Examining the molecular clock hypothesis for the contemporary evolution of the rabies virus

Authors: Rowan Durrant[1]*, Christina A. Cobbold[1,2], Kirstyn Brunker[1], Kathryn Campbell[1], Jonathan Dushoff[3,4,5], Elaine A. Ferguson[1], Gurdeep Jaswant[1,6,7,8], Ahmed Lugelo[1,8,9], Kennedy Lushasi[8], Lwitiko Sikana[8], and Katie Hampson[1]

Affiliations:

1. Boyd Orr Centre for Population and Ecosystem Health, School of Biodiversity, One Health & Veterinary Medicine, College of Medical, Veterinary & Life Sciences, University of Glasgow, Glasgow, UK
2. School of Mathematics and Statistics, University of Glasgow, Glasgow, UK
3. Department of Biology, McMaster University, Hamilton, Ontario, Canada
4. Department of Mathematics and Statistics, McMaster University, Hamilton, Ontario, Canada
5. M. G. DeGroote Institute for Infectious Disease Research, McMaster University, Hamilton, Ontario, Canada
6. University of Nairobi Institute of Tropical and Infectious Diseases (UNITID), Nairobi, Kenya.
7. Tanzania Industrial Research Development Organisation (TIRDO), Dar es salaam, Tanzania
8. Environmental Health and Ecological Sciences Department, Ifakara Health Institute, Ifakara, Tanzania
9. Global Animal Health Tanzania, Arusha, Tanzania

***Corresponding author**: Rowan Durrant: r.durrant.1@research.gla.ac.uk.

Author contributions: Conceptualisation: RD, CC, KH. Formal analysis: RD. Funding acquisition: KH. Investigation: RD. Methodology: RD, CC, JD, EF, KH. Project administration: KB, KH. Resources: GJ, AL, KL, LS. Software: RD, JD, EF. Supervision: CC, KH. Visualisation: RD. Writing - original draft preparation: RD, EF. Writing - review & editing: RD, CC, KB, KC, JD, EF, KH.

Data accessibility statement: All the source code used to produce the results in this study, as well as input and output data, have been uploaded to a GitHub repository: https://github.com/RowanDurrant/Rabies-Mutation

Keywords: incubation period, generation interval, root-to-tip divergence, Lyssavirus, modelling, substitution

Abstract

The molecular clock is a method for measuring the rate of virus evolution but the key assumption that mutations accumulate on the genome at a constant rate over time does not always hold true. While modelling approaches exist to accommodate deviations from a strict molecular clock, assumptions about rate variation may not fully represent the underlying evolutionary processes and can affect the accuracy of clock calibration and divergence time estimates. There is considerable variability in rabies virus (RABV) incubation periods, ranging from days to over a year, during which viral replication may be reduced. This prompts the question of whether modelling RABV on a per infection generation basis might be more appropriate. We investigate how variable incubation periods affect root-to-tip divergence under per-unit time and per-generation models of mutation. Additionally, we assess how well these models represent root-to-tip divergence in time-stamped RABV sequences. We find that at low mutation rates (<1 substitution per genome per generation) divergence patterns between these models are difficult to distinguish, while above this threshold differences become apparent across a range of sampling rates. Using a Tanzanian dataset, we calculate the mean substitution rate to be 0.17 substitutions per genome per generation. At RABV's substitution rate, the per-generation substitution model is unlikely to represent rabies evolution substantially differently than the molecular clock model when examining contemporary outbreaks; over enough generations for any divergence to accumulate, extreme incubation periods average out. However, measuring substitution rates per-generation holds potential in applications such as inferring transmission trees and predicting lineage emergence.

Introduction

The molecular clock is a method of measuring the rate of evolution of organisms, based on the assumption that their genomes accumulate neutral mutations at a constant rate over time, either across all lineages (the "strict molecular clock") or within each individual lineage but with variation between them (the "relaxed molecular clock") (1,2). The ability to sample viral sequences through time, and the application of the molecular clock hypothesis to these sequences, has led to massive advances in using viral genetic data to investigate disease outbreaks (3). The clock rate, measured in substitutions per site per unit time, can be used to estimate how long ago pathogens diverged (4), and the date of infection of individual infected hosts (5). Combining the analysis of epidemiological and genetic data has allowed further insights into the history of outbreaks (6), and the introduction of geographic data provides predictions as to rates of spread and the frequency and source of introductions (7,8). However, in order to conduct these phylogenetic analyses, genetic divergence must measurably increase over time in the dataset under investigation (9).

The rabies virus (RABV) is a negative-strand RNA virus, with a genome size of approximately 12 kilobases. While RNA viruses generally have high mutation rates due to a lack of proofreading by RNA polymerases, RABV has a relatively low clock rate of between 1×10^{-4} and 5×10^{-4} substitutions per site per year (10–12). This may be due to strong purifying selection (10), or to an unusual feature of RABV: that infections can exhibit extended incubation periods within the host. The median generation interval (the time between one individual becoming infected and then infecting another) is estimated to be 17.3 days in domestic dogs (13), with other studies estimating mean serial intervals of 26.3 days (14) and 45.0 days (15). Symptoms, infectivity, and death from rabies, however, can occasionally occur years after the initial infection event (16). The length of this incubation period is influenced by the route of exposure, with bites to the head and neck leading to more

82 rapid disease progression than bites to lower extremities (17). RABV can remain in the
83 muscle at the bite site for significant lengths of time before invading the host's motor neurons
84 and progressing through the nervous system, with limited, if any, infection of other muscle
85 fibres (18). While some replication in the muscle cells has been observed (19), RABV
86 replication at the inoculation site is not necessary for neural invasion (20). It is currently
87 unknown precisely how the RABV replication rate in the host muscle cells and peripheral
88 nervous system compares to the massive replication rate within the cells of the central
89 nervous system and brain. However, work suggests that RABV replication in muscle cells
90 may be reduced (21), and RABV replication in cultured rat sensory neurons may be 10- to
91 100-fold lower than replication rates in rat and mouse CNS neurons (22). Rabies infections
92 that involve long incubation periods may, therefore, not lead to more accumulated mutations
93 than those with shorter incubation periods, as viral mutation is strongly influenced by the
94 replication process (24).

95 Changes in mutation rates through time due to long incubation periods may affect how we
96 analyse RABV sequence data and interpret these analyses. A relaxed molecular clock is
97 usually required to carry out phylogenetic analyses on rabies datasets, and it is not
98 uncommon for there to be difficulties in applying these analyses due to "insufficient temporal
99 signal"; usually referring to either no or a negative relationship between genetic divergence
100 and time, or this relationship having a very low R^2 (25–29). RABV clock rates can differ
101 significantly between hosts (30), geographic locations (31) and circulating lineages (12),
102 which may be driven in part by differences in incubation periods. If the variable incubation
103 period of rabies infections does cause deviation from the molecular clock model (exceeding
104 the variation captured by relaxed or modified clock models), this may negatively affect the
105 accuracy of time-scaled phylogenetic trees and emergence date predictions. Conversely, if
106 mutation does continue at a consistent rate during the incubation period, attention should be

paid to extremely long incubators which could drive the emergence of new variants, as seen recently in chronic SARS-CoV-2 infections (32,33).

We hypothesised that reduced replication (and thus mutation) during the incubation period could cause rabies evolution to be better represented by a per-generation model of mutation than by the molecular clock model. We aim to clarify the nature of contemporary RABV evolution using in silico methods, comparing the root-to-tip divergence of sequences generated from synthetic outbreaks under per-unit time or per-generation mutation models, and comparing these to RABV genomic data from Tanzania. We also aim to calculate a per-generation substitution rate for RABV for future use as a parameter in transmission tree reconstruction algorithms.

