

# THE LANCET Microbe

## **Supplementary appendix 1**

This appendix formed part of the original submission and has been peer reviewed.  
We post it as supplied by the authors.

Supplement to: Ghafari M, Kemp SA, Hall M, et al. SARS-CoV-2 genomic diversity and within-host evolution in individuals with persistent infection in the UK: an observational, longitudinal, population-based surveillance study. *Lancet Microbe* 2025. <https://doi.org/10.1016/j.lanmic.2025.101154>

# Appendix 1

## Table of content

1. Sequencing
2. Association between infection type and SARS-CoV-2 lineage
3. Nucleotide diversity
4. Estimating within-host genetic distance
5. Estimating within-host evolutionary rate
6. Estimating within- and between-lineage rates at the between-host level
7. Handling of missing data for the evolutionary rate analysis
8. Divergence rate from putative founder
9. Supplementary Table 1: Model comparison for estimating within-host evolutionary rates
10. Supplementary Table 2: Associations between various host factors and within-host evolutionary rates
11. Supplementary Table 3: Recurrent mutations found in four or more persistent infections
12. Supplementary Table 4: Recurrent mutations with known drug-resistant properties
13. Supplementary Figure 1: Number of persistent infections identified as a function of the threshold number of cases for calling a rare SNP
14. Supplementary Figure 2: Number of days elapsed since the last time a persistently infected individual had a negative PCR test
15. Supplementary Figure 3: Site frequency spectrum
16. Supplementary Figure 4: Rates of genome-wide, nonsynonymous, and synonymous evolution in 13 persistently infected individuals
17. Supplementary Figure 5: Temporal allele frequency dynamics in nine persistent infections
18. Supplementary Figure 6: Between-host fitness effect and prevalence of recurrent mutations identified in persistently infected individuals
19. Supplementary Figure 7: Rates of genome-wide, nonsynonymous, and synonymous evolution in all persistently infected individuals
20. References
21. The COVID-19 Infection Survey Group author list
22. The COVID-19 Genomics UK (COG-UK) Consortium author list

## Sequencing

Samples were sequenced at one of five sequencing centres, University of Oxford (OXON), Northumbria University and associated NHS foundation trusts (NORT), National Infection Service Public Health England (PHEC), Quadram Institute Bioscience, Norwich (NORW), and Wellcome Sanger Institute (Sanger). The great majority of samples were sequenced on Illumina Novaseq, with the rest using Oxford Nanopore GridION or MINION. The standard consensus FASTA sequences for all ONS-CIS samples were generated using the ARTIC Nextflow processing pipeline (v1) (1), or veSeq, an RNA sequencing protocol based on a quantitative targeted enrichment strategy (2,3) with consensus sequences produced using Shiver (v1.5.8) (4). For additional information about the survey, sequencing protocol, and FASTA consensus sequence protocol see (2,5).

## Association between infection type and SARS-CoV-2 lineage

By identifying a list of all high viral load persistent and non-persistent infections recorded through the ONS-CIS (**Figure 1**), we were able to assess whether the likelihood of persistent versus non-persistent infections differs by SARS-CoV-2 lineage, taking into account the overall number of infections per lineage.

We performed a chi-square test to compare the distribution of persistent and non-persistent infections across the major SARS-CoV-2 lineages in our study (Alpha, Delta, BA.1, BA.2, BA.2.75, BA.4, BA.5, and BQ.1). To ensure unbiased results, we excluded the XBB lineage from this analysis because the wave of infections with XBB was still ongoing at the end of the ONS-CIS survey (2). Additionally, we separated BA.2.75 and BQ.1 (and their descendants) from their parental BA.2 and BA.5 lineages, respectively, since these sublineages formed distinct and near-complete waves of infection in the UK by the time of our study.

The chi-square test yielded weak evidence of an association between infection type (persistent vs. non-persistent) and SARS-CoV-2 lineage ( $\chi^2 = 18.023$ ,  $df = 7$ ,  $p\text{-value} = 0.01187$ ). This suggests that the distribution of persistent and acute infections across lineages is not entirely uniform. However, when we restricted the analysis to lineages with fully completed waves of infection in the UK during the study period (Alpha, Delta, BA.1, BA.2 excluding BA.2.75, BA.4, and BA.5 excluding BQ.1), the evidence for an association weakened ( $\chi^2 = 9.9218$ ,  $df = 5$ ,  $p\text{-value} = 0.07748$ ).

This result aligns with the biological expectation, as we do not anticipate any specific lineage to be intrinsically more or less likely to cause persistent versus acute infections. Instead, the apparent association in the full dataset may reflect variability in sampling or incomplete data for lineages with ongoing or overlapping waves, such as BA.2.75, BQ.1, and XBB.

## Nucleotide diversity

Nucleotide diversity was calculated using the  $\pi$  statistic, which is the common measure of diversity least affected by the number of sequences used in the analysis (6). For each persistent infection, nucleotide diversity at a given time point is given by:

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

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

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

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

## Estimating within-host genetic distance

We used differences in mutant allele frequencies between two sequences from the same infection to calculate the genetic distance between the sequences. This is similar to an approach that has been used to measure within-host evolutionary rates of influenza A in a chronically infected individual (7). We calculated changes in allele frequency relative to the first sequenced time point in each persistent infection. Synonymous and nonsynonymous distance was determined by whether the mutant allele would result in the same (synonymous) or a different (nonsynonymous) amino acid being coded for compared to the first time point in the infection.

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

Following this definition of evolutionary distance, a mutant allele  $i$ , present at frequency  $f_i(t_0)$  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 pairwise distance between the two sequences. More generally, if the pair of sequences has  $M$  mutant alleles, the total genetic distance between them is

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

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

### Estimating within-host evolutionary rate

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

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

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

where  $r$  is the evolutionary rate and  $e$  is the y-intercept, which represents the expected amount of noise when measuring genetic distance. The noise could arise from either sequencing error or undiagnosed population structure (7). If a persistent infection does not have a detectable mutant allele that reaches frequency  $\geq 20\%$ , we exclude that individual from evolutionary rate analysis as we cannot quantify the contribution of noise in frequency change of alleles.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Our analysis showed, based on the lowest BIC value, that the random slope with one fixed intercept regression model (iii) best explains genome-wide and nonsynonymous evolutionary rates while the random intercept regression model (ii)

best explains synonymous rate for persistent infections. The lines of best fit for all the persistent infections with measurable evolution is shown in **Supplementary Figure 7**.

#### Handling of missing data for the evolutionary rate analysis

Missing data arose in two ways: (1) Viral samples from some infections had no measurable evolutionary rate (measurable genetic distance) based on our criteria, and (2) some samples were excluded due to poor genomic coverage. To ensure the reliability of our measurements, we excluded 82 persistent infections (out of 576) with no measurable viral evolution or whose samples all failed our quality-control thresholds -- this is now clearly laid out in our flow diagram (**Figure 1**). For the remaining 494 individuals, only samples that passed the genomic coverage criteria were used in the analysis, resulting in varying numbers of samples per individual.

A complete-case analysis approach was effectively applied at the level of samples. After filtering, only infections with at least two high-quality (>50% genomic coverage) samples and at least one reliable measurement of genetic distance (>50% genomic overlap between consensus sequences from the baseline and subsequent timepoints) were included.

Multiple imputation was not used because the missing data were not random. Missingness was due to technical reasons, such as poor genomic coverage or lack of measurable evolution. Imputing genetic distance values for excluded samples could have introduced bias into the analysis, as these excluded data were inherently unreliable.

The final dataset included 494 infections with varying numbers of samples per individual, all of which passed our quality-control thresholds. The random slope with a fixed intercept structure of our preferred mixed-effects model accounts for this variation in the number of samples across individuals, ensuring that all reliable data contributed appropriately to the estimation of evolutionary rates.

In terms of covariates, we considered age, sex, prior infection, vaccination status, viral load dynamics and duration of infection (**Supplementary Table 2**). There was no missing data in these data, although we accept that there could be ascertainment bias (particularly for example in terms of identifying prior infections, determined on the basis of tests in the survey and linked data from the national testing programmes).

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

To assess the saltatory evolution of SARS-CoV-2 at the between-host level, we used a previously identified representative sample from the ONS-CIS dataset (2). This dataset covered sequences from the Alpha, Delta, Omicron BA.1, BA.2 (excluding BA.2.75), BA.2.75, BA.4, BA.5 (excluding BQ.1), and BQ.1 lineages. We then constructed the ancestral sequence for each major lineage using TreeTime (8) and

calculated total, nonsynonymous, and synonymous Hamming distances between samples from each major lineage relative to the ancestral sequence of the same major lineage. Finally, to estimate the between-lineage rate, we calculated the total, nonsynonymous, and synonymous Hamming distances between the Wuhan reference sequence (NC\_045512.2) and the ancestral sequence for each major lineage.

#### Divergence rate from putative founder

To explore evolutionary rate variation across the genome, we next assumed the consensus sequence at baseline represents the founder virus, and that the start of infection occurred at the midpoint between the last negative PCR test and the first sampled time point of the persistent infection. For the majority of infections, the last negative PCR test was taken between 20 to 40 days before the baseline (see **Supplementary Figure 2**). Using the estimated infection start dates, we calculated an evolutionary rate for each region of the genome, aggregating across all individuals. We called this the divergence rate to distinguish it from the approach we took to measure evolutionary rates per individual, because most infections had only a limited number of mutations, which precluded a calculation of a per-individual rate per gene or gene segment. This commonly used approach to measuring within-host divergence rates comes with two key disadvantages compared to the intra-infection evolutionary rates. First, it requires estimating the time elapsed since the start of the infection rather than using only known sample collection dates. Second, this method has a tendency to ascribe any changes in allele frequencies, or their absence, to substitution rates rather than to sampling noise.

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

While this method is frequently used for calculating within-host divergence rates for viruses like HIV (9), it will miss early fixation events that might have shifted the consensus sequence away from the true founding virus by the time the first sample was collected; assumes the founding viral population was genetically homogeneous (10); does not control for noise which could bias estimates of the divergence rate. Nonetheless, aggregating across a large number of individuals should help mitigate these effects.

This approach involved treating each measurement of divergence from the putative founder at any given time point,  $t$ , as an independent observation, regardless of its associated persistent infection. The divergence from the founder for each genomic segment at any time point, including baseline, was defined as the cumulative



frequency of all mutant alleles within that segment at time  $t$ . For example, if there were no mutant alleles within a genomic segment at a given time point, we recorded a divergence of zero. Subsequently, we used a linear regression with a zero y-intercept at the start time of infection to calculate the divergence rate from the putative founder for each genomic segment. This can be expressed as  $d^{(n)}_i(t) = r_i t + \varepsilon^{(n)}_i(t)$ , where  $r_i$  is the divergence rate for genomic segment  $i$ , and  $d^{(n)}_i(t)$  is calculated as the genetic divergence of sample  $n$  from its putative founder within segment  $i$  at time  $t$ . Each sample,  $n$ , from a persistent infection represents one measurement of  $d^{(n)}_i(t)$ . If a sample is collected at time  $t=t^*$  and has no mutant alleles within segment  $i$ , then  $d^{(n)}_i(t^*)=0$ . For each sample, the estimated start of infection is taken as  $t=0$ . Each sample from an individual acts as an independent observation of genetic distance for segment  $i$ . The error term  $\varepsilon^{(n)}_i(t)$  represents the residual unexplained variability in distance,  $d^{(n)}_i(t)$ , for sample  $n$ .

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

## 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), recorded 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 (detectable viral rebound: 34; no rebound: 460). Duration of infection is classed as a continuous variable ranging from 26 to 316 days per infection. **(b)** The same analysis as in (a) is performed for viral load dynamics specifically within a subset of 79 persistent infections with at least three PCR tests (rebound: 32; no rebound: 47). See Figure 1. **(c)** Comparing the BIC values for a subset of infections with durations lasting longer than 36 days (160 infections) and 56 days (75 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)**

Null model	-1662	-1641	-1694
Viral load dynamics	-1658	-1638	-1689

**(c)**

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 Table 3: Recurrent mutations found in four or more persistent infections and their known and/or predicted phenotypic impact.**

Nuc. mutation	Gene	AA mutation	n	Lineage	Description	Reversion*	Ref
A22775G	S	N405D	18	BA.2, BA.4, BA.5	Reversion of a lineage-defining mutation 405N to wild-type 405D; recurrent in chronic infections	Yes	(11)
T19587A	ORF1ab	T6441T	13	BA.2, BA.5	NSP14:T516; rarely found in the general population; has negative between-host fitness effect	No	(12,13)
T19183G	ORF1ab	C6307G	10	BA.2, BA.5	NSP14:C382G; mutation is within the N7-methyltransferase active domain which may impact mRNA stability and translation efficiency	No	(14)
A1045C	ORF1ab	A260A	9	BA.2, BA.4, BA.5	NSP2:A80A; rarely found in the general population; has negative between-host fitness effect	No	(12,13)
C29510A	N	R413S	9	BA.2, BA.5	Commonly found in all SARS-CoV-2 lineages prior to Omicron BA.2	Yes	(12)
T19557G	ORF1ab	F6431L	8	BA.2, BA.5	NSP14:F506L; mutation is within the N7-methyltransferase active domain and potentially impacts mRNA stability and efficient translation	No	(14)
T28105A	ORF8	I71N	8	BA.1, BA.2, BA.5, XBB	Residues 71 to 75 form assemblies that may mediate immune suppression activities	No	(15,16)
T22917G/A	S	L452R/Q/M	7	BA.2	Lineage-defining for BA.4, BA.5 and JN.1	No	(17)
T9534C	ORF1ab	I3090T	7	BA.2, BA.5	NSP4:I327T; commonly found in all SARS-CoV-2 lineages prior to Omicron BA.2	Yes	(12)
C23202A	S	T547K	6	BA.2, BA.5	Lineage-defining for BA.1	No	(17)
T8987C	ORF1ab	F2908L	5	BA.1, BA.2, BA.5	NSP4:F145L; rarely found in the general population; has negative between-host fitness effect	No	(12,13)
C11750T	ORF1ab	L3829F	5	Delta, BA.2, BA.5	NSP6:L260F; recurrent after Nirmatrelvir and Ritonavir treatment	No	(11,18)
C27509T	ORF7a	T39I	5	BA.1, BA.2, BA.5	Recurrent in persistent infections	No	(19)
T10492A/G	ORF1ab	G3409G	5	BA.1, BA.2, BA.5	NSP5:G146G; rarely found in the general population; has negative between-host fitness effect	No	(12,13)
C823T	ORF1ab	V186V	4	Alpha, Delta, BA.2, BA.5	NSP2:V6V; recurrent in chronic infections	No	(20)
C4829T	ORF1ab	Q1522*	4	Delta, BA.2	NSP3:Q704*; a premature stop codon mutation which can give rise to defective viral particles, modulate viral replication, and immune interactions	No	(21)
C5178A/T	ORF1ab	T1638I	4	BA.2, BA.5	NSP3:T820N/I; recurrent in chronic infections	No	(20)
T6573C	ORF1ab	F2103S	4	BA.2	NSP3:F1285S; commonly found in BA.2	Yes	(12)
T8991C	ORF1ab	V2909A	4	BA.2	NSP4:V146A; commonly found in Delta and BA.2	Yes	(12)
C10369T	ORF1ab	R3368R	4	Delta, BA.2, BA.5	NSP5:R105R; recurrent in chronic infections	No	(20)
C11454T	ORF1ab	A3730V	4	Delta, BA.1, BA.2	NSP6:A161V; found in many sequences from Alpha and Delta	No	(12)
C12439T	ORF1ab	P4058P	4	BA.2, BA.5	NSP8:P116P; found in many sequences from Delta	No	(12)
T17978C	ORF1ab	L5905F	4	BA.1, BA.2, BA.5	NSP13:L581P; found in many sequences from Delta	No	(12)
G21701A	S	V47I	4	BA.1, BA.2, BA.5	Increased ACE2 binding in BA.2	No	(22,23)
A22629C	S	K356T	4	BA.2, BA.5	Lineage-defining for JN.1	No	(17)
C24213T	S	S884F	4	BA.2, BA.4, BA.5	Increased ACE2 binding in BA.2	No	(22,23)
C24588T	S	T1009I	4	Alpha, BA.2	Found in many BF.1 (sublineage of BA.5) sequences	No	(12)
C26408T	E	S55F	4	BA.1, BA.2, BA.5	Found in many sequences from Delta and BA.2	No	(12)
C28724T	N	P151S	4	Delta, BA.1, BA.5	Known to be under multilevel selection	No	(24)

