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**).

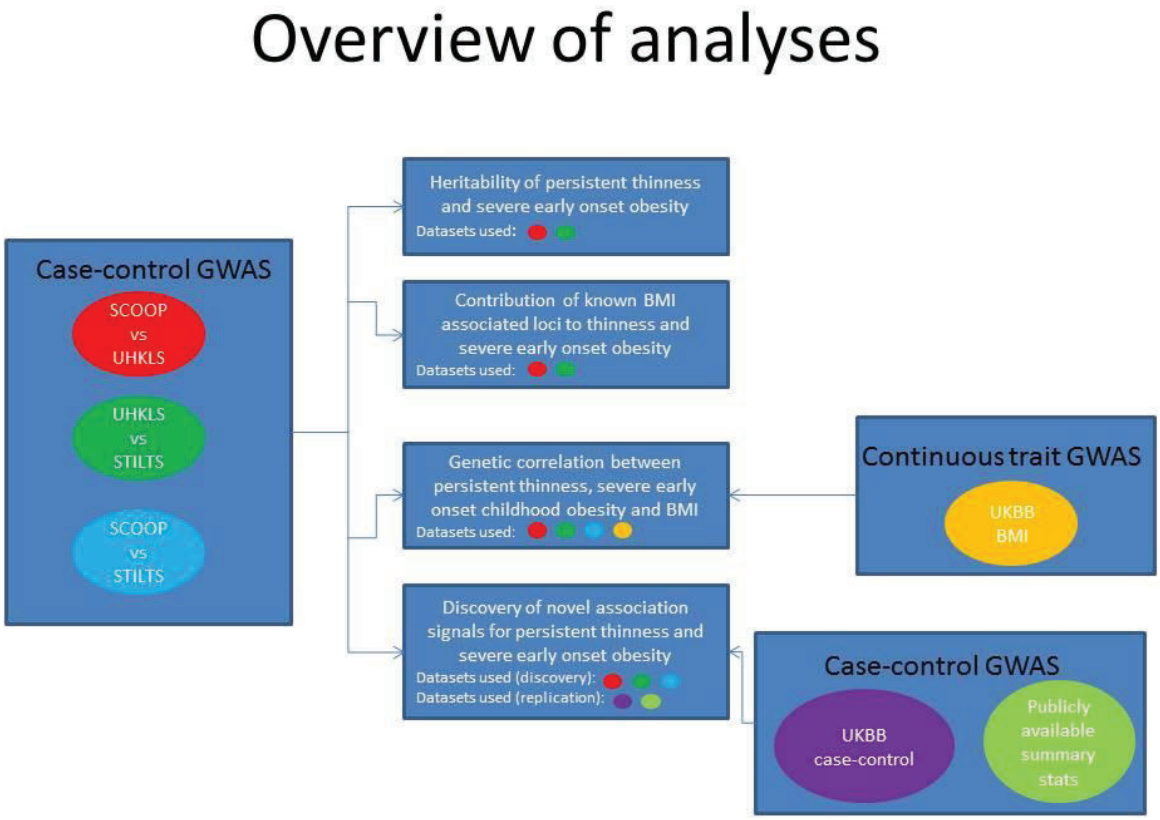

Figure 2.1: Overview of cohorts and analyses.

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.