Methods

We investigate two contrasting mutational models for RABV – i.e., substitutions occurring on a per-generation vs. per-unit-time basis – using a simulation approach. We first generated synthetic RABV outbreaks using a branching process model (13) and then simulated these two mutation processes over the resulting transmission trees. From the resulting synthetic sequences, we examined root-to-tip divergence and calculated variance explained (R^2) from linear regressions, and compared these to the root-to-tip divergence of a set of RABV whole genome sequences from Tanzania. Finally, we developed a method to estimate the per-generation substitution rate for RABV and tested this on synthetic data before applying it to the real RABV dataset.

Rabies outbreak simulation

We simulated RABV mutation on branching-process simulations of rabies outbreaks. Outbreaks were simulated 100 times over a spatially explicit representation of Mara Region in northern Tanzania. In Serengeti District, where contact tracing data were available, the model was initialised with the three cases that occurred in the mean generation interval (g=27 days, based on contact tracing data) prior to 2017 (simulations were run over a dog population representing that in Mara region between 2017 and 2024). In the rest of Mara region, where there were no data to guide initialisation, we seeded with $(0.01Dg)/365$ cases, where D is the initial dog population in that area. If R_e=1 (endemic transmission), this results in roughly 1% of the population becoming rabid over a year; contact tracing data suggest that incidence typically does not exceed that level (34). This led to a total of 273 initial cases in the region. Each case was assigned a number of offspring cases drawn from a negative binomial distribution (34) with mean (R_0)=1.05 and dispersion parameter=1.33. The R_0 value was chosen to result in a median number of cases each month that was roughly constant over time (over the 100 simulations), mimicking endemic disease. Movement of rabid dogs from their home locations to and between transmission locations followed a random walk with step lengths drawn from a Weibull distribution (shape=0.41, scale=0.13). We simulated occasional long-distance transport of dogs to a random location prior to their first transmission in 2% of cases (13). At each of a rabid dog's transmission locations, another dog was randomly selected within the local 1km2 grid cell. If this dog was susceptible (i.e., not vaccinated or already incubating infection from a prior transmission event), rabies was transmitted. A generation interval was drawn for each new infection from a lognormal distribution (meanlog=2.96, sdlog=0.82), describing the time delay before it also became rabid and made its assigned transmissions. The step-length and generation-interval distributions were fitted using contact tracing data from Serengeti District, Tanzania (13). Branching process simulations were continued until 7 years had passed or rabies went extinct. Each synthetic

case was assigned an individual ID, and for every case (except initial seed cases) we recorded the ID of the associated progenitor case. Dates of infection and transmission were recorded for each case.

We isolated complete transmission trees descending from each of the 273 initial cases from within one randomly selected synthetic outbreak. Transmission trees that contained over 100 cases (9 out of 273 trees in total, that ranged in size from 533 - 19,382 cases) were then used to generate synthetic sequence data. For each of the 9 trees the index case was assigned an initial 12kb genome sequence. Under the per-unit time mutation model, we determined the expected number of mutations by multiplying the clock rate, the genome length and the length of the generation interval, for each case along the resulting transmission tree. The realised number of mutations was then chosen from a Poisson distribution, with this mean. We then randomly chose positions to change and new nucleotides to change them to. The resulting synthetic sequence data is referred to as the "time-based sequence data". The generation-based model of mutation works as above, with the exception that the expected number of substitutions in a generation is constant and produces the synthetic "generation-based sequence data".

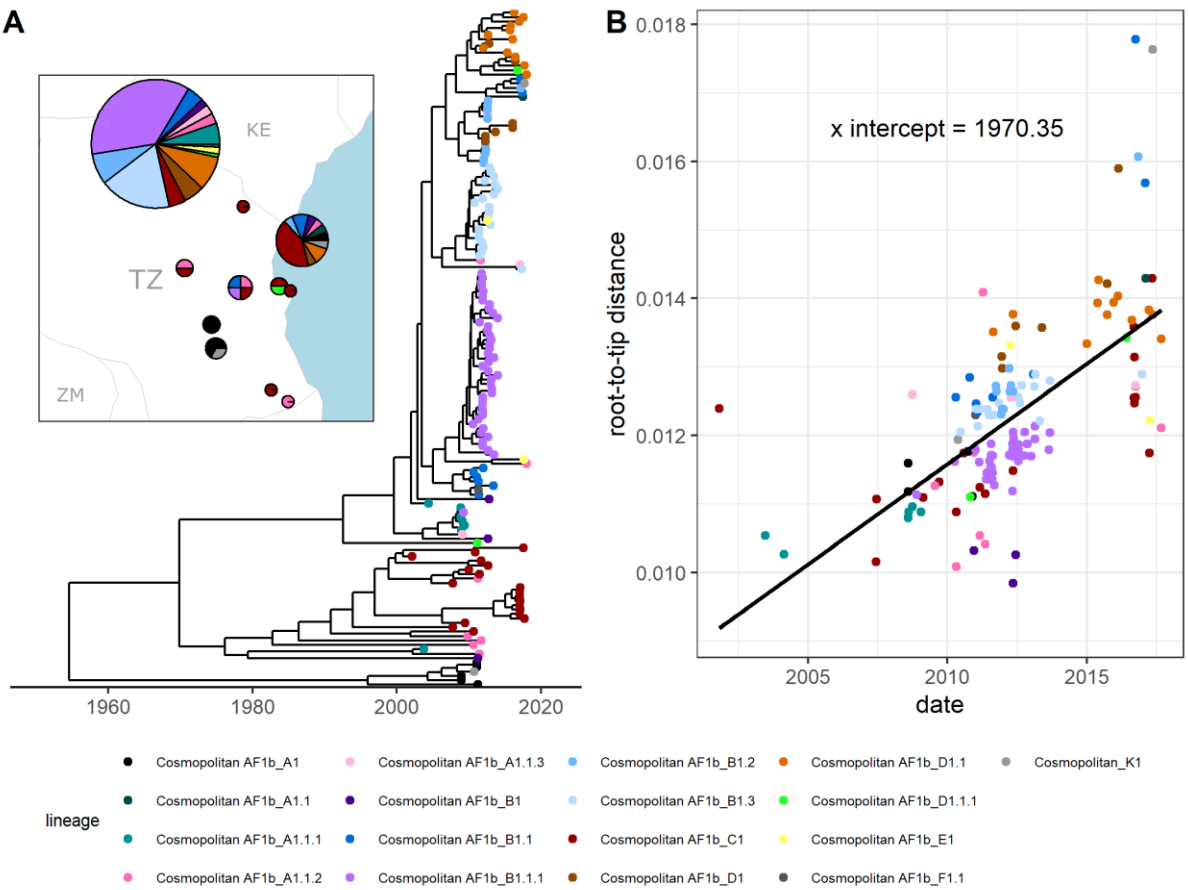
Divergence rate analysis

To investigate patterns of temporal divergence under the mutation models described above, we generated synthetic data with values of substitution rates ranging from 0.05 to 3 substitutions per genome per generation (or the per unit time substitution rate equivalent) and 4 population sampling regimes (from 1% of cases to 20%, informed by a previous study that estimated that routine surveillance for rabies rarely confirms more than 10% of circulating cases (35)). We calculated the genetic divergence as the number of nucleotide differences

179 from the index case to each sampled case. For each of the nine transmission trees, we then
180 compared genetic divergence with time under each scenario (substitution rate and sampling
181 regime combination), using linear regression through the origin.

