

1 Inferring demographic history using two-locus statistics

2 Aaron P. Ragsdale^{1,*} and Ryan N. Gutenkunst^{2,**}

¹Program in Applied Mathematics, ²Department of Molecular and Cellular Biology,
University of Arizona, Tucson, Arizona 85721

*aragsdale@math.arizona.edu, **rgutenk@email.arizona.edu

3 February 14, 2017

4 Abstract

5 Population demographic history may be learned from contemporary genetic variation data.
6 Methods based on aggregating the statistics of many single loci into an allele frequency spec-
7 trum (AFS) have proven powerful, but such methods ignore potentially informative patterns of
8 linkage disequilibrium (LD) between neighboring loci. To leverage such patterns, we developed
9 a composite-likelihood framework for inferring demographic history from aggregated statistics of
10 pairs of loci. Using this framework, we show that two-locus statistics are indeed more sensitive
11 to demographic history than single-locus statistics such as the AFS. In particular, two-locus
12 statistics escape the notorious confounding of depth and duration of a bottleneck, and they
13 provide a means to estimate effective population size based on the recombination rather than
14 mutation rate. We applied our approach to a Zambian population of *Drosophila melanogaster*.
15 Notably, using both single- and two-locus statistics, we found substantially lower estimates of
16 effective population size than previous works. Together, our results demonstrate the broad
17 potential for two-locus statistics to enable powerful population genetic inference.

18 Introduction

19 Patterns of genetic variation within a population are shaped by the evolutionary and demographic
20 history of that population, so observed variation encodes information about that history. Knowing
21 population demographic history serves as an important control for learning about natural selection
22 (Bustamante et al., 2001; Boyko et al., 2008) and understanding the relative efficacy of selection

23 as populations change in size (Lohmueller et al., 2008; Henn et al., 2016). One particularly in-
24 formative statistic used to summarize genetic polymorphism data is the allele frequency spectrum
25 (AFS), which stores the distribution of observed single-locus allele frequencies from a sample of
26 the population. The shape of the AFS is sensitive to demographic history, and fitting the expected
27 AFS under parameterized demographic models to the observed AFS is a powerful approach for
28 learning about demographic history (Marth et al., 2004; Williamson et al., 2005; Gutenkunst et al.,
29 2009; Kamm et al., 2016b).

30 For unlinked loci, the AFS is a sufficient statistic of the data and completely describes observed
31 patterns of variation (Lohmueller et al., 2009). The expected sample frequency spectrum under
32 arbitrary single- or multi-population histories can be efficiently calculated with either coalescent
33 (Kingman, 1982; Tajima, 1983) or diffusion (Kimura, 1964; Williamson et al., 2005; Gutenkunst
34 et al., 2009) approaches. Poisson random field theory (Sawyer and Hartl, 1992) can then be used
35 to calculate the likelihood of the data given model parameters. A key assumption of the Poisson
36 random field framework is that of independence between segregating loci, so that allele frequency
37 trajectories are uncorrelated. However, neighboring loci are physically linked on the chromosome,
38 and their allele frequencies are thus correlated. Recombination serves to reduce this correlation,
39 with a higher rate of recombination between two loci more rapidly breaking down that associa-
40 tion. For any two linked SNPs, their linkage disequilibrium is a measure of their non-independence.
41 Furthermore, as with allele frequencies, patterns of linkage disequilibrium are shaped by histori-
42 cal demographic events such as bottlenecks, growth, and admixture, and therefore they are also
43 informative about history (Pritchard and Przeworski, 2001).

44 For linked sites the distribution of linkage disequilibrium carries additional information to the
45 allele frequency spectrum about past demography (Myers et al., 2008), and the joint distribution of
46 allele frequencies and linkage disequilibrium between pairs of SNPs should afford greater power for
47 demographic inferences than those based on allele frequencies alone. Characterizing two-locus allele
48 frequency dynamics and calculating their sampling probabilities has attracted a large body of work.
49 Kimura considered the case of genetic drift at multi-allelic loci using a diffusion approximation,
50 and he calculated the time to fixation for one of the alleles when more than two alleles are present
51 (Kimura, 1955). This approach was expanded over the following decade to explicitly consider the
52 two-locus setting with two alleles at each locus (Kimura, 1963; Hill and Robertson, 1966; Karlin

53 and McGregor, 1968; Ohta and Kimura, 1969; Watterson, 1970). These studies were generally
54 interested in the probability and rates of fixation under arbitrary recombination between the two
55 loci and in characterizing the expectation and variance of linkage disequilibrium.

56 More recently, sampling probabilities for two neutral linked loci were directly calculated under
57 equilibrium demography (Golding, 1984; Hudson, 1985; Ethier and Griffiths, 1990), often using
58 the recursion approach due to Golding (1984). Hudson (2001) extended these results to gener-
59 ate those sampling probabilities with knowledge of the ancestral state and proposed a composite
60 likelihood approach for fine-scale estimation of recombination rates across the genome, which has
61 been implemented to infer recombination maps and identify hotspots in human and *Drosophila*
62 populations (McVean et al., 2004; Auton and McVean, 2007; Chan et al., 2012). Xie (2011) used
63 a diffusion approach to calculate the sample frequency spectrum for two completely linked loci
64 under neutrality or equal levels of selection, while Ferretti et al. (2016) recently used a coalescent
65 approach to calculate the expected frequency spectrum for two completely linked neutral loci, and
66 neutral sampling probabilities were developed under the coalescent with recombination for moder-
67 ate to large recombination rates and constant population size (Jenkins and Song, 2009, 2010, 2012;
68 Bhaskar and Song, 2012). Recently, Kamm et al. (2016a) developed a coalescent approach to gener-
69 ate two-locus sampling probabilities under arbitrary demography and recombination and found
70 that accounting for demographic history improves accuracy in composite likelihood approaches for
71 estimating fine-scale recombination rates.

72 Here, we characterize the increase in power of demographic inference from using two-locus allele
73 frequency statistics versus using the single-locus AFS. In particular, the depth and duration of a bot-
74 tleneck are confounded when using the AFS, but we show they can be independently inferred using
75 two-locus statistics. To enable our analyses, we developed a numerical solution to the diffusion ap-
76 proximation for two-locus allele frequencies with arbitrary recombination. We packaged this method
77 in a two-locus composite likelihood framework that can be used to infer single-population demo-
78 graphic histories. Moreover, this framework allows for an estimate of the effective population size
79 based on recombination that is independent from estimates based on levels of diversity. Using this
80 approach, we inferred demographic history for a highly studied Zambian *Drosophila melanogaster*
81 population, finding a smaller effective population size than previous analyses ($N_e \sim 1.5 - 3 \times 10^5$)
82 and a demographic history of recent modest growth and no severe bottlenecks.

83 Theory and Methods

84 A discrete two-locus model with influx of new mutations

85 We used a diffusion approximation to a two-locus model that allows for two alleles at each locus,
86 which are separated by recombination fraction r (Karlin and McGregor, 1968; Watterson, 1970).
87 We allow the left locus to carry alleles A and a , while the right locus permits alleles B and b . Then
88 four haplotypes are possible, AB , Ab , aB , and ab , with frequencies n_{AB} , n_{Ab} , n_{aB} and n_{ab} that sum
89 to $2N$ (Fig. 1A). Frequencies in the subsequent generation are found by considering the random
90 pairing of haplotypes and the probability of a given pairing passing on each type to their offspring.
91 These probabilities depend on current haplotype frequencies and the recombination rate and are
92 described in Table 1 of Watterson (1970). For example, a parent carrying haplotypes AB/Ab will
93 pass on AB with probability $\frac{1}{2}$ and Ab with probability $\frac{1}{2}$, even with recombination. On the other
94 hand, a parent with AB/ab will pass on AB or ab each with probability $\frac{1}{2}(1-r)$ and Ab or aB each
95 with probability $\frac{1}{2}r$. The numbers $(n'_{AB}, n'_{Ab}, n'_{aB}, n'_{ab})$ of each haplotype in the next generation
96 are then pulled from the multinomial distribution for sampling $2N$ haplotypes with probabilities
97 found by considering random pairing of haplotypes and recombination.

98 New two-locus pairings, with two alleles segregating at both sites, arise when a new mutation
99 occurs at one unmutated locus when the other locus is already polymorphic. Suppose, without
100 loss of generality, that the right locus is already polymorphic, with derived allele B at frequency
101 $x_B = n_B/2N$, and ancestral allele b at frequency $x_b = 1 - x_B$. Then a new A mutation at the
102 left locus begins at frequency $x_A = 1/2N$ and occurs on the B haplotype with probability x_B or
103 on the b haplotype with probability x_b . Two-locus frequencies then evolve under the multinomial
104 process described above until one or both loci are fixed for either the ancestral or derived allele, at
105 which point we stop tracking that two-locus pair. The frequencies x_B are drawn from the popula-
106 tion distribution of one-locus frequencies $f(x)$, which can be approximated using diffusion theory
107 (Kimura, 1964). Thus, new independent two-locus pairs enter the population with frequencies
108 $(x_{AB}, x_{Ab}, x_{aB}) = (1/2N, 0, x_B - 1/2N)$ with rate proportional to $x_B f(x_B)$ and $(0, 1/2N, x_B)$ with
109 rate proportional to $(1 - x_B) f(x_B)$.

110 The density $\phi(x_1, x_2, x_3)$ of two-locus haplotype frequencies, where x_1 , x_2 and x_3 are the relative
111 frequencies of haplotypes AB , Ab and aB , respectively (Figure 1B), can be approximated using

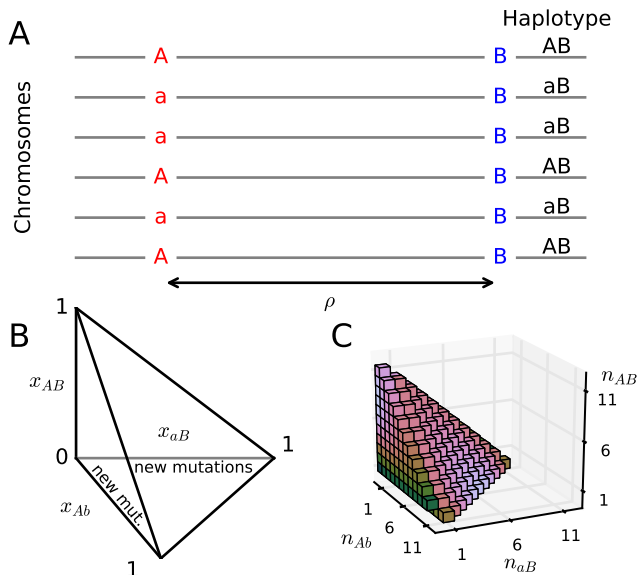


Figure 1: **Two-locus model and frequency spectrum** (A) Two loci with two alleles each are separated by recombination distance $\rho = 4N_e r$. Four haplotypes are possible, and we track the frequencies of the three derived haplotypes. (B) Frequencies change within a tetrahedral domain, with corners of the domain corresponding to one of the four haplotypes fixed in the population. New two-locus pairs occur when a new mutation A occurs against the B/b background, or when B occurs against the A/a background, so we inject density along the Ab or aB axes proportional to the background one-locus allele frequencies. (C) A sample two-locus haplotype frequency spectrum for a sample size of $n = 12$.

