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 with 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 reduction in the number of lineages due to coalescence is larger than the increase in the number of lineages due to selection, and the number of contributing lineages 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, we can compute likelihoods for different demographic scenarios. Unfortunately, the AFS calculations can be time consuming with complex demographic models, for example with multiple populations 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) with complex demographic models.

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. To describe the a sample of size n, we need to recursively consider samples of size $n' \leq n$. The decrease in the number of contributing lineages can be framed in terms of coalescent events (Kingman, 1982). 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, and 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 want to know how many parental lineages

(n_p) have been sampled from time $t-1$ (Fig. 1). Under a neutral coalescent model (Fig. 1A), the number of contributing parental lineages at $t-1$ is $n_p \leq n_o$, as the number of lineages decreases due to coalescent events.

To model the interplay of selection and drift, we consider a two-stage selection scheme (Fig. 1B). First, in a neutral process, n_p parents at $t-1$ produce a (potentially infinite) number n_g gametes for an intermediate $t-\frac{1}{2}$ generation. Second, the n_g gametes are sampled (with rejection) into n_o offspring at t. The number of rejected samples depends on the strength of negative selection, $s<0$. With stronger negative selection, more gametes will be rejected, so that $n_o \leq n_g$. We want to show that as n_o increases, asymptotically $n_o \leq n_p$.

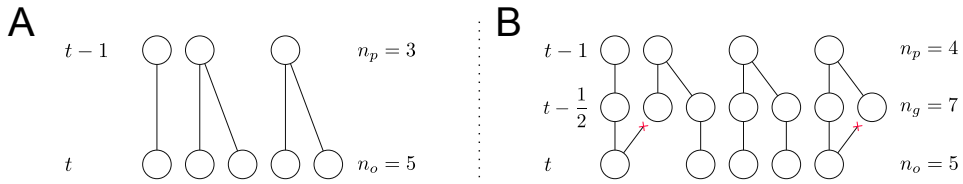

Figure 1: Realizations of sampling parental lineages under neutrality (A) and selection (B). **A** Under neutrality, possible coalescent events imply that number of parental lineages n_p at $t-1$ is less than or equal to n_o offspring lineages at t. **B** With selection, we add an intermediate gamete 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 the Kingman coalescent, only a single coalescent event is allowed per generation, in approximation that sample size is much smaller than the population size: $n_o \ll N$. This implies that under neutrality $n_p \in [n_o - 1, n]$. However, Bhaskar et al. (2014) show that with increasing sample size, higher order coalescent terms contribute more substantially. This means that to describe the sample of size n_o, we need to consider n_p potentially in the range $n_p \in [1, n]$.

We want to describe the time-evolution of the allele-frequency spectrum (AFS) in a sample size n_o at time t, which we denote as $\Phi_{n_o}^{(t)}$. We construct this recursively in terms of smaller sample sizes, n'. In this section, we follow the exposition in Jouganous et al. (2017), using drift ($\mathcal{D}_{n' \to n}$) and selection ($\mathcal{S}_{n' \to n}$) operators. These operators are sparse matrices that describe changes in allele frequencies in going from sample size n' to sample size n due to coalescent and selection events, respectively (Jouganous et al., 2017).

Under neutrality (Fig. 1A), we have $n_p \in [1, n_o]$, therefore:

$$\Phi_{n_o}^{(t)} = \sum_{n_p=1}^{n_o} \mathcal{D}_{n_p \to n_o} \Phi_{n_p}^{(t-1)} \tag{1}$$

This equation is closed with respect to the sample size n_o.

To include the effect of selection, we consider first the production of gametes from the parental generation as a neutral process. Changing the subscripts in (1) to refer to 1B, we have:

$$\Phi_{n_g}^{(t-\frac{1}{2})} = \sum_{n_p=1}^{n_g} \mathcal{D}_{n_p \to n_g} \Phi_{n_p}^{(t-1)}$$

Then, the produced gametes are sampled with rejection into n_o offspring:

$$\Phi_{n_o}^{(t)} = \sum_{n_g=n_o}^{\infty} \mathcal{S}_{n_g \to n_o} \Phi_{n_g}^{(t-\frac{1}{2})}$$

Combining the two expressions above, we get:

$$\Phi_{n_o}^{(t)} = \sum_{n_g=n_o}^{\infty} \mathcal{S}_{n_g \to n_o} \sum_{n_p=1}^{n_g} \mathcal{D}_{n_p \to n_g} \Phi_{n_p}^{(t-1)} \tag{2}$$

Since we need to consider a potentially infinite number of gametes produced (n_g), the equation (2) is no longer closed with respect to sample size.