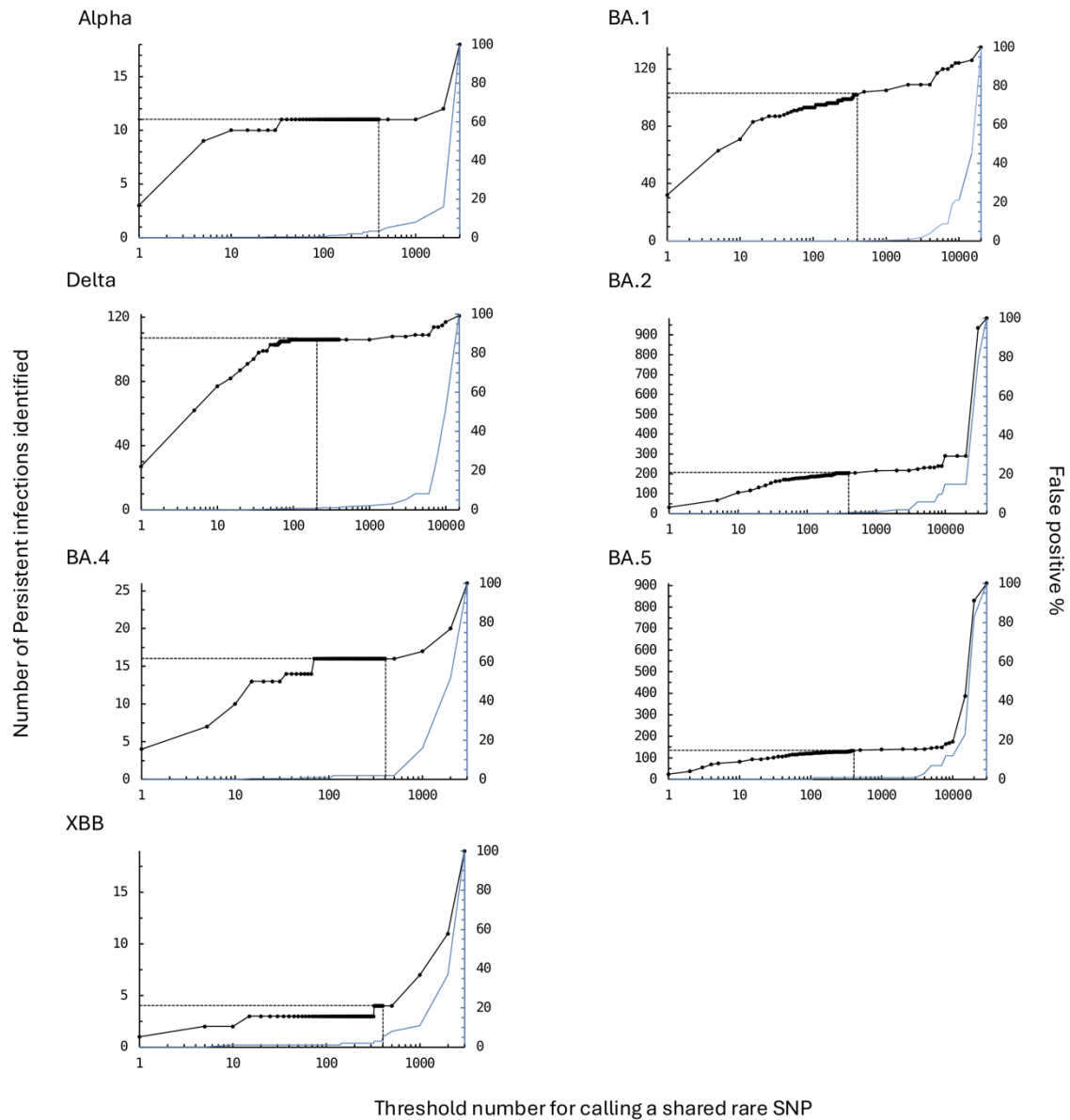
\*De novo mutations during persistent infections that reverted the consensus nucleotide at baseline (which differed from the Wuhan-Hu reference) back to the Wuhan-Hu reference sequence (NC\_045512.2).

**Supplementary Table 4: Recurrent mutations found in three or more persistent infections and reported to have drug-resistance properties.**

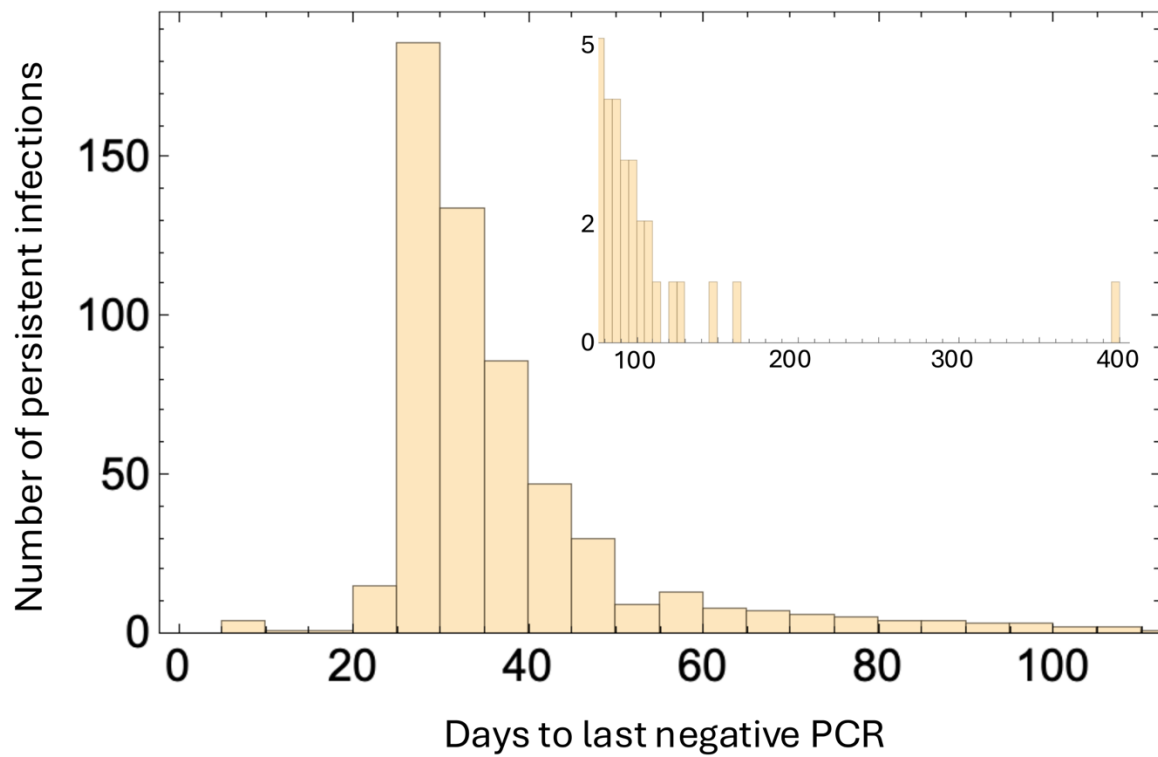
Nuc. mutation	Gene	AA mutation	n	Lineage	Description	Reversion*	Ref
A22628C A22629C/G	S	K356Q/P K356T/R	7	BA.2, BA.5	S:K356T is associated with sotrovimab resistance	No	(25–27)
C11750T	NSP6	L260F	5	Delta, BA.2, BA.5	NSP6:L260F frequently occurs in patients treated with Nirmatrelvir and Ritonavir	No	(18)
G22899T/A G22898A	S	G446V/D G446S	5	Delta, BA.2, BA.5	S:G446V is associated with resistance to Casirivimab and Imdevimab. G446S is a molnupiravir-associated mutation	No	(28,29)
G22580A A22581T A22582C	S	E340K E340V E340D	5	BA.1, BA.2, BA.5	S:E340K/V/D is associated with sotrovimab resistance. E340K is also a molnupiravir-associated mutation	No	(27–29)
C21588T/A C21587A	S	P9L/Q P9T	5	Delta, BA.2	S:P9L is a molnupiravir-associated mutation	No	(30)
C10965T	NSP5	T304I	3	Delta, BA.1, BA.5	NSP5:T304I is associated with nirmatrelvir resistance	No	(31,32)
G25019T A25020G	S	D1153Y D1153G	3	BA.2, BA.5	S:D1153Y is known to escape CC9.104 and CC67.105 antibodies in BA.2	No	(23)

\*De novo mutations during persistent infections that reverted the consensus nucleotide at baseline (which differed from the Wuhan-Hu reference) back to the Wuhan-Hu reference sequence (NC\_045512.2).

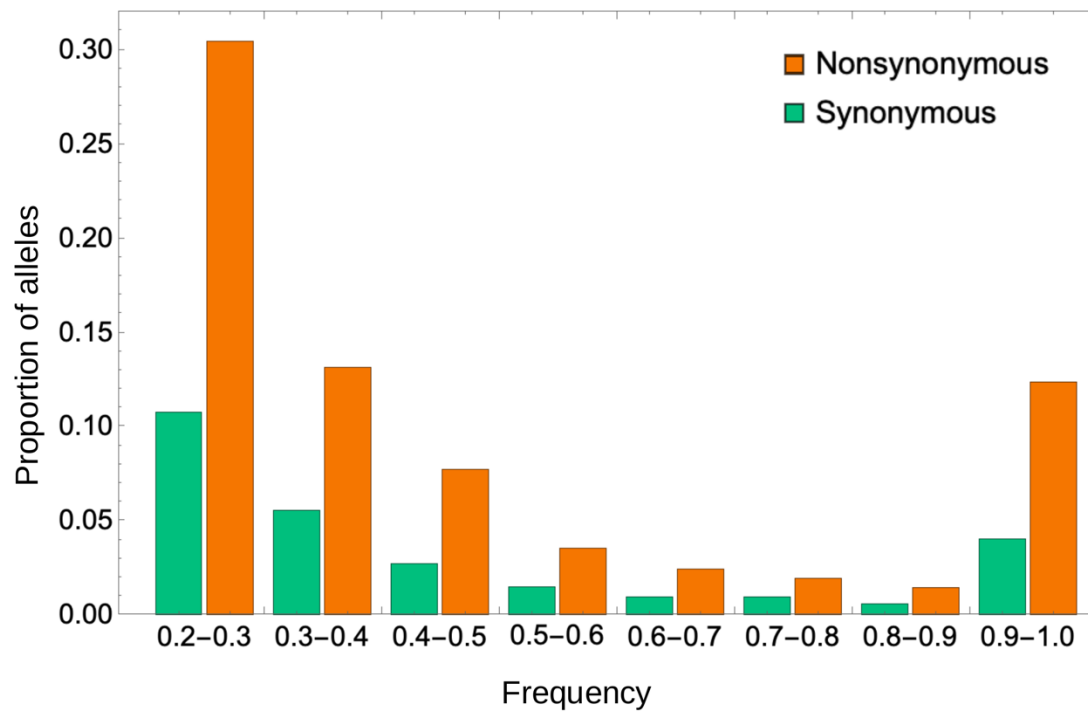
## Supplementary Figures



**Supplementary Figure 1: 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 study (5), 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.

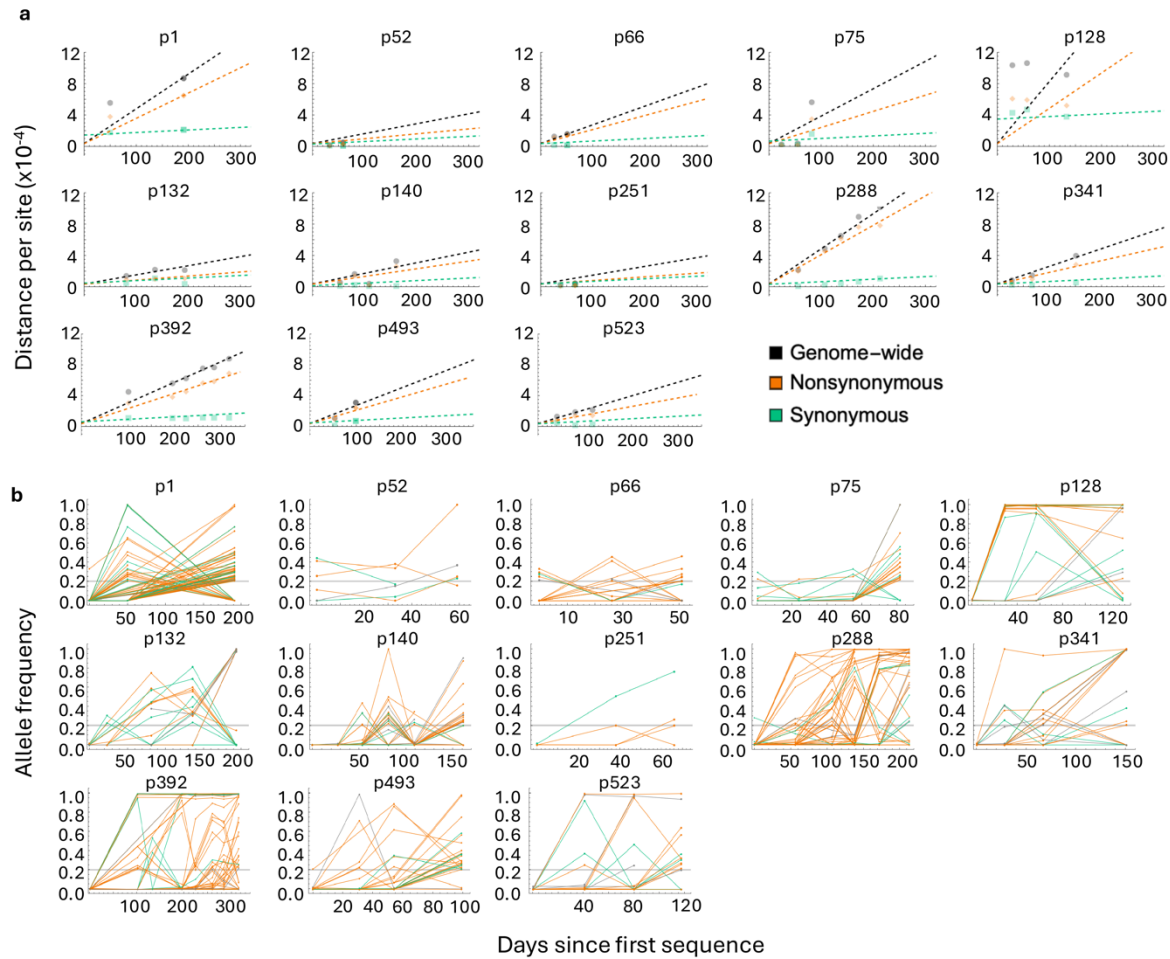


**Supplementary Figure 2: Number of days elapsed since the last time a persistently infected individual had a negative PCR test.** The histogram plot includes all 576 identified persistent infections.

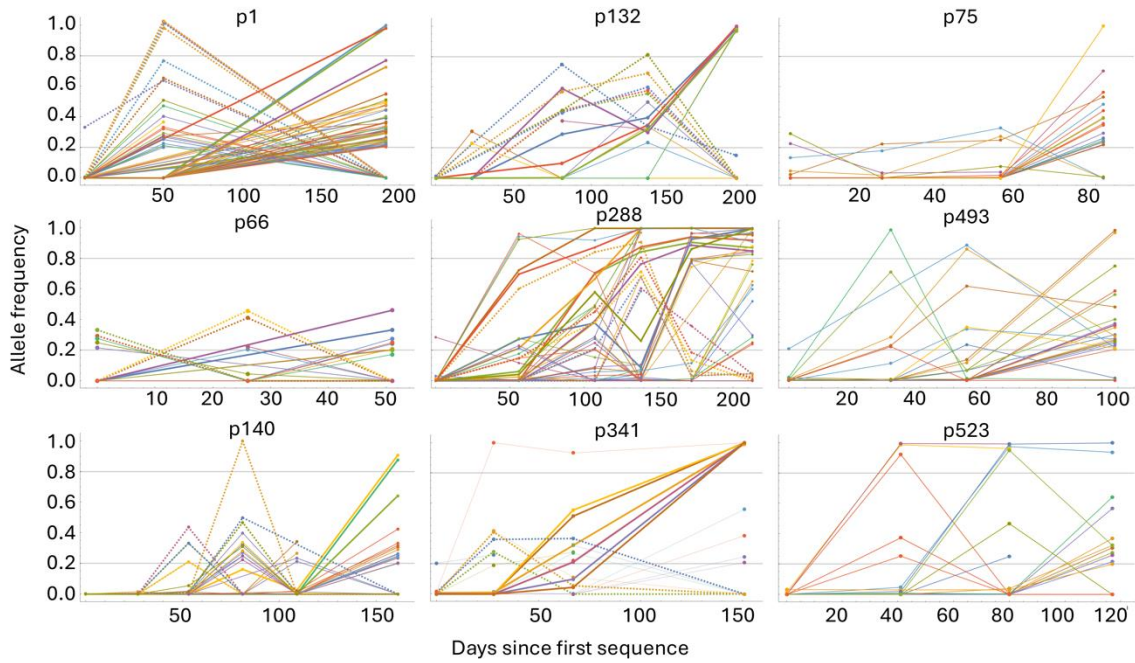


**Supplementary Figure 3: Site frequency spectrum.** Proportion of synonymous (green) and nonsynonymous (orange) mutations in persistent infections across all frequency bands.

