Distinguishing among coalescent models using two-site allele frequency spectra

Daniel P. Rice

August 7, 2018

Abstract

[NOTE: OLD] The genetic diversity of a population reflects its demographic and evolutionary history. Methods for inferring this history typically assume that the ancestry of a sample can be modeled by the Kingman coalescent process. A defining feature of the Kingman coalescent is that it generates genealogies that are binary trees: no more than two ancestral lineages may coalesce at the same time. However, this assumption breaks down under several scenarios. For example, pervasive natural selection, rapid spatial range expansion, and extreme variation in offspring number can all generate genealogies with "multiple-merger" events in which more than two lineages coalesce instantaneously. Therefore, detecting multiple mergers is important both for understanding which forces have shaped the diversity of a population and for avoiding fitting misspecified models to data. Current methods to detect multiple mergers rely on the average site frequency spectrum (SFS). However, the signatures of multiple mergers in the average SFS are also consistent with a Kingman coalescent process with a time-varying population size. Here, I present a new method for detecting multiple mergers based on the mutual information of the joint allele frequency spectrum at pairs of linked sites. Unlike the average SFS, the mutual information depends mostly on the topologies of genealogies rather than their branch lengths and is therefore robust to most demographic effects.

Introduction

The genetic diversity of a population reflects its demographic and evolutionary history. Learning about this history from contemporary genetic data is the domain of modern population genetics (see M. Hahn 2018). The fundamental tools of the trade are toy models, which are used to study how historical forces shape genetic diversity, and which

form the basis of parametric inference methods. However, populations are complicated and, moreover, varied in their complications. No simple model can capture the processes governing every species' evolution, and fitting a misspecified model will generate misleading inferences. It is therefore crucial to understand the limits of our models and to be able to assess when a model is appropriate for a particular data set.

One of the most important and widely used models is the Kingman coalescent (**Kingman1982a**; **Kingman1982b**; **Kingman1982c**; **Hudson1983**; Tajima 1983). The Kingman coalescent is a stochastic process that generates gene genealogies, trees representing the patterns of shared ancestry of sampled individuals. Inference methods use these genealogies as latent variables linking demographic parameters to observable data (Rosenberg and Nordborg 2002). The Kingman coalescent has a number of convenient properties that allow for both analytical calculations (e.g., **Tajima1989**) and efficient stochastic simulations (e.g., Hudson 2002): tree topologies are independent of waiting times; waiting times are generated by a Markov process; and neutral mutations are modeled as a Poisson process conditionally independent of the tree. Moreover, the model may be extended to study a variety of biological phenomena including recombination, population structure, and variation in sex ratios or ploidy (see generally Wakeley 2009).

In its simplest form, the Kingman coalescent has a single parameter, the coalescent rate, which determines the branch lengths of genealogies (**Kingman1982a**). Under a variety of conditions, the coalescent rate is inversely proportional to the size of the population (**Kingman1982b**). Accordingly, a growing or shrinking population may be modeled by a time-varying rate (R. C. Griffiths and S. Tavaré 1994; R C Griffiths and Simon Tavaré 1998). This observation is the basis for a variety of methods to infer historical population sizes from genetic data.

With modern population genomics data sets, full-data likelihood models are impractical, so population-size inference is typically done on informative summary statistics. One widely used statistic is the site frequency spectrum (SFS): the number of mutations observed as a function of their allele frequency in the sample. The expected number of mutations at a given frequency depends on the branch lengths, and hence, on the coalescent rate. Thus, one can infer the coalescent rate by integrating over tree topologies, weighted by their probabilities under the Kingman coalescent. Some methods perform this integration by Monte Carlo simulation (e.g., Coventry et al. 2010; Excoffier et al. 2013). Others (e.g., Nielsen 2000, Bhaskar, Wang, and Yun S Song 2015), compute the expected site frequency spectrum directly for simple demographic models using the results of R C Griffiths and Simon Tavaré 1998 or Polanski and Kimmel 2003. (Another class of SFS-based methods are based on corresponding forward-time models rather than the coalescent (Ryan N. Gutenkunst et al. 2009; Lukić, Hey, and

Chen 2011; Ragsdale and Ryan N Gutenkunst 2017; Jouganous et al. 2017).)

A serious problem for this inference procedure is that different models of evolution generate different relationships between historical population sizes and genetic diversity. One of the basic assumptions of the Kingman coalescent is that natural selection is negligible in determining the distribution of genealogies. When this assumption is violated, Kingman-based inference methods are misspecified. For example, Schrider, Shanku, and Kern 2016 recently demonstrated how population-size inference can be distorted by selective sweeps. This effect is present in SFS-based methods as well as in sequentially Markov coalescent methods (e.g., Li and Durbin 2011). In a similar vein, Cvijović, Good, and Desai 2018, showed that purifying selection at linked sites that is sufficient to reduce overall genetic diversity is also sufficient to distort the SFS, leading to a false signal of population growth. Genomic evidence from multiple species suggests that such violations of the neutral model underlying the Kingman coalescent may be widespread (Sella et al. 2009; Corbett-Detig, Hartl, and Sackton 2015; Kern and M. W. Hahn 2018).

An important extension of the Kingman coalescent is a family of models known as *multiple merger coalescents* (**Pitman1999**; **Sagitov1999**; **DonnellyKurtz1999**) (reviewed in Eldon 2016), which arise in a variety of contexts both with and without selection. Whereas in the Kingman coalescent lineages may coalesce only pairwise, multiple merger coalescents permit more than two lineages to coalesce in a single event. The even more general class of simultaneous multiple merger coalescents (**Schweinsberg2000**; **MohleSagitov2001**; **Sagitov2003**) permits more than one multiple merger event at the same time. These models are relevant for species with "sweepstakes" reproductive events (Eldon and Wakeley 2006; Sargsyan and Wakeley 2008), fat-tailed offspring number distributions (**Schweinsberg2003**), recurring selective sweeps at linked sites (**CoopRalph**; Durrett and Schweinsberg 2005), rapid adaptation (**NeherHallatscheck2013**; **DesaiEtAl**), and purifying selection at sufficiently dense sites (**Seger**; **Good**; **Nicholaisen**). In each of these contexts, the coalescent timescale is not necessarily proportional to the population size. For example, with fat-tailed offspring distributions the rate of coalescence is a power law in the population size (**Schweinsberg2003**), while with linked sweeps it is determined by rate of linked sweeps (Durrett and Schweinsberg 2005). In these settings, interpreting the level of genetic diversity in terms of an "effective population size" is misleading and inferences based on the Kingman coalescent may be *qualitatively* incorrect. It is therefore important to determine whether the Kingman model is appropriate for a given data set before performing demographic inference.

