Methods to fit Allele Frequency Fluctuations

In the main text, we calculate the chance of pairwise identity by state F, which we defined as the probability of coalescence before a long distance or mutation event. These chances of recent co-ancestry can be only indirectly observed as covariances of allele frequencies, and it is not immediately clear how to best fit these results to observed data. The fitting is also further complicated by the fact that mean allele frequencies are usually unknown. Fitting all of them would introduce many new parameters beyond the small number of demographic parameters and likely lead to to over-fitting of the data, as naive allele frequency estimates will be biased towards the most common direction of the allele frequency fluctuations. There are different possible approaches to deal with these problems. We decided to implement and test three different ways to fit allele frequency fluctuations based on our model. Their full implementations are available on the github repository: `https://github.com/hringbauer/Harald.git`

Method 1: Gaussian Random Field Method

In Computer Science and Machine Learning, the so called Gaussian Random Field method is widely used to fit covariance structures (Rasmussen and Williams 2006). The goal is to fit the covariances themselves, and then use this fit to make new predictions based on the fitted covariance structure. Here, we adapt this method to fit allele frequency covariances. A similar approach has been recently used by Bradburd *et al.* (2013).

Summarizing briefly, in the Gaussian Random Field method the observed data y_i are modeled to depend on known parameters $\vec{x_i}$ and to covary depending on these known parameters. The covariances affect latent, unobserved variables f_i. These unobserved variables are drawn from a multivariate normal distribution with some mean m and covariance matrix K : $f \sim N(m, K)$. The entries K_{ij} of the covariance matrix depend on x_i and x_j, and a set of so called hyper parameters θ:

$$K_{ij} = K(x_i, x_j, \theta)$$

The Gaussian Random Field approach utilizes the fact that a multivariate Gaussian distribution is fully determined by its mean and its covariance matrix. It therefore possible to write down a full likelihood of the observed data given the covariance matrix, by integrating over all latent variables f:

$$L(y, \theta) = \int_f P(y|f)P(f|K(x, \theta)) \tag{1}$$

If one assumes that the data y are drawn as a Gaussian around the latent variable, this integral can be solved analytically due to convenient properties of Gaussian probability densities. One can then easily calculate the marginal likelihood of the observed data, and fit the hyper-parameters via maximizing this likelihood.

If the observations are restricted to binary discrete values (w.l.o.g. 0 or 1), it is still possible to apply the Gaussian Random field model. One typically

transforms the latent variables f using a so called link function $p(f)$ to take values p_i between 0 and 1 (most commonly the logit or the probit function), and then models the discrete observed values y_i to be drawn binomially with mean p_i. However, integral 1 cannot be solved analytically anymore. As it is very high dimensional, direct numerical integration is also computationally infeasible. Therefore, several analytical approximations to 1 are widely used (Nickisch and Rasmussen 2008). For genotype data, we decided to utilize a custom implementation of the Laplace method (Rasmussen and Williams 2006). This widely used approach is based on a second order Taylor approximation around the most likely latent variables f_i. These and the Hessian are found numerically, and these calculations can be done relatively fast. Using this approximation allows for a analytical approximation of the total likelihood. A full description of the method can be found in Rasmussen and Williams (2006).

Genetic data with geographic information consists of discrete genotypes y_i sampled at positions $\vec{x_i}$. W.l.o.g. biallelic markers have values 0 or 1. Diploids can be split up into two haploid genotype data points. The Gaussian Random Field method can then be adapted to fit covariance structure within such data, but one has to deal with some peculiarities. Importantly, the magnitude of the allele frequency covariances depends on the mean allele frequency $\bar{p}$:

$$\mathrm{Cov}(y_i, y_j)) = \bar{p}\,(1-\bar{p})\,F(x_i, x_j), \tag{2}$$

while our model predicts the $F(x_i, x_j)$. In order to account for the additional terms, we introduce a custom link function. We utilize the inverse Fisher's angular transformation of allele frequency (Fisher *et al.* 1947):

$$p(f) := \sin^2\left(\frac{f}{2}\right)$$

This is a valid link function, as its image is confined to the interval $[0, 1]$. Its usefulness stems from the fact that it solves the following differential equation:

$$p'(f) = \sqrt{p(f)\,(1-p(f))}.$$

Allele frequency fluctuations are usually small, and a first order approximation $p(f) \approx p(f_0) + p'(f_0)\Delta f$ yields:

$$\mathrm{Cov}\left((p(f_1), p(f_2)\right) \approx p'(f_0)^2 \mathrm{Cov}(f_1, f_2) = p_0(1-p_0))\mathrm{Cov}(f_1, f_2)$$

Comparing with Eq. 2 shows that this link function together with the F-Matrix as Covariance kernel model the covariance structure of discrete genotypes. As we can directly calculate $F(x_i, x_j)$ with our model, this approach can be used to fit the demographic parameters θ to the data.

