Inferring Demography and Selection in Organisms Characterized by Skewed Offspring Distributions

Andrew M. Sackman,[1] Rebecca B. Harris, and Jeffrey D. Jensen
School of Life Sciences, Center for Evolution and Medicine, Arizona State University, Tempe, Arizona 85287

ABSTRACT The recent increase in time-series population genomic data from experimental, natural, and ancient populations has been accompanied by a promising growth in methodologies for inferring demographic and selective parameters from such data. However, these methods have largely presumed that the populations of interest are well-described by the Kingman coalescent. In reality, many groups of organisms, including viruses, marine organisms, and some plants, protists, and fungi, typified by high variance in progeny number, may be best characterized by multiple-merger coalescent models. Estimation of population genetic parameters under Wright-Fisher assumptions for these organisms may thus be prone to serious mis-inference. We propose a novel method for the joint inference of demography and selection under the Ψ-coalescent model, termed Multiple-Merger Coalescent Approximate Bayesian Computation, or MMC-ABC. We first demonstrate mis-inference under the Kingman, and then exhibit the superior performance of MMC-ABC under conditions of skewed offspring distributions. In order to highlight the utility of this approach, we reanalyzed previously published drug-selection lines of influenza A virus. We jointly inferred the extent of progeny-skew inherent to viral replication and identified putative drug-resistance mutations.

KEYWORDS time-sampled inference; selection; population genetics; coalescent theory; sweepstakes reproduction

ELUCIDATION of the underlying processes of evolution through the measurement of temporal changes in allele frequencies has remained a major focus of population genetics since the founding of the field (Fisher 1930; Wright 1931). Advancements in sequencing technologies over the last decade have dramatically increased the availability of genome-wide time-sampled polymorphism data for a wide variety of organisms, and several methods have been developed to analyze such data (Malaspinas *et al.* 2012; Mathieson and McVean 2013; Foll *et al.* 2014a; Lacerda and Seoighe 2014; Steinrücken *et al.* 2014; Ferrer-Admetlla *et al.* 2016; Schraiber *et al.* 2016; Shim *et al.* 2016; Rousseau *et al.* 2017). Of primary interest is the estimation of site-specific selection coefficients, and new methods account for nonequilibrium demography and environmental fluctuations by, for example, accounting for effective population size, population structure, and changing selection intensities.

Time-series polymorphism data are generally available from three sources: experimentally evolved populations, clinical patient samples, and ancient specimens. Viruses are well-represented among such data, owing both to their obvious clinical relevance, as well as their short generation times, small genomes, and relatively high mutation rates. However, aspects of viral biology render the application of standard population genetic inference methods problematic. Namely, existing methodologies for analyzing time-sampled polymorphism data are generally developed around the Kingman coalescent framework and the Wright-Fisher (WF) model (Wright 1931; Kingman 1982) and are of questionable applicability to organisms typified by large variances in offspring distributions, or so-called "sweepstakes reproduction," including not only viruses but many classes of prokaryotes, fungi, plants, and animals (reviewed in Tellier and Lemaire 2014; Irwin *et al.* 2016).

In particular, the WF model assumes constant population size, random mating, nonoverlapping generations, and Poisson offspring distributions with equal mean and variance. The Kingman coalescent is derived in the limit of the WF model and shares its assumptions. Reassuringly, population genetic

doi: https://doi.org/10.1534/genetics.118.301684
Manuscript received October 11, 2018; accepted for publication January 15, 2019; published Early Online January 16, 2019.
Supplemental material available at Figshare: https://doi.org/10.25386/genetics.7579943.
[1]Corresponding author: School of Life Sciences, Center for Evolution and Medicine, Arizona State University, 427 E. Tyler Mall, Tempe, AZ 85287. E-mail: amsackma@asu.edu

statistics and methods developed under the Kingman have been shown to be robust to many violations of WF assumptions (Möhle 1998, 1999), and have been extended to incorporate selection, migration, and population structure (Neuhauser and Krone 1997; Nordborg 1997; Wilkinson-Herbots 1998). However, large variance in offspring number (Eldon and Wakeley 2006; Matuszewski *et al.* 2018), strong selection (Neher and Hallatschek 2013; Schweinsberg 2017), large sample sizes (Wakeley and Takahashi 2003; Bhaskar *et al.* 2014), and recurrent selective sweeps (Durrett and Schweinsberg 2004, 2005) may violate the critical assumption underlying the Kingman coalescent that only two lineages may coalesce at a time. Such a violation may produce genealogies that are characterized by multiple-lineage mergers. Thus, the analysis of genomic data from organisms characterized by highly skewed offspring distributions—such as viruses—may be prone to serious misinference if examined with traditional WF and Kingman based approaches, even under neutrality. In particular, the neutral multiple merger events induced by the reproductive biology of these organisms may be mistaken for multiple-merger events induced by positive selection (Hallatschek 2018).

Though not widely utilized for inference, an alternative class of multiple-merger coalescent (MMC) models have been developed that are more general than the Kingman (*e.g.*, Bolthausen and Sznitman 1998; Pitman 1999; Sagitov 1999; Schweinsberg 2000; Möhle and Sagitov 2001), many being derived from Moran models generalized to allow multiple offspring per individual. Many of the recently derived MMC models form specific sub-classes of the Λ-coalescent, of which the Kingman is also a specific case, in which only two lineages are allowed to merge in a generation (Donnelly and Kurtz 1999; Pitman 1999; Sagitov 1999). It has been demonstrated that expectations under MMC models differ from those of the Kingman coalescent in several significant ways: effective population size (N_e) does not scale linearly with census size (N) as it does under the Kingman (Huillet and Möhle 2011); the site frequency spectrum (SFS) is skewed toward an excess of low- and high-frequency variants relative to the standard WF expectations, even under equilibrium neutrality (Eldon and Wakeley 2006; Blath *et al.* 2016); and the fixation probability of new beneficial mutations approaches one as population size increases (Der *et al.* 2011).

Eldon and Wakeley (2006, 2008, 2009) introduced a specific case of the broader class of Λ MMC models, the Ψ-coalescent, under which the parameter Ψ describes the proportion of offspring in the population originating from a single parent in the previous generation. The Ψ-coalescent has been used in several instances to infer the strength and frequency of sweepstake events in marine organisms typified by Type-III survivorship curves (Eldon and Wakeley 2006; Birkner *et al.* 2013; Blath *et al.* 2016; Matuszewski *et al.* 2018), and the expected SFS has been determined under both standard and nonequilibrium demography (Matuszewski *et al.* 2018).

Thus, we here introduce a novel statistical inference approach, termed Multiple-Merger Coalescent Approximate Bayesian Computation (MMC-ABC), for inferring population genetic parameters from time-sampled polymorphism data in populations subject to sweepstakes reproduction. MMC-ABC first characterizes the neutral demography of the population by generating genome-wide estimates of N and Ψ. It then estimates site-specific selection coefficients under the inferred sweepstakes model. We demonstrate that failing to account for skewed offspring distributions results in strong mis-inference of both demography and selection, and that MMC-ABC is capable of accurate joint estimation of offspring skew and selection coefficients even when the population size is not precisely known.

