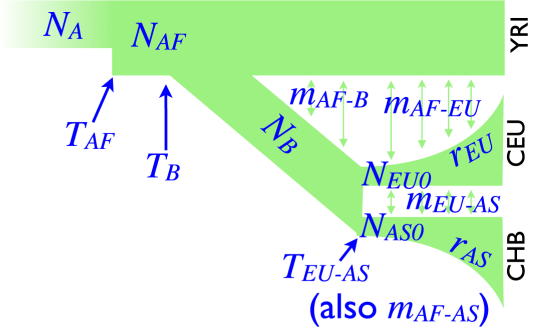
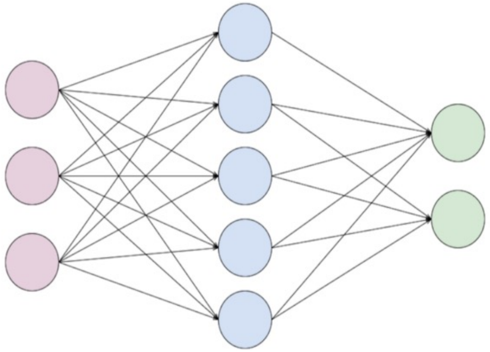
Computationally Efficient Demographic Inference with Supervised Machine Learning





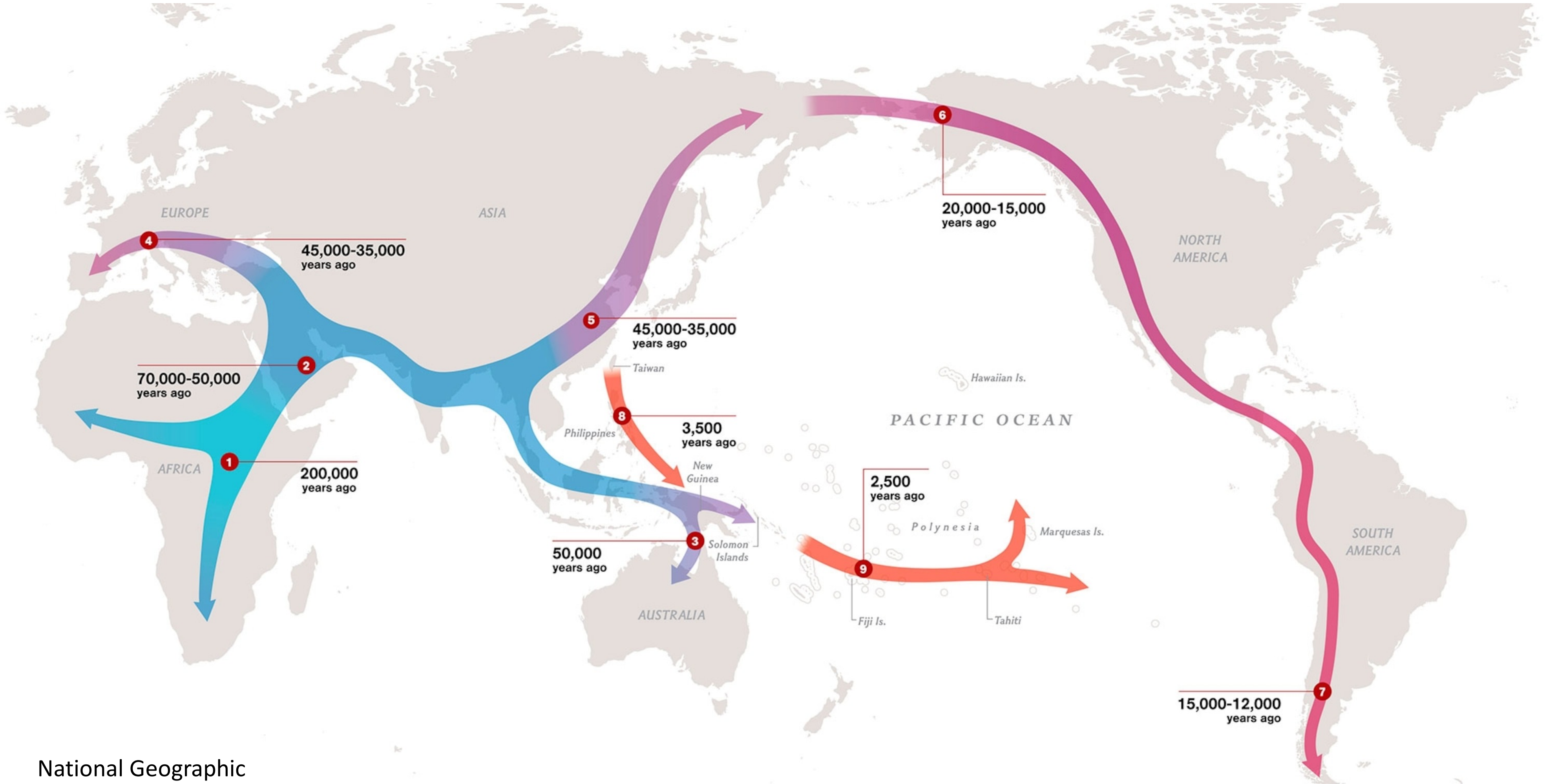
Linh N. Tran, Connie Sun, Mathews Sajan, Ryan Gutenkunst
University of Arizona

SMBE everywhere GS2
July 12, 2022

THE UNIVERSITY OF ARIZONA

NIH NIGMS

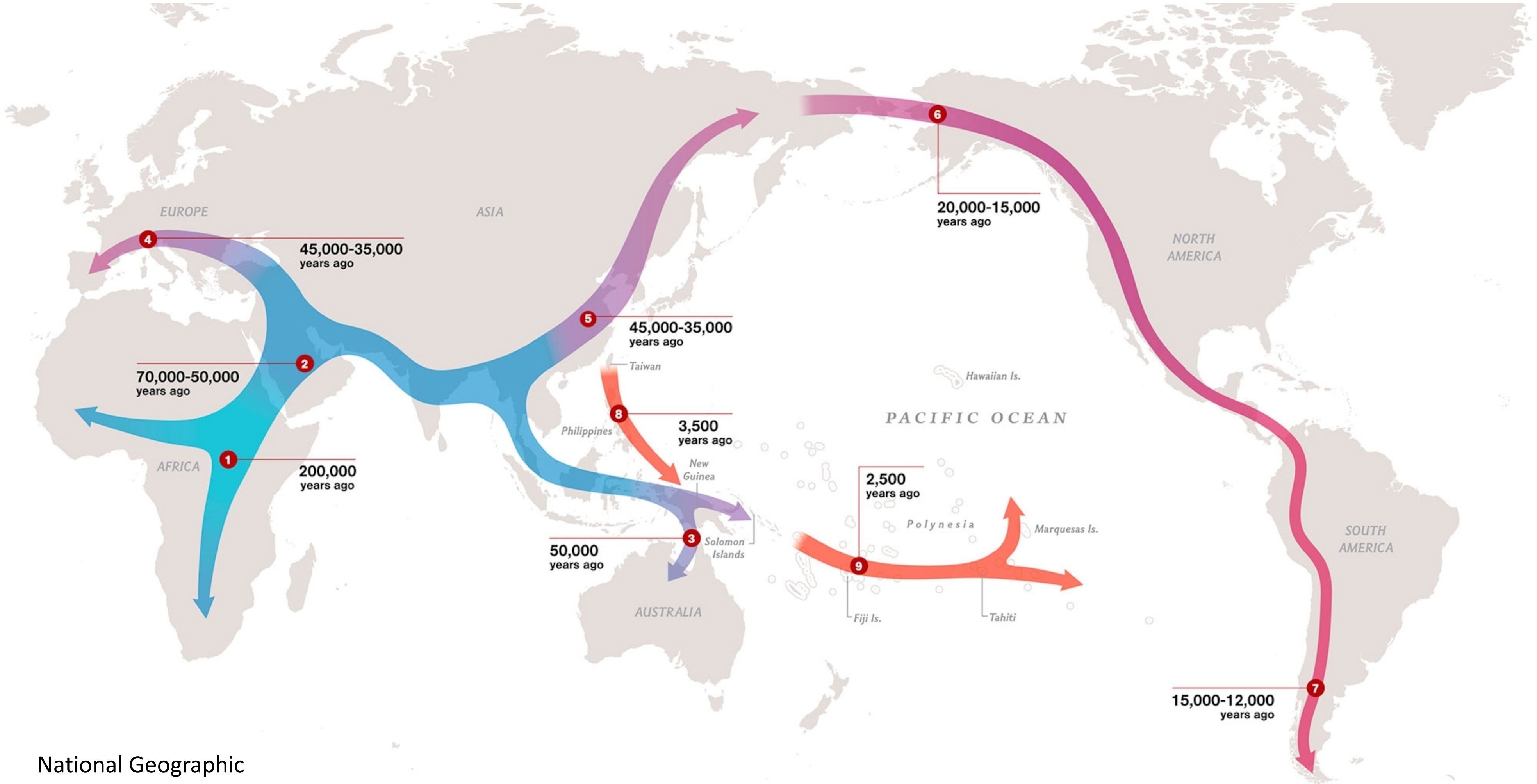
Demographic history inference



National Geographic

Demographic history inference – Motivation

- Understand population history: bottlenecks, gene flow, etc.

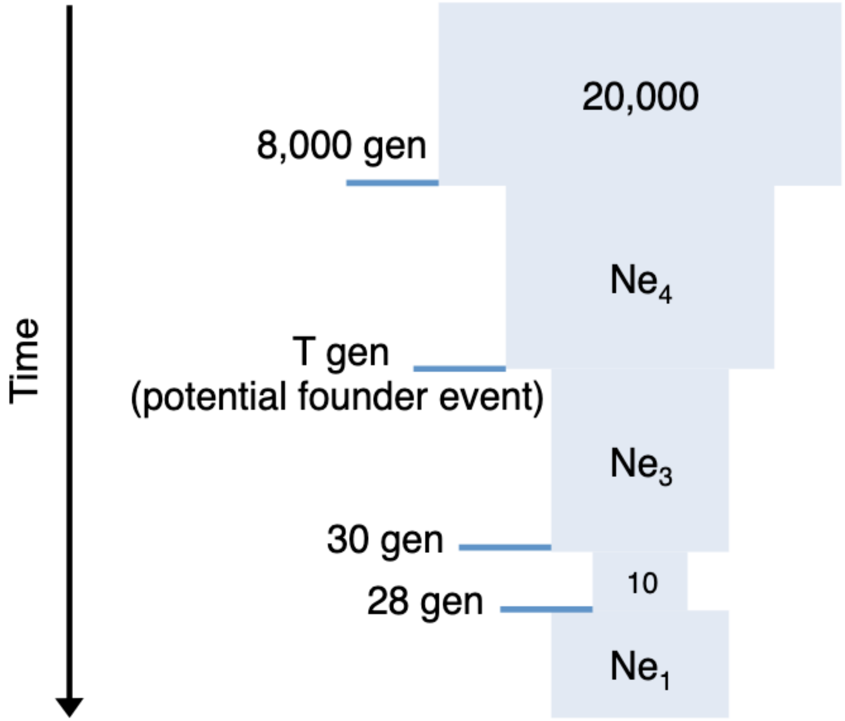


National Geographic

Demographic history inference – Motivation

National Geographic

- Understand population history: bottlenecks, gene flow, etc.
- Conservation: historical genetic diversity of endangered species

