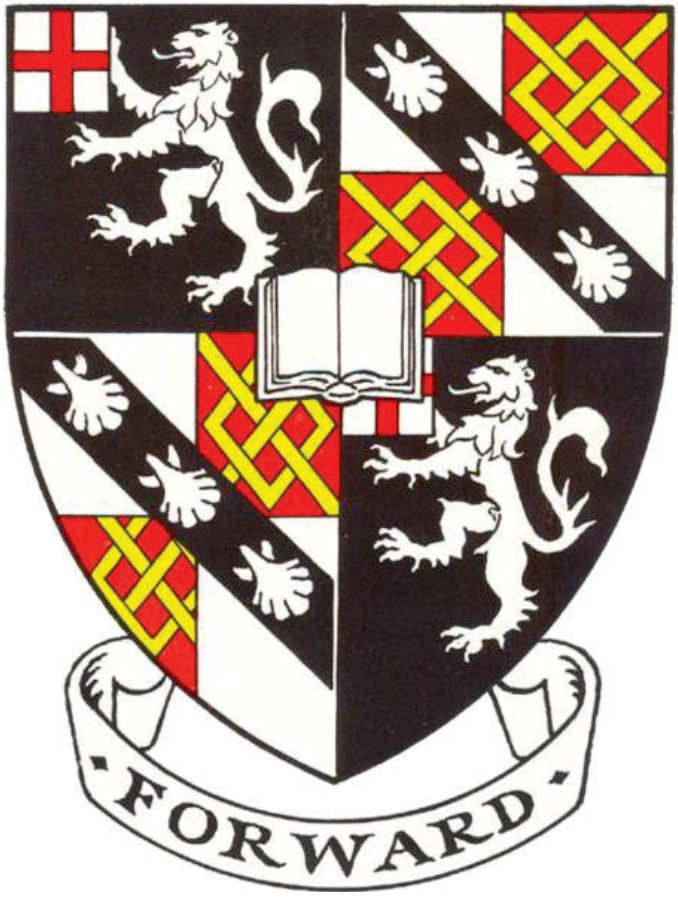
Genetic studies of cardiometabolic traits



Fernando Riveros Mckay Aguilera

Churchill College

University of Cambridge

Wellcome Sanger Institute

September 2018

This dissertation is submitted for the degree of Doctor of Philosophy

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the contributions section within each chapter and/or specified in the text. It is not being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution. 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 or similar institution. It does not exceed the prescribed word limit for the Faculty of Biology.

Fernando Riveros Mckay Aguilera

September 2018

“The world is full of lies. Memory is fuzzy and unreliable. Words we say are often transformed and what ends up in the pages of history is an amalgamation of people’s perception of us through time. But science…man science is cool”

Winston Churchill

Abstract

Diet and lifestyle have changed dramatically in the last few decades, leading to an increase in prevalence of obesity, defined as a body mass index >30Kg/m^2, dyslipidaemias (defined as abnormal lipid profiles) and type 2 diabetes (T2D). Together, these cardiometabolic traits and diseases, have contributed to the increased burden of cardiovascular disease, the leading cause of death in Western societies.

Complex traits and diseases, such as cardiometabolic traits, arise as a result of the interaction between an individual's predisposing genetic makeup and a permissive environment. Since 2005, genome-wide association studies (GWAS) have been successfully applied to complex traits leading to the discovery of thousands of trait-associated variants. Nonetheless, much is still to be understood regarding the genetic architecture of these traits, as well as their underlying biology. This thesis aims to further explore the genetic architecture of cardiometabolic traits by using complementary approaches with greater genetic and phenotype resolution, ranging from studying clinically ascertained extreme phenotypes, deep molecular profiling, or sequence level data.

In chapter 2, I investigated the genetic architecture of healthy human thinness (N=1,471) and contrasted it to that of severe early onset childhood obesity (N=1,456). I demonstrated that healthy human thinness, like severe obesity, is a heritable trait, with a polygenic component. I identified a novel BMI-associated locus at *PKHD1*, and found evidence of association at several loci that had only been discovered using large cohorts with >40,000

individuals demonstrating the power gains in studying clinically ascertained extreme phenotypes.

In chapter 3, I coupled high-resolution nuclear magnetic resonance (NMR) measurements in healthy blood donors, with next-generation sequencing to establish the role of rare coding variation in circulating metabolic biomarker biology. In gene-based analysis, I identified *ACSL1*, *MYCN*, *FBXO36* and *B4GALNT3* as novel gene-trait associations ($P<2.5x10^{-6}$). I also found a novel link between loss-of-function mutations in the "regulation of the pyruvate dehydrogenase (PDH) complex" pathway and intermediate-density lipoprotein (IDL), low-density lipoprotein (LDL) and circulating cholesterol measurements. In addition, I demonstrated that rare "protective" variation in lipoprotein metabolism genes was present in the lower tails of four measurements which are CVD risk factors in this healthy population, demonstrating a role for rare coding variation and the extremes of healthy phenotypes.

In chapter 4, I performed a genome-wide association study of fructosamine, a measurement of total serum protein glycation which is useful to monitor rapid changes in glycaemic levels after treatment, as it reflects average glycaemia over 2-3 weeks. In contrast to HbA1c, which reflects average glucose concentration over the life-span of the erythrocyte (~3 months), fructosamine levels are not predicted to be influenced by factors affecting the erythrocyte. Surprisingly, I found that in this dataset fructosamine had low heritability (2% vs 20% for HbA1c), and was poorly correlated with HbA1c and other glycaemic traits. Despite this, I found two loci previously associated with glycaemic or albumin traits, *G6PC2* and *FCGRT* respectively *($P<5x10^{-8}$),* associated with fructosamine suggesting shared genetic influence.

Altogether my results demonstrate the utility of higher resolution genotype and phenotype data in further elucidating the genetic architecture of a range of cardiometabolic traits, and the power advantages of study designs that focus on individuals at the extremes of phenotype distribution. As large cohorts and national biobanks with sequencing and deep multi-dimensional phenotyping become more prevalent, we will be moving closer to understanding the multiple aetiological mechanisms leading to CVD, and subsequently improve diagnosis and treatment of these conditions.

Acknowledgements

Ok, so from my understanding this is pretty much the free flow section of this work. So yeah, firstly, all of these you're about to read (or skim through) would not be at all possible without the dynamic duo of Inês Barroso and Eleftheria Zeggini. Initially I thought I might even have to transform acknowledgements into its own chapter with a subsection for each because there's so much I want to say (yet so little that realistically would come out). Yet if I could present an abstract for the "acknowledgements chapter" that only exists in my head is this: Inês, thank you for the amazing support, for the constant challenges and helping me feel like this process and piece of work is truly an achievement. Ele, thank you so much for making me feel like part of your team, Volos and everything it represents for my career and life ambitions. Both of you are top notch mentors and role models and the science discussions I've had with both of you in and out of work have truly enriched my experience. It also must be said that I truly appreciate all the emotional support throughout the hardships endured this four years. And on the topic of mentorship and support, I cannot end this paragraph without thanking Eleanor Wheeler who was incredibly patient and supportive as my day to day supervisor and really helped me get started here and was the filter for my dumbest questions so that Inês or Ele wouldn't have to deal with those. Ellie, you rock.

Moving on to other people in my science life, thank you to my thesis committee Nicole Soranzo and Adam Butterworth. I really appreciate your input into my work and for integrating me into some of the research done by you to get an early feel on big collaborative research. Thanks also Carl Anderson for the discussions and his excellent role as head of Grad Programme. Thanks Darren Logan for being so helpful ever since the first interview. Thanks past and present members of team35 and team144 for all the help in multiple stages of my PhD. It is truly a privilege to be surrounded by such a diverse group of smart people with different expertise. At Churchill College, special shout out to Rebecca Sawalmeh who was incredibly important in my college life. Also Rita and Barry for the mentorship dinners.

Now let's get a bit more personal (just a bit). I formed an amazing group of friends here at Sanger, met a bunch of super cool people and these people made life sooooo much easier. I'll start kind of historically. But thanks Ximena and Martin, for being there for me since my first jetlagged interview and help me get settled here and teaching me what it's like to be a Mexican in Cambridge. Sophie, the first friend I made here (and very wisely done), the funniest woman I know and a pillar of emotional support throughout, and the cakes..oh god the cakes. Thank you Neneh, my big sister whose calming and soothing demeanour are incredibly contagious. Patrick Short: the man, the legend. What would life have been like if we hadn't shared that car to work. Thank you to Loukas, my life mentor who has taught me

so much and yet I'm still constantly learning more. Thank you to Arthur, I can't believe I didn't hang out with you more often in the early days; you're like jalapeno to my life adding a much needed spice. My best fitbit friend forever (bfff) Katharina for keeping me fit and all the fitbit walks. Miguel, thanks for the jamming sessions and the active lifestyle guidance. Thanks to all my friends back in Mexico, especially: Yoshi, Yann, Sebas and Palas. Then finally the gang: Dim, Lil, Veli, Alice, GM and Mash. The centre of my social life and the people that made me want to go to work every day. Thanks all of you for being there in my hour of need as well. Thanks GM for trying to get me out of my house constantly when all I want is sulk. Thanks Alice for being there for me when I needed to talk. Thanks Veli for the closeness we've achieved this past year and all the life advice. Dim..D-bone. Dude..duude. Thanks for being there through the best and the worst, literally. Lili, thanks for all the selfless acts of kindness, for the excellent company on my way to Sanger and for calling me a "stupid idiot" when I needed it. And thanks Mash…just coz.

Last but not least, I'd like to thank my family. I've been incredibly lucky to have been born with the parents I have. I'm lucky to have shared most of my life with them and my siblings. I've been really lucky to get the random combination of genetic variants and environment they provided to form a complex human being that is about to get a PhD (maybe?). So yeah…thank you infinitely.

Ok, so that should be it. But also, one thing I must say is that this thesis has a soundtrack. Every chapter was written with an album playing in the background. You don't have to listen to it yourself, but …I mean..if you want to :

- Chapter 1: Gaslight Anthem – Handwritten
- Chapter 2: Sum 41 – Underclass hero
- Chapter 3: PXNDX – Para ti con desprecio
- Chapter 4: Hamilton (the musical)
- Chapter 5: My Chemical Romance – Danger days …

Perfect. So..now..if you read all of this..sorry ..and thanks. Enjoy the rest of the thesis (or bye if you only read this).

P.S: I'm just kidding Mash, you know you're the best friend in the world. I would not have survived Cambridge, especially not the last year without you. You're key to my sanity. Thank you so much for everything.

Publications

From this thesis:

Riveros-Mckay F, Oliver-Williams C, Karthikeyan S, Walter K, Kundu K, et al. Sequencing reveals role of rare variation in circulating metabolic biomarkers. (In preparation)

Riveros-Mckay F*, Mistry V, Bounds R, Hendricks A, Keogh JM, et al. Genetic architecture of human thinness compared to severe obesity. *PLOS Genetics*. (in press)

Arising elsewhere:

Huckins LM*, Hatzikotoulas K*, Southam L, Thornton LM, Steinberg J, **Aguilera-McKay F**, et al.. (2018) Investigation of common, low-frequency and rare genome-wide variation in anorexia nervosa. *Mol Psychiatry*. 2018 May;23(5):1169-1180. doi: 10.1038/mp.2017.88. Epub 2017 Jul 25.

Astle WJ*, Elding H*, Jiang T*, Allen D, Ruklisa D, Mann AL, Mead D, Bouman H, **Riveros-Mckay F**, et al. (2016) The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease.*Cell*. 2016 Nov 17;167(5):1415-1429.e19.doi: 10.1016/j.cell.2016.10.042

Table of Contents

List of Figures

List of Tables

List of Abbreviations

Abbreviation	Full name
1000G	1000 Genomes
AC	Allele count
ALSPAC	Avon Longitudinal Study of Parents and Children
AN	Allele number
BMI	Body mass index
CD/CV	Common disease/common variant
CD/RV	Common disease/rare variant
CHD	Coronary heart disease
CI	Confidence intervals
CNV	Copy number variant
CVD	Cardiovascular disease
DZ	Dizygotic
EA	Effect allele
EAF	Effect allele frequency
eQTL	Expression quantitative trait locus
FG	Fasting glucose
FI	Fasting insulin
GG	Glycation gap
GOOS	Genetics Of Obesity Study
GRS	Genetic risk score
GWAS	Genome-wide association study
GxE	Gene-by-environment
H^2	Broad-sense heritability
h^2	Narrow-sense heritability
HbA1c	Glycated haemoglobin
HDL-C	High-density lipoprotein cholesterol
HOMA-B	Beta cell function by homeostasis model assessment
HOMA-IR	Insulin resistance by homeostasis model assessment
HRC	Hapoltype Reference Consortium
HWE	Hardy-Weinberg equilibrium
IMS	Institute of Metabolic Sciences
LD	Linkage disequilibrium
LDL-C	Low-density lipoprotein cholesterol
LoF	Loss-of-function
MAC	Minor allele count
MAF	Minor allele frequency
MI	Myocardial infarction
MR	Mendelian randomisation
MZ	Monozygotic
N	Sample size
NEA	Non-effect allele

NGS	Next-generation sequencing
NMR	Nuclear magnetic resonance
OGTT	Oral glucose tolerance test
OR	Odds ratio
p	P-value
PCA	Principal component analysis
PheWAS	Phenome-wide association study
QC	Quality control
r	Correlation coefficient
r^2	Coefficient of determination
RG	Genetic correlation
SCOOP	Severe Childhood Onset Obesity Project
SE	Standard error
SNP	Single nucleotide polymorphism
STILTS	STudy Into Lean and Thin Subjects
T2D	Type 2 diabetes
TC	Total cholesterol
TG	Triglycerides
UKBB	UK Biobank
UKHLS	UK household longitudinal study
WES	Whole-exome sequencing
WGS	Whole-genome sequencing
WHR	Waist-to-hip ratio
WSI	Wellcome Sanger Institute

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].

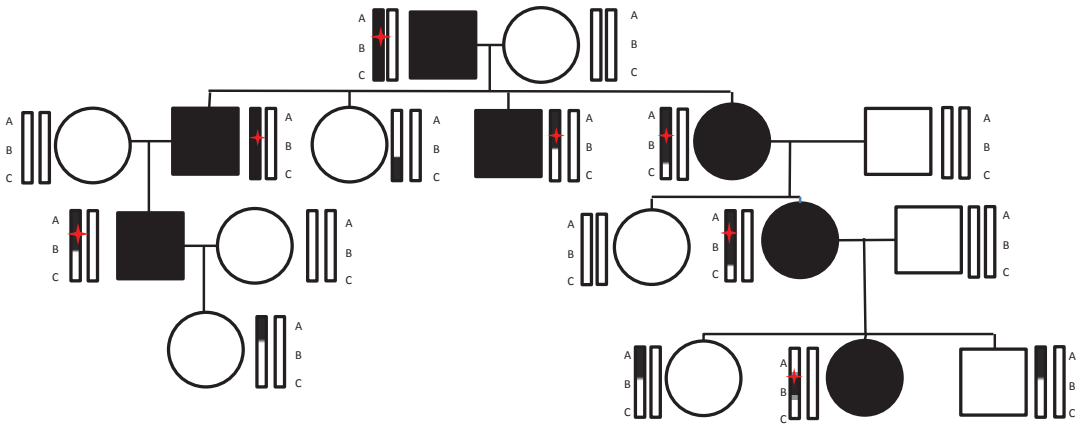


Figure 1.1: Principles of linkage analysis. A family pedigree is shown from a typical linkage analysis study for a Mendelian dominant disorder. Square (males) and circles (females) in black indicate affected individuals whereas symbols with no fill indicate unaffected individuals. Rectangles next to 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.

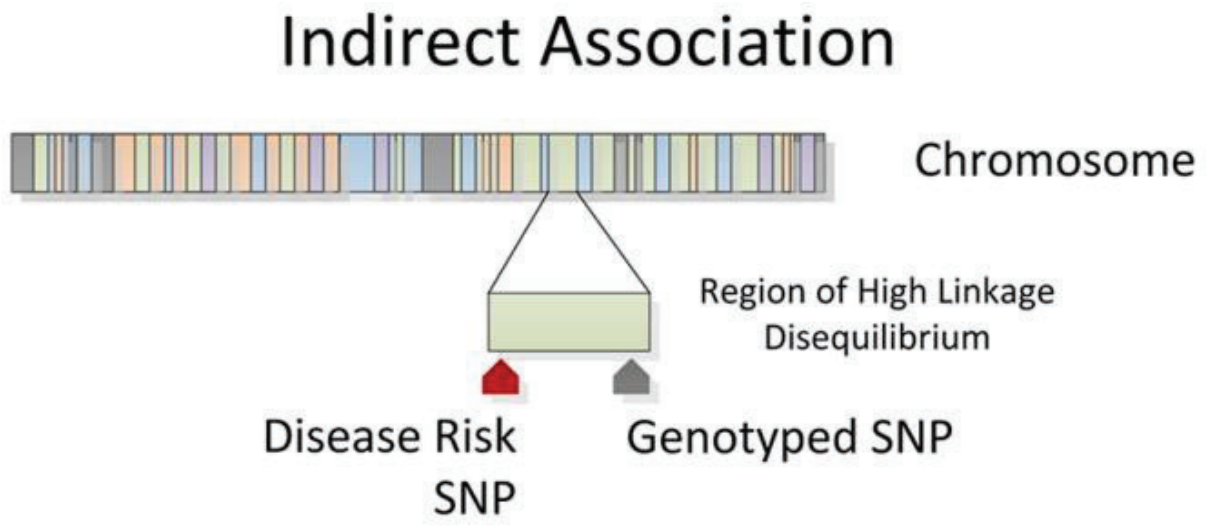


Figure 1.2: Indirect association. In a GWAS more often than not, the tested allele is not the causal allele. GWAS takes advantage of LD to identify regions of the genome associated with a phenotype by using SNPs in high LD with the causal allele. Figure extracted from Bush W.S and Moore J.H (2012) [62].

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.

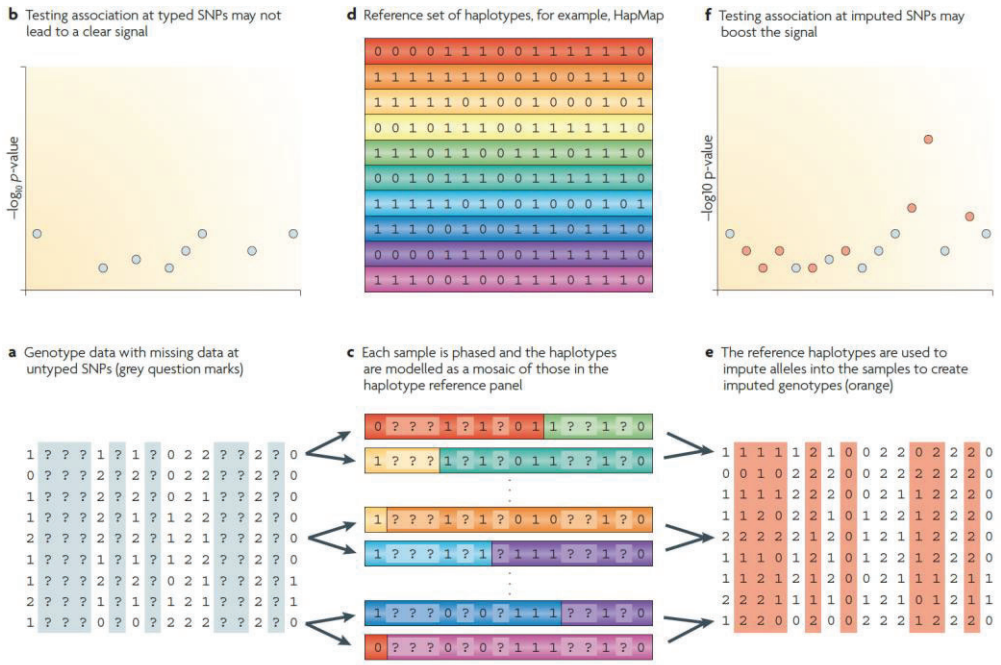


Figure 1.3: Genotype imputation process. A) Genotype data from individuals is collected with missingness at certain sites. B) Testing association only at directly genotyped sites may not lead to a significant signal. C) Samples are phased and haplotypes are modelled as mosaics of the haplotypes present 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**).

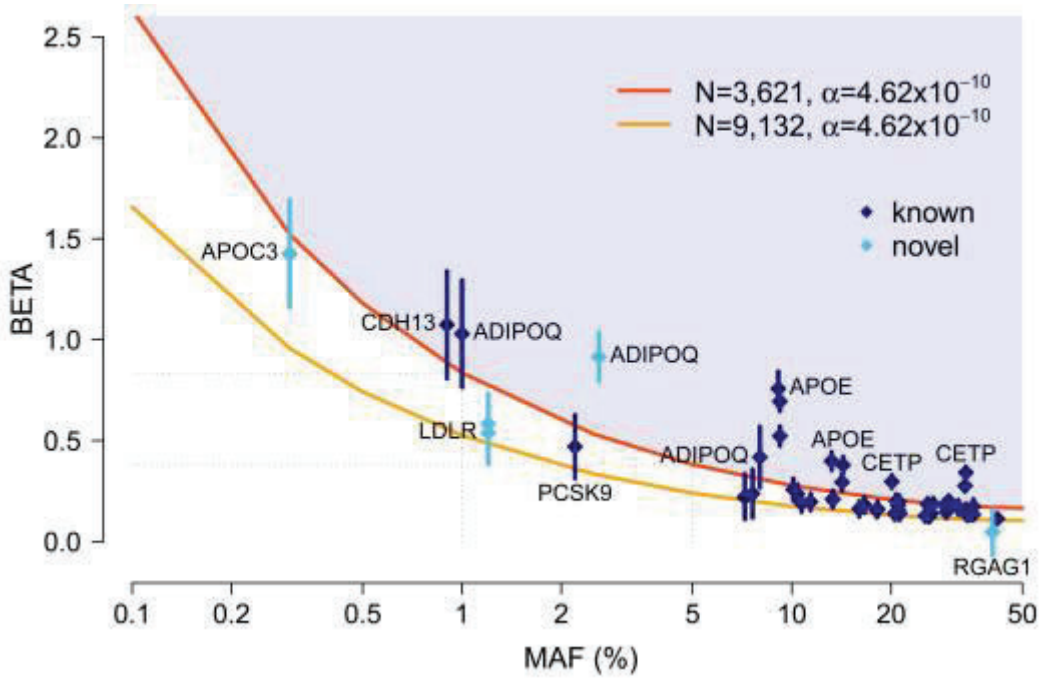


Figure 1.4: Results from single point association analysis in UK10K for 31 core traits shared between TwinsUK and ASLPAC cohorts. Minor allele frequency of variants is plotted on the X axis and effect sizes are plotted on the Y axis. Known associations are coloured in dark blue whereas novel associations 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].

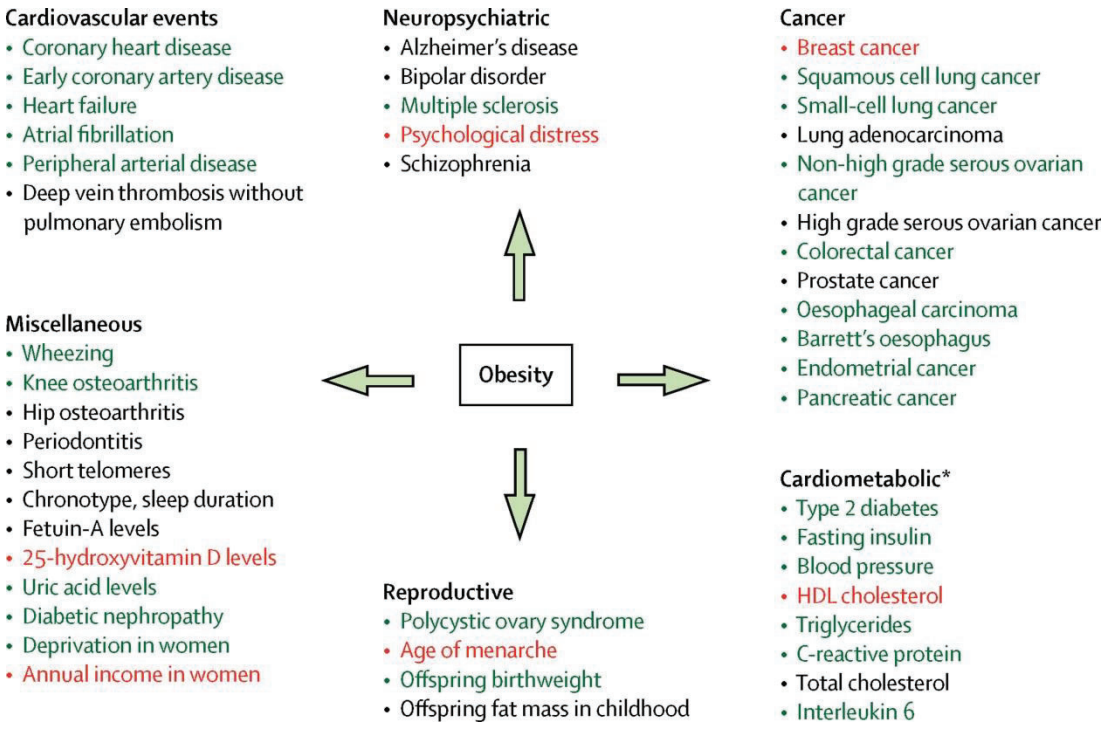


Figure 1.5: Inferences of causality of obesity derived from Mendelian randomisation studies. Only phenotypes with most consistent evidence are shown. Phenotypes in green are those for which there is a positive causal association of obesity whereas phenotypes in red are those with a negative causal association. 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**).

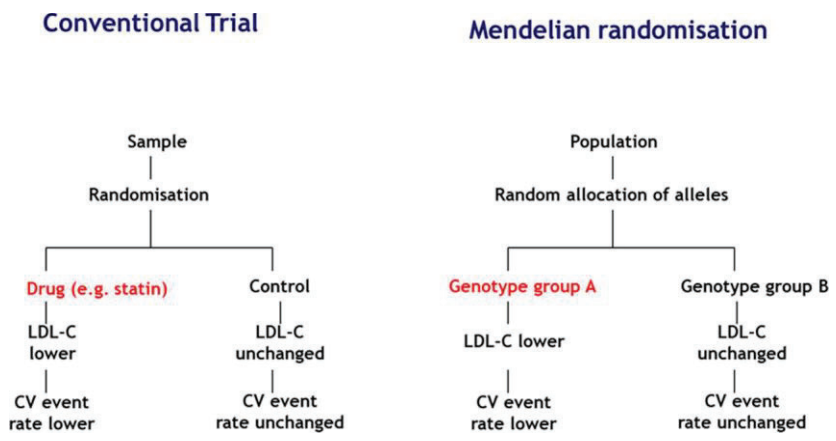


Figure 1.6:Comparison of conventional clinical trial with a Mendelian randomisation (MR) study. In a conventional trial, trait reducing treatment (in this case statins and LDL-C) is randomly allocated in a population and comparing the treated and untreated group allows you to assess if the trait (LDL-C) 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.

2 Chapter 2: The Genetic Architecture of Human Thinness

2.1 Introduction

Obesity, defined as a body mass index (BMI) greater than 30kg/m^2, is one of the leading causes of preventable death worldwide [175]. In recent years, the prevalence of obesity has risen and this has been linked to an increasingly "obesogenic" environment (i.e an environment promoting the consumption of high calorie foods and reduced levels of physical activity [176]). However, within a given environment, there is considerable variation in body weight; some people are particularly susceptible to severe obesity, whilst others remain thin [177, 178]. Indeed BMI heritability estimates from multiple family, twin and adoption studies range from 40% to 70% which suggests that genetic factors play a major role in the development of obesity [179]. To date, most studies aimed at understanding the aetiology of obesity have focused on BMI as a continuous trait, and have identified more than >900 common and low-frequency obesity-susceptibility loci [80, 93, 180-184]. Additionally, studies of people at one extreme of the distribution (severe obesity) have led to the identification of rare, penetrant genetic variants that affect key molecular and neural pathways involved in human energy homeostasis[185-192]. These findings have provided a rationale for targeting these pathways for therapeutic benefit. One such example is the development of drugs targeting *MC4R* [193] which harbours both, rare highly penetrant variation [194, 195] and downstream common variation with modest effect size [93, 196]. In contrast, little is known about the specific genetic characteristics of persistently thin individuals (thinness defined using WHO criteria BMI$\leq$18kg/m^2).

A small number of previous studies have found that thinness appears to be a trait that is at least as stable and heritable as obesity [197-200]. A large study of 7,078 UK children and adolescents, found that the strongest predictor of child/adolescent thinness was parental weight status. The prevalence of thinness was highest (16.2%) when both parents were thin and progressively lower when both parents were normal weight, overweight or obese [201]. There is also some evidence for gene dosage playing a role in both tails of the BMI distribution. A deletion in 16p11.2 has been shown to associate with a highly penetrant form of obesity, whereas its reciprocal duplication is associated with underweight status [202]. Another copy number variant in 20q13.3 is associated with less severe forms of obesity and thinness, with deletions observed in obese, and duplications observed in thin probands (defined in this particular study as BMI <= 23 kg/m^2) [203].

Despite evidence for genetic factors contributing to the phenotypic variance at both tails of the BMI distribution, at the time of this study, GWAS approaches that had included thin individuals had either used them exclusively as controls to contrast with extreme obesity [204], and/or they had not ascertained for healthy thinness [205]. Understanding the mechanisms underlying thinness/resistance to obesity may highlight novel anti-obesity targets for future drug development [206]. To do this there are two possible study designs, each with its own advantages and disadvantages. One approach uses a population-based cohort, where data for participants at the tails of the distribution are extracted, and each is compared to the other in a case-control analysis. This approach was used effectively by Berndt et al 2013 [207] who analysed the top and bottom 5% of each cohort that contributed to the original GIANT BMI meta-analysis [208]. One of the biggest advantages of this approach is that it is less susceptible to age, sex and other environmental effects

influencing observations. The disadvantage is that, by their very definition, such population based cohorts often contain a limited number of people at the "extremes" (i.e. severe obesity and thinness) [207]. For example, in the full UK Biobank release (N= 487,411), there are only 626 individuals with a comparable level of obesity as those present in children from the Severe Childhood Onset Obesity Project (SCOOP) cohort (BMI standard deviation score >3, age of onset <10yr) which has been previously used to identify novel loci associated with obesity [160]. The second approach is particularly useful for complex disorders where environmental exposure can have a strong influence on the development of the condition (e.g. asthma, type 2 diabetes and obesity). Here, one maximises genetic load in the cases by carefully selecting affected individuals that are less likely to have been exposed to environmental risk factors. For example, one might select individuals with early age of onset for the condition which lessens the amount of time they would have been exposed to environmental factors [160, 209]. To complement this approach to the selection of cases, controls are also selected to increase the chances that they do not have the disease or are unlikely to develop the disease later in life [204]. This is normally done by selecting contrasting controls, or "super-controls". The advantages of this approach as a way to increase power have been shown in multiple studies [210-212] including the previously mentioned study performed by our group using the SCOOP cohort uncovering new loci that had been missed by conventional BMI GWAS at the time [160]. One of the limitations of this approach is that it is more susceptible to differential effects of age, sex and other environmental factors between cases and controls.

In this chapter, I describe a genetic study that used this case-"super control" design to begin to dissect the genetic architecture of healthy human thinness. To do this our group

collaborated with Professor Sadaf Farooqi's group who recruited a new cohort of healthy thin individuals from the UK (STudy Into Lean and Thin Subjects, STILTS cohort; mean BMI = 17.6 kg/m^2) and who had previously recruited the SCOOP cohort. My work focused on all analytical elements of the study.

2.2 Chapter aims

The overall aim of this chapter is to contrast the genetic architecture of persistent healthy thinness with that of severe early onset obesity. In this chapter I use genome-wide directly genotyped and imputed data from two clinically ascertained cohorts (STILTS and SCOOP) and two population cohorts (the UK household longitudinal study (UKHLS) and UK Biobank (UKBB)) to:

I. Assess the heritability of persistent healthy thinness.
II. Identify the contribution of established BMI loci at the extremes of the phenotype distribution.
III. Discover novel loci associated with either tail of the BMI distribution.

2.3 Methods

2.3.1 Cohorts

SCOOP, STILTS and UKHLS cohorts were used for the heritability, genetic correlation, genetic risk score and association analyses with established BMI loci, as well as, used as a discovery cohort in the genome-wide association study (GWAS). UK Biobank samples were used for genetic correlation analysis and in the replication stages of the GWAS. ALSPAC was used to for sensitivity analyses in SCOOP vs UKHLS comparisons (**Figure 2.1**).

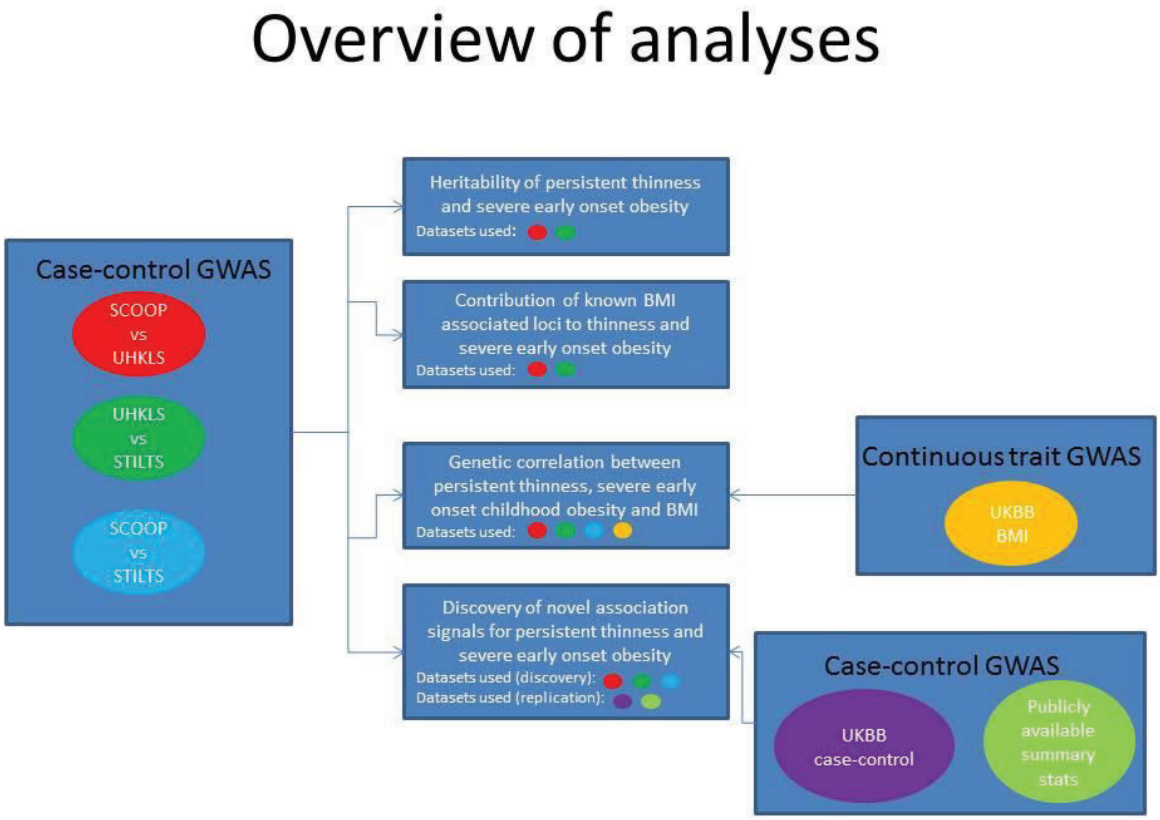


Figure 2.1: Overview of cohorts and analyses.

2.3.1.1 STudy Into Lean and Thin Subjects (STILTS)

Recruitment was performed by Professor Sadaf Farooqi's group at the Wellcome Trust-MRC Institute of Metabolic Science (IMS). The aim was to recruit a new cohort of UK European ancestry individuals who were thin (defined as a body mass index $\leq$ 18kg/m^2) and well. After ethical committee approval (12/EE/0172), they worked with the NIHR Primary Care Research Network (PCRN) to collaborate with 601 GP practices in England. Each practice searched their electronic health records using the inclusion criteria (age 18-65 years, BMI$\leq$18 kg/m^2) and exclusion criteria (medical conditions that could potentially affect weight (chronic renal, liver, gastrointestinal problems, metabolic and psychiatric disease,

known eating disorders). The case notes of each potential participant were reviewed by the GP or a senior nurse with clinical knowledge of the participant to exclude other potential causes of low body weight in discussion with the study team. Through this approach, 25,000 individuals were identified who fitted the inclusion criteria in the study. These individuals were invited to participate in the study; approximately 12% (2,900) replied consenting to take part. The team obtained a detailed medical and medication history, screened for eating disorders using a questionnaire (SCOFF) that has been validated against more formal clinical assessment [213] and excluded those who exercised vigorously (>6 metabolic equivalents (METs); http://www.who.int/dietphysicalactivity/physical_activity_intensity/en/). Prof Farooqi's group also excluded people who were thin only at a certain point in their lives (often as young adults), to focus on those who were persistently thin/always thin throughout life as this group would likely be enriched for genetic factors contributing to their thinness. The participants were asked this specific question to identify persistently thin individuals: "have you always been thin?" Only those who answered positively were included. Questionnaires were manually checked by senior clinical staff for these parameters and for reported ethnicity (non-European ancestry excluded). A small number of individuals (N=43) with a BMI of 19 kg/m^2 were included as they had a strong family history of thinness. 74% of the STILTS cohort have a family history of persistent thinness, suggesting there is an enrichment for genetically driven thinness. DNA was extracted from salivary samples obtained from these individuals using the Oragene 500 kit according to manufacturer's instructions.

2.3.1.2 Severe Childhood Onset Obesity Project (SCOOP)

The Severe Childhood Onset Obesity Project (SCOOP, N~4,800) cohort [160] is a sub-cohort of the Genetics Of Obesity Study (GOOS, N~7,000) [214] comprised of those individuals of British self-reported European ancestry. As for GOOS, all SCOOP participants recruited into the cohort have a BMI standard deviation score (SDS) > 3 and onset of obesity before the age of 10 years. SCOOP individuals likely to have congenital leptin deficiency were excluded by measurement of serum leptin, and individuals with mutations in the melanocortin 4 receptor gene (*MC4R*) (the most common genetic form of penetrant obesity) were excluded by prior Sanger sequencing. The cohort has ethical committee approval (MREC 97/5/21).

2.3.1.3 UK household longitudinal study (UKHLS)

United Kingdom Household Longitudinal Study (UKHLS) also known as Understanding Society (https://www.understandingsociety.ac.uk) is a longitudinal household study designed to capture economic, social and health information from 40.000 UK households (England, Scotland, Wales and Northern Ireland) representative of the UK population [215]. A subset of 10,484 individuals was selected for genome-wide array genotyping. Genetic and phenotype data was obtained through The Understanding Society Data Access Committee (DAC) application system. The United Kingdom Household Longitudinal Study has been approved by the University of Essex Ethics Committee and informed consent was obtained from every participant. This cohort was used as a control dataset with SCOOP and STILTS cases. UKHLS data is available for download in EGA with accession code EGAS00001001232.

2.3.1.4 UK Biobank (UKBB)

This study includes approximately 488,377 participants with genetic data released (including ~50,000 from the UKBiLEVE cohort [216]) of the total 502,648 individuals from UK BioBank (UKBB). UKBB samples were genotyped on the UK Biobank Axiom array at the Affymetrix Research Services Laboratory in Santa Clara, California, USA. The full release was imputed to the Haplotype Reference Consortium (HRC) [127]. UKBiLEVE samples were genotyped on the UK BiLEVE array which is a previous version of the UK Biobank Axiom array sharing over 95% of the markers. At the time of this study, 487,411 samples with directly genotyped and imputed data were available and data was downloaded using tools provided by UK Biobank. Extensive data from health and lifestyle questionnaires is available as well as linked clinical records. BMI, as well as other physical measurements were taken on attendance of recruitment centre. Severely obese participants in the available data were defined as those with BMI ≥ 40 kg/m^2 (N=9,706) and thin individuals were defined as those with BMI ≤ 19 kg/m^2 (N=4,538). For sensitivity analyses, to more closely match thin individuals in UKBB to the STILTS cohort, I also used ICD10 clinical records as well as self-reported medical data to exclude individuals whose low BMI could be explained by a medical condition (**Supplementary Tables 12-13 of Riveros-Mckay et al 2018 [217] (Appendix A))**. This resulted in a subset of 2,518 thin individuals who met the same health criteria as those in the STILTS cohort. Given that it has been previously shown that type I error rate for variants with a low minor allele count (MAC) is inadequately controlled for in very unbalanced case-control scenarios [218], I randomly subsampled 35,000 individuals from the original 487,411 genotyped individuals and removed those with BMI≤19 or BMI ≥30, to generate an independent control set. The 25,856 participants remaining after BMI exclusions from the

tails, generated a non-extreme set of individuals kept as putative controls. The other 452,411 genotyped samples were kept as the BMI dataset for downstream analyses (**Table 2.1**). An interim release consisting of a subset 152,249 individuals from UKBB was released in May 2015. This interim release was imputed to a combined UK10K and 1000G Phase 3 reference panel and contains several variants which are not currently present in the HRC panel, as such it was used in some of the analyses described.

	Thin (BMI ≤ 19)	Obese (BMI ≥ 40)	Controls (19 < BMI ≤ 30)	BMI Dataset
Initial sample sets	4,538	9,706	35,000	452,411
Final sample sets post QC	3,532	7,526	20,720 (BMI range 19-30)	387,164
Sex				
Male	719 (20%)	2,468 (33%)	9,467 (46%)	178,029 (46%)
Female	2,813 (80%)	5,058 (67%)	11,253 (54%)	209,134 (54%)

Table 2.1:Summary of UKBB sample sets

2.3.1.5 Avon Longitudinal Study of Parents and Children (ALSPAC)

The Avon Longitudinal Study of Parents and Children (ALSPAC) [219, 220], also known as Children of the 90s, is a prospective population-based British birth cohort study. Ethical approval for the study was obtained from the ALSPAC Ethics and Law Committee and the Local Research Ethics Committees. The study website contains details of all the data that is available through a fully searchable data dictionary (http://www.bris.ac.uk/alspac/researchers/data-access/data-dictionary/). ALSPAC children were genotyped using the Illumina HumanHap550 quad chip genotyping platforms by 23andme subcontracting the Wellcome Sanger Institute (WSI), Cambridge, UK and the Laboratory Corporation of America, Burlington, NC, US. Genotypes were imputed against

the 1000G Phase 3 reference panel using IMPUTE V2.2.2 [221, 222]. In the current study, analysis was restricted to a subset of unrelated (identity-by-state < 0.05 [39]) children with genetic data and BMI measured between the age of 12 and 17 years (n=4,964, 48.5% male). The mean age of the children was 14 years and the mean BMI 20.5.

2.3.2 Genotyping and quality control

2.3.2.1 SCOOP, STILTS and UKHLS

For the SCOOP cohort, DNA was extracted from whole blood as previously described [160]. For the STILTS cohort, DNA was extracted from saliva using the Oragene saliva DNA kits (online protocol) and quantified using Qubit. All samples from SCOOP, STILTS and UKHLS were typed across 30 SNPs on the Sequenom® platform (Sequenom® Inc. California, USA) for sample quality control by the Genotyping Facility at WSI. Of the 3,607 SCOOP and STILTS samples submitted for Sequenom genotyping, 3,280 passed quality controls filters which were i) degraded samples, ii) gender inference failure, iii) Sequenom failure or iv) low concentration (90.9% pass rate). Of the 10,433 UKHLS samples, 9,965 passed Sequenom sample quality control (95.5% pass rate). Subsequently, UKHLS controls were genotyped on the Illumina HumanCoreExome-12v1-0 Beadchip. The 3,280 SCOOP and STILTS samples, and 48 overlapping UKHLS samples (to test for possible array version effects) were genotyped on the Illumina HumanCoreExome-12v1-1 Beadchip by the Genotyping Facility at the WSI. Genotype calling was performed centrally for all batches at the WSI using GenCall. I excluded samples based on the following criteria: i) concordance against Sequenom genotypes <90%; ii) for each pair of sample duplicates, exclude one with highest missingness; iii) sex inferred from genetic data different from stated sex ; iv) sample call rate

<95%; v) sample autosome heterozygosity rate >3 SD from mean done separately for low (<1%) and high MAF(>1%) bins; vi) magnitude of intensity signal in both channels <90%; and vii) for each pair of related individuals (proportion of IBD (PI_HAT) >0.05), the individual with the lowest call rate was excluded. I performed SNP QC using PLINK v1.07 [223]. Criteria for excluding SNPs was: i) Hardy-Weinberg equilibrium (HWE) $p<1x10^{-6}$; ii) Call rate <95% for MAF≥5%, call rate <97% for 1% ≤MAF<5%, and call rate <99% for MAF <1%. SMARTPCA v10210 [224] was used for principal component analysis (PCA). To verify the absence of array version effects I used PCA on the subset of shared controls genotyped on both versions of the array. Cutoffs for samples that diverged from the European cluster were chosen manually after inspecting the PCA plot. SNPs with discordant MAFs in the different versions of the array were excluded. After removal of non-European samples and 13 samples due to cryptic relatedness, 1,456 SCOOP and 1,471 STILTS samples remained for analysis. For UKHLS, 82 samples were removed after applying a strict European filter and 680 related samples were removed by Vanisha Mistry after applying a '3rd degree" kinship filter in KING [225]. A total of 9,203 samples remained, of which 6,460 had a BMI >19 and <30 ("non-extremes").

2.3.2.2 UK Biobank

Sample QC was performed using all 487,411 samples using the sample QC file provided by UK Biobank. I used the following criteria to exclude samples: i) supplied and genetically inferred sex mismatches; ii) heterozygosity and missingness outliers; iii) not used in kinship estimation; iv) non-European British individuals; v) samples that withdrew consent and vi) for each pair of related individuals (KING kinship coefficient >=0.0442), I preferentially kept

cases (BMI ≥ 40 or BMI <=19), otherwise, I randomly selected one individual out of the pair. After sample QC, thirteen individuals with very extreme BMI values were also removed (BMI <14 or BMI >74). One of them had no genotype data, whereas the remaining twelve had underlying health conditions that could influence their BMI such as eating disorders, abnormal weight loss and COPD for eleven underweight individuals and hypothyroidism for one extremely obese individual. In the end, 7,526 obese (BMI ≥ 40), 3,532 thin (BMI ≤ 19) and 20,720 non-extreme controls (19 < BMI ≤ 30) remained for case-control analyses. In addition, 387,164 samples remained for analysis of BMI as a continuous trait. There was an overlap of 10,282 samples (~2.6% of the BMI dataset) with obese and thin cases (**Figure 2.2**). The same procedure was performed on the interim release of 152,249 UKBB samples to produce a set of 2,799 obese, 1,212 thin, 8,193 controls and 127,672 individuals for the independent BMI dataset. All genome-wide association analyses on UKBB were also performed on this subset to query variants that are not currently available in the full UKBB release.

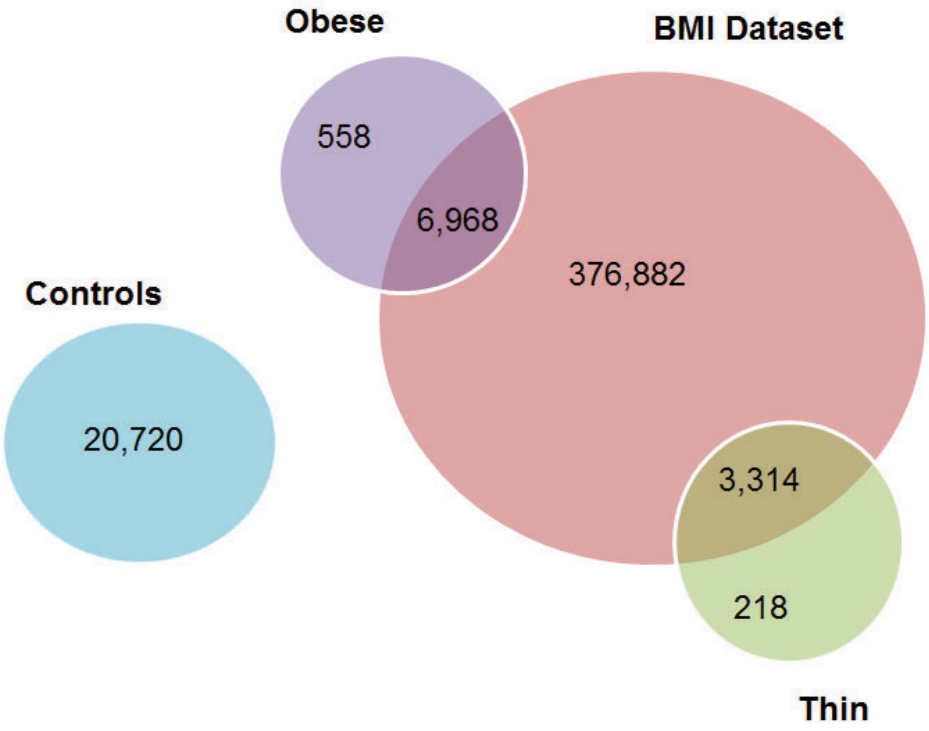


Figure 2.2: Summary of the UKBB sample sets after QC. Venn Diagram showing sample numbers and overlap between UKBB sample sets used in genetic correlation (BMI dataset) and GWAS replication (obese, controls, lean) analyses.

2.3.3 Imputation and genome-wide association analyses

2.3.3.1 SCOOP, STILTS and UKHLS association analysis

Imputation and genome-wide association analyses for SCOOP, STILTS and UKHLS were performed by Vanisha Mistry. Genotypes from SCOOP, STILTS and UKHLS controls were phased together with SHAPEITv2, and subsequently imputed with IMPUTE2 [221, 222] to the merged UK10K and 1000G Phase 3 reference panel [126], containing ~91.3 million autosomal and chromosome X sites, from 6,285 samples. More than 98% of variants with MAF ≥0.5% had an imputation quality score of r^2≥0.4, however variants with MAF <0.1% had a poor imputation quality with only 27% variants with r^2≥0.4. First-pass single-variant

association tests were done for all variants irrespective of MAF, or imputation quality score (see below). Analyses of 1,456 SCOOP, 1,471 STILTS and 6,460 controls (BMI range 19-30) of European ancestry were based on the frequentist association test, using the EM algorithm, as implemented in SNPTEST v2.5 [226], under an additive model and adjusting for six PCs and sex as covariates.

2.3.3.2 UKBB BMI dataset single-variant association analysis

For the BMI dataset, I used BOLT-LMM [227] to perform an association analysis with BMI using sex, age, 10 PCs and UKBB genotyping array as covariates.

2.3.4 Heritability estimates and genetic correlation

Summary statistics from the SCOOP vs. UKHLS, STILTS vs. UKHLS, UKBB obese vs controls, UKBB thin vs controls and UKBB BMI analyses were filtered and a subset of 1,197,969 of the 1,217,312 HapMap3 SNPs was kept in each dataset since HapMap3 reference panel markers are common and normally well-imputed variants. Using LD score regression [228] I first calculated the heritability of severe childhood obesity (SCOOP vs UKHLS) and persistent thinness (STILTS vs UKHLS). For severe childhood obesity, I estimated a prevalence of 0.15% using the BMI centile equivalent to 3SDS in children [229]. In the case of persistent thinness (BMI<=19), I used a General Practice (GP) based cohort for our prevalence estimates: CALIBER [230]. The CALIBER database consists of 1,173,863 records derived from GP practices. For the heritability analysis, I used a prevalence estimate of 2.8% for BMI<=19 (Claudia Langenberg and Harry Hemingway, personal communication). I also used LD score regression to calculate the genetic correlation of SCOOP with STILTS, SCOOP with BMI and

STILTS with BMI. The genetic correlation between obesity and persistent thinness with anorexia was estimated using the summary statistics from SCOOP vs UKHLS and STILTS vs. UKHLS, and summary statistics available from the Genetic Consortium for Anorexia Nervosa (GCAN) in LD Hub [231]. The same analysis was repeated for UKBB obese vs controls and UKBB thin vs controls. Genetic correlation estimates for BMI vs Overweight, Obesity Class 1, Obesity Class 2 and Obesity Class 3 were also extracted from LD Hub (http://ldsc.broadinstitute.org/ldhub/).

2.3.5 Comparison with established GIANT BMI associated loci

I obtained the list of 97 established BMI associated loci from the latest publicly available data from the GIANT consortium at the time of this study [92]. I used this list as I wanted to focus on established common variation in Europeans with accurate effect sizes. In order to test whether there was evidence of enrichment of nominally significant signals with consistent direction of effect, I performed a binomial test using the subset of signals with nominal significance in the SCOOP vs UKHLS, and STILTS vs UKHLS analyses. Variance explained was calculated using the rms package [232] v4.5.0 in R [233] and Nagelkerke's R^2 is reported. Power calculations were performed using Quanto [234].

2.3.6 Analysis of potential age effects in SCOOP

To investigate if differences in the observed OR from our SCOOP vs UKHLS analysis were influenced by age differences between cases (SCOOP, mean age ~ 11) and controls (UKHLS,

mean age ~52), I obtained BMI summary statistics from Nicholas Timpson and Laura Corbin for the ALSPAC cohort. To calculate ORs and SE from the ALSPAC BMI summary statistics I used genotype counts from SNPTEST output. I then used a z-test to test for significant differences between the OR calculated using genotype counts of SCOOP and ALSPAC against the SCOOP vs. UKHLS OR.

2.3.7 Simulations under an additive model

I created 10,000 simulations of 1 million individuals for the 97 GIANT BMI loci randomly sampling alleles based on the allele frequency from UKHLS using an R script. For each simulated genotype, phenotypes were simulated with DISSECT [235] using the effect size in GIANT and then removed all samples from the lower tail where the phenotype was <3SDS to better reproduce the actual BMI distribution. Afterwards I randomly sampled 1,471 individuals from the bottom 1.6% and 1,456 from top 0.15% and compared against a random set of 6,460 controls from the equivalent percentiles to BMI 19-30 in UKHLS. Finally, for each of these loci, I calculated the absolute difference between our observed OR and the mean OR from the simulations and counted how many times an equal or larger absolute difference in the simulated data was observed and assigned a p-value. This was done separately for SCOOP vs UKHLS and STILTS vs UKHLS. The high accuracy of the 97 GIANT BMI loci allowed me to assess significant differences between the observed and expected ORs.

2.3.8 Genetic Risk Score

For this analysis, Vanisha Mistry calculated the GRS scores, Audrey Hendricks performed ordinal regression statistical analyses and I compared BMI category GRS scores with simulations. The R package GTX (https://CRAN.R-project.org/package=gtx) was used to transpose genotype probabilities into dosages, and a combined dosage score, weighted by the effect size from GIANT, for 97 BMI SNPs [92] was calculated and standardised. An ordinal relationship between the genetic risk score and BMI category (i.e. thin, normal, or obese) was checked using ordinal logistic regression with the clm function in the ordinal R package. For each of the 10,000 simulations, a genetic risk score was created and an ordinal logistic regression was run. Audrey compared the observed test statistic testing whether the odds were the same by BMI category to the 10,000 simulation test statistics. Audrey calculated the p-value as the number of simulations with a test statistic larger than that observed in the real data. I also calculated a mean genetic risk score for each BMI category (obese, thin and controls) across the 10,000 simulations. I used a t-test to test whether the mean observed GRS score in each category was significantly different from the one estimated using the simulations.

2.3.9 Discovery stage GWAS

First pass single-variant association analyses results were used as discovery datasets for the GWAS. After association analysis performed by Vanisha Mistry, I removed variants with MAF<0.5%, an INFO score <0.4, and HWE $p<1x10^{-6}$, as these highlighted regions of the genome that were problematic, including CNV regions with poor imputation quality.

Quantile-quantile plots indicated that the genomic inflation was well controlled for in SCOOP-UKHLS (λ=1.06) and STILTS-UKHLS (λ =1.04), and slightly higher for SCOOP-STILTS (λ =1.08). I used LD score regression [228] to correct for inflation not due to polygenicity. To identify distinct loci, I performed clumping as implemented in PLINK [223] using summary statistics from the association tests and LD information from the imputed data, clumping variants 250kb away from an index variant and with an r^2>0.1. In order to further identify a set of likely independent signals I performed conditional analysis of the lead SNPs in SNPTEST to take into account long-range LD. A total of 135 autosomal variants with p<$1x10^{-5}$ in any of the three case-control analyses were taken forward for replication in UKBB. All case-control results are reported with the lower BMI group as reference.

2.3.10 UKBB association analysis

I tested 72,355,667 SNPs for association under an additive model in SNPTEST using sex, age, 10 PCs and UKBB genotyping array as covariates. Three comparisons were done: obese vs thin, obese vs controls and controls vs thin. Variants with an INFO score <0.4, HWE *p*<$1x10^{-6}$ were filtered out from the results. Inflation factors were calculated for variants with MAF>0.5%. Inflation factors were calculated using HapMap3 reference panel markers. The LD score regression intercepts were 1.0074 in obese vs thin, 1.0057 in obese vs controls and 1.009 in thin vs controls. I used all thin individuals, regardless of health status, as a replication cohort to maximize power.

2.3.11 GIANT, EGG and SCOOP 2013 summary statistics

Summary statistics for the GIANT Extremes obesity meta-analysis [207] were obtained from http://portals.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium_data_files. Summary statistics for EGG [236] were obtained from http://egg-consortium.org/childhood-obesity.html. I used summary statistics from our group's previous study of 1,509 early-onset obesity SCOOP cases compared to 5,380 publicly available WTCCC2 controls (SCOOP 2013) [160]. Data for the SCOOP cases is available to download from the European Genome-Phenome Archive (EGA) using accession number EGAD00010000594. The control samples are available to download using accession numbers EGAD00000000021 and EGAD00000000023. These replication studies are largely non-overlapping with our discovery datasets and each-other. When a lead variant was not available in a replication cohort, a proxy ($r^2 \geq 0.8$) was used in the meta-analysis.

2.3.12 Replication meta-analysis

I meta-analysed summary statistics for the 135 variants reaching $p<1x10^{-5}$ in SCOOP vs STILTS, SCOOP vs UKHLS, and UKHLS vs STILTS with the corresponding results from UKBB and study specific replication cohorts. For obese vs. thin and obese vs. controls comparisons I used fixed-effects meta-analysis correcting for unknown sample overlap in replication cohorts using METACARPA [237]. For thin vs. controls I used a fixed-effects meta-analysis in METAL [238]. Heterogeneity was assessed using Cochran's Q-test heterogeneity p-value in METAL. A signal was considered to replicate if it met all of the following criteria: i) consistent direction of effect; ii) $p<0.05$ in at least one replication cohort; and iii) the meta-analysis p-value reached standard genome-wide significance ($p<5x10^{-8}$). Application of a more

stringent p-value cutoff of $p \leq 1.17 \times 10^{-8}$ which would take into account the additional variants on the lower allele frequency spectrum (and hence increased number of independent tests) [239] only affected one previously established signal (*SULT1A1*, rs3760091, $p=2.65 \times 10^{-8}$) in the obese vs. controls analysis that fell just above this threshold (**Table 2.6**). rs4440960 was later removed from final results (SCOOP vs UKHLS and STILTS vs UKHLS) after close examination revealed it was present in a CNV region with poor imputation quality.

2.3.13 Comparison of newly established candidate loci and UKBB independent BMI dataset

To evaluate whether the number of associated signals in SCOOP vs STILTS, SCOOP vs UKHLS and UKHLS vs STILTS that were directionally consistent and nominally significant in the independent UKBB BMI analysis were more than expected by chance, I performed a binomial test (**Table 2.9**).

2.3.14 Lookup of previously identified obesity-related signals in our discovery datasets

I took all signals reaching genome-wide significance, or identified for the first time in the GIANT Extremes obesity meta-analysis [207], with either the tails of BMI or obesity classes, and in childhood obesity studies [160, 236] and performed look-up of those signals in all three of our discovery analyses (SCOOP vs STILTS, SCOOP vs UKHLS and UKHLS vs STILTS) (**Supplementary Table 10 of Riveros-Mckay et al 2018 [217] (Appendix A))**.

2.4 Results

2.4.1 Discovery cohorts characteristics

The discovery cohorts consisted of genotype data for 1,622 persistently thin healthy individuals (STILTS), 1,985 severe childhood onset obesity cases (SCOOP) and 10,433 population based individuals (UKHLS) used as a common set of control. A summary of cohort characteristics is presented in **Table 2.2**. I tested for significant differences between discovery cohorts that could affect interpretation of association results. Using a Fisher's test I determined that there's a significant sex difference (p<0.001) in STILTS vs SCOOP and UKHLS reflecting a low prevalence of thinness in men as defined by our BMI threshold. I also found significant differences in the ages of all cohorts using a t-test (p<0.001). This difference was partly by design in SCOOP since ascertainment based on young age was done deliberately to minimize time of exposure to Western obesogenic environments. After sample and variant quality control, I retained 1,471 thin individuals, 1,456 obese individuals, 6,460 control individuals in the BMI range 19-30 kg/m2 (non-extremes).

	STILTS (thin)		SCOOP (obese)		UKHLS (controls)	
N total	1622		1985		10433	
	Female	Male	Female	Male	Female	Male
N	1325 (81.69%)*	297 (18.31%)*	1083 (54.56%)	902 (45.44%)	5837 (55.95%)	4596 (44.05%)
Age**	36.64 ± 14.33 (mean ± SD)	35.17 ± 14.50 (mean ± SD)	10.74 ± 7.44 (mean ± SD)	10.93 ± 7.09 (mean ± SD)	52.02 ± 16.73 (mean ± SD)	52.67 ± 17.31 (mean ± SD)
BMI	17.56 ± 0.93 (mean ± SD)	17.56 ± 1.06 (mean ± SD)	33.66 ± 9.47 (mean ± SD)	34.41 ± 10.61 (mean ± SD)	26.98 ± 7.94 (mean ± SD)	26.86 ± 7.83 (mean ± SD)
BMI sds (children)			3.70 ± 0.83 (mean ± SD)	3.83 ± 0.87 (mean ± SD)		

Table 2.2:Summary of discovery sample sets before QC. *Significantly different in STILTS compared to SCOOP and UKHLS p<0.001. **Significantly different across all cohorts p<0.001.

2.4.2 Heritability of persistent thinness and severe early onset obesity

In my first analysis I contrasted the heritability of thinness to that of severe early onset childhood obesity. To this end genotypes for SCOOP, STILTS and UKHLS were imputed together to a combined UK10K+1000G reference panel by Vanisha Mistry and logistic regression results from SNPTEST for SCOOP vs UKHLS and STILTS vs UKHLS were used. I used LD score regression to estimate heritability explained by common variation (MAF >5%) using a subset of 1,197,969 HapMap3 markers (**Methods 2.3.4**). Using prevalence estimates previously described (**Methods 2.3.4**), I estimated that common variation accounted for 32.33% (95% CI 23.75%-40.91%) of the phenotypic variance on the liability scale in severe early onset obesity, and 28.07% (95% CI 13.80%-42.34%) in persistent thinness, suggesting both traits are similarly heritable.

2.4.3 Contribution of known BMI associated loci to thinness and severe early onset obesity

To investigate the role of common variant European BMI-associated loci in persistent thinness vs severe early onset obesity, I focused on the 97 loci from GIANT [92] available at the start of this study. Three-way association analyses were performed by Vanisha Mistry: SCCOP vs. STILTS, SCOOP vs UKHLS, UKHLS vs. STILTS (**Methods 2.3.3.1**). After quality control, 41,266,535 variants remained for association analyses in the three cohorts: SCOOP vs STILTS, SCOOP vs UKHLS and UKHLS vs STILTS.

Of these 97 established BMI associated loci, I found that 40 were nominally significant (p<0.05) in SCOOP vs UKHLS and 15 in UKHLS vs STILTS (**Table 2.3, Supplementary Table 2 of Riveros-Mckay et al 2018 [217] (Appendix A)**). Direction of effect was consistent for all of

these loci, which was more than expected by chance (binomial p=9.09x10^{-13} and binomial p=3.05x10^{-5}, respectively). Overall, the proportion of phenotypic variance explained by the 97 established BMI associated loci was 10.67% in SCOOP vs UKHLS, and 4.33% in STILTS vs UKHLS (**Methods 2.3.5**). However, evaluation of association results in thin (STILTS) and obese (SCOOP) individuals, compared to the same controls (UKHLS), highlighted that the results are not a mirror image of each other (**Figure 2.3**).

rsID	Gene	GIANT				SCOOP vs. UKHLS			UKHLS vs. STILTS		
		EA	EAF	Beta	P value	EAF	OR	P value	EAF	OR	P value
rs1558902	*FTO*	A	0.41	0.08	7.5X10^{-153}	0.41	**1.42**	**1.25X10^{-17}**	0.38	**1.17**	**2.78X10^{-4}**
rs6567160	*MC4R*	C	0.23	0.05	3.93X10^{-53}	0.24	**1.30**	**7.91X10^{-9}**	0.22	**1.25**	**1.38X10^{-5}**
rs13021737	*TMEM18*	G	0.82	0.06	1.11X10^{-50}	0.83	**1.35**	**3.89X10^{-7}**	0.82	**1.21**	**4.44X10^{-4}**
rs10938397	*GNPDA2*	G	0.43	0.04	3.21X10^{-38}	0.44	**1.18**	**4.50X10^{-5}**	0.42	1.08	6.24X10^{-2}
rs543874	*SEC16B*	G	0.19	0.04	2.62X10^{-35}	0.21	**1.20**	**2.22X10^{-4}**	0.20	**1.17**	**3.11X10^{-3}**
rs2207139	*TFAP2B*	G	0.17	0.04	4.13X10^{-29}	0.17	**1.17**	**2.70X10^{-3}**	0.16	1.11	6.21X10^{-2}
rs11030104	*BDNF*	A	0.79	0.04	5.56X10^{-28}	0.79	**1.14**	**1.27X10^{-2}**	0.79	**1.12**	**2.43X10^{-2}**
rs3101336	*NEGR1*	C	0.61	0.03	2.66X10^{-26}	0.60	**1.19**	**3.66X10^{-5}**	0.59	1.05	2.07X10^{-1}
rs7138803	*BCDIN3D*	A	0.38	0.03	8.15X10^{-24}	0.37	**1.21**	**4.68X10^{-6}**	0.36	1.03	4.47X10^{-1}
rs10182181	*ADCY3*	G	0.46	0.03	8.78X10^{-24}	0.49	**1.20**	**9.30X10^{-6}**	0.48	**1.18**	**6.81X10^{-5}**
rs3888190	*ATP2A1*	A	0.40	0.03	3.14X10^{-23}	0.40	**1.12**	**3.87X10^{-3}**	0.39	1.03	4.34X10^{-1}
rs1516725	*ETV5*	C	0.87	0.04	1.89X10^{-22}	0.86	**1.15**	**1.89X10^{-2}**	0.85	**1.18**	**5.03X10^{-3}**
rs12446632	*GPRC5B*	G	0.86	0.04	1.48X10^{-18}	0.85	1.09	1.24X10^{-1}	0.85	**1.19**	**2.20X10^{-3}**
rs16951275	*MAP2K5*	T	0.78	0.03	1.91X10^{-17}	0.77	**1.13**	**1.43X10^{-2}**	0.77	1.05	2.80X10^{-1}
rs3817334	*MTCH2*	T	0.40	0.02	5.15X10^{-17}	0.41	**1.09**	**3.52X10^{-2}**	0.40	**1.09**	**3.29X10^{-2}**
rs12566985	*FPGT-TNNI3K*	G	0.44	0.02	3.28X10^{-15}	0.43	**1.20**	**1.04X10^{-5}**	0.42	1.03	3.96X10^{-1}
rs3810291	*ZC3H4*	A	0.66	0.02	4.81X10^{-15}	0.67	**1.13**	**4.69X10^{-3}**	0.66	1.07	1.15X10^{-1}
rs7141420	*NRXN3*	T	0.52	0.02	1.23X10^{-14}	0.51	**1.11**	**1.11X10^{-2}**	0.50	1.00	9.48X10^{-1}
rs13078960	*CADM2*	G	0.19	0.03	1.74X10^{-14}	0.20	0.99	9.08X10^{-1}	0.20	**1.19**	**9.49X10^{-4}**
rs17024393	*GNAT2*	C	0.04	0.06	7.03X10^{-14}	0.02	**1.56**	**1.26X10^{-4}**	0.02	1.09	5.20X10^{-1}
rs13107325	*SLC39A8*	T	0.07	0.04	1.83X10^{-12}	0.08	**1.28**	**4.84X10^{-4}**	0.07	**1.20**	**2.89X10^{-2}**
rs17405819	*HNF4G*	T	0.70	0.02	2.07X10^{-11}	0.70	**1.12**	**1.19X10^{-2}**	0.69	1.08	6.30X10^{-2}
rs2365389	*FHIT*	C	0.58	0.02	1.63X10^{-10}	0.59	**1.09**	**3.94X10^{-2}**	0.58	1.06	1.80X10^{-1}
rs205262	*C6orf106*	G	0.27	0.02	1.75X10^{-10}	0.26	**1.16**	**1.14X10^{-3}**	0.26	1.05	3.12X10^{-1}
rs2820292	*NAV1*	C	0.55	0.02	1.83X10^{-10}	0.56	1.03	4.74X10^{-1}	0.56	**1.09**	**3.47X10^{-2}**
rs9641123	*CALCR*	C	0.42	0.01	2.08X10^{-10}	0.41	**1.09**	**3.19X10^{-2}**	0.40	1.03	4.09X10^{-1}

rsID	Gene	GIANT				SCOOP vs. UKHLS			UKHLS vs. STILTS		
		EA	EAF	Beta	P value	EAF	OR	P value	EAF	OR	P value
rs12016871	*MTIF3*	T	0.20	0.03	$2.29X10^{-10}$	0.17	**1.15**	**$7.09X10^{-3}$**	0.17	0.96	$4.84X10^{-1}$
rs16851483	*RASA2*	T	0.06	0.04	$3.55X10^{-10}$	0.06	**1.20**	**$2.17X10^{-2}$**	0.06	1.17	$8.83X10^{-2}$
rs1928295	*TLR4*	T	0.54	0.01	$7.91X10^{-10}$	0.56	**1.10**	**$2.00X10^{-2}$**	0.56	0.99	$8.13X10^{-1}$
rs2650492	*SBK1*	A	0.30	0.02	$1.92X10^{-9}$	0.29	**1.17**	**$2.93X10^{-4}$**	0.29	1.05	$2.42X10^{-1}$
rs12940622	*RPTOR*	G	0.57	0.01	$2.49X10^{-9}$	0.55	**1.12**	**$7.20X10^{-3}$**	0.55	1.06	$1.28X10^{-1}$
rs11847697	*PRKD1*	T	0.04	0.04	$3.99X10^{-9}$	0.04	**1.25**	**$1.72X10^{-2}$**	0.04	1.24	$5.05X10^{-2}$
rs4740619	*C9orf93*	T	0.54	0.01	$4.56X10^{-9}$	0.54	1.05	$2.10X10^{-1}$	0.54	**1.12**	**$5.88X10^{-3}$**
rs11191560	*NT5C2*	C	0.08	0.03	$8.45X10^{-9}$	0.07	**1.23**	**$4.23X10^{-3}$**	0.07	1.00	$9.98X10^{-1}$
rs1000940	*RABEP1*	G	0.32	0.01	$1.28X10^{-8}$	0.30	**1.11**	**$1.47X10^{-2}$**	0.29	1.06	$2.04X10^{-1}$
rs2836754	*ETS2*	C	0.61	0.01	$1.61X10^{-8}$	0.65	1.05	$2.42X10^{-1}$	0.64	**1.12**	**$1.03X10^{-2}$**
rs9400239	*FOXO3*	C	0.68	0.01	$1.61X10^{-8}$	0.70	**1.11**	**$1.84X10^{-2}$**	0.70	**1.09**	**$4.31X10^{-2}$**
rs29941	*KCTD15*	G	0.66	0.01	$2.41X10^{-8}$	0.67	**1.13**	**$5.77X10^{-3}$**	0.66	1.02	$5.56X10^{-1}$
rs9374842	*LOC285762*	T	0.74	0.01	$2.67X10^{-8}$	0.77	**1.16**	**$3.41X10^{-3}$**	0.76	1.05	$2.53X10^{-1}$
rs6477694	*EPB41L4B*	C	0.36	0.01	$2.67X10^{-8}$	0.35	**1.10**	**$2.73X10^{-2}$**	0.34	1.04	$3.53X10^{-1}$
rs7899106	*GRID1*	G	0.05	0.04	$2.96X10^{-8}$	0.05	**1.24**	**$1.48X10^{-2}$**	0.05	0.94	$5.90X10^{-1}$
rs2245368	*PMS2L11*	C	0.18	0.03	$3.19X10^{-8}$	0.16	**1.22**	**$2.73X10^{-4}$**	0.16	0.98	$7.82X10^{-1}$
rs17203016	*CREB1*	G	0.19	0.02	$3.41X10^{-8}$	0.20	**1.13**	**$1.32X10^{-2}$**	0.20	0.98	$7.28X10^{-1}$
rs17724992	*PGPEP1*	A	0.74	0.01	$3.42X10^{-8}$	0.74	**1.15**	**$2.99X10^{-3}$**	0.73	1.04	$3.90X10^{-1}$
rs9540493	*MIR548X2*	A	0.45	0.01	$4.97X10^{-8}$	0.45	**1.12**	**$9.92X10^{-3}$**	0.44	1.00	$9.28X10^{-1}$

Table 2.3: BMI-associated loci that were nominally significant in either. SCOOP vs UKHLS or UKHLS vs STILTS.EA= Effect allele (BMI increasing allele); EAF = Effect allele frequency. Only loci that are nominally significant ($p<0.05$) in at least one comparison are shown. Nominally significant loci ($p<0.05$) are highlighted in bold for each comparison

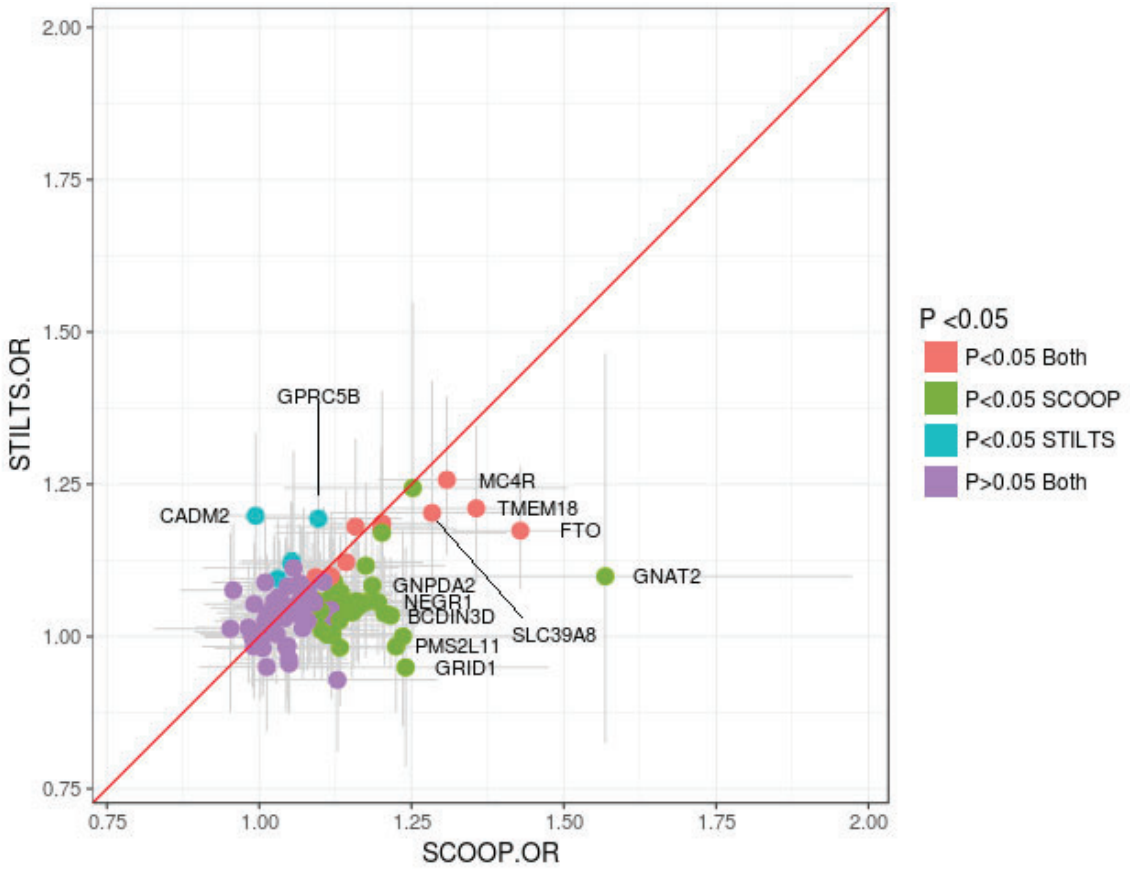


Figure 2.3: Odds ratio comparison for the 97 BMI associated loci. Odds ratios for SCOOP vs UKHLS (x-axis) and UKHLS vs STILTS (y-axis) comparisons are shown for the 97 known BMI loci from GIANT. Colours of data points represent nominal significance in both analyses (red), only SCOOP vs. UKHLS (green), only STILTS vs UKHLS (blue) or in neither analysis (purple). Error bars represent 95% confidence intervals for the odds ratios for SCOOP vs UKHLS (x-axis) and for UKHLS vs STILTS (y-axis). A subset of data points with larger separation from the red diagonal line (x=y) are labelled.

