A Likelihood-Free Inference Framework for Population Genetic Data using Exchangeable neural networks

Gunnar Nelson-Virignia Tech

Introduction

- Explosion of DNA sequencing from the past decade
- Human Genome project
- Genetics has inference methods for model classes
- 1.) Is populations at a exchangeable
- 2.) Methods are usable to exploit data
- 3.) Challenges are tackled by likelihood free methods of scientific simulations of datasets

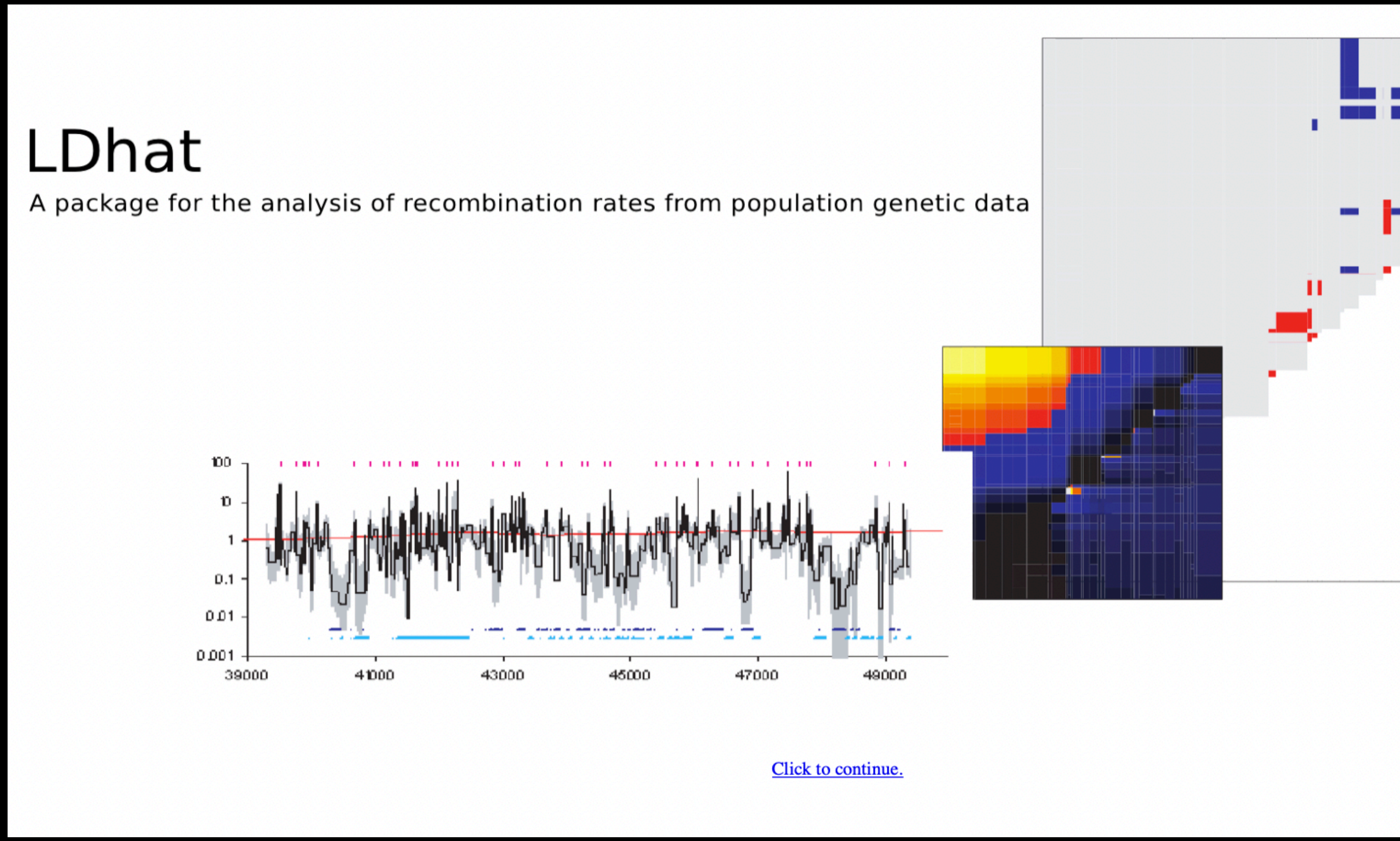
Related Work

- Focus rom the micro to macroscopic standpoint
- High focus on population genetics for disruption of data based on finite amount of data
- ABC interest of the population
- Status are compared and a summary of real data
- Simulated aparmaters for weighted according to similarity of stats
- However, there needs to be a better metric for understanding the whole population, while understanding the accuracies and exploingin symmetries int the input data

Related Work

Continued

- The team demonstrates the idea of exploiting the whole population data genomics
- Te team wanted to focus on hotspot testing and estimation fo a population based on the scalable whole genomic pattern
- Such as LDhot (https://github.com/auton1/LDhot)
- Current relies on likelihood on sparse pollution



LDhat

A package for the analysis of recombination rates from population genetic data

Click to continue.

