A large effective population size for within-host influenza virus infection

Casper K Lumby[1], Lei Zhao[1], Judy Breuer[2,3], Christopher J R Illingworth[1,4,5]

[1]Department of Genetics, Downing Street, University of Cambridge, Cambridge

[2]Great Ormond Street Hospital, Great Ormond Street, London, UK

[3]Division of Infection and Immunity, University College London, London, UK

[4]Department of Applied Mathematics and Theoretical Physics, Clarkson Road, University of Cambridge, Cambridge

[5]Department of Computer Science, Institute of Biotechnology, University of Helsinki, 00014 Helsinki, Finland

Strains of the influenza virus form coherent global populations, yet exist at the level of single infections in individual hosts. The relationship between these scales is a critical topic for understanding viral evolution. Here we investigate the within-host relationship between selection and the stochastic effects of genetic drift, estimating an effective population size of infection N_e for influenza infection. Examining whole-genome sequence data describing a chronic case of influenza B in a severely immunocompromised child we infer an N_e of 2.5×10^7 (95% confidence range 1.0×10^7 to 9.0×10^7) suggesting the importance of genetic drift to be minimal. Our result, supported by the analysis of data from influenza A infection, suggests that positive selection during within-host infection is primarily limited by the typically short period of infection. Atypically long infections may have a disproportionate influence upon global patterns of viral evolution.

The evolution of the influenza virus may be considered across a broad range of scales. On a global level, populations exhibit coherent behaviour[1–3], evolving rapidly under collective host

28 immune pressure[4,5]. On another level, these global populations are nothing more than very
29 large numbers of individual host infections, separated by transmission events.
30
31 Despite the clear role for selection in global influenza populations, recent studies of within-
32 host infection have suggested that positive selection does not strongly influence evolution at
33 this smaller scale[6–8]. Contrasting explanations have been given for this, with suggestions
34 either that selection at the within-host level is intrinsically inefficient, being dominated by
35 stochastic processes[7], or that while selection is efficient, a mismatch in timing between the
36 peak viral titre and the host adaptive immune response prevents selection from taking effect[8].
37
38 To resolve this issue, we evaluated the relative importance of selection and genetic drift during
39 a case of influenza infection. The balance between these factors is determined by the effective
40 size of the population, denoted N_e. If N_e is high, selection will outweigh genetic drift, even
41 where differences in viral fitness are small[9]. By contrast, if N_e is low, less fit viruses are more
42 likely to outcompete their fitter compatriots.
43
44 Estimating N_e is a difficult task, with a long history of method development in this area[10–12]. A
45 simple measure of N_e may be calculated by matching the genetic change in allele frequencies
46 in a population with the changes occurring in an idealised population evolving under genetic
47 drift[13]. However, such estimates are vulnerable to distortion, for example being reduced by
48 the effect of positive selection in a population. Where the global influenza A/H3N2 population
49 is driven by repeated selective sweeps[14–16] a neutral estimation method suggests a value for
50 N_e not much greater than 100[17]. While methods for jointly estimating N_e and selection exist,
51 they are limited to systems with only a few loci of interest[18–22]. Non-trivial population structure
52 can also affect estimates[23]; a growing body of evidence supports the existence of such
53 structure in within-host influenza infection[24–27]. While careful experimental techniques can
54 reduce sequencing error[28], noise from sequencing and unrepresentative sample collection

55 combine[29], potentially confounding estimates of N_e in viral populations[30]. If N_e is high, any
56 signal of drift can be obscured by noise.
57
58 We here estimate an effective population size for within-host influenza B infection using data
59 collected from a severely immunocompromised host. While the viral load of the infection was
60 not unusual for a hospitalised childhood infection[31], an absence of cell-mediated immunity led
61 to the persistence of the infection for several months[32]. Given extensive sequence data
62 collected during infection, the reduced role of positive selection, combined with novel methods
63 to account for noise and population structure, enabled an improved inference of N_e. The large
64 effective size we infer suggests that selection acts in an efficient manner during within-host
65 influenza infection. The influence of positive selection is limited only by the duration of
66 infection.
67
68 **Results and Discussion**
69
70 Viral samples from the population were collected at 41 time points spanning 8 months during
71 the course of an influenza B infection in a severely immunocompromised host (Fig. 1A).
72 Clinical details of the case, and the use of viral sequence data in evaluating the effectiveness
73 of clinical intervention, have been described elsewhere[32]. After unsuccessful treatment with
74 oseltamivir, zanamivir and nitazoxanide, a bone marrow transplant and favipiravir combination
75 therapy led to the apparent clearance of infection. Apart from a single exception, biweekly
76 samples tested negative for influenza across a period of close to two months. A subsequent
77 resurgence of zanamivir-resistant infection was cleared by favipiravir and zanamivir in
78 combination.
79
80 Phylogenetic analysis of whole-genome viral consensus sequences showed the existence of
81 non-trivial population structure, with at least two distinct clades (Fig. 1B, Fig. 1S1); we term
82 these clades A and B. While being phylogenetically separated the two clades persisted across

several months of infection. Haplotype reconstruction showed that samples from clade B were comprised of distinct viral haplotypes to those from clade A (Fig. 1C, 1S2). Clade B slowly evolved away from the initial consensus sequence (Fig. 1D), while viruses in clade A stayed close to the initial consensus. The cladal structure suggests the existence of spatially distinct viral populations in the host, samples stochastically representing one population or the other.

To estimate the effective population size, we analysed genome-wide sequence data from samples in clade A collected before first use of favipiravir. A method of linear regression was used to quantify the rate of viral evolution, measuring the genetic distance between samples as a function of increasing time between dates of sample collection. We inferred a rate of 0.051 substitutions per day (97.5% confidence interval 0.034 to 0.068) (Fig. 2A), equivalent to 7.94 substitutions genome-wide across 157 days of evolution. The vertical intercept of this line provides an estimate of the contribution of noise to the measured distance between samples, for example arising from sequencing error or undiagnosed population structure. The identified value of close to 40 substitutions is equivalent to a between-sample allele frequency difference of approximately +/- 0.3% per locus. While considerable noise affects each sample, the dataset as a whole provides a clear signal of evolutionary change.

A simulation based analysis, measuring the extent of evolution in idealised Wright-Fisher populations[13], inferred an effective population size of 2.5×10^7 (95% confidence range 1.0×10^7 to 9.0×10^7) for viruses in clade A before the use of favipiravir(Fig. 2B). This value is substantially larger than estimates made recently for within-host HIV infection[33,34], and suggests that even weak selection could easily overcome genetic drift. Data from clade B gave a lower estimated value of 2×10^6, (95% confidence range 4×10^5 to 2×10^8) perhaps reflecting the less frequent observation of samples in that clade (Fig. 2C, D).

The value of Ne might be lowered by population structure within the influenza infection[35]. The partial onset of zanamivir resistant alleles[36], sporadically observed at intermediate frequency

after the administration of the drug (Fig. 2S1), is suggestive of population structure going beyond our simple division into clades; the random sampling of viruses from resistant and susceptible subpopulations would produce this behaviour.

Positive selection might have led us to underestimate N_e. While viral evolution was generally not driven by positive selection (Fig. 2S2), any such selection (e.g. for zanamivir resistance) would increase the rate of viral evolution, lowering our inferred value. Despite this, our result is clear. Once an infection is established, selection will dominate the stochastic effects of drift upon within-host evolution.

The dataset we considered is particularly suited to our calculation. The long period of infection combined with frequent sampling allowed for the characterisation of a slow rate of evolution amidst population structure and noise in the data. Further, the absence of strong selection reduced the error intrinsic to our inference approach, which assumed an idealised neutral population. To provide further validation we repeated our approach on data describing long-term influenza A/H3N2 infection in four immunocompromised adults[37]. The estimates for N_e we obtained, of between 3×10^5 and 1×10^6 (Fig. 2S3), while high, were smaller than for our flu B case, potentially being reduced by an increased influence of selection.

We believe that our study provides a first realistic estimate of within-host effective population size for severe influenza infection in humans. The viral load in the influenza B case was high, representative of hospitalised cases of childhood influenza infection. However, the magnitude of our inferred effective size, of order 10^7, suggests that selection will predominate even at lower viral loads. Our result supports the idea that the observed lack of within-host variation in typical cases of influenza[6,7] can be explained by the short period of infection; the stochastic effects of genetic drift do not limit the impact of positive selection. Rather, as influenza infections are founded by small numbers of viral particles[7,38], the majority of low-frequency variants must arise through *de novo* mutation. In a typical infection, very strong selection is

required for such variants to reach a substantial frequency in the population[39]. We suggest that, while not being confounded by drift, selection does not usually have time to fix novel variants in the population. Clinical evidence suggests that in cases of longer infection, or in the emergence of antiviral resistance, selection does influence the evolutionary outcome of infection[37,40–44].

Our result highlights the potential importance of longer infections in the adaptation of global influenza populations, particularly where some adaptive immune response remains. A newly emergent variant under strong positive selection increases faster than linearly in frequency[45]. Given a large N_e, implying efficient selection, additional days of infection will have a disproportionate influence upon the potential transmission of adaptive variants. This does not imply that longer infections are the sole driving force behind global viral adaptation[37]; selective effects affecting viral transmissibility[30] would provide an alternative explanation. However, our work suggests that longer-term infections may be an important area of study in the quest to better understand global influenza virus evolution.

Methods

Summary

In a single-locus haploid system, the expected change in a variant allele with frequency q caused by genetic drift is given by the formula[46]

$$E[\Delta q] = \sqrt{\frac{q(1-q)}{N_e}}$$

This fact has been exploited to evaluate the size of transmission bottlenecks in influenza infection, comparing statistics of genome sequence data collected before and after a

166 transmission event[30,47]. Such a calculation may be affected by noise in the sequencing of a
167 population, particularly where the extent of noise outweighs the genuine change in a
168 population[30]. Noise in sequence data may be caused by unrepresentative sampling of a
169 population or by error in the experimental process itself[29].
170
171 Because of noise, the comparison of two sequence samples is not a good way of establishing
172 Ne if this statistic may be high. Here we look at an alternative statistic, namely the sum change
173 in variant frequencies in a population; this statistic describes the rate of evolution of the
174 population as a whole. We apply a method of linear regression to multiple samples from the
175 population to establish this rate. We then use simulations to identify the value of N_e that, given
176 the diversity of the population, reproduces this rate of evolutionary change. Our approach is
177 robust to noise in the data, the regression calculation identifying the underlying rate of the
178 evolution of the population rather than the simple observed distance between samples.
179
180 *Sequence data and bioinformatics*
181
182 Sequence data describing the evolution of the infection was generated as part of a previous
183 study[32]. Data, edited to remove human genome sequence data, have been deposited in the
184 Sequence Read Archive with BioProject ID PRJNA601176. The HCV data used in validating
185 the sequencing pipeline (see below) was previously deposited in the Sequence Read Archive
186 with BioProject ID PRJNA380188. Processed files describing raw variant frequencies for both
187 datasets are available, along with code used in this project, at https://github.com/cjri/FluBData.
188
189 Short-read data were aligned first to a broad set of influenza sequences. Sequences from this
190 set to which the highest number of reads aligned were identified and used to carry out a
191 second short-read alignment. The SAMFIRE software package was then used to filter the
192 short-read data with a PHRED score cutoff of 30, to identify consensus sequences, and to

193 calculate the number of each nucleotide found at each position in the genome. SAMFIRE is
194 available from https://github.com/cjri/samfire.
195
196 *Calculation of evolutionary rates*
197
198 Variant frequencies at different time points during infection were used to calculate a rate of
199 change in the population over time. For each locus in the genome for which data were
200 collected, we calculated the nucleotide frequencies , describing the frequency at each locus *i*
201 of each of the nucleotides at the time of sampling *t*. We then calculated differences in the
202 frequencies observed at each locus *i* using a generalisation of the Hamming distance
203

$$d_i(t_1, t_2) = \frac{1}{2} \sum_{a \in \{A,C,G,T\}} |q_i^a(t_1) - q_i^a(t_2)|$$

204
205
206 where the term inside the sum indicates the absolute difference between the frequency of
207 allele *a* at locus *i*. The statistic d_i is equal to one in the case of a substitution, for example
208 where only A nucleotides are observed in one sample and only G nucleotides in another.
209 However, in contrast to the Hamming distance it further captures smaller changes in allele
210 frequencies, lesser changes producing values between zero and one, such that a change of
211 a variant frequency from 45% to 55% at a two-allele locus would equate to a distance of 0.1,
212 representing half of the sum of the absolute changes in each of the two frequencies.
213
214 Having calculated d_i statistics for each locus, the total distance between two samples was
215 calculated as
216

$$D(t_1, t_2) = \sum_i d_i(t_1, t_2)$$

217
218 where the sum over *i* was conducted over all loci in the viral genome.

