2 Methods

Disclaimer

The UK10K project is conducted in a collaborative nature. The WGS sequencing data was produced by a dedicated data production team, with similar strategies and tools as those used for 1000GP. My contribution included helping with WGS data QC, being the single major person for creating UK10K imputation reference panel and its evaluation, and conducted all the statistical analysis for all of the 13 CVD traits unless for a few centrally run analyses which will be explicitly mentioned throughout the according chapters.

2.1 Introduction

There are two major topics for this thesis: I first describes the development and evaluation of a novel imputation panel based on WGS dataset from the UK10K cohorts arm (Chapter 3), and then focus on phenotype-genotype associations for three separate trait groups where both sequenced and imputed data are used (Chapters 4-6). The three trait chapters (Chapters 4-6) employ similar data and analytical approaches, therefore, I describe here in this Methods chapter the generation and generalised analytic details of WGS based association studies for analyses applied in these chapters. Many of these methods were proposed and adopted centrally by the UK10K study (The UK10K Consortium 2015) so as to effectively handle multiple analyses and to allow cross-comparison of association results. For specific methods that are only applied to one or a small number of traits, I will further describe them in the method section of each of the three trait chapters (Chapters 4-6).

2.2 Study samples

Here I provide summary information of all cohorts that contributed to the analyses described in Chapters 3-6. Additional information relative to specific phenotype traits are given within the respective chapters.

2.2.1 UK10K WGS cohorts

ALSPAC. The Avon Longitudinal Study of Parents and Children (ALSPAC) is a long-term health research project. More than 14,000 mothers enrolled during pregnancy in 1991 and 1992, and the health and development of their children has been followed in great detail ever since (Golding et al. 2001). A random sample of 2,040 study participants was selected for WGS. The ALSPAC Genetics Advisory Committee approved the study and all participants gave signed consent to the study.

TwinsUK. The Department of Twin Research and Genetic Epidemiology (DTR), is the UK's only twin registry of 11,000 identical and non-identical twins between the ages of 16 and 85 years (Moayyeri et al. 2012). The database used to study the genetic and environmental aetiology of age-related complex traits and diseases. The St Thomas's Hospital Ethics Committee approved the study and all participants gave signed consent to the study.

2.2.2 UK10K GWA cohorts

For ALSPAC, a total of 8,365 samples were genotyped in Illumina 550k. Besides the WGS samples, there were another 6,557 samples available (Bonnelykke et al. 2013). For TwinsUK, there were another 2,575 samples that were unrelated to the sequence dataset (IBS>0.125) with genotypes on Illumina HumanHap300 or Illumina Human610 arrays (Soranzo et al. 2009). Imputed TwinsUK data, although unrelated to those samples selected for WGS, did contain related individuals (mainly co-twins) which would require an association test that adjusts for the relatedness. Both datasets passed QC criteria (gender check, heterozygosity, European ancestry, relatedness (ALSPAC) and zygosity (TwinsUK).

Variants discovered through WGS of the TwinsUK and ALSPAC cohorts were imputed into the full GWAS genotyped cohorts. Of note, for TwinsUK, 2,040 samples were genotyped in Illumina317K and 3,614 samples were gentoyped in Illumina610k. the 317K SNP array was first imputed to the 610K SNP array and then the two datasets were merged to create a single dataset with 610K SNPs. Typically, the two recommended approaches to deal with two SNP-arrays from two different genotyping platforms are: 1. Keep only those common SNPs and create a single dataset, which usually remove a lot of SNP data from at least one of the two panels. 2. Impute the two SNP arrays separately and perform all downstream analyses separately. For TwinsUK, I evaluated various designs and eventually adopted a third option, to impute TwinsUK 300K to 600K so that I got a single dataset with 600K SNPs for downstream imputation and evaluation. This was made possible because the following two reasons: first, more than 95% of SNPs in the 300K panel is in the 600K panel. So, the 300K panel is almost an exact subset of 600K. The design of Illumina SNP panels is mainly based on tagging approach, which is different from Affymetrix's random selection approach. I found out that the haplotypes tagged by the 300K SNPs are almost identical to those tagged by the 600K SNP panel. Second, there are more than 400 twin-pairs where one twin is in 300K panel while the other is in 600K panel. This made imputation from 300K to 600K with very high accuracy. After adopting this imputation approach, I run association studies for a few traits by adding a dummy variable to indicate the status of being in the 300K or 600K panel, and found that the results were almost identical as that obtained without using the dummy variable.

2.2.3 Expanded discovery cohorts

1958 Birth Cohort. Participants to the cohort have been followed-up regularly since birth with prospective information collected on a wide range of indicators related to health, health behaviour, lifestyle, growth and development. There have been 9 contacts with the participants since their birth (ages 7, 11, 16, 23, 33, 41, 45, 47, and 50 years). The biomedical survey at age 45 years included collection of blood samples and DNA from about 8000 participants. The survey was approved by the South East multicentre research ethics committee (MREC). There was an informed consent process conducted by the National Centre for Social Research (Power and Elliott).

INGI-Val Borbera. The INGI-Val Borbera population is a collection of 1,785 genotyped samples collected in the Val Borbera Valley, a geographically isolated valley located within the Appennine Mountains in Northwest Italy (Traglia et al.). The valley is inhabited by about 3,000 descendants from the original population, living in 7 villages along the valley and in the mountains. Participants were healthy people 18-102 years of age that had at least one grandfather living in the valley. A standard battery of tests were performed by the laboratory of ASL 22 - Novi Ligure (AL), on sera from fasting blood collected in the morning. The project was approved by the Ethical committee of the San Raffaele Hospital and of the Piemonte Region. All participants signed an informed consent.

INGI FVG. The INGI Friuli Venezia Giulia (FVG) cohort comprised of about 1700 samples from six isolated villages covering a total area of 7858 km^2 in a hilly part of Friuli-Venezia Giulia (FVG) county located in north-eastern Italy (Esko et al.). Genotyping and phenotypic data for 1590 samples are available. Participants were randomly selected people 3-92 years of age. People with age < 18 were excluded from analyses. Ethics approval was obtained from the Ethics Committee of the Burlo Garofolo children hospital in Trieste. Written informed consent was obtained from every participant to the study.

INGI Carlantino. Carlantino is a small village in the Province of Foggia in southern Italy. Genetic analyses of chromosome Y haplotypes as well as mitochondrial DNA show that Carlantino is a genetically homogeneous population and not only a geographically isolated village (Lanzara et al. 2015). Participants were randomly selected in a range of 15 – 90 years of age. Genotyping and phenotypic data are available for 630 individuals. People with age < 18 were excluded from analyses. The local administration of Carlantino, the Health Service of Foggia Province, Italy, and ethical committee of the IRCCS Burlo-Garofolo of Trieste approved the project. Written informed consent was obtained from every participant to the study.

INCIPE. For the INCIPE study, 6200 randomly chosen individuals, all Caucasians and at least 40 years of age as of 1 January 2006, received a letter inviting them to participate in the study. A total of 3870 subjects (62%) accepted and were enrolled. Two studies were included in the analysis: **1.** INCIPE1: Individuals genotyped on Affymetrix 500k**; 2.** INCIPE2: Individuals genotyped on HumanCoreExome-12v1**.** The ethics committees of the involved institutions approved the study protocol.

