Inferring population genetics parameters of evolving viruses using time-series data

Tal Zinger,[1,†] Maoz Gelbart,[1,‡] Danielle Miller,[1] Pleuni S. Pennings,[2] and Adi Stern[1,*]

[1]Department of Molecular Microbiology and Biotechnology, School of Molecular Cell Biology and Biotechnology, Haim Levanon Str., Tel-Aviv University, Tel-Aviv 69978, Israel and [2]Department of Biology, San Francisco State University, 1600 Holloway Ave, San Francisco, CA 94132, USA

*Corresponding author: E-mail: sternadi@tau.ac.il

†http://orcid.org/0000-0002-5580-5282

‡http://orcid.org/0000-0001-9268-9404

Abstract

With the advent of deep sequencing techniques, it is now possible to track the evolution of viruses with ever-increasing detail. Here, we present Flexible Inference from Time-Series (FITS)—a computational tool that allows inference of one of three parameters: the fitness of a specific mutation, the mutation rate or the population size from genomic time-series sequencing data. FITS was designed first and foremost for analysis of either short-term Evolve & Resequence (E&R) experiments or rapidly recombining populations of viruses. We thoroughly explore the performance of FITS on simulated data and highlight its ability to infer the fitness/mutation rate/population size. We further show that FITS can infer meaningful information even when the input parameters are inexact. In particular, FITS is able to successfully categorize a mutation as advantageous or deleterious. We next apply FITS to empirical data from an E&R experiment on poliovirus where parameters were determined experimentally and demonstrate high accuracy in inference.

Key words: fitness landscape; mutation rate; experimental evolution.

1. Introduction

Evolutionary biology has traditionally relied on inferring evolutionary processes using data from one time point, namely from the present. With the advent of evermore accurate next-generation sequencing (NGS) techniques, it is now possible to observe virus evolution in action—either through Evolve and Resequence (E&R) experiments (Acevedo et al. 2014; Foll et al. 2014; Stern et al. 2017) or from clinical samples obtained from patients (Ramachandran et al., 2011; Dunn et al., 2015; Zanini et al., 2015). Recent development of novel NGS techniques allows detection of ultra-rare alleles, even at frequencies of 10^{-4} or lower (Jabara et al. 2011; Meacham et al. 2011; Lou et al. 2013; Yang et al. 2013; Acevedo et al. 2014; Zhou et al. 2015; Gelbart et al. 2018; Salk et al. 2018). This allows tracking the fate of a mutation from the moment it is created, in particular in RNA viruses that have high mutation rates.

The dynamics of allele frequency over time depends on the following factors: (1) the relative fitness (w) of the allele as compared to the wild-type (WT) allele, (2) the population-wide mutation rate (μ), and (3) the population size (N) (Kimura, 1964) (Fig. 1A). These factors hence determine the probabilities of allele frequency trajectories (Fig. 1B). For large enough populations, an allele of low fitness will remain at low frequencies
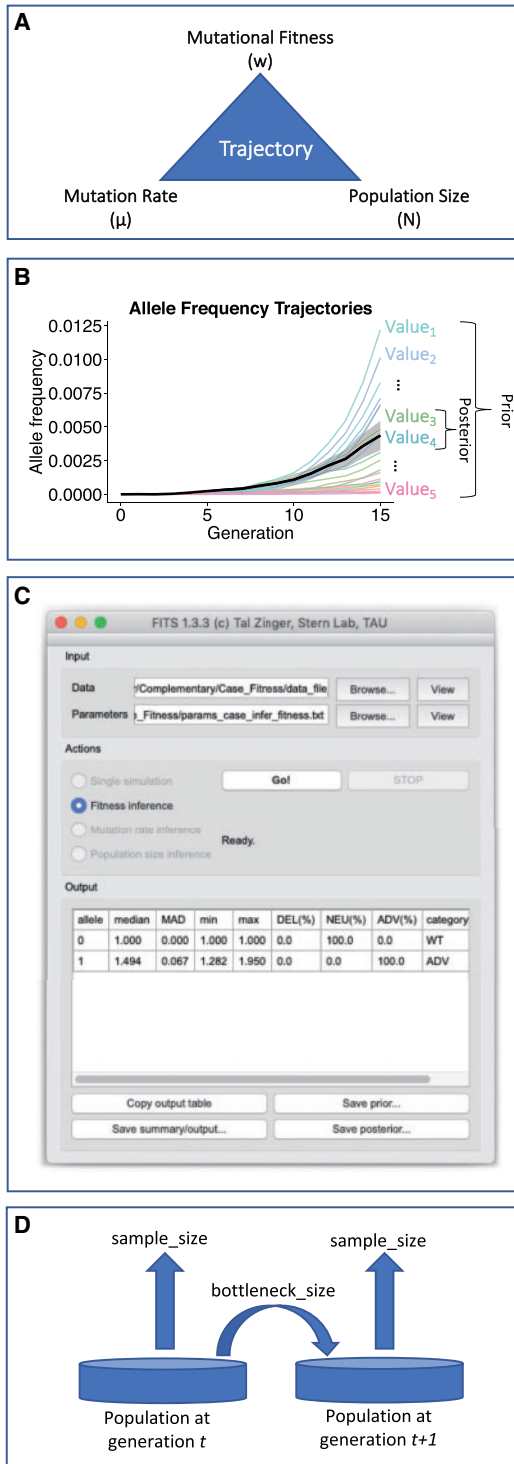
Figure 1. FITS overview. (A) An allele frequency trajectory is affected by the allele's fitness, mutation rate, and population size. FITS can infer the value of one of these factors if information is present about the other two. (B) Rejection-based ABC in FITS works by simulating trajectories, using sampled values for the missing factor from a prior distribution. Distance from the observed data (black line) is then measured for each trajectory and used as a summary statistic. The sampled values used to generate the trajectories closest to the observed data (shaded area) constitute an approximation for the posterior distribution. (C) FITS offers a user-friendly graphical user interface. The basic input required from the user is a *data file* with observed allele frequencies, and a *parameters file*, defining the parameters to be used for inference. Here, we see inference results for fitness, where the mutant allele is found to be advantageous. (D) FITS can account for bottlenecks in the size of the replicating population (parameter *bottleneck_size*), as well as sampling effects (e.g. a sample taken for sequencing) (parameter *sample_size*).

based on mutation-selection balance, where for haploid populations we expect the frequency f to be equal to $\frac{\mu}{1-w}$. Accordingly, a genome bearing a lethal mutation ($w = 0$) that is able to be packaged yet is unable to initiate a new infection is expected to be maintained at exactly the mutation rate μ—it is re-introduced each generation with a probability of μ and eliminated by selection at the end of each generation. The population size will alter the extent to which random genetic drift affects the trajectory, with small populations being more susceptible to random fluctuations of frequencies than large populations. The mutation rate μ determines the rate at which new mutations are introduced into the population at each generation. For viruses, a generation is defined here as a replication cycle.

There have been many notable advances in the development of approaches to infer selection and/or population size from time-series data (Bollback et al. 2008; Illingworth et al. 2012; Acevedo et al. 2014; Renzette et al. 2014; Feder et al. 2014; Foll et al. 2014, 2015; Ferrer-Admetlla et al. 2016; Jónás et al. 2016; Khatri 2016; Steinrücken et al. 2014; Schraiber et al. 2016; Terhorst et al. 2015; Topa et al. 2015). However, many of these methods either ignore genetic drift, are not designed for very low frequency alleles, or allow for only two alleles per locus. Recent advances in sequencing accuracy revealed that in virus populations, all four alleles (nucleotides) often co-segregate at the same position at very low frequencies (Acevedo et al. 2014). As we show below, considering this extra information can improve the inference.