One possible approach to identifying multiple mergers in genomic

data is to use the SFS as a summary statistic. (**Koskela2015** developed a full-likelihood method based on importance sampling, but, as with other "exact" inference methods, it does not scale to genomic data.) Birkner, Blath, and Eldon 2013; Blath et al. 2016; Spence, John A Kamm, and Yun S Song 2016 derived methods for computing the expected site frequency spectrum of (simultaneous-)multiple merger coalescents. Eldon, Birkner, et al. 2015 showed that it is possible to distinguish between a multiple merger coalescent of the beta family and the Kingman coalescent with exponential growth using the SFS. Rödelsperger et al. 2014 detected non-Kingman signatures of widespread linked selection in the nematode *Pristionchus pacificus* by demonstrating that the site frequency spectrum is non-monotonic, a signature of multiple mergers (**NeherHallatscheck2013**; Birkner, Blath, and Eldon 2013).

However, existing methods have limitations for distinguishing multiple mergers from general models of population-size change. The primary signature of multiple mergers in the SFS is an overabundance of low-frequency mutations relative to the Kingman expectation, which is also the signature of population growth. Eldon, Birkner, et al. 2015 were able to reject exponential growth in favor of multiple mergers with sufficient data, but a more flexible model of growth may be able to fit the multiple mergers SFS (see Myers, Fefferman, and Patterson 2008; Bhaskar and Yun S Song 2014). The non-mononotic SFS identified by Rödelsperger et al. 2014 is a more robust signature of multiple mergers, but detecting it in data requires knowledge of the ancestral allele at each site. High-frequency mutations are typically much rarer than low-frequency mutations, so misidentifying even a small fraction of ancestral alleles can generate a non-monotonic SFS at high frequencies.

Here, we propose that summary statistics based on the spectrum of mutation frequencies at pairs of nearby sites—the 2-SFS (**Hudson2001**; **FerrettiEtal2018**))—are useful for distinguishing between the Kingman coalescent with population growth and multiple merger coalescents. We show that this is true for both perfectly linked sites and sites separated by a moderate genetic distance. These statistics may be calculated efficiently from genomic single nucleotide polymorphism data. Furthermore, they do not require phasing, recombination maps, or ancestral allele identification and are informative even with small sample sizes. Together, these properties make the 2-SFS useful for demographic model-checking in a wide range of species. Finally, we demonstrate this model-checking procedure on genomic diversity data from *Drosophila melanogaster* (Lack et al. 2015).

Definitions and previously known results

The SFS and 2-SFS

Consider a sample of n chromosomes taken from a population of size N with mutation rate μ per site per generation. Following the notation of Fu 1995, the sample site frequency spectrum is $\{\xi_i : 1 \leq i < n\}$, where ξ_i is the fraction of sites containing a mutation with derived allele count i in the sample. In many cases, the ancestral allele is unknown and so the allele in i samples and the complimentary allele in $n-i$ samples are indistinguishable. Therefore, we will mostly consider the *folded* site frequency spectrum $\{\eta_i = \xi_i + (1 - \delta_{i,n-i})\xi_{n-i} : 1 \leq i \leq \lfloor n/2 \rfloor\}$, where $\delta_{k,k'}$ is the Kronecker delta.

The SFS and folded SFS are single-site statistics. As such, they can be calculated from a set of single nucleotide polymorphisms (SNPs) without any information about their relative locations in the genome. We define the two-site frequency spectrum (2-SFS) at distance $d > 0$, $\{\xi_{ij}(d) : 1 \leq i, j < n\}$, as the fraction of pairs of sites separated by d bases for which there is a mutation with derived allele count i at one site and a second mutation with derived allele count j at the other site. (Note that $\xi_{ij}(d) = \xi_{ji}(d)$ by symmetry.) This object has been studied for non-recombining sites by Ferretti et al. 2018 in a neutral model and **Xie2011** in a model with selection. By analogy to the folded SFS, we define the folded 2-SFS, as

$$\begin{aligned}\eta_{ij}(d) = \xi_{ij}(d) \\ &+ (1 - \delta_{i,n-i})\xi_{n-i,j}(d) \\ &+ (1 - \delta_{j,n-j})\xi_{i,n-j}(d) \\ &+ (1 - \delta_{i,n-i})(1 - \delta_{j,n-j})\xi_{n-i,n-j}(d).\end{aligned}$$

In this section, we will be considering non-recombining sites and so will suppress the distance d.

In the limit of low per-site mutation rate ($\mu \to 0$), all polymorphic sites are bi-allelic and the expected SFS and 2-SFS are related to moments of the branch length distribution by

$$\langle \xi_i \rangle = \mu \langle \tau_i \rangle \tag{1}$$

$$\langle \xi_{ij} \rangle = \mu^2 \langle \tau_i \tau_j \rangle , \tag{2}$$

where τ_i is the total length of branches subtending i leaves of a gene genealogy and $\langle \cdot \rangle$ represents the expectation over the distribution of gene genealogies defined by a coalescent model. Thus, the SFS and 2-SFS depend on the distribution of coalescent times as well as the distribution of tree topologies.

Fu 1995 calculated the first and second moments of the branch-length distribution for non-recombining infinite sites locus under the

standard time-homogeneous Kingman coalescent. He found that $\langle \tau_i \tau_j \rangle < \langle \tau_i \rangle \langle \tau_j \rangle$ for all $j \neq i, (n-i)$. This result, combined with Eq. (1) and (2), implies a negative correlation between mutations at different frequencies: trees generating a mutation with derived allele count i are less likely than average to generate a second mutation with derived allele count $j \neq i, (n-i)$. (There are positive correlations between mutations at complimentary frequencies induced by genealogies whose root node partitions the tree into subtrees of size i and $n-i$.)

Birkner, Blath, and Eldon 2013 extended Fu's calculation to a family of multiple merger coalescents called beta coalescents. This one-parameter family interpolates between the Kingman coalescent and the Bolthausen-Sznitman coalescent as the parameter, α, varies from 2 to 1. Beta coalescents arise in models with fat-tailed offspring distributions (**Schweinsberg2003**), and the Bolthausen-Sznitman coalescent is the limiting distribution of genealogies in rapidly adapting populations (**NeherHallatscheck2013**). Like Fu, Birkner et al. were primarily concerned with computing the sample variance of SFS-based summary statistics such as Tajima's D (**Tajima1989**). As a result, they were mostly interested in the diagonal terms of the SFS covariance matrix, which dominate that calculation. However, Figures 5 and 6 of Birkner, Blath, and Eldon 2013 show positive correlations between ξ_i and ξ_j for $j \neq i, n-i$. Thus, unlike the standard Kingman coalescent, the beta coalescent can generate positive associations between mutations with different minor allele counts.

Our goal is to demonstrate that this distinction in the 2-SFS between the multiple merger and Kingman coalescents: (i) applies to Kingman coalescents with time-varying coalescent rates; (ii) is robust to recombination between the sites; (iii) is also a feature of forward–time models with selection, and (iv) can form the basis of a model-checking procedure for demographic inference methods. To this end, we will make use of the fact that the 2-SFS may be interpreted as a joint probability distribution to define the following useful transformation.

Frequency pointwise mutual information

Given a coalescent model, the minor allele count at an arbitrary site is a random variable with probability mass function $p(i) = \langle \eta_i \rangle$, where we define η_0 to be the fraction of monomorphic sites. Similarly, the minor allele counts at a pair of sites separated by d bases is a pair of random variables with joint probability mass function $p_d(i,j) = \langle \eta_{ij}(d) \rangle$. We are interested in quantifying the dependence between minor allele counts at a given pair of sites, particularly for $i \neq j$.

