Distinguishing multiple-merger from Kingman coalescence using two-site frequency spectra

Daniel P. Rice[1], John Novembre[1,*], and Michael M. Desai[2,*]

[1]Department of Human Genetics, University of Chicago, Chicago, IL
[2]Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA
[*]These authors contributed equally to this work.

August 29, 2018

Abstract

Population genetics methods for demographic inference 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 and extreme variation in offspring number can both 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 in genomic data rely on the site frequency spectrum (SFS). However, the signatures of multiple mergers in the SFS are also consistent with a Kingman coalescent process with a time-varying population size. Here, we present a new method for detecting multiple mergers based on the pointwise mutual information of the two-site frequency spectrum for pairs of linked sites. Unlike the SFS, the pointwise mutual information depends mostly on the topologies of genealogies rather than their branch lengths and is therefore insensitive to population size change. This statistic is global in the sense that it can detect when the genome-wide genetic diversity is inconsistent with the Kingman coalescent, rather than detecting outlier regions, as in selection scan methods.

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 Hahn 2018). The fundamental tools of the trade are simplified mathematical models, which connect unobserved quantities such as the population size to observable features of genetic data. 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 widely used models is the Kingman coalescent (Kingman 1982b; Kingman 1982a; Hudson 1983; 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 genetic data (Rosenberg and Nordborg 2002). The Kingman coalescent has a number of convenient properties that allow for both analytical calculations (e.g., Tajima 1989) 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 can be extended to study a variety of biological phenomena including recombination, population structure, and variation in sex ratios or ploidy (see generally Wakeley 2009).

An important application of the Kingman coalescent is inferring historical population sizes from genetic data (Schraiber and Akey 2015). In its simplest form, the model has a single parameter, the coalescent rate, which determines the branch lengths of genealogies (Kingman 1982b). Under a variety of conditions, the coalescent rate is inversely proportional to the population size (Kingman 1982a). Accordingly, a growing or shrinking population may be modeled by a time-varying rate (Griffiths and Tavaré 1994; Griffiths and Tavaré 1998). Patterns of genetic diversity depend on the ratio of the coalescent rate to other evolutionary rate parameters. For example, the *site frequency spectrum* (SFS)—the number of mutations segregating at different frequencies in a sample—is determined by the ratio of the mutation rate to the (time-varying) coalescent rate. Kingman coalescent–based inference methods solve the inverse problem of determining the population size history that best explains particular features of the data, such as the SFS (e.g., Bhaskar, Wang, et al. 2015) or variations in heterozygocity along a chromosome (e.g., Li and Durbin 2011).

A serious problem for this class of inference methods is that different models of evolution generate different relationships between historical population sizes and genetic diversity. For example, 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 instance, when a beneficial mutation increases rapidly in frequency, it distorts the genealogies at nearby sites (see e.g., Coop and Ralph 2012). If these "selective sweeps" occur regularly, they may be the dominant factor determining the distribution of genealogies. In this case, the average coalescent rate is proportional to the number of beneficial mutations introduced per generation, which is itself *directly* rather than inversely proportional to the population size. It follows that the relationship between the population size and the expected number of neutral mutations in a sample is inverted: larger populations will be less diverse than smaller populations.

While the example above is extreme, it is well-established that violations of the neutrality assumption can distort or mask the signatures of population size changes. For example, Schrider et al. 2016 recently demonstrated that several popular inference methods give misleading results in the presence of selective sweeps. In a similar vein, Cvijović et al. 2018, showed that purifying selection at linked sites that is sufficient to reduce genetic diversity is also sufficient to distort the SFS, leading to a false signal of population growth. Moreover, genomic evidence from multiple species suggests that such violations of neutrality may be widespread (Sella et al. 2009; Corbett-Detig et al. 2015; Kern and Hahn 2018).

An important extension of the Kingman coalescent is a family of models known as *multiple merger coalescents* (Pitman 1999; Sagitov 1999; Donnelly and Kurtz 1999) (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 (Schweinsberg 2000; Möhle and Sagitov 2001; Sagitov 2003) permits more than one distinct 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 (Schweinsberg 2003), recurring selective sweeps at linked sites (Durrett and Schweinsberg 2005; Coop and Ralph 2012), rapid adaptation (Neher and Hallatschek 2013; Desai et al. 2013), and purifying selection at sufficiently dense sites (Seger et al. 2010; Nicolaisen and Desai 2012).

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 (Schweinsberg 2003), while with linked sweeps it is determined by rate of linked sweeps, as described above (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. This task is distinct from "selection scan" methods designed to detect particular regions of the genome that are under selection (see Vitti et al. 2013). These methods typically assume that most of the genome is evolving neutrally and that the genome-wide distribution of summary statistics reflects demographic factors. Genomic regions that are outliers from this distribution are presumed to be under selection. In contrast, we are interested in detecting when the genome-wide background *is not* well-modeled by the Kingman coalescent.

One approach to identifying multiple mergers in genomic data is to use the SFS as a summary statistic. Birkner et al. 2013; Blath et al. 2016; Spence et al. 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 widespread linked selection in the nematode *Pristionchus pacificus* by demonstrating that the site frequency spectrum is non-monotonic, a signature of multiple mergers (Neher and Hallatschek 2013; Birkner et al. 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 et al. 2008; Bhaskar and 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 two-site frequency spectrum (2-SFS)—the generalization of the SFS to pairs of nearby sites (Hudson 2001; Ferretti et al. 2018))—are useful for dis-

tinguishing between the Kingman coalescent with population growth and multiple merger coalescents. These statistics may be calculated efficiently from genomic single nucleotide variant 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.

Following the notation of Fu 1995, the site frequency spectrum of a sample of n haploid genomes 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 complementary 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 can be calculated from a set of single nucleotide polymorphisms (SNPs) without any information about their relative locations in the genome.

