5 Conclusions and future directions

Genetic studies of complex traits have advanced our understanding of complex disease by revealing the polygenic architecture of most of these traits, uncovering biological mechanisms contributing to phenotypic variance, and in some cases highlighting novel potential therapeutic targets. Most of these advances have been through the exploration of common variation in the population through array-based genotyping. As the field has moved forward, there has been an increasing interest in understanding the contribution of rare variation to common genetic traits and diseases, facilitated by improved imputation reference panels [127, 152, 392], and decreasing costs of sequencing. Parallel to this, the range of studied phenotypes has continued to expand by including higher resolution measurements (high dimensional molecular phenotypes), focusing on extremes of the phenotype distribution, and measuring various correlated traits in the same individuals to gain novel insights into the pathophysiology of disease.

In this thesis, I have provided further knowledge on the genetic architecture of a distinct number of cardiometabolic traits (Chapters 2, 3 and 4) by combining a variety of approaches with diverse genotypic and phenotypic resolution. These ranged from analysis of rare coding variation (Chapter 3) to common variants (Chapters 2 and 4), as well as, different degrees of phenotypic resolution, including biomarkers of cardiovascular disease obtained from NMR measurements (Chapter 3), extremes of continuous phenotypes (BMI) clinically ascertained (Chapter 2), and exploration of a glycaemic biomarker hitherto little explored (Chapter 4).

I and others first explored the genetic architecture of persistently thin and healthy individuals using a clinically ascertained cohort: STILTS (**Chapter 2**). This allowed me to establish the heritability of healthy thinness for the first time and show that this estimate is similar to that of early onset severe obesity. I and others also performed a GWAS of persistent healthy thinness vs. severe obesity with a total sample size of 2,927. We were able to find evidence of association in loci that had only just been discovered at the time of this work, using large cohorts with >40,000 individuals highlighting the added value of a clinical extreme approach. Finally, results from this study also showed that thinness falls on the lower end of the polygenic BMI spectrum, although incomplete genetic correlation with BMI suggests it is plausible additional loci influencing thinness might be found by focusing on clinically ascertained persistent and healthy thinness, and further investigating the rarer allele frequency spectrum. The work from this chapter provides a valuable resource for future studies into body mass index, where further studies on similarly ascertained clinical extremes can be combined with these datasets to increase power to detect novel loci and/or investigate non-additive effects of established loci at the extremes of the distribution. Loci exerting their effect mostly through the lower tail of the BMI distribution might highlight protective variation aiding the search for anti-obesity therapeutic targets.

In the next two chapters I studied the genetics of circulating biomarkers in a population of healthy blood donors (INTERVAL). In **Chapter 3**, I studied the influence of rare variation on 226 serum lipoproteins, lipids and amino acids measured on a subset of this population with WES and/or WGS data (N_{total}=7,142). Gene-based analyses recapitulated established associations in lipoprotein metabolism genes (*APOB*, *APOC3*, *PCSK9*, *SCARB1* and *LIPC*) and amino acid metabolism genes (*HAL*, *PAH*, *ALDH1L1*) and highlighted four genes (*ACSL1*, *MYCN*, *FBXO36* and *B4GALNT3*) potentially involved in lipoprotein metabolism that merit further replication in additional studies using similar high resolution measurements.

Expanding the analysis to gene sets, I found a novel association of rare loss-of-function variants in the regulation of pyruvate dehydrogenase (PDH) complex pathway with intermediate and low density lipoprotein metabolism. Finally, focusing on genes near GWAS signals for traditionally measured lipid traits, after removing loci where the effector transcript is known, I found an enrichment of rare variant associations in genes near HDL-C GWAS signals in esterified and total cholesterol in extra-large HDL suggesting this gene set is enriched for effector transcripts. Exploring the tails of the distribution of these measurements, I also found an enrichment of predicted deleterious variants in lipoprotein disorder and metabolism gene sets at the lower tails of four lipoprotein measurements. This finding demonstrates that rare "protective" variation with strong effects is a significant contributor to lipoprotein levels in a healthy population. Overall, I showed that the increased genotypic resolution gained by using sequencing data allowed us to unveil the contribution of rare variation to the extremes of the distribution of circulating biomarkers, the identification of a novel pathway influencing these measurements, and to highlight the enrichment of effector transcripts near HDL GWAS signals, all findings which had not been addressed in previous work using array-based genotyping platforms on larger sample sizes on the same NMR platform (e.g Kettunen et al. (2016) [173] N=24,925).

