Counting parental contribution - how large sample size makes strong selection weak

Ivan Krukov, Simon Gravel

Abstract

Neutral models of genetic diversity tend to be easier to analyze compared to models including selection. Because lineages are exchangeable in the neutral Wright-Fisher model, for example, the number of lineages that are relevant to the ancestry of a sample at a single locus can only decrease as we go back in time. As a consequence, useful recursion equations can be derived for patterns of polymorphism. Under negative selection, by contrast, the number of relevant lineages can increase as we go back in time, and the equivalent recursion equations do not close. Given a large enough sample size, however, the reduction in the number of lineages due to genetic drift is larger than the increase in the number of lineages due to natural selection, and the number of relevant lineages is unlikely to increase. We use this observation to derive asymptotically closed recursion equations for the distribution of allele frequencies. We show that this approach is accurate under strong drift and strong natural selection. We derive several asymptotic results to understand when the sample size is sufficiently large to overcome the influence of selection.

1. Introduction

The calculation of the allele frequency spectrum (AFS) is an important tool for the inference of demographic histories and other population genetic parameters. In the absence of selection, the number of parental lineages that contribute to the sample decreases back in time due to coalescent events. This means that the equations are closed with respect to the sample size under neutrality. This has paved the way for moment-based recursions of the allele frequencies [1, 2, 3]. Recently, a number of successful methods, including [4, 5, 6], have become popular for this purpose. At their core, these methods describe the time-evolution of the AFS. The goal of these approaches is to obtain the probability of observing a given number of derived alleles in a finite sample conditional on the state of the parental lineages.

Preprint submitted to Theoretical Population Biology *March 28, 2020*

Despite many successes, considering selection has been problematic within this framework. Under negative selection, the number of parental lineages that contribute to a sample can be larger that the sample size itself, due to selective death events. As a consequence, the equation lose the closure property [5].

An extension to the Kingman's coalescent [7], the ancestral selection graph (ASG), [8] considers the ancestry of a sample in the presence of selection. The ASG framework can be used to study the ancestry of highly deleterious alleles [9] under certain assumptions. Another possible resolution is to use an uncontrolled jack-knife approximation to add extra lineages - the method proposed in [5].

Unlike the ASG, the present treatment does not need to assume an infinitely large population size, and also explicitly tracks multiple coalescent events, which is important with large sample sizes [10]. Our approach can be combined with the jackknife, but we improve on it by deriving bounds on the performance our approximation.

2. Background

Consider the behavior of a single biallelic locus in a haploid Wright-Fisher model with a population of size N. We consider a sample of size n lineages from within the population. We want to know how many parental lineages r have contributed to the present sample from one generation in the past. Going back in time, the random variable $\mathcal{R}$ is the number of lineages that contribute to the present day sample. Figure 1 shows examples the different models that we consider below – standard coalescent (1A), coalescent with multiple merges (1B), and coalescent with multiple merges and selection (1C).

First, consider a recursion describing the evolution of the expected allele frequency spectrum (AFS) $\Phi_n^{(t)}$, in the standard coalescent without selection (Figure 1A). At generation t, the sample consists of n lineages, and at $t-1$ there are r parental lineages. In what follows, we closely follow the exposition of [5]. The standard coalescent model allows at most one event per generation, so we need to consider only two cases: namely $r = n$ if no coalescent occurs, and $r = n - 1$ if a single event took place.

Without a coalescent event, r parental lineages are sampled randomly into the present generation, so the allele frequencies remain the same: $\Phi_n^{(t)} = \Phi_n^{(t-1)}$. With a single coalescent event, the contribution comes from $r = (n-1)$ parental lineages: $\Phi_n^{(t)} = \mathcal{D}(\Phi_{n-1}^{(t-1)})$, where $\mathcal{D}$ is a sparse

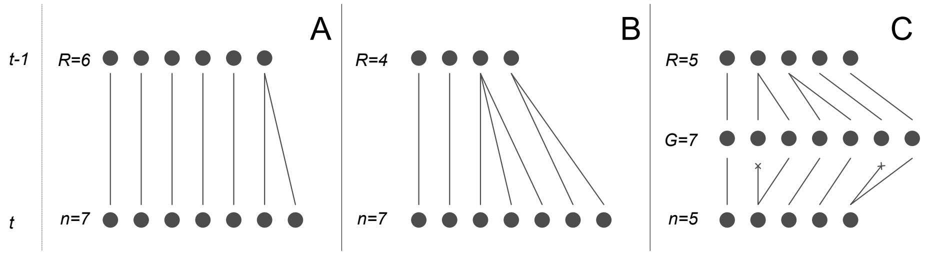

Figure 1: Realizations of sampling models, showing $\mathcal{R} = r$ parental lineages at $t-1$ contributing to n lineages in present generation t. **A** Standard coalescent - at most 1 coalescent event per generation; $r = [n-1, n]$. **B** Coalescent with multiple merges - parents *3* and *4* have 3 and 2 offspring, respectively; $r \leq n$ **C** Including selection - each parent produces a random number of gametes ($\mathcal{G}$), which then may survive to produce offspring; $r \leq n$, $r \leq g2n$

$n \times (n-1)$ matrix describing the effect of drift (we demonstrate construction of $\mathcal{D}$ and related terms in appendix A). Combining the two terms, we have:

$$\Phi_n^{(t)} = \Phi_n^{(t-1)} + \mathcal{D}\Phi_{n-1}^{(t-1)} \tag{1}$$

In other words, the AFS of size n at time t can be obtained if we know the AFS of sample sizes n and $n-1$ in the previous generation. This gives rise to a recursion formula for the frequency spectrum that can be solved efficiently [5]. Here we would like to generalize such an approach to cases including natural selection and large sample sizes.

We first incorporate the effects of large sample size. As previously shown in [10, 11], the coalescent approximation may not be adequate in this setting, since multiple coalescent events can take place within a single generation (1B). With a slight abuse of notation we denote $\mathcal{D}_i$ as the i^{th}-order diffusion matrix in which i lineages are lost due to genetic drift. For example, $\mathcal{D}_2$ includes both three-way coalescent and double two-way coalescent. In Appendix A, we demonstrate an efficient dynamic programming algorithm to exhaustively enumerate all the events for a drift-only model.

