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

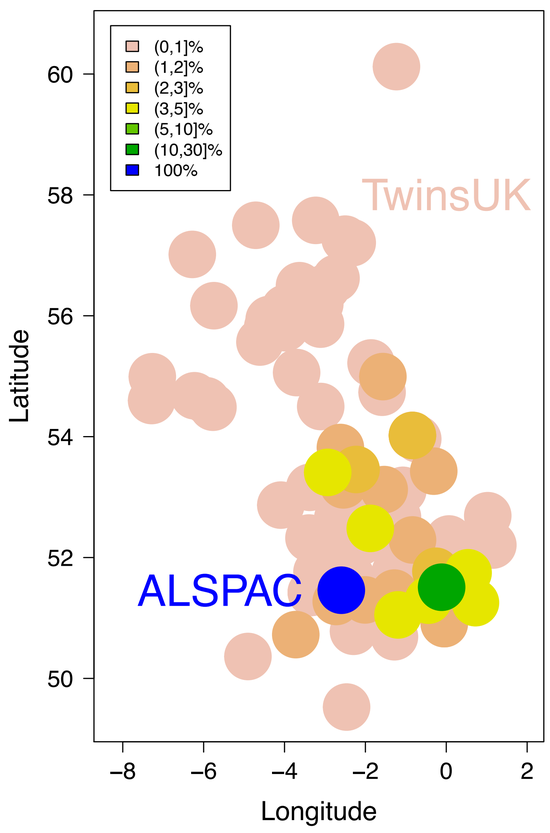
Klaudia Walter
on behalf of the UK10K Consortium Cohorts Group (http://www.uk10k.org/studies/cohorts.html)

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 3,621 study participants.

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.

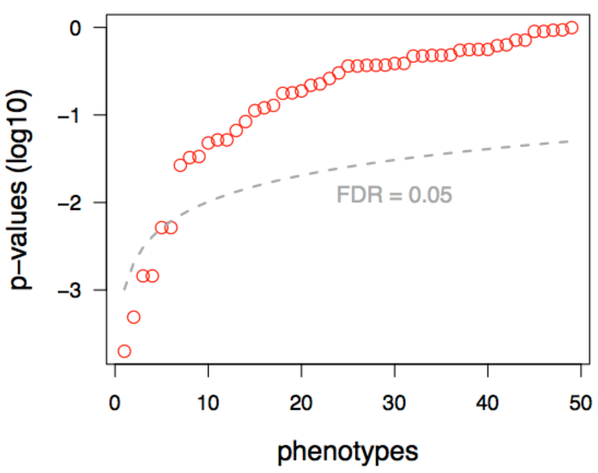
Phenotype data

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 $f_i(x_i)$ to satisfy:

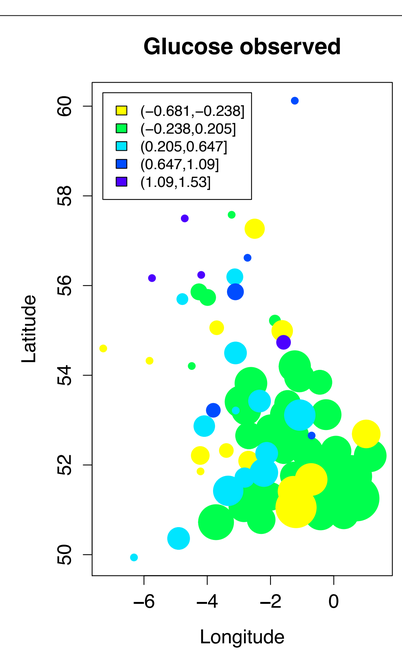
$$g(E(Y)) = \beta_0 + f_1(x_1) + f_2(x_2) + \ldots + f_m(x_m)$$

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

GAM models were fitted for each trait against geographical location. Then the significance of the smoothing functions were tested using ANOVA.



