Models of strong selection in large samples

Ivan Krukov, Simon Gravel

Abstract

Neutral models of genetic diversity tend to be easier to study than models including selection. In the Wright-Fisher model, the number of parental lineages that contribute to ancestry of a sample is lower than the number of offspring. As a consequence, useful recursion equations can be derived for patterns of polymorphism. By contrast, under negative selection, the number of relevant lineages can increase as we go back in time, due to selective deaths. As a result, the equivalent recursion equations do not close. However, given a sufficiently large sample size, the expected reduction in the number of contributing lineages due to coalescence is larger than the increase due to selection, so the net number is unlikely to increase. We use this observation to derive asymptotically closed recursion equations for the distribution of allele frequencies in finite samples. We show that this approach is accurate under strong drift and strong natural selection. We derive several asymptotic results to determine when the sample size is sufficiently large for drift to overcome the effect of selection.

1. Introduction

The allele frequency spectrum (AFS) is an important summary of genetic diversity that is commonly used to infer demographic history and natural selection (). Given a demographic scenario of population size histories and migrations, the diffusion approximation or coalescent simulations can be used to obtain a predicted AFS (). By comparing predictions to the observed AFS, one can compute likelihoods for different demographic scenarios. Unfortunately, the AFS calculations can be time consuming with complex demographic models, for example with multiple populations, or with large sample sizes ().

In the absence of selection, efficient computational shortcuts can be used. In particular, recursion equations have been derived for moments of the allele frequency distribution (Kimura and Crow,

1964; Ewens, 1972; Jouganous et al., 2017). Recently, these recursions have been useful in fitting complex demographic models to genetic data (Jouganous et al., 2017; Kamm et al., 2017).

In the presence of natural selection, the corresponding recursion equations do not close (**?**Jouganous et al., 2017) – they form an infinite set of coupled ordinary differential equations. Moment-based closure approximation have been developed (Jouganous et al., 2017), but these are not robust to strong selection and their convergence properties are not well understood.

Closure of the moment equations under the neutral Wright-Fisher model occurs because the number of parental lineages that contribute to the present day sample is equal to or smaller than the sample size, which is due to coalescent events back in time (Kingman, 1982). This does not hold under negative selection – due to selective deaths, the number of parental lineages that contribute can be larger than the sample size, n. As we demonstrate later, this leads to a potentially infinite number of terms in the equations. The potential increase in the number of relevant lineages as we go back in time is explicit in the ancestral selection graphs (ASG), (Krone and Neuhauser, 1997).

The actual increase in the number of lineages as we go back in time depends on the relative importance of drift and selection. The number of coalescent events per generation is proportional to the square of the sample size n, while the number of selective deaths is linear in n. This suggests that with sufficiently large samples, the effect of selection would be smaller than that of drift, which would prevent the increase in the number of lineages – at least most of the times. Thus recursion equations can become almost-closed, in a sense that we will explore below.

An additional complication is multiple and/or simultaneous coalescent events – which emerge with large sample sizes (Bhaskar et al., 2014). The standard coalescent model only allows one event per generation, but we also need to consider higher-order events, *e.g.* multiple two-lineage or three-lineage mergers. These multiple-lineage coalescent events oppose the effect of selection by rapidly decreasing the number of contributing lineages (Nelson et al., 2019).

In this article we derive these asymptotically-closed recursions in the Wright-Fisher model, and study their behaviour and applications for modelling the distribution of allele frequencies under strong selection.

2. Background

We consider a haploid Wright-Fisher model of size N, focusing on a single biallelic locus. For a present-day sample with n_o offspring lineages at time t, we will be looking for recursion equations

for the allele frequency spectrum $(\Phi_{n_o}^{(t)})$ by considering the sampling process in a finite sample under drift and selection. We define the AFS as a probability of observing a given allele frequency for a single locus.

To model selection, we imagine that all parents generate a large number of gametes, and that offspring pick gametes at random. Draws from the deleterious allele are rejected with probability s, triggering a re-draw from another parent (Fig.1B). This leads to a $1:1-s$ advantage in favour of the advantageous allele, and makes explicit the number of lineages that need to be drawn to generate a sample of size n_o. In particular, the number n_g of gametes that are drawn increases with s.
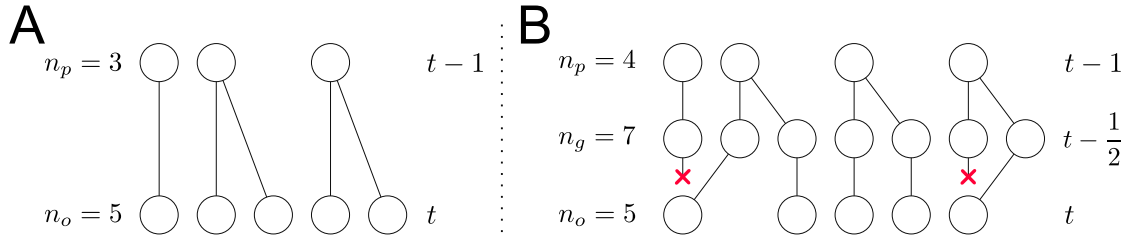

Figure 1: Realizations of sampling parental lineages under neutrality (A) and selection (B). **A** Under neutrality, possible coalescent events imply that number of parental lineages n_p at $t-1$ is less than or equal to n_o offspring lineages at t. **B** With selection, we add an intermediate gamete n_g generation at $t-\frac{1}{2}$. Production of gametes is neutral, so $n_p \leq n_g$. Gametes are sampled with rejection into offspring, so $n_o \leq n_g$. Rejected samples shown with red crosses. n_p - parental sample size (at $t-1$), n_g - number of gametes (at $t-\frac{1}{2}$), n_o - offspring (current) sample size (at t).

In this selection model, the number n_p of distinct parental lineages drawn is $n_p = n_o + n_r - n_c$, with n_r the number of rejections and n_c the number of coalescent events. Thus n_p can be smaller than n_o (if there is more drift than selection), or larger (if there is more selection than drift).

To express the allele-frequency spectrum $\Phi_{n_o}^{(t)}$ for n_o offspring in terms of the AFS at time $t-1$, we can sum over the random variables n_p, the number of distinct parental lineages selected, and j, the number of derived lineages among them:

$$\Phi_{n_o}^{(t)}(i) = \sum_{n_p,j} P_{n_o}(i,j,n_p) = \sum_{n_p,j} P_{n_o}(i|j,n_p)P(j,n_p), \tag{1}$$

where the subscript n_o indicates probabilities that depend on n_o. The event (j, n_p) means

3

that our Wright-Fisher sampling for n_o offspring selected exactly n_p distinct ancestors, of which j are derived. Under Wright-Fisher sampling, the order in which (previously unsampled) parental lineages are drawn is random. That is, we could have performed a random permutation of the parental population prior to starting the sampling process, and sampled new parental lineages in order from this permutation. Thus the event (j, n_p) can be reformulated as the joint events that the first n_p parental alleles from the random permutation carry j derived alleles *and* that exactly n_p distinct parents were drawn in the Wright-Fisher sampling of the first n_o offspring. Now, let $r(j, n_p)$ be the (less specific) event that the first n_p parental alleles from the random permutation carry j derived alleles. Then $P(r(j, n_p)) = \Phi_{n_p}^{(t)}(j)$, and we can write:

$$
\begin{aligned}
\Phi_{n_o}^{(t)}(i) &= \sum_{n_p, j} P_{n_o}(i | r(j, n_p), n_p) P_{n_o}(n_p | r(j, n_p)) P(r(j, n_p)) \\
&= \sum_{n_p=1}^{n_{p,max}} \sum_{j=0}^{n_p} P_{n_o}(i, n_p | r(j, n_p)) \Phi_{n_p}^{(t-1)}(j) \\
&\equiv \sum_{n_p=1}^{n_{p,max}} \mathbf{T}_{n_p, n_o} \Phi_{n_p}^{(t-1)}
\end{aligned}
\tag{2}
$$

where the $\mathbf{T}_{n_p, n_o}$ are $(n_o + 1) \times (n_p + 1)$ matrices whose row and column indices correspond to the number of derived alleles in the offspring and contributing parental lineages, respectively.

