Project title: Joint multivariate modelling and prediction in genetic data
Principal supervisor: Kaustubh Adhikari
Second supervisor: Fadlalla Elfadaly
Discipline: Statistics
Research area/keywords: Multivariate analysis, Bayesian modelling, machine learning, prediction
Suitable for: Full time applicants

Project background and description
With an explosion of projects producing data from thousands of people over millions of genetic markers [1], the need for advanced statistical methods to deal with high-dimensional data became more prominent. With this explosion in genetic studies came an explosion in proposed statistical tools too, but most of the proposed methods [2] do not deal with joint multivariate analysis of the input data.

Genome-wide association studies (GWAS) require two main types of input data on sampled individuals: millions of genetic markers (genotypes) and physical characteristics (phenotypes). Traditional analysis only performed univariate linear regression analysis by assessing one genetic marker against one phenotype at a time [3]. An increasing number of proposed statistical methods propose more sophisticated models such as Bayesian methods [4], but most methods do not leverage the multivariate structure in the data, for example the correlated nature of the variables.

The proposed PhD project will therefore aim to develop multivariate statistical models for joint analysis of the genotype-phenotype data. Methods suitable for high-dimensional data, such as Bayesian sparse priors [5], and machine learning [6], will be explored. Algorithms developed will have a focus on computational efficiency [7] 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 [8].

A large number of studies have made available summary statistics (effect sizes and p-values) for their analysis [9], even though individual-level data might not be available. New methods will be developed to integrate them during modelling and prediction to increase the effective sample size and power.

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 / Matlab etc.) and programming experience would be useful. No prior knowledge of genetics or biology is required.

Background reading/references
How to apply for this project

1. Read the Guide for applicants to check eligibility, especially entry and English language requirements.

2. Informal enquiries can be directed to the Director of Research.

3. Complete an application form, and send to the Director of Research by 8th March 2019.