112 diffusion theory, as described in the next section. The two-locus haplotype frequency spectrum
 113 stores the counts of derived haplotypes in a sample, where one or both loci carry the derived allele.
 114 To obtain the two-locus spectrum F for n samples from the density function ϕ (Fig. 1C), we sample
 115 against the multinomial sampling distribution:

$$F_{i,j,k} \propto \int \int \int_{\substack{x_i \geq 0 \forall i \\ x_1 + x_2 + x_3 \leq 1}} \phi(x_1, x_2, x_3) \binom{n}{i, j, k} x_1^i x_2^j x_3^k (1 - x_1 - x_2 - x_3)^{n-i-j-k} dx_1 dx_2 dx_3. \quad (1)$$

116 Here, $\binom{n}{i, j, k}$ is the multinomial coefficient, defined as $n! / (i! j! k! (n - i - j - k)!)$. Because we assume
 117 that two-locus pairs are independent realizations of this process, Poisson random field theory tells
 118 us that if we observe data $D(i, j, k)$, each entry in the observed two-locus spectrum is a Poisson
 119 random variable with mean $F(i, j, k)$. This allows the application of likelihood theory to compare

120 observed data to model expectations.

121 The two-locus diffusion approximation

122 We solved the multiallelic diffusion equation for ϕ to obtain the expected sample two-locus spec-
123 trum. Measuring time τ in units of $2N_a$ generations, where N_a is the ancestral reference population
124 size, the forward diffusion equation describes the evolution of the probability density of two-locus
125 frequencies and is written as

$$\begin{aligned} \frac{\partial \phi}{\partial \tau} = & \frac{1}{2} \sum_{1 \leq i \leq 3} \frac{\partial^2}{\partial x_i^2} \left(\frac{x_i(1-x_i)\phi}{\nu(\tau)} \right) - \sum_{1 \leq i < j \leq 3} \frac{\partial^2}{\partial x_i \partial x_j} \left(\frac{x_i x_j \phi}{\nu(\tau)} \right) \\ & + \frac{\rho}{2} \left[\frac{\partial}{\partial x_1} (D\phi) - \frac{\partial}{\partial x_2} (D\phi) - \frac{\partial}{\partial x_3} (D\phi) \right]. \end{aligned} \quad (2)$$

126 Here, $D = x_1(1-x_1-x_2-x_3) - x_2x_3$ is the linkage disequilibrium, given haplotype frequencies
127 (x_1, x_2, x_3) , and $\nu(\tau) = \frac{N(\tau)}{N_A}$ is a function for the relative population size to the ancestral population
128 size at time τ . The population scaled recombination rate between the A/a and B/b loci is $\rho = 4N_A r$,
129 where r is the recombination rate per generation per meiotic event. The action of recombination
130 is readily interpretable in the diffusion equation; recombination acts directionally on the haplotype
131 frequencies x_i , pushing them toward linkage equilibrium ($D = 0$) at a rate directly proportional to
132 the recombination rate ρ .

133 The domain of the two-locus diffusion equation is the tetrahedron with $0 \leq x_i \leq 1$ for $i = 1, 2, 3$,
134 and $\sum_i x_i \leq 1$ (Fig. 1B). If the recombination rate $\rho = 0$ and there is no recurrent mutation, then
135 all boundary surfaces of the domain are absorbing, so if one of the haplotypes is lost from the
136 population it remains lost. However, with $\rho > 0$, the boundary is not necessarily absorbing, as
137 recombination may reintroduce a previously absent haplotype. For example, if only Ab and aB
138 types are found in the population, a recombination event between the two loci may create either
139 an ab or AB type in an individual in the next generation. Some of the edges of the domain are
140 absorbing, since once one of either A/a or B/b fixes at the left or right locus, respectively, that
141 two-locus pair remains fixed in the absence of recurrent mutation.

142 We numerically solved Eq. 2 using finite differencing in a framework similar to Ragsdale et al.
143 (2016). We split the diffusion operator into mixed and non-mixed terms, using an implicit alter-

144 nating direction scheme for the non-mixed spatial derivatives (Chang and Cooper, 1970) and a
145 standard explicit scheme for the mixed spatial derivatives. We used equal numbers of uniformly
146 spaced grid points for each spatial dimension, so that grid points coincided directly on the off-axes
147 surface of the domain. This allowed for density to be accurately integrated along the surface and
148 interior of the domain. As discussed in Ragsdale et al. (2016) and detailed in the Supporting Infor-
149 mation, naively applying finite differencing along the off-axes surface led to numerical error in the
150 solution to ϕ . Thus, we instead accounted for density moving between the interior of the domain
151 and that surface by directly moving density between the two each timestep.

152 Because the diffusion equation is linear, it can be used to solve for the density of all two-locus
153 frequencies in the population by allowing for the influx of new mutations each generation. For the
154 single locus diffusion equation, this amounts to the injection of density at rate $\theta/2$ at frequency
155 $1/(2N)$, with the appropriate limit taken to allow $N \rightarrow \infty$. In the two-locus model, one of the
156 two loci will already be polymorphic (suppose the right B/b locus), and a mutation occurs at the
157 other (left) locus. As described above, the new mutation A at the left locus initially has frequency
158 $1/(2N)$, while the right locus carries derived allele B with frequency $x \in (0, 1)$ depending on the
159 single-locus population allele frequency spectrum $f(x)$, which will itself depend on the population
160 size function $\nu(\tau)$. Allele A falls on the B background with probability x and the b background with
161 probability $1 - x$. Thus, we inject density into the two-locus diffusion equation by simultaneously
162 tracking the single locus allele frequency density function f and setting the influx of density into
163 ϕ proportional to f along the x_2 and x_3 axes (Fig. 1B). To solve for the two-locus spectrum under
164 a nonequilibrium demographic model $\nu(\tau)$, we first solve for ϕ at equilibrium and then integrate
165 forward according to ν . We then sample ϕ against the multinomial sampling distribution with
166 sample size n (Eq. 1) to obtain the two-locus spectrum.

167 Composite likelihood estimation and demographic inference

168 We follow the composite likelihood approach outlined by Hudson (2001), in which we consider
169 pairs of loci and their sampling distribution. Reducing the full likelihood for more than two linked
170 loci to the composite likelihood over all possible pairs of polymorphisms leads to the loss of in-
171 formation. However, computing two-locus sampling statistics retains a considerable amount of
172 information regarding both allele frequencies and patterns of linkage disequilibrium between them.

173 For recombination distances $\rho \in [\rho_{\min}, \rho_{\max}]$, we consider all pairs of loci separated by each ρ ,
 174 and store sampling frequencies in the two-locus frequency spectrum for this range of ρ . In prac-
 175 tice, recombination distances vary continuously over any interval, so we are required to bin our
 176 data within subintervals of ρ by defining intervals $[\rho_0, \rho_1), [\rho_1, \rho_2), \dots, [\rho_{n-1}, \rho_n]$. For fine enough
 177 subintervals, we approximated the expected two-locus spectrum for an interval $[\rho_{i-1}, \rho_i)$ using our
 178 diffusion approach with the mean recombination rate over that interval $\rho = (\rho_{i-1} + \rho_i)/2$.

179 For a given ρ -interval, we made the assumption that all pairs of loci contributing to the two-locus
 180 spectrum are independent, approximating the full likelihood by the composite likelihood across all
 181 pairs of loci. The two-locus frequency spectrum then forms a Poisson random field, so for sample
 182 data D and expected model M calculated under model parameters Θ , the likelihood of the data
 183 $\mathcal{L}(\Theta|D)$ can be calculated by assuming each data entry D_i is a Poisson random variable with mean
 184 M_i . Thus, the likelihood function for a single ρ -bin is

$$\mathcal{L}(\Theta|D) = \prod_i \frac{e^{-M_i} M_i^{D_i}}{D_i!}. \quad (3)$$

185 We allowed the population mutation rate θ to be an implicit parameter for each bin, which scales
 186 the total size of the frequency spectrum while retaining its shape. The maximum likelihood value
 187 for θ is then $\hat{\theta} = \left(\frac{\sum D_i}{\sum \tilde{M}_i}\right)^{1/2}$, where \tilde{M} is the model spectrum with θ set to one. The square arises
 188 because mutations that are paired to existing variant sites arise proportional to rate θ , but those
 189 existing mutations also arise proportional to rate θ , so that the total rate of influx of new two-locus
 190 pairs occur at a rate proportional to θ^2 .

191 We simultaneously considered all bin intervals of $\rho \in [\rho_{\min}, \rho_{\max}]$, and so for bin centers
 192 $(\rho_{1/2}, \rho_{1+1/2}, \dots)$, the likelihood function is

$$\mathcal{L}(\Theta|D_{\rho_j}, j = 1/2, 1 + 1/2, \dots) = \prod_j \prod_i \frac{e^{-M_{j,i}} M_{j,i}^{D_{j,i}}}{D_{j,i}!}, \quad (4)$$

193 where j indexes the ρ -bins, and i indexes the frequency spectrum entries for a given ρ_j . In reality,
 194 pairs of loci are not independent, so we used the Godambe Information Matrix (GIM) to estimate
 195 parameter uncertainties (Coffman et al., 2016), which adjusts the composite likelihood statistics
 196 to account for linkage between data. This required bootstrapping the data, and we did so by
 197 dividing the autosomal genome into 1,000 bins of equal length and resampling these regions with

198 replacement.

199 We fit single-population demographic models to the data, which are defined by the population
200 size history function $\nu(\tau)$ (Eq. 2). We considered simplified demographic models that may be
201 described by a handful of parameters, rather than inferring a parameter free function $\nu(\tau)$ as in Liu
202 and Fu (2015). For example, in an instantaneous expansion model, the parameters are the relative
203 change in size ν and the time T in the past that the population changed size.

