Figure 1: Example of the coarse m_e approximation when partial divergence matters.

Approximation of the effective migration rate in windows

We approximate the gff in window i of map length L_i by

$$g_i = g_{ii} \prod_{j \neq i} g_{ij}$$

where

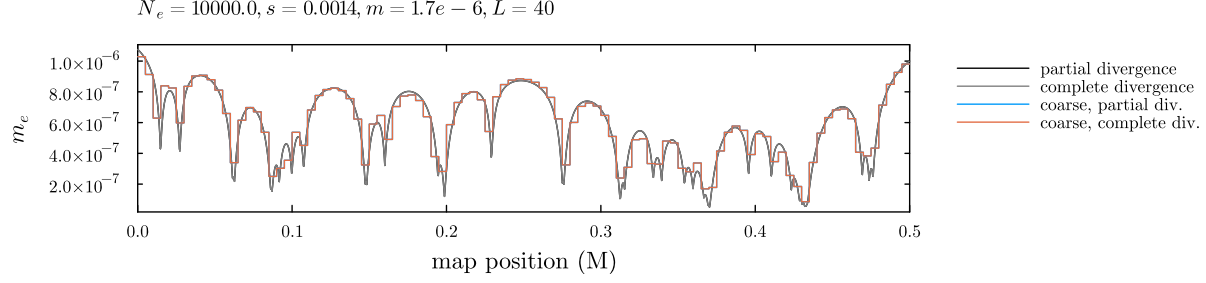$$g_{ii} = \frac{1}{L_i} \int_0^{L_i} \exp\left(-\sum_{j=1}^{X_i} \frac{s\Delta}{s\Delta + m + r(x_j, x)}\right) dx$$

and

$$g_{ij} = \exp\left(-\frac{s\Delta X_j}{s\Delta + m + \bar{r}_{ij}}\right)$$

Here Δ is a measure of the allele frequency divergence at selected loci between the two populations. In the detailed model of Zwaenepoel, Sachdeva, and Fraïsse (2024), expected allele frequency divergence due to genetic drift is predicted at each selected locus individually. In this coarse-grained approximation, we account for partial divergence through a single quantity Δ instead. Note that one could in principle also work with the expected divergence for each window. Setting $\Delta = 1$ amounts to assuming complete divergence, i.e. selection is strong relative to drift ($N_e s \gg 1$).

A crude way to estimate a single Δ is by using the harmonic mean recombination rate among selected loci and solving for the allele frequencies using the fixed-point iteration outlined in Sachdeva (2022) and Zwaenepoel, Sachdeva, and Fraïsse (2024).

Note: Accounting for partial divergence in this way is likely not very relevant, as it mostly amounts to substituting some $s_e = s\Delta$ for s in the complete divergence model. This is not entirely true of course, since Δ depends on s in a complicated way. It would become more relevant if we allow for heterogeneous selection coefficients across the genome, in which case we may have that some loci are subject to swamping while other remain differentiated.

$N_e = 10000.0, s = 0.0014, m = 1.7e-6, L = 40$

Figure 2: Example of the coarse m_e approximation when divergence is essentially complete.

Bayesian inference using MCMC

For observed data y, we have the following generative model

$$\theta \sim \text{SomePrior}(\dots)$$
$$\nu \sim \text{Gamma}(\alpha, \beta)$$
$$X_i|\nu \sim \text{Poisson}(\nu L_i)$$
$$y_i|X, s, m, \dots \sim \text{Evolution}(\dots)$$

where ν is the density of selected sites per unit of map length, X_i the number of selected sites in window i, L_i the map length of window i, y_i the data in window i, and θ lumps together all other (hyper)parameters $(m, s, u, N_e, \dots)$.