The number of significant terms in the outer summation depends on the strength of negative selection $s < 0$ – described here by $\mathcal{S}$. Stronger negative selection will result in more resampling (Fig. 1B). The number of significant terms in the inner summation, however, depends on the sample size – larger sample sizes allow for more coalescent events. Above, this is opaquely included in $\mathcal{D}$.

IK: I don't like this explanation, since I feel that the equation doesn't add much to the figure.

We want to show that as the sample size increases, the number of significant terms in (2) decreases due to a large number of coalescent events.

The opposing effects of drift and selection on the number of lineages can be clearly seen in the context of the size of the ancestral selection graph (ASG): in which the number of lineages is described by continuous-time a birth-death process (Krone and Neuhauser, 1997; Wakeley, 2009).

$$n \to \begin{cases} n+1 & \text{at rate } \frac{\sigma n}{2} \qquad \text{(selection)} \\ n-1 & \text{at rate } \frac{n(n-1)}{2} \qquad \text{(coalescence)} \end{cases} \tag{3}$$

IK: The reason I use continuous time rates here is that the coalescent term is obviously quadratic in n. If I use discrete generations, it will be $\frac{\sigma}{\sigma+n-1}$ for selection, and $\frac{n-1}{\sigma+n-1}$ for neutrality, which is a little less obvious.

where σ is a population-scaled selection coefficient. The coalescence term is quadratic with respect to the sample size n, while the selection term is linear. The rate of coalescence is higher than the rate of selective events if the number of lineages $n > \sigma + 1$.

Our goal is to further investigate the interplay of selection and drift, and their effect on the *AFS*. First, we propose a construction that allows us to calculate the *AFS* under strong selection and large sample size, accounting for high-order coalescent terms. Second, we show that with increasing sample size, the system becomes asymptotically closed. We construct exact and approximate probability distributions that describe the number of contributing lineages. Additionally, we derive a normal approximation that allows us to calculate the quantile function of the sample size for a desired degree of closure.

IK: –SNIP–

3. Results

3.1. Markov process construction

SG: It seems like it would make the most sense to derive the transition matrices in this section (without limiting ourselves to a square matrix), and talking about the truncation approximation only in the next section. You could then have a symbol for the overall transition matrix Q, and the truncated matrix (the closure approximation) We first define a recursion equation for the distribution of allele frequency in a sample of size n from a haploid Wright-Fisher population of size N. Given that the sample in parental generation has j copies of the derived allele, we seek to calculate the probability that the sample will contain i derived copies in the following generation.

SG: The following could be clarifiedIn the neutral case, this transition probability can be calculated if we know the transition probabilities in the smaller samples of size $n' \in [1, n-1]$ (similar to (**??**)) (Bhaskar et al., 2014). We do not construct the intermediate matrices $\mathcal{D}_\rangle$ of (**??**) explicitly,

but instead calculate a matrix $P((j,n) \to (i,n))$, giving the probabilities of transitioning from j to i derived alleles in sample of size n within one generation. In brief, we can construct the transition probability matrix $P((j,n) \to (i,n))$ if we know the matrix for $n-1$, $P((\cdot,n-1) \to (\cdot,n-1))$, and then use the conditional probabilities for each type of an event. Starting from the base cases for a sample size of 1, we build up a set of square transition probability matrices. By using a dynamic programming algorithm, we can achieve reasonable performance for realistic sample sizes. The full derivation is shown in appendix A.

The case of selection is slightly more complicated, since now we need to consider a larger set of states that can lead to transition from j to i copies in a sample size of n. In particular, there can be a large number of gametes (fig. 1C) that can contribute from n when there is no selection, to a potentially infinite number under strong negative selection. In our calculation, we only consider up to one selective death event per lineage, so the number of contributing gametes is between n and $2n$. SG: Perfect: you can skip discussion of the approximation in the previous section, since you just introduce it here.