204 Phased and unphased data

205 For data with phased chromosomes, determining haplotype frequencies is straightforward counting
206 of haplotypes for a given pair of loci. Using an aligned outgroup, the ancestral state for each SNP
207 may be determined, so that the two-locus spectrum stores derived two-locus allele frequencies. The
208 ancestral state for each locus may be misidentified, potentially due to sequencing error or recurrent
209 mutation along the lineage leading to the outgroup, and this can distort the two-locus spectrum
210 (Hernandez et al., 2007). To account for ancestral misidentification, we included the probability
211 $p_{\text{mis}} \in [0, 1]$ that a given SNP had a misidentified state in our model fitting. Thus, with probability
212 $p_{\text{mis}}(1 - p_{\text{mis}})$ the A allele was misidentified but the B allele was correctly identified, and with
213 the same probability the B allele was misidentified and the A allele was correctly identified. Both
214 alleles A and B were misidentified with probability p_{mis}^2 . In our demographic model fits to data,
215 we fit p_{mis} along with the parameters from the demographic model.

216 When data is unphased, as is the case for many genomic datasets, observed haplotypes can not
217 be tallied. Rather, we are left with counts of genotypes in individuals, $(n_{AABB}, n_{AABb}, n_{AAbb}, n_{AaBB}, \dots)$.
218 The composite linkage disequilibrium statistic \hat{D} is an unbiased estimator for D (Weir, 1979; Zaykin,
219 2004),

$$\hat{D} = \frac{1}{n} \left(2n_{AABB} + n_{AABb} + n_{AaBB} + \frac{1}{2}n_{AaBb} \right) - 2pq, \quad (5)$$

220 where n is the number of sampled individuals. One possible approach to summarize observed data
221 might be to work with the joint statistics $p = n_A$, $q = n_B$, and \hat{D} . Instead, we directly used
222 genotype counts in the “genotype frequency spectrum” G . In genotype data, individuals may carry
223 AA , Aa , or aa at the left locus, and BB , Bb , or bb at the right locus. Thus, there are nine possible
224 two-locus genotypes ($AABB$, $AABb$, $AAbb$, $AaBB$, ...) that could be observed to be carried by
225 an individual, so that G is an eight dimensional object with size $(n + 1)^8$. However, G is sparse

226 and can be stored efficiently. Each genotype can only be formed by the pairing of two specific
227 haplotypes (e.g. $AABb$ can only be from one haplotype of each AB and Ab), except for $AaBb$,
228 which could be formed by $AB+ab$ or $Ab+aB$. Thus, we expected G to still carry information about
229 demography through the joint patterns of allele frequencies and linkage disequilibrium. Expected
230 genotype frequencies can be calculated from expected haplotype frequencies, and we detail our
231 approach in the Supporting Information.

232 *Drosophila* sequence data and recombination map

233 As an application, we considered a single Zambian population of fruit flies, using data from phase 3
234 of the *Drosophila* Population Genomics Project (DPGP3), available from the *Drosophila* Genome
235 Nexus (Lack et al., 2015). The data consisted of 197 sequenced haploid embryos, so genomes were
236 necessarily phased. We used Annovar (Wang et al., 2010) to annotate all biallelic SNPs across
237 the genome, and we used intronic and intergenic regions in our two-locus analysis. We determined
238 the ancestral allele for each SNP using the alignment to *D. simulans* (April 2006, dm3 aligned to
239 droSim1, downloaded from the UCSC genome browser), by assuming the *D. simulans* allele was
240 ancestral. If the *D. melanogaster* site had no alignment, or if the *D. simulans* allele was different
241 than the two *melanogaster* alleles, we discarded that site.

242 For each chromosome, we considered all pairs of biallelic SNPs in intergenic and intronic regions
243 for which an ancestral state could be determined, within recombination distance ρ_{\max} . We deter-
244 mined recombination distances using the recombination map inferred by Comeron et al. (2012),
245 which reports cumulative recombination rates in units of cM over 100,000 bp intervals along each
246 chromosome. We converted to $\rho = 4N_e r$ by taking the map distance d (in cM) separating the two
247 SNPs and multiplying by $4N_e/100$. This required an estimate for N_e , so we used neutral demo-
248 graphic fits to intronic and intergenic single-locus data, which provided an estimate for $\theta = 4N_e\mu L$.
249 Here, μ is the mutation rate, and we used $\mu = 5.5 \times 10^{-9}$ (Schrider et al., 2013). The total length
250 of sequences that were included in our analysis was $L \approx 3.93 \times 10^7$. Then $N_e = \theta/(4\mu L) \approx 3 \times 10^5$.
251 For each two-locus pair, we counted the number of AB , Ab , aB , and ab haplotypes across all 197
252 samples and then subsampled to a sample size of $n = 20$. In the supporting information, we show
253 how to project data to a smaller sample size, but for the sample sizes in our dataset the full projec-
254 tion would have required more memory than we had available. This allowed for more pairs to be

255 included in the data, as any pair of loci without missing haplotype data for at least 20 samples was
256 included, and a smaller sample size allowed for more rapid evaluation of the expected frequency
257 spectrum for optimization.

258 **Independent inference of N_e**

259 Two-locus statistics are binned by the populations size-scaled recombination rate $\rho = 4N_e r$, where
260 r is the recombination rate per meiotic event per generation. Thus, given a recombination map we
261 require an accurate estimate for N_e to appropriately bin the data. In the case that the effective
262 population size is unknown, N_e may be left as a parameter to be fit during optimization of the
263 model to the data. In this approach, we guess an initial effective population size N_0 to first bin
264 the data by $\rho_0 = 4N_0 r$ (for example, 10^4 for human populations, or 10^6 for *Drosophila*) and then
265 allow the ρ -value for each bin to be rescaled by α_N as $\rho = 4N_0 r \alpha_N$. If the best fit $\alpha_N = 1$, then
266 N_0 turned out to be the best fit effective population size, while if α_N is larger or smaller than one,
267 then the best fit N_e is inferred to be larger or smaller than N_0 by that factor. We rescaled the ρ
268 value for each bin of data instead of reassigning data to fixed bins for fair comparison of likelihoods
269 across varying values of α_N , and because reassigning two-locus data each iteration of optimization
270 would be computationally burdensome.

271 **Results and Discussion**

272 **Numerical accuracy of solution to two-locus allele frequency spectrum**

273 We first compared our numerical solution for two-locus statistics for a population in demographic
274 equilibrium to those calculated by Hudson (2001). Our solution matched those using Hudson's
275 algorithm across all values of ρ , from completely linked ($\rho = 0$) to loose linkage ($\rho = 100$) (Fig. 2,
276 top row). To verify our numerical solution for nonequilibrium demography, we compared it to
277 simulations of the discrete two-locus process with an influx of mutations. We simulated a population
278 of $N = 1000$ diploid, randomly mating, individuals for independent pairs of loci separated by a
279 given recombination rate. New two-locus pairs entered the population at a rate proportional to
280 Eqs. S3 and S4. We allowed the simulation to proceed for $20N$ generations and then applied
281 specified population size changes, sampling two-locus haplotype frequencies from the population

282 after each simulation completed. Our nonequilibrium solution matched the simulated two-locus
 283 statistics (Fig. 2, bottom row). See Supporting Information for further details regarding simulation
 284 and numerical accuracy.

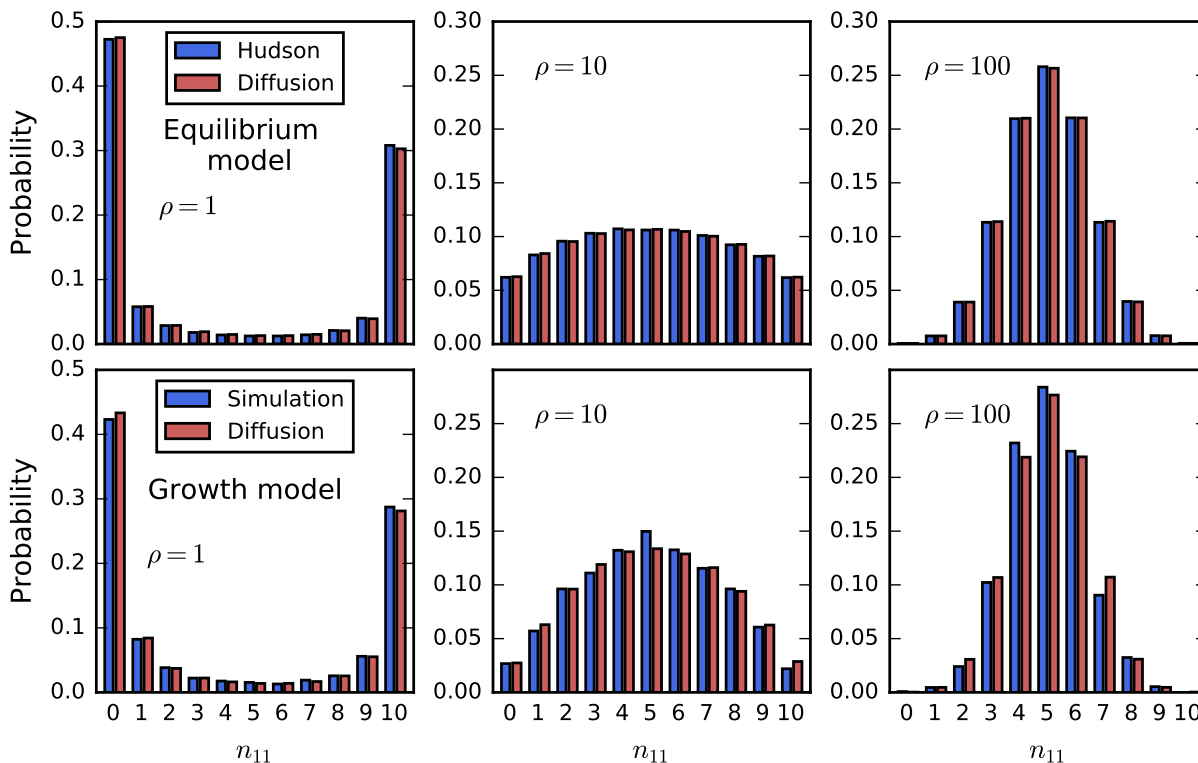


Figure 2: **Verification of numerical solution.** For sample size $n = 30$, the distribution of n_{AB} is shown, when the frequencies of A and B are $p = 10$ and $q = 15$ and ρ is varied. Top row: Comparison to equilibrium statistics from Hudson (2001). Bottom row: Comparison to discrete simulation under growth model.