In contrast, the 2-SFS is a statistic of *pairs* of sites. We define the 2-SFS, $\{\xi_{ij}(d) : d > 0; 1 \leq i, j < n\}$, 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 Xie 2011 in a model with selection. We define the folded 2-SFS, $\eta_{ij}(d)$, by analogy to the folded SFS, categorizing pairs of sites by their minor allele frequencies. (For non-recombining sites, the 2-SFS is independent of the distance and so we will suppress the d in our notation.)

In the limit of low per-site mutation rate ($\mu \to 0$), and no recombination, all polymorphic sites are bi-allelic and the expected SFS and 2-SFS are related to moments of the genealogical 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 $j \notin \{i, (n-i)\}$. 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 $j \notin \{i, (n-i)\}$. (There are positive correlations between mutations at complementary frequencies induced by genealogies whose root node partitions the tree into subtrees of size i and $n-i$.)

Birkner et al. 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 (Schweinsberg 2003; Steinrücken et al. 2013), and the Bolthausen-Sznitman coalescent is the limiting distribution of genealogies in rapidly adapting populations (Neher and Hallatschek 2013). Like Fu, Birkner et al. were primarily concerned with computing the sample variance of SFS-based summary statistics such as Tajima's D (Tajima 1989). 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 et al. 2013 show positive correlations between ξ_i and ξ_j for $j \notin \{i, n-i\}$. Thus, unlike the standard Kingman coalescent, the beta coalescent can generate positive associations between mutations with different minor allele counts.

In the following, we demonstrate that the positive associations between mutations at different frequencies distinguish the multiple-merger from the Kingman coalescent and that this distinction: (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 do so, we introduce a transformation of the 2-SFS that we call frequency pointwise mutual information (fPMI). We use a combination of numerical calculations and stochastic simulations to explore the properties of this statistic under different coalescent models. Finally, we demonstrate the use of fPMI in a coalescent model-checking procedure on genomic diversity data from *Drosophila melanogaster* (Lack et al. 2015).

Methods

Frequency pointwise mutual information

We are interested in quantifying the dependence between minor allele counts at a pair of sites, particularly for $i \neq j$. To do so, we define the

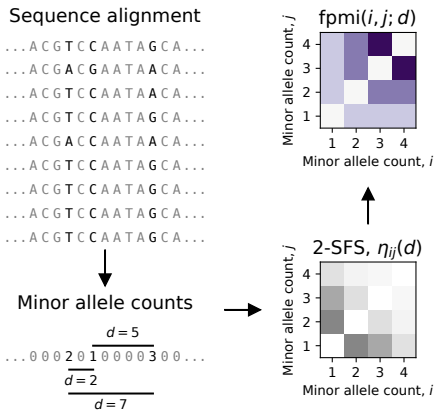

Figure 1: Summarizing population genetic data with frequency pointwise information. First, a sequence alignment is reduced to the minor allele frequencies and positions of segregating mutations. Then, for a given distance, d, the 2-SFS gives the number of pairs of sites with each combination of minor allele frequencies. Finally, fPMI is calculated from the 2-SFS according to Eq. (3). [OPTIONS: Could add (1) weighted fPMI, and/or (2) binned 2-SFS]

frequency pointwise mutual information as

$$\mathrm{fPMI}(i,j;d) = \log \frac{\langle \eta_{i,j}(d) \rangle}{\langle \eta_i \rangle \langle \eta_j \rangle}. \tag{3}$$

Figure 1 shows the steps to compute fPMI from a sample of sequences.

In information theory, the pointwise mutual information (PMI) of a pair of random variables, X and Y, is a transformation of the probability mass function, $p(x, y)$, given by

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

where $p(x) = \sum_y p(x, y)$ and vice versa (Church and Hanks 1990). PMI measures the change in the probability that $X = x$ given knowledge that $Y = y$. When X and Y are independent, $\mathrm{PMI}(x, y) = 0$ for all x, y. On the other hand, when X and Y are not independent, $\mathrm{PMI}(x, y) > 0$ implies that $p(x|y) > p(x)$, and conversely for $\mathrm{PMI}(x, y) < 0$. (The expectation of PMI over the joint distribution of X and Y is known as the mutual information of X and Y (Cover and Thomas 1991).)

In our setting, the 2-SFS may be interpreted as a joint probability mass function over minor allele counts. Given a coalescent model, the minor allele count at an arbitrary site is a random variable over $\{0, \ldots, \lfloor n/2 \rfloor\}$ 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 $p_d(i, j) = \langle \eta_{ij}(d) \rangle$. Comparing, Eq. (3) with Eq. (4) shows that fPMI is the standard pointwise mutual information for pairs of minor allele counts.

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 imply that $\text{fPMI}(i, j) < 0$ for $i \neq j$ for non-recombining sites under the time-homogeneous Kingman coalescent.

We claim that the second benefit of computing fPMI from the 2-SFS is that it normalizes for distortions in the coalescent time distribution. 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 fPMI reflects additional information about the distribution of genealogies, once distortions in coalescent times are accounted for. That is, we expect fPMI to be sensitive to multiple mergers, but invariant under changes in the coalescent rate.

Furthermore, the same normalization renders fPMI insensitive to ascertainment bias. Variant detection methods typically compare sampled sequences to a reference sequence and "call" a variant site when there is sufficient evidence that at least one sample differs from the reference at that site. As a result, the ascertainment probability, p_i, of a mutation is an increasing function of its true allele count i. This effect distorts the expected ascertained SFS ($\langle \eta_i \rangle_{\text{asc}} = \mu \langle \tau_i \rangle p_i$) toward high-frequency mutations. However, provided that variant detection at each site is approximately independent, the primary effect of ascertainment bias in Eq. (3) is to multiply the numerator and denominator by a common factor of $p_i p_j$. Because these factors cancel, fPMI is less sensitive than the raw SFS and 2-SFS to ascertainment effects.

