Whole-genome Sequencing-based Association Studies of Cardiovascular Biomarkers

Jie Huang

This dissertation is submitted to the University of Cambridge
Faculty of Biology for the degree of Doctor of Philosophy

Darwin College

University of Cambridge

February 2015

PREFACE

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text.

This dissertation is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University of similar institution except as declared in the Preface and specified in the text.

The dissertation does not exceed the page limit of 300 specified by the Biology Degree Committee

ABSTRACT

Background: Genome-wide association studies (GWAS) have significantly advanced the genetic study of complex human traits. With the advent of whole-genome sequencing (WGS) technologies and the increased capacity to identify rare variants, GWAS that use WGS data are expected to provide further opportunities for the discovery of variants that have larger and even causal effects. The UK10K project is one of the largest studies that use WGS to investigate the contribution of low frequency and rare genetic variants to medical traits.

Research aims: My research aims to address the utility of WGS-based imputation and associations for identifying the genetic determinants of a select quantitative traits that are associated with cardiovascular risks. Under the UK10K project framework, I study a suite of circulating biomarkers that have been reported for association with CVD. Specifically, I seek to evaluate the following three broad aspects: 1. what are the characteristics of phasing and imputation with WGS data? 2. what novel analytic methods could be applied to a large scale WGS based association study on a rich of phenotypes? 3. can I identify novel and potentially stronger effect genetic variants that are associated with the chosen CVD traits?

Methods: My study leverages existing WGS data from the UK10K project (N = ~4,000) and further uses it as a reference to impute more samples (N > 10,000) that have genome-wide SNP array data. In doing so, I first evaluate the quality of the WGS data and its utility for imputation, by comparing it to WGS data from the 1000 Genomes Project. Then, I examine the associations between genotypes and phenotypes for 13 quantitative traits, first in samples having WGS and then in samples having imputed data. The 13 CVD related biomarkers include four lipid traits (high-density lipoprotein cholesterol (HDL), low-density lipoprotein cholesterol (LDL), total cholesterol (TC), triglycerides (TG)), one inflammatory biomarker (C-reactive protein (CRP)), and eight haematological traits (hemoglobin (HGB), mean corpuscular hemoglobin (MCH), mean corpuscular hemoglobin concentration (MCHC), mean corpuscular volume (MCV), packed cell volume (PCV), platelet counts (PLT), red blood cell counts (RBC), white blood cell counts (WBC)).

To my dear parents: YuanYu Huang & Youquan Deng

To my wife: Weilin Chen

To my daughter Valerie and my son Jimmy

ACKNOWLEDGEMENTS

From the three years of my PhD study there are a lot of people I want to mention and thank for giving me help and advice. The first and most important person is Dr. Nicole Soranzo, my PhD thesis supervisor who gave me an opportunity to come to Cambridge and Sanger Institute to pursue my PhD study and work on a most important and frontier project on WGS. Her dedication to science and attentive guidance to trainees inspires me to become a good researcher and scientific leader myself.

I would like to thank the other faculty members who served on my thesis committee and provided strategic guidance for my PhD project: Dr. Richard Durbin, Dr. Carl Anderson, Dr. Eleftheria Zeggini from Sanger Institute, and Dr. Adam Butterworth from University of Cambridge.

I would like to thank our wonderful teammates (team 151): Lu Chen, Klaudia Walter, Louella Vasquez, Valentina Iotchkova, Massimiliano Cocca, Matthias Geihs, Yasin Memari, So-Youn Shin. Over the past three years, we sit close together and worked even closely with each other. Whenever there is a question, I feel that I could simply turn around my chair and get insightful feedback and help instantly.

I would also like to thank the larger community of the UK10K project and outsider collaborators who contributed data for our large meta-analysis, particularly to Drs. Nic Timpson and Josine Min from Bristol University.

Finally, I would like to thank my former supervisor, Dr. Chris O'Donnell at Framingham Heart Study, for offering me research experience in cardiovascular genetics and continuous support for my PhD study and career growth.

Table of Contents

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

1000GP	1000 Genomes Project
ADH	Autosomal dominant hypercholesterolemia
ALSPAC	Avon Longitudinal Study of Parents and Children
Apo-A1	apolipoprotein A-I
Apo-B	apolipoprotein B
Apo-E	apolipoprotein E
AMD	age-related macular degeneration
BF	Bayes' factor
BGI	Beijing Genomics Institute
BP	blood pressure
CAD	coronary artery disease
CBR	Cambridge BioResource
CHD	coronary heart disease
CNV	copy number variation
CKD	chronic kidney disease
CRP	C-reactive protein
CVD	cardiovascular disease
DALYs	disability-adjusted life years
DHS	DNaseI hypersensitive sites
EAF	effect allele frequency
EMR	electronic medical records
ERFC	Emerging Risk Factors Collaboration
FHS	Framingham Heart Study
FVG	Friuli Venezia Giulia
GWAS	genome-wide association studies
HDL	high-density lipoprotein
HGB	hemoglobin
HELIC	HELlenic Isolated Cohorts study
HMM	hidden markov model
HWE	hardy-weinberg equilibrium
IBD	identify by descent

IBS	identify by state
InDel	insertion/deletion polymorphism
INGI	Italian Network of Genetic Isolates
LD	linkage disequilibrium
LDL	low-density lipoprotein
LMT	lipid modification therapies
LoF	loss of function
LOLIPOP	London Life Sciences Population study
LURIC	Ludwigshafen Risk and Cardiovascular Health
MAF	minor allele frequency
HGB	haemoglobin
MCH	mean corpuscular hemoglobin
MCHC	mean corpuscular hemoglobin concentration
MCV	mean cell volume
MDS	multidimensional scaling
MI	myocardial infarction
MR	mendelian randomisation
OR	odds ratio
PCA	principle component analysis
PCV	packed cell volume
PLT	platelet count
PROCARDIS	Precocious Coronary Artery Disease Study
QC	quality control
RBC	red blood cell
RCT	reverse cholesterol transport
SKAT	sequence kernel association test
SKAT-O	sequence kernel association test - optimized
SNP	single nucleotide polymorphism
SNV	single nucleotide variation
TC	total cholesterol
TFBS	transcription factor binding sites
TG	triglycerides
TSS	transcription start site

TwinsUK	UK Adult Twin Registry
UK10K	10,000 UK genome sequencing project
UTR	untranslated regions
VB	Val Borbera
WBC	white blood cell
WGS	Whole Genome Sequencing
WTCCC	Wellcome Trust Case Control Consortium
WTSI	Wellcome Trust Sanger Institute

PUBLICATIONS ARISING FROM THIS DISSERTATION

- ** Co-first author*
- *For papers with more than 10 authors, my name is listed together with the first 3 and the last 3 authors. When there are more than 3 co-starred first-authors, all of them are listed.*

1. Gormley P*, Downes K*, **Huang J***, Kettunen J, Aki S, …, Palotie A, Ripatti S, Soranzo N. A polygenic panel of platelet-associated SNPs is associated with risk of incident ischaemic stroke. (*submitted*)
2. Walter K*, Min M*, **Huang J***, Lucy Crooks*, …, Timpson NJ, Durbin R , Soranzo N. The UK10K project: rare variants in health and disease. *(under revision)*
3. **Huang J**, Howie B, Memari M, …, Timpson NJ, Marchini J, Soranzo N, UK10K Project. A reference panel of 3,781 genomes from the UK10K Project increases imputation performance over the 1000 Genomes Project. *Nature Communications*. (*accepted*)
4. Taylor P, Porcu E, Chew S, … **Huang J**, …, Soranzo N, Timpson NJ, Wilson S, the UK10K Consortium. Whole genome sequence based analysis of thyroid function. *Nature Communications*. 2015 Mar 6;6:5681
5. Timpson NJ, Walter K, Min JL, …, **Huang J**, ..., Humphries SE, Zeggini E, Soranzo N; UK1OK consortium members. A novel low-frequency variant near APOC3 is associated with plasma triglyceride and VLDL levels in Europeans. *Nature Communications*. 2014 Sep 16;5:4871
6. O'Connell J, Gurdasani D, Delaneau O, …, **Huang J**, …, Soranzo N, Sandhu MS, Marchini J. A general approach for haplotype phasing across the full spectrum of relatedness. *PLOS Genetics* 2014 Apr 17;10(4):e1004234

PUBLICATIONS ARISING ELSEWHERE (from 2012-01 to 2015-01)

- * *Co-first author*
- *For papers with more than 10 authors, my name is listed together with the first 3 and the last 3 authors. When there are more than 3 co-starred first-authors, all of them are listed.*

1. Baumert J*, **Huang J***, McKnight B*, Sabater-Lleal M*, Steri M*, …, Strachan DP, Peters A, Smith NL. No evidence for genome-wide interactions on plasma fibrinogen by smoking, alcohol consumption and body mass index: results from meta-analyses of 80,607 subjects. *PLoS One*. December 31, 2014 DOI: 10.1371
2. Shin SY, Fauman EB, Petersen AK, …, **Huang J**, …, Kastenmüller G, Spector TD, Soranzo N. An atlas of genetic influences on human metabolism. *Nature Genetics* 2014 Jun;46(6):543-50
3. Han B, Luo H, Raelson J, **Huang J**, Li Y, Tremblay J, Hu B, Qi S, Wu J. TGFBI (βIG-H3) is a diabetes risk gene based on mouse and human genetic studies. *Hum Mol Genet*. 2014 Apr 11
4. **Huang J**, Huffman JE, Yamkauchi M, …, Lowenstein CJ, Strachan DP, O'Donnell CJ; CHARGE Consortium Hemostatic Factor Working Group. Genome-wide association study for circulating tissue plasminogen activator levels and functional follow-up implicates endothelial STXBP5 and STX2. *Arteriosclerosis,Thrombosis, and Vascular Biology*. 2014 Feb 27
5. Sabater-Lleal M*, **Huang J***, Chasman D*, Naitza S*, Dehghan A*, …, Strachan DP, Hamsten A, O'Donnell CJ. Multiethnic meta-analysis of genome-wide association studies in >100 000 subjects identifies 23 fibrinogen-associated Loci but no strong evidence of a causal association between circulating fibrinogen and cardiovascular disease. *Circulation*. 2013 Sep 17;128(12):1310-24.
6. **Huang J**, Liu Y, Welch R, Willer C, Hindorff LA, Li Y. WikiGWA: an open platform for collecting and using genome-wide association (GWA) results. *European Journal of Human Genetics.* 2013 Apr;21(4):471-3
7. O'Seaghdha CM, Wu H, Yang Q, …, **Huang J**, …, Bonny O, Fox CS, Bochud M. Meta-analysis of genome-wide association studies identifies six new loci for serum calcium concentrations. *PLoS Genetics*. 2013 Sep;9(9):e1003796
8. Kleber ME, Seppälä I, Pilz S, …, **Huang J**, …, Lehtimäki T, März W, Meinitzer A. Genome-wide association study identifies three genomic loci significantly associated with

serum levels of homoarginine – The AtheroRemo Consortium. *Circulation Cardiovascular Genetics.* 2013 Sep 18

9. McGrath LM, Cornelis MC, …, **Huang J**, …, Sullivan P, Perlis RH, Smoller JW. Genetic predictors of risk and resilience in psychiatric disorders: a cross-disorder genomewide association study of functional impairment in major depressive disorder, bipolar disorder, and schizophrenia. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*. 2013 Sep 13
10. ALSGEN Consortium, Ahmeti KB, Ajroud-Driss S, …, **Huang J**, …, Veldink JH, Yang Y, Zheng JG. Age of onset of amyotrophic lateral sclerosis is modulated by a locus on 1p34.1. *Neurobiology of Aging*. 2013 Jan;34(1):357.e7-19
11. Chen M-H*, **Huang J***, Chen W-M, Larson MG, Fox CS, Vasan RS, Seshadri S, O'Donnell CJ, Yang Q. Using family-based imputation in genome-wide association studies with large complex pedigrees: the Framingham Heart Study. *PLoS ONE*. 2012;7(12):e51589. (*co-first author)
12. **Huang J**, Sabater-Lleal M, Asselbergs FW, …, Liu Y, O'Donnell CJ, Hamsten A. Genome-wide association study for circulating levels of plasminogen activator inhibitor-1 (PAI-1) provides novel insights into the regulation of PAI-1. *Blood.* 2012 Dec 6;120(24):4873-81
13. **Huang J**, Ellinghaus D, Franke A, Howie B, Li Y. 1000 Genomes-based imputation identifies novel and refined associations for the Wellcome Trust Case Control Consortium phase 1 data. *European Journal of Human Genetics.* 2012 Jul;20(7):801-5
14. Bis JC, DeCarli C, Smith AV, …, **Huang J**, …, Launer LJ, Ikram MA, Seshadri S; Cohorts for Heart and Aging Research in Genomic Epidemiology Consortium. Common variants at 12q14 and 12q24 are associated with hippocampal volume. *Nature Genetics*. 2012 Apr 15;44(5):545-51.
15. Willour VL1, Seifuddin F, Mahon PB, …, **Huang J**, …, Gurling H, Purcell S, Smoller JW, Craddock N, DePaulo JR Jr, Schulze TG, McMahon FJ, Zandi PP, Potash JB. A genome-wide association study of attempted suicide. *Mol Psychiatry.* 2012 Apr;17(4):433-44.

1 Introduction

1.1 The burden of cardiovascular disease in modern society

Over the past few decades, improved sanitation and medical advances have led to a considerable decrease in mortality from infectious diseases. At the same time, chronic conditions such as cardiovascular disease (CVD) became the principal cause of mortality in the developed world (Kuller 1976). Although mortality from CVD has been decreasing, it is still the number one cause of mortality among chronic diseases. CVD refers to all the diseases of the heart and circulation system, including coronary heart disease (CHD), stroke, angina, heart attack, congenital heart disease. CHD and stroke are the two most common forms of CVD and both are mainly caused by atherosclerosis, a condition where arteries become narrowed by a gradual build-up of fatty material (i.e., atheroma) within artery walls. When the arteries become too narrow and there is inadequate oxygen-rich blood delivered to the heart, it causes angina, manifested by a pain or discomfort in the chest. When an atheroma or part of it in the arteries breaks away, it causes clotting in the circulation and cutting off the supply of oxygen-rich blood to heart muscle, leading to myocardial infarction (MI), commonly known as heart attack. When the blood clot blocks an artery that carries blood to the brain, it causes an ischaemic stroke. Another form of stroke is haemorrhagic stroke, caused by the rupture of a blood vessel in the brain.

Based on the World Health Organization's report of global status on non-communicable diseases (year 2010), an estimated 17.3 million people died from CVD in 2008, representing 30% of all global deaths. It was projected that that this number would reach 23.3 million by 2030, making CVD remain to be the single leading cause of death over the next decade. For the two most common forms of CVD, CHD and stroke accounted for an estimated 7.3 million and 6.2 million of the total death respectively. Over 80% of CVD deaths take place in low- and middle-income countries. CVD is responsible for 10% of Disability-adjusted life years (DALYs) lost in low- and middle-income countries and 18% in high-income countries. DALYs is used more often to estimate the total burden of a disease, as opposed to simply count the number of resulting deaths.

1.2 Established and emerging risk factors for CVD

The term “risk factor” was first coined in Dr. Kannel’s 1961 report of the association between circulating low-density lipoprotein cholesterol (LDL) and CVD (Kannel et al. 1961). Risk prediction is mainly used for disease prevention, defined as actions directed to avoid illness and promoting health to reduce the need for secondary and tertiary health care. Risk factors are important for assessing disease risk and therefore for disease prevention, while intermediate phenotypes usually reflect disease progression and are important markers for disease intervention and treatment. Risk factors were usually first identified through epidemiological studies. For example, the Framingham Heart Study (FHS) used a prospective design and identified age, male sex, smoking status, diabetes mellitus, hypertension, and serum cholesterol level as the most important risk factors for developing CVD (Dawber et al. 1959, Kannel et al. 1964). The INTERHEART study is based on a case-control design and reported a longer list of factors that account for most of the MI risk in 52 countries (Yusuf et al. 2004). There are more than 100 risk factors reported for association with CVD (Brotman et al. 2005). The criteria for being an established CVD risk factor include: a significant independent impact on the risk of CVD, a high prevalence in many populations, and a reduced level of CVD by the treatment and control of the risk factor. LDL is the first established risk factor for CVD. The decrease in mortality from CVD since 1980s was closely associated with lowering underlying risk factors especially LDL, which accounted for more than one-third of the observed decrease in mortality from CHD (Hunink et al. 1997).

Classical CVD risk factors include dyslipidemia (Kannel et al. 1961, Anderson et al. 1987), hypertension (Kannel et al. 1980), obesity (Lavie and Milani 2003), smoking (Service. 1983, Lavie and Milani 2003, Yusuf et al. 2004, Teo et al. 2006), alcohol drinking (Stampfer et al. 1988, Rimm et al. 1991), and physical inactivity (Pate et al. 1995). New risk factors include inflammatory markers especially C-reactive protein (CRP) (Koenig et al. 2004, Cushman et al. 2005), heamostasis markers such as figrinogen (Kannel et al. 1987), white blood cell count (WBC) (Kannel et al. 1992), homocysteine (Selhub et al. 1995), lipoprotein (a) (Bostom et al. 1996, Helfand et al. 2009), and uric acid (Kim et al. 2010) (**Figure 1.1**). CRP and WBC will be described in detail in later chapters. Risk factors initiated the atherosclerotic process and continued to be present throughout the cardiovascular disease continuum (CVDC). The concept of CVDC was originally described by Dzau and colleagues

in 1991 (Dzau and Braunwald 1991), later on validated by clinical evidence of improved patient outcomes (Dzau et al. 2006). In CVDC, a chain of events are precipitated by several risk factors, which eventually cause end-stage heart failure and death if untreated (**Figure 1.2)**. Most CVD could be prevented by addressing modifiable risk factors such as smoking, unhealthy diet and physical inactivity, hypertension, and dyslipidemia.

Risk factors have been used to estimate the onset of both non-fatal and fatal cardiovascular events through the calculation of a risk score. Among them are the Framingham risk score (Wilson et al. 1998), the Joint British Societies risk charts (British Cardiac et al. 2005), the ASSIGN score (Tunstall-Pedoe et al. 2006), the Systematic COronary Risk Evaluation (SCORE) risk charts (Graham et al. 2007), and the Reynolds Risk Score (Ridker et al. 2007). There are differences among these scoring approaches. For example, the Framingham risk score is based on data from a single community, while the SCORE risk charts were based on data from 12 European countries. These epidemiologic risk profiling did not address the fact that risks can differ between regions and countries due to different life styles, life expectancy and genetic predisposition. Therefore, these risk prediction algorithms need to evolve over time. An updated Framingham risk score in 2008 predicted risk for more CVD outcomes including cerebrovascular events, peripheral artery disease and heart failure (D'Agostino et al. 2008), compared to the one first developed in 1998. Type-2 diabetes (T2D) was dropped from the updated Framingham risk score because it was considered to be a disease outcome itself, with similar risk factors as that for CVD. These risk scores are used to determine who should be offered preventive drugs such as those lowering blood pressure or cholesterol levels. Individuals with <10%, 10-20%, and >20% CVD risks are considered low, intermediate, and high risk respectively.

The term “biomarker”, as used in the title of this thesis, focuses more on the biologically measurable risk factors. It is meant to distinguish from lifestyle related risk factors such as smoking, drinking, and nutrition. The term biomarker was established as a medical subject heading term in 1989, meaning “measurable and quantifiable biological parameter (e.g. specific enzyme concentration, specific hormone concentration, specific gene phenotype distribution in a population, presence of biological substance) which serves as index for health- and physiology-related assessments, such as disease risk, psychiatric disorders, environmental exposure and its effects, disease diagnosis, metabolic processes, substance abuse, pregnancy, cell line development, epidemiologic studies, etc.” In 2001, an updated definition of biomarker is given by the US National Institutes of Health 2001, as “a

characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacological responses to a therapeutic intervention" (Biomarkers Definitions Working 2001). This definition made the term biomarker more inclusive. In this thesis, the studied cardiovascular biomarkers are all biological molecules existing in circulatory system.

Figure 1.1 Established and new/emerging risk factors for CVD

This figure is adopted from (Badimon and Vilahur 2012) as is.

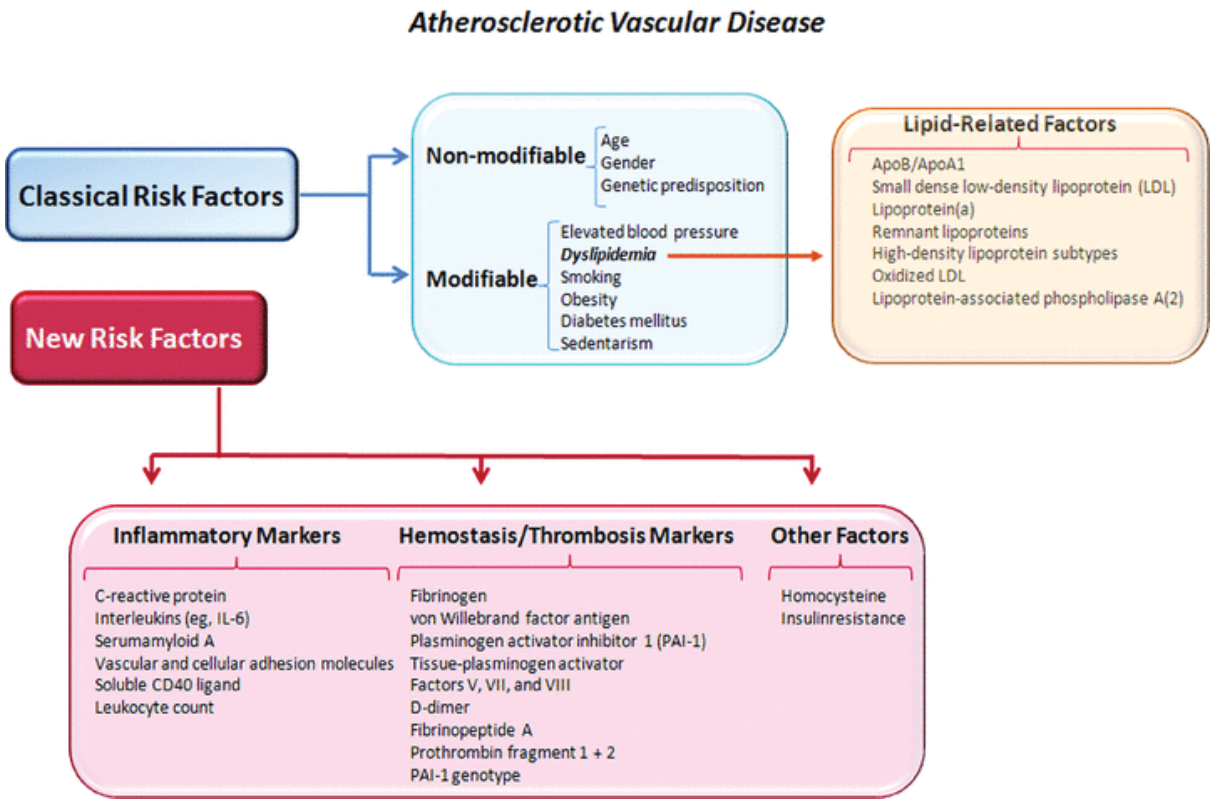
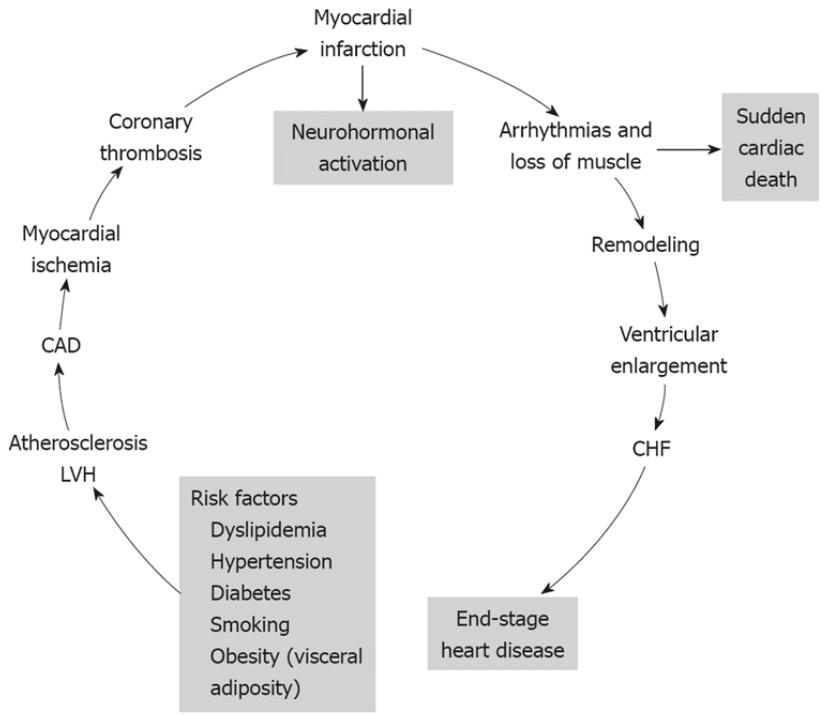


Figure 1.2 The cardiovascular disease continuum

This figure was adapted from Dzau et.al as it (Dzau and Braunwald 1991). LVH indicates left ventricular hypertrophy. CHF indicates congestive heart failure. The major risk factors leading to CVDC are listed at the bottom. All these risk factors, with the exception of smoking, constitute the metabolic syndrome.



1.3 The allelic architecture of complex traits

Population genetics is the study of the distributions and changes of allele frequency in a population, while the population is subject to evolutionary processes. Study areas of population genetics include recombinations, Mendelian inheritance, genetic linkage and linkage disequilibrium (LD), population stratification, etc. Allelic architecture refers to the number and frequencies of susceptibility alleles underlying complex diseases. Diseases with high prevalence in the general population such as T2D and CHD are polygenic, i.e., determined by multiple genetic variants, together with lifestyle and environmental factors. This is also the case for complex, quantitative risk factors. Although there is distinct difference of allelic architecture between high prevalent complex diseases and low prevalent Mendelian diseases, these two are not completely disconnected. Recently, a study linked complex diseases to unique collections of Mendelian loci by showing that common variants associated with complex diseases are enriched in the genes with Mendelian patterns of inheritance (Blair et al. 2013).

Genetic research on complex traits began with surveying candidate variants or regions of the genome, followed by analysis analyses that scan the whole genome with limited resolution, and then genome-wide association studies (GWAS) over the past ~10 years. Due to the nature of "hypothesis driven", candidate gene studies used a very liberal *P* value (such as $P<0.05$) threshold to claim significance, which could lead to a high level of reported false positives (Masicampo and Lalande 2012). Actually, less than 5% of associations identified in candidate gene studies were replicated in larger GWAS (Ioannidis et al., 2011). Linkage analysis is suitable for detecting rare and highly penetrant variants causative for rare diseases with classical Mendelian patterns of inheritance. Early success example of linkage studies included the identification of causal mutations for cystic fibrosis (Kerem et al. 1989) and Huntington disease (MacDonald et al. 1992). In general, linkage analysis is not suitable for detecting common alleles of unusually large effects for complex diseases, but there are a few exceptions, including the successful discoveries of the *INS* locus in T1D (Bell et al. 1984) and the *ApoE* locus in early onset Alzheimer's disease (St George-Hyslop et al. 1987, Goate et al. 1991). The LOD score (logarithm (base 10) of odds) is a statistical test often used for linkage analysis (Morton 1955). It compares the likelihood of obtaining the test data if the two loci are indeed linked, to the likelihood of observing the same data purely by chance. A LOD score of 3.3 or higher has been shown to correspond to a statistical significance level of

0.05. There are two main algorithms used to calculate LOD score: the Elston–Stewart algorithm (Elston and Stewart 1971), and the Lander– Green algorithm (Lander and Green 1987). The major difference is whether the recursion took place over individuals in a pedigree (computing increases linearly with pedigree size but exponentially with the number of loci) or over loci (computing increases linearly with the number of loci but exponentially with pedigree size). The Elston–Stewart algorithm is applicable to very large pedigrees while the Lander–Green algorithm can accommodate thousands of markers on a chromosome.

Before GWAS approach was widely used, there were two theories for explaining genetic underpinning of complex diseases with high prevalence: common disease common variant (CDCV) and common disease rare variant (CDRV). The CDCV theory hypothesised that a small number of common variants could explain a large proportion of phenotypic variation for common traits (Lander 1996, Reich and Lander 2001, Pritchard and Cox 2002, Botstein and Risch 2003). This CDCV theory has been well supported by GWAS where many common variants are identified for association with common diseases and complex traits (Hindorff et al. 2009). However, common variants did not explain common variation fully (Manolio et al. 2009), and this led to a slightly modified version of CDCV - the infinitesimal model. The infinitesimal model highlighted the role of a much larger number of common variants with much smaller effects. This model was also supported by GWAS especially large scale meta-analysis with adequate power for both diseases traits (International Schizophrenia et al. 2009) and quantitative traits (Yang et al. 2011). In contrast to CDCV and infinitesimal model, the CDRV theory hypothesized that a large number of rare variants with large effects could explain a large proportion of heritability (Cirulli and Goldstein 2010). It is worth noting that very rare variants would not be common enough to explain large variance or reach genome-wide significance even if they are causal and have large effects in a small proportion of studied samples. Statistical simulations have shown that CDCV and CDRV are not necessarily mutually exclusive, with both rare and common variants underlying a polygenic genetic architecture for complex traits (Hemani et al. 2013). Other models such as the broad sense heritability model (Eichler et al. 2010) looked beyond genetic variants by considering the combined effects of genotypic, environmental and epigenetic interactions.

1.4 Genome-wide association studies (GWAS)

The completion of the human genome project (Lander et al. 2001, Venter et al. 2001) and the rapid improvement of technologies for ascertaining and analysing the human genome set the stage for GWAS, which has changed the landscape of genetic study on complex diseases. In 2005, only a few dozen loci were reported for association with a handful of complex diseases. By the end of 2011, the NHGRI GWAS catalogue has reported over 2,000 association signals for over 200 complex traits. Actually, the idea of GWAS was not new, proposed as early as in 1996, when association testing was found to have greater power than linkage analysis especially for detecting variants with modest effect sizes (Risch and Merikangas 1996). Risch and colleagues suggested that creating high-density genome-wide polymorphism maps would allow well-powered association testing across all genes. Although the concept and analytic methods for GWAS were ready at that time, it was only implemented around 2005 when genome-wide SNP array were commercialized and were affordable for research projects with large sample size (Syvanen 2005). The genetic polymorphism selection by major vendors was mainly based on data generated from the International HapMap project (International HapMap et al. 2007, International HapMap et al. 2010). For the two biggest vendors, Affymetrix used a strategy of randomly selected SNPs while Illumina used tagging methods that maximize coverage in European populations (Barrett and Cardon 2006). The early versions of SNP arrays usually include less than 1 million common variants, which could be imputed to up to 3 million variants discovered from the HapMap project. When a common set of haplotype variants are analysed by most individual cohorts, results could be cross-examined and meta-analysed in large collaborative consortia.

Compared to candidate gene studies and linkage analysis, GWAS scan the whole genome in a systematic manner for detecting genetic variants susceptible to diseases and quantitative traits (Hirschhorn and Daly 2005). Since GWAS became available, large advances have been made. One of the early successes of GWAS was the identification of the *Complement Factor H* gene as a major risk factor for age-related macular degeneration (AMD) (Haines et al. 2005, Klein et al. 2005), in studies of relatively small sample size (~100 cases) and employing a sparse SNP array (~110K). These studies not only identified strongly associated genetic variants, but also proved that common variants included in genome-wide SNP array could tag underlying causal variants, a key assumption for GWAS.

Follow-up resequencing studies revealed a functional polymorphism that is in high linkage disequilibrium (LD) with the discovered GWAS signal. However, the AMD genetic variants identified in these two studies are rare examples where common variants (MAF >5%) have large effects (OR > 4). In general, the identification of genetic variants linked to complex traits would require many more samples and variants to tag the whole genome and survive the large number of multiple testing. In 2007, a landmark GWAS study with ~17,000 subjects typed on half a million variant SNP array (Wellcome Trust Case Control Consortium 2007) identified 24 independent association signals for seven common diseases. This first WTCCC study was the largest set of GWAS of its time, costing a total of $9 million. It identified 21 loci, of which 14 were novel. All these associations has been confirmed in later meta-analyses. Later on, many other studies conducted extensive replication for suggestive signals coming from this WTCCC study and identified many more novel loci, for type 1 diabetes (Todd et al. 2007), type 2 diabetes (Zeggini et al. 2007), rheumatoid arthritis (Thomson et al. 2007, Barton et al. 2008), and Crohn's disease (Parkes et al. 2007). This in a way established the importance of performing independent replication for modern GWAS. This study also provided a first strong indication of differences in allelic architecture for different traits, with many more associations detected for autoimmune diseases as opposed to hypertension or CAD. Besides novel findings, a number of novel techniques and protocols used in this study became standards in GWAS since then, for example, systematic assessing and adjusting for population stratification, and using the HapMap reference panel for genotype imputation. This study also characterised other types of genomic variations including copy number variants (CNV) and large insertions and deletions. The second landmark genomic study from the WTCCC concluded that most common CNVs are well tagged by common SNPs and are unlikely to discover novel findings for common human diseases (Wellcome Trust Case Control et al. 2010). However, rare CNV and large deletions have been reported for association with other categories of complex diseases including autism and schizophrenia (International Schizophrenia 2008, Glessner et al. 2009).

The subsequent widespread implementation of imputation analysis based on common reference maps (HapMap2 mainly) has been instrumental in the completion of powered meta-analyses of GWAS studies, allowing reaching sample sizes necessary for robust genetic discoveries. As of September 2014, more than 2,000 robust associations with complex traits have been reported (Hindorff et al. 2009), which revealed important biological pathways and defined novel therapeutic hypotheses (Visscher et al. 2012). For example, GWAS on T2D

have played an important role in shifting research focus away from insulin resistance towards insulin production (McCarthy and Zeggini 2009) and led to the identification of many new drug targets (Wolfs et al. 2009). Another example is the discovery of *BCL11A* as a major modifier of disease severity in haemoglobinopathies (Akinsheye et al. 2011), which led to the development of new treatment options for sickle cell disease and beta-thalassemia (Bauer and Orkin 2011).

1.5 GWAS studies of CVD events and cardiovascular biomarkers

The heritability for CHD and stroke was established to be 50% (Fischer et al. 2005) and 32% (Bak et al. 2002) respectively. Although the prevalence of the metabolic syndrome has greatly increased in the past decades due to lifestyle changes, a large portion of the phenotypic variation in cardio-metabolic traits between individuals is still due to genetic variation (van Dongen et al. 2013). GWAS have been widely used to study both end points and intermediate phenotypes of CVD. As mentioned above, the first WTCCC study studied CAD and hypertension together with five other diseases. It reported one locus for coronary CAD but none for hypertension (Wellcome Trust Case Control Consortium 2007). Over the past few years, collaborative efforts have made it possible to conduct large meta-analysis of GWAS with the sample size up to tens of times of the original WTCCC study. Two published large meta-analysis on CAD reported a total of 46 genetic loci for association with CAD (Schunkert et al. 2011, Consortium et al. 2013). The 2013 study reported that 12 and 5 of these 46 CAD loci show significant associations with lipids and BP respectively. It further reported that the four most significant pathways mapping to networks comprising 85% of these putative genes are linked to lipid metabolism and inflammation, underscoring the causal role of lipids and inflammation in the genetic aetiology of CAD. The latest efforts on CAD GWAS used a similar sample size as that in the 2013 study (60,801 cases and 123,504 controls vs. 63,746 CAD cases and 130,681 controls), but used the 1000GP data as imputation reference panel so that it interrogated 6.7 million common (MAF>0.05) and 2.7 million low frequency (0.005<MAF<0.05) (CARDIoGRAMplusC4D Consortium 2015). In addition to confirming most known CAD loci, this study identified 10 novel loci, eight

additive and two recessive. However, this study suggested a lack of evidence of low frequency variants with larger effects and no evidence of synthetic association and suggested that the genetic susceptibility of CAD is largely determined by common SNPs of small effect size.

It was proven challenging that the CAD loci discovered from GWAS could add improvement for risk prediction (Buijsse et al. 2011, Companioni et al. 2011) as compared to other phenotypes such as AMD (Seddon et al. 2009). In general, using genetic loci for risk prediction has unique advantages because genetics do not change over an individual's lifetime and are not affected by other risk factors. Therefore, risk prediction can be carried out much further in advance. In the past 15 years, interest has grown on predicting CVD risk at longer-term (for example, 30-year or lifetime). Genetic information shall benefit such efforts to improve communication of risk, and motivate risk-factor modification especially in young patients (Wong 2014). Also, Mendelian Randomization (MR) studies using genetic variants as instrumental variables could resolve epidemiological problems of establishing causality, which established the causal role for LDL to CVD (Linsel-Nitschke et al. 2008), but not for high-density lipoprotein cholesterol (HDL) (Voight et al. 2012). This approach could also be used to perform retrospective drug trials, for example, the establishment of IL6R as a drug target for CVD (Interleukin-6 Receptor Mendelian Randomisation Analysis et al. 2012).

As stated above, CVD risk factors are critical for the initiation and progression of CVD events. From the point view of genetic research, quantitatively measured risk factors are also preferred to dichotomous CVD events due to increased power and an often more interpretable outcome. For example, assays for LDL levels are precise and standardized around the world, but the diagnosis and clinical criteria for CHD might differ significantly. The beta statistics of a particular variant indicates a unit change in LDL level per allele, but such a statistic for disease outcome would be less intuitive for interpretation. Once genetic variants for quantitative variants are discovered, they could provide clinical insights to the associated diseases (Teslovich et al. 2010). Compared to the disease end points, meta-analyses for quantitative traits have identified many more loci and explained much larger proportion of phenotypic variance. A GWAS meta-analysis for plasma lipids identified 95 loci that explain ~12% of phenotypic variance for high density lipoprotein (HDL), LDL, and total cholesterol (TC). The large sample size is proving powerful for identifying genetic variants with small effect size. Compared to the first WTCCC study that included ~2,000 cases and ~3,000 controls for studying hypertension and discovered no associated locus, the

largest GWAS on BP included more than 200,000 samples identified a total of 29 loci (16 novel) for association with BP. A genetic risk score based on these 29 variants are associated with hypertension, left ventricular wall thickness, stroke and CAD (Ehret et al. 2011). This effectively demonstrates the value of using quantitative risk factors for genetic study of CVD events.

1.6 Rare variants and the motivation for whole genome sequencing (WGS)

Common variants identified by GWAS have proven highly informative to identify novel biological processes underlying common disease (Hindorff et al. 2009). But GWAS is only well powered to detect associations that are well covered by common tag SNPs. Populations with different LD to the HapMap populations, or meta-analyses across populations with different patterns of LD, can confound the tag SNP approach (Teo et al. 2010). Also, low frequency variants are not well tagged by common SNPs (International HapMap et al. 2010). So far, common variants discovered from first generation GWAS explained only a small proportion of phenotypic variance for most common traits and there is a lack of proven added predictive value in clinical usage by including GWAS signals on top of risk factors already known. The missing heritability theory (Manolio et al. 2009) hypothesized that GWAS might have missed variants that have large effects but too low frequency to be detected by SNP array. This is also supported by the evolution theory that alleles susceptible to diseases and their risks are likely to be deleterious and could not reach high frequency due to purifying selection (Pritchard 2001, Goldstein et al. 2013). Although it is debatable on whether, and how much, synthetic associations from variants could explain common variants effects, it was already shown that rare copy number variants contribute to several complex neurodevelopmental disorders (International Schizophrenia 2008, Glessner et al. 2009). The variants with low to rare frequency (shown in light blue in **Figure 1.3**) could be where a large proportion of missing heritability resides. This is a key underlying reasoning for the new generation of population genetic studies where sequencing technologies are used for discovering low frequency (defined here as MAF between 1-5%) and rare variants (defined here as MAF <1%). Sequencing could identify low frequency and rare SNPs, various types of structural variations, as well as more common variants (~ 10-15%) that are not well tagged by SNP arrays (Flannick et al. 2012). Sequencing studies could also

potentially discover causal functional variants that could not be well interrogated on SNP array or imputation (Cirulli and Goldstein 2010).

The desire to study low frequency and rare variants in a genome-wide fashion was met by fast development in sequencing technologies. In 2004, the 454 pyrosequencing method pioneered the field by allowing hundreds of thousands of sequencing reactions to be carried out in parallel (Langaee and Ronaghi 2005). In 2006, the Solexa reversible termination sequencing method was commercialized by Illumina. In 2007, the Oligonucleotide Ligation and Detection (SOLiD) technology was introduced by ABI (now Life Tech). By 2007, it was possible to sequence over 500Mb a day on a single machine (Mardis 2008), and that was when the 1000 Genomes Project (1000GP) was founded to perform low-coverage (2-4X) sequencing on up to 2,500 human genomes. Since 2008, more sequencing technologies are developed, including Ion torrent, pacific biosciences, Illumina's MiSeq (Quail et al. 2012). In January 2010, Illumina unveiled the HiSeq 2000 sequencing system. It initially generated two billion paired-end reads and 200Gb of quality filtered data in a single run, which allows researchers to obtain 30-fold coverage of two human genomes in a single run. This is the sequencing technology adopted by the UK10K project, which is funded by the Wellcome Trust in March 2010.

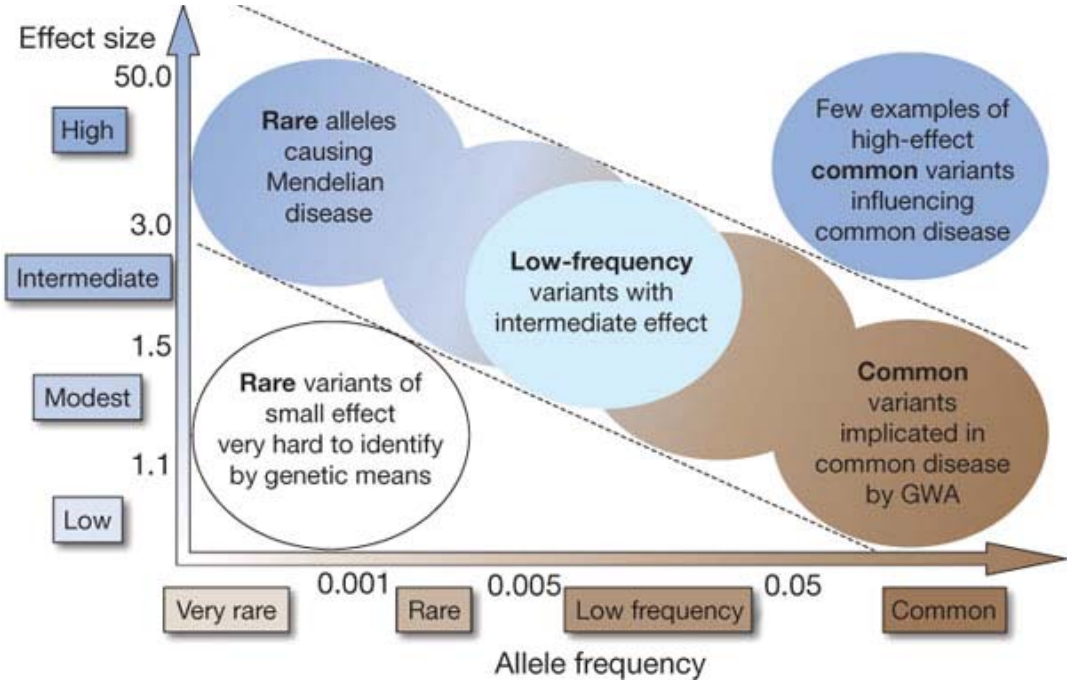
While WGS is still prohibitively expensive for large population based studies, the development of sequence capture technology enabled sequencing of the whole exome (Albert et al. 2007), which covers ~1.5% of the human genome (Lander et al. 2001). Compared to WGS, whole exome sequencing (WES) studies have been conducted at an even greater scale over the past several years, due to cost efficiency as well as data analysis efficiency where genomic boundaries and annotations could be defined straightforward and therefore the results are easier to be interpreted. WES became the dominant method for discovering causal variants for Mendelian diseases (Bamshad et al. 2011), while WGS should discover a lot more biologically relevant variants for common complex traits. This is consistent with findings from the ENCODE project that most variants that control protein biochemistry are non-coding and are not within exons (Pennisi 2012). Currently, most WGS technology sequence the whole genome in low depth, sometimes complemented by high-depth sequencing of the whole exome (Abecasis et al. 2012).

Finally, the increased availability of whole-genome and whole-exome sequencing data is bringing linkage analysis once again to the forefront of genetic research, owing to the development of powerful methods to detect rare variants and the use of family-based data."

In association studies, population stratification can lead to an increased number of false-positive results if not properly accounted for. However, this is not a problem in linkage analysis because the family structure instead of the population genotype frequencies dictates a proband's genotypes. Given that large and complete pedigree is usually hard to get for genetic studies, it is preferable to combine positive aspects of linkage and association analysis by using family-based rather than population-based control individuals. Although the transmission disequilibrium test (TDT) tests have already used such family-based controls .it is only powerful when there is both linkage and association. The TDT test was recently extended (the rare variant-TDT (RV-TDT)) to WGS data, with several rare variant association tests methods implemented (He et al. 2014). Linkage analysis not only effectively adjusts for population stratifications, but also provides statistical evidence for disease aetiology. Over the past couple of years, linkage analysis coupled with WGS have identified many new disease susceptibility genes, with a sample size that is much smaller that would be needed for a population based genome-wide scan. In the future, linkage analysis of WGS data is expected to be even more widely used (Yan et al. 2013, Santos-Cortez et al. 2014).

Figure 1.3 The allelic spectrum of human disease predisposition

This figure is copied as is from Maniolio et al. 2009 (Manolio et al. 2009). It illustrates the relationship between frequency and effect size for genetic variants contributing to human disease, from common to rare. The focus of WGS based studies aim to low-frequency to rare alleles with modest effect sizes, as shown by the light blue circle in the figure.



1.7 The UK10K Project

In 2010, the Wellcome Trust found the largest WGS study at the time - the UK10K project, with a £10.5 million funding support. The UK10K project aims to better understand the link between low frequency and rare genetic variants and their impact on health and diseases (The UK10K Consortium 2015). The full UK10K project conducted sequencing for ~10,000 samples: the cohort arm (referred as UK10K-Cohorts) conducted WGS for ~4,000 population based samples; the disease arm conducted high-depth WES for ~6,000 affected individuals. For the ~4,000 samples included in the cohort arm, ~2,000 each are from two well established population studies in UK: TwinsUK (Spector and Williams 2006) and The Avon Longitudinal Study of Parents and Children (ALSPAC) (Golding et al. 2001). TwinsUK is a general population throughout UK (Moayyeri et al. 2012) while ALSPAC is a population-based birth cohort study that recruited more than 13,000 pregnant women resident in Bristol (formerly Avon) UK. For both cohorts, study participants were selected to maximise phenotypic coverage, previous genome-wide array genotyping, coverage with other "-omic" datasets (transcriptomic, metabolomic) and consent to WGS, but were otherwise representative of the original population samples.

Using low-depth WGS in UK10K-Cohorts is a cost-effective approach when high-depth WGS is still prohibitively expensive for thousands of samples. For example, it was shown that sequencing 3,000 individuals at low-depth (4X) provides similar power to sequencing of >2,000 individuals at high depth (30X) for disease-associated variants with frequency >0.2%, but the low-depth approach only requires ~20% of the sequencing resources (Li et al. 2011). An average sequencing depth of 7X in the UK10K-Cohorts project enables the identification of almost all accessible SNPs, Insertion/Deletion polymorphism (InDel) and other structural variants down to MAF of 0.1% (Le and Durbin 2011). This is one magnitude higher resolution compared to the 1000 genome project (1000GP) that fully characterize variants down to MAF of 1% (Abecasis et al. 2012). The low-depth sequencing was proven sensitive for detecting rare variants, which detected more than 70% of singletons and more than 90% of doubletons that are discovered in the UK10K high-depth (80X) WES arm. The UK10K WGS approach also discovered a lot of rare variants that could be potentially characteristic of the UK population. Roughly, only 10% of singletons discovered in UK10K WGS were previously discovered by 1000GP (The UK10K Consortium 2015).

Besides the ~4,000 samples directly sequenced, the two cohorts in UK10K cohort arm (TwinsUK and ALSPAC) have an additional ~10,000 samples with genome-wide SNP array data, which could be imputed into the full set of variants discovered from WGS. All variants with MAF down to 0.1% should be imputable, where minor alleles occur more than five times in the study sample and the definition of a shared haplotype between study sample and reference sample is possible. A total of 64 biomedically relevant traits (60 quantitative traits and four binary traits) were measured in these two cohorts and were analysed in UK10K-Cohorts, 31 of which exist in both cohorts and are their initial association results were presented in the UK10K flagship paper (The UK10K Consortium 2015). The sample size for each of the 64 traits is listed in **Table 1.1**. My PhD thesis concentrate on a total of 13 CVD related biomarkers, including four lipid traits (HDL, LDL, TC, TG), one inflammatory biomarker (CRP), and eight haematological traits (Hemoglobin (HGB), Mean corpuscular hemoglobin (MCH), Mean corpuscular hemoglobin concentration (MCHC), Mean corpuscular volume (MCV), Packed cell volume (PCV), Platelet counts (PLT), Red blood cell counts (RBC), White blood cell counts (WBC)).

The large number of traits measured on the same individuals in the UK10K-Cohorts provided a good opportunity to learn about the general allelic architecture especially rare variants architecture of those traits. Since single marker association tests are typically underpowered for rare variants (MAF <1%), the UK10K-Cohort projects adopted an integrative framework of variance component method and burden tests implemented in sequence kernel association test (SKAT) and SKAT optimized (SKAT-O) (Wu et al. 2011, Liu and Leal 2012) . The details of these association tests will be described in Chapter 2.

Table 1.1 List of traits in UK10K-Cohorts

The 64 traits were grouped into categories based on biomedical relevance. WGS means those samples sequenced, GWA means those samples with SNP-array data, imputed to the WGS reference panel.

Category	Name	TwinsUK WGS	ALSPAC WGS	Total WGS	TwinsUK GWA	ALSPAC GWA	Total GWAS	Total
Obesity /anthropometry	**BMI**	1747	1791	3538	2330	4101	6431	9969
	Height	1747	1794	3541	2331	4103	6434	9975
	Weight	1747	1812	3559	2330	4132	6462	10021
	Hip circumference	1266	1808	3074	1623	4115	5738	8812
	Waist circumference	1265	1807	3072	1624	4121	5745	8817
	Waist hip ratio	1265	1806	3071	1620	4116	5736	8807
	Total fat mass	1716	1683	3399	2095	3815	5910	9309
	Total lean mass	1716	1683	3399	2095	3815	5910	9309
	Trunk fat mass	1514	1683	3197	547	3815	4362	7559
	Forearm length	-	1760	1760	-	4367	4367	6127
	Head circumference	-	1762	1762	-	4388	4388	6150
	Leg length	-	1764	1764	-	4386	4386	6150
	Sitting height	-	1764	1764	-	4387	4387	6151
	Upperarm length	-	1762	1762	-	4369	4369	6131
	Adiponectin	864	1461	2325	737	2772	3509	5834
	Leptin	958	1459	2417	663	2765	3428	5845
Diabetes Biochemistry	**Glucose**	1701	1224	2925	2202	1701	3903	6828
	HOMA-B	1669	1219	2888	1671	1697	3368	6256
	HOMA-IR	1577	1219	2796	1659	1695	3354	6150
	Insulin	1676	1220	2896	1927	1693	3620	6516
Heart function	**Heart rate (ECG+pulse)**	1385	1590	2975	939	2932	3871	6846
CVD hypertension	**DBP**	1536	1773	3309	1457	4046	5503	8812
	SBP	1536	1773	3309	1457	4046	5503	8812
CVD Biochemistry	**HDL**	1713	1497	3210	1896	2820	4716	7926
	LDL	1696	1495	3191	1870	2815	4685	7876
	TC	1711	1495	3206	1895	2817	4712	7918
	TG	1705	1497	3202	1882	2820	4702	7904
	VLDL	1700	1497	3197	1874	2820	4694	7891
	Apolipoprotein A1	1449	1465	2914	995	2772	3767	6681
	Apolipoprotein B	1443	1468	2911	989	2765	3754	6665
	Homocysteine	1279	93	1372	799	184	983	2355
	CRP	879	1167	2046	1017	2226	3243	5289
Blood Biochemistry	**HGB**	1553	1524	3077	1056	2882	3938	7015
	MCH	1549	-	1549	1061	-	1061	2610
	MCHC	942	-	942	947	-	947	1889
	MCV	1548	-	1548	1058	-	1058	2606
	PCV	1555	-	1555	1062	-	1062	2617
	PLT	1553	-	1553	1070	-	1070	2623
	RBC	1561	-	1561	1062	-	1062	2623
	WBC	1551	-	1551	1065	-	1065	2616
	Interleukin 6	-	1480	1480	-	2779	2779	4259
Liver Function	**Albumin**	1713	-	1713	1700	-	1700	3413
	Alkaline phosphatase	1702	-	1702	1636	-	1636	3338
	Bilirubin	1702	-	1702	1637	-	1637	3339
	Gamma glutamyl transpeptidase	1699	-	1699	1594	-	1594	3293

Table 1.1 List of traits in UK10K-Cohorts (*continued*)

Category	Name	TwinsUK WGS	ALSPAC WGS	Total WGS	TwinsUK GWA	ALSPAC GWA	Total GWAS	Total
Renal Function	**Bicarbonate**	1714	-	1714	1676	-	1676	3390
	Creatinine	1707	-	1707	1629	-	1629	3336
	Phosphate	1392	-	1392	1691	-	1691	3083
	Sodium	1683	-	1683	1677	-	1677	3360
	Urea	1697	-	1697	1617	-	1617	3314
	Uric acid	1305	-	1305	1588	-	1588	2893
Lung Function	**FEV/FVC ratio**	1676	1604	3280	1892	3521	5413	8693
	Forced Expiratory Capacity	1679	1606	3285	1896	3522	5418	8703
	Forced Expiratory Volume	1681	1606	3287	1896	3522	5418	8705
Birth	**Birth weight**	-	1691	1691	-	5327	5327	7018
	Birth length	-	1137	1137	-	3470	3470	4607
	Gestational age	-	1712	1712	-	5390	5390	7102
	Ponderal index	-	1122	1122	-	3421	3421	4543
	Placental weight	-	703	703	-	2166	2166	2869
Dynamic	**Grip strength**	1514	1682	3196	901	3465	4366	7562
	Ever broken bone*	-	1756	1756	-	3657	3657	5413
	Eye preference*	-	1671	1671	-	4158	4158	5829
	Handedness tasks*	-	1700	1700	-	3972	3972	5672
	Handedness drawing*	-	1676	1676	-	3875	3875	5551

* binary traits

1.8 This thesis

In this chapter, I have reviewed the research on complex disease genetics in general, and the genetics of cardiovascular biomarkers in particular. I also laid out the motivation for WGS based studies and gave a description of the UK10K project. My main hypothesis is that applying WGS to deeply phenotyped population samples is capable of discovering rare but highly penetrant genetic variants. The main research aim is to utilize large-scale WGS data and WGS imputed data to identify novel genetic variants that contribute to CVD related traits. As it is still not clear whether some of the selected biomarkers are direct mediators of the disease or merely markers of disease manifestation, I hope to identify highly penetrant genetic determinants of these biomarkers that can, in the future, be used to assess genetic risk and causal effects. I have contributed to the whole UK10K-Cohorts study and will elaborate on some general lessons learned from this study in the general discussion section. In the following chapters, I describe methods and results for WGS based imputation (chapter 3) and the deep analysis of 13 CVD biomarkers (chapters 4-6). Specifically, I seek to evaluate the following three broad aspects: 1. what are the characteristics of phasing and imputation with WGS data? 2. what novel analytic methods could be applied to a large scale WGS based association study on a rich of phenotypes? 3. can I identify novel and potentially stronger effect genetic variants that are associated with the chosen CVD traits?

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).

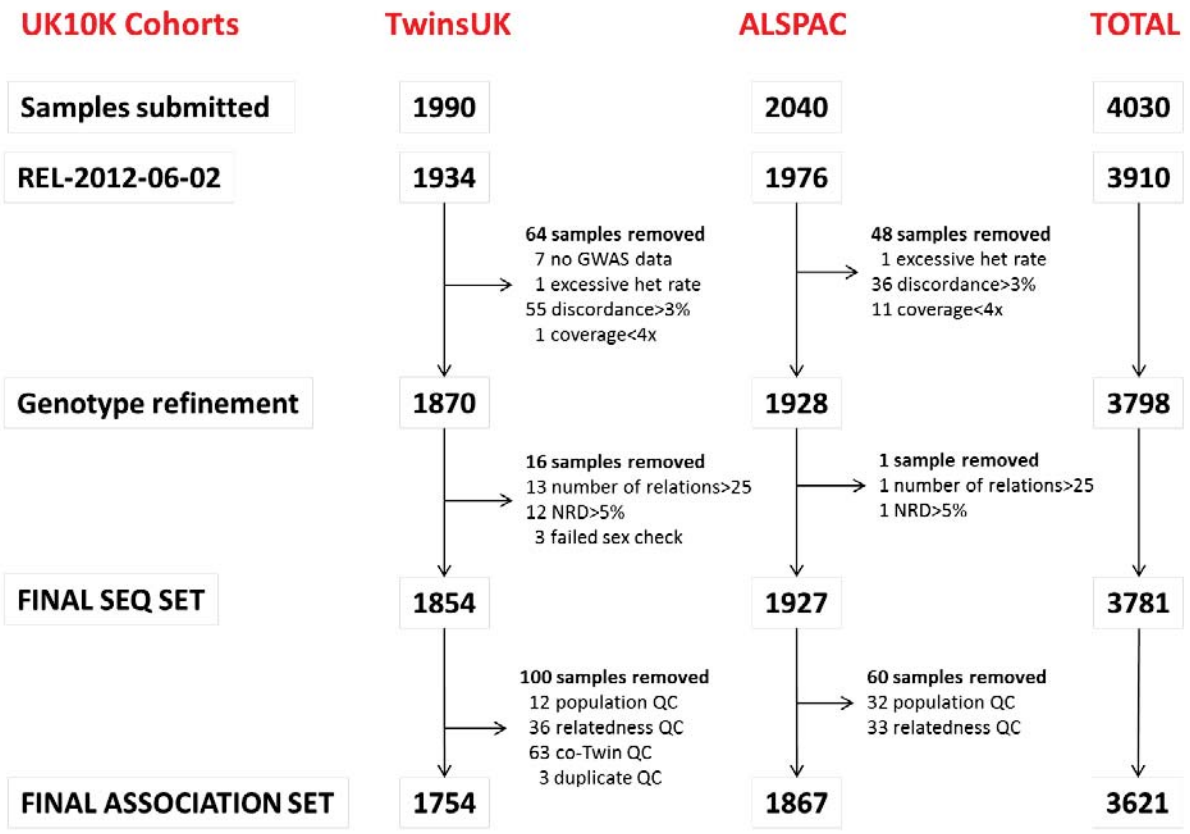
2.3 Genetic data

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.

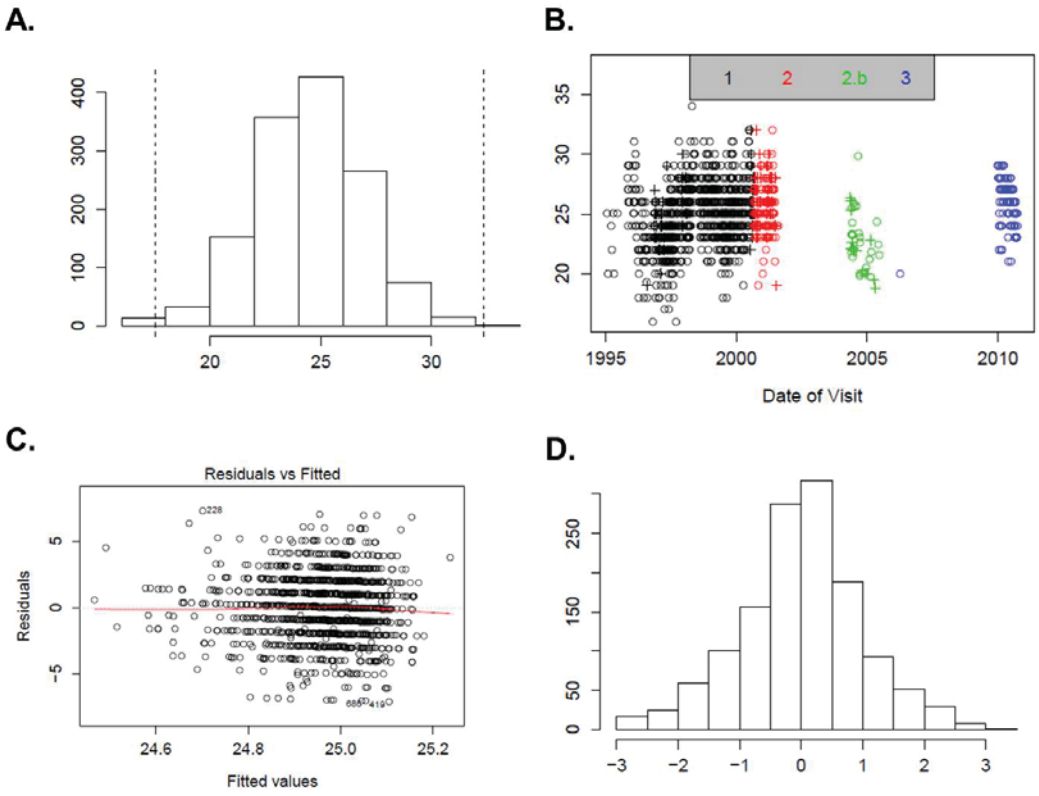
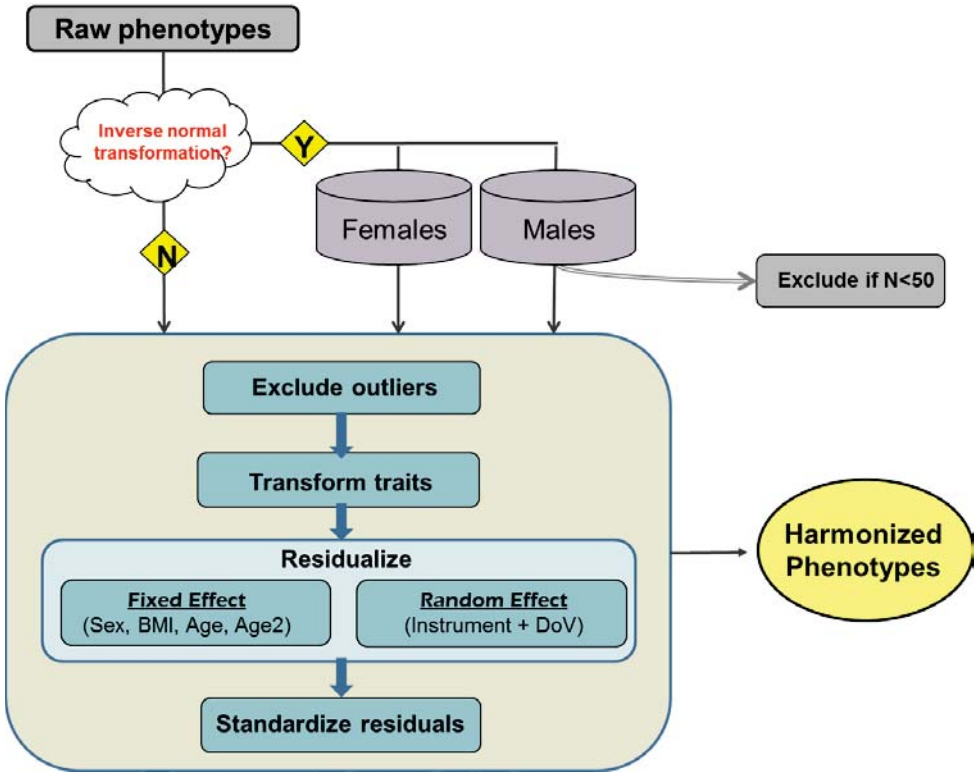


Figure 2.3 Phenotype harmonization protocol

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

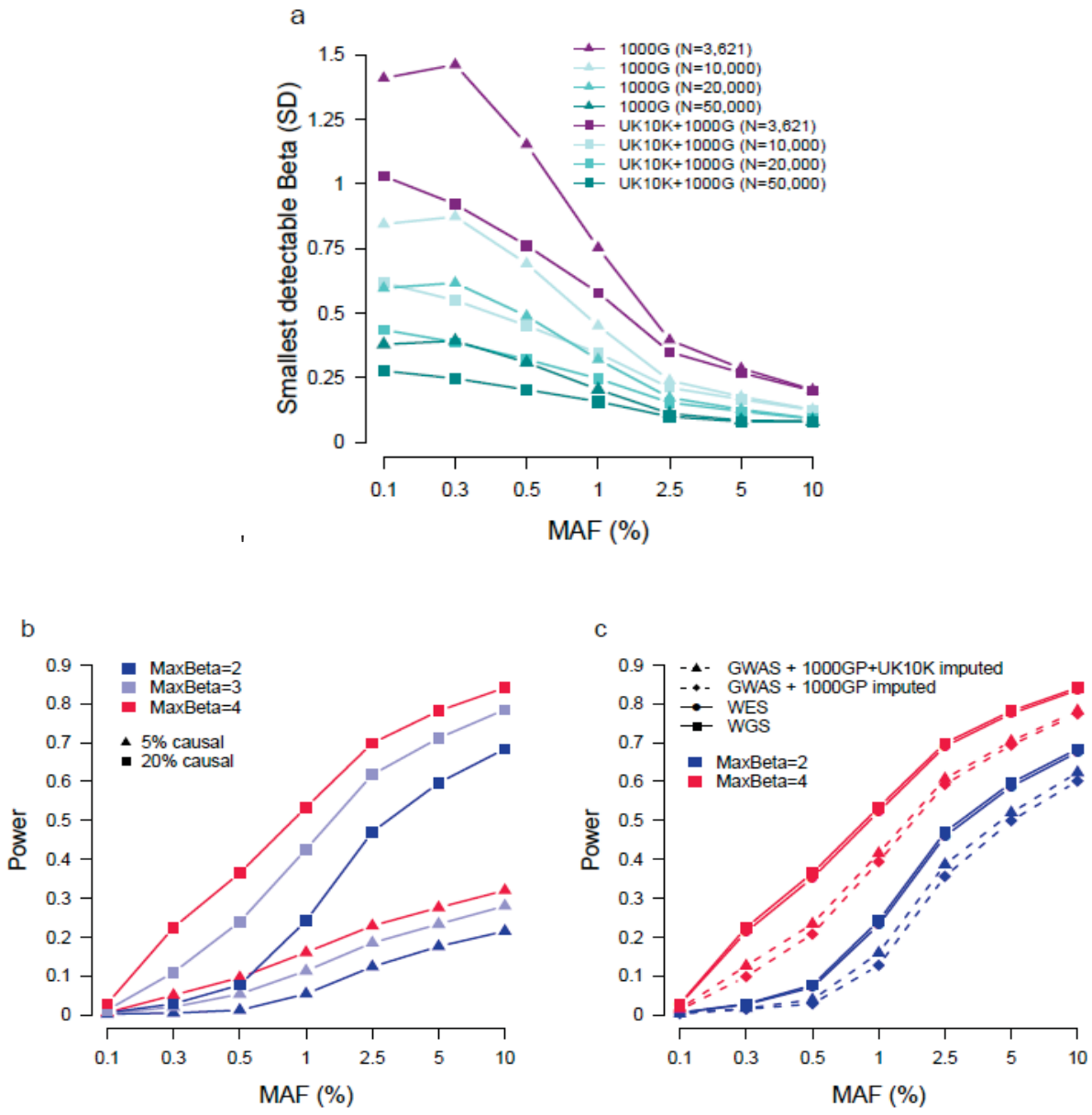
= 0.3%, there is equivalent power when imputing from UK10K+1000GP into a 3,621 sample as when using the 1000GP imputation panel alone with 10,000 samples.

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:

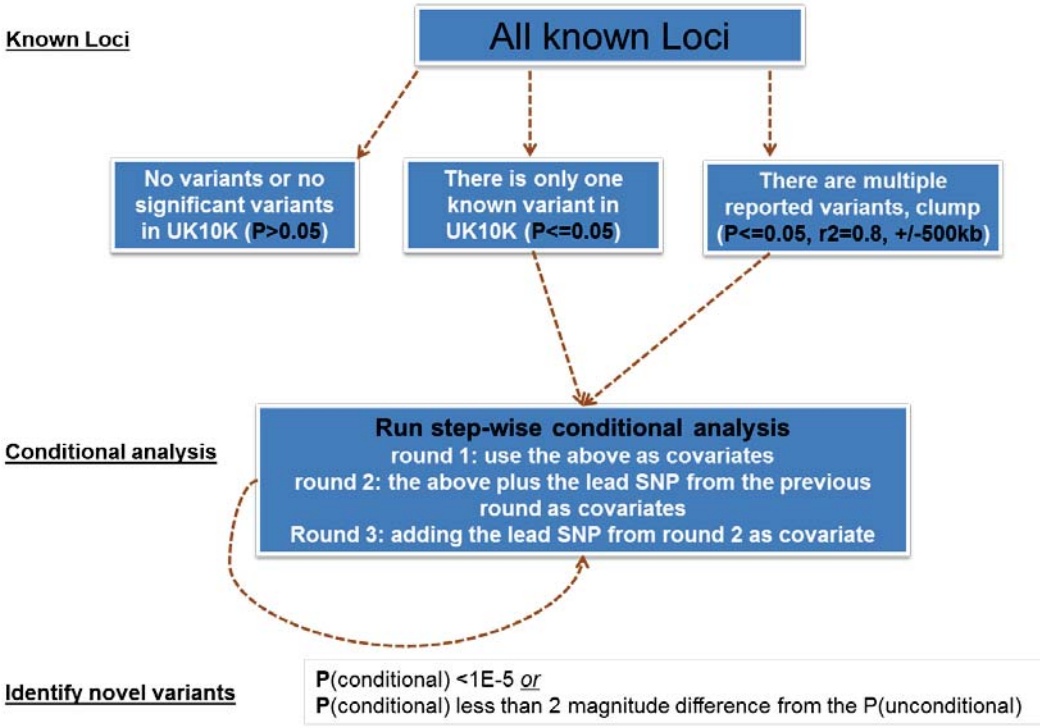
1. For each studied trait, I compiled a list of published variants as positive controls by selecting all SNPs associated with a trait of interest from the NHGRI GWAS Catalog (http://www.genome.gov/gwastudies/) ($P \leq$ 5E-08 last updated in May 2014),

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.

3 Imputation

Disclaimer

The content of this chapter is now published as a paper (Huang et al. 2015). Text written in this chapter might overlap substantially with text in the published paper. In this chapter, I use "I" for the work that was mainly done by myself alone, while indicate clearly for work done by others. Bryan Howie, the co-first author of the published paper, implemented the IMPUTE2 software for merging reference panels and for using a new metric to sample haplotypes.

3.1 Introduction

3.1.1 How imputation works

Imputation is a statistical inference of missing genotypes, where genotyped markers from SNP arrays are used to infer unobserved genotypes from haplotype panels. Although there are quite a few different software for running imputation, the common underlying method is based on a hidden Markov model (HMM) that treats a sample haplotype as a mosaic of a pool of reference haplotypes and uses haplotype patterns in a reference panel to predict unobserved genotypes in a study dataset (Li and Stephens 2003, Scheet and Stephens 2006, Marchini et al. 2007, Browning and Browning 2009, Li et al. 2009). Imputation using large reference panels such as 1000GP has been made computationally efficient by pre-phasing of GWAS samples (Howie et al. 2012) and approximations that select a subset of reference haplotypes (Howie et al. 2011).

3.1.2 Use of imputation in GWAS

Imputation has been instrumental to the discovery of thousands of complex trait loci in genome-wide association studies (GWAS) (Howie et al. 2009). Imputation not only boosts genetic data through a most cost-effective approach and therefore increases statistical power, but also generates datasets with common list of SNPs that facilitate broad collaboration. By imputing individual SNP array dataset with customized content to the common set of variants in HapMap (International HapMap et al. 2007, International HapMap et al. 2010), the international society has been able to look at a common set of ~3 million variants across different cohorts and projects.

3.1.3 Imputation with WGS reference panels

Those variants in the HapMap reference panel are mainly common across populations, defined as MAF >5%. Although WGS provides near-complete characterization of genetic variation, it is still prohibitive for researchers to conduct WGS on large number of samples

that are needed to study the effect on phenotypic variation by rare variants. Instead, using publically available WGS data as reference panels to impute existing datasets with genome-wide SNP array data would be a most cost-effective alternative. Built upon from the HapMap project, the 1000GP provides phased haplotypes for more than a thousand samples from diverse worldwide populations, thereby boosting variant coverage and imputation quality, particularly for variants with MAF of 1-5% (Abecasis et al. 2012). In my early work, I showed that imputations using the 1000GP data could identify novel genetic variants that were not identified in SNP arrays or through HapMap based imputation (Huang et al. 2012).

The 1000GP imputation reference panel currently widely used (Phase 1 version 3) includes a total of 1092 samples, 381 of which are European. In contrast, the UK10K project conducted WGS for 3,781 European samples with higher depth (~7X), and is powered to detect and impute variants with MAF down to 0.1% (The UK10K Consortium 2015). Using the UK10K panel or using the combination of UK10K and 1000GP are expected to provide more accurate imputation for low frequency and rare variants, which are a most effective approach for increasing statistical power along with a large sample size. Here I evaluate the utility of the UK10K WGS dataset as an imputation reference panel, above and beyond the WGS data from 1000GP.

3.1.4 Aims of this study

As the imputation reference panel includes thousands of reference haplotypes and tens of millions of variants, for each of the thousands of samples on the target panel to be imputed, the ideal scenario is that the best matched haplotype exists in the reference haplotype pool while the imputation program does not need to scan all haplotypes in order to use it for imputation. Combining multiple reference panels could improve the representativeness of the reference haplotype pool, while designing an algorithm to quickly narrow down the best matched haplotypes would substantially save computation time and cost. Therefore, the evaluation steps aims to find a preferred imputation strategy that maximizes haplotype representativeness and minimizes computational resources.

Evaluation on performance of WGS reference panels

Recently, a new option in the IMPUTE2 software (Howie et al. 2009, Howie et al. 2011) allowed two sets of haplotypes to be combined to form a single set of haplotypes at the union set of sites. Imputation into GWAS samples can then be carried out using this combined panel. This method can be used to combine two sets of haplotypes from two distinct population cohorts, such as UK10K and 1000GP, as described in this chapter. The results from my evaluation of UK10K and 1000GP should also help investigators who wish to use their own WGS data instead of UK10K to merge with 1000GP data.

The main difficulty in combining reference panels is that some sites will only have data in one or other of the panels. This could be due to population specific alleles, low-coverage of the non-reference allele, or cohort specific site filtering that removed the site from consideration. The new option in IMPUTE2 software uses HMM to impute the unobserved alleles in each panel while the other panel is used as reference. Once the two reference panels are imputed up to the union of their variants, the best-guess haplotypes are used to impute a GWAS cohort in the same way as using only one reference panel. IMPUTE2 could output the haplotypes of the merged reference panel so that they are used for future imputation without repeating this merging step. This new functionality is available in IMPUTE2 v2.3.0 or newer version (https://mathgen.stats.ox.ac.uk/impute/impute_v2.html).

Evaluation on approximation of haplotype sampling

Genotype imputation in GWAS has always been a computationally intensive task. Recent developments like pre-phasing have greatly reduced the computational cost of imputation, but growing reference panels continue to challenge existing methods. Previously, IMPUTE2 chose a different subset of k_{hap} reference haplotypes (by default, 500) for each GWAS haplotype. The matching was based on an approximation of hamming distance metric. When this subset includes the most informative reference haplotypes, it can speed up the imputation calculations without sacrificing much accuracy. The cost of imputation with pre-phased GWAS data scales linearly with the number of reference haplotypes N, so the speedup expected from this approximation is roughly N / k_{hap} after accounting for the overhead of reading in a large data set. This speed-up would matter significantly since there are around ~10,000 haplotypes in the combined UK10K and 1000GP reference panel.

