Composite likelihood approach for the migration 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.

Heliconius example

We tested this approach by simulating data in windows along chromosome 18 of *Heliconius melpomene/cydno* using `msprime` under the IM model, assuming $\mu = 2 \times 10^{-9}$, $T = 4$My and the average recombination rate in each window. We assume the effective population sizes estimated by gIMble in each window, and use for simplicity a single population size for both A and B equal to the mean of the gIMble estimates for both populations. We assumed $m_i = \mathbb{E}\left[m_{e,i}\right]$, where the latter are obtained using the Gibbs sampler fitted to the gIMble output (with $\sigma = 1/4$). We then used the following probabilistic program (using `Turing.jl` Ge *et al.* (2018)) to conduct inference of the coalescence and migration rates in windows using the CL approach described above. We sampled down the simulated data so that we have an average number of sites (including non-polymorphic sites) of about 5000. The approach appears to work surprisingly well, see fig. 1.

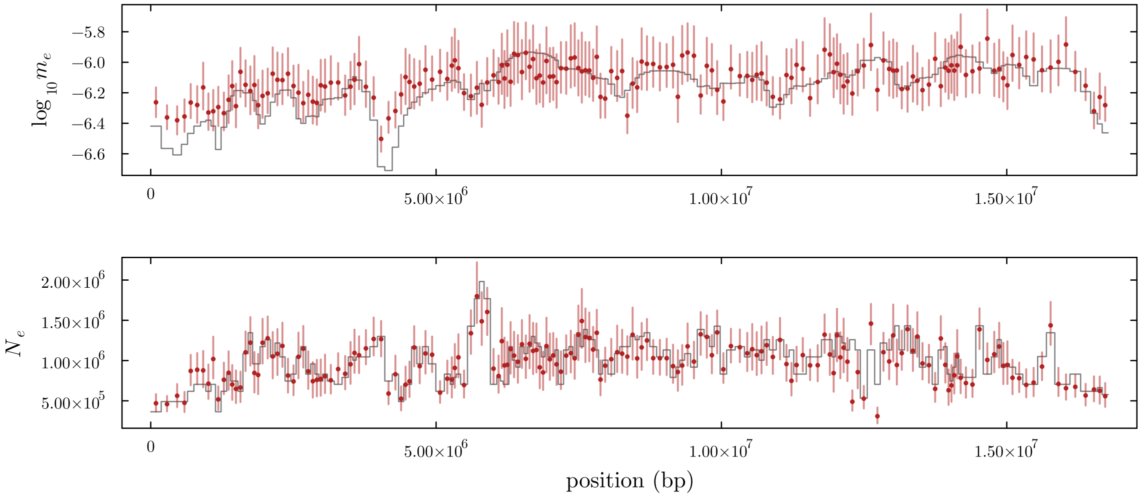


Figure 1: Marginal posterior estimates of window-specific migration and coalecence rates for `msprime` simulations on *Heliconius* chromosome 18.

```julia
@model function tmodel(data, u=2e-9)
    n = length(data)
    tau ~ Exponential(1)  # controls variation across windows
    m ~ Normal(5.5, 5.0)  # mean log-scale migration rate
    l ~ Normal(5.5, 5.0)  # mean log-scale coalescence rate
    lms ~ MvNormal(fill(m, n), tau)  # log-scale migration rates
    lls ~ MvNormal(fill(l, n), tau)  # log-scale coalescence rates
    ms = exp.(lms)  # vector of migration rates for each window
    ls = exp.(lls)  # vector of coalescence rates for each window
    for i=1:length(data)  # for each window i
        # calculate the expected proportion of each site pattern
        ps = Barriers.probs(ms[i]*u, u, ls[i]*u, ls[i]*u)
        if !isprobvec(ps)  # handle numerical issues
            Turing.@addlogprob! -Inf
        else  # Multinomial likelihood
            data[i] ~ Multinomial(sum(data[i]), ps)
        end
    end
end
```

Note that this approach assumes an equilibrium two-deme model, ignoring the divergence of the two populations at time $T = 4$My ago. The results from Laetsch *et al.* (2023) suggest that for this data set, this does not really matter much. Of course one wonders whether one could include T in the calculation of the site pattern probabilities.

We seem to overestimate m rather consistently. It would be good to find out the source of this bias. We expect to overestimate m when we ignore the fact that $T_{div} < \infty$ (see below).

