1 Chapter 1: Introduction

1.1 Complex traits

Complex diseases and traits are phenotypes that, in contrast to simple Mendelian disorders, are not explained by the action of one single gene within any given person or family. Instead, complex diseases and traits arise from the action of independent genetic factors, environmental factors and gene-by-environment (GxE) interactions. The independent genetic factors often provide small contributions to the overall risk of a disease or to the variability of a continuous trait [1].

Height and weight are two examples of human complex traits. Early studies looking at family resemblance suggest that these two traits have a strong genetic component and that there is no single major locus influencing these traits [2-4]. Welfare components such as nutritional quality and health also have a high impact on these traits [5, 6]. As such, individuals could have a strong genetic background of trait increasing alleles but never realize their genetic "potential" if not placed in a permissive environment. This is a key difference with traditional Mendelian disorders where a single mutation within a given family is considered necessary and/or sufficient to cause the phenotype.

1.1.1 Cardiometabolic traits and impact on human health

Cardiovascular diseases (CVDs) are a group of mostly complex diseases that affect the heart and blood vessels including: coronary heart disease (CHD), cerebrovascular disease,

peripheral arterial disease, rheumatic heart disease, congenital heart disease, deep vein thrombosis and pulmonary embolism [7]. CVDs account for most deaths globally [7] and it is estimated that 90% of these diseases are preventable [8].

In recent years, CVDs have been increasing in prevalence in developing countries [9-11] which makes them a continuing global public health priority in the years to come. Risk factors for cardiovascular disease include: family history [12], age [13], sex [13], tobacco use [14], physical inactivity [14], diet (e.g high trans-fat intake [15], high salt intake [16]), heavy alcohol consumption [17], high blood pressure [18], diabetes [18] , obesity [19] and excess circulating lipids (hyperlipidaemia) [20].

Many of these risk factors are not completely independent of each other. Obesity, defined as a body mass index (BMI) greater than 30Kg/m^2, often co-occurs with type 2 diabetes (T2D) and/or hyperlipidaemia and confers a ~3 fold increase in risk for coronary heart disease in men younger than 65 even after adjusting for other risk factors [21]. The increased risk is also observed in women but with a smaller relative risk [22]. Besides CVD, obesity is a risk factor for other medical conditions such as hypertension, osteoarthritis and certain cancers [23]. Furthermore, obesity has an overall adverse impact in quality of life as on top of some secondary physical factors arising from obesity, there is a social stigmatization of the condition that can result in discriminatory behaviours towards obese individuals [24]. More details about obesity are described in **Chapter 2**.

Diabetes is a group of disorders characterised by excess levels of sugar in a person's blood over a long period of time. Over 90% of the cases of diabetes are T2D cases [25]. T2D arises as a result of insufficient insulin production from pancreatic beta cells when an individual develops insulin resistance, a condition characterised by the cells' inability to respond

properly to insulin. Obesity is considered one of the most important factors leading to T2D as it is tightly linked to development of insulin resistance [26]. Given diabetes is a lifelong condition, chronic mismanagement of the condition leads to early mortality, and particularly, cardiovascular death. This risk is exacerbated by medical complications linked to the condition such as renal complications [27]. More details about diabetes are described in **Chapter 4**.

Hyperlipidaemia encompasses conditions such as hypercholesterolaemia (excess levels of cholesterol) and hypertriglyceridaemia (excess levels of triglycerides). Cumulative exposure to hyperlipidaemia in young adulthood is associated with an increased risk of CHD in a dose-dependent fashion after adjusting for other cardiac risk factors [20]. Hyperlipidaemia can be divided into primary or secondary. Primary hyperlipidaemias are also called familial hyperlipidaemias and are characterised by genetic alterations leading to abnormally high levels of lipids [28]. Secondary hyperlipidaemias, also known as acquired hyperlipidaemias, arise from underlying disorders leading to alterations in lipid levels. T2D is one of the most common causes of acquired hyperlipidaemias [29]. More details about circulating lipids are described in **Chapter 3**.

1.1.2 Heritability

Heritability is defined as the proportion of variance of a trait that can be explained due to genetic factors. This measurement captures the resemblance between parent and offspring. So traits with high heritability have high resemblance between parents and offspring whereas traits with a low heritability have low resemblance [30]. Heritability can be divided into broad sense heritability and narrow sense heritability. Broad sense heritability (H^2)

reflects all genetic contributions to a phenotype including additive (average effects of alleles at a locus), dominant (interaction between alternative alleles at a single locus) and epistatic effects (interactions between different loci) and it is defined as H^2=Var(*G*)/Var(*P*), where Var(*G*) is the variance of genotypic effects and Var(*P*) is the variance of the phenotype . Most of the genetic variance in populations is thought to be driven by additive effects [31]. Therefore another widely used estimate of heritability is that of narrow sense heritability (h^2) which is defined as h^2=Var(*A*)/Var(*P*) where Var(*A*) is the additive variance component of the genetic variance.

To estimate heritability, studies in human populations have mostly focused on related individuals. Traditionally studies calculated heritability looking at correlations amongst family members (e.g parent-offspring, full siblings, twins) [30] or adoption studies [32]. Amongst these studies, the most common study design is a twin study design that looks at phenotypic correlation between monozygotic (MZ) twin pairs and dizygotic (DZ) twin pairs [33]. The rationale behind these studies is that differences in trait correlation between monozygotic twin pairs compared to dizygotic twin pairs are driven primarily via genetic effects since twins tend to share the same environments. These studies are also particularly helpful to disentangle shared and unique environmental effects. Shared environmental effects can be extracted by subtracting the heritability estimate contribution from the observed twin phenotypic correlation (r_{MZ}-h^2 in MZ twins where r_{MZ} is phenotypic correlation in MZ twins and r_{DZ}-(h^2/2) in DZ twins where r_{DZ} is phenotypic correlation in DZ twins), i.e the percentage of the observed correlation that is not explained by genetic effects. Unique environmental effects are obtained by quantifying the observed difference

in MZ twins ($1-r_{MZ}$), i.e, the degree to which the observed correlation in MZ twins differs from 1.