3.2 Methods

The various evaluations to be conducted aim to address the two key questions stated above: 1. Does UK10K reference panel perform better than 1000GP, or combining these two panels together would perform even better? 2. Is there a cost-effective approach for sampling only some of the reference haplotypes for imputing each sample? For the testing evaluations described in this chapter, the reference panels are UK10K WGS and 1000GP WGS, and the target panels are two pseudo-GWAS where some genetic variants are masked out to mimic the content of a SNP array panel. The masked out variants would then be used as "true" data to compare with the imputed data for the same sites and same sample. The evaluation was done sequentially. Once a preferred metric or design is identified in one round, the less preferred metrics or designs will not be evaluated again in the following rounds.

3.2.1 WGS Reference Haplotypes

UK10K WGS

The UK10K WGS data included 3,781 samples and contained over 42 million SNV and ~3.5 million insertion/deletion polymorphisms. To assess the quality of genotype data from low-depth sequencing, the UK10K study compared the variant sites and genotypes of 61 TwinsUK individuals with high-coverage exome data. A high level of concordance was observed (**Table 3.1**). Originally, the UK10K WGS panel was phased by Beagle during the genotype refinement step. In 2013, it was reported that re-phasing the 1000GP WGS panel using SHAPEIT v2 led to improved imputation quality (Delaneau et al. 2013), I therefore used SHAPEIT v2 for re-phasing the UK10K reference haplotypes. Per the recommendation of this software, the mean size of the windows in which conditioning haplotypes are defined is set to 0.5MB, instead of 2MB used for pre-phasing GWAS. Due to the significantly higher number of variants in the WGS data, the re-phasing was conducted by 3MB chunk with 250kb buffering regions, rather than by whole chromosomes as for the pseudo-GWAS. Imputation was carried out on the same chunks with the same flanking regions. To re-phase the UK10K final release sequencing data, I first converted the VCF files into PLINK binary format, each chromosomes split into 3MB chunks with +/-250kb flanking regions. I then used SHAPEIT v2 to re-phrase the haplotypes for each 3MB chunks with +/-250kb flanking

regions. Although the chunk files could be used as reference panels directly, I also created whole chromosome files based on these re-phased chunks. To do that, phasing information from the SHAPEIT output was copied back to the original VCF files, by using the vcf-phased-join program from the VCFTOOLS package (Danecek et al.).

To merge UK10K reference panel with 1000GP reference panel for creating a combined reference panel, I first identified sites that need to be excluded. For UK10K, the following sites were excluded: 18,180,633 singletons that do not exist in 1000GP, 1,064,168 multi-allelic sites and 214,631mis-matched alleles sites. For 1000GP, the following sites were excluded: 7,053,246 singletons that do not exist in UK10K, 23,932 sites with a SNP and an INDEL at the same position and 443 within large structural deletions (**Table 3.2**). To identify these variants, I first used VCF-QUERY to get the summary statistics of the two sets of VCF files, including chromosome, position, reference and alternative alleles, and then compare the two summary statistics files against each other. I then used VCFTOOLS to exclude those sites to create a new set of VCF files. Finally, I used VCF-QUERY to convert the new VCF files into phased haplotypes and legend files that could be fed directly to IMPUTE2 for running imputation.

1000GP WGS

The 1000GP Phase I integrated variant set release (v3) for low-coverage whole-genomes in NCBI build 37 (hg19) coordinates was downloaded from 1000GP FTP site (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20110521/, 23 Nov 2010 data freezes). This callset includes phased haplotypes for 1,092 individuals and 39,527,072 variants (22 autosome and chromosome X). The haplotypes were inferred from a combination of low-coverage genome sequence data, and they contain SNPs, short INDELs, and large deletions. As mentioned above, the following sites were excluded: 7,053,246 singletons that do not exist in UK10K, 23,932 sites with a SNP and an INDEL at the same position and 443 within large structural deletions. The final reference panel included all 1,092 samples and 32,449,428 sites.

Merging two WGS reference panels

The following 3 steps were used to merge two WGS reference panels using IMPUTE2 (version 2.3 and later):

1. Impute the variants that are specific to panel 1 (1000GP) into panel 2 (UK10K).
2. Impute the variants that are specific to panel 2 (UK10K) into panel 1 (1000GP).
3. Treat the imputed haplotypes in both panels (with the union of variants from both) as known (i.e., take the best-guess haplotypes) and impute the GWAS cohort in the usual way.

Data access

UK10K reference haplotypes are available from the European Genome-phenome archive (EGA study: EGAS00001000713, EGA dataset: EGAD00001000776) under managed access conditions (see http://www.uk10k.org/data_access).

3.2.2 Test GWAS datasets

UK10K Pseudo-GWAS

A random set of 500 samples passing QC filters were chosen from the TwinsUK (N=1,854) and ALSPAC (N=1,927) WGS datasets. Genotypes for a total of 13,413 sites (corresponding to the content of the Illumina HumanHap610 SNP-array) on chromosome 20 were extracted from the UK10K WGS data in these 1,000 samples.

INCIPE Pseudo-GWAS

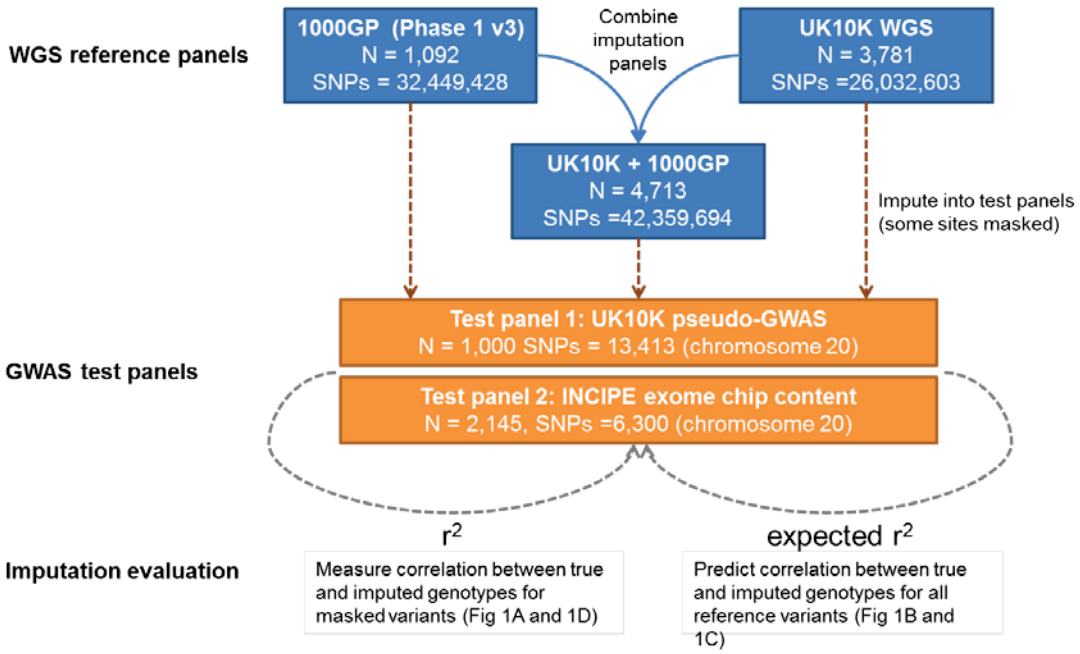
For the INCIPE study, 6,200 Caucasian participants were randomly chosen from the lists of registered patients of 62 randomly selected general practitioners based in four geographical areas in the Veneto region, north-eastern Italy (Gambaro et al. 2010). A total of total of 2,258 samples were genotyped with the HumanCoreExome-12v1-1 platform and were subject to further quality control (QC) evaluation as follows to determine sample and SNP quality. The details of QC for this dataset is presented elsewhere (Huang et al.). At the end, there are a total of 346,941 polymorphic variants on autosomes and 8,822 of those on chromosome 20 were retained for analysis. For the imputation evaluation, 2,522 exonic variants (i.e. those corresponding to the exome selected part of the array) on chromosome 20

were masked out. The remaining 6,300 SNPs were retained as a pseudo-GWAS imputation panel.

3.2.3 Running imputation

Prior to imputation, the two pseudo-GWAS datasets were pre-phased using SHAPEIT v2 (Delaneau et al. 2013) to increase phasing accuracy. The UK10K pseudo-GWAS panel was phased jointly with those samples in UK10K WGS. The INCIPE pseudo-GWAS of 2,145 participants was pre-phased separately. Imputation of genotypes from the three phased reference panels (UK10K, 1000GP and UK10K+1000GP) into the two test panels was carried out on chromosome 20, split in 3MB chunks with 250kb buffer regions. Imputation was performed using standard parameters with IMPUTE2. The accuracy of imputed variants was calculated as the Pearson correlation coefficient (r^2) between imputed genotype dosages in [0-2] and masked sequence genotypes in (0,1,2). The results were stratified into non-overlapping MAF bins for plotting. The overall flow of imputation evaluation is shown in **Figure 3.1**.

Figure 3.1 imputation evaluation workflow



3.3 Results

3.3.1 Characteristics of UK10K WGS panel

The UK10K Cohorts Project (http://www.uk10k.org/studies/cohorts.html) includes two population samples from the UK. The TwinsUK registry comprises unselected, mostly female volunteers ascertained from the general population through national media campaigns in the UK (Moayyeri et al. 2012). The Avon Longitudinal Study of Parents and Children (ALSPAC) is a population-based birth cohort study that recruited more than 13,000 pregnant women resident in Bristol (formerly Avon) UK (Golding et al. 2001). A total of 1,990 individuals from TwinsUK and 2,040 individuals from ALSPAC were consented for sequencing. Variant sites and genotype likelihoods were called using SAMtools (Li et al. 2009), and genotypes were refined and phased using Beagle (Browning and Browning 2009), following similar procedures to the 1000GP (Abecasis et al. 2012). After quality control, 45,492,035 variant sites were retained (**Table 3.2**) in 1,854 and 1,927 individuals in the TwinsUK and ALSPAC panels, respectively. I downloaded the phased haplotypes of 1000GP (Phase 1 integrated v3), which include a total of 39,527,072 sites. For imputation, I removed multi-allelic sites and further excluded variants seen only once in the combined 1000GP+UK10K dataset. A total of 26,032,603 sites were retained for the imputation reference panel of UK10K panel, and 32,449,428 sites for the imputation reference panel of 1000GP. Given that 16,122,337 exist in both panels, combining the two reference panels results in a total of 42,359,694 sites (**Table 3.2**).

Table 3.1 Sequence quality and variation metrics for UK10K Cohorts

This table was adopted from the UK10K study. The numbers in the table was provided by Klaudia Walter. For 61 overlapping TwinsUK individuals, the UK10K study compared the variant sites and genotypes of the low-coverage sequences with high-coverage exome data by non-overlapping AF bins (WGS versus Exomes). It considered 74,621 shared sites in non-overlapping AF bins, and calculated (i) the fraction of concordant over total sites, (ii) Non-Ref genotypes, (NRD, %) = number of non-reference genotypes and non-reference genotype discordance (NRD, in %) between WGS and Exomes; (iii) False discovery rate (FDR = FP=(FP + TP)), where it considered the exomes as the truth set; (iv) number of false positives (FP) and FDR for sites that are or not shared with the 1000 Genomes Project, PhaseI (1000GP); (v) false negative rate (FNR = FN=(FN + TP)), where AF bins were defined based on the 61 exomes. Furthermore, it compared 22 monozygotic (MZ) twin pairs at 880,280 bi-allelic SNV sites on chromosome 20, reporting (i) the percentage of concordant genotypes, non-reference genotypes and NRD. AF are from the set of 3,621 samples, which contains at most one of the two MZ twins from each pair. The discrepancies can be caused by errors in either twin, so the expected NRD to the truth would be half the NRD value given.

	WGS vs. Exomes						**MZ Twins**	
AF	**Total sites (concordant, %)**	**Non-Ref genotypes (NRD, %)**	**FP (FDR, %)**	**FP in 1000GP (FDR, %)**	**FP not in 1000GP (FDR, %)**	**FNR (%)**	**Total sites (concordant, %)**	**Non-Ref genotypes (NRD, %)**
AC=1	2,963 (99.999)	2,965 (0.1)	125 (4.0)	11 (3.8)	114 (4.1)	n.a.	411,583 (99.995)	3,534 (12.7)
AC=2	1,566 (99.998)	1,577 (0.1)	147 (8.6)	25 (7.9)	122 (8.7)	n.a.	101,116 (99.989)	1,594 (15.1)
0:03-1%	16,303 (99.928)	21,114 (3.3)	1,160 (6.6)	766 (5.5)	394 (11.3)	27.2	193,531 (99.954)	19,034 (10.2)
1-5%	16,356 (99.829)	53,165 (3.2)	1,038 (6.0)	980 (5.7)	58 (68.2)	6.4	50,360 (99.776)	56,554 (4.4)
>5%	37,433 (99.688)	1,151,178 (0.6)	2,668 (6.7)	2,653 (6.6)	15 (46.9)	7.3	123,690 (99.574)	1,382,934 (0.8)

Table 3.2 Descriptive for imputation reference panels

For UK10K, the following sites were excluded: 18,180,633singletons that do not exist in 1000GP, 1,064,168 multi-allelic sites and 214,631mis-matched alleles sites. For 1000GP, the following sites were excluded: 7,053,246 singletons that do not exist in UK10K, 23,932 sites with a SNP and an INDEL at the same position and 443 within large structural deletions.

	UK10K	**1000GP (Phase 1 v3)**	**Combined**	**Overlap**
N samples (% European)	3,781 (100%)	1,092 (34.7%)	4,873	--
N total sites in final release	45,492,035	39,527,072	--	
N total sites after filtering	26,032,603	32,449,428	42,359,694	16,122,337
Autosome SNPs	23,411,635	29,797,220	38,238,102	14,970,753
Autosome INDELs	1,698,262	1,370,819	2,407,858	661,223
Chr X SNPs	858,380	1,223,328	1,612,230	469,478
Chr X INDELs	64,326	58,061	101,504	20,883

3.3.2 Imputation evaluation on UK10K vs. 1000GP reference panels

As a first assessment of the UK10K reference panel, I performed a leave-one-out cross-validation on a sub-sample of 1,000 individuals from the UK10K WGS dataset (500 from TwinsUK and 500 from ALSPAC). I removed each sample from the reference panel in turn, selected 13,413 sites on chromosome 20 from the Illumina 610k bead chip, and imputed all other sites on this chromosome from a given reference panel. The imputation was conducted with three haplotype reference panels: the 1000GP panel, the "original" UK10K panel produced by initial genotype refinement and haplotyping with BEAGLE, and a "re-phased" UK10K panel that was generated by using SHAPEIT2 to estimate haplotypes from the BEAGLE genotypes. The accuracy of imputed variants was calculated as the Pearson correlation coefficient (r^2) between imputed genotype dosages in [0-2] and masked sequence genotypes in [0,1,2]. The results were stratified into non-overlapping MAF bins for plotting.

The results of this experiment are shown in **Figure 3.2A**, which focuses on variants with MAF<5%. Both UK10K reference panels (blue dotted and solid lines) produced higher accuracy than the 1000GP panel (black line), with greater gains at lower frequencies. These trends were expected due to the larger sample size and better ancestry matching of the UK10K reference panel to the pseudo-GWAS data. Notably, the UK10K reference panel yielded much higher imputation accuracy after re-phasing with SHAPEIT2 (solid vs. dotted blue lines): the mean r^2 at low frequencies increased by more than 0.1 (20%) after re-phasing, which implies a substantial boost in the power to detect associations. A large imputation panel is a resource that can inform a variety of association studies, so these results suggest that taking the time to improve a WGS panel's haplotype quality could have substantial downstream benefits. Most recently, I evaluated the added value of using UK10K WGS reference panel on top of the latest 1000GP reference panel (phase 3), based on a US population (FHS samples). I observed significant improvement when adding the UK10K panel on top of 1000GP. At the MAF of 0.002, 0.01, 0.1, the mean r2 value increased from 0.438, 0.522, 0.844 to 0.532, 0.621, 0.876 respectively. This evaluation was based on pseudo-GWAS of 320 FHS WGS samples.

3.3.3 Evaluation of metrics for choosing reference haplotypes

I noticed that some rare variants were imputed much better when using the entire UK10K reference panel to drive imputation, yet poorly when using IMPUTE2's k_{hap} approximation. This approximation reduces the computational cost of imputation by using a region-wide (e.g., across a 3MB imputation chunk) Hamming distance metric to reduce the number of reference haplotypes used by a given GWAS haplotype. The investigation of these variants led to the development of a new approximation that uses local (rather than region-wide) haplotype sharing to choose a subset of reference haplotypes. This was done by Bryan Howie. This approximation delivers the same substantial speed boost as the existing k_{hap} approximation, but it does not sacrifice imputation accuracy at rare and low-frequency variants. For example, **Figure 3.2 B** shows the results of imputing the INCIPE pseudo-GWAS data with the UK10K reference panel. The full UK10K panel produced the highest accuracy (solid blue line), while the k_{hap} approximation based on Hamming distance (solid orange line) was less accurate for SNPs with MAF<5%. By contrast, the new approximation based on haplotype tract sharing (dashed orange line) was nearly as accurate as the full reference panel, at ~10% of the computing time. Further speed improvements are possible for a modest price in accuracy. The evaluation of different K_hap (500 vs. 7562) and different sampling algorithm (tract sharing vs. hamming distance) was only run using the Italian isolates data. This is because imputing the UK10K pseudo-GWAS would need the leave-one-out approach, which would add an extra layer of complexity to the evaluation. Of note, the INCIPE pseudo-GWAS was generated from a SNP array data, not from WGS. Therefore, the number of variants masked out is much smaller and that in the UK10K pseudo-GWAS, and the r2 value between the two plots should be compared with this in mind.

The goal behind this new approximation is to ensure that each site in a study haplotype has the opportunity to copy the reference haplotype with the longest shared tract of allelic identity. The algorithm works as follows, from the point of view of a single GWAS haplotype:

1. For each reference haplotype, identify sets of contiguous sites that show no allele mismatches with the study haplotype; store these shared haplotype tracts for each reference haplotype.
2. At each site, generate a hash table whose keys are shared tract lengths (in genetic map units) and whose values are indices of the corresponding reference haplotypes. A given key can map to multiple values.

3. At each site, use the hash table created in the previous step to generate a list of reference haplotype indices ranked in descending order of shared tract length. Ties are broken at random.
4. Add the top-ranked haplotype index at each site to a list of unique reference haplotype indices; these states are marked for copying by the current study haplotype.
5. Go to the next-ranked haplotype index ("level") and repeat Step 4 until k_{hap} distinct reference haplotypes have been identified. If the number of selected haplotypes exceeds k_{hap} at a particular level, choose a random subset of the reference indices at that level such that the total number of selected haplotypes is k_{hap}.

The advantage of the newly proposed tract sharing metric was illustrated in **Figure 3.3**. The computational cost of imputing a study haplotype with the Hamming distance approximation is *O(MN)*, where *M* is the number of sites shared between the study and reference panels and *N* is the number of reference haplotypes. By comparison, the cost of this new tract length approximation is roughly *O(4MN)* – the factor of four appears because this approximation scans the sites in a region multiple times. While the tract sharing approximation requires more calculations, it is still linear in *M* and *N*, and the Hamming distance approximation accounts for less than 0.2% of a typical imputation run (as determined by profiling the IMPUTE2 C++ code when imputing the INCIPE pseudo-GWAS with the UK10K reference panel). In summary, the new tract sharing approximation has a similar computational cost to the Hamming distance approximation of (Howie et al. 2011), but it is better at maintaining imputation accuracy for low-frequency and rare SNPs. This will be a useful approach as imputation reference panels continue to grow.

Figure 3.2 Imputation performance for different reference panels and strategies

(A) Imputation accuracy in the UK10K pseudo-GWAS test panel using reference panels from 1000GP (black) and UK10K (blue). **(B).** Imputation accuracy in the INCIPE pseudo-GWAS panel using the UK10K reference panel and different imputation approximations.

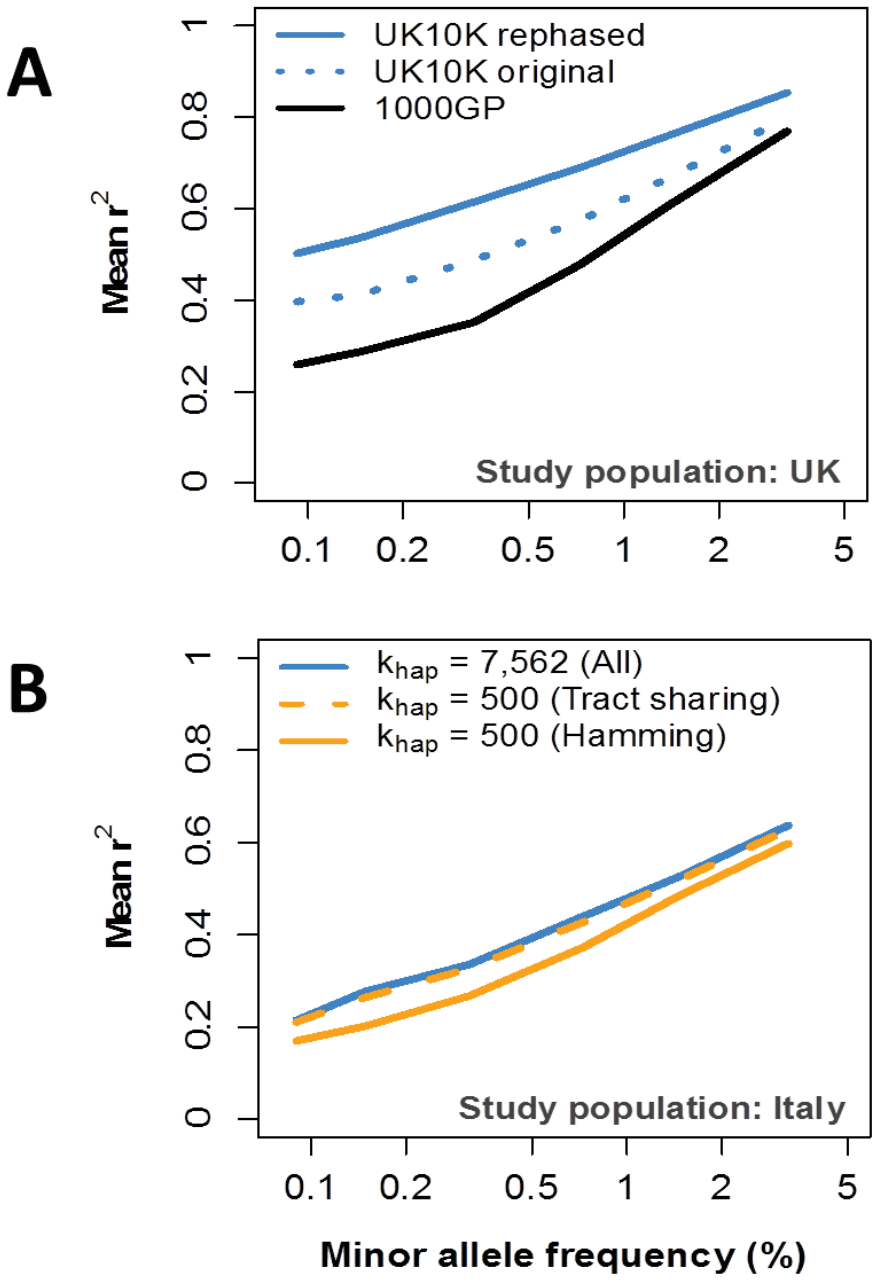
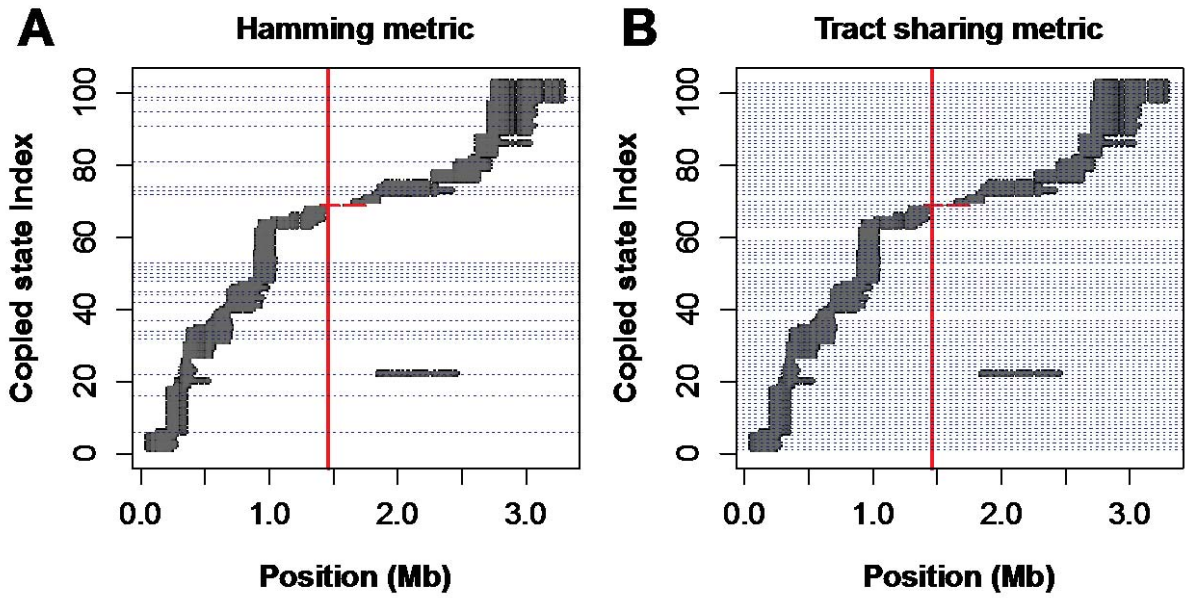


Figure 3.3 Illustration of reference states (haplotypes) copied by IMPUTE2

This figure is based on imputing one INCIPE pseudo-GWAS haplotype from the UK10K reference panel in a 3Mb region on chromosome 20. Points at each position on the chromosome (x-axis) represent reference haplotypes that were copied with marginal (per-site) posterior probabilities of at least 0.01 when using the full UK10K reference panel (7,562 haplotypes). Copied reference haplotypes are ordered on the y-axis by the position at which they first surpassed this threshold. The location of the SNP examined is marked by a vertical red line, and points belonging to the haplotype that carries this variant are also coloured red. Subsets of reference states selected by different approximations are marked by dotted blue lines. **(A)** Reference states selected with k_{hap}=500 under a Hamming distance approximation. Of the 103 copied states in this plot, 25 (24%) were chosen under this approximation. **(B)** Reference states selected with k_{hap}=500 under a tract sharing approximation. Of the 103 copied states in this plot, 96 (93%) were chosen under this approximation.

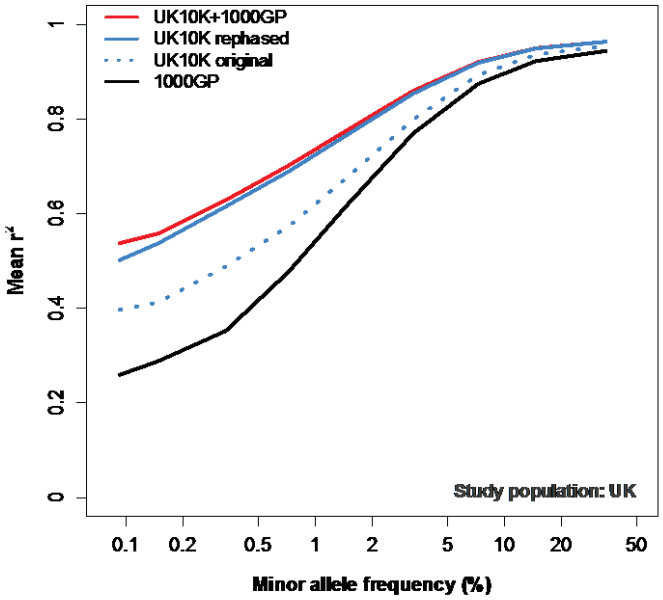


3.3.4 Evaluation of combining two reference panels

Figure 3.4 shows how a combined 1000GP+UK10K panel (red) produced by this method performed against each panel separately (1000GP, black; UK10K, blue) when imputing a pseudo-GWAS of UK ancestry. The combined and UK10K panels produced very similar numbers of high-confidence (predicted r^2>0.8) variants at MAFs of 0.5% and higher, implying that the combined panel is neither helpful nor harmful for imputing common and low-frequency variants when a large, population-specific panel is available. On chromosome 20, the combined panel added 2,263 high-confidence rare variants that were not captured by the UK10K panel (MAF<0.5%; 4% increase), which could reflect mutations that have drifted to very low frequencies in the UK but persist on the same haplotype background elsewhere in Europe (Howie et al. 2011, Jewett et al. 2012). A similar result was observed when the imputation was run for a population in northern Italy (INCIPE cohort). The INCIPE cohort was newly genotyped in this study, using Illumina HumanCoreExome-12v1-1 arrays. After stringent quality control, the genotype data of chromosome 20 was split into an imputation panel (containing 6,300 SNPs genotyped in 2,145 study participants) and a test panel, corresponding to the exome content of the array (2,522 SNPs, all with MAF≤5%). In this dataset the UK10K reference panel outperformed the 1000GP panel in all frequency bins, despite the fact that the 1000GP includes a panel (TSI, or "Toscani in Italia") that is genetically more similar to the study population. As before, the combined 1000GP+UK10K panel yielded a larger number of high-confidence imputed variants than the UK10K panel alone – here, the combined panel added 3,729 well-imputed variants with MAF<0.5%, for a 20% increase in rare variants over the UK10K panel. These results suggest that it can be especially useful to combine the strengths of multiple panels when a large, population-specific reference set is not available for a particular GWAS population.

Figure 3.4 Performance of combining UK10K and 1000GP panels

Imputation accuracy in the UK10K pseudo-GWAS test panel using reference panels from 1000GP (black), UK10K (blue), and UK10K+1000GP (red) across all MAFs. The rephased UK10K panel was combined with the 1000GP panel to produce the UK10K+1000GP panel.



3.4 Conclusion & Discussion

As WGS becomes a standard tool for population and disease genetics, there will be many questions about how to design sequencing studies, how to process the data, how to combine data across studies, and how to limit the computational costs of downstream analysis. With data from one of the most ambitious population sequencing studies to date, the above evaluations have demonstrated the value of a large, UK-specific reference panel for imputation in British cohorts and in other European populations. I showed that the UK10K reference panel greatly increases accuracy and coverage of low-frequency variants relative to a panel of 1,092 individuals from the 1000GP. The results show that state-of-the-art phasing methods like SHAPEIT v2 are essential for creating high-quality haplotype panels. Combining WGS data across studies is a desirable goal, which is now available in IMPUTE2 that can integrate sets of phased haplotypes to produce a unified reference panel. The combined panel is much larger than the 28.6 million imputable sites in the UK10K panel or 32.5 million imputable sites in the 1000GP panel. Finally, due to observations from my evaluation, a new approximation in IMPUTE2 was implemented that helps reduce the trade-off between imputation speed and accuracy as reference panels continue to grow.

As shown in chapter 2, sizable reductions in the magnitude of the effect sizes can be identified at any sample size through the use of the UK10K reference panel and the improved imputation quality. For instance, for a variant of MAF = 0.3% we have equivalent power when imputing from UK10K+1000GP into a 3,621 sample as we have when using the 1000GP imputation panel alone with 10,000 samples (**Figure 2.4a**). Similar, although weaker, increases in power were seen for region-based tests of rare variants. Although absolute power in **Figure 2.4b** is generally poor, there is demonstrable power improvements when data are better imputed or are directly sequenced (**Figure 2.4c**). The benefits of combining two reference panels in improving imputation for rare variants, as demonstrated in this study, could provide a good reference to future efforts that aim to combine a lot more WGS datasets. For example, the Haplotype Reference Consortium (http://www.haplotype-reference-consortium.org/) so far combined WGS from 20 cohorts with more than 30,000 whole genomes. This is expected to significantly improve imputation especially for samples whose ancestries are not as well represented in the 1000GP or in UK10K.

In summary, my recommendation for future WGS based imputation would include the following: 1. pre-phase WGS panel with SHAPEIT; 2. combining two reference panels; 3. if computation cost is not an issue, use all haplotypes, otherwise, using the new IMPUTE2 to pick the top haplotypes; 4. run evaluations and check output data to confirm that the best strategy was adopted and the desirable imputation performance was achieved.

4 Lipids

4.1 An introduction to lipids.

4.1.1 Biology and physiology circulating lipids

Lipids are a group of naturally occurring molecules that include fats, sterols, fat-soluble vitamins, triglycerides (TG), phospholipids, and others. The main biological functions of lipids include storing energy, signalling, and acting as structural components of cell membranes. The most familiar type of animal sterol is **cholesterol**, which is vital to animal cell membrane structure and function and a precursor to fat-soluble vitamins and steroid hormones. Cholesterol is transported inside **lipoproteins**. Lipoproteins are named based on their size and density; the lower the density, the larger the particle (Lusis and Pajukanta 2008, Ramasamy 2014). The density of lipoprotein is positively determined by the protein to lipid ratios. In order of increasing density, lipoproteins include chylomicrons, very-low-density lipoprotein (VLDL), LDL, intermediate-density lipoprotein (IDL), and high-density lipoprotein (HDL) (Olson 1998). Lipoproteins contain **apolipoproteins**, which bind to specific receptors on cell membranes and determine the starting and ending points of cholesterol transport. Chylomicrons, the least dense cholesterol transport molecules, carry fats from the intestine to muscle and other tissues in need of fatty acids for energy or fat production. Unused cholesterol remains in cholesterol-rich chylomicron remnants and is taken up to the bloodstream by the liver.

LDL particles are the major blood cholesterol carriers. Its molecule shells contain apolipoprotein B100, which is recognized by LDL receptors in peripheral tissues. The identification of the LDL receptor dramatically improved our understanding of cholesterol metabolism (Brown and Goldstein 1976). Excessive LDL molecules not bound by LDL receptors appear in blood circulation. When oxidized and taken up by macrophages, these LDL molecules become engorged and form foam cells, which often become trapped in the walls of blood vessels to form atherosclerotic plaques. HDL particles transport cholesterol back to the liver for excretion or for other tissues that synthesize hormones, in a process known as reverse cholesterol transport (RCT) (Lewis and Rader 2005). Because of the

function of HDL and LDL particles, the enzymatically measured HDL and LDL levels are often referred to as “good” and “bad” cholesterol, respectively.

TG is an ester derived from glycerol and three fatty acids, and it is the main constituents of vegetable oil (typically more unsaturated) and animal fats (typically more saturated). As a blood lipid, TG enables the bidirectional transference of adipose fat and blood glucose from the liver, playing an important role in metabolism as energy sources and transporters of dietary fat. Lipoprotein lipases on the walls of blood vessels break down TG into free fatty acids and glycerol so that it can pass through cell membranes. Fatty acids can then be taken up by cells via the fatty acid transporter.

4.1.2 Lipids as risk factors for CVD

TC and LDL as CVD risk factors

Large epidemiological studies have established serum level of total cholesterol (TC) especially LDL as major risk factors for CHD (Arsenault et al. 2011). This was later confirmed by MR studies (Cohen et al. 2006) and clinical trials (Shepherd et al. 1995, Downs et al. 1998, Heart Protection Study Collaborative 2002, Badimon et al. 2010). It was estimated that 1 mmol/L reduction in LDL level is associated with a 23% reduction in CHD events (Cholesterol Treatment Trialists et al. 2010), a 12% reduction in all-cause mortality, a 19% reduction in CHD-related mortality (Baigent et al. 2005). The association is log linear with no threshold below which benefit ceases. However, the association of TC or LDL with stroke is not as strong as that with CHD. One study reported that TC was weakly positively related to ischaemic and total stroke mortality in early middle age (40-59 years), and the association could be largely accounted for by the association between TC and blood pressure (Prospective Studies et al. 2007). The weak association with stroke could be due to the fact that stroke is a heterogeneous condition and various causes of ischemic stroke may have different associations with cholesterol (Amarenco et al. 2004, Amarenco and Steg 2007). Nevertheless, randomized trials of statin therapy have shown that reduction of LDL by about 1.5 mmol/L could reduce by about a third the incidence not only of ischemic heart disease but also of ischemic stroke, independently of age, BP or pre-randomization lipid concentrations (Baigent et al. 2005). Statin is the most widely used cholesterol lowering drug, developed

based on the discovery of the fungal metabolite ML-236A and ML-236B (Endo et al. 1976, Kuroda et al. 1979). These lipid modification therapies (LMTs) have revolutionised contemporary approaches to primary and secondary prevention of CVD (Webb et al. 2013).

The understanding that all cholesteryl esters transported by lipoproteins other than HDL (including LDL, VLDL, IDL, and chylomicron remnants) are atherogenic has led to the concept that non-HDL-c levels (TC minus HDL-c) might be more strongly associated with CVD risk than LDL-c alone (Robinson 2009). Several investigators have shown that the ratio between these particles predicts CVD risk better than isolated lipoprotein sub-fractions (Lemieux et al. 2001, Ingelsson et al. 2007, Kannel et al. 2008, Arsenault et al. 2009). The most widely used ratios including TC/HDL, followed by TG/HDL (Castelli 1988). In clinical trials, measuring Apo-B, or Apo-B/Apo-AI ratio also has advantages to assess the efficacy of lipid-lowering therapies.

HDL as CVD risk factors

The FHS first reported that HDL had an inverse association with the incidence of CHD (Gordon et al. 1977). This was later confirmed by other studies (Assmann et al. 1996, Goldbourt et al. 1997). It was estimated that 1 mg/dL increase of HDL is associated with a 1.9 to 2.3% reduction in cardiovascular risk in men and 3.2% in women. This relationship holds even for individuals with low level of LDL (Gordon et al. 1989). The atheroprotective effect of HDL has been mainly attributed to RCT. Over the past few years, other features of HDL have been suggested, including anti-inflammatory, immunomodulatory, antioxidant, antithrombotic, and endothelial cell repair effects (Choi et al. 2006, Ibanez et al. 2007, Badimon et al. 2010).

Although several lifestyle related approaches have demonstrated the ability to increase HDL and improve CVD outcomes (Choi et al. 2006), Mendelian randomization using variants associated with HDL at the *LCAT*, *CETP*, *APOA1*, *ABCA1*, *LIPC*, and *LIPG* loci have largely failed to support a strong causal relationship between HDL and risk of CAD (Frikke-Schmidt et al. 2008, Johannsen et al. 2009, Ridker et al. 2009, Haase et al. 2012, Voight et al. 2012). In clinical trials, Torcetrapib, an inhibitor for cholesteryl ester transfer protein (CETP), showed a significant increase in HDL-c levels but also led to an increase in cardiovascular events and total mortality (Barter et al. 2007, Barter 2009). Small peptides that mimic some of the properties of apolipoprotein A-I (Apo-AI) have been shown to improve HDL function and reduce atherosclerosis without altering overall HDL levels (Navab et al. 2011). It was reasoned that the quality of HDL, rather than the quantity, may influence its

atheroprotective effects. In a more recent clinical trial, a high dose of quinazoline molecule RVX-208 was used to stimulate increased synthesis of endogenous Apo-AI and provided some encouraging results (Nicholls et al. 2011). Detailed proteomic and lipidomic analyses are needed to provide further new insights into the heterogeneous efforts of various HDL compositions. Novel pharmaco-therapeutic strategies directed at HDL include augmenting Apo-AI levels directly and indirectly, mimicking the functionality of Apo-AI, and enhancing steps in the RCT pathways (Degoma and Rader 2011).

TG as CVD risk factor

Serum TG level has been reported for positive association with incidence of CVD (Bansal et al. 2007, Nordestgaard et al. 2007, Sarwar et al. 2007). In 2009, a large meta-analysis based on more than 300,000 individual from 68 long-term prospective studies reported that TG was no longer an independent risk factor for CVD (including non-fatal MI, CHD death, stroke) after adjustment for other risk factors (Emerging Risk Factors et al. 2009). This study indicated that CVD outcomes might be influenced by correlates of TG (such as non-HDL, HDL, or LDL) and TG is a marker instead of a risk factor for CVD. In the same year, another meta-analysis of 31 studies reported a positive association between TG and stroke, with a note for the need for additional large prospective studies especially in stroke subtypes to firmly establish the independent nature of the effect (Labreuche et al. 2009).

There is more evidence for a causal role of TG from MR studies. In 2010, the Triglyceride Coronary Disease Genetics Consortium and Emerging Risk Factors Collaboration first showed a causal association between triglyceride-mediated pathways and coronary heart disease (Triglyceride Coronary Disease Genetics et al. 2010). The instrumental variable used in this study is a single SNP in the promoter of the *APOA5* gene (-1131T>C，rs662799), which directly affects TG metabolism while is only indirectly associated with other lipid parameters including LDL. Another MR study included 185 common variants in a model that accounted for effects on HDL and LDL and also concluded the causal role of TG (Do et al. 2013). A recent WES study for early-onset MI found that carriers of rare non-synonymous mutations in *APOA5* had higher plasma TG and increased risk for MI (Do et al. 2014). Rare mutations that disrupt *APOC3*, a gene in close proximity to and functionally related to *APOA5*, were also associated with a lower level of TG and a reduced risk for CHD (Tg et al. 2014) and ischemic CVD (Jorgensen et al. 2014). These

evidences support that disordered metabolism of TG-rich lipoproteins contributes to CVD risk.

4.1.3 Genetic determinants of lipids levels

Disruptions in the lipoprotein metabolism can cause many different kinds of dyslipidemias depending on the particle or enzyme that is affected. Most of these lipid related syndromes are caused by a mutation in a single gene, i.e., monogenic, and are inherited based on Mendelian laws. There are two major groups of lipid related syndromes: hyperlipidemias and lipoprotein deficiency disorders. Hyperlipidemias are syndromes where lipoprotein levels are elevated in blood and are further classified into different categories (Fredrickson and Lees 1965). It is estimated that genetic and environmental factors have a roughly equal impact on the variation of plasma levels of lipids, with heritability around 50% (Beekman et al. 2002, Pilia et al. 2006, Weiss et al. 2006, Goode et al. 2007). The discovery of genetic factors influencing or even causing lipid level variations is very important for translational medical advances. For example, low-frequency coding variants in *PCSK9* were found to play a causal role in lowering LDL level and protecting against risk of CHD (Abifadel et al. 2003, Allard et al. 2005), which led to the development of a new class of drugs for lowering plasma LDL level (Stein et al. 2012).

Findings from candidate gene and linkage analysis

So far, a total of 26 monogenic genes with causative mutations for dyslipidemia were reported (Kuivenhoven and Hegele 2014) (**Table 4.1**). About half of these were discovered through candidate gene studies with *a priori* knowledge of the protein products. Another ~ 20% of causative gene mutations for monogenic dyslipidemias were found using genetic mapping approaches such as linkage analysis. The availability of patients and families with extreme dyslipidemia is essential in these studies. High throughput approaches including WES have confirmed the role of previously established genes and identified a small number of new causes of monogenic dyslipidemias. Out of 20 loci for genes causing severe changes in lipid metabolism, 16 have also shown association in GWAS, and four of these overlapping loci include genes that are known drug targets (**Figure 4.1**).

Findings from first generation GWAS

Since 2007, a total of 34 GWAS studies have been conducted to discover genetic variations underlying lipids, most of them are based on individuals of European ancestry (**Table 4.2**). The two biggest one are published in 2010 (Teslovich et al. 2010) and in 2013 (Global Lipids Genetics et al. 2013). The former reported 95 loci in total while the latter added 62 more loci with nearly ~200,000 samples, leading to a total of 157 loci. Among the 62 new loci, 32 have some previous connection within lipoprotein metabolism. Among the 157 GWAS loci, 65 show significant associations with two or more of the four main lipid traits, four of which (*CETP, TRIB1, FADS1-2-3, APOA1*) show associations with all lipids traits. However, there is still an overall lack of new knowledge of lipids, given the adequate power of these studies. The phenotypic variation explained by these new GWAS loci is also low, with ~2% of the variation explained by the 62 new loci, which increases the total explained by all GWAS loci to ~15% (Global Lipids Genetics et al. 2013). Nevertheless, further functional studies have begun to emerge and showed promising results. Besides reporting the largest number of novel lipids loci based on statistical significance, the Global Lipids Genetics study also conducted further functional analyses including association with mRNA expression levels and pathway analyses to uncover relationships between lipids loci and those of genes and other functional elements in the genome. The results provided direction for biological and therapeutic research into risk factors for CAD.

Findings from next generation sequencing

Next generation sequencing (on both DNA and RNA) are yielding tremendous successes for discovering novel genes and novel mutations underlying single gene syndromic disorders across a wide range of disease entities and disciplines (Boycott et al. 2013). For lipids, sequencing studies on candidates genes revealed a burden of rare missense or nonsense variants for individuals with low plasma HDL-c levels in the general population (Cohen et al. 2004) and patients with hypertriglyceridemia (Johansen et al. 2010). Next generation sequencing especially WES was first applied to patients with familial dyslipidemia, but has thus far mostly confirmed already known loci instead of finding novel mutations (**Table 4.3**). A recent WES study on 2,005 individuals including 554 with extreme levels of LDL identified significant associations of rare or low frequency variants in known LDL modifying genes such as *PCSK9, LDLR,* and *APOB,* as well as for a novel gene *PNPLA5*. This study

reported that the effect sizes for the burden of rare variants for each associated gene were substantially higher than those observed for individual SNPs identified from GWASs (Lange et al. 2014). Exome chip is a cost-effective alternative to WES. An exome-chip based study with > 200,000 low-frequency and rare coding sequence variants in 56,538 individuals identified new low-frequency variants in four known genes with large effects on HDL-C and/or triglycerides (Peloso et al. 2014). None of these four variants was associated with risk for CHD, suggesting that examples of low-frequency coding variants with robust effects on both lipids and CHD will be limited. Another recent exome-chip based study with ~80,000 coding variants in 5,643 individuals identified a variant that encodes p.Glu167Lys for association with TC and the risk of MI. It is within a locus previously known as *NCAN-CILP2-PBX4* or 19p13 (Holmen et al. 2014).

Based on limited studies reported so far, applying NGS to general healthy population did not yield many novel findings either. Nevertheless, the effect sizes from the burden of rare variants are substantially higher than those from single marker based analysis, therefore supporting a strategy for rare variants aggregation tests. WGS study on lipids was first reported in 2013, with ~1,000 samples with 6X coverage sequencing (Morrison et al. 2013). This study estimated that common and low frequency variation contributes more to heritability of HDL levels (61.8%) than rare variation (7.8%). It also highlighted the value of regulatory and non-protein-coding regions of the genome in addition to protein-coding regions.

Table 4.1 Gene discovery in monogenic dyslipidemias

This table is adopted from (Kuivenhoven and Hegele 2014), listing the single gene causes for the main dyslipidemia states encountered in the clinic, subdivided according to the primary lipid disturbance.

Gene	Discovery	References
Elevated LDL		
ABCG5/G8	Linkage mapping	(Berge et al. 2000)
APOB	A priori knowledge of protein	(Soria et al. 1989)
LDLRAP1	Linkage mapping	(Garcia et al. 2001)
LDLR	A priori knowledge of protein	(Lehrman et al. 1985)
LIPA	WES plus a priori knowledge of protein	(Stitziel et al. 2013)
PCSK9	Linkage analysis	(Abifadel et al. 2009)
Depressed LDL		
ANGPTL3	Mouse studies plus WES	(Musunuru et al. 2010)
APOB	A priori knowledge of protein	(Young et al. 1987)
PCSK9	Linkage analysis plus sequencing	(Cohen et al. 2005)
MTTP	A priori knowledge of protein	(Sharp et al. 1993)
SAR1B	Linkage mapping	(Jones et al. 2003)
MYLIP (IDOL)	In vitro studies (Zelcer et al. 2009)	(Sorrentino et al. 2013)
Elevated HDL		
CETP	A priori knowledge of protein	(Brown et al. 1989)
LIPC	A priori knowledge of protein	(Hegele et al. 1991)
Depressed HDL		
APOA1	A priori knowledge of protein	(von Eckardstein et al. 1989)
LCAT	A priori knowledge of protein	(Funke et al. 1991)
ABCA1	Linkage mapping	(Rust et al. 1999)
Elevated TG		
APOA5	Bioinformatics	(Marcais et al. 2005)
APOC2	A priori knowledge of protein	(Cox et al. 1978)
APOE	A priori knowledge of protein	(Cladaras et al. 1987)
GPD1	Linkage mapping	(Basel-Vanagaite et al. 2012)
GPIHBP1	mutant mouse	(Beigneux et al. 2009)
LMF1	mouse study	(Peterfy et al. 2007)
LPL	A priori knowledge of protein	(Emi et al. 1990)
SLC25A49	Linkage studies plus WES	(Rosenthal et al. 2013)
Depressed TG		
APOC3	GWAS in isolate	(Pollin et al. 2008)

Figure 4.1 Lipids loci overlap between candidate gene studies and GWAS

This figure is modified and updated from (Kathiresan and Srivastava 2012)

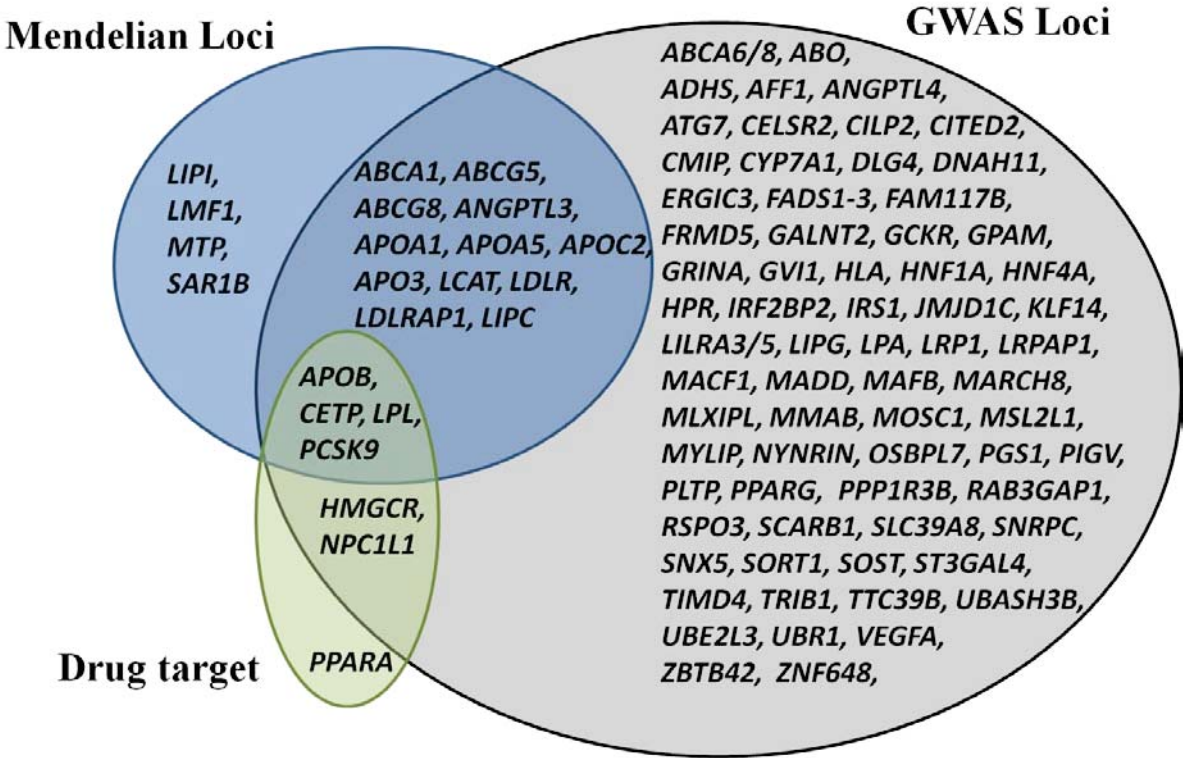


Table 4.2 GWAS studies of lipids

Date is for publication date. Samples are all European ancestry unless explicitly specified otherwise: FIN for Finnish, CHN for Chinese, KOR for Korean, JAP for Japanese, AA for African American, MEX for Mexican, HIS for Hispanics. The sample size before "+" is for discovery while the sample size after "+" is for replication.

Date	Sample size	Main findings	Reference
2007-04	1464 T2D +1467	A locus in *GCKR* with TG	(Saxena et al. 2007)
2007-09	1,087 + ~8,100	No replicated associations	(Kathiresan et al. 2007)
2008-01	1,955 + 2,033	Replicated PSRC1 and CELSR2	(Wallace et al. 2008)
2008-01	8,656+11,437	11 known loci	(Willer et al. 2008)
2008-01	2,758+18,544	6 new loci	(Kathiresan et al. 2008)
2008-01	1,005+6,827	A missense SNP in MLXIPL for TG	(Kooner et al. 2008)
2008-02	11,685+4,979	2 novel variants for LDL	(Sandhu et al. 2008)
2008-09	2,346 Kosrae	3 SNPs in HMGCR for LDL	(Burkhardt et al. 2008)
2008-10	4,274+15,873	CETP and LPL for HDL	(Heid et al. 2008)
2008-10	6,382 + 970	5 novel loci for lipids	(Chasman et al. 2008)
2008-12	19,840+20,623	30 loci including 11 novel	(Kathiresan et al. 2009)
2008-12	4,763 FIN	9 novel loci	(Sabatti et al. 2009)
2008-12	21,848 and 714	6 novel and 16 known for lipids	(Aulchenko et al. 2009)
2008-12	809 + 698 Amish	A null mutation in APOC3	(Pollin et al. 2008)
2009-02	18,245	SNPs at CETP predicts MI risk	(Ridker et al. 2009)
2009-04	900 + 1,810 JAP	variants at CETP for HDL	(Hiura et al. 2009)
2009-11	17,296 + 2700	10 novel loci for lipids	(Chasman et al. 2009)
2010-01	656 + 3,282	2 novel loci	(Igl et al. 2010)
2010-02	8,993 JAP	46 novel loci for blood and lipids traits	(Kamatani et al. 2010)
2010-04	6,078 + 1,231	2 novel loci for lipids	(Ma et al. 2010)
2010-08	100,184	59 novel and 36 known loci	(Teslovich et al. 2010)
2010-09	17,723 + 37,774	4 novel loci for lipids	(Waterworth et al. 2010)
2011-09	12,545+30,395 KOR	10 novel loci for metabolic traits	(Kim et al. 2011)
2011-11	32,225 + 11,509	1 new locus for TC	(Surakka et al. 2011)
2011-12	1,999+1,496 CHN	1 novel locus	(Tan et al. 2012)
2012-01	8,330 FIN	11 novel loci for metabolic traits	(Kettunen et al. 2012)
2012-08	1867 EMR based	A strong protective variant in APOE	(Rasmussen-Torvik et al. 2012)
2012-12	1,720 + 1,261 twins	1 locus related to variability of HDL	(Surakka et al. 2012)
2013-03	2,240 + 2,121 MEX	A novel locus for TG	(Weissglas-Volkov et al. 2013)
2013-05	7,917 AA, 3,506 HIS	striking similarities across populations	(Coram et al. 2013)
2013-09	1,782 + 1,719 FIL	2 known loci: APOE, APOA5	(Wu et al. 2013)
2013-09	839+5,248 Sorbs	1 novel locus	(Keller et al. 2013)
2013-10	94,595 + 93,982	62 novel and 95 known loci	(Willer et al. 2013)
2013-12	3,451 + 8,830 CHN	Replicated 8 known loci	(Zhou et al. 2013)

Table 4.3 NGS studies on lipids

There are five small scale sequencing studies on patients with familial dyslipidemia and three studies on healthy populations with relatively large sample size. WES, WGS, and exome-chip technologies were used for each of the three studies on healthy population. Samples are all European ancestry unless explicitly specified otherwise.

Date	Sample size	Main findings	Reference
Familial dyslipidemia			
2010-10	WES on 2	ANGPTL3 mutations for familial combined hypolipidemia	(Musunuru et al. 2010)
2010-11	WGS of 1	two nonsense mutations in ABCG5 caused sitosterolemia	(Rios et al. 2010)
2012-03	WES on 1 family	novel APOB mutation for ADH	(Motazacker et al. 2012)
2012-10	WES on 14	heterozygous in-frame deletion in the APOE gene for ADH	(Marduel et al. 2013)
2013-09	WES on 3	a homozygous splicing mutation in LIPA for hypercholesterolemia	(Stitziel et al. 2013)
Healthy population			
2013-06	WGS of 962	HDL Heritability mainly explained by common variants	(Morrison et al. 2013)
2014-01	WES of 2,005	LDL and the burden of rare variants in PNPLA5	(Lange et al. 2014)
2014-03	X-chip of 5,771	causal variant in *TM6SF2* influencing TC and MI	(Holmen et al. 2014)

.

4.1.4 Aims of this study

Under the framework of the UK10K project (The UK10K Consortium 2015), this study aims to identify novel genetic variants that are associated with plasma lipids levels and also fine map known lipids loci with WGS data. The current study is by far the largest WGS based association study of lipids, with up to 3,210 WGS samples and more than 22,000 samples with WGS imputed data. I first analyse the WGS samples aiming to discover rare and low frequency variants with large effect sizes. Then I analyse a much larger group of cohorts with imputed data to discover novel associations across the full MAF spectrum. Besides single marker based genome-wide scan, this study is able to fine map known loci and investigate the association and contribution of rare variants to serum lipids variance. This work will not only contribute to the understanding of the allelic architecture of lipid variation in healthy population but also provide a good reference for using WGS data to study complex traits in general.

4.2 Methods

4.2.1 Cohorts & phenotype measurements

There were a total of 14 cohorts included for the expanded discovery, including both WGS and the SNP-array imputed samples for TwinsUK and ALSPAC, plus 10 other cohorts where genome-wide SNP data and raw lipids phenotypes were made available (**Table 4.4**). There were 11 more cohorts included for stage-1 replication. Some of them had genome-wide results as well, but only the top hits from the expanded discovery were queried from the replicate data. For the final few replicated variants, I used the WHI data for a further replication. The details of these cohorts were given in chapter 2.

Lipids measurement methods were as following: for **ALSPAC**, plasma levels of TC, HDL and TG were measured with enzymatic colorimetric assays (Roche) on a Hitachi Modular P Analyser. LDL was derived from the following formula: TC- (HDL+TG/2.19); for **TwinsUK,** Enzymatic colorimetric assays were used to measure serum levels of TC, HDL and TG were measured using three analysing devices (Cobas Fara; Roche Diagnostics, Lewes, UK; Kodak Ektachem dry chemistry analysers (Johnson and Johnson Vitros Ektachem machine, Beckman LX20 analysers, Roche P800 modular system)); for **1958BC**, serum TG, TC and HDL were measured in serum by Olympus model AU640 autoanalyser in a central lab in Newcastle. Enzymatic colorimetric determination GPO-PAP method was used to determine TG, CHOD-PAP method for TC and for HDL; for **INGI-VB**, lipids were measured using HITACHI 917 ROCHE and Unicel Dx-C 800 BECKMAN devices; for **INGI-FVG** and **INGI-Carl**, lipids were measured using BIOTECNICA BT-3000 TARGA chemistry analyser; for **INCIPE**, enzymatic determination of TC and TG was performed on Dimension RxL apparatus (Siemens Diagnostics). HDL cholesterol was determined by the homogeneous method; LDL cholesterol by the Friedewald formula (Friedewald et al. 1972); for **LURIC**, TC and TG were obtained by ß-quantification from serum and measured enzymatically using WAKO reagents on a WAKO 30R analyser (Neuss, Germany). LDL and HDL were measured after separating lipoproteins with a combined ultracentrifugation-precipitation method; for **HELIC Manolis** and **HELIC Pomak** and **Teenage**, TC, HDL, TG were assessed using enzymatic colorimetric assays and while LDL levels were calculated according to Friedewald equation (Friedewald et al. 1972). For WHI, HDL, LDL, and TG

measurements were performed at the University of Minnesota by standard biochemical methods on the Roche Modular P Chemistry analyzer (Roche Diagnostics): HDL was measured in serum by the HDL-C plus third generation direct method; TG was measured in serum by Triglyceride GB reagent, and total cholesterol (TC) was measured in serum by a cholesterol oxidase method. LDL was calculated in serum specimens having a TG value < 400 mg/dl according to the formula of Friedewald et al. [Based on the LDL-lowering effects of statins, we estimated the pretreatment LDL value for individuals on lipid-lowering medication by dividing treated LDL values by 0.75.

For phenotype harmonization, extra care was given to the TwinsUK cohorts given there was random efforts of different dates of visits and different instrumental measurements (**Table 4.5)**. For ALSPAC and other cohorts in expanded discovery and replication, the same phenotype protocol was used. Inverse normal transformation was applied to all cohorts. For each cohort, the residuals with confounding variables regressed out were standardized so that the phenotype had a mean of 0 and a standard deviation of 1.

Table 4.4 Characteristics of participating cohorts

All cohorts are population based, except for TwinsUK. Imputation was conducted with the 1000G and UK10K combined reference panel, unless otherwise specified. Age is in mean (range). Traits values are in the format of mean (SD). For each trait of each cohort, the residuals with confounding variables regressed out were standardized so that the phenotype has a mean of 0 and a standard deviation of 1.

	Study	N	Country	Age	% Female	HDL	LDL	TG	TC
discovery	ALSPAC WGS	1,497	UK	10 (9-11)	50.3	1.40 (0.01)	2.31 (0.01)	1.14 (0.01)	4.24 (0.02)
	TwinsUK WGS	1,713	UK	56 (17-85)	100.0	1.79 (0.01)	3.16 (0.02)	1.12 (0.01)	5.48 (0.03)
	ALSPAC GWA	2,820	UK	10 (9-12)	49.2	1.40 (0.01)	2.36 (0.01)	1.14 (0.01)	4.28 (0.01)
	TwinsUK GWA	1,896	UK	50 (16-83)	81.1	1.51 (0.01)	3.33 (0.03)	1.18 (0.02)	5.38 (0.03)
	1958 BC	5,493	UK	44 (44-44)	52.3	1.56 (0.01)	3.42 (0.01)	2.07 (0.02)	5.88 (0.01)
	INGI-Carl	413	Italy	50 (18-83)	60.0	--	--	1.48 (0.04)	5.30 (0.06)
	INGI-FVG	1,394	Italy	52 (18-92)	58.2	1.38 (0.01)	3.71 (0.03)	1.30 (0.02)	5.69 (0.03)
	INGI-VB	1,776	Italy	55 (18-102)	56.3	1.52 (0.01)	3.23 (0.02)	1.19 (0.02)	5.3 (0.03)
	INCIPE1	653	Italy	60 (35-89)	54.4	1.49 (0.01)	3.49 (0.03)	1.18 (0.03)	5.52 (0.04)
	INCIPE2	1,382	Italy	58 (26-95)	50.9	1.49 (0.01)	3.39 (0.02)	1.10 (0.02)	5.39 (0.03)
	LURIC-Ctrl	983	Germany	61 (17-91)	60.8	1.07 (0.01)	3.21 (0.03)	1.82 (0.04)	5.22 (0.03)
	HELIC MANOLIS	1,264	Greece	62 (18-99)	57.2	1.32 (0.01)	3.22 (0.03)	1.56 (0.03)	5.57 (0.08)
	HELIC POMAK	999	Greece	43 (13-87)	72.1	1.15 (0.01)	3.15 (0.03)	1.52 (0.03)	5.01 (0.03)
	TEENAGE	557	Greece	13 (11-18)	55.9	1.44 (0.01)	2.33 (0.02)	0.67 (0.01)	4.09 (0.03)
replication	LOLI-EW610	905	UK	56 (35-75)	26.8	1.42 (0.01)	3.46 (0.03)	1.54 (0.04)	5.57 (0.03)
	LOLI-EWA	566	UK	55 (23-75)	13.1	1.30 (0.01)	3.16 (0.04)	1.70 (0.05)	5.21 (0.05)
	LOLI-EWP	610	UK	56 (32-67)	0.0	1.26 (0.01)	3.06 (0.04)	1.83 (0.06)	5.13 (0.04)
	RS-1	2981	NL	69 (48-75)	41.2	1.06 (0.01)	3.21 (0.04)	1.262 (0.06)	6.06 (0.04)
	RS-2	1823	NL	67 (51-75)	47.7	1.29 (0.01)	3.22 (0.03)	1.23 (0.03)	6.12 (0.04)
	GoT2D	2076	UK	NA	NA	NA	NA	NA	NA
	InChianti	621	Italy	56 (47-71)	56.3	1.53 (0.01)	3.36 (0.03)	1.28 (0.02)	4.99 (0.03)
	FinRisk	817	Finland	56 (47-68)	46.8	1.4 (0.03)	3.11 (0.05)	1.68 (0.04)	5.78 (0.05)
	Fenland	8701	UK	65 (47-77)	46.2	1.43 (0.01)	3.21 (0.02)	1.65 (0.02)	5.12 (0.03)
	UCLEB-BRHS	2742	UK	69 (58-81)	0.0	1.15 (0.01)	3.89 (0.02)	2.05 (0.03)	6.36 (0.02)
	UCLEB-BWHHS	3309	UK	71 (60-81)	100.0	1.62 (0.01)	4.14 (0.03)	1.91 (0.02)	6.62 (0.03)
	WHI	10,999	US	51 (44-69)	100.0	1.36 (0.02)	3.11 (0.04)	1.93 (0.06)	5.27 (0.05)

Table 4.5 Phenotype harmonization protocol for lipids traits

Analyser was tested as a random effect variable, while the others including age and age^2 are tested as fixed effect covariates.

Dataset	Trait	Transformation	Gender stratified	Co-variates tested	Filter	Analyser
ALSPAC WGS+GWA	HDL	inverse normal	yes	age, age^2	5 SD	--
TwinsUK GWA	HDL	inverse normal	yes	age,age^2,analyser	4 SD	yes
TwinsUK WGS	HDL	inverse normal	--	age, age^2	5 SD	yes
ALSPAC WGS+GWA	LDL	inverse normal	yes	age, age^2	5 SD	--
TwinsUK GWA	LDL	inverse normal	yes	age,age^2,analyser	4 SD	yes
TwinsUK WGS	LDL	inverse normal	--	age, age^2	5 SD	yes
ALSPAC WGS+GWA	TC	inverse normal	yes	age, age^2	5 SD	--
TwinsUK GWA	TC	inverse normal	yes	age,age^2,analyser	4 SD	yes
TwinsUK WGS	TC	inverse normal	--	age, age^2	5 SD	yes
ALSPAC WGS+GWA	TG	inverse normal	yes	age, age^2	5 SD	--
TwinsUK GWA	TG	inverse normal	yes	age,age^2,analyser	4 SD	yes
TwinsUK WGS	TG	inverse normal	--	age, age^2	5 SD	yes

4.2.2 Single marker based discovery and follow-up

For single marker tests, I first fitted linear models on standardised trait residuals to test associations of allele dosages with 13,074,236 SNVs and 1,122,542 biallelic InDels (MAF≥0.1%) in the two WGS samples (TwinsUK and ALSPAC), using SNPTEST. Then I run the same analysis for 12 more cohorts with imputed data to identify novel variants across the allele frequency spectrum with a much larger sample size and increased power. Among the 12 additional cohorts, SNPTEST was used for population based samples while GEMMA was used for genetic isolates and cohorts with family structure. Meta-analyses were performed using GWAMA v2.1 (Magi and Morris 2010), assuming a fixed effect model adjusted genomic control to the summary statistics for both input and output data. Meta-analysis was first run for two WGS cohorts, to generate the WGS only based "2-way" results. Meta-analyses were then run for all 14 cohorts with genome-wide association results, leading to "14-way" results as an expanded discovery. Given the poor imputation quality and weak statistical power for rare variants, I chose to exclude the variants that did not pass a low allele frequency threshold (MAF<0.1%). For imputed cohorts, the variants with INFO score <0.4 were also excluded.

Given a large number of lipids loci already reported by previous GWAS with much larger sample size than this study, a rigorous loci selection was conducted to select putative novel loci that are statistically truly novel. The core of this loci selection process was a step-wise conditional analysis as described in chapter 2. Initially, GWAS Catalog and literature review were used to identify known variants. For those variants that survived the conditional tests, they were further checked against the full genome-wide results of the two largest GWAS (Teslovich et al. 2010, Global Lipids Genetics et al. 2013) (available at http://csg.sph.umich.edu/locuszoom/) to ensure their true novelty. As described in chapter 2, I excluded those variants that did not survive the step-wide conditional analyses or those having modest to high LD (r2>0.1) with known variants. For putative novel variants discovered from above, I conducted meta-analysis for replication cohorts and further performed a joint meta-analysis that calculated the statistics of all discovery and replication cohorts combined together.

4.2.3 Rare variant aggregation based discovery and follow-up

I first evaluated the associations of rare variants by considering genes as functional units of analysis. I applied two separate statistical models with different properties to rare variants (MAF<1%): SKAT and burden tests, both implemented in a unified software SKAT-O. As described in chapter 2, in *naïve* tests, all variants in exons, untranslated regions (UTRs) and essential splice sites were considered, and were given equal weight of being causal (50,214 windows for 35,709 genes, mean=35 variants, median=38 variants per window). In functional tests, only loss of function (LoF) and predicted functional variants were included (15,528 gene windows with ≥ 5 variants, mean=18, median=14 variants per gene). Finally, I run the locus-based analysis genome-wide in an agonistic fashion, by constructing ~1.8 million windows of 3 kb each, overlapping by half (median 35 SNVs/window, MAF<1%), assigning an equal weight to all variants.

For replication of locus based top hits, we used rareMetal (Feng et al. 2014) to reconstruct gene-level test statistics from single marker score statistics (Liu et al. 2014). The single maker score statistics were calculated with the Cochran-Mantel-Haenszel method. RareMetal works for meta-analysis of results from burden tests as well as SKAT tests. The windows with *P*<1E-5 in GW and *P* <1E-4 for EW based were taken forward for replication. Replications were conducted in three cohorts: GoT2D, FinRisk, InChianti. Finally, for those replicated loci, I explored a “drop-one” approach to determine whether the aggregation association was mainly driven by a single contributing variant. This worked by sequentially dropping one variant at a time and re-run SKAT-O for the same region with the same parameters. A variant was found to be contributing to the SKAT signal when dropping it causes a significant change of the SKAT-O P, usually from significant to non-significant. When more than one variant were found to be contributing, LD patterns were examined to evaluate the independence of those variants. In cases where a single variant with main effect could explain the association, usually the single marker was not sufficiently powered to detect an association in the same region.

4.2.4 Fine-mapping of known loci

For lipids, there were a total of 157 known loci reported. Many of those loci were significant in multiple lipids traits. I identified a total of 282 trait-specific regions for carrying out fine-mapping analysis to assess the probability of each variant being causal given other variants in the region. Within each signal I included SNPs in high LD (defined as all variants having $r^2 \geq 0.8$ with the most associated variants in the region), apart for *APOE* where an extended analysis interval was considered. As described in chapter 2, for each lipids trait I first created a list of fine-mapping regions based on HapMap estimates of recombination rates. I then analysed each region separately for each of the 14 participating cohort using Bayesian linear additive models, by accounting for covariates as in the general single point association analyses. At the end, the resulting BFs for each variant were multiplied to obtain a joint BF measure of association, with the assumption that each cohort is independent. These BFs were then used to calculate posterior probabilities, based on the assumption that there was exactly one causal SNP in each region. In addition, 95% and 99% credible sets were constructed in order to assess the uncertainty of the fine-mapping analysis.

The fine-mapped variants were further overlapped with four liver-essential TFBS data(Ballester et al. 2014). In brief, the genome-wide occupancy of four transcription factors (HNF4A, CEBPA, ONECUT1, and FOXA1) was determined in primary liver in five species (*Homo sapiens, Macaca mulatta, Canis familiaris, Mus musculus,* and *Rattus norvegicus*) using chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq). After mapping and peak calling, the regions of the genomes with the various combinations of the transcription factor binding events were analysed to determine the extent that binding events are shared across species and the characteristics of the shared and non-shared binding sites.

4.4 Results

4.4.1 Novel loci and novel variants from single marker analysis

WGS for low frequency and rare variants

The assessment of associations based on imputation or WES has been incomplete. I thus sought to investigate if additional low-frequency or rare variants with strong effects could be detected from the WGS dataset. I first tested association results using solely the WGS dataset in order to identify whether these variants existed. Associations were carried out in 13,074,236 SNVs and 1,122,542 biallelic InDels (MAF≥0.1%) using linear regression and data from the two WGS cohorts was meta-analysed.

Based on the meta-analysis of two UK10K WGS cohorts, there were a total of 267 trait-specific associations reaching the generally used genome-wide significance *P*<5.0E-08. All but two of these associations were previously reported, mapped to five known loci (*PCSK9, CELSR2, SID2, CETP, APOE)* (**Figure 4.2**). The first putative novel association is rs1505058, an intergenic variants on chromosome 5, for association with HDL (MAF=0.1%, beta=2.26, *P*=2.9E-09). The second putative novel association is rs185450930, an intronic variant within *SEMA3A* on chromosome 7, for association with TG (MAF=0.1%, beta=2.92, *P*=2.3E-08).

To look at suggestive associations, I used a less stringent threshold and discovered 117 more variants (a total of 384) having *P*<1E-6. Among all 384 variants, 90 variants have MAF between 0.1% and 5% and 22 are independent of known variants, i.e., either having no positive controls within 1Mb or surviving the conditional analysis and LD pruning with known variants within 1Mb. This list of 22 variants included the two variants with *P*<5E-08 described above, and are considered putative novel variants based on the two WGS cohorts. One de-novo genotyped cohort (Fenland) and three external WGS cohorts included in the expanded discovery (GoT2D, InChianti, FinRisk) were used as replication datasets for these 22 putative novel variants based on UK10K WGS, although not all these four cohorts have association results for these 22 variants. Their association summary statistics and replication results for these 22 variants were given in **Table 4.6**. The replication results for each of the four individual cohorts were given in **Table 4.7**. Based on the limited replication, only one variant within *LDLR* (rs72658867, EAF=1.2% (A), beta=-0.584) was replicated with a

consistent and comparable effect size (beta=-0.471, *P*=4.8E-12). Of note, the rare splice variant (rs138326449) in the *APOC3* gene was recently reported by us and others as associated with TG and coronary artery disease risk (Timpson et al. , Jorgensen et al. 2014, The TG and HDL Working Group of the Exome Sequencing Project 2014), therefore, it is viewed as a positive control instead of a novel locus.

Give the low power of single marker based replication for variants with low to rare frequency, the rare variants based tests (implemented in SKAT-O) were conducted for the 21 windows that include 21 variants except the variant on chromosome X (**Table 4.8).** Ten windows have SKAT-O *P* <2.3E-3 (i.e., 0.05/22), much more than expected. For all these 21 windows, the SKAT-O *P* is not much more significant than SKAT *P*, indicating that the signals are mainly driven by SKAT test instead of burden test. Indeed, for each of those five windows with SKAT *P*<1E-5, the SKAT signal was found to be driven by a single variant through a drop-one SKAT-O analysis.

Figure 4.2. Single point association results of lipids on WGS samples

X-axis is for chromosome and positions (build 37). Y-axis is for –log10(*P*). Variants passing threshold of 5E-08 and 1E-06 are shown in red and blue, respectively. For those passing threshold of 5E-08, known loci were marked in green text while putative novel loci were marked in red text.

