Notes on modeling heterogeneous gene flow across the genome

Arthur Zwaenepoel

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.

But: when there is heterogeneous selection, calculating/estimating a single Δ may not work at all. We could and should check how the coarse approximation behaves in the presence of heterogeneous selection and swamping (Himani also stressed this – i.e. is the case with

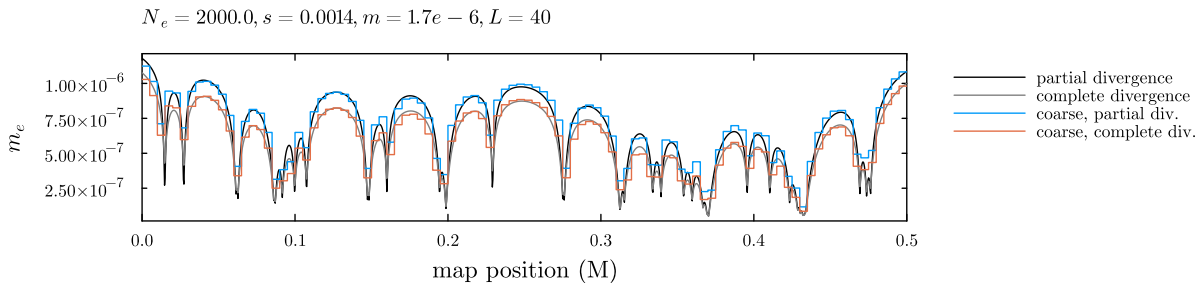

Figure 1: Example of the coarse m_e approximation when partial divergence matters.

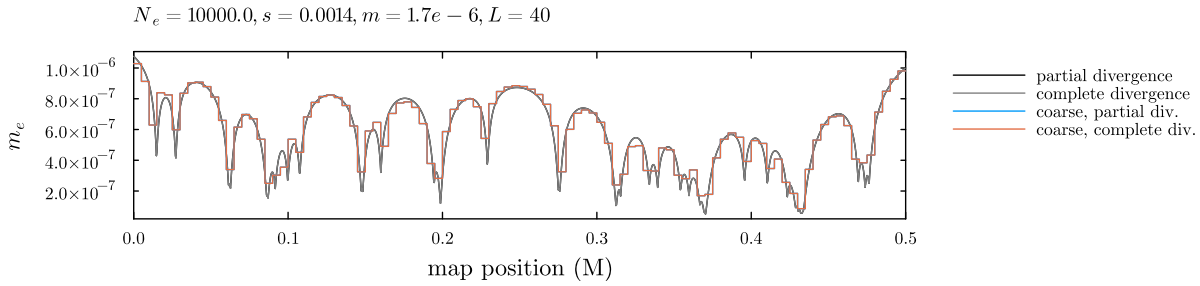

Figure 2: Example of the coarse m_e approximation when divergence is essentially complete.

homogeneous selection relevant at all? I guess I find it so, as answering the question "if s were constant, how many loci are there under selection" is already nontrivial to answer...).

Note: The coarse-grained model with partial divergence, if it works (cfr. comments above), is tricky to combine with the mixture model approach to account for N_e variation (due to BGS). This is because partial divergence depends on drift, so in the marginalization that is performed at each site, one would have to recalculate m_e for everey component, and likely this involves a whole-genome calculation (unless I am overlooking something)... So if we want to combine m_e and N_e heterogeneity, I think one would have to condition on a specified N_e profile, and not use some marginalized mixture model.

Heterogeneous selection coefficients

When we ignore heterogeneity in selection coefficients, it is unclear what we will infer. However, the coarse model outlined above with $\bar{s}$ for s yields not too different predictions – it mostly seems to exaggerate the effect of weakly selected loci (fig. **??**). But again, in an inference setting, where the data is heterogeneous but we use the coarse homogeneous model, it is unclear what this would mean for our inference, and to what extent one could hope to think of $\hat{s}$ as an estimator of $\bar{s}$.

