Supporting Text, Figures, and Tables
for Inferring demographic history using two-locus statistics
Aaron P. Ragsdale and Ryan N. Gutenkunst

Two-locus solution numerics

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

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

where

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

and

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

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

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

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

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

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

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

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

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

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

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

Binning data by ρ

Differences in the two-locus frequency spectra for varying values of ρ are more pronounced at small ρ. (For example, the differences between spectra for $\rho = 1$ and 2 are much more pronounced than the differences between spectra for $\rho = 49$ and 50.) Thus, we used tighter bins for low recombination rates and wider bins for higher recombination rates. We partitioned data into 28 bins, chosen to match the number of cores on a node of our compute cluster, and computation of spectra for each 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, 1.6, 1.8, 2, 3, 5, 7, 9, 11, 14, 17, 20, 25, 30, 35, 40, and 50.

Details of simulation using `ms`

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

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

and for the bottleneck model was

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

Projection

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

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

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

where C_i^n indicates the binomial coefficient with parameters n and i. The computational cost of this projection is $O(n^3 n_{proj}^3)$, because every cell of the large spectrum (of size n_{proj}) must be projected downward and, in principle, contributes to every cell of the small spectrum (of size n).

Genotype frequency expectations from haplotype frequencies

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

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

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

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

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 must also count the total number of ways that the colored balls may be distributed. Here, n_{RR} is the number of pure red ball pairings in the set,

n_{GY} is the number of pairs of a green and yellow ball paired together, and so forth. First, for pure-colored (e.g. red) pairings, there are $\binom{n_R}{2n_{RR}}$ ways to assign red balls between pure and mixed pairings. Of the pure pairings, there are Pairings($2n_{RR}$) (Eq. S6) ways to split the pure red balls into pairs. (The other three colors follow the same calculations.) $n_{RG} + n_{RB} + n_{RY} = n_R - 2n_{RR}$ red balls will be paired with non-red balls. For these red balls in mixed pairings, there are $\binom{n_{RG}+n_{RB}+n_{RY}}{n_{RG},n_{RB}}$ ways to split them into the given number of RG, RB, and RY pairs, where $\binom{n}{i,j}$ is the trinomial coefficient, defined as $\frac{n!}{i!\,j!\,k!}$ with $k = n - i - j$. Finally, for red balls that will be paired with green balls, there are $n_{RG}!$ permutations of these possible pairings. Again, the other colors follow the same calculation.

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

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

References

Coffman, A. J., Hsieh, P. H., Gravel, S., and Gutenkunst, R. N. (2016). Computationally efficient composite likelihood statistics for demographic inference. *Molecular Biology and Evolution*, 33(2):591–593.

Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H., and Bustamante, C. D. (2009). Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genetics*, 5(10):e1000695.

Marth, G. T., Czabarka, E., Murvai, J., and Sherry, S. T. (2004). The allele frequency spectrum in genome-side human variation data reveals signals of differential demographic history in three large world populations. *Genetics*, 166(1):351–372.

Ragsdale, A. P., Coffman, A. J., Hsieh, P., Struck, T. J., and Gutenkunst, R. N. (2016). Triallelic population genomics for inferring correlated fitness effects of same site nonsynonymous mutations. *Genetics*, 203(1):513–523.

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. The last two rows show results from an earlier analysis in which we misinterpreted the units of the recombination map, so that we effectively fixed N_e at half the intended value. Confidence intervals for the two-locus data three-epoch model fit are large, because the likelihood surface was locally rough, which interferes with the numerical derivative calculations used in calculating the Godambe Information Matrix. We believe this roughness arose from our procedure of initializing diffusion equation integrations from cached equilibrium ϕ densities, because the model output changes slightly but discontinuously when changes in N_e result in the use of different cached initial ϕ densities.

