

SCIENTIFIC REPORTS



OPEN

Allele Age Under Non-Classical Assumptions is Clarified by an Exact Computational Markov Chain Approach

Bianca De Sanctis¹, Ivan Krukov² & A. P. Jason de Koning^{1,2,3} 

Determination of the age of an allele based on its population frequency is a well-studied problem in population genetics, for which a variety of approximations have been proposed. We present a new result that, surprisingly, allows the expectation and variance of allele age to be computed exactly (within machine precision) for any finite absorbing Markov chain model in a matter of seconds. This approach makes none of the classical assumptions (e.g., weak selection, reversibility, infinite sites), exploits modern sparse linear algebra techniques, integrates over all sample paths, and is rapidly computable for Wright-Fisher populations up to $N_e = 100,000$. With this approach, we study the joint effect of recurrent mutation, dominance, and selection, and demonstrate new examples of “selective strolls” where the classical symmetry of allele age with respect to selection is violated by weakly selected alleles that are older than neutral alleles at the same frequency. We also show evidence for a strong age imbalance, where rare deleterious alleles are expected to be substantially older than advantageous alleles observed at the same frequency when population-scaled mutation rates are large. These results highlight the under-appreciated utility of computational methods for the direct analysis of Markov chain models in population genetics.

Allele age is generally defined as the duration of time a mutant allele has been segregating in a population. The problem of calculating the expected age of an allele given its current population frequency is an important problem in population genomics (e.g., ref.¹) with a long history of theoretical investigations (e.g., refs^{2–6}; reviewed in ref.⁷). One reason that allele age remains an important problem is that the effects of selection and age can be highly confounded in terms of their influence on population frequency. That is, an allele may be at low frequency because it is deleterious or simply because it is young. Inferences about the fitness effects of segregating polymorphisms must therefore make some consideration of allele age, either explicitly or implicitly, and methods for inferring fitness impacts based on allele ages have even been proposed⁸.

The first theoretical analysis of allele age was developed by Kimura and Ohta³ using a continuous-time diffusion approximation to the age of a neutral allele in a finite population. Later work added consideration of selection⁹, yielding the well-known result that allele age is expected to be symmetric with respect to the direction of selection, and that neutral alleles are expected to be older than selected alleles observed at the same frequency (the “Maruyama effect” hereafter). Recently, an interesting exception to these classical results has been pointed out^{10–12}. Mafessoni *et al.*¹² showed that weakly selected rare alleles are expected to be about 5% older than neutral alleles observed at the same frequency, when heterozygote fitness is non-additive. This phenomenon appears to be an example of a more general behaviour recently termed ‘stochastic slowdown’¹⁰, where weak selection counter-intuitively prolongs, rather than shortens, the average time to absorption. It is important to understand the generality of these findings, since, as Mafessoni *et al.*¹² point out, many new mutations arising in a population

¹University of Calgary, Cumming School of Medicine, Dept. of Biochemistry and Molecular Biology, Calgary, Alberta, Canada. ²University of Calgary, Cumming School of Medicine, Department of Biochemistry and Molecular Biology Graduate Program (Bioinformatics stream), Calgary, Alberta, Canada. ³University of Calgary, Cumming School of Medicine, Department of Medical Genetics, and Alberta Children’s Hospital Research Institute, Calgary, Alberta, Canada. Correspondence and requests for materials should be addressed to A.P.J.d. (email: jason.dekoning@ucalgary.ca)

are expected to be recessive and weakly deleterious, and it is conceivable that this slowdown effect could thereby mislead attempts to make inferences about natural selection.

Previous investigations of allele age, and classical approaches in population genetics more generally, have required that mutation rates are assumed to be so slow that no additional mutations can occur during the segregation of an initial variant (implying that the population-scaled mutation rate, θ , is very small or ≈ 0). However, cases where this assumption is violated in nature are increasingly being reported, and it is likely in such cases that classical population genetic theory will be unreliable at best¹³. While in most eukaryotes, θ is estimated to be $\ll 0.05$, several examples of so-called hyperdiverse eukaryotes are known with $\hat{\theta}$ between 0.05 and 0.15¹⁴. In bacteria, it is not uncommon for estimates of θ to be at the high end of this range or significantly larger. For example, Sung *et al.*¹⁵ reported average estimates taken from the literature of $\theta = 0.15$ for *Helicobacter pylori* and 0.12 for *Salmonella enterica*, both of significant biomedical interest. Hughes *et al.*¹⁶ also reported $\hat{\theta}$ in *Pseudomonas syringae* to be 0.55. In addition, θ in some organisms including viruses and pathogens has been estimated to be much larger, by a variety of analytical methods, with estimates often exceeding 1. For example, θ in HIV-1 has been estimated to be between 10 and 369¹⁷ in one study, and >1 using the effective population size estimated by Pennings *et al.*^{18,19} together with mutation rates from other studies; similarly, θ in macaque monkeys infected with RT-SHIV (an engineered simian immunodeficiency virus encoding human HIV-1 reverse transcriptase) has been estimated to be greater than one²⁰. Other arguments that classical assumptions about θ may be violated in nature have also been recently put forward. For example, Messer and Petrov²¹ have highlighted that most known cases of molecular adaptation across diverse organisms show signatures of soft selective sweeps (but see ref.²²), where adaptive alleles have multiple origins either by recurrent mutation or migration. These findings are potentially unexpected if evolution is strongly mutation-limited and may indicate that the effective population-scaled mutation rate is underestimated in many cases and/or that adaptation may tend to occur during periods of episodically large population size (and thus, high θ)²³. We therefore decided to revisit the problem of calculating allele age based on population frequency under non-classical assumptions, and in particular to examine the impact of large values of θ on the expected age of an allele. For beneficial variants, the values of θ that we consider are expected to produce adaptive fixations that may have either multiple mutational origins or single origins²⁴.

To study the effects of non-classical parameter ranges on allele age, we develop a new exact approach capable of rapidly computing moments of the allele age distribution under any absorbing discrete-time Markov chain model of population genetics. This approach exploits sparsity, parallelism, and modern computational architectures²⁵, and is completely general with respect to the underlying model. It therefore requires none of the classical simplifying assumptions (e.g., weak selection, weak mutation, infinite sites, etc). For the purposes of the present study, we assumed a biallelic diploid Wright-Fisher model²⁶ including bidirectional mutation, selection and dominance. Computationally, our solution mainly relies on back-substitutions using an LU decomposition of a sparse matrix derived from the model's transition matrix, and does not use any matrix-matrix multiplications, which are computationally expensive. This computational implementation is similar to that in ref.²⁵, where we applied sparse matrix techniques to the calculation of population genetic quantities such as the probability of fixation and sojourn times (but not allele age). To the best of our knowledge, this is the first computationally feasible, exact approach for computing allele age (or its moments) to be proposed. Calculation of the expected value and variance of allele age is fast, exact and scales easily to realistic population sizes ($N_e \approx 10^5$ for Wright-Fisher type models, and much larger for Moran models due to their greater sparsity; see Discussion). We have implemented this method in our software package Wright-Fisher Exact Solver, WFES²⁵ (available at <https://github.com/dekoning-lab/wfes/>).