One important feature about heritability is that it is not constant in time or space. The heritability of foetal length, for example, increases during later developmental stages [34]. Changes in environmental factors within a population can also affect heritability estimates as in the case of intelligence measurements [35]. Changes in allele frequency during selection or introduction of new alleles in a population via migration can also alter a trait's heritability in a given population.

Heritability is an important parameter as the power of most studies to discover loci associated with a trait is positively correlated with the heritability of the trait [36]. For Mendelian disorders, heritability is straightforward as the disorder only manifests itself if you have alterations in one gene (or in very few cases a small number) and discovery of this gene, or genes, can be assessed in families with affected individuals by observing the patterns of co- inheritance of the disease and genetic markers (described in more detail in **Section 1.1.3**). For complex traits, heritability estimates can be taken into account when selecting a population in which to study the genetic basis of a particular trait. For example, BMI is a trait where heritability is higher during childhood [37] so if one wants to boost power for locus discovery, one might opt to choose a population where environment has a lesser impact on the variance of the trait. With the development of improved technologies for human molecular phenotyping at scale, population studies of traits such as high resolution measurements of circulating lipid and lipoprotein subclasses have become feasible in genetic studies. As the overall heritability of these traits is higher compared to traditionally measured lipid traits in the clinic (e.g. large-density lipoprotein (LDL)

cholesterol or triglycerides (TG)) they can be used for lipid metabolism locus discovery with smaller sample sizes and to shed light on more detailed biological aspects of lipid metabolism [38] (more details in **Chapter 3**).

More recently, with the advent of genome-wide array technologies (described in more detail in **Section 1.2**), new methods have been developed to estimate heritability using genome-wide genotype data [39-43]. These are routinely used to both estimate the heritability of traits, and the proportion of this heritability that can be explained by mostly common genetic variants. These methods will not be discussed in further detail in this thesis.

1.1.3 Genetic studies of complex traits

Genetic studies of Mendelian disorders used linkage and candidate gene approaches to find the underlying genes with mutations causal of the disease in question. Linkage of two loci occurs when these are transmitted together from parent to offspring more often than expected by chance under random assortment. A collection of loci along a chromosome region that are often inherited together is called a haplotype. Using linkage information one can identify genetic markers that co-occur with a disease in family pedigrees. After identifying co-inherited genetic markers, one uses this information to narrow down the region where the causal gene likely lies by finding the smallest haplotype that is co-inherited in affected individuals (**Figure 1.1**). Before high-throughput sequencing approaches were possible, once this interval was identified, selection of plausible candidate genes within the region was done based on biological knowledge. Candidate genes were then sequenced in patients to find the mutations associated with the trait. One of the first success stories for

linkage studies was the identification of the cystic fibrosis gene [44, 45] where a three base pair deletion accounts for 70% of all cystic fibrosis cases observed. Other genes successfully identified via linkage analysis were the Duchenne muscular dystrophy (DMD) gene [46], the Fanconi's anaemia gene [47] and the Huntington disease gene [48, 49].

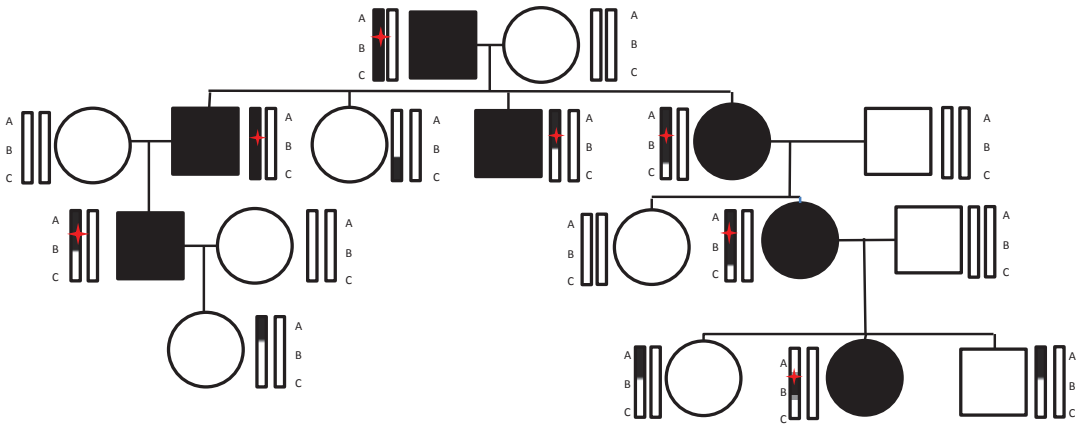

Figure 1.1: Principles of linkage analysis. A family pedigree is shown from a typical linkage analysis study for a Mendelian dominant disorder. Square (males) and circles (females) in black indicate affected individuals whereas symbols with no fill indicate unaffected individuals. Rectangles next to the symbols represent a fraction of a chromosome with the haplotype containing the associated gene where black filled sections represent the same specific alleles at marker polymorphisms. Letters A, B and C represent genetic markers and the red star is the unknown causal mutation.