Materials and Methods

Forward simulation of populations under the Ψ-coalescent

Eldon and Wakeley (2006) described a model, the Ψ-coalescent, where each reproductive event in a population of size N is either, with probability $1-\epsilon$, a standard WF event yielding a single offspring, or, with probability ϵ, a multiple-merger event yielding ΨN offspring. The probability $\epsilon = 1/N^{\gamma}$ such that the coalescent history of a sample is dominated by multiple-merger events when $0 < \gamma < 2$, and $\gamma \geq 2$ produces a coalescent history typical of the Kingman. The rate at which k out of n lineages merge under the Ψ-coalescent is therefore (Tellier and Lemaire 2014):

$$\lambda_{n,k} = \binom{n}{k} \Psi^{k} (1-\Psi)^{n-k}, \text{ with } 0 < \Psi < 1$$

Under this model, Ψ has a straightforward biological interpretation. Namely, it is equal to the proportion of individuals in generation t_i who are the offspring of a single individual in t_{i-1} (Eldon and Wakeley 2006). We simulated populations evolving under a Ψ-coalescent model with SLiM version 3 (Haller and Messer 2019). To circumvent the WF framework of SLiM, we utilized a system of subpopulations with migration to achieve the same effect as sweepstakes reproduction events. Each generation consists of three steps:

1. One individual is chosen from the population (A) and placed in a separate subpopulation (B) of size $N = 1$. The unidirectional migration rate from B to A is set to Ψ.
2. One WF generation occurs, with migration from subpopulation B resulting in the chosen individual contributing $N\Psi$ of the individuals of the next generation of A. A series of mate choice callbacks within SLiM force the migration rate to be exact, rather than stochastic (see source code in the Supplemental Materials). Thus, each generation is a mix of $N(1-\Psi)$ individual WF reproductive events and a single sweepstakes event of magnitude $N\Psi$.
3. Subpopulation B is removed, and the next generation begins.

N_e-based ABC method

The data X consist of allele frequency trajectories measured at L loci: x_i $(i = 1, \ldots, L)$. The N_e-based ABC methodology

(modified from the method of Foll *et al.* 2014a) infers genome-wide values of N and Ψ and L locus-specific selection coefficients $s_i (i = 1, \ldots, L)$. At a particular locus i, we can approximate the joint posterior distribution as:

$$P(N, \Psi, s_i|X) \approx P(N, \Psi|T(X))P(s_i|N, \Psi, U(X_i))$$

where $T(X) = T(X_1, \ldots, X_L)$ denotes summary statistics chosen to be informative about N and Ψ that are a function of all loci, and $U(X_i)$ denotes locus-specific summary statistics chosen to be informative about s_i. A two-step ABC algorithm as proposed by Bazin *et al.* (2010) is used to approximate this posterior:

Step 1. Obtain an approximation of the density

$$P(N, \Psi|T(X)) \approx P(N, \Psi|X)$$

1. Simulate L trajectories for J populations $X_{i,j}$ using the starting frequencies from the first time point in each trajectory x_i, with N and Ψ for each trajectory sampled randomly from their priors, and J equal to the total number of simulation replicates.
2. Compute $T(X_{i,j})$ for each simulated population.
3. Retain the simulations with the smallest Euclidian distance between $T(X)$ and $T(x)$ to obtain a sample from an approximation to $P(N, \Psi|T(X)) \approx P(N, \Psi|X)$.

Step 2. For loci $i = 1$ to $i = L$:

1. Simulate K trajectories $X_{i,k}$ from a Ψ-coalescent model, with s_i sampled randomly from its prior, and N and Ψ from the joint density obtained in Step 1.
2. Compute $U(X_{i,k})$ for each simulated trajectory.
3. Retain the simulations with the smallest Euclidian distance between $U(X_i)$ and $U(x_i)$ to obtain a sample from an approximation to $P(s_i|N, \Psi, X_i)P(N, \Psi|X) = P(N, \Psi, s_i|X)$.

In Step 1 of MMC-ABC, a population of size N and skew Ψ (chosen from their prior distributions) is evolved with variants matching those described by the empirical data. The starting frequencies $x_{i,1} (i = 1, \ldots, L)$ are identical to those observed during the first sampled time point. The frequency of each allele under consideration is output at each generation of the trajectories in X.

As in the WF-ABC methodology of Foll *et al.* (2014b), we define $T(X)$ as a single statistic, Fs', an unbiased estimator of N_e under the WF model, given by Jorde and Ryman (2007):

$$Fs' = \frac{1}{t} \frac{Fs\left[1 - 1/\left(2\tilde{n}\right)\right] - 2/\tilde{n}}{(1 + Fs/4)[1 - 1/(n_y)]} \text{ with } Fs = \frac{(x-y)^2}{z(1-z)}$$

where x and y are the minor allele frequencies at the two time points separated by t generations, $z = (x + y)/2$, and $\tilde{n}$ is the harmonic mean of the sample sizes n_x and n_y at the two time points expressed in the number of chromosomes (twice the number of individuals for diploids). We averaged Fs' values over sites and times to obtain a genome-wide estimator of $N_e = 1/Fs'$ for haploids and $N_e = 1/2Fs'$ for diploids (Jorde and Ryman 2007). Note that we use the common notation where N_e corresponds to the effective number of individuals, and the corresponding number of chromosomes for diploids is $2N_e$.

In Step 2 of MMC-ABC, simulations are performed for each site with an initial allele frequency and sample size matching those observed, and with N and Ψ drawn from a joint posterior derived during Step 1 and the selection coefficient s chosen from its prior. At each site we utilize two summary statistics derived from Fs': $U(X_i) = (Fsd'_i, Fsi'_i)$, with Fsd' and Fsi' calculated, respectively, between pairs of time points where the allele considered is decreasing and increasing in frequency, such that, at a given site, $Fs' = Fsd' + Fsi'$. For the diploid model, we define the relative fitness as $w_{AA} = 1 + s$, $w_{Aa} = 1 + sh$, and $w_{aa} = 1$, where h denotes the dominance ratio (1 = dominant, 0.5 = codominance, 0 = recessive), and as $w_A = 1 + s$ and $w_a = 1$ for the haploid model (Ewens 2004).

Simulated data sets for testing performance of MMC-ABC

The data used for testing the performance of MMC-ABC were generated in one of two ways:

1. A diploid population of size N was first evolved under standard, neutral WF conditions for a burn-in period of 50,000 generations, and then evolved for a period of time under sweepstakes conditions. The frequencies of every segregating allele were output at the onset of sweepstakes conditions and at predetermined intervals for a set number of generations, including mutations present at the start of output as well as mutations that arose or fixed during the output period. Trajectories meeting minimum criteria (at least three informative time points, at least two consecutive time points with frequency >0.01, and at least one time point with frequency higher than 0.025) were retained. Data were generated in this manner for testing the performance of Step 1 of MMC-ABC (joint estimation of N and Ψ). Unfiltered single-time point population data were used to generate the observed SFS data in Figure 1.
2. Individual trajectories of mutations of a given starting frequency with selection coefficient s were modeled in a diploid population of size N with free recombination so that all sites were unlinked, with allele frequency trajectories and sweepstakes dynamics beginning in generation one. Trajectories generated in this manner were pooled into larger data sets for use in testing the performance of Step 2 of MMC-ABC (estimation of site-specific selection coefficients), with all allele trajectories beginning at a minor allele frequency of 10%—a frequency low enough that most neutral mutations should not fix, but high enough to ensure the availability of multiple informative time points for most trajectories.

