Early Life Determinants of Metabolic and Reproductive Health

Benjamin Carver Hollis

University of Cambridge
St. Edmund's College

December 2019

This thesis is submitted for the degree of Doctor of Philosophy

Declaration

This thesis is my own work and includes nothing which is the outcome of work done in collaboration except as declared in the ‘Contributions and Collaborations’ sections at the beginning of each chapter or as specified in the text.

No part of this thesis is substantially similar to any work that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution.

It does not exceed the limit of 60,000 words prescribed by the Degree Committee of the Faculties of Clinical Medicine and Veterinary Medicine.

Benjamin Carver Hollis MRC-Epidemiology Unit

Dissertation title: "Early Life Determinants of Metabolic and Reproductive Health"

Events in early life have consistently been associated with health outcomes in later life. The 'developmental origins' theory first hypothesised that adverse conditions in-utero can lead to physiological adaptations in the developing foetus which have long lasting influences on health. This concept has been extended to early childhood and adolescence, whereby exposures during critical periods of development can impact health throughout the life course of an individual. In particular, a secular trend for a decreasing age of puberty onset has been linked to the global increases in prevalence of cardio-metabolic diseases and cancer. It has been suggested that childhood obesity and lifetime sex hormone exposure may act as key mediating factors in this relationship. As the prevalence of obesity continues to rise globally and consequent comorbidities place an increasing burden on healthcare systems, understanding the mechanisms that link early life events to later life health have become of increasing importance.

While environmental exposures are often cited as being highly influential on growth and development, the role of genetics has become gradually more apparent in recent years. This has been aided by the availability of increasingly large data resources. Genetic studies have shown that many developmental traits are highly heritable and share genetic determinants with metabolic and reproductive health outcomes.

In this thesis I use data from large-scale, population-based resources to further elucidate the role of genetic and epigenetic factors in explaining observed associations between developmental traits and later life metabolic and reproductive health. I begin by examining the genetic aetiology of puberty timing in men, a key stage of sexual development which is understudied compared to women. I identify 29 novel genes involved in the control of puberty timing, implicating new biological pathways and demonstrate genetic correlations between earlier age of puberty and adverse health in adulthood. I then expand on the theme of genetic discovery by conducting genome-wide association studies (GWAS) for reproductive traits in the UK Biobank study. These outcomes have important societal and public health impacts but many have not previously been investigated from a genetic perspective. I identify over 800 variant-trait associations, highlighting genomic regions with highly pleiotropic influences on a diverse range of reproductive traits. This data is then leveraged to construct a framework for conducting phenome-wide association studies (PheWAS), which is used to explore the extent to which both BMI and sex hormone exposure act as mediating factors to explain the link between earlier puberty and heightened reproductive health risks.

I go on to examine mechanisms linking early life markers of development to adult health. I investigate the association between weight at birth and body composition in adulthood, determining that foetal and maternal-specific genetic determinants of birth weight have differential influences on fat and lean mass distribution. Downstream analyses suggest that these operate through distinct biological pathways, adding to our understanding of the association between low birth weight and poor health. Finally, I conduct an epigenome-wide association study (EWAS) as a complementary approach to GWAS for both BMI and puberty timing to identify additional genomic loci associated with both traits. Using EWAS data I present evidence for a casual epigenetic effect on puberty timing, functioning through both BMI-mediated and independent pathways.

The findings from this thesis contribute to the understanding of the genetic determinants of early life developmental processes and their relationship with later health, which have important implications for the improvement of individualised disease prevention and management.

Acknowledgements

This thesis is the culmination of three delightful, hard-working years at the MRC Epidemiology Unit, and I am grateful to have worked alongside a number of inspiring, brilliant and talented individuals whom I'm honoured to consider my peers and friends.

To my supervisor John Perry, thank you for giving me this opportunity and for your guidance throughout the process. Additionally to the members of the Growth and Development group, principally Ken Ong for his input and assistance with various aspects of this dissertation.

To my colleagues and collaborators, your knowledge and enthusiasm is greatly appreciated.

To the friends I've made along the way, you have made this experience ever more enjoyable and I will cherish our times spent together these last three years.

To my family, thank you for your unwavering support.

To Alice, for putting up with the ups and downs (mostly ups though), for your love and support that helps keep me going.

And finally to Felix – this dissertation might have materialised without you, but I choose not to imagine what it would have been like without your guidance, tutelage and friendship.

Publications

The following papers related to chapters of this thesis have been published or are in the process of submission for publication at the time of writing:

Hollis B., Day F.R., Busch, A.S., Thompson D., Soares A.L.G., Timmers P.R.J.H., Kwong A., Easton D.P., Joshi P.K., Timpson N.J., The PRACTICAL CONSORTIUM, 23andMe Research Team, Ong K.K., Perry J.R.B. Expanded genomic analysis for male voice-breaking highlights a shared phenotypic and genetic basis between puberty timing and natural hair colour. *Nat Commun.* 11; 1536 (2020).

Busch A.S., **Hollis B**., Day F.R., Sorensen K, Aksglaede L., Perry J.R.B., Ong K.K., Juul A., Hagen C.P. Voice break in boys – temporal relations with other pubertal milestones and likely causal effects of BMI. *Human Reproduction*. 34(8);1514-1522 (2019).

Ruth K.S., Day F.R., Tyrrell J, Thompson D.J., [6 authors], **Hollis B**., [6 authors], Murray A., Ong K.K., Frayling T.M., Perry J.R.B. Using human genetics to understand the disease impacts of testosterone in men and women. (*Accepted at Nature Medicine*).

Commonly Used Abbreviations

AAM	Age at menarche
AFB	Age at first birth
AFS	Age at first sexual intercourse
BMI	Body mass index
Bp	Base-pair
Chr	Chromosome
CI	Confidence interval
CVD	Cardiovascular disease
DEXA	Dual energy x-ray absorptiometry
DOHaD	Developmental Origins of Health and Disease
EA	Effect allele
EAF	Effect allele frequency
eQTL	Expression quantitative trait loci
GIANT	Genetic Investigation of Anthropometric Traits
GTEx	Genotype-Tissue Expression
FDR	False discovery rate
FH	Facial Hair
GRS	Genetic risk score
GWAS	Genome-wide association study
HES	Hospital Episode Statistics
HLA	Human leukocyte antigen
HRC	Haplotype Reference Consortium
HWE	Hardy-Weinberg equilibrium
IVW	Inverse variance weighted
LD	Linkage disequilibrium
MAF	Minor allele frequency
MB	Mega-base
MR	Mendelian randomisation
OA	Other allele
OR	Odds ratio
PC	Principal component
QC	Quality control
SD	Standard deviation
SE	Standard error
SNP	Single nucleotide polymorphism
T2D	Type 2 diabetes

UKBB……………………….. UK Biobank

VB…………………………….Voice breaking

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

1.1 The global burden of obesity

The excessive accumulation of body fat, manifesting in overweight and obesity, is among the most significant threats to global public health. In 2016, the World Health Organisation (WHO) estimated that worldwide 39% of adults over the age of 18 years were overweight as classified by a body mass index (BMI) greater than 25kg/m^2. Of these, 13% are classified as obese (BMI>30kg/m^2)[1]. In the UK it has been estimated that, among adults, 65% of men and 58% of women are either overweight or obese[2]. The trends for the prevalence of obesity in childhood are particularly striking, as globally an estimated 41 million children under the age of 5 and 340 million children aged 5 to 19 are overweight or obese, which represents a nearly ten-fold increase since 1975[3]. Among children in the UK, evidence suggests that as many as 1 in 3 of those aged 2 to 15 years are overweight[4].

These trends in the prevalence of obesity are mirrored by increases in comorbidities such as type 2 diabetes[5] (T2D), cardiovascular disease[6] (CVD) and particular cancer types[7]. The economic costs of obesity and its associated comorbidities are substantial. While estimates vary, it has been reported that the total cost of treatment for obesity and related conditions was $149.4 billion in the US in 2014[8], while the National Health Service (NHS) in the UK is estimated to have spent £6.1 billion on obesity-related ill-health in 2014-15[9]. The individual and societal costs demonstrate the immense scale of this issue and underscore the importance of identifying the mechanisms linking adiposity to disease processes in order to combat this epidemic.

1.2 Reproductive health and disease

Reproductive health encompasses a wide spectrum of traits and diseases, as evidenced by the WHO definition of reproductive health as:

> "a state of complete physical, mental and social wellbeing…in all matters related to the reproductive system and its function and processes"[10].

This definition comprises a broad and diverse range of health outcomes, including diseases of the reproductive organs and tissues, pregnancy-related outcomes, sexually transmitted infections (STIs) and violence against women. As such, the most significant reproductive health issues are likely to be geographically varied and population specific. In developing nations a substantial proportion of the overall burden of reproductive disease is from infectious diseases, with a disproportionate number of the approximately 37.9 million individuals suffering from HIV/AIDS residing in developing nations[11]. In contrast, in more developed countries chronic reproductive diseases have a considerably greater impact on public health. For example polycystic ovary syndrome (PCOS), one of the most common chronic reproductive health conditions in women, has an estimated prevalence of 5.5% in Caucasian populations[12] and its treatment is estimated to cost the NHS over £200 million annually[13]. Endometriosis, another common chronic reproductive condition which causes painful menstruation and infertility in approximately 1 in 10 women of reproductive age[14], has been estimated to have an annual economic impact of over £8 billion when factoring in treatment costs and work hours lost[15]. Furthermore, cancers of the reproductive organs represent a substantial proportion of cancer diagnoses, with breast (15% of new cases), prostate (13%), uterine (3%) and ovarian (2%) cancers consistently ranking among the types with the highest incidence in the UK[16] (**Figure 1.1**).

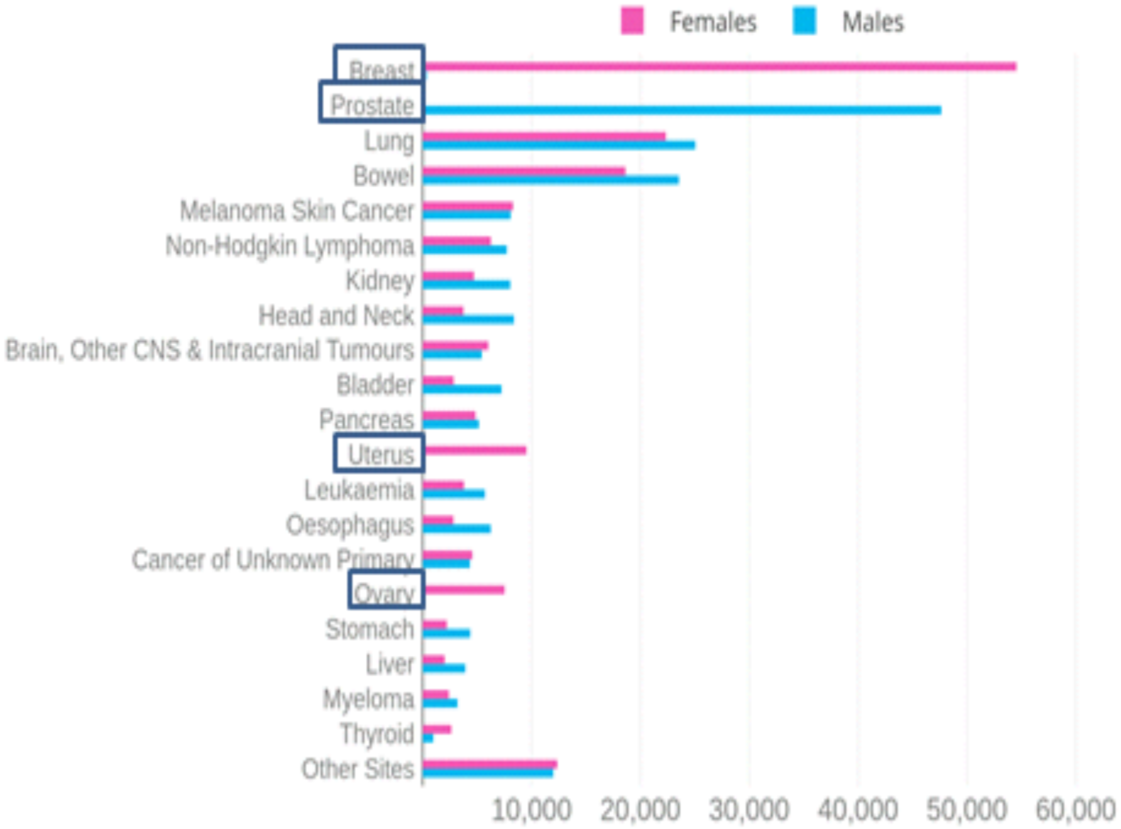


Figure 1.1: Incidence of the 20 most common cancer types in the United Kingdom in 2016. Reproductive cancers are highlighted in blue boxes. Source: Cancer Research UK (https://www.cancerresearchuk.org/health-professional/cancer-statistics/incidence/common-cancers-compared).

1.2.1 Obesity and reproductive health

Adiposity has a strong influence on the function of the reproductive system. This is due in large part to the fact that sex hormones are fat soluble, to the extent that adipose tissue is often considered to be an endocrine organ and is a major site for the storage and release of androgens and oestrogens (the principal sex hormones in men and women, respectively)[17]. Because of this, measures of adiposity have consistently been associated with reproductive traits and outcomes. Obese women are more susceptible to ovarian dysregulation, and have been shown to be at a nearly three-fold higher risk of infertility compared to women of normal weight[18,19]. Higher BMI is also a significant risk factor for PCOS, and weight reduction has been shown to improve pregnancy and general health outcomes in women suffering from the condition[20,21]. However the relationship between higher BMI and poor reproductive health is not consistent across all outcomes, as complex and condition-specific associations have been reported. For example there is evidence that higher BMI is protective against the risk of endometriosis, though the mechanisms for this remain unclear[22]. The complex relationship between BMI and reproductive health is further highlighted by cancers of the reproductive system. Higher BMI has been associated with increased risk of ovarian[23,24] and endometrial[25,26] cancers in populations of diverse ancestry. Conversely, other studies have presented evidence of heterogeneity of the effect of adiposity based on different clinical indicators. For instance higher BMI has been associated with a decreased risk of breast cancer in pre-menopausal women while increasing risk after menopause[27,28]. Similarly, certain subtypes of ovarian cancer have shown differential risks related to BMI[29]. Such observations highlight the need for caution when making inferences from observational studies, as the aetiological mechanisms often require more nuanced consideration.

1.3 Genetic approaches to the study of complex disease

Over the past decade, the role of genetic factors in complex disease aetiology has become increasingly clear. This has been made possible by the widespread use of genome-wide association studies (GWAS), which have demonstrated the association between common genetic variation in the population and disease risk[30]. What began with studies of a few thousand individuals genotyped on sparse arrays, capturing small fractions of the totality of genetic variants within the genome, has expanded dramatically. This has been enabled by large-scale population studies and increasingly efficient and cost-effective high-throughput DNA sequencing technologies[30]. Studies such as the UK Biobank cohort, which has collected comprehensive phenotype information and conducted genome-wide DNA sequencing and imputation of genetic variation in over 500,000 individuals, have been a primary driver of genetic discovery[31]. In addition, consumer-based genetic companies such as 23andMe have also contributed substantial amounts of data which has been made available to researchers[32]. By combining data from multiple sources in large consortia, recent GWAS have obtained sample sizes in excess of 1 million participants[33,34], providing statistical power to identify associations of small effect with great precision at large numbers of genetic loci. As a further advantage, in examining the combined effects of genetic variants across multiple studies the traditional requirement for replication of observed effects becomes unnecessary as this is built into the study design.

1.3.1 Genetic discovery for reproductive health traits

Large-scale GWAS have been conducted for several reproductive health traits of interest, identifying hundreds of associated genetic variants which have contributed to a deeper understanding of the biology of reproduction and reproductive disease. Among the best characterised traits are the timings of reproductive events including menarche and menopause, where GWAS have identified 389 and 106 associated genetic loci, respectively[35,36]. Downstream analyses based on these loci have been used to identify underlying biological pathways and confirm observational associations with a number of health and social outcomes. Reproductive cancers have also received much attention, with consortia oversampling disease cases to identify associated variants with high statistical power for discovery[37–39] (**Table 1.1**).

Table 1.1: Examples of large-scale GWAS for reproductive traits

Trait	Study	Sample Size	Associated Variants
Age at menarche	Day *et al.* (2017)[35]	368,888 (European)	389
Age at menopause	Perry *et al.* (2014)[36]	182,416 (European)	106
Endometriosis	Sapkota *et al.* (2017)[40]	208,641 (European)	14
PCOS	Day et al. (2018)[41]	113,238 (European)	14
Breast Cancer	Michailidou *et al.* (2017)[42]	256,123 (European and East Asian)	119
Prostate Cancer	Schumacher *et al.* (2018)[43]	140,306 (European)	163
Ovarian Cancer	Fehringer et al. (2016)[44]	123,671 (European)	26

1.3.2 Causal inference using genetic variants

Identification of genetic variants associated with health traits has enabled the development of techniques to make causal inferences about the effects of exposures on health outcomes. Mendelian randomisation (MR) analyses are based on the principle that alleles segregate randomly and independently to gametes during meiosis (i.e., Mendel's first and second laws of inheritance). If variants with a known association to an exposure of interest are used as instrumental variables, then this can be thought of as approximating a randomised control trial for the effect of the exposure on the outcome[45]. Since its conception, MR has become a commonly used and widely accepted method for causal inference in epidemiological studies. Throughout this thesis, MR is used to make such inferences regarding early life exposures and reproductive and metabolic outcomes. MR has various caveats and assumptions which must be considered, and extensions of the MR method which allow for more robust testing of observed associations are also used; these are discussed in Chapter 2.

1.4 Early life exposures as predictors of later-life health outcomes

The hypothesis that exposures in early life can manifest as disease in adulthood dates back to the 1980s, based on work from David Barker and colleagues which found a strong correlation between temporal and geographic rates of infant mortality and ischemic heart disease[46]. These associations lead to the development of the 'Foetal Origins of Disease' theory which postulated that stressors in the intra-uterine environment, such as maternal malnutrition, lead to adaptations by the developing foetus which can have long-lasting impacts on individual health[47,48]. This theory was derived largely from evidence in famine-exposed populations, in particular the 1944-45 Dutch Hunger Winter cohort[49,50]. Long-term follow up of these individuals indicated that children born to mothers during the last years of the second world war, a time of extreme food scarcity in the Netherlands, were at increased risk of multiple adverse cardio-metabolic outcomes including obesity, poor glucose tolerance[51] and high blood pressure[52]. These associations have subsequently been confirmed experimentally in animal models[53].

The Foetal Origins theory has since evolved to incorporate developmental exposures more generally. In its current form, the Developmental Origins of Health and Disease (DOHaD) hypothesis extends the theory from in-utero exposures to include exposures in infancy, childhood and adolescence[54,55]. The field of life course epidemiology has arisen from this, which is aimed at investigation of the biological, behavioural and social exposures of early life as determinants of adult disease[56].

1.4.1 Mechanisms for the 'Developmental Origins' theory

In recent years, the focus of studies related to DOHaD have shifted towards the identification of mechanisms which underlie the observed associations between early life exposures and adult disease. This research has highlighted the role played by epigenetic factors in health and disease, and is currently an area of significant interest in developmental biology[57–59]. Epigenetics refers to a group of covalent modifications to the structure of DNA which do not alter the underlying sequence of bases which make up the genetic code. Among these modifications are DNA methylation, histone modifications and RNA-mediating silencing. All of these mechanisms ultimately change the level of expression of genes in certain tissues, which can alter the activity of the encoded proteins and resulting in changes to physiological processes.

In this thesis, focus is placed on two early life exposures which have been identified as key determinants of disease risk in adulthood: birth weight and puberty timing.

1.5 Birth weight

Birth weight is an important indicator of pre-natal health, and is often used in epidemiological studies as an indicator of intra-uterine exposures. Birth weight is variable within the population and has been shown to be ancestry-specific. The range for 'normal' birth weight is typically considered to be between 2500g and 4000g in individuals of European ancestry, while lower cut-offs used for East and South Asian populations[60]. Birth weight is often reported relative to expected values for gestational age, as babies born pre-term (i.e. birth occurring at <37 weeks of gestation) are expected to be lighter on average than those born at full-term[61]. Thus, the terms 'small for gestational age' (SGA) and 'large for gestational age' (LGA) are often used to describe infants born outside the normal range relative to their expected size at birth (**Figure 1.2**).

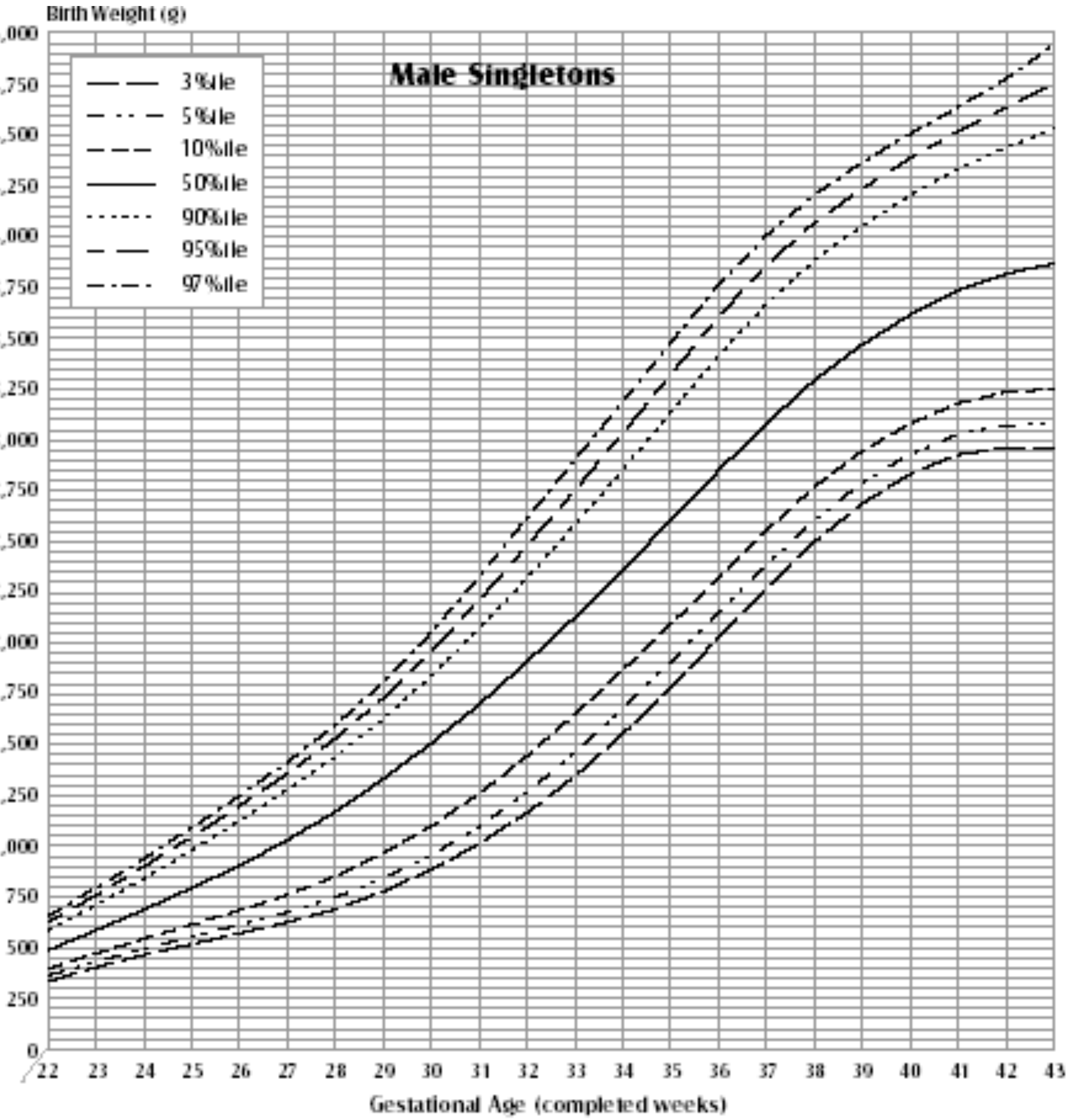


Figure 1.2: Centiles of weight for gestational age of male singletons in a European-ancestry population. Illustrative example showing that individuals born at a weight above the 90th centile are considered large for gestational age (LGA) while those born below the 10th centile are considered small for gestational age (SGA). Source: Canadian Perinatal Surveillance System (https://www.canada.ca/en/public-health/services/injury-prevention/health-surveillance-epidemiology-division/maternal-infant-health/birth-weight-gestational.html)

1.5.1 Environmental determinants of birth weight

As described above, maternal malnutrition was the impetus behind the development of the DOHaD hypothesis and has consistently been associated with lower offspring birth weight[62,63]. Initial observations were based on extremes of low caloric intake during pregnancy in famine-exposed mothers, and this continues to be an important public health issue particularly in low income countries[64,65]. However, even when food scarcity is not of primary concern reduced energy availability to the developing foetus can pose a threat to pre-natal health. As an example women with coeliac disease, a condition characterized by malabsorption of nutrients upon exposure to gluten, are much more likely to give birth to low birth weight babies when gluten is present in the diet[66]. Other associated environmental exposures include smoking[67] and maternal stress[68], both of which have been shown to contribute to low birth weight.

1.5.2 Genetic determinants of birth weight

Genetic determinants are also known to contribute significantly to birth weight. Twin studies have produced estimates for the foetal genetic contribution to singleton birth weight of between 25% and 53%[69–71]. However, the study of the genetic determinants of birth weight is complicated by the fact that both the foetal and maternal genome contribute this variation. Moreover, these genetic components are correlated (r~0.5), making it difficult to disentangle the specific genetic contributions. Recently, Warrington and colleagues published a study which performed GWAS on both own (foetal) and offspring (maternal) birth weight[72]. Combining these estimates in a meta-analysis, they identified over 300 variants which were associated with birth weight at a genome-wide significance level[72]. SNP-captured heritability, which is generally expected to be lower than estimates from twin studies due to insufficient statistical power to detect variants with small effect sizes and rare alleles not being adequately captured in the arrays[73], was estimated to be ~40%. Using structural equation models (SEM) to partition the effects into foetal and maternal specific components, they determined that the foetal genetic component accounts for the majority of this variance (~28%). Maternal variation, which may influence birth weight via pre-natal intra-uterine effects as well as influencing the post-natal environment, accounted for ~8% for the variation.

1.6 Puberty Timing

Puberty is defined as the transition from childhood to the age of physical and sexual maturity, where the body becomes capable of sexual reproduction. The pubertal process is a gradual one and its progress is often measured by distinct physiological markers. The onset of puberty is triggered by re-activation of the hypothalamic-pituitary-gonadal (HPG) axis, which enters a period of quiescence after early activity in the post-natal period[74]. Stimulation of increased oestrogen secretion (in females) and testosterone secretion (in males) leads to the development of secondary sexual characteristics. In females these include breast growth (thelarche), hip growth and the beginning of menstruation (menarche). In males, changes include growth of the testicles and penis, facial hair growth and voice deepening.

1.6.1 Measurement of puberty timing

Physiological changes associated with puberty occur gradually, often over the course of 4-5 years and in a characteristic temporal pattern. Marshall and Tanner first described the typical presentation of pubertal changes, and the Tanner scale which they developed is regularly used to chart progression through pubertal development[75,76]. The age at which puberty is said to have occurred is therefore somewhat arbitrary as it is dependent on the choice of marker. Several studies have attempted to measure puberty longitudinally in cohort studies, for example the Avon Longitudinal Study of Parents and Children (ALSPAC)[77] which has collected detailed follow-up of various developmental markers in up to 14,500 children over the course of 20+ years. This level of detail is not often possible to replicate however, particularly in large biobank-sized cohorts which often recruit participants many years after puberty has completed. The most common measurement of puberty timing in retrospective studies is age at menarche (AAM), a relatively late stage of puberty in women but one which is often well recalled[35]. In men, the age at voice breaking has previously been used as a reasonably well recalled marker for puberty timing[78].