With multiple coalescent events per generation, (1) becomes:

$$\Phi_n^{(t)} = \Phi_n^{(t-1)} + \sum_{i=1}^{n} \mathcal{D}_i \Phi_{n-i}^{(t-1)} \tag{2}$$

The equation (2) is still closed in terms of the sample size, since Φ_n^t only depends on $r = (n-i) < n$ parental lineages.

If we now consider selective death events, we must also account for lineages that were not transmitted due to selection. We use the following model to describe this (1C). Each generation, a random number of $\mathcal{R} = r$ parental lineages produce a large number of gametes, $\mathcal{G} = g$. Then n individuals are formed by randomly sampling gametes, without replacement. The probability of successfully sampling a particular gamete is $1 - xs$, where x is the frequency of the derived allele in the parental generation. Then, with probability xs, a new gamete is re-drawn. This sampling scheme allow us to consider drift ($r \to g$) and selection ($g \to n$) within the same generation as distinct processes.

For example, in case of a single selective death event, we have $\Phi_n^{(t)} = \mathcal{S}(\Phi_{n+1}^{(t-1)})$, with selection matrix $\mathcal{S}$ (see appendix B). Multiple selection events are possible per generation, but we restrict our attention to the case where each lineage experiences at most one selective death event. This still allows us to consider strong selection, with at most $r \leq 2n$ parental lineages contributing:

$$\Phi_n^{(t)} = \Phi_n^{(t-1)} + \sum_{j=0}^{2n} \sum_{i=1}^{n} \mathcal{S}_j \mathcal{D}_i \Phi_{n-i+j}^{(t-1)} \tag{3}$$

The closure no longer holds for (3), as up to $g \leq 2n$ gametes can contribute to the sample. However, the effect of a large sample size counteracts the additional lineages needed due to selection. The opposite effects of drift and selection on the number of lineages relevant to a sample are particularly clear in the context of the ancestral selection graph (ASG): in which the number of lineages relevant to a sample can be described as a birth-death process [8, 9]:

$$n \to \begin{cases} n+1 & \text{at rate } \frac{Ns}{2} n \text{ (selection)} \\ n-1 & \text{at rate } \binom{n}{2} \text{ (coalescence)} \end{cases} \tag{4}$$

The rate of coalescence is higher than the rate of selective deaths if the number of lineages $Ns < n - 1$. While building an ancestral selection graph, there is no particular constraint on the number of lineages that exist, as long as we eventually find a common ancestor.

Our goal here is to define recursions generalizing equation (1). For the equations to be closed, we need to ensure that the rate of coalescence is large enough that it overcomes the rate of selection not only on average, but *almost always.*

The rest of the paper is organized into two sections. In the first section, we construct a recursion to track the number of derived lineages in a large sample from a Wright-Fisher model, similar to [5, 6]. In this, we fully account for multiple coalescent events per generation, and show that we restore closure with increasing sample size. In the second part, we derive a number of asymptotic results to get a better understanding of the process. We construct an exact probability distribution for the number of contributing parental lineages, together with several approximations. Importantly, we derive a normal approximation that allows us to calculate a quantile of the sample size where the system is approximately closed.

3. Results

3.1. Markov process construction

We first define a recursion equation for the distribution of allele frequency in a sample of size n from a haploid Wright-Fisher population of size N,

In the neutral case, we want to compute the matrices $\mathcal{D}_i$ from Equation (2). SG: Or, if you want to compute the square matrix, define it above? These transition probabilities are conceptually simple: $(D_{i,jk}$, is the probability that SG: .... However, the large number of combinatoric possibilities when accounting for multiple events make the computation cumbersome.

We therefore use a dynamic programming approach where we construct every transition probability matrix in for the (parental) sample sizes $p \in [1, n-1]$. Then the transition probabilities $P((j,n) \to (i,n))$ can be obtained from $P((j,n-1) \to (i,n-1))$. We derive this equation in appendix A.

In the case with selection, the number of parental lineages p can exceed n: $p \in [1, \infty]$. We restrict our attention to the cases where parents of each lineage in the sample experience at most one selective death event (*i.e.* $s \ll 1$, but Ns can still be much larger than 1), so we consider the range of $p \in [1, 2n]$.

We therefore want to find $Q((j,m) \to (i,n))$ SG: define(note that the current and parental sample sizes, n and m can be different) in terms of $Q((j',m') \to (i',n-1))$, where $m' \in [1, 2n]$, $j', i' \in [1, m']$. The list of events that contribute to $Q((j,m) \to (i,n))$ is shown in Figure 2, and the full derivation is in appendix B. SG: This needs a bit more guidance.

In addition to every transition probability matrix of size $p \in [1, n-1]$, the calculation in the case with selection additionally requires the calculation of rectangular matrices where the number