285 Two-locus statistics are sensitive to demography

286 To assess the increase in statistical power for demographic history inference using the two-locus
 287 spectrum versus the single-locus spectrum, we used the information theoretical measure Kullback-
 288 Leibler (KL) divergence (Kullback and Leibler, 1951). The KL divergence measures the amount
 289 of information lost if an incorrect demographic model M_0 is used to approximate the true model
 290 M_{true} , and it can be interpreted as the expected likelihood ratio statistic for testing M_{true} against

291 M_0 . For discrete distributions, such as frequency spectra, KL divergence is defined as

$$D_{KL}(M_{\text{true}}\|M_0) = \sum_i M_{\text{true}}(i) \log \frac{M_{\text{true}}(i)}{M_0(i)}. \quad (6)$$

292 In our comparisons, we took M_0 to be a model of constant demography and compared the KL diver-
293 gence for two demographic models, an instantaneous growth model and a bottleneck and recovery
294 model, between two-locus and single-locus frequency spectra (Fig. 3). A larger KL divergence in-
295 dicated that more information is contained in the data to reject the constant size model. For the
296 two model types, we considered varying recovery times T since the demographic event, so in the
297 growth model T is the time since the instantaneous expansion ($\nu = 2$), and in the bottleneck model
298 T is the time since recovery from the bottleneck ($\nu_B = 0.1, T_B = 0.05$). In all cases, the two-locus
299 spectrum is more informative about the demography per pair of linked loci than are two unlinked
300 loci in the single-locus frequency spectrum.

301 We considered the KL divergence for varying values of recombination rate ρ from completely
302 linked ($\rho = 0$) to loose linkage ($\rho = 100$). For large ρ , KL divergence from two-locus statis-
303 tics converged to the measure for unlinked single-locus data, which is to be expected as $\rho \rightarrow \infty$
304 implies unlinked loci. Importantly, the most informative recombination distance varied between
305 demographic models and recovery times T since demographic events. As T increases, lower recom-
306 bination rates are relatively more sensitive, because higher recombination rates will restore levels of
307 linkage disequilibrium faster than lower recombination rates. Therefore, loosely linked loci are more
308 informative about recent demographic events, while tightly linked loci are more informative about
309 deeper events. We performed the KL divergence analysis on genotype data as well (Figure 3, red
310 curves), and we found that two-locus statistics at the genotype level are also more sensitive than
311 one-locus statistics. For the growth model, the KL divergence of genotype data was intermediate
312 between the KL divergences of one-locus and haplotype data, but for the bottleneck model, very
313 little sensitivity is lost when using genotype data instead of haplotype data.

314 **Fits to simulated data**

315 To further validate our model and to explore efficient and informative ways to collate two-locus
316 statistics, we simulated single-population demographic history under neutrality with realistic human
317 mutation and recombination rates for many large (1 Mb) regions using `ms` (Hudson, 2002) (details

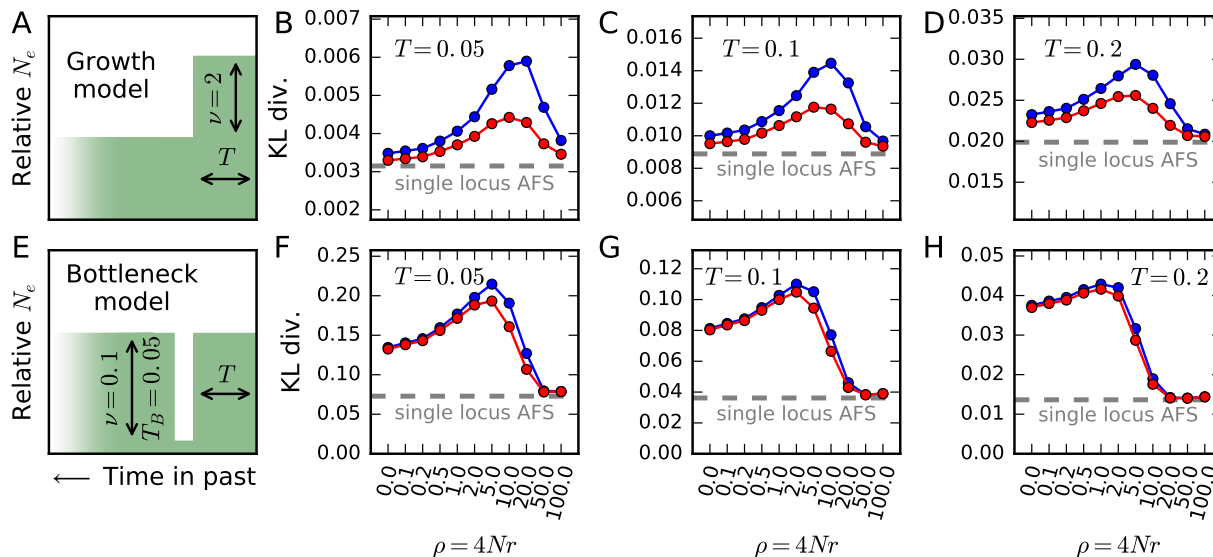


Figure 3: **Sensitivity to demography.** We compared KL divergence measures between two-locus statistics and the single-locus frequency spectrum for a simple growth model (A, top row) and a bottleneck model (E, bottom row). The blue curve shows the KL divergence for phased (haplotype) data, while the red curve is for unphased (genotype) data. In each comparison, we considered the KL divergence between the specified demographic model and a null model of constant population size. (A) In the instantaneous growth model, the population doubled in size some time T in the past, and we considered (B) $T = 0.05$, (C) 0.1, and (D) 0.2. (E) In the bottleneck model, the population shrank to 1/10 its original size for $T_B = 0.05$ genetic time units and then recovered to its original size T genetic units ago for (F) $T = 0.05$, (G) 0.1, and (H) 0.2. In all cases, and across all values of ρ , KL divergence was greater for two-locus statistics than the corresponding single locus statistics of the same number of unlinked sites. The two-locus spectrum is thus more sensitive to demographic history than the single-locus spectrum.

318 in Supporting Information). Using sets of 100 simulated 1 Mb regions, we simulated a simple
 319 growth model (instantaneous expansion by a factor of 2, 0.1 time units before present) and fit
 320 the demography to both simulated single- and two-locus statistics (Supporting Information). We
 321 repeated this simulation and fitting process 50 times and checked how accurately and precisely
 322 we recovered the simulated demographic parameters. We used the same simulations to check the
 323 accuracy of our fits to genotype data, by pairing chromosomes to create diploid individuals. Fig. 4
 324 shows our fits to simulated data, with two-locus genotype statistics more precisely recovering the
 325 true demographic model than single-locus statistics, and haplotype statistics more precisely than
 326 genotype statistics. When we allowed N_e to vary, we also accurately recovered the simulated

327 parameters including α_N (Fig. 4B).

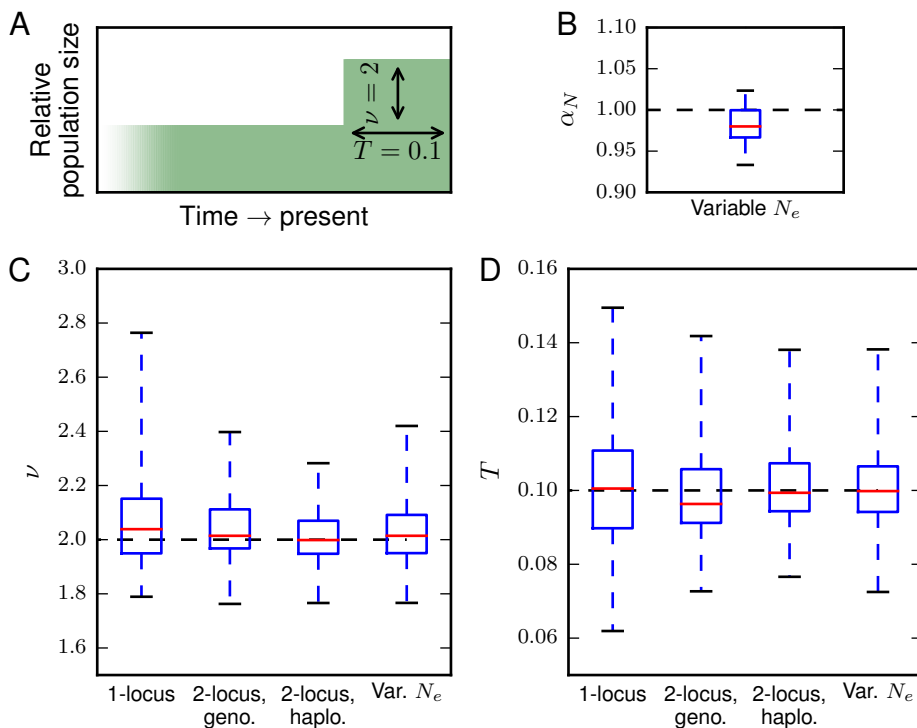


Figure 4: **Fits to data from simulated growth model.** A: We simulated 50 replicate data sets with length 100Mb under an instantaneous growth model using *ms* and checked how accurately we recovered the simulated parameters for both single- and two-locus data, including allowing N_e to vary (B). C-D: For both ν and T , fits to the two-locus frequency spectrum were more accurate than single-locus fits. Here, the median values and top and bottom quartiles are indicated by the boxes, and the whiskers extend to the largest and smallest inferred values from the simulated datasets.