Binned allele 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 using a binned form of the SFS and 2-SFS:

$$\eta_{\mathrm{lo}}(i_c) = \sum_{i=1}^{i_c-1} \eta_i \tag{5}$$

$$\eta_{\mathrm{hi}}(i_c) = \sum_{i=i_c}^{\lfloor n/2 \rfloor} \eta_i \tag{6}$$

$$\eta_{\mathrm{hi,lo}}(d; i_c) = \sum_{i=1}^{i_c-1} \sum_{j=i_c}^{\lfloor n/2 \rfloor} \eta_{ij}(d), \tag{7}$$

where i_c is an arbitrary cutoff between high and low minor allele frequency. Binning 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. It is possible to define less drastic coarse-graining schemes that strike a balance between sampling noise and preserving more detailed information about allele frequencies.

We can also compute the pointwise mutual information in this binned distribution as

$$\mathrm{hiloPMI}(d; i_c) = \log \frac{\langle \eta_{\mathrm{hi,lo}}(d; i_c) \rangle}{\langle \eta_{\mathrm{lo}}(i_c) \rangle \langle \eta_{\mathrm{hi}}(i_c) \rangle}. \tag{8}$$

One could similarly calculate five other pointwise mutual information statistics from the binning 2-SFS (e.g., the PMI between monomorphic sites and sites with high minor allele counts), but we will not use these statistics here.

Weighted fPMI

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

$$\mathrm{wfPMI}(i, j : d) = \frac{\langle \eta_{ij} \rangle}{\mu^2 \langle T_2 \rangle^2} \mathrm{fPMI}(i, j; d), \tag{9}$$

where T_2 is the coalescence time for a sample of size two. The numerator of the weighting factor in Eq. (9) serves to emphasize the most common pairs of minor allele counts. The denominator, which is proportional to the square of the expected pairwise diversity, Π, ensures that wfPMI is invariant under changes in the overall mutation and coalescence rates.

Computing branch-length moments

We implemented numerical computations of the moments of the branch lengths $\langle \tau_i \rangle$ and $\langle \tau_{ij} \rangle$. For the Kingman coalescent with time-varying coalescent rate, we used equations (1)-(12) of Živković and Wiehe 2008. For the beta coalescent, we implemented the recursion described by Birkner et al. 2013.

Functions for computing the branch-length moments are implimented in `python` using the numerical package `numpy` (**numpy**) and 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 Griffiths and Tavaré 1994. They are therefore only practical for sample sizes up to $n \approx 40$. The recursion of Birkner et al. 2013 is $\mathcal{O}(n^4)$ and is also only practical for samples up to $n \approx 50$.

Coalescent simulations

We ran coalescent simulations using a custom version of `msprime` (Kelleher et al. 2016) capable of multiple mergers, based on modifications made by Joe Zhu [GITHUB]. 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].

For each coalescent model, we simulated at least 10^4 independent infinite-sites loci, each with length d and per basepair recombination rate r resulting in a per locus total recombination rate dr. (See SI for parameter combinations.) We chose values of dr so that the realized values of $dr \langle T_2 \rangle$ varied over several orders of magnitude. 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 $\{\langle \tau_{ij} \rangle\}$ by averaging over independent loci. These expectations allow us to calculate all of the 2-SFS statistics defined above.

Forward-time simulations of selective sweeps

We simulated a model of recurring selective sweeps using the software `SLiM` (Messer 2013). 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 $d_{\text{sweep}} \sim (rT_{\text{sweep}})^{-1} \approx$

2×10^6 basepairs long. Thus $d_{\text{sweep}} \ll L$, which will minimize edge effects of simulating 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

We analyzed sequence data from the DPGP3 data set, which consists of haploid consensus sequences from ~ 200 flies, obtained via the haploid embryo method of Langley et al. 2011. The SNP calls underlying these sequences have been subjected to a variety of quality filters described in Lack et al. 2015. 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 by sample name. (As a result, the data for each chromosome arm is from a slightly different subset of the individuals.)

We calculated the average pairwise diversity, Π, as a function of position for each autosomal chromosome arm (Fig. 10). Pairwise diversity is high in the middle of each chromosome arm and lower near the centromeres and telomeres. Our modeling—and coalescent-based demographic inference methods in general—assumes that the distribution of gene genealogies is homogeneous along the chromosome. Therefore, we selected a 13-16 MB "central" region of each arm with relatively homogeneous values of Π for further analysis. The boundaries of these central regions are given in Table 1.

In order to ensure that the segregating mutations reflect true genetic diversity and not variation in calling errors, we excluded sites where fewer than 90 of the 100 samples have genotype calls. This leaves over 90% of all sites and does not substantially alter the fraction of sites that are polymorphic in the sample (Table 2)

We fit a demographic model to the folded SFS of fourfold degenerate sites for each chromosome arm separately using `fastNeutrino` (Bhaskar, Wang, et al. 2015). Following Ragsdale and Gutenkunst

2017 we fit a three-epoch piecewise constant–N model:

$$N(t) = \begin{cases} N_1 & 0 \leq t \leq t_1 \\ N_2 & t_1 < t \leq t_2 \\ N_{\text{anc}} & t > t_2, \end{cases} \tag{10}$$

estimating both change-points (t_1 ¡ t_2) and both population sizes (N_1, N_2). We specified the ancient population size $N_{\text{anc}} = 3 \times 10^5$, as in Ragsdale and Gutenkunst 2017. For all four chromosomes, `fastNeutrino` inferred similar population growth. Fitted parameters are presented in Table 3. We simulated the SFS and 2-SFS under the fitted parameters using `msprime`.