Notably, a striking difference was observed in association results in the *FTO* locus where the lead intronic obesity risk variant, rs1558902, showed a moderate effect size and modest evidence of association in controls compared to thin individuals (UKHLS vs STILTS)(p=0.00027, OR=1.17, 95% CI [1.08,1.28], EAF=0.39), despite having a large effect and being associated at genome-wide significance levels in obese compared to control individuals (SCOOP vs UKHLS) (p=1.25x10^{-17}, OR=1.43, 95% CI [1.32,1.55], EAF=0.41) (**Figure 2.3, Table 2.3**). *GNAT2* also showed a larger effect and significance in the analysis of SCOOP vs UKHLS (p=1.26x10^{-4}, OR=1.57, 95% CI [1.25, 1.97], EAF=0.03), than in UKHLS vs STILTS (p=0.52, OR=1.10, 95% CI [0.82, 1.47], EAF=0.02) (**Figure 2.3, Table 2.3**). This discrepancy in

association strength and effect size was also seen at the opposite end of the BMI spectrum in *CADM2* where the lead SNP, rs13078960, showed evidence of association in UKHLS vs STILTS (p= 9.48x10^{-4}, OR=1.2, 95% CI [1.08, 1.33], EAF=0.20) but no association in SCOOP vs UKHLS (p>0.05). In contrast to results at the *FTO* and *CADM2* loci, for *MC4R* the results are more comparable, with genome-wide significant association in SCOOP vs UKHLS (rs6567160, p=7.91x10^{-9}, OR=1.31, 95% CI [1.19, 1.43], EAF=0.25) and highly significant association results in UKHLS vs STILTS(p=1.38x10^{-5}, OR=1.26, 95% CI [1.13, 1.39], EAF=0.23)**.** One possible explanation for these observed differences is that they arose as a result of randomly sampling a small subset of individuals at the two extremes of the distribution and/or by the different degree of extremeness of the phenotype. To formally test if these results were significantly different from those expected under a model where loci act additively across the BMI distribution, I simulated 10,000 different populations of 1 million individuals with genotypes for the 97 established BMI loci using allele frequencies in UKHLS, and then simulated a phenotype using the effect sizes in GIANT (**Methods 2.3.7**). These simulations detected fourteen loci with nominally significant deviation from an additive model, however none remained significant after correction for the number of tests (p=0.05/97*2 = ~0.0002, **Table 2.4**). However, *CADM2* was nominally significant in both SCOOP vs UKHLS and STILTS vs UKHLS analyses, with slightly lower OR detected in SCOOP vs UKHLS compared to simulated data, and slightly higher OR detected in UKHLS vs STILTS compared to simulated data (**Table 2.4**). Since both SCOOP and STILTS are significantly younger than UKHLS, I used summary statistics from the ALSPAC cohort which consists of 4,964 children aged 13-16 to test if the OR differences observed in SCOOP vs UKHLS were due to age effects. For the 97 GIANT loci overall there were no significant differences (z-test, p>0.05) except for rs2245368 (*PMS2L11* locus, z-test p=3.81x10^{5}, **Supplementary Table 4 of**

Riveros-Mckay et al 2018 [217] (Appendix A)). In combination, these results suggest that the observed differences in ORs and p-values could have arisen because our severe obese cases are much more extreme (i.e. deviate more from the mean) than the healthy thin individuals. Results also suggest our obese and thin sample sizes gave us limited power to detect significant differences compared to the additive model given the wide confidence intervals observed in simulations.

SCOOP			
Gene	p-val	observed OR	mean simulation OR
QPCTL	0.0471	1.02	1.14
FPGT-TNNI3K	0.0161	1.21	1.09
CADM2	0.0177	0.99	1.12
STXBP6	0.0379	0.99	1.09
HSD17B12	0.0113	0.96	1.08
ZBTB10	0.0166	0.95	1.14
STILTS			
Gene	p-val	observed OR	mean simulation OR
MC4R	0.0137	1.26	1.12
ADCY3	0.0059	1.19	1.06
CADM2	0.0148	1.20	1.06
LINGO2	0.0436	0.96	1.05
TCF7L2	0.0337	0.96	1.05
C9orf93	0.0398	1.12	1.04
SCARB2	0.0473	0.95	1.06
ETS2	0.0479	1.12	1.03
CLIP1	0.0311	0.93	1.06

Table 2.4: Nominally significant loci for non-additive effect in extremes.

In addition to analysing established BMI loci on an individual basis, I also looked at genetic risk scores (GRS) generated from the 97 BMI associated loci from GIANT [92] to analyse the contribution of these loci as a whole. To this end, Vanisha Mistry generated weighted GRS scores and Audrey Hendricks ran an ordinal logistic regression testing the association of the GRS scores on BMI category (i.e. thin (STILTS), normal (UKHLS), obese (SCOOP)). As expected, the standardised BMI genetic risk score was strongly associated with BMI

category (weighted score p=8.59x10^{-133}). The effect of a one standard deviation increase in the standardised BMI genetic risk score was significantly larger for obese vs. (thin & normal) than for (obese & normal) vs. thin (p=7.48x10^{-11}) with odds ratio and 95% confidence intervals of 1.94 (1.83, 2.07) and 1.50 (1.42, 1.59), respectively. However, using the simulations described above (**Methods 2.3.7**), confirmed that the larger OR for obese vs. (thin & normal) was not significantly different (p=0.41) than what we would expect given an additive genetic model, and the different degrees of "extremeness" in our thin and obese cases. A BMI genetic score excluding the *FTO* variant produced similar results (data not shown). I also tested whether the mean GRS in each BMI category was significantly different from that predicted via simulations and found no significant difference (**Figure 2.4**). As a sanity check, I also compared controls against simulations and no significant difference was observed (p=0.18).

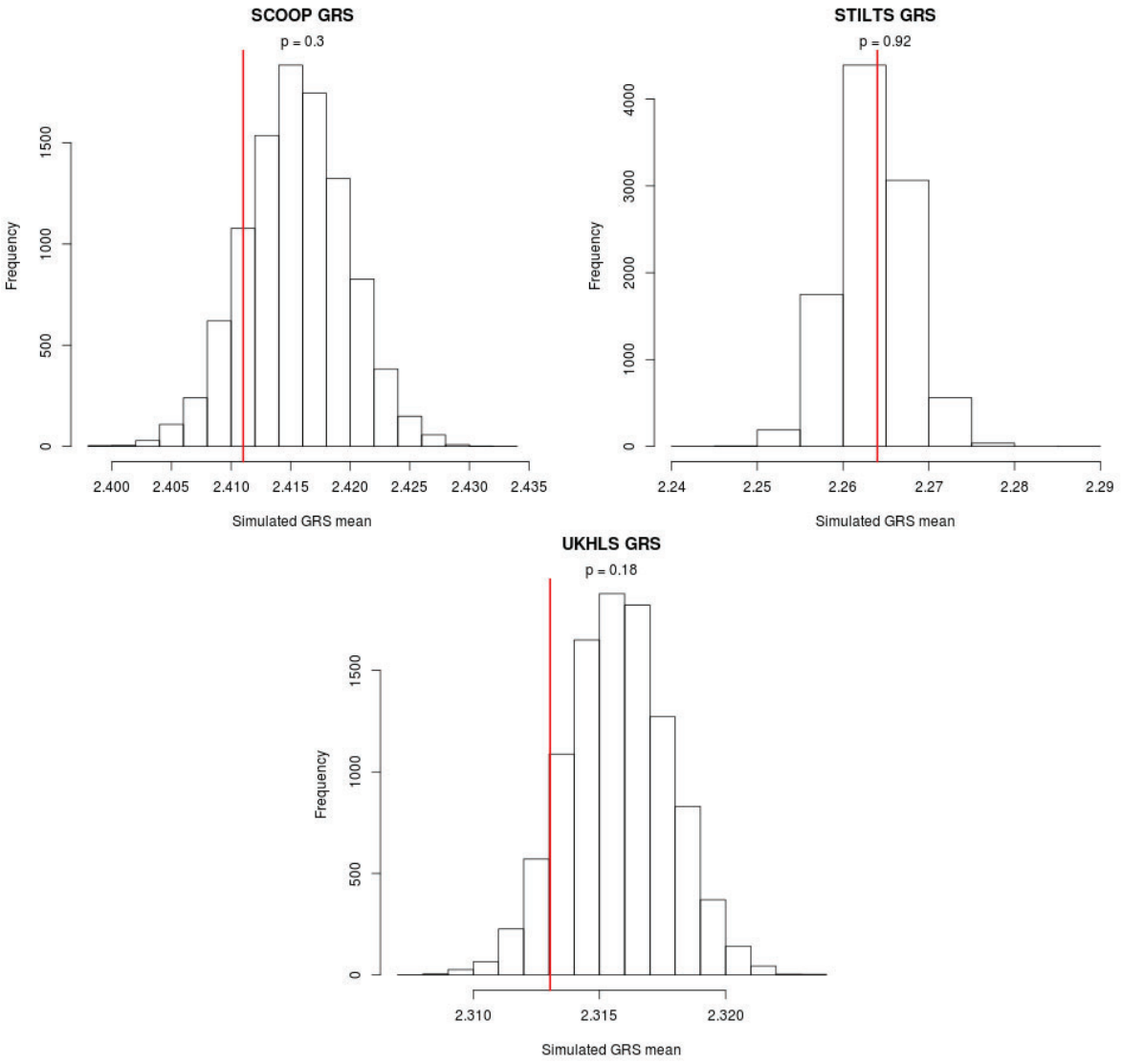


Figure 2.4: Mean GRS for SCOOP, STILTS and UKHLS compared to simulations. Histogram represents mean GRS scores for each BMI category across 10,000 simulations. Vertical red line highlights the observed value in real data.

2.4.4 Genetic correlation between persistent thinness, severe early onset childhood obesity and BMI

Given the observed differences in association results from thin (STILTS) and obese (SCOOP) individuals, compared to the same set of control individuals (UKHLS), I next explored the genetic correlation of severe early onset obesity, persistent thinness and BMI using LD score

regression (**Methods 2.3.4**). For this, I used summary statistics from the SCOOP vs UKHLS, STILTS vs UKHLS and BMI data from participants in UK Biobank (UKBB). As expected from the association results, the genetic correlation of severe early onset obesity and BMI was high (r=0.86, 95% CI [0.74, 0.98], p=1.86x10^{-43}). I also detected weaker negative correlation between persistent thinness and BMI (r=-0.63, 95% CI [-0.44,-0.82], p=3.54x10^{-11}), and between persistent thinness and severe obesity (r=-0.49, 95% CI [-0.17,-0.82], p=0.003). In contrast with previously described obesity classes, severe early onset obesity and persistent thinness were not completely correlated with BMI (**Figure 2.5**). As an inverse genetic correlation between BMI, obesity and anorexia nervosa (a disorder that is characterised by thinness and complex behavioural manifestations) has been reported [228], I also tested for genetic correlation with anorexia nervosa, and found that neither severe early onset obesity, nor persistent thinness, were significantly correlated with anorexia nervosa (r=-0.05, 95% CI [-0.15,0.05], p=0.33 and r=0.13, 95% CI [-0.02,0.28], p=0.09, respectively).

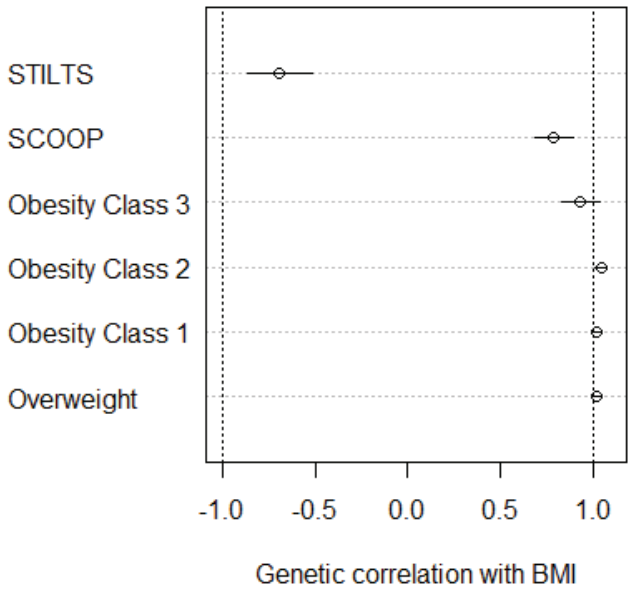


Figure 2.5: Genetic correlation of traits and BMI. Genetic correlation estimates and 95% CI for severe early-onset childhood obesity (SCOOP), healthy persistent thinness (STILTS), Obesity Class 3, Obesity Class 2, Obesity Class 1 and Overweight. Dotted lines represent complete genetic correlation.

2.4.5 Discovery of novel association signals for persistent thinness and severe early onset obesity

After the initial association analysis, I sought evidence for novel signals associated with either end of the BMI distribution (persistent thinness or severe early onset obesity; **Methods 2.3.9**). In all three analyses, in addition to loci mapping to established BMI and obesity loci, I identified *PIGZ* and *C3orf38*, two novel loci in the thin vs control analysis, that reached conventional genome-wide significance (GWS) ($p \leq 5 \times 10^{-8}$) (**Table 2.5, Figure 2.6**). However, an additional 125 SNPs, in 118 distinct loci, reached the arbitrary threshold of $p \leq 10^{-5}$ in at least one analysis, for which I sought replication to assess if promising signals are true signals or likely false-positives that could have arisen by confounding effects such as

genotyping batch effects (**Supplementary Tables 5-7 of Riveros-Mckay et al 2018 [217] (Appendix A)**).

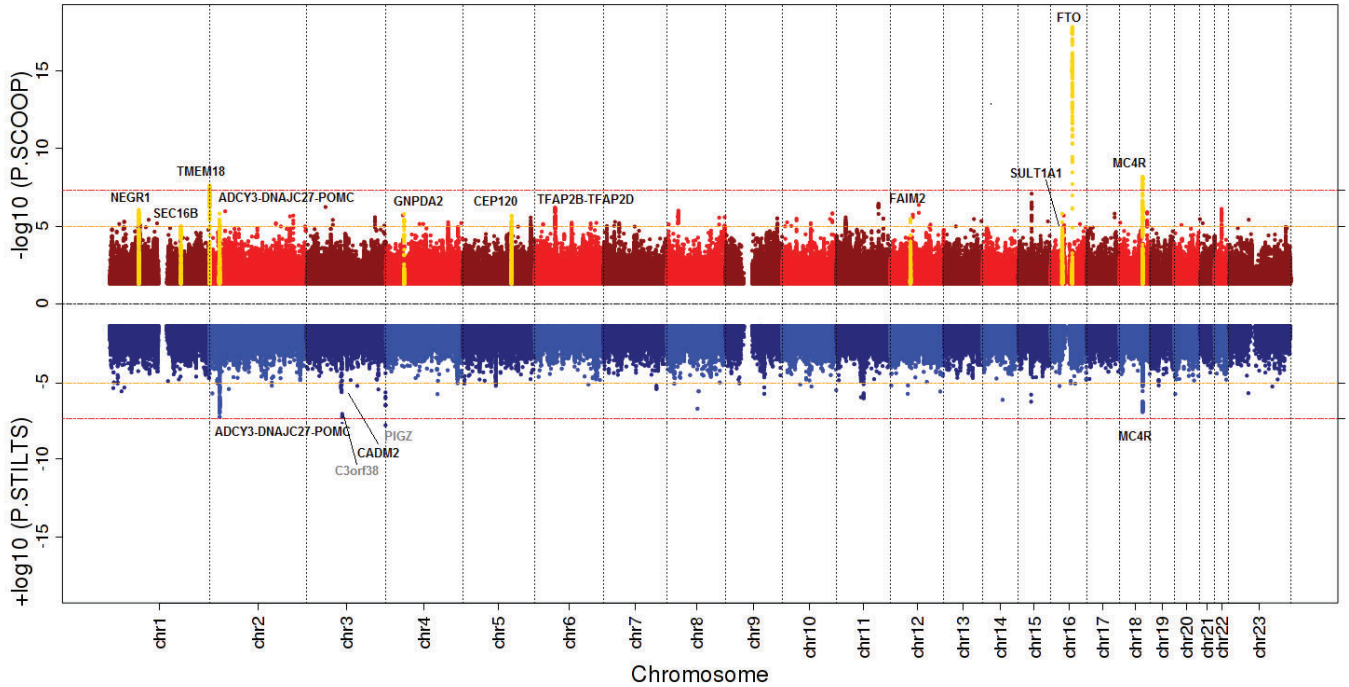


Figure 2.6: Miami plot of SCOOP vs. UKHLS and STILTS vs. UKHLS. Miami plot produced in EasyStrata [23], Red=SCOOP vs. UKHLS; Blue=STILTS vs. UKHLS. Red lines indicate genome-wide significance threshold at $p=5x10^{-8}$. Orange lines indicate discovery significance threshold at $p=1x10^{-5}$. Black labels highlight known BMI/obesity loci that were taken forward for replication and yellow peaks indicate those that met genome-wide significance after replication. Grey labels highlight novel loci that did not replicate.

Obese vs. thin							
rsID	**Nearest gene**	**EA**	**NEA**	**OR (95% CI)**	**P value**	**EAF Obese**	**EAF Thin**
rs9930333	*FTO*	G	T	1.70(1.52,1.90)	$2.30X10^{-20}$	49.59%	37.46%
rs2168711	*MC4R*	C	T	1.66(1.45,1.89)	$8.29X10^{-14}$	28.90%	19.95%
rs6748821	*TMEM18*	G	A	1.65(1.42,1.91)	$9.45X10^{-11}$	86.69%	79.84%
rs506589	*SEC16B*	C	T	1.46(1.27,1.67)	$5.42X10^{-8}$	23.98%	18.07%
rs6738433	*ADCY3-DNAJC27*	C	G	1.43(1.28,1.60)	1.71X10-^{10}	47.31%	43.92%
rs62107261	*FAM150B*	T	C	2.37(1.75,3.20)	$2.07X10^{-8}$	96.37%	93.38%
Obese vs. controls							
rsID	**Nearest gene**	**EA**	**NEA**	**OR (95% CI)**	**P value**	**EAF Obese**	**EAF Controls**
rs9928094	*FTO*	G	A	1.44(1.33,1.57)	$1.42X10^{-18}$	49.50%	41.32%
rs35614134	*MC4R*	AC	A	1.31(1.20,1.44)	$6.27X10^{-9}$	29.01%	23.69%
rs66906321	*TMEM18*	C	T	1.40(1.24,1.57)	$2.35X10^{-8}$	85.78%	81.35%
Controls vs. thin							
rsID	**Nearest gene**	**EA**	**NEA**	**OR (95% CI)**	**P value**	**EAF Controls**	**EAF Thin**
rs117638949	*PIGZ*	T	A	3.5 (2.27,5.4)	$1.50X10^{-8}$	99.50%	98.55%
rs75937976	*C3orf38*	G	C	2.95 (2.02,4.32)	$2.43X10^{-8}$	99.20%	98.25%

Table 2.5: Genome-wide significant loci in discovery analysis. EA= Effect allele (BMI increasing allele); EAF = Effect allele frequency.

As our obese and thin cases (SCOOP and STILTS) lie at the very extreme tails of the BMI distribution, there are few comparable replication datasets. I therefore used the UKBB dataset and selected individuals at the top (BMI>=40, N=7,526) and bottom end of the distribution (BMI≤19, N=3,532) to more closely match the BMI criteria of our clinically ascertained thin and obese individuals. I used 20,720 samples from the rest of the UKBB cohort as a control set (**Methods 2.3.2.2, Figure 2.2**). As previously mentioned (**Methods 2.3.2.2**), I used all thin individuals regardless of health status in this analysis. However, using ICD10 codes and self-reported illness data **(Supplementary Tables 12-13 of Riveros-Mckay et al 2018 [217] (Appendix A))** to remove individuals who had a relevant medical diagnosis before date of attendance at UKBB recruitment centre, yielded materially equivalent results (**Figure 2.7**), so I have elected to keep the original results with all thin participants as my primary analysis. In cases where lead variants or proxies (r^2>0.8) were not, at the time of this study, available in the full UKBB genetic release I used results from the interim release

using 2,799 individuals with BMI>=40, 1,212 with BMI<=19 and 8,193 controls (**Methods 2.3.2.2**). There was a significant negative genetic correlation for the obese replication cohort with anorexia nervosa (r= -0.24, 95% CI [-0.37,-0.11], *p*=0.01) and a positive genetic correlation for the thin replication cohort (r=0.49, 95% CI [0.22-0.76] *p*=0.0003). The positive genetic correlation for the thin replication cohort was still observed after using ICD10 codes and self-reported illness data to clean the phenotype (r=0.62, 95% CI [0.30,0.92], *p*=0.0001).

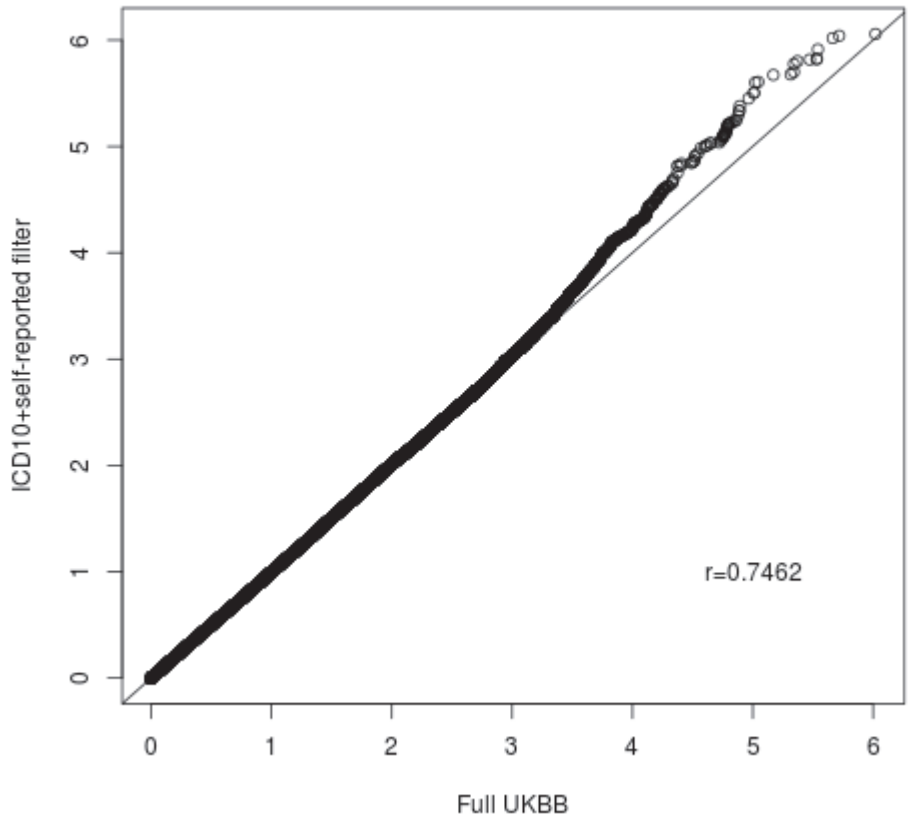


Figure 2.7: Quantile-quantile plots for UKBB case-control analysis with different exclusion criteria for thin individuals. Q-Q plot using all thin individuals as cases (Full UKBB) and removing individuals based on ICD10 and self-reported data (ICD10+self-reported filter). Correlation for –log10 p-values is shown (r=0.7462).

To further increase power, I took advantage of publicly available summary statistics from the GIANT Extremes obesity meta-analysis [207], the EGG childhood obesity study [236],

and our group's previous study on non-overlapping SCOOP participants (SCOOP 2013) [160], as additional replication datasets. For SCOOP vs. STILTS I used the GIANT BMI tails meta-analysis results [207] (up to 7,962 cases/8,106 controls from the upper/lower 5th percentiles of the BMI trait distribution). For SCOOP vs. UKHLS I used the GIANT obesity class III summary statistics [207] (up to 2,896 cases with BMI ≥40kg/m^2 vs 47,468 controls with BMI <25 kg/m^2), the EGG childhood obesity study [236] (children with BMI ≥95th percentile of BMI vs 8,318 children with BMI <50th percentile of BMI) and SCOOP 2013 [160]. Fixed effect meta-analyses yielded genome-wide significant signals at well-known BMI associated loci in both the obese vs. thin, and obese vs. control analyses, and both the *PIGZ* and *C3orf38* loci identified at the discovery stage failed to replicate when combined with additional data **(Table 2.6, Supplementary Tables 5-7 of Riveros-Mckay et al 2018 [217] (Appendix A).** However, the *SNRPC* locus described here (rs75398113), though not independent from the previously described *SNRPC*/*C6orf106* locus (rs205262, r^2= 0.29) [92], appears to be driving the previously reported association at this locus (rs205262 conditioned on rs75398113, $p_{conditioned}$=0.7, **Table 2.7**). Both SNPs are eQTLs for C6or106 and UHRF1BP1 in multiple tissues including brain and colon tissues on GTEx however neither of these are obvious biological candidates linked to energy homeostasis.

Obese vs. thin						Discovery cohort				Replication cohorts					Combined analysis		
rsID	Nearest gene	Chr	Position (bp)	EA	NEA	OR (95% CI)	P value	EAF Ob	EAF Th	Cohort	OR (95% CI)	P value	EAF Ob	EAF Th	OR (95% CI)	P value	HetPVal
rs9930333	*FTO*	16	53799977	G	T	1.70 (1.52,1.90)	$2.30X10^{-20}$	49.59%	37.46%	UKBB	1.46 (1.38,1.55)	$3.60X10^{-36}$	48.26%	38.93%	1.48 (1.42,1.54)	$8.52X10^{-76}$	$3.34X10^{-2}$
										GIANT	1.43 (1.34,1.54)	$8.10X10^{-25}$					
rs2168711	*MC4R*	18	57848531	C	T	1.66 (1.45,1.89)	$8.29X10^{-14}$	28.90%	19.95%	UKBB	1.23 (1.15,1.32)	$2.19X10^{-9}$	26.75%	22.90%	1.27 (1.21,1.33)	$2.02X10^{-21}$	$1.12X10^{-4}$
										GIANT	1.20 (1.10,1.30)	$1.80X10^{-5}$					
rs6748821	*TMEM18*[a]	2	629601	G	A	1.65 (1.42,1.91)	$9.45X10^{-11}$	86.69%	79.84%	UKBB	1.27 (1.18,1.37)	$1.31X10^{-9}$	85.00%	81.69%	1.32 (1.24,1.39)	$7.76X10^{-21}$	$2.81X10^{-3}$
										GIANT	1.26 (1.14,1.39)	$9.90X10^{-6}$					
rs506589	*SEC16B*	1	177894287	C	T	1.46 (1.27,1.67)	$5.42X10^{-8}$	23.98%	18.07%	UKBB	1.25 (1.17,1.35)	$5.44X10^{-10}$	23.11%	19.16%	1.28 (1.21,1.35)	$3.14X10^{-20}$	$1.21X10^{-1}$
										GIANT	1.25 (1.14,1.37)	$2.70X10^{-6}$					
rs6738433	*ADCY7*[b]	2	25159501	C	G	1.43 (1.28,1.60)	$1.71X10^{-10}$	47.31%	43.92%	UKBB	1.21 (1.14,1.28)	$2.74X10^{-10}$	50.70%	45.96%	1.19 (1.14,1.24)	$3.19X10^{-17}$	$6.25X10^{-3}$
										GIANT	1.10 (1.03,1.17)	$5.70X10^{-3}$					
rs7132908	*FAIM2*	12	50263148	A	G	1.31 (1.17,1.47)	$2.26X10^{-6}$	42.45%	36.27%	UKBB	1.18 (1.11,1.25)	$5.43X10^{-8}$	41.11%	37.39%	1.20 (1.15,1.26)	$1.93X10^{-16}$	$2.52X10^{-1}$
										GIANT	1.20 (1.10,1.30)	$6.60X10^{-6}$					
rs62107261	*FAM150B*	2	422144	T	C	2.37 (1.75,3.20)	$2.07X10^{-8}$	96.37%	93.38%	UKBB	1.54 (1.35,1.76)	$3.57X10^{-10}$	96.28%	94.36%	1.65 (1.46,1.87)	$1.15X10^{-15}$	$1.07X10^{-2}$
rs12507026	*GNPDA2*[c]	4	45181334	T	A	1.30 (1.17,1.46)	$3.69X10^{-6}$	47.29%	40.92%	UKBB	1.14 (1.08,1.21)	$8.76X10^{-6}$	45.30%	41.98%	1.18 (1.13,1.23)	$5.53X10^{-15}$	$4.06X10^{-2}$
										GIANT	1.20 (1.12,1.28)	$3.10X10^{-7}$					
rs75398113	*SNRPC*	6	34728071	C	A	1.53 (1.27,1.85)	$8.91X10^{-6}$	11.95%	8.04%	UKBB	1.24 (1.12,1.37)	$2.07X10^{-5}$	10.47%	8.52%	1.30 (1.19,1.42)	$5.19X10^{-9}$	$5.56X10^{-2}$
rs13135092	*SLC39A8*	4	103198082	G	A	1.58 (1.30,1.93)	$4.70X10^{-6}$	10.50%	7.24%	UKBB	1.25 (1.12,1.39)	$5.57X10^{-5}$	9.24%	7.52%	1.32 (1.20,1.45)	$1.06X10^{-8}$	$3.59X10^{-2}$

Obese vs. controls						Discovery cohort				Replication cohorts					Combined analysis		
rsID	Nearest gene	Chr	Position (bp)	EA	NEA	OR (95% CI)	P value	EAF Ob	EAF Co	Cohort	OR (95% CI)	P value	EAF Ob	EAF Co	OR (95% CI)	P value	HetPVal
rs9928094	*FTO*	16	53799905	G	A	1.44 (1.33,1.57)	$1.42X10^{-18}$	49.50%	41.32%	UKBB	1.30 (1.25,1.35)	$2.74X10^{-41}$	48.34%	41.91%	1.32 (1.29,1.36)	$5.94X10^{-101}$	$4.41X10^{-5}$
										SCOOP 2013	1.46 (1.34,1.60)	$4.88X10^{-17}$					
										EGG	1.21 (1.15,1.28)	$7.20X10^{-13}$					
										GIANT	1.43 (1.34,1.54)	$6.60X10^{-25}$					
rs35614134	*MC4R*[d]	18	57832856	AC	A	1.31 (1.20,1.44)	$6.27X10^{-9}$	29.01%	23.69%	UKBB	1.22 (1.16,1.27)	$1.25X10^{-18}$	26.72%	23.15%	1.23 (1.20,1.27)	$1.57X10^{-43}$	$3.55X10^{-1}$
										SCOOP 2013	1.32 (1.19,1.46)	$1.22X10^{-7}$					
										EGG	1.22 (1.15,1.30)	$1.27X10^{-10}$					
										GIANT	1.20 (1.10,1.30)	$1.70X10^{-5}$					
rs66906321	*TMEM18*[e]	2	630070	C	T	1.40 (1.24,1.57)	$2.35X10^{-8}$	85.78%	81.35%	UKBB	1.17 (1.11,1.24)	$3.44X10^{-9}$	84.44%	82.20%	1.25 (1.21,1.29)	$9.72X10^{-35}$	$1.33X10^{-2}$
										SCOOP 2013	1.39 (1.24,1.57)	$6.65X10^{-8}$					
										EGG	1.28 (1.19,1.37)	$5.15X10^{-12}$					
										GIANT	1.27 (1.15,1.40)	$3.40X10^{-6}$					
rs7132908	*FAIM2*[f]	12	50263148	A	G	1.22 (1.12,1.32)	$3.27X10^{-6}$	42.45%	37.82%	UKBB	1.15 (1.10,1.19)	$5.37X10^{-12}$	41.11%	37.71%	1.17 (1.14,1.21)	$2.38X10^{-31}$	$4.86X10^{-1}$
										SCOOP 2013	1.23 (1.12,1.35)	$8.89X10^{-6}$					
										EGG	1.18 (1.11,1.25)	$1.24X10^{-8}$					
										GIANT	1.20 (1.10,1.30)	$6.60X10^{-6}$					
rs2384060	*ADCY3*[g]	2	25135438	G	A	1.23 (1.13,1.34)	$1.53X10^{-6}$	43.52%	38.90%	UKBB	1.11 (1.07,1.15)	$4.89X10^{-8}$	47.67%	44.93%	1.14 (1.11,1.17)	$9.39X10^{-23}$	$1.13X10^{-1}$
										SCOOP 2013	1.09 (1.00,1.19)	$5.01XX10^{-2}$					

Obese vs. controls						Discovery cohort				Replication cohorts					Combined analysis		
rsID	Nearest gene	Chr	Position (bp)	EA	NEA	OR (95% CI)	P value	EAF Ob	EAF Co	Cohort	OR (95% CI)	P value	EAF Ob	EAF Co	OR (95% CI)	P value	HetPVal
										EGG	1.18 (1.12,1.24)	$1.02X10^{-9}$					
										GIANT	1.12 (1.04,1.19)	$1.60X10^{-3}$					
rs11209947	*NEGR1*[h]	1	72808551	A	T	1.30 (1.17,1.44)	$8.51X10^{-7}$	76.58%	72.18%	UKBB	1.11 (1.05,1.16)	$4.53X10^{-5}$	81.18%	79.76%	1.17 (1.13,1.21)	$5.17X10^{-20}$	$7.26X10^{-5}$
										SCOOP 2013	1.46 (1.30,1.63)	$2.21X10^{-10}$					
										EGG	1.13 (1.06,1.22)	$4.60X10^{-4}$					
										GIANT	1.22 (1.11,1.35)	$5.60X10^{-5}$					
rs12735657	*SEC16B*[i]	1	177809133	C	T	1.24 (1.13,1.37)	$9.72X10^{-6}$	24.26%	20.46%	UKBB	1.12 (1.07,1.17)	$1.48X10^{-6}$	22.87%	20.94%	1.15 (1.12,1.19)	$7.26X10^{-19}$	$1.79X10^{-1}$
										SCOOP 2013	1.20 (1.07,1.33)	$1.18X10^{-3}$					
										EGG	1.14 (1.06,1.21)	$1.52X10^{-4}$					
										GIANT	1.22 (1.11,1.34)	$1.80X10^{-5}$					
rs13104545	*GNPDA2*	4	45184907	A	G	1.27 (1.15,1.40)	$1.61X10^{-6}$	27.41%	23.45%	UKBB	1.07 (1.02,1.11)	$5.35X10^{-3}$	24.36%	23.26%	1.13 (1.09,1.17)	$1.47X10^{-11}$	$9.39X10^{-5}$
										EGG	1.13 (1.04,1.22)	$3.39X10^{-3}$					
										GIANT	1.34 (1.20,1.49)	$1.20X10^{-7}$					
rs112446794	*CEP120*[j]	5	122665465	T	C	1.23 (1.13,1.35)	$2.08X10^{-6}$	33.15%	28.69%	UKBB	1.07 (1.02,1.11)	$2.55X10^{-3}$	29.47%	28.21%	1.09 (1.06,1.13)	$3.45X10^{-10}$	$3.33X10^{-2}$
										SCOOP 2013	1.08 (0.98,1.19)	$1.38X10^{-1}$					
										EGG	1.12 (1.06,1.18)	$1.22X10^{-4}$					
										GIANT	1.05 (0.97,1.13)	$2.40X10^{-1}$					

Obese vs control						Discovery cohort				Replication cohorts					Combined analysis		
rsid	Nearest gene	Chr	Position (bp)	EA	NEA	OR (95% CI)	P value	EAF Ob	EAF Co	Cohort	OR (95% CI)	P value	EAF Ob	EAF Co	OR (95% CI)	P value	HetPVal
rs3760091	*SULT1A1*	16	28620800	C	G	1.24 (1.14,1.35)	$1.56X10^{-6}$	64.89%	60.23%	UKBB	1.09 (1.04,1.14)	$1.19X10^{-4}$	63.49%	61.44%	1.12 (1.07,1.16)	$2.65X10^{-8}$	$8.49X10^{-3}$

Table 2.6: GWAS results for SNPs meeting p<5x10-8 in all three analyses. EA= Effect allele (BMI increasing allele); NEA= Non-effect allele; OR = Odds ratio; 95% CI = 95% confidence interval for the odds ratio; EAF = effect allele frequency. Positions mapped to hg19, Build 37. a rs12995480 used as proxy in GIANT. b rs2384054 used as proxy in GIANT. c rs12641981 used as proxy in GIANT. d rs663129 used as proxy in GIANT, EGG and SCOOP 2013. e rs13007080 used as proxy in GIANT, EGG and SCOOP 2013. f rs7138803 used as proxy in SCOOP 2013. g rs6722587 used as proxy in GIANT, EGG and SCOOP 2013. h rs4132288 used as proxy in GIANT, EGG and SCOOP 2013. I rs1460940 used as proxy in GIANT, EGG and SCOOP 2013. j rs1366333 used as proxy in GIANT, EGG and SCOOP 2013.

SNPID	p-value	OR	conditioned p-value	conditioned OR	conditioned on
rs75398113*	$5.44X10^{-6}$	1.53	$2.94X10^{-4}$	1.5	rs205262**
rs205262**	$5.59X10^{-3}$	1.19	$7.09X10^{-1}$	1.03	rs75398113*

Table 2.7: Reciprocal conditional analysis of rs75398113 (SNRPC) and rs205262 (C6orf106) in SCOOP vs STILTS analysis. r2=0.29. p-values and ORs are shown without any LD correction applied. *Top signal in this study. **Previously established locus.

This is also the case for the *CEP120* locus (rs112446794) in the obese vs. controls analysis where reciprocal conditional analysis reveals the locus described here is driving the association observed at the reported locus (rs4308481 conditioned on rs112446794, $p_{conditioned}$=0.08,**Table 2.8**).

SNPID	p-value	OR	conditioned p-value	conditioned OR	conditioned on
rs112446794*	$1.94X10^{-6}$	1.23	$6.39X10^{-3}$	1.16	rs4308481**
rs4308481**	$1.89X10^{-5}$	1.2	$7.82X10^{-2}$	1.1	rs112446794*

Table 2.8: Reciprocal analysis of rs112446794 (CEP120) and rs4308481 (PRDM6-CEP120) in SCOOP vs UKHLS analysis. r2=0.36. p-values and ORs are shown without any LD correction applied. . *Top signal in this study. **Previously established locus

Finally, I used the independent BMI dataset from UKBB (**Methods 2.3.2.2**) to investigate whether any of the loci meeting our arbitrary $p \leq 10^{-5}$ in discovery efforts, were independently associated with BMI as a continuous trait. This identified a novel BMI-associated locus near *PKHD1* (SCOOP vs. STILTS $p=5.99x10^{-6}$, SCOOP vs. UKHLS $p=2.13x10^{-6}$, BMI $p=2.3x10^{-13}$, **Table 2.9**). Furthermore, there was an excess of nominally significant and directionally consistent signals in variants taken for replication in the obese vs. thin, and obese vs. controls analyses, after removing known signals and *PKHD1*, when comparing against a GWAS performed on the BMI dataset from UKBB (binomial $p=4.88x10^{-4}$, and binomial $p=9.77x10^{-3}$, respectively, **Methods, Table 2.9**).

Despite the smaller sample size, the SCOOP vs STILTS comparison had increased power to detect some loci, including the locus *FAM150B* (**Table 2.6)**, which did not meet our $p<10^{-5}$ threshold to be taken forward for replication in SCOOP vs UKHLS analysis ($p=2.36x10^{-4}$).

SCOOP vs. STILTS SNP	Nearest Gene	Effect	Other	EAF UKBB	Beta UKBB	SE UKBB	P value UKBB	Binomial P value
rs654240	*CCND1*	T	C	0.41	0.05	0.01	$1.50X10^{-5}$	$4.88X10^{-4}$
rs4447506	*PIK3C3*	G	A	0.39	0.05	0.01	$1.50X10^{-6}$	
rs2425853*	*CDH22*	C	G	0.69	0.06	0.01	$8.30X10^{-7}$	
rs2836760	*LOC400867*	T	G	0.09	0.05	0.02	$8.70X10^{-3}$	
rs6711131**	*BAZ2B*	A	G	0.63	0.06	0.02	$1.80X10^{-3}$	
rs375252497**	*SEMA3B*	AAATAAT AATAAT	A	0.67	0.10	0.02	$1.80X10^{-6}$	
rs11792928	*LMX1B*	T	C	0.29	0.03	0.01	$1.10X10^{-2}$	
rs516579	*MTCL1*	G	T	0.80	0.03	0.01	$2.30X10^{-2}$	
rs73145387	*ABI3BP*	C	G	0.97	0.07	0.03	$2.90X10^{-2}$	
rs11185396	*LOC100129138*	C	T	0.10	0.04	0.02	$2.60X10^{-2}$	
rs599291	*SLC44A5*	T	C	0.45	0.02	0.01	$2.50X10^{-2}$	
rs2784243***	*PKHD1*	T	C	0.58	0.07	0.01	$2.70X10^{-11}$	
SCOOP vs. UKHLS SNP	**Nearest Gene**	**Effect**	**Other**	**EAF UKBB**	**Beta UKBB**	**SE UKBB**	**P value UKBB**	**Binomial P value**
rs144435735	*LINC00682*	A	G	0.02	0.09	0.04	$1.20X10^{-2}$	$9.77X10^{-3}$
rs8096590	*LINC01541*	A	G	0.31	0.04	0.01	$7.90X10^{-4}$	
rs10944524	*MIR4643*	T	C	0.15	0.03	0.02	$2.80X10^{-2}$	
rs115474151	*SLC7A14*	A	T	0.01	0.18	0.09	$3.70X10^{-2}$	
rs11563327	*HOXA1*	C	T	0.71	0.02	0.01	$4.30X10^{-2}$	
rs1571570	*PBX3*	C	G	0.07	0.05	0.02	$1.90X10^{-2}$	
rs5873242**	*RANBP17*	A	T	0.32	0.08	0.02	$7.80X10^{-5}$	
rs75809547***	*PTBP2*	C	T	0.01	-0.15	0.06	$1.30X10^{-2}$	
rs898708	*PNOC*	C	T	0.69	0.02	0.01	$3.30X10^{-2}$	
rs2237402	*POU6F2*	G	A	0.66	0.05	0.01	$1.20X10^{-6}$	
rs10456655***	*PKHD1*	G	C	0.17	0.10	0.01	$2.30X10^{-13}$	
UKHLS vs. STILTS SNP	**Nearest Gene**	**Effect**	**Other**	**EAF UKBB**	**Beta UKBB**	**SE UKBB**	**P value UKBB**	**Binomial P value**
rs514529	*LRP5*	T	A	0.53	0.03	0.01	$5.10X10^{-3}$	$3.75X10^{-1}$
rs138251346	*LOC101929452*	A	G	0.99	0.13	0.07	$3.50X10^{-2}$	
rs553440779****	*KCNJ3*	T	C	0.01	-0.16	0.07	$2.20X10^{-2}$	

Table 2.9: Consistency of the direction of effect in candidate loci meeting p<1x10-5 in the discovery stages with BMI dataset GWAS. *Proxy for rs10546790. **Interim release used in UKBB for these signals. N=127,672. ***Novel signal – excluded from enrichment test. ****Opposite direction of effect. Effect=Effect allele (BMI increasing allele); Other=Other allele; Beta UKBB=Beta in UKBB BMI GWAS; SE UKBB=SE in UKBB BMI GWAS, P value UKBB=P value in UKBB BMI GWAS. Binomial P value=P value for binomial test).

2.5 Discussion

In this chapter, I and others performed the largest, at the time of completion, GWAS on healthy individuals with persistent thinness, and provided the first insights into the genetic architecture of this trait. I first show, using genome-wide data, that persistent healthy thinness is a heritable trait (h^2=28.07%) with a comparable heritability estimate to that of severe childhood obesity (h^2=32.33%). I also show a negative and incomplete genetic correlation between persistent healthy thinness and severe childhood obesity (r=-0.49, 95% CI [-0.17,-0.82], *p*=0.003). The incomplete genetic overlap between the two clinically ascertained traits is highlighted by the fact that some established BMI loci are more strongly associated at one end of the clinical BMI distribution compared to the other (e.g. *FTO* and *CADM2*), while others, appear to exert effects across the entire BMI spectrum (e.g. *MC4R* [184, 240, 241]). However, further exploration by simulation demonstrated some of these differences are likely to be due to the different degrees of "extremeness" of the two clinical cohorts (i.e. the difference in mean BMI between controls and obese individuals is larger than that of controls and thin individuals) and not due to a deviation from additive effects of the tested loci on BMI. It is worth noting that *CADM2* was not detected even at nominal significance in the previous SCOOP effort (*p*=0.41, OR=1.04 [160]), nor is it detected in the EGG study of childhood obesity (*p*=0.06, OR=1.06 [236]) which suggests that in this case the departure from expected OR (**Table 2.4**) may reflect a true finding. Variants in *CADM2* have also been associated with habitual physical activity [242]. GRS results also showed that overall genetic effects of the established loci do not deviate significantly from an additive model. This is in contrast with earlier studies which suggested larger effects at the higher

end of the BMI distribution [243, 244] but in agreement with more recent observations contrasting the bottom 5% and top 5% of the BMI tails where associated loci were also consistent with additive effects [207]. This is also in contrast with a previous study on height, where a deviation from additivity was found, but only for short individuals in the bottom 1.5% of the distribution [245], which suggests that analysis focused just on the most extreme individuals may be warranted.

Focusing on the 97 BMI associated loci [92] studied here, I show that the percentage of phenotypic variance explained by these loci is lower in persistently thin (4.33%) compared to obese individuals (10.67%) which is higher than previous estimates for BMI (~2.7% variance) using the same loci [92] and for severe obesity based on a subset of 32 loci (5.5% of the variance) [207]. Even though I partially addressed the possibility of age influencing these results by using data from the ASLPAC cohort, one cannot exclude the possibility that different effects of age and sex in our discovery cohorts (**Table 2.2**), and gene-by-environment interactions, could be influencing some of the results observed. For example, gene-by-environment interactions and age effects have been previously reported at the *FTO* locus [246-249] where a larger effect is detected in younger adults.

In studying thin individuals there are often concerns regarding the prevalence of eating disorders, notably anorexia nervosa, amongst participants. Prof Farooqi's group sought to carefully exclude eating disorders at two phases of recruitment (by medical history and by questionnaire). Additionally, in this work I demonstrate that in our cohort of healthy thin individuals, anorexia nervosa is unlikely to be a confounder as the two traits do not exhibit significant genetic correlation (r=0.13, 95% CI [-0.02,0.28], p=0.09). This was not the case for the UKBB replication cohort where a positive genetic correlation was observed (r= 0.49

95% CI [0.22-0.76] *p*=0.0003). The positive genetic correlation with anorexia was still observed after removing individuals with medical conditions that could explain their low BMI (r=0.62, 95% CI [0.30,0.92], *p*=0.0001**)**. These results highlight the importance of the careful phenotyping performed in the recruitment phase and the utility of the STILTS cohort as a resource to study healthy and persistent thinness.

In the genome-wide association analyses amongst the signals I took forward for replication, in addition to detecting established BMI-associated loci, I find a novel BMI-association at *PKHD1* in the UKBB BMI dataset (rs10456655, beta=0.10, p=2.3x10^{-13}, **Table 2.9**), where a proxy for this variant (rs2579994, r^2=1 in 1000G Phase 3 CEU) has been previously nominally associated with waist and hip circumference (*p*=5.60x10^{-5} and *p*=4.40x10^{-4} respectively) [250]. In addition, I found associations at loci that had only recently been established at the time of this study, using very large sample sizes. *FAM150B*, was only suggestively associated at discovery stage in Tachmazidou *et al* (2017) [251] (N=47,476, *p*=2.57×10^{-5}) whereas it reached genome-wide significance when contrasting SCOOP vs STILTS (N=2,927, *p*=2.07x10^{-8}, **Table 2.6**). Also, *PRDM6-CEP120* [180] was discovered in a Japanese study with a sample size of 173,430 and had not been previously reported in a European population. In this study, a signal near the locus (rs112446794, r^2=0.36) showed suggestive evidence of association in SCOOP vs UKHLS (*p*=2.08x10^{-6}, **Table 2.6**) with a significantly smaller sample size. Conditional analysis revealed the lead SNP in this study drives the association of the previously established signal (**Table 2.8**). *CEP120* codes for centrosomal protein 120 and variants near this locus have been previously associated with height [252] and waist circumference in East Asians [253]. Missense variants in the gene itself have been associated with rare ciliopathies [254, 255]. Lastly, amongst the signals taken forward for replication

from our case-control analyses, and after removing known and newly established loci, an enrichment of directionally consistent and nominal associations in the analysis of BMI as a continuous trait is observed, suggesting that some of these results may warrant additional investigation, in particular in similarly ascertained thin and obese cohorts. One such example is rs4447506, near *PIK3C3,* which was not only nominally significant and consistent in the independent UKBB BMI analysis (p=1.5x10^{-6}, **Table 2.9**), but also in the Locke et al. (2015) [92] BMI results (p= 0.01), and in the GIANT BMI tails analysis I used as replication (**Supplementary Table 5 of Riveros-Mckay et al 2018 [217] (Appendix A)**). Despite not reaching genome-wide significance in our discovery cohorts, directionally consistent suggestive associations were observed at a number of loci previously associated with BMI tails and with different obesity classes [207] (**Supplementary Table 10 of Riveros-Mckay et al 2018 [217] (Appendix A)**). One important limitation of this study design is that most replication cohorts are population derived and not clinically ascertained cohorts like our discovery dataset which could be a source for phenotype heterogeneity and subsequently reduced power to replicate signals.

It is also worth noting that these clinically ascertained extremes display evidence of incomplete genetic correlation with BMI, in contrast to previously described obesity classes (**Figure 2.5**) which supports the hypothesis that additional loci might be uncovered by focusing on these clinical extremes. Altogether, these results highlight some power advantages of using clinically ascertained extremes of the phenotype distribution to detect associations. However, a consequence of their very specific clinical ascertainment is that the conclusions we draw here cannot be straightforwardly extrapolated to the general population.

In summary, analyses performed in this chapter suggest that further genetic studies focused on persistently thin individuals are warranted. The STILTS cohort is an excellent resource in which to conduct such additional genetic exploration. Further genetic and phenotypic studies focused on persistently thin individuals may provide new insights into the mechanisms regulating human energy balance, and may uncover potential anti-obesity drug targets.

2.6 Future directions

Some outstanding questions remain from the work presented in this chapter, which could be addressed with some additional analyses. Namely, the possibility remains that the observed ORs in the UKHLS vs STILTS analysis could have been influenced by the significant age difference between the two cohorts. An analysis using only a subset of UKHLS samples with a similar age distribution to those in STILTS could provide a better estimate to explore differences in effect sizes on the tails of the BMI distribution.

Additionally, it would be of interest to assess the genetic correlation of extreme obesity and healthy persistent thinness with additional diseases and traits. These analyses would be feasible using summary statistics for >500 traits from UK Biobank participants recently made available (http://www.nealelab.is/uk-biobank/).

Lastly, for future studies it would be of interest to explore multiple BMI cutoffs for obesity in adults from UK Biobank and calculate genetic correlation with SCOOP to find the optimal BMI cutoff for future replication studies in adults when pursuing findings originating from the SCOOP cohort.

3 Chapter 3: The Role of Rare Variation in Circulating Metabolic Biomarkers

3.1 Introduction

Metabolic measurements reflect an individual's endogenous biochemical processes and environmental exposures [256, 257]. Many circulating lipids, lipoproteins and metabolites have been previously implicated in the development of cardiovascular disease (CVD) [258-261] or used as biomarkers for disease diagnosis or prognosis [262, 263]. High circulating levels of total cholesterol (TC) and low-density lipoprotein (LDL) cholesterol, for example, have been associated with increased risk of coronary heart disease (CHD)[264]. On the other hand, circulating levels of high-density lipoprotein (HDL) cholesterol have been regarded as protective factors against CHD [265]. Despite the observed association between low HDL levels and increased CHD risk, a causal role for HDL levels was more unclear before genetic studies, due to potential confounding by other CHD risk factors correlated with low HDL, like increased plasma triglycerides (TG) [266].

In the diagnostic setting, metabolites like creatinine and branched chain amino acids (valine, leucine and isoleucine) are helpful biomarkers for diseases such a kidney disease [267] , or hyperinsulinism [268-270]. Understanding the genetic influence on circulating levels of these metabolic biomarkers can help us gain insight into the biological processes regulating these traits, lead to improve aetiological understanding of CVD and identify novel potential therapeutic drug targets. Notable examples of candidate drug targets identified via genetic approaches are *LDLR* [271, 272], *APOB* [273, 274] and *PCSK9* [275, 276]. Mipomersen, a commercially available *APOB* inhibitor, has already shown association with reduction in cardiovascular events in patients with hypercholesterolaemia [277] and two

PCSK9 inhibitors: alirocumab and evolocumab have been shown to reduce risk of myocardial infranction (MI) and stroke in clinical trials [278].

Genome-wide association studies (GWAS) focusing on traditionally measured lipid traits have greatly expanded our knowledge of lipid biology and to date 250 loci have been robustly associated with total cholesterol (TC), high-density lipoprotein cholesterol (HDL-C), low-density lipoprotein cholesterol (LDL-C), and/or triglycerides (TG) [84, 116, 279-285]. Through these studies it has been found that most loci identified in European populations contribute to the genetic architecture of lipid traits across global populations [116], that there are metabolic links between blood lipids and type 2 diabetes, blood pressure, waist-hip-ratio and BMI [280], and more recently that some mechanisms of lowering LDL-C increase type 2 diabetes (T2D) risk [84]. Mendelian randomisation (MR) approaches have also used information gained through GWAS to examine the causal link between low HDL levels and CVD where findings suggest that low HDL levels are not causal for CVD since many studies report no association between CVD and genetically lowered levels of HDL [110-114]. These MR approaches have also been used to identify a potential causal link between increased plasma urate levels and CVD [286], although other studies measuring serum urate levels have not found that link [287].

In addition to this, more detailed metabolic profiling using high resolution nuclear magnetic resonance (NMR) measurements, has proven helpful to find additional lipid and small molecule metabolism-associated loci with smaller sample sizes, and to assess pleiotropic effects of previously established loci [38, 173, 288]. An example of this, is a novel link between the *LPA* locus and very-low-density lipoprotein (VLDL) metabolism (measured by using high resolution NMR), with effect sizes twice as large as those found for traditionally

measured lipid traits like LDL-C and TC, suggesting these measurements are better at capturing the underlying biological processes in lipid metabolism than traditionally measured lipid traits. In this same study, by constructing a genetic risk score using variants associated with Lp(a) levels and using a Mendelian randomisation approach the authors were able to demonstrate a causal link between increased Lp(a) levels and overall lipoprotein metabolism [173].

Despite the increased usage of exome arrays which have been used at scale to capture low-frequency and rare coding variation contributing to lipid and amino acid metabolism [84, 282-284, 288, 289], large-scale sequencing studies have the added value of assessing rare variation at single nucleotide resolution across the whole genome, or exome, including the detection of private variants which could have large effects on protein function. These approaches enabled, for example, the discovery of inactivating variants in key proteins which are models for drug target antagonism such as ANGPTL4, where carriers of a missense E40K variant and other inactivating variants had reduced risk of CHD [290, 291].

Notwithstanding the progress made in recent years in understanding the genetic aetiology of a number of traditional lipid traits, at the time of this analysis, there were no studies coupling NMR measurements with sequencing data to explore the role of rare genetic variants in the metabolism of high resolution lipid, lipoprotein and metabolite traits. In this chapter, I address this gap in knowledge by examining the contribution of rare variation (MAF <1%) to 226 serum metabolic measurements in the INTERVAL cohort which consist of healthy blood donors residing in the UK. This project was done in collaboration with Dr Adam Butterworth's group at the University of Cambridge. My work involved QCing of sequencing and phenotype data as well as all analytical aspects of the study.

3.2 Chapter aims

The overall aim of this chapter is to explore how coupling next generation sequencing (NGS) and high resolution metabolic measurements can help us gain new insights into metabolic biomarker biology through rare variant analyses. To do this, I took advantage of the INTERVAL cohort, which is comprised of healthy blood donors who have been deeply phenotyped and who also have genome-wide array data. In my project I used data from a subset of 7,142 participants with NMR measurements and NGS data to:

I. Identify novel loci, genes and/or gene sets associated with metabolic biomarkers.
II. Identify effector transcripts at established GWAS loci for traditionally measured lipid traits.
III. Assess the contribution of genes known to be involved in lipoprotein metabolism to the tails of the phenotype distribution of lipoprotein and glyceride traits in a healthy population.

3.3 Methods

3.3.1 Participants

The INTERVAL cohort consists of 47,393 predominantly healthy blood donors in the UK [292]. This study was the result of a collaboration between the Universities of Cambridge and Oxford and the NHS Blood and Transplant Unit. The study was set up to determine the optimum intervals between donations for men and women without affecting the overall health of blood donors. Individuals were asked to fill an online general questionnaire every six months containing basic lifestyle and health-related information. At the time of this

study, a different set of biomarker assays were performed on blood samples collected on the first visit and those collected on the 2 year follow-up visit. All individuals have been genotyped using the Affymetrix UK Biobank Axiom Array and imputed using a combined UK10K-1000G Phase III imputation panel [293]. A subset of 4,502 individuals was selected for whole-exome sequencing (WES) [294] and another subset of 3,762 was selected for whole-genome sequencing (WGS). There was an overlap of 54 individuals in both datasets.

3.3.2 Sequencing and genotype calling

WES and WGS were performed at the Wellcome Sanger Institute (WSI) sequencing facility, with read alignment and genotype calling performed by the Human Genetics Informatics (HGI) group at Sanger. For WES sheared DNA was prepared for Illumina paired-end sequencing and enriched for target regions using Agilent's SureSelect Human All Exon V5 capture technology (Agilent Technologies; Santa Clara, California, USA). The exome capture library preparation was sequenced using the Illumina HiSeq 2000 platform as paired-end 75 bp reads. Reads were aligned to the GRCh37 human reference genome using BWA (v0.5.10) [295]. GATK HaplotypeCaller v3.4 [296] was used for variant calling and recalibration. For WGS sheared DNA was prepared for Illumina paired-end sequencing. Sequencing was performed using the Illumina HiSeq X platform as paired-end 75 bp reads. Reads were aligned to the GRCh38 human reference genome using mostly BWA (v.0.7.12) although a subset of samples was aligned with v.0.7.13 or v.0.7.15. GATK HaplotypeCaller v3.5 was used for variant calling and recalibration. I extracted coordinates from the VCF files that mapped to regions targeted in the WES. I then used custom scripts to transform coordinates of variants to GRCh37 human reference.

3.3.3 Sample QC

I performed sample QC for WES using the same filters Tarjinder Singh used on a previous release of the INTERVAL WES dataset [294]. Sample QC for WGS was performed by Kousik Kundu, Klaudia Walter and I. For WES data I filtered out samples based on the following criteria: i) withdrawn consent; ii) estimated contamination >3% according to the software VerifyBamID [297]; iii) sex inferred from genetic data different from sex supplied ; iv) non-European samples after manual inspection of clustering in 1000G principal components analysis (PCA) and choosing cutoffs on the first 2 PCs; v) heterozygosity outliers (samples +/- 3 SD away from the mean number of heterozygous counts); vi) non-reference homozygosity outliers (samples +/- 3 SD away from the mean number of non-reference homozygous counts); vii) outlier Ti/TV rates (transition to transversion ratio +/- 3 SD away from the mean ratio); viii) excess singletons (number of singleton variants >3 SD from the cohort mean). After quality control 4,070 WES samples were kept for downstream analyses. For WGS data we filtered out samples based on the following criteria: i) estimated contamination >2% according to software VerifyBamID; ii) non-reference discordance (NRD) with genotype data on the same samples >4%; iii) European population outliers from PCA (PC1>0 and minimum PC2); iv) heterozygosity outliers (samples +/- 3 SD away from the mean number of heterozygous counts); v) number of third-degree relatives (proportion IBD (PI_HAT) >0.125) > 18, vi) overlap with WES. After quality control 3,670 WGS samples were kept for further analyses.

3.3.4 Variant QC

For variants with MAF>1% I used the following thresholds to exclude variants: i) VQSR: 99.90% tranche for WES and 99% tranche for WGS; ii) missingness >3%; iii) HWE $p<1x10^{-5}$. For variants with MAF≤1% the following thresholds were used: i) VQSR: 99.90% tranche for WES, 99% tranche for WGS SNPs and 90% tranche for WGS indels; ii) GQ: <20 for SNPs and <60 for indels; iii) DP <2; iv) AB>15 & <80 for heterozygous variants. After genotype-level QC (GQ,DP,AB) only variants with <3% missingness were kept. 1,716,946 variants were kept in the final WES release and 1,724,250 in the final WGS release.

3.3.5 Phenotype QC

A total of 230 metabolic biomarkers were produced by the serum NMR metabolomics platform (Nightingale Health Ltd.) [298] on 46,097 blood samples from the INTERVAL cohort collected on the first visit. Phenotyping was performed by Antti J. Kangas (Nightingale Health Ltd.). I performed phenotype QC on the raw phenotypes. Glucose, lactose, pyruvate and acetate were excluded initially due to unreliable measurements according to platform provider. Conjugated linoleic acid and conjugated linoleic acid to total fatty acid ratio were set to missing for 3,585 samples showing signs of peroxidation. Creatinine levels were set to missing for 1,993 samples with isopropyl alcohol signals. Glutamine levels were set to missing for 347 samples that showed signs of glutamine to glutamate degradation. Samples with more than 30% missingness or identified as EDTA plasma were removed. After this step, for each pair of related samples (PI_HAT>0.125) I kept only one, preferentially keeping samples with the lowest missingness in WES or lowest NRD in WGS. Phenotypes were rank-based inverse normalised for all individuals. Clare Oliver-Williams assessed which technical

covariates influenced phenotype levels and determined centre, processing duration and month of donation were possible sources of batch effects. I then separately performed linear regression for WES and WGS adjusting for age, gender, centre, processing duration, month of donation and 10 PCs. Residuals from both WES and WGS linear regressions were used as the outcome variables in all subsequent analyses. After this final step I kept 3,741 samples in the WES dataset and 3,420 samples in the WGS dataset for downstream analyses.

3.3.6 Single point analyses

Power calculations to define MAF threshold for single point analyses were done using Quanto [234]. I used the WES data as a discovery dataset and performed association analyses using SNPTEST v2.5.2 [226] under an additive model. Variants were taken forward for validation if $p < 1x10^{-5}$. I then performed association analyses using SNPTEST on the WGS data which was used as a validation dataset. Results were subsequently meta-analysed using a fixed-effects model in METAL [238]. Genome-wide significance threshold was calculated as: $0.05/ (276,563*19)=9.52x10^{-9}$, where 276,563 is the number of tested variants with MAF>0.1% and 19 is the number of PCs explaining >95% of the variance of 226 metabolic biomarkers , an approach previously used in similar studies using the same NMR platform [38, 173]. A signal was considered to replicate if after meta-analysis it met the following criteria: i) it met the defined genome-wide significance threshold ($9.52x10^{-9}$); and ii) it was nominally significant ($p<0.05$) in the validation dataset (WGS). After this step, to define loci, I performed clumping using PLINK [223] based on the lowest p for each variant on any trait-association using an $r^2=0.2$ and a window size of 1Mb.

3.3.7 Gene-based analyses

I annotated coding variant consequences with VEP [50] using Ensembl gene set version 75 for the hg19/GRCh37 human genome assembly. Loss-of–function (LoF) variants were annotated with a VEP plugin: LOFTEE (https://github.com/konradjk/loftee). This plugin uses distance to end of transcript and other in-frame splice sites, non-canonical splice site information and size of introns to remove LoF that are less likely to have a damaging impact on protein structure. I downloaded M-CAP scores and extracted all missense variants with AC>=1 in the WES or WGS datasets [51]. Two different nested tests were used to group rare variants into testable gene units: predicted to be high confidence LoF by LOFTEE in any transcript of the gene, and the same LoF variants plus rare (MAF <1%) missense variants, mapping to any transcript of the gene, predicted to be likely deleterious by M-CAP (M-CAP score >0.025) (MCAP+LoF). M-CAP uses a machine learning algorithm integrating multiple annotations (e.g base-pair conservation, amino acid conservation, chemical properties of substituted amino acid, etc) to predict the pathogenicity of rare (MAF <1%) missense variants.

I performed rare-variant aggregation tests as implemented in the SKAT-O R package [52, 53]. For the LoF tests, I performed a burden test (rho=1) whereas for the MCAP+LoF tests I used the optimal unified approach (method="optimal.adj"). Genes were taken forward for validation if $p<5x10^{-3}$.

To increase power in my analyses I also implemented a strategy to incorporate information from the multiple phenotypes measured in our dataset, by adjusting for correlated phenotypes, which has been shown to increase power in single point association analyses [30]. To minimise chances of a false positive association I only adjusted for phenotypes as

covariates at the validation stage ensuring evidence of association in discovery stage was present without adjustment for covariates. In order for a metabolic biomarker to be selected as a covariate in the validation stage, the following conditions had to be met: i) no evidence of genetic correlation ($p>0.05$) with outcome using publicly available summary statistics from Kettunen et al. (2016) [25]; ii) phenotypic correlation in our dataset >10%; iii) not belonging to same metabolic biomarker supergroup as outcome (**Table 3.1**). This approach resulted in 99 eligible NMR traits for which other traits could be used as covariates. METASKAT [54] was used to perform meta-analysis using the same parameters as in discovery. A signal was considered to replicate if: i) it met the Bonferroni corrected gene-level significance threshold ($p < 1.32x10^{-7}$); ii) >2 variants were tested; iii) it was nominally significant in the unadjusted test for WGS (i.e without adjusting for correlated traits). The Bonferroni corrected gene-level significance threshold was chosen after adjusting the standard gene-level significance threshold ($2.5x10^{-6}$) for 19 PCs. To test if a single variant was driving an observed association, I performed leave-one-out analysis for all variants contributing to the test. An association was considered to be driven by a single variant if, after removing it, the test resulted in a non-significant association ($p>0.05$).

