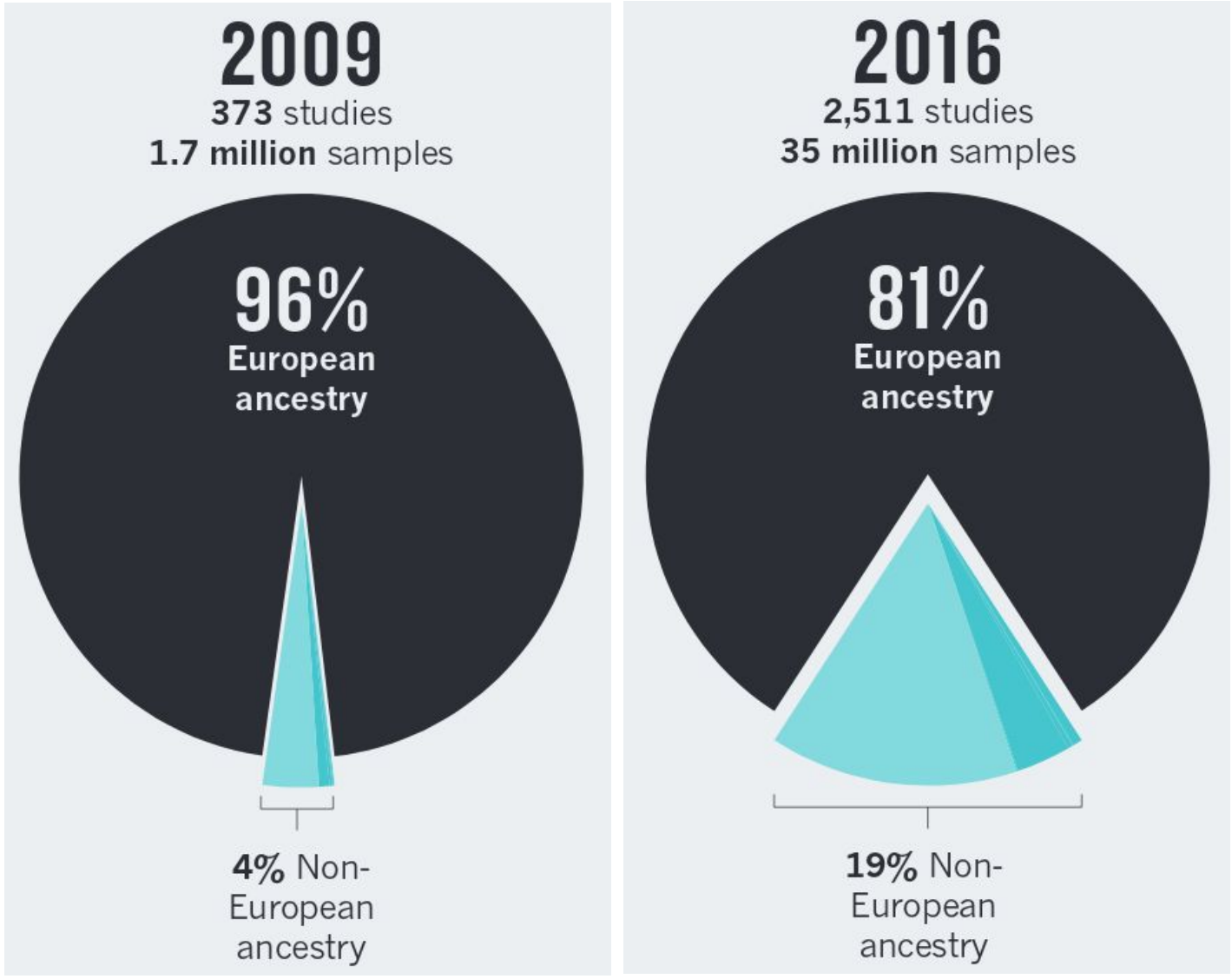
An HMM-CNN Method for Inferring Natural Selection Strengths in Evolutionary History

Nhung Hoang and Hyong Hark Lee
Advisor: Sara Mathieson
Department of Computer Science, Swarthmore College

Motivation

Genes under natural selection in European populations are well studied. We know where they are located on the chromosome and which of three broad categories they fall under (diet, pigmentation, and immunity). However, it is unclear whether this categorization holds for other populations, which have been historically less studied.