We devise a sampling scheme which makes use of the conjugacy of the Poisson and Gamma distributions. Specifically, we use a Gibbs sampler that cycles through the following updates:

$$\theta|\nu, X, y$$
$$\nu|X, \theta$$
$$X_i|\nu, \theta, y$$

where for the first sampling step we use a Metropolis-Hastings update, for the second we sample from the analytically available posterior (exploiting conjugacy) and for the third we calculate the posterior probabilities

$$\Pr\{X_i = k|y, \nu, \theta\} = \frac{f(y|X_i = k, \nu, \theta) \Pr\{X_i = k|\nu\}}{\sum_{j=0}^{l} f(y|X_i = j, \nu, \theta) \Pr\{X_i = j|\nu\}}$$

for $X_i = \{0, 1, \dots, l\}$, where l is chosen so that $\Pr\{X_i > l|\nu\} < \epsilon$ for some suitably chosen ϵ. This sampler is reasonably efficient.

Composite likelihood for the two-population model

We outline a composite likelihood approach similar to the one used by Elyashiv et al. (2016) and Murphy et al. (2022), but in the context of a two-deme model.

We assume two demes, labeled A and B, connected by unidirectional migration forward in time from B to A (so that, backward in time, lineages move from A to B at rate m).

2

Each deme is assumed to follow neutral Wright-Fisher dynamics, with coalescence rates $\lambda_A = 1/2N_A$ and $\lambda_B = 1/2N_B$ in population A and B respectively. We assume an infinite-sites model where mutations occur at rate μ per site, and each mutation occurs at a previously unmutated site.

Considering a sample of two haplotypes from each population (or a single diploid individual in each population), we can distinguish five different states or site patterns

F	fixed	all samples are fixed for the same allele
FD	fixed difference	samples from the different populations are fixed for alternative alleles.
HA	heterozygous in A	the samples from A have different alleles, whereas those from B have identical alleles
HB	heterozygous in B	the samples from B have different alleles, whereas those from A have identical alleles
HAB	heterozygous in A and B	the samples within each population have different alleles

Under the stated model, one can determine in a relatively straightforward manner the probability of each states given the relative rates of mutation, coalescence and migration, as these are competing exponential processes, and the state is determined as soon as a mutation happens (as a consequence of the infinite sites assumption). The expressions are highly unwieldy, but are readily found using a computer algebra system. A simple composite (CL) likelihood approach then suggests itself: count site patterns in observed data and use a Multinomial likelihood. Setting the mutation rate to some reasonable estimate, one can then estimate parameters on the time-scale set by the mutation rate.

A similar CL approach may be possible under the IM model, where the site-pattern likelihoods could be obtained using the generating function approach (perhaps symbolically?).

Bayesian inference with the composite likelihood

Fig. 3 shows results from a forward simulation with 100 loci under selection along a 10M chromosome (a one-chromosome genome, say). The detailed theory of Zwaenepoel, Sachdeva, and Fraïsse (2024) fits the observations quite nicely. Partial divergence appears to matter, but there seems to be no swamping and we assume a homogeneous architecture, so one can accomodate this by inferring a smaller selection coefficient I think.
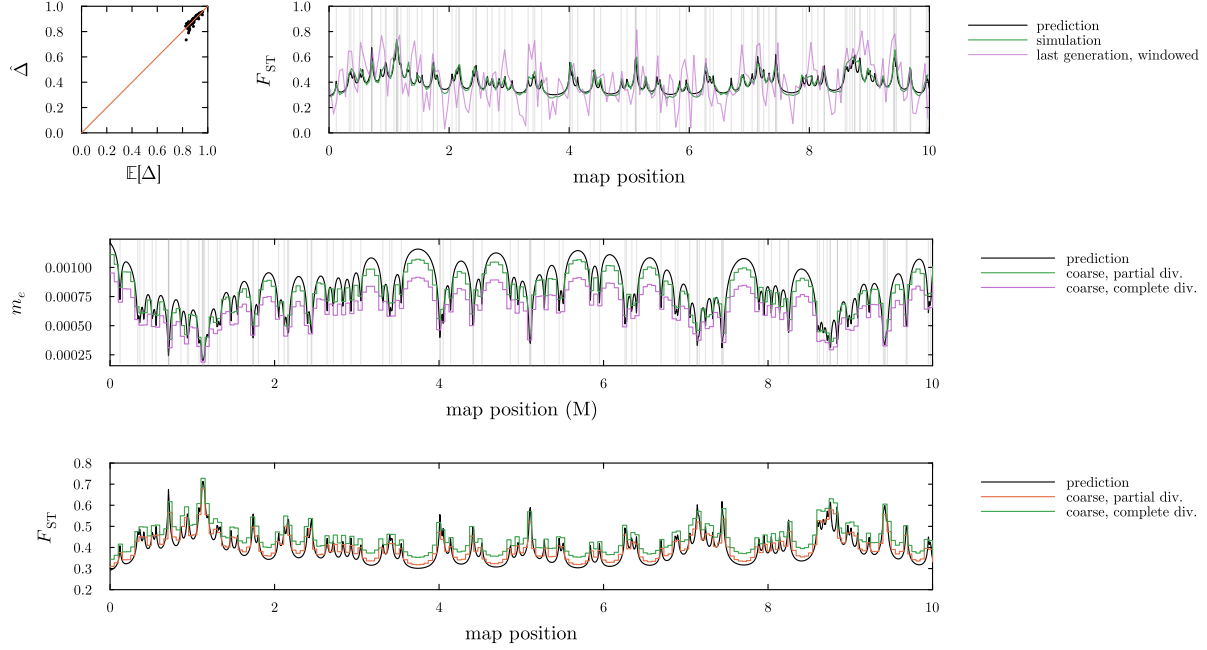
We conduct inference for the following model:

$$m \sim \text{Exponential}(0.008)$$
$$s \sim \text{Exponential}(0.01)$$
$$\lambda \sim \text{Exponential}(1/500)$$
$$\alpha \sim \text{Exponential}(1)$$
$$\nu \sim \text{Gamma}(10, 1)$$
$$X_i | \nu \sim \text{Poisson}(\nu L_i)$$
$$y_i | X, s, m, \lambda, \alpha \sim \text{Multinomial} \left(n_i, p(u, m_{e,i}, \lambda) \right)^c$$
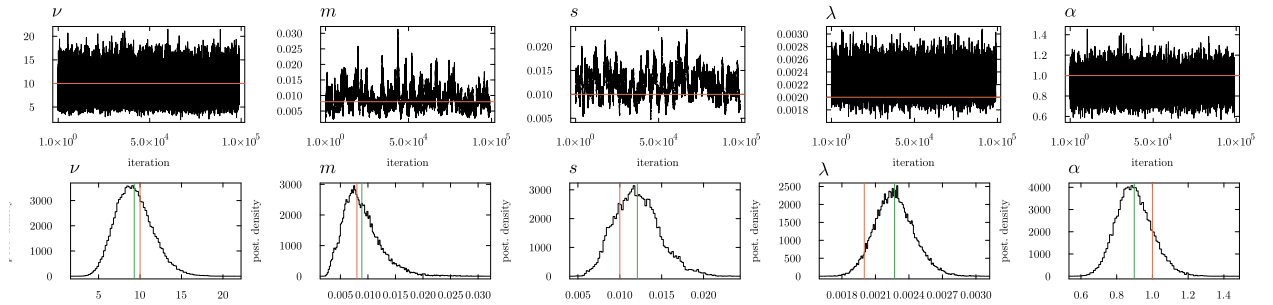
The mutation rate is assumed known (one could also assume a known scaled mutation rate I guess). We compsoite the likelihood over all available 2×2 samples from the A and B populations. The ^{c} indicates that we use a power likelihood to calibrate the composite likelihood. For a $k_A \times k_B$ sample, we used $c = (k_A(k_A - 1)/2 \times k_B(k_B - 1)/2)^{-1}$, i.e. the reciprocal of the number of 2×2 comparisons. Here we assumed the available data to consist of $k_A = 5$ samples form A and $k_B = 5$ from B. Fig. 4 shows trace plots and marginal posterior densities. Fig. 5 shows the inferred number of selected sites in each window and the inferred m_e profile. It seems that we do recover a reasonable m_e profile.
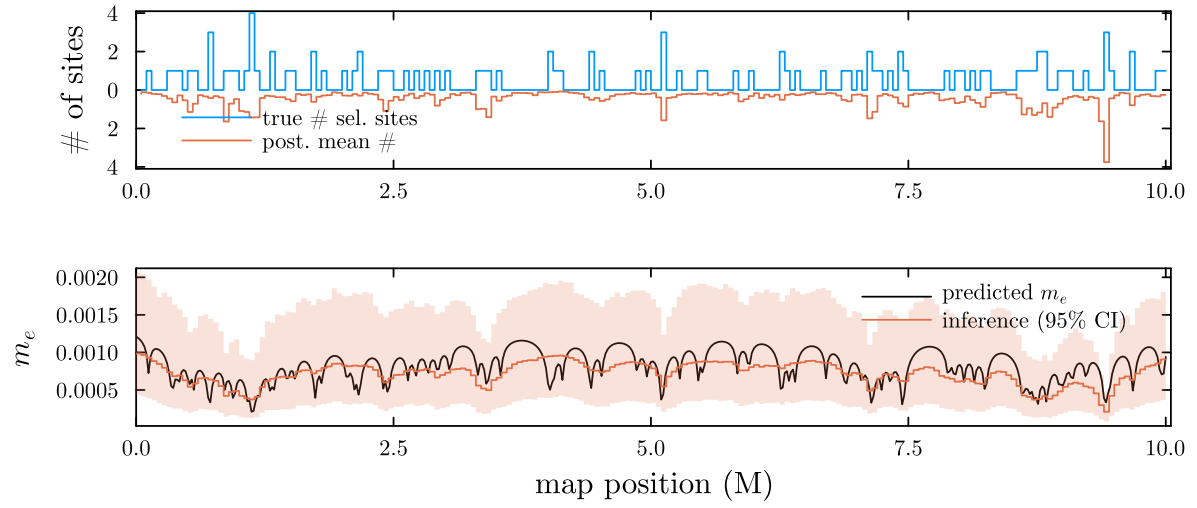
References