In order to calculate a rate of evolution, between-sample distances were plotted against the separation in time between samples. Linear regression was performed using the Mathematica 11 software package, using the same package to calculate a 97.5% confidence interval for the result. In this case the gradient of the linear model gives the rate of change in time of the genetic distance between samples, averaged across the dataset, providing an estimate of the rate of evolution of the population. The intersection of the line with the vertical axis, equal to the nominal distance between samples at time zero, gives an indication of the extent of 'noise' in the data, which may arise from artefacts in either the sampling or sequencing of viruses from the host[29].

Calculation of synonymous and non-synonymous rates of evolution in the population were calculated in the same manner, with the exception that the distances d_i were calculated over individual nucleotides rather than in a per-locus manner. We calculate

$$d_i^{NS}(t_1,t_2) = \frac{1}{2}\sum_{a\in A_{N,i}} |q_i^a(t_1) - q_i^a(t_2)|,$$

and

$$d_i^{S}(t_1,t_2) = \frac{1}{2}\sum_{a\in A_{S,i}} |q_i^a(t_1) - q_i^a(t_2)|,$$

where $A_{N,i}$ and $A_{S,i}$ are the sets of nucleotides at position *i* in the genome which respectively induce non-synonymous and synonymous changes in the consensus sequence. Synonymous and non-synonymous variants were identified with respect to influenza B protein sequences; a nucleotide substitution was defined as being non-synonymous if it induced a change in the coded protein in at least one viral protein sequence. Mean rates of synonymous

and non-synonymous evolution were expressed as mean values per nucleotide, reflecting the differing numbers of each type of potential substitution in the viral genome.

Estimating the effective population size: Wright-Fisher simulation

We used a Wright-Fisher model to simulate the evolution of viral populations, identifying the population size that gave an equivalent rate of evolution to the real data[13]. Data from the viral population gives an estimated rate of evolution per day whereas a Wright-Fisher simulation gives an estimated rate of evolution per generation. We therefore scaled the former to match the experimentally ascertained estimate of 10 hours per generation for influenza B[53].

To conduct our simulation we constructed a population of N viruses. Each simulated virus had a genome comprised of eight segments, each identical in length to the corresponding segment of the influenza B virus sampled from the patient. The genetic composition of the viral population at the beginning of the simulation was determined by the observed frequencies of non-consensus alleles in a random sample collected from the population. At each locus, a multinomial sample of viruses were chosen to be assigned each of the non-consensus alleles in accordance with the observed frequencies. Variant alleles were assigned independently for each locus, with no intrinsic association between alleles. The sample collected on 23rd November 2017 was excluded as a starting point from this analysis due to its low read depth; all other samples had a mean read depth in excess of 2000-fold coverage.

We simulated a single generation of the evolution of our population under genetic drift, generating a random sample of N viruses from the population. We calculated allele frequency data from the initial and final populations, using these to calculate the distance in sequence space through which the population had evolved according to the modified Hamming distance described above.

For each population size tested, our simulation was run 400 times, using the data to produce a 97.5% confidence interval for the extent of evolutionary change at a given effective population size. The extent of evolution of the real population was then compared to the results from our simulated populations, giving an inference of the effective size of the viral population.

Amendments were made to this basic approach.

Accounting for false-positive variants in sequencing: Estimating a false positive rate

The evolutionary distance calculated by our method is dependent upon the extent of diversity in the viral population. Given a greater number of polymorphic alleles in a system, the evolutionary distance, calculated as the sum of allele frequency changes, will also increase. While the experimental pipeline we used has been shown to perform well in capturing within-host viral diversity[54], the possibility remains that sequencing could contribute additional diversity to the initial populations used in our simulation. We therefore made an estimate of the extent to which our sequencing process led to the false identification of variants.

To achieve this we used data from a previous study describing the repeat sequencing of hepatitis C virus (HCV) samples from a host[29]; data in this previous study were collected using the same sequencing pipeline as that used to collect the data considered here and therefore provide a generic measure of the level of false positive variation. The data we analysed, coded as HCV01 in the original study, comprised four clinical HCV samples, each of which was split following nucleic acid extraction. Some replicate samples were processed using a DNase depletion method before all samples went through cDNA synthesis, library preparation and sequencing. DNase depletion led to samples with lower read depth; we here compared sequence data collected from the non-depleted replicates of each sample. Variant

299 frequencies within this dataset, where variation was observed in more than one sample, are
300 shown in Fig. 2S4.
301
302 Considering the real viral sample, we note that at any given genetic locus, a minority variant
303 either exists or does not exist according to some well-defined criterion. (For the moment the
304 way in which variation is defined is not important; methods for defining variation, which include
305 the use of a frequency threshold, are discussed later.) We denote the possible states of a
306 locus as P and N, according to whether the locus is positive or negative for variation. We
307 suppose that the probability that a random locus in the genome has a minority variant is given
308 by P_P, leading to the equivalent statistic $P_N = 1- P_P$.
309
310 Sequencing of a specific position in the genome results in the observation or non-observation
311 of a variant. In our data we have sets of two replicate observations of each position in the
312 genome, giving for each minority variant the possible outcomes VV, VX, XV, and XX, where
313 V corresponds to the observation of a variant, and X corresponds to the non-observation of a
314 variant. These observations contain errors; we denote the true positive, false positive, true
315 negative and false negative rates of the variant identification process by $P_{V|P}$, $P_{V|N}$, $P_{X|N}$, and
316 $P_{X|P}$ respectively. In this notation, V|P indicates the observation of a variant conditional on the
317 variant being a true positive.
318
319 The underlying purpose of our calculation is to remove falsely detected variation from the
320 population. We begin by assuming that the false negative rate of detecting variants is equal
321 to zero. That is, where we do not see a variant in the sequence data, we assume that a variant
322 is never actually present. This is a conservative step in so far as we never add unobserved
323 variation to the population. Our assumption gives the result that the false negative rate, $P_{X|P}$
324 $= 0$. In so far that a variant is never unobserved it follows that the true positive rate $P_{V|P} = 1$.
325

We may now construct expressions for the probabilities of observing each of the four possible outcomes. Noting that $P_{V|N} + P_{X|N} = 1$ we obtain

$$P_{VV} = P_P P_{V|P}^2 + (1 - P_P) P_{V|N}^2 = P_P + (1 - P_P) P_{V|N}^2$$

$$P_{VX} = P_{XV} = P_P P_{X|P} P_{V|P} + (1 - P_P) P_{X|N} P_{V|N} = (1 - P_P)(1 - P_{V|N}) P_{V|N}$$

$$P_{XX} = P_P P_{X|P}^2 + (1 - P_P) P_{X|N}^2 = (1 - P_P)(1 - P_{V|N})^2$$

Thus the outcome probabilities may be expressed in terms of the underlying probability of a position having a variant, P_P, and the false positive rate $P_{V|N}$.

We next processed our sequence replicate data, considering only sites that were sequenced to a read depth of at least 2000-fold coverage. For each locus in a dataset, we calculated the observed frequency of each of the nucleotides A, C, G, and T, generating pairs which described these frequencies in each of our two replicate datasets. Removing pairs in which an allele has a frequency of more than 0.5 in either of the two datasets, we obtained a list of minority variants from each locus, generally comprising three allele frequency pairs per locus. If it is correct that two of the three minority alleles have very low frequencies, the frequencies are close to being statistically independent; the existence of a very few alleles of one minority type does not greatly affect the probability of another variant allele being observed in another read. We note that, of the more than 73 thousand sites sequenced, only 56, fewer than 0.1%, had more than one minority variant at a frequency greater than 1%. We proceeded on the assumption that each pair of minority frequencies was statistically independent of the others.

From the repeated observations of sites, we may count the number of observations of each of the four outcomes; given a total of N pairs we denote these as N_{VV}, N_{VX}, N_{XV}, and N_{XX}. Under

our model of independent pairs we constructed the multinomial log likelihood of the underlying variant and false positive rates.

$$L(P_P, P_{V|N}) = \log\left[\binom{N}{N_{VV}N_{VX}N_{XV}N_{XX}} P_{VV}^{N_{VV}} P_{VX}^{N_{VX}} P_{XV}^{N_{XV}} P_{XX}^{N_{XX}}\right]$$

where the terms P_{ab} are constructed from P_P and $P_{V|N}$ according to the equations above.

Given a set of paired observations, we calculated the maximum likelihood values of P_P and $P_{V|N}$. From these statistics we are able to calculate the positive predictive value of sequencing, namely the proportion of observed variants that are true positives. This is achieved by dividing the probability that a true positive was detected (as $P_{V|P}$ = 1, equal to the probability that a locus has a minority variant), by the probability that a variant was detected:

$$\mathrm{PPV} = \frac{P_P}{P_P + (1 - P_P)P_{V|N}}$$

Frequency dependence of false-positive variant calling

Within our data, our expectation was that minority variants at higher allele frequencies would be more likely to be observed as variants in both replicate samples. We note that, where a frequency cutoff is applied to identify variants, care is required in the above protocol. For example, if a hard threshold was applied, in which variants were called at 1% frequency, a variant that was detected at frequencies of 1.01% and 0.99% would be regarded as having been observed in one case, and not observed in the other, although it likely represents a consistent observation.

In order to assess the frequency dependence of our true positive rate we defined minimum and maximum variant frequency thresholds q^{min} and q^{max}, and denoted the replicate observations of a minority variant frequency as q^A and q^B in the two samples. We further defined the frequency q^{cut} according to the formula

$$q^{cut} = \min\left\{q^{min}, \max\left\{\frac{q^{min}}{2}, 0.001\right\}\right\}$$

We then defined regions of frequency space as follows:

$$VV: \begin{array}{c} q^A \geq q^{cut}; \quad q^B \geq q^{cut}; \quad q^A + q^B \geq \frac{3q^{min}}{2} \\ q^A < q^{max}; \quad q^B < q^{max}; \quad q^A + q^B < \frac{3q^{max}}{2} \end{array}$$

$$VX: q^{min} \leq q^A < q^{max}; \quad q^B < q^{cut}$$

$$XV: q^A < q^{cut}; \quad q^{min} \leq q^B < q^{max}$$

$$XX: q^A < q^{cut}; \quad q^B < q^{cut}; \quad q^A + q^B < \frac{3q^{min}}{2}$$

These inequalities are illustrated in Fig. 2S5.

In the above, q^{cut} functions to slightly harshen the criteria for detecting variants at low frequencies. If a variant is observed in one sample at frequency greater than q^{min}, then if q^{min} is greater than 0.2%, the frequency in the second sample had to be at least half q^{min} to be counted. If q^{min} was between 0.1% and 0.2%, the frequency in the second sample had to be at least 0.1%, while if q^{min} was less than 0.1%, the frequency in the second sample had to be at least q^{min}.