1.6.2 Age at puberty

The timing of pubertal milestones is variable within the population and across ancestries[79,80]. Studies from across the globe have described a secular trend for decreasing age at menarche over the last century and more[81–84]. In the UK, the average AAM for women born between 1990 and 1993 was 12.3 years, compared to women born in the earlier part of the century who showed an average age of 13.5 years[81]. This represents a marked decrease over a relatively short time scale. There have been various explanations proposed for this trend, however it is commonly held that in recent years the lowering age at puberty is linked to an increased prevalence of childhood obesity[85,86]. The trend for earlier age at puberty is of concern for public health, as puberty has consistently been associated with risks for adverse health outcomes[87]. Several hypotheses have been suggested for this, including increased lifetime exposure to sex hormones and genetic overlap with BMI[35]. The identification and characterization of the mechanisms linking puberty to health risks is an area of on-going research.

1.6.3 Genetic determinants of puberty timing

Early genetic studies identified several rare variants which are associated with Mendelian disorders of puberty, and included mutations at the *MKRN3* and *KISS1*/*KISS1R* loci causing precocious puberty[88]. GWAS have greatly expanded our knowledge of the genetic determinants of puberty, with the most recent study on AAM revealing a highly heritable and polygenic architecture with 389 associated genomic loci[35]. Studies in men have been far less insightful, as the largest male-specific study to date identified only 14 associated loci[78]. Genetic correlations reveal a strong shared genetic architecture between puberty timing in men and women, suggesting that the large discrepancy between the number of associated loci is largely due to insufficient statistical power resulting from the much lower sample sizes of GWAS in men (~55,000 compared to ~370,000 in women).

1.7 Summary and areas of opportunity

The role of early life exposures in later life health has been an area of intense research and debate, and much evidence has been generated in support of the DOHaD hypothesis since it was first proposed. Much remains unknown however, particularly regarding the aetiological mechanisms underlying many of the observed associations. Important open questions related to the developmental origins of disease include:

1) What are the genetic determinants of puberty timing in men?
2) Does birth weight influence disease risk by altering body fat distribution?
3) Does DNA methylation causally influence BMI and puberty timing?
4) What are the genetic determinants of overall reproductive health, and to what extent are these shared between traits?
5) To what extent does puberty timing causally influence reproductive health outcomes, and how much of this can be explained by genetic associations with BMI and sex hormones traits?

Through the application of genetic methods, this thesis aims to address these questions, contributing to the understanding of the ways in which early life exposures influence the health outcomes, with a particular emphasis on reproductive health.

Chapter 2: Data sources and common methods

2.1 DATA SOURCES

2.1.1 The UK Biobank Study

Study design and recruitment of participants

The UK Biobank study is a large prospective cohort study, established in the UK in 2005 with the aim of investigating the genetic and non-genetic determinants of the diseases of middle and old-age[31]. Eligible participants were any UK resident aged between 40 and 69 years who were registered with an NHS GP service. Between 2006 and 2010, a total of 503,325 participants were recruited and attended one of 22 assessment centres across the UK for baseline data collection (**Figure 2.1**).

Informed consent was obtained from all participants during their visit to the assessment centres. Ethical approval was granted by the UK National Research Ethics Committee North West, in accordance with the principles of the World Medical Association Declaration of Helsinki.

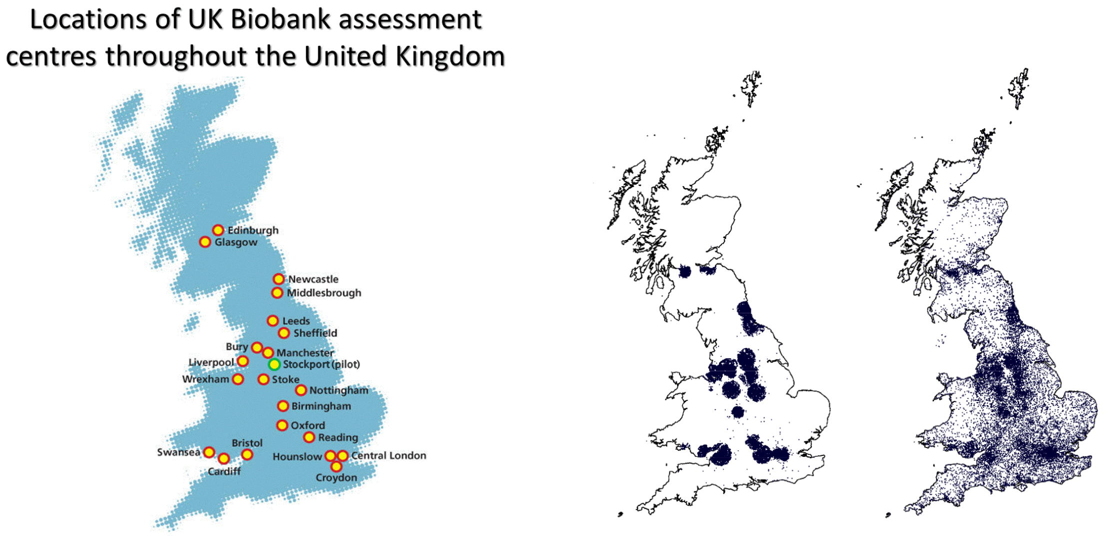


Figure 2.1: Location of UK Biobank assessment centres (left panel) and participants by current living address and place of birth (right panel). Sources: The UK Biobank (https://biobank.ctsu.ox.ac.uk/crystal/exinfo.cgi?src=UKB_centres_map) and Abdellaoiu *et al.*[89]

Baseline and follow-up assessments

All participants enrolled in the study attended an initial 2 to 3-hour assessment, which comprised completion of an online touchscreen questionnaire, a face-to-face interview and collection of physical measurements and biological samples. The touchscreen questionnaire was designed to collect information on current and past health, lifestyle health exposures (e.g. diet, smoking, alcohol consumption and physical activity), socioeconomic factors, psychological well-being and cognitive function. This was followed by a computer-assisted interview with a trained nurse to further assess personal health and family health history. Physical measurements included anthropometric measurements, hand grip strength, bone density, spirometry, arterial stiffness measurements, and physical fitness levels[90]. For each participant, 45mL of blood and 9mL of urine were collected. Blood and urine samples were stored in vacutainer tubes with appropriate anti-coagulant and preservatives added, and transported to a central laboratory facility for processing and storage[91]. Details on individual measurements relevant to specific studies in this thesis are described in the corresponding chapter.

In addition to data collected by researchers during the assessment centre visits, information on incident and prevalent healthcare events are available via linkage to NHS records. This includes Hospital Episode Statistics (HES) dating back to 1996, and more recently primary care data has been added.

Genotyping and imputation

DNA was extracted from whole blood buffy coat aliquot contained in EDTA-vacutainers[92] and used for genotyping of participants using one of two arrays. A subset of the cohort (N=49,950) were genotyped at 807,411 genetic markers using the UK Biobank Lung Exome Variant Evaluation (UK BiLEVE) from Affymetrix, which was enriched for variants related to lung function among heavy and non-smoking UK Biobank participants[93]. The remaining participants (N=438,427) were genotyped at 825,927 markers using the UK Biobank Axiom Array from Affymetrix. These two arrays share 95% of the same markers[94]. Genotyping was conducted in 106 batches of approximately 4700 samples per batch, resulting in genotype information on 489,212 participants at 812,428 bi-allelic SNPs and insertions/deletion (indels). Poor quality markers were excluded based on missing rate and heterozygosity adjusted for population ancestry. Further quality control exclusions included non-XX or XY sex chromosome karyotype. Following all quality control measures, the total sample comprised 488,377 and 805,426 markers.

Markers present on both genotyping arrays were taken forward for haplotype imputation. Markers that failed QC in multiple batches, had a missing rate >5% or a MAF<0.0001 were also removed. Phasing was then conducted on autosomes using SHAPEIT3 software[95] and 1000 Genomes phase 3 as the reference panel[96]. Imputation of haplotypes was performed using the Haplotype Reference Consortium (HRC) reference panel[97], with separate imputation using the combined UK10K[98] and 1000 Genomes phase 3 reference panels using IMPUTE4 software[99]. The final dataset comprised 93,095,623 autosomal markers and 3,963,705 X chromosome makers in 487,442 individuals. Reference SNP IDs (RSIDs) were obtained using the GRCh37 genome build.

2.1.2 The Fenland Study

Study design and recruitment of participants

The Fenland study is a population-based cohort study based in Cambridgeshire, UK, designed to investigate the genetic and lifestyle influences on cardio-metabolic disease[100]. Eligible participants were anyone born between 1950 and 1975 and registered with a GP practice in Cambridge at the time of recruitment. Between 2004 and 2014, 12,242 participants were recruited, with an overall response rate of 27%. Exclusion criteria were: clinical diagnosis of diabetes mellitus, inability to walk without assistance, terminal illness (≤1 year life expectancy), clinical diagnosis of psychiatric disorder, and current pregnancy or breastfeeding. Participants attended one of three MRC Epidemiology Unit testing centres for initial assessment. Written consent was given by all study participants, and the study was approved by the Cambridge Local Research Ethics Committee.

Baseline and follow-up assessments

Phase 1 of the study was completed in 2015, while Phase 2 began in 2014 and is ongoing. Participants attending testing centres completed questionnaires on lifestyle and general health, had clinical, anthropometric and physical health measurements taken, as well as wearing physical activity monitors. Body composition was assessed by dual energy x-ray absorptiometry (DEXA). Finally, blood and urine samples were collected from each participant for further metabolic assessment and genotyping.

Genotyping and imputation

Participants from the Fenland cohort were genotyped on one of three platforms. The majority of participants were genotyped using the Affymetrix Axiom Biobank chip (N=8,994), while a further 1,402 samples were genotyped using the Affymetrix SNP 5.0 chip. Additionally, 1,060 samples were genotyped using the Infinium Core Exome 24 version 1 chip. Genotypes were imputed using a combined reference panel comprising HRC, UK10K and 1000 Genomes phase 3 using IMPUTE2 software[101].

2.1.3 The EPIC-Norfolk Study

Study design and recruitment of participants

The European Prospective Study on Nutrition and Cancer (EPIC) is a prospective cohort study of over 500,000 individuals in multiple centres throughout western Europe and Scandanavia[102]. Primarily designed to investigate the role of diet in the development of cancer, the cohort collects data on multiple exposures and outcomes allowing for the investigation of environmental and genetic influences on chronic disease risk. EPIC-Norfolk is one of two UK-based sub-cohorts, which began recruitment of participants in Norwich and surrounding areas in 1993[103]. All individuals aged 40 to 79 years who were registered at a participating general practice surgery (N=35) were eligible, unless deemed unsuitable by the practitioner. Between 1993 and 1997, a total of 25,639 participants were enrolled in the study (overall response rate of 33%)[104].

Baseline and follow-up assessments

Participants attended an initial health check (1HC) upon recruitment where anthropometric and physical measurements were taken, lifestyle exposures were assessed by questionnaire and blood samples were collected[104]. Participants have been invited to follow-up assessments at 3 (2HC), 13 (3HC) and 20 (4HC) years after initial assessment where further data was collected, including DEXA scanning. Data on attrition rates have been published, showing that individuals who attended 3HC were generally younger, more physically active and of higher socioeconomic status than those who attended only 1HC[104], indicative of a healthy participant selection bias in follow-up assessments.

Genotyping and imputation

DNA obtained from blood samples was extracted using standard protocols. Genome-wide genotyping was performed using the Affymetrix UK Biobank Axiom array and imputed using a combined reference panel using HRC, UK10K and 1000 Genomes Phase 3, and imputed using IMPUTE2 software.

2.1.4 The 23andMe Study

Study design and recruitment of participants

23andMe (Mountain View, CA, USA) is a personal genomics company providing direct-to-consumer genetic testing. Participants provide saliva samples which are processed by the 23andMe research team, who provide reports which are made available to consumers on their personal ancestry and health. All participants have the option to provide informed consent to have their information used for health related research, with a protocol approved by Ethical and

Independent Review Services, accredited by the Association for the Accreditation of Human Research Protection Programs[78].

Genotyping and imputation

The methods for DNA extraction from saliva samples has been described previously[105]. Genotyping was performed across one of four platforms. The V1 and V2 platforms genotyped approximately 560,000 SNPs using a custom Illumina HumanHap550+ BeadChip (Illumina, CA, USA). The V3 platform genotyped ~950,000 SNPs using the Illumina OmniExpress+ BeadChip, while the V4 platform uses the same chip as V3 with a less dense coverage of ~560,000 SNPs[106]. For the analyses in this thesis, imputation was performed using the 1000 Genomes phase 3 reference panel after exclusion of SNPs with Hardy-Weinberg equilibrium ($P < 10^{-20}$), a call rate <95% or with excessive allele frequency discrepancies compared to 1000 Genome reference panels (described in Day *et al.*)[78].

2.2 COMMON METHODS

2.2.1 Linear mixed models

Population stratification and cryptic relatedness can confound associations between genetic variants and phenotypes, and can bias GWAS estimates when not adequately controlled for in genetic discovery. Commonly, principal component analysis (PCA) is used to mitigate these effects with genetically determine PCs included as covariates in linear regression models[107]. However, these methods are not fully robust to fine-scale population structure and genetic relatedness among participants. Linear mixed models (LMMs) are now predominantly used in genetic discovery in order to explicitly control for population stratification and cryptic relatedness. For the genetic discovery analyses in this thesis, LMMs were implemented using BOLT-LMM (v2.3.4)[108]. Additive per-allele effects were assumed and effect estimates and P-values from the infinitesimal model were used, which makes the implicit assumption that all variants are causal with small, normally distributed effect sizes.

By default, BOLT-LMM uses a linear model for both quantitative and categorical traits (i.e. it does not perform logistic regression). Therefore for binary and categorical outcomes, log odds ratios and standard errors are obtained from the linear regression estimates using the following formulae:

$$\log OR = \beta/(\mu \times (1-\mu)) \qquad [1]$$

$$SE_{OR} = SE/(\mu \times (1-\mu)) \qquad [2]$$

2.2.2 Fixed-effects meta-analysis and heterogeneity estimates

Combination of GWAS summary statistics was performed using fixed-effects inverse-variance weighted meta-analysis implemented in METAL[109]. In addition, METAL was used to calculate heterogeneity of effect sizes between studies based on the heterogeneity I^2 statistic and P-value produced by the software.

2.2.3 Estimating genetic linkage

Where it was necessary to calculate the degree of linkage disequilibrium (LD) between genetic variants, estimates were obtained based on reference values from 1000 Genomes Phase 3 European samples, implemented in LD Link (versions 3.2 to 3.7.2)[110]. In addition, regional association plots were generated using LocusZoom using data from 1000 Genome phase 1 Europeans as reference[111].

2.2.4 Gene-set enrichment analysis using MAGENTA

Gene-set enrichment analysis (GSEA) allows for further characterisation of the functionality of GWAS hits by examining whether genome-wide trait associations are enriched for genes within defined biological pathways. In this thesis, Meta-Analysis Gene set ENrichment of varianT Associations (MAGENTA v2.4)[112] was used to assess enrichment based on GWAS summary statistics in pre-defined pathways (gene sets). MAGENTA assigns a score to each gene in the genome (gene scores) by assigning a single index SNP which is defined as the autosomal SNP with the lowest P-value within a 140kb window. Scores are adjusted for gene size, SNP density and LD and then ranked. For each pathway, the number of gene scores which rank in the 95th and 75th percentile for all genes in that set are counted and compared to the expected number, which is determined from a distribution derived from 1 million randomly-permuted pathways.

Gene sets were obtained from five different publicly available resources: Gene Ontology (GO), KEGG, PANTHER, Reactome and BioCarta. Pathways significance was determined based on a false discovery rate (FDR) ≤0.05 at either the 75th or 95th percentile.

2.2.5 LD score regression

LD score regression (LDSC) is a technique developed by Bulik-Sullivan and colleagues[113] which can be applied to genome-wide summary statistics to distinguish polygenic effects from confounding factors such as population stratification. This is achieved by computing population-specific LD scores and regressing these against GWAS summary statistics for each variant. Extensions of the method allow for inferences to be made on SNP-based heritability and genetic correlations (r_g) between traits[113]. LDSC is applied throughout this thesis, using both the stand-alone software implemented in Python and the LD Hub database, a web-based application which has compiled GWAS summary statistics from multiple sources[114]. Study specific applications are described in the appropriate chapters. In all cases, only variants available in HapMap Phase III[115] were included in LDSC analyses.

2.2.6 Mendelian randomization

Mendelian randomization (MR) was first proposed by Katan[116] as a method for controlling for reverse causation and confounding, which provide a significant barrier to casual inference in traditional epidemiological studies. The idea is based on (and named after) Mendel's first two laws of inheritance, whereby alleles segregate during meiosis (first law) and are independently assorted into gametes (second law). If the allele is associated with a risk factor of interest, this can be thought of as a natural randomized trial, and because the haplotype is generally not modifiable this circumvents the problem of reverse causation. The widespread use of MR was championed by Davey-Smith and Ebrahim[45], and since the advent of GWAS it has been used extensively in genetic epidemiology as a tool for making causal inference. Many extensions of the method have been developed, all of which are fundamentally based on the use of genetic variants as instrumental variables for the exposure (**Figure 2.2**). MR analyses rely on three key assumptions: 1) the genetic instrument must be associated with the risk factor; 2) the genetic instrument must not be associated with any confounders; and 3) the genetic instrument must not be independently associated with the outcome except via the risk factor of interest.

This thesis makes extensive use of the two-sample MR technique first described by Pierce and Burgess[117], whereby genetic risk scores obtained from GWAS summary statistics for the exposure of interest are investigated for their effects in an independent sample measuring their effects on the outcome. Primary MR analyses employ the inverse-variance weighted (IVW) regression model, where the regression of the effect of each variant on the outcome is weighted by the precision (inverse of the variance of the effect estimate) on the outcome.

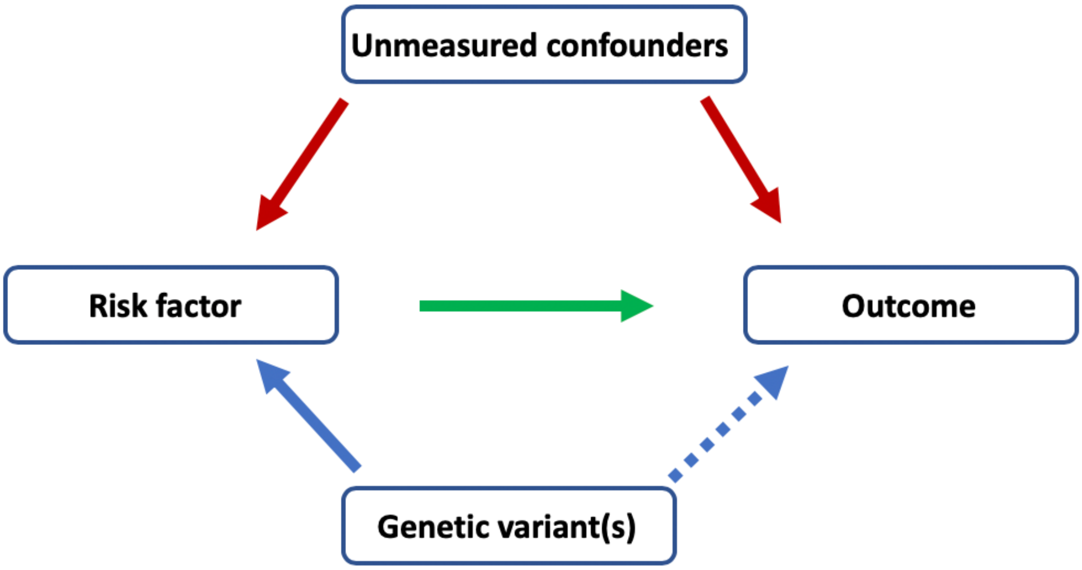


Figure 2.2: Schematic representation of Mendelian randomization theory. The desired value for the association between the risk factor and the outcome (green arrow) may be confounded by unmeasured latent factors (red arrows). Genetic variants which are associated with the risk factor may serve as instrumental variables for the exposure (solid blue arrow), and their effect on the outcome (dashed blue arrow) can provide evidence for a causal association with the outcome provided that the MR assumptions are met.

Sensitivity analyses

Various methods to test the robustness of MR associations have been proposed and are used in the literature. These methods are designed to account for violations of the assumptions of MR. Horizontal pleiotropy may arise if genetic variants are associated with the outcome via pathways independent of the outcome. If the combined pleiotropic effects of a polygenic risk score are not balanced (i.e. they do not sum to zero – so called 'directional pleiotropy') then the second MR assumption is violated and the causal estimate from the IVW method may be biased away from the null[118]. To mitigate these effects, MR-Egger regression is applied as a sensitivity analysis to test for the presence of directional pleiotropy which may bias the results of MR[119]. MR-Egger allows for the possibility of multiple invalid instruments by not forcing the regression line to pass through the origin. While MR-Egger has been shown to be generally underpowered compared to IVW methods[120], an y-axis intercept which is significantly different from zero is indicative of pleiotropic effects and suggests caution must be taken with the interpretation of results of the IVW estimate. In these cases, the use of the MR-Egger estimate may be more appropriate.

Additional sensitivity analyses are also commonly employed. These include the calculation of the Cochrane's Q statistic and P-value, where significant values (P<0.05) provide further evidence of horizontal pleiotropy. In addition weighted median approaches have been developed, providing an alternative method to MR-Egger by weighting effect estimates towards the median value to account for pleiotropic effects of variants[121]. In this thesis a further newly-developed method, MR-PRESSO[122], is used in several instances. MR-PRESSO detects and accounts for the presence of horizontal pleiotropy by removing outlier variants from the analysis, while providing an estimate for the distortion of the results caused by the outliers before their removal.

Multivariable MR

A further method to account for the pleiotropic effects of SNPs is multivariable MR. In instances where one or multiple variants have a causal influence on the outcome via multiple pathways, including the predicted effects of the variant for each risk factor as covariates in the same model can allow the effects to be partitioned across the pathways[123]. This allows for the assessment of the direct, and theoretically unconfounded, effects for the risk factor of interest to be obtained. Similar assumptions as apply to univariate MR analyses also apply here, only in this case the genetic variant may be associated with multiple risk factors under study[124].

Chapter 3: Genetic determinants of male puberty timing

Contributions and collaborations

All analyses described in this chapter were conducted by me with the exception of the following: Dr. John Perry (MRC Epidemiology Unit, University of Cambridge) conducted the GWAS for the UK Biobank phenotypes (early and late voice breaking; early and late first facial hair) while summary statistics for age at voice breaking were provided from the 23andMe Research Team (23andMe Inc., Mountain View, CA); replication of signals using the genetic risk scores in ALSPAC was conducted by Dr. Ana Goncalves Soares (MRC Integrative Epidemiology Unit, University of Bristol); Dr. Deborah Thompson (Centre for Cancer Genetic Epidemiology, Department of Public Health and Primary Care, University of Cambridge) conducted the prostate cancer MR analyses; and Dr. Felix Day (MRC Epidemiology Unit, University of Cambridge) performed the phenotypic regressions with hair colour in UK Biobank. In addition, Professor Ken Ong (MRC Epidemiology Unit, University of Cambridge) assisted with the annotation and interpretation of implicated genes and non-synonymous variants.

SUMMARY

Puberty represents a critical window of development, and the timing of puberty is highly variable and is associated with a number of health outcomes in later life. To date, most of our understanding of the genetic control of puberty timing is based on studies in women, as age at menarche is clearly defined and often well recalled. This chapter reports a multi-trait genome-wide association study meta-analysis for puberty timing in men, based on recalled timing of voice breaking and first appearance of facial hair. With an effective sample size of 205,354 men, this represents a nearly four-fold increase from the largest previously reported. 76 independent signals for male puberty are identified, including 29 signals not previously associated with puberty in either sex. Earlier puberty timing in men is shown to be genetically correlated with several adverse health outcomes, and Mendelian randomisation analyses provide evidence for a causal relationship between earlier puberty and higher risk of shorter lifespan. In addition, a novel mechanism linking puberty timing to natural hair colour is reported, possibly mediated by the shared influence of pituitary hormones on both traits. These findings demonstrate the value of sex-specific studies of puberty timing, and serve as an important example of how early-life developmental exposures can affect long-term health.

3.1 BACKGROUND

3.1.1 Puberty: Physiological changes and triggers

Puberty describes the transition from childhood to the age of physical and sexual maturity, where individuals become capable of sexual reproduction. Changes in levels of sex hormones, triggered by activation of the hypothalamic-pituitary-gonadal axis, lead to physiological changes and development of secondary sex characteristics. In boys these are primarily brought on by increased exposure to androgens, particularly testosterone, and include increase in testes size, growth of skeletal muscle, growth of armpit and facial hair, and deepening of the voice due to elongation of the larynx[125]. In girls, increased production of oestrogens leads to breast and ovary development, changes in body morphology and fat distribution[126], and the beginning of menses.

The duration of puberty varies between individuals, and is associated with the age of onset[127]. Progression through various stages of puberty may be measured by several means, however the Tanner scale is the most commonly used metric. This divides pubertal progression into five stages, based on testicular enlargement in boys, breast development in girls, and pubic hair growth for both sexes[75,76].

3.1.2 Puberty timing and health: population trends and disease associations

The triggers for puberty onset are an area of intense scientific interest. Puberty timing varies widely in the population, and is determined by a combination of environmental, nutritional, genetic and social factors[128–130]. Population studies on puberty timing describe a secular trend of decreasing age of onset worldwide, with the average age at menarche (AAM) decreasing from approximately 17 years in the early nineteenth century to a 12.5 years today[131]. Explanations for this marked decrease include endocrine disrupting chemicals[132], chronic stress[133] and dietary changes[134]. Notably, increased body fat, as approximated by BMI, has consistently been associated with early puberty[88,135–137] and it is widely accepted that the increasing prevalence of childhood obesity is likely to explain much of this trend.

The decreasing age of puberty onset has important implications for public health, as early puberty timing has been shown to increase the risk of many cancers, cardiovascular and metabolic conditions, as well as gynaecological, gastrointestinal and musculoskeletal disorders[87]. It is therefore of importance to understand the aetiology and mechanisms controlling pubertal onset, as this will inform public health policy and clinical decisions on management strategies and targeted interventions.

3.1.3 Genetic studies of puberty timing

Early studies on the genetic control of puberty timing largely focused on extreme cases of precocious or delayed puberty, and identified rare variants causing Mendelian disorders[128]. More recently, large-scale population-based studies using genome-wide methods have elucidated more of the genetics underlying common variation in puberty timing, and indicate a highly heritable and polygenic architecture. The largest such study to date, with a sample size of over 370,000 women, identified 389 independent signals for AAM in women while estimating the overall SNP-based heritability of the trait to be approximately 32%[35]. Downstream analyses identified gene set enrichment in retinoic pathways (among others), and provided robust evidence for causal links between early puberty and poor adult health, corroborating observational evidence.

Large-scale genetic studies on puberty timing conducted to date have primarily focused on women, due to the fact that AAM is often a more memorable and easily recalled marker for puberty timing than male pubertal milestones, and is therefore regularly recorded in large cohorts. The largest male-specific study on puberty genetics was conducted in a considerably smaller sample of ~55,000 men and used recalled age of voice breaking as a marker for puberty onset[78]. The study identified 14 independently associated signals for male puberty timing, and individual effect sizes were strongly correlated with AAM effects (r_g=0.74), providing strong support for voice breaking as a valid maker for puberty.

Despite the strong overall genetic correlation between markers for voice breaking and AAM, notable exceptions were also identified with stronger effects in one sex compared to the other. The most striking example occurs at the *SIM1/MCHR2* locus, where the puberty delaying allele in females is associated with earlier puberty in men. Furthermore, studies in mice have identified sex-specific patterns of expression for puberty-associated genes[138].

3.1.4 Areas of opportunity

The lack of large-scale genetic studies for puberty timing in men, along with the examples of variants with sex-specific effects, highlights the need for further research into male puberty genetics. The UK Biobank study represents one of the best resources available for genetic analysis, containing data for over 500,000 individuals with genetic and broad phenotypic data in their second release[31]. This includes information on two male-specific puberty markers: relative age at voice breaking and relative age at first facial hair. In this chapter I describe a study which incorporates this information with the previously analysed data on voice breaking from the 23andMe study, employing a method which allows for meta-analysis of related traits from potentially overlapping samples. This expanded analysis substantially increases existing knowledge on the genetic architecture of male puberty and provides strong evidence in support of puberty timing as an important early life risk factor for health and disease outcomes.