A general measure of the dependence between a pair of random variables is the *mutual information* (**CoverThomas1991**). The mu-

tual information between random variables X and Y is a functional of their joint probability mass function, $p(x, y)$, defined as

$$I[p(x,y)] = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}, \tag{3}$$

where, with slight abuse of notation, $p(x) = \sum_y p(x, y)$ and vice versa. Equivalently, the mutual information is the Kullback-Leibler divergence between the true joint distribution of X and Y and the joint distribution of two independent random variables with the same marginal distributions as X and Y (**CoverThomas1991**). Thus, mutual information captures generic, possibly nonlinear, dependence between X and Y.

We would like to measure the association between alleles at particular frequencies, rather than the overall dependence between allele frequencies at nearby sites (which would be a measure of linkage disequilibrium). Equation (3) shows that $I[p(x, y)]$ is the expectation, relative to $p(x, y)$, of a quantity known as the *pointwise mutual information*, where

$$\text{pmi}(x,y) = \log \frac{p(x,y)}{p(x)p(y)} = \log \frac{p(x|y)}{p(x)} = \log \frac{p(y|x)}{p(y)} \tag{4}$$

(**ChurchHanks1990**). Pointwise mutual information measures the change in the probability that $X = x$ given knowledge that $Y = y$. When X and Y are independent, $\text{pmi}(x, y) = 0$ for all x, y. On the other hand, when X and Y are not independent, $\text{pmi}(x, y) > 0$ implies that $p(x|y) > p(x)$, and conversely for $\text{pmi}(x, y) < 0$.

As we have shown above, the expected folded 2-SFS is a joint probability mass function over minor allele counts. We may thus define the *frequency pointwise mutual information* as

$$\text{fpmi}(i,j;d) = \log \frac{\langle \eta_{i,j}(d) \rangle}{\langle \eta_i \rangle \langle \eta_j \rangle}. \tag{5}$$

This transformation of the 2-SFS has several useful properties. First, because it is based on the minor allele frequencies, we may compute fpmi from data without knowing the ancestral allele. Moreover, the results of Fu 1995 show that $\text{fpmi}(i, j) < 0$ for $i \neq j$ for non-recombining sites under the time-homogeneous Kingman coalescent.

The second benefit of computing fpmi from the 2-SFS is that it normalizes for distortions in the SFS. Population-size variation distorts the SFS by changing the distribution of waiting times. However, the time-inhomogeneous Kingman coalescent generates the same distribution of tree topologies as the time-homogeneous version. On the other hand, multiple mergers alter both the coalescent time distribution and

the distribution of topologies. These two effects are convolved in the SFS, making it difficult to distinguish between population growth and multiple mergers. Our hope is that the 2-SFS contains additional information about the distribution of genealogies, once distortions in coalescent times are accounted for.

For plotting purposes, we will also use a weighted version of the fpmi:

$$\mathrm{wfpmi}(i, j : d) = \frac{\langle \eta_{ij} \rangle}{\mu^2 \langle T_2 \rangle^2} \, \mathrm{fpmi}(i, j; d), \tag{6}$$

where T_2 is the coalescence time for a sample of size two. This weighting emphasizes the most common pairs of minor allele counts, while maintaining invariance to the mutation and pairwise coalescence rates.

Coarse-grained frequencies

With finite data, estimates of the 2-SFS will be noisy. This is particularly true for $i, j \gg 1$ because $\langle \eta_{ij} \rangle$ decays like $(ij)^{-1}$ for the standard Kingman (Fu 1995) and faster with growth or multiple mergers. We show in Results that a positive association between mutations with high minor allele counts and mutations with low minor allele counts is a signature of multiple mergers. This suggests a coarse-grained SFS and 2-SFS:

$$\eta_{\mathrm{lo}}(i_c) = \sum_{i=1}^{i_c - 1} \eta_i \tag{7}$$

$$\eta_{\mathrm{hi}}(i_c) = \sum_{i=i_c}^{\lfloor n/2 \rfloor} \eta_i \tag{8}$$

$$\eta_{\mathrm{hi,lo}}(d; i_c) = \sum_{i=1}^{i_c - 1} \sum_{j=i_c}^{\lfloor n/2 \rfloor} \eta_{ij}(d), \tag{9}$$

where i_c is an arbitrary cutoff between high and low minor allele frequency. Coarse-graining allows us to estimate the 2-SFS stably for large sample sizes, because we may adjust i_c to ensure a large number of sites in both the high–minor allele count and low–minor allele count bins.

We can also compute the pointwise mutual information in this coarse-grained distribution as

$$\mathrm{hilopmi}(d; i_c) = \log \frac{\langle \eta_{\mathrm{hi,lo}}(d; i_c) \rangle}{\langle \eta_{\mathrm{lo}}(i_c) \rangle \langle \eta_{\mathrm{hi}}(i_c) \rangle}. \tag{10}$$

One could similarly calculate five other pointwise mutual information statistics from the coarse-grained 2-SFS (e.g., the PMI between

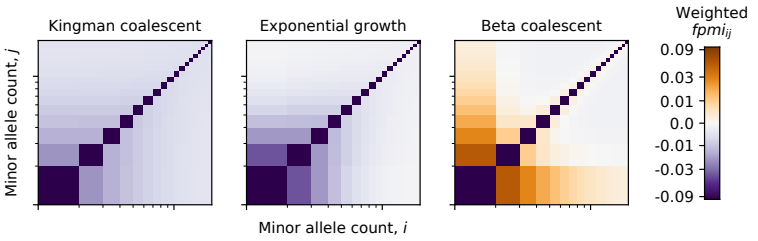

Figure 1: Weighted frequency pointwise mutual information for three coalescent models: the time-homogeneous Kingman coalescent, the Kingman coalescent with exponential growth ($g = 4$), and the beta coalescent ($\alpha = 1.45$). For all models $n = 39$. Growth and α parameters were chosen to generate average SFS with similar Tajima's D. The diagonal $i = j$ is masked.

monomorphic sites and sites with high minor allele counts), but we will not use these statistics here.

Results

Population growth vs. the beta coalescent in non-recombining loci

In this section, we compare the fpmi of the Kingman coalescent with and without population growth to the fpmi of the beta coalescent, for pairs of sites without recombination. We wrote a `python` package to compute the SFS, 2-SFS, and fpmi numerically. This package uses the results of Fu 1995 for the time-homogeneous Kingman coalescent, Živković and Wiehe 2008 for the time-inhomogeneous Kingman coalescent, and Birkner, Blath, and Eldon 2013 for the beta coalescent. (See Methods for details.)

Figure 1 shows the wfpmi for constant-N Kingman coalescent; the Kingman coalescent with exponential growth, $N(t) = N_0 \exp(-g\frac{t}{N_0})$; and a beta coalescent intermediate between the Kingman and Bolthausen-Sznitman coalescents. We find that while exponential growth with $g = 4$ has a substantial effect on the SFS (Tajima's D $= -0.39\pi$, where π is the average pairwise diversity), it does not qualitatively change the fpmi. In particular, $\text{fpmi}_{i,j} < 0$ for all $i \neq j$, just as in the constant-N Kingman coalescent.