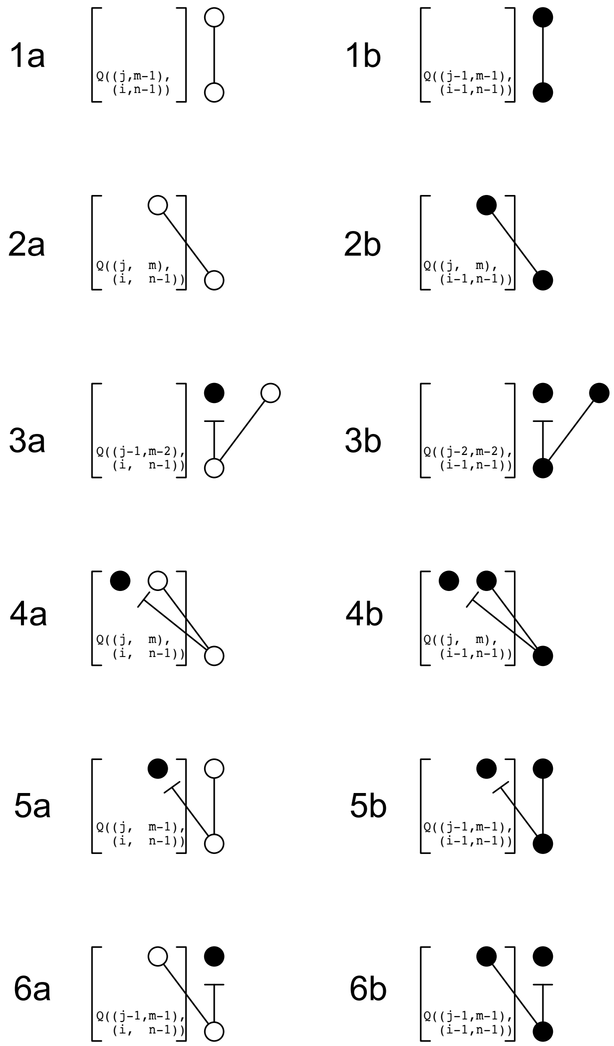

Figure 2: Contribution of parental configurations to a present sample. The transition probability at $Q((j,m) \rightarrow (i,n))$ is the sum of 12 terms shown in the figure. The filled circles indicate derived alleles, empty circles - ancestral alleles, lines show ancestral descent. The square brackets and the align within indicate the parental configuration that contributes to the entry. Full derivation is shown in appendix B.

of parental contributors ($p' \in [1, 2n]$) is not equal to the sample size. As a result, the calculation time is of the order of $O(n^4)$ for the selection case, while it is only $O(n^3)$ for the neutral case.

3.2. Calculation of site frequency spectra

Once the matrix Q is constructed, it can be used to calculate the site frequency spectrum within a sample. For the infinite sites model at equilibrium, we can calculate the *SFS* Φ as a solution to a linear system:

$$\Phi = \Phi Q + n\mu e_1 \tag{5}$$

SG: is Q square, now? Have you truncated it? This needs to be clarified! where μ is the per-site mutation rate, and e_1 is the first column of the identity matrix of size n. Figure 3 shows the comparison of the *AFS* calculated from Equation (5), the diffusion approximation [12, eq. 9.23], and the calculation performed in `Moments` [5]. Panel A shows a comparison at $Ns = 100$ SG: caption says otherwise, with the population size ($N = 2000$), which is substantially larger than the sample size ($n = 200$). There is a small deviation between the approaches at large allele frequencies. However, since highly deleterious alleles are unlikely to reach these frequencies, the difference is immaterial SG: I would not say this, since this will eventually speak to the fixation probability of the deleterious alleles – might be impactful over long time scales. At stronger selection coefficients, `Moments` suffers from numerical instability, while the diffusion approximation performs well (not shown).

If the sample size is the same as the population size ($n = N = 200$) (Fig. 3B), the diffusion approximation and `Moments` perform poorly, while our approach remains stable. This is expected, since the diffusion framework does not perform well if multiple coalescent events contribute. SG: Furthermore, if our sample size is the entire population, we expect recursion equations to be closed, since by definition we cannot need information about more than N samples! SG: Show that exact transition as validation?

3.3. Closure properties

To show the closure properties of Q, we can calculate the total probability that more that n parental lineages contribute to the sample of a given size. By construction, the sum of rows of Q should correspond to the total probability mass that included configurations contribute (Fig. 2).

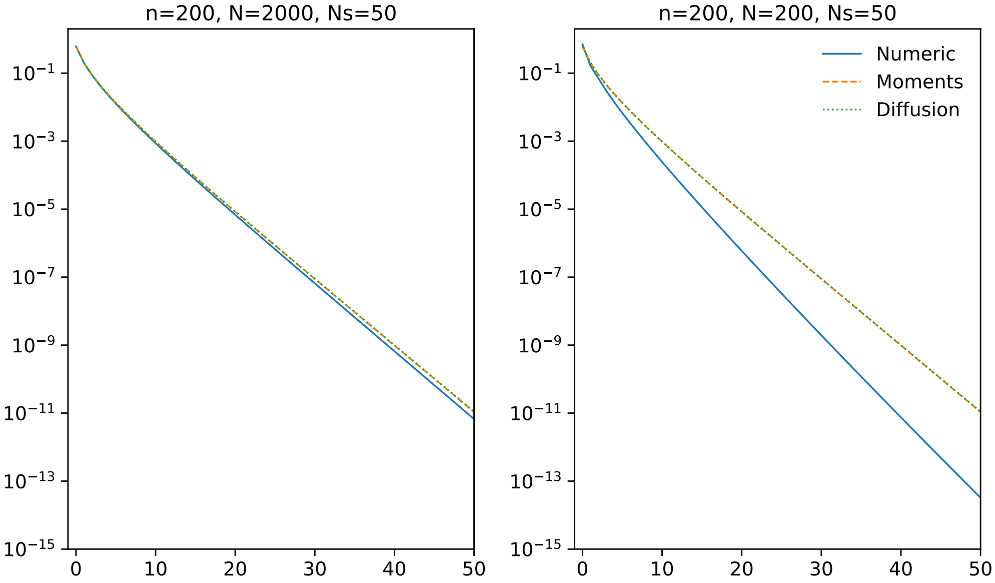

Figure 3: Site frequency spectra in a sample of size $n = 200$, for highly deleterious alleles ($Ns = -50$). (A) shows the frequency spectrum in a sample from a large population ($N = 2000$), (B) in a small population ($N = 200$). Both panels are truncated at 10^{-15}, to show only sufficiently high allele frequencies. Y-axis on a logarithmic scale. SG: No need to specify the axis – people can see. Are the SFS' normalized?