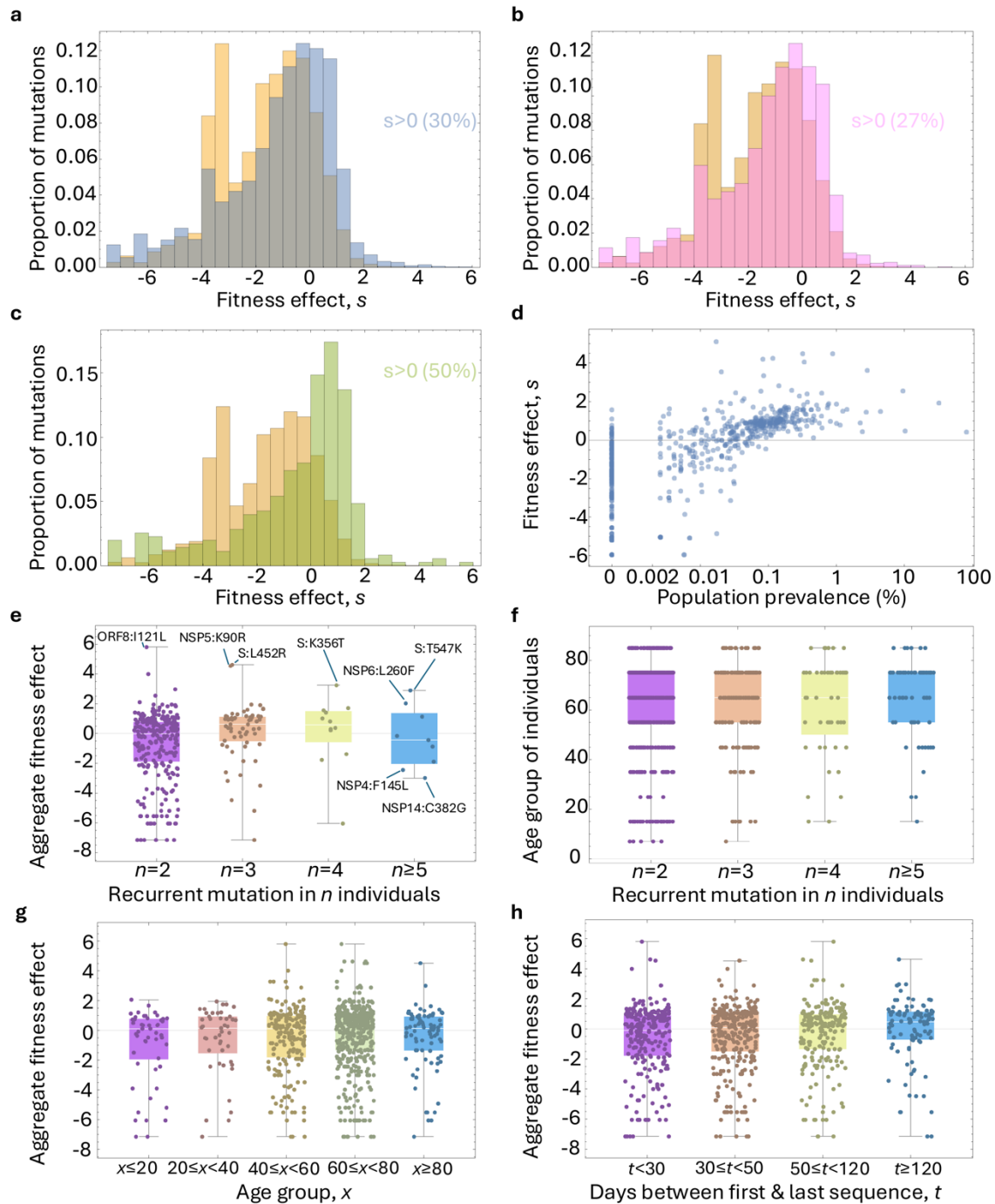




**Supplementary Figure 4: Rates of genome-wide, nonsynonymous, and synonymous evolution in 13 persistently infected individuals.** (a) Illustrates the evolutionary distance over time for a subset of 13 persistently infected individuals, each characterised by a minimum of three temporal data points and 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. (c) 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.

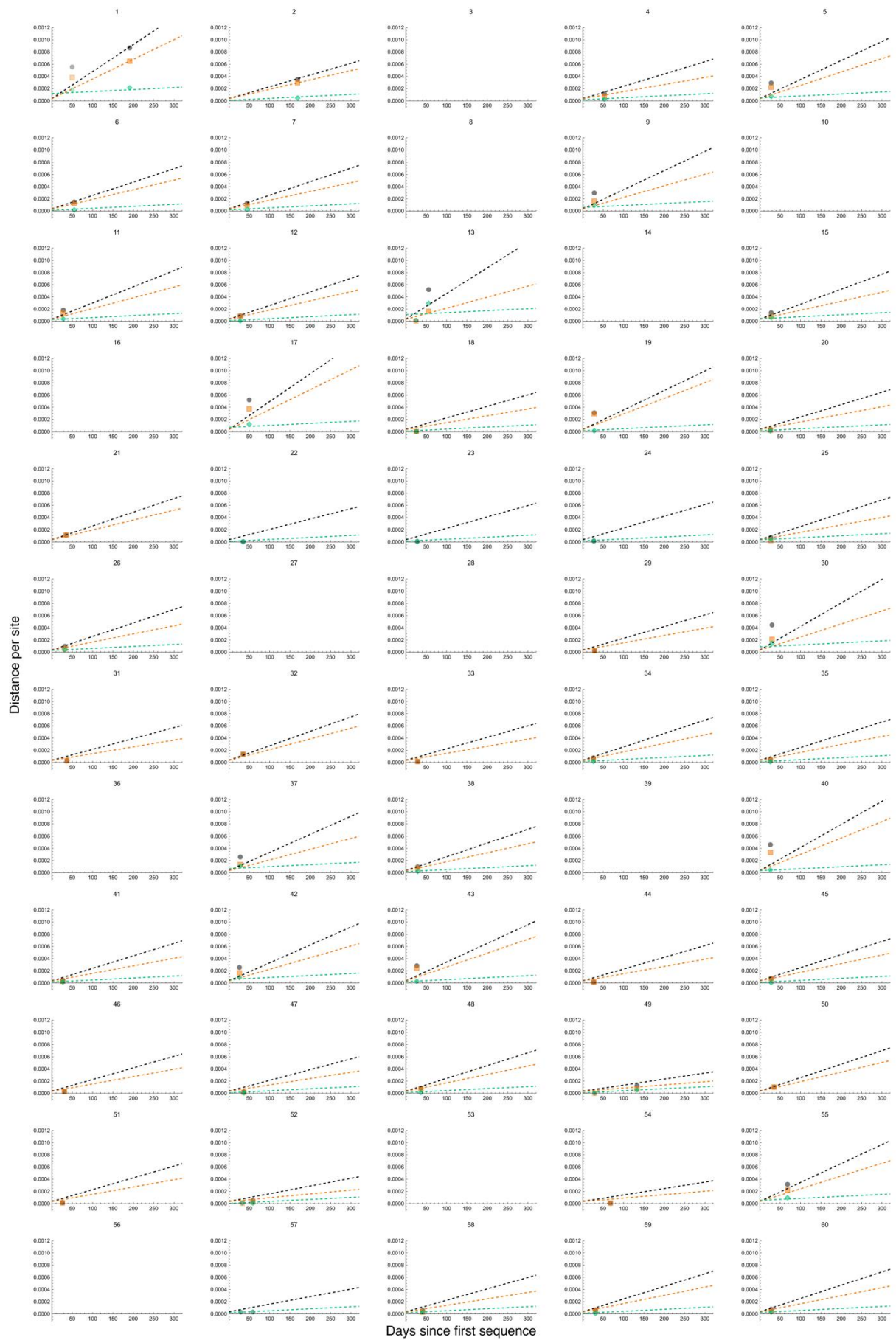


**Supplementary Figure 5: Temporal allele frequency dynamics in nine persistent infections.** The figure illustrates two distinct patterns of allele dynamics over time. In the left column (infections p1, p66, and p140), we observe transient allele groups that emerge at one time point, with some reaching 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 individuals (5). 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.

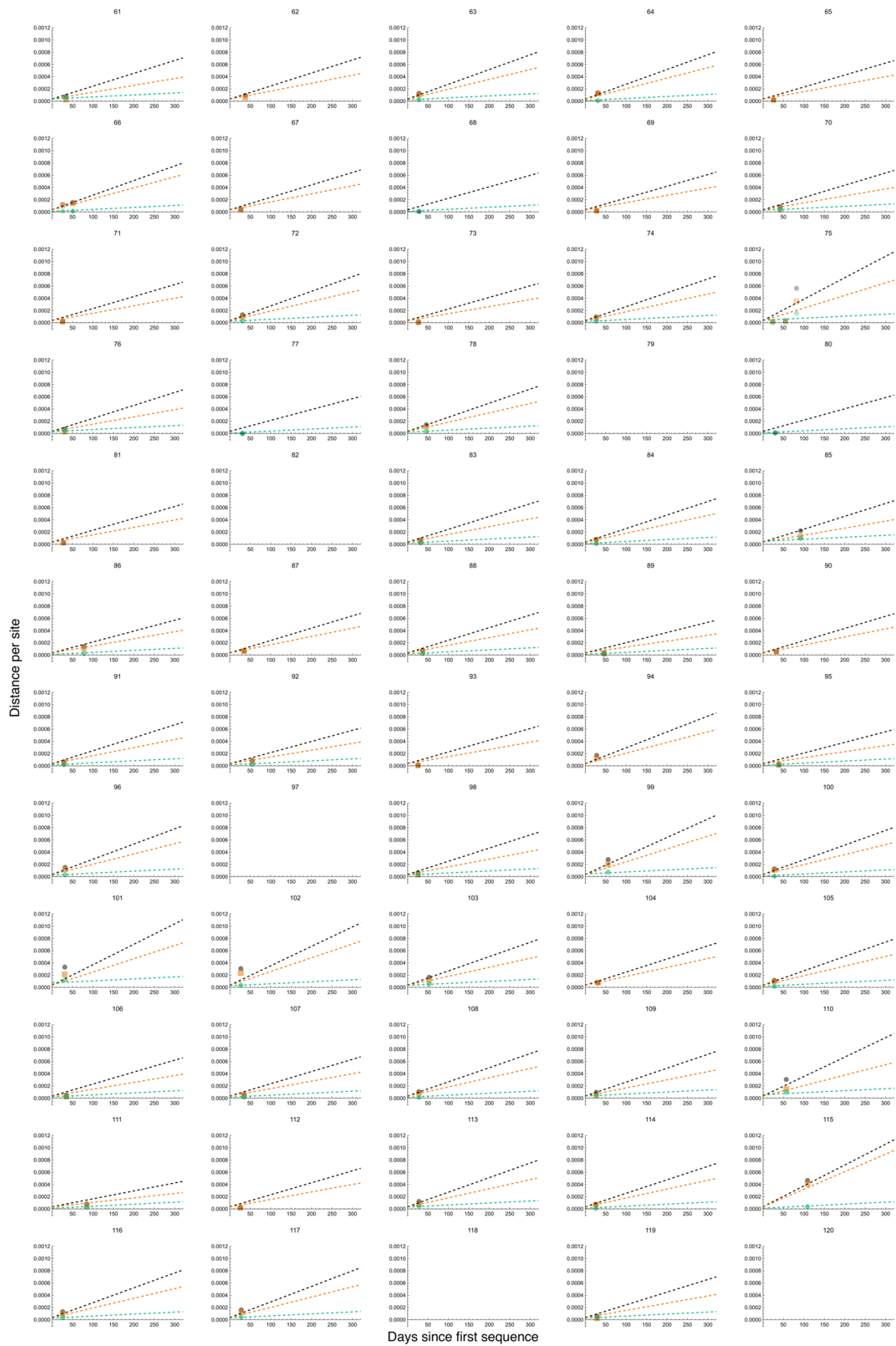


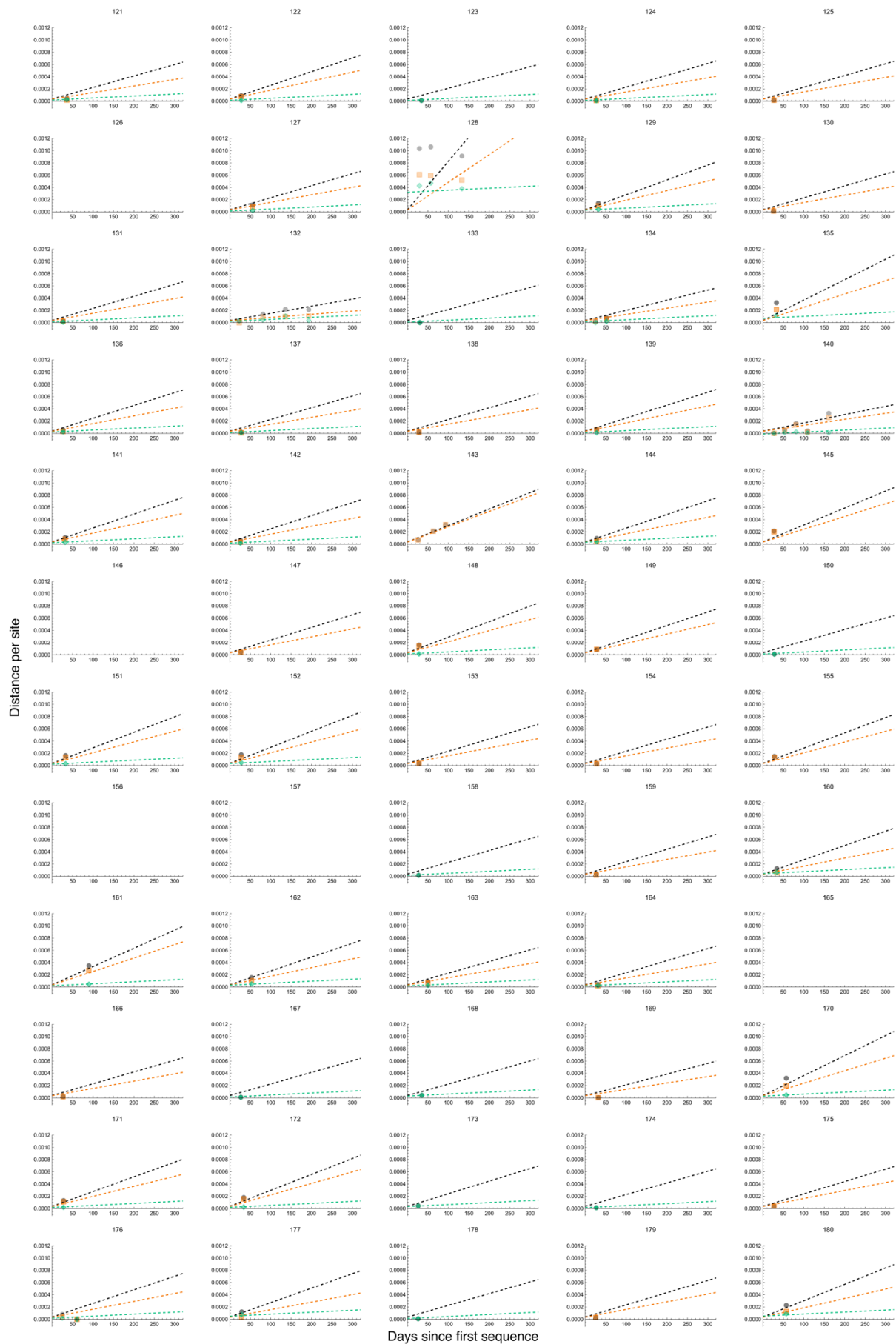
**Supplementary Figure 6: Between-host fitness effect and prevalence of mutations identified in persistently-infected individuals.** (a) Distribution of between-host fitness effects of all SARS-CoV-2 mutations on a global phylogeny (orange), between-host fitness of all mutations found in persistently 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  $n$

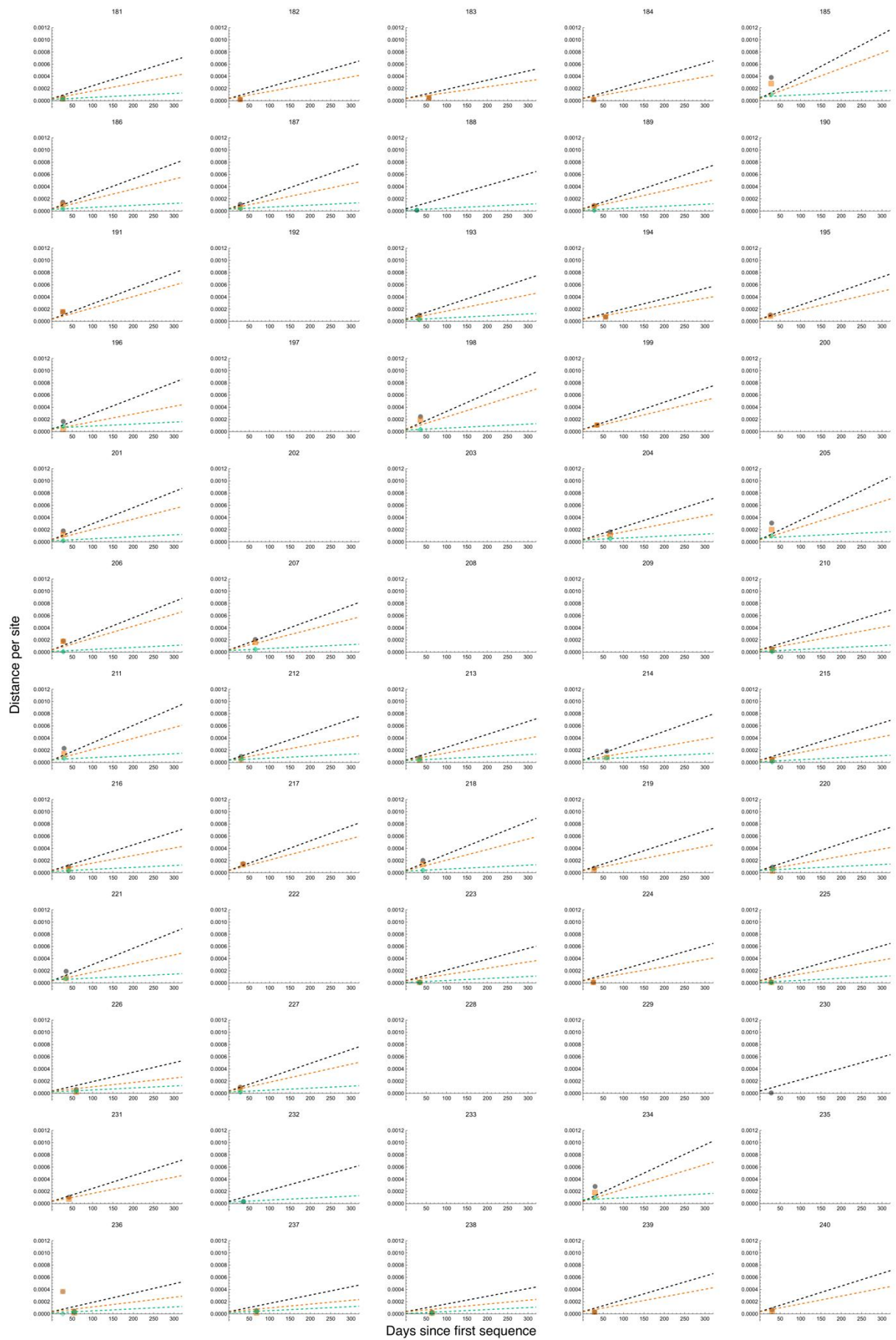
persistent infections. Some of the mutations with extremely high and low fitness effects are highlighted. **(f)** Age-group of all individuals which share  $n$  recurrent mutations. **(g)** Aggregate fitness effect of recurrent mutations per age group. **(h)** Aggregate fitness effect of recurrent mutations 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.csv](https://github.com/jbloomlab/SARS2-mut-fitness/blob/main/results_public_2024-04-19/nt_fitness/ntmut_fitness_by_clade.csv) (13).