On the other hand, a beta coalescent that generates similar distortions in the SFS ($\alpha = 1.45$, Tajima's D $= -0.42\pi$) also generates positive fpmi. The effect of multiple mergers is strongest on fpmi between high and low minor allele counts. This justifies coarse-graining

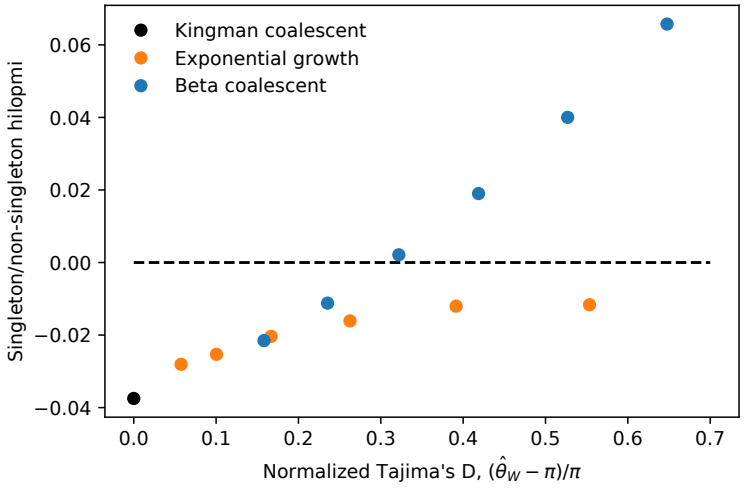

Figure 2: Pointwise mutual information between singletons and non-singletons for Kingman coalescents with exponential growth and beta coalescents. The exponential growth rate, g, ranges from 0.25 to 8.0 in coalescent time units. The beta coalescent parameter, α, ranges from 1.75 (nearly Kingman) to 1.25 (nearly Bolthausen-Sznitman).

the SFS and 2-SFS into high– and low–minor allele count bins as defined above.

Accordingly, we computed the coarse-grained hilopmi between singletons and non-singletons ($i_c = 1$) for a range of g and α. For the same parameters, we also calculated a normalized Tajima's D statistic: $(\hat{\theta}_W - \pi)/\pi$, where $\hat{\theta}_W$ is Watterson's estimator of the coalescent scaled mutation rate (**Watterson19??**). The latter is a measure of the distortions in the SFS induced by population growth or multiple mergers and equals zero for the constant-N Kingman coalescent. Figure 2 shows that as the population growth rate increases, distorting the SFS, hilopmi increases relative to the constant-N Kingman, but plateaus at a negative value. On the other hand, beta coalescents that generate similar distortions in the SFS, generate larger changes in hilopmi, including positive values. Thus, the 2-SFS in general, and hilopmi in particular, are capable of capturing the effects of multiple mergers beyond the distortions in branch lengths.

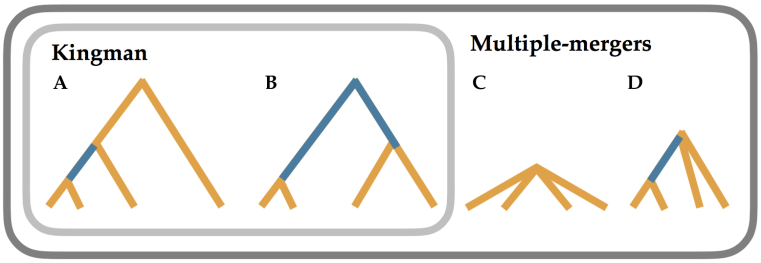

Figure 3: Genealogies with a sample size of $n = 4$. Opportunities for singleton/tripleton mutations are in orange. Opportunities for doubletons are in blue. (Note: there is a third possible multiple-merger topology, not shown.)

Samples of four chromosomes

We can get an intuitive understanding for the result above by considering a sample four chromosomes. In the Kingman coalescent, there are only two possible tree topologies (Figure 3). Furthermore, the total branch length is independent of the topology Wakeley 2009. As a result, there is a trade-off between the length of branches leading to singleton/tripleton mutations on one hand, and the branch length leading to doubletons on the other. Genealogies with topology (A) will have more opportunities for the former and loci with topology (B) will have more opportunities for the latter. Conditional on observing a doubleton at a site, it is thus more likely that the genealogy has topology (B) and so the expected number of singletons at sites with the same genealogy is lower than average. In terms of the 2-SFS, we have $\langle \eta_{12} \rangle < \langle \eta_1 \rangle \langle \eta_2 \rangle$.

On the other hand, multiple mergers induce correlations between the tree topology and the total branch length. For example, topology (C) has less opportunity for singletons *and* less opportunity for doubletons than (A) or (B), even though the expected proportion of singletons is higher. Thus, observing *any mutation at all* makes topology (C) less likely and the expected number of other mutations at all frequencies higher. If multiple-mergers events are frequent enough, this effect may dominate the tradeoff between (A) and (B) so that $\langle \eta_{12} \rangle > \langle \eta_1 \rangle \langle \eta_2 \rangle$.

As argued above, $\langle \eta_{12} \rangle$ is also distorted by changes in the coalescent time distribution induced by population growth. Figure 4 demonstrates that fpmi accounts for this fact by normalizing by the SFS. Figure 4 plots fpmi_{12} against the ratio of singletons to doubletons η_1/η_2 for the beta coalescent and two models of population growth: exponential growth and a piecewise-constant model with two epochs. In the latter model, we vary both the fold-change in N and the time of the change.

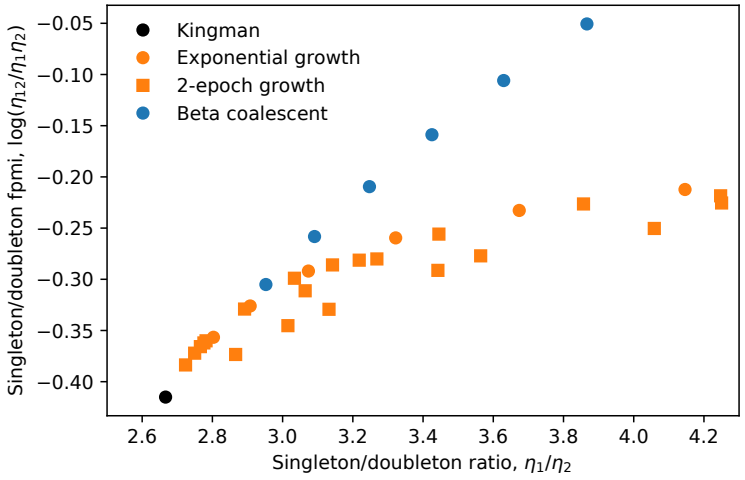

Figure 4: Distortions in fpmi vs. distortions in the SFS for $n = 4$. Parameters for beta coalescents and exponential growth are as in Fig. 2. For the piecewise-constant–N model, the fold-change in N varies from 2 to 2^4 and the time of this change varies from 2^-4 to 2 in coalescent time units.

