A delay-deterministic model for inferring fitness effects from time-resolved genome sequence data

Nuno R. Nené¹, Alistair S. Dunham¹, and Christopher J. R. Illingworth¹

¹Department of Genetics, University of Cambridge, Cambridge, UK

ABSTRACT

A common challenge arising from the observation of an evolutionary system over time is to infer the magnitude of selection acting upon a specific genetic variant, or variants, within the population. The inference of selection may be confounded by the effects of genetic drift in a system, leading to the development of inference procedures to account for these effects. However, recent work has suggested that deterministic models of evolution may be effective in capturing the effects of selection even under complex models of demography, suggesting the more general application of deterministic approaches to inference. Responding to this literature, we here note a case in which a deterministic model of evolution may give highly misleading inferences, resulting from the non-deterministic properties of mutation in a finite population. We propose an alternative approach which corrects for this error, which we denote the delay-deterministic model. Applying our model to a simple evolutionary system we demonstrate its performance in quantifying the extent of selection acting within that system. We further consider the application of our model to sequence data from an evolutionary experiment. We outline scenarios in which our model may produce improved results for the inference of selection, noting that such situations can be easily identified via the use of a regular deterministic model.

Introduction

Fitness landscapes describe the relationship between the genome of an organism and its evolutionary fitness. Evolutionary fitness encompasses a broad range of important phenotypes of an organism, making the inference of details of fitness landscapes a topic of broad biological interest. In some important biological systems, adaptation occurs as a rapid and ongoing process. Where multiple beneficial mutations arise in a population simultaneously, linkage between mutations has a substantial impact upon adaptation; a considerable body of literature has characterised the implications of such effects for adaptation. Where adaptation is sufficiently rapid to be observed, time-resolved sequence data may be of assistance in measuring the extent to which a variant is under selection. Under the assumption of a large population size, the evolution of a single beneficial allele over time can be described by deterministic differential equations. Given sufficient observations of a population under study, the simplicity of this deterministic framework allows it to be extended to infer selection in far more complicated evolutionary scenarios; fitting a deterministic model to data provides an estimate of the magnitude of selection acting upon one or very many alleles. In other situations, genetic drift is an important factor to account for; in a small population, changes in allele frequency occurring via drift may outweigh those caused by selection. In this situation, a variety of methods have therefore been developed to consider the evolution of a single-locus, two-allele system, estimating in a joint calculation the effective size of a population, and the magnitude of selection acting upon a variant allele. In a similar calculation, one may estimate whether or not a change in the frequency of an allele has arisen through selection or genetic drift. Genetic drift induces an uncertainty in the future frequency of an allele; accounting for this, alleles which have changed by more than a given threshold may be identified, enabling the attribution of selection to genetic variants. A similar approach has been applied to the case where a population is large, but measurements of allele frequency are uncertain; model selection procedures discriminate ‘neutral’ from ‘selected’ behaviour in an allele frequency trajectory.

Where genetic drift is incorporated into a model, a variety of approaches to modelling Wright-Fisher propagation have been adopted. Numerical solution of the stochastic dynamics of the population may be computationally intensive, inspiring the development of more rapid propagation methods, and the consideration of potential alternative solutions. In a recent work, examining a range of potential models for the demographic history of a population, it was concluded that deterministic approximations to evolution under drift can produce accurate estimates of the magnitude of selection. Such models of selection, mutation, and recombination have been used to generate insights into viral adaptation. Time-resolved sequence data describing pathogenic populations is becoming increasingly available; in so far as demographic effects can be ignored in such systems, evolutionary inference becomes possible at far-reduced computational cost, making this an important area for
methodological development and application.

While acknowledging the potential for deterministic models to generate biological insight, we here present an important case in which a deterministic inference of selection from population sequence data produces a severely deficient result. In this case, a stochastic approach to inference produces a correct result, albeit with additional prior knowledge of the system and at the cost of a substantial amount of computational time. However, the use of what we term a delay-deterministic model, including a single extra model parameter, goes a substantial way to correcting the error in the deterministic calculation. We propose that under a range of circumstances we go on to describe, the delay-deterministic model combines the speed of a deterministic modelling framework with the accuracy provided by more computationally intensive models.