Applying the principles of linkage analysis to complex traits has been a more difficult task and has led to many false positive results [50, 51]. As mentioned previously, complex traits are often the result of the action of many independent genes, each one contributing to a small degree to disease development/trait variability [1]. Other factors that made linkage studies for complex disease and traits difficult were the variable degree of expressivity,

incomplete penetrance and variable age of onset affecting a trait/disease, making it hard to properly define phenotype or choose the right population to study [52]. When applying linkage analysis to complex phenotypes, these factors combined result in linked regions with very wide 95% confidence intervals (CI) making the prioritisation of genes extremely difficult as intervals could encompass hundreds of genes. Sample sizes required to reduce the standard error in the positional estimate were prohibitively large (>1,600 families) and denser marker maps could only provide marginal benefits towards identifying plausible causal genes. This is important since most linkage studies at the time (1990-2000) were done using very small sample sizes [53]. Significance thresholds were also very lenient at the time which contributed to the generation of false positive results [54]. When using more stringent significance threshold, it was found that 66.3% of the linkage studies on complex traits as of December 2000 showed no significant linkage [55]. For these reasons, genetic association studies were proposed as a better suited technique to analyse complex traits [56].

1.2 GWAS of complex traits

Genome-wide association studies (GWAS) have been crucial to our understanding of complex traits. The shift from family studies to population based studies was in great part motivated by the common disease/common variant (CD/CV) hypothesis that states that common disease in the population is mostly influenced by common genetic variation in the population [57]. Given that allele frequency of disease associated alleles and prevalence of disease are strongly correlated, the CD/CV hypothesis would suggest that most of the common variation associated with disease would have low penetrance. To find these common variants with low penetrance one would need to test a wide number of variants

across the human genome. To this end, GWAS makes use of linkage disequilibrium (LD). The phenomenon of LD occurs when in a population, alleles at a number of loci co-occur more than expected by chance. The human genome can then be divided into blocks of haplotypes with differing degrees of LD [58, 59]. Population phenomena such as migration, bottlenecks and genetic drift can alter the patterns of LD in the genome and as such, one expects differences in LD block size across different populations. African populations for example, tend to have smaller LD blocks than European ones mainly due to the more recent arrival of humans in Europe allowing less time for recombination events to take place [60]. Therefore, instead of attempting to test all variation across the genome, one could just test polymorphic sites in a population that capture the majority of variation within an LD block. The most common polymorphism in the genome are single nucleotide polymorphisms (SNPs), and these became the preferred variant to test in genetic studies as they could be accurately genotyped with ease. SNPs that capture variation within an LD block are called tagging SNPs or tag SNPs, as they "tag" or capture information on that particular LD block. In GWAS, testing the causal allele for a phenotype is very unlikely and therefore testing for polymorphisms in LD with the causal allele can lead to identification of genomic regions associated with a trait (**Figure 1.2**) [61]. In a case-control study, a GWAS tests if an allele is observed more than expected by chance in individuals with a disease compared to a set of controls. For a quantitative trait, in the most basic scenario, a GWAS tests if the presence of a certain allele is a statistically significant predictor of the outcome variable (i.e. the quantitative trait) under a linear regression.

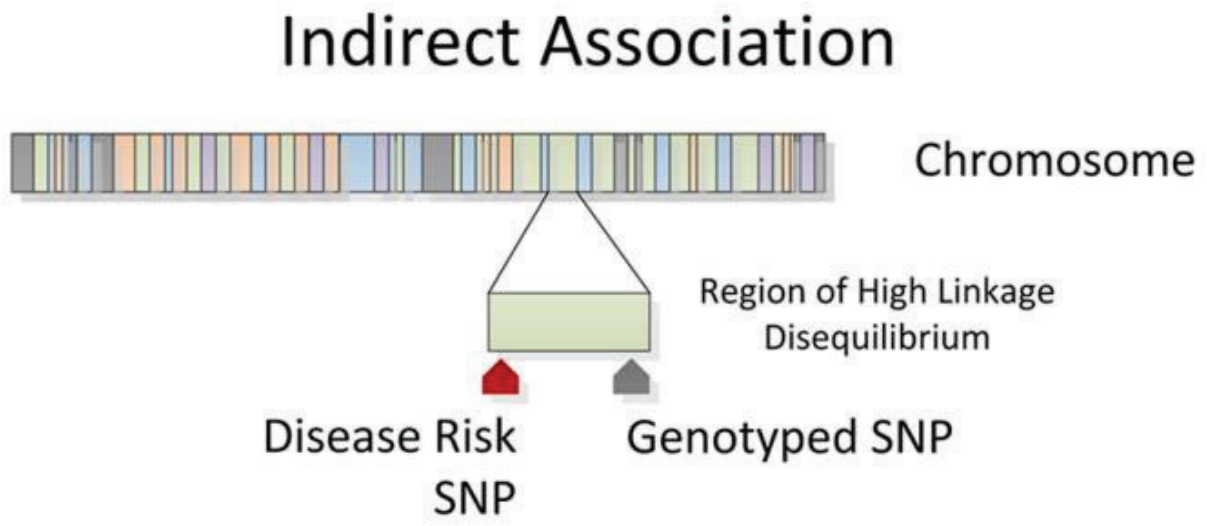

Figure 1.2: Indirect association. In a GWAS more often than not, the tested allele is not the causal allele. GWAS takes advantage of LD to identify regions of the genome associated with a phenotype by using SNPs in high LD with the causal allele. Figure extracted from Bush W.S and Moore J.H (2012) [62].

The International HapMap project was a major milestone for association studies as it provided the first comprehensive collection of SNPs covering the human genome [63]. By capturing variation at millions of sites within the human genome, the HapMap project allowed the examination of the correlation of SNPs in different populations and the identification of tag SNPs. One important insight gained from the HapMap project is that in European and Asian populations, one can capture >80% of common variation (MAF >= 0.05) across the genome using only a subset of SNPs between 500,000 and 1,000,000 [64]. Before the HapMap project, technologies to simultaneously assay a few thousand SNPs in the genome had already started being developed [65]. The first decade after the development of the first genotyping array saw an increase in number of sites tested ranging from a few thousand in the first array to more than a million in the latest versions in great part thanks to the HapMap project [66] and later projects such as the 1000 genomes project (see **Sections 1.2.1.1**).

