Joint estimation of migration rates and effective population sizes from trio coalescence times in inferred tree sequences

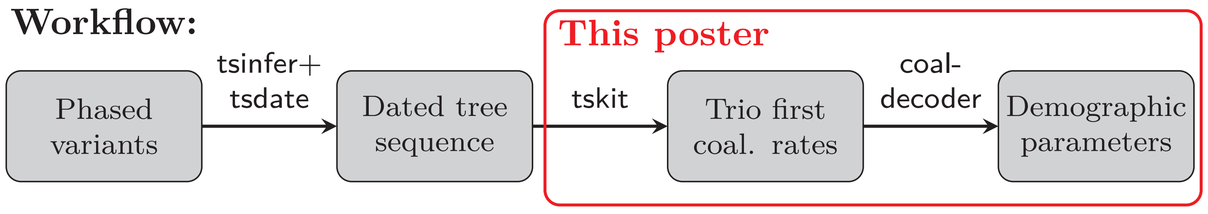
Nathaniel S. Pope (nspope@utexas.edu) Department of Entomology, Pennsylvania State University

Background

Tree sequences as a basis for demographic inference

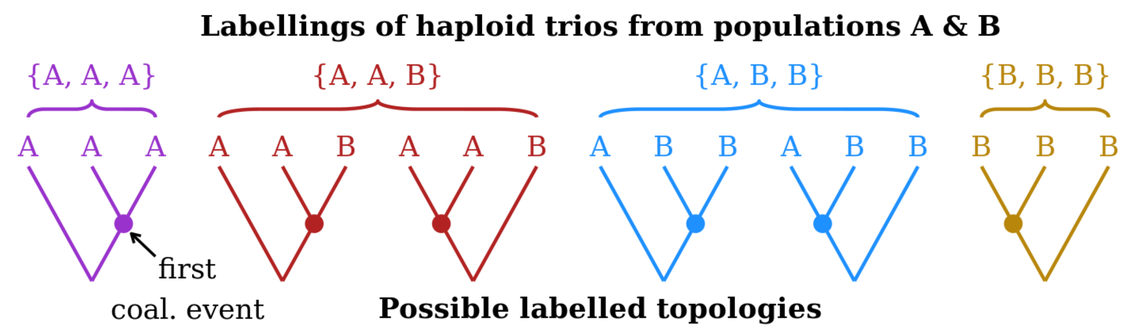
Tree sequences are efficient data structures for storing the genealogical relationships among large numbers of recombinant DNA sequences (e.g. the Ancestral Recombination Graph), that can be directly inferred from phased genetic polymorphisms (Kelleher *et al* 2019). These "empirical" ARGs are very informative about population histories, but probabilistic models for the full graph are intractable under even moderately complicated demographic scenarios. Here I propose a composite likelihood method (coaldecoder) for estimating a piecewise-constant demographic history for a set of populations, using statistics (trio first coalescence rates) that can be efficiently extracted from tree sequences using the tskit library.

Workflow:



Trio first coalescence rates

In a panmictic population, the inverse of the haploid effective population size equals the rate of coalescence between pairs of lineages. This duality has previously been used to estimate piecewise-constant population size histories from tree sequences (Speidel *et al* 2019), but does not hold when there is population structure. However, simple pairwise rates can be extended by considering first coalescence events for different population-labelled trios, which are then informative about asymmetric migration over time. For example, with two populations there are four possible population-labelled trios and six distinct events/possible topologies:



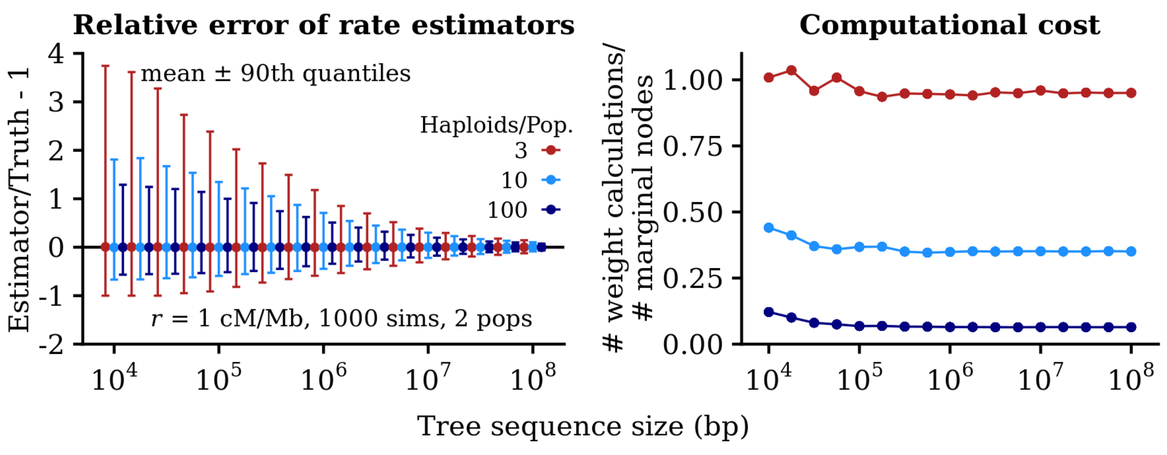
while for P populations there are $\binom{P+2}{P-1}$ labellings and $P + 3\binom{P}{3} + 6\binom{P}{2}$ distinct topologies (e.g. different trio first coalescence rates).