We propose an HMM-CNN integrated method, a machine learning-based tool that takes advantage of the growing accessibility to genetic data from geographically distributed human populations. Our method improves upon summary statistics, the traditional method for quantifying information about a population based on genetic samples from the population.



Background

Genetic Sequencing

- Sequencing technology has rapidly improved over the years, opening the gate to an abundance of genetic data available for study
- The evolutionary history of humans can be pieced together by observing patterns of genetic variations across geographically scattered populations
- By focusing on specific loci on chromosomes, population geneticists have found correlations between genes and geographic location

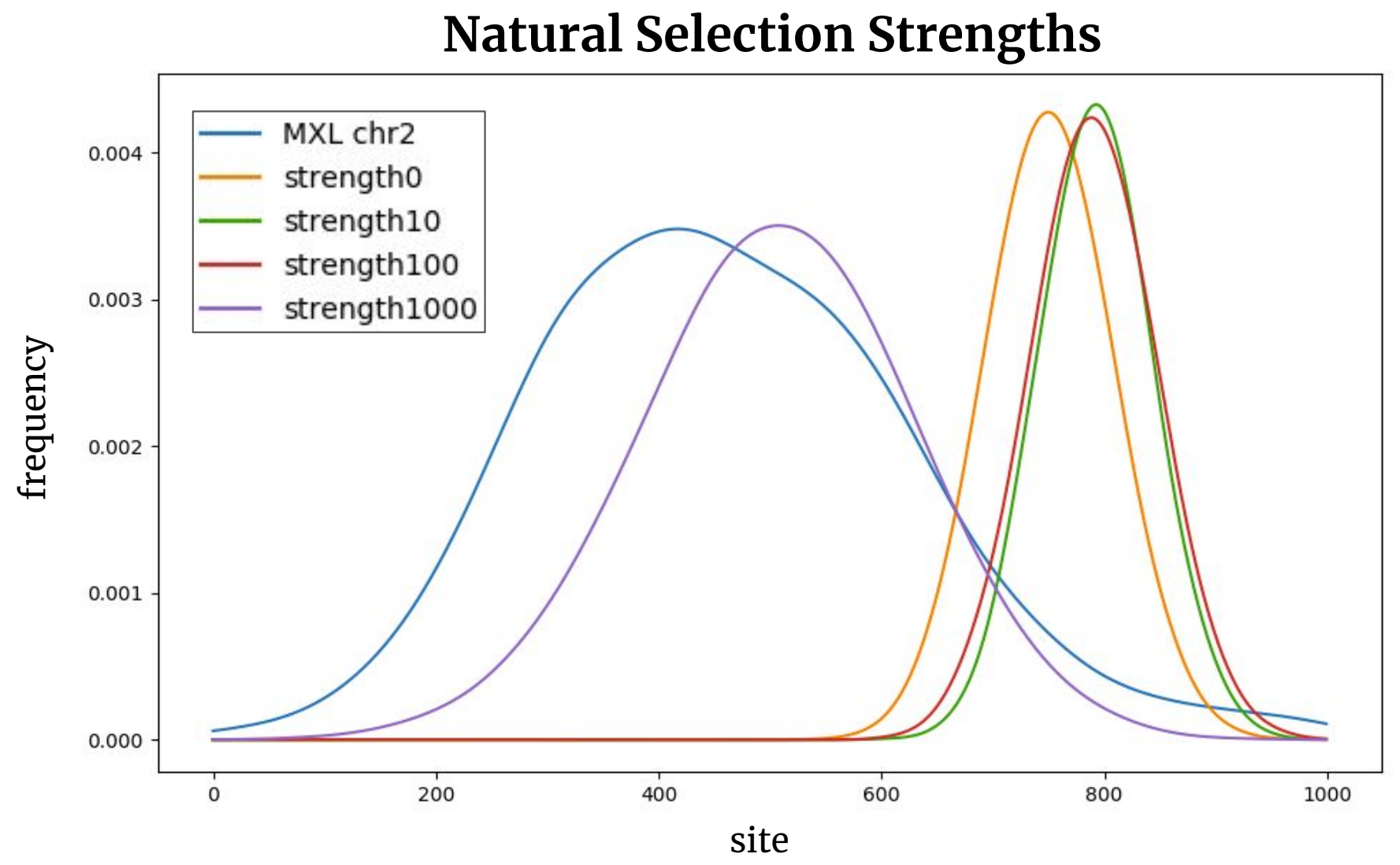
Natural Selection

Yum! Green beetles! Our favorite!
...generations later...
...generations later...

- Positive natural selection occurs when an advantageous variant from a novel mutation spreads
- Regions of the chromosome under selection are observable as drops in variation