The **Ludwigshafen Risk and Cardiovascular Health (LURIC) study.** The LURIC study is a prospective study of more than 3,300 individuals of German ancestry in whom cardiovascular and metabolic phenotypes (CAD, MI, dyslipidaemia, hypertension, metabolic syndrome and diabetes mellitus) have been defined or ruled out using standardised methodologies in all study completed participants. A 10-year clinical follow-up for total and cause specific mortality has been completed. (Winkelmann et al.) From 1997 to 2002 about 3,800 patients were recruited at the Heart Center of Ludwigshafen (Rhein). Inclusion criteria were: German ancestry, clinical stability (except for acute coronary syndromes) and existence of a coronary angiogram. Exclusion criteria were: any acute illness other than acute coronary syndromes, any chronic disease where non-cardiac disease predominated and a history of malignancy within the last five years. The study was approved by the ethics review committee at the Landesärztekammer Rheinland-Pfalz in Mainz, Germany, and written informed consent was obtained from the participants.

CBR: Cambridge BioResource: CBR is a collection of pseudo-anonymised DNA samples from 8,000 healthy blood donors that has been established in 2008 and 2010 by the NIHR funded Cambridge Biomedical Research Centre in collaboration with NHS Blood and Transplant for use in genotype-phenotype association studies (Dendrou et al. 2009). Four thousand donors each were enrolled during 2007 and 2009. Full blood counts (FBCs) were obtained from EDTA anticoagulated samples of blood drawn from the pouches of the donation collection sets. FBCs performed on an ABX Pentra 60 automated haematology analyser (ABX Diagnostics, Montpellier, France) or on a Sysmex XE-2100. For the purpose of calibration measurements, 500 blood samples were performed on both the Beckman-Coulter and Sysmex instruments. Measurements were performed between 16-24 hours after phlebotomy.

HELIC-MANOLIS. The HELIC (Hellenic Isolated Cohorts; www.helic.org) MANOLIS (Minoan Isolates) collection focuses on Anogia and surrounding Mylopotamos villages. Recruitment of this population-based sample was primarily carried out at the village medical centres. All individuals were older than 17 years and had to have at least one parent from the Mylopotamos area. The study includes biological sample collection for DNA extraction and lab-based blood measurements, and interview-based questionnaire filling. The phenotypes collected include anthropometric and biometric measurements, clinical evaluation data, biochemical and haematological profiles, self-reported medical history, demographic,

socioeconomic and lifestyle information. The study was approved by the Harokopio University Bioethics Committee and informed consent was obtained from every participant.

HELIC-Pomak. The HELIC (Hellenic Isolated Cohorts; www.helic.org) Pomak collection focuses on the Pomak villages, a set of isolated mountainous villages in the North of Greece. Recruitment of this population-based sample was primarily carried out at the village medical centres. The study includes biological sample collection for DNA extraction and lab-based blood measurements, and interview-based questionnaire filling. The phenotypes collected include anthropometric and biometric measurements, clinical evaluation data, biochemical and haematological profiles, self-reported medical history, demographic, socioeconomic and lifestyle information. The study was approved by the Harokopio University Bioethics Committee and informed consent was obtained from every participant.

TEENAGE. Participants were drawn from the TEENAGE (TEENs of Attica: Genes and Environment) study. A random sample of 857 adolescent students attending public secondary schools located in the wider Athens area of Attica in Greece were recruited in the study from 2008 to 2010. Our sample comprised 707 (55.9% females) adolescents of Greek origin aged 13.42 ± 0.88 years. Details of recruitment and data collection have been described elsewhere (Ntalla et al.). Prior to recruitment all study participants gave their verbal assent along with their parents'/guardians' written consent forms. The study was approved by Harokopio University Bioethics Committee and the Greek Ministry of Education, Lifelong Learning and Religious Affairs.

LOLIPOP: London Life Sciences Prospective Population Study **(**LOLIPOP) is an ongoing community cohort of approximately 30,000 individuals aged 35-75 years, recruited in West London, UK to study the environmental and genetic factors that contribute to cardiovascular disease among UK Indian Asians. The study includes both European and Indian Asian subjects. For the current study, only white individuals were included in the primary meta-analysis. Three studies were included in the analysis: **(1).** LOLIPOP - EWA: European whites from the general population, genotyped on Affymetrix 500K arrays. **(2).** LOLIPOP - EWP: European whites from the general population, genotyped on Perlegen custom array. **(3).** LOLIPOP - EW610: European whites from the general population, genotyped on Illumnia Human610 array.

FENLAND: The Fenland Study is a community-based cohort of individuals born between 1950 and 1975 and residing in East Cambridgeshire or Fenland, UK. The goal of the

Fenland Study is to study the interactions between diet, lifestyle, and genetic factors and risk of diabetes and obesity.

FHS: The Framingham Heart Study started in 1948 with 5,209 randomly ascertained participants from Framingham, Massachusetts, US, who had undergone biannual examinations to investigate cardiovascular disease and its risk factors. In 1971, the Offspring cohort (comprising 5,124 children of the original cohort and the children's spouses) and in 2002, the Third Generation (consisting of 4,095 children of the Offspring cohort) were recruited. FHS participants in this study are of European ancestry. The methods of recruitment and data collection for the Offspring and Third Generation cohorts have been described (Feinleib et al. 1975).

The **Precocious Coronary Artery Disease Study (PROCARDIS) cases and controls cohorts**: The PROCARDIS (Clarke et al. 2009) study consists of coronary artery disease (CAD) cases and controls from four European countries (UK, Italy, Sweden and Germany). CAD (defined as myocardial infarction, acute coronary syndrome, unstable or stable angina, or need for coronary artery bypass surgery or percutaneous coronary intervention) was diagnosed before 66 years of age and 80% of cases had a sibling fulfilling the same criteria for CAD. Subjects with self-reported non-European ancestry were excluded. Among the “genetically-enriched” CAD cases, 70% had suffered myocardial infarction (MI). In the UK, patients were identified from hospital records used previously to recruit patients for large-scale trials of cholesterol-lowering therapy. Patients were identified in Italy through hospitals that had collaborated in the GISSI studies, in Sweden through existing registries of cases that had contracted MI at a young age or through the central database of the Stockholm County Council, and in Germany through the PROCAM and related databases. Controls with nopersonal or sibling history of CAD before age 66 years were contemporaneously recruited using the same infrastructure. For each of the CAD cases, one control was recruited of the same sex,ethnicity and within 5 years of age, with no personal or sibling history of CAD before age of 66 years.

Women’s Health Initiative (**WHI**): WHI is one of the largest (n=161,808) studies of women's health ever undertaken in the U.S (The Women’s Health Initiative Study Group 1998). There are two major components of WHI: (1) a Clinical Trial (CT) that enrolled and randomized 68,132 women ages 50 – 79 into at least one of three placebo-control clinical trials (hormone therapy, dietary modification, and calcium/vitamin D); and (2) an Observational Study (OS) that enrolled 93,676 women of the same age range into a parallel

prospective cohort study. A diverse population including 26,045 (17%) women from minority groups were recruited from 1993-1998 at 40 clinical centers across the U.S. The design has been published (Anderson et al. 2003, Hays et al. 2003). For the CT and OS participants enrolled in WHI and who had consented to genetic research, DNA was extracted by the Specimen Processing Laboratory at the Fred Hutchinson Cancer Research Center (FHCRC) using specimens that were collected at the time of enrollment in to the study (between 1993 and 1998).