182 In order to compare our synthetic patterns of divergence over time to real rabies data, a root-
183 to-tip divergence plot was also generated for a dataset of real RABV sequences (data from
184 (36); Figure 1A) using TempEst (v1.5.3 (37)), with the best-fit root located (Figure 1B).
185 These rabies cases occurred between 2001 and 2017 and were primarily from the Serengeti
186 district and Pemba Island, with the remaining sequences from elsewhere in Tanzania (Figure
187 1A inset). Sequence acquisition and tree building methods are detailed in (36).

188



189

Fig 1. Phylogeny of real rabies virus whole genome sequences from Tanzania and root-to-tip divergence. (A) The time-scaled tree (36) used to generate the root-to-tip divergence plot and to calculate the per-generation substitution rate. The inset map shows the approximate locations that the samples were collected from, and the lineages present in each location. Map point size represents the number of sequences in this dataset from district centroid locations. **(B)** The corresponding root-to-tip divergence plot. Point colours represent RABV lineage.

Calculating the per-generation substitution rate

We developed a method of calculating the per-generation substitution rate using the clock rate, the genome length, and the mean generation interval. We assessed this method's accuracy using the synthetic outbreak sequence data, before applying it to the aforementioned set of RABV whole genome sequences.

To estimate the mean per-generation substitution rate, we analysed sequence data with BEAST, and multiplied clock rate values from the BEAST posterior by the generation-interval lengths sampled from the posterior of a simple Bayesian analysis and then multiplied again by the genome length. The mean and 95% confidence interval of the prediction of the per-generation substitution rate for the RABV dataset was calculated by taking the mean and the 2.5% and 97.5% percentiles of the resulting multiplied posteriors.

To evaluate the accuracy of this method in predicting the mean per-generation substitution rate, we also applied it to synthetic sequence data generated from outbreaks using the per-generation mutation model as described above, under different substitution rates (11 values ranging from 0.05 substitutions per generation to 3 substitutions per generation) and case sampling rates (1%, 5%, 10% or 20% of cases sampled) across the 9 transmission trees that

contained at least 100 cases. A subsample of synthetic sequences was selected and saved to a FASTA file. BEAST log files were generated from these sequences using BEASTGen version 1.0.2 and BEAST version 1.10.4 (38). We chose to use a JC substitution model with a strict clock, no site heterogeneity due to our per-generation mutation model used in the simulations having equal probability of any site or base being chosen and assumed constant population size. We used a chain length of 10 million and tracelog frequency of 1000 and used a 25% burn-in period. We applied the substitution rate calculation method to these phylogenetic trees, and assessed the accuracy of the resulting mean per-generation substitution rates by comparing them to the parameter values used to generate the synthetic sequences, using the natural log of the ratio:

$$Deviation = \ln\left(\frac{M_p}{M_a}\right)$$

where M_p is the mean predicted per-generation substitution rate and M_a is the actual substitution rate, where a deviation of zero means perfect accuracy.

The same method was applied to the dataset of 153 RABV sequences sampled from across Tanzania (data from (36); Figure 1A). The mean per-generation substitution rate was calculated, and distributions were fitted from the multiplied generation interval and clock rate posteriors (generation interval posteriors based on values from (13) for the Tanzanian dataset, extracted directly from the lognormal distribution used in simulations, and clock rate posteriors taken from the BEAST log file of the time-scaled tree from Lushasi et al. (36)) and genome length as described above. We compared different distributions (Gamma and Lognormal) for predicting substitution rates and selected the best fitting distribution by AIC. We also calculated the probabilities of between 0 and 10 SNP differences occurring across 1, 5 or 10 infection generations. For this calculation we simulated mutations arising at a Poisson

238 rate with lambda drawn from the fitted substitution rate distribution. The means and 95%
239 confidence intervals were calculated from the 10,000 simulations.

240

241 Data and code availability

242 All code is available at https://github.com/RowanDurrant/Rabies-Mutation. Analyses were
243 conducted using the R programming language (39). The beta regression curve and prediction
244 interval in Figure 2C was generated using the 'betareg' R package (40). RABV lineages were
245 assigned using MADDOG (41).

246

247 **Results**

248 Root-to-tip divergence analysis

249 At higher per-generation substitution rates (1 substitution per genome per generation and
250 above), distinct differences can be seen between root-to-tip divergence plots from the two
251 models of mutation (Figure 2A). The synthetic data generated from the per-generation
252 mutation model shows "stray" clusters or ridges of points both above and below the main
253 funnel of points, illustrated in the example shown in Figure 2A. Divergence plots from
254 synthetic data generated from the time-based model of mutation have less variance and do not
255 exhibit this pattern. At lower substitution rates (below 1 substitution per generation), no such
256 pattern is clearly distinguishable (Figure 2B). When the cases represented by the high-
257 divergence points from the per-generation model in Fig. 2A are visualised in a transmission
258 tree, they are mainly confined to a single chain (Supplementary Figure S1).