Note that this is analogous to the outer summation boundaries in eq. (2). Again, we do not calculate the matrix $\mathcal{S}_i$ directly, but rather construct a transition probability $Q((j,n) \to (i,n))$ SG: I don't understand what Q is, here.. We can define these transitions recursively if we know the transitions of $Q((\cdot,m) \to (\cdot,n-1))$. Note that with selection we can have $m \in [1,2n]$ lineages contribute, whereas we only needed $n' \in [1,n-1]$ in the neutral case.

To retain the closure property under selection, the Markov process needs to take $2n$ lineages in the parental generation to n lineages in the present. However, such rectangular transition probability matrix is not suited for our purposes, since we want to describe the behavior of a sample with constant size n. Instead, we only calculate the *truncated* transition probabilities for a $n \times n$ matrix Q. This means that under very strong selection, some transitions will be unaccounted for. However, since the total sum of transition probabilities sums to 1, we can easily calculate the total missing probabilities. As we show in the rest of this work, this missing probability tends to 0 as sample size n increases.

The construction of the full and truncated transition probability matrices is implemented via a dynamic programming approach similar to the neutral case, and is described fully in appendix B. Because we need to account for additional lineages in the selection case, the calculation time is of the order of $O(n^4)$, while it is only $O(n^3)$ for the neutral case. The increase in complexity

makes this approach less suitable for large sample size, but we derive several approximations in the following sections.

3.2. Calculation of allele frequency spectra

Once the truncated matrix Q is constructed, it can be used to calculate the allele frequency spectrum. For the infinite sites model at equilibrium, we can SG: calculate-approximate the SG: equilibriumAFS Φ as a solution to a linear system:

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

where μ is the per-site mutation rate, and e_1 is the first column of the identity matrix of size n. Figure 2 shows the comparison of the AFS calculated from Equation (4), the diffusion approximation (Ewens, 2004, eq. 9.23), and the calculation performed in `Moments` (Jouganous et al., 2017). Panel A shows a comparison at $Ns = 50$, with the population size ($N = 2000$), which is substantially larger than the sample size ($n = 200$). There is a small deviation between the approaches at large allele frequencies. At stronger selection coefficients, `Moments` suffers from numerical instability, while the diffusion approximation performs well (not shown SG: why?).

If the sample size is the same as the population size ($n = N = 200$) (Fig. 2B), the diffusion approximation and `Moments` perform poorly, while our approach SG: no need to make it about ourselves here remains stable. This is expected, since the diffusion framework does not perform well if multiple coalescent events contribute SG: cite bhaskar. Furthermore, if our sample size is the entire population, we expect recursion equations to be closed SG: This deserves more clarification. To confirm this, we compare our result to the AFS calculated from a whole-population haploid Wright-Fisher model, with $N = 200$. (Fig. 2B) shows that our calculation is close to the full Wright-Fisher model. The discrepancy between the curves is due to a difference in the way the selection coefficients are calculated SG: What does that mean?.

3.3. Closure properties

To SG: show-investigate the closure properties of Q, we can calculate the total probability that more that n parental lineages contribute to the sample of a given size. By construction, the sum of rows of Q should correspond to the total probability mass that included configurations contribute (Fig. **??** SG: no figure). Thus, the probability that some number of configurations are unaccounted