In addition to the average SFS used to fit the model, we computed the average 2-SFS for pairs of sites at distances distances between 3 bp and 5 Kb. Because we are using four-fold degenerate sites, we only computed the 2-SFS for distances that are multiples of 3. For comparison between data and simulations, we scale the distances in basepairs by a critical distance $d_c = (r \langle T_2 \rangle)^{-1}$. We used a genome-wide recombination rate of $r = 2 \times 10^{-8}$ per basepair per generation (Comeron et al. 2012). We estimated $\langle T_2 \rangle$ by $\Pi/2\mu$. For Π, we used the average pairwise diversity at fourfold degenerate sites in the central region of each chromosome arm. For μ, we used a genome-wide mutation rate of 3×10^{-9} per basepair per generation (Keightley et al. 2014). These estimates are not precise, but only serve to scale genetic distances to the correct order of magnitude.

A `snakemake` pipeline, `python` scripts, and `jupyter` notebooks to replicate our data processing, model fitting, simulations, and analysis of the DPGP3 data are available [GITHUB].

Results

Population growth versus the beta coalescent in non-recombining loci

We compared 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 are interested in whether fPMI can distinguish beta from Kingman coalescent models that produce similar distortions in the SFS. To this end, we computed a version of Tajima's D (Tajima 1989) normalized to be invariant under changes in the mean coalescent rate: $D = (\Pi - \hat{\theta}_W)/\Pi$, where $\hat{\theta}_W$ is Watterson's theta, a linear function of the SFS (Watterson 1975). Negative values of D indicate an overabundance of low-frequency mutations relative to the time-homogeneous Kingman expectation. All results are computed

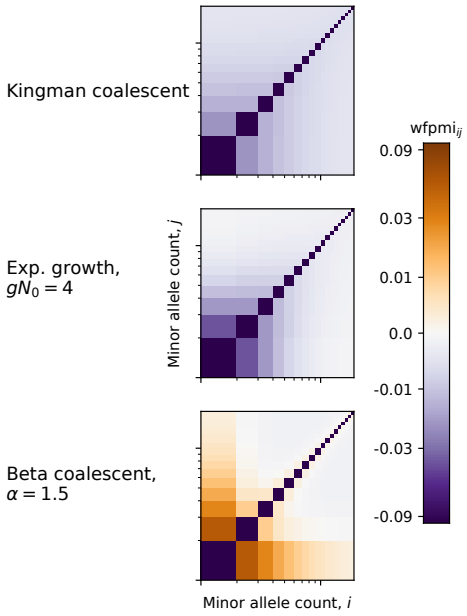

Figure 2: Weighted frequency pointwise mutual information for three coalescent models: the time-homogeneous Kingman coalescent, the Kingman coalescent with exponential growth ($g = 4$), and the beta coalescent ($\alpha = 1.45$). For all models $n = 39$. Growth and α parameters were chosen to generate average SFS with similar Tajima's D. The diagonal $i = j$ is masked.

numerically using the results of Fu 1995, Živković and Wiehe 2008, and Birkner et al. 2013 (see Methods).

As a first example, Fig. 2 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. Exponential growth with $g = 4$ and the beta coalescent with $\alpha = 1.45$ both generate similar substantial distortions in the SFS $D = -0.4$. However, exponential growth 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, the beta coalescent generates positive fPMI. The effect of multiple mergers is strongest on fPMI between high and low minor allele counts. This justifies binning the SFS and 2-SFS into high– and low–minor allele count bins as defined above.

To generalize this finding, we computed the binned hiloPMI between singletons and non-singletons ($i_c = 1$) for a range of g and α. Figure 3 shows that as the population growth rate increases, distort-

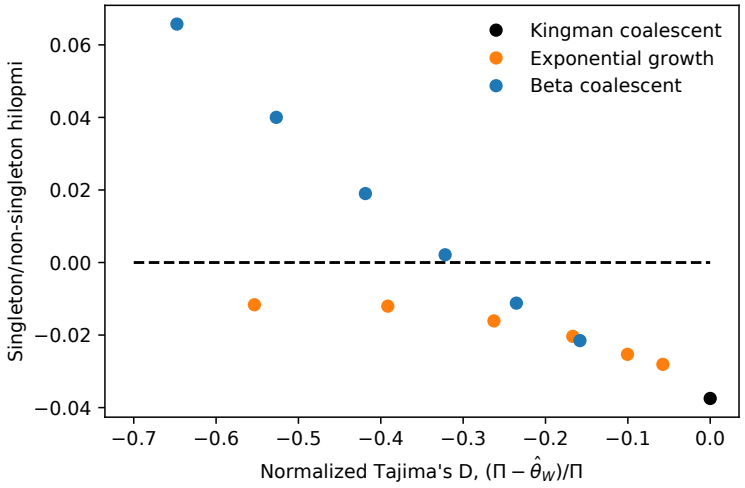

Figure 3: Pointwise mutual information between singletons and non-singletons for Kingman coalescents with exponential growth and beta coalescents. The exponential growth rate, g, ranges from 0.25 to 8.0 in coalescent time units. The beta coalescent parameter, α, ranges from 1.75 (nearly Kingman) to 1.25 (nearly Bolthausen-Sznitman).

ing 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.