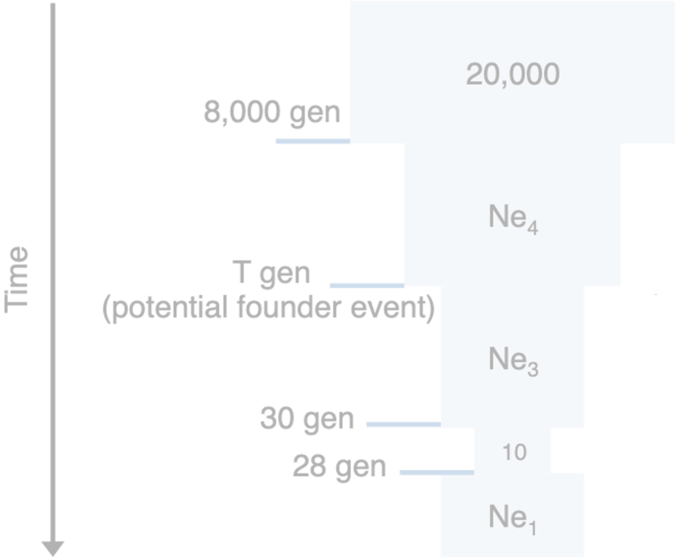


Robinson et al. (2016) *Current Biology*

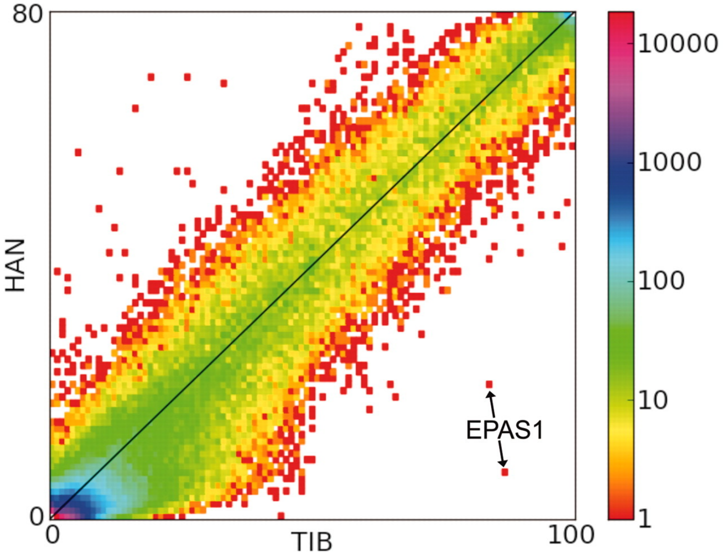
Demographic history inference – Motivation

National Geographic

- Understand population history: bottlenecks, gene flow, etc.
- Conservation: historical genetic diversity of endangered species
- Selection: sets neutral background for detecting adaptation



Robinson et al. (2016) *Current Biology*

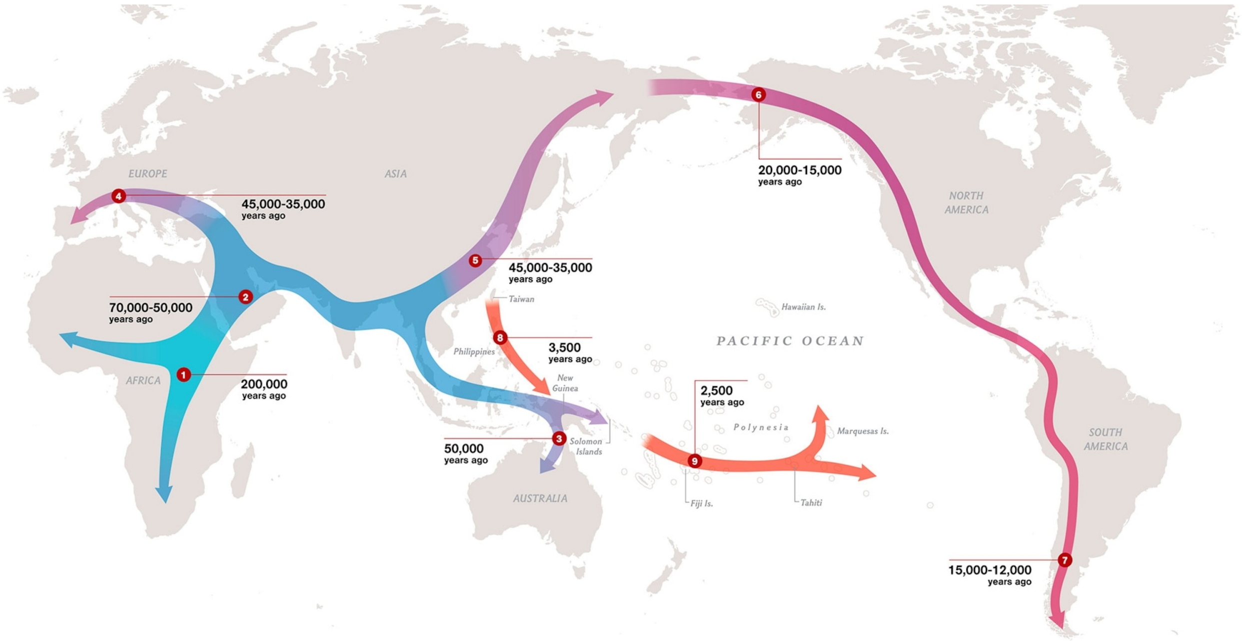


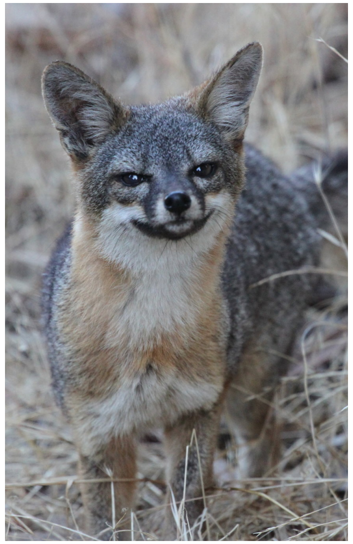
Yi et al. (2010) *Science*

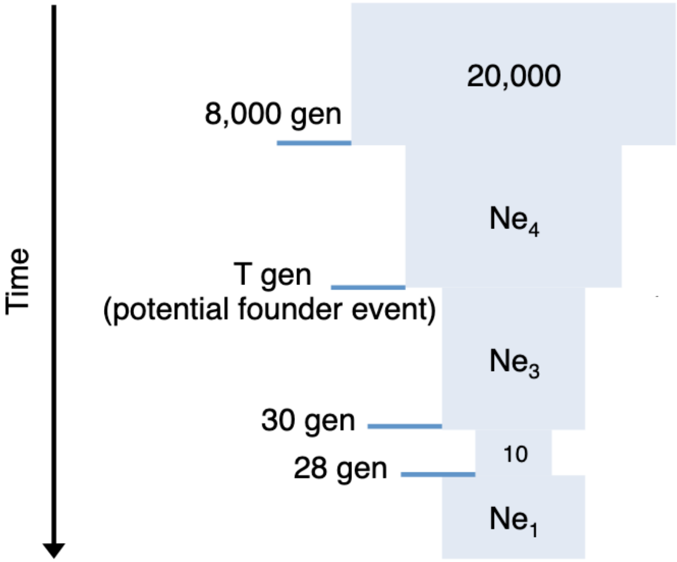
Demographic history inference – Motivation

- Understand population history: bottlenecks, gene flow, etc.
- Conservation: historical genetic diversity of endangered species
- Selection: sets neutral background for detecting adaptation

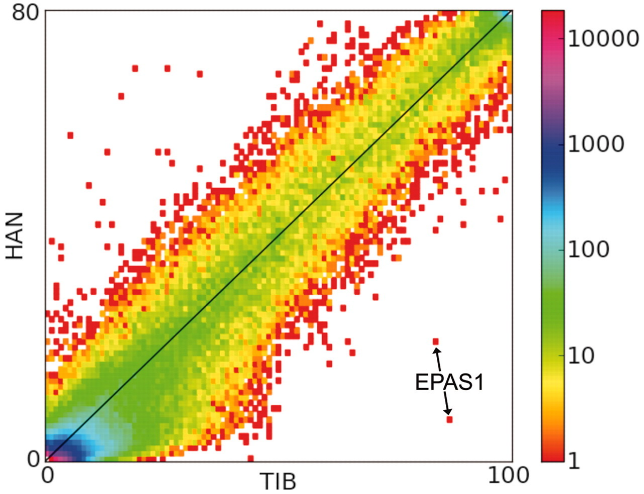
National Geographic







Robinson et al. (2016) *Current Biology*



Yi et al. (2010) *Science*

Diffusion approximation for demographic inference (dadi)

PLOS GENETICS

BROWSE PUBLISH ABOUT

SEARCH

advanced search

OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data

Ryan N. Gutenkunst, Ryan D. Hernandez, Scott H. Williamson, Carlos D. Bustamante

Published: October 23, 2009 • https://doi.org/10.1371/journal.pgen.1000695

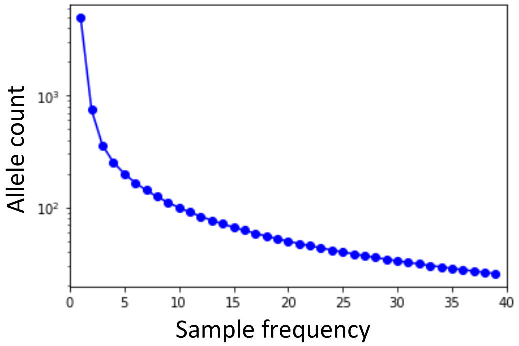
Groups

☆ dadi-user 391 members

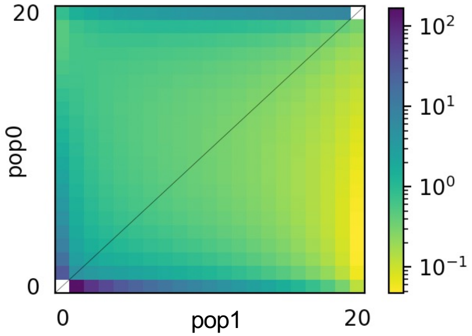
Diffusion approximation for demographic inference (dadi)

Data

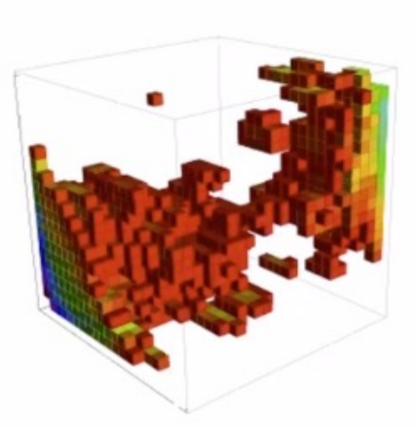
Data Summary



1D – 1 population



2D – 2 populations

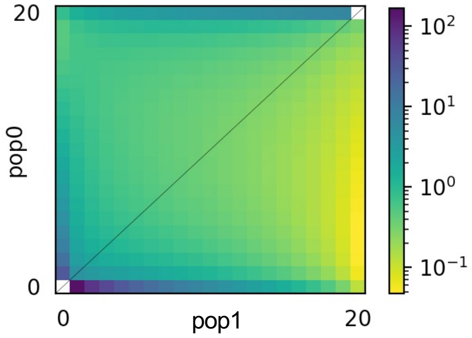


3D – 3 populations

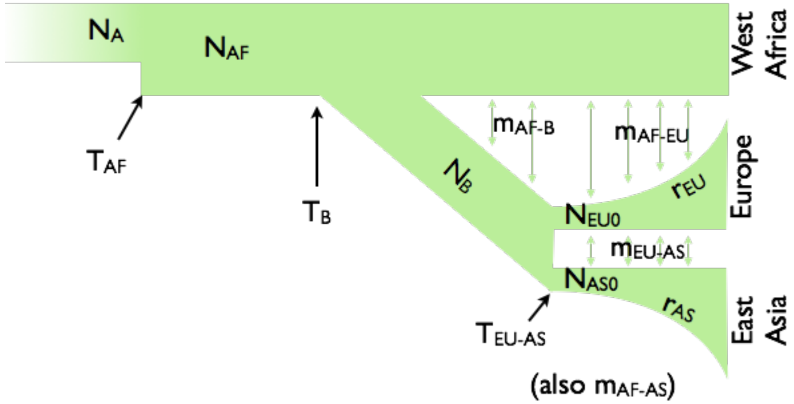
Diffusion approximation for demographic inference (dadi)