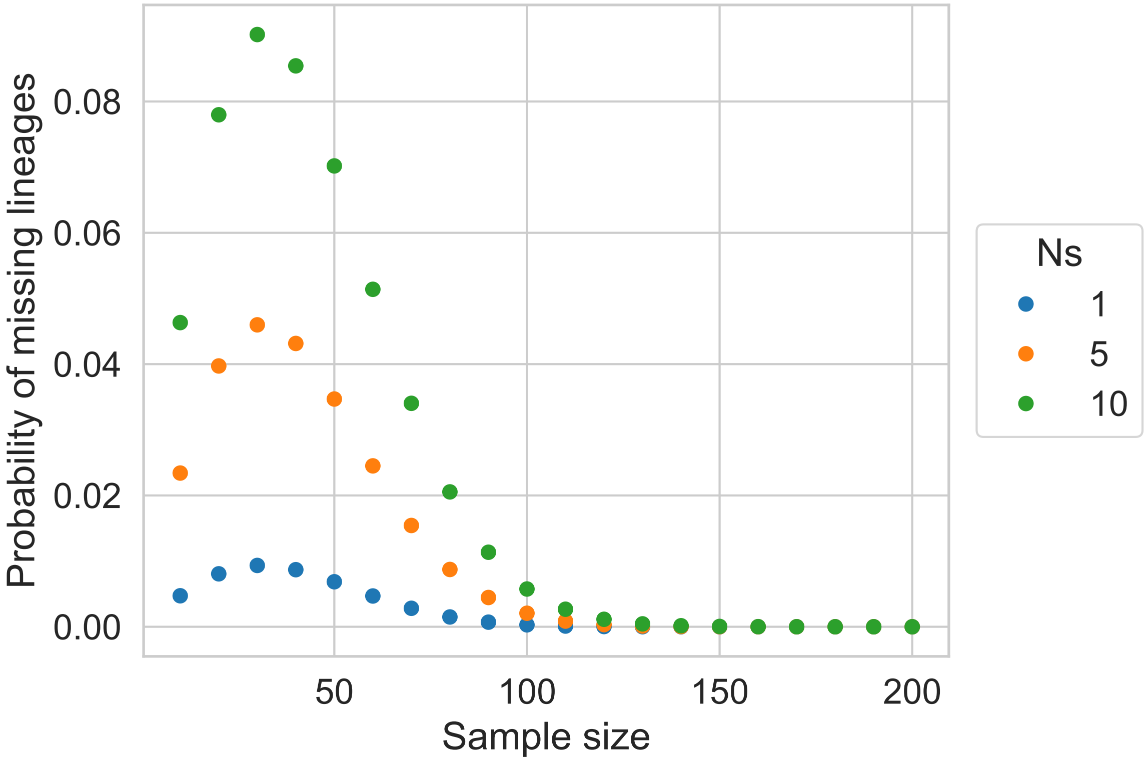

Figure 4: Probability that unaccounted lineages contribute to the transition probabilities. The probabilities are calculated as 1 minus the sum of probabilities for the state where every allele is derived. SG: Did we already discuss having a log y? Include N in caption!

Thus, the probability that some number of configurations are unaccounted for, with j derived alleles in the sample SG: Parents?, is given by $1 - \sum_{i=0}^{n} Q_{i,j}$. This probability depends on the number of derived alleles carried by the parental sample: the more derived alleles, the higher the likelihood of a selective event. Figure 4 shows the probability of missing configurations in a sample size of $n = 200$ in the worst-case scenario, with $j = 200$ derived lineages.

Since the expected number of drift events increases quadratically and the number of selective events increases only linearly, the probability that we need additional lineages decreases rapidly with sample sizes.

4. Asymptotic closure properties

We now want to determine what sample size is sufficient so that the number of coalescent events due to drift is almost always larger than the number of selection events, such that the system remains closed (3). We derive several approximations to the model proposed in the first section,in order to get a better understanding of this behavior.

As a first order approximation, we consider the mean number of lineages that contribute via the two processes. Then, we construct a full probability distribution of the number of contributing lineages one generations into the past. Finally, we propose a normal approximation to this distribution, in order to derive a simple quantile function for the number of used lineages.

We want to know the upper bound on the number of lineages used. Since the maximum number of lineages will be resampled when all lineages are derived, we will usually assume $x = 1$ in the following calculations. This also allows us to treat the lineages as exchangeable [9]. Note that in section 3.1, we did not assume exchangeability of lineages, which led to a considerably more complex formulation.

For a given sample size, the probability that p parents have contributed is:

$$Pr(\mathcal{P} = p|n) = \sum_{\mathcal{P}} Pr(\mathcal{P} = p|\mathcal{G} = g)Pr(\mathcal{G} = g|n) \tag{6}$$

Where $\mathcal{P}$ and $\mathcal{G}$ are random variables denoting the number of contributing parents and gametes, respectively.

Before deriving the distribution formally, we seek to obtain several approximate results.

4.1. Expected number of lineages used

First, we seek an approximate expression for the expectation of the total number of lineages used. This can be approximated as the sum of expectations of the number of lineages sampled under drift plus the number of lineages rejected by selection (selective deaths). The number of parents that contribute to n gametes (drift) will be:

$$\hat{E}[\mathcal{P}|n] = N(1 - \left(1 - \frac{1}{N}\right)^n) \tag{7}$$

The probability of selecting a particular parent is $\frac{1}{N}$, so the probability of selecting different parents for n individuals is $(1-\frac{1}{N})^n$. Then one minus this value is the probability that the same parent was picked at least once by any of the n individuals.

For selection, we want to consider the expected number of gametes that are rejected by selection to form a sample size of n. If the probability of rejection is xs, the scheme is described by the negative binomial distribution, where the random variable is the number of failures, given n successes. The expectation of this parameterization of negative binomial is:

$$\hat{E}[\mathcal{G}-n|n] = n\left(\frac{xs}{1-xs}\right) \tag{8}$$

Then summing the expectations of the two random variables yields:

$$\begin{aligned}
\hat{E}[\mathcal{P}+\mathcal{G}-n] &= \hat{E}[\mathcal{G}-n|n] + \hat{E}[\mathcal{P}|n] && (9)\\
&= N(1-\left(1-\frac{1}{N}\right)^n) + n\left(\frac{xs}{1-xs}\right) && (10)\\
&\underset{N\gg n}{\approx} \frac{nxs}{1-xs} - \frac{n^2}{2N} && (11)
\end{aligned}$$

The second approximation is made under the assumption that the sample size is much smaller than the population size. We can see that the expected number of lineages sampled will be increased by selection as a linear term. Drift tends to decrease the number of lineages as a quadratic term with respect to the sample size. This is analogous to the results from the ancestral selection graph [8], but now includes sample size directly.