Methods

Problem

- Given data X in relative sequence with population
- Each population will have a genetic datapoint which correspond to the individual and columns indicating the presence of a Single Nucleotide Polymorphism (SNP)
- Goal is to learn posterior in hotspots genetic
- 1. Simulation on the Fly: Sampel a mini bath
- 2. Exchange able neural network to match the mapping and input of the label as well

Methods

Exchange Neural Networks

- Goal is to learn the function where the parameters and distribution of the population can also be exchanged with binary matrix to apply the symmetrical function of the output of each map of the function

- The exhfnagle neural network has advnagges with the neural network given the flexibility of the models and efficiency of data augmentation

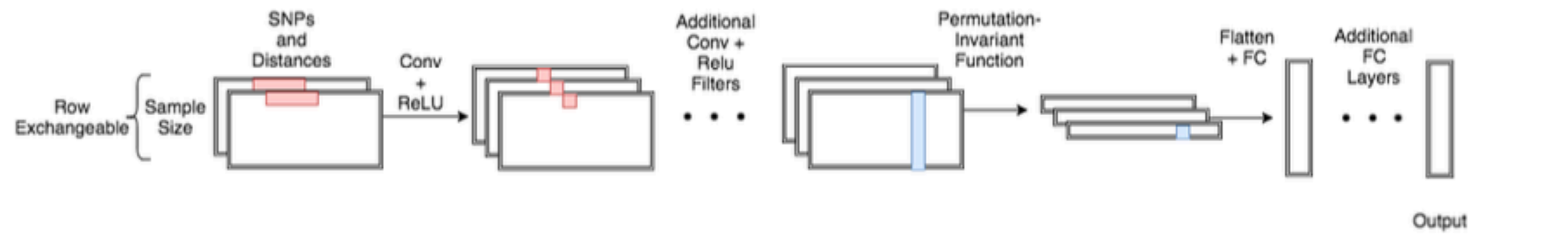
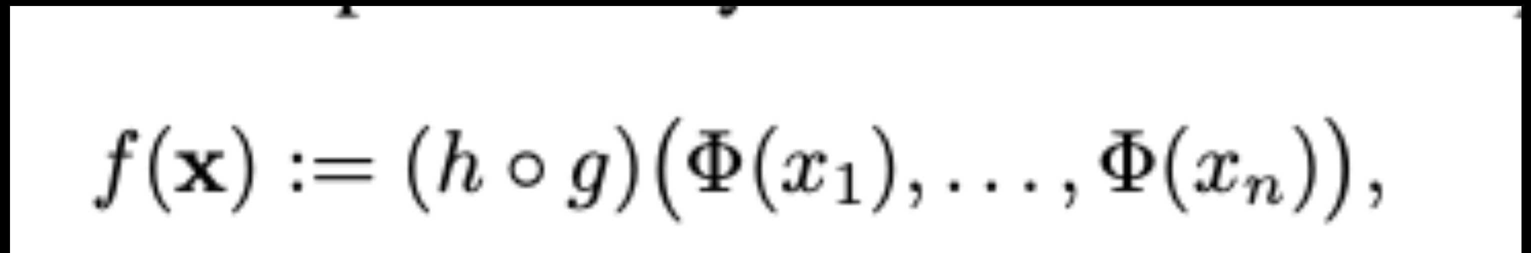


Figure 1: A cartoon schematic of the exchangeable architecture for population genetics.



