Genomics and Deep Learning to Unveil the Social Stratification in Urban Admixing Populations

Author: Ben Nouhan, bjn20@ic.ac.uk

Lead Supervisor: Dr Matteo Fumagalli, m.fumagalli@imperial.ac.uk

Co-Supervisor: Dr Alex Mas Sandoval, a.mas-sandoval@imperial.ac.uk

Department of Life Sciences
Imperial College London

December 19, 2020

Population Genomics, Deep Learning, Genetic Admixture, Assortative Mating, Social Stratification, Anthropological Genetics

1 Introduction

Certain genetic phenomena, including assortative mating and sex bias, have the potential to alter the structure of human populations. This in turn modifies genomic variation, reflected in a population's genomic data which can be used to infer said phenomena. Various factors, cultural or socio-economic, can cause them to arise or to manifest themselves. Historically, the social stratification of societies by wealth, power and perceived race, alongside explicit racial segregation policies, have modulated human mating behaviour away from random mating.

In the past century, geopolitical and economic landscape has experienced and will continue to experience intercontinental migration towards areas of high population density, engendered by the likes of globalization, industrialisation, shifting demographics, the fallout of colonialism and global warming. This mass migration occurring within a relatively small timescale has founded new, diverse societies with complex and stratified urban population structures.

The modern era is not the first in which this has occurred, indeed the convoluted layers of ancient processes of migration and subsequent admixture, which shaped populations around the world over millennia, are shown to have been sex-biased in many cases, and may additionally have been impacted by localised assortative mating patterns.(Goldberg et al., 2014; Skoglund & Mathieson, 2018)

This project seeks to utilise deep learning algorithms and widely available genomic data in the elucidation of how complex human mating behaviours have been effected by both social and economic conditions, which stem from the genetic structure of different historic and current admixing populations.(Sheehan & Song, 2016)

2 Methods

Deep Learning

I will be building and optimising neural networks using the Keras and TensorfFow deep learning frameworks and Python.(Abadi et al., 2016) Different combinations of these networks and both the multi-class classification and regression methods for mutliple parameter estimation will then be tested and compared by parameter estimation accuracy Simulations will be carried out with SliM, an open-source evolutionary simulation framework, to simulate training and validation datasets in preparation for whichever real-life dataset is chosen.(Haller & Messer, 2019)

Data

I plan to use whole-human genome samples from the region we will be studying. Once the neural networks I will be using are fully optimised and trained, they will be applied to populations from the American continents.

