Models of strong selection in large samples

Ivan Krukov, Simon Gravel

Abstract

Neutral models of genetic diversity tend to be easier to analyze than models including selection. Under the neutral Wright-Fisher model, the number of lineages that contribute to ancestry of a sample decreases back in time due to coalescent events. 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.

haha!

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, due to coalescent events back in time (Kingman, 1982). To describe a sample of size n, we need to recursively consider samples of size $n' \leq n$. This does not hold under negative selection – due to selective deaths, the number of parental lineages n' can be larger than n. As we demonstrate later, this leads to a potentially infinite number of terms in the equations. This is similar to the ancestral selection graphs (ASG), (Krone and Neuhauser, 1997), where the number of relevant lineages can increase back in time.

The interplay of drift and selection is important to consider. In large sample sizes, there are many more common ancestry events than selective deaths, so the number of contributing lineages is unlikely to increase back in time. This suggests that large sample sizes can lead to almost-closed recursion equations, as we will demonstrate here.

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 behavior and applications for modeling 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 sample with n_o offspring lineages at time t, we will be looking for recursion equations for

the allele frequency distribution by considering the sampling process for a finite sample under drift and selection.

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 (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.
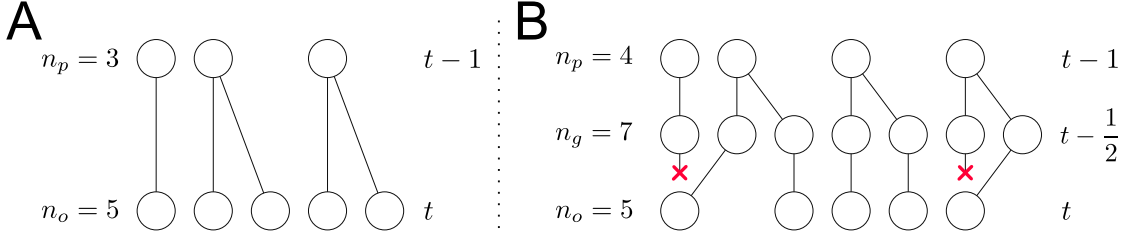

Figure 1: Realizations of sampling parental lineages under neutrality (A) and selection (B). **A** Under neutrality, possible coalescent events imply that number of parental lineages n_p at $t - 1$ is less than or equal to n_o offspring lineages at t. **B** With selection, we add an intermediate gamete 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 coalescences. 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 $\Phi_{n_o}^{(t)}$ in a sample size n_o at time t in terms of the parental AFS $\Phi_{n_p}^{(t-1)}$, we take advantage of the exchangeability of the parents in the drawing process. If we perform the Wright-Fisher sampling sequentially, one offspring at a time, the order in which (previously unsampled) parental lineages are drawn is random. That is, we could have selected a random permutation of the parental population prior to starting the sampling process, and decided to sample new parental lineages in order from this permutation. We can therefore condition on the random result of this permutation.

The probability that we draw i derived alleles in the offspring can be written as a sum over

two intermediate random variables: the number n_p of distinct parental lineages drawn, and the number j of derived parental lineages in a random sample of n_p parental lineages. The event $r(j,n)$ that we draw j derived alleles in a random sample of n parental lineages has probability $P(r(j,n)) = \Phi_n^{(t-1)}(j)$. We can therefore write

$$\Phi_{n_o}^{(t)}(i) = \sum_{n_p,j} P(i, n_p, r(j, n_p)) = \sum_{n_p,j} P(i, n_p | r(j, n_p)) \Phi_{n_p}^{(t-1)}(j).$$

SG: Alternative derivation: To estimate the allele-frequency spectrum $\Phi_{n_o}^{(t)}$ in a sample size n_o at time t, 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:

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

The event (j, n_p) means that our Wright-Fisher sampling for n_o offspring selected exactly n_p distinct ancestors, of which j are derived. If we perform the Wright-Fisher sampling sequentially, one offspring at a time, 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 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. Let $r(j, n_p)$ be the 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