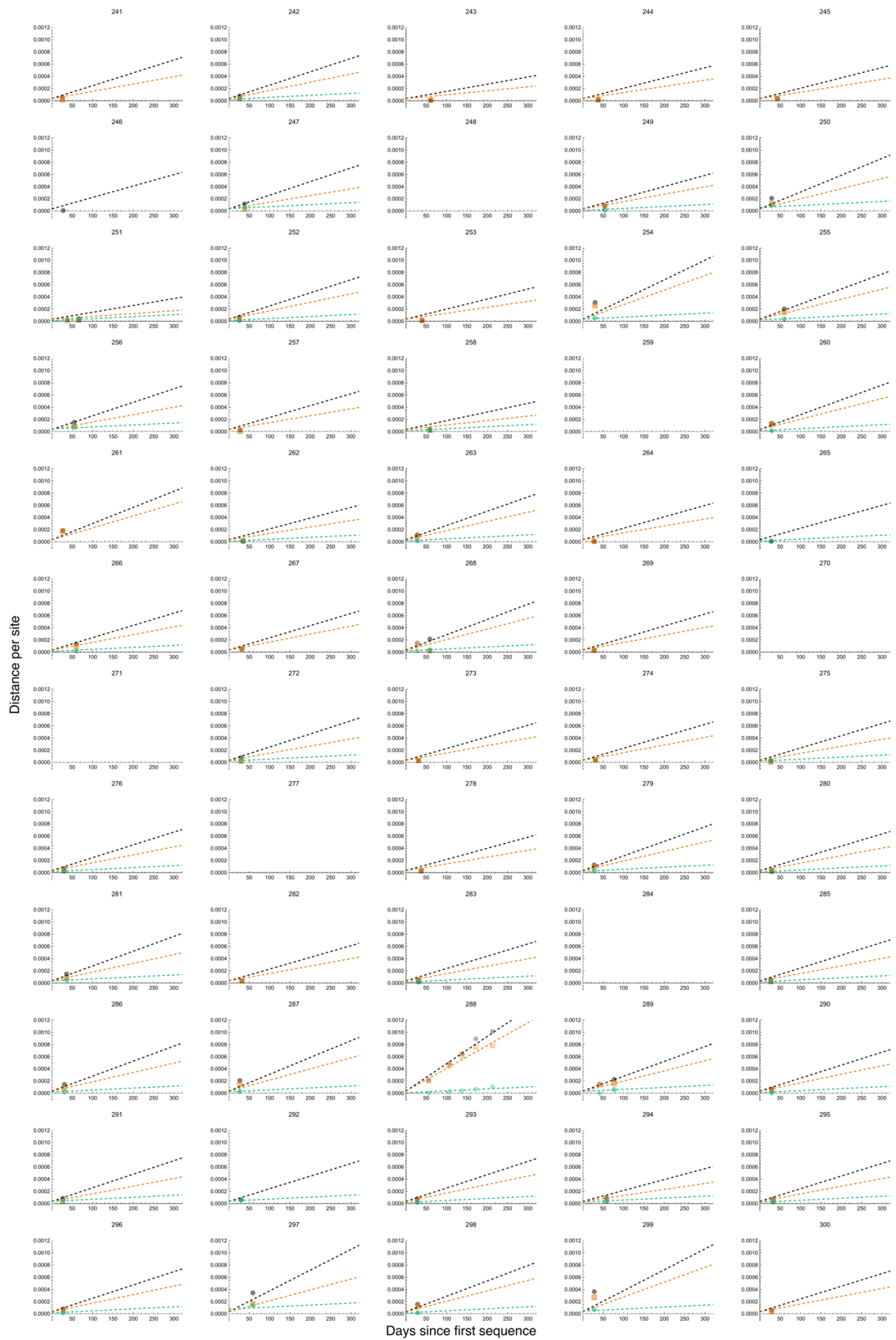


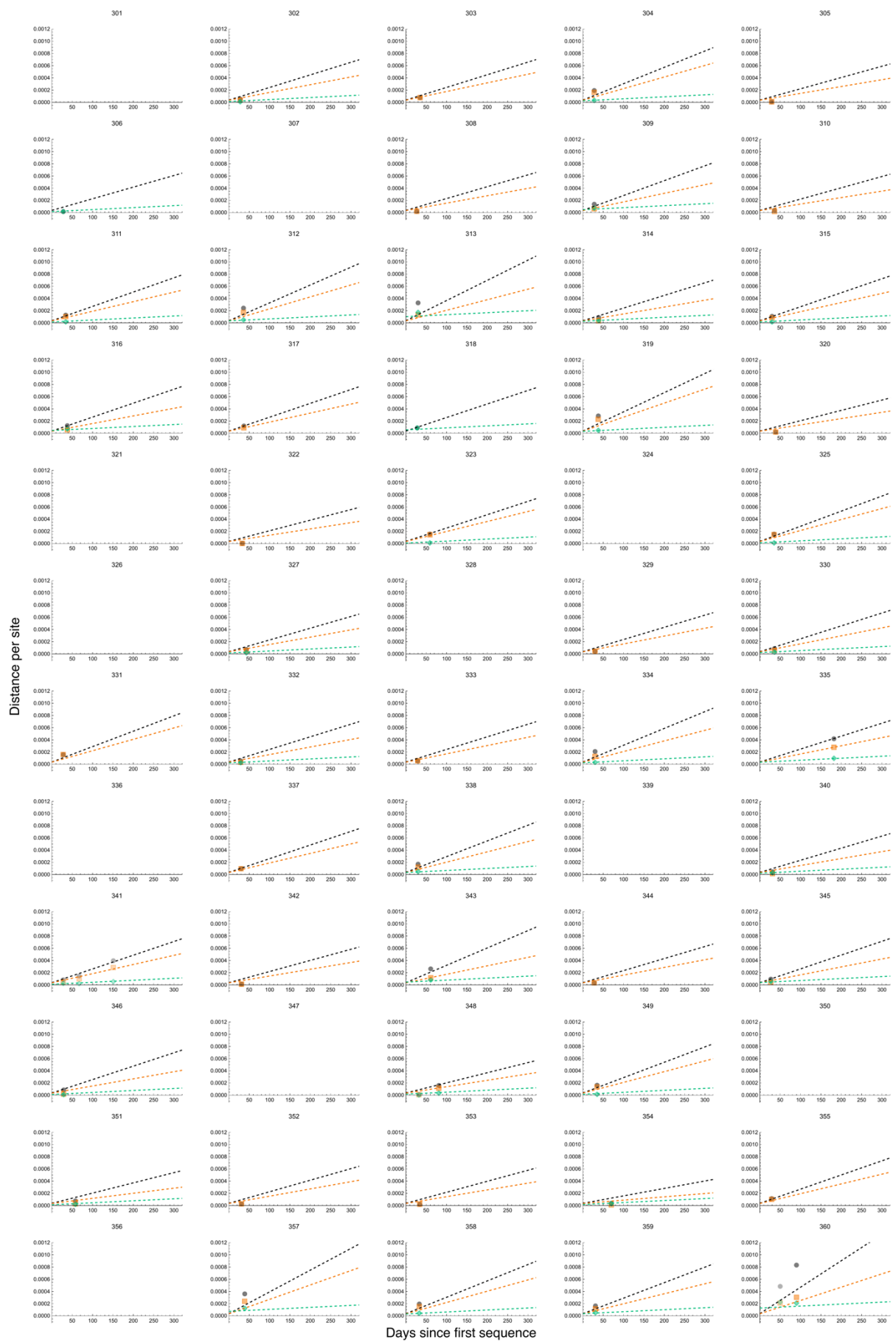


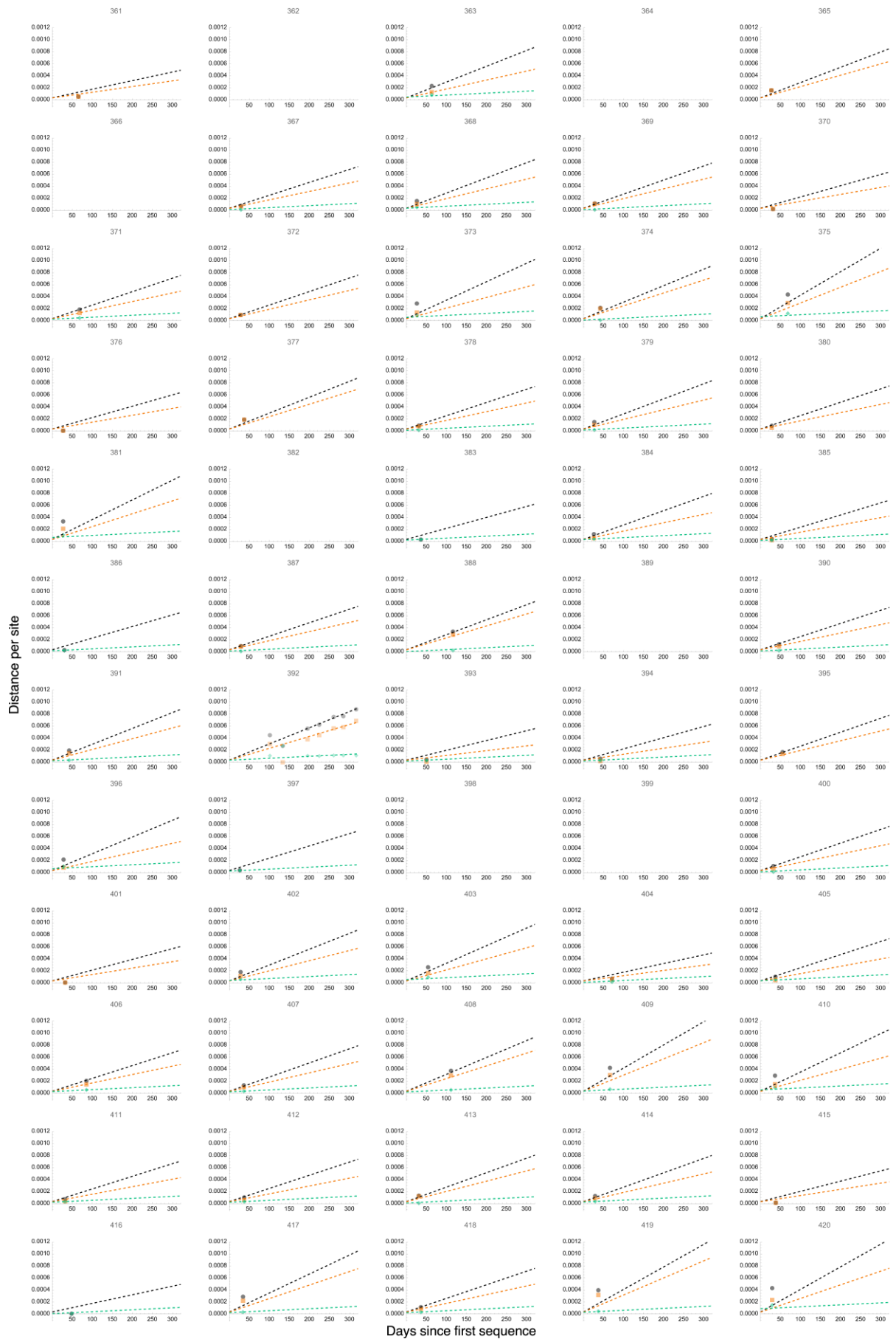


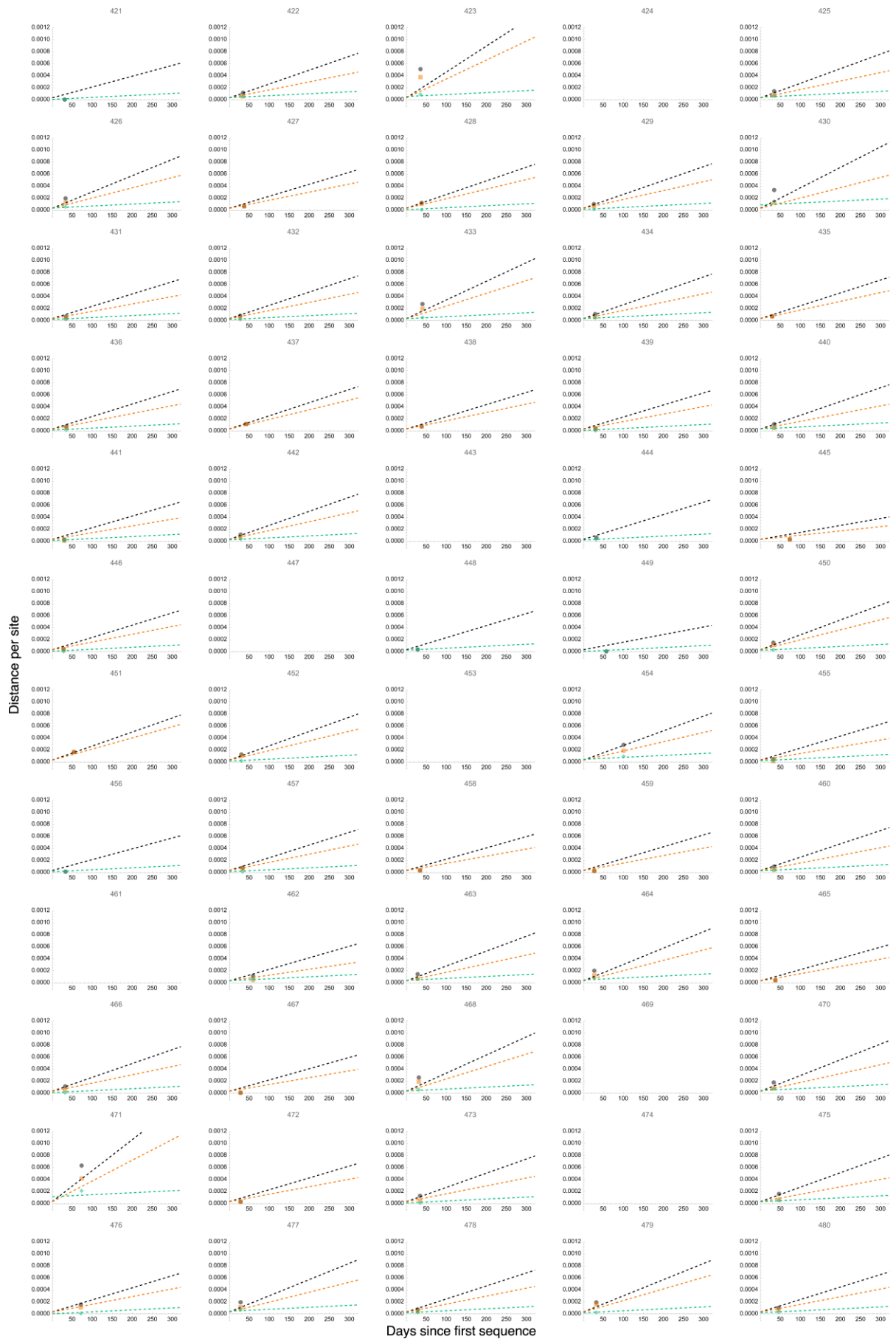
Days since first sequence

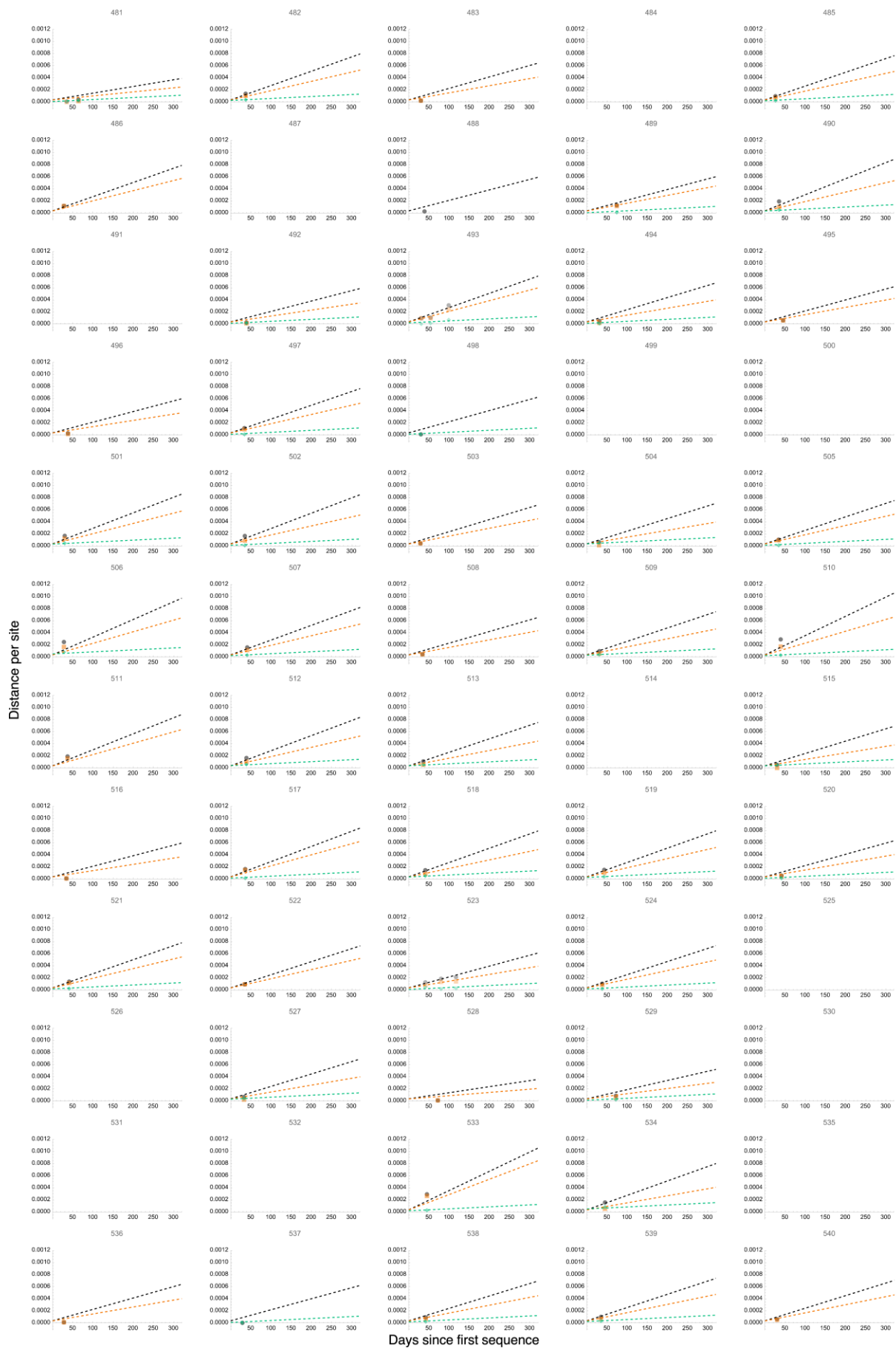


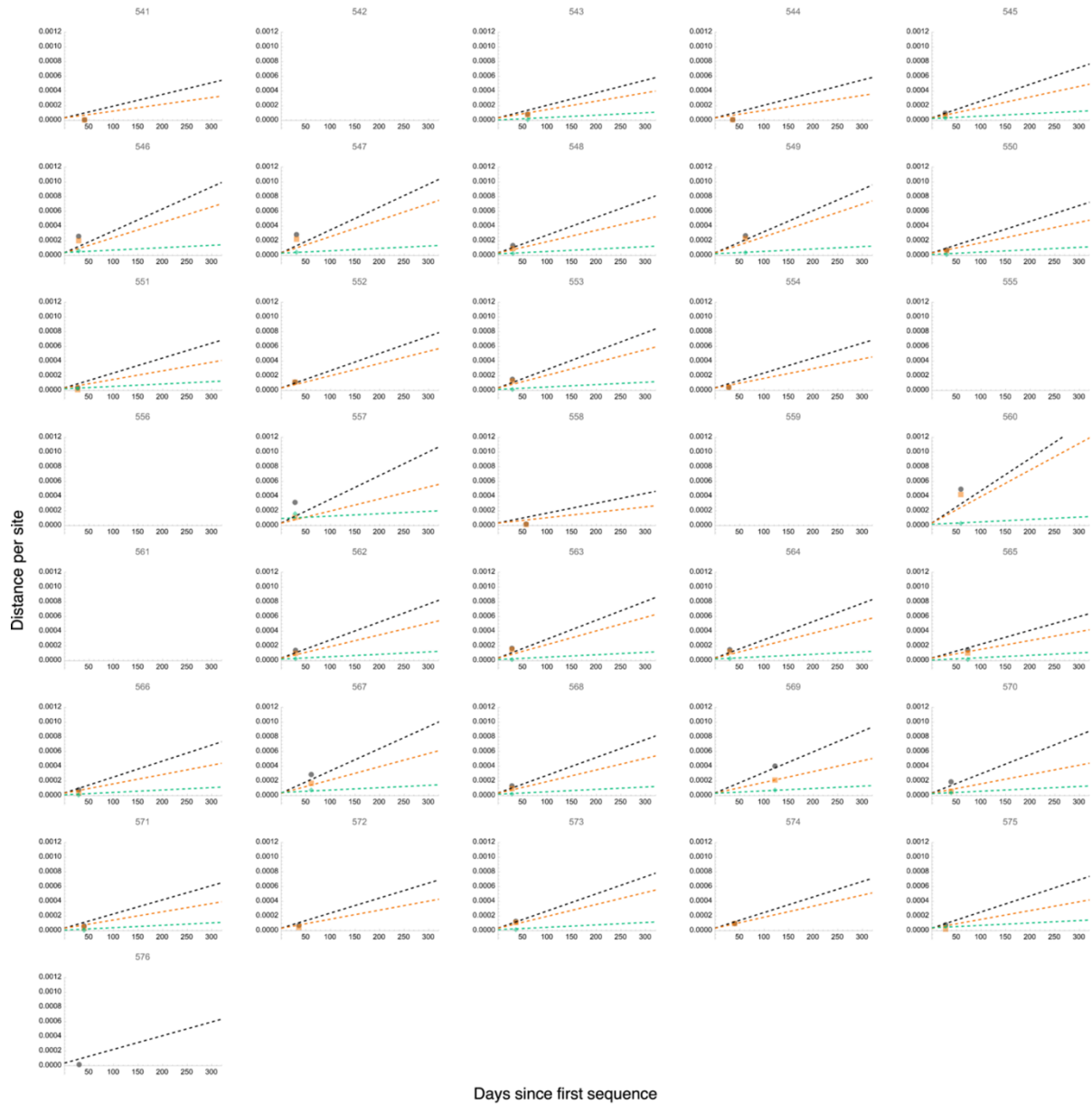












**Supplementary Figure 7: Rates of genome-wide, nonsynonymous, and synonymous evolution in all persistently infected individuals with measurable rates.** The evolutionary distance over time for 494 persistently infected individuals with measurable genome-wide rate (black), 457 nonsynonymous rate (orange), 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.



## References

1. 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 [Internet]. Github; [cited 2023 Jan 18]. Available from: <https://github.com/connor-lab/ncov2019-artic-nf>
2. Lythgoe KA, Golubchik T, Hall M, House T, Cahuantzi R, MacIntyre-Cockett G, et al. Lineage replacement and evolution captured by 3 years of the United Kingdom Coronavirus (COVID-19) Infection Survey. *Proc Biol Sci* [Internet]. 2023 Oct 25;290(2009). Available from: <http://dx.doi.org/10.1098/rspb.2023.1284>
3. Bonsall D, Golubchik T, de Cesare M, Limbada M, Kosloff B, MacIntyre-Cockett G, et al. A comprehensive genomics solution for HIV surveillance and clinical monitoring in low-income settings. *J Clin Microbiol* [Internet]. 2020 Sep 22;58(10). Available from: <http://dx.doi.org/10.1128/jcm.00382-20>
4. Wymant C, Blanquart F, Golubchik T, Gall A, Bakker M, Bezemer D, et al. Easy and accurate reconstruction of whole HIV genomes from short-read sequence data with shiver. *Virus Evol.* 2018 Jan 1;4(1):vey007.
5. Ghafari M, Hall M, Golubchik T, Ayoubkhani D, House T, MacIntyre-Cockett G, et al. Prevalence of persistent SARS-CoV-2 in a large community surveillance study. *Nature*. 2024 Feb 29;626(8001):1094–101.
6. Zhao L, Illingworth CJR. Measurements of intrahost viral diversity require an unbiased diversity metric. *Virus Evol.* 2019 Jan 1;5(1):vey041.
7. Lumby CK, Zhao L, Breuer J, Illingworth CJR. A large effective population size for established within-host influenza virus infection. *Elife*. 2020 Aug 10;9:e56915.
8. Sagulenko P, Puller V, Neher RA. TreeTime: Maximum-likelihood phylodynamic analysis. *Virus Evol.* 2018 Jan 1;4(1):vex042.
9. Druelle V, Neher RA. Reversions to consensus are positively selected in HIV-1 and bias substitution rate estimates. *Virus Evol.* 2023 Jan 9;9(1):veac118.
10. Raghwani J, Redd AD, Longosz AF, Wu CH, Serwadda D, Martens C, et al. Evolution of HIV-1 within untreated individuals and at the population scale in Uganda. *PLoS Pathog.* 2018 Jul;14(7):e1007167.
11. Ip JD, Chu WM, Chan WM, Chu AWH, Leung RCY, Peng Q, et al. The significance of recurrent de novo amino acid substitutions that emerged during chronic SARS-CoV-2 infection: an observational study. *EBioMedicine*. 2024 Sep;107(105273):105273.
12. Chen C, Nadeau S, Yared M, Voinov P, Xie N, Roemer C, et al. CoV-Spectrum: analysis of globally shared SARS-CoV-2 data to identify and characterize new variants. *Bioinformatics*. 2022 Mar 4;38(6):1735–7.
13. Bloom JD, Neher RA. Fitness effects of mutations to SARS-CoV-2 proteins. *Virus Evol.* 2024 Mar 27;9(2):vead055.
14. Baddock HT, Broluh S, Yosaatmadja Y, Ratnaweera M, Bielinski M, Swift LP, et al. Characterization of the SARS-CoV-2 ExoN (nsp14ExoN-nsp10) complex: implications for its role in viral genome stability and inhibitor identification. *Nucleic Acids Res.* 2022 Feb 22;50(3):1484–500.
15. Zhang Y, Chen Y, Li Y, Huang F, Luo B, Yuan Y, et al. The ORF8 protein of SARS-CoV-2 mediates immune evasion through down-regulating MHC-I. *Proc Natl Acad Sci U S A*. 2021 Jun 8;118(23):e2024202118.