Results

Using the approach outlined above and described fully in the Methods, we considered allele age and related quantities in a biallelic Wright-Fisher model including bidirectional mutation, selection, and dominance. For selection coefficient s and dominance coefficient h , the homozygous wildtype fitness was defined as 1, heterozygote fitness as $1 + sh$, and homozygous mutant fitness as $1 + s$ (following standard definitions²⁶). Bi-directional mutation was modelled in the Wright-Fisher transition matrix²⁶, with extinction and fixation states assumed to be absorbing. This assumption implies a return process such that when a mutant frequency of 1 is attained, the population is returned to a frequency of 0 (equivalent to swapping the labels for the wild-type and mutant states); this allows properties of average trajectories to be easily calculated based on their starting or ending states.

In a biallelic diploid model, each individual may be either wild-type or mutant at each locus and chromosome. We define an “allele” here explicitly as the mutant genotype. Thus, allele age refers to how long the mutant state has been segregating in the population, starting from a population that was monomorphic for the wild-type state. By allowing mutation, we assume that an arbitrary number of new mutations could potentially arise in the population while an initial mutant is segregating, and thus the assumption of shared ancestry of all segregating mutants is not necessarily made. In the context of classical theory it may seem unnatural to consider allele age while including mutation. However, this is because classical theory makes the assumption that mutation cannot be recurrent, while there is no such prohibition in nature. Furthermore, even when θ is large, allele trajectories include long periods of time spent at the boundaries, and it therefore remains reasonable to demarcate the behaviour of such trajectories based on their visits to the boundaries. This may no longer be true when θ is so large that a population always contains all possible alleles ($\theta \gg 1$).

Except where otherwise specified, all results that follow are for a rare allele observed in $x = 10$ copies, sampled from an effective population size of $N_e = 10,000$ diploids. Forward and backward mutation rates were assumed equal. We consider a range of population-scaled mutation rates, $\theta = 4N_e\mu$, between $\theta = 0.0048$ and $\theta = 0.96$, where μ is the mutation rate per site per chromosome, and N_e the effective population size. Results obtained using values of θ that were two orders of magnitude smaller than $\theta = 0.0048$ were largely similar (not shown).

N_e	x	Diffusion	Exact
1,000	10	106.5	103.73
5,000	10	138.29	134.99
10,000	10	152.09	148.56
20,000	10	165.92	162.16
50,000	10	184.23	180.16
1,000	100	630.68	628.65
5,000	100	930.34	927.8
10,000	100	1,064.99	1,062.22
20,000	100	1,201.30	1,198.30
50,000	100	1,382.93	1,379.63
1,000	1,000	2,772.59	2,771.02
5,000	1,000	5,116.86	5,115.03
10,000	1,000	6,306.80	6,304.77
20,000	1,000	7,566.93	7,564.69
50,000	1,000	9,303.37	9,300.85

Table 1. Expected neutral allele age determined by exact computation (this study) and by Kimura and Ohta's³ diffusion approximation. No selection or mutation were assumed in the underlying Wright-Fisher model to ensure that the assumptions of both methods were compatible.

θ	$2N_e s$	h	Simulation		Exact	
			Mean	Std. Dev.	Mean	Std. Dev.
0.01	0	NA	106.67	391.64	106.39	389.57
0.05	0	NA	118.21	433.01	117.99	431.67
0.1	0	NA	135.17	493.47	134.91	491.67
0.5	0	NA	477.75	1,531.31	477.67	1,531.58
0.96	0	NA	3,315.99	7,775.71	3,320.94	7,791.84
0.01	-3	0.0	116.69	449.90	116.42	449.90
0.01	-3	0.5	97.04	319.34	96.86	317.74
0.01	-3	1.0	84.51	238.28	84.38	237.13
0.01	3	0.0	91.62	275.07	91.47	273.91
0.01	3	0.5	96.63	317.56	96.46	316.12
0.01	3	1.0	100.73	369.59	100.53	367.69
0.96	-3	0.0	4,746.56	9,021.43	4,742.61	9,011.70
0.96	-3	0.5	2,994.45	5,757.48	2,990.35	5,746.30
0.96	-3	1.0	1,996.64	3,813.41	1,994.75	3,808.97
0.96	3	0.0	729.07	2,090.27	728.74	2,088.35
0.96	3	0.5	773.22	2,759.72	773.03	2,758.71
0.96	3	1.0	933.51	4,004.62	932.89	4,004.79

Table 2. Representative expected allele age and variance including selection, dominance and mutation determined by simulation and exact computation. A diploid population of $N_e = 1,000$ was assumed with $p = 1$ and $x = 10$.

Validation by comparison to other methods. We first examined the correspondence between expected allele age determined by exact computation with the Wright-Fisher Markov model and the expected allele age approximated using Kimura and Ohta's diffusion approach³. Since Kimura and Ohta's method assumes no selection and no mutation, we ran our computations on a Wright-Fisher model having these same assumptions. Across a range of effective populations sizes and observed allele counts, the methods exhibited close correspondence (Table 1), where Kimura and Ohta's method consistently overestimated allele age by a few generations.

We next validated our method and its implementation by comparing results to allele age simulations that included selection, dominance, and mutation (Table 2). Allele age probability distributions can be approximated by simulation by reversing the direction of time in a Wright-Fisher model that is modified to have the same stationary distribution as the original (forward-time) transition matrix²⁷. "Forward time" simulations of this reversed model can then be performed starting at the observed frequency, $x/(2N_e)$, and running until the beginning of the sample path ($p/(2N_e)$; see Methods for details). Simulations performed in this manner agreed well with the model-based computations across the entire parameter range. Allele-frequency probability distributions approximated by simulation are shown for a subset of cases in Fig. 1.