328 In an identical fashion, we also simulated a bottleneck model, in which the population size
 329 shrank by a factor of 0.1 for 0.05 genetic time units and then recovered to its original size for 0.2
 330 time units until sampling at present (Fig. 5). For this demography, the fits to single-locus statistics
 331 were inconsistent, and many replicates did not converge to reasonable parameter values, with ν_B
 332 tending to 0. The two-locus haplotype fits more accurately recovered the modeled parameters,
 333 although the inferred values of ν_B were consistently slightly elevated. The fits to genotype data
 334 were also more accurate than using single-locus data, consistent with our KL divergence results
 335 (Figure 3). Disentangling the depth and duration of a bottleneck from allele frequency data is
 336 notoriously challenging (Keinan et al., 2007; Bunnefeld et al., 2015), and jointly incorporating

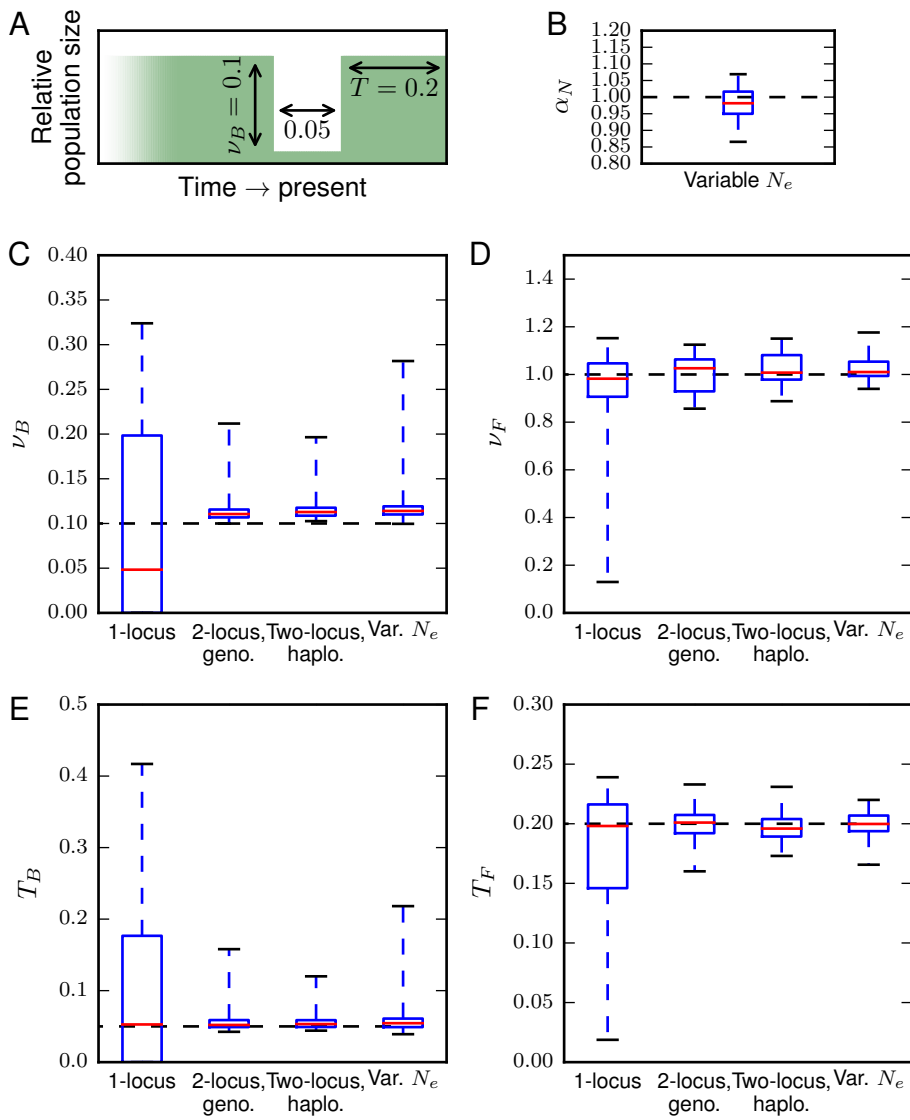


Figure 5: **Fits to data from simulated bottleneck model.** A: We simulated 50 replicate data sets with length 100Mb under a bottleneck and recovery demographic history, in which the population declined to 0.1 its original size for $T = 0.05$ genetic time units and then recovered to its original size for 0.2 time units. C-F: Demographic inferences using single-locus data alone could not consistently recover the true parameters. However, using genotype or haplotype two-locus data allowed for precise inference of model parameters, including when N_e was allowed to vary (B).

337 information about linkage disequilibrium dramatically improves parameter identifiability.

338 Demographic inference of a *Zambian Drosophila* population

339 As an application of our approach, we considered the demographic history of a Zambian population
340 of *Drosophila melanogaster*, which is thought to be a close proxy to the ancestral population (Lack
341 et al., 2015). We first fit two- and three-epoch single-population demographic models to intronic
342 and intergenic single-locus data in order to estimate θ and N_e (Table 1). We inferred the ancestral
343 effective population size to be approximately 3×10^5 , which is somewhat lower than previously
344 suggested sizes for *D. melanogaster* (Keightley et al., 2014; Garud and Petrov, 2016). Using the
345 recombination map of Comeron et al. (2012), we determined distances in ρ between pairs of loci,
346 assuming an effective population size of 3×10^5 , and we binned two-locus data as described above.
347 We then fit the two- and three-epoch models to the two-locus data, with and without varying
348 N_e (Table 1) and calculated parameter uncertainties using the Godambe Information Matrix (Ta-
349 ble S1). For all fits, we subsampled the data to 20 samples for computational speed, and additional
350 speed-up was afforded by calculating each ρ -bin’s expected frequency spectrum in parallel.

Table 1: **Point estimates from fits to *Drosophila* data.** Reported log-likelihoods (LL) are for two-locus data using the demographic history parameters from each fit. 95% confidence intervals are given in Table S1.

Data statistics (Model)	ν_1	ν_2	T_1	T_2	p_{mis}	N_e	LL
One-locus (2-epoch)	4.23		0.329		0.0476	302,900	-1068200
One-locus (3-epoch)	2.35	10.7	0.388	0.0938	0.0496	291,500	-1404200
Two-locus (fix N_e , 2-epoch)	3.83		0.371		0.0449	3×10^5	-1025600
Two-locus (fix N_e , 3-epoch)	34.3	1.69	0.220	0.053	0.0434	3×10^5	-844200
Two-locus (var. N_e , 2-epoch)	4.02		0.379		0.0456	179,900	-851700
Two-locus (var. N_e , 3-epoch)	1.53	4.58	0.352	0.286	0.0473	170,000	-825600

351 For the two-epoch model, parameter values inferred using single- and two-locus data were quite
352 similar (Table 1). For the three-epoch model, however, inferred values were quite different. In
353 particular, the two-locus fit with fixed N_e inferred a large population size increase followed by a
354 sharp decline, but the single-locus fit and the two-locus fit with variable N_e both inferred two-stage
355 increases with qualitatively similar estimates. When we allowed N_e to be simultaneously fit to the
356 data, we found the best-fit value was smaller (1.7×10^5), and the variable N_e three-epoch model
357 best fit the two-locus data. The disagreement of inferred parameters for the fixed- N_e fit is likely
358 due to the model attempting to fit observed LD but being constrained by an N_e larger than the
359 optimal value. This suggests that scaling the recombination map by a fixed estimate for N_e may

360 introduce significant bias into downstream parameter estimates.

361 All of the inferred models fit the single-locus frequency spectrum well (Fig. S2), but they varied
362 in their ability to capture patterns of LD (Fig. 6). The two-locus data fit with a three-epoch model
363 including variable N_e fit the LD decay curve much better than any of the other model fits, although
364 it still underestimated long-range LD. Previous models of *D. melanogaster* demographic history
365 also underestimated long-range LD (Garud and Petrov, 2016). While a more complex demography
366 might be able to better fit the LD curve, factors aside from single-population demography may
367 be critical to generating the pattern of long-range elevated LD, including population substructure,
368 recent admixture, or the effects of linked selection.

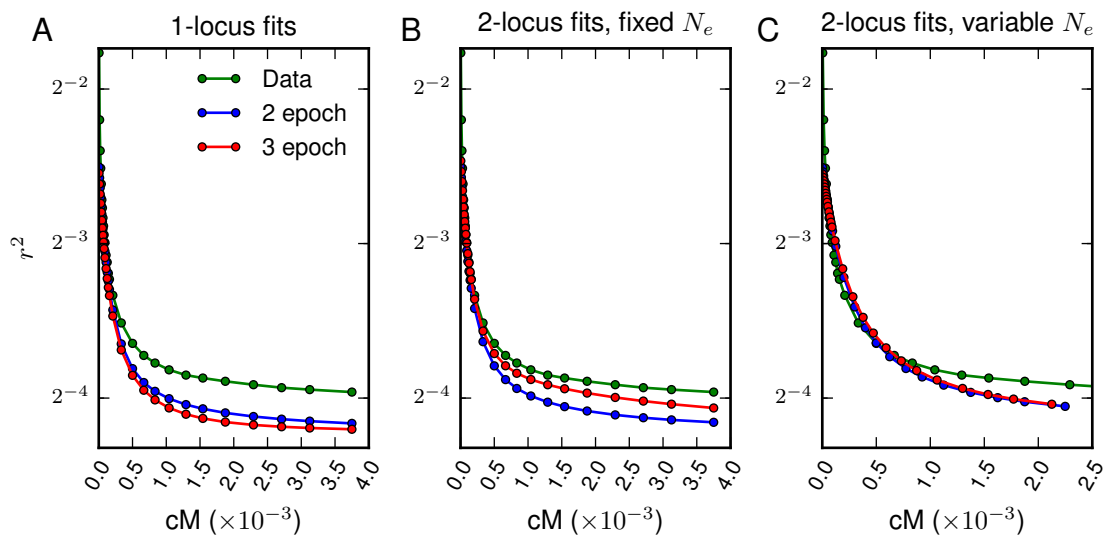


Figure 6: **Fits to LD-decay from *Drosophila* data.** LD-decay curves for two-locus models compared to observed decay curves from the data. (A) The two-locus model using the best fit parameters from single-locus data, (B) the two-locus model fit with N_e set to 3×10^5 , and (C) the two-locus model with N_e allowed to vary. Each of the models underestimates long-range LD decay, as also observed by Garud and Petrov (2016), although the two-locus fits that allow variable N_e attempt to compensate for the poor fit to observed levels of LD (C).

369 Our estimates of the ancestral effective population size of *D. melanogaster* are notably smaller
370 than previous estimates. Keightley et al. (2014) estimated the spontaneous mutation rate by
371 sequencing a family of two parents and 12 full-sibling offspring and used their estimation to infer
372 $N_e \sim 1.4 \times 10^6$. The effective population size may also be estimated from observed levels of

373 diversity, and Charlesworth (2015) estimated $N_e \sim 0.7 \times 10^6$ using observed synonymous site
374 diversity. Furthermore, N_e is often assumed to be $\geq 10^6$ in many population genetic studies of *D.*
375 *melanogaster* (Sella et al., 2009; Garud et al., 2015; Garud and Petrov, 2016). Our estimates for
376 N_e were substantially lower. Using levels of diversity for intronic and intergenic loci, we estimated
377 $N_e \sim 3 \times 10^5$ through our demographic fits to the single-locus AFS (Table 1). In an alternative
378 approach, we allowed N_e to vary in the two-locus inference, and we estimated a smaller value of
379 $N_e \sim 1.7 \times 10^5$. This approach is based on the rescaling of the recombination map without assuming
380 a fixed mutation rate, and it thus provides an independent inference of the effective population size.
381 Together, our results suggest that ancestral N_e for *D. melanogaster* may be substantially lower than
382 previously estimated, and studies that require an assumed effective population size should consider
383 a wider range of possible N_e values. Notably, it has been suggested that linked selection is common
384 throughout the genome of *D. melanogaster* (Garud and Petrov, 2016), and linked selection is known
385 to increase the variance in offspring distribution, which in turn decreases the effective population
386 size (Leffler et al., 2012).