16. Flower TG, Buffalo CZ, Hooy RM, Allaire M, Ren X, Hurley JH. Structure of SARS-CoV-2 ORF8, a rapidly evolving immune evasion protein. *Proc Natl Acad Sci U S A*. 2021 Jan 12;118(2):e2021785118.
17. Hodcroft EB. CoVariants: SARS-CoV-2 Mutations and Variants of Interest [Internet]. 2021 [cited 2025 Jan 14]. Available from: <https://covariants.org/>
18. Carlin AF, Clark AE, Chaillon A, Garretson AF, Bray W, Porrachia M, et al. Virologic and immunologic characterization of Coronavirus disease 2019 recrudescence after nirmatrelvir/ritonavir treatment. *Clin Infect Dis*. 2023 Feb 8;76(3):e530–2.
19. Gonzalez-Reiche AS, Alshammary H, Schaefer S, Patel G, Polanco J, Carreño JM, et al. Sequential intrahost evolution and onward transmission of SARS-CoV-2 variants. *Nat Commun*. 2023 Jun 3;14(1):3235.
20. Harari S, Tahor M, Rutsinsky N, Meijer S, Miller D, Henig O, et al. Drivers of adaptive evolution during chronic SARS-CoV-2 infections. *Nat Med*. 2022 Jul;28(7):1501–8.
21. Zhao L, Lythgoe KA. The social role of defective viral genomes in chronic viral infections: a commentary on Leeks et al. 2023. *J Evol Biol*. 2023 Nov;36(11):1577–81.
22. Dadonaite B, Brown J, McMahon TE, Farrell AG, Figgins MD, Asarnow D, et al. Spike deep mutational scanning helps predict success of SARS-CoV-2 clades. *Nature*. 2024 Jul;631(8021):617–26.
23. Dadonaite B, Crawford KHD, Radford CE, Farrell AG, Yu TC, Hannon WW, et al. A pseudovirus system enables deep mutational scanning of the full SARS-CoV-2 spike. *Cell*. 2023 Mar 16;186(6):1263–78.e20.
24. Vinicius Bonetti Franceschi EV. Phylogenetic signatures reveal multilevel selection and fitness costs in SARS-CoV-2 [Internet]. [cited 2025 Jan 14]. Available from: <https://wellcomeopenresearch.org/articles/9-85/v2>
25. Huygens, Sammy, Corine GeurtsvanKessel, Arvind Gharbharan, Susanne Bogers, Nathalie Worp, Marjan Boter, Hannelore I. Bax et al. Clinical and Virological Outcome of Monoclonal Antibody Therapies Across SARS-CoV-2 Variants in 245 Immunocompromised Patients: A Multicenter Prospective Cohort Study. *Clinical Infectious Diseases*. 78:1514–21.
26. Vellas C, Trémeaux P, Del Bello A, Latour J, Jeanne N, Ranger N, et al. Resistance mutations in SARS-CoV-2 omicron variant in patients treated with sotrovimab. *Clin Microbiol Infect*. 2022 Sep;28(9):1297–9.
27. Ragonnet-Cronin M, Nutalai R, Huo J, Dijokaite-Guraliuc A, Das R, Tuekprakhon A, et al. Generation of SARS-CoV-2 escape mutations by monoclonal antibody therapy. *Nat Commun*. 2023 Jun 7;14(1):3334.
28. Huygens S, GeurtsvanKessel C, Gharbharan A, Bogers S, Worp N, Boter M, et al. Clinical and virological outcome of monoclonal antibody therapies across SARS-CoV-2 variants in 245 immunocompromised patients: A multicenter prospective cohort study. *Clin Infect Dis*. 2024 Jun 14;78(6):1514–21.
29. Rockett R, Basile K, Maddocks S, Fong W, Agius JE, Johnson-Mackinnon J, et al. Resistance mutations in SARS-CoV-2 delta variant after sotrovimab use. *N Engl J Med*. 2022 Apr 14;386(15):1477–9.
30. Sanderson T, Hisner R, Donovan-Banfield I 'ah, Hartman H, Løchen A, Peacock TP, et al. A molnupiravir-associated mutational signature in global SARS-CoV-2 genomes. *Nature*. 2023 Nov;623(7987):594–600.
31. Iketani S, Mohri H, Culbertson B, Hong SJ, Duan Y, Luck MI, et al. Multiple pathways for



SARS-CoV-2 resistance to nirmatrelvir. *Nature*. 2023 Jan;613(7944):558–64.

32. Ip JD, Wing-Ho Chu A, Chan WM, Cheuk-Ying Leung R, Umer Abdullah SM, Sun Y, et al. Global prevalence of SARS-CoV-2 3CL protease mutations associated with nirmatrelvir or ensitrelvir resistance. *EBioMedicine*. 2023 May;91(104559):104559.