Traits	Description	Supergroup	Single point tests	Gene tests	Gene-set tests	Enrichment near GWAS signals tests	Tails tests
XXL-VLDL-P	Concentration of chylomicrons and extremely large VLDL particles	Lipid and lipoprotein	X	X	X		X
XXL-VLDL-L	Total lipids in chylomicrons and extremely large VLDL	Lipid and lipoprotein	X	X	X		X
XXL-VLDL-PL	Phospholipids in chylomicrons and extremely large VLDL	Lipid and lipoprotein	X	X	X		X
XXL-VLDL-C	Total cholesterol in chylomicrons and extremely large VLDL	Lipid and lipoprotein	X	X	X	X	X
XXL-VLDL-CE	Cholesterol esters in chylomicrons and extremely large VLDL	Lipid and lipoprotein	X	X	X	X	X
XXL-VLDL-FC	Free cholesterol in chylomicrons and extremely large VLDL	Lipid and lipoprotein	X	X	X	X	X
XXL-VLDL-TG	Triglycerides in chylomicrons and extremely large VLDL	Lipid and lipoprotein	X	X	X	X	X
XL-VLDL-P	Concentration of very large VLDL particles	Lipid and lipoprotein	X	X	X		X
XL-VLDL-L	Total lipids in very large VLDL	Lipid and lipoprotein	X	X	X		X
XL-VLDL-PL	Phospholipids in very large VLDL	Lipid and lipoprotein	X	X	X		X

Traits	Description	Supergroup	Single point tests	Gene tests	Gene-set tests	Enrichment near GWAS signals tests	Tails tests
XL-VLDL-C	Total cholesterol in very large VLDL	Lipid and lipoprotein	X	X	X	X	X
XL-VLDL-CE	Cholesterol esters in very large VLDL	Lipid and lipoprotein	X	X	X	X	X
XL-VLDL-FC	Free cholesterol in very large VLDL	Lipid and lipoprotein	X	X	X	X	X
XL-VLDL-TG	Triglycerides in very large VLDL	Lipid and lipoprotein	X	X	X	X	X
L-VLDL-P	Concentration of large VLDL particles	Lipid and lipoprotein	X	X	X		X
L-VLDL-L	Total lipids in large VLDL	Lipid and lipoprotein	X	X	X		X
L-VLDL-PL	Phospholipids in large VLDL	Lipid and lipoprotein	X	X	X		X
L-VLDL-C	Total cholesterol in large VLDL	Lipid and lipoprotein	X	X	X	X	X
L-VLDL-CE	Cholesterol esters in large VLDL	Lipid and lipoprotein	X	X	X	X	X
L-VLDL-FC	Free cholesterol in large VLDL	Lipid and lipoprotein	X	X	X	X	X
L-VLDL-TG	Triglycerides in large VLDL	Lipid and lipoprotein	X	X	X	X	X
M-VLDL-P	Concentration of medium VLDL particles	Lipid and lipoprotein	X	X	X		X
M-VLDL-L	Total lipids in medium VLDL	Lipid and lipoprotein	X	X	X		X
M-VLDL-PL	Phospholipids in medium VLDL	Lipid and lipoprotein	X	X	X		X
M-VLDL-C	Total cholesterol in medium VLDL	Lipid and lipoprotein	X	X	X	X	X
M-VLDL-CE	Cholesterol esters in medium VLDL	Lipid and lipoprotein	X	X	X	X	X
M-VLDL-FC	Free cholesterol in medium VLDL	Lipid and lipoprotein	X	X	X	X	X
M-VLDL-TG	Triglycerides in medium VLDL	Lipid and lipoprotein	X	X	X	X	X
S-VLDL-P	Concentration of small VLDL particles	Lipid and lipoprotein	X	X	X		X
S-VLDL-L	Total lipids in small VLDL	Lipid and lipoprotein	X	X	X		X
S-VLDL-PL	Phospholipids in small VLDL	Lipid and lipoprotein	X	X	X		X
S-VLDL-C	Total cholesterol in small VLDL	Lipid and lipoprotein	X	X	X	X	X
S-VLDL-CE	Cholesterol esters in small VLDL	Lipid and lipoprotein	X	X	X	X	X
S-VLDL-FC	Free cholesterol in small VLDL	Lipid and lipoprotein	X	X	X	X	X
S-VLDL-TG	Triglycerides in small VLDL	Lipid and lipoprotein	X	X	X	X	X
XS-VLDL-P	Concentration of very small VLDL particles	Lipid and lipoprotein	X	X	X		X
XS-VLDL-L	Total lipids in very small VLDL	Lipid and lipoprotein	X	X	X		X
XS-VLDL-PL	Phospholipids in very small VLDL	Lipid and lipoprotein	X	X	X		X
XS-VLDL-C	Total cholesterol in very small VLDL	Lipid and lipoprotein	X	X	X	X	X
XS-VLDL-CE	Cholesterol esters in very small VLDL	Lipid and lipoprotein	X	X	X	X	X
XS-VLDL-FC	Free cholesterol in very small VLDL	Lipid and lipoprotein	X	X	X	X	X
XS-VLDL-TG	Triglycerides in very small VLDL	Lipid and lipoprotein	X	X	X	X	X
IDL-P	Concentration of IDL particles	Lipid and lipoprotein	X	X	X		
IDL-L	Total lipids in IDL	Lipid and lipoprotein	X	X	X		
IDL-PL	Phospholipids in IDL	Lipid and lipoprotein	X	X	X		
IDL-C	Total cholesterol in IDL	Lipid and lipoprotein	X	X	X	X	
IDL-CE	Cholesterol esters in IDL	Lipid and	X	X	X	X	

Traits	Description	Supergroup	Single point tests	Gene tests	Gene-set tests	Enrichment near GWAS signals tests	Tails tests
		lipoprotein					
IDL-FC	Free cholesterol in IDL	Lipid and lipoprotein	X	X	X	X	
IDL-TG	Triglycerides in IDL	Lipid and lipoprotein	X	X	X	X	
L-LDL-P	Concentration of large LDL particles	Lipid and lipoprotein	X	X	X	X	X
L-LDL-L	Total lipids in large LDL	Lipid and lipoprotein	X	X	X	X	X
L-LDL-PL	Phospholipids in large LDL	Lipid and lipoprotein	X	X	X	X	X
L-LDL-C	Total cholesterol in large LDL	Lipid and lipoprotein	X	X	X	X	X
L-LDL-CE	Cholesterol esters in large LDL	Lipid and lipoprotein	X	X	X	X	X
L-LDL-FC	Free cholesterol in large LDL	Lipid and lipoprotein	X	X	X	X	X
L-LDL-TG	Triglycerides in large LDL	Lipid and lipoprotein	X	X	X	X	X
M-LDL-P	Concentration of medium LDL particles	Lipid and lipoprotein	X	X	X	X	X
M-LDL-L	Total lipids in medium LDL	Lipid and lipoprotein	X	X	X	X	X
M-LDL-PL	Phospholipids in medium LDL	Lipid and lipoprotein	X	X	X	X	X
M-LDL-C	Total cholesterol in medium LDL	Lipid and lipoprotein	X	X	X	X	X
M-LDL-CE	Cholesterol esters in medium LDL	Lipid and lipoprotein	X	X	X	X	X
M-LDL-FC	Free cholesterol in medium LDL	Lipid and lipoprotein	X	X	X	X	X
M-LDL-TG	Triglycerides in medium LDL	Lipid and lipoprotein	X	X	X	X	X
S-LDL-P	Concentration of small LDL particles	Lipid and lipoprotein	X	X	X	X	X
S-LDL-L	Total lipids in small LDL	Lipid and lipoprotein	X	X	X	X	X
S-LDL-PL	Phospholipids in small LDL	Lipid and lipoprotein	X	X	X	X	X
S-LDL-C	Total cholesterol in small LDL	Lipid and lipoprotein	X	X	X	X	X
S-LDL-CE	Cholesterol esters in small LDL	Lipid and lipoprotein	X	X	X	X	X
S-LDL-FC	Free cholesterol in small LDL	Lipid and lipoprotein	X	X	X	X	X
S-LDL-TG	Triglycerides in small LDL	Lipid and lipoprotein	X	X	X	X	X
XL-HDL-P	Concentration of very large HDL particles	Lipid and lipoprotein	X	X	X	X	X
XL-HDL-L	Total lipids in very large HDL	Lipid and lipoprotein	X	X	X	X	X
XL-HDL-PL	Phospholipids in very large HDL	Lipid and lipoprotein	X	X	X	X	X
XL-HDL-C	Total cholesterol in very large HDL	Lipid and lipoprotein	X	X	X	X	X
XL-HDL-CE	Cholesterol esters in very large HDL	Lipid and lipoprotein	X	X	X	X	X
XL-HDL-FC	Free cholesterol in very large HDL	Lipid and lipoprotein	X	X	X	X	X
XL-HDL-TG	Triglycerides in very large HDL	Lipid and lipoprotein	X	X	X	X	X
L-HDL-P	Concentration of large HDL particles	Lipid and lipoprotein	X	X	X	X	X
L-HDL-L	Total lipids in large HDL	Lipid and lipoprotein	X	X	X	X	X
L-HDL-PL	Phospholipids in large HDL	Lipid and lipoprotein	X	X	X	X	X
L-HDL-C	Total cholesterol in large HDL	Lipid and lipoprotein	X	X	X	X	X
L-HDL-CE	Cholesterol esters in large HDL	Lipid and lipoprotein	X	X	X	X	X
L-HDL-FC	Free cholesterol in large HDL	Lipid and lipoprotein	X	X	X	X	X

Traits	Description	Supergroup	Single point tests	Gene tests	Gene-set tests	Enrichment near GWAS signals tests	Tails tests
L-HDL-TG	Triglycerides in large HDL	Lipid and lipoprotein	X	X	X	X	X
M-HDL-P	Concentration of medium HDL particles	Lipid and lipoprotein	X	X	X	X	X
M-HDL-L	Total lipids in medium HDL	Lipid and lipoprotein	X	X	X	X	X
M-HDL-PL	Phospholipids in medium HDL	Lipid and lipoprotein	X	X	X	X	X
M-HDL-C	Total cholesterol in medium HDL	Lipid and lipoprotein	X	X	X	X	X
M-HDL-CE	Cholesterol esters in medium HDL	Lipid and lipoprotein	X	X	X	X	X
M-HDL-FC	Free cholesterol in medium HDL	Lipid and lipoprotein	X	X	X	X	X
M-HDL-TG	Triglycerides in medium HDL	Lipid and lipoprotein	X	X	X	X	X
S-HDL-P	Concentration of small HDL particles	Lipid and lipoprotein	X	X	X	X	X
S-HDL-L	Total lipids in small HDL	Lipid and lipoprotein	X	X	X	X	X
S-HDL-PL	Phospholipids in small HDL	Lipid and lipoprotein	X	X	X	X	X
S-HDL-C	Total cholesterol in small HDL	Lipid and lipoprotein	X	X	X	X	X
S-HDL-CE	Cholesterol esters in small HDL	Lipid and lipoprotein	X	X	X	X	X
S-HDL-FC	Free cholesterol in small HDL	Lipid and lipoprotein	X	X	X	X	X
S-HDL-TG	Triglycerides in small HDL	Lipid and lipoprotein	X	X	X	X	X
XXL-VLDL-PL_%	Phospholipids to total lipds ratio in chylomicrons and extremely large VLDL	Lipid and lipoprotein	X	X	X		
XXL-VLDL-C_%	Total cholesterol to total lipids ratio in chylomicrons and extremely large VLDL	Lipid and lipoprotein	X	X	X	X	
XXL-VLDL-CE_%	Cholesterol esters to total lipids ratio in chylomicrons and extremely large VLDL	Lipid and lipoprotein	X	X	X	X	
XXL-VLDL-FC_%	Free cholesterol to total lipids ratio in chylomicrons and extremely large VLDL	Lipid and lipoprotein	X	X	X	X	
XXL-VLDL-TG_%	Triglycerides to total lipids ratio in chylomicrons and extremely large VLDL	Lipid and lipoprotein	X	X	X	X	
XL-VLDL-PL_%	Phospholipids to total lipds ratio in very large VLDL	Lipid and lipoprotein	X	X	X		
XL-VLDL-C_%	Total cholesterol to total lipids ratio in very large VLDL	Lipid and lipoprotein	X	X	X	X	
XL-VLDL-CE_%	Cholesterol esters to total lipids ratio in very large VLDL	Lipid and lipoprotein	X	X	X	X	
XL-VLDL-FC_%	Free cholesterol to total lipids ratio in very large VLDL	Lipid and lipoprotein	X	X	X	X	
XL-VLDL-TG_%	Triglycerides to total lipids ratio in very large VLDL	Lipid and lipoprotein	X	X	X	X	
L-VLDL-PL_%	Phospholipids to total lipds ratio in large VLDL	Lipid and lipoprotein	X	X	X		
L-VLDL-C_%	Total cholesterol to total lipids ratio in large VLDL	Lipid and lipoprotein	X	X	X	X	
L-VLDL-CE_%	Cholesterol esters to total lipids ratio in large VLDL	Lipid and lipoprotein	X	X	X	X	
L-VLDL-FC_%	Free cholesterol to total lipids ratio in large VLDL	Lipid and lipoprotein	X	X	X	X	
L-VLDL-TG_%	Triglycerides to total lipids ratio in large VLDL	Lipid and lipoprotein	X	X	X	X	
M-VLDL-PL_%	Phospholipids to total lipds ratio in medium VLDL	Lipid and lipoprotein	X	X	X		
M-VLDL-C_%	Total cholesterol to total lipids ratio in medium VLDL	Lipid and lipoprotein	X	X	X	X	
M-VLDL-CE_%	Cholesterol esters to total lipids ratio in medium VLDL	Lipid and lipoprotein	X	X	X	X	
M-VLDL-FC_%	Free cholesterol to total lipids ratio in medium VLDL	Lipid and lipoprotein	X	X	X	X	
M-VLDL-TG_%	Triglycerides to total lipids ratio in medium VLDL	Lipid and lipoprotein	X	X	X	X	
S-VLDL-PL_%	Phospholipids to total lipds ratio in small VLDL	Lipid and lipoprotein	X	X	X		
S-VLDL-C_%	Total cholesterol to total lipids ratio in	Lipid and	X	X	X	X	

Traits	Description	Supergroup	Single point tests	Gene tests	Gene-set tests	Enrichment near GWAS signals tests	Tails tests
	small VLDL	lipoprotein					
S-VLDL-CE_%	Cholesterol esters to total lipids ratio in small VLDL	Lipid and lipoprotein	X	X	X	X	
S-VLDL-FC_%	Free cholesterol to total lipids ratio in small VLDL	Lipid and lipoprotein	X	X	X	X	
S-VLDL-TG_%	Triglycerides to total lipids ratio in small VLDL	Lipid and lipoprotein	X	X	X	X	
XS-VLDL-PL_%	Phospholipids to total lipds ratio in very small VLDL	Lipid and lipoprotein	X	X	X		
XS-VLDL-C_%	Total cholesterol to total lipids ratio in very small VLDL	Lipid and lipoprotein	X	X	X	X	
XS-VLDL-CE_%	Cholesterol esters to total lipids ratio in very small VLDL	Lipid and lipoprotein	X	X	X	X	
XS-VLDL-FC_%	Free cholesterol to total lipids ratio in very small VLDL	Lipid and lipoprotein	X	X	X	X	
XS-VLDL-TG_%	Triglycerides to total lipids ratio very small VLDL	Lipid and lipoprotein	X	X	X	X	
IDL-PL_%	Phospholipids to total lipds ratio in IDL	Lipid and lipoprotein	X	X	X		
IDL-C_%	Total cholesterol to total lipids ratio in IDL	Lipid and lipoprotein	X	X	X	X	
IDL-CE_%	Cholesterol esters to total lipids ratio in IDL	Lipid and lipoprotein	X	X	X	X	
IDL-FC_%	Free cholesterol to total lipids ratio in IDL	Lipid and lipoprotein	X	X	X	X	
IDL-TG_%	Triglycerides to total lipids ratio in IDL	Lipid and lipoprotein	X	X	X	X	
L-LDL-PL_%	Phospholipids to total lipds ratio in large LDL	Lipid and lipoprotein	X	X	X	X	
L-LDL-C_%	Total cholesterol to total lipids ratio in large LDL	Lipid and lipoprotein	X	X	X	X	
L-LDL-CE_%	Cholesterol esters to total lipids ratio in large LDL	Lipid and lipoprotein	X	X	X	X	
L-LDL-FC_%	Free cholesterol to total lipids ratio in large LDL	Lipid and lipoprotein	X	X	X	X	
L-LDL-TG_%	Triglycerides to total lipids ratio in large LDL	Lipid and lipoprotein	X	X	X	X	
M-LDL-PL_%	Phospholipids to total lipds ratio in medium LDL	Lipid and lipoprotein	X	X	X	X	
M-LDL-C_%	Total cholesterol to total lipids ratio in medium LDL	Lipid and lipoprotein	X	X	X	X	
M-LDL-CE_%	Cholesterol esters to total lipids ratio in medium LDL	Lipid and lipoprotein	X	X	X	X	
M-LDL-FC_%	Free cholesterol to total lipids ratio in medium LDL	Lipid and lipoprotein	X	X	X	X	
M-LDL-TG_%	Triglycerides to total lipids ratio in medium LDL	Lipid and lipoprotein	X	X	X	X	
S-LDL-PL_%	Phospholipids to total lipds ratio in small LDL	Lipid and lipoprotein	X	X	X	X	
S-LDL-C_%	Total cholesterol to total lipids ratio in small LDL	Lipid and lipoprotein	X	X	X	X	
S-LDL-CE_%	Cholesterol esters to total lipids ratio in small LDL	Lipid and lipoprotein	X	X	X	X	
S-LDL-FC_%	Free cholesterol to total lipids ratio in small LDL	Lipid and lipoprotein	X	X	X	X	
S-LDL-TG_%	Triglycerides to total lipids ratio in small LDL	Lipid and lipoprotein	X	X	X	X	
XL-HDL-PL_%	Phospholipids to total lipds ratio in very large HDL	Lipid and lipoprotein	X	X	X	X	
XL-HDL-C_%	Total cholesterol to total lipids ratio in very large HDL	Lipid and lipoprotein	X	X	X	X	
XL-HDL-CE_%	Cholesterol esters to total lipids ratio in very large HDL	Lipid and lipoprotein	X	X	X	X	
XL-HDL-FC_%	Free cholesterol to total lipids ratio in very large HDL	Lipid and lipoprotein	X	X	X	X	
XL-HDL-TG_%	Triglycerides to total lipids ratio in very large HDL	Lipid and lipoprotein	X	X	X	X	
L-HDL-PL_%	Phospholipids to total lipds ratio in large HDL	Lipid and lipoprotein	X	X	X	X	
L-HDL-C_%	Total cholesterol to total lipids ratio in large HDL	Lipid and lipoprotein	X	X	X	X	
L-HDL-CE_%	Cholesterol esters to total lipids ratio in large HDL	Lipid and lipoprotein	X	X	X	X	

Traits	Description	Supergroup	Single point tests	Gene tests	Gene-set tests	Enrichment near GWAS signals tests	Tails tests
L-HDL-FC_%	Free cholesterol to total lipids ratio in large HDL	Lipid and lipoprotein	X	X	X	X	
L-HDL-TG_%	Triglycerides to total lipids ratio in large HDL	Lipid and lipoprotein	X	X	X	X	
M-HDL-PL_%	Phospholipids to total lipds ratio in medium HDL	Lipid and lipoprotein	X	X	X	X	
M-HDL-C_%	Total cholesterol to total lipids ratio in medium HDL	Lipid and lipoprotein	X	X	X	X	
M-HDL-CE_%	Cholesterol esters to total lipids ratio in medium HDL	Lipid and lipoprotein	X	X	X	X	
M-HDL-FC_%	Free cholesterol to total lipids ratio in medium HDL	Lipid and lipoprotein	X	X	X	X	
M-HDL-TG_%	Triglycerides to total lipids ratio in medium HDL	Lipid and lipoprotein	X	X	X	X	
S-HDL-PL_%	Phospholipids to total lipds ratio in small HDL	Lipid and lipoprotein	X	X	X	X	
S-HDL-C_%	Total cholesterol to total lipids ratio in small HDL	Lipid and lipoprotein	X	X	X	X	
S-HDL-CE_%	Cholesterol esters to total lipids ratio in small HDL	Lipid and lipoprotein	X	X	X	X	
S-HDL-FC_%	Free cholesterol to total lipids ratio in small HDL	Lipid and lipoprotein	X	X	X	X	
S-HDL-TG_%	Triglycerides to total lipids ratio in small HDL	Lipid and lipoprotein	X	X	X	X	
VLDL-D	Mean diameter for VLDL particles	Lipid and lipoprotein	X	X	X		
LDL-D	Mean diameter for LDL particles	Lipid and lipoprotein	X	X	X	X	X
HDL-D	Mean diameter for HDL particles	Lipid and lipoprotein	X	X	X	X	X
Serum-C	Serum total cholesterol	Lipid and lipoprotein	X	X	X	X	X
VLDL-C	Total cholesterol in VLDL	Lipid and lipoprotein	X	X	X	X	X
Remnant-C	Remnant cholesterol (non-HDL, non-LDL-cholesterol)	Lipid and lipoprotein	X	X	X	X	X
LDL-C	Total cholesterol in LDL	Lipid and lipoprotein	X	X	X	X	X
HDL-C	Total cholesterol in HDL	Lipid and lipoprotein	X	X	X	X	X
HDL2-C	Total cholesterol in HDL2	Lipid and lipoprotein	X	X	X	X	X
HDL3-C	Total cholesterol in HDL3	Lipid and lipoprotein	X	X	X	X	X
EstC	Esterified cholesterol	Lipid and lipoprotein	X	X	X	X	X
FreeC	Free cholesterol	Lipid and lipoprotein	X	X	X	X	X
Serum-TG	Serum total triglycerides	Lipid and lipoprotein	X	X	X	X	X
VLDL-TG	Triglycerides in VLDL	Lipid and lipoprotein	X	X	X	X	X
LDL-TG	Triglycerides in LDL	Lipid and lipoprotein	X	X	X	X	X
HDL-TG	Triglycerides in HDL	Lipid and lipoprotein	X	X	X	X	X
DAG	Diacylglycerol	Lipid and lipoprotein	X	X	X		
DAG/TG	Ratio of diacylglycerol to triglycerides	Lipid and lipoprotein	X	X	X	X	
TotPG	Total phosphoglycerides	Lipid and lipoprotein	X	X	X		
TG/PG	Ratio of triglycerides to phosphoglycerides	Lipid and lipoprotein	X	X	X	X	
PC	Phosphatidylcholine and other cholines	Lipid and lipoprotein	X	X	X		
SM	Sphingomyelins	Lipid and lipoprotein	X	X	X		
TotCho	Total cholines	Lipid and lipoprotein	X	X	X		
ApoA1	Apolipoprotein A--I *	Lipid and lipoprotein	X	X	X		
ApoB	Apolipoprotein B *	Lipid and	X	X	X		

Traits	Description	Supergroup	Single point tests	Gene tests	Gene-set tests	Enrichment near GWAS signals tests	Tails tests
		lipoprotein					
ApoB/ApoA1	Ratio of apolipoprotein B to apolipoprotein A--I	Lipid and lipoprotein	X	X	X		
TotFA	Total fatty acids	Lipid and lipoprotein	X	X	X		
FALen	Estimated description of fatty acid chain length, not actual carbon number	Lipid and lipoprotein	X	X	X		
UnsatDeg	Estimated degree of unsaturation	Lipid and lipoprotein	X	X	X		
DHA	22:6, docosahexaenoic acid	Lipid and lipoprotein	X	X	X		
LA	18:2, linoleic acid	Lipid and lipoprotein	X	X	X		
CLA	Conjugated linoleic acid	Lipid and lipoprotein	X	X	X		
FAw3	Omega--3 fatty acids	Lipid and lipoprotein	X	X	X		
FAw6	Omega--6 fatty acids	Lipid and lipoprotein	X	X	X		
PUFA	Polyunsaturated fatty acids	Lipid and lipoprotein	X	X	X		
MUFA	Monounsaturated fatty acids; 16:1, 18:1	Lipid and lipoprotein	X	X	X		
SFA	Saturated fatty acids	Lipid and lipoprotein	X	X	X		
DHA/FA	Ratio of 22:6 docosahexaenoic acid to total fatty acids	Lipid and lipoprotein	X	X	X		
LA/FA	Ratio of 18:2 linoleic acid to total fatty acids	Lipid and lipoprotein	X	X	X		
CLA/FA	Ratio of conjugated linoleic acid to total fatty acids	Lipid and lipoprotein	X	X	X		
FAw3/FA	Ratio of omega--3 fatty acids to total fatty acids	Lipid and lipoprotein	X	X	X		
FAw6/FA	Ratio of omega--6 fatty acids to total fatty acids	Lipid and lipoprotein	X	X	X		
PUFA/FA	Ratio of polyunsaturated fatty acids to total fatty acids	Lipid and lipoprotein	X	X	X		
MUFA/FA	Ratio of monounsaturated fatty acids to total fatty acids	Lipid and lipoprotein	X	X	X		
SFA/FA	Ratio of saturated fatty acids to total fatty acids	Lipid and lipoprotein	X	X	X		
Ala	Alanine	Aminoacid	X	X	X		
Gln	Glutamine	Aminoacid	X	X	X		
Gly	Glycine	Aminoacid	X	X	X		
His	Histidine	Aminoacid	X	X	X		
Ile	Isoleucine	Aminoacid	X	X	X		
Leu	Leucine	Aminoacid	X	X	X		
Val	Valine	Aminoacid	X	X	X		
Phe	Phenylalanine	Aminoacid	X	X	X		
Tyr	Tyrosine	Aminoacid	X	X	X		
AcAce	Acetoacetate	Ketone bodies	X	X	X		
Crea	Creatinine	Fluid balance	X	X	X		
Alb	Albumin	Fluid balance	X	X	X		
Gp	Glycoprotein acetyls, mainly a1-acid glycoprotein	Inflammation	X	X	X		

Table 3.1: List of traits and analyses where they were used

3.3.8 Gene-set analyses

To perform gene set analysis I obtained a curated gene-disease list from DisGeNET [299, 300] and gene lists of metabolic pathways from KEGG [301-303] and Reactome [304, 305]. The gene-disease list obtained from DisGeNET, combines expert curated gene-disease associations from the following databases: a) CTD (Comparative Toxicogenomics Database); b) UNIPROT; c) ORPHANET (an online rare disease and orphan drug data base); d) PSYGENET (Psychiatric disorders Gene association NETwork); and e) HPO (Human Phenotype Ontology). I limited analysis to gene sets with more than three genes resulting in 7,150 total gene sets to test. Finally, I extracted loss-of-function variants from genes in the gene sets and ran SKAT-O (method="optimal.adj") for each of the traits. Similarly to the gene-based analysis, I used WES data as discovery, and took signals forward for validation in WGS if $p < 0.01$. Covariate selection for correlated traits was performed as described in the gene-based analyses (**Methods 3.3.7**). The gene-set-wide significance threshold was calculated by first estimating the effective number of gene sets tested given the high overlap amongst them. Using PCA I estimated that 1094 PCs explain > 95% of the variance in gene sets. The significance threshold was therefore calculated as: $0.05/(1094*19)=2.41\text{x}10^{-6}$ where 19 corresponds to the effective number of phenotypes tested as described above. A signal was considered to replicate if after meta-analysis: i) it met the defined gene-set-wide significance threshold ($p_{meta} < 2.41\text{x}10^{-6}$); ii) >2 variants were tested; iii) it was nominally significant ($p_{validation}<0.05$) in the unadjusted test for WGS (i.e without adjusting for correlated traits).

3.3.9 Genes near GWAS signals

GWAS catalog data files (release 27-09-2017) were downloaded from https://www.ebi.ac.uk/gwas/docs/file-downloads [79]. I focused on GWAS loci associated with HDL cholesterol, LDL cholesterol, total cholesterol and triglycerides. I extracted all reported genes for GWAS loci that were associated at genome-wide significance ($p<5x10^{-8}$) excluding cases where the "REPORTED GENE" value was: i) NR (not reported); ii) intergenic; iii) APO(APOE) cluster; iv) HLA-area (**Table 3.2**). For this analysis, I ran SKAT-O using the optimal unified approach (method="optimal.adj") on the four gene sets (HDLC reported, LDLC reported, TC reported, TG reported, **Table 3.2**). The list of genes known to be involved in conditions leading to abnormal lipid levels was created extracting relevant genes from the DisGeNET and Reactome gene lists. Afterwards, I conducted a manual review of the published literature to remove genes where functional work in mouse or human has revealed a direct role of the gene in HDL metabolism (**Table 3.2**). The search terms used were "[*gene name*] loss of function HDL" and "[*gene name*] knockout HDL". Significance threshold ($p < 0.005$) was determined by correcting for 10 PCs explaining >95% of the variance of the traits used in this analysis.

HDLC reported*	HDLC reported (known removed)**	LDLC reported*	TC reported*	TG reported*	Known genes
ABCA1	*ACAD11*	*ABCG5*	*ABCA1*	*AFF1*	*ABCA1*
ABCA8	*ADH5*	*ABCG8*	*ABCB11*	*AKR1C4*	*ABCA8*
AC016735.2	*ALDH1A2*	*ABO*	*ABCG5*	*ALDH2*	*AC016735.2*
ACAD11	*ANGPTL1*	*ACAD11*	*ABCG8*	*ANGPTL3*	*ANGPTL4*
ADH5	*ATG7*	*ANGPTL3*	*ABO*	*ANGPTL4*	*ANGPTL8*
ALDH1A2	*CITED2*	*APOA1*	*ADAMTS3*	*APOA1*	*APOA1*
ANGPTL1	*CMIP*	*APOB*	*ANGPTL3*	*APOA5*	*APOA5*
ANGPTL4	*COBLL1*	*APOC1*	*APOA1*	*APOB*	*APOB*
ANGPTL8	*COPB1*	*APOE*	*APOB*	*APOC1*	*APOC3*

HDLC reported*	HDLC reported (known removed)**	LDLC reported*	TC reported*	TG reported*	Known genes
APOA1	*CPS1*	*BRAP*	*APOE*	*APOE*	*APOE*
APOA5	*DAGLB*	*BRCA2*	*ASAP3*	*BAI3*	*ARL15*
APOB	*FADS1*	*CELSR2*	*BRAP*	*LMBRD1*	*C12orf51*
APOC3	*FAM13A*	*CETP*	*C6orf106*	*CAPN3*	*C6orf106*
APOE	*GPAM*	*CILP2*	*CELSR2*	*CCR6*	*CD300LG*
ARL15	*GSK3B*	*CMTM6*	*CETP*	*CEP68*	*CD36*
ATG7	*HAS1*	*CSNK1G3*	*CILP2*	*CETP*	*CETP*
C12orf51	*IKZF1*	*CYP7A1*	*CMTM6*	*CILP2*	*FTO*
C6orf106	*KAT5*	*DLG4*	*CSNK1G3*	*CITED2*	*GALNT2*
CD300LG	*LACTB*	*DNAH11*	*CYP7A1*	*COBLL1*	*HNF4A*
CD36	*LRP4*	*EHBP1*	*DLG4*	*CTF1*	*IGHVII-33-1*
CETP	*LRRC29*	*FAM117B*	*DNAH11*	*CYP26A1*	*IRS1*
CITED2	*MADD*	*FN1*	*DOCK7*	*DNAH17*	*KLF14*
CMIP	*MC4R*	*FRK*	*ERGIC3*	*DOCK7*	*LCAT*
COBLL1	*MLXIPL*	*GATA6*	*EVI5*	*ERGIC3*	*LILRA3*
COPB1	*MVK*	*GPAM*	*FAM117B*	*FADS1*	*LIPC*
CPS1	*MYL2*	*HFE*	*FN1*	*FRMD5*	*LIPG*
DAGLB	*OR4C46*	*HLA*	*FRK*	*FTO*	*LOC100996634*
FADS1	*PDE3A*	*HLA-C*	*GCKR*	*GALNT2*	*LOC55908*
FAM13A	*PEPD*	*HMGCR*	*GPAM*	*GCKR*	*LPA*
FTO	*PGS1*	*HNF1A*	*GPR146*	*GPR85*	*LPL*
GALNT2	*RBM5*	*HPR*	*HBS1L*	*HLA*	*LRP1*
GPAM	*RSPO3*	*IDOL*	*HFE*	*INSR*	*MSL2L1*
GSK3B	*SBNO1*	*INSIG2*	*HLA*	*IRS1*	*PABPC4*
HAS1	*SEMA3C*	*IRF2BP2*	*HLA-C*	*JMJD1C*	*PLTP*
HNF4A	*SETD2*	*LDLR*	*HMGCR*	*KLHL8*	*PPP1R3B*
IGHVII-33-1	*SLC39A8*	*LDLRAP1*	*HNF1A*	*LIPC*	*PRKAG3*
IKZF1	*SNX13*	*LOC84931*	*HNF4A*	*LPA*	*RMI2*
IRS1	*STAB1*	*LPA*	*HPR*	*LPL*	*RP-11-115*
KAT5	*STARD3*	*LRPAP1*	*IDOL*	*LRP1*	*SCARB1*
KLF14	*TMEM176A*	*MAFB*	*INSIG2*	*LRPAP1*	*SIK3*
LACTB	*TRPS1*	*MIR148A*	*IRF2BP2*	*MAP3K1*	*TRIB1*
LCAT	*UBASH3B*	*MOSC1*	*KCNK17*	*MAU2*	*TTC39B*
LILRA3	*ZNF648*	*MTHFD2L*	*LDLR*	*MET*	*UBE2L3*
LIPC		*MTMR3*	*LDLRAP1*	*MIR148A*	*VEGFA*
LIPG		*MYLIP*	*LIPC*	*MLXIPL*	*ZNF664*
LOC100996634		*NCAN*	*LIPG*	*MPP3*	
LOC55908		*NPC1L1*	*LPA*	*MSL2L1*	
LPA		*OSBPL7*	*LRPAP1*	*NAT2*	

HDLC reported*	HDLC reported (known removed)**	LDLC reported*	TC reported*	TG reported*	Known genes
LPL		*PCSK9*	*MAFB*	*PDXDC1*	
LRP1		*PFAS*	*MAMSTR*	*PEPD*	
LRP4		*PLEC1*	*MARCH8*	*PINX1*	
LRRC29		*PPARA*	*MIR148A*	*PLA2G6*	
MADD		*PPARG*	*MOSC1*	*PLTP*	
MC4R		*PPP1R3B*	*MTHFD2L*	*PROX1*	
MLXIPL		*SMARCA4*	*MYLIP*	*RSPO3*	
MSL2L1		*SNX5*	*NAT2*	*SIK3*	
MVK		*SORT1*	*NCAN*	*TIMD4*	
MYL2		*SOX17*	*NPC1L1*	*TM4SF5*	
OR4C46		*SPTLC3*	*OSBPL7*	*TP53BP1*	
PABPC4		*ST3GAL4*	*PCSK9*	*TRIB1*	
PDE3A		*TIMD4*	*PHLDB1*	*TYW1B*	
PEPD		*TOP1*	*PLEC1*	*VEGFA*	
PGS1		*TRIB1*	*PPARA*	*ZNF664*	
PLTP		*VLDLR*	*PPARG*		
PPP1R3B		*ZNF274*	*PPP1R3B*		
PRKAG3			*PXK*		
RBM5			*RAB3GAP1*		
RMI2			*RAF1*		
RP-11-115			*RP11-115*		
J16.1			*J16.1*		
RSPO3			*SAMM50*		
SBNO1			*SNX5*		
SCARB1			*SORT1*		
SEMA3C			*SOX17*		
SETD2			*SPTY2D1*		
SIK3			*ST3GAL4*		
SLC39A8			*TIMD4*		
SNX13			*TMEM57*		
STAB1			*TOP1*		
STARD3			*TRIB1*		
TMEM176A			*TRPS1*		
TRIB1			*TTC39B*		
TRPS1			*UBASH3B*		
TTC39B			*UGT1A1*		
UBASH3B			*VLDLR*		
UBE2L3					
VEGFA					

HDLC reported*	HDLC reported (known removed)**	LDLC reported*	TC reported*	TG reported*	Known genes
ZNF648					
ZNF664					

Table 3.2: Gene sets used for enrichment of genes near GWAS signals analyses. HDL reported -Genes reported associated with "HDL cholesterol" unambiguously ; HDLC reported (known removed) - Genes reported associated with "HDL cholesterol" unambiguously but with known genes involved in HDL metabolism or lipid abnormalities removed; LDLC reported - Genes reported associated with "LDL cholesterol" unambiguously; TC reported - Genes reported associated with "Cholesterol, total" unambiguously; TG reported - Genes reported associated with "Triglycerides" unambiguously; Known genes - Genes removed for sensitivity analysis that are known to be involved in lipid abnormalities or HDL metabolism based on literature review; *Gene sets used in analyses running SKAT-O on gene sets.; **Gene sets used in sensitivity analyses.

3.3.10 Analysis of tails of phenotype distribution

For this analysis, I used all lipoprotein and lipid traits but excluded derived measures (lipid ratios) resulting in 106 traits (**Table 3.1**). I focused on likely deleterious missense and loss-of-function variation in lipid metabolism and disease gene sets (**Table 3.3**) with an allele count <10 in each dataset. I chose an arbitrary cutoff of 10 individuals with the highest and lowest values for the traits to define tails for all 106 traits.

Gene Set	Source
Abnormality_of_lipid_metabolism	DisGeneNet
Dyslipidaemias	DisGeneNet
HDL_assembly	Reactome
HDL_clearance	Reactome
HDL_remodeling	Reactome
Hyperlipidaemia	DisGeneNet
Hypertriglyceridaemia_CTD	DisGeneNet
Hypertriglyceridaemia_HPO	DisGeneNet
LDL_clearance	Reactome
LDL_remodeling	Reactome
Triglyceride_biosynthesis	Reactome
Triglyceride_catabolism	Reactome
Triglyceride_metabolism	Reactome
VLDL_assembly	Reactome
VLDL_clearance	Reactome

Table 3.3: List of gene sets used for tails analyses.

Given the high phenotypic correlation of these traits, there was a high overlap of individuals at the tails of the distributions so I removed traits that shared >=8 individuals with any other trait reducing the number of tested traits to 50. For each trait, total deleterious allele count from each gene set for upper and lower tails was obtained and an empirical *p* was calculated by performing 10,000 permutations extracting 10 random individuals from the phenotype distribution and counting the number of deleterious alleles from the gene set. The significance threshold (p = 0.00037) was chosen by correcting for 9 PCs explaining >95% of the traits variance and 15 pathways. Meta-analysis was done using Stouffer's method [306] as implemented in the metap package [307] in R.

3.4 Results

3.4.1 Single point analyses

I first explored whether I could recapitulate known associations with NMR traits, as well as, potentially identify novel associations with rarer variants not previously tested in GWAS arrays. To this end, I performed single-point association analysis for 226 NMR metabolic biomarkers using WES data from 3,741 healthy blood donors from the INTERVAL cohort as a discovery dataset (**Methods 3.3.6**). Power calculations showed very limited power to detect associations for variants on the rare allele frequency spectrum with this sample size (power=4.6% to find an association with $p<1x10^{-5}$ -threshold to take forward for validation- with beta=1 and variant with MAF=0.1%). I therefore focused on variants with MAF>=0.1%. After association analyses for all traits I took forward for validation 494 variants associated with at least one trait with $p<1x10^{-5}$. I performed validation using whole-genome sequence (WGS) data from 3,401 independent individuals from the same cohort. After meta-analysis,

34 unique loci were associated with at least one trait (**Table 3.4**). All of these associations had already been previously described [38, 173, 308].

Rsid	Gene	most severe consequence	top trait	EA	NEA	discov p	validation p	meta-p	beta	se	EAF	n assoc traits
rs1047891	*CPS1*	missense_variant (Thr1412Asn)	Gly	a	c	$1.48x10^{-68}$	$4.47x10^{-54}$	$2.09x10^{-125}$	0.42	0.02	32.47%	1
rs1077834	*LIPC,ALDH1A2*	intron_variant	L-HDL-TG	t	c	$2.52x10^{-16}$	$6.90x10^{-21}$	$1.11x10^{-35}$	-0.25	0.02	21.41%	35
rs11076176	*CETP*	intron_variant	M-HDL-TG	t	g	$5.82x10^{-7}$	$6.62x10^{-6}$	$1.65x10^{-11}$	-0.15	0.02	16.92%	6
rs11591147	*PCSK9*	missense_variant (Arg46Leu)	IDL-FC	t	g	$7.31x10^{-12}$	$2.20x10^{-5}$	$2.96x10^{-15}$	-0.48	0.06	1.73%	45
rs116843064	*ANGPTL4*	missense_variant (Glu40Lys)	S-VLDL-TG	a	g	$7.81x10^{-7}$	$2.67x10^{-6}$	$9.11x10^{-12}$	-0.40	0.06	1.89%	17
rs1184865	*DOCK7*	intron_variant	M-HDL-TG	a	g	$6.59x10^{-6}$	$5.66x10^{-5}$	$1.45x10^{-9}$	-0.10	0.02	36.13%	1
rs12191266	*SLC16A10*	intron_variant	Tyr	t	c	$4.68x10^{-6}$	$2.42x10^{-5}$	$4.48x10^{-10}$	-0.15	0.02	14.43%	1
rs1260326	*GCKR*	missense_variant (Leu446Pro)	MUFA	t	c	$1.20x10^{-6}$	$5.31x10^{-6}$	$2.61x10^{-11}$	0.12	0.02	39.85%	17
rs138326449	*APOC3*	splice_donor_variant (2^{nd} exon)	S-VLDL-TG	a	g	$7.91x10^{-6}$	$8.80x10^{-6}$	$2.90x10^{-10}$	-1.10	0.17	0.23%	6
rs17231506	*CETP*	upstream_gene_variant	HDL2-C	t	c	$6.73x10^{-17}$	$4.65x10^{-18}$	$1.35x10^{-33}$	0.21	0.02	31.83%	38
rs174476	*RAB3IL1*	intron_variant	UnsatDeg	t	c	$2.05x10^{-9}$	$1.48x10^{-5}$	$1.95x10^{-13}$	0.12	0.02	41.71%	1
rs174547	*FADS1,FADS2*	intron_variant	UnsatDeg	t	c	$1.03x10^{-41}$	$5.96x10^{-38}$	$9.02x10^{-80}$	0.33	0.02	33.71%	8
rs174602	*FADS2*	non_coding_transcript_exon_variant	UnsatDeg	t	c	$1.21x10^{-11}$	$5.64x10^{-7}$	$4.97x10^{-17}$	0.17	0.02	20.16%	2
rs1912826	*KLKB1*	intron_variant	His	a	g	$7.80x10^{-11}$	$5.54x10^{-9}$	$2.04x10^{-18}$	0.15	0.02	48.89%	2
rs2072560	*APOA5*	intron_variant	XS-VLDL-TG_%	t	c	$1.15x10^{-8}$	$2.06x10^{-7}$	$1.07x10^{-14}$	0.27	0.04	5.90%	30
rs2228671	*LDLR*	non_coding_transcript_exon_variant	IDL-FC	t	c	$2.04x10^{-7}$	$6.27x10^{-7}$	$5.55x10^{-13}$	-0.18	0.03	12.26%	38
rs2295601	*ELOVL2*	synonymous_variant	DHA/FA	a	g	$1.54x10^{-10}$	$6.61x10^{-9}$	$4.69x10^{-18}$	-0.17	0.02	22.90%	2
rs2575876	*ABCA1*	intron_variant	HDL3-C	a	g	$1.92x10^{-6}$	$8.30x10^{-8}$	$8.12x10^{-13}$	-0.14	0.02	24.65%	1
rs2657879	*GLS2*	3_prime_UTR_variant	Gln	a	g	$1.16x10^{-11}$	$1.72x10^{-15}$	$1.50x10^{-25}$	0.23	0.02	18.07%	1
rs283813	*PVRL2*	intron_variant	S-LDL-C_%	a	t	$3.08x10^{-8}$	$1.20x10^{-5}$	$2.20x10^{-12}$	-0.23	0.03	6.90%	22
rs28399637	*BCAM*	intron_variant	S-LDL-CE_%	a	g	$4.95x10^{-9}$	$8.59x10^{-7}$	$2.02x10^{-14}$	0.14	0.02	31.77%	25
rs28399654	*BCAM*	missense_variant (Val196Ile)	S-LDL-C_%	a	g	$1.38x10^{-11}$	$8.80x10^{-8}$	$8.29x10^{-18}$	-0.40	0.05	3.37%	34
rs328	*LPL*	stop_gained (Ser474Ter)	TG/PG	c	g	$1.08x10^{-8}$	$1.44x10^{-7}$	$7.00x10^{-15}$	0.22	0.03	10.09%	19
rs3798220	*LPA*	missense_variant (Ile1891Met)	XL-VLDL-CE	t	c	$3.04x10^{-6}$	$4.55x10^{-13}$	$6.15x10^{-17}$	0.55	0.07	1.76%	16
rs386606006	*APOB*	synonymous_variant	ApoB	a	g	$9.37x10^{-6}$	$2.97x10^{-6}$	$1.17x10^{-10}$	0.11	0.02	48.80%	1

Rsid	Gene	most severe consequence	top trait	EA	NEA	discov p	validation p	meta-p	beta	se	EAF	n assoc traits
rs429358	*APOE*	missense_variant (Cys130Arg)	S-LDL-PL_%	t	c	$9.37x10^{-17}$	$1.20x10^{-17}$	$4.69x10^{-33}$	0.27	0.02	15.07%	61
rs435306	*PLTP*	intron_variant	L-HDL-PL_%	t	g	$4.90x10^{-7}$	$4.17x10^{-7}$	$8.84x10^{-13}$	0.14	0.02	25.50%	1
rs4804573	*KANK2*	3_prime_UTR_variant	S-LDL-PL_%	a	g	$1.49x10^{-7}$	$6.26x10^{-5}$	$4.66x10^{-11}$	0.11	0.02	47.05%	9
rs5880	*CETP*	missense_variant (Ala390Pro)	HDL-C	c	g	$7.97x10^{-7}$	$3.05x10^{-8}$	$1.17x10^{-13}$	-0.28	0.04	4.87%	8
rs61937878	*HAL*	missense_variant (Val549Met)	His	t	c	$7.41x10^{-14}$	$3.75x10^{-8}$	$2.01x10^{-20}$	0.95	0.10	0.66%	1
rs693672	*FADS3*	intron_variant	UnsatDeg	t	c	$1.44x10^{-10}$	$1.36x10^{-9}$	$8.97x10^{-19}$	-0.19	0.02	16.76%	1
rs7412	*APOE*	missense_variant (Arg176Cys)	S-LDL-CE_%	t	c	$8.55x10^{-63}$	$1.82x10^{-58}$	$5.97x10^{-124}$	-0.71	0.03	7.80%	89
rs76075198	*CEACAM19*	synonymous_variant	S-LDL-CE_%	t	c	$6.76x10^{-7}$	$5.25x10^{-8}$	$1.72x10^{-13}$	-0.41	0.06	2.20%	10
rs7679	*PCIF1*	3_prime_UTR_variant	L-HDL-PL_%	t	c	$5.43x10^{-18}$	$1.14x10^{-19}$	$2.23x10^{-36}$	-0.27	0.02	18.05%	19

Table 3.4: Single point association analyses results. Most severe consequence=most severe consequence predicted by VEP on CANONICAL transcript. top trait=trait with the lowest p-value. EA=effect allele. NEA=non-effect allele discov p=p-value for top trait in discovery cohort (WES). validation p=p-value for top trait in validation cohort (WGS). meta-p= p-value for top trait. beta=beta for top trait after meta-analysis. se=se for top trait after meta-analysis. EAF=effect allele frequency. n assoc traits=number of associated traits.

3.4.2 Gene-based analyses

I next sought to discover new gene-trait associations using rare-variant aggregate tests. After running association tests using two nested approaches to group rare variants (LoF and MCAP+LoF, **Methods 3.3.7**), genes were taken forward for validation if they reached the arbitrary threshold of p <5x10^{-3} (**Supplementary Tables 1-2 of Riveros-Mckay et al (in preparation, Appendix B)**). A burden test was used when testing only LoF whereas the optimal unified approach was used when adding predicted deleterious missense variants (MCAP+LoF). This is because I expected most high confidence LoF variants to influence a trait with the same direction of effect and therefore the burden test should be better powered than the optimal unified approach to detect an association. When including missense variants one could expect different directions of effect and therefore the optimal unified approach should be better powered. As previously suggested, to boost discovery power I adjusted for correlated metabolic biomarkers [309, 310]. However, to minimise the possible collider bias this could incur, I only did this at the validation stage. This was to ensure there was at least suggestive evidence for association in the discovery stage without adjusting for any metabolite (**Methods 3.3.7**). After meta-analysis, five genes (*APOB*, *APOC3*, *PCSK9*, *PAH*, *HAL*) associated with 92 different traits with p < 1.32x10^{-7}, which is the stringent significance threshold after correcting for the effective number of tested phenotypes (**Table 3.5, Methods 3.3.7**). All five genes have been previously associated with their respective traits [38, 308, 311]. As expected, I found that there was a significant increase in the strength of association signal for traits for which I used other correlated traits as covariates when compared to the unadjusted tests [309, 310]**,** with the most notable example being a >30 order of magnitude increase in association strength for *PAH*

and phenylalanine (**Table 3.5**). In total, 32 of the 92 known gene-trait associations met the stringent significance threshold (p<1.32x10^{-7}) only after adjusting for correlated traits (**Supplementary Tables 1-2 of Riveros-Mckay et al (in preparation, Appendix B)**).

LoF								
Gene	**Top trait**	**p-value (covs)**	**p-value (raw)**	**N WES**	**N WGS**	**N overlap**	**N traits associated**	**Driven by single variant?**
APOB	**IDL-TG**	**3.20x10^{-13}**	**1.72x10^{-10}**	**6**	**5**	**0**	**45 (57)**	**No**
APOC3	**XS-VLDL-TG**	**6.10x10^{-13}**	**3.58x10^{-12}**	**3**	**2**	**2**	**46 (56)**	**No**
PAH	**Phe**	**5.82x10^{-11}**	**8.25x10^{-3}**	**4**	**3**	**1**	**1 (1)**	**Yes**
MCAP+LoF								
Gene	**Top trait**	**p-value (covs)**	**p-value (raw)**	**N WES**	**N WGS**	**N overlap**	**N traits associated**	**Driven by single variant?**
PAH	**Phe**	**8.33x10^{-63}**	**1.67x10^{-28}**	**39**	**41**	**18**	**1 (1)**	**No**
HAL	**His**	**NA**	**3.72x10^{-42}**	**48**	**37**	**22**	**1 (1)**	**No**
APOC3	**XS-VLDL-TG**	**5.46x10^{-11}**	**2.15x10^{-10}**	**6**	**6**	**3**	**26 (40)**	**No**
PCSK9	**IDL-FC**	**2.39x10^{-10}**	**1.11x10^{-7}**	**15**	**17**	**3**	**29 (34)**	**No**
ACSL1	IDL-P	1.82x10^{-7}	1.76x10^{-4}	4	6	2	0 (1)	Yes
MYCN	M-VLDL-L	6.20x10^{-7}	3.97x10^{-6}	8	8	3	0 (5)	No
ALDH1L1	Gly	NA	4.56x10^{-7}	39	38	19	0 (1)	No
SCARB1	XL-HDL-FC	NA	6.93x10^{-7}	25	18	10	0 (6)	No
FBXO36	IDL-CE_%	NA	1.98x10^{-6}	5	2	1	0 (1)	Yes
B4GALNT3	L-VLDL-FC_%	NA	7.59x10^{-7}	28	22	13	0 (1)	No
LIPC	XXL-VLDL-C_%	NA	9.03x10^{-7}	28	29	11	0 (2)	No

Table 3.5:Genes significantly associated (p<2.5x10-6) with at least one trait in gene-based analyses focusing on loss-of-function (LoF) or predicted deleterious missense by M-CAP plus loss-of-function (MCAP+LoF). Genes that meet gene-level significance after adjusting for multiple phenotypes (p<1.32x10^{-7}) are highlighted in bold. Top trait: trait with the smallest p-value after meta-analysis adjusting for correlated metabolites. p-value (covs): p-value of meta-analysis after adjusting for correlated metabolites for top trait. If NA, this analysis was not performed for this trait due to no metabolic biomarkers meeting the criteria to be included as covariates in meta-analysis. p-value (raw): p-value of meta-analysis without adjusting for correlated metabolites for top trait. N WES: number of tested variants in WES. N WGS: number of tested variants in WGS. N overlap: number of variants present in both WES and WGS. N traits associated: number of traits that meet gene-level significance after adjusting for multiple phenotypes (p<1.32x10^{-7}), traits meeting standard gene-level significance (2.5x10^{-6}) in parenthesis. Driven by single variant?: Yes if after conditioning on top associated variant the meta-analysis association disappears (p>0.05). IDL-TG: Triglycerides in IDL. XS-VLDL-TG: Triglycerides in very small VLDL. Phe: Phenylalanine. His: Histidine. IDL-FC: Free cholesterol in IDL. IDL-P: Concentration of IDL particles. M-VLDL-L: Total lipids in medium VLDL. Gly:Glycine. XL-HDL-FC: Free cholesterol in very large HDL. IDL-CE_%: Cholesterol esters to total lipids ratio in IDL. L-VLDL-FC%: Free cholesterol to total lipids ratio in large VLDL. XXL-VLDL-C_%: Total cholesterol to total lipids ratio in extremely large VLDL.

In addition to established genes, I found 15 gene-trait associations in seven genes meeting standard gene-level significance before adjusting for multiple traits ($p< 2.5x10^{-6}$) which also had nominal evidence of association in the validation cohort ($p< 0.05$). Nine of these were gene-trait associations in three established genes (*ALDH1L1*, *SCARB1*, *LIPC*, **Table 3.5**), suggesting that other results achieving this significance threshold may warrant being prioritised for additional follow-up to establish their validity. In particular amongst the remaining four genes, the association between IDL particle concentration (IDL-P) and *ACSL1* ($p = 1.82x10^{-7}$), as well as, the associations of multiple very-low-density lipoprotein (VLDL) traits to *MYCN* (min $p = 6.20x10^{-7}$) merit further exploration as both genes have been previously linked to lipid metabolism in mouse studies [312-314].

3.4.3 Gene set analyses

To find links between predicted loss-of-function rare variants and metabolic biomarker biology, I next explored associations of these variants in 7,150 gene sets. To this end, I used two biological pathway databases (Reactome, KEGG) and one database that contains expert curated disease associated genes (DisGeNET) (**Methods3.3.8**). Gene set analysis yielded 163 gene-set-trait associations with 14 unique gene sets (**Supplementary Table 4 of Riveros-Mckay et al (in preparation, Appendix B)**). Given that 143 gene-set-trait associations were with 13 gene sets that included two genes with a well-established role in lipid biology (*APOB* and *APOC3*), I repeated the test removing variants in these genes. After removal, there is residual evidence of association ($p<0.05$) in 102 of 143 gene-set-trait signals representing 12 of 13 gene sets. Of the 163 gene-set-trait associations, the remaining 20 gene-set-trait associations (in gene sets not containing either *APOB* or *APOC3)* represent associations of

various lipoprotein related metabolic biomarkers with the "regulation of pyruvate dehydrogenase (PDH) complex" pathway in REACTOME (R-HSA-204174, min p= 7.85x10^{-7}, trait=phospholipids in intermediate density lipoproteins (IDL-PL), **Table 3.6**). These associations encompassed 12 LoF variants in WES and four in WGS (**Figure 3.1**). Upon further inspection, I found that most variants in this pathway were contributing to the association suggesting the signal was not driven by a single gene, in addition they all have the same direction of effect (i.e. the rho(ρ) value in the SKAT-O test was one in both the WES and the WGS analyses). Two variants were of particular interest as they were present in both WES and WGS datasets, rs113309941 in Pyruvate Dehydrogenase Complex Component X (*PDHX*), and rs201013643 in Pyruvate Dehydrogenase Phosphatase Regulatory Subunit (*PDPR*). In *PDHX, rs*113309941 leads to a premature stop mutation (Gln248Ter), it has an allele count (AC) of one in each of WES and WGS, and is very rare in gnomAD[1]. rs201013643 in PDRP also leads to a premature stop (Arg714Ter) and is present in a single heterozygous individual in the WES dataset and two heterozygous in the WGS. This variant is also rare in gnomAD[2]. The five individuals with these two variants had higher than average values (upper percentile range from 44.1% to 0.03%) for measurements that are CVD risk factors such as cholesterol in intermediate-density lipoproteins (IDL-C) and low-density lipoproteins (LDL-C) suggesting these variants may have a deleterious impact on lipid metabolism and cardiovascular risk. Notably, one of the carriers of the *PDHX* Gln248Ter variant was in the top 0.03% for LDL-C in INTERVAL (4.086 mmol/l) and had no predicted deleterious missense mutations in known hypercholesterolaemia genes *PCSK9, APOB* or *LDLR* suggesting this novel protein truncating variant may be contributing to their high LDL-

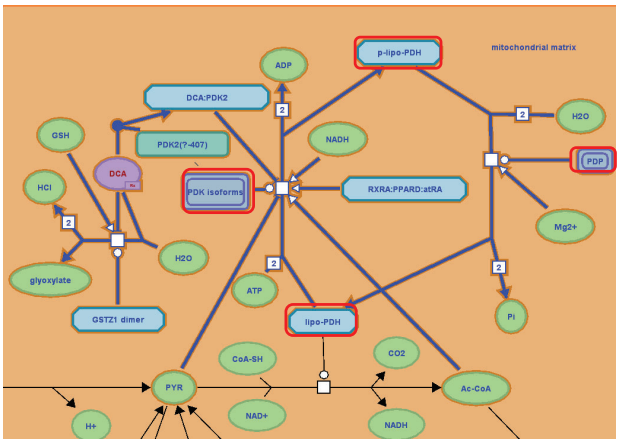
[1] AC (all gnomAD)=3, allele number (AN) (all gnomAD)=246,116, AC (Non-Finnish European (NFE))=2 AN (NFE)=116,604.
[2] AC (all gnomAD)=141, AN (all gnomAD)=275,988, AC (NFE) =8, AN (NFE)=126,382.

C levels. The other carrier was in the top 19.3% percentile of the cohort. None of the genes in this pathway have been previously associated to these traits and therefore this study links these genes collectively to intermediate and low density lipoprotein metabolism and circulating cholesterol for the first time.

Gene set id	Trait	WES p	N WES	WGS p	N WGS	Meta-p	Description	Source
R-HSA-204174	IDL-PL	0.005939	12	0.000503	4	7.85×10^{-7}	Regulation of pyruvate dehydrogenase (PDH) complex	Reactome
R-HSA-204174	M-LDL-PL	0.002671	12	0.000594	4	1.01×10^{-6}	Regulation of pyruvate dehydrogenase (PDH) complex	Reactome
R-HSA-204174	EstC	0.004754	12	0.001175	4	1.09×10^{-6}	Regulation of pyruvate dehydrogenase (PDH) complex	Reactome
R-HSA-204174	IDL-P	0.003992	12	0.000593	4	1.17×10^{-6}	Regulation of pyruvate dehydrogenase (PDH) complex	Reactome
R-HSA-204174	L-LDL-P	0.004822	12	0.000258	4	1.20×10^{-6}	Regulation of pyruvate dehydrogenase (PDH) complex	Reactome
R-HSA-204174	L-LDL-PL	0.004853	12	0.000423	4	1.21×10^{-6}	Regulation of pyruvate dehydrogenase (PDH) complex	Reactome
R-HSA-204174	IDL-L	0.004313	12	0.000574	4	1.21×10^{-6}	Regulation of pyruvate dehydrogenase (PDH) complex	Reactome
R-HSA-204174	SerumC	0.005999	12	0.001071	4	1.24×10^{-6}	Regulation of pyruvate dehydrogenase (PDH) complex	Reactome
R-HSA-204174	L-LDL-L	0.005082	12	0.000275	4	1.35×10^{-6}	Regulation of pyruvate dehydrogenase (PDH) complex	Reactome
R-HSA-204174	IDL-C	0.00475	12	0.001019	4	1.40×10^{-6}	Regulation of pyruvate dehydrogenase (PDH) complex	Reactome
R-HSA-204174	L-LDL-FC	0.00681	12	0.0003	4	1.46×10^{-6}	Regulation of pyruvate dehydrogenase (PDH) complex	Reactome
R-HSA-204174	L-LDL-C	0.006489	12	0.000275	4	1.87×10^{-6}	Regulation of pyruvate dehydrogenase (PDH) complex	Reactome
R-HSA-204174	M-LDL-P	0.006409	12	0.000132	4	1.96×10^{-6}	Regulation of pyruvate dehydrogenase (PDH) complex	Reactome
R-HSA-204174	L-LDL-CE	0.006486	12	0.000277	4	2.01×10^{-6}	Regulation of pyruvate dehydrogenase (PDH) complex	Reactome
R-HSA-204174	S-LDL-L	0.006413	12	0.000115	4	2.13×10^{-6}	Regulation of pyruvate dehydrogenase (PDH) complex	Reactome
R-HSA-204174	S-LDL-P	0.005994	12	0.000113	4	2.13×10^{-6}	Regulation of pyruvate dehydrogenase (PDH) complex	Reactome
R-HSA-204174	M-LDL-L	0.006416	12	0.000164	4	2.13×10^{-6}	Regulation of pyruvate dehydrogenase (PDH) complex	Reactome
R-HSA-204174	LDL-C	0.007809	12	0.000177	4	2.17×10^{-6}	Regulation of pyruvate dehydrogenase (PDH) complex	Reactome
R-HSA-204174	ApoB	0.00504	12	0.000803	4	2.20×10^{-6}	Regulation of pyruvate dehydrogenase (PDH) complex	Reactome
R-HSA-204174	IDL-FC	0.009798	12	0.000399	4	2.22×10^{-6}	Regulation of pyruvate dehydrogenase (PDH) complex	Reactome

Table 3.6: Gene set analyses results. WES p = p-value in WES dataset. N WES = number of variants tested in WES dataset. WGS p = p-value in WGS dataset. N WGS = number of variants tested in WGS dataset. Meta-p = Meta-analysis p-value after removing APO genes from gene sets (APOB and APOC3).

a)



b)

Gene	Consequence	AC
Pyruvate dehydrogenase (PDH) complex		
DLAT	Splice acceptor (2nd exon)	WES=1
DLD	Frameshift (Val212SerfsTer32)	WES=1
PDHA2	Stop gain (Tyr28Ter)	WES=1
PDHA2	Frameshift (Val297GlnfsTer14)	WES=1
PDHA2	Stop gain (Gln78Ter)	WES=1
PDHA2	Frameshift (Lys83IlefsTer20)	WES=1
PDHA2	Stop gain (Tyr80Ter)	WES=1
PDHX	Splice donor (2nd exon)	WES=1
PDHX	Stop gain (Gln248Ter)	WES=1 WGS=1
Pyruvate dehydrogenase phosphatase (PDP)		
PDP2	Frameshift (Asn33IlefsTer5)	WES=1
PDP2	Stop gain (Gln352Ter)	WES=1
PDPR	Stop gain (Trp402Ter)	WES=1
PDPR	Stop gain (Arg714Ter)	WES=1 WGS=2
Pyruvate dehydrogenase kinase (PDK)_		
PDK1	Stop gain (Arg66Ter*)	WGS=1

Figure 3.1: Loss-of-function (LoF) variants in regulation of pyruvate dehydrogenase (PDH) complex pathway. a) Figure adapted from REACTOME pathway browser (https://reactome.org/PathwayBrowser/) [315]. Highlighted in red are protein complexes that carry LoF variants in INTERVAL WES or WGS. b) List of genes, consequences and allele count (AC) of LoF variants in the different protein complexes in the pathway.

3.4.4 Enrichment of rare variant associations in genes near established GWAS signals in lipoprotein related metabolic biomarkers

Next, I conducted analyses to investigate whether genes near GWAS index variants associated with traditional lipid traits (HDL-C, LDL-C, TC and TG) were enriched for rare variant associations computationally predicted to affect protein sequence and function with high resolution lipoprotein measurements, which could suggest enrichment of effector transcripts (i.e. transcripts/genes likely to be causal of the original association) in the gene set. Given that this was a hypothesis driven approach using established signals, to boost discovery power I pooled together both WES and WGS data into a single dataset of 7,179 individuals. First, I extracted from the GWAS catalog (release 27-09-2017) the "reported genes" near signals that have been associated with HDL-C, LDL-C, TC or TG and created four gene sets (**Table 3.2**). I only focused on genes that were reported unambiguously (i.e where only one gene is reported) since for associations where more than one gene is reported, it is possible that only one will be the effector gene and rare variants from the non-effector genes will only add noise to the analysis and therefore reduce power. I grouped rare coding variants in the gene set using two nested approaches (LoF and MCAP+LoF) and ran SKAT-O on the gene sets for 157 lipoprotein and lipid traits. Using this approach I found associations ($p < 0.005$, correcting for effective number of tests, **Methods 3.3.9**) for genes near HDL GWAS signals with 18 HDL-related traits (**Table 3.7**), the strongest association being with esterified cholesterol in extra-large HDL (XL-HDL-CE, $p=2.83 \times 10^{-5}$, MCAP+LoF). Associations ($p < 0.005$, **Methods 3.3.9**) in two extra-large HDL cholesterol related traits remained after removing variants in genes known to be involved in conditions leading to abnormal lipid levels or genes where functional work has shown an effect on HDL-C (**Table 3.7**) suggesting there is a contribution to the phenotypic variance of these traits by rare

coding variants in genes, near GWAS signals, without a known role in HDL metabolism, which may represent novel effector transcripts.

Trait	GWAS signal gene set	LoF p-value	MCAP+LoF p-value	LoF p-value (known removed)	MCAP+LoF p-value (known removed)
HDL2-C	HDL-C	$9.03x10^{-3}$	$4.72x10^{-3}$	$4.73x10^{-1}$	$1.41x10^{-1}$
HDL-D	HDL-C	$6.29x10^{-3}$	$2.55x10^{-3}$	$6.88x10^{-1}$	$3.46x10^{-1}$
L-HDL-C_%	HDL-C	$1.49x10^{-3}$	$6.04x10^{-2}$	$4.45x10^{-1}$	$8.78x10^{-1}$
L-HDL-FC_%	HDL-C	$1.67x10^{-4}$	$5.40x10^{-4}$	$1.40x10^{-1}$	$3.52x10^{-1}$
L-HDL-FC	HDL-C	$9.21x10^{-3}$	$3.14x10^{-3}$	$3.95x10^{-1}$	$2.63x10^{-1}$
L-HDL-TG_%	HDL-C	$2.27x10^{-3}$	$1.30x10^{-1}$	$3.40x10^{-1}$	$7.25x10^{-1}$
M-HDL-TG_%	HDL-C	$6.76x10^{-4}$	$1.18x10^{-3}$	$9.98x10^{-2}$	$7.19x10^{-1}$
S-HDL-TG_%	HDL-C	$4.68x10^{-3}$	$4.37x10^{-3}$	$4.37x10^{-1}$	$7.76x10^{-1}$
S-HDL-TG	HDL-C	$1.61x10^{-3}$	$5.47x10^{-3}$	$3.47x10^{-1}$	$3.73x10^{-1}$
XL-HDL-CE	HDL-C	$2.86x10^{-2}$	$2.83x10^{-5}$	1.00	$3.69x10^{-4}$
XL-HDL-C	HDL-C	$1.85x10^{-2}$	$4.43x10^{-5}$	$8.48x10^{-1}$	$9.03x10^{-4}$
XL-HDL-FC	HDL-C	$6.41x10^{-3}$	$2.44x10^{-4}$	$7.43x10^{-1}$	$1.11x10^{-2}$
XL-HDL-L	HDL-C	$1.14x10^{-2}$	$1.75x10^{-4}$	$7.00x10^{-1}$	$7.07x10^{-3}$
XL-HDL-P	HDL-C	$1.17x10^{-2}$	$1.91x10^{-4}$	$6.92x10^{-1}$	$7.56x10^{-3}$
XL-HDL-PL	HDL-C	$8.07x10^{-3}$	$9.94x10^{-4}$	$5.12x10^{-1}$	$1.11x10^{-1}$

Table 3.7:Significant results (p<0.005) in SKAT-O analysis on gene sets built from lists of genes near established GWAS loci. LoF p-value: SKAT-O results for analysis focusing on loss-of-function variants in gene set. MCAP+LoF p-value: SKAT-O results for analysis focusing on rare missense variants (MAF <1%) predicted to be likely deleterious (M-CAP score >0.025) and loss-of-function variants in gene set. LoF p-value (known removed) = SKAT-O results for LoF approach after removing genes known to be involved in lipoprotein metabolism. MCAP+LoF p-value (known removed) = SKAT-O results for MCAP+LoF approach after removing genes known to be involved in lipoprotein metabolism.

3.4.5 Enrichment of rare variation in tails of the phenotypic distribution of lipoprotein and glyceride related traits

Finally, I aimed to investigate whether individuals at the extreme tail of the phenotype distribution for 106 lipoprotein and lipid traits harboured rare coding variants likely to be contributing to their phenotype. I used the WES dataset as a discovery dataset and the WGS dataset as validation. An arbitrary cutoff of 10 individuals at each tail was used to define the tails for all of the 106 traits (**Methods 3.3.10**). After meta-analysis, I found an enrichment of predicted deleterious rare variation ($p < 0.00037$, **Methods 3.3.10, Table 3.8,**

Supplementary Table 9 of Riveros-Mckay et al (in preparation, Appendix B)) in hyperlipidaemia related genes on the lower tail of cholesterol in small VLDL (S-VLDL-C), esterified cholesterol in small VLDL (S-VLDL-CE) and concentration of extra small VLDL particles (XS-VLDL-P), and rare variation on HDL remodelling related genes on the lower tail of concentration of small HDL particles (S-HDL-P). I still observed nominal evidence of association in the WES and WGS datasets for the S-VLDL-C and XS-VLDL-P results using a 0.5% percentile cut-off for the tails but no evidence of association was found when using a 1% percentile cut-off (**Supplementary Table 10 of Riveros-Mckay et al (in preparation, Appendix B)**). This is likely due to the fact that by extending the number of individuals taken from the tails, we are decreasing the average distance to the mean and diluting signal coming from true extreme values.

Upper tails				
Trait	**WES P**	**WGS P**	**Meta-P**	**Gene Set**
S-VLDL-FC	$3.3x10^{-2}$	$2.37x10^{-2}$	$3.45x10^{-3}$	Hypertriglyceridemia_HPO
XS-VLDL-C	$3.3x10^{-2}$	$2.37x10^{-2}$	$3.45x10^{-3}$	Hypertriglyceridemia_HPO
Lower tails				
Trait	**WES P**	**WGS P**	**Meta-P**	**Gene Set**
S-VLDL-C	**$5.8x10^{-3}$**	**$2.31x10^{-3}$**	**$7.61x10^{-5}$**	**Hyperlipidaemia**
XS-VLDL-P	**$1.85x10^{-2}$**	**$7x10^{-4}$**	**$9.42x10^{-5}$**	**Hyperlipidaemia**
S-VLDL-CE	**$5.8x10^{-3}$**	**$6.75x10^{-3}$**	**$2.07x10^{-4}$**	**Hyperlipidaemia**
S-HDL-P	**$2.72x10^{-3}$**	**$1.84x10^{-2}$**	**$2.89x10^{-4}$**	**HDL_remodeling**
S-HDL-P	$4.10x10^{-2}$	$3.92x10^{-2}$	$8.x24x10^{-3}$	Hypertriglyceridemia_CTD

Table 3.8:Gene sets where there is a nominally significant enrichment of rare variation in the tails of a lipid or lipoprotein measurement (p<0.05) in both WES and WGS. Highlighted in bold are gene sets that are significant after meta-analysis using Stouffer's method [306] and after adjusting for multiple traits (p<=0.00037). WES P: permutation p in WES. WGS P: permutation p in WGS. Meta-P: p after meta-analysis using Stouffer's method. S-VLDL-FC: Free cholesterol in small VLDL. XS-VLDL-C : Cholesterol in very small VLDL. S-VLDL-C: Cholesterol in small VLDL. XS-VLDL-P: Concentration of very small VLDL particles. S-VLDL-CE: Cholesterol esters in small VLDL. S-HDL-P: Concentration of small HDL particles.

3.5 Discussion

Exploring rare coding variation provides an opportunity to gain insights into biological processes regulating the circulating levels of metabolic biomarkers. Here I took advantage of the combination of sequencing data and high-resolution NMR measurements to elucidate how this variation influences multiple metabolic measurements in a healthy cohort of UK blood donors.

To identify variants, genes and gene sets associated with metabolic biomarkers, I used a two-stage approach using WES data in discovery ($N_{discovery}$=3,741), and WGS data for validation ($N_{validation}$=3,401). I first performed single-point association analysis to assess whether I was able to recapitulate established associations with metabolic biomarkers, and potentially identify novel associated rare variants. This yielded associations at 34 previously established loci. The lack of novel findings was expected given the smaller sample size compared to similar studies using the same NMR platform (INTERVAL N=7,142, Kettunen et al. (2016) [173] N=24,925) and the limited power to detect associations with rare variants. As an example, for 7,142 individuals, I only had 2.5% power to detect a significant association (p<9.51x10^{-9} in a combined analysis, **Methods 3.3.6**) with an effect size of 1 for variants with MAF 0.1%. This study was part of a collaboration with Dr Adam Butterworth's group in the University of Cambridge. As such, array based genotype data for the full INTERVAL cohort was analysed by them and will form part of a large-scale meta-analysis collaborative effort. For this reason, I did not explore these results further.

Rare-variant aggregation tests were used to identify genes harbouring multiple rare coding variants associated with metabolic biomarkers. To gain power at the validation stage I adjusted analyses for correlated traits, an approach previously described for single-point analysis [310]. This yielded significant power gains, namely at the known *PAH* association with phenylalanine levels, where adjusting for 71 phenotypically correlated traits resulted in a greater than 30-fold magnitude change in the statistical evidence of association after meta-analysis. This approach therefore can benefit similar studies with multiple phenotypes measured in the same individuals. And, in future efforts, use of association data from these traits in the INTERVAL cohort, instead of publicly available summary statistics, to determine which traits are not genetically correlated could also be used to increase power for many of the measurements that had no publicly available summary statistics, including all derived lipid ratios. Overall, this approach yielded 4,114 gene-trait associations taken forward for validation ($p_{discovery}<5x10^{-3}$). After meta-analysis besides recapitulating previous associations in eight known genes (*APOB, APOC3*, *PAH*, *HAL*, *PCSK9*, *ALDH1L1*, *SCARB1* and *LIPC,* **Table 3.5**), this method also identified four genes (*ACSL1*, *MYCN, B4GALNT3*, *FBXO36*) that met standard gene-level significance ($p<2.5x10^{-6}$, **Table 3.5**) in at least one gene-trait association test. Of these, *ACSL1* and *MYCN* have been previously linked to lipid metabolism [312-314], suggesting that among the gene-level significant findings there may be additional true positives which will merit additional follow-up.

ACSL1, which encodes long-chain-fatty-acid—CoA ligase 1, is the predominant isoform of *ACSL* in the liver. The gene was associated with concentration of IDL particles in this study ($p = 6.20x10^{-7}$), and its deficiency in the liver has been shown to reduce synthesis of triglycerides and beta oxidation, and alter the fatty acid composition of major phospholipids

[316]. An intronic variant (rs60780116) in *ACSL1* has been associated with T2D [317] and elevated expression of *ACSL1* has been shown to be an independent risk factor for acute myocardial infraction [318].

