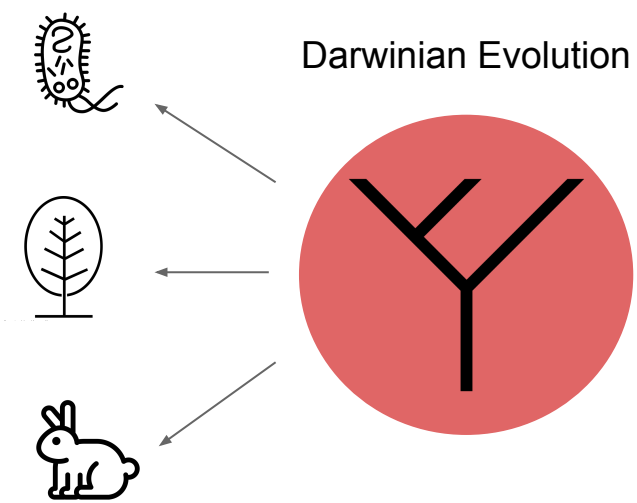
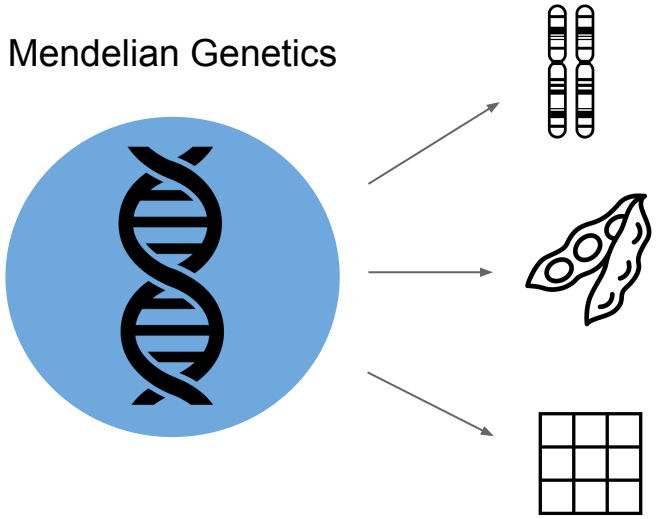
Review: *Deep Learning for Population Genetic Inference*

(Sheehan & Song)

Presented by Kartik Sawhney, Patrick O'Grady, Rishab Mehra, Thomas Liu & Alexandra Bourdillon

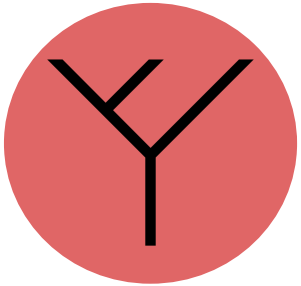
Population Genetics





Population Genetics

Darwinian Evolution



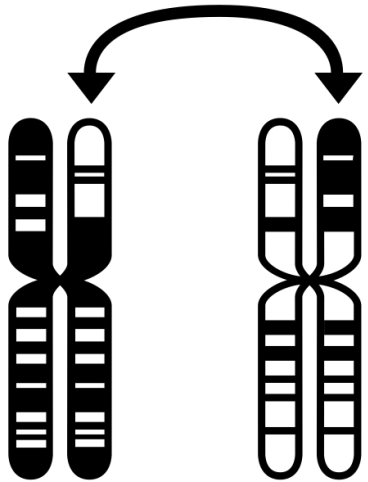
The Hardy-Weinberg Principle

Evolution stagnates when...

- No mutations
- Random mating only
- No immigration/emigration
- No natural selection factors
- Large population

Mendelian Genetics

Statistical Inference



- Genealogy provides insight about evolutionary + demographic forces
- Evolutionary forces include:
  - Mutation, recombination, types of selection
- Demographic forces include:
  - Changes in population size, migration, isolation

Our History

ancestry.com

Types of Positive Selection

- **Hard Sweep:** A beneficial mutation
  - Rare but frequency increases rapidly from single copy
- **Soft Sweep:** A previously neutral mutation becomes favorable due to environmental change
  - Multiple copies contribute to genetic spike
- **Balancing Selection:** Higher than expected ratio of gene is maintained
  - i.e. due to heterozygote advantage
  - Allows for genetic polymorphism
- Researchers also used * **Neutral Selection**

Core Challenge

- **Joint inference of natural selection and demography** (in the form of a population size change history)
- Separate the **global nature of demography** from the **local nature of genetic selection** without needing sequential / distinct steps

Core Challenge

- **Joint inference of natural selection and demography** (in the form of a population size change history)
- Separate the **global nature of demography** from the **local nature of genetic selection** without needing sequential / distinct steps