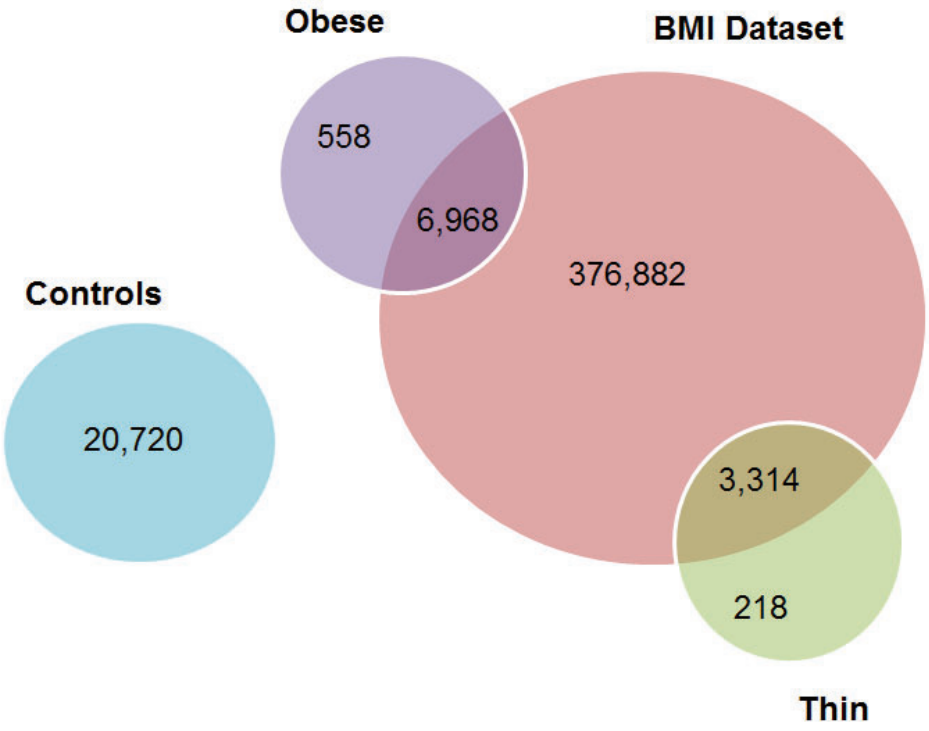

Figure 2.2: Summary of the UKBB sample sets after QC. Venn Diagram showing sample numbers and overlap between UKBB sample sets used in genetic correlation (BMI dataset) and GWAS replication (obese, controls, lean) analyses.