Under neutral evolution, $n_{p,max} \leq n_o$. Since we can obtain smaller *AFS* from larger *AFS* through downsampling (*i.e.*, $\Phi_{n_p}^{(t)} = \mathbf{H}_{n_p, n_o} \Phi_{n_o}^{(t)}$ for hypergeometric projection matrix $\mathbf{H}_{n_p, n_o}$ if $n_p < n_o$), Equation (2) provides a closed form recursion for Φ_{n_o}. This property was used in Jouganous et al. (2017) to efficiently compute distributions of allele frequencies under the large sample size limit.

Under selection, $n_{p,max}$ can be larger than n_o, leading to an infinite set of coupled equations. Our goal here is to take advantage of the fact that selective events can be treated exactly in the large sample size limit as long as $n_p < n_o$ or, equivalently, as long as there are more coalescent than selection events: $n_r \leq n_c$. If we have high confidence that this will be the case, we can restore closure by truncating Eq. (2):

$$\Phi_{n_o}^{(t)}(i) \simeq \sum_{n_p=1}^{n_o} \mathbf{T}_{n_p,n_o} \Phi_{n_p}^{(t-1)}$$

$$= \sum_{n_p=1}^{n_o} \mathbf{T}_{n_p,n_o} \mathbf{H}_{n_p,n_o} \Phi_{n_o}^{(t-1)} \tag{3}$$

$$\equiv \mathbf{Q}_{n_o} \Phi_{n_o}^{(t-1)}$$

A jackknife approximation can be used to simulate the drawing of a small number of additional lineages and improve upon this closure approximation ($\Phi_{n_p}^{(t)} \simeq J_{n_p,n_o} \Phi_{n_o}^{(t)}$ with $n_p > n_o$). Jouganous et al. (2017) used this to derive approximate recursion equations under weak selection. We will show below that closure is asymptotically maintained for large sample sizes even without requiring a jackknife approximation.

Our first goal is to obtain an explicit recursion for the matrices in (2). To do so, we will need to account for multiple coalescent events, which will require some careful bookkeeping.

2.1. Constructing the transition matrix

Even though T_{n_p,n_o} is a simple combinatorial probability describing a single generation, we were unable to compute an analytical expression for it while allowing for multiple coalescences and multiple selective events. However, we can obtain fairly simple recursions in terms of the sample size n_o, by assuming that we know T_{n_p,n_o-1} conditional on the last drawn offspring.

IK: Need to reconcile n_p (parents) vs n_c (contributors) notation.

Figure 2 shows the recursion equation in the selection case. The summands correspond to types of draws, represented graphically on the right. Filled circles correspond to derived alleles, empty to ancestral alleles. A line connecting two circles represents a successful draw, a capped broken line - a selective event. Double line represents potentially multiple draws. Square brackets represent the events in a smaller ($n_o - 1$) sample size.

The term $T_r \begin{bmatrix} i_c, n_c \\ i_o, n_o \end{bmatrix}$ represents the probability of drawing i_o derived alleles into the offspring generation from i_c derived in the parental generation, with exactly r selective events for that lineage. n_c and n_o are the sample sizes of parental and offspring generations, respectively.

Each of the summands (A,B,C,D) represents a particular type of a draw, whereas the recursive term describes the events in a smaller sample size. To account for potentially multiple selection

$$T_{r=0}\begin{bmatrix} i_c, n_c \\ i_o, n_o \end{bmatrix} = \left(1 - \frac{n_c-1}{N}\right)\frac{n_c-i_c}{n_c}\sum_{r=0}^{r_{max}} T_r \begin{bmatrix} i_c, n_c-1 \\ i_o, n_o-1 \end{bmatrix} \qquad (A)$$

$$+ \frac{n_c-i_c}{N}\sum_{r=0}^{r_{max}} T_r \begin{bmatrix} i_c, n_c \\ i_o, n_o-1 \end{bmatrix} \qquad (B)$$

$$+ \left(1 - \frac{n_c-1}{N}\right)\frac{i_c}{n_c}(1-s)\sum_{r=0}^{r_{max}}\left(\frac{s}{1-s}\right)^{\delta_{r,r_{max}}} T_r \begin{bmatrix} i_c-1, n_c-1 \\ i_o-1, n_o-1 \end{bmatrix} \qquad (C)$$

$$+ \frac{i_c}{N}(1-s)\sum_{r=0}^{r_{max}}\left(\frac{s}{1-s}\right)^{\delta_{r,r_{max}}} T_r \begin{bmatrix} i_c, n_c \\ i_o-1, n_o-1 \end{bmatrix} \qquad (D)$$

$$T_r\begin{bmatrix} i_c, n_c \\ i_o, n_o \end{bmatrix} = \left(1 - \frac{n_c-1}{N}\right)\frac{i_c}{n_c}sT_{r-1}\begin{bmatrix} i_c-1, n_c-1 \\ i_o, n_o \end{bmatrix} \qquad (rC)$$

$$+ \frac{i_c}{N}sT_{r-1}\begin{bmatrix} i_c, n_c \\ i_o, n_o \end{bmatrix} \qquad (rD)$$

Figure 2: Recursion construct for transition probability with selection, accounting for multiple kinds of coalescent events. The right hand panel represents summands on the left graphically. Filled and empty circles represent derived and ancestral alleles. Solid and broken lines are successful lineage draws and re-draws due to selection. Double lines represent potnetially multiple draws. Square brackets represent events in a smaller sample size ($n_o - 1$). Summands (A-D) are successful draws where the last lineage is ancestral (A,B) or derived (C,D). Note that these terms depend on the probability that there were between 0 to r_{max} in sample size $n_o - 1$. Terms rC and rD represent the probability that last selection re-draw was due to a lineage from outside (rC) or within (rD) the sample.

events per lineage, we also need to consider that the smaller sample size $n_o - 1$ had anywhere between 0, and r_{max} selective re-draw evants. The probability for r re-draws $T_r\begin{bmatrix} \cdot,\cdot \\ \cdot, n_o-1 \end{bmatrix}$ is the sum of terms rC and rD.

There are other ways to write down similar recursion equations. In the appendix, we show a simpler model for neutral alleles, and a more direct model where $r_{max} = 1$.

Using a dynamic programming algorithm, we implement the construction of the exact neutral transition matrices $\mathbf{T}$ in $O(n^3)$ operations. The set of selection transition probability matrices $\mathbf{T}$ requires $O(n^4 + r)$ operations.

In practice, we set the maximum number of re-draws per lineage to $r_{max} = 4$, which appears to be sufficient to consider strong negative selection up to $Ns = 50$.

Since these calculations are expensive, we also propose a number of approximations in the following sections.

2.2. Missing probability

Truncation of the recursion means that some events are not accounted for. Given j derived alleles in a sample of size n_o, probability lost due to truncation is simply $1 - \sum_i Q_{n_o}(i,j)$, with the maximum loss of probability occurring for $j = n_o$. Thus the maximum loss of probability occurs for very deleterious alleles at high frequency – which is an unusual occurrence. An alternative method of measuring lost probability is by weighing the missed probability by $\phi_{n_o}(j)$.

In numerical applications, we re-normalize each row of Q to ensure a proper probability transition, and we track $1 - \sum_i Q_{n_o}(i, n_o)$ to ensure that it remains below a threshold SG: What did we use?.

3. Results

3.1. Markov process construction

SG: Do we really need the following two paragraphs?

