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

The Open University

2024 PhD Projects

Project title	Reproducible and Accessible Statistical Genetics Workflows tested in Evolutionary Psychiatry
Principal supervisor	Colette Christiansen (Mathematics & Statistics)
Co-supervisors	Wendi Bacon (Life, Health & Chemical Sciences) and Stefanie Biedermann (Mathematics & Statistics)
Discipline	Statistics
Research area/keywords	Statistics, Genetics, Evolutionary Psychiatry, Human computer interaction, Computer Science
Location	The Open University, Milton Keynes, United Kingdom
Suitable for	Full time applicants only

Project background and description

The majority of statistical genetics calculations, for example Genome Wide Association Studies and Polygenic Risk Scores are carried out using High Performance computing environments which require users to have excellent computing and programming skills. It is also difficult for these calculations to be reproduced as a number of steps are required using adhoc programmes. As our knowledge grows, these calculations will become more mainstream, and we have a greater ability to gain knowledge from investigating genomic differences between individuals. This makes it vital that a wide variety of users can access this field, not just those with advanced statistical and programming skills. The Galaxy Platform allows users to analyse data without programming skills enabling those with deep biological expertise to be able to carry out vital research to move the field of genetics forward.

The Galaxy Training Network2 and the annual GTN Smörgåsbord (led by Dr Hiltemann, Stubbs lab, Erasmus MC) provides a platform for high quality bioinformatics training using Galaxy. However, users can still struggle to apply the analyses to their own messy data, and are ever limited by the tools that currently exist in Galaxy. In this project, the student will build a Psychiatric Genetics statistical workbench, including polygenic risks scores and comparisons of single nucleotide polymorphisms. They will use ancient DNA samples as a small test case and use the created workflows to calculate polygenic risk scores (PRS), genomic evolutionary rate profiling (GERP) scores, and compare genomic differences to gain insights in the evolution of psychiatric disorders. In using the workflows, the student will evaluate the pipeline for accuracy, reproducibility and useability by individuals with a wide range of backgrounds. The student will employ user testing to assess and improve the tools, training, and most importantly, the decision-making by users, with an emphasis on engaging non-mathematicians in the vital statistics they use.

Background reading/references

1. Prüfer, K., Posth, C., Yu, H. et al. A genome sequence from a modern human skull over 45,000 years old from Zlatý kůň in Czechia. *Nat Ecol Evol* **5**, 820–825 (2021). https://doi.org/10.1038/s41559-021-01443-x
2. Batut, B. et al. Community-driven data analysis training for biology. *Cell Syst*. **6** , 752-758.e1 (2018).