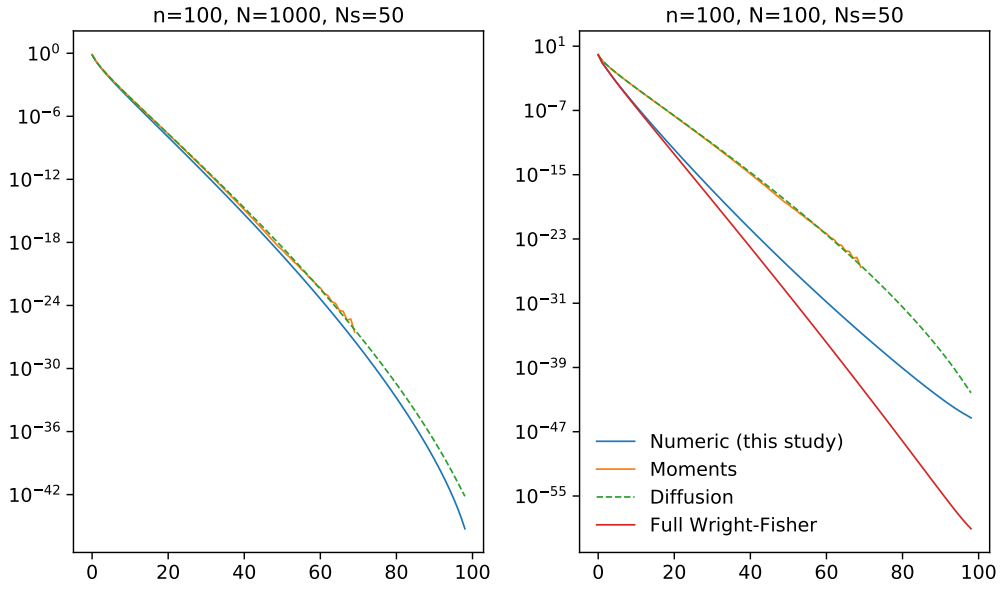

Figure 2: Normalized allele frequency spectra in a sample of size $n = 200$, for highly deleterious alleles ($Ns = -50$). (A) shows the frequency spectrum in a sample from a large population ($N = 2000$), (B) in a small population ($N = 200$). Both panels are truncated at 10^{-15}, to show only moderately high allele frequencies.

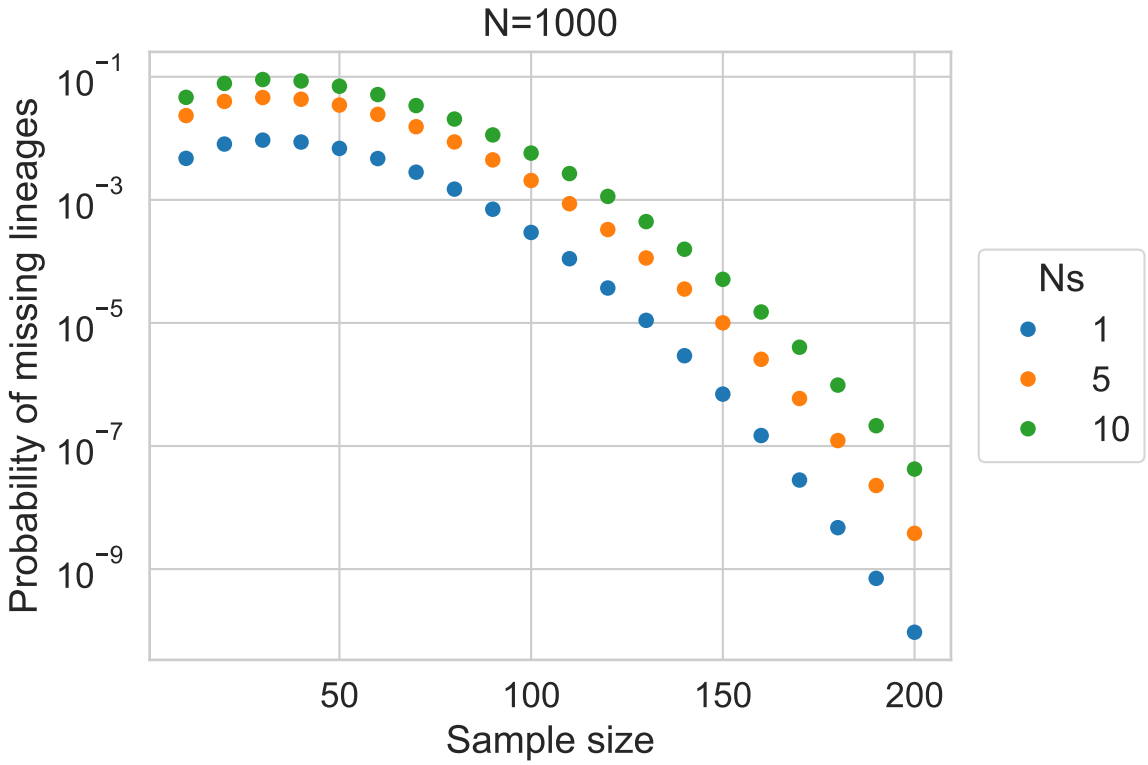

Figure 3: Probability that unaccounted lineages contribute to the transition probabilities. The probabilities are calculated as 1 minus the sum of probabilities for the state where every allele is derived. IK: Need to keep N consistent

for, with j derived alleles in the parental sample, is given by $1-\sum_{i=0}^{n} Q_{i,j}$. This probability depends on the number of derived alleles carried by the parental sample: the more derived alleles, the higher the likelihood of a selective event. Figure 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 (2). We derive several approximations to the model proposed in the first section, in order to get a better understanding of this behavior.