We now want to ask when the expected number of lineages is less that the sample size:

$$\begin{aligned}
\hat{E}[\mathcal{P}] &< n\\
\frac{nxs}{1-xs} - \frac{n^2}{2N} &< n && (12)\\
n &\geq \frac{2Nxs}{1-xs}\\
&\approx 2Nxs && (13)
\end{aligned}$$

This allows us to derive a simple expression for the sample size where drift overcomes selection. Figure 5 shows this for several selection coefficients, assuming the entirety of the sample is derived

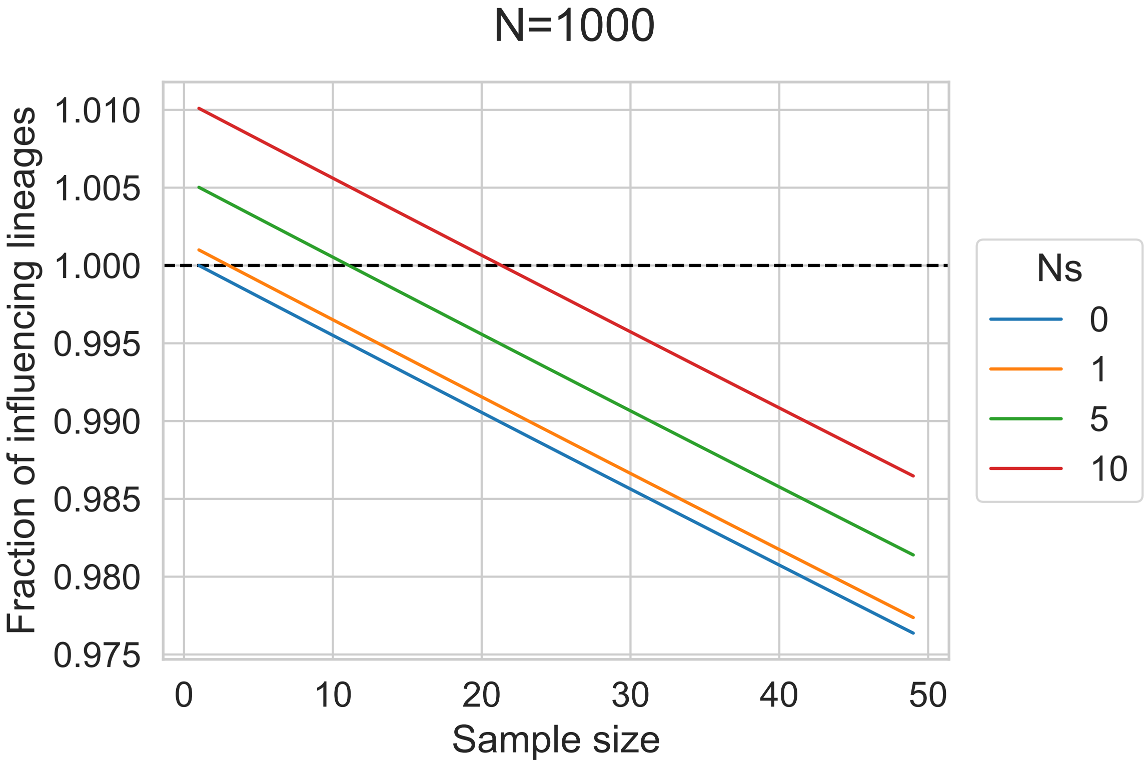

Figure 5: Critical sample size for different selection coefficients. The Y axis shows the fraction of parental lineages over the sample size, $\frac{p}{n}$, each line corresponds to a different selection coefficient. Above $\frac{p}{n} \geq 1$, selection dominates, below – drift. The critical sample size, where the expected number of parental contributing lineages is smaller than the sample size is well-approximated by $2Ns$.

in a population of $N = 1,000$. The Y axis shows the fraction of contributing parental lineages to the sample size, $\frac{p}{n}$. Above the horizontal line $\frac{p}{n} > 1$, selection dominates. Below, drift reduces the number of used lineages. The intercept of the line with $\frac{p}{n} = 1$ is the critical sample size, which is well-approximated by $2Ns$.

Using the same equation, we can track the expected number of used parental lineages back in time, which we denote as n_{t-1}:

$$n_{t-1} = \frac{n_t xs}{1 - xs} - \frac{n_t^2}{2N} \tag{14}$$

We solve this recurrence going back in time 10,000 generations, producing figure 6. The equilib-

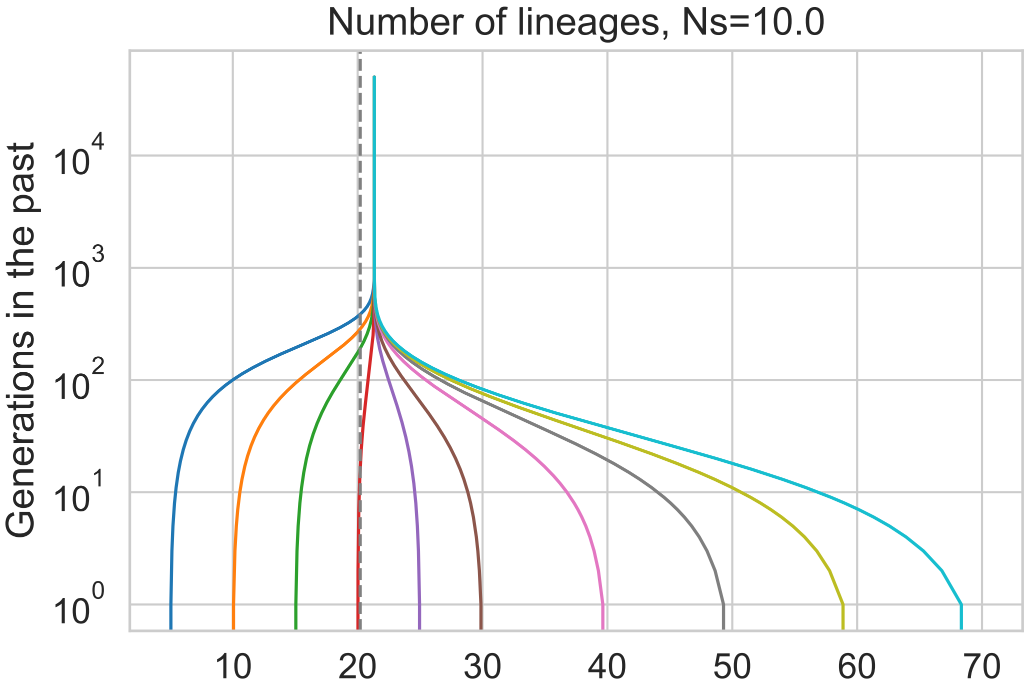

Figure 6: Expected number of contributing parental lineages back in time. Starting at a given sample size, the number of contributions is tracked with align 14. The Y axis shows time on a logarithmic scale, X axis is the sample size. $N = 1000$, $Ns = 10$.