Data

Data Summary



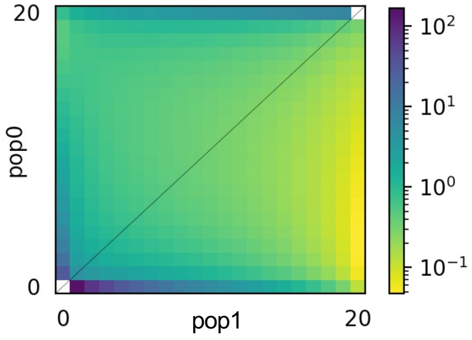
Input parameterized model

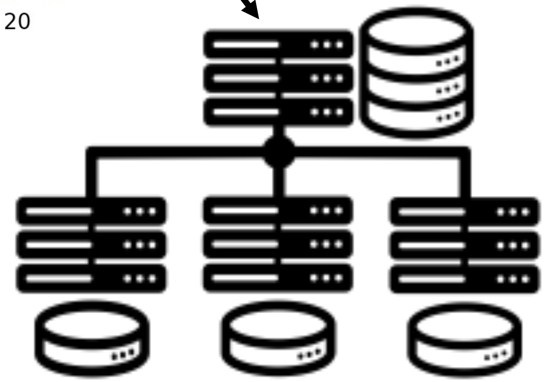


Diffusion approximation for demographic inference (dadi)

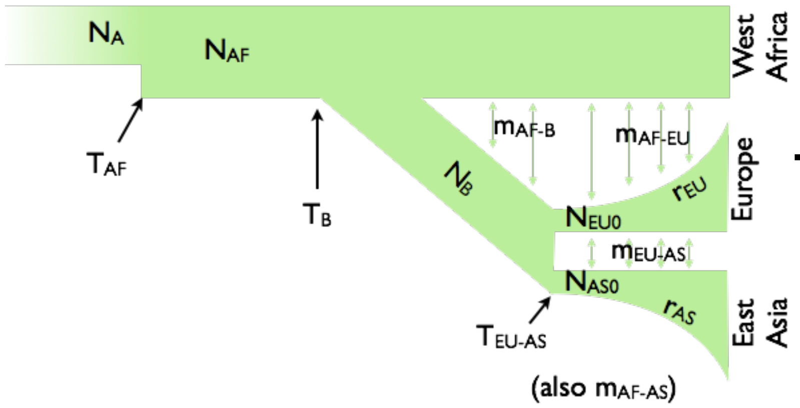
Data

Data Summary





Input parameterized model

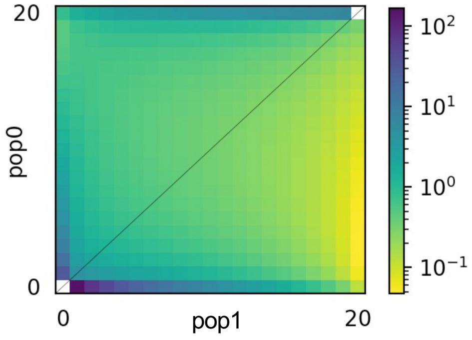


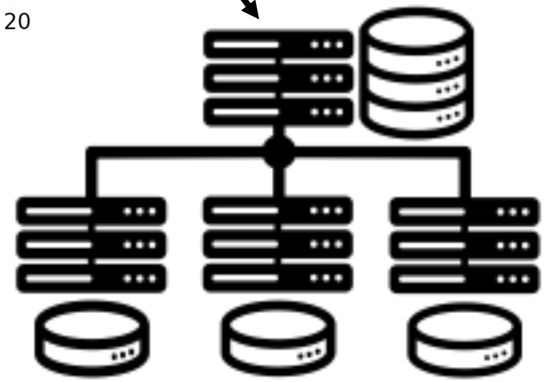
Likelihood Optimization

Diffusion approximation for demographic inference (dadi)

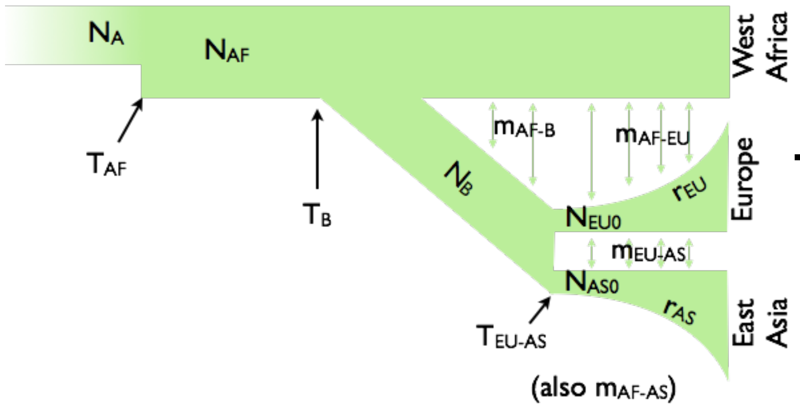
Data

Data Summary



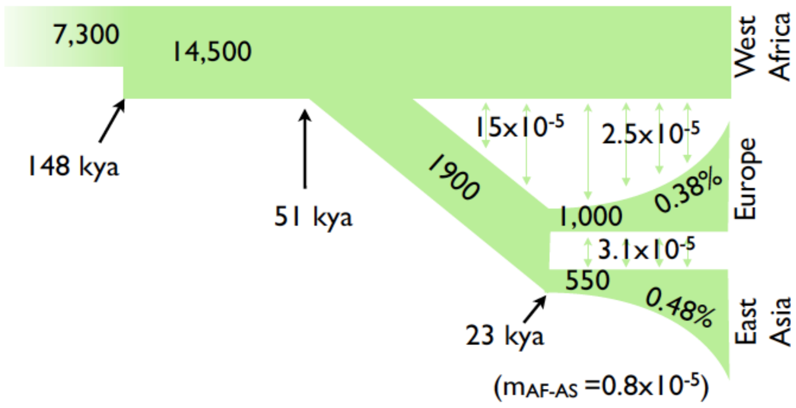


Input parameterized model

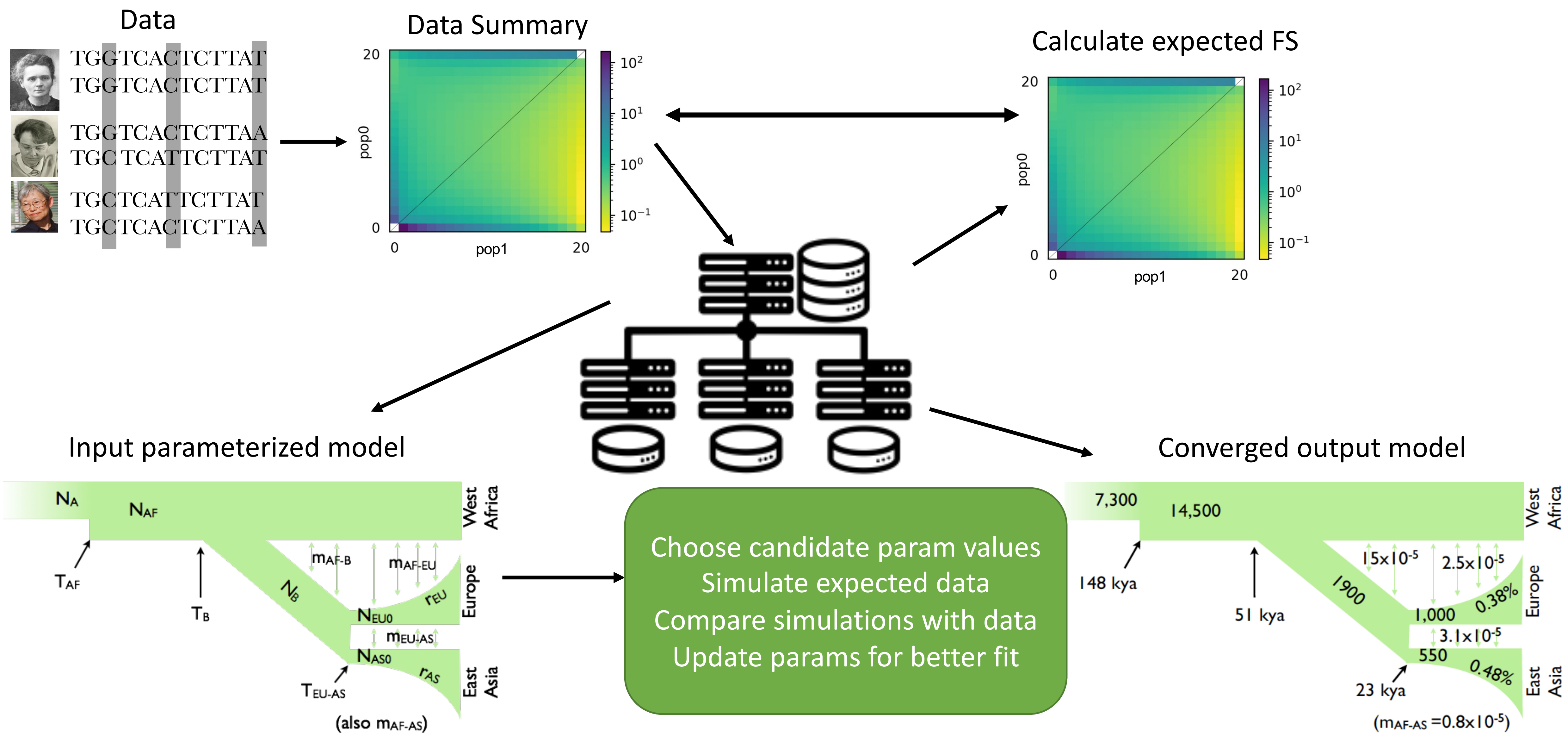


Likelihood Optimization

Converged output model



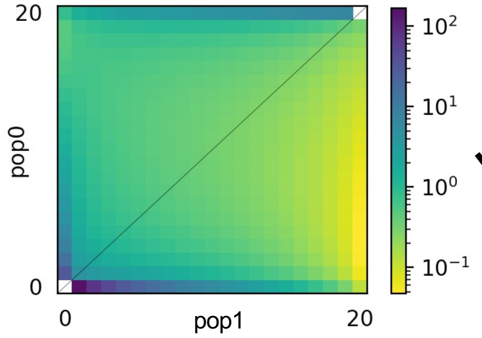
Diffusion approximation for demographic inference (dadi)



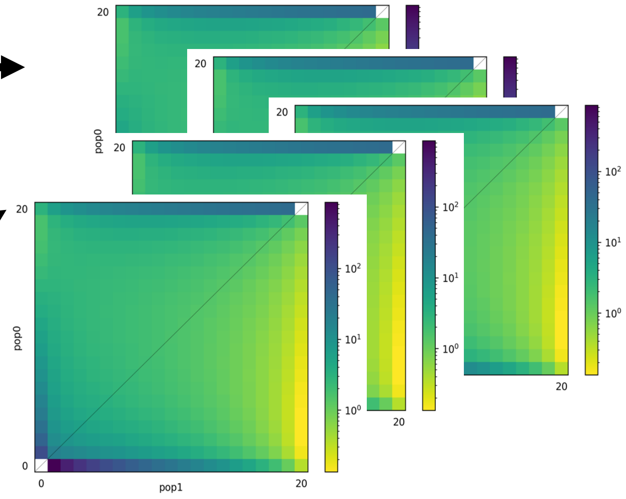
Diffusion approximation for demographic inference (dadi)