3 Anticipated Outcome

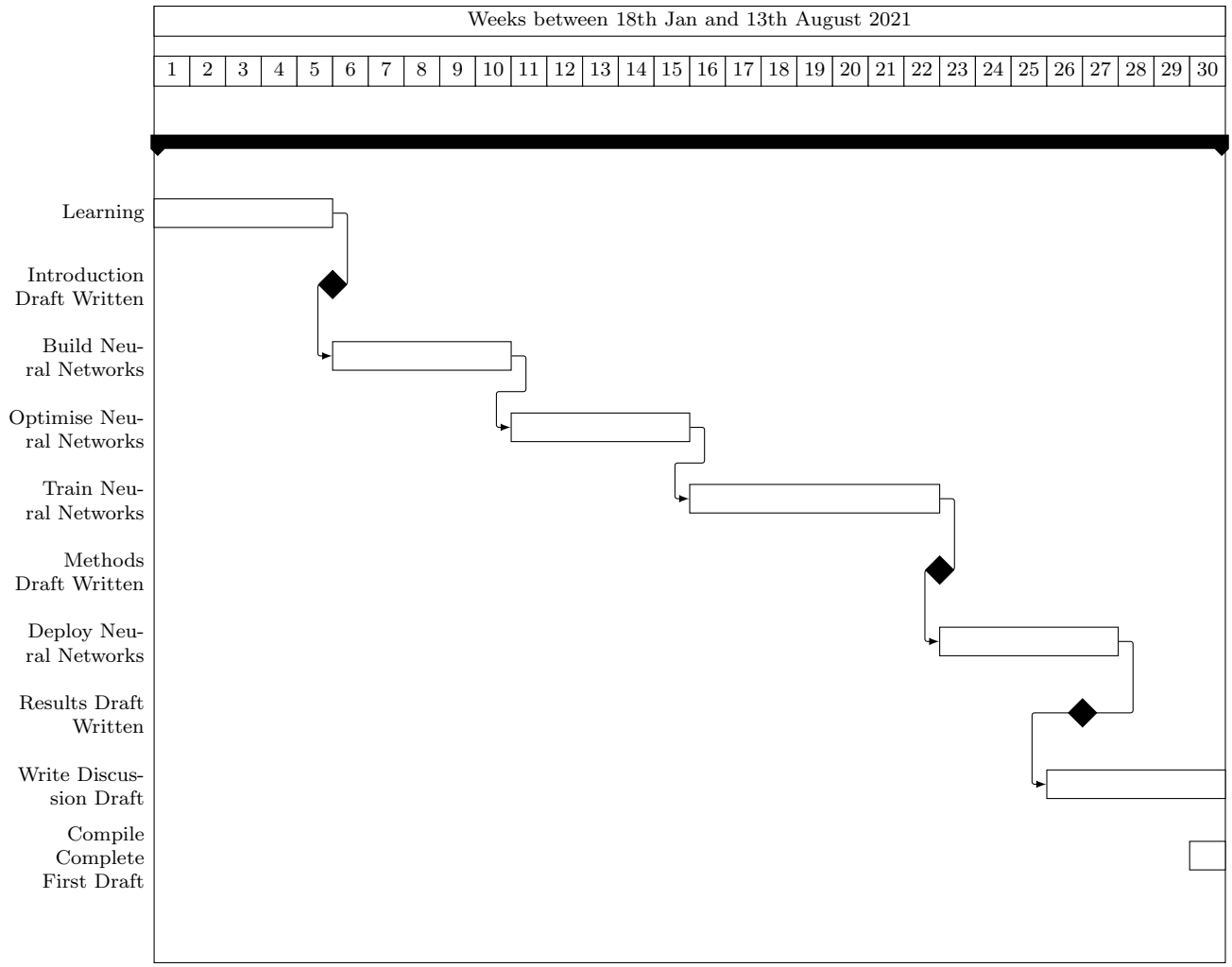
The project's outcome will be in two forms. I will be comparing the efficacy of the different neural network combinations outlined above based on their accuracy estimating multiple parameters. I will then be using highest-performing method to infer parameters, such as assortative mating and sex bias, from genomic data and ultimately integrate my findings into the phenotypic and cultural landscape of the studied region.

4 Budget

No purchases are expected to be required.

5 Project Feasibility

Starting 18th Jan, there are 30 weeks before I should hand in a second draft to my supervisor on the 13th August, for his final thoughts and my finishing touches before the 26th August deadline. My plan for those 30 weeks is shown in the chart below.



6 Supervisor Statement

Supervisor: Fumagalli Matteo

I have seen and approved the proposal and the budget.

Signature:

- -

Date: 19/12/20

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., ... Zheng, X. (2016). TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems arXiv 1603.04467. http://arxiv.org/abs/1603.04467

Goldberg, A., Verdu, P., & Rosenberg, N. A. (2014). Autosomal admixture levels are informative about sex bias in admixed populations. *Genetics*, *198*(3), 1209–1229. https://doi.org/10.1534/genetics.114.166793

Haller, B. C., & Messer, P. W. (2019). SLiM 3: Forward Genetic Simulations Beyond the Wright-Fisher Model. *Molecular Biology and Evolution*, *36*(3), 632–637. https://doi.org/10.1093/molbev/msy228

Sheehan, S., & Song, Y. S. (2016). Deep Learning for Population Genetic Inference. *PLoS Computational Biology*, *12*(3), 1–28. https://doi.org/10.1371/journal.pcbi.1004845

Skoglund, P., & Mathieson, I. (2018). Ancient genomics of modern humans: The first decade. *Annual Review of Genomics and Human Genetics*, *19*, 381–404. https://doi.org/10.1146/annurev-genom-083117-021749