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

SG: check if AFS is defined as probability or counts, in which case there is a scaling factor where $r(j, n)$ is the event that j out of the first n sampled alleles in the population are derived. $r(j, n_p)$ follows a hypergeometric distribution and is distinct from j, the event of drawing j parental derived alleles in the Wright-Fisher sampling process for the first n_o offspring. In words, the probability that we draw exactly n_p parental alleles, of which j are derived, is equal to the probability that a random subset of n_p alleles drawn from the ancestral population contains j derived alleles, times

4

the probability that we draw exactly n_p parental samples in that case. Thus we can write

$$\Phi_{n_o}^{(t)}(i) = \sum_{n_p=1}^{n_{p,max}} \sum_{j=0}^{n_p} P(i, n_p | r(j, n_p)) \Phi_{n_p}^{(t-1)}(j). \tag{1}$$

Under neutral evolution, $n_{p,max} \leq n_0$. Since we can obtain smaller AFS from larger AFS through downsampling (i.e., $\Phi_{n'}^{(t)} = P_{n',n} \Phi_n^{(t)}$ for hypergeometric projection matrix SG: We need to use a variable other than P for all these matrices and probabilities :) $P_{n',n}$ and $n' < n$) (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. A jackknife approximation can be used to simulate the drawing of additional lineages while preserving closure. Jouganous et al. (2017) also derived approximate recursion equations under selection, but these required multiple approximations: because Jouganous et al worked in a limit where the number of coalescent and selection events per generation is much smaller than one, every selective event led to $n_p > n_o$, requiring a closure approximation. 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 $n_s - n_d \leq 0$.

In the large population size and weak selection limit, the probability of observing a coalescence event in a generation is $\frac{n_o(n_o-1)}{2N}$, whereas the probability of observing a selection event is $2n_o f_p s$, with f_p the parental deleterious allele frequency. These rates are familiar from the ancestral selection graph (Krone and Neuhauser, 1997). Since the probability of coalescence events grows quadratically with the sample size and the rate of selective events grows linearly, we expect more drift events than selective events when $n_o > 2N f_p s$.

Our goal is to obtain an explicit recursion from equation (2) and show that its asymptotic behavior for large n_o allows for almost-exact solutions even in the strong selection regime. To do so, we will need to account for multiple coalescent events, which will require some careful bookkeeping.

2.1. Constructing the transition matrix

We can write Equation (2) in matrix form as SG: placeholder notation

$$\Phi_{n_o}^{(t)} = \sum_{n_p=1}^{n_{p,max}} \mathbf{P}_{n_p,n_o} \Phi_{n_p}^{(t-1)}. \tag{2}$$

5

Given that a sample of n_p parents from the parental population has j derived alleles ,the $(i, j)^{th}$ entry of $\mathbf{P}_{n_p, n_o}$ can be interpreted as the probability that Wright-Fisher sampling of n_o offspring draws exactly n_p parents, and that the offspring has exactly i derived alleles. In the multiple-coalescence and strong selection regimes, we were unable to obtain an analytical expression that accounts for all the combinatorial ways to achieve this.

However, we can obtain fairly simple recursions in terms of the sample size n_o. Intuitively, $\mathbf{P}_{n_p, n_o-1}$ provides information about the first $n_o - 1$ offsprings and the parental samples. We can then draw an additional offspring, and update the transition matrix by summing over a small number of coalescence and selection possibilities for that offspring lineage. The derivation itself is somewhat tedious, and presented in Appendix X.

SG: Perhaps include description of recursion, jackknife, and convergence?

3. Results

3.1. Markov process construction

The maximum sample size is $n = N$, the size of the entire population. We expect that in the regime where $n \ll N$, 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 $\mathbf{P}_s$ 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 $\mathbf{P}_s$. While it should in principle be possible to include additional selective events, we do not pursue this here.

The above restriction implies that under strong selection with large number of derived lineages i, there is a possibility that some probability is *missing* from $\mathbf{P}_s$. However, since this model forms a proper Markov chain, we can calculate the total missing probability, by subtracting the total probability for a given derived allele count i from 1. This gives a rather natural proxy to checking how many extra lineages are required by selection. As we show further, this missing probability tends to 0 with increasing sample size.

