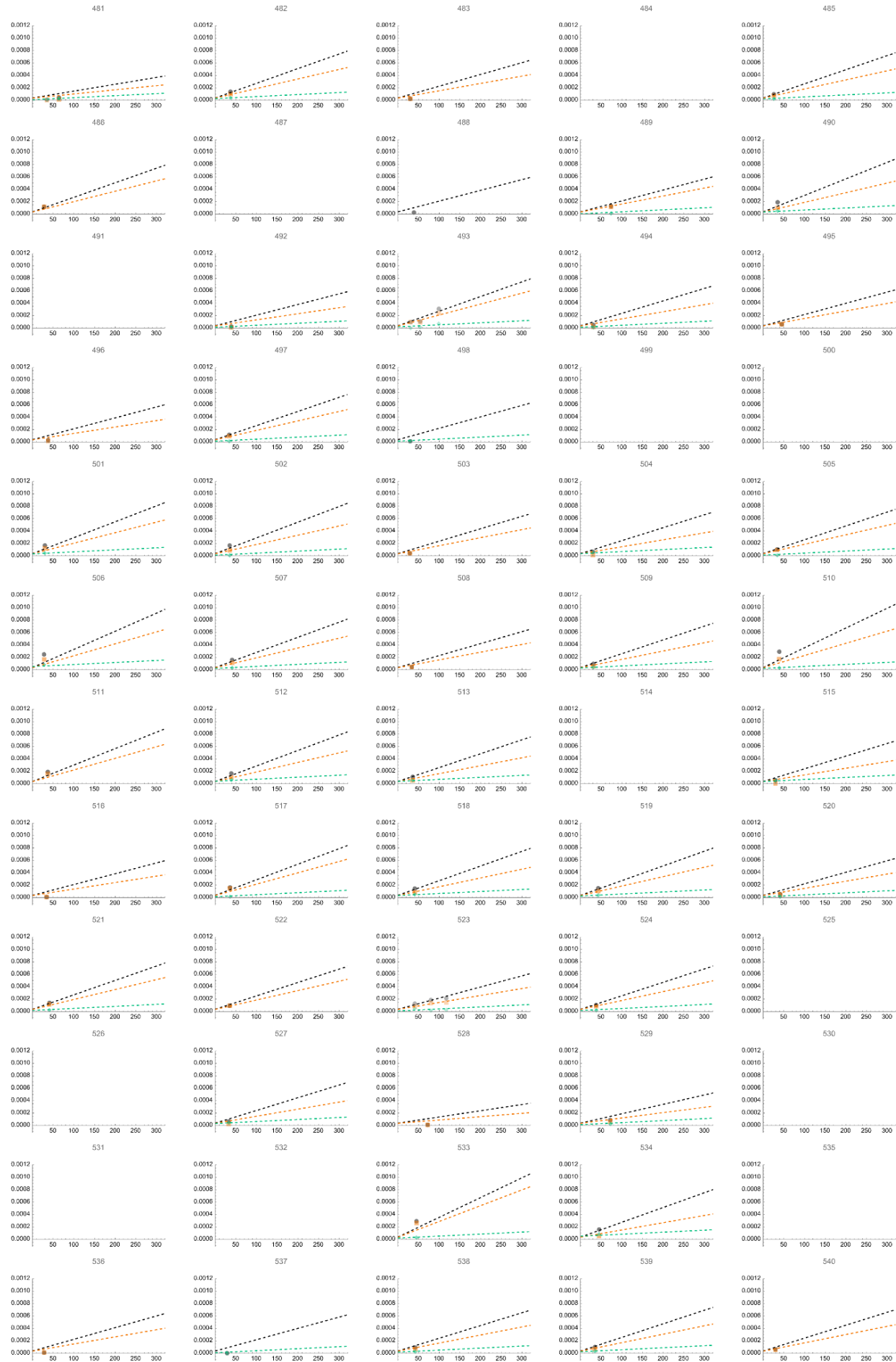
Days since first sequence

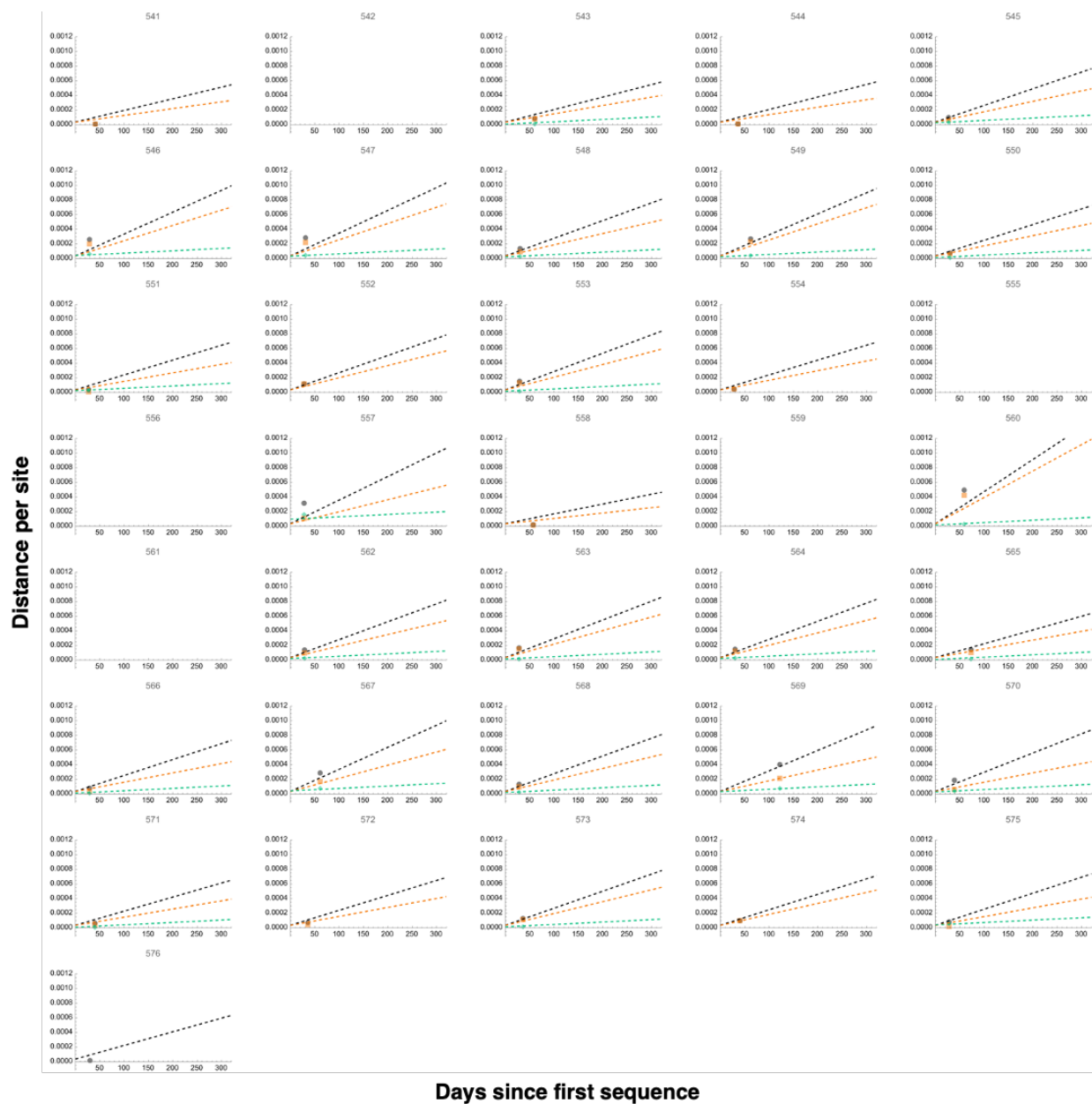












Supplementary Figure 7: Rates of genome-wide, nonsynonymous, and synonymous evolution in all persistently infected individuals with measurable rates. The evolutionary distance over time for 494 persistently infected individuals with measurable genome-wide rate (black), 457 nonsynonymous rate (orange), and 368 synonymous rate (green). Points on the graph represent the total genetic distance from the consensus sequence at the initial time point, calculated based on allele frequency changes over time. Dashed lines indicate the regression lines that best fit these data.