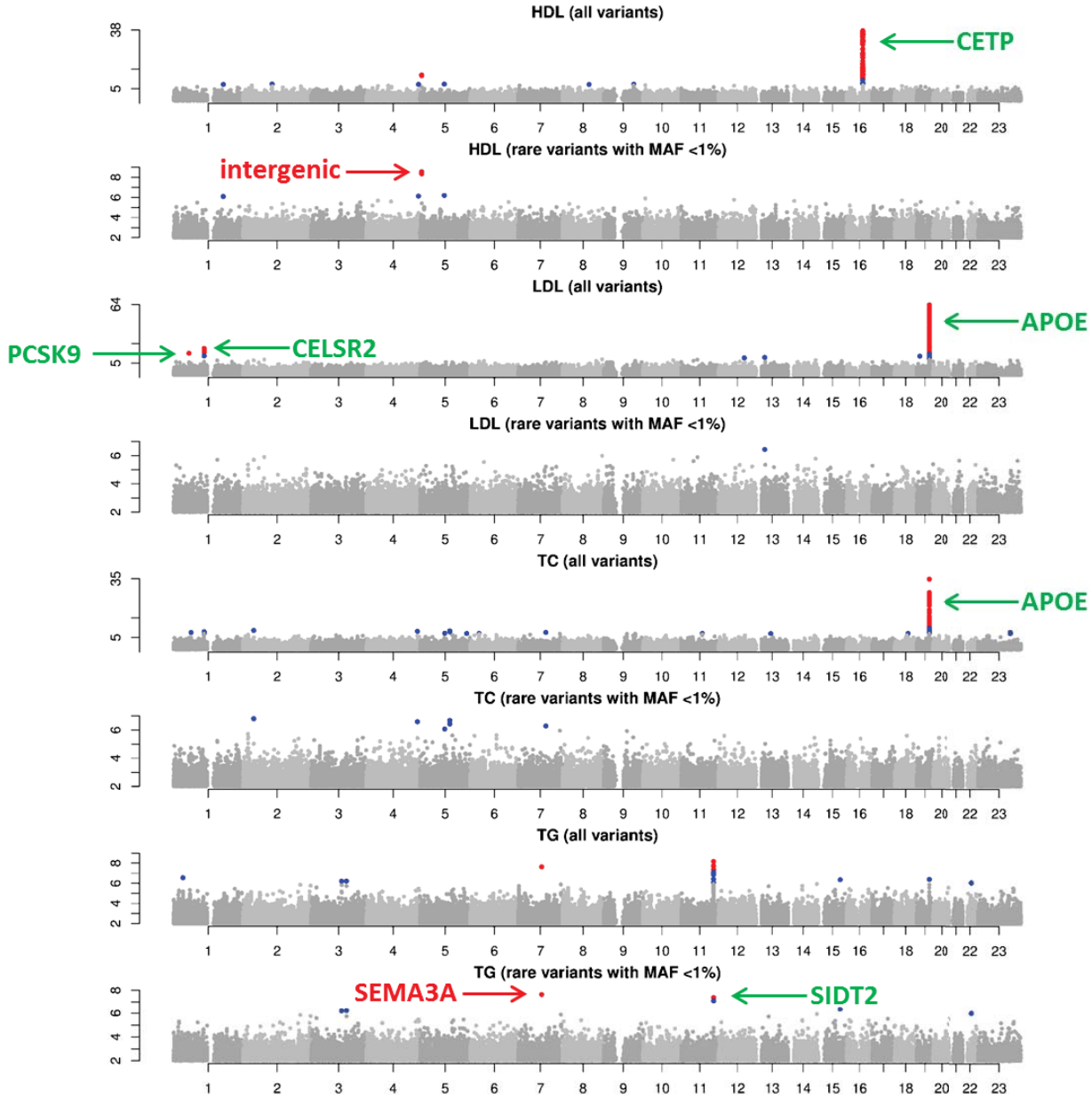


Table 4.6 Putative novel variants of low or rare frequency from UK10K WGS

WGS variants (P<1E-6) either have no positive controls within 1Mb or are independently significant from known variants.

Six have low frequency (MAF between 1-5%) and could be imputed with fair accuracy.

					UK10K WGS						Replication (WGS, de novo)			
Low frequency (6 variants)														
trai	**rsID**	**CHR**	**POS**	**Gene**	**EA**	**NEA**	**EAF**	**beta**	**SE**	**P**	**Beta**	**SE**	**P**	**N**
HD	rs72831743	2	103,690,744	Intergenic	A	C	0.012	-0.572	0.115	6.4E-07	-	0.05	7.8E-01	12161
TC	rs139029427	7	98,664,474	*SMURF1*	C	T	0.010	-0.643	0.128	5.1E-07	-	0.06	3.3E-01	12233
HD	rs150103869	8	94,349,833	*LINC00535*	A	C	0.010	0.632	0.128	8.1E-07	-	0.06	8.9E-03	12159
LD	rs77198522	12	91,641,075	Intergenic	T	C	0.030	-0.384	0.076	4.8E-07	0.004	0.03	9.1E-01	11948
LD	rs72658867	19	11,231,203	*LDLR*	A	G	0.012	-0.584	0.112	1.7E-07	-	0.06	4.8E-12	12215
TC	chrX:117293318	X	117,293,318	Intergenic	G	GGA	0.013	-0.869	0.172	4.6E-07	-	0.35	1.3E-02	614
Rare (16 variants)														
HD	rs184490209	1	178,071,554	*RASAL2*	A	G	0.001	1.958	0.396	8.0E-07	0.266	0.27	3.3E-01	2750
TC	chr2:37882057	2	37,882,057	*CDC42EP3*	GA	G	0.007	-0.752	0.143	1.6E-07	0.206	0.30	5.0E-01	2247
TC	rs143755400	2	37,883,627	*CDC42EP3*	A	G	0.007	-0.747	0.142	1.6E-07	-	0.09	8.0E-01	11618
TG	rs147039106	3	108,844,173	*MORC1*	C	T	0.007	0.799	0.160	6.2E-07	0.008	0.05	8.7E-01	12332
TG	chr3:126360068	3	126,360,068	*TXNRD3*	C	T	0.003	-1.138	0.228	6.0E-07	-	0.13	1.6E-01	12438
TC	chr4:182413170	4	182,413,170	*RP11-433O3.1*	A	G	0.001	-1.803	0.350	2.6E-07	0.001	0.17	9.9E-01	10818
HD	chr4:186058963	4	186,058,963	*SLC25A4*	G	T	0.004	-1.044	0.210	7.4E-07	-	0.11	2.4E-01	11758
HD	rs1505058	5	6,558,466	Intergenic	C	A	0.001	2.258	0.379	2.9E-09	-	0.20	7.3E-01	8776
HD	chr5:87396789	5	87,396,789	Intergenic	T	C	0.001	-1.887	0.378	6.5E-07	0.039	0.19	8.4E-01	10878
TC	rs183893710	5	88,977,348	Intergenic	G	C	0.005	-0.872	0.177	8.6E-07	-	0.07	6.3E-01	12467
TC	chr5:107200309	5	107,200,309	*FBXL17*	T	C	0.001	-2.571	0.495	2.1E-07	--	--	--	--
TG	rs185450930	7	83,755,035	*SEMA3A*	A	G	0.001	2.923	0.523	2.3E-08	--	--	--	--
TG	chr11:117053959	11	117,053,959	*SIDT2*	A	G	0.003	-1.359	0.248	4.2E-08	--	--	--	--
LD	chr13:31087680	13	31,087,680	*HMGB1*	C	T	0.002	-1.378	0.271	3.7E-07	-	0.19	9.9E-01	10755
TG	chr15:78513033	15	78,513,033	*ACSBG1*	T	C	0.001	-2.851	0.564	4.4E-07	--	--	--	--
TG	rs191808700	22	30,633,306	*LIF*	G	A	0.001	2.458	0.500	9.0E-07	--	--	--	--

* chr11:117053959 is in close proximity with the *APOC3* variants rs138326449 (chr11:116701354), with modest LD (r2 = 0.644), and is not independent significant based on conditional analysis.

Table 4.7 Replication results of WGS top hits

GoT2D, InChianti, and FinRisk used WGS, while Fenland used de-novo genotyping. For each set of results, the effect allele frequency (EAF), beta, standard deviation (SE), *P* value, and the total sample size were presented. Records with $P < 0.05$ are highlighted in red text.

		GoT2D					InChianti					FinRisk					Fenland				
trait	**rsID**	**EAF**	**Beta**	**SE**	**P**	**N**	**EAF**	**Beta**	**SE**	**P**	**N**	**EAF**	**beta**	**SE**	**P**	**N**	**EAF**	**beta**	**SE**	**P**	**N**
HDL	rs72831743	0.009	0.145	0.159	3.6E-01	2129	0.006	-0.197	0.380	6.0E-01	621	0.006	0.298	0.282	2.9E-01	856	0.012	-0.037	0.084	6.6E-01	5760
TC	rs139029427	0.010	-0.069	0.156	6.6E-01	2247	0.011	-0.377	0.280	1.8E-01	614	0.012	0.032	0.210	8.8E-01	856	0.010	-0.042	0.094	6.5E-01	5729
HDL	rs150103869	0.006	-0.396	0.214	6.3E-02	2129	0.010	-0.370	0.280	1.9E-01	621	0.003	-0.011	0.417	9.8E-01	856	0.012	-0.057	0.084	5.0E-01	5764
LDL	rs77198522	0.070	0.041	0.062	5.1E-01	2076	0.022	-0.259	0.197	1.9E-01	621	0.084	0.064	0.086	4.6E-01	817	0.035	0.021	0.052	6.9E-01	5653
LDL	rs72658867	0.006	-0.426	0.203	3.5E-02	2076	0.013	-0.579	0.252	2.2E-02	621	0.001	-0.024	0.697	9.7E-01	817	0.010	-0.473	0.076	4.9E-10	8701
TC	chrX:117293318	--	--	--	--	--	0.007	-0.882	0.354	1.3E-02	614	--	--	--	--	--	--	--	--	--	--
HDL	rs184490209	0.002	-0.126	0.536	8.1E-01	2129	0.008	0.404	0.318	2.0E-01	621	--	--	--	--	--	--	--	--	--	--
TC	chr2:37882057	0.003	0.206	0.303	5.0E-01	2247	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
TC	rs143755400	0.002	0.223	0.322	4.9E-01	2247	0.005	0.305	0.410	4.6E-01	614	--	--	--	--	--	0.006	-0.066	0.100	5.1E-01	8757
TG	rs147039106	0.033	-0.021	0.089	8.2E-01	2190	0.002	-0.013	0.708	9.9E-01	614	0.047	0.116	0.110	2.9E-01	856	0.011	-0.020	0.074	7.9E-01	8672
TG	chr3:126360068	0.002	-0.335	0.415	4.2E-01	2190	0.002	-0.466	0.578	4.2E-01	614	0.002	0.018	0.545	9.7E-01	856	0.003	-0.165	0.149	2.7E-01	8778
TC	chr4:182413170	0.001	0.759	0.470	1.1E-01	2247	--	--	--	--	--	--	--	--	--	--	0.002	-0.122	0.190	5.2E-01	8571
HDL	chr4:186058963	0.002	-0.093	0.406	8.2E-01	2129						0.001	-0.126	0.658	8.5E-01	856	0.004	-0.139	0.122	2.5E-01	8773
HDL	rs1505058	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	0.001	-0.072	0.209	7.3E-01	8776
HDL	chr5:87396789	0.001	-1.660	0.619	7.4E-03	2129	--	--	--	--	--	--	--	--	--	--	0.001	0.232	0.209	2.7E-01	8749
TC	rs183893710	0.003	0.184	0.283	5.2E-01	2247	0.008	0.006	0.319	9.8E-01	614	0.003	0.027	0.437	9.5E-01	856	0.009	-0.057	0.079	4.7E-01	8750
TC	chr5:107200309	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
TG	rs185450930	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
TG	chr11:117053959	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
LDL	chr13:31087680	0.001	-0.286	0.576	6.2E-01	2076	--	--	--	--	--	--	--	--	--	--	0.001	0.035	0.209	8.7E-01	8679
TG	chr15:78513033	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
TG	rs191808700	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

Table 4.8 SKAT results for single point test top hits

For each of the 22 top hits based on WGS single marker analysis, the selected SKAT-O window included the index variant. For genome-wide SKAT-O analysis with overlapping windows, when there are two windows include a variant, the one with the lower *P* value is listed. For SKAT-O test, *P* <2.3E-3 (i.e., 0.05/22) are shown in red.

trait	rsID	GW SKAT region	GW SKAT	GW SKATO	EW SKAT Region	EW SKAT	EW SKATO
HDL	rs72831743	chr2:103690501-103693500	2.66E-01	4.11E-01	--	--	--
TC	rs139029427	chr7:98664001-98667000	1.11E-06	2.70E-06	SMURF1.w3	8.25E-01	1
HDL	rs150103869	chr8:94348501-94351500	7.23E-01	5.06E-01	--	--	--
LDL	rs77198522	chr12:91641001-91644000	2.53E-01	4.02E-01	--	--	--
LDL	rs72658867	chr19:11230501-11233500	4.47E-02	3.04E-02	LDLR.w3	2.51E-01	3.75E-01
TC	chrX:117293316	--	--	--	--	--	--
HDL	rs184490209	chr1:178071001-178074000	5.32E-03	9.23E-03	RASAL2.w1	7.06E-01	8.66E-01
TC	chr2:37882057	chr2:37881001-37884000	4.01E-06	9.74E-06	CDC42EP3.w4	2.51E-02	7.73E-03
TC	rs143755400	chr2:37882501-37885500	6.53E-04	1.30E-03	CDC42EP3.w4	2.51E-02	7.73E-03
TG	rs147039106	chr3:108843001-108846000	9.09E-07	2.69E-06	--	--	--
TG	chr3:126360068	chr3:126360001-126363000	1.21E-06	2.90E-06	TXNRD3.w3	4.58E-01	6.46E-01
TC	chr4:182413170	chr4:182412001-182415000	1.18E-04	2.57E-04	--	--	--
HDL	chr4:186058963	chr4:186058501-186061500	2.12E-03	4.56E-03	--	--	--
HDL	rs1505058	chr5:6558001-6561000	3.77E-05	7.67E-05	--	--	--
HDL	chr5:87396789	chr5:87396001-87399000	7.51E-02	1.27E-01	--	--	--
TC	rs183893710	chr5:88977001-88980000	8.74E-07	2.46E-06	--	--	--
TC	chr5:107200309	chr5:107199001-107202000	6.36E-02	1.12E-01	FBXL17.w2	3.95E-01	1.38E-01
TG	rs185450930	chr7:83754001-83757000	5.13E-02	9.32E-02	SEMA3A.w3	2.01E-01	3.24E-01
TG	chr11:117053959	chr11:117052501-117055500	1.66E-03	3.58E-03	SIDT2.w2	6.64E-01	8.62E-01
LDL	chr13:31087680	chr13:31087501-31090500	2.77E-04	5.72E-04	HMGB1.w3	8.59E-01	2.04E-01
TG	chr15:78513033	chr15:78513001-78516000	8.70E-03	1.67E-02	ACSBG1.w4	8.30E-01	3.89E-01
TG	rs191808700	chr22:30633001-30636000	2.33E-03	4.25E-03	--	--	--

Meta-analysis for identifying novel variants of all allele spectrums

Given the enhanced imputation quality with the UK10K WGS reference panel as demonstrated in chapter 3, I included 12 more cohorts with imputed data for an expanded discovery, to increase power for discover variants across all allele frequency spectrum. As mentioned earlier in the methods section, variants with MAF <0.1% or imputation INFO <0.4 were not included. This effort yielded 5,306 variants with *P* <1E-07, 5,023 of which reached genome-wide significant threshold (*P* <5E-08) (**Figure 4.3**). I carried out step-wise conditional analysis to identify putative novel associations, as described in chapter 2 and the methods section of this chapter. All but four associations did not survive the novelty test, i.e, either association singles going away after conditional on known variants or in modest to high LD with known variants (r2>0.1). Two of these associations don't have positive controls within 1Mb. For the other two with position controls within 1Mb, chr16: 66926255 is conditioned on the four known variants (chr16:67708897, chr16:67902070, chr16:68013471, chr16:68024995) and its conditional *P* is 1.2E-07; rs72658867 is conditioned on four known variants (chr19:11195030, chr19:11202306, chr19:11224265, chr19:11227602) and its conditional *P* is 6.2E-10.

The four putative novel variants were taken forward in two rounds of replications that included genotypes from WGS, imputation and *de novo* genotyping. The association results including discovery and two rounds of replications for these four variants were reported in **Table 4.9**. The cohort specific results for these four variants were given in **Table 4.10**. The first variant is a common variant (MAF of 16.5%, rs57367316) on chromosome 2, for association with TG. It did not survive the first round of replication. Its best proxy rs4404266 (chr2:107712732, 12,462bp apart, r^2=0.63) has *P*=0.91 in the Global lipids study (Global Lipids Genetics et al. 2013). As shown in **Table 4.10**, this variant is only marginally significant in one replication cohort (FinRisk, *P*=0.046) but with an opposite effect size. Therefore, this variant is most likely to be false positive. The second variant chr16: 66926255 has an overall MAF of 0.003 and *P*=6.9E-08. However, this variant did not show evidence for replication either. Upon further inspection, the signal in the expanded discovery was mostly driven by a single cohort (HELIC-Manolis, beta (SE) = 1.491(0.236), *P*=9.7E-10), a genetic isolate of Greek origin, where its MAF is much higher (0.009) than the remaining cohorts. Failure to replicate this variant may be due to either a false positive in the Greek discovery cohort, or insufficient power in the non-isolate cohorts where the variant has low MAF. The third novel association detected was with variant rs72658867 within *LDLR*,

associated with LDL levels. This variant is annotated to be in a splice region, with MAF of 0.01 and meta-analysis *P*=1.49E-10. This variant is replicated in both rounds of replication, with *P*= 2.9E-11 and *P*=2.5E-02 respectively (**Table 4.9**). The combined meta-analysis result is: EAF=0.10 (A), beta (SE) =-0.326 (0.035), *P*=1.50E-20, N=51,757. This variant is independent of (LD r2<0.01) neighboring variants previously reported for association with CHD or lipids phenotypes (**Figure 4.4**). Previously, this variant was annotated as in intron 14 of *LDLR* under the name of "2140+5G>A", reported to have no effect on plasma cholesterol levels (Whittall et al. 2002) in a control sample with ~700 subjects. The fourth novel association, a common, X-linked variant associated with LDL (rs5985471, chrX:109703961, MAF=0.403, beta=0.050, *P*= 7.37E-08). This association is also replicated in two rounds of replication, with *P*= 6.6E-05 and *P*=2.8E-04 respectively. The combined meta-analysis result is: EAF=0.40 (T); beta (SE) = -0.042 (0.005), *P*= 2.02E-14, N=50,929. A sex-stratified analysis based on two cohorts with large number of males and females (ALSPAC and 1985BC) found that this association is significant in both males and females, therefore, not sex-specific. Within +/-500kb of rs5985471, there are two known associations, both of which are in high LD with rs5985471 (r2>0.8). The first one is rs5943057 (chrX:109939205), previously reported for association with CAD (*P* = 8.66E-07) in the C4D study (Coronary Artery Disease Genetics 2011). The minor allele for rs5985471 in this study is associated with a decreased level of LDL, i.e., protective. In the C4D study, the minor allele of rs5943057 is associated with a decreased level of CAD. The other known variant in strong LD is rs1573036 (chrX:109820068), previously reported for association with sex hormone-binding globulin levels (Coviello et al. 2012).

Figure 4.3 Association results of 14-way meta-analysis of the four main lipid traits

X-axis is for chromosome and positions (build 37). Y-axis is for –log10(*P*). Variants passing threshold of 5E-08 and 1E-07 are shown in red and blue, respectively. For those passing threshold of 5E-08, known loci were marked in green text while putative novel loci were marked in red text.

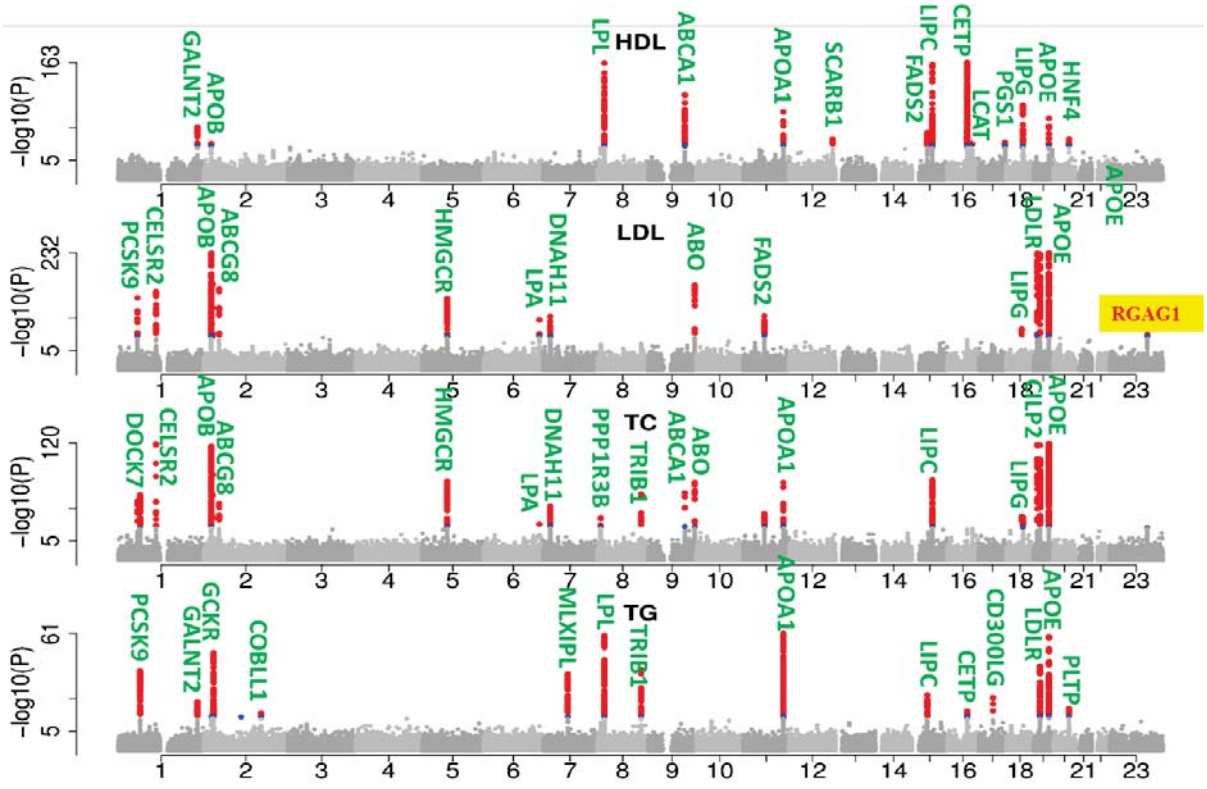


Table 4.9 Expanded discovery(14-way meta-analysis) top hits

This table shows the results of the expanded discovery meta-analysis (i.e., 14-way), followed by the two round of replications. For each set of results, the effect allele frequency (EAF), beta, standard deviation (SE), *P* value, and the total sample size were presented.

						14-way				
Trait	**rsID**	**CHR**	**POS**	**Gene**	**EA**	**EAF**	**Beta**	**SE**	**P**	**N**
TG	rs57367316	2	107,725,194	intergenic	A/G	0.165	0.074	0.014	6.9E-08	22,727
HDL	16:66926255	16	66,926,255	PDP2	T/A	0.003	-0.556	0.102	6.9E-08	22,385
LDL	rs72658867	19	11,231,203	LDLR	A/G	0.010	-0.342	0.053	1.5E-10	22,013
LDL	rs5985471	X	109,703,961	RGAG1	T/C	0.406	-0.047	0.009	7.4E-08	20,217

		Stage 1 replication					Stage 2 replication				
Trait	**rsID**	**EAF**	**Beta**	**SE**	**P**	**N**	**EAF**	**Beta**	**SE**	**P**	**N**
TG	rs57367316	0.156	-0.016	0.012	0.175	25599	--	--	--	--	--
HDL	16:66926255	0.002	-0.438	0.304	1.5E-01	4941	--	--	--	--	--
LDL	rs72658867	0.008	-0.390	0.059	2.9E-11	19099	0.010	-0.185	0.077	2.5E-02	10645
LDL	rs5985471	0.393	-0.034	0.008	6.6E-05	20066	0.406	-0.055	0.014	2.8E-04	10646

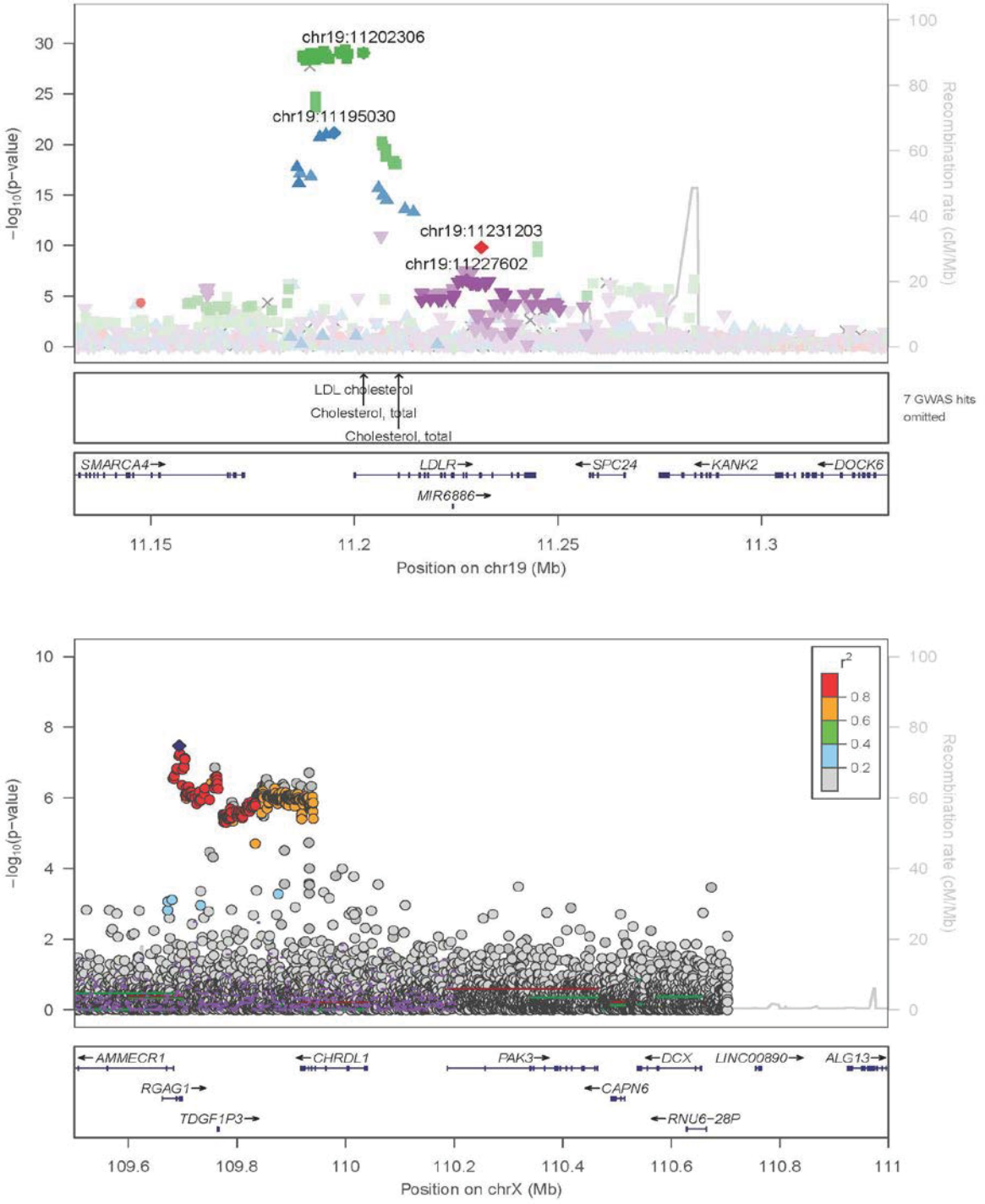
Table 4.10 Cohort specific results for four top variants based on 14-way meta-analysis

For each set of results, the effect allele frequency (EAF), beta, standard deviation (SE), *P* value, sample size (N), and imputation INFO score were presented. Records with *P* < 0.05 are highlighted in red text.

	rs57367316, TG						chr16: 66926255, HDL						rs72658867, LDL						rs5985471, LDL					
Cohort	**EAF**	**Beta**	**SE**	**P**	**N**	**Info**	**EAF**	**Beta**	**SE**	**P**	**N**	**Info**	**EAF**	**Beta**	**SE**	**P**	**N**	**Info**	**EAF**	**Beta**	**SE**	**P**	**N**	**Info**
ALSPAC WGS	0.154	0.113	0.052	3.0E-02	1497	0.99	0.002	-1.233	0.498	1.3E-02	1497	0.88	0.012	-0.713	0.157	5.8E-06	1495	0.99	0.404	-0.049	0.029	9.7E-02	1495	1.00
TwinsUK WGS	0.156	0.124	0.047	8.3E-03	1705	1.00	0.002	-0.838	0.372	2.5E-02	1713	0.88	0.012	-0.452	0.159	4.5E-03	1696	0.96	0.399	0.049	0.035	1.6E-01	1696	1.00
ALSPAC GWA	0.154	0.077	0.039	5.1E-02	2820	0.90	0.003	-0.191	0.328	5.6E-01	2820	0.63	0.009	-0.524	0.157	8.6E-04	2815	0.83	0.392	-0.077	0.023	6.4E-04	2815	1.00
TwinsUK GWA	0.154	0.002	0.048	9.6E-01	1882	0.92	0.003	-0.581	0.331	8.0E-02	1896	0.63	0.009	-0.245	0.189	1.9E-01	1870	0.83	0.406	-0.051	0.031	9.8E-02	1870	1.00
1958BC	0.154	0.081	0.028	3.8E-03	5485	0.91	0.003	0.080	0.214	7.1E-01	5493	0.65	0.011	-0.360	0.102	4.1E-04	5186	0.83	0.394	-0.087	0.016	1.5E-07	5186	1.00
INGI-CARL	0.143	0.077	0.121	5.3E-01	412	0.78	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--
INGI-FVG	0.168	0.156	0.058	7.6E-03	1377	0.84	--	--	--	--	--	--	0.005	-0.049	0.343	8.9E-01	1377	0.61	0.452	-0.041	0.032	2.2E-01	1377	0.96
INGI-VB	0.178	0.006	0.053	9.1E-01	1776	0.78	0.004	-0.106	0.458	8.2E-01	1776	0.44	0.014	-0.238	0.193	2.2E-01	1775	0.59	0.379	-0.033	0.034	3.3E-01	1775	0.80
HELIC-A	0.238	0.018	0.052	7.4E-01	1245	0.86	0.009	-1.491	0.236	9.7E-10	1247	0.89	0.017	-0.206	0.205	3.2E-01	1253	0.60	0.384	0.007	0.035	8.5E-01	1253	1.00
HELIC-P	0.151	-0.033	0.071	6.4E-01	964	0.88	0.002	-0.953	0.559	8.9E-02	976	0.91	0.002	-0.904	0.512	7.9E-02	976	0.78	0.393	0.082	0.042	5.3E-02	976	0.99
INCIPE-1	0.191	0.092	0.071	2.0E-01	653	0.88	0.003	-0.358	0.634	5.7E-01	653	0.63	0.015	0.056	0.290	8.5E-01	653	0.69	0.462	-0.079	0.048	9.7E-02	653	0.99
INCIPE-2	0.175	0.010	0.056	8.6E-01	1382	0.80	0.004	-0.413	0.312	1.9E-01	1382	0.89	0.012	0.284	0.213	1.8E-01	1380	0.69	0.442	-0.018	0.031	5.6E-01	1380	0.99
LURIC	0.177	0.167	0.063	8.2E-03	983	0.85	0.002	-0.101	0.634	8.7E-01	983	0.52	0.009	-0.255	0.289	3.8E-01	983	0.71	0.409	-0.038	0.037	3.0E-01	960	1.00
Teenage	0.171	0.236	0.087	6.8E-03	551	0.85	0.004	0.231	0.616	7.1E-01	557	0.56	0.007	-0.304	0.520	5.6E-01	557	0.50	0.413	-0.065	0.051	2.1E-01	557	0.99
Fenland	0.162	-0.017	0.021	4.2E-01	8660	1.00	--	--	--	--	--	--	0.010	-0.473	0.076	4.9E-10	8701	1.00	0.392	-0.041	0.013	1.5E-03	8590	1.00
FinRisk	0.130	-0.133	0.067	4.6E-02	856	1.00	--	--	--	--	--	--	0.001	-0.024	0.697	9.7E-01	817	1.00	--	--	--	--	--	--
GoT2D	0.151	-0.007	0.042	8.7E-01	2190	--	--	--	--	--	--	--	0.006	-0.426	0.203	3.5E-02	2076	--	--	--	--	--	--	--
InChianti	0.204	-0.035	0.074	6.4E-01	614	1.00	--	--	--	--	--	--	0.013	-0.579	0.252	2.2E-02	621	1.00	0.383	-0.129	0.048	7.0E-03	621	1.00
Lolipop EW610	0.167	-0.064	0.065	3.2E-01	927	0.90	--	--	--	--	--	--	0.016	-0.166	0.200	4.1E-01	905	0.91	--	--	--	--	--	--
Lolipop EWA	0.148	-0.065	0.085	4.4E-01	582	0.89	--	--	--	--	--	--	0.004	-0.360	0.598	5.5E-01	566	0.66	--	--	--	--	--	--
Lolipop EWP	0.160	-0.024	0.084	7.7E-01	642	0.83	--	--	--	--	--	--	0.013	0.125	0.267	6.4E-01	610	0.89	--	--	--	--	--	--
RS-1	0.157	-0.021	0.037	5.6E-01	3108	0.90	0.001	0.102	0.582	8.6E-01	3081	0.48	0.005	-0.578	0.226	1.1E-02	2981	0.72	0.396	-0.010	0.022	6.4E-01	2981	1.00
RS-2	0.158	0.026	0.048	5.9E-01	1847	0.89	0.003	-0.640	0.356	7.2E-02	1861	0.69	0.006	0.199	0.269	4.6E-01	1823	0.59	0.381	0.009	0.028	7.5E-01	1823	0.99
UCLEB BRHS	0.149	0.025	0.038	5.1E-01	2785	1.00	--	--	--	--	--	--	--	--	--	--	--	--	0.399	-0.049	0.020	1.3E-02	2742	1.00
UCLEB BWHS	0.146	-0.018	0.034	6.0E-01	3388	1.00	--	--	--	--	--	--	--	--	--	--	--	--	0.397	-0.020	0.025	4.3E-01	3309	1.00
WHI garnet	0.163	0.048	0.032	1.4E-01	3755	0.93	0.003	0.078	0.240	7.5E-01	3781	0.75	0.011	-0.136	0.128	2.9E-01	3726	0.77	0.400	-0.066	0.023	4.5E-03	3726	0.99
WHI hipfx	0.149	-0.039	0.075	6.0E-01	799	0.89	--	--	--	--	--	--	0.012	0.177	0.290	5.4E-01	639	0.80	0.401	-0.046	0.057	4.2E-01	639	0.99
WHI mopmap	0.164	0.120	0.071	9.2E-02	768	0.94	--	--	--	--	--	--	0.006	-0.100	0.410	8.1E-01	745	0.63	0.380	-0.023	0.053	6.7E-01	745	0.99
WHI whims	0.163	0.038	0.027	1.6E-01	5546	0.92	0.003	0.152	0.223	5.0E-01	5580	0.71	0.011	-0.268	0.104	9.8E-03	5537	0.79	0.415	-0.052	0.019	6.7E-03	5537	1.00

Figure 4.4 Regional plots of two loci with replicated novel associations

The top plot is for association with LDL in the *LDLR* region. The bottom plot is for the novel locus on chromosome X. Both are for association with LDL and *P* values are based on the 14-way meta-analysis. For the LDLR locus, the novel variant is shown in red text, while the SNPs tagged by previously reported variants are known in other colors. For the chromosome X region, there were no previously reported variants.



4.3.2 Fine mapping of known and novel loci

To fine-map lipid-associated regions, I implemented the method of Maller et al. (Maller et al. 2012), as described in chapter 2 and the Methods section above. For 41 out of a total of 282 regions examined, there are sufficient resolution to limit the number of possible causal variants to a small informative set (log10BF>5 and # of variants <20). The distribution of the number of causal variants within these 41 loci is shown in **Figure 4.5**.

To further characterize the predicted functional consequence of the FM variants, the fine-mapping regions were overlapped with four liver-essential TFBS data (Ballester et al. 2014). Ten variants that are in the 95% credible set of these 41 fine-mapped regions also overlapped with a TFBS (**Table 4.11**). These 10 variants should be considered as good candidates for further functional and causality studies. By further overlapping these 10 variants with liver expression of quantitative trait loci (eQTL) data on GTEx (http://www.gtexportal.org/), I identified two variants have significant eQTL signal (eQTL *P*<5E-08). The first one is rs12740374 in *SORT1*, which was previously identified as causal (Musunuru K, et. al. 2010, Nature). The second one is rs10438978 (A/G alleles) close to *LIPG*, with eQTL *P*=1.96E-10 and motif change of CTCF_disc3. The discovery of a causal variant in *SORT1* locus demonstrated the proof-of-concept for this approach.

Figure 4.1 Number of putative causal variants within fine-mapped loci

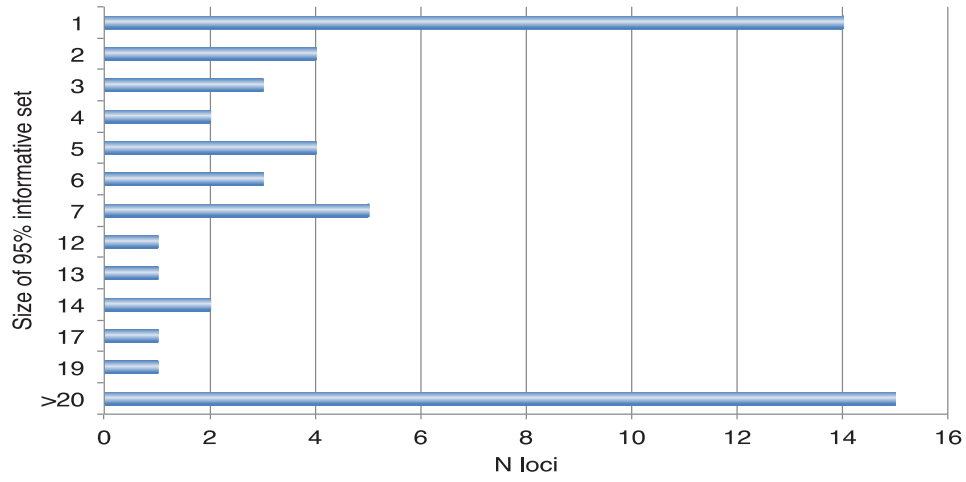


Table 4.11 Predictive causal variants based on fine mapping

This table lists 10 putative causal variants within the 41 fine-mapped regions that overlap with a TFBS.

BF: bayes factor, PP: posterior probability

Trait	SNP	Chr	Pos	log10BF	PP	gene
LDL	rs12740374	1	109,817,590	24.33	0.15	CELSR2 3_prime
LDL	rs4245791	2	44,074,431	7.41	0.30	ABCG8 intron
HDL	rs4100654	9	107,669,241	9.13	0.71	ABCA1 intron
HDL	rs1077834	15	58,723,479	25.83	0.10	LIPC:upstream
HDL	rs1800588	15	58,723,675	26.26	0.25	LIPC:upstream
HDL	rs2070895	15	58,723,939	26.36	0.33	LIPC:upstream
HDL	rs10438978	18	47,158,186	10.72	0.18	LIPG
HDL	rs9304381	18	47,158,234	10.93	0.29	LIPG
LDL	rs58542926	19	19,379,549	25.24	0.15	TM6SF2 missense
TG	rs483082	19	45,416,178	15.76	0.26	APOE upstream

			WGS				14-way			
Trait	SNP	EA	EAF	beta	SE	P	EAF	beta	SE	P
LDL	rs12740374	T	0.211	-0.178	0.030	3.2E-09	0.218	-0.139	0.012	2.9E-32
LDL	rs4245791	T	0.658	-0.033	0.025	1.9E-01	0.663	-0.080	0.010	4.3E-15
HDL	rs4100654	C	0.098	-0.205	0.042	1.3E-06	0.096	-0.128	0.017	1.4E-14
HDL	rs1077834	C	0.204	0.136	0.031	1.5E-05	0.215	0.146	0.012	7.6E-35
HDL	rs1800588	T	0.202	0.137	0.031	1.3E-05	0.211	0.148	0.012	1.9E-35
HDL	rs2070895	A	0.204	0.134	0.031	2.0E-05	0.215	0.147	0.012	2.4E-35
HDL	rs10438978	C	0.819	0.048	0.033	1.5E-01	0.835	0.098	0.013	5.2E-14
HDL	rs9304381	T	0.819	0.048	0.033	1.5E-01	0.836	0.099	0.013	3.5E-14
LDL	rs58542926	T	0.073	-0.140	0.048	3.4E-03	0.074	-0.190	0.018	6.6E-25
TG	rs483082	T	0.242	0.126	0.029	1.6E-05	0.213	0.130	0.012	2.7E-27

4.3.3 Novel loci based on rare variants aggregation test

The above are for single marker base tests, which has limited power to detect associations for low frequency and rare variants given the current number of samples with WGS. Here I show association results based on rare variants aggregation tests. As stated in the Methods section, three types of SKAT-O analyses were run: genome-wide sliding window, exome-wide gene based, and exome-wide with only functional variants. Overall, the statistics of these tests follow the expected distribution assuming a NULL association, where the lambda is close to 1 and the tail does not significantly deviate from the expected (**Figure 4.6**). Of note, the QQ plots are not based on SKAT-O *P* value because that is a statistic after comparing two tests (SKAT and burden). The genome-wide significance thresholds are predefined as 6.8E-08, 1.2E-06, 1E-05 respectively for genome-wide, exome-wide, and functional variants based SKAT-O. There are four loci surpassing these significance thresholds (**Figure 4.7)**. These four windows and another 103 windows with *P*<1E-5 in GW and *P* <1E-4 for EW based were taken forward for replication in three cohorts (GoT2D, FinRisk, InChianti). At the most liberal threshold of replication P<0.05, 19 windows have evidence for replication by either SKAT or burden statistics. However, only the *APOC3* region has an adequate replication (*P* < 0.0005) that survived the multiple tests on 107 windows, with combined SKAT *P*=1.36E-08. The only other window with a combined SKAT *P*<5E-08 is chr4:110946001-110949000 for TG (SKAT *P* =2.23E-08). As shown in **Figure 4.8**, the peak of the SKAT signal lies between the *EGF* and *ELOVL6* gene. The full name for *ELOVL6* is ELOVL Fatty Acid Elongase 6, whose function is to catalyze the synthesis of saturated and monounsaturated fatty acids. It is certainly a plausible gene for impacting circulating lipids levels. The best single marker variant within this region is rs184358074, AF=0.6%, *P*=5.3E-04, which would be considered non-significant based on the pre-defined threshold. Drop-one anlaysis confirmed that this signal is not driven by any single variants that were included in the SKAT-O analysis.

Figure 4.6 QQ plots of SKAT tests for lipids

The four columns are for HDL LDL TC TG; each pairs of rows are for genome-wide, exome-wide, and functional variants.

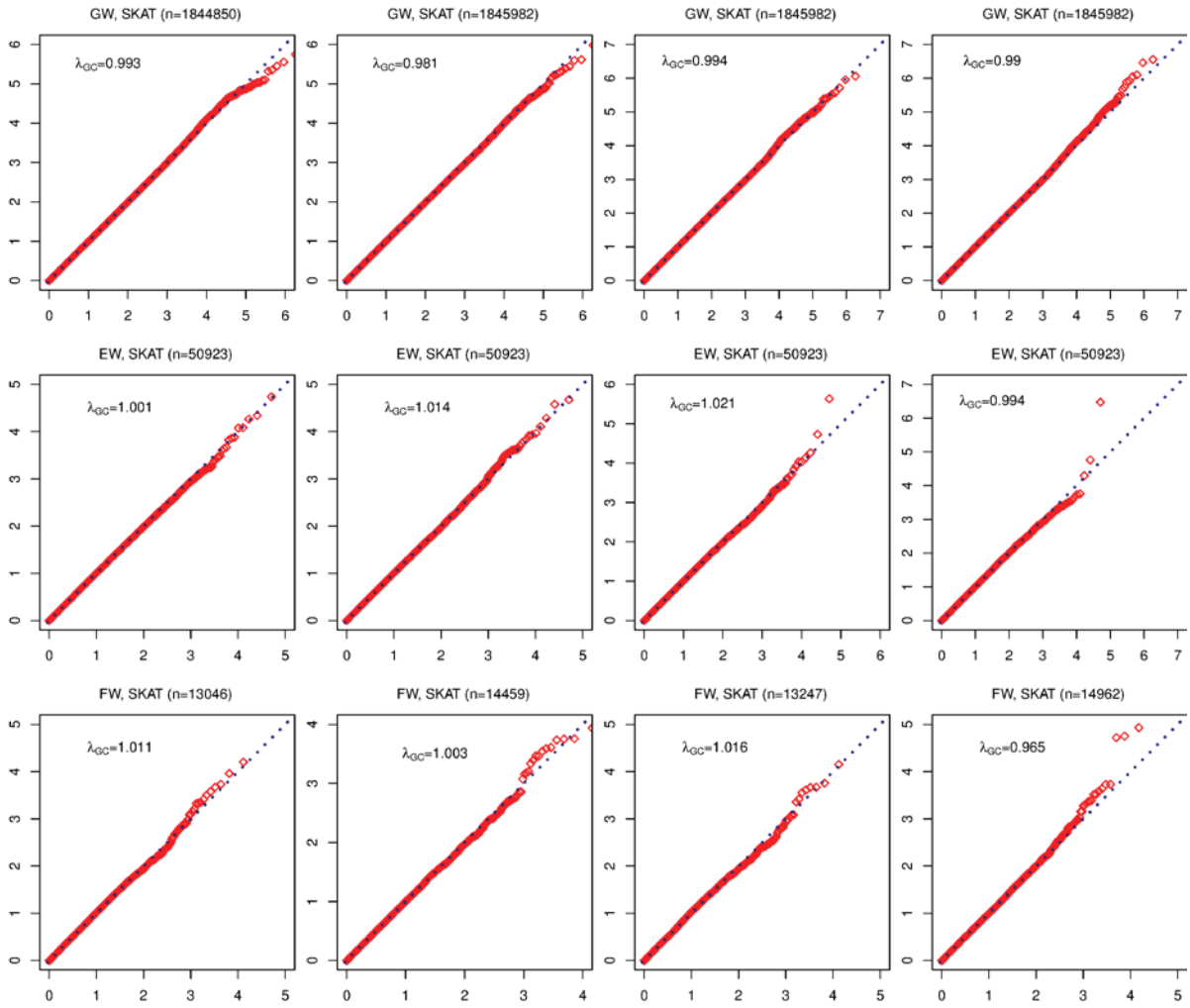


Figure 4.7 Rare variants aggregation test results for lipids

The genome-wide significant signals are shown in red, with threshold of $P < 6.8E-08$, 1.2E-06, 1E-05 respectively for genome-wide, exome-wide, and functional variants based SKAT-O. Suggestive signals are shown in blue, with threshold of $P < 1E-05$, 1E-04, 1E-04 respectively for genome-wide, exome-wide, and functional variants based SKAT-O.

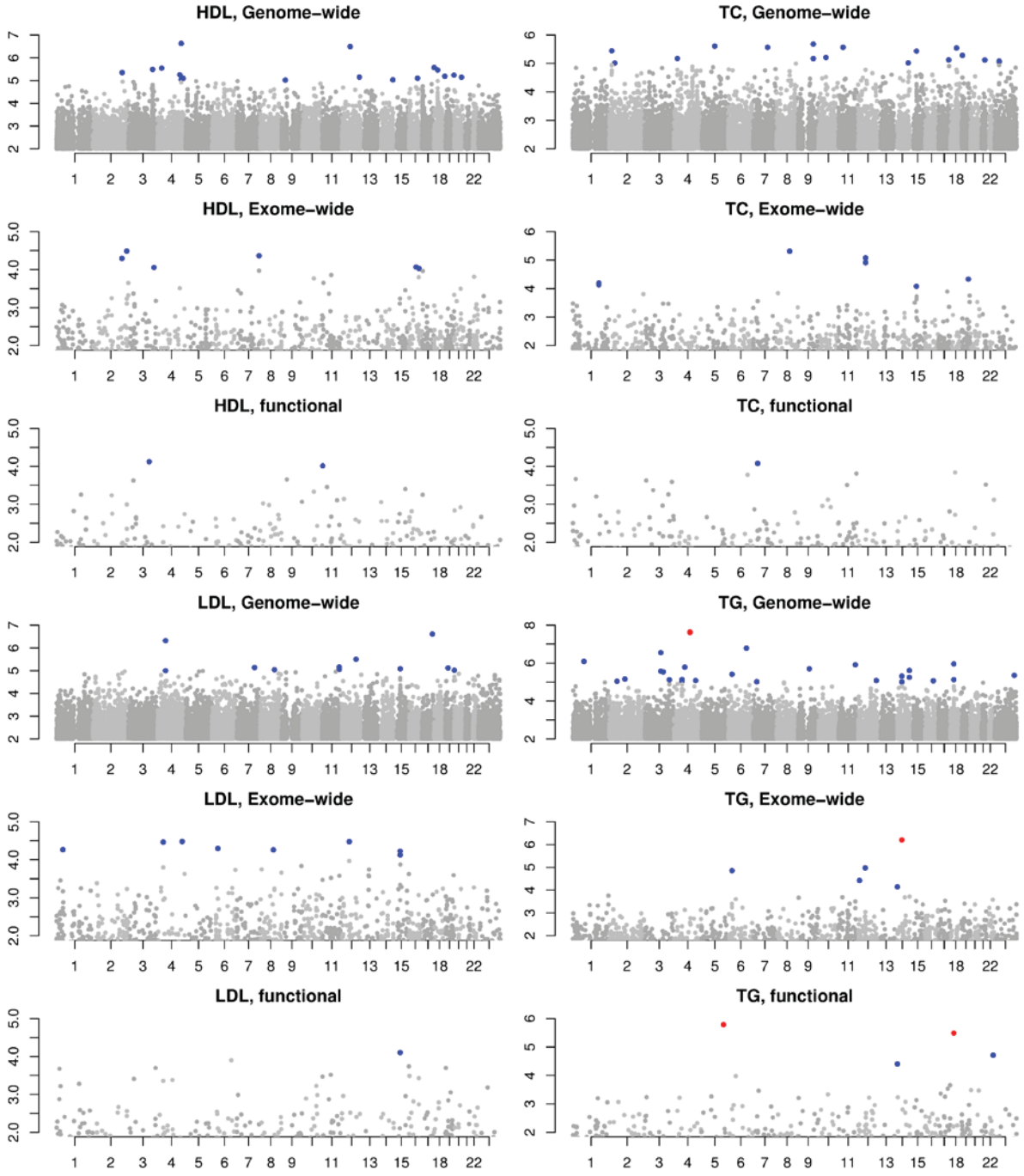
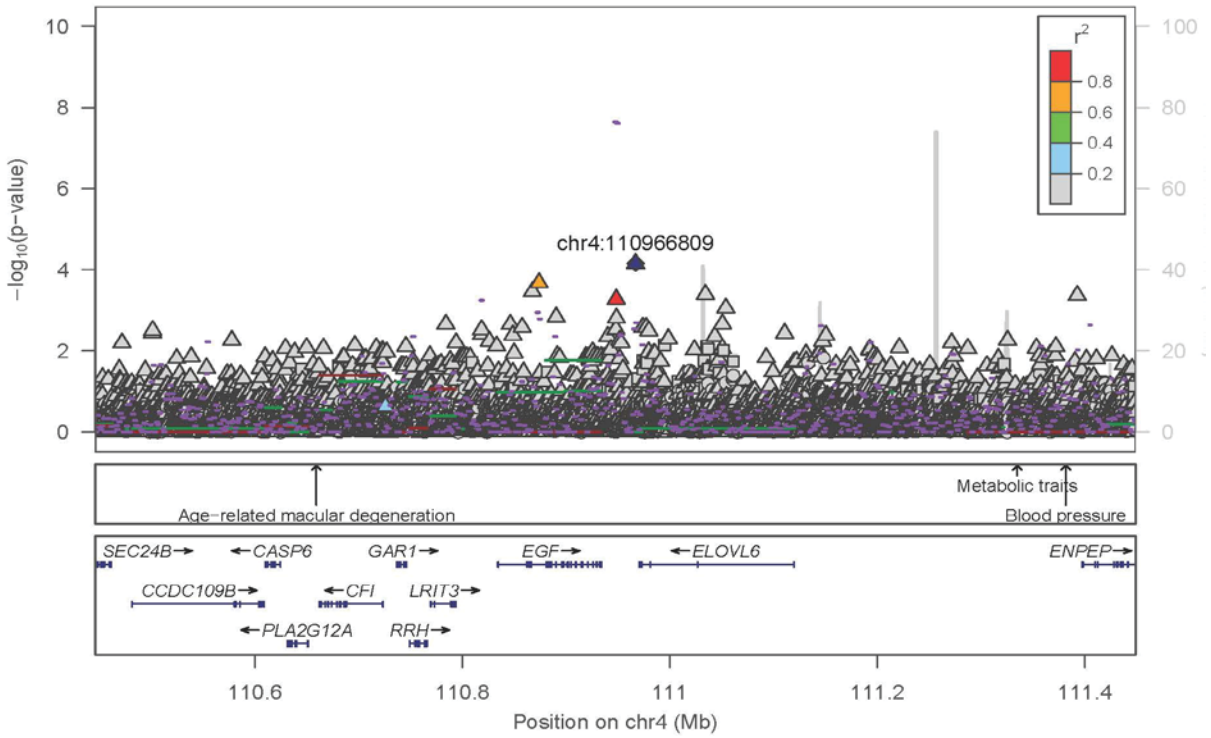


Figure 4.8 Regional plot of SKAT-O locus *EGF-ELOVL6*

The UK10K WGS single marker results are shown in points, where circle, cube, and triangle are used for common, low frequency, and rare variants. The UK10K SKAT-O results are shown in horizontal lines, where purple, green, brown are used for genome-wide SKAT, exome-wide SKAT, and functional variants exome-SKAT.



4.4 Conclusion & Discussion

4.4.1 Summary of main findings

This is by far the largest genome-wide scan on identifying genetic variants of plasma lipids using WGS data. Although the total sample size is much smaller than that in Global lipids study, the sequencing generated data and WGS imputed data provide an unprecedented opportunity to uncover rare and causal variants and their associations, as demonstrated by the example of *APOC3*, *LDLR*, and the novel locus on chromosome X. Although the clinical relevance of the *LDLR* variant (rs72658867) is yet to be confirmed, the *APOC3* variant (rs138326449, IVS2+1G→A) was already reported to be strongly associated with reduced CHD risk. In two studies that established the causality of rare variants within *APOC3*, one used high-depth WES (Tg et al. 2014) and the other used targeted re-sequencing (Jorgensen et al. 2014). The UK10K data is the first low-coverage WGS data that discovered this variant through both single marker based test and rare variant aggregation test.

Recently, there was an exome-array based study reported four rare variants for association with HDL or TG with large effect sizes (Peloso et al. 2014). But only one variant, rs186808413 within *PAFAH1B2*, is marginally significant in the UK10K WGS based results, P=0.018. This variant is in low LD with the reported splice variant within *APOC3* (rs138326449), r2=0.18, 341kb apart. Another WES based study reported an association between LDL and the burden of rare and low-frequency variants in *PNPLA5* (Lange et al. 2014). However, this result is not replicated in our exome-wide based SKAT-O test (P >0.05).

4.4.2 Interpretation of results

A wealth of novel lipid loci have been identified through a variety of approaches focused on common and low-frequency variation and collaborative meta-analyses in multi-ethnic populations. Despite progress in identification of loci, the task of determining causal variants remains challenging. This work will undoubtedly be enhanced by improved understanding of regulatory DNA at a genome-wide level as well as new methodologies for interrogating the relationships between noncoding SNPs and regulatory regions. Equally

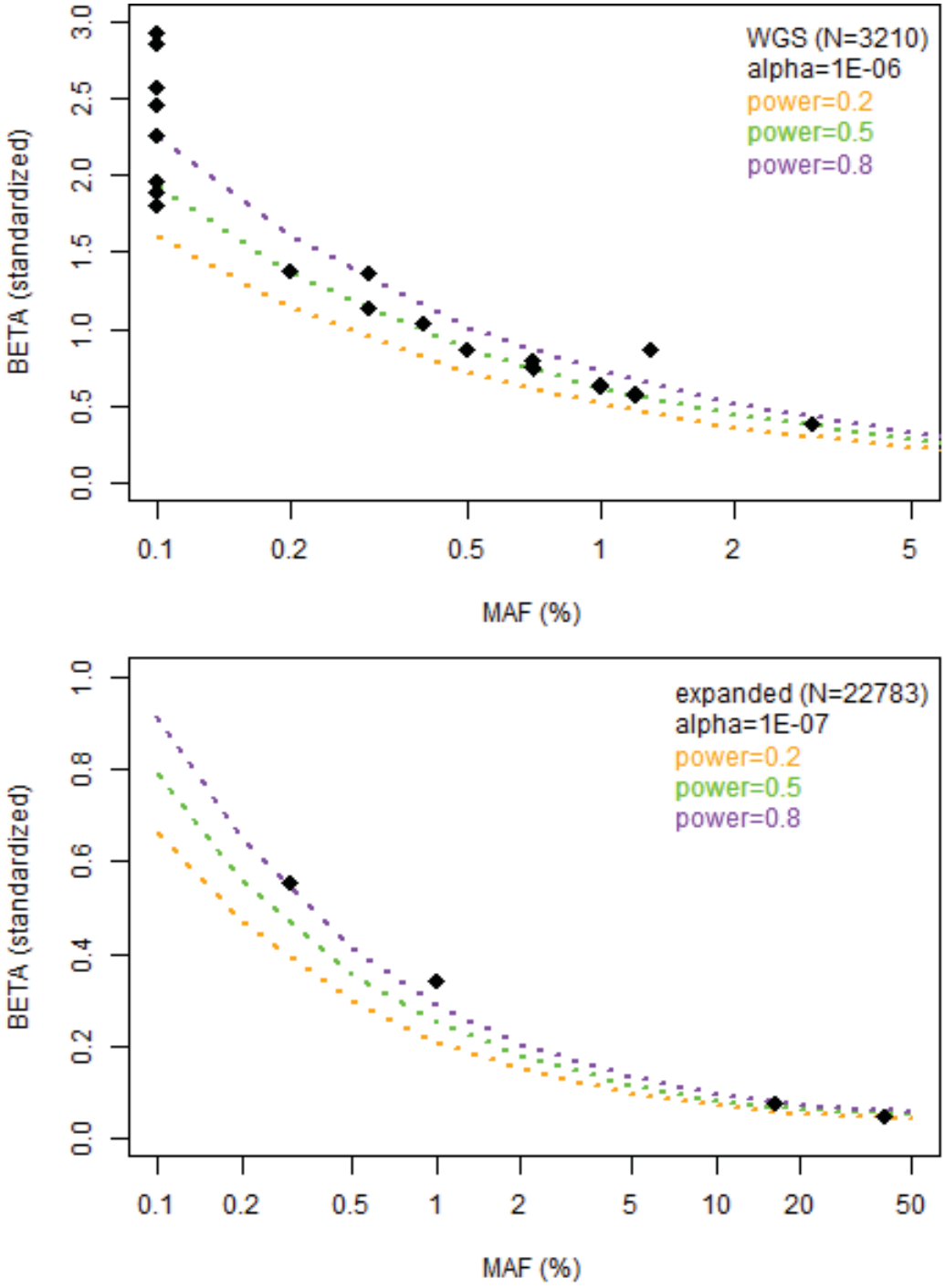
challenging is the identification of causal genes at novel loci. Additional insights will be gleaned from focusing on low-frequency and rare coding variation at candidate loci in large populations.

The single marker association testing of four lipids follows closely the expected relationship between EAF and effect size (beta) as dictated by study power (Park et al. 2011), as shown in **Figure 4.9**. Low frequency alleles of very high penetrance (beta ~1 SD) are unlikely to exist within this allelic space in the general European-ancestry population. Examples such as the rare *APOC3* or *LDLR* variants, with sufficient individual effect sizes to be clinically informative, are beginning to emerge (Flannick et al. 2012), but these findings are likely to be exceptions rather than a paradigm. Greater power than the current study will be required for capturing a greater proportion of missing heritability through either increases in sample size or genotyping accuracy and SNV density. The assessment of rare variants using a range of single-marker, exome-based and genome-based tests suggests that naïve and even functional scans were broadly underpowered to detect associations with high certainty, requiring extensive follow-up replication studies (Zuk et al. 2014). Deep sequencing will be needed to discover and fully assess this frequency range, which contains highly penetrant, potentially clinically important variants not accessible through imputation.

Finally, based on **Table 4.1** and **Figure 4.1**, there are five genes that were discovered by both linkage analysis and GWAS: ABCA1, ABCG5, ABCG8, LDLRAP1, PCSK9. However, none of these gene regions is significant based on exome-wide SKAT-O analyses (P >0.05). In single variant based analysis, there are no variants with MAF <5% in these genes have a P-value that surpassed the pre-defined threshold of 1.0E-07. This could be very likely due to the limited power of the current study to detect association signals for low frequency and rare variants.

Figure 4.9 Statistical power and novel variants from single marker analysis

The top and bottom plots are for WGS samples and expanded discovery samples respectively. Y-axis is a variant's effect, expressed in standard deviation units. X-axis is MAF of effect alleles. Colored lines indicate 20%, 50%, and 80% power. Alpha is set at *P*<1E-06 for WGS and *P*<1E-07 for expanded discovery respectively. The 16 putative novel WGS variants are shown in the top power plot for WGS, and the four putative novel variants from expanded discovery are shown in the bottom power plot for expanded discovery.



4.4.3 Future direction

Presently, there are still challenges in applying statistical methods to rare variants based analysis, especially when the sample size is small. During phenotype harmonization, samples with values that are more than three standard deviation of the mean are excluded. This is justifiable given that the focus of this study is on quantitative traits in healthy populations. However, this approach might have prevented the identification of a small group of individuals who carry rare variants with large effects that are linked with Mendelian conditions, as that reported by the Morrison study (Morrison et al. 2013).

As the field of lipid genetics moves beyond GWAS to focusing on identification of causal variants, causal loci, and biological mechanisms underlying novel genes, the study of low frequency and rare variants with large sample sizes and integrating genomic data with functional data would be critical. For common noncoding variants that are within (or in high LD with) defined promoter or known regulatory regions of nearby genes, one could assess the underlying effects of them through gene reporter assays, binding affinity for specific transcription factors, and related functional approaches. Such efforts have been done for a limited number of lipid-associated variants, such as for the causal role of *SORT1* to LDL and CVD risk (Musunuru et al. 2010), where the minor allele of the causal variant within a cis-regulatory region was found to create a de novo C/EBP TFBS that caused C/EBP-dependent upregulation of expression of the nearby genes. Another approach is to overlay GWAS variants with regions with chromatin marks or regions of DNase I hypersensitivity, suggesting open chromatin and active transcription (Maurano et al. 2012). Finally, in vivo overexpression or knockdown of candidate genes at a locus in animal models would provide most convincing causal evidence. The large lipids GWAS in 2010 reported such work for three candidate genes influencing HDL: *GALNT2*, *PPP1R3B* and *TTC39B* (Teslovich et al. 2010).

5 Full Blood Counts

5.1 An introduction to full blood counts

5.1.1 Biology and physiology of FBC

Blood cells play major roles for a variety of essential physiological functions. Among their many functions, red blood cells (RBC) transport oxygen, white blood cells (WBC) are engaged with some of the immune and inflammatory responses, and platelets (PLT) form blood clots to prevent excessive bleeding. RBC, WBC, PLT are also called erythrocytes, leukocytes, thrombocytes, respectively. All these blood cells, also called hematocytes, are produced by hematopoiesis (Orkin and Zon 2008). Circulation levels of blood cells are commonly measured in clinical visits and regular physical check-ups, because they are easily measured and an abnormal number or size or feature of the them are indicators of multiple human diseases. Very low level of RBC and hemoglobin (HGB) is the direct causes of anemia; rapid production of abnormal white blood cells causes leukemia; low level of PLT counts causes thrombocytopenia. There are a few other commonly measured RBC related traits, including haemoglobin (HGB), mean cell haemoglobin (MCH), mean cell haemoglobin concentration (MCHC), mean cell volume (MCV), packed cell volume (PCV). Although these traits are highly correlated, assaying multiple traits simultaneously could provide refined insights into path-physiological process. For example, a decrease of both MCV and MCH suggests a problem in hemoglobin production caused by iron deficiency or ineffective synthesis of globin polypeptides. WBCs are classified into five subtypes based on their morphology and functions, including neutrophils, basophils, eosinophils, lymphocytes and monocytes. Determination of platelet size, usually via quantification of mean platelet volume (MPV), is a simple and easy method of accurately assessing platelet function. In some genetic studies, both MPV and PLT were used as phenotypes.

Although environmental factors especially poor nutrition and infections casuse abnormal blood cells, genetics play a major role for both severe blood disorders and normal

variation of blood cell levels in healthy individuals. For example, mutations in *G6PD* cause chronic hemolytic anemia, and mutations in oncogenes or tumor suppressor genes cause leukemia.

5.1.2 FBC traits as risk factors for CVD

FBC is a commonly used screening for indicators of health and disease.

RBC traits and risk for CVD

RBC is directly related to cardiovascular performance. The use of exogenous EPO has been reported in athletes to boost performance. Anemia, defined as HGB <11 g/dL in women or <13 g/dL in men, is the most common form that ranges from mild fatigue to heart failure (Greenburg 1996). The World Health Organization estimates that anemia affects 1.62 billion people in the world, as of the end of 2013. The main causes of anemia are poor nutrition and iron deficiency, infections (e.g., malaria) and RBC diseases including hemoglobinopathies. Since anemia is mostly frequent in Africa and South-East Asia, it is critical to search for genetic associations with hemoglobin levels in these populations.

WBC and risk for CVD

WBC count is used as a clinical marker of inflammation status. Patients with elevated WBC have been shown to be in a higher risk of developing acute MI and acute coronary and vascular events. Measuring WBC and its sub-phenotypes could be used for a better way of risk stratification of patients admitted with acute vascular events (Hoffman et al. 2004). High WBC has been associated with an increased risk of CVD (Danesh et al. 1998), cancer mortality (Shankar et al. 2006) and all-cause mortality (Ruggiero et al. 2007). Elevated WBC is also associated with disease risk factors including increasing age, high BP, cigarette smoking, adiposity and increasing plasma inflammatory markers (Nieto et al. 1992). The association of WBC with cardiovascular risk factors may either represent manifestation of subclinical disease or suggest that WBC is part of the causal chain leading to atherosclerosis. More recently, it was reported that WBC count is also a predictor of fatal and nonfatal ischemic vascular disease independent of other CHD risk factors (Campbell et al. 2012).

PLT and CVD

Coronary atherosclerosis is a highly complex chronic inflammatory disease that may convert into an acute clinical event, especially in acute coronary syndromes (ACS) which occur secondary to atherosclerotic plaque rupture and subsequent vessel ischaemia (Tiong and Brieger 2005). PLT not only contribute to acute thrombotic vascular occlusion but also participate in the inflammatory and matrix-degrading processes of coronary atherosclerosis itself. Platelet- endothelial cell interactions at lesion-prone sites might trigger an inflammatory response in the vessel wall early in the genesis of atherosclerosis and contribute to destabilization of advanced atherosclerotic lesions (Massberg et al. 2003). PLT is also involved in the pathology of acute stroke, since early platelet adhesion/activation mechanisms are critical pathogenic factors in infarct development and trigger a thrombo-inflammatory cascade in acute stroke that results in infarct growth.

There is an abundance evidence for PLT's involvement and association with CVD. In 1986, it was first reported that a decrease of PLT and an increase of MPV correlated with infarct size (Glud et al. 1986). Abnormalities of platelet function may contribute to the relatively poor prognosis of myocardial infarction in patients with diabetes (Hendra et al. 1988), and vascular and nonvascular death (Thaulow et al. 1991). Some other associations are especially with MPV but not PLT count. Larger platelets have a greater mass and a greater prothrombotic potential than smaller platelets. The larger and more reactive platelets are enriched in individuals with known CAD risk factors including hypercholesterolaemia (Pathansali et al. 2001) and hypertension (Nadar et al. 2004), and might be causally related to ongoing coronary artery obstruction in unstable angina (Pizzulli et al. 1998). However, in spite of the strong link between MPV and increased CAD risk, there is no data from clinical trials to show that reducing MPV could bring favourable CAD outcomes.

5.1.3 Genetic determinants of FBC

It is estimated that the heritability is 0.67, 0.38, 0.53 for RBC, WBC, PLT respectively, based on a study with >6,000 healthy Sardinians (Pilia et al. 2006). A Twin study showed slightly different numbers especially for WBC, with 0.37, 0.42, 0.62, and 0.57 for HGB, RBC, WBC, PLT respectively (Garner et al. 2000). Blood cell traits are particularly well-suited for

genetic association studies and functional follow-up because they are usually available in most cohorts or biobanks and there are well-developed cell culture systems or model organisms. Large-scale gene silencing and other functional experiments in fruit flies, zebrafish and mice were already shown to be effective for validating genetic loci identified by GWAS (Gieger et al. 2011, van der Harst et al. 2012).

Findings from candidate gene and linkage analysis

Candidate gene studies identified a few loci for association with FBC. The first well studied gene is HBB (β-globin). Mutations in this gene are implicated with several genetic disorders such as sickle-cell disease and beta thalassemia. Other mutations in this gene also bring beneficial effects such as genetic resistance to malaria (Kwiatkowski 2005). Mutations were also found in two other genes: mutations in *EPOR* (erythropoietin receptor) causing familial erythrocytosis (Watowich et al. 1999, Zeng et al. 2001), and mutations in *HFE* (hemochromatosis) causing hereditary hemochromatosis (McLaren et al. 2007). Linkage studies also identified a few reproducible signals, most notably a linkage peak that encompasses the MYB transcription factor (Lin et al. 2007, Menzel et al. 2007).

Findings from first generation GWAS

As shown in **Table 5.1**, a total of 25 GWAS have been conducted for FBC related traits since 2008. The largest studies of blood cells, based on individuals of European ancestry, have so far identified 75, 10 and 68 SNPs for RBC (van der Harst et al. 2012), WBC (Nalls et al. 2011), and platelet traits (Gieger et al. 2011) respectively. There are much fewer associated loci for WBC because its GWAS had a smaller sample size and there is heterogeneity among WBC sub-phenotypes. Like GWAS for other quantitative traits such as lipids, the variants discovered from blood cell GWAS explained a small fraction of the heritable variation (<10%). Also, like the lipids traits, most loci are associated with a single blood cell trait while a few presented pleiotropic effects. This includes two loci (*SH2B3*, *HBS1L-MYB*) associated with all three blood cell traits, both of which have clear biological impact on hematopoiesis.

Again, like lipids traits, many variants discovered through GWAS for association with FBC are within or near genes that are causal for Mendelian hematological disorders, for

example, SNPs near *TMPRSS6*, *HFE*, *TRF2* (for iron deficiency), *HK1* (for hemolytic anemia), and *TBUU1* (for throbocytopenia). Due to the much denser scanning of the genome compared to linkage studies, GWAS was able to pinpoint stronger candidate genes for some of these overlapping loci. Unlike lipids traits or many other traits, where GWAS loci and effects are comparable among multiple ethnic groups (Monda et al. 2013), there are notable exceptions for FBC traits. For example, genetic variants near the gens of α-globin, β-globin and *G6PD* are much more common in African populations because they provide a selective advantage against malaria infections.

Findings from next generation sequencing

No studies have been reported using next generation sequencing.

Table 5.1 GWAS studies on FBC traits

Date is for publication date. Samples are all European ancestry unless explicitly specified otherwise: IND for Indian, JAP for Japanese, AA for African American. The sample size before "+" is for discovery while the sample size after "+" is for replication.

Date	Sample	Main findings	Reference
2008-11	1,062	No SNPs associated with FBC traits at P<5E-08	(Yang et al. 2007)
2008-12	411 from families and 459 twins	Variants in TF and HFE explain ~40% of genetic variation in serum-transferrin levels	(Benyamin et al. 2009)
2008-12	1,606+8,617	Identified 3 loci associated with MPV	(Meisinger et al. 2009)
2009-02	1,221+7,365	A variant on 7q22.3 for MPV and PLT	(Soranzo et al. 2009)
2009-10	4,627+9,316	22 loci for 8 hematological parameters	(Soranzo et al. 2009)
2009-10	16,001 EA and IND	Missense variant in TMPRSS6 for HGB	(Chambers et al. 2009)
2009-10	4,818+3470	Variants in TMPRSS6 are associated with iron status	(Benyamin et al. 2009)
2009-10	3,477+1543	3 loci for monocyte counts and erythrocyte volume	(Ferreira et al. 2009)
2009-10	24,167+9,456	5 know loci, 18 novel loci	(Ganesh et al. 2009)
2010-02	14,700 JAP	46 new and 43 known associations	(Kamatani et al. 2010)
2010-09	3012	demonstrate feasibility of using EMR for GWAS	(Kullo et al. 2010)
2011-03	679+232	2 replicated loci for iron deficiency	(McLaren et al. 2011)
2011-06	8,794+5998 JAP	nine novel loci associated with WBC subtypes	(Okada et al. 2011)
2011-06	16,388 AA	CXCL2, CDK6, PSMD3-CSF3 associated with WBC	(Reiner et al. 2011)
2011-07	19,509+11,823	7 loci associated with WBC	(Nalls et al. 2011)
2011-10	13,923	2 loci each for EA and AA, for WBC	(Crosslin et al. 2012)
2011-11	~18,600+18838	68 loci reliably for PLT and MPV	(Gieger et al. 2011)
2012-03	16,388 AA	5 novel loci for PLT	(Qayyum et al. 2012)
2012-12	62,553 +63506	75 loci for RBC	(van der Harst et al. 2012)
2012-12	62,34EA and 7943 AA	5 novel loci for EA RBC, 1 novel for AA PLT	(Li et al. 2013)
2013-02	16,485	Extended several RBC loci from EA to AA	(Chen et al. 2013)
2013-03	11,014	4 novel loci for monocyte count	(Crosslin et al. 2013)
2013-05	1,904+411 AA	malaria resistance variants associated with RBC	(Ding et al. 2013)
2013-07	1,664+2,200	Identified TAF3 as a gene for MCHC	(Pistis et al. 2013)
2013-09	13,582	EMR based, no new loci reported	(Shameer et al. 2014)

5.1.4 Aims of this study

To discover novel variants, especially those with low or rare frequency but large effects, this study used WGS data from the UK10K project for an upgraded genome-wide scan on eight FBC traits (RBC, HGB, MCH, MCHC, MCV, PCV, PLT, WBC). The current study is by far the largest WGS based association study of FBC traits, with up to 1,497 WGS samples and more than 21,000 samples with WGS imputed data. I first analysed the WGS samples aiming to discover rare and low frequency variants with large effect sizes. Then I analysed a much larger group of cohorts with imputed data to discover novel associations across the full MAF spectrum. Besides standard approaches including single marker based test and rare

variants collapsing test, this study also explored a few novel methods for a comprehensive assessment on the genetics of FBC. This included fine-mapping of known loci to identify causal variants, assessing enrichment in various functional and regulatory features, and an exploring of relationship between genetic variants associated with FBC and host response to infectious diseases including tuberculosis and malaria (Ding et al. 2013, McMorran et al. 2013)

5.2 Methods

5.2.1 Cohorts & phenotype measurements

The phenotype harmonization protocol for the FBC traits in TwinsUK was presented in **Table 5.2**. For TwinsUK, previously I separated it into TwinsUK WGS samples and TwinsUK imputed sample for lipids analysis. As mentioned in **Section 2.6**, after running an evaluation on lipids traits, I combined the genetic data for TwinkUK WGS and imputed samples together so that the co-Twins could also be included for analysis. This is necessary since there was a relative small number of samples available for FBC traits. A total of 12 cohorts were included for the expanded discovery for FBC, and six WHI cohorts of European ancestry were included for replication (**Table 5.3**). All these WHI cohorts had genome-wide results available.