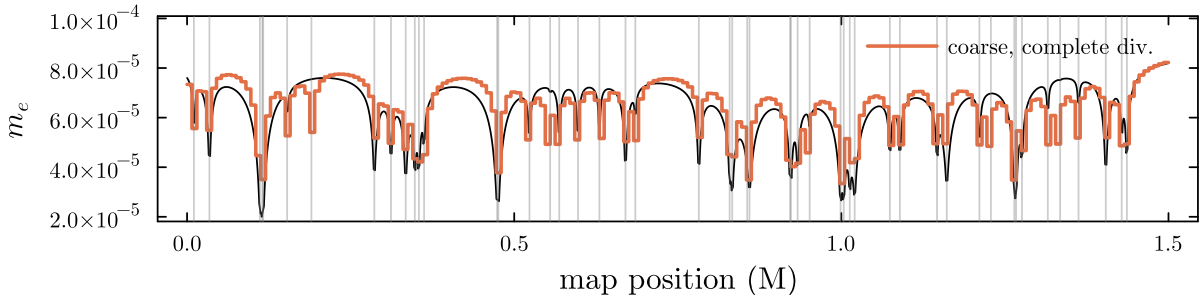

Figure 3: $L = 50$ from an exponential DFE in a setting where complete divergence is OK an assumption.

Bayesian inference using MCMC

For observed data y, we have the following generative model

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

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:

$$\begin{aligned} &\theta|\nu, X, y \\ &\nu|X, \theta \\ &X_i|\nu, \theta, y \end{aligned}$$

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). 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.

Important challenge: 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. 4 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 accommodate this by inferring a smaller selection coefficient I think.

We conduct inference for the following model:

$$\begin{aligned} m &\sim \text{Exponential}(0.008) \\ s &\sim \text{Exponential}(0.01) \\ \lambda &\sim \text{Exponential}(1/500) \\ \alpha &\sim \text{Exponential}(1) \\ \nu &\sim \text{Gamma}(10, 1) \\ X_i|\nu &\sim \text{Poisson}(\nu L_i) \\ y_i|X, s, m, \lambda, \alpha &\sim \text{Multinomial}\left(n_i, p(u, m_{e,i}, \lambda)\right)^c \end{aligned}$$

The mutation rate is assumed known (one could also assume a known scaled mutation rate I guess). We composite 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. This would be an overcorrection, as it corresponds roughly to the assumption that a single site provides a single site pattern count, whereas it should cooresponds to an effective number of data points which is somehow in between 1 and c^{-1}.[1] [2] Here we assumed the available data to consist of $k_A = 5$ samples form A and $k_B = 5$ from B. Fig. 5 shows trace plots and marginal posterior densities. Fig. 6 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.

Empirical calibration of the composite likelihood

The composite likelihood yields an unbiased approximation to the true likelihood (in the sense that MLE's are unbiased), but yields to narrow credible intervals in a Bayesian context. This is because we treat the data (sites, 2×2 comparisons) as independent, wherease they are not (because of linkage and the genealogical tree relating all samples, respectively). We can use a power likelihood to correct for this overconfidence, i.e. we raise the lieklihood to some power c, where $c < 1$. One can think of this as rescaling the likelihood by figuring out how many effectively independent data points one has (roughly: we have c effective data points per site that enters the composite likelihood calculation).

The question remains: how to choose c? Recall that we have chopped the genome in windows. We assume that compositing window likelihoods is OK (linkage is sufficiently weak between windows). A possible empirical approach to obtain a decent value of c is the following:

1. For a given parameter set $\theta = (m, N_{e,A}, N_{e,B}, u)$, simulate an $n \times n$ sample y from the coalescent with recombination for a stretch of genome corresponding to the window

[1] Perhaps it would be better to use something like $((k_A + k_B)/2)/(k_A(k_A - 1)/2 \times k_B(k_B - 1)/2)$. Somehow the number of effective comparisons should be related to the number of nodes in the genealogy, which is linear in the number of leaves, but it is complicated by the two-population setting, so should it be proportional to the number of nodes in the subpopulation genealogies? It would be good to check with our empirical calibration approach.

[2] Note that gimble uses the $n_A \times n_B$ diploid comparisons. Not exactly sure why, is it because they don't assume phased data, but do use blocks of multiple sites?