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

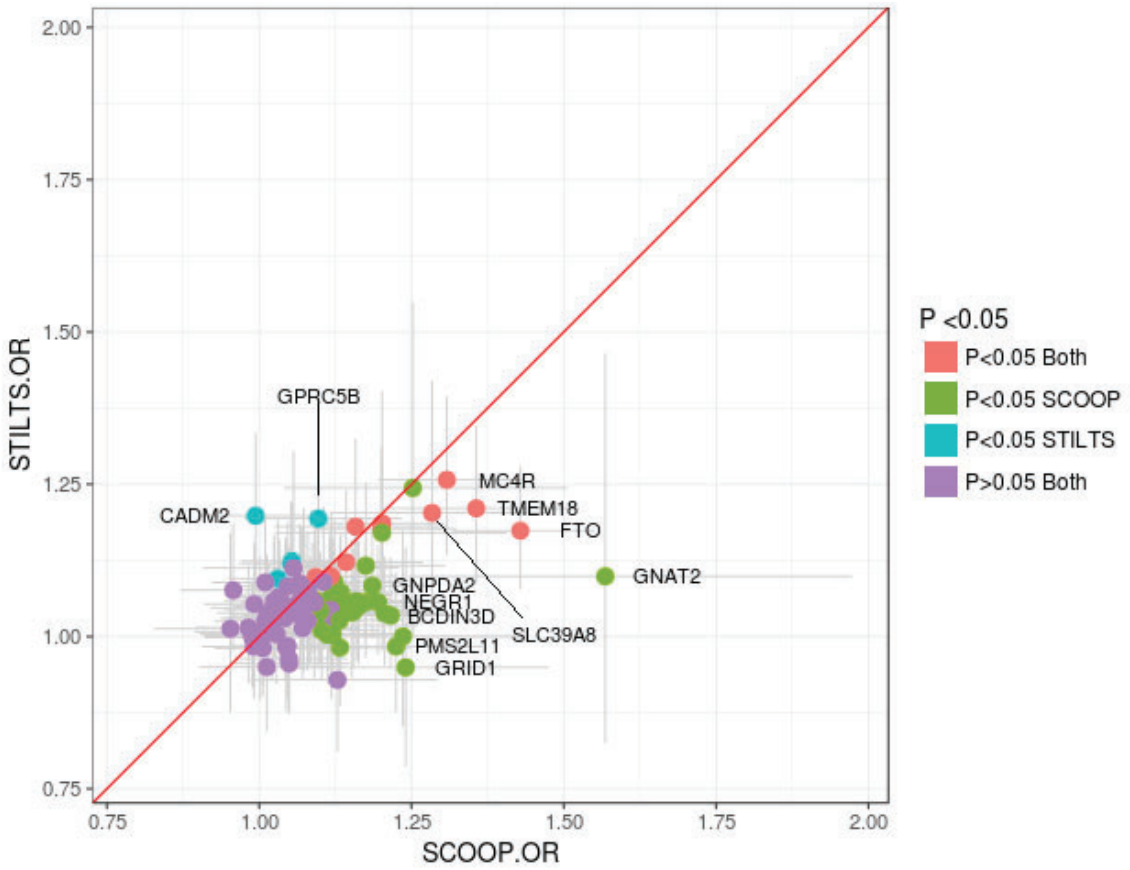

Figure 2.3: Odds ratio comparison for the 97 BMI associated loci. Odds ratios for SCOOP vs UKHLS (x-axis) and UKHLS vs STILTS (y-axis) comparisons are shown for the 97 known BMI loci from GIANT. Colours of data points represent nominal significance in both analyses (red), only SCOOP vs. UKHLS (green), 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).