The sample size n_o can range between 1 and N, the size of the entire population. We expect that when $n \ll N$, the calculations with the present model do not differ substantially from the prediction under the standard coalescent, or (Jouganous et al., 2017). In the limit where $n \to N$, we expect the transition probabilities in SG: Call this matrix something else? R? T? $\mathbf{P}_s$ to approach those in the full Wright-Fisher model (*i.e.* (Ewens, 2004, eq. 1.58)).

Similar to (**??**), the transition probability matrix with selection ($\mathbf{P}_s$), involves considering a large number of intermediate lineages. In practice, we limit the number of selective events to at most one per lineage – which means that the maximum number of contributing lineages is $2n_o$ – one selective death for each offspring. This allows us to model strong selection, while limiting the number of coalescent configurations we need to consider when building $\mathbf{P}_s$.

3.2. Calculation of allele frequency spectra

Once the truncated matrix $\mathbf{Q}_s$ is constructed, it can be used to calculate the allele frequency spectrum. For example, in the infinite sites model at equilibrium, we can approximate the equilibrium AFS Φ as a solution to a linear system:

$$\Phi = Q\Phi + n\mu e_1 \tag{4}$$

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 (**??**), the diffusion approximation (Ewens, 2004, eq. 9.23), and the calculation performed in `Moments` (Jouganous et al., 2017). SG: Should we give an example for very large sample sizes on panels A and B to show the finite sample effect? IK: Is $n = 100, N = 100$ good here?. Panels A and B show the *AFS* under neutrality, C and D under strong negative selection. Under neutrality, all the models agree, yeilding similar *AFS*.

Figure 3C shows a comparison at strong negative $Ns = 50$, with the population size ($N = 1000$), which is substantially larger than the sample size ($n = 100$). There is a small deviation between the approaches at large allele frequencies. If the sample size is the same as the population size ($n = N = 100$) (Fig. 3D), the diffusion approximation and `Moments` depart from the Wright-Fisher prediction. The approach presented here (**??**) shows a better match to the Wright-Fisher model.

3.3. Closure properties

SG: I added the closure derivation higher up. So here we can focus on the results.

To investigate the closure properties of $\mathbf{P}_s$, we can calculate the total probability that more that n_o parental lineages contribute to a sample of size n_o. SG: I'm not sure I understand what you describe below. My take: Since $P_{n_o,n_p}(i,j)$ is a probability distribution over n_p and i, we can easily compute the probability lost to truncation as $1 - \sum_{i=0}^{n_o} \sum_{n_p=0}^{n_o} P_{n_o,n_p}(i,j)$.

By construction, the sum of rows of $\mathbf{P}_s$ should correspond to the total probability mass that included configurations contribute (Fig. 9). Thus, the probability that some number of configurations are unaccounted for, with j derived alleles in the parental sample, is given by $1 - \sum_{i=0}^{n} \mathbf{P}_s(i,j)$. 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.
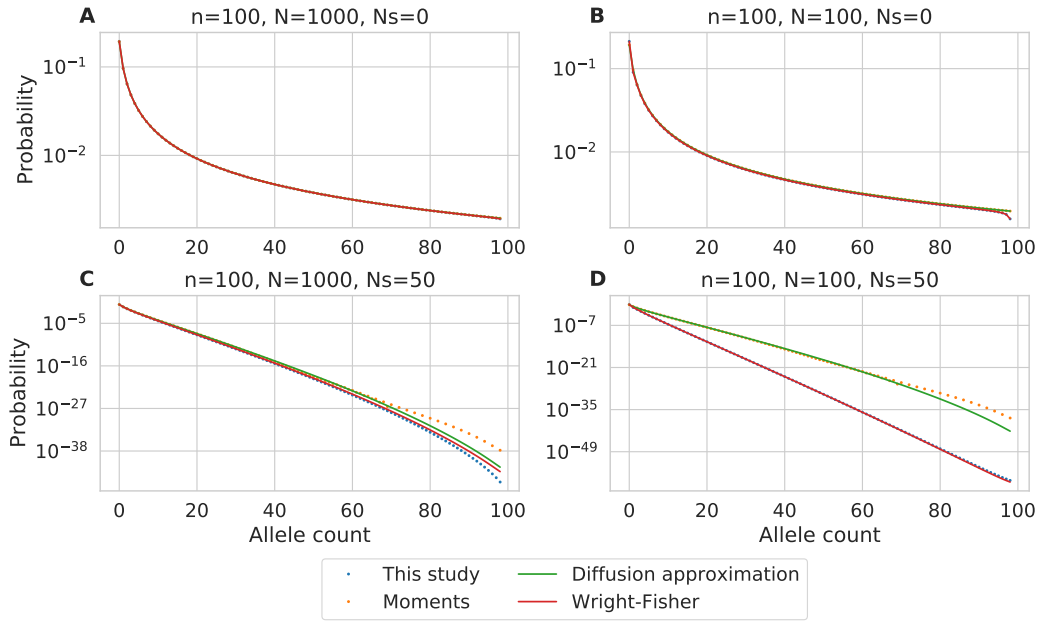
Figure 3: Normalized allele frequency spectra in a sample of size $n = 100$ (B,D), for neutral (Ns=0) (A,B), or highly deleterious alleles ($Ns = 50$) (C,D). (A) shows the frequency spectrum in a sample from a large population ($N = 1000$), (B) in a small population ($N = 100$). (C,D) same, for deleterious allele. Note that for a strongly deleterious allele, the probabilities are extremely small.
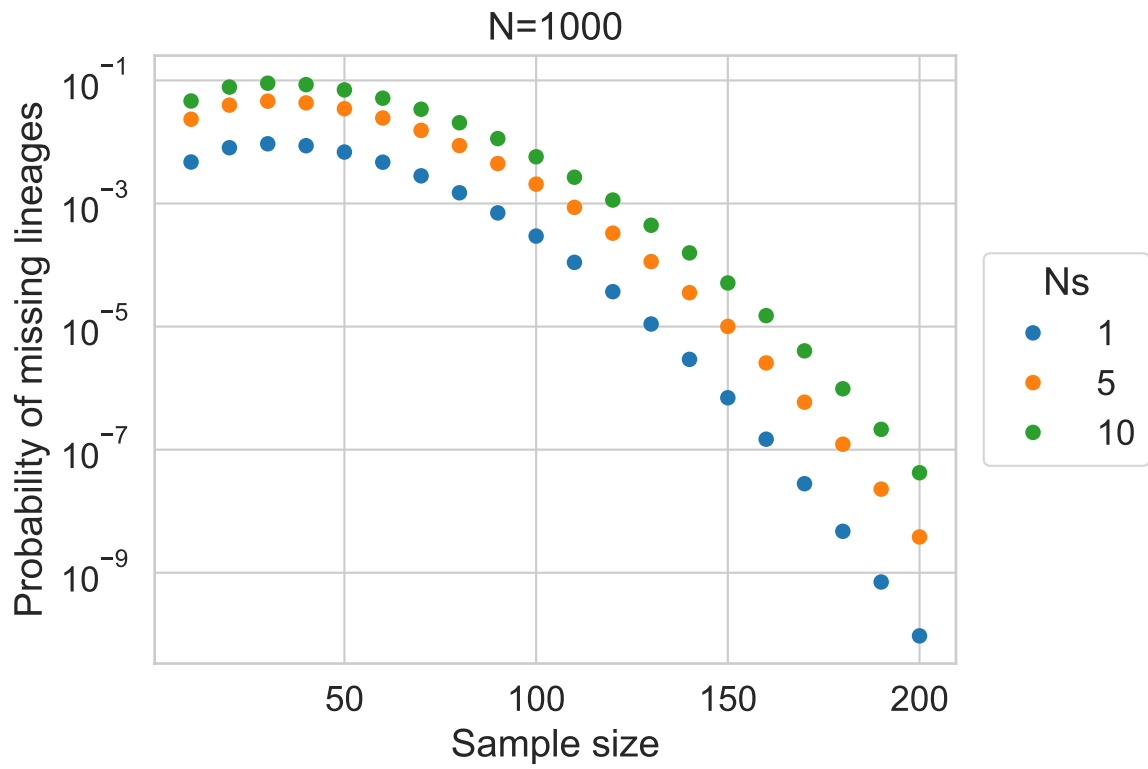
9

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: It could be cool to have both the worst case missing lineages (with $f = 1$) and the average number of missing lineages at equilibrium.