Like Tajima's D in the previous section, the singleton/doubleton ratio measures distortion the SFS relative to the constant-N Kingman coalescent. (In fact, with $n = 4$, this ratio captures *all* of the distortion in the SFS.) As with the larger sample size (Fig. 2), mulitple mergers generate larger distortions in fpmi than population growth does, accounting for the distortions in the SFS. Moreover, the two different models of growth collapse onto one another. This suggests that for the Kingman coalescent with growth, $\text{fpmi}_{12} \approx f(\eta_1/\eta_2)$ for some function f.

Pointwise mutual information between recombining sites

The previous sections have shown that fpmi between pairs of non-recombining sites can discriminate between the Kingman coalescent with population growth and the beta coalescent. However, most demographic inference is performed on regions of the genome with non-zero recombination rates. In fact, for large non-recombining loci, it is sometimes possible to reconstruct the gene genealogy exactly and calculate tree imbalance statistics directly (see e.g., **Seger2010**). The advantage of the fpmi approach is that it can be computed on a genomic scale, as with the SFS. Therefore, it is important to assess the

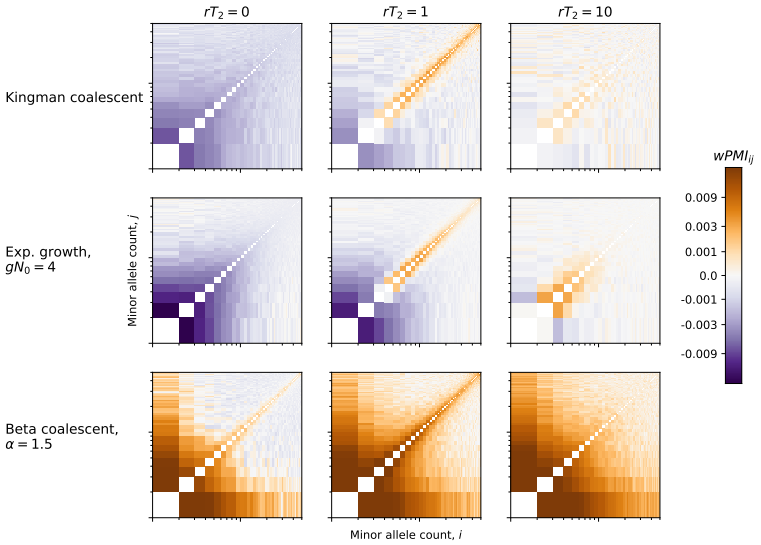

Figure 5: Weighted frequency pointwise information for three recombination rates with sample size $n = 100$. As in Figure 1, the diagonal elements are masked, and the growth and multiple-merger parameters were chosen to generate average SFS with similar Tajima's D.

robustness of our approach to recombination between sites.

To measure the relationship between fpmi and recombination, we ran coalescent simulations using a version of the program `msprime` **KehellerEtAl201?** modified to allow for multiple mergers. In particular, we simulated a model where coalescent events are drawn from the distribution specified by the beta coalescent, as in our numerical calculations above. In this model, marginal genealogies will follow the beta coalescent distribution, and the average SFS will be given by the formula in Birkner, Blath, and Eldon 2013. We also simulated data from two models of population growth: exponential growth and a two-epoch piecewise-constant model.

For each coalescent model, we simulated at least 10^4 independent infinite-sites loci, each with total recombination rate, r (See SI for parameter combinations). We chose values of r so that the combined parameter $r \langle T_2 \rangle$ varied over several orders of magnitude, where $\langle T_2 \rangle$ was measured from the simulation output. For each locus, we measured $\{\tau_i : i = 1, \ldots, n-1\}$ in the two genealogies at each end of the loci. We then calculated the expectations $\{\langle \tau_i \} \rangle$ and $\{\langle \tau_i j \rangle\}$ by averaging over independent loci. These expectations allow us to calculate all of the 2-SFS statistics defined above.

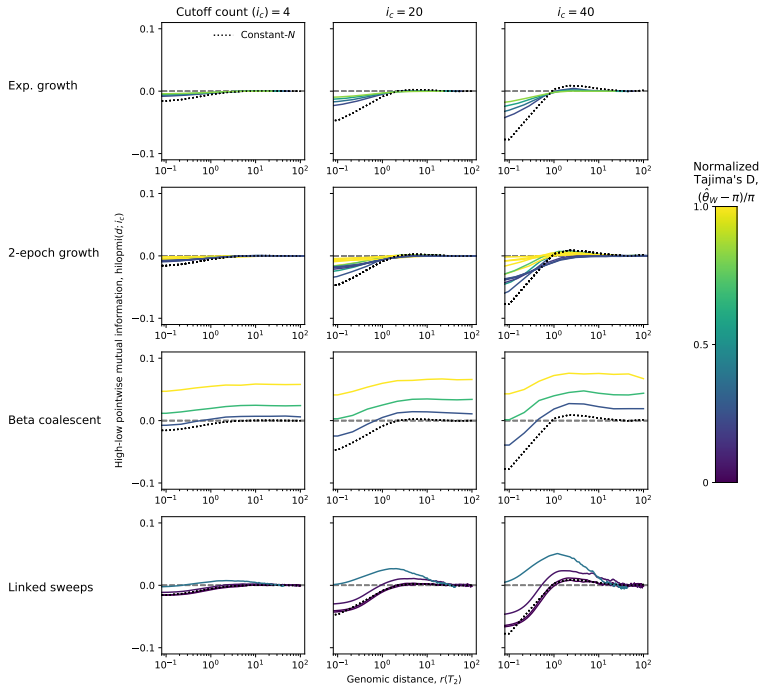

Figure 6: Pointwise mutual information between high– and low–minor allele count mutations (hilopmi) versus genetic distance. The first three rows show coalescent simulations in `msprime`. The fourth row shows forward-time `SLiM` simulations with selection at linked sites. Lines are colored by the distortion in the SFS, as measured by Tajima's D. [List parameters here?]

Figure 5 shows the weighted fpmi for three genetic distances in a constant-N Kingman coalescent, a Kingman coalescent with exponential growth, and a beta coalescent. As in the non-recombining case, the constant-N and exponential-growth Kingman models have similar fpmi. In both models, fpmi > 0 for $i \approx j$ for $r\langle T_2 \rangle \sim 1$. This is presumably because trees at nearby sites contain clades with similar numbers of leaves. These positive correlations, however, do not extend to $i \gg j$, which is the signal of multiple mergers in our coarse-grained hilopmi. On the other hand, the positive fpmi in the beta coalescent persists for $r\langle T_2 \rangle > 1$.

Figure 6 shows hilopmi(d, i_c) in both models of growth and the beta coalescent, for a range of $d = r\langle T_2 \rangle$ and three different choices of i_c. Each curve represents a particular parameter combination and coalescent model and is colored by the distortion in the average SFS, which is independent of the recombination rate. At low recombination

rates, hilopmi may be greater in growing populations than in constant-N populations (dotted lines), but is always less than zero, consistent with the results for non-recombining sites. When $r\,\langle T_2\rangle \geq 1$, hilopmi may be slightly positive for large i_c, but is smaller in growing populations than in the constant-N Kingman model. In all Kingman models, hilopmi decays to zero for $r\,\langle T_2\rangle \gg 1$.