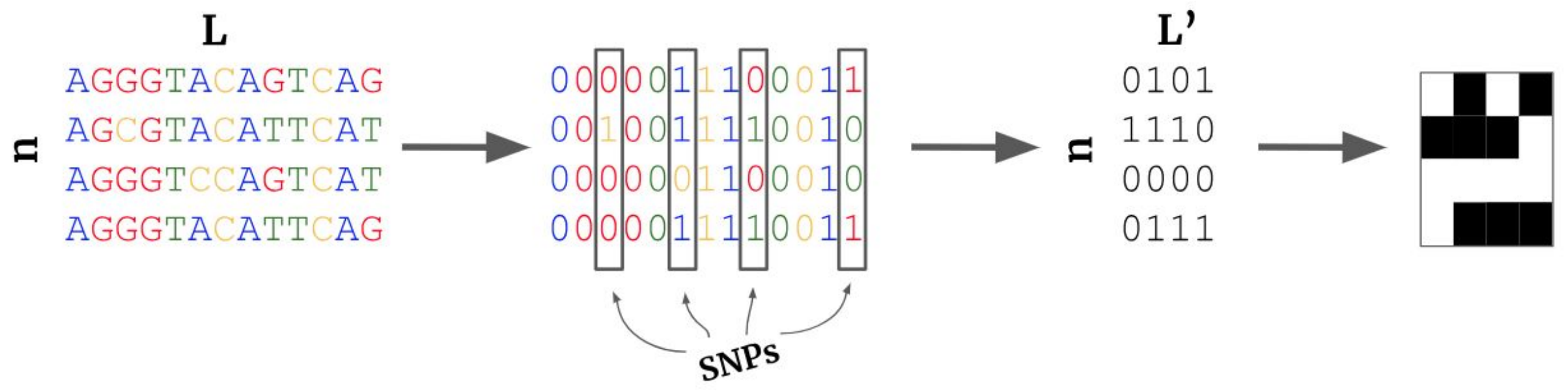
Summary Statistics

Distribution of SNPs for Natural Selection Strengths



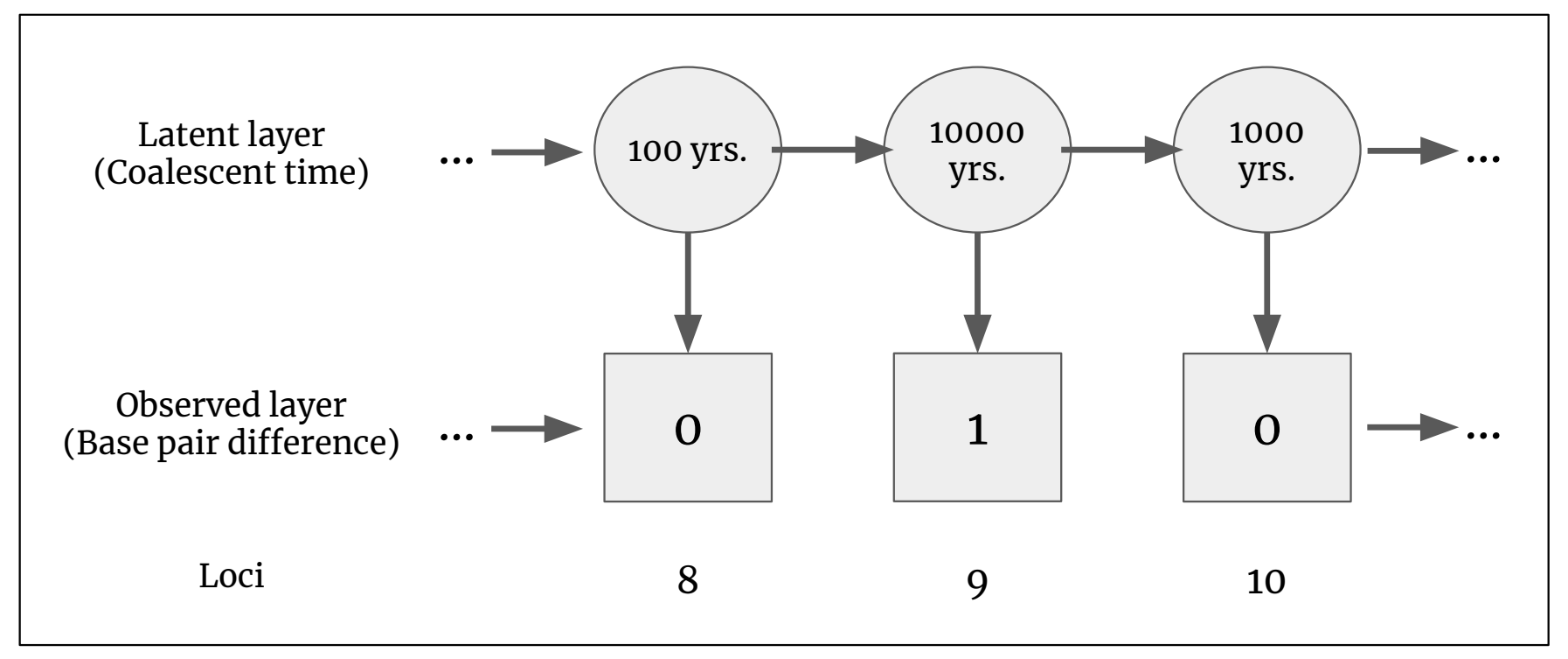
- Numerical summary of patterns within a set of genetic data, interpreted against well-studied but theoretical values
- Yields simplistic and nearly comprehensive results, but it is computationally expensive
- Usually undermined by confounding variables and lossy compression
- SNP count is a good summary statistic for detecting natural selection, but it is not sufficient when we want to know the strength of selection