MYCN encodes N-myc proto-oncogene protein and its amplification can lead to tumorigenesis [319, 320]. Previous animal studies have shown that inhibition of *MYCN* can lead to accumulation of intracellular lipid droplets in tumour cells [314]. Here I find association between *MYCN* and concentration of lipids, phospholipids and triglycerides in medium VLDL, total particle concentration of medium VLDL and triglycerides in small VLDL (min $p = 6.20 \times 10^{-7,}$ **Table 3.5, Supplementary Table 2 of Riveros-Mckay et al (in preparation, Appendix B)**).

The other two genes do not have any obvious link to lipid metabolism. *B4GALNT3* encodes beta-1,4-N-acetyl-galactosaminyl transferase 3. This protein mediates the N,N'-diacetyllactosediamine formation on gastric mucosa [321]. Mouse knockouts have been associated with abnormal tail movements, abnormal retinal pigmentation and increased circulating alkaline phosphatase levels [322] and variants near the gene have been associated with height and hip circumference adjusted for BMI in human GWAS [94, 323]. *FBXO36* is a member of the F-box protein family. F-box proteins are known to be involved in protein ubiquitination [324]. Replication of these signals in additional studies would represent a novel link between these genes and lipid metabolism.

In gene set analysis, the “regulation of pyruvate dehydrogenase (PDH) complex” pathway was newly associated with 20 traits, mostly related to IDL and LDL lipoproteins. None of the genes in this pathway have been previously linked to any of these phenotypes, and this data suggests the signal arises from a cumulative effect of predicted loss-of-function variants in

different genes in the pathway (**Figure 3.1**), which represents a novel link between this pathway and lipoprotein metabolism. Of note, a carrier of a rare stop gain mutation (Gln248Ter) in *PDHX* had very high levels of LDL-C (4.086 mmol/l, top 0.03% of full INTERVAL cohort) with no other rare mutation in genes known to harbour rare mutations causative of hypercholesterolaemia (*PCSK9*, *APOB*, *LDLR*). The other carrier of this variant had slightly increased LDL-C levels but within normal clinical range (1.823 mmol/l, top 19.3% of the full INTERVAL cohort). Since we lack information on medication, specifically, lipid lowering medication, the degree to which this variant influences the observed LDL-C levels is difficult to assess. The PDH complex has been shown to be crucial for metabolic flexibility, i.e. the capacity to adjust fuel oxidation based on nutrient availability, which itself has been shown to play a role in cardiovascular disease [325].

In analyses aiming at identifying effector transcripts at established GWAS loci associated with traditional lipid measurements (HDL-C, LDL-C, TC and TG), I established that reported genes mapping near HDL-C associated loci were enriched for rare coding variants associated with multiple HDL-related measurements. The results remained significant ($p<0.005$) after removing genes known to be directly involved in HDL metabolism, suggesting rare coding variants in this gene set contribute to variation in these traits, and that this gene set is potentially enriched for additional effector transcripts, though common variants in the same haplotype as these rare variants could also account for some of the signal we observe. One of the major limitations of this approach is that most of the times, the reported gene is the closest gene and we may miss the true causal gene if the GWAS signal is regulating a more distant gene. It is also important to note that an enrichment of rare variant associations

near reported genes does not necessarily mean that they solely explain the GWAS non-coding association and other genes might also be contributing to the signal.

Finally, I showed that one can detect enrichment of rare variation in genes involved in lipoprotein metabolism in phenotypic extremes of some of these NMR measurements. Specifically, I showed enrichment of rare variants in hyperlipidaemia related genes in individuals with very low levels of cholesterol and esterified cholesterol in small VLDL, total small VLDL particle concentration, and enrichment of rare variants in HDL remodelling genes in individuals with very low levels of small HDL particles. Given that high levels of small HDL particles have been previously associated with higher incidence of ischemic stroke (IS) [326] some of these variants could have protective effects. These results are in agreement with previous work on LDL-C [285] and HDL-C [327] that show that common polygenic signals seem to have a higher impact on the higher extremes of lipid traits whereas there is evidence for a higher prevalence of rare variation on the lower extremes [327]. This is also expected since the INTERVAL cohort consists predominantly of healthy blood donors and therefore the distribution of many of these traits might be truncated and depleted of individuals with rare "damaging" variants. Another factor that could contribute to the observed results is that each trait will have a different distribution and given the fact I am choosing an arbitrary number of participants at the top and bottom of the distribution, these participants will not represent equivalent "extremes".

A major limitation of rare variant association analyses to date is that, despite the advances in computational methods predicting the pathogenicity of rare variants, many of these predicted deleterious variants appear to exert little to no effect as evidenced by the non-significant associations with known positive controls where one should be well powered to

detect association if most of these variants were sufficiently deleterious. Some reported gene-based associations may be due to a few population specific variants, making those findings hard to replicate. As an example, a study using the same NMR platform and performing gene-based analysis using exome-chip data found a significant association of *LIPG* with many HDL subclass traits (min p=3.8x10^{-17}, all protein-truncating and missense variants, $N_{variants}$=5 in a Finnish population [288] whereas in this study the same gene was only nominally significant in triglycerides in medium HDL (p=0.049) querying 19 missense and LoF variants predicted to be deleterious. Power in our study was ~ 82% to find an association at $p < 0.001$ if 50% of the variants included in the test were causal and had the same direction and maximum beta is 1.1, this dropped to ~75% power if 20% of those variants had opposite directions of effect. Upon further inspection, the burden in the original study is mostly driven by one LoF variant (rs200435657, p=4x10^{-6}), and one missense variant (rs201922257, p=8.6x10^{-9}) that are almost monomorphic in Non-Finnish Europeans (gnomAD AC=1 and 7 respectively, AN= 126,228 and 126,712) but have an increased AC in Finnish populations (gnomAD AC=43 and 44 respectively, AN= 25,782 and 25,784). Another missense variant contributing to the association (rs77960347, p=4.8x10^{-6}) is low frequency in NFE (INTERVAL MAF=1.6%) and therefore was not included in our analysis, but it is worth noting that this variant is predicted to be tolerated by SIFT and only possible damaging by PolyPhen. Another study using the same platform but focusing on amino acids [289] found a burden of rare variants in *BCAT2* (p=7.4x10^{-7}, all protein-truncating and missense variants, $N_{variants}$=3) affecting valine levels where one of the two variants driving the association (rs199999090, p=5.36x10^{-4}) was monomorphic in our data and the other variant (rs117048185, p=4.12x10^{-4}) was also similarly associated in my dataset (p=3.x89x10^{-3}) but was not predicted to be deleterious by MCAP (or other similar algorithms

like PolyPhen and SIFT) and therefore was not included in the burden test that included eight variants (p_{burden}=0.76). Other examples of non-significant associations from traditionally measured lipid traits include a *PNPLA5* association with LDL-C [328] and a *TEAD2* association with HDL-C [284]. In the case of *PNPLA5* we tested 10 predicted deleterious variants and found no association p=0.59. However, the reported association with *PNPLA5*, was driven by an African American signal [283]. In the case of *TEAD2* the SNP driving the signal, rs142665148, is monomorphic in the European population and was found in a Chinese population, although unlike *LIPG*, *BCAT2* and *PNPLA5,* this gene is not a known effector transcript and might represent a false positive.

Further work on the INTERVAL cohort incorporating proteomics data could help better understand the potential functional consequences of rare coding variation and help bridge the gap between the rare variant analyses associations presented in this chapter and the observed consequences to circulating metabolic biomarkers.Altogether, my results showed that focusing on rare variation and deep metabolic phenotyping provides new insights into circulating metabolic biomarker biology. This argues for the expansion of deeper molecular phenotyping as part of large cohort sequencing efforts to gain further understanding on the role of rare coding variation on circulating metabolic biomarkers which may potentially lead to novel drug target discovery and/or provide additional genetic validation for specific targets.

4 Chapter 4: The heritability of fructosamine and its genetic relationship to HbA1c.

4.1 Introduction

Hyperglycaemia (high blood glucose) is the defining characteristic of diabetes. Normal fasting glucose (FG) levels in the blood range between 4.0 to 5.4 mmol/L (72-99 mg/dL), while post-prandial levels range up to 7.8 mmol/L (140 mg/dL) two hours after eating (2hr glucose) [329]. The most common tests to diagnose diabetes are fasting glucose (FG) and 2hr glucose after an oral glucose tolerance test (OGTT) (**Fig 4.1**). Both tests require a fasting period between 8 and 12 hours. OGTT involves taking a blood sample after the fasting period and then patients are given a very sweet drink containing 75g of glucose. Another blood sample is taken after two hours and this sample is the one used for diagnosis (2hr glucose post-OGTT). In some cases, blood samples are also taken at regular intervals between the intake of the sugar drink and the 2hr blood sample. Concordance between FG and 2hr glucose is not complete [330] and interestingly, individuals diagnosed using both criteria have higher cardiovascular disease risk than those only diagnosed using FG [331], and cardiovascular and all-cause mortality are increased in individuals diagnosed using 2hr glucose when compared to individuals diagnosed using FG [332].

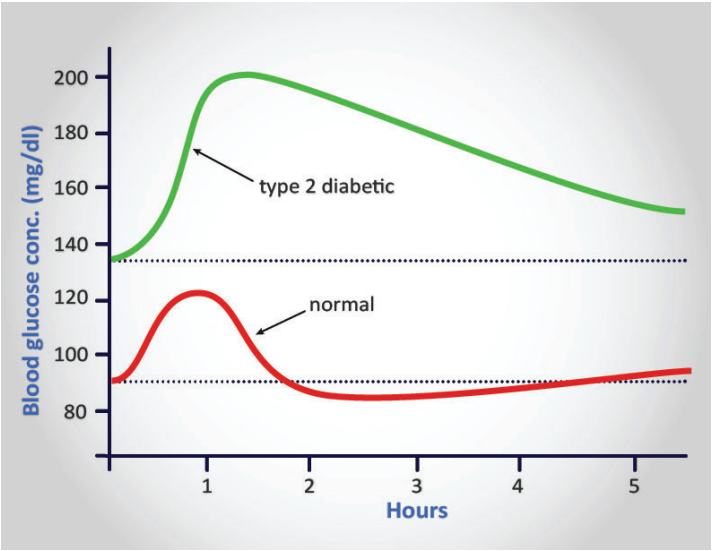


Figure 4.1: Diagnosis of type 2 diabetes. Blood glucose levels are shown for unaffected individuals (red line) and individuals with diabetes (green line) over a time course of 5 hours after glucose challenge. Fasting glucose (FG) test only measures glucose at the first time point. OGTT measures the first time point and the 2hr mark. Extracted from: https://themedicalbiochemistrypage.org/diabetes.php

Timely diagnosis of diabetes is important as uncontrolled high blood glucose levels can lead to clinical complications such as retinopathy, kidney failure or heart disease [25]. Undiagnosed diabetes can lead to damage to tissues due to hyperglycaemia which occurs over time without individuals displaying any marked symptoms. The degree of hyperglycaemia, reflecting the amount of damage to either insulin secretion or insulin response mechanisms (**Fig 4.2**), has an impact on the action course for treatment of the condition. Individuals with elevated glucose levels, but below the established threshold for diabetes diagnosis (FG=5.5 to 6.9 mmol/l (100 to 125 mg/dl), 2hr glucose=7.8 to 11.0 mmol/l (140 to 199 mg/dl)), are referred to as having prediabetes and can often manage their glucose levels by a combination of weight loss, physical activity and in rare cases oral glucose reducing medication [333]. When insulin secretion systems are severely damaged, individuals require insulin injections. This is the case for type 1 diabetes patients, who suffer from complete destruction of their beta-cells due to autoimmunity. For T2D cases there are

a host of oral treatments (e.g. metformin, sulphonylureas, thiazolidinediones (TZDs), dipeptidyl peptidase-4 (DPP-4) inhibitors and sodium glucose transporterase (SLGT-2) inhibitors) [334-336] but eventually many patients require insulin treatment.

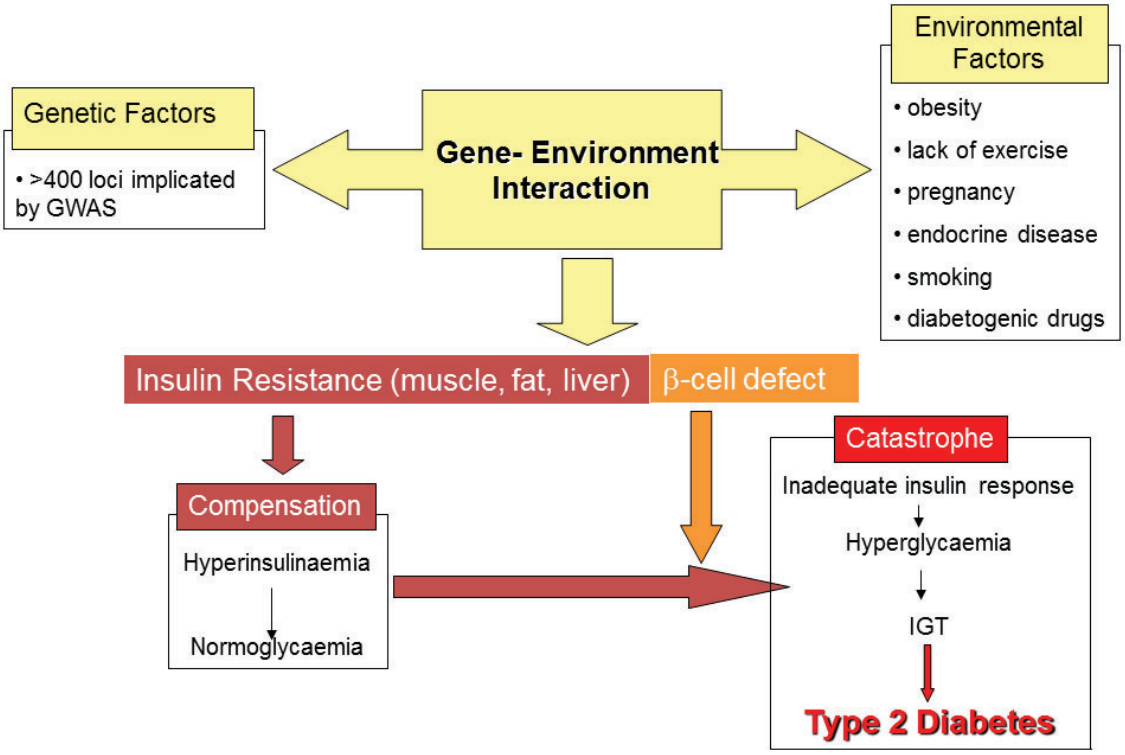


Figure 4.2: Aetiology of T2D. Interplay between genetic and environmental factors determines an individual's susceptibility to T2D. T2D arises from impaired beta-cell function and insulin resistance which occurs primarily in muscle, fat and liver tissues (i.e. insulin target organs). Under normal beta-cell function, in the setting of insulin resistance, insulin production is increased to increase the uptake of glucose by these tissues and glucose levels in the blood are kept within normal ranges. If this fails, glucose levels increase in the blood leading to impaired glucose tolerance (IGT). Individuals with untreated IGT have a high risk of developing T2D and cardiovascular disease. Image provided by Inês Barroso.

Another measure of glucose levels in the blood is glycated haemoglobin (HbA1c), which is the proportion of haemoglobin in the blood that has been glycated, and reflects average glucose levels over the life-span of an erythrocyte (~3 months). HbA1c is widely used to assess glycaemic control in patients with diabetes [337, 338]. As a diagnostic tool, HbA1c has a lower sensitivity than FG, but its negative predictive value is high, suggesting that low HbA1c levels provide strong evidence to discard a diabetes diagnosis [339]. However, HbA1c

levels as a diagnostic tool can be influenced by conditions that affect the lifespan of erythrocytes such as sickle cell trait and anaemia [340, 341], or by ethnic differences [342] (**Figure 4.3**). For example, ethnic minorities in the US such as Hispanics, Asians, American Indians and blacks have on average higher HbA1c levels than whites after adjusting for factors affecting glycaemia. This could affect the utility of HbA1c for T2D diagnosis, especially in populations of different ancestry [343]. Twin studies have estimated that heritability for HbA1c is high ranging from 62% to 75% [344, 345]. Multiple GWAS of HbA1c have looked into the genetic component of this trait, with a total of 60 loci identified to date [121, 346-350]. Lookups of association results of HbA1c-associated loci, with publicly available summary statistics for additional glycaemic (FG, 2h glucose and fasting insulin) and blood cell traits, in addition to conditional analyses adjusting for glycaemic traits (FG, 2hr glucose) or blood cell traits (haemoglobin levels, mean corpuscular volume, mean corpuscular haemoglobin) classified these loci as those mostly influencing HbA1c through glycaemic, erythrocytic, or unclassifiable pathways [121]. Understanding the pathway through which these variants affect HbA1c levels is important as it may influence their effect on T2D risk, diagnosis and treatment. For example, in a recent study from the MAGIC investigators [121] the authors described a missense variant in *G6PD* that lowers HbA1c levels through non-glycaemic pathways. This means that the lower HbA1c levels in *G6PD* variant carriers no longer reflect ambient glycaemia and therefore individuals with this variant could remain undiagnosed for T2D if this information were not taken into account and if they were screened by a single HbA1c measurement (see **Chapter 1, Section 1.2.2**). Understanding HbA1c genetics, which is studied in healthy non diabetic individuals to avoid confounding by diabetes and its treatment, could therefore help improve T2D diagnosis in populations of different ancestry.

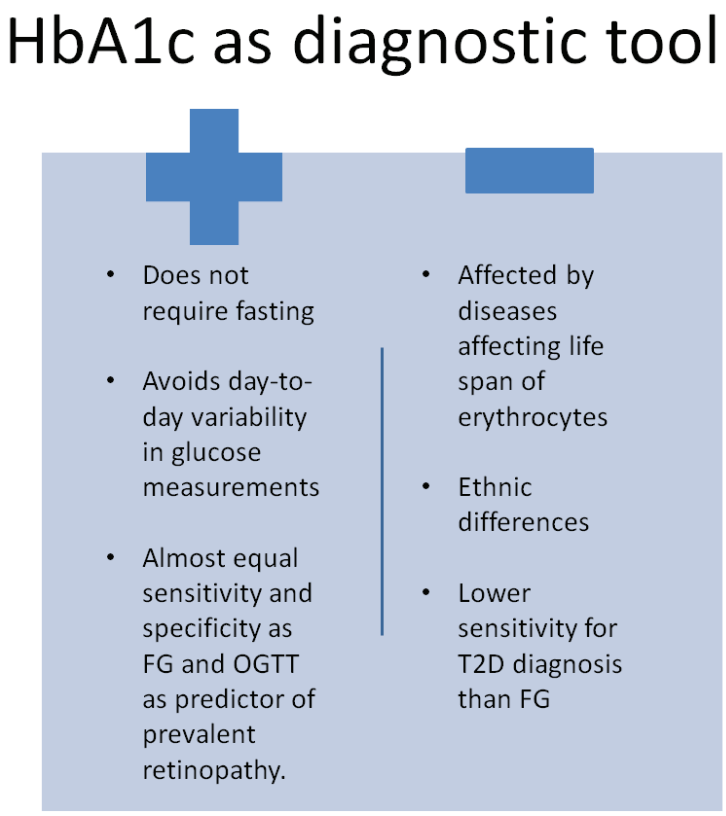


Figure 4.3: Advantages and disadvantages of HbA1c as a diagnostic tool.

Fructosamine is a measurement of glycation of total serum proteins. Since the most abundant serum protein is albumin, fructosamine normally reflects glycation of albumin [351]. In contrast to HbA1c, fructosamine measures short-term glycaemic control (from two to three weeks) and has been suggested as a useful marker for monitoring quick changes in glycaemic levels after treatment in individuals with diabetes [352]. As it is independent of haemoglobin, fructosamine levels are not affected by blood disorders and therefore is less likely to be influenced by erythrocytic traits, making it a viable alternative to HbA1c in the presence of anaemia or other blood disorders to monitor glycaemic levels [353]. Despite its potential advantages, it is not widely used as a measure of glucose control and has not been as standardised as HbA1c [354]. This lack of standardisation is problematic if fructosamine is

to be implemented as a diagnostic tool as accurate cutoffs need to be defined making sure variability within and between different labs is due to individual differences and are not assay or laboratory dependent. Nevertheless, studies have found association of fructosamine levels with diabetes incidence [355], retinopathy and chronic kidney disease [356], independently of baseline fasting glucose and HbA1c. Furthermore, there is a high correlation of fructosamine levels with HbA1c levels in patients with diabetes (r=0.7-0.8; [356-359]). The discordance in individuals between HbA1c levels and those levels predicted by its regression on fructosamine has been termed "glycation gap (GG)" [360]. Differences in FG and HbA1c as T2D diagnostic tools can be influenced by the GG as shown in a study where individuals were classified into three groups based on GG tertiles: low, medium and high glycators. Individuals diagnosed with diabetes by FG/2hr glucose and diagnosed as normoglycaemic by HbA1c had a significant depletion of individuals in the upper tertile of the GG suggesting individuals with low propensity for haemoglobin glycation are less likely to be diagnosed as diabetic by HbA1c criteria. In fact, the optimal HbA1c cutoff (i.e, the value misclassifying fewest patients) for low glycators was lower than that of high glycators (5.75% vs 6.25%) and had reduced sensitivity (54% vs 70%) [361]. To date, only one small study has looked at heritability of fructosamine in twins (40 monozygotic and 46 dizygotic) concluding it was not significantly inherited although a model including additive genetic effects and unique environmental influences could not be excluded [362]. A recent fructosamine GWAS performed on 8,951 mostly normoglycaemic white individuals ($N_{discovery}$=7,647, $N_{replication}$=1,304) found one single replicating association near *RCN3* (rs34459162, p=5.3x10^{-9}) after meta-analysis. This study did not examine the heritability of fructosamine and there was no significant evidence of genetic correlation with FG or HbA1c although there was some evidence of association for three established FG and/or HbA1c loci

(*TCF7L2*, *GCK* and *SLC2A2*) [363]. Elucidating non-glycaemic genetic influences of measurements used in T2D diagnosis, such as HbA1c and fructosamine, can help improve the diagnostic accuracy of these tests as well as provide insights into the different mechanisms by which these traits increase risk of diabetes comorbidities such as cardiovascular disease or chronic kidney disease independently of other glycaemic traits.

In this chapter, I performed a fructosamine GWAS to gain further insights into the genetic influences on fructosamine levels, and explore its genetic relationship with HbA1c and other glycaemic traits.

4.2 Chapter aims

The overall aim of this chapter is to explore the genetic basis of fructosamine. I use genome-wide genetic data available on up to ~19M SNPs on 24,586 individuals from the INTERVAL cohort to:

I. Assess the heritability of fructosamine.
II. Find novel loci associated with fructosamine.
III. Explore the genetic correlation of fructosamine with other glycaemic traits.
IV. Explore the effects of established glycaemic loci on fructosamine.

4.3 Methods

4.3.1 Participants

Work in this chapter was done using the INTERVAL cohort which consists of 47,394 predominantly healthy blood donors in the UK (more details in **Chapter 3 Methods 3.3.1**).

4.3.2 Genotyping, variant quality control and imputation

Genotyping, variant quality control and imputation on the INTERVAL cohort were performed by collaborators and full details can be found in Astle et al 2017 [293]. INTERVAL participants (N_{total}=48,813) were genotyped in ten batches on the UK Biobank Affymetrix Axiom Array. Standard QC procedures were implemented by Affymetrix during the genotype calling pipeline. Samples were excluded if signal intensity was poor (dish QC <0.82) or if call rate was low (<97%). Variants were excluded if a) call rate was low (<95%), b) there were more than three genotype clusters, c) cluster statistics (Fisher's linear discriminant, heterozygous cluster strength, homozygote ratio offset) indicated poor quality or d) they were complicated multi-allelic variants . Extra QC steps were performed by Tao Jiang and Heather Elding. Variants from a batch were failed if: a) fewer than ten minor allele homozygotes were called, b) the cluster plot contained at least one sample with an intensity at least twice as far from the origin as the next most extreme sample, c) the outlying sample(s) had an extreme polar angle (< 15° or > 75°) in the direction of the minor allele. Next, duplicate and non-European samples were excluded. Non-Europeans were defined based on PC1 and PC2 score thresholds defined after visual inspection of a PCA with 1000G major ancestry populations. Within batch variant QC was then performed discarding variants based on deviation from HWE ($p < 5x10^{-6}$) and a low call rate (<97%). If variants failed any of these last two filters or any of the Affymetrix filters in four out of ten batches, variants were discarded across all batches. After merging all batches, sample contamination was estimated using a contamination estimate based on allele frequency and probeset intensity [297]. Samples were excluded if this estimate was >10% or >3% if the sample also had more than ten first- or second-degree relatives (PI_HAT > 0.1875). Samples were then

excluded if they were heterozygosity outliers (>3 SD from mean), they had missing phenotypic sex or if supplied sex mismatched genetically inferred sex. Variants were removed if they had a MAF range >0.05 across all batches, if they were monomorphic in one or more batches and MAF > 0.01 in another batch, or if they had differing minor allele between batches (for variants with max MAF <0.475). An extra 69 across-batch duplicates were removed after merging batches. A global HWE filter ($p < 5x10^{-6}$) and a stringent call rate filter (>99% on non-failed batches and >75% globally) were applied to select variants used for imputation. Dataset was phased using SHAPEIT3 [364] and then imputed to a combined UK10K-1000G Phase III imputation reference panel using the Sanger Imputation Server [127].

4.3.3 Phenotyping

Phenotyping was performed by Star-SHL lab (http://www.star-shl.nl/). Fructosamine was measured using a colorimetric assay (Roche/Hitachi MODULAR P analyser system) and HbA1c was measured using a high performance liquid chromatography assay (Tosoh Automated Glycohemoglobin Analyser HLC-723G8 system) using serum collected on the 2 year follow-up visit. Fructosamine and HbA1c measurements as well as questionnaire data, technical variables and blood cell trait measurements were provided by the data administrator (University of Cambridge). I performed phenotype quality control in R to prepare the data for association analysis.

4.3.4 Association analysis, heritability and genetic correlation

Residuals obtained after phenotype quality control (**Results 4.4.1**) for fructosamine and HbA1c were used in this analysis. BOLT-LMM [227] was used to run genome-wide

association analysis on 19,100,024 variants with MAF > 0.1% and INFO score >0.4. LD score regression [42] was used to establish the heritability and genetic correlation of both traits. LD score regression was also used to compute genetic correlation of fructosamine and HbA1c with blood cell traits. A subset of 1,142,170 of the 1,217,312 HapMap3 SNPs with non-missing betas was kept in each dataset to perform these analyses. HbA1c summary statistics for European individuals from Wheeler et al 2017 [121] were obtained from the MAGIC consortium website (https://www.magicinvestigators.org/downloads/). Blood cell traits summary statistics were downloaded from http://www.bloodcellgenetics.org/. Genetic correlation analyses with glycaemic traits and albumin was performed using LD Hub [231].

4.3.5 Fructosamine discovery GWAS

LD score regression results showed no signs of inflation so no genomic correction was performed (LD intercept=1.01). I performed clumping as implemented in PLINK [223] to establish unique loci. Variants were clumped if they were 250kb away from the lead signal and if r^2>0.1. Conditional analysis was also used to identify distinct signals within a locus after clumping. To compare effect sizes of the lead variant near *RCN3* (rs34459162) found in a previous study [363], I reran normalisation on fructosamine matching the transformation done in that study and correcting for sex, donation centre, height, weight, processing date, number of donations and attendance date. SNPTEST v2.5.2 was used to rerun association analysis under an additive model and used for reciprocal conditional analyses. To estimate the significance of the difference in effect sizes I used a Z-test. SNPTEST v2.5.2 was also

used to test for association of the lead signal near *RCN3* in this study with serum albumin levels.

4.3.6 Lookup of established glycaemic loci

A list of established glycaemic loci was obtained from Eleanor Wheeler. This list was curated by Eleanor Wheeler and Gaëlle Marenne and last updated in March 2018. I first removed from the list chromosome X variants, variants monomorphic in the European population, and one variant not present in the INTERVAL data. To extract index variants per locus, I extracted LD information from European individuals in 1000G using LDlink (https://ldlink.nci.nih.gov/) [365]. For variants that were not biallelic in 1000G, I calculated LD in the INTERVAL dataset. For each pair of variants with $r^2 > 0.1$, I kept variants that had the lowest p-value in any of the association analysis with fasting glucose, 2 hr glucose, fasting insulin or HbA1c performed in European individuals in the latest trans-ethnic MAGIC unpublished analyses. Association data was provided by Ji Chen. The full list of index SNPs is presented in **Table 4.1**. Significance threshold for association was established by dividing 0.05 by the number of loci tested ($0.05/142 = 3.5 \times 10^{-4}$). To assess if there was an enrichment of directionally consistent and nominally significant signals in fructosamine that have been previously associated with glycaemic traits I performed a binomial test. Of the set of variants previously associated with glycaemic traits that were nominally significant in fructosamine, signals near *USP4* and *ANK1* were removed in this test since their association status with HbA1c is through non-glycaemic pathways and I chose to focus only on those influencing HbA1c through glycaemic pathways [121]. I also removed rs9727115 and rs150781447 as these were only associated with proinsulin levels adjusted for fasting glucose and late-phase proinsulin levels and the expected relationship between directions

of effect of a variant affecting proinsulin and fructosamine is not obvious. Direction of effect was determined using the HbA1c raising allele as reference (i.e we expect HbA1c raising variants to also raise fructosamine). If the variant was not associated with HbA1c, I used FG as reference and for rs9884482, which was not associated with HbA1c or FG, I used FI.

snp	chr	pos	gene	trait
rs340874	1	214159256	PROX1-AS1	FG [83], FG_adjBMI [366]
rs1886686	1	67390468	WDR78	FG_adjBMI *
rs2820436	1	219640680	LYPLAL1	FI [367], FI_adjBMI [367]
rs141203811	1	229772141	URB2	FI_adjBMI [368]
rs2375278	1	25529038	SYF2	HbA1c [121]
rs267738	1	150940625	CERS2	HbA1c [121], HbA1c_adjBMI *
rs6684514	1	156255456	TMEM79	HbA1c [346]
rs857725	1	158607935	SPTA1	HbA1c_adjBMI *
rs9727115	1	99177253	SNX7	Proinsulin_adjFG [369]
rs1260326	2	27730940	GCKR	2hG_adjBMI [370], FG_adjBMI [367], FI_adjBMI [367]
rs895636	2	45188353	SIX3	FG [371], FG_adjBMI [372]
rs1371614	2	27152874	DPYSL5	FG_adjBMI [366], FG_BMI30 [366]
rs3736594	2	27995781	MRPL33	FG_adjBMI [366]
rs35720761	2	43519977	THADA	FG_adjBMI *, HbA1c_adjBMI *
rs138726309	2	169763262	G6PC2	FG_adjBMI [368]
rs2232323	2	169764141	G6PC2	FG_adjBMI *, HbA1c_adjBMI *
rs146779637	2	169764368	G6PC2	FG_adjBMI *, HbA1c_adjBMI *
rs552976	2	169791438	ABCB11	FG_adjBMI [367], HbA1c [367]
rs733331	2	173546313	RAPGEF4-AS1	FG_adjBMI [372]
rs1530559	2	135755629	MAP3K19	FI [367], FI_adjBMI [367]
rs10195252	2	165513091	COBLL1	FI [367], FI_adjBMI [367]
rs1983210	2	220421417	OBSL1	FI_adjBMI *
rs2943645	2	227099180	LOC646736	FI_adjBMI [367]
rs17509001	2	24021231	ATAD2B	HbA1c [121]
rs12621844	2	48414735	FOXN2	HbA1c [121]
rs3755157	2	169792171	ABCB11	HbA1c [346]
rs17256082	2	175292364	SCRN3	HbA1c [121]
rs11708067	3	123065778	ADCY5	2hG_adjBMI [367], FG [83], FG_adjBMI [366], HbA1c [121]
rs7651090	3	185513392	IGF2BP2	2hG_adjBMI [367], FG [367], FG_adjBMI [367]
rs17036328	3	12390484	PPARG	FI_adjBMI [367]
rs7616006	3	12267648	SYN2	HbA1c [121]
rs9818758	3	49382925	USP4	HbA1c [121]
rs8192675	3	170724883	SLC2A2	HbA1c [121]

snp	chr	pos	gene	trait
rs4894799	3	171795540	FNDC3B	HbA1c [121]
rs35726701	3	49740895	RNF123	HbA1c_adjBMI *
rs9884482	4	106081636	TET2	FI [367], FI_adjBMI [367]
rs3822072	4	89741269	FAM13A	FI_adjBMI [367]
rs6822892	4	157734675	PDGFC	FI_adjBMI [367]
rs17046216	4	166255704	MSMO1	FI_adjBMI [378], HOMA-IR [378]
rs13134327	4	144659795	FREM3	HbA1c [121]
rs2237051	4	110901198	EGF	HbA1c_adjBMI *
rs1019503	5	96254817	ERAP2	2hG [367], 2hG_adjBMI [367]
rs146886108	5	14751305	ANKH	FG_adjBMI *
rs7708285	5	76425867	ZBED3-AS1	FG_adjBMI [367]
rs7713317	5	95716722	PCSK1	FG_adjBMI [367]
rs4865796	5	53272664	ARL15	FI [367], FI_adjBMI [367]
rs6450057	5	51647364	PELO	FI_adjBMI [373]
rs459193	5	55806751	LOC101928448	FI_adjBMI [367]
rs31244	5	75594743	SV2C	HbA1c_adjBMI *
rs35658696	5	102338811	PAM	Insulinogenic index [374]
rs10305492	6	39046794	GLP1R	FG [375], FG_adjBMI [368]
rs17762454	6	7213200	RREB1	FG_adjBMI [367]
rs35742417	6	7247344	RREB1	FG_adjBMI [368], HbA1c_adjBMI *
rs2745353	6	127452935	RSPO3	FI [367], FI_adjBMI [367]
rs6912327	6	34764922	UHRF1BP1	FI_adjBMI [367]
rs7756992	6	20679709	CDKAL1	HbA1c [121]
rs1800562	6	26093141	HFE	HbA1c [367], HbA1c_adjBMI *
rs11964178	6	109562035	LOC100996634	HbA1c [121]
rs9399137	6	135419018	HBS1L	HbA1c [346]
rs592423	6	139840693	LOC645434	HbA1c [121]
rs1799945	6	26091179	HFE	HbA1c_adjBMI *
rs1799884	7	44229068	GCK	1hG [376], 2hG [376], FG [376], HbA1c [350]
rs2191349	7	15064309	AGMO	FG [83], FG_adjBMI [366], HbA1c [121]
rs6943153	7	50791579	GRB10	FG [367], FG_adjBMI [367]
rs6947345	7	101071933	COL26A1	FG [377]
rs194524	7	89861832	STEAP2	FG_adjBMI *
rs1167800	7	75176196	HIP1	FI [367], FI_adjBMI [367]
rs35332062	7	73012042	MLXIPL	HbA1c_adjBMI *
rs11558471	8	118185733	SLC30A8	FG [367], FG_adjBMI [366], HbA1c [121], Proinsulin [369]
rs4841132	8	9183596	LOC157273	FG_adjBMI [366], FG_BMI30 [366], FI_adjBMI [366], FI_BMI30 [366]
rs4737009	8	41630405	ANK1	HbA1c [367]
rs6980507	8	42383084	SLC20A2	HbA1c [121]
rs34664882	8	41543675	ANK1	HbA1c_adjBMI *
rs7034200	9	4289050	GLIS3	FG [83], FG_adjBMI [366]
rs10811661	9	22134094	CDKN2B-AS1	FG [367], FG_adjBMI [366]

snp	chr	pos	gene	trait
rs16913693	9	111680359	IKBKAP	FG [367], FG_adjBMI [367]
rs651007	9	136153875	ABO	FG [375]
rs3829109	9	139256766	DNLZ	FG [367], FG_adjBMI [367]
rs7040409	9	91503236	C9orf47	HbA1c [121]
rs1467311	9	110536932	KLF4	HbA1c [121]
rs11557154	9	34107505	DCAF12	HbA1c_adjBMI *
rs3824420	9	712766	KANK1	Proinsulin AUC 0-30 [374]
rs7903146	10	114758349	TCF7L2	2hG [367], 2hG_adjBMI [367], FG [83], FG_adjBMI [366], FI [367], FI_adjBMI [366], HbA1c [367], Proinsulin [369]
rs701865	10	95381773	PDE6C	FG_adjBMI *
rs11195502	10	113039667	ADRA2A	FG_adjBMI [367]
rs7077836	10	132751498	MIR378C	FI_adjBMI [378], HOMA-IR [378]
rs16926246	10	71093392	HK1	HbA1c [350]
rs906220	10	71060610	HK1	HbA1c_adjBMI *
rs11605924	11	45873091	CRY2	FG [83], FG_adjBMI [366]
rs11603334	11	72432985	ARAP1	FG [367], FG_adjBMI [366], FG_BMI30 [366], HbA1c [121], Proinsulin [369]
rs1387153	11	92673828	MTNR1B	FG_adjBMI [367], HbA1c [367]
rs3782123	11	205198	BET1L	HbA1c [121]
rs2237896	11	2858440	KCNQ1	HbA1c [121]
rs174577	11	61604814	FADS2	HbA1c [121]
rs11224302	11	100456604	ARHGAP42	HbA1c [121]
rs415895	11	9769562	SWAP70	HbA1c_adjBMI *
rs117706710	11	10508903	AMPD3	HbA1c_adjBMI *
rs643788	11	118967758	DPAGT1	HbA1c_adjBMI *
rs10501320	11	47293799	MADD	Proinsulin [369]
rs10838687	11	47312892	MADD	Proinsulin [369]
rs17331697	12	97868906	RMST	FG [377]
rs10747083	12	133041618	FBRSL1	FG [367], FG_adjBMI [367]
rs2657879	12	56865338	GLS2	FG_adjBMI [367]
rs145878042	12	48143315	RAPGEF3	FI_adjBMI *
rs860598	12	102898446	IGF1	FI_adjBMI [367]
rs2110073	12	7075882	PHB2	HbA1c [121]
rs2408955	12	48499131	SENP1	HbA1c [121]
rs3184504	12	111884608	SH2B3	HbA1c_adjBMI *
rs2650000	12	121388962	HNF1A-AS1	Insulinogenic index [374]
rs150781447	12	65224220	TBC1D30	Proinsulin AUC 30-120 [374]
rs11619319	13	28487599	PDX1-AS1	FG [367], FG_adjBMI [367], HbA1c [121]
rs576674	13	33554302	KL	FG [367], FG_adjBMI [367], HbA1c [121]
rs282587	13	113351662	ATP11A	HbA1c [121]
rs9604573	13	114542858	GAS6, GAS6-AS1	HbA1c [121]
rs3783347	14	100839261	WARS	FG [367], FG_adjBMI [367]

snp	chr	pos	gene	trait
rs229587	14	65263300	SPTB	HbA1c_adjBMI *
rs4502156	15	62383155	C2CD4A	2hG_adjBMI [367], FG_adjBMI [367], Proinsulin [369]
rs2018860	15	99258710	IGF1R	FG_adjBMI [372]
rs1549318	15	71109147	LARP6	Proinsulin [369]
rs11248914	16	293562	ITFG3	HbA1c [121]
rs1558902	16	53803574	FTO	HbA1c [121]
rs4783565	16	68750190	CDH3	HbA1c [121]
rs837763	16	88853729	PIEZO1	HbA1c [121]
rs3747481	16	30666367	PRR14	HbA1c_adjBMI *
rs201226914	16	88798919	PIEZO1, LOC100289580	HbA1c_adjBMI *
rs72839768	17	7129898	DVL2	2hG_adjBMI *
rs61741902	17	2282779	SGSM2	Fasting proinsulin [374]
rs9914988	17	27183104	ERAL1	HbA1c [121]
rs12602486	17	42241929	C17orf53	HbA1c [346]
rs2073285	17	76117361	TMC6	HbA1c [121]
rs1046896	17	80685533	FN3KRP	HbA1c [367]
rs2748427	17	76121864	TMC6	HbA1c_adjBMI *
rs7225887	17	80904844	B3GNTL1	HbA1c_adjBMI *
rs4790333	17	2262703	SGSM2	Proinsulin [369]
rs1800437	19	46181392	GIPR	2hG_adjBMI *
rs731839	19	33899065	PEPD	FI [367], FI_adjBMI [367]
rs11667918	19	17232499	MYO9B	HbA1c [346]
rs17533903	19	17256523	MYO9B	HbA1c [121]
rs35413309	19	33167837	RGS9BP	HbA1c_adjBMI *
rs6113722	20	22557099	LINC00261	FG [367], FG_adjBMI [367]
rs17265513	20	39832628	ZHX3	FG_adjBMI [368]
rs855791	22	37462936	TMPRSS6	HbA1c [367], HbA1c_adjBMI *

Table 4.1: Index variants for established glycaemic loci per trait. 1hG= 1 hr Glucose. 2hG_adjBMI= 2 hr glucose adjusted for BMI. FG=Fasting glucose. FG_adjBMI=Fasting glucose adjusted for BMI. FI=Fasting insulin. FI_adjBMI=Fasting insulin adjusted for BMI. HbA1c=Glycated haemoglobin. HbA1c_adjBMI=Glycated haemoglobin adjusted for BMI. HOMA-IR= Insulin resistance homeostasis model assessment. Proinsulin AUC 0-30= Early phase proinsulin. Proinsulin AUC 30-120= Late phase proinsulin. Proinsulin_adjFG=Proinsulin adjusted for fasting glucose. Sources=Publication where the index SNP in the table was first associated with its respective trait. *MAGIC unpublished: based on European results from meta-analysis of data genotyped on the ExomeChip array.

4.4 Results

4.4.1 Phenotype quality control

Biological measurements can be susceptible to technical variation. To prepare the data for association analysis I performed quality control to assess if there were any effects of

technical variables on the fructosamine and HbA1c measurements. Fructosamine was measured on 28,310 individuals of the INTERVAL cohort and HbA1c was measured on 5,811 individuals out of which 5,420 had both measurements.

First, I inspected if measured values were concordant with what is expected based on available literature on the subject. Median fructosamine levels were high (294 µmol/L) compared to the normal expected range in healthy individuals (202-285 µmol/L) [379]. In contrast, HbA1c median levels were within range (median=35 mmol/mol, expected value=31-42 mmol/mol) [380]. Correlation between fructosamine and HbA1c was lower than expected (r=0.11 (this study) vs r=0.61[360], **Fig 4.4**).

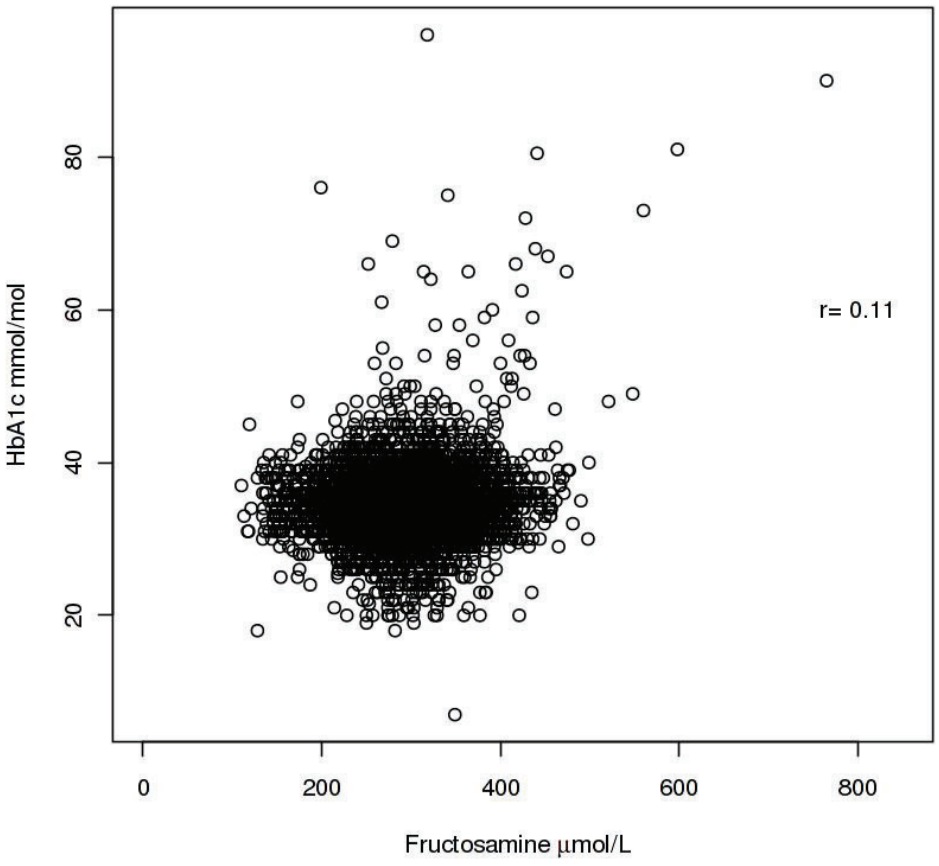


Figure 4.4: Correlation between fructosamine and HbA1c levels. r=correlation between fructosamine and HbA1c.

Next, I performed linear regression to determine which biometric and technical variables were significantly associated with fructosamine and HbA1c measurements. I determined sex, donation centre, use of glucose medication, height, weight, processing date and number of donations were significantly associated with both; while age, number of low haemoglobin deferrals, use of lipid lowering medication and use of blood pressure medication were associated with HbA1c exclusively and attendance date with fructosamine (**Table 4.2**). Individuals that reported use of glucose medication were subsequently removed.

Variable	Fructosamine	HbA1c
Age		X
Sex	X	X
Height	X	X
Weight	X	X
Attendance date	X	
Processing date	X	X
Donation centre	X	X
Number of donations	X	X
Number of low haemoglobin deferrals		X
Use of glucose medication	X	X
Use of lipid medication		X
Use of blood pressure medication		X

Table 4.2: Variables significantly associated with fructosamine and HbA1c. An X marks variables significantly associated (p<0.05 in linear regression).

After adjusting for relevant biometric and technical variables, residuals were extracted and inverse rank normalised. Correlation between fructosamine and HbA1c remained unchanged after adjusting for covariates (**Fig 4.5**).

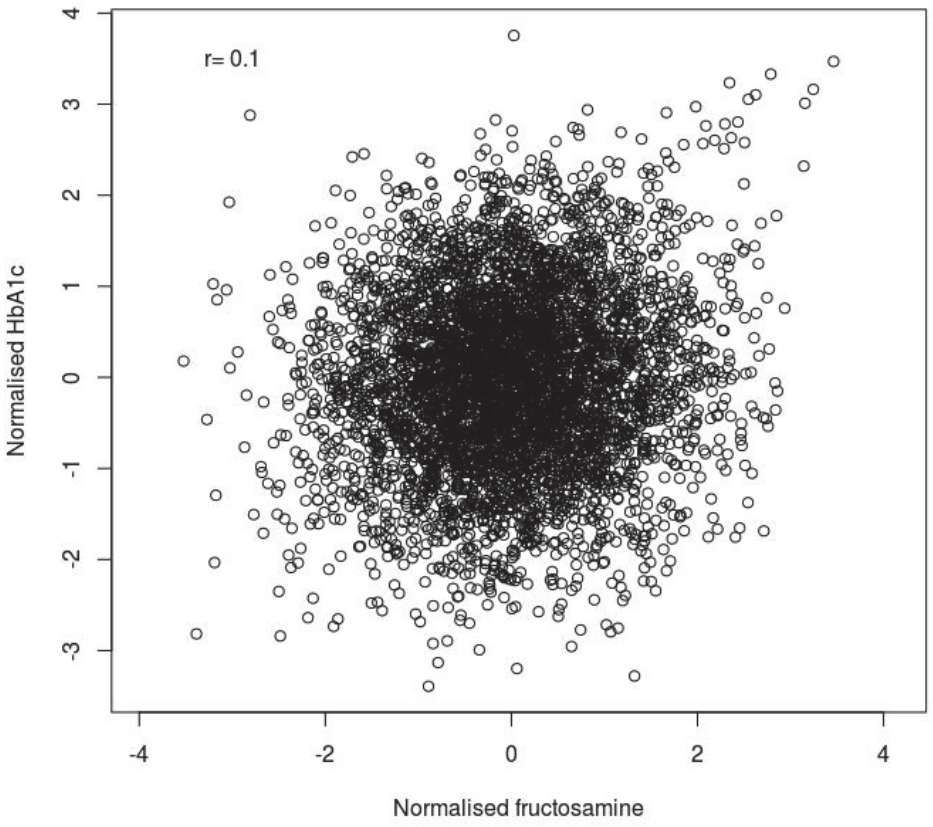


Figure 4.5: Correlation between normalised fructosamine and HbA1c levels after adjusting for biometric and technical variables. r=correlation between fructosamine and HbA1c.

4.4.2 Heritability of fructosamine and genetic correlation results

After inverse rank normalisation, 24,586 individuals with fructosamine measurements and 5,153 with HbA1c measurements were mapped to the genetic data. I used this data to run genetic association analyses with BOLT-LMM. LD score regression was used on association results to estimate common SNP (MAF>5%) heritability for both traits. The heritability estimate for fructosamine was very low (2% (95% CI -2%-5%)) which is in contrast to that for HbA1c (17% (95% CI 0%-35%)). No evidence of genetic correlation was observed between the two traits (Genetic correlation (RG)=0.40, SE=0.63, *p*=0.52). Using LD Hub, I estimated genetic correlation with other glycaemic traits. Fructosamine was not significantly

correlated with any other glycaemic trait (p>0.05, **Table 4.3**). HbA1c was significantly correlated with T2D (RG=0.47, p=0.01) and FG (RG=0.45, p=3.6x10^{-3}, **Table 4.3**). There was also significant and almost complete genetic correlation between HbA1c in this study and HbA1c results from previous MAGIC efforts [121] (RG=0.88, SE=0.20, p=1.34x10^{-5}) supporting the reliability of HbA1c measurements in this study. I also calculated genetic correlation between fructosamine and serum albumin and found no evidence of genetic correlation. Finally, given the established role of HbA1c-associated variants with some blood cell traits [121], I also calculated genetic correlation with twelve blood cell traits (**Table 4.3**). I found a positive genetic correlation of HbA1c with mean corpuscular haemoglobin concentration (MCHC, RG=0.37, SE=0.14, p=6.1x10^{-3}), mean corpuscular volume (MCV, RG=0.21, SE=0.11,p=0.04) and mean corpuscular haemoglobin (MCH,RG=0.28, SE=0.11, p=0.01) and a negative genetic correlation with red cell distribution width (RDW, RG=-0.24, SE=0.11, p=0.04).

Trait	Fructosamine			HbA1c		
	RG	SE	*P*	RG	SE	*p*
T2D	-0.04	0.19	0.88	0.47	0.18	0.01
FG_adjBMI	0.60	0.35	0.09	0.45	0.15	3.6x10^{-3}
FI_adjBMI	-0.43	0.32	0.17	-0.15	0.17	0.36
Fasting proinsulin	-0.10	0.37	0.78	-0.01	0.30	0.98
2hG_adjBMI	-0.76	0.49	0.12	0.05	0.28	0.87
Albumin	0.49	0.49	0.32	0.18	0.30	0.55
HCT	0.52	0.29	0.07	0.13	0.11	0.25
HGP	0.45	0.26	0.08	0.22	0.12	0.07
HLR	0.30	0.22	0.18	0.15	0.11	0.18
HLR%	0.23	0.20	0.24	0.17	0.11	0.14
IRF	0.30	0.24	0.21	0.01	0.12	0.90
MCHC	-0.18	0.21	0.41	0.37	0.14	6.1x10^{-3}
MCH	-0.03	0.14	0.83	0.28	0.11	0.01
MCV	0.04	0.14	0.75	0.21	0.11	0.04
RBC	0.35	0.21	0.09	-0.07	0.11	0.54
RDW	-0.18	0.16	0.27	-0.24	0.11	0.04
RET	0.25	0.19	0.20	0.19	0.12	0.10

Trait	Fructosamine			HbA1c		
	RG	SE	*P*	RG	SE	*p*
RET%	0.18	0.17	0.30	0.21	0.12	0.07

Table 4.3: Genetic correlation results for fructosamine and HbA1c. RG=genetic correlation. SE=standard error. P=p-value FG_adjBMI=Fasting glucose adjusted by BMI, FI_adjBMI=Fasting insulin adjusted by BMI, 2hG_adjBMI=2hr glucose adjusted by BMI. HCT=Haematocrit. HGB=Haemoglobin concentration. HLR=High light scatter reticulocyte count. HLR%=High light scatter percentage of red cells. IRF=Immature fraction of reticulocytes. MCHC=Mean corpuscular haemoglobin concentration. MCH=Mean corpuscular haemoglobin. MCV=Mean corpuscular volume. RBC=Red blood cell count. RDW=Red cell distribution width. RET=Reticulocyte count. RET%=Reticulocyte fraction of red cells. Highlighted in yellow, significant genetic correlation estimates. Genetic correlation analyses with glycaemic traits and albumin was performed using LD Hub [231]. Blood cell traits summary statistics for genetic correlation obtained from Astle et al 2017 [291].

4.4.3 Discovery of novel loci associated with fructosamine

Fructosamine association analysis yielded two associated loci at genome-wide significance ($p<5x10^{-8}$). The first association signal was rs853777 near *G6PC2* (beta=0.06, $p=1.7x10^{-10}$). This locus is also associated with HbA1c and FG [83]. The lead SNP in Dupuis et al 2010 [83] is rs560887 (r^2 with rs853777=0.63) with an effect size of 0.032 (%) for HbA1c and 0.075 (mmol/L) for FG. The effect size for rs853777 on the untransformed fructosamine values was 2.54 µmol/L. The second association signal was rs111476047 near *RCN3* (beta=0.09, $p=4.8x10^{-14}$). This locus was also previously associated with fructosamine in Loomis et al 2018 [363]. The lead signal in their study, rs34459162, is in moderate LD with my index variant (rs111476047 r^2= 0.28), and has the same direction of effect. To compare effect sizes, I repeated the association analysis for rs34459162 transforming the fructosamine measurements using natural log transformation as done in the previous study instead of inverse rank normalisation. The effect size was smaller in this study but the difference was not significant ($beta_{INTERVAL}$=0.015, $beta_{reported}$=0.02, p_{diff}=0.24). Reciprocal conditional analysis suggested that the lead signal found in this study was more tightly linked to the true causal variant than the previously reported signal (**Table 4.4**).

rsid	p-value	beta	conditioned p-value	conditioned beta	conditioned on
rs111476047*+	$2.19x10^{-11}$	0.014	$3.42x10^{-6}$	0.011	rs34459162**
rs34459162+**	$2.08x10^{-7}$	0.015	$5.35x10^{-2}$	0.007	rs111476047*
rs111476047*	$5.23x10^{-14}$	0.087	$8.90x10^{-9}$	0.074	rs739347***
rs739347***	$7.61x10^{-8}$	0.079	$2.09x10^{-2}$	0.038	rs111476047*
rs111476047*	$5.23x10^{-14}$	0.087	$1.51x10^{-12}$	0.083	rs34010237****
rs34010237****	$1.75x10^{-4}$	0.047	$6.24x10^{-3}$	0.034	rs111476047*

Table 4.4: Reciprocal conditional analysis of lead variant near *RCN3*. r^2=0.28.+Analysis was performed using log transformed fructosamine values. *Lead signal in this study. **Lead signal in Loomis et al 2018[363]. ***Lead signal for albumin GWAS in Franceschini et al (2012) [381]. ****Lead signal for albumin GWAS in Kanai et al 2018[382]. Numbers might differ slightly from main text due to difference in software use for association analysis (**Methods 4.3.5**).

Two additional SNPs mapping near the *RCN3* locus, rs739347 (r2=0.25 with rs111476047) and rs34010237 (r^2=0.02 with rs111476047), have been previously associated with serum albumin levels in European [381] and Japanese individuals [382], respectively. Both rs739347 and rs34010237 variants were significantly associated with fructosamine levels in this study (p=6.9x10^{-8} and p=2.2x10^{-4}, respectively) though reciprocal conditional analyses showed the rs739347 signal was heavily attenuated after conditioning on the lead signal in this study suggesting this association was mostly driven by the lead signal in this study (**Table 4.4**). Finally, to assess the effect of my index variant (rs111476047) on serum albumin levels, I used NMR measurements available from the first visit (**Chapter 3 Methods 3.3.1**), and found a significant and directionally consistent association with albumin levels in these data (beta=0.02, p=6.2x10^{-3}).

4.4.4 Evaluation of the effects of established glycaemic loci on fructosamine levels

Finally, I explored the influence of established glycaemic loci on fructosamine levels. For this analysis I used a list curated by Eleanor Wheeler and Gaëlle Marenne (**Table 4.1**). In total,

142 unique SNPs associated with at least one glycaemic trait were extracted. I found significant associations (Bonferroni-corrected $p < 3.5 \times 10^{-4}$, **Methods 4.3.6, Table 4.5**) in four loci: *ADCY5*, *GCK*, *G6PC2* and *MTNR1B*.

rsid	Gene	Chr	Pos	EA	NEA	EAF	B	SE	P
rs11708067	*ADCY5*	3	123065778	A	G	0.76	0.04	0.01	1.5×10^{-4}
rs730497	*GCK*	7	44223721	T	C	0.18	0.04	0.01	1.4×10^{-4}
rs1387153	*MTNR1B*	11	92673828	T	C	0.29	0.04	0.01	1.9×10^{-4}
rs552976	*G6PC2**	2	169791438	G	A	0.64	0.06	0.01	8.3×10^{-9}

Table 4.5: Associations of established glycaemic loci on fructosamine. Chr=chromosome. Pos = position in GRCH37. EA=Effect allele. NEA=Non-effect allele. EAF=Effect allele frequency. B=effect size SE=standard error of effect. P=p-value. *Nearest gene is *ABCB11*, but *G6PC2* is known to be the effector transcript at the locus [368].

I also found an enrichment of nominally significant and directionally consistent glycaemic signals in the fructosamine association results suggesting that these loci also have an effect on fructosamine (binomial $p=5.6 \times 10^{-3}$, **Table 4.6**).

rsid	Chr	Pos	EA-FR	EAF	B	SE	P	Gene	Trait	EA-T
rs10811661	9	22134094	T	0.83	0.04	0.01	1.9×10^{-3}	*CDKN2B-AS1*	HbA1c	T
rs11603334	11	72432985	G	0.85	0.03	0.01	7.9×10^{-3}	*ARAP1*	HbA1c	G
rs11708067	3	123065778	A	0.76	0.04	0.01	1.5×10^{-4}	*ADCY5*	HbA1c	A
rs1387153	11	92673828	T	0.29	0.04	0.01	1.4×10^{-4}	*MTNR1B*	HbA1c	T
rs17265513	20	39832628	C	0.19	0.03	0.01	1.0×10^{-2}	*ZHX3*	FG	C
rs1799884	7	44229068	T	0.18	0.04	0.01	1.4×10^{-5}	*GCK*	HbA1c	T
rs2232323	2	169764141	A	0.99	0.13	0.06	2.3×10^{-2}	*G6PC2*	HbA1c	A
rs3829109	9	139256766	G	0.73	0.02	0.01	3.3×10^{-2}	*DNLZ*	HbA1c	G
rs552976	2	169791438	G	0.64	0.06	0.01	8.3×10^{-9}	*ABCB11*	HbA1c	G
rs7651090	3	185513392	G	0.32	0.03	0.01	1.2×10^{-3}	*IGF2BP2*	HbA1c	G
rs7708285	5	76425867	G	0.31	0.02	0.01	2.2×10^{-2}	*ZBED3-AS1*	HbA1c	G
rs9884482	4	106081636	C	0.37	0.02	0.01	2.4×10^{-2}	*TET2*	FI	C

Table 4.6: Nominally significant and directionally consistent established glycaemic loci. Table legend: Chr=chromosome. Pos=position in GRCH37. EA-FR=effect allele in fructosamine. B=effect size in fructosamine. SE=standard error of effect size in fructosamine. P=p-value in fructosamine. Gene=nearest gene. Trait=Trait where the association of the SNP was previously reported. EA-T=Effect allele of the associated trait. Binomial $p=5.6 \times 10^{-3}$.

4.5 Discussion

In this chapter, as is standard for studies of glycaemic measures, I examined the genetic influences on fructosamine levels in a healthy population where one can explore these influences in a non-diabetic setting where measures are unaffected by disease and its treatment. Specifically, I sought to quantify the heritability of fructosamine, identify loci affecting the trait and explore its genetic relationship with other glycaemic traits. Overall, my results show that, in contrast with HbA1c, the heritability for fructosamine is low, despite some evidence for shared genetic aetiology. Results also highlight a variant potentially regulating fructosamine levels through pathways that also regulate circulating albumin.

Firstly I established that, in agreement with previous twin studies [362], fructosamine appears to be a lowly heritable trait (2% (95% CI -2%-5%)) suggesting most of the variation of the trait in this population is due to environmental factors which is not surprising given the fact that this is a trait normally used to measure short term changes in glycaemia after treatment . Fructosamine also does not show evidence of significant genetic correlation with other glycaemic traits including HbA1c (p>0.05) which is somewhat surprising given the fact that both traits normally have a high phenotypic correlation (~0.61[360]) and reflect similar biological processes, namely, the glycation of serum proteins. This lack of genetic correlation was also observed in Loomis et al 2018 [363]. The HbA1c heritability estimate in this study (17% (95% CI 0%-35%) was higher than those reported in LD Hub (7% (95% CI 4%-9%))[231] and the one obtained using summary statistics from the latest published MAGIC effort [121] (6% (95% CI 5%-8%)) but this difference was not statistically significant due to the wide confidence intervals in this study. It is likely though, that the estimate from this

study is inflated and the actual heritability estimate is closer to the one obtained from the MAGIC HbA1c data as this dataset has a much larger sample size (>20 times as large) and therefore is better powered to obtain a more accurate estimate. HbA1c was genetically correlated with both glycaemic traits (FG and T2D) and erythrocytic traits (MCHC, MCH, MCV and RDW), which is consistent with what is known about the biology of HbA1c [121].

Despite the low heritability, I was able to detect two loci associated with fructosamine levels at genome-wide significance ($p<5x10^{-8}$). The first locus was rs853777, near *G6PC2*. *G6PC2* codes for Glucose-6-Phosphatase Catalytic Subunit 2 and it is a well-established locus in HbA1c and fasting glucose metabolism [83]. This protein is produced specifically in islet beta cells and is involved in regulation of insulin secretion [383]. A mouse knockout of this gene exhibits mild metabolic phenotypes (reduction of blood glucose with no impact on cholesterol, glycerol, insulin and glucagon concentrations or body weight) and enhanced islet responsiveness to blood glucose levels[384, 385], making it a feasible therapeutic target given that no deleterious consequences were observed after the knockout. Interestingly, this locus was only nominally associated with fructosamine levels in Loomis et al 2018 [363] (rs1402837, $p=0.016$, r^2 with rs853777=0.17). This therefore represents a novel association of this locus with fructosamine levels.

The other associated SNP was rs111476047 located downstream of *RCN3*. *RCN3* codes for Reticulocalbin 3, which is an EF-hand calcium-binding protein of poorly understood function [386]. This locus was previously associated with fructosamine levels and conditional analysis shows that the locus found in this study is possible more tightly linked to the true causal variant (**Table 4.4**). There was no expression information for the lead signal in GTEx but rs113886122, the second strongest SNP in this locus ($p=7.3x10^{-14}$, r^2 with

rs111476047=0.62) is an eQTL for *FCGRT* in tibial nerve, subcutaneous adipose, transverse colon, skin and transformed fibroblasts tissues [387]. *FCGRT* codes for Fc Fragment of IgG Receptor and Transporter which plays a role in maintenance of albumin levels protecting it from degradation [388]. In addition to this, mouse studies have shown that hepatic levels of this protein regulate albumin homeostasis and susceptibility to liver injury [389]. These results combined with the suggestive evidence of association with albumin levels (p=6.2x10^{-3}) suggest that the locus found in this study could influence fructosamine levels through pathways that also regulate albumin.

Finally, lookups of previously established glycaemic loci suggest that factors affecting other glycaemic traits such as HbA1c, fasting glucose and fasting insulin also influence fructosamine levels reflecting a shared genetic aetiology for these traits. As sample sizes increase, it is likely some of these signals will reach genome-wide significance.

The results in this chapter need further exploration given a few limitations. Firstly, fructosamine levels were unusually high (median=294 µmol/L) compared to established reference ranges (202-285 µmol/L). Secondly, the correlation of fructosamine levels and HbA1c was unexpectedly low (r=0.1) and this discrepancy appears to not be driven by unreliability of HbA1c measurements as these fell within expected ranges and were supported by genetic correlation results. Phenotype quality control did not address either of these issues which suggests that there might be other factors influencing these observations such as machine calibration issues. Nevertheless, fructosamine measurements seem reliable enough to produce biologically plausible association results.

Future studies on the genetic architecture of fructosamine will shed more light into the different glycaemic and non-glycaemic mechanisms that can affect fructosamine levels, and

potentially identify mechanisms that affect risk of comorbidities independently of other glycaemic traits such as fasting glucose and HbA1c. Furthermore, increasing the number of individuals with both HbA1c and fructosamine measured could help identify variants associated with protein glycation by examining the genetic influences of the glycation gap. Screening for these variants could be potentially useful in the clinic when testing for T2D using HbA1c as a diagnostic tool.

4.6 Future directions

While working on this analysis, the first GWAS on fructosamine levels was published [363]. This presents an opportunity to use summary statistics from that study for meta-analysis with my data. To my knowledge, ARIC and CARDIA, the cohorts used in this previous study, are the only cohorts with available fructosamine and genetic data therefore the only other dataset that can be combined with mine.

To further explore the influence of established glycaemic loci on fructosamine levels, I will build a GRS score for T2D, FG, albumin and HbA1c and test them on the fructosamine dataset for association. In addition to this, I can also explore whether there is an enrichment of rare variant associations in known glycaemic loci using the WES and WGS data in the INTERVAL cohort ($N_{WES+WGS}$=5,874).

Another possible avenue to explore is to perform multi-trait analysis with fructosamine and HbA1c, or T2D, to identify pleiotropic effects or to boost power in identification of loci affecting fructosamine levels exclusively. This can be achieved using a method that uses summary statistics as input such as MTAG [390] so I can combine summary statistics from this study with data from the MAGIC and DIAGRAM consortia.

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.

References

1. Lander, E.S. and N.J. Schork, *Genetic dissection of complex traits.* Science, 1994. **265**(5181): p. 2037-48.
2. Livshits, G., *Growth and development of bodyweight, height and head circumference during the first two years of life: quantitative genetic aspects.* Ann Hum Biol, 1986. **13**(4): p. 387-96.
3. Mueller, W.H., *Parent-child correlations for stature and weight among school aged children: A review of 24 studies.* Hum Biol, 1976. **48**(2): p. 379-97.
4. Rao, D.C., et al., *Analysis of family resemblance. V. Height and weight in northeastern Brazil.* Am J Hum Genet, 1975. **27**(4): p. 509-20.
5. Bogin, B. and L. Rios, *Rapid morphological change in living humans: implications for modern human origins.* Comp Biochem Physiol A Mol Integr Physiol, 2003. **136**(1): p. 71-84.
6. Singhal, A., et al., *Early nutrition and leptin concentrations in later life.* Am J Clin Nutr, 2002. **75**(6): p. 993-9.
7. World Health Organization. *Cardiovascular diseases (CVDs). [online]* 2018 27 Jul. 2018]; Available from: http://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds).
8. McGill, H.C., Jr., C.A. McMahan, and S.S. Gidding, *Preventing heart disease in the 21st century: implications of the Pathobiological Determinants of Atherosclerosis in Youth (PDAY) study.* Circulation, 2008. **117**(9): p. 1216-27.
9. Gersh, B.J., et al., *Novel therapeutic concepts: the epidemic of cardiovascular disease in the developing world: global implications.* Eur Heart J, 2010. **31**(6): p. 642-8.
10. Teo, K.K. and H. Dokainish, *The Emerging Epidemic of Cardiovascular Risk Factors and Atherosclerotic Disease in Developing Countries.* Can J Cardiol, 2017. **33**(3): p. 358-365.
11. Gaziano, T.A., et al., *Growing epidemic of coronary heart disease in low- and middle-income countries.* Curr Probl Cardiol, 2010. **35**(2): p. 72-115.
12. Kathiresan, S. and D. Srivastava, *Genetics of human cardiovascular disease.* Cell, 2012. **148**(6): p. 1242-57.
13. Finegold, J.A., P. Asaria, and D.P. Francis, *Mortality from ischaemic heart disease by country, region, and age: statistics from World Health Organisation and United Nations.* Int J Cardiol, 2013. **168**(2): p. 934-45.
14. Borrell, L.N., *The effects of smoking and physical inactivity on advancing mortality in U.S. adults.* Ann Epidemiol, 2014. **24**(6): p. 484-7.
15. de Souza, R.J., et al., *Intake of saturated and trans unsaturated fatty acids and risk of all cause mortality, cardiovascular disease, and type 2 diabetes: systematic review and meta-analysis of observational studies.* BMJ, 2015. **351**: p. h3978.
16. Baldo, M.P., S.L. Rodrigues, and J.G. Mill, *High salt intake as a multifaceted cardiovascular disease: new support from cellular and molecular evidence.* Heart Fail Rev, 2015. **20**(4): p. 461-74.
17. Fernandez-Sola, J., *Cardiovascular risks and benefits of moderate and heavy alcohol consumption.* Nat Rev Cardiol, 2015. **12**(10): p. 576-87.
18. O'Donnell, M.J., et al., *Risk factors for ischaemic and intracerebral haemorrhagic stroke in 22 countries (the INTERSTROKE study): a case-control study.* Lancet, 2010. **376**(9735): p. 112-23.
19. Eckel, R.H., *Obesity and heart disease: a statement for healthcare professionals from the Nutrition Committee, American Heart Association.* Circulation, 1997. **96**(9): p. 3248-50.
20. Navar-Boggan, A.M., et al., *Hyperlipidemia in early adulthood increases long-term risk of coronary heart disease.* Circulation, 2015. **131**(5): p. 451-8.
21. Rimm, E.B., et al., *Body size and fat distribution as predictors of coronary heart disease among middle-aged and older US men.* Am J Epidemiol, 1995. **141**(12): p. 1117-27.

22. Evans, M.F. and J. Frank, *Body weight and mortality among women.* Can Fam Physician, 1997. **43**: p. 455.
23. Khaodhiar, L., K.C. McCowen, and G.L. Blackburn, *Obesity and its comorbid conditions.* Clin Cornerstone, 1999. **2**(3): p. 17-31.
24. Taylor, V.H., et al., *The impact of obesity on quality of life.* Best Pract Res Clin Endocrinol Metab, 2013. **27**(2): p. 139-46.
25. World Health Organization. *Diabetes mellitus [online]*. 2018 27 Jul. 2018]; Available from: http://www.who.int/mediacentre/factsheets/fs138/en/.
26. Al-Goblan, A.S., M.A. Al-Alfi, and M.Z. Khan, *Mechanism linking diabetes mellitus and obesity.* Diabetes Metab Syndr Obes, 2014. **7**: p. 587-91.
27. Tancredi, M., et al., *Excess Mortality among Persons with Type 2 Diabetes.* N Engl J Med, 2015. **373**(18): p. 1720-32.
28. Teramoto, T., et al., *Primary hyperlipidemia.* J Atheroscler Thromb, 2008. **15**(2): p. 49-51.
29. Chait, A. and J.D. Brunzell, *Acquired hyperlipidemia (secondary dyslipoproteinemias).* Endocrinol Metab Clin North Am, 1990. **19**(2): p. 259-78.
30. Wray, N. and P. Visscher, *Estimating trait heritability.* Nature Education, 2008. **1**(1): p. 29.
31. Hill, W.G., M.E. Goddard, and P.M. Visscher, *Data and theory point to mainly additive genetic variance for complex traits.* PLoS Genet, 2008. **4**(2): p. e1000008.
32. Plomin, R. and D. Daniels, *Why are children in the same family so different from one another?* Int J Epidemiol, 2011. **40**(3): p. 563-82.
33. Boomsma, D., A. Busjahn, and L. Peltonen, *Classical twin studies and beyond.* Nat Rev Genet, 2002. **3**(11): p. 872-82.
34. Mook-Kanamori, D.O., et al., *Heritability estimates of body size in fetal life and early childhood.* PLoS One, 2012. **7**(7): p. e39901.
35. Sauce, B. and L.D. Matzel, *The paradox of intelligence: Heritability and malleability coexist in hidden gene-environment interplay.* Psychol Bull, 2018. **144**(1): p. 26-47.
36. Burton, P.R., M.D. Tobin, and J.L. Hopper, *Key concepts in genetic epidemiology.* Lancet, 2005. **366**(9489): p. 941-51.
37. Elks, C.E., et al., *Variability in the heritability of body mass index: a systematic review and meta-regression.* Front Endocrinol (Lausanne), 2012. **3**: p. 29.
38. Kettunen, J., et al., *Genome-wide association study identifies multiple loci influencing human serum metabolite levels.* Nat Genet, 2012. **44**(3): p. 269-76.
39. Yang, J., et al., *GCTA: a tool for genome-wide complex trait analysis.* Am J Hum Genet, 2011. **88**(1): p. 76-82.
40. Speed, D., et al., *Reevaluation of SNP heritability in complex human traits.* Nat Genet, 2017. **49**(7): p. 986-992.
41. Yang, J., et al., *Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index.* Nat Genet, 2015. **47**(10): p. 1114-20.
42. Bulik-Sullivan, B.K., et al., *LD Score regression distinguishes confounding from polygenicity in genome-wide association studies.* Nat Genet, 2015. **47**(3): p. 291-5.
43. Loh, P.R., et al., *Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis.* Nat Genet, 2015. **47**(12): p. 1385-92.
44. Kerem, B., et al., *Identification of the cystic fibrosis gene: genetic analysis.* Science, 1989. **245**(4922): p. 1073-80.
45. Riordan, J.R., et al., *Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA.* Science, 1989. **245**(4922): p. 1066-73.
46. Koenig, M., et al., *Complete cloning of the Duchenne muscular dystrophy (DMD) cDNA and preliminary genomic organization of the DMD gene in normal and affected individuals.* Cell, 1987. **50**(3): p. 509-17.
47. Strathdee, C.A., et al., *Cloning of cDNAs for Fanconi's anaemia by functional complementation.* Nature, 1992. **356**(6372): p. 763-7.

