**Principal Investigator**

Dr Mark Iles

**Address**

University of Leeds, Section of Epidemiology and Biostatistics, Cancer Genetics Building, St James's Hospital, Leeds LS9 7TF, United Kingdom

**Summary of research**

GWAS, stratification, principal components, genetic epidemiology

**Application Lay Summary:**

**1a:** Genetic association studies aim to identify genetic variants differing in frequency between cases and controls. If cases and controls are from distinct locations, any genetic frequency differences between the two will generate spurious association. Statistical methods exist to adjust for this, but will be less effective as studies become larger and more subtle levels of 'population stratification' (as in the UK) affect results. This project aims to quantify the extent to which population stratification in the UK can affect genetic association studies and develop statistical approaches to correct for it. This is applicable to any common genetically-influenced trait.

**1b:** It is vital that genetic association studies are correctly designed and analysed so that they do not generate spurious results. This would have an adverse impact on the understanding of disease genetics and potentially on areas such as genetic risk prediction. Our research aims to develop methods to avoid this.

**1c:** I will simulate case-control data unrelated to disease by randomly sampling individuals (and their genotypes) from UK Biobank. Population stratification can be simulated by sampling an excess of cases from one collection centre and an

excess of controls from another. Statistical tests of genetic association can be conducted on the dataset, with and without correction for stratification. This whole process of random sampling followed by analysis is then repeated many times to give an estimate of how often false positive disease/gene results are generated under different degrees of stratification, oversampling and stratification-correction.

**1d:** To investigate subtle stratification effects requires a large sample size. I previously thought the entire dataset of 500,000 would be unmanageable. However, recent emails suggest this is not so. The advantage of having the whole dataset is that GWAS-sized case-control sets (several thousand people) from specific centres/areas can be studied and validated in independent samples from the same centres/areas. These numbers will ensure sufficient control samples from each centre, when investigating the effect of stratification when genuine genetic case-control signals are present (with 'genuine' cases). Thus I would like genotype data on all samples.