To deal with the problem of over-fitting by estimating a potentially large number of mean allele frequencies, we adapted the Gaussian Random Field approach. Unknown allele frequencies are not estimated directly, but only the variance of the unknown distribution of mean allele frequency: We model that mean latent variables are randomly drawn from a distribution with Variance σ^2 around some overall mean and that then the latent variables f_i are drawn with

covariance matrix K around this means. If the means are drawn from a normal distribution, the overall covariance will also be distributed as a multivariate normal distribution:

$$f \sim N(0, K + J\sigma^2)$$

where J denotes the unit matrix, whose entries are all 1. Using this approach, we can fit the effects of unknown distribution of mean allele frequencies as a single hyper-parameter of the covariance matrix. For multiple, independent (unlinked) genotypes, the marginal likelihoods can be multiplied.

Summarizing, the Gaussian Random Field approach allows us to calculate an approximate marginal likelihood of genotype data given the expected co-ancestry structure F based on some demographic hyper parameters θ. Using standard methods to maximize likelihoods, we can find maximum likelihood estimates of these θ. After experimenting with several methods, we found that the standard Nelder-Mead method works very reliably, and used it in all our implementations.

Our approach has two sources of error. First, it is not immediately clear how accurate the Laplace approximation is for genotype data, in particular since allele frequency correlations are typically weak. Second, allele frequency data will not be always distributed as a multivariate Gaussian. Under the model of diffusion of ancestry, there will also be higher order moments. For instance, having recent co-ancestry with individuals in one geographic direction makes it less likely to have shared co-ancestry with individuals from the opposite direction, and this effect is not captured well by the Gaussian Random Field model. Calculating these higher order moments would go far beyond the pairwise diffusion model that we outline in this work. However, the multivariate Gaussian approximation can be expected to be an accurate approximation as long as fluctuations remain small (Barton *et al.* 2013).

Method 2: Maximizing Pairwise Likelihoods

This method maximizes the likelihood of observing the three states of pairwise genotypes. Given two markers, there are four possible states: 00, 10, 01 and 11. Using our calculations for the co-ancestry matrix F, it is straightforward to write down the probability for each of these for states. Denoting the mean allele frequency of marker 1 by p and marker 0 by q:

$$\begin{aligned}
P(00) &= F \cdot p + (1 - F) \cdot p^2 \\
P(10) &= P(01) = (1 - F) \cdot p \cdot q \\
P(11) &= F \cdot q + (1 - F) \cdot q^2.
\end{aligned}$$

As the mean allele frequency is usually unknown, we integrate over their unknown distribution:

$$\begin{aligned}
P(00) &= F \cdot \bar{p} + (1 - F) \cdot \left(\mathrm{Var}(p) + \bar{p}^2\right) \\
P(10) &= P(01) = (1 - F) \cdot \left(\bar{p} - \mathrm{Var}(p) - \bar{p}^2\right) \\
P(11) &= F \cdot \bar{q} + (1 - F) \cdot \left(\mathrm{Var}(q) + \bar{q}^2\right)
\end{aligned}$$

This approach introduces one additional parameter: $v := \mathrm{Var}(p) = \mathrm{Var}(q)$. By multiplying all pairwise likelihoods one gets at a composite likelihood that depends on the demographic parameters θ and the variance parameter v. These pairwise likelihoods are not independent - realized co-ancestry with one individual also increases the probability of co-ancestry with other individuals near the related one. Therefore, multiplying pairwise likelihoods does not yield the total likelihood of the observed data. However, this composite likelihood should be seen as a way to fit the data, and this approach will give consistent parameter estimates in the limit of large amounts of sufficiently uncorrelated data. We implemented the maximization of this likelihood by using the GenericLikelihoodModel class of the Python package statsmodels.