In my last project, I performed the largest GWAS to date on fructosamine levels on 24,586 individuals from the INTERVAL cohort (**Chapter 4**). Here I characterised the heritability of the trait and found it to be very low (~2%), which is consistent with what would be expected from a trait measuring short term changes in glycaemia. In addition to this, I discovered one novel locus (*G6PC2*) associated with fructosamine that has been previously linked to other glycaemic traits [367], and another locus *(RCN3*) that had been previously linked to fructosamine through non-glycaemic pathways [363]. I also found some shared genetic aetiology between fructosamine and other glycaemic traits such as glycated haemoglobin, fasting glucose and fasting insulin (binomial p=5.6x10^{-3} for enrichment of nominally significant signals with consistent direction of effect) but no evidence of genome-wide genetic correlation (p>0.05 for all estimates). Fructosamine, as a glycaemic trait, has been understudied and only very recently the first genetic study was published [363]. Future work on this dataset will aim to provide more clarity into the genetic relationship of this trait with T2D, its comorbidities and other glycaemic traits.

Altogether, the different approaches used in this thesis shed light on specific components of the genetic architecture of the studied cardiometabolic traits. Varying levels of genotypic resolution allowed me to explore the impact of variation across the allele frequency spectrum to the genetic architecture of these traits. Contribution of common variation was assessed via genome-wide imputed data (**Chapters 2 and 4**) whereas contribution of rare variation was assessed via sequencing data (**Chapter 3**). I also tested various levels of phenotypic detail to capture different aspects of cardiometabolic trait biology (more on this on **Section 5.1**). The diverse study designs employed in this thesis showcase the utility of combining datasets with different degrees of genotypic and phenotypic resolution to gain novel biological insights.

5.1 Expanding the range of phenotypic measurements

Cardiovascular disease can be impacted by a wide diversity of risk factors. Understanding the genetic bases of each can help us better recognise the causality networks leading to disease and the heterogeneity in presentation of symptoms, comorbidities and outcomes. The choice of phenotype to focus on will lead to a different snapshot of these complex networks of interactions. In this thesis I have explored different resolutions of phenotypes from anthropometric measurements (extremes of BMI distribution), to measurement of a relatively unexplored glycaemic trait (fructosamine), to high resolution circulating biomarker measurements (NMR data). Each of these projects allowed me to understand different biological aspects of these traits tightly linked to cardiovascular disease.

As demonstrated in previous efforts [38, 173, 288] and this thesis, higher resolution measurements of many circulating lipid, lipoprotein and amino acids can provide novel metabolic insights as many of these measurements are better at capturing underlying biology. Having a single large cohort with these measurements provides a huge advantage in avoiding between-study heterogeneity not due to biological variables. In future, coupling high resolution measurements with sequence data and electronic health records (EHR) has the potential benefit of assessing *in-silico* effects of protein inactivation on circulating biomarker metabolism and unexpected (positive or negative) medical side-effects. This can be achieved by testing the effect of loss-of-function variants (mimicking drug targeting) on different circulating biomarkers and medical conditions through mediation analysis. Population cohorts such as the UK Biobank (and other large cohorts that may accrue relevant data) will provide a unique opportunity to explore these types of questions as they accrue sequencing data and high resolution NMR measurements [393, 394].

In parallel with the development of large national biobanks, studies of carefully selected clinical cases can add a powerful dimension to the study of the genetic architecture of common traits. In particular carefully ascertained individuals on the extremes of the phenotype distribution, especially as sample sizes increase and the genetic resolution increases to sequence based studies, may reveal additional rare variants of larger effect exerting effects on these traits and highlight possible new therapeutics. Studies in height and lipid traits have shown a higher polygenic component in the upper tail of the distribution and have suggested a role for rare variation in the lower tail [245, 327]. It is possible then, that WES data on the STILTS cohort might generate further insights into the genetic causes of persistent and healthy thinness.

5.2 Assessing pleiotropy in complex disease

Deep phenotyping (i.e, the simultaneous measurement of multiple detailed phenotypes) also allows exploration of biological questions involving multiple correlated traits. The correlation structure of phenotypes can aid genetic studies in two ways: increase power to detect associations by capturing noise due to environmental variation and identification of shared genetic effects between traits (pleiotropy). The former was discussed in **Chapter 3** and the latter is a feature of complex traits whose better understanding is key for the future of precision medicine.

Pleiotropy occurs when a single gene affects more than one trait simultaneously. One way to assess pleiotropy is by testing a single variant against a wide number of phenotypes simultaneously in a phenome-wide association study (PheWAS) [144]. Another way to test for pleiotropy that does not pinpoint the associated loci but gives an overall sense of genetic relationship between two traits is through genome-wide genetic correlation analyses [228, 395]. Through these approaches, it has been shown that pleiotropic effects in the human phenome are pervasive.