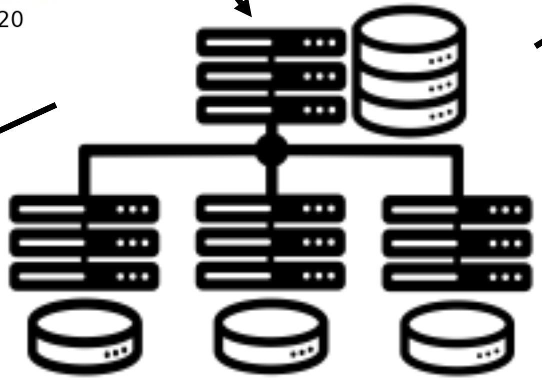
Data

Data Summary

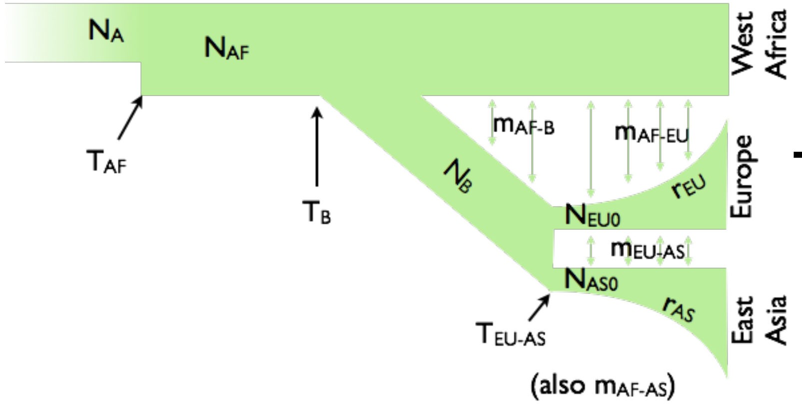


Calculate many expected FS





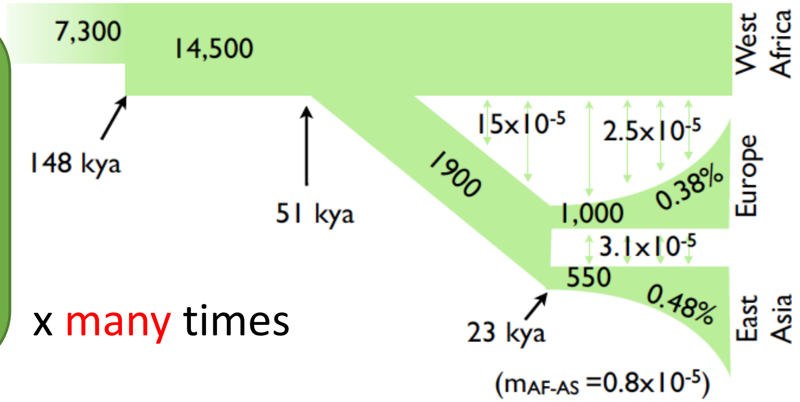
Input parameterized model



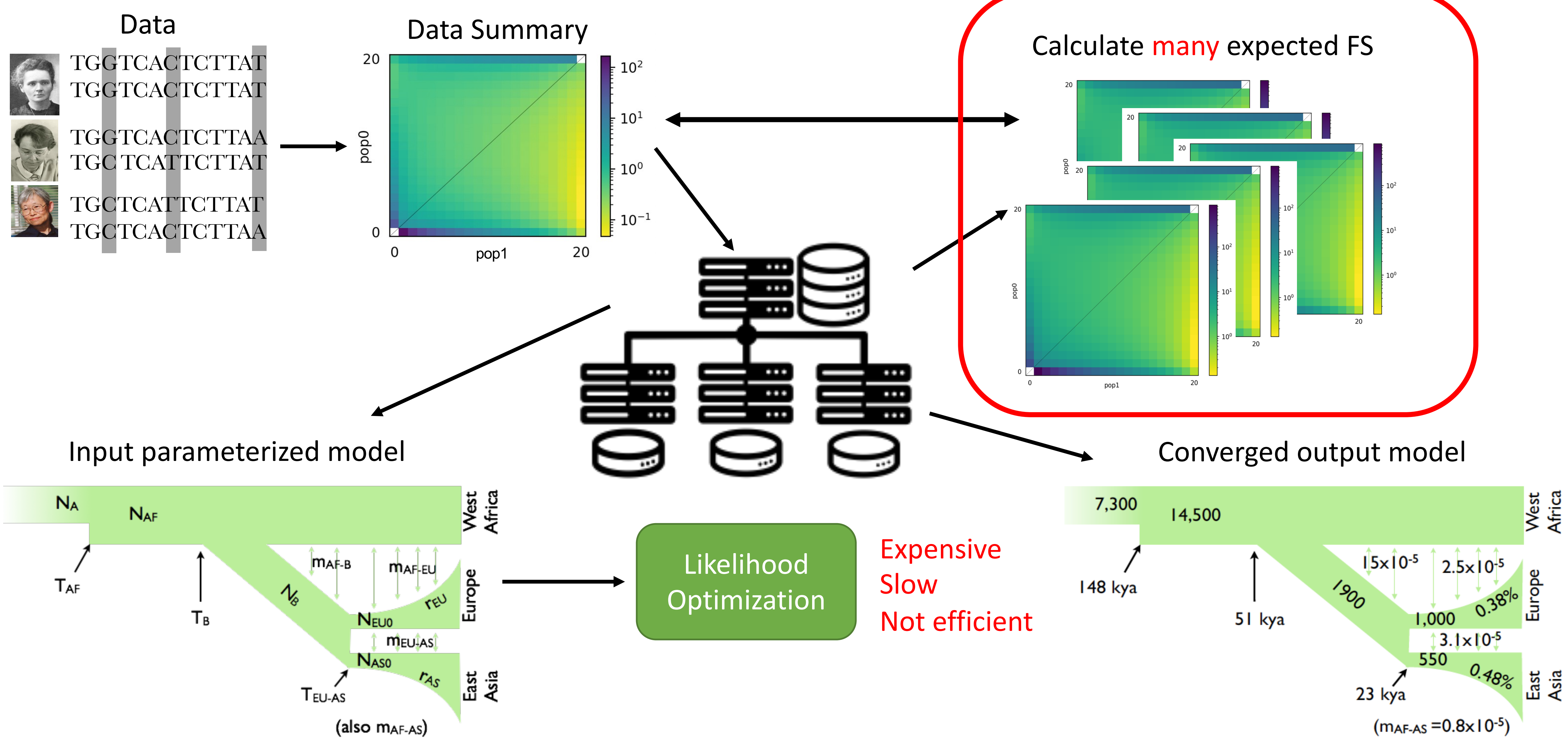
Choose candidate param values
Simulate expected data
Compare simulations with data
Update params for better fit

x many times

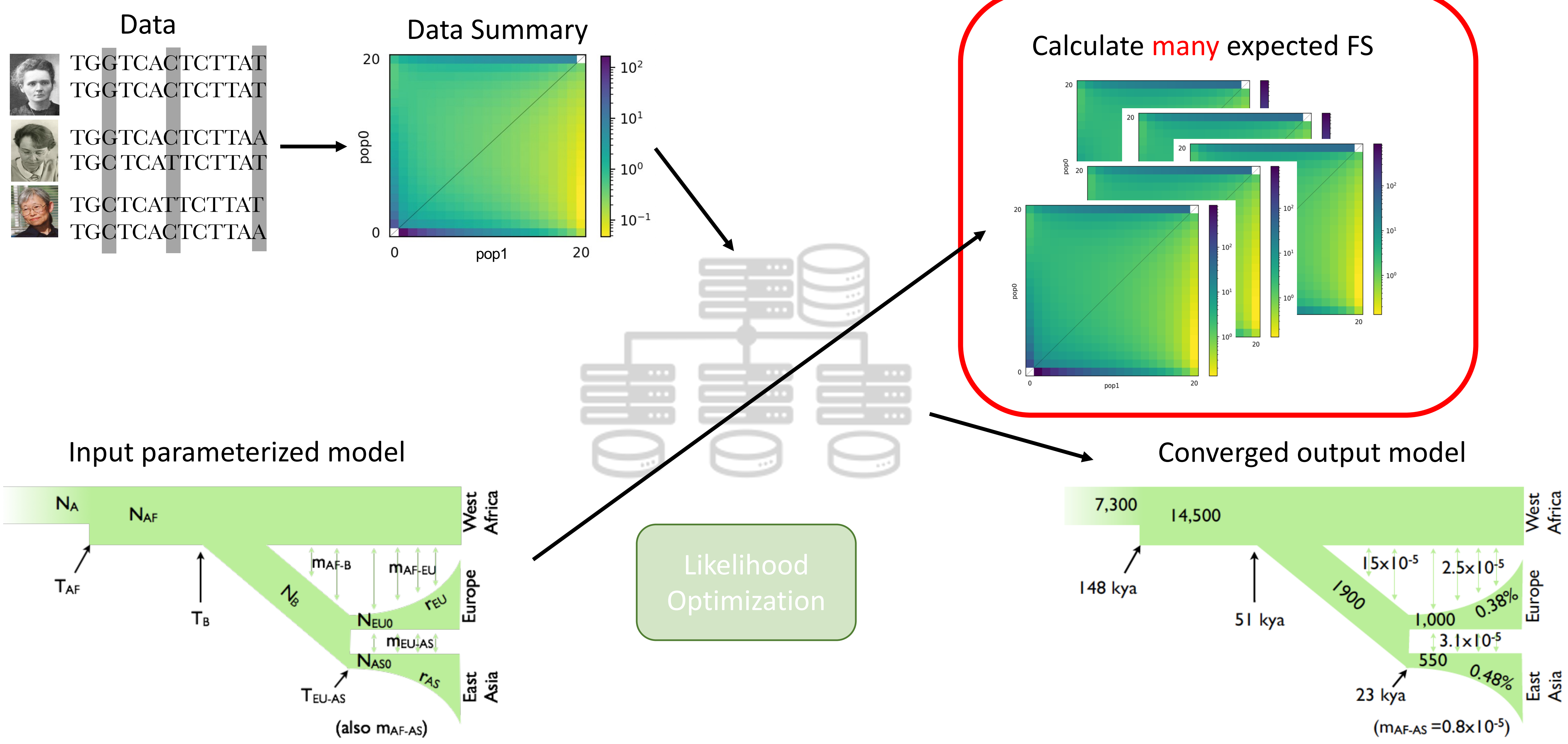
Converged output model



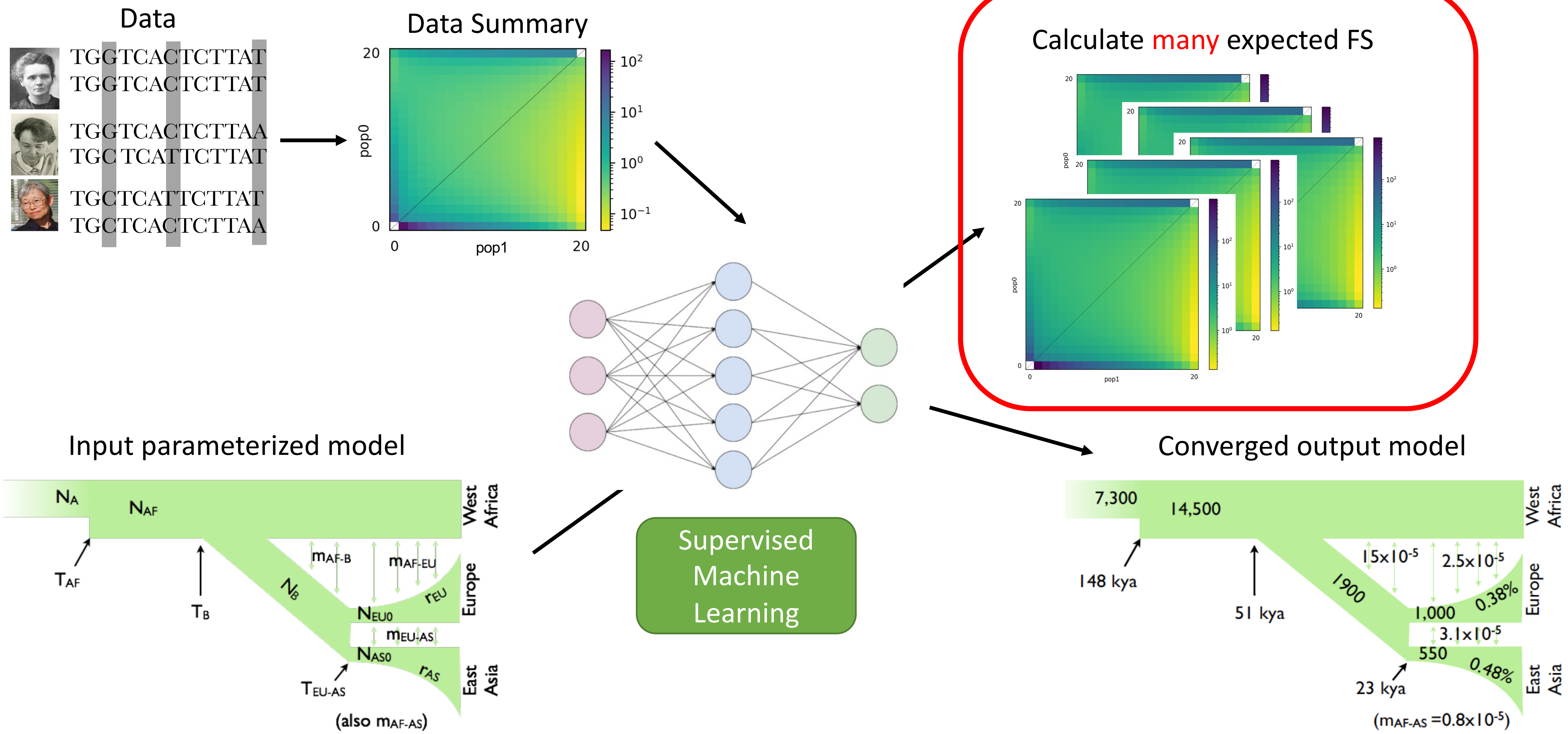
Diffusion approximation for demographic inference (dadi)