Method 3: Pairwise Homozygosity

One can also fit identity-by-descent probabilities F based on the signal of pairwise homozygosity. As stated in the main text, the chance of pairwise homozygosity for a single marker is given by:

$$h = F + (1 - F) \cdot (\bar{p}^2 + (1 - \bar{p}^2)$$

Summing over all markers gives the expected fraction of pairwise homozygotes:

$$\mathrm{E}(h) = F + (1 - F) \cdot \underbrace{\sum_{p_i} \Pr(p_i) \cdot (p_i^2 + (1 - p_i)^2)}_{:=s}$$

In order to fit this signal, we minimize the sum of squared difference between the expected and observed pairwise homozygosity for all pairs:

$$\bar{\theta} = \min_{\theta} \sum_{k<l} (\bar{h}_{kl}(\theta, s) - h_{kl})^2$$

In our implementation we use the method curvefit from the Python package Scipy.

Performance on simulated Data

To test which method performs best in scenarios with realistic parameters, we tested them on simulated data sets. We used the simulation scheme outlined in the main text to generate data with known demographic parameters, and applied the three methods described above. We first simulated and fitted scenarios without a barrier, in order to test the general capability of the methods to accurately fit allele frequency fluctuations. The outcome is visualized in Figure Fig. S1. Our results show that the Gaussian Random Field method (Method 1) has a significant downward bias when estimating the neighborhood size, whereas the pairwise likelihood and pairwise homozygosity method are approximately unbiased. Our results also indicate that these two inference methods produce highly correlated estimates and have similar estimation variances. We also found that using the pairwise homozygosity method is a factor of 10 quicker than using the pairwise likelihood method.

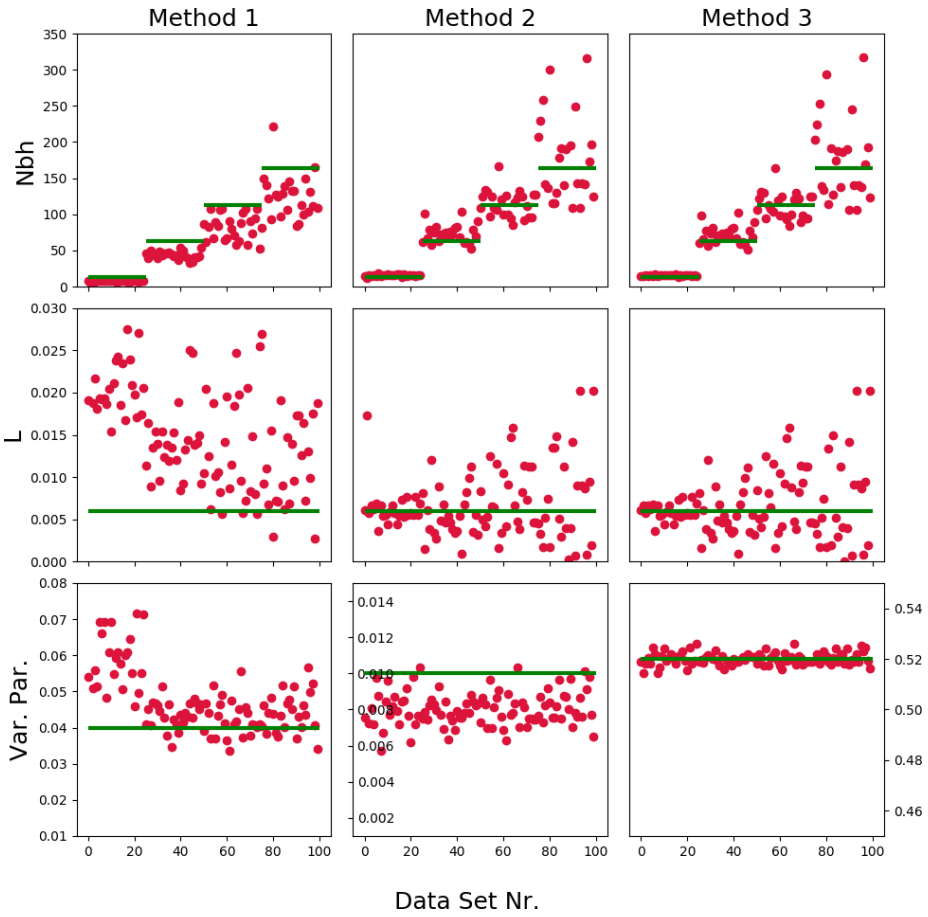

Figure S1: Fit to allele frequency fluctuations: We tested the three methods on synthetic datasets generated using the method described in the main text. The datasets consist of 1000 haploid individuals, situated on a grid of 50×20 individuals spaced 2 dispersal units apart along each axis, with genotype information for 200 loci. We simulated 25 replicate data sets for four different neighborhood sizes (and consequently different magnitudes of allele frequency fluctuations). Throughout, $m = 0.006$ and a random distribution of mean allele frequencies with $\sigma(p) = 0.1$. The lower row depicts the estimates for the Variance Parameter that fits the fluctuations of mean allele frequencies, whose true value is different for all three used models.