1 Materials and Methods

1.1 Simulated trajectories under a Wright-Fisher propagation model

Simulated data from a population was generated according to a model of sequential mutation and selection steps. We consider a population of \( N \) individuals occupying a linear network of \( L + 1 \) distinct haplotypes, each haplotype being separated from the previous one by a single mutation. We model the fitness of each haplotype as continually increasing with linear distance, such that the fitness of haplotype \( i \) is given by \( w_i = 1 + si \) for some parameter \( s \) (Figure 1). We denote the number of individuals of haplotype \( i \) in the population after \( t \) generations by \( n_i(t) \). Within each generation, propagation of the system was conducted using a simple model of mutation and selection. Mutation was modelled as occurring between adjacent haplotypes; for a given mutation rate \( \mu \), the number of mutants \( m_{ij} \) from haplotype \( i \) into an adjacent haplotype \( j \) was given by the Poisson distribution

\[
P(m_{ij} = k) = \frac{(\mu n_i(t) \Delta t)^k}{k!} e^{-\mu n_i(t) \Delta t}
\]  

Subsequently, the next generation was drawn via a multinomial process

\[
P(n(t+1)) = \frac{N!}{\prod_i n_i(t+1)} \prod_i \left( \frac{\bar{n}_i(t) \Delta t}{N \bar{w}} \right)^{n_i(t+1)}
\]  

where \( \bar{n}_i(t) \) is the number of individuals of haplotype \( i \), \( \bar{w} \) is the mean fitness of the population. Sequencing of the population was simulated via a multinomial emission model with sequencing read depth \( N_d \). We consider the behaviour of our system with a variety of parameters. Simulations were conducted with \( \mu = 10^{-5} \), population size \( N \in \{10^5, 10^9\} \), \( s \in \{0.1, 0.2, \ldots, 0.9\} \). A sequencing depth of \( N_d = 10^3 \) was simulated, with sampling from the population every generation; such a depth of sampling is likely unrealistic for a biological system but allows for clearer comparison of different inference methods. Our model therefore simulates the effect of strong selection, with \( Ns >> 1 \), but is not restricted to the strong mutation paradigm of \( \mu N >> 1 \). In each case the initial state of the system was defined by \( n_i(0) = N \) and \( n_i(0) = 0 \) for all \( i > 0 \).

![Figure 1](image-url)

**Figure 1.** Linear haplotype model used for simulation. The \( \ell^{th} \) haplotype has fitness \( w_{\ell} = 1 + is \); mutation occurs between adjacent haplotypes with constant rate \( \mu \).

1.2 Inference methods

Inferences of selection were conducted using an evolutionary model to generate inferred haplotype frequencies \( q_i(t) \) across time. Given a set of such frequencies, a multinomial log likelihood was calculated for the system

\[
L(\{q_i(t)\}) = \sum_k \log \frac{N_d!}{\prod_i n_i^k(t_k)!} \prod_i (q_i(t_k))^{n_i^k(t_k)}
\]  

where \( n_i^k(t_k) \) is the number of observations of the haplotype \( i \) at time \( t_k \), and the sum is calculated over data from all observed time points. The parameters \( q_i(t_0) \) and \( s \) were optimised in each case. Three models were used to generate inferred frequencies.
1.2.1 Deterministic inference model
In the first model, haplotype frequencies were modelled under the assumption of an infinite population size. As such, in each generation a fraction of each haplotype was subject to mutation

\[ M(q_i(t)) = (1 - \mu)q_i(t) + \mu \sum_j q_j(t) \] (4)

where as in the simulations the sum is conducted over all haplotypes \( j \) that differ from \( i \) by a single allele. Selection was included in a similarly deterministic manner:

\[ S(q_i(t)) = \frac{w_i q_i(t)}{\sum_j w_j q_j(t)} \] (5)

where the sum was calculated over all haplotypes. The next generation is given by

\[ q(t+1) = S(M(q(t))) \] (6)

1.2.2 Stochastic inference model
In the second model, allele frequencies were propagated in exactly the same way as in the model used for simulation. Stochastic simulations of viral populations have been used to explore the potential range of outcomes occurring in viral systems\(^{42}\). In order to sample the space of potential outcomes of the model, 1000 replicates of the model were run for each proposed set of parameters; the mean value of the likelihoods for the replicates, each calculated according to Eq.3, was maximised in order to identify parameters. In order to deal with the stochasticity of the likelihood function, an optimisation method was implemented that prevented the resampling of previously-tested model parameters.