6

The recursive nature of these calculations makes them relatively inefficient. Using a dynamic programming algorithm, we implement the construction of the neutral matrix $\mathbf{P}$ in the order of $O(n^3)$ operations. The selection transition probability matrix $\mathbf{P}_s$ needs $O(n^4)$ operations, due to the need to consider intermediate lineages. The increase in complexity makes this approach unsuitable for large sample sizes, but we derive several approximations in the following sections.

3.2. Calculation of allele frequency spectra

Once the truncated matrix $\mathbf{P}_s$ 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:

$$\Phi = \Phi Q + n\mu e_1 \tag{3}$$

where μ is the per-site mutation rate, and e_1 is the first column of the identity matrix of size n. Figure 2 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). 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. When the sample size is equal to the population size (Fig. 2B), we also include the AFS from the full Wright-Fisher model, which the other three calculations agree with.

Figure 2C 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. At stronger selection coefficients, `Moments` suffers from numerical instability, while the diffusion approximation performs well.

If the sample size is the same as the population size ($n = N = 100$) (Fig. 2D), the diffusion approximation and `Moments` perform detract SG: departs? from the Wright-Fisher prediction. The approach presented here (**??**) shows a better match to the Wright-Fisher model at small allele frequencies, but deviates later. We believe that the disagreement is due to the fact that we only allow at most one selective event per lineage, while there is no such limitation in the Wright-Fisher model.
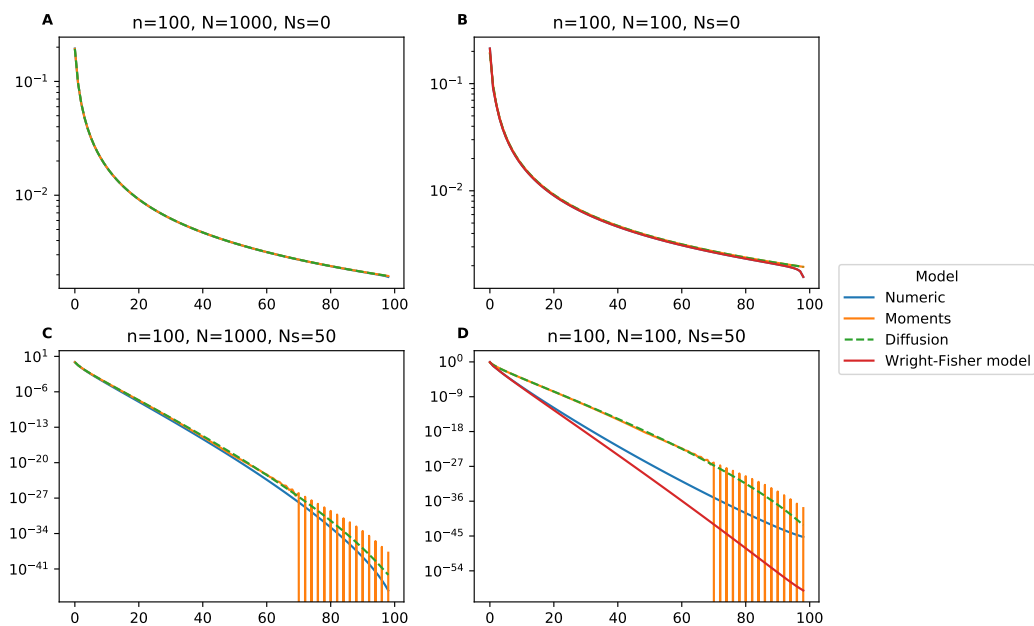
7

Figure 2: Normalized allele frequency spectra in a sample of size $n = 100$, for highly deleterious alleles ($Ns = 50$). (A) shows the frequency spectrum in a sample from a large population ($N = 1000$), (B) in a small population ($N = 100$). SG: The caption does not correspond to the figure. what is going on with the stripes? Presumably some log artefact? Maybe clarify that "numeric" is this work. Since the genome is only $3 \times 10^9 bases$, it would make sense to cutoff at 10^-9. Would also highlight differences better. Axis labels missing. PErhaps put the x label in proportions, and cut off for strong selection?