rium point is well-approximated by $2Ns$, shown as a dashed line here. This is the same as solving equation 12 explicitly. Non-withstanding of the starting sample size, we converge to the equilibrium relatively quickly.

4.2. Distribution

We now construct a probability distribution of the number of contributing lineages one generation into the past.

The number of parental lineages used by drift can be modelled by the modified occupancy (Arfwedson) distribution [9, 13, 14]. This is given by:

$$P(\mathcal{P} = p|\mathcal{G} = g) = \frac{S_2(g,p)N!}{(N-p)!N^g} \tag{15}$$

where $S_2(g,p)$ is a Stirling number of the second kind, which is the number of ways to partition g objects into p categories. A typical statement of the occupancy distribution is that we have N urns and g colored balls, and we want to know the probability that exactly p of the urns will be occupied (see [14] section 10.4 for a thorough treatment). In our case, N is the population size, urns correspond to the parents, colored balls to gametes. Note that the under drift, the number of parents will be smaller or equal to the number of gametes $p \leq g$.

The occupancy distribution is not simple to evaluate, but good performance can be achieved by pre-computing a table of reduced occupancy numbers, using the algorithm of [13].

As stated before, the number of lineages sampled under selection is described with a negative binomial distribution. Unlike 8, however, we are looking for the total number of lineages sampled, not simply the number of failed trials. In this parameterization, the probability of the negative binomial is given by:

$$P(\mathcal{G} = g|n) = \binom{g-1}{n-1}(1-xs)^n(xs)^{g-n} \tag{16}$$

Here, the number of gametes can be larger that the sample size $n \leq g$, if selection is present $(s < 0)$.

Combining the two distributions together through 6, we get:

$$Pr(\mathcal{P} = p|n) = \sum_{g=1}^{\infty} \frac{S_2(g,p)N!}{(N-p)!N^g}\binom{g-1}{n-1}(1-xs)^n(xs)^{g-n} \tag{17}$$

Unfortunately, this distribution does not have a simple analytical form. In certain parameter regimes, this can be approximated by the normal distribution [14, 13], which we describe in the next section.

Figure 7 shows the distribution of the number of contributing parental lineages for several selection coefficients for a sample $n = 20$. In the absence of selection, the distribution has zero probability above $n = 20$, as no extra lineages can be sampled. As the strength of selection is

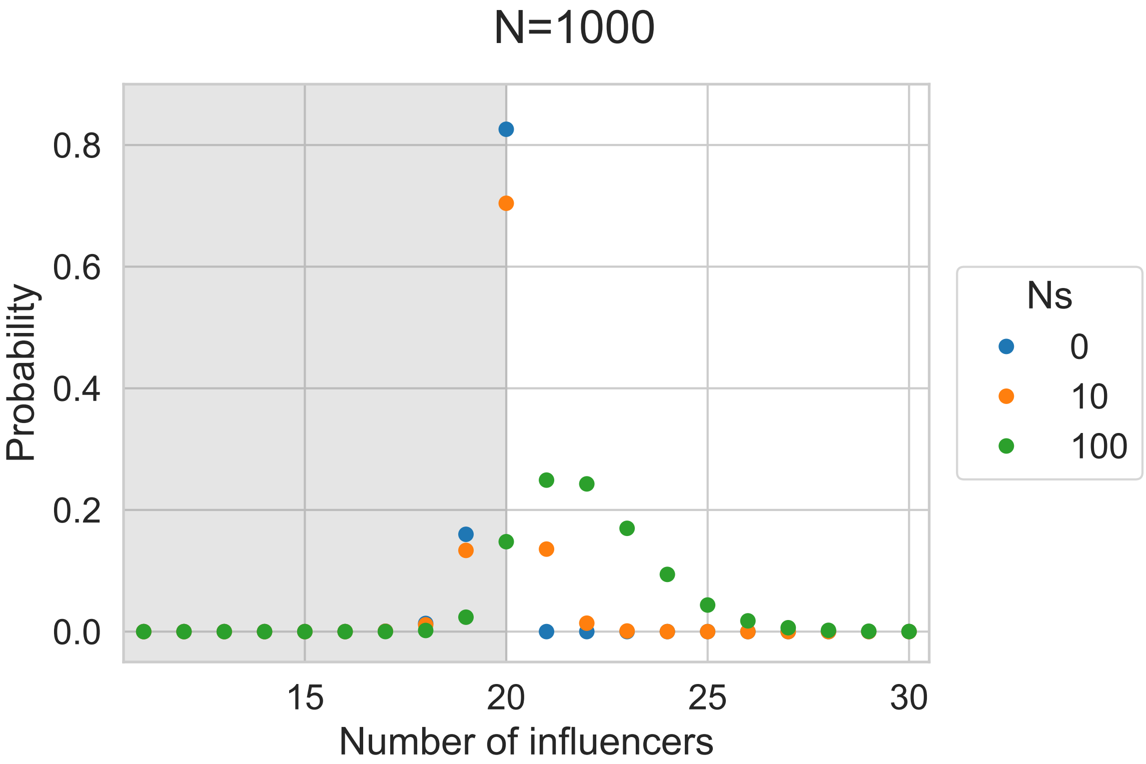

Figure 7: The distribution of the number of parental contributing lineages one generation into the past ($n = 20$, $N = 1000$). Shaded area shows the drift-dominated regime, where the number of lineages is smaller than the sample size.