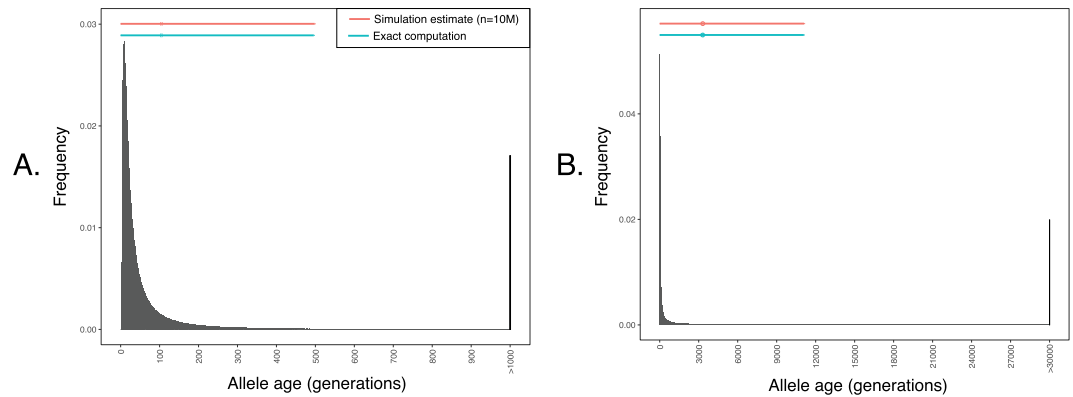


Figure 1. Representative neutral allele age probability distributions determined by simulation. Simulations for $N_e = 1,000$ were performed 10 million times with $s = 0$. As the allele age distributions have very long tails, the undisplayed portion of the tail is accumulated in the final bin. **(A)** $\theta = 0.01$. **(B)** $\theta = 0.96$. Intermediate values of θ are shown in Figs S2 and S3.

	Simulation*	Exact†
θ	Time (sec.)	Time (sec.)
0.01	911.12	0.58
0.05	1,021.56	0.61
0.1	1,190.46	0.60
0.5	5,143.14	0.44
0.96	46,723.40	0.64

Table 3. Representative run-times (wall clock) for parallel computation of neutral allele age. *Simulations were run on 32 cores, parallelized across 10 M replicates. †Exact calculation of the mean and variance of allele age was performed on the same machine using 8 threads. The number of threads in each case was chosen to approximately minimize wall clock time needed by each method.

Computational advantages of the exact approach. Allele age simulations were implemented in C++ and parallelized, so that their runtimes would be reasonably fast (see https://github.com/dekoning-lab/allele_age_simulator/). Simulations were much more time consuming than the direct computation of the moments using our approach (e.g., 15 minutes versus 0.6 seconds for $\theta = 0.01$; Table 3). As θ was increased, the simulations took increasingly more time both because the allele trajectories grew longer on average and because higher mutation rates also increased the variance in the duration of allele age trajectories. For $\theta = 0.96$, running a 10 million replicate simulation over 32 cores took approximately 13 hours. On the other hand, the runtime for the exact matrix method was constant across different mutation rates and averaged about 0.5 seconds. Thus, when moments provide sufficient information, they can be obtained much more efficiently using our exact approach.

Direct demonstration of classical results. Several classical results pertaining to allele age can be directly obtained by examining expected allele age and variance as a function of selection (Fig. 2). It should be emphasized that these plots are neither probability distributions nor estimates. Rather, they are the exact moments of allele age derived directly from the Wright-Fisher model, as explained in the Methods section.

For rare alleles, the expected allele age has a large variance relative to the mean and the mean age is roughly symmetric with respect to the sign of the selection coefficient, with neutral alleles expected to be older than selected alleles (Fig. 2B, leftmost column; the Maruyama effect⁹). The symmetry of allele age with respect to the direction of selection is among the most conspicuous classical findings on allele age, and has been the subject of recent study, where different authors have both supported it using population genomic data⁸ and argued against it using simulations that included linkage²⁸.

Selective strolls and stochastic slowdowns. Recent work^{10–12} has convincingly demonstrated, using primarily simulation and diffusion theory methods, that weakly selected alleles are sometimes expected to be older than neutral alleles observed at the same frequency when fitness in heterozygotes is non-additive. This idea was termed “selective strolls” by Mafessoni *et al.*¹², referring to the observation that selected variants may sometimes persist in a population slightly longer than neutral ones. Here we directly reproduce this effect for rare recessive alleles ($h = 0$), where it can be seen that weakly deleterious alleles are expected to be older than neutral alleles at the same frequency (Fig. 2A, leftmost column), and for dominant alleles ($h = 1$), where it can be seen that weakly advantageous alleles are expected to be older than neutral alleles at the same frequency (Fig. 2C, leftmost column). Consistent with the findings of Mafessoni *et al.*¹², it is apparent that the selective stroll effect size is not very large and is on the order of about 5%.