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 (**??**). We derive several approximations to the model proposed in the first section, in order to get a better understanding of this behavior.

In the following derivations, we are assuming that the derived allele is present at frequency x, as opposed to explicitly modeling the count of derive alleles **??**, which simplifies the calculations. When looking for the upper bound on the number of rejected lineages, we take $x = 1$, since only derived alleles experience selection.

4.1. Mean number of contributing lineages

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

$$Pr(n_p|n_o) = \sum_{n_g} Pr(n_p|n_g)Pr(n_g|n_o) \tag{5}$$

Where n_p and n_g is the number of contributing parents and gametes, respectively (Fig. 1B).

Using the law of total expectation, we can write the expectation $E[n_p - n_o|n_o]$ as

$$E[n_p - n_o|n_o] = E[n_p|n_o] - n_o$$
$$= E_{n_g}E_{n_p}[n_p|n_g] - n_o.$$

Assuming $x = 1$, the expectation over n_p simply given by the occupancy distribution .

$$\hat{E}[n_p - n_o|n_o] = E_{n_g}\left[N\left[1 - \left(1 - \frac{1}{N}\right)^{n_g}\right]\right] - n_o$$
$$= N - NE_{n_g}\left[\left(1 - \frac{1}{N}\right)^{n_g}\right] - n_o.$$

Since $n_g - n_o$ follows a negative binomial distribution with n_g trials and success rate $1 - s$, we can use the moment generating function to show that for any constant x,

$$E_{n_g}[x^{n_g}] = x^{n_o}\left(\frac{1 - s}{1 - sx}\right)^{n_o}. \tag{6}$$

Thus we can write

$$\hat{E}[n_p - n_o|n_o] = N - N\left(1 - \frac{1}{N}\right)^{n_o}\left(\frac{1 - s}{1 - s\left(1 - \frac{1}{N}\right)}\right)^{n_o} - n_o.$$

11

Taking the terms of order $\frac{1}{N}$ gives

$$\hat{E}[n_p - n_o|n_o] = n_o s - \frac{n_o(n_o - 1)}{2N}$$

We thus have the usual result that the increase of the number of lineages due to selection is linear, whereas the reduction due to drift is quadratic. Solving for $\hat{E}[n_p - n_o|n_o] < 0$ yields, to leading order,

$$n_o^* \geq 2Nxs. \tag{7}$$

Figure 5 shows the critical sample size for several selection coefficients, assuming the entirety of the sample is derived ($x = 1$) in a population of $N = 1,000$. The Y axis shows the fraction of contributing parental lineages to the sample size, $\frac{n_p}{n_o}$. Above the horizontal line $\frac{n_p}{n_o} > 1$, selection dominates. Below, drift reduces the number of used lineages. The intercept of the line with $\frac{n_p}{n_o} = 1$ is the critical sample size, which is well-approximated by $2Nsx$.

4.2. Distribution of number of contributing lineages

To ensure approximate closure of the recursion, we need a stricter constraint, that is, $n_p \leq n_o$ almost always (rather than on average. We therefore need to study the distribution $P(n_p|n_0)$.

We can obtain the variance of this distribution by using the law of total variance:

$$\mathrm{Var}\,[n_p - n_o] = \mathrm{Var}_{n_g}\,[E\,[n_p - n_o|n_g]] + E_{n_g}\,[Var\,[n_p - n_o|n_g]] \tag{8}$$

The expectation in the first term can be derived from the occupancy distribution and the identity 6:
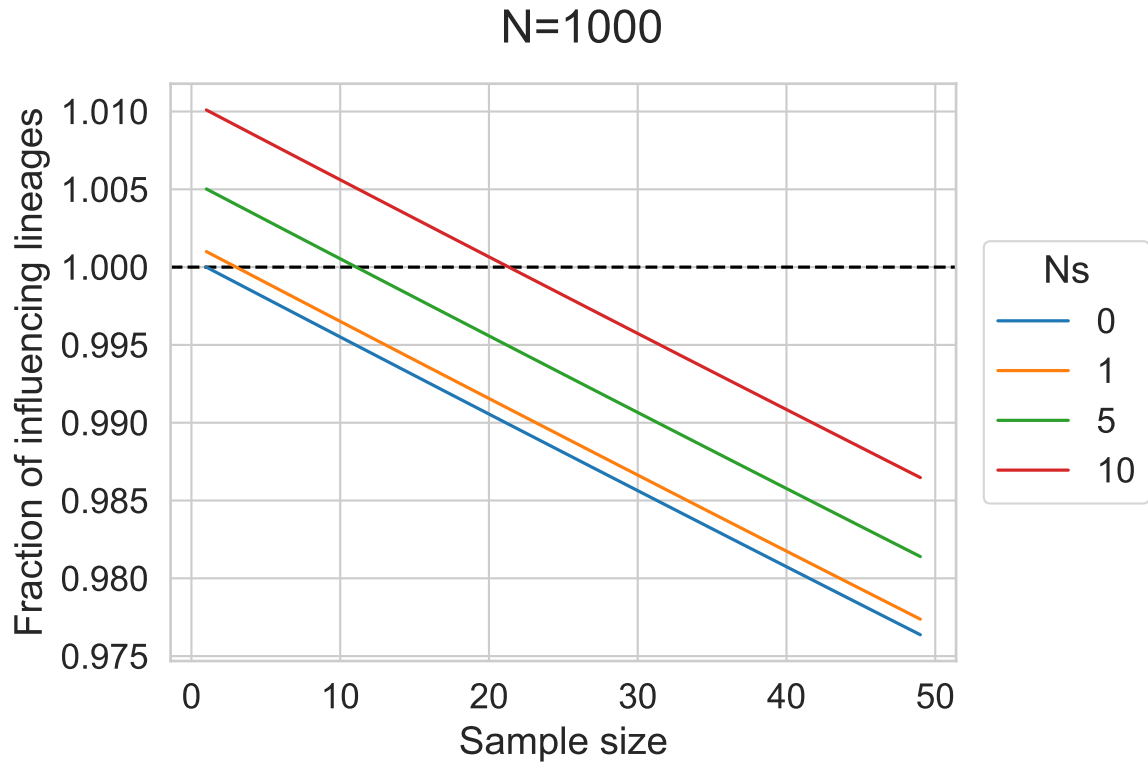
Figure 5: Critical sample size for different selection coefficients. The Y axis shows the fraction of parental lineages over the sample size, $\frac{n_p}{n_o}$, each line corresponds to a different selection coefficient. Above $\frac{n_p}{n_o} \geq 1$, selection dominates, below – drift. The critical sample size, where the expected number of parental contributing lineages is smaller than the sample size is well-approximated by $2Ns$.

$$\mathrm{Var}_{n_g}\left[E\left[n_p - n_o | n_g\right]\right] = \mathrm{Var}_{n_g}\left[E\left[n_p | n_g\right]\right]$$

$$= \mathrm{Var}_{n_g}\left[N\left(1 - (1 - \frac{1}{N})^{n_g}\right)\right]$$

$$= N^2\,\mathrm{Var}_{n_g}\left[(1 - \frac{1}{N})^{n_g}\right]$$

$$= N^2\left(E_{n_g}\left[(1 - \frac{1}{N})^{2n_g}\right] - E_{n_g}\left[(1 - \frac{1}{N})^{n_g}\right]^2\right)$$

$$= N^2\left(\left(1 - \frac{1}{N}\right)^{2n_o}\left(\frac{1-s}{1 - s\left(1 - \frac{1}{N}\right)^2}\right)^{n_o} - \left(1 - \frac{1}{N}\right)^{2n_o}\left(\frac{1-s}{1 - s\left(1 - \frac{1}{N}\right)}\right)^{2n_o}\right)$$

$$(9)$$

The variance of the second term is the variance of the occupancy distribution :

$$E_{n_g}\left[Var\left[n_p - n_o | n_g\right]\right] = E_{n_g}\left[N((N-1)(1 - 2/N)^{n_g} + (1 - 1/N)^{n_g} - N(1 - 1/N)^{2n_g})\right]$$

$$= N(N-1)\left(1 - \frac{2}{N}\right)^{n_o}\left(\frac{1-s}{1 - s\left(1 - \frac{2}{N}\right)}\right)^{n_o} + \left(1 - \frac{1}{N}\right)^{n_o}\left(\frac{1-s}{1 - s\left(1 - \frac{1}{N}\right)}\right)^{n_o}$$

$$- N\left(1 - \frac{1}{N}\right)^{2n_o}\left(\frac{1-s}{1 - s\left(1 - \frac{1}{N}\right)^2}\right)^{n_o}.$$

$$(10)$$

The two terms can be put together to provide an analytical expression for the variance of n_p.