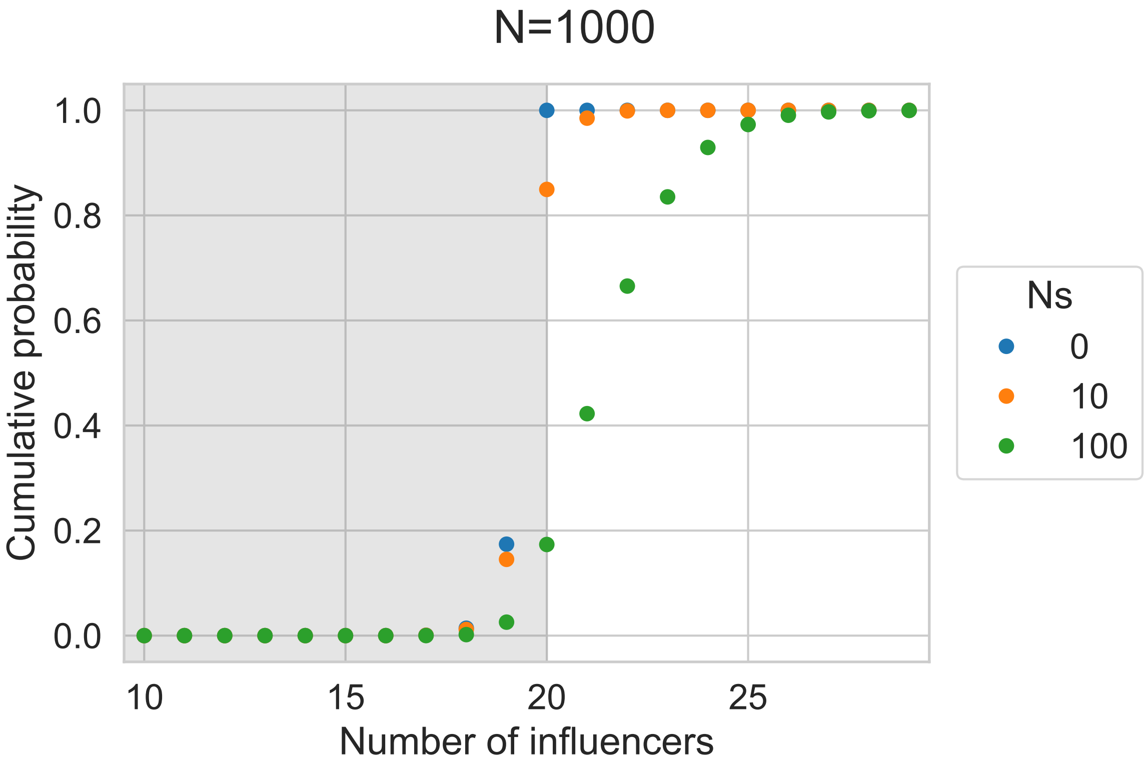

Figure 8: The cumulative distribution of the number of parental contributing lineages one generation into the past ($n = 20$, $N = 1000$). Shaded area shows the drift-dominated regime, where the number of lineages is smaller than the sample size.

increased, we begin requiring larger number of lineages. At the equilibrium point ($Ns = 10$, (12)), the distribution is symmetric.

We note that at the critical sample size, the probability that we will have a sufficient number of lineages is only 50%. In order to guarantee that drift will out-pace selection, we can calculate the cumulative distribution - Figure 8. This shows that a sample size in which the majority of lineages are accounted for can be substantially larger than the critical sample size of equation (12). To derive a convenient expression, we turn to the normal approximation in the next section.

4.3. Normal approximation

Finally, we can construct a normal approximation to the distribution of the number of contributing lineages. The occupancy distribution is approximated by the normal [13] when $n \ll N$.

Likewise, the number of failures (eq. (8)) before a given number of successes, can be approximated by the normal distribution. In the case of large population size, as required by the approximation of the occupancy by the normal, we can approximate the total number of contributing lineages as the sum of lineages contributed by the two distributions. The random variable which is a sum of two normally-distributed random variables is also normal, with $\mu = \mu_1 + \mu_2$ and $\sigma^2 = \sigma_1^2 + \sigma_2^2$. By combining the required expectations and variance, we find that the normal approximation then has the form:

$$Pr(\mathcal{P} = p|n) \approx \mathcal{N}(\mu = (sn)/(1-s) + N(1-(1-1/N)^n), \tag{18}$$

$$\sigma = \sqrt{N\left((N-1)\left(1-\frac{2}{N}\right)^n + \left(1-\frac{1}{N}\right)^n - N\left(1-\frac{1}{N}\right)^{2n}\right) + \frac{ns}{(1-s)^2})} \tag{19}$$

Figure 9 shows the quantiles of the normal approximation. We see that up to 99% of the lineages will be contained within the sample of 200 with $Ns = 20$. Larger percentiles will require larger sample sizes.

5. Conclusion

In this work we show that with the increasing sample size, the effect of drift overcomes the effect of selection. As a result, it is possible to construct asymptotically closed solutions to coalescent with selection, provided the sample size is sufficiently large.

The sample size where the expected number of extra lineages required by selection is less than the sample size is well approximated by $2Nxs$. However, the sample size that guarantees that almost no extra lineages are required is considerably larger (8).

Using this observation, we can construct a Markov model that describes the number of derived alleles in the sample. With a sufficiently large sample size, such Markov chains are closed, and can be used for the calculation of the allele frequency spectra with strong selection.