1.2.3 Delay-deterministic model
Finally a delay-deterministic model was implemented, identical to the deterministic model described above, but with the addition of a delay representing the time for establishment of individuals with a novel haplotype. Specifically, the mutation function of Eq.4 was modified, with mutation out of a haplotype occurring only if the frequency of that haplotype was greater than a specific threshold.

\[ M(q_i(t)) = (1 - I_i(t)\mu)q_i(t) + \mu \sum_j I_j(t)q_j(t) \] (7)

where the index function

\[ I_i(t) = \begin{cases} 1 : q_i(t) \geq \beta \\ 0 : otherwise \end{cases} \] (8)

The parameter \( \beta \) was optimised to identify the maximum likelihood model.

We note that, of the three inference models, the stochastic model requires an estimation of the total population size, \( N \); for the sake of computational time we used the correct value of this parameter in our inferences. Neither the deterministic or delay-deterministic models require an estimate of population size.

1.2.4 Application to experimental data
To explore the use of our approach with experimental data, the deterministic and delay-deterministic models were applied to influenza sequence data collected from an evolutionary experiment conducted in ferrets\(^{43}\). A previous analysis of these data using a deterministic approach has been described in an earlier publication\(^{35}\); for simplicity we here considered within-host data from a single animal in the study, denoted F3501 in the original work, for which the initial population diversity was relatively low. Genome sequence data were processed as described in the previous publication\(^{35}\), identifying loci in the HA segment at which significant change in allele frequency was observed, then processing short read data spanning these loci into a set of multi-locus variant calls and inferring haplotype frequencies which best fit the observed data using a maximum likelihood model. Mutation was initially modelled as occurring deterministically between haplotypes, identifying an optimal model of haplotype fitness using model selection. Fitness parameters within this model were then re-inferred using the delay-deterministic framework, comparing inferred fitnesses from under each approach. In common with the default model in the original study, the rate of mutation was modelled as \( \mu = 10^{-5} \).
2 Results

The stochastic and delay-deterministic models produced the most accurate inferences of selection. Selection coefficients inferred from the three different models are shown in Figure 2. For each of the values of $\mu N$, the deterministic model underestimates the magnitude of selection, $s$, for values of $L$ greater than one (that is, where there were three or more haplotypes in total), with an increasing degree of underestimation as $L$ increases. Where $L = 5$, the gradient of a linear model fitted to the mean inferred frequencies was equal to 0.095; roughly one tenth of what would be given by a correct inference model. The results obtained also depend upon the value of $\mu N$; where this statistic is larger, the extent to which the deterministic model underestimated the true magnitude of selection was reduced; here for the case $L = 5$ the gradient of the fitted linear model was 0.39. Results from the stochastic inference model show a good reproduction of the correct fitness values; linear gradients varied between 0.98 and 1.02 for the case $\mu N = 1$, indicating an accurate reproduction of selection coefficients as would be expected given the identical models of propagation. The delay-deterministic method was close in performance to the stochastic model, with gradients between 0.88 and 0.97 indicating a small underestimate of the strength of selection. This underestimate was not reproduced in the case $\mu N = 10000$, for which gradients fitted to the delay-deterministic outputs were either side of 1.

Figure 2. The deterministic model substantially underestimates the correct magnitude of selection for models with multiple haplotypes. Parameters are shown for values of $L$ between 1 and 5 with (A) $\mu N = 1$. (B) $\mu N = 10000$. Points show mean inferred selection coefficients; error bars were determined from a set of 100 replicate calculations. Lines show the outputs of a linear regression model fitted to the mean values.