3.2 METHODS

3.2.1 Participating studies and assessment of puberty timing phenotypes

Age at voice breaking was determined in the 23andMe study by response to the question: "How old were you when your voice began to crack/deepen?", administered in an online questionnaire. Male participants chose from one of seven pre-defined age bins: under 9 years, 9-10 years, 11-12 years, 13-14 years, 15-16 years, 17-18 years, or 19 years or older. These were then re-scaled to one year age bins using a previously validated method[139]. In the UK Biobank study, timing of voice breaking and of first appearance of facial hair were determined by responses to the touch-screen questions "When did your voice break?" and "When did you start to grow facial hair?", respectively. Participants were required to choose from one of five possible options for both questions: younger than average, about average, older than average, do not know, or prefer not to answer.

3.2.2 Genome-wide association studies for puberty timing phenotypes

Details on genotyping and imputation methods for the 23andMe and UK Biobank studies are described in Chapter 2. In the 23andMe study, genetic and phenotypic data were available in up to 55,871 unrelated men across 13.7 million imputed variants. Genetic associations with age at voice breaking were obtained by linear regression models, including age and five genetically determined principle components as covariates to account for population structure, with additive allelic effects assumed. P-values for SNP associations were computed using likelihood ratio tests.

In the UK Biobank study, genetic and phenotypic data were available in up to 191,270 men for age at voice breaking and 198,731 men for age at first facial hair, across 18.5 million variants. For this analysis, we restricted our sample to unrelated individuals of self-reported white European ancestry. Respondents who answered either 'older than average' or 'younger than average' for each phenotype were compared in separate models using logistic regression, with the 'about average' group used as a reference in each case. Effect estimates for genetic variants were obtained using regression models implemented in BOLT-LMM v2.2 (as described in Chapter 2), with age, genotyping chip and ten principle components included as covariates.

3.2.3 Multi-trait meta-analysis for puberty timing in men

GWAS summary results for each of the five strata (23andMe age at voice breaking; UK Biobank relatively early and relatively late voice breaking; and UK Biobank relatively early and relatively late first facial hair) were meta-analysed using Multi-Trait Analysis of GWAS (MTAG), developed by Turley *et al.*[140]. MTAG uses GWAS summary statistics from multiple correlated traits to effectively increase sample size and statistical power to detect genetic associations, by estimating a variance-covariance matrix to correlate the effect sizes of each trait using a moments-based method. In addition, MTAG calculates a variance-covariance matrix for the GWAS estimation error using LD score regressions. SNP effect estimates are then calculated for each trait using a moments-based function in a generalisation of standard inverse-variance weighted meta-analysis. Effect estimates and P-values can then be used in downstream analysis, provided three key assumptions are met:

1. The variance-covariance matrix for effect sizes for all SNPs is homogeneous across all traits;
2. Sampling variation can be ignored; and
3. Sample overlap is adequately captured.

Violation of the assumption of homogeneity of effect sizes may be plausible in this setting if some SNPs have an effect on voice breaking but not on facial hair (or vice-versa). Therefore the upper bound for the false discovery rate (maxFDR) was calculated, as recommended by the MTAG authors. Prior to meta-analysis, we removed rare variants (MAF<0.01). In addition, due to the large difference between the number of cases and controls in the four UK Biobank strata, we calculated the effective sample size using the equation:

$$N_{eff} = \frac{2}{\frac{1}{N_{cases}} + \frac{1}{N_{controls}}}$$

Effective sample sizes for early and late voice breaking were 15,711 and 21,217, respectively and for early and late facial hair were 17,391 and 23,011, respectively. For our genetic discovery and subsequent analyses, we chose effect estimates from the 23andMe study as the base trait so that effect estimates were on a continuous scale.

3.2.4 Identification of independent and novel loci and gene annotation

The genome-wide P-value threshold of $P<5 \times 10^{-8}$ was used to determine statistically significant SNP associations. Independent signals were identified using a combination of distance-based clumping and linkage disequilibrium metrics: the SNP with the lowest P-value within a 1 MB window which was not in LD ($r^2<0.05$) with another genome-side significant SNP was considered the association signal at that locus. For each independent locus, all previously reported AAM and voice breaking loci within 1 MB of that variant were identified. A locus was considered novel if there were no previously reported puberty loci (for either AAM or VB) within 1 MB, or if any previously reported loci within 1 MB were not in LD with it (r^2<0.05). Gene annotation was performed using a combination of methods. Information on the nearest gene, as well as all non-synonymous variants in LD with each signal, was obtained from HaploReg v4.1 (https://pubs.broadinstitute.org/mammals/haploreg/haploreg.php). In addition, other genes in the region were visualised using LocusZoom. The most likely causal variant was determined by combining this information with existing knowledge.

3.2.5 Replication of male puberty timing signals in ALSPAC

In collaboration with colleagues at the University of Bristol, replication of the identified signals was sought in an independent cohort. The Avon Longitudinal Study of Parents and Children (ALSPAC) recruited pregnant women resident in the Avon area of the UK with an expected delivery date between 1st April 1991 and 31st December 1992[77]. Since then, mothers, partners, and offspring have been followed up regularly through questionnaires and clinical assessments. The offspring cohort consists of 14,775 live-born children (75.7% of the eligible live births). Full details of recruitment, follow-up and data collection have been reported previously. Ethical approval for the study was obtained from the ALSPAC Ethics and Law committee and the Local Research Ethics Committees. A series of nine postal questionnaires regarding pubertal development were administered annually from the time the participant was aged 8 until the age of 17. The questionnaires, which were responded to by either the parents or the participant, had schematic drawings and verbal descriptions of secondary sexual characteristics (genitalia and pubic hair development) based on the Tanner staging system, as well as information on armpit hair growth and voice change. Age at voice change was considered the age at which the adolescent reported his voice to be "occasionally a lot lower or to have changed completely". Weight and height were measured annually up to age 13 years, then at ages 15 and 17 years by a trained research team. Age at peak height velocity (PHV) was estimated using Superimposition by Translation And Rotation (SITAR) mixed effects growth curve analysis[141].

A genetic risk score (GRS) was calculated based on all available SNPs identified in the multi-trait GWAS, weighted by the effect size for that SNP as reported in the ReproGen Consortium AAM GWAS. The GRS was standardised, and results presented are the increase in the phenotype per standard deviation increase in the GRS. Linear (continuous phenotype) and logistic (binary phenotype) regression analyses were performed in models both unadjusted and adjusted for age and ten genetically-determined principle components, with the except for age at PHV, age at voice change and age of first armpit hair which only controlled for PCs.

3.2.6 Heterogeneity in puberty timing effects between sexes

Lookups of summary statistics for independent loci identified in the male puberty meta-analysis were obtained from AAM GWAS data from the ReproGen consortium[35]. In addition, summary statistics for the 389 independent AAM loci were obtained from the MTAG meta-analysis. Heterogeneity of effect sizes for both sets of variants was calculated from the I^2 statistic and P-value, implemented in METAL[109].

3.2.7 Gene expression analysis

To determine tissue-specific expression of genes identified in the male puberty meta-analysis, data from the GTEx project[142] was incorporated with our summary statistics. Transcription levels of genes in 53 different tissue types were investigated, using a Bonferroni-corrected threshold of $P<9.4\times10^{-4}$ (=0.05/53) to determine significance.

3.2.8 Gene set enrichment analysis

Genome-wide meta-analysis summary statistics for male puberty timing were investigated for gene-set enrichment in known biological pathways using MAGENTA[112]. Testing was performed on 3,216 pathways from four databases (PANTHER, KEGG, Gene Ontology and Ingenuity), with significance determined based on an FDR<0.05 for genes in the 75th percentile.

3.2.9 Genetic correlations between male puberty timing and health-related traits

LD score regression[113] was used to determine genetic correlations (r_g) using meta-analysis summary statistics for male puberty timing and 751 health-related traits which were publicly available from the LD Hub database[114].

3.2.10 Mendelian randomisation analyses

In order to determine the likelihood of a causal effect of puberty timing in men on health outcomes of interest, two-sample MR analysis using summary statistics was employed. Two exemplar traits were chosen based on previously reported associations with puberty timing: prostate cancer and overall lifespan (longevity)[35].

Prostate cancer

GWAS summary statistics for prostate cancer risk were obtained from the PRACTICAL/ELLIPSE consortium[143], based on GWAS meta-analysis of 65,044 cases and 48,344 controls (all of European ancestry) and genotyped using either the iCOGS or OncoArray chips. These analyses were repeated using summary statistics from a subset of 9,640 cases with advanced disease and 45,704 controls. Advanced cases were defined as those who had at least one of: a Gleason score of 8+, prostate cancer death, metastatic disease or a prostate specific antigen (PSA) level >100. Two-sample MR analyses were conducted using weighted linear regression of the SNP-prostate cancer log odds ratio (logOR) on the top signals for SNP-puberty beta coefficients. The relationship was further tested using MR-Egger regression to assess evidence for unbalanced horizontal pleiotropy. Finally, multivariate MR using genetically-predicted SNP effects on BMI was used to separate the total risk of puberty timing on prostate cancer risk into direct (i.e. BMI-independent) and indirect (SNPs operating via BMI) effects.

Longevity

Summary statistics for longevity were obtained from colleagues at the University of Edinburgh, using data from the UK Biobank study and Lifegen consortium[144]. GWAS summary statistics of parent survivorship in 1,012,050 parent lifespans were obtained using a Cox proportional hazard model and were meta-analysed in an inverse-variance weighed meta-analysis. All participants were unrelated and of white European descent. Hazard ratios and their standard errors were taken forward for 74 male puberty loci which were available in this dataset, and two-sample MR analyses were performed to assess the causal effect of male puberty timing on longevity. As a sensitivity analysis, the MR was repeated using MR-PRESSO in order to remove SNPs exhibiting a significant amount of heterogeneity between the datasets.

3.2.11 Association between hair colour and puberty timing

Due to an enrichment for genes related to pigmentation arising from the puberty timing meta-analysis, it was decided to investigate the relationship between natural hair colour and puberty timing in both men and women. Information on natural hair colour for UK Biobank participants was collected via touchscreen questionnaire, in response to the question “What best describes your natural hair colour? (If your hair colour is grey, the colour before you went grey)”. Participants chose from one of 6 possible colours: blond, red, light brown, dark brown, black or other. For our analyses, we restricted this to include only non-related individuals of white European ancestry, totalling 190,845 men and 238,179 women. Hair colours were assigned numerical values from lightest (blond) to darkest (black) in order to perform ordered logistic regression of hair colour for both relative age at voice breaking in men and AAM in women. In both cases, blond hair was used as the reference group and models were adjusted for the top 40 principle components to account for population structure. In men this produces an effect estimate as an odds ratio for early puberty (relative to blond-haired individuals), while in women the effect estimate is on a continuous scale for AAM (in years) relative to the mean AAM for those with blond hair.

To further investigate the possibility of shared biological factors influencing both hair colour and puberty timing, two-sample MR analysis was performed. Summary statistics for natural hair colour were obtained from publicly available data from a recently published GWAS[145]. Summary statistics for male puberty timing were obtained from the 23andMe-only estimates of the independent loci identified in the MTAG meta-analysis. This decision was taken in order to ensure the phenotype was measured on a continuous and uniformly measured scale. For female puberty timing, SNP effect estimates for the 389 independent AAM loci were obtained from the ReproGen consortium. Inverse-variance weighted MR was used for both male and female associations, with MR-Egger, weighted median and penalised weighted median being used as sensitivity analysis to account for heterogeneity in effect estimates and unbalanced horizontal pleiotropy. In addition, because there was partial sample overlap between the male puberty timing GWAS and the hair colour GWAS, which both included 23andMe participants, we performed a further sensitivity analysis in non-overlapping samples. This consisted of a more limited 5-SNP instrument that did not include data from 23andMe or UK Biobank[105]. This was then assessed for effects on puberty timing in UK Biobank men and women in an individual-level MR analysis, controlling for geographical (assessment centre) and as well as 40 principle components to account for genetic ancestry.

3.3 RESULTS

3.3.1 Validation of recalled age of first facial hair as marker for puberty timing

In order to validate recalled age of onset of facial hair as a valid marker of puberty timing in men for inclusion in the meta-analysis, the co-occurrence of relatively early and relatively late first facial hair was compared to that of the similarly dichotomised age at voice breaking among the same individuals in UK Biobank. Recalled age at first facial hair showed high concordance with age at voice breaking, with 87.1% of respondents reporting the same relative age for both traits and only 0.2% reporting the traits occurring in the opposite direction (**Table 3.1**).

The ability of facial hair traits to detect puberty timing loci was then tested by comparing the association with previously reported AAM loci to the association with relative age of first facial hair. Of the 328 autosomal AAM signals for which genotype data were available in UK Biobank, 266 showed directionally-concordant associations with relatively early facial hair and 276 with relatively late facial hair. This represents a significantly higher proportion than expected by chance (binomial test P-value=1.2×10^{-31} for early and P=2.1×10^{-38} for late facial hair onset). Furthermore, substantially more AAM signals showed at least nominally significant associations (P<0.05) with relatively-early (102, 31.1%) and relatively-late (152, 46.3%) compared to the ~16 predicted by chance for each outcome. It was therefore concluded that relative age of first facial hair was a valid phenotypic marker for puberty timing in men.

Table 3.1: Concordance of self-reported relative ages of voice breaking and first facial hair among UK Biobank participants.

		Voice Breaking			
		Younger than average	**About average age**	**Older than average**	**Total**
Facial Hair	**Younger than average**	5,952	6,254	125	12,331
	About average age	1,980	150,303	1,832	154,115
	Older than average	236	14,180	9,300	23,716
	Total	8,168	170,737	11,257	190,162

Green boxes show individuals reporting the same relative ages for both traits, orange boxes represent those reporting average for one trait and relatively earlier or later for the other, and red boxes represent individuals who report traits in opposite directions.

3.3.2 Discovery of independent and novel signals for male puberty timing

The combination of the five GWAS strata in MTAG yielded an effective sample size of 205,354 men for meta-analysis of pubertal timing. Genetic correlations between the traits, obtained from MTAG output using LD score regression, were high (r_g ranging from 0.57 and 0.91). A total of 7,897 variants were associated with male puberty timing at genome-wide statistical significance ($P<5\times10^{-8}$), comprising 76 independent signals (**Figure 3.1**). The maxFDR calculation from MTAG was 7.9×10^{-4}, indicating that test statistics are unlikely to be biased due to violation of the homogeneity assumption. The most significantly associated variant (rs11156429, $P=3.5\times10^{-52}$) was located in/near *LIN28B*, consistent with previously reported studies in men and women[35,78]. Associations with voice breaking loci at the *NR4A2* ($P=8.6\times10^{-36}$), *TMEM38B* ($P=1.3\times10^{-32}$) and *LEPR* ($P=6.1\times10^{-22}$) loci were also replicated in this analysis[78]. Of the 76 signals, 29 were not within 1 MB or in linkage disequilibrium ($r^2>0.05$) with a previously reported AAM or voice breaking SNP and were therefore considered novel (**Table 3.2**). Two of the 76 lead variants associated with male pubertal timing were non-synonymous: a previously reported AAM signal in *KDM4C* (rs913588), encoding a lysine-specific demethylase, and a novel male-specific signal in *ALX4* (rs3824915), encoding a homeobox gene involved in fibroblast growth factor (FGF) signalling that is mutated in rare disorders of cranium/central neural system (CNS) development with male-specific hypogonadism[146]. A further 10 lead variants were in strong LD ($r^2>0.8$) with one or more non-synonymous variants, of which three represent novel signals for puberty timing: *FGF11*, which encodes a FGF expressed in the developing CNS and promotes peripheral androgen receptor expression[147]; *TFAP4,* which encodes a transcription factor of the basic helix-loop-helix-zipper family[148]; and *GCKR*, which encodes a regulatory protein that inhibits glucokinase in liver and pancreatic islets and is associated with a range of cardio-metabolic traits[149]. A further 7 are reported signals for AAM, but were not previously reported for voice-breaking. These missense variants are in the following genes: *SRD5A2*, encoding for Steroid 5-alpha-reductase, which converts testosterone into the more potent androgen dihydrotestosterone; *LEPR*, encoding the receptor for appetite and reproduction hormone leptin; *SMARCAD1*, encoding a mediator of histone H3/H4 deacetylation; *BDNF, FNDC9, FAM118A,* and *ZNF446*.

Table 3.2: Novel genes associated with puberty timing in males

Variant	Chr	Position	Alleles#	EAF	Nearest gene	MTAG Beta (S.E.)	MTAG P-value
rs71578952	7	131,001,466	C/T	0.495	*MKLN1*	0.035 (0.003)	8.4×10^{-28}
rs2222746	17	44,222,019	T/G	0.165	*KIAA1267*	0.048 (0.004)	9.0×10^{-28}
rs73182377	3	181,512,034	C/T	0.227	*SOX2OT*	0.040 (0.004)	1.9×10^{-24}
rs3824915	11	44,331,509	G/C	0.496	*ALX4*	0.030 (0.003)	1.0×10^{-20}
rs77578010	1	11,035,758	A/G	0.776	*C1orf127*	0.036 (0.004)	1.1×10^{-20}
rs7402990	15	28,384,491	G/A	0.916	*HERC2*	0.051 (0.006)	1.4×10^{-18}
rs17833789	17	55,230,628	C/A	0.547	*AKAP1*	0.028 (0.003)	5.9×10^{-18}
rs12203592	6	396,321	C/T	0.170	*IRF4*	0.035 (0.004)	9.6×10^{-16}
rs35063026	16	89,736,157	T/C	0.069	*C16orf55*	0.051 (0.006)	2.1×10^{-15}
rs9690350	7	547,800	C/G	0.419	*PDGFA*	0.025 (0.003)	3.2×10^{-14}
rs6560353	9	76,375,544	G/T	0.161	*ANXA1*	0.033 (0.004)	9.7×10^{-14}
rs7905367	10	54,334,653	G/C	0.219	*MBL2*	0.027 (0.004)	4.7×10^{-12}
rs7136086	12	114,129,719	C/T	0.734	*RBM19*	0.025 (0.004)	1.4×10^{-11}
rs10110581	8	60,691,207	G/A	0.263	*CA8*	0.025 (0.004)	2.3×10^{-11}
rs2842385	6	19,078,274	G/A	0.193	*MIR548A1*	0.027 (0.004)	2.5×10^{-11}
rs11836880	12	91,243,529	C/G	0.040	*C12orf37*	0.053 (0.008)	1.1×10^{-10}
rs10164550	2	121,159,205	A/G	0.657	*INHBB*	0.022 (0.003)	1.7×10^{-10}
rs12895406	14	36,998,950	A/G	0.536	*NKX2-1*	0.021 (0.003)	2.8×10^{-10}
rs835648	3	136,671,504	A/T	0.696	*NCK1*	0.022 (0.004)	5.2×10^{-10}
rs780094	2	27,741,237	T/C	0.413	*GCKR*	0.020 (0.003)	7.8×10^{-10}
rs1979835	5	135,689,839	A/G	0.876	*TRPC7*	0.029 (0.005)	2.5×10^{-9}
rs12930815	16	4,348,635	C/T	0.521	*TFAP4*	0.019 (0.003)	2.6×10^{-9}
rs12940636	17	53,400,110	C/T	0.655	*HLF*	0.020 (0.003)	3.6×10^{-9}
rs60856990	17	7,337,853	A/G	0.629	*TMEM102*	0.019 (0.003)	1.3×10^{-8}
rs17193410	17	32,474,149	G/A	0.880	*ACCN1*	0.028 (0.005)	1.5×10^{-8}
rs11761054	7	46,076,649	G/C	0.291	*IGFBP3*	0.020 (0.004)	2.2×10^{-8}
rs10765711	11	94,879,318	C/G	0.415	*ENDOD1*	0.018 (0.003)	2.9×10^{-8}
rs61168554	15	99,286,980	A/G	0.361	*IGF1R*	0.019 (0.003)	3.8×10^{-8}
rs12983109	19	49,579,710	G/A	0.742	*KCNA7*	0.020 (0.004)	3.8×10^{-8}

# Effect allele/other allele

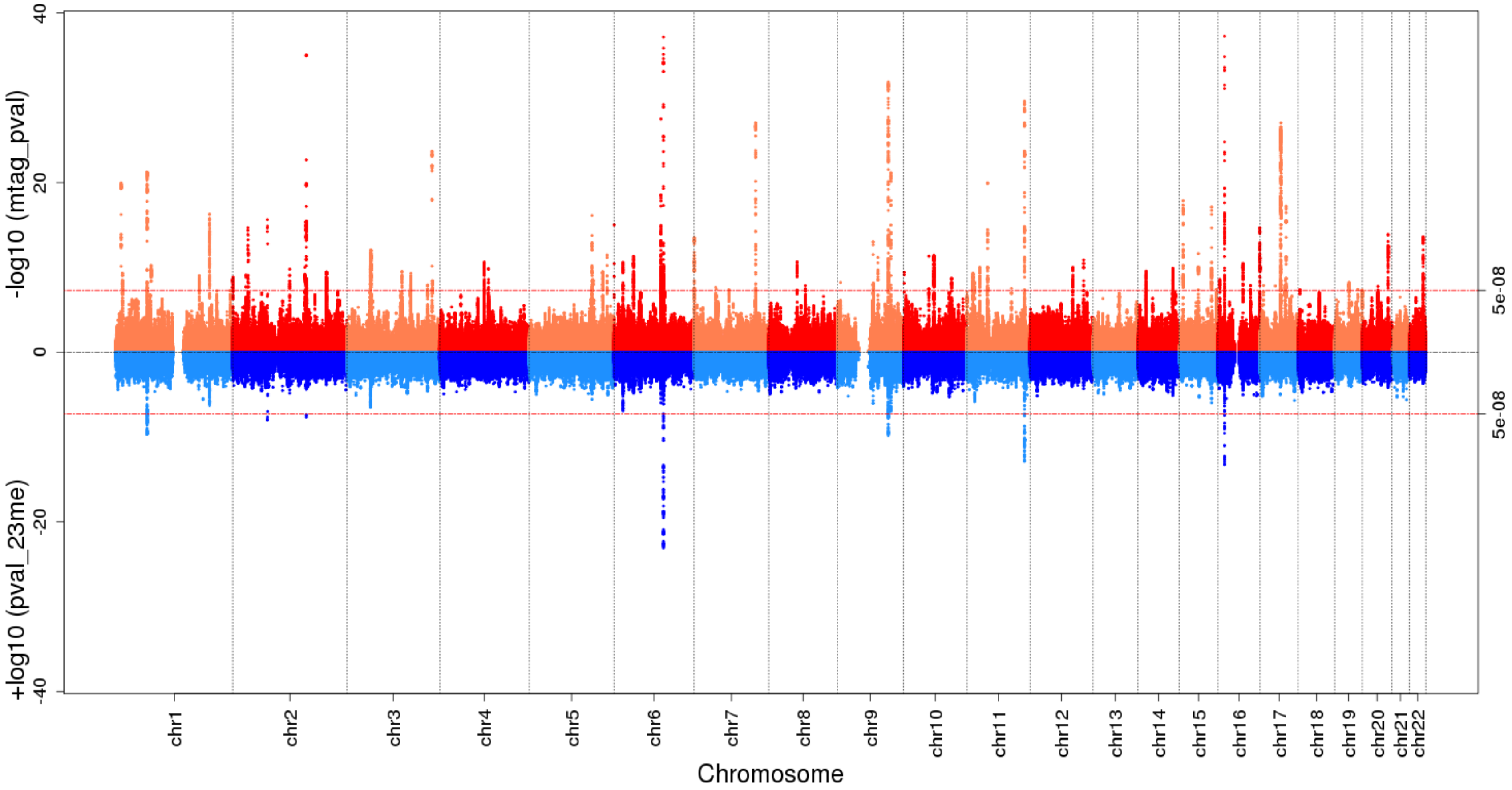


Figure 3.1: Miami plot for multi-trait GWAS of male puberty timing. MTAG –$\log_{10}$ P-values for SNP associations with male puberty timing (top half, red shades) and previously reported –$\log_{10}$ P-values for age at voice breaking from the 23andMe study GWAS (bottom half, blue shades) are displayed. Red dashed lines indicate the genome-wide statistical significance threshold ($P<5 \times 10^{-8}$).

3.3.3 Replication of signals using a polygenic risk score in ALSPAC

Collective confirmation of signals for puberty timing in men was sought by constructing a polygenic risk score (PRS) for puberty timing (aligned to later age of puberty onset) based on the 73 independent signals which were also available in ALSPAC. Results are summarised in **Table 3.3**. The PRS was associated with older ages of peak height velocity ($P=1.6\times10^{-14}$), age at voice break ($P=4.5\times10^{-3}$) and age of armpit hair appearance ($P=7.3\times10^{-5}$) in a directionally-consistent manner for the age-combined variables. When age-specific variables were considered, the PRS showed no significant association with Tanner staging of genital and pubic hair development or for occurrence of voice breaking in the youngest age category (mean age= 9.6 years). This is likely due to low statistical power as very few individuals will have reached puberty at this age. For mean age 13.1 years, the PRS is negatively associated with Tanner stage variables and occurrence of voice breaking, indicating a lower likelihood of attaining these developmental milestones at this age. In the oldest age category (mean age 17.0 years) the PRS remains inversely associated with Tanner stage indicators but is positively associated with occurrence of voice breaking (P=0.02). While underpowered in the age-specific categories, the associations are largely directionally consistent while the age-combined variables are all significantly associated and in the expected directions, thus providing support for the validity of the variants identified using MTAG.

Table 3.3: Associations of polygenic risk score for male puberty timing with milestones for puberty in ALSPAC

Phenotype	Mean Age	N	Beta	SE	P-value	r^2 (%)
Age-combined variables†						
Age at peak height velocity (years)	-	2,028	0.157	0.020	1.6×10^{-14}	3.3
Age at voice break (years)	-	2,343	0.101	0.036	4.5×10^{-3}	0.6
Age of armpit hair appearance (years)	-	2,403	0.120	0.030	7.3×10^{-5}	1.1
Age-specific variables						
Tanner stage of genital development	9.6	2,191	-0.009	0.039	0.82	0.4
Tanner stage of pubic hair development	9.6	2,139	-0.028	0.057	0.62	0.7
Voice break (Y/N)	9.6	2,280	0.014	0.103	0.89	1.0
Tanner stage of genital development	13.1	1,555	-0.240	0.047	2.4×10^{-7}	0.9
Tanner stage of pubic hair development	13.1	1,749	-0.260	0.042	7.5×10^{-10}	1.0
Voice break (Y/N)	13.1	1,886	-0.247	0.047	1.4×10^{-7}	1.6
Appearance of armpit hair	13.1	1,861	-0.197	0.048	4.3×10^{-5}	1.4
Tanner stage of genital development	17.0	1,127	-0.214	0.067	1.3×10^{-3}	1.3
Tanner stage of pubic hair development	17.0	1,268	-0.269	0.071	1.6×10^{-4}	1.7
Voice break (Y/N)	17.0	1,143	0.161	0.070	0.02	1.4
Appearance of armpit hair	17.0	1,274	-0.063	0.199	0.75	7.5
†adjusted only for PCs						

3.3.4 Genetic heterogeneity of puberty timing between sexes

Consistent with previous studies[35], a moderately strong genome-wide genetic correlation was observed between pubertal timing in men and women (r_g=0.68, P=2.6×10^{-213}) (**Figure 3.2**). However there were exceptions to this overall trend, with 5 of 76 male puberty signals and 15 of 387 AAM signals showing significant heterogeneity in their effects between sexes. As previously reported[35], one signal at the *SIM1/PRDM13/MCHR2* locus showed significant and directionally opposite effects where the allele that conferred earlier puberty timing in men delayed AAM in women (rs6931884/T: $\beta_{\text{voice-breaking}}$= -0.064 years/allele; β_{menarche}=0.059 years/allele; $P_{\text{heterogeneity}}$=2.6×10^{-14}). Two variants located near to genes that are disrupted in rare disorders of puberty showed no effect or weaker effect in males than in females: rs184950120, 5'UTR to *MKRN3* ($\beta_{\text{voice-breaking}}$=0.085 years/allele; β_{menarche}=0.396 years/allele, $P_{\text{heterogeneity}}$=3.6×10^{-3}), and rs62342064, one of 3 AAM variants in/near *TACR3* ($\beta_{\text{voice-breaking}}$= -0.017 years/allele; β_{menarche}=0.057 years/allele, $P_{\text{heterogeneity}}$=4.2×10^{-5}).