$$f(\mathbf{x}) := (h \circ g)\big(\Phi(x_1), \ldots, \Phi(x_n)\big),$$

Methods

Simulation on the fly

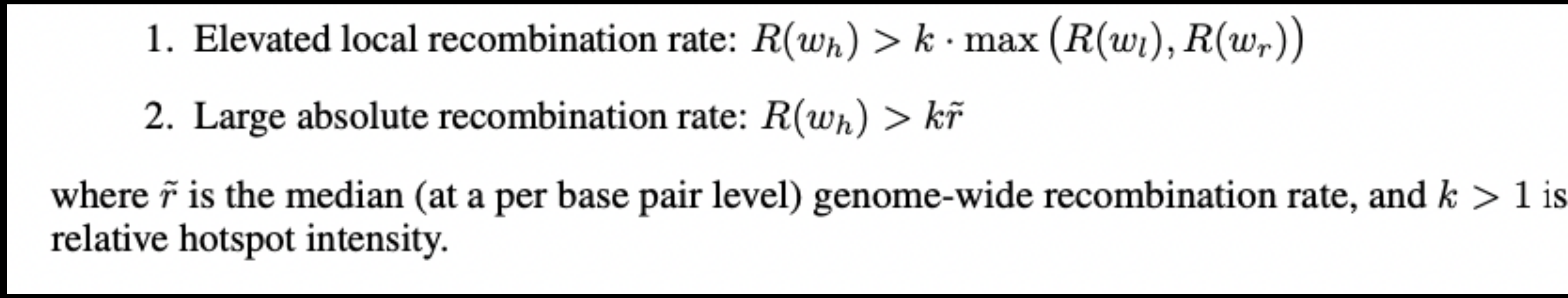
- Supervised learning methods (also machine learning) used fixed training set and mutable passes over a convolution
- During scientific simulations the amount fo training data avaialbe is limited to amount of computing time available for simulation so the team proposes simulating each training datapoint afresh each one by the training data
- Wanted to focus when the model is subject to the data of the pollution
- Given data is unpredictability of the populations

Empirical Study

- There were few hotspots with testing the reocmibnaiton fo hotspots to test the following and
- validated the methods fo the data to showcase the learning rate and batch count with each oft methods to focus on the learning rate
- Did a comparison study of the ground truth of the methods against LDhot to supplement the classification of the data
- U = 1.1 x 10^-8 per generation of nucleotide,
- **Convolution path length of 5 SNPs, 32 and 64 convolutions**

Recombination hotspots

- Recombination hotspots are short regions of the genome with high recombination rate
- They play an import role in complex disease inheritance patterns
- The team wanted to evaluate:
  - 1. Local recombination rate
  - 2. Absolute recombination rate



1. Elevated local recombination rate: $R(w_h) > k \cdot \max\big(R(w_l), R(w_r)\big)$
2. Large absolute recombination rate: $R(w_h) > k\tilde{r}$

where $\tilde{r}$ is the median (at a per base pair level) genome-wide recombination rate, and $k > 1$ is relative hotspot intensity.

Evaluation of exchange neural networks

- Comparison of 2D convolutions with varying patch heights
- Accuracy under human-like population genetic parameters will vary 2D patch heights to show the left of Figure 2
- The team found the accuracy decreased with a faster per batch computation time(not surprising)
- The accuracy silll remains about 90% during test time for sample sizes wrought 0.1-20x the training sample size

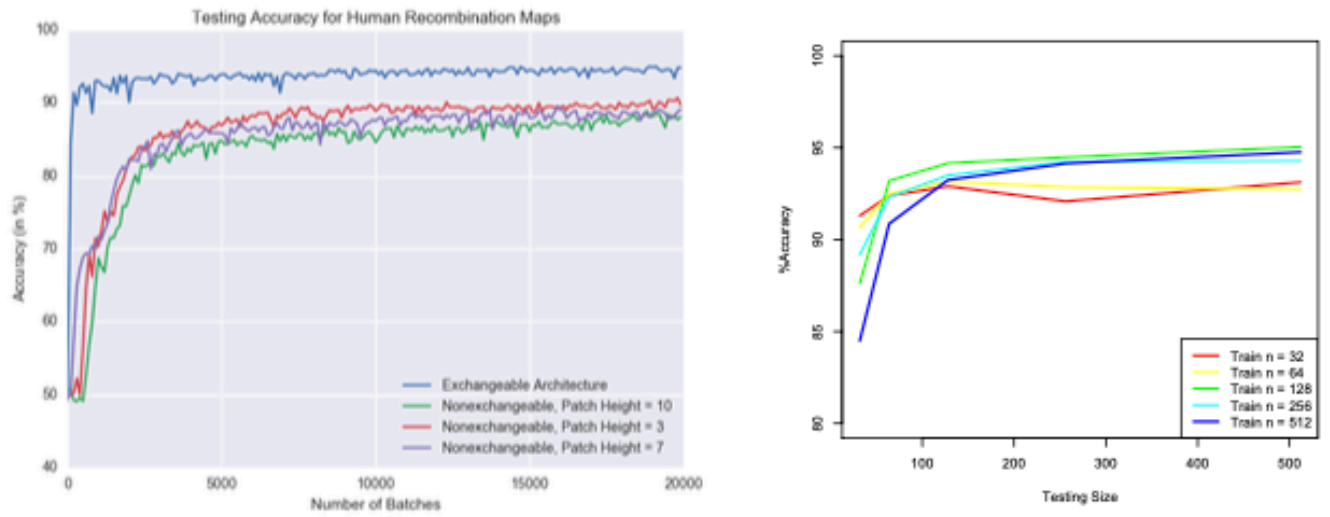


Figure 2: (Left)Accuracy comparison between exchangeable vs nonexchangeable architectures (Right)Performance of changing the number of individuals at test time for varying training sample sizes.