We frequently used a fixed value of $\Psi = 0.1$ throughout our study, as this is close to the value estimated for experimentally evolved lines of influenza analyzed below. Additionally, at this

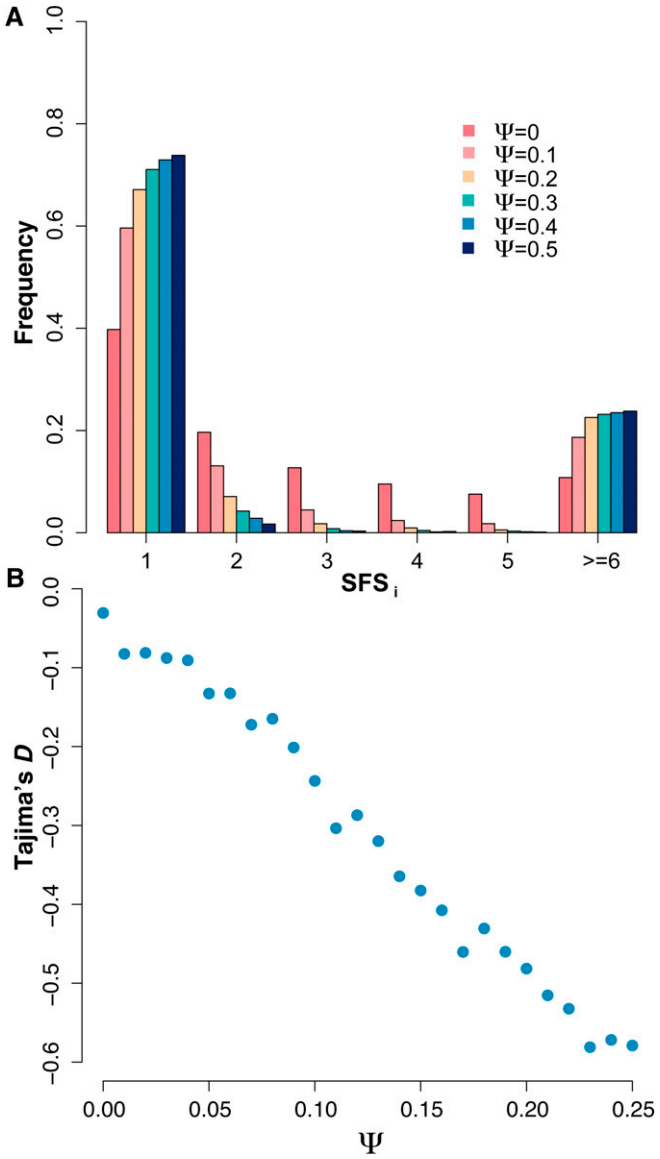

Figure 1 (A) The site frequency spectrum (SFS) for $\Psi \in \{0, 0.1, 0.2, 0.3, 0.4, 0.5\}$, derived from values averaged over 100 replicate simulations at $N = 1000$ with sample size $n = 250$. (B) The value of Tajima's D for $0 \leq \Psi \leq 0.25$ averaged across 100 replicate simulated populations with sample size $n = 30$. As shown, offspring skew strongly biases commonly used summary statistics, even under equilibrium neutrality.

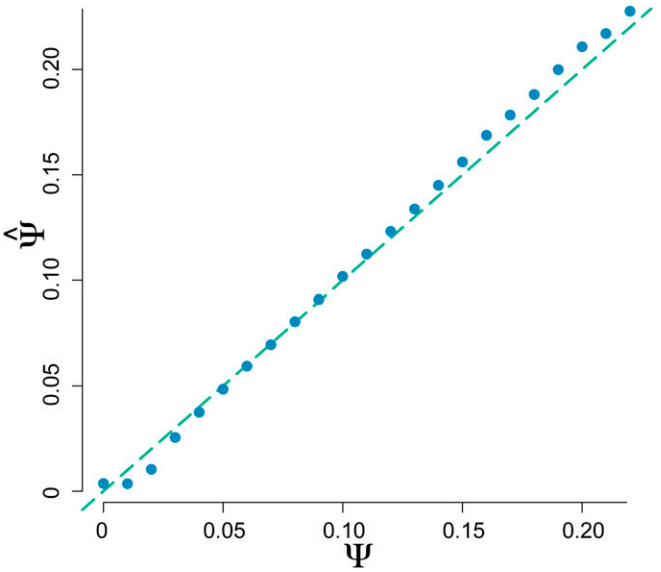

Figure 2 Estimation of Ψ by MMC-ABC. Estimates were averaged over 1000 replicate populations of size $N = 1000$, with an average of 300 polymorphic sites per population tracked at 20 time points over 200 generations with a minimum of 8 informative time points, with the correct value of N specified to MMC-ABC. True values of Ψ are indicated by the dashed line. Thus, MMC-ABC accurately estimates the value of Ψ from time-series data when the true value of N is known.

level of skew, multiple mergers should dominate the coalescent history of a population without entirely eliminating all segregating variation. When $\Psi > 0.25$, variation is generally eliminated from the population more quickly than it can be generated, and we therefore restricted most of our analyses performed over a range of Ψ to values of $\Psi < 0.25$.

Analysis of drug-resistance in influenza A virus

We applied MMC-ABC to time-series polymorphism data from experimentally evolved populations of influenza A virus, originally described by Foll *et al.* (2014a). The data consist of population genomic sequencing from two control lineages and two lineages exposed to exponentially increasing concentrations of the influenza drug oseltamivir, reared on Madin-Darby canine kidney (MDCK) cells and sampled every 13 generations. The data were previously analyzed with WF-ABC and putative drug-resistance mutations were identified. We reanalyzed the data with MMC-ABC for comparison.

Data availability

The source code and manual for MMC-ABC, along with the SLiM and python scripts used to generate our simulated data, are publicly available at https://github.com/sackmana/MMC-ABC/. The raw data from the experimentally evolved influenza virus populations can be found at the ALiVE repository at http://bib.umassmed.edu/influenza/. Supplemental material available at Figshare: https://doi.org/10.25386/genetics.7579943.

Results and Discussion

Effects of skewed offspring distributions on variation within populations

To underscore the importance of properly accounting for skewed offspring distributions when inferring selection from population genetic data, we briefly illustrate the effects of sweepstakes reproduction on two population genetic summary statistics. Under a model of sweepstakes reproduction where the variable Ψ describes the proportion of individuals in a generation that are the offspring of a single individual in the previous generation, we summarize in Figure 1 the SFS