For different ranges of frequency values, q^{min} and q^{max}, the proportion of observed variants that were true positives was calculated according to the maximum likelihood method above, using these categorisations. Results are shown in Fig. 2S6. In the process of setting up the initial state of our Wright-Fisher simulation variants observed in the sequence data were considered in turn, drawing a Bernoulli random variable for each variant. Variants were included in the initial simulated population with probability equal to the proportion of observed variants that were estimated to be true positives.

Accounting for mutation-selection balance

To account for our neglect of mutation, a frequency cutoff was applied to our simulation data. Under a pure process of genetic drift, low-frequency variants in our population are likely to die out, reaching a frequency of zero. In a real population, this would not occur, variants being sustained at low frequencies by a balance of mutation and purifying selection[55,56]. To correct for this we post-processed the initial and final frequency values from our simulations before calculating our distance, imposing a minimum minority allele frequency of 0.1%. All changes in allele frequency below this threshold were ignored, such that, for example, if a variant changed from 0.5% to 0%, this was processed after the fact so that the variant changed from 0.5% to 0.1%. The choice of threshold here is conservative; leading to a conservatively low estimate of N_e.

Confidence intervals

Confidence intervals for the effective population size were calculated as the overlap of 97.5% confidence intervals for the evolutionary rates in the observed data, calculated from the regression for the real data, and estimated from the simulated statistics. The overlap of these values gives an approximate 95% confidence interval for N_e.

Approximations in the Wright-Fisher model

In the calculation to set up an initial viral population, the assignment of minority alleles to sequences becomes slow at large population sizes. Our code simulated viral genomes; a variant allele was included into the population by choosing an appropriate proportion of genomes to which the variant was assigned. For greater computational efficiency we used a pseudo-random approach for choosing genomes. Given a population size N, we generated a set P of prime numbers that were each larger than N. Given some desired allele frequency q we wish to choose qN genomes to which to assign the variant. We therefore calculated the set of numbers

$a^k \pmod p$

where p is a prime number sampled at random from the set P, and a is a randomly chosen primitive root of p. Given this choice of a and p, the values a^k (where k is an integer between 1 and p-1) form a pseudorandom permutation of the numbers from 1 to p-1. We constructed a set of qN genomes by choosing genomes indexed in turn by the elements of this set, beginning from k=1, and discarding values greater than N.

To achieve calculations for population sizes larger than 10^7 we implemented a statistical averaging method. We generated a single population of size 10^6, then generated 200 outcomes of a single generation of the same size, recording allele frequencies in each case. In order to simulate a value of N of size r x 10^6 we compared the frequencies of the initial population to the mean frequencies of a random set of r outcomes. This is equivalent of simulating transmission from a population of size r x 10^6 in which the initial population contains r copies of each of one of 10^6 genotypes.

459 *Phylogenetic analysis*

460

461 Consensus sequences of data were analysed using the BEAST2 software package[48].
462 Consensus sequences from each viral segment were concatenated then aligned using
463 MUSCLE[49] before performing a phylogenetic analysis on the whole genome sequence
464 alignment. The B/Venezuela/02/2016 sequence was used to root the alignment, the
465 haemagglutinin segment of this virus having been identified as being very close to those from
466 the patient. Trees were generated using the HKY substitution model[50]. A Monte Carlo process
467 was run for 10 million iterations, generating a consensus tree with TreeAnnotator using the
468 first 10% of trees as burn-in. Figures were made using the FigTree package
469 (http://tree.bio.ed.ac.uk/software/figtree/).

470

471 *Haplotype reconstruction*

472

473 Haplotype reconstruction was performed using multi-locus polymorphism data generated by
474 the SAMFIRE software package[51]. Variant loci in the genome were identified as those at
475 which a change in the consensus nucleotide was observed between the initial and the final
476 consensus. The short-read data were then processed, converting reads into strings of alleles
477 observed at these loci; a single paired-end read may describe alleles at none, one, or multiple
478 loci. Next, these strings were combined using a combinatorial algorithm to construct a list of
479 single-segment haplotypes, sufficient to explain all of the observed data; no frequencies were
480 inferred at this point. Finally, a Dirichlet-multinomial model was used to infer the maximum
481 likelihood frequencies of each haplotype given the data from each time point[52]. Formally, we
482 divided reads into sets, according to the loci at which they described alleles. A multi-locus
483 variant consists of an observation of some specific alleles at the loci in question. By way of
484 notation, we denote by the number of reads in set i which describe the multi-locus variant a,
485 and denote the total number of reads in the set as N_i. Given a set of haplotypes with
486 frequencies given by the elements of the vector $\boldsymbol{q}$, we write as the summed frequencies of

haplotypes that match each multi-locus variant *a* in set *i*. For example, the haplotypes ATA and ATG would both match the multi-locus variant AT- describing alleles at only the first two loci. We now express a likelihood for the haplotype frequencies:

$$\mathcal{L}(\boldsymbol{q}) = \sum_i \log \frac{\Gamma(N_i+1)}{\prod_a \Gamma(n_i^a+1)} \frac{\Gamma(\sum_a Cq_i^a)}{\Gamma(\sum_a n_i^a + Cq_i^a)} \prod_a \frac{\Gamma(n_i^a + Cq_i^a)}{\Gamma(Cq_i^a)}$$

Here the parameter C describes the extent of noise in the sequence data, a lower value indicating a lower confidence in the sequence data. Haplotype reconstruction was performed by finding the maximum likelihood value of the vector of haplotype frequencies **q**. A value of C=200 was chosen for the calculation, representing a conservative estimate given the prior performance of the sequencing pipeline used in this study[29]. In contrast to previous calculations in which an evolutionary model was fitted to data[52], haplotype frequencies for each time point and for each viral segment were in this case inferred independently, with no underlying evolutionary model.

Data describing influenza A/H3N2 infection

Our analysis of data describing long-term influenza A/H3N2 infection was performed on data from a previous study[37]. As our method does not require an exceptional quality of sequencing data to calculate a rate of evolution more samples were included in our analysis than were examined in the original study. Using the codes established in the previous study, we used samples from patient W from days 0, 7, 14, 21, 28, 56, 62, 67 and 76; from patient X from days 0, 7, 14, 21, 28, 42, and 72; from patient Y from days 0, 7, 14, 21, 28, 35, 48, 56, and 70; from patient Z from days 14, 15, 20, 25, 41, 48, 55, 62, and 69. An identical procedure to that used to estimate Ne from the influenza B data was applied, calculating a rate of evolution per day from sequence data, scaling this to a rate per generation (in this case a seven hour generation time was modelled[53]), and then running simulations to estimate N_e. We note that

514 the estimates of false positive rate generated for the influenza B data were applied equally in
515 this case, due to not having equivalent data to re-estimate these values. Examining the data
516 from patient W, our distance measurements suggested potential population structure involving
517 the samples collected on days 62 and 69; these samples were excluded from our regression
518 analysis.
519

Figure Legends

Figure 1. Population structure of the influenza infection. A. CT values from viral samples collected over time indicate the viral load of the infection; a higher number corresponds to a lower viral load. Drug information, above, shows the times during which oseltamivir (green), zanamivir (yellow), nitazoxanide (blue) and favipiravir (purple) were prescribed. Black dots show samples from which viral sequence data were collected; gray dots show samples from which viral sequence data were not collected. The green box shows the window of time over which samples were analysed, preceding the use of favipiravir in January. The mean viral load (dashed horizontal line, red) was close to the mean reported for a set of samples from hospitalised children with influenza (dashed horizontal blue line)[31]. A black arrow shows the date of a bone marrow transplant (BMT). **B.** A phylogeny of whole-genome viral consensus sequences identified two distinct clades in the viral population. Clade B featured three samples, distributed across the period of infection, with the remaining samples contained in Clade A. **C.** Sub-consensus structure of the viral population inferred via a haplotype reconstruction algorithm using data from the neuraminidase segment. The same division of sequences into two clades is visible, with samples being comprised of distinct viral genotypes. The area of each circle is proportional to the inferred frequency of the corresponding haplotype in the viral population. Haplotypes reaching a frequency of at least 10% in at least one time point are shown. Multiple drugs were administered to the patient through time, with a favipiravir/zanamivir combination first causing a temporary reduction of the population to undetectable levels, then finally clearing the infection. Haplotypes spanned the loci 96, 170, 177, 402, 403, 483, 571, 653, 968, 973, 1011, 1079, 1170, and 1240 in the NA segment. **D.** Evolutionary relationship between the haplotypes; clade B is distinct from and evolves away from those sequences comprising the initial infection. Numbers refer to the distinct haplotypes identified within the population.

548 **Figure 2 : Measuring rates of evolution in the viral population. A.** Computed rate of
549 evolution for viruses in clade A up to the time of the first use of favipiravir. The distance
550 between two sequences is calculated as the total absolute difference in four-allele frequencies
551 measured across the genome. The calculated rate per generation is based upon a generation
552 time for influenza of 10 hours[53]. **B.** Distribution of evolutionary distances in influenza
553 populations simulated using a Wright-Fisher model compared to the distance per generation
554 calculated in the regression fit. A solid blue line shows the mean, with shading indicating an
555 approximate 97.5% confidence interval around the mean. Statistics were calculated from sets
556 of 400 simulations conducted at each value of N_e. The dashed black line shows the rate of
557 evolution of the real population; gray shading shows a 97.5% confidence interval for this
558 statistic. **C.** Calculated rate of evolution for viruses in clade B. For the purposes of calculating
559 a rate of evolution the first sample collected from the patient was included as part of clade B.
560 **D.** Estimation of N_e for clade B. The results of simulations shown here are identical to those
561 in part B of the figure.
562
563 **Figure 1 supplement 1.** Complete phylogeny of whole-genome viral consensus sequences,
564 coloured by clade.
565
566 **Figure 1 supplement 2: A.** Sub-consensus structure of the viral population inferred via a
567 haplotype reconstruction algorithm using data from the haemagglutinin segment. A division
568 of sequences into two clades is visible, with samples including largely distinct viral genotypes.
569 The area of each circle is proportional to the amount of virus in each clade. Haplotypes
570 reaching a frequency of at least 10% in at least one time point are shown. Haplotypes spanned
571 the loci 258, 261, 364, 451, 521, 541, 635, and 641 in the HA segment. **B.** Evolutionary
572 relationship between the haplotypes; clade B is distinct from and evolves away from those
573 sequences comprising the initial infection. Numbers refer to the distinct haplotypes identified
574 within the population.
575

576 **Figure 2 supplement 1:** Amino acids present at codon 117 of the neuramindase segment of
577 the virus after the first administration of zanamivir. The consensus glutamate nucleotide (blue)
578 was sometimes replaced by glycine (green), valine (yellow), and alanine (red). Glycine and
579 alanine are associated with zanamivir resistance in influenza B.
580
581 **Figure 2 supplement 2: A.** Comparison of rates of synonymous and non-synonymous
582 evolution for viruses in clade A up to the time of the administration of favipiravir. The distance
583 between two samples is calculated as the mean absolute difference in allele frequency, as
584 averaged over synonymous and non-synonymous positions in the genome. B. Comparison
585 of rates of synonymous and non-synonymous evolution for viruses in clade B. The rate of
586 evolution in both clades was slower at non-synonymous sites than at synonymous sites,
587 suggesting a general pattern of purifying selection at non-synonymous sites. Change in the
588 population was not as a whole driven by positive selection.
589
590 **Figure 2 supplement 3:** Estimates of the effective population size for data from a study of
591 long-term influenza A/H3N2 infection in four patients. Patients are denoted with the letters
592 assigned them in the original study[37]. Rates of evolution within each patient were calculated
593 by linear regression, conducted on a plot of evolutionary versus temporal distance between
594 samples. The inferred regression line is shown in red for each dataset. For Patient W samples
595 collected at two time points appear as outliers in the distance plot; distances involving these
596 samples, shown in yellow, were excluded from the calculation. Accompanying plots show
597 distances inferred via simulation compared to the inferred rates. A solid blue line shows the
598 mean, with shading indicating an approximate 97.5% confidence interval around the mean.
599 Statistics were calculated from sets of 400 simulations conducted at each value of N_e.
600
601 **Figure 2 supplement 4:** Frequencies of minority variant alleles identified in the HCV01
602 dataset used to evaluate the accuracy of variant calling in our sequencing pipeline. Samples
603 in this dataset were split following RNA extraction with replicate sets of RNA being processed

and sequenced independently. Variants at higher frequencies were identified at more consistent frequencies than variants at lower frequencies.