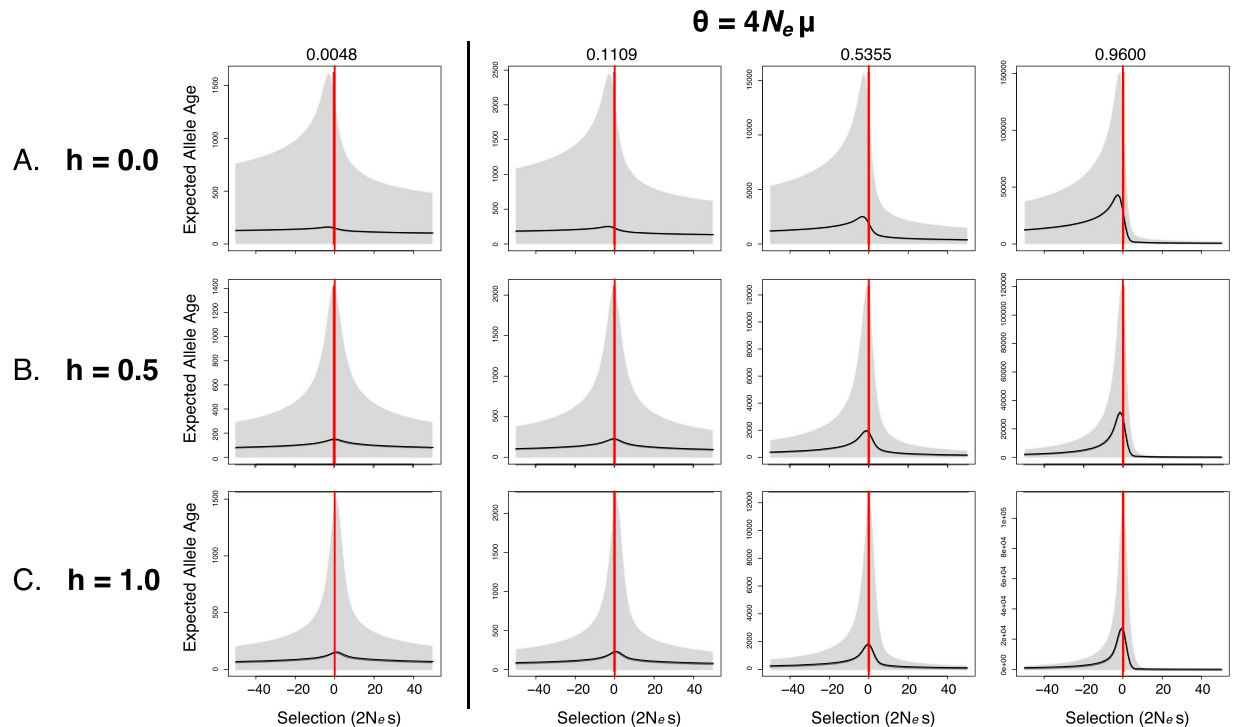


Figure 2. Expected allele age and variance as a function of selection, dominance, and mutation rate. All calculations were made for a rare allele ($x = 10$) assuming $N_e = 10,000$ diploids. When heterozygote fitness is non-additive, weakly selected alleles are expected to be older than neutral alleles observed at the same frequency ((A,C) left). When mutation is weak and heterozygote fitness is additive, allele age is symmetric with respect to the direction of selection ((B) left). When the mutation rate increases, an age imbalance with respect to the direction of selection appears (left to right). Full results over a larger grid of θ can be found in Fig. S1.

Recurrent mutation and age imbalance. Contrary to the Maruyama effect, for population-scaled mutation rates approaching $\theta \approx 1$ the mean allele age becomes strongly asymmetric around $s = 0$ (Fig. 2, c.f. left to right) such that weakly to moderately deleterious alleles can on average be substantially older than advantageous alleles at the same frequency. We refer to this previously unobserved phenomenon as “age imbalance”.

Under age imbalance, slightly deleterious alleles are also expected to generally be older than neutral alleles at the same frequency. This new example of stochastic slowdown is observed even when heterozygote fitness is additive (i.e., with $h = 0.5$). The effect size in this case is substantially larger than for the previously noted slowdowns with small θ (or $\theta = 0$; ref.¹²). For example, expected extinction times for the oldest alleles with $h = 0.5$ are approximately 22.7% longer than for neutral alleles.

Rare recessive alleles ($h = 0$) under recurrent mutation and large θ (Fig. 2, right) experience the same effect but to an even greater degree. Recessivity and fast mutation appear to have a similar and mutually reinforcing effect on both age imbalance and the stochastic slowdown under weak selection. Both selective stroll and age imbalance results appear to be explained primarily by the average time to extinction (Fig. 3, left), which indicates that when mutation rates are bidirectionally fast, weakly deleterious alleles counter-intuitively take longer to go extinct than do advantageous (or neutral) alleles. For $h = 0$ extinction times are even longer for deleterious recessive alleles than for those with $h = 0.5$, but now the expected fixation times also show a similar imbalance with respect to the direction of selection (Fig. 3, c.f. A and B), which accentuates the stochastic slowdown further. Remarkably, the expected time to extinction for the oldest, weakly selected recessive alleles is about 66.9% longer than for neutral alleles (Fig. 2A, left). The same results for $h = 1$ are shown in Fig. 3C, where fixation times are shifted to the right rather than the left, which seems to largely cancel out the stochastic slowdown caused by the left-shifted extinction times.

To help explain Fig. 3, we also calculated the conditional sojourn times for mutants that go to extinction, and compared these to sojourn times for neutral variants (Fig. 4). For deleterious alleles, we see that the time spent at low frequencies increases as we move away from $2N_e s = 0$ until $2N_e s = -2.53$ is reached; the stronger selection is within this parameter range, the more extinction sojourns are dominated by residency at lower frequencies compared to neutral. While this trend is expected since negative selection opposes increases in allele frequency, it is surprising that the net change in non-neutral sojourn times is positive. That is, the increased time at low frequencies surpasses the decreased time spent at high frequencies, resulting in longer sojourns overall. This phenomenon has also been reported for previously noted stochastic slowdowns¹¹.

Allowing the starting number of copies to vary. When population-scaled mutation rates are very high it can become plausible that an originating mutation enters the population in several copies (i.e., that it

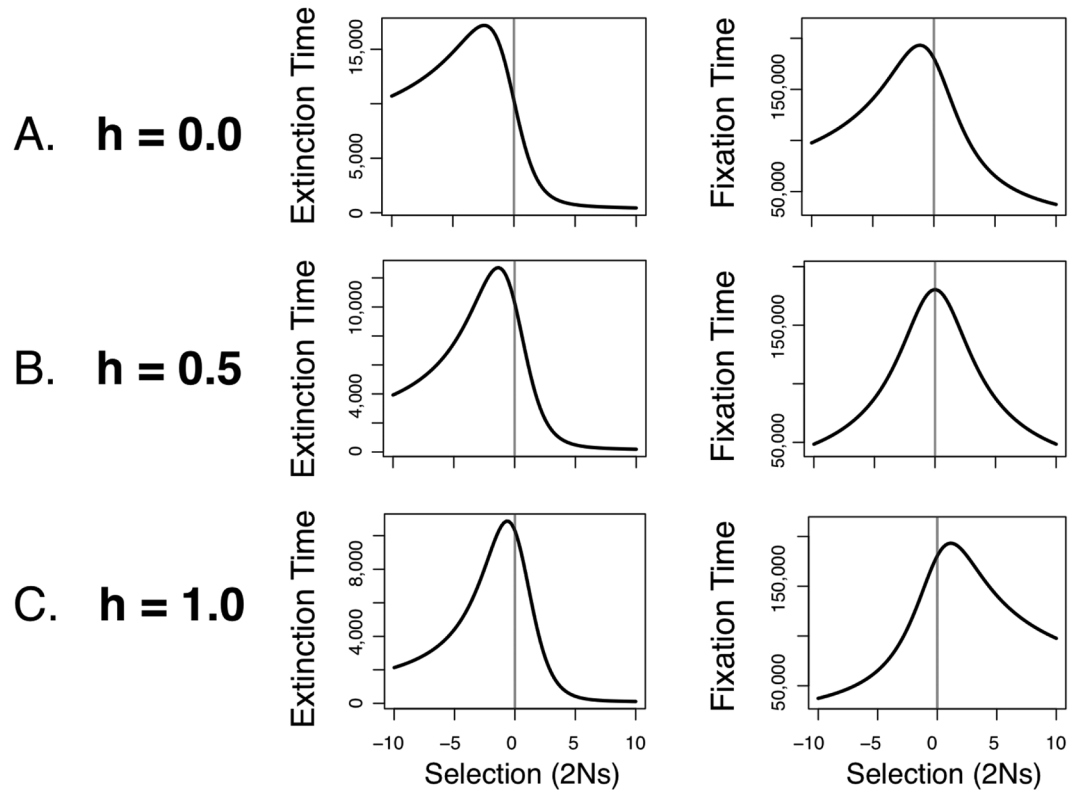


Figure 3. Expected extinction and fixation times when mutation is strong ($\theta = 0.96$) calculated by exact computation with the Wright-Fisher Markov model²⁵. Parameter values used were the same as in Fig. 2. Note the strong asymmetry with respect to the direction of selection, contrary to the classical result of Maruyama and Kimura³².