Elyashiv, Eyal, Shmuel Sattath, Tina T Hu, Alon Strutsovsky, Graham McVicker, Peter Andolfatto, Graham Coop, and Guy Sella. 2016. "A Genomic Map of the Effects of Linked Selection in Drosophila." *PLoS Genetics* 12 (8): e1006130.

Murphy, David A, Eyal Elyashiv, Guy Amster, and Guy Sella. 2022. "Broad-Scale Variation in Human Genetic Diversity Levels Is Predicted by Purifying Selection on Coding and Non-Coding Elements." *Elife* 12: e76065.

Sachdeva, Himani. 2022. "Reproductive Isolation via Polygenic Local Adaptation in Sub-Divided Populations: Effect of Linkage Disequilibria and Drift." *PLoS Genetics* 18 (9): e1010297.

Zwaenepoel, Arthur, Himani Sachdeva, and Christelle Fraïsse. 2024. "The Genetic Architecture of Polygenic Local Adaptation and Its Role in Shaping Barriers to Gene Flow." *Genetics*, iyae140.

Figure 3: Example of a forward simulation. We assume the two-island model, with two populations of size 500, assuming unidirectional migration at rate $m = 0.008$ and 100 selected biallelic loci uniformly scattered across the genome, each with selection coefficient $s = 0.01$. In addition, we track evolution at 100.000 neutral biallelic loci. Top row, left: predicted vs. observed allele frequency divergence at selected sites. Vertical lines are the locations of selected sites. Top row, right: predicted vs. observed F_{ST} (neutral sites, averaged in windows). Middle row: different m_e predictions. Bottom row: different F_{ST} predictions.

Figure 4: Trace plots for inference $m, s, \lambda = 1/N_{e,A}, \alpha = N_{e,A}/N_{e,B}$ and ν for a 5×5 sample of the simulated data (cfr. fig. 3). We use $c = 0.01$ as power likelihood. The mutation rate u is assumed known. A Gamma(10,1) prior is assumed for ν, and exponential priors with mean set to the true values for the other parameters.

Figure 5: Marginal posterior distribution for the number of selected sites and m_e in each window, based on the same sample as displayed in fig. 4.