Quality control in the UK10K cohorts project: Low coverage whole genome sequencing in 2,432 samples

Klaudia Walter, Shane McCarthy, Petr Danecek, James Stalker, Nicole Soranzo and Richard Durbin on behalf of the UK10K Consortium Cohorts Group (http://www.uk10k.org/studies/cohorts.html)

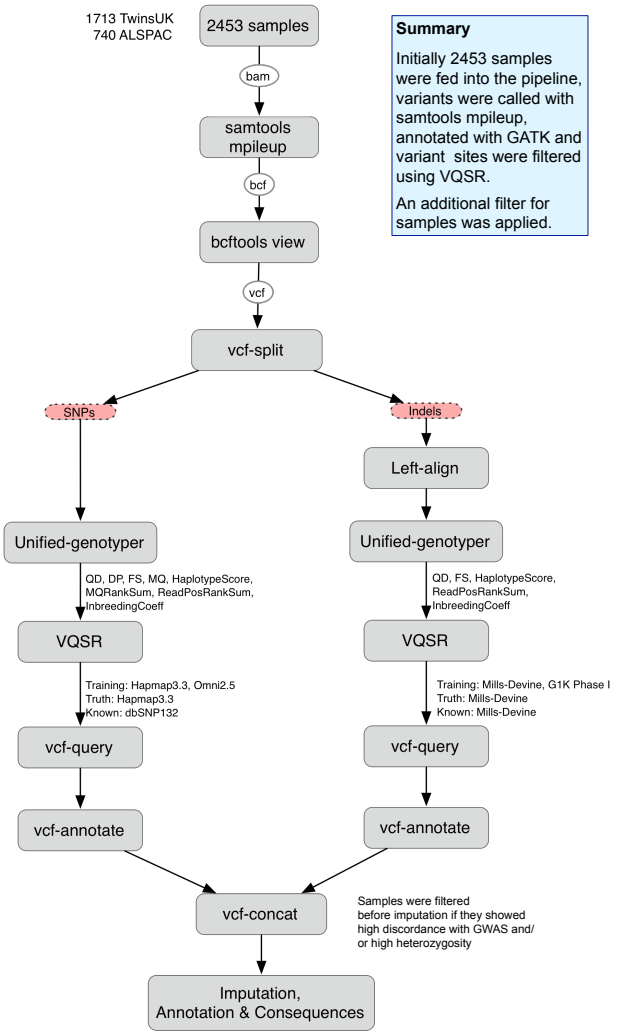
Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK
Email: kw8@sanger.ac.uk

Background

The UK10K project is a collaboration between the Wellcome Trust Sanger Institute and multiple research centres in the UK and Finland. As part of the **UK10K cohorts project**, 4,000 whole genomes are being sequenced by next-generation sequencing technologies at low coverage (average 6x) from two population based UK studies with rich clinical and molecular phenotype data, **TwinsUK** and **ALSPAC**.

The project will ultimately aim to assess the association of newly identified common and rare variants with **~50 cardiometabolic and anthropometric traits**. We describe here the current release of 2,432 whole genome sequences (1,692 from TwinsUK and 740 from ALSPAC).

Production workflow



Summary

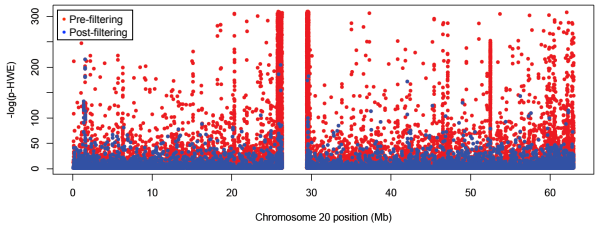
Initially 2453 samples were fed into the pipeline, variants were called with samtools mpileup, annotated with GATK and variant sites were filtered using VQSR.

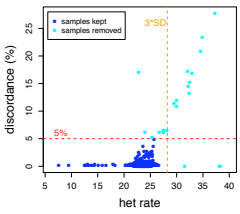
An additional filter for samples was applied.

Genotype data

Filtering of variant sites using VQSR

Extremely low *p*-values from Hardy-Weinberg Equilibrium tests cluster mainly around the centromere and in telomeric regions (shown for chr20).
Variant quality score recalibration (VQSR) appears to be an efficient filter for candidate variant sites.
http://www.broadinstitute.org/gsa/wiki/index.php/Variant_quality_score_recalibration





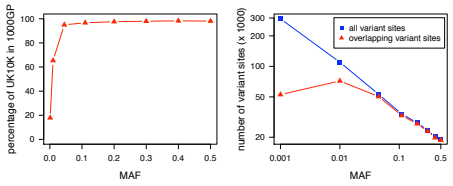
Filtering of 44 samples for excess heterozygosity

Discordance of samples with GWAS genotypes is highly correlated with excess heterozygosity. Likely cause is DNA contamination.

April 2012 Release

The table shows the numbers of samples and variants that passed QC.

Number of samples	2,432
Number of SNPs	32,767,011
Number of INDELs	4,656,017
Number of large deletions	22,320



Overlap of UK10K sites with 1000GP

Concordance is very high for common variants.

Accuracy of genotypes

Low-coverage genotypes are highly concordant with exome and GWAS genotypes (142 overlapping exome samples; 600 GWAS and 3500 exome variant sites; chr20 only).

	Exome	GWAS
Low-Cov	99.65	99.75
Exome		99.97

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