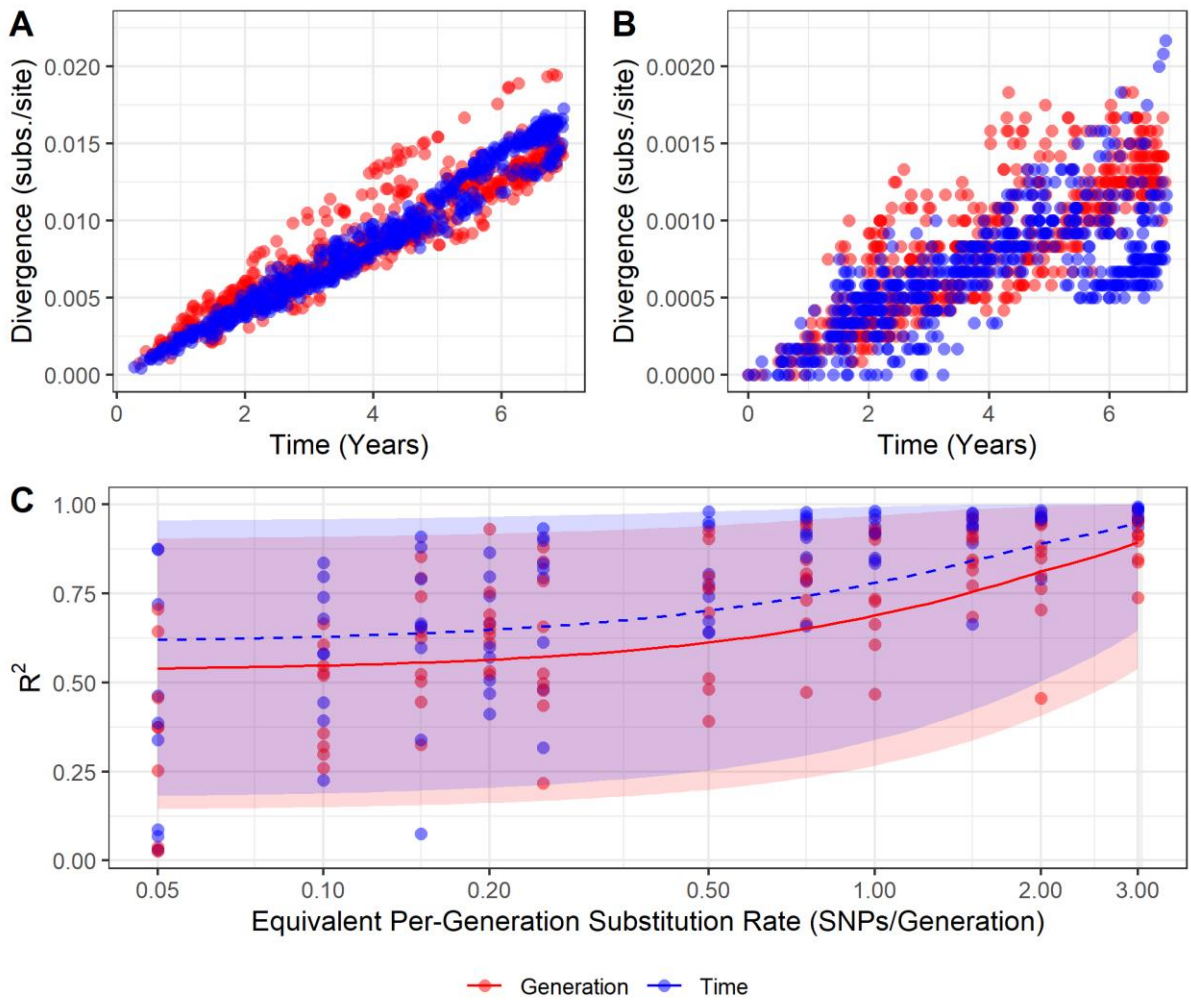


Fig. 2: Temporal genetic divergence varies under two models of mutation. (A) Root-to-tip divergence plots for synthetic sequences produced using time-based and generation-based mutation models, equivalent to 2 substitutions per genome per generation and **(B)** equivalent to 0.2 substitutions per genome per generation. Note that the y-axis scales differ by an order of magnitude between A and B. These data are from running mutation models over the same single transmission tree and have a case sampling rate of 5% (i.e., 621 cases sampled of 12,434 total). **(C)** The R^2 values obtained from regression through the origin of root-to-tip divergence of synthetic data from the time-based and generation-based models. Point colour indicates the mutation model used to generate the data. Lines represent beta regressions with logit links fit to data points, and shading represents the 95% prediction interval. The X axis is

log scaled. 5% of cases were sampled here; sampling rate had little effect on R2 (Supplementary Figure S2).

Root-to-tip divergence plots derived from synthetic transmission trees using the time-based mutation model had, on average, higher R^2 values than those from synthetic transmission trees using the per-generation mutation model, although this is more difficult to distinguish below a substitution rate of 0.5 substitutions per genome per generation (Figure 2C). As the substitution rate increases, the R^2 values across both mutation models increase. The case sampling rate appears to have little effect on R^2 (Supplementary Figure S2).

The root-to-tip divergence plot of the Tanzanian RABV dataset more closely resembles those of lower substitution rate simulations, where it is difficult to determine any difference between the models of mutation (Figure 1). While most lineages surround the regression line, some (for example Cosmopolitan AF1b_B1) group below the line, but without forming a distinguishable “ridge”.

<u>Substitution rate calculation</u>

The accuracy of our method used to calculate per-generation substitution rate remains similar at all but the lowest values of substitution rate (Figure 3), with a tendency to underestimate the substitution rate (meaning that the predicted substitution rate is below the substitution rate parameter used to generate the synthetic data; mean natural log of the ratio of -0.13 and root-mean-square of 0.54, where values of 0 indicate perfect prediction). Accuracy appears to be more influenced by the number of sequences used in the BEAST analysis than by the case sampling rate itself; the mean natural log of the ratio falls to -0.32 when fewer than 50 sequences are used (root-square-mean of 0.91). The reverse is also observed, as when more

than 2000 sequences are used the mean natural log of the ratio rises to an overestimation of 0.38 (root-mean-square of 0.85). This is likely due to the chain length of 10,000,000 being insufficient for such a large number of sequences, and so should be taken into consideration if using this method on very large datasets.

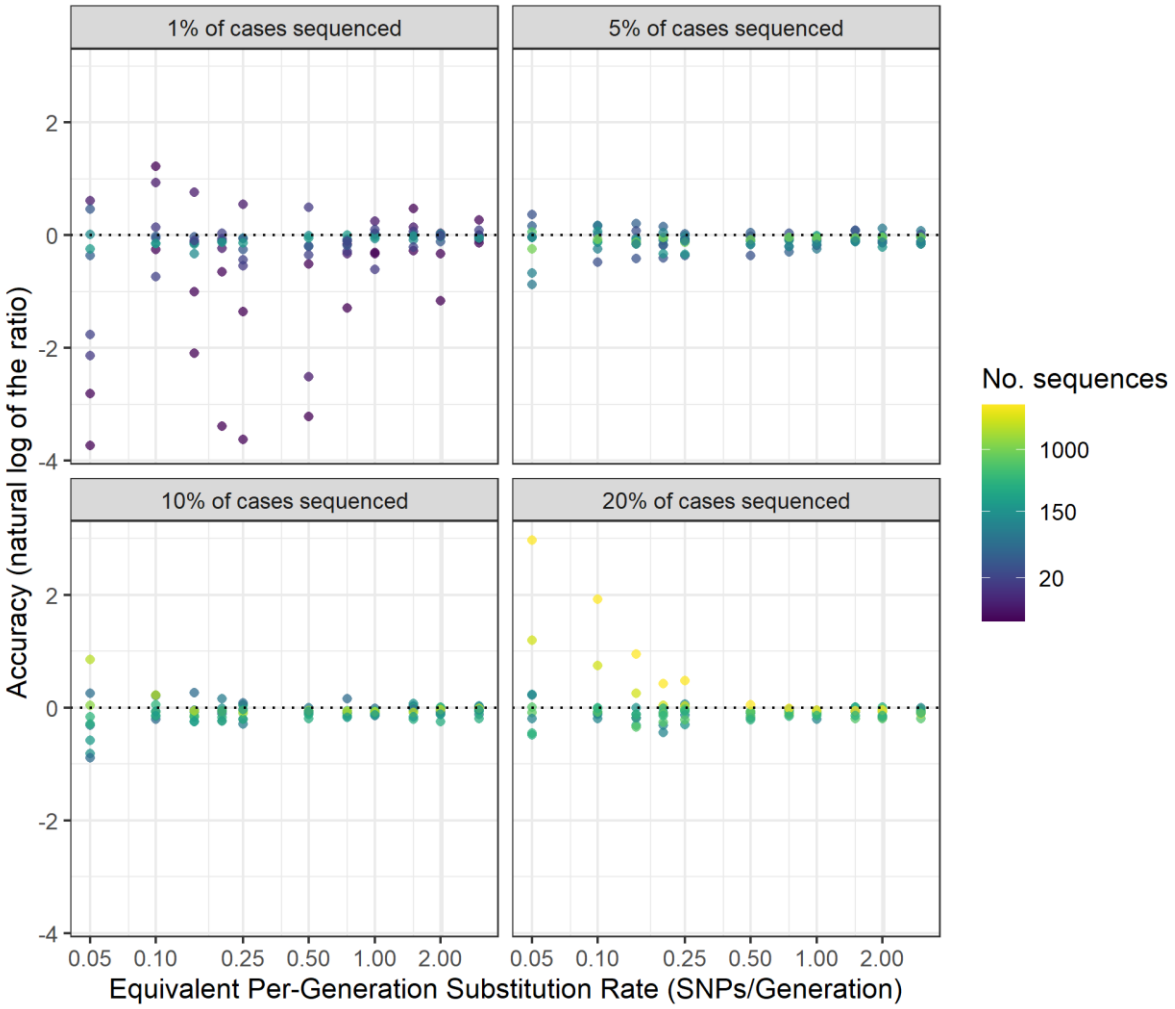


Fig. 3: Accuracy of per-generation substitution rate predictions for different numbers of sequences, substitution rates and sampling rates. Facets indicate case sampling rate. The dotted line represents perfect accuracy. X axis and colour scale are log transformed.