For ALSPAC, HGB were measured with Hemocue Hb201+ analyser. For all eight FBC traits in TwinsUK and all other discovery cohorts, the traits were measured with Beckman Coulters, except for WHI, where HGB, HCT, WBC, and PLT were determined at local laboratories using automated hematology cell counters and standardized quality assurance procedures (Margolis et al. 2005). Different phenotype transformation protocol was applied to the eight FBC phenotypes: inverse normal transformation for HGB, PCV, PLT, square root for MCH, natural log for WBC, and no transformation for MCHC, MCV, RBC. For each trait of each cohort, the residuals with confounding variables regressed out were standardized so that the phenotype has a mean of 0 and a standard deviation of 1.

Table 5.2 Phenotype harmonization protocol for FBC traits

Analyser and visit were tested as random effect variables, while the others including age and age^2 are tested as fixed effect covariates.

Dataset	Trait	Transformation	Gender	Co-variates tested	Filter	Analyser	Visit
TwinsUK GWA	WBC	Natural log	no	age, age^2, sex,dov	3 SD	--	no
TwinsUK WGS	WBC	Natural log	--	age, age^2, dov (2 and 3 periods)	3 SD	--	no
TwinsUK GWA	MCH	Square	no	age, age^2, sex,dov	3 SD	--	yes
TwinsUK WGS	MCH	Square	--	age, age^2, dov (2 and 3 periods)	3 SD	--	yes (3 periods)
TwinsUK GWA	MCHC	untransformed	no	age, age^2, sex,dov	3 SD	--	no
TwinsUK WGS	MCHC	untransformed	--	age, age^2, dov (2 and 3 periods)	3 SD	--	no
TwinsUK GWA	MCV	untransformed	no	age, age^2, sex,dov	3 SD	--	yes
TwinsUK WGS	MCV	untransformed	--	age, age^2, dov (2 and 3 periods)	3 SD	--	yes (3 periods)
TwinsUK GWA	PCV	inverse normal	no	age, age^2, sex,dov	3 SD	--	yes
TwinsUK WGS	PCV	inverse normal	--	age, age^2, dov (2 and 3 periods)	3 SD	--	yes (3 periods)
TwinsUK GWA	PLT	inverse normal	no	age, age^2, sex,dov	3 SD	--	yes
TwinsUK WGS	PLT	inverse normal	--	age, age^2, dov (2 and 3 periods)	3 SD	--	yes (2 periods)
TwinsUK GWA	RBC	untransformed	no	age, age^2, sex,dov	3 SD	--	yes
TwinsUK WGS	RBC	untransformed	--	age, age^2, dov (2 and 3 periods)	3 SD	--	yes (3 periods)

5.2.2 Single marker based discovery and follow-up

To discover variants of low and rare frequency with big effect size, I first run genome-wide association for the TwinUK WGS cohort, with up to 1,497 samples (**Table 5.3**). For HGB, genome-wide association for the ALSPAC WGS samples were also run and were then meta-analyzed with the TwinUK WGS results. Variants with *P*<1E-6 are deemed of interest for follow-up and further characterization. To discover novel variants across the full MAF spectrum, I included up to 10 more cohorts with imputed data in a 12-way meta-analysis, followed by a replication meta-analysis with up to six independent cohorts from WHI (**Table 5.3**). The WHI data only included four phenotypes: HGB, PCV, PLT, WBC. The 12-way meta-analysis included up to 21,519 samples, while the 6-way replication meta-analysis included up to 20,038 samples. Due to the relatively small number of sample size in the 12-way expanded discovery and given the availability of the full genome-wide results of the 6-way replication cohorts, I also run a further expanded discovery meta-analysis for those four traits (HGB, PCV, PLT, WBC) in a 18-way meta-analysis with up to 41,557 samples. For this 18-way meta-analysis, there was no further data for replication.

The TwinsUK WGS and GWA samples were imputed and analyzed together with GEMMA by adjusting for sample genotype status. As described in the Methods chapter, this included all TwinsUK samples for the association analysis and showed better power than analyzing WGS and imputed samples separately where related samples across the two datasets would have to be excluded. A few in-house GWAS results (from the HaemGen consortium) on these traits were also made available to serve as a more comprehensive list of positive controls.

5.2.3 Rare variant aggregation based discovery and follow-up

To evaluate the aggregation effects of rare variants, I used SKAT-O to discover genomic regions that harbour rare variants with large efforts but those effects could be picked up by single marker based analysis. The method for rare variant aggregation based test was the same as that used for lipids, except that the meta-analysis was only run for HGB since it was the only FBC trait measured and analysed in both TwinsUK and ALSPAC. I first evaluated the associations of rare variants by considering genes as functional units of analysis.

I applied two separate statistical models with different properties to rare variants (MAF<1%): SKAT and burden tests, both implemented in a unified software SKAT-O. As described in chapter 2, in *naïve* tests, all variants in exons, untranslated regions (UTRs) and essential splice sites were considered, and were given equal weight of being causal (50,214 windows for 35,709 genes, mean=35 variants, median=38 variants per window). In functional tests, only loss of function (LoF) and predicted functional variants were included (15,528 gene windows with ≥ 5 variants, mean=18, median=14 variants per gene). Finally, I run the locus-based analysis genome-wide in an agonistic fashion, by constructing ~1.8 million windows of 3 kb each, overlapping by half (median 35 SNVs/window, MAF<1%), assigning an equal weight to all variants. There was no external data available for rareMetal analysis to replicate windows of interest for the FBC traits.

5.2.4 Fine-mapping of known loci

The fine-mapping method was described in chapter 2 and it is the same as that used for lipids. Within each signal I included SNPs in high LD (defined as all variants having $r^2 \geq 0.8$ with the most associated variants in the region). For each FBC trait I first created a list of fine-mapping regions based on HapMap estimates of recombination rates. I then analysed each region separately for each of 10 participating cohort using Bayesian linear additive models, by accounting for covariates as in the general single point association analyses. At the end, the resulting BFs for each variant were multiplied to obtain a joint BF measure of association, with the assumption that each cohort is independent. These BFs were 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 were constructed in order to assess the uncertainty of the fine-mapping analysis.

Table 5.3 Characteristics of participating cohorts

All cohorts are population based, except for TwinsUK. Imputation was conducted with the 1000G and UK10K combined reference panel unless otherwise specified. For each trait of each cohort, the residuals were standardized so that the phenotype has a mean of 0 and a standard deviation of 1.

	Cohort	N	Country	Age	% Female	HGB (g/dl)	RBC (10^{12}/l)
Discovery	TwinsUK	1,497	UK	56 (17-85)	97.3	13.34 (1.02)	4.47 (0.35)
	ALSPAC WGS	1,713	UK	10 (9-11)	50.3	14.22 (1.10)	--
	ALSPAC GWA	1,896	UK	10 (9-12)	49.2	13.98 (0.97)	--
	CBR	5,493	UK	45 (34-67)	58.2	14.73 (0.93)	4.97 (0.34)
	INGI-CARL	413	Italy	50 (18-83)	60.0	15.11 (0.96)	4.65 (0.31)
	INGI-FVG	1,377	Italy	52 (18-92)	58.2	14.56 (0.87)	4.62 (0.28)
	INGI-VB	1,776	Italy	55 (18-102)	56.3	13.96 (0.87)	4.30 (0.27)
	HELIC-Manolis	1,247	Greece	62 (18-99)	57.2	15.10 (0.92)	4.41 (0.30)
	HELIC-Pomak	976	Greece	43 (13-87)	72.1	14.65 (0.86)	4.33 (0.29)
	UKBS	2,070	UK	43 (35-62)	54.1	14.03 (0.77)	4.35 (0.27)
	LURIC-Case	1,633	Germany	61 (17-91)	60.8	13.95 (0.91)	4.82 (0.37)
	LURIC-Ctrl	1,428	Germany	62 (18-92)	59.7	14.02 (1.01)	4.61 (0.35)
Replication	WHI-Garnet	3,821	US	65 (50-79)	100.0	14.01 (0.89)	--
	WHI-Gecco1	1,992	US	65 (50-79)	100.0	13.76 (0.93)	--
	WHI-Gecco2	1,737	US	65 (50-79)	100.0	14.04 (0.99)	--
	WHI-Hipfx	3,825	US	65 (50-79)	100.0	14.03 (1.00)	--
	WHI-Mopmap	3,031	US	65 (50-79)	100.0	13.55 (0.79)	--
	WHI-Whims	5,632	US	65 (50-79)	100.0	13.98 (0.93)	--

	Cohort	MCH (pg)	MCHC (g/dl)	MCV (fl)	PCV (l/l)	PLT (10^9/l)	WBC (10^9/l)
Discovery	TwinsUK	29.92 (1.76)	32.35 (1.38)	83.65 (4.01)	0.43 (0.05)	253.9 (63.1)	6.10 (1.81)
	ALSPAC WGS	--	--	--	--	--	--
	ALSPAC GWA	--	--	--	--	--	--
	CBR	29.73 (1.73)	33.19 (1.03)	89.57 (3.98)	0.49 (0.03)	232.9 (50.9)	6.34 (1.52)
	INGI-CARL	26.34 (1.65)	34.43 (0.99)	92.11 (3.67)	0.47 (0.04)	287.5 (48.8)	6.33 (1.45)
	INGI-FVG	25.82 (1.39)	35.12 (1.10)	87.56 (3.01)	0.46 (0.04)	301.0 (59.1)	7.01 (1.44)
	INGI-VB	30.11 (1.76)	32.78 (1.03)	89.22 (2.99)	0.44 (0.06)	297.4 (49.7)	5.43 (1.21)
	HELIC-Manolis	27.82 (1.68)	33.94 (0.89)	94.01 (3.22)	0.47 (0.05)	221.9 (53.2)	6.10 (1.70)
	HELIC-Pomak	29.01 (1.59)	34.21 (1.21)	89.76 (2.78)	0.50 (0.04)	254.7 (55.4)	7.02 (1.81)
	UKBS	29.88 (1.72)	31.27 (1.03)	93.21 (3.21)	0.48 (0.06)	261.2 (57.3)	5.06 (1.20)
	LURIC-Case	30.21 (1.81)	34.77 (0.95)	87.45 (3.04)	0.45 (0.05)	277.4 (61.0)	6.43 (1.43)
	LURIC-Ctrl	29.01 (1.77)	32.19 (0.97)	91.02 (3.66)	0.45 (0.07)	310.7 (67.3)	5.98 (1.55)
Replication	WHI-Garnet	--	--	--	0.46 (0.05	302.0 (58.9)	4.97 (1.09)
	WHI-Gecco1	--	--	--	0.47 (0.06)	298.3 (54.0)	6.03 (1.42)
	WHI-Gecco2	--	--	--	0.49 (0.03	276.9 (48.8)	6.11 (1.39)
	WHI-Hipfx	--	--	--	0.43 (0.05)	320.1 (59.7)	7.02 (1.83)
	WHI-Mopmap	--	--	--	0.47 (0.04)	300.7 (56.3)	5.32 (1.56)
	WHI-Whims	--	--	--	0.48 (0.05)	288.4 (48.7)	5.05 (1.76)

5.3 Results

5.3.1 Novel loci and novel variants from single marker analysis

WGS for low frequency and rare variants

Here I sought to investigate if low-frequency or rare variants with strong effects could be detected from the WGS dataset. I first tested association results using solely the WGS dataset in order to identify whether these variants existed. Associations were carried out in 13,074,236 SNVs and 1,122,542 biallelic InDels (MAF≥0.1%) using linear regression. For HGB, data from TwinsUK and ALSPAC was meta-analysed.

A total of 60 variants have *P*<5E-08, based on TwinsUK WGS samples alone (**Figure 5.1**). 57 of these variants are in the *HBS1L* (HBS1-like translational GTPase) region. *HBS1L* encodes a member of the GTP-binding elongation factor family, mostly expressed in heart and skeletal muscle. The intergenic region between *HBS1L* and *MYB* is a quantitative trait locus (QTL) that controls fetal hemoglobin level and influences erythrocyte, platelet, and monocyte counts. The other three variants with *P*<5E-08 were not previously reported for associations with blood cell traits. The first one is a low frequency variant for association with PCV (rs114119841, chr2:38831057, EA=C, EAF=0.020, *P*=3.20E-08). This is annotated as a regulatory region variant for *HNRPLL* (heterogeneous nuclear ribonucleoprotein L-like). *HNRNPLL* is a master regulator of activation-induced alternative splicing in T cells. It alters the splicing of a tyrosine phosphatase that is essential for T-cell development and activation (Oberdoerffer et al. 2008). However, this variant was not replicated either, with *P*>0.05 in the 10-way meta-analysis. The second one is a common variant within *COL23A1*, for association with HGB. The index SNP is rs4976769 (chr5:177808188, EA=G, EAF=0.065, *P*=3.85E-08). This variant has a meta-analysis *P*=3.75E-04, based on a total of 10 cohorts and 16,687 samples. Given its allele frequency and the sample size in the 10-way meta-analysis, this signal was not replicated and might be a false positive. The third one is a rare variant for association with MCHC (rs145884292, chr9:24195910, EA=C, EAF=0.008, *P*=2.91E-08). The index SNP (rs145884292) is an intergenic variant, ~5 Mb away from *ELAV2* (embryonic lethal, abnormal vision, Drosophila-like 2), which has no apparent relevance to blood traits.

To look at suggestive associations, I used a less stringent threshold and discovered a further 155 variants have *P*<1E-06, as highlighted blue in **Figure 5.1**. For these 215 variants in total, 25 have MAF between 0.005 and 0.05 (**Table 5.4**). For the given number of WGS samples for FBC traits and the sequencing coverage, there was a high probability (>98%) of detecting variants down to MAF of 0.5% (Li et al. 2011). Among these 25 variants, rs62064540 (for association with MCH) is in proximity to a previously report association with MCHC (rs689992), and rs113833421 is in proximity to previously reported variant rs11672923 (for association with RBC). But there was no LD between the current study's index SNPs and previously reported variants (r2< 0.01) in both cases. The rest 23 variants were not within 1MB of any positive controls. Given the lack of independent replication cohorts with directly sequenced or de novo genotyped data for these variants, I presented the expanded discovery meta-analysis (12-way) results for these variants (**Table 5.4)**. However, none of these 25 variants became more significant in the expanded discovery meta-analysis, all of which had *P*>1E-4 in the 12-way meta-analysis. This could be due to poor imputation or lack of power for replication, or these signals are false positive. Preferably, WGS or directly typed genotype should be used as replication for this set of variants, when resources become available.

Figure 5.1 Association results for WGS based samples for FBC traits

X-axis is for chromosome and positions (build 37). Y-axis is for –log10(*P*). Variants passing threshold of 5E-08 and 1E-06 are shown in red and blue, respectively. For those passing threshold of 5E-08, known loci were marked in green text while putative novel loci were marked in red text.

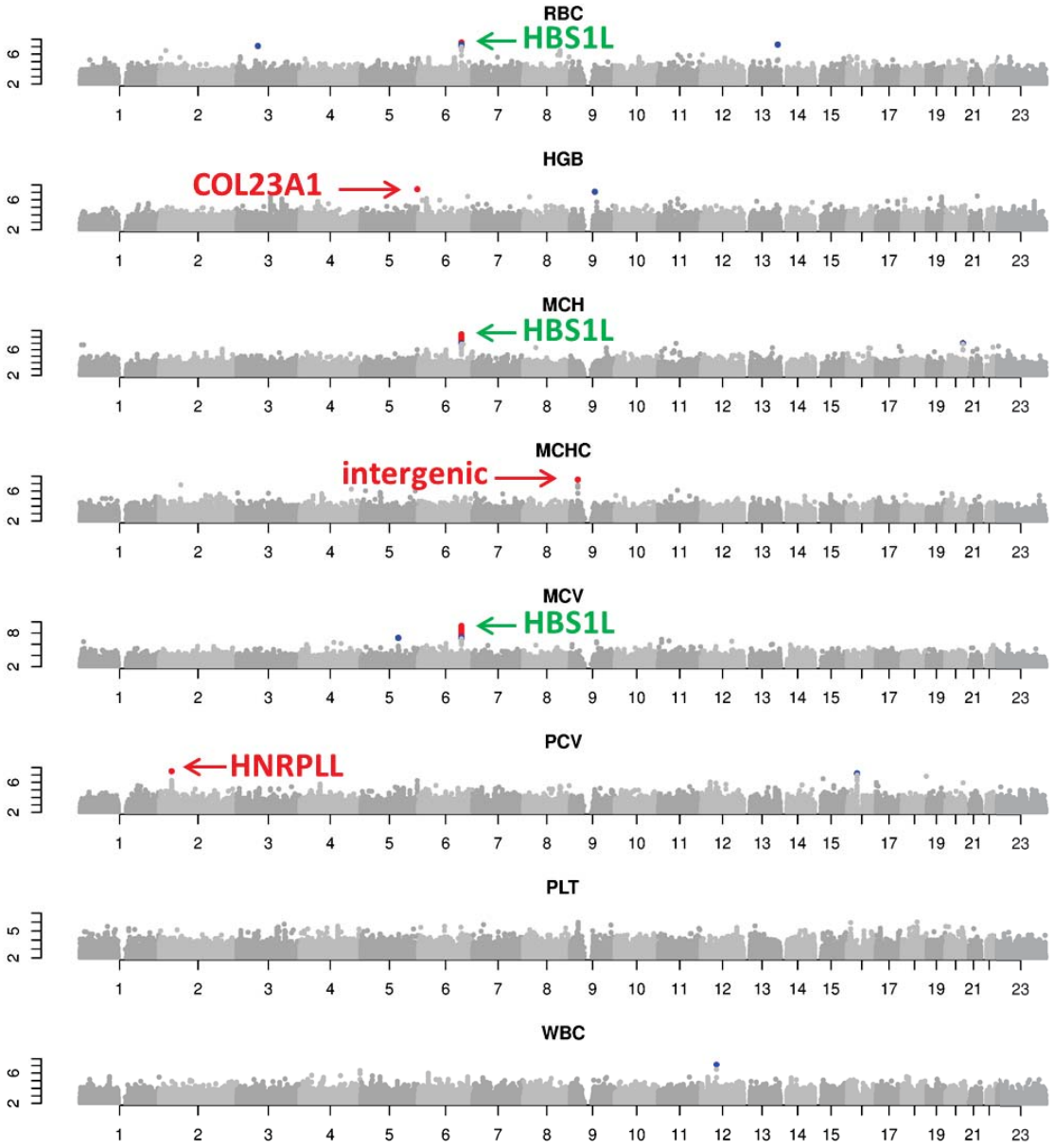


Table 5.4 Putative novel variants of low or rare frequency from UK10K WGS

25 WGS variants (P<1E-6) either have no positive controls within 1Mb or are independently significant from known variants. Six have low frequency (MAF between 1-5%) and could be imputed with fair accuracy.

						WGS				12-way meta-analysis				
trait	**rsID**	**CHR**	**POS**	**EA**	**NEA**	**EAF**	**beta**	**SE**	**P**	**EAF**	**beta**	**SE**	**P**	**N**
RBC	rs76777478	2	20383810	T	G	0.048	0.427	0.083	3.23E-07	0.047	0.042	0.025	9.7E-02	13944
PCV	rs114119841	2	38831057	C	A	0.020	0.705	0.127	3.20E-08	0.025	0.059	0.037	1.2E-01	15350
MCHC	rs146621801	2	67557307	G	C	0.013	-1.095	0.207	1.5E-07	0.014	-0.030	0.055	5.9E-01	12891
MCH	rs186149310	2	197653260	G	A	0.012	-0.845	0.170	7.8E-07	0.010	-0.121	0.094	2.0E-01	12189
RBC	rs189761618	3	66373898	A	G	0.015	0.821	0.152	8.1E-08	0.009	0.083	0.064	2.0E-01	12956
HGB	rs11917207	3	105964555	G	A	0.046	0.296	0.059	6.3E-07	0.047	0.087	0.025	4.7E-04	19751
MCV	rs145802933	4	106800944	G	C	0.008	1.030	0.203	4.6E-07	0.007	0.208	0.094	2.9E-02	15280
MCHC	rs74339994	4	161780587	A	T	0.008	-1.325	0.263	5.4E-07	0.012	-0.039	0.067	5.6E-01	12892
WBC	rs76070316	4	189101798	G	C	0.007	1.044	0.206	4.6E-07	0.014	-0.018	0.026	4.9E-01	15340
MCHC	rs188771831	5	15237510	G	T	0.015	-0.962	0.195	9.2E-07	0.009	-0.033	0.079	6.8E-01	12892
MCV	rs72663338	5	118080521	G	A	0.042	-0.514	0.095	7.3E-08	0.042	-0.065	0.032	4.6E-02	15281
MCHC	rs74964545	5	171263478	T	C	0.012	1.092	0.221	9.4E-07	0.012	0.051	0.070	4.7E-01	12891
MCH	rs6862184	5	177396364	A	G	0.959	-0.442	0.090	9.7E-07	0.963	-0.072	0.041	7.9E-02	12190
MCV	rs181579991	6	87074470	A	G	0.009	0.985	0.198	7.3E-07	0.006	0.159	0.096	1.0E-01	14327
HGB	rs62434477	6	155327584	T	C	0.017	-0.509	0.100	3.5E-07	0.017	-0.131	0.047	5.6E-03	19752
MCHC	rs145884292	9	24195910	C	T	0.008	-1.452	0.260	2.9E-08	0.007	-0.142	0.086	1.0E-01	12892
HGB	rs75472650	9	77788213	T	C	0.008	-0.798	0.149	8.3E-08	0.007	-0.217	0.071	2.5E-03	19749
HGB	rs72914272	11	61376274	T	C	0.033	0.358	0.072	7.8E-07	0.027	0.052	0.036	1.6E-01	19751
PCV	rs11829947	12	28334475	C	T	0.012	-0.784	0.159	9.2E-07	0.011	-0.055	0.053	3.0E-01	15350
RBC	rs117125854	13	106362073	A	G	0.006	1.242	0.227	5.3E-08	0.006	0.090	0.071	2.1E-01	13942
PCV	rs67824122	15	26248846	T	A	0.007	-1.422	0.277	3.2E-07	0.007	-0.221	0.091	1.6E-02	15349
MCH	rs62064540 *	17	72171888	C	T	0.008	-1.066	0.208	3.1E-07	0.007	-0.181	0.096	6.0E-02	12190
HGB	rs62087096	18	8998320	T	A	0.037	0.326	0.066	9.1E-07	0.032	0.028	0.029	3.4E-01	19750
PCV	rs148652300	18	76407934	T	C	0.006	1.227	0.233	1.5E-07	0.006	0.203	0.105	5.6E-02	14867
HGB	rs113833421 *	19	46421564	T	C	0.011	0.607	0.120	4.2E-07	0.013	0.092	0.046	4.6E-02	19751

* rs62064540 (for association with MCH) is in proximity to previously reported variant rs689992 (for association with MCHC). rs113833421 is in proximity to previously reported variant rs11672923 (for association with RBC). But the LD between the current study's index SNPs and previously reported variants are less than 0.01 in both cases.

Meta-analysis for identifying novel variants of all allele spectrums

Given the enhanced imputation quality with the UK10K WGS reference panel as demonstrated in chapter 3, I included up to 10 more cohorts with imputed data for an expanded discovery, to increase power for discover variants across all allele frequency spectrum. Only HGB was measured in ALSPAC WGS and ALSPAC imputed data and have a total of 12 cohorts for meta-analysis, while the other FBC traits included only 10 cohorts for meta-analysis. Variants with MAF <0.1% or imputation INFO <0.4 were not included. The genome-wide results for the expanded discovery was presented in **Figure 5.2**. A total of 3,952 variants passed the pre-defined threshold for genome-wide and suggestive significance (*P*<1E-07). Through the step-wise conditional analysis as described in chapter 2 and the methods section of this chapter, nine loci were found to be putative novel, and three known loci harboured novel variants (**Table 5.5**). The detailed results for each participating cohort are shown in **Table 5.6**. For the nine putative novel loci, three of them didn't have any other variants with *P*<1E-5 within 1Mb and there was a lack of supporting signals from SKAT-O test. These three are rare with MAF< 0.5%. Therefore, they are most likely to be false positive or would be difficult to be replicated. For the other six loci, further replication would be needed to confirm the association.

Given the availability of the genome-wide results for the six replication cohorts (for four traits: HGB, PCV, PLT, WBC), I run an 18-way meta-analysis that included the 12 discovery cohorts and six replication cohorts. Based on this 18-way meta-analysis, I identified a total of 12 associations that have *P*<5E-08 while their associations did not meet the significance threshold (*P*<1E-07) pre-defined for the 12-way analysis (**Table 5.7**). Although further independent replication is needed to confirm these associations, two signals have such strong associations that might not need further replication. The first one is the association of WBC in the *HLA* locus. A recent trans-ethnic GWAS meta-analysis on WBC reported an association within this region (Keller et al. 2014), but the reported lead SNP (rs2853946, chr6:31 247 203, EUR MAF=0.348) is in low LD with the lead SNP of this study (rs113164910, chr6:32427005, LD r2=0.08). The other strong signal from the 18-way is the (growth factor independent 1B transcription repressor) locus for association with PLT. The lead SNP (rs150813342) is a rare (18-way MAF=0.007) synonymous SNP within *GFI1B*, which encodes a zinc-finger containing transcriptional regulator that is primarily expressed in cells of hematopoietic lineage. The encoded protein complexes with numerous other transcriptional regulatory proteins to control expression of genes involved in the development

and maturation of erythrocytes and megakaryocytes. Mutations in this gene are the cause of the autosomal dominant platelet disorder, platelet-type bleeding disorder-17 (Monteferrario et al. 2014).

For the three putative novel variants that are less than 1Mb away from previously reported variants for association with FBC traits, the association details of known variants and their LD with the putative novel variants in LD are listed in **Table 5.8**. The first one is the *CCND3* locus on chromosome 6, where three common variants were reported for association with MCV. All significantly associated SNPs within the *CCND3* locus are tagged by previously reports variants, except for rs112233623 and another SNP in high LD (rs113267280, chr6:41952511, LD r2=0.74) (**Figure 5.3)**. The second known locus with novel variants is on chromosome 11. Upon further examination of individual cohort results, I found that this association in the meta-analysis was mainly driven by one isolated population, HELIC-Pomak. The lead SNP rs11821302 has an EAF of 0.001 in TwinsUK but an EAF of 0.05 in HELIC-Pomak. For the HELIC-Pomak cohort, the lead SNP in this region is rs7116019 (chr11:4618606) (Zeggini 2014), but it is not significant (P>0.05) in TwinsUK or any other cohorts included in the 10-way meta-analysis. In this locus, there is a variant associated with protective immunity against severe malaria (rs11036238), which might offer some clue on the genetic isolate's response to malaria infection (Jallow et al. 2009). The third locus with novel variants is within *NPRL3* (nitrogen permease regulator-like 3) on chromosome 16. The two known variants are much more common and they are within a different gene *ITFG3* (integrin alpha FG-GAP repeat containing 3). Within 1Mb region, there is no other SNP in high LD with the index SNP rs117747069 (**Figure 5.3**). However, based on the Regulome database (http://regulome.stanford.edu), the functional evidence for rs117747069 is much stronger than the two known variants in this region (rs7189020 and rs1122794). The Regulome score is "2b" (supporting data from TFBS, motif, DNase footprint, and DNase peak) for rs117747069 and "5" (supporting evidence from TFBS or DNase peak) for rs7189020, while there is no functional data available for rs1122794. The regional plots for the two strongest associations based on 18-way meta-analysis were shown in **Figure 5.4.**

Figure 5.2 Results for 12-way meta-analysis

X-axis is for chromosome and positions (build 37). Y-axis is for –log10(*P*). Variants passing threshold of 5E-08 and 1E-07 are shown in red and blue, respectively. For those passing threshold of 5E-08, known loci were marked in green text while putative novel loci were marked in red text.

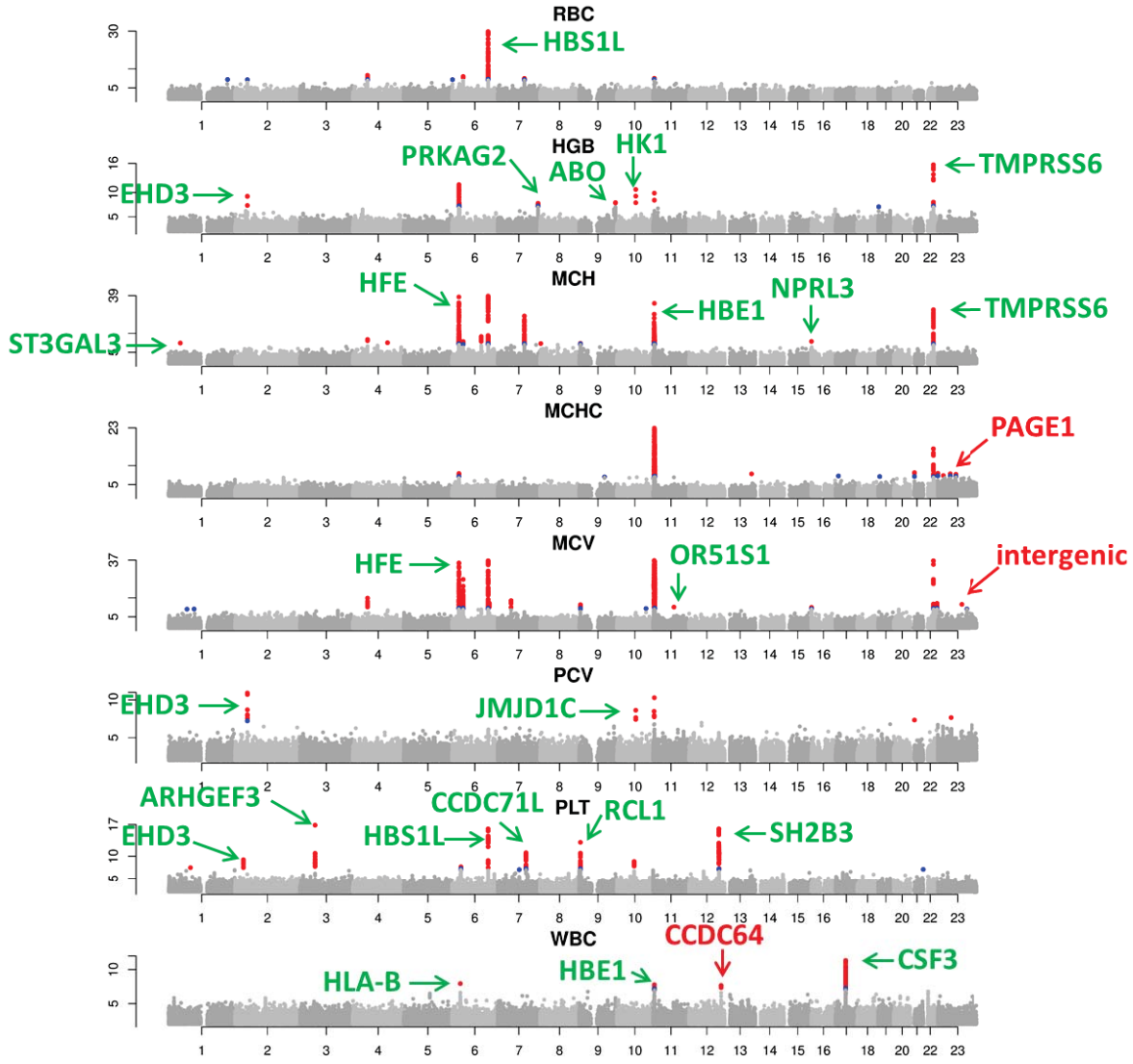


Table 5.5 Novel FBC variants based on expanded discovery (12-way meta-analysis)

The top part listed the index SNP for 9 putative novel loci. The bottom part listed three variants that have positive controls within 1Mb.

For the index SNVs in the nine novel loci, three are lonely variants and have no supporting SKAT signal, as labelled with * in the table.

							12-way meta-analysis					Replication				
Type	**Trait**	**rsID**	**CHR**	**POS**	**Gene**	**EA**	**EAF**	**Beta**	**SE**	**P**	**N**	**EAF**	**Beta**	**SE**	**P**	**N**
Putative Novel Loci	MCV	chr1:69249341*	1	69,249,341	intergenic	C	0.001	-2.024	0.376	9.87E-08	8,321	--	--	--	--	--
	MCV	rs189931100 *	1	96,028,784	intergenic	G	0.001	-1.989	0.369	9.18E-08	9,827	--	--	--	--	--
	RBC	chr6:1906294	6	1,906,294	GMDS	T	0.002	0.786	0.145	7.06E-08	13,944	--	--	--	--	--
	MCV	rs189443777	10	109,452,247	intergenic	A	0.008	-0.410	0.075	6.51E-08	14,804	--	--	--	--	--
	WBC	rs74853946	12	120,501,797	CCDC64	T	0.018	-0.181	0.032	2.14E-08	15,342	0.021	0.035	0.036	0.426	20,062
	MCHC	rs144022851	21	14,589,985	intergenic	T	0.090	0.196	0.033	7.16E-09	12,893	--	--	--	--	--
	PLT	rs200989541 *	21	47,565,506	FTCD	A	0.004	0.821	0.152	7.85E-08	8,703	0.001	-0.065	0.358	0.872	9,418
	MCHC	rs143473229	X	49,514,596	PAGE1	G	0.016	-0.255	0.045	1.47E-08	10,858	--	--	--	--	--
	MCV	rs73221860	X	111,785,547	--	G	0.207	0.075	0.014	3.99E-08	14,173	--	--	--	--	--
Putative Novel variants	MCV	rs112233623	6	41,924,998	*CCND3*	T	0.011	0.384	0.062	9.15E-10	15,277	--	--	--	--	--
	MCV	rs11821302	11	4,868,158	*OR51S1*	T	0.009	-1.161	0.094	1.38E-34	6,893	--	--	--	--	--
	MCH	rs117747069	16	170,076	*NPRL3*	C	0.032	-0.280	0.049	1.33E-08	12,189	--	--	--	--	--

Table 5.6 Cohort specific results of top hits from expanded discovery analysis

For each set of results, the effect allele frequency (EAF), beta, standard deviation (SE), *P* value, sample size (N), and imputation INFO score were presented. Records with *P* < 0.05 are highlighted in red text.

	MCV, chr1:69249341						MCV, chr1:96028784						RBC, chr6:1906294						MCV, chr6:41924998					
cohort	**EAF**	**beta**	**SE**	**P**	**N**	**Info**	**EAF**	**Beta**	**SE**	**P**	**N**	**Info**	**EAF**	**beta**	**SE**	**P**	**N**	**Info**	**EAF**	**beta**	**SE**	**P**	**N**	**Info**
TwinsUK WGS	0.001	-1.976	0.710	5.4E-03	1548	0.99	-	-	-	-	-	-	0.003	0.061	0.315	8.5E-01	1561	1.00	0.011	0.249	0.181	1.7E-01	1548	0.92
CARL	0.000	18.54	44.05	6.8E-01	474	0.05	0.001	12.780	17.89	4.8E-01	474	0.16	0.001	2.803	2.776	3.2E-01	480	0.35	0.003	2.029	5.027	6.9E-01	474	0.36
CBR	-	-	-	-	-	-	0.002	-1.777	0.635	5.2E-03	1033	0.51	0.003	0.670	0.439	1.3E-01	1033	0.76	0.013	0.415	0.211	4.9E-02	1033	0.88
FVG	0.001	-1.515	12.827	9.0E-01	1374	0.06	0.001	-1.475	4.795	7.6E-01	1374	0.43	0.001	0.746	0.369	4.3E-02	1396	0.50	0.006	1.091	1.047	3.0E-01	1374	0.91
HA	0.002	-0.959	1.619	5.6E-01	979	0.15	0.002	-0.321	1.506	8.3E-01	979	0.14	0.002	0.503	0.600	4.0E-01	989	0.62	0.007	0.769	0.293	8.9E-03	979	0.88
HP	0.001	-3.749	2.041	6.7E-02	954	0.15	0.001	-3.115	2.501	2.2E-01	954	0.10	0.001	2.845	7.515	7.0E-01	968	0.02	0.028	0.483	0.157	2.2E-03	954	0.90
LURIC-1	0.001	-3.462	0.745	3.7E-06	1428	0.55	0.001	-1.368	0.851	1.1E-01	1428	0.36	-	-	-	-	-	-	0.011	0.257	0.195	1.9E-01	1428	0.84
LURIC-2	-	-	-	-	-	-	-	-	-	-	-	-	0.003	0.945	0.367	1.0E-02	1633	0.80	0.014	0.201	0.162	2.2E-01	1633	0.87
TwinsUKall	0.001	-1.464	0.465	1.7E-03	3586	0.75	0.001	-2.666	0.584	7.8E-06	3586	0.41	0.002	0.550	0.307	7.4E-02	3609	0.75	0.012	0.308	0.121	1.1E-02	3586	0.90
TwinsUK	0.000	28.194	22.13	2.0E-01	1058	0.68	0.001	-3.186	0.988	1.4E-03	1058	0.38	0.001	1.601	0.832	5.6E-02	1062	0.71	0.010	0.364	0.259	1.6E-01	1058	0.88
UKBS	0.001	0.239	0.557	6.7E-01	2065	0.84	-	-	-	-	-	-	0.003	0.709	0.325	2.9E-02	2067	0.75	0.010	0.354	0.169	3.6E-02	2065	0.89
VB	0.000	-11.98	9.322	2.0E-01	1755	0.10	0.000	-11.145	6.128	6.9E-02	1755	0.03	0.003	1.479	0.453	1.2E-03	1770	0.62	0.007	0.769	0.243	1.6E-03	1755	0.79

	MCV, chr10:109452247						MCV, chr11:4868158						WBC, chr12:120501797						MCH, chr16:170076					
cohort	**EAF**	**beta**	**SE**	**P**	**N**	**Info**	**EAF**	**beta**	**SE**	**P**	**N**	**Info**	**EAF**	**beta**	**SE**	**P**	**N**	**Info**	**EAF**	**beta**	**SE**	**P**	**N**	**Info**
TwinsUK WGS	0.006	-1.135	0.233	1.2E-06	1548	0.97	-	-	-	-	-	-	0.018	-0.061	0.137	6.6E-01	1551	0.98	0.037	-0.326	0.096	6.8E-04	1549	0.96
CARL	0.000	11.043	44.917	8.1E-01	474	0.06	0.000	2.549	219.43	9.8E-01	474	0.01	0.010	-0.355	0.093	2.6E-04	484	0.73	0.028	27.678	35.395	4.3E-01	473	0.58
CBR	0.007	-0.432	0.271	1.1E-01	1033	0.92	-	-	-	-	-	-	0.023	-0.216	0.147	1.4E-01	1033	0.96	0.039	-0.096	0.140	4.9E-01	1033	0.64
FVG	0.010	-0.250	0.813	7.6E-01	1374	0.91	0.001	-2.683	2.809	3.4E-01	1374	0.87	0.010	-0.159	0.047	8.4E-04	1387	0.83	0.029	-17.80	13.921	2.0E-01	1357	0.54
HA	0.006	-0.290	0.362	4.2E-01	979	0.69	0.005	-0.637	0.344	6.6E-02	979	1.00	0.014	-0.318	0.245	1.9E-01	990	0.69	0.022	0.190	0.261	4.7E-01	981	0.36
HP	0.001	0.089	0.706	9.0E-01	954	0.84	0.052	-1.185	0.100	1.3E-26	954	1.00	0.001	0.567	1.111	6.1E-01	963	0.75	0.030	0.026	0.182	8.8E-01	949	0.58
LURIC-1	0.008	-0.147	0.222	5.1E-01	1428	0.86	0.001	0.398	0.681	5.6E-01	1428	0.84	0.024	-0.205	0.131	1.2E-01	1428	0.86	-	-	-	-	-	-
LURIC-2	0.010	-0.246	0.192	2.0E-01	1633	0.87	-	-	-	-	-	-	0.023	-0.216	0.127	9.0E-02	1633	0.85	-	-	-	-	-	-
TwinsUKall	0.007	-0.498	0.158	1.7E-03	3586	0.92	0.001	-1.618	0.510	1.6E-03	3586	0.95	0.018	-0.091	0.096	3.4E-01	3597	0.92	0.038	-0.348	0.073	2.4E-06	3587	0.71
TwinsUK	0.008	-0.021	0.313	9.5E-01	1058	0.92	0.001	-1.932	0.698	6.0E-03	1058	0.95	0.019	-0.086	0.179	6.3E-01	1065	0.91	0.035	-0.298	0.141	3.5E-02	1061	0.70
UKBS	0.008	-0.518	0.177	3.4E-03	2065	0.91	-	-	-	-	-	-	0.023	-0.108	0.105	3.0E-01	2053	0.92	0.033	-0.194	0.109	7.5E-02	2061	0.63
VB	0.012	-0.554	0.174	1.5E-03	1755	0.96	0.000	-33.81	163.74	8.4E-01	1755	0.00	0.015	-0.167	0.164	3.1E-01	1774	0.79	0.026	-0.673	0.141	2.1E-06	1749	0.61

Table 5.6 Cohort specific results of top hits from expanded discovery analysis (*continued*)

	MCHC, chr21:47565506						PLT, chr21:14589985						MCHC, chrX:49514596						MCV, chrX:111785547					
Cohort	**EAF**	**beta**	**SE**	**P**	**N**	**Info**	**EAF**	**beta**	**SE**	**P**	**N**	**Info**	**EAF**	**beta**	**SE**	**P**	**N**	**Info**	**EAF**	**beta**	**SE**	**P**	**N**	**Info**
TwinsUK WGS	-	-	-	-	-	-	-	-	-	-	-	-	0.005	-0.056	0.339	8.7E-01	942	1.00	0.181	0.023	0.047	6.3E-01	1548	1.00
CARL	-	-	-	-	-	-	0.014	0.148	0.446	7.5E-01	483	0.89	0.005	1.437	58.397	9.8E-01	483	0.00	0.171	12.204	6.031	4.5E-02	474	0.01
CBR	-	-	-	-	-	-	0.027	0.160	0.210	4.5E-01	1033	0.42	-	-	-	-	-	-	-	-	-	-	-	-
FVG	0.001	0.834	1.233	5.0E-01	1375	0.15	0.668	0.228	0.040	2.5E-08	1391	0.63	0.099	-0.236	0.048	1.0E-06	1391	0.92	0.281	0.088	0.150	5.5E-01	1374	0.80
HA	0.001	1.182	1.792	5.1E-01	991	0.20	0.043	0.241	0.171	1.6E-01	994	0.41	0.001	3.485	2.416	1.5E-01	994	0.27	0.213	0.121	0.048	1.3E-02	979	0.99
HP	0.024	0.852	0.163	3.4E-07	968	0.91	0.036	-0.060	0.232	7.9E-01	963	0.28	0.000	-9.721	9.655	3.2E-01	963	0.01	0.297	0.168	0.046	3.1E-04	954	0.98
LURIC-1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.198	0.092	0.037	1.4E-02	1392	0.99
LURIC-2	-	-	-	-	-	-	0.015	-0.059	0.158	7.1E-01	1633	0.78	0.005	-0.204	0.242	4.0E-01	1594	0.43	0.202	0.048	0.033	1.5E-01	1594	0.99
TwinsUKall	0.001	0.582	0.453	2.0E-01	3602	0.51	0.011	0.472	0.142	9.6E-04	2565	0.82	0.004	-0.382	0.275	1.6E-01	2565	0.70	0.191	-0.007	0.032	8.3E-01	3586	1.00
TwinsUK	0.001	-0.029	0.794	9.7E-01	1070	0.46	0.011	0.593	0.238	1.3E-02	947	0.81	0.005	-0.653	0.382	8.8E-02	947	0.63	0.213	-0.069	0.055	2.1E-01	1058	1.00
UKBS	-	-	-	-	-	-	0.019	0.096	0.150	5.2E-01	2059	0.58	0.005	-0.494	0.198	1.3E-02	2059	0.74	0.183	0.111	0.033	7.7E-04	2065	0.99
VB	0.001	0.354	1.989	8.6E-01	1767	0.11	0.018	-0.072	0.136	6.0E-01	1772	0.89	0.003	-0.643	0.405	1.1E-01	1772	0.37	0.177	0.069	0.038	7.1E-02	1755	1.00
WHI-garnet	0.001	0.552	0.680	4.2E-01	3802	0.26	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
WHI_gecco1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
WHI_gecco2	0.001	0.734	1.008	4.7E-01	1733	0.44	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
WHI_hipfx	0.001	0.210	0.666	7.5E-01	3807	0.33	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
WHI_mopmap	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
WHI_whims	0.001	-0.303	0.422	4.7E-01	5617	0.36	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Table 5.7 Top hits from a further expanded discovery (18-way meta-analysis)

For each set of results, the effect allele frequency (EAF), beta, standard deviation (SE), *P* value, and the total sample size were presented. For positive controls within 1Mb, only the one in highest LD is shown when there are multiple ones. The information includes trait name, rsID, CHR:POS, and LD measured in r2.

							WGS			
Trait	**CHRPOS**	**rsID**	**Positive Controls within 1Mb**	gene	**EA**	**NEA**	**EAF**	**beta**	**SE**	**P**
PCV	chr1:9,077,128	rs769904	--	SLC2A7	C	T	0.002	-0.223	0.409	5.9E-01
PLT	chr1:24,743,879	rs760968	PLT, rs592372, chr1:25636197, NA	C1orf201	T	C	0.246	0.008	0.042	8.6E-01
HGB	chr2:159,916,661	rs113682276	--	TANC1	A	G	0.009	0.087	0.141	5.4E-01
PLT	chr3:56,929,498	rs200858303	PLT, rs1354034, chr3:56849749, 0.06	ARHGEF3	T	TTA	--	--	--	--
WBC	chr6:32,427,005	rs113164910	HGB, rs9272219, chr6:32602269, 0.036	HLA-DRB9	A	AAC	0.327	-0.081	0.038	3.2E-02
PLT	chr9:91,459,039	rs141068793	PLT, rs11142062, chr9:90658749, NA	--	C	T	0.062	-0.113	0.078	1.5E-01
PLT	chr9:135,864,513	rs150813342	HGB, rs4128808, chr9:136065229, 0.011	GFI1B	T	C	0.004	-0.229	0.291	4.3E-01
WBC	chr17:7,231,792	rs9905997	--	NEURL4	G	A	0.44	0.095	0.035	7.5E-03
PLT	chr17:64,195,431	rs75003668	--	PSMD7P1	G	A	0.033	0.221	0.115	5.4E-02
HGB	chr20:22,110,210	rs138233587	--	--	A	AT	0.046	-0.078	0.06	1.9E-01
PLT	chr21:36,474,114	rs2834764	--	RUNX1	A	G	0.415	-0.036	0.036	3.2E-01
PLT	chr22:50,570,755	rs75570992	RBC, rs140522, chr22:50971266, 0.00	MOV10L1	C	G	0.072	0.113	0.069	1.0E-01

		12-way					6-way					18-way				
Trait	**CHRPOS**	**EAF**	**beta**	**SE**	**P**	**N**	**EAF**	**beta**	**SE**	**P**	**N**	**EAF**	**beta**	**SE**	**P**	**N**
PCV	chr1:9077128	0.002	-0.415	0.155	8.0E-03	15,351	0.005	-0.434	0.078	2.2E-06	20063	0.003	-0.430	0.070	1.1E-08	35,414
PLT	chr1:24743879	0.229	0.045	0.014	1.7E-03	15,327	0.239	0.064	0.012	1.7E-06	19948	0.234	0.056	0.009	5.2E-09	35,275
HGB	chr2:159916661	0.006	0.269	0.075	3.6E-04	19,749	0.007	0.342	0.072	6.6E-05	20034	0.007	0.307	0.052	4.1E-08	39,783
PLT	chr3:56929498	0.420	0.047	0.012	1.4E-04	15,326	0.434	0.062	0.010	1.2E-07	19948	0.428	0.056	0.008	2.7E-11	35,274
WBC	chr6:32427005	0.289	-0.035	0.009	8.7E-05	15,342	0.325	-0.096	0.011	5.8E-14	20062	0.309	-0.059	0.007	2.4E-16	35,404
PLT	chr9:91459039	0.078	-0.093	0.022	3.4E-05	15,326	0.067	-0.086	0.020	1.9E-04	19948	0.072	-0.089	0.015	2.2E-08	35,274
PLT	chr9:135864513	0.007	-0.398	0.080	8.8E-07	15,326	0.008	-0.485	0.061	3.9E-12	19950	0.007	-0.453	0.049	2.4E-18	35,276
WBC	chr17:7231792	0.454	0.032	0.007	4.9E-06	15,342	0.453	0.048	0.010	9.3E-05	20064	0.453	0.037	0.006	1.5E-09	35,406
PLT	chr17:64195431	0.029	0.157	0.041	1.4E-04	15,327	0.028	0.164	0.036	5.3E-05	19948	0.028	0.161	0.027	1.8E-08	35,275
HGB	chr20:22110210	0.056	-0.069	0.023	2.6E-03	19,750	0.053	-0.125	0.023	4.0E-06	20035	0.054	-0.097	0.016	2.3E-08	39,785
PLT	chr21:36474114	0.415	-0.046	0.012	1.2E-04	15,327	0.421	-0.045	0.010	9.7E-05	19949	0.419	-0.046	0.008	3.0E-08	35,276
PLT	chr22:50570755	0.059	0.120	0.027	8.6E-06	14,844	0.061	0.116	0.023	6.3E-06	19946	0.060	0.118	0.017	1.4E-10	34,790

Table 5.8 LD of three putative novel variants in known locus

For each locus, the 10-way association statistics and the LD for all known variants within 1Mb of the putative novel variants are listed.

Novel variants	Known variants	Associated traits	CHR:POS	MAF	10-way P	LD (r2)
rs112233623 (chr6:41924998)	rs3218097	MCV	chr6:41905275	0.247	6.72E-11	0.027
	rs9349205	MCV	chr6:41925159	0.233	2.01E-11	0.028
	rs11970772	MCV	chr6:41925290	0.214	7.91E-14	0.002
rs11821302 (chr11:4868158)	rs7116019	MCV	chr11:4618606	0.012	3.11E-31	0
	rs11036238	Malaria	chr11:5225635	0.272	--	0
	rs2071348	Beta thalassemia/hemoglubin E	chr11:5264146	0.340	--	0
	rs4910742	Fetal hemoglobin levels	chr11:5306509	0.051	--	0
rs117747069 (chr16:170076)	rs7189020	MCV	chr16:304803	0.376	1.29E-04	0.015
	rs1122794	MCH	chr16:309155	0.181	2.12E-05	0.006

Figure 5.3 Regional plots of two known loci with putative novel variants

The top plot is for the *CCND3* locus for association with MCV. The bottom plot is for *NPRL3* locus for association with MCH. The *P* values are based on the 10-way meta-analysis. The novel variant is shown in red text, while the SNPs tagged by previously reported variants are known in other colors.

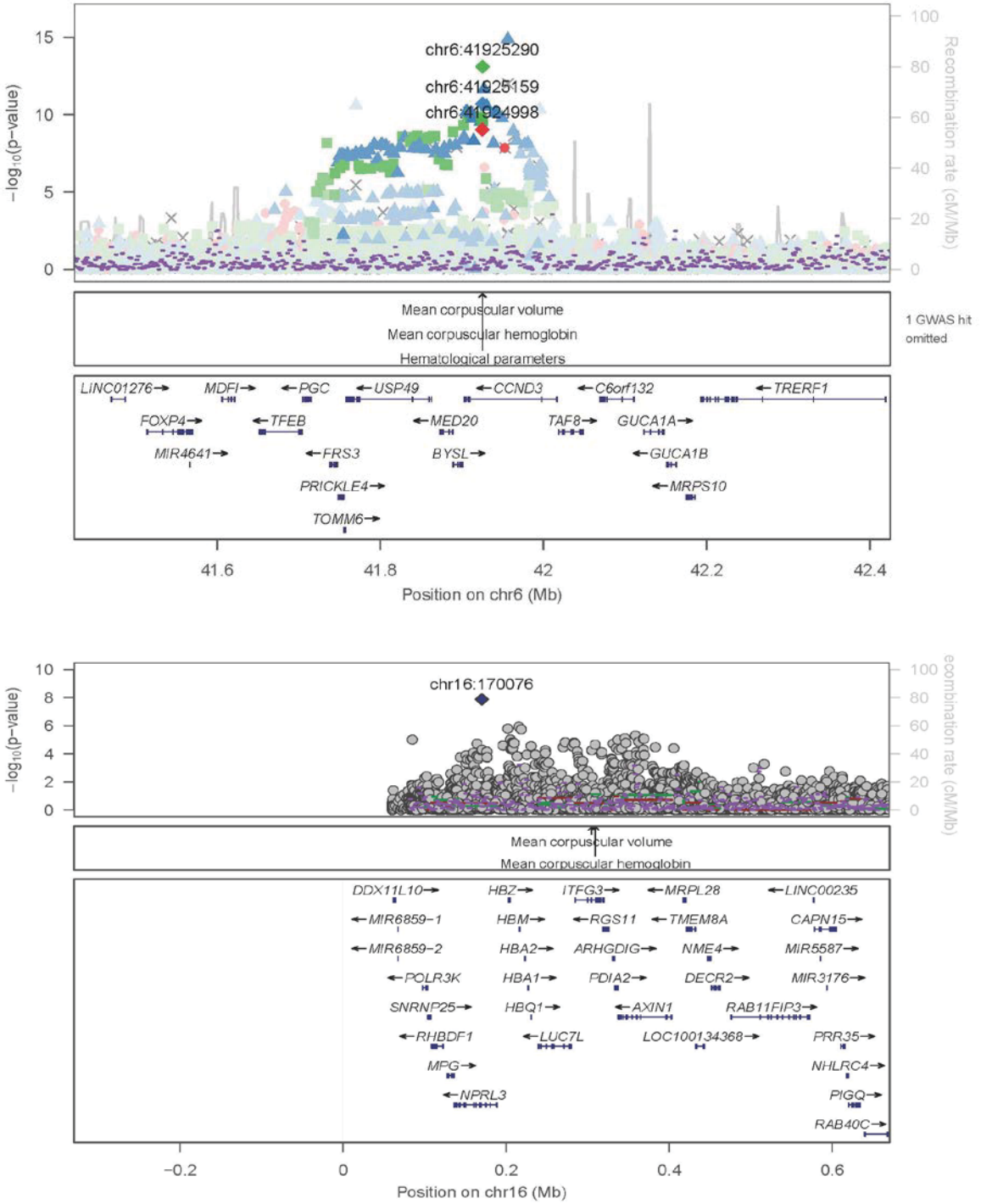
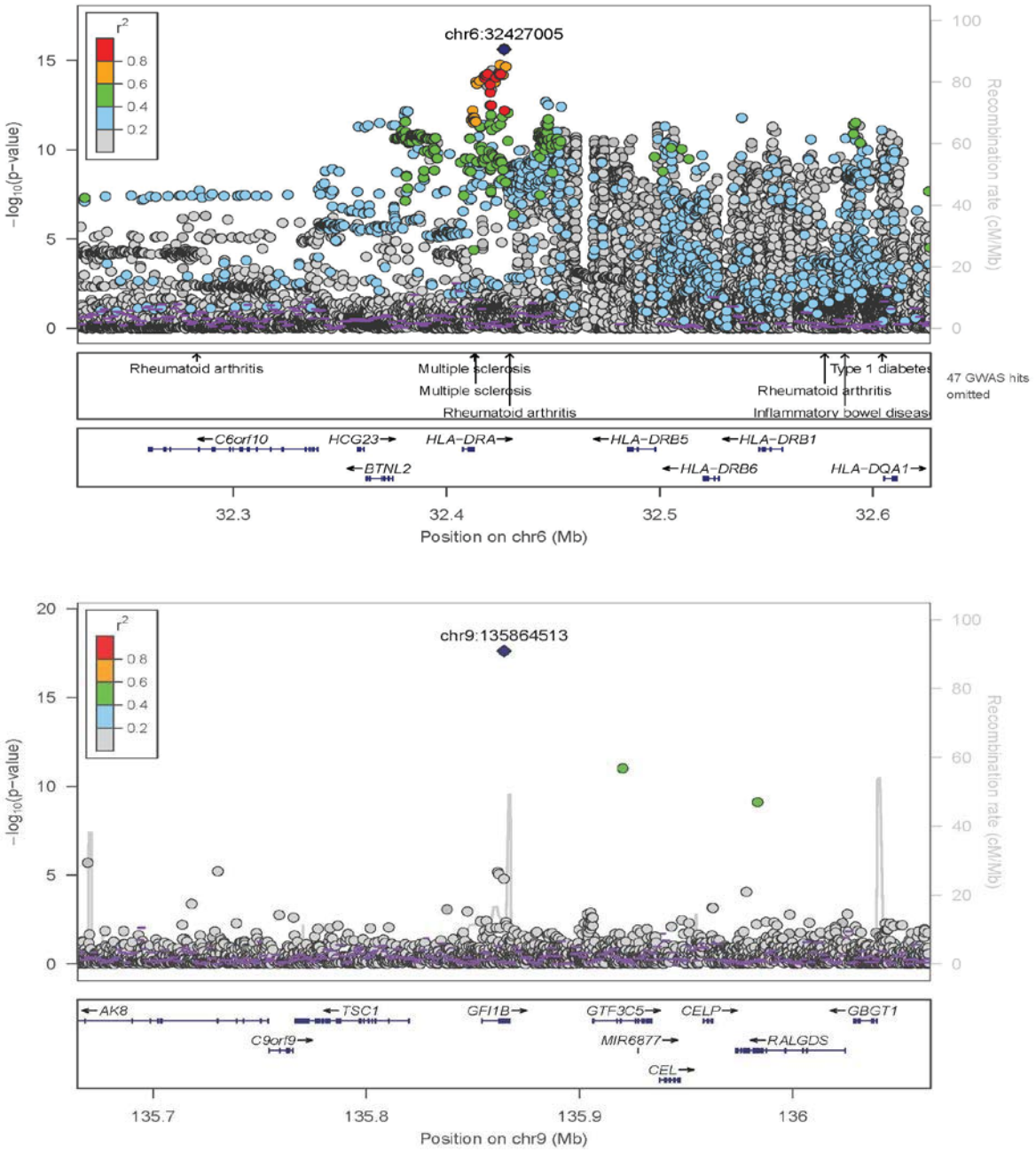


Figure 5.4 Regional plots of top hits from 18-way meta-analysis

The top plot is for the HLA locus for association with WBC, and the bottom plot is for GFI1B locus for association with PLT.



5.3.2 Fine mapping of known and novel loci

The availability of WGS compared on GWAS based on sparse datasets allows one to evaluate statistically the plausibility of each variant in an association signal to be causally associated with a trait. To fine-map FBC associated regions, I implemented the method of Maller et al. (Maller et al. 2012), as described in chapter 2 and the Methods section above. For seven known loci, there are sufficient resolution to limit the number of possible causal variants to a small informative set (log10BF>5 and # of variants <20) (**Table 5.9**). There are a total of 22 putative causal variants in these seven loci, three of which are previously reported known variants. Based on Regulome database, rs115740542 has the strongest evidence for functionality, with a score of "1a" (supporting evidence from TF binding, matched TF motif, matched DNase footprint, DNase peak), while rs198851 and rs12005199 have modest evidence for functionality (supporting evidence from TF binding, any motif, DNase footprint, DNase peak). The rest variants all have a score greater than 4, indicating weak support of functionality.

Table 5.9 Putative causal variants based on fine mapping

BP: Bayes factor, PP: posterior probability. Three previously reported known variant are labelled with *.

		Fine-mapping					WGS					Meta-analysis					
trait	**region**	**rsID**	**CHRPOS**	**GWAVA**	**log10(BF)**	**PP**	**EA**	**EAF**	**BETA**	**SE**	**P**	**EA**	**EAF**	**beta**	**se**	**P**	**N**
PLT	ARHGEF	rs1354034	chr3:56849749 *	Intron	9.89	1.00	--	--	--	--	--	T	0.421	-0.104	0.012	1.38E-17	15328
MCH	HFE, chr6:25343245-26589359	rs80215559	chr6:25918225	Intron	19.94	0.03	C	0.069	0.352	0.072	9.66E-07	T	0.941	-0.319	0.032	4.02E-23	12190
		rs1800562	chr6:26093141	Missense	20.25	0.07	A	0.070	0.341	0.070	1.32E-06	G	0.939	-0.316	0.031	3.30E-24	12190
		rs79220007	chr6:26098474	3_prime_UTR	20.48	0.12	C	0.069	0.338	0.070	1.71E-06	T	0.940	-0.318	0.031	3.31E-24	12189
		rs115740542	chr6:26123502	Upstream	21.28	0.74	C	0.067	0.339	0.072	2.49E-06	T	0.941	-0.323	0.032	3.00E-24	12190
MCV	HFE chr6:25600233-26589359	rs1799945	chr6:26091179	Missense	12.51	0.39	G	0.140	0.173	0.053	1.09E-03	C	0.846	-0.162	0.018	5.88E-20	15280
		rs2032451	chr6:26092170	Upstream	11.38	0.03	T	0.142	0.175	0.053	8.67E-04	G	0.845	-0.156	0.018	1.04E-18	15279
		rs1800562	chr6:26093141 *	Missense	11.30	0.02	A	0.070	0.263	0.070	1.98E-04	G	0.942	-0.238	0.028	1.44E-17	15281
		rs79220007	chr6:26098474	3_prime_UTR	11.35	0.03	C	0.069	0.259	0.071	2.57E-04	T	0.943	-0.239	0.028	1.38E-17	15278
		rs198851	chr6:26104632	Downstream	12.50	0.38	G	0.859	-0.171	0.053	1.19E-03	T	0.153	0.163	0.018	7.00E-20	15281
		rs198846	chr6:26107463	Downstream	11.59	0.05	G	0.853	-0.167	0.052	1.27E-03	A	0.158	0.156	0.017	5.71E-19	15279
		rs198833	chr6:26114508	Downstream	11.46	0.04	A	0.854	-0.165	0.052	1.44E-03	G	0.158	0.155	0.017	1.03E-18	15280
		rs115740542	chr6:26123502	Upstream	11.41	0.03	C	0.067	0.258	0.072	3.61E-04	T	0.945	-0.240	0.028	6.20E-17	15281
PLT	intergenic chr9:4740135-4903034	rs385893	chr9:4763176 *	Regulatory	6.11	0.03	C	0.511	0.091	0.036	1.23E-02	T	0.493	-0.081	0.012	2.02E-11	15328
		rs12005199	chr9:4763491	Regulatory	7.68	0.94	A	0.291	0.134	0.039	5.81E-04	G	0.727	-0.105	0.014	8.40E-14	15328
MCH	OR52A1, chr11:4810830-5765688	chr11:5042074	chr11:5042074	Downstream	18.06	0.31	A	0.001	-0.146	0.710	8.37E-01	A	0.003	-1.846	0.200	5.04E-20	4568
		rs181392259	chr11:5054906	Upstream	17.20	0.04	T	0.004	-0.017	0.315	9.57E-01	T	0.003	-0.894	0.140	1.93E-10	11716
		chr11:5126515	chr11:5126515	--	17.97	0.25	--	--	--	--	--	T	0.997	1.675	0.189	1.15E-18	8623
		chr11:5180087	chr11:5180087	Intron	18.08	0.33	T	0.001	-0.138	0.710	8.46E-01	T	0.003	-1.868	0.203	5.20E-20	4568
		rs183952362	chr11:5196364	Upstream	17.13	0.04	G	0.004	-0.367	0.319	2.51E-01	G	0.004	-0.636	0.122	2.25E-07	11717
MCH	RAB11FIP3, chr16:442805-602595	rs143109032	chr16:536959	Upstream	7.31	1.00	T	0.005	0.394	0.279	1.58E-01	C	0.994	0.221	0.103	3.26E-02	12189
MCH	TMPRSS6 chr22:37366826-37510072	rs855791	chr22:37462936 *	Missense	26.38	0.98	G	0.555	0.182	0.036	4.02E-07	A	0.444	-0.161	0.014	7.62E-29	12190

5.3.3 Novel loci based on rare variants aggregation test

The above are for single marker base tests, which has limited power to detect associations for low frequency and rare variants given the current number of samples with WGS. Here I show association results based on rare variants aggregation tests. As stated in the Methods, there types of SKAT-O analyses were run: genome-wide sliding window, exome-wide gene based, and exome-wide with only functional variants. Overall, the statistics of these tests follow the expected distribution assuming a NULL association, and there is a lack of signals meeting pre-defined genome-wide significance threshold (**Figure 5.5**). Nevertheless, there are six regions that meet our pre-defined significance threshold for follow-up (P<6.8E-08 for genome-wide SKAT-O, P<1.2E-06 for exome-wide SKAT-O, P<1.0E-05 for functional variants SKAT-O) (**Table 5.10)**. For three of these loci, the SKAT P value is much less significant than the SKAT-O P value, indicating that the signals are mainly driven by burden tests. Although independent replication is needed to confirm the rare variants aggregation based association with these six regions, the *RHBDL2* locus for association with PLT is a biologically plausible. It was reported that *RHBDL2* and thrombomodulin have important roles in wound healing via the release of soluble *RHBDL2* from keratinocytes and that may function as an autocrine/paracrine signal promoting wound healing (Cheng et al. 2011). The most strongly associated variant based on WGS data alone in this locus is chr1:39384826 (MAF=0.008, P=2.21E-05) (**Figure 5.6**). However, this variant is not significant in the 10-way meta-analysis (P<0.05). This locus also harbours a variant (rs4246511, chr1:39380385) previously reported for associated with menopause age at onset (Stolk et al. 2012). However, the rare variants based association for this region needs to be validated and replicated to drive further interpretation on this locus.

Figure 5.5 Rare variants aggregation test results for FBC traits

There are eight rows, each row for one of the eight traits as indicated in the plot title. The genome-wide significant signals are shown in red, with threshold of $P < 6.8E-08$, 1.2E-06, 1E-05 respectively for genome-wide, exome-wide, and functional variants based SKAT-O. Suggestive signals are shown in blue, with threshold of $P < 1E-05$, 1E-04, 1E-04 respectively for genome-wide, exome-wide, and functional variants based SKAT-O.

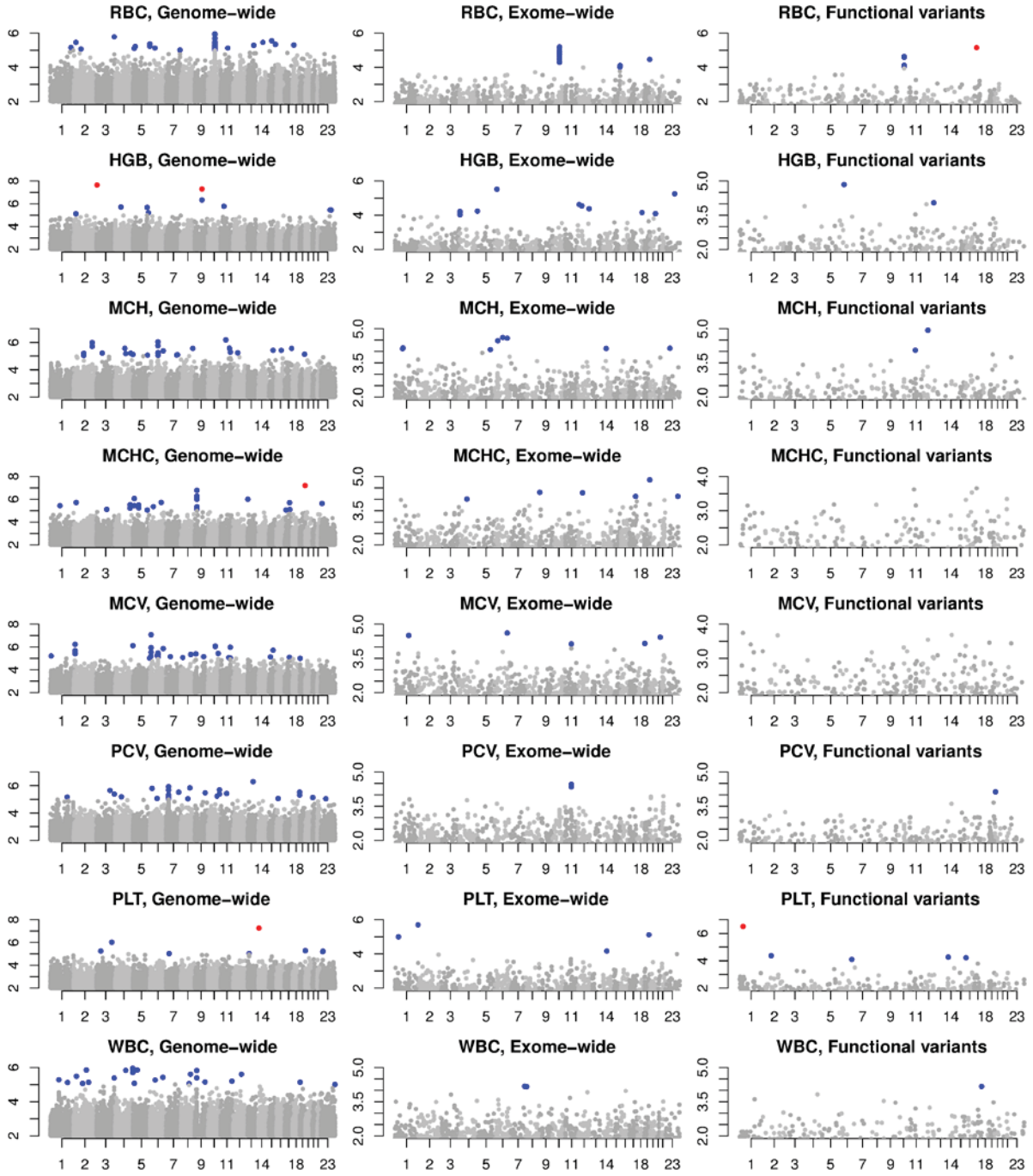


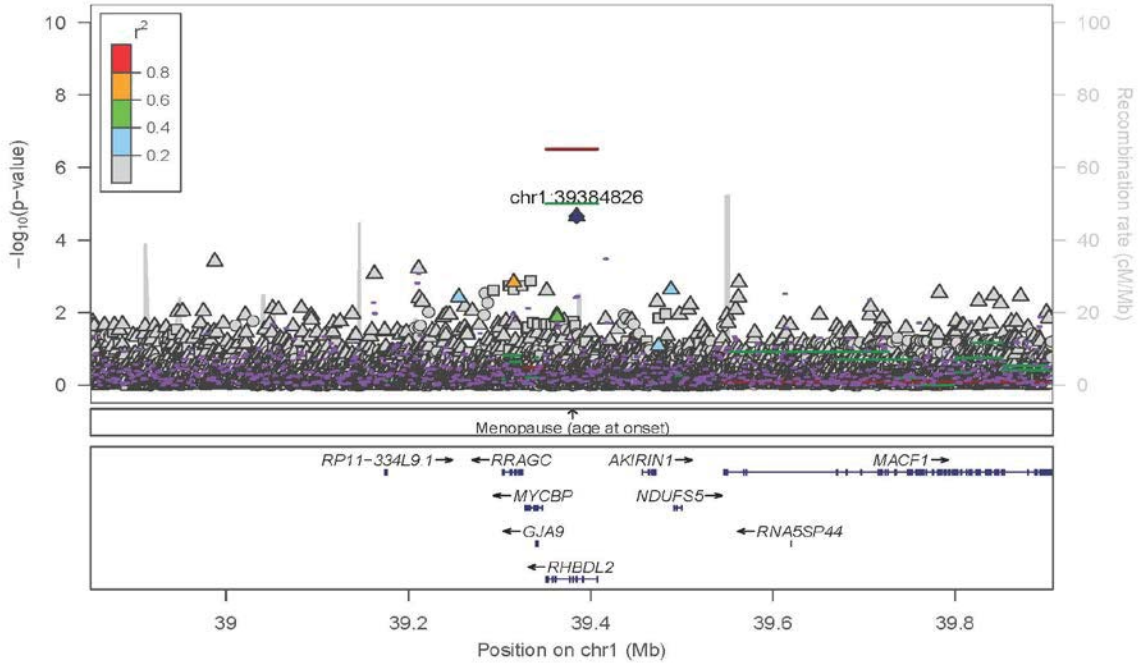
Table 5.10 Rare variants aggregation tests based top hits for FBC traits

For the three locus marked with *, the SKAT *P* is much less significant than the SKAT-O *P*, indicating that the signals are mainly driven by burden tests.

trait	**Type**	**locus**	**chr**	**start**	**End**	**TwinsUK**	**ALSPAC**	**SKAT**	**TwinsUK**	**ALSPAC**	**SKAT-O**
PLT	Functional variants	RHBDL2	1	39,351,479	39,407,471	7.52E-06	--	7.52E-06	3.11E-07	--	3.11E-07
HGB *	Genome-wide	GRM7	3	6,463,501	6,466,500	1.02E-01	3.09E-03	5.40E-04	1.75E-01	6.35E-03	2.28E-08
HGB	Genome-wide	OSTF1	9	77,787,001	77,790,000	1.74E-05	3.73E-04	1.68E-08	4.49E-05	6.24E-04	5.03E-08
PLT *	Genome-wide	DHRS4	14	24,462,001	24,465,000	1.01E-04	--	1.01E-04	5.48E-08	--	5.48E-08
RBC	Functional variants	PIGS	17	26,880,401	26,898,890	8.09E-05	--	8.09E-05	6.99E-06	--	6.99E-06
MCHC *	Genome-wide	ZSCAN5A	19	56,883,001	56,886,000	1.01E-04	--	1.20E-05	6.26E-08	--	6.26E-08

Figure 5.6 Regional plots of *RHBDL2*

The single marker results are based on TwinsUK WGS. The horizontal dashed lines are SKAT-O, purple, green, red for genome-wide SKAT-O, exome-wide SKAT-O, and functional variants based SKAT-O respectively.



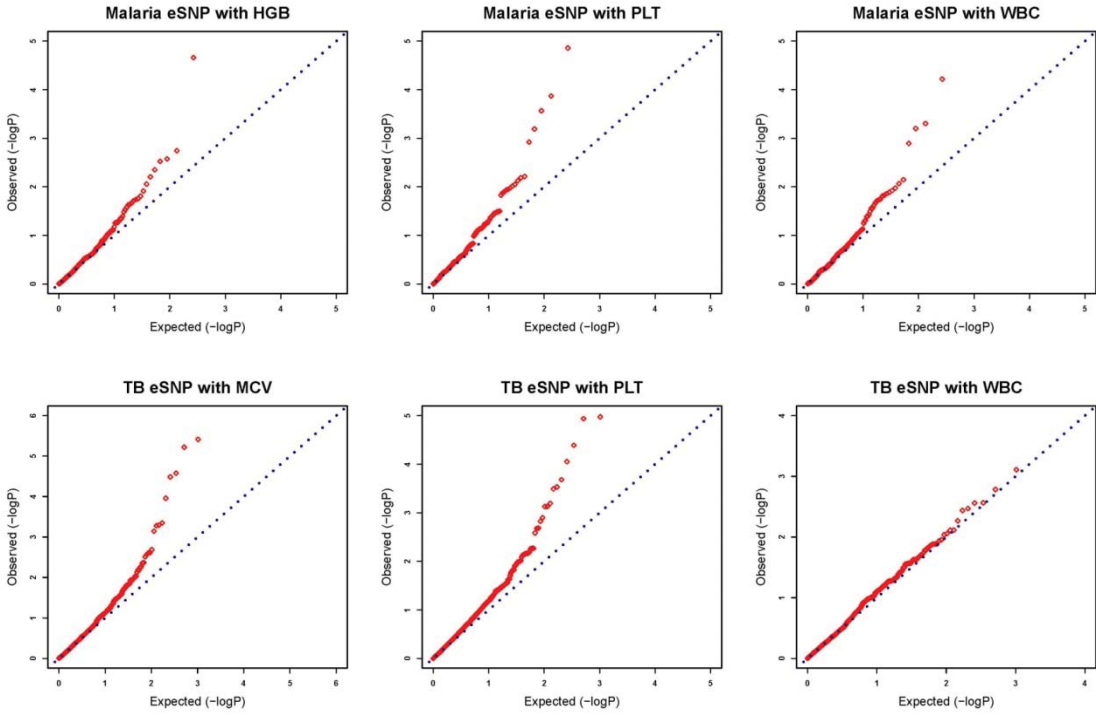
5.3.4 Host-response eQTL

Given the role of blood parameters in the host response to bacterial infection, I explored an approach to show whether genetic variants associated with host-response are enriched for association with FBC traits. I included 268 eSNP associated with gene expression of host response to malaria infection (Idaghdour et al. 2012) and 40 loci implied for response to severe malaria (Timmann et al. 2012). For tuberculosis, I used 1,046 eSNPs (720 for infected, 756 for unaffected) (Barreiro et al. 2012). Given the overall low number of variants tested, I did not perform a formal enrichment test, but used QQ plots to see whether the SNPs associated with host-response follow a NULL distribution for association with FBC traits. As shown in **Figure 5.7**, both HGB and PLT are enriched for eSNPs associated with host response to both Malaria and TB. WBC is enriched for eSNP associated with host response to Malaria but not to TB.