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)	$3.99 - 4.10$		$0.362 - 0.380$		$0.0448 - 0.0461$	3×10^5 (fixed)
2-loc (fix N_e, 3-ep)	$1.66 - 1.86$	$4.61 - 4.72$	$0.297 - 0.307$	$0.2421 - 0.2523$	$0.0462 - 0.0478$	3×10^5 (fixed)
2-loc (var N_e, 2-ep)	$3.93 - 4.13$		$0.353 - 0.385$		$0.0447 - 0.0461$	$289,900 - 311,000$
2-loc (var N_e, 3-ep)	$1.25 - 1.61$	$4.33 - 4.79$	$0.228 - 0.522$	$0.291 - 0.299$	$0.0463 - 0.0479$	$168,700 - 406,100$
2-loc (fix N_e, 2-ep)	$3.69 - 3.96$		$0.358 - 0.383$		$0.0437 - 0.0460$	1.5×10^5 (fixed)
2-loc (fix N_e, 3-ep)	$9.03 - 59.6$	$1.64 - 1.75$	$0.209 - 0.231$	$0.0524 - 0.0536$	$0.0422 - 0.0446$	1.5×10^5 (fixed)

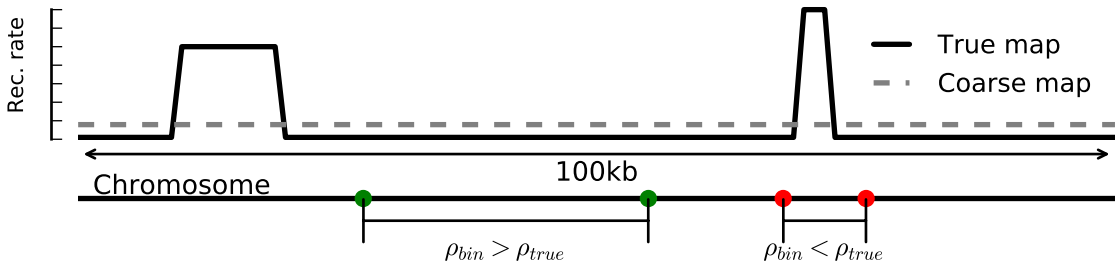

Figure S1: **Effects of recombination map resolution.** A coarsened recombination map may not capture fine-scale recombination hotspots, so the recombination rate between two loci may be underestimated (red loci) or overestimated (green loci).

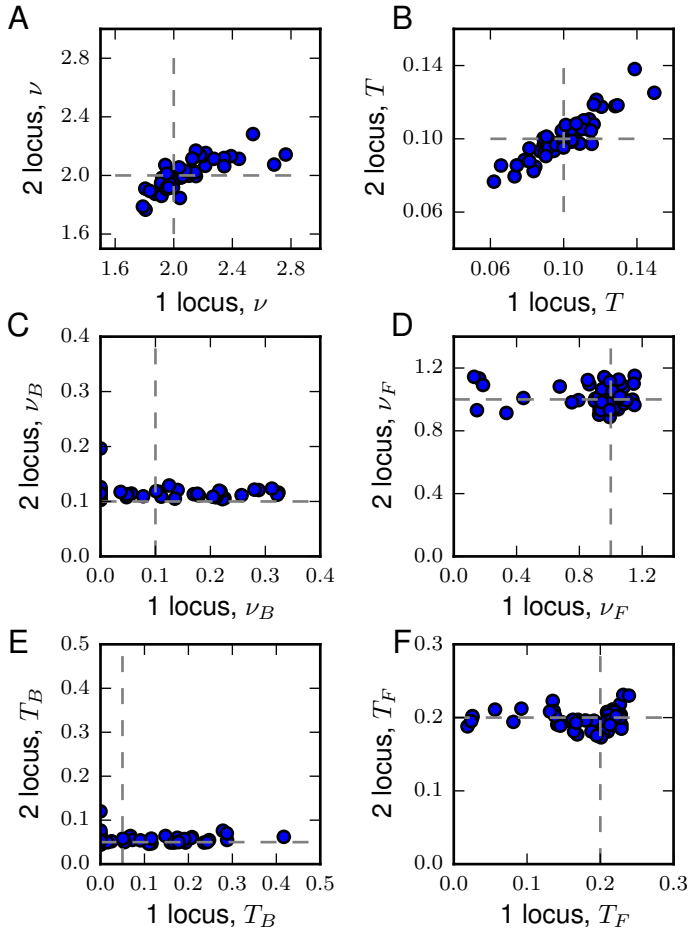

Figure S2: **Correlation of parameters fit to simulated data.** Plotted are the inferred parameters from the same simulated data sets with single- and two-locus statistics. Gray dashed lines indicate the true simulated values. (A-B) For the simple growth model, inferred values of ν and T from the two approaches were correlated, although the two-locus approach was more precise. (C-F) For the bottleneck model, inferred parameter values were largely uncorrelated. Here, we results from the haplotype, fixed-N_e two-locus fits.