- **Computational & theoretical constraints** due to complex population genetic models
- Relatively unexplored applications of DL in population genomics
- Demography and selection can **leave similar genomic signals**.
- Lack of existing research focus on joint inference of these two factors

Current Landscape

- **Existing machine learning focus on selective forces only**; particularly on classifying genome into neutral vs. selected regions
- Methods highlighting robustness to various demographic scenarios, not inferring anything about selection or demography
- Implicit **assumption about the irrelevance of selection** on demography

Core Challenge

African *Drosophila melanogaster*

Core Challenge

African *Drosophila melanogaster*

- **Pervasive selection** of their genome **confounds demographic analysis; vice versa**
- Existing literature and research of their demography to serve as benchmark and basis of comparison for deep learning models

Methods

Approximate Bayesian Computation (ABC)

Deep Learning

Approximate Bayesian Computation (ABC)

- Simulate many datasets under a prior for the desired parameters of interest
- Reduce datasets to vector of summary statistics
- Find closest summary statistics to that of target dataset
- Use corresponding parameters to find the posterior distribution

Advantages and Disadvantages of ABC

Advantages

- Easy to use
- Returns a posterior distribution

Disadvantages

- Curse of dimensionality
- Uses rejection algorithm
- Challenging to use for a combination of continuous parameters and categorical distributions

Deep Learning

Why Deep Learning?

- Makes full use of datasets
- Handles correlation between summary statistics
- Produces interpretable features

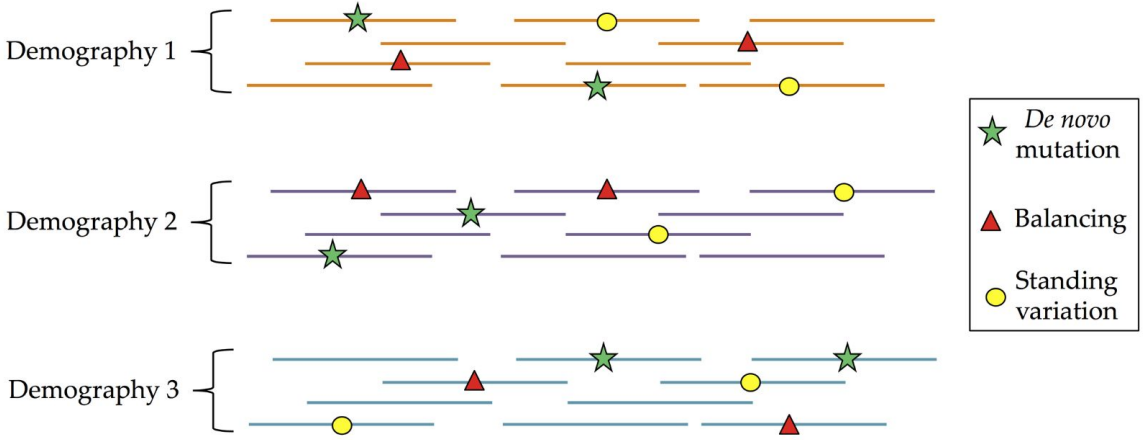
Dataset

3 Population Sizes:

- Recent Effective Population (N1)
- Bottleneck (N2)
- Ancient (N3)

4 Selection Classes:

- No Selection (Neutral)
- Positive Directional Selection (Hard Sweep)
- Selection on standing Variation (Soft Sweep)
- Balancing Selection



Simulations

- Used msms to simulate demographic histories
- Downsampled the Zambia Drosophila Melnogaster dataset
- Repeated 2500 times
- Sampled 160 regions for each demographic history
- This led to a total of 400,000 datasets

Pre-training using autoencoders

Why?

- Random initialization leads to falling into local minimas

Details

- Activation Function: Logistic Function
- Sparsity Constraint
- Used the learned weights as initial weights of the first layer