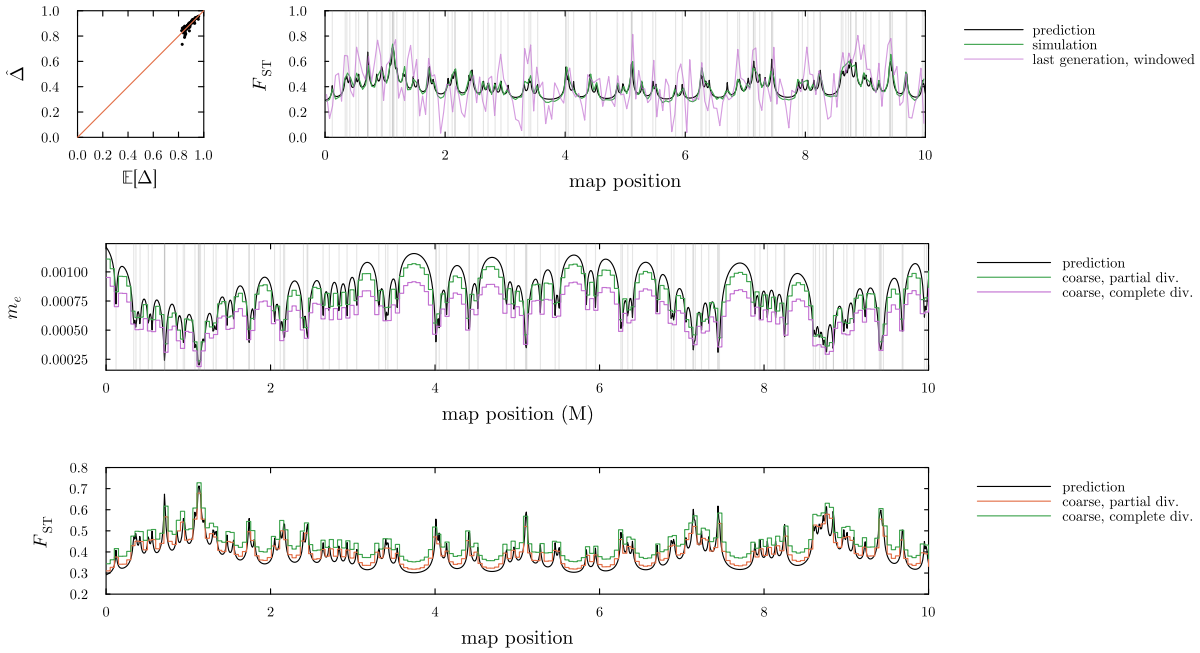

Figure 4: Example of a forward simulation. We assume the two-island model, with two populations of size 500, assuming unidirectional migration at rate $m = 0.008$ and 100 selected biallelic loci uniformly scattered across the genome, each with selection coefficient $s = 0.01$. In addition, we track evolution 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.

size.

2. Determine the posterior distribution $p(m|y)$ using ABC (conditioning on the true values for the other parameters).
3. Choose $\hat{c} = \mathrm{argmin} KL(p(m|y), p_c(m|y))$, where $p_c(m|y)$ is the posterior with the composite likelihood as sampling distribution.

The overall idea/motivation is: full-scale ABC is infeasible, but we can do ABC on a small scale (single window, single parameter) to inform the composite likelihood calibration.

There are a couple of issues with this.

1. It is noisy, as it depends on a single realization of the coalescent (y) which is noisy. (possible solution: we can repeat the procedure and get the mean of $\hat{c}$).
2. How should one choose the other parameters? Do they matter (much)? Some preliminary simulations suggest they do matter.

Implementation details of the calibration approach

There are many ways to do the ABC-based calibration outlined above. A basic approach which appears to work is as follows:

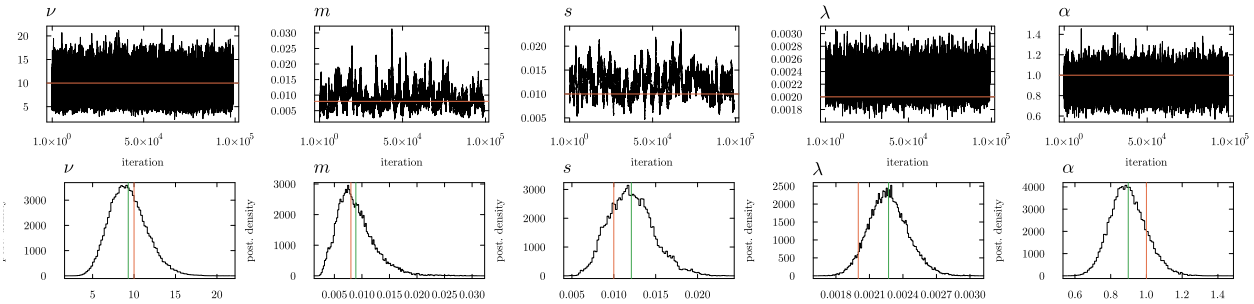

Figure 5: Trace plots for inference $m, s, \lambda = 1/N_{e,A}, \alpha = N_{e,A}/N_{e,B}$ and ν for a 5×5 sample of the simulated data (cfr. fig. 4). We use $c = 0.01$ as power likelihood. The mutation rate u is assumed known. A Gamma(10,1) prior is assumed for ν, and exponential priors with mean set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