48. *A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. The Huntington's Disease Collaborative Research Group.* Cell, 1993. **72**(6): p. 971-83.
49. Gusella, J.F., et al., *A polymorphic DNA marker genetically linked to Huntington's disease.* Nature, 1983. **306**(5940): p. 234-8.
50. Hutchison, K.E., et al., *Population stratification in the candidate gene study: fatal threat or red herring?* Psychol Bull, 2004. **130**(1): p. 66-79.
51. Todorov, A.A. and D.C. Rao, *Trade-off between false positives and false negatives in the linkage analysis of complex traits.* Genet Epidemiol, 1997. **14**(5): p. 453-64.
52. Tabor, H.K., N.J. Risch, and R.M. Myers, *Candidate-gene approaches for studying complex genetic traits: practical considerations.* Nat Rev Genet, 2002. **3**(5): p. 391-7.
53. Roberts, S.B., et al., *Replication of linkage studies of complex traits: an examination of variation in location estimates.* Am J Hum Genet, 1999. **65**(3): p. 876-84.
54. Lander, E. and L. Kruglyak, *Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results.* Nat Genet, 1995. **11**(3): p. 241-7.
55. Altmuller, J., et al., *Genomewide scans of complex human diseases: true linkage is hard to find.* Am J Hum Genet, 2001. **69**(5): p. 936-50.
56. Risch, N. and K. Merikangas, *The future of genetic studies of complex human diseases.* Science, 1996. **273**(5281): p. 1516-7.
57. Reich, D.E. and E.S. Lander, *On the allelic spectrum of human disease.* Trends Genet, 2001. **17**(9): p. 502-10.
58. Daly, M.J., et al., *High-resolution haplotype structure in the human genome.* Nat Genet, 2001. **29**(2): p. 229-32.
59. Goldstein, D.B., *Islands of linkage disequilibrium.* Nat Genet, 2001. **29**(2): p. 109-11.
60. Gabriel, S.B., et al., *The structure of haplotype blocks in the human genome.* Science, 2002. **296**(5576): p. 2225-9.
61. Shifman, S., et al., *Linkage disequilibrium patterns of the human genome across populations.* Hum Mol Genet, 2003. **12**(7): p. 771-6.
62. Bush, W.S. and J.H. Moore, *Chapter 11: Genome-wide association studies.* PLoS Comput Biol, 2012. **8**(12): p. e1002822.
63. International HapMap, C., *The International HapMap Project.* Nature, 2003. **426**(6968): p. 789-96.
64. Li, M., C. Li, and W. Guan, *Evaluation of coverage variation of SNP chips for genome-wide association studies.* Eur J Hum Genet, 2008. **16**(5): p. 635-43.
65. Wang, D.G., et al., *Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome.* Science, 1998. **280**(5366): p. 1077-82.
66. LaFramboise, T., *Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances.* Nucleic Acids Res, 2009. **37**(13): p. 4181-93.
67. Haines, J.L., et al., *Complement factor H variant increases the risk of age-related macular degeneration.* Science, 2005. **308**(5720): p. 419-21.
68. Wellcome Trust Case Control, C., *Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls.* Nature, 2007. **447**(7145): p. 661-78.
69. Ioannidis, J.P., *Genetic associations: false or true?* Trends Mol Med, 2003. **9**(4): p. 135-8.
70. Ioannidis, J.P., N.A. Patsopoulos, and E. Evangelou, *Heterogeneity in meta-analyses of genome-wide association investigations.* PLoS One, 2007. **2**(9): p. e841.
71. Stephens, M., N.J. Smith, and P. Donnelly, *A new statistical method for haplotype reconstruction from population data.* Am J Hum Genet, 2001. **68**(4): p. 978-89.
72. Genomes Project, C., et al., *A map of human genome variation from population-scale sequencing.* Nature, 2010. **467**(7319): p. 1061-73.
73. Marchini, J. and B. Howie, *Genotype imputation for genome-wide association studies.* Nat Rev Genet, 2010. **11**(7): p. 499-511.

74. Willer, C.J., et al., *Six new loci associated with body mass index highlight a neuronal influence on body weight regulation.* Nat Genet, 2009. **41**(1): p. 25-34.
75. Zeggini, E., et al., *Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes.* Nat Genet, 2008. **40**(5): p. 638-45.
76. Prokopenko, I., et al., *Variants in MTNR1B influence fasting glucose levels.* Nat Genet, 2009. **41**(1): p. 77-81.
77. Willer, C.J., et al., *Newly identified loci that influence lipid concentrations and risk of coronary artery disease.* Nat Genet, 2008. **40**(2): p. 161-9.
78. Consortium, C.A.D., et al., *Large-scale association analysis identifies new risk loci for coronary artery disease.* Nat Genet, 2013. **45**(1): p. 25-33.
79. MacArthur, J., et al., *The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog).* Nucleic Acids Res, 2017. **45**(D1): p. D896-D901.
80. Yengo, L., et al., *Meta-analysis of genome-wide association studies for height and body mass index in approximately 700000 individuals of European ancestry.* Hum Mol Genet, 2018.
81. Yang, J., et al., *Common SNPs explain a large proportion of the heritability for human height.* Nat Genet, 2010. **42**(7): p. 565-9.
82. Fuchsberger, C., et al., *The genetic architecture of type 2 diabetes.* Nature, 2016. **536**(7614): p. 41-47.
83. Dupuis, J., et al., *New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk.* Nat Genet, 2010. **42**(2): p. 105-16.
84. Liu, D.J., et al., *Exome-wide association study of plasma lipids in >300,000 individuals.* Nat Genet, 2017. **49**(12): p. 1758-1766.
85. Johansen, C.T., et al., *Excess of rare variants in genes identified by genome-wide association study of hypertriglyceridemia.* Nat Genet, 2010. **42**(8): p. 684-7.
86. Chong, J.X., et al., *The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities.* Am J Hum Genet, 2015. **97**(2): p. 199-215.
87. LeMaire, S.A., et al., *Genome-wide association study identifies a susceptibility locus for thoracic aortic aneurysms and aortic dissections spanning FBN1 at 15q21.1.* Nat Genet, 2011. **43**(10): p. 996-1000.
88. Newton-Cheh, C., et al., *Common variants at ten loci influence QT interval duration in the QTGEN Study.* Nat Genet, 2009. **41**(4): p. 399-406.
89. Hegele, R.A., *Plasma lipoproteins: genetic influences and clinical implications.* Nat Rev Genet, 2009. **10**(2): p. 109-21.
90. Antonarakis, S.E., et al., *Mendelian disorders and multifactorial traits: the big divide or one for all?* Nat Rev Genet, 2010. **11**(5): p. 380-4.
91. Consortium, U.K., et al., *The UK10K project identifies rare variants in health and disease.* Nature, 2015. **526**(7571): p. 82-90.
92. Locke, A.E., et al., *Genetic studies of body mass index yield new insights for obesity biology.* Nature, 2015. **518**(7538): p. 197-206.
93. Turcot, V., et al., *Protein-altering variants associated with body mass index implicate pathways that control energy intake and expenditure in obesity.* Nat Genet, 2018. **50**(1): p. 26-41.
94. Shungin, D., et al., *New genetic loci link adipose and insulin biology to body fat distribution.* Nature, 2015. **518**(7538): p. 187-196.
95. Frayling, T.M., *Genome-wide association studies provide new insights into type 2 diabetes aetiology.* Nat Rev Genet, 2007. **8**(9): p. 657-62.
96. Morris, A.P., et al., *Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes.* Nat Genet, 2012. **44**(9): p. 981-90.
97. Lyssenko, V., et al., *Common variant in MTNR1B associated with increased risk of type 2 diabetes and impaired early insulin secretion.* Nat Genet, 2009. **41**(1): p. 82-8.

98. Renstrom, F., et al., *Season-dependent associations of circadian rhythm-regulating loci (CRY1, CRY2 and MTNR1B) and glucose homeostasis: the GLACIER Study.* Diabetologia, 2015. **58**(5): p. 997-1005.
99. Pasquali, L., et al., *Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants.* Nat Genet, 2014. **46**(2): p. 136-143.
100. Lotta, L.A., et al., *Integrative genomic analysis implicates limited peripheral adipose storage capacity in the pathogenesis of human insulin resistance.* Nat Genet, 2017. **49**(1): p. 17-26.
101. Musunuru, K., et al., *From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus.* Nature, 2010. **466**(7307): p. 714-9.
102. Knowles, J.W. and E.A. Ashley, *Cardiovascular disease: The rise of the genetic risk score.* PLoS Med, 2018. **15**(3): p. e1002546.
103. Romanos, J., et al., *Improving coeliac disease risk prediction by testing non-HLA variants additional to HLA variants.* Gut, 2014. **63**(3): p. 415-22.
104. Sharp, S.A., et al., *Clinical and research uses of genetic risk scores in type 1 diabetes.* Curr Opin Genet Dev, 2018. **50**: p. 96-102.
105. Loos, R.J.F. and A. Janssens, *Predicting Polygenic Obesity Using Genetic Information.* Cell Metab, 2017. **25**(3): p. 535-543.
106. Goodarzi, M.O., *Genetics of obesity: what genetic association studies have taught us about the biology of obesity and its complications.* Lancet Diabetes Endocrinol, 2018. **6**(3): p. 223-236.
107. Afzal, S., et al., *Genetically low vitamin D concentrations and increased mortality: Mendelian randomisation analysis in three large cohorts.* BMJ, 2014. **349**: p. g6330.
108. Cho, Y., et al., *Alcohol intake and cardiovascular risk factors: A Mendelian randomisation study.* Sci Rep, 2015. **5**: p. 18422.
109. Hagg, S., et al., *Adiposity as a cause of cardiovascular disease: a Mendelian randomization study.* Int J Epidemiol, 2015. **44**(2): p. 578-86.
110. Voight, B.F., et al., *Plasma HDL cholesterol and risk of myocardial infarction: a mendelian randomisation study.* Lancet, 2012. **380**(9841): p. 572-80.
111. Haase, C.L., et al., *LCAT, HDL cholesterol and ischemic cardiovascular disease: a Mendelian randomization study of HDL cholesterol in 54,500 individuals.* J Clin Endocrinol Metab, 2012. **97**(2): p. E248-56.
112. Agerholm-Larsen, B., et al., *Elevated HDL cholesterol is a risk factor for ischemic heart disease in white women when caused by a common mutation in the cholesteryl ester transfer protein gene.* Circulation, 2000. **101**(16): p. 1907-12.
113. Andersen, R.V., et al., *Hepatic lipase mutations,elevated high-density lipoprotein cholesterol, and increased risk of ischemic heart disease: the Copenhagen City Heart Study.* J Am Coll Cardiol, 2003. **41**(11): p. 1972-82.
114. Frikke-Schmidt, R., et al., *Association of loss-of-function mutations in the ABCA1 gene with high-density lipoprotein cholesterol levels and risk of ischemic heart disease.* JAMA, 2008. **299**(21): p. 2524-32.
115. Bennett, D.A. and M.V. Holmes, *Mendelian randomisation in cardiovascular research: an introduction for clinicians.* Heart, 2017. **103**(18): p. 1400-1407.
116. Teslovich, T.M., et al., *Biological, clinical and population relevance of 95 loci for blood lipids.* Nature, 2010. **466**(7307): p. 707-13.
117. Pollin, T.I., et al., *A null mutation in human APOC3 confers a favorable plasma lipid profile and apparent cardioprotection.* Science, 2008. **322**(5908): p. 1702-5.
118. Flannick, J., et al., *Loss-of-function mutations in SLC30A8 protect against type 2 diabetes.* Nat Genet, 2014. **46**(4): p. 357-63.
119. Pearson, E.R., et al., *Variation in TCF7L2 influences therapeutic response to sulfonylureas: a GoDARTs study.* Diabetes, 2007. **56**(8): p. 2178-82.

120. Aslibekyan, S., et al., *Variants identified in a GWAS meta-analysis for blood lipids are associated with the lipid response to fenofibrate.* PLoS One, 2012. **7**(10): p. e48663.
121. Wheeler, E., et al., *Impact of common genetic determinants of Hemoglobin A1c on type 2 diabetes risk and diagnosis in ancestrally diverse populations: A transethnic genome-wide meta-analysis.* PLoS Med, 2017. **14**(9): p. e1002383.
122. Price, A.L., C.C. Spencer, and P. Donnelly, *Progress and promise in understanding the genetic basis of common diseases.* Proc Biol Sci, 2015. **282**(1821).
123. Pritchard, J.K., *Are rare variants responsible for susceptibility to complex diseases?* Am J Hum Genet, 2001. **69**(1): p. 124-37.
124. Boyle, E.A., Y.I. Li, and J.K. Pritchard, *An Expanded View of Complex Traits: From Polygenic to Omnigenic.* Cell, 2017. **169**(7): p. 1177-1186.
125. *The Haplotype Reference Consortium*. Available from: http://www.haplotype-reference-consortium.org.
126. Huang, J., et al., *Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel.* Nat Commun, 2015. **6**: p. 8111.
127. McCarthy, S., et al., *A reference panel of 64,976 haplotypes for genotype imputation.* Nat Genet, 2016. **48**(10): p. 1279-83.
128. Bustamante, C.D., et al., *Natural selection on protein-coding genes in the human genome.* Nature, 2005. **437**(7062): p. 1153-7.
129. Lim, E.T., et al., *Distribution and medical impact of loss-of-function variants in the Finnish founder population.* PLoS Genet, 2014. **10**(7): p. e1004494.
130. *UK Biobank Axiom Array*. Available from: http://www.ukbiobank.ac.uk/scientists-3/uk-biobank-axiom-array/.
131. Kotowski, I.K., et al., *A spectrum of PCSK9 alleles contributes to plasma levels of low-density lipoprotein cholesterol.* Am J Hum Genet, 2006. **78**(3): p. 410-22.
132. Cohen, J.C., et al., *Multiple rare alleles contribute to low plasma levels of HDL cholesterol.* Science, 2004. **305**(5685): p. 869-72.
133. Romeo, S., et al., *Population-based resequencing of ANGPTL4 uncovers variations that reduce triglycerides and increase HDL.* Nat Genet, 2007. **39**(4): p. 513-6.
134. Sudlow, C., et al., *UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age.* PLoS Med, 2015. **12**(3): p. e1001779.
135. Sankar, P.L. and L.S. Parker, *The Precision Medicine Initiative's All of Us Research Program: an agenda for research on its ethical, legal, and social issues.* Genet Med, 2017. **19**(7): p. 743-750.
136. Chen, Z., et al., *China Kadoorie Biobank of 0.5 million people: survey methods, baseline characteristics and long-term follow-up.* Int J Epidemiol, 2011. **40**(6): p. 1652-66.
137. Higgins, J.P., et al., *Measuring inconsistency in meta-analyses.* BMJ, 2003. **327**(7414): p. 557-60.
138. Nakaoka, H. and I. Inoue, *Meta-analysis of genetic association studies: methodologies, between-study heterogeneity and winner's curse.* J Hum Genet, 2009. **54**(11): p. 615-23.
139. Lin, D.Y. and P.F. Sullivan, *Meta-analysis of genome-wide association studies with overlapping subjects.* Am J Hum Genet, 2009. **85**(6): p. 862-72.
140. Galesloot, T.E., et al., *A comparison of multivariate genome-wide association methods.* PLoS One, 2014. **9**(4): p. e95923.
141. O'Reilly, P.F., et al., *MultiPhen: joint model of multiple phenotypes can increase discovery in GWAS.* PLoS One, 2012. **7**(5): p. e34861.
142. Furlotte, N.A. and E. Eskin, *Efficient multiple-trait association and estimation of genetic correlation using the matrix-variate linear mixed model.* Genetics, 2015. **200**(1): p. 59-68.
143. Denny, J.C., et al., *Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data.* Nat Biotechnol, 2013. **31**(12): p. 1102-10.

144. Denny, J.C., et al., *PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations.* Bioinformatics, 2010. **26**(9): p. 1205-10.
145. Ritchie, M.D., et al., *Genome- and phenome-wide analyses of cardiac conduction identifies markers of arrhythmia risk.* Circulation, 2013. **127**(13): p. 1377-85.
146. Tyrrell, J., et al., *Gene-obesogenic environment interactions in the UK Biobank study.* Int J Epidemiol, 2017. **46**(2): p. 559-575.
147. Moore, R., et al., *A linear mixed model approach to study multivariate gene-environment interactions.* bioRxiv, 2018.
148. Emdin, C.A., et al., *Analysis of predicted loss-of-function variants in UK Biobank identifies variants protective for disease.* Nat Commun, 2018. **9**(1): p. 1613.
149. Dhandapany, P.S., et al., *A common MYBPC3 (cardiac myosin binding protein C) variant associated with cardiomyopathies in South Asia.* Nat Genet, 2009. **41**(2): p. 187-91.
150. Kurki, M.I., et al., *High risk population isolate reveals low frequency variants predisposing to intracranial aneurysms.* PLoS Genet, 2014. **10**(1): p. e1004134.
151. Consortium, S.T.D., et al., *Association of a low-frequency variant in HNF1A with type 2 diabetes in a Latino population.* JAMA, 2014. **311**(22): p. 2305-14.
152. Genomes Project, C., et al., *An integrated map of genetic variation from 1,092 human genomes.* Nature, 2012. **491**(7422): p. 56-65.
153. Hatzikotoulas, K., A. Gilly, and E. Zeggini, *Using population isolates in genetic association studies.* Brief Funct Genomics, 2014. **13**(5): p. 371-7.
154. Asimit, J.L., et al., *Trans-ethnic study design approaches for fine-mapping.* Eur J Hum Genet, 2016. **24**(9): p. 1330-6.
155. Hu, Y., et al., *Discovery and fine-mapping of loci associated with MUFAs through trans-ethnic meta-analysis in Chinese and European populations.* J Lipid Res, 2017. **58**(5): p. 974-981.
156. Kuo, J.Z., et al., *Trans-ethnic fine mapping identifies a novel independent locus at the 3' end of CDKAL1 and novel variants of several susceptibility loci for type 2 diabetes in a Han Chinese population.* Diabetologia, 2013. **56**(12): p. 2619-28.
157. Magi, R., et al., *Trans-ethnic meta-regression of genome-wide association studies accounting for ancestry increases power for discovery and improves fine-mapping resolution.* Hum Mol Genet, 2017. **26**(18): p. 3639-3650.
158. Sebat, J., et al., *Strong association of de novo copy number mutations with autism.* Science, 2007. **316**(5823): p. 445-9.
159. Walsh, T., et al., *Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia.* Science, 2008. **320**(5875): p. 539-43.
160. Wheeler, E., et al., *Genome-wide SNP and CNV analysis identifies common and low-frequency variants associated with severe early-onset obesity.* Nat Genet, 2013. **45**(5): p. 513-7.
161. Glessner, J.T., et al., *A genome-wide study reveals copy number variants exclusive to childhood obesity cases.* Am J Hum Genet, 2010. **87**(5): p. 661-6.
162. Gonzalez, J.R., et al., *A common 16p11.2 inversion underlies the joint susceptibility to asthma and obesity.* Am J Hum Genet, 2014. **94**(3): p. 361-72.
163. Mace, A., et al., *CNV-association meta-analysis in 191,161 European adults reveals new loci associated with anthropometric traits.* Nat Commun, 2017. **8**(1): p. 744.
164. Dajani, R., et al., *CNV Analysis Associates AKNAD1 with Type-2 Diabetes in Jordan Subpopulations.* Sci Rep, 2015. **5**: p. 13391.
165. Hayes, J.L., et al., *Diagnosis of copy number variation by Illumina next generation sequencing is comparable in performance to oligonucleotide array comparative genomic hybridisation.* Genomics, 2013. **102**(3): p. 174-81.
166. Hehir-Kwa, J.Y., et al., *Genome-wide copy number profiling on high-density bacterial artificial chromosomes, single-nucleotide polymorphisms, and oligonucleotide microarrays: a platform comparison based on statistical power analysis.* DNA Res, 2007. **14**(1): p. 1-11.

167. Feng, Y., et al., *Improved molecular diagnosis by the detection of exonic deletions with target gene capture and deep sequencing.* Genet Med, 2015. **17**(2): p. 99-107.
168. Russo, C.D., et al., *Comparative study of aCGH and Next Generation Sequencing (NGS) for chromosomal microdeletion and microduplication screening.* J Prenat Med, 2014. **8**(3-4): p. 57-69.
169. Fromer, M., et al., *Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth.* Am J Hum Genet, 2012. **91**(4): p. 597-607.
170. Handsaker, R.E., et al., *Discovery and genotyping of genome structural polymorphism by sequencing on a population scale.* Nat Genet, 2011. **43**(3): p. 269-76.
171. Shin, S.Y., et al., *An atlas of genetic influences on human blood metabolites.* Nat Genet, 2014. **46**(6): p. 543-550.
172. Suhre, K., et al., *Human metabolic individuality in biomedical and pharmaceutical research.* Nature, 2011. **477**(7362): p. 54-60.
173. Kettunen, J., et al., *Genome-wide study for circulating metabolites identifies 62 loci and reveals novel systemic effects of LPA.* Nat Commun, 2016. **7**: p. 11122.
174. Folkersen, L., et al., *Mapping of 79 loci for 83 plasma protein biomarkers in cardiovascular disease.* PLoS Genet, 2017. **13**(4): p. e1006706.
175. Global, B.M.I.M.C., et al., *Body-mass index and all-cause mortality: individual-participant-data meta-analysis of 239 prospective studies in four continents.* Lancet, 2016. **388**(10046): p. 776-86.
176. Ogden, C.L., M.D. Carroll, and K.M. Flegal, *Prevalence of obesity in the United States.* JAMA, 2014. **312**(2): p. 189-90.
177. Wardle, J., et al., *Evidence for a strong genetic influence on childhood adiposity despite the force of the obesogenic environment.* Am J Clin Nutr, 2008. **87**(2): p. 398-404.
178. Silventoinen, K., et al., *Heritability of body size and muscle strength in young adulthood: a study of one million Swedish men.* Genet Epidemiol, 2008. **32**(4): p. 341-9.
179. Allison, D.B., et al., *The heritability of body mass index among an international sample of monozygotic twins reared apart.* Int J Obes Relat Metab Disord, 1996. **20**(6): p. 501-6.
180. Akiyama, M., et al., *Genome-wide association study identifies 112 new loci for body mass index in the Japanese population.* Nat Genet, 2017. **49**(10): p. 1458-1467.
181. Grarup, N., et al., *Loss-of-function variants in ADCY3 increase risk of obesity and type 2 diabetes.* Nat Genet, 2018. **50**(2): p. 172-174.
182. Justice, A.E., et al., *Genome-wide meta-analysis of 241,258 adults accounting for smoking behaviour identifies novel loci for obesity traits.* Nat Commun, 2017. **8**: p. 14977.
183. Minster, R.L., et al., *A thrifty variant in CREBRF strongly influences body mass index in Samoans.* Nat Genet, 2016. **48**(9): p. 1049-1054.
184. Pigeyre, M., et al., *Recent progress in genetics, epigenetics and metagenomics unveils the pathophysiology of human obesity.* Clin Sci (Lond), 2016. **130**(12): p. 943-86.
185. Ramachandrappa, S., et al., *Rare variants in single-minded 1 (SIM1) are associated with severe obesity.* J Clin Invest, 2013. **123**(7): p. 3042-50.
186. Doche, M.E., et al., *Human SH2B1 mutations are associated with maladaptive behaviors and obesity.* J Clin Invest, 2012. **122**(12): p. 4732-6.
187. O'Rahilly, S. and I.S. Farooqi, *Human obesity: a heritable neurobehavioral disorder that is highly sensitive to environmental conditions.* Diabetes, 2008. **57**(11): p. 2905-10.
188. Saeed, S., et al., *Loss-of-function mutations in ADCY3 cause monogenic severe obesity.* Nat Genet, 2018. **50**(2): p. 175-179.
189. Clement, K., et al., *A mutation in the human leptin receptor gene causes obesity and pituitary dysfunction.* Nature, 1998. **392**(6674): p. 398-401.
190. Krude, H., et al., *Severe early-onset obesity, adrenal insufficiency and red hair pigmentation caused by POMC mutations in humans.* Nat Genet, 1998. **19**(2): p. 155-7.

191. Montague, C.T., et al., *Congenital leptin deficiency is associated with severe early-onset obesity in humans.* Nature, 1997. **387**(6636): p. 903-8.
192. Gray, J., et al., *Hyperphagia, severe obesity, impaired cognitive function, and hyperactivity associated with functional loss of one copy of the brain-derived neurotrophic factor (BDNF) gene.* Diabetes, 2006. **55**(12): p. 3366-71.
193. Collet, T.H., et al., *Evaluation of a melanocortin-4 receptor (MC4R) agonist (Setmelanotide) in MC4R deficiency.* Mol Metab, 2017. **6**(10): p. 1321-1329.
194. Yeo, G.S., et al., *A frameshift mutation in MC4R associated with dominantly inherited human obesity.* Nat Genet, 1998. **20**(2): p. 111-2.
195. Vaisse, C., et al., *A frameshift mutation in human MC4R is associated with a dominant form of obesity.* Nat Genet, 1998. **20**(2): p. 113-4.
196. Loos, R.J., et al., *Common variants near MC4R are associated with fat mass, weight and risk of obesity.* Nat Genet, 2008. **40**(6): p. 768-75.
197. Bulik, C.M. and D.B. Allison, *The genetic epidemiology of thinness.* Obes Rev, 2001. **2**(2): p. 107-15.
198. Costanzo, P.R. and S.S. Schiffman, *Thinness--not obesity--has a genetic component.* Neurosci Biobehav Rev, 1989. **13**(1): p. 55-8.
199. Magnusson, P.K. and F. Rasmussen, *Familial resemblance of body mass index and familial risk of high and low body mass index. A study of young men in Sweden.* Int J Obes Relat Metab Disord, 2002. **26**(9): p. 1225-31.
200. Laskarzewski, P.M., et al., *Familial obesity and leanness.* Int J Obes, 1983. **7**(6): p. 505-27.
201. Whitaker, K.L., et al., *The intergenerational transmission of thinness.* Arch Pediatr Adolesc Med, 2011. **165**(10): p. 900-5.
202. Jacquemont, S., et al., *Mirror extreme BMI phenotypes associated with gene dosage at the chromosome 16p11.2 locus.* Nature, 2011. **478**(7367): p. 97-102.
203. Hasstedt, S.J., et al., *A Copy Number Variant on Chromosome 20q13.3 Implicated in Thinness and Severe Obesity.* J Obes, 2015. **2015**: p. 623431.
204. Hinney, A., et al., *Genome wide association (GWA) study for early onset extreme obesity supports the role of fat mass and obesity associated gene (FTO) variants.* PLoS One, 2007. **2**(12): p. e1361.
205. Scannell Bryan, M., et al., *Genome-wide association studies and heritability estimates of body mass index related phenotypes in Bangladeshi adults.* PLoS One, 2014. **9**(8): p. e105062.
206. Braud, S., M. Ciufolini, and I. Harosh, *'Energy expenditure genes' or 'energy absorption genes': a new target for the treatment of obesity and Type II diabetes.* Future Med Chem, 2010. **2**(12): p. 1777-83.
207. Berndt, S.I., et al., *Genome-wide meta-analysis identifies 11 new loci for anthropometric traits and provides insights into genetic architecture.* Nat Genet, 2013. **45**(5): p. 501-12.
208. Speliotes, E.K., et al., *Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index.* Nat Genet, 2010. **42**(11): p. 937-48.
209. Cornish, K.M., et al., *Association of the dopamine transporter (DAT1) 10/10-repeat genotype with ADHD symptoms and response inhibition in a general population sample.* Mol Psychiatry, 2005. **10**(7): p. 686-98.
210. Emond, M.J., et al., *Exome sequencing of extreme phenotypes identifies DCTN4 as a modifier of chronic Pseudomonas aeruginosa infection in cystic fibrosis.* Nat Genet, 2012. **44**(8): p. 886-9.
211. Lanktree, M.B., et al., *Extremes of unexplained variation as a phenotype: an efficient approach for genome-wide association studies of cardiovascular disease.* Circ Cardiovasc Genet, 2010. **3**(2): p. 215-21.

212. Zhou, Y.J., Y. Wang, and L.L. Chen, *Detecting the Common and Individual Effects of Rare Variants on Quantitative Traits by Using Extreme Phenotype Sampling.* Genes (Basel), 2016. **7**(1).
213. Morgan, J.F., F. Reid, and J.H. Lacey, *The SCOFF questionnaire: assessment of a new screening tool for eating disorders.* BMJ, 1999. **319**(7223): p. 1467-8.
214. Bochukova, E.G., et al., *Large, rare chromosomal deletions associated with severe early-onset obesity.* Nature, 2010. **463**(7281): p. 666-70.
215. University of Essex. Institute for Social and Economic Research and NatCen Social Research, *Understanding Society: Waves 1-5, 2009-2014 [computer file]*. 7th Edition. Colchester, Essex: UK Data Archive [distributor] November 2015 SN: 6614.
216. Wain, L.V., et al., *Novel insights into the genetics of smoking behaviour, lung function, and chronic obstructive pulmonary disease (UK BiLEVE): a genetic association study in UK Biobank.* Lancet Respir Med, 2015. **3**(10): p. 769-81.
217. Riveros-Mckay, F., et al., *Genetic architecture of human thinness compared to severe obesity.* PLoS Genet, 2018. **[in press]**.
218. Ma, C., et al., *Recommended joint and meta-analysis strategies for case-control association testing of single low-count variants.* Genet Epidemiol, 2013. **37**(6): p. 539-50.
219. Boyd, A., et al., *Cohort Profile: the 'children of the 90s'--the index offspring of the Avon Longitudinal Study of Parents and Children.* Int J Epidemiol, 2013. **42**(1): p. 111-27.
220. Fraser, A., et al., *Cohort Profile: the Avon Longitudinal Study of Parents and Children: ALSPAC mothers cohort.* Int J Epidemiol, 2013. **42**(1): p. 97-110.
221. Howie, B., J. Marchini, and M. Stephens, *Genotype imputation with thousands of genomes.* G3 (Bethesda), 2011. **1**(6): p. 457-70.
222. Howie, B.N., P. Donnelly, and J. Marchini, *A flexible and accurate genotype imputation method for the next generation of genome-wide association studies.* PLoS Genet, 2009. **5**(6): p. e1000529.
223. Purcell, S., et al., *PLINK: a tool set for whole-genome association and population-based linkage analyses.* Am J Hum Genet, 2007. **81**(3): p. 559-75.
224. Price, A.L., et al., *Principal components analysis corrects for stratification in genome-wide association studies.* Nat Genet, 2006. **38**(8): p. 904-9.
225. Manichaikul, A., et al., *Robust relationship inference in genome-wide association studies.* Bioinformatics, 2010. **26**(22): p. 2867-73.
226. Marchini, J., et al., *A new multipoint method for genome-wide association studies by imputation of genotypes.* Nat Genet, 2007. **39**(7): p. 906-13.
227. Loh, P.R., et al., *Efficient Bayesian mixed-model analysis increases association power in large cohorts.* Nat Genet, 2015. **47**(3): p. 284-90.
228. Bulik-Sullivan, B., et al., *An atlas of genetic correlations across human diseases and traits.* Nat Genet, 2015. **47**(11): p. 1236-41.
229. *Measuring and interpreting BMI in Children :: Public Health England Obesity Knowledge and Intelligence team*. 19th December 2016]; Available from: https://www.noo.org.uk/NOO_about_obesity/measurement/children.
230. Denaxas, S.C., et al., *Data resource profile: cardiovascular disease research using linked bespoke studies and electronic health records (CALIBER).* Int J Epidemiol, 2012. **41**(6): p. 1625-38.
231. Zheng, J., et al., *LD Hub: a centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis.* Bioinformatics, 2017. **33**(2): p. 272-279.
232. Harrell, F.E., *rms: R functions for biostatistical/epidemiologic modeling, testing, estimation, validation, graphics, prediction, and typesetting by storing enhanced model design attributes in the fit, 2013. Implements methods in Regression Modeling Strategies, New York:Springer, 2001.*

233. R Development Core Team, *R: A Language and Environment for Statistical Computing*. 2011, R Foundation for Statistical Computing: Vienna, Austria.
234. Gauderman, W. and J. Morrison, *QUANTO 1.1: A computer program for power and sample size calculations for genetic-epidemiology studies, http://hydra.usc.edu/gxe*. 2006.
235. Canela-Xandri, O., et al., *A new tool called DISSECT for analysing large genomic data sets using a Big Data approach.* Nat Commun, 2015. **6**: p. 10162.
236. Bradfield, J.P., et al., *A genome-wide association meta-analysis identifies new childhood obesity loci.* Nat Genet, 2012. **44**(5): p. 526-31.
237. Southam, L., et al., *Whole genome sequencing and imputation in isolated populations identify genetic associations with medically-relevant complex traits.* Nat Commun, 2017. **8**: p. 15606.
238. Willer, C.J., Y. Li, and G.R. Abecasis, *METAL: fast and efficient meta-analysis of genomewide association scans.* Bioinformatics, 2010. **26**(17): p. 2190-1.
239. Xu, C., et al., *Estimating genome-wide significance for whole-genome sequencing studies.* Genet Epidemiol, 2014. **38**(4): p. 281-90.
240. Hinney, A., A.L. Volckmar, and N. Knoll, *Melanocortin-4 receptor in energy homeostasis and obesity pathogenesis.* Prog Mol Biol Transl Sci, 2013. **114**: p. 147-91.
241. Geller, F., et al., *Melanocortin-4 receptor gene variant I103 is negatively associated with obesity.* Am J Hum Genet, 2004. **74**(3): p. 572-81.
242. Klimentidis, Y.C., et al., *Genome-wide association study of habitual physical activity in over 377,000 UK Biobank participants identifies multiple variants including CADM2 and APOE.* Int J Obes (Lond), 2018.
243. Mitchell, J.A., et al., *Obesity-susceptibility loci and the tails of the pediatric BMI distribution.* Obesity (Silver Spring), 2013. **21**(6): p. 1256-60.
244. Beyerlein, A., et al., *Genetic markers of obesity risk: stronger associations with body composition in overweight compared to normal-weight children.* PLoS One, 2011. **6**(4): p. e19057.
245. Chan, Y., et al., *Common variants show predicted polygenic effects on height in the tails of the distribution, except in extremely short individuals.* PLoS Genet, 2011. **7**(12): p. e1002439.
246. Young, A.I., F. Wauthier, and P. Donnelly, *Multiple novel gene-by-environment interactions modify the effect of FTO variants on body mass index.* Nat Commun, 2016. **7**: p. 12724.
247. Winkler, T.W., et al., *The Influence of Age and Sex on Genetic Associations with Adult Body Size and Shape: A Large-Scale Genome-Wide Interaction Study.* PLoS Genet, 2015. **11**(10): p. e1005378.
248. Qi, Q., et al., *FTO genetic variants, dietary intake and body mass index: insights from 177,330 individuals.* Hum Mol Genet, 2014. **23**(25): p. 6961-72.
249. Bjornland, T., et al., *Assessing gene-environment interaction effects of FTO, MC4R and lifestyle factors on obesity using an extreme phenotype sampling design: Results from the HUNT study.* PLoS One, 2017. **12**(4): p. e0175071.
250. Shungin, D., et al., *New genetic loci link adipose and insulin biology to body fat distribution.* Nature, 2015. **518**(7538): p. 187-96.
251. Tachmazidou, I., et al., *Whole-Genome Sequencing Coupled to Imputation Discovers Genetic Signals for Anthropometric Traits.* Am J Hum Genet, 2017. **100**(6): p. 865-884.
252. Lango Allen, H., et al., *Hundreds of variants clustered in genomic loci and biological pathways affect human height.* Nature, 2010. **467**(7317): p. 832-8.
253. Wen, W., et al., *Genome-wide association studies in East Asians identify new loci for waist-hip ratio and waist circumference.* Sci Rep, 2016. **6**: p. 17958.
254. Shaheen, R., et al., *A founder CEP120 mutation in Jeune asphyxiating thoracic dystrophy expands the role of centriolar proteins in skeletal ciliopathies.* Hum Mol Genet, 2015. **24**(5): p. 1410-9.

255. Roosing, S., et al., *Mutations in CEP120 cause Joubert syndrome as well as complex ciliopathy phenotypes.* J Med Genet, 2016. **53**(9): p. 608-15.
256. Athersuch, T.J. and H.C. Keun, *Metabolic profiling in human exposome studies.* Mutagenesis, 2015. **30**(6): p. 755-62.
257. Lydic, T.A. and Y.H. Goo, *Lipidomics unveils the complexity of the lipidome in metabolic diseases.* Clin Transl Med, 2018. **7**(1): p. 4.
258. Arsenault, B.J., S.M. Boekholdt, and J.J. Kastelein, *Lipid parameters for measuring risk of cardiovascular disease.* Nat Rev Cardiol, 2011. **8**(4): p. 197-206.
259. Nordestgaard, B.G. and A. Varbo, *Triglycerides and cardiovascular disease.* Lancet, 2014. **384**(9943): p. 626-635.
260. Varbo, A., M. Benn, and B.G. Nordestgaard, *Remnant cholesterol as a cause of ischemic heart disease: evidence, definition, measurement, atherogenicity, high risk patients, and present and future treatment.* Pharmacol Ther, 2014. **141**(3): p. 358-67.
261. Wyler von Ballmoos, M.C., B. Haring, and F.M. Sacks, *The risk of cardiovascular events with increased apolipoprotein CIII: A systematic review and meta-analysis.* J Clin Lipidol, 2015. **9**(4): p. 498-510.
262. Geng, P., et al., *Serum mannose-binding lectin is a strong biomarker of diabetic retinopathy in chinese patients with diabetes.* Diabetes Care, 2015. **38**(5): p. 868-75.
263. Trpkovic, A., et al., *Oxidized low-density lipoprotein as a biomarker of cardiovascular diseases.* Crit Rev Clin Lab Sci, 2015. **52**(2): p. 70-85.
264. *Strategies for the prevention of coronary heart disease: a policy statement of the European Atherosclerosis Society.* Eur Heart J, 1987. **8**(1): p. 77-88.
265. Gordon, T., et al., *High density lipoprotein as a protective factor against coronary heart disease. The Framingham Study.* Am J Med, 1977. **62**(5): p. 707-14.
266. Hokanson, J.E. and M.A. Austin, *Plasma triglyceride level is a risk factor for cardiovascular disease independent of high-density lipoprotein cholesterol level: a meta-analysis of population-based prospective studies.* J Cardiovasc Risk, 1996. **3**(2): p. 213-9.
267. UK, N., *'Chronic kidney disease - Diagnosis - NHS Choices', [online] Available from: https://www.nhs.uk/conditions/kidney-disease/diagnosis/ (Accessed 02 July 2018). .* 2018.
268. Chaussain, J.L., et al., *Serum branched-chain amino acids in the diagnosis of hyperinsulinism in infancy.* J Pediatr, 1980. **97**(6): p. 923-6.
269. Adeva, M.M., et al., *Insulin resistance and the metabolism of branched-chain amino acids in humans.* Amino Acids, 2012. **43**(1): p. 171-81.
270. Wang, Q., et al., *Genetic Support for a Causal Role of Insulin Resistance on Circulating Branched-Chain Amino Acids and Inflammation.* Diabetes Care, 2017. **40**(12): p. 1779-1786.
271. Hobbs, H.H., M.S. Brown, and J.L. Goldstein, *Molecular genetics of the LDL receptor gene in familial hypercholesterolemia.* Hum Mutat, 1992. **1**(6): p. 445-66.
272. Shichiri, M., A. Tanaka, and Y. Hirata, *Intravenous gene therapy for familial hypercholesterolemia using ligand-facilitated transfer of a liposome:LDL receptor gene complex.* Gene Ther, 2003. **10**(9): p. 827-31.
273. Soria, L.F., et al., *Association between a specific apolipoprotein B mutation and familial defective apolipoprotein B-100.* Proc Natl Acad Sci U S A, 1989. **86**(2): p. 587-91.
274. Gebhard, C., et al., *Apolipoprotein B antisense inhibition--update on mipomersen.* Curr Pharm Des, 2013. **19**(17): p. 3132-42.
275. Abifadel, M., et al., *Mutations in PCSK9 cause autosomal dominant hypercholesterolemia.* Nat Genet, 2003. **34**(2): p. 154-6.
276. Duff, C.J. and N.M. Hooper, *PCSK9: an emerging target for treatment of hypercholesterolemia.* Expert Opin Ther Targets, 2011. **15**(2): p. 157-68.
277. Duell, P.B., et al., *Long-term mipomersen treatment is associated with a reduction in cardiovascular events in patients with familial hypercholesterolemia.* J Clin Lipidol, 2016. **10**(4): p. 1011-1021.

278. Karatasakis, A., et al., *Effect of PCSK9 Inhibitors on Clinical Outcomes in Patients With Hypercholesterolemia: A Meta-Analysis of 35 Randomized Controlled Trials.* J Am Heart Assoc, 2017. **6**(12).
279. Asselbergs, F.W., et al., *Large-scale gene-centric meta-analysis across 32 studies identifies multiple lipid loci.* Am J Hum Genet, 2012. **91**(5): p. 823-38.
280. Willer, C.J., et al., *Discovery and refinement of loci associated with lipid levels.* Nat Genet, 2013. **45**(11): p. 1274-1283.
281. Albrechtsen, A., et al., *Exome sequencing-driven discovery of coding polymorphisms associated with common metabolic phenotypes.* Diabetologia, 2013. **56**(2): p. 298-310.
282. Peloso, G.M., et al., *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, 2014. **94**(2): p. 223-32.
283. Surakka, I., et al., *The impact of low-frequency and rare variants on lipid levels.* Nat Genet, 2015. **47**(6): p. 589-97.
284. Tang, C.S., et al., *Exome-wide association analysis reveals novel coding sequence variants associated with lipid traits in Chinese.* Nat Commun, 2015. **6**: p. 10206.
285. Natarajan, P., et al., *Deep-coverage whole genome sequences and blood lipids among 16,324 individuals.* bioRxiv, 2017.
286. White, J., et al., *Plasma urate concentration and risk of coronary heart disease: a Mendelian randomisation analysis.* Lancet Diabetes Endocrinol, 2016. **4**(4): p. 327-36.
287. Keenan, T., et al., *Causal Assessment of Serum Urate Levels in Cardiometabolic Diseases Through a Mendelian Randomization Study.* J Am Coll Cardiol, 2016. **67**(4): p. 407-416.
288. Davis, J.P., et al., *Common, low-frequency, and rare genetic variants associated with lipoprotein subclasses and triglyceride measures in Finnish men from the METSIM study.* PLoS Genet, 2017. **13**(10): p. e1007079.
289. Teslovich, T.M., et al., *Identification of seven novel loci associated with amino acid levels using single-variant and gene-based tests in 8545 Finnish men from the METSIM study.* Hum Mol Genet, 2018. **27**(9): p. 1664-1674.
290. Dewey, F.E., et al., *Inactivating Variants in ANGPTL4 and Risk of Coronary Artery Disease.* N Engl J Med, 2016. **374**(12): p. 1123-33.
291. Dewey, F.E., et al., *Distribution and clinical impact of functional variants in 50,726 whole-exome sequences from the DiscovEHR study.* Science, 2016. **354**(6319).
292. Moore, C., et al., *The INTERVAL trial to determine whether intervals between blood donations can be safely and acceptably decreased to optimise blood supply: study protocol for a randomised controlled trial.* Trials, 2014. **15**: p. 363.
293. Astle, W.J., et al., *The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease.* Cell, 2016. **167**(5): p. 1415-1429 e19.
294. Singh, T., et al., *Rare loss-of-function variants in SETD1A are associated with schizophrenia and developmental disorders.* Nat Neurosci, 2016. **19**(4): p. 571-7.
295. Li, H. and R. Durbin, *Fast and accurate short read alignment with Burrows-Wheeler transform.* Bioinformatics, 2009. **25**(14): p. 1754-60.
296. McKenna, A., et al., *The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data.* Genome Res, 2010. **20**(9): p. 1297-303.
297. Jun, G., et al., *Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data.* Am J Hum Genet, 2012. **91**(5): p. 839-48.
298. Soininen, P., et al., *High-throughput serum NMR metabonomics for cost-effective holistic studies on systemic metabolism.* Analyst, 2009. **134**(9): p. 1781-5.
299. Pinero, J., et al., *DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants.* Nucleic Acids Res, 2017. **45**(D1): p. D833-D839.
300. Pinero, J., et al., *DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes.* Database (Oxford), 2015. **2015**: p. bav028.

301. Kanehisa, M., et al., *KEGG: new perspectives on genomes, pathways, diseases and drugs.* Nucleic Acids Res, 2017. **45**(D1): p. D353-D361.
302. Kanehisa, M. and S. Goto, *KEGG: kyoto encyclopedia of genes and genomes.* Nucleic Acids Res, 2000. **28**(1): p. 27-30.
303. Kanehisa, M., et al., *KEGG as a reference resource for gene and protein annotation.* Nucleic Acids Res, 2016. **44**(D1): p. D457-62.
304. Fabregat, A., et al., *The Reactome Pathway Knowledgebase.* Nucleic Acids Res, 2018. **46**(D1): p. D649-D655.
305. Milacic, M., et al., *Annotating cancer variants and anti-cancer therapeutics in reactome.* Cancers (Basel), 2012. **4**(4): p. 1180-211.
306. Zaykin, D.V., *Optimally weighted Z-test is a powerful method for combining probabilities in meta-analysis.* J Evol Biol, 2011. **24**(8): p. 1836-41.
307. Dewey, M., *metap: meta-analysis of significance values*. 2018.
308. Rhee, E.P., et al., *An exome array study of the plasma metabolome.* Nat Commun, 2016. **7**: p. 12360.
309. Aschard, H., et al., *Adjusting for heritable covariates can bias effect estimates in genome-wide association studies.* Am J Hum Genet, 2015. **96**(2): p. 329-39.
310. Aschard, H., et al., *Covariate selection for association screening in multiphenotype genetic studies.* Nat Genet, 2017. **49**(12): p. 1789-1795.
311. Drenos, F., et al., *Metabolic Characterization of a Rare Genetic Variation Within APOC3 and Its Lipoprotein Lipase-Independent Effects.* Circ Cardiovasc Genet, 2016. **9**(3): p. 231-9.
312. Grevengoed, T.J., et al., *Acyl-CoA synthetase 1 deficiency alters cardiolipin species and impairs mitochondrial function.* J Lipid Res, 2015. **56**(8): p. 1572-82.
313. Yan, S., et al., *Long-chain acyl-CoA synthetase in fatty acid metabolism involved in liver and other diseases: an update.* World J Gastroenterol, 2015. **21**(12): p. 3492-8.
314. Zirath, H., et al., *MYC inhibition induces metabolic changes leading to accumulation of lipid droplets in tumor cells.* Proc Natl Acad Sci U S A, 2013. **110**(25): p. 10258-63.
315. Gopinathrao, G., *Image for "Regulation of pyruvate dehydrogenase (PDH) complex". Reactome, release 65, doi:10.3180/REACT_12528.1.* 2007.
316. Li, L.O., et al., *Liver-specific loss of long chain acyl-CoA synthetase-1 decreases triacylglycerol synthesis and beta-oxidation and alters phospholipid fatty acid composition.* J Biol Chem, 2009. **284**(41): p. 27816-26.
317. Scott, R.A., et al., *An Expanded Genome-Wide Association Study of Type 2 Diabetes in Europeans.* Diabetes, 2017. **66**(11): p. 2888-2902.
318. Yang, L., et al., *High expression of long chain acyl-coenzyme A synthetase 1 in peripheral blood may be a molecular marker for assessing the risk of acute myocardial infarction.* Exp Ther Med, 2017. **14**(5): p. 4065-4072.
319. Brodeur, G.M., et al., *Amplification of N-myc in untreated human neuroblastomas correlates with advanced disease stage.* Science, 1984. **224**(4653): p. 1121-4.
320. Emanuel, B.S., et al., *N-myc amplification in multiple homogeneously staining regions in two human neuroblastomas.* Proc Natl Acad Sci U S A, 1985. **82**(11): p. 3736-40.
321. Sato, T., et al., *Molecular cloning and characterization of a novel human beta 1,4-N-acetylgalactosaminyltransferase, beta 4GalNAc-T3, responsible for the synthesis of N,N'-diacetyllactosediamine, galNAc beta 1-4GlcNAc.* J Biol Chem, 2003. **278**(48): p. 47534-44.
322. Morgan, H., et al., *EuroPhenome: a repository for high-throughput mouse phenotyping data.* Nucleic Acids Res, 2010. **38**(Database issue): p. D577-85.
323. Wood, A.R., et al., *Defining the role of common variation in the genomic and biological architecture of adult human height.* Nat Genet, 2014. **46**(11): p. 1173-86.
324. Jin, J., et al., *Systematic analysis and nomenclature of mammalian F-box proteins.* Genes Dev, 2004. **18**(21): p. 2573-80.

325. Zhang, S., et al., *The pivotal role of pyruvate dehydrogenase kinases in metabolic flexibility.* Nutr Metab (Lond), 2014. **11**(1): p. 10.
326. Holmes, M.V., et al., *Lipids, Lipoproteins, and Metabolites and Risk of Myocardial Infarction and Stroke.* J Am Coll Cardiol, 2018. **71**(6): p. 620-632.
327. Dron, J.S., et al., *Polygenic determinants in extremes of high-density lipoprotein cholesterol.* J Lipid Res, 2017. **58**(11): p. 2162-2170.
328. Lange, L.A., et al., *Whole-exome sequencing identifies rare and low-frequency coding variants associated with LDL cholesterol.* Am J Hum Genet, 2014. **94**(2): p. 233-45.
329. Diabetes.co.uk. *Normal and Diabetic Blood Sugar Level Ranges*. 2018 [Accessed: 11-08-2018]; Available from: https://www.diabetes.co.uk/diabetes_care/blood-sugar-level-ranges.html.
330. Drzewoski, J. and L. Czupryniak, *Concordance between fasting and 2-h post-glucose challenge criteria for the diagnosis of diabetes mellitus and glucose intolerance in high risk individuals.* Diabet Med, 2001. **18**(1): p. 29-31.
331. Faerch, K., et al., *Trajectories of cardiometabolic risk factors before diagnosis of three subtypes of type 2 diabetes: a post-hoc analysis of the longitudinal Whitehall II cohort study.* Lancet Diabetes Endocrinol, 2013. **1**(1): p. 43-51.
332. Decode Study Group, t.E.D.E.G., *Glucose tolerance and cardiovascular mortality: comparison of fasting and 2-hour diagnostic criteria.* Arch Intern Med, 2001. **161**(3): p. 397-405.
333. American Diabetes, A., *Diagnosis and classification of diabetes mellitus.* Diabetes Care, 2014. **37 Suppl 1**: p. S81-90.
334. Paschou, S.A., et al., *Type 2 Diabetes and Osteoporosis: A Guide to Optimal Management.* J Clin Endocrinol Metab, 2017. **102**(10): p. 3621-3634.
335. Phung, O.J., et al., *Early combination therapy for the treatment of type 2 diabetes mellitus: systematic review and meta-analysis.* Diabetes Obes Metab, 2014. **16**(5): p. 410-7.
336. Saulsberry, W.J., et al., *Comparative efficacy and safety of antidiabetic drug regimens added to stable and inadequate metformin and thiazolidinedione therapy in type 2 diabetes.* Int J Clin Pract, 2015. **69**(11): p. 1221-35.
337. Koenig, R.J., et al., *Correlation of glucose regulation and hemoglobin AIc in diabetes mellitus.* N Engl J Med, 1976. **295**(8): p. 417-20.
338. Sidorenkov, G., et al., *A longitudinal study examining adherence to guidelines in diabetes care according to different definitions of adequacy and timeliness.* PLoS One, 2011. **6**(9): p. e24278.
339. Ghazanfari, Z., et al., *A Comparison of HbA1c and Fasting Blood Sugar Tests in General Population.* Int J Prev Med, 2010. **1**(3): p. 187-94.
340. Roberts, W.L., et al., *Effects of hemoglobin C and S traits on glycohemoglobin measurements by eleven methods.* Clin Chem, 2005. **51**(4): p. 776-8.
341. Son, J.I., et al., *Hemoglobin a1c may be an inadequate diagnostic tool for diabetes mellitus in anemic subjects.* Diabetes Metab J, 2013. **37**(5): p. 343-8.
342. Herman, W.H., et al., *Differences in A1C by race and ethnicity among patients with impaired glucose tolerance in the Diabetes Prevention Program.* Diabetes Care, 2007. **30**(10): p. 2453-7.
343. Cohen, R.M., S. Haggerty, and W.H. Herman, *HbA1c for the diagnosis of diabetes and prediabetes: is it time for a mid-course correction?* J Clin Endocrinol Metab, 2010. **95**(12): p. 5203-6.
344. Simonis-Bik, A.M., et al., *The heritability of HbA1c and fasting blood glucose in different measurement settings.* Twin Res Hum Genet, 2008. **11**(6): p. 597-602.
345. Snieder, H., et al., *HbA(1c) levels are genetically determined even in type 1 diabetes: evidence from healthy and diabetic twins.* Diabetes, 2001. **50**(12): p. 2858-63.
346. Chen, P., et al., *Multiple nonglycemic genomic loci are newly associated with blood level of glycated hemoglobin in East Asians.* Diabetes, 2014. **63**(7): p. 2551-62.

347. Franklin, C.S., et al., *The TCF7L2 diabetes risk variant is associated with HbA(1)(C) levels: a genome-wide association meta-analysis.* Ann Hum Genet, 2010. **74**(6): p. 471-8.
348. Pare, G., et al., *Novel association of HK1 with glycated hemoglobin in a non-diabetic population: a genome-wide evaluation of 14,618 participants in the Women's Genome Health Study.* PLoS Genet, 2008. **4**(12): p. e1000312.
349. Ryu, J. and C. Lee, *Association of glycosylated hemoglobin with the gene encoding CDKAL1 in the Korean Association Resource (KARE) study.* Hum Mutat, 2012. **33**(4): p. 655-9.
350. Soranzo, N., et al., *Common variants at 10 genomic loci influence hemoglobin A(1)(C) levels via glycemic and nonglycemic pathways.* Diabetes, 2010. **59**(12): p. 3229-39.
351. Danese, E., et al., *Advantages and pitfalls of fructosamine and glycated albumin in the diagnosis and treatment of diabetes.* J Diabetes Sci Technol, 2015. **9**(2): p. 169-76.
352. Armbruster, D.A., *Fructosamine: structure, analysis, and clinical usefulness.* Clin Chem, 1987. **33**(12): p. 2153-63.
353. Sacks, D.B., *A1C versus glucose testing: a comparison.* Diabetes Care, 2011. **34**(2): p. 518-23.
354. Little, R.R., et al., *Status of hemoglobin A1c measurement and goals for improvement: from chaos to order for improving diabetes care.* Clin Chem, 2011. **57**(2): p. 205-14.
355. Juraschek, S.P., et al., *Alternative markers of hyperglycemia and risk of diabetes.* Diabetes Care, 2012. **35**(11): p. 2265-70.
356. Selvin, E., et al., *Fructosamine and glycated albumin for risk stratification and prediction of incident diabetes and microvascular complications: a prospective cohort analysis of the Atherosclerosis Risk in Communities (ARIC) study.* Lancet Diabetes Endocrinol, 2014. **2**(4): p. 279-88.
357. Hom, F.G., B. Ettinger, and M.J. Lin, *Comparison of serum fructosamine vs glycohemoglobin as measures of glycemic control in a large diabetic population.* Acta Diabetol, 1998. **35**(1): p. 48-51.
358. Smart, L.M., et al., *Comparison of fructosamine with glycosylated hemoglobin and plasma proteins as measures of glycemic control.* Diabetes Care, 1988. **11**(5): p. 433-6.
359. Zafon, C., et al., *Variables involved in the discordance between HbA1c and fructosamine: the glycation gap revisited.* PLoS One, 2013. **8**(6): p. e66696.
360. Cohen, R.M., et al., *Discordance between HbA1c and fructosamine: evidence for a glycosylation gap and its relation to diabetic nephropathy.* Diabetes Care, 2003. **26**(1): p. 163-7.
361. Rodriguez-Segade, S., et al., *Influence of the glycation gap on the diagnosis of type 2 diabetes.* Acta Diabetol, 2015. **52**(3): p. 453-9.
362. Cohen, R.M., et al., *Evidence for independent heritability of the glycation gap (glycosylation gap) fraction of HbA1c in nondiabetic twins.* Diabetes Care, 2006. **29**(8): p. 1739-43.
363. Loomis, S.J., et al., *Genome-Wide Association Study of Serum Fructosamine and Glycated Albumin in Adults Without Diagnosed Diabetes: Results From the Atherosclerosis Risk in Communities Study.* Diabetes, 2018. **67**(8): p. 1684-1696.
364. O'Connell, J., et al., *Haplotype estimation for biobank-scale data sets.* Nat Genet, 2016. **48**(7): p. 817-20.
365. Machiela, M.J. and S.J. Chanock, *LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants.* Bioinformatics, 2015. **31**(21): p. 3555-7.
366. Manning, A.K., et al., *A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance.* Nat Genet, 2012. **44**(6): p. 659-69.
367. Scott, R.A., et al., *Large-scale association analyses identify new loci influencing glycemic traits and provide insight into the underlying biological pathways.* Nat Genet, 2012. **44**(9): p. 991-1005.

368. Mahajan, A., et al., *Identification and functional characterization of G6PC2 coding variants influencing glycemic traits define an effector transcript at the G6PC2-ABCB11 locus.* PLoS Genet, 2015. **11**(1): p. e1004876.
369. Strawbridge, R.J., et al., *Genome-wide association identifies nine common variants associated with fasting proinsulin levels and provides new insights into the pathophysiology of type 2 diabetes.* Diabetes, 2011. **60**(10): p. 2624-34.
370. Saxena, R., et al., *Genetic variation in GIPR influences the glucose and insulin responses to an oral glucose challenge.* Nat Genet, 2010. **42**(2): p. 142-8.
371. Kim, Y.J., et al., *Large-scale genome-wide association studies in East Asians identify new genetic loci influencing metabolic traits.* Nat Genet, 2011. **43**(10): p. 990-5.
372. Hwang, J.Y., et al., *Genome-wide association meta-analysis identifies novel variants associated with fasting plasma glucose in East Asians.* Diabetes, 2015. **64**(1): p. 291-8.
373. Liu, C.T., et al., *Trans-ethnic Meta-analysis and Functional Annotation Illuminates the Genetic Architecture of Fasting Glucose and Insulin.* Am J Hum Genet, 2016. **99**(1): p. 56-75.
374. Huyghe, J.R., et al., *Exome array analysis identifies new loci and low-frequency variants influencing insulin processing and secretion.* Nat Genet, 2013. **45**(2): p. 197-201.
375. Wessel, J., et al., *Low-frequency and rare exome chip variants associate with fasting glucose and type 2 diabetes susceptibility.* Nat Commun, 2015. **6**: p. 5897.
376. Go, M.J., et al., *New susceptibility loci in MYL2, C12orf51 and OAS1 associated with 1-h plasma glucose as predisposing risk factors for type 2 diabetes in the Korean population.* J Hum Genet, 2013. **58**(6): p. 362-5.
377. Horikoshi, M., et al., *Discovery and Fine-Mapping of Glycaemic and Obesity-Related Trait Loci Using High-Density Imputation.* PLoS Genet, 2015. **11**(7): p. e1005230.
378. Chen, G., et al., *Genome-wide association study identifies novel loci association with fasting insulin and insulin resistance in African Americans.* Hum Mol Genet, 2012. **21**(20): p. 4530-6.
379. Henny, J., et al., *Detetermination of reference values for a colorimetric fructosamine assay*. Vol. 38. 1992. 153-160.
380. Sherwani, S.I., et al., *Significance of HbA1c Test in Diagnosis and Prognosis of Diabetic Patients.* Biomark Insights, 2016. **11**: p. 95-104.
381. Franceschini, N., et al., *Discovery and fine mapping of serum protein loci through transethnic meta-analysis.* Am J Hum Genet, 2012. **91**(4): p. 744-53.
382. Kanai, M., et al., *Genetic analysis of quantitative traits in the Japanese population links cell types to complex human diseases.* Nat Genet, 2018. **50**(3): p. 390-400.
383. Martin, C.C., et al., *Cloning and characterization of the human and rat islet-specific glucose-6-phosphatase catalytic subunit-related protein (IGRP) genes.* J Biol Chem, 2001. **276**(27): p. 25197-207.
384. Wang, Y., et al., *Deletion of the gene encoding the islet-specific glucose-6-phosphatase catalytic subunit-related protein autoantigen results in a mild metabolic phenotype.* Diabetologia, 2007. **50**(4): p. 774-8.
385. Pound, L.D., et al., *G6PC2: a negative regulator of basal glucose-stimulated insulin secretion.* Diabetes, 2013. **62**(5): p. 1547-56.
386. Honore, B. and H. Vorum, *The CREC family, a novel family of multiple EF-hand, low-affinity Ca(2+)-binding proteins localised to the secretory pathway of mammalian cells.* FEBS Lett, 2000. **466**(1): p. 11-8.
387. Consortium, G.T., *The Genotype-Tissue Expression (GTEx) project.* Nat Genet, 2013. **45**(6): p. 580-5.
388. Chaudhury, C., et al., *The major histocompatibility complex-related Fc receptor for IgG (FcRn) binds albumin and prolongs its lifespan.* J Exp Med, 2003. **197**(3): p. 315-22.
389. Pyzik, M., et al., *Hepatic FcRn regulates albumin homeostasis and susceptibility to liver injury.* Proc Natl Acad Sci U S A, 2017. **114**(14): p. E2862-E2871.

390. Turley, P., et al., *Multi-trait analysis of genome-wide association summary statistics using MTAG.* Nat Genet, 2018. **50**(2): p. 229-237.
391. Dimas, A.S., et al., *Impact of type 2 diabetes susceptibility variants on quantitative glycemic traits reveals mechanistic heterogeneity.* Diabetes, 2014. **63**(6): p. 2158-71.
392. Gurdasani, D., et al., *The African Genome Variation Project shapes medical genetics in Africa.* Nature, 2015. **517**(7534): p. 327-32.
393. UK Biobank, *Nightingale Health and UK Biobank announces major initiative to analyse half a million blood samples to facilitate global medical research*. 2018, [Press Release] Retrieved from: http://www.ukbiobank.ac.uk/2018/06/nightingale-health-and-uk-biobank-announces-major-initiative-to-analyse-half-a-million-blood-samples-to-facilitate-global-medical-research/.
394. UK Biobank, *Whole genome sequencing will 'transform the research landscape for a wide range of diseases'*. 2018, [Prese Release] Retrieved from:: http://www.ukbiobank.ac.uk/2018/04/whole-genome-sequencing-will-transform-the-research-landscape-for-a-wide-range-of-diseases/.
395. Cross-Disorder Group of the Psychiatric Genomics, C., et al., *Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs.* Nat Genet, 2013. **45**(9): p. 984-94.
396. Andreassen, O.A., et al., *Genetic pleiotropy between multiple sclerosis and schizophrenia but not bipolar disorder: differential involvement of immune-related gene loci.* Mol Psychiatry, 2015. **20**(2): p. 207-14.
397. Moutsianas, L., et al., *The power of gene-based rare variant methods to detect disease-associated variation and test hypotheses about complex disease.* PLoS Genet, 2015. **11**(4): p. e1005165.
398. Fowler, D.M. and S. Fields, *Deep mutational scanning: a new style of protein science.* Nat Methods, 2014. **11**(8): p. 801-7.
399. Chen, S., et al., *An interactome perturbation framework prioritizes damaging missense mutations for developmental disorders.* Nat Genet, 2018. **50**(7): p. 1032-1040.
400. Hendricks, A.E., et al., *Rare Variant Analysis of Human and Rodent Obesity Genes in Individuals with Severe Childhood Obesity.* Sci Rep, 2017. **7**(1): p. 4394.
401. Singh, T., et al., *The contribution of rare variants to risk of schizophrenia in individuals with and without intellectual disability.* Nat Genet, 2017. **49**(8): p. 1167-1173.
402. Short, P.J., et al., *De novo mutations in regulatory elements in neurodevelopmental disorders.* Nature, 2018. **555**(7698): p. 611-616.
403. Siggens, L. and K. Ekwall, *Epigenetics, chromatin and genome organization: recent advances from the ENCODE project.* J Intern Med, 2014. **276**(3): p. 201-14.
404. Roadmap Epigenomics, C., et al., *Integrative analysis of 111 reference human epigenomes.* Nature, 2015. **518**(7539): p. 317-30.
405. Martens, J.H. and H.G. Stunnenberg, *BLUEPRINT: mapping human blood cell epigenomes.* Haematologica, 2013. **98**(10): p. 1487-9.
406. Li, X., et al., *The impact of rare variation on gene expression across tissues.* Nature, 2017. **550**(7675): p. 239-243.
407. Inoue, F. and N. Ahituv, *Decoding enhancers using massively parallel reporter assays.* Genomics, 2015. **106**(3): p. 159-164.

Appendix

Genetic architecture of human thinness compared to severe obesity

Fernando Riveros-McKay[1*], Vanisha Mistry[2*], Rebecca Bounds[2], Audrey Hendricks[1,3], Julia M. Keogh[2], Hannah Thomas[2], Elana Henning[2], Laura J. Corbin[4,5], Understanding Society Scientific Group, Stephen O'Rahilly[2], Eleftheria Zeggini[1], Eleanor Wheeler[1], Inês Barroso[1,2], I. Sadaf Farooqi[2].

Affiliations

[1]Wellcome Sanger Institute, Cambridge, UK; [2]University of Cambridge Metabolic Research Laboratories and NIHR Cambridge Biomedical Research Centre, Wellcome Trust-MRC Institute of Metabolic Science, Addenbrooke's Hospital, Cambridge, UK; [3]Department of Mathematical and Statistical Sciences, University of Colorado-Denver, Denver, CO 80204, USA; [4]MRC Integrative Epidemiology Unit at University of Bristol, Bristol, BS8 2BN, UK; [5]Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, BS8 2BN, UK. Correspondence should be addressed to: I. Sadaf Farooqi (isf20@cam.ac.uk) and Inês Barroso (ib1@sanger.ac.uk). *These authors contributed equally.

Abstract

The variation in weight within a shared environment is largely attributable to genetic factors. Whilst many genes/loci confer susceptibility to obesity, little is known about the genetic architecture of healthy thinness. Here, we characterise the heritability of thinness which we found was comparable to that of severe obesity (h^2=28.07 vs 32.33% respectively), although with incomplete genetic overlap (r=-0.49, 95% CI [-0.17, -0.82], p=0.003). In a genome-wide association analysis of thinness (n=1,471) vs severe obesity (n=1,456), we identified 10 loci previously associated with obesity, and demonstrate enrichment for established BMI-associated loci ($p_{binomial}$=3.05x10^{-5}). Simulation analyses showed that different association results between the extremes were likely in agreement with additive effects across the BMI distribution, suggesting different effects on thinness and obesity could be due to their different degrees of extremeness. In further analyses, we detected a novel obesity and BMI-associated locus at *PKHD1* (rs2784243, obese vs. thin p=5.99x10^{-6}, obese vs. controls p=2.13x10^{-6} p_{BMI}=2.3x10^{-13}), associations at loci recently discovered with much larger sample sizes (e.g. *FAM150B* and *PRDM6-CEP120*), and novel variants driving associations at previously established signals (e.g. rs205262 at the *SNRPC*/*C6orf106* locus and rs112446794 at the *PRDM6-CEP120* locus). Our ability to replicate loci found with much larger sample sizes demonstrates the value of clinical extremes and suggest that characterisation of the genetics of thinness may provide a more nuanced understanding of the genetic architecture of body weight regulation and may inform the identification of potential anti-obesity targets.

38 **Author Summary**

39 Obesity-associated disorders are amongst the leading causes of morbidity and mortality
40 worldwide. Most genome-wide association studies (GWAS) have focused on body mass index (BMI=
41 weight in Kg divided by height squared (m^2)) and obesity, but to date no genetic association study
42 testing thin and healthy individuals has been performed. In this study, we recruited a first of its kind
43 cohort of 1,471 clinically ascertained thin and healthy individuals and contrasted the genetic
44 architecture of the trait with that of severe early onset obesity. We show that thinness, like obesity,
45 is a heritable trait with a polygenic component. In a GWAS of persistent healthy thinness vs. severe
46 obesity with a total sample size of 2,927, we are able to find evidence of association in loci that
47 have only been recently discovered using large cohorts with >40,000 individuals. We also find a
48 novel BMI-associated locus at *PKHD1* in UK Biobank highlighted by our association study. This work
49 illustrates the value and increased power brought upon by using clinically ascertained extremes to
50 study complex traits and provides a valuable resource on which to study resistance to obesity in an
51 increasingly obesogenic environment.

Introduction

The rising prevalence of obesity is driven by changes in the environment including the consumption of high calorie foods and reduced levels of physical activity [1]. However, within a given environment, there is considerable variation in body weight; some people are particularly susceptible to severe obesity, whilst others remain thin [2,3]. Family, twin and adoption studies have consistently demonstrated that 40-70% of the variation in body weight can be attributed to heritable factors [4]. As a result, many studies have focused on the genetic basis of body mass index (BMI) and/or obesity. To date >250 common and low-frequency obesity-susceptibility loci have been identified [5-10]. Additionally, studies of people at one extreme of the distribution (severe obesity) have led to the identification of rare, penetrant genetic variants that affect key molecular and neural pathways involved in human energy homeostasis [11-14]. These findings have provided a rationale for targeting these pathways for therapeutic benefit. In contrast, little is known about the specific genetic characteristics of persistently thin individuals (thinness defined using WHO criteria BMI$\leq$18kg/m^2). Understanding the mechanisms underlying thinness/resistance to obesity may highlight novel anti-obesity targets for future drug development.

A small number of previous studies have found that thinness appears to be a trait that is at least as stable and heritable as obesity [15-18]. A large study of 7,078 UK children and adolescents, found that the strongest predictor of child/adolescent thinness was parental weight status. The prevalence of thinness was highest (16.2%) when both parents were thin and progressively lower when both parents were normal weight, overweight or obese [19].

One approach to studying thinness is to study individuals from a population-based cohort for a quantitative or continuous trait. For example, it is possible to generate a "case-control" study by taking the extremes of the population distribution for a continuous trait such as BMI, an approach used effectively by Berndt *et al.* 2013 [20] who analysed the top and bottom 5% in cohorts participating in the GIANT Consortium. However, by their very definition, such population-based cohorts often contain a limited number of people at the "extremes" (i.e. severe obesity and thinness) [20]. To date, other GWAS approaches that included thin individuals have either used them exclusively as controls to contrast with extreme obesity [21], or have not ascertained for healthy thinness [22]. Here, we use a different study design, and one that has been used to increase power to detect genetic association, in particular for disorders where there is a large environmental component (e.g. asthma, type 2 diabetes and obesity), enriching our case series with affected individuals that may be more genetically loaded. This selection is usually done by selecting individuals who may have a more extreme form of disease, are younger (less time for environment to impact their disease) and perhaps have family members also affected with the same condition. To complement this approach to the selection of cases, controls are also selected to increase the chances that they do not have the disease or are unlikely to develop the disease later in life [21]. This is normally done by selecting contrasting controls, or "super-controls". However, the low prevalence of thinness in countries such as the UK and the fact that people who are well but constitutionally thin do not routinely come to medical attention, poses challenges to recruitment of a cohort of healthy thin individuals. We were able to take advantage of the UK National Health Service (NHS) research infrastructure to recruit from primary care (**Methods**) using body mass index

(BMI: weight in kg/height in metres^2) criteria and personal review of individual case files to identify a cohort of approximately 2000 UK European descent thin adults (STudy Into Lean and Thin Subjects, STILTS cohort; mean BMI = 17.6 kg/m^2) who are well, without medical conditions or eating disorders (**Methods**). 74% of the STILTS cohort have a family history of persistent thinness throughout life, suggesting we have enriched for genetically driven thinness.

