Single point and rare variant analysis by whole genome sequencing in 4,000 samples in the UK10K cohorts project

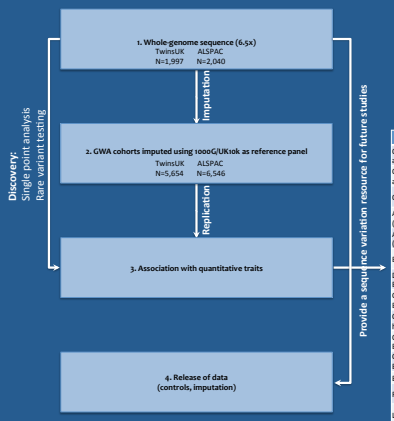
Josine L. Min[1], The UK10K Consortium (Cohorts Group)
[1]School of Social and Community Medicine, University of Bristol, Bristol, UK

Background

The UK10K project is a collaboration between multiple research centres mainly in the UK aiming to uncover rare genetic variants contributing to disease and health status by sequencing 10,000 people. As part of UK10K, 4,000 individuals from two deeply phenotyped cohorts – **TwinsUK** and the **Avon Longitudinal Study of Parents and Children (ALSPAC)** – have been sequenced to average 6.5x coverage using next-generation sequencing technology. The project will ultimately aim to assess the association of newly identified common and rare variants with **~50 cardiometabolic and anthropometric traits** (Fig 1). We describe here the current release of 2,432 whole genome sequences (1,692 from TwinsUK and 740 from ALSPAC).

Production workflow

Initially 2,453 samples were sequenced and variants were called with samtools mpileup, annotated with GATK and variants were filtered with Variant quality score recalibration (VQSR). In the current release of 2,432 whole genome sequences, we removed individuals with excessive heterozygosity and high discordance with GWA genotypes or high singleton rate.



Category	Phenotype (Bold present in both cohorts)
Obesity/ anthropometry	**Height, Weight, BMI**
Obesity/ anthropometry	**Hip, Waist, Waist-Hip Ratio**
Obesity (extended)	**Adiponectin, Leptin**
Anthropometry (extended)	**Total Fat Mass, Total Lean Mass**, Trunk Fat Mass
Anthropometry (extended)	Forearm length, Upper arm length, Head circumference, Sitting height, Leg length
Early Life	Birthweight, Birth Length, Placental Weight, Gestational Age, Method of Delivery, Ponderal Index
Diabetes Biochemistry	**Fasting Glucose, Fasting Insulin, HOMA-B, HOMA-IR**
Cardiovascular / ECG	QT interval, RR interval, Heart Rate (ECG), Heart Rate (pulse)
Cardiovascular / hypertension	**Repeat Diastolic BP, Repeat Systolic BP**
Cardiovascular / Biochem	**HDL, LDL, TG, TC**
Cardiovascular / Biochem	**sCRP, homocysteine, ApoA1, ApoB**
Blood Biochem	**Hemoglobin**, MCH, MCHC, MCV, PCV, PC, RBC, WBC, IL6
Renal Function	Uric Acid, Bicarbonate,Creatinine, Phosphate, Sodium, Urea, Cortisol
Liver Function	Albumin, Alkaline Phosphatase, Bilirubin, Gamma Glutamyl Transpeptidase
Lung Function	**Forced Expiratory Volume, Forced Vital Capacity, FEV/FVC Ratio**
Dynamic	Grip strength ,Ever broken bone, Eye preference, Handedness

Fig 1: UK10K Strategy

Statistical analyses

In the first stage of analyses, we tested single variants with MAF>1% for association with 46 quantitative traits using SNPTEST and sets of coding (cds) and coding with functional consequence (csq) variants using two regression-based methods (SKAT and ridge regression) on 12 known loci with common and/or rare variants for HDL, LDL, Triglycerides (TG) and Total Cholesterol (TC) using MAF cutoffs of 0.1%,1%, 5% and ALL and non-overlapping windows of 10 and 50 SNPs (Table 1).

Table 1: Sets of cds and csq variants for twelve known lipid genes.