## **The COVID-19 Infection Survey Group**

Tina Thomas<sup>1</sup>, Dawid Pienaar<sup>1</sup>, Joy Preece<sup>1</sup>, Sarah Crofts<sup>1</sup>, Lina Lloyd<sup>1</sup>, Michelle Bowen<sup>1</sup>, Russell Black<sup>1</sup>, Antonio Felton<sup>1</sup>, Megan Crees<sup>1</sup>, Joel Jones<sup>1</sup>, Esther Sutherland<sup>1</sup>, Derrick W. Crook<sup>2</sup>, Emma Pritchard<sup>2</sup>, Karina-Doris Vihta<sup>2</sup>, Alison Howarth<sup>2</sup>, Brian D. Marsden<sup>2</sup>, Kevin K. Chau<sup>2</sup>, Lucas Martins Ferreira<sup>2</sup>, Wanwisa Dejnirattisai<sup>2</sup>, Juthathip Mongkolsapaya<sup>2</sup>, Sarah Hoosdally<sup>2</sup>, Richard Cornall<sup>2</sup>, David I Stuart<sup>2</sup>, Gavin Screaton<sup>2</sup>, John N Newton<sup>3</sup>, John I Bell<sup>4</sup>, Stuart Cox<sup>5</sup>, Kevin Paddon<sup>5</sup>, Tim James<sup>5</sup>, Julie Robotham<sup>6</sup>, Paul Birrell<sup>6</sup>, Helena Jordan<sup>7</sup>, Tim Sheppard<sup>7</sup>, Graham Athey<sup>7</sup>, Dan Moody<sup>7</sup>, Leigh Curry<sup>7</sup>, Pamela Brereton<sup>7</sup>, Ian Jarvis<sup>8</sup>, Anna Godsmark<sup>8</sup>, George Morris<sup>8</sup>, Bobby Mallick<sup>8</sup>, Phil Eeles<sup>8</sup>, Jodie Hay<sup>9</sup>, Harper VanSteenhouse<sup>9</sup>, Jessica Lee<sup>10</sup>, Sean White<sup>11</sup>, Tim Evans<sup>11</sup>, Lisa Bloemberg<sup>11</sup>, Katie Allison<sup>12</sup>, Anouska Pandya<sup>12</sup>, Sophie Davis<sup>12</sup>, David I Conway<sup>13</sup>, Margaret MacLeod<sup>13</sup>, Chris Cunningham<sup>13</sup>

<sup>1</sup> Office for National Statistics, Newport, UK.

<sup>2</sup> Nuffield Department of Medicine, University of Oxford, Oxford, UK.

<sup>3</sup> Office for Health Improvement and Disparities, London, UK

<sup>4</sup> Office of the Regius Professor of Medicine, University of Oxford, Oxford, UK

<sup>5</sup> Oxford University Hospitals NHS Foundation Trust, Oxford, UK

<sup>6</sup> UK Health Security Agency, London, UK

<sup>7</sup> IQVIA, London, UK

<sup>8</sup> National Biocentre, Milton Keynes, UK.

<sup>9</sup> Glasgow Lighthouse Laboratory, London, UK

<sup>10</sup> Department of Health and Social Care, London, UK

<sup>11</sup> Welsh Government, Cardiff, UK

<sup>12</sup> Scottish Government, Edinburgh, UK

<sup>13</sup> Public Health Scotland, Edinburgh, UK

## **The COVID-19 Genomics UK (COG-UK) Consortium**

<https://www.cogconsortium.uk>

**Funding acquisition, Leadership and supervision, Metadata curation, Project administration, Samples and logistics, Sequencing and analysis, Software and analysis tools, and Visualisation:**

Dr Samuel C Robson PhD <sup>13, 84</sup>

**Funding acquisition, Leadership and supervision, Metadata curation, Project administration, Samples and logistics, Sequencing and analysis, and Software and analysis tools:**

Dr Thomas R Connor PhD <sup>11, 74</sup> and Prof Nicholas J Loman PhD <sup>43</sup>

**Leadership and supervision, Metadata curation, Project administration, Samples and logistics, Sequencing and analysis, Software and analysis tools, and Visualisation:**

Dr Tanya Golubchik PhD <sup>5</sup>

**Funding acquisition, Leadership and supervision, Metadata curation, Samples and logistics, Sequencing and analysis, and Visualisation:**

Dr Rocio T Martinez Nunez PhD <sup>46</sup>

**Funding acquisition, Leadership and supervision, Project administration, Samples and logistics, Sequencing and analysis, and Software and analysis tools:**

Dr David Bonsall PhD <sup>5</sup>

**Funding acquisition, Leadership and supervision, Project administration, Sequencing and analysis, Software and analysis tools, and Visualisation:**

Prof Andrew Rambaut DPhil <sup>104</sup>

**Funding acquisition, Metadata curation, Project administration, Samples and logistics, Sequencing and analysis, and Software and analysis tools:**

Dr Luke B Snell MSc, MBBS <sup>12</sup>

**Leadership and supervision, Metadata curation, Project administration, Samples and logistics, Software and analysis tools, and Visualisation:**

Rich Livett MSc <sup>116</sup>

**Funding acquisition, Leadership and supervision, Metadata curation, Project administration, and Samples and logistics:**

Dr Catherine Ludden PhD <sup>20, 70</sup>

**Funding acquisition, Leadership and supervision, Metadata curation, Samples and logistics, and Sequencing and analysis:**

Dr Sally Corden PhD <sup>74</sup> and Dr Eleni Nastouli FRCPATH <sup>96, 95, 30</sup>

**Funding acquisition, Leadership and supervision, Metadata curation, Sequencing and analysis, and Software and analysis tools:**

Dr Gaia Nebbia PhD, FRCPATH <sup>12</sup>

**Funding acquisition, Leadership and supervision, Project administration, Samples and logistics, and Sequencing and analysis:**

**Leadership and supervision, Metadata curation, Project administration, Samples and logistics, and Sequencing and analysis:**

Prof Katrina Lythgoe PhD <sup>5</sup>, Dr M. Estee Torok FRCP <sup>19, 20</sup> and Prof Ian G Goodfellow PhD <sup>24</sup>

**Leadership and supervision, Metadata curation, Project administration, Samples and logistics, and Visualisation:**

Dr Jacqui A Prieto PhD <sup>97, 82</sup> and Dr Kordo Saeed MD, FRCPath <sup>97, 83</sup>

**Leadership and supervision, Metadata curation, Project administration, Sequencing and analysis, and Software and analysis tools:**

Dr David K Jackson PhD <sup>116</sup>

**Leadership and supervision, Metadata curation, Samples and logistics, Sequencing and analysis, and Visualisation:**

Dr Catherine Houlihan PhD <sup>96, 94</sup>

**Leadership and supervision, Metadata curation, Sequencing and analysis, Software and analysis tools, and Visualisation:**

Dr Dan Frampton PhD <sup>94, 95</sup>

**Metadata curation, Project administration, Samples and logistics, Sequencing and analysis, and Software and analysis tools:**

Dr William L Hamilton PhD <sup>19</sup> and Dr Adam A Witney PhD <sup>41</sup>

**Funding acquisition, Samples and logistics, Sequencing and analysis, and Visualisation:**

Dr Giselda Bucca PhD <sup>101</sup>

**Funding acquisition, Leadership and supervision, Metadata curation, and Project administration:**

Dr Cassie F Pope PhD <sup>40, 41</sup>

**Funding acquisition, Leadership and supervision, Metadata curation, and Samples and logistics:**

Dr Catherine Moore PhD <sup>74</sup>

**Funding acquisition, Leadership and supervision, Metadata curation, and Sequencing and analysis:**

Prof Emma C Thomson PhD, FRCP <sup>53</sup>

**Funding acquisition, Leadership and supervision, Project administration, and Samples and logistics:**

Dr Teresa Cutino-Moguel PhD <sup>2</sup>, Dr Ewan M Harrison PhD <sup>116, 102</sup>

**Funding acquisition, Leadership and supervision, Sequencing and analysis, and Visualisation:**

Prof Colin P Smith PhD <sup>101</sup>

**Leadership and supervision, Metadata curation, Project administration, and Sequencing and analysis:**

Fiona Rogan BSc <sup>77</sup>

**Leadership and supervision, Metadata curation, Project administration, and Samples and logistics:**  
Shaun M Beckwith MSc <sup>6</sup>, Abigail Murray Degree <sup>6</sup>, Dawn Singleton HNC <sup>6</sup>, Dr Kirstine Eastick PhD,  
FRCPath <sup>37</sup>, Dr Liz A Sheridan PhD <sup>98</sup>, Paul Randell MSc, PgD <sup>99</sup>, Dr Leigh M Jackson PhD <sup>105</sup>, Dr Cristina  
V Ariani PhD <sup>116</sup> and Dr Sónia Gonçalves PhD <sup>116</sup>

**Leadership and supervision, Metadata curation, Samples and logistics, and Sequencing and analysis:**

Dr Derek J Fairley PhD <sup>3, 77</sup>, Prof Matthew W Loose PhD <sup>18</sup> and Joanne Watkins MSc <sup>74</sup>

**Leadership and supervision, Metadata curation, Samples and logistics, and Visualisation:**

Dr Samuel Moses MD <sup>25, 106</sup>

**Leadership and supervision, Metadata curation, Sequencing and analysis, and Software and analysis tools:**

Dr Sam Nicholls PhD <sup>43</sup>, Dr Matthew Bull PhD <sup>74</sup> and Dr Roberto Amato PhD <sup>116</sup>

**Leadership and supervision, Project administration, Samples and logistics, and Sequencing and analysis:**

Prof Darren L Smith PhD <sup>36, 65, 66</sup>

**Leadership and supervision, Sequencing and analysis, Software and analysis tools, and Visualisation:**

Prof David M Aanensen PhD <sup>14, 116</sup> and Dr Jeffrey C Barrett PhD <sup>116</sup>

**Metadata curation, Project administration, Samples and logistics, and Sequencing and analysis:**

Dr Beatrix Kele PhD <sup>2</sup>, Dr Dinesh Aggarwal MRCP<sup>20, 116, 70</sup>, Dr James G Shepherd MBCHB, MRCP <sup>53</sup>, Dr  
Martin D Curran PhD <sup>71</sup> and Dr Surendra Parmar PhD <sup>71</sup>

**Metadata curation, Project administration, Sequencing and analysis, and Software and analysis tools:**

Dr Matthew D Parker PhD <sup>109</sup>

**Metadata curation, Samples and logistics, Sequencing and analysis, and Software and analysis tools:**

Dr Catryn Williams PhD <sup>74</sup>

**Metadata curation, Samples and logistics, Sequencing and analysis, and Visualisation:**

Dr Sharon Glaysher PhD <sup>68</sup>

**Metadata curation, Sequencing and analysis, Software and analysis tools, and Visualisation:**

Dr Anthony P Underwood PhD <sup>14, 116</sup>, Dr Matthew Bashton PhD <sup>36, 65</sup>, Dr Nicole Pacchiarini PhD <sup>74</sup>, Dr  
Katie F Loveson PhD <sup>84</sup> and Matthew Byott MSc <sup>95, 96</sup>

**Project administration, Sequencing and analysis, Software and analysis tools, and Visualisation:**

Dr Alessandro M Carabelli PhD <sup>20</sup>

**Funding acquisition, Leadership and supervision, and Metadata curation:**

Dr Kate E Templeton PhD <sup>56, 104</sup>

**Funding acquisition, Leadership and supervision, and Project administration:**

Prof Sharon J Peacock PhD <sup>20, 70</sup>, Dr Thushan I de Silva PhD <sup>109</sup>, Dr Dennis Wang PhD <sup>109</sup>, Dr Cordelia F Langford PhD <sup>116</sup> and John Sillitoe BEng <sup>116</sup>

**Funding acquisition, Leadership and supervision, and Samples and logistics:**

Prof Rory N Gunson PhD, FRCPath <sup>55</sup>

**Funding acquisition, Leadership and supervision, and Sequencing and analysis:**

Dr Simon Cottrell PhD <sup>74</sup>, Dr Justin O'Grady PhD <sup>75, 103</sup> and Prof Dominic Kwiatkowski PhD <sup>116, 108</sup>

**Leadership and supervision, Metadata curation, and Project administration:**

Dr Patrick J Lillie PhD, FRCP <sup>37</sup>

**Leadership and supervision, Metadata curation, and Samples and logistics:**

Dr Nicholas Cortes MBCHB <sup>33</sup>, Dr Nathan Moore MBCHB <sup>33</sup>, Dr Claire Thomas DPhil <sup>33</sup>, Phillipa J Burns MSc, DipRCPPath <sup>37</sup>, Dr Tabitha W Mahungu FRCPath <sup>80</sup> and Steven Liggett BSc <sup>86</sup>

**Leadership and supervision, Metadata curation, and Sequencing and analysis:**

Angela H Beckett MSc <sup>13, 81</sup> and Prof Matthew TG Holden PhD <sup>73</sup>

**Leadership and supervision, Project administration, and Samples and logistics:**

Dr Lisa J Levett PhD <sup>34</sup>, Dr Husam Osman PhD <sup>70, 35</sup> and Dr Mohammed O Hassan-Ibrahim PhD, FRCPath <sup>99</sup>

**Leadership and supervision, Project administration, and Sequencing and analysis:**

Dr David A Simpson PhD <sup>77</sup>

**Leadership and supervision, Samples and logistics, and Sequencing and analysis:**

Dr Meera Chand PhD <sup>72</sup>, Prof Ravi K Gupta PhD <sup>102</sup>, Prof Alistair C Darby PhD <sup>107</sup> and Prof Steve Paterson PhD <sup>107</sup>

**Leadership and supervision, Sequencing and analysis, and Software and analysis tools:**

Prof Oliver G Pybus DPhil <sup>23</sup>, Dr Erik M Volz PhD <sup>39</sup>, Prof Daniela de Angelis PhD <sup>52</sup>, Prof David L Robertson PhD <sup>53</sup>, Dr Andrew J Page PhD <sup>75</sup> and Dr Inigo Martincorena PhD <sup>116</sup>

**Leadership and supervision, Sequencing and analysis, and Visualisation:**

Dr Louise Aigrain PhD <sup>116</sup> and Dr Andrew R Bassett PhD <sup>116</sup>

**Metadata curation, Project administration, and Samples and logistics:**

Dr Nick Wong DPhil, MRCP, FRCPath <sup>50</sup>, Dr Yusri Taha MD, PhD <sup>89</sup>, Michelle J Erkiert BA <sup>99</sup> and Dr Michael H Spencer Chapman MBBS <sup>116, 102</sup>

**Metadata curation, Project administration, and Sequencing and analysis:**

Dr Rebecca Dewar PhD <sup>56</sup> and Martin P McHugh MSc <sup>56, 111</sup>

**Metadata curation, Project administration, and Software and analysis tools:**

Siddharth Mookerjee MPH <sup>38, 57</sup>

**Metadata curation, Project administration, and Visualisation:**

Stephen Aplin <sup>97</sup>, Matthew Harvey <sup>97</sup>, Thea Sass <sup>97</sup>, Dr Helen Umpleby FRCP <sup>97</sup> and Helen Wheeler <sup>97</sup>

**Metadata curation, Samples and logistics, and Sequencing and analysis:**

Dr James P McKenna PhD <sup>3</sup>, Dr Ben Warne MRCP <sup>9</sup>, Joshua F Taylor MSc <sup>22</sup>, Yasmin Chaudhry BSc <sup>24</sup>, Rhys Izuagbe <sup>24</sup>, Dr Aminu S Jahun PhD <sup>24</sup>, Dr Gregory R Young PhD <sup>36, 65</sup>, Dr Claire McMurray PhD <sup>43</sup>, Dr Clare M McCann PhD <sup>65, 66</sup>, Dr Andrew Nelson PhD <sup>65, 66</sup> and Scott Elliott <sup>68</sup>

**Metadata curation, Samples and logistics, and Visualisation:**

Hannah Lowe MSc <sup>25</sup>

**Metadata curation, Sequencing and analysis, and Software and analysis tools:**

Dr Anna Price PhD <sup>11</sup>, Matthew R Crown BSc <sup>65</sup>, Dr Sara Rey PhD <sup>74</sup>, Dr Sunando Roy PhD <sup>96</sup> and Dr Ben Temperton PhD <sup>105</sup>

**Metadata curation, Sequencing and analysis, and Visualisation:**

Dr Sharif Shaaban PhD <sup>73</sup> and Dr Andrew R Hesketh PhD <sup>101</sup>

**Project administration, Samples and logistics, and Sequencing and analysis:**

Dr Kenneth G Laing PhD<sup>41</sup>, Dr Irene M Monahan PhD <sup>41</sup> and Dr Judith Heaney PhD <sup>95, 96, 34</sup>

**Project administration, Samples and logistics, and Visualisation:**

Dr Emanuela Pelosi FRCPATH <sup>97</sup>, Siona Silveira MSc <sup>97</sup> and Dr Eleri Wilson-Davies MD, FRCPATH <sup>97</sup>

**Samples and logistics, Software and analysis tools, and Visualisation:**

Dr Helen Fryer PhD <sup>5</sup>

**Sequencing and analysis, Software and analysis tools, and Visualization:**

Dr Helen Adams PhD <sup>4</sup>, Dr Louis du Plessis PhD <sup>23</sup>, Dr Rob Johnson PhD <sup>39</sup>, Dr William T Harvey PhD <sup>53, 42</sup>, Dr Joseph Hughes PhD <sup>53</sup>, Dr Richard J Orton PhD <sup>53</sup>, Dr Lewis G Spurgin PhD <sup>59</sup>, Dr Yann Bourgeois PhD <sup>81</sup>, Dr Chris Ruis PhD <sup>102</sup>, Áine O'Toole MSc <sup>104</sup>, Marina Gourtovaia MSc <sup>116</sup> and Dr Theo Sanderson PhD <sup>116</sup>