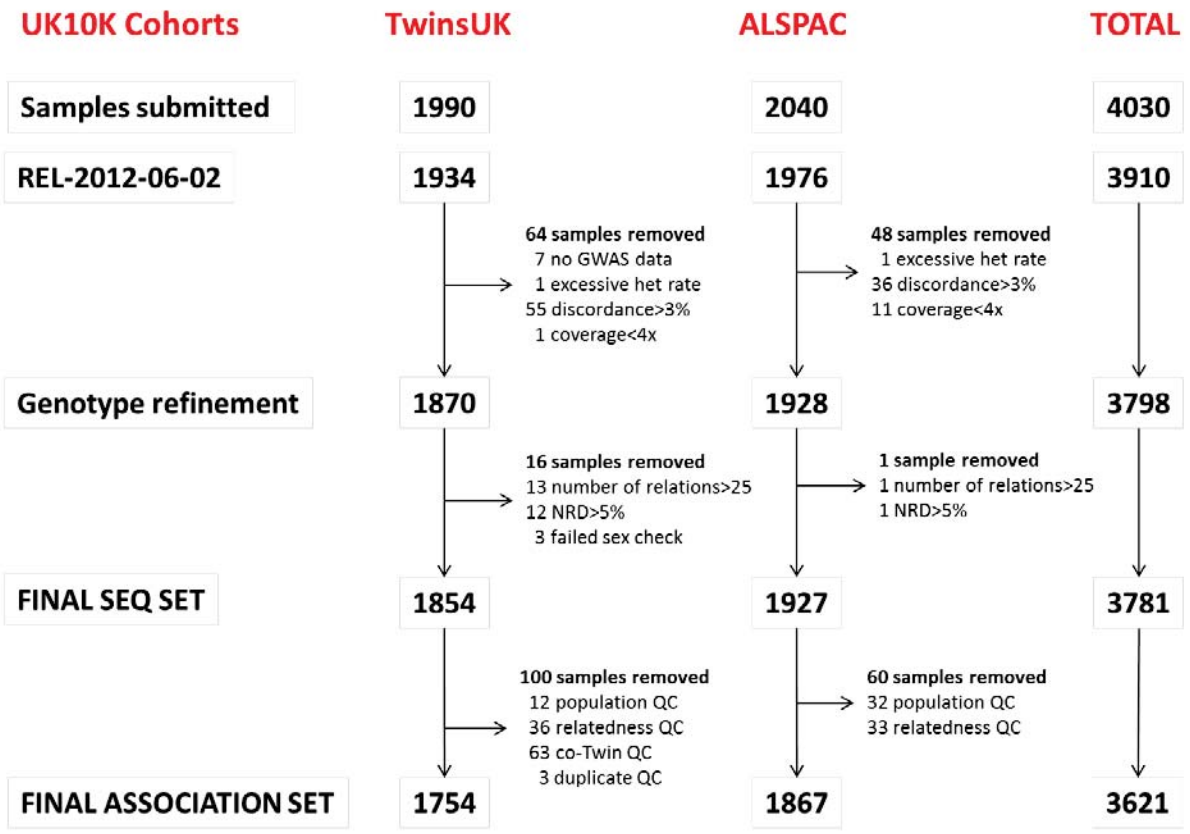
2.3 Genetic data

2.3.1 UK10K WGS data

The details of UK10K WGS data production was presented in the UK10K flagship paper supplementary (The UK10K Consortium 2015). In summary, low read-depth WGS was performed at both the Wellcome Trust Sanger Institute (WTSI) and the Beijing Genomics Institute (BGI) from Jan 2011 to March 2012. The data production was done with similar procedures as that for the 1000GP (Abecasis et al. 2012), and was almost fully handled by a dedicated data production team within UK10K. My contribution included re-phasing of the UK10K WGS data using SHAPEIT v2 (Delaneau et al. 2013) to generate an improved imputation reference panel and investigating the batch effects between samples assayed at the two sequencing centers: WTSI vs. BGI. The motivation and procedures for re-phasing the UK10K WGS data will be presented in chapter 3. For investigating batch effects, I used multidimensional scaling analysis (MDS) on a pruned set of independent markers (n =2,203,581). Based on this work, a total of 335,982 SNVs with significant association with sequencing centre ($P \leq 0.01$) were removed, resulting ~42 million single nucleotide variation (SNV) and ~3.5 million InDels. The number of variants excluded due to potential batch effects resulting from two sequencing center comprised less than 1% of the total number of variants. Nevertheless, this exclusion could be avoided by adding sequencing center as a covariate in the downstream association studies. For a total of 3,910 samples that had WGS performed, 3,798 went to genotype refinement step and 3,781 are in the final dataset for

UK10K formal release. These 3,781 samples made the dataset used for imputation reference panel. Finally, 3,621 of these 3,781 samples were included for association studies, after excluding those samples of non-European ancestry or failed relatedness check (**Figure 2.1**).

Figure 2.1 UK10K WGS samples data production



2.3.2 Imputation using WGS reference panel

There are ~9,000 samples (6,557 for ALSPAC and 2,575 for TwinsUK) that have genome-wide SNP-array data but don't have WGS data. These samples were imputed into the full set of WGS variants, initially by using the UK10K WGS reference panel alone. Later on, with the availability of a new software functionality (IMPUTE version 2.1.3 and later) and after a comprehensive evaluation, I designed a preferred imputation strategy to impute these ~9,000 samples and many more external cohorts. Details of the imputation evaluation and selection of final strategy were described in Chapter 3. As listed in **section 2.2.3** earlier, a few genetic isolates from Italy and Greece were used as expanded discovery cohorts and they were imputed using the same strategy designed for non-isolates. Population isolates have reduced phenotypic, environmental and genetic heterogeneity, and rare variants present in the founders drift up in frequency as the population expands. These characteristics make genetic isolates preferable for the detection of rare variants associated with complex traits (Zeggini 2014). The success of using population isolates to discover common and rare variants were exemplified in association studies conducted in the Icelandic population (Holm et al. 2011), the Greenlandic founder population (Moltke et al. 2014), and Finnish population (Lim et al. 2014), and the Greek isolates (Tachmazidou et al. 2013).

2.4 Phenotype harmonization

In genome-wide genotype-phenotype association studies, the curation of genetic data is given a large amount of attention, given the large data volume, high cost for sequencing, and lengthy computational process for data production and QC. However, phenotype data is equally important and its harmonization is a key for the design and success of the association studies as well. Many published GWAS intended to use simplified approaches, usually a logarithm transformation of phenotype and an adjusting on age and sex and sometimes principle components. For the UK10K project in general and the traits that I analysed, a more comprehensive phenotype harmonization process was implemented. A particular reason for this comprehensive approach was that the TwinsUK phenotypes were measured by different

analysts and instruments and spanned across a few years due to historical reasons. Therefore, extra consideration was needed to address potential batch effects for these phenotypes.