Limited Number of Loci and Individuals

Different methods are expected to perform differently when information is limited. Therefore, we tested the three methods on datasets with a varying amount of data. We simulated two types of data sets: One with a varying number of loci, and one with a varying number of individuals. Our results are visualized in Fig. S2 and Fig. S3.

Interestingly, the Gaussian Random Field method remains biased when the number of loci increases; however this bias vanishes with increasing number of individuals. The estimator variance of the other two methods decreases slowly with increasing information. However, neither increasing the number of loci nor increasing the number of individuals seem to yield dramatic increases in estimation accuracy.

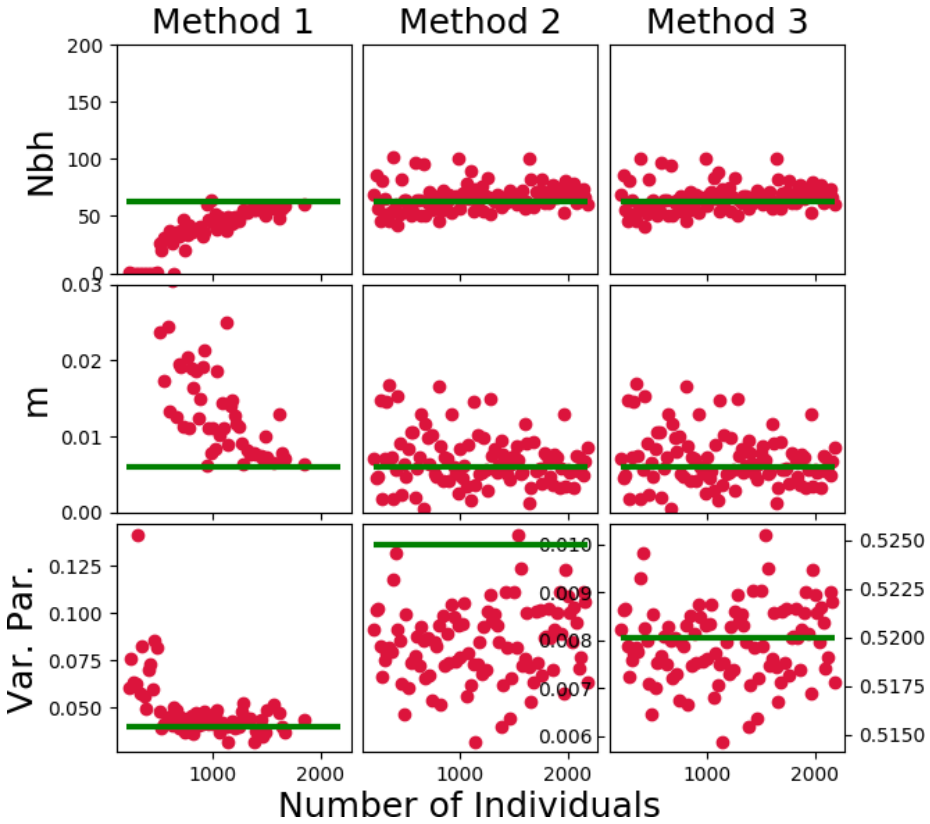

Figure S2: Testing the methods on datasets with varying numbers of individuals. We randomly sub-sampled the target number (200−2200 individuals) from a grid of 100×40 haploid individuals spaced 1 dispersal units apart along each axis. We simulated independent data sets with genotype information for 200 loci ($m = 0.006$, $Nbh = 4\pi 5 \approx 62.83$ and $\sigma(p) = 0.1$).