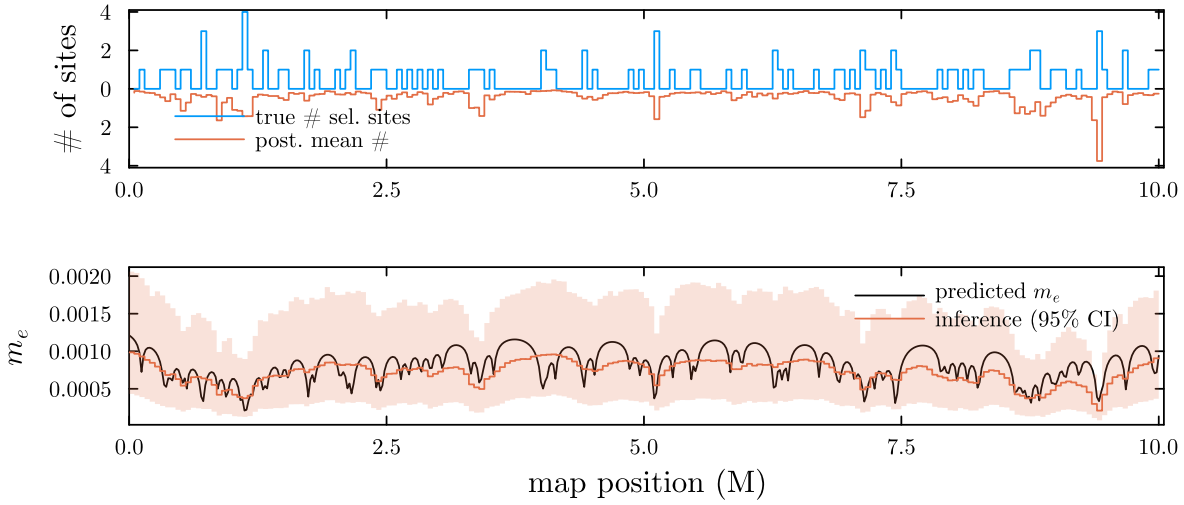

Figure 6: Marginal posterior distribution for the number of selected sites and m_e in each window, based on the same sample as displayed in fig. 5.

1. Simulate a data set from $y|\theta, m$, calculate the jSFS. Call this the calibration data set.
2. Simulate n_{ABC} data sets from $y|\theta$, i.e. integrating over the prior (draw m from the prior, draw $y|\theta, m$ as in (1), using the same sample size as in (1). Calculate the jSFS, and calculate the sum of squared differences SSD between this jSFS and the jSFS of the calibration data set.
3. Obtain the subset of (2) for which $SSD < \epsilon$, where ϵ is for instance the 0.05 percentile of the SSD values. The associated m values are an ABC approximation to the posterior $p(m|\theta, y)$.
4. Fit a Lognormal (or Gamma) approximation to $p(m|\theta, y)$.
5. Find $\hat{c} = \text{argmin}_c KL(p(m|\theta, y), p_c(m|\theta, y))$
6. Repeat this a number of times and take the mean of the $\hat{c}$ values.

Doing this for the simulation data set analyzed above, we get $\hat{c} \approx 0.0005$ for a 10×10 sample (fig. 7), which is more or less accidentally about the reciprocal of the number of comparisons.