In contrast, the beta coalescent generates hilopmi that is consistently greater than the constant-N Kingman. As with non-recombining sites, the hilopmi is also greater in the beta coalescent model then in models of population growth than generate similar distortions in the SFS. [IDEA: A plot of hilopmi at a fixed r vs Tajima's D for all the models together. Could make several plots, varying r and i_c. This might be more convincing than the current figure.] This is true across recombination rates and high/low cutoff minor allele counts. These results demonstrate that hilopmi is capable of discriminating between coalescent models even when there is recombination between sites.

There is one other salient feature of Fig. 6: hilopmi does not decay to zero at long distances in the beta coalescent. This is related to the results of **EldonWakeley20??** who showed that a model with "jackpot" reproductive events can generate infinite-range linkage disequilibrium. They showed that different scalings of rates of mutation, pairwise coalescence, multiple merger coalescence, and recombination lead to different behaviors of diversity and linkage disequilibrium. Our implementation of the beta coalescent model with recombination corresponds to a particular scaling limit where recombination does not have time to decorrelate trees during multiple mergers events, even at infinite genetic distances. We do not expect this behavior to be universal to multiple mergers coalescents with recombination.

Linked selective sweeps

Various authors have shown that natural selection can generate multiple merger coalescents at linked neutral sites (e.g., **CoopRalph**; **NeherHallatscheck2013**; **DesaiEtAl**; **Seger**; Durrett and Schweinsberg 2005). However, our simulations of the beta coalescent with recombination are of an explicitly neutral model. Thus, they are at best an approximation to the selective models cited above. It is therefore important to verify that selection at linked sites can, in fact, generate the sorts of signals in fpmi that we have detected in the beta coalescent.

To test this proposition, we performed forward-time simulations of recurring selective sweeps using the software `SLiM` (**MesserEtAl201?**). In these simulations, we simulated individual chromosomes with homogenous recombination and two types of mutations: neutral and beneficial with fixed selection coefficient s. Both types of mutations occurred at random, uniformly distributed across the length of the

chromosome. By varying the recombination rate, mutation rates, population size, and selection coefficient, we were able to vary the rate of selective sweeps linked to neutral sites. After the populations reached steady-state, we sampled individuals and calculated $\{\eta_i\}$ and $\{\eta_{ij}(d)\}$ for the neutral mutations. (See Methods for details.)

The bottom row of Fig. 6 shows the results of these simulations. For intermediate genetic distances, $dr\,\langle T_2\rangle \sim 1$, the effects of linked sweeps are qualitatively similar to the effects of the beta coalescent. That is, when sweeps are sufficiently frequent to distort the SFS, as measured by Tajima's D, they also increase hilopmi. The primary difference between the forward-time sweeps and beta coalescent simulations, is that the distortions caused by sweeps decay to zero for $dr\,\langle T_2\rangle \gg 1$. This is reasonable because the effect of a particular sweep on genealogies should be localized around the beneficial mutation.

Coalescent model checking: application to *Drosophila melanogaster*

Our results above show that fpmi and its coarse-grained analog, hilopmi are useful for distinguishing population growth from the effects of multiple mergers. This is true even when the population growth generates similar distortions in the SFS. We therefore propose the following model-checking procedure for demographic inference methods:

(i) Fit a demographic model to data using any relevant method.

(ii) Simulate genealogies under the fitted model (using `msprime` or other coalescent simulator). Calculate the fpmi(d) and hilopmi($d; i_c$) predicted by the model.

(iii) Calculate fpmi(d) and hilopmi($d; i_c$) from the data.

(iv) Compare true to predicted statistics. Evaluate model fit.

This procedure checks whether the model is consistent with a dimension the data that was used in fitting the model. Inconsistency suggests that the inferred $N(t)$ may be an artifact of natural selection, skewed offspring distributions, etc., rather than reflecting the true historical population size.

In this section, we illustrate the procedure just outlined by using genomic diversity data from the *Drosophila melanogaster* DGPG3 panel Lack et al. 2015. The DPGP3 data consists of haploid consensus sequences from ~ 200 wild-caught flies from a Zambian population known to be relatively free of cosmopolitan admixture. Recently, several groups have used the DPGP3 data to estimate the population-size history of *D. melanogaster* (Terhorst, John A. Kamm, and Yun S. Song 2017; Ragsdale and Ryan N Gutenkunst 2017). On the other hand, it is widely believed that the genetic diversity of *Drosophila* is strongly

shaped by natural selection (e.g., **Elyashiv??201?**; **others?**; Garud and Petrov 2016) Thus, this data is a good candidate for demonstrating the utility of fpmi for assessing coalescent model fit.

After filtering for coverage, removing chromosome arms with known inversions, downsampling to $n = 100$ samples per autosomal chromosome arm, and identifying 4-fold degenerate sites, we selected the central region of each chromosome that have consistent high diversity (Methods). Because the average pairwise diversity varies between arms—possibly reflecting selection or different sets of segregating inversions—we performed all subsequent calculations on each arm independently. We fit a demographic model to the site frequency spectra of these central regions using `fastNeutrino` **BhaskarEtal2015** We fit a 3-epoch piecewise-constant model, with four free parameters: two changepoints and two population size ratios. We report our fitted parameters and those of a similar published model (Ragsdale and Ryan N Gutenkunst 2017) in Table 4 [Make table]. We then simulated under our fitted model using `msprime` and computed the expected and observed SFS, fpmi, and hilopmi (Fig. 7).

The first row of Fig. 7 shows that the expected SFS under the fit demographic models agree with the observed SFS, demonstrating that a time-varying $N(t)$ can explain this aspect of the data well. In contrast, the second row shows the expected and observed weighted fpmi for averaged over distances less than $15d_c$ apart, where $d_c = 2/(\pi\mu)$ is the distance scale corresponding to $r\langle T_2\rangle = 1$. Here, the data shows strong positive associations between nearby alleles at different frequencies, while the model of population growth predicts weak negative associations except just off of the diagonal. This pattern extends across a range of genomic distances, $d/d_c \in (10^{-1}, 10^2)$ (Fig. 7, third row). As a result, we may conclude that the data is not well explained by the Kingman coalescent with population growth.

Note that the hilopmi decays toward zero at large distances. This matches the expectation from simulations with selective sweeps rather than the beta coalescent. However, we caution against concluding that sweeps are necessarily responsible for the deviations from the Kingman expectation.