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. 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 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`

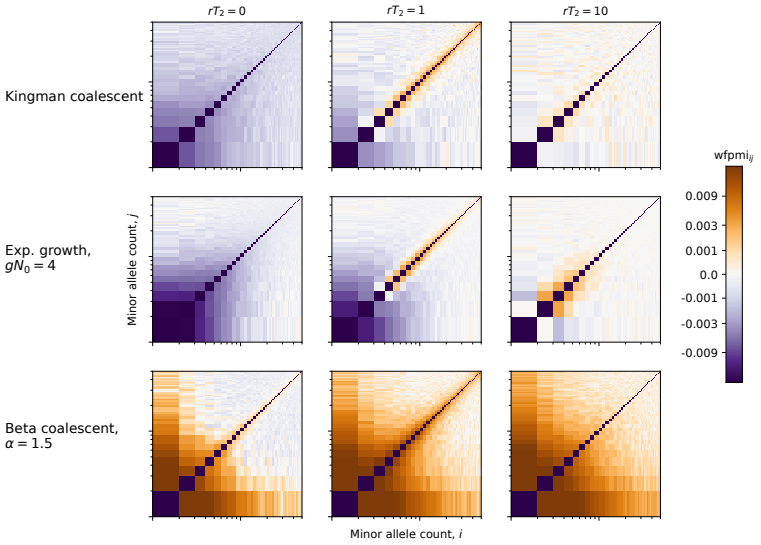

Figure 4: Weighted frequency pointwise information for three recombination rates with sample size $n = 100$. As in Figure 2, the diagonal elements are masked, and the growth and multiple-merger parameters were chosen to generate average SFS with similar Tajima's D.

KehellerEtAl201? modified to allow for multiple mergers (See Methods). In particular, we simulated a model where the sizes of 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 et al. 2013. We also simulated data from two models of population growth: exponential growth and a two-epoch piecewise-constant model.

Figure 4 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 $dr\,\langle T_2 \rangle \sim 1$. This is presumably because trees at nearby sites contain clades with similar numbers of leaves. These positive correlations do not extend to $i \gg j$, which is the signal of multiple mergers in our binned hiloPMI. On the other hand, the positive fPMI in the beta coalescent persists for $dr\,\langle T_2 \rangle > 1$. Thus, the signal of multiple mergers in fPMI is robust to recombination.

Figure 5 shows $\text{hiloPMI}(d, i_c)$ in both models of growth and the

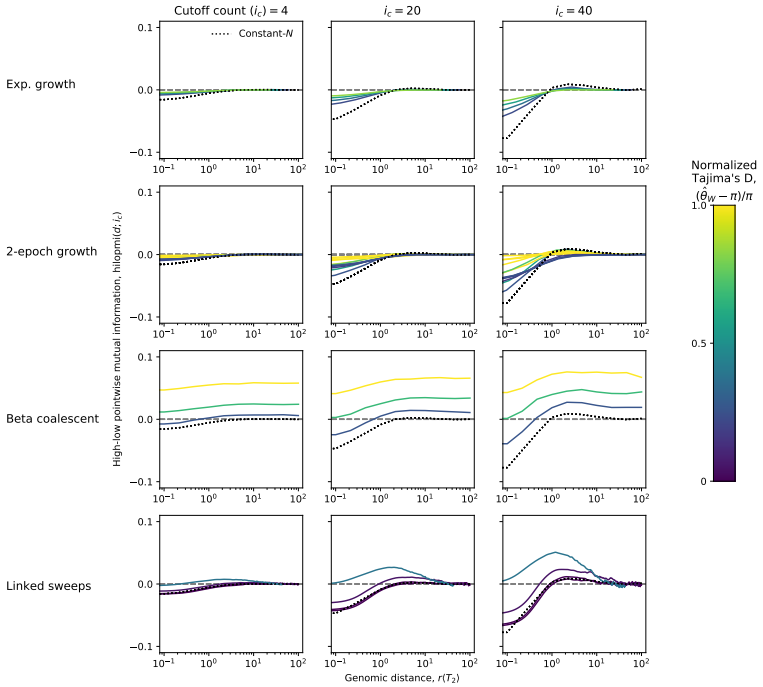

Figure 5: Pointwise mutual information between high– and low–minor allele count mutations (hiloPMI) versus genetic distance. The first three rows show coalescent simulations in `msprime`. Exponential growth rate g ranges from 0.01 to 8.0 in coalescent time units. Piecewise constant change-points, range from 0.01 to 1 in coalescent time units, and the ratio of ancestral to modern population size ranges from 0.01 to 0.2. The fourth row shows forward-time `SLiM` simulations with selective sweeps at linked sites (beneficial mutation fraction $f_{\text{sel}} \in \{10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}\}$, see Methods). Lines are colored by the distortion in the SFS, as measured by Tajima's D. [TODO: SPLIT SWEEPS PANELS OFF INTO THEIR OWN FIGURE.]

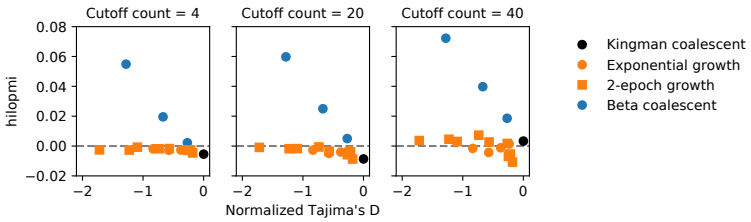

Figure 6: Pointwise mutual information between high– and low–minor allele count mutations (hiloPMI) versus normalized Tajima's D for simulations with recombination. All values are for $dr\,\langle T_2\rangle \sim 1$ and $n = 100$. Orange points show Kingman coalescent simulations with population growth. Blue points show beta coalescent simulations. [Q: Should this be the bottom row of the previous figure?]

beta coalescent, for a range of $dr\,\langle T_2\rangle$ 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 $dr\,\langle T_2\rangle \geq 1$, 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 $dr\,\langle T_2\rangle \gg 1$.

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. 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.

Figure 6 shows the relative invariance of hiloPMI to a time-varying coalescent rate. While population growth in a Kingman coalescent model can strongly distort the SFS, as measured by Tajima's D, it has very little effect on the hiloPMI for sites $dr\,\langle T_2\rangle \sim 1$ apart. This holds for exponential as well as piecewise constant growth. In contrast, the beta coalescent induces large positive hiloPMI for the same values of Tajima's D. The figure also showes that this result is robust to the choice of cutoff minor allele count for binning.