387 Conclusions

388 Based on the continuous approximation to a two-allele two-locus discrete Wright-Fisher model
389 with recombination, we developed a numerical solution to the two-locus diffusion equation that
390 handles arbitrary recombination rates and demographic history. We used this method to develop a
391 composite likelihood framework to infer demographic history from observed two-locus data, which
392 can handle data sampled as either haplotypes or genotypes. While two-locus statistics have been
393 successfully and extensively used to infer fine-scale recombination maps for many organisms, we
394 focused on quantifying the additional power afforded by two-locus over single-locus statistics for
395 demographic history inference. We found that two-locus statistics do provide substantial additional
396 power. For example, while inferring the parameters of a bottleneck model from single-locus data
397 is notoriously difficult (Keinan et al., 2007), we were able to precisely and consistently recover the
398 correct demographic parameters using two-locus statistics. Moreover, for at least some scenarios,
399 little power is lost when data are unphased and genotype frequencies are fit. Finally, we turned
400 to data from a Zambian fruit fly population, and we found that using two-locus statistics to infer
401 demographic history provided a much better fit to both the allele frequency spectrum and observed

402 patterns of LD. The demographic history that we inferred still underestimates the observed long-
403 range levels of LD, which has been previously observed in this population (Garud and Petrov,
404 2016). Moreover, using two independent approaches, one based on levels of diversity and the
405 other based on scaling the recombination map, we inferred the ancestral effective population size
406 to be substantially lower than previous inferences. It is likely that additional factors to single
407 population demography are at play, including potentially complicated demographic features such
408 as substructure and admixture, and the effects of linked selection.

409 Acknowledgments

410 AR thanks Gleb Zhelezov for helpful discussions regarding various numerical approaches for this
411 class of PDEs. The authors also thank Nandita Garud for sharing recombination map data and
412 useful discussion regarding the demographic history of the fly population. This work was supported
413 by the National Science Foundation (DEB-1146074 to RG).

414 Supporting Information

415 Two-locus solution numerics

416 Our numerical solution to two-locus diffusion equation (Eq. 2) uses finite differences, closely fol-
417 lowing the numerical methods described in Ragsdale et al. (2016). We separately apply the mixed
418 and non-mixed spatial derivatives, using an alternating direction implicit (ADI) method for non-
419 mixed terms and a standard explicit term for the mixed terms. The grid spacing is uniform with
420 equal number M of grid points in each direction x_i , so that grid spacing $\Delta = 1/(M - 1)$. For the
421 ADI method, each direction was sequentially integrated forward in time. For the x_1 direction, we
422 discretized Eq. 2 as

$$\begin{aligned} \frac{\phi_{i,j,k}^{n+1} - \phi_{i,j,k}^n}{\Delta\tau} = & \frac{1}{2\nu_\tau} \frac{1}{\Delta} \left(\frac{V_{i+1}\phi_{i+1,j,k}^{n+1} - V_i\phi_{i,j,k}^{n+1}}{\Delta} - \frac{V_i\phi_{i,j,k}^{n+1} - V_{i-1}\phi_{i-1,j,k}^{n+1}}{\Delta} \right) \\ & - \frac{1}{2} \frac{1}{\Delta} \left(M_{i+1/2,j,k} \left(\phi_{i+1,j,k}^{n+1} + \phi_{i,j,k}^{n+1} \right) - M_{i-1/2,j,k} \left(\phi_{i,j,k}^{n+1} + \phi_{i-1,j,k}^{n+1} \right) \right), \end{aligned} \quad (\text{S1})$$

where

$$V_i = x_i(1 - x_i)$$

and

$$M_{i,j,k} = -\frac{\rho}{2} [x_i(1 - x_i - x_j - x_k) - x_j x_k].$$

423 The x_2 and x_3 discretizations were similar, but with the opposite sign for $M_{i,j,k}$. For the mixed
424 derivative terms, we sequentially applied an explicit scheme over the (x_1, x_2) , (x_1, x_3) , and (x_2, x_3)
425 planes. In the (x_1, x_2) direction, we used the discretization

$$\frac{\phi_{i,j,k}^{n+1} - \phi_{i,j,k}^n}{\Delta\tau} = -\frac{(C\phi^n)_{i+1,j+1,k} - (C\phi^n)_{i+1,j-1,k} - (C\phi^n)_{i-1,j+1,k} + (C\phi^n)_{i-1,j-1,k}}{4\Delta^2}. \quad (\text{S2})$$

426 The (x_1, x_3) and (x_2, x_3) planes were analogous.

427 Sequentially applying the ADI and explicit mixed derivative methods along the off-axes surface
428 resulted in significant error, with an excess of density pushed to the surface. Again, similar to
429 Ragsdale et al. (2016) we integrated ϕ forward in time using the methods described above for
430 all grid points not on the off-axes surface. For each grid point near that surface, we calculated
431 the amount of density that should be lost to the surface each time step and directly moved that
432 density to the surface. This density from a grid point at (x_1, x_2, x_3) may be found by numerically
433 integrating the analogous one-dimensional process forward one time unit from a point mass placed
434 at $x = x_1 + x_2 + x_3$ and measuring the amount of density that fixes at $x = 1$. We similarly
435 directly moved density from the surface back into the interior of the domain each time step due
436 to recombination events along that surface. Each time step we also integrated the density on the
437 surface forward in time using Eqs. S1 and S2 for the analogous three state process.

438 To model the influx of new mutations, we coupled our numerical solution to the two-locus dif-
439 fusion equation to single-locus models ϕ^{bi} for the background allele frequencies. These simulations
440 were carried out using $\partial a \partial i$ (Gutenkunst et al., 2009), and densities ϕ^{bi} were added to the two-locus
441 solution ϕ along the x_2 and x_3 axes, corresponding to the new haplotype starting at low frequency
442 after mutation. Specifically, suppose B/b alleles are already segregating at the right locus with the
443 frequency of B as x , and a new A mutation occurs at the left locus. The mutation A lands on the
444 B background with probability x and lands on the b background with probability $1 - x$. We thus

445 added the amount

$$\frac{\theta}{2} \frac{1}{\Delta^3} \Delta\tau \phi_k^{\text{bi}} (1 - x_k) \quad (\text{S3})$$

446 to $\phi_{0,1,k}$, and

$$\frac{\theta}{2} \frac{1}{\Delta^3} \Delta\tau \phi_k^{\text{bi}} x_k \quad (\text{S4})$$

447 to $\phi_{1,0,k}$. The injection for B onto A/a was analogous, adding to $\phi_{0,j,1}$ and $\phi_{1,j,0}$.

448 The diffusion equation is valid in the limit of large population size N_e , so we extrapolated on
449 grid spacing Δ to approximate the solutions for $\Delta \rightarrow 0$. In practice, the number of grid points
450 should exceed the number of samples in the frequency spectrum. With a sample size of 20, we
451 typically used grid spacings with $M = 40, 50$, and 60. We also found that accuracy was improved
452 by extrapolating on $\Delta\tau$ as well, and we used $\Delta\tau = [0.005, 0.0025, 0.001]$ for these grid spacings.

453 **Binning data by ρ**

454 Differences in the two-locus frequency spectra for varying values of ρ are more pronounced at small
455 ρ . (For example, the differences between spectra for $\rho = 1$ and 2 are much more pronounced than
456 the differences between spectra for $\rho = 49$ and 50.) Thus, we used tighter bins for low recombination
457 rates and wider bins for higher recombination rates. We partitioned data into 28 bins, chosen to
458 match the number of cores on a node of our compute cluster, and computation of spectra for each
459 bin was parallelized. The bin edges were $\rho = 0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1, 1.2, 1.4,$
460 $1.6, 1.8, 2, 3, 5, 7, 9, 11, 14, 17, 20, 25, 30, 35, 40$, and 50.

461 **Details of simulation using `ms`**

462 We simulated two demographic models using `ms`: a growth model and bottleneck model, as described
463 in the Results and Discussion. Each simulation consisted of 100 1Mb regions, and we repeated each
464 simulation 50 times, with a sample size of 20 chromosomes. For both demographics, we set the
465 per-base recombination rate to $r = 2.5 \times 10^{-8}$ and the mutation rate to $\mu = 2.5 \times 10^{-8}$. The input
466 command for the growth model was

```
467 ./ms 20 5000 -t 800 -r 400 1000000 -p 6 -eN 0.025 .5,
```

468 and for the bottleneck model was

```
469 ./ms 20 5000 -t 400 -r 400 1000000 -p 6 -eN 0.1 0.1 -eN 0.125 1.0.
```

470 Projection

471 In many genomic data sets, some SNPs might not be called in every individual. Moreover, SNPs
472 will vary in the number of individuals for which data exists. Instead of discarding those SNPs with
473 missing data, by projecting the frequency spectrum down to a smaller sample size n_{proj} , all data
474 called in at least n_{proj} sampled chromosomes may be included (Marth et al., 2004). To project
475 the single-locus frequency spectrum from a sample size of n to a smaller sample size of n_{proj} , one
476 averages over all possible ways of picking subsamples of size n_{proj} from the n observed samples
477 using the hypergeometric function (Marth et al., 2004).

478 For two-locus statistics, we only included data when both the left and right alleles were
479 called in an individual. To project from n observed samples to n_{proj} , with $n_{\text{proj}} < n$, we aver-
480 aged over all possible ways of subsampling the n observed haplotypes. For data with sampled
481 haplotype counts $(n_{AB}, n_{Ab}, n_{aB}, n_{ab}), \sum n_{**} = n$, we counted the number of ways to sample
482 $(\tilde{n}_{AB}, \tilde{n}_{Ab}, \tilde{n}_{aB}, \tilde{n}_{ab}), \sum \tilde{n}_{**} = n_{\text{proj}}$ from that collection of n samples. The probability that we
483 choose $(\tilde{n}_{**}) = (i, j, k, l)$ haplotypes from (n_{**}) can be expressed as

$$P(i, j, k, l) = C_i^{n_{AB}} C_j^{n_{Ab}} C_k^{n_{aB}} C_l^{n_{ab}} / C_{n_{\text{proj}}}^n, \quad (\text{S5})$$

484 where C_i^n indicates the binomial coefficient with parameters n and i .