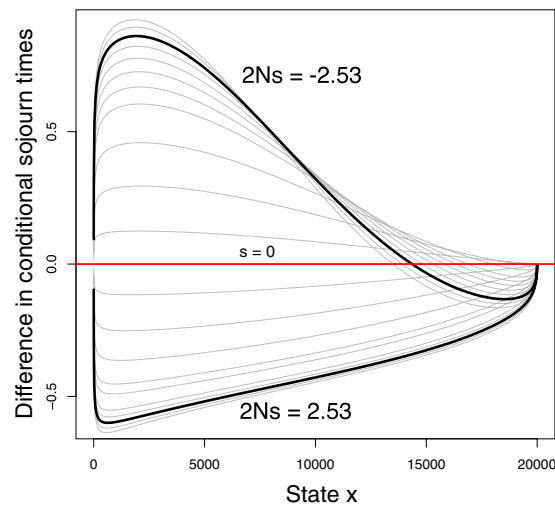


Figure 4. Difference in conditional sojourn times (compared to neutral) for selected alleles going to extinction. Curves span approximately $2N_e s = \{-3, 3\}$; $h = 0$, $\theta = 0.96$. Top: increasing selection against the mutant allele up to the critical point $2N_e s = -2.53$ counter-intuitively increases sojourn times by prolonging residency in low frequency classes. Bottom: increasing selection favouring the mutant allele decreases the length of extinction sojourns. Bold: $2N_e s = \pm 2.53$. The maximum of the extinction time curve in Fig. 3A is at -2.53 .

simultaneously occurs in several individuals). For example, when $\theta = 0.96$, the average number of mutations entering the population per generation is $0.96/2 = 0.48$, so on average there will be a new mutation every two generations. The probability of a population generating multiple copies of the mutant allele in a single generation, assuming mutations are Poisson distributed, is ≈ 0.38 . This may pose problems for any method for calculating

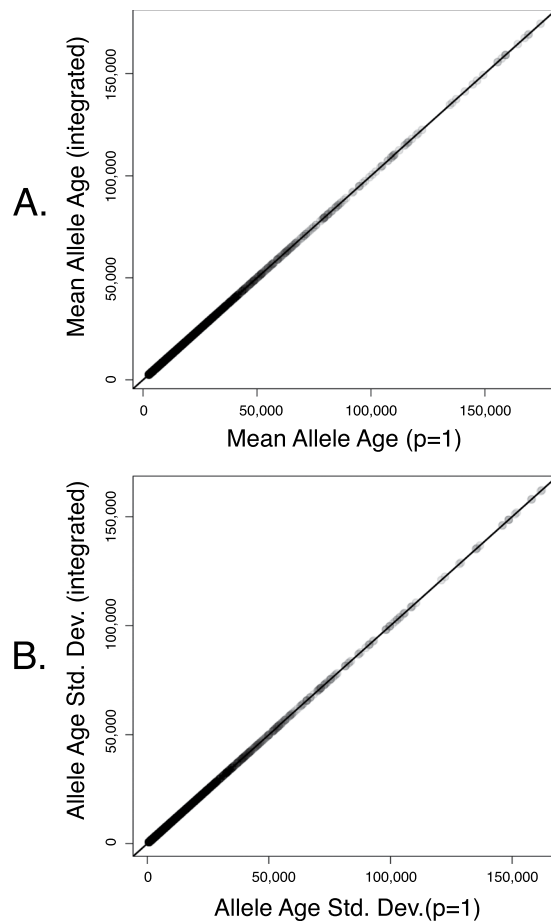


Figure 5. Effect of integrating out uncertainty in p . The integral (summation) was taken to a finite number of terms such that all values with $p \geq 10^{-5}$ were considered. Points represent all parameter combinations considered in Fig. 2.

allele age, since when the likelihood of a population simultaneously generating more than one mutant becomes non-negligible, the starting number of copies, p , should be integrated out.

To integrate over p we consider the probability of starting in p copies, given that $p \sim \text{Poisson}(\lambda = \theta/2)$. This can be easily implemented in our computational procedure starting at Equation 8 by reusing the LU decomposition of $(I - Q)^T$, which does not depend on p (see Methods). Since this decomposition is by far the most computationally expensive operation, the integrated solution is trivially harder than when assuming a single starting copy. In addition, since the probability of large numbers of mutations occurring in the same generation will typically be negligible, we define a threshold ε such that only starting configurations with a probability greater than ε are considered. Below, we assumed $\varepsilon = 10^{-5}$.

In Fig. 5 we show the effect of numerically integrating over p when $\theta = 0.96$ for the range of mutation rates, selection coefficients, and dominance coefficients considered throughout the manuscript. In most cases, the results were identical at better than three to four decimal places, and only began to diverge slightly when θ was very large (i.e., $\theta = 0.96$). It is possible that other statistics of the Markov process might change more than this as a function of p , and thus to be conservative one may choose to always integrate over p (particularly since this adds only seconds to the compute time). However, we conclude that assuming $p = 1$ (as is done by convention in all previous studies of allele age that we are aware of) is likely to introduce no bias unless θ is quite large (i.e., $\gg 1$).

Discussion

Computational population genetics approaches offer the relatively straightforward ability to explore parameter ranges or assumptions that may be inaccessible to classical theory. Usually simulations are used to address scenarios where the assumptions of classical theory may be violated. However, simulations can often be slow, require long runtimes to obtain precise estimates for rare events, and can scale poorly to large populations. An alternative computational approach is to find a class of models whose properties can be interrogated directly, without the need for simulation. For example, Steinruecken *et al.*²⁹ recently showed how the transition density function of biallelic Wright-Fisher diffusions³⁰ could be approximately computed, eliminating the need for a variety of simulations (although allele age has not been considered in this framework). Here we have shown that even the exact computational analysis of biallelic Markov models (including Wright-Fisher models) can be made efficient enough to often eliminate the need for either simulations or diffusion approximations in the first place. Markov

chain models are typically discounted early in the lifecycle of a population genetic investigation in favour of diffusion approximations, since they are widely viewed as impractical to work with due to their large and potentially unwieldy state spaces. Contrariwise, here and elsewhere²⁵, we have shown that judicious computation, sparsity, and parallelism can be together exploited to rather surprising effect, making exact computation under general Markov models not only tractable but capable of generating new insights with ease. Working directly with the underlying Markov models of population genetics has a number of advantages. For example, when strong mutation is included, absorbing boundaries can artificially become inaccessible in a diffusion. There is no corresponding problem when studying the unapproximated Markov chain. In addition, diffusion approaches cannot easily describe behaviours at the absorbing boundaries (but see ref.³¹).