There is one other salient feature of Fig. 5: hiloPMI does not decay to zero at long distances in the beta coalescent. This is related to the

results of Eldon and Wakeley 2006, 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 in 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., Durrett and Schweinsberg 2005; Coop and Ralph 2012; Neher and Hallatschek 2013; Desai et al. 2013; Seger et al. 2010). 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` (Messer 2013). 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.

The bottom row of Fig. 5 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.

Application to *Drosophila melanogaster*: Coalescent model checking

Our results above show that fPMI and its binned 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.

(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 not 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 $\sim$ 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 et al. 2017; Ragsdale and 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 et al. 2016; 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` Bhaskar, Wang, et al. 2015. We fit a 3-epoch piecewise-constant model, with four free parameters: two changepoints and two population size ratios. We report our fitted parameters in Table 3. We then simulated under our fitted

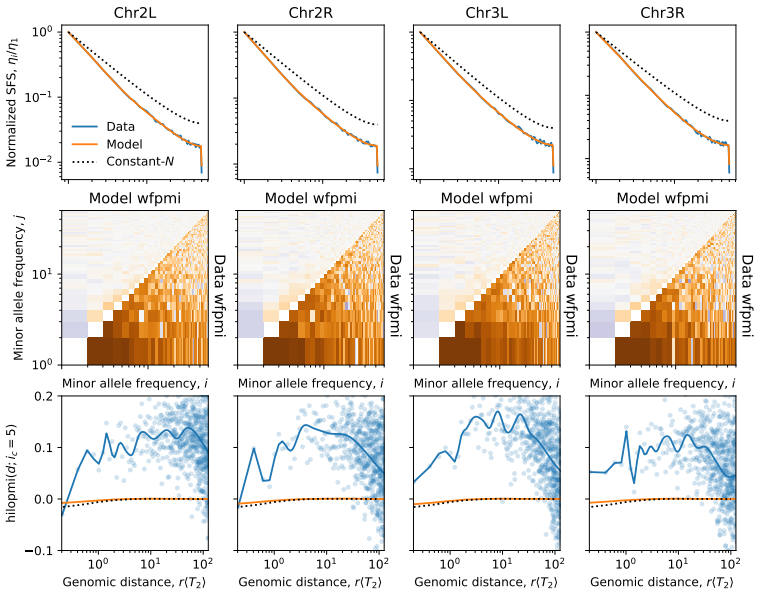

Figure 7: DPGP3 data and coalescent model predictions. First row: observed (blue) and expected (orange) site frequency spectrum compared to constant-N Kingman coalescent (dotted lines). Second row: Expected (upper triangle) and observed (lower triangle) weighted fPMI averaged over all pairs of sites 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]

model using `msprime` and computed the expected and observed SFS, fPMI, and hiloPMI (Fig. 7, See methods for details.).

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 $d_c r \langle T_2 \rangle = 1$. 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.

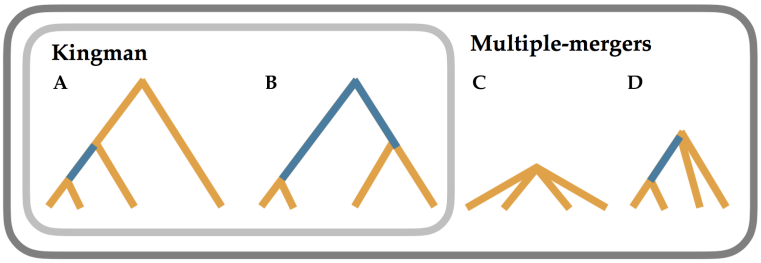

Figure 8: Genealogies with a sample size of $n = 4$. Opportunities for singleton/tripleton mutations are in orange. Opportunities for doubletons are in blue. (Note: there is a third possible multiple-merger topology, not shown.)

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

We have shown that fPMI and its binned analog hiloPMI are sensitive to multiple mergers, but relatively invariant under population growth in the Kingman coalescent. These properties make them well-suited for coalescent model checking. We have demonstrated a model-checking procedure on data from *D. melanogaster*, which is believed to be strongly shaped by natural selection, and found evidence that population growth can not explain the positive associations between high and low frequency mutations.

We can get an intuitive understanding for why fPMI distinguishes among coalescent models by considering a sample of four chromosomes. In the Kingman coalescent, there are only two possible tree topologies (Figure 8). 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 $\langle \eta_{12} \rangle < \langle \eta_1 \rangle \langle \eta_2 \rangle$.

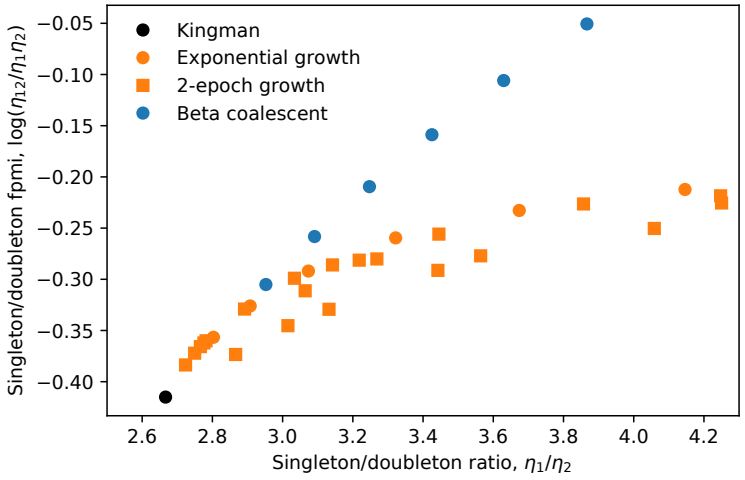

Figure 9: Distortions in fPMI vs. distortions in the SFS for $n = 4$. Parameters for beta coalescents and exponential growth are as in Fig. 3. For the piecewise-constant–N model, the fold-change in N varies from 2 to 2^4 and the time of this change varies from 2^{-4} to 2 in coalescent time units.