Data Simulation



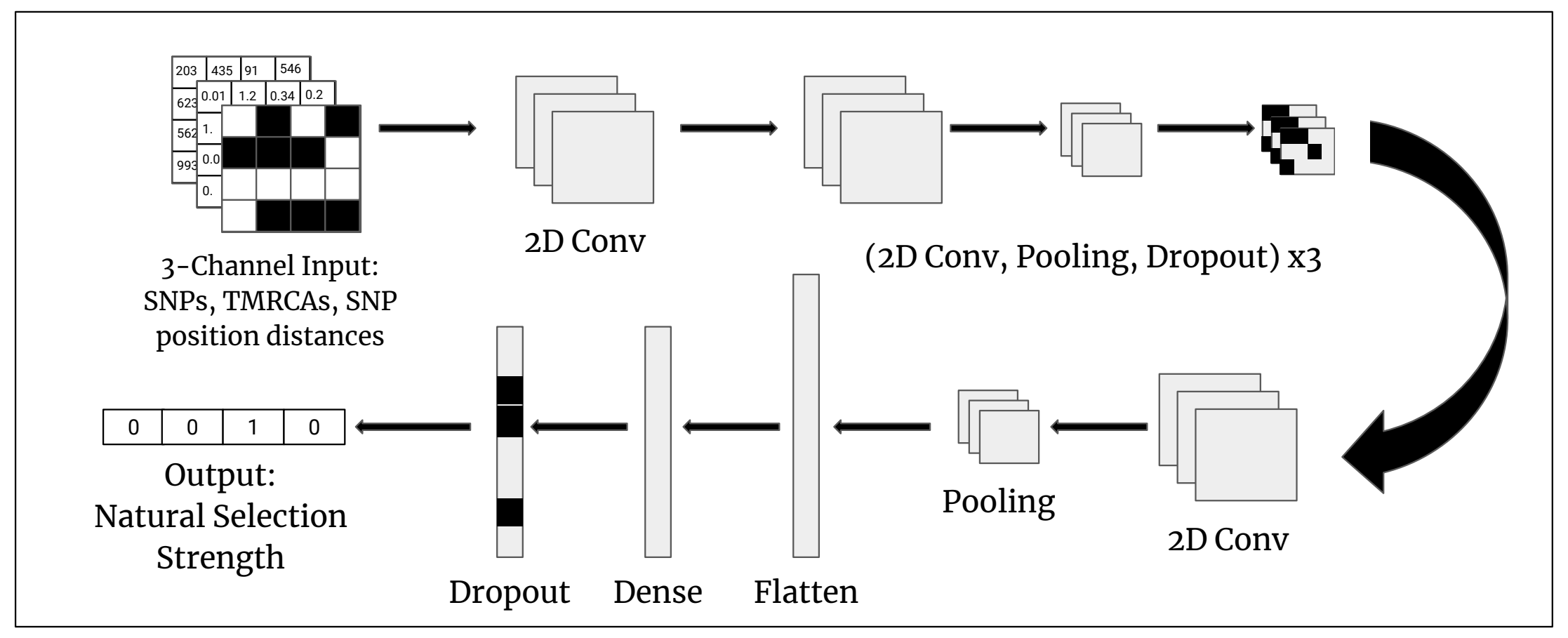
- Understanding of human evolutionary history is not yet comprehensive enough to evaluate new methods against
- Common practice to use simulated data for preliminary experiments
- Here, the simulated data of varying natural selection strengths are generated using a coalescent simulator
- Able to process the genetic data as if they were black & white images by first binarizing the SNP occurrences

Integrated Method



Hidden Markov Model (HMM)

unsupervised learning algorithm used here to link what is observed in a pair of sequences (same or different base pairs) to the predicted time to most recent ancestor (TMRCA)

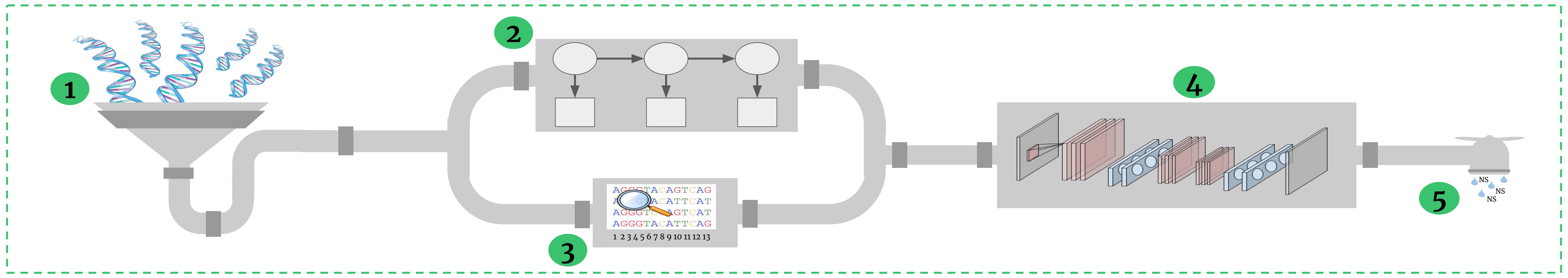


Convolutional Neural Network (CNN)