Here, we introduce Flexible Inference from Time-Series (FITS), a user-friendly tool that allows the user to infer either the fitness of a mutation, or the mutation rate or the population size, from time-series data. FITS builds upon previous work (Foll et al. 2014) but incorporates several important improvements such as allowing for recurrent mutation and allowing for the inference of mutation rates and population size, and not only fitness. FITS also allows running single locus simulations under a Wright–Fisher model with selection, mutation, and drift, as described in the Section 2.1. Particularly, FITS is available either through a user-friendly graphical user interface (GUI, Fig. 1C) or as a command-line tool, allowing parallel processing of genome-wide data.

2. Methods

2.1 Overview—inferring parameters using ABC

FITS relies on the rejection-Approximate Bayesian Computation (ABC) method, which has gained popularity in recent years (Beaumont 2010; Csilléry et al. 2010; Sunnåker et al. 2013). We start off with empirical data of allele frequencies over time. Possible values for the factor in question (fitness, mutation rate, or population size) are sampled from a prior distribution and used for simulating trajectories (Fig. 1B). Simulations begin with the frequency of the first time-point given by the user in the input data file; this first time point will define the WT allele (most common allele) versus the mutant allele/s.

Simulated trajectories are generated using the two-step Wright–Fisher with selection model with k alleles. The first step of the model applies selection. The frequency of allele i at generation $t + 1$ before random genetic drift is $f_i^{t+1} = \sum_{j=1}^{k} \frac{w_i}{\overline{w}} f_i^t \mu_{j \rightarrow i}$, where w_i is the relative fitness of allele i, $\overline{w}$ is the mean fitness of the population with alleles frequencies $\overrightarrow{f^t}$, and $\mu_{j \rightarrow i}$ is the mutation rate from allele j to i. Next, random genetic drift is

Box 1. Limitations of FITS

FITS is designed to allow inference of fitness/mutation rate/population size from single-position time-series allele frequency data. Single position data is typical for next generation sequencing of virus populations, which is usually based on short reads, rendering the inference of haplotypes very difficult. FITS therefore does not take into account linkage among sites and we recommend using FITS in the following cases:

- Viruses (or other microbes) whose recombination rates are higher than the combined action of selection and mutation, and thus linkage will be broken down. This is true under some conditions for many (but not all) RNA viruses (Worobey and Holmes 1999).
- E&R experiments or *in vivo* data that are based on one to few founder viruses evolving for limited number of generations. For example, given a mutation rate of 10^{-5} for poliovirus and a clonal population at the beginning of the experiments, we expect that >90 per cent of the virus genomes will bear only one mutation after seven replication cycles, and >70 per cent after fourteen cycles as explored herein. Notably, for organisms with lower mutation rates this will be true for many more generations.

We recommend cautions when using FITS if:

- There is a high probability that there are many linked mutations on the same genome. This will occur, for example, if selection is strong enough to exceed the rate by which recombination breaks down linked sites (see discussion).
- The data is unreliable (for example, small sample size leading to uncertainty in allele frequency estimates, or sequencing errors that mask expected mutation frequencies).

applied, by using binomial sampling: $Pr\left(X_i^{t+1} = x \mid f_i^t, N\right) = \binom{N}{Nx}f_i^{t+1^{Nx}}\left(1 - f_i^{t+1}\right)^{N-Nx}$ where x denotes the number of alleles of type i sampled at generation $t+1$, leading to an allele frequency of x/N after both selection and drift. When performing serial passaging, a population bottleneck may be imposed upon the population every generation or every few generations, based on the user input. Furthermore, only a fraction of the genomes may be sampled for sequencing. FITS is able to account for both these bottlenecks (Fig. 1D) by performing additional binomial sampling steps.

After running the simulations, we measure the ℓ_1 distance between the observed and simulated trajectories. Namely, for each generation $t \in \{1 \ldots n\}$, the *sim*ulated frequency of each non-WT allele $i \in \{1 \ldots k-1\}$ is subtracted from the *obs*erved frequency and the absolute value is taken: $\ell_1 = \sum_{t=1}^{n}\sum_{i=1}^{k-1}\left|f_{i,sim}^t - f_{i,obs}^t\right|$. The top 1 per cent trajectories (i.e., with the minimal distance) are used as an approximation of the posterior distribution of the inferred factor (Sunnåker et al. 2013; Foll et al. 2014). From this distribution, we take the *median* as a point estimate for the inferred factor (Csilléry et al. 2010; Aeschbacher et al. 2013; van der Vaart et al. 2015). Finally, we test whether the posterior distribution is significantly narrower than the prior distribution (see Fig. 1B). This is done by applying Levene's test (van der Vaart et al. 2015). In the next sections, we will discuss how FITS can be used to infer (1) fitness, (2) mutation rates, or (3) population size. We also discuss how to use FITS to infer information from multiple independent loci.

2.2 Inferring fitness values

The relative fitness of an allele is a measure of its *advantageous* ($w > 1$), *deleterious* ($0 \leq w < 1$) or *neutral* ($w = 1$) effect on the reproductive success of the allele-bearing individuals, relative to individuals with the WT allele (for which $w := 1$). FITS assumes that $w \in [0, 2]$, assuming that most advantageous alleles tend to be no more than twice as fit as the WT. FITS may assume a uniform prior distribution in this interval, for which the user may define the bounds. Since fitness tends to be biased toward deleterious mutations (Sanjuán 2010; Huber et al. 2017; Peck and Lauring 2018), FITS uses as default a prior distribution that is based on empirical measurements of the fitness of distribution effects in viruses (Sanjuán et al. 2004, 2010) (see Supplementary Fig. S1). FITS also offers the ability to input a binned user-defined distribution as a prior.

2.3 Inferring the fitness category

Often a user may be less interested in the exact fitness value of an allele but will be more interested in broadly classifying an allele as deleterious, advantageous, or neutral. Therefore, we define an allele as *advantageous* (ADV) if at least 95 per cent of its posterior distribution is greater than 1, and *deleterious* (DEL) if at least 95 per cent of its posterior distribution is smaller than 1. This classification is based on the measure of significance for a posterior distribution previously proposed (Beaumont and Balding 2004), referred to as a "Bayesian P-value" (Foll et al. 2014). We can potentially classify an allele as neutral (NEU) when 95 per cent of the posterior distribution is equal to 1. Yet, we realize that even if an allele is neutral the posterior distribution will likely include values near one but not equal to 1. When more than 50 per cent (but less than 95%) of the posterior distribution is positioned within the appropriate interval, FITS ambivalently classifies the allele as ?ADV, ?NEU, or ?DEL.

2.4 Inferring mutation rates

Many models utilize a single mutation rate μ to describe the probability of an allele to change, that is for every single allele, $\mu = Pr(A_i \rightarrow A_{j \neq i})$. Nevertheless, several studies have recently measured mutation rates that vary quite a lot between different pairs of alleles (Abram et al. 2010; Acevedo et al. 2014; Zanini et al. 2017). FITS has been therefore designed to infer the mutation rate between any pair of alleles separately, given input on the fitness of the allele and the population size. FITS samples a value for the exponent (n) from a uniform prior, such that $\mu_{A_i \rightarrow A_j} = 10^n$. This allows obtaining a general idea of the order of magnitude of the mutation rate; more exact inference is challenging as we show below and may often be unnecessary when one is interested in a general estimate of mutation rates. Setting the range of the prior to $[-7, -2]$ captures most mutation rates of viruses (Sanjuán et al. 2010; Acevedo et al. 2014). When possible, we recommend using multiple loci with mutations known to bear the same fitness effects (e.g. synonymous mutations) for inferring mutation rates.