Here, we present a new, and the largest-to-date, GWAS focused on persistent healthy thinness and contrast the genetic architecture of this trait with that of severe early onset obesity ascertained in the clinic. We explored whether the genetic loci influencing thinness are the same as those influencing obesity, i.e., are these two clinically ascertained traits reverse sides of the same "coin", or whether there are important genetic differences between them. We show that persistent thinness and severe early onset obesity are both heritable traits (h^2=28.07% and h^2=32.33%, respectively) that share a number of associated loci, and both are enriched for established BMI associated loci (binomial p=3.05x10^{-5} and 9.09x10^{-13}, respectively). Nonetheless, we also detected important differences, with some loci more strongly associated at the upper clinical end of the BMI distribution (e.g. *FTO*), some at the lower end (e.g. *CADM2*), whilst other loci are equivalently associated with both clinical ends of the BMI spectrum (e.g. *MC4R*). Simulation tests showed that these results did not significantly deviate from additive effects and most likely reflect the different degrees of extremeness present in our clinically ascertained cohorts, where severely obese individuals represent a more significant deviation from the mean than healthy thin individuals do (the same degree of thinness may not be compatible with healthy human life). These data support expansion of genetic studies of persistent thinness as an approach to gain further insights into the

biology underlying human energy homeostasis, and as an alternative approach to uncovering potential anti-obesity targets for drug development.

Results

Heritability of persistent thinness and severe early onset obesity

To investigate the heritability of healthy thinness and contrast it with that of severe early onset childhood obesity we obtained genotype data for 1,622 persistently thin healthy individuals (STILTS), 1,985 severe childhood onset obesity cases (SCOOP; European ancestry individuals from the GOOS cohort) and 10,433 population-based individuals (UKHLS) used as a common set of controls (**Methods, S1 Table**). All participants were genotyped on the Illumina Core Exome array, including 551,839 markers. After sample and variant quality control, we retained 1,471 thin individuals, 1,456 obese individuals, 6,460 control individuals in the BMI range 19-30 kg/m^2 (non-extremes). 477,288 directly genotyped variants were included in the analysis (**Methods**); 54% common variants (minor allele frequency (MAF) ≥1% amongst controls) and 46% rare variants (MAF<1% amongst controls), of which most were protein-coding (96.8%). We then imputed genotypes to a combined UK10K+1000G reference panel and, using LD score regression, we estimated that a subset of 1,197,969 HapMap3 markers accounted for 32.33% (95% CI 23.75%-40.91%) of the phenotypic variance on the liability scale in severe early onset obesity, and 28.07% (95% CI 13.80%-42.34%) in persistent thinness, suggesting both traits are similarly heritable (**Methods**). The heritability estimates reported here were used mainly to establish the fact that

thinness is a heritable trait; we expect our liability scale estimates to be mostly unbiased given the study design [23]. However, given the low prevalence of the traits presented here, these estimates may represent upper bounds.

Contribution of known BMI associated loci to thinness and severe early onset obesity

To investigate the role of established common variant European BMI associated loci, we studied the 97 loci from GIANT [24] in persistent thinness vs severe early onset obesity and performed three-way association analyses: obese vs. thin, obese vs controls, controls vs. thin (**Methods, S1 Table**). After quality control, 41,266,535 variants remained for association analyses in the three cohorts: SCOOP vs STILTS, SCOOP vs UKHLS and UKHLS vs STILTS. Of the 97 established BMI associated loci from GIANT [24], we found that 40 were nominally significant (p<0.05) in SCOOP vs UKHLS and 15 in UKHLS vs STILTS (**S2 Table**). Direction of effect was consistent for all of these loci, which was more than expected by chance (binomial p=9.09x10^{-13} and binomial p=3.05x10^{-5}, respectively). Overall, the proportion of phenotypic variance explained by the 97 established BMI associated loci was 10.67% in SCOOP vs UKHLS, and 4.33% in STILTS vs UKHLS (**Methods**). Evaluation of association results in thin (STILTS) and obese (SCOOP) individuals, compared to the same controls (UKHLS), suggested that the results are not a mirror image of each other (**Figs 1-2**), **however we found little evidence of non-additive effects at the loci explaining this discrepancy (see below).** We observed a striking difference in association results in the *FTO* locus where the lead intronic obesity risk variant, rs1558902, showed a moderate effect size and modest evidence of association in controls

compared to thin individuals from STILTS (p=0.00027, OR=1.17, 95% CI [1.08,1.28], EAF=0.39), despite having a large effect and being associated at genome-wide significance levels in SCOOP (p=1.25x10^{-17}, OR=1.43, 95% CI [1.32,1.55], EAF=0.41), and *GNAT2* also showed a larger effect and significance in the analysis of obese compared to control individuals (p=1.26x10^{-4}, OR=1.57, 95% CI [1.25, 1.97], EAF=0.03), than in the thin analysis (p=0.52, OR=1.10, 95% CI [0.82, 1.47], EAF=0.02, **Fig 1**, **S2 Table**). This discrepancy in association strength and effect size was also seen at the opposite end of the BMI spectrum in *CADM2* where the lead SNP, rs13078960, showed evidence of association in STILTS (p= 9.48x10^{-4}, OR=1.2, 95% CI [1.08, 1.33], EAF=0.20) but no association in SCOOP (p>0.05). In contrast to results at the *FTO* and *CADM2* loci, for *MC4R* the results are more comparable, with genome-wide significant association in obese individuals (rs6567160, p=7.91x10^{-9}, OR=1.31, 95% CI [1.19, 1.43], EAF=0.25) and highly significant association results in thin individuals (p=1.38x10^{-5}, OR=1.26, 95% CI [1.13, 1.39], EAF=0.23, **S2 Table**)**.** To formally test if these results were significantly different from those expected under a model where loci act additively across the BMI distribution, we simulated 10,000 different populations of 1 million individuals with genotypes for the 97 established BMI loci using allele frequencies in the European population, and then simulated a phenotype using the effect sizes in GIANT (**Methods**). These simulations detected fourteen loci with nominally significant deviation from an additive model, however none remained significant after correction for the number of tests (p=0.05/97*2 = ~0.0002, **S3 Table**), though *CADM2* was nominally significant in both SCOOP and STILTS analyses, with slightly lower OR detected in SCOOP compared to simulated data, and slightly higher OR detected in STILTS compared to simulated data (**S3 Table**). **Recent work in mouse knockouts has shown *CADM2* plays**

175 **an important role in systemic energy homeostasis [25] and variants near the gene have also been**
176 **recently linked to habitual physical activity in humans [26].** Since SCOOP participants are
177 significantly younger than UKHLS, we used summary statistics from a subset of the ALSPAC cohort
178 [27] which consists of 4,964 children aged 13-16 to test if the observed OR differences in SCOOP vs
179 UKHLS, compared to STILTS vs UKHLS, were due to age effects in SCOOP (**Methods**). For the 97
180 GIANT loci overall there were no significant differences in the ORs when comparing SCOOP to
181 UKHLS or SCOOP to ALSPAC (z-test, p>0.05) except for rs2245368 (*PMS2L11* locus, z-test
182 p=3.81x10^{5}, **S4 Table**). In combination, these results suggest that the observed differences in ORs
183 and p-values could have arisen because our severe obese cases are much more extreme (i.e.
184 deviate more from the mean) than the healthy thin individuals, and that our obese and thin sample
185 sizes gave us limited power to detect significant differences compared to the additive model.

186 **Fig 1. Odds ratio comparison for established BMI associated loci.** Odds ratios for SCOOP vs UKHLS
187 (x-axis) and UKHLS vs STILTS (y-axis) comparisons are shown for the 97 known BMI loci from GIANT
188 [24]. Colours of data points represent nominal significance in both analyses (red), only SCOOP vs.
189 UKHLS (green), only STILTS vs UKHLS (blue) or in neither analysis (purple). Error bars represent 95%
190 confidence intervals for the odds ratios for SCOOP vs UKHLS (x-axis) and for UKHLS vs STILTS (y-
191 axis). A subset of data points with larger separation from the red diagonal line (x=y) are labelled.

192

193 Next we investigated the association of a genetic risk score, generated from the 97 BMI associated
194 loci from GIANT [24] on BMI category (i.e. thin, normal, obese) using an ordinal logistic regression

(**Methods**). As expected, the standardised BMI genetic risk score was strongly associated with BMI category (weighted score p=8.59x10^{-133}). We found that the effect of a one standard deviation increase in the standardised BMI genetic risk score was significantly larger for obese vs. (thin & normal) than for (obese & normal) vs. thin (p=7.48x10^{-11}, **S1 Appendix**) with odds ratio and 95% confidence intervals of 1.94 (1.83, 2.07) and 1.50 (1.42, 1.59) respectively. However, using the simulations described above (**Methods**), we confirm that the larger OR for obese vs. (thin & normal) is not significantly different (p=0.41) than what we would expect given an additive genetic model, and the different degrees of extremeness in our thin and obese cases. Mean GRS in each BMI category was also not significantly different from that predicted via simulations (**S1 Fig**, **Methods**).

Genetic Correlation between persistent thinness, severe early onset childhood obesity and BMI

Given the observed differences in association results from thin and obese individuals, compared to the same set of control individuals, we next explored the genetic correlation of severe early onset obesity, persistent thinness and BMI using LD score regression (**Methods**). For this, we used summary statistics from the SCOOP vs UKHLS, STILTS vs UKHLS and BMI data from participants in UK Biobank (UKBB, **Methods**). As expected from the association results, the genetic correlation of severe early onset obesity and BMI was high (r=0.79, 95% CI [0.69, 0.89], p=1.14x10^{-52}). We also observed weaker negative correlation between persistent thinness and BMI (r=-0.69, 95% CI [-0.86, -0.51], p= 1.17x10^{-14}), and between persistent thinness and severe obesity (r=-0.49, 95% CI [-0.17, -0.82], p=0.003). As an inverse genetic correlation between BMI, obesity and anorexia nervosa (a

215 disorder that is characterised by thinness and complex behavioural manifestations) has recently
216 been reported [28], we also tested for genetic correlation with anorexia nervosa, and found that
217 neither severe early onset obesity, nor persistent thinness, were significantly correlated with
218 anorexia nervosa (r=-0.05, 95% CI [-0.15,0.05], p=0.33 and r=0.13, 95% CI [-0.02,0.28], p=0.09,
219 respectively; **Methods**).

220

221 **Association signals for persistent thinness and severe early onset obesity replicate established**
222 **BMI associated loci**

223 Given available genome-wide directly genotyped and imputed data we sought evidence for novel
224 signals associated with either end of the BMI distribution (persistent thinness or severe early onset
225 obesity; **Methods**) **but found no novel replicating loci (details below).** In all three **discovery**
226 analyses, in addition to loci mapping to established BMI and obesity loci, we identified *PIGZ* and
227 *C3orf38*, two **putative** novel loci in the thin vs control analysis, that reached conventional genome-
228 wide significance (GWS) ($p \leq 5x10^{-8}$) (**Tables S5-S7, Fig 2**). However, an additional 125 SNPs, in 118
229 distinct loci, reached the arbitrary threshold of $p \leq 10^{-5}$ in at least one analysis, for which we sought
230 replication (**Tables S5-S7**)**.**

231 **Fig 2. Miami plot of SCOOP vs. UKHLS and STILTS vs. UKHLS.** Miami plot produced in EasyStrata
232 [29], Red=SCOOP vs. UKHLS; Blue=STILTS vs. UKHLS. Red lines indicate genome-wide significance
233 threshold at $p=5x10^{-08}$. Orange lines indicate discovery significance threshold at $p=1x10^{-05}$. Black
234 labels highlight known BMI/obesity loci that were taken forward for replication and yellow peaks

indicate those that met genome-wide significance after replication. Grey labels highlight novel loci with $p<5x10^{-08}$ that did not replicate.

As our obese and thin cases (SCOOP and STILTS) lie at the very extreme tails of the BMI distribution, there are few comparable replication datasets. We therefore used the UKBB dataset and selected individuals at the top (BMI>=40, N= 7,526) and bottom end of the distribution (BMI≤19, N= 3,532) to more closely match the BMI criteria of our clinically ascertained thin and obese individuals. We used 20,720 samples from the rest of the UKBB cohort as a control set (**Methods**, **S2 Fig**). In cases where lead variants or proxies ($r^2>0.8$) were not currently available in the full UKBB genetic release we used results from the interim release using 2,799 individuals with BMI>=40, 1,212 with BMI<=19 and 8,193 controls (**Methods**). We noted a significant negative genetic correlation for our obese replication cohort with anorexia nervosa (r= -0.24, 95% CI [-0.37,-0.11], p=0.01) and a positive genetic correlation for our thin cohort (r=0.49, 95% CI [0.22-0.76] p=0.0003). We also observed significant genetic correlation between obesity in the discovery and replication cohorts (r=0.84, 95% CI [0.65-1] $p=5.05x10^{-17}$) and between thinness in the discovery and replication cohorts (r= 0.62, 95% CI [0.20-1] p=0.004).

To further increase power, we took advantage of publicly available summary statistics from the GIANT Extremes obesity meta-analysis [20], the EGG childhood obesity study [30], and our own previous study on non-overlapping SCOOP participants (SCOOP 2013) [31], as additional replication datasets. For SCOOP vs. STILTS we used the GIANT BMI tails meta-analysis results [20] (up to 7,962

255 cases/8,106 controls from the upper/lower 5th percentiles of the BMI trait distribution). For SCOOP
256 vs. UKHLS we used the GIANT obesity class III summary statistics [20] (up to 2,896 cases with BMI
257 ≥40kg/m^2 vs 47,468 controls with BMI <25 kg/m^2), the EGG childhood obesity study [30] (children
258 with BMI ≥95th percentile of BMI vs 8,318 children with BMI <50th percentile of BMI) and SCOOP
259 2013 [31]. Fixed effect meta-analyses yielded genome-wide significant signals at well-known BMI
260 associated loci in both the obese vs. thin, and obese vs. control analyses, and both the *PIGZ* and
261 *C3orf38* loci identified at the discovery stage failed to replicate when combined with additional data
262 **(Table 1, S7 Table).** However, the *SNRPC* locus described here (rs75398113), though not
263 independent from the previously described *SNRPC*/*C6orf106* locus (rs205262, r^2= 0.29) [24],
264 appears to be driving the previously reported association at this locus (rs205262 conditioned on
265 rs75398113, $p_{conditioned}$=0.7, **S8 Table**). Both SNPs are eQTLs for *C6or106* and *UHRF1BP1* in multiple
266 tissues including brain and colon tissues on GTEx however neither of these are obvious biological
267 candidates linked to energy homeostasis.

268

269 **Table 1 - GWAS results for SNPs meeting *p*<5x10^{-8} in all three analyses.** EA= Effect allele (BMI
270 increasing allele); NEA= Non-effect allele; OR = Odds ratio; 95% CI = 95% confidence interval for the
271 odds ratio; EAF = effect allele frequency. Positions mapped to hg19, Build 37. [a]rs12995480 used as
272 proxy in GIANT. [b]rs2384054 used as proxy in GIANT. [c]rs12641981 used as proxy in GIANT. [d]rs663129
273 used as proxy in GIANT, EGG and SCOOP 2013. [e]rs13007080 used as proxy in GIANT, EGG and
274 SCOOP 2013. [f]rs7138803 used as proxy in SCOOP 2013. [g]rs6722587 used as proxy in GIANT, EGG
275 and SCOOP 2013. [h]rs4132288 used as proxy in GIANT, EGG and SCOOP 2013. [i]rs1460940 used as

proxy in GIANT, EGG and SCOOP 2013. [j]rs1366333 used as proxy in GIANT, EGG and SCOOP 2013. [k]GIANT BMI tails [20]. [l]GIANT obesity class III [20].

Finally, we used the independent BMI dataset from UKBB (**Methods**) to investigate whether any of the loci meeting our arbitrary $p \leq 10^{-5}$ in discovery efforts, were independently associated with BMI as a continuous trait. This identified a novel BMI-associated locus near *PKHD1* (SCOOP vs. STILTS $p=5.99\times10^{-6}$, SCOOP vs. UKHLS $p=2.13\times10^{-6}$, BMI $p=2.3\times10^{-13}$, **S9 Table**). Furthermore, we note that when comparing the signals we took for replication (based on case control analyses) with association results with BMI as a continuous trait derived from an independent set of samples from UKBB, there are more directionally consistent and nominally significant associations with BMI than expected by chance suggesting that amongst these loci, there may be additional real associations (binomial $p=4.88\times10^{-4}$, and binomial $p=9.77\times10^{-3}$, respectively, Methods, S9 Table)."

Despite the smaller sample size, the obese vs thin comparison had increased power to detect some loci (**S3 Fig**), including a recently discovered variant near *FAM150B* [32] (rs62107261, MAF= ~5%), which did not meet our $p<10^{-5}$ threshold to be taken forward for replication in obese vs controls analysis ($p=2.36\times10^{-4}$).

Discussion

Here we present results from the largest to-date GWAS performed on healthy individuals with persistent thinness and provide the first insights into the genetic architecture of this trait. To our

296 knowledge, there are only two other studies using thin individuals with comparable mean BMIs
297 **[21,22]**. The study by Hinney *et al*. **[21]** (N=442), was only able to detect *FTO* at genome-wide
298 significance level with rs1121980 having a similar effect to that which we report (OR=1.66 vs OR=
299 1.69 in our data). In the Scannell Bryan *et al***. [22]** study, Bangladeshi individuals were reportedly
300 thin and malnourished, and a single suggestive association was found with an intronic variant in
301 *NRXN3* (rs12882679, $p=9.57x10^{-7}$) which is not significant in our study (p=0.77).

302 Using genome-wide genotype data we show that persistent healthy thinness, similar to severe
303 obesity (h^2=32.33%), is a heritable trait (h^2=28.07%). Persistent healthy thinness and severe
304 childhood obesity are negatively correlated (r=-0.49, 95% CI [-0.17, -0.82], *p*=0.003), and share a
305 number of genetic risk loci. Nonetheless, the genetic overlap between the two clinically ascertained
306 traits appears to be incomplete, as highlighted by some loci which were more strongly associated
307 at one end of the BMI distribution (e.g. *CADM2*), while others, appeared to exert effects across the
308 entire BMI spectrum (e.g. *MC4R* [9,33,34]). Further exploration by simulation demonstrated that
309 these differences are likely to be due to the different degrees of extremeness of the two clinical
310 cohorts (i.e. a similar degree of thinness to that of the obese cohort may not be compatible with
311 healthy human life) and not due to a deviation from additive effects of the tested loci on BMI, with
312 the possible exception of *CADM2* which deviated from expectation with nominal significance in
313 both the obese and the thin analysis (**S3 Table**). This is in contrast with earlier studies which
314 suggested larger effects at the higher end of the BMI distribution [35,36] but in agreement with
315 more recent observations contrasting the bottom 5% and top 5% of the BMI tails where associated
316 loci were also consistent with additive effects [20]. This is also in contrast with a previous study on

317 height, where a deviation from additivity was found, but only for short individuals in the bottom
318 1.5% of the distribution [37], which suggests that analysis focused just on the most extreme
319 individuals may be warranted.

320

321 Focusing on the 97 previously established BMI associated loci [24], we show that the percentage of
322 phenotypic variance explained by these loci is lower in persistently thin (4.33%) compared to obese
323 individuals (10.67%), and that the effect of an increase/decrease in the BMI genetic risk score was
324 much larger, on average, for obese individuals than for thin individuals (one standard deviation
325 increase in the standardised BMI genetic risk score of 1.94, 95% CI (1.83, 2.07) and 1.50, 95% CI
326 (1.42, 1.59), respectively) which is consistent with the difference in BMI units amongst categories.
327 And, although our analysis using age-matched controls from ALSPAC suggested that the observed
328 differences in ORs, comparing obese vs control individuals to controls vs thin individuals, was
329 unlikely to be due to age effects, we cannot completely exclude the possibility that different effects
330 of age and sex in our discovery cohorts (**S1 Table**), and gene-by-environment interactions, could be
331 influencing some of the results we observe. For example, gene-by-environment interactions and
332 age effects have been previously reported at the *FTO* locus [38-41] where a larger effect is detected
333 in younger adults. **It is worth noting though that non-additive effects have also been observed in**
334 **the *FTO* locus [42].**

335

In studying thin individuals there are often concerns regarding the prevalence of eating disorders, notably anorexia nervosa amongst participants. We sought to carefully exclude eating disorders at two phases of recruitment (by medical history and by questionnaire). Additionally, we demonstrate that in our cohort of healthy thin individuals, anorexia nervosa is unlikely to be a confounder as the two traits are genetically only weakly correlated (r=0.13, 95% CI [-0.02,0.28], p=0.09). This was not the case for the UKBB replication cohort where a positive genetic correlation was observed (r= 0.49 95% CI [0.22-0.76] p=0.0003). The positive genetic correlation with anorexia was still observed after removing individuals with medical conditions that could explain their low BMI (r=0.62, 95% CI [0.30,0.92], p=0.0001, **Methods)**. These results highlight the importance of the careful phenotyping performed in the recruitment phase and the utility of the STILTS cohort as a resource to study healthy and persistent thinness.

In the genome-wide association analyses amongst the signals we took forward for replication, in addition to detecting established BMI-associated loci, we find a novel BMI-association at *PKHD1* in the UKBB BMI dataset (rs10456655, β=0.10, p=2.3x10^{-13}, **S9 Table**), where a proxy for this variant (rs2579994, r^2=1 in 1000G Phase 3 CEU) has been previously nominally associated with waist and hip circumference (p=5.60x10^{-5} and p=4.40x10^{-4} respectively) [43]. In addition, we found associations at loci that have only recently been established using very large sample sizes. *FAM150B*, was only suggestively associated at discovery stage in Tachmazidou *et al.* (2017) [32] (n=47,476, p=2.57×10^{-5}) whereas it reached genome-wide significance when contrasting SCOOP vs STILTS (n=2,927, p=2.07x10^{-8}, **S5 Table**). Also, *PRDM6-CEP120* [5] was recently discovered in a Japanese study with a sample size of 173,430 and has not been previously reported in a European

357 population. In our study, a signal near the locus (rs112446794, r^2=0.36) showed suggestive evidence
358 of association in SCOOP vs UKHLS (p=2.08x10^{-6}, **S6 Table**) with a significantly smaller sample size.
359 Conditional analysis reveals the lead SNP in this study drives the association of the previously
360 established signal (**S8 Table**). *CEP120* codes for centrosomal protein 120. Variants near this locus
361 have been previously associated with height [44] and waist circumference in East Asians [45].
362 Missense variants in the gene itself have been associated with rare ciliopathies [46,47]. Lastly,
363 amongst the signals we took for replication, and after removing known and newly established loci,
364 we still observe an enrichment of directionally consistent and nominal associations in the analysis
365 of BMI as a continuous trait, suggesting that some of these results may warrant additional
366 investigation, in particular in similarly ascertained thin and obese cohorts. One such example is
367 rs4447506, near *PIK3C3,* which was not only nominally significant and consistent in the
368 independent UKBB BMI analysis (p=1.5x10^{-6}, **S9 Table**), but also in the Locke *et al*. (2015) [24] BMI
369 results (p= 0.01), and in the GIANT BMI tails analysis we used as replication (**S5 Table**). We also
370 note, that despite not reaching genome-wide significance in our discovery cohorts, we observe
371 directionally consistent suggestive associations at a number of loci previously associated with BMI
372 tails and with different obesity classes [20] (**S10 Table**)**.** Altogether, these results highlight some
373 power advantages of using clinically ascertained extremes of the phenotype distribution to detect
374 associations and suggest that healthy thinness falls at the lower end of the polygenic BMI spectrum.
375 It is worth noting though that these clinically ascertained extremes display evidence of incomplete
376 genetic correlation with BMI, in contrast to previously described obesity classes (S4 Fig), so it is
377 plausible that additional loci might be uncovered by focusing on clinical extremes.

As our results were based on clinically ascertained participants which met very specific criteria, it is worth noting these conclusions cannot be straightforwardly extrapolated to the general population. Experiments in animals have identified loci/genes associated with thinness/decreased body weight due to reduced food intake/increased energy expenditure/resistance to high fat diet-induced obesity [48,49], mechanisms that we hypothesise may contribute to human thinness. The STILTS cohort, being uncorrelated to anorexia nervosa, is an excellent resource in which to conduct such additional genetic exploration. Further genetic and phenotypic studies focused on persistently thin individuals may provide new insights into the mechanisms regulating human energy balance and may uncover potential anti-obesity drug targets.

Methods

ETHICS STATEMENT

The study was reviewed and approved by the South Cambridgeshire Research Ethics Committee (12/EE/0172). All participants provided written informed consent prior to inclusion.

COHORTS

SCOOP, STILTS and UKHLS cohorts were used for the heritability, genetic correlation, genetic risk score and association analyses with established BMI loci, as well as, used as a discovery cohort in the genome-wide association study (GWAS) and gene-based tests. UK Biobank samples were used for genetic correlation analysis and in the replication stages of the GWAS and gene-based tests. ALSPAC was used as an additional control dataset to UKHLS for comparison against SCOOP in the established BMI loci analysis.

STILTS

The aim was to recruit a new cohort of UK European people who are thin (defined as a body mass index $\leq$ 18kg/m^2) and well. After ethical committee approval (12/EE/0172), we worked with the NIHR Primary Care Research Network (PCRN) to collaborate with 601 GP practices in England. Each practice searched their electronic health records using our inclusion criteria (age 18-65 years, BMI$\leq$18 kg/m^2) and exclusion criteria (medical conditions that could potentially affect weight (chronic renal, liver, gastrointestinal problems, metabolic and psychiatric disease, known eating disorders). A small number of individuals (n=43) with a BMI of 19.0 kg/m^2 were included as they

had a strong family history of thinness. The case notes of each potential participant were reviewed by the GP or a senior nurse with clinical knowledge of the participant to exclude other potential causes of low body weight in discussion with the study team. Through this approach we identified 25,000 individuals who fitted our criteria for inclusion in the study. These individuals were invited to participate in the study; approximately 12% (2,900) replied consenting to take part. We obtained a detailed medical and medication history, screened for eating disorders using a questionnaire (SCOFF) that has been validated against more formal clinical assessment [50]. We excluded all participants who stated that they exercised every day/more than 3 times a week/whose reported activity exceeded 6 metabolic equivalents (METs) for any duration or frequency (http://www.who.int/dietphysicalactivity/physical_activity_intensity/en/). With these rather strict criteria for exercise, we sought to limit the contribution of exercise as a contributor to the thinness of participants in the STILTS cohort. We excluded people who were thin only at a certain point in their lives (often as young adults) to focus on those who were persistently thin/always thin throughout life as we hypothesised that this group would be enriched for genetic factors contributing to their thinness. We asked a specific question to identify these individuals: “have you always been thin?” Only those who answered positively were included. Questionnaires were manually checked by senior clinical staff for these parameters and for reported ethnicity (non-European ancestry excluded). DNA was extracted from salivary samples obtained from these individuals using the Oragene 500 kit according to manufacturer’s instructions (**S1 Table**).

SCOOP

With ethical committee approval (MREC 97/5/21), we have recruited 7,000 individuals with severe early-onset obesity (BMI standard deviation score (SDS) > 3; onset of obesity before the age of 10 years) to the Genetics of Obesity Study (GOOS) [51]. The Severe Childhood Onset Obesity Project (SCOOP) cohort [31] is a sub-cohort of GOOS comprised of ~4,800 British individuals of European ancestry; **S1 Table**). SCOOP individuals likely to have congenital leptin deficiency, a treatable cause of severe obesity, were excluded by measurement of serum leptin, and individuals with mutations in the melanocortin 4 receptor gene (*MC4R*) (the most common genetic form of penetrant obesity) were excluded by prior Sanger sequencing.

UKHLS

Understanding Society (UKHLS) is a longitudinal household study designed to capture economic, social and health information from UK individuals[52]. A subset of 10,484 individuals was selected for genome-wide array genotyping. This cohort was used as a control dataset with SCOOP and STILTS cases (**S1 Table**).

UK BIOBANK (UKBB)

This study includes approximately 487,411 participants with genetic data released (including ~50,000 from the UKBiLEVE cohort [53]) of the total 502,648 individuals from UK BioBank (UKBB). UKBB samples were genotyped on the UK Biobank Axiom array at the Affymetrix Research Services Laboratory in Santa Clara, California, USA and imputed to the Haplotype Reference Consortium (HRC) panel [54]. UKBiLEVE samples were genotyped on the UK BiLEVE array which is a previous

449 version of the UK Biobank Axiom array sharing over 95% of the markers. To date, 487,411 samples
450 with directly genotyped and imputed data are available and data was downloaded using tools
451 provided by UK Biobank. Extensive data from health and lifestyle questionnaires is currently
452 available as well as linked clinical records. BMI, as well as other physical measurements were taken
453 on attendance of recruitment centre. Severely obese participants in the available data were defined
454 as those with BMI ≥ 40 kg/m^2 (N=9,706) and thin individuals were defined as those with BMI ≤ 19
455 kg/m^2 (N=4,538). Given that it has been previously shown that type I error rate for variants with a
456 low minor allele count (MAC) is inadequately controlled for in very unbalanced case-control
457 scenarios[55], we randomly subsampled 35,000 individuals from the original 487,411 genotyped
458 individuals and removed those with BMI≤19 or BMI ≥30, to generate an independent control set.
459 The 25,856 participants remaining after BMI exclusions from the tails, generated a non-extreme set
460 of individuals kept as putative controls (**S2 Fig**). The other 452,411 genotyped samples were kept as
461 the BMI dataset for downstream analyses (**S11 Table, S2 Fig**). An interim release consisting of a
462 subset 152,249 individuals from UKBB was released in May 2015. This interim release was imputed
463 to a combined UK10K and 1000G Phase 3 reference panel and contains several variants which are
464 not currently present in the HRC panel, as such it was used in some of the analyses described.
465
466 **ALSPAC**
467 The Avon Longitudinal Study of Parents and Children (ALSPAC) [27,56], also known as Children of
468 the 90s, is a prospective population-based British birth cohort study. Ethical approval for the study
469 was obtained from the ALSPAC Ethics and Law Committee and the Local Research Ethics

Committees. Please note that the study website contains details of all the data that is available through a fully searchable data dictionary (http://www.bris.ac.uk/alspac/researchers/data-access/data-dictionary/). Further information about this cohort, including details of the genotyping and imputation procedures, can be found in **S2 Appendix**. This analysis was restricted to a subset of unrelated (identity-by-state < 0.05 [57]) children with genetic data and BMI measured between the age of 12 and 17 years (n=4,964, 48.5% male). The mean age of the children was 14 years and the mean BMI 20.5.

GENOTYPING AND QUALITY CONTROL

SCOOP, STILTS and UKHLS

For the SCOOP cohort, DNA was extracted from whole blood as previously described [31]. For the STILTS cohort, DNA was extracted from saliva using the Oragene saliva DNA kits (online protocol) and quantified using Qubit. All samples from SCOOP, STILTS and UKHLS were typed across 30 SNPs on the Sequenom platform (Sequenom Inc. California, USA) for sample quality control. Of the 3,607 SCOOP and STILTS samples submitted for Sequenom genotyping, 3,280 passed quality controls filters (90.9% pass rate). Of the 10,433 UKHLS samples, 9,965 passed Sequenom sample quality control (95.5% pass rate). Subsequently, UKHLS controls were genotyped on the Illumina HumanCoreExome-12v1-0 Beadchip. The 3,280 SCOOP and STILTS samples, and 48 overlapping UKHLS samples (to test for possible array version effects) were genotyped on the Illumina HumanCoreExome-12v1-1 Beadchip by the Genotyping Facility at the Wellcome Sanger Institute (WSI). Genotype calling was performed centrally for all batches at the WSI using GenCall. Criteria

for excluding samples were as follows: i) concordance against Sequenom genotypes <90%; ii) for each pair of sample duplicates, exclude one with highest missingness; iii) sex inferred from genetic data different from stated sex ; iv) sample call rate <95%; v) sample autosome heterozygosity rate >3 SDS from mean done separately for low (<1%) and high MAF(>1%) bins; vi) magnitude of intensity signal in both channels <90%; and vii) for each pair of related individuals (proportion of IBD (PI_HAT) >0.05), the individual with the lowest call rate was excluded. We performed SNP QC using PLINK v1.07[58]. Criteria for excluding SNPs was: i) Hardy-Weinberg equilibrium (HWE) $p<1x10^{-6}$; ii) Call rate <95% for MAF≥5%, call rate <97% for 1% ≤MAF<5%, and call rate <99% for MAF <1%. SMARTPCA v10210 [59] was used for principal component analysis (PCA). To verify the absence of array version effects we used PCA on the subset of shared controls genotyped on both versions of the array. Cut-offs for samples that diverged from the European cluster were chosen manually after inspecting the PCA plot. SNPs with discordant MAFs in the different versions of the array were excluded. After removal of non-European samples and 13 samples due to cryptic relatedness, 1,456 SCOOP and 1,471 STILTS samples remained for analysis. For UKHLS, 82 samples were removed after applying a strict European filter and 680 related samples were removed after applying a "3^{rd} degree" kinship filter in KING[60]. A total of 9,203 samples remained, of which 6,460 had a BMI >19 and <30 ("controls").

UK BIOBANK

Sample QC was performed using all 487,411 samples. Criteria for excluding samples were as follows: i) supplied and genetically inferred sex mismatches; ii) heterozygosity and missingness

512 outliers according to centrally provided sample QC files; iii) samples not used in kinship estimation
513 by UKBB; iv) individuals that did not identify as "white british" or did not cluster with other "white
514 british" in PCA analysis ; v) samples that withdrew consent and vi) for each pair of related
515 individuals (KING kinship estimate>0.0442), we randomly selected an individual preferentially
516 keeping cases if one related individual is a control. After sample QC, thirteen individuals with
517 underlying health conditions that could influence their BMI were also removed, twelve had BMI<14,
518 and one had BMI>74. In the end, 7,526 obese, 3,532 thin and 20,720 non-extreme controls
519 remained for case-control analyses. In addition, 387,164 samples remained for analysis of BMI as a
520 continuous trait. There is an overlap of 10, 282 samples (~2.6% of the BMI dataset) with obese and
521 thin cases (**S2 Fig**). The same procedure was performed on the interim release of 152,249 UKBB
522 samples to produce a set of 2,799 obese, 1,212 thin, 8,193 controls and 127,672 individuals for the
523 independent BMI dataset. All subsequent analyses on UKBB were also performed on this subset to
524 query variants that are not currently available in the full UKBB release.
525
526 **IMPUTATION AND GENOME-WIDE ASSOCIATION ANALYSES**
527 **SCOOP, STILTS and UKHLS single-variant association analysis**
528 Genotypes from SCOOP, STILTS and UKHLS controls were phased together with SHAPEITv2 [61], and
529 subsequently imputed with IMPUTE2 [62,63] to the merged UK10K and 1000G Phase 3 reference
530 panel [64], containing ~91.3 million autosomal and chromosome X sites, from 6,285 samples. More
531 than 98% of variants with MAF ≥0.5% had an imputation quality score of r^2≥0.4, however variants
532 with MAF <0.1% had a poor imputation quality with only 27% variants with r^2≥0.4 (**S5 Fig**). First-

pass single-variant association tests were done for all variants irrespective of MAF, or imputation quality score (see below). Analyses of 1,456 SCOOP, 1,471 STILTS and 6,460 controls (BMI range 19-30) of European ancestry were based on the frequentist association test, using the EM algorithm, as implemented in SNPTEST v2.5 [65], under an additive model and adjusting for six PCs and sex as covariates.

UKBB BMI dataset single-variant association analysis

For the BMI dataset, we used BOLT-LMM [66] to perform an association analysis with BMI using sex, age, 10 PCs and UKBB genotyping array as covariates.

Heritability estimates and genetic correlation

Summary statistics from the SCOOP vs. UKHLS, STILTS vs. UKHLS, UKBB obese vs controls, UKBB thin vs controls and UKBB BMI analyses were filtered and a subset of 1,197,969 HapMap3 SNPs was kept in each dataset. Using LD score regression [67] we first calculated the heritability of severe childhood obesity (SCOOP vs UKHLS) and persistent thinness (STILTS vs UKHLS). For severe childhood obesity, we estimated a prevalence of 0.15% using the BMI centile equivalent to 3SDS in children [68]. In the case of persistent thinness (BMI<=19), we used a GP based cohort for our prevalence estimates: CALIBER [69]. The CALIBER database consists of 1,173,863 records derived from GP practices. For the heritability analysis, we used a prevalence estimate of 2.8% for BMI<=19 (Claudia Langenberg and Harry Hemingway, personal communication). We also used LD score

553 regression to calculate the genetic correlation of SCOOP with STILTS, SCOOP with UKBB obese,
554 SCOOP with BMI, STILTS with UKBB thin and STILTS with BMI. The genetic correlation between
555 obesity and persistent thinness with anorexia was estimated using the summary statistics from
556 SCOOP vs UKHLS and STILTS vs. UKHLS, and summary statistics available from the Genetic
557 Consortium for Anorexia Nervosa (GCAN) in LD Hub [70]. The same analysis was repeated for UKBB
558 obese vs controls and UKBB thin vs controls. Genetic correlation estimates for BMI vs Overweight,
559 Obesity Class 1, Obesity Class 2 and Obesity Class 3 were also extracted from LD Hub (**S4 Fig**).
560
561 **Comparison with established GIANT BMI associated loci**
562 We obtained the list of 97 established BMI associated loci from the publicly available data from the
563 GIANT consortium [24]. We used this list as we wanted to focus on established common variation in
564 Europeans with accurate effect sizes for simulations. In order to test whether there is evidence of
565 enrichment of nominally significant signals with consistent direction of effect, we performed a
566 binomial test using the subset of signals with nominal significance in the SCOOP vs UKHLS, and
567 STILTS vs UKHLS analyses. Variance explained was calculated using the rms package [71] v4.5.0 in R
568 [72] and Nagelkerke's R^2 is reported. Power calculations were performed using Quanto [73]. To
569 calculate ORs and SE from the ALSPAC BMI summary statistics we used genotype counts from
570 SNPTEST output. We then used a z-test to test for significant differences between the OR calculated
571 using genotype counts of SCOOP and ALSPAC against the SCOOP vs. UKHLS OR.
572
573 **Simulations under an additive model**

We created 10,000 simulations of 1 million individuals for the 97 GIANT BMI loci randomly sampling alleles based on the allele frequency from the sex-combined European dataset reported in Locke *et al*. [24] using an R script. For each simulated genotype, we simulated phenotypes with DISSECT [74] using the effect size in GIANT and then removed all samples from the lower tail where the phenotype was <3SDs to better reproduce the actual BMI distribution. Afterwards we randomly sampled 1,471 individuals from the bottom 2.8% and 1,456 from top 0.15% and compared against a random set of 6,460 controls from the equivalent percentiles to BMI 19-30. Finally, for each of these loci, we calculated the absolute difference between our observed OR and the mean OR from the simulations and counted how many times we saw an equal or larger absolute difference in the simulated data and assigned a p-value. This was done separately for SCOOP vs UKHLS and STILTS vs UKHLS.

Genetic Risk Score

The R package GTX (https://cran.r-project.org/web/packages/gtx/index.html) was used to transpose genotype probabilities into dosages, and a combined dosage score, weighted by the effect size from GIANT, for 97 BMI SNPs [24] was calculated and standardised. We checked whether there was an ordinal relationship between the genetic risk score and BMI category (i.e. thin, normal, or obese) using ordinal logistic regression with the clm function in the ordinal R package. While the assumption of equal variance appears to hold (**S6 Fig**)**,** the proportional odds assumption indicating equal odds between thin, normal, and obese groups is violated for the BMI genetic risk score and some of the principal component covariates (i.e., PC2, PC3, and PC6). As our primary

model, we ran a partial proportional odds model adjusting for PC1, PC4, and PC5 and allowing the BMI genetic score, PC2, PC3, and PC6 to vary between BMI category. To check for consistency, we ran a partial proportional odds model adjusting for the first six PCs and allowing only the BMI genetic score to vary between BMI group and a full proportional odds model allowing all six PCs and the BMI genetic score to vary between BMI group (**S1 Appendix**). Using ANOVA, we formally tested the proportional odds assumption for the BMI genetic risk score. A genetic risk score was created and an ordinal logistic regression was run for each of the 10,000 simulations. We compared the observed test statistic testing whether the odds were the same by BMI category to the 10,000 simulation test statistics. We calculated the p-value as the number of simulations with a test statistic larger than that observed in the real data. A mean genetic risk score was also calculated for each BMI category (obese, thin and controls) across the 10,000 simulations. A t-test was used to test whether the mean observed GRS score in each category was significantly different from the one estimated using the simulations.

Discovery stage GWAS

First pass single-variant association analyses results were used as discovery datasets for the GWAS. After association analysis, we removed variants with MAF<0.5%, an INFO score <0.4, and HWE $p<1x10^{-6}$, as these highlighted regions of the genome that were problematic, including CNV regions with poor imputation quality. Quantile-quantile plots indicated that the genomic inflation was well controlled for in SCOOP-UKHLS (λ=1.06) and STILTS-UKHLS (λ =1.04), and slightly higher for SCOOP-STILTS (λ =1.08**, S7 Fig**). We used LD score regression [67] to correct for inflation not due to

polygenicity. To identify distinct loci, we performed clumping as implemented in PLINK [58] using summary statistics from the association tests and LD information from the imputed data, clumping variants 250kb away from an index variant and with an r^2>0.1. In order to further identify a set of likely independent signals we performed conditional analysis of the lead SNPs in SNPTEST to take into account long-range LD. A total of 135 autosomal variants with $p<1x10^{-5}$ in any of the three case-control analyses were taken forward for replication in UKBB. All case-control results are reported with the lower BMI group as reference.

UKBB association analysis

We tested 1,208,692 SNPs for association under an additive model in SNPTEST using sex, age, 10 PCs and UKBB genotyping array as covariates. Three comparisons were done: obese vs thin, obese vs controls and controls vs thin. Variants with an INFO score <0.4, HWE $p<1x10^{-6}$ were filtered out from the results. Inflation factors were calculated using HapMap markers. The LD score regression intercepts were 1.0074 in obese vs thin, 1.0057 in obese vs controls and 1.009 in thin vs controls. We used all thin individuals, regardless of health status, as our replication cohort to maximize power. However, using ICD10 codes and self-reported illness data (**Tables S12 and S13**) to remove individuals who had a relevant medical diagnosis before date of attendance at UKBB recruitment centre, yielded 2,518 thin individuals and materially equivalent results (**S8 Fig**).

GIANT, EGG and SCOOP 2013 summary statistics

We obtained summary statistics for the GIANT Extremes obesity meta-analysis [20] from http://portals.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium_data_files. Summary statistics for EGG [30] were obtained from http://egg-consortium.org/childhood-obesity.html. We used summary statistics from our previous study of 1,509 early-onset obesity SCOOP cases compared to 5,380 publicly available WTCCC2 controls (SCOOP 2013) [31]. Data for the SCOOP cases is available to download from the European Genome-Phenome Archive (EGA) using accession number EGAD00010000594. The control samples are available to download using accession numbers EGAD00000000021 and EGAD00000000023. These replication studies are largely non-overlapping with our discovery datasets and each-other. When a lead variant was not available in a replication cohort, a proxy ($r^2 \geq 0.8$) was used in the meta-analysis.

Replication meta-analysis

We meta-analysed summary statistics for the 135 variants reaching $p<1x10^{-5}$ in SCOOP/STILTS/UKHLS with the corresponding results from UKBB and study specific replication cohorts (**Tables S5-S7**). For obese vs. thin and obese vs. controls comparisons we used fixed-effects meta-analysis correcting for unknown sample overlap in replication cohorts using METACARPA [75]. For thin vs. controls we used a fixed-effects meta-analysis in METAL [76]. Heterogeneity was assessed using Cochran's Q-test heterogeneity p-value in METAL. A signal was considered to replicate if it met all the following criteria: i) consistent direction of effect; ii) $p<0.05$ in at least one replication cohort; and iii) the meta-analysis p-value reached standard genome-wide significance ($p<5x10^{-8}$). Given that we are querying additional variants on the lower allele frequency spectrum,

one could also use a more strict genome-wide significance threshold taking into account the increased number of tests ($p \leq 1.17 \times 10^{-8}$) [77]. In practice, this only affected one previously established signal (*SULT1A1*, rs3760091) in our obese vs. controls analysis that fell just below this threshold (**S6 Table**). rs4440960 was later removed from final results (SCOOP vs UKHLS and STILTS vs UKHLS) after close examination revealed it was present in a CNV region with poor imputation quality.

Comparison of newly established candidate loci and UKBB independent BMI dataset

We identified eleven signals in SCOOP vs STILTS, nine in SCOOP vs UKHLS and two in UKHLS vs STILTS that were nominally significant in the UKBB BMI dataset GWAS, and directionally consistent. A binomial test was used to check for enrichment of signals with consistent direction of effect (**S9 Table**).

Lookup of previously identified obesity-related signals in our discovery datasets

We took all signals reaching genome-wide significance, or identified for the first time in the GIANT Extremes obesity meta-analysis [20], with either the tails of BMI or obesity classes, and in childhood obesity studies [30,31] and performed look-up of those signals in all three of our discovery analyses (SCOOP vs STILTS, SCOOP vs UKHLS and UKHLS vs STILTS). ORs and p-values from the previous studies and look-up results from our discovery datasets are reported in **S10 Table**.

677 **Data availability**

678 Summary statistics for the discovery analyses will be available to download from EGA
679 (EGAS00001002624). UKHLS data is available for download in EGA with accession code
680 EGAS00001001232.

Appendix A

Table 1

Obese vs. thin						Discovery cohort										Replication cohorts					Combined analysis		
rsID	Nearest gene	Chr.	Position (bp)	EA	NEA	OR (95% CI)	P value	EAF Ob	EAF Th	Cohort	OR (95% CI)	P value	EAF Ob	EAF Th	OR (95% CI)	P value	HetPVal
rs9930333	*FTO*	16	53799977	G	T	1.70(1.52,1.90)	2.30E-20	49.59%	37.46%	UKBB	1.46(1.38,1.55)	3.60E-36	48.26%	38.93%	1.48(1.42,1.54)	8.52E-76	3.34E-02
										GIANT[k]	1.43(1.34,1.54)	8.10E-25					
rs2168711	*MC4R*	18	57848531	C	T	1.66(1.45,1.89)	8.29E-14	28.90%	19.95%	UKBB	1.23(1.15,1.32)	2.19E-09	26.75%	22.90%	1.27(1.21,1.33)	2.02E-21	1.12E-04
										GIANT[k]	1.20(1.10,1.30)	1.80E-05					
rs6748821	*TMEM18*[a]	2	629601	G	A	1.65(1.42,1.91)	9.45E-11	86.69%	79.84%	UKBB	1.27(1.18,1.37)	1.31E-09	85.00%	81.69%	1.32(1.24,1.39)	7.76E-21	2.81E-03
										GIANT[k]	1.26(1.14,1.39)	9.90E-06					
rs506589	*SEC16B*	1	177894287	C	T	1.46(1.27,1.67)	5.42E-08	23.98%	18.07%	UKBB	1.25(1.17,1.35)	5.44E-10	23.11%	19.16%	1.28(1.21,1.35)	3.14E-20	1.21E-01
										GIANT[k]	1.25(1.14,1.37)	2.70E-06					
rs6738433	*ADCY3-DNAJC27*[b]	2	25159501	C	G	1.43(1.28,1.60)	1.71E-10	47.31%	43.92%	UKBB	1.21(1.14,1.28)	2.74E-10	50.70%	45.96%	1.19(1.14,1.24)	3.19E-17	6.25E-03
										GIANT[k]	1.10(1.03,1.17)	5.70E-03					
rs7132908	*FAIM2*	12	50263148	A	G	1.31(1.17,1.47)	2.26E-06	42.45%	36.27%	UKBB	1.18(1.11,1.25)	5.43E-08	41.11%	37.39%	1.20(1.15,1.26)	1.93E-16	2.52E-01
										GIANT[k]	1.20(1.10,1.30)	6.60E-06					
rs62107261	*FAM150B*	2	422144	T	C	2.37(1.75,3.20)	2.07E-08	96.37%	93.38%	UKBB	1.54(1.35,1.76)	3.57E-10	96.28%	94.36%	1.65(1.46,1.87)	1.15E-15	1.07E-02
rs12507026	*GNPDA2*[c]	4	45181334	T	A	1.30(1.17,1.46)	3.69E-06	47.29%	40.92%	UKBB	1.14(1.08,1.21)	8.76E-06	45.30%	41.98%	1.18(1.13,1.23)	5.53E-15	4.06E-02
										GIANT[k]	1.20(1.12,1.28)	3.10E-07					
rs75398113	*SNRPC*	6	34728071	C	A	1.53(1.27,1.85)	8.91E-06	11.95%	8.04%	UKBB	1.24(1.12,1.37)	2.07E-05	10.47%	8.52%	1.30(1.19,1.42)	5.19E-09	5.56E-02
rs13135092	*SLC39A8*	4	103198082	G	A	1.58(1.30,1.93)	4.70E-06	10.50%	7.24%	UKBB	1.25(1.12,1.39)	5.57E-05	9.24%	7.52%	1.32(1.20,1.45)	1.06E-08	3.59E-02
Obese vs. controls																	
rsID	Nearest gene	Chr.	Position (bp)	EA	NEA	OR (95% CI)	P value	EAF Ob	EAF Co	Cohort	OR (95% CI)	P value	EAF Ob	EAF Co	OR (95% CI)	P value	HetPVal

Appendix A

rs9928094	*FTO*	16	53799905	G	A	1.44(1.33,1.57)	1.42E-18	49.50%	41.32%	UKBB	1.30(1.25,1.35)	2.74E-41	48.34%	41.91%	1.32(1.29,1.36)	5.94E-101	4.41E-05
										SCOOP 2013	1.46(1.34,1.60)	4.88E-17					
										EGG	1.21(1.15,1.28)	7.20E-13					
										GIANT[i]	1.43(1.34,1.54)	6.60E-25					
rs35614134	*MC4R*[d]	18	57832856	AC	A	1.31(1.20,1.44)	6.27E-09	29.01%	23.69%	UKBB	1.22(1.16,1.27)	1.25E-18	26.72%	23.15%	1.23(1.20,1.27)	1.57E-43	3.55E-01
										SCOOP 2013	1.32(1.19,1.46)	1.22E-07					
										EGG	1.22(1.15,1.30)	1.27E-10					
										GIANT[i]	1.20(1.10,1.30)	1.70E-05					
rs66906321	*TMEM18*[e]	2	630070	C	T	1.40(1.24,1.57)	2.35E-08	85.78%	81.35%	UKBB	1.17(1.11,1.24)	3.44E-09	84.44%	82.20%	1.25(1.21,1.29)	9.72E-35	1.33E-02
										SCOOP 2013	1.39(1.24,1.57)	6.65E-08					
										EGG	1.28(1.19,1.37)	5.15E-12					
										GIANT[i]	1.27(1.15,1.40)	3.40E-06					
rs7132908	*FAIM2*[f]	12	50263148	A	G	1.22(1.12,1.32)	3.27E-06	42.45%	37.82%	UKBB	1.15(1.10,1.19)	5.37E-12	41.11%	37.71%	1.17(1.14,1.21)	2.38E-31	4.86E-01
										SCOOP 2013	1.23(1.12,1.35)	8.89E-06					
										EGG	1.18(1.11,1.25)	1.24E-08					
										GIANT[i]	1.20(1.10,1.30)	6.60E-06					
rs2384060	*ADCY3-DNAJC27*[g]	2	25135438	G	A	1.23(1.13,1.34)	1.53E-06	43.52%	38.90%	UKBB	1.11(1.07,1.15)	4.89E-08	47.67%	44.93%	1.14(1.11,1.17)	9.39E-23	1.13E-01
										SCOOP 2013	1.09(1.00,1.19)	5.01XE-02					
										EGG	1.18(1.12,1.24)	1.02E-09					
										GIANT[i]	1.12(1.04,1.19)	1.60E-03					
rs11209947	*NEGR1*[h]	1	72808551	A	T	1.30(1.17,1.44)	8.51E-07	76.58%	72.18%	UKBB	1.11(1.05,1.16)	4.53E-05	81.18%	79.76%	1.17(1.13,1.21)	5.17E-20	7.26E-05
										SCOOP 2013	1.46(1.30,1.63)	2.21E-10					
										EGG	1.13(1.06,1.22)	4.60E-04					
										GIANT[i]	1.22(1.11,1.35)	5.60E-05					
rs12735657	*SEC16B*[i]	1	177809133	C	T	1.24(1.13,1.37)	9.72E-06	24.26%	20.46%	UKBB	1.12(1.07,1.17)	1.48E-06	22.87%	20.94%	1.15(1.12,1.19)	7.26E-19	1.79E-01

										SCOOP 2013	1.20(1.07,1.33)	1.18E-03					
										EGG	1.14(1.06,1.21)	1.52E-04					
										GIANT[i]	1.22(1.11,1.34)	1.80E-05					
rs13104545	*GNPDA2*	4	45184907	A	G	1.27(1.15,1.40)	1.61E-06	27.41%	23.45%	UKBB	1.07(1.02,1.11)	5.35E-03	24.36%	23.26%	1.13(1.09,1.17)	1.47E-11	9.39E-05
										EGG	1.13(1.04,1.22)	3.39E-03					
										GIANT[i]	1.34(1.20,1.49)	1.20E-07					
rs112446794	***CEP120***[j]	5	122665465	T	C	1.23(1.13,1.35)	2.08E-06	33.15%	28.69%	UKBB	1.07(1.02,1.11)	2.55E-03	29.47%	28.21%	1.09(1.06,1.13)	3.45E-10	3.33E-02
										SCOOP 2013	1.08(0.98,1.19)	1.38E-01					
										EGG	1.12(1.06,1.18)	1.22E-04					
										GIANT[i]	1.05(0.97,1.13)	2.40E-01					
rs3760091	*SULT1A1*	16	28620800	C	G	1.24(1.14,1.35)	1.56E-06	64.89%	60.23%	UKBB	1.09(1.04,1.14)	1.19E-04	63.49%	61.44%	1.12(1.07,1.16)	2.65E-08	8.49E-03

Acknowledgements

We are indebted to the participants of the STILTS cohort and the patients and families involved in the Genetics of Obesity Study (GOOS) cohort. We thank the staff of the NIHR Primary Care Research Network, the GPs, Physicians and nurses involved in identifying and recruiting participants to STILTS and GOOS. These data are from Understanding Society: The UK Household Longitudinal Study, which is led by the Institute for Social and Economic Research at the University of Essex and funded by the Economic and Social Research Council. The data were collected by NatCen and the genome wide scan data were analysed by the Wellcome Sanger Institute. The Understanding Society DAC have an application system for genetics data and all use of the data should be approved by them. This research has been conducted using the UK Biobank Resource (Application Number 14069). Data on the childhood obesity trait has been contributed by EGG Consortium and has been downloaded from www.egg-consortium.org. We are extremely grateful to all the families who took part in the ALSPAC study, the midwives for their help in recruiting them, and the whole ALSPAC team, which includes interviewers, computer and laboratory technicians, clerical workers, research scientists, volunteers, managers, receptionists and nurses. The data used for the analyses described in this manuscript were obtained from the GTEx Portal on 08/Feb/2018. The authors would like to thank Emma Gray, Michelle Dignam and staff of the WSI Sample Management and Genotyping facilities for their contribution, as well as, Konstantinos Hatzikotoulas and Ioanna Tachmazidou for their assistance in the QC of UK Biobank data. Understanding Society Scientific Group members: Michaela Benzeval[1], Jonathan Burton[1], Nicholas Buck[1], Annette Jäckle[1], Meena

Kumari[1], Heather Laurie[1], Peter Lynn[1], Stephen Pudney[1], Birgitta Rabe[1], Dieter Wolke[2]. 1) Institute for Social and Economic Research, University of Essex, UK; 2) University of Warwick, UK.

References

1. Ogden CL, Carroll MD, Flegal KM (2014) Prevalence of obesity in the United States. JAMA 312: 189-190.
2. Wardle J, Carnell S, Haworth CM, Plomin R (2008) Evidence for a strong genetic influence on childhood adiposity despite the force of the obesogenic environment. Am J Clin Nutr 87: 398-404.
3. Silventoinen K, Magnusson PK, Tynelius P, Kaprio J, Rasmussen F (2008) Heritability of body size and muscle strength in young adulthood: a study of one million Swedish men. Genet Epidemiol 32: 341-349.
4. Allison DB, Kaprio J, Korkeila M, Koskenvuo M, Neale MC, et al. (1996) The heritability of body mass index among an international sample of monozygotic twins reared apart. Int J Obes Relat Metab Disord 20: 501-506.
5. Akiyama M, Okada Y, Kanai M, Takahashi A, Momozawa Y, et al. (2017) Genome-wide association study identifies 112 new loci for body mass index in the Japanese population. Nat Genet 49: 1458-1467.
6. Grarup N, Moltke I, Andersen MK, Dalby M, Vitting-Seerup K, et al. (2018) Loss-of-function variants in ADCY3 increase risk of obesity and type 2 diabetes. Nat Genet 50: 172-174.
7. Justice AE, Winkler TW, Feitosa MF, Graff M, Fisher VA, et al. (2017) Genome-wide meta-analysis of 241,258 adults accounting for smoking behaviour identifies novel loci for obesity traits. Nat Commun 8: 14977.
8. Minster RL, Hawley NL, Su CT, Sun G, Kershaw EE, et al. (2016) A thrifty variant in CREBRF strongly influences body mass index in Samoans. Nat Genet 48: 1049-1054.
9. Pigeyre M, Yazdi FT, Kaur Y, Meyre D (2016) Recent progress in genetics, epigenetics and metagenomics unveils the pathophysiology of human obesity. Clin Sci (Lond) 130: 943-986.
10. Turcot V, Lu Y, Highland HM, Schurmann C, Justice AE, et al. (2018) Protein-altering variants associated with body mass index implicate pathways that control energy intake and expenditure in obesity. Nat Genet 50: 26-41.
11. Ramachandrappa S, Raimondo A, Cali AM, Keogh JM, Henning E, et al. (2013) Rare variants in single-minded 1 (SIM1) are associated with severe obesity. J Clin Invest 123: 3042-3050.
12. Doche ME, Bochukova EG, Su HW, Pearce LR, Keogh JM, et al. (2012) Human SH2B1 mutations are associated with maladaptive behaviors and obesity. J Clin Invest 122: 4732-4736.
13. O'Rahilly S, Farooqi IS (2008) Human obesity: a heritable neurobehavioral disorder that is highly sensitive to environmental conditions. Diabetes 57: 2905-2910.
14. Saeed S, Bonnefond A, Tamanini F, Mirza MU, Manzoor J, et al. (2018) Loss-of-function mutations in ADCY3 cause monogenic severe obesity. Nat Genet 50: 175-179.
15. Bulik CM, Allison DB (2001) The genetic epidemiology of thinness. Obes Rev 2: 107-115.
16. Costanzo PR, Schiffman SS (1989) Thinness--not obesity--has a genetic component. Neurosci Biobehav Rev 13: 55-58.

17. Magnusson PK, Rasmussen F (2002) Familial resemblance of body mass index and familial risk of high and low body mass index. A study of young men in Sweden. Int J Obes Relat Metab Disord 26: 1225-1231.
18. Laskarzewski PM, Khoury P, Morrison JA, Kelly K, Mellies MJ, et al. (1983) Familial obesity and leanness. Int J Obes 7: 505-527.
19. Whitaker KL, Jarvis MJ, Boniface D, Wardle J (2011) The intergenerational transmission of thinness. Arch Pediatr Adolesc Med 165: 900-905.
20. Berndt SI, Gustafsson S, Magi R, Ganna A, Wheeler E, et al. (2013) Genome-wide meta-analysis identifies 11 new loci for anthropometric traits and provides insights into genetic architecture. Nat Genet 45: 501-512.
21. Hinney A, Nguyen TT, Scherag A, Friedel S, Bronner G, et al. (2007) Genome wide association (GWA) study for early onset extreme obesity supports the role of fat mass and obesity associated gene (FTO) variants. PLoS One 2: e1361.
22. Scannell Bryan M, Argos M, Pierce B, Tong L, Rakibuz-Zaman M, et al. (2014) Genome-wide association studies and heritability estimates of body mass index related phenotypes in Bangladeshi adults. PLoS One 9: e105062.
23. Lee SH, Wray NR, Goddard ME, Visscher PM (2011) Estimating missing heritability for disease from genome-wide association studies. Am J Hum Genet 88: 294-305.
24. Locke AE, Kahali B, Berndt SI, Justice AE, Pers TH, et al. (2015) Genetic studies of body mass index yield new insights for obesity biology. Nature 518: 197-206.
25. Yan X, Wang Z, Schmidt V, Gauert A, Willnow TE, et al. (2018) Cadm2 regulates body weight and energy homeostasis in mice. Mol Metab 8: 180-188.
26. Klimentidis YC, Raichlen DA, Bea J, Garcia DO, Wineinger NE, et al. (2018) Genome-wide association study of habitual physical activity in over 377,000 UK Biobank participants identifies multiple variants including CADM2 and APOE. Int J Obes (Lond).
27. Boyd A, Golding J, Macleod J, Lawlor DA, Fraser A, et al. (2013) Cohort Profile: the 'children of the 90s'--the index offspring of the Avon Longitudinal Study of Parents and Children. Int J Epidemiol 42: 111-127.
28. Bulik-Sullivan B, Finucane HK, Anttila V, Gusev A, Day FR, et al. (2015) An atlas of genetic correlations across human diseases and traits. Nat Genet 47: 1236-1241.
29. Winkler TW, Kutalik Z, Gorski M, Lottaz C, Kronenberg F, et al. (2015) EasyStrata: evaluation and visualization of stratified genome-wide association meta-analysis data. Bioinformatics 31: 259-261.
30. Bradfield JP, Taal HR, Timpson NJ, Scherag A, Lecoeur C, et al. (2012) A genome-wide association meta-analysis identifies new childhood obesity loci. Nat Genet 44: 526-531.
31. Wheeler E, Huang N, Bochukova EG, Keogh JM, Lindsay S, et al. (2013) Genome-wide SNP and CNV analysis identifies common and low-frequency variants associated with severe early-onset obesity. Nat Genet 45: 513-517.
32. Tachmazidou I, Suveges D, Min JL, Ritchie GRS, Steinberg J, et al. (2017) Whole-Genome Sequencing Coupled to Imputation Discovers Genetic Signals for Anthropometric Traits. Am J Hum Genet 100: 865-884.
33. Hinney A, Volckmar AL, Knoll N (2013) Melanocortin-4 receptor in energy homeostasis and obesity pathogenesis. Prog Mol Biol Transl Sci 114: 147-191.
34. Geller F, Reichwald K, Dempfle A, Illig T, Vollmert C, et al. (2004) Melanocortin-4 receptor gene variant I103 is negatively associated with obesity. Am J Hum Genet 74: 572-581.
35. Mitchell JA, Hakonarson H, Rebbeck TR, Grant SF (2013) Obesity-susceptibility loci and the tails of the pediatric BMI distribution. Obesity (Silver Spring) 21: 1256-1260.

36. Beyerlein A, von Kries R, Ness AR, Ong KK (2011) Genetic markers of obesity risk: stronger associations with body composition in overweight compared to normal-weight children. PLoS One 6: e19057.
37. Chan Y, Holmen OL, Dauber A, Vatten L, Havulinna AS, et al. (2011) Common variants show predicted polygenic effects on height in the tails of the distribution, except in extremely short individuals. PLoS Genet 7: e1002439.
38. Young AI, Wauthier F, Donnelly P (2016) Multiple novel gene-by-environment interactions modify the effect of FTO variants on body mass index. Nat Commun 7: 12724.
39. Winkler TW, Justice AE, Graff M, Barata L, Feitosa MF, et al. (2015) The Influence of Age and Sex on Genetic Associations with Adult Body Size and Shape: A Large-Scale Genome-Wide Interaction Study. PLoS Genet 11: e1005378.
40. Qi Q, Kilpelainen TO, Downer MK, Tanaka T, Smith CE, et al. (2014) FTO genetic variants, dietary intake and body mass index: insights from 177,330 individuals. Hum Mol Genet 23: 6961-6972.
41. Bjornland T, Langaas M, Grill V, Mostad IL (2017) Assessing gene-environment interaction effects of FTO, MC4R and lifestyle factors on obesity using an extreme phenotype sampling design: Results from the HUNT study. PLoS One 12: e0175071.
42. Wood AR, Tyrrell J, Beaumont R, Jones SE, Tuke MA, et al. (2016) Variants in the FTO and CDKAL1 loci have recessive effects on risk of obesity and type 2 diabetes, respectively. Diabetologia 59: 1214-1221.
43. Shungin D, Winkler TW, Croteau-Chonka DC, Ferreira T, Locke AE, et al. (2015) New genetic loci link adipose and insulin biology to body fat distribution. Nature 518: 187-196.
44. Lango Allen H, Estrada K, Lettre G, Berndt SI, Weedon MN, et al. (2010) Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature 467: 832-838.
45. Wen W, Kato N, Hwang JY, Guo X, Tabara Y, et al. (2016) Genome-wide association studies in East Asians identify new loci for waist-hip ratio and waist circumference. Sci Rep 6: 17958.
46. Shaheen R, Schmidts M, Faqeih E, Hashem A, Lausch E, et al. (2015) A founder CEP120 mutation in Jeune asphyxiating thoracic dystrophy expands the role of centriolar proteins in skeletal ciliopathies. Hum Mol Genet 24: 1410-1419.
47. Roosing S, Romani M, Isrie M, Rosti RO, Micalizzi A, et al. (2016) Mutations in CEP120 cause Joubert syndrome as well as complex ciliopathy phenotypes. J Med Genet 53: 608-615.
48. Morton NM, Nelson YB, Michailidou Z, Di Rollo EM, Ramage L, et al. (2011) A stratified transcriptomics analysis of polygenic fat and lean mouse adipose tissues identifies novel candidate obesity genes. PLoS One 6: e23944.
49. Simoncic M, Horvat S, Stevenson PL, Bunger L, Holmes MC, et al. (2008) Divergent physical activity and novel alternative responses to high fat feeding in polygenic fat and lean mice. Behav Genet 38: 292-300.
50. Morgan JF, Reid F, Lacey JH (1999) The SCOFF questionnaire: assessment of a new screening tool for eating disorders. BMJ 319: 1467-1468.
51. Bochukova EG, Huang N, Keogh J, Henning E, Purmann C, et al. (2010) Large, rare chromosomal deletions associated with severe early-onset obesity. Nature 463: 666-670.
52. University of Essex. Institute for Social and Economic Research and NatCen Social Research Understanding Society: Waves 1-5, 2009-2014 [computer file]. 7th Edition. Colchester, Essex: UK Data Archive [distributor] November 2015 SN: 6614.
53. Wain LV, Shrine N, Miller S, Jackson VE, Ntalla I, et al. (2015) Novel insights into the genetics of smoking behaviour, lung function, and chronic obstructive pulmonary disease (UK BiLEVE): a genetic association study in UK Biobank. Lancet Respir Med 3: 769-781.

54. McCarthy S, Das S, Kretzschmar W, Delaneau O, Wood AR, et al. (2016) A reference panel of 64,976 haplotypes for genotype imputation. Nat Genet 48: 1279-1283.
55. Ma C, Blackwell T, Boehnke M, Scott LJ, Go TDi (2013) Recommended joint and meta-analysis strategies for case-control association testing of single low-count variants. Genet Epidemiol 37: 539-550.
56. Fraser A, Macdonald-Wallis C, Tilling K, Boyd A, Golding J, et al. (2013) Cohort Profile: the Avon Longitudinal Study of Parents and Children: ALSPAC mothers cohort. Int J Epidemiol 42: 97-110.
57. Yang J, Lee SH, Goddard ME, Visscher PM (2011) GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet 88: 76-82.
58. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81: 559-575.
59. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, et al. (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38: 904-909.
60. Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, et al. (2010) Robust relationship inference in genome-wide association studies. Bioinformatics 26: 2867-2873.
61. Delaneau O, Marchini J, Zagury JF (2011) A linear complexity phasing method for thousands of genomes. Nat Methods 9: 179-181.
62. Howie BN, Donnelly P, Marchini J (2009) A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet 5: e1000529.
63. Howie B, Marchini J, Stephens M (2011) Genotype imputation with thousands of genomes. G3 (Bethesda) 1: 457-470.
64. Huang J, Howie B, McCarthy S, Memari Y, Walter K, et al. (2015) Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. Nat Commun 6: 8111.
65. Marchini J, Howie B, Myers S, McVean G, Donnelly P (2007) A new multipoint method for genome-wide association studies by imputation of genotypes. Nat Genet 39: 906-913.
66. Loh PR, Tucker G, Bulik-Sullivan BK, Vilhjalmsson BJ, Finucane HK, et al. (2015) Efficient Bayesian mixed-model analysis increases association power in large cohorts. Nat Genet 47: 284-290.
67. Bulik-Sullivan BK, Loh PR, Finucane HK, Ripke S, Yang J, et al. (2015) LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat Genet 47: 291-295.
68. Felix JF, Bradfield JP, Monnereau C, van der Valk RJ, Stergiakouli E, et al. (2016) Genome-wide association analysis identifies three new susceptibility loci for childhood body mass index. Hum Mol Genet 25: 389-403.
69. Denaxas SC, George J, Herrett E, Shah AD, Kalra D, et al. (2012) Data resource profile: cardiovascular disease research using linked bespoke studies and electronic health records (CALIBER). Int J Epidemiol 41: 1625-1638.
70. Zheng J, Erzurumluoglu AM, Elsworth BL, Kemp JP, Howe L, et al. (2017) LD Hub: a centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis. Bioinformatics 33: 272-279.
71. Harrell FE rms: R functions for biostatistical/epidemiologic modeling, testing, estimation, validation, graphics, prediction, and typesetting by storing enhanced model design attributes in the fit, 2013. Implements methods in *Regression Modeling Strategies*, New York:Springer, 2001.
72. R Development Core Team (2011) R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing.
73. Gauderman W, Morrison J (2006) QUANTO 1.1: A computer program for power and sample size calculations for genetic-epidemiology studies, http://hydra.usc.edu/gxe.
74. Canela-Xandri O, Law A, Gray A, Woolliams JA, Tenesa A (2015) A new tool called DISSECT for analysing large genomic data sets using a Big Data approach. Nat Commun 6: 10162.

75. Southam L, Gilly A, Suveges D, Farmaki AE, Schwartzentruber J, et al. (2017) Whole genome sequencing and imputation in isolated populations identify genetic associations with medically-relevant complex traits. Nat Commun 8: 15606.
76. Willer CJ, Li Y, Abecasis GR (2010) METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics 26: 2190-2191.
77. Xu C, Tachmazidou I, Walter K, Ciampi A, Zeggini E, et al. (2014) Estimating genome-wide significance for whole-genome sequencing studies. Genet Epidemiol 38: 281-290.

Supporting information captions

S1 Appendix. **Assessing equal vs. unequal effects for the genetic risk score.**

S2 Appendix. **The Avon Longitudinal Study of Parents and Children.**

S1 Fig. **Mean GRS for SCOOP and STILTS compared to simulations**. Histogram represents mean GRS scores for each BMI category across 10,000 simulations. Vertical red line highlights the observed value in real data. p=p-value of difference**.**

S2 Fig. **Summary of the UKBB sample sets after QC.** Venn Diagram showing sample numbers and overlap between UKBB sample sets used in genetic correlation (BMI dataset) and GWAS replication (obese, controls, thin) analyses.

S3 Fig. Manhattan plot of SCOOP vs STILTS. Manhattan plot produced in EasyStrata, red line indicates genome-wide significance threshold at p=5x10-08. Orange line indicates discovery significance threshold at p=1x10-05. Black labels highlight known BMI/obesity loci that were taken forward for replication and yellow peaks indicate those that met genome-wide significance after replication.

S4 Fig. Genetic correlation of traits and BMI. Genetic correlation estimates and 95% CI for severe early-onset childhood obesity (SCOOP), healthy persistent thinness (STILTS), Obesity Class 3, Obesity Class 2, Obesity Class 1 and Overweight. Dotted lines represent complete genetic correlation.

S5 Fig. Quality of UK10K+1000G imputed genotypes. Percentage of variants with INFO score (r^2)>0.4, as derived from the IMPUTE2 imputation algorithm, stratified by minor allele frequency across all samples (SCOOP, STILTS and UKHLS).

S6 Fig. Box and density plots of risk score weighted by effect size for 97 BMI associated SNPs from GIANT. A weighted genetic risk score for each individual was obtained by summing genotype dosages multiplied by the effect (beta) estimates from GIANT for each of the 97 SNPs. To check the equal variance assumption, we used a box plot (left) and density plot (right). Density plot: Green = STILTS; Blue = UKHLS; Red = SCOOP.

S7 Fig. Quantile-quantile plots of three discovery analysis cohorts. Q-Q plots of LD Score Regression-corrected p-values for the three analysis cohorts used for the discovery analysis, produced in EasyStrata. Red=SCOOP vs. STILTS; Black=SCOOP vs. UKHLS, Blue=STILTS vs. UKHLS. Variants passing QC and with MAF >=0.5% are shown. LD Score regression intercept (λLD) values before correction are shown for each analysis.