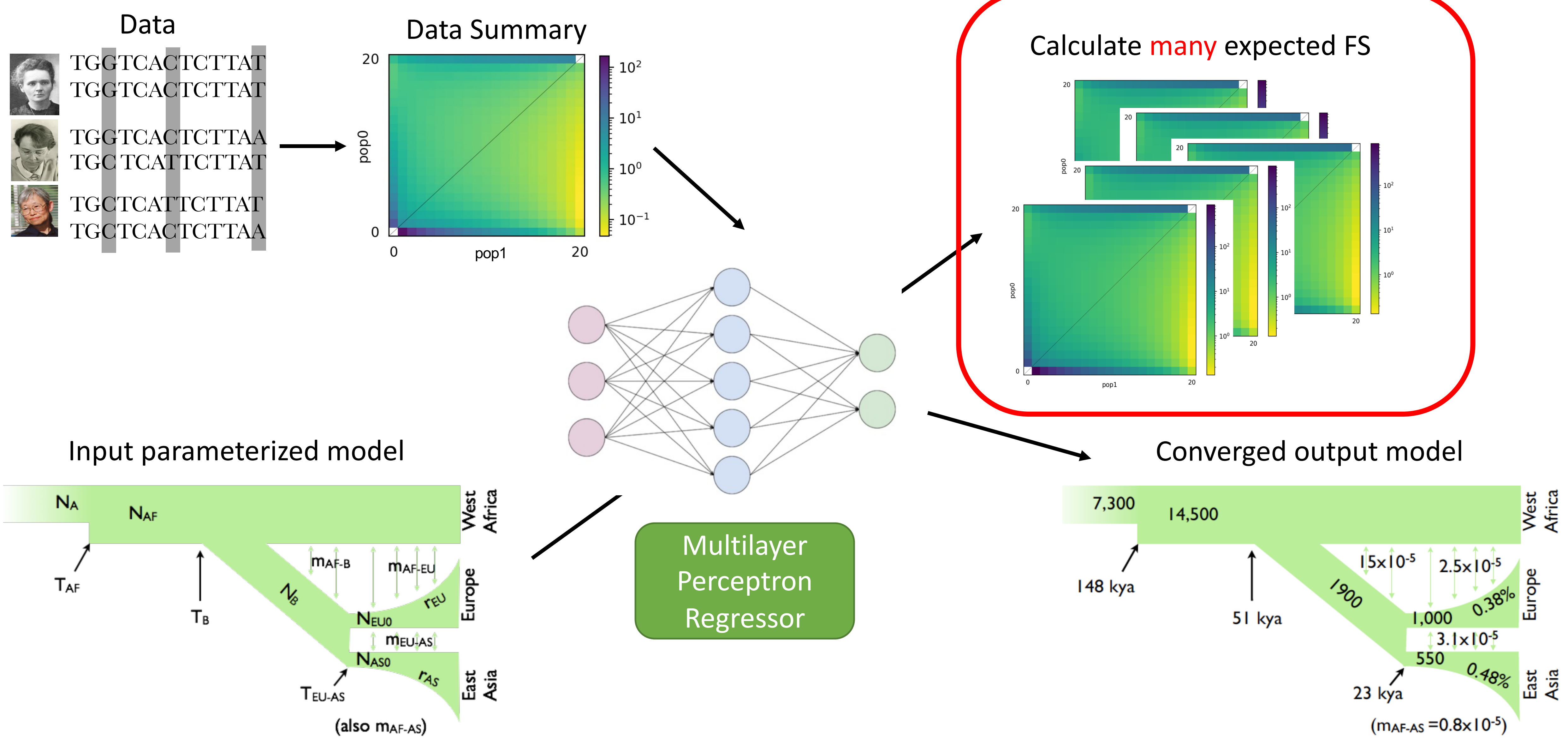
Diffusion approximation for demographic inference (dadi)



Diffusion approximation for demographic inference (dadi)



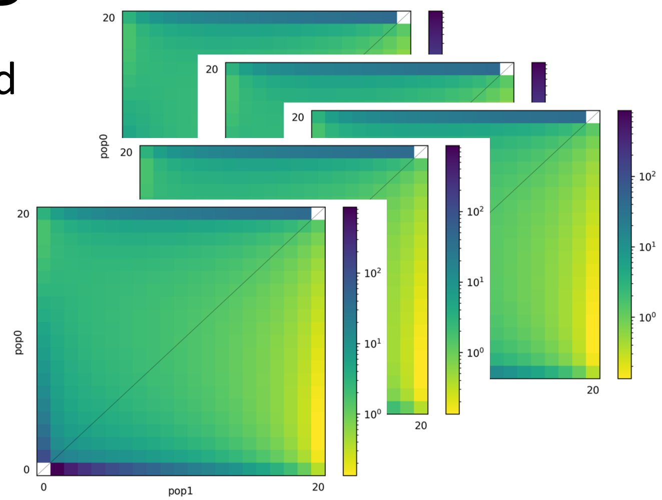
Diffusion approximation for demographic inference (dadi)



dadi-machine learning workflow

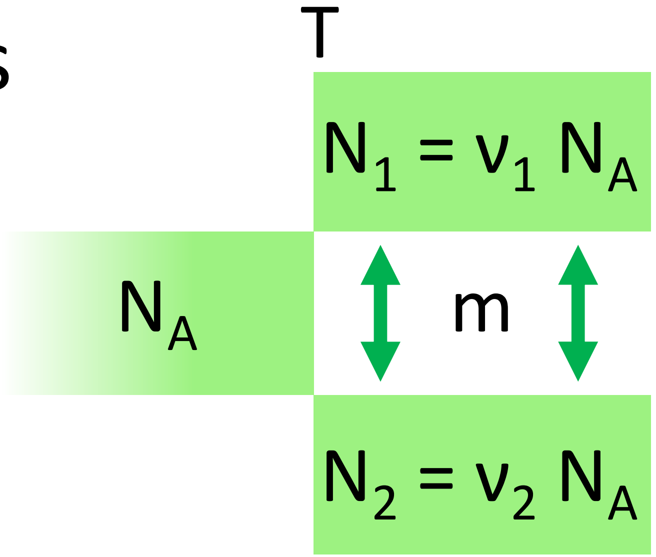
Training data

dadi-simulated spectra (1000)



Labels

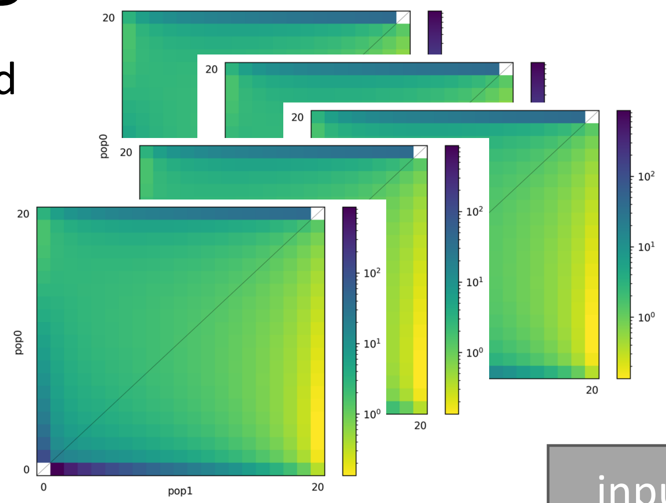
parameter values

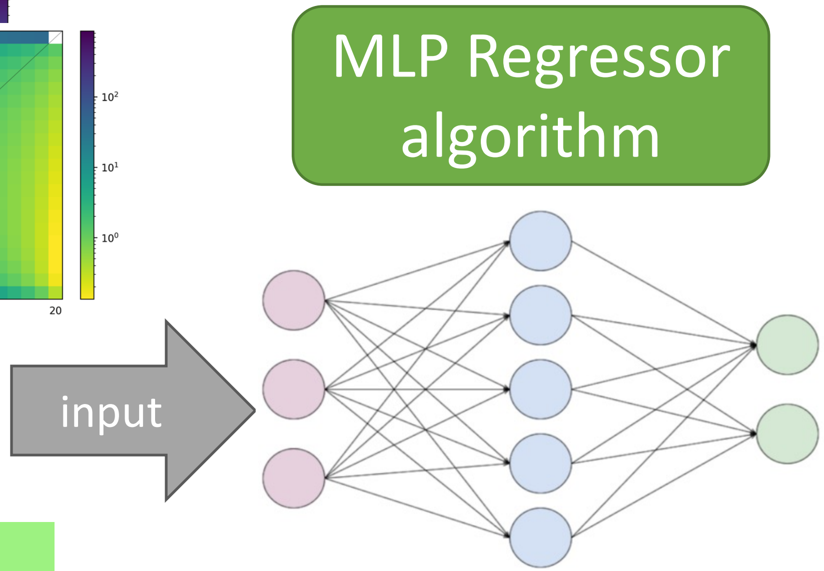


dadi-machine learning workflow

Training data

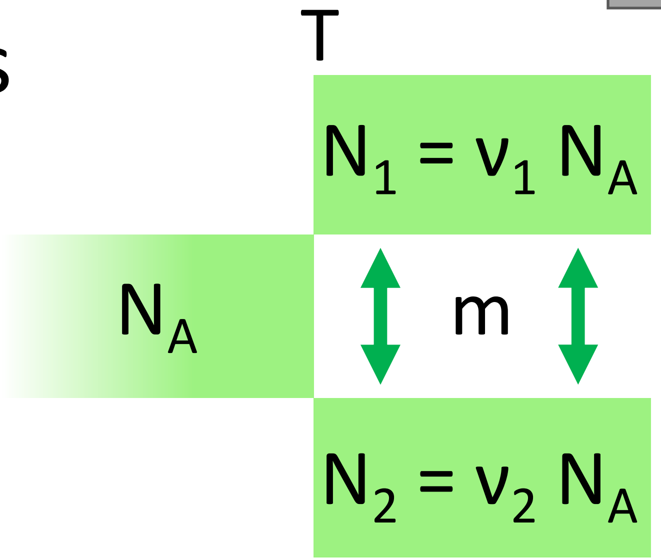
dadi-simulated spectra (1000)





Labels

parameter values

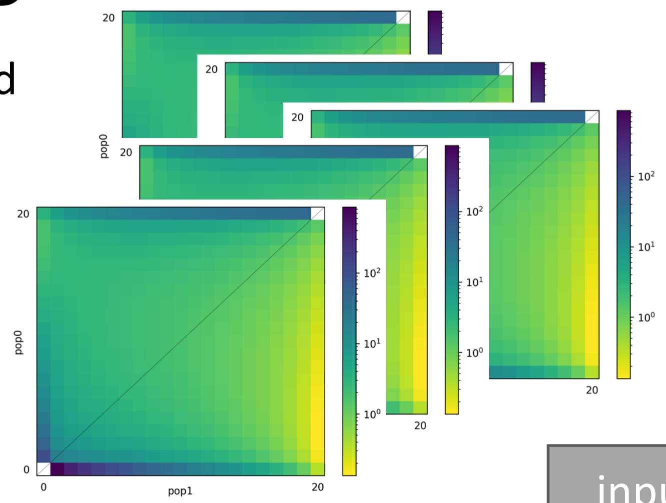


Learns:
Finds pattern of association

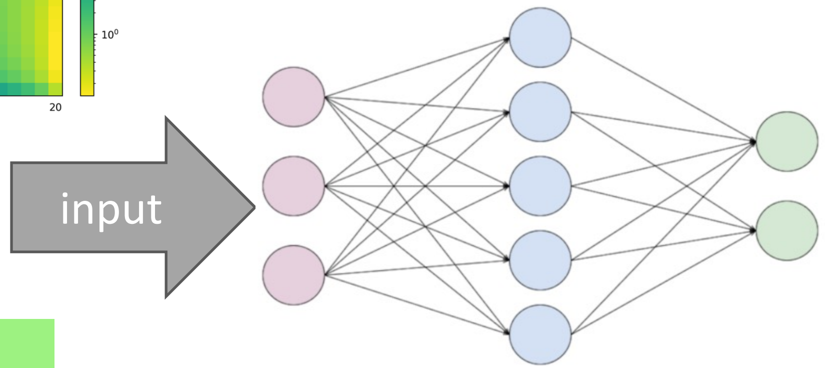
dadi-machine learning workflow

Training data

dadi-simulated spectra (1000)