The results we obtained can be intuitively understood via a plot of the evolutionary dynamics of the linear system (Figure 3). Given a deterministic model with the correct selection coefficient, the population propagates through the haplotypes substantially faster than does the stochastic model. In the Wright-Fisher model, given that $N\mu = 1$, a mean of one individual mutates from haplotype 0 to haplotype 1 in the first generation. Following the second generation the probability of an individual being found in haplotype 2 is therefore approximately $\mu$. In so far as double mutations are ignored within our model framework, at least one individual is required to occupy a haplotype before the next haplotype can be founded via mutation; this leads to a delay of multiple generations before a single individual reaches the final haplotype, at which point selection of magnitude of $1 + Ls$ ensures the eventual fixation of this haplotype. By contrast, in the deterministic model, mutation propagates the population rapidly through the system; after $L$ generations the final haplotype is deterministically occupied by a frequency of the population of order $\mu^L$. The increased fitness of this final haplotype therefore takes effect on the system more rapidly, leading to the observed faster propagation. When the deterministic model is optimised, a lower fitness parameter $s$ is inferred to compensate for this effect, which increases dependent upon the number of haplotypes in the system. By contrast, the delay-deterministic model corrects for the error of the deterministic model. By imposing a delay on the rate at which new haplotypes are founded by mutation, the rate of propagation through the haplotypes is controlled, giving an improved fit to the data, and therefore a more accurate inference of selection.

At higher values of $\mu N$, differences between the stochastic and deterministic systems are reduced (Supporting Figure S1).
As $N$ tends to infinity, the number of individuals mutating between haplotypes per generation approaches the deterministic limit and the time at which a haplotype becomes established decreases, with the consequence that less of a reduction in $s$ is required to fit the model to the data. The deterministic model therefore provides a good description of the behaviour of the discrete system as $\mu L N$ in the discrete model approaches a value much larger than 1. For a within-host model of influenza, where $\mu$ may be of the order $10^{-444}$, and $N$ potentially of the order $10^{1442}$, this implies that the value of $L$ of 4 or greater could lead to failure of the deterministic model; we next consider the application of our models to data from an experimental evolution study.

Application of the deterministic and delay-deterministic methods to data from an evolutionary experiment showed only small differences between inferred parameters. Details of each inference are given in Supporting Table S1. Within this system evolution proceeds exceptionally fast, with mutation into new and highly advantageous haplotypes being inferred to drive the adaptation of the system over the course of an infection (Figures 4A, 4B). Application of the delay-deterministic value gave a marginally improved fit to the data, with a maximum likelihood value 0.63 units better than the deterministic model without accounting for the additional parameter. The value of $\beta$ was inferred to be $1.8 \times 10^{-10}$, smaller than any inferred initial haplotype frequency that was greater than zero. However, the inferred haplotype fitness values were very similar between the models, with deviations in fitness of not much greater than 1% (Figure 4C). This final result can be understood in terms of the arrangement of haplotypes within the system; although some haplotypes were inferred to have initially zero frequency, being created by mutation from other haplotypes, there was insufficient time for more distant haplotypes, at least two mutations away from initially occupied haplotypes, to increase to an appreciable frequency. This result is informative for calculations performed on biological datasets; even where selection for novel variants is extreme in nature, a delay-deterministic model is unlikely to be required to generate correct inferences of selection on timescales of 4-5 days. Previous inferences of selection for within-host influenza adaptation using deterministic methods are therefore unlikely to be negatively affected by the use of a deterministic model of mutation. Rather, the value of the method will arise over longer timescales, where the population grows under selection into haplotypes that are separated by multiple variants from sequences in the initial population.

### 3 Discussion

Deterministic models of adaptation have been proposed as a rapid and effective method for inferring selection from time-resolved sequence data. We have here highlighted a limitation of such frameworks whereby a deterministic model may severely underestimate the magnitude of selection in a system. This underestimation results from delays in the propagation of a finite population towards mutationally distant haplotypes; at least one individual is required to occupy a haplotype before mutation out of that haplotype may occur. As shown here, this delay operates even at high values of $N \mu$; we suggest that a population size satisfying $N \mu L > 1$ would be required to remove this effect. As a solution to this problem we propose an alternative inference procedure, which we term a delay-deterministic model. Under this model the progress of a nominally infinite population through the system is delayed via the use of an additional model parameter, matching the behaviour of the stochastic system; the additional parameter recreates the behaviour of a finite population, giving a correct inference of selection.