2.5 Inferring population size

The population size (N) affects the extent to which either genetic drift or selection exert their effect on the allele frequency trajectory. We sample the exponent (n) from a uniform distribution, such that $N = 10^n$, once again allowing the user to obtain the order of magnitude of the population size rather than an exact value which is very challenging to infer. A range of $[2, 8]$ should capture many experimental and natural settings. We note that for large values of N the allele frequency trajectory will differ ever so slightly, regardless of the precise value of N. For this reason, the aim of FITS is not necessarily to give an
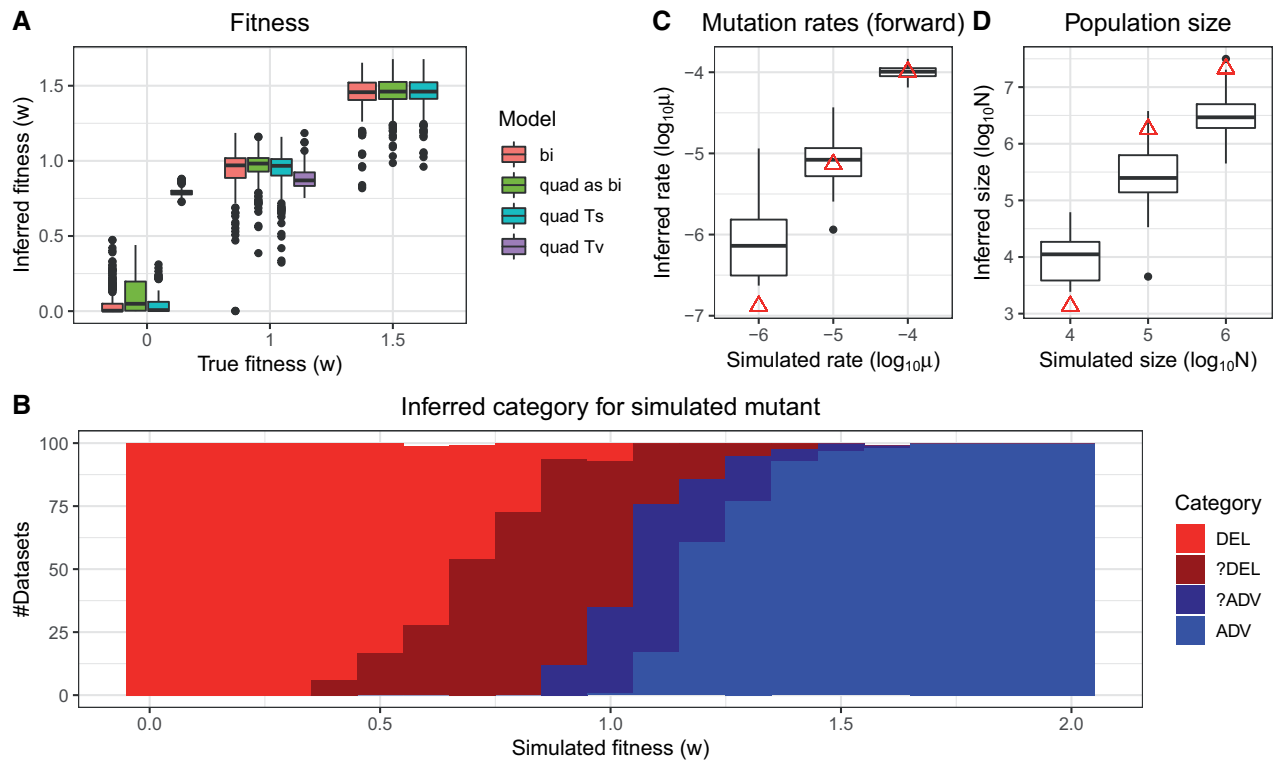
Figure 2. Results of FITS on simulated data. (A) Accuracy of fitness inference for a total of 2,700 simulated biallelic and quadrallelic datasets. Plotted separately are inferences for transitions under the biallelic model (bi), transitions or transversion under the quadrallelic model (quad Ts, quad Tv), and quadrallelic data that were collapsed to biallelic data (quad as bi). (B) Accuracy of FITS in classifying alleles with 100 datasets simulated for a range of fitness values. For each fitness value, we show the proportion of datasets that were classified in each category. Only inferences with significant P value were taken and as a result, some fitness values are represented by less than 100 datasets. (C and D) Accuracy of inference for forward mutation rate (C) and population size (D) inference, based on a set of neutral mutations. Boxplots show the distribution of inferences across different simulated values. Triangles represent joint multi-locus inference using all of the loci used for individual inference. Boxplots in (A), (C), and (D) display boxes spanning the lower and upper quartiles (interquartile range (IQR)). The middle band represents the median, and whiskers extend to 1.5 times the IQR. Points beyond this range are outliers.

accurate estimate of N, but rather to give upper or lower bounds on N.

2.6 Joint inference with multiple independent loci

FITS allows inferring a joint fitness value, mutation rate, or population size for multiple independent loci. In this case, the other two input parameters are assumed to be shared across all loci. Simulations are performed for each locus independently as described above, and the ℓ_1 distance is calculated. Next, however, the median of all ℓ_1 distances is used as a summary statistic to generate the posterior distribution for the missing parameter (see Fig. 2).

3. Results

3.1 Simulated datasets

In order to validate the accuracy of FITS, we tested it using simulated data. We began by simulating frequency trajectories using FITS under a biallelic model over 15 generations, with different combinations of parameter values: $N = \{10^4, 10^5, 10^6\}$, $\mu = \{10^{-6}, 10^{-5}, 10^{-4}\}$ and $w = \{0, 1.0, 1.5\}$. Simulations all began from an initial mutant allele frequency of zero, mimicking a situation where the population starts off without genetic variation. This is the case in many virus infections that are initiated by a very limited number of virus particles (Keele et al. 2008; Bull et al. 2011) and for many experimental setups (Pepin and

Wichman 2008; Acevedo et al. 2014; Lind et al. 2015; Stern et al. 2017; Hiltunen et al. 2018). For each parameter combination, 100 replicate datasets were generated, yielding a total of 2, 700 datasets. We then used FITS to infer parameters using these datasets.

3.2 Accuracy of fitness estimates

We analyzed how many datasets yielded a posterior distribution significantly narrower than the prior distribution based on Levene's test (van der Vaart et al. 2015). Results show that inference tends to be most reliable when $N\mu \geq 1$, with 99 per cent to 100 per cent of datasets analyzed yielding a narrowed posterior (Supplementary Fig. S2A). This most likely derives from the fact that when $N\mu < 1$, a new mutation may not be created at all, and allele frequencies will remain at zero for most generations, regardless of the fitness. Therefore, FITS outputs a warning on unreliable inference when given input parameters where $N\mu < 1$. We next turned to testing the effect of the number of simulations (i.e. the number of samples from the prior distribution) on reliability of inference (Nakagome et al. 2013). Indeed, we saw significantly more narrowed posteriors when increasing the number of simulations from 10^4 to 10^5 (t-test, $P = 0.0002$). However, increasing from 10^5 to 10^6 simulations did not lead to a significant improvement (t-test, $P = 0.27$, Supplementary Fig. S2B).

