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

present

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

2 Background

We want to describe the behavior of n biallelic loci from a Wright-Fisher model with population size N. Specifically, we seek an expression for the transition probability matrix $\mathbf{P}$ for the time evolution of the *AFS*, Φ:

$$\Phi_n^{(t)} = \mathbf{P}\Phi_n^{(t-1)} \tag{1}$$

$\mathbf{P}$ is a square $n \times n$ matrix, and it enumerates the number of derived alleles (i) at a biallelic locus in a sample of size n. We will need to keep track of these quantities in the offspring (o), parents (p), and contributing (c) lineages for a given generation (table X). For example, $\frac{i_p}{n_p}$ (read as "i_p out of n_p") denotes a sample of size n with i derived in the parental generation (at $t-1$).

Instead of using the summation formulation presented in the main text (**??**), we use recursive conditional probabilities, where the transitions are defined in terms of transitions in smaller sample sizes. Each entry of the matrix, $\mathbf{P}_{(i_o,i_p)}$ is a a transition probability from i_p to i_o derived alleles with sample sizes n_p and n_o, respectively. The key observation is that to construct such probabilities, we can condition on a set of coalescent events, $\{\Lambda\}$:

$$\begin{aligned}
P\left[\frac{i_o}{n_o}\middle|\frac{i_p}{n_p}\right] &= \sum_{\lambda\in\{\Lambda\}} P(\lambda)P\left[\frac{i_o}{n_o}\middle|\frac{i_p}{n_p},\lambda\right] \\
&= \sum_{\lambda\in\{\Lambda\}} P(\lambda)P\left[\frac{i_o-|\lambda|_{i_o}}{n_o-|\lambda|_{n_o}}\middle|\frac{i_p-|\lambda|_{i_p}}{n_p-|\lambda|_{n_p}}\right]
\end{aligned}$$

Above, $|\lambda|$ denotes the size of the coalescent event. The second equation ascertains that conditioning on a coalescent event is equivalent to subtracting relevant lineages from a sample. This yields a recurrence where transition probabilities in a sample of size n can be described recursively in terms of transition probabilities in smaller sample sizes, $n' \in [1, n-1]$.

3 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

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

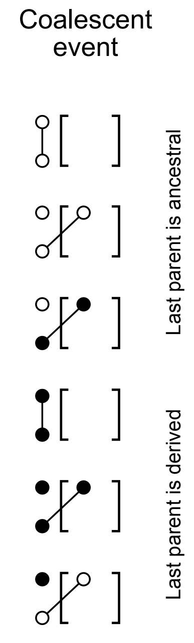

Figure 1: Recurrence defining transition probabilities in a model without selection. Right panel shows coalescent events corresponding to each summand. Each transition probability is defined in terms of transition in a smaller sample size. First three terms are conditional on the last parent having an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

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 $\mathbf{P}$, $P\left[\frac{\cdot}{n}\middle|\frac{\cdot}{n}\right]$, 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.

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 1. 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.

When calculating a single entry in 1, the variables have the following ranges.

$$\begin{aligned}
n_p &= n \\
i_p &\in [0, n] \\
n_o &= n \\
i_o &\in [0, n]
\end{aligned} \tag{2}$$

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

$$
\begin{aligned}
P\left[\frac{1}{1}\middle|\frac{1}{1}\right] &= 1\\
P\left[\frac{0}{1}\middle|\frac{0}{1}\right] &= 1\\
P\left[\frac{0}{1}\middle|\frac{1}{1}\right] &= 0\\
P\left[\frac{1}{1}\middle|\frac{0}{1}\right] &= 0
\end{aligned}
$$

4 Selection case

Due to selective deaths, the number of lineages (n_c) that contribute to the current generation can be significantly larger than the number of offspring (n_o), and especially so with strong selection. 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.

$$
P_s\left[\frac{i_o}{n}\middle|\frac{i_p}{n}\right] = \sum_{i_c, n_c} P_s\left[\frac{i_o}{n}\middle|\frac{i_c}{n_c}\right] P_s\left[\frac{i_c}{n_c}\middle|\frac{i_p}{n}\right] \tag{3}
$$

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 2, 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 2 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:

$$
\begin{aligned}
n_p &= n\\
i_p &\in [0, n]\\
n_c &\in [1, n]\\
i_c &\in [0, n_c]
\end{aligned} \tag{4}
$$

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

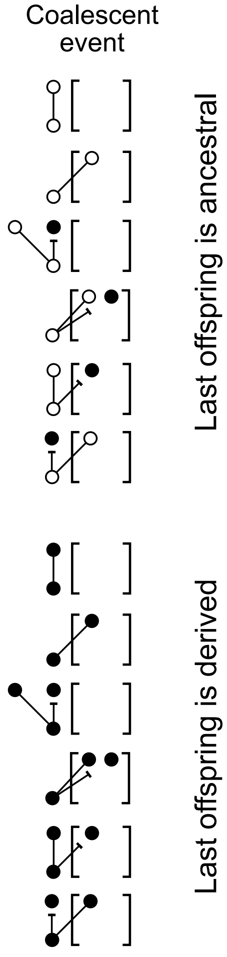

Figure 2: Recurrence defining transition probabilities in a model with selection. Right panel shows coalescent events corresponding to each summand. Each transition probability is defined in terms of transition in a smaller sample size. First six terms are conditional on the last offspring having an ancestral state, last six – derived. Filled circles – derived alleles; empty circles - ancestral alleles; square brackets – smaller sample.

Note that unlike in the neutral case n_c is now variable. The base cases of the recurrence are:

$$\begin{aligned}
P\left[\begin{array}{c|c}1&1\\=&=\\1&1\end{array}\right] &= 1-s\\
P\left[\begin{array}{c|c}0&0\\=&=\\1&1\end{array}\right] &= 1\\
P\left[\begin{array}{c|c}1&2\\=&=\\1&2\end{array}\right] &= s\\
P\left[\begin{array}{c|c}0&1\\=&=\\1&2\end{array}\right] &= \frac{1}{s}\\
\text{otherwise}\quad 0
\end{aligned}$$