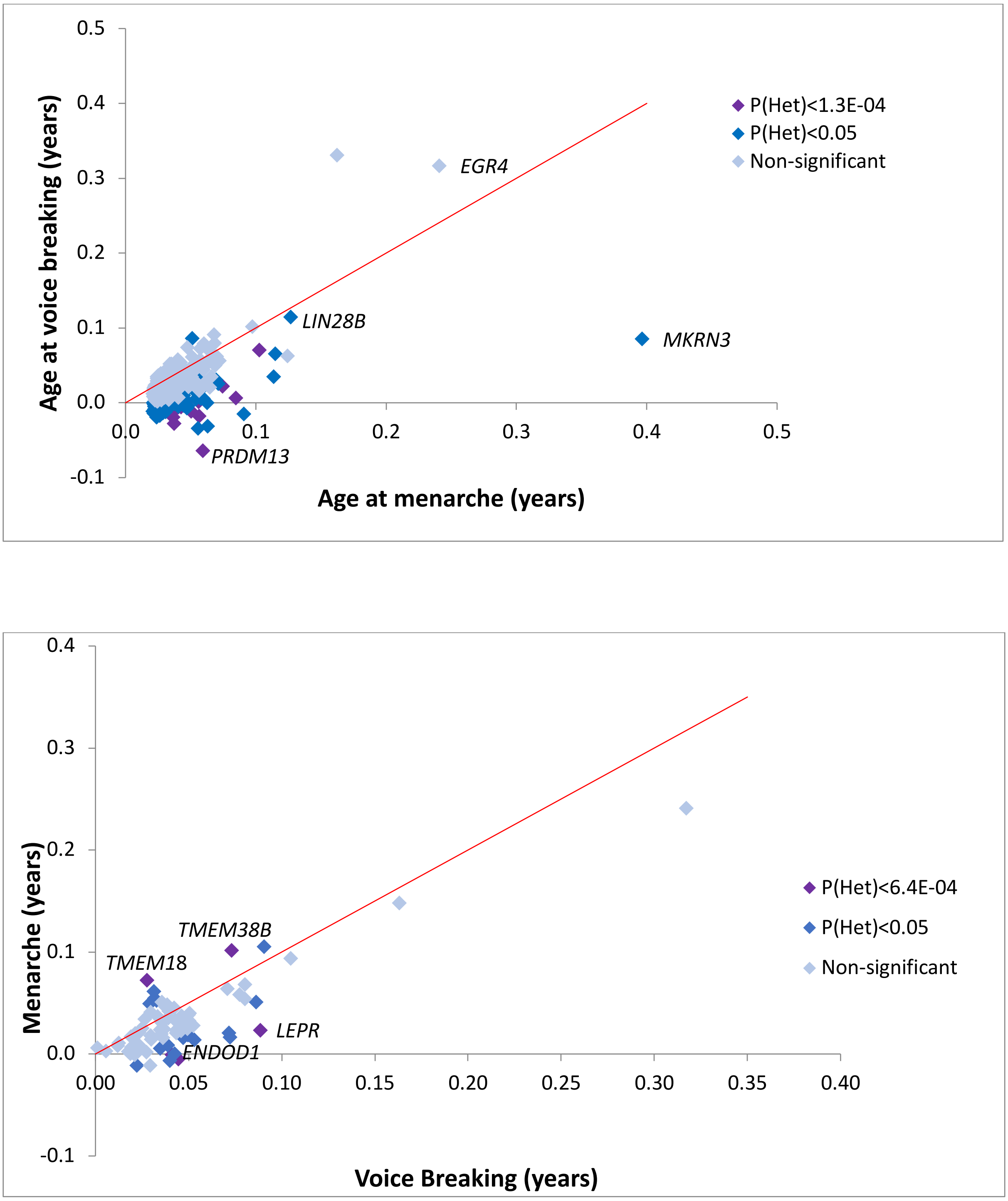


Figure 3.2: Scatterplots comparing effect sizes of AAM loci (top panel) and VB loci (bottom panel). SNPs are not independent across the two panels. Variants are coloured based on heterogeneity (P-value) between women and men. Red lines indicate perfect agreement (i.e. X=Y).

3.3.5 Tissue-specific expression of puberty timing genes

Gene expression analysis for general tissue categories in GTEx indicated that central nervous system (CNS) tissues were the most strongly enriched for genes which co-locate near to male puberty timing signals (enrichment P=4.4×10^{-8}). Significant enrichment was also found for adrenal and pancreas tissues (P=3.2×10^{-6}), skeletal muscle (P=1.2×10^{-4}), connective and bone tissue (P=4.7×10^{-4}) as well as liver (P=1.4×10^{3}) and kidney (P=4.3×10^{-3}) (**Figure 3.3**). When specific tissues were considered, none of the 53 GTEx tissues analysed showed statistically significant enrichment after applying a Bonferroni correction to account for multiple tests. Strongest enrichment occurred in the cerebellar hemisphere (P=6.9×10^{-3}), cerebral cortex (P=0.01) and hypothalamus (P=0.02), while reproductive tissues showed lower levels of enrichment (ovary P=0.13; testis P=0.53) (**Figure 3.4**).

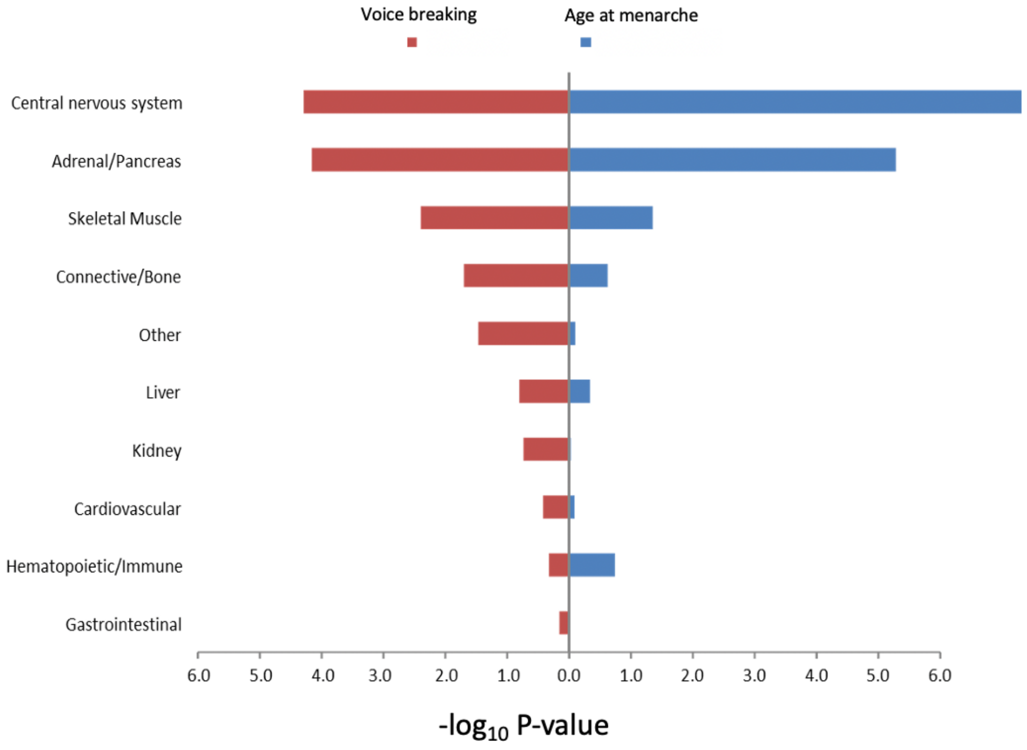


Figure 3.3: Tissue expression for male puberty loci by general GTEx tissue categories. Values on the left side (in red) are for voice breaking, while the right side is for age at menarche (in blue).

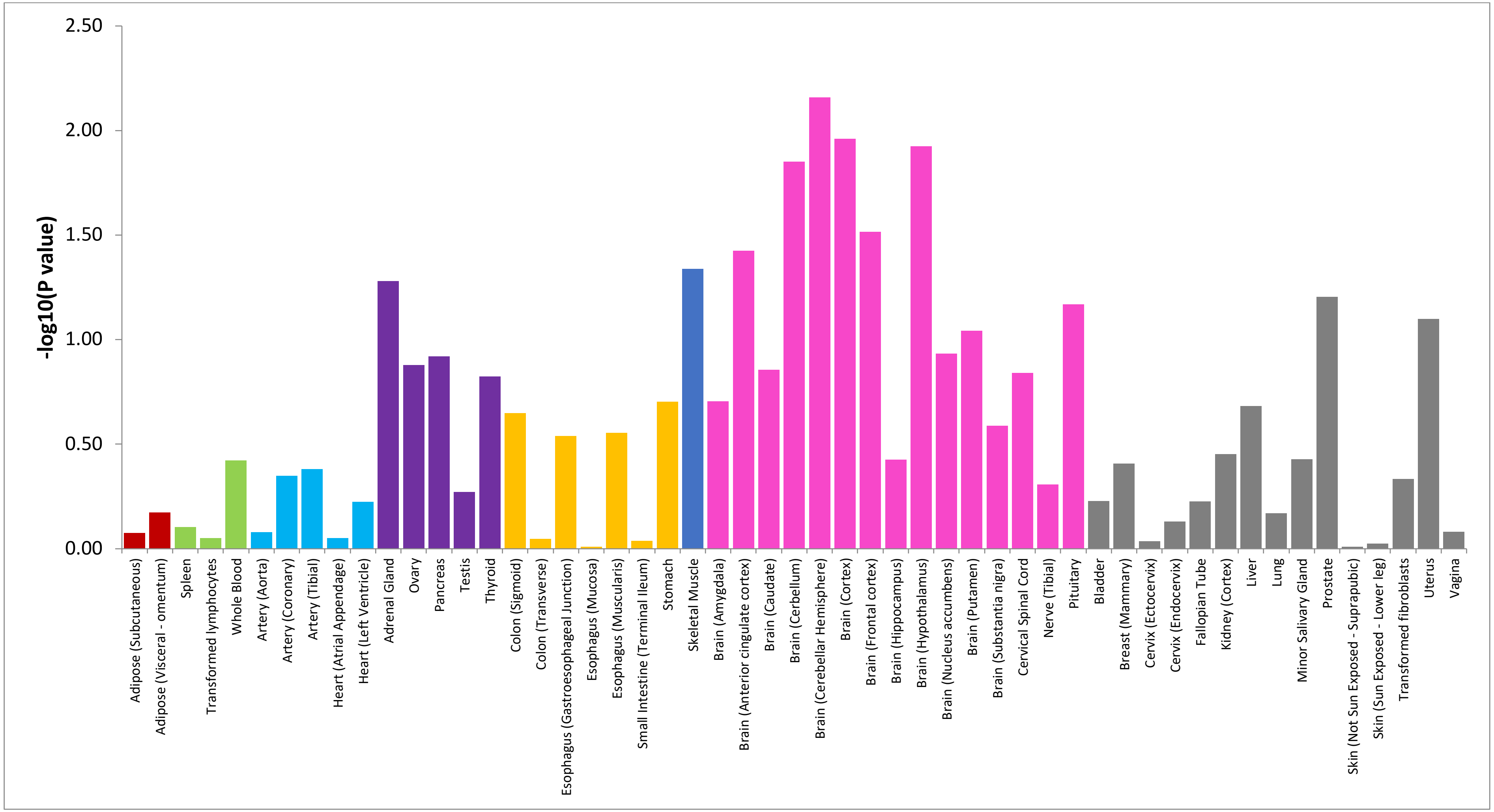


Figure 3.4: Tissue expression for male puberty loci by specific GTEx tissues. –log$_{10}$ p-values for gene expression are depicted on y-axis. Tissues are colour-coded by categories: adipose (red); blood (green); cardiovascular (light blue); endocrine (purple); gastrointestinal (yellow); skeletal muscle (dark blue); central nervous system (pink); other (grey). Note that no significance threshold is indicated, as all P-values were below the Bonferroni-corrected threshold

3.3.6 Gene set enrichment implicates multiple pathways

To identify mechanisms that regulate pubertal timing in males, we tested all SNPs genome-wide for enrichment of voice breaking associations with genes in pre-defined biological pathways. Four pathways showed evidence of enrichment: histone methyltransferase complex (FDR=0.01); regulation of transcription (FDR=0.02); ATP binding (FDR=0.03); and cAMP biosynthetic process (FDR=0.03).

3.3.7 Genetic correlations with male puberty timing and other health-related traits

To assess the extent of shared heritability between male puberty timing and other complex health and behavioural traits, genome-wide genetic correlations with 751 phenotypes of interest were calculated using LD score regression. Male puberty timing showed the strongest positive genetic correlations with adolescent growth (r_g=0.77) and AAM (r_g=0.68). Additionally, later male puberty was positively correlated with social and behaviour traits including age at first live birth (r_g=0.26), attainment of university-level degree (r_g=0.18) and fluid intelligence (r_g=0.13). Furthermore, overall health rating showed strong positive correlation with later puberty (r_g=0.22) (**Figure 3.5**). In contrast, the data suggest that individuals who go through puberty earlier are at increased risk for a number of adverse health outcomes in later life based on negative genetic correlations, including hypertension (r_g=-0.16), diabetes (r_g=-0.22) and osteoarthritis (r_g=-0.24). In addition, earlier puberty is genetically correlated with health risk behaviours including pack-years of smoking (r_g=-0.20) and alcohol intake frequency (r_g=-0.20). Overall, this indicates a trend for earlier puberty being disadvantageous for health in adulthood.

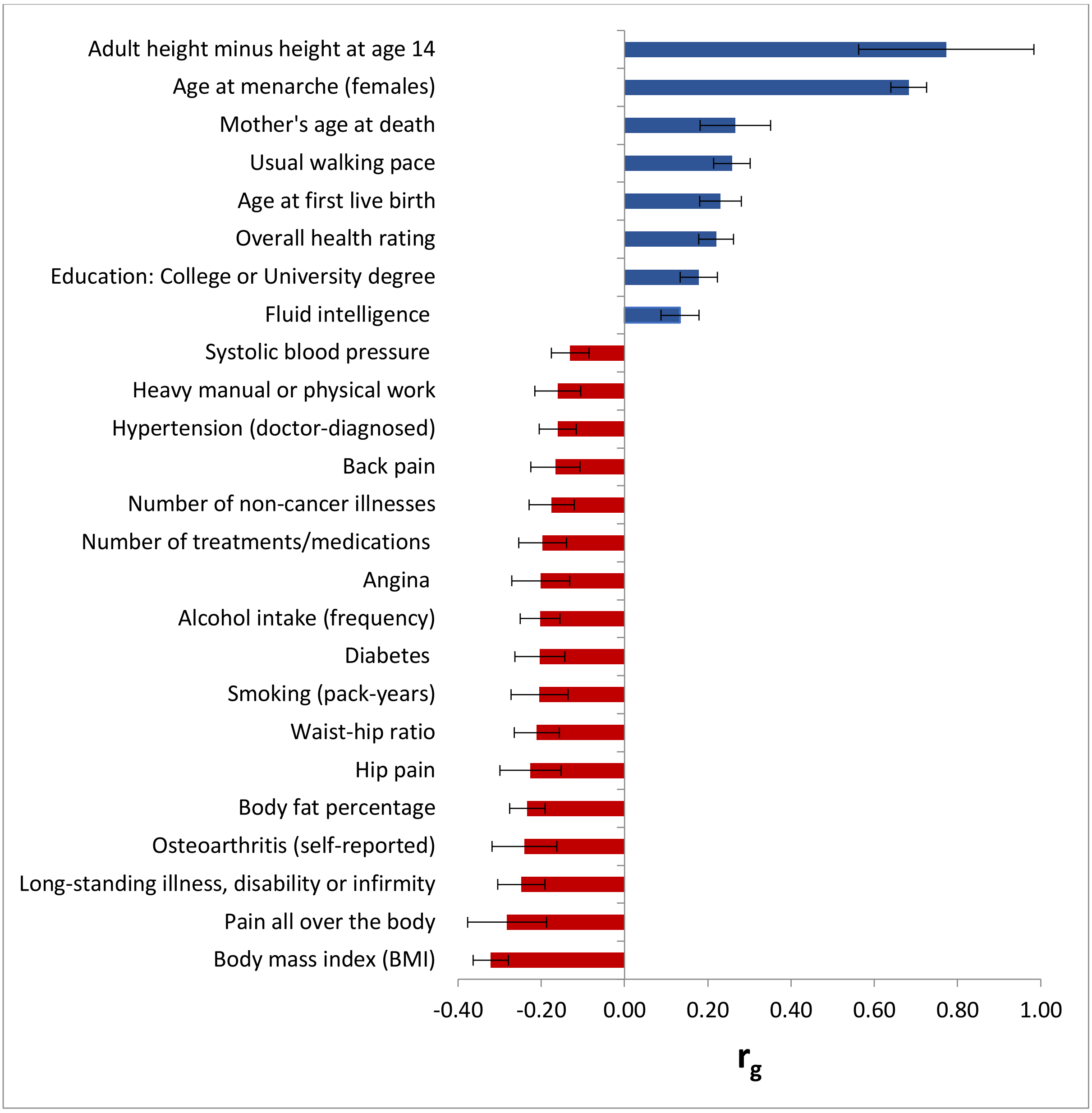


Figure 3.5: Genetic correlations between male puberty timing and selected anthropometric and health-related traits. Calculated using LD Score regression. Positive correlations are shown in blue and negative correlations in red. Error bars represent 95% confidence intervals.

3.3.8 Causal effects of male puberty timing on prostate cancer and overall lifespan

Previous studies have reported evidence for causal relationships between earlier puberty (AAM) and higher risk for hormone-sensitive cancers (e.g. breast cancer)[35]. This effect may be mitigated to some extent by the protective effects of higher BMI, which must be accounted for by inclusion of genetically-predicted BMI as a covariate. In two-sample, multivariate approach a protective effect for later puberty timing on prostate cancer was observed when adjusting for BMI, however this was only statistically significant in cases on advanced prostate cancer (OR= 0.89, P=0.06 for all cases; OR=0.82, P=0.05 for advanced cases) (**Table 3.4**). Similarly, higher BMI shows suggestive evidence of a protective effect against prostate cancer when considering all cases, though this is not statistically significant (OR=0.90, P=0.06). For all analyses there was evidence for significant heterogeneity (minimum I^2=29.3%, P=0.01 for advanced cases adjusted for BMI), while MR-Egger tests to account for unbalanced horizontal pleiotropy of SNPs were significant only for advanced cases. In summation, these results present weak evidence for a causal effect of male puberty timing on increased prostate cancer risk.

In contrast, strong evidence was found for a causal effect of male puberty timing on lifespan, corresponding to ~9 months longer life per year later puberty (IVW P=6.7×10^{-4}) (**Figure 3.6**). This result was consistent after removal of one outlier SNP using MR-PRESSO (P=8.3×10^{-4}) while weighted median MR analysis, which is robust to the effects of heterogeneous SNPs, also supported this (P=0.02).

Table 3.4: Mendelian randomisation analyses for association between male puberty timing and prostate cancer

	Odds Ratio	95% CI	P-value	I^2 (%)	P_{het}	MR Egger (P-value)
All prostate cancers						
Puberty timing	0.94	0.84-1.05	0.27	48.0	2.7×10^{-6}	0.20
Puberty timing (adj. BMI)	0.89	0.79-1.01	0.06	46.4	8.2×10^{-6}	0.10
BMI	0.90	0.81-1.00	0.06	49.3	4.5×10^{-8}	0.36
Advanced prostate cancers						
Puberty timing	0.87	0.72-1.05	0.13	30.7	0.01	0.02
Puberty timing (adj. BMI)	0.82	0.68-1.00	0.05	29.3	0.01	0.01
BMI	0.94	0.79-1.13	0.53	35.7	4.1×10^{-4}	0.32

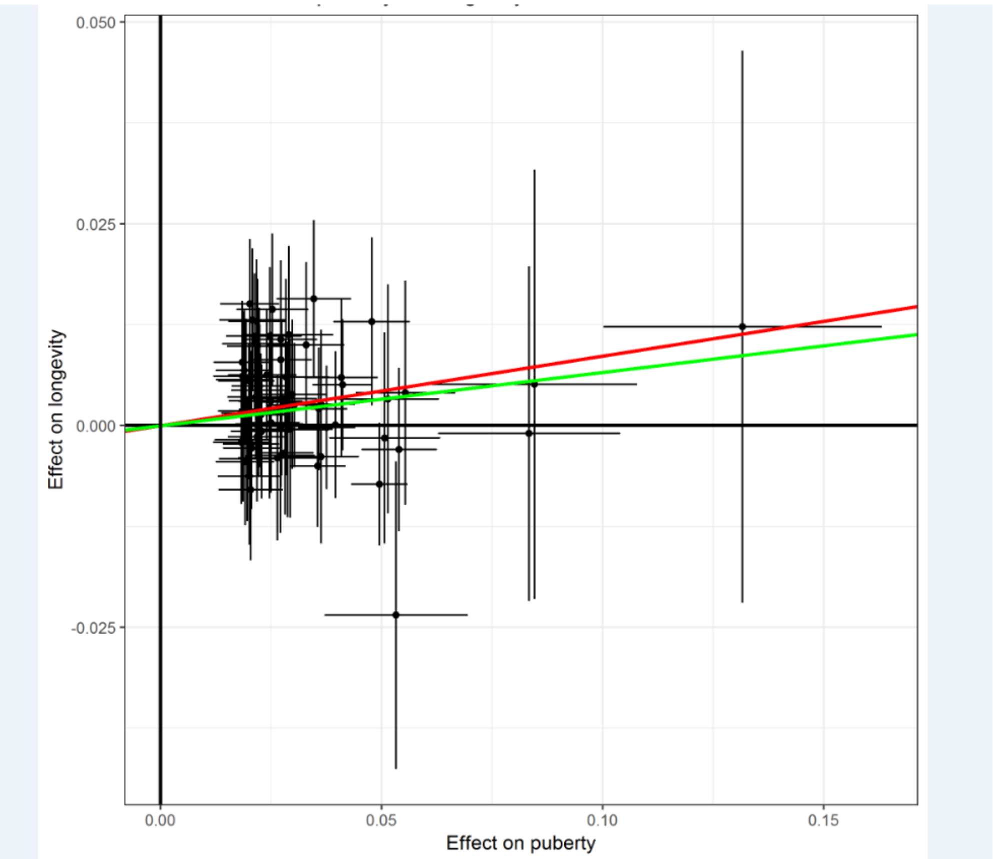


Figure 3.6: Scatterplot of 73 male puberty loci (pruned for heterogeneity) comparing effect sizes for puberty timing (from MTAG) and longevity. Lines show results for different MR models: IVW (red), weighted median (green).

3.3.9 Genetic and phenotypic links between hair colour and puberty timing

Gene annotation of the 76 independent loci for male puberty timing revealed three novel loci which were located nearby to genes previously associated with pigmentation: *HERC2, IRF4,* and *C16orf55*[150–152]. A biological link between puberty timing and pigmentation may be plausible, given previously published evidence for a progressive darkening of skin and hair colour during adolescence[153,154]. Together, these lines of evidence prompted an investigation of potential links between inter-individual variation in natural hair colour and puberty timing. In men, those with red, dark brown and black hair showed progressively higher odds of early puberty timing relative to men with blond hair. In women, while there was an overall effect of women with darker hair colours having early puberty timing relative to those with blond hair, the trend of increased risk for earlier puberty with progressively darker hair was not observed (**Table 3.5**).

Table 3.5: Phenotypic associations between natural hair colour and puberty timing in UK Biobank men and women

Natural hair colour	Men (n=179,549)			Women (n=238,195)		
	Effect	95% CI	P-value	Effect	95% CI	P-value
Blond	(ref)			(ref)		
Red	1.17	1.02, 1.35	0.02	-0.100	-0.134, -0.066	6.2×10^{-9}
Light brown	1.00	0.91, 1.09	0.92	-0.026	-0.047, -0.005	0.01
Dark brown	1.45	1.34, 1.58	6.7×10^{-18}	-0.093	-0.144, -0.072	6.7×10^{-18}
Black	1.63	1.46, 1.81	1.2×10^{-9}	-0.059	-0.113, -0.005	0.03

To further explore this association using genetic data, the effects of genetic variants associated with natural hair colour were systematically assessed for their corresponding effect on puberty timing in a two-sample MR approach. In contrast to more typical application of MR to infer causality between an exposure and an outcome, in this setting the approach can provide more explicit evidence of shared biological mechanisms linking these traits. The findings from this analysis are largely in line with the phenotypic observations, as IVW MR in men suggests that darker natural hair colour is associated with earlier puberty timing in men (β= -0.044 years per ordered category, $P=7\times10^{-3}$) (**Figure 3.7**). Cochrane's Q statistic show no evidence for heterogeneity (P=0.99) while the MR-Egger intercept suggests no evidence of horizontal pleiotropy (P=0.99). The effect is directionally consistent in women (β_{IVW}= -0.017, $P=3.64\times10^{-3}$) though with more heterogeneity of effects between traits (Cochrane's Q: P=0.038).

Sensitivity analysis using the 5-SNP score from the non-overlapping sample was highly consistent in men, with dark hair associated with earlier puberty (OR=1.16, $P=1.72\times10^{-19}$); however the result was not replicated in women (β= -0.006, P=0.23), leaving open the possibility of the effects observed in the main analysis being driven in part by sample overlap in women.

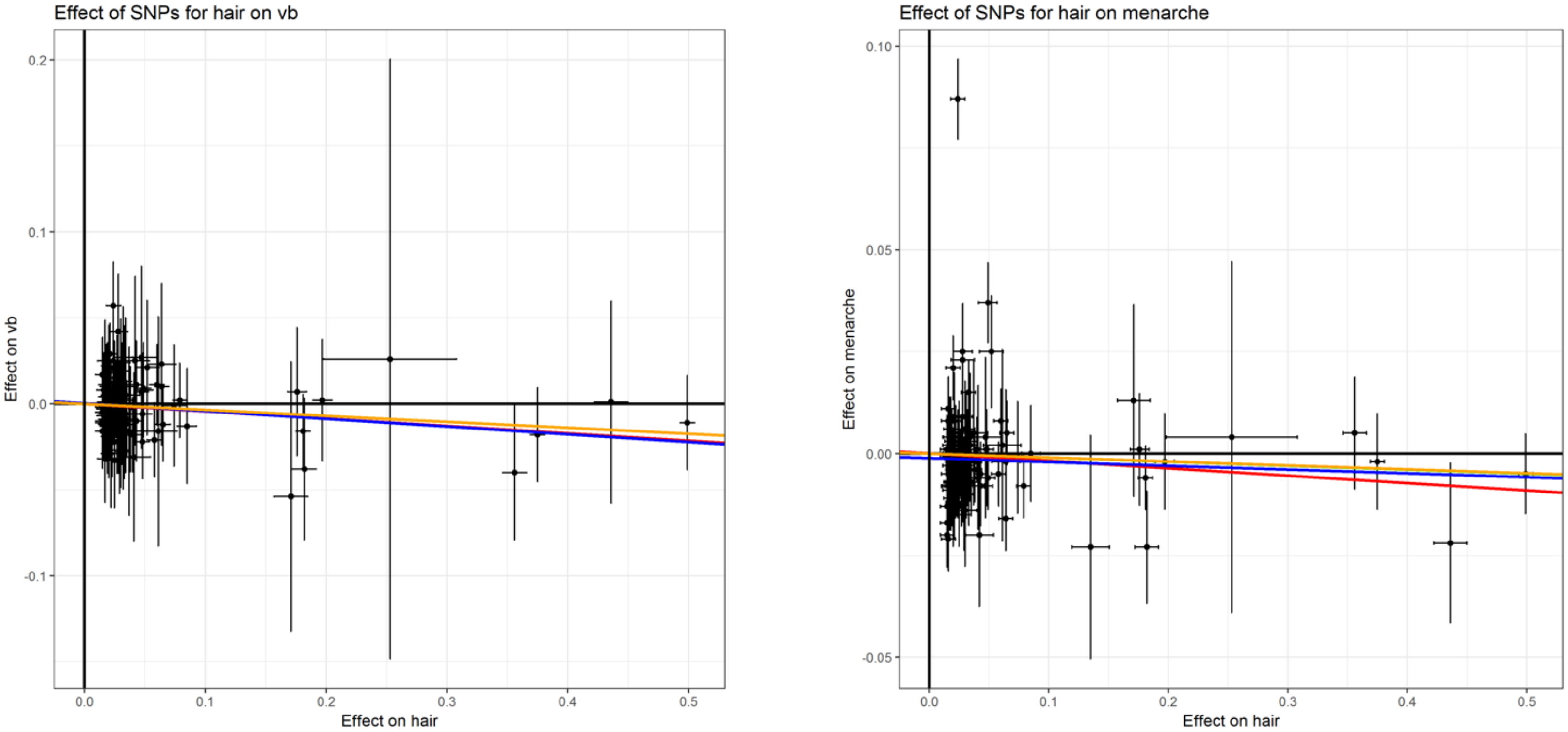


Figure 3.7: Mendelian randomisation analyses for effect of hair pigmentation on puberty timing in men and women. SNP effect size on hair colour is plotted on the x-axis, while corresponding effect on age at voice breaking in men (left panel) and age at menarche in women (right panel) is plotted on the y-axis. Error bars reflect standard errors. Regression lines are shown for inverse variance-weighted (red), MR-Egger (blue) and penalised weighed median (orange) methods.