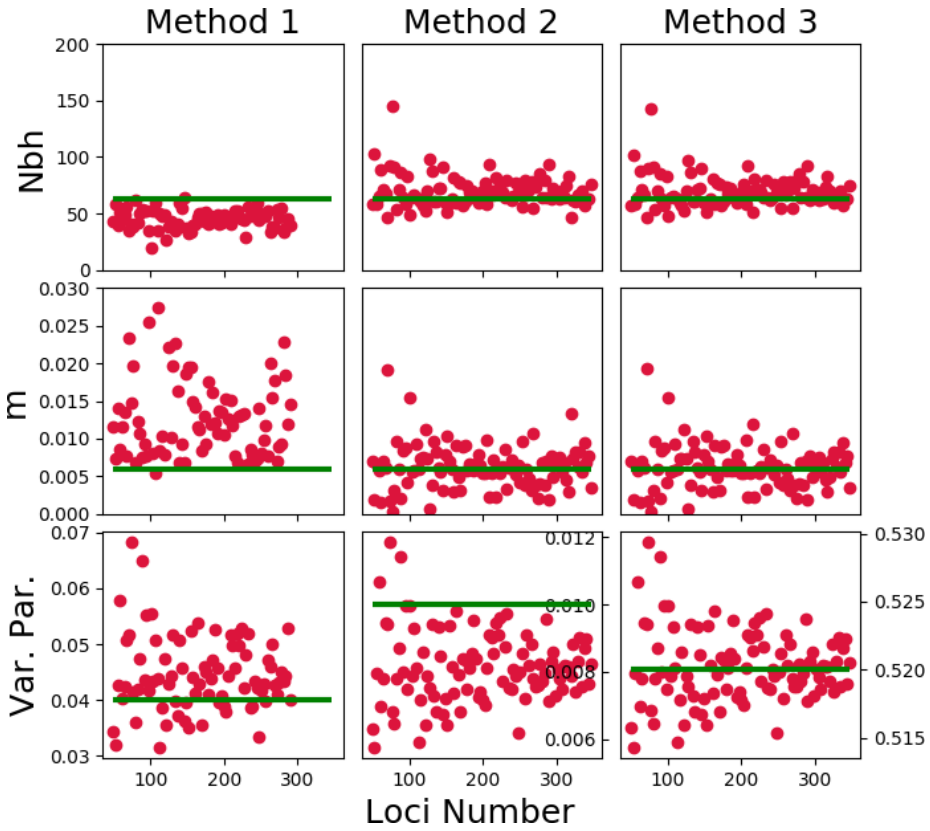

Figure S3: Testing the methods on datasets with varying number of loci. We simulated independent data sets with varying number of independent loci (50-350). Haploid individuals were spaced on a grid of 50×20 with a spacing of 1 dispersal unit along each axis. ($m = 0.006$, Nbh $= 4\pi 5 \approx 62.83$ and $\sigma(p) = 0.1$)

Fitting Barriers with limited Data

One can fit the barrier strength parameter γ while keeping the other parameters fixed. This can be for instance useful if one wants to test the hypothesis of a barrier at a specific subset of loci. One can then estimate the demographic parameters using all loci, and proceed to fit γ based only on the subset of markers.

We therefore tested this approach. We find that fitting the barrier strength alone does not markedly improve inference for estimating γ, at least in the tested scenario, in which there is sufficient information to accurately fit the isolation by distance pattern (Fig. S4).

Our results also indicate that even with 2400 individuals and a strong barrier ($\gamma = 0.05$), one would need at least a few dozen independent biallelic markers to reliably estimate a strong barrier. The required number of markers and individuals for a given scenario will of course depend on the exact sampling scheme as well as the strength and shape of isolation by distance in the data.

Binning Indivduals into Demes

Method 3 can be used to analyze deme data, as outlined in the main text. Binning into demes of k individuals each speeds up calculations by a factor of k^2, as all pairwise comparisons for individuals between two demes reduce to a single comparison. On the other hand, binning nearby individuals is not expected to have a big effect on the inference scheme, as only information for pairs within demes is lost. To confirm this intuition we tested our method on simulated data (Fig. S5). Our results indicate that small scale binning (with bins are extended up to a few dispersal units) does not have a major effect on the parameter estimates. The variance of the inferred parameters increases as expected, but this increase is slow.

Conclusion

The application to simulate data indicates that the methods based on fitting pairwise statistics (Method 2 and Method 3) are more accurate and less biased than the computationally more elaborate Gaussian Random Field approach (Method 1). As outlined above, the latter suffers from two potential errors: The Laplace approximation and also the multivariate Gaussian approximation could be inaccurate for the spatial covariance patterns typically observed in genotype data. Our datasets were simulated under an explicit population genetics model (see main text) with parameters chosen to match typical isolation by distance patterns, and our findings imply that the Gaussian Random Field approach with a Laplace approximation is not suited well to fit those, unless the number of sampled individuals gets very high (Fig. S2).