detect patterns within a dataset and represent the data by its most informative features in ways that often times appear unintuitive to humans

summarizes global trends

extracts local details



1

A region of genetic sequences from multiple individuals of a single population are fed into the pipeline.

2

Two of the sequences from the population sample are fed into the HMM. The TMRCA of each base pair of the two sequences are calculated and stored as a CNN input channel.

3

The SNPs and the distances between their local positions (from the entire population sample) are stored as two separate CNN input channels.

4

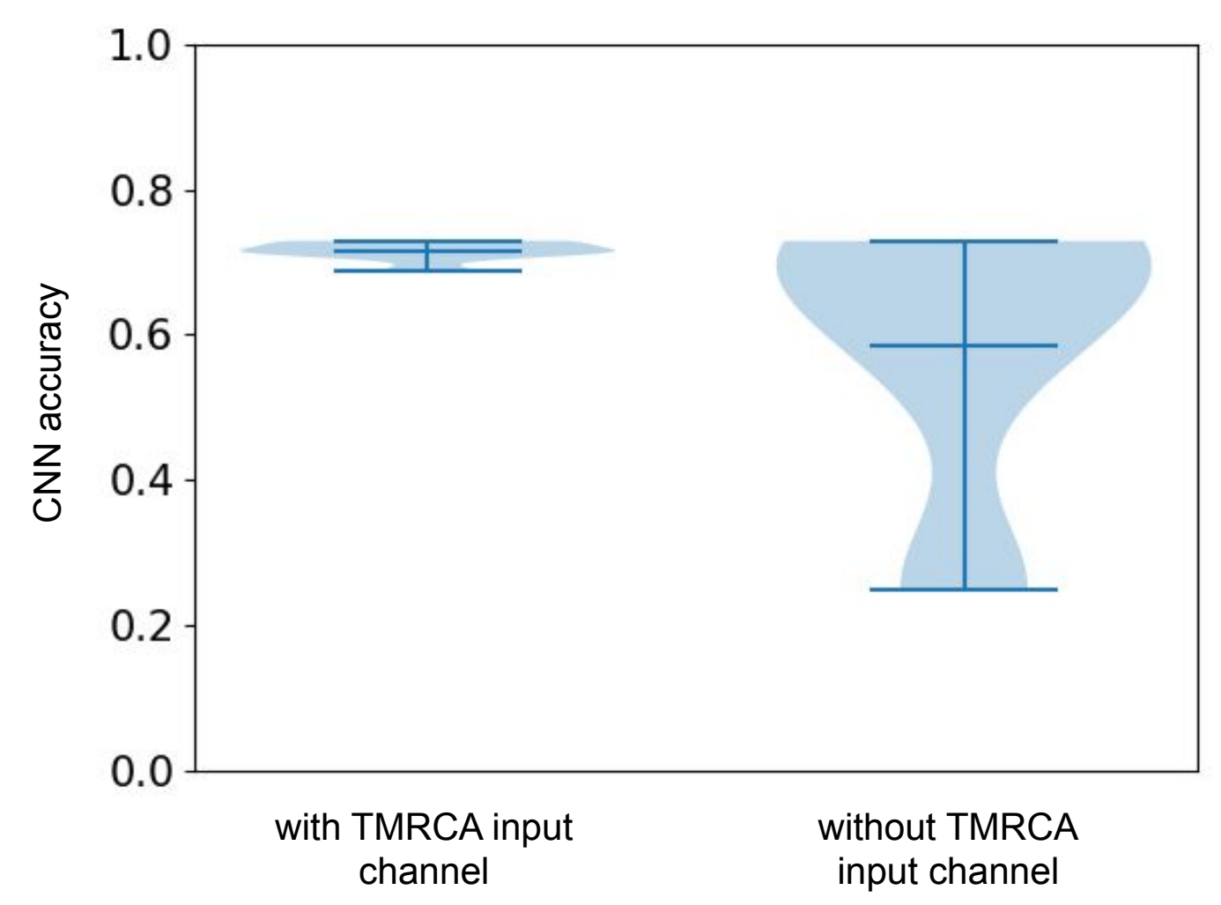
The three channels (in order of SNPs, TMRCAs, and position distances) are fed into the CNN as a single input.