Coalescence rate estimation from tree sequences

In tree sequences, a given node may occur in multiple marginal genealogies. Let t_i be the time of node i measured in generations; let $\mathcal{G}_i$ be the set of genealogies containing i; and let π_j be the fraction of the tree sequence spanned by genealogy j. We discretise time so that epoch e has the time interval $(l_e, r_e]$ where $r_e = l_{e+1}$. Let $u_1u_2|u_3$ denote the trio topology $((u_1, u_2), u_3)$ where $u_k = 1 \dots P$ are population indices. Rates of first coalescence events for $u_1u_2|u_3$ in epoch e can be estimated by:

$$y_e^{(u_1u_2|u_3)} = \frac{\sum_i \mathbb{I}[r_e \geq t_i > l_e] \sum_{j \in \mathcal{G}_i} \pi_j w_{ij}^{(u_1u_2|u_3)}}{(r_e - l_e) \sum_i \mathbb{I}[t_i > l_e] \sum_{j \in \mathcal{G}_i} \pi_j w_{ij}^{(u_1u_2|u_3)}},$$

where the 'weights' $w_{ij}^{(u_1u_2|u_3)}$ count the number of $(u_1u_2|u_3)$-trios in genealogy j where node i is the first coalescence event. Simulations indicate that this estimator is unbiased and consistent, and is efficiently calculated for large numbers of samples using tree edge differences via tskit.



https://github.com/nspope/coaldecoder

Implementation

State-space formulation and composite likelihood

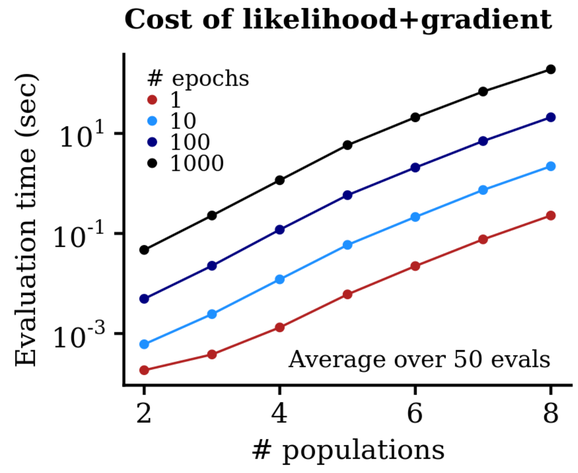
To estimate demographic parameters, we use a continuous-time Markov process to model the migration of three lineages from an initial population labelling $\ell = \{u_1, u_2, u_3\}$ back until the first coalescence event:

- The P^3 transitory states are assignments of lineages to populations.
- The 3 absorbing states are pairwise coalescences between lineages.
- The infinitesimal generator $\mathbf{Q}(\mathbf{M}_e)$ is parameterized by epoch-specific demographic parameters $\mathbf{M}_e$ for epochs $e = 1, 2, \dots$
- We penalize squared differences of $\mathbf{M}$ between adjacent epochs.
- The initial state probability vectors $\mathbf{x}_0^{(\ell)}$ are unitary, each with a non-zero element that corresponds to the transitory state matching ℓ.

This leads to a non-linear state-space model for the 'empirical' rates:

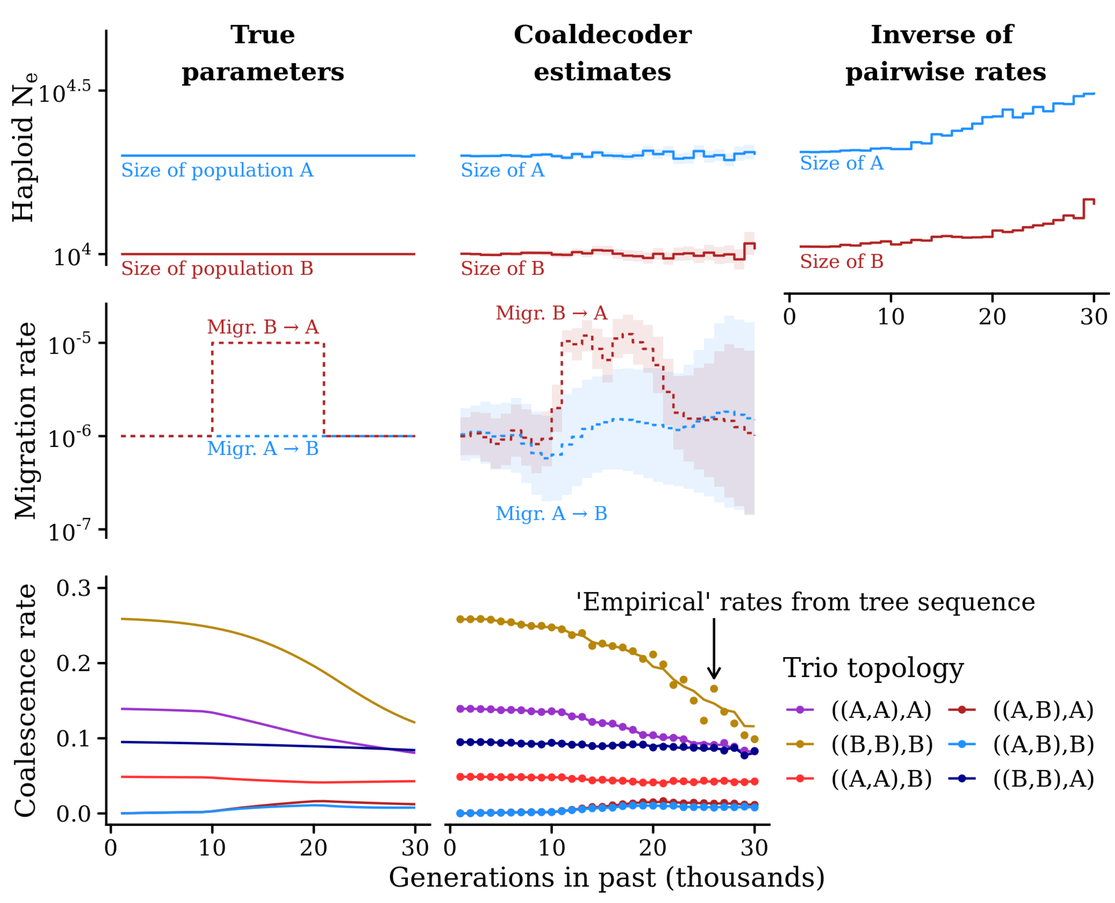
$$p(\mathbf{y}_e^{(\ell)} | \mathbf{M}_{1\dots e}) \propto \exp\{-2^{-1}(\mathbf{y}_e^{(\ell)} - \hat{\mathbf{y}}_e^{(\ell)})^T \mathbf{S}_e^{(\ell)} (\mathbf{y}_e^{(\ell)} - \hat{\mathbf{y}}_e^{(\ell)})\},$$
$$\hat{\mathbf{y}}_e^{(\ell)} = f_\ell(\mathbf{x}_e^{(\ell)}, \mathbf{x}_{e-1}^{(\ell)}), \;\; \mathbf{x}_e^{(\ell)} = \mathrm{expm}\{(l_e - r_e)\mathbf{Q}(\mathbf{M}_e)\}^T \mathbf{x}_{e-1}^{(\ell)},$$
$$\mathbf{M}_{e,i} = \mathbf{M}_{e-1,i} + \epsilon_{e,i}, \;\; \epsilon_{e,i} \sim \mathcal{N}(0, \lambda_i^{-2}),$$

where the composite likelihood is $\prod_{\ell,e} p(\mathbf{y}_e^{(\ell)} | \mathbf{M}_{1\dots e})$, $\mathbf{y}_e$ are 'empirical' rates calculated from the tree sequence, $f_\ell(\cdot)$ maps state vectors for a given ℓ onto the appropriate trio rates, and $\mathbf{S}_e$ are precision matrices estimated by block bootstrapping $\mathbf{y}_e$. The state-space structure enables very fast gradient calculation via reverse algorithmic differentiation, and we use L-BFGS-B to optimize $\mathbf{M}_e$.



Proof of concept with two-population model

I used msprime to simulate data from a two-population island model with an asymmetric pulse of migration from 10 to 20 thousand generations in the past. The inputs consisted of ten 50 Mb tree sequences, with twenty haploid samples per population and $r = 1$ cM/Mb. Cross-validation was used to choose a (global) smoothness penalty λ. Using trio first coalescence rates, coaldecoder correctly recovers this pulse and the constant population sizes over 1000-generation epochs. In contrast, inverting the within-population pairwise coalescence rates results in biased effective population size estimates, falsely giving the impression of a population decline.



References

Kelleher J, et al. 2019. Inferring whole-genome histories in large population datasets. *Nature Genetics* 51: 1330-1338.

Speidel L, et al. 2019. A method for estimating genome-wide genealogies for thousands of samples. *Nature Genetics* 51: 1321-1329.