Figure 2 supplement 5: Regions of frequency space used to define observations and non-observations of allele frequencies. V indicates the identification of a variant, while X indicates the non-identification of a variant. Combinations of V and X indicate observations made in two replicate samples.

Figure 2 supplement 6: Positive predictive value for minority variants under our sequencing pipeline, calculated at different frequency ranges. While high frequency variants were very reliably identified, the reliability of identifying variants was significantly impaired at lower frequencies.

References

1. Buonagurio, D. A. *et al.* Evolution of human influenza A viruses over 50 years: rapid, uniform rate of change in NS gene. *Science* **232**, 980–982 (1986).
2. Fitch, W. M., Bush, R. M., Bender, C. A. & Cox, N. J. Long term trends in the evolution of H(3) HA1 human influenza type A. *Proc. Natl. Acad. Sci. U. S. A.* **94**, 7712–7718 (1997).
3. Bedford, T. *et al.* Global circulation patterns of seasonal influenza viruses vary with antigenic drift. *Nature* **523**, 217–220 (2015).
4. Ferguson, N. M., Galvani, A. P. & Bush, R. M. Ecological and immunological determinants of influenza evolution. *Nature* **422**, 428–433 (2003).
5. Grenfell, B. T. *et al.* Unifying the epidemiological and evolutionary dynamics of pathogens. *Science* **303**, 327–332 (2004).

631 6. Debbink, K. *et al.* Vaccination has minimal impact on the intrahost diversity of H3N2
632 influenza viruses. *PLoS Pathog.* **13**, e1006194 (2017).
633 7. McCrone, J. T. *et al.* Stochastic processes constrain the within and between host
634 evolution of influenza virus. *Elife* **7**, (2018).
635 8. Han, A. X., Maurer-Stroh, S. & Russell, C. A. Individual immune selection pressure has
636 limited impact on seasonal influenza virus evolution. *Nat Ecol Evol* **3**, 302–311 (2019).
637 9. Rouzine, I. M., Rodrigo, A. & Coffin, J. M. Transition between stochastic evolution and
638 deterministic evolution in the presence of selection: general theory and application to
639 virology. *Microbiol. Mol. Biol. Rev.* **65**, 151–185 (2001).
640 10. Wright, S. Size of population and breeding structure in relation to evolution. *Science* **87**,
641 430–431 (1938).
642 11. Wang, J., Santiago, E. & Caballero, A. Prediction and estimation of effective population
643 size. *Heredity* **117**, 193–206 (2016).
644 12. Khatri, B. S. & Burt, A. Robust Estimation of Recent Effective Population Size from
645 Number of Independent Origins in Soft Sweeps. *Mol. Biol. Evol.* **36**, 2040–2052 (2019).
646 13. Kimura, M. & Crow, J. F. The Measurement of Effective Population Number. *Evolution*
647 vol. 17 279 (1963).
648 14. Fitch, W. M., Leiter, J. M., Li, X. Q. & Palese, P. Positive Darwinian evolution in human
649 influenza A viruses. *Proc. Natl. Acad. Sci. U. S. A.* **88**, 4270–4274 (1991).
650 15. Rambaut, A. *et al.* The genomic and epidemiological dynamics of human influenza A
651 virus. *Nature* vol. 453 615–619 (2008).
652 16. Strelkowa, N. & Lässig, M. Clonal interference in the evolution of influenza. *Genetics*
653 **192**, 671–682 (2012).
654 17. Bedford, T., Cobey, S., Beerli, P. & Pascual, M. Global migration dynamics underlie
655 evolution and persistence of human influenza A (H3N2). *PLoS Pathog.* **6**, e1000918
656 (2010).
657 18. Bollback, J. P., York, T. L. & Nielsen, R. Estimation of 2Nes from temporal allele
658 frequency data. *Genetics* **179**, 497–502 (2008).

659 19. Feder, A. F., Kryazhimskiy, S. & Plotkin, J. B. Identifying signatures of selection in
660 genetic time series. *Genetics* **196**, 509–522 (2014).
661 20. Foll, M. *et al.* Influenza virus drug resistance: a time-sampled population genetics
662 perspective. *PLoS Genet.* **10**, e1004185 (2014).
663 21. Terhorst, J., Schlötterer, C. & Song, Y. S. Multi-locus analysis of genomic time series
664 data from experimental evolution. *PLoS Genet.* **11**, e1005069 (2015).
665 22. Rousseau, E. *et al.* Estimating virus effective population size and selection without
666 neutral markers. *PLoS Pathog.* **13**, e1006702 (2017).
667 23. Laporte, V. & Charlesworth, B. Effective population size and population subdivision in
668 demographically structured populations. *Genetics* **162**, 501–519 (2002).
669 24. Lakdawala, S. S. *et al.* The soft palate is an important site of adaptation for
670 transmissible influenza viruses. *Nature* **526**, 122–125 (2015).
671 25. Sobel Leonard, A. *et al.* The effective rate of influenza reassortment is limited during
672 human infection. *PLoS Pathog.* **13**, e1006203 (2017).
673 26. Richard, M., Herfst, S., Tao, H., Jacobs, N. T. & Lowen, A. C. Influenza A Virus
674 Reassortment Is Limited by Anatomical Compartmentalization following Coinfection via
675 Distinct Routes. *J. Virol.* **92**, (2018).
676 27. Hamada, N. *et al.* Intrahost emergent dynamics of oseltamivir-resistant virus of
677 pandemic influenza A (H1N1) 2009 in a fatally immunocompromised patient. *J. Infect.*
678 *Chemother.* **18**, 865–871 (2012).
679 28. McCrone, J. T. & Lauring, A. S. Measurements of Intrahost Viral Diversity Are Extremely
680 Sensitive to Systematic Errors in Variant Calling. *J. Virol.* **90**, 6884–6895 (2016).
681 29. Illingworth, C. J. R. *et al.* On the effective depth of viral sequence data. *Virus Evol* **3**,
682 vex030 (2017).
683 30. Lumby, C. K., Nene, N. R. & Illingworth, C. J. R. A novel framework for inferring
684 parameters of transmission from viral sequence data. *PLoS Genet.* **14**, e1007718
685 (2018).
686 31. Wishaupt, J. O. *et al.* Pitfalls in interpretation of CT-values of RT-PCR in children with

687 acute respiratory tract infections. *J. Clin. Virol.* **90**, 1–6 (2017).

688 32. Lumby, C. K. *et al.* Favipiravir and Zanamivir Cleared Infection with Influenza B in a
689 Severely Immunocompromised Child. *Clin. Infect. Dis.* (2020) doi:10.1093/cid/ciaa023.

690 33. Pennings, P. S., Kryazhimskiy, S. & Wakeley, J. Loss and recovery of genetic diversity
691 in adapting populations of HIV. *PLoS Genet.* **10**, e1004000 (2014).

692 34. Rouzine, I. M., Coffin, J. M. & Weinberger, L. S. Fifteen years later: hard and soft
693 selection sweeps confirm a large population number for HIV in vivo. *PLoS Genet.* **10**,
694 e1004179 (2014).

695 35. Whitlock, M. C. & Barton, N. H. The effective size of a subdivided population. *Genetics*
696 **146**, 427–441 (1997).

697 36. Jackson, D., Barclay, W. & Zürcher, T. Characterization of recombinant influenza B
698 viruses with key neuraminidase inhibitor resistance mutations. *Journal of Antimicrobial*
699 *Chemotherapy* vol. 55 162–169 (2005).

700 37. Xue, K. S. *et al.* Parallel evolution of influenza across multiple spatiotemporal scales.
701 *Elife* **6**, (2017).

702 38. Valesano, A. L. *et al.* Influenza B Viruses Exhibit Lower Within-Host Diversity than
703 Influenza A Viruses in Human Hosts. *J. Virol.* **94**, (2020).

704 39. Zhao, L., Abbasi, A. B. & Illingworth, C. J. R. Mutational load causes stochastic
705 evolutionary outcomes in acute RNA viral infection. *Virus Evol* **5**, vez008 (2019).

706 40. Gubareva, L. V., Matrosovich, M. N., Brenner, M. K., Bethell, R. C. & Webster, R. G.
707 Evidence for Zanamivir Resistance in an Immunocompromised Child Infected with
708 Influenza B Virus. *The Journal of Infectious Diseases* vol. 178 1257–1262 (1998).

709 41. Snydman, D. R. Oseltamivir Resistance During Treatment of Influenza A (H5N1)
710 Infection. *Yearbook of Medicine* vol. 2006 70–71 (2006).

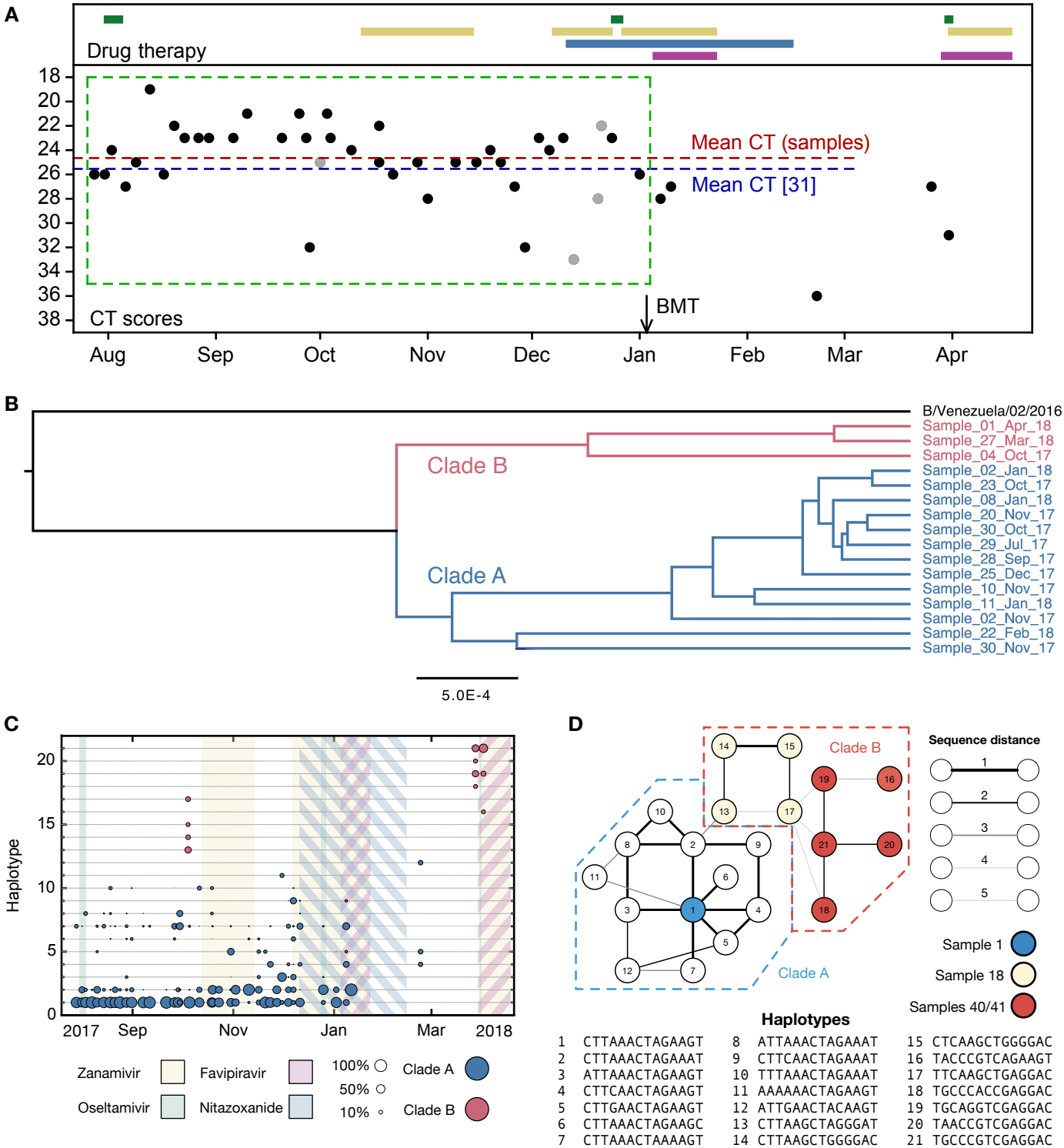
711 42. Centers for Disease Control and Prevention (CDC). Oseltamivir-resistant novel
712 influenza A (H1N1) virus infection in two immunosuppressed patients - Seattle,
713 Washington, 2009. *MMWR Morb. Mortal. Wkly. Rep.* **58**, 893–896 (2009).