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. 8). 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 3 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 (??). 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:

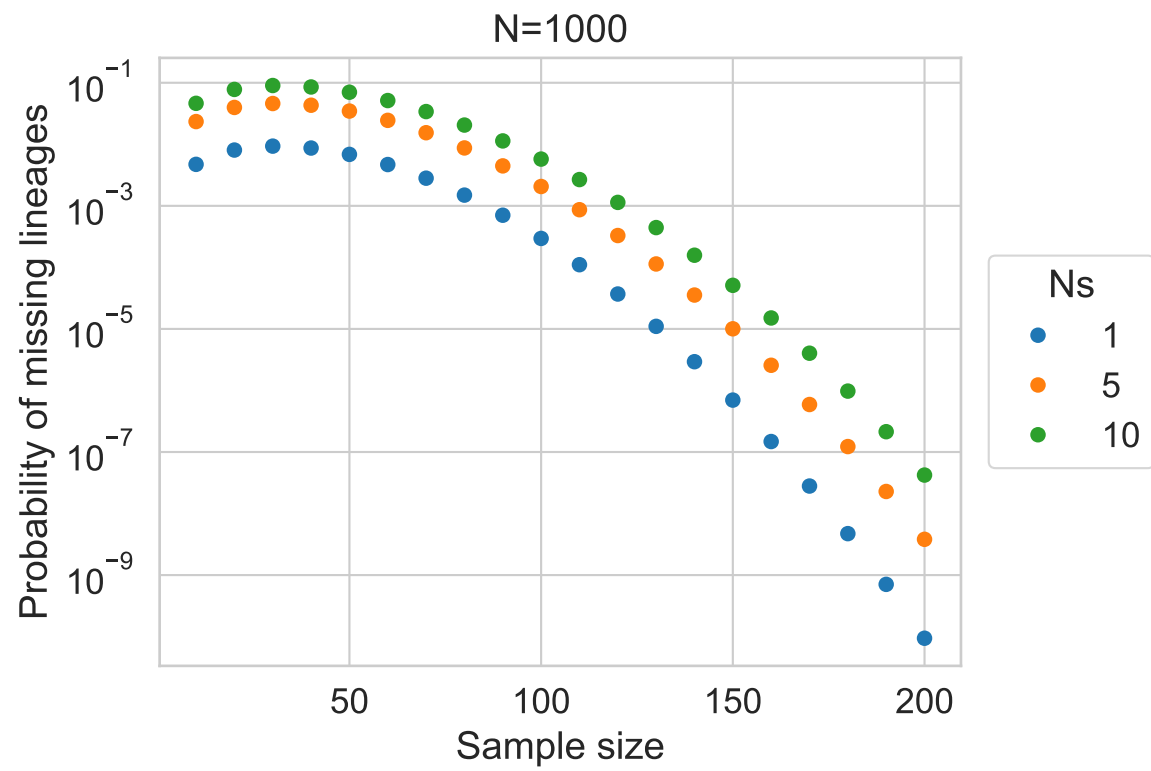$$Pr(n_p|n_o) = \sum_{n_g} Pr(n_p|n_g)Pr(n_g|n_o) \tag{4}$$

Figure 3: 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.

Where n_p and n_g is the number of contributing parents and gametes, respectively (Fig. 1B). Note that we consider this backward in time, so n_o is given, and we ask the probability of n_p conditional on n_o.