As a first order approximation, we consider the mean number of contributing lineages. Then, we construct a full probability distribution of the number of lineages contributing from the parental generation. Finally, we propose a normal approximation, which has a simple quantile function. This allows us to calculate the number of required lineages for the system to be closed with a given measure of certainty.

In the following derivations, we are assuming that the derived allele is present in the parental sample at frequency x, as opposed to explicitly modeling the count of derive alleles **??**, which considerably simplifies the calculations. If we seek the upper bound for the number of "lost" lineages, the maximal value will occur with $x = 1$, since only the derived lineages experience selection.

4.1. Mean number of contributing lineages

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

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

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

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

4.2. Expected number of lineages used

As a first order approximation, we can model $E[n_p|n]$ as the sum of lineages used under drift $E[n_p|n]$ plus the number of extra lineages required by selection, $E[n_p - n|n]$.

$$\begin{aligned}\hat{E}[n_p|n] &= \hat{E}[n_g - n|n] + \hat{E}[n_p|n] \\ &= N(1 - \left(1 - \frac{1}{N}\right)^n) + n\left(\frac{xs}{1-xs}\right) \\ &\underset{N \gg n}{\approx} \frac{nxs}{1-xs} - \frac{n^2}{2N}\end{aligned}$$

The expectations can be derived directly or from the corresponding probability distributions (**??**).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. (3).

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

$$\begin{aligned} \hat{E}[n_p|n] &< n \\ \frac{nxs}{1-xs} - \frac{n^2}{2N} &< n \\ n &\geq \frac{2Nxs}{1-xs} \\ &\approx 2Nxs \end{aligned}$$

This gives a simple expression for the sample size where drift overcomes selection: $n \geq 2Nxs$. Figure 4 shows this 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, $\frac{r}{n}$. Above the horizontal line $\frac{r}{n} > 1$, selection dominates. Below, drift reduces the number of used lineages. The intercept of the line with $\frac{r}{n} = 1$ is the critical sample size, which is well-approximated by $2Nsx$.

4.3. Distribution of number of contributing lineages

We now construct a probability distribution of the number of contributing lineages one generation into the past 1C, (5).

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:

$$P(\mathcal{R} = r|\mathcal{G} = g) = \frac{S_2(g,r)N!}{(N-r)!N^g} \tag{6}$$

where $S_2(g,r)$ is a Stirling number of the second kind, which is the number of ways to partition g gametes into r 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 $r \leq g$.

The distribution of the number of gametes, n_g is given by the negative binomial, parameterized by the total number of trials before n successes:

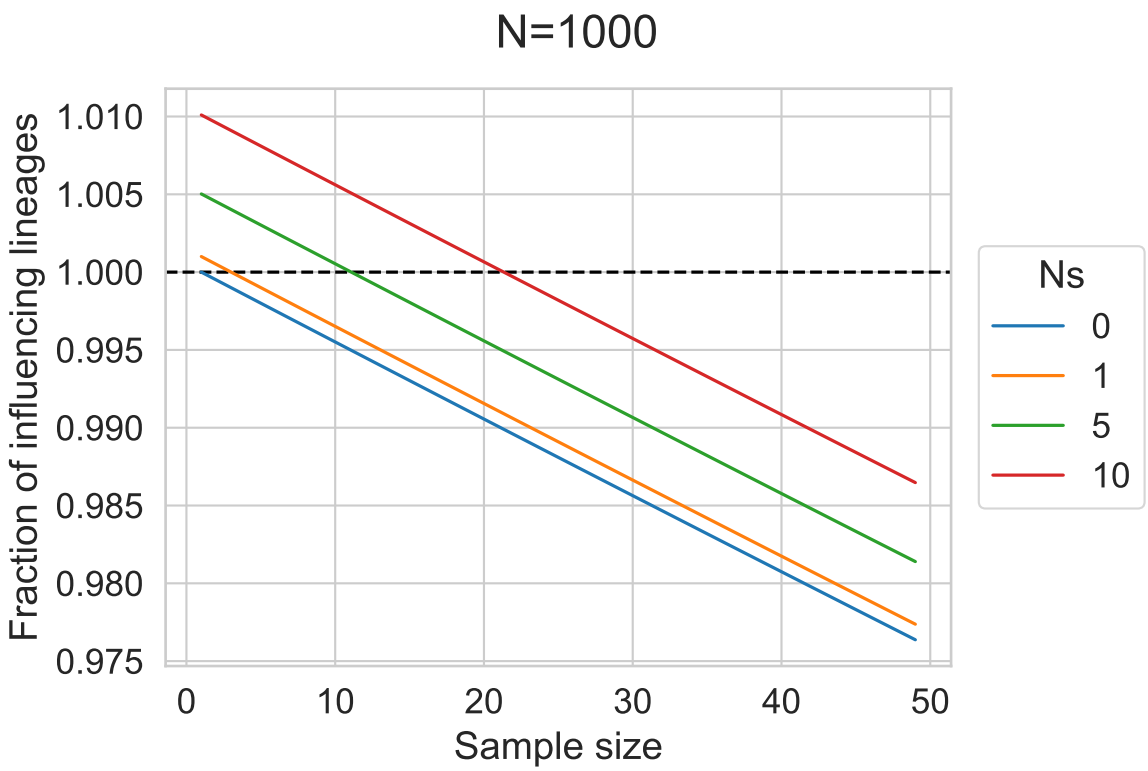

Figure 4: Critical sample size for different selection coefficients. The Y axis shows the fraction of parental lineages over the sample size, $\frac{r}{n}$, each line corresponds to a different selection coefficient. Above $\frac{r}{n} \geq 1$, selection dominates, below – drift. The critical sample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

$$P(\mathcal{G} = g|n) = \binom{g-1}{n-1}(1-xs)^n(xs)^{g-n} \tag{7}$$

Here, the number of gametes can be larger that the sample size $n \leq g$, if selection is present ($s < 0$) SG: Are you not using $s > 0$?.

Combining the two distributions together through 5, we get:

$$Pr(\mathcal{R} = r|n) = \sum_{g=1}^{\infty} \frac{S_2(g,r)N!}{(N-r)!N^g}\binom{g-1}{n-1}(1-xs)^n(xs)^{g-n} \tag{8}$$

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 $E[n_p|n] = n$. However, the distributions in **??** show that there is a large probability that $n_p > n$ at $n_{crit} = 2n = 20$. In order to guarantee that drift will out-pace selection, we can calculate the cumulative distribution - Figure 6. This shows that a sample size in which the majority of lineages are accounted for can be substantially larger than the critical sample size of equation (4.2). To derive a convenient analytical approximation, we turn to the normal approximation in the next section.

4.4. Normal approximation

Finally, we can construct a normal approximation to the distribution of the number of contributing lineages. The occupancy distribution is approximated by the normal (O'Neill, 2019) when $n \ll N$. Likewise, the number of failures (eq. (**??**)) 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 SG: What does that mean? Why do you need an approximation? Were you not computing a bound?. 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:

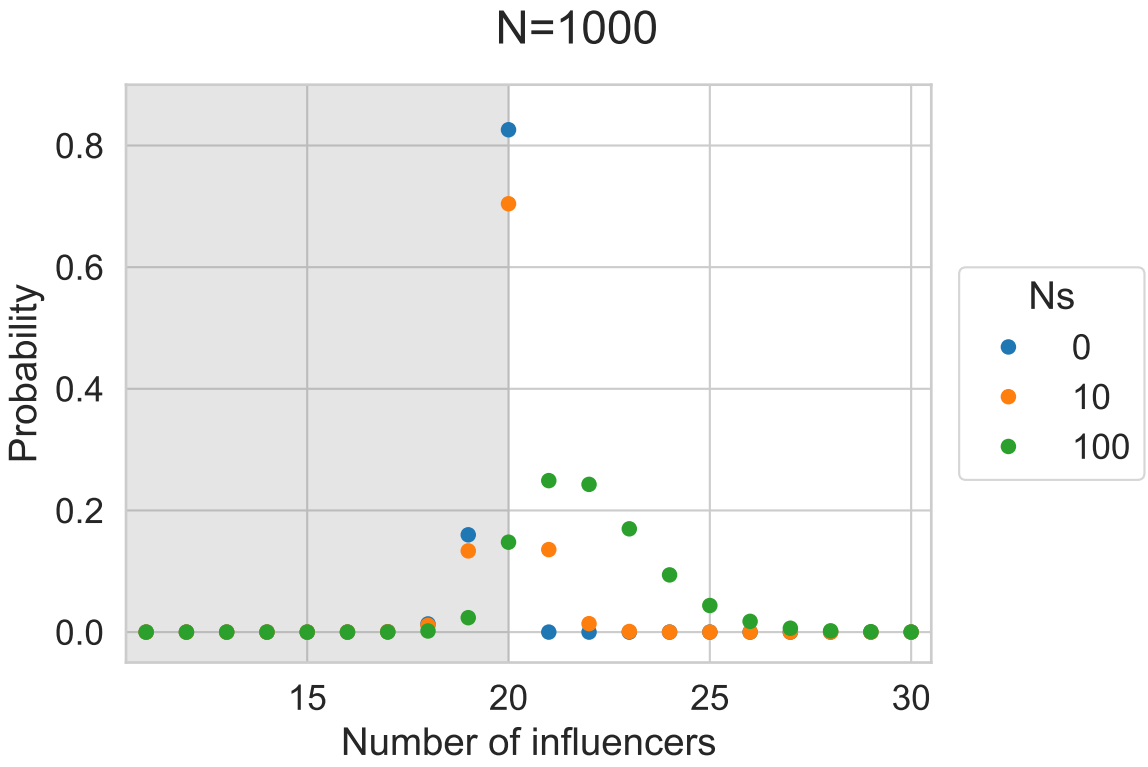

Figure 5: The distribution of the number of parental contributing lineages one generation into the past ($n = 20$, $N = 1000$). Shaded area shows the drift-dominated regime, where the number of lineages is smaller than the sample size.

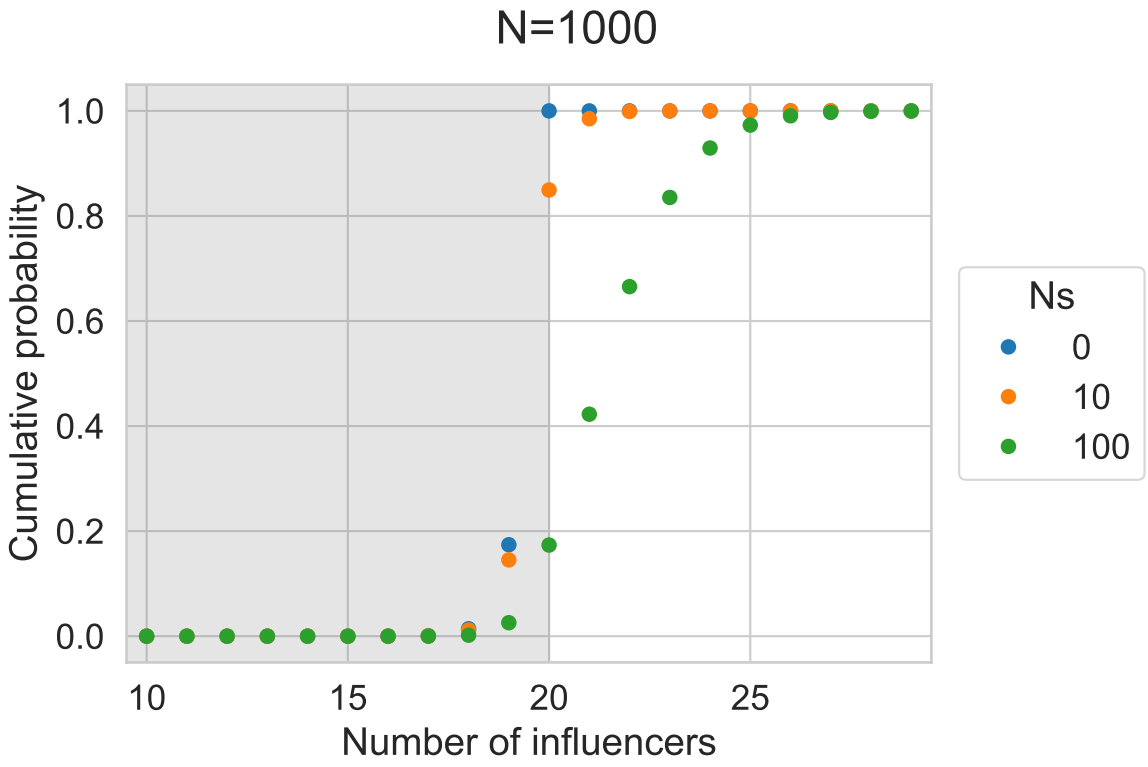

Figure 6: The cumulative distribution of the number of parental contributing lineages one generation into the past ($n = 20$, $N = 1000$). Shaded area shows the drift-dominated regime, where the number of lineages is smaller than the sample size. IK: This should be a two-panel with the previous figure

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

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

SG: Tell people N to make self contained 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.