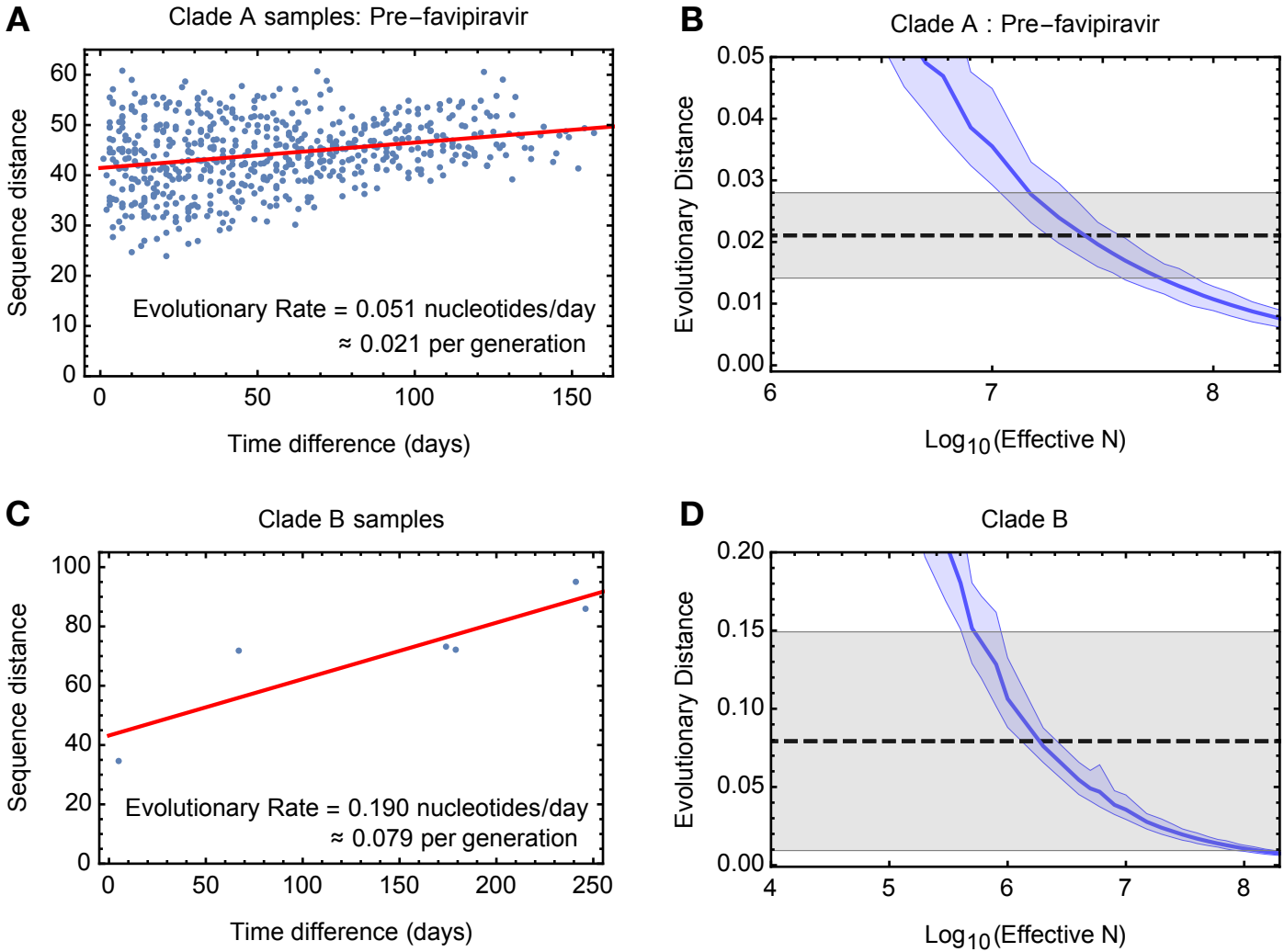
714 43. Imai, M. *et al.* Influenza A variants with reduced susceptibility to baloxavir isolated from

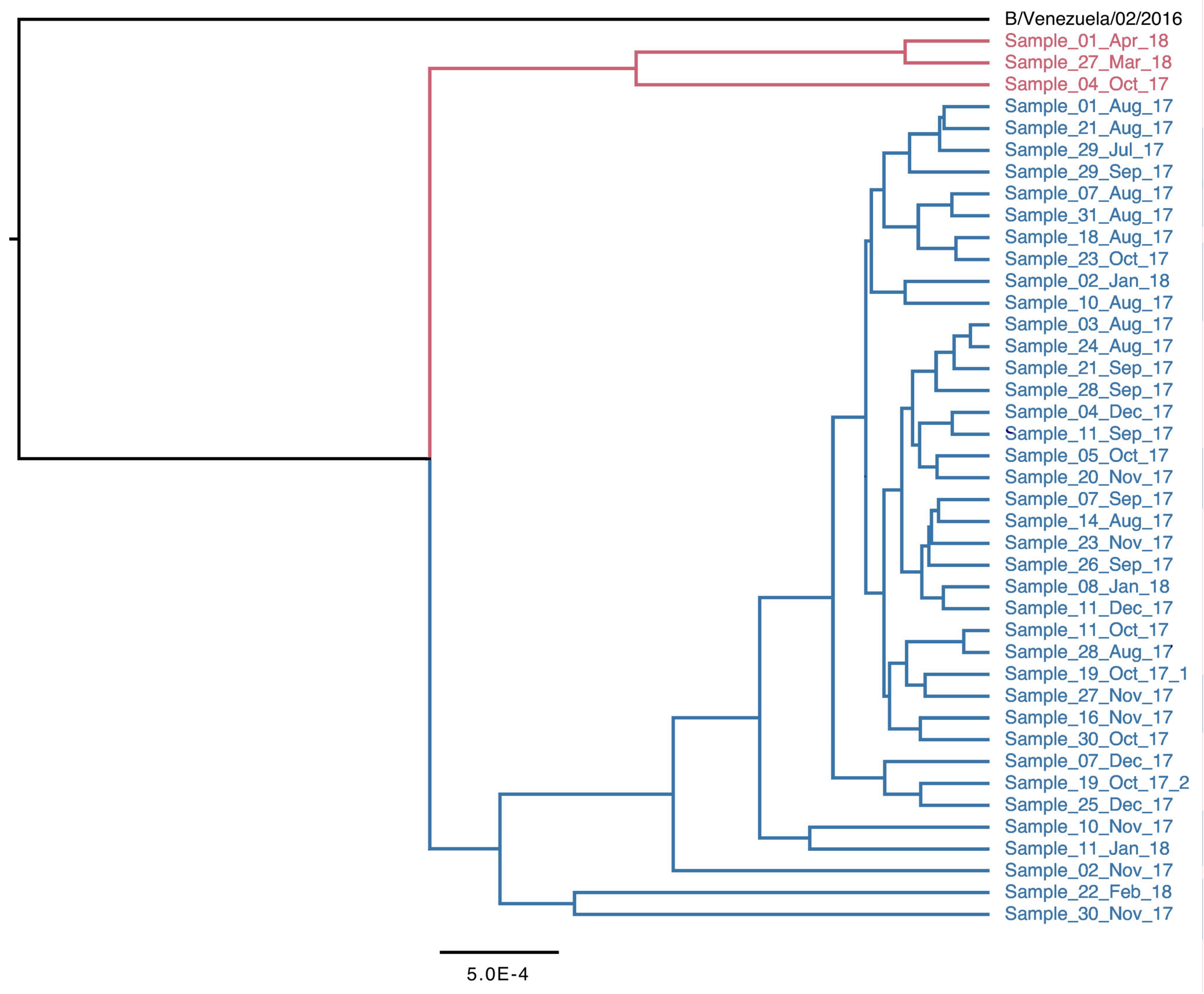
715 Japanese patients are fit and transmit through respiratory droplets. *Nat Microbiol* (2019)
716 doi:10.1038/s41564-019-0609-0.
717 44. Rogers, M. B. *et al.* Intrahost Dynamics of Antiviral Resistance in Influenza A Virus
718 Reflect Complex Patterns of Segment Linkage, Reassortment, and Natural Selection.
719 *mBio* vol. 6 (2015).
720 45. Haldane, J. B. S. A mathematical theory of natural and artificial selection. Part I. *Trans.*
721 *Camb. Phil. Soc* **23**, 19–41 (1924).
722 46. Charlesworth, B. Fundamental concepts in genetics: effective population size and
723 patterns of molecular evolution and variation. *Nat. Rev. Genet.* **10**, 195–205 (2009).
724 47. Sobel Leonard, A., Weissman, D. B., Greenbaum, B., Ghedin, E. & Koelle, K.
725 Transmission Bottleneck Size Estimation from Pathogen Deep-Sequencing Data, with
726 an Application to Human Influenza A Virus. *J. Virol.* **91**, (2017).
727 48. Bouckaert, R. *et al.* BEAST 2: A Software Platform for Bayesian Evolutionary Analysis.
728 *PLoS Computational Biology* vol. 10 e1003537 (2014).
729 49. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high
730 throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
731 50. Hasegawa, M., Kishino, H. & Yano, T. Dating of the human-ape splitting by a molecular
732 clock of mitochondrial DNA. *J. Mol. Evol.* **22**, 160–174 (1985).
733 51. Illingworth, C. J. R. SAMFIRE: multi-locus variant calling for time-resolved sequence
734 data. *Bioinformatics* **32**, 2208–2209 (2016).
735 52. Illingworth, C. J. R. Fitness Inference from Short-Read Data: Within-Host Evolution of a
736 Reassortant H5N1 Influenza Virus. *Mol. Biol. Evol.* **32**, 3012–3026 (2015).
737 53. Nobusawa, E. & Sato, K. Comparison of the Mutation Rates of Human Influenza A and
738 B Viruses. *Journal of Virology* vol. 80 3675–3678 (2006).
739 54. Thomson, E. *et al.* Comparison of Next-Generation Sequencing Technologies for
740 Comprehensive Assessment of Full-Length Hepatitis C Viral Genomes. *J. Clin.*
741 *Microbiol.* **54**, 2470–2484 (2016).
742 55. Haldane, J. B. S. The Effect of Variation of Fitness. *The American Naturalist* vol. 71