The Tanzanian RABV dataset from which we estimated the per-generation substitution rate contains 153 sequences in total, and the accompanying time-scaled phylogenetic tree has a root-to-tip height of approximately 65 years, although the sequences spanned just 16 years as they were sampled from 2001 to 2017 (with 46.7% from years 2011-2012). These sequences were largely complete; 98% of sequences were >95% complete (>11,327 kb in length). The mean per generation substitution rate of RABV in this dataset was estimated to be 0.171 (95% confidence interval: 0.127 - 0.219). The best fitting distribution by AIC to the output of the multiplied Bayesian posteriors was a Gamma distribution with the parameters shape = 51.69 and rate = 301.8 (table 1).

Table 1: Parameters and distributions used to estimate the per generation substitution rate. The distributions fitted to the multiplied Bayesian output are shown, along with their parameters and AIC.

Distribution	**Parameters**	**Δ AIC**
Gamma	shape = 51.69, rate = 301.8	0
Lognormal	meanlog = -1.774, sdlog = 0.1402	1277.4

Using the calculated per generation per genome substitution rates, we calculated the probability of different numbers of substitutions occurring over 1, 5 and 10 generations, drawing the per-generation substitution rate (λ) from the above distribution (Figure 4). Over many generations it is still quite likely for zero substitutions to occur; after 10 generations, the probability of zero substitutions having occurred is 0.19.

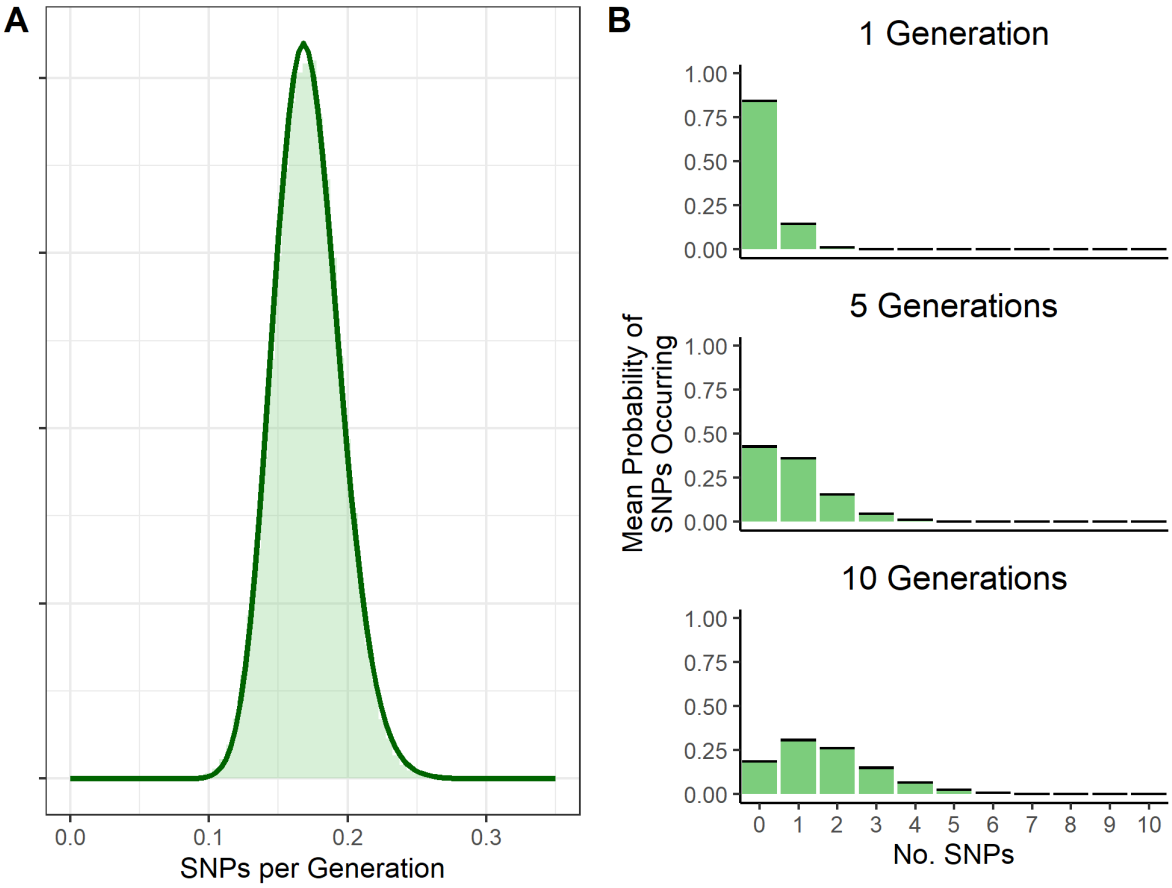


Fig. 4: Probability distributions of the mean per-generation substitution rate and substitutions occurring over generations. **(A)** predicted probability distribution of the per genome per generation substitution rate from Tanzanian RABV sequences, with underlying histogram of multiplied Bayesian posteriors of clock rate and generation interval. **(B)** probability distribution of SNPs occurring over 1, 5 and 10 generations. The λ value for a Poisson rate of SNP occurrence is drawn from the SNPs per generation distribution fitted in Figure 4A. Black bars represent the 95% confidence intervals (which are very tight).

Discussion

It can be difficult to get sufficient temporal signal for RABV sequence datasets, potentially due to its variable incubation periods. We hypothesised that a per-generation model of

335 mutation may be more representative of RABV evolution than a purely time-based model.
336 We found that substantial differences in root-to-tip divergence patterns between synthetic
337 outbreaks using generation-based and time-based models of mutation could be observed only
338 at high underlying substitution rates. The substitution rate for the Tanzanian RABV
339 sequences examined (~0.17 substitutions per genome per generation) was in the range where
340 divergence patterns in the two models were extremely similar. We can thus assume that the
341 two models will give extremely similar results on the relevant time scale. As we observed
342 increasing divergence over time with reasonable R^2 values within this substitution rate range,
343 it implies that variable incubation periods alone may not fully account for the challenge in
344 obtaining temporal signal. Other factors such as sampling biases (i.e., insufficient sequence
345 diversity captured over a short timescale) may be responsible. This is an important
346 consideration for analysing RABV sequences from new outbreaks, or from endemic areas
347 where sampling is opportunistic.

348 The observation of little difference between root-to-tip divergence plots derived from the two
349 mutation models at substitution rates below 1 substitution per genome per generation is likely
350 because of averaging; multiple generations of infection are expected to have occurred per
351 substitution that arises on the viral genome. Over the many generations needed before
352 significant levels of viral genetic diversity are reached, the influence of any unusually long
353 incubation periods will be damped by the opposite influence of unusually short incubation
354 periods, eventually becoming indistinguishable from clock-like behaviour. On the other hand,
355 at higher substitution rates ridges form on the root-to-tip divergence plots under the per-
356 generation model of mutation but not under the per unit time model. While not affecting the
357 overall clock rate, these ridges reduce the overall R^2, and may be better analysed using a
358 separate local clock (42). The cases in these ridges almost all descend from a common
359 ancestor (Supplementary Figure S1), suggesting that a single unusually long or short

360 incubation period can affect which phylogenetic analyses we perform. Studies examining the
361 number of substitutions occurring between successive sequenced cases, and whether this
362 increases when the secondary case's incubation period is unusually long, could clarify the
363 exact relationship between substitutions, generations, and time. More detailed data will be
364 required to investigate this further.