Evaluation of the fly

- The team simulates-on-the-fly comparison to the standard fixed training set with the fixed set of 10,000 to run 20,000 training batches with a test side of 5000

- The team then investigated posterior calibration

- Measuring whiter there is any bias of uncertainty output from the neural network

- Form the 25000 datapoints

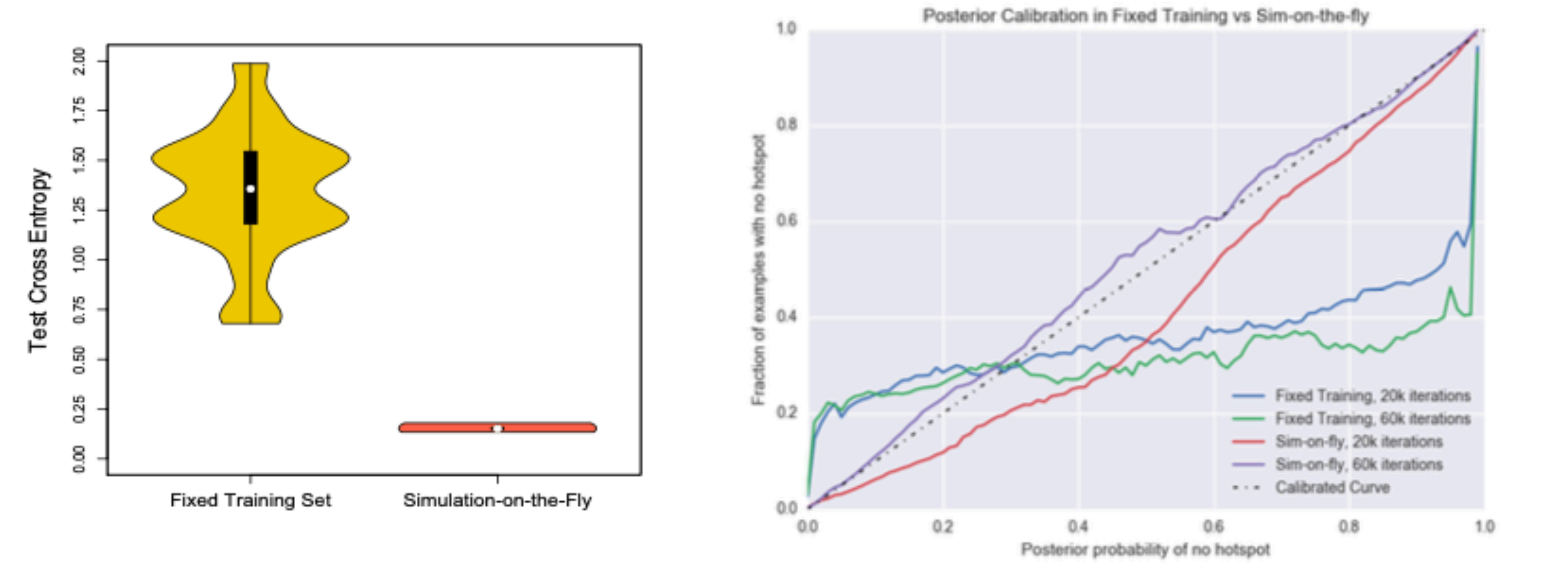


Figure 3: (Left)Comparison between the test cross entropy of a fixed training set of size 10000 an simulation-on-the-fly. (Right)Posterior calibration. The black dashed line is a perfectly calibrate curve. The red and purple lines are for simulation-on-the-fly after 20k and 60k iterations; the blu and green lines for a fixed training set of 10k points, for 20k and 60k iterations.

Discussion/Conclusion

- The team proposes the first likiehood free inference method for exhnagle population genetic data to not rely solely on handcrafted summary statics
- Design neural netoewks which learn an exhangeable representation of population genetic data and map the posterior
- Development and application of neural netokws harness raw sequences to address important chanlalgnes of machine learning to population enetics
- The methods prove to be a major advance in likelihood free inference insulation where ABC is too inaccurate.

Thank you, questions.