**Funding acquisition, and Leadership and supervision:**

Dr Christophe Fraser PhD <sup>5</sup>, Dr Jonathan Edgeworth PhD, FRCPATH <sup>12</sup>, Prof Judith Breuer MD <sup>96, 29</sup>, Dr Stephen L Michell PhD <sup>105</sup> and Prof John A Todd PhD <sup>115</sup>

**Funding acquisition, and Project administration:**

Michaela John BSc <sup>10</sup> and Dr David Buck PhD <sup>115</sup>

**Leadership and supervision, and Metadata curation:**

Dr Kavitha Gajee MBBS, FRCPATH <sup>37</sup> and Dr Gemma L Kay PhD <sup>75</sup>

**Leadership and supervision, and Project administration:**

David Heyburn <sup>74</sup>

**Leadership and supervision, and Samples and logistics:**

Dr Themoula Charalampous PhD <sup>12, 46</sup>, Adela Alcolea-Medina <sup>32, 112</sup>, Katie Kitchman BSc <sup>37</sup>, Prof Alan McNally PhD <sup>43, 93</sup>, David T Pritchard MSc, CSci <sup>50</sup>, Dr Samir Dervisevic FRCPATH <sup>58</sup>, Dr Peter Muir PhD <sup>70</sup>, Dr Esther Robinson PhD <sup>70, 35</sup>, Dr Barry B Vipond PhD <sup>70</sup>, Newara A Ramadan MSc, CSci, FIBMS <sup>78</sup>, Dr Christopher Jeanes MBBS <sup>90</sup>, Danni Weldon BSc <sup>116</sup>, Jana Catalan MSc <sup>118</sup> and Neil Jones MSc <sup>118</sup>

#### **Leadership and supervision, and Sequencing and analysis:**

Dr Ana da Silva Filipe PhD <sup>53</sup>, Dr Chris Williams MBBS <sup>74</sup>, Marc Fuchs BSc <sup>77</sup>, Dr Julia Miskelly PhD <sup>77</sup>, Dr Aaron R Jeffries PhD <sup>105</sup>, Karen Oliver BSc <sup>116</sup> and Dr Naomi R Park PhD <sup>116</sup>

#### **Metadata curation, and Samples and logistics:**

Amy Ash BSc <sup>1</sup>, Cherian Koshy MSc, CSci, FIBMS <sup>1</sup>, Magdalena Barrow <sup>7</sup>, Dr Sarah L Buchan PhD <sup>7</sup>, Dr Anna Mantzouratou PhD <sup>7</sup>, Dr Gemma Clark PhD <sup>15</sup>, Dr Christopher W Holmes PhD <sup>16</sup>, Sharon Campbell MSc <sup>17</sup>, Thomas Davis MSc <sup>21</sup>, Ngee Keong Tan MSc <sup>22</sup>, Dr Julianne R Brown PhD <sup>29</sup>, Dr Kathryn A Harris PhD <sup>29, 2</sup>, Stephen P Kidd MSc <sup>33</sup>, Dr Paul R Grant PhD <sup>34</sup>, Dr Li Xu-McCrae PhD <sup>35</sup>, Dr Alison Cox PhD <sup>38, 63</sup>, Pinglawathee Madona <sup>38, 63</sup>, Dr Marcus Pond PhD <sup>38, 63</sup>, Dr Paul A Randell MBChB <sup>38, 63</sup>, Karen T Withell FIBMS <sup>48</sup>, Cheryl Williams MSc <sup>51</sup>, Dr Clive Graham MD <sup>60</sup>, Rebecca Denton-Smith BSc <sup>62</sup>, Emma Swindells BSc <sup>62</sup>, Robyn Turnbull BSc <sup>62</sup>, Dr Tim J Sloan PhD <sup>67</sup>, Dr Andrew Bosworth PhD <sup>70, 35</sup>, Stephanie Hutchings <sup>70</sup>, Hannah M Pymont MSc <sup>70</sup>, Dr Anna Casey PhD <sup>76</sup>, Dr Liz Ratcliffe PhD <sup>76</sup>, Dr Christopher R Jones PhD <sup>79, 105</sup>, Dr Bridget A Knight PhD <sup>79, 105</sup>, Dr Tanzina Haque PhD, FRCPATH <sup>80</sup>, Dr Jennifer Hart MRCP <sup>80</sup>, Dr Dianne Irish-Tavares FRCPATH <sup>80</sup>, Eric Witeles MSc <sup>80</sup>, Craig Mower BA <sup>86</sup>, Louisa K Watson DipHE <sup>86</sup>, Jennifer Collins BSc <sup>89</sup>, Gary Eltringham BSc <sup>89</sup>, Dorian Crudgington <sup>98</sup>, Ben Macklin <sup>98</sup>, Prof Miren Iturriza-Gomara PhD <sup>107</sup>, Dr Anita O Lucaci PhD <sup>107</sup> and Dr Patrick C McClure PhD <sup>113</sup>

#### **Metadata curation, and Sequencing and analysis:**

Matthew Carlile BSc <sup>18</sup>, Dr Nadine Holmes PhD <sup>18</sup>, Dr Christopher Moore PhD <sup>18</sup>, Dr Nathaniel Storey PhD <sup>29</sup>, Dr Stefan Rooke PhD <sup>73</sup>, Dr Gonzalo Yebra PhD <sup>73</sup>, Dr Noel Craine DPhil <sup>74</sup>, Malorie Perry MSc <sup>74</sup>, Dr Nabil-Fareed Alikhan PhD <sup>75</sup>, Dr Stephen Bridgett PhD <sup>77</sup>, Kate F Cook MScR <sup>84</sup>, Christopher Fearn MSc <sup>84</sup>, Dr Salman Goudarzi PhD <sup>84</sup>, Prof Ronan A Lyons MD <sup>88</sup>, Dr Thomas Williams MD <sup>104</sup>, Dr Sam T Haldenby PhD <sup>107</sup>, Jillian Durham BSc <sup>116</sup> and Dr Steven Leonard PhD <sup>116</sup>

#### **Metadata curation, and Software and analysis tools:**

Robert M Davies MA (Cantab) <sup>116</sup>

#### **Project administration, and Samples and logistics:**

Dr Rahul Batra MD <sup>12</sup>, Beth Blane BSc <sup>20</sup>, Dr Moira J Spyder PhD <sup>30, 95, 96</sup>, Perminder Smith MSc <sup>32, 112</sup>, Mehmet Yavus <sup>85, 109</sup>, Dr Rachel J Williams PhD <sup>96</sup>, Dr Adhyana IK Mahanama MD <sup>97</sup>, Dr Buddhini Samaraweera MD <sup>97</sup>, Sophia T Girgis MSc <sup>102</sup>, Samantha E Hansford CSci <sup>109</sup>, Dr Angie Green PhD <sup>115</sup>, Dr Charlotte Beaver PhD <sup>116</sup>, Katherine L Bellis <sup>116, 102</sup>, Matthew J Dorman <sup>116</sup>, Sally Kay <sup>116</sup>, Liam Prestwood <sup>116</sup> and Dr Shavanthi Rajatileka PhD <sup>116</sup>

#### **Project administration, and Sequencing and analysis:**

Dr Joshua Quick PhD <sup>43</sup>

#### **Project administration, and Software and analysis tools:**

Radoslaw Poplawski BSc <sup>43</sup>

#### **Samples and logistics, and Sequencing and analysis:**



Dr Nicola Reynolds PhD <sup>8</sup>, Andrew Mack MPhil <sup>11</sup>, Dr Arthur Morriss PhD <sup>11</sup>, Thomas Whalley BSc <sup>11</sup>, Bindi Patel BSc <sup>12</sup>, Dr Iliana Georgana PhD <sup>24</sup>, Dr Myra Hosmillo PhD <sup>24</sup>, Malte L Pinckert MPhil <sup>24</sup>, Dr Joanne Stockton PhD <sup>43</sup>, Dr John H Henderson PhD <sup>65</sup>, Amy Hollis HND <sup>65</sup>, Dr William Stanley PhD <sup>65</sup>, Dr Wen C Yew PhD <sup>65</sup>, Dr Richard Myers PhD <sup>72</sup>, Dr Alicia Thornton PhD <sup>72</sup>, Alexander Adams BSc <sup>74</sup>, Tara Annett BSc <sup>74</sup>, Dr Hibo Asad PhD <sup>74</sup>, Alec Birchley MSc <sup>74</sup>, Jason Coombes BSc <sup>74</sup>, Johnathan M Evans MSc <sup>74</sup>, Laia Fina <sup>74</sup>, Bree Gatica-Wilcox MPhil <sup>74</sup>, Lauren Gilbert <sup>74</sup>, Lee Graham BSc <sup>74</sup>, Jessica Hey BSc <sup>74</sup>, Ember Hilvers MPH <sup>74</sup>, Sophie Jones MSc <sup>74</sup>, Hannah Jones <sup>74</sup>, Sara Kumziene-Summerhayes MSc <sup>74</sup>, Dr Caoimhe McKerr PhD <sup>74</sup>, Jessica Powell BSc <sup>74</sup>, Georgia Pugh <sup>74</sup>, Sarah Taylor <sup>74</sup>, Alexander J Trotter MRes <sup>75</sup>, Charlotte A Williams BSc <sup>96</sup>, Leanne M Kermack MSc <sup>102</sup>, Benjamin H Foulkes MSc <sup>109</sup>, Marta Gallis MSc <sup>109</sup>, Hailey R Hornsby MSc <sup>109</sup>, Stavroula F Louka MSc <sup>109</sup>, Dr Manoj Pohare PhD <sup>109</sup>, Paige Wolverson MSc <sup>109</sup>, Peijun Zhang MSc <sup>109</sup>, George MacIntyre-Cockett BSc <sup>115</sup>, Amy Trebes MSc <sup>115</sup>, Dr Robin J Moll PhD <sup>116</sup>, Lynne Ferguson MSc <sup>117</sup>, Dr Emily J Goldstein PhD <sup>117</sup>, Dr Alasdair Maclean PhD <sup>117</sup> and Dr Rachael Tomb PhD <sup>117</sup>

#### **Samples and logistics, and Software and analysis tools:**

Dr Igor Starinskij MSc, MRCP <sup>53</sup>

#### **Sequencing and analysis, and Software and analysis tools:**

Laura Thomson BSc <sup>5</sup>, Joel Southgate MSc <sup>11, 74</sup>, Dr Moritz UG Kraemer DPhil <sup>23</sup>, Dr Jayna Raghvani PhD <sup>23</sup>, Dr Alex E Zarebski PhD <sup>23</sup>, Olivia Boyd MSc <sup>39</sup>, Lily Geidelberg MSc <sup>39</sup>, Dr Chris J Illingworth PhD <sup>52</sup>, Dr Chris Jackson PhD <sup>52</sup>, Dr David Pascall PhD <sup>52</sup>, Dr Sreenu Vattipally PhD <sup>53</sup>, Timothy M Freeman MPhil <sup>109</sup>, Dr Sharon N Hsu PhD <sup>109</sup>, Dr Benjamin B Lindsey MRCP <sup>109</sup>, Dr Keith James PhD <sup>116</sup>, Kevin Lewis <sup>116</sup>, Gerry Tonkin-Hill <sup>116</sup> and Dr Jaime M Tovar-Corona PhD <sup>116</sup>

#### **Sequencing and analysis, and Visualisation:**

MacGregor Cox MSci <sup>20</sup>

#### **Software and analysis tools, and Visualisation:**

Dr Khalil Abudahab PhD <sup>14, 116</sup>, Mirko Menegazzo <sup>14</sup>, Ben EW Taylor MEng <sup>14, 116</sup>, Dr Corin A Yeats PhD <sup>14</sup>, Afrida Mukaddas BTech <sup>53</sup>, Derek W Wright MSc <sup>53</sup>, Dr Leonardo de Oliveira Martins PhD <sup>75</sup>, Dr Rachel Colquhoun DPhil <sup>104</sup>, Verity Hill <sup>104</sup>, Dr Ben Jackson PhD <sup>104</sup>, Dr JT McCrone PhD <sup>104</sup>, Dr Nathan Medd PhD <sup>104</sup>, Dr Emily Scher PhD <sup>104</sup> and Jon-Paul Keatley <sup>116</sup>

#### **Leadership and supervision:**

Dr Tanya Curran PhD <sup>3</sup>, Dr Sian Morgan FRCPATH <sup>10</sup>, Prof Patrick Maxwell PhD <sup>20</sup>, Prof Ken Smith PhD <sup>20</sup>, Dr Sahar Eldirdiri MBBS, MSc, FRCPATH <sup>21</sup>, Anita Kenyon MSc <sup>21</sup>, Prof Alison H Holmes MD <sup>38, 57</sup>, Dr James R Price PhD <sup>38, 57</sup>, Dr Tim Wyatt PhD <sup>69</sup>, Dr Alison E Mather PhD <sup>75</sup>, Dr Timofey Skvortsov PhD <sup>77</sup> and Prof John A Hartley PhD <sup>96</sup>

#### **Metadata curation:**

Prof Martyn Guest PhD <sup>11</sup>, Dr Christine Kitchen PhD <sup>11</sup>, Dr Ian Merrick PhD <sup>11</sup>, Robert Munn BSc <sup>11</sup>, Dr Beatrice Bertolusso Degree <sup>33</sup>, Dr Jessica Lynch MBCHB <sup>33</sup>, Dr Gabrielle Vernet MBBS <sup>33</sup>, Stuart Kirk MSc <sup>34</sup>, Dr Elizabeth Wastnedge MD <sup>56</sup>, Dr Rachael Stanley PhD <sup>58</sup>, Giles Idle <sup>64</sup>, Dr Declan T Bradley PhD <sup>69, 77</sup>, Nicholas F Killough MSc <sup>69</sup>, Dr Jennifer Poyner MD <sup>79</sup> and Matilde Mori BSc <sup>110</sup>

#### **Project administration:**

Owen Jones BSc <sup>11</sup>, Victoria Wright BSc <sup>18</sup>, Ellena Brooks MA <sup>20</sup>, Carol M Churcher BSc <sup>20</sup>, Dr Laia Delgado Callico PhD <sup>20</sup>, Mireille Fragakis HND <sup>20</sup>, Dr Katerina Galai PhD <sup>20, 70</sup>, Dr Andrew Jermy PhD <sup>20</sup>,

Sarah Judges BA <sup>20</sup>, Anna Markov BSc <sup>20</sup>, Georgina M McManus BSc <sup>20</sup>, Kim S Smith <sup>20</sup>, Peter M D Thomas-McEwen MSc <sup>20</sup>, Dr Elaine Westwick PhD <sup>20</sup>, Dr Stephen W Attwood PhD <sup>23</sup>, Dr Frances Bolt PhD <sup>38, 57</sup>, Dr Alisha Davies PhD <sup>74</sup>, Elen De Lacy MPH <sup>74</sup>, Fatima Downing <sup>74</sup>, Sue Edwards <sup>74</sup>, Lizzie Meadows MA <sup>75</sup>, Sarah Jeremiah MSc <sup>97</sup>, Dr Nikki Smith PhD <sup>109</sup> and Luke Foulser <sup>116</sup>