Method 3, which is based on fitting pairwise homozygosity, is additionally faster by a factor of about 10 than Method 2. Our results here indicate that it can be used on binned data without much loss of accuracy. Therefore, we decided to base inference in the main text on this method.

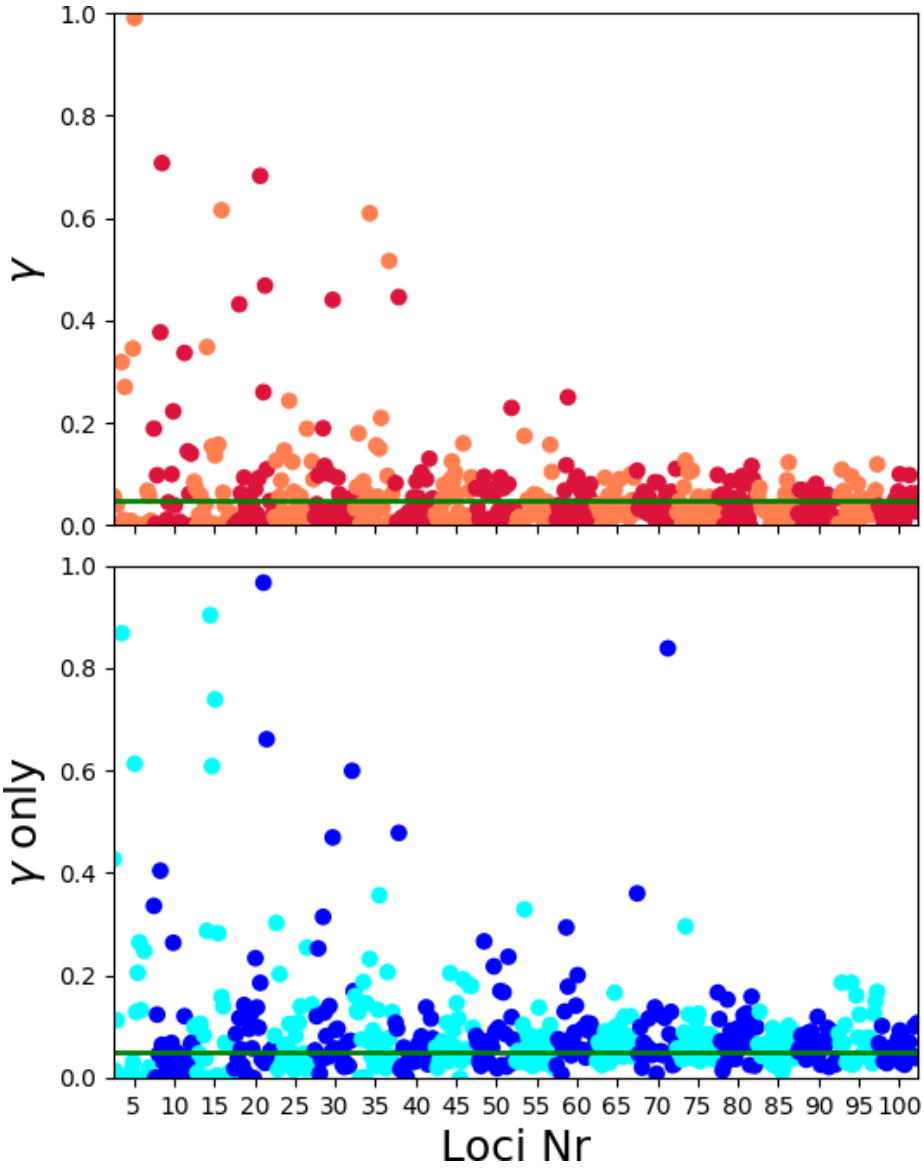

Figure S4: Fitting only the barrier parameter: We simulated a dataset of 60×40 individuals spaced on a grid with step size 1σ ($m = 0.006$, Nbh $= 4\pi 5 \approx 62.83$ and $\sigma(p) = 0.1$ and a strong barrier at the middle of the x-axis with $\gamma = 0.05$). We simulated 25 replicates for different loci numbers $(5, 10, \cdots 100)$. We applied Method 3 to fit the barrier strength γ by estimating all parameters (upper figure) and to fit only the barrier strength and the fluctuation parameter, with the demographic parameters fixed to their true value (lower figure).