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, $2Nsx$ is a critical sample size, where the decrease of lineages due to coalescent back in time out-competes the increase due to selection (eq. (4.2)). Further, we derive the full probability distribution for the number lineages needed with given selection coefficient and 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

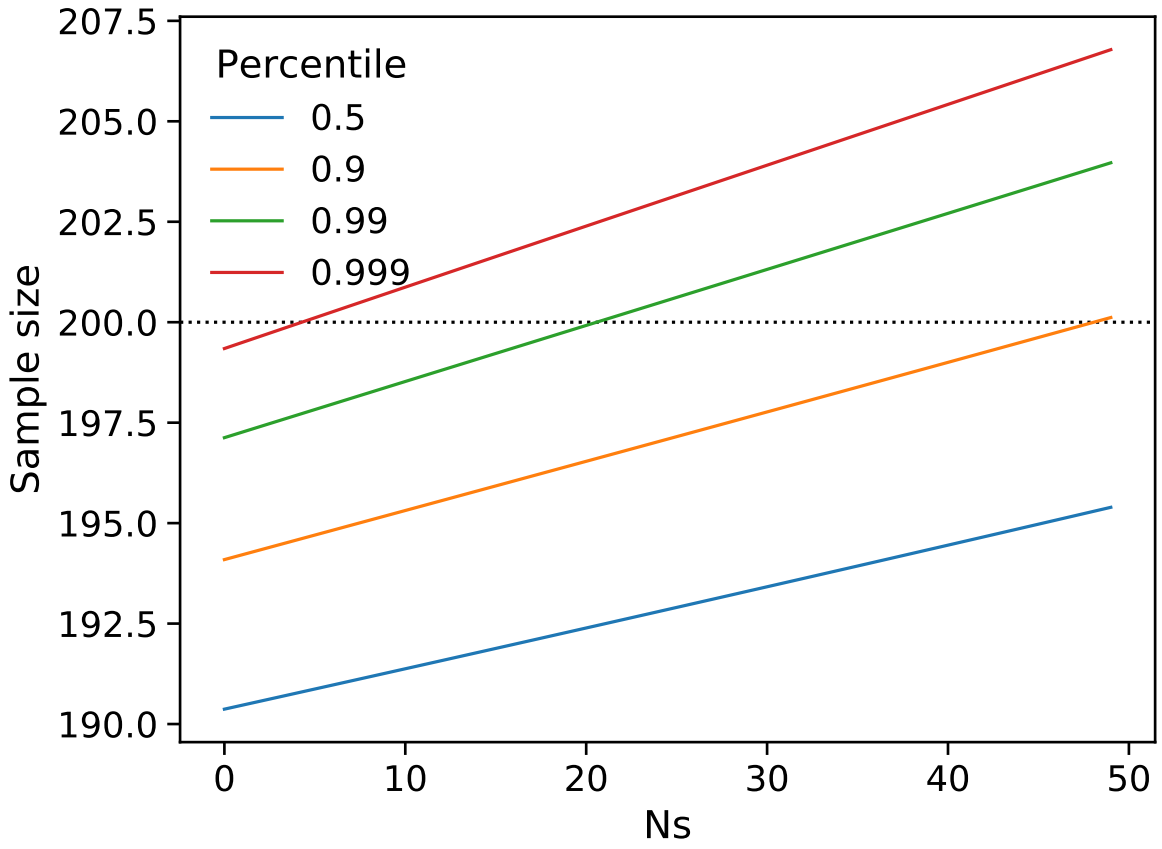

Figure 7: The quantile function of the closure of the sample SG: What is that?. Each line corresponds to different percentile of the normal approximation. Black dashed line shows the reference sample size $n = 200$ SG: does it play a special role? If not why mention it (or have this line, really)?. SG: 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 algo?

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.