3.4 DISCUSSION

Puberty timing has important consequences for long-term health, and while previous studies have demonstrated a substantial overlap in the genetic architecture underling puberty timing between sexes, we risk missing key information for variants with heterogeneous effects by not conducting sex-specific analyses. While inferences have been drawn from large-scale genomic studies conducted on AAM in women, this study represents a substantial increase in our understanding of male-specific puberty timing genetic variation with a sample size nearly four-fold larger than any male-only study previously conducted.

3.4.1 Novel loci implicate new and unexpected pathways

This multi-trait meta-analysis, which incorporates phenotypes from the UK Biobank and 23andMe studies in an effective sample of 205,354 men, identified 76 independent loci for puberty timing in men of which 29 had not previously been associated with puberty in either sex. These novel loci implicate genes with known links to endocrine function, such as the *INHBB* gene which encodes the beta-B subunit of the Inhibin B hormone and influences the onset of spermatogenesis by inhibiting follicle stimulating hormone (FSH) secretion. Similarly, *SRD5A2* which encodes the steroid hormone 5-alpha-reductase is responsible for the conversion of testosterone to dihydrotestosterone, is essential for development of secondary sex characteristics associated with puberty. Genome-wide expression data shows that male puberty genes are enriched in CNS tissues, consistent with findings in AAM[35] and is expected given important role of the hypothalamic-pituitary-gonadal axis which controls aspects of the reproductive lifecycle. Indeed, hypothalamus and pituitary tissues were among the most enriched for expression. This may also reflect an enrichment for BMI-mediated pathways, as BMI is known to have a strong influence on puberty timing and has a CNS-enriched expression pattern related to appetite regulation[155].

In addition to genes and pathways with known reproductive function, the novel loci also revealed unexpected insights into factors related to puberty timing. While a darkening of hair and skin pigmentation has consistently been reported in children on European ancestry[153,154], this has never been systematically assessed using genetic methods. The results of the MR analyses for natural hair colour to puberty timing yielded robust evidence for widespread shared biological mechanisms influencing these two traits. The overall effect sizes are larger in men, and the graded trend for earlier puberty with darkening hair is not apparent in women. This is perhaps to be expected as it has been reported that androgens have a stronger stimulatory effect on melanogenesis compared to oestrogens[156]. A plausible biological link exists in the pituitary propeptide pro-opiomelanocortin (POMC), which contains several peptides which influence melanogenesis including ACTH, α-MSH, and β-MSH. The cleavage of the pro-hormone into constituent parts is achieved through the action of the convertase enzymes PC-1 and PC-2, which are both signals for AAM[35].

3.4.2 Genetics underpinnings of associations between puberty and adult health

The link between earlier puberty and adverse health outcomes in later life has been well-established in the literature, though again much of this is based on puberty data in women. This may be explained by increased duration of exposure to sex hormones for those undergoing earlier puberty, as these are known to increase the risk for many metabolic diseases and cancer. It may also reflect a 'common soil' relationship with BMI, as overweight children are more susceptible to early puberty and are more likely to remain overweight into adulthood. Genetic correlations from this male puberty meta-GWAS confirm many of these findings, including strong inverse associations with diabetes, cardiovascular conditions and adult BMI, compared to positive correlations with favourable health and social outcomes including perceived health and educational attainment.

MR analyses demonstrated a causal influence of earlier male puberty on decreased lifespan. Earlier puberty has also previously been shown to be causally related to prostate cancer when accounting for BMI[35], though the MR analyses conducted here did not find evidence to support this. However the directions of effect were consistent with expectations based on previous findings; therefore it is premature to dismiss the possibility of a causal relationship based on borderline-insignificant P-values. If these results can be confirmed in larger studies, this represents an interesting and complex dynamic, whereby increased BMI has directionally opposite effects of prostate cancer risk: an indirect deleterious effect via its association with earlier puberty, balanced by a direct protective effect.

3.4.3 Strengths and Limitations

This study presents a vast improvement in the sample size for discovery of genetic variants associated with male puberty timing, a trait which has previously been poorly characterised. The MTAG framework allows incorporation of data from multiple and potentially overlapping cohorts from UK Biobank in addition to the 23andMe data used previously. MTAG also allows for the inclusion of data on other markers for puberty timing (i.e. facial hair onset) to the previously validated age of voice breaking, and the four-fold increase in sample size resulting from this represents a significant advance in statistical power for discovery. This is evidenced by the large increase in the number of genetic loci associated with male puberty timing, from 6 to 76 independent genomic associations.

Results from MTAG must be interpreted cautiously, as the meta-analysis relies on a number of quite strong assumptions. Simulations involving extreme examples have demonstrated that the assumptions of ignoring sampling variation and the adequate capture of sample overlap may safely be made for most practical applications of MTAG. However the primary assumption of homogeneity of SNP effects across all traits is potentially violated in this study due to the inclusion of facial hair onset to the other traits measuring voice breaking. To mitigate this, great care was taken to ensure that effect estimates produced from MTAG were not being driven exclusively by a single stratum, and the maxFDR calculation provided strong justification that this key assumption was not violated.

Due to the size of this study, replication of individual SNPs in an independent cohort was intractable as there are few, if any, cohorts with sufficient numbers of participants with puberty data to provide reliable replication of individual SNPs. While ALSPAC has thoroughly assessed phenotype data in their cohort, the sample sizes are substantially smaller than the discovery set and replication relied on combining all SNPs into a polygenic risk to improve power. Therefore, an objective for future follow-up studies should be to replicate the associations with individuals SNPs.

3.4.4 Conclusion

In the context of early life determinants for adult disease, puberty may be considered to represent the culmination of 'early-life' and transition into adulthood. This chapter summarises the importance of puberty timing on an individual's health and well-being throughout their life course, and presents results from a vastly expanded genetic discovery on male-specific genetic variation in puberty timing. Annotation and statistical analyses of the functions of newly identified variants confirms previously reported associations linking early puberty to poor health outcomes, while identifying novel insights into underlying biological mechanisms.

Chapter 4: Birth weight and adult body composition

Contributions and collaborations

This study was conceived in collaboration with colleagues at the MRC Epidemiology Unit, including Dr. Ken Ong, Dr. Felix Day and Dr. John Perry. In addition, Dr. Laura Wittemans conducted the GWAS for DEXA-derived compartmental body composition. All other work described here was completed by me, including the meta-analysis of GWAS summary statistics, hierarchical clustering and Mendelian randomisation analyses.

SUMMARY

Birth weight is an important early life exposure which has been shown to predict health outcomes in adulthood. Yet despite the consistently reported associations between low birth weight and common comorbidities of obesity (including high blood pressure, cardiovascular disease and diabetes), observational studies have not shown a consistent relationship between lower weight at birth and higher BMI in adulthood. It has been suggested that low birth weight may instead lead to adaptations which influence the distribution of body fat towards a more central and metabolically unhealthy pattern. While there is some evidence for this from observational studies, it has not been systemically investigated using genetic methods. In this chapter I describe such a study, using genetic variants associated with birth weight in combination with results from a GWAS on compartmental fat and lean mass composition obtained from DEXA scans. By using genetic data, robust causal associations are identified for the influence of birth weight on body composition and fat distribution. By partitioning the heritable components of birth weight into foetal and maternal-specific effects, inferences can be made regarding the relative impact of foetal and maternal genomes. These results inform the ongoing debate surrounding mechanisms of the DOHaD hypothesis, and provide new insights into the role of genetic factors in explaining the associations between early life exposures and adult disease.

4.1 BACKGROUND

4.1.1 Birth weight and associations with metabolic health outcomes

Exposures which influence health outcomes can begin very early in the life course. This is exemplified by observed associations between birth weight and risk of cardio-metabolic disease in later-life. Associations between low birth weight and higher blood pressure, T2D and dyslipidaemia in adulthood have consistently been reported[157–161]. These effects of low birth weight may begin to manifest even in early childhood, with metabolic abnormalities including insulin resistance and abnormal inflammatory markers having been reported in children as young as two years[162,163]. At the opposite end of the spectrum, high birth weight has also been shown to be a risk factor for poor metabolic health, with studies demonstrating increased prevalence of T2D and CVD among individuals born at high weight for gestational age (HGA)[164,165]. Such findings highlight the importance of healthy birth weight in normal growth and development, and identify abnormal birth weight as a key early-life exposure which can have a long-lasting impact on individual health.

4.1.2 Foetal programming or shared genetic effects?

Low birth weight is often purported to be a reflection of adverse conditions in the intra-uterine environment. As discussed in Chapter 1, the DOHaD hypothesis posits that adaptations made by the foetus in response to stressors such as maternal malnutrition can explain the heightened health risks that are regularly observed in low birth weight individuals[166,167]. This has been termed 'foetal programming', and is supported primarily by evidence indicating increased prevalence of obesity and other indicators of poor cardio-metabolic health in offspring born to mothers in famine-exposed populations. However, in recent years genetic studies have called elements of this theory into question. Birth weight has been shown to be highly heritable, with common genetic variants explaining ~40% of the population-level variation in birth weight[168]. Genetic influences on birth weight are the result of a combination of the individual's own genotype as well as the correlated genotype of the mother (r=~0.5). A recent study by Warrington *et al.* identified over 300 genetic variants associated with birth weight in a combined GWAS of own (foetal) and offspring (maternal) birth weight, and using structural equation models (SEM) they were able to quantify the contributions of foetal and maternal-specific effects of the associated variants[72]. Use of these variants in MR analyses showed that foetal birth weight-lowering genetic effects are causally associated with increased blood pressure in the individual. Conversely, maternal birth weight-lowering genetic effects, which may be thought of as approximating an adverse intra-uterine environment, showed no such association with blood pressure in offspring. Based on this the authors suggested that shared genetic factors, rather than adverse intra-uterine conditions, may be the primary attributable risk factor for poor cardio-metabolic health in low birth weight individuals. The relative contributions of foetal programming and shared genetic architecture between birth weight-lowering variants and metabolic outcomes remains an ongoing area of debate in the field.

4.1.3 Birth weight, BMI and body composition

Studies investigating associations between birth weight and BMI in adulthood have produced mixed results. Many have reported positive associations (i.e. higher birth weight conferring higher adult BMI)[165,169,170] while others have shown evidence of U- or J-shaped associations[171–173]. Whilst BMI is a convenient proxy for overall adiposity due to its ease of measurement in large cohorts, it is limited by the fact that it is also associated with both lean and bone mass. Fat distribution has been proposed to be a more accurate indicator of health risk independent of overall adiposity, and particularly fat accumulated in the abdominal region (so-called "central fat")[174,175]. Several studies have investigated associations between birth weight and body composition. Higher birth weight has shown positive associations with measures of fat-free mass (FFM; i.e. lean and bone mass) in children and adults[176–178]. Positive associations have also been reported for total fat mass (FM)[164,171,179] and for measures of central adiposity[180–182] in both high and low birth weight individuals. However, the majority of these studies have relied on approximations of central fat such as skinfold thickness and waist-to-hip ratio (WHR). While these provide readily-measured indicators of fat distribution, research has suggested that further categorisation of central fat into visceral adipose tissue (VAT) and subcutaneous adipose tissue (SAT) may be necessary to fully resolve the effects of central adiposity on metabolic health[183–185]. This is due to the greater metabolic activity of VAT as compared to SAT, making it a stronger predictor of metabolic dysfunction. Studies specifically investigating the effects of VAT and SAT have suggested that birth weight is inversely correlated with VAT in adults[100,186,187], which may partly explain the associations between low birth weight and poor health. However this has not been tested using genetic methods, which may provide further insights by offering evidence for a causal relationship between birth weight and patterns of fat distribution.

4.1.4 Areas of opportunity

The well-established associations between adiposity and metabolic health suggest this as a likely mechanism linking population variation in birth weight with differential disease risk. While studies using BMI as the measure of adiposity have produced mixed results, there is evidence for an effect of birth weight on patterns of fat distribution, which is a more meaningful clinical indicator of health risk. However, to date no study has systematically analysed this association using genetic methods. Moreover, investigations into regional fat distribution have largely relied on approximations using relatively crude metrics such as WHR and skinfolds. Dual-energy x-ray absorptiometry (DEXA) scanning affords the ability to measure body composition in far greater detail, by partitioning fat and lean mass into specific body compartments. In this chapter, I describe a study which utilises DEXA scan data from several large-scale cohort studies to investigate the association between birth weight and adult body composition. To begin, I perform meta-analysis of summary statistics from three individual GWAS of fat and lean mass in 14 body compartments to identify genetic determinants of body composition. In combination with published summary statistics of birth weight-associated variants, I use these data to provide new insights into the relationship between birth weight and body composition in adults. The specific objectives are:

1) To identify subsets of birth weight-associated genetic variants which have specific effects on compartmental fat and lean mass distribution;
2) To determine whether these identified subsets are enriched in unique biological pathways which may explain associated disease aetiology; and
3) To investigate whether maternal and foetal-specific genetic effects on birth weight have differential effects on compartmental body composition.

More broadly, the application of genetic methods to investigate the effect of birth weight on body composition may offer new insights into the roles of foetal programming and overlapping genetic determinants on health outcomes, further informing the debate.

4.2 METHODS

4.2.1 Participating studies

This study uses data from the UK Biobank, Fenland and EPIC-Norfolk studies. Descriptions of study participants and data collection methods are described in Chapter 2 of this thesis.

4.2.2 Assessment of overall and regional body fat and lean mass composition

Overall and regional body composition was assessed using DEXA imaging technology. DEXA technology is able to differentiate between fat, bone and lean tissue by estimating the differential attenuation of X-rays at two different energy levels[188]. While magnetic resonance imaging (MRI) and computerised tomography (CT) are considered the "gold standard" for measurement of fat and lean mass compartments, the time, cost and reduced radiation exposure of DEXA scans makes them more suitable for use in large-scale population-based studies. Furthermore, validation studies comparing fat and lean mass estimates derived from DEXA with MRI estimates show a high correlation of whole body fat and lean mass measurements between the two techniques (R=0.99 and 0.97, respectively)[189].

Software developed by DEXA manufacturers delimits seven distinct body regions: android, gynoid, legs, arms, trunk, head and pelvis[190]. These regions are depicted in **Figure 4.1**. In the android region, subcutaneous fat thickness is also estimated and used to calculate a variable for visceral fat mass by subtracting subcutaneous android fat mass from total android fat mass (**Figure 4.1 B**). In total, body composition measurements were available for nine fat mass compartments (android, gynoid, trunk, leg, arm, peripheral, visceral, subcutaneous and total body) and for six lean mass compartments (android, gynoid, arm, leg, appendicular and total body). Peripheral fat and appendicular lean mass were derived by adding the estimates from the arms and legs for fat and lean mass, respectively.

DEXA scanning for the Fenland and EPIC-Norfolk studies were conducted using the Lunar Prodigy advanced beam scanner (GE Healthcare, Madison, WI, USA), with enCORE software provided by the manufacturer used to assess images (version 14.10.022, GE Healthcare). Images were manually reviewed by trained technicians, with low quality images discarded. In Fenland, scanning was performed on 11,869 participants, with 11,741 passing quality control while in EPIC-Norfolk 5,547 (of 5,568) participants passed quality control. In the UK Biobank, DEXA scanning is performed as part of the ongoing UK Biobank Imaging Study[191] using the GE-Lunar iDXA (GE Healthcare, Madison, WI, USA), with images processed using the enCORE software. Data on 4,995 participants was available after application of quality control screening.

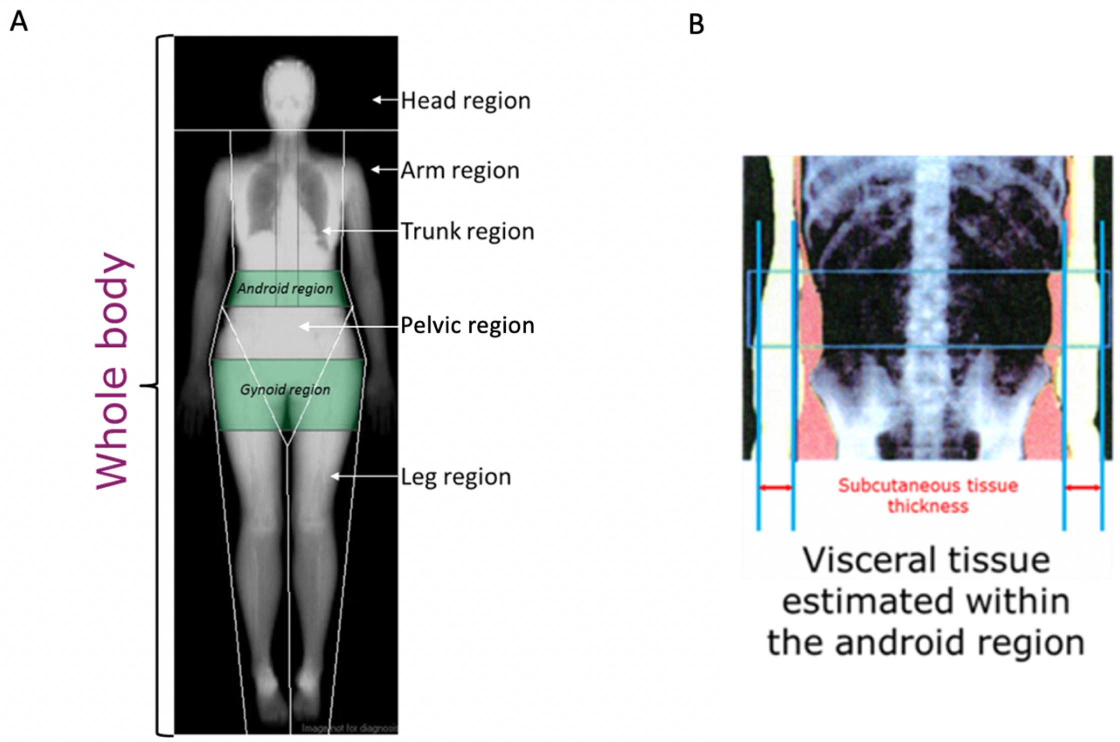


Figure 4.1: Body composition using DEXA. Panel **A** depicts whole body scan of soft tissues (i.e. fat and lean mass). The seven regions defined by the DEXA software are shown. Panel **B** shows a detailed image of the android region, used to measure subcutaneous fat thickness (are between the blue lines) and visceral adipose tissue, which is the subcutaneous fat subtracted from the total android fat mass. Source – The DAPA Measurement Toolkit (https://dapa-toolkit.mrc.ac.uk/).

4.2.3 GWAS and meta-analysis

Genotyping and imputation methods for Fenland, EPIC-Norfolk and UK Biobank are described in Chapter 2. All traits were natural log-transformed prior to GWAS and adjusted for age and study-specific covariates. The residuals from linear regression models were rank-based inverse normally transformed. All transformations were performed separately in men and women. Additionally, to account for overall body size as a potential confounding factor models were adjusted for total fat mass for fat compartments and for height squared in lean mass compartments.

GWAS on autosomal genetic variants were conducted separately for each of the studies, with separate analyses for each of the different genotyping platforms in the Fenland study (Axiom Biobank, SNP 5.0 and Core Exome – see Chapter 2). In the Fenland and EPIC-Norfolk studies GWAS were conducted using BGENIE v1.2[94] and adjusted for the first four genetic principle components to account for population substructure, while in UK Biobank GWAS was conducted using BOLT-LMM v2.3 and adjusted for the first ten genetic principle components as well as the genotyping chip. Quality control exclusions included: 1) variants with minor allele count <10; 2) insertions or deletions; 3) variants with >2 alleles present at a locus; 4) a HWE p-value < 10^{-8}; and 5) absolute value of beta or standard error >5. Further quality control measures were conducted using the R package 'Easy QC'[192] including comparison of observed allele frequencies and allele frequencies from 1000 Genomes CEU population, scatter plots of P-value versus Z-scores, quantile-quantile plots and genomic inflation plots.

To combine estimates across the three studies, fixed-effect meta-analysis were performed in METAL[109] using sex-combined models for each trait. Variants were excluded if they were not available in at least two of the datasets or were not in at least 50% of maximum sample size.

4.2.4 Clustered heat maps

Lookups of body composition GWAS summary statistics were performed for each of 305 variants which were independently associated with own (foetal) or offspring (maternal) birth weight ($P<5.0\times10^{-8}$) identified by Warrington *et al.*[72] Of these, 4 variants were on the X chromosome and one variant (rs7553582) had no effect estimate for maternal birth weight and were excluded. Effect estimates for the remaining variants were aligned separately for both foetal and maternal effects, such that the effect allele was considered to be the birth weight-increasing allele. For each birth weight variant, effect estimates and standard errors were used to construct Z-scores (= β/SE) from summary statistics for all body composition variables. These were then compiled to construct a distance matrix in order to perform unsupervised hierarchical clustering. Similarity between points was estimated by Euclidian distance, with a complete-linkage clustering method applied. Clustering was performed and heat maps were generated using the 'pheatmap' package in R.

As a sensitivity analysis, hierarchical clustering was also performed on birth weight variants classified as having foetal-only (N=83) or maternal-only effects (N=45) based on the SEM partitioning by Warrington et al.[72]

4.2.6 Mendelian randomisation analyses

To investigate whether birth weight has a causal influence on adult body composition, univariate and multivariate Mendelian randomisation analyses were conducted. The 300 birth weight autosomal variants were used to construct the genetic instruments for the exposure, modelling foetal and maternal effects separately. Estimates for the variants' effects on birth weight were obtained from the foetal and maternal-specific birth weight GWASs, and aligned such that the birth weight-increasing allele was considered the effect allele. For the univariate MRs, a two-sample IVW approach was used to regress variants' effect on birth weight against the effect estimate for each body composition variable (see Chapter 2). In the multivariate analyses, the effect of the foetal or maternal genome is modelled as a potential mediating effect as described by Burgess *et al.*[123]:

$$\beta_{Y_j} = \theta_F \beta_{F_j} + \theta_M \beta_{M_j} + \varepsilon_{D_j,} \qquad \varepsilon_{D_j} \sim N(0, se(\beta_{Y_j})^2) \quad [1]$$

Where β_{Y_j} , β_{F_j} and β_{M_j} are the estimates for the association between the jth variant and the outcome (i.e. the DEXA variable), foetal birth weight and maternal birth weight, respectively. θ_F and θ_M are the effects of the foetal and maternal genetic variants on the outcome, while ε_{D_j} represents the inverse variance weighting. Hence, by including the foetal or maternal genetic effects as a mediator, the specific (direct) effects of the maternal and foetal genetic variants on the outcome can be estimated.

4.3 RESULTS

4.3.1 Associations for individual birth weight variants with body composition variables

Each of 300 birth weight variants passed quality control measures for all body composition trait meta-analyses, and thus were available for use in this study. Associations between individual birth weight variants and body composition variables were limited, with 10 genome-wide significant variant-trait associations ($P<5.0\times10^{-8}$) (**Table 4.1**). Six of these associations were with 2 SNPs located in/near the gene for high motility group AT-hook 1 (HMGA1), a chromatin protein that is involved in regulation of gene transcription which has been associated with risk for T2D[193]. The birth weight-increasing alleles at both variants at this locus were associated with decreased peripheral fat and increased trunk fat. In addition the variant rs75034466 was associated with decreased leg fat and increased visceral fat. Another variant, rs1480470 locating in/near the gene for HMGA2, also showed associations with fat mass in multiple compartments. Here, the birth weight-increasing allele was associated with decreased trunk and android fat and increased peripheral fat. Variants at this locus have also been associated with a number of relevant anthropometric and metabolic traits including BMI-adjusted hip circumference[194], T2D[195] and polycystic ovary syndrome[196]. Finally, a foetal-only SNP located in/near *LOC339894/CCNL1* (rs1482852) was positively associated with subcutaneous fat mass.

Table 4.1: Birth weight variants showing significant associations with body composition variables

Variant	Chr:Pos	Nearest Gene	SEM annotation	Associated trait	β (S.E.)*	P-value
rs1480470	12:66,412,130	*HMGA2*	Foetal-only	Android fat	-0.06 (0.01)	4.9×10^{-9}
				Peripheral fat	0.06 (0.01)	1.9×10^{-9}
				Trunk fat	-0.06 (0.01)	5.0×10^{-9}
rs75034466	6:34,199,815	*HMGA1*	Foetal and Maternal	Leg fat	-0.14 (0.02)	1.9×10^{-8}
				Peripheral fat	-0.16 (0.02)	7.0×10^{-11}†
				Visceral fat	0.14 (0.03)	5.0×10^{-8}
				Trunk fat	0.16 (0.02)	4.5×10^{-10} †
rs75104038	6:34,190,104	*HMGA1*	Foetal and Maternal	Peripheral fat	-0.13 (0.02)	6.9×10^{-9}
				Trunk fat	0.12 (0.02)	2.1×10^{-8}
rs1482852	3:156,798,294	*LOC339894/CCNL1*	Foetal-only	Subcutaneous fat	0.08 (0.01)	7.2×10^{-14}

*Effects aligned to birth weight-increasing allele

† Significant after application of Bonferroni-correction for number of tests ($P<1.7\times10^{-10}$, $=5.0\times10^{-8}/300$)

4.3.2 Clustering identifies subsets of variants with specific effects on body composition

Unsupervised hierarchical clustering was applied to Z-score transformed association statistics for 290 birth weight loci across the 14 body composition variables (10 loci had opposite directions of effect in foetal and maternal birth weight GWAS and were excluded from the analysis). The clustered heat map is shown in **Figure 4.2**, with all effect estimates aligned to the birth weight-increasing allele. As expected, the primary clustering of body composition variables along the x-axis separates central fat compartments (visceral, android and trunk fat mass) from all other variables while a secondary cluster separates peripheral fat mass from lean mass compartments. Along the y-axis the optimal number of clusters was determined by the silhouette method, which revealed two distinct clusters. The first cluster (Cluster 1) was comprised of 124 SNPs which generally showed negative associations with central fat compartments and positive associations with peripheral fat mass compartments. Within this cluster, a subset of variants showing additional positive associations with lean mass compartment was also observed. The larger second cluster (Cluster 2; N=166 SNPs) appears to be more heterogeneous, with one subset of variants showing generally positive associations with central fat compartments and negative associations with peripheral fat mass and lean mass compartments, while other variants seem to have little or no effect on any of the variables.

