Population stratification in the UK10K cohorts project: Rare variant analysis by whole genome sequencing in 3,621 samples

Klaudia Walter
on behalf of the UK10K Consortium Cohorts Group

Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK
Email: kw8@sanger.ac.uk

Background

The UK10K Cohorts project aims to research the relationship between rare and common genetic variants with a comprehensive set of quantitative measures relevant to cardiovascular and metabolic disease. For this purpose, the genomes of nearly 4,000 individuals from two large population samples in the UK (TwinsUK, N=1,754 and ALSPAC, N=1,867) were sequenced at low pass (median coverage 6x). After stringent QC steps, the data set comprises 44.5 million SNPs, 3.5 million INDELs and more than 20,000 large deletions across genomes of nearly 4,000 individuals from two large population samples in the relevant to cardiovascular and metabolic disease. For this purpose, the UK10K Cohorts project aims to research the relationship between rare and common variants.

Population

The cohorts were chosen to have differing geographic properties, with ALSPAC participants originating from geographically restricted area (Avon) in the South West of the country, while TwinsUK participants have UK-wide origin.

Population structure is a confounder of association studies based on common variants, but the influence of rare variants has been less well studied.

We exploited the properties of the population to study the extent to which geographic stratification exists at rare variants.

Phenotype data

The GAM model (generalized additive model by Hastie and Tibshirani) specifies a distribution (such as a normal distribution, or a binomial distribution) and a link function \( g \) relating the expected value of the distribution to the \( m \) predictor variables, and attempts to fit functions \( f(x) \) to satisfy:

\[
g(E(Y)) = \beta_0 + f_1(x_1) + f_2(x_2) + \ldots + f_m(x_m)
\]

The functions \( f(x) \) may be fit using parametric or non-parametric means, thus providing the potential for better fits to data than other methods.

GAM models can be used to account for regional differences as shown for the trait Glucose which was averaged within small regions for the TwinsUK samples.

The plots show the distribution before and after fitting a GAM model.

Genotype data

We implemented a method similar to the one proposed by Mathieson & McVean (Nature Genetics 2012) to assess the extent at which rare alleles are shared locally. From our analysis it seems there is no obvious geographical effect on the sharing of rare alleles. However, the effect could be masked by the number of comparisons made.

UK10K Cohorts Team

UK10K Chair: Richard Durbin (WTSI)
UK10K Cohorts Chairs: Nicole Soranzo (WTSI), Nicholas Timpson (Bristol University), Brent Richards (McGill University)