Training Details

Input: 5 statistics from each dataset

Output: 2 response variables

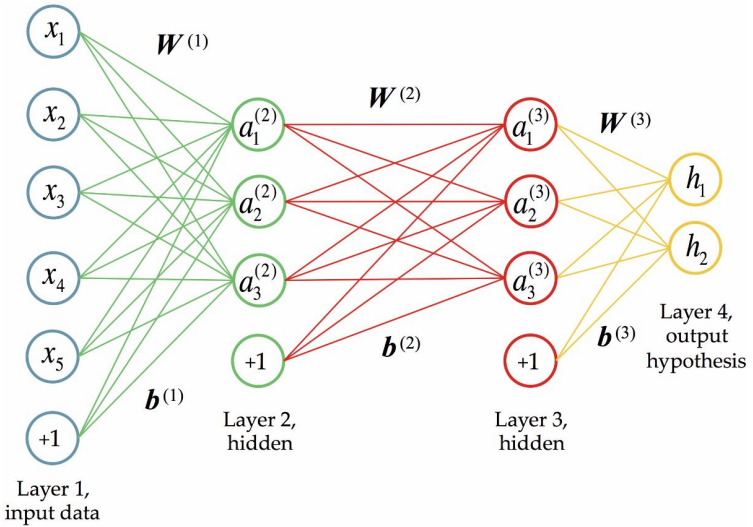
Layers: 4

Activation Function: Logistic Function

Loss: L2 squared

Output Layer: Linear activation function for population size and Softmax for selection

Optimization Technique: BFGS (Newton Method Variant)



Testing

Effective Population Sizes:

- Average stat: we average the statistics for each region within the dataset (corresponding to the same demography), then run this average through the trained network.
- Final: we run the statistics for each region through the trained network separately, then average the results.
- Neutral regions: we use the same method as Final, but instead of using all the regions, we only use the ones we predict as being neutral.

Selection class prediction:

We obtain a probability distribution over the classes for each region, then select the class with the highest probability.

Results

Dataset	Method	N_1 error	N_2 error	N_3 error
Full summary statistics	ABCtoolbox	0.062	0.043	0.218
	Deep learning	0.044	0.028	0.221
Filtered summary statistics	ABCtoolbox	0.161	0.035	0.311
	Deep learning	0.065	0.055	0.319

Deep learning predictions	N_1 error	N_2 error	N_3 error
Average stat prediction	0.051	0.074	0.487
Final prediction	0.098	0.077	0.569
Neutral regions prediction	0.072	0.083	0.566

Results (Cont.)

Initialization Type	N_1 error	N_2 error	N_3 error
Random	0.429	0.421	0.710
Autoencoder	0.061	0.166	0.577

		Called Class		
True Class	Neutral	Hard Sweep	Soft Sweep	Balancing
		Random Initialization		
Neutral	**1.000**	0.000	0.000	0.000
Hard Sweep	**0.978**	0.007	0.000	0.015
Soft Sweep	**1.000**	0.000	0.000	0.000
Balancing	**1.000**	0.000	0.000	0.000
		Autoencoder Initialization		
Neutral	**1.000**	0.000	0.000	0.000
Hard Sweep	0.145	**0.831**	0.004	0.021
Soft Sweep	0.011	0.001	**0.987**	0.000
Balancing	0.030	0.028	0.001	**0.941**

doi:10.1371/journal.pcbi.1004845.t006

Runtime

Task	ABCtoolbox	Deep Learning
Simulating data	370 hrs (10 ~ 15 cores)	370 hrs (10 ~ 15 cores)
Computing summary statistics	1800 hrs (1 core)	1800 hrs (1 core)
Demography only (1000 × 160 datasets)		
Training and testing (filtered statistics)	114 hrs (1 core)	3.75 hrs (20 cores)
Training and testing (unfiltered)	336 hrs (1 core)	11 hrs (20 cores)
Demography & selection (2500 × 160 datasets)		
Training	N/A	74 hrs (20 cores)
Testing	N/A	3 min (1 core)

Summary

Goal of Paper: Demonstrate potential of deep learning in population genomic analysis

Why?:

- Can build model to interpret complex population model better than direct statistical inference (difficult theoretically and computationally)
- Distinguish uninformative and informative summary statistics

Takeaways

- Could combine with other methods to compare results
- Using deep learning to classify regions as neutral or selected is very appealing for subsequent analysis
- Deep learning can make efficient use of even a limited number of simulated datasets

Future Work

- Learn how various summary statistics better relate to parameters
- Incorporate datasets with wider range of selection onset times
- Effect of ratio of recombination to mutation rate on simulations
- Use ABC MCMC to simulate data
- Use deep learning to select informative statistics for ABC
- Apply techniques to population structure and splits

Future Work (cont.)

- Use deep learning for continuous parameter inference in population genetics and other fields
- Combine “black-box” models with coalesent modeling we know to be realistic

Citations

Sheehan, Sara, and Yun S. Song. ”Deep learning for population genetic inference.” PLoS Comput Biol 12.3 (2016): e1004845.

Zhou H, Hu S, Matveev R, Yu Q, Li J, Khaitovich P, et al. A Chronological Atlas of Natural Selection in the Human Genome during the Past Half-million Years; 2015. BioRxiv preprint, http://dx.doi.org/10. 1101/018929.

For Icons: https://thenounproject.com