365 We calculated RABV's mean per-generation substitution rate to be approximately 0.17
366 substitutions per genome per transmission generation. This estimate is lower than predictions
367 for other RNA viruses, such as SARS (2 substitutions per genome per human passage (43)),
368 SARS-CoV-2 (0.52 substitutions per genome per 5.8-day generation interval (44)) and Ebola
369 virus (0.875 substitutions per genome per 14-day generation interval (45)). RNA viruses that
370 undergo periods of reduced replication or complete latency often show reduced substitution
371 rates, with one extreme example being HTLV-1/2 (46,47). However, we would not expect this
372 to affect the per-generation rate. The low per-generation substitution rate seen in rabies is
373 therefore likely due to mutation being constrained by other factors, such as strong purifying
374 selection (10).

375 We can predict from the estimated per-generation substitution rate that identical sequences
376 are likely to have less than 5 intermediate generations between them (probability of fewer
377 than five generations occurring before a mutation occurs > 0.49 by repeated sampling of a
378 Poisson distribution with a lambda of 0.17), but still have a non-negligible probability of
379 being more distantly related. While the low substitution rate means that comparing the
380 number of SNPs between sequences alone may not be an effective method of determining
381 infector-infectee relationships, it could be used in conjunction with temporal and location
382 data to make more accurate predictions of transmission events by ruling out relationships
383 between more distantly related transmission chains co-circulating in the same area, as in (48).
384 Notably, our Poisson distribution of the number of substitutions occurring in one generation

is visually very similar to the genetic signature distribution reported in Cori *et al*. ((48), Fig S1), despite different methods and RABV datasets being used in their calculations. The mean per-generation substitution rate produced through the methods described could also potentially be used in the future to estimate the size of rabies outbreaks from genetic data and calculate predicted rates of RABV lineage emergence based on the classification guidelines of MADDOG (41).

While the Jukes-Cantor model was the most appropriate to use on our synthetic data due to the simplicity of the mutation models, phylogenetic analyses on real RABV genome data usually use a more complex model, such as the GTR + G substitution model used to generate the Tanzanian tree shown in this study (36). This, along with the simplicity of our mutation model as well as sampling biases in the real dataset, may affect how comparable synthetic root-to-tip divergence plots are to the real data. We also observed that when a very large number of sequences (> 2000) were used to predict the per-generation substitution rate, the prediction accuracy decreased. Upon inspecting the BEAST log files using Tracer (49), it became apparent that our one-size-fits-all approach of using a BEAST chain length of 10,000,000 was insufficient for such large amounts of data. While it is unlikely that this number of sequences will be used for this purpose in the future, the prediction accuracy is likely to be increased simply by using an appropriate chain length.

While the molecular clock has proven critical for gaining insights into the history and dynamics of disease outbreaks, the epidemiological characteristics of a virus should be considered when choosing how to measure viral evolution. In this study, we determine that the per-generation model is not likely to produce substantially different results from the molecular clock model when analysing contemporary RABV evolution. We also estimate the

mean per-generation substitution rate of RABV for future use in transmission tree reconstruction and efforts to estimate outbreak sizes and lineage emergence rates. Given that many different lineages circulating simultaneously is seemingly a common occurrence in areas with endemic rabies, it is important to investigate whether these lineages vary in evolutionary rate and generation interval length, and ascertain the potential effects on phylogenetic analyses.

Acknowledgments

This work was supported by the EPSRC DTP (RD) [EP/T517896/1], Institutional Strategic Support Fund grants at the University of Glasgow (KB) [204820], a Genomics and Modelling to the Control of Virus Pathogens (GeMVi) fellowship funded by National Institute for Health Research (NIHR) (GJ) [176382], Wellcome (KH) (207569/Z/17/Z, 224520/Z/21/Z), and the MRC (KB) [MR/X002047/1].

References

1. Gojobori T, Moriyama EN, Kimura M. Molecular clock of viral evolution, and the neutral theory. Proc Natl Acad Sci. 1990;87(24):10015–8.

2. Drummond AJ, Ho SYW, Phillips MJ, Rambaut A. Relaxed Phylogenetics and Dating with Confidence. PLOS Biol. 2006;4(5):e88.

3. Drummond A, Oliver G. P, Rambaut A. Inference of Viral Evolutionary Rates from Molecular Sequences. Adv Parasitol. 2003;54:331–58.

4. Pybus OG, Rambaut A. Evolutionary analysis of the dynamics of viral infectious disease. Nat Rev Genet. 2009;10(8):540–50.

5. Wróbel B, Torres-Puente M, Jiménez N, Bracho MA, García-Robles I, Moya A, et al. Analysis of the Overdispersed Clock in the Short-Term Evolution of Hepatitis C Virus: Using the E1/E2 Gene Sequences to Infer Infection Dates in a Single Source Outbreak. Mol Biol Evol. 2006;23(6):1242–53.

6. Grenfell BT, Pybus OG, Gog JR, Wood JLN, Daly JM, Mumford JA, et al. Unifying the Epidemiological and Evolutionary Dynamics of Pathogens. Science. 2004;303(5656):327–32.

7. Gire SK, Goba A, Andersen KG, Sealfon RSG, Park DJ, Kanneh L, et al. Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. Science. 2014;345(6202):1369–72.

8. Kamath PL, Foster JT, Drees KP, Luikart G, Quance C, Anderson NJ, et al. Genomics reveals historic and contemporary transmission dynamics of a bacterial disease among wildlife and livestock. Nat Commun. 2016;7(1):11448.

9. Drummond AJ, Pybus OG, Rambaut A, Forsberg R, Rodrigo AG. Measurably evolving populations. Trends Ecol Evol. 2003;18(9):481–8.

10. Holmes EC, Woelk CH, Kassis R, Bourhy H. Genetic Constraints and the Adaptive Evolution of Rabies Virus in Nature. Virology. 2002;292(2):247–57.

11. Biek R, Pybus OG, Lloyd-Smith JO, Didelot X. Measurably evolving pathogens in the genomic era. Trends Ecol Evol. 2015;30(6):306–13.

12. Layan M, Dellicour S, Baele G, Cauchemez S, Bourhy H. Mathematical modelling and phylodynamics for the study of dog rabies dynamics and control: A scoping review. PLoS Negl Trop Dis. 2021;15(5):e0009449.

13. Mancy R, Rajeev M, Lugelo A, Brunker K, Cleaveland S, Ferguson EA, et al. Rabies shows how scale of transmission can enable acute infections to persist at low prevalence. Science. 2022;376(6592):512–6.

14. Hayes S, Lushasi K, Sambo M, Changalucha J, Ferguson EA, Sikana L, et al. Understanding the incidence and timing of rabies cases in domestic animals and wildlife in south-east Tanzania in the presence of widespread domestic dog vaccination campaigns. Vet Res. 2022;53(1):106.

15. Kurosawa A, Tojinbara K, Kadowaki H, Hampson K, Yamada A, Makita K. The rise and fall of rabies in Japan: A quantitative history of rabies epidemics in Osaka Prefecture, 1914–1933. PLoS Negl Trop Dis. 2017;11(3):e0005435.