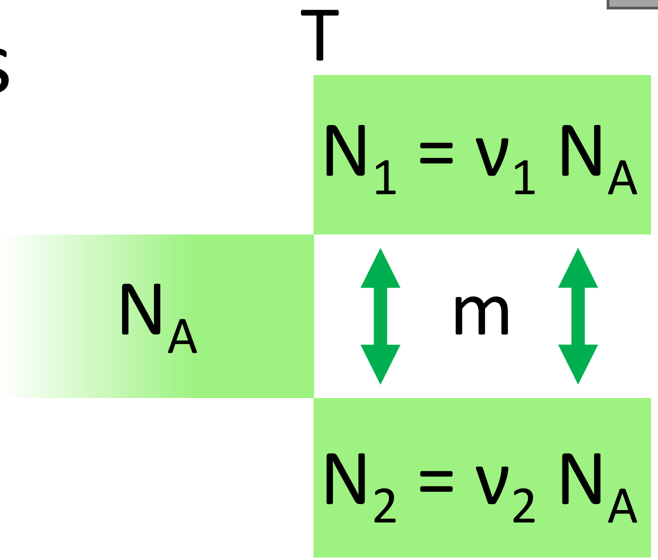


Trained MLP Regressor



Labels

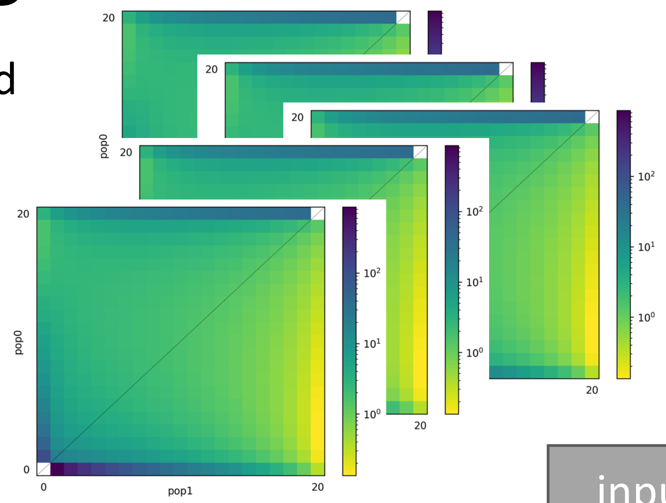
parameter values

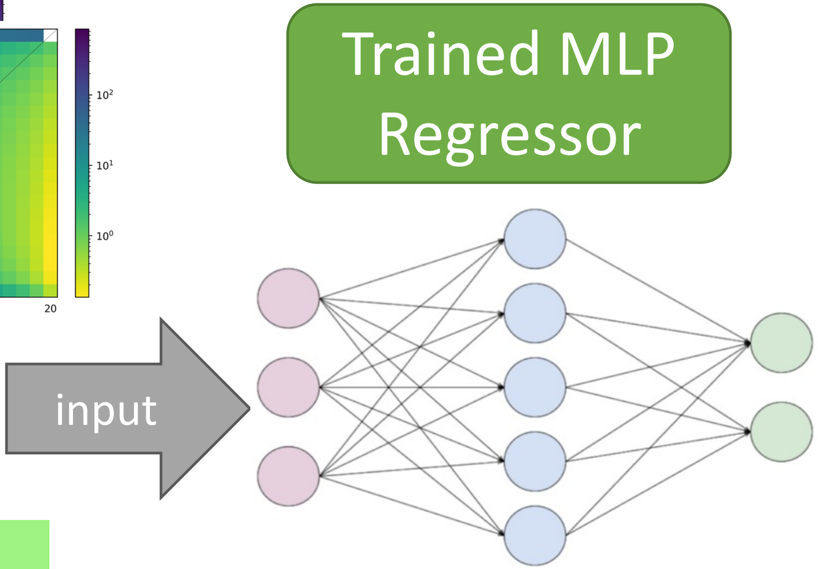


dadi-machine learning workflow

Training data

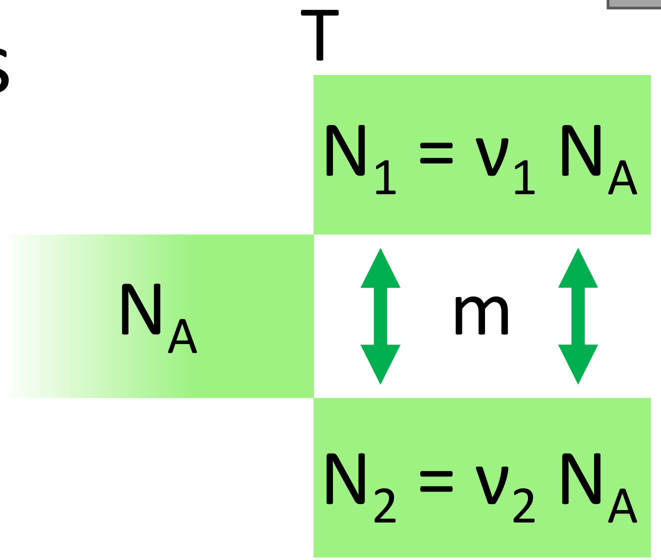
dadi-simulated spectra (1000)





Labels

parameter values



Strength of frequency spectra simulation with dadi:

- Faster and less expensive than coalescent simulations
- Organism-agnostic and transferability

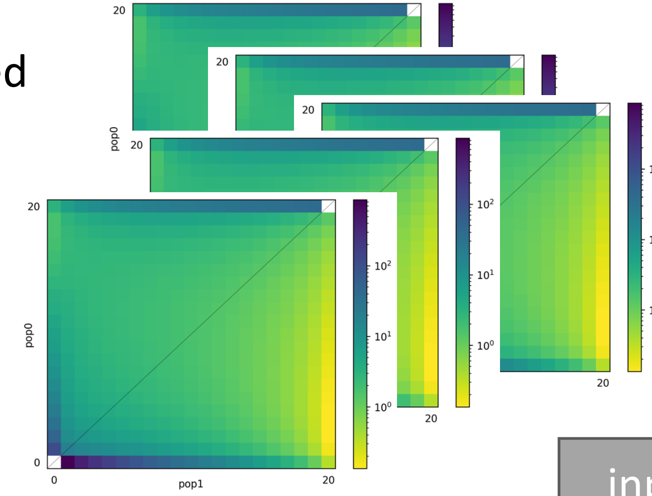
Weakness:

- Ignores linkage

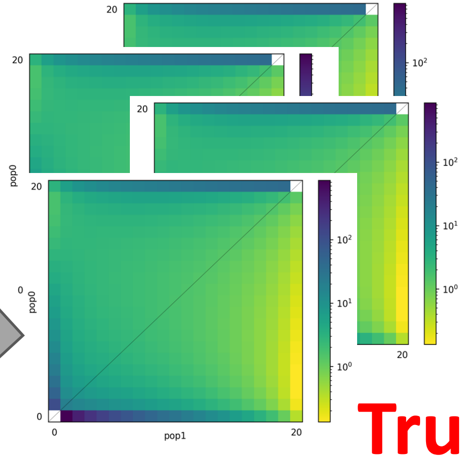
dadi-machine learning workflow

Training data

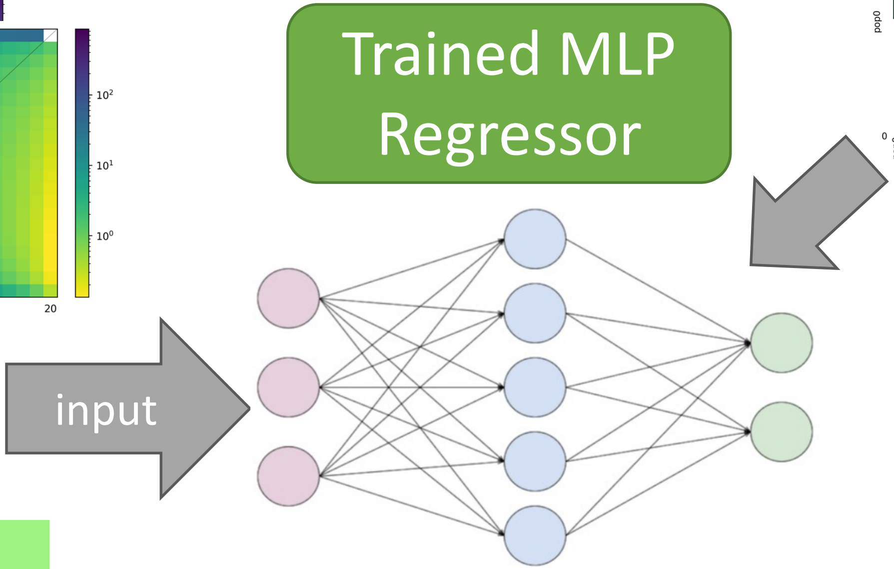
dadi-simulated spectra (1000)



Test data



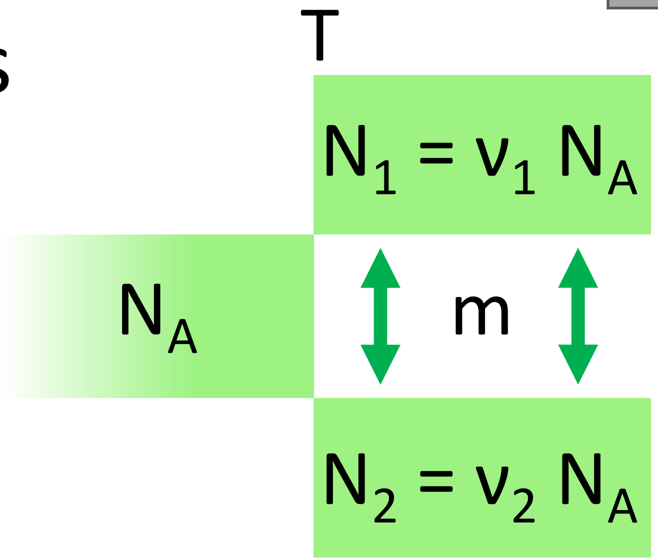
dadi- and msprime-simulated spectra (100-200)



True params: set aside for comparison

Labels

parameter values



Strength of frequency spectra simulation with dadi:

- Faster and less expensive than coalescent simulations
- Organism-agnostic and transferability

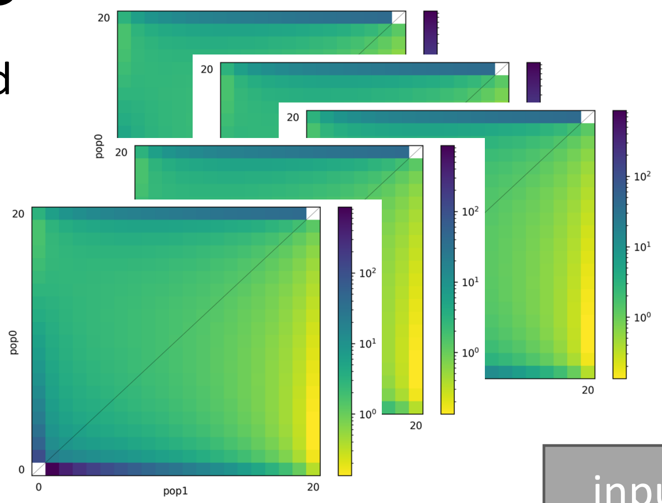
Weakness:

- Ignores linkage

dadi-machine learning workflow

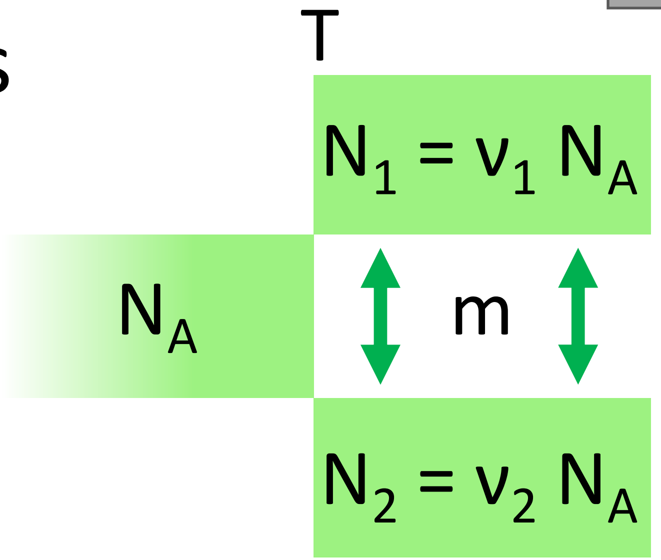
Training data

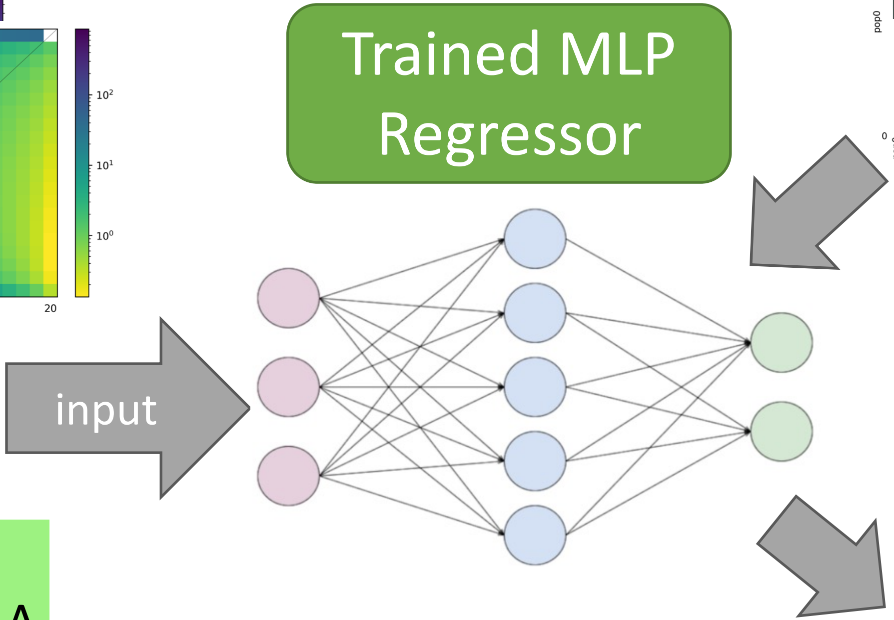
dadi-simulated spectra (1000)



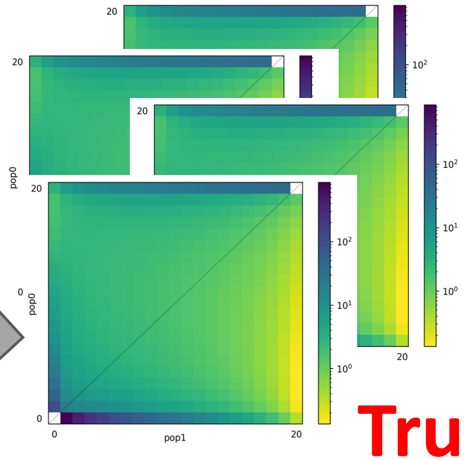
Labels

parameter values





Test data



dadi- and msprime-simulated spectra (100-200)

True params: set aside for comparison

Prediction

param values for each test spectrum

Predict :

$\nu_1 = 0.1$

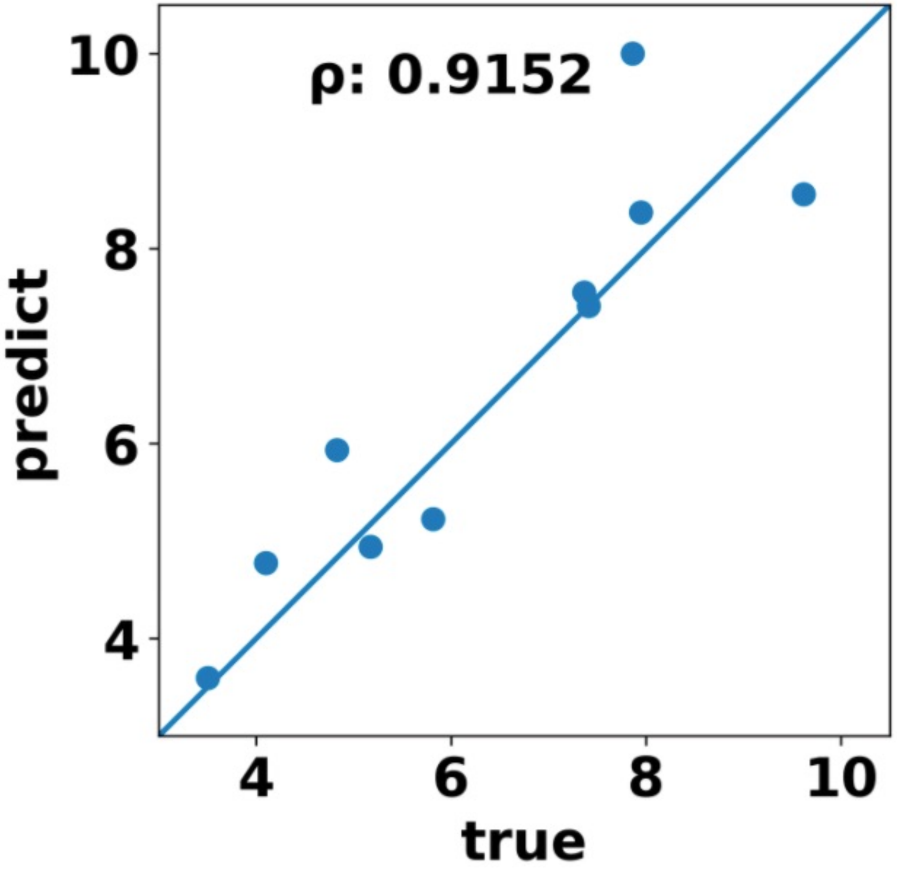
$\nu_2 = 0.5$

$T = 1.3$

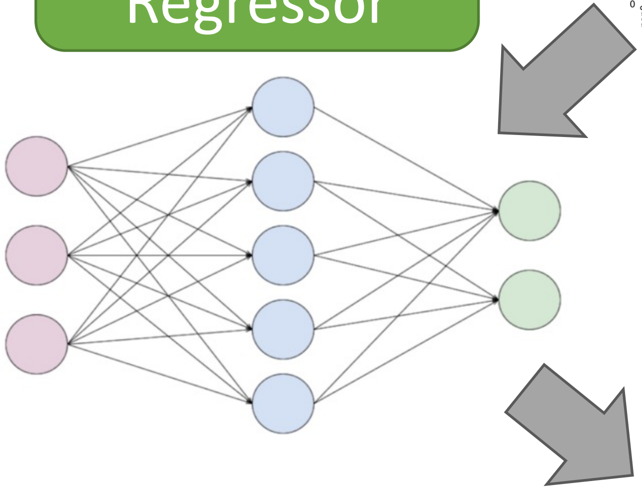
$m = 5$

Testing and validation

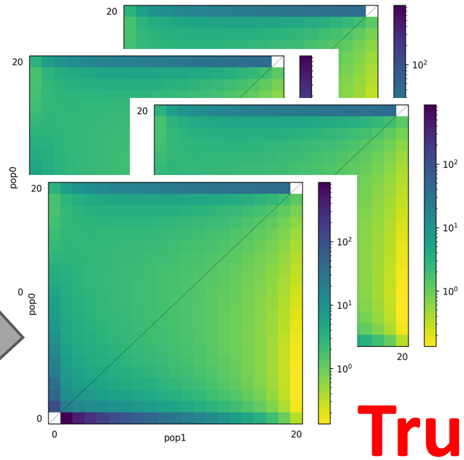
Prediction accuracy



Trained MLP Regressor



Test data



dadi- and msprime-simulated spectra (100-200)

True params: set aside for comparison

Prediction

param values for each test spectrum

Predict :