Discussion

- Review results.
- Availability of code: models for computing moments numerically, scripts for running simulations. Notebooks for data analysis.
- We focus here on coalescent-based methods, but also relevant for diffusion methods (because they imply a coalescent model)

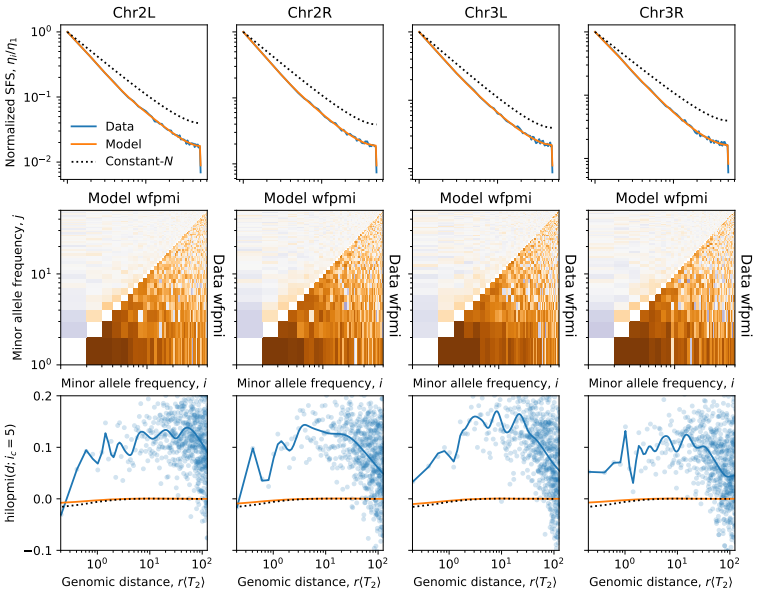

Figure 7: DPGP3 data and coalescent model predictions. First row: observed (blue) and expected (orange) site frequency spectrum compared to constant-N Kingman coalescent (dotted lines). Second row: Expected (upper triangle) and observed (lower triangle) weighted fpmi averaged over all pairs of sites less than $15d_c$ apart. Third row: Observed (blue) and expected (orange) hilopmi versus genetic distance. Cutoff minor allele count, $i_c, = 5$. Solid blue curves show cubic spline fit. Dotted lines show the expectation under the constant-N Kingman coalescent. [Middle row needs colorbar]

- Mention methods to fit multiple mergers coalescent to data (computationally difficult). Not really what we're trying to do here. Our goal is that you can exploit this extra dimension of the data you already have to check your model and make sure you're not fitting a signal of selection, etc with a demographic model.
- Could also develop a formal hypothesis testing procedure, using the bootstrap, etc.
- Compare with: Single-locus tests that rely on building trees (Seger whale lice, Neher with viruses)
- Compare with: Multi-locus tests that compare a particular region to genomic background. (What if the background isn't neutral/Kingman?)
- But: we could extend our method to a local method that looks for variation in "multiple-merger-ness"
- Multiple-mergers coalescent and recombination: open area of research
- fpmi(d) should be a standard posterior check for demographic models. Possibly incorporate into model fitting packages.
- Interesting to compare across species.

Methods

Computing branch-length moments

We implemented numerical computations of the moments of the branch lengths $\langle \tau_i \rangle$ and $\langle \tau_{ij} \rangle$ using the `python` numerical package `numpy` (**numpy**). For the Kingman coalescent with time-varying coalescent rate, we used equations (1)-(12) of Živković and Wiehe 2008. To calculate the second moments of the coalescent time distribution for exponentially growing populations, we used Gaussian quadrature. For the beta coalescent, we implemented the recursion described by Birkner, Blath, and Eldon 2013.

Functions for compute the branch-length moments are available [GITHUB]. The Kingman coalescent code currently can compute moments for exponentially growing and two-epoch piecewise constant models, but could be extended to allow for other models. The formulas of Živković and Wiehe 2008 exhibit numerical instability related to the instability of R. C. Griffiths and S. Tavaré 1994. They are therefore only practical for sample sizes up to $n \approx 40$. The recursion of Birkner, Blath, and Eldon 2013 is $\mathcal{O}(n^4)$ [CHECK THIS] and is also only practical for samples up to $n \approx 50$.

Coalescent simulations

Our beta coalescent simulation code is based on modifications to the `msprime` package made by Joe Zhu [link to github] to allow for multiple mergers. This code, together with `python` wrapper scripts and utility functions to run simulations and calculate the 2-SFS and fpmi from the `msprime` output is available [GITHUB].

[Possibly move some of the description here]

Forward-time simulations of selective sweeps

We simulated a model of recurring selective sweeps using the software `SLiM` (**MesserEtAl201?**). In all of these simulations, we simulated a population of 500 diploids for 10^4 generations. Each haploid genome consisted of a single genomic element $L = 10^8$ basepairs long with recombination rate per basepair $r = 10^{-8}$ and overall mutation rate $\mu = 10^{-7}$. We simulated two types of mutations: neutral mutations and beneficial mutations with additive effects and selection coefficient $s = 0.1$. With these parameters, $2Ns = 100$, so beneficial mutations are strongly selected and will sweep in $T_{\text{sweep}} \sim s^{-1} \log Ns \approx 50$ generations. Such sweeps will effect a region of the chromosome $d_{\text{sweep}} \sim (rT_{\text{sweep}})^{-1} \approx 2 \times 10^6$ basepairs long. Thus $d_{\text{sweep}} \ll L$, which will minimize edge effects of having a finite chromosome.

In order to vary the effects of sweeps on neutral diversity, we varied the fraction of mutations that are beneficial f_{sel} over several orders of magnitude: $f_{\text{sel}} \in \{10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}\}$. For each f_{sel}, we ran 100 independent replicate simulations and computed η_i and η_{ij} of neutral mutations averaged over all replicates.

`SLiM` parameter files, `python` wrapper scripts for parsing output, and `snakemake` files for running simulations are available [GITHUB].

Analysis of *D. melanogaster* data

The DPGP3 data set consists of haploid consensus sequences from $\sim$ 200 files, obtained via the haploid embryo method of **LangleyEtal2011** We obtained the DPGP3 consensus sequence files version 1.1 from `www.johnpool.net/genomes.html`. These files contain the sequence alignments of all flies in the sample on all chromosome arms. We also downloaded the Nov. 3, 2016 spreadsheet of inversions available at the link above. For each chromosome arm, we excluded any samples with an inversion in that arm and then down-sampled to $n = 100$ by taking the first 100 remaining samples in alphanumeric order of sample name. (This means that the data for each chromosome arm is from a slightly different subset of the individuals.)

- data pre-processing: throw out inversions, downsampling to 100 chromosome, 4-D sites, filter for coverage
- picking the central regions
- fitting demographic model with fastNeutrino: fit 3-epoch piecewise constant to SFS. Report model parameters
- simulating demographic model
- computing statistics

References

Bhaskar, Anand and Yun S Song (2014). "Descartes' Rule of Signs and the Identifiability of Population Demographic Models from Genomic Variation Data". *Ann. Stat.* 42.6, pp. 2469–2493.

Bhaskar, Anand, Y X Rachel Wang, and Yun S Song (2015). "Efficient Inference of Population Size Histories and Locus-Specific Mutation Rates from Large-Sample Genomic Variation Data". *Genome Res.* 25.2, pp. 268–279.

Birkner, Matthias, Jochen Blath, and Bjarki Eldon (2013). "Statistical Properties of the Site-Frequency Spectrum Associated with Lambda-Coalescents". *Genetics* 195.3, pp. 1037–1053.