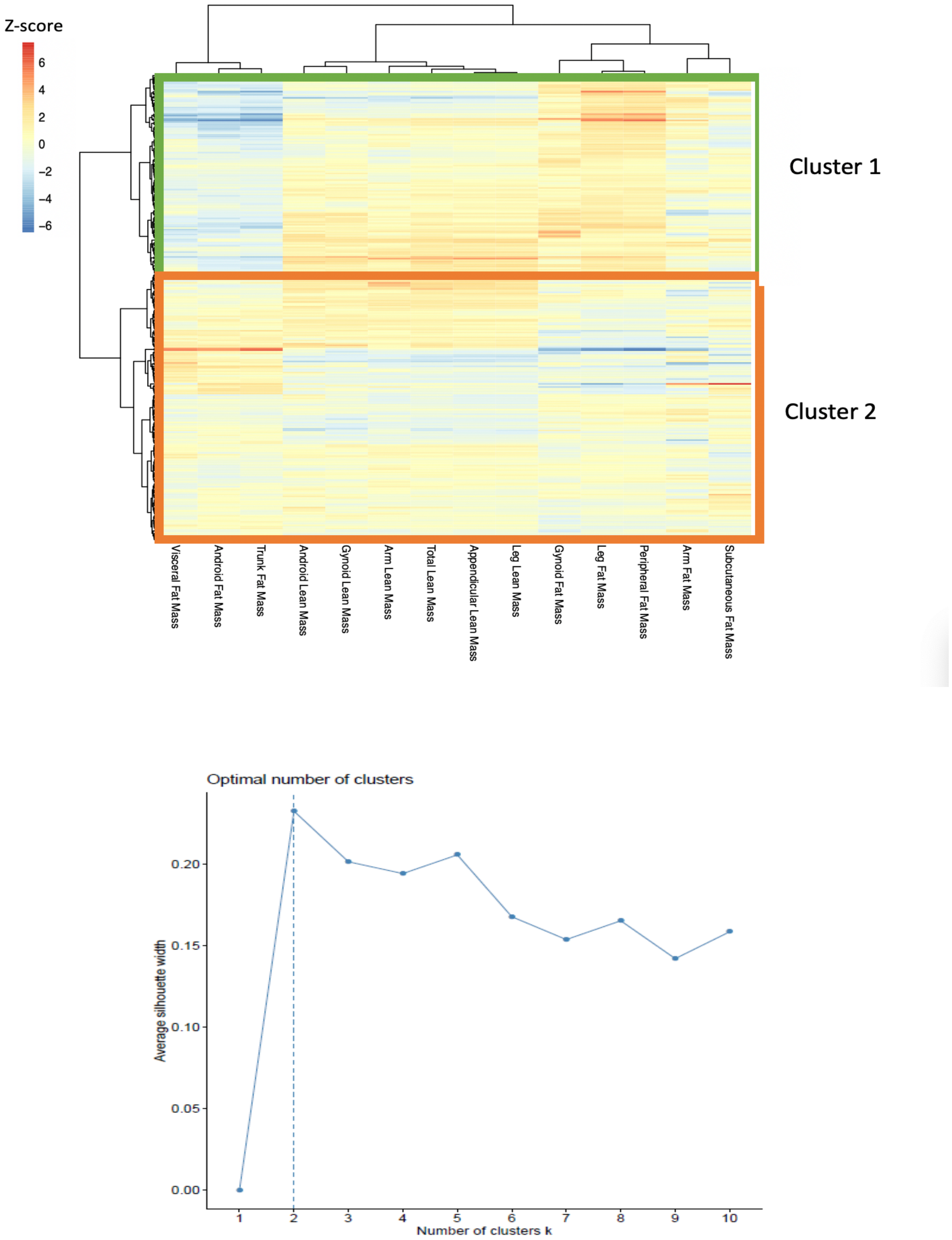


Figure 4.2: Clustered heat map of birth weight variants by body composition traits (top panel) with average silhouette plot for SNP cluster (bottom panel). The 290 birth weight variants with directionally concordant effects in maternal and foetal GWASs were used. Cell colours correspond to Z-scores for variant association with body composition variables, with red shades indicating positive associations and blue shades indicating inverse associations. All effects are aligned to the birth weight-increasing allele. The two distinct clusters of variants (along the y-axis) identified by the average silhouette method plot are depicted by the green box (cluster 1) and the orange box (cluster 2).

4.3.3 Functional enrichment analysis show distinct biological pathways underlying clusters

In order to make biological inferences about the subsets of variants identified by hierarchical clustering, functional enrichment analysis was conducted based on nearest gene annotation of SNPs belonging to each of the primary clusters. Enrichment analysis was conducted using STRING (v11.0)[197]. While this method is likely underpowered compared to gene-set enrichment analysis tools such as MAGENTA and DEPICT which use genome-wide data to identify enrichment, STRING is advantageous in this setting as it allows inference based on the smaller number of associated genes. STRING contains a database of protein-protein interactions derived from multiple sources including experimental work and text-mining of the literature. For this analysis, evidence for functional enrichment was based on functional enrichment in KEGG curated pathways. Genes in Cluster 1 showed evidence of enrichment (FDR<0.05) in 95 KEGG pathways, while Cluster 2 genes were enriched in 122 KEGG pathways (**Supplementary Table 4.1**). Of these, 20 pathways were unique to Cluster 1 and 47 were unique to Cluster 2. Among the pathways unique to Cluster 1 (the pathway characterised by variants with decrease central fat mass deposition in favour of peripheral deposition) are insulin signalling ($\text{FDR}=2.0\times10^{-3}$), carbohydrate absorption ($\text{FDR}=9.8\times10^{-3}$) and oxytocin signalling (FDR=0.04). Pathways unique to Cluster 2, characterised by variants associated with increased central fat deposition while decreasing peripheral fat mass and lean mass, included cocaine addiction ($\text{FDR}=4.3\times10^{-3}$), inflammatory bowel disease ($\text{FDR}=7.5\times10^{-3}$) and thyroid hormone synthesis (FDR=0.01)

4.3.4 Foetal and maternal-specific effects

Birth weight variants with foetal and maternal-specific effects did not appear to segregate based on the subsets of variants identified by hierarchical clustering. The proportion of variants classified as ‘foetal-only’ and ‘maternal-only’ by the SEM models were not significantly different between Clusters 1 and 2 (χ^2 p-value=0.98). To further investigate foetal and maternal-specific effects on body composition, hierarchical clustering was performed using the subsets of SNPs classified as ‘maternal-only’ (N=45 SNPs) or ‘foetal-only’ (N=83 SNPs) by Warrington *et al*[198]. Primary clustering of body composition variables (x-axis) was consistent with the full analysis for both foetal and maternal subsets, with central fat compartments separating from the remaining variables. Foetal birth weight-increasing SNPs appeared to show a general effect of decreasing fat mass in central compartments, compared with maternal SNPs which had a more heterogeneous effect on central fat (**Figure 4.3 a) and b)**). Maternal SNPs did appear to generally be associated with increased lean mass, with the primary clustering of maternal SNPs splitting those increasing from those decreasing lean mass.

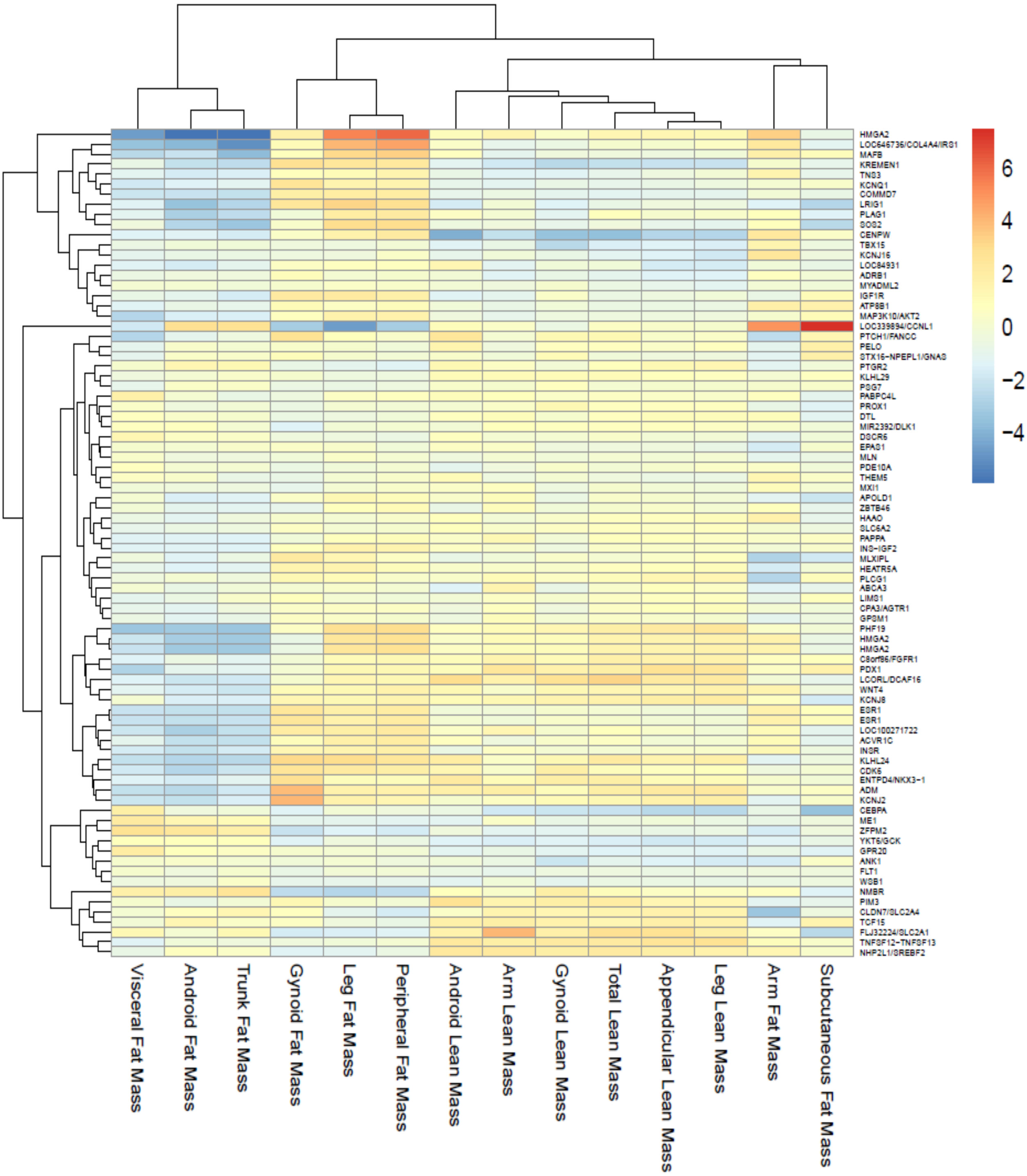


Figure 4.3 a): Clustered heat map of foetal-only birth weight variants by associations with body composition variables. A continuous chromatic scale is used to depict Z-score-transformed associations for birth weight variants across the body composition variables. Red shades depict positive associations while blue shades show negative associations.

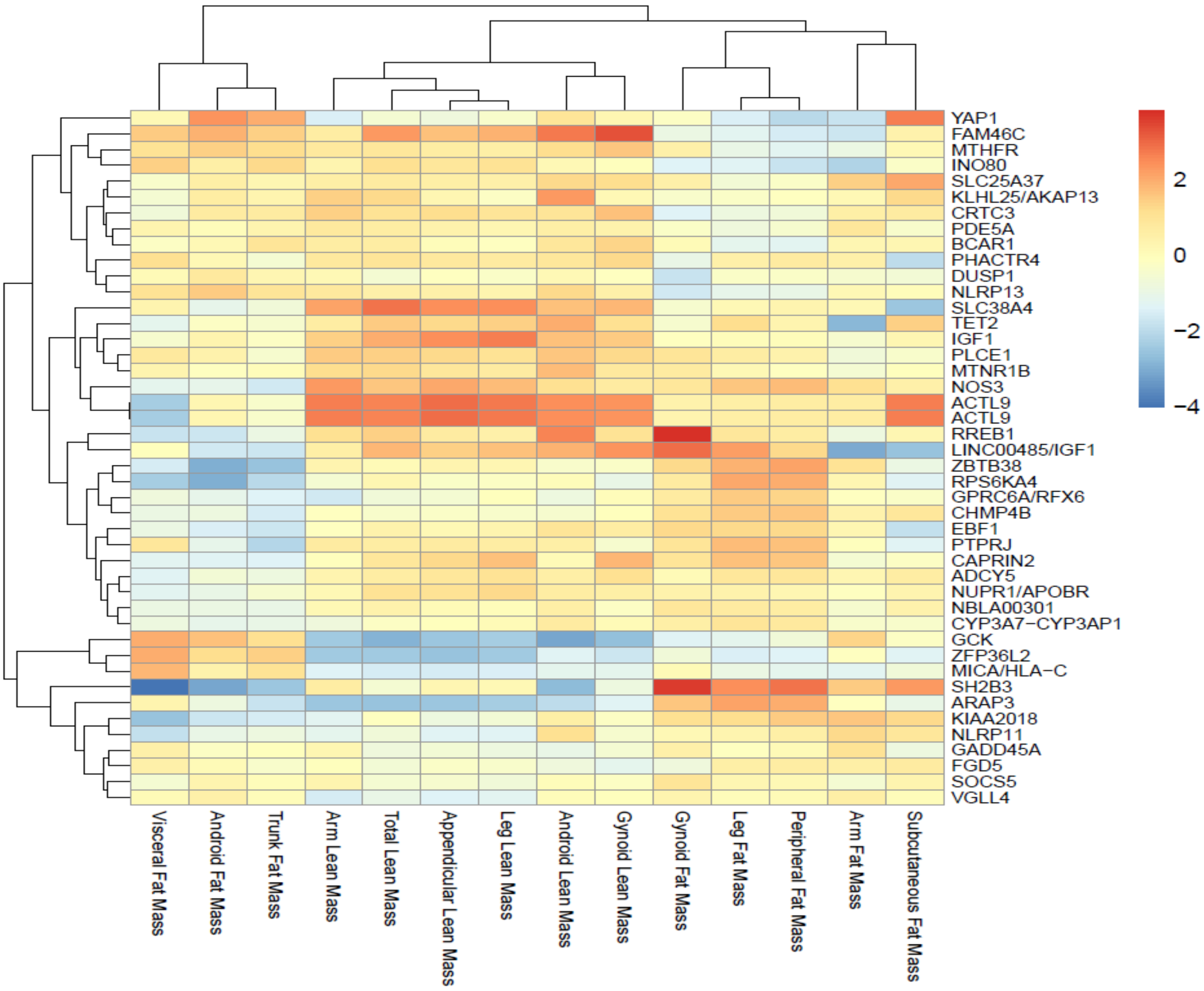


Figure 4.3 b): Clustered heat map of maternal-only birth weight variants by associations with body composition variables.

4.3.5 Mendelian randomisation analyses

Clustering methods are able to identify subsets of variants with similar patterns of association across multiple phenotypes. This type of analysis can yield valuable insights into the effects of individual and groups of variants. However, such techniques do not allow inference to be made regarding the overall effect of the exposure on the outcome of interest. In order to more generally assess the effect of birth weight on compartmental body composition, MR analyses were conducted using birth weight-associated variants as instrumental variables. Firstly, univariate analyses were run using effect estimates from foetal and maternal-specific birth weight GWAS as weights in separate models in order to gain an overall picture of the effect of birth weight on each body composition variable. However, as described above there is an approximately 50% overlap in the genomes of mother and offspring. Therefore in addition to the univariate models, multivariate models were also run in which the foetal or maternal genome effects on birth weight (depending on the model being considered) were adjusted for. Such analyses enable the investigation of the specific influence of foetal and maternal genetic effects of birth weight on compartmental body composition

Foetal effects

In univariate MR analyses, higher genetically predicted foetal birth weight was strongly associated with decreased fat mass in central body compartments ($\beta_{android}$= -0.32, P= 2.0×10^{-20}; β_{trunk}= -0.35, P= 1.9×10^{-24}; $\beta_{visceral}$= -0.29, P= 2.3×10^{-16}). In contrast, significant positive associations were observed for all lean mass compartments and all peripheral fat mass compartments with the exception of arm fat (P=0.21) and subcutaneous fat (P=0.88) (**Table 4.2**). Upon adjustment for maternal effects in the multivariate MR models, the significant effects observed for all fat mass compartments were completely attenuated (P>3.6×10^{-3}, =0.05/14 tests), with the exception of arm fat which became statistically significant (β= -0.23, P=3.0×10^{-3}). For lean mass variables, adjustment for maternal effects also attenuated the effects considerably; however suggestive positive associations (P<0.05) remained for android, gynoid, trunk and total lean mass.

Maternal effects

Univariate MR analyses using genetically predicted maternal birth weight as the exposure showed directionally consistent effects to the foetal instrument for all compartmental body composition traits, though with reduced effect sizes across all variables (**Table 4.2**). When adjusted for foetal effects, the associations with all lean mass variables are completely attenuated (minimum P=0.21). However, the associations with the majority of fat mass variables remained strongly significant in the multivariate MR models. Higher genetically predicted maternal birth weight conferred lower fat mass in central compartments including the android (β= -0.41, P=1.0×10^{-7}), trunk (β= -0.44, P=9.5×10^{-8}) and visceral (β= -0.37, P=1.5×10^{-6}) regions, independently of effects of the foetal genome. In contrast, higher maternally-driven birth weight was associated with increased fat deposition in peripheral compartments including the arm (β= 0.21, P=3.2×10^{-3}) and leg (β= 0.37, P=3.1×10^{-6}) regions as well as increased overall peripheral fat mass (β= 0.41, P=4.1×10^{-7}). The lone exception to this trend is gynoid fat mass which did not remain significantly associated after correction for the number of tests, though it remained suggestively positively associated, directionally consistent with the other peripheral fat compartments (β= 0.18, P=0.04).

Table 4.2: MR analyses for foetal and maternal specific effects on compartmental body composition

	Foetal Effects				Maternal Effects			
	Univariate[a]		Multivariate[b]		Univariate[a]		Multivariate[b]	
Trait	Beta (SE)	P-value	Beta (SE)	P-value	Beta (SE)	P-value	Beta (SE)	P-value
Central Fat Mass Compartments								
Android	-0.32 (0.04)	2.0×10^{-20}	0.12 (0.08)	0.13	-0.22 (0.04)	2.2×10^{-9}	-0.41 (0.08)	1.0×10^{-7}
Trunk	-0.35 (0.04)	1.9×10^{-24}	0.12 (0.09)	0.18	-0.26 (0.04)	1.1×10^{-11}	-0.44 (0.08)	9.5×10^{-8}
Visceral	-0.29 (0.04)	2.3×10^{-16}	0.11 (0.08)	0.17	-0.20 (0.04)	1.3×10^{-7}	-0.37 (0.07)	1.5×10^{-6}
Peripheral Fat Mass Compartments								
Arm	0.04 (0.04)	0.21	-0.23 (0.08)	3.0×10^{-3}	-0.05 (0.04)	0.17	0.21 (0.07)	3.2×10^{-3}
Gynoid	0.17 (0.04)	5.7×10^{-7}	0.02 (0.08)	0.78	0.16 (0.04)	3.4×10^{-5}	0.18 (0.07)	0.04
Leg	0.33 (0.04)	4.5×10^{-21}	-0.07 (0.09)	0.43	0.25 (0.04)	2.2×10^{-11}	0.37 (0.08)	3.1×10^{-6}
Peripheral.	0.32 (0.04)	1.2×10^{-20}	-0.12 (0.09)	0.16	0.23 (0.04)	1.6×10^{-9}	0.41 (0.08)	4.1×10^{-7}
Subcutan.	-0.01 (0.04)	0.88	-0.02 (0.07)	0.83	-0.01 (0.04)	0.77	0.01 (0.07)	0.93
Lean Mass Compartments								
Android	0.19 (0.04)	2.1×10^{-8}	0.19 (0.07)	0.01	0.24 (0.04)	3.0×10^{-10}	0.06 (0.07)	0.37
App.	0.16 (0.04)	2.8×10^{-6}	0.12 (0.07)	0.10	0.18 (0.04)	8.4×10^{-7}	0.07 (0.07)	0.28
Arm	0.12 (0.04)	2.0×10^{-20}	0.11 (0.07)	0.11	0.15 (0.04)	7.4×10^{-5}	0.04 (0.07)	0.54
Gynoid	0.13 (0.04)	1.5×10^{-4}	0.14 (0.07)	0.04	0.17 (0.04)	8.8×10^{-6}	0.03 (0.07)	0.68
Leg	0.16 (0.04)	2.6×10^{-6}	0.11 (0.07)	0.12	0.18 (0.04)	1.2×10^{-6}	0.08 (0,07)	0.23
Total	0.20 (0.04)	5.2×10^{-9}	0.16 (0.07)	0.03	0.23 (0.04)	5.0×10^{-10}	0.09 (0.07)	0.21
Trunk	0.20 (0.04)	1.7×10^{-8}	0.18 (0.07)	0.01	0.24 (0.04)	3.2×10^{-10}	0.06 (0.07)	0.35

a- Univariate models use effect estimates for foetal and maternal birth weight GWAS, aligned to the birth weight-increasing direction in each case.

b- Multivariate models were adjusted for the effect of genetically-predicted foetal or maternal birth weight.

4.4 DISCUSSION

Low birth weight is a significant risk factor for metabolic disease in adulthood, but the mechanisms driving the association are not completely understood. It has been hypothesised that foetal programming as an adaptation to adverse intra-uterine conditions drives correlations with disease, though this has been challenged by the results from recent genetic studies which have shown that shared genetic factors may underlie these associations. Results from non-genetic studies have suggested that individuals born at low birth weight are more susceptible to an unhealthy fat distribution, with a tendency to store fat centrally. This represents another plausible mechanism explaining the associations between variance in birth weight and disease risk, though this has not been assessed in a causal framework using genetic techniques. Here, I have leveraged data from GWAS for birth weight and regional body composition to provide new insights.

4.4.1 Effects of individual and subsets of birth weight variants on body composition

Individual birth weight variants showed few associations with the 14 compartmental body composition variables, with 10 genome-wide significant associations across 4 SNPs. Hierarchical clustering of birth weight variants was more informative, revealing subsets of SNPs with distinct patterns of association with fat and lean mass distribution. The clustering of body composition variables based on associations with birth weight separated the phenotypes into expected groups, with central fat, peripheral fat and lean mass forming distinct clusters. This provides confidence that the clustering of SNPs that arise are also likely to be biologically meaningful. Primary clustering of genetic variants appeared to be driven by two subsets with opposite directions of effect on central fat mass compartments (android, trunk and visceral) and peripheral fat mass in the gynoid and leg regions. Pathway analysis in STRING found non-overlapping pathways between these clusters, suggesting that different biological mechanisms may underlie the observed associations. Of interest, genes belonging to the cluster defined by inverse associations with central fat and a positive relationship with central fat were enriched in pathways related to insulin signalling and carbohydrate metabolism. Maternal glucose levels have been implicated as a key determinant of foetal growth via stimulation of foetal insulin production, with lower maternal glucose levels associated with lower birth weight in offspring[199]. Thus, shared genetic determinants may exist which govern lower foetal insulin and increased central fat, though further studies will be needed to confirm this.

4.4.2 Evidence for a casual effect of intra-uterine environment on central fat mass

Multivariate MR analyses suggested that higher genetically predicted birth weight conferred lower central fat mass (in the android, trunk and visceral regions) and higher peripheral fat mass (in the arm and leg regions as well as total peripheral fat). This clear delineation of predicted fat distribution between central and peripheral compartments provides support to the argument for foetal programming, as maternally-predicted birth weight adjusted for foetal effects can be thought of as representing the influence of the intra-uterine environment. This is in line with evidence from twin studies, which have shown that among monozygotic twin pairs, the heavier twin at birth tended to have a more favourable body composition as adults[200]. Thus, this line of evidence supports the idea that non-genetic developmental factors play an important role in determining body composition.

Conversely, foetal effects on birth weight only showed significant inverse association with arm fat mass, but not with any other fat mass compartment after adjustment for the effect of maternal influences. Foetal effects were positively correlated with all lean mass variables, with android, trunk and total lean mass showing nominally significant associations though none remained so after application of a more stringent threshold to account for the number of tests.

4.4.3 Strengths and limitations

This study expands on previous attempts to examine associations between birth weight and body composition. The primary strength of this study was the use of DEXA scans in the measurement of body composition, which allows for examination of specific patterns on fat and lean mass distribution which are unattainable though other methods used previously. In addition, the use of foetal and maternal GWAS data has allowed for the investigation of specific effects relating to shared genetics and the intra-uterine environment, by implementing multivariate MR analyses which has enabled us to make inferences regarding the specific contributions of these two factors to body composition.

Several limitations must also be mentioned which may affect interpretation of the presented results. Firstly, while adjusting for total body fat amongst fat mass variables and height for lean mass variables is necessary for the assessment of body compartments independent of total body size, the high heritability of these traits introduces the possibility of collider bias influencing the results[201]. Secondly, the weight limit for DEXA scanning (140kg) means that analysis is not possible for participants with extreme body mass. Finally, limiting the sample to only individuals of European descent prevents inference on other ethnic groups.

4.4.4 Conclusion

This chapter presents evidence for a causal association between low birth weight and a less metabolically healthy distribution of fat mass in adulthood. This appears to be driven largely by maternal factors which may influence birth weight via the intra-uterine environment. This provides support for the involvement of foetal programming in mediating the association between low birth weight and poor cardio-metabolic health in later life.

Chapter 5: Epigenetic Determinants of Body Mass Index and Puberty Timing

Contributions and Collaborations

This project was conceived in collaboration with members of the MRC Epidemiology Unit genetics forum. Much of the data used for this analysis was obtained from publicly available resources, which is indicated in the text. The meta-analysis of UK Biobank and GIANT GWAS data for BMI was conducted by Dr. Robert Scott (MRC Epidemiology Unit, University of Cambridge). The EWAS-FUSION pipeline was developed by Dr. Jing-Hua Zhao and Dr. Alexia Cardona (MRC Epidemiology Unit, University of Cambridge), who also generated the results for the BMI analysis. In addition, re-analysis of the genetic risk score for methylation was conducted by Dr. Felix Day (MRC Epidemiology Unit, University of Cambridge) and is included here for completeness. All other analyses were conducted by me, including lookups of meQTLs from the BIOS database, MR analyses, analysis and implementation of EWAS-FUSION data for BMI and puberty timing.

SUMMARY

The regulation of gene expression is a central theme of biology, and plays a critical role in development from conception through to adulthood. Epigenetic mechanisms, which involve covalent modifications to DNA structure without altering the underlying genetic code, are known to play an important role in this process, and include DNA methylation, histone modification and RNA-mediated silencing. Epigenetic changes are conserved through mitosis and meiosis, and are therefore heritable between generations. Additionally, new modifications can accumulate throughout an individual's life course, with factors such as smoking, diet and adverse intra-uterine conditions implicated. It has therefore been theorised that epigenetic modifications may at least partly explain the associations between adverse early life exposures and health risks. However, the nature of the association between epigenetic factors and early life exposures is still unclear. In addition, the direction of causality between epigenetic modifications and health traits is an area of ongoing debate. In this chapter, I address these two open questions and present evidence for a causal role of DNA methylation on both BMI and puberty timing by combining publicly available data sources with a new method for imputing methylation. In addition, I perform a re-analysis of previously reported results which have suggested an overarching directional effect of BMI on methylation. In doing so I demonstrate that this interpretation may have been an oversimplification, and provide evidence that a bi-directional relationship is more likely.

5.1 BACKGROUND

5.1.1 Epigenetic mechanisms in growth and development

Epigenetics describes a group of modifications to DNA that do not involve editing of the underlying sequence of base pairs that make up the genetic code[202]. Epigenetic modifications influence physiology by altering levels of gene expression, and are involved in many aspects of normal growth and development including cell differentiation[203], X chromosome inactivation[204] and genetic imprinting[205].

Mechanisms for epigenetic modifications to DNA include methylation, histone modifications and RNA-mediated silencing of genes. Among these, DNA methylation is perhaps the most widely studied and best characterised in humans. DNA methylation involves the addition of a methyl group (-CH_3) to cytosine (C) bases in DNA, via enzymes called DNA methyltransferases (DNMTs). The targeted cytosine bases are typically adjacent to guanine (G) bases, and are hence known as CpG sites, with the 'p' representing the phosphate linkage between the two. In the human genome, CpG sites are often clustered in the promoter and enhancer regions of genes in so-called 'CpG islands'. Methylation at these sites interferes with the binding of transcription factors, which can inhibit the initiation of transcription thereby reducing the level of mRNA produced for that gene and ultimately the abundance of the encoded protein (**Figure 5.1**).

