Using ancestral recombination graphs to study reproductive isolation

Francisco Campuzano Jiménez[1], Arthur Zwaenepoel[1], Els De Keyzer[1], Hannes Svardal[1,2]

[1]Evolutionary Ecology Group, University of Antwerp, Belgium; [2]Naturalis Biodiversity Center, Leiden, Netherlands

Introduction

- The complete ancestry of a sample of genomes is described by a graph structure known as the ancestral recombination graph (ARG).
- Recent progress in algorithms makes ARG reconstruction computationally feasible for genome-scale data.

How well do reconstructed ARGs capture the genomic signatures of reproductive isolation and can they be used to study it effectively?

Case study: isolation-with-migration (IM)

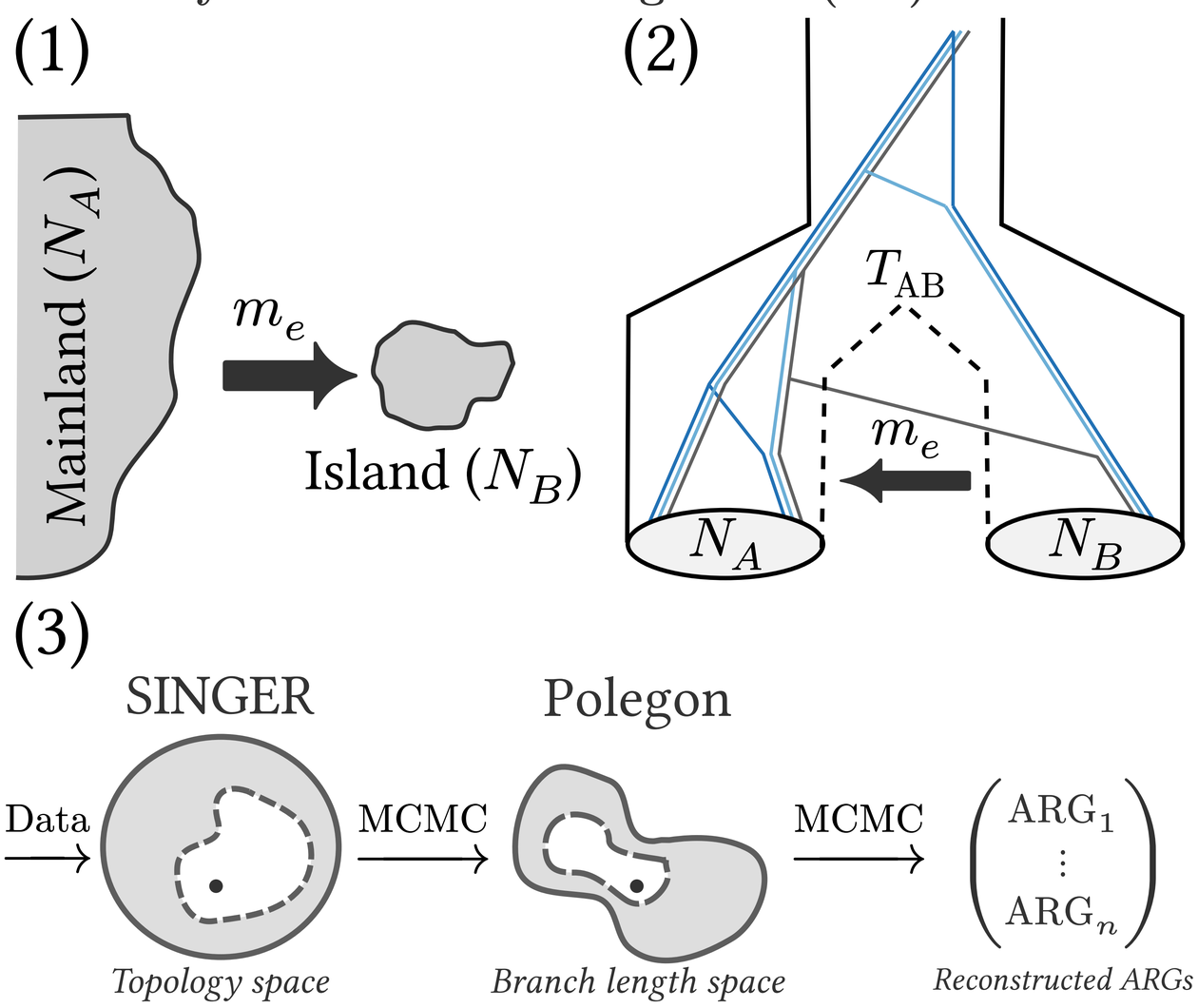


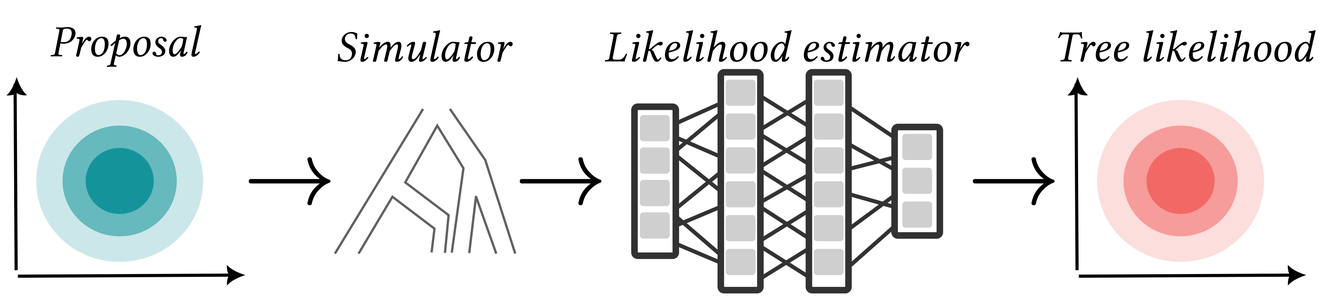
Figure 1: (1) We performed population genetic simulations under the scenario shown using `SLiM` and `msprime` and (2) recorded the true ARG. (3) We then used a state-of-the-art ARG reconstruction method, `SINGER+Polegon`, to obtain a sample of inferred ARGs. This approach uses uninformative priors (solid lines) that do not favor ARGs that are more likely under an IM model (dash lines).

Inference framework

Likelihood calculation for ARGs remains an unsolved problem and is especially hard for structured populations. A previous attempt, SCAR, was built upon assumptions that fail in the regime of migration rate that is relevant for speciation.

We address this issue with a novel and highly scalable approximate inference scheme that naturally captures local differences arising from selection.