For each of the 13 CVD traits in TwinsUK, I manually examined the statistical distribution to determine the appropriate threshold for outlier exclusion and identified the best fit transformation (natural log, inverse normal, square root, inverse, or non-transformed). Then I evaluated the list of confounding covariates that need to be adjusted for (including age, age*age, sex, BMI, batch effect). All these covariates were fit into a linear model and only those significantly associated with the traits are included in the linear regression model. To address the confounding effect of instruments and dates of visits, I created a categorical variable that combined the information of these two variables and then added this categorical variable into the linear mixed model as a random effect. When inverse-normal transformation was used, the samples were divided into males and females for transformation and covariates adjustment separately. **Figure 2.2** showed four snapshots of the phenotype harmonization results for RBC trait. As shown in panel B, there was an instrumental effect for the raw phenotype. After adjusting for batch effects and other cofounding factors, the regressed and standardized residuals followed a normal distribution. The use of standardization as the last step of the phenotype harmonization facilitated meta-analysis and cross-traits examination of effect sizes. The general outline applied for phenotype harmonization was summarized in **Figure 2.3**.

Figure 2.2 Evaluation of batch effects and trait distribution

An example of assessing batch effects within the TwinsUK RBC trait. A) Raw trait distribution; B) Trait value per individual as a function of measurement date (x-axis) and instrument type (coded in 4 colours); C) Linear mixed modelling with covariates; D) Distribution of harmonized phenotype residuals.

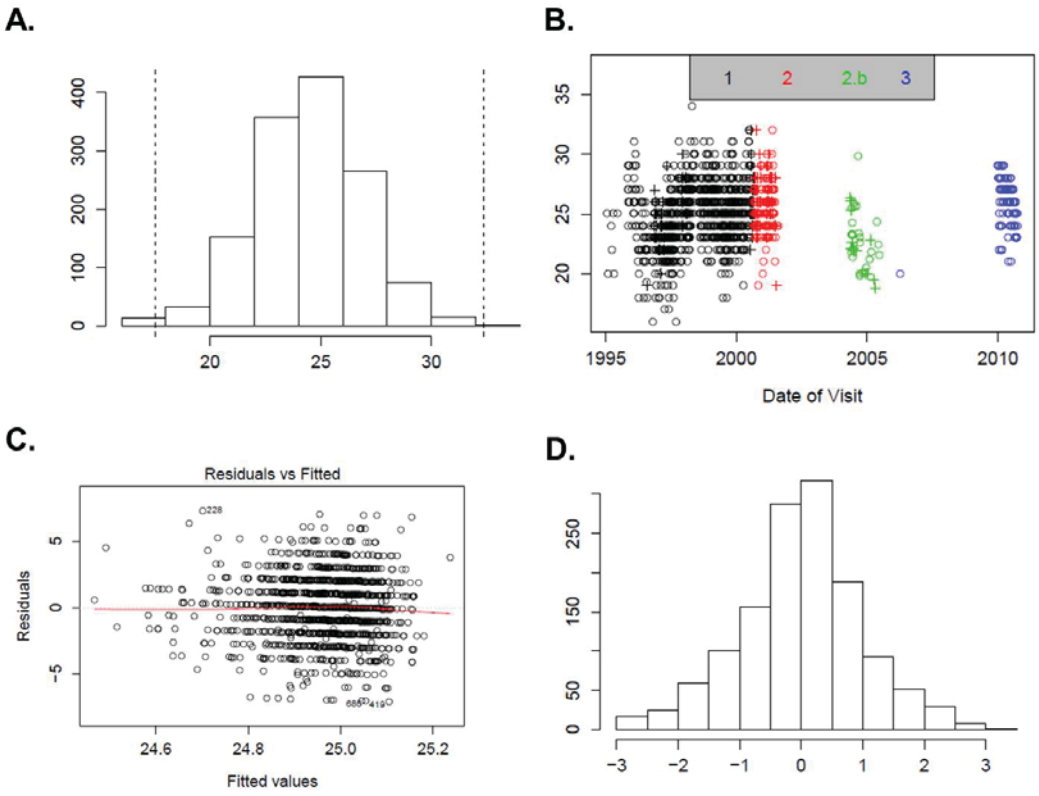
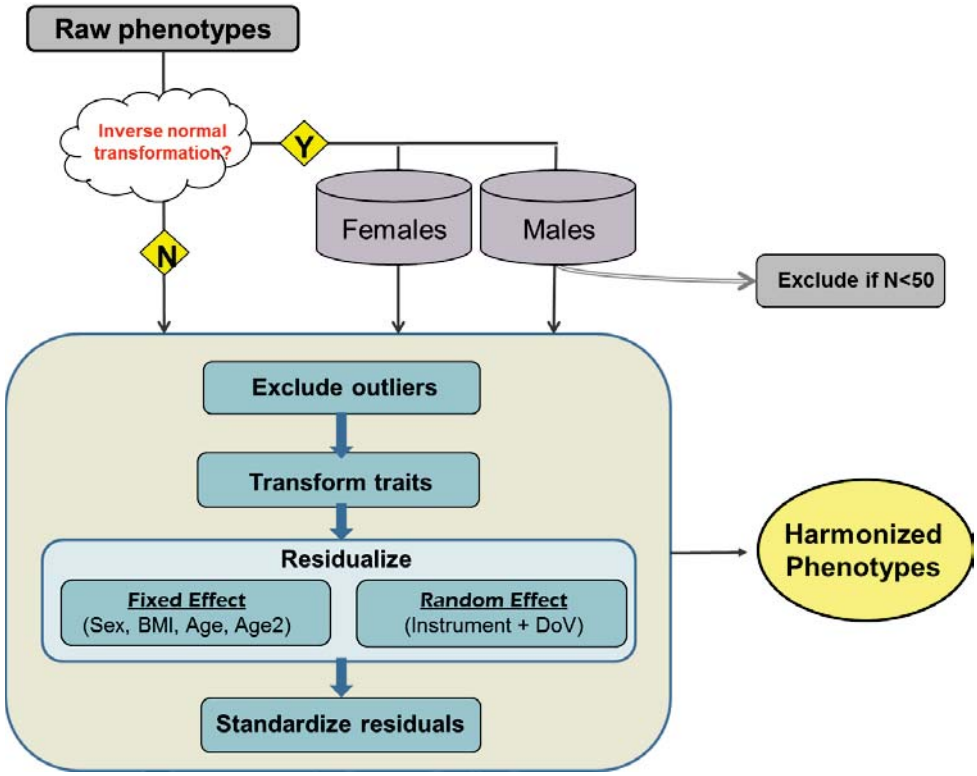


Figure 2.3 Phenotype harmonization protocol

The first step is to identify outlier filtering threshold and decide a transformation metrics. The next step is to adjust for potentially confounding factors, which includes age, age^2, gender, and body mass index (BMI), dependent on trait. All these covariates are fit into a linear model and only those significantly associated with the traits are included in the final model. When inverse-normal transformation is used, the samples are divided into males and females for transformation and covariates adjustment separately.



2.5 Statistical methods for association studies

Compared to GWAS based on SNP array data, statistical challenges for WGS data include but not limited to: choices of statistical tests, selecting analysis intervals from whole genome, statistical methods for structural variations, correcting for population stratification and family relatedness at rare variants, and adjusting for multiple testing. There are well established methods for estimating and correcting for population stratification for common variants (McCarthy et al. 2008), but there is not yet an established assessment for low frequency and rare variants. Over the course of the UK10K project, a few high throughput computational pipelines were developed to analyse many traits in parallel. These standardised protocols enforce consistent statistical approaches and facilitate the parallel evaluation of a large number of quantitative traits.