In demonstrating the application of our model, we have chosen the simplest possible situation in which the effects we are studying apply; that of a linear set of haplotypes separated by single mutations. Such a system has previously been considered in an application to cancer, calculating the time at which a novel haplotype might arise; while the system we consider is similar, our research question differs from this earlier study. We note that our approach is not limited to the linear set in so far as it can be applied to the inference of fitness landscapes of greater complexity. We further note that our model is not the only approach that would give a correct inference of selection under the circumstances of a population entering mutationally distant haplotypes. For example, inferring an ‘establishment time’ for each haplotype, at which a haplotype enters a population at a frequency above the selection-drift threshold, would likely generate correct results, albeit at the cost of learning a single parameter per haplotype in the system. The use of a model of time-dependent selection, in which selection only begins to affect a haplotype at a specific point in time, would also give an approximately correct inference of selection, delaying the impact of selection to a point at which the inferred trajectory with fit the model. However, this again would incur a computational cost, and would require the imposition on the system of the potentially incorrect assumption of a change in the magnitude of selection with time. Considering approximations to the stochastic system an approach based upon a reproduction of the stochastic distribution of allele frequencies would likely give a correct result albeit this has not to our knowledge been achieved. The delay-deterministic model we present here gives a computationally rapid approach to correctly infer the magnitude of selection.

While accurately reproducing the magnitude of selection, our model has an advantage over the stochastic approach in utilising a framework of deterministic propagation. Whereas, in the manner we implemented it, the stochastic model required 1000 replicate propagations of the model for each set of parameters tested, the delay-deterministic approach requires only
Figure 3. Time-dependent dispersion of trajectories for the case $\mu N = 1$, $L = 5$ and $s = 0.5$. Frequencies of each haplotype are shown reading left to right from the top. In each case the solid gray line (sometimes obscured) shows the mean haplotype frequency of the simulated data across time, calculated across 100 simulations. The region within one standard deviation of this frequency is indicated by gray dashed lines and is shaded. The black line shows the propagation of the deterministic model in the case where $s = 0.5$ and the population starts at the same haplotype distribution as the simulation, while the blue line shows the results of a maximum likelihood fit between the deterministic model and the mean data. The mean maximum likelihood fit of the delay-deterministic model to the data is shown as an orange line.
Figure 4. (A) Fitness landscape inferred from experimental data using the deterministic method. Inferred fitnesses are given for haplotypes for which the inferred frequency reached at least 1% within the course of within-host evolution. Given haplotypes describe the sequence of the HA segment of the virus at genome positions 339, 496, 728, 738, 788, 1018, 1020 and 1144. The haplotypes 9, 10, 13, and 15, marked in darker outline, were inferred to have zero initial frequencies. (B) Inferred changes in haplotype frequency over time. Haplotype frequencies are shown in gray, with the exceptions of haplotypes 1 (blue), 2 (green), 6 (yellow), 12 (purple) and 16 (red). (C) Differences in inferred haplotype fitness values between the deterministic and delay-deterministic methods shown proportional to the value inferred under the deterministic model.
the optimisation of a single additional parameter. The relative cost of this is likely to vary considerably depending upon the complexity of the system in question; in the application to influenza data considered here, where the model contained tens of parameters to be optimised, a single additional parameter is not likely to add substantially to the computational cost, provided the optimisation procedure is implemented in an efficient manner. We note that faster implementations of the stochastic framework are likely to be achievable; the recent demonstration of such methods for the two-allele case suggest the extension to more generalised population models to be a valuable avenue for exploration\textsuperscript{30,32}.

Applying our model to data from an evolutionary experiment, we identified very similar results between the deterministic and delay-deterministic methods; despite very high magnitudes of selection being inferred to act upon haplotypes in this case, little difference in the model inferences was found. We therefore propose that our method will be of relevance in cases where selection is strong and acts over longer time periods than those of the experiment considered, for which adaptation was observed over only a 5-day window. The identification of cases for which the deterministic model will produce correct or incorrect results is likely to be possible via application of the model itself; wherever a substantial proportion of a population is inferred to evolve into a haplotype more than two mutations distant from a haplotype occupied by the initial population, a delay-deterministic or similar approach should be considered.