### **Samples and logistics:**

Amita Patel BSc <sup>12</sup>, Dr Louise Berry PhD <sup>15</sup>, Dr Tim Boswell PhD <sup>15</sup>, Dr Vicki M Fleming PhD <sup>15</sup>, Dr Hannah C Howson-Wells PhD <sup>15</sup>, Dr Amelia Joseph PhD <sup>15</sup>, Manjinder Khakh <sup>15</sup>, Dr Michelle M Lister PhD <sup>15</sup>, Paul W Bird MSc, MRes <sup>16</sup>, Karlie Fallon <sup>16</sup>, Thomas Helmer <sup>16</sup>, Dr Claire L McMurray PhD <sup>16</sup>, Mina Odedra BSc <sup>16</sup>, Jessica Shaw BSc <sup>16</sup>, Dr Julian W Tang PhD <sup>16</sup>, Nicholas J Willford MSc <sup>16</sup>, Victoria Blakey BSc <sup>17</sup>, Dr Veena Raviprakash MD <sup>17</sup>, Nicola Sheriff BSc <sup>17</sup>, Lesley-Anne Williams BSc <sup>17</sup>, Theresa Feltwell MSc <sup>20</sup>, Dr Luke Bedford PhD <sup>26</sup>, Dr James S Cargill PhD <sup>27</sup>, Warwick Hughes MSc <sup>27</sup>, Dr Jonathan Moore MD <sup>28</sup>, Susanne Stonehouse BSc <sup>28</sup>, Laura Atkinson MSc <sup>29</sup>, Jack CD Lee MSc <sup>29</sup>, Dr Divya Shah PhD <sup>29</sup>, Natasha Ohemeng-Kumi MSc <sup>32, 112</sup>, John Ramble MSc <sup>32, 112</sup>, Jasveen Sehmi MSc <sup>32, 112</sup>, Dr Rebecca Williams BMBS <sup>33</sup>, Wendy Chatterton MSc <sup>34</sup>, Monika Pusok MSc <sup>34</sup>, William Everson MSc <sup>37</sup>, Anibolina Castigador IBMS HCPC <sup>44</sup>, Emily Macnaughton FRCPath <sup>44</sup>, Dr Kate El Bouzidi MRCP <sup>45</sup>, Dr Temi Lampejo FRCPath <sup>45</sup>, Dr Malur Sudhanva FRCPath <sup>45</sup>, Cassie Breen BSc <sup>47</sup>, Dr Graciela Sluga MD, MSc <sup>48</sup>, Dr Shazaad SY Ahmad MSc <sup>49, 70</sup>, Dr Ryan P George PhD <sup>49</sup>, Dr Nicholas W Machin MSc <sup>49, 70</sup>, Debbie Binns BSc <sup>50</sup>, Victoria James BSc <sup>50</sup>, Dr Rachel Blacow MBCHB <sup>55</sup>, Dr Lindsay Coupland PhD <sup>58</sup>, Dr Louise Smith PhD <sup>59</sup>, Dr Edward Barton MD <sup>60</sup>, Debra Padgett BSc <sup>60</sup>, Garren Scott BSc <sup>60</sup>, Dr Aidan Cross MBCHB <sup>61</sup>, Dr Mariyam Mirfenderesky FRCPath <sup>61</sup>, Jane Greenaway MSc <sup>62</sup>, Kevin Cole <sup>64</sup>, Phillip Clarke <sup>67</sup>, Nichola Duckworth <sup>67</sup>, Sarah Walsh <sup>67</sup>, Kelly Bicknell <sup>68</sup>, Robert Impey MSc <sup>68</sup>, Dr Sarah Wyllie PhD <sup>68</sup>, Richard Hopes <sup>70</sup>, Dr Chloe Bishop PhD <sup>72</sup>, Dr Vicki Chalker PhD <sup>72</sup>, Dr Ian Harrison PhD <sup>72</sup>, Laura Gifford MSc <sup>74</sup>, Dr Zoltan Molnar PhD <sup>77</sup>, Dr Cressida Auckland FRCPath <sup>79</sup>, Dr Cariad Evans PhD <sup>85, 109</sup>, Dr Kate Johnson PhD <sup>85, 109</sup>, Dr David G Partridge FRCP, FRCPath <sup>85, 109</sup>, Dr Mohammad Raza PhD <sup>85, 109</sup>, Paul Baker MD <sup>86</sup>, Prof Stephen Bonner PhD <sup>86</sup>, Sarah Essex <sup>86</sup>, Leanne J Murray <sup>86</sup>, Andrew I Lawton MSc <sup>87</sup>, Dr Shirelle Burton-Fanning MD <sup>89</sup>, Dr Brendan Al Payne MD <sup>89</sup>, Dr Sheila Waugh MD <sup>89</sup>, Andrea N Gomes MSc <sup>91</sup>, Maimuna Kimuli MSc <sup>91</sup>, Darren R Murray MSc <sup>91</sup>, Paula Ashfield MSc <sup>92</sup>, Dr Donald Dobie MBCHB <sup>92</sup>, Dr Fiona Ashford PhD <sup>93</sup>, Dr Angus Best PhD <sup>93</sup>, Dr Liam Crawford PhD <sup>93</sup>, Dr Nicola Cumley PhD <sup>93</sup>, Dr Megan Mayhew PhD <sup>93</sup>, Dr Oliver Megram PhD <sup>93</sup>, Dr Jeremy Mirza PhD <sup>93</sup>, Dr Emma Moles-Garcia PhD <sup>93</sup>, Dr Benita Percival PhD <sup>93</sup>, Megan Driscoll BSc <sup>96</sup>, Leah Ensell BSc <sup>96</sup>, Dr Helen L Lowe PhD <sup>96</sup>, Laurentiu Maftei BSc <sup>96</sup>, Matteo Mondani MSc <sup>96</sup>, Nicola J Chaloner BSc <sup>99</sup>, Benjamin J Cogger BSc <sup>99</sup>, Lisa J Easton MSc <sup>99</sup>, Hannah Huckson BSc <sup>99</sup>, Jonathan Lewis MSc, PgD, FIBMS <sup>99</sup>, Sarah Lowdon BSc <sup>99</sup>, Cassandra S Malone MSc <sup>99</sup>, Florence Munemo BSc <sup>99</sup>, Manasa Mutingwende MSc <sup>99</sup>, Roberto Nicodemi BSc <sup>99</sup>, Olga Podplomyk FD <sup>99</sup>, Thomas Somassa BSc <sup>99</sup>, Dr Andrew Beggs PhD <sup>100</sup>, Dr Alex Richter PhD <sup>100</sup>, Claire Cormie <sup>102</sup>, Joana Dias MSc <sup>102</sup>, Sally Forrest BSc <sup>102</sup>, Dr Ellen E Higginson PhD <sup>102</sup>, Mailis Maes MPhil <sup>102</sup>, Jamie Young BSc <sup>102</sup>, Dr Rose K Davidson PhD <sup>103</sup>, Kathryn A Jackson MSc <sup>107</sup>, Dr Alexander J Keeley MRCP <sup>109</sup>, Prof Jonathan Ball PhD <sup>113</sup>, Timothy Byaruhanga MSc <sup>113</sup>, Dr Joseph G Chappell PhD <sup>113</sup>, Jayasree Dey MSc <sup>113</sup>, Jack D Hill MSc <sup>113</sup>, Emily J Park MSc <sup>113</sup>, Arezou Fanaie MSc <sup>114</sup>, Rachel A Hilson MSc <sup>114</sup>, Geraldine Yaze MSc <sup>114</sup> and Stephanie Lo <sup>116</sup>

### **Sequencing and analysis:**

Safiah Afifi BSc <sup>10</sup>, Robert Beer BSc <sup>10</sup>, Joshua Maksimovic FD <sup>10</sup>, Kathryn McCluggage Masters <sup>10</sup>, Karla Spellman FD <sup>10</sup>, Catherine Bresner BSc <sup>11</sup>, William Fuller BSc <sup>11</sup>, Dr Angela Marchbank BSc <sup>11</sup>, Trudy Workman HNC <sup>11</sup>, Dr Ekaterina Shelest PhD <sup>13, 81</sup>, Dr Johnny Debebe PhD <sup>18</sup>, Dr Fei Sang PhD <sup>18</sup>, Dr Sarah Francois PhD <sup>23</sup>, Bernardo Gutierrez MSc <sup>23</sup>, Dr Tetyana I Vasylyeva DPhil <sup>23</sup>, Dr Flavia Flaviani PhD <sup>31</sup>, Dr Manon Ragonnet-Cronin PhD <sup>39</sup>, Dr Katherine L Smollett PhD <sup>42</sup>, Alice Broos BSc <sup>53</sup>, Daniel

Mair BSc <sup>53</sup>, Jenna Nichols BSc <sup>53</sup>, Dr Kyriaki Nomikou PhD <sup>53</sup>, Dr Lily Tong PhD <sup>53</sup>, Ioulia Tsatsani MSc <sup>53</sup>, Prof Sarah O'Brien PhD <sup>54</sup>, Prof Steven Rushton PhD <sup>54</sup>, Dr Roy Sanderson PhD <sup>54</sup>, Dr Jon Perkins MBChB <sup>55</sup>, Seb Cotton MSc <sup>56</sup>, Abbie Gallagher BSc <sup>56</sup>, Dr Elias Allara MD, PhD <sup>70,102</sup>, Clare Pearson MSc <sup>70,102</sup>, Dr David Bibby PhD <sup>72</sup>, Dr Gavin Dabrera PhD <sup>72</sup>, Dr Nicholas Ellaby PhD <sup>72</sup>, Dr Eileen Gallagher PhD <sup>72</sup>, Dr Jonathan Hubb PhD <sup>72</sup>, Dr Angie Lackenby PhD <sup>72</sup>, Dr David Lee PhD <sup>72</sup>, Nikos Manesis <sup>72</sup>, Dr Tamyo Mbisa PhD <sup>72</sup>, Dr Steven Platt PhD <sup>72</sup>, Katherine A Twohig <sup>72</sup>, Dr Mari Morgan PhD <sup>74</sup>, Alp Aydin MSci <sup>75</sup>, David J Baker BEng <sup>75</sup>, Dr Ebenezer Foster-Nyarko PhD <sup>75</sup>, Dr Sophie J Prosolek PhD <sup>75</sup>, Steven Rudder <sup>75</sup>, Chris Baxter BSc <sup>77</sup>, Sílvia F Carvalho MSc <sup>77</sup>, Dr Deborah Lavin PhD <sup>77</sup>, Dr Arun Mariappan PhD <sup>77</sup>, Dr Clara Radulescu PhD <sup>77</sup>, Dr Aditi Singh PhD <sup>77</sup>, Miao Tang MD <sup>77</sup>, Helen Morcrette BSc <sup>79</sup>, Nadua Bayzid BSc <sup>96</sup>, Marius Cotic MSc <sup>96</sup>, Dr Carlos E Balcazar PhD <sup>104</sup>, Dr Michael D Gallagher PhD <sup>104</sup>, Dr Daniel Maloney PhD <sup>104</sup>, Thomas D Stanton BSc <sup>104</sup>, Dr Kathleen A Williamson PhD <sup>104</sup>, Dr Robin Manley PhD <sup>105</sup>, Michelle L Michelsen BSc <sup>105</sup>, Dr Christine M Sambles PhD <sup>105</sup>, Dr David J Studholme PhD <sup>105</sup>, Joanna Warwick-Dugdale BSc <sup>105</sup>, Richard Eccles MSc <sup>107</sup>, Matthew Gemmell MSc <sup>107</sup>, Dr Richard Gregory PhD <sup>107</sup>, Dr Margaret Hughes PhD <sup>107</sup>, Charlotte Nelson MSc <sup>107</sup>, Dr Lucille Rainbow PhD <sup>107</sup>, Dr Edith E Vamos PhD <sup>107</sup>, Hermione J Webster BSc <sup>107</sup>, Dr Mark Whitehead PhD <sup>107</sup>, Claudia Wierzbicki BSc <sup>107</sup>, Dr Adrienn Angyal PhD <sup>109</sup>, Dr Luke R Green PhD <sup>109</sup>, Dr Max Whiteley PhD <sup>109</sup>, Emma Betteridge BSc <sup>116</sup>, Dr Iraad F Bronner PhD <sup>116</sup>, Ben W Farr BSc <sup>116</sup>, Scott Goodwin MSc <sup>116</sup>, Dr Stefanie V Lensing PhD <sup>116</sup>, Shane A McCarthy <sup>116,102</sup>, Dr Michael A Quail PhD <sup>116</sup>, Diana Rajan MSc <sup>116</sup>, Dr Nicholas M Redshaw PhD <sup>116</sup>, Carol Scott <sup>116</sup>, Lesley Shirley MSc <sup>116</sup> and Scott AJ Thurston BSc <sup>116</sup>

#### **Software and analysis tools:**

Dr Will Rowe PhD<sup>43</sup>, Amy Gaskin MSc <sup>74</sup>, Dr Thanh Le-Viet PhD <sup>75</sup>, James Bonfield BSc <sup>116</sup>, Jennifer Liddle <sup>116</sup> and Andrew Whitwham BSc <sup>116</sup>

**1** Barking, Havering and Redbridge University Hospitals NHS Trust, **2** Barts Health NHS Trust, **3** Belfast Health & Social Care Trust, **4** Betsi Cadwaladr University Health Board, **5** Big Data Institute, Nuffield Department of Medicine, University of Oxford, **6** Blackpool Teaching Hospitals NHS Foundation Trust, **7** Bournemouth University, **8** Cambridge Stem Cell Institute, University of Cambridge, **9** Cambridge University Hospitals NHS Foundation Trust, **10** Cardiff and Vale University Health Board, **11** Cardiff University, **12** Centre for Clinical Infection and Diagnostics Research, Department of Infectious Diseases, Guy's and St Thomas' NHS Foundation Trust, **13** Centre for Enzyme Innovation, University of Portsmouth, **14** Centre for Genomic Pathogen Surveillance, University of Oxford, **15** Clinical Microbiology Department, Queens Medical Centre, Nottingham University Hospitals NHS Trust, **16** Clinical Microbiology, University Hospitals of Leicester NHS Trust, **17** County Durham and Darlington NHS Foundation Trust, **18** Deep Seq, School of Life Sciences, Queens Medical Centre, University of Nottingham, **19** Department of Infectious Diseases and Microbiology, Cambridge University Hospitals NHS Foundation Trust, **20** Department of Medicine, University of Cambridge, **21** Department of Microbiology, Kettering General Hospital, **22** Department of Microbiology, South West London Pathology, **23** Department of Zoology, University of Oxford, **24** Division of Virology, Department of Pathology, University of Cambridge, **25** East Kent Hospitals University NHS Foundation Trust, **26** East Suffolk and North Essex NHS Foundation Trust, **27** East Sussex Healthcare NHS Trust, **28** Gateshead Health NHS Foundation Trust, **29** Great Ormond Street Hospital for Children NHS Foundation Trust, **30** Great Ormond Street Institute of Child Health (GOS ICH), University College London (UCL), **31** Guy's and St. Thomas' Biomedical Research Centre, **32** Guy's and St. Thomas' NHS Foundation Trust, **33** Hampshire Hospitals NHS Foundation Trust, **34** Health Services Laboratories, **35** Heartlands Hospital, Birmingham, **36** Hub for Biotechnology in the Built Environment, Northumbria University, **37** Hull University Teaching Hospitals NHS Trust, **38** Imperial College Healthcare NHS Trust, **39** Imperial College London, **40** Infection Care Group, St George's University Hospitals NHS Foundation Trust, **41** Institute for Infection and Immunity, St George's University of London, **42** Institute of Biodiversity, Animal Health & Comparative Medicine, **43** Institute of Microbiology and Infection, University of Birmingham, **44** Isle of Wight NHS Trust, **45** King's College Hospital NHS Foundation Trust, **46** King's College London, **47**

Liverpool Clinical Laboratories, **48** Maidstone and Tunbridge Wells NHS Trust, **49** Manchester University NHS Foundation Trust, **50** Microbiology Department, Buckinghamshire Healthcare NHS Trust, **51** Microbiology, Royal Oldham Hospital, **52** MRC Biostatistics Unit, University of Cambridge, **53** MRC-University of Glasgow Centre for Virus Research, **54** Newcastle University, **55** NHS Greater Glasgow and Clyde, **56** NHS Lothian, **57** NIHR Health Protection Research Unit in HCAI and AMR, Imperial College London, **58** Norfolk and Norwich University Hospitals NHS Foundation Trust, **59** Norfolk County Council, **60** North Cumbria Integrated Care NHS Foundation Trust, **61** North Middlesex University Hospital NHS Trust, **62** North Tees and Hartlepool NHS Foundation Trust, **63** North West London Pathology, **64** Northumbria Healthcare NHS Foundation Trust, **65** Northumbria University, **66** NU-OMICS, Northumbria University, **67** Path Links, Northern Lincolnshire and Goole NHS Foundation Trust, **68** Portsmouth Hospitals University NHS Trust, **69** Public Health Agency, Northern Ireland, **70** Public Health England, **71** Public Health England, Cambridge, **72** Public Health England, Colindale, **73** Public Health Scotland, **74** Public Health Wales, **75** Quadram Institute Bioscience, **76** Queen Elizabeth Hospital, Birmingham, **77** Queen's University Belfast, **78** Royal Brompton and Harefield Hospitals, **79** Royal Devon and Exeter NHS Foundation Trust, **80** Royal Free London NHS Foundation Trust, **81** School of Biological Sciences, University of Portsmouth, **82** School of Health Sciences, University of Southampton, **83** School of Medicine, University of Southampton, **84** School of Pharmacy & Biomedical Sciences, University of Portsmouth, **85** Sheffield Teaching Hospitals NHS Foundation Trust, **86** South Tees Hospitals NHS Foundation Trust, **87** Southwest Pathology Services, **88** Swansea University, **89** The Newcastle upon Tyne Hospitals NHS Foundation Trust, **90** The Queen Elizabeth Hospital King's Lynn NHS Foundation Trust, **91** The Royal Marsden NHS Foundation Trust, **92** The Royal Wolverhampton NHS Trust, **93** Turnkey Laboratory, University of Birmingham, **94** University College London Division of Infection and Immunity, **95** University College London Hospital Advanced Pathogen Diagnostics Unit, **96** University College London Hospitals NHS Foundation Trust, **97** University Hospital Southampton NHS Foundation Trust, **98** University Hospitals Dorset NHS Foundation Trust, **99** University Hospitals Sussex NHS Foundation Trust, **100** University of Birmingham, **101** University of Brighton, **102** University of Cambridge, **103** University of East Anglia, **104** University of Edinburgh, **105** University of Exeter, **106** University of Kent, **107** University of Liverpool, **108** University of Oxford, **109** University of Sheffield, **110** University of Southampton, **111** University of St Andrews, **112** Viapath, Guy's and St Thomas' NHS Foundation Trust, and King's College Hospital NHS Foundation Trust, **113** Virology, School of Life Sciences, Queens Medical Centre, University of Nottingham, **114** Watford General Hospital, **115** Wellcome Centre for Human Genetics, Nuffield Department of Medicine, University of Oxford, **116** Wellcome Sanger Institute, **117** West of Scotland Specialist Virology Centre, NHS Greater Glasgow and Clyde, **118** Whittington Health NHS Trust.