2.5.1 Power estimation

Power for single marker tests was calculated based on the non-centrality parameter of the chi-squared distribution, i.e., $NCP = 2(N-1)p(1-p)\beta^2 r^2$ (Chapman et al. 2003, Spencer et al. 2009), where N is the sample size, p is the minor allele frequency (MAF), β is the standardised effect of a SNV on a continuous phenotype (standardised so that β is the effect per standard deviation of the phenotype), and r^2 is the square of the correlation between a true genotype and a genotype measured with error. The UK10K study calculated power from a non-central chi-squared distribution for the a genome-wide significance threshold of 1.1E-08, the estimated genome-wide significance for WGS studies (Xu et al.), for a range of values of r, and for sample size N=3,621 (The UK10K Consortium 2015) (**Figure 2.4a**). This significance threshold takes into account the large number of variants identified by WGS. **Figure 2.4a** showed that the low pass WGS design had 80% power to detect associations of SNVs of low frequency and rare down to ~MAF 0.5%, for alleles with Betas ≥ ~1.2 standard deviations. This is a MAF range poorly tagged by older-generation imputation panels based on HapMap. **Figure 2.4a** also shows sizable reductions in the magnitude of the effect sizes that can be identified at any sample size through use of the UK10K reference panel, when added to the 1000GP panel. For instance, for a variant of MAF

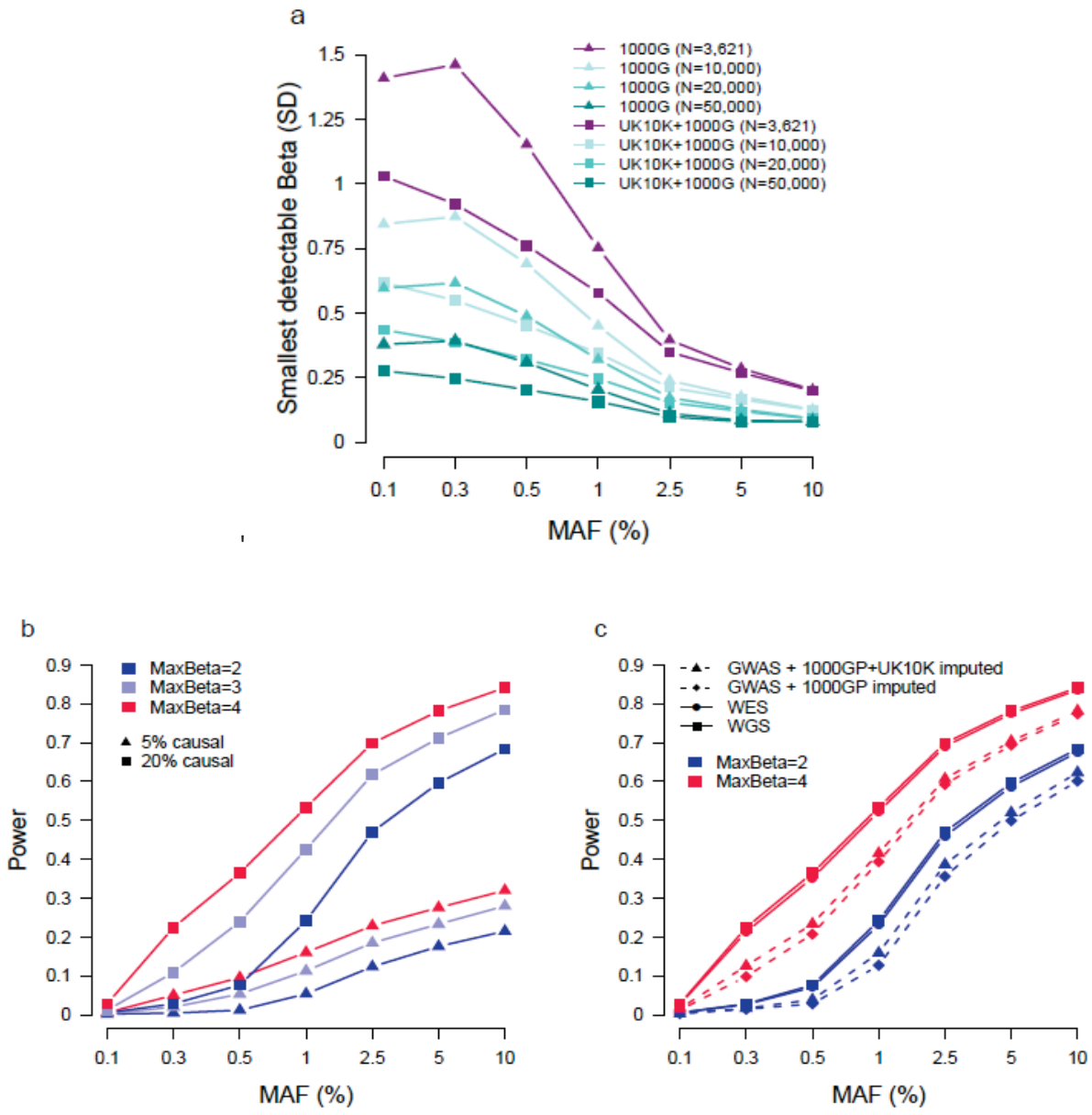
= 0.3%, there is equivalent power when imputing from UK10K+1000GP into a 3,621 sample as when using the 1000GP imputation panel alone with 10,000 samples.

Power for the SKAT rare variant tests (Wu et al. 2011, Lee et al. 2012) was calculated by assuming a causal model for the relationship between the SNVs and the phenotype. The calculation used ten regions of 30 variants randomly sampled from each autosome, and then genotype errors were randomly added to the data following observed r^2 values between genotypes from data imputed from different sources (WGS, high depth WES**,** GWAS+imputation against 1000GP, GWAS+imputation against the combined reference panel of 1000GP and UK10K), and matching the MAF of each variant using the same parameters as in **Figure 2.4b.** Relative power is the ratio of the power with $r^2 = 1$ divided by power when $r^2 < 1$.

Figure 2.4 Power calculation in the UK10K cohorts

This plot is adopted from the UK10K main paper (The UK10K Consortium 2015), made by Klaudia Walter.

a. Strength of single-variant associations detectable at 80% power as a function of MAF and sample size. **b.** Power of region-based tests in the UK10K-cohorts sample. Evaluations assume N=3,621, alpha = 6.7×10^{-8} and that the proportion of causal variants in the regions is either 5% or 20%, for maximum association (maxBeta) in a region =2,3,4. **c**. Power of region-based tests and the impact of genotype imputation, with the proportion of causal variants in the regions set to 20%.



2.5.2 Single-variant based association studies

One of the most powerful tools for the analysis of genome-wide data has been a single marker based test of association with one degree of freedom. For variants with MAF ≥ 0.1%, I conducted single marker based association test genome-wide for each of the studied traits, first on WGS data and then on imputed data. The exclusion of variants with MAF < 0.1% is based on statistical power calculation. Each variant was fitted into a regression model, where the independent variant is standardized phenotype residuals (with covariate regressed out) and the dependent variable is genotype dosage. The genotype dosage represents the predicted dosage of the non-reference allele given the data available, i.e. the probability of being heterozygote plus two times of the probability of being non-reference allele homozygote . It has a value between 0 and 2 and gives an indication of how well the genotype is supported by the imputation process of the sequence data. Genotype dosage has also been used in SNP array based GWAS to account for imputation uncertainly. Although WGS data was supposed to be directly assayed, the WGS data obtained from low-depth sequencing had gone through imputation process to derive the final genetic reads.