The number of parental lineages used by drift can be modelled by the modified occupancy (Arfwedson) distribution (Wakeley, 2009; O'Neill, 2019; Johnson et al., 2005). This is given by:

$$P(n_p | n_g) = \frac{S_2(n_g, n_p)N!}{(N - n_p)!N^{n_g}} \tag{11}$$

where $S_2(n_g, n_p)$ is a Stirling number of the second kind, which is the number of ways to partition n_g gametes into n_p parents (see Johnson et al. (2005) section 10.4 for a thorough treatment). Note that the under drift, the number of parents will be smaller or equal to the number of gametes $n_p \leq n_g$.

The distribution of the number of gametes, n_g is given by the negative binomial, parameterized by probability of resample s, and the total number of trials before n_o successes (*i.e.* $n_r + n_o$):
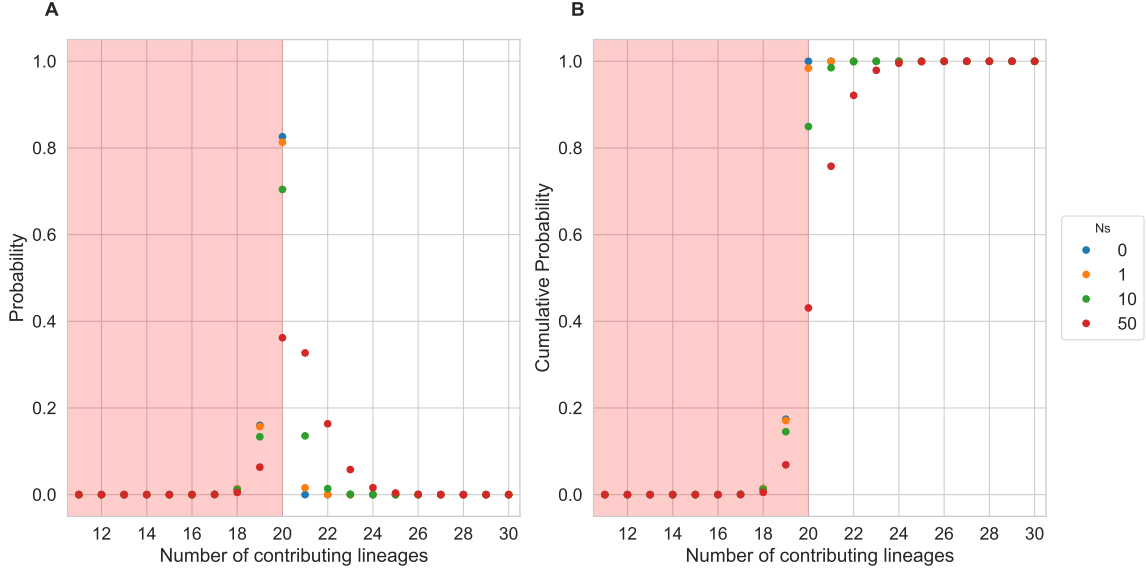
14

Figure 6: **A** The distribution and **B**, cumulative distribution of the number of parental contributing lineages one generation into the past ($n = 20$, $N = 1000$). Shaded red area shows the drift-dominated regime, where the number of lineages is smaller than the sample size.

$$P(n_g|n_o) = \binom{n_g - 1}{n_o - 1}(1 - xs)^{n_o}(xs)^{n_g - n_o} \tag{12}$$

Here, the number of gametes can be larger that the sample size $n_o \leq n_g$, if selection is present. Combining the two distributions together through 5, we get:

$$Pr(n_p|n_o) = \sum_{n_g=1}^{\infty} \frac{S_2(n_g, n_p)N!}{(N - n_p)!N^{n_g}} \binom{n_g - 1}{n_o - 1}(1 - xs)^{n_o}(xs)^{n_g - n_o} \tag{13}$$

This distribution does not appear to have a simple analytical form. However, it can be computed efficiently using methods presented in (O'Neill, 2019). Figure 6 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 increased, we begin requiring larger number of lineages.

We defined the critical sample size as $n_o^* > E[n_p|n_o]$. However, the distributions in 6 show that there is a large probability that $n_p > n_o$ at $n_o^* = 2n = 20$. In order to guarantee that drift

15

will out-pace selection, we can calculate the cumulative distribution. This implies that a sample size in which the *majority* of lineages are accounted for can be substantially larger than the critical sample size of equation (7). To derive a convenient analytical approximation, we turn to the normal approximation in the next section.

4.3. Normal approximation

We can construct a normal approximation to the distribution of the number of contributing lineages. This approximation will allow us to calculate sample size where, for example, the number of contributing lineages is smaller that the sample size 95% of the time, instead of 50%, as given by n_o^* in equation (7).

The occupancy distribution is approximated by the normal (O'Neill, 2019) when $n_o \ll N$. Likewise, the number of failures (n_r) 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:

$$Pr(\mathcal{R} = r|n) \approx \mathcal{N}(\mu = [(sn)/(1-s) + N(1 - (1 - 1/N)^n)], \tag{14}$$

$$\sigma = \sqrt{N\left((N-1)\left(1 - \frac{2}{N}\right)^n + \left(1 - \frac{1}{N}\right)^n - N\left(1 - \frac{1}{N}\right)^{2n}\right) + \frac{ns}{(1-s)^2}}) \tag{15}$$

Figure 7 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.