We went on to test the accuracy of FITS for the biallelic model on a subset of the simulated data, in which $N = 10^5$ and $\mu = 10^{-5}$ (Fig. 2A). Results of FITS were found to be quite satisfactory: the fitness of lethal alleles (true fitness equals to 0) was inferred as 0.05 ± 0.09 (mean and SD), the fitness of neutral alleles (simulated fitness equal to 1) was inferred as 0.9 ± 0.1 and the fitness of advantageous alleles (simulated fitness equal to 1.5) was inferred as 1.4 ± 0.1 (Fig. 2A). Thus, the fitness of lethal alleles was slightly overestimated whereas neutral and advantageous alleles were slightly underestimated. Finally, we tested using the ℓ_2 (Euclidean) distance as an alternative summary statistic and did not notice any improvement in inference (Supplementary Fig. S3).

We next sought to test the performance of FITS using the quadrallelic model (using four alleles). Our goal was to mimic as much as possible a biologically plausible dataset of virus alleles (Supplementary Table S1), with different mutation rates for transitions (10^{-5}) and transversions (10^{-6}), based on a transition/transversion ratio that is often around ten (Stoltzfus and Norris 2016). For the transition allele, 99 per cent of the posteriors were narrowed compared to 24 per cent for each of the transversion alleles. In general, FITS inferred the fitness of lethal transitions as 0.05 ± 0.08 and lethal transversions as 0.8 ± 0.03, the fitness of neutral transitions as 0.9 ± 0.1 and neutral transversions as 0.9 ± 0.1 and the fitness of advantageous transitions as 1.5 ± 0.1 (Fig. 2A). As expected by our results on the effects of $N\mu$, the accuracy of inference was strongly affected by the type of the mutation, with fitness of transition alleles inferred more accurately. To test if it is better to neglect transversions, we 'collapsed' our quadrallelic data into two alleles, by removing the transversions and normalizing the transitions and WT frequencies to one (Fig. 2A). This led to inferior inference of the lethal alleles, since they tended to be overestimated, suggesting that despite the inaccuracy in inferring transversions, taking the additional information present in transversions into account helps to increase the accuracy of the transitions.

Our results show that fitness values of neutral and advantageous alleles are slightly underestimated. This effect seems to be related to the stochastic effects of copy number: in the initial generations, the copy number of a newly born mutation is almost always very low (depending on N), and thus an allele may be lost and regenerated over several generations till it 'takes off' due to selection. This will lead to lower allele frequencies in general, which will resemble simulations with lower fitness values.

3.3 Classifying allele fitness

Although FITS gives a point estimate as an output, for many researchers, the category of the allele's fitness (DEL, ADV) is more important than the exact value. We thus set out to see how accurately FITS categorizes alleles. In order to do so, we generated datasets by simulating trajectories for 20 different fitness values ranging from zero to two, assuming $N = 10^5$ and $\mu = 10^{-5}$. For each fitness value, we generated 100 replicates and used FITS to classify the fitness of the mutant allele (Fig. 2B). Our results showed that in general, FITS is able to quite accurately classify the allele, in particular when including the ambivalent ?ADV and ?DEL labels as well. Reassuringly we found that the mutant allele was classified as advantageous only when it was indeed so. Only a few datasets (7/100) yielded the ?ADV ambivalent labeling, and none yielded ADV when the actual fitness value was $w \leq 1$. This is consistent with FITS' conservative estimation of advantageous alleles, which is a desired behavior for many users.

Finally, we set out to test FITS on simulated datasets based on multi-locus models that also take into account factors such as recombination and linked selection. We used FFPopSim (Zanini and Neher 2012), using the program's parameters of an HIV population replicating for 180 generations ($\sim$1 year). We tested the inference of FITS under two scenarios: dense sampling of the first ten generations or sparse sampling every ten generations of generations 10 through 180 (Fig. 3A). In general, FITS was quite successful in estimating the distribution of fitness values, with increased accuracy when more generations were taken into account (Fig. 3A). Deleterious alleles appeared to be sometimes underestimated, manifested as residual plots shifted to the left (Fig. 3B). Neutral and advantageous alleles were mostly inferred quite accurately (Fig. 3B) although we noted a consistent slight underestimation that become more pronounced when more generations were taken into account, possibly due to background selection. When focusing on classification of alleles into DEL/ADV, we noted that for alleles with simulated fitness up to 0.98, 39 per cent were classified correctly as DEL based on ten generations (this goes up to 93% if considering also ?DEL), and 99 per cent were classified correctly based on 180 generations. For advantageous alleles with fitness of 1.02 and higher, 12 per cent were classified correctly based on ten generations (this goes up to 82% if considering also ?ADV), and 71 per cent were classified correctly based on 180 generations (this goes up to 82% if considering also ?ADV).

3.4 Sampling effects

We considered that often the sample of genomes sequences may not correctly represent the allele frequencies in the population, as has been previously noted (Illingworth et al. 2017). This will be especially pronounced for rare alleles, which may likely not be sampled at all if the sample size is very small. In such cases, we would like to avoid incorrect inference by FITS. To this end, FITS can take into account the sample size in the simulations it performs. We tested FITS inference of fitness on a population size of 10^5, a mutation rate of 10^{-5}, yet with a sample size of 200, manifested by setting the FITS parameter *sample_size* to 200. Out of 100 datasets with $w = 0$ and with $w = 1$, we noted that Levene's test failed in more than half of the inferences or gave borderline results (P value that bordered 0.01). Moreover, the posterior distribution in these cases often spanned most of the range of the prior distribution. This was in stark contrast to all our previous results where Levene's test most often gave highly significant results ($P < 10^{-5}$) and a quite tight posterior distribution. We hence suggest that users take caution when observing results with a borderline Levene's test result and a wide spread of the posterior distribution.

3.5 Comparison with other tools

To further evaluate the accuracy of FITS, we set out to compare it with previously published tools. We attempted to analyze our simulated datasets using WFABC (Foll et al. 2015) and failed to get sufficient/reliable results for comparison under the conditions of large population size and rare alleles as explored herein (Supplementary Material). We further attempted to run another fitness inference method based on maximum-likelihood (Lacerda and Seoighe 2014). As stated by the authors, this method is less suitable for mutant allele frequencies approaching 0 or 1. We ran the R code (supplied by the authors) on our datasets and indeed got very inaccurate values (see Supplementary Table S2) for lethal and neutral alleles;
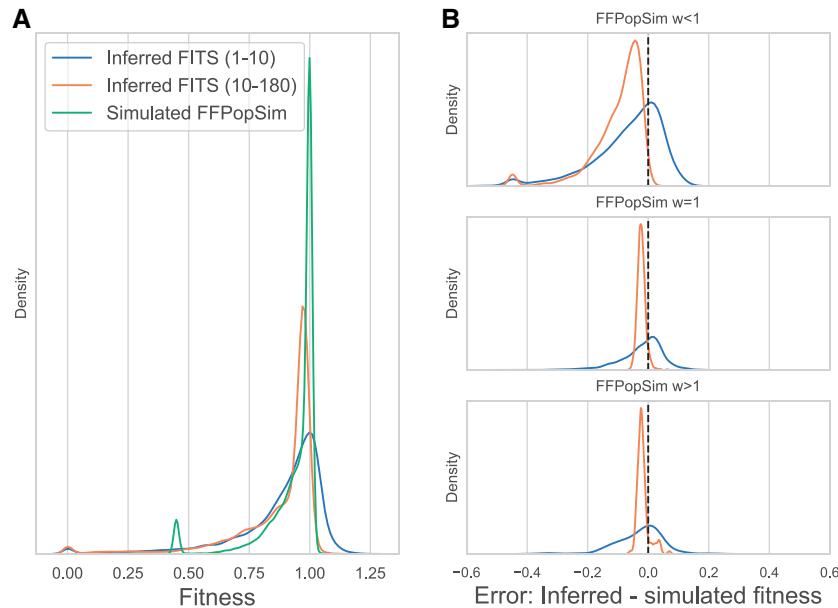
Figure 3. FITS inference on simulated multi-locus populations of HIV. Blue: inference based on the first ten generations; Orange: inference based on 180 generations; Green: "true" distribution, as simulated by FFPopSim. (A) Density plots of simulated ("true") versus inferred fitness values. (B) Residual density plots showing inference errors, defined as the difference between the inferred fitness value and the simulated "true" fitness value, segregated based on the "true" fitness value as given by FFPopSim.

advantageous alleles ($w = 1.5$) were inferred accurately. The inability to perform full direct comparison between FITS and the methods described emphasize FITS's novelty in its ability to correctly handle rare alleles.