It was soon after the development of genotyping arrays querying hundreds of thousands of sites that the first GWAS was published in 2005 [67]. This GWAS was a case-control study looking at age-related macular degeneration (ARMD) and found two SNPs that were significantly associated with the condition. Two years after, the Wellcome Trust Case Control Consortium (WTCCC) demonstrated that one can use shared controls in GWAS to find associations at multiple common diseases [68].

1.2.1 Meta-analysis

Similar to linkage analysis, one of the key limiting factors to detect signals in association studies is sample size [69]. Combining different studies for a trait under a meta-analysis framework provides multiple advantages for association studies. Firstly, combining studies increases sample size, therefore increasing power to detect association, especially at variants on the lower frequency allelic spectrum (minor allele frequency (MAF) 1-5%) which normally can only be detected if there is a large effect size which is rare in polygenic conditions. Secondly, it helps reduce false positives by testing for evidence of association at the same locus in multiple independent datasets. One major development that made meta-analysis of several different studies possible was genotype imputation.

One of the drawbacks of meta-analysis is that between-study heterogeneity can arise due to study specific factors such as different LD structure in populations, different environmental exposures or phenotype classification. Identifying sources of heterogeneity though, can reveal some interesting biological features underlying the association results [70].

1.2.1.1 Imputation

Imputation consists of mathematically inferring the most likely genotype at a given position given information of SNPs surrounding the position (**Figure 1.3**) [71]. LD information from populations of interest is used to maximise accuracy of these predictions. This technique allows comparison of genotypes at the same position in two studies that might have used different genotyping arrays and therefore might not have typed exactly the same variants. Imputation normally requires a "reference panel" which is a set of SNPs for which we know LD information in a given population. Besides the HapMap project, another initiative that provided a key boost to the field was the 1000 genomes project (1000G) [72]. The goal of this project was to sequence the genome of ~1000 individuals from diverse ethnic backgrounds using sequencing technologies that were developed during the time of the study. When used as a reference panel for imputation, 1000G project provides haplotype information for several million of variants across the human genome.

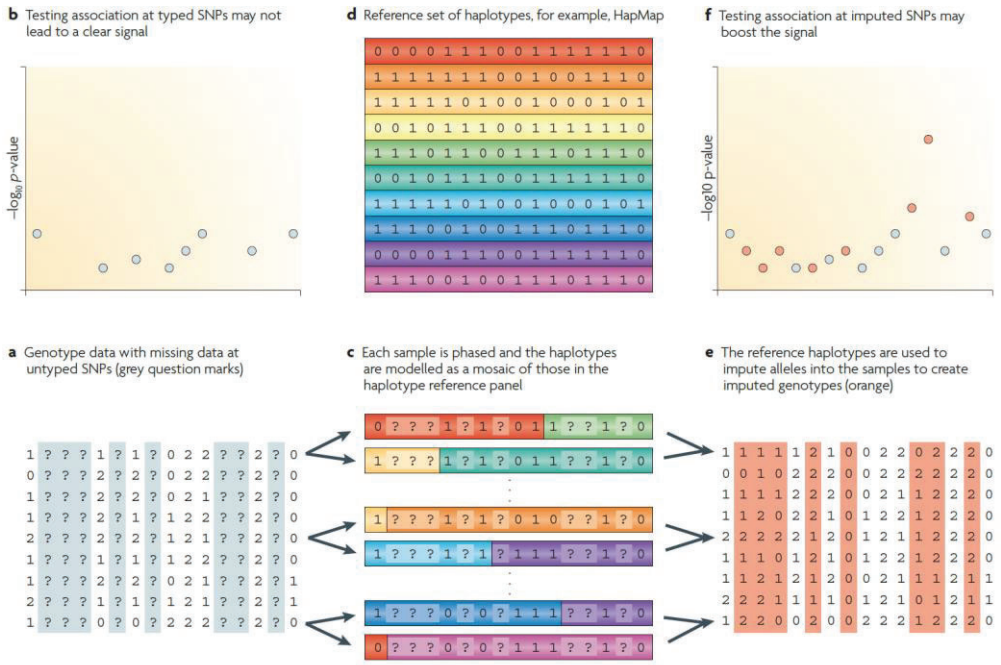

Figure 1.3: Genotype imputation process. A) Genotype data from individuals is collected with missingness at certain sites. B) Testing association only at directly genotyped sites may not lead to a significant signal. C) Samples are phased and haplotypes are modelled as mosaics of the haplotypes present in a reference panel. D) A reference panel is used to impute missing variants. E) After imputation, sites with missingness for which the reference panel has information are mathematically inferred. F) Testing association on the imputed dataset might boost signal. Figure extracted from Marchini J and Howie B (2010) [73].

Advances in imputation technologies facilitated the collaboration amongst many research groups to study complex traits and led to the creation of several consortia to perform large scale GWAS. Examples of these consortia focused on cardiometabolic traits are presented in **Table 1.1**.

Consortium	Traits of interest	First publication
GIANT	anthropometric traits (e.g height, BMI)	Willer et al (2009) [74]
DIAGRAM	type 2 diabetes	Zeggini et al (2008) [75]
MAGIC	glycaemic traits (e.g fasting glucose, fasting insulin, two hour glucose, glycated haemoglobin (HbA1c), amongst others)	Prokopenko et al (2009) [76]
GLGC	lipid traits (e.g HDL cholesterol, LDL cholesterol)	Willer et al (2008) [77]
CARDIoGRAMplusC4D	coronary artery disease and myocardial infarction	CARDIoGRAMplusC4D (2013) [78]

Table 1.1:Examples of large cardiometablic GWAS consortia.

1.2.2 Insights gained from GWAS of complex traits

In the past 13 years since the publication of the first GWAS, this study design has become the standard in the field of human genetics to study complex traits. The CD/CV hypothesis received early support from GWAS with most trait-associated loci being indexed by common variants (median allele frequency of 40%) with small to modest effect sizes (median odds ratio (OR)=1.19) [79]. Furthermore most associations found as of July 2018, have been associations in non-coding regions (~94.7%) [79].