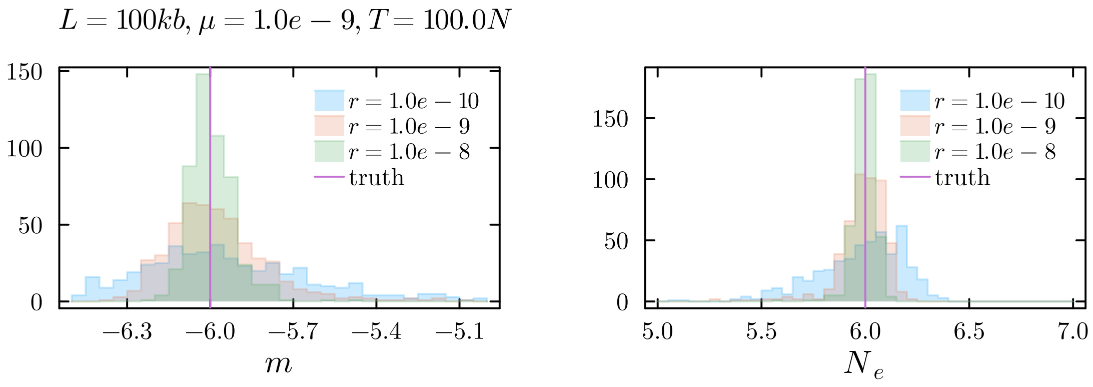


Figure 2: Maximum likelihood estimates of m and N_e for a 100kb region for different recombination rates. x-axis is on a $\log_{10}$ scale.

IM model

The above assumes the M(igration) model, can we derive similar quantities for the IM model? This model has one more parameter, but it does not act like just another competing exponential.

More detailed simulations

It is worthwhile to explore in what parameter regimes this works.

Fig. 2 suggests that we get roughly unbiased estimates, but with a larger variance when linkage gets tight. This is roughly what I expected (there might be a slight bias for N_e.

Fig. 3 suggests that when we wrongly assume an equilibrium migration model for the case $Nm = 1$, we strongly overestimate the migration rate when $T < 2N$, but not so much when $T \geq 2N$. The reason for overestimation is obvious: when T is small, the different lineages merge into one population where they can freely coalesce rather quickly, this looks the same as having a high migration rate in the equilibrium two-island model. Note that N_e is estimate very accurately, but this may partly be because we assume all three populations have the same N_e in our simulations. In fig. 5 we relax this assumption and get a bias in the B population size, which is the sink population, forward in time. Note that the ancestral population has the same size as the source population in our simulations.

I think this suggests that if (1) we are wishing to estimate parameters on a scale of say 100kb windows (the window size is not too important), with (2) $T \geq 2N_e$ and (3) $1/Nm < T$ (where T is expressed in units of N).

This is fairly restrictive, but nevertheless may be relevant, especially in systems with high rates of gene flow.

Notes on composite likelihood

See (**hudson2001?**), Murphy *et al.* (2022), *etc.*

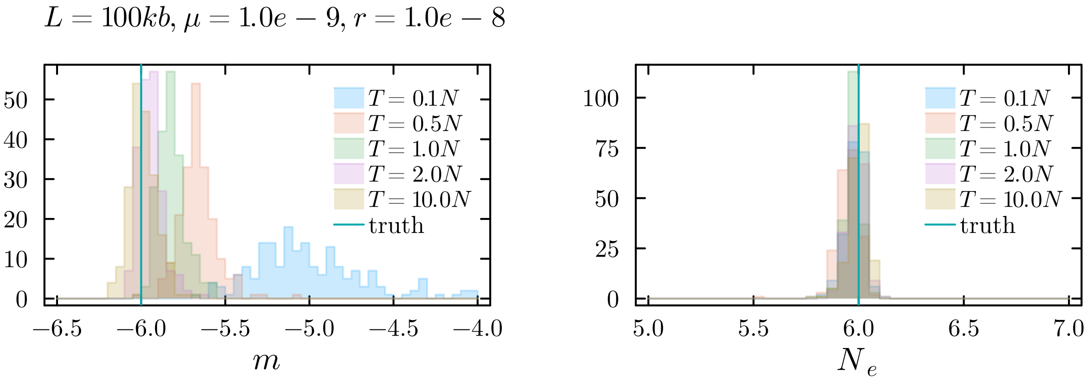


Figure 3: Maximum likelihood estimates of m and N_e for a 100kb region for different T_{div} (note that the CL expression assumes $T_{div} \to \infty$). x-axis is on a $\log_{10}$ scale.