One of the most appealing aspects of this computational population genetics approach is that it is general with respect to underlying modelling assumptions, as long as they can be expressed as a finite absorbing Markov chain. This approach also has several advantages over simulations, including fast runtimes that are relatively insensitive to modelling assumptions (Table 3), and exact results (within machine precision) even for small effects or rare events that would otherwise require long-run, high replicate simulations to study. For a population size of $N_e = 10,000$, exact calculation of the expected allele age and variance, absorption probabilities and times, and conditional sojourn times, takes only about 6.5 seconds using 16 Intel E5-2670 cores (2.60 GHz) in our reference implementation²⁵. Models with greater sparsity are even faster and can scale much better. For example, the same analysis under a comparable Moran model takes only about 0.25 seconds²⁵.

The method proposed for calculating allele age is based on the efficient computation of the moments of the probability distribution of allele ages. It is therefore appropriate to view these quantities not as estimates, but as exact results for a given model. An advantage of this approach is that the expected value of the allele age probability distribution will more often be much closer to the true allele age than would a maximum likelihood estimator, since the age distributions are both highly skewed and very long tailed (see Fig. 1). A potential disadvantage is that we must assume that the true population frequency is known without error. In cases where it is not, error in the observed frequency could be accounted for by computing allele age for a range of population frequencies centred on the observed value.

As shown in Fig. 2, classical allele age results^{3,9,32} can be easily obtained for general population genetic models with our approach. We also reproduced exact representations of recently discovered effects, such as “selective strolls”, which have a smaller effect on expected allele age when mutation rates are low (also see ref.¹²). By exploiting the generality of our approach, we discovered new evidence for a stochastic slowdown that occurs when bidirectional mutation is fast, such that rare, weakly deleterious alleles are expected to be substantially older than neutral alleles. In the most extreme case, average extinction times for the oldest alleles were 22% and 68% longer than for neutral alleles (for $h = 0.5$ and $h = 0$, respectively). Finally, we found that when relaxing the assumption of weak mutation, a large age imbalance arises with respect to selection, such that rare deleterious alleles are expected to be old and rare advantageous alleles very young. This may be explained in part by the expectation that with strong mutation pressure and positive selection, allele frequencies will rise rapidly following origination. When this is true, the best explanation for a beneficial allele being rare is that it only arose quite recently. This expected rapid rise in mutant frequency under strong mutation and positive selection may also be responsible for the much faster extinction times for beneficial alleles compared to deleterious ones (Fig. 3: left), since the longer beneficial alleles persist, the more likely their frequencies are to be pushed upwards towards fixation. Consequently, the mutants that go to extinction are most likely to do so quickly.

A potential limitation of our approach to calculating allele age is that we have assumed equilibrium demography with constant population size. However, this is a limitation of our implementation rather than of the method itself. One solution to this problem is to consider instantaneous switches among different population sizes under a Markov-modulated model. By virtue of our sparse linear algebra approach, this would only be linearly more difficult than the constant population size approach. It could also have advantages over existing diffusion theory methods³³, for example, by faithfully modelling an increase in the population mutation rate during population growth that includes the effect of recurrent mutation. Such considerations may be important for understanding adaptation in organisms with “boom and bust” population dynamics²³. We leave exploration of these ideas for future work.

Methods

Theory. Let $X(t)$ be an absorbing discrete-time Markov chain with known transition matrix P and state-space defined by the number of copies of a mutant allele in a population of N_e effective diploid individuals. Let Q be the submatrix of P that contains only transient-to-transient state transitions. Assume that the current number of mutant alleles x is a transient state, so the allele in question is neither extinct nor fixed. We also assume that the allele entered the population at a specific frequency $p/(2N_e)$, where p is a transient state (we later show how this assumption can be relaxed). In practice, we consider $p = 1$ unless stated otherwise.

The probability of transitioning from state p to state x in time t is simply $P_{p,x}^t$, or equivalently $Q_{p,x}^t$ since both p and x are transient states. Since the Markov chain is absorbing,

$$\sum_{t=0}^{\infty} Q_{p,x}^t = (I - Q)_{p,x}^{-1} \quad (1)$$

is finite³⁴, where I is the identity matrix. This finiteness allows us to fix x and p and specify a probability distribution of the allele age.

$$f_{p,x}(t) = \frac{Q_{p,x}^t}{\sum_{t=0}^{\infty} Q_{p,x}^t} = \frac{Q_{p,x}^t}{(I - Q)_{p,x}^{-1}} \tag{2}$$

A complete measure theoretic construction of this distribution can be found in the supplementary material S1 Appendix. The exact moments of this distribution can be written in terms of the matrix Q by using matrix sum identities. We show the first three below using $[A]_{bc}$ to denote the entry in the b -th row and c -th column of matrix A .

$$\mu_1 = \sum_{t=0}^{\infty} t f_{p,x}(t) = \frac{\sum_{t=0}^{\infty} t Q_{p,x}^t}{\sum_{t=0}^{\infty} Q_{p,x}^t} = \frac{[Q(I - Q)^{-2}]_{p,x}}{[(I - Q)^{-1}]_{p,x}} \tag{3}$$

$$\mu_2 = \sum_{t=0}^{\infty} t^2 f_{p,x}(t) = \frac{\sum_{t=0}^{\infty} t^2 Q_{p,x}^t}{\sum_{t=0}^{\infty} Q_{p,x}^t} = \frac{[Q(I + Q)(I - Q)^{-3}]_{p,x}}{[(I - Q)^{-1}]_{p,x}} \tag{4}$$

$$\mu_3 = \sum_{t=0}^{\infty} t^3 f_{p,x}(t) = \frac{\sum_{t=0}^{\infty} t^3 Q_{p,x}^t}{\sum_{t=0}^{\infty} Q_{p,x}^t} = \frac{[Q(Q^2 + 4Q + 1)(I - Q)^{-4}]_{p,x}}{[(I - Q)^{-1}]_{p,x}} \tag{5}$$

The expected allele age is given by μ_1 , and the variance is given by $\mu_2 - \mu_1^2$.

It is interesting, and relevant if the reader wishes to compute higher moments than those listed above, to notice that the k -th moment μ_k is closely linked to the matrix polylogarithm function $Li_{-k}(Q)$ by the following equation.

$$\mu_k = \frac{[Li_{-k}(Q)]_{p,x}}{[(I - Q)^{-1}]_{p,x}} \tag{6}$$

where

$$Li_{-s}(z) = \sum_{k=1}^{\infty} z^k k^s = \left(z \frac{\partial}{\partial z} \right)^s (z(1 - z)^{-1}) \tag{7}$$

Combining equations 6 and 7 therefore allows for the rapid symbolic computation of the closed-form expressions for any moment μ_k .