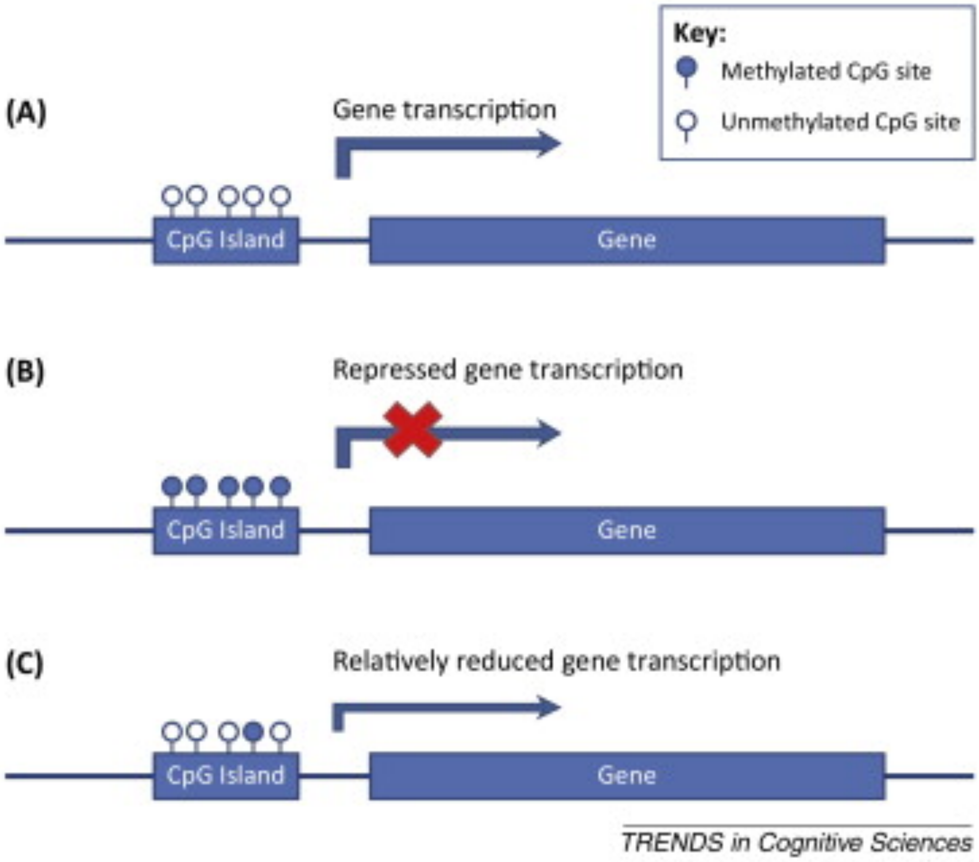


Figure 5.1: Schematic representation of the role of DNA methylation in the regulation of gene expression. Source: Nikolova *et al*.[206]

Epigenetic epidemiology

DNA methylation is highly conserved throughout mitosis in humans, and is also maintained through meiotic divisions in the production of gametes[207]. Consequently, DNA methylation is heritable between generations. De novo methylation can also occur throughout an individual's life course[208], with factors including smoking[209], diet[210] and intrauterine stress[211,212] influencing mean methylation levels at certain loci. It is estimated that between 60 and 80 percent of CpG sites in the human genome are methylated[202]. However this can differ over time, between individuals and in different tissues within the same individual[213]. The associations between DNA methylation and certain diseases are well established, as methylation is known to play a role in the development of cancer[214], systemic lupus erythromatosus[215] and fragile X syndrome[216]. In recent years, the development of high throughput technologies and increasingly sensitive assays has allowed sequencing of greater proportions of the human methylome (i.e. the pattern of methylation within the genome). Inter-individual variation in methylation levels at different genetic loci forms the basis for the emerging field of epigenetic epidemiology[217], facilitating the investigation into the role that variation in DNA methylation plays in common and complex health traits.

5.1.2 DNA methylation as a mechanism for DOHaD

DNA methylation and obesity

As discussed in Chapter 1, epigenetic factors are widely considered to be a mediator of the associations between adverse developmental conditions and subsequent risks of obesity and poor cardio-metabolic health (as described by the DOHaD hypothesis). The measurement of genome-wide variation in methylation levels has allowed this to be more rigorously investigated in recent years through the development of epigenome-wide association studies (EWAS). These studies have performed hypothesis-free tests at hundreds of thousands of differentially methylated loci, yielding new insights into the role of DNA methylation in determining BMI and body composition[218,219]. Dick *et al*. described the first such study, which identified three differentially methylated regions (DMRs) in the hypoxia-inducible factor 3 alpha (*HIF3A*) gene that were independently associated with BMI[220]. More recent studies have demonstrated the extent to which variation in methylation across the genome can influence BMI. The largest such study to date by Wahl *et al*. identified 187 genetic loci where differential methylation is associated with BMI[221,222]. In the context of early life determinants of health, the extensive variation of methylation within the population suggests that adverse early life exposures need not be extreme to produce lasting impacts on health.

Imprinting and puberty timing

Genomic imprinting describes a process whereby genes from a particular parent are expressed, while those from the other parent are silenced. As mentioned above, this process is known to be mediated by DNA methylation[223]. Genetic studies of AAM have demonstrated that genes associated with puberty timing are enriched in known imprinted regions of the genome, including rare variants in the *MKRN3* and *DLK1* loci which show associations with puberty timing when inherited from the father but not when inherited from the mother[35]. While further work is required to determine the downstream effects and health consequences of this observed association, this provides another line of evidence linking DNA methylation to early life exposures.

5.1.3 Cause or effect?

A fundamental question of epigenetic epidemiology relates to the direction of causality between methylation and health outcomes. Taking BMI as an example, it is plausible that increased methylation leads to changes in metabolic processes which facilitate the accumulation of excess fat. An alterative explanation is that increased adiposity causes differential levels of methylation at certain loci, and it has been suggested that this may explain (at least in part) the downstream metabolic consequences of obesity[221]. Disentangling this potential for reverse causality is an ongoing area of research, with several studies having attempted to provide clarity. Richmond *et al*. assessed the directionality of the association between methylation and BMI at the aforementioned *HIF3A* gene using long-term longitudinal follow-up, determining that methylation at this locus was likely the consequence rather than the cause of increased BMI[224]. Others have attempted to employ instrumental variable analysis to address the question of causality. Mendelson *et al*. used such an approach to identify one locus in the *SREBF1* gene which showed evidence of altered methylation having a causal influence on both BMI and coronary artery disease; though methylation was reported to be secondary to BMI at a greater number of loci[222]. In their EWAS study, Wahl *et al*. performed bi-directional MR analyses and concluded that methylation was largely a consequence of increased BMI rather than a cause[221]. In summation, the overall consensus in the literature appears to be for a direction of effect in which BMI is the causal factor and methylation the consequence. However it must be noted that these study designs have key limitations including relatively low sample sizes, low epigenome coverage and in some instances flawed methodology. As such, a definitive answer on the direction of causality remains elusive.

5.1.4 Areas of opportunity

Epigenetic modifications to DNA appear to be a plausible potential mediator of the association between an adverse developmental environment and adult disease. However, important questions remain unanswered. Associations between DNA methylation and both adiposity and puberty timing have been reported[225]. While considerable evidence for the association with BMI has been uncovered, the influence of genome-wide DNA methylation on puberty timing has not been studied. Furthermore, the question of the direction of causality between DNA methylation and health traits of interest remains open. Understanding the nature of these associations is essential if the findings from epigenetic studies are to be translated to potential clinical or public health benefits. The interest in the field of epigenetics has led to the generation of publicly available resources linking genetic variation to methylation levels. In this chapter I make use of several of these resources, as well as applying a novel method for the detection of genetic variants associated with methylation, to investigate the role of epigenetic factors in the association between early-life exposures and later-life health effects. The specific aims of this study are:

1) To re-analyse previously published data on the association between methylation and BMI, in order to test the consensus of an overarching causal effect of BMI on methylation;
2) To apply a novel method for performing EWAS to BMI which facilitates causal inference, in order to detect further genetic loci where methylation is casually associated with BMI; and

3) To apply the same novel EWAS method using GWAS summary statistics for age at voice breaking, enabling the first genome-wide discovery for genetic variants at which methylation is casually associated with puberty timing.

These analyses can yield new insights into the effects of DNA methylation, and provide further evidence for or against the hypothesis that early-life determinants of adult disease are mediated by epigenetic mechanisms.

5.2 METHODS

5.2.1 Identification of methylation loci causally associated with BMI

We first sought to identify individual genetic loci at which DNA methylation has a causal effect on BMI. Methylation variable positions (MVPs) previously reported to be associated with BMI were used in this analysis (BMI-MVPs). This comprised the 187 loci identified in an EWAS of individuals of European and Indian-Asian ancestry from the LOLIPOP, EPICOR and KORA studies (total N=10,261), as described by Wahl *et al*[221]. In their study, MVPs were considered significant if they met the following criteria: 1) reached epigenome-wide significance ($P<1.0\times10^{-7}$) in the discovery sample (N=5,387 individuals); 2) were associated with BMI with nominal significance ($P<0.05$) and directional consistency in the replication sample (N=4,874 individuals); and 3) achieved epigenome-wide significance in the combined analysis.

5.2.2 Identification of genetic variants associated with methylation

For each of the 187 BMI-MVPs, we next identified genetic variants which are associated with methylation at these positions. Data from the publicly available Biobank-based Integrative Omics Study (BIOS) browser (https://genenetwork.nl/biosqtlbrowser/) was used to identify *cis*-acting methylation quantitative trait loci (*cis*-meQTLs) associated with the BMI-MVPs. The methods for this study have been described previously[226]. In brief, DNA methylation in whole blood was analysed in 3,841 unrelated individuals from five Dutch biobanks. At least one *cis*-meQTL effect was observed at 139,566 of 405,709 MVPs, with a total of 272,037 independent *cis*-meQTL effects identified (up to 16 per MVP). Lookups were performed for each of the 187 BMI-MVPs in the BIOS database. For each, the *cis*-meQTL with the lowest P-value for association with methylation was selected as the sentinel SNP at that position and was brought forward for subsequent stages of analysis.

5.2.3 Meta-analysis of GIANT and UK Biobank BMI GWAS

In order to maximise the available sample size for MR analyses, a meta-analysis of GWAS summary statistics for BMI from the GIANT consortium[155] and UK Biobank[31] study (first release) was conducted by colleagues at the MRC Epidemiology Unit. Details on the GWAS methods for each individual study have been published previously. Summary statistics were obtained from each of these studies and meta-analysed using an inverse-variance weighted model in METAL[109]. Variants with and MAF < 0.01 were excluded. Summary statistics from this meta-analysis were used for subsequent analyses in this study.

5.2.4 Mendelian randomisation analyses for causal effects of methylation on BMI

To assess whether methylation has a causal effect on BMI, single gene Mendelian randomisation (MR) analyses were performed. For each of the *cis*-meQTLs identified for BMI-MVPs, the effect estimate for methylation at that locus was obtained from BIOS data, with the corresponding effect estimates and standard errors for each SNP on BMI obtained from the UK Biobank-GIANT meta-analysis. SNPs were aligned such that the effect allele was considered to be the one which increased methylation levels. Estimates for the causal effect of methylation on BMI, along with the standard errors, were then obtained for each variant using the following formulae:

$$\beta_{IVW} = \frac{\sum X_k Y_k \sigma_{Y_k}^{-2}}{\sum X_k^2 \sigma_{Y_k}^{-2}} \qquad [1]$$

$$SE = \sqrt{\frac{1}{X_k^2 \sigma_{Y_k}^{-2}}} \qquad [2]$$

where β_{IVW} is the inverse variance weighted estimate of the effect of methylation on BMI, X_k is the effect of the SNP k on methylation of the MVP, Y_k is the effect of the SNP on BMI, and $\sigma_{Y_k}^{-2}$ is the inverse variance of the effect of the SNP on BMI. Z-scores were produced based on these estimates by dividing the effect size by the standard error. A Bonferroni-corrected P-value threshold of based on the number of tests performed was used to assess statistical significance of individual variants.

5.2.5 Imputation of putatively causal methylation sites using EWAS-FUSION

Gusev *et al.*[227] and Mancuso *et al.*[228] have described a method called Functional Summary-based Imputation (FUSION), whereby measurements of gene expression are combined with summary association statistics from GWAS to identify associations between expression of genes and complex traits. Colleagues at the MRC Epidemiology Unit have extended this method to be applicable to methylation data, allowing for a more powerful approach for identifying putatively causal methylation-trait associations. A description of the method they developed is given here.

For a given methylation site, an EWAS statistic can be estimated as follows:

$$Z_{EWAS} = \frac{w\prime_{me} z_T}{\sqrt{w\prime_{me} V w_{me}}} \qquad [3]$$

where w_{me} is the weight associated with methylation, z_T is the GWAS summary z-statistic and V is the covariance matrix for z_T.

Weights were computed using methylation data from 1,117 individuals, profiled using Illumina Infinium Human Methylation 450k beadchips. Initial quality control was performed per manufacturer specifications, with methylation intensity values corrected using the Illumina Background Correction algorithm implemented using the R package '*minfi*'. Methylation intensities with a P-value ≥0.01 were set to missing and beta values were calculated for each methylation marker per sample. Marker call rates were calculated as the proportion of missing data at each position, and those with a call rate ≤0.95 were excluded from the analysis (N=8,775). Sites with multimodal distributions of signal intensity were identified using the R package '*Enmix',* with 3,295 sites excluded. Finally, a further 18,874 sites were excluded on the basis of probes mapping to >1 genomic location. The final dataset comprised methylation intensities at 442,920 autosomal MVPs.

Genotyping on the same 1,117 individuals was performed with the Affymetrix BioBank Axiom chips HapMap2 SNPs available from the genetic sequencing of these individuals were extracted using PLINK2 software according to the *cis*-positions of each probe. These were subsequently used to build weights analogous to gene expression as implemented in the computer software TWAS. Heritability of probes was calculated using GCTA software[229], with probes filtered based on a significance level of P<0.01. In total, 79,569 MVPs reached the threshold for significance.

5.2.6 Implementation of EWAS-FUSION for BMI and puberty timing

EWAS-FUSION software (https://github.com/jinghuazhao/EWAS-fusion) combines methylation weights with GWAS summary statistics to identify sites where methylation is associated with the trait of interest. The pipeline produces both an unadjusted model (full model) as well as a conditional (joint) model where associated variants are adjusted for the effects of neighbouring loci. EWAS-FUSION is analogous to an MR analysis, as the weights used to predict methylation are genetically predicted by SNPs, thus approximating an instrumental variable analysis. Therefore, any methylation-trait associations identified by EWAS-FUSION may be interpreted as being causal. We applied summary statistics from both the GIANT-UK Biobank BMI meta-analysis as well as the voice breaking meta-analysis (described in Chapter 3) to EWAS-FUSION, in order to identify loci which are causally associated with these traits. A Bonferroni-corrected P-value threshold of $P<6.3 \times 10^{-7}$ (=0.05/79,569) was used to determine statistical significance.

5.3 RESULTS

5.3.1 MR analyses of known BMI-MVPs reveal new casual associations

For the 187 known BMI-MVPs queried in the BIOS data, a total of 279 significantly associated *cis*-meQTLs ($P<2.7\times10^{-4}$, =0.05/187) were identified. 131 (70%) of the BMI-MVPs had at least one associated *cis*-meQTL. Of these, two of the 131 sentinel *cis*-meQTLs were not available in the GIANT-UK Biobank BMI meta-analysis and were excluded, leaving a total of 129 *cis*-meQTLs with data available for both methylation and BMI effects (**Supplementary Table 5.1**). Twenty-two of the sentinel meQTLs were associated with BMI with nominal significance ($P<0.05$), though none remained associated at a genome-wide significance level ($P<5.0\times10^{-8}$).

Single-SNP MR analyses identified 22 loci showing at least nominal ($P<0.05$) evidence for a causal effect of methylation on BMI (**Table 5.1**). Three of these were statistically significant after application of a Bonferroni-corrected P-value threshold: rs115616784 (meQTL for cg26663590 in/near *NFATC2IP*, $P=2.7\times10^{-7}$); rs6778735 (meQTL for cg00108715 in/near *NT5DC2*, $P= 1.1\times10^{-6}$); and rs2747429 (meQTL for cg00094412 in/near *GABBR1*, $P=8.5\times10^{-6}$). Methylation at the *NFATC2IP* locus has previously been implicated as having a causal effect on BMI[221], while *NT5DC2* and *GABBR1* represent novel findings. Both *NFAT2CIP* and *NT5DC2* show positive associations with BMI, (i.e. increased methylation at these loci confer higher BMI). Conversely the association for *GABBR1,* which encodes the receptor for gamma-aminobutyric acid (GABA, the main inhibitory neurotransmitter) was in the opposite direction (increased methylation leading to decreased BMI). In summary, this repeated analysis of existing data using an independent sample to identify meQTLs as instrumental variables confirmed the one casual association of methylation on BMI (*NFAT2CIP*) reported in the analysis of the same data by Wahl *et al*[221]., while identifying two further examples with evidence for causation in this direction (at *NT5DC2* and *GABBR1*).

Table 5.1: **Previously identified BMI-MVPs showing causal effect on BMI using BIOS data**

					IVW MR estimate	
MVP	Chr: pos	Nearest gene	Sentinel meQTL	EA/OA	Z-score	P-value
cg26663590	16: 28,959,310	*NFATC2IP*	rs115616784	A/G	5.18	2.2× 10^{-7}†
cg00108715	3: 52,565,015	*NT5DC2*	rs6778735	T/C	4.87	1.1× 10^{-6}†
cg00094412	6: 29,592,854	*GABBR1*	rs2747429	C/T	-4.45	8.5× 10^{-6}†
cg06559575	12: 53,490,352	*IGFBP6*	rs10876407	A/T	-3.47	5.2× 10^{-4}
cg07682160	19: 18,959,935	*UPF1*	rs7250622	G/A	-3.05	2.3× 10^{-3}
cg05063895	16: 2,073,518	*SLC9A3R2*	rs28698483	C/T	3.00	2.7× 10^{-3}
cg27614723	15: 92,399,897	*SLCO3A1*	rs7165398	C/T	2.83	4.6× 10^{-3}
cg03523676	14: 24,540,235	*CPNE6*	rs7146599	G/A	2.73	6.3× 10^{-3}
cg16578636	10: 92,987,457	*PCGF5*	rs2648718	T/C	2.57	0.01
cg10717869	1: 205,780,912	*SLC41A1*	rs4396169	T/C	2.50	0.01
cg08726900	16: 89,550,474	*ANKRD11*	rs2019604	G/T	2.45	0.01
cg00431050	10: 103,985,730	*ELOVL3*	rs61873698	T/G	-2.44	0.01
cg23232188	3: 121,556,543	*EAF2*	rs9854539	G/A	2.26	0.02
cg12593793	1: 156,074,135	*LMNA*	rs915179	G/A	2.26	0.02
cg08443038	16: 89,006,877	*CBFA2T3*	rs491224	A/G	2.14	0.03
cg00863378	16: 56,549,757	*BBS2*	rs72814488	C/T	-2.12	0.03
cg24679890	19: 17,246,356	*MYO9B*	rs111366154	T/C	2.10	0.04
cg16163382	2: 37,938,640	*CDC42EP3*	rs9808547	C/G	-2.07	0.04
cg26403843	5: 158,634,085	*RNF145*	rs6556405	C/T	2.06	0.04
cg06192883	15: 52,554,171	*MYO5C*	rs71472932	A/G	-2.05	0.04
cg10438589	4: 14,531,493	*LINC00504*	rs16890352	G/A	2.05	0.04
cg09777883	11: 112,093,696	*BCO2*	rs67245191	A/G	2.03	0.04
cg02650017	17: 47,301,614	*PHOSPHO1*	rs850523	C/T	-1.97	0.05
cg16846518	3: 128,062,608	*EEFSEC*	rs2687729	G/A	1.93	0.05
cg18219562	17: 41,773,643	*MEOX1*	rs1107747	C/T	-1.93	0.05

† Significant after correction for multiple tests.
EA = effect allele; OA = other allele

5.3.2 Novel BMI-associated loci identified in EWAS-FUSION

In the full (unadjusted) EWAS-FUSION model, 762 MVPs were significantly associated with BMI (P< 6.3×10^{-7}), of which 68 were independently associated in the conditional (joint) model (**Table 5.2 and Figure 5.2**). Of these, 53 are mapped to a distance of greater than 1MB from any of the 187 previously identified MVPs and are therefore likely to represent novel associations; though this could not be confirmed by measurement of LD as methylation data is not conducive to this type of analysis. Among the loci showing the strongest associations were two MVPs located in/near genes which have been associated with central regulation of appetite and energy balance in relation to obesity. Cg13314394 in/near *TMEM18*[230] (P_{joint}=7.6×10^{-79}) is proximal to the top

GIANT-UKBB GWAS SNP in the locus (rs13021737; **Table 5.2**) while the most significant GWAS SNP for this MVP as identified by EWAS-FUSION, rs12714415 is in strong LD ($r^2 = 0.99$) with this SNP (**Figure 5.3**). Additionally, cg09781307 is proximal to rs11030104 in/near the antisense strand of brain-derived neurotrophic factor (*BDNF-AS*; $P_{joint}=2.4\times10^{-41}$). Increased methylation at both of these loci conferred increased BMI.

MVPs were enriched in known BMI-associated genomic regions, as the nearest gene for 22 of the 68 independent loci was also one of the 97 BMI-associated loci identified by Locke *et al*[155]. In total, 29 (42.6%) of the independent BMI-MVPs identified in EWAS-FUSION are located within 1 MB of a known BMI GWAS signal, indicating significant enrichment compared to the background rate of 7.8% for all MVPs (Fisher's exact test P-value = 8.2×10^{-16}).

5.3.3 Lookups of known BMI-MVPs in EWAS-FUSION shows consistent direction of effects

The 187 previously reported BMI-MVPs were queried in the EWAS-FUSION BMI results, of which 59 were available. Directional consistency was observed for 34 MVPs (57.6%) in EWAS-FUSION. Three of these previously reported BMI-MVPs showed evidence for a causal effect of methylation on BMI. Two of these: cg26663590 (in/near *NFAT2CIP*, $P=5.0\times10^{-19}$) and cg00108715 (in/near *NT5DC2*, $P=1.0\times10^{-5}$), were identified as significantly associated with BMI in the single-SNP MR analyses and had consistent effect sizes between the two methods. The third MVP identified from the single-SNP MRs, cg00094412 (in/near *GABBR1*) was not available in EWAS-FUSION. However another variant, cg26403843 (in/near *RNF145*), was significantly associated with BMI in EWAS-FUSION ($P=5.7\times10^{-4}$) with a directionally consistent effect. Therefore in total, 4 out of 187 known BMI-MVPs show evidence for a causal effect of methylation on BMI between the two methods applied in this study, compared to the one reported in the paper.

Table 5.2 MVPs associated with BMI from EWAS-FUSION conditional model

			EWAS-FUSION (Joint Model)			Nearest BMI SNP[b]		
MVP	Chr: pos	Nearest Gene	Beta (S.E.)	P-value	≤1 MB from known BMI-MVP[a]	SNP	Gene	Dist. (b.p.)
cg13314394	2: 628,958	*TMEM18*	-19.0 (1.0)	7.6×10^{-79}	-	rs13021737	*TMEM18*	3,390
cg09781307	11: 27,648,324	*BDNF-AS*	-13.0 (1.0)	2.4×10^{-41}	-	rs11030104	*BDNF*	36,193
cg06941159	16: 28,875,210	*SH2B1*	12.0 (1.0)	9.1×10^{-35}	Yes	rs2650492	*SBK1*	541,799
cg17351154	19: 46,181,428	*GIPR*	12.1 (1.0)	7.0×10^{-34}	Yes	rs2287019	*QPCTL*	20,744
cg12177314	1: 177,841,709	*SEC16B*	11.4 (1.0)	6.0×10^{-30}	-	rs543874	*SEC16B*	47,771
cg00337662	15: 68,113,328	*SKOR1*	-11.0 (1.0)	4.1×10^{-29}	-	rs16951275	*MAP2K5*	36,160
cg08875605	3: 185,798,025	*ETV5*	11.2 (1.0)	7.0×10^{-29}	-	rs1516725	*ETV5*	25,979
cg15604733	6: 51,295,164	*PKHD1*	-11.0 (1.0)	3.8×10^{-28}	-	rs2207139	*TFAP2B*	449,674
cg06117072	6: 50,791,385	*TFAP2B*	-11.0 (1.0)	3.4×10^{-27}	-	rs2207139	*TFAP2B*	54,105
cg00760992	12: 50,223,054	*LOC100286844*	9.9 (1.0)	5.3×10^{-23}	-	rs7138803	*BCDIN3D*	24,414
cg18129748	3: 49,941,408	*MST1R*	-9.4 (1.0)	7.4×10^{-21}	-	> 1 MB	-	-
cg10061532	1: 201,886,748	*LMOD1*	9.1 (1.0)	5.8×10^{-20}	Yes	rs2820292	*NAV1*	102,461
cg09256413	1: 72,566,690	*NEGR1*	-8.4 (1.0)	3.2×10^{-17}	-	rs3101336	*NEGR1*	184,495
cg24210813	17: 46,669,644	*HOXB-AS3*	8.2 (1.0)	3.0×10^{-16}	Yes	> 1 MB	-	-
cg27466709	17: 1,844,037	*RTN4RL1*	7.9 (1.0)	2.4×10^{-15}	Yes	rs9914578	*SMG6*	161,099
cg22814740	1: 78,442,554	*FUBP1*	7.7 (1.0)	1.5×10^{-14}	-	rs12401738	*FUBP1*	223,205
cg14672496	18: 21,087,552	*C18orf8*	-7.4 (1.0)	1.5×10^{-13}	-	rs1808579	*C18orf8*	17,336
cg15770687	7: 76,625,569	*DTX2P1-UPK3BP1-PMS2P11*	7.3 (1.0)	2.8×10^{-13}	-	rs2245368	*PMS2L11*	17,426
cg00927699	21: 46,493,289	*SSR4P1*	7.2 (1.0)	6.6×10^{-13}	-	> 1 MB	-	-
cg06632762	1: 110,230,545	*GSTM1*	-7.2 (1.0)	6.7×10^{-13}	Yes	rs17024393	*GNAT2*	75,857
cg16926213	1: 1,841,314	*CALML6*	7.1 (1.0)	1.7×10^{-12}	Yes	> 1 MB	-	-
cg10026317	1: 47,694,919	*TAL1*	6.8 (1.0)	1.4×10^{-11}	-	rs977747	*TAL1*	10,242
cg16046214	5: 170,289,848	*RANBP17*	-6.7 (1.0)	2.6×10^{-11}	Yes	> 1 MB	-	-
cg13699650	9: 126,102,244	*CRB2*	-6.6 (1.0)	4.5×10^{-11}	-	> 1 MB	-	-

Table 5.2 (Continued)