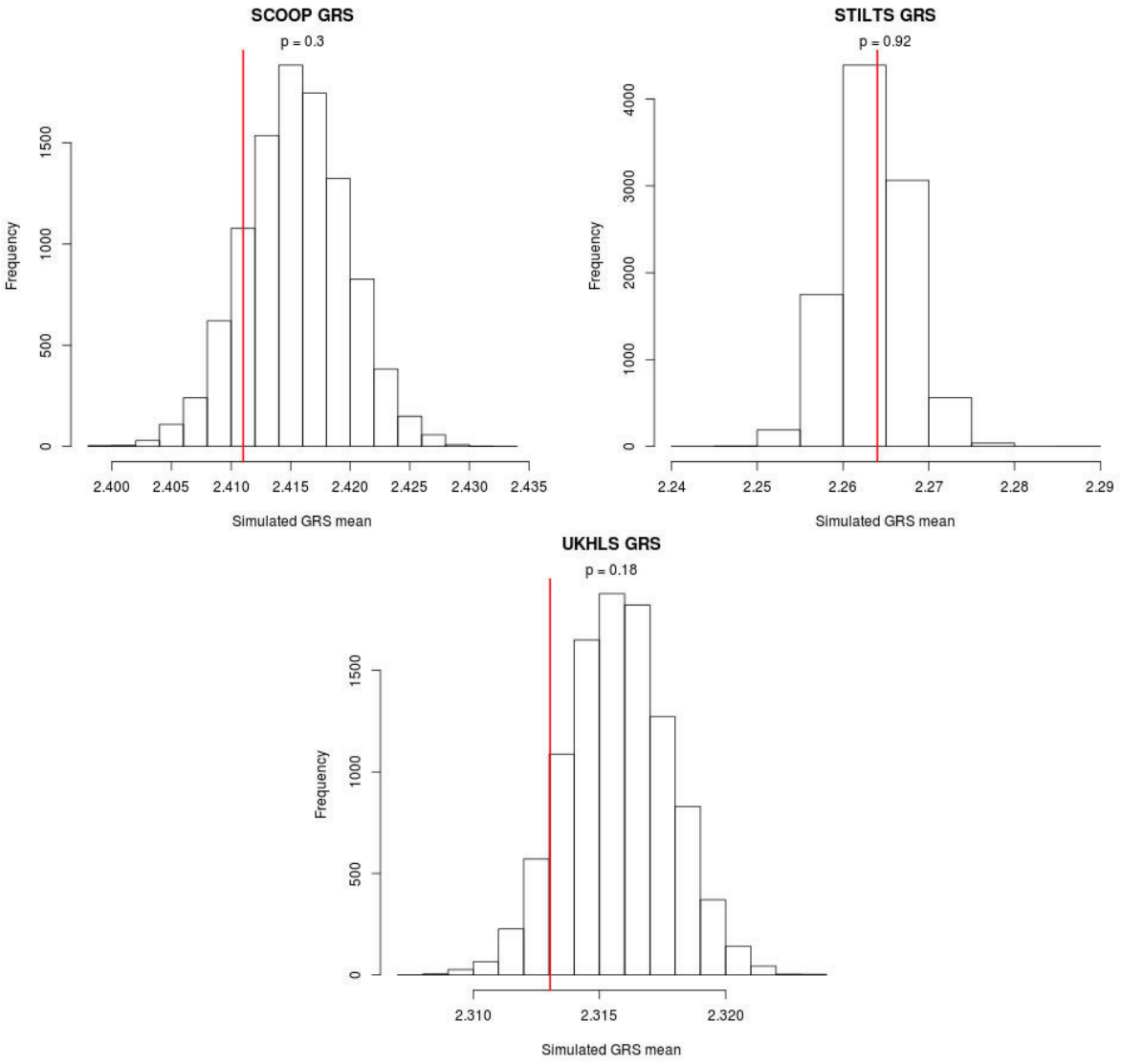

Figure 2.4: Mean GRS for SCOOP, STILTS and UKHLS compared to simulations. Histogram represents mean GRS scores for each BMI category across 10,000 simulations. Vertical red line highlights the observed value in real data.

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).

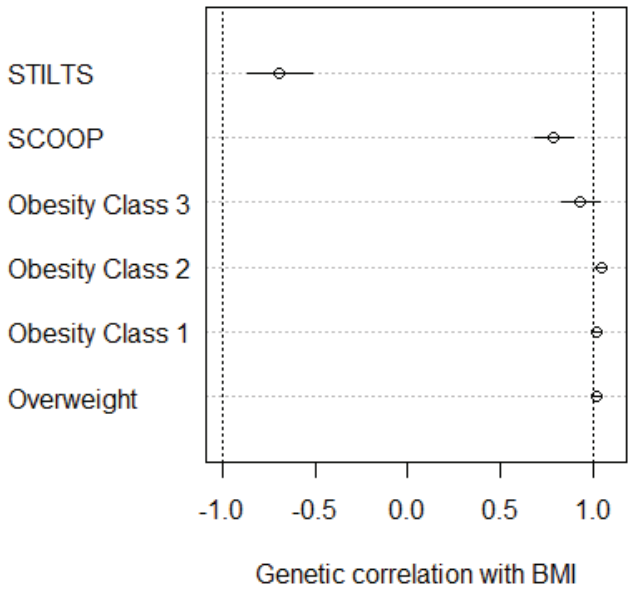

Figure 2.5: Genetic correlation of traits and BMI. Genetic correlation estimates and 95% CI for severe early-onset childhood obesity (SCOOP), healthy persistent thinness (STILTS), Obesity Class 3, Obesity Class 2, Obesity Class 1 and Overweight. Dotted lines represent complete genetic correlation.

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)**).