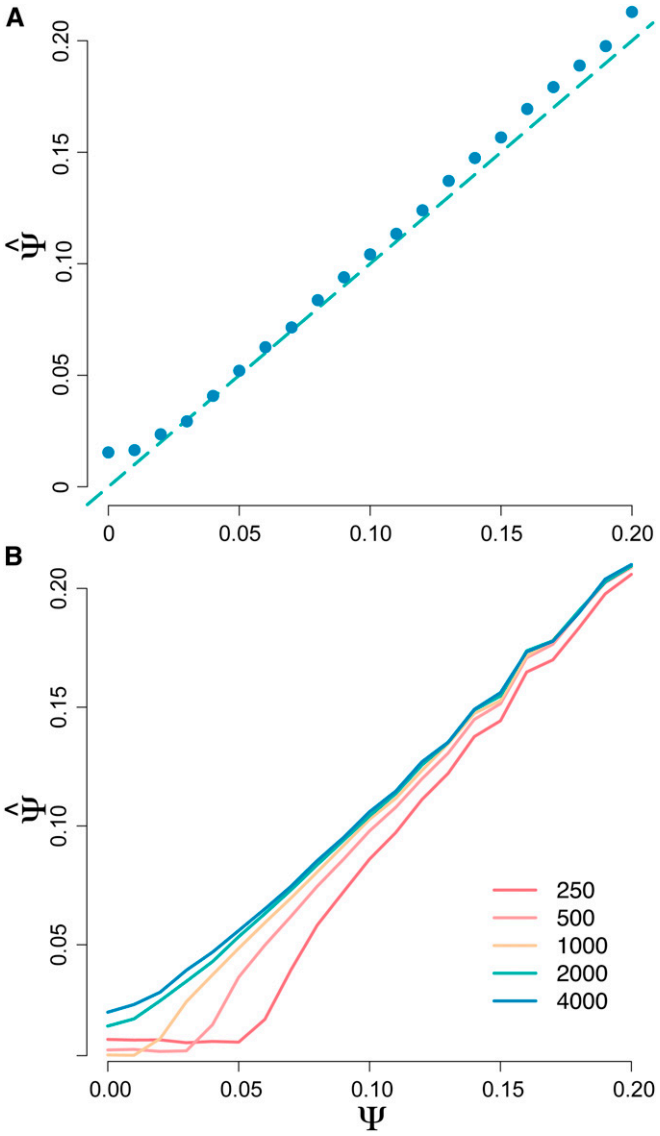

Figure 3 (A) Estimation of Ψ by MMC-ABC. Estimates were averaged over 100 replicate populations of size $N = 1000$ with values of Ψ and N drawn from priors $\sim U[0, 0.3]$ and $\sim U[250, 4000]$, demonstrating the robustness of MMC-ABC to mis-specification of census size. True values of Ψ are indicated by the dashed line. (B) Estimation of Ψ by MMC-ABC, with estimates averaged over 100 replicate populations of size $N = 1000$ with either the correct value of N or an incorrect value of N ($N \in \{250, 500, 1000, 2000, 4000\}$) specified, demonstrating the nonlinear relationship between N and N_e under the Ψ-coalescent, with mis-specification of N having little effect on the accurate estimation of Ψ when Ψ is large.

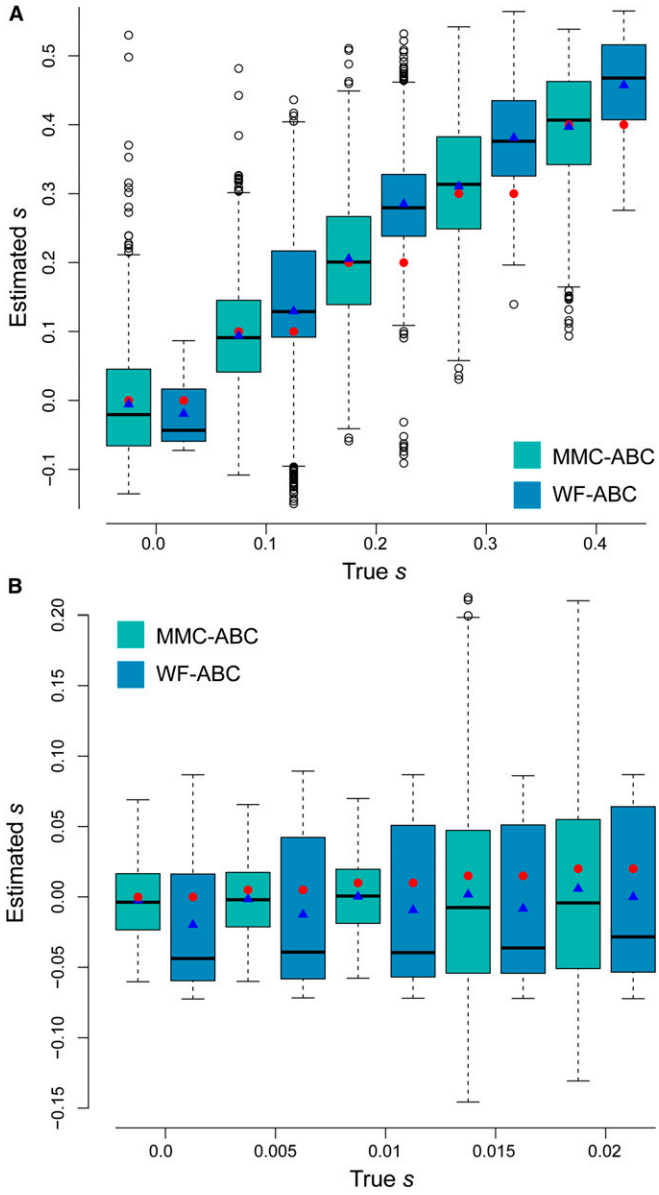

Figure 4 (A) Estimation of s by MMC-ABC and WF-ABC for 1000 sites under selection for $s \in \{0, 0.1, 0.2, 0.3, 0.4\}$ with the true values of $N = 1000$ and $\Psi = 0.1$ provided to MMC-ABC and the true value of N provided to WF-ABC. Results presented in a standard box plot with the box as the first, second, and third quartiles, and the whiskers as the lowest and highest datum within the 1.5 interquartile range of the lower and upper quartiles, respectively. Red circles indicate the true value of s, and blue triangles indicate the sample mean. (B) Estimation for $s \in \{0, 0.005, 0.01, 0.015, 0.02\}$ with the same conditions as above. WF-ABC tends to underestimate s for neutral alleles and overestimate s under strong positive selection under sweepstakes reproduction.

and Tajima's D, averaged over 100 replicate populations of size $N = 1000$ under a broad range of Ψ.

The primary points of note are that, under equilibrium neutrality, nonzero values of Ψ skew the SFS toward an excess of singletons and high-frequency variants, and that Tajima's D is negatively correlated with Ψ. The reader may note that Tajima's D is slightly negative for $\Psi = 0$, as should be expected given that Tajima's D is a biased summary of the SFS dependent upon the recombination rate (Thornton 2005).

Hence, it is clear that failure to account for offspring skew may result in mis-inference, as null model expectations differ strongly from those of the WF model. In the following sections, we will demonstrate that accounting for sweepstakes reproduction simply as a decrease in N_e (as in WF-ABC) results in highly biased estimates of selection. However, explicitly incorporating the underlying processes of MMC events can

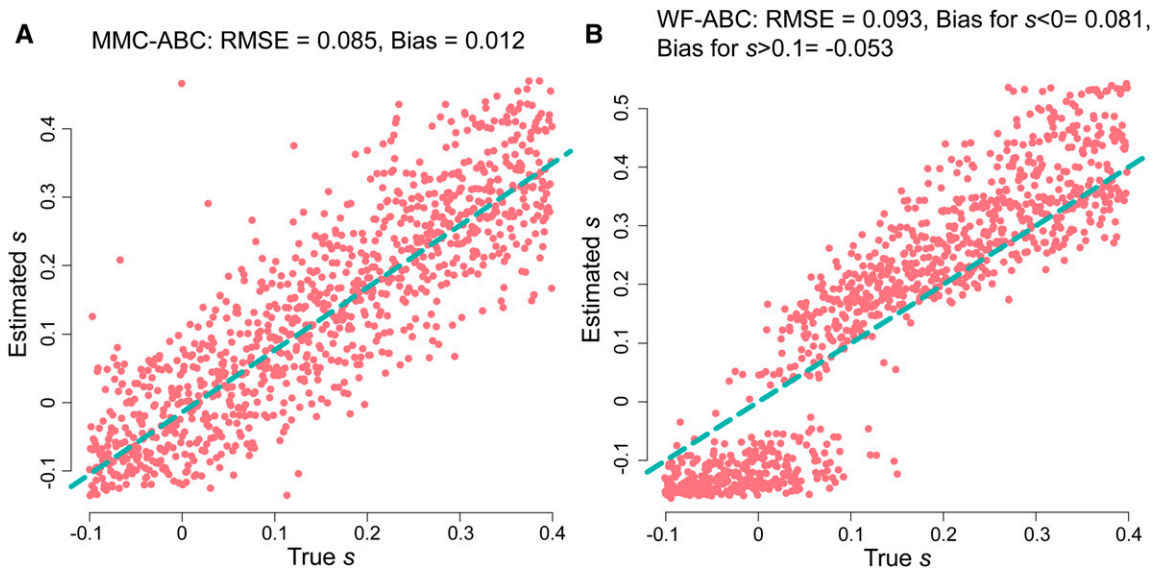

Figure 5 Estimation of s by MMC-ABC (A) and WF-ABC (B). Dots represent 1000 sites under selection with the true s ranging from -0.1 to 0.4. For each site, we estimated s from 10,000 simulations with a uniform prior $\sim U[-0.2, 0.6]$. MMC-ABC was provided with the correct values of $\Psi = 0.1$ and N. As shown, MMC-ABC is a relatively unbiased estimator of s under offspring skew, while WF-ABC strongly overestimates s for positively selected sites, and underestimates s for neutral and negatively selected sites.

