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.

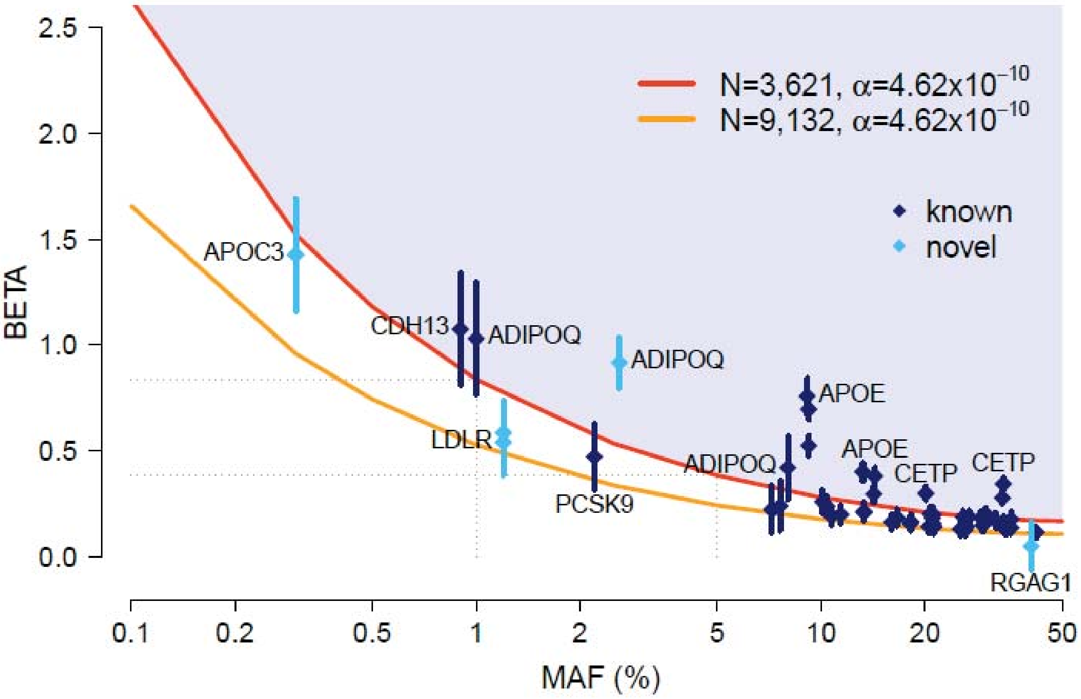
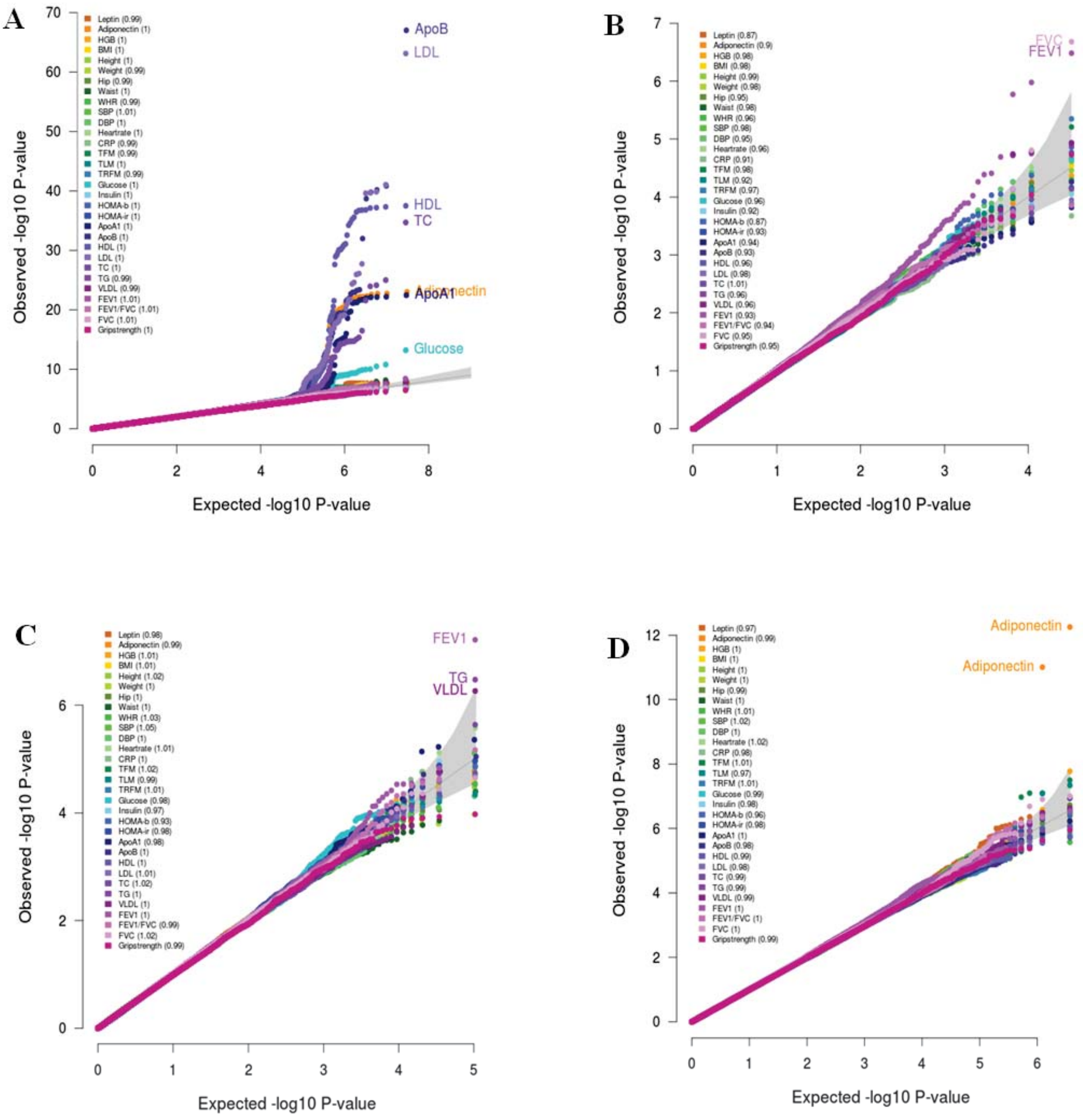


Figure 7.2 QQ plot of association tests for 31 UK10K core traits

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.