School of Mathematics and Statistics
Faculty of Science, Technology, Engineering and Mathematics

The Open University

2024 PhD Projects

Project title	AI and Machine Learning in Human Genetics
Principal supervisor	Kaustubh Adhikari
Discipline	Statistics
Research area/keywords	Artificial Intelligence (AI), Machine Learning, Multivariate analysis, Genetic Data
Suitable for	Full time applicants, Part time applicants

Project background and description

In recent years, the field of human genetics has grown immensely in terms of data production and analysis, now routinely analysing data from thousands of people over millions of genetic markers [1] and thousands of variables [2]. Consequently, statistical methods are being developed to handle such high-dimensional data effectively [3,4]. In particular, AI (artificial intelligence) and machine learning models have been popular approaches [5,6].

The proposed PhD project will aim to develop multivariate statistical models for the analysis of large numbers of genetic markers (genotypes) and physical characteristics (phenotypes), such as skin and eye colour [7]. Latest methods that are suitable for high-dimensional data, spanning both AI and machine learning domains [5,6] and classical statistical procedures [8,9], will be explored. Algorithms developed will have a focus on computational efficiency [10] to handle large-scale datasets. Such methods will be useful in discovery of new genes associated with new phenotypes, but also in better prediction of physical characteristics from genetic data, useful for forensic reconstructions [7].

There will also be the scope of studying ancient human samples such as the Neanderthals by studying genetic data from ancient DNA obtained from prehistoric human remains [11].

As an applied Statistics project, it will involve both theoretical and computational work. The candidate should have a strong knowledge of statistics with a suitable degree. Computational (R / Python etc.) and programming experience would be useful. No prior knowledge of genetics or biology is required - all necessary training will be provided.

An associated theme of the work would be about increasing the diversity of genetics research [12], discussing the gains brought about in science by broadening the perspective. There will be substantial scope for the student to engage in outreach and public engagement, if interested.

Background reading/references

1. Thompson, S. G., Willeit, P., UK Biobank comes of age. Lancet, (2015). 386(9993): p. 509-10.
2. Claes P, et al., genome-wide mapping of global-to-local genetic effects on human facial shape. Nat Genet (2018); 50:414–23.
3. Yang, J., Lee, S.H., Goddard, M.E., Visscher, P.M., GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet 88, 76-82 (2011).

4. The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447, 661–678 (2007).

5. Giardina A, Paria SS, Adhikari K, A naive method to discover directions in the StyleGAN2 latent space. arXiv:2203.10373 (2022).

6. Yelmen B, Jay F, An Overview of Deep Generative Models in Functional and Evolutionary Genomics. Annual Review of Biomedical Data Science (2023), 6:173-189.

7. Adhikari, K., et al., A genome-wide association scan in admixed Latin Americans identifies loci influencing facial and scalp hair features. Nat Commun, (2016). 7: p. 10815.

8. Turley, P. et al., Multi-trait analysis of genome-wide association summary statistics using MTAG. Nat. Genet. 50, 229–237 (2018).

9. Lloyd-Jones, L.R., Zeng, J., Sidorenko, J. et al., Improved polygenic prediction by Bayesian multiple regression on summary statistics. Nat Commun, 10, 5086 (2019).

10. Runcie DE, Crawford L. Fast and flexible linear mixed models for genome-wide genetics. PLoS Genet (2019), 15(2): e1007978.

11. Racimo, F., et al. The spatiotemporal spread of human migrations during the European Holocene. PNAS 117, 16 (2020), pp. 8989-9000.

12. Sirugo, G. et al., The Missing Diversity in Human Genetic Studies. Cell 177, (2019).