485 Genotype frequency expectations from haplotype frequencies

For a given entry (i, j, k) in the two-locus spectrum with haplotype frequencies

$$(n_{AB}, n_{Ab}, n_{aB}, n_{ab}) = (i, j, k, n - i - j - k),$$

486 we determined expected genotype frequencies by counting all possible ways that the haplotypes
487 could be paired. To calculate pairing probabilities and visualize the computation, consider pairing
488 a collection of n (even) colored balls that could be any of four colors (red, green, blue, and yellow),
489 where n_R is the number of red balls, n_G the number of green, and so forth. The total number of
490 ways than n objects can be paired is

$$\text{Pairings}(n) = \frac{n!}{(n/2)! 2^{n/2}}. \quad (\text{S6})$$

491 For a given configuration $(n_{**}) = (n_{RR}, n_{GG}, n_{BB}, n_{YY}, n_{RG}, n_{RB}, n_{RY}, n_{GB}, n_{GY}, n_{BY})$, we
 492 must also count the total number of ways that the colored balls may be distributed. Here, n_{RR} is
 493 the number of pure red ball pairings in the set, n_{GY} is the number of pairs of a green and yellow ball
 494 paired together, and so forth. First, for pure-colored (e.g. red) pairings, there are $\binom{n_R}{2n_{RR}}$ ways to
 495 assign red balls between pure and mixed pairings. Of the pure pairings, there are $\text{Pairings}(2n_{RR})$
 496 (Eq. S6) ways to split the pure red balls into pairs. (The other three colors follow the same
 497 calculations.) $n_{RG} + n_{RB} + n_{RY} = n_R - 2n_{RR}$ red balls will be paired with non-red balls. For these
 498 red balls in mixed pairings, there are $\binom{n_{RG} + n_{RB} + n_{RY}}{n_{RG}, n_{RB}, n_{RY}}$ ways to split them into the given number of
 499 RG , RB , and RY pairs, where $\binom{n}{i, j, k}$ is the trinomial coefficient, with $i + j + k = n$, defined as
 500 $\frac{n!}{i!j!k!}$. Finally, for red balls that will be paired with green balls, there are $n_{RG}!$ permutations of
 501 these possible pairings. Again, the other colors follow the same calculation.

502 Now, the probability that haplotypes with frequencies (n_R, n_G, n_B, n_Y) will be paired as (n_{**})
 503 is the number of ways that unique pairings lead to that configuration of genotypes, divided by the
 504 total number of possible pairings:

$$\begin{aligned}
 P((n_{**})|(n_R, n_G, n_B, n_Y)) &= \frac{1}{\text{Pairings}(n)} \binom{n_R}{2n_{RR}} \text{Pairings}(2n_{RR}) \binom{n_G}{2n_{GG}} \text{Pairings}(2n_{GG}) \\
 &\quad \binom{n_B}{2n_{BB}} \text{Pairings}(2n_{BB}) \binom{n_Y}{2n_{YY}} \text{Pairings}(2n_{YY}) \\
 &\quad \binom{n_{RG} + n_{RB} + n_{RY}}{n_{RG}, n_{RB}, n_{RY}} \binom{n_{RG} + n_{GB} + n_{GY}}{n_{RG}, n_{GB}, n_{GY}} \binom{n_{RB} + n_{GB} + n_{GY}}{n_{RB}, n_{GB}, n_{GY}} \\
 &\quad \binom{n_{RY} + n_{GY} + n_{BY}}{n_{RY}, n_{GY}, n_{BY}} n_{RG}! n_{RB}! n_{RY}! n_{GB}! n_{GY}! n_{BY}!. \tag{S7}
 \end{aligned}$$

505

Table S1: **95% confidence intervals from fits to *Drosophila* data.** We used the Godambe Information Matrix (Coffman et al., 2016) to estimate uncertainties for our best fit parameter values.

Data (Model)	ν_1	ν_2	T_1	T_2	p_{mis}	N_e
1-loc (2-ep)	4.16 – 4.30		0.321 – 0.337		0.0468 – 0.0486	295,600 – 310,700
1-loc (3-ep)	2.26 – 2.44	8.2 – 13.2	0.374 – 0.402	0.084 – 0.107	0.0488 – 0.0504	284,500 – 299,000
2-loc (fix N_e , 2-ep)	3.69 – 3.96		0.358 – 0.383		0.0437 – 0.0460	
2-loc (fix N_e , 3-ep)	9.03 – 59.6	1.64 – 1.75	0.209 – 0.231	0.0524 – 0.0536	0.0422 – 0.0446	
2-loc (var N_e , 2-ep)	3.94 – 4.10		0.370 – 0.388		0.0450 – 0.0462	179,500 – 180,500
2-loc (var N_e , 3-ep)	1.30 – 1.76	4.37 – 4.79	0.347 – 0.357	0.242 – 0.330	0.0460 – 0.0486	169,000 – 171,000

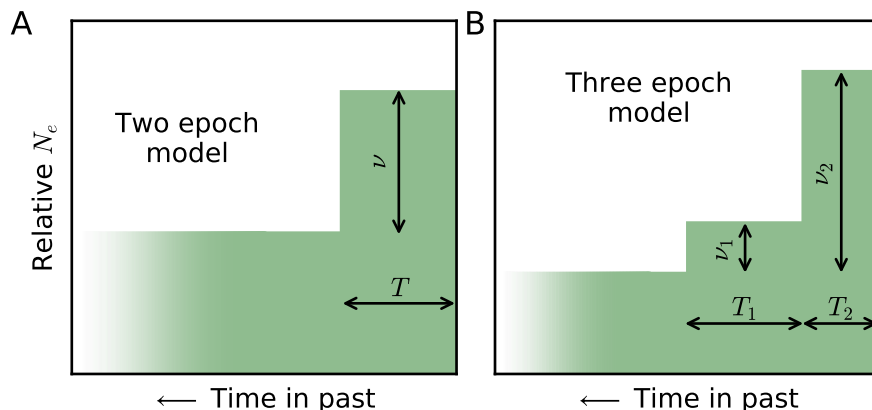


Figure S1: **Demographic models fit to data.** The two single-population models we simulated data under and then fit to the observed *D. melanogaster* data. (A) The two epoch model has a relative size change ν some time T in the past, while (B) the three epoch model includes two periods of recent size change with sizes ν_1 and ν_2 relative to the ancestral population size and lasting for times T_1 and T_2 , resp.

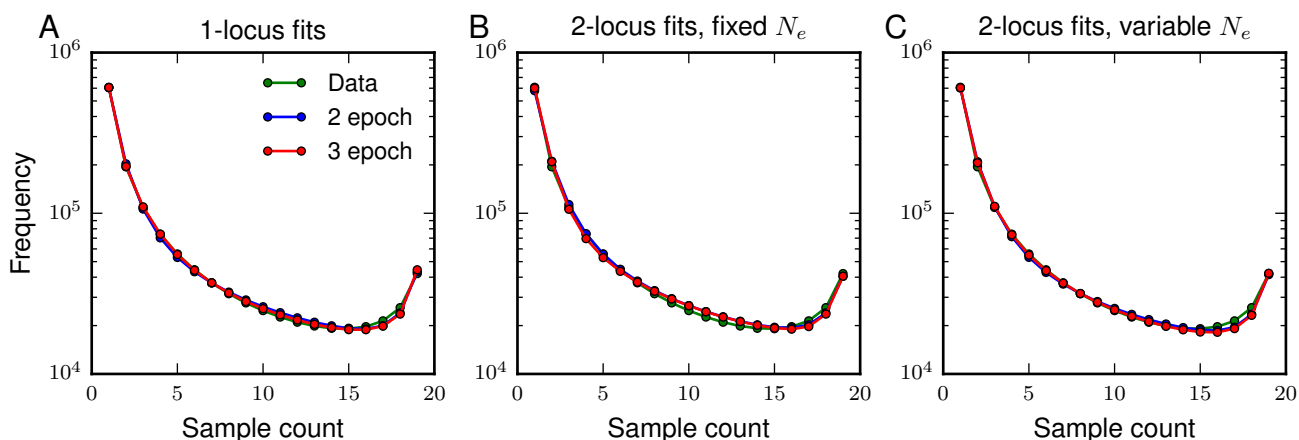


Figure S2: **Fits to single-locus AFS.** All inferred models fit the single-locus data well. (A) We fit two- and three-epoch models to the single-locus AFS, including a parameter to account for ancestral misidentification that causes the over-representation of high frequency alleles. (B) We fit those same models to two-locus data and fixed $N_e = 3 \times 10^5$, which was inferred from our fits to the single-locus data. (C) N_e was allowed to vary, rescaling the effective recombination rates.

506 References

- 507 Auton, A. and McVean, G. (2007). Recombination rate estimation in the presence of hotspots.
 508 *Genome Research*, 17(8):1219–1227.

- 509 Bhaskar, A. and Song, Y. S. (2012). Closed-form asymptotic sampling distributions under the
510 coalescent with recombination for an arbitrary number of loci. *Advances in Applied Probability*,
511 44(2):391–407.
- 512 Boyko, A. R., Williamson, S. H., Indap, A. R., Degenhardt, J. D., Hernandez, R. D., Lohmueller,
513 K. E., Adams, M. D., Schmidt, S., Sninsky, J. J., Sunyaev, S. R., White, T. J., Nielsen, R.,
514 Clark, A. G., and Bustamante, C. D. (2008). Assessing the evolutionary impact of amino acid
515 mutations in the human genome. *PLoS Genetics*, 4(5):e1000083.
- 516 Bunnefeld, L., Frantz, L. A. F., and Lohse, K. (2015). Inferring bottlenecks from genome-wide
517 samples of short sequence blocks. *Genetics*, 201(3):1157–1169.
- 518 Bustamante, C. D., Wakeley, J., Sawyer, S., and Hartl, D. L. (2001). Directional selection and the
519 site-frequency spectrum. *Genetics*, 159(4):1779–1788.
- 520 Chan, A. H., Jenkins, P. A., and Song, Y. S. (2012). Genome-Wide Fine-Scale Recombination Rate
521 Variation in *Drosophila melanogaster*. *PLoS Genetics*, 8(12):e1003090.
- 522 Chang, J. S. and Cooper, G. (1970). A Practical Difference Scheme for Fokker-Planck Equations.
523 *Journal of Computational Physics*, 6(1):1–16.
- 524 Charlesworth, B. (2015). Causes of natural variation in fitness: Evidence from studies of *Drosophila*
525 populations. *Proceedings of the National Academy of Sciences*, 112(6):1662–1669.
- 526 Coffman, A. J., Hsieh, P. H., Gravel, S., and Gutenkunst, R. N. (2016). Computationally Efficient
527 Composite Likelihood Statistics for Demographic Inference. *Molecular Biology and Evolution*,
528 33(2):591–593.
- 529 Comeron, J. M., Ratnappan, R., and Bailin, S. (2012). The Many Landscapes of Recombination
530 in *Drosophila melanogaster*. *PLoS Genetics*, 8(10):e1002905.
- 531 Ethier, S. N. and Griffiths, R. C. (1990). On the two-locus sampling distribution. *Journal of*
532 *Mathematical Biology*, 29(2):131–159.
- 533 Ferretti, L., Klassmann, A., Raineri, E., Wiehe, T., Ramos-Onsins, S. E., and Achaz, G. (2016).
534 The expected neutral frequency spectrum of linked sites. *arXiv*, 1604.06713.