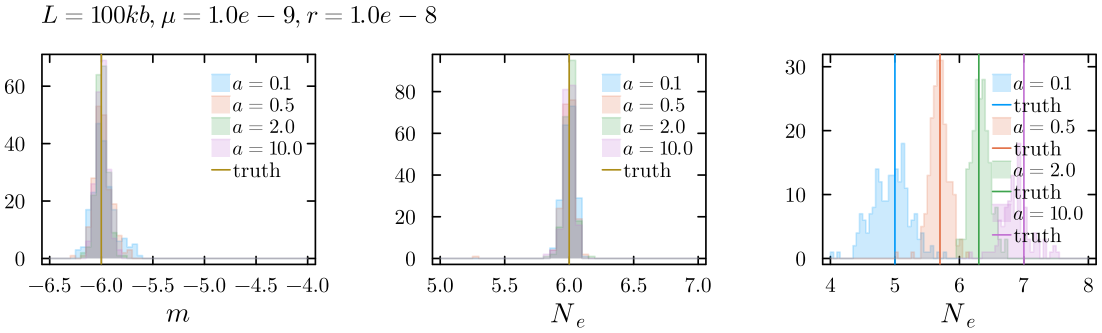


Figure 4: With different population sizes in A and B (C has the B population size). Here $T = 10N_A$.

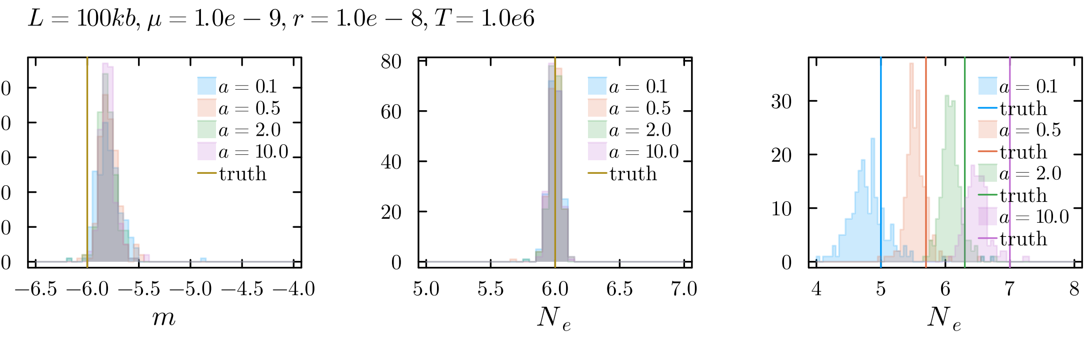
References

Elyashiv E., S. Sattath, T. T. Hu, A. Strutsovsky, G. McVicker, *et al.*, 2016 A genomic map of the effects of linked selection in drosophila. PLoS genetics 12: e1006130.

Ge H., K. Xu, and Z. Ghahramani, 2018 Turing: A language for flexible probabilistic inference, pp. 1682–1690 in *International conference on artificial intelligence and statistics*, PMLR.

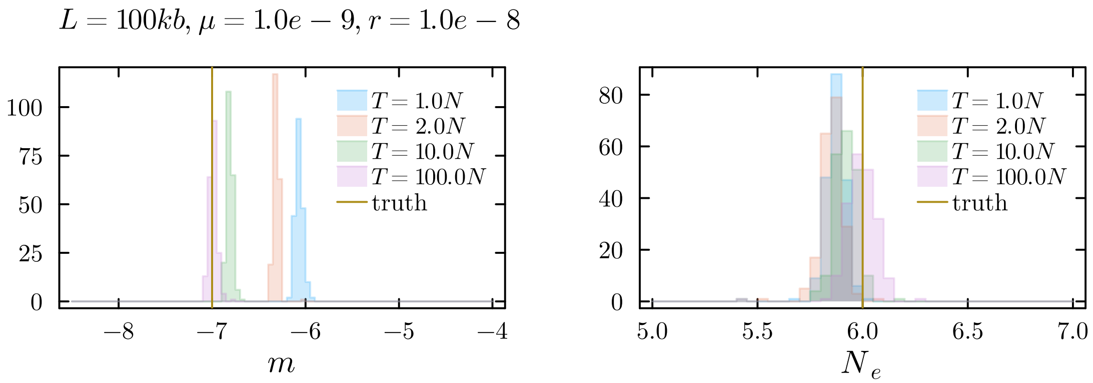
Laetsch D. R., G. Bisschop, S. H. Martin, S. Aeschbacher, D. Setter, *et al.*, 2023 Demographically explicit scans for barriers to gene flow using gIMble. PLoS genetics 19: e1010999.

Murphy D. A., E. Elyashiv, G. Amster, and G. Sella, 2022 Broad-scale variation in human genetic diversity levels is predicted by purifying selection on coding and non-coding elements. Elife 12: e76065.



$L = 100kb, \mu = 1.0e-9, r = 1.0e-8, T = 1.0e6$

Figure 5: With $T = N_A$ and asymmetric population sizes



$L = 100kb, \mu = 1.0e-9, r = 1.0e-8$

Figure 6: With $Nm = 0.1$ (instead of $Nm = 1$ above), this shows that we need $Nm > 1/T$, as expected.