As a first order approximation, we can model the expectation $E[n_p|n_o]$ as the sum of lineages used under drift $\tilde{E}[n_p|n_o]$ (Fig. 1A) plus the number of extra lineages required by selection, $E[n_r|n_o]$.

$$\hat{E}[n_p|n_o] = \tilde{E}[n_p|n_o] + \hat{E}[n_r|n_o]$$
$$= N\left[1 - \left(1 - \frac{1}{N}\right)^n\right] + n\left(\frac{xs}{1 - xs}\right)$$
$$\underset{N \gg n}{\approx} \frac{nxs}{1 - xs} - \frac{n^2}{2N}$$

The expectations can be derived directly or from the corresponding probability distributions 4.2. The second approximation is made under the assumption that the sample size is much smaller than the population size. The increase of the number of lineages due to selection is linear. Drift decreases 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 (Krone and Neuhauser, 1997), eq. (??). Solving (4.1) for $n_o^* > \hat{E}[n_p|n_o]$ yields:

$$n_o^* \geq 2Nxs \tag{5}$$

This represents a critical sample size, above which drift outpaces selection. Note again the similarity to (??).

Figure 4 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

We now construct a probability distribution of the number of contributing lineages one generation into the past 1B, (4).

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:
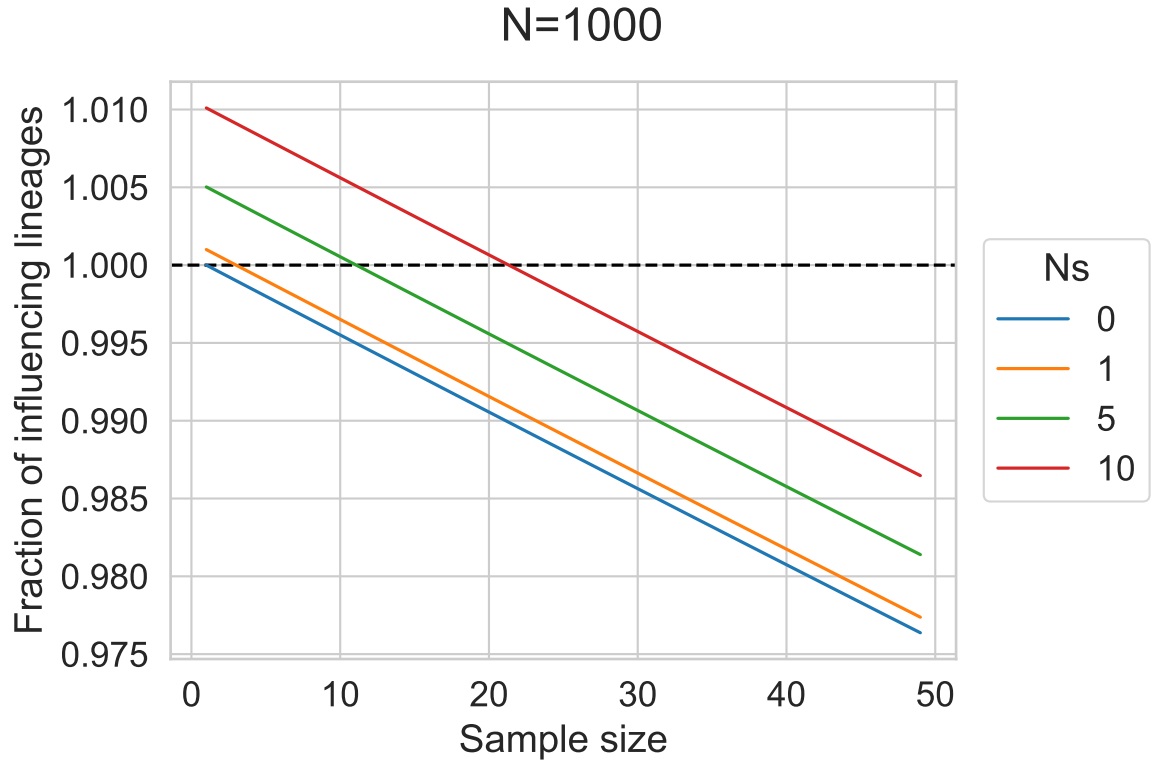
Figure 4: Critical sample size for different selection coefficients. The Y axis shows the fraction of parental lineages over the sample size, $\frac{n_p}{n_o}$, each line corresponds to a different selection coefficient. Above $\frac{n_p}{n_o} \geq 1$, selection dominates, below – drift. The critical sample size, where the expected number of parental contributing lineages is smaller than the sample size is well-approximated by $2Ns$.