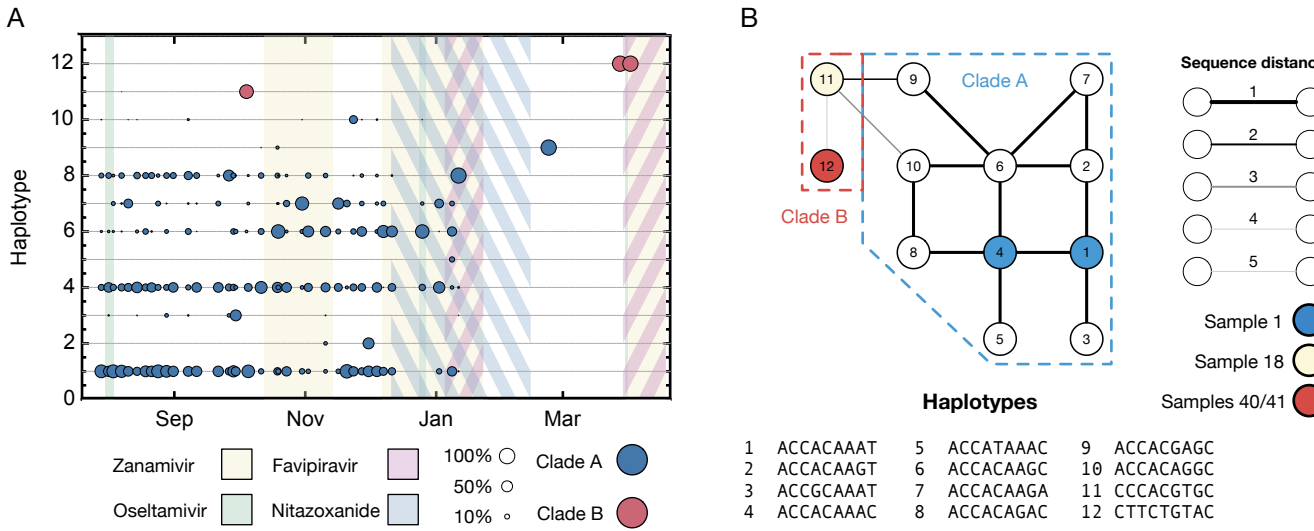
337–349 (1937).

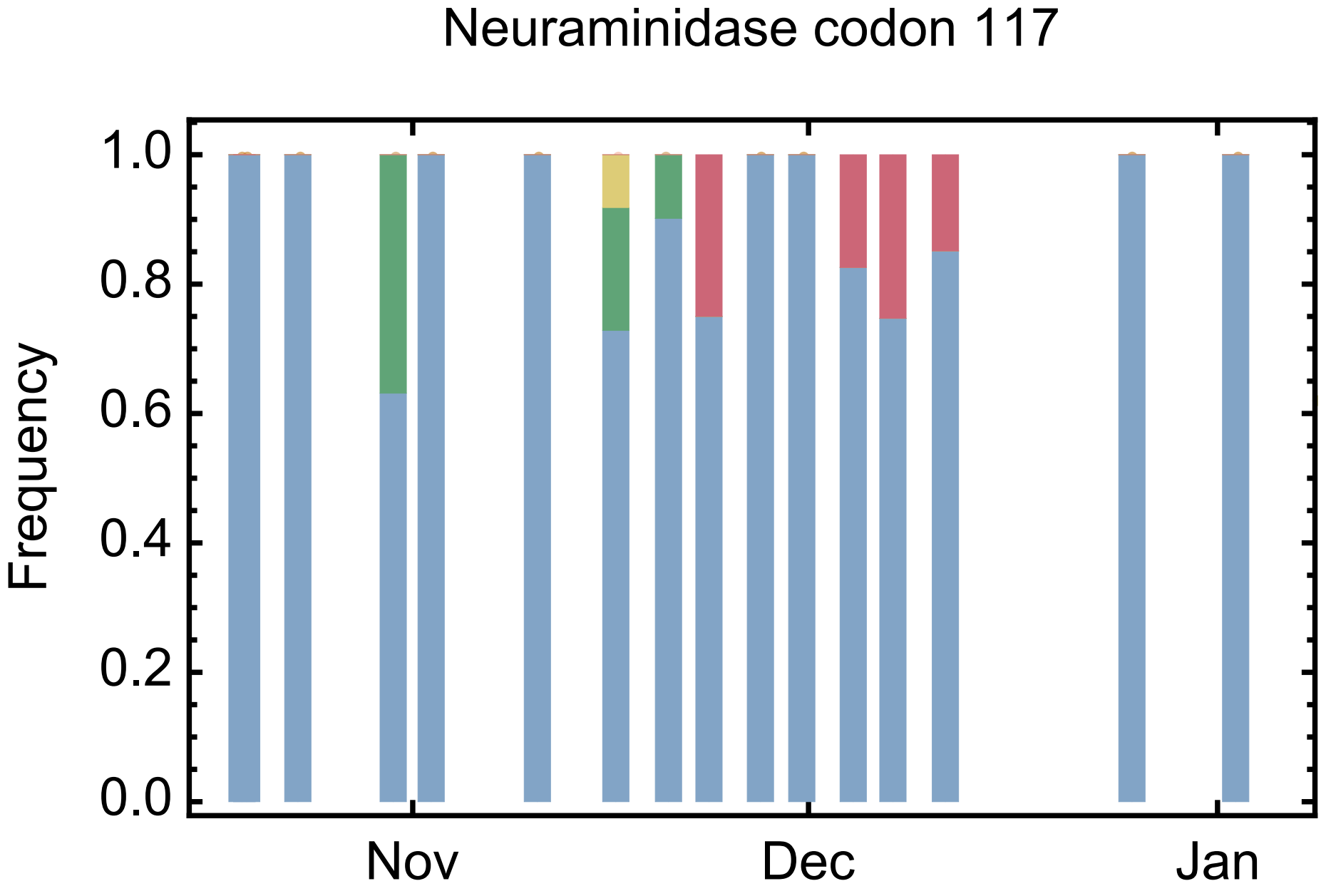
56. Haigh, J. The accumulation of deleterious genes in a population—Muller's Ratchet. *Theoretical Population Biology* vol. 14 251–267 (1978).

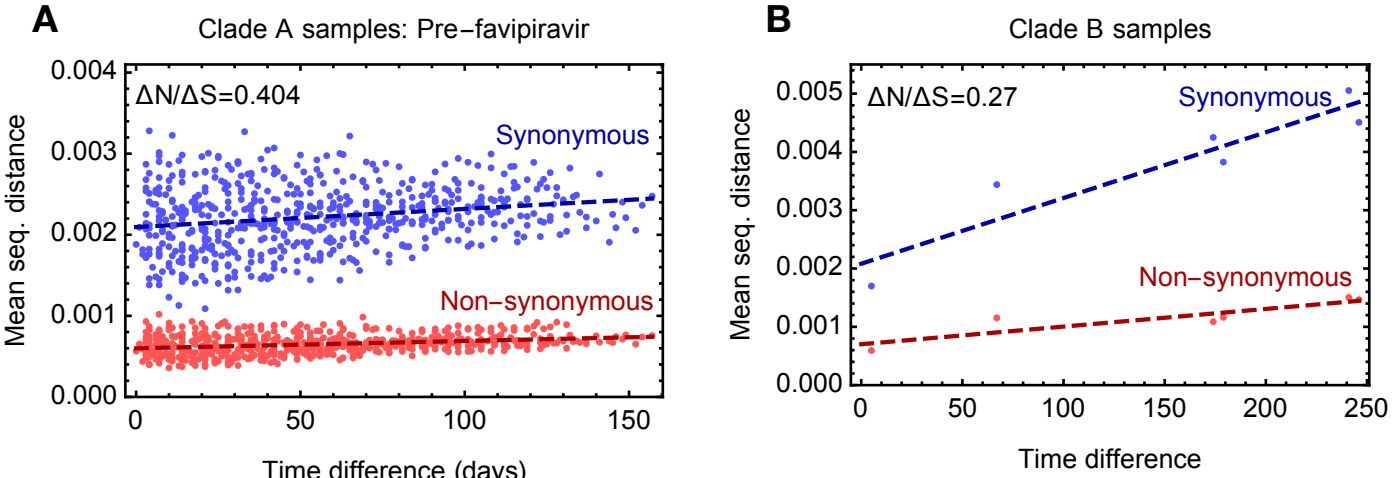

A
Drug therapy
18
20
22
24
26
28
30
32
34
36
38
CT scores
Mean CT (samples)
Mean CT [31]
BMT
Aug
Sep
Oct
Nov
Dec
Jan
Feb
Mar
Apr
B
B/Venezuela/02/2016
Sample_01_Apr_18
Sample_27_Mar_18
Sample_04_Oct_17
Sample_02_Jan_18
Sample_23_Oct_17
Sample_08_Jan_18
Sample_20_Nov_17
Sample_30_Oct_17
Sample_29_Jul_17
Sample_28_Sep_17
Sample_25_Dec_17
Sample_10_Nov_17
Sample_11_Jan_18
Sample_02_Nov_17
Sample_22_Feb_18
Sample_30_Nov_17
Clade B
Clade A
5.0E-4
C
Haplotype
0
5
10
15
20
2017
Sep
Nov
Jan
Mar
2018
Zanamivir
Oseltamivir
Favipiravir
Nitazoxanide
100%
50%
10%
Clade A
Clade B
D
Clade B
Clade A
Sequence distance
1
2
3
4
5
Sample 1
Sample 18
Samples 40/41
Haplotypes
1 CTTAAACTAGAAGT
2 CTTAAACTAGAAAT
3 ATTAAACTAGAAGT
4 CTTCAACTAGAAGT
5 CTTGAACTAGAAGT
6 CTTAAACTAGAAGC
7 CTTAAACTAAAAGT
8 ATTAAACTAGAAAT
9 CTTCAACTAGAAAT
10 TTTAAACTAGAAAT
11 AAAAAACTAGAAGT
12 ATTGAACTACAAGT
13 CTTAAGCTAGGGAT
14 CTTAAGCTGGGGAC
15 CTCAAGCTGGGGAC
16 TACCCGTCAGAAGT
17 TTCAAGCTGAGGAC
18 TGCCCACCGAGGAC
19 TGCAGGTCGAGGAC
20 TAACCGTCGAGGAC
21 TGCCCGTCGAGGAC

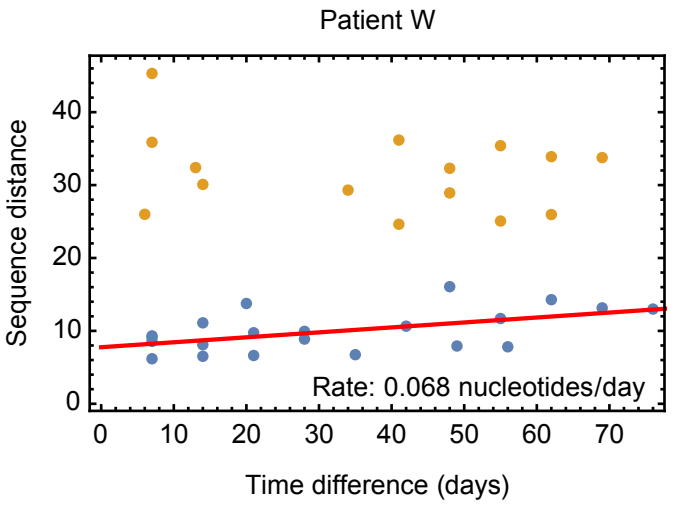

A
Clade A samples: Pre−favipiravir
Sequence distance
Time difference (days)
Evolutionary Rate = 0.051 nucleotides/day
≈ 0.021 per generation
B
Clade A : Pre−favipiravir
Evolutionary Distance
Log10(Effective N)
C
Clade B samples
Sequence distance
Time difference (days)
Evolutionary Rate = 0.190 nucleotides/day
≈ 0.079 per generation
D
Clade B
Evolutionary Distance
Log10(Effective N)