Implementation. Computation of the moments in Equations 3, 4 and 5 can be greatly simplified. This simplification requires obtaining a single LU decomposition of a sparse matrix and using it to solve multiple linear systems by back-substitution. This computational approach is similar to our approach in ref.²⁵, where it was applied to the calculation of quantities such as the probability of fixation and sojourn times.

The first step is to calculate the LU decomposition of $(I - Q)^T$, where T denotes transpose. LU decomposition has a theoretical time complexity on the same order as matrix multiplication, and thus can be as large as $O(n^3)$ floating point operations for a dense $n \times n$ matrix. However, much faster solutions are possible for sparse matrices, which scale in terms of the number of non-zero entries (e.g., refs^{35,36}). For Wright-Fisher models, Q and hence $(I - Q)^T$, are typically very sparse (at machine precision), and thus a potentially large time savings can be obtained by exploiting this sparsity. Computation of the LU decomposition is by far the most time-intensive step, but we find it is still feasible for population sizes around 10^5 on typical workstation computers as of the time of writing²⁵. As noted earlier, much larger effective population sizes can be easily considered with the more sparse Moran model.

The second step is to use forward and back substitution to solve multiple linear systems. Given the LU decomposition, this is quite fast and typically requires only a few seconds. First we solve for M_1 in

$$(I - Q)^T M_1 = e_p \tag{8}$$

where e_p is the p -th column of the identity matrix. Note that M_1^T is the p -th row of $(I - Q)^{-1}$, so that the x -th entry of M_1 is in fact $(I - Q)_{p,x}^{-1}$ as required in the denominator of Equations 3 and 4.

Next, we use the same LU decomposition to solve for M_2 in

$$(I - Q)^T M_2 = M_1 \tag{9}$$

Notice that

$$((I - Q)^2)^T M_2 = (I - Q)^T (I - Q)^T M_2 = (I - Q)^T M_1 = e_p \tag{10}$$

so that M_2^T is actually the p -th row of $(I - Q)^{-2}$. We next take the dot product of M_2^T with the x -th column of Q , which we call Q_x .

$$M_2^T \cdot Q_x = [(I - Q)^{-2} Q]_{p,x} = [Q(I - Q)^{-2}]_{p,x} \tag{11}$$

which is what was required in the numerator of Equation 3.

We repeat the procedure and solve for M_3 in

$$(I - Q)^T M_3 = M_2 \quad (12)$$

Again, we have

$$((I - Q)^3)^T M_3 = (I - Q)^T (I - Q)^T M_2 = (I - Q)^T M_1 = e_p \quad (13)$$

so that M_3^T is the p -th row of $(I - Q)^{-3}$. In order to compute the numerator of the second moment, we also need the x -th column of $Q(I + Q)$, which we call A_x . Note this does not in any way necessitate a full matrix multiplication, as we require only the x -th column. Although this is potentially an expensive $O(n^2)$ computation, in practice, sparsity makes it trivially easy. Now we have

$$M_3^T \cdot A_x = [(I - Q)^{-3} Q(I + Q)]_{p,x} = [Q(I + Q)(I - Q)^{-3}]_{p,x} \quad (14)$$

as required in the numerator of Equation 4.

Hence we have calculated all necessary components of the expected value and variance as given in Equations 3 and 4.

The computation of higher moments can be easily implemented as well. To do this, one would first use equations 6 and 7 to obtain closed-form expressions for the needed moments. We recommend using a factored form of the expression so that matrix multiplication is never required in the implementation (it is a convenient property of the polylogarithm that all closed-form expressions of $Li_{-s}(z)$ factor completely over the reals). The implementation would then require extending the above algorithm as needed, i.e. iteratively solving

$$(I - Q)^T M_{k+1} = M_k \quad (15)$$

for M_{k+1} , where M_k^T is the p -th row of $(I - Q)^{-k}$.

We have implemented this approach for the first two moments in our software package Wright-Fisher Exact Solver, WFES²⁵ (available at <https://github.com/dekoning-lab/wfes/>). In practice it takes only seconds to minutes to calculate the relevant quantities for population sizes under $N_e = 100,000$.

As an aside, we note that the full probability distribution can also be feasibly approximated for small N_e to an arbitrary degree of precision by taking the summation in equation 2 to some large finite value.

Simulations

In order to simulate a distribution of allele ages, we must reverse the process, i.e. use the reversed absorbing Markov chain. Specifically, the simulation will start at state x and essentially run backwards in time until it hits state p . It will then either keep going, or stop with a probability equal to the probability that the current visit to state p is the beginning of the chain (when the mutation first entered the population). This backwards simulation can be done by creating a reversed transition matrix and running it in a forwards simulation.

We use the method presented in Chae *et al.*²⁷, which is as follows. The states of the reversed absorbing Markov chain are $\{1, 2, \dots, 2N_e - 2, 2N_e - 1, \text{stop}\}$, where the stop state is absorbing and all others are transient. The reversed Markov chain does not regard fixation or extinction as absorbing states, and in fact does not allow transition to these states at all.

Let P' be the matrix of transition probabilities of the reversed absorbing Markov chain. In its canonical form,

$$P' = \begin{pmatrix} Q' & R' \\ 0 & I \end{pmatrix} \quad (16)$$

We have

$$Q'_{j,k} = \frac{Q_{k,j} N_{p,k}}{N_{p,j}} \quad \text{and} \quad R'_{j,i} = \begin{cases} N_{p,p}^{-1} & \text{if } j = p, i = \text{stop} \\ 0 & \text{otherwise} \end{cases} \quad (17)$$

where Q and N are the transient-to-transient state transition matrix and the fundamental matrix, respectively, of the original Markov chain. (Note that N here is used by convention to represent the fundamental matrix and has no relationship to N_e defined above).

References

1. Fu, W. *et al.* Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature* **493**, 216–20 (2013).
2. Griffiths, R. C. & Tavaré, S. The age of a mutation in a general coalescent tree. *Commun. Statist. - Stochastic Models* **14**, 273–275 (1998).
3. Kimura, M. & Ohta, T. The age of a neutral mutant persisting in a finite population. *Genetics* **75**, 199–212 (1973).
4. Li, W. H. The first arrival time and mean age of a deleterious mutant gene in a finite population. *Am. J. Hum. Genet.* **27**, 274–286 (1975).
5. Maruyama, T. The age of an allele in a finite population. *Genet. Res.* **23**, 137–143 (1974).
6. Watterson, G. Reversibility and the age of an allele ii. two-allele models, with selection and mutation. *Theor. Popul. Biol.* **12**, 179–196 (1977).
7. Slatkin, M. & Rannala, B. Estimating allele age. *Annu. Rev. Genomics Hum. Genet.* **2000**, 225–249 (2000).
8. Kiezun, A. *et al.* Deleterious alleles in the human genome are on average younger than neutral alleles of the same frequency. *PLoS Genet.* **9**, e1003301 (2013).
9. Maruyama, T. The age of a rare mutant gene in a large population. *Am. J. Hum. Genet.* **26**, 669–673 (1974).
10. Altrock, P. M., Gokhale, C. S. & Traulsen, A. Stochastic slowdown in evolutionary processes. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **82**, 011925 (2010).
11. Altrock, P. M., Traulsen, A. & Galla, T. The mechanics of stochastic slowdown in evolutionary games. *J. Theor. Biol.* **311**, 94–106 (2012).