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 9 demonstrates that fPMI accounts for this fact by normalizing by the SFS. Figure 9 plots fPMI$(1, 2)$ 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. Like Tajima's D, 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. 3), mulitple mergers generate larger distortions in fPMI than population growth does, accounting for the distortions in the SFS. Moreover, the results for two different models of growth coincide. This suggests that for the Kingman coalescent with

growth, $\text{fPMI}(1,2) \approx f(\eta_1/\eta_2)$ for some function f.

We focus here on demographic inference methods that explicitly use the Kingman coalescent model for calculations. However, another popular class of methods are based on forward-time models, such as the Wright-Fisher model (e.g., Gutenkunst et al. 2009, Sheehan et al. 2013). We believe that fPMI is useful for model-checking with these methods as well. This is because forward-time neutral models each have an associated dual coalescent model Etheridge 2011, which can be used to compute the expected fPMI as outlined in this paper. Furthermore, non-neutral models also generate predictions about the 2-SFS and thus fPMI. In principle, any fitted model that allows calculation or simulation of the 2-SFS may be checked using fPMI. However, the suitability of fPMI for discriminating among arbitrary families of models is unknown.

In this paper, we have outlined a graphical model-checking procedure in the spirit of Anscombe 1973. One could extend our work to implement a formal hypothesis-testing framework for rejecting the best-fit Kingman coalescent model according to some inference method. Numerical calculations would require higher moments of the branch-length distribution, which may be computationally intractable for large samples. On the other hand, it would be straightforward to develop a bootstrap-style procedure, using coalescent simulations to estimate the variance in test statistics under the fit model. In any case, we believe that it would be useful for simulation packages such as `msprime` and population genetics analysis libraries to include standard functions for computing the 2-SFS, fPMI, and hiloPMI. We provide such functions in the [GITHUB] repository associated with this paper.

Our coalescent simulations implement a particular version of the multiple mergers coalescent with recombination. However, the precise correspondence between this model and any particular forward-time process is unclear. In order to compare with our numerical calculations based on Birkner et al. 2013, we use the beta coalescent for the distribution of marginal genealogies, but the explicit forward-time models that have been studied by others generate simultaneous multiple mergers. (See e.g., Durrett and Schweinsberg 2005 in a model with selective sweeps, but note that our forward-time simulations suggest that this distinction is not important to our main results.) Moreover, the long-range correlations observed in Fig. 5 likely depend on an implicit choice regarding the scaling of recombination, coalescent, and multiple merger rates (Eldon and Wakeley 2006). These issues are poorly understood, and more theoretical work is required to understand the interactions between multiple mergers and recombination.

There are two interesting potential empirical applications of our work: assessing the evidence for variation in multiple mergers coalescence within genomes and between species. For example, one could

compute the fPMI in different regions of a large genome and look for a relationship between the strength of non-Kingman coalescence and genomic properties such as the recombination rate and functional density. Alternatively, one could survey multiple species using a data set such as the diversity data compiled by Corbett-Detig et al. 2015. Either would reveal new information about the suitability of population genetic models, and the forces that determine genetic diversity.

References

Anscombe, F. J. (1973). "Graphs in Statistical Analysis". *The American Statistician* 27.1, pp. 17–21.

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.

Church, Kenneth Ward and Patrick Hanks (1990). "Word Association Norms, Mutual Information, and Lexicography". *Comput. Linguist.* 16.1, pp. 22–29.

Comeron, Josep M., Ramesh Ratnappan, and Samuel Bailin (2012). "The Many Landscapes of Recombination in Drosophila Melanogaster". *PLOS Genetics* 8.10, e1002905.

Coop, Graham and Peter Ralph (2012). "Patterns of Neutral Diversity Under General Models of Selective Sweeps". *Genetics* 192.1, pp. 205–224.

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.

Cover, Thomas M. and Joy A. Thomas (1991). *Elements of Information Theory*. New York, NY, USA: Wiley-Interscience.

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

Desai, Michael M., Aleksandra M. Walczak, and Daniel S. Fisher (2013). "Genetic Diversity and the Structure of Genealogies in Rapidly Adapting Populations". *Genetics* 193.2, pp. 565–585.

Donnelly, Peter and Thomas G. Kurtz (1999). "Particle Representations for Measure-Valued Population Models". *The Annals of Probability* 27.1, pp. 166–205.

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.

Elyashiv, Eyal et al. (2016). "A Genomic Map of the Effects of Linked Selection in Drosophila". *PLoS Genetics* 12.8.

Etheridge, Alison (2011). "Mutation and Random Genetic Drift". In: *Some Mathematical Models from Population Genetics: École d'Été de Probabilités de Saint-Flour XXXIX-2009*. Ed. by Alison Etheridge. Lecture Notes in Mathematics. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 5–32.

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 Com-

mon in Drosophila Melanogaster". *Genetics* 203.2, pp. 863–880.

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.

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.

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, R. R. (1983). "Properties of a Neutral Allele Model with Intragenic Recombination". *Theoretical Population Biology* 23.2, pp. 183–201.

Hudson, Richard R. (2001). "Two-Locus Sampling Distributions and Their Application". *Genetics* 159.4, pp. 1805–1817.

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

Keightley, Peter D. et al. (2014). "Estimation of the Spontaneous Mutation Rate per Nucleotide Site in a Drosophila Melanogaster Full-Sib Family". *Genetics* 196.1, pp. 313–320.

