THE UNIVERSITY OF ARIZONA

Computationally efficient demographic history inference from allele frequencies with supervised machine learning

Linh N. Tran[1,2], Connie K. Sun[2], Travis J. Struck[2], Mathews Sajan[2], Ryan N. Gutenkunst[2]

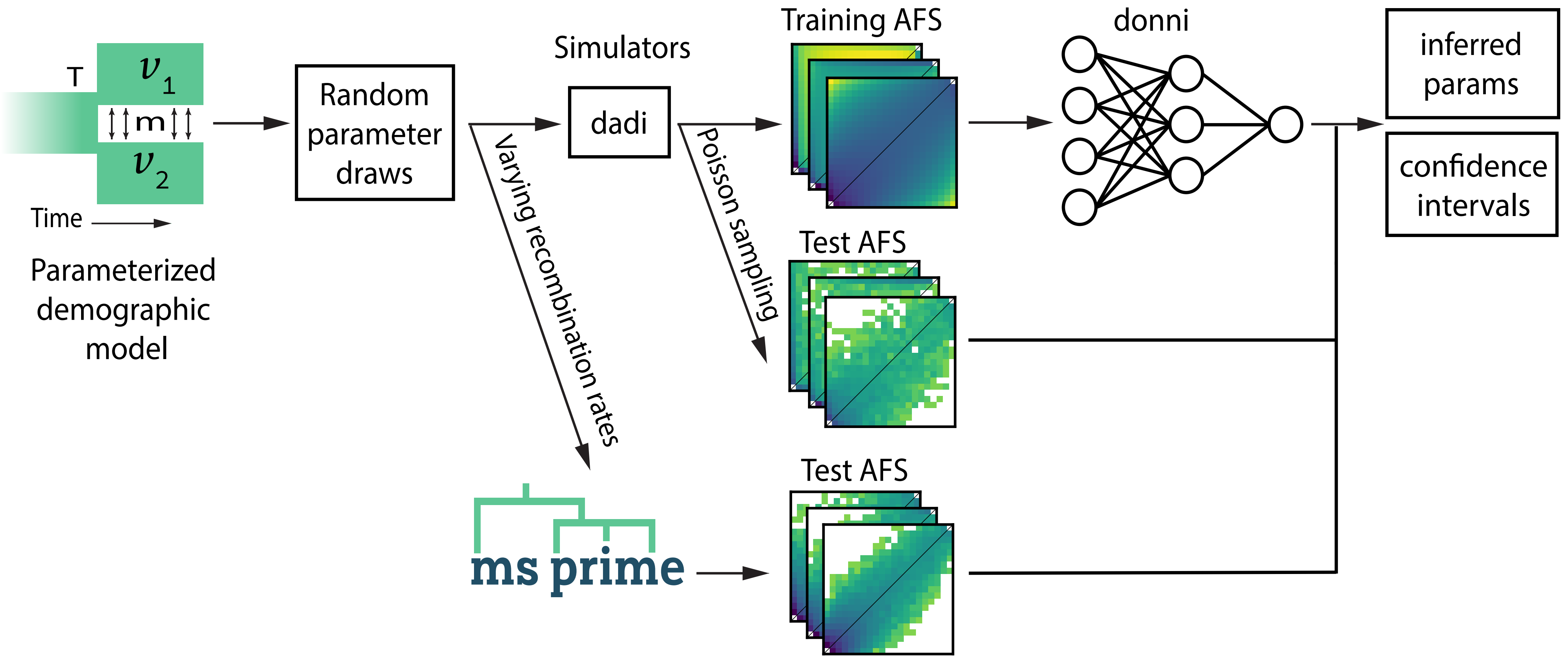
[1] Genetics Graduate Interdisciplinary Program, [2] Department of Molecular and Cellular Biology, University of Arizona; lnt@arizona.edu; http://gutengroup.mcb.arizona.edu

Introduction

Inferring demographic history of natural populations from genomic data is of central concern in many studies across research fields. Previously, our group had developed dadi, a widely-used inference method based on the allele frequency spectrum (AFS) and maximum likelihood optimization. To circumvent dadi's expensive likelihood optimization procedure, we developed donni (demography optimization via neural network inference), an inference method based on dadi and supervised machine learning that is more efficient while maintaining comparable inference accuracy.