correctly adjust for their effects and yield accurate and precise estimates of s from time-series data.

Estimation of Ψ with MMC-ABC

In Step 1 of MMC-ABC, the trajectories of all sites included in the data are used to estimate N_e using the unbiased estimator of Jorde and Ryman (2007). In the case where the census size or harmonic mean of the population size across all time points is known, as is often the case in experimental lineages, populations of census size N with sweepstakes parameter Ψ drawn from its prior and mutational frequencies matching those at the first time point of the data are simulated for the same number of generations as the original data. The best 1% of simulations are retained to generate a posterior for Ψ.

MMC-ABC is able to accurately infer Ψ over a broad parameter space. Figure 2 shows the mean of the posterior distribution of Ψ averaged over 1000 replicate populations each at $\Psi \in \{0, 0.01, \ldots 0.25\}$ in the case where the correct value of N is specified. These illustrative parameter values were chosen to match general features of common viral experimental evolution studies (*e.g.*, Foll *et al.* 2014a; Bank *et al.* 2016; Ormond *et al.* 2017).

Although in cases of experimental evolution precise measurements of N may be available to inform the prior used in Step 1 of MMC-ABC, knowledge of the size of the population in question may not be available. Therefore, we determined the power of MMC-ABC to accurately estimate Ψ in the absence of knowledge about the true value of N. In this case, both N and Ψ are drawn from priors, and MMC-ABC generates a joint posterior for the two parameters. We found that MMC-ABC is a good estimator of Ψ even when a large, uniform prior is used (Figure 3). MMC-ABC likewise performs well in the case where a single, incorrect value of N is specified, particularly for high values of Ψ, at which $\hat{\Psi}$ converges at the true value due to the nonlinear relationship between Ψ and N_e (Figure 3).

We assessed the performance of MMC-ABC over a range of data types, including cases with 5, 11, or 21 time points over a span of 100 generations, as well as for sample sizes of 25, 100, and 250 for populations of $N = 1000$ at $\Psi \in \{0, 0.05, 0.1, 0.15, 0.2\}$ (Supplemental Material, Figures S1 and S2). As expected, the estimation of Ψ improves with larger sample sizes and more densely sampled time points. However, MMC-ABC remains a good estimator of Ψ, even with as few as five time points or a sample size of 25.

To assess the ability of MMC-ABC to perform accurate inference from time-series data including ancient samples, we estimated Ψ for 500 replicate simulated populations with data from 10 time points spaced 10 generations apart, and a single time point 500 generations in the past. The true value of Ψ was 0.1 for all populations, with roughly one third of sites being nonzero at the ancient time point. The average value of Ψ estimated across all replicate populations was 0.10 (Figure S3).

Estimation of site-specific selection coefficients

In the second step of MMC-ABC, the posterior distributions of N and Ψ obtained in Step 1 are used to simulate 10,000 trajectories at each site x_i $(i = 1, \ldots, L)$ with the alleles introduced in the population at the initial frequency $x_{i,1}$ provided in the data. The best 1% of simulations are retained to generate a posterior for s. Foll *et al.* (2014b) previously demonstrated WF-ABC to be a good estimator of genome-wide N_e and site-specific selection coefficients in populations well-described by the Kingman coalescent. The performance of WF-ABC matched or exceeded that of similar methods. Therefore, we restrict our comparison of the performance

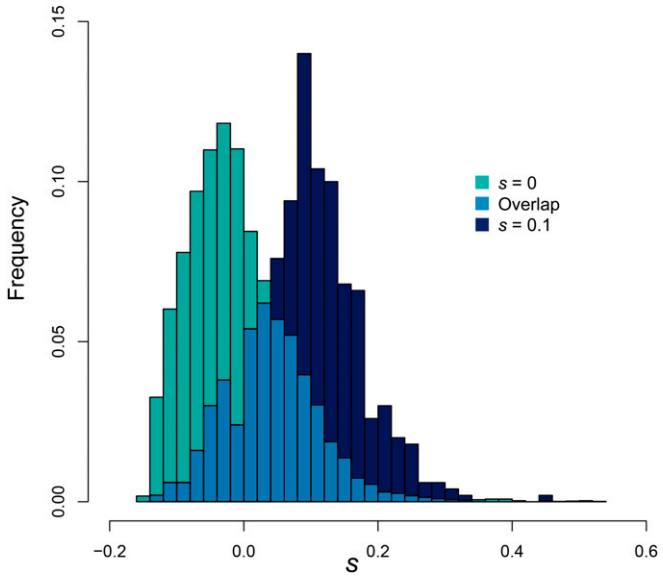

Figure 6 Estimation of s by MMC-ABC for 9500 neutral sites and 500 sites for which $s = 0.1$ with $N = 1000$ and $\Psi = 0.1$, with N estimated over a uniform prior $\sim U[250, 2000]$, Ψ estimated from the prior $\sim U[0, 0.3]$, and s estimated over $\sim U[-0.2, 0.6]$. Note that we display the relative frequencies for estimated values of s for each class of mutation, for which there were unequal numbers of total sites. These results demonstrate the ability of MMC-ABC to jointly and accurately estimate N, Ψ, and s from genomic data, even when a large number of sites are under positive selection.

of MMC-ABC to that of WF-ABC. For a detailed comparison of the performance of WF-ABC with that of other methods, including those of Bollback *et al.* (2008), Malaspinas *et al.* (2012), and Mathieson and McVean (2013), see the results of Foll *et al.* (2014b).

To compare the ability of MMC-ABC and WF-ABC to infer site-specific selection coefficients, we estimated s for 1000 trajectories simulated under the Ψ-coalescent with $N = 1000$ and $\Psi = 0.1$ for $s \in \{0, 0.1, 0.2, 0.3, 0.4\}$. All allele trajectories began from a minor allele frequency of 10%. Because the summary statistics used by MMC-ABC and WF-ABC assume that the majority of sites are neutral, we provided true values of N and Ψ to MMC-ABC and of N to WF-ABC in this initial comparison. As shown in Figure 4, MMC-ABC is very accurate at estimating s under recurrent and strong sweepstakes reproduction, while WF-ABC consistently overestimates selection coefficients for positively selected sites and underestimates s for neutral sites. The same is true for small values of $s \in \{0, 0.005, 0.01, 0.015, 0.02\}$. The results of the same analysis performed over a broad range of $0 \leq \Psi \leq 0.2$ demonstrate that the performance of WF-ABC rapidly deteriorates when Ψ is as large as 0.04 (Figure S4).