Kelleher, Jerome, Alison M. Etheridge, and Gilean McVean (2016). "Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes". *PLOS Computational Biology* 12.5, e1004842.

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.

Kingman, J. F. C. (1982a). "On the Genealogy of Large Populations". *Journal of Applied Probability* 19.A, pp. 27–43.

— (1982b). "The Coalescent". *Stochastic Processes and their Applications* 13.3, pp. 235–248.

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.

Langley, Charles H. et al. (2011). "Circumventing Heterozygosity: Sequencing the Amplified Genome of a Single Haploid Drosophila Melanogaster Embryo". *Genetics* 188.2, pp. 239–246.

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

Messer, Philipp W. (2013). "SLiM: Simulating Evolution with Selection and Linkage". *Genetics* 194.4, pp. 1037–1039.

Möhle, Martin and Serik Sagitov (2001). "A Classification of Coalescent Processes for Haploid Exchangeable Population Models". *The Annals of Probability* 29.4, pp. 1547–1562.

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

Neher, R. A. and O. Hallatschek (2013). "Genealogies of Rapidly Adapting Populations". *Proceedings of the National Academy of Sciences* 110.2, pp. 437–442.

Nicolaisen, Lauren E. and Michael M. Desai (2012). "Distortions in Genealogies Due to Purifying Selection". *Molecular Biology and Evolution* 29.11, pp. 3589–3600.

Pitman, Jim (1999). "Coalescents With Multiple Collisions". *The Annals of Probability* 27.4, pp. 1870–1902.

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.

Sagitov, Serik (1999). "The General Coalescent with Asynchronous Mergers of Ancestral Lines". *Journal of Applied Probability* 36.4, pp. 1116–1125.

— (2003). "Convergence to the Coalescent with Simultaneous Multiple Mergers". *Journal of Applied Probability* 40.4, pp. 839–854.

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.

Schraiber, Joshua G. and Joshua M. Akey (2015). "Methods and Models for Unravelling Human Evolutionary History". *Nature Reviews Genetics* 16.12, pp. 727–740.

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.

Schweinsberg, Jason (2000). "Coalescents with Simultaneous Multiple Collisions". *Electronic Journal of Probability* 5.

— (2003). "Coalescent Processes Obtained from Supercritical Galton–Watson Processes". *Stochastic Processes and their Applications* 106.1, pp. 107–139.

Seger, Jon et al. (2010). "Gene Genealogies Strongly Distorted by Weakly Interfering Mutations in Constant Environments". *Genetics* 184.2, pp. 529–545.

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

Sheehan, Sara, Kelley Harris, and Yun S. Song (2013). "Estimating Variable Effective Population Sizes from Multiple Genomes: A Sequentially Markov Conditional Sampling Distribution Approach". *Genetics* 194.3, pp. 647–662.

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

Steinrücken, Matthias, Matthias Birkner, and Jochen Blath (2013). "Analysis of DNA Sequence Variation within Marine Species Using Beta-Coalescents". *Theoretical Population Biology*. Coalescent Theory 87, pp. 15–24.

Tajima, F. (1989). "Statistical Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism." *Genetics* 123.3, pp. 585–595.

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.

Vitti, Joseph J., Sharon R. Grossman, and Pardis C. Sabeti (2013). "Detecting Natural Selection in Genomic Data". *Annual Review of Genetics* 47.1, pp. 97–120.

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

Watterson, G. A. (1975). "On the Number of Segregating Sites in Genetical Models without Recombination". *Theoretical Population Biology* 7.2, pp. 256–276.

Xie, Xiaohui (2011). "The Site-Frequency Spectrum of Linked Sites". *Bulletin of Mathematical Biology* 73.3, pp. 459–494.

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

Supplementary Figures and Tables

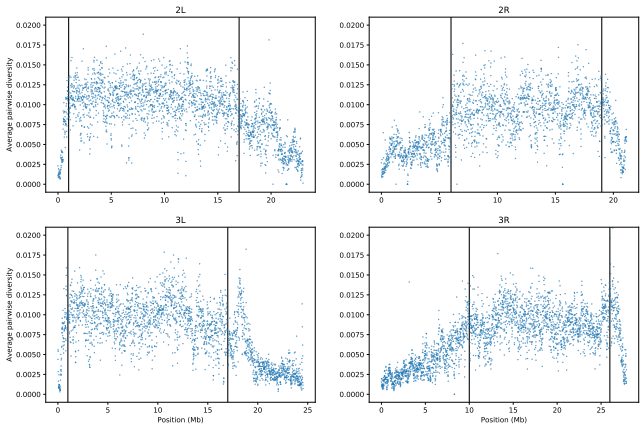

Figure 10: Average pairwise diversity, Π, of 100 haploid samples from DPGP3 panel. Each point represents the average over a 10 KB window. Vertical lines show the boundaries of the central regions defined in Table 1.

Table 1: Boundaries of the homogeneous central regions of chromosomes (shown in 10).

Chr.	**Start position (MB)**	**End position (MB)**
2L	1	17
2R	6	19
3L	1	17
3R	10	26

Table 2: Fraction of sites with at least 90 genotyped samples

Chr.	**All sites**	**Polymorphic sites**
2L	0.913	0.924
2R	0.922	0.925
3L	0.919	0.920
3R	0.933	0.936

Table 3: Piecewise-constant model parameters fit to the SFS of each chromosome arm (see Eq. (10), $N_{\text{anc}} = 3 \times 10^5$).

Chr.	$N_1(\times 10^5)$	$N_2(\times 10^5)$	$t_1(\times 10^4)$	$t_2(\times 10^4)$
2L	10.7	4.6	2.8	39.9
2R	9.1	3.9	1.6	8.8
3L	7.3	3.8	1.3	6.4
3R	9.7	5.1	2.2	8.6