$$P(n_p|n_g) = \frac{S_2(n_g, n_p)N!}{(N - n_p)!N^{n_g}} \tag{6}$$

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.* $n_r + n_o$):

$$P(n_g|n_o) = \binom{n_g - 1}{n_o - 1}(1 - xs)^{n_o}(xs)^{n_g - n_o} \tag{7}$$

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 4, we get:

$$Pr(n_p|n_o) = \sum_{n_g=1}^{\infty} \frac{S_2(n_g, n_p)N!}{(N - n_p)!N^{n_g}}\binom{n_g - 1}{n_o - 1}(1 - xs)^{n_o}(xs)^{n_g - n_o} \tag{8}$$

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 5 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 5 show that there is a large probability that $n_p > n_o$ at $n_o^* = 2n = 20$. In order to guarantee that drift 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 (5). 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
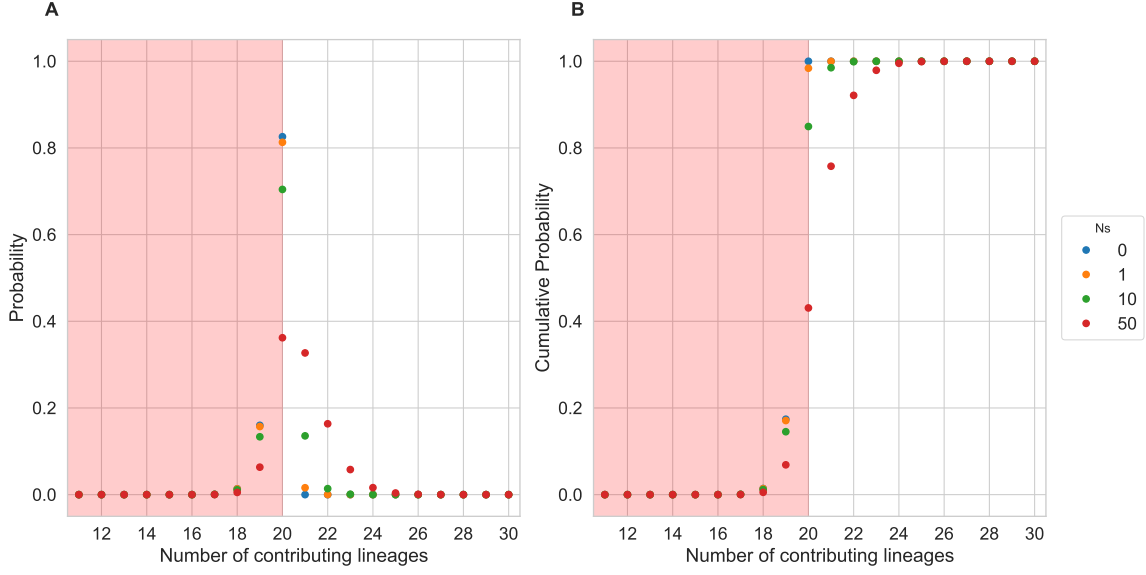
Figure 5: **A** The distribution and **B**, cumulative distribution of the number of parental contributing lineages one generation into the past ($n = 20$, $N = 1000$). Shaded red area shows the drift-dominated regime, where the number of lineages is smaller than the sample size.

of contributing lineages is smaller that the sample size 95% of the time, instead of 50%, as given by n_o^* in equation (5).