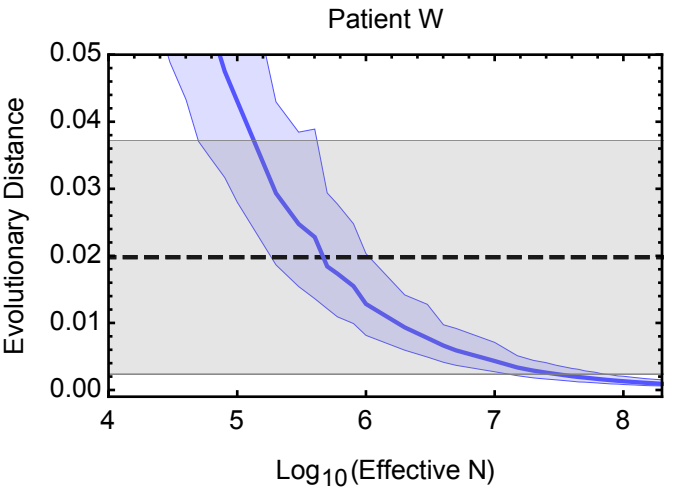

B/Venezuela/02/2016
Sample_01_Apr_18
Sample_27_Mar_18
Sample_04_Oct_17
Sample_01_Aug_17
Sample_21_Aug_17
Sample_29_Jul_17
Sample_29_Sep_17
Sample_07_Aug_17
Sample_31_Aug_17
Sample_18_Aug_17
Sample_23_Oct_17
Sample_02_Jan_18
Sample_10_Aug_17
Sample_03_Aug_17
Sample_24_Aug_17
Sample_21_Sep_17
Sample_28_Sep_17
Sample_04_Dec_17
Sample_11_Sep_17
Sample_05_Oct_17
Sample_20_Nov_17
Sample_07_Sep_17
Sample_14_Aug_17
Sample_23_Nov_17
Sample_26_Sep_17
Sample_08_Jan_18
Sample_11_Dec_17
Sample_11_Oct_17
Sample_28_Aug_17
Sample_19_Oct_17_1
Sample_27_Nov_17
Sample_16_Nov_17
Sample_30_Oct_17
Sample_07_Dec_17
Sample_19_Oct_17_2
Sample_25_Dec_17
Sample_10_Nov_17
Sample_11_Jan_18
Sample_02_Nov_17
Sample_22_Feb_18
Sample_30_Nov_17
5.0E-4

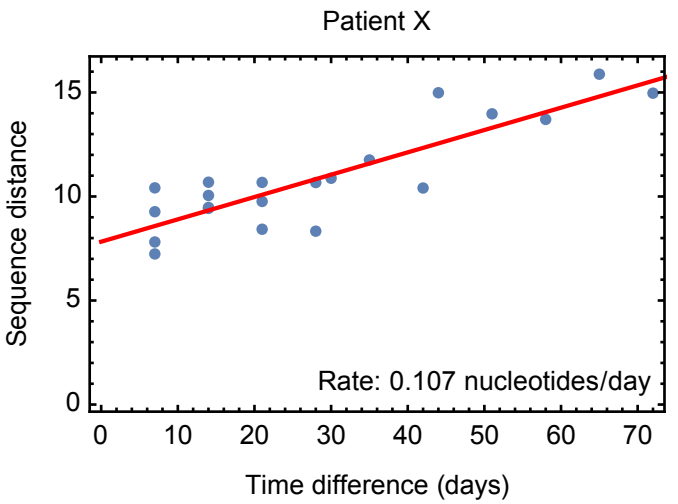

A
Haplotype
0
2
4
6
8
10
12
Sep
Nov
Jan
Mar
Zanamivir
Favipiravir
Oseltamivir
Nitazoxanide
100%
50%
10%
Clade A
Clade B
B
11
9
7
12
10
6
2
8
4
1
5
3
Clade A
Clade B
Haplotypes
Sequence distance
1
2
3
4
5
Sample 1
Sample 18
Samples 40/41
1 ACCACAAAT
2 ACCACAAGT
3 ACCGCAAAT
4 ACCACAAAC
5 ACCATAAAC
6 ACCACAAGC
7 ACCACAAGA
8 ACCACAGAC
9 ACCACGAGC
10 ACCACAGGC
11 CCCACGTGC
12 CTTCTGTAC

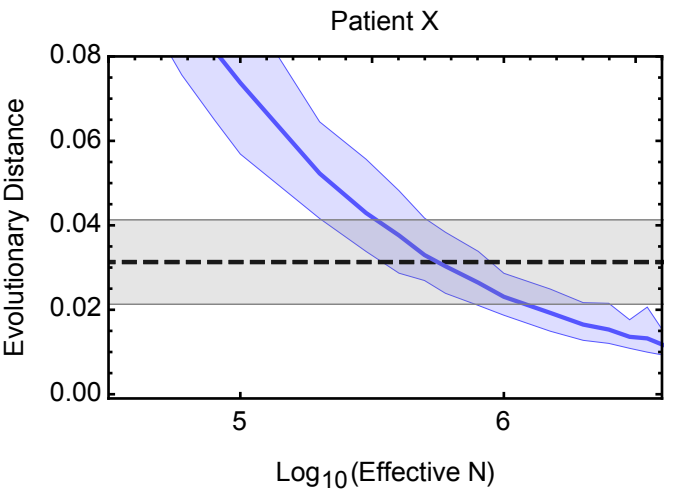

Neuraminidase codon 117
Frequency
1.0
0.8
0.6
0.4
0.2
0.0
Nov
Dec
Jan

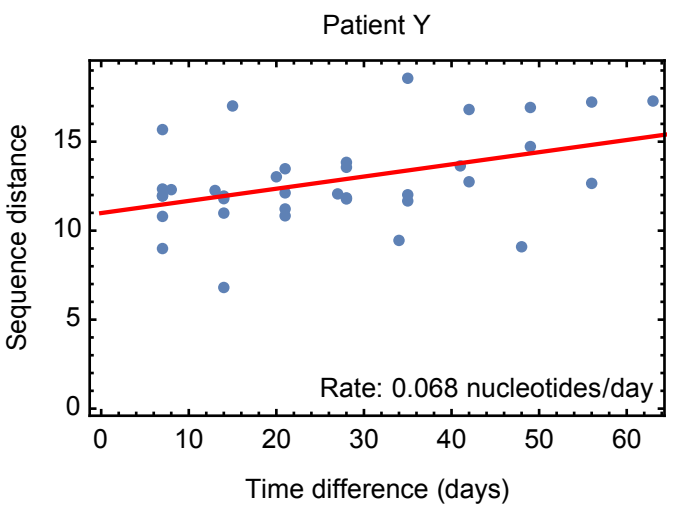

A
Clade A samples: Pre−favipiravir
ΔN/ΔS=0.404
Synonymous
Non−synonymous
Mean seq. distance
Time difference (days)
0.004
0.003
0.002
0.001
0.000
0
50
100
150
B
Clade B samples
ΔN/ΔS=0.27
Synonymous
Non−synonymous
Mean seq. distance
Time difference
0.005
0.004
0.003
0.002
0.001
0.000
0
50
100
150
200
250

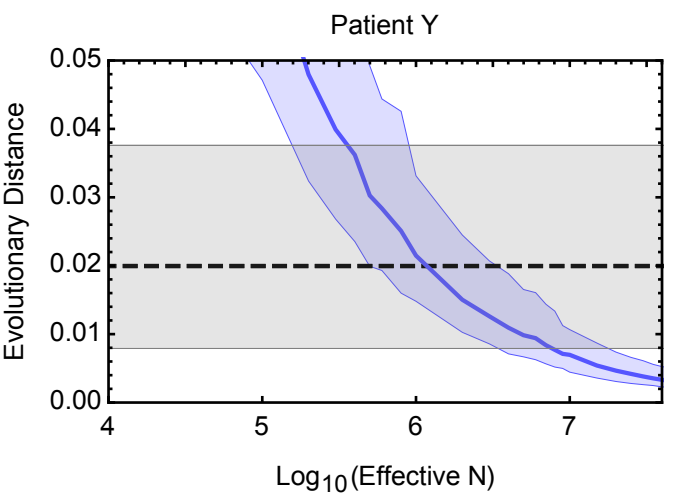

Patient W
Sequence distance
Time difference (days)
Rate: 0.068 nucleotides/day

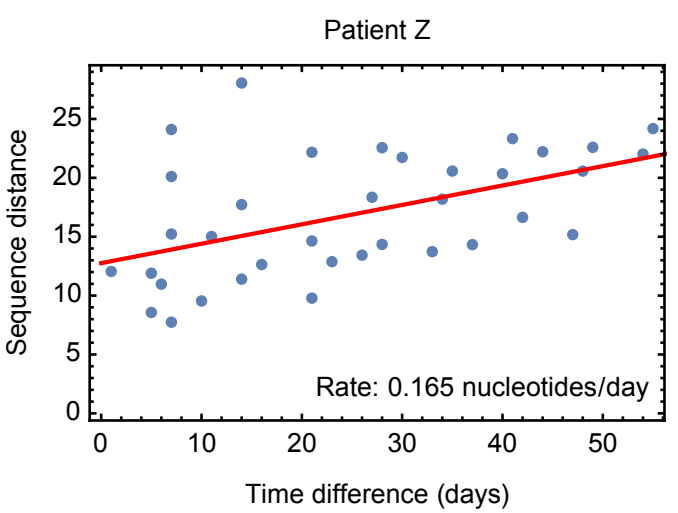

Patient W
Evolutionary Distance
Log10(Effective N)