(1) Training amortized tree likelihood estimator $\mathcal{L}_{\text{nn}}(\theta \mid t)$



(2) Approximate inference with a composite likelihood

$$\mathcal{L}(\theta|\text{ARG}) \propto \prod_{\text{every } t \text{ tree}} \mathcal{L}(\theta|t) \cdot \text{span}(t) \approx \mathcal{L}_{\text{nn}}(\theta|t) \cdot \text{span}(t)$$

Figure 2: Our approach is based on two different approximations. (1) Exact likelihood computation under the structured coalescent involves solving a high-dimensional system of ODEs. We instead train neural networks on sufficient summary statistics to approximate these likelihoods. This amortized approach achieves high accuracy and scales linearly with the number of lineages. (2) We substitute the intractable likelihood of the ARG with a composite likelihood across every marginal tree, weighted by the tree span.

Neutral demographic inference

We accurately infer demographic history, with a slight bias toward panmixia perhaps due to the use of uninformative prior that reduces statistical power. However, the appropriate scaling factor for obtaining calibrated posteriors remains unclear.

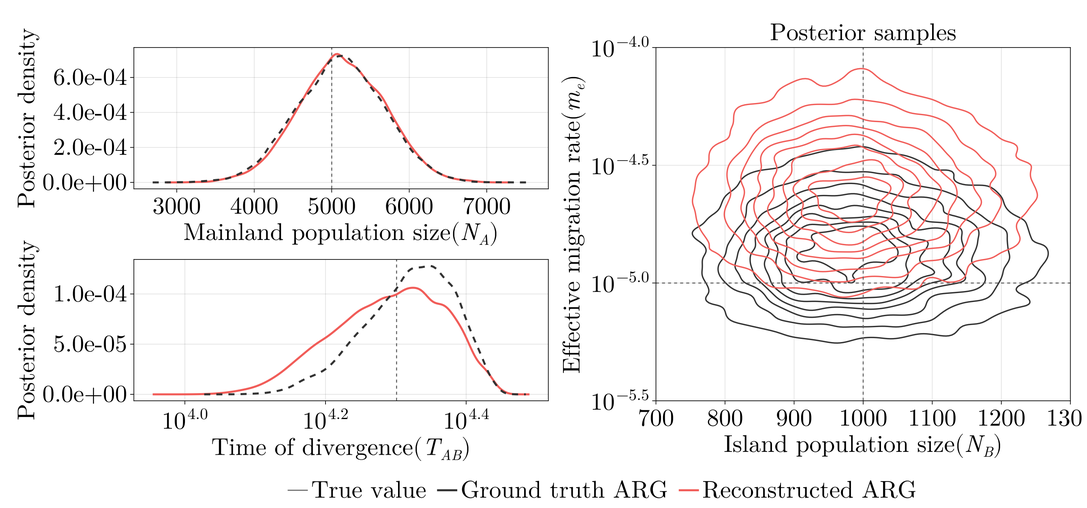


Figure 3: Genome-wide demography was inferred using a composite power-scale likelihood. The power-scale (that controls the concentration) was arbitrarily set such that the effective number of independent trees ≈ 10. Parameters were estimated via variational inference on a strictly neutral, 1-Morgan chromosome simulated for 10 diploid individuals.