For unrelated samples

For unrelated samples (including ALSPAC WGS and most population based cohorts in the expanded discovery and replication), I used SNPTEST v 2.4.0 (Marchini et al. 2007) to conduct single marker based analysis on genome-wide scale. SNPTEST was used in many GWAS studies including the landmark WTCCC 2007 study (Wellcome Trust Case Control 2007). I used the option of "-frequentist 1" for the additive model, "-method expected" for using genotype dosage, and "-use_raw_phenotypes" to disable the default quantile normalization since the phenotype residuals were already standardized. For each single marker i, the statistical model is expressed as: $y_i = \beta_0 + \beta_1 x_i + e$.

For related samples

For samples with relatedness (TwinsUK imputed samples and genetic isolates), I used GEMMA v0.94 (Zhou and Stephens 2012) to conduct single marker based association test. GEMMA uses a standard linear mixed model that takes familiar relatedness into consideration. This makes exact genome-wide association analysis computationally practical

and approximations unnecessary. Before running GEMMA for association analysis, I first used GEMMA to calculate a kinship matrix with the centered genotype model, based on the genome-wide SNP array data. By default, GEMMA filters out variants with missingness > 0.05, MAF <0.01, r2<0.9999. I used "-maf 0 -miss 1 -r2 1" to force all variants to be included for analysis.

Meta-analysis of single marker summary statistics

The WGS and imputed cohorts that I used present an ideal scenario for meta-analysis, because all cohorts were imputed to the same reference panel and went through the same protocol of phenotype harmonization (including outlier exclusion, transformation, covariates regression, and standardization). Meta-analyses of individual cohort summary statistics were performed using GWAMA v 2.1 (Magi and Morris 2010), which was based on a fixed effect model. Compared to another widely used meta-analysis software - METAL (Willer et al. 2010), GWAMA has the following advantages: (i) random effect model included; (ii) output two heterogeneity statistics, the Cochran's *Q* statistics and I^2; (ii) perform genomic control correction for the meta-analyzed statistics as well as on individual GWAS. The statistical calculation of effect B_j and variance V_j for GWAMA is given as below, where β_{ij} represents the effect of the reference allele at the *j*-th single marker in the *i*-th study, and w_{ij} represents the inverse of the variance of the estimated allelic effect:

$$B_j = \frac{\sum_{i=1}^{N} \beta_{ij} w_{ij}}{\sum_{i=1}^{N} w_{ij}} \qquad V_j = \left(\sum_{i=1}^{N} w_{ij}\right)^{-1}$$

2.5.3 Loci selection for single marker results

I conducted loci selection for single marker based analyses, first for WGS results and then for meta-analysis results, in the following steps:

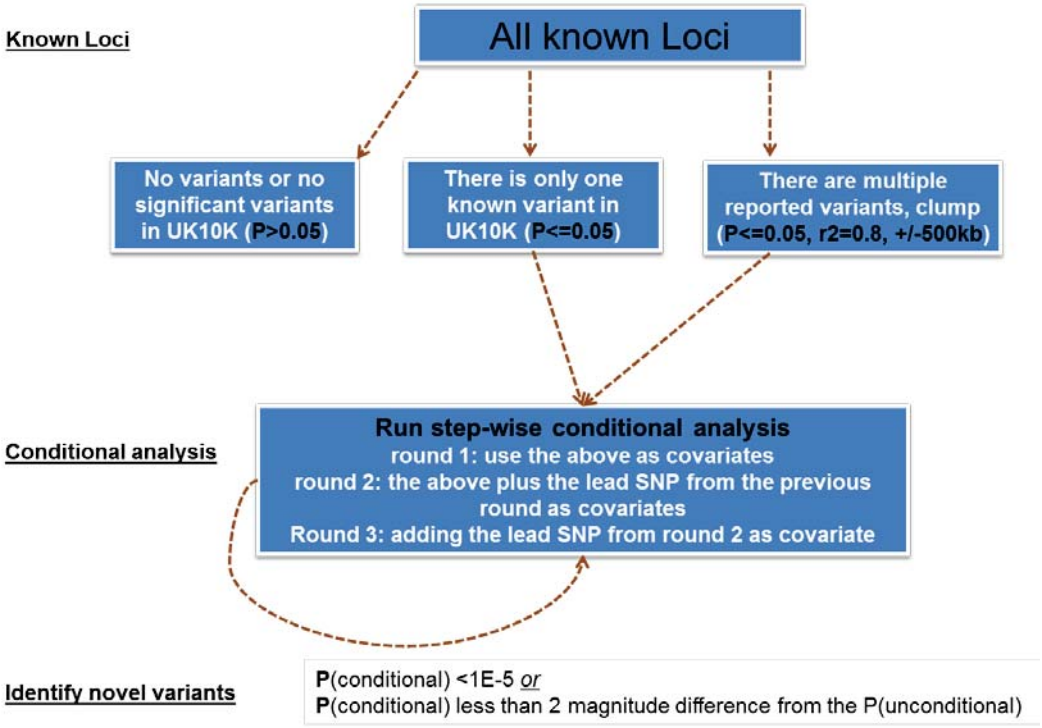
1. For each studied trait, I compiled a list of published variants as positive controls by selecting all SNPs associated with a trait of interest from the NHGRI GWAS Catalog (http://www.genome.gov/gwastudies/) ($P \leq$ 5E-08 last updated in May 2014),

supplemented by manual curation of all associations reported in the literature reaching the same significance threshold.

2. I then identified significant and borderline significance variants from single marker based tests. The genome-wide significant threshold was set as 5.0E-08, while the borderline significant threshold was set as 1.0E-06 and 1.0E-07 for WGS and meta-analysis respectively. These thresholds were chosen to select a reasonable number of SNPs for further follow-up. Also, several of the phenotype specific QQ plots showed some evidence of a change-point at approximately these thresholds.
3. For all variants selected above, I run sequential conditional analyses to identify putative novel variants, conditional on all positive controls of the same traits within 1Mb of the top variants. I only included those positive controls with at least a marginal significance in the UK10K project (P<0.05). Where a known locus reported multiple correlated variants, I clumped the set of variants to remove highly correlated ones (using a LD metric r2 >0.8 applied to within a 1MB sliding window from each known index SNP). This avoided collinearly errors when a variant is conditioned against multiple highly correlated variants. In the initial round of conditional analysis, all selected top variants were conditioned on the clumped known variants if there was any known variant within 1Mb. In further rounds, associations were conditioned against the same set of known variants plus the variant with the most significant P value identified in the previous round of conditional analysis. The conditional analysis was tested independently for each cohort and a meta-analysis was conducted at the end of each round until the conditional association P value was no longer significant (P >1E-05). The steps for this sequencing conditional analyses was summarized in **Figure 2.5**.
4. A variant was considered independent if the conditional $P \leq$ 10-5 or it is less than 100 times of the unconditional P. Variants were classified as **known** (denoting either a known variant, or a variant for which the association signal disappears after conditioning on the known locus) or **novel** (denoted as variant which still is significant after conditional on known loci). For novel signals, the variant with the lowest conditional P between multiple associated variants was reported.
5. Some of the studied traits have the full GWAS results publically available. For example, the full GWAS results of lipids are posted at http://csg.sph.umich.edu/locuszoom/. For any putative novel lipids variants that

survived the above steps, I run clumping analysis to make sure that the novel variants to be reported are not tagged by any of the publically posted variants with even a modest association (P <0.01).