3.6 Mutation rate accuracy

In order to infer the mutation rate, one must begin by studying an allele whose fitness is known. This may be assumed to be the case for synonymous mutations, which are most often neutral, or for an allele where external information is available regarding fitness. Here, we tested the ability of FITS to infer mutation rates given a neutral allele, by simulating datasets with varying mutation rates, while retaining $N = 10^5$ and $w = 1.0$. Forward and back mutation rates were set to be the same value. Not surprisingly, FITS was mostly unable to infer the back mutation rate, as manifested in only 12 out of 300 dataset analysis yielding narrowed posterior distributions for the back mutation (compared to 300 out of 300 for the forward mutation). This is likely because in our context, back mutations operate on the mutant allele, which exists at a very low copy number. We therefore focused only on inference of the forward mutation rate. Mean values across individual loci for the forward mutation rate exponent were measured at -4.0 ± 0.07, -5.1 ± 0.3, and -6.1 ± 0.4 for $log_{10}\mu = -4, -5,$ and -6, respectively (Fig. 2C). We next used our joint multiple loci approach, which estimates one posterior distribution for all loci at once. The median of this distribution is shown as triangles in Fig. 2C; Estimates are very accurate for the higher mutation rates, but this approach appears to be less accurate for low mutation rates as compared to aggregating results from single loci. Similar accuracy was found when simulating deleterious alleles and using them to infer the mutation rate (Supplementary Fig. S4).

We next set out to compare mutation rate inferences of FITS to inferences obtained using alternative methods on empirical data. Acevedo et al. (2014) used highly accurate sequencing to infer the frequency of lethal mutations in poliovirus type 1 and

based on mutation–selection balance used these frequencies to infer mutation rates. We used FITS to infer mutation rates either by inferring the mutation rate for each synonymous mutation independently and displaying boxplots with the median as the inferred mutation rate, or by using our multiple loci approach (see Section 2.6; Supplementary Fig. S5). We used synonymous mutations only. In general, the two methods agreed quite well on the mutation rates across almost all of the different transversions. However, there was discrepancy in the inference of the transitions: FITS inferred the transition rate as $\sim 10^{-5}$, whereas Acevedo et al. inferred it as $\sim 10^{-4}$. This discrepancy held even when synonymous mutations were filtered using various filters (high- or low-frequency mutations, mutations that reside in secondary structures, different metrics and summary statistics; Supplementary Fig. S6). Finally, we tested the difference in inference under a biallelic model versus a quadrallelic model and found that the quadrallelic inference led to less variance in the inferred rates, leading to a better separation between transitions and transversions (Supplementary Fig. S7).

3.7 Population size accuracy

The fitness of the allele (as well as the mutation rate) must be also known in order to infer the population size. Once again, we here mimicked inference given a neutral allele, by simulating $w = 1$ and $\mu = 10^{-5}$ over 100 datasets and inferred the population size using FITS (Fig. 2D). In terms of narrowed posterior distributions, we got fractions of 100/100, 98/100, and 87/100 for $N = 10^4, 10^5,$ and 10^6, respectively. Our point estimates of population size were 3.71 ± 0.66, 5.41 ± 0.55, and 6.49 ± 0.5 for $log_{10}N = 4, 5,$ and 6, respectively.

3.8 FITS inference given noisy input parameters

In many setups, population parameters may be imprecisely estimated. For example, virus population size may actually be smaller or larger by an order of magnitude due to either a

Table 1. Summary of possible inference errors obtained when inputting incorrect mutation rates or population sizes.

True category	Input error	Result
ADV	Too low N	Underestimation; may look like neutral
LETHAL	Too low N	Overestimation; may look like neutral
ADV	Too high μ	Underestimation; may look like neutral
NEU	Too high μ	Underestimation; may look like deleterious
LETHAL	Too low μ	Overestimation; may look like neutral

simple experimental error or due to inherent difficulty to infer it. Therefore, it is of great interest to see how FITS is affected by incorrect input parameter values.

We took a subset of the simulated datasets used for demonstrating the accuracy of FITS and ran the fitness inference again, intentionally using wrong values for the population size (Fig. 4A) and mutation rate (Fig. 4B). When focusing on incorrect population size used as input, we observed that our inference of fitness remained quite robust for both neutral and advantageous alleles. Fitness was slightly underestimated in these cases when the input population size was an order of magnitude lower than the true value, making FITS more conservative in estimation of adaptive alleles. For lethal alleles, we observed very accurate inference even when the input population size was too high. However, FITS overestimated the fitness of lethal alleles when the input population size was too low: lethal alleles were on average estimated as having a fitness of 0.2 ± 0.1 when the input population size was half of the real size and 0.8 ± 0.1 when the input population size was one tenth of the real size.

On the other hand, inputting wrong mutation rates to FITS had a more complex effect on the accuracy. For neutral and advantageous alleles, giving as input an extremely low mutation rate had little effect, and FITS quite accurately inferred the fitness values in these cases. However, too high an input for the mutation rates caused FITS to strongly underestimate the fitness values of neutral and advantageous alleles. In fact, neutral alleles were often estimated as lethal if the mutation rate given as input was an order of magnitude higher than the real value (0.1 ± 0.2). This is consistent with predictions from mutation–selection balance theory: a lethal allele is expected to be maintained at a frequency of the mutation rate. If the erroneously given mutation rate is very high, neutral alleles will remain below this mutation rate over the short time frame simulated and will hence be classified as lethal. While this is a critical point to notice, it still emphasizes that FITS remains conservative for advantageous alleles and will not report false positives. Similar to the case with inaccurate population sizes, fitness of lethal alleles tends to also be overestimated when too low a mutation rate is given as input. The consequences of inference with incorrect input values are summarized generally in Table 1.

In summary, when the mutation rate or population size are twice as high or twice as low as the real value (i.e. same order of magnitude), the inference of FITS is still quite robust. However, when FITS receives as input a parameter that is an order of magnitude higher or lower than the real value, this has a pronounced effect on inference of lethal (and presumably also non-lethal deleterious) alleles. Importantly, FITS remains conservative with advantageous alleles and tends to not overestimate their fitness.

3.9 Case Study – OPV2 Quadrallelic Analysis

We next set out to use FITS to analyze empirical data obtained from sequencing of oral poliovirus type 2 (OPV2) that we have previously performed (Stern et al. 2017). Briefly, OPV2 was serially passaged at 39.5˚C for seven passages, corresponding to fourteen generations. During the experiment, a population of $N = 10^6$ infectious virus particles (plaque-forming unit (PFU)) were seeded onto about 10^7 cells grown in tissue culture. Each passage was sequenced using highly accurate CirSeq sequencing (Acevedo et al. 2014), allowing the detection of mutations at a frequency as low as 10^{-6}. Coverage (number of reads covering a locus) spanned between 10^5 and 10^6 across all sequenced passages.