5

The CNN outputs its prediction for what the population sample's natural selection strength is.

Results

Effect of TMRCA on CNN Accuracy



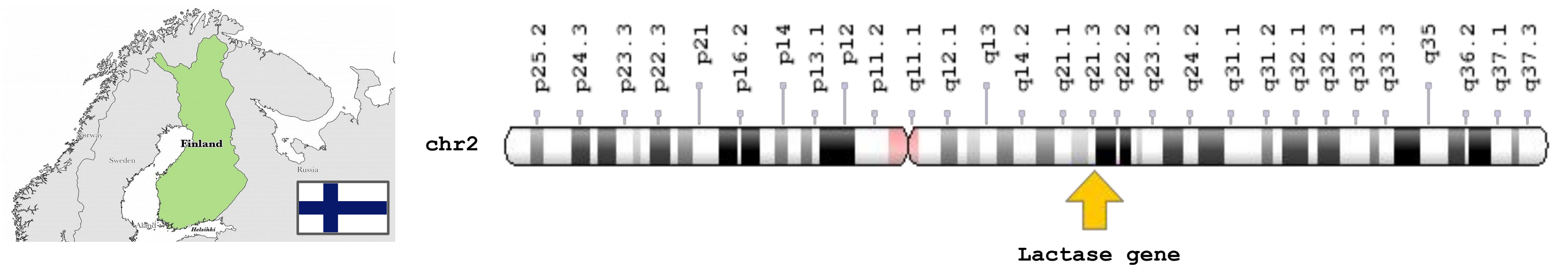
We found that CNNs which included TMRCA as an added channel consistently scored high accuracies (70%), while CNNs without a TMRCA channel had inconsistent accuracies.

Confusion Matrix on Natural Selection Strength (msms test data)

		Network-Predicted			
		0	10	100	1000
Actual	0	4443	1437	111	2
	10	3257	2563	107	9
	100	1048	1295	3646	39
	1000	0	8	51	5984

This confusion matrix was derived from a CNN with a TMRCA channel. The network correctly predicted the natural selection strength most of the time. Furthermore, the incorrect predictions tend to be near the ground truths.

Next Steps



Our next step is to retrain our method using more realistic simulated data. The end goal is to be able to run real human population data through the method to extract novel information about such data.

The lactase gene is well understood, particularly for European populations that saw intense growth after the domestication of cows. We hope to eventually be able to use the lactase gene of the FIN population to validate our method.

Acknowledgements

We would like to extend our deepest gratitudes towards Professor Sara Mathieson for her guidance and support on this research project. We would also like to thank the Dean's Office at Swarthmore College for supporting us with research grants.

References

J. Chan, V. Perrone, J. P. Spence, P. A. Jenkins, S. Mathieson, and Y. S. Song, "A Likelihood-Free Inference Framework for Population Genetic Data using Exchangeable Neural Networks," 2018.

H. Li and R. Durbin, "Inference of human population history from individual whole-genome sequences," *Nature*, 2011.

D. R. Schrider and A. D. Kern, "Supervised Machine Learning for Population Genetics: A New Paradigm," Trends in Genetics, vol. 34, no. 4, pp. 301–312, 2018.