Studies of pleiotropy can reveal unknown molecular links between seemingly unrelated phenotypes such as multiple sclerosis and schizophrenia [396] or childhood obesity and ulcerative colitis [228]. Given that in complex disease, a risk factor can be regulated by several different genetic variants representing different pathways, understanding how these variants impact disease risk could potentially add a new dimension to patient risk stratification beyond the sole measurement of the risk factor. For lipid and glycaemic traits in particular, there has been an increasing amount of evidence showing how cardiovascular disease and T2D risk changes depending on the pathway through which risk factors are increased or decreased, for example, only some HDL-C raising genetic mechanisms have an effect on CVD risk [110](see **Chapter 1 Section 1.2.2**). My findings in **Chapter 3** were consistent with what has been previously reported in literature [38, 311] of pleiotropic effects of genes such as *APOB*, *APOC3* and *PCSK9* that have been previously associated with traditionally measured lipid traits on multiple detailed measurements of lipoprotein metabolism. In **Chapter 4**, I show that similarly to what has been previously shown for HbA1c [121, 350], fructosamine levels can be increased via glycaemic or non-glycaemic pathways.

Further pleiotropic studies on CVD risk factors are warranted to get a clearer picture on the influence of these traits on cardiovascular disease and T2D risk and potentially identify optimal drug targets (e.g targets without a detrimental impact on another trait).

5.3 Exploring the contribution of rare variation to cardiometabolic traits

Rare variant analyses are currently underpowered to detect associations at gene-wide significance ($2.5x10^{-6}$) with sample sizes similar to the ones in many current studies (~10,000 samples), especially in case-control studies [397]. It is therefore not surprising that gene-based tests in **Chapter 3** did not yield novel associations that remained significant after correcting for multiple traits. As mentioned in the discussion of the aforementioned chapter (see **Chapter 3 Discussion 3.5**), pathogenicity scores are an important tool to help prioritise variants but still, these are not perfect. Integration of information from human interactome networks and techniques such as deep mutational scanning in the future, will potentially lead to improvement in prediction of deleteriousness of protein coding variants [398, 399]. In the end, the balance between stringency of filters used in variant selection for the analysis and the number of variants included in it determines the outcome of the test. Since this information is usually not known *a priori*, it is not uncommon to use various sets of filters in gene-based tests to maximise power [91, 288, 400]. Since high confidence loss-of-function variants are rare, an approach that has been used before with success is testing gene sets instead of individual genes [401]. This approach was also successful in my own data. The downside to this approach is that it is harder to pinpoint causal genes.

As whole-genome sequencing becomes more prevalent, it will become an even bigger challenge to develop scores to prioritise variants to be included in rare variant aggregation tests as consequences of non-coding variation are less well understood than those in coding variation where one can more easily interpret the impact on the affected protein sequence. Attempts at scoring non-coding variants have been shown to fail to differentiate neutral variation from highly deleterious variation [402]. Generation of epigenomic maps for distinct cell types such as the ENCODE [403], ROADMAP EPIGENOME [404] and BLUEPRINT projects [405] will provide additional data to functionally categorise non-coding variation and refine these functional scoring algorithms that mostly rely on machine-learning approaches. Previous efforts to improve annotation of non-coding variants also include usage of expression data from the GTEx consortium to generate an algorithm that predicts regulatory effects of rare variants [406]. Another technique that should allow for improvements in identification of regulatory elements is massively parallel reporter assays [407]. These assays allow testing for activity of thousands of regulatory elements in a single experiment making it ideal for this endeavour.

On-going improvement of pathogenicity scores for coding and non-coding variation will not only aid in the discovery of novel gene-trait associations but will also be crucial when incorporating sequencing data from patients in the clinic by differentiating likely causal mutations for a given phenotype from neutral variation, therefore influencing provision of diagnosis and in time influencing treatment choice.

5.4 Concluding remarks

The field of complex disease genetics has been undergoing a major transformation with increasing sample sizes, establishment of large deeply phenotyped cohorts and decreasing costs of sequencing. GWAS studies have helped us get a better understanding of complex disease but there are still many gaps in the knowledge of the biological underpinnings of a wide number of traits. During my PhD I have addressed some of these gaps by focusing on understudied phenotypes, in particular, risk factors for T2D and cardiovascular disease and using a combination of imputed and sequencing data to study them. I provided the first evaluation of the genetic architecture of persistent and healthy thinness, insights into the contribution of rare variants to circulating biomarkers levels and novel findings regarding the genetic architecture of fructosamine regulation. Nevertheless, many questions still remain that can only be addressed by increasing sample sizes (preferably with sequencing data), expanding studies to include more samples of non-European origin, exploring other forms of genetic variation that are currently understudied (e.g. structural variation), expanding the number of phenotypes tested and functional follow-up of associated loci. Some of the outstanding questions in the field include but are not limited to:

- How many independent loci influence these risk factors?
- What are the causal variants in associated loci?
- What is the contribution of structural variation to trait heritability?
- What proportion of these loci are shared between risk factors?
- Can we identify protective rare variation in genes not highlighted in association studies that only occurs in the tails of the phenotype distribution?
- Which genes represent ideal drug targets?
- What is the biological consequence of associated non-coding loci?
- How do genetic variants associated with disease or trait mechanistically impact pathophysiology/ physiology?

Answering these questions is necessary if one aims to be able to use genetic data in standard clinical practice. Precision medicine will rely on these on-going advancements in the field to improve quality of patient care.