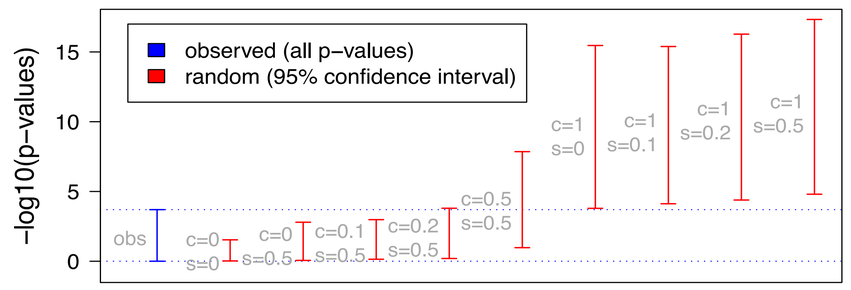
Trait	p-value
Urea (BMIadj)	0.00020
Glucose	0.00049
Height	0.00145
Height (std)	0.00145
Leptin	0.00517
Leptin (std)	0.00517



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.

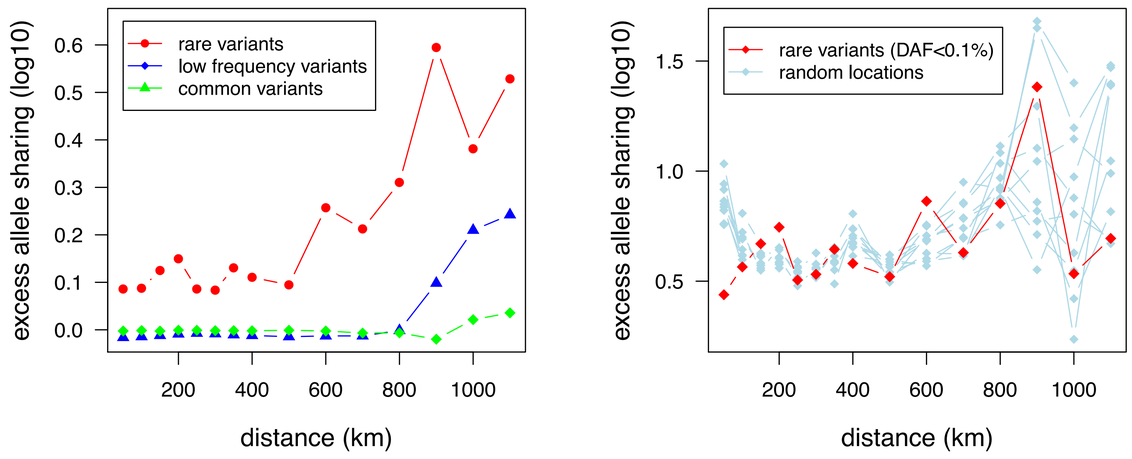
To test the significance of the GAM model p-values, random traits were generated using a normal distribution *N*(0,1), adding a regional spike and a north-south cline from 0.1 to 0.5 standard deviations (sd).



Observed p-values behave similar to simulated data with a north-south cline of 0.2 sd and a spike of 0.5 sd.

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.



We observed some regional stratification regarding the phenotypes, but so far no genotypic stratification which seems more challenging to analyse.

UK10K Cohorts Team

UK10K Chair: Richard Durbin (WTSI)

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

WTSI: Aaron Day-Williams, Andrew Brown, Audrey Hendricks, Chris Franklin, Dawn Muddyman, Eleftheria Zeggini, Ines Barroso, Ioanna Tachmazidou, Jie Huang, Jim Stalker, Julian Hughes, Kalliope Panoutsopoulou, Kim Wong, Klaudia Walter, Lorraine Southam, Lu Chen, Margarida Lopes, Petr Danecek, Shane McCarthy, So-Youn Shin, Yasin Memari; **Kings College London:** Alireza Moayyeri, Feng Zhang, Genevieve Lachance, John Perry, Kerrin Small, Kirsten Ward, Lydia Quaye, Massimo Mangino, Pirro Hysi, Sarah Metrustry, Scott Wilson, Tim Spector, Yalda Jamshidi; **University of Bristol:** Beate St Pourcain, Chris Boustred, Dave Evans, George Davey-Smith, Ghazaleh Fatemifar, Ian Day, John Kemp, Josine Min, Lavinia Paternoster, Tom Gaunt; **McGill University:** Celia Greenwood, Houfeng Zheng, Rui Li; **University of Leicester:** Louise Wain, Martin Tobin; **BGI Shenzen:** Jing Tian, Jun Wang, Sifei He, Yingrui Li; **EBI:** Graham Ritchie, Paul Flicek; **University of Oxford:** Jonathan Marchini