Estimating s for single trajectories of mutations covering a wider range of true selection coefficients from -0.1 to 0.4, it is evident that MMC-ABC is not only a good estimator under sweepstakes reproduction of selection for sites under positive selection and neutrality, but is also accurate for sites under negative selection. WF-ABC, however, in addition to having a strong bias toward overestimation of s for sites under positive selection, is negatively biased for neutral and negatively selected sites (Figure 5). Inference under the Kingman for organisms that violate the assumption of small variance in progeny distributions is thus prone to serious over- or under-estimation of selection coefficients, while correctly accounting for reproductive skew produces accurate estimates of selective strength. This mis-inference under the WF model results from the acceleration of transit times under sweepstakes reproduction, which are interpreted by WF-ABC as an amplification of positive or negative selection.

As with the estimation of Ψ, we assessed the performance of Part 2 of MMC-ABC over a range of data types, including cases with 5, 11, or 21 time points over a span of 100 generations, as well as for sample sizes of 25, 100, and 250 for populations of $N = 1000$ at $s \in \{0, 0.1, 0.2, 0.3, 0.4\}$ (Figures S5 and S6). Again, as expected, the estimation of s improves with larger sample sizes and more densely sampled time points. However, MMC-ABC is a reasonably good estimator of s even with as few as five time points or a sample size of 25.

Joint estimation of N, Ψ, and s

We simulated trajectories for 9500 neutral loci and 500 selected loci for which $s = 0.1$, under conditions in which $N = 1000$ and $\Psi = 0.1$. MMC-ABC estimated first N and Ψ over priors of $\sim U[250, 2000]$ and $\sim U[0, 0.3]$, respectively, and then estimated s for each site with values of N and Ψ drawn from the joint posterior (Figure 6). The estimated value of $\hat{\Psi} = 0.101$, the mean estimated value of s for neutral sites was -0.008, and for positively selected sites was 0.0924, highlighting the ability of MMC-ABC to jointly estimate the magnitude of skewed offspring distributions and site-specific selection coefficients with accuracy, even when a relatively large proportion (5% in this case) of sites are under strong positive selection.

Results of similar analyses comparing the performance of MMC-ABC and WF-ABC for sets of simulated trajectories of 1900 neutral and 100 selected loci with $\Psi \in \{0, 0.1, 0.2\}$ and $s \in \{0.1, 0.2, 0.3, 0.4\}$ demonstrate good performance of MMC-ABC over a broad range of Ψ, and poor performance of WF-ABC when $\Psi > 0$ (Figures S7 and S8).

These results are notable, given that both recurrent positive selection and skewed progeny distributions can result in coalescent trees dominated by multiple-mergers (Durrett and Schweinsberg 2004, 2005). Different features of the data—resulting from the localized effects of selection and the genome-wide effects of sweepstakes reproduction—allow us to disentangle the MMC behavior of neutral offspring skew from that of non-neutral offspring skew generated by positive selection.

Application to data from influenza A

We applied MMC-ABC to time-series data from the experimental evolution of influenza A. These data were collected under standard culture conditions and during a period of exposure to exponentially increasing concentrations

Table 1 Influenza A virus mutations identified as significantly beneficial by MMC-ABC

	Segment	Position	Substitution type	Initial freq. (%)	Final freq. (%)	WF-ABC s estimates (99% HPDIs)	MMC-ABC s estimates (99% HPDIs)
Control 1	HA	1395	Nonsynonymous	0.03	92.5	0.12 (0.05, 0.19)	0.14 (0.06, 0.21)
Control 2	HA	1211	Nonsynonymous	0.04	100.0	0.20 (0.08, 0.35)	0.23 (0.15, 0.32)
Drug 1	PA	2194	Synonymous	1.4	36.7	0.09 (0.02, 0.17)	0.11 (0.05, 0.18)
	HA	48	Synonymous	0.1	92.3	0.14 (0.06, 0.27)	0.16 (0.05, 0.24)
	HA	1395	Nonsynonymous	0.06	99.9	0.22 (0.08, 0.34)	0.27 (0.13, 0.42)
	NA	582	Synonymous	0.02	98.3	0.29 (0.15, 0.45)	0.43 (0.28, 0.56)
	NA	823[a]	Nonsynonymous	0.04	99.5	0.15 (0.06, 0.24)	0.18 (0.08, 0.28)
Drug 2	NA	823[a]	Nonsynonymous	0.04	90.3	0.27 (0.12, 0.48)	0.26 (0.13, 0.42)

[a] Sites of known drug-resistance mutations

of the drug oseltamivir (Foll *et al.* 2014a; Renzette *et al.* 2014).

The data consist of time-sampled minor allele frequencies for two control lineages and two drug-selected lineages. Using WF-ABC, Foll *et al.* (2014a) previously estimated the effective population sizes of the control and selected populations to be 176 and 226, respectively, with values of N_e derived from the harmonic means of the population sizes during passaging being 737 and 696, respectively. They hypothesized that the discrepancies in measurements of N_e were likely due to the large variance in viral burst sizes, yielding skewed offspring distributions. These experimentally evolved populations are therefore well-suited to the application of MMC-ABC.

We first obtained estimates of Ψ for each population, using the harmonic population size means as a prior for N. We then obtained posterior distributions of s for all mutations segregating in at least two time points and with a minimum frequency of 2.5% for at least one time point. We define Bayesian "P-values" for s as $P(s < 0|x)$ and consider a trajectory to be "significant at level p" if its equal-tailed $100(1-p)\%$ posterior interval excludes zero (Beaumont and Balding 2004).

The mean posterior estimate of Ψ for the two control lines was 0.067, and the mean value of Ψ across both drug-treatment lines was 0.084. MMC-ABC recovered two of the same six control line mutations and 7 of the 15 mutations from the drug selection lines identified by Foll *et al.* (2014a) as being beneficial at the level $p = 0.01$ (Table 1, summarizing all 8 mutations significant under MMC-ABC and 8 of the 20 significant under WF-ABC, sites of known drug-resistance mutations shown in bold font). The mutations of significant beneficial effect under WF-ABC had an average effect of $s = 0.1$ for control line mutations and $s = 0.13$ for drug-selection mutations. The same sets of mutations (including those that did not achieve significance under MMC-ABC) had average effects of $s = 0.11$ and $s = 0.17$, as estimated by MMC-ABC.

The beneficial mutations identified in the control lines are likely adaptations to the MDCK cells used in serial passaging. One mutation at nucleotide position 1395 of the hemagglutinin segment, which rose to high frequency in the first control and drug lines, has been widely observed across influenza strains and is a common adaptation to tissue culture (Daniels *et al.* 1985; Reed *et al.* 2009; Foll *et al.* 2014a). Another mutation, which reached high frequency in the second control line, has likewise been associated with adaptation to culture conditions (Lin *et al.* 1997; Ilyushina *et al.* 2007). Notably, the mutation at position 823 of the neuraminidase segment (identified as H275Y under the N2 numbering system) achieved high frequency in both drug lineages, and is a well-documented resistance mutation for oseltamivir (Sha and Luo 1997; Arzt *et al.* 2001; Collins *et al.* 2008).

Six of the eight synonymous mutations found to be significantly beneficial by WF-ABC were not significantly beneficial under MMC-ABC. By estimating an appropriate neutral null model under the Ψ-coalescent, we reduced the list of candidate resistance mutations, thus likely minimizing the rate of false positives and excluding many hitchhiking mutations (as the synonymous sites are likely to be). This is supported by an analysis of the proportion of neutral and positively selected mutations that are significantly beneficial or deleterious under WF-ABC and MMC-ABC, which demonstrated a high false-positive rate of neutral mutations classified as strongly beneficial by WF-ABC under moderately strong offspring skew (Table S2). Several experimentally validated mutations known to improve either infectivity in tissue culture or resistance to oseltamivir were retained under MMC-ABC, as were a handful of other potential candidate resistance mutations.