Trait	Gene name	No SNPs ALL (cds/csq)	No SNPs MAF<0.05 (cds/csq)	No SNPs MAF<0.01 (cds/csq)	No SNPs MAF<0.001 (cds/csq)	Chr	Gene start	Gene end	Known variants Common/rare
TG	ANGPTL3	10/6	10/6	10/6	8/4	1	63063187	63071180	Common+Rare
TG	APOB	387/230	372/218	365/212	145/95	2	21224301	21266945	Common+Rare
TG	GCKR	23/20	22/19	22/19	15/14	2	27719706	27746550	Common+Rare
TG, HDL	LPL	14/7	11/6	8/4	7/4	8	19796582	19824770	Rare(TG), Common(TG, HDL)
HDL, TC	ABCA1	64/35	55/31	50/29	40/24	9	107543283	107690518	Common(HDL, TC), Rare(HDL)
TG	ANGPTL5	12/7	12/7	10/6	9/5	11	101761405	101787253	Rare
TG	APOA5	5/2	3/1	2/0	2/0	11	116660086	116663136	Rare
HDL, TC, LDL, TG	CETP	17/12	16/11	13/9	11/8	16	56995762	57017757	Common
HDL	LCAT	7/5	7/5	5/4	5/4	16	67973787	67978015	Common+Rare
HDL, TC	LIPG	10/6	9/5	8/4	7/3	18	47088427	47119278	Rare(HDL), Common(HDL, TC)
TG, HDL	ANGPTL4	11/8	9/7	9/7	6/5	19	8429011	8439257	Rare(TG), Common(HDL)
LDL, TC	LDLR	36/11	29/11	26/10	23/9	19	11200038	11244492	Common

Results

Singlepoint analyses

Genotype accuracy is high and there is no evidence for inflation of association summary statistics (Fig 2). Several variants were found confirming previously known genome-wide association results, including CETP for HDL, UGT1A1 for Bilirubin, SLC2A9 for Uric acid, ABO for alkaline phosphatase, APOE for Total cholesterol and LDL and HBS1L-MYB for several blood traits.

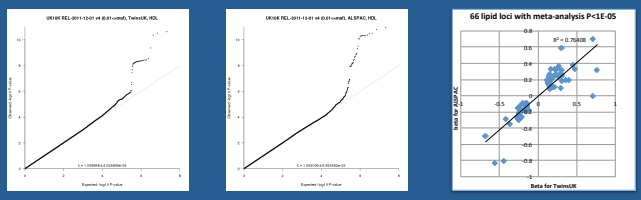


Fig 2: QQplots for results of singlepoint analysis of HDL (MAF>1%) Left TwinsUK Right: ALSPAC

Rare variant analyses

We tested sets of coding variants and coding variants with functional consequence from 12 known lipid genes for association with HDL, LDL, TG and TC using ridge regression (lambda 10) and SKAT. Suggestive associations between CETP and HDL (Fig 3), LPL and TG and GCKR and TC were found (p<0.001).

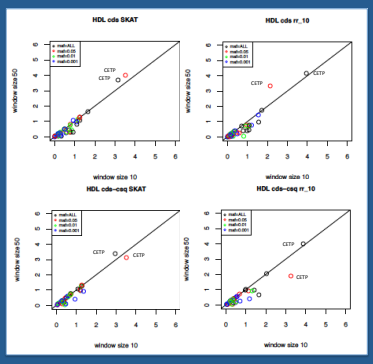


Fig 3: Rare variant analyses with ridge regression and SKAT using cds and csq sets of variants. Variant sets were filtered for different MAF cut-offs and analysed in fixed windows of 10 and 50 variants. CETP is significant (p<0.001) for ALL and MAF<0.05 for both collapsing tests. Three cds/csq contributing SNPs (p<0.05) are in moderate LD (r2<0.13,D'>0.33) with GWA top signal rs1800775. Conditioning on rs1800775 reduces significance from pval $1e^{-4}$ to $1e^{-3}$ for SKAT and ridge regression (window size 50, MAF=ALL). Remaining signal overlaps with secondary GWA signal.

Conclusions

Initial novel signals are being followed up. Currently analysis of a final set of 3,910 samples is under way, and variants discovered through genome-wide sequencing of the TwinsUK and ALSPAC cohorts are being imputed into the full genotyped cohorts to increase the power of the UK10K study. Our results provide insights into how large-scale whole-genome sequencing efforts are likely to reveal for the genetic architecture of complex traits and might significantly increase the initial estimates of explained heritability.

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, Maria Soler Artigas; **BGI Shenzen:** Jing Tian, Jun Wang, Sifei He, Yingrui Li; **EBI:** Graham Ritchie, Paul Flicek; **University of Oxford:** Jonathan Marchini