S8 Fig. Quantile-quantile plots for UKBB case-control analysis with different exclusion criteria for thin individuals. Q-Q plot using all thin individuals as cases (Full UKBB) and removing individuals based on ICD10 and self-reported data (ICD10+self-reported filter). Correlation for –log10 p-values is shown (r=0.7462).

S1 Table. Summary of discovery sample sets.

S2 Table. 97 BMI SNPs from the GIANT consortium study and their summary statistics in our three analysis cohorts.

S3 Table. Nominally significant loci for non-additive effect in extremes.

S4 Table. Difference in SCOOP OR when using ALSPAC as control dataset vs. UKHLS.

S5 Table. Discovery, replication and meta-analysis results for 32 SNPs meeting P<10-5 in discovery association results of SCOOP vs STILTS analysis.

S6 Table. Discovery, replication and meta-analysis results for 66 SNPs meeting P<10-5 in discovery association results of SCOOP vs UKHLS analysis.

S7 Table. Discovery, replication and meta-analysis results for 37 SNPs meeting P<10-5 in discovery association results of UKHLS vs STILTS analysis.

S8 Table: Reciprocal analysis of previously established signals and lead signals in this study.

S9 Table. Consistency of the direction of effect in candidate loci meeting p<1x10-5 in the discovery stages with BMI dataset GWAS.

S10 Table. Published loci from GIANT, EGG and SCOOP 2013 not reaching genome-wide significance in our study

S11 Table. Summary of UKBB sample sets.

S12 Table. ICD10 codes used to exclude thin individuals in UKBB

S13 Table. Self-reported illness codes used to exclude thin individuals in UKBB

S2 Table. 97 BMI SNPs from the GIANT consortium study and their summary statistics in our three analysis cohorts.

rsID	Chr.	Position (bp)*	Nearest gene(s)	GIANT						SCOOP vs. STILTS							SCOOP vs. UKHLS							UKHLS vs. STILTS						
				Effect	Other	EAF	Beta	SE	P value	Effect	Other	EAF	OR	95% CI	P value	Direction	Effect	Other	EAF	OR	95% CI	P value	Direction	Effect	Other	EAF	OR	95% CI	P value	Direction
rs1558902	16	53803574	*FTO*	A	T	0.415	0.082	0.003	7.51E-153	A	T	0.419	1.689	(1.509,1.89)	7.31E-20	+	A	T	0.410	1.429	(1.317,1.55)	1.25E-17	+	A	T	0.387	1.174	(1.077,1.28)	2.78E-04	+
rs6567160	18	57829135	*MC4R*	C	T	0.236	0.056	0.004	3.93E-53	C	T	0.244	1.643	(1.44,1.875)	1.77E-13	+	C	T	0.246	1.308	(1.194,1.433)	7.91E-09	+	C	T	0.229	1.257	(1.134,1.394)	1.38E-05	+
rs13021737	2	632348	*TMEM18*	G	A	0.828	0.060	0.004	1.113E-50	G	A	0.832	1.586	(1.365,1.844)	1.81E-09	+	G	A	0.832	1.356	(1.206,1.526)	3.89E-07	+	G	A	0.820	1.211	(1.088,1.347)	4.44E-04	+
rs10938397	4	45182527	*GNPDA2*	G	A	0.434	0.040	0.003	3.205E-38	G	A	0.442	1.302	(1.164,1.457)	4.19E-06	+	G	A	0.440	1.186	(1.093,1.287)	4.50E-05	+	G	A	0.429	1.084	(0.996,1.179)	6.24E-02	+
rs543874	1	177889480	*SEC16B*	G	A	0.193	0.048	0.004	2.618E-35	G	A	0.210	1.452	(1.267,1.664)	8.57E-08	+	G	A	0.213	1.201	(1.09,1.324)	2.22E-04	+	G	A	0.202	1.177	(1.056,1.311)	3.11E-03	+
rs2207139	6	50845490	*TFAP2B*	G	A	0.177	0.045	0.004	4.126E-29	G	A	0.171	1.292	(1.116,1.496)	5.91E-04	+	G	A	0.170	1.175	(1.057,1.306)	2.70E-03	+	G	A	0.163	1.116	(0.994,1.253)	6.21E-02	+
rs11030104	11	27684517	*BDNF*	A	G	0.792	0.041	0.004	5.557E-28	A	G	0.796	1.320	(1.151,1.513)	7.31E-05	+	A	G	0.799	1.143	(1.029,1.269)	1.27E-02	+	A	G	0.792	1.122	(1.015,1.24)	2.43E-02	+
rs3101336	1	72751185	*NEGR1*	C	T	0.613	0.033	0.003	2.661E-26	C	T	0.613	1.245	(1.111,1.395)	1.58E-04	+	C	T	0.605	1.195	(1.098,1.3)	3.66E-05	+	C	T	0.595	1.056	(0.97,1.15)	2.07E-01	+
rs7138803	12	50247468	*BCDIN3D*	A	G	0.384	0.032	0.003	8.153E-24	A	G	0.379	1.286	(1.149,1.441)	1.31E-05	+	A	G	0.371	1.216	(1.118,1.322)	4.68E-06	+	A	G	0.360	1.035	(0.948,1.13)	4.47E-01	+
rs10182181	2	25150296	*ADCY3*	G	A	0.462	0.031	0.003	8.777E-24	G	A	0.490	1.415	(1.268,1.58)	6.34E-10	+	G	A	0.496	1.202	(1.108,1.304)	9.30E-06	+	G	A	0.480	1.186	(1.09,1.29)	6.81E-05	+
rs3888190	16	28889486	*ATP2A1*	A	C	0.403	0.031	0.003	3.14E-23	A	C	0.407	1.136	(1.017,1.271)	2.46E-02	+	A	C	0.401	1.129	(1.04,1.226)	3.87E-03	+	A	C	0.395	1.035	(0.95,1.126)	4.34E-01	+
rs1516725	3	185824004	*ETV5*	C	T	0.872	0.045	0.005	1.886E-22	C	T	0.859	1.303	(1.112,1.526)	1.05E-03	+	C	T	0.863	1.158	(1.024,1.309)	1.89E-02	+	C	T	0.856	1.180	(1.051,1.325)	5.03E-03	+
rs12446632	16	19935389	*GPRC5B*	G	A	0.865	0.040	0.005	1.477E-18	G	A	0.849	1.279	(1.094,1.495)	2.03E-03	+	G	A	0.856	1.097	(0.975,1.235)	1.24E-01	+	G	A	0.850	1.194	(1.066,1.337)	2.20E-03	+
rs2287019	19	46202172	*QPCTL*	C	T	0.804	0.036	0.004	4.585E-18	C	T	0.814	1.038	(0.898,1.2)	6.14E-01	+	C	T	0.816	1.024	(0.919,1.142)	6.61E-01	+	C	T	0.814	1.047	(0.939,1.168)	4.06E-01	+
rs16951275	15	68077168	*MAP2K5*	T	C	0.784	0.031	0.004	1.911E-17	T	C	0.779	1.173	(1.028,1.339)	1.79E-02	+	T	C	0.776	1.133	(1.025,1.251)	1.43E-02	+	T	C	0.770	1.056	(0.957,1.164)	2.80E-01	+
rs3817334	11	47650993	*MTCH2*	T	C	0.407	0.026	0.003	5.145E-17	T	C	0.408	1.202	(1.073,1.347)	1.46E-03	+	T	C	0.414	1.093	(1.006,1.187)	3.52E-02	+	T	C	0.405	1.098	(1.008,1.196)	3.29E-02	+
rs2112347	5	75015242	*POC5*	T	G	0.629	0.026	0.003	6.191E-17	T	G	0.633	1.087	(0.969,1.219)	1.55E-01	+	T	G	0.637	1.032	(0.949,1.124)	4.60E-01	+	T	G	0.634	1.065	(0.977,1.16)	1.51E-01	+
rs12566985	1	75002193	*FPGT-TNNI3K*	G	A	0.446	0.024	0.003	3.282E-15	G	A	0.448	1.273	(1.138,1.424)	2.48E-05	+	G	A	0.439	1.206	(1.11,1.311)	1.04E-05	+	G	A	0.429	1.038	(0.952,1.132)	3.96E-01	+
rs3810291	19	47569003	*ZC3H4*	A	G	0.666	0.028	0.004	4.812E-15	A	G	0.670	1.198	(1.067,1.345)	2.19E-03	+	A	G	0.671	1.134	(1.039,1.237)	4.69E-03	+	A	G	0.663	1.072	(0.983,1.17)	1.15E-01	+
rs7141420	14	79899454	*NRXN3*	T	C	0.527	0.024	0.003	1.23E-14	T	C	0.518	1.151	(1.03,1.287)	1.33E-02	+	T	C	0.513	1.112	(1.025,1.208)	1.11E-02	+	T	C	0.507	1.003	(0.922,1.091)	9.48E-01	+
rs13078960	3	85807590	*CADM2*	G	T	0.196	0.030	0.004	1.737E-14	G	T	0.192	1.170	(1.018,1.344)	2.74E-02	+	G	T	0.206	0.994	(0.899,1.099)	9.08E-01	-	G	T	0.201	1.198	(1.076,1.334)	9.49E-04	+
rs10968576	9	28414339	*LINGO2*	G	A	0.320	0.025	0.003	6.607E-14	G	A	0.322	1.026	(0.91,1.155)	6.78E-01	+	G	A	0.316	1.049	(0.96,1.145)	2.88E-01	+	G	A	0.316	0.964	(0.881,1.055)	4.22E-01	-
rs17024393	1	110154688	*GNAT2*	C	T	0.040	0.066	0.009	7.029E-14	C	T	0.029	1.802	(1.284,2.529)	6.58E-04	+	C	T	0.026	1.568	(1.246,1.973)	1.26E-04	+	C	T	0.023	1.099	(0.824,1.465)	5.20E-01	+
rs657452	1	49589847	*AGBL4*	A	G	0.394	0.023	0.003	5.482E-13	A	G	0.381	1.103	(0.981,1.24)	1.01E-01	+	A	G	0.384	1.035	(0.95,1.126)	4.31E-01	+	A	G	0.380	1.037	(0.95,1.132)	4.14E-01	+
rs12429545	13	54102206	*OLFM4*	A	G	0.133	0.033	0.005	1.094E-12	A	G	0.134	1.076	(0.916,1.264)	3.72E-01	+	A	G	0.130	1.119	(0.992,1.261)	6.73E-02	+	A	G	0.127	0.982	(0.866,1.115)	7.82E-01	-
rs12286929	11	115022404	*CADM1*	G	A	0.523	0.022	0.003	1.31E-12	G	A	0.529	1.143	(1.024,1.276)	1.74E-02	+	G	A	0.530	1.069	(0.985,1.16)	1.11E-01	+	G	A	0.524	1.063	(0.978,1.155)	1.49E-01	+
rs13107325	4	103188709	*SLC39A8*	T	C	0.072	0.048	0.007	1.825E-12	T	C	0.081	1.605	(1.309,1.967)	5.14E-06	+	T	C	0.081	1.284	(1.116,1.477)	4.84E-04	+	T	C	0.075	1.203	(1.019,1.42)	2.89E-02	+
rs11165643	1	96924097	*PTBP2*	T	C	0.583	0.022	0.003	2.07E-12	T	C	0.588	1.022	(0.913,1.144)	7.03E-01	+	T	C	0.591	1.017	(0.936,1.105)	6.97E-01	+	T	C	0.589	1.030	(0.947,1.12)	4.94E-01	+
rs7903146	10	114758349	*TCF7L2*	C	T	0.713	0.023	0.003	1.112E-11	C	T	0.717	1.015	(0.899,1.147)	8.07E-01	+	C	T	0.712	1.049	(0.959,1.149)	2.94E-01	+	C	T	0.711	0.956	(0.872,1.048)	3.34E-01	-
rs10132280	14	25928179	*STXBP6*	C	A	0.682	0.023	0.003	1.141E-11	C	A	0.704	0.968	(0.855,1.095)	6.02E-01	-	C	A	0.703	0.991	(0.906,1.084)	8.48E-01	-	C	A	0.704	0.984	(0.897,1.079)	7.25E-01	-
rs17405819	8	76806584	*HNF4G*	T	C	0.700	0.022	0.003	2.07E-11	T	C	0.705	1.269	(1.125,1.433)	1.09E-04	+	T	C	0.706	1.124	(1.026,1.23)	1.19E-02	+	T	C	0.699	1.089	(0.995,1.191)	6.30E-02	+
rs6091540	20	51087862	*ZFP64*	C	T	0.723	0.019	0.004	2.154E-11	C	T	0.711	1.145	(1.015,1.292)	2.76E-02	+	C	T	0.711	1.067	(0.975,1.168)	1.60E-01	+	C	T	0.707	1.035	(0.945,1.133)	4.62E-01	+
rs1016287	2	59305625	*LINC01122*	T	C	0.287	0.023	0.003	2.253E-11	T	C	0.311	1.007	(0.895,1.131)	9.13E-01	+	T	C	0.298	1.070	(0.98,1.168)	1.29E-01	+	T	C	0.298	0.928	(0.849,1.016)	1.05E-01	-
rs4256980	11	8673939	*TRIM66*	G	C	0.646	0.021	0.003	2.9E-11	G	C	0.650	1.031	(0.919,1.157)	6.04E-01	+	G	C	0.654	1.013	(0.93,1.104)	7.68E-01	+	G	C	0.652	1.039	(0.952,1.134)	3.94E-01	+
rs17094222	10	102395440	*HIF1AN*	C	T	0.211	0.025	0.004	5.942E-11	C	T	0.210	1.133	(0.985,1.303)	8.06E-02	+	C	T	0.212	1.043	(0.942,1.155)	4.15E-01	+	C	T	0.209	1.045	(0.939,1.163)	4.25E-01	+
rs12401738	1	78446761	*FUBP1*	A	G	0.352	0.021	0.003	1.145E-10	A	G	0.378	0.988	(0.881,1.106)	8.30E-01	-	A	G	0.377	1.007	(0.925,1.096)	8.75E-01	+	A	G	0.377	0.998	(0.916,1.089)	9.71E-01	-
rs7599312	2	213413231	*ERBB4*	G	A	0.724	0.022	0.003	1.173E-10	G	A	0.732	1.083	(0.957,1.225)	2.08E-01	+	G	A	0.731	1.047	(0.955,1.148)	3.32E-01	+	G	A	0.729	0.985	(0.897,1.081)	7.45E-01	-
rs2365389	3	61236462	*FHIT*	C	T	0.582	0.020	0.003	1.629E-10	C	T	0.595	1.185	(1.058,1.328)	3.43E-03	+	C	T	0.594	1.092	(1.004,1.187)	3.94E-02	+	C	T	0.589	1.060	(0.973,1.154)	1.80E-01	+
rs205262	6	34563164	*C6orf106*	G	A	0.273	0.022	0.004	1.753E-10	G	A	0.275	1.186	(1.049,1.341)	6.38E-03	+	G	A	0.269	1.163	(1.062,1.274)	1.14E-03	+	G	A	0.262	1.051	(0.955,1.156)	3.12E-01	+
rs2820292	1	201784287	*NAV1*	C	A	0.555	0.020	0.003	1.834E-10	C	A	0.562	1.084	(0.969,1.212)	1.57E-01	+	C	A	0.567	1.031	(0.949,1.119)	4.74E-01	+	C	A	0.563	1.095	(1.007,1.191)	3.47E-02	+
rs12885454	14	29738038	*PRKD1*	C	A	0.642	0.021	0.003	1.943E-10	C	A	0.631	1.084	(0.967,1.214)	1.65E-01	+	C	A	0.633	1.084	(0.995,1.126)	4.35E-01	+	C	A	0.630	1.033	(0.948,1.125)	4.61E-01	+
rs9641123	7	93197732	*CALCR*	C	G	0.429	0.019	0.004	2.077E-10	C	G	0.411	1.161	(1.036,1.302)	1.03E-02	+	C	G	0.410	1.096	(1.008,1.193)	3.19E-02	+	C	G	0.404	1.037	(0.951,1.131)	4.09E-01	+
rs12016871	13	28017782	*MTIF3*	T	C	0.203	0.030	0.005	2.291E-10	T	C	0.188	1.089	(0.947,1.252)	2.32E-01	+	T	C	0.179	1.152	(1.039,1.278)	7.09E-03	+	T	C	0.176	0.962	(0.862,1.073)	4.84E-01	-
rs16851483	3	141275436	*RASA2*	T	G	0.066	0.048	0.008	3.548E-10	T	G	0.066	1.428	(1.137,1.795)	2.23E-03	+	T	G	0.067	1.202	(1.027,1.406)	2.17E-02	+	T	G	0.064	1.170	(0.977,1.403)	8.83E-02	+
rs1167827	7	75163169	*HIP1*	G	A	0.553	0.020	0.003	6.333E-10	G	A	0.569	1.167	(1.046,1.302)	5.65E-03	+	G	A	0.572	1.075	(0.99,1.167)	8.67E-02	+	G	A	0.566	1.075	(0.989,1.169)	8.89E-02	+
rs758747	16	3627358	*NLRC3*	T	C	0.265	0.023	0.004	7.473E-10	T	C	0.270	1.138	(1.007,1.287)	3.88E-02	+	T	C	0.280	1.010	(0.923,1.106)	8.29E-01	+	T	C	0.275	1.089	(0.991,1.197)	7.65E-02	+
rs1928295	9	120378483	*TLR4*	T	C	0.548	0.019	0.003	7.91E-10	T	C	0.576	1.116	(0.998,1.249)	5.38E-02	+	T	C	0.568	1.103	(1.015,1.197)	2.00E-02	+	T	C	0.563	0.990	(0.911,1.076)	8.13E-01	-
rs9925964	16	31129895	*KAT8*	A	G	0.620	0.019	0.003	8.108E-10	A	G	0.642	1.036	(0.924,1.161)	5.47E-01	+	A	G	0.646	1.010	(0.927,1.1)	8.27E-01	+	A	G	0.644	1.025	(0.939,1.119)	5.77E-01	+
rs11126666	2	26928811	*KCNK3*	A	G	0.283	0.021	0.003	1.332E-09	A	G	0.255	1.001	(0.883,1.134)	9.89E-01	+	A	G	0.254	0.992	(0.903,1.089)	8.60E-01	-	A	G	0.255	1.004	(0.913,1.104)	9.33E-01	+
rs2650492	16	28333411	*SBK1*	A	G	0.303	0.021	0.004	1.915E-09	A	G	0.306	1.211	(1.076,1.363)	1.48E-03	+	A	G	0.299	1.175	(1.077,1.281)	2.93E-04	+	A	G	0.292	1.056	(0.964,1.158)	2.42E-01	+
rs6804842	3	25106437	*RARB*	G	A	0.575	0.019	0.003	2.476E-09	G	A	0.580	1.118	(1,1.25)	4.99E-02	+	G	A	0.580	1.083	(0.996,1.177)	6.17E-02	+	G	A	0.574	1.037	(0.953,1.128)	3.96E-01	+
rs12940622	17	78615571	*RPTOR*	G	A	0.575	0.018	0.003	2.494E-09	G	A	0.560	1.185	(1.061,1.324)	2.62E-03	+	G	A	0.559	1.121	(1.031,1.218)	7.20E-03	+	G	A	0.551	1.068	(0.981,1.161)	1.28E-01	+
rs7164727	15	73093991	*LOC100287559*	T	C	0.686	0.018	0.003	3.915E-09	T	C	0.676	0.940	(0.835,1.058)	3.04E-01	-	T	C	0.669	0.992	(0.909,1.082)	8.56E-01	-	T	C	0.672	0.949	(0.867,1.04)	2.61E-01	-
rs11847697	14	30515112	*PRKD1*	T	C	0.042	0.049	0.008	3.99E-09	T	C	0.045	1.420	(1.093,1.846)	8.68E-03	+	T	C	0.045	1.252	(1.041,1.507)	1.72E-02	+	T	C	0.042	1.244	(1,1.548)	5.05E-02	+
rs4740619	9	15634326	*C9orf93*	T	C	0.542	0.018	0.003	4.564E-09	T	C	0.537	1.183	(1.059,1.321)	2.94E-03	+	T	C	0.548	1.054	(0.971,1.143)	2.10E-01	+	T	C	0.540	1.124	(1.034,1.221)	5.88E-03	+
rs492400	2	219349752	*USP37*	C	T	0.423	0.016	0.003	6.784E-09	C	T	0.426	1.097	(0.983,1.225)	9.79E-02	+	C	T	0.431	1.033	(0.951,1.121)	4.42E-01	+	C	T	0.426	1.063	(0.976,1.158)	1.58E-01	+
rs13191362	6	163033350	*PARK2*	A	G	0.879	0.028	0.005	7.339E-09	A	G	0.878	1.164	(0.981,1.38)	8.18E-02	+	A	G	0.878	1.073	(0.945,1.219)	2.78E-01	+	A	G	0.876	1.069	(0.943,1.211)	2.98E-01	+
rs3736485	15	51748610	*DMXL2*	A	G	0.454	0.018	0.003	7.412E-09	A	G	0.472	1.021	(0.914,1.14)	7.18E-01	+	A	G	0.460	1.084	(0.998,1.177)	5.48E-02	+	A	G	0.458	0.961	(0.884,1.045)	3.54E-01	-
rs17001654	4	77129568	*SCARB2*	G	C	0.153	0.031	0.005	7.76E-09	G	C	0.153	0.991	(0.849,1.158)	9.12E-01	-	G	C	0.149	1.012	(0.901,1.137)	8.37E-01	+	G	C	0.150	0.950	(0.844,1.069)	3.94E-01	-
rs11191560	10	104869038	*NT5C2*	C	T	0.089	0.031	0.005	8.446E-09	C	T	0.082	1.255	(1.025,1.538)	2.83E-02	+	C	T	0.076	1.236	(1.069,1.429)	4.23E-03	+	C	T	0.073	1.000	(0.852,1.174)	9.98E-01	+
rs2080454	16	49062590	*CBLN1*	C	A	0.405	0.017	0.003	8.604E-09	C	A	0.385	1.011	(0.903,1.131)	8.50E-01	+	C	A	0.386	0.989	(0.91,1.075)	7.96E-01	-	C	A	0.387	0.998	(0.917,1.087)	9.70E-01	-
rs7715256	5	153537893	*GALNT10*	G	T	0.421	0.016	0.003	8.851E-09	G	T	0.423	1.180	(1.055,1.318)	3.60E-03	+	G	T	0.428	1.067	(0.982,1.158)	1.26E-01	+	G	T	0.421	1.087	(0.998,1.185)	5.56E-02	+
rs2176040	2	227092802	*LOC646736*	A	G	0.365	0.014	0.003	9.989E-09	A	G	0.364	1.025	(0.913,1.151)	6.75E-01	+	A	G	0.350	1.080	(0.993,1.175)	7.37E-02	+	A	G	0.350	0.930	(0.853,1.013)	9.77E-02	-
rs1528435	2	181550962	*UBE2E3*	T	C	0.631	0.018	0.003	1.196E-08	T	C	0.636	1.013	(0.904,1.136)	8.21E-01	+	T	C	0.624	1.087	(0.999,1.183)	5.27E-02	+	T	C	0.622	0.944	(0.866,1.029)	1.90E-01	-
rs2075650	19	45395619	*TOMM40*	A	G	0.848	0.026	0.005	1.247E-08	G	A	0.147	0.968	(0.828,1.131)	6.82E-01	-	A	G	0.853	1.002	(0.893,1.123)	9.77E-01	+	A	G	0.853	1.007	(0.896,1.132)	9.04E-01	+
rs1000940	17	5283252	*RABEP1*	G	A	0.320	0.019	0.003	1.284E-08	G	A	0.307	1.155	(1.026,1.302)	1.75E-02	+	G	A	0.304	1.115	(1.022,1.218)	1.47E-02	+	G	A	0.298	1.061	(0.968,1.164)	2.04E-01	+
rs11583200	1	50559820	*ELAVL4*	C	T	0.396	0.018	0.003	1.479E-08	C	T	0.376	1.083	(0.965,1.214)	1.75E-01	+	C	T	0.381	1.024	(0.943,1.113)	5.70E-01	+	C	T	0.377	1.058	(0.971,1.153)	2.01E-01	+
rs7239883	18	40147671	*LOC284260*	G	A	0.393	0.016	0.003	1.511E-08	G	A	0.380	1.029	(0.917,1.154)	6.25E-01	+	G	A	0.377	1.029	(0.946,1.119)	5.08E-01	+	G	A	0.376	1.004	(0.921,1.093)	9.33E-01	+
rs2836754	21	40291740	*ETS2*	C	T	0.612	0.016	0.003	1.605E-08	C	T	0.643	1.134	(1.011,1.272)	3.22E-02	+	C	T	0.651	1.053	(0.966,1.147)	2.42E-01	+	C	T	0.645	1.120	(1.027,1.221)	1.03E-02	+
rs9400239	6	108977663	*FOXO3*	C	T	0.688	0.019	0.003	1.613E-08	C	T	0.706	1.214	(1.074,1.371)	1.85E-03	+	C	T	0.709	1.118	(1.019,1.226)	1.84E-02	+	C	T	0.702	1.099	(1.003,1.205)	4.31E-02	+
rs10733682	9	129460914	*LMX1B*	A	G	0.478	0.017	0.003	1.83E-08	A	G	0.479	1.084	(0.968,1.212)	1.61E-01	+	A	G	0.475	1.071	(0.986,1.165)	1.06E-01	+	A	G	0.471	1.014	(0.93,1.105)	7.55E-01	+
rs11688816	2	63053048	*EHBP1*	G	A	0.525	0.017	0.003	1.893E-08	G	A	0.542	1.033	(0.924,1.154)	5.69E-01	+	G	A	0.544	0.988	(0.91,1.074)	7.83E-01	-	G	A	0.544	1.001	(0.92,1.09)	9.76E-01	+
rs11057405	12	122781897	*CLIP1*	G	A	0.901	0.031	0.006	2.019E-08	G	A	0.898	1.070	(0.892,1.284)	4.67E-01	+	G	A	0.891	1.129	(0.985,1.293)	8.09E-02	+	G	A	0.890	0.929	(0.81,1.065)	2.89E-01	-
rs9914578	17	2005136	*SMG6*	G	C	0.211	0.020	0.004	2.072E-08	G	C	0.208	1.044	(0.91,1.197)	5.39E-01	+	G	C	0.207	1.042	(0.943,1.152)	4.17E-01	+	G	C	0.205	1.030	(0.929,1.143)	5.74E-01	+
rs977747	1	47684677	*TAL1*	T	G	0.391	0.017	0.003	2.182E-08	T	G	0.412	1.099	(0.983,1.23)	9.84E-02	+	T	G	0.416	1.043	(0.961,1.133)	3.13E-01	+	T	G	0.411	1.050	(0.965,1.143)	2.54E-01	+

Appendix A

rs2121279	2	143043285	*LRP1B*	T	C	0.152	0.025	0.004	2.313E-08	T	C	0.133	1.063	(0.904,1.252)	4.59E-01	+	T	C	0.127	1.118	(0.991,1.26)	6.99E-02	+	T	C	0.126	0.958	(0.845,1.085)	4.99E-01	-
rs29941	19	34309532	*KCTD15*	G	A	0.669	0.018	0.003	2.407E-08	G	A	0.677	1.187	(1.055,1.336)	4.37E-03	+	G	A	0.671	1.132	(1.037,1.236)	5.77E-03	+	G	A	0.665	1.027	(0.94,1.123)	5.56E-01	+
rs11727676	4	145659064	*HHIP*	T	C	0.910	0.036	0.006	2.55E-08	C	T	0.092	0.999	(0.813,1.228)	9.94E-01	-	T	C	0.904	1.056	(0.907,1.23)	4.80E-01	+	T	C	0.905	0.899	(0.766,1.054)	1.90E-01	-
rs3849570	3	81792112	*GBE1*	A	C	0.359	0.019	0.003	2.601E-08	A	C	0.348	1.040	(0.926,1.169)	5.05E-01	+	A	C	0.346	1.021	(0.937,1.113)	6.37E-01	+	A	C	0.346	0.988	(0.905,1.078)	7.80E-01	-
rs9374842	6	120185665	*LOC285762*	T	C	0.748	0.019	0.004	2.673E-08	T	C	0.775	1.222	(1.073,1.393)	2.54E-03	+	T	C	0.772	1.160	(1.05,1.281)	3.41E-03	+	T	C	0.766	1.058	(0.96,1.166)	2.53E-01	+
rs6477694	9	111932342	*EPB41L4B*	C	T	0.365	0.017	0.003	2.673E-08	C	T	0.356	1.161	(1.035,1.303)	1.07E-02	+	C	T	0.353	1.101	(1.011,1.198)	2.73E-02	+	C	T	0.347	1.043	(0.955,1.139)	3.53E-01	+
rs4787491	16	30015337	*INO80E*	G	A	0.509	0.016	0.003	2.696E-08	G	A	0.538	1.014	(0.908,1.132)	8.08E-01	+	G	A	0.536	1.006	(0.927,1.092)	8.87E-01	+	G	A	0.537	0.981	(0.902,1.067)	6.56E-01	-
rs1441264	13	79580919	*MIR548A2*	A	G	0.609	0.018	0.003	2.959E-08	A	G	0.590	1.082	(0.963,1.215)	1.86E-01	+	A	G	0.590	1.051	(0.963,1.146)	2.68E-01	+	A	G	0.587	1.049	(0.961,1.146)	2.86E-01	+
rs7899106	10	87410904	*GRID1*	G	A	0.052	0.040	0.007	2.96E-08	G	A	0.056	1.269	(0.998,1.613)	5.17E-02	+	G	A	0.051	1.240	(1.043,1.475)	1.48E-02	+	G	A	0.050	0.949	(0.786,1.147)	5.90E-01	-
rs2176598	11	43864278	*HSD17B12*	T	C	0.251	0.020	0.004	2.971E-08	T	C	0.237	1.055	(0.926,1.201)	4.19E-01	+	T	C	0.247	0.957	(0.871,1.053)	3.68E-01	-	T	C	0.246	1.076	(0.976,1.187)	1.41E-01	+
rs2245368	7	76608143	*PMS2L11*	C	T	0.180	0.032	0.006	3.187E-08	C	T	0.178	1.190	(1.025,1.382)	2.27E-02	+	C	T	0.167	1.225	(1.098,1.366)	2.73E-04	+	C	T	0.162	0.984	(0.875,1.105)	7.82E-01	-
rs17203016	2	208255518	*CREB1*	G	A	0.197	0.021	0.004	3.406E-08	G	A	0.213	1.128	(0.987,1.289)	7.77E-02	+	G	A	0.206	1.133	(1.026,1.25)	1.32E-02	+	G	A	0.202	0.982	(0.886,1.088)	7.28E-01	-
rs17724992	19	18454825	*PGPEP1*	A	G	0.746	0.019	0.004	3.415E-08	A	G	0.744	1.196	(1.055,1.356)	5.09E-03	+	A	G	0.741	1.155	(1.05,1.271)	2.99E-03	+	A	G	0.734	1.042	(0.949,1.144)	3.90E-01	+
rs7243357	18	56883319	*GRP*	T	G	0.812	0.022	0.004	3.857E-08	T	G	0.825	1.182	(1.02,1.368)	2.56E-02	+	T	G	0.826	1.106	(0.989,1.236)	7.66E-02	+	T	G	0.821	1.090	(0.978,1.214)	1.19E-01	+
rs16907751	8	81375457	*ZBTB10*	C	T	0.916	0.035	0.007	3.888E-08	C	T	0.906	0.966	(0.797,1.171)	7.28E-01	-	C	T	0.908	0.953	(0.828,1.097)	5.01E-01	-	C	T	0.909	1.013	(0.876,1.171)	8.63E-01	+
rs1808579	18	21104888	*C18orf8*	C	T	0.534	0.017	0.003	4.169E-08	C	T	0.522	1.096	(0.981,1.224)	1.05E-01	+	C	T	0.517	1.079	(0.994,1.172)	6.90E-02	+	C	T	0.513	1.026	(0.943,1.115)	5.53E-01	+
rs13201877	6	137675541	*IFNGR1*	G	A	0.142	0.023	0.005	4.285E-08	G	A	0.141	1.181	(1.006,1.385)	4.18E-02	+	G	A	0.141	1.091	(0.971,1.225)	1.43E-01	+	G	A	0.138	1.056	(0.932,1.196)	3.95E-01	+
rs2033732	8	85079709	*RALYL*	C	T	0.747	0.019	0.004	4.889E-08	C	T	0.743	1.008	(0.89,1.142)	8.95E-01	+	C	T	0.744	0.982	(0.895,1.078)	7.08E-01	-	C	T	0.744	1.015	(0.923,1.117)	7.62E-01	+
rs9540493	13	66205704	*MIR548X2*	A	G	0.456	0.017	0.003	4.971E-08	A	G	0.460	1.130	(1.005,1.27)	4.13E-02	+	A	G	0.454	1.120	(1.028,1.222)	9.92E-03	+	A	G	0.449	1.004	(0.92,1.096)	9.28E-01	+
rs1460676	2	164567689	*FIGN*	C	T	0.173	0.020	0.004	4.978E-08	C	T	0.158	1.022	(0.879,1.187)	7.81E-01	+	C	T	0.155	1.044	(0.934,1.168)	4.46E-01	+	C	T	0.154	0.983	(0.876,1.103)	7.66E-01	-
rs6465468	7	95169514	*ASB4*	T	G	0.304	0.017	0.004	4.98E-08	T	G	0.308	1.005	(0.887,1.139)	9.36E-01	+	T	G	0.301	1.047	(0.955,1.149)	3.24E-01	+	T	G	0.301	0.955	(0.868,1.049)	3.36E-01	-
rs751414**	6	40350030	TDRG1	T	G	0.258	0.018	0.004	1.58E-05	T	G	0.283	1.16813	(1.033,1.32)	1.29E-02	+	T	G	0.287	1.04676	(0.957,1.145)	3.18E-01	+	T	G	0.283	1.08231	(0.986,1.188)	9.67E-02	+

*GRCh37/hg19 coordinates
**Proxy for rs2033529
Effect = Effect allele (BMI increasing allele); Other = Other allele; EAF = Effect allele frequency

S4 Table. Difference in SCOOP OR when using ALSPAC as control dataset vs. UKHLS

SNP	Locus	OR.UKHLS	OR.ALSPAC	P.Diff
rs1558902	FTO	1.4287329	1.3427721	2.94E-01
rs6567160	MC4R	1.3080991	1.3604779	5.52E-01
rs13021737	TMEM18	1.3563998	1.2974696	6.00E-01
rs10938397	GNPDA2	1.1857281	1.1860919	9.96E-01
rs543874	SEC16B	1.2010834	1.2045657	9.67E-01
rs2207139	TFAP2B	1.1750903	1.1546588	8.18E-01
rs11030104	BDNF	1.1428476	1.1088972	6.90E-01
rs3101336	NEGR1	1.1948946	1.2385984	5.57E-01
rs7138803	BCDIN3D	1.215858	1.2146898	9.87E-01
rs10182181	ADCY3	1.2020002	1.2265576	7.31E-01
rs3888190	ATP2A1	1.1293237	1.0144525	7.22E-02
rs1516725	ETV5	1.158149	1.026153	1.74E-01
rs12446632	GPRC5B	1.0971063	1.0185721	3.86E-01
rs2287019	QPCTL	1.0244956	1.0421619	8.25E-01
rs16951275	MAP2K5	1.1325514	1.092782	6.20E-01
rs3817334	MTCH2	1.0927407	1.1358904	5.16E-01
rs2112347	POC5	1.0324305	1.004322	6.53E-01
rs12566985	FPGT-TNNI3K	1.2061603	1.1713434	6.23E-01
rs3810291	ZC3H4	1.1339907	1.0873902	5.08E-01
rs7141420	NRXN3	1.1124525	1.1058898	9.20E-01
rs13078960	CADM2	0.99411	1.031164	6.16E-01
rs10968576	LINGO2	1.048838	1.0523973	9.57E-01
rs17024393	GNAT2	1.5681554	1.5545372	9.58E-01
rs657452	AGBL4	1.0346724	1.0741845	5.39E-01
rs12429545	OLFM4	1.1186482	1.1316867	8.93E-01
rs12286929	CADM1	1.0687761	1.0658373	9.63E-01
rs13107325	SLC39A8	1.2837186	1.3563332	5.90E-01
rs11165643	PTBP2	1.0166116	1.0013239	8.01E-01
rs7903146	TCF7L2	1.049512	1.1068024	4.16E-01
rs10132280	STXBP6	0.9912485	0.9586591	6.05E-01
rs17405819	HNF4G	1.1236114	1.0863413	6.08E-01
rs6091540	ZFP64	1.067074	1.1034806	6.08E-01
rs1016287	LINC01122	1.0702148	1.0895905	7.78E-01
rs4256980	TRIM66	1.0129686	0.9606069	3.92E-01
rs17094222	HIF1AN	1.0431979	1.0554176	8.73E-01
rs12401738	FUBP1	1.0068534	0.9709964	5.52E-01
rs7599312	ERBB4	1.0466985	0.9901823	4.06E-01
rs2365389	FHIT	1.0918292	1.1451163	4.30E-01
rs205262	C6orf106	1.1634375	1.0784589	2.46E-01
rs2820292	NAV1	1.0305774	0.9731171	3.36E-01
rs12885454	PRKD1	1.0343118	0.9851811	4.27E-01
rs9641123	CALCR	1.0963951	1.0743197	7.35E-01
rs9581854	MTIF3	1.1523104	1.0643572	2.87E-01
rs16851483	RASA2	1.2018139	1.2290979	8.43E-01
rs11167827	HIP1	1.0745054	1.0968666	7.30E-01
rs758747	NLRC3	1.0100159	1.0528825	5.26E-01
rs1928295	TLR4	1.1026364	1.0470854	3.86E-01
rs9925964	KAT8	1.009594	1.0531275	4.94E-01
rs11126666	KCNK3	0.9916191	1.0011154	8.88E-01
rs2650492	SBK1	1.1745464	1.1002881	3.00E-01
rs6804842	RARB	1.0826074	1.0744722	9.00E-01
rs12940622	RPTOR	1.1210032	1.0859058	5.96E-01
rs7164727	LOC100287559	0.9919261	0.9667406	6.83E-01
rs11847697	PRKD1	1.2522288	1.1977594	7.40E-01
rs4740619	C9orf93	1.053687	1.0122709	4.98E-01
rs492400	USP37	1.0326143	1.0502736	7.75E-01
rs13191362	PARK2	1.0730103	1.1335706	5.49E-01
rs3736485	DMXL2	1.0839278	1.0843441	9.95E-01
rs17001654	SCARB2	1.0123387	0.9533294	4.70E-01
rs11191560	NT5C2	1.2358896	1.1978983	7.66E-01
rs2080454	CBLN1	0.9890632	0.9957939	9.11E-01
rs7715256	GALNT10	1.0667076	1.0659983	9.91E-01
rs2176040	LOC646736	1.0800962	1.0561891	7.15E-01
rs1528435	UBE2E3	1.0871075	1.0592421	6.71E-01
rs2075650	TOMM40	1.0017207	0.9436388	4.72E-01
rs1000940	RABEP1	1.1155025	1.1561709	5.73E-01
rs11583200	ELAVL4	1.0244068	1.0473277	7.14E-01
rs7239883	LOC284260	1.0288501	0.9941999	5.73E-01
rs2836754	ETS2	1.0526894	1.0507287	9.76E-01
rs9400239	FOXO3	1.1176888	1.0694002	5.06E-01
rs10733682	LMX1B	1.0712934	1.0545872	7.92E-01
rs11688816	EHBP1	0.988461	0.9758464	8.29E-01
rs11057405	CLIP1	1.1285696	1.0499994	4.62E-01
rs9914578	SMG6	1.0423268	1.0775537	6.46E-01
rs977747	TAL1	1.0432587	1.0166069	6.64E-01
rs2121279	LRP1B	1.1174631	1.0960597	8.23E-01
rs29941	KCTD15	1.1320951	1.0502216	2.39E-01
rs11727676	HHIP	1.0563982	1.0477135	9.38E-01
rs3849570	GBE1	1.0208987	0.9989013	7.25E-01
rs9374842	LOC285762	1.159775	1.1754326	8.52E-01
rs6477694	EPB41L4B	1.100504	1.0672617	6.17E-01
rs4787491	INO80E	1.0059218	1.0361099	6.17E-01
rs1441264	MIR548A2	1.0505803	1.0275484	7.19E-01
rs7899106	GRID1	1.240441	1.3269198	5.94E-01
rs2176598	HSD17B12	0.9573421	0.9531657	9.49E-01
rs2245368	PMS2L11	1.2247062	0.8928163	3.81E-05
rs17203016	CREB1	1.1327383	1.1718323	6.35E-01
rs17724992	PGPEP1	1.1553103	1.156706	9.86E-01
rs7243357	GRP	1.1056075	1.0651079	6.41E-01
rs16907751	ZBTB10	0.9527799	1.0182715	5.11E-01
rs1808579	C18orf8	1.0793057	1.0641407	8.11E-01
rs13201877	IFNGR1	1.0907763	1.0761121	8.71E-01
rs2033732	RALYL	0.9823586	0.9258369	3.80E-01
rs9540493	MIR548X2	1.1204384	1.0826008	5.72E-01
rs1460676	FIGN	1.0443083	1.0803401	6.74E-01
rs6465468	ASB4	1.0475326	0.949917	1.33E-01
rs2033529	TDRG1	0.9553109	0.9790073	7.05E-01

OR.UKHLS= OR when using UKHLS as control group
OR.ALSPAC= OR when using age-matched ALSPAC as control group
P.Diff=p value for difference

Appendix A

S5 Table. Discovery, replication and meta-analysis results for 32 SNPs meeting P<10-5 in discovery association results of SCOOP vs STILTS analysis.

						SCOOP				UKBB					GIANT BMI Tails				Combined analysis		
rsID	Nearest gene	Chr.	Position (bp)	EA	NEA	OR (95% CI)	P value	EAF Obese	EAF Thin	proxy rsID	OR (95% CI)	P value	EAF Obese	EAF Thin	proxy rsID	r2	OR (95% CI)	P value	OR (95% CI)	P value	HetPVal
rs9930333	FTO	16	53799977	G	T	1.70(1.52,1.90)	2.30E-20	49.59%	37.46%		1.46(1.38,1.55)	3.60E-36	48.26%	38.93%			1.43(1.34,1.54)	8.10E-25	1.48(1.42,1.54)	8.52E-76	3.34E-02
rs2168711	MC4R	18	57848531	C	T	1.66(1.45,1.89)	8.29E-14	28.90%	19.95%		1.23(1.15,1.32)	2.19E-09	26.75%	22.90%			1.20(1.10,1.30)	1.80E-05	1.27(1.21,1.33)	2.02E-21	1.12E-04
rs6748821	TMEM18	2	629601	G	A	1.65(1.42,1.91)	9.45E-11	86.69%	79.84%		1.27(1.18,1.37)	1.31E-09	85.00%	81.69%	rs12995480	0.998	1.26(1.14,1.39)	9.90E-06	1.32(1.24,1.39)	7.76E-21	2.81E-03
rs506589	SEC16B	1	177894287	C	T	1.46(1.27,1.67)	5.42E-08	23.98%	18.07%		1.25(1.17,1.35)	5.44E-10	23.11%	19.16%			1.25(1.14,1.37)	2.70E-06	1.28(1.21,1.35)	3.14E-20	1.21E-01
rs6738433	ADCY3-DNAJC27	2	25159501	C	G	1.43(1.28,1.60)	1.71E-10	47.31%	43.92%		1.21(1.14,1.28)	2.74E-10	50.70%	45.96%	rs2384054	0.968	1.10(1.03,1.17)	5.70E-03	1.19(1.14,1.24)	3.19E-17	6.25E-03
rs7132908	FAIM2	12	50263148	A	G	1.31(1.17,1.47)	2.26E-06	42.45%	36.27%		1.18(1.11,1.25)	5.43E-08	41.11%	37.39%			1.20(1.10,1.30)	6.60E-06	1.20(1.15,1.26)	1.93E-16	2.52E-01
rs62107261	FAM150B	2	422144	T	C	2.37(1.75,3.20)	2.07E-08	96.37%	93.38%		1.54(1.35,1.76)	3.57E-10	96.28%	94.36%	NA	NA	NA	NA	1.65(1.46,1.87)	1.15E-15	1.07E-02
rs12507026	GNPDA2	4	45181334	T	A	1.30(1.17,1.46)	3.69E-06	47.29%	40.92%		1.14(1.08,1.21)	8.76E-06	45.30%	41.98%	rs12641981	0.998	1.20(1.12,1.28)	3.10E-07	1.18(1.13,1.23)	5.53E-15	4.06E-02
rs75398113	SNRPC	6	34728071	C	G	1.53(1.27,1.85)	8.91E-06	11.95%	8.04%		1.24(1.12,1.37)	2.07E-05	10.47%	8.52%	NA	NA	NA	NA	1.30(1.19,1.42)	5.19E-09	5.56E-02
rs13135092	SLC39A8	4	103198082	G	A	1.58(1.30,1.93)	4.70E-06	10.50%	7.24%		1.25(1.12,1.39)	5.57E-05	9.24%	7.52%	NA	NA	NA	NA	1.32(1.20,1.45)	1.06E-08	3.59E-02
rs57988840	TFAP2B	6	50817748	T	A	1.69(1.39,2.05)	1.27E-07	92.53%	88.81%		1.13(1.02,1.24)	1.65E-02	91.05%	90.04%	rs7769978	1	1.20(1.03,1.39)	1.90E-02	1.22(1.13,1.31)	3.86E-07	2.87E-04
rs4447506	PIK3C3	18	39510074	G	A	1.32(1.17,1.48)	4.21E-06	41.83%	36.39%		1.07(1.01,1.14)	2.60E-02	39.34%	37.71%			1.10(1.02,1.18)	7.80E-03	1.11(1.06,1.16)	1.46E-06	7.85E-03
rs375252497*	SEMA3B	3	50310286	AAATAATAATAAT	A	1.35(1.20,1.53)	1.74E-06	37.22%	31.78%		1.13(1.02,1.26)	2.50E-02	34.30%	31.95%	NA	NA	NA	NA	1.22(1.13,1.32)	1.49E-06	3.05E-02
rs7927262	HIPK3	11	33384447	C	T	1.41(1.24,1.60)	1.81E-07	41.78%	35.78%		1.08(0.89,1.32)	4.43E-01	97.52%	97.37%	NA	NA	NA	NA	1.31(1.17,1.45)	1.58E-06	2.87E-02
rs654240	CCND1	11	69448373	T	C	1.35(1.21,1.52)	2.99E-07	43.85%	37.39%		1.05(0.99,1.12)	9.25E-02	41.43%	40.23%			1.08(1.00,1.16)	5.30E-02	1.10(1.05,1.15)	2.10E-05	6.81E-04
rs35403928*	PRDM6	5	122416569	GT	G	1.39(1.23,1.56)	6.79E-08	39.85%	32.94%		1.05(0.95,1.16)	3.61E-01	37.64%	36.49%	NA	NA	NA	NA	1.18(1.09,1.28)	2.46E-05	4.77E-04
rs516579	MTCL1	18	8801634	G	T	1.40(1.22,1.61)	2.07E-06	82.14%	77.25%		1.03(0.96,1.11)	4.52E-01	80.35%	80.05%	rs518561	0.998	1.15(1.04,1.27)	6.40E-03	1.11(1.05,1.18)	9.70E-05	1.11E-04
rs397859802*	FLJ26850	19	50556007	C	CA	1.92(1.45,2.53)	4.49E-06	6.02%	3.44%		1.11(0.86,1.44)	4.28E-01	4.25%	3.78%	NA	NA	NA	NA	1.43(1.18,1.73)	2.12E-04	4.77E-03
rs2784243	PKHD1	6	51454640	T	C	1.30(1.16,1.45)	5.99E-06	61.89%	56.06%		1.07(1.01,1.13)	2.90E-02	58.99%	57.34%	rs2784187	0.988	1.02(0.95,1.10)	5.40E-01	1.08(1.04,1.13)	3.14E-04	2.55E-03
rs11792928	LMX1B	9	129401550	T	C	1.36(1.20,1.53)	1.32E-06	32.13%	26.91%		1.05(0.99,1.12)	1.17E-01	29.94%	29.01%			1.03(0.95,1.11)	5.00E-01	1.08(1.03,1.13)	8.05E-04	5.19E-04
rs6711131*	BAZ2B	2	160407777	A	G	1.31(1.17,1.47)	4.30E-06	65.12%	58.62%		1.02(0.92,1.13)	6.81E-01	63.33%	63.04%	NA	NA	NA	NA	1.14(1.05,1.23)	8.90E-04	1.33E-03
rs73145387	ABI3BP	3	100813661	C	G	2.48(1.67,3.69)	7.36E-06	98.00%	96.42%		1.15(0.96,1.37)	1.29E-01	97.55%	97.19%	NA	NA	NA	NA	1.31(1.11,1.54)	1.29E-03	5.19E-04
rs599291	SLC44A5	1	75691616	T	C	1.31(1.17,1.47)	2.35E-06	47.71%	41.63%		1.02(0.96,1.08)	4.95E-01	44.55%	44.01%			1.04(0.97,1.11)	2.20E-01	1.06(1.02,1.11)	3.44E-03	4.01E-04
rs11185396	LOC100129138	1	104754536	C	T	1.50(1.26,1.80)	8.13E-06	12.78%	9.21%		1.06(0.97,1.17)	2.13E-01	10.37%	9.65%			1.01(0.89,1.14)	9.20E-01	1.10(1.03,1.18)	6.95E-03	8.13E-04
rs2836760	LOC400867	21	40300052	T	G	1.65(1.33,2.03)	3.28E-06	10.33%	7.12%		1.03(0.93,1.14)	5.92E-01	9.14%	8.91%			1.07(0.93,1.23)	3.50E-01	1.11(1.03,1.20)	9.44E-03	3.30E-04
rs11159277	SPTLC2	14	78932957	A	T	1.35(1.20,1.53)	1.56E-06	71.04%	66.32%		1.03(0.95,1.08)	6.53E-01	68.83%	68.55%	NA	NA	NA	NA	1.08(1.02,1.14)	9.74E-03	4.58E-05
rs10546790	CDH22	20	44910100	C	CAT	1.34(1.19,1.52)	1.91E-06	72.94%	66.87%	rs2425853	1.03(0.97,1.10)	3.42E-01	70.11%	69.59%	rs2425853	0.998	1.00(0.93,1.08)	9.90E-01	1.06(1.01,1.11)	1.35E-02	1.57E-04
rs11319985*	CNTN6	3	1377810	T	TA	1.29(1.15,1.45)	9.85E-06	61.56%	56.63%		0.97(0.88,1.07)	5.75E-01	57.91%	58.39%	NA	NA	NA	NA	1.10(1.02,1.18)	1.38E-02	2.03E-04
rs4790399	RAP1GAP2	17	2883199	C	T	1.33(1.18,1.51)	6.95E-06	74.28%	69.50%		1.02(0.96,1.09)	5.43E-01	71.22%	70.85%			0.99(0.91,1.08)	8.30E-01	1.05(1.00,1.10)	4.40E-02	2.50E-04
rs536093	PDE10A	6	165945641	T	C	1.38(1.22,1.58)	1.01E-06	27.05%	21.65%		0.97(0.90,1.03)	3.17E-01	24.39%	25.03%			1.06(0.97,1.15)	2.00E-01	1.05(1.00,1.10)	6.84E-02	9.29E-06
rs936249	CACNA1B	9	140971315	T	C	2.41(1.66,3.49)	3.81E-06	6.31%	4.63%		1.01(0.88,1.15)	9.30E-01	4.78%	4.77%	NA	NA	NA	NA	1.11(0.98,1.27)	9.53E-02	1.62E-05
rs1692144	GJA5	1	147281349	C	T	1.37(1.19,1.57)	8.19E-06	81.52%	77.06%		1.04(0.97,1.12)	2.90E-01	79.54%	79.06%			0.92(0.84,1.01)	7.00E-02	1.04(0.99,1.10)	1.29E-01	1.68E-05

***Interim release used in UKBB for these signals. Nobese=2,799. Nthin=1,212**

EA= Effect allele (BMI increasing allele); NEA= Non-effect allele; OR = Odds ratio; 95% CI = 95% confidence interval for the odds ratio; EAF = effect allele frequency. HetPVal= Heterozygosity p-value

Positions mapped to hg19

Blue line: Conventional genome-wide significant threshold (p<5E-08) in combined analysis.

S6 Table. Discovery, replication and meta-analysis results for 66 SNPs meeting P<10-5 in discovery association results of SCOOP vs UKHLS analysis.

						SCOOP				UKBB					SCOOP 2013				EGG				GIANT obesity class III				Combined analysis		
rsID	Nearest gene	Chr.	Position (bp)	EA	NEA	OR (95% CI)	P value	EAF Obese	EAF Non-extremes	proxy rsID	OR (95% CI)	P value	EAF Obese	EAF Non-extremes	proxy rsID	r2	OR (95% CI)	P value	proxy rsID	r2	OR (95% CI)	P value	proxy rsID	r2	OR (95% CI)	P value	OR (95% CI)	P value	HetPVal
rs9928094	FTO	16	53799905	G	A	1.44(1.33,1.57)	1.42E-18	49.50%	41.32%		1.30(1.25,1.35)	2.74E-41	48.34%	41.91%			1.46(1.34,1.60)	4.88E-17			1.21(1.15,1.28)	7.20E-13			1.43(1.34,1.54)	6.60E-25	1.32(1.29,1.36)	5.94E-101	4.41E-05
rs35614134	MC4R	18	57832856	AC	A	1.31(1.20,1.44)	6.27E-09	29.01%	23.69%	rs663129	1.22(1.16,1.27)	1.25E-18	26.72%	23.15%	rs663129	0.99814	1.32(1.19,1.46)	1.22E-07	rs663129	0.99814	1.22(1.15,1.30)	1.27E-10	rs663129	0.99814	1.20(1.10,1.30)	1.70E-05	1.23(1.20,1.27)	1.57E-43	3.55E-01
rs66906321	TMEM18	2	630070	C	T	1.40(1.24,1.57)	2.35E-08	85.78%	81.35%		1.17(1.11,1.24)	3.44E-09	84.44%	82.20%	rs13007080	0.95655	1.39(1.24,1.57)	6.65E-08	rs13007080	0.956548	1.28(1.19,1.37)	5.15E-12	rs13007080	0.956548	1.27(1.15,1.40)	3.40E-06	1.25(1.21,1.29)	9.72E-35	1.33E-02
rs7132908	FAIM2	12	50263148	A	G	1.22(1.12,1.32)	3.27E-06	42.45%	37.82%		1.15(1.10,1.19)	5.37E-12	41.11%	37.71%	rs7138803	0.90165	1.23(1.12,1.35)	8.89E-06			1.18(1.11,1.25)	1.24E-08			1.20(1.10,1.30)	6.60E-06	1.17(1.14,1.21)	2.38E-31	4.86E-01
rs2384060	ADCY3-DNAJC27	2	25135438	G	A	1.23(1.13,1.34)	1.53E-06	43.52%	38.90%	rs6722587	1.11(1.07,1.15)	4.89E-08	47.67%	44.93%	rs6722587	0.90829	1.09(1.00,1.19)	5.01E-02	rs6722587	0.908287	1.18(1.12,1.24)	1.02E-09	rs6722587	0.908287	1.12(1.04,1.19)	1.60E-03	1.14(1.11,1.17)	9.39E-23	1.13E-01
rs11209947	NEGR1	1	72808551	A	T	1.30(1.17,1.44)	8.51E-07	76.58%	72.18%	rs1460940	1.11(1.05,1.16)	4.53E-05	81.18%	79.76%	rs1460940	0.80203	1.46(1.30,1.63)	2.21E-10	rs1460940	0.802029	1.13(1.06,1.22)	4.60E-04	rs1460940	0.802029	1.22(1.11,1.35)	5.60E-05	1.17(1.13,1.21)	5.17E-20	7.26E-05
rs12735657	SEC16B	1	177809133	C	T	1.24(1.13,1.37)	9.72E-06	24.26%	20.46%		1.12(1.07,1.17)	1.48E-06	22.87%	20.94%	rs4132288	0.9914	1.20(1.07,1.33)	1.18E-03	rs4132288	0.991399	1.14(1.06,1.21)	1.52E-04	rs4132288	0.991399	1.22(1.11,1.34)	1.80E-05	1.15(1.12,1.19)	7.26E-19	1.79E-01
rs13104545	GNPDA2	4	45184907	A	G	1.27(1.15,1.40)	1.61E-06	27.41%	23.45%		1.07(1.02,1.11)	5.35E-03	24.36%	23.26%	NA	NA	NA	NA			1.13(1.04,1.22)	3.39E-03			1.34(1.20,1.49)	1.20E-07	1.13(1.09,1.17)	1.47E-11	9.39E-05
rs112446794	CEP120	5	122665465	T	C	1.23(1.13,1.35)	2.08E-06	33.15%	28.69%		1.07(1.02,1.11)	2.55E-03	29.47%	28.21%	rs1366333	0.95876	1.08(0.98,1.19)	1.38E-01	rs1366333	0.958762	1.12(1.06,1.18)	1.22E-04	rs1366333	0.958762	1.05(0.97,1.13)	2.40E-01	1.09(1.06,1.13)	3.45E-10	3.33E-02
rs3760091	SULT1A1	16	28620800	C	G	1.24(1.14,1.35)	1.56E-06	64.89%	60.23%		1.09(1.04,1.14)	1.19E-04	63.49%	61.44%	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1.12(1.07,1.16)	2.65E-08	8.49E-03
rs115474151	SLC7A14	3	170272967	A	T	1.36(1.20,1.55)	2.63E-06	12.44%	9.53%		1.27(0.86,1.87)	2.36E-01	0.26%	0.20%	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1.35(1.20,1.53)	1.40E-06	7.24E-01
rs8096590	LINC01541	18	69231235	A	G	1.24(1.14,1.35)	1.26E-06	34.39%	29.80%		1.03(0.99,1.07)	1.68E-01	31.28%	30.75%	rs7228473	0.98094	1.08(0.99,1.19)	9.70E-02	rs7228473	0.980939	1.02(0.97,1.08)	4.30E-01	rs7228473	0.980939	1.14(1.05,1.23)	1.40E-03	1.07(1.04,1.10)	7.38E-06	7.32E-04
rs201388971*	CDKAL1	6	20514945	T	TTG	1.33(1.17,1.50)	7.17E-06	85.90%	82.52%		1.10(1.01,1.20)	3.58E-02	84.15%	82.90%	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1.17(1.09,1.25)	1.77E-05	1.36E-02
rs141442356	FAT1	4	187723286	T	TG	1.23(1.13,1.35)	6.26E-06	42.33%	38.18%		0.97(0.76,1.24)	8.27E-01	99.39%	99.42%	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1.20(1.10,1.31)	2.98E-05	7.57E-02
rs10456655	PKHD1	6	51817204	G	C	1.30(1.16,1.44)	2.13E-06	20.35%	16.84%		1.06(1.01,1.12)	2.00E-02	17.45%	16.53%	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1.10(1.05,1.16)	3.34E-05	1.10E-03
rs7752955	TFAP2D	6	50402660	T	C	1.28(1.16,1.40)	6.46E-07	76.66%	72.16%		1.04(1.00,1.09)	6.33E-02	73.38%	72.61%			1.07(0.97,1.18)	1.89E-01			1.06(1.00,1.12)	6.05E-02			1.00(0.92,1.09)	9.50E-01	1.06(1.03,1.10)	3.90E-05	2.37E-03
rs898708	PNOC	8	28119194	C	T	1.26(1.15,1.38)	9.94E-07	73.10%	68.50%		1.02(0.98,1.07)	2.77E-01	69.52%	69.11%			1.06(0.96,1.17)	2.51E-01			1.04(0.98,1.10)	1.53E-01			1.08(0.99,1.16)	7.10E-02	1.06(1.03,1.09)	9.27E-05	2.38E-03
rs34208875	CASC17	17	69195603	C	T	2.95(1.90,4.59)	1.55E-06	1.33%	0.59%		1.18(0.90,1.55)	2.29E-01	0.52%	0.43%	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1.53(1.21,1.93)	3.78E-04	5.57E-04
rs540249707	METTL7B	12	56072632	G	A	3.60(2.13,6.08)	1.77E-06	1.03%	0.41%		1.17(0.78,1.75)	4.46E-01	0.23%	0.21%	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1.78(1.29,2.46)	4.25E-04	9.06E-04
rs10944524	MIR4643	6	91898609	T	C	1.28(1.15,1.43)	5.88E-06	18.12%	14.73%		1.05(0.99,1.10)	9.69E-02	15.62%	15.13%	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1.09(1.04,1.14)	4.94E-04	8.25E-04
rs60581051	ANKS1B	12	100147984	T	C	1.30(1.16,1.45)	7.81E-06	16.53%	13.40%		1.05(0.99,1.11)	8.71E-02	14.04%	13.55%	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1.09(1.04,1.15)	5.02E-04	1.01E-03
rs147725108	LINC00354	13	112579895	C	G	1.89(1.44,2.48)	4.66E-06	3.10%	1.92%		1.11(0.93,1.32)	2.61E-01	1.20%	1.07%	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1.30(1.12,1.51)	5.74E-04	1.27E-03
rs142450848	CARD18	11	105111566	C	T	2.34(1.69,3.25)	3.54E-07	2.08%	0.97%		1.10(0.92,1.33)	3.00E-01	1.07%	0.96%	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1.33(1.13,1.57)	5.89E-04	9.62E-05
rs5873242*	RANBP17	5	170321490	A	T	1.25(1.14,1.38)	2.73E-06	72.23%	67.87%		1.03(0.96,1.10)	4.25E-01	68.78%	68.25%	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1.10(1.04,1.17)	6.63E-04	8.90E-04
rs67529790*	LINC00977	8	130055514	AT	A	1.26(1.14,1.39)	2.78E-06	42.00%	37.87%		1.02(0.95,1.10)	5.13E-01	38.60%	38.10%	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1.10(1.04,1.17)	8.05E-04	8.21E-04
rs138019013	ALS2CL	3	46737713	T	G	3.92(2.29,6.69)	5.75E-07	1.01%	0.40%		0.97(0.58,1.60)	8.90E-01	0.14%	0.14%	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1.87(1.29,2.70)	8.55E-04	1.86E-04
rs28581396*	SLC38A11	2	165876796	C	T	2.14(1.54,2.98)	6.32E-06	2.01%	1.06%		1.07(0.80,1.42)	6.50E-01	1.42%	1.34%	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1.44(1.16,1.79)	9.60E-04	1.84E-03
rs144209184	SCD5	4	83646583	C	T	2.40(1.64,3.51)	7.09E-06	1.70%	0.88%		1.14(0.90,1.44)	2.74E-01	0.68%	0.59%	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1.40(1.15,1.71)	1.01E-03	1.15E-03
rs116931808*	CSMD1	8	4035481	C	G	2.29(1.59,3.29)	7.45E-06	1.60%	0.73%		1.01(0.70,1.47)	9.45E-01	0.72%	0.72%	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1.54(1.19,2.00)	1.10E-03	2.13E-03
rs11563327	HOXA1	7	27071888	C	T	1.24(1.13,1.36)	9.41E-06	74.64%	70.42%		1.03(0.99,1.08)	1.16E-01	71.80%	71.12%	NA	NA	NA	NA			1.03(0.97,1.09)	3.30E-01			1.03(0.95,1.11)	5.30E-01	1.05(1.02,1.08)	1.21E-03	4.77E-03
rs540119135	PTPRR	12	71094332	G	A	3.17(2.03,4.95)	4.21E-07	1.24%	0.45%		1.05(0.75,1.46)	7.90E-01	0.33%	0.33%	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1.55(1.19,2.03)	1.22E-03	9.42E-05
rs66671632	ARC	8	143680772	T	C	1.31(1.17,1.47)	2.62E-06	16.48%	13.14%		1.04(0.98,1.10)	2.32E-01	12.61%	12.26%	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1.09(1.03,1.15)	1.35E-03	2.70E-04
rs141214244*	RAB11FIP2	10	119652851	T	GTGTG	1.32(1.18,1.49)	3.45E-06	21.01%	17.88%		1.02(0.94,1.12)	5.92E-01	19.37%	18.93%	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1.12(1.04,1.20)	1.51E-03	5.94E-04
rs2237402	POU6F2	7	39449768	G	A	1.23(1.12,1.34)	6.31E-06	69.78%	65.38%		1.04(1.00,1.08)	7.72E-02	66.68%	65.95%	rs2237403	0.99506	1.02(0.93,1.12)	7.31E-01	rs2237403	0.995062	1.01(0.95,1.06)	7.75E-01	rs2237403	0.995062	1.04(0.96,1.11)	3.30E-01	1.04(1.02,1.07)	1.78E-03	5.14E-03
rs147345620	GPC6	13	94017255	G	A	3.02(1.89,4.82)	3.46E-06	1.25%	0.60%		1.12(0.85,1.47)	4.20E-01	0.51%	0.45%	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1.44(1.14,1.83)	2.37E-03	3.17E-04
rs11589523	FAM46C	1	118270980	T	C	1.22(1.12,1.32)	6.46E-06	36.66%	32.31%		1.02(0.98,1.07)	2.64E-01	33.42%	32.89%	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1.06(1.02,1.10)	3.26E-03	3.27E-04
rs556954774*	LOC100130992	10	22549699	C	CT	1.33(1.17,1.51)	9.52E-06	16.87%	14.13%		1.01(0.91,1.12)	8.35E-01	14.57%	14.53%	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1.12(1.04,1.22)	3.54E-03	8.29E-04
rs34515326*	MIR99AHG	21	17987843	CA	C	1.95(1.46,2.60)	6.33E-06	3.12%	2.02%		1.00(0.79,1.26)	9.72E-01	2.17%	2.16%	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1.30(1.09,1.57)	4.58E-03	4.39E-04
rs139198909	ZFHX3	16	72929809	C	A	2.50(1.68,3.73)	7.18E-06	1.50%	0.74%		1.08(0.85,1.36)	5.37E-01	0.68%	0.64%	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1.34(1.09,1.64)	5.10E-03	3.66E-04
rs191312158*	CCNY	10	35871176	C	T	2.76(1.80,4.23)	3.30E-06	1.42%	0.74%		0.96(0.69,1.35)	8.27E-01	0.98%	1.04%	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1.44(1.11,1.88)	6.68E-03	1.54E-04
6:153400217_AT_A	RGS17	6	153400217	AT	A	1.24(1.13,1.36)	6.04E-06	62.00%	57.94%		1.02(0.97,1.07)	4.13E-01	74.18%	73.83%	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1.06(1.02,1.10)	6.69E-03	1.98E-04
rs191634319	SOS1	2	39318135	C	G	3.54(2.13,5.88)	1.09E-06	1.04%	0.41%		1.05(0.78,1.40)	7.66E-01	0.43%	0.43%	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1.42(1.10,1.83)	7.11E-03	4.52E-05
rs62179502	SPATS2L	2	201146942	G	A	2.03(1.51,2.73)	2.31E-06	98.06%	96.28%		1.06(0.95,1.19)	3.12E-01	97.12%	96.98%	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1.15(1.04,1.28)	8.37E-03	5.18E-05
rs543354609	MDS2	1	23930694	C	T	2.79(1.78,4.37)	7.11E-06	1.20%	0.52%		1.04(0.80,1.36)	7.73E-01	0.51%	0.50%	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1.35(1.07,1.70)	1.04E-02	2.15E-04
rs144759478	PTH2R	2	209323283	T	C	1.64(1.34,2.02)	2.06E-06	5.13%	3.42%		1.02(0.92,1.14)	7.04E-01	3.29%	3.23%	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1.13(1.03,1.25)	1.10E-02	5.66E-05
rs6496033	LOC440311	15	95245915	A	C	1.47(1.25,1.73)	4.82E-06	8.81%	6.65%		1.00(0.93,1.08)	9.73E-01	7.37%	7.35%	NA	NA	NA	NA			1.07(0.96,1.19)	2.48E-01			1.19(0.96,1.47)	1.10E-01	1.07(1.02,1.13)	1.16E-02	3.80E-04
rs1571570	PBX3	9	128748249	C	G	1.41(1.22,1.63)	3.08E-06	9.45%	6.91%		1.08(1.00,1.16)	4.00E-02	7.33%	6.82%	NA	NA	NA	NA			0.97(0.89,1.06)	5.17E-01			0.96(0.83,1.12)	6.30E-01	1.07(1.01,1.12)	1.28E-02	1.40E-04
rs28700201*	TP53TG3D	16	32275881	A	G	1.33(1.18,1.50)	2.04E-06	49.33%	45.88%		0.96(0.88,1.06)	4.72E-01	54.34%	54.74%	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1.10(1.02,1.18)	1.36E-02	3.76E-05
rs571984391*	LMX1B	9	129414279	TG	T	1.22(1.12,1.33)	6.58E-06	44.94%	40.48%		0.99(0.93,1.06)	7.89E-01	42.42%	42.68%	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1.07(1.01,1.12)	1.54E-02	1.39E-04
rs524349	UNC13C	15	54689366	A	G	1.29(1.17,1.41)	8.01E-08	73.26%	67.99%		1.01(0.97,1.05)	7.10E-01	69.42%	69.20%	rs563881	0.9026	1.00(0.91,1.11)	9.41E-01	rs563881	0.902595	1.02(0.96,1.08)	5.48E-01	rs563881	0.902595	1.02(0.95,1.11)	5.80E-01	1.04(1.01,1.07)	1.64E-02	8.42E-05
rs144435735	LINC00682	4	41881253	A	G	1.98(1.49,2.61)	1.96E-06	3.26%	2.06%		0.95(0.80,1.13)	5.51E-01	1.25%	1.33%	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1.16(1.00,1.34)	4.81E-02	1.23E-05
rs73676140	NXPH1	7	8729124	G	A	1.59(1.29,1.95)	9.74E-06	4.75%	3.16%		0.99(0.88,1.10)	7.95E-01	3.07%	3.12%	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1.10(1.00,1.21)	6.06E-02	5.89E-05
rs139813768	SNX5	20	17828396	A	G	2.47(1.66,3.67)	8.32E-06	1.57%	0.83%		0.94(0.75,1.19)	6.21E-01	0.65%	0.68%	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1.21(0.99,1.48)	6.46E-02	4.32E-05
rs6806545	LOC647323	3	193660056	T	C	1.21(1.11,1.32)	9.08E-06	41.15%	36.68%		0.99(0.95,1.03)	6.64E-01	37.72%	37.84%	NA	NA	NA	NA			1.03(0.98,1.09)	2.74E-01			1.00(0.94,1.07)	9.20E-01	1.02(1.00,1.05)	8.16E-02	4.24E-04
rs146883791	DCHS2	4	155121474	A	ATAAG	1.74(1.37,2.21)	5.48E-06	3.42%	2.02%	rs7680244	0.94(0.83,1.07)	3.43E-01	2.26%	2.41%	rs7680244	0.99439	1.43(1.04,1.96)	2.95E-02	rs7680244	0.994386	0.99(0.80,1.23)	9.25E-01	NA	NA	NA	NA	1.09(0.99,1.20)	8.79E-02	3.30E-05
rs9863931	ADAMTS9-AS2	3	65014778	T	C	1.31(1.17,1.47)	3.89E-06	15.79%	12.47%		1.01(0.96,1.07)	6.52E-01	12.40%	12.17%	NA	NA	NA	NA			0.98(0.91,1.06)	6.73E-01			0.99(0.88,1.12)	9.20E-01	1.04(0.99,1.08)	9.31E-02	2.84E-04
rs7069309	CAMK1D	10	12351089	T	C	1.20(1.11,1.30)	5.59E-06	49.24%	45.94%		1.01(0.97,1.05)	5.94E-01	50.11%	49.66%	NA	NA	NA	NA			0.99(0.94,1.05)	7.17E-01			0.99(0.92,1.07)	8.10E-01	1.02(1.00,1.05)	9.91E-02	3.72E-04
rs79325679	LARGE	22	33929898	T	C	2.07(1.55,2.76)	7.71E-07	2.50%	1.26%		0.93(0.80,1.09)	3.70E-01	1.47%	1.59%	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1.12(0.97,1.28)	1.16E-01	1.79E-06
rs11872477		18	2037477	T	C	1.26(1.14,1.39)	5.92E-06	37.93%	34.04%		0.99(0.94,1.03)	6.07E-01	32.52%	32.80%	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1.03(0.99,1.08)	1.44E-01	1.55E-05
rs190999252	CCSER1	4	92847036	A	G	2.07(1.51,2.84)	7.12E-06	2.25%	1.27%		0.93(0.79,1.11)	4.34E-01	1.26%	1.34%	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1.12(0.96,1.30)	1.52E-01	1.49E-05
rs75809547	PTBP2	1	97331151	C	T	2.61(1.74,3.93)	3.85E-06	1.46%	0.65%		0.86(0.68,1.09)	2.21E-01	0.62%	0.72%	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1.14(0.93,1.41)	2.00E-01	4.10E-06
rs147842378	AGO4	1	36276202	C	G	2.01(1.49,2.71)	5.08E-06	3.16%	2.10%		0.91(0.78,1.07)	2.67E-01	1.48%	1.63%	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	1.08(0.94,1.25)	2.54E-01	5.19E-06
11:23234822_ATGG_A	MIR8054	11	23234822	ATGG	A	1.22(1.13,1.33)	2.68E-06	43.06%	38.46%	rs2403819	0.99(0.96,1.03)	7.41E-01	39.17%	39.33%	rs2403819	0.97568	1.01(0.92,1.11)	8.49E-01	rs2403819	0.975677	0.96(0.91,1.01)	1.04E-01	rs2403819	0.975677	1.03(0.96,1.10)	4.00E-01	1.01(0.98,1.04)	4.19E-01	5.19E-05
rs12261064	GPR26	10	125429706	A	G	1.55(1.30,1.86)	1.48E-06	6.21%	4.14%		0.90(0.82,0.99)	2.71E-02	4.32%	4.72%			0.86(0.69,1.05)	1.44E-01			0.92(0.82,1.04)	2.00E-01			1.05(0.86,1.28)	6.20E-01	1.02(0.96,1.09)	4.68E-01	2.40E-06
rs1188854	ECT2L	6	139196524	G	A	1.39(1.21,1.61)	6.57E-06	10.13%	7.69%		0.97(0.90,1.03)	3.14E-01	8.20%	8.43%	NA	NA	NA	NA			1.00(0.91,1.10)	9.58E-01			0.94(0.81,1.11)	4.80E-01	1.02(0.97,1.07)	5.48E-01	8.42E-05
rs3104056	COL13A1	10	71510039	A	G	1.53(1.27,1.84)	6.31E-06	5.64%	3.78%		0.85(0.76,0.94)	1.16E-03	3.55%	4.09%			1.05(0.83,1.32)	6.83E-01			0.98(0.84,1.14)	8.09E-01			1.20(0.93,1.54)	1.60E-01	1.00(0.94,1.08)	9.08E-01	1.17E-06

***Interim release used in UKBB for these signals. Nobese=2,799. Ncontrols=8,193**

EA= Effect allele (BMI increasing allele); NEA= Non-effect allele; OR = Odds ratio; 95% CI = 95% confidence interval for the odds ratio; EAF = effect allele frequency. HetPVal= Heterozygosity p-value

Positions mapped to hg19

S7 Table. Discovery, replication and meta-analysis results for 37 SNPs meeting P<10-5 in discovery association results of UKHLS vs STILTS analysis.