16. Boland TA, McGuone D, Jindal J, Rocha M, Cumming M, Rupprecht CE, et al. Phylogenetic and epidemiologic evidence of multiyear incubation in human rabies. Ann Neurol. 2014;75(1):155–60.

17. Dimaano EM, Scholand SJ, Alera MTP, Belandres DB. Clinical and epidemiological features of human rabies cases in the Philippines: a review from 1987 to 2006. Int J Infect Dis. 2011;15(7):e495–9.

18. Charlton KM, Nadin-Davis S, Casey GA, Wandeler AI. The long incubation period in rabies: delayed progression of infection in muscle at the site of exposure. Acta Neuropathol (Berl). 1997;94(1):73–7.

19. Yamaoka S, Ito N, Ohka S, Kaneda S, Nakamura H, Agari T, et al. Involvement of the Rabies Virus Phosphoprotein Gene in Neuroinvasiveness. J Virol. 2013;87(22):12327–38.

20. Shankar V, Dietzschold B, Koprowski H. Direct entry of rabies virus into the central nervous system without prior local replication. J Virol. 1991;65(5):2736–8.

21. Schnell MJ, McGettigan JP, Wirblich C, Papaneri A. The cell biology of rabies virus: using stealth to reach the brain. Nat Rev Microbiol. 2010;8(1):51–61.

22. Lycke E, Tsiang H. Rabies virus infection of cultured rat sensory neurons. J Virol. 1987;61(9):2733–41.

23. Faber M, Pulmanausahakul R, Hodawadekar SS, Spitsin S, McGettigan JP, Schnell MJ, et al. Overexpression of the Rabies Virus Glycoprotein Results in Enhancement of Apoptosis and Antiviral Immune Response. J Virol. 2002;76(7):3374–81.

24. Belshaw R, Gardner A, Rambaut A, Pybus OG. Pacing a small cage: mutation and RNA viruses. Trends Ecol Evol. 2008;23(4):188–93.

25. Fusaro A, Monne I, Salomoni A, Angot A, Trolese M, Ferrè N, et al. The introduction of fox rabies into Italy (2008–2011) was due to two viral genetic groups with distinct phylogeographic patterns. Infect Genet Evol. 2013;17:202–9.

26. Wang L, Wu X, Bao J, Song C, Du J. Phylodynamic and transmission pattern of rabies virus in China and its neighboring countries. Arch Virol. 2019;164(8):2119–29.

27. Zhang Y, Vrancken B, Feng Y, Dellicour S, Yang Q, Yang W, et al. Cross-border spread, lineage displacement and evolutionary rate estimation of rabies virus in Yunnan Province, China. Virol J. 2017;14(1):102.

28. Faye M, Faye O, Paola ND, Ndione MHD, Diagne MM, Diagne CT, et al. Rabies surveillance in Senegal 2001 to 2015 uncovers first infection of a honey-badger. Transbound Emerg Dis. 2022;69(5):e1350–64.

29. Caraballo DA, Lema C, Novaro L, Gury-Dohmen F, Russo S, Beltrán FJ, et al. A Novel Terrestrial Rabies Virus Lineage Occurring in South America: Origin, Diversification, and Evidence of Contact between Wild and Domestic Cycles. Viruses. 2021;13(12):2484.

499 30. Troupin C, Dacheux L, Tanguy M, Sabeta C, Blanc H, Bouchier C, et al. Large-Scale
500 Phylogenomic Analysis Reveals the Complex Evolutionary History of Rabies Virus in
501 Multiple Carnivore Hosts. PLOS Pathog. 2016;12(12):e1006041.

502 31. Streicker DG, Lemey P, Velasco-Villa A, Rupprecht CE. Rates of Viral Evolution Are
503 Linked to Host Geography in Bat Rabies. PLOS Pathog. 2012;8(5):e1002720.

504 32. Kemp SA, Collier DA, Datir RP, Ferreira IATM, Gayed S, Jahun A, et al. SARS-CoV-2
505 evolution during treatment of chronic infection. Nature. 2021;592(7853):277–82.

506 33. Choi B, Choudhary MC, Regan J, Sparks JA, Padera RF, Qiu X, et al. Persistence and
507 Evolution of SARS-CoV-2 in an Immunocompromised Host. N Engl J Med.
508 2020;383(23):2291–3.

509 34. Hampson K, Dushoff J, Cleaveland S, Haydon DT, Kaare M, Packer C, et al.
510 Transmission Dynamics and Prospects for the Elimination of Canine Rabies. PLOS Biol.
511 2009;7(3):e1000053.

512 35. Townsend SE, Lembo T, Cleaveland S, Meslin FX, Miranda ME, Putra AAG, et al.
513 Surveillance guidelines for disease elimination: A case study of canine rabies. Comp
514 Immunol Microbiol Infect Dis. 2013;36(3):249–61.

515 36. Lushasi K, Brunker K, Rajeev M, Ferguson EA, Jaswant G, Baker LL, et al. Integrating
516 contact tracing and whole-genome sequencing to track the elimination of dog-mediated
517 rabies: an observational and genomic study. eLife. 2023;12:e85262.

518 37. Rambaut A, Lam TT, Max Carvalho L, Pybus OG. Exploring the temporal structure of
519 heterochronous sequences using TempEst (formerly Path-O-Gen). Virus Evol.
520 2016;2(1):vew007.

521 38. Suchard MA, Lemey P, Baele G, Ayres DL, Drummond AJ, Rambaut A. Bayesian
522 phylogenetic and phylodynamic data integration using BEAST 1.10. Virus Evol.
523 2018;4(1):vey016.

524 39. R Core Team. R: A Language and Environment for Statistical Computing [Internet].
525 Vienna, Austria; 2020. Available from: https://www.R-project.org/

526 40. Cribari-Neto F, Zeileis A. Beta Regression in R. J Stat Softw. 2010;34(1):1–24.

527 41. Campbell K, Gifford RJ, Singer J, Hill V, O'Toole A, Rambaut A, et al. Making genomic
528 surveillance deliver: A lineage classification and nomenclature system to inform rabies
529 elimination. PLOS Pathog. 2022;18(5):e1010023.

530 42. Featherstone LA, Rambaut A, Duchene S, Wirth W. Clockor2: Inferring global and local
531 strict molecular clocks using root-to-tip regression. BioRxiv. 2023;548947.

532 43. Vega VB, Ruan Y, Liu J, Lee WH, Wei CL, Se-Thoe SY, et al. Mutational dynamics of the
533 SARS coronavirus in cell culture and human populations isolated in 2003. BMC Infect Dis.
534 2004;4:32.

535 44. Braun K, Moreno G, Wagner C, Accola MA, Rehrauer WM, Baker D, et al. Limited
536 within-host diversity and tight transmission bottlenecks limit SARS-CoV-2 evolution in
537 acutely infected individuals. BioRxiv. 2021;440988.

538 45. Kinganda-Lusamaki E, Black A, Mukadi D, Hadfield J, Mbala-Kingebeni P, Pratt CB, et
539 al. Operationalizing genomic epidemiology during the Nord-Kivu Ebola outbreak,
540 Democratic Republic of the Congo. MedRxiv. 2020;20125567.

541 46. Holmes EC. Molecular Clocks and the Puzzle of RNA Virus Origins. J Virol.
542 2003;77(7):3893–7.