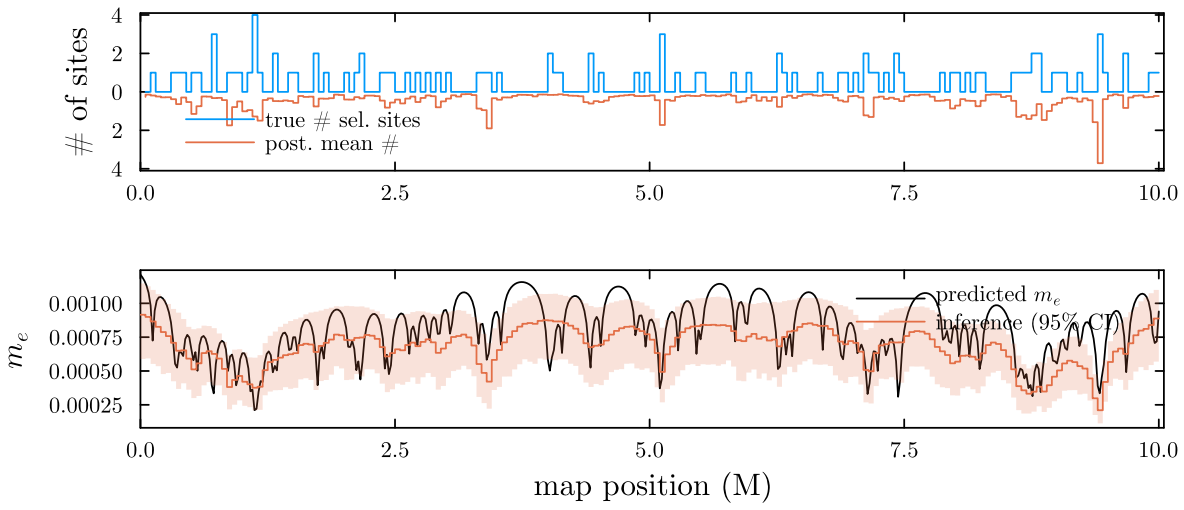

Figure 7: Results with the calibrated CL.

Does it suggest that this heuristic (the reciprocal of the number of 2×2 samples) is a good one to start from?

Segmentation

One idea I suddenly had is that the site pattern counts provide a 'sequence' which could be segmented as one does DNA sequence data. The idea that one can segment other types of sequence using such an approach I first encountered in a very different context Nakatani and McLysaght (2017) while working on my PhD. This is interesting, since one would then segment based on empirical homogeneity in the site patterns, which sounds like it would give more relevant windows for inferring parameters in than choosing an arbitrary constant window size. I applied my old implementation of this (which I implemented for the inference of macrosynteny) and it worked out of the box (fig. 8).

Note: I wonder to what extent these segments correspond to segments spanned by a single marginal tree in an inferred ARG?

Note: In principle it would be possible to do everything jointly: instead of the Dirichlet-Multinomial likelihood, one could use the two deme coalescent composite likelihood. In fact, the power likelihood approach probably amounts somehow to a Dirichlet-Multinomial thing where the Dirichlet parameters are determined by the Coalescent composite likelihood + some pseudocount. Doing everything jointly is perhaps not feasible or very interesting, but this perspective of using a Dirichlet 'layer' in between is perhaps more helpful than the power likelihood view?

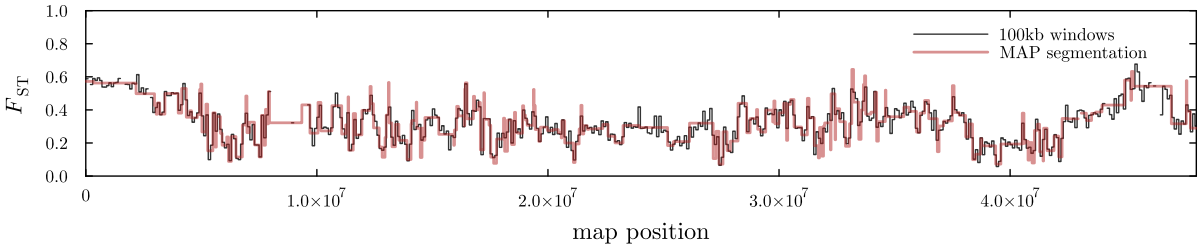

Figure 8: Segment F_{ST} estimates for the MAP segmentation vs. constant 100kb window segmentation. (minimum window size 25kb, $\alpha = 10$, we got 235 segments instead of the 481 windows of 100kb).

Simulations

Full forward simulations

These would have to include

1. Divergent selection
2. BGS
3. Neutral loci

We should probably use SLiM, so that we don't have to do (3) explicitly (use tree sequence recording). If I implement tree sequence recording in my forward simulations, I'll end up rewriting $x\%$ of `tskit` I fear...

For (1) we should focus on the near complete divergence setting. For (2) we should focus on the 'strong' BGS limit, where the Hudson & Kaplan theory applies (and a rescaling of N_e is fine). We should ignore more complicated regimes with selective interference initially.

Backward simulations based on m_e approximations

One could take the m_e predictions from the coarse or detailed approximations in windows, and simulate data backward in time within these windows. I guess this is of interest to study the performance of the inference machinery, under the assumption that the many modeling assumptions hold.

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.

Nakatani, Yoichiro, and Aoife McLysaght. 2017. "Genomes as Documents of Evolutionary History: A Probabilistic Macrosynteny Model for the Reconstruction of Ancestral Genomes."

Bioinformatics 33 (14): i369–78. https://doi.org/10.1093/bioinformatics/btx259.

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.