Figure 2.5 Flow of step-wise conditional analysis



2.5.4 Rare variants aggregation analysis

Due to the nature of low frequency of rare variants, traditional single marker based analysis lacks power (Asimit and Zeggini 2010). A better alternative is to collapse or to aggregate rare variants within a functional unit, for example, a gene or pathway. Then the aggregated functional unit could be fit into a regression model just as that done in the single marker based association test. The simplest such approach is the burden test (Morgenthaler and Thilly 2007, Li and Leal 2008). Various burden tests exist and they differ mainly in the way that they take into account allele frequencies of individual variants and whether they take weighted combinations of variants based on *a priori* information (Price et al. 2010). However, burden tests are limited for their assumptions that all or most rare variants within each tested unit influence the phenotypes in the same direction with the same magnitude (unless known weights are incorporated). They have been shown poor statistic power across most plausible allelic architectures, where many common and rare variants within a region have little or no effect and when there are a combination of variants with opposite effects (Ladouceur et al. 2012).

Some other aggregation methods did not assume that all tested rare variants act in the same direction, including the C-alpha test (Neale et al. 2011), SKAT (Wu et al. 2011) and the estimated regression coefficient test (EREC) (Lin and Tang 2011). For the traits that I studied, I used SKAT-O that runs both SKAT and burden tests (Lee et al. 2012). SKAT is a variance-component multiple regression test which retains power in settings where neutral variants or variants with opposite direction of effects could result in loss of power. SKAT-O represents the best linear combination of SKAT and burden tests, which is supposed to maximize power. Therefore, the SKAT-O statistics is generally more significant than SKAT. I excluded singletons or variants with MAF>1% from SKAT and SKAT-O tests. For those variants whose SKAT *P* is very close to SKAT-O *P*, the associations would be predominantly driven by a single rare variant within the window, which is insensitive to burden test. For lipids and CRP that have WGS data in both TwinsUK and ALSPAC, meta-analyses of summary statistics was performed using MetaSKAT v0.27 with default options (Lee et al. 2012). Klaudia Water did the variants selection and window selection, which served as a central resource for rare variants aggregation tests for all UK10K traits. The SKAT-O tests were run by grouping variants in the following three ways:

Genome-wide: The availability of WGS data opens a window for conducting rare variants aggregation tests across the genome, even though there is still a lack of good strategy to group rare variants outside of gene regions. Mainly as an exploratory experiment, the UK10K project designed an agnostic approach where ~1.8 million windows of equal size (3kb) were constructed across the entire genome, with one window overlapping with the next by half. This approach is agnostic to function and therefore has less power to detect true signals than those with reliable prior knowledge of genomic function, but it has the potential to capture groups of putatively functionally correlated rare variants within any regulatory feature. On average, each sliding window has 35 variants. Based on simulation studies, the genome-wide significance threshold for this approach is P <6.8E-08.

Exome-wide: For exome-wide tests, all variants in exons, untranslated regions (UTRs) and essential splice sites were included and were given equal weight of being causal. Through this approach, a total of 50,746 windows were constructed for 26,212 genes from GENCODE v15 (Harrow et al. 2012). Each window has an average of 35 variants and a maximum of 50 variants. Based on simulation studies, the genome-wide significance threshold for this approach is P <1.2e-6.

Functional variants based: These tests only included missense variants and those predicted to be loss of function. Across the genome, 15,528 gene windows were constructed, each with five or more missense and loss of function variants. On average there are 17 variants per gene.

2.5.5 Loci selection for rare variant aggregation results

In general, there is a lack of optimal approach for following up regions of interest identified by rare variants aggregation tests. First, there is a lack of independent WGS cohorts that could be used for replication, because usually external WGS cohorts would want to get their primary discovery published before serving as a replication cohort. Secondly, unlike SNP array data, the number of variants in each rare variants aggregation window is different among different cohorts, due to the difference of allele frequencies especially for rare variants and due to sequencing quality and QC filtering. Therefore, a same window would include different set of variants across multiple cohorts. For the traits that I studied, I only managed to get replication data for lipids traits. The strategy for replication will be detailed in chapter 4.

2.5.6 Other statistical methods

Besides association analyses that aimed to identify single variants or single gene regions of interest, a few more statistical analyses were conducted to explore some general properties of allelic architecture of the studied traits.

2.5.6.1 Percentage of variance explained

Under an evolutionary neutral model, variance explained (VE) follows a uniform distribution as a function of MAF, meaning that variants with MAF < X% explain X% of heritability. In reality, however, lots of traits are related to fitness and have been under natural selection to some extent (Visscher et al. 2012). Therefore, it's interesting to quantify the VE for biomedically relevant traits such as the CVD traits included in this thesis. Morrison et al. estimated that common variants (MAF > 1%) explain 61.8% (SE = 14.2) of the variance in HDL levels and rare variants (MAF < 1%) explain an additional 7.8% (SE = 9.8) of the variance. However, due to the small sample size and the large SE, this estimation needs to be confirmed.

The UK10K study used the Restricted Maximum Likelihood (REML) method implemented in GCTA (http://www.complextraitgenomics.com/software/gcta/reml.html) (Yang et al. 2010) to estimate phenotypic variance explained by SNV sets in the UK10K WGS data (The UK10K Consortium 2015). It used SNV with MAF≥1% and calculated VE for variants from different reference panels: i.e., HapMap2 (Variant N=2,331,713), Hapmap3 (N=1,168,695), 1000GP (N=7,475,230) and the entire UK10K reference panel (N=8,317,582). There was evidence for improvement in VE with increasing SNV density for a subset of the traits including lipids. While only reaching suggestive levels of associations given power, those loci are enriched for true associations as shown from the FDR values, potentially informing prioritization strategies for follow-up studies. This finding provided a basis for focusing attention on low frequency and rare variants selected using more liberal *P* value thresholds.

2.5.6.2 Fine mapping of known loci and functional enrichment analysis

GWAS have been increasingly fruitful in discovering genotype-phenotype associations. The mechanisms underlying these associations, however, are still largely unknown as only a small fraction of these SNPs directly alter protein-coding genes. The interpretation of functional consequences of non-coding variants has been greatly enhanced by large-scale efforts to identify regulatory genomic regions (e.g ENCODE and NIH Roadmap Epigenome Project). It is expected that a more accurate classification of enrichment patters might lead to biological insights and help prioritise variants for follow-up studies. Common approaches for integrating GWAS with functional data are the so called enrichment analyses, which take genetic variants statistically important to a phenotype and characterise the degree to which they appear in various genomic regions. Characterizing the non-random patterns of association of GWAS signals to functional information is important at least for two reasons. Firstly, characterizing enrichment patterns for a given phenotype with a given non-coding mark in a given cell provides insights into (potentially unknown) biological processes. Secondly, it can provide rules for interpreting putative functional consequences of genetic variants and for designing follow-up experiments.