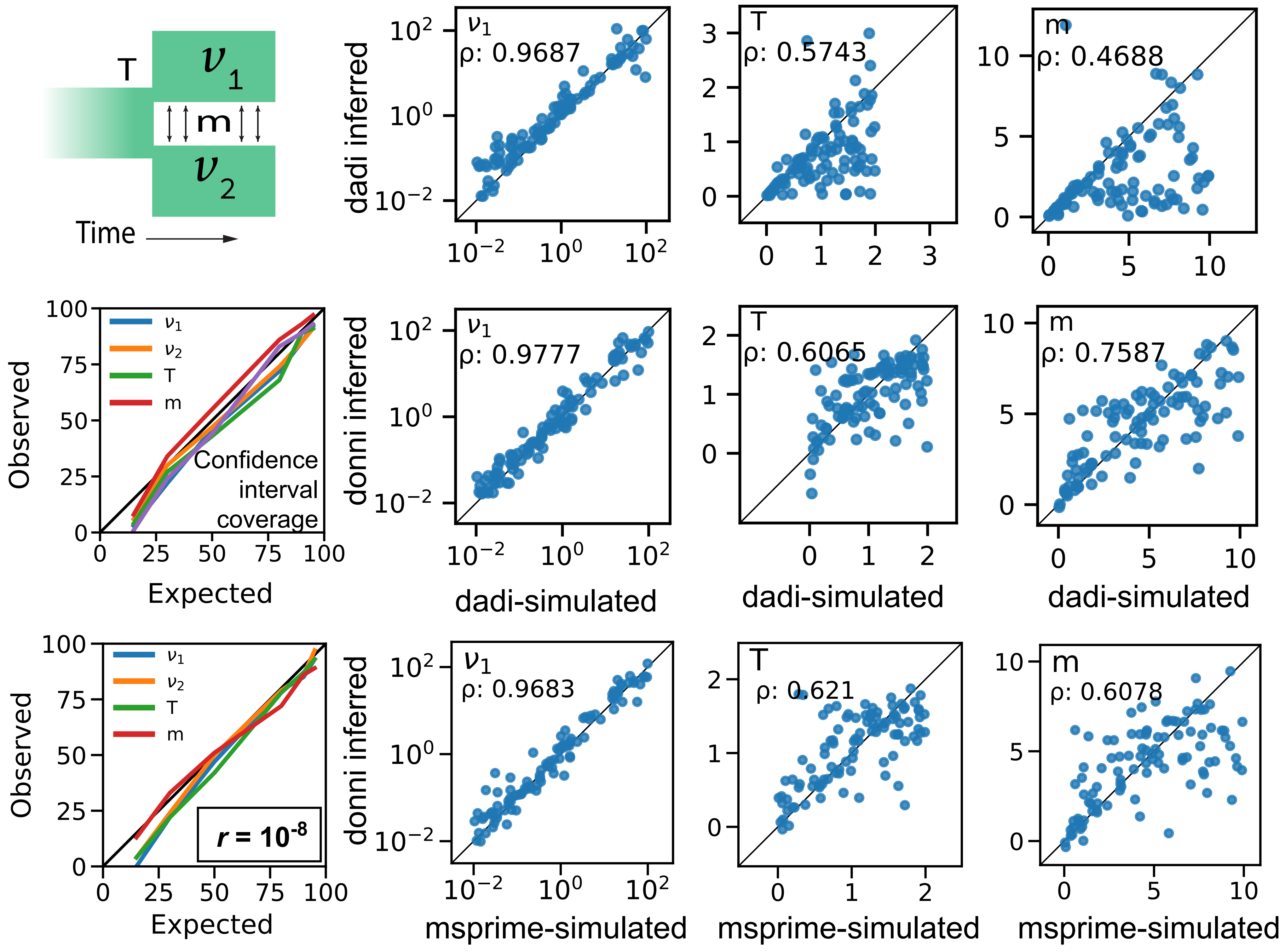
Implementation

For a given demographic model, we drew sets of model parameters from a biologically relevant range. Each parameter set represents a demographic history and corresponds to an expected AFS. We used the expected AFS simulated with dadi and their corresponding parameters as training data for the scikit-learn multi-layer perceptron (MLP) regressor algorithm. We generated test data either by Poisson sampling from dadi-simulated AFS or by varying recombination rates with msprime, resulting in a change in variance compared to training AFS. We also implemented an uncertainty quantification method using the software package MAPIE. Therefore, the output of donni's trained networks includes both inferred parameters and their confidence intervals. For automatic hyperparameter tuning, we implemented the HyperBand algorithm using scikit-learn successive halving random search.