References
1. de Visser, J. A. G. M. & Krug, J. Empirical fitness landscapes and the predictability of evolution. Nat Rev Genet. 15, 480–490 (2014).
2. Buonagurio, D. A. et al. Evolution of human influenza A viruses over 50 years: rapid, uniform rate of change in NS gene. Sci. 232, 980–982 (1986).
3. Bergland, A. O., Behrman, E. L., O’Brien, K. R., Schmidt, P. S. & Petrov, D. A. Genomic Evidence of Rapid and Stable Adaptive Oscillations over Seasonal Time Scales in Drosophila. PLoS Genet. 10, e1004775–19 (2014).
4. Barton, N. H. Linkage and the limits to natural selection. Genet. 140, 821–841 (1995).
5. Gerrish, P. J. & Lenski, R. E. The fate of competing beneficial mutations in an asexual population. Genet. 102-103, 127–144 (1998).
6. Gillespie, J. H. Is the population size of a species relevant to its evolution? Evol. 55, 2161–2169 (2001).
7. Rouzine, I. M., Wakeley, J. & Coffin, J. M. The solitary wave of asexual evolution. Proc. Natl. Acad. Sci. 100, 587–592 (2003).
8. Desai, M. M. & Fisher, D. S. Beneficial mutation–selection balance and the effect of linkage on positive selection. Genet. 176, 1759–1798 (2007).
9. Schiffels, S., Szollosi, G. J., Mustonen, V. & Lassig, M. Emergent Neutrality in Adaptive Asexual Evolution. Genet. 189, 1361–1375 (2011).
10. Good, B. H., Rouzine, I. M., Balick, D. J., Hallatschek, O. & Desai, M. M. Distribution of fixed beneficial mutations and the rate of adaptation in asexual populations. Proc. Natl. Acad. Sci. 109, 4950–4955 (2012).
11. Rouzine, I. M. & Weinberger, L. S. The quantitative theory of within-host viral evolution. J. Stat. Mech. Theory Exp. 2013, P01009 (2013).
12. Hartl, D. & Clark, A. Principles of Population Genetics (Sinauer, 2007).
13. Illingworth, C. J. & Mustonen, V. Distinguishing Driver and Passenger Mutations in an Evolutionary History Categorized by Interference. Genet. 189, 989–1000 (2011).
14. Illingworth, C. J. & Mustonen, V. Quantifying selection in evolving populations using time-resolved genetic data. J. Stat. Mech. Theory Exp. P01004–17 (2013).
15. Rouzine, I. M., Rodrigo, A. & Coffin, J. M. Transition between stochastic evolution and deterministic evolution in the presence of selection: General theory and application to virology. Micro. Mol. Biol. Rev 151–181 (2001).
16. O’Hara, R. B. Comparing the effects of genetic drift and fluctuating selection on genotype frequency changes in the scarlet tiger moth. Proc. Royal Soc. B: Biol. Sci. 272, 211–217 (2005).
17. Bollback, J. P., York, T. L. & Nielsen, R. Estimation of 2Nes From Temporal Allele Frequency Data. Genet. 179, 497–502 (2008).
18. Malaspinas, A.-S., Malaspinas, O., Evans, S. N. & Slatkin, M. Estimating allele age and selection coefficient from time-serial data. *Genet.* **192**, 599–607 (2012).
19. Mathieson, I. & McVean, G. Estimating Selection Coefficients in Spatially Structured Populations from Time Series Data of Allele Frequencies. *Genet.* **193**, 973–984 (2013).
20. Lacerda, M. & Seoighe, C. Population genetics inference for longitudinally-sampled mutants under strong selection. *Genet.* **198**, 1237–1250 (2014). DOI 10.1534/genetics.114.167957.
21. Foll, M. *et al.* Influenza virus drug resistance: A time-sampled population genetics perspective. *PLOS Genet.* **10**, 1–17 (2014). DOI 10.1371/journal.pgen.1004185.
22. Schraiber, J. G., Evans, S. N. & Slatkin, M. Bayesian inference of natural selection from allele frequency time series. *Genet.* **203**, 493–511 (2016).
23. Ferrer-Admetlla, A., Leuenberger, C., Jensen, J. D. & Wegmann, D. An approximate markov model for the wright-fisher diffusion and its application to time series data. *Genet.* genetics–115 (2016).
24. Kimura, M. Solution of a process of random genetic drift with a continuous model. *Proc. Natl. Acad. Sci. United States Am.* **41**, 144–150 (1955).
25. Feder, A. F., Kryazhimskiy, S. & Plotkin, J. B. Identifying signatures of selection in genetic time series. *Genet.* **196**, 509–522 (2014). DOI 10.1534/genetics.113.158220.