- 535 Garud, N. R., Messer, P. W., Buzbas, E. O., and Petrov, D. A. (2015). Recent Selective Sweeps
536 in North American *Drosophila melanogaster* Show Signatures of Soft Sweeps. *PLoS Genetics*,
537 11(2):1–32.
- 538 Garud, N. R. and Petrov, D. A. (2016). Elevated linkage disequilibrium and signatures of soft
539 sweeps are common in *Drosophila melanogaster*. *Genetics*, 203(2):863–880.
- 540 Golding, G. B. (1984). The sampling distribution of linkage disequilibrium. *Genetics*, 108(1):257–
541 274.
- 542 Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H., and Bustamante, C. D. (2009). Inferring
543 the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency
544 Data. *PLoS Genetics*, 5(10):e1000695.
- 545 Henn, B. M., Botigué, L. R., Peischl, S., Dupanloup, I., Lipatov, M., Maples, B. K., Martin, A. R.,
546 Musharoff, S., Cann, H., Snyder, M. P., Excoffier, L., Kidd, J. M., and Bustamante, C. D.
547 (2016). Distance from sub-Saharan Africa predicts mutational load in diverse human genomes.
548 *Proceedings of the National Academy of Sciences*, 113(4):E440–E449.
- 549 Hernandez, R. D., Hubisz, M. J., Wheeler, D. A., Smith, D. G., Ferguson, B., Rogers, J., Nazareth,
550 L., Indap, A., Bourquin, T., McPherson, J., Muzny, D., Gibbs, R., Nielsen, R., and Bustamante,
551 C. D. (2007). Demographic Histories and Patterns of Linkage Disequilibrium in Chinese and
552 Indian Rhesus Macaques. *Science*, 316(5822):240–243.
- 553 Hill, W. G. and Robertson, A. (1966). The effect of linkage on limits to artificial selection. *Genetical*
554 *Research*, 8(3):269–94.
- 555 Hudson, R. R. (1985). The sampling distribution of linkage disequilibrium under an infinite allele
556 model without selection. *Genetics*, 109(3):611–631.
- 557 Hudson, R. R. (2001). Two-locus sampling distributions and their application. *Genetics*,
558 159(4):1805–1817.
- 559 Hudson, R. R. (2002). Generating samples under a Wright-Fisher neutral model of genetic variation.
560 *Bioinformatics*, 18(2):337–338.

- 561 Jenkins, P. A. and Song, Y. S. (2009). Closed-form two-locus sampling distributions: Accuracy
562 and universality. *Genetics*, 183(3):1087–1103.
- 563 Jenkins, P. A. and Song, Y. S. (2010). An asymptotic sampling formula for the coalescent with
564 recombination. *Annals of Applied Probability*, 20(3):1005–1028.
- 565 Jenkins, P. A. and Song, Y. S. (2012). Padé approximants and exact two-locus sampling distribu-
566 tions. *Annals of Applied Probability*, 22(2):576–607.
- 567 Kamm, J. A., Spence, J. P., Chan, J., and Song, Y. S. (2016a). Two-Locus Likelihoods Under
568 Variable Population Size and Fine-Scale Recombination Rate Estimation. *Genetics*, 203(3):1381–
569 1399.
- 570 Kamm, J. A., Terhorst, J., and Song, Y. S. (2016b). Efficient computation of the joint sample
571 frequency spectra for multiple populations. *Journal of Computational and Graphical Statistics*,
572 pages 1–37.
- 573 Karlin, S. and McGregor, J. (1968). Rates and Probabilities of Fixation for Two Locus Random
574 Mating Finite Populations without Selection. *Genetics*, 58(1):141–159.
- 575 Keightley, P. D., Ness, R. W., Halligan, D. L., and Haddrill, P. R. (2014). Estimation of the spon-
576 taneous mutation rate per nucleotide site in a *Drosophila melanogaster* full-sib family. *Genetics*,
577 196(1):313–320.
- 578 Keinan, A., Mullikin, J. C., Patterson, N., and Reich, D. (2007). Measurement of the human allele
579 frequency spectrum demonstrates greater genetic drift in East Asians than in Europeans. *Nature*
580 *Genetics*, 39(10):1251–1255.
- 581 Kimura, M. (1955). Random Genetic Drift in Multi-Allelic Locus. *Evolution*, 9(4):419–435.
- 582 Kimura, M. (1963). A probability method for treating inbreeding systems, especially with linked
583 genes. *Biometrics*, 19(1):1–17.
- 584 Kimura, M. (1964). Diffusion Models in Population Genetics. *Journal of Applied Probability*,
585 1(2):177–232.
- 586 Kingman, J. (1982). The coalescent. *Stochastic Processes and their Applications*, 13(3):235–248.

- 587 Kullback, S. and Leibler, R. A. (1951). On Information and Sufficiency. *The Annals of Mathematical*
588 *Statistics*, 22(1):79–86.
- 589 Lack, J. B., Cardeno, C. M., Crepeau, M. W., Taylor, W., Corbett-Detig, R. B., Stevens, K. a.,
590 Langley, C. H., and Pool, J. E. (2015). The Drosophila Genome Nexus: A Population Genomic
591 Resource of 623 *Drosophila melanogaster* Genomes, Including 197 from a Single Ancestral Range
592 Population. *Genetics*, 199(4):1229–1241.
- 593 Leffler, E. M., Bullaughey, K., Matute, D. R., Meyer, W. K., Ségurel, L., Venkat, A., Andolfatto,
594 P., and Przeworski, M. (2012). Revisiting an Old Riddle: What Determines Genetic Diversity
595 Levels within Species? *PLoS Biology*, 10(9):e1001388.
- 596 Liu, X. and Fu, Y.-X. (2015). Exploring population size changes using SNP frequency spectra.
597 *Nature Genetics*, 47(5):555–559.
- 598 Lohmueller, K. E., Bustamante, C. D., and Clark, A. G. (2009). Methods for Human Demographic
599 Inference Using Haplotype Patterns From Genomewide Single-Nucleotide Polymorphism Data.
600 *Genetics*, 182(1):217–231.
- 601 Lohmueller, K. E., Indap, A. R., Schmidt, S., Boyko, A. R., Hernandez, R. D., Hubisz, M. J.,
602 Sninsky, J. J., White, T. J., Sunyaev, S. R., Nielsen, R., Clark, A. G., and Bustamante, C. D.
603 (2008). Proportionally more deleterious genetic variation in European than in African popula-
604 tions. *Nature*, 451(7181):994–997.
- 605 Marth, G. T., Czabarka, E., Murvai, J., and Sherry, S. T. (2004). The Allele Frequency Spectrum
606 in Genome-Wide Human Variation Data Reveals Signals of Differential Demographic History in
607 Three Large World Populations. *Genetics*, 166(1):351–372.
- 608 McVean, G., Myers, S., and Hunt, S. (2004). The Fine-Scale Structure of Recombination Rate
609 Variation in the Human Genome. *Science*, 304(5670):581–584.
- 610 Myers, S., Fefferman, C., and Patterson, N. (2008). Can one learn history from the allelic spectrum?
611 *Theoretical Population Biology*, 73(3):342–348.
- 612 Ohta, T. and Kimura, M. (1969). Linkage disequilibrium due to random genetic drift. *Genetical*
613 *Research*, 13(01):47–55.

- 614 Pritchard, J. K. and Przeworski, M. (2001). Linkage Disequilibrium in Humans: Models and Data.
615 *The American Journal of Human Genetics*, 69(1):1–14.
- 616 Ragsdale, A. P., Coffman, A. J., Hsieh, P., Struck, T. J., and Gutenkunst, R. N. (2016). Trial-
617 lelic Population Genomics for Inferring Correlated Fitness Effects of Same Site Nonsynonymous
618 Mutations. *Genetics*, 203(1):513–523.
- 619 Sawyer, S. A. and Hartl, D. L. (1992). Population genetics of polymorphism and divergence.
620 *Genetics*, 132(4):1161–1176.
- 621 Schrider, D. R., Houle, D., Lynch, M., and Hahn, M. W. (2013). Rates and genomic consequences
622 of spontaneous mutational events in *Drosophila melanogaster*. *Genetics*, 194(4):937–954.
- 623 Sella, G., Petrov, D. A., Przeworski, M., and Andolfatto, P. (2009). Pervasive Natural Selection in
624 the *Drosophila* Genome? *PLoS Genetics*, 5(6):e1000495.
- 625 Tajima, F. (1983). Evolutionary relationship of DNA sequences in finite populations. *Genetics*,
626 105(2):437–460.
- 627 Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants
628 from high-throughput sequencing data. *Nucleic Acids Research*, 38(16):e164–e164.
- 629 Watterson, G. (1970). The Effect of Linkage in a Finite Population. *Theoretical Population Biology*,
630 1:72–87.
- 631 Weir, B. S. (1979). Inferences about Linkage Disequilibrium. *Biometrics*, 35(1):235–254.
- 632 Williamson, S. H., Hernandez, R., Fledel-alon, A., Zhu, L., Nielsen, R., and Bustamante, C. D.
633 (2005). Simultaneous inference of selection and population growth from patterns of variation in
634 the human genome. *PNAS*, 102(22):7882–7887.
- 635 Xie, X. (2011). The Site-Frequency Spectrum of Linked Sites. *Bulletin of Mathematical Biology*,
636 73(3):459–494.
- 637 Zaykin, D. V. (2004). Bounds and normalization of the composite linkage disequilibrium coefficient.
638 *Genetic Epidemiology*, 27(3):252–257.