Conclusions

The revolution in sequencing technology has increased the availability of time-series polymorphism data by orders of magnitude, but the utility of such data relies upon the derivation and development of appropriate inference methodologies. The neutral biology of large swaths of the tree of life renders the most common class of method based on the Kingman coalescent of questionable use. We have demonstrated here that performing inference under the assumptions of the Wright-Fisher model and the Kingman coalescent leads to an incorrect understanding of both population size and selection coefficients in such organisms. Matuszewski *et al.* (2018) have also shown this to be true for the demographic history of the population. Fortunately, the theoretical details are in place to develop similar inference of demography and

selection under biologically appropriate alternative coalescent models (Wakeley 2013).

We have shown that MMC-ABC is able to jointly estimate N, Ψ, and site-specific selection coefficients accurately, even under high levels of reproductive skew and with an unknown population size. Notably, we were able to distinguish selection-induced offspring skew from skew originating from the neutral reproductive biology of populations, largely due to the genome-wide scale of MMC events relative to the localized effects of selection. We were also able to differentiate drift-induced effects imposed by small population sizes from those induced by sweepstakes reproduction events.

Very little is known regarding the extent of progeny skew across groups of viruses, bacteria, and plants, or the extent of skew artificially induced by domestication and cultivation. However, this work demonstrates that, at least with time-sampled allele frequency data, such inference is now possible. Moreover, our method will allow for the construction of much more accurate neutral null models in these organisms, which will greatly reduce false-positive rates in scans for selection, provide a more accurate picture of demographic history, and reveal previously hidden details regarding variance in offspring number.

Acknowledgments

We thank Stefan Laurent for helpful discussion. This work was funded by grants from the European Research Council, the Swiss National Science Foundation, and the U.S. Department of Defense to J.D.J.

Literature Cited

Arzt, S., F. Baudin, A. Barge, P. Timmins, W. P. Burmeister *et al.*, 2001 Combined results from solution studies on intact influenza virus M1 protein and from a new crystal form of its N-terminal domain show that M1 is an elongated monomer. Virology 279: 439–446. https://doi.org/10.1006/viro.2000.0727

Bank, C., N. Renzette, P. Liu, S. Matuszewski, H. Shim *et al.*, 2016 An experimental evaluation of drug-induced mutational meltdown as an antiviral treatment strategy. Evolution 70: 2470–2484. https://doi.org/10.1111/evo.13041

Bazin, E., K. J. Dawson, and M. A. Beaumont, 2010 Likelihood-free inference of a population structure and local adaptation in a Bayesian hierarchical model. Genetics 185: 587–602. https://doi.org/10.1534/genetics.109.112391

Beaumont, M., and D. Balding, 2004 Identifying adaptive genetic divergence among population from genome scans. Mol. Ecol. 13: 969–980. https://doi.org/10.1111/j.1365-294X.2004.02125.x

Bhaskar, A., A. G. Clark, and Y. S. Song, 2014 Distortion of genealogical properties when the sample is very large. Proc. Natl. Acad. Sci. USA 111: 2385–2390. https://doi.org/10.1073/pnas.1322709111

Birkner, M., J. Blath, and B. Eldon, 2013 Statistical properties of the site-frequency spectrum associated with Λ-coalescents. Genetics 195: 1037–1053. https://doi.org/10.1534/genetics.113.156612

Blath, J., M. C. Cronjäger, B. Eldon, and M. Hammer, 2016 The site-frequency spectrum associated with Ξ-coalescents. Theor. Popul. Biol. 110: 36–50. https://doi.org/10.1016/j.tpb.2016.04.002

Bollback, J. P., T. L. York, and R. Nielsen, 2008 Estimation of 2Nes from temporal allele frequency data. Genetics 179: 497–502. https://doi.org/10.1534/genetics.107.085019

Bolthausen, E., and A. Sznitman, 1998 On Ruelle's probability cascades and an abstract cavity method. Commun. Math. Phys. 197: 247–276. https://doi.org/10.1007/s002200050450

Collins, P. J., L. F. Haire, Y. P. Lin, J. Liu, R. J. Russell *et al.*, 2008 Crystal structures of oseltamivir-resistant influenza virus neuraminidase mutants. Nature 453: 1258–1261. https://doi.org/10.1038/nature06956

Daniels, R. S., J. C. Downie, A. J. Hay, M. Knossow, J. J. Skehel *et al.*, 1985 Fusion mutants of the influenza virus hemagglutinin glycoprotein. Cell 40: 431–439. https://doi.org/10.1016/0092-8674(85)90157-6

Der, R., C. L. Epstein, and J. B. Plotkin, 2011 Generalized population models and the nature of genetic drift. Theor. Popul. Biol. 80: 80–99. https://doi.org/10.1016/j.tpb.2011.06.004

Donnelly, P., and T. G. Kurtz, 1999 Particle representations for measure-valued population models. Ann. Probab. 27: 166–205. https://doi.org/10.1214/aop/1022677258

Durrett, R., and J. Schweinsberg, 2004 Approximating selective sweeps. Theor. Popul. Biol. 66: 129–138. https://doi.org/10.1016/j.tpb.2004.04.002

Durrett, R., and J. Schweinsberg, 2005 A coalescent model for the effect of advantageous mutations on the genealogy of a population. Stochastic Process. Appl. 115: 1628–1657. https://doi.org/10.1016/j.spa.2005.04.009

Eldon, B., and J. Wakeley, 2006 Coalescent processes when the distribution of offspring number among individuals is highly skewed. Genetics 172: 2621–2633. https://doi.org/10.1534/genetics.105.052175

Eldon, B., and J. Wakeley, 2008 Linkage disequilibrium under skewed offspring distribution among individuals in a population. Genetics 178: 1517–1532. https://doi.org/10.1534/genetics.107.075200

Eldon, B., and J. Wakeley, 2009 Coalescence times and FST under a skewed offspring distribution among individuals in a population. Genetics 181: 615–629. https://doi.org/10.1534/genetics.108.094342

Ewens, W. J., 2004 *Mathematical Population Genetics: Theoretical Introduction*. Springer, New York. https://doi.org/10.1007/978-0-387-21822-9

Ferrer-Admetlla, A., C. Leuenberger, J. D. Jensen, and D. Wegmann, 2016 An approximate Markov model for the Wright-Fisher diffusion and its application to time series data. Genetics 203: 831–846. https://doi.org/10.1534/genetics.115.184598

Fisher, R. A., 1930 *The Genetical Theory of Natural Selection*. Oxford University Press, Oxford. https://doi.org/10.5962/bhl.title.27468

Foll, M., Y.-P. Poh, N. Renzette, A. Ferrer-Admetlla, C. Bank *et al.*, 2014a Influenza virus drug resistance: a time-sampled population genetics perspective. PLoS Genet. 10: e1004185. https://doi.org/10.1371/journal.pgen.1004185

Foll, M., H. Shim, and J. D. Jensen, 2014b WFABC: a Wright-Fisher ABC-based approach for inferring effective population sizes and selection coefficients from time-sampled data. Mol. Ecol. Resour. 15: 87–98. https://doi.org/10.1111/1755-0998.12280