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:

14

$$Pr(\mathcal{R} = r|n) \approx \mathcal{N}(\mu = [(sn)/(1-s) + N(1-(1-1/N)^n)], \tag{9}$$

$$\sigma = \sqrt{N\left((N-1)\left(1-\frac{2}{N}\right)^n + \left(1-\frac{1}{N}\right)^n - N\left(1-\frac{1}{N}\right)^{2n}\right) + \frac{ns}{(1-s)^2}}) \tag{10}$$

Figure 6 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. 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:
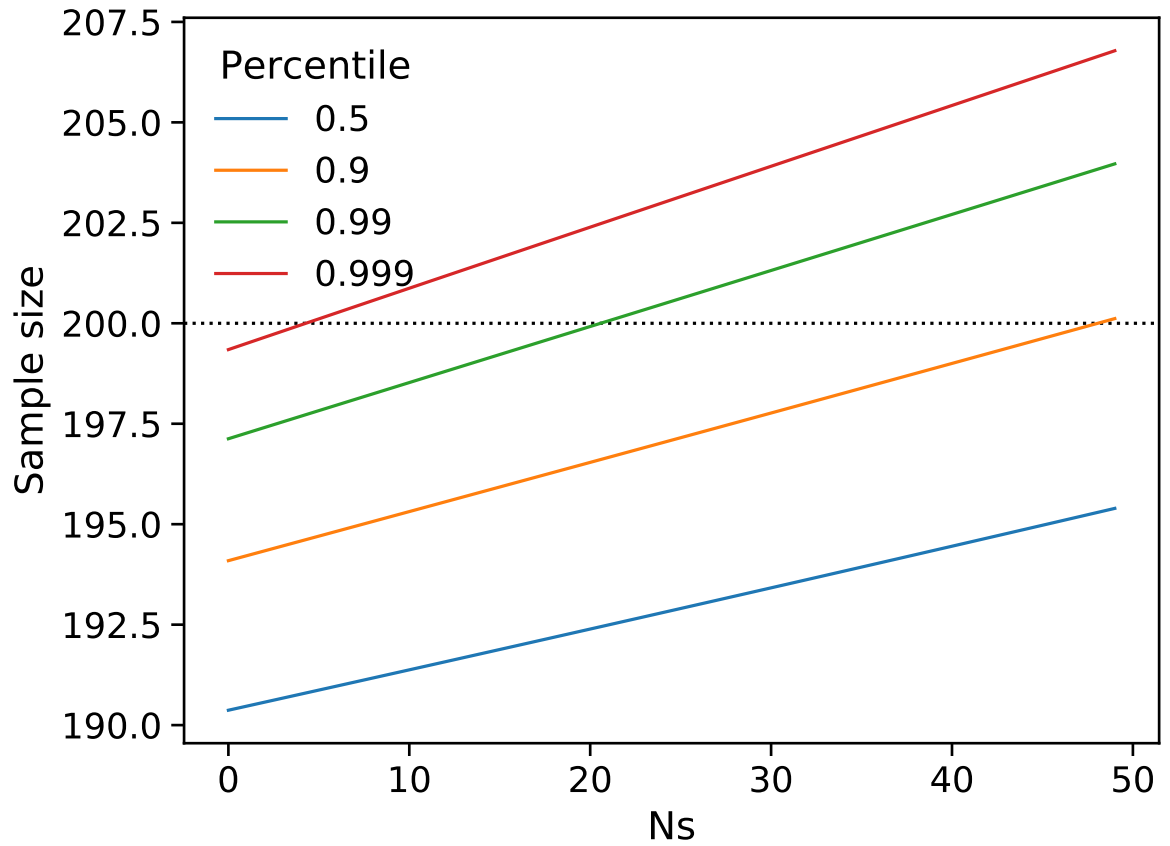
Figure 6: IK: TODO: need to resolve this still The quantile function of the closure of the sample (9). Each line corresponds to different percentile of the normal approximation. Black dashed line shows the reference sample size $n_o^* = 200$ SG: does it play a special role? If not why mention it (or have this line, really)?. SG: It also seems like showing the cumulative distributions themselves would be more intuitive. E.g log($missing p$). Also would be nice to have the numerical calculation. Could you get the cumulative distribution for the occupancy distribution from the Oneil algorithm?