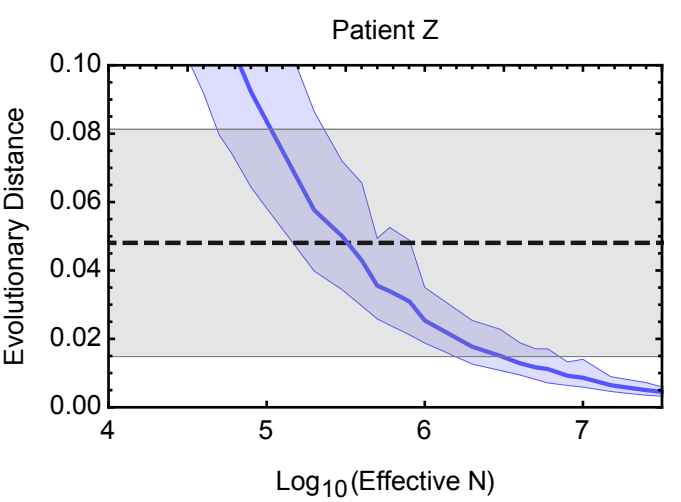

Patient X
Sequence distance
Time difference (days)
Rate: 0.107 nucleotides/day

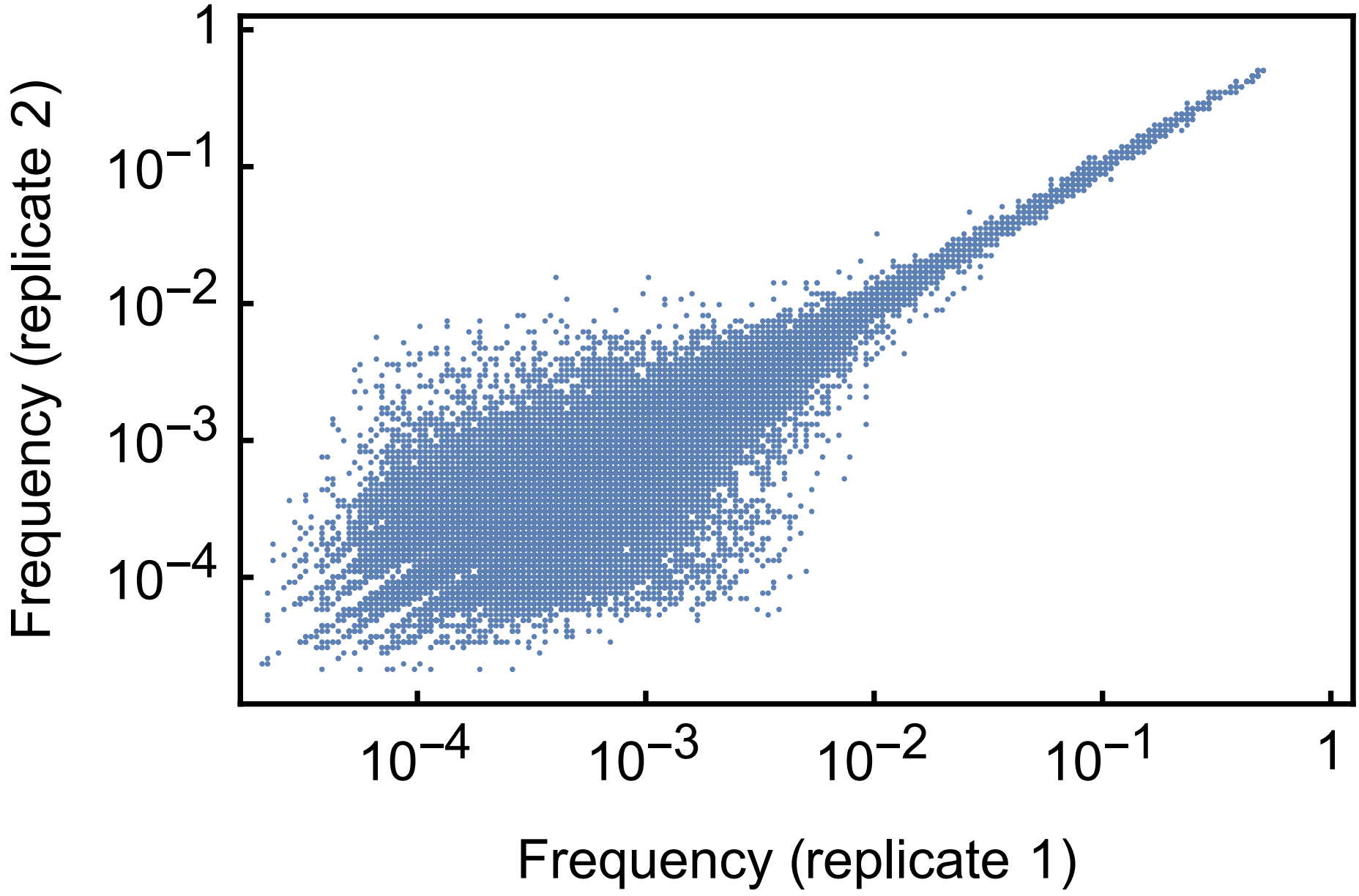

Patient X
Evolutionary Distance
Log10(Effective N)

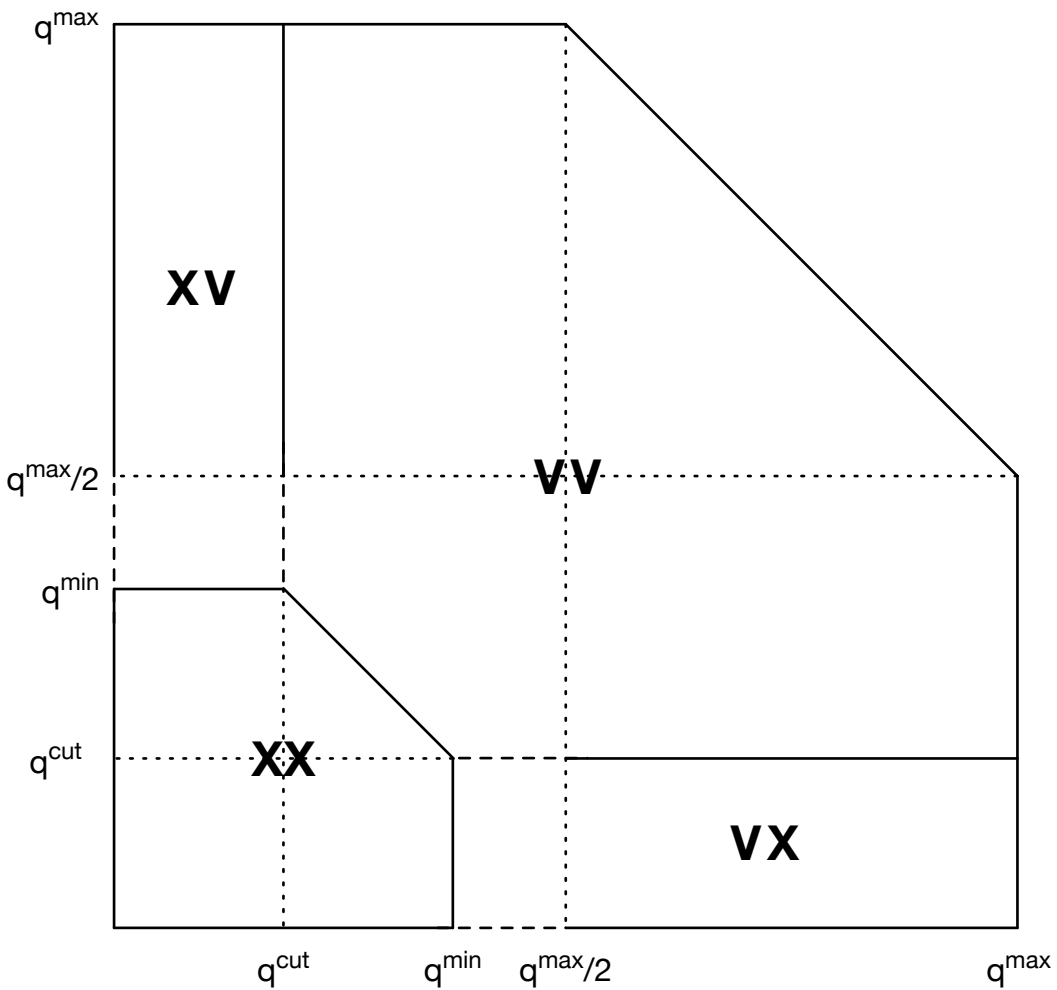

Patient Y
Sequence distance
Time difference (days)
Rate: 0.068 nucleotides/day

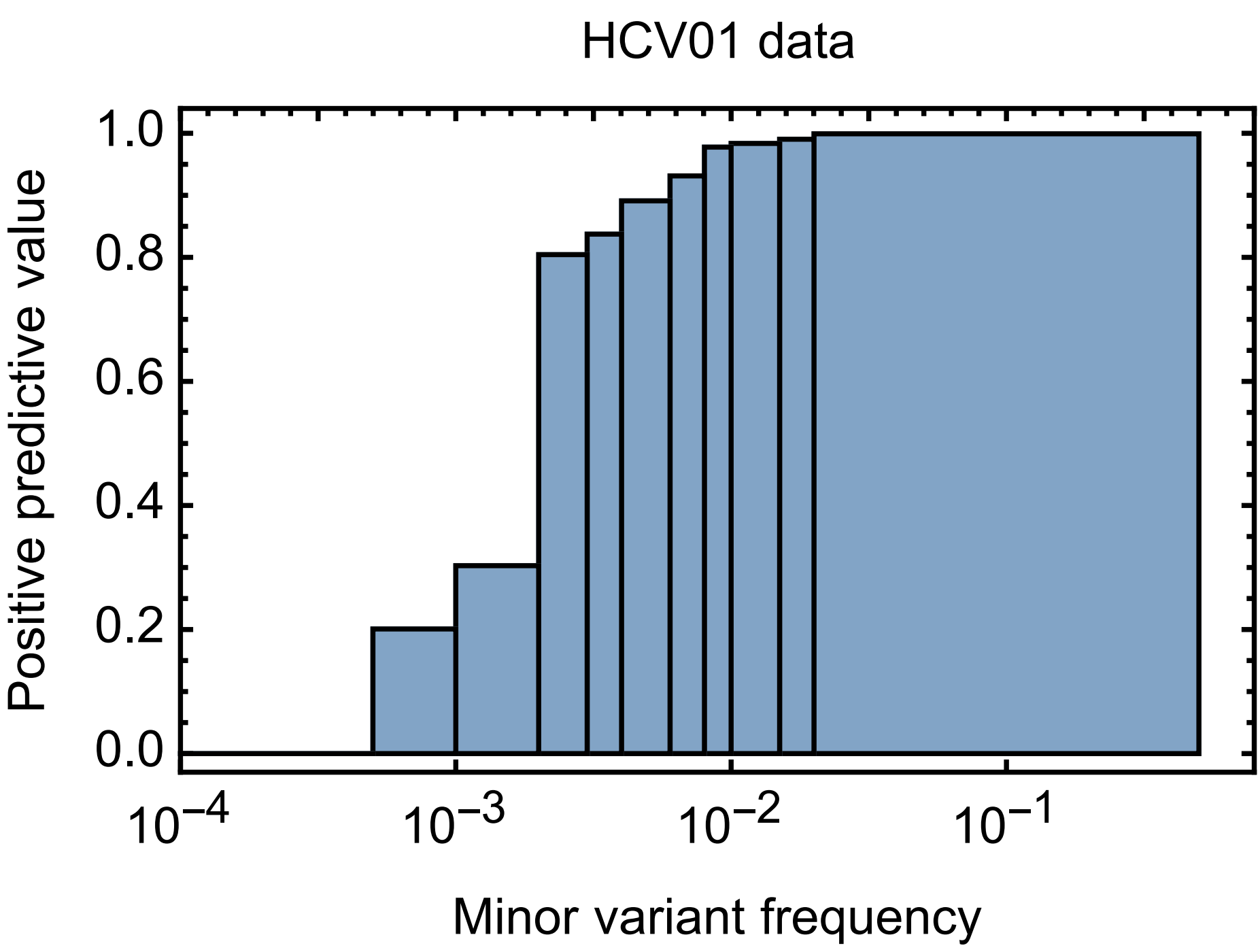

Patient Y
Evolutionary Distance
Log10(Effective N)


Patient Z
Sequence distance
Time difference (days)
Rate: 0.165 nucleotides/day


Patient Z
Evolutionary Distance
Log10(Effective N)


Frequency (replicate 2)
1
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-4}
10^{-3}
10^{-2}
10^{-1}
1
Frequency (replicate 1)


q^{max}
XV
$q^{max}/2$
VV
q^{min}
q^{cut}
XX
VX
q^{cut}
q^{min}
$q^{max}/2$
q^{max}


HCV01 data
Positive predictive value
1.0
0.8
0.6
0.4
0.2
0.0
10^{-4}
10^{-3}
10^{-2}
10^{-1}
Minor variant frequency