Blath, Jochen et al. (2016). "The Site-Frequency Spectrum Associated with Ξ-Coalescents". *Theor. Popul. Biol.* 110, pp. 36–50.

Corbett-Detig, Russell B., Daniel L. Hartl, and Timothy B. Sackton (2015). "Natural Selection Constrains Neutral Diversity across A Wide Range of Species". *PLoS Biology* 13.4.

Coventry, Alex et al. (2010). "Deep Resequencing Reveals Excess Rare Recent Variants Consistent with Explosive Population Growth". *Nature Communications* 1, p. 131.

Cvijović, Ivana, Benjamin H. Good, and Michael M. Desai (2018). "The Effect of Strong Purifying Selection on Genetic Diversity". *Genetics*, genetics.301058.2018.

Durrett, Rick and Jason Schweinsberg (2005). "A Coalescent Model for the Effect of Advantageous Mutations on the Genealogy of a Population". *Stochastic Processes and their Applications* 115.10, pp. 1628–1657.

Eldon, Bjarki (2016). "Inference Methods for Multiple Merger Coalescents". In: *Evolutionary Biology: Convergent Evolution, Evolution of Complex Traits, Concepts and Methods*. Ed. by Pierre Pontarotti. Cham: Springer International Publishing, pp. 347–371.

Eldon, Bjarki, Matthias Birkner, et al. (2015). "Can the Site-Frequency Spectrum Distinguish Exponential Population Growth from Multiple-Merger Coalescents?" *Genetics* 199.3, pp. 841–856.

Eldon, Bjarki and John Wakeley (2006). "Coalescent Processes When the Distribution of Offspring Number Among Individuals Is Highly Skewed". *Genetics* 172.4, pp. 2621–2633.

Excoffier, Laurent et al. (2013). "Robust Demographic Inference from Genomic and SNP Data". *PLOS Genetics* 9.10, e1003905.

Ferretti, Luca et al. (2018). "The Neutral Frequency Spectrum of Linked Sites". *Theoretical Population Biology*.

Fu, Y X (1995). "Statistical Properties of Segregating Sites". *Theor. Popul. Biol.* 48.2, pp. 172–197.

Garud, Nandita R and Dmitri A Petrov (2016). "Elevated Linkage Disequilibrium and Signatures of Soft Sweeps Are Common in Drosophila Melanogaster". *Genetics* 203.2, pp. 863–880.

Griffiths, R. C. and S. Tavaré (1994). "Sampling Theory for Neutral Alleles in a Varying Environment". *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 344.1310, pp. 403–410.

Griffiths, R C and Simon Tavaré (1998). "The Age of a Mutation in a General Coalescent Tree". *Communications in Statistics. Stochastic Models* 14.1-2, pp. 273–295.

Gutenkunst, Ryan N. et al. (2009). "Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data". *PLOS Genetics* 5.10, e1000695.

Hahn, M.W. (2018). *Molecular Population Genetics*. Sinauer Series. Oxford University Press.

Hudson, Richard R. (2002). "Generating Samples under a Wright–Fisher Neutral Model of Genetic Variation". *Bioinformatics* 18.2, pp. 337–338.

Jouganous, Julien et al. (2017). "Inferring the Joint Demographic History of Multiple Populations: Beyond the Diffusion Approximation". *Genetics* 206.3, pp. 1549–1567.

Kern, Andrew D. and Matthew W. Hahn (2018). "The Neutral Theory in Light of Natural Selection". *Molecular Biology and Evolution* 35.6, pp. 1366–1371.

Lack, Justin B et al. (2015). "The Drosophila Genome Nexus: A Population Genomic Resource of 623 Drosophila Melanogaster Genomes, Including 197 from a Single Ancestral Range Population". *Genetics* 199.4, pp. 1229–1241.

Li, Heng and Richard Durbin (2011). "Inference of Human Population History from Individual Whole-Genome Sequences". *Nature* 475.7357, pp. 493–496.

Lukić, Sergio, Jody Hey, and Kevin Chen (2011). "Non-Equilibrium Allele Frequency Spectra via Spectral Methods". *Theoretical Population Biology* 79.4, pp. 203–219.

Myers, Simon, Charles Fefferman, and Nick Patterson (2008). "Can One Learn History from the Allelic Spectrum?" *Theoretical Population Biology* 73.3, pp. 342–348.

Nielsen, Rasmus (2000). "Estimation of Population Parameters and Recombination Rates From Single Nucleotide Polymorphisms". *Genetics* 154.2, pp. 931–942.

Polanski, A and M Kimmel (2003). "New Explicit Expressions for Relative Frequencies of Single-Nucleotide Polymorphisms with Application to Statistical Inference on Population Growth". *Genetics* 165.1, pp. 427–436.

Ragsdale, Aaron P and Ryan N Gutenkunst (2017). "Inferring Demographic History Using Two-Locus Statistics". *Genetics* 206.2, pp. 1037–1048.

Rödelsperger, Christian et al. (2014). "Characterization of Genetic Diversity in the Nematode Pristionchus Pacificus from Population-Scale Resequencing Data". *Genetics* 196.4, pp. 1153–1165.

Rosenberg, Noah A. and Magnus Nordborg (2002). "Genealogical Trees, Coalescent Theory and the Analysis of Genetic Polymorphisms". *Nature Reviews Genetics* 3.5, pp. 380–390.

Sargsyan, Ori and John Wakeley (2008). "A Coalescent Process with Simultaneous Multiple Mergers for Approximating the Gene Genealogies of Many Marine Organisms". *Theor. Popul. Biol.* 74.1, pp. 104–114.

Schrider, Daniel R., Alexander G. Shanku, and Andrew D. Kern (2016). "Effects of Linked Selective Sweeps on Demographic Inference and Model Selection". *Genetics* 204.3, pp. 1207–1223.

Sella, Guy et al. (2009). "Pervasive Natural Selection in the Drosophila Genome?" *PLOS Genetics* 5.6, e1000495.

Spence, Jeffrey P, John A Kamm, and Yun S Song (2016). "The Site Frequency Spectrum for General Coalescents". *Genetics* 202.4, pp. 1549–1561.

Tajima, Fumio (1983). "Evolutionary Relationship of DNA Sequences in Finite Populations". *Genetics* 105.2, pp. 437–460.

Terhorst, Jonathan, John A. Kamm, and Yun S. Song (2017). "Robust and Scalable Inference of Population History from Hundreds of Unphased Whole Genomes". *Nature Genetics* 49.2, pp. 303–309.

Wakeley, John (2009). *Coalescent Theory: An Introduction*. Greenwood Village, Colorado: Roberts & Company. 352 pp.

Živković, Daniel and Thomas Wiehe (2008). "Second-Order Moments of Segregating Sites Under Variable Population Size". *Genetics* 180.1, pp. 341–357.

Tables

Table 1: [Table of parameters underlying Figure 2]

Table 2: [Table of parameters underlying Figure 4]

Table 3: [Table of parameters underlying Figure 6]

Table 4: [DPGP3 model fit]