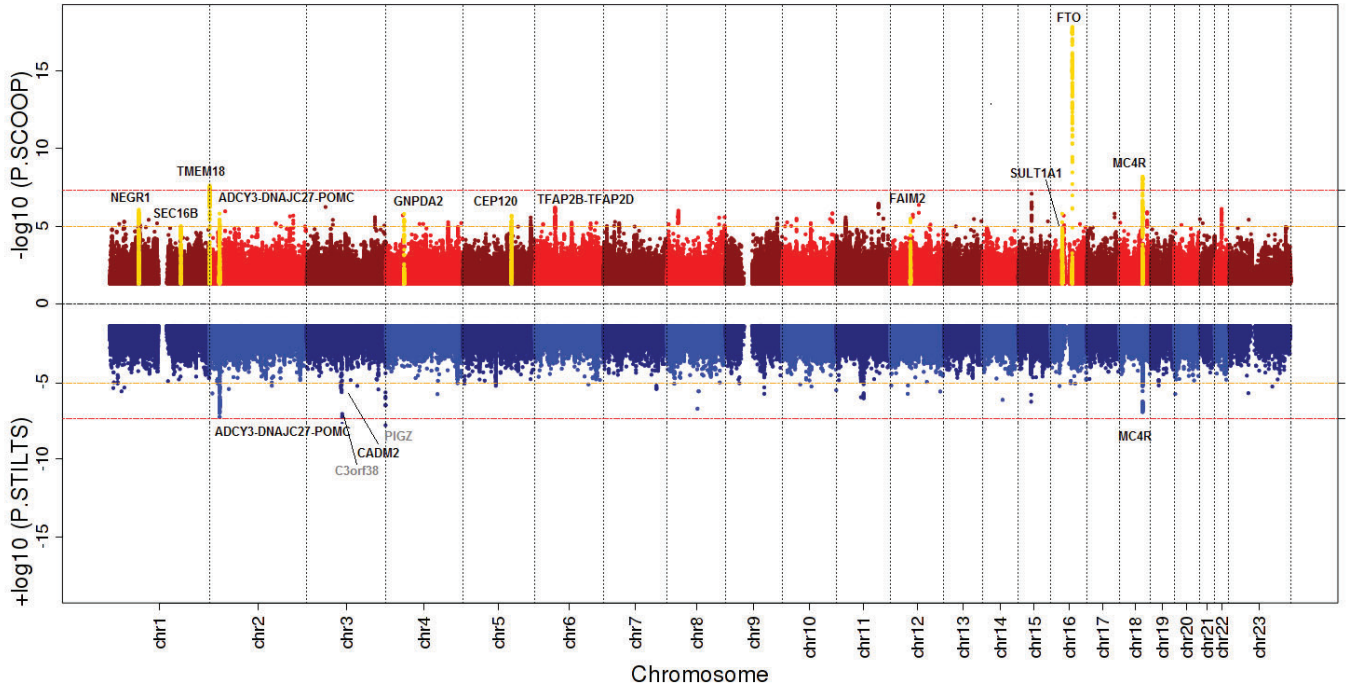

Figure 2.6: Miami plot of SCOOP vs. UKHLS and STILTS vs. UKHLS. Miami plot produced in EasyStrata [23], Red=SCOOP vs. UKHLS; Blue=STILTS vs. UKHLS. Red lines indicate genome-wide significance threshold at $p=5x10^{-8}$. Orange lines indicate discovery significance threshold at $p=1x10^{-5}$. Black labels 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).

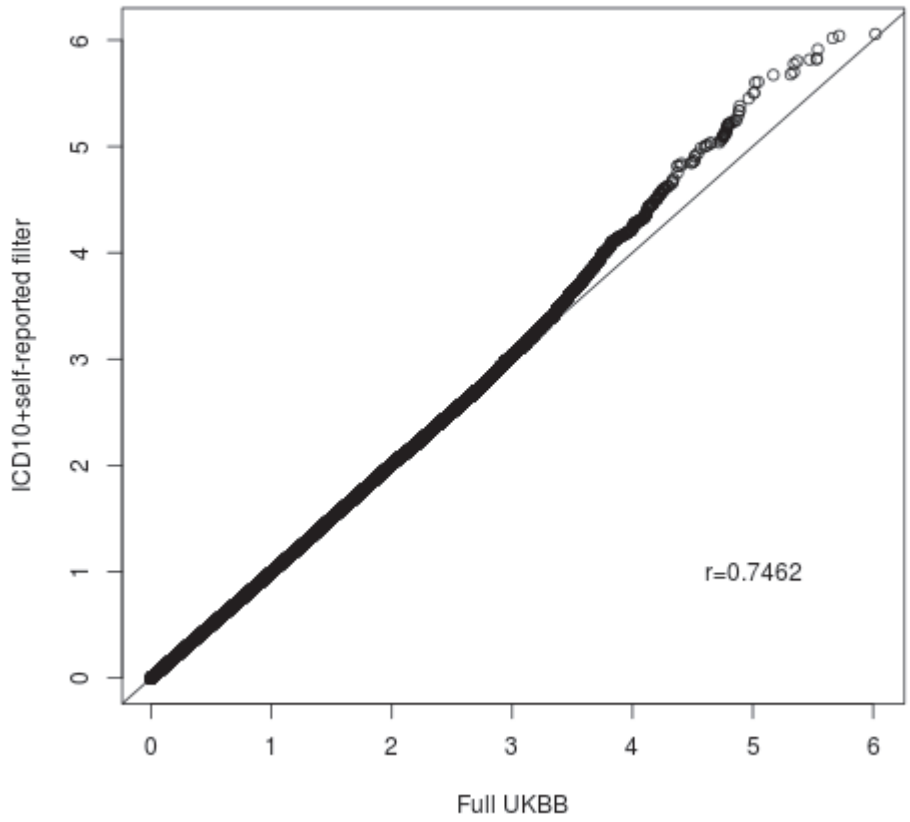

Figure 2.7: Quantile-quantile plots for UKBB case-control analysis with different exclusion criteria for thin individuals. Q-Q plot using all thin individuals as cases (Full UKBB) and removing individuals based on ICD10 and self-reported data (ICD10+self-reported filter). Correlation for –log10 p-values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