Hallatschek, O., 2018 Selection-like biases emerge in population models with recurrent jackpot events. Genetics 210: 1053–1073. https://doi.org/10.1534/genetics.118.301516

Haller, B. C., and P. W. Messer, 2019 Slim 3: forward genetic simulations beyond the Wright-Fisher model. Mol. Biol. Evol. https://doi.org/10.1093/molbev/msy228

Huillet, T., and M. Möhle, 2011 Population genetics models with skewed fertilities: a forward and backward analysis. Stoch. Models 27: 521–554. https://doi.org/10.1080/15326349.2011.593411

Ilyushina, N. A., E. A. Govorkova, C. J. Russell, E. Hoffmann, and R. G. Webster, 2007 Contribution of H7 haemagglutinin to amantadine resistance and infectivity of influenza virus. J. Gen. Virol. 88: 1266–1274. https://doi.org/10.1099/vir.0.82256-0

Irwin, K., S. Laurent, S. Matuszewski, S. Vuilleumier, L. Ormond *et al.*, 2016 On the importance of skewed offspring distributions and background selection in virus population genetics. Heredity 117: 393–399. https://doi.org/10.1038/hdy.2016.58

Jorde, P. E., and N. Ryman, 2007 Unbiased estimator for genetic drift and effective population size. Genetics 177: 927–935. https://doi.org/10.1534/genetics.107.075481

Kingman, J. F. C., 1982 The coalescent. Stochastic Process. Appl. 13: 235–248. https://doi.org/10.1016/0304-4149(82)90011-4

Lacerda, M., and C. Seoighe, 2014 Population genetics inference of longitudinally-sampled mutants under strong selection. Genetics 198: 1237–1250. https://doi.org/10.1534/genetics.114.167957

Lin, Y. P., S. A. Wharton, J. Martin, J. J. Skehel, D. C. Wiley *et al.*, 1997 Adaptation of egg-grown and transfectant influenza viruses for growth in mammalian cells: selection of hemagglutinin mutants with elevated pH of membrane fusion. Virology 233: 402–410. https://doi.org/10.1006/viro.1997.8626

Malaspinas, A.-S., O. Malaspinas, S. N. Evans, and M. Slatkin, 2012 Estimating allele age and selection coefficient from time-serial data. Genetics 192: 599–607. https://doi.org/10.1534/genetics.112.140939

Mathieson, I., and G. McVean, 2013 Estimating selection coefficients in spatially structured populations from time series data of allele frequencies. Genetics 193: 973–984. https://doi.org/10.1534/genetics.112.147611

Matuszewski, S., M. E. Hildebrandt, G. Achaz, and J. D. Jensen, 2018 Coalescent processes with skewed offspring distributions and nonequilibrium demography. Genetics 208: 323–338. https://doi.org/10.1534/genetics.117.300499

Möhle, M., 1998 Robustness results for the coalescent. J. Appl. Probab. 35: 438–447. https://doi.org/10.1239/jap/1032192859

Möhle, M., 1999 Weak convergence to the coalescent in neutral population models. J. Appl. Probab. 36: 446–460. https://doi.org/10.1239/jap/1032374464

Möhle, M., and S. Sagitov, 2001 A classification of coalescent processes for haploid exchangeable population models. Ann. Probab. 29: 1547–1562.

Neher, R. A., and O. Hallatschek, 2013 Genealogies of rapidly adapting populations. Proc. Natl. Acad. Sci. USA 110: 437–442. https://doi.org/10.1073/pnas.1213113110

Neuhauser, C., and S. M. Krone, 1997 The genealogy of samples in models with selection. Genetics 145: 519–534.

Nordborg, M., 1997 Structured coalescent processes on different time scales. Genetics 146: 1501–1514.

Ormond, L., P. Liu, S. Matuszewski, N. Renzette, C. Bank *et al.*, 2017 The combined effect of oseltamivir and favipiravir on influenza A virus evolution. Genome Biol. Evol. 9: 1913–1924. https://doi.org/10.1093/gbe/evx138

Pitman, J., 1999 Coalescents with multiple collisions. Ann. Probab. 27: 1870–1902. https://doi.org/10.1214/aop/1022874819

Reed, M. L., H. Yen, R. M. DuBois, O. A. Bridges, R. Salomon *et al.*, 2009 Amino acid residues in the fusion peptide pocket regulate the pH of activation of the H5N1 influenza virus hemagglutinin protein. J. Virol. 83: 3568–3580. https://doi.org/10.1128/JVI.02238-08

Renzette, N., D. R. Caffrey, K. B. Zeldovich, P. Liu, G. R. Gallagher *et al.*, 2014 Evolution of the influenza A virus genome during development of oseltamivir resistance *in vitro*. J. Virol. 88: 272–281. https://doi.org/10.1128/JVI.01067-13

Rousseau, E., B. Moury, L. Malleret, R. Senoussi, A. Palloix *et al.*, 2017 Estimating virus effective population size and selection without neutral markers. PLoS Pathog. 13: e1006702. https://doi.org/10.1371/journal.ppat.1006702

Sagitov, S., 1999 The general coalescent with asynchronous mergers of ancestral lines. J. Appl. Probab. 36: 1116–1125. https://doi.org/10.1239/jap/1032374759

Schraiber, J. G., S. N. Evans, and M. Slatkin, 2016 Bayesian inference of natural selection from allele frequency time series. Genetics 203: 493–511. https://doi.org/10.1534/genetics.116.187278

Schweinsberg, J., 2000 Coalescents with simultaneous multiple collisions. Electron. J. Probab. 5: 50. https://doi.org/10.1214/EJP.v5-68

Schweinsberg, J., 2017 Rigorous results for a population model with selection II: genealogy of the population. Electron. J. Probab. 22: 54.

Sha, B., and M. Luo, 1997 Structure of a bifunctional membrane-RNA binding protein, influenza virus matrix protein M1. Nat. Struct. Mol. Biol. 4: 239–244. https://doi.org/10.1038/nsb0397-239

Shim, H., S. Laurent, S. Matuszewski, M. Foll, and J. D. Jensen, 2016 Detecting and quantifying changing selection intensities from time-sampled polymorphism data. G3 (Bethesda) 6: 893–904. https://doi.org/10.1534/g3.115.023200

Steinrücken, M., A. Bhaskar, and Y. S. Song, 2014 A novel spectral method for inferring general diploid selection from time series genetic data. Ann. Appl. Stat. 8: 2203–2222. https://doi.org/10.1214/14-AOAS764

Tellier, A., and C. Lemaire, 2014 Coalescence 2.0: a multiple branching of recent theoretical developments and their applications. Mol. Ecol. 23: 2637–2652. https://doi.org/10.1111/mec.12755

Thornton, K., 2005 Recombination and the properties of Tajima's D in the context of approximate-likelihood calculation. Genetics 171: 2143–2148. https://doi.org/10.1534/genetics.105.043786

Wakeley, J., 2013 Coalescent theory has many new branches. Theor. Popul. Biol. 87: 1–4. https://doi.org/10.1016/j.tpb.2013.06.001

Wakeley, J., and T. Takahashi, 2003 Gene genealogies when the sample size exceeds the effective size of the population. Mol. Biol. Evol. 20: 208–213. https://doi.org/10.1093/molbev/msg024

Wilkinson-Herbots, H. M., 1998 Genealogy and subpopulation differentiation under various models of population structure. J. Math. Biol. 37: 535–585. https://doi.org/10.1007/s002850050140

Wright, S. G., 1931 Evolution in Mendelian populations. Genetics 15: 97–159.

Communicating editor: M. Beaumont