4.4. Integrating over few generations

To increase the chances that the number of lineages ancestral to a sample of size n_o is less than n_o, we can also chose to compute a transition matrix over more than a single generation.
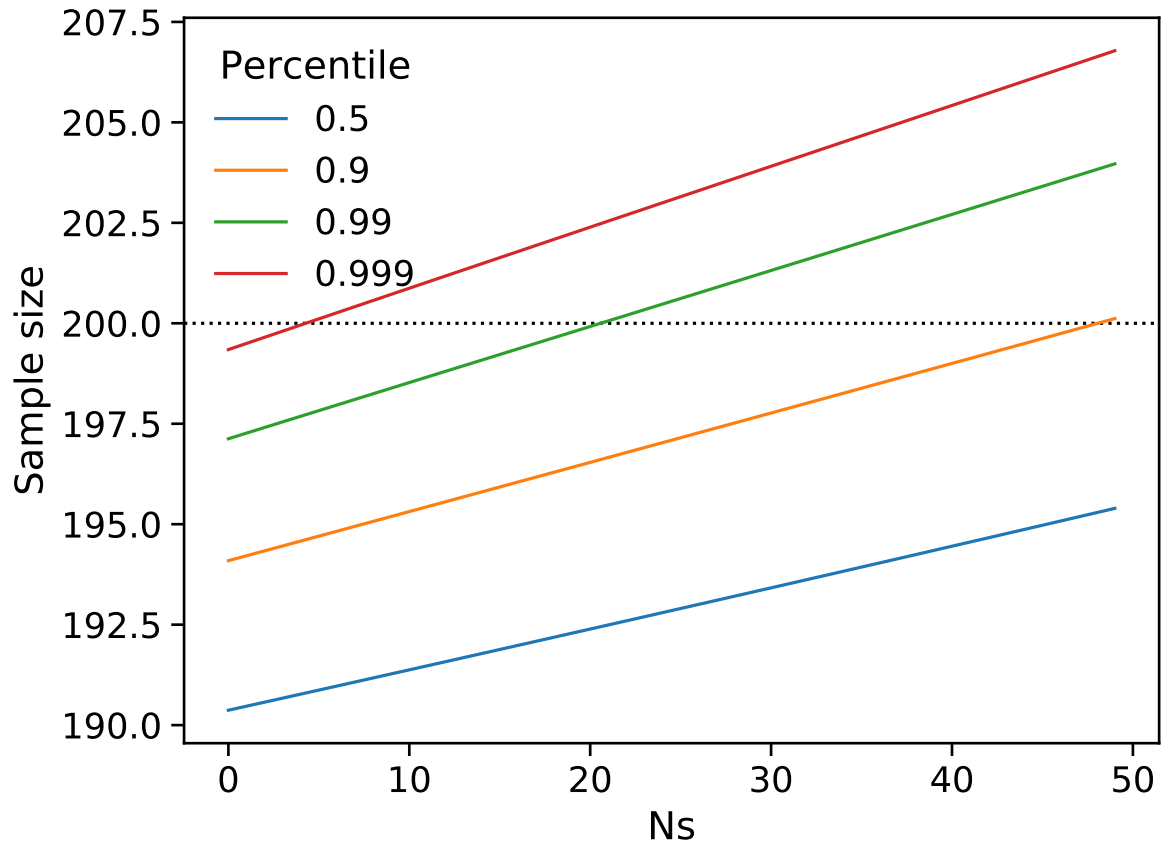
Figure 7: IK: TODO: need to resolve this still The quantile function of the closure of the sample (14). Each line corresponds to different percentile of the normal approximation. Black dashed line shows the reference sample size $n_o^* = 200$ SG: does it play a special role? If not why mention it (or have this line, really)?. SG: It also seems like showing the cumulative distributions themselves would be more intuitive. E.g $\log(missingp)$. Also would be nice to have the numerical calculation. Could you get the cumulative distribution for the occupancy distribution from the Oneil algorithm?

In this case, lineages gained by selective death at one generation can can be compensated by loss to genetic drift at another generation. If the expected number of drift events is higher than the expected number of selective deaths, this can reduce the likelihood of chance events. Asymptotic results suggest ...

Such transition matrices can also be computed iteratively. If n_g is the number of contributing lineages in the grandparental generation, and k is the number of derived among those, and we define $P(i, n_g | q(k, n_g))$, as above, as the probability of drawing i derived alleles and using exactly n_g grandparental alleles given the event $q(k, n_g)$ k of the first n_g grandparental alleles are derived, we can condition on the number of contributing alleles n_c and the number of derived alleles j among them in the parental generation:

$$
\begin{aligned}
P(i, n_g | r(k, n_g)) &= \sum_{j, n_c} P(i, n_g; j, n_c r(j, n_c) | q(k, n_g)) \\
&= \sum_{j, n_c} P(i, n_c, r(j, n_c); n_g | q(k, n_g)) \\
&= \sum_{j, n_c} P(i, n_c | r(j, n_c); n_g; q(k, n_g)) P(r(j, n_c); n_g | q(k, n_g)) \qquad (16) \\
&= \sum_{j, n_c} P(i, n_c | r(j, n_c)) P(r(j, n_c); n_g | q(k, n_g)) \\
&= \sum_{j, n_c} P(i, n_c | r(j, n_c)) P(r(j, n_c); n_g | q(k, n_g))
\end{aligned}
$$

and SG: despite the crap notation, this is just the product of quantities we have already computed.

5. Conclusion

Classically, the coalescent considers models in the absence of natural selection. Since selection can increase the number of contributing lineages back in time, the coalescent can no longer be represented by trees, but instead acquires a graph structure. The ancestral selection graphs (Krone and Neuhauser, 1997) deal with this in the limit of large population size (N).

The large population size approximation implies that the sample size n is much smaller than the whole population ($n \ll N$), so it is unlikely that more than one coalescent event will happen per generation. However, recent work (Bhaskar et al., 2014; Nelson et al., 2019) pointed out that this assumption is unreasonable with sample sizes pertinent to modern experiments. As a results,

models that consider multiple coalescent events per generation are gaining increased relevance in the field (**?**).

In this work we show that increasing the sample size has another unexpected consequence. As sample size increases, the larger number of lineages needed due to selection can be masked by coalescent events. In this sense, the large sample size rescues the model from effect of selection. This means that recursion equations needed to calculate sample properties are asymptotically closed with large population size.

At first approximation, $n_o^* = 2Nsx$ is a critical sample size, where the decrease of lineages due to coalescent back in time out-competes the increase due to selection (eq. (7)). Further, we derive the full probability distribution for the number lineages needed with given selection coefficient and sample size (eq. (13)). Unfortunately, the distribution does not have a closed form, so we derive a normal approximation to the number of lineages that contribute to a sample (eq. (14)). The normal approximation then allows us to get a quantile function that we use to find if the model preserves closure with some confidence level.

This work has several implications. First, we can combine the model described here with the jackknife approximation (Jouganous et al., 2017). This will allow us to construct a more robust inference framework that can account for large sample size and strong selection.

Further, the results here suggest that effect of weak selection may be detectable in studies with large sample sizes. This may open up a way for new investigations of natural selection in population genetics.

References

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

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

Ewens, W.J., 2004. Mathematical Population Genetics: I. Theoretical Introduction.. volume 27 of *Interdisciplinary Applied Mathematics*. 2 ed., Springer New York, New York. OCLC: 958522782.

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

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

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

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

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

Krone, S.M., Neuhauser, C., 1997. Ancestral processes with selection. Theoretical Population Biology 51, 210–237. doi:`10.1006/tpbi.1997.1299`.

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

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

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

6. Appendix

6.1. Table of symbols

Symbol	Generation	Meaning
p	$t-1$	Parental generation
c	$t-\frac{1}{2}$	Contributing (intermediate, selection only)
o	t	Offspring (current) generation

Symbol	Meaning
n	Sample size
i	Number of derived alleles in sample n.
$\dfrac{i}{n}$	i derived *out of* n total alleles

6.2. Neutral case

We want to construct the entries in the probability matrix $P\left[\left.\dfrac{\cdot}{n_o}\right|\dfrac{\cdot}{n_p}\right]$ in terms transition probabilities in smaller sample sizes. Under neutrality, the number of contributing parental lineages n_p' (at $t-1$) can not be larger than the number of offspring lineages n_o. Since $\mathbf{P}$ is square, $max(n_p) = max(n_o) = n$. Thus, our aim is to express every entry of $\mathbf{P}$, $P\left[\left.\dfrac{\cdot}{n}\right|\dfrac{\cdot}{n}\right]$, in terms of $P\left[\left.\dfrac{\cdot}{n-1}\right|\dfrac{\cdot}{n-1}\right]$. Since we never require extra lineages ($n_p \leq n_o$), the recurrence is closed.