As a future direction, we want to combine the jackknife approximation [5] with the results presented here. Since the jackknife is an uncontrolled approximation, the current results provide a more sensible approach. However, we can still employ the jackknife in the cases where extra

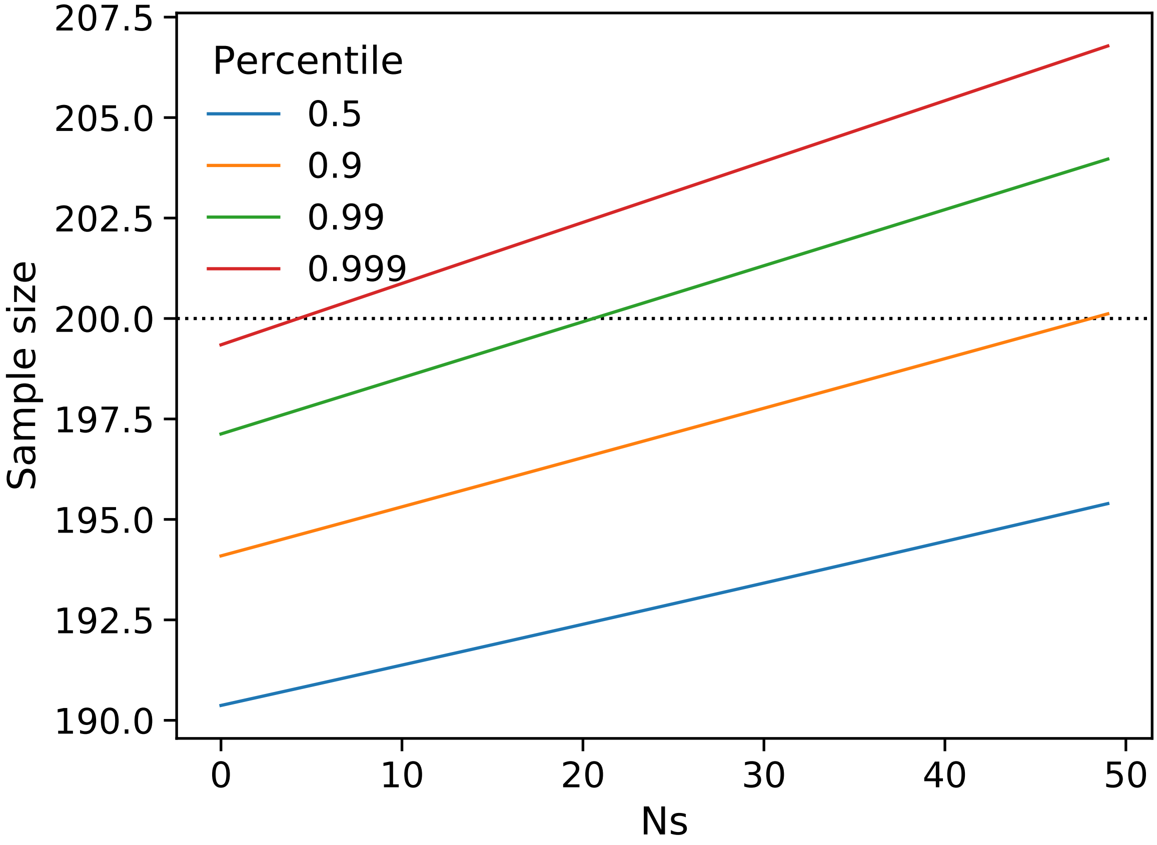

Figure 9: The quantile function of the closure of the sample. Each line corresponds to different percentile of the normal approximation. Black dashed line shows the reference sample size $n = 200$.

lineages are still required in the current approach. This has the potential of further improving the accuracy of the model and computational efficiency.

References

[1] M. Kimura, J. F. Crow, The number of alleles that can be maintained in a finite population, Genetics 49 (4) (1964) 725–738.

[2] W. J. Ewens, The sampling theory of selectively neutral alleles, Theoretical Population Biology 3 (1) (1972) 87–112. doi:10.1016/0040-5809(72)90035-4.

[3] P. Donnelly, T. G. Kurtz, Particle representations for measure-valued population models, The Annals of Probability 27 (1) (1999) 166–205. doi:10.1214/aop/1022677258.

[4] R. N. Gutenkunst, R. D. Hernandez, S. H. Williamson, C. D. Bustamante, Inferring the joint demographic history of multiple populations from multidimensional snp frequency data, PLOS Genetics 5 (10) (2009) e1000695. doi:10.1371/journal.pgen.1000695.

[5] J. Jouganous, W. Long, A. P. Ragsdale, S. Gravel, Inferring the joint demographic history of multiple populations: Beyond the diffusion approximation, Genetics 206 (3) (2017) 1549–1567. doi:10.1534/genetics.117.200493.

[6] J. A. Kamm, J. Terhorst, Y. S. Song, Efficient computation of the joint sample frequency spectra for multiple populations, Journal of Computational and Graphical Statistics 26 (1) (2017) 182–194. doi:10.1080/10618600.2016.1159212.

[7] J. F. C. Kingman, The coalescent, Stochastic Processes and their Applications 13 (3) (1982) 235–248. doi:10.1016/0304-4149(82)90011-4.

[8] S. M. Krone, C. Neuhauser, Ancestral processes with selection, Theoretical Population Biology 51 (3) (1997) 210–237. doi:10.1006/tpbi.1997.1299.

[9] J. Wakeley, Coalescent Theory - an Introduction, W. H. Freeman, New York, 2009.

[10] A. Bhaskar, A. G. Clark, Y. S. Song, Distortion of genealogical properties when the sample is very large, Proceedings of the National Academy of Sciences 111 (6) (2014) 2385–2390. doi:10.1073/pnas.1322709111.

[11] D. Nelson, J. Kelleher, A. P. Ragsdale, G. McVean, S. Gravel, Coupling wright-fisher and coalescent dynamics for realistic simulation of population-scale datasets, bioRxiv (2019) 674440doi:10.1101/674440.

[12] W. J. Ewens, Mathematical Population Genetics: I. Theoretical Introduction., 2nd Edition, Vol. 27 of Interdisciplinary Applied Mathematics, Springer New York, New York, 2004, oCLC: 958522782.

[13] B. O'Neill, The classical occupancy distribution: Computation and approximation, The American Statistician (2019) 1–12doi:10.1080/00031305.2019.1699445.

[14] N. Johnson, A. Kemp, S. Kotz, Occupancy distributions, in: Univariate Discrete Distributions, 3rd Edition, Wiley Series in Probability and Statistics, John Wiley & Sons, Ltd, 2005.