For traits like height and BMI, there are now >3000 and >900 established loci respectively [80]. These loci explain ~24.6% of the variance in height [80] and~6% of the variance in BMI [80] which leaves much room to identify additional loci in the future explaining some of the remaining heritability. However, heritability estimates using genome-wide imputed data suggest that much of the remaining heritability for both traits can be explained by common variation with smaller effects than those discovered so far and therefore the rest of the associated loci will be uncovered by just increasing sample size [41, 81]. This also appears to be the case for T2D where large-scale sequencing studies support the hypothesis that most of the genetic predisposition to T2D arises from common variation [82]. For other glycaemic traits, association studies have highlighted potential differences in genetic architecture for these traits. Beta cell function by homeostasis model assessment (HOMA-B) and insulin resistance by homeostasis model assessment (HOMA-IR), for example, are two traits with similar heritability estimates (26% and 27% respectively) and despite only slight differences in sample sizes (N_{HOMAB}=36,466, N_{HOMAIR}=37,037), GWAS found more significant associations with HOMA-B (>12 associations) than for HOMA-IR (two associations) suggesting differences in effect sizes, allele frequency of variants, number of loci or

environmental modification between these traits [83]. For lipid traits, more than 250 loci have already been identified associated with high-density lipoprotein cholesterol (HDL-C), low-density lipoprotein cholesterol (LDL-C), total cholesterol (TC) and/or triglycerides (TG) [84]. The genetic architecture of some of these traits like TG features a complementary role of common variation with small effects and rare variation with large effects affecting the trait as evidenced by the enrichment of rare variation (MAF<1%) found in known GWAS genes associated with elevated levels of TG [85].

Overlap of genes found in linkage studies of Mendelian forms of disease and GWAS performed on related cardiometabolic traits has been commonly observed in the field suggesting that many genes responsible for severe phenotypes also play an important role in complex traits [86-88]. For example, in studies of T2D, rare variation influencing disease risk, appears to be enriched in genes implicated in Mendelian forms of diabetes or altered glucose metabolism [82] providing evidence for genetic overlap between the more common and rarer forms of disease. Similarly to T2D, GWAS for lipid traits have found associations with common variants near genes involved in Mendelian forms of dyslipidemia such as *APOB, LDLR, APOE, PCSK9* , *CETP, LIPC* and *LCAT* amongst others[89].

Furthermore, evidence for low-frequency variants with effects larger than those found in common variants but lower than those found in Mendelian disorders (so called “Goldilocks” alleles)[90] so far have not been found for most complex traits except lipid traits [91] (**Figure 1.4**).

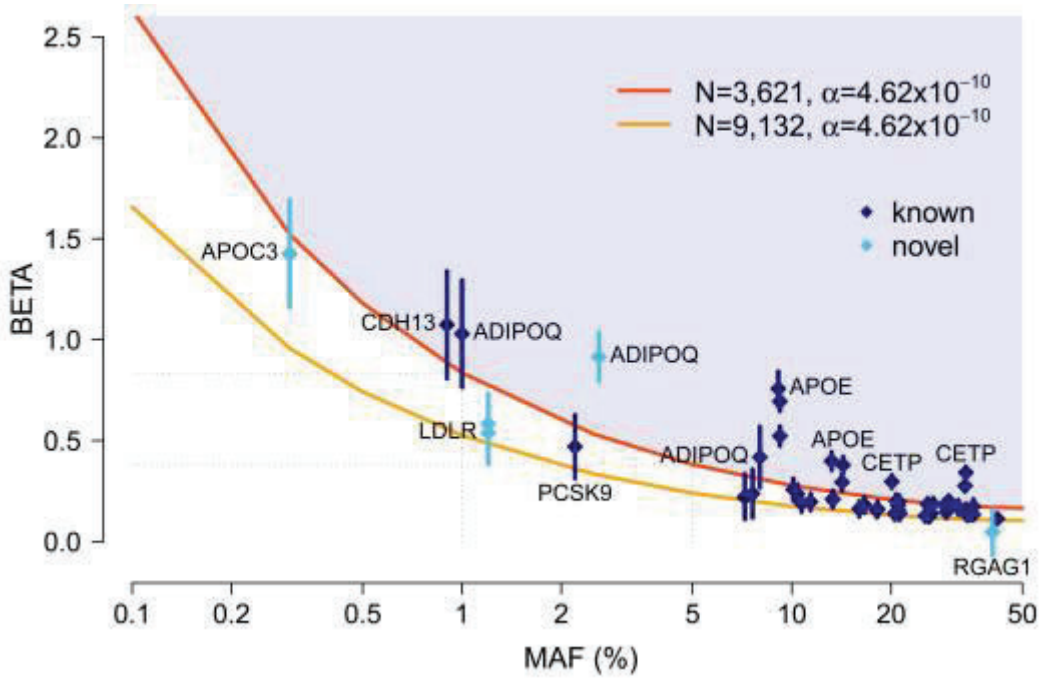

Figure 1.4: Results from single point association analysis in UK10K for 31 core traits shared between TwinsUK and ASLPAC cohorts. Minor allele frequency of variants is plotted on the X axis and effect sizes are plotted on the Y axis. Known associations are coloured in dark blue whereas novel associations are coloured in light blue with error bars being proportional to the standard error of the beta. Red and orange lines indicate 80% power at experiment-wide significance level ($p < 4.62 \times 10^{-10}$) for the maximum theoretical sample size for the WGS sample and WGS+GWA, respectively. The notable absence of loci in the middle part of the figure suggests "Goldilocks" alleles are a rare occurrence. Figure extracted from UK10K Consortium (2015) [91].