We used FITS as follows: first, we ran FITS on each locus (independently) to infer the fitness of each allele, assuming a population size of 10^6. Mutation rates given as input were based on estimates obtained previously based on linear regression of synonymous mutation frequencies, under the assumption that they are mostly neutral (Stern et al. 2017). Next, in order to test how FITS infers mutation rates, we ran FITS independently on each presumably neutral synonymous mutation. Accordingly, FITS was given $w = 1$ as input, and once again N was set to 10^6. Results were compared to the linear regression results obtained previously. Finally, we used FITS to also infer the population size, by running FITS independently on each (once again presumably neutral) synonymous mutation. Accordingly, FITS was given $w = 1$ as input, and the mutation rates were set to the values obtained from the linear regression. The results of all these analyses are described below.

3.10 Inferring fitness of each mutation in the genome of OPV2

We first ran an analysis with FITS using the biallelic model, applied to transition mutations only. Next, we ran an analysis using the quadrallelic model, applied to loci where all four nucleotides were observed. FITS was run on each locus independently. The results of both analyses give the distribution of fitness effects (DFE; Fig. 5) of the virus. Notably, this is a unique in-depth view of genomic evolution, enabled due to (1) the very high sequencing depth in the experiment, (2) the very high rate of mutation of the viral populations, and (3) the highly accurate sequencing approach used. Our results show a clear difference in the distribution of fitness effects obtained with transitions versus transitions + transversions. Transversions tended to be enriched with more non-lethal deleterious variants, whereas transitions were far less deleterious in general (Fig. 5). Indeed, this is in line with the genetic code structure, since transitions will more often create synonymous mutations, and when creating non-synonymous mutations, transitions often create more similar amino acids (Sella and Ardell 2002).

3.11 Inferring the population-wide mutation rates and population size of OPV2

We next set out to infer the mutation rate for the transition mutations of OPV2. Only loci where a synonymous transition mutation was observed were used for the analysis. The forward mutation rate estimates ranged between 10^{-6} and 10^{-5}. This is in agreement with the transition mutation rates we inferred previously using linear regression (Stern et al. 2017), which spanned $5 \times 10^{-6} - 10^{-5}$. In a similar manner, we inferred the population size of the virus, based on independent inference of the population size at each locus where a synonymous
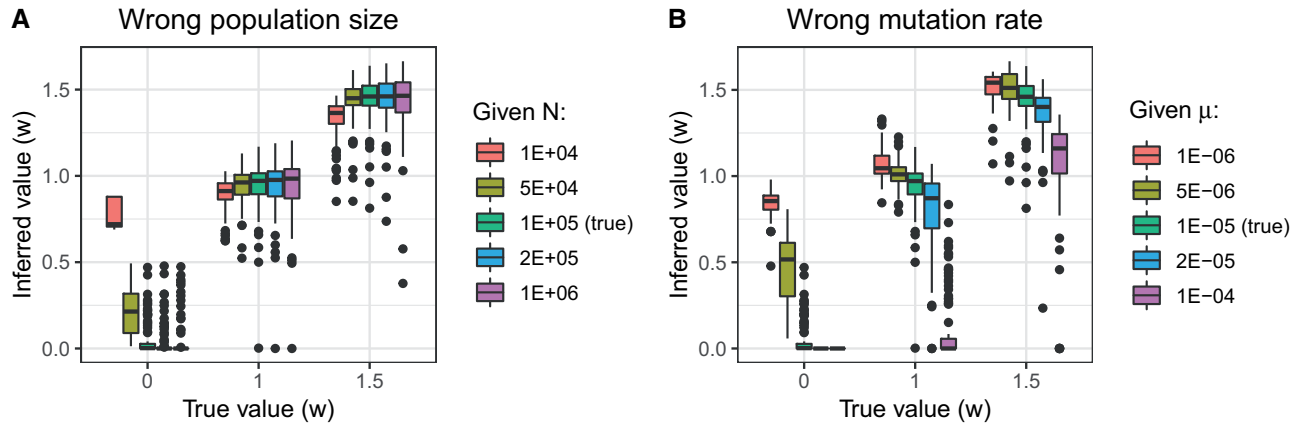
Figure 4. Effects of incorrect input values on FITS inference of fitness . Boxplots are as described in Fig. 2A. (A) Incorrect input population size, while mutation rate and fitness are fixed to true values. (B) Incorrect input mutation rate, while population size and fitness are fixed to true values. See main text for details.
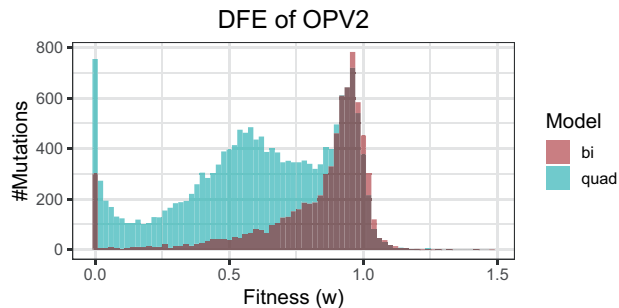

Figure 5. Inferred distribution of fitness effects (DFE) across all loci in the genome of OPV2 based on FITS under a biallelic or quadrallelic model. Using the quadrallelic model (transitions + transversions) extends the one of the biallelic model (transition mutations only), revealing additional deleterious alleles.

transition mutation was observed. Inferred population sizes ranged mostly between 10^5 and 10^6, which is largely in agreement with the experimental protocol used to seed 10^6 PFUs at each passage (Stern et al. 2017).

4. Discussion

We have developed FITS, a generic method that allows analyzing time-series data, and inferring the key parameters that shaped the evolutionary trajectory of an allele in an experiment or in real life settings. The program was designed with recent evolutionary experiments of RNA virus populations in mind (Acevedo et al. 2014; Stern et al. 2014, 2017). These experiments monitor a population of viruses that begins as a clonal entity and accumulates genetic diversity rapidly due to the high rate of mutation of the RNA viruses. In the initial setup of the experiment, genetic drift plays a prominent role, since mutations are born and present at low copy numbers (Supplementary Fig. S8). However, FITS is generic enough to be used to analyze other types of data, essentially any evolutionary experiment that tracks the population frequency of a trait over time.

Some of the key advantages of FITS include the fact that through the simulations, FITS is able to mimic the true biology of an allele as it is created and spreads in the population. FITS incorporates the mutation rate, allowing the introduction of new mutations along time. This is especially vital for new

arising mutations. Moreover, by directly modeling stochastic effects, FITS takes into account fluctuations in allele frequencies, which may be quite prominent in the first few generations. FITS is able to model four alleles, and we show that neglecting this information may result in some overestimation of lethal alleles and less robust estimation of mutation rates. On the other hand, we see no added value for the incorporation of back mutations (data not shown), suggesting that this is a parameter that the user does not need to supply. Reassuringly, our results show a very high level of accuracy even in some cases of very noisy simulated data. Finally, our results suggest FITS can successfully infer neutral and advantageous alleles even in the case where mutations are not independent, as shown by our simulations of HIV genomes. We do note that we seem to overestimate deleterious and lethal alleles; presumably this might occur due to hitchhiking effects. Nevertheless, these alleles are still categorized as deleterious alleles by FITS. In general, we hope the fact that we have delineated the conditions where FITS tends to err, will allow users to be cautious when interpreting the result.