For functional enrichment analysis, genomic fine-mapping was usually conducted first to select a most informative subset of SNPs that are predicted to contain the causal variants. It is well accepted that the SNPs showing the strongest association are not necessarily the causal variants, due to sampling variation and LD. Nevertheless, the dense coverage of the WGS increased the likelihood that causal variants are assayed. Bayesian fine-mapping approaches have been widely used to narrow down a credible set of putative causal variants, which could then be used for studying functional insights. In a recent fine-mapping and enrichment analysis study on T1D (Onengut-Gumuscu et al. 2015), the Bayesian approach was found to be more informative than the r2-based approach to select credible sets of SNPs, where SNPs in the credible sets were found to be strongly enriched in enhancer chromatin states in immunologically relevant tissues. The same fine-mapping method (Wellcome Trust Case Control et al. 2012) was also used in my study.

After choosing an informative set of SNPs through fine-mapping, choosing an informative set of functional annotations relevant to the studied traits is also important. Recently, a novel hierarchical model for jointly analyzing GWASs and genomic annotations was proposed, which uses association statistics computed across the genome to identify

classes of genomic elements that are enriched with or depleted of loci influencing a trait (Pickrell 2014). When applied to 18 diseases and traits including lipids and hematological traits, this model was shown able to identify the relevant types of genomic information from a set of 450 genome annotations.

Fine mapping of known loci

For the known regions of each trait, the availability of WGS data provided an opportunity for fine-mapping, so as to identify functional and potentially causal variants. I used the fine-mapping method described by Maller and colleagues (Wellcome Trust Case Control et al. 2012), which was based on Bayesian linear additive modelling. The Bayes' factors (BF) for each SNP in a fine-mapped region were multiplied to obtain a joint BF measure of association, with the assumption that cohorts are independent. These BFs are then used to calculate posterior probabilities, based on the assumption that there is exactly one causal SNP in each region. In addition, 95% and 99% credible sets are constructed in order to assess the uncertainty of the fine-mapping analysis. BF ratios are also computed as the ratio between each variant in the region of interest and the best scoring (fine-mapped) variant. This measure allows for direct inference on the usefulness of the fine-mapping experiment between various variants sets (e.g. UK10K vs 1000GP vs HapMap data). Also, a BF ratio between each variant and each positive control is computed to show the relative advantage of the fine-mapped variant when compared to the currently reported variant.

The boundaries of each region were chosen to be at a distance of at least 0.1 centi-morgan either side of the positive control variants. In Maller's original paper, two additional conditions were used to expand these boundaries, namely to include variants in LD with the positive control of $r^2 > 0.2$ and variants with *P* value within 2 orders of magnitude of the positive control *P* value. However, since the original paper reported that in almost all cases these two conditions did not change the boundaries, I did not implement these two additional conditions. For all variants predicted to be causal, their annotation information is added, based on the Variant Effect Predictor (VEP) tool from Ensembl (McLaren et al. 2010). Functional variants are defined as falling into one of these eight categories: frameshift_variant, stop_gained, splice_donor_variant, splice_acceptor_variant, missense_variant, inframe_deletion, inframe_insertion, initiator_codon_variant, stop_lost.

2.6 Conclusion & Discussion

After a few years into WGS based studies, many of the methods described in this chapter now become quite standard with ready-to-use software and tools. However, there is still a lot more to be explored in terms of statistical methods and data integration, in order to get the most out of a rich collection of WGS data. The following are a few recommended approaches/practices based on my ~3 years of work on the UK10K project:

1. Maximize power with better quality genetic data and larger sample size. Given that WGS samples are still costly to get, datasets with much larger sample sizes could be added to the analysis by optimized imputation approaches. To boost sample size, I combined the genetic data for TwinksUK WGS and imputed samples together so that the co-Twins could also be included for analysis. Otherwise, they would violate the independent nature of different cohorts and be excluded. For lipids traits, I found this approach significantly increased power, where positive controls become more significant with the combined approach. This approach of combining WGS and imputed samples was adopted for full blood counts traits and CRP but not for lipids, because the sample size was relatively larger for lipids and the association studies for lipids traits were conducted at a much earlier stage.
2. Given that functional annotation for a large portion of the full genome is limited, it is necessary to combine agnostic hypothesis-free approaches with targeted approaches. For example, the genome-wide SKAT-O tests took an agonistic approach while the exome-wide SKAT-O tests utilized existing knowledge to include only functional variants within gene regions.
3. Use consistent terminology and software across the project. For example, use CHRPOS instead of rsID as the identifier of genetic variants because rsID could evolve over the time and sometimes ambiguous. Many mainstream software have the same underlying algorithm and conduct the same calculation. For example, both METAL and GWAMA does inverse variance based meta-analysis. While each research has his/her own preference, it is recommended to use one to assume the consistency of input and output files.

Many of the methods and approaches described in this chapter are derived from the framework for the overall UK10K projects. For a large-scale collaborative project like this one, I did manage to work independently and also collaboratively. For those centrally adopted methods, I run the analyses for all of the traits that were included in this thesis, unless explicitly credited to others. I also developed slightly different approaches where they are appropriate.

First, in the UK10K flagship paper, only the UK10K reference panel is used for imputation, which led to an exclusion of ~4.3 million variants due to batch effect and failing of other QC metrics. For my traits, I used the UK10K plus 1000GP panel for imputation. Most of those variants excluded from the UK10K alone panel did exist and passed QC in 1000GP and were therefore included in the imputed datasets and downstream meta-analyses. One reason for this design difference is that the software functionality for merging reference panels was developed at a rather later stage. The number of samples is much larger for my studied traits as well. The UK10K project reported association results based on WGS plus the imputed samples in the remaining part of TwinsUK and ALSPAC. However, for the CVD biomarkers that I studied, there are many more cohorts included in the meta-analysis, for example, a total of 14 for lipids.

Second, my strategy for loci selection is different from that used in the UK10K main study. The UK10K study first run clumping to narrow down a list of index SNPs and then run conditional analysis. This was because clumping is a well-established procedure, while conditional analysis was brought into the project much later after a rather extensive discussion on the selection of software and the decision of various thresholds. I included all variants passing a liberal significance threshold (P<1E-7 in meta-analysis) for conditional analysis. This avoids filtering out too many variants in the clumping step. The LD clumping is based on UK10K WGS data only, which could be accurate for the UK10K main study, but might not be accurate when my study included many non-UK cohorts. My approach of conditional analysis was further boosted by using the raw genotype and phenotype data of all participating cohorts, instead of using summary statistics as that done in GCTA.

Finally, the significance threshold that I used is different. In the UK10K main study, variants with P <1E-5 in WGS were selected for initial *in-silico* follow-up. Then those reaching P <1E-7 in the meta-analysis were considered as top hits. In my study, for WGS results, I put a more stringent threshold of P <1E-06 and took forward only those variants with MAF <5%, which might not be well imputed. For meta-analysis, I only applied one

threshold *P*<1E-7 without limiting to those having a certain level of significance level in WGS (such as WGS *P* <1E-5 used in UK10K flagship paper). This is because the WGS sample is now much smaller compared with the total number of samples in my meta-analyses. Also, it is practical and cost-effective to follow-up a lot more variants through *in-silico* methods.