It is well established that genetic loci associated with resistance to malaria (for example, HBB, HBA1/HBA2, and G6PD) are associated with RBC traits (Ding et al. 2013). This is consistent with the fact that the malaria parasites grow in the human red cells. In 2012, a research team at Duke University discovered that human microRNA found in sickle red cells directly participate in the gene regulation of malaria parasites (LaMonte et al. 2012). The study showed that when two different microRNAs were introduced at higher levels in normal red cells, the parasite growth also was decreased. Another surprise in this investigation was the presence of a chimera, a fusion of human microRNA with the parasites' mRNAs, which represents a unique form of host-parasite interaction. This may reflect either a novel form of host-cell immunity or a mechanism by which the parasite is able to adapt to the host-cell environment. Although WBC changes during infections to TB, there was no reported evidence that the genetic loci associated with TB resistance is also associated with WBC. Similarly, platelet phagocytosis may contribute to thrombocytopenia found in vivax malaria (Coelho et al. 2013), but the preliminary data presented in **Figure 5.7** is the first to imply that genetics is involved between the phenotypic variation of FBC traits and the host response to infection of malaria and TB.

Figure 5.7 eSNPs associated with host response to TB and Malaria

Y-axis is the observed *P* value of eSNPs previously reported for association with Malaria (the first row) and TB (the second row), for association with HGB (first column), PLT (second column), WBC (third column). These *P* values are from the 12-way meta-analysis. The X-axis is the expected *P* value under the NULL hypothesis of no association.



5.4 Conclusion & Discussion

5.4.1 Summary of main findings

So far, there are no reported studies on FBC that used WGS data. With a modest WGS sample size (N=1,497), I identified three putative novel variants, but they were not replicated based on a few imputed datasets made available for replication. A total of 25 variants with MAF between 0.5% and 5% have *P* <1e-06 based on TwinsUK WGS, but replication is needed to establish any of these signals. Nevertheless, the association of rs115740542 within *NPRL3* with MCH was already supported by epigenomic annotation. To boost study power, I included a total of 12 cohorts for discovery and 6 cohorts for replication. I further conducted a meta-analysis with all 18 cohorts with a sample size up to 41,557. Based on the 12-way meta-analysis, a total of nine novel loci and three novel variants within known loci were discovered at a pre-defined *P*<1E-07. However, replication data is only available for two of these variants with non-replicated results. Based on the 18-way meta-analysis, there are two strong associations: the *HLA* locus for association with WBC, and the *GFI1B* locus for association with PLT. Given the function of these two regions for the according phenotypes and given the strength of the association signals, there two associations are most likely to be true and deserve further investigation. Fine-mapping analysis identified one SNP rs115740542 within *HFE* to be highly likely causal, with supporting evidence of functional data (RegulomeDB). By running a systematic enrichment analysis, I observed that hematological traits associated SNVs are significantly enriched in key epigenomic features including chromatin state, histone modification, and TFBS. Through rare variant aggregation analysis, I discovered that the aggregated functional variants in *RHBDL2* are strongly associated with PLT, which is biologically plausible.

5.4.2 Interpretation of results

The single marker association testing of eight lipids follows closely the expected relationship between EAF and effect size (beta) as dictated by study power (Park et al. 2011),

as shown in **Figure 5.8**. Given the relatively small sample size and yet the encouraging finding of two strong signals based on the 18-way analysis, more truly novel associations are expected to be found with larger sample sizes.

GWAS on FBC traits has already brought translational outcome. As we know, the β-globin gene (*HBB*) is silent prior to birth and the β-globin subunits are encoded by the γ-globin gene (*HBG1* and *HBG2*) to form fetal hemoglobin (HbF). The switch from HbF to HbA production is a transcriptionally and epigenetically tightly regulated process (Sankaran et al. 2010). The association of *BCL11A* with HbF levels were first reported through GWAS (Menzel et al. 2007, Uda et al. 2008). Later on, *BCL11A* was found to be a potent transcriptional repressor of γ-globin gene expression and that its inactivation in the erythroid lineage can treat sickle cell disease in mouse model through re-activation of HbF production (Sankaran et al. 2008, Xu et al. 2011). This model was confirmed by targeted deletion of the enhancer through genome engineering that blocked *BCL11A* expression and re-activated γ-globin gene expression and HbF production (Sankaran et al. 2012). As genome editing methods are rapidly improving, this proof-of-concept experiment suggests a new therapeutic strategy for β-thalassemia and sickle cell diseases with mutations in *HBB* (Bauer and Orkin 2011, Hardison and Blobel 2013).

5.4.3 Future direction

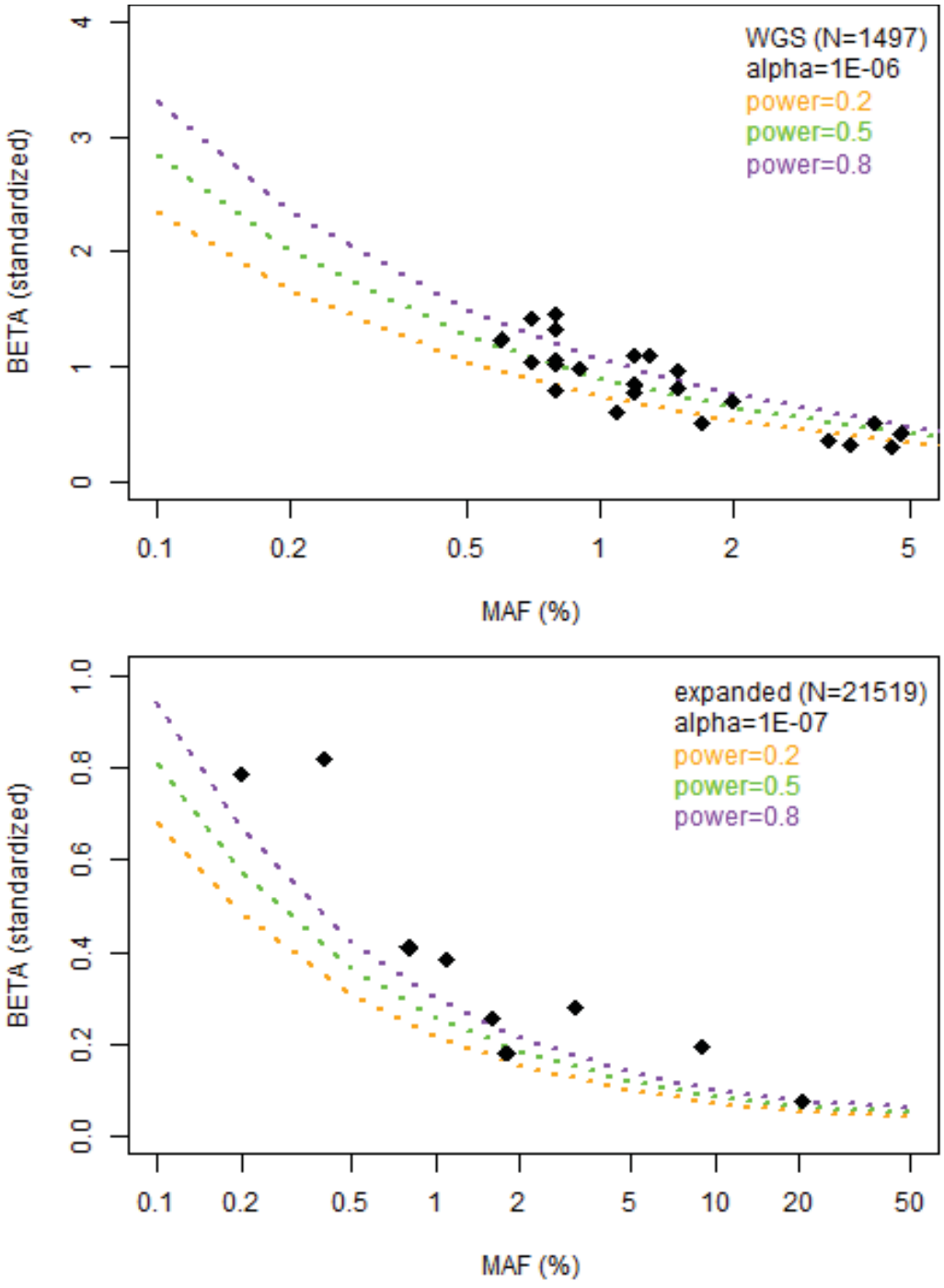
Compared to many other complex traits, future larger studies on FBC traits with WGS dataset might be more achievable given these traits are widely measured in clinical settings for evaluation health and diseases. FBC traits are also preferred phenotypes for the study the genetics of complex human diseases because they could be easily manipulated in vitro and discovered genes could be assessed in cell cultures and model organisms. It is not surprising that there is an overall lack of novel loci discovered given the sample size in the current study compared to previously conducted GWAS on these traits. The lack of loci for WBC could be also due to phenotype heterogeneity because the major populations of white blood cells (lymphocytes, granulocytes, monocytes) differ markedly in their roles and lifespans.

Besides increasing the number of samples of European ancestry, including samples of diverse ethnicity could also boost the genetic findings for FBC traits. For many complex traits, African samples have been used to fine map genetic loci discovered from European

samples, due to longer haplotypes in Africans. However, for FBC traits, sometimes a very strong genetic association in African population might not have any association in the European population. The associations of variants in *DARC* with WBC (Reich et al. 2009) and the association of variants in *HBA2* with RBC (Chen et al. 2013) are only observed in Africans while those variants are almost monomorphic in Europeans. The former variation protects against *Plasmodium vivax* while the latter protests against malaria infections, which are common in Africa. This study demonstrated similar phenomenon for genetic isolates. The signal on chromosome 11 is marginally significant in TwinsUK while strongly significant in HELIC-Pomak. Also, the signal on chromosome 21 (chr21:14589985 for association with PLT) mainly came from an Italian isolate: INGI-FVG. Its MAF in TwinsUK is ~1%, but ~4% in two Greek isolates, and ~33% in INGI-FVG. Once these are confirmed to be true signals in the general population, the use of genetic isolates would be proven valuable for identifying these associations, which would otherwise require a much larger sample size for detection of the association.

Figure 5.8 Statistical power and novel variants from single marker analysis

The top and bottom plots are for WGS samples and expanded discovery samples respectively. Y-axis is a variant's effect, expressed in standard deviation units. X-axis is MAF of effect alleles. Colored lines indicate 20%, 50%, and 80% power. Alpha is set at *P*<1E-06 for WGS and *P*<1E-07 for expanded discovery respectively. The 25 putative novel WGS variants are shown in the top power plot for WGS, and the nine putative novel variants from expanded discovery are shown in the bottom power plot for expanded discovery.



6 CRP

6.1 An introduction on CRP

6.1.1 Biology and physiology of circulating CRP

CRP is an acute-phase protein of hepatic origin, composed of five 23-kDa subunits. CRP is synthesized by the liver in response to factors released by macrophages and adipocytes (Pepys and Hirschfield 2003), and its level increases following interleukin-6 secretion from macrophages and T cells. CRP was named so because it was first identified as a substance in the serum of patients with acute inflammation that reacted with the C-polysaccharide of Pneumococcus. It binds to lysophosphatidylcholine expressed on the surface of dead or dying cells and some types of bacteria in order to activate the complement system (Thompson et al. 1999).

CRP is associated with multiple aspects of atherosclerosis such as adhesion molecule expression, effects on fibrinolysis and alteration of endothelial function (Szmitko et al. 2003). Both lipid crystals and the infiltration of inflammatory cells are characteristic features of atherosclerosis that can be detected at the earliest stages of plaque development. Inflammatory mechanisms couple dyslipidaemia to atheroma formation. Crystalline cholesterol acts as an endogenous danger signal and its deposition in arteries or elsewhere is an early cause rather than a late consequence of inflammation (Duewell et al. 2010). Leukocyte recruitment and expression of pro-inflammatory cytokines characterize early atherogenesis. Moreover, inflammatory pathways promote thrombosis, a late and dreaded complication of atherosclerosis responsible for MI and most strokes. Identifying the triggers for inflammation and unravelling the details of inflammatory pathways may eventually furnish new therapeutic targets (Libby 2002).

A few inflammatory biomarkers have been studied for their potential link and role in atherosclerosis, with CRP being the one most widely studied and having strongest evidence

for added value to CVD risk prediction (Koenig et al. 2004, Cushman et al. 2005). The other inflammatory biomarkers include interleukin-6 (IL-6) (Ridker et al. 2000), and lipoprotein-associated phospholipase A2 (Lp-PLA2) (Persson et al. 2007), P-selectin (Ridker et al. 2001), tumour necrosis factor alpha (TNF-α), the inter-cellular adhesion molecule 1 (ICAM-1) and vascular cell adhesion molecule 1 (VCAM-1) (Malik et al. 2001) are also associated with CVD risk; however, these markers are less stable than CRP and hence are less reliable indicators.

6.1.2 CRP as risk factors for CVD

In 1994, CRP was first reported to predict a poor outcome in patients with unstable angina (Liuzzo et al. 1994). In 1997, CRP was deemed a plausible risk factor when Ridker and colleagues reported that baseline CRP predicted the risk of future MI and stroke (Ridker et al. 1997). But CRP is a relatively moderate predictor of CHD compared to established risk factors including lipids level and blood pressure (Danesh et al. 2004). Several epidemiological studies have shown that the addition of CRP to traditional risk factors only raises the *c* statistic by less than 0.015 (Folsom et al. 2006, Lloyd-Jones et al. 2006, Melander et al. 2009). Therefore CRP assessment would only have a small effect on treatment decisions (Boekholdt and Kastelein 2010).

The Emerging Risk Factors Collaboration (ERFC) conducted by far the largest epidemiological study on CRP, which combined dataset on more than 160,000 subjects and comprises 1.31 million person-years at risk and ~28,000 fatal or non-fatal disease outcomes (Emerging Risk Factors et al. 2010). This study reported that CRP concentrations were associated with the risk of CVD (including CHD, ischaemic stroke, vascular mortality, and non-vascular mortality), most established CVD risk factors, and other inflammatory markers. One year later, another study reported that CRP were associated with an increased CHD risk, after adjusting for more variables including waist circumference, physical activity, smoking, diabetes, SBP, HDL and LDL, hormone replacement therapy in women (Rana et al. 2011). CRP is also confounded by other non-established CVD risk factors. For example, the age-related variation in CRP and IL-6 is largely explained by differences in visceral adipose tissue (Cartier et al. 2009).

As heart-healthy diets, weight loss, and physical activity all reduce CRP levels as well as other CVD risk factors, the AHA/CDC guidelines suggest that a finding of elevated CRP can be used to reinforce basic messages for lifestyle change. Statin therapy may be effective in the primary prevention of coronary events among those with relatively low lipid levels but with elevated levels of CRP (Ridker et al. 2001) (Ridker et al. 2008) (Ridker et al. 2009), therefore making CRP an attractive biomarker to identify patients who are likely to benefit from statin therapy. The guidelines on CVD prevention (Expert Panel on Detection and Treatment of High Blood Cholesterol in 2001) recommend that individuals at high risk should be treated whereas additional information is needed for those at intermediate risk. Some reported that CRP level could fine-tune the choice of treatment for those predicted with intermediate risk (Pearson et al. 2003, Ridker et al. 2008, Rana et al. 2009) (Koenig et al. 2004), but this is not supported in other studies including the FHS (Wilson et al. 2005, Sattar et al. 2007).

CRP is not an established risk factor, because MR studies did not establish its causal role to CVD, by using a single cis- variant (Casas et al. 2006, Lawlor et al. 2008) or multiple cis- variants (Elliott et al. 2009) as instrumental variables. The causality of CRP to CVD is complicated by the fact that CRP is also synthesised in smooth muscle cells within diseased atherosclerotic arteries. Inflammation may play a causal role via upstream effectors rather than the downstream marker of CRP. Certain factors more proximal in regulation of CRP could play a causal role (Brunner et al. 2008, Elliott et al. 2009), and such connections have been established for two other inflammatory biomarkers (IL6 for CVD, and IL1 for T2DM).

6.1.3 Genetic determinants of CRP

The heritability of serum CRP level is up to 52% (MacGregor et al. 2004), providing a strong case for discovering genetic determinants of CRP.

Findings from candidate gene and linkage analysis

In 2008, a linkage study was performed on a few inflammatory biomarkers including CRP, IL-6, and TNF-α in 764 subjects enrolled in the Quebec family study (Ruchat et al. 2008). The reported linkage signal was very modest and none remained significant after

adjustment for body mass index. The result suggested that several QTLs influence plasma levels of CRP partly via their effects on adiposity.

Findings from first generation GWAS

Since 2008, a total of 14 GWAS have been performed to discover genetic variants for association with CRP (**Table 6.1**). In 2008, the first large-scale GWAS scan on CRP led to a discovery of seven loci (*LEPR*, *CRP*, *IL6R*, *GCKR*, 12q23.2, *HNF1A*, *APOE*) (Ridker et al. 2008). The protein products for six of these loci are directly involved in metabolic syndrome, insulin resistance, beta cell function, weight homeostasis, and premature atherothrombosis. The largest GWAS of CRP was performed in 2011. It included more than 80,000 subjects of European ancestry and identified 11 novel loci (Dehghan et al. 2011). These loci are related to metabolic syndrome, immune system, and pathways previously unknown for chronic inflammation. GWAS on CRP was also conducted on non-European population, where novel findings included *TREM2* in African-American females (Reiner et al. 2012) and *IL6* (rs2097677) in Japanese individuals (Okada et al. 2011). These studies have been focusing on common variants genotyped on SNP arrays with imputation based on HapMap reference panel.

Findings from next generation sequencing

So far, there is no reported study on CRP that used high-throughput next generation sequencing technologies.

Table 6.1 GWAS studies of CRP

Date is for publication date. Samples are all European ancestry unless explicitly specified otherwise: KOR for Korean, FIL for Filipino, AA for African American, HIS for Hispanics, SAR for Sardinian, ASN for Asian. The sample size before "+" is for discovery while the sample size after "+" is for replication.

Date	Samples	Main findings	References
2008-04	6,345	4 loci (*LEPR*,*HNF1A*, *IL6R*, *GCKR*)	(Ridker et al. 2008)
2008-04	909+5,106	*HNF1A* intron 1	(Reiner et al. 2008)
2009-07	17,967+13,615	5 loci, no causal role of CRP for CHD	(Elliott et al. 2009)
2010-12	10,112+2,742 JAP	pleiotropic associations in IL6 gene	(Okada et al. 2011)
2011-02	66,185+16,540	7 known and 11 novel loci	(Dehghan et al. 2011)
2011-06	1,709 FIL	Interaction of CRP and HNF1A	(Wu et al. 2012)
2012-01	1,092	changes in response to fenofibrate treatment	(Aslibekyan et al. 2012)
2012-01	4,694 + 1392 SAR	3 novel loci	(Naitza et al. 2012)
2012-04	837AA	EA signals transferable to AA, AA data can fine-map of EA signal.	(Doumatey et al. 2012)
2012-07	8,842 KOR	CRP and WBC have distinct genetic components	(Kong and Lee 2013)
2012-08	8,280 AA 3,548 HIS	a common TREM2 variant	(Reiner et al. 2012)
2013-07	~7,500 ASN	EA variants are also detected in Asian	(Dorajoo et al. 2013)
2014-03	7,570 AA	a novel locus in CD36	(Ellis et al. 2014)
2014-04	7627 + 903 KOR	A novel variants in the ARG1	(Vinayagamoorthy et al. 2014)

6.1.4 Aims of this study

Under the framework of the UK10K project (The UK10K Consortium 2015), this study aimed to identify novel genetic variants that are associated with serum CRP levels and also fine map known CRP loci with WGS data. The current study is by far the largest WGS based association study of CRP, with 2,046 WGS samples and more than 32,000 samples with WGS imputed data. I first analysed the WGS samples aiming to discover rare and low frequency variants with large effect sizes. Then I analysed a much larger group of cohorts with imputed data to discover novel associations across the full MAF spectrum. Besides single marker based genome-wide scan, this study was able to fine map known loci and investigate the association and contribution of rare variants to serum lipids variance.

6.2 Methods

6.2.1 Cohorts & phenotype measurements

Like lipids, CRP was measured in both TwinsUK and ALSPAC. For WGS based analysis, I conducted a 2-way meta-analysis. For expanded discovery analysis, I included an additional 14 cohorts while used a single TwinsUK dataset as I did for FBC traits analyses. This leads to a 15-way expanded discovery analysis. The six WHI cohorts used for replication for FBC traits were also used for replication for CRP. But for CRP, I obtained another six cohorts as stage-2 replication (**Table 6.2**).

CRP was measured by high-sensitivity immunology assay in all participating cohorts. CRP measurement methods are as following: for **ALSPAC**, CRP was measured by Latex enhanced assay; for **TwinsUK**, CRP was measured by automated particle-enhanced immunoturbidimetric assay (Roche UK, Welwyn Garden City, UK); for **1958BC**, CRP antigen levels were measured by high sensitivity nephelometric assay using latex particles coated with monoclonal antibodies to human CRP in the BN Prospec protein analyzer (Dade Behring, Marburg, Germany). For **HELIC-MANOLIS** and **HELIC-Pomak**, CRP was measured using an immunoturbidimetric assay on a COBAS 8000 analyser (Roche). For

WHI, CRP was measured using a latex-particle enhanced immunoturbidimetric assay kit (Roche Diagnostics, Indianapolis, IN). For **FHS** and the rest of discovery cohorts, CRP was measured in fasting serum samples using various versions of high-sensitivity assay, mostly the Dade Behring BN100.

For phenotype harmonizaiton, I first excluded abnormal values of CRP, defined as <0.1mg/L or >10mg/L. For TwinsUK, the phenotype harmonization was conducted for WGS and GWA samples separately. Inverse normal transformation was applied to the full dataset without gender specific transformation. Regression test found no significant effects of dates of visits or analysers. BMI was not included as a covariate. For each trait of each cohort, the residuals with confounding variables regressed out were standardized so that the phenotype has a mean of 0 and a standard deviation of 1.

6.2.2 Single marker based discovery and follow-up

To discover variants of low and rare frequency with big effect size, I first run genome-wide association for the TwinUK WGS and ALSPAC WGS. I used SNPTEST to fit linear models on standardised trait residuals to test associations of allele dosages with 13,074,236 SNVs and 1,122,542 biallelic InDels (MAF≥0.1%) in the two WGS samples, followed by a meta-analysis to produce the 2-way meta-analysis, which has a total sample size of 2,046 (**Table 6.2**). Variants with 2-way meta-analysis *P*<1E-6 are deemed of interest for follow-up and further characterization. For the expanded discovery meta-analysis, I used all TwinsUK samples as a single cohort, the same way as I did for the FBC traits, in order to bring the co-Twins into the analysis. There are a total of 15 cohorts included for this expanded discovery analysis with a total sample size of 32,624 (**Table 6.2**). For each individual cohort, SNPTEST was used for population based samples while GEMMA was used for genetic isolates and cohorts with family structure. A 15-way meta-analysis was conducted using GWAMA v2.1, assuming a fixed effect model and adjusting genomic control to the summary statistics for both input and output data. Given the poor imputation quality and weak statistical power for rare variants, I chose to exclude the variants that did not pass a low allele frequency threshold (MAF<0.1%). For imputed cohorts, the variants with INFO <0.4 were also excluded.

Given the availability of the genome-wide results for the six replication cohorts from WHI, I run a 21-way meta-analysis that included the 15 discovery cohorts and six replication cohorts, as what I did for four FBC traits. This time, I have an additional six cohorts for stage-2 replication. The total sample size is 32,624 for 15-way discovery, 12,868 for 6-way (WHI) replication, and 27,726 for stage-2 replication.

6.2.3 Rare variant aggregation based discovery and follow-up

To evaluate the aggregation effects of rare variants, I used SKAT-O to discover genomic regions that harbour rare variants with large efforts but those effects could be picked up by single marker based analysis. I first evaluated the associations of rare variants by considering genes as functional units of analysis. I applied two separate statistical models with different properties to rare variants (MAF<1%): SKAT and burden tests, both implemented in a unified software SKAT-O. As described in chapter 2, in *naïve* tests, all variants in exons, untranslated regions (UTRs) and essential splice sites were considered, and were given equal weight of being causal (50,214 windows for 35,709 genes, mean=35 variants, median=38 variants per window). In functional tests, only loss of function (LoF) and predicted functional variants were included (15,528 gene windows with ≥ 5 variants, mean=18, median=14 variants per gene). Finally, I run the locus-based analysis genome-wide in an agonistic fashion, by constructing ~1.8 million windows of 3 kb each, overlapping by half (median 35 SNVs/window, MAF<1%), assigning an equal weight to all variants. For CRP, there is no replication data available for rare variants aggregation based tests.

6.2.4 Fine-mapping of known loci

For a total of 37 previously established CRP loci, I carried out fine-mapping analysis to assess the probability of each variant being causal given other variants in the region. Within each signal I included SNPs in high LD (defined as all variants having $r^2 \geq 0.8$ with the most associated variants in the region). As described in chapter 2, I first created a list of fine-mapping regions based on HapMap estimates of recombination rates. I then analysed each

region separately for each of the 15 participating cohort using Bayesian linear additive models, by accounting for covariates as in the general single point association analyses. At the end, the resulting BFs for each variant were multiplied to obtain a joint BF measure of association, with the assumption that each cohort is 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.

Table 6.2 Characteristics of participating cohorts

All cohorts are population based, except for TwinsUK. Imputation was conducted with the 1000G and UK10K combined reference panel unless otherwise specified. For the expanded discovery analysis, "TwinUK WGS" were not included because it is already in "TwinsUK all".

	Cohort	N	Country	Age	% Female	CRP (mg/L)
Discovery	ALSPAC WGS	1,167	UK	10 (9-11)	50.3	1.01 (0.25)
	TwinsUK WGS	879	UK	56 (17-85)	100.0	1.42 (0.32)
	ALSPAC GWA	2,226	UK	10 (9-12)	49.2	0.78 (0.21)
	TwinkUK all	2,512	UK	50 (16-83)	97.3	0.94 (0.22)
	1958BC	4,910	UK	44 (44-44)	52	1.00 (0.24)
	FHS	6,320	Italy	49 (31-72)	53	0.62 (0.21)
	INGI-FVG	411	Italy	52 (18-92)	58.2	1.34 (0.13)
	INGI-VB	1,162	Italy	55 (18-102)	56.3	1.02 (0.31)
	HELIC-Manolis	1,093	Greece	62 (18-99)	57.2	1.33 (0.22)
	HELIC-Pomak	839	Greece	43 (13-87)	72.1	0.98 (0.16)
	INCIPE-1	807	Italy	60 (35-89)	54	0.79 (0.21)
	INCIPE-2	1,332	Italy	58 (26-95)	51	0.82 (0.17)
	LURIC-Ctrl	1,228	Germany	62 (18-92)	59.7	1.01 (0.20)
	LURIC-Case	1,202	Germany	61 (17-91)	60.8	1.65 (0.22)
	Procardis-Case	3,732	Sweden	43 (13-87)	49.1	1.54 (0.22)
	Procardis-Ctrl	3,683	Sweden	63 (51-78)	55.8	1.06 (0.19)
Replication stage 1	WHI-Garnet	3,388	US	65 (50-79)	100.0	0.88 (0.20)
	WHI-Gecco1	780	US	65 (50-79)	100.0	1.07 (0.30)
	WHI-Gecco2	1,072	US	65 (50-79)	100.0	0.87 (0.21)
	WHI-Hipfx	1,716	US	65 (50-79)	100.0	0.99 (0.26)
	WHI-Mopmap	721	US	65 (50-79)	100.0	1.31 (0.32)
	WHI-Whism	5,191	US	65 (50-79)	100.0	1.22 (0.19)
Replication stage 2	Rotterdam Study	5,455	Netherlands	69 (48-75)	41.2	0.91 (0.32)
	LOLIPOP-EWA	505	UK	56 (35-75)	26.8	0.77 (0.19)
	LOLIPOP-EWP	564	UK	55 (23-75)	13.1	0.89 (0.24)
	LOLIPOP-EW610	834	UK	56 (32-67)	0.0	0.94 (0.18)
	Fenland	8,178	UK	65 (47-77)	46.2	1.08 (0.30)
	Lifelines	12,190	Netherlands	NA	NA	NA

6.3 Results

6.3.1 Novel loci and novel variants from single marker analysis

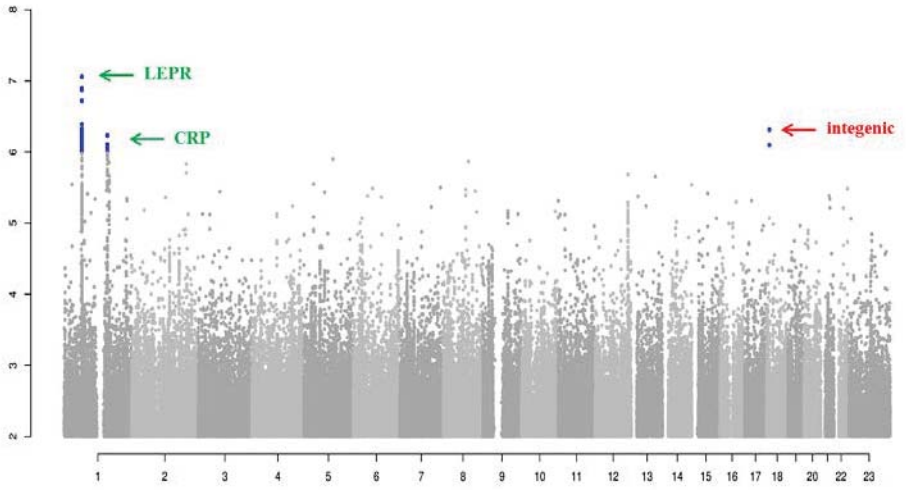
WGS for low frequency and rare variants

The assessment of associations based on imputation or WES has been incomplete. I thus sought to investigate if additional low-frequency or rare variants with strong effects could be detected from the WGS dataset. I first tested association results using solely the WGS dataset in order to identify whether these variants existed. Associations were carried out in 13,074,236 SNVs and 1,122,542 biallelic InDels (MAF≥0.1%) using linear regression and data from the two WGS cohorts was meta-analysed.

There are a total of 61 variants from UK10K WGS that have *P*<1E-6, but none of these reached *P*<5.0E-08 (**Figure 6.1**). 59 of these are common variants within the well-established *LEPR* and *CRP* loci, while the other two have low frequency (MAF=0.02) in an intergenic region on chromosome 18. These two variants are in high LD (r2=0.97). The first one is rs112734184 (chr18:10441718, EA=G, EAF=0.022, beta=-0.540, *P*=4.94E-07) and the second one is rs112155044 (chr18:10445499, EA=T, EAF=0.022, beta=-0.524, *P*= 8.06E-07). These two variants are non-significant in the large meta-analysis with 15 cohorts and ~45,000 samples (*P*>0.05), as described later. Therefore, they are most likely to be specific to the two UK10K cohorts or false positive.

Figure 6.1 Association Results of CRP based on WGS samples

X-axis is for chromosome and positions (build 37). Y-axis is for –log10(*P*). Variants passing threshold of 5E-08 and 1E-06 are shown in red and blue, respectively. For those passing threshold of 5E-08, known loci were marked in green text while putative novel loci were marked in red text.



Meta-analysis for identifying novel variants of all allele spectrums

Given the enhanced imputation quality with the UK10K WGS reference panel as demonstrated in chapter 3, I included an additional of 13 cohorts with imputed data for an expanded discovery, to increase power for discover variants across all allele frequency spectrum. As mentioned earlier in the methods section, variants with MAF <0.1% or imputation INFO <0.4 were not included. This effort yielded a total of 1,303 variants with *P*<1E-07, six of which are deemed novel after conditional analysis and LD pruning with positive controls (**Figure 6.2**, **Table 6.3**). Initially, six European cohorts from the Women's Genome Initiative (WHI) were made available for *in-silico* replication, but none of the six variants from the 15-way discovery were replicated. I then run a meta-analysis including all 15 discovery cohorts and six replication cohorts in a 21-way meta-analysis, where two novel variants passed the genome-wide significant threshold of 5E-08. These two variants are listed at the bottom of **Table 6.3**. The individual cohort results for these variants are presented in **Table 6.4**.

I took forward these eight variants into a stage 2 replication with six independent cohorts. Two of the eight variants were replicated at *P* <0.05. The regional plots of these two novel loci are shown in **Figure 6.3**. For the first locus, the lead SNP rs9393691 (chr6:26272829) is a common variant (MAF=0.383) within *HIST1H3G* (Histone cluster 1, H3g). This gene is found in the large histone gene cluster on chromosome 6. Histones are basic nuclear proteins that are responsible for the nucleosome structure of the chromosomal fiber in eukaryotes. Two molecules of each of the four core histones (H2A, H2B, H3, and H4) form an octamer, around which approximately 146 bp of DNA is wrapped in repeating units, called nucleosomes. The linker histone, H1, interacts with linker DNA between nucleosomes and functions in the compaction of chromatin into higher order structures. The association barely met the pre-defined threshold of *P*<1E-07, with 15-way *P*=9.90E-08. This region has been reportedly associated with many phenotypes including hematological traits and CHD risk factors, but the current lead SNP rs9393691 is not in LD with any of the known variants (r2<0.1) expect for one variant reported for association with height (rs10946808, r2=0.47) (**Table 6.5**). This variant exists in the published largest GWAS on CRP (Dehghan et al. 2011), but it was not significant (beta=0.0104, SE=0.0065, *P*=0.106). For the second locus, the lead SNP rs117410733 (chr15:52655560) is an intronic variant within *MYO5A*, which is a class of actin-based motor proteins involved in cytoplasmic vesicle transport and anchorage, spindle-

polealignment and mRNA translocation. Currently, there is no evidence in the literature supporting this gene's role in affecting circulating CRP level.

I also compared the summary statistics of 17 variants reported in the published largest GWAS (Dehghan et al. 2011). Six of those 17 variants are marginally significant in UK10K WGS (P <0.05) and five of them are genome-wide significant in 15-way meta-analysis (P <5E-08). Although the statistical significances differ, the effect size and directions are comparable between the previous GWAS, TwinsUK WGS and ALSPAC WGS. For the majority of the studied phenotypes, they are inverse normal transformed followed by a standardization of residuals. So, the phenotypes used in the association studies all have a normal distribution, with a mean of 0 and 1. This accounted for a lot of heterogeneity between individual GWAS that were included in the meta-analysis. For CRP for example, the mean (standard deviation) values for the raw phenotypes are 3.38(6.51) for TwinsUK and 0.84 (3.09) for ALSPAC. For the 17 positive controls, the effect sizes of 17 positive controls are very comparable between these two cohorts even though the raw phenotype values differ significantly.

Figure 6.2 Single marker association results of CRP from expanded meta-analysis

From top to bottom, the three plots are for 15-way, 6-way replication (WHI cohorts), and 21-way combined, respectively. X-axis is for chromosome and positions (build 37). Y-axis is for –log10(*P*). Variants passing threshold of 5E-08 and 1E-07 are shown in red and blue, respectively. For those passing threshold of 5E-08, known loci were marked in green text while putative novel loci were marked in red text.

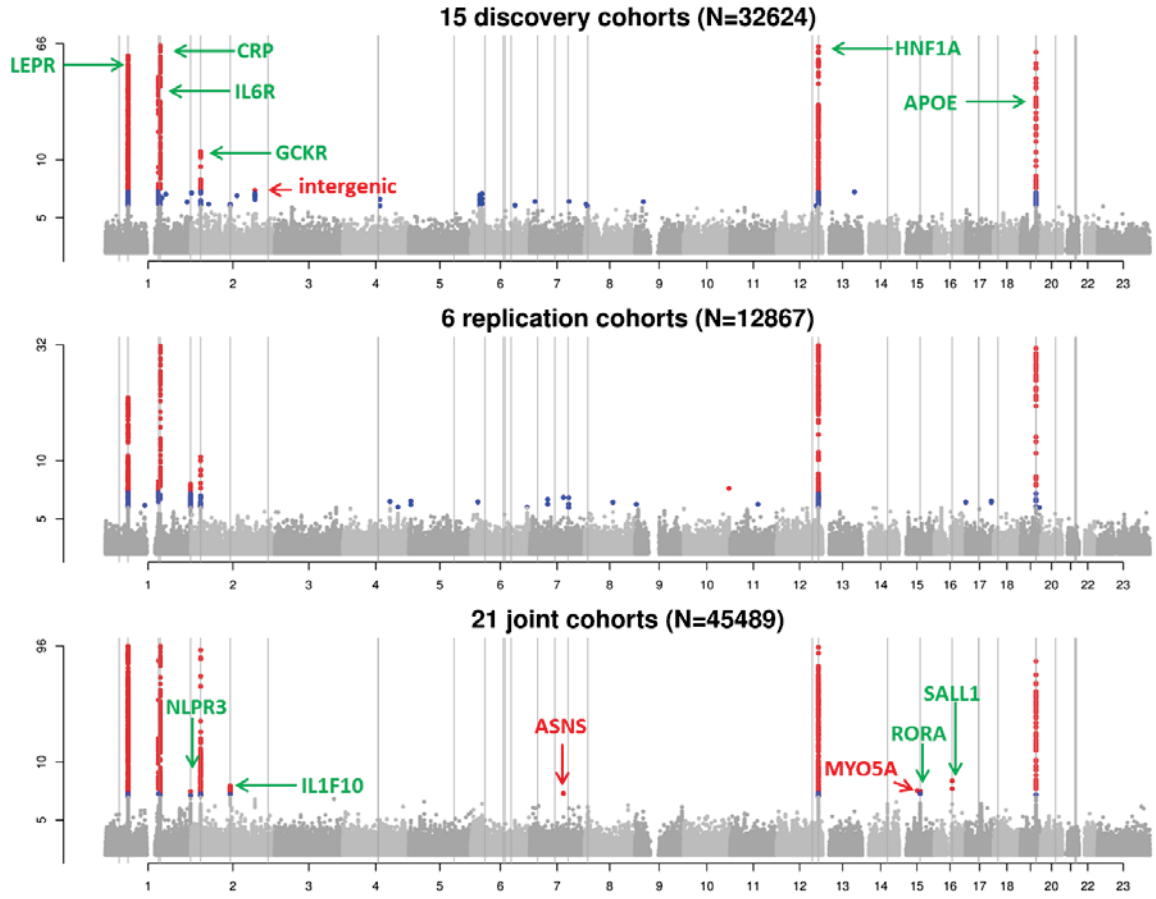


Table 6.3 Novel associations of CRP from expanded discovery meta-analysis

The first six variants are putative novel based on the 15-way expanded discovery with *P* <1E-07. The last two variants are putative novel based on the 21-way further expanded meta-analysis with *P* <5E-08.

					15-way					WHI Replication (with genome-wide data)				
rsID	**CHR**	**POS**	**Gene**	**EA**	**EAF**	**Beta**	**SE**	**P**	**N**	**EAF**	**Beta**	**SE**	**P**	**N**
rs137929481	1	176,045,265	*RFWD2*	G/A	0.002	0.584	0.109	9.06E-08	30615	0.004	-0.139	0.132	0.36	12865
rs35993482	2	1,320,638	*SNTG2*	A/G	0.021	0.207	0.038	7.00E-08	31454	0.025	-0.063	0.051	0.28	12865
rs76870040	2	185,422,774	*ZNF804A*	G/A	0.015	0.186	0.034	4.13E-08	32623	0.018	0.078	0.048	0.15	12868
rs9393691	6	26,272,829	*HIST1H3G*	C/T	0.383	-0.045	0.008	9.90E-08	32622	0.389	0.004	0.012	0.76	12866
rs9269303	6	32,539,581	*HLA*	T/G	0.476	-0.059	0.011	8.41E-08	30648	0.432	-0.001	0.015	0.95	12868
rs186492213	13	92,240,699	*GPC5*	G/A	0.002	-0.544	0.100	5.77E-08	30040	0.002	-0.107	0.157	0.55	12083
P<5E-08 in 21-way														
chr7:97545859	7	97545859	*ASNS*	A/G	0.009	-0.143	0.052	6.37E-03	32621	0.009	-0.449	0.075	1.36E-07	12865
rs117410733	15	52655560	*MYO5A*	G/A	0.009	-0.189	0.045	2.77E-05	32623	0.011	-0.241	0.060	4.01E-04	12865

21-way					Stage 2 replication					Combined				
EAF	**Beta**	**SE**	**P**	**N**	**EAF**	**Beta**	**SE**	**P**	**N**	**EAF**	**Beta**	**SE**	**P**	**N**
0.003	0.293	0.084	8.08E-04	43,480	0.003	0.722	0.452	0.11	1903	0.003	0.307	0.082	1.99E-04	46222
0.022	0.111	0.031	5.21E-04	44,319	0.311	0.000	0.014	0.98	7998	0.066	0.020	0.013	1.20E-01	58412
0.016	0.150	0.028	1.68E-07	45,491	0.015	0.003	0.035	0.93	21630	0.016	0.093	0.022	1.57E-05	73216
0.384	-0.029	0.007	8.09E-05	45,488	0.374	-0.045	0.009	5.82E-07	21529	0.381	-0.035	0.005	1.71E-11	73112
0.463	-0.039	0.009	2.28E-05	43,516	0.479	-0.011	0.041	0.80	1903	0.464	-0.038	0.009	1.23E-05	45419
0.002	-0.419	0.084	1.82E-06	42,123	0.002	0.004	0.159	0.98	7493	0.002	-0.326	0.074	1.17E-05	56966
P<5E-08 in 21-way														
0.009	-0.244	0.043	4.90E-08	45,486	0.010	0.056	0.064	0.38	7998	0.009	-0.151	0.036	2.21E-05	59579
0.009	-0.208	0.036	2.75E-08	45,488	0.006	-0.094	0.047	4.32E-02	7998	0.009	-0.165	0.028	5.78E-09	59581

Table 6.4 Cohort specific results of novel associations from expanded discovery

For each set of results, the effect allele frequency (EAF), beta, standard deviation (SE), *P* value, sample size (N), and imputation INFO score were presented. Records with *P* < 0.05 are highlighted in red text.

	chr1:176045265						chr2:1320638						chr2:185422774						chr6:26272829					
Cohort	**EAF**	**beta**	**SE**	**P**	**N**	**Info**	**EAF**	**beta**	**SE**	**P**	**N**	**Info**	**EAF**	**beta**	**SE**	**P**	**N**	**Info**	**EAF**	**beta**	**SE**	**P**	**N**	**Info**
ALSPAC WGS	--	--	--	--	--	--	--	--	--	--	--	--	0.014	0.118	0.171	4.9E-01	1167	0.99	0.376	-0.056	0.043	1.9E-01	1167	1.00
TwinsUK WGS	--	--	--	--	--	--	--	--	--	--	--	--	0.015	0.456	0.192	1.7E-02	879	1.00	0.378	-0.008	0.050	8.7E-01	879	1.00
ALASPAC GWA	0.002	0.439	0.406	2.8E-01	2226	0.57	0.019	0.238	0.14	8.9E-02	2226	0.62	0.016	0.256	0.119	3.2E-02	2226	1.00	0.377	-0.098	0.031	1.4E-03	2226	1.00
1958BC	0.003	0.589	0.245	1.6E-02	4910	0.58	0.019	0.281	0.093	2.7E-03	4910	0.61	0.015	0.198	0.083	1.7E-02	4910	0.99	0.358	-0.062	0.021	3.1E-03	4910	1.00
FHS	0.003	0.548	0.253	3.1E-02	6320	0.48	0.026	0.128	0.076	9.3E-02	6320	0.6	0.015	0.133	0.078	8.8E-02	6320	0.98	0.389	-0.019	0.020	3.3E-01	6320	0.99
INGI-FVG	0.001	4.623	2.773	9.7E-02	411	0.87	0.028	0.181	0.275	5.1E-01	411	0.54	0.016	0.091	0.269	7.4E-01	411	0.97	0.433	-0.074	0.067	2.7E-01	411	1.00
HELIC-A	0.002	0.837	0.661	2.1E-01	1093	0.46	0.063	0.421	0.132	1.8E-03	1093	0.48	0.011	0.136	0.211	5.2E-01	1093	1.00	0.502	-0.039	0.045	3.8E-01	1093	1.00
HELIC-P	0	0.681	5.495	9.0E-01	839	0.05	0.018	0.149	0.291	6.1E-01	839	0.4	0.040	0.102	0.129	4.3E-01	839	0.99	0.339	-0.047	0.054	3.8E-01	839	1.00
Incipe-1	0.003	0.272	0.778	7.3E-01	807	0.39	0.028	0.311	0.212	1.4E-01	807	0.48	0.012	0.522	0.211	1.3E-02	807	0.99	0.428	-0.042	0.050	4.0E-01	807	1.00
Incipe-2	0.002	0.732	0.68	2.8E-01	1332	0.36	0.033	-0.156	0.157	3.2E-01	1332	0.51	0.012	0.096	0.182	6.0E-01	1332	0.99	0.414	-0.057	0.039	1.4E-01	1332	1.00
LURIC-Ctrl	0.004	0.547	0.456	2.3E-01	1228	0.55	0.024	0.478	0.175	6.3E-03	1228	0.57	0.013	0.089	0.182	6.2E-01	1228	0.98	0.399	-0.116	0.041	4.8E-03	1228	1.00
LURIC-Case	0.004	0.742	0.451	1.0E-01	1202	0.53	0.024	0.11	0.17	5.2E-01	1202	0.61	0.017	0.173	0.162	2.9E-01	1202	0.98	0.390	-0.049	0.043	2.5E-01	1202	1.00
Procardis-case	0.001	0.694	0.301	2.1E-02	3732	0.88	0.011	0.139	0.167	4.1E-01	3732	0.46	0.016	0.196	0.092	3.3E-02	3732	1.00	0.376	-0.040	0.024	1.0E-01	3732	1.00
Procardis-ctrl	0.001	0.464	0.376	2.2E-01	3683	0.88	0.011	0.683	0.232	3.7E-03	3683	0.46	0.016	0.395	0.141	5.3E-03	3683	1.00	0.376	-0.025	0.036	5.0E-01	3683	1.00
TwinsUKvall	0.003	0.534	0.298	7.3E-02	2512	0.66	0.02	0.122	0.122	3.2E-01	2512	0.72	0.012	0.200	0.131	1.3E-01	2512	0.99	0.368	-0.009	0.031	7.6E-01	2512	1.00
TwinsUK GWA	0.004	1.08	0.464	2.1E-02	1017	0.65	0.021	-0.032	0.191	8.7E-01	1017	0.67	0.013	0.005	0.202	9.8E-01	1017	0.99	0.355	0.008	0.048	8.7E-01	1017	1.00
INGI-VBI	0.002	0.662	0.843	4.4E-01	1162	0.29	0.019	0.136	0.227	5.5E-01	1162	0.47	0.011	0.098	0.212	6.5E-01	1162	0.99	0.355	0.033	0.045	4.7E-01	1162	1.00
WHI-Garnet	0.004	0.079	0.25	7.5E-01	3388	0.63	0.027	-0.081	0.094	3.8E-01	3388	0.64	0.014	0.127	0.101	2.1E-01	3388	1.00	0.391	0.033	0.025	1.7E-01	3388	1.00
WHI-Gecco1	0.005	0.555	0.466	2.3E-01	780	0.62	0.025	-0.134	0.228	5.6E-01	780	0.56	0.017	0.426	0.206	3.9E-02	780	0.99	0.381	0.012	0.057	8.3E-01	780	1.00
WHI-Gecco2	0.004	0.218	0.532	6.8E-01	1072	0.42	0.022	0.077	0.189	6.8E-01	1072	0.61	0.018	0.011	0.160	9.5E-01	1072	1.00	0.398	0.024	0.042	5.7E-01	1072	1.00
WHI-Hipfx	0.004	0.04	0.398	9.2E-01	1716	0.49	0.024	-0.082	0.139	5.6E-01	1716	0.63	0.020	0.127	0.123	3.0E-01	1716	1.00	0.393	-0.014	0.034	6.9E-01	1716	1.00
WHI-Mopmap	0.005	-0.037	0.429	9.3E-01	721	0.75	0.03	0.137	0.195	4.8E-01	721	0.65	0.026	-0.06	0.169	7.2E-01	721	1.00	0.379	-0.054	0.054	3.2E-01	721	1.00
WHI-Whims	0.004	-0.592	0.216	6.1E-03	5191	0.54	0.024	-0.096	0.084	2.5E-01	5191	0.59	0.018	0.029	0.075	7.0E-01	5191	0.99	0.389	-0.007	0.020	7.5E-01	5191	1.00

Table 6.4. Cohort specific results of meta-analysis top hits (continued)

	chr6:32539581						chr13:92240699						chr7:97545859						chr15:52655560					
cohort	**EAF**	**beta**	**SE**	**P**	**N**	**Info**	**EAF**	**beta**	**SE**	**P**	**N**	**Info**	**EAF**	**beta**	**SE**	**P**	**N**	**Info**	**EAF**	**beta**	**SE**	**P**	**N**	**Info**
ALSPAC WGS	--	--	--	--	--	--	0.001	-0.533	0.563	3.4E-01	1167	0.91	0.008	-0.186	0.232	4.2E-01	1167	0.95	0.011	-0.1	0.198	6.2E-01	1167	0.99
TwinsUK WGS	--	--	--	--	--	--	0.003	-0.691	0.418	9.9E-02	879	0.97	0.007	0.088	0.302	7.7E-01	879	0.95	0.006	0.058	0.333	8.6E-01	879	0.93
alspac	0.492	-0.095	0.038	1.1E-02	2226	0.64	0.002	-0.297	0.422	4.8E-01	2226	0.81	0.006	-0.190	0.218	3.8E-01	2226	0.80	0.007	-0.075	0.186	6.9E-01	2226	0.94
b58c	0.503	-0.058	0.026	2.3E-02	4910	0.62	0.001	-0.303	0.312	3.3E-01	4910	0.74	0.008	-0.193	0.126	1.3E-01	4910	0.83	0.008	-0.115	0.118	3.3E-01	4910	0.97
fhs	0.536	-0.058	0.024	1.5E-02	6320	0.66	0.003	-0.619	0.202	2.2E-03	6320	0.75	0.008	-0.258	0.119	3.0E-02	6320	0.77	0.009	-0.187	0.105	7.6E-02	6320	0.93
fvg	0.285	-0.106	0.087	2.2E-01	411	0.69	0	11.67	11.452	3.1E-01	411	0.03	0.006	0.015	0.480	9.7E-01	411	0.77	0.007	0.18	0.397	6.5E-01	411	0.98
HA	0.447	0.082	0.056	1.4E-01	1093	0.68	0.003	0.275	0.527	6.0E-01	1093	0.53	0.001	-1.126	2.126	6.0E-01	1093	0.45	0.014	-0.186	0.195	3.4E-01	1093	0.96
HP	0.424	-0.033	0.060	5.9E-01	839	0.67	0	61.77	25.695	1.7E-02	839	0.01	0.001	0.551	1.029	5.9E-01	839	0.48	0.005	0.167	0.416	6.9E-01	839	0.76
incipe1	--	--	--	--	--	--	0.001	-0.257	0.699	7.1E-01	807	0.84	0.005	0.446	0.416	2.8E-01	807	0.77	0.013	-0.149	0.224	5.1E-01	807	0.94
incipe2	0.390	-0.091	0.048	5.8E-02	1332	0.68	--	--	--	--	--	--	0.005	-0.222	0.343	5.2E-01	1332	0.65	0.01	-0.115	0.215	5.9E-01	1332	0.88
luric1	0.491	-0.084	0.050	9.5E-02	1228	0.64	0.001	-0.561	0.669	4.0E-01	1228	0.71	0.008	-0.340	0.266	2.0E-01	1228	0.75	0.018	-0.41	0.148	5.8E-03	1228	0.99
luric2	0.473	-0.072	0.051	1.6E-01	1202	0.63	0.004	-1.094	0.336	1.2E-03	1202	0.93	0.008	0.290	0.273	2.9E-01	1202	0.68	0.011	-0.75	0.202	2.1E-04	1202	0.98
procase	0.450	-0.036	0.029	2.2E-01	3732	0.69	0.002	-0.407	0.264	1.2E-01	3732	0.96	0.018	-0.129	0.117	2.7E-01	3732	0.60	0.009	-0.106	0.124	3.9E-01	3732	0.97
proctrl	0.450	-0.083	0.043	5.5E-02	3683	0.69	0.002	-0.888	0.303	3.5E-03	3683	0.96	0.018	-0.074	0.163	6.5E-01	3683	0.60	0.009	-0.332	0.167	4.7E-02	3683	0.97
TwinsUKall	0.416	-0.066	0.036	6.8E-02	2512	0.70	0.003	-0.295	0.309	3.4E-01	2512	0.82	0.008	-0.144	0.169	3.9E-01	2512	0.86	0.005	0.062	0.202	7.6E-01	2512	0.99
TwinsUK	0.409	-0.048	0.055	3.8E-01	1017	0.70	0.002	-0.246	0.613	6.9E-01	1017	0.78	0.009	-0.219	0.270	4.2E-01	1017	0.85	0.006	0.415	0.291	1.5E-01	1017	0.99
vb	0.515	-0.076	0.201	7.0E-01	1162	0.05	0.004	-0.827	0.432	5.7E-02	1162	0.61	0.001	1.818	0.839	3.1E-02	1162	0.34	0.004	-0.233	0.344	5.0E-01	1162	0.97
whi_garnet	0.352	-0.016	0.029	5.8E-01	3388	0.78	0.001	-0.259	0.36	4.7E-01	3388	0.84	0.010	-0.300	0.134	2.5E-02	3388	0.85	0.012	-0.238	0.113	3.5E-02	3388	0.98
whi_GECCO1	0.486	0.034	0.067	6.2E-01	780	0.63	--	--	--	--	--	--	0.009	-0.691	0.326	3.4E-02	780	0.77	0.012	-0.06	0.256	8.2E-01	780	0.93
whi_GECCO2	0.507	0.121	0.053	2.2E-02	1072	0.64	0.002	-0.067	0.484	8.9E-01	1072	0.84	0.009	-0.608	0.240	1.1E-02	1072	0.78	0.01	-0.63	0.214	3.3E-03	1072	0.97
whi_hipfx	0.496	0.053	0.043	2.2E-01	1716	0.63	0.002	-0.092	0.4	8.2E-01	1716	0.78	0.008	-0.383	0.212	7.1E-02	1716	0.77	0.01	-0.452	0.172	8.6E-03	1716	0.97
whi_mopmap	0.317	0.033	0.061	5.9E-01	721	0.87	0.003	0.448	0.502	3.7E-01	721	0.93	0.007	-0.858	0.388	2.7E-02	721	0.63	0.013	-0.293	0.238	2.2E-01	721	0.99
whi_whims	0.455	-0.041	0.024	8.1E-02	5191	0.68	0.002	-0.175	0.233	4.5E-01	5191	0.83	0.008	-0.481	0.121	7.6E-05	5191	0.82	0.012	-0.12	0.093	2.0E-01	5191	0.98

Figure 6.3 Regional plots of two novel associations of CRP

The *P* value in the plot is from the 15-way expanded discovery meta-analysis. The top plot shows the *HIST1H3G* locus. The lead SNP rs9393691 (chr6:26272829) is significant in 15-way (*P* =9.90E-08). Its combined 27-way meta-analysis *P*=1.71E-11. The bottom plots show the MYO5A locus. The lead SNP rs117410733 (chr15:52655560) is not significant in 15-way (*P*=2.77E-05), but in 21-way (*P*=2.75E-08). Its combined 27-way meta-analysis P=5.78E-09.

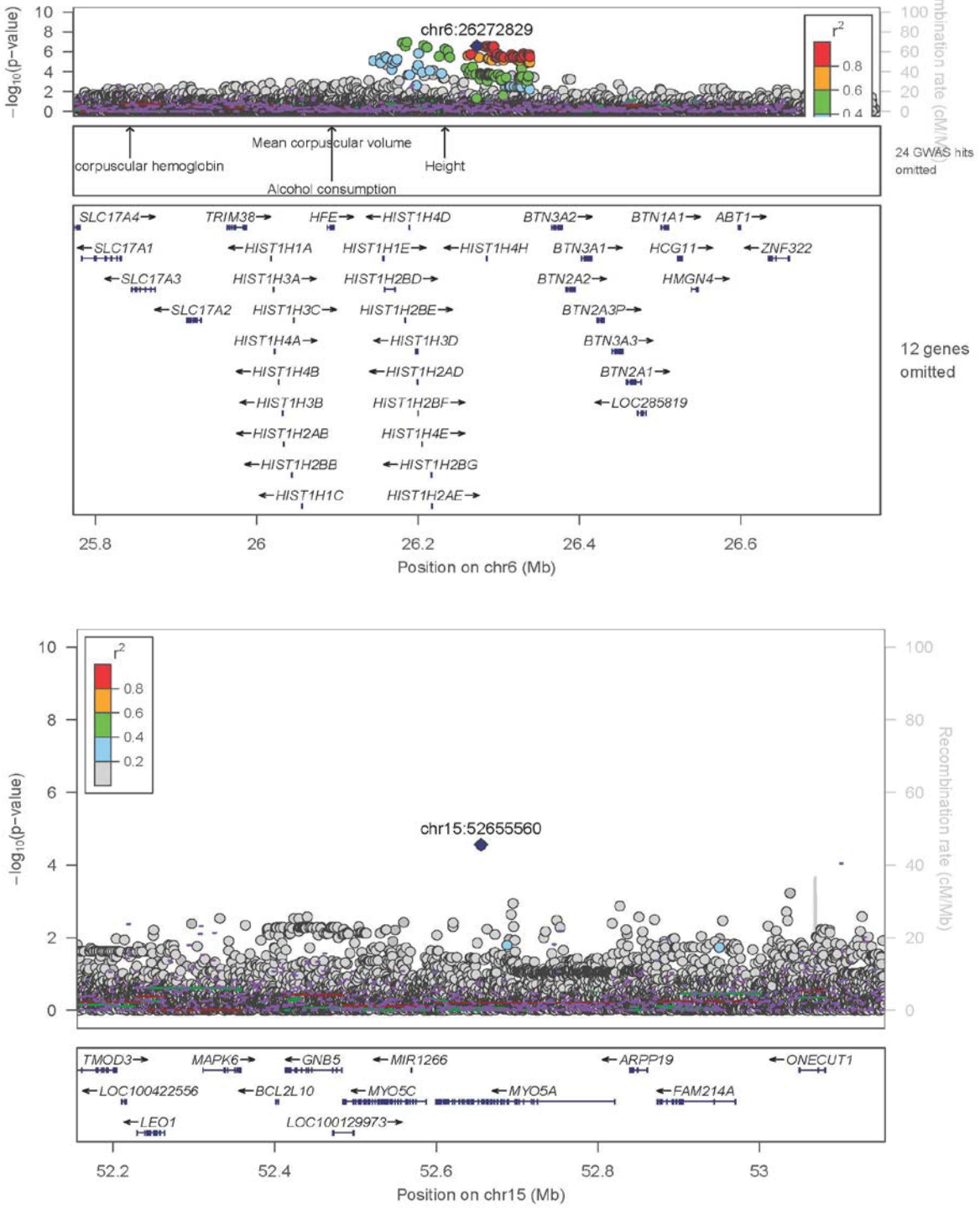


Table 6.5 LD between novel and known variants in *HIST1H3G*

This table lists 14 associations reported in GWAS Catalog that are within 1Mb of rs9393691. The LD of each variant with rs9393691 is shown in the last column, based on the WGS data of UK10K.

SNP	Trait	Chr	Pos	r2 with rs9393691
rs11754288	Cardiovascular disease risk factors	6	25776949	0.00
rs1165196	Urate levels	6	25813150	0.00
rs17342717	Iron status biomarkers	6	25821770	0.02
rs1183201	Uric acid levels	6	25823444	0.00
rs1408272	Mean corpuscular hemoglobin	6	25842951	0.02
rs1165205	Urate levels	6	25870542	0.00
rs1799945	Diastolic blood pressure	6	26091179	0.02
rs1799945	Iron levels	6	26091179	0.02
rs1800562	Hemoglobin	6	26093141	0.03
rs1800562	Cardiovascular disease risk factors	6	26093141	0.03
rs1800562	LDL cholesterol	6	26093141	0.03
rs198846	Blood pressure	6	26107463	0.02
rs198846	Hemoglobin	6	26107463	0.02
rs10946808	Height	6	26233387	0.47

6.3.2 Fine mapping of known and novel loci

The availability of WGS compared on GWAS based on sparse datasets allows one to evaluate statistically the plausibility of each variant in an association signal to be causally associated with a trait. To fine-map lipid-associated regions, I implemented the method of Maller et al. (Maller et al. 2012), as described in chapter 2 and the Methods section above. For a total of 37 regions examined, there are sufficient resolution to limit the number of possible causal variants to a small informative set for three regions (log10BF>5 and # of variants <20) (**Table 6.6**).

First for the *CRP* locus, a single variant rs3091244 is predicted to be causal with posterior probability of 1. This variant was reported in the first GWAS study on CRP (Ridker et al. 2008) and it was the lead SNP in the *CRP* locus. It was reported as a tri-allelic SNP, with the common allele G and two less-common alleles of A and T. This variant was filtered out from the UK10K WGS data, but was imputed as bi-allelic for all other imputed cohorts. In the 15-way meta-analysis, the frequency for the minor A is 0.33 and there is no allele of T. rs3091244 is the only fine-mapped CRP variant that overlaps with a TFBS binding site. This might provide a functional explanation for its causality. Second, for the *HFN1A* locus, a total of 13 variants together explain 95% of the posterior probability. Based on Regulome database (http://regulome.stanford.edu), two variants (rs2259816, rs1169313) have a high score of "1f" with supporting functional data from eQTL, TF binding, DNase peak, and a third variant (rs1169310) has a score of "2b", meaning with supporting functional data from TF binding, any motif, DNase Footprint, and DNase peak. Lastly, for the *APOE* locus, although the lead SNP rs429358 based on the 15-way meta-analysis is a missense variant, fine-mapping predicted rs1065853 with a higher posterior probability for causality, posterior probability of 0.73 for rs1065853 vs. 0.23 for rs429358. Based on the Regulome database, the score is "2b" for rs1065853, meaning supporting evidence from TF binding, any motif, DNase Footprint, and DNase peak, while the score for rs429358 is "5", meaning supporting evidence only from TF binding or DNase peak. The LD among these two variants are modest (r2=0.76). rs429358 has been reported for association with Alzheimer's diseases (Kim et al. 2011, Ramanan et al. 2014), but there was no reported association for rs1065853.

Table 6.6 Putative causal variants based on fine mapping

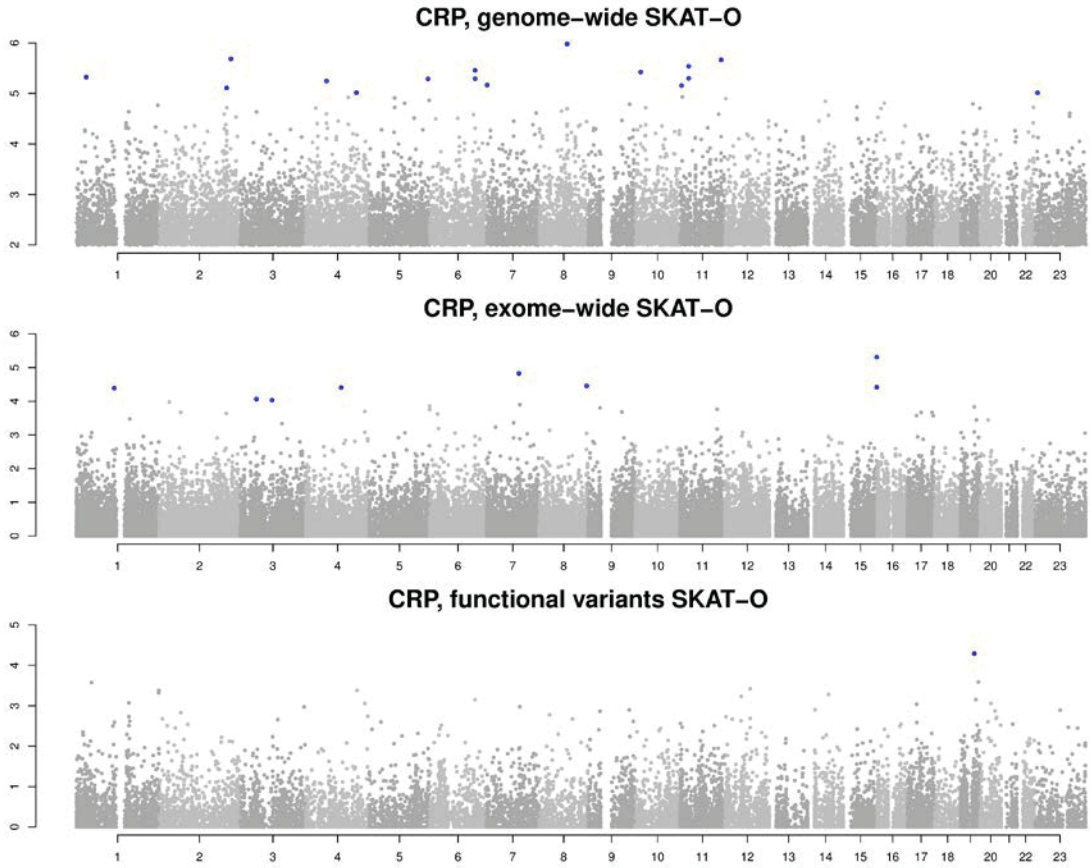
Fine-mapping						WGS 2-way					15-way				
loci	**rsID**	**CHRPOS**	**GWAVA**	**BF**	**PPA**	**EA**	**EAF**	**beta**	**SE**	**P**	**EAF**	**beta**	**SE**	**P**	**N**
CRP	rs3091244	chr1:159684665	Intronic	28.03	1.00	A	--	--	--	--	0.334	0.142	0.009	1.30E-54	31456
HNF1A	rs2264782	chr12:121432603	Upstream	19.07	0.03	T	0.354	-0.118	0.032	2.47E-04	0.372	-0.106	0.009	1.20E-34	32623
	rs2259852	chr12:121434833	3_prime_UTR	19.03	0.03	A	0.354	-0.119	0.032	2.34E-04	0.372	-0.107	0.009	5.58E-35	32622
	rs2464195	chr12:121435475	3_prime_UTR	19.03	0.03	A	0.354	-0.117	0.032	2.69E-04	0.372	-0.107	0.009	5.22E-35	32624
	rs2259816	chr12:121435587	3_prime_UTR	19.02	0.03	T	0.353	-0.118	0.032	2.58E-04	0.372	-0.106	0.009	6.63E-35	32623
	rs1169306	chr12:121438311	Downstream	18.99	0.03	T	0.357	-0.116	0.032	3.12E-04	0.374	-0.106	0.009	1.52E-34	32623
	rs735396	chr12:121438844	Downstream	19.48	0.09	C	0.353	-0.118	0.032	2.52E-04	0.373	-0.107	0.009	3.31E-35	32624
	rs1169309	chr12:121439192	3_prime_UTR	19.07	0.03	T	0.354	-0.119	0.032	2.33E-04	0.372	-0.106	0.009	6.17E-35	32623
	rs1169310	chr12:121439433	3_prime_UTR	19.48	0.09	A	0.354	-0.120	0.032	1.95E-04	0.373	-0.107	0.009	2.49E-35	32622
	rs1169311	chr12:121440731	3_prime_UTR	19.49	0.09	T	0.355	-0.119	0.032	2.03E-04	0.373	-0.107	0.009	2.53E-35	32623
	rs1169312	chr12:121441461	3_prime_UTR	19.39	0.07	T	0.356	-0.119	0.032	2.17E-04	0.375	-0.107	0.009	1.63E-35	32623
	rs1169313	chr12:121442670	Downstream	19.73	0.15	C	0.356	-0.118	0.032	2.25E-04	0.376	-0.107	0.009	1.41E-35	32623
	rs112249815	chr12:121444441	Exon	19.79	0.17	C	0.357	-0.116	0.032	3.19E-04	0.376	-0.107	0.009	1.47E-35	32620
	rs2257962	chr12:121445808	upstream_gene	19.69	0.14	C	0.356	-0.119	0.032	2.13E-04	0.375	-0.108	0.009	9.66E-36	32623
APOE	rs429358	chr19:45411941	Missense	14.83	0.23	C	0.142	-0.169	0.046	2.75E-04	0.142	-0.193	0.012	4.77E-55	32621
	rs1065853	chr19:45413233	upstream_gene	15.33	0.73	G	0.133	-0.232	0.053	1.09E-05	0.121	-0.202	0.014	2.96E-45	32620

6.3.3 Novel loci based on rare variants aggregation test

No variants are significant from the three types of SKAT-O tests, by using genome-wide significance threshold of $P < 6.8E-08$, 1.2E-06, 1E-05 respectively for genome-wide, exome-wide, and functional variants based SKAT-O (**Figure 6.4**). For those regions reaching less stringent threshold for suggestive association, as highlighted in blue in **Figure 6.4**, none of them are within 1Mb of known CRP loci. This could indicate truly a lack of rare variant that have large effects on serum CRP level, or due to inadequate power of the WGS samples used in this study.

Figure 6.4 Rare variants aggregation test results for CRP

The genome-wide significant signals are shown in red, with threshold of $P < 6.8E-08$, 1.2E-06, 1E-05 respectively for genome-wide, exome-wide, and functional variants based SKAT-O. Suggestive signals are shown in blue, with threshold of $P < 1E-05$, 1E-04, 1E-04 respectively for genome-wide, exome-wide, and functional variants based SKAT-O.



6.4 Conclusion & Discussion

6.4.1 Summary of main findings

Using 2,046 samples with WGS data and CRP levels measured (879 for TwinsUK, 1167 for ALSPAC), I applied a combination of approaches to conduct a genome-wide discovery of novel variants of low frequency associated with CRP level. Here, I identified two low frequency novel variants (MAF =2%) with *P*< 1E-6 but they were not replicated using imputed data. Then, I included up to ~73,000 samples with mostly imputed data to discover novel CRP associations across the full allele frequency spectrum. Here, I was able to discovery two novel associations. The first one is a common variant in the *HIST1H3G* locus (rs9393691, MAF=0.383, 27-way meta-analysis *P*=1.71E-11). The second one is a low frequency intronic variant within *MYO5A (*rs117410733, MAF=0.009, 27-way meta-analysis *P*=5.78E-09*)*. Fine-mapping analysis coupled with functional annotation narrowed down to putative causal variants within *CRP* and *APOE*. Rare variants aggregation tests did not identify putative novel loci that meet pre-defined genome-wide significance threshold.

6.4.2 Interpretation of results

The single marker association testing of CRP follows closely the expected relationship between EAF and effect size (beta) as dictated by study power (Park et al. 2011), as shown in **Figure 6.5**. Low frequency alleles of very high penetrance (beta ~1 SD) are unlikely to exist within this allelic space in the general European-ancestry population. Given that the genome-wide 21-way meta-analysis with a sample size more than 45,000, a number comparable to the previously published largest GWAS on CRP, using a combination of WGS samples and WGS imputed samples does not seem to be able to discover substantially more novel associations, either common or rare. The strongest association signal from single marker based analysis is a common variant within a gene-rich region, *HIST1H3G*. This association stood out only in the 27-way meta-analysis with a sample size of ~73,000. This implies that increasing sample size to this level for genome-wide association analysis could still be valuable. Given the gene-rich nature of this region and its association with many CVD related traits including lipids and FBC, targeted resequencing of this region and further

functional annotations are important steps for firmly establish this association and the understanding the underlying biology.

6.4.3 Future direction

Due to the constraint of time and resource, there is no independent WGS data was obtained to replicate some of the loci with suggestive evidence for rare variants aggregation based association. This could be an area worth further research efforts. Across, analysing multiple inflammatory traits together in a multivariate approach might discover common associations and pathways under the inflammation process in general. This could include the study of CRP and WBC traits together. IL-6 was also one of the 64 traits in UK10K, but its sample size is limited, only existing in ALSPAC and few of the external cohorts that were made available for expanded discovery and replication.

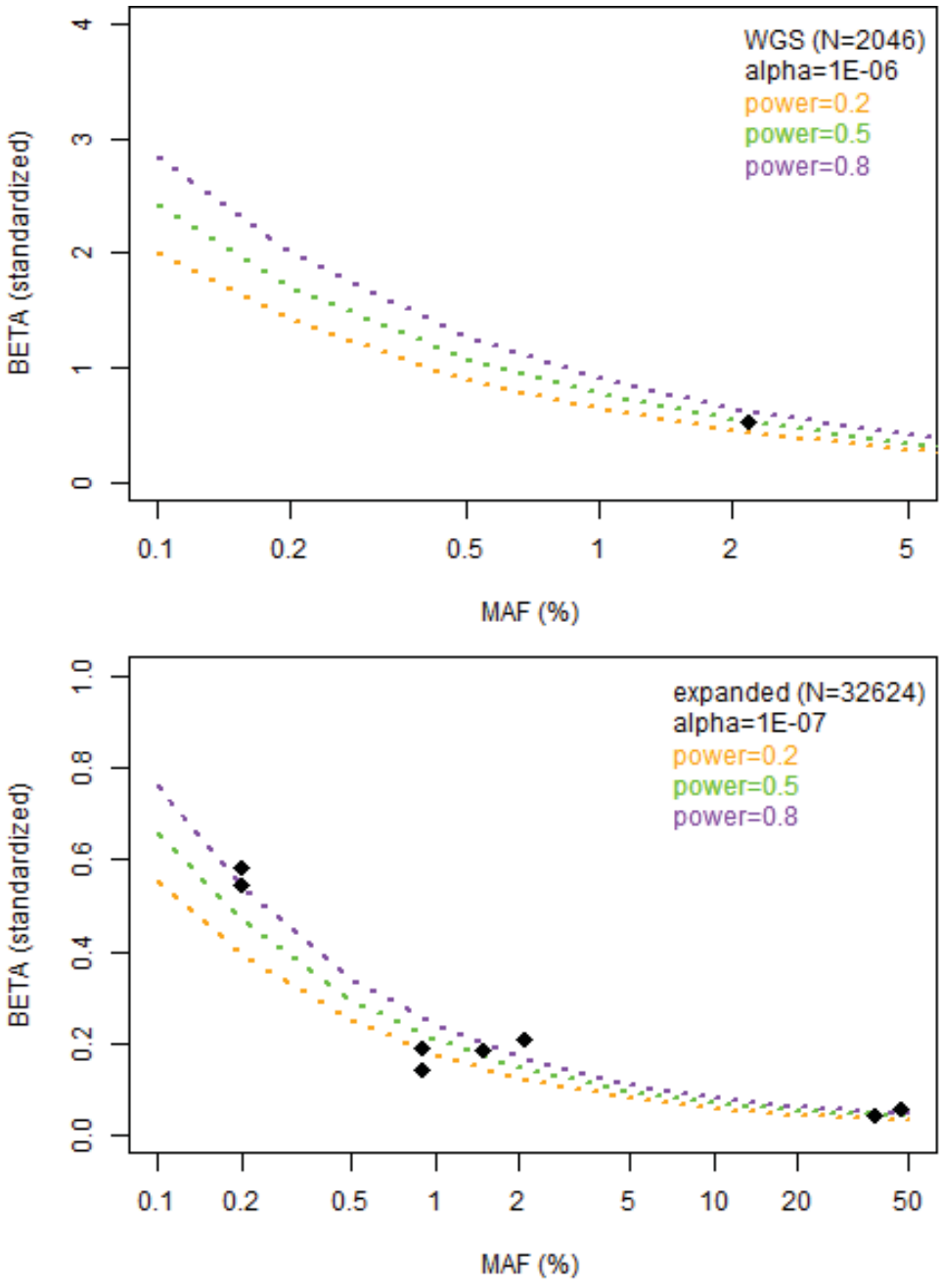
In this thesis, the study of CRP is overall separate from the other 12 CVD biomarkers. However, in the future a study combining CRP and lipids especially LDL would be desirable to fully understand their joint effects and interactions. It was reported that adding CRP to LDL in cell culture systems stimulated formation of foam cells, a typical feature of atherosclerotic plaques (Zwaka et al. 2001). However, it is not known whether this reflects opsonization of the LDL particles by CRP or an effect of CRP on the phagocytic cells themselves. Binding of CRP to lipids, especially lecithin (phosphatidyl choline), and to plasma lipoproteins has been known for decades. This could suggest new measurement of lipids bound CRP as the studied trait in genotype-phenotype association studies that aim to discover genetic factors underlying inflammatory process and CVD risk in general. Also, the co-analysis of CRP and WBC could also be explored. Previously, many studies have investigated both CRP and WBC for association with the risk of various diseases and aging prognosis (Keskin et al. 2004, Santos et al. 2004, Peltola et al. 2006, Willems et al. 2010).