Barrier to gene-flow scan

We found strong agreement between estimates from the reconstructed and true ARGs, suggesting that this approach could offer unprecedented resolution to quantify selection against gene flow.

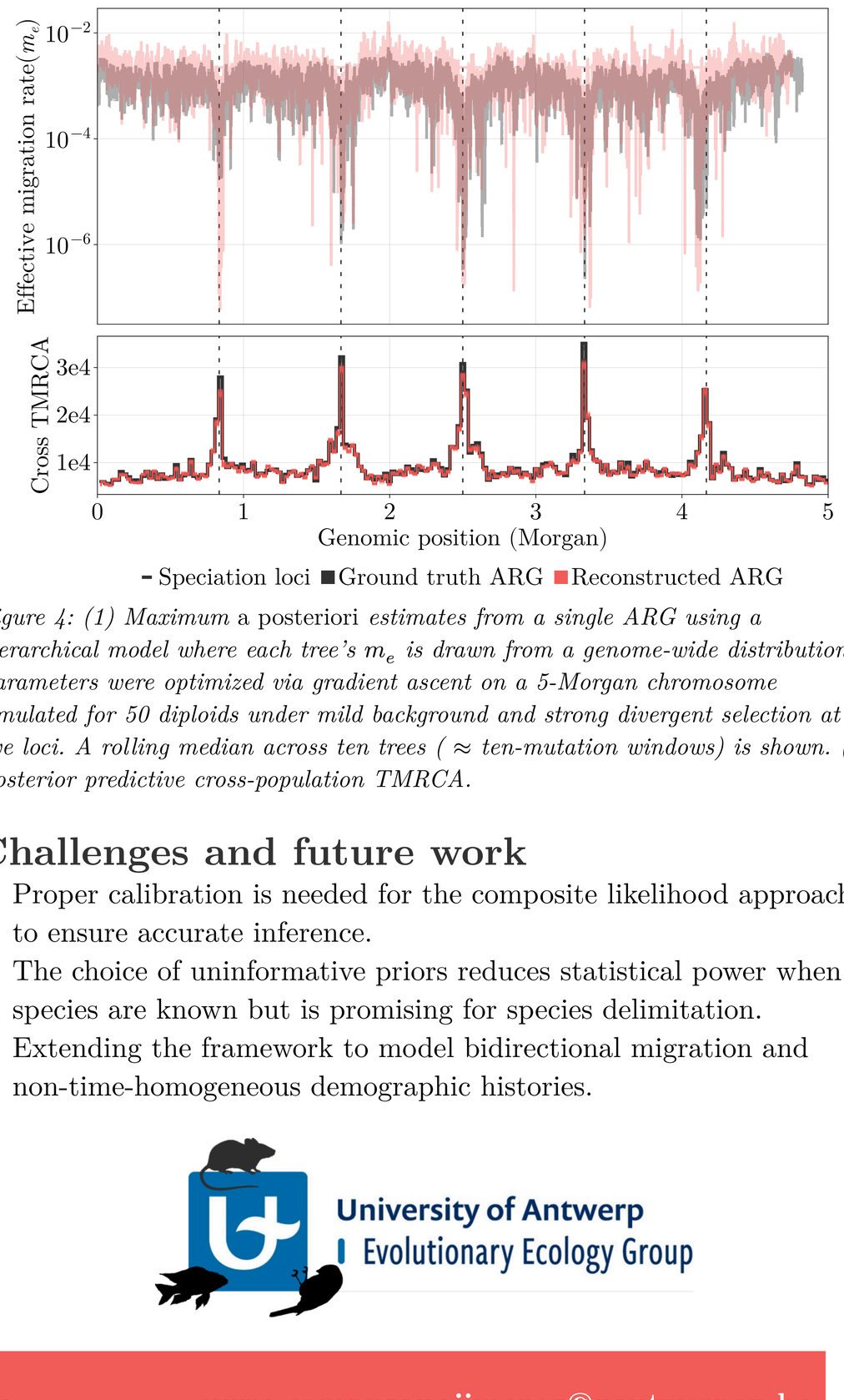


Figure 4: (1) Maximum a posteriori *estimates from a single ARG using a hierarchical model where each tree's m_e is drawn from a genome-wide distribution. Parameters were optimized via gradient ascent on a 5-Morgan chromosome simulated for 50 diploids under mild background and strong divergent selection at five loci. A rolling median across ten trees ($\approx$ ten-mutation windows) is shown. (2) Posterior predictive cross-population TMRCA.*

Challenges and future work

- Proper calibration is needed for the composite likelihood approach to ensure accurate inference.
- The choice of uninformative priors reduces statistical power when species are known but is promising for species delimitation.
- Extending the framework to model bidirectional migration and non-time-homogeneous demographic histories.

University of Antwerp
Evolutionary Ecology Group

You can find me on social media as @currocam currocam.github.io curro.campuzanojimenez@uantwerpen.be