$$P(i, n_g | r(k, n_g)) = \sum_{j,n_c} P(i, n_g; j, n_c r(j, n_c) | q(k, n_g))$$

$$= \sum_{j,n_c} P(i, n_c, r(j, n_c); n_g | q(k, n_g))$$

$$= \sum_{j,n_c} P(i, n_c | r(j, n_c); n_g; q(k, n_g)) P(r(j, n_c); n_g | q(k, n_g)) \quad (11)$$

$$= \sum_{j,n_c} P(i, n_c | r(j, n_c)) P(r(j, n_c); n_g | q(k, n_g))$$

$$= \sum_{j,n_c} P(i, n_c | r(j, n_c)) P(r(j, n_c); n_g | q(k, n_g))$$

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. (5)). Further, we derive the full probability distribution for the number lineages needed with given selection coefficient and

17

sample size (eq. (8)). 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. (9)). 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.
$\frac{i}{n}$	i derived *out of* n total alleles

6.2. Neutral case

We want to construct the entries in the probability matrix $P\left[\frac{\cdot}{n_o}\middle|\frac{\cdot}{n_p}\right]$ 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 **P** is square, $max(n_p) = max(n_o) = n$. Thus, our aim is to express every entry of **P**, $P\left[\frac{\cdot}{n}\middle|\frac{\cdot}{n}\right]$, in terms of $P\left[\frac{\cdot}{n-1}\middle|\frac{\cdot}{n-1}\right]$. Since we never require extra lineages ($n_p \leq n_o$), the recurrence is closed.

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

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

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

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

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

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

Last parent is ancestral

Last parent is derived

Figure 7: 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 $n-1$.

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 7. 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

20

When calculating a single entry in 7, the variables have the following ranges.

$$n_p = n$$
$$i_p \in [0, n]$$
$$n_o = n$$
$$i_o \in [0, n]$$

(12)

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

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

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.

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

(13)

21

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

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

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

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

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

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

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

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

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

$$(14)$$

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.

$$P(t; E_{n_o,t}(i, n_p)|r(j, n_p)) = \sum_c P(c, t; E_{n_o,t}(i, n_p)|r(j, n_p)). \tag{15}$$

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

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

$$= sP(c = True, a = D, t; E_{n_o,t-1}(i - 1, n_p)|r(j, n_p)) \tag{16}$$

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

SG: I think this might want to be:

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

$$= sP(c = True, a = D, t; E_{n_o,t-1}(i, n_p)|r(j, n_p)) \tag{17}$$

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

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

22

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

$$= sP(c = False, a = D, t; E_{n_o,t-1}(i, n_p - 1)|r(j, n_p)) \qquad (18)$$

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

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 $i \in \{0, \ldots, n_0\}$, $j \in \{0, \ldots, n_p\}$, and $n_p \in \{1, \ldots, n_{p,max}\}$. 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 $t_{m}ax$ for each successful draw, the number of terms we must compute is of order $t_{m}axn_p^4$. 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 $t_{m}ax$ cutoff, we modify the Wright-Fisher model by imposing a successful draw after $t_{m}ax - 1$ 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.

$$P_s\left[\frac{i_o}{n}\middle|\frac{i_p}{n}\right] = \sum_{i_c,n_c} P_s\left[\frac{i_o}{n}\middle|\frac{i_c}{n_c}\right] P_s\left[\frac{i_c}{n_c}\middle|\frac{i_p}{n}\right] \qquad (19)$$

The probability conditional on the contributing lineages ($P_s\left[\frac{i_o}{n}\middle|\frac{i_c}{n_c}\right]$) is given by equation 8, while $P_s\left[\frac{i_c}{n_c}\middle|\frac{i_p}{n}\right]$ 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 8 is analogous to the neutral case, and gives $P_s\left[\frac{i_o}{n}\middle|\frac{i_c}{n_c}\right]$, 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:

$$
\begin{aligned}
n_p &= n \\
i_p &\in [0, n] \\
n_c &\in [1, n] \\
i_c &\in [0, n_c]
\end{aligned}
$$
(20)

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

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

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

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

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

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

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

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

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

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

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

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

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

Last offspring is ancestral

Last offspring is derived

Figure 8: 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.

25