Although the focus of this PhD thesis is on CVD related biomarkers, CRP could build a bridge between CVD and the other chronic disease with tremendous public health burden, cancer. Epidemiologic studies suggest that in patients with several types of solid cancers, elevated circulating levels of CRP are associated with poor prognosis, whereas in apparently healthy individuals from the general population, elevated levels of CRP are associated with increased future risk of cancer of any type. While most MR studies have failed to establish a causal role of serum CRP level to the development of CVD, a recent MR study provided

promising results for establishing a causal role of serum CRP levels to colorectal cancer (Nimptsch et al. 2015).

Figure 6.5 Statistical power and novel variants from single marker analysis

The top and bottom plots are for WGS samples and expanded discovery samples respectively. Y-axis is a variant's effect, expressed in standard deviation units. X-axis is MAF of effect alleles. Colored lines indicate 20%, 50%, and 80% power. Alpha is set at *P*<1E-06 for WGS and *P*<1E-07 for expanded discovery respectively. The two putative novel WGS variants are shown in the top power plot for WGS, and the eight putative novel variants from expanded discovery are shown in the bottom power plot for expanded discovery.



Chapter 7. Summary & Discussion

7.1 This thesis

The aims of UK10K-Cohort study include a direct genetic association studies with well-phenotyped samples and providing the UK10K WGS data as a resource for imputing external cohorts. Overall, these two aims are achieved as shown in my thesis.

For imputation, this thesis provided a full evaluation and thereafter recommended a best practice guide for running imputations. In particular, the implementation of using tract sharing algorithm to pick haplotypes was due to a direct observation that sampling more haplotypes (than the default number of 500) by the previously established k_hap approach improved imputation for low frequency and rare variants.

This study conducted genome-wide association studies for 13 CVD related quantitative traits, which used both directly sequenced data and imputed data. Compared to GWAS or WES, WGS is able to obtain an unbiased glimpse of the relative contributions of rare and common variation to the heritability of a complex trait.

7.2 Implication of findings for genetics of complex traits

A striking observation from single-marker association studies of 13 CVD traits was that - within the bounds of this study's statistical power - no alleles with stronger contribution to variance than classical lipid alleles are observed. The observed distribution of MAF and effect size for associated SNVs is compatible with expectations for polygenic models of inheritance, and suggests that low frequency alleles of very high penetrance (beta ~1 SD) are unlikely to exist within this allelic space in the general European-ancestry population. Examples such as the rare *APOC3* or *LDLR* variants, with sufficient individual effect sizes to be clinically informative, are beginning to emerge. However, greater power than the current study will be required for capturing a greater proportion of missing heritability through either

increases in sample size (most effective for common variants) or genotyping accuracy and SNV density (most effective for low frequency and rare variants).

Overall, this study suggests a paucity of variants of low frequencies with strong effects that were not identified by previous GWAS approaches. Even if this could be viewed as a negative picture, this knowledge was not clear at the beginning of the UK10K study. Therefore, this is still valuable knowledge and reference for investigators who are planning their own WGS based studies. Overall, for WGS studies with samples at this size (<4,000) or even much smaller, published studies have reported very few novel findings. So, at least for traits where WGS has already been conducted, future studies would need more power before taking off. Although the current study mainly examined quantitative traits, this overall lack of finding for rare variants with strong effects is also true for cardiovascular diseases traits, including MI (Holmen et al. 2014) and early-onset MI (Do et al. 2015). Also for common autoimmune disease, rare variants at known loci were reported to have a negligible role in diseases susceptibility and missing heritability (Hunt et al. 2013). These observations are generalised to all other UK10K traits, as shown in **Figure 7.1** and **Figure 7.2**.

Figure 7.1 Allelic spectrum for single marker association results in UK10K

This plot is adopted from the UK10K main paper, made by Klaudia Walter. Allelic spectrum for single marker association results for independent variants identified in the single-variant analysis for 31 core traits in UK10K-cohorts. A variant's effect (absolute value of Beta, expressed in standard deviation units) is given as a function of minor allele frequency (MAF, x-axis). Error bars are proportional to the standard error of the beta, variants identifying known loci are dark blue and variants identifying novel signals replicated in independent studies are coloured in light blue. The red and orange lines indicate 80% power at experiment-wide significance level (p-value ≤ 4.62x10-10) for the maximum theoretical sample size for the WGS sample and WGS+GWA respectively. Thus, the WGS-based association study has 80% power to detect loci with Beta-MAF values falling on the lavender shading.

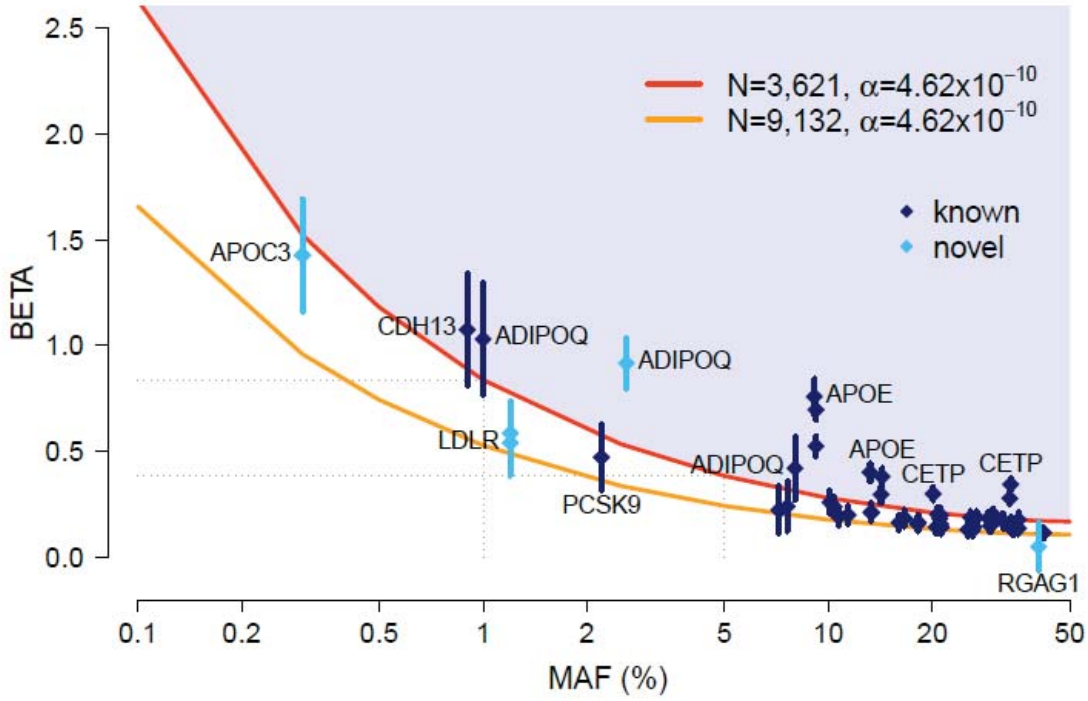
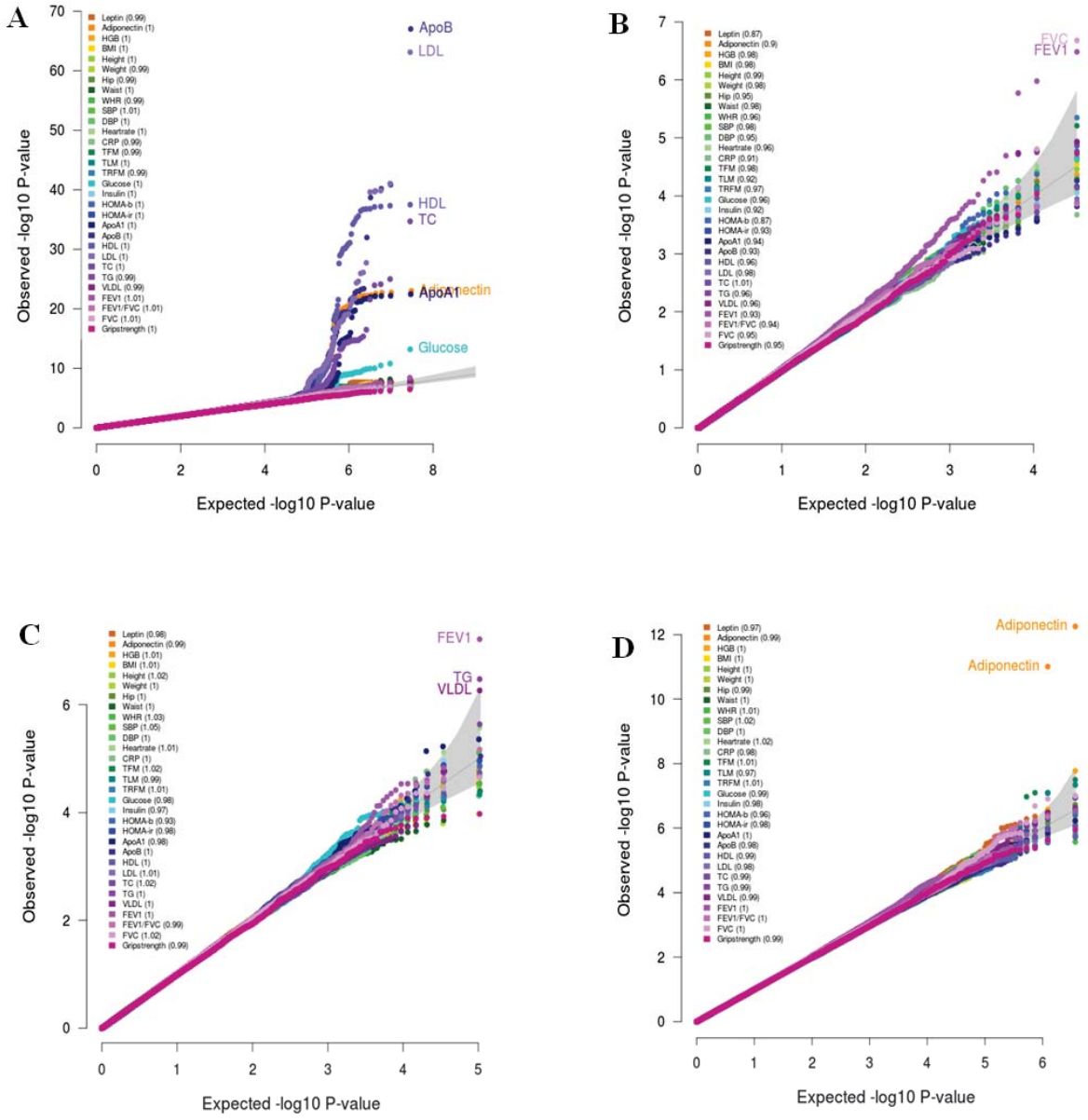


Figure 7.2 QQ plot of association tests for 31 UK10K core traits

This plot is adopted from the UK10K main paper, made by Klaudia Walter.

The four plots A-D are for single marker association tests, exome-based rare variant tests (SKAT, functional scan), exome-based rare variant tests (SKAT, naïve scan), genome-wide rare variant tests (SKAT, 3-kb windows), respectively.



7.3 Strength and limitations of the current study

The Strength of the current study included at least the following three aspects. First, this is a pioneering exploration of using WGS in association studies for a large number of CVD biomarkers. The UK10K study is one of the largest WGS based study on a large set of highly correlated phenotypes. The lipids WGS described in chapter four is the largest WGS for these traits so far. The association studies using WGS for full blood counts and CRP are the first ones for these traits. Second, a large imputation reference panel and new feature of a major imputation software was developed from this work. This addressed two key issues for imputation: a. the combining of WGS based reference panels; b. the strategy for sampling the mostly matched haplotypes to get the optimal results for achieving imputation accuracy while retaining the computing time. The discoveries of additional associations imply that these imputation panels will aid future discoveries. Third, analyses are standardized by the development of high-throughput pipelines and an integrated suite of analytic approaches. Through this project, I have developed pipelines for running imputation, genome-wide association tests, work-flow for loci prioritization, and visualization of genome-wide statistics. The highly automated pipelines facilitate scaling and independent cross checking, which are important for genome-wide analyses with large volume data from WGS.

The following four limitations are worth noting for this study. First, the sample size is still limited given the nature of discovering and replication rare variants. It is suggested that a discovery sample of at least 25,000 subjects and a substantial replication set is needed for a well-powered study that aims to identify rare variants (Zuk et al. 2014). This could be addressed by joining larger consortisum and by following up a more comprehensive set of variants that pass a less stringent statistical threshold. Second, although low-depth sequencing has been proven quite effective in characterizing the whole genome, high depth coverage (up to 80X) might significantly improve accuracy of detecting rare and particularly singleton variants. This in turn could significantly increase power of rare variant tests. Third, the phenotype is currently analysed individually, whereas more integrative approaches such as multivariate analysis could be applied, for both lipids and blood traits. In additional to the power gained, adopting a multivariate approach allows estimation of the amount co-heritability, or pleiotropy across traits. Fourth, a further exploration of rare variants test. For regions within genes, I need to deal with different gene sizes, regions with dense and overlapping genes. For intergenic and noncoding regions, the current approach of sliding

window is agnostic, therefore, there is space for better methods implementing better aggregation strategies based on biological priors.

7.4 Recommendations for future research in the field

Robinson and colleagues made six recommendations for explaining additional genetic variation in complex traits (Robinson et al. 2014). I ordered them based on my perception of their importance, with the first one being the most important. They are: 1. increase sample size to address limited power; 2. collect more and better phenotypes to address poorly described phenotypes; 3. imputation and direct sequencing to address poor allele frequency coverage; 4. use endophenotypes, expression, and pathway information to address poor integration of functional data; 5. uultivariate analysis for addressing ignored pleiotropy; 6. use CNVs and mitochondrial SNPs to address structure variants that was usually ignored in first generation GWAS. In my view, sample size is still the number 1 limiting factor that most sequencing studies conducted so far have failed to discover a lot of novel association signals. At this moment, I am getting more samples for some of the 13 studied traits, and more novel signals begin to emerge.

7.4.1 Larger sample size with increased power

Height is a model trait for understanding how human genetics of complex traits works, it has a high heritability (~80%) and is easily measured in large samples. The international Genetic Investigation of Anthropometric Traits (GIANT) Consortium now built the largest sample to date (N> 250,000) and pinned down 697 variants (in 424 gene regions) associated with height (Wood et al. 2014), the largest number to date associated with any trait or disease. These loci now explained 20 percent of the heritability of height, up from about 12 percent when a GWAS with 183,727 individuals identified 180 loci (Lango Allen et al. 2010). The study also narrows down the genomic regions that contain a substantial proportion of remaining variation to be discovered with even larger sample sizes. The results are consistent with a genetic architecture for human height that is characterized by a very large but finite number (thousands) of causal variants, located throughout the genome but clustered in both a biological and genomic manner. This pseudo-infinitesimal model of genetic architecture may characterize many other polygenic traits and diseases.

It has been argued that larger GWAS will provide limited new biological insights even though they identify more loci and explain more missing heritability because the range

of implicated genes and pathways will lose specificity and cover essentially the entire genome (Goldstein 2009). On the contrary, this largest GWAS on height showed that the identification of many hundred and even thousand associated variants can continue to provide biologically relevant information and prioritize many additional new and relevant genes. The observations that genes and especially pathways implicated by multiple variants suggests that the larger set of results retain biological specificity but that, at some point, a new set of associated variants will largely highlight the same genes, pathways and biological mechanisms as have already been seen. However, this endpoint has not reached for height, not to mention GWAS studies of other complex traits with much less sample size.
On the basis of the results of large genetic studies of height, it is anticipated that increasing the number of associated loci for other traits and diseases could yield similarly rich lists that would generate new biological hypotheses and motivate future research into the basis of human biology and disease. There is also strong evidence of multiple alleles at the same locus segregating in the population and for associated loci overlapping with mendelian forms, suggesting a large but finite genomic mutational target with effect sizes ranging from minute (~0.01 s.d.) to gigantic (>3 s.d.; in the case of monogenic mutations). This is in line with the findings of rare variants with large effects within *APOC3* and *LDLR*.

7.4.2 High genotyping accuracy through high-depth WGS

The systematic genome-wide evaluation of low frequency and rare variants over a large number of representative traits has implications for future studies of complex traits. For common variants (MAF≥5%), variation within Europe is fully captured by current low depth sequencing and current imputation approach, and increase sample size would be most beneficial. For example, the identification of the chrX signal for LDL was mainly driven by sample size increasing. For low frequency and rare variants down to approximately 0.1% MAF, substantial relative power gains can be achieved through increases in genotyping accuracy. For example, power gains of as much as 22-fold could be observed under some scenarios (SNVs of MAFs=0.1-0.5% and effect sizes of 0.6-1.2 standard deviations) when genotype accuracy improved from r^2<0.5 to 1 (The UK10K Consortium 2015). Future increases in the number of haplotypes in imputation reference panels are expected to improve imputation accuracy for alleles down to around 0.1%, and could lead to novel discoveries in

this frequency range. For example, the *APOC3* rare variant (MAF=0.2%) was significant in the WGS alone even though the sample size is modest (Timpson et al. 2014).

Based on UK10K data, the power increases as much as 22-fold when genotype accuracy was improved from r2<0.5 to r2=1. But for common variants, the UK10K study also showed that variation within Europe is fully and adequately captured by low-coverage sequencing and adding sequencing depth would not be much valuable. This is in line with the lack of novel findings for common variants from the traits that I studied. There is compelling evidence that the classical lipid alleles (and notably the *APOE* variant rs7412) represent extremes of genetic risk for a wide range of biomedical traits where our sample is fully powered (blue shading in **Figure 7.1**). Given the high degree of coverage of the human genome achieved in the UK10K study, results here do suggest that across these traits future "low hanging fruit" discoveries of low frequency variants of high penetrance (as defined by study power) are highly unlikely.

7.4.3 Better methods for rare variants aggregation test and replication

The assessment of rare variants using both exome-based and genome-based tests suggests that both naïve and functional scans were broadly underpowered to detect associations with high certainty (Zuk et al. 2014). Genetic variants at this frequency range potentially include those of high penetrance and clinically functional. The UK10K study used both low-depth WGS and high-depth WES. For fully capturing rare variants for aggregation based tests, high depth WGS might be the preferable approach. Furthermore, accounting for the observed heterogeneity in allelic architecture between loci is likely to remain the biggest challenge in assessing the contribution of rare variants to phenotypic variance. For this thesis, I was only able to get rare variants based replication data for four lipids traits but not for CRP and eight FBC traits. More data for both discovery and replication would enable a more comprehensive evaluation of the rare variants aggregation methods and results.

7.4.4 System biology approach that integrates various functional data

Since 2010, when massively parallel sequencing has become largely available, also when the U10K study was initiated, no major new insights into genes governing lipid metabolism have been reported. This is probably because the etiologies of true Mendelian lipid disorders with overt clinical complications have been largely resolved. In the meantime, proving the importance of new candidate genes is challenging due to very low frequencies of large impact variants in the population. For example, a loss of two functional *LCAT* alleles causes near HDL deficiency but the DNA of 100,000 individuals was needed simply to statistically link LCAT to HDL cholesterol levels (Teslovich et al. 2010). Also, *in silico* program do not consider other aspects of protein biochemistry such as post-translational modification, protein-protein interactions (Tchernitchko et al. 2004). It was therefore suggested that to refocus efforts on direct functional analysis of the genes that have already been discovered (Kuivenhoven and Hegele 2014). It has now become possible to identify the downstream effects of disease-associated SNPs through meta-analysis of eQTL (Westra et al. 2013). Another promising strategy is to identify novel key regulators of proteins that have previously been shown to interact with gene products that have established roles, through the use of proteomic network analyses to create phenomes or interactomes that shed new light on the origin of human diseases (Lage et al. 2007). Finally, the combination of rare and common variants as well as comparing across different populations could also lead to novel discovery. A good example is *PCSK9*, where the initial finding of a very low frequency functional mutation in *ADH* (Abifadel et al. 2003) and discoveries of more common variants in larger multi-ethnical populations led to the discovery of common sequence variations with large effects on plasma cholesterol levels in selected populations (Cohen et al. 2005).

7.4.5 Pleiotropy analysis

Previously, I have developed methods for pleiotropy analyses to analyze multiple correlated phenotypes in a unified framework, for psychiatric disorders (Huang et al. 2010) and for cardio-metabolic traits (Huang et al. 2011). More recently, Stephens and colleges developed a framework for assessing associations between multiple related outcome variables and a single explanatory variable of interest, based on Bayesian model comparison and model averaging for multivariate regressions (Stephens 2013). This framework unifies several common approaches to address the issues of testing multiple related phenotypes, with both standard univariate and standard multivariate association tests included as special cases. The other advantage of this newly proposed framework is that it unifies the problems of testing

for associations and explaining associations. I plan to adopt methods like this one to test the 4 lipids traits and the 8 hematological traits in a unified manner.

7.4.6 Thinking genetics in the context of the trend of metabolic syndrome.

Environment (i.e., the trend of metabolic syndrome such as increasing prevalence of obesity) may be playing an increasing role, but at the same time this trend offers the unique opportunity for longitudinal studies like FHS and TwinsUK and newer large cohorts like UK Biobank, to study secular trends in the contribution of genetic variation to cardiometabolic traits and the specific contribution of gene by environment interactions to cardiometabolic traits. The genetic variation identified in the backdrop of this trend would be more relevant to the current trend of metabolic syndrome such as increasing prevalence of obesity. That is, we are more likely to identify those genetic variants that will have effects on phenotypes only when environmental risk factors exist. Therefore, these genetic variants could be used more effectively to identify and benefit those whose could minimize the environmental risk factors and maintain a healthy lifestyle. There is also increasing recognition of the importance of different patterns of obesity and tissue depots of fat, and the genetics of these traits may differ (WHR vs. BMI). For example, abdominal adiposity is more connected with metabolic syndrome. Finally, this trend demands a more rigorous phenotype harmonization process for phenotype-genotype association studies. For example, to tease apart the modulation effect of BMI to type-2 diabetes, BMI should be regressed out from the phenotype before the association analysis.

References

1 . Abecasis, G. R., A. Auton, L. D. Brooks, M. A. DePristo, R. M. Durbin, R. E. Handsaker, . . . G. A. McVean (2012). "An integrated map of genetic variation from 1,092 human genomes." Nature **491**(7422): 56-65.

2 . Abifadel, M., J. P. Rabes, M. Devillers, A. Munnich, D. Erlich, C. Junien, . . . C. Boileau (2009). "Mutations and polymorphisms in the proprotein convertase subtilisin kexin 9 (PCSK9) gene in cholesterol metabolism and disease." Hum Mutat **30**(4): 520-529.

3 . Abifadel, M., M. Varret, J. P. Rabes, D. Allard, K. Ouguerram, M. Devillers, . . . C. Boileau (2003). "Mutations in PCSK9 cause autosomal dominant hypercholesterolemia." Nat Genet **34**(2): 154-156.

4 . Akinsheye, I., A. Alsultan, N. Solovieff, D. Ngo, C. T. Baldwin, P. Sebastiani, . . . M. H. Steinberg (2011). "Fetal hemoglobin in sickle cell anemia." Blood **118**(1): 19-27.

5 . Albert, T. J., M. N. Molla, D. M. Muzny, L. Nazareth, D. Wheeler, X. Song, . . . R. A. Gibbs (2007). "Direct selection of human genomic loci by microarray hybridization." Nat Methods **4**(11): 903-905.

6 . Allard, D., S. Amsellem, M. Abifadel, M. Trillard, M. Devillers, G. Luc, . . . J. P. Rabes (2005). "Novel mutations of the PCSK9 gene cause variable phenotype of autosomal dominant hypercholesterolemia." Hum Mutat **26**(5): 497.

7 . Amarenco, P., P. Lavallee and P. J. Touboul (2004). "Stroke prevention, blood cholesterol, and statins." Lancet Neurol **3**(5): 271-278.

8 . Amarenco, P. and P. G. Steg (2007). "The paradox of cholesterol and stroke." Lancet **370**(9602): 1803-1804.

9 . Anderson, G. L., J. Manson, R. Wallace, B. Lund, D. Hall, S. Davis, . . . R. L. Prentice (2003). "Implementation of the Women's Health Initiative study design." Ann Epidemiol **13**(9 Suppl): S5-17.

10 . Anderson, K. M., W. P. Castelli and D. Levy (1987). "Cholesterol and mortality. 30 years of follow-up from the Framingham study." JAMA **257**(16): 2176-2180.

11 . Arsenault, B. J., S. M. Boekholdt and J. J. Kastelein (2011). "Lipid parameters for measuring risk of cardiovascular disease." Nat Rev Cardiol **8**(4): 197-206.

12 . Arsenault, B. J., J. S. Rana, E. S. Stroes, J. P. Despres, P. K. Shah, J. J. Kastelein, . . . K. T. Khaw (2009). "Beyond low-density lipoprotein cholesterol: respective contributions of non-high-density lipoprotein cholesterol levels, triglycerides, and the total cholesterol/high-density lipoprotein cholesterol ratio to coronary heart disease risk in apparently healthy men and women." J Am Coll Cardiol **55**(1): 35-41.

13 . Asimit, J. and E. Zeggini (2010). "Rare variant association analysis methods for complex traits." Annu Rev Genet **44**: 293-308.

14 . Aslibekyan, S., E. K. Kabagambe, M. R. Irvin, R. J. Straka, I. B. Borecki, H. K. Tiwari, . . . D. K. Arnett (2012). "A genome-wide association study of inflammatory biomarker changes in response to fenofibrate treatment in the Genetics of Lipid Lowering Drug and Diet Network." Pharmacogenet Genomics **22**(3): 191-197.

15 . Assmann, G., H. Schulte, A. von Eckardstein and Y. Huang (1996). "High-density lipoprotein cholesterol as a predictor of coronary heart disease risk. The PROCAM experience and pathophysiological implications for reverse cholesterol transport." Atherosclerosis **124 Suppl**: S11-20.

16 . Aulchenko, Y. S., S. Ripatti, I. Lindqvist, D. Boomsma, I. M. Heid, P. P. Pramstaller, . . . E. Consortium (2009). "Loci influencing lipid levels and coronary heart disease risk in 16 European population cohorts." Nat Genet **41**(1): 47-55.

17 . Badimon, J. J., C. G. Santos-Gallego and L. Badimon (2010). "[Importance of HDL cholesterol in atherothrombosis: how did we get here? Where are we going?]." Rev Esp Cardiol **63 Suppl 2**: 20-35.

18 . Badimon, L. and G. Vilahur (2012). "LDL-cholesterol versus HDL-cholesterol in the atherosclerotic plaque: inflammatory resolution versus thrombotic chaos." Ann N Y Acad Sci **1254**: 18-32.

19 . Baigent, C., A. Keech, P. M. Kearney, L. Blackwell, G. Buck, C. Pollicino, . . . C. Cholesterol Treatment Trialists (2005). "Efficacy and safety of cholesterol-lowering treatment: prospective meta-

analysis of data from 90,056 participants in 14 randomised trials of statins." Lancet **366**(9493): 1267-1278.
20 . Bak, S., D. Gaist, S. H. Sindrup, A. Skytthe and K. Christensen (2002). "Genetic liability in stroke: a long-term follow-up study of Danish twins." Stroke **33**(3): 769-774.
21 . Ballester, B., A. Medina-Rivera, D. Schmidt, M. Gonzalez-Porta, M. Carlucci, X. Chen, . . . M. D. Wilson (2014). "Multi-species, multi-transcription factor binding highlights conserved control of tissue-specific biological pathways." Elife **3**: e02626.
22 . Bamshad, M. J., S. B. Ng, A. W. Bigham, H. K. Tabor, M. J. Emond, D. A. Nickerson and J. Shendure (2011). "Exome sequencing as a tool for Mendelian disease gene discovery." Nat Rev Genet **12**(11): 745-755.
23 . Bansal, S., J. E. Buring, N. Rifai, S. Mora, F. M. Sacks and P. M. Ridker (2007). "Fasting compared with nonfasting triglycerides and risk of cardiovascular events in women." JAMA **298**(3): 309-316.
24 . Barreiro, L. B., L. Tailleux, A. A. Pai, B. Gicquel, J. C. Marioni and Y. Gilad (2012). "Deciphering the genetic architecture of variation in the immune response to Mycobacterium tuberculosis infection." Proc Natl Acad Sci U S A **109**(4): 1204-1209.
25 . Barrett, J. C. and L. R. Cardon (2006). "Evaluating coverage of genome-wide association studies." Nat Genet **38**(6): 659-662.
26 . Barter, P. (2009). "Lessons learned from the Investigation of Lipid Level Management to Understand its Impact in Atherosclerotic Events (ILLUMINATE) trial." Am J Cardiol **104**(10 Suppl): 10E-15E.
27 . Barter, P. J., M. Caulfield, M. Eriksson, S. M. Grundy, J. J. Kastelein, M. Komajda, . . . I. Investigators (2007). "Effects of torcetrapib in patients at high risk for coronary events." N Engl J Med **357**(21): 2109-2122.
28 . Barton, A., W. Thomson, X. Ke, S. Eyre, A. Hinks, J. Bowes, . . . J. Worthington (2008). "Rheumatoid arthritis susceptibility loci at chromosomes 10p15, 12q13 and 22q13." Nat Genet **40**(10): 1156-1159.
29 . Basel-Vanagaite, L., N. Zevit, A. Har Zahav, L. Guo, S. Parathath, M. Pasmanik-Chor, . . . R. Shamir (2012). "Transient infantile hypertriglyceridemia, fatty liver, and hepatic fibrosis caused by mutated GPD1, encoding glycerol-3-phosphate dehydrogenase 1." Am J Hum Genet **90**(1): 49-60.
30 . Bauer, D. E. and S. H. Orkin (2011). "Update on fetal hemoglobin gene regulation in hemoglobinopathies." Curr Opin Pediatr **23**(1): 1-8.
31 . Beekman, M., B. T. Heijmans, N. G. Martin, N. L. Pedersen, J. B. Whitfield, U. DeFaire, . . . D. I. Boomsma (2002). "Heritabilities of apolipoprotein and lipid levels in three countries." Twin Res **5**(2): 87-97.
32 . Beigneux, A. P., R. Franssen, A. Bensadoun, P. Gin, K. Melford, J. Peter, . . . S. G. Young (2009). "Chylomicronemia with a mutant GPIHBP1 (Q115P) that cannot bind lipoprotein lipase." Arterioscler Thromb Vasc Biol **29**(6): 956-962.
33 . Bell, G. I., S. Horita and J. H. Karam (1984). "A polymorphic locus near the human insulin gene is associated with insulin-dependent diabetes mellitus." Diabetes **33**(2): 176-183.
34 . Benyamin, B., M. A. Ferreira, G. Willemsen, S. Gordon, R. P. Middelberg, B. P. McEvoy, . . . J. B. Whitfield (2009). "Common variants in TMPRSS6 are associated with iron status and erythrocyte volume." Nat Genet **41**(11): 1173-1175.
35 . Benyamin, B., A. F. McRae, G. Zhu, S. Gordon, A. K. Henders, A. Palotie, . . . P. M. Visscher (2009). "Variants in TF and HFE explain approximately 40% of genetic variation in serum-transferrin levels." Am J Hum Genet **84**(1): 60-65.
36 . Berge, K. E., H. Tian, G. A. Graf, L. Yu, N. V. Grishin, J. Schultz, . . . H. H. Hobbs (2000). "Accumulation of dietary cholesterol in sitosterolemia caused by mutations in adjacent ABC transporters." Science **290**(5497): 1771-1775.
37 . Biomarkers Definitions Working, G. (2001). "Biomarkers and surrogate endpoints: preferred definitions and conceptual framework." Clin Pharmacol Ther **69**(3): 89-95.

38 . Blair, D. R., C. S. Lyttle, J. M. Mortensen, C. F. Bearden, A. B. Jensen, H. Khiabanian, . . . A. Rzhetsky (2013). "A nondegenerate code of deleterious variants in Mendelian loci contributes to complex disease risk." Cell **155**(1): 70-80.
39 . Boekholdt, S. M. and J. J. Kastelein (2010). "C-reactive protein and cardiovascular risk: more fuel to the fire." Lancet **375**(9709): 95-96.
40 . Bonnelykke, K., M. C. Matheson, T. H. Pers, R. Granell, D. P. Strachan, A. C. Alves, . . . C. Lifecourse Epidemiology (2013). "Meta-analysis of genome-wide association studies identifies ten loci influencing allergic sensitization." Nat Genet **45**(8): 902-906.
41 . Bostom, A. G., L. A. Cupples, J. L. Jenner, J. M. Ordovas, L. J. Seman, P. W. Wilson, . . . W. P. Castelli (1996). "Elevated plasma lipoprotein(a) and coronary heart disease in men aged 55 years and younger. A prospective study." JAMA **276**(7): 544-548.
42 . Botstein, D. and N. Risch (2003). "Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease." Nat Genet **33 Suppl**: 228-237.
43 . Boycott, K. M., M. R. Vanstone, D. E. Bulman and A. E. MacKenzie (2013). "Rare-disease genetics in the era of next-generation sequencing: discovery to translation." Nat Rev Genet **14**(10): 681-691.
44 . British Cardiac, S., S. British Hypertension, U. K. Diabetes, U. K. Heart, S. Primary Care Cardiovascular and A. Stroke (2005). "JBS 2: Joint British Societies' guidelines on prevention of cardiovascular disease in clinical practice." Heart **91 Suppl 5**: v1-52.
45 . Brotman, D. J., E. Walker, M. S. Lauer and R. G. O'Brien (2005). "In search of fewer independent risk factors." Arch Intern Med **165**(2): 138-145.
46 . Brown, M. L., A. Inazu, C. B. Hesler, L. B. Agellon, C. Mann, M. E. Whitlock, . . . et al. (1989). "Molecular basis of lipid transfer protein deficiency in a family with increased high-density lipoproteins." Nature **342**(6248): 448-451.
47 . Brown, M. S. and J. L. Goldstein (1976). "Receptor-mediated control of cholesterol metabolism." Science **191**(4223): 150-154.
48 . Browning, B. L. and S. R. Browning (2009). "A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals." Am J Hum Genet **84**(2): 210-223.
49 . Brunner, E. J., M. Kivimaki, D. R. Witte, D. A. Lawlor, G. Davey Smith, J. A. Cooper, . . . M. Kumari (2008). "Inflammation, insulin resistance, and diabetes--Mendelian randomization using CRP haplotypes points upstream." PLoS Med **5**(8): e155.
50 . Buijsse, B., R. K. Simmons, S. J. Griffin and M. B. Schulze (2011). "Risk assessment tools for identifying individuals at risk of developing type 2 diabetes." Epidemiol Rev **33**(1): 46-62.
51 . Burkhardt, R., E. E. Kenny, J. K. Lowe, A. Birkeland, R. Josowitz, M. Noel, . . . J. L. Breslow (2008). "Common SNPs in HMGCR in micronesians and whites associated with LDL-cholesterol levels affect alternative splicing of exon13." Arterioscler Thromb Vasc Biol **28**(11): 2078-2084.
52 . Campbell, P. J., C. MacLean, P. A. Beer, G. Buck, K. Wheatley, J. J. Kiladjian, . . . A. R. Green (2012). "Correlation of blood counts with vascular complications in essential thrombocythemia: analysis of the prospective PT1 cohort." Blood **120**(7): 1409-1411.
53 . CARDIoGRAMplusC4D Consortium (2015). "A Comprehensive 1000 Genomes-based GWAS meta-analysis of Coronary Artery Disease." under review.
54 . Cartier, A., M. Cote, I. Lemieux, L. Perusse, A. Tremblay, C. Bouchard and J. P. Despres (2009). "Age-related differences in inflammatory markers in men: contribution of visceral adiposity." Metabolism **58**(10): 1452-1458.
55 . Casas, J. P., T. Shah, J. Cooper, E. Hawe, A. D. McMahon, D. Gaffney, . . . A. D. Hingorani (2006). "Insight into the nature of the CRP-coronary event association using Mendelian randomization." Int J Epidemiol **35**(4): 922-931.
56 . Castelli, W. P. (1988). "Cholesterol and lipids in the risk of coronary artery disease--the Framingham Heart Study." Can J Cardiol **4 Suppl A**: 5A-10A.

57 . Chambers, J. C., W. Zhang, Y. Li, J. Sehmi, M. N. Wass, D. Zabaneh, . . . J. S. Kooner (2009). "Genome-wide association study identifies variants in TMPRSS6 associated with hemoglobin levels." Nat Genet **41**(11): 1170-1172.
58 . Chapman, J. M., J. D. Cooper, J. A. Todd and D. G. Clayton (2003). "Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power." Hum Hered **56**(1-3): 18-31.
59 . Chasman, D. I., G. Pare, S. Mora, J. C. Hopewell, G. Peloso, R. Clarke, . . . P. M. Ridker (2009). "Forty-three loci associated with plasma lipoprotein size, concentration, and cholesterol content in genome-wide analysis." PLoS Genet **5**(11): e1000730.
60 . Chasman, D. I., G. Pare, R. Y. Zee, A. N. Parker, N. R. Cook, J. E. Buring, . . . P. M. Ridker (2008). "Genetic loci associated with plasma concentration of low-density lipoprotein cholesterol, high-density lipoprotein cholesterol, triglycerides, apolipoprotein A1, and Apolipoprotein B among 6382 white women in genome-wide analysis with replication." Circ Cardiovasc Genet **1**(1): 21-30.
61 . Chen, Z., H. Tang, R. Qayyum, U. M. Schick, M. A. Nalls, R. Handsaker, . . . A. P. Reiner (2013). "Genome-wide association analysis of red blood cell traits in African Americans: the COGENT Network." Hum Mol Genet **22**(12): 2529-2538.
62 . Cheng, T. L., Y. T. Wu, H. Y. Lin, F. C. Hsu, S. K. Liu, B. I. Chang, . . . H. L. Wu (2011). "Functions of rhomboid family protease RHBDL2 and thrombomodulin in wound healing." J Invest Dermatol **131**(12): 2486-2494.
63 . Choi, B. G., G. Vilahur, J. F. Viles-Gonzalez and J. J. Badimon (2006). "The role of high-density lipoprotein cholesterol in atherothrombosis." Mt Sinai J Med **73**(4): 690-701.
64 . Cholesterol Treatment Trialists, C., C. Baigent, L. Blackwell, J. Emberson, L. E. Holland, C. Reith, . . . R. Collins (2010). "Efficacy and safety of more intensive lowering of LDL cholesterol: a meta-analysis of data from 170,000 participants in 26 randomised trials." Lancet **376**(9753): 1670-1681.
65 . Cirulli, E. T. and D. B. Goldstein (2010). "Uncovering the roles of rare variants in common disease through whole-genome sequencing." Nat Rev Genet **11**(6): 415-425.
66 . Cladaras, C., M. Hadzopoulou-Cladaras, B. K. Felber, G. Pavlakis and V. I. Zannis (1987). "The molecular basis of a familial apoE deficiency. An acceptor splice site mutation in the third intron of the deficient apoE gene." J Biol Chem **262**(5): 2310-2315.
67 . Clarke, R., J. F. Peden, J. C. Hopewell, T. Kyriakou, A. Goel, S. C. Heath, . . . M. Farrall (2009). "Genetic variants associated with Lp(a) lipoprotein level and coronary disease." N Engl J Med **361**(26): 2518-2528.
68 . Coelho, H. C., S. C. Lopes, J. P. Pimentel, P. A. Nogueira, F. T. Costa, A. M. Siqueira, . . . M. V. Lacerda (2013). "Thrombocytopenia in Plasmodium vivax malaria is related to platelets phagocytosis." PLoS One **8**(5): e63410.
69 . Cohen, J., A. Pertsemlidis, I. K. Kotowski, R. Graham, C. K. Garcia and H. H. Hobbs (2005). "Low LDL cholesterol in individuals of African descent resulting from frequent nonsense mutations in PCSK9." Nat Genet **37**(2): 161-165.
70 . Cohen, J. C., E. Boerwinkle, T. H. Mosley, Jr. and H. H. Hobbs (2006). "Sequence variations in PCSK9, low LDL, and protection against coronary heart disease." N Engl J Med **354**(12): 1264-1272.
71 . Cohen, J. C., R. S. Kiss, A. Pertsemlidis, Y. L. Marcel, R. McPherson and H. H. Hobbs (2004). "Multiple rare alleles contribute to low plasma levels of HDL cholesterol." Science **305**(5685): 869-872.
72 . Companioni, O., F. Rodriguez Esparragon, A. M. Fernandez-Aceituno and J. C. Rodriguez Perez (2011). "[Genetic variants, cardiovascular risk and genome-wide association studies]." Rev Esp Cardiol **64**(6): 509-514.
73 . Consortium, C. A. D., P. Deloukas, S. Kanoni, C. Willenborg, M. Farrall, T. L. Assimes, . . . N. J. Samani (2013). "Large-scale association analysis identifies new risk loci for coronary artery disease." Nat Genet **45**(1): 25-33.

74 . Coram, M. A., Q. Duan, T. J. Hoffmann, T. Thornton, J. W. Knowles, N. A. Johnson, . . . H. Tang (2013). "Genome-wide characterization of shared and distinct genetic components that influence blood lipid levels in ethnically diverse human populations." Am J Hum Genet **92**(6): 904-916.
75 . Coronary Artery Disease Genetics, C. (2011). "A genome-wide association study in Europeans and South Asians identifies five new loci for coronary artery disease." Nat Genet **43**(4): 339-344.
76 . Coviello, A. D., R. Haring, M. Wellons, D. Vaidya, T. Lehtimaki, S. Keildson, . . . J. R. Perry (2012). "A genome-wide association meta-analysis of circulating sex hormone-binding globulin reveals multiple Loci implicated in sex steroid hormone regulation." PLoS Genet **8**(7): e1002805.
77 . Cox, D. W., W. C. Breckenridge and J. A. Little (1978). "Inheritance of apolipoprotein C-II deficiency with hypertriglyceridemia and pancreatitis." N Engl J Med **299**(26): 1421-1424.
78 . Crosslin, D. R., A. McDavid, N. Weston, S. C. Nelson, X. Zheng, E. Hart, . . . N. Genomics (2012). "Genetic variants associated with the white blood cell count in 13,923 subjects in the eMERGE Network." Hum Genet **131**(4): 639-652.
79 . Crosslin, D. R., A. McDavid, N. Weston, X. Zheng, E. Hart, M. de Andrade, . . . N. Genomics (2013). "Genetic variation associated with circulating monocyte count in the eMERGE Network." Hum Mol Genet **22**(10): 2119-2127.
80 . Cushman, M., A. M. Arnold, B. M. Psaty, T. A. Manolio, L. H. Kuller, G. L. Burke, . . . R. P. Tracy (2005). "C-reactive protein and the 10-year incidence of coronary heart disease in older men and women: the cardiovascular health study." Circulation **112**(1): 25-31.
81 . D'Agostino, R. B., Sr., R. S. Vasan, M. J. Pencina, P. A. Wolf, M. Cobain, J. M. Massaro and W. B. Kannel (2008). "General cardiovascular risk profile for use in primary care: the Framingham Heart Study." Circulation **117**(6): 743-753.
82 . Danecek, P., A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, . . . G. Genomes Project Analysis (2011). "The variant call format and VCFtools." Bioinformatics **27**(15): 2156-2158.
83 . Danesh, J., R. Collins, P. Appleby and R. Peto (1998). "Association of fibrinogen, C-reactive protein, albumin, or leukocyte count with coronary heart disease: meta-analyses of prospective studies." JAMA **279**(18): 1477-1482.
84 . Danesh, J., J. G. Wheeler, G. M. Hirschfield, S. Eda, G. Eiriksdottir, A. Rumley, . . . V. Gudnason (2004). "C-reactive protein and other circulating markers of inflammation in the prediction of coronary heart disease." N Engl J Med **350**(14): 1387-1397.
85 . Dawber, T. R., W. B. Kannel, N. Revotskie, J. Stokes, 3rd, A. Kagan and T. Gordon (1959). "Some factors associated with the development of coronary heart disease: six years' follow-up experience in the Framingham study." Am J Public Health Nations Health **49**: 1349-1356.
86 . Degoma, E. M. and D. J. Rader (2011). "Novel HDL-directed pharmacotherapeutic strategies." Nat Rev Cardiol **8**(5): 266-277.
87 . Dehghan, A., J. Dupuis, M. Barbalic, J. C. Bis, G. Eiriksdottir, C. Lu, . . . D. I. Chasman (2011). "Meta-analysis of genome-wide association studies in >80 000 subjects identifies multiple loci for C-reactive protein levels." Circulation **123**(7): 731-738.
88 . Delaneau, O., J. F. Zagury and J. Marchini (2013). "Improved whole-chromosome phasing for disease and population genetic studies." Nat Methods **10**(1): 5-6.
89 . Dendrou, C. A., V. Plagnol, E. Fung, J. H. Yang, K. Downes, J. D. Cooper, . . . L. S. Wicker (2009). "Cell-specific protein phenotypes for the autoimmune locus IL2RA using a genotype-selectable human bioresource." Nat Genet **41**(9): 1011-1015.
90 . Ding, K., M. de Andrade, T. A. Manolio, D. C. Crawford, L. J. Rasmussen-Torvik, M. D. Ritchie, . . . I. J. Kullo (2013). "Genetic variants that confer resistance to malaria are associated with red blood cell traits in African-Americans: an electronic medical record-based genome-wide association study." G3 (Bethesda) **3**(7): 1061-1068.
91 . Do, R., N. O. Stitziel, H. Won, A. B. Jorgensen, S. Duga, P. Angelica Merlini, . . . S. Kathiresan (2014). "Exome sequencing identifies rare LDLR and APOA5 alleles conferring risk for myocardial infarction." Nature.

92 . Do, R., N. O. Stitziel, H. H. Won, A. B. Jorgensen, S. Duga, P. Angelica Merlini, . . . S. Kathiresan (2015). "Exome sequencing identifies rare LDLR and APOA5 alleles conferring risk for myocardial infarction." Nature **518**(7537): 102-106.
93 . Do, R., C. J. Willer, E. M. Schmidt, S. Sengupta, C. Gao, G. M. Peloso, . . . S. Kathiresan (2013). "Common variants associated with plasma triglycerides and risk for coronary artery disease." Nat Genet **45**(11): 1345-1352.
94 . Dorajoo, R., R. Li, M. K. Ikram, J. Liu, P. Froguel, J. Lee, . . . Y. Friedlander (2013). "Are C-reactive protein associated genetic variants associated with serum levels and retinal markers of microvascular pathology in Asian populations from Singapore?" PLoS One **8**(7): e67650.
95 . Doumatey, A. P., G. Chen, F. Tekola Ayele, J. Zhou, M. Erdos, D. Shriner, . . . C. N. Rotimi (2012). "C-reactive protein (CRP) promoter polymorphisms influence circulating CRP levels in a genome-wide association study of African Americans." Hum Mol Genet **21**(13): 3063-3072.
96 . Downs, J. R., M. Clearfield, S. Weis, E. Whitney, D. R. Shapiro, P. A. Beere, . . . A. M. Gotto, Jr. (1998). "Primary prevention of acute coronary events with lovastatin in men and women with average cholesterol levels: results of AFCAPS/TexCAPS. Air Force/Texas Coronary Atherosclerosis Prevention Study." JAMA **279**(20): 1615-1622.
97 . Duewell, P., H. Kono, K. J. Rayner, C. M. Sirois, G. Vladimer, F. G. Bauernfeind, . . . E. Latz (2010). "NLRP3 inflammasomes are required for atherogenesis and activated by cholesterol crystals." Nature **464**(7293): 1357-1361.
98 . Dzau, V. and E. Braunwald (1991). "Resolved and unresolved issues in the prevention and treatment of coronary artery disease: a workshop consensus statement." Am Heart J **121**(4 Pt 1): 1244-1263.
99 . Dzau, V. J., E. M. Antman, H. R. Black, D. L. Hayes, J. E. Manson, J. Plutzky, . . . W. Stevenson (2006). "The cardiovascular disease continuum validated: clinical evidence of improved patient outcomes: part II: Clinical trial evidence (acute coronary syndromes through renal disease) and future directions." Circulation **114**(25): 2871-2891.
100 . Ehret, G. B., P. B. Munroe, K. M. Rice, M. Bochud, A. D. Johnson, D. I. Chasman, . . . L. Lightstone (2011). "Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk." Nature **478**(7367): 103-109.
101 . Eichler, E. E., J. Flint, G. Gibson, A. Kong, S. M. Leal, J. H. Moore and J. H. Nadeau (2010). "Missing heritability and strategies for finding the underlying causes of complex disease." Nat Rev Genet **11**(6): 446-450.
102 . Elliott, P., J. C. Chambers, W. Zhang, R. Clarke, J. C. Hopewell, J. F. Peden, . . . J. S. Kooner (2009). "Genetic Loci associated with C-reactive protein levels and risk of coronary heart disease." JAMA **302**(1): 37-48.
103 . Ellis, J., E. M. Lange, J. Li, J. Dupuis, J. Baumert, J. D. Walston, . . . L. A. Lange (2014). "Large multiethnic Candidate Gene Study for C-reactive protein levels: identification of a novel association at CD36 in African Americans." Hum Genet **133**(8): 985-995.
104 . Elston, R. C. and J. Stewart (1971). "A general model for the genetic analysis of pedigree data." Hum Hered **21**(6): 523-542.
105 . Emerging Risk Factors, C., E. Di Angelantonio, N. Sarwar, P. Perry, S. Kaptoge, K. K. Ray, . . . J. Danesh (2009). "Major lipids, apolipoproteins, and risk of vascular disease." JAMA **302**(18): 1993-2000.
106 . Emerging Risk Factors, C., S. Kaptoge, E. Di Angelantonio, G. Lowe, M. B. Pepys, S. G. Thompson, . . . J. Danesh (2010). "C-reactive protein concentration and risk of coronary heart disease, stroke, and mortality: an individual participant meta-analysis." Lancet **375**(9709): 132-140.
107 . Emi, M., D. E. Wilson, P. H. Iverius, L. Wu, A. Hata, R. Hegele, . . . J. M. Lalouel (1990). "Missense mutation (Gly----Glu188) of human lipoprotein lipase imparting functional deficiency." J Biol Chem **265**(10): 5910-5916.

108 . Endo, A., M. Kuroda and K. Tanzawa (1976). "Competitive inhibition of 3-hydroxy-3-methylglutaryl coenzyme A reductase by ML-236A and ML-236B fungal metabolites, having hypocholesterolemic activity." FEBS Lett **72**(2): 323-326.
109 . Esko, T., M. Mezzavilla, M. Nelis, C. Borel, T. Debniak, E. Jakkula, . . . P. D'Adamo (2013). "Genetic characterization of northeastern Italian population isolates in the context of broader European genetic diversity." Eur J Hum Genet **21**(6): 659-665.
110 . Expert Panel on Detection, E. and A. Treatment of High Blood Cholesterol in (2001). "Executive Summary of The Third Report of The National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, And Treatment of High Blood Cholesterol In Adults (Adult Treatment Panel III)." JAMA **285**(19): 2486-2497.
111 . Feinleib, M., W. B. Kannel, R. J. Garrison, P. M. McNamara and W. P. Castelli (1975). "The Framingham Offspring Study. Design and preliminary data." Prev Med **4**(4): 518-525.
112 . Feng, S., D. Liu, X. Zhan, M. K. Wing and G. R. Abecasis (2014). "RAREMETAL: fast and powerful meta-analysis for rare variants." Bioinformatics.
113 . Ferreira, M. A., J. J. Hottenga, N. M. Warrington, S. E. Medland, G. Willemsen, R. W. Lawrence, . . . D. I. Boomsma (2009). "Sequence variants in three loci influence monocyte counts and erythrocyte volume." Am J Hum Genet **85**(5): 745-749.
114 . Fischer, M., U. Broeckel, S. Holmer, A. Baessler, C. Hengstenberg, B. Mayer, . . . H. Schunkert (2005). "Distinct heritable patterns of angiographic coronary artery disease in families with myocardial infarction." Circulation **111**(7): 855-862.
115 . Flannick, J., J. M. Korn, P. Fontanillas, G. B. Grant, E. Banks, M. A. Depristo and D. Altshuler (2012). "Efficiency and power as a function of sequence coverage, SNP array density, and imputation." PLoS Comput Biol **8**(7): e1002604.
116 . Folsom, A. R., L. E. Chambless, C. M. Ballantyne, J. Coresh, G. Heiss, K. K. Wu, . . . A. R. Sharrett (2006). "An assessment of incremental coronary risk prediction using C-reactive protein and other novel risk markers: the atherosclerosis risk in communities study." Arch Intern Med **166**(13): 1368-1373.
117 . Fredrickson, D. S. and R. S. Lees (1965). "A System for Phenotyping Hyperlipoproteinemia." Circulation **31**: 321-327.
118 . Friedewald, W. T., R. I. Levy and D. S. Fredrickson (1972). "Estimation of the concentration of low-density lipoprotein cholesterol in plasma, without use of the preparative ultracentrifuge." Clin Chem **18**(6): 499-502.
119 . Frikke-Schmidt, R., B. G. Nordestgaard, M. C. Stene, A. A. Sethi, A. T. Remaley, P. Schnohr, . . . A. Tybjaerg-Hansen (2008). "Association of loss-of-function mutations in the ABCA1 gene with high-density lipoprotein cholesterol levels and risk of ischemic heart disease." JAMA **299**(21): 2524-2532.
120 . Funke, H., A. von Eckardstein, P. H. Pritchard, J. J. Albers, J. J. Kastelein, C. Droste and G. Assmann (1991). "A molecular defect causing fish eye disease: an amino acid exchange in lecithin-cholesterol acyltransferase (LCAT) leads to the selective loss of alpha-LCAT activity." Proc Natl Acad Sci U S A **88**(11): 4855-4859.
121 . Gambaro, G., T. Yabarek, M. S. Graziani, A. Gemelli, C. Abaterusso, A. C. Frigo, . . . I. S. Group (2010). "Prevalence of CKD in northeastern Italy: results of the INCIPE study and comparison with NHANES." Clin J Am Soc Nephrol **5**(11): 1946-1953.
122 . Ganesh, S. K., N. A. Zakai, F. J. van Rooij, N. Soranzo, A. V. Smith, M. A. Nalls, . . . J. P. Lin (2009). "Multiple loci influence erythrocyte phenotypes in the CHARGE Consortium." Nat Genet **41**(11): 1191-1198.
123 . Garcia, C. K., K. Wilund, M. Arca, G. Zuliani, R. Fellin, M. Maioli, . . . H. H. Hobbs (2001). "Autosomal recessive hypercholesterolemia caused by mutations in a putative LDL receptor adaptor protein." Science **292**(5520): 1394-1398.
124 . Garner, C., T. Tatu, J. E. Reittie, T. Littlewood, J. Darley, S. Cervino, . . . S. L. Thein (2000). "Genetic influences on F cells and other hematologic variables: a twin heritability study." Blood **95**(1): 342-346.

125 . Gieger, C., A. Radhakrishnan, A. Cvejic, W. Tang, E. Porcu, G. Pistis, . . . N. Soranzo (2011). "New gene functions in megakaryopoiesis and platelet formation." Nature **480**(7376): 201-208.
126 . Glessner, J. T., K. Wang, G. Cai, O. Korvatska, C. E. Kim, S. Wood, . . . H. Hakonarson (2009). "Autism genome-wide copy number variation reveals ubiquitin and neuronal genes." Nature **459**(7246): 569-573.
127 . Global Lipids Genetics, C., C. J. Willer, E. M. Schmidt, S. Sengupta, G. M. Peloso, S. Gustafsson, . . . G. R. Abecasis (2013). "Discovery and refinement of loci associated with lipid levels." Nat Genet **45**(11): 1274-1283.
128 . Glud, T., E. B. Schmidt, S. D. Kristensen and T. Arnfred (1986). "Platelet number and volume during myocardial infarction in relation to infarct size." Acta Med Scand **220**(5): 401-405.
129 . Goate, A., M. C. Chartier-Harlin, M. Mullan, J. Brown, F. Crawford, L. Fidani, . . . et al. (1991). "Segregation of a missense mutation in the amyloid precursor protein gene with familial Alzheimer's disease." Nature **349**(6311): 704-706.
130 . Goldbourt, U., S. Yaari and J. H. Medalie (1997). "Isolated low HDL cholesterol as a risk factor for coronary heart disease mortality. A 21-year follow-up of 8000 men." Arterioscler Thromb Vasc Biol **17**(1): 107-113.
131 . Golding, J., M. Pembrey and R. Jones (2001). "ALSPAC--the Avon Longitudinal Study of Parents and Children. I. Study methodology." Paediatr Perinat Epidemiol **15**(1): 74-87.
132 . Golding, J., M. Pembrey, R. Jones and A. S. Team (2001). "ALSPAC--the Avon Longitudinal Study of Parents and Children. I. Study methodology." Paediatr Perinat Epidemiol **15**(1): 74-87.
133 . Goldstein, D. B. (2009). "Common genetic variation and human traits." N Engl J Med **360**(17): 1696-1698.
134 . Goldstein, D. B., A. Allen, J. Keebler, E. H. Margulies, S. Petrou, S. Petrovski and S. Sunyaev (2013). "Sequencing studies in human genetics: design and interpretation." Nat Rev Genet **14**(7): 460-470.
135 . Goode, E. L., S. S. Cherny, J. C. Christian, G. P. Jarvik and M. de Andrade (2007). "Heritability of longitudinal measures of body mass index and lipid and lipoprotein levels in aging twins." Twin Res Hum Genet **10**(5): 703-711.
136 . Gordon, D. J., J. L. Probstfield, R. J. Garrison, J. D. Neaton, W. P. Castelli, J. D. Knoke, . . . H. A. Tyroler (1989). "High-density lipoprotein cholesterol and cardiovascular disease. Four prospective American studies." Circulation **79**(1): 8-15.
137 . Gordon, T., W. P. Castelli, M. C. Hjortland, W. B. Kannel and T. R. Dawber (1977). "High density lipoprotein as a protective factor against coronary heart disease. The Framingham Study." Am J Med **62**(5): 707-714.
138 . Graham, I., D. Atar, K. Borch-Johnsen, G. Boysen, G. Burell, R. Cifkova, . . . G. European Society of Cardiology Committee for Practice (2007). "European guidelines on cardiovascular disease prevention in clinical practice: executive summary: Fourth Joint Task Force of the European Society of Cardiology and Other Societies on Cardiovascular Disease Prevention in Clinical Practice (Constituted by representatives of nine societies and by invited experts)." Eur Heart J **28**(19): 2375-2414.
139 . Greenburg, A. G. (1996). "Pathophysiology of anemia." Am J Med **101**(2A): 7S-11S.
140 . Haase, C. L., A. Tybjaerg-Hansen, A. A. Qayyum, J. Schou, B. G. Nordestgaard and R. Frikke-Schmidt (2012). "LCAT, HDL cholesterol and ischemic cardiovascular disease: a Mendelian randomization study of HDL cholesterol in 54,500 individuals." J Clin Endocrinol Metab **97**(2): E248-256.
141 . Haines, J. L., M. A. Hauser, S. Schmidt, W. K. Scott, L. M. Olson, P. Gallins, . . . M. A. Pericak-Vance (2005). "Complement factor H variant increases the risk of age-related macular degeneration." Science **308**(5720): 419-421.
142 . Hardison, R. C. and G. A. Blobel (2013). "Genetics. GWAS to therapy by genome edits?" Science **342**(6155): 206-207.

143 . Harrow, J., A. Frankish, J. M. Gonzalez, E. Tapanari, M. Diekhans, F. Kokocinski, . . . T. J. Hubbard (2012). "GENCODE: the reference human genome annotation for The ENCODE Project." Genome Res **22**(9): 1760-1774.
144 . Hays, J., J. R. Hunt, F. A. Hubbell, G. L. Anderson, M. Limacher, C. Allen and J. E. Rossouw (2003). "The Women's Health Initiative recruitment methods and results." Ann Epidemiol **13**(9 Suppl): S18-77.
145 . He, Z., B. J. O'Roak, J. D. Smith, G. Wang, S. Hooker, R. L. Santos-Cortez, . . . S. M. Leal (2014). "Rare-variant extensions of the transmission disequilibrium test: application to autism exome sequence data." Am J Hum Genet **94**(1): 33-46.
146 . Heart Protection Study Collaborative, G. (2002). "MRC/BHF Heart Protection Study of cholesterol lowering with simvastatin in 20,536 high-risk individuals: a randomised placebo-controlled trial." Lancet **360**(9326): 7-22.
147 . Hegele, R. A., J. A. Little and P. W. Connelly (1991). "Compound heterozygosity for mutant hepatic lipase in familial hepatic lipase deficiency." Biochem Biophys Res Commun **179**(1): 78-84.
148 . Heid, I. M., E. Boes, M. Muller, B. Kollerits, C. Lamina, S. Coassin, . . . F. Kronenberg (2008). "Genome-wide association analysis of high-density lipoprotein cholesterol in the population-based KORA study sheds new light on intergenic regions." Circ Cardiovasc Genet **1**(1): 10-20.
149 . Helfand, M., D. I. Buckley, M. Freeman, R. Fu, K. Rogers, C. Fleming and L. L. Humphrey (2009). "Emerging risk factors for coronary heart disease: a summary of systematic reviews conducted for the U.S. Preventive Services Task Force." Ann Intern Med **151**(7): 496-507.
150 . Hemani, G., J. Yang, A. Vinkhuyzen, J. E. Powell, G. Willemsen, J. J. Hottenga, . . . P. M. Visscher (2013). "Inference of the genetic architecture underlying BMI and height with the use of 20,240 sibling pairs." Am J Hum Genet **93**(5): 865-875.
151 . Hendra, T. J., G. A. Oswald and J. S. Yudkin (1988). "Increased mean platelet volume after acute myocardial infarction relates to diabetes and to cardiac failure." Diabetes Res Clin Pract **5**(1): 63-69.
152 . Hindorff, L. A., P. Sethupathy, H. A. Junkins, E. M. Ramos, J. P. Mehta, F. S. Collins and T. A. Manolio (2009). "Potential etiologic and functional implications of genome-wide association loci for human diseases and traits." Proc Natl Acad Sci U S A **106**(23): 9362-9367.
153 . Hirschhorn, J. N. and M. J. Daly (2005). "Genome-wide association studies for common diseases and complex traits." Nat Rev Genet **6**(2): 95-108.
154 . Hiura, Y., C. S. Shen, Y. Kokubo, T. Okamura, T. Morisaki, H. Tomoike, . . . N. Iwai (2009). "Identification of genetic markers associated with high-density lipoprotein-cholesterol by genome-wide screening in a Japanese population: the Suita study." Circ J **73**(6): 1119-1126.
155 . Hoffman, M., A. Blum, R. Baruch, E. Kaplan and M. Benjamin (2004). "Leukocytes and coronary heart disease." Atherosclerosis **172**(1): 1-6.
156 . Holm, H., D. F. Gudbjartsson, P. Sulem, G. Masson, H. T. Helgadottir, C. Zanon, . . . K. Stefansson (2011). "A rare variant in MYH6 is associated with high risk of sick sinus syndrome." Nat Genet **43**(4): 316-320.
157 . Holmen, O. L., H. Zhang, Y. Fan, D. H. Hovelson, E. M. Schmidt, W. Zhou, . . . C. J. Willer (2014). "Systematic evaluation of coding variation identifies a candidate causal variant in TM6SF2 influencing total cholesterol and myocardial infarction risk." Nat Genet **46**(4): 345-351.
158 . Holmen, O. L., H. Zhang, W. Zhou, E. Schmidt, D. H. Hovelson, A. Langhammer, . . . C. J. Willer (2014). "No large-effect low-frequency coding variation found for myocardial infarction." Hum Mol Genet **23**(17): 4721-4728.
159 . Howie, B., C. Fuchsberger, M. Stephens, J. Marchini and G. R. Abecasis (2012). "Fast and accurate genotype imputation in genome-wide association studies through pre-phasing." Nat Genet **44**(8): 955-959.
160 . Howie, B., J. Marchini and M. Stephens (2011). "Genotype imputation with thousands of genomes." G3 (Bethesda) **1**(6): 457-470.
161 . Howie, B. N., P. Donnelly and J. Marchini (2009). "A flexible and accurate genotype imputation method for the next generation of genome-wide association studies." PLoS Genet **5**(6): e1000529.

162 . Huang, J., D. Ellinghaus, A. Franke, B. Howie and Y. Li (2012). "1000 Genomes-based imputation identifies novel and refined associations for the Wellcome Trust Case Control Consortium phase 1 Data." Eur J Hum Genet.
163 . Huang, J., B. Howie, S. McCarthy, Y. Memari, K. Walter, J. Min, . . . N. Soranzo (2015). "A reference panel of 3,781 genomes from the UK10K Project increases imputation performance of low frequency variants." Nature Communications (Under peer review).
164 . Huang, J., A. D. Johnson and C. J. O'Donnell (2011). "PRIMe: a method for characterization and evaluation of pleiotropic regions from multiple genome-wide association studies." Bioinformatics **27**(9): 1201-1206.
165 . Huang, J., R. H. Perlis, P. H. Lee, A. J. Rush, M. Fava, G. S. Sachs, . . . J. W. Smoller (2010). "Cross-disorder genomewide analysis of schizophrenia, bipolar disorder, and depression." Am J Psychiatry **167**(10): 1254-1263.
166 . Hunink, M. G., L. Goldman, A. N. Tosteson, M. A. Mittleman, P. A. Goldman, L. W. Williams, . . . M. C. Weinstein (1997). "The recent decline in mortality from coronary heart disease, 1980-1990. The effect of secular trends in risk factors and treatment." JAMA **277**(7): 535-542.
167 . Hunt, K. A., V. Mistry, N. A. Bockett, T. Ahmad, M. Ban, J. N. Barker, . . . D. A. van Heel (2013). "Negligible impact of rare autoimmune-locus coding-region variants on missing heritability." Nature **498**(7453): 232-235.
168 . Ibanez, B., G. Vilahur and J. J. Badimon (2007). "Plaque progression and regression in atherothrombosis." J Thromb Haemost **5 Suppl 1**: 292-299.
169 . Idaghdour, Y., J. Quinlan, J. P. Goulet, J. Berghout, E. Gbeha, V. Bruat, . . . P. Awadalla (2012). "Evidence for additive and interaction effects of host genotype and infection in malaria." Proc Natl Acad Sci U S A **109**(42): 16786-16793.
170 . Igl, W., A. Johansson, J. F. Wilson, S. H. Wild, O. Polasek, C. Hayward, . . . E. Consortium (2010). "Modeling of environmental effects in genome-wide association studies identifies SLC2A2 and HP as novel loci influencing serum cholesterol levels." PLoS Genet **6**(1): e1000798.
171 . Ingelsson, E., E. J. Schaefer, J. H. Contois, J. R. McNamara, L. Sullivan, M. J. Keyes, . . . R. S. Vasan (2007). "Clinical utility of different lipid measures for prediction of coronary heart disease in men and women." JAMA **298**(7): 776-785.
172 . Interleukin-6 Receptor Mendelian Randomisation Analysis, C., A. D. Hingorani and J. P. Casas (2012). "The interleukin-6 receptor as a target for prevention of coronary heart disease: a mendelian randomisation analysis." Lancet **379**(9822): 1214-1224.
173 . International HapMap, C., D. M. Altshuler, R. A. Gibbs, L. Peltonen, D. M. Altshuler, R. A. Gibbs, . . . J. E. McEwen (2010). "Integrating common and rare genetic variation in diverse human populations." Nature **467**(7311): 52-58.
174 . International HapMap, C., K. A. Frazer, D. G. Ballinger, D. R. Cox, D. A. Hinds, L. L. Stuve, . . . J. Stewart (2007). "A second generation human haplotype map of over 3.1 million SNPs." Nature **449**(7164): 851-861.
175 . International Schizophrenia, C. (2008). "Rare chromosomal deletions and duplications increase risk of schizophrenia." Nature **455**(7210): 237-241.
176 . International Schizophrenia, C., S. M. Purcell, N. R. Wray, J. L. Stone, P. M. Visscher, M. C. O'Donovan, . . . P. Sklar (2009). "Common polygenic variation contributes to risk of schizophrenia and bipolar disorder." Nature **460**(7256): 748-752.
177 . Jallow, M., Y. Y. Teo, K. S. Small, K. A. Rockett, P. Deloukas, T. G. Clark, . . . N. Malaria Genomic Epidemiology (2009). "Genome-wide and fine-resolution association analysis of malaria in West Africa." Nat Genet **41**(6): 657-665.
178 . Jewett, E. M., M. Zawistowski, N. A. Rosenberg and S. Zollner (2012). "A coalescent model for genotype imputation." Genetics **191**(4): 1239-1255.
179 . Johannsen, T. H., P. R. Kamstrup, R. V. Andersen, G. B. Jensen, H. Sillesen, A. Tybjaerg-Hansen and B. G. Nordestgaard (2009). "Hepatic lipase, genetically elevated high-density lipoprotein, and risk of ischemic cardiovascular disease." J Clin Endocrinol Metab **94**(4): 1264-1273.

180 . Johansen, C. T., J. Wang, M. B. Lanktree, H. Cao, A. D. McIntyre, M. R. Ban, . . . R. A. Hegele (2010). "Excess of rare variants in genes identified by genome-wide association study of hypertriglyceridemia." Nat Genet **42**(8): 684-687.
181 . Jones, B., E. L. Jones, S. A. Bonney, H. N. Patel, A. R. Mensenkamp, S. Eichenbaum-Voline, . . . C. C. Shoulders (2003). "Mutations in a Sar1 GTPase of COPII vesicles are associated with lipid absorption disorders." Nat Genet **34**(1): 29-31.
182 . Jorgensen, A. B., R. Frikke-Schmidt, B. G. Nordestgaard and A. Tybjaerg-Hansen (2014). "Loss-of-Function Mutations in APOC3 and Risk of Ischemic Vascular Disease." N Engl J Med **371**(1): 32-41.
183 . Kamatani, Y., K. Matsuda, Y. Okada, M. Kubo, N. Hosono, Y. Daigo, . . . N. Kamatani (2010). "Genome-wide association study of hematological and biochemical traits in a Japanese population." Nat Genet **42**(3): 210-215.
184 . Kannel, W. B., K. Anderson and P. W. Wilson (1992). "White blood cell count and cardiovascular disease. Insights from the Framingham Study." JAMA **267**(9): 1253-1256.
185 . Kannel, W. B., T. R. Dawber, G. D. Friedman, W. E. Glennon and P. M. McNamara (1964). "Risk Factors in Coronary Heart Disease. An Evaluation of Several Serum Lipids as Predictors of Coronary Heart Disease; the Framingham Study." Ann Intern Med **61**: 888-899.
186 . Kannel, W. B., T. R. Dawber, A. Kagan, N. Revotskie and J. Stokes, 3rd (1961). "Factors of risk in the development of coronary heart disease--six year follow-up experience. The Framingham Study." Ann Intern Med **55**: 33-50.
187 . Kannel, W. B., T. R. Dawber and D. L. McGee (1980). "Perspectives on systolic hypertension. The Framingham study." Circulation **61**(6): 1179-1182.
188 . Kannel, W. B., R. S. Vasan, M. J. Keyes, L. M. Sullivan and S. J. Robins (2008). "Usefulness of the triglyceride-high-density lipoprotein versus the cholesterol-high-density lipoprotein ratio for predicting insulin resistance and cardiometabolic risk (from the Framingham Offspring Cohort)." Am J Cardiol **101**(4): 497-501.
189 . Kannel, W. B., P. A. Wolf, W. P. Castelli and R. B. D'Agostino (1987). "Fibrinogen and risk of cardiovascular disease. The Framingham Study." JAMA **258**(9): 1183-1186.
190 . Kathiresan, S., A. K. Manning, S. Demissie, R. B. D'Agostino, A. Surti, C. Guiducci, . . . L. A. Cupples (2007). "A genome-wide association study for blood lipid phenotypes in the Framingham Heart Study." BMC Med Genet **8 Suppl 1**: S17.
191 . Kathiresan, S., O. Melander, C. Guiducci, A. Surti, N. P. Burtt, M. J. Rieder, . . . M. Orho-Melander (2008). "Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans." Nat Genet **40**(2): 189-197.
192 . Kathiresan, S. and D. Srivastava (2012). "Genetics of human cardiovascular disease." Cell **148**(6): 1242-1257.
193 . Kathiresan, S., C. J. Willer, G. M. Peloso, S. Demissie, K. Musunuru, E. E. Schadt, . . . L. A. Cupples (2009). "Common variants at 30 loci contribute to polygenic dyslipidemia." Nat Genet **41**(1): 56-65.
194 . Keller, M., D. Schleinitz, J. Forster, A. Tonjes, Y. Bottcher, A. Fischer-Rosinsky, . . . P. Kovacs (2013). "THOC5: a novel gene involved in HDL-cholesterol metabolism." J Lipid Res **54**(11): 3170-3176.
195 . Keller, M. F., A. P. Reiner, Y. Okada, F. J. van Rooij, A. D. Johnson, M. H. Chen, . . . G. BioBank Japan Project Working (2014). "Trans-ethnic meta-analysis of white blood cell phenotypes." Hum Mol Genet **23**(25): 6944-6960.
196 . Kerem, B., J. M. Rommens, J. A. Buchanan, D. Markiewicz, T. K. Cox, A. Chakravarti, . . . L. C. Tsui (1989). "Identification of the cystic fibrosis gene: genetic analysis." Science **245**(4922): 1073-1080.
197 . Keskin, O., R. E. Ulusoy, M. Kalemoglu, M. H. Us, I. Yildirim, O. Tarcin, . . . N. Ardic (2004). "White blood cell count and C-reactive protein predict short-term prognosis in acute myocardial infarction." J Int Med Res **32**(6): 646-654.