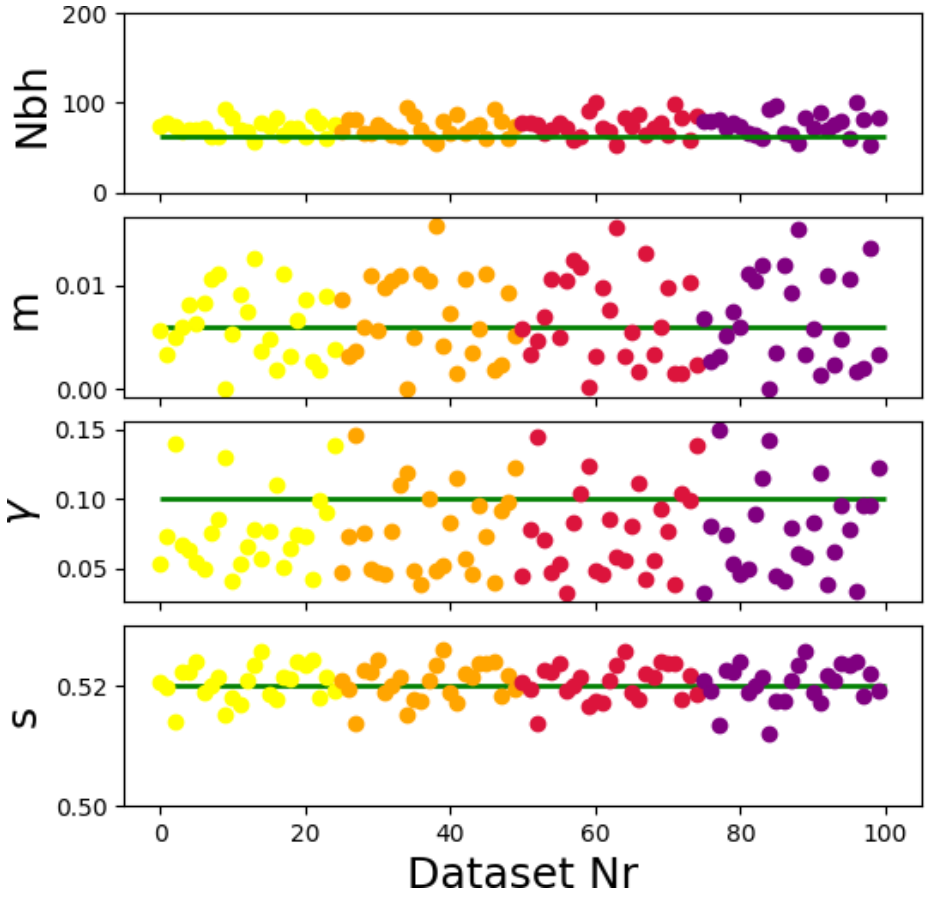

Figure S5: Various degrees of binning: We simulated 25 datasets of 60×40 individuals spaced one dispersal unit apart ($m = 0.006$, Nbh $= 4\pi 5 \approx 62.83$ and $\sigma(p) = 0.1$ and a barrier at the middle of the x-axis with $\gamma = 0.1$). For analysis, we binned individual data into demes of 1×1 (yellow), 2×2 (orange), 3×3 (red), 4×4 (purple) individuals. We used Method 3 to analyze the resulting data sets.

Literature Cited

Barton, N., A. Etheridge, J. Kelleher, and A. Véber, 2013 Inference in two dimensions: allele frequencies versus lengths of shared sequence blocks. Theoretical population biology **87**: 105–119.

Bradburd, G. S., P. L. Ralph, and G. M. Coop, 2013 Disentangling the effects of geographic and ecological isolation on genetic differentiation. Evolution **67**: 3258–3273.

Fisher, R. A. *et al.*, 1947 219: The spread of a gene in natural conditions in a colony of the moth panaxia dominula l. .

Nickisch, H. and C. E. Rasmussen, 2008 Approximations for binary gaussian process classification. Journal of Machine Learning Research **9**: 2035–2078.

Rasmussen, C. E. and C. K. Williams, 2006 *Gaussian processes for machine learning*, volume 1. MIT press Cambridge.