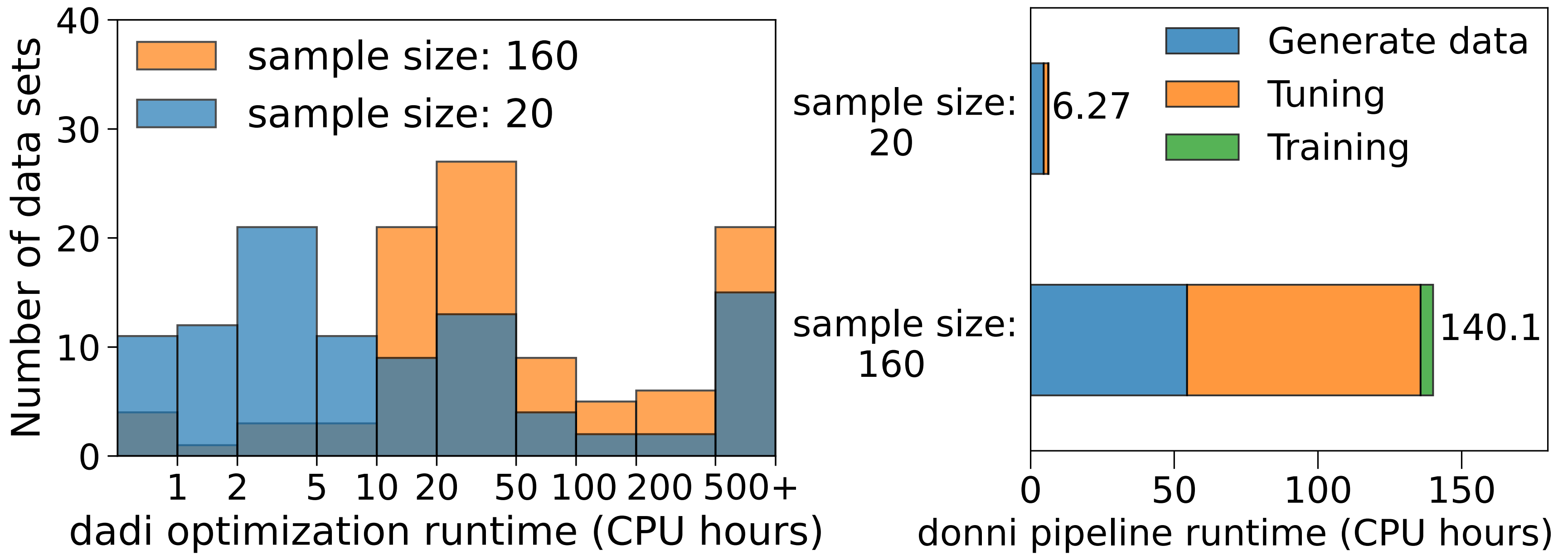
Validation

We validated donni's accuracy by comparing with dadi likelihood optimization for one- to three-population models. As shown for a split-migration model below, donni's accuracy is comparable to dadi's. donni also performed well on AFS generated with msprime under similar demographies but include linkage. donni's uncertainty quantification method is also well-calibrated for for both dadi- and msprime-simulated test AFS.



Benchmarking

To compare the efficiency of donni and dadi, we benchmarked the compute time required by each method. As shown below, there was a spread of optimization runtime among the 100 test AFS for dadi, with several difficult spectra requiring 500+ CPU hours to reach convergence. By comparison, the computation required for training MLPs with donni was less than the average time required for running dadi optimization on a single AFS. Inferring demographic parameters with donni's trained MLPs is nearly instantaneous. This result suggests that donni may benefit many cases where dadi optimization can take a long time to reach convergence.



Distribution

We have produced a suite of trained MLPs for a large collection demographic history models for sample sizes 10, 20, 40, 80, and 160 chromosomes per population, which our software automatically downloads from CyVerse. Users can also use our pipeline to train custom models that support a different sample size. Our pipeline and user code are open-source and available on GitHub: https://github.com/lntran26/donni.

References & Acknowledgements

Scikit-learn: https://scikit-learn.org/
MAPIE: https://github.com/scikit-learn-contrib/MAPIE
HyperBand: Li, L. et al., 2017. J. of ML Research

This research was supported by the National Institute of General Medical Sciences of the NIH (R01GM127348 to RNG).