12. Mafessoni, F. & Lachmann, M. Selective strolls: fixation and extinction in diploids are slower for weakly selected mutations than for neutral ones. *Genetics* **201**, 1581–1589 (2015).
13. Charlesworth, B. & Jain, K. Purifying selection, drift, and reversible mutation with arbitrarily high mutation rates. *Genetics* **198**, 1587–1602 (2014).
14. Cutter, A. D., Jovelin, R. & Dey, A. Molecular hyperdiversity and evolution in very large populations. *Mol. Ecol.* **22**, 2074–2095 (2013).
15. Sung, W., Ackerman, M. S., Miller, S. F., Doak, T. G. & Lynch, M. Drift-barrier hypothesis and mutation-rate evolution. *Proc. Natl. Acad. Sci. USA* **109**(45), 18488–18492 (2012).
16. Hughes, A. L., Friedman, R., Rivaille, P. & French, J. O. Synonymous and nonsynonymous polymorphisms versus divergences in bacterial genomes. *Mol. Biol. Evol.* **25**(10), 2199–2209 (2008).
17. Maldarelli, F. *et al.* HIV populations are large and accumulate high genetic diversity in a nonlinear fashion. *J. Virol.* **87**, 10313–10323 (2013).
18. Pennings, P., Kryazhimskiy, S. & Wakeley, J. Loss and recovery of genetic diversity in adapting populations of HIV. *PLoS Genet.* **10**(1), e1004000 (2014).
19. Rouzine, I. M., Coffin, J. M. & Weinberger, L. S. Fifteen years later: hard and soft selection sweeps confirm a large population number for HIV *in vivo*. *PLoS Genet.* **10**(2), e1004179 (2014).
20. Feder, A. F., Kline, C., Polacino, P., Cottrell, M. & Kashuba, A. D. M. *et al.* A spatio-temporal assessment of simian/human immunodeficiency virus (SHIV) evolution reveals a highly dynamic process within the host. *PLoS Pathogens* **15**(3), e1006358 (2017).
21. Messer, P. M. & Petrov, D. Population genomics of rapid adaptation by soft selective sweeps. *TREE* **28**, 659–669 (2013).
22. Jensen, J. D. On the unfounded enthusiasm for soft selective sweeps. *Nat. Commun.* **5**, 5281 (2014).
23. Karasov, T., Messer, P. & Petrov, D. Evidence that adaptation in drosophila is not limited by mutation at single sites. *PLoS Genet.* **6**, e1000924 (2010).
24. Pennings, P. S. & Hermisson, J. Soft sweeps II—molecular population genetics of adaptation from recurrent mutation or migration. *Mol. Biol. Evol.* **23**, 1076–1084 (2006).
25. Krukov, I., De Sanctis, B. D. & de Koning, A. P. J. Wright–Fisher exact solver (WFES): scalable analysis of population genetic models without simulation or diffusion theory. *Bioinformatics* **33**, 1416–1417 (2017).
26. Ewens, W. J. *Mathematical Population Genetics 1: Theoretical Introduction* Edn. 2 (New York: Springer-Verlag, USA, 2004).
27. Chae, K. C. & Kim, T. S. Reversed absorbing Markov chain: A sample path approach. *Oper. Res. Lett.* **16**, 41–46 (1994).
28. Qiu, S. & Fedorov, A. Maruyama's allelic age revised by whole-genome GEMA simulations. *Genomics* **105**, 282–287 (2015).
29. Steinrücken, M., Jewett, E. M. & Song, Y. S. Spectraltdf: transition densities of diffusion processes with time-varying selection parameters, mutation rates and effective population sizes. *Bioinformatics* **32**, 795–797 (2016).
30. Song, Y. & Steinrücken, M. A simple method for finding explicit analytic transition densities of diffusion processes with general diploid selection. *Genetics* **190**, 1117–1129 (2012).
31. Zhao, L., Yue, X. & Waxman, D. Complete numerical solution of the diffusion equation of random genetic drift. *Genetics* **194**, 973–985 (2013).
32. Maruyama, T. & Kimura, M. A note on the speed of gene frequency changes in reverse directions in a finite population. *Evolution* **28**, 161–163 (1974).
33. Evans, S. N., Shvets, Y. & Slatkin, M. Non-equilibrium theory of the allele frequency spectrum. *Theor. Popul. Biol.* **71**, 109–119 (2007).
34. Snell, J. L. & Kemeny, J. G. *Finite Markov Chains* (Van Nostrand, Princeton, NJ, USA, 1960).
35. Amestoy, P. R., Duff, I. S. & L'Excellent, J. Y. Multifrontal parallel distributed symmetric and unsymmetric solvers. *Comput. Methods Appl. Mech. Eng.* **184**, 501–520 (2000).
36. Schenk, O., Gartner, K., Fichtner, W. & Stricker, A. PARDISO: a high-performance serial and parallel sparse linear solver in semiconductor device simulation. *Future Gener. Comput. Syst.* **18**, 69–78 (2001).

Acknowledgements

We thank Nathan Bryans for helpful comments on the manuscript and German Luna Patiarray for mathematical insight regarding measure theory. This work was supported by a Discovery Grant to APJdK from the Natural Sciences and Engineering Research Council of Canada (NSERC DG 03651), by an NSERC USRA award (BDS), and by an Alberta Innovates doctoral fellowship to IK. The authors gratefully acknowledge infrastructure support from the Canada Foundation for Innovation (CFI LOF #31908, APJdK) and the Alberta Children's Hospital Research Institute.

Author Contributions

A.P.J.d.K. designed the research, B.D.S. formulated the methodology, A.P.J.d.K. implemented the methodology and performed the analyses, I.K. implemented the simulation methodology and performed analyses, and A.P.J.d.K. and B.D.S. wrote the manuscript. All authors reviewed the manuscript.

Additional Information

Supplementary information accompanies this paper at <https://doi.org/10.1038/s41598-017-12239-0>.

Competing Interests: The authors declare that they have no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2017