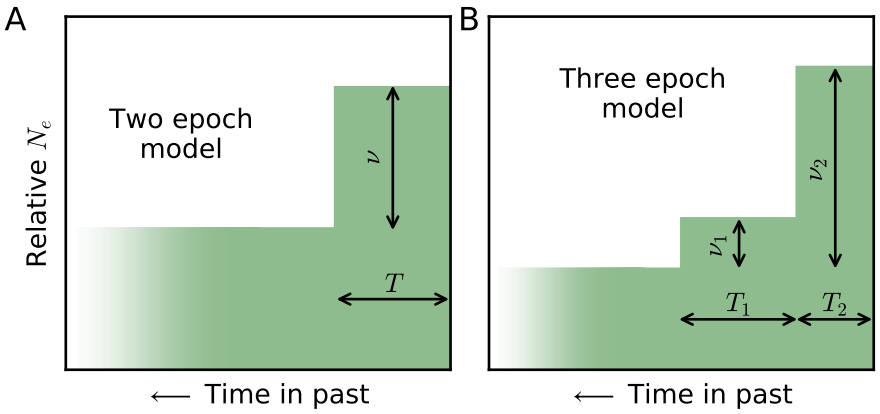

Figure S3: **Demographic models fit to data.** The two single-population models we simulated data under and then fit to the observed *D. melanogaster* data. (A) The two epoch model has a relative size change ν some time T in the past, while (B) the three epoch model includes two periods of recent size change with sizes ν_1 and ν_2 relative to the ancestral population size and lasting for times T_1 and T_2, resp.

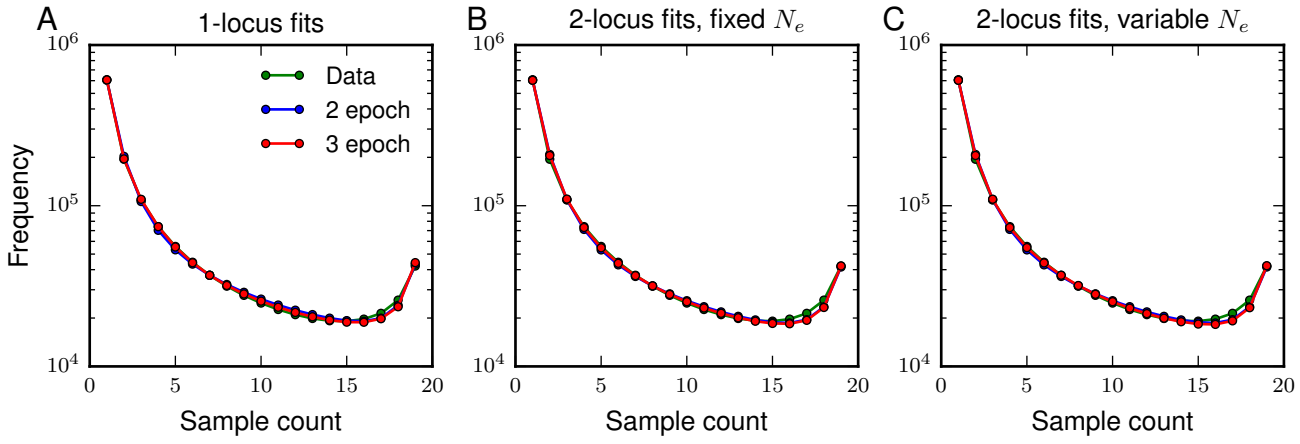

Figure S4: **Fits to single-locus AFS.** All inferred models fit the single-locus data well. (A) We fit two- and three-epoch models to the single-locus AFS, including a parameter to account for ancestral misidentification that causes the over-representation of high frequency alleles. (B) We fit those 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.