Our results on mutation rate inference were sometimes different than those obtained previously based on mutation–selection balance of lethal alleles in poliovirus (Acevedo et al. 2014). While Acevedo inferred transition rates of ~10^{-4}, FITS inferred transition rates around 10^{-5}. One reason why this discrepancy may arise from the same data is the fact that FITS takes into account the time-series nature of the data whilst Acevedo et al. do not. In fact, the sequencing protocol includes a stage where very high multiplicity of infection (MOI) infection is performed; such high MOI may allow for complementation and hence an increase in the frequencies of lethal mutations (Stern et al. 2014). Accordingly, taking into account the change in frequency across time may mitigate the artificial inflation of frequencies at each time point. We further note that varying transition rates between 10^{-5} and 10^{-4} have been reported previously for poliovirus (de la Torre et al. 1990, 1992; Sanjuán et al. 2010), and suggesting that the 'real' value of the transition mutation rate is unclear and may depend on the method of measurement and inference.

It is important to delineate the limitations of FITS, which represent the assumptions of the Wright–Fisher model used for simulations and the framework used herein. It has been recently suggested that some of the Wright–Fisher assumptions,

such as a Poisson distribution of offspring, may not be appropriate for viruses (Sackman et al. 2019); this awaits further experimental investigation. Next, FITS assumes that loci evolve independently, and hence each locus is analyzed separately; accordingly, phenomena such as linkage, or epistasis, are not taken into account. We note that while we promote FITS for use with viruses with high recombination rates, strong selection that exceeds the recombination rate, as has been observed previously for CTL escape in HIV (Kessinger et al. 2013; Garcia et al. 2016), may still lead to strong effects of linkage and hence potentially erroneous inference by FITS. Future work will be required to perform direct modeling of non-independent evolution among sites in FITS. A second limitation of FITS has to do with the amount of information present in the experiment. Our simulation results showed that when the copy number of the allele is low, as reflected by a low $N\mu$, or when sampling results in loss of information on the allele copy number, there is not enough information to infer fitness with FITS. Moreover, reliable accurate sequencing is central when inferring parameters such as the mutation rate or low fitness alleles that segregate at very low frequencies. Importantly, one of the features of FITS is the ability of the program to detect unreliable inference, both when $N\mu$ is too low, and also when the posterior distribution yields no additional information over the prior distribution, and to output a warning to the user.

We note that FITS is designed to infer parameter values regarding one specific locus in a genome. However, FITS should be more robust when multiple loci are used to infer a specific parameter that is supposedly shared across many loci, such as the population size or mutation rate of a specific category of sites. This is also true for fitness—while naturally a user may be interested in the fitness of one particular allele, fitness inferred for a class of alleles (e.g. a particular type of non-synonymous mutations) will likely yield more robust results. This is evident when viewing the empirical data inferences from individual mutations (Supplementary Fig. S5) that often span an order of magnitude. When using more loci, a clearer view emerges as to where the mass of the distribution resides (e.g. the median) (Supplementary Figs S5 and S6). Notably, we also allow for joint inference using multiple loci, which most often agreed with the median of the individual inferences. However, we noted that when the number of loci was limited (<20), the joint inference approach often yielded less satisfactory results than site by site inference (data not shown).

To summarize, FITS is a generic tool that may be used for inferring fitness, mutation rates, or population size. We suggest that when genomic data is available, an iterative approach may be used: first, synonymous loci can be used to infer the mutation rates and the population size. Next, the inferred mutation rates and population size can be used as input to infer the fitness of each mutation. Finally, the mutations inferred as neutral can used to re-assess the mutation rates and population size. Future work will be required to test whether such an iterative scheme is robust and whether multiple parameters can be inferred at once. To the best of our knowledge FITS is the first available tool for inferring mutation rates and population sizes (but see (Ferrer-Admetlla et al. 2016) who allow inference of population-size scaled selection), and the first user-friendly tool for inferring fitness of mutations in virus populations. We have made a great effort in making FITS intuitive for understanding and for use, hopefully making this another milestone in making the tools of contemporary computational biology available to all virologists.

Acknowledgements

Data availability

FITS is written in C++ and is available both with a user-friendly graphical user interface but also as a command line program that allows parallel high throughput analyses. Source code, binaries (Windows, Mac and Linux) and complementary scripts, are available on GitHub at https://github.com/SternLabTAU/FITS.

Supplementary data

Supplementary data are available at *Virus Evolution* online.

Conflict of interest: None declared.

References

Abram, M. E. et al. (2010) 'Nature, Position, and Frequency of Mutations Made in a Single Cycle of HIV-1 Replication', *Journal of Virology*, 84: 9864–78.

Acevedo, A., Brodsky, L., and Andino, R. (2014) 'Mutational and Fitness Landscapes of an RNA Virus Revealed through Population Sequencing', *Nature*, 505: 686–90.

Aeschbacher, S., Futschik, A., and Beaumont, M. A. (2013) 'Approximate Bayesian Computation for Modular Inference Problems with Many Parameters: The Example of Migration Rates', *Molecular Ecology*, 22: 987–1002.

Beaumont, M. A., and Balding, D. J. (2004) 'Identifying Adaptive Genetic Divergence among Populations from Genome Scans', *Molecular Ecology*, 13: 969–80.

—— (2010) 'Approximate Bayesian Computation in Evolution and Ecology', *Annual Review of Ecology, Evolution, and Systematics*, 41: 379–406.

Bollback, J. P., York, T. L., and Nielsen, R. (2008) 'Estimation of 2Nes from Temporal Allele Frequency Data', *Genetics*, 179: 497–502.

Bull, R. A. et al. (2011) 'Sequential Bottlenecks Drive Viral Evolution in Early Acute Hepatitis C Virus Infection', *PLoS Pathogens*, 7: 1–14.

Csilléry, K. et al. (2010) 'Approximate Bayesian Computation (ABC) in Practice', *Trends in Ecology and Evolution*, 25: 410–8.

Dunn, G. et al. (2015) 'Twenty-Eight Years of Poliovirus Replication in an Immunodeficient Individual: Impact on the Global Polio Eradication Initiative', *PLoS Pathogens*, 11: e1005114.

Feder, A. F., Kryazhimskiy, S., and Plotkin, J. B. (2014) 'Identifying Signatures of Selection in Genetic Time Series', *Genetics*, 196: 509–22.

Ferrer-Admetlla, A. et al. (2016) 'An Approximate Markov Model for the Wright–Fisher Diffusion and Its Application to Time Series Data', *Genetics*, 203: 831–46.

Foll, M. et al. (2014) 'Influenza Virus Drug Resistance: A Time-Sampled Population Genetics Perspective', *PLoS Genetics*, 10: e1004185.

——, Shim, H., and Jensen, J. D. (2015) 'WFABC: A Wright-Fisher ABC-Based Approach for Inferring Effective Population Sizes and Selection Coefficients from Time-Sampled Data', *Molecular Ecology Resources*, 15: 87–98.

Garcia, V., Feldman, M. W., and Regoes, R. R. (2016) 'Investigating the Consequences of Interference between Multiple CD8+ T Cell Escape Mutations in Early HIV Infection', *PLoS Computational Biology*, 12: 1–23.

Gelbart, M. et al. (2018) 'AccuNGS: detecting ultra-rare variants in viruses from clinical samples', *bioRxiv*, doi: 10.1101/349498.

Hiltunen, T. et al. (2018) 'Dual-Stressor Selection Alters Eco-Evolutionary Dynamics in Experimental Communities', *Nature Ecology & Evolution*, 2: 1974–81.