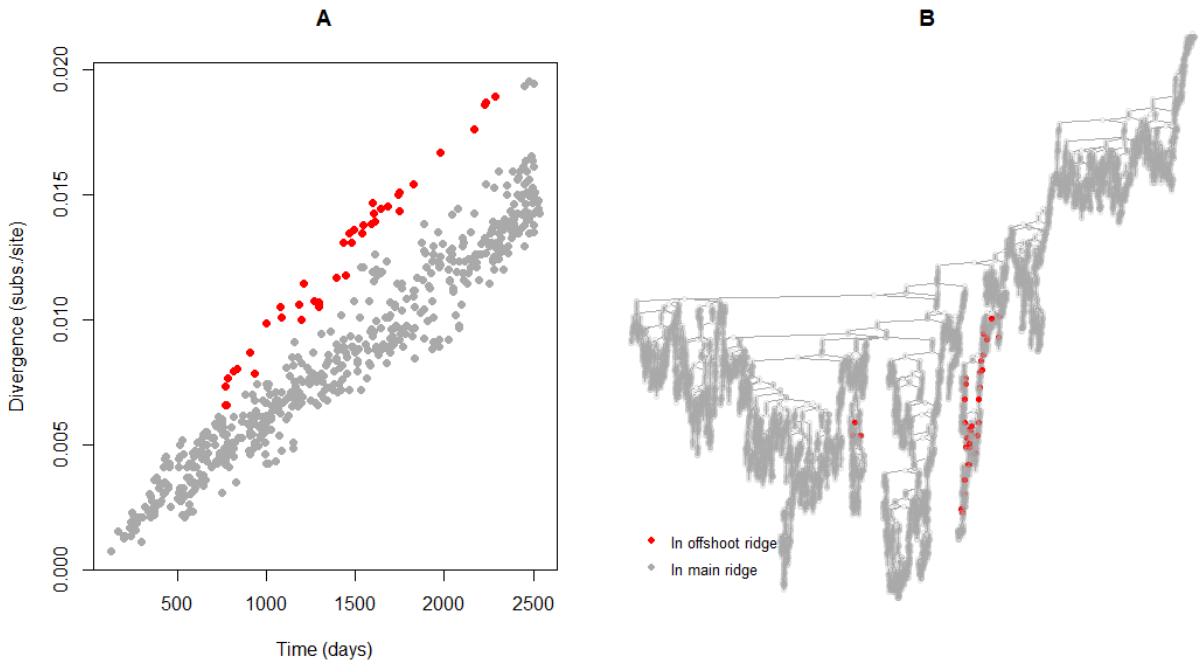
543 47. Van Dooren S, Salemi M, Vandamme AM. Dating the Origin of the African Human T-Cell
544 Lymphotropic Virus Type-I (HTLV-I) Subtypes. Mol Biol Evol. 2001;18(4):661–71.

545 48. Cori A, Nouvellet P, Garske T, Bourhy H, Nakouné E, Jombart T. A graph-based evidence
546 synthesis approach to detecting outbreak clusters: An application to dog rabies. PLOS
547 Comput Biol. 2018;14(12):e1006554.

548 49. Rambaut A, Drummond AJ, Xie D, Baele G, Suchard MA. Posterior Summarization in
549 Bayesian Phylogenetics Using Tracer 1.7. Syst Biol. 2018;67(5):901–4.

550

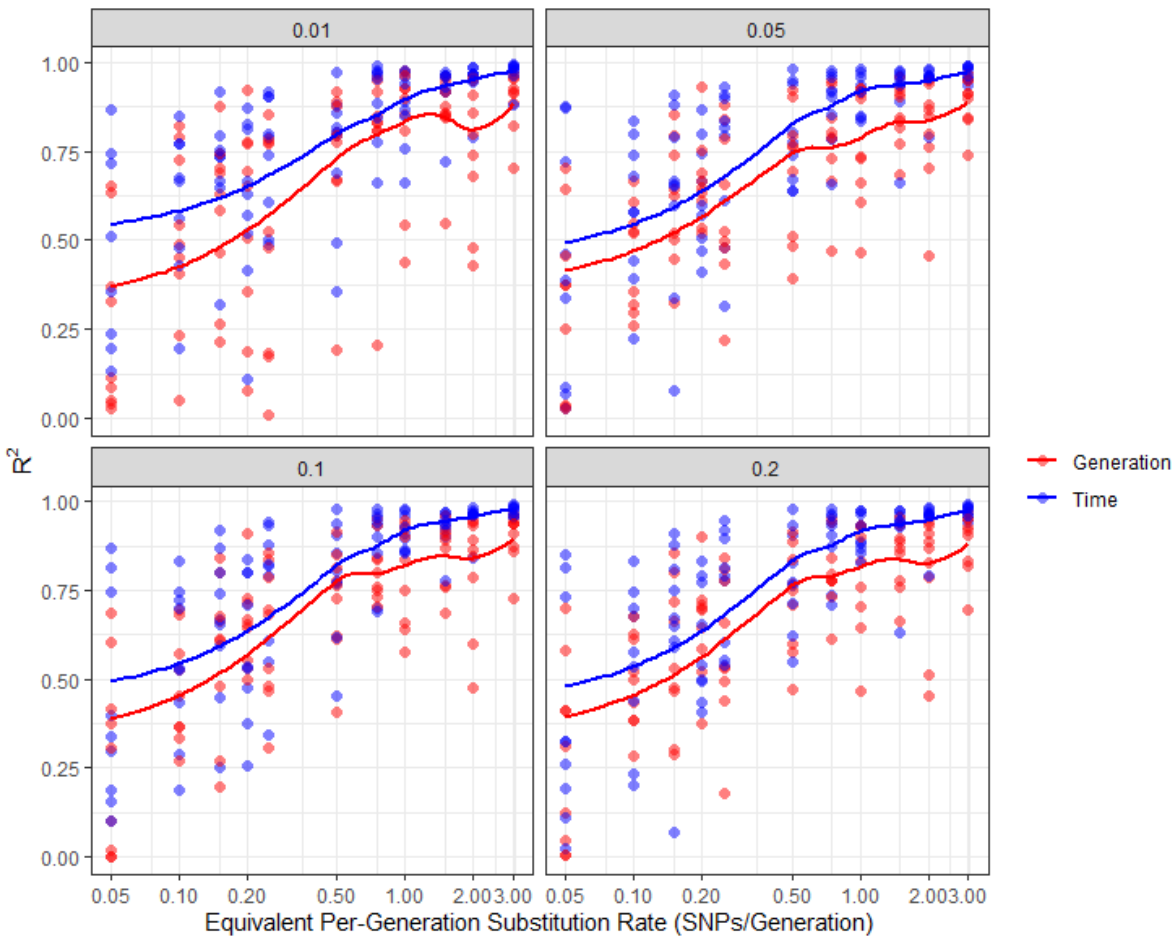
551 **Supplementary Information**



552

553 **Supp. Fig. S1: points in the offshoot ridge predominantly occur in one transmission tree.**
554 **(A)** root-to-tip divergence plot (2 SNPs/genome/generation, 5% sampling rate) with offshoot
555 ridge points highlighted in red. Offshoot ridge points are defined in this plot as having a
556 divergence rate above $8x10^{-6}$ substitutions/day and occurring after day 750. **(B)** transmission
557 tree of the simulated outbreak with offshoot ridge cases highlighted in red. Graph edge length
558 is not proportional to time or divergence.

559



Supp. Fig. S2: Sampling rate does not impact the R^2 of root-to-tip divergence plots from synthetic data. Plot is faceted by sampling rate as a proportion of the total number of cases in the outbreak, point colour represents mutation model.