198 . Kettunen, J., T. Tukiainen, A. P. Sarin, A. Ortega-Alonso, E. Tikkanen, L. P. Lyytikainen, . . . S. Ripatti (2012). "Genome-wide association study identifies multiple loci influencing human serum metabolite levels." Nat Genet **44**(3): 269-276.
199 . Kim, S., S. Swaminathan, L. Shen, S. L. Risacher, K. Nho, T. Foroud, . . . I. Alzheimer's Disease Neuroimaging (2011). "Genome-wide association study of CSF biomarkers Abeta1-42, t-tau, and p-tau181p in the ADNI cohort." Neurology **76**(1): 69-79.
200 . Kim, S. Y., J. P. Guevara, K. M. Kim, H. K. Choi, D. F. Heitjan and D. A. Albert (2010). "Hyperuricemia and coronary heart disease: a systematic review and meta-analysis." Arthritis Care Res (Hoboken) **62**(2): 170-180.
201 . Kim, Y. J., M. J. Go, C. Hu, C. B. Hong, Y. K. Kim, J. Y. Lee, . . . Y. S. Cho (2011). "Large-scale genome-wide association studies in East Asians identify new genetic loci influencing metabolic traits." Nat Genet **43**(10): 990-995.
202 . Klein, R. J., C. Zeiss, E. Y. Chew, J. Y. Tsai, R. S. Sackler, C. Haynes, . . . J. Hoh (2005). "Complement factor H polymorphism in age-related macular degeneration." Science **308**(5720): 385-389.
203 . Koenig, W., H. Lowel, J. Baumert and C. Meisinger (2004). "C-reactive protein modulates risk prediction based on the Framingham Score: implications for future risk assessment: results from a large cohort study in southern Germany." Circulation **109**(11): 1349-1353.
204 . Kong, M. and C. Lee (2013). "Genetic associations with C-reactive protein level and white blood cell count in the KARE study." Int J Immunogenet **40**(2): 120-125.
205 . Kooner, J. S., J. C. Chambers, C. A. Aguilar-Salinas, D. A. Hinds, C. L. Hyde, G. R. Warnes, . . . J. F. Thompson (2008). "Genome-wide scan identifies variation in MLXIPL associated with plasma triglycerides." Nat Genet **40**(2): 149-151.
206 . Kuivenhoven, J. A. and R. A. Hegele (2014). "Mining the genome for lipid genes." Biochim Biophys Acta **1842**(10): 1993-2009.
207 . Kuller, L. H. (1976). "Epidemiology of cardiovascular diseases: current perspectives." Am J Epidemiol **104**(4): 425-496.
208 . Kullo, I. J., K. Ding, H. Jouni, C. Y. Smith and C. G. Chute (2010). "A genome-wide association study of red blood cell traits using the electronic medical record." PLoS One **5**(9).
209 . Kuroda, M., Y. Tsujita, K. Tanzawa and A. Endo (1979). "Hypolipidemic effects in monkeys of ML-236B, a competitive inhibitor of 3-hydroxy-3-methylglutaryl coenzyme A reductase." Lipids **14**(6): 585-589.
210 . Kwiatkowski, D. P. (2005). "How malaria has affected the human genome and what human genetics can teach us about malaria." Am J Hum Genet **77**(2): 171-192.
211 . Labreuche, J., P. J. Touboul and P. Amarenco (2009). "Plasma triglyceride levels and risk of stroke and carotid atherosclerosis: a systematic review of the epidemiological studies." Atherosclerosis **203**(2): 331-345.
212 . Ladouceur, M., Z. Dastani, Y. S. Aulchenko, C. M. Greenwood and J. B. Richards (2012). "The empirical power of rare variant association methods: results from sanger sequencing in 1,998 individuals." PLoS Genet **8**(2): e1002496.
213 . Lage, K., E. O. Karlberg, Z. M. Storling, P. I. Olason, A. G. Pedersen, O. Rigina, . . . S. Brunak (2007). "A human phenome-interactome network of protein complexes implicated in genetic disorders." Nat Biotechnol **25**(3): 309-316.
214 . LaMonte, G., N. Philip, J. Reardon, J. R. Lacsina, W. Majoros, L. Chapman, . . . J. T. Chi (2012). "Translocation of sickle cell erythrocyte microRNAs into Plasmodium falciparum inhibits parasite translation and contributes to malaria resistance." Cell Host Microbe **12**(2): 187-199.
215 . Lander, E. S. (1996). "The new genomics: global views of biology." Science **274**(5287): 536-539.
216 . Lander, E. S. and P. Green (1987). "Construction of multilocus genetic linkage maps in humans." Proc Natl Acad Sci U S A **84**(8): 2363-2367.

217 . Lander, E. S., L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, . . . C. International Human Genome Sequencing (2001). "Initial sequencing and analysis of the human genome." Nature **409**(6822): 860-921.
218 . Langaee, T. and M. Ronaghi (2005). "Genetic variation analyses by Pyrosequencing." Mutat Res **573**(1-2): 96-102.
219 . Lange, L. A., Y. Hu, H. Zhang, C. Xue, E. M. Schmidt, Z. Z. Tang, . . . N. G. O. E. S. Project (2014). "Whole-exome sequencing identifies rare and low-frequency coding variants associated with LDL cholesterol." Am J Hum Genet **94**(2): 233-245.
220 . Lango Allen, H., K. Estrada, G. Lettre, S. I. Berndt, M. N. Weedon, F. Rivadeneira, . . . J. N. Hirschhorn (2010). "Hundreds of variants clustered in genomic loci and biological pathways affect human height." Nature **467**(7317): 832-838.
221 . Lanzara, C., A. d'Adamo and M. Montico (2015). "Use of an Italian isolated population for studying complex diseases. The Carlantino project: study design and preliminary results. ." Slovenian J Pub Health(in press).
222 . Lavie, C. J. and R. V. Milani (2003). "Obesity and cardiovascular disease: the hippocrates paradox?" J Am Coll Cardiol **42**(4): 677-679.
223 . Lawlor, D. A., R. M. Harbord, N. J. Timpson, G. D. Lowe, A. Rumley, T. R. Gaunt, . . . G. D. Smith (2008). "The association of C-reactive protein and CRP genotype with coronary heart disease: findings from five studies with 4,610 cases amongst 18,637 participants." PLoS One **3**(8): e3011.
224 . Le, S. Q. and R. Durbin (2011). "SNP detection and genotyping from low-coverage sequencing data on multiple diploid samples." Genome Res **21**(6): 952-960.
225 . Lee, S., M. J. Emond, M. J. Bamshad, K. C. Barnes, M. J. Rieder, D. A. Nickerson, . . . X. Lin (2012). "Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies." Am J Hum Genet **91**(2): 224-237.
226 . Lee, S., M. C. Wu and X. Lin (2012). "Optimal tests for rare variant effects in sequencing association studies." Biostatistics **13**(4): 762-775.
227 . Lehrman, M. A., J. L. Goldstein, M. S. Brown, D. W. Russell and W. J. Schneider (1985). "Internalization-defective LDL receptors produced by genes with nonsense and frameshift mutations that truncate the cytoplasmic domain." Cell **41**(3): 735-743.
228 . Lemieux, I., B. Lamarche, C. Couillard, A. Pascot, B. Cantin, J. Bergeron, . . . J. P. Despres (2001). "Total cholesterol/HDL cholesterol ratio vs LDL cholesterol/HDL cholesterol ratio as indices of ischemic heart disease risk in men: the Quebec Cardiovascular Study." Arch Intern Med **161**(22): 2685-2692.
229 . Lewis, G. F. and D. J. Rader (2005). "New insights into the regulation of HDL metabolism and reverse cholesterol transport." Circ Res **96**(12): 1221-1232.
230 . Li, B. and S. M. Leal (2008). "Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data." Am J Hum Genet **83**(3): 311-321.
231 . Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, . . . S. Genome Project Data Processing (2009). "The Sequence Alignment/Map format and SAMtools." Bioinformatics **25**(16): 2078-2079.
232 . Li, J., J. T. Glessner, H. Zhang, C. Hou, Z. Wei, J. P. Bradfield, . . . P. M. Sleiman (2013). "GWAS of blood cell traits identifies novel associated loci and epistatic interactions in Caucasian and African-American children." Hum Mol Genet **22**(7): 1457-1464.
233 . Li, N. and M. Stephens (2003). "Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data." Genetics **165**(4): 2213-2233.
234 . Li, Y., C. Sidore, H. M. Kang, M. Boehnke and G. R. Abecasis (2011). "Low-coverage sequencing: implications for design of complex trait association studies." Genome Res **21**(6): 940-951.
235 . Li, Y., C. Willer, S. Sanna and G. Abecasis (2009). "Genotype imputation." Annu Rev Genomics Hum Genet **10**: 387-406.
236 . Libby, P. (2002). "Inflammation in atherosclerosis." Nature **420**(6917): 868-874.

237 . Lim, E. T., P. Wurtz, A. S. Havulinna, P. Palta, T. Tukiainen, K. Rehnstrom, . . . P. Sequencing Initiative Suomi (2014). "Distribution and medical impact of loss-of-function variants in the Finnish founder population." PLoS Genet **10**(7): e1004494.
238 . Lin, D. Y. and Z. Z. Tang (2011). "A general framework for detecting disease associations with rare variants in sequencing studies." Am J Hum Genet **89**(3): 354-367.
239 . Lin, J. P., C. J. O'Donnell, L. Jin, C. Fox, Q. Yang and L. A. Cupples (2007). "Evidence for linkage of red blood cell size and count: genome-wide scans in the Framingham Heart Study." Am J Hematol **82**(7): 605-610.
240 . Linsel-Nitschke, P., A. Gotz, J. Erdmann, I. Braenne, P. Braund, C. Hengstenberg, . . . C. Cardiogenics (2008). "Lifelong reduction of LDL-cholesterol related to a common variant in the LDL-receptor gene decreases the risk of coronary artery disease--a Mendelian Randomisation study." PLoS One **3**(8): e2986.
241 . Liu, D. J. and S. M. Leal (2012). "Estimating genetic effects and quantifying missing heritability explained by identified rare-variant associations." Am J Hum Genet **91**(4): 585-596.
242 . Liu, D. J., G. M. Peloso, X. Zhan, O. L. Holmen, M. Zawistowski, S. Feng, . . . G. R. Abecasis (2014). "Meta-analysis of gene-level tests for rare variant association." Nat Genet **46**(2): 200-204.
243 . Liuzzo, G., L. M. Biasucci, J. R. Gallimore, R. L. Grillo, A. G. Rebuzzi, M. B. Pepys and A. Maseri (1994). "The prognostic value of C-reactive protein and serum amyloid a protein in severe unstable angina." N Engl J Med **331**(7): 417-424.
244 . Lloyd-Jones, D. M., K. Liu, L. Tian and P. Greenland (2006). "Narrative review: Assessment of C-reactive protein in risk prediction for cardiovascular disease." Ann Intern Med **145**(1): 35-42.
245 . Lusis, A. J. and P. Pajukanta (2008). "A treasure trove for lipoprotein biology." Nat Genet **40**(2): 129-130.
246 . Ma, L., J. Yang, H. B. Runesha, T. Tanaka, L. Ferrucci, S. Bandinelli and Y. Da (2010). "Genome-wide association analysis of total cholesterol and high-density lipoprotein cholesterol levels using the Framingham heart study data." BMC Med Genet **11**: 55.
247 . MacDonald, M. E., A. Novelletto, C. Lin, D. Tagle, G. Barnes, G. Bates, . . . et al. (1992). "The Huntington's disease candidate region exhibits many different haplotypes." Nat Genet **1**(2): 99-103.
248 . MacGregor, A. J., J. R. Gallimore, T. D. Spector and M. B. Pepys (2004). "Genetic effects on baseline values of C-reactive protein and serum amyloid a protein: a comparison of monozygotic and dizygotic twins." Clin Chem **50**(1): 130-134.
249 . Magi, R. and A. P. Morris (2010). "GWAMA: software for genome-wide association meta-analysis." BMC Bioinformatics **11**: 288.
250 . Malik, I., J. Danesh, P. Whincup, V. Bhatia, O. Papacosta, M. Walker, . . . D. Haskard (2001). "Soluble adhesion molecules and prediction of coronary heart disease: a prospective study and meta-analysis." Lancet **358**(9286): 971-976.
251 . Maller, J. B., G. McVean, J. Byrnes, D. Vukcevic, K. Palin, Z. Su, . . . P. Donnelly (2012). "Bayesian refinement of association signals for 14 loci in 3 common diseases." Nat Genet **44**(12): 1294-1301.
252 . Manolio, T. A., F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, . . . P. M. Visscher (2009). "Finding the missing heritability of complex diseases." Nature **461**(7265): 747-753.
253 . Marcais, C., B. Verges, S. Charriere, V. Pruneta, M. Merlin, S. Billon, . . . P. Moulin (2005). "Apoa5 Q139X truncation predisposes to late-onset hyperchylomicronemia due to lipoprotein lipase impairment." J Clin Invest **115**(10): 2862-2869.
254 . Marchini, J., B. Howie, S. Myers, G. McVean and P. Donnelly (2007). "A new multipoint method for genome-wide association studies by imputation of genotypes." Nat Genet **39**(7): 906-913.
255 . Mardis, E. R. (2008). "The impact of next-generation sequencing technology on genetics." Trends Genet **24**(3): 133-141.
256 . Marduel, M., K. Ouguerram, V. Serre, D. Bonnefont-Rousselot, A. Marques-Pinheiro, K. Erik Berge, . . . M. Varret (2013). "Description of a large family with autosomal dominant hypercholesterolemia associated with the APOE p.Leu167del mutation." Hum Mutat **34**(1): 83-87.

257 . Margolis, K. L., J. E. Manson, P. Greenland, R. J. Rodabough, P. F. Bray, M. Safford, . . . G. Women's Health Initiative Research (2005). "Leukocyte count as a predictor of cardiovascular events and mortality in postmenopausal women: the Women's Health Initiative Observational Study." Arch Intern Med **165**(5): 500-508.
258 . Masicampo, E. J. and D. R. Lalande (2012). "A peculiar prevalence of p values just below .05." Q J Exp Psychol (Hove) **65**(11): 2271-2279.
259 . Massberg, S., C. Schulz and M. Gawaz (2003). "Role of platelets in the pathophysiology of acute coronary syndrome." Semin Vasc Med **3**(2): 147-162.
260 . Maurano, M. T., R. Humbert, E. Rynes, R. E. Thurman, E. Haugen, H. Wang, . . . J. A. Stamatoyannopoulos (2012). "Systematic localization of common disease-associated variation in regulatory DNA." Science **337**(6099): 1190-1195.
261 . McCarthy, M. I., G. R. Abecasis, L. R. Cardon, D. B. Goldstein, J. Little, J. P. Ioannidis and J. N. Hirschhorn (2008). "Genome-wide association studies for complex traits: consensus, uncertainty and challenges." Nat Rev Genet **9**(5): 356-369.
262 . McCarthy, M. I. and E. Zeggini (2009). "Genome-wide association studies in type 2 diabetes." Curr Diab Rep **9**(2): 164-171.
263 . McLaren, C. E., J. C. Barton, V. R. Gordeuk, L. Wu, P. C. Adams, D. M. Reboussin, . . . I. Iron Overload Screening Study Research (2007). "Determinants and characteristics of mean corpuscular volume and hemoglobin concentration in white HFE C282Y homozygotes in the hemochromatosis and iron overload screening study." Am J Hematol **82**(10): 898-905.
264 . McLaren, C. E., C. P. Garner, C. C. Constantine, S. McLachlan, C. D. Vulpe, B. M. Snively, . . . G. D. McLaren (2011). "Genome-wide association study identifies genetic loci associated with iron deficiency." PLoS One **6**(3): e17390.
265 . McLaren, W., B. Pritchard, D. Rios, Y. Chen, P. Flicek and F. Cunningham (2010). "Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor." Bioinformatics **26**(16): 2069-2070.
266 . McMorran, B. J., G. Burgio and S. J. Foote (2013). "New insights into the protective power of platelets in malaria infection." Commun Integr Biol **6**(3): e23653.
267 . Meisinger, C., H. Prokisch, C. Gieger, N. Soranzo, D. Mehta, D. Rosskopf, . . . A. Doring (2009). "A genome-wide association study identifies three loci associated with mean platelet volume." Am J Hum Genet **84**(1): 66-71.
268 . Melander, O., C. Newton-Cheh, P. Almgren, B. Hedblad, G. Berglund, G. Engstrom, . . . T. J. Wang (2009). "Novel and conventional biomarkers for prediction of incident cardiovascular events in the community." JAMA **302**(1): 49-57.
269 . Menzel, S., C. Garner, I. Gut, F. Matsuda, M. Yamaguchi, S. Heath, . . . S. L. Thein (2007). "A QTL influencing F cell production maps to a gene encoding a zinc-finger protein on chromosome 2p15." Nat Genet **39**(10): 1197-1199.
270 . Menzel, S., J. Jiang, N. Silver, J. Gallagher, J. Cunningham, G. Surdulescu, . . . S. L. Thein (2007). "The HBS1L-MYB intergenic region on chromosome 6q23.3 influences erythrocyte, platelet, and monocyte counts in humans." Blood **110**(10): 3624-3626.
271 . Moayyeri, A., C. J. Hammond, D. J. Hart and T. D. Spector (2012). "The UK Adult Twin Registry (TwinsUK Resource)." Twin Res Hum Genet: 1-6.
272 . Moltke, I., N. Grarup, M. E. Jorgensen, P. Bjerregaard, J. T. Treebak, M. Fumagalli, . . . T. Hansen (2014). "A common Greenlandic TBC1D4 variant confers muscle insulin resistance and type 2 diabetes." Nature **512**(7513): 190-193.
273 . Monda, K. L., G. K. Chen, K. C. Taylor, C. Palmer, T. L. Edwards, L. A. Lange, . . . C. A. Haiman (2013). "A meta-analysis identifies new loci associated with body mass index in individuals of African ancestry." Nat Genet **45**(6): 690-696.
274 . Monteferrario, D., N. A. Bolar, A. E. Marneth, K. M. Hebeda, S. M. Bergevoet, H. Veenstra, . . . B. A. Van der Reijden (2014). "A dominant-negative GFI1B mutation in the gray platelet syndrome." N Engl J Med **370**(3): 245-253.

275 . Morgenthaler, S. and W. G. Thilly (2007). "A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST)." Mutat Res **615**(1-2): 28-56.
276 . Morrison, A. C., A. Voorman, A. D. Johnson, X. Liu, J. Yu, A. Li, . . . C. Aging Research in Genetic Epidemiology (2013). "Whole-genome sequence-based analysis of high-density lipoprotein cholesterol." Nat Genet **45**(8): 899-901.
277 . Morton, N. E. (1955). "Sequential tests for the detection of linkage." Am J Hum Genet **7**(3): 277-318.
278 . Motazacker, M. M., J. Pirruccello, R. Huijgen, R. Do, S. Gabriel, J. Peter, . . . S. W. Fouchier (2012). "Advances in genetics show the need for extending screening strategies for autosomal dominant hypercholesterolaemia." Eur Heart J **33**(11): 1360-1366.
279 . Musunuru, K., J. P. Pirruccello, R. Do, G. M. Peloso, C. Guiducci, C. Sougnez, . . . S. Kathiresan (2010). "Exome sequencing, ANGPTL3 mutations, and familial combined hypolipidemia." N Engl J Med **363**(23): 2220-2227.
280 . Musunuru, K., A. Strong, M. Frank-Kamenetsky, N. E. Lee, T. Ahfeldt, K. V. Sachs, . . . D. J. Rader (2010). "From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus." Nature **466**(7307): 714-719.
281 . Nadar, S., A. D. Blann and G. Y. Lip (2004). "Platelet morphology and plasma indices of platelet activation in essential hypertension: effects of amlodipine-based antihypertensive therapy." Ann Med **36**(7): 552-557.
282 . Naitza, S., E. Porcu, M. Steri, D. D. Taub, A. Mulas, X. Xiao, . . . F. Cucca (2012). "A genome-wide association scan on the levels of markers of inflammation in Sardinians reveals associations that underpin its complex regulation." PLoS Genet **8**(1): e1002480.
283 . Nalls, M. A., D. J. Couper, T. Tanaka, F. J. van Rooij, M. H. Chen, A. V. Smith, . . . S. K. Ganesh (2011). "Multiple loci are associated with white blood cell phenotypes." PLoS Genet **7**(6): e1002113.
284 . Navab, M., S. T. Reddy, B. J. Van Lenten and A. M. Fogelman (2011). "HDL and cardiovascular disease: atherogenic and atheroprotective mechanisms." Nat Rev Cardiol **8**(4): 222-232.
285 . Neale, B. M., M. A. Rivas, B. F. Voight, D. Altshuler, B. Devlin, M. Orho-Melander, . . . M. J. Daly (2011). "Testing for an unusual distribution of rare variants." PLoS Genet **7**(3): e1001322.
286 . Nicholls, S. J., A. Gordon, J. Johansson, K. Wolski, C. M. Ballantyne, J. J. Kastelein, . . . S. E. Nissen (2011). "Efficacy and safety of a novel oral inducer of apolipoprotein a-I synthesis in statin-treated patients with stable coronary artery disease a randomized controlled trial." J Am Coll Cardiol **57**(9): 1111-1119.
287 . Nieto, F. J., M. Szklo, A. R. Folsom, R. Rock and M. Mercuri (1992). "Leukocyte count correlates in middle-aged adults: the Atherosclerosis Risk in Communities (ARIC) Study." Am J Epidemiol **136**(5): 525-537.
288 . Nimptsch, K., K. Aleksandrova, H. Boeing, J. Janke, Y. A. Lee, M. Jenab, . . . T. Pischon (2015). "Association of CRP genetic variants with blood concentrations of C-reactive protein and colorectal cancer risk." Int J Cancer **136**(5): 1181-1192.
289 . Nordestgaard, B. G., M. Benn, P. Schnohr and A. Tybjaerg-Hansen (2007). "Nonfasting triglycerides and risk of myocardial infarction, ischemic heart disease, and death in men and women." JAMA **298**(3): 299-308.
290 . Ntalla, I., M. Giannakopoulou, P. Vlachou, K. Giannitsopoulou, V. Gkesou, C. Makridi, . . . G. V. Dedoussis (2014). "Body composition and eating behaviours in relation to dieting involvement in a sample of urban Greek adolescents from the TEENAGE (TEENs of Attica: Genes & Environment) study." Public Health Nutr **17**(3): 561-568.
291 . Oberdoerffer, S., L. F. Moita, D. Neems, R. P. Freitas, N. Hacohen and A. Rao (2008). "Regulation of CD45 alternative splicing by heterogeneous ribonucleoprotein, hnRNPLL." Science **321**(5889): 686-691.

292 . Okada, Y., T. Hirota, Y. Kamatani, A. Takahashi, H. Ohmiya, N. Kumasaka, . . . N. Kamatani (2011). "Identification of nine novel loci associated with white blood cell subtypes in a Japanese population." PLoS Genet **7**(6): e1002067.
293 . Okada, Y., A. Takahashi, H. Ohmiya, N. Kumasaka, Y. Kamatani, N. Hosono, . . . N. Kamatani (2011). "Genome-wide association study for C-reactive protein levels identified pleiotropic associations in the IL6 locus." Hum Mol Genet **20**(6): 1224-1231.
294 . Olson, R. E. (1998). "Discovery of the lipoproteins, their role in fat transport and their significance as risk factors." J Nutr **128**(2 Suppl): 439S-443S.
295 . Onengut-Gumuscu, S., W. M. Chen, O. Burren, N. J. Cooper, A. R. Quinlan, J. C. Mychaleckyj, . . . S. S. Rich (2015). "Fine mapping of type 1 diabetes susceptibility loci and evidence for colocalization of causal variants with lymphoid gene enhancers." Nat Genet **47**(4): 381-386.
296 . Orkin, S. H. and L. I. Zon (2008). "Hematopoiesis: an evolving paradigm for stem cell biology." Cell **132**(4): 631-644.
297 . Park, J. H., M. H. Gail, C. R. Weinberg, R. J. Carroll, C. C. Chung, Z. Wang, . . . N. Chatterjee (2011). "Distribution of allele frequencies and effect sizes and their interrelationships for common genetic susceptibility variants." Proc Natl Acad Sci U S A **108**(44): 18026-18031.
298 . Parkes, M., J. C. Barrett, N. J. Prescott, M. Tremelling, C. A. Anderson, S. A. Fisher, . . . C. G. Mathew (2007). "Sequence variants in the autophagy gene IRGM and multiple other replicating loci contribute to Crohn's disease susceptibility." Nat Genet **39**(7): 830-832.
299 . Pate, R. R., M. Pratt, S. N. Blair, W. L. Haskell, C. A. Macera, C. Bouchard, . . . et al. (1995). "Physical activity and public health. A recommendation from the Centers for Disease Control and Prevention and the American College of Sports Medicine." JAMA **273**(5): 402-407.
300 . Pathansali, R., N. Smith and P. Bath (2001). "Altered megakaryocyte-platelet haemostatic axis in hypercholesterolaemia." Platelets **12**(5): 292-297.
301 . Pearson, T. A., G. A. Mensah, R. W. Alexander, J. L. Anderson, R. O. Cannon, 3rd, M. Criqui, . . . A. American Heart (2003). "Markers of inflammation and cardiovascular disease: application to clinical and public health practice: A statement for healthcare professionals from the Centers for Disease Control and Prevention and the American Heart Association." Circulation **107**(3): 499-511.
302 . Peloso, G. M., P. L. Auer, J. C. Bis, A. Voorman, A. C. Morrison, N. O. Stitziel, . . . L. A. Cupples (2014). "Association of low-frequency and rare coding-sequence variants with blood lipids and coronary heart disease in 56,000 whites and blacks." Am J Hum Genet **94**(2): 223-232.
303 . Peltola, V., J. Mertsola and O. Ruuskanen (2006). "Comparison of total white blood cell count and serum C-reactive protein levels in confirmed bacterial and viral infections." J Pediatr **149**(5): 721-724.
304 . Pennisi, E. (2012). "Genomics. ENCODE project writes eulogy for junk DNA." Science **337**(6099): 1159, 1161.
305 . Pepys, M. B. and G. M. Hirschfield (2003). "C-reactive protein: a critical update." J Clin Invest **111**(12): 1805-1812.
306 . Persson, M., B. Hedblad, J. J. Nelson and G. Berglund (2007). "Elevated Lp-PLA2 levels add prognostic information to the metabolic syndrome on incidence of cardiovascular events among middle-aged nondiabetic subjects." Arterioscler Thromb Vasc Biol **27**(6): 1411-1416.
307 . Peterfy, M., O. Ben-Zeev, H. Z. Mao, D. Weissglas-Volkov, B. E. Aouizerat, C. R. Pullinger, . . . M. H. Doolittle (2007). "Mutations in LMF1 cause combined lipase deficiency and severe hypertriglyceridemia." Nat Genet **39**(12): 1483-1487.
308 . Pickrell, J. K. (2014). "Joint analysis of functional genomic data and genome-wide association studies of 18 human traits." Am J Hum Genet **94**(4): 559-573.
309 . Pilia, G., W. M. Chen, A. Scuteri, M. Orru, G. Albai, M. Dei, . . . D. Schlessinger (2006). "Heritability of cardiovascular and personality traits in 6,148 Sardinians." PLoS Genet **2**(8): e132.
310 . Pistis, G., S. U. Okonkwo, M. Traglia, C. Sala, S. Y. Shin, C. Masciullo, . . . D. Toniolo (2013). "Genome wide association analysis of a founder population identified TAF3 as a gene for MCHC in humans." PLoS One **8**(7): e69206.

311 . Pizzulli, L., A. Yang, J. F. Martin and B. Luderitz (1998). "Changes in platelet size and count in unstable angina compared to stable angina or non-cardiac chest pain." Eur Heart J **19**(1): 80-84.
312 . Pollin, T. I., C. M. Damcott, H. Shen, S. H. Ott, J. Shelton, R. B. Horenstein, . . . A. R. Shuldiner (2008). "A null mutation in human APOC3 confers a favorable plasma lipid profile and apparent cardioprotection." Science **322**(5908): 1702-1705.
313 . Power, C. and J. Elliott (2006). "Cohort profile: 1958 British birth cohort (National Child Development Study)." Int J Epidemiol **35**(1): 34-41.
314 . Price, A. L., G. V. Kryukov, P. I. de Bakker, S. M. Purcell, J. Staples, L. J. Wei and S. R. Sunyaev (2010). "Pooled association tests for rare variants in exon-resequencing studies." Am J Hum Genet **86**(6): 832-838.
315 . Pritchard, J. K. (2001). "Are rare variants responsible for susceptibility to complex diseases?" Am J Hum Genet **69**(1): 124-137.
316 . Pritchard, J. K. and N. J. Cox (2002). "The allelic architecture of human disease genes: common disease-common variant...or not?" Hum Mol Genet **11**(20): 2417-2423.
317 . Prospective Studies, C., S. Lewington, G. Whitlock, R. Clarke, P. Sherliker, J. Emberson, . . . R. Collins (2007). "Blood cholesterol and vascular mortality by age, sex, and blood pressure: a meta-analysis of individual data from 61 prospective studies with 55,000 vascular deaths." Lancet **370**(9602): 1829-1839.
318 . Qayyum, R., B. M. Snively, E. Ziv, M. A. Nalls, Y. Liu, W. Tang, . . . A. P. Reiner (2012). "A meta-analysis and genome-wide association study of platelet count and mean platelet volume in african americans." PLoS Genet **8**(3): e1002491.
319 . Quail, M. A., M. Smith, P. Coupland, T. D. Otto, S. R. Harris, T. R. Connor, . . . Y. Gu (2012). "A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers." BMC Genomics **13**: 341.
320 . Ramanan, V. K., S. L. Risacher, K. Nho, S. Kim, S. Swaminathan, L. Shen, . . . I. Alzheimer's Disease Neuroimaging (2014). "APOE and BCHE as modulators of cerebral amyloid deposition: a florbetapir PET genome-wide association study." Mol Psychiatry **19**(3): 351-357.
321 . Ramasamy, I. (2014). "Recent advances in physiological lipoprotein metabolism." Clin Chem Lab Med **52**(12): 1695-1727.
322 . Rana, J. S., B. J. Arsenault, J. P. Despres, M. Cote, P. J. Talmud, E. Ninio, . . . S. M. Boekholdt (2011). "Inflammatory biomarkers, physical activity, waist circumference, and risk of future coronary heart disease in healthy men and women." Eur Heart J **32**(3): 336-344.
323 . Rana, J. S., M. Cote, J. P. Despres, M. S. Sandhu, P. J. Talmud, E. Ninio, . . . S. M. Boekholdt (2009). "Inflammatory biomarkers and the prediction of coronary events among people at intermediate risk: the EPIC-Norfolk prospective population study." Heart **95**(20): 1682-1687.
324 . Rasmussen-Torvik, L. J., J. A. Pacheco, R. A. Wilke, W. K. Thompson, M. D. Ritchie, A. N. Kho, . . . R. L. Chisholm (2012). "High density GWAS for LDL cholesterol in African Americans using electronic medical records reveals a strong protective variant in APOE." Clin Transl Sci **5**(5): 394-399.
325 . Reich, D., M. A. Nalls, W. H. Kao, E. L. Akylbekova, A. Tandon, N. Patterson, . . . J. G. Wilson (2009). "Reduced neutrophil count in people of African descent is due to a regulatory variant in the Duffy antigen receptor for chemokines gene." PLoS Genet **5**(1): e1000360.
326 . Reich, D. E. and E. S. Lander (2001). "On the allelic spectrum of human disease." Trends Genet **17**(9): 502-510.
327 . Reiner, A. P., M. J. Barber, Y. Guan, P. M. Ridker, L. A. Lange, D. I. Chasman, . . . R. M. Krauss (2008). "Polymorphisms of the HNF1A gene encoding hepatocyte nuclear factor-1 alpha are associated with C-reactive protein." Am J Hum Genet **82**(5): 1193-1201.
328 . Reiner, A. P., S. Beleza, N. Franceschini, P. L. Auer, J. G. Robinson, C. Kooperberg, . . . H. Tang (2012). "Genome-wide association and population genetic analysis of C-reactive protein in African American and Hispanic American women." Am J Hum Genet **91**(3): 502-512.

329 . Reiner, A. P., G. Lettre, M. A. Nalls, S. K. Ganesh, R. Mathias, M. A. Austin, . . . J. G. Wilson (2011). "Genome-wide association study of white blood cell count in 16,388 African Americans: the continental origins and genetic epidemiology network (COGENT)." PLoS Genet **7**(6): e1002108.
330 . Ridker, P. M., J. E. Buring and N. Rifai (2001). "Soluble P-selectin and the risk of future cardiovascular events." Circulation **103**(4): 491-495.
331 . Ridker, P. M., J. E. Buring, N. Rifai and N. R. Cook (2007). "Development and validation of improved algorithms for the assessment of global cardiovascular risk in women: the Reynolds Risk Score." JAMA **297**(6): 611-619.
332 . Ridker, P. M., M. Cushman, M. J. Stampfer, R. P. Tracy and C. H. Hennekens (1997). "Inflammation, aspirin, and the risk of cardiovascular disease in apparently healthy men." N Engl J Med **336**(14): 973-979.
333 . Ridker, P. M., E. Danielson, F. A. Fonseca, J. Genest, A. M. Gotto, Jr., J. J. Kastelein, . . . J. S. Group (2008). "Rosuvastatin to prevent vascular events in men and women with elevated C-reactive protein." N Engl J Med **359**(21): 2195-2207.
334 . Ridker, P. M., E. Danielson, F. A. Fonseca, J. Genest, A. M. Gotto, Jr., J. J. Kastelein, . . . J. T. S. Group (2009). "Reduction in C-reactive protein and LDL cholesterol and cardiovascular event rates after initiation of rosuvastatin: a prospective study of the JUPITER trial." Lancet **373**(9670): 1175-1182.
335 . Ridker, P. M., G. Pare, A. Parker, R. Y. Zee, J. S. Danik, J. E. Buring, . . . D. I. Chasman (2008). "Loci related to metabolic-syndrome pathways including LEPR,HNF1A, IL6R, and GCKR associate with plasma C-reactive protein: the Women's Genome Health Study." Am J Hum Genet **82**(5): 1185-1192.
336 . Ridker, P. M., G. Pare, A. N. Parker, R. Y. Zee, J. P. Miletich and D. I. Chasman (2009). "Polymorphism in the CETP gene region, HDL cholesterol, and risk of future myocardial infarction: Genomewide analysis among 18 245 initially healthy women from the Women's Genome Health Study." Circ Cardiovasc Genet **2**(1): 26-33.
337 . Ridker, P. M., N. P. Paynter, N. Rifai, J. M. Gaziano and N. R. Cook (2008). "C-reactive protein and parental history improve global cardiovascular risk prediction: the Reynolds Risk Score for men." Circulation **118**(22): 2243-2251, 2244p following 2251.
338 . Ridker, P. M., N. Rifai, M. Clearfield, J. R. Downs, S. E. Weis, J. S. Miles, . . . I. Air Force/Texas Coronary Atherosclerosis Prevention Study (2001). "Measurement of C-reactive protein for the targeting of statin therapy in the primary prevention of acute coronary events." N Engl J Med **344**(26): 1959-1965.
339 . Ridker, P. M., N. Rifai, M. J. Stampfer and C. H. Hennekens (2000). "Plasma concentration of interleukin-6 and the risk of future myocardial infarction among apparently healthy men." Circulation **101**(15): 1767-1772.
340 . Rimm, E. B., E. L. Giovannucci, W. C. Willett, G. A. Colditz, A. Ascherio, B. Rosner and M. J. Stampfer (1991). "Prospective study of alcohol consumption and risk of coronary disease in men." Lancet **338**(8765): 464-468.
341 . Rios, J., E. Stein, J. Shendure, H. H. Hobbs and J. C. Cohen (2010). "Identification by whole-genome resequencing of gene defect responsible for severe hypercholesterolemia." Hum Mol Genet **19**(22): 4313-4318.
342 . Risch, N. and K. Merikangas (1996). "The future of genetic studies of complex human diseases." Science **273**(5281): 1516-1517.
343 . Robinson, J. G. (2009). "Are you targeting non-high-density lipoprotein cholesterol?" J Am Coll Cardiol **55**(1): 42-44.
344 . Robinson, M. R., N. R. Wray and P. M. Visscher (2014). "Explaining additional genetic variation in complex traits." Trends Genet **30**(4): 124-132.
345 . Rosenthal, E. A., J. Ranchalis, D. R. Crosslin, A. Burt, J. D. Brunzell, A. G. Motulsky, . . . G. P. Jarvik (2013). "Joint linkage and association analysis with exome sequence data implicates SLC25A40 in hypertriglyceridemia." Am J Hum Genet **93**(6): 1035-1045.

346 . Ruchat, S. M., J. P. Despres, S. J. Weisnagel, Y. C. Chagnon, C. Bouchard and L. Perusse (2008). "Genome-wide linkage analysis for circulating levels of adipokines and C-reactive protein in the Quebec family study (QFS)." J Hum Genet **53**(7): 629-636.
347 . Ruggiero, C., E. J. Metter, A. Cherubini, M. Maggio, R. Sen, S. S. Najjar, . . . L. Ferrucci (2007). "White blood cell count and mortality in the Baltimore Longitudinal Study of Aging." J Am Coll Cardiol **49**(18): 1841-1850.
348 . Rust, S., M. Rosier, H. Funke, J. Real, Z. Amoura, J. C. Piette, . . . G. Assmann (1999). "Tangier disease is caused by mutations in the gene encoding ATP-binding cassette transporter 1." Nat Genet **22**(4): 352-355.
349 . Sabatti, C., S. K. Service, A. L. Hartikainen, A. Pouta, S. Ripatti, J. Brodsky, . . . L. Peltonen (2009). "Genome-wide association analysis of metabolic traits in a birth cohort from a founder population." Nat Genet **41**(1): 35-46.
350 . Sandhu, M. S., D. M. Waterworth, S. L. Debenham, E. Wheeler, K. Papadakis, J. H. Zhao, . . . V. Mooser (2008). "LDL-cholesterol concentrations: a genome-wide association study." Lancet **371**(9611): 483-491.
351 . Sankaran, V. G., L. S. Ludwig, E. Sicinska, J. Xu, D. E. Bauer, J. C. Eng, . . . H. F. Lodish (2012). "Cyclin D3 coordinates the cell cycle during differentiation to regulate erythrocyte size and number." Genes Dev **26**(18): 2075-2087.
352 . Sankaran, V. G., T. F. Menne, J. Xu, T. E. Akie, G. Lettre, B. Van Handel, . . . S. H. Orkin (2008). "Human fetal hemoglobin expression is regulated by the developmental stage-specific repressor BCL11A." Science **322**(5909): 1839-1842.
353 . Sankaran, V. G., J. Xu and S. H. Orkin (2010). "Advances in the understanding of haemoglobin switching." Br J Haematol **149**(2): 181-194.
354 . Santos-Cortez, R. L., K. Lee, A. P. Giese, M. Ansar, M. Amin-Ud-Din, K. Rehn, . . . S. M. Leal (2014). "Adenylate cyclase 1 (ADCY1) mutations cause recessive hearing impairment in humans and defects in hair cell function and hearing in zebrafish." Hum Mol Genet **23**(12): 3289-3298.
355 . Santos, S., T. W. Rooke, K. R. Bailey, J. P. McConnell and I. J. Kullo (2004). "Relation of markers of inflammation (C-reactive protein, white blood cell count, and lipoprotein-associated phospholipase A2) to the ankle-brachial index." Vasc Med **9**(3): 171-176.
356 . Sarwar, N., J. Danesh, G. Eiriksdottir, G. Sigurdsson, N. Wareham, S. Bingham, . . . V. Gudnason (2007). "Triglycerides and the risk of coronary heart disease: 10,158 incident cases among 262,525 participants in 29 Western prospective studies." Circulation **115**(4): 450-458.
357 . Sattar, N., H. M. Murray, A. McConnachie, G. J. Blauw, E. L. Bollen, B. M. Buckley, . . . P. S. Group (2007). "C-reactive protein and prediction of coronary heart disease and global vascular events in the Prospective Study of Pravastatin in the Elderly at Risk (PROSPER)." Circulation **115**(8): 981-989.
358 . Saxena, R., B. F. Voight, V. Lyssenko, N. P. Burtt, P. I. de Bakker, H. Chen, . . . S. Purcell (2007). "Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels." Science **316**(5829): 1331-1336.
359 . Scheet, P. and M. Stephens (2006). "A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase." Am J Hum Genet **78**(4): 629-644.
360 . Schunkert, H., I. R. Konig, S. Kathiresan, M. P. Reilly, T. L. Assimes, H. Holm, . . . N. J. Samani (2011). "Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease." Nat Genet **43**(4): 333-338.
361 . Seddon, J. M., R. Reynolds, J. Maller, J. A. Fagerness, M. J. Daly and B. Rosner (2009). "Prediction model for prevalence and incidence of advanced age-related macular degeneration based on genetic, demographic, and environmental variables." Invest Ophthalmol Vis Sci **50**(5): 2044-2053.

362 . Selhub, J., P. F. Jacques, A. G. Bostom, R. B. D'Agostino, P. W. Wilson, A. J. Belanger, . . . I. H. Rosenberg (1995). "Association between plasma homocysteine concentrations and extracranial carotid-artery stenosis." N Engl J Med **332**(5): 286-291.
363 . Service., U. P. H. (1983). "The Health Consequences of Smoking: Cardiovascular Disease: A Report of the Surgeon General. ." DHHS (PHS) 84-50204.
364 . Shameer, K., J. C. Denny, K. Ding, H. Jouni, D. R. Crosslin, M. de Andrade, . . . I. J. Kullo (2014). "A genome- and phenome-wide association study to identify genetic variants influencing platelet count and volume and their pleiotropic effects." Hum Genet **133**(1): 95-109.
365 . Shankar, A., J. J. Wang, E. Rochtchina, M. C. Yu, R. Kefford and P. Mitchell (2006). "Association between circulating white blood cell count and cancer mortality: a population-based cohort study." Arch Intern Med **166**(2): 188-194.
366 . Sharp, D., L. Blinderman, K. A. Combs, B. Kienzle, B. Ricci, K. Wager-Smith, . . . et al. (1993). "Cloning and gene defects in microsomal triglyceride transfer protein associated with abetalipoproteinaemia." Nature **365**(6441): 65-69.
367 . Shepherd, J., S. M. Cobbe, I. Ford, C. G. Isles, A. R. Lorimer, P. W. MacFarlane, . . . C. J. Packard (1995). "Prevention of coronary heart disease with pravastatin in men with hypercholesterolemia. West of Scotland Coronary Prevention Study Group." N Engl J Med **333**(20): 1301-1307.
368 . Soranzo, N., A. Rendon, C. Gieger, C. I. Jones, N. A. Watkins, S. Menzel, . . . W. H. Ouwehand (2009). "A novel variant on chromosome 7q22.3 associated with mean platelet volume, counts, and function." Blood **113**(16): 3831-3837.
369 . Soranzo, N., F. Rivadeneira, U. Chinappen-Horsley, I. Malkina, J. B. Richards, N. Hammond, . . . P. Deloukas (2009). "Meta-analysis of genome-wide scans for human adult stature identifies novel Loci and associations with measures of skeletal frame size." PLoS Genet **5**(4): e1000445.
370 . Soranzo, N., T. D. Spector, M. Mangino, B. Kuhnel, A. Rendon, A. Teumer, . . . C. Gieger (2009). "A genome-wide meta-analysis identifies 22 loci associated with eight hematological parameters in the HaemGen consortium." Nat Genet **41**(11): 1182-1190.
371 . Soria, L. F., E. H. Ludwig, H. R. Clarke, G. L. Vega, S. M. Grundy and B. J. McCarthy (1989). "Association between a specific apolipoprotein B mutation and familial defective apolipoprotein B-100." Proc Natl Acad Sci U S A **86**(2): 587-591.
372 . Sorrentino, V., S. W. Fouchier, M. M. Motazacker, J. K. Nelson, J. C. Defesche, G. M. Dallinga-Thie, . . . N. Zelcer (2013). "Identification of a loss-of-function inducible degrader of the low-density lipoprotein receptor variant in individuals with low circulating low-density lipoprotein." Eur Heart J **34**(17): 1292-1297.
373 . Spector, T. D. and F. M. Williams (2006). "The UK Adult Twin Registry (TwinsUK)." Twin Res Hum Genet **9**(6): 899-906.
374 . Spencer, C. C., Z. Su, P. Donnelly and J. Marchini (2009). "Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip." PLoS Genet **5**(5): e1000477.
375 . St George-Hyslop, P. H., R. E. Tanzi, R. J. Polinsky, J. L. Haines, L. Nee, P. C. Watkins, . . . et al. (1987). "The genetic defect causing familial Alzheimer's disease maps on chromosome 21." Science **235**(4791): 885-890.
376 . Stampfer, M. J., G. A. Colditz, W. C. Willett, F. E. Speizer and C. H. Hennekens (1988). "A prospective study of moderate alcohol consumption and the risk of coronary disease and stroke in women." N Engl J Med **319**(5): 267-273.
377 . Stein, E. A., S. Mellis, G. D. Yancopoulos, N. Stahl, D. Logan, W. B. Smith, . . . G. D. Swergold (2012). "Effect of a monoclonal antibody to PCSK9 on LDL cholesterol." N Engl J Med **366**(12): 1108-1118.
378 . Stephens, M. (2013). "A unified framework for association analysis with multiple related phenotypes." PLoS One **8**(7): e65245.
379 . Stitziel, N. O., S. W. Fouchier, B. Sjouke, G. M. Peloso, A. M. Moscoso, P. L. Auer, . . . G. O. E. S. P. Blood Institute (2013). "Exome sequencing and directed clinical phenotyping diagnose cholesterol

ester storage disease presenting as autosomal recessive hypercholesterolemia." Arterioscler Thromb Vasc Biol **33**(12): 2909-2914.
380 . Stolk, L., J. R. Perry, D. I. Chasman, C. He, M. Mangino, P. Sulem, . . . K. L. Lunetta (2012). "Meta-analyses identify 13 loci associated with age at menopause and highlight DNA repair and immune pathways." Nat Genet **44**(3): 260-268.
381 . Surakka, I., A. Isaacs, L. C. Karssen, P. P. Laurila, R. P. Middelberg, E. Tikkanen, . . . E. Consortium (2011). "A genome-wide screen for interactions reveals a new locus on 4p15 modifying the effect of waist-to-hip ratio on total cholesterol." PLoS Genet **7**(10): e1002333.
382 . Surakka, I., J. B. Whitfield, M. Perola, P. M. Visscher, G. W. Montgomery, M. Falchi, . . . E. P. Genom (2012). "A genome-wide association study of monozygotic twin-pairs suggests a locus related to variability of serum high-density lipoprotein cholesterol." Twin Res Hum Genet **15**(6): 691-699.
383 . Syvanen, A. C. (2005). "Toward genome-wide SNP genotyping." Nat Genet **37 Suppl**: S5-10.
384 . Szmitko, P. E., C. H. Wang, R. D. Weisel, J. R. de Almeida, T. J. Anderson and S. Verma (2003). "New markers of inflammation and endothelial cell activation: Part I." Circulation **108**(16): 1917-1923.
385 . Tachmazidou, I., G. Dedoussis, L. Southam, A. E. Farmaki, G. R. Ritchie, D. K. Xifara, . . . E. Zeggini (2013). "A rare functional cardioprotective APOC3 variant has risen in frequency in distinct population isolates." Nat Commun **4**: 2872.
386 . Tan, A., J. Sun, N. Xia, X. Qin, Y. Hu, S. Zhang, . . . J. Xu (2012). "A genome-wide association and gene-environment interaction study for serum triglycerides levels in a healthy Chinese male population." Hum Mol Genet **21**(7): 1658-1664.
387 . Tchernitchko, D., M. Goossens and H. Wajcman (2004). "In silico prediction of the deleterious effect of a mutation: proceed with caution in clinical genetics." Clin Chem **50**(11): 1974-1978.
388 . Teo, K. K., S. Ounpuu, S. Hawken, M. R. Pandey, V. Valentin, D. Hunt, . . . I. S. Investigators (2006). "Tobacco use and risk of myocardial infarction in 52 countries in the INTERHEART study: a case-control study." Lancet **368**(9536): 647-658.
389 . Teo, Y. Y., K. S. Small and D. P. Kwiatkowski (2010). "Methodological challenges of genome-wide association analysis in Africa." Nat Rev Genet **11**(2): 149-160.
390 . Teslovich, T. M., K. Musunuru, A. V. Smith, A. C. Edmondson, I. M. Stylianou, M. Koseki, . . . S. Kathiresan (2010). "Biological, clinical and population relevance of 95 loci for blood lipids." Nature **466**(7307): 707-713.
391 . Tg, N. H. L. Hdl Working Group of the Exome Sequencing Project, I. Blood, J. Crosby, G. M. Peloso, P. L. Auer, . . . S. Kathiresan (2014). "Loss-of-function mutations in APOC3, triglycerides, and coronary disease." N Engl J Med **371**(1): 22-31.
392 . Thaulow, E., J. Erikssen, L. Sandvik, H. Stormorken and P. F. Cohn (1991). "Blood platelet count and function are related to total and cardiovascular death in apparently healthy men." Circulation **84**(2): 613-617.
393 . The TG and HDL Working Group of the Exome Sequencing Project, N. H. L. B., Institute (2014). "Loss-of-Function Mutations in APOC3, Triglycerides, and Coronary Disease." N Engl J Med **371**(1): 22-31.
394 . The UK10K Consortium (2015). "The UK10K project: rare variants in health and disease." *submitted*.
395 . The Women's Health Initiative Study Group (1998). "Design of the Women's Health Initiative clinical trial and observational study. The Women's Health Initiative Study Group." Control Clin Trials **19**(1): 61-109.
396 . Thompson, D., M. B. Pepys and S. P. Wood (1999). "The physiological structure of human C-reactive protein and its complex with phosphocholine." Structure **7**(2): 169-177.
397 . Thomson, W., A. Barton, X. Ke, S. Eyre, A. Hinks, J. Bowes, . . . J. Worthington (2007). "Rheumatoid arthritis association at 6q23." Nat Genet **39**(12): 1431-1433.

398 . Timmann, C., T. Thye, M. Vens, J. Evans, J. May, C. Ehmen, . . . R. D. Horstmann (2012). "Genome-wide association study indicates two novel resistance loci for severe malaria." Nature **489**(7416): 443-446.
399 . Timpson, N., K. Walter, M. JL, I. Tachmazidou, G. Malerba, S.-Y. Shin, . . . N. Soranzo "A novel low-frequency variant near APOC3 is associated with plasma triglyceride and VLDL levels in Europeans." Nature Communications (Under peer review).
400 . Timpson, N. J., K. Walter, J. L. Min, I. Tachmazidou, G. Malerba, S. Y. Shin, . . . U. O. C. Members (2014). "A rare variant in APOC3 is associated with plasma triglyceride and VLDL levels in Europeans." Nat Commun **5**: 4871.
401 . Tiong, A. Y. and D. Brieger (2005). "Inflammation and coronary artery disease." Am Heart J **150**(1): 11-18.
402 . Todd, J. A., N. M. Walker, J. D. Cooper, D. J. Smyth, K. Downes, V. Plagnol, . . . D. G. Clayton (2007). "Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes." Nat Genet **39**(7): 857-864.
403 . Traglia, M., C. Sala, C. Masciullo, V. Cverhova, F. Lori, G. Pistis, . . . D. Toniolo (2009). "Heritability and Demographic Analyses in the Large Isolated Population of Val Borbera Suggest Advantages in Mapping Complex Traits Genes." PLoS ONE **4**(10): e7554.
404 . Triglyceride Coronary Disease Genetics, C., C. Emerging Risk Factors, N. Sarwar, M. S. Sandhu, S. L. Ricketts, A. S. Butterworth, . . . J. Danesh (2010). "Triglyceride-mediated pathways and coronary disease: collaborative analysis of 101 studies." Lancet **375**(9726): 1634-1639.
405 . Tunstall-Pedoe, H., M. Woodward and S. g. o. r. estimation (2006). "By neglecting deprivation, cardiovascular risk scoring will exacerbate social gradients in disease." Heart **92**(3): 307-310.
406 . Uda, M., R. Galanello, S. Sanna, G. Lettre, V. G. Sankaran, W. Chen, . . . A. Cao (2008). "Genome-wide association study shows BCL11A associated with persistent fetal hemoglobin and amelioration of the phenotype of beta-thalassemia." Proc Natl Acad Sci U S A **105**(5): 1620-1625.
407 . van der Harst, P., W. Zhang, I. Mateo Leach, A. Rendon, N. Verweij, J. Sehmi, . . . J. C. Chambers (2012). "Seventy-five genetic loci influencing the human red blood cell." Nature **492**(7429): 369-375.
408 . van Dongen, J., G. Willemsen, W. M. Chen, E. J. de Geus and D. I. Boomsma (2013). "Heritability of metabolic syndrome traits in a large population-based sample." J Lipid Res **54**(10): 2914-2923.
409 . Venter, J. C., M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, . . . X. Zhu (2001). "The sequence of the human genome." Science **291**(5507): 1304-1351.
410 . Vinayagamoorthy, N., H. J. Hu, S. H. Yim, S. H. Jung, J. Jo, S. H. Jee and Y. J. Chung (2014). "New variants including ARG1 polymorphisms associated with C-reactive protein levels identified by genome-wide association and pathway analysis." PLoS One **9**(4): e95866.
411 . Visscher, P. M., M. A. Brown, M. I. McCarthy and J. Yang (2012). "Five years of GWAS discovery." Am J Hum Genet **90**(1): 7-24.
412 . Visscher, P. M., M. E. Goddard, E. M. Derks and N. R. Wray (2012). "Evidence-based psychiatric genetics, AKA the false dichotomy between common and rare variant hypotheses." Mol Psychiatry **17**(5): 474-485.
413 . Voight, B. F., G. M. Peloso, M. Orho-Melander, R. Frikke-Schmidt, M. Barbalic, M. K. Jensen, . . . S. Kathiresan (2012). "Plasma HDL cholesterol and risk of myocardial infarction: a mendelian randomisation study." Lancet **380**(9841): 572-580.
414 . von Eckardstein, A., H. Funke, A. Henke, K. Altland, A. Benninghoven and G. Assmann (1989). "Apolipoprotein A-I variants. Naturally occurring substitutions of proline residues affect plasma concentration of apolipoprotein A-I." J Clin Invest **84**(6): 1722-1730.
415 . Wallace, C., S. J. Newhouse, P. Braund, F. Zhang, M. Tobin, M. Falchi, . . . P. B. Munroe (2008). "Genome-wide association study identifies genes for biomarkers of cardiovascular disease: serum urate and dyslipidemia." Am J Hum Genet **82**(1): 139-149.
416 . Waterworth, D. M., S. L. Ricketts, K. Song, L. Chen, J. H. Zhao, S. Ripatti, . . . M. S. Sandhu (2010). "Genetic variants influencing circulating lipid levels and risk of coronary artery disease." Arterioscler Thromb Vasc Biol **30**(11): 2264-2276.

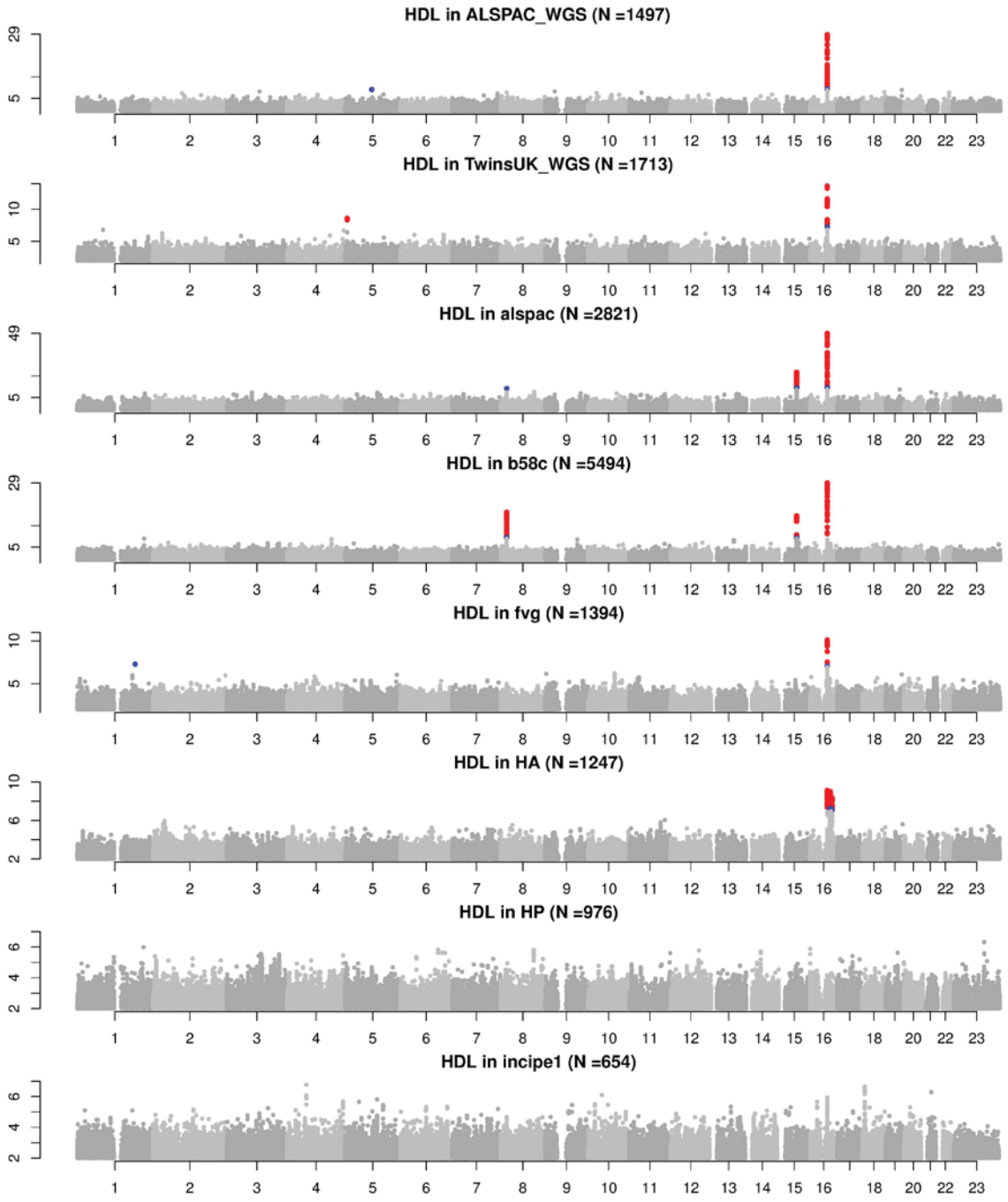
417 . Watowich, S. S., X. Xie, U. Klingmuller, J. Kere, M. Lindlof, S. Berglund and A. de la Chapelle (1999). "Erythropoietin receptor mutations associated with familial erythrocytosis cause hypersensitivity to erythropoietin in the heterozygous state." Blood **94**(7): 2530-2532.
418 . Webb, J., H. Gonna and K. K. Ray (2013). "Lipid management: maximising reduction of cardiac risk." Clin Med **13**(6): 618-620.
419 . Weiss, L. A., L. Pan, M. Abney and C. Ober (2006). "The sex-specific genetic architecture of quantitative traits in humans." Nat Genet **38**(2): 218-222.
420 . Weissglas-Volkov, D., C. A. Aguilar-Salinas, E. Nikkola, K. A. Deere, I. Cruz-Bautista, O. Arellano-Campos, . . . P. Pajukanta (2013). "Genomic study in Mexicans identifies a new locus for triglycerides and refines European lipid loci." J Med Genet **50**(5): 298-308.
421 . Wellcome Trust Case Control, C. (2007). "Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls." Nature **447**(7145): 661-678.
422 . Wellcome Trust Case Control, C., N. Craddock, M. E. Hurles, N. Cardin, R. D. Pearson, V. Plagnol, . . . P. Donnelly (2010). "Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls." Nature **464**(7289): 713-720.
423 . Wellcome Trust Case Control, C., J. B. Maller, G. McVean, J. Byrnes, D. Vukcevic, K. Palin, . . . P. Donnelly (2012). "Bayesian refinement of association signals for 14 loci in 3 common diseases." Nat Genet **44**(12): 1294-1301.
424 . Wellcome Trust Case Control Consortium (2007). "Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls." Nature **447**(7145): 661-678.
425 . Westra, H. J., M. J. Peters, T. Esko, H. Yaghootkar, C. Schurmann, J. Kettunen, . . . L. Franke (2013). "Systematic identification of trans eQTLs as putative drivers of known disease associations." Nat Genet **45**(10): 1238-1243.
426 . Whittall, R. A., S. Matheus, T. Cranston, G. J. Miller and S. E. Humphries (2002). "The intron 14 2140+5G>A variant in the low density lipoprotein receptor gene has no effect on plasma cholesterol levels." J Med Genet **39**(9): e57.
427 . Willems, J. M., S. Trompet, G. J. Blauw, R. G. Westendorp and A. J. de Craen (2010). "White blood cell count and C-reactive protein are independent predictors of mortality in the oldest old." J Gerontol A Biol Sci Med Sci **65**(7): 764-768.
428 . Willer, C. J., Y. Li and G. R. Abecasis (2010). "METAL: fast and efficient meta-analysis of genomewide association scans." Bioinformatics **26**(17): 2190-2191.
429 . Willer, C. J., S. Sanna, A. U. Jackson, A. Scuteri, L. L. Bonnycastle, R. Clarke, . . . G. R. Abecasis (2008). "Newly identified loci that influence lipid concentrations and risk of coronary artery disease." Nat Genet **40**(2): 161-169.
430 . Willer, C. J., E. M. Schmidt, S. Sengupta, G. M. Peloso, S. Gustafsson, S. Kanoni, . . . G. R. Abecasis (2013). "Discovery and refinement of loci associated with lipid levels." Nat Genet **45**(11): 1274-1283.
431 . Wilson, P. W., R. B. D'Agostino, D. Levy, A. M. Belanger, H. Silbershatz and W. B. Kannel (1998). "Prediction of coronary heart disease using risk factor categories." Circulation **97**(18): 1837-1847.
432 . Wilson, P. W., B. H. Nam, M. Pencina, R. B. D'Agostino, Sr., E. J. Benjamin and C. J. O'Donnell (2005). "C-reactive protein and risk of cardiovascular disease in men and women from the Framingham Heart Study." Arch Intern Med **165**(21): 2473-2478.
433 . Winkelmann, B. R., W. Marz, B. O. Boehm, R. Zotz, J. Hager, P. Hellstern, . . . L. S. Group (2001). "Rationale and design of the LURIC study--a resource for functional genomics, pharmacogenomics and long-term prognosis of cardiovascular disease." Pharmacogenomics **2**(1 Suppl 1): S1-73.
434 . Wolfs, M. G., M. H. Hofker, C. Wijmenga and T. W. van Haeften (2009). "Type 2 Diabetes Mellitus: New Genetic Insights will Lead to New Therapeutics." Curr Genomics **10**(2): 110-118.
435 . Wong, N. D. (2014). "Epidemiological studies of CHD and the evolution of preventive cardiology." Nat Rev Cardiol **11**(5): 276-289.

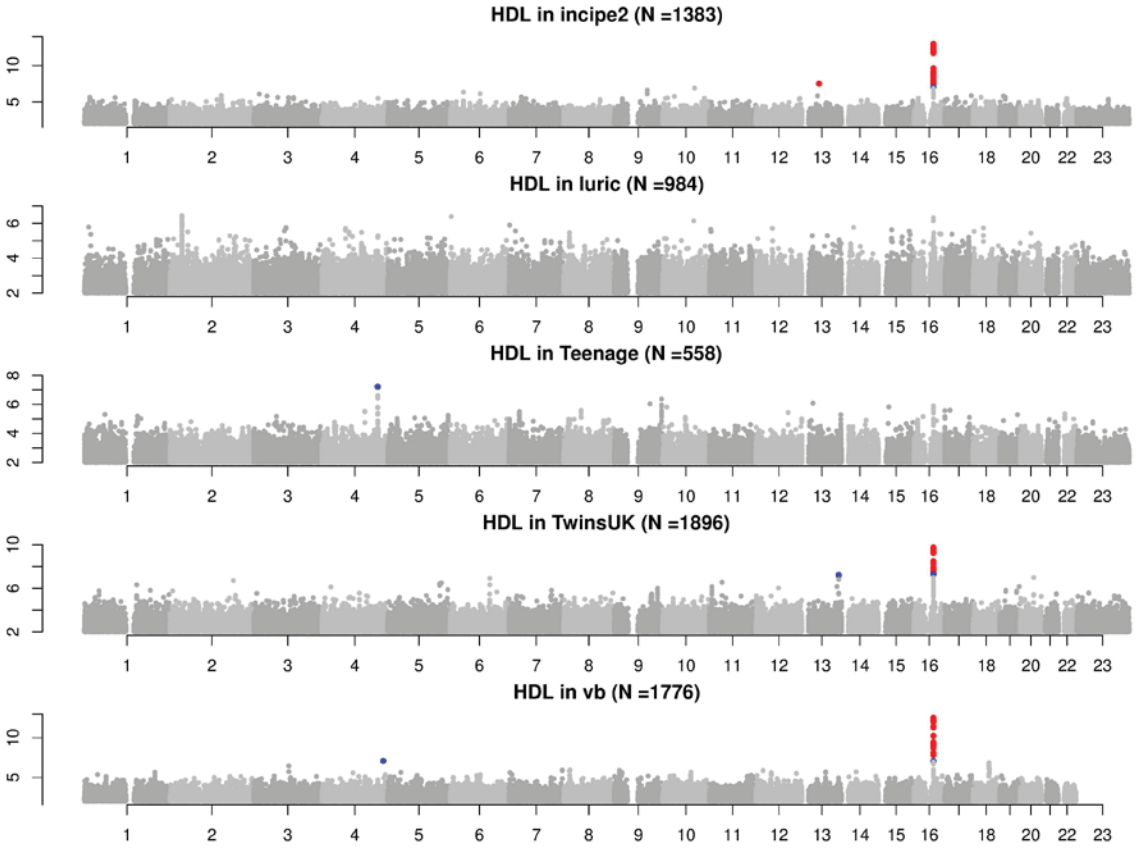
436 . Wood, A. R., T. Esko, J. Yang, S. Vedantam, T. H. Pers, S. Gustafsson, . . . T. M. Frayling (2014). "Defining the role of common variation in the genomic and biological architecture of adult human height." Nat Genet **46**(11): 1173-1186.
437 . Wu, M. C., S. Lee, T. Cai, Y. Li, M. Boehnke and X. Lin (2011). "Rare-variant association testing for sequencing data with the sequence kernel association test." American journal of human genetics **89**(1): 82-93.
438 . Wu, M. C., S. Lee, T. Cai, Y. Li, M. Boehnke and X. Lin (2011). "Rare-variant association testing for sequencing data with the sequence kernel association test." Am J Hum Genet **89**(1): 82-93.
439 . Wu, Y., A. F. Marvelle, J. Li, D. C. Croteau-Chonka, A. B. Feranil, C. W. Kuzawa, . . . K. L. Mohlke (2013). "Genetic association with lipids in Filipinos: waist circumference modifies an APOA5 effect on triglyceride levels." J Lipid Res **54**(11): 3198-3205.
440 . Wu, Y., T. W. McDade, C. W. Kuzawa, J. Borja, Y. Li, L. S. Adair, . . . L. A. Lange (2012). "Genome-wide association with C-reactive protein levels in CLHNS: evidence for the CRP and HNF1A loci and their interaction with exposure to a pathogenic environment." Inflammation **35**(2): 574-583.
441 . Xu, C., I. Tachmazidou, K. Walter, A. Ciampi, E. Zeggini, C. M. T. Greenwood and t. U. K. Consortium (2014). "Estimating Genome-Wide Significance for Whole-Genome Sequencing Studies." Genetic Epidemiology: n/a-n/a.
442 . Xu, J., C. Peng, V. G. Sankaran, Z. Shao, E. B. Esrick, B. G. Chong, . . . S. H. Orkin (2011). "Correction of sickle cell disease in adult mice by interference with fetal hemoglobin silencing." Science **334**(6058): 993-996.
443 . Yan, J., T. Takahashi, T. Ohura, H. Adachi, I. Takahashi, E. Ogawa, . . . A. Koizumi (2013). "Combined linkage analysis and exome sequencing identifies novel genes for familial goiter." J Hum Genet **58**(6): 366-377.
444 . Yang, J., B. Benyamin, B. P. McEvoy, S. Gordon, A. K. Henders, D. R. Nyholt, . . . P. M. Visscher (2010). "Common SNPs explain a large proportion of the heritability for human height." Nat Genet **42**(7): 565-569.
445 . Yang, J., T. A. Manolio, L. R. Pasquale, E. Boerwinkle, N. Caporaso, J. M. Cunningham, . . . P. M. Visscher (2011). "Genome partitioning of genetic variation for complex traits using common SNPs." Nat Genet **43**(6): 519-525.
446 . Yang, Q., S. Kathiresan, J. P. Lin, G. H. Tofler and C. J. O'Donnell (2007). "Genome-wide association and linkage analyses of hemostatic factors and hematological phenotypes in the Framingham Heart Study." BMC Med Genet **8 Suppl 1**: S12.
447 . Young, S. G., S. J. Bertics, L. K. Curtiss, B. W. Dubois and J. L. Witztum (1987). "Genetic analysis of a kindred with familial hypobetalipoproteinemia. Evidence for two separate gene defects: one associated with an abnormal apolipoprotein B species, apolipoprotein B-37; and a second associated with low plasma concentrations of apolipoprotein B-100." J Clin Invest **79**(6): 1842-1851.
448 . Yusuf, S., S. Hawken, S. Ounpuu, T. Dans, A. Avezum, F. Lanas, . . . I. S. Investigators (2004). "Effect of potentially modifiable risk factors associated with myocardial infarction in 52 countries (the INTERHEART study): case-control study." Lancet **364**(9438): 937-952.
449 . Zeggini, E. (2014). "Genetic characterisation of Greek population isolates reveals strong genetic drift at missense and trait-associated variants." under review.
450 . Zeggini, E. (2014). "Using genetically isolated populations to understand the genomic basis of disease." Genome Med **6**(10): 83.
451 . Zeggini, E., M. N. Weedon, C. M. Lindgren, T. M. Frayling, K. S. Elliott, H. Lango, . . . A. T. Hattersley (2007). "Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes." Science **316**(5829): 1336-1341.
452 . Zelcer, N., C. Hong, R. Boyadjian and P. Tontonoz (2009). "LXR regulates cholesterol uptake through Idol-dependent ubiquitination of the LDL receptor." Science **325**(5936): 100-104.
453 . Zeng, S. M., J. Yankowitz, J. A. Widness and R. G. Strauss (2001). "Etiology of differences in hematocrit between males and females: sequence-based polymorphisms in erythropoietin and its receptor." J Gend Specif Med **4**(1): 35-40.

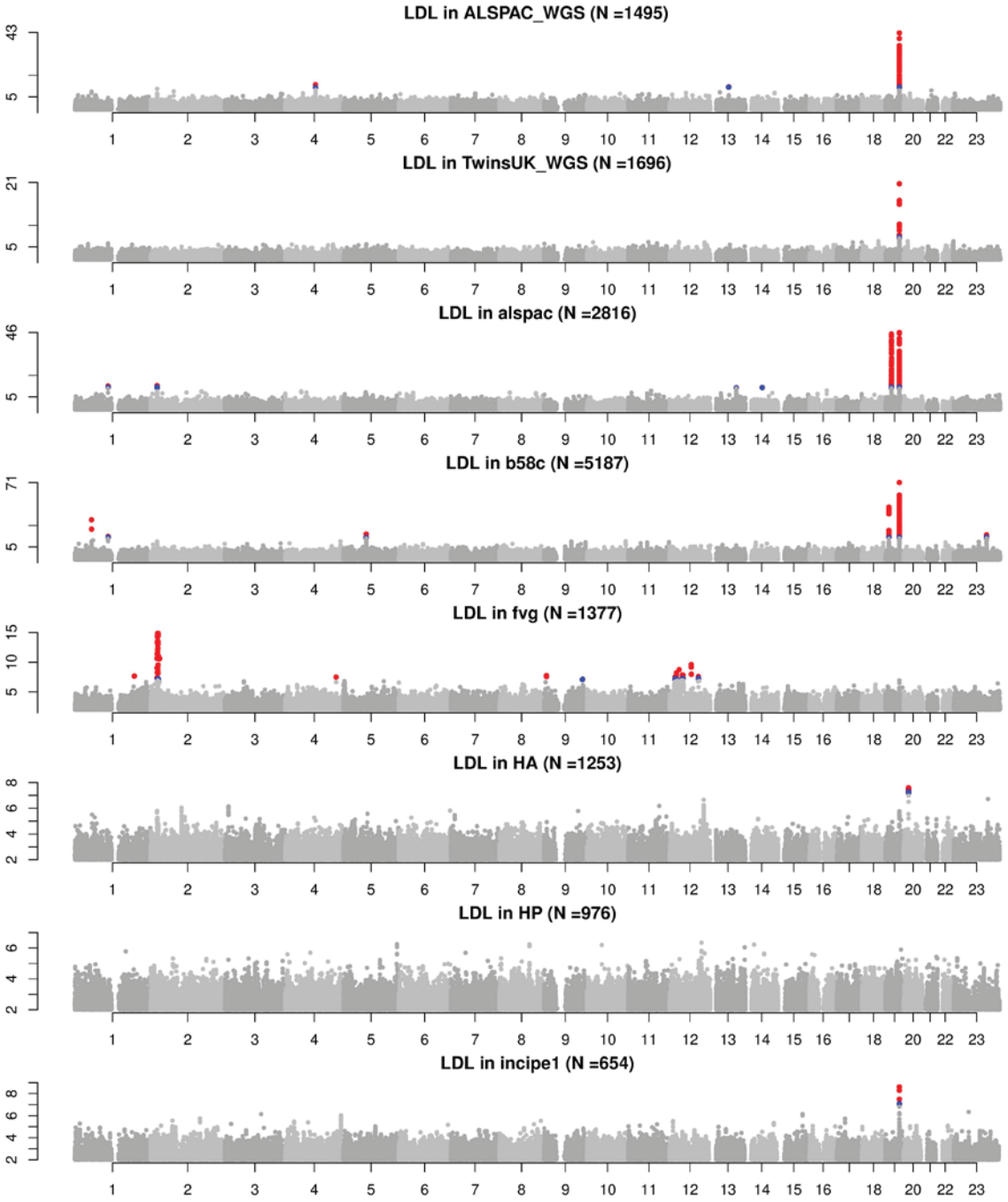
454 . Zhou, L., M. He, Z. Mo, C. Wu, H. Yang, D. Yu, . . . T. Wu (2013). "A genome wide association study identifies common variants associated with lipid levels in the Chinese population." PLoS One **8**(12): e82420.
455 . Zhou, X. and M. Stephens (2012). "Genome-wide efficient mixed-model analysis for association studies." Nat Genet **44**(7): 821-824.
456 . Zuk, O., S. F. Schaffner, K. Samocha, R. Do, E. Hechter, S. Kathiresan, . . . E. S. Lander (2014). "Searching for missing heritability: designing rare variant association studies." Proc Natl Acad Sci U S A **111**(4): E455-464.
457 . Zwaka, T. P., V. Hombach and J. Torzewski (2001). "C-reactive protein-mediated low density lipoprotein uptake by macrophages: implications for atherosclerosis." Circulation **103**(9): 1194-1197.