Results from GWAS have also led to novel insights into the biological pathways involved in the development of complex diseases. For genes near BMI associated loci, an enrichment in pathways related to synaptic plasticity and glutamate receptor activity has been observed which has highlighted the key role of central appetite control in the aetiology of common obesity [92]. Analysis focusing on low-frequency and rare variants have also implicated pathways related to insulin action and adipocyte/lipid metabolism [93]. For related measures of adiposity such as waist-to-hip ratio (WHR), there has been evidence of significant sexual dimorphism and an enrichment of genes expressed in adipose tissue depots [94]. Results from GWAS show that, as expected, T2D can arise due to alterations in

pathways affecting pancreatic beta cell formation and function or via pathways involved with regulation of fasting glucose as well as obesity [95, 96]. Some associations have also highlighted the role of genes involved in circadian rhythm pathways in glucose metabolism and T2D development such as *MTNR1B* [76, 97] and *CRY2* [83]. Interestingly, subsequent work found that these associations were season-dependent [98]. Other unanticipated enriched pathways that have been highlighted by these approaches include pathways related to the CREBBP-related transcription factor activity, cell cycle regulation and adipocytokine signalling [96]. Results also show an enrichment of pancreatic islet enhancer clusters in T2D and fasting glucose (FG) associated loci showcasing how integration of genetic information with knowledge of regulatory features can help identify processes affecting traits and aid in fine-mapping and finding causal variants [99]. Integrative approaches looking at mechanisms underlying insulin resistance have also revealed a pivotal role of storage capacity of peripheral adipose tissue in insulin-resistant cardiometabolic disease [100]. Loci identified via GWAS have also highlighted novel regulatory pathways involved in lipoprotein metabolism like in the case of *SORT1*, a locus harbouring variants associated with LDL-C and myocardial infraction (MI), which was shown to modulate hepatic VLDL secretion in mouse [101].

Our increased understanding of the biology behind many of these traits through GWAS has also led to clinically relevant applications. One important genetic tool in this context is the genetic risk score (GRS). For any given complex trait, GRS are often constructed by summing the number of risk alleles present in an individual and usually weighing this sum by the effect size of each one of these risk alleles. In cases like CVD, GRS can now outperform traditional risk factors for risk prediction which makes incorporation of genetic testing in the

clinic a valuable addition [102]. GRS for coeliac disease also show improvements in risk prediction over traditional methods [103]. With the increasing prevalence of obesity in younger individuals, GRS scores for T1D can be used to discriminate between T1D and T2D diagnosis as the genetic overlap between these two traits is very low [104]. In cases like obesity, traditional risk factors such as family history and childhood obesity are still outperforming GRS for risk prediction [105]. Nevertheless, obesity GRS has been helpful in Mendelian randomisation approaches to identify phenotypes where obesity is causal, therefore clarifying the relationship between obesity and many of its co-morbidities (**Figure 1.5**) [106].

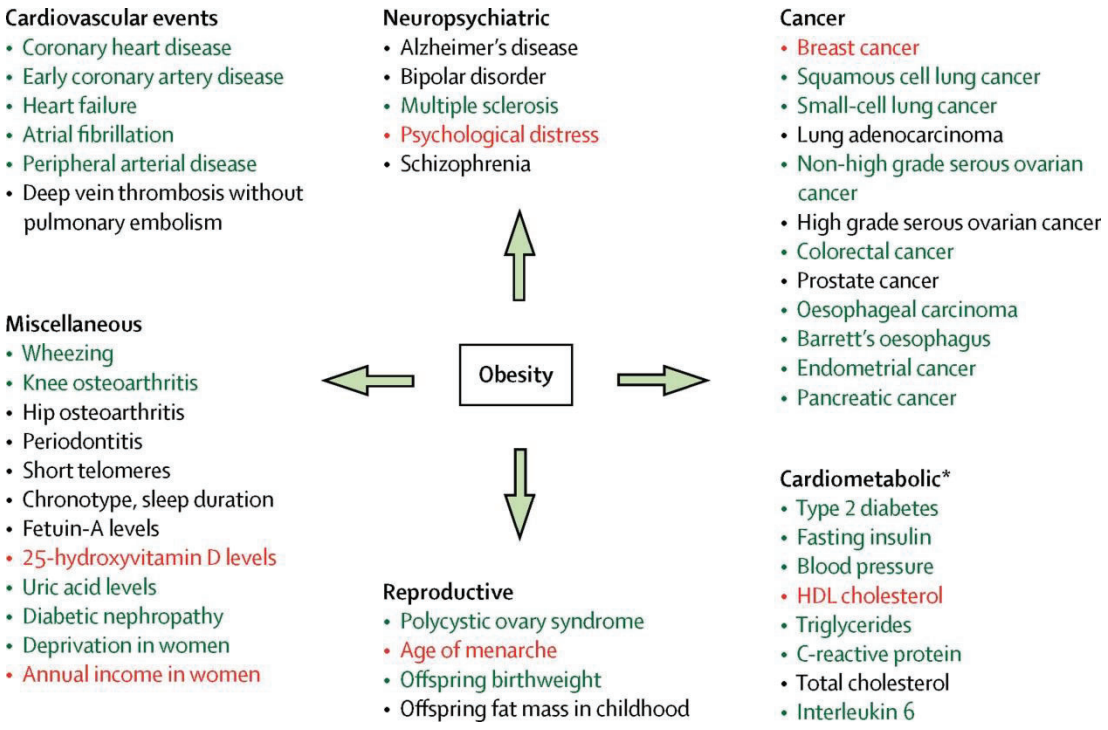

Figure 1.5: Inferences of causality of obesity derived from Mendelian randomisation studies. Only phenotypes with most consistent evidence are shown. Phenotypes in green are those for which there is a positive causal association of obesity whereas phenotypes in red are those with a negative causal association. Phenotypes in black are those where mendelian randomisation approaches have shown no causal role of obesity. Figure extracted from Goodarzi, M.O (2018) [106].