$\nu_1 = 0.1$

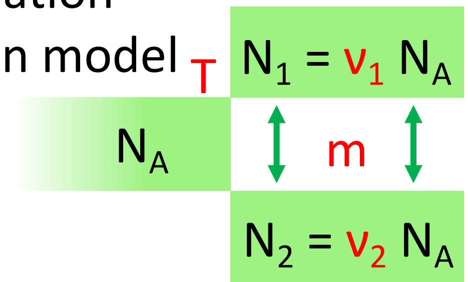
$\nu_2 = 0.5$

$T = 1.3$

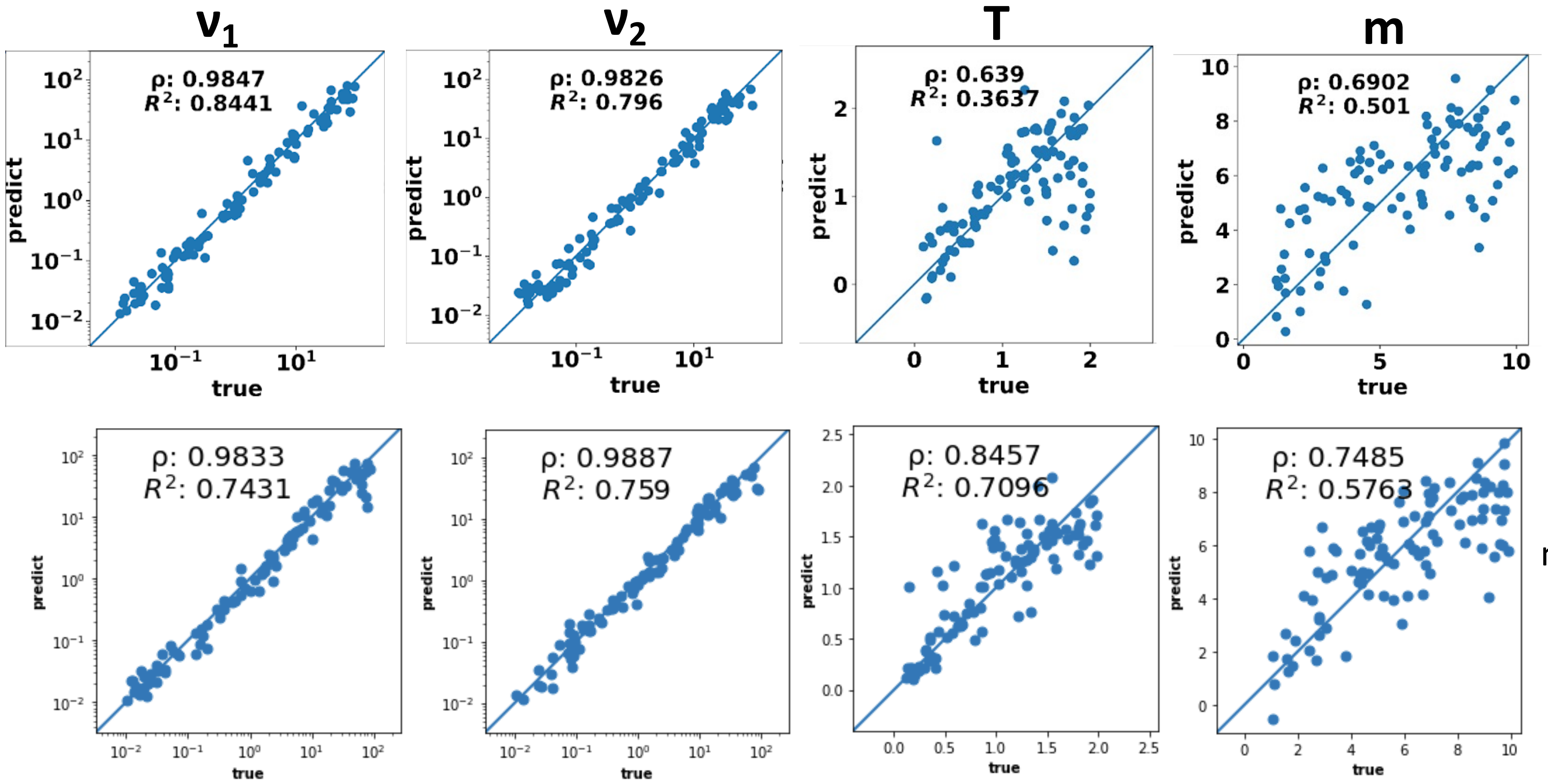
$m = 5$

Results – prediction accuracy

two-population
split-migration model



MLPR

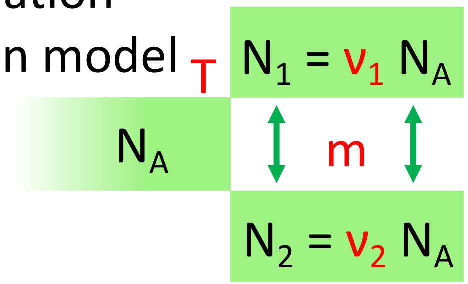


dadi-simulated
data sets

msprime-simulated
data sets

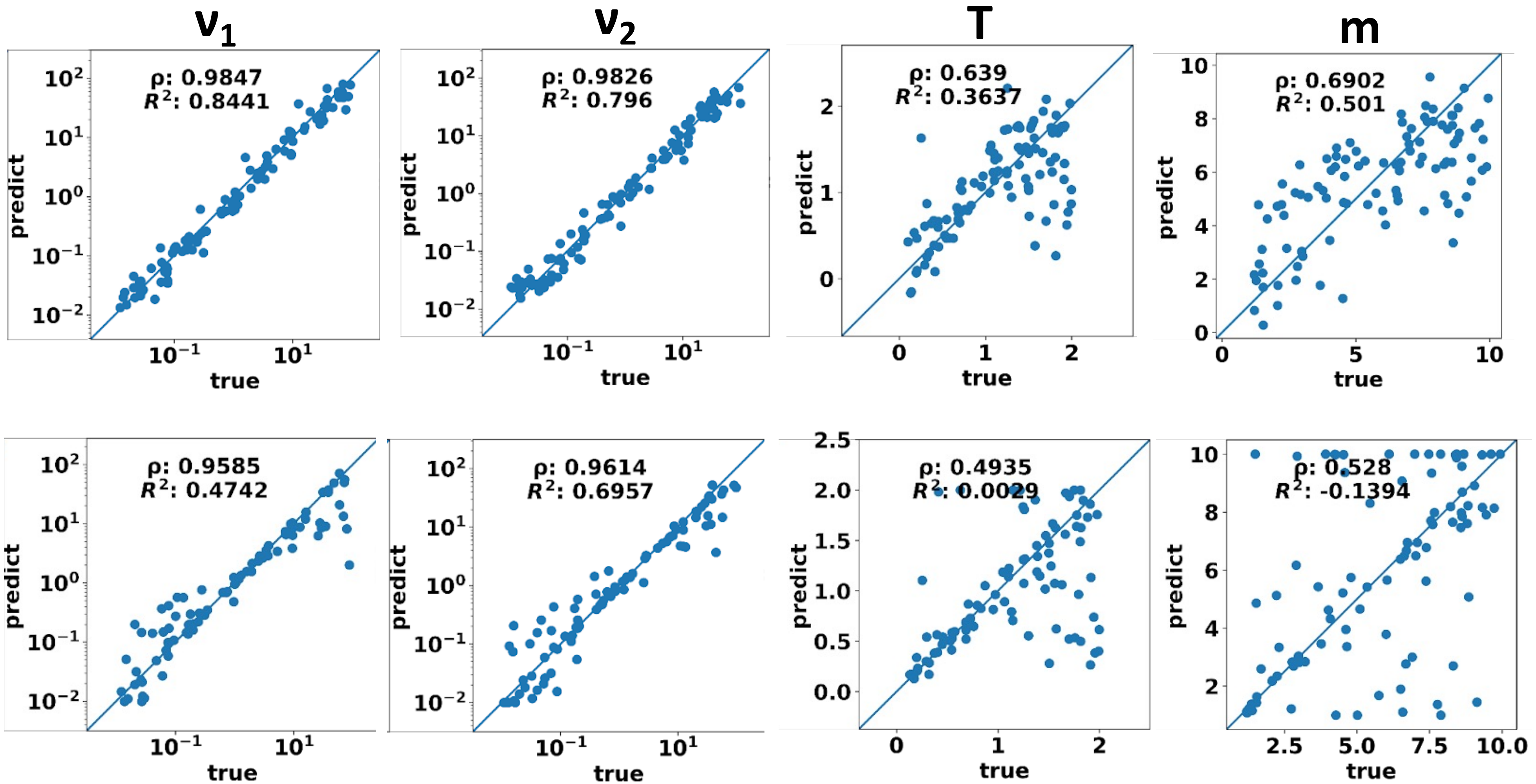
Results – prediction accuracy

two-population
split-migration model



MLPR

dadi

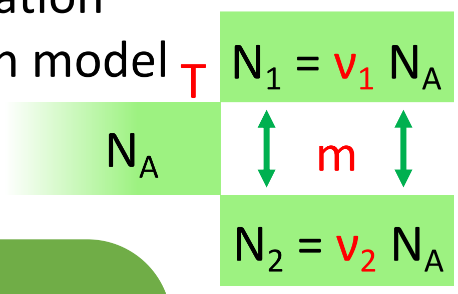


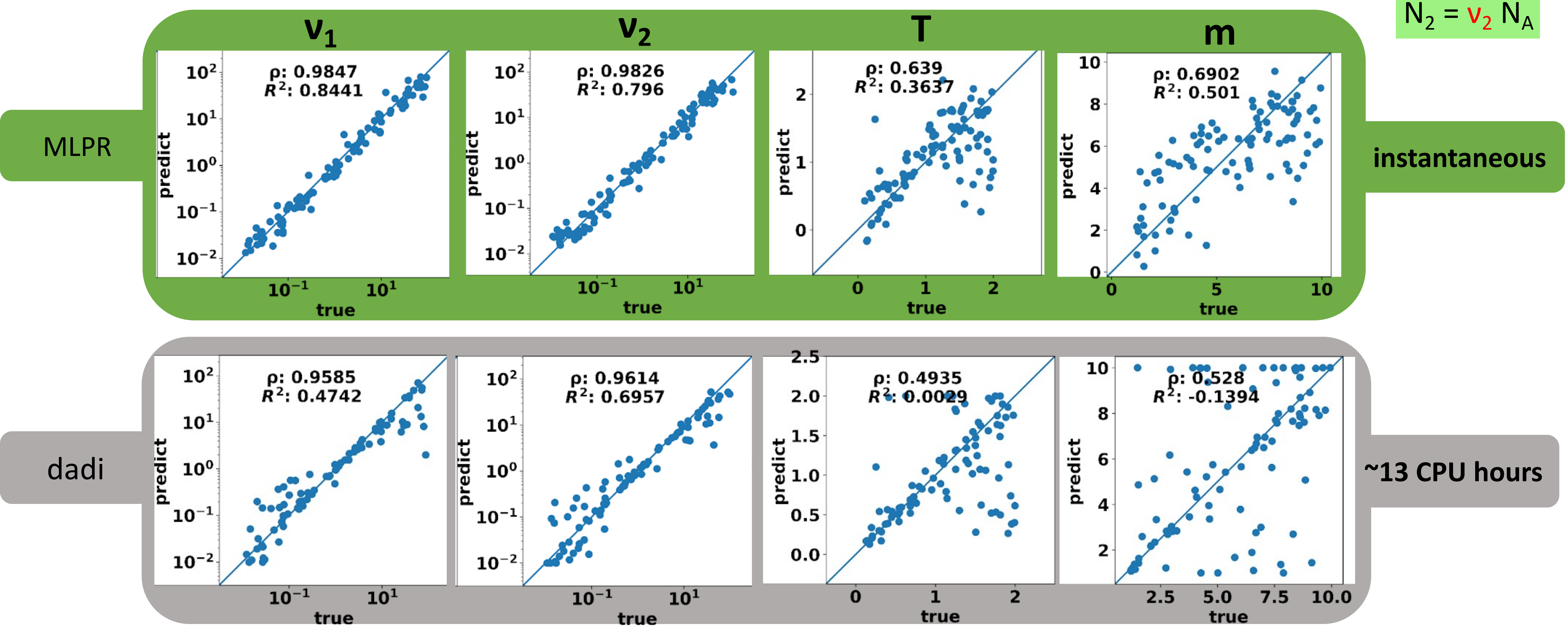
dadi-simulated data sets

dadi-simulated data sets

Results – prediction accuracy and computational efficiency

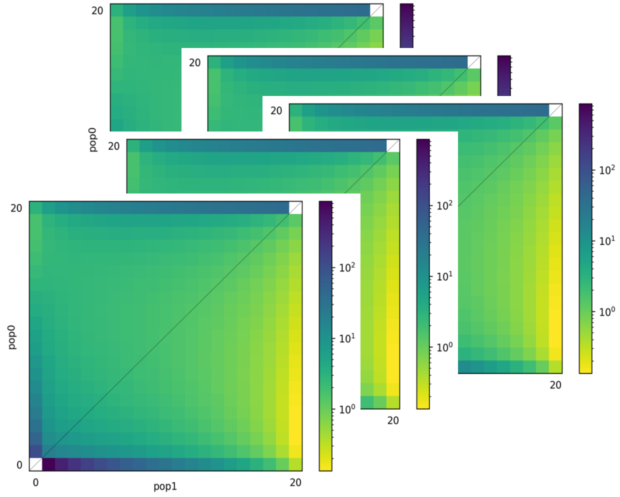
two-population split-migration model

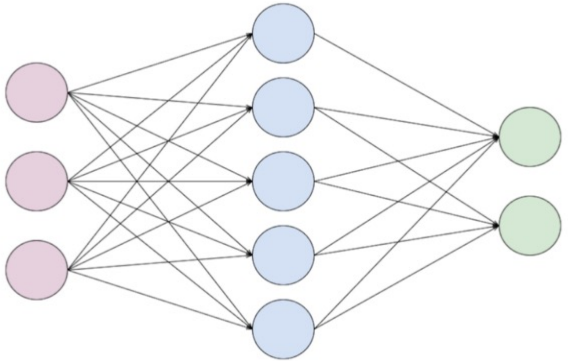


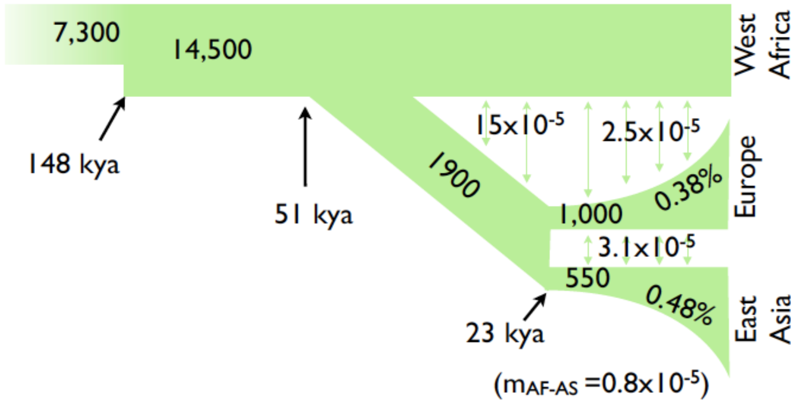


Summary

- Machine learning approach to improve the computational efficiency of a widely used likelihood-based demographic inference method
- Comparable accuracy with significantly reduced computational cost & complexity
- Robust to test simulations generated with linkage
- Accompanying uncertainty quantification method that provides prediction intervals
- Trained MLPRs for frequently used demographic models and workflow will be distributed







@LnTran26 http://gutengroup.arizona.edu