To calculate the transition probabilities, we first condition on the state of the last parental allele drawn, and then on the coalescent event that last offspring lineage participates in. The recurrence is shown in figure 8. The panel on the right depicts the coalescent event for each term. Empty circles represent ancestral alleles, filled circles – derived. The square box represents a sample of size $n-1$. The first three terms in the sum correspond to the cases where the last parent that we drew was ancestral, last three – derived. SG: Would it be worth presenting the non-square transition probabilites as well to prepare the reader for what comes next with selection

When calculating a single entry in 8, the variables have the following ranges.

$$
\begin{aligned}
n_p &= n \\
i_p &\in [0, n] \\
n_o &= n \\
i_o &\in [0, n]
\end{aligned}
\tag{17}
$$

$$P\left[\frac{i_o}{n_o}\middle|\frac{i_p}{n_p}\right] = \left(\frac{n-i_o}{n}\right)\Bigg\{ \left(1-\frac{n-1}{N}\right)P\left[\frac{i_o}{n_o-1}\middle|\frac{i_p}{n_p-1}\right]$$

$$+ \left(\frac{i_o}{N}\right)P\left[\frac{i_o-1}{n_o-1}\middle|\frac{i_p}{n_p-1}\right]$$

$$+\left(\frac{n-i_o-1}{N}\right)P\left[\frac{i_o}{n_o-1}\middle|\frac{i_p}{n_p-1}\right]\Bigg\}$$

$$\left(\frac{i_o}{n_o}\right)\Bigg\{ \left(1-\frac{n-1}{N}\right)P\left[\frac{i_o-1}{n_o-1}\middle|\frac{i_p-1}{n_p-1}\right]$$

$$+ \left(\frac{i_o-1}{N}\right)P\left[\frac{i_o-1}{n_o-1}\middle|\frac{i_p-1}{n_p-1}\right]$$

$$+ \left(\frac{n-i_o}{N}\right)P\left[\frac{i_o}{n_o-1}\middle|\frac{i_p-1}{n_p-1}\right]\Bigg\}$$

Last parent is ancestral

Last parent is derived

Figure 8: Recurrence defining transition probabilities in a model without selection. Right panel shows coalescent events corresponding to each summand. Each transition probability is defined in terms of transition in a smaller sample size. First three terms are conditional on the last parent having an ancestral state, last three – derived. Filled circles – derived alleles; empty circles - ancestral alleles; square brackets – sample of size $n-1$.

The recurrence is calculated while $n > 1$, with the following base cases:

$$P\begin{bmatrix} 1 & \Big| & 1 \\ 1 & \Big| & 1 \end{bmatrix} = 1$$

$$P\begin{bmatrix} 0 & \Big| & 0 \\ 1 & \Big| & 1 \end{bmatrix} = 1$$

$$P\begin{bmatrix} 0 & \Big| & 1 \\ 1 & \Big| & 1 \end{bmatrix} = 0$$

$$P\begin{bmatrix} 1 & \Big| & 0 \\ 1 & \Big| & 1 \end{bmatrix} = 0$$

6.3. Selection case

Due to selective deaths, the number of lineages (n_c) that contribute to the current generation can be larger than the number of offspring (n_o), and especially so with strong selection. Because the number of sampling configurations can be large, we use dynamic programming to estimate $\mathbf{P}_{n_p, n_o}$ by summing over the possibilities for the last successful draw. Using the probability interpretation of the transition matrix, $\mathbf{P}_{n_p, n_o}(i, j) = P(i, n_p | r(j, n_p))$, the probability that we draw i derived offspring and exactly n_p parental offspring given that j of the first n_p sampled parental alleles are derived. The last successful draw event can be specified by the number $t \in \{0, \infty\}$ of prior failed draws due to selection since the last successful draw, the allele $a \in A, D$ selected, and the event c of whether or not the sampled parental allele was previously drawn $c \in \{True, False\}$. We also consider the event $s \in \{True, False\}$ of whether the last draw was successful. Finally, let us define the event $E_{n_o, t}(i, n_p)$ that we have drawn i derived offspring among n_o successful draws followed by t failures, and that this required exactly n_p parental lineages.

$$P(i, n_p | r(j, n_p)) = P(E_{n_0, 0}(i, n_p) | r(j, n_p)) = \sum_{a, c, t} P(a, c, t; E_{n_o, 0}(i, n_p) | r(j, n_p)) \tag{18}$$

23

Let us consider the term $a = A$, $c = False$ SG: We could use a better notation here, eg using tikz.

$$P(a = A, c = False, t; E_{n_o,0}(i, n_p)|r(j, n_p)) = P(a = A, c = False, t, s = True; E_{n_o,t}(i, n_p - 1)|r(j, n_p))$$

$$= P(a = A, c = False, t; E_{n_o,t}(i, n_p - 1)|r(j, n_p))$$

$$= P(a = A, c = False, t; E_{n_o,t}(i, n_p - 1); r(j, n_p - 1)|r(j, n_p))$$

$$= P(a = A, c = False, t; E_{n_o,t}(i, n_p - 1)|r(j, n_p - 1)r(j, n_p))P(r(j, n_p - 1)|r($$

$$= P(c = False, t; E_{n_o,t}(i, n_p - 1)|r(j, n_p - 1)r(j, n_p))\frac{n_p - j}{n_p}$$

$$= P(c = False|t; E_{n_o,t}(i, n_p - 1)|r(j, n_p - 1)r(j, n_p))P(t; E_{n_o,t}(i, n_p - 1)|r(j,$$

$$= \left(1 - \frac{n_p - 1}{N}\right) P(t; E_{n_o,t}(i, n_p - 1)|r(j, n_p - 1)r(j, n_p))\frac{n_p - j}{n_p}$$

$$= \left(1 - \frac{n_p - 1}{N}\right) P(t; E_{n_o,t}(i, n_p - 1)|r(j, n_p - 1))\frac{n_p - j}{n_p}$$

$$(19)$$

where the fourth line uses Bayes rule, and most other lines are exercises in rewriting the same event in different ways. Other combinations of a and c also yield expressions in terms of probabilities $P(t; E_{n_o,t}(i, n_p)|r(j, n_p))$ for the state prior to the successful draw SG: write down final results?.

These can be similarly expressed as recursions over the last draw. Selection only affects derived alleles, but it can occur after both coalescence and non-coalescence events.

$$P(t; E_{n_o,t}(i, n_p)|r(j, n_p)) = \sum_c P(c, t; E_{n_o,t}(i, n_p)|r(j, n_p)). \tag{20}$$

For example, the $c = True$ term can be written as

$$P(c = True, t; E_{n_o,t}(i, n_p)|r(j, n_p)) = P(c = True, a = D, s = False, t; E_{n_o,t}(i, n_p)|r(j, n_p))$$

$$= sP(c = True, a = D, t; E_{n_o,t-1}(i - 1, n_p)|r(j, n_p)) \tag{21}$$

$$= s\frac{j}{N}P(t; E_{n_o,t-1}(i - 1, n_p)|r(j, n_p)),$$

SG: I think this might want to be:

$$P(c = True, t; E_{n_o,t}(i, n_p)|r(j, n_p)) = P(c = True, a = D, s = False, t; E_{n_o,t}(i, n_p)|r(j, n_p))$$

$$= sP(c = True, a = D, t; E_{n_o,t-1}(i, n_p)|r(j, n_p)) \tag{22}$$

$$= s\frac{j}{N}P(t; E_{n_o,t-1}(i, n_p)|r(j, n_p)),$$

and similarly for $c = False$ SG: Write out? TODO, not complete.

$$P(c = False, t; E_{n_o,t}(i, n_p)|r(j, n_p)) = P(c = False, a = D, s = False, t; E_{n_o,t}(i, n_p)|r(j, n_p))$$

$$= sP(c = False, a = D, t; E_{n_o,t-1}(i, n_p - 1)|r(j, n_p)) \qquad (23)$$

$$= s\frac{j}{N}P(t; E_{n_o,t-1}(i, n_p)|r(j, n_p)),$$

Putting this all together SG: pseudocode?, we can perform an iteration over all n_o. For each n_o, we will compute all terms of the form $P(E_{n_0,0}(i, n_p)|r(j, n_p))$, for $i \in \{0, \dots, n_0\}$, $j \in \{0, \dots, n_p\}$, and $n_p \in \{1, \dots, n_{p,max}\}$. We further need to iterate over the possible number of failed selective events. If we only allow a maximum amount of failed selected events of t_max for each successful draw, the number of terms we must compute is of order $t_max n_p^4$. The number of computations for each term is constant and only depends on previously computed terms.