Mendelian randomisation analysis is a method that uses genetic instruments to assess the causality of a modifiable exposure on an outcome of interest [107-110] (**Figure 1.6**). In addition to ascertaining the causal role of obesity on its co-morbidities, this approach has also been used to identify the causal relationship between additional traits and disease. For example, it has demonstrated that the influence of lipid measurements such as LDL-C and HDL-C on T2D [84] and CVDs [111-114] risk is dependent on the particular pathway involved. That is, only some pathways that reduce LDL-C have an impact on T2D incidence [84] and only some genetic mechanisms that increase HDL-C have an impact on CVD risk [110, 112] (more details presented in **Chapter 3**).

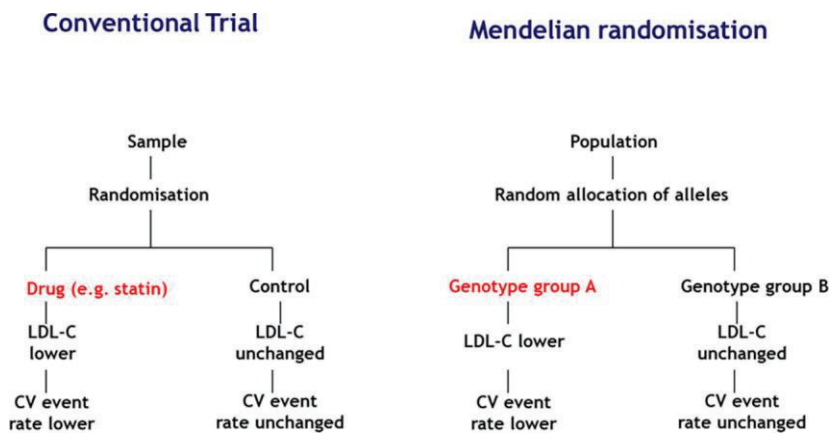

Figure 1.6:Comparison of conventional clinical trial with a Mendelian randomisation (MR) study. In a conventional trial, trait reducing treatment (in this case statins and LDL-C) is randomly allocated in a population and comparing the treated and untreated group allows you to assess if the trait (LDL-C) has an impact on the outcome (CV event). In a MR study, we look at the random allocation of alleles in a population at birth and use associated genetic variants as an instrument to assess the impact of the trait on the outcome. Extracted from Bennet D.A et al (2017) [115].

GWAS has also helped identify potential drug targets. Even though common variation near a gene identified via GWAS can have a very small effect on the trait, targeting the gene itself might lead to potential clinical benefits (e.g common variation near *HMGRC* has a small effect on LDL-C but its targeting via statins [116] had been previously shown to successfully treat hypercholesterolaemia). Loss-of-function (LoF) variants in *APOC3* have been associated

with a favourable lipid profile and reduced CVD risk suggesting the gene is a good candidate for lipid lowering drugs [117]. Another gene where protective LoF variants have been identified is *SLC30A8*, where carriers of rare protein-truncating variants have 65% reduced T2D risk highlighting this gene as a potential T2D drug target as well [118]. Not only can GWAS help identify drug targets, it can also influence treatment choice for certain conditions. For example, response to treatment of T2D via sulfonylureas can be influenced by variants near *TCF7L2* [119]. Another example is response to fenofibrate, a lipid lowering medication, which can be influenced by variants near *APOA1* [120].

Finally, another way GWAS could be used in the clinical setting is by identifying alleles that can influence accuracy of disease diagnostics. One notable example is potential improvement in T2D diagnosis using HbA1c in individuals with African American ancestry. HbA1c is a measurement of protein glycation reflecting average glucose concentration in the blood during the lifespan of an erythrocyte (~ 3 months). Usage of HbA1c as a T2D diagnostic tool can sometimes be hampered by the fact that HbA1c levels can be affected via conditions altering lifespan of eyrthrocytes independent of blood glucose levels (more details in **Chapter 4**). A GWAS on HbA1c has identified a variant with high prevalence in individuals with African American ancestry (MAF=11%) near *G6PD* that affects HbA1c levels by shortening the life span of red blood cells. It is estimated that screening for this variant would avoid 650,000 false negative T2D diagnoses in African Americans in the US [121].

1.2.3 Open questions/ unresolved issues:

Despite greater understanding of the genetic architecture of many traits, the proportion of heritability explained remains below 10-15% for most, and causal variants for associated loci are mostly unknown [122]. Early on, one possible explanation for this “missing heritability”

was that a substantial proportion of the heritability of complex traits can be explained by rare variants with large effects that aren't captured by standard genotyping platforms [123]. This is also known as the common disease / rare variant (CD/RV) hypothesis in contrast to the CD/CV hypothesis. At the time of this thesis though, data does not support this hypothesis and accumulating evidence suggests that for traits like height and BMI, most of the heritability will be explained by common variation (see **Section 1.2.2**). Another model that attempts to explain gaps in knowledge and suggest future directions for association studies is the "omnigenic model" that argues that a large number of loci will affect a given trait through indirect effects in regulatory networks affecting a core number of genes that affect the disease directly [124]. To address the "missing heritability" problem, several approaches have been proposed. Larger imputation reference panels such as combined UK10K [91] and 1000G Phase III [72] or the haplotype reference consortium (HRC) [125] have greatly increased imputation accuracy, especially for low-frequency and rare variants achieving good correlations (r2 >0.6) between imputed genotype dosages and masked genotypes for variants with a MAF as low as 0.5% in UK10K and 0.1% in HRC [126, 127].