						STILTS				UKBB				Combined analysis		
rsID	Nearest gene	Chr.	Position (bp)	EA	NEA	OR (95% CI)	P value	EAF Non-extremes	EAF Thin	OR (95% CI)	P value	EAF Non-extremes	EAF Thin	OR (95% CI)	P value	HetPVal
rs13262703	PI15	8	75819902	A	T	4.15 (2.42,7.11)	2.32E-07	99.62%	98.88%	1.69(0.98,2.91)	5.68E-02	99.84%	99.74%	2.66(1.81,3.89)	5.46E-07	2.15E-02
rs558258836*	LOC102724874	8	78716821	T	A	4.04 (2.25,7.26)	3.07E-06	99.60%	99.04%	1.69 (0.89,3.2)	1.07E-01	99.50%	99.28%	2.71 (1.76,4.18)	5.99E-06	4.90E-02
rs2123163	CADM2	3	85243797	T	G	1.68 (1.35,2.1)	4.90E-06	6.40%	4.20%	1.14(1.00,1.29)	4.22E-02	5.12%	4.60%	1.25(1.12,1.39)	6.18E-05	2.76E-03
rs150756788	SLC2A7	1	9050295	G	T	2.09 (1.52,2.87)	4.96E-06	98.21%	97.25%	1.20(0.91,1.60)	2.00E-01	99.26%	99.11%	1.54(1.25,1.90)	6.41E-05	1.07E-02
rs545797179*	FOXN2	2	48546924	AT	A	2.79 (1.8,4.32)	4.31E-06	99.24%	98.48%	1.15 (0.65,2.05)	6.23E-01	99.18%	99.07%	2.02 (1.42,2.86)	7.77E-05	1.65E-02
rs117638949*	PIGZ	3	196692722	T	A	3.5 (2.27,5.4)	1.50E-08	99.50%	98.55%	0.54 (0.27,1.09)	8.60E-02	99.30%	99.62%	2.09 (1.44,3.02)	9.25E-05	8.97E-06
rs576762972*	CACNA1C	12	2244717	T	C	2.17 (1.55,3.05)	7.23E-06	98.99%	98.03%	1.16 (0.78,1.72)	4.69E-01	98.87%	98.73%	1.66 (1.29,2.15)	1.05E-04	1.79E-02
rs138454709*	COL8A2	1	36592131	A	G	2.58 (1.72,3.88)	5.29E-06	99.03%	98.33%	1.11 (0.65,1.9)	7.04E-01	99.04%	99.00%	1.89 (1.37,2.62)	1.17E-04	1.41E-02
rs75937976	C3orf38	3	88321976	G	C	2.95 (2.02,4.32)	2.43E-08	99.20%	98.25%	1.10(0.84,1.44)	4.96E-01	99.13%	99.05%	1.53(1.23,1.91)	1.52E-04	3.33E-05
rs190051670	PHF2	9	96460947	C	T	2.55 (1.73,3.76)	2.11E-06	99.25%	98.35%	1.19(0.91,1.56)	2.00E-01	99.18%	99.05%	1.53(1.23,1.91)	1.68E-04	1.59E-03
rs56152157	EDIL3	5	83171742	G	A	1.21 (1.11,1.31)	6.91E-06	47.95%	42.99%	1.04(0.99,1.10)	1.21E-01	47.37%	46.44%	1.09(1.04,1.14)	2.11E-04	2.85E-03
rs139226692*	ASAH1	8	17928720	C	CA	2.82 (1.78,4.46)	9.46E-06	99.56%	98.81%	1.11 (0.63,1.92)	7.24E-01	99.37%	99.34%	1.93 (1.35,2.75)	2.73E-04	1.08E-02
rs112958625*	KNDC1	10	134969737	G	A	2.8 (1.81,4.33)	3.61E-06	99.00%	98.39%	1.01 (0.59,1.72)	9.73E-01	98.93%	98.94%	1.86 (1.33,2.62)	3.02E-04	3.75E-03
rs68090520*	ZMAT3	3	178717361	C	A	1.24 (1.13,1.36)	4.04E-06	54.37%	50.43%	1.02 (0.93,1.11)	7.45E-01	53.49%	53.13%	1.12 (1.05,1.2)	4.39E-04	2.72E-03
rs17544568	ONECUT1	15	53321119	G	A	2.04 (1.54,2.7)	6.53E-07	97.94%	96.67%	1.09(0.93,1.29)	2.78E-01	97.58%	97.40%	1.28(1.11,1.47)	5.80E-04	1.74E-04
rs143866745*	LOC101927495	11	61356693	C	T	1.31 (1.17,1.46)	1.26E-06	60.23%	56.57%	0.89 (0.51,1.55)	6.88E-01	99.16%	99.21%	1.84 (1.29,2.61)	6.65E-04	9.33E-04
rs184273748*	PTPRU	1	29562801	G	A	2.53 (1.71,3.73)	3.04E-06	99.12%	98.24%	0.65 (0.33,1.26)	2.00E-01	99.17%	99.35%	1.79 (1.28,2.5)	7.11E-04	5.41E-04
rs115861768	MIR4426	16	60885992	C	T	3.27 (1.93,5.52)	9.57E-06	99.53%	98.98%	1.14(0.73,1.76)	5.68E-01	99.65%	99.64%	1.76(1.25,2.46)	1.04E-03	2.46E-03
rs191980904*	UQCRC2	16	21946517	C	T	2.98 (1.85,4.79)	6.69E-06	99.56%	98.93%	0.84 (0.46,1.55)	5.84E-01	99.36%	99.44%	1.85 (1.27,2.7)	1.28E-03	1.39E-03
rs11665052	MC4R	18	57908675	G	A	1.31 (1.18,1.44)	1.40E-07	27.11%	22.61%	1.02(0.96,1.08)	5.63E-01	26.22%	25.78%	1.09(1.03,1.14)	1.48E-03	2.22E-05
2:25411587_C_CT*	POMC	2	25411587	C	CT	1.36 (1.21,1.51)	6.52E-08	83.75%	79.76%	1.01(0.95,1.08)	7.22E-01	82.18%	82.10%	1.10(1.04,1.16)	1.76E-03	9.91E-06
rs137887309	CDH23	10	73211425	G	A	2.66 (1.74,4.07)	6.24E-06	99.48%	98.66%	1.03(0.71,1.50)	8.67E-01	99.54%	99.50%	1.56(1.18,2.06)	1.94E-03	9.94E-04
rs11757467	EYA4	6	133808153	A	T	1.71 (1.35,2.17)	8.55E-06	97.57%	96.11%	1.05(0.90,1.24)	5.26E-01	97.48%	97.33%	1.23(1.08,1.40)	2.43E-03	9.08E-04
rs148209625	ZNF664-FAM101A	12	124681051	C	T	2.2 (1.58,3.07)	2.97E-06	99.02%	97.95%	1.04(0.82,1.32)	7.63E-01	98.89%	98.85%	1.35(1.11,1.64)	2.74E-03	3.19E-04
rs71515311*	TMEM72-AS1	10	45116672	A	ATAT	1.25 (1.13,1.38)	8.55E-06	70.65%	66.71%	0.99 (0.89,1.09)	7.72E-01	69.59%	70.05%	1.11 (1.04,1.19)	2.84E-03	9.00E-04
rs11557769	ACTN1	14	69341653	T	A	1.95 (1.5,2.55)	8.61E-07	98.33%	96.94%	0.91(0.70,1.17)	4.48E-01	98.83%	98.98%	1.31(1.09,1.57)	4.41E-03	4.45E-05
rs142441937	KLF15	3	126030681	G	A	2.53 (1.69,3.79)	6.78E-06	99.16%	98.44%	1.02(0.78,1.34)	8.75E-01	99.08%	99.11%	1.36(1.08,1.70)	8.09E-03	2.71E-04
rs117944743	ZNF93	19	20060211	C	G	2.06 (1.5,2.82)	7.33E-06	98.81%	97.84%	1.01(0.82,1.25)	9.34E-01	98.49%	98.47%	1.26(1.05,1.50)	1.06E-02	2.28E-04
rs142425331	CHCHD3	7	132583478	G	A	1.6 (1.31,1.96)	4.36E-06	96.58%	94.82%	0.98(0.84,1.13)	7.44E-01	96.61%	96.71%	1.16(1.03,1.30)	1.59E-02	8.80E-05
rs138251346	LOC101929452	2	7279064	A	G	2.99 (1.9,4.7)	2.23E-06	99.27%	98.62%	0.77(0.50,1.19)	2.42E-01	99.53%	99.64%	1.46(1.07,2.00)	1.66E-02	2.20E-05
rs1435711	ADAMTS20	12	43429113	G	A	1.32 (1.18,1.49)	2.11E-06	86.34%	83.30%	0.99(0.92,1.07)	8.17E-01	85.74%	85.93%	1.08(1.01,1.15)	1.76E-02	3.92E-05
rs553440779	KCNJ3	2	155835504	T	C	2.67 (1.74,4.09)	6.98E-06	99.27%	98.53%	0.72(0.45,1.15)	1.66E-01	99.61%	99.70%	1.46(1.07,2.00)	1.78E-02	4.88E-05
rs77709566	INTU	4	128466995	A	G	2 (1.5,2.66)	2.01E-06	98.42%	97.20%	0.97(0.81,1.17)	7.51E-01	97.98%	98.03%	1.20(1.03,1.40)	2.02E-02	3.20E-05
rs514529	LRP5	11	68090836	T	A	1.23 (1.13,1.34)	1.09E-06	53.61%	51.60%	0.99(0.94,1.05)	7.62E-01	52.12%	52.27%	1.05(1.01,1.10)	2.04E-02	1.68E-05
rs200275909*	ADAMTS20	12	43954570	A	AT	3.21 (1.93,5.35)	7.29E-06	99.42%	98.76%	0.88 (0.78,1)	5.84E-02	85.23%	86.54%	1.1 (1.01,1.2)	2.38E-02	4.69E-06
rs73085383	ZNF343	20	2503465	C	T	2.13 (1.56,2.92)	2.05E-06	98.63%	97.60%	0.85(0.66,1.10)	2.28E-01	98.79%	98.97%	1.23(1.01,1.50)	3.92E-02	8.79E-06
rs527595266	ADAMTS16	5	5341419	C	G	2.91 (1.81,4.68)	9.95E-06	99.42%	98.79%	0.93(0.68,1.27)	6.52E-01	99.30%	99.30%	1.31(1.01,1.71)	4.00E-02	8.21E-05

***Interim release used in UKBB for these signals. Nthin=1,212. Ncontrols=8,193**

****rs4665779 was used as a proxy in UKBB**

EA= Effect allele (BMI increasing allele); NEA= Non-effect allele; OR = Odds ratio; 95% CI = 95% confidence interval for the odds ratio; EAF = effect allele frequency. HetPVal= Heterozygosity p-value

Positions mapped to hg19

Red line: Strict genome-wide significant threshold (p<1.17E-08) in combined analysis. Blue line: Conventional genome-wide significant threshold (p<5E-08) in combined analysis.

S10 Table. Published loci from GIANT, EGG and SCOOP 2013 not reaching genome-wide significance in our study

Known BMI loci with meta p <5E-8 in GIANT BMI tails study but not in this study (obese vs thin)

rsID	Gene	OR GIANT BMI tails Stage 1	P GIANT BMI tails Stage 1	OR SCOOP/STILTS	P SCOOP/STILTS	OR SCOOP/UKHLS	P SCOOP/UKHLS	OR UKHLS/STILTS	P UKHLS/STILTS
rs2568958	NEGR1	1.17 (1.12,1.23)	6.80E-10	1.25 (1.11,1.39)	1.00E-04	1.19(1.09,1.29)	5.65E-05	1.06(0.97,1.16)	1.73E-01
rs987237	TFAP2B	1.20 (1.12,1.28)	4.30E-07	1.31 (1.14,1.50)	2.00E-04	1.17(1.05,1.29)	3.25E-03	1.14(1.01,1.27)	2.72E-02
rs2030323	BDNF	1.21 (1.13,1.30)	5.20E-08	1.31 (1.13,1.50)	7.46E-05	1.15(1.03,1.27)	1.03E-02	1.10(1.00,1.22)	4.92E-02
rs1516725	ETV5	1.30 (1.19, 1.42)	2.10E-08	1.30 (1.11,1.52)	8.00E-04	1.16(1.02,1.31)	1.89E-02	1.18(1.05,1.33)	5.03E-03

Loci identified in S.I. Berndt, et al. (2013)

rsID	Gene	OR GIANT Stage 1	P GIANT Stage 1	OR SCOOP/STILTS	P SCOOP/STILTS	OR SCOOP/UKHLS	P SCOOP/UKHLS	OR UKHLS/STILTS	P UKHLS/STILTS	Reported Trait
rs7989336	HS6ST3	1.12	5.88E-09	1.13(1.01,1.26)	3.17E-02	1.03(0.95,1.12)	4.42E-01	1.09(1.00,1.19)	4.15E-02	Obesity class 2
rs17381664	ZZZ3	1.11	7.61E-08	1.00(0.89,1.12)	9.86E-01	0.98(0.91,1.07)	6.99E-01	1.03(0.95,1.12)	4.82E-01	Obesity class 2
rs17024258	GNAT2	1.23	1.41E-06	1.80(1.29,2.53)	6.27E-04	1.57(1.25,1.97)	1.18E-04	1.10(0.82,1.46)	5.32E-01	Obesity class 1
rs4735692	HNF4G	1.07	5.03E-08	1.08(0.97,1.21)	1.57E-01	1.00(0.92,1.09)	9.87E-01	1.07(0.98,1.16)	1.27E-01	Obesity class 1
rs13041126	MRPS33P4	1.07	3.05E-07	1.14(1.01,1.28)	3.88E-02	1.07(0.97,1.17)	1.71E-01	1.03(0.94,1.13)	5.43E-01	Obesity class 1
rs2531995	ADCY9	1.06	3.17E-06	1.14(1.01,1.28)	3.22E-02	1.06(0.97,1.16)	2.06E-01	1.08(0.98,1.18)	1.04E-01	Obesity class 1
rs4735692	HNF4G	1.05	6.13E-09	1.08(0.97,1.21)	1.57E-01	1.00(0.92,1.09)	9.87E-01	1.07(0.98,1.16)	1.27E-01	Overweight
rs7503807	RPTOR	1.04	4.20E-06	1.18(1.06,1.32)	2.90E-03	1.11(1.03,1.21)	1.04E-02	1.07(0.98,1.16)	1.24E-01	Overweight

Loci indentified in J.P. Bradfield, H.R. Taal, et al. (2012)

rsID	Gene	OR EGG Stage 1	P EGG Stage 1	OR SCOOP/STILTS	P SCOOP/STILTS	OR SCOOP/UKHLS	P SCOOP/UKHLS	OR UKHLS/STILTS	P UKHLS/STILTS
rs9568856	OLFM4	1.21	6.58E−7	1.09(0.93,1.28)	2.71E-01	1.14(1.01,1.28)	2.99E-02	0.97(0.86,1.10)	6.41E-01
rs9299	HOXB5	1.14	9.12E−7	1.18(1.05,1.32)	6.46E-03	1.03(0.94,1.12)	5.68E-01	1.09(1.00,1.19)	5.34E-02

Loci identified in E. Wheeler, et al. (2013)

rsID	Gene	OR SCOOP 2013 Stage 1	P SCOOP 2013 Stage 1	OR SCOOP/STILTS	P SCOOP/STILTS	OR SCOOP/UKHLS	P SCOOP/UKHLS	OR UKHLS/STILTS	P UKHLS/STILTS
rs1993709	NEGR1	1.46	1.98E-12	1.30(1.13,1.50)	2.54E-04	1.29(1.16,1.44)	4.45E-06	1.03(0.93,1.14)	6.16E-01
rs1957894	PRKCH	1.64	1.01E-08	1.25(1.03,1.51)	2.61E-02	1.17(1.02,1.35)	2.40E-02	1.01(0.87,1.18)	8.79E-01
rs11208659	LEPR	1.63	1.16E-10	1.22(1.01,1.48)	4.33E-02	1.28(1.12,1.48)	4.35E-04	0.95(0.81,1.10)	4.90E-01
rs564343	PACS1	1.25	5.81E-08	1.01(0.90,1.13)	9.18E-01	1.04(0.95,1.13)	4.12E-01	0.96(0.89,1.05)	4.12E-01
rs11109072	RMST	1.79	1.48E-07	0.87(0.63,1.20)	3.83E-01	0.95(0.74,1.21)	6.74E-01	0.97(0.76,1.24)	8.13E-01

S12 Table. ICD10 codes used to exclude thin individuals in UKBB

Code	ICD10 Code Description
A071	A07.1 Giardiasis [lambliasis]
A150	A15.0 Tuberculosis of lung, confirmed by sputum microscopy with or without culture
A151	A15.1 Tuberculosis of lung, confirmed by culture only
A152	A15.2 Tuberculosis of lung, confirmed histologically
A159	A15.9 Respiratory tuberculosis unspecified, confirmed bacteriologically and histologically
A162	A16.2 Tuberculosis of lung, without mention of bacteriological or histological confirmation
A169	A16.9 Respiratory tuberculosis unspecified, without mention of bacteriological or histological confirmation
B181	B18.1 Chronic viral hepatitis B without delta-agent
B182	B18.2 Chronic viral hepatitis C
B203	B20.3 HIV disease resulting in other viral infections
B204	B20.4 HIV disease resulting in candidiasis
B238	B23.8 HIV disease resulting in other specified conditions
B24	B24 Unspecified human immunodeficiency virus [HIV] disease
C01	C01 Malignant neoplasm of base of tongue
C099	C09.9 Tonsil, unspecified
C108	C10.8 Overlapping lesion of oropharynx
C109	C10.9 Oropharynx, unspecified
C155	C15.5 Lower third of oesophagus
C159	C15.9 Oesophagus, unspecified
C169	C16.9 Stomach, unspecified
C172	C17.2 Ileum
C180	C18.0 Caecum
C182	C18.2 Ascending colon
C184	C18.4 Transverse colon
C187	C18.7 Sigmoid colon
C189	C18.9 Colon, unspecified
C20	C20 Malignant neoplasm of rectum
C210	C21.0 Anus, unspecified
C211	C21.1 Anal canal
C220	C22.0 Liver cell carcinoma
C221	C22.1 Intrahepatic bile duct carcinoma
C250	C25.0 Head of pancreas
C258	C25.8 Overlapping lesion of pancreas
C259	C25.9 Pancreas, unspecified
C341	C34.1 Upper lobe, bronchus or lung
C342	C34.2 Middle lobe, bronchus or lung
C343	C34.3 Lower lobe, bronchus or lung
C349	C34.9 Bronchus or lung, unspecified
C411	C41.1 Mandible
C435	C43.5 Malignant melanoma of trunk
C436	C43.6 Malignant melanoma of upper limb, including shoulder
C437	C43.7 Malignant melanoma of lower limb, including hip
C439	C43.9 Malignant melanoma of skin, unspecified
C441	C44.1 Skin of eyelid, including canthus
C442	C44.2 Skin of ear and external auricular canal
C443	C44.3 Skin of other and unspecified parts of face
C444	C44.4 Skin of scalp and neck
C445	C44.5 Skin of trunk
C446	C44.6 Skin of upper limb, including shoulder
C447	C44.7 Skin of lower limb. including hip
C482	C48.2 Peritoneum, unspecified
C503	C50.3 Lower-inner quadrant of breast
C504	C50.4 Upper-outer quadrant of breast
C505	C50.5 Lower-outer quadrant of breast
C509	C50.9 Breast, unspecified
C541	C54.1 Endometrium
C56	C56 Malignant neoplasm of ovary
C61	C61 Malignant neoplasm of prostate
C64	C64 Malignant neoplasm of kidney, except renal pelvis
C66	C66 Malignant neoplasm of ureter
C679	C67.9 Bladder, unspecified
C719	C71.9 Brain, unspecified
C73	C73 Malignant neoplasm of thyroid gland
C770	C77.0 Lymph nodes of head, face and neck
C771	C77.1 Intrathoracic lymph nodes
C772	C77.2 Intra-abdominal lymph nodes
C773	C77.3 Axillary and upper limb lymph nodes
C779	C77.9 Lymph node, unspecified
C780	C78.0 Secondary malignant neoplasm of lung
C786	C78.6 Secondary malignant neoplasm of retroperitoneum and peritoneum
C787	C78.7 Secondary malignant neoplasm of liver
C788	C78.8 Secondary malignant neoplasm of other and unspecified digestive organs
C793	C79.3 Secondary malignant neoplasm of brain and cerebral meninges
C795	C79.5 Secondary malignant neoplasm of bone and bone marrow
C80	C80 Malignant neoplasm without specification of site
C829	C82.9 Follicular non-Hodgkin's lymphoma, unspecified
C844	C84.4 Peripheral T-cell lymphoma
C859	C85.9 Non-Hodgkin's lymphoma, unspecified type
C880	C88.0 Waldenstr m's macroglobulinaemia
C910	C91.0 Acute lymphoblastic leukaemia
C920	C92.0 Acute myeloid leukaemia
D001	D00.1 Oesophagus
D013	D01.3 Anus and anal canal
D033	D03.3 Melanoma in situ of other and unspecified parts of face
D371	D37.1 Stomach
D374	D37.4 Colon
D375	D37.5 Rectum
D377	D37.7 Other digestive organs
D381	D38.1 Trachea, bronchus and lung
D391	D39.1 Ovary
D400	D40.0 Prostate
D430	D43.0 Brain, supratentorial
D432	D43.2 Brain, unspecified
D443	D44.3 Pituitary gland
D45	D45 Polycythaemia vera
D469	D46.9 Myelodysplastic syndrome, unspecified
D471	D47.1 Chronic myeloproliferative disease
D477	D47.7 Other specified neoplasms of uncertain or unknown behaviour of lymphoid, haematopoietic and related tissue
D481	D48.1 Connective and other soft tissue
D485	D48.5 Skin
D486	D48.6 Breast
D500	D50.0 Iron deficiency anaemia secondary to blood loss (chronic)
D508	D50.8 Other iron deficiency anaemias
D509	D50.9 Iron deficiency anaemia, unspecified
D510	D51.0 Vitamin B12 deficiency anaemia due to intrinsic factor deficiency
D519	D51.9 Vitamin B12 deficiency anaemia, unspecified
D529	D52.9 Folate deficiency anaemia, unspecified
D539	D53.9 Nutritional anaemia, unspecified
E039	E03.9 Hypothyroidism, unspecified
E041	E04.1 Non-toxic single thyroid nodule
E042	E04.2 Non-toxic multinodular goitre
E049	E04.9 Non-toxic goitre, unspecified
E050	E05.0 Thyrotoxicosis with diffuse goitre
E052	E05.2 Thyrotoxicosis with toxic multinodular goitre
E059	E05.9 Thyrotoxicosis, unspecified
E063	E06.3 Autoimmune thyroiditis
E079	E07.9 Disorder of thyroid, unspecified
E101	E10.1 With ketoacidosis
E103	E10.3 With ophthalmic complications
E104	E10.4 With neurological complications
E105	E10.5 With peripheral circulatory complications
E109	E10.9 Without complications
E110	E11.0 With coma
E112	E11.2 With renal complications
E113	E11.3 With ophthalmic complications
E114	E11.4 With neurological complications
E115	E11.5 With perioheral circulatory comolications
E118	E11.8 With unspecified complications
E162	E16.2 Hypoglycaemia, unspecified
E210	E21.0 Primary hyoerparathyroidism
E211	E21.1 Secondary hyperparathyroidism, not elsewhere classified
E212	E21.2 Other hyperparathyroidism
E213	E21.3 Hyoerparathyrodism, unspecified
E222	E22.2 Syndrome of inappropriate secretion of antidiuretic hormone
E229	E22.9 Hyperfunction of pituitary gland, unspecified
E230	E23.0 Hypopituitarism
E232	E23.2 Diabetes insipidus
E236	E23.6 Other disorders of pituitary gland
E237	E23.7 Disorder of pituitary gland, unspecified
E271	E27.1 Primary adrenocortical insufficiency
E272	E27.2 Addisonian crisis
E274	E27.4 Other and unspecified adrenocortical insufficiency
E279	E27.9 Disorder of adrenal gland, unspecified
E348	E34.8 Other specified endocrine disorders
E349	E34.9 Endocrine disorder, unspecified
E46	E46 Unspecified protein-energy malnutrition
E538	E53.8 Deficiency of other specified B group vitamins
E559	E55.9 Vitamin D deficiency, unspecified
E804	E80.4 Gilbert's syndrome
E831	E83.1 Disorders of iron metabolism
E833	E83.3 Disorders of phosphorus metabolism
E834	E83.4 Disorders of magnesium metabolism
E835	E83.5 Disorders of calcium metabolism
E853	E85.3 Secondary systemic amyloidosis
E854	E85.4 Organ-limited amyloidosis
E859	E85.9 Amvloidosis, unspecified
E86	E86 Volume depletion
E870	E87.0 Hyperosmolality and hypernatraemia
E871	E87.1 Hypo-osmolality and hyponatraemia
E872	E87.2 Acidosis
E873	E87.3 Alkalosis
E875	E87.5 Hyperkalaemia
E876	E87.6 Hypokalaemia
E878	E87.8 Other disorders of electrolyte and fluid balance. not elsewhere classified
E880	E88.0 Disorders of plasma-protein metabolism, not elsewhere classified
E881	E88.1 Lipodystrophy, not elsewhere classified
E890	E89.0 Postoracedural hvoothvroidism
E891	E89.1 Postprocedural hypoinsulinaemia
F009	F00.9 Dementia in Alzheimer's disease, unspecified
F019	F01.9 Vascular dementia. unspecified
F03	F03 Unspecified dementia
F059	F05.9 Delirium, unspecified
F067	F06.7 Mild cognitive disorder
F069	F06.9 Unspecified mental disorder due to brain damage and dysfunction and to physical disease
F072	F07.2 Postconcussional syndrome
F09	F09 Unspecified organic or symptomatic mental disorder
F100	F10.0 Acute intoxication
F101	F10.1 Harmful use
F102	F10.2 Dependence syndrome
F103	F10.3 Withdrawal state
F104	F10.4 Withdrawal state with delirium

F105	F10.5 Psychotic disorder
F106	F10.6 Amnesic syndrome
F109	F10.9 Unspecified mental and behavioural disorder
F110	F11.0 Acute intoxication
F111	F11.1 Harmful use
F112	F11.2 Dependence syndrome
F115	F11.5 Psychotic disorder
F121	F12.1 Harmful use
F122	F12.2 Dependence syndrome
F170	F17.0 Acute intoxication
F171	F17.1 Harmful use
F172	F17.2 Dependence syndrome
F173	F17.3 Withdrawal state
F191	F19.1 Harmful use
F193	F19.3 Withdrawal state
F200	F20.0 Paranoid schizophrenia
F206	F20.6 Simple schizophrenia
F208	F20.8 Other schizophrenia
F209	F20.9 Schizophrenia, unspecified
F220	F22.0 Delusional disorder
F230	F23.0 Acute polymorphic psychotic disorder without symptoms of schizophrenia
F231	F23.1 Acute polymorphic psychotic disorder with symptoms of schizophrenia
F239	F23.9 Acute and transient psychotic disorder, unspecified
F258	F25.8 Other schizoaffective disorders
F29	F29 Unspecified nonorganic psychosis
F300	F30.0 Hypomania
F309	F30.9 Manic episode, unspecified
F310	F31.0 Bipolar affective disorder, current episode hypomanic
F312	F31.2 Bipolar affective disorder, current episode manic with psychotic symptoms
F315	F31.5 Bipolar affective disorder, current episode severe depression with psychotic symptoms
F317	F31.7 Bipolar affective disorder, currently in remission
F319	F31.9 Bipolar affective disorder, unspecified
F320	F32.0 Mild depressive episode
F321	F32.1 Moderate depressive episode
F322	F32.2 Severe deoressive episode without osychotic symotoms
F323	F32.3 Severe depressive episode with psychotic symptoms
F328	F32.8 Other depressive episodes
F329	F32.9 Deoressive episode. unsoecified
F330	F33.0 Recurrent depressive disorder, current episode mild
F331	F33.1 Recurrent depressive disorder, current episode moderate
F332	F33.2 Recurrent depressive disorder. current episode severe without osychotic symotoms
F333	F33.3 Recurrent depressive disorder, current episode severe with psychotic symptoms
F334	F33.4 Recurrent depressive disorder, currently in remission
F339	F33.9 Recurrent depressive disorder, unspecified
F341	F34.1 Dysthymia
F380	F38.0 Other single mood [affective] disorders
F402	F40.2 Specific (isolated) phobias
F410	F41.0 Panic disorder [episodic paroxysmal anxiety]
F411	F41.1 Generalised anxiety disorder
F412	F41.2 Mixed anxiety and depressive disorder
F419	F41.9 Anxiety disorder, unspecified
F420	F42.0 Predominantlv obsessional thouehts or ruminations
F428	F42.8 Other obsessive-compulsive disorders
F429	F42.9 Obsessive-compulsive disorder, unspecified
F430	F43.0 Acute stress reaction
F431	F43.1 Posttraumatic stress disorder
F432	F43.2 Adjustment disorders
F439	F43.9 Reaction to severe stress. unspecified
F458	F45.8 Other somatoform disorders
F500	F50.0 Anorexia nervosa
F501	F50.1 Atvpical anorexia nervosa
F502	F50.2 Bulimia nervosa
F508	F50.8 Other eating disorders
F509	F50.9 Eating disorder, unspecified
F522	F52.2 Failure of genital response
F530	F53.0 Mild mental and behavioural disorders associated with the puerperium, not elsewhere classified
F55	F55 Abuse of non-dependence-producing substances
F603	F60.3 Emotionally unstable personality disorder
F605	F60.5 Anankastic personality disorder
F606	F60.6 Anxious [avoidant] personality disorder
F607	F60.7 Dependent personality disorder
F609	F60.9 Personality disorder. unspecified
F633	F63.3 Trichotillomania
F640	F64.0 Transsexualism
F801	F80.1 Expressive laneuaee disorder
F819	F81.9 Developmental disorder of scholastic skills, unspecified
F911	F91.1 Unsocialised conduct disorder
F99	F99 Mental disorder. not otherwise soecified
G309	G30.9 Alzheimer's disease, unspecified
G35	G35 Multiple sclerosis
J440	J44.0 Chronic obstructive oulmonary disease with acute lower resoiratorv infection
J441	J44.1 Chronic obstructive pulmonary disease with acute exacerbation, unspecified
J448	J44.8 Othersoecified chronic obstructive pulmonary disease
J449	J44.9 Chronic obstructive pulmonary disease, unspecified
K500	K50.0 Crohn's disease of small intestine
K501	K50.1 Crohn's disease of large intestine
K508	K50.8 Other Crohn's disease
K509	K50.9 Crohn's disease, unspecified
K510	K51.0 Ulcerative (chronic) enterocolitis
K512	K51.2 Ulcerative (chronic) proctitis
K513	K51.3 Ulcerative (chronic) rectosigmoiditis
K518	K51.8 Other ulcerative colitis
K519	K51.9 Ulcerative colitis, unspecified
K521	K52.1 Toxic gastro-enteritis and colitis
K528	K52.8 Other soecified non-infective gastro-enteritis and colitis
K529	K52.9 Non-infective gastro-enteritis and colitis, unspecified
K580	K58.0 Irritable bowel syndrome with diarrhoea
K589	K58.9 Irritable bowel svndrome without diarrhoea
K700	K70.0 Alcoholic fatty liver
K703	K70.3 Alcoholic cirrhosis of liver
K709	K70.9 Alcoholic liver disease. unspecified
K720	K72.0 Acute and subacute hepatic failure
K729	K72.9 Hepatic failure, unspecified
K740	K74.0 Hepatic fibrosis
K743	K74.3 Primary biliary cirrhosis
K744	K74.4 Secondary biliary cirrhosis
K745	K74.5 Biliary cirrhosis, unspecified
K746	K74.6 Other and unspecified cirrhosis of liver
K750	K75.0 Abscess of liver
K760	K76.0 Fatty (change of) liver, not elsewhere classified
K766	K76.6 Portal hypertension
K767	K76.7 Hepatorenal syndrome
K768	K76.8 Other specified diseases of liver
K769	K76.9 Liver disease, unspecified
K770	K77.0 Liver disorders in infectious and parasitic diseases classified elsewhere
K900	K90.0 Coeliac disease
K904	K90.4 Malabsorption due to intolerance, not elsewhere classified
K909	K90.9 Intestinal malabsorption. unspecified
K910	K91.0 Vomiting following gastro-intestinal surgery
K911	K91.1 Postgastric surgery syndromes
K912	K91.2 Postsurgical malabsorption, not elsewhere classified
K913	K91.3 Postoperative intestinal obstruction
K914	K91.4 Colostomy and enterostomy malfunction
K918	K91.8 Other postprocedural disorders of digestive system, not elsewhere classified
K920	K92.0 Haematemesis
K921	K92.1 Melaena
K922	K92.2 Gastro-intestinal haemorrhage, unspecified
K928	K92.8 Other specified diseases of digestive system
K929	K92.9 Disease of digestive svstem. unsoecified
N180	N18.0 End-stage renal disease
N185	N18.5 Chronic kidney disease, stage 5
N188	N18.8 Other chronic renal failure
N189	N18.9 Chronic renal failure, unspecified
N19	N19 Unspecified renal failure
Q02	Q02 Microcephaly
Q874	Q87.4 Marfan's syndrome
R630	R63.0 Anorexia
R633	R63.3 Feeding difficulties and mismanagement
R634	R63.4 Abnormal weight loss
R64	R64 Cachexia
Y835	Y83.5 Amputation of limb(s)
Z511	Z51.1 Chemotherapy session for neoplasm
Z512	Z51.2 Other chemotherapy
Z800	Z80.0 Family history of malignant neoplasm of digestive organs
Z850	Z85.0 Personal history of malignant neoplasm of digestive organs
Z851	Z85.1 Personal history of malignant neoplasm of trachea, bronchus and lung
Z853	Z85.3 Personal history of malignant neoplasm of breast
Z854	Z85.4 Personal history of malignant neoplasm of genital organs
Z855	Z85.5 Personal history of malienant neoplasm of urinary tract
Z856	Z85.6 Personal history of leukaemia
Z857	Z85.7 Personal history of other malignant neoplasms of lymphoid, haematopoietic and related tissues
Z858	Z85.8 Personal history of malienant neoplasms of other oreans and svstems
Z860	Z86.0 Personal history of other neoplasms
Z864	Z86.4 Personal history of psychoactive substance abuse
Z865	Z86.5 Personal history of other mental and behavioural disorders
Z895	Z89.5 Acquired absence of leg at or below knee
Z896	Z89.6 Acquired absence of leg above knee
Z899	Z89.9 Acquired absence of limb. unspecified
Z901	Z90.1 Acquired absence of breast(s)
Z902	Z90.2 Acquired absence of lung [part of]
Z903	Z90.3 Acquired absence of part of stomach
Z904	Z90.4 Acquired absence of other parts of digestive tract
Z905	Z90.5 Acquired absence of kidney
Z906	Z90.6 Acquired absence of other parts of urinary tract
Z907	Z90.7 Acquired absence of genital organ(s)
Z992	Z99.2 Dependence on renal dialysis
Z993	Z99.3 Dependence on wheelchair

S13 Table. Self-reported illness codes used to exclude thin individuals in UKBB

	Psychiatric
1286	depression
1287	anxiety/panic attacks
1288	nervous breakdown
1289	schizophrenia
1290	deliberate self-harm/suicide attempt
1291	mania/bipolar disorder/manic depression
1469	post-traumatic stress disorder
1470	anorexia/bulimia/other eating disorder
1614	stress
1615	obsessive compulsive disorder (ocd)
1616	insomnia
1408	alcohol dependency
1409	opioid dependency
1410	other substance abuse/dependency
1531	post-natal depression
	Liver
1136	liver/biliary/pancreas problem
1155	hepatitis
1158	liver failure/cirrhosis
1159	bile duct disease
1161	gall bladder disease
1164	pancreatic disease
1507	haemochromatosis
1508	jaundice (unknown cause)
1156	infective/viral hepatitis
1157	non-infective hepatitis
1578	hepatitis a
1579	hepatitis b
1580	hepatitis c
1581	hepatitis d
1582	hepatitis e
1506	primary biliary cirrhosis
1604	alcoholic liver disease / alcoholic cirrhosis
1160	bile duct obstruction/ascending cholangitis
1475	sclerosing cholangitis
1165	pancreatitis
	Cardiac
1076	heart failure/pulmonary odema
	Renal
1192	renal/kidney failure
1193	renal failure requiring dialysis
1194	renal failure not requiring dialysis
1405	other renal/kidney problem
1196	urinary tract infection/kidney infection
1515	pyelonephritis
1427	polycystic kidney
1519	kidney nephropathy
1608	nephritis
1520	iga nephropathy
1607	diabetic nephropathy
1609	glomerulnephritis
	Gut
1154	irritable bowel syndrome
1456	malabsorption/coeliac disease
1457	duodenal ulcer
1459	colitis/not chrons or ulcerative colitis
1461	inflammatory bowel disease
1502	appendicitis
1503	anal problem
1599	constipation
1600	bowel / intestinal perforation
1601	bowel / intestinal infarction
1602	bowel / intestinal obstruction
1603	rectal prolapse
1462	crohns disease
1463	ulcerative colitis
	Abdominal
1400	peptic ulcer
	Endocrine
1224	thyroid problem (not cancer)
1229	parathyroid gland problem (not cancer)
1232	disorder of adrenal gland
1237	disorder of pituitary gland
1239	cushings syndrome
1432	carcinoid syndrome
1682	benign insulinoma
1221	gestational diabetes
1222	type 1 diabetes
1225	hyperthyroidism/thyrotoxicosis
1226	hypothyroidism/myxoedema
1228	thyroid radioablation therapy
1428	thyroiditis
1522	grave's disease
1610	thyroid goitre
1230	parathyroid hyperplasia/adenoma
1611	hyperparathyroidism
1233	adrenal tumour
1234	adrenocortical insufficiency/addison's disease
1235	hyperaldosteronism/conn's syndrome
1236	phaeochromocytoma
1238	pituitary adenoma/tumour
1429	acromegaly
1430	hypopituitarism
1431	hyperprolactinaemia
	COPD
1112	COPD
	Infections
1439	hiv/aids
1567	infectious mononucleosis / glandular fever / epstein barr virus (ebv)
1440	tuberculosis (tb)
1575	herpes simplex

Cancer (responded yes to "Have you ever been diagnosed with cancer?")

Appendix B

Supplementary Tables 1 and 2 are too large to print. They are located here:

Supplementary Table 1

https://docs.google.com/spreadsheets/d/1HYbX5qI81pvMjAM7bn8yWlN34OGtwudpDJLzLVUbu5A/edit?usp=sharing

Supplementary Table 2

https://docs.google.com/spreadsheets/d/19s_C6eb7uX4etbaTQ0M-XUvYYhOeTutiyIJwM1XwJ4A/edit?usp=sharing

Supplementary Table 4: Gene set analyses results

Gene set id	Trait	Meta-p	Meta-p (no APO)	WES p	N WES	WGS p	N WGS	Description	Source
C0020445	lhdlfc_	2.31E-10	0.02813214	1.01E-05	35	7.62E-06	21	Hypercholesterolemia Familial	DisGeneNet
C0020476	lhdlfc_	1.58E-11	0.000932652	2.39E-06	14	7.77E-07	7	Hyperlipoproteinemias	DisGeneNet
C0020476	hdld	1.81E-10	0.000279994	0.000496	14	1.80E-08	7	Hyperlipoproteinemias	DisGeneNet
C0020476	lhdlc_	2.90E-08	0.00385449	2.23E-05	14	0.000201	7	Hyperlipoproteinemias	DisGeneNet
C0020476	lhdlpl_	2.15E-06	0.002200132	0.000977	14	0.000793	7	Hyperlipoproteinemias	DisGeneNet
C0342881	idltg	2.02E-11	0.015485781	2.03E-09	11	0.002838	8	Familial hypercholesterolemia - homozygous	DisGeneNet
C0342881	xsvldlp	3.79E-10	0.014275635	4.03E-07	11	0.00085	8	Familial hypercholesterolemia - homozygous	DisGeneNet
C0342881	ldltg	7.64E-10	0.006844523	9.76E-09	11	0.004302	8	Familial hypercholesterolemia - homozygous	DisGeneNet
C0342881	xsvldltg	1.08E-09	0.023413237	1.84E-07	11	0.006007	8	Familial hypercholesterolemia - homozygous	DisGeneNet
C0342881	lldltg	3.58E-09	0.005062039	8.20E-08	11	0.003857	8	Familial hypercholesterolemia - homozygous	DisGeneNet
C0342881	apob	7.72E-09	0.005089742	2.38E-07	11	0.002934	8	Familial hypercholesterolemia - homozygous	DisGeneNet
C0342881	svldlfc	3.18E-08	0.012250296	2.71E-05	11	0.002389	8	Familial hypercholesterolemia - homozygous	DisGeneNet
C0342881	mldltg	7.07E-08	0.013478956	5.24E-08	11	0.029378	8	Familial hypercholesterolemia - homozygous	DisGeneNet
C0342881	sldltg	8.59E-08	0.016697804	5.88E-08	11	0.026173	8	Familial hypercholesterolemia - homozygous	DisGeneNet
C0342881	mufa	1.10E-07	0.018070242	0.00013	11	0.007047	8	Familial hypercholesterolemia - homozygous	DisGeneNet
C0342881	idll	1.75E-07	0.010999563	3.69E-06	11	0.003782	8	Familial hypercholesterolemia - homozygous	DisGeneNet
C0342881	apobapoa1	2.15E-07	0.004237918	1.04E-06	11	0.00795	8	Familial hypercholesterolemia - homozygous	DisGeneNet
C0342881	lldlp	2.43E-07	0.009922028	4.65E-07	11	0.012224	8	Familial hypercholesterolemia - homozygous	DisGeneNet
C0342881	svldlpl	2.48E-07	0.0089107	4.13E-05	11	0.002879	8	Familial hypercholesterolemia - homozygous	DisGeneNet
C0342881	lldll	2.84E-07	0.010485712	8.43E-07	11	0.013297	8	Familial hypercholesterolemia - homozygous	DisGeneNet
C0342881	xsvldlpl	3.71E-07	0.03201298	1.63E-07	11	0.004467	8	Familial hypercholesterolemia - homozygous	DisGeneNet
C0342881	idlp	3.89E-07	0.009724476	1.56E-06	11	0.002886	8	Familial hypercholesterolemia - homozygous	DisGeneNet
C0342881	idlpl	4.91E-07	0.012464279	3.13E-06	11	0.008312	8	Familial hypercholesterolemia - homozygous	DisGeneNet
C0342881	ldlc	6.95E-07	0.013848465	1.34E-06	11	0.026768	8	Familial hypercholesterolemia - homozygous	DisGeneNet
C0342881	lldlpl	7.04E-07	0.013332371	3.51E-06	11	0.018528	8	Familial hypercholesterolemia - homozygous	DisGeneNet
C0342881	lldlce	7.20E-07	0.01120345	2.29E-06	11	0.018631	8	Familial hypercholesterolemia - homozygous	DisGeneNet
C0342881	mldlpl	7.91E-07	0.012623335	1.12E-06	11	0.030335	8	Familial hypercholesterolemia - homozygous	DisGeneNet
C0342881	totfa	9.12E-07	0.020070097	2.66E-05	11	0.006704	8	Familial hypercholesterolemia - homozygous	DisGeneNet
C0342881	lldlc	9.44E-07	0.01233823	3.16E-06	11	0.018568	8	Familial hypercholesterolemia - homozygous	DisGeneNet
C0342881	mldlp	9.49E-07	0.012045521	2.99E-07	11	0.043184	8	Familial hypercholesterolemia - homozygous	DisGeneNet
C0342881	mldll	1.10E-06	0.011701482	4.54E-07	11	0.047026	8	Familial hypercholesterolemia - homozygous	DisGeneNet
C0342881	mldlfc	1.64E-06	0.03871723	2.76E-06	11	0.044593	8	Familial hypercholesterolemia - homozygous	DisGeneNet
C0342883	lhdlfc_	9.97E-14	0.001782186	6.12E-07	9	1.04E-09	4	Cholesteryl Ester Transfer Protein Deficiency	DisGeneNet
C0342883	tgpg	9.85E-10	0.016207152	5.21E-05	9	2.13E-06	4	Cholesteryl Ester Transfer Protein Deficiency	DisGeneNet
C0542037	lhdlfc_	3.57E-13	0.003632137	6.12E-07	9	1.74E-09	3	Hypotriglyceridaemia	DisGeneNet
C0542037	tgpg	3.23E-09	0.01845352	5.21E-05	9	2.21E-06	3	Hypotriglyceridaemia	DisGeneNet
C0745103	idltg	1.90E-10	0.008406138	1.83E-08	21	0.010046	17	Hyperlipoproteinemia Type IIa	DisGeneNet
C0745103	xsvldltg	2.03E-10	0.001916666	3.13E-07	21	0.008834	17	Hyperlipoproteinemia Type IIa	DisGeneNet
C0745103	svldlfc	1.22E-09	0.001385636	3.49E-05	21	0.001432	17	Hyperlipoproteinemia Type IIa	DisGeneNet
C0745103	xsvldlp	3.75E-09	0.014609129	3.02E-06	21	0.00142	17	Hyperlipoproteinemia Type IIa	DisGeneNet
C0745103	svldlpl	4.15E-09	0.000606073	3.76E-05	21	0.001927	17	Hyperlipoproteinemia Type IIa	DisGeneNet
C0745103	svldll	1.06E-08	0.001568385	7.76E-05	21	0.002428	17	Hyperlipoproteinemia Type IIa	DisGeneNet
C0745103	svldlp	1.49E-08	0.001319162	8.22E-05	21	0.003239	17	Hyperlipoproteinemia Type IIa	DisGeneNet
C0745103	mufa	1.26E-07	0.00369098	0.000211	21	0.003996	17	Hyperlipoproteinemia Type IIa	DisGeneNet
C0745103	ldltg	2.00E-07	0.020495788	2.63E-07	21	0.013265	17	Hyperlipoproteinemia Type IIa	DisGeneNet
C0745103	lldltg	4.58E-07	0.020209296	1.50E-06	21	0.01609	17	Hyperlipoproteinemia Type IIa	DisGeneNet
C0745103	sldltg	5.19E-07	0.01743014	2.02E-06	21	0.032317	17	Hyperlipoproteinemia Type IIa	DisGeneNet
C0745103	apob	1.19E-06	0.006937804	2.87E-06	21	0.001865	17	Hyperlipoproteinemia Type IIa	DisGeneNet
C0745103	apobapoa1	1.34E-06	0.004883344	1.88E-05	21	0.010105	17	Hyperlipoproteinemia Type IIa	DisGeneNet
C0745103	mvldlfc	1.68E-06	0.000836725	0.000477	21	0.019124	17	Hyperlipoproteinemia Type IIa	DisGeneNet
C0745103	totfa	1.71E-06	0.006830084	6.15E-05	21	0.006047	17	Hyperlipoproteinemia Type IIa	DisGeneNet
C1848486	xsvldlpl	5.53E-08	0.004265067	6.53E-07	11	0.005985	9	Premature arteriosclerosis	DisGeneNet
C1848486	sldltg	2.10E-07	0.026356402	9.08E-08	11	0.036179	9	Premature arteriosclerosis	DisGeneNet
C1848486	mldltg	8.28E-07	0.02980488	2.03E-07	11	0.044683	9	Premature arteriosclerosis	DisGeneNet
C4280503	xsvldlpl	5.53E-08	0.004265067	6.53E-07	11	0.005985	9	Premature hardening of arteries	DisGeneNet
C4280503	sldltg	2.10E-07	0.026356402	9.08E-08	11	0.036179	9	Premature hardening of arteries	DisGeneNet
C4280503	mldltg	8.28E-07	0.02980488	2.03E-07	11	0.044683	9	Premature hardening of arteries	DisGeneNet
R-HSA-204174	idlpl	7.85E-07	7.85E-07	0.005939	12	0.000503	4	Regulation of pyruvate dehydrogenase (PDH) complex	Reactome
R-HSA-204174	mldlpl	1.01E-06	1.01E-06	0.002671	12	0.000594	4	Regulation of pyruvate dehydrogenase (PDH) complex	Reactome
R-HSA-204174	estc	1.09E-06	1.09E-06	0.004754	12	0.001175	4	Regulation of pyruvate dehydrogenase (PDH) complex	Reactome
R-HSA-204174	idlp	1.17E-06	1.17E-06	0.003992	12	0.000593	4	Regulation of pyruvate dehydrogenase (PDH) complex	Reactome
R-HSA-204174	lldlp	1.20E-06	1.20E-06	0.004822	12	0.000258	4	Regulation of pyruvate dehydrogenase (PDH) complex	Reactome
R-HSA-204174	lldlpl	1.21E-06	1.21E-06	0.004853	12	0.000423	4	Regulation of pyruvate dehydrogenase (PDH) complex	Reactome
R-HSA-204174	idll	1.21E-06	1.21E-06	0.004313	12	0.000574	4	Regulation of pyruvate dehydrogenase (PDH) complex	Reactome
R-HSA-204174	serumc	1.24E-06	1.24E-06	0.005999	12	0.001071	4	Regulation of pyruvate dehydrogenase (PDH) complex	Reactome
R-HSA-204174	lldll	1.35E-06	1.35E-06	0.005082	12	0.000275	4	Regulation of pyruvate dehydrogenase (PDH) complex	Reactome
R-HSA-204174	idlc	1.40E-06	1.40E-06	0.00475	12	0.001019	4	Regulation of pyruvate dehydrogenase (PDH) complex	Reactome
R-HSA-204174	lldlfc	1.46E-06	1.46E-06	0.00681	12	0.0003	4	Regulation of pyruvate dehydrogenase (PDH) complex	Reactome
R-HSA-204174	lldlc	1.87E-06	1.87E-06	0.006489	12	0.000275	4	Regulation of pyruvate dehydrogenase (PDH) complex	Reactome
R-HSA-204174	mldlp	1.96E-06	1.96E-06	0.006409	12	0.000132	4	Regulation of pyruvate dehydrogenase (PDH) complex	Reactome
R-HSA-204174	lldlce	2.01E-06	2.01E-06	0.006486	12	0.000277	4	Regulation of pyruvate dehydrogenase (PDH) complex	Reactome
R-HSA-204174	sldll	2.13E-06	2.13E-06	0.006413	12	0.000115	4	Regulation of pyruvate dehydrogenase (PDH) complex	Reactome
R-HSA-204174	sldlp	2.13E-06	2.13E-06	0.005994	12	0.000113	4	Regulation of pyruvate dehydrogenase (PDH) complex	Reactome
R-HSA-204174	mldll	2.13E-06	2.13E-06	0.006416	12	0.000164	4	Regulation of pyruvate dehydrogenase (PDH) complex	Reactome
R-HSA-204174	ldlc	2.17E-06	2.17E-06	0.007809	12	0.000177	4	Regulation of pyruvate dehydrogenase (PDH) complex	Reactome
R-HSA-204174	apob	2.20E-06	2.20E-06	0.00504	12	0.000803	4	Regulation of pyruvate dehydrogenase (PDH) complex	Reactome
R-HSA-204174	idlfc	2.22E-06	2.22E-06	0.009798	12	0.000399	4	Regulation of pyruvate dehydrogenase (PDH) complex	Reactome
R-HSA-8866423	xsvldlp	1.49E-12	0.027026999	2.06E-09	8	0.000246	7	VLDL assembly	Reactome
R-HSA-8866423	xsvldll	6.57E-12	0.029658511	3.13E-09	8	0.000254	7	VLDL assembly	Reactome
R-HSA-8866423	xsvldlpl	3.27E-10	0.047296943	4.87E-10	8	0.000529	7	VLDL assembly	Reactome
R-HSA-8866423	idlp	5.94E-10	0.012821521	2.79E-09	8	0.000385	7	VLDL assembly	Reactome
R-HSA-8866423	apob	9.00E-10	0.035805827	1.23E-09	8	0.001105	7	VLDL assembly	Reactome
R-HSA-8866423	lldlpl_	1.21E-09	0.003361758	2.31E-11	8	0.006697	7	VLDL assembly	Reactome
R-HSA-8866423	idll	2.82E-09	0.014169646	1.95E-08	8	0.000547	7	VLDL assembly	Reactome
R-HSA-8866423	ldlc	2.02E-08	0.015814492	6.36E-09	8	0.003754	7	VLDL assembly	Reactome

R-HSA-8866423	lldlp	2.09E-08	0.010925413	1.78E-09	8	0.001674	7	VLDL assembly	Reactome
R-HSA-8866423	remnantc	6.95E-08	0.005468158	4.44E-09	8	0.00083	7	VLDL assembly	Reactome
R-HSA-8866423	lldlfc	1.75E-07	0.012845409	7.16E-08	8	0.002439	7	VLDL assembly	Reactome
R-HSA-8866423	idlpl	1.84E-07	0.011330913	2.70E-08	8	0.000806	7	VLDL assembly	Reactome
R-HSA-8866423	xsvldlfc	1.96E-07	0.037068974	1.64E-07	8	0.002008	7	VLDL assembly	Reactome
R-HSA-8866423	lldlpl	2.05E-07	0.009682997	1.12E-08	8	0.00295	7	VLDL assembly	Reactome
R-HSA-8866423	lldlce	2.22E-07	0.012777309	8.81E-09	8	0.002613	7	VLDL assembly	Reactome
R-HSA-8866423	lldlc	2.28E-07	0.012348304	1.40E-08	8	0.002473	7	VLDL assembly	Reactome
R-HSA-8866423	lldll	2.53E-07	0.010991255	3.93E-09	8	0.00178	7	VLDL assembly	Reactome
R-HSA-8866423	idlfc	2.64E-07	0.020705061	2.07E-07	8	0.001976	7	VLDL assembly	Reactome
R-HSA-8866423	xsvldlc	2.81E-07	0.010931813	7.53E-06	8	0.000766	7	VLDL assembly	Reactome
R-HSA-8866423	idlc	3.52E-07	0.018406604	6.83E-07	8	0.001729	7	VLDL assembly	Reactome
R-HSA-8866423	serumc	4.74E-07	0.023383675	6.02E-07	8	0.008607	7	VLDL assembly	Reactome
R-HSA-8866423	idlce	5.22E-07	0.00201215	1.58E-06	8	0.001804	7	VLDL assembly	Reactome
R-HSA-8866423	mldlp	5.32E-07	0.019315992	6.25E-09	8	0.008059	7	VLDL assembly	Reactome
R-HSA-8866423	mldll	5.50E-07	0.017018952	9.19E-09	8	0.008598	7	VLDL assembly	Reactome
R-HSA-8866423	estc	5.80E-07	0.024024992	4.43E-07	8	0.012954	7	VLDL assembly	Reactome
R-HSA-8866423	freec	6.04E-07	0.027416347	6.56E-06	8	0.004008	7	VLDL assembly	Reactome
R-HSA-8866423	idlpl_	6.58E-07	0.039687097	7.17E-07	8	0.01239	7	VLDL assembly	Reactome
R-HSA-8866423	mldlpl	7.13E-07	0.015426761	1.99E-08	8	0.010748	7	VLDL assembly	Reactome
R-HSA-8866423	xsvldlce	7.44E-07	0.009844208	4.89E-05	8	0.000835	7	VLDL assembly	Reactome
R-HSA-8866423	sldlc	7.51E-07	0.024307244	4.84E-09	8	0.017042	7	VLDL assembly	Reactome
R-HSA-8866423	sldlp	7.67E-07	0.027289638	2.54E-09	8	0.015185	7	VLDL assembly	Reactome
R-HSA-8866423	pufa	7.71E-07	0.08454695	1.50E-06	8	0.008925	7	VLDL assembly	Reactome
R-HSA-8866423	vldlc	8.93E-07	0.052364901	1.41E-05	8	0.002975	7	VLDL assembly	Reactome
R-HSA-8866423	sldlce	9.02E-07	0.007812149	2.95E-09	8	0.01715	7	VLDL assembly	Reactome
R-HSA-8866423	sldll	9.22E-07	0.027649486	1.60E-09	8	0.015745	7	VLDL assembly	Reactome
R-HSA-8866423	mldlc	1.18E-06	0.019562719	3.67E-08	8	0.012777	7	VLDL assembly	Reactome
R-HSA-8866423	mldlce	1.27E-06	0.021261762	4.76E-08	8	0.012747	7	VLDL assembly	Reactome
R-HSA-8866423	svldlce	1.44E-06	0.024795814	8.30E-05	8	0.001773	7	VLDL assembly	Reactome
R-HSA-8866423	mldlfc	2.02E-06	0.005231542	3.73E-08	8	0.016205	7	VLDL assembly	Reactome
R-HSA-8866423	sldlfc	2.14E-06	0.015190169	6.26E-09	8	0.027399	7	VLDL assembly	Reactome
R-HSA-8963888	xsvldlp	2.49E-14	0.206996778	2.15E-10	10	2.02E-05	11	Chylomicron assembly	Reactome
R-HSA-8963888	svldlc	3.38E-14	0.378917505	1.71E-09	10	2.65E-05	11	Chylomicron assembly	Reactome
R-HSA-8963888	xsvldll	2.89E-13	0.204758667	3.87E-10	10	5.27E-05	11	Chylomicron assembly	Reactome
R-HSA-8963888	apobapoa1	8.43E-13	0.167387417	1.72E-09	10	3.83E-06	11	Chylomicron assembly	Reactome
R-HSA-8963888	vldlc	2.12E-11	0.280931433	3.23E-09	10	4.19E-05	11	Chylomicron assembly	Reactome
R-HSA-8963888	lldlfc_	7.49E-11	0.195504351	2.15E-10	10	0.000573	11	Chylomicron assembly	Reactome
R-HSA-8963888	svldlce	2.15E-10	0.147784098	1.40E-08	10	0.000173	11	Chylomicron assembly	Reactome
R-HSA-8963888	mvldlce	2.15E-10	0.281624878	1.40E-07	10	0.000117	11	Chylomicron assembly	Reactome
R-HSA-8963888	remnantc	5.86E-10	0.085210798	2.92E-08	10	0.000799	11	Chylomicron assembly	Reactome
R-HSA-8963888	ldltg	9.59E-10	0.396710914	6.92E-08	10	0.000441	11	Chylomicron assembly	Reactome
R-HSA-8963888	lldltg	3.23E-09	0.294471306	3.09E-07	10	0.000602	11	Chylomicron assembly	Reactome
R-HSA-8963888	xsvldlfc	3.29E-09	0.081395683	2.97E-08	10	0.002742	11	Chylomicron assembly	Reactome
R-HSA-8963888	mufa	2.93E-08	0.388099899	7.78E-06	10	0.002762	11	Chylomicron assembly	Reactome
R-HSA-8963888	xsvldlce	2.94E-08	0.069144815	2.02E-06	10	0.001391	11	Chylomicron assembly	Reactome
R-HSA-8963888	xsvldlc	3.01E-08	0.06186095	4.02E-07	10	0.001125	11	Chylomicron assembly	Reactome
R-HSA-8963888	idlpl_	3.01E-08	0.098183672	7.18E-08	10	0.005842	11	Chylomicron assembly	Reactome
R-HSA-8963888	apob	6.39E-08	0.24090487	6.48E-09	10	0.001438	11	Chylomicron assembly	Reactome
R-HSA-8963888	xsvldlpl	6.41E-08	0.303509995	6.64E-09	10	0.000858	11	Chylomicron assembly	Reactome
R-HSA-8963888	lhdltg_	2.38E-07	0.248013219	6.88E-05	10	0.000773	11	Chylomicron assembly	Reactome
R-HSA-8963888	lldlpl_	4.25E-07	0.046875529	2.37E-07	10	0.02747	11	Chylomicron assembly	Reactome
R-HSA-8963888	xlvldlpl	6.11E-07	0.195788822	0.000226	10	0.001796	11	Chylomicron assembly	Reactome
R-HSA-8963888	xlvldll	6.12E-07	0.215946625	0.000218	10	0.001764	11	Chylomicron assembly	Reactome
R-HSA-8963888	xlvldlp	6.27E-07	0.336990987	0.000219	10	0.001784	11	Chylomicron assembly	Reactome
R-HSA-8963888	xsvldltg_	6.28E-07	0.089073028	0.00775	10	1.70E-05	11	Chylomicron assembly	Reactome
R-HSA-8963888	xlvldlce	1.14E-06	0.309746179	0.000159	10	0.001166	11	Chylomicron assembly	Reactome
R-HSA-8963888	xxlvldltg	1.16E-06	0.176848043	0.000474	10	0.002696	11	Chylomicron assembly	Reactome
R-HSA-8963888	xlvldltg	1.18E-06	0.321278515	0.000222	10	0.002044	11	Chylomicron assembly	Reactome
R-HSA-8963888	xlvldlc	1.20E-06	0.309130119	0.000172	10	0.001447	11	Chylomicron assembly	Reactome
R-HSA-8963888	xxlvldlce	1.21E-06	0.210451215	0.000326	10	0.001414	11	Chylomicron assembly	Reactome
R-HSA-8963888	xxlvldll	1.29E-06	0.192140471	0.000347	10	0.002374	11	Chylomicron assembly	Reactome
R-HSA-8963888	lvldlpl_	1.43E-06	0.551285259	8.06E-06	9	0.014211	11	Chylomicron assembly	Reactome
R-HSA-8963888	xxlvldlp	2.07E-06	0.186176084	0.000354	10	0.00242	11	Chylomicron assembly	Reactome
R-HSA-8963888	totfa	2.11E-06	0.511678701	8.44E-07	10	0.045133	11	Chylomicron assembly	Reactome
R-HSA-8963888	xxlvldlc	2.14E-06	0.219048098	0.000325	10	0.001807	11	Chylomicron assembly	Reactome
R-HSA-8963888	xlvldlfc	2.19E-06	0.314011612	0.000228	10	0.002124	11	Chylomicron assembly	Reactome
R-HSA-8963898	xsvldltg	9.97E-10	0.237293907	2.89E-08	23	0.003889	19	Plasma lipoprotein assembly	Reactome
R-HSA-8963898	svldlpl	6.28E-07	0.819781918	1.17E-06	23	0.002978	19	Plasma lipoprotein assembly	Reactome
R-HSA-8963898	svldlfc	6.79E-07	1	1.11E-06	23	0.004052	19	Plasma lipoprotein assembly	Reactome
R-HSA-8963898	svldlp	1.19E-06	1	1.35E-06	23	0.004011	19	Plasma lipoprotein assembly	Reactome
R-HSA-8963898	svldll	1.25E-06	1	1.13E-06	23	0.004472	19	Plasma lipoprotein assembly	Reactome
R-HSA-8963901	hdld	9.72E-10	0.001414545	0.000108	12	8.68E-06	12	Chylomicron remodeling	Reactome
R-HSA-8963901	xlhdlfc	3.04E-09	0.004188796	0.000336	12	6.07E-05	12	Chylomicron remodeling	Reactome
R-HSA-8963901	lhdlc_	1.01E-08	0.003841981	4.60E-05	12	0.000594	12	Chylomicron remodeling	Reactome
R-HSA-8963901	xlhdlpl	1.13E-08	0.007480561	0.000162	12	4.45E-05	12	Chylomicron remodeling	Reactome
R-HSA-8963901	xlhdlc	1.76E-07	0.011331821	0.002411	12	0.000666	12	Chylomicron remodeling	Reactome
R-HSA-8964058	tgpg	5.88E-10	0.006630914	1.81E-05	17	2.46E-06	8	HDL remodeling	Reactome

Meta-p= Meta-analysis p-value
Meta-p (no APO) = Meta-analysis p-value after removing APO genes from gene sets (APOB and APOC3)
WES p = p-value in WES dataset
N WES = number of variants tested in WES dataset
WGS p = p-value in WGS dataset
N WGS = number of variants tested in WGS dataset

Supplementary Table 9: Detailed results for gene sets with enriched rare variation in tails of lipoprotein traits

S-VLDL-C lower tail outliers. Hyperlipidemia gene set.

gene	snp	dataset	MAC	residuals in all carriers
AGL	rs200459772	WES	5	2.36762834154852,-0.334045067074641,0.431527558983269,-0.838811852821138,-3.05000388882286
APOB	2:21236148	WES	1	-2.661258903
APC	rs150973053	WES	1	-3.089584993
APC	rs201830995	WES	3	-2.87066740721444,-0.787318922230483,0.420463200843388
CYP19A1	rs141305220	WES	2	-3.494055746710 22,-1.2437172570647
CYP19A1	rs200111039	WES	9	-2.975904533000663,0.253051068847167,0.795701074251656,1.01065228811834,-0.403431340606028,-0.144560598282279,-3.08958499345356,0.741693060794646,-0.4324749949308
NPHS1	rs368988883	WES	1	-3.374778926
GCG	2:163003928	WGS	1	-3.123944186
APC	5:112173509	WGS	2	-3.543944498814 21,-0.121786221385006
APC	5:112174919	WGS	2	-3.543944498814 21,-0.121786221385006
APC	5:112178070	WGS	2	-3.543944498814 21,-0.121786221385006
APC	5:112179437	WGS	2	-3.543944498814 21,-0.121786221385006
NOS3	7:150698995	WGS	2	-0.18836307152566,-2.7037021818663
NOS3	7:150706632	WGS	5	-0.304682642445497,-0.26183599116571,-0.48084454181164,-2.83884053615798,-0.759193154129386
CETP	rs150236668	WGS	2	-1.08925353711354,-3.543944498814 21
NPHS1	19:36342715	WGS	3	0.920578019402659,-0.398163133632229,-2.93246487402699

XS-VLDL-P lower tail outliers. Hyperlipidemia gene set.

gene	snp	dataset	MAC	effects
APOB	2:21236148	WES	1	-3.436640493
APC	rs150973053	WES	1	-3.202863174
APC	rs201830995	WES	3	-2.86524374013287,-0.168052075323293,0.077312983454771
NOS3	rs141170595	WES	7	-1.18611115166881,-2.955898255405 99,-0.246085238062439,1.13215214546922,0.154253491587311,-0.217108986788457,-1.74689283105004
CYP19A1	rs141305220	WES	2	-3.205389847208 28,-0.876451886676179
CYP19A1	rs200111039	WES	9	-2.640986452125 51,0.545406135777777,1.01054251354388,0.763927578916 17,-0.245365417517183,-0.682701582154253,-3.20286317422441,0.569449665319146,-0.251861115970539
NPHS1	rs368988883	WES	1	-3.318749511
NPHS2	1:179520511	WGS	2	0.213855282424323,-2.73710673267041
NPHS2	1:179530462	WGS	6	-2.85736273031488,0.500033274189366,0.129175297645043,0.476908535573381,-0.94191940828643,0.175183524144263
APOB	2:21255263	WGS	1	-2.965806851
GCG	2:163003928	WGS	1	-3.430062283
APC	5:112173509	WGS	2	-2.949070995374 61,-0.259525678305062
APC	5:112174919	WGS	2	-2.949070995374 61,-0.259525678305062
APC	5:112178070	WGS	2	-2.949070995374 61,-0.259525678305062
APC	5:112179437	WGS	2	-2.949070995374 61,-0.259525678305062
NOS3	7:150698995	WGS	2	-0.466020371752611,-2.83066719904639
CETP	rs150236668	WGS	2	-0.633213631434203,-2.94907099537461

S-VLDL-CE lower tail outliers. Hyperlipidemia gene set.

gene	snp	dataset	MAC	effects
AGL	rs200459772	WES	5	2.36762834154852,-0.334045067074641,0.431527558983269,-0.838811852821138,-3.05000388882286
APOB	2:21236148	WES	1	-2.661258903
APC	rs150973053	WES	1	-3.089584993
APC	rs201830995	WES	3	-2.87066740721444,-0.787318922230483,0.420463200843388
CYP19A1	rs141305220	WES	2	-3.494055746710 22,-1.2437172570647
CYP19A1	rs200111039	WGS	9	-2.975904533000663,0.253051068847167,0.795701074251656,1.01065228811834,-0.403431340606028,-0.144560598282279,-3.08958499345356,0.741693060794646,-0.4324749949308
NPHS1	rs368988883	WES	1	-3.374778926
GCG	2:163003928	WGS	1	-3.123944186
APC	5:112173509	WGS	2	-3.543944498814 21,-0.121786221385006
APC	5:112174919	WGS	2	-3.543944498814 21,-0.121786221385006
APC	5:112178070	WGS	2	-3.543944498814 21,-0.121786221385006
APC	5:112179437	WGS	2	-3.543944498814 21,-0.121786221385006
NOS3	7:150698995	WGS	2	-0.18836307152566,-2.7037021818663
NOS3	7:150706632	WGS	5	-0.304682642445497,-0.26183599116571,-0.48084454181164,-2.83884053615798,-0.759193154129386
CETP	rs150236668	WGS	2	-1.08925353711354,-3.543944498814 21
NPHS1	19:36342715	WGS	3	0.920578019402659,-0.398163133632229,-2.93246487402699

S-HDL-P lower tail outliers. HDL remodeling gene set

gene	snp	dataset	MAC	effects
CETP	rs140547417	WES	10	0.578651608406939,0.610798008574449,0.292679415486239,0.395395459347943,-1.113868535475629,-2.932639377408 99,-0.0998285578608295,-0.0864903418646204,-0.318903381965163,0.775064146714445
LIPG	18:47107925	WES	1	-3.234598237
APOE	rs199768005	WES	7	-0.540244700238687,1.92520088605348,0.92978260411206,-3.02709326825206,-0.789305787720864,0.121976706457689,-1.34111543948004
ABCG1	rs148226451	WES	1	-2.932639377
APOA1	11:11670686:	WGS	1	-3.003505735
APOA1	rs199759119	WGS	7	-1.005363495326 65,-2.86126280384725,-0.582922966059555,-0.693164051709012,1.64763326906752,-2.39922951075938,-1.79280325835833
CETP	rs142750310	WGS	1	-3.046409293

Supplementary Table 10: Sensitity analyses for rare variant enrichment in tails analysis using different percentile cutoffs to define tails of the phenotypic distribution

.5% Percentile upper tails

trait	p.wes	p.wgs	meta-p	Gene set
lldlc	0.00432	0.03209	0.0007737	LDL_clearance
vldld	0.02887	0.00607	0.0009188	VLDL_clearance

.5% Percentile lower tails

trait	p.wes	p.wgs	meta-p	Gene set
svldlce	0.02992	0.01634	0.0022477	Hyperlipidemia
svldlfc	0.01287	0.00676	0.0004448	Hyperlipidemia
xsvldlp	0.02992	0.0024	0.0004422	Hyperlipidemia
idltg	0.00032	0.01528	4.02E-05	LDL_remodeling
idltg	1.00E-05	0.01621	2.97E-06	VLDL_assembly

1 Percentile lower tails

trait	p.wes	p.wgs	meta-p	Gene set
mhdltg	0.04487	0.00976	0.0021777	Hyperlipidemia

p.wes: permutation p-value in WES
p.wgs: permutation p-value in WGS
meta-p: p-value after meta-analysis using Stouffer's method
Highlighted in yellow are gene sets that are significant after meta-analysis using Stouffer's method and after adjusting for multiple traits (p<=0.00037).