Huber, C. D. et al. (2017) 'Determining the Factors Driving Selective Effects of New Nonsynonymous Mutations', *Proceedings of the National Academy of Sciences*, 114: 4465–70.

Illingworth, C. J. R. et al. (2012) 'Quantifying Selection Acting on a Complex Trait Using Allele Frequency Time Series Data', *Molecular Biology and Evolution*, 29: 1187–97.

—— et al. (2017) 'On the Effective Depth of Viral Sequence Data', *Virus Evolution*, 3: 1–9.

Jabara, C. B. et al. (2011) 'Accurate Sampling and Deep Sequencing of the HIV-1 Protease Gene Using a Primer ID', *Proceedings of the National Academy of Sciences*, 108: 20166–71.

Jónás, Á. et al. (2016) 'Estimating the Effective Population Size from Temporal Allele Frequency Changes in Experimental Evolution', *Genetics*, 204: 723–35.

Keele, B. F. et al. (2008) 'Identification and Characterization of Transmitted and Early Founder Virus Envelopes in Primary HIV-1 Infection', *Proceedings of the National Academy of Sciences*, 105: 7552–7.

Kessinger, T. A., Perelson, A. S., and Neher, R. A. (2013) 'Inferring HIV Escape Rates from Multi-Locus Genotype Data', *Frontiers in Immunology*, 4: 1–13.

Khatri, B. S. (2016) 'Quantifying Evolutionary Dynamics from Variant-Frequency Time Series', *Scientific Reports*, 6: 1–12.

Kimura, M. (1964) 'Diffusion Models in Population Genetics', *Journal of Applied Probability*, 1: 177–232.

de la Torre, J. C. et al. (1992) 'High Frequency of Single-Base Transitions and Extreme Frequency of Precise Multiple-Base Reversion Mutations in Poliovirus', *Proceedings of the National Academy of Sciences*, 89: 2531–5.

——, Wimmer, E., and Holland, J. J. (1990) 'Very High Frequency of Reversion to Guanidine Resistance in Clonal Pools of Guanidine-Dependent Type 1 Poliovirus', *Journal of Virology*, 64: 664–71.

Lacerda, M., and Seoighe, C. (2014) 'Population Genetics Inference for Longitudinally-Sampled Mutants under Strong Selection', *Genetics*, 198: 1237–50.

Lind, P. A., Farr, A. D., and Rainey, P. B. (2015) 'Experimental Evolution Reveals Hidden Diversity in Evolutionary Pathways', *eLife*, 4: e07074.

Lou, D. I. et al. (2013) 'High-Throughput DNA Sequencing Errors Are Reduced by Orders of Magnitude Using Circle Sequencing',

Proceedings of the National Academy of Sciences of the United States of America, 110: 19872–7.

Meacham, F. et al. (2011) 'Identification and Correction of Systematic Error in High-Throughput Sequence Data', *BMC Bioinformatics*, 12:

Nakagome, S., Fukumizu, K., and Mano, S. (2013) 'Kernel Approximate Bayesian Computation in Population Genetic Inferences', *Statistical Applications in Genetics and Molecular Biology*, 12: 667–78.

Peck, K. M., and Lauring, A. S. (2018) 'Complexities of Viral Mutation Rates', *Journal of Virology*, 92: 1–8.

Pepin, K. M., and Wichman, H. A. (2008) 'Experimental Evolution and Genome Sequencing Reveal Variation in Levels of Clonal Interference in Large Populations of Bacteriophage ϕX174', *BMC Evolutionary Biology*, 8: 1–14.

Ramachandran, S. et al. (2011) 'Temporal Variations in the Hepatitis C Virus Intrahost Population during Chronic Infection', *Journal of Virology*, 85: 6369–80.

Renzette, N. et al. (2014) 'Evolution of the Influenza a Virus Genome during Development of Oseltamivir Resistance in Vitro', *Journal of Virology*, 88: 272–81.

Sackman, A. M., Harris, R. B., and Jensen, J. D. (2019) 'Inferring Demography and Selection in Organisms Characterized by Skewed Offspring Distributions', *Genetics*, 211: 1019–28.

Salk, J. J., Schmitt, M. W., and Loeb, L. A. (2018) 'Enhancing the Accuracy of Next-Generation Sequencing for Detecting Rare and Subclonal Mutations', *Nature Reviews Genetics*, 19: 269–85.

Sanjuán, R. (2010) 'Mutational Fitness Effects in RNA and Single-Stranded DNA Viruses: Common Patterns Revealed by Site-Directed Mutagenesis Studies', *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365: 1975–82.

——, Moya, A., and Elena, S. F. (2004) 'The Distribution of Fitness Effects Caused by Single-Nucleotide Substitutions in an RNA Virus', *Proceedings of the National Academy of Sciences of the United States of America*, 101: 8396–401.

—— et al. (2010) 'Viral Mutation Rates', *Journal of Virology*, 84: 9733–48.

Schraiber, J. G., Evans, S. N., and Slatkin, M. (2016) 'Bayesian Inference of Natural Selection from Allele Frequency Time Series', *Genetics*, 203: 493–511.

Sella, G., and Ardell, D. H. (2002) 'The Impact of Message Mutation on the Fitness of a Genetic Code', *Journal of Molecular Evolution*, 54: 638–51.

Steinrücken, M., Bhaskar, A., and Song, Y. S. (2014) 'A Novel Spectral Method for Inferring General Diploid Selection from Time Series Genetic Data', *The Annals of Applied Statistics*, 8: 2203–22.

Stern, A. et al. (2014) 'Costs and Benefits of Mutational Robustness in RNA Viruses', *Cell Reports*, 8: 1026–36.

—— et al. (2017) 'The Evolutionary Pathway to Virulence of an RNA Virus', *Cell*, 169: 35–46.e19.

Stoltzfus, A., and Norris, R. W. (2016) 'On the Causes of Evolutionary Transition: Transversion Bias', *Molecular Biology and Evolution*, 33: 595–602.

Sunnåker, M. et al. (2013) 'Approximate Bayesian Computation', *PLoS Computational Biology*, 9: e1002803.

Terhorst, J., Schlötterer, C., and Song, Y. S. (2015) 'Multi-Locus Analysis of Genomic Time Series Data from Experimental Evolution', *PLoS Genetics*, 11: e1005069.

Topa, H. et al. (2015) 'Gaussian Process Test for High-Throughput Sequencing Time Series: Application to Experimental Evolution', *Bioinformatics*, 31: 1762–70.

van der Vaart, E. et al. (2015) 'Calibration and Evaluation of Individual-Based Models Using Approximate Bayesian Computation', *Ecological Modelling*, 312: 182–90.

Worobey, M., and Holmes, E. C. (1999) 'Evolutionary Aspects of Recombination in RNS Viruses', *Journal of General Virology*, 80: 2535–43.

Yang, X. et al. (2013) 'V-Phaser 2: Variant Inference for Viral Populations', *BMC Genomics*, 14: 674.

Zanini, F. et al. (2015) 'Population Genomics of Intrapatient HIV-1 Evolution', *eLife*, 4: e11282.

——, and Neher, R. A. (2012) 'FFPopSim: An Efficient Forward Simulation Package for the Evolution of Large Populations', *Bioinformatics*, 28: 3332–3.

—— et al. (2017) 'In Vivo Mutation Rates and the Landscape of Fitness Costs of HIV', *Virus Evolution*, 3: vex003.

Zhou, S. et al. (2015) 'Primer ID Validates Template Sampling Depth and Greatly Reduces the Error Rate of Next-Generation Sequencing of HIV-1 Genomic RNA Populations', *Journal of Virology*, 89: 8540–55.