To ensure that probabilities do sum to one despite the t_max cutoff, we modify the Wright-Fisher model by imposing a successful draw after $t_max - 1$ attempt. Thus terms

$P(a = D, c, t_{max}; E_{n_o,0}(i, n_p)|r(j, n_p))$ will lose a factor $(1 - s)$.

We use n_c, the intermediate number of lineages at time $t - \frac{1}{2}$, which can potentially be much larger than the number of parents, n_p. This is analogous to the gamete intermediates, as presented in the main text. However, the two are not equivalent, since in this formulation we apply selection *and* drift on the intermediate lineages. We model the intermediate contributing alleles as a random sample from n_p alleles, without replacement.

$$P_s\begin{bmatrix} \dfrac{i_o}{n} & \dfrac{i_p}{n} \end{bmatrix} = \sum_{i_c, n_c} P_s\begin{bmatrix} \dfrac{i_o}{n} & \dfrac{i_c}{n_c} \end{bmatrix} P_s\begin{bmatrix} \dfrac{i_c}{n_c} & \dfrac{i_p}{n} \end{bmatrix} \qquad (24)$$

The probability conditional on the contributing lineages ($P_s\begin{bmatrix} \dfrac{i_o}{n} & \dfrac{i_c}{n_c} \end{bmatrix}$) is given by equation 9, while $P_s\begin{bmatrix} \dfrac{i_c}{n_c} & \dfrac{i_p}{n} \end{bmatrix}$ is given by the hypergeometric distribution. The support of the hypergeometric distribution means that we can not have $n_c > n$. Note that while $i_c \leq n_c \leq n$, we can still have $i_c > i_p$ if i_p is small. A formulation where a n_c is potentially infinitely large will be desirable.

Under the current definition, P_s is not closed, since the cases where $n_c > n$ are not accounted for. However, as we show in the main text, the formulation is asymptotically closed, as n increases.

The recursive definition in 9 is analogous to the neutral case, and gives $P_s\begin{bmatrix} \dfrac{i_o}{n} & \dfrac{i_c}{n_c} \end{bmatrix}$, the probability that i_o out of n lineages are derived, given that i_c out of n_c contributed to it. To construct

this probability, we condition on the coalescent events involving the last offspring allele. We limit the model to at most 1 selective death per lineage. However, in the entire sample, there still can be a large number of selective deaths. There are 6 distinct coalescent events with 0 or 1 selective deaths, with distinct probabilities based on whether the last offspring allele is ancestral or derived. This gives 12 different cases:
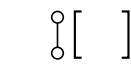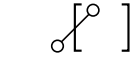
For each calculation, the ranges of the variables are:

$$
\begin{aligned}
n_p &= n \\
i_p &\in [0, n] \\
n_c &\in [1, n] \\
i_c &\in [0, n_c]
\end{aligned}
\tag{25}
$$
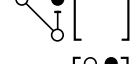
Note that unlike in the neutral case n_c is now variable. The base cases of the recurrence are:
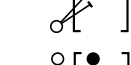
$$
\begin{aligned}
P \begin{bmatrix} 1 & 1 \\ \hline 1 & 1 \end{bmatrix} &= 1 - s \\
P \begin{bmatrix} 0 & 0 \\ \hline 1 & 1 \end{bmatrix} &= 1 \\
P \begin{bmatrix} 1 & 2 \\ \hline 1 & 2 \end{bmatrix} &= s \\
P \begin{bmatrix} 0 & 1 \\ \hline 1 & 2 \end{bmatrix} &= \frac{1}{s} \\
\text{otherwise} \quad & 0
\end{aligned}
$$

Coalescent
event
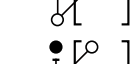
$$P_s\left[\frac{i_o}{n_o}\middle|\frac{i_c}{n_c}\right] =$$

$$\left(1 - \frac{n_c-1}{N}\right)\frac{n_c-i_c}{n_c}P_s\left[\frac{i_o}{n_o-1}\middle|\frac{i_c}{n_c-1}\right]$$

$$+\frac{n_c-i_c}{N}P_s\left[\frac{i_o}{n_o-1}\middle|\frac{i_c}{n_c}\right]$$

$$+\left(1-\frac{n_c-2}{N}\right)\frac{i_c}{n_c}s\left(1-\frac{n_c-1}{N}\right)\frac{n_c-i_c}{n_c-1}P_s\left[\frac{i_o}{n_o-1}\middle|\frac{i_c-1}{n_c-2}\right]$$

$$+\frac{i_c}{N}s\frac{n_c-i_c}{N}P_s\left[\frac{i_o}{n_o-1}\middle|\frac{i_c}{n_c}\right]$$

$$+\frac{i_c}{N}s\left(1-\frac{n_c-1}{N}\right)\frac{n_c-i_c}{n_c}P_s\left[\frac{i_o}{n_o-1}\middle|\frac{i_c}{n_c-1}\right]$$

$$+\left(1-\frac{n_c-1}{N}\right)\frac{i_c}{n_c}s\left(\frac{n_c-i_c}{N}\right)P_s\left[\frac{i_o}{n_o-1}\middle|\frac{i_c-1}{n_c-1}\right]$$

$$+\left(1-\frac{n_c-1}{N}\right)\frac{i_c}{n_c}(1-s)P_s\left[\frac{i_o-1}{n_o-1}\middle|\frac{i_c-1}{n_c-1}\right]$$

$$+\frac{i_c}{N}(1-s)P_s\left[\frac{i_o-1}{n_o-1}\middle|\frac{i_c}{n_c}\right]$$

$$+\left(1-\frac{n_c-2}{N}\right)\frac{i_c}{n_c}s\left(1-\frac{n_c-1}{N}\right)\frac{i_c-1}{n_c-1}P_s\left[\frac{i_o-1}{n_o-1}\middle|\frac{i_c-2}{n_c-2}\right]$$

$$+\frac{i_c}{N}s\frac{i_c-1}{N}P_s\left[\frac{i_o-1}{n_o-1}\middle|\frac{i_c}{n_c}\right]$$

$$+\frac{i_c}{N}s\left(1-\frac{n_c-1}{N}\right)\frac{i_c-1}{n_c}P_s\left[\frac{i_o-1}{n_o-1}\middle|\frac{i_c-1}{n_c-1}\right]$$

$$+\left(1-\frac{n_c-1}{N}\right)\frac{i_c}{n_c}s\frac{i_c-1}{N}P_s\left[\frac{i_o-1}{n_o-1}\middle|\frac{i_c-1}{n_c-1}\right]$$

$$+\left(1-\frac{n_c-1}{N}\right)\frac{i_c}{n_c}s\frac{1}{N}P_s\left[\frac{i_o-1}{n_o-1}\middle|\frac{i_c-1}{n_c-1}\right]$$

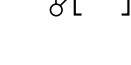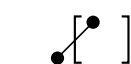$$+\left(\frac{i_c}{N}\right)s\frac{1}{N}P_s\left[\frac{i_o-1}{n_o-1}\middle|\frac{i_c}{n_c}\right]$$
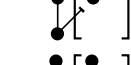
Last offspring is ancestral

Last offspring is derived

Derived, resample

Figure 9: Recurrence defining transition probabilities in a model with selection SG: I think there are some problems with the way fractions are defined, e.g. the first term should start with $1-\frac{n_c-1}{N}$, not $\frac{1-(n_c-1)}{N}$. Right panel shows coalescent events corresponding to each summand. Each transition probability is defined in terms of transition in a smaller sample size. First six terms are conditional on the last offspring having an ancestral state, last six – derived. Filled circles – derived alleles; empty circles - ancestral alleles; square brackets – smaller sample.

27