26. Topa, H., Jónás, Ā., Koller, R., Kosiol, C. & Honkala, A. Gaussian process test for high-throughput sequencing time series: application to experimental evolution. *Bioinforma.* **31**, 1762–1770 (2015). DOI 10.1093/bioinformatics/btv014.
27. Terhorst, J., Schlötterer, C. & Song, Y. S. Multi-locus analysis of genomic time series data from experimental evolution. *PLoS Genet.* **11**, 1–29 (2015). DOI 10.1371/journal.pgen.1005069.
28. Illingworth, C. J., Fischer, A. & Mustonen, V. Identifying selection in the within-host evolution of influenza using viral sequence data. *PLoS Comput. Biol.* **10**, e1003755 (2014).
29. Tataru, P., Simonsen, M., Bataillon, T. & Hobolth, A. Statistical inference in the wright-fisher model using allele frequency data. *Syst. Biol.* **66**, e30 (2016). DOI 10.1093/sysbio/syw056.
30. Khatri, B. S. Quantifying evolutionary dynamics from variant-frequency time series. *Sci. Reports* **6**, 32497 (2016).
31. Nené, N. R., Mustonen, V. & Illingworth, C. J. R. Evaluating genetic drift in time-series evolutionary analysis. *J. Theor. Biol.* **437**, 51–57 (2018).
32. Krukov, I., de Sanctis, B. & de Koning, A. P. J. Wright-Fisher exact solver (WFES): scalable analysis of population genetic models without simulation or diffusion theory. *Bioinforma.* **33**, 1416–1417 (2017).
33. Jewett, E. M., Steinrücker, M. & Song, Y. S. The effects of population size histories on estimates of selection coefficients from time-series genetic data. *Mol. Biol. Evol.* **33**, 3002–3027 (2016).
34. Ganusov, V. V. *et al.* Fitness Costs and Diversity of the Cytotoxic T Lymphocyte (CTL) Response Determine the Rate of CTL Escape during Acute and Chronic Phases of HIV Infection. *J. Virol.* **85**, 10518–10528 (2011).
35. Illingworth, C. J. R. Fitness inference from short-read data: Within-host evolution of a reassortant h5n1 influenza virus. *Mol. Biol. Evol.* **32**, 3012–3026 (2015). DOI 10.1093/molbev/msv171.
36. Sobel Leonard, A. *et al.* The effective rate of influenza reassortment is limited during human infection. *PLoS Pathog.* **13**, e1006203–26 (2017).
37. Shankarappa, R. *et al.* Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection. *J. Virol.* **73**, 10489–10502 (1999).
38. Zanini, F. *et al.* Population genomics of intrapatient HIV-1 evolution. *eLife* **4** (2015).
39. Xue, K. S. *et al.* Parallel evolution of influenza across multiple spatiotemporal scales. *eLife* **6**, e26875 (2017).
40. Houldcroft, C. J., Beale, M. A. & Breuer, J. Clinical and biological insights from viral genome sequencing. *Nat. reviews. Microbiol.* **15**, 183–192 (2017).
41. Park, S.-C., Simon, D. & Krug, J. The speed of evolution in large asexual populations. *J. Stat. Phys.* **138**, 381–410 (2010). DOI 10.1007/s10955-009-9915-x.
42. Russell, C. A. *et al.* The potential for respiratory droplet–transmissible a/h5n1 influenza virus to evolve in a mammalian host. *Sci.* **336**, 1541–1547 (2012). DOI 10.1126/science.1222526.
43. Wilker, P. R. et al. Selection on haemagglutinin imposes a bottleneck during mammalian transmission of reassortant H5N1 influenza viruses. Nat. Commun. 4, 1–11 (2013).

44. Pauly, M. D., Procario, M. C. & Lauring, A. S. A novel twelve class fluctuation test reveals higher than expected mutation rates for influenza A viruses. eLife 6, 686 (2017).

45. Beerenwinkel, N. et al. Genetic progression and the waiting time to cancer. PLoS Comput. Biol. 3, 2239–2246 (2007).

46. Illingworth, C. J. R. & Mustonen, V. A method to infer positive selection from marker dynamics in an asexual population. Bioinforma. 28, 831–837 (2012).

47. Kessinger, T. A., Perelson, A. S. & Neher, R. A. Inferring HIV escape rates from multi-locus genotype data. Immune system modeling analysis 348 (2015).

48. Martin, G. & Lambert, A. A simple, semi-deterministic approximation to the distribution of selective sweeps in large populations. Theor. population biology 101, 40–46 (2015).