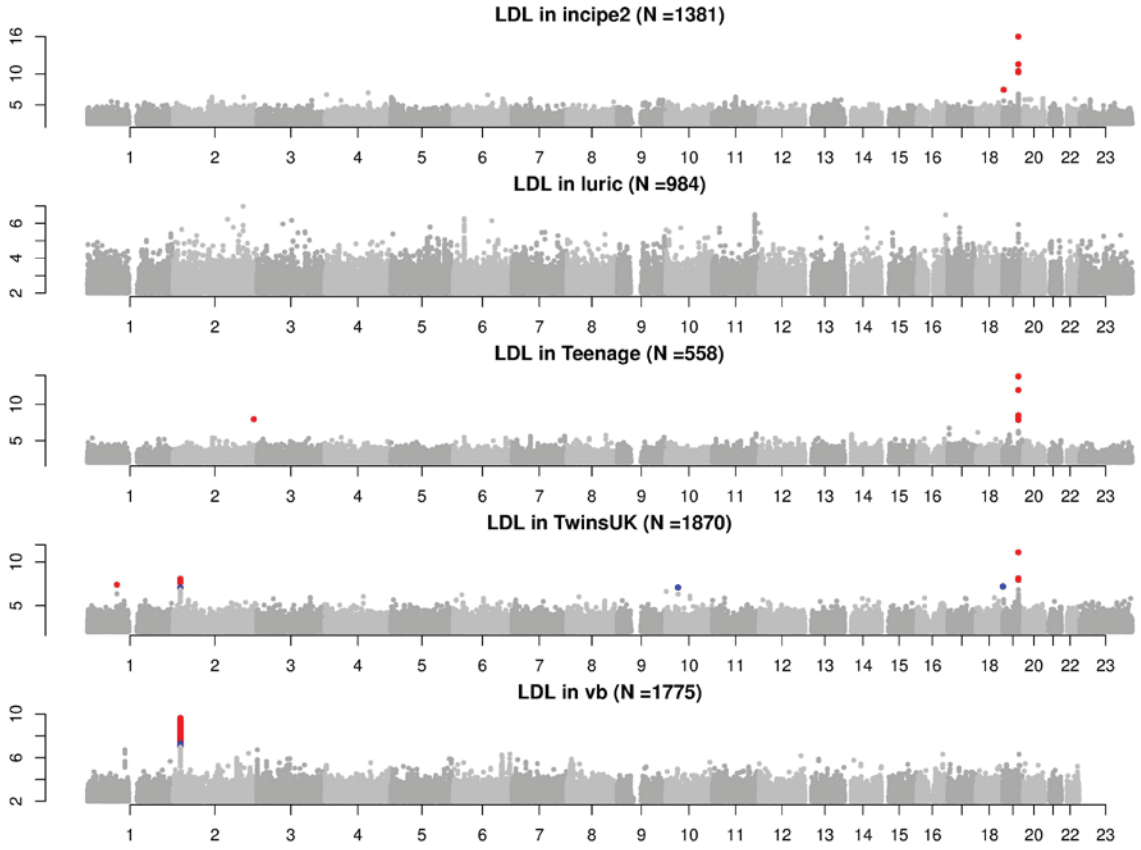
Appendix.

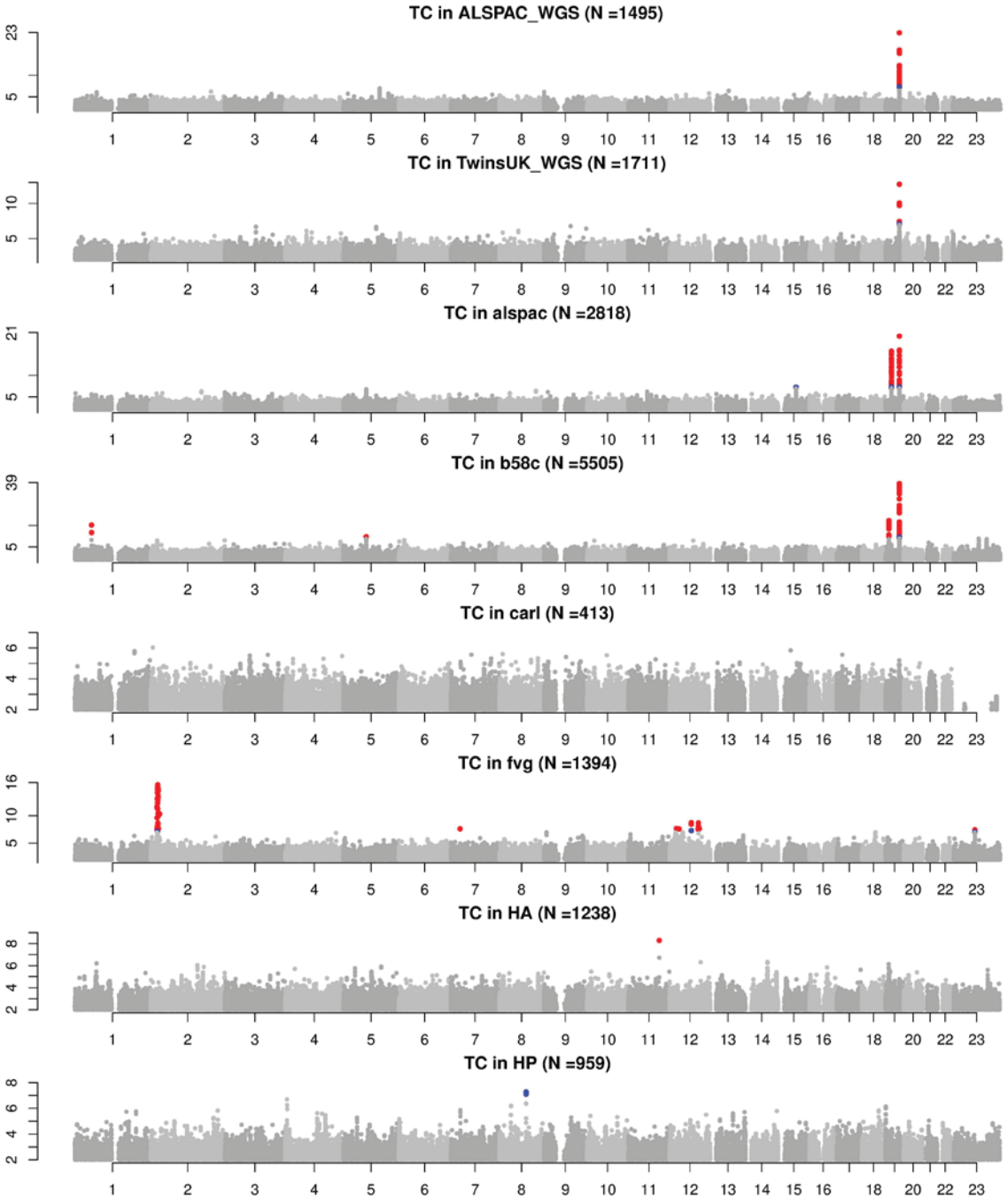
Appendix 1 Manhattan plots of individual GWA

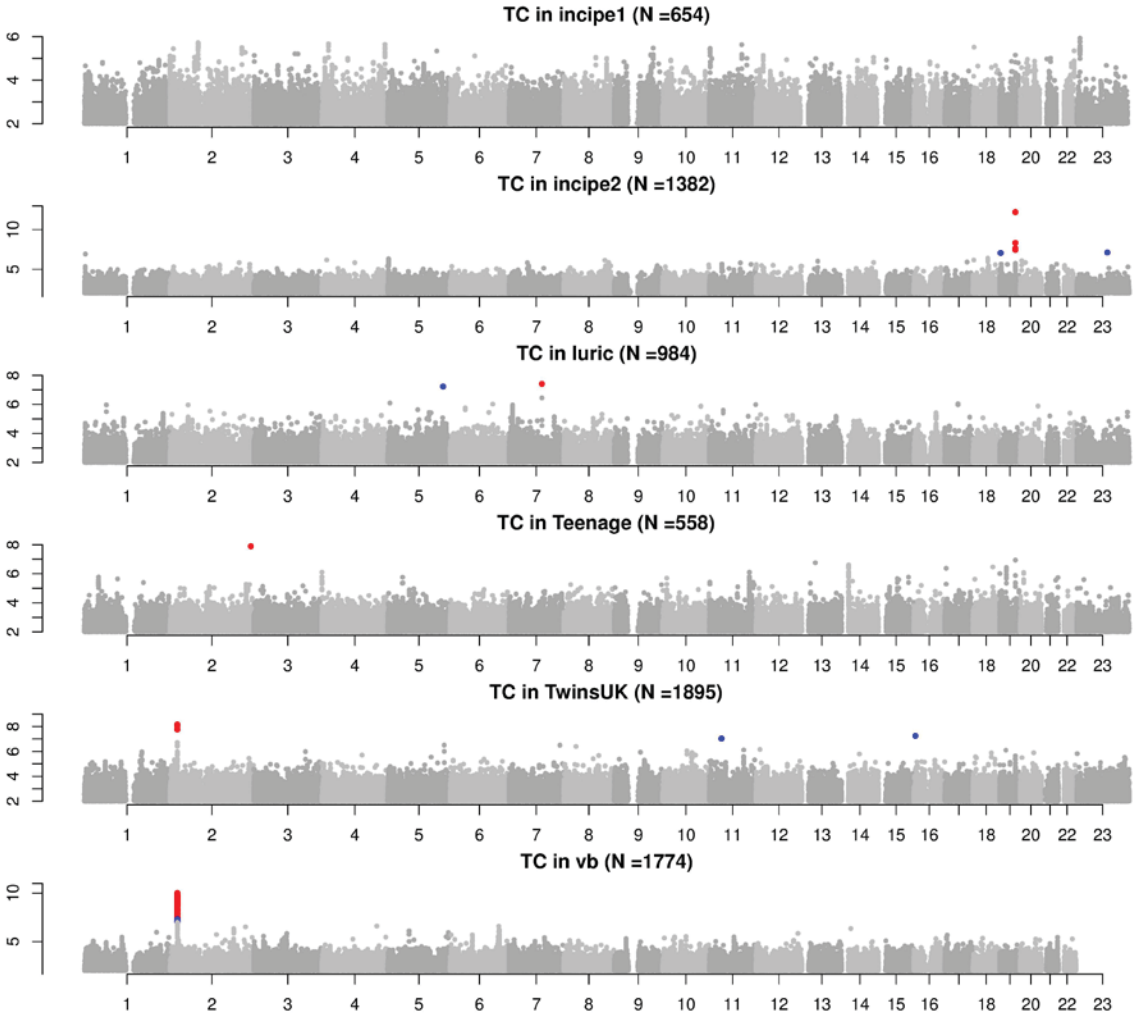

HDL in ALSPAC_WGS (N =1497)
HDL in TwinsUK_WGS (N =1713)
HDL in alspac (N =2821)
HDL in b58c (N =5494)
HDL in fvg (N =1394)
HDL in HA (N =1247)
HDL in HP (N =976)
HDL in incipe1 (N =654)
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 18 20 22 23

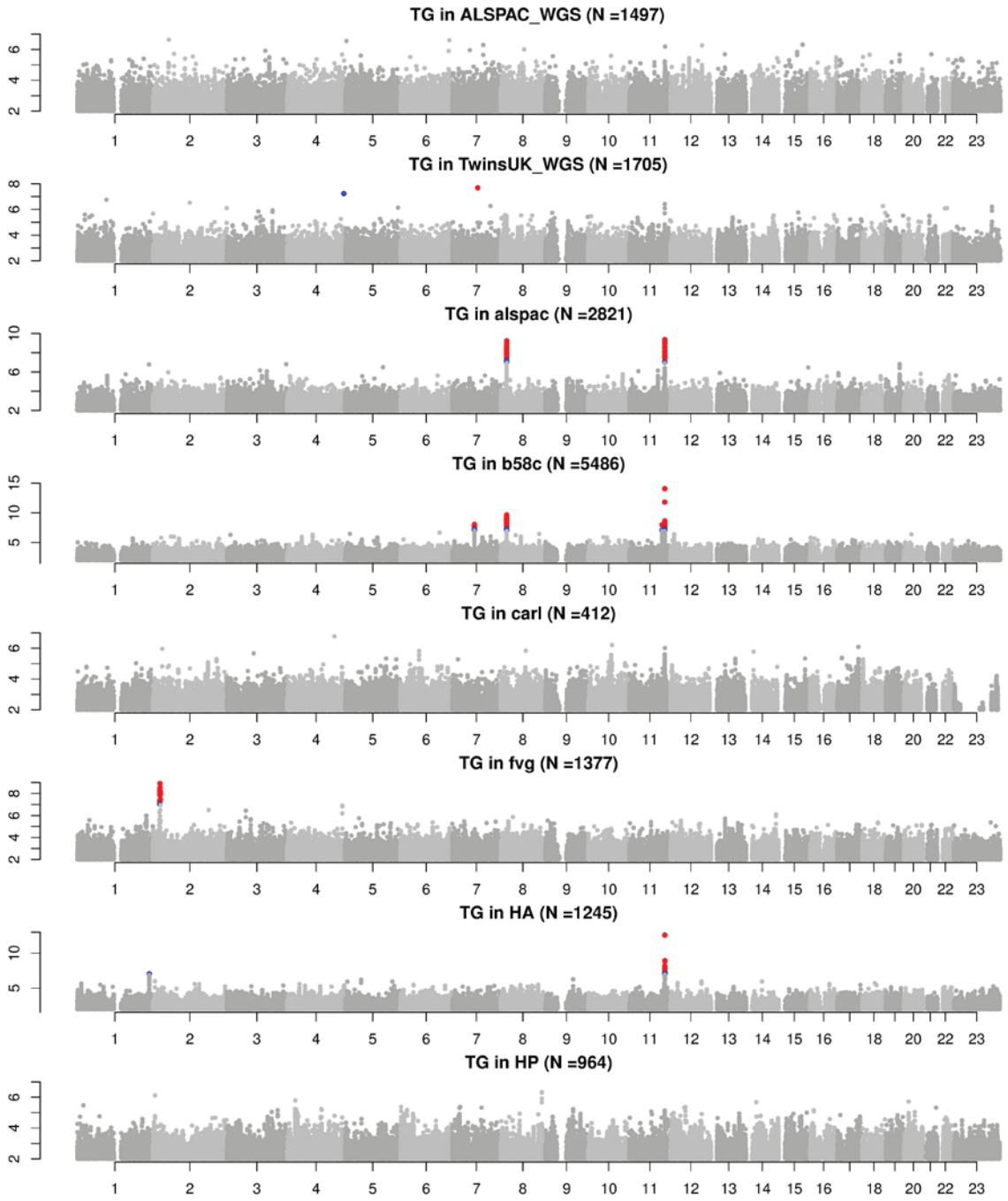

HDL in incipe2 (N =1383)
HDL in luric (N =984)
HDL in Teenage (N =558)
HDL in TwinsUK (N =1896)
HDL in vb (N =1776)
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 18 20 22 23

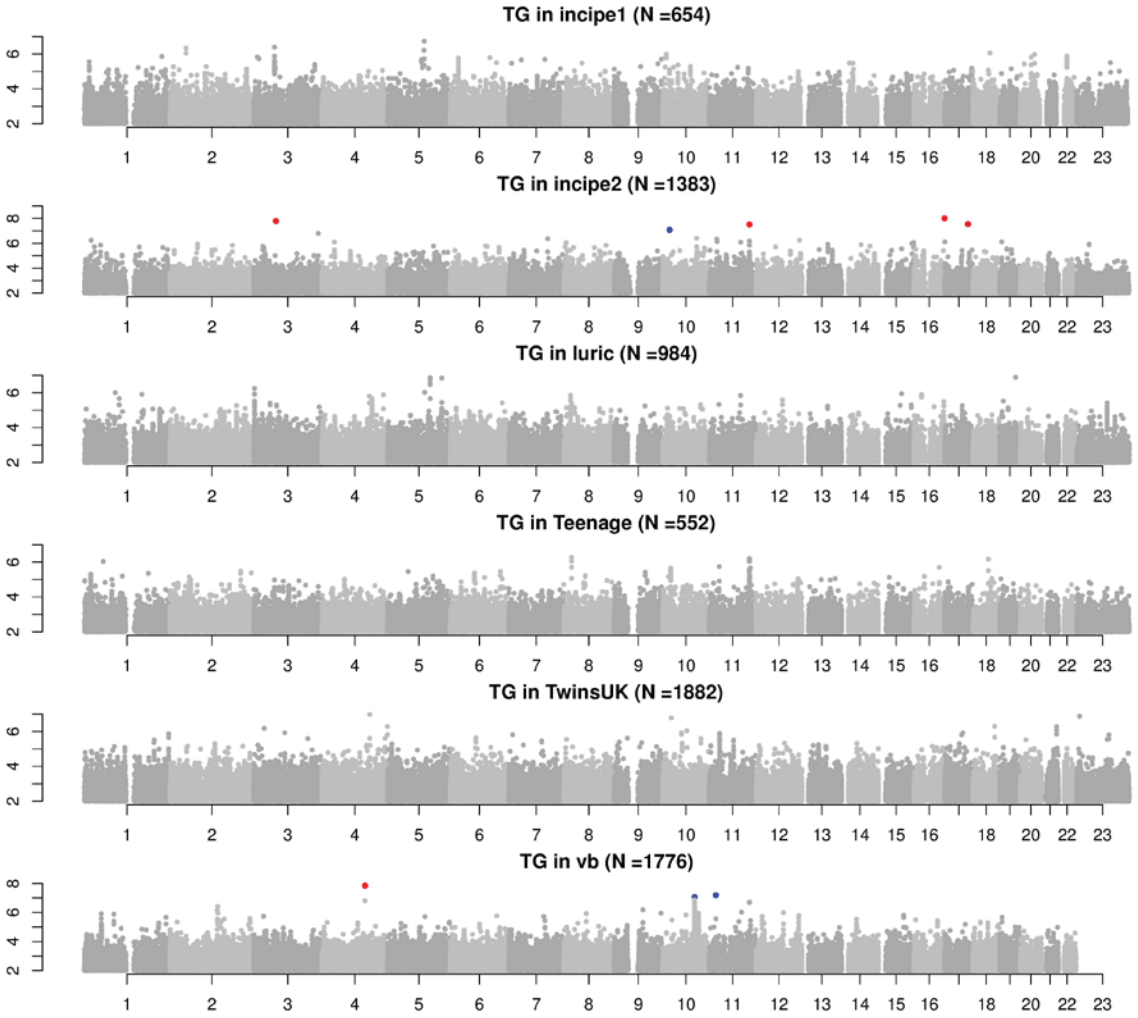

LDL in ALSPAC_WGS (N =1495)
LDL in TwinsUK_WGS (N =1696)
LDL in alspac (N =2816)
LDL in b58c (N =5187)
LDL in fvg (N =1377)
LDL in HA (N =1253)
LDL in HP (N =976)
LDL in incipe1 (N =654)
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 18 20 22 23

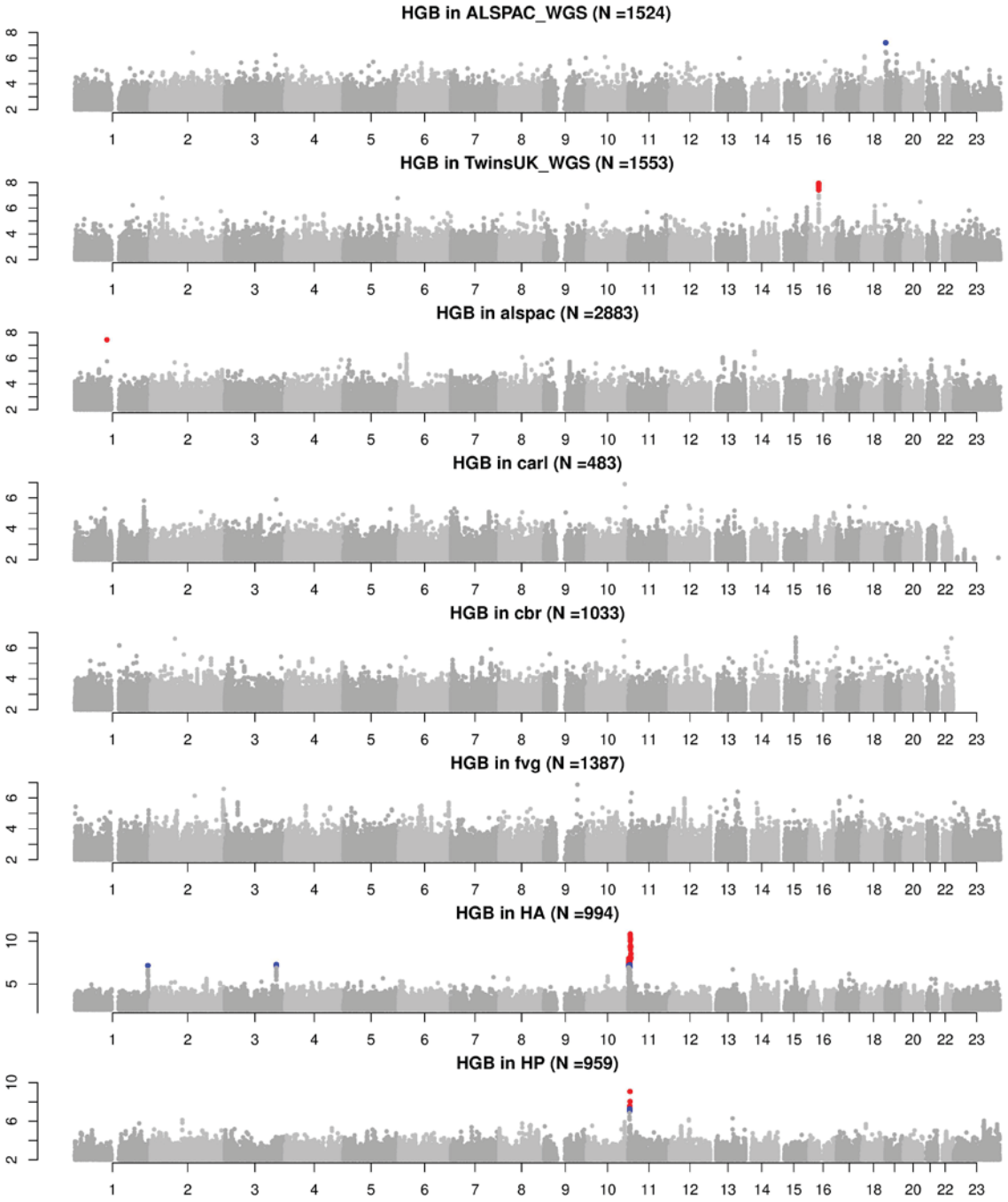

LDL in incipe2 (N =1381)
LDL in luric (N =984)
LDL in Teenage (N =558)
LDL in TwinsUK (N =1870)
LDL in vb (N =1775)
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 18 20 22 23

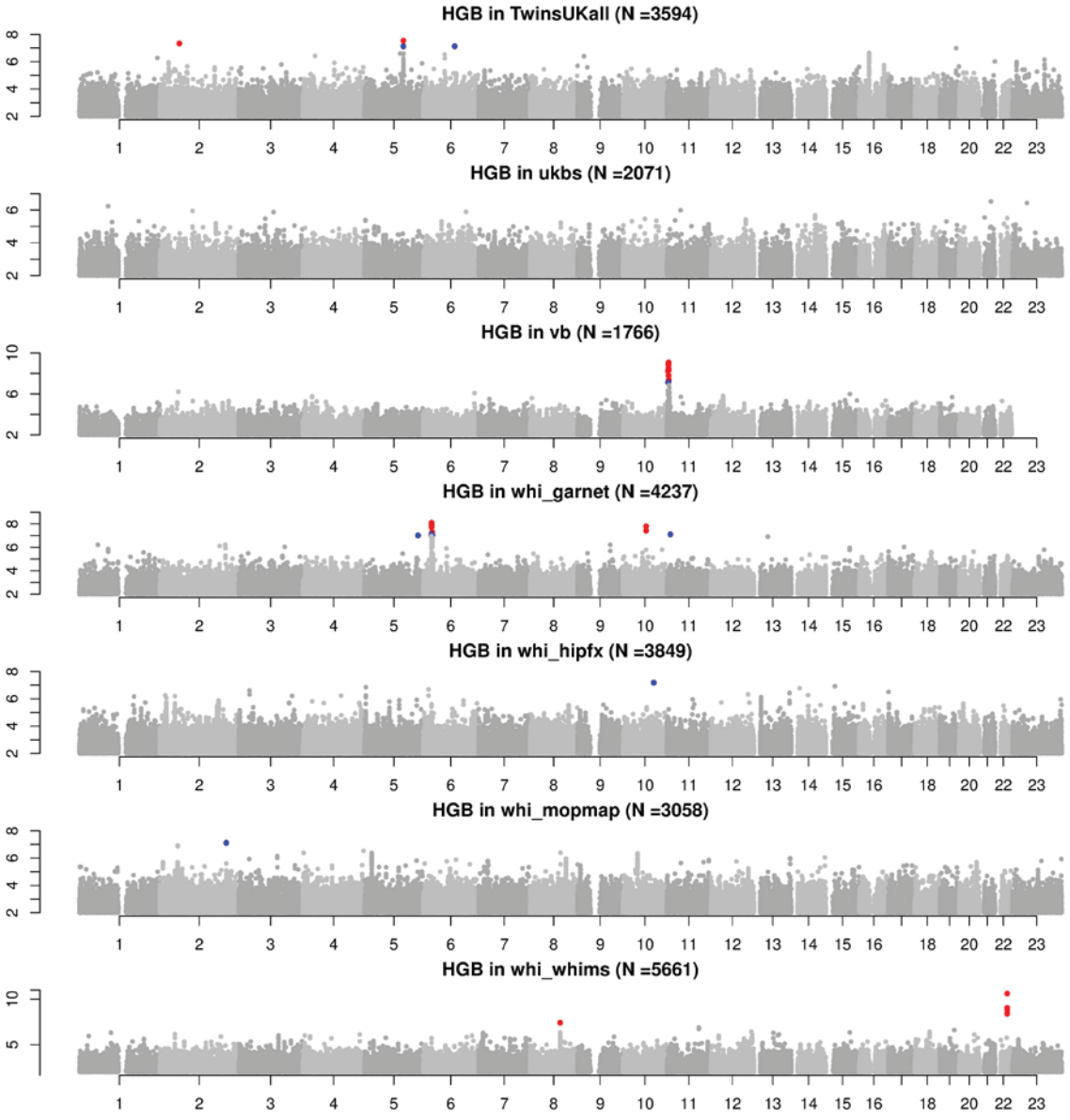

TC in ALSPAC_WGS (N =1495)
TC in TwinsUK_WGS (N =1711)
TC in alspac (N =2818)
TC in b58c (N =5505)
TC in carl (N =413)
TC in fvg (N =1394)
TC in HA (N =1238)
TC in HP (N =959)
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 18 20 22 23

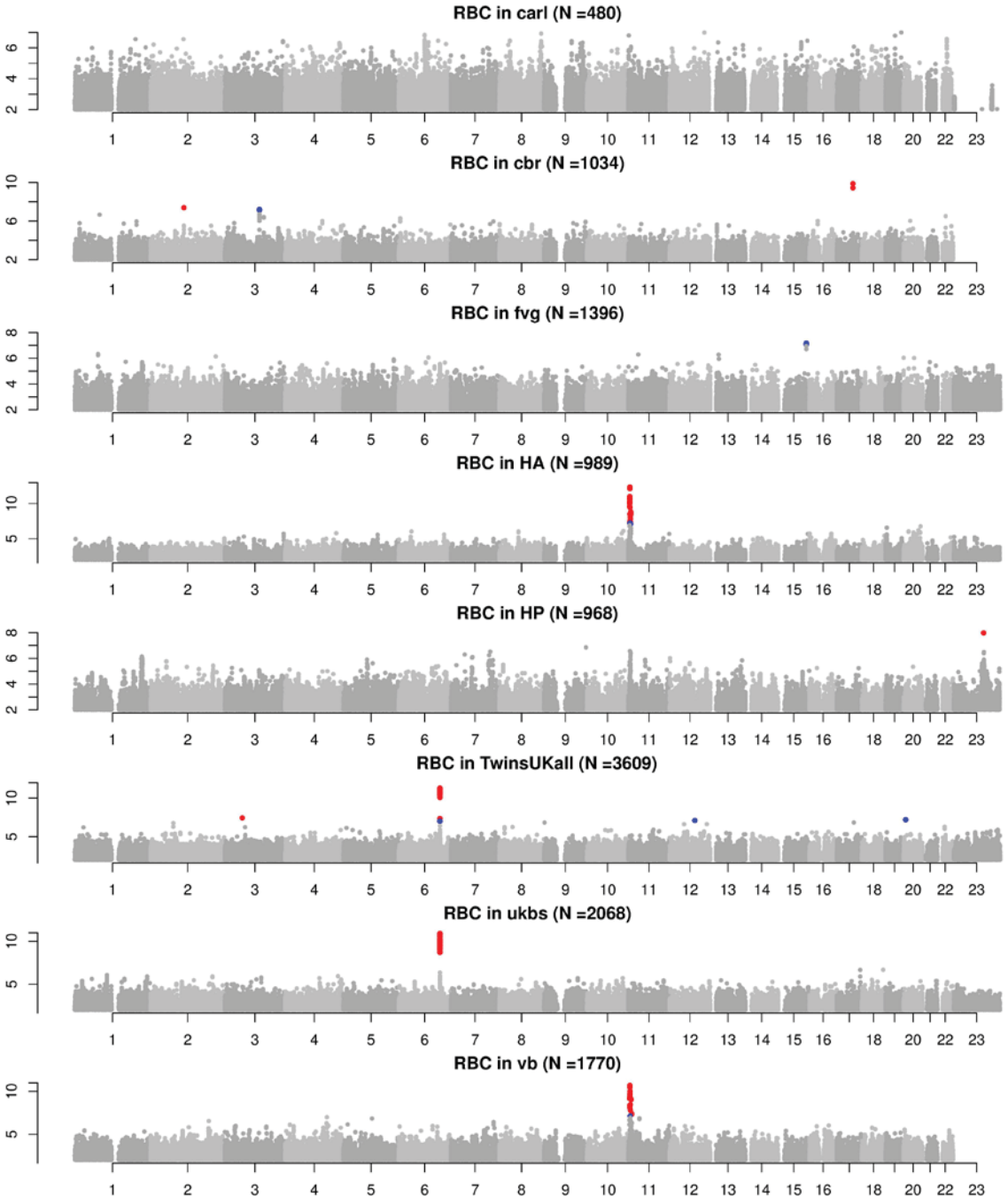

TC in incipe1 (N =654)
TC in incipe2 (N =1382)
TC in luric (N =984)
TC in Teenage (N =558)
TC in TwinsUK (N =1895)
TC in vb (N =1774)
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 18 20 22 23

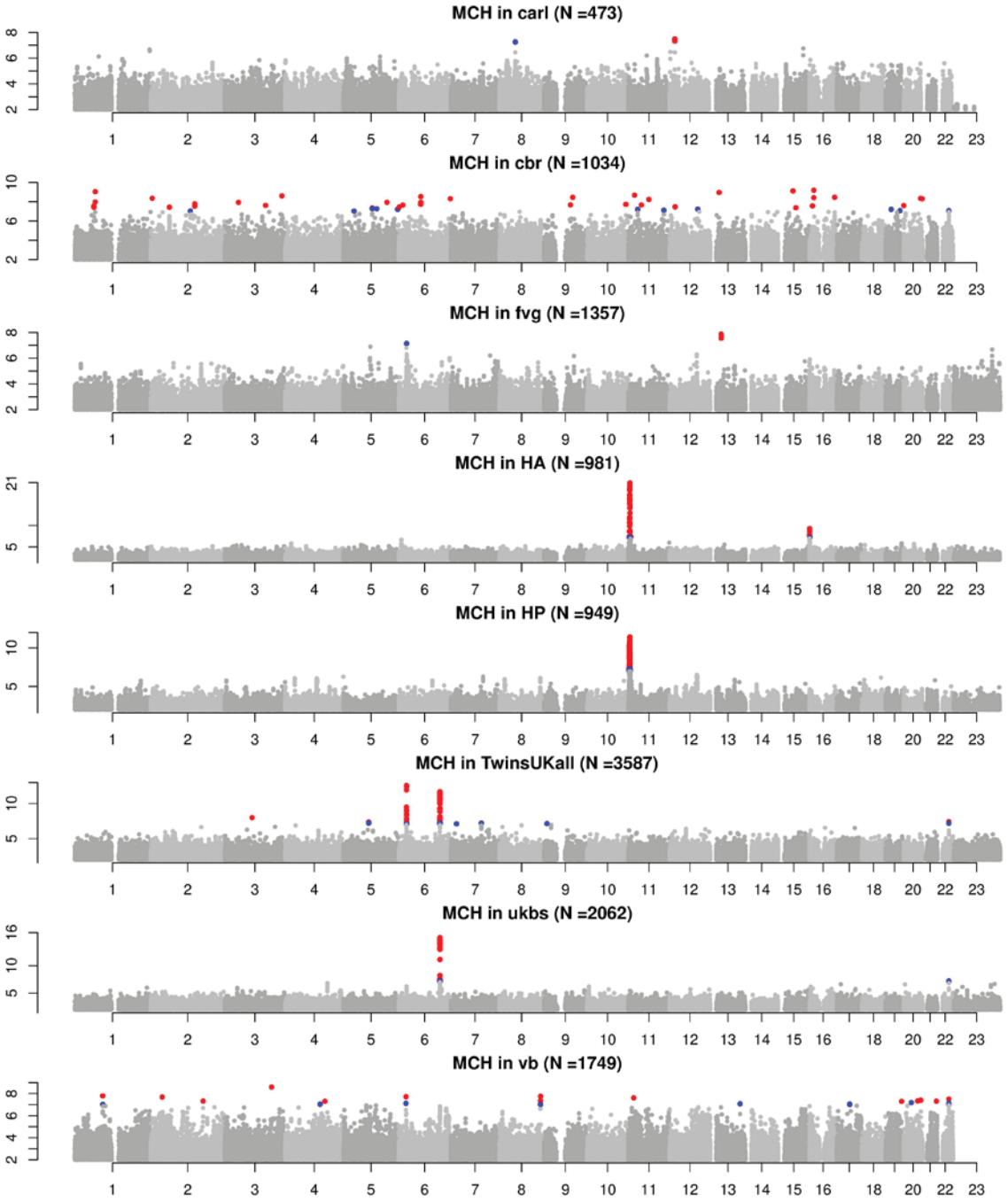

TG in ALSPAC_WGS (N =1497)
TG in TwinsUK_WGS (N =1705)
TG in alspac (N =2821)
TG in b58c (N =5486)
TG in carl (N =412)
TG in fvg (N =1377)
TG in HA (N =1245)
TG in HP (N =964)
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 18 20 22 23

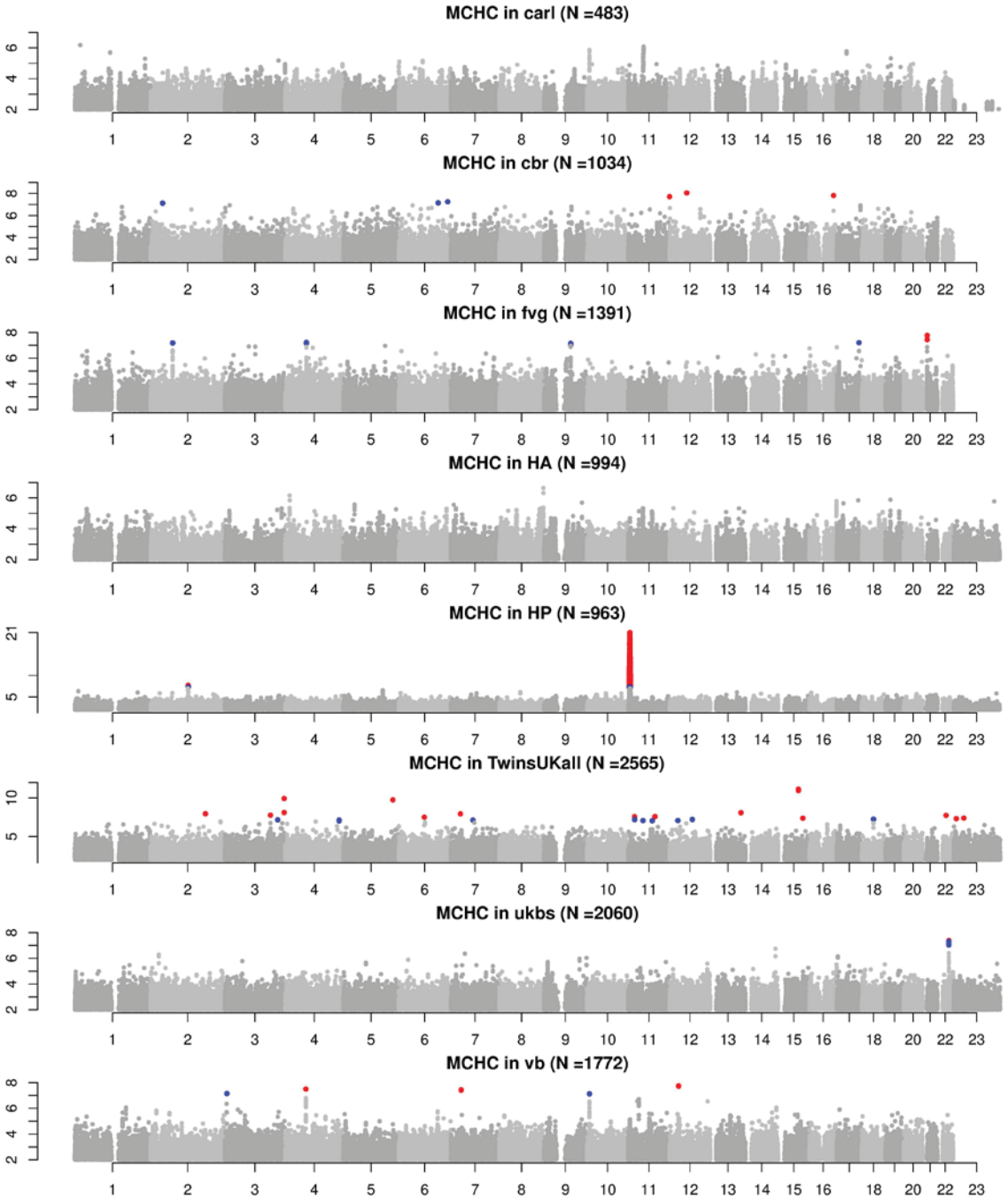

TG in incipe1 (N =654)
TG in incipe2 (N =1383)
TG in luric (N =984)
TG in Teenage (N =552)
TG in TwinsUK (N =1882)
TG in vb (N =1776)
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 18 20 22 23

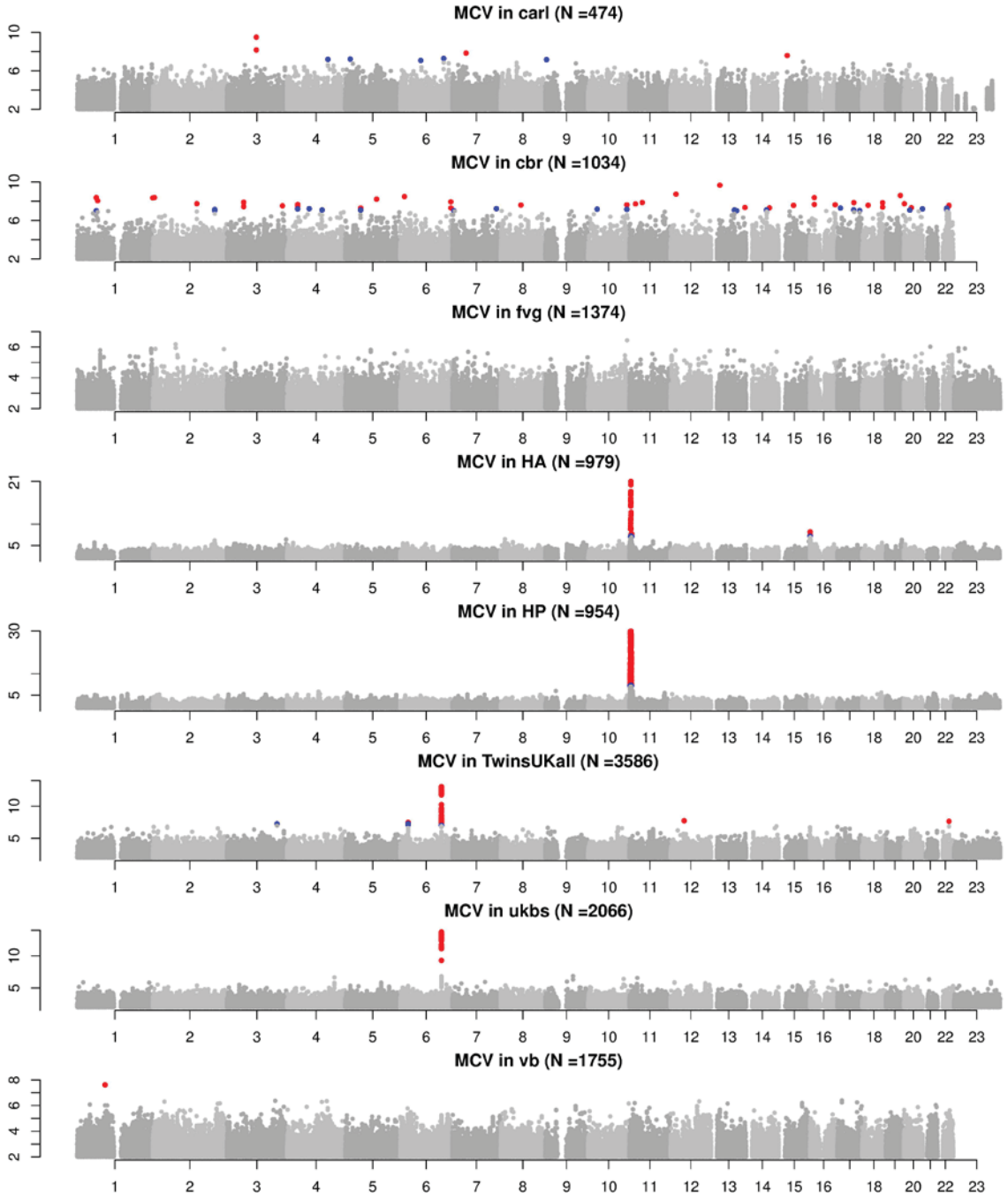

HGB in ALSPAC_WGS (N =1524)
HGB in TwinsUK_WGS (N =1553)
HGB in alspac (N =2883)
HGB in carl (N =483)
HGB in cbr (N =1033)
HGB in fvg (N =1387)
HGB in HA (N =994)
HGB in HP (N =959)

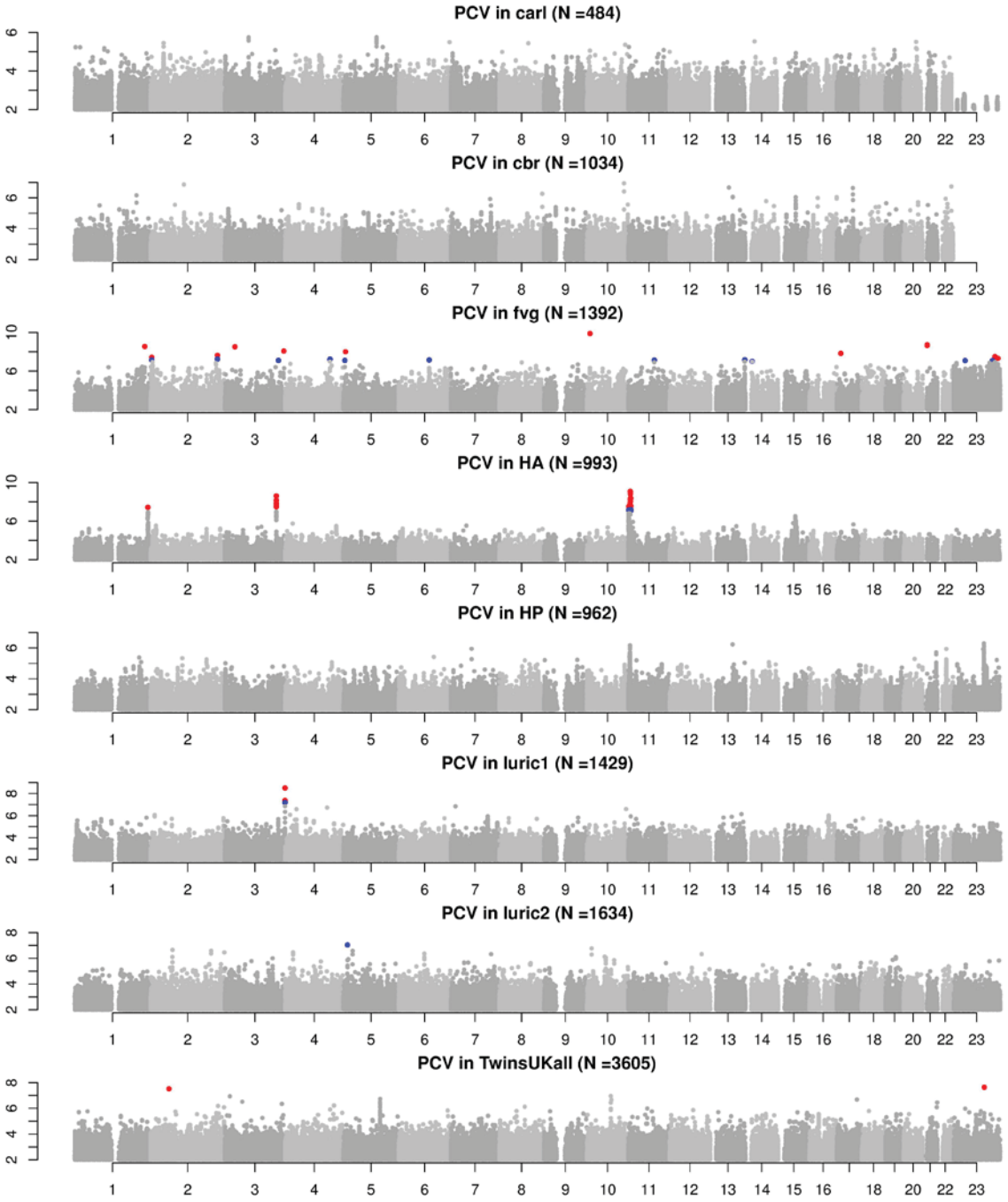

HGB in TwinsUKall (N =3594)
HGB in ukbs (N =2071)
HGB in vb (N =1766)
HGB in whi_garnet (N =4237)
HGB in whi_hipfx (N =3849)
HGB in whi_mopmap (N =3058)
HGB in whi_whims (N =5661)

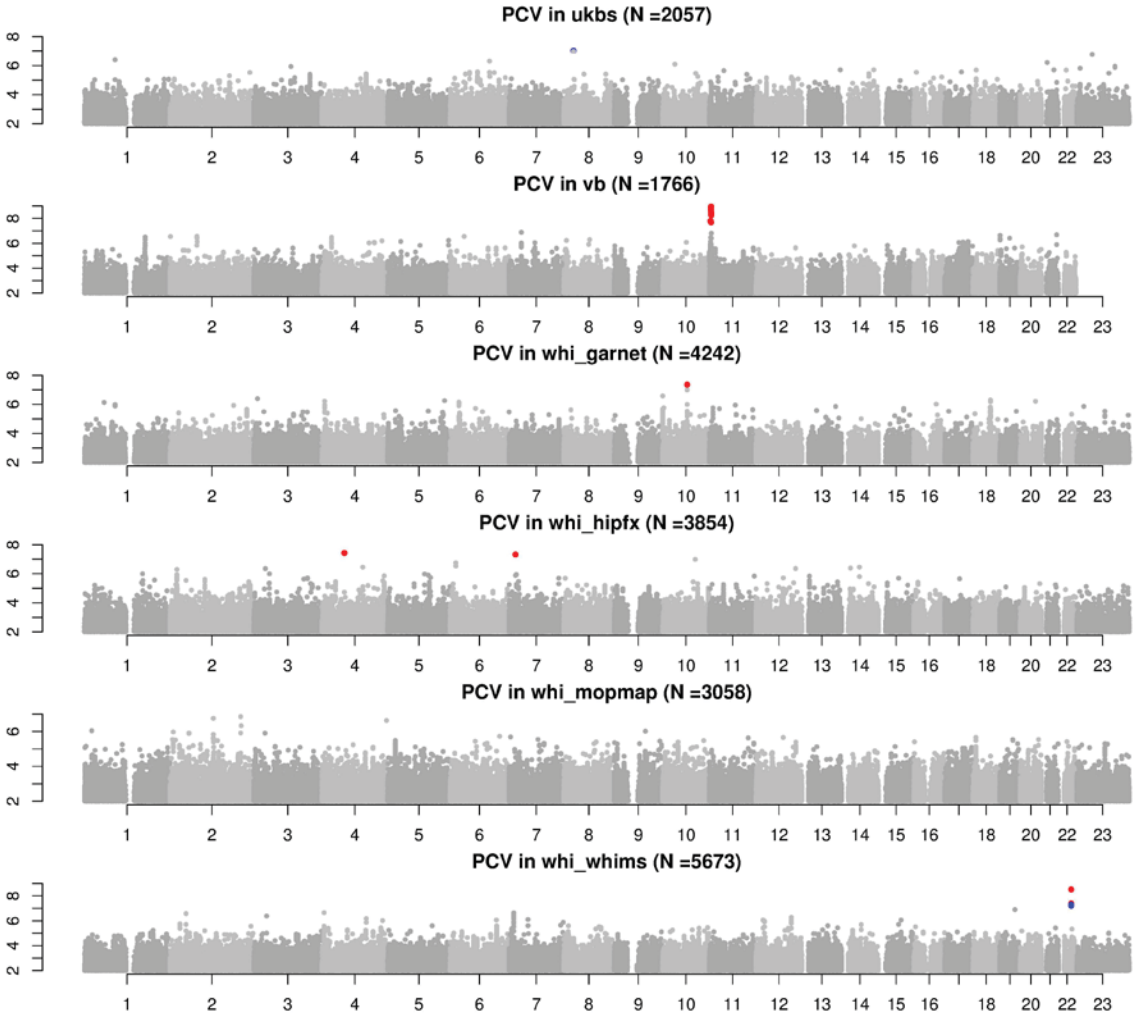

RBC in carl (N =480)
RBC in cbr (N =1034)
RBC in fvg (N =1396)
RBC in HA (N =989)
RBC in HP (N =968)
RBC in TwinsUKall (N =3609)
RBC in ukbs (N =2068)
RBC in vb (N =1770)
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 18 20 22 23

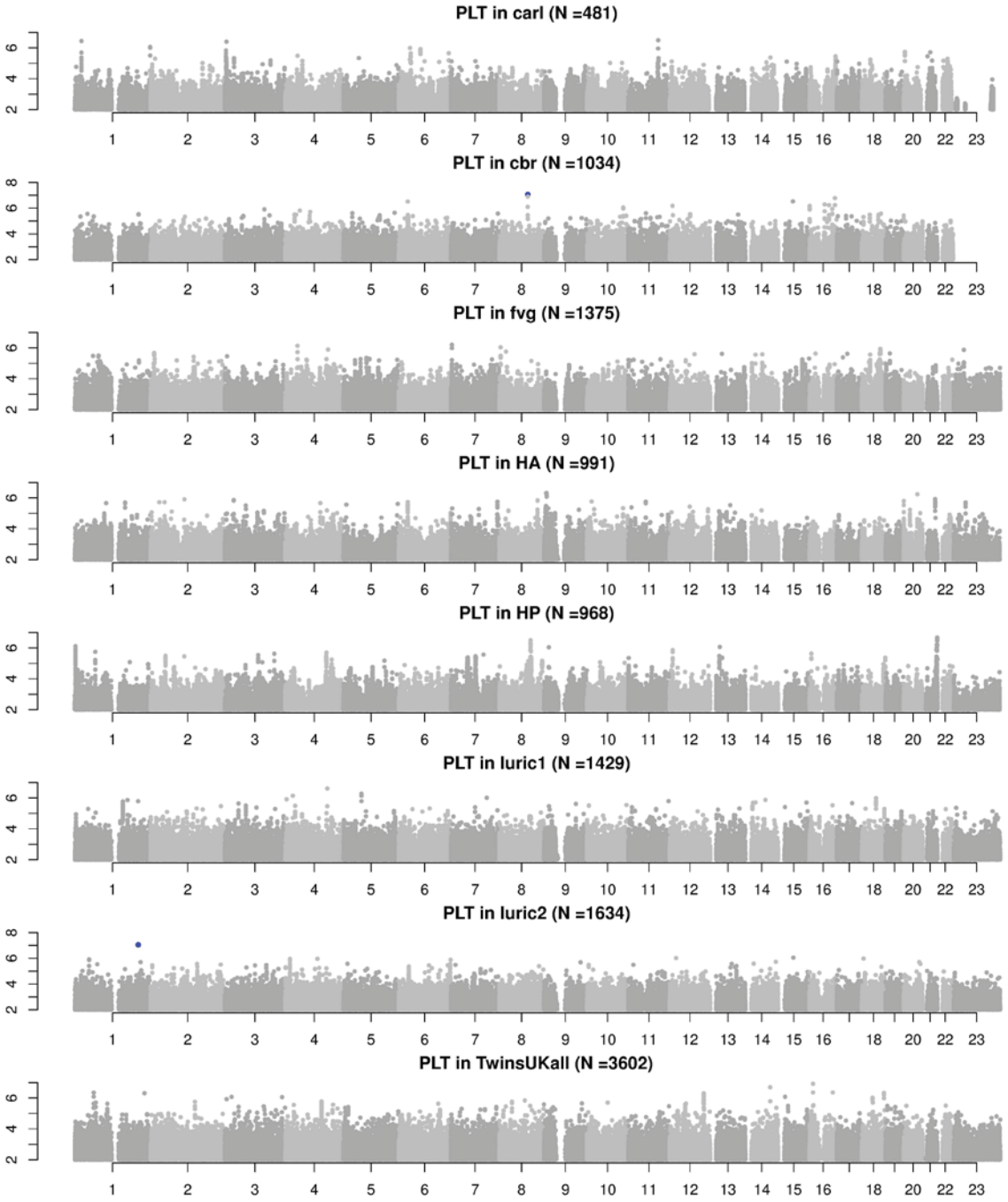

MCH in carl (N =473)
MCH in cbr (N =1034)
MCH in fvg (N =1357)
MCH in HA (N =981)
MCH in HP (N =949)
MCH in TwinsUKall (N =3587)
MCH in ukbs (N =2062)
MCH in vb (N =1749)

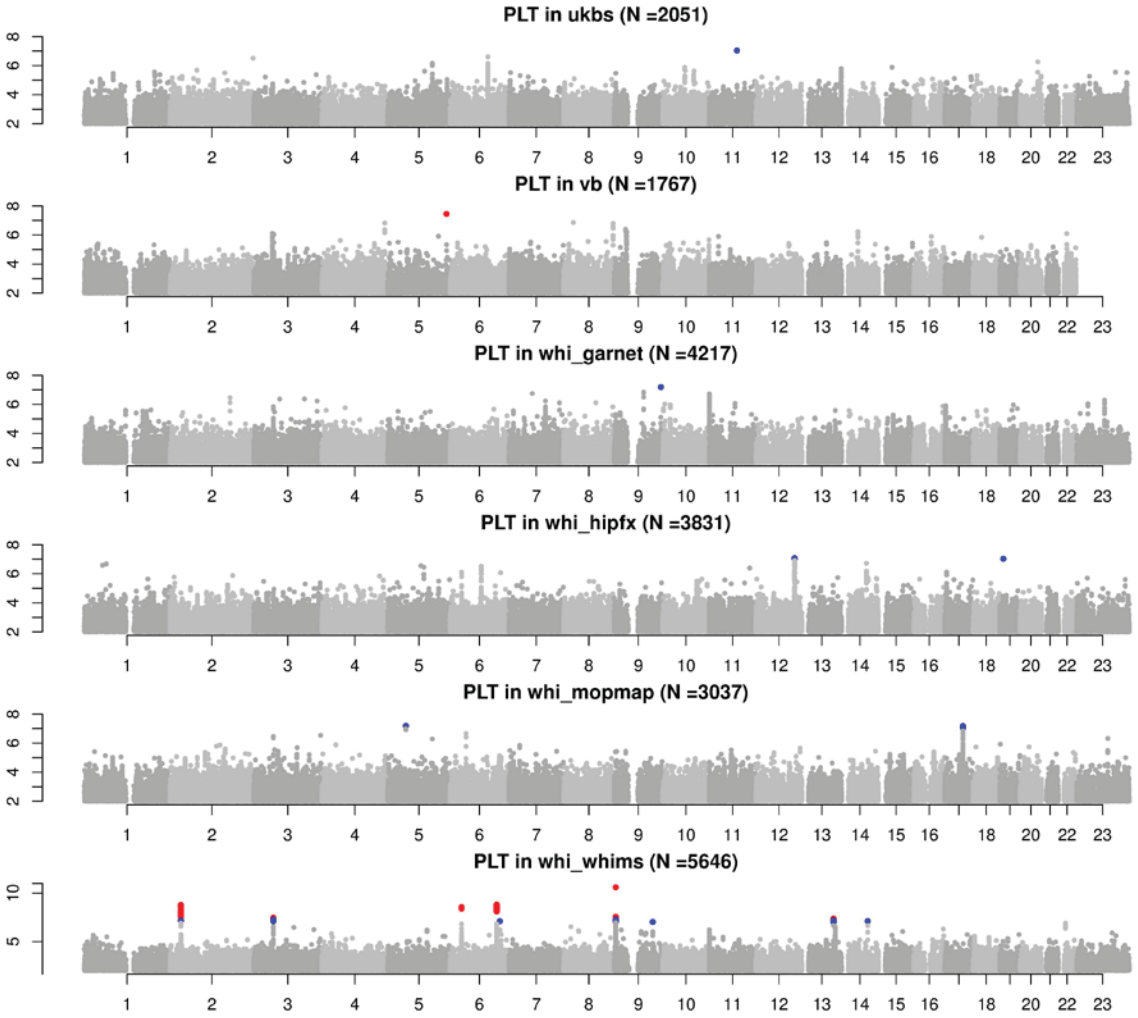

MCHC in carl (N =483)
MCHC in cbr (N =1034)
MCHC in fvg (N =1391)
MCHC in HA (N =994)
MCHC in HP (N =963)
MCHC in TwinsUKall (N =2565)
MCHC in ukbs (N =2060)
MCHC in vb (N =1772)
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 18 20 22 23

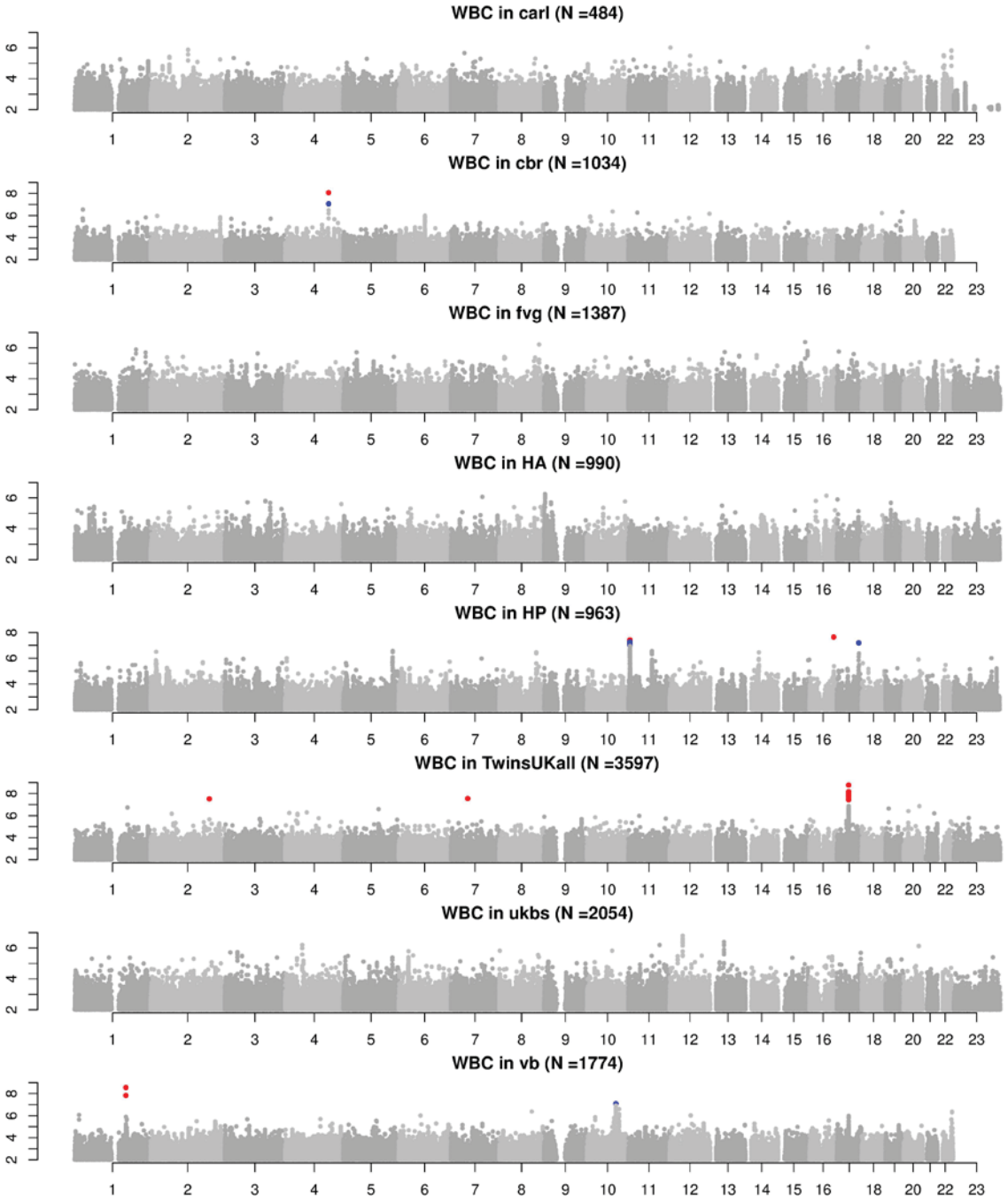

MCV in carl (N =474)
MCV in cbr (N =1034)
MCV in fvg (N =1374)
MCV in HA (N =979)
MCV in HP (N =954)
MCV in TwinsUKall (N =3586)
MCV in ukbs (N =2066)
MCV in vb (N =1755)
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 18 20 22 23

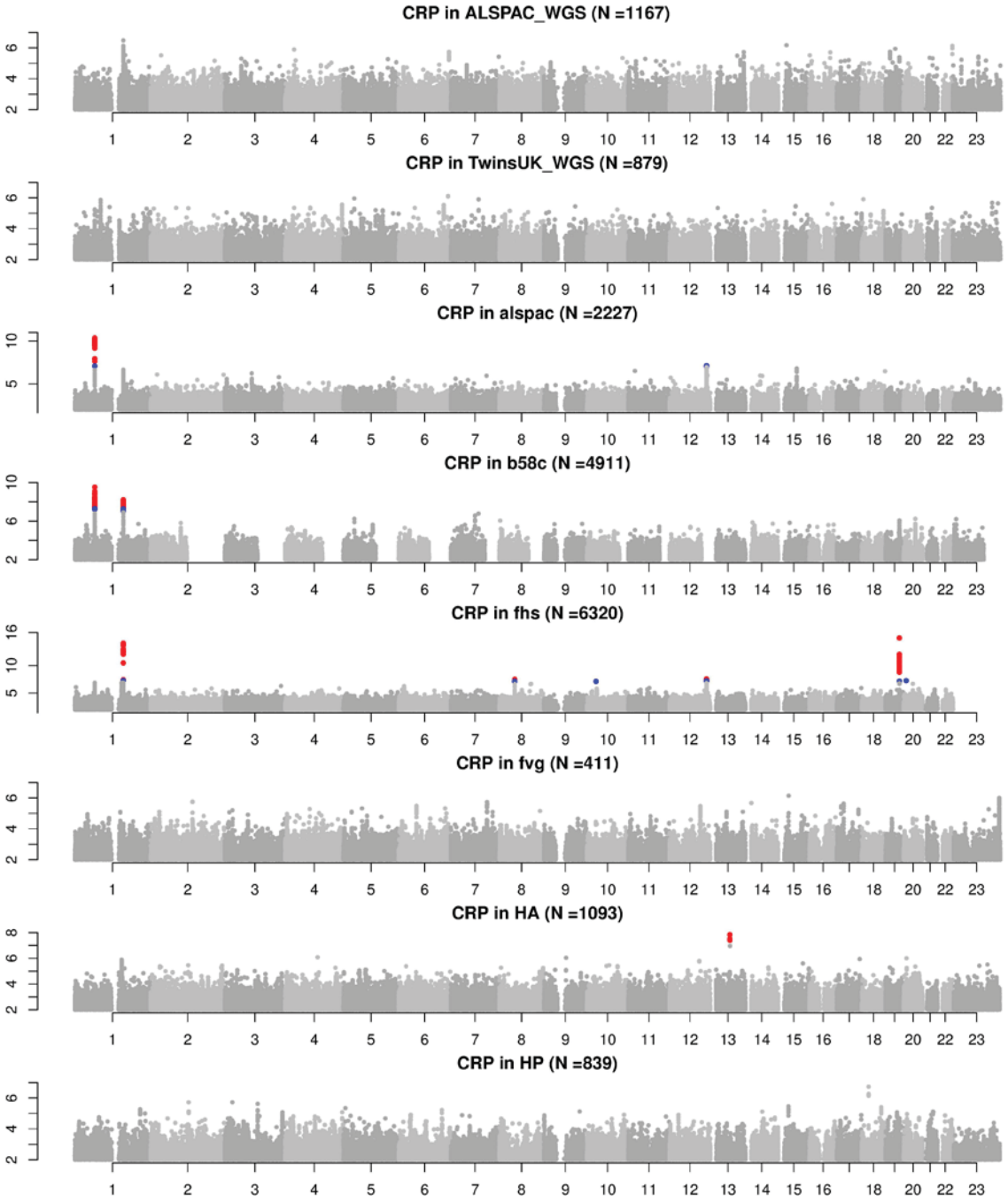

PCV in carl (N =484)
PCV in cbr (N =1034)
PCV in fvg (N =1392)
PCV in HA (N =993)
PCV in HP (N =962)
PCV in luric1 (N =1429)
PCV in luric2 (N =1634)
PCV in TwinsUKall (N =3605)
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 18 20 22 23

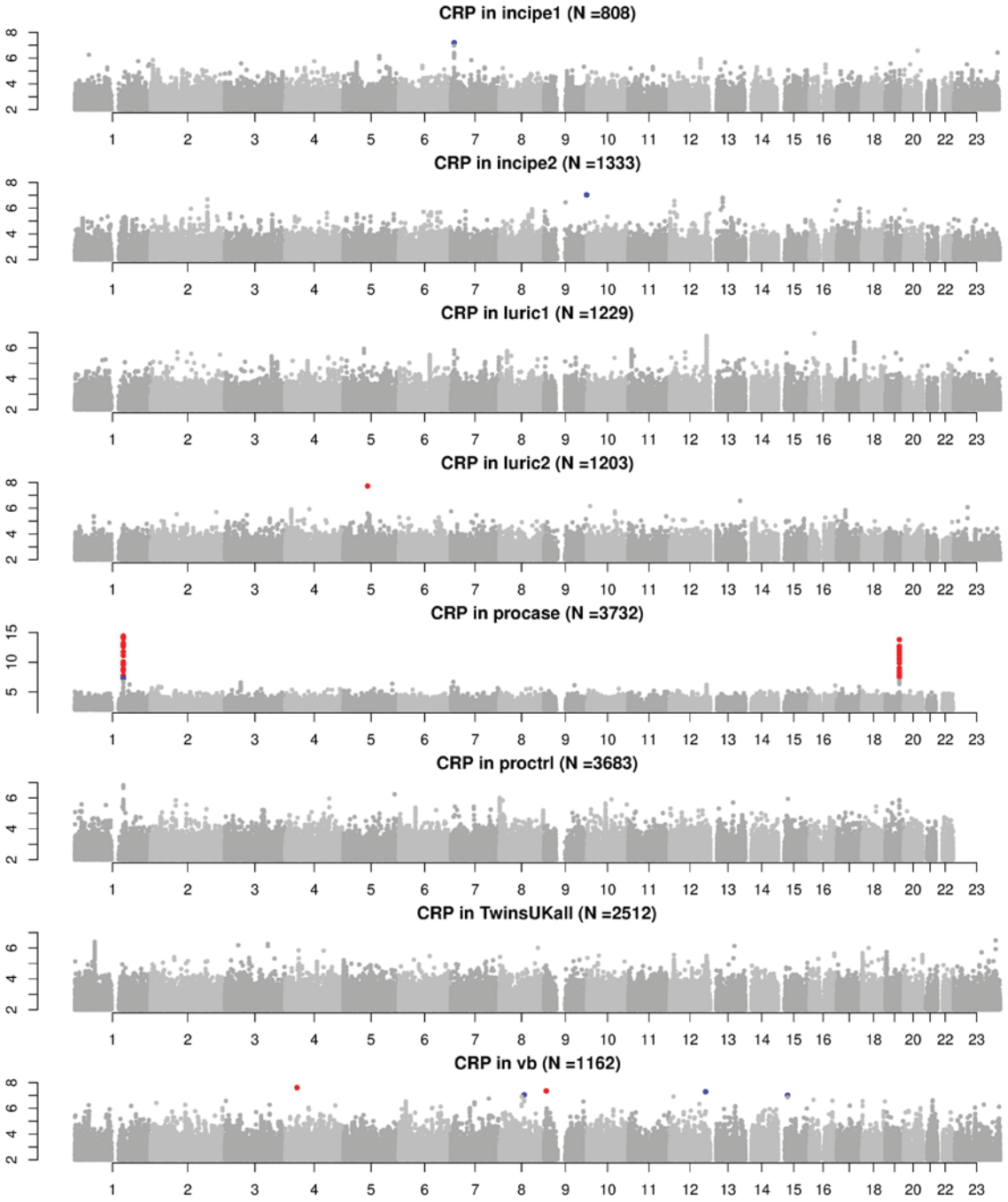

PCV in ukbs (N =2057)
PCV in vb (N =1766)
PCV in whi_garnet (N =4242)
PCV in whi_hipfx (N =3854)
PCV in whi_mopmap (N =3058)
PCV in whi_whims (N =5673)
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 18 20 22 23

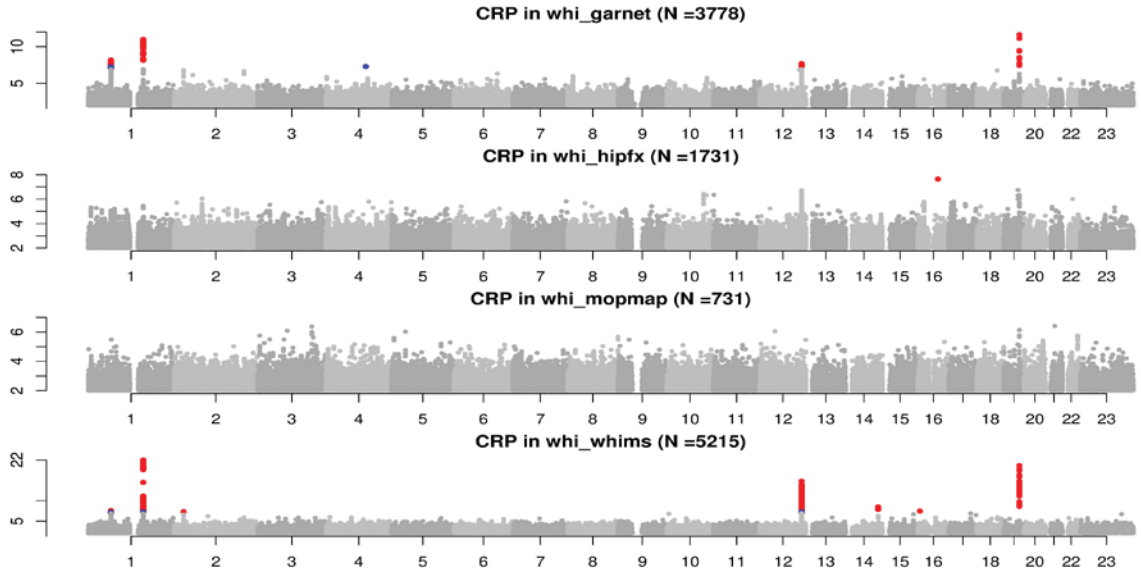

PLT in carl (N =481)
PLT in cbr (N =1034)
PLT in fvg (N =1375)
PLT in HA (N =991)
PLT in HP (N =968)
PLT in luric1 (N =1429)
PLT in luric2 (N =1634)
PLT in TwinsUKall (N =3602)
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 18 20 22 23

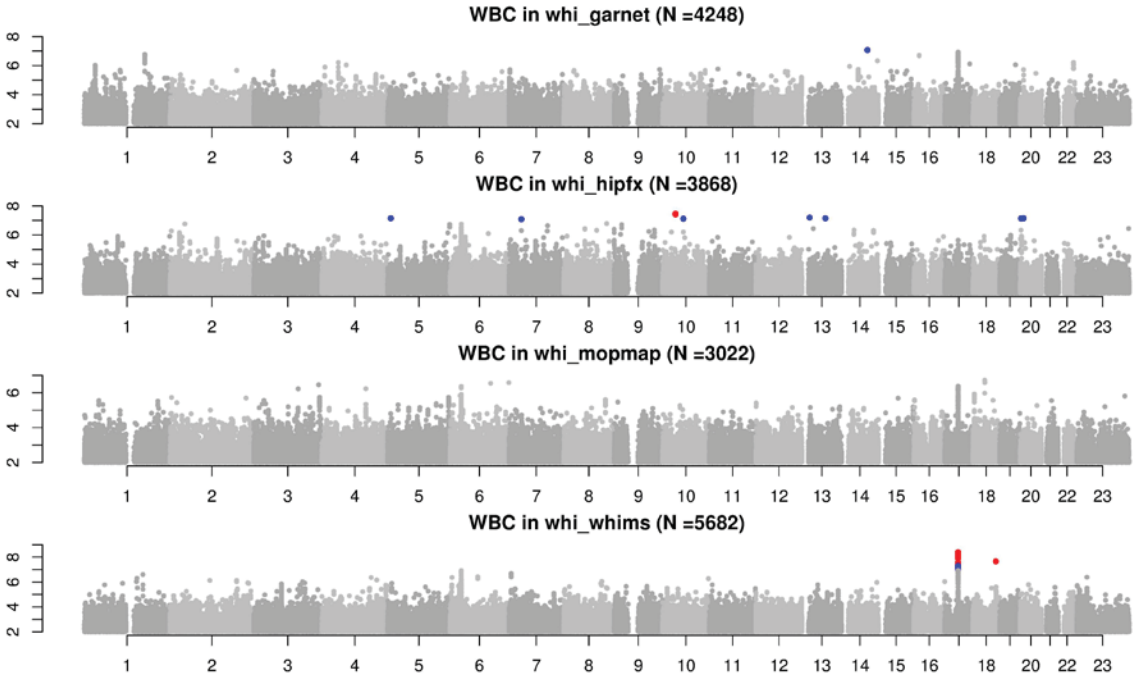

PLT in ukbs (N =2051)
PLT in vb (N =1767)
PLT in whi_garnet (N =4217)
PLT in whi_hipfx (N =3831)
PLT in whi_mopmap (N =3037)
PLT in whi_whims (N =5646)
2 4 6 8
5 10
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 18 20 22 23


WBC in carl (N =484)
WBC in cbr (N =1034)
WBC in fvg (N =1387)
WBC in HA (N =990)
WBC in HP (N =963)
WBC in TwinsUKall (N =3597)
WBC in ukbs (N =2054)
WBC in vb (N =1774)
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 18 20 22 23


CRP in ALSPAC_WGS (N =1167)
CRP in TwinsUK_WGS (N =879)
CRP in alspac (N =2227)
CRP in b58c (N =4911)
CRP in fhs (N =6320)
CRP in fvg (N =411)
CRP in HA (N =1093)
CRP in HP (N =839)
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 18 20 22 23


CRP in incipe1 (N =808)
CRP in incipe2 (N =1333)
CRP in luric1 (N =1229)
CRP in luric2 (N =1203)
CRP in procase (N =3732)
CRP in proctrl (N =3683)
CRP in TwinsUKall (N =2512)
CRP in vb (N =1162)
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 18 20 22 23


CRP in whi_garnet (N =3778)
CRP in whi_hipfx (N =1731)
CRP in whi_mopmap (N =731)
CRP in whi_whims (N =5215)
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 18 20 22 23


WBC in whi_garnet (N =4248)
WBC in whi_hipfx (N =3868)
WBC in whi_mopmap (N =3022)
WBC in whi_whims (N =5682)