Denser genotyping arrays enriched for low-frequency variants in coding regions are also powerful approaches since variants in these regions normally have a high phenotypic impact and are therefore under selective pressure [91, 128, 129]. Some arrays like the UK Biobank Axiom Array [130] combine the "exome component" with a "GWAS component" designed to enhance genome-wide imputation of common and low-frequency variants in a specific population. Another way to analyse rare coding variation is by doing whole-exome sequencing (WES) which uses target-enrichment methods to selectively capture exonic regions during library preparation before sequencing. As next-generation sequencing

technologies costs continue to decrease, whole-genome sequence (WGS) becomes a viable alternative that allows us to explore noncoding variation at a higher resolution. An important finding highlighting the relevance of honing in on low-frequency and rare coding variation is that variants identified via these approaches are better than common coding variants at identifying enriched gene sets associated with traits such as BMI suggesting that we are more likely to find causal variants with these approaches [93]. Sequencing studies have found multiple rare variants in candidate genes such as variants in *PCSK9* associated with LDL-C [131], variants in *ABCA1*, *APOA1* and *LCAT* associated with low HDL-C [132] or variants in *ANGPTL4* associated with reduced TG and high HDL-C [133] suggesting an important role of rare variants in the genetic architecture of these traits. These approaches have also helped increase the number of known effector transcripts associated with T2D [82].

Population-scale studies coupled with these approaches allow increases in power especially when it comes to the analysis of rare variants. Several of these cohorts have already started appearing in different countries such as UK Biobank (UKBB) which consists of 500,000 deeply phenotyped UK individuals with genotype data currently available and sequencing data in the near future [134]; the All of Us Research Program which aims to recruit 1,000,000 United States individuals that will have genotype and whole genome sequencing data [135] or the China Kadoorie Biobank which has a similar sample size as the UK Biobank (~510,000 individuals) and has also been deeply phenotyped and genotyped on a custom array for Asian populations [136]. The availability of individual level genotype and deep phenotyping in these large datasets provides several advantages. Firstly, having a very large dataset instead of meta-analysing various small studies is more convenient in terms of

dealing with between-study heterogeneity [137, 138], or sample overlap [139]. Secondly, it enables multi-trait analyses across multiple potentially correlated traits, which is more powerful than combining results from univariate analysis even when genetic correlation of the traits is weak [140, 141]. It also provides extra information on the covariance of these traits that would be missed when comparing summary statistics from different studies [142]. The availability of linked medical health records facilitates the study of pleiotropy (i.e the influence of one locus across multiple phenotypes) of genetic variants using methods such as phenome wide association studies (PheWAS) [143-145]. PheWAS are studies where a variant or subset of variants (normally previously linked to a trait of interest) are tested against a wide number of phenotypes simultaneously to examine the pleiotropic effects of these variants. Availability of linked medical health records also allows inferences to be made regarding the causality of traits in certain diseases. Finally, we can also evaluate GxE interactions by collecting multiple environmental data for these individuals [146, 147]. Recent work in UK Biobank, has been able to find predicted LoF variation protective against diseases such as T2D, asthma and coronary artery disease in the UK population bolstering the case for usage of large-scale population studies with dense genome-wide genetic data to identify potential drug targets [148]. Sequencing data in these large cohorts will provide new opportunities to explore the impact of rare variation in the aetiology of complex traits.

Another area of on-going improvement is that of diversity in studied populations. To date, most association studies have been performed in individuals of European ascent. But there are several advantages to be gained by increasing diversity. Firstly, effect sizes can vary between populations due to differing environmental factors which is crucial if one wants to use genetic information in the clinic to assess disease risk in non-European individuals. As

highlighted also by trans-ethnic HbA1c work [121], allele frequency also can differ widely between populations and some prevalent variants in a specific population are of particular value in the diagnostic setting. These differences in allele frequency also have aided in identifying associations of different cardiometabolic traits such as T2D and cardiomyopathies with variants that are rare or monomorphic in European populations [149-151]. Population isolates in particular are helpful to study rare variation as population events such as bottlenecks, genetic drift and endogamy can lead to an enrichment of rare alleles[152, 153]. Finally, the differing LD structure between populations can be helpful in fine-mapping efforts to identify causal variants [154-157].

Structural variations, such as CNVs, have also been currently underexplored but several links of structural variation to complex traits have been found such as autism [158], schizophrenia [159], severe childhood obesity [160, 161], asthma and obesity [162],several anthropometric traits[163] and T2D [164]. Currently array-based comparative genomic hybridisation (aCGH) is considered the gold standard for CNV detection [165] although platform-dependent differences in sensitivity have been a source for concern [166]. Usage of sequencing as a viable alternative has been explored [167, 168] and as WES and WGS becomes more prevalent, long-read sequencing technology improves and algorithms to analyse such data continue being developed [169, 170], the number of studies exploring structural variant association with complex traits will likely increase significantly.

Improvement in measurement resolution for many quantitative traits is also a promising avenue moving forward. GWAS studies using over 500 metabolites measured on the Metabolon platform or high resolution nuclear magnetic resonance (NMR) measurements of lipoprotein and lipid traits have found associations with effect sizes that are unusually

large for GWAS and enrichment of druggable targets in metabolomics loci [38, 171-173]. In addition to this, proteomics platforms such as OLINK have been helpful to identify variants regulating proteins that have been previously implicated in cardiovascular disease [174].

1.3 Thesis aims

In this thesis, the overarching aim is to gain further insights into the genetic architecture of different cardiometabolic traits through a combination of approaches with greater genotypic and phenotypic resolution. The main aim for each of the three results chapters in this thesis is described below:

1. In chapter 2, the aim is to characterise the genetic architecture of persistent and healthy thinness and contrast it to that of severe early onset obesity in two clinically ascertained cohorts.
2. In chapter 3, the aim is to gain novel insights into metabolic biomarker biology by analysing the contribution of rare variants to high resolution metabolic measurements.
3. In chapter 4, the aim is to characterise the genetic architecture of fructosamine, a measurement of total serum protein glycation, and explore the influence of previously established glycaemic loci on the trait.