			EWAS-FUSION (Joint Model)			Nearest BMI SNP[b]		
MVP	Chr: pos	Nearest Gene	Beta (S.E.)	P-value	≤1 MB from known BMI-MVP[a]	SNP	Gene	Dist. (b.p.)
cg02370877	8: 76,316,876	*HNF4G*	6.6 (1.0)	5.0×10^{-11}	-	rs17405819	*HNF4G*	489,708
cg00865973	4: 102,709,336	*BANK1*	-6.6 (1.0)	5.2×10^{-11}	-	rs13107325	*SLC39AB*	479,373
cg13634994	9: 129,452,574	*LMX1B*	-6.6 (1.0)	5.5×10^{-11}	-	rs10733682	*LMX1B*	8,340
cg15979035	14: 93,698,870	*UBR7*	-6.5 (1.0)	7.3×10^{-11}	-	> 1 MB	-	-
cg09002922	5: 87,956,389	*LINC00461*	-6.7 (1.1)	2.2×10^{-10}	-	> 1 MB	-	-
cg08321942	19: 34,310,625	*KCTD15*	-6.3 (1.0)	2.9×10^{-10}	-	rs29941	*KCTD15*	1,093
cg11784887	8: 8,729,819	*MFHAS1*	-6.3 (1.0)	3.0×10^{-10}	-	> 1 MB	-	-
cg22399598	7: 74,570,700	*NCF1C*	-6.6 (1.1)	6.7×10^{-10}	Yes	rs1167827	*HIP1*	592,469
cg11822372	1: 151,115,635	*SEMA6C*	-6.2 (1.0)	7.6×10^{-10}	-	> 1 MB	-	-
cg07177395	14: 79,968,686	*NRXN3*	-6.1 (1.0)	8.9×10^{-10}	-	rs7141420	*NRXN3*	69,232
cg04231319	10: 21,824,447	*MLLT10*	-6.1 (1.0)	1.0×10^{-9}	-	>1 MB	-	-
cg14263021	9: 16,039,921	*CCDC171*	6.1 (1.0)	1.4×10^{-9}	-	rs4740619	*C9orf93*	405,595
cg14418213	9: 33,814,450	*UBE2R2*	6.0 (1.0)	2.3×10^{-9}	Yes	> 1 MB	-	-
cg23702848	14: 104,172,109	*XRCC3*	6.0 (1.0)	2.6×10^{-9}	-	> 1 MB	-	-
cg03641066	10: 77,542,423	*C10orf11*	5.9 (1.0)	3.4×10^{-9}	-	> 1 MB	-	-
cg02407048	7: 77,046,134	*PION*	5.9 (1.0)	4.5×10^{-9}	-	> 1 MB	-	-
cg26115667	14: 103,294,656	*TRAF3*	5.8 (1.0)	8.5×10^{-9}	-	> 1 MB	-	-
cg09716613	13: 33,000,534	*N4BP2L1*	-5.7 (1.0)	1.2×10^{-8}	-	> 1 MB	-	-
cg26487582	4: 25,322,200	*ZCCHC4*	-5.7 (1.0)	1.5×10^{-8}	-	> 1 MB	-	-
cg04839131	7: 150,644,715	*KCNH2*	-5.6 (1.0)	1.7×10^{-8}	-	> 1 MB	-	-
cg16409650	13: 96,744,161	*HS6ST3*	-5.6 (1.0)	1.7×10^{-8}	-	> 1 MB	-	-
cg22722731	4: 140,781,201	*MAML3*	5.6 (1.0)	2.3×10^{-8}	-	> 1 MB	-	-
cg22274539	12: 103,696,209	*C12orf42*	5.6 (1.0)	2.3×10^{-8}	-	> 1 MB	-	-
cg09335911	10: 100,027,962	*LOXL4*	-5.5 (1.0)	3.8×10^{-8}	-	> 1 MB	-	-

Table 5.2 (Continued)

			EWAS-FUSION (Joint Model)			Nearest BMI SNP[b]		
MVP	Chr: pos	Nearest Gene	Beta (S.E.)	P-value	≤1 MB from known BMI-MVP[a]	SNP	Gene	Dist. (b.p.)
cg25574965	4: 52,728,090	*DCUN1D4*	-5.5 (1.0)	3.9×10^{-8}	-	> 1 MB	-	-
cg19365176	12: 116,963,394	*LINC00173*	-5.5 (1.0)	4.2×10^{-8}	-	> 1 MB	-	-
cg03309328	12: 99,525,864	*ANKS1B*	5.5 (1.0)	4.7×10^{-8}	-	> 1 MB	-	-
cg08529931	10: 88,124,564	*GRID1*	-5.5 (1.0)	4.8×10^{-8}	-	rs7899106	*GRID1*	713,660
cg18884555	22: 38,610,234	*MAFF*	-5.5 (1.0)	5.0×10^{-8}	Yes	NA	*NA*	NA
cg11469321	4: 104,021,294	*BDH2*	5.4 (1.0)	7.0×10^{-8}	-	rs13107325	*SLC39A8*	832,585
cg24860938	13: 27,998,603	*GTF3A*	5.4 (1.0)	8.3×10^{-8}	-	> 1 MB	-	-
cg15400367	13: 112,171,862	*TEX29*	-5.4 (1.0)	8.5×10^{-8}	-	> 1 MB	-	-
cg14191369	9: 130,996,177	*DNM1*	5.3 (1.0)	1.0×10^{-7}	-	> 1 MB	-	-
cg05198960	12: 133,309,102	*ANKLE2*	-5.2 (1.0)	1.8×10^{-7}	-	> 1 MB	-	-
cg00647317	7: 50,633,725	*DDC*	-5.2 (1.0)	2.1×10^{-7}	-	> 1 MB	-	-
cg14750066	12: 41,582,063	*PDZRN4*	5.2 (1.0)	2.1×10^{-7}	Yes	> 1 MB	-	-
cg10869879	12: 56,474,569	*ERBB3*	5.2 (1.0)	2.4×10^{-7}	-	> 1 MB	-	-
cg02138778	20: 25,172,755	*ENTPD6*	5.1 (1.0)	2.6×10^{-7}	-	> 1 MB	-	-
cg16888547	7: 3,2339,497	*PDE1C*	-5.1 (1.0)	2.9×10^{-7}	-	> 1 MB	-	-
cg22466678	12: 89,749,033	*DUSP6*	5.1 (1.0)	3.4×10^{-7}	-	> 1 MB	-	-
cg19619956	5: 176,967,557	*FAM193B*	-5.0 (1.0)	4.7×10^{-7}	Yes	> 1 MB	-	-
cg19076659	5: 137,688,057	*KDM3B*	5.0 (1.0)	5.1×10^{-7}	-	> 1 MB	-	-
cg06670463	4: 103,749,966	*UBE2D3*	5.0 (1.0)	6.0×10^{-7}	-	rs13107325	*SLC39A8*	561,257
cg08309041	10: 134,002,714	*DPYSL4*	5.0 (1.0)	6.0×10^{-7}	-	> 1 MB	-	-

a - based on MVPs identified in EWAS by Wahl *et al.* (ref. 221) within 500kb either side of CpG location
b - based on nearest BMI-associated SNP identified in GIANT by Locke *et al.* (ref. 155)

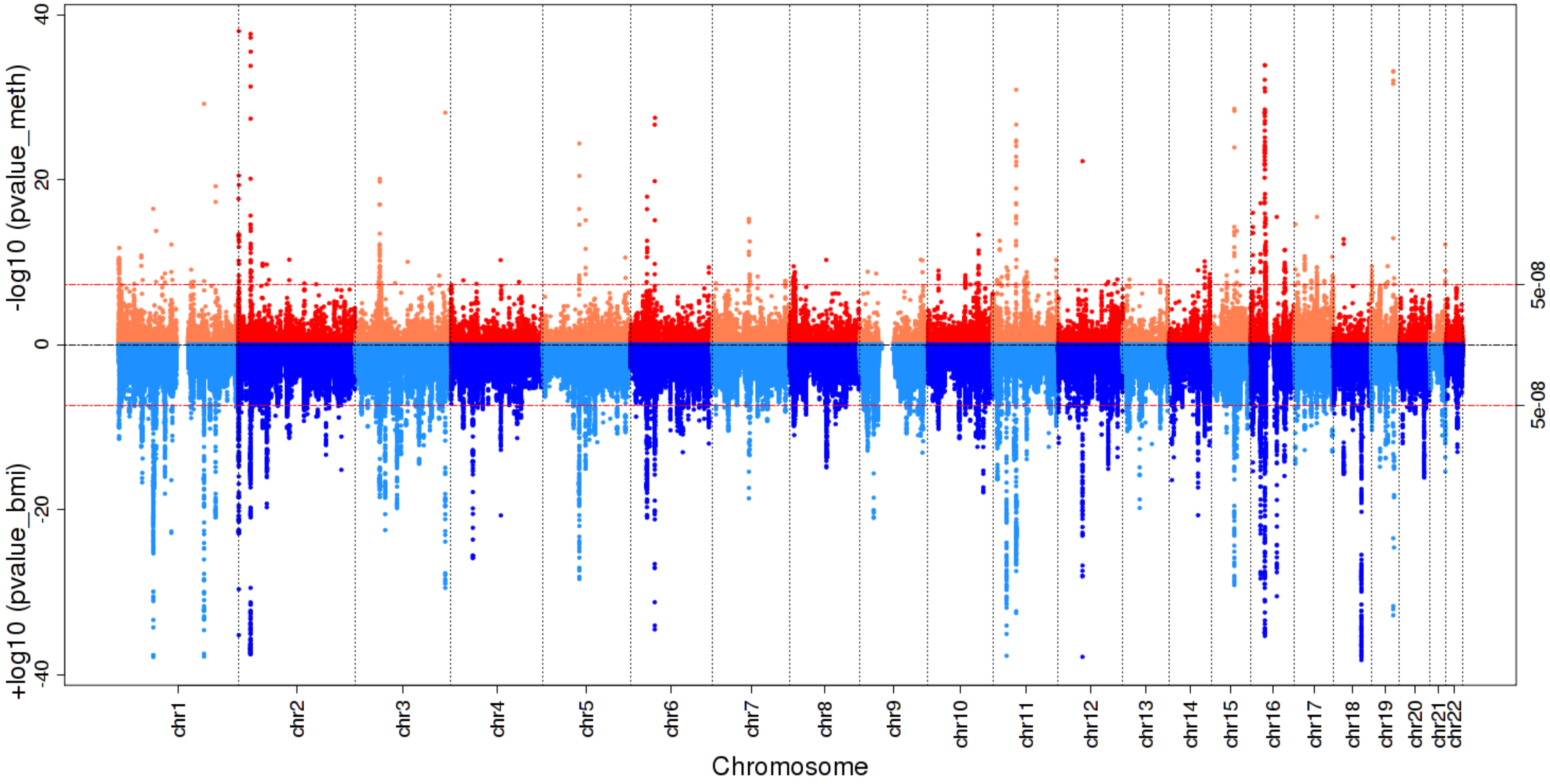


Figure 5.2: Miami plot comparing effect estimates from EWAS-FUSION (red shades) and GIANT-UKBB BMI meta-analysis (blue shades). $-\log_{10}$ P-values are plotted for each variant. Red dashed lines indicate genome-wide significance threshold ($P<5.0\times10^{-8}$).

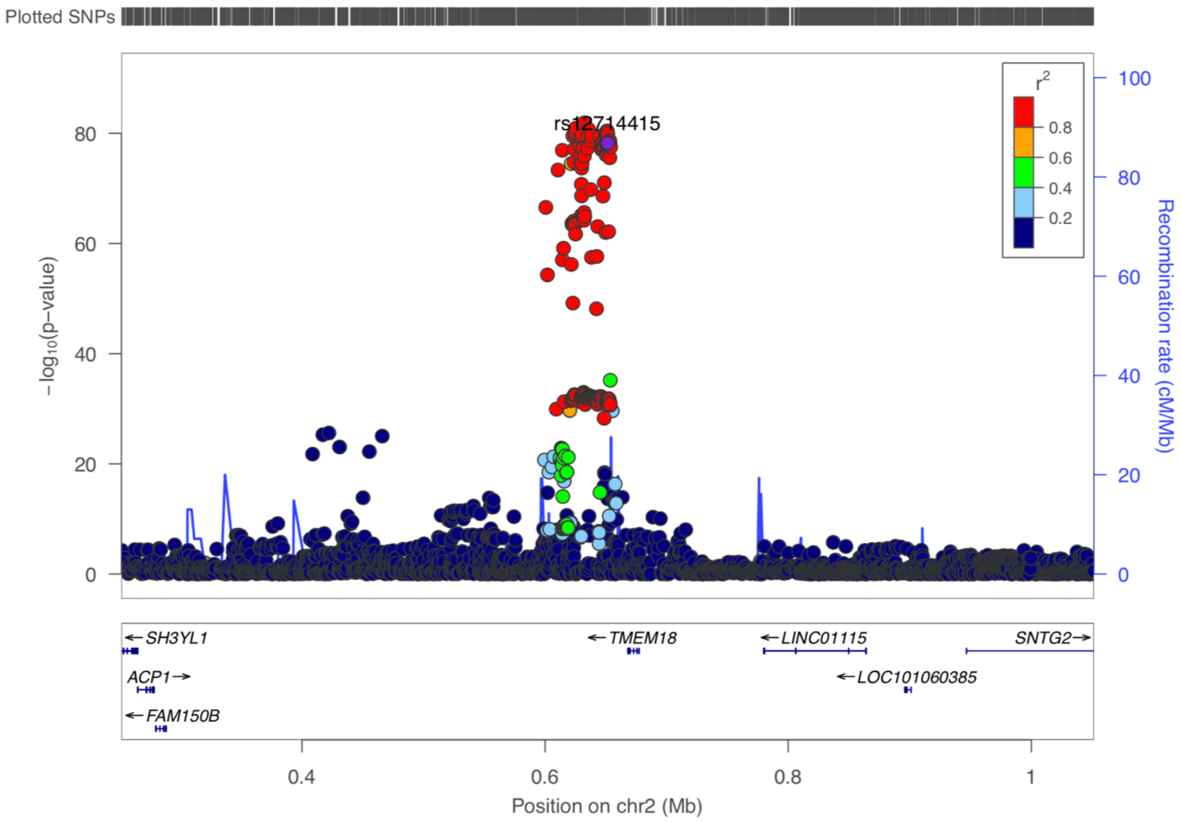


Figure 5.3: Regional association plot for the *TMEM18* locus from the GIANT-UKBB BMI meta-analysis. The highlighted SNP (rs12714415) represents the most significant GWAS SNP in the locus as identified by EWAS-FUSION, which is in strong LD with the top GIANT-UKBB GWAS in the locus.

5.3.4 Re-analysis of genetic risk score for BMI shows substantial reduction in correlation

In their paper on the association between methylation and BMI, Wahl *et al*. constructed a genetic risk score for BMI and used this in MR analyses to predict the effect of BMI on observed levels of methylation in their cohort. They observed a strong correlation between their BMI risk score and methylation (reported r_g =0.81, **Figure 5.4 b**). This association formed one of the major points of evidence used to substantiate their claim that BMI was the primary causal exposure for methylation and not the other way around. However, this analysis was based on loci with known associations with BMI, and as methylation levels were not aligned to express positive values (as is best practice when conducting such analyses) it is likely that this effect is artificially inflated. This analysis was therefore repeated using the same data, with all GRS-predicted effects on methylation expressed as positive values, which removes the artificial gap at the centre. Whereas the originally reported correlation was strongly positive (R^2=0.65), the effect of aligning the GRS-predicted effects substantially reduces the strength of this correlation (R^2=0.25) (**Figure 5.5**). This suggests that the initial report of the impact of BMI on DNA methylation was substantially overestimated, and further weakens the case for an overarching causal chain in this direction.

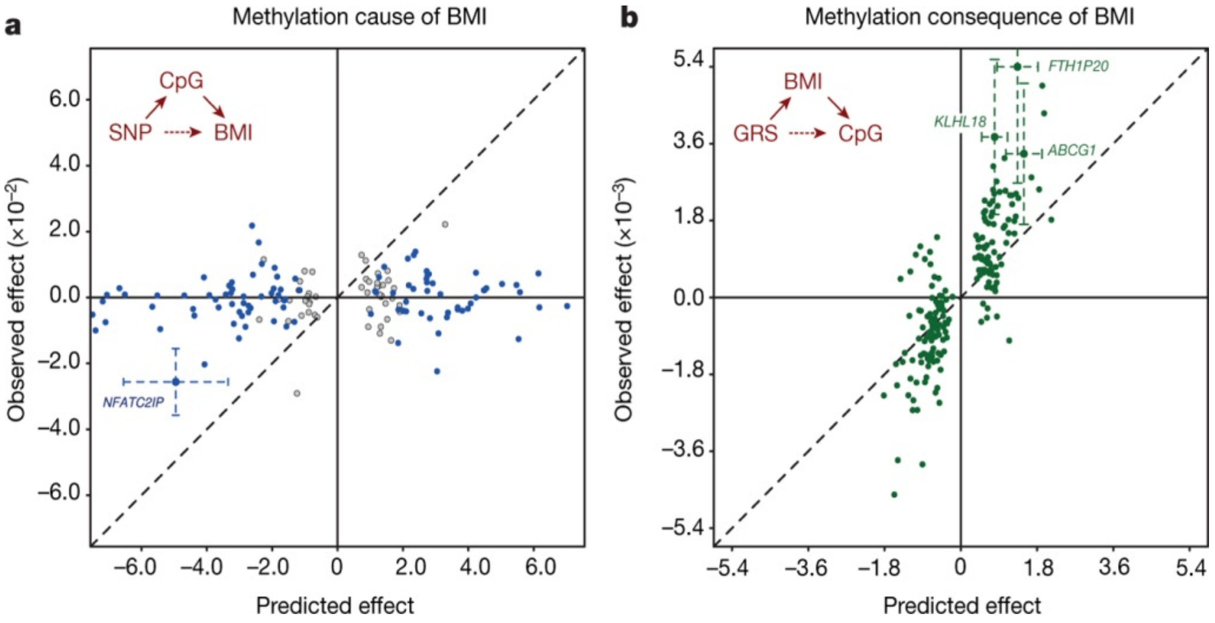
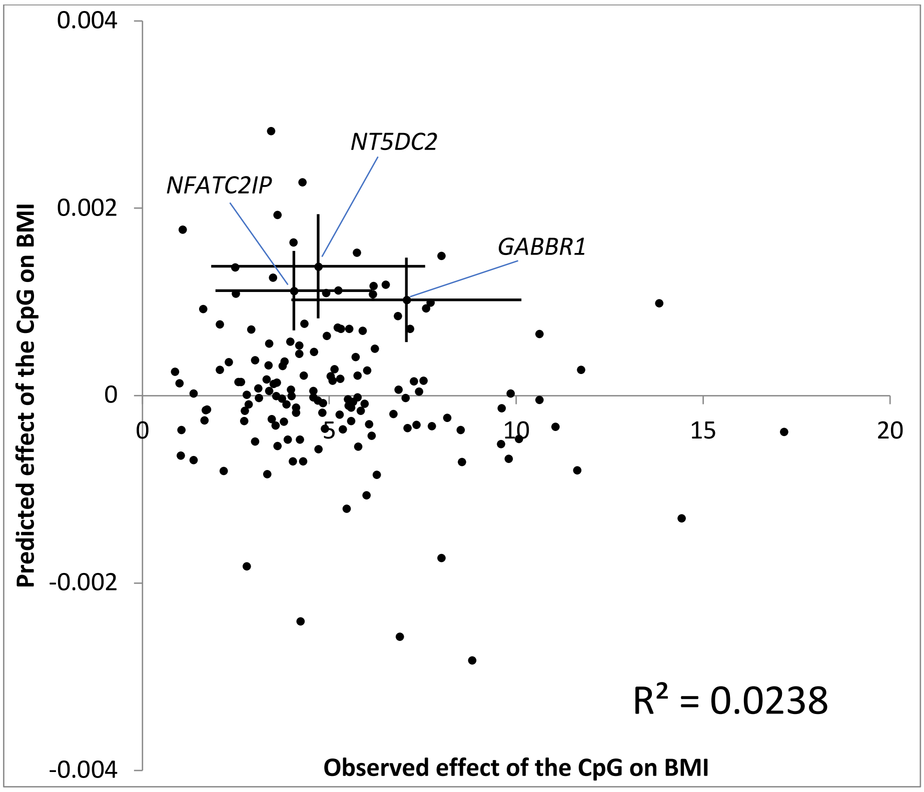


Figure 5.4: Figure reproduced from Wahl *et al.*[221] showing correlation between single-SNP effects on methylation (a) and methylation predicted by BMI-GRS compared to observed effects (b). In both cases, the effect sizes have not been aligned to the methylation-increasing direction, creating an artificial gap about the origin and which has the effect of biasing correlation estimates upwards.



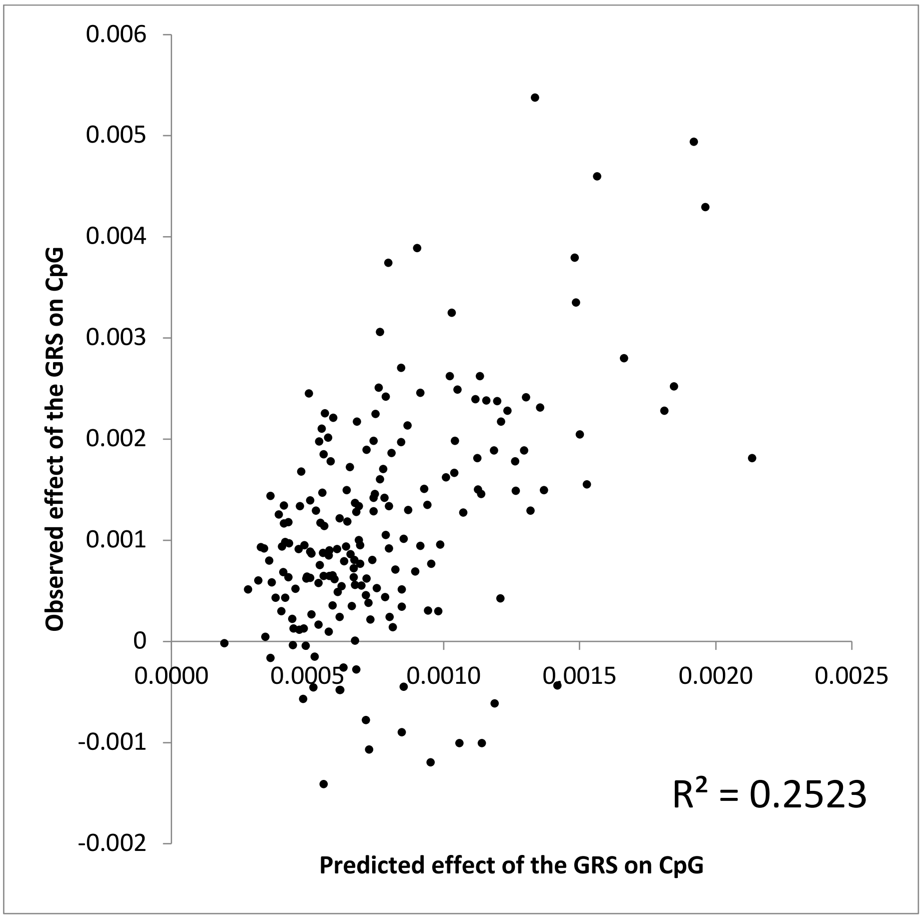


Figure 5.5 Re-analysis of the data from Wahl *et al.*[221] with effects aligned to methylation increasing alleles. Removal of the artificial gap leads to substantial reduction in estimated R² values.

5.3.5 EWAS-FUSION links methylation at multiple loci with puberty timing

Implementation of EWAS-FUSION using male puberty summary statistics identified 168 MVPs associated with puberty timing, of which 38 were significant in the conditional (joint) model (**Table 5.3**). The majority of these (29 of the 38, 76%) were located within 1 MB of a previously reported puberty variant. The remaining 9 loci were not significantly associated with puberty in the largest GWAS for either AAM or voice breaking, indicating that these loci are associated with puberty only through epigenetic mechanisms.

5.3.6 Influence of BMI on puberty-associated MVPs

To assess the extent to which puberty-associated MVPs were mediated through BMI, lookups of the 38 identified MVPs were performed in the BMI EWAS-FUSION results. Fourteen of the MVPs showed at least nominal associations with BMI in EWAS-FUSION (40%; 3 MVPs were not imputed in the BMI analysis), with four showing significant effects after Bonferroni correction (**Table 5.3**). For the majority of MVPs with at least nominal evidence for association with BMI (i.e. those with P<0.05), methylation at these loci showed the expected relationship between BMI and puberty timing (i.e. conferring higher BMI and earlier puberty, or vice versa). However for 4 loci the effects were opposite to expected associations. cg16228356 and cg118826563 were negatively associated with both puberty and BMI, with increased methylation causing earlier puberty onset but lower BMI. Conversely cg17623882 and cg14378231 were each positively associated with puberty and BMI, with increased methylation at these loci associated with delayed puberty and higher BMI, opposite to the general effect. These variants are likely to influence puberty timing and BMI via different pathways, and further investigation is warranted to clarify these effects.

5.3.7 Puberty-MVPs in imprinted regions

The two previously reported AAM loci which showed parent-of-origin effects, at *DLK1* and *MKRN3*, were not among the 38 puberty-associated MVPs (i.e. not within 1 MB). To further investigate whether imprinted regions are implicated amongst puberty MVPs, we looked up the 38 loci to see if any were proximal to one of the 42 imprinted regions identified in GTEx data as reported by Baran *et al*[231]. None of the puberty MVPs were among these known imprinted regions based on annotation of nearest gene. However two MVPs, cg03473532 in/near *MKLN1* and cg27513965 in/near *TFAP4*, were both within 1MB of known imprinted SNPs at *MEST* and *ZNF597*, respectively. This leaves open the possibility for an effect of imprinting at these loci, though maternal and paternal-specific models of inheritance will be needed to confirm this.

Table 5.3 MVPs associated with male puberty timing from EWAS-FUSION

			FUSION (Voice breaking)		FUSION (BMI)	
MVP	Chr: pos	Nearest Gene	Beta (S.E.)	P-value	Z-score	P-value
cg01176694	16: 14,289,232	-	14.2 (1.0)	1.4×10^{-45}	-0.69	0.49
cg03473532	7: 130,659,283	*MKLN1*	-10.4 (1.0)	1.7×10^{-25}	-1.02	0.31
cg16786949	3: 182,994,098	-	10.4 (1.0)	4.7×10^{-25}	-0.33	0.74
cg02851062	11: 122,354,193	*BSX*	-10.8 (1.0)	6.0×10^{-25}	1.72	0.09
cg16228356	17: 41,204,727	-	-10.1 (1.0)	1.2×10^{-23}	-1.93	0.05
cg27631724	1: 10,962,954	*C1orf127*	9.8 (1.0)	1.5×10^{-22}	0.67	0.50
cg04744409	6: 105,494,884	-	-24.5 (2.6)	4.1×10^{-21}	2.79	5.2×10^{-3}
cg25161029	11: 122,351,182	-	-8.3 (1.0)	2.3×10^{-15}	-1.12	0.26
cg07114886	3: 51,720,986	*GRM2*	-7.7 (1.0)	2.3×10^{-14}	4.13	3.6×10^{-5}
cg25657713	6: 100,145,429	-	7.8 (1.0)	1.3×10^{-13}	-1.37	0.17
cg19984742	20: 54,257,897	*MC3R*	7.2 (1.0)	7.5×10^{-13}	-0.76	0.45
cg27554954	15: 58,478,887	*ANXA2*	6.9 (1.0)	6.5×10^{-12}	0.02	0.98
cg14184400	3: 49,435,061	*AMT*	-6.9 (1.0)	8.5×10^{-12}	0.90	0.37
cg17623882	6: 41,881,589	*USP49*	6.7 (1.0)	2.1×10^{-11}	2.65	8.0×10^{-3}
cg19226099	20: 54,257,492	*MC3R*	6.6 (1.0)	4.2×10^{-11}	0.64	0.52
cg04106006	11: 27,699,030	*BDNF*	-6.4 (1.0)	1.2×10^{-10}	N/A	N/A
cg11882563	14: 35,668,029	-	-6.4 (1.0)	2.1×10^{-10}	-2.38	0.02
cg22156842	3: 138,019,859	*TMEM22*	6.2 (1.0)	4.7×10^{-10}	-4.82	1.4×10^{-6}
cg08976101	11: 44,284,999	*ALX4*	6.2 (1.0)	6.1×10^{-10}	-0.86	0.39
cg11618529	19: 63,643,920	*ZNF132*	-5.9 (1.0)	2.9×10^{-9}	-0.90	0.37
cg12537728	11: 13,266,172	*ARNTL*	5.9 (1.0)	3.3×10^{-9}	-7.33	2.4×10^{-13}
cg22456251	15: 86,956,077	*MIR7-2*	5.9 (1.0)	3.7×10^{-9}	-0.17	0.87
cg14515364	2: 626,606	-	-5.8 (1.0)	7.3×10^{-9}	16.20	2.6×10^{-59}
cg00741624	14: 92,967,142	*KIAA1409*	5.7 (1.0)	1.2×10^{-8}	-5.46	4.7×10^{-8}
cg27513965	16: 4,261,434	*TFAP4*	-5.6 (1.0)	2.2×10^{-8}	0.52	0.61
cg15129506	2: 27,812,216	-	5.6 (1.0)	2.5×10^{-8}	-4.14	3.5×10^{-5}
cg11777420	2: 120,486,403	*EPB41L5*	-5.4 (1.0)	6.8×10^{-8}	N/A	N/A
cg09953425	20: 32,914,810	*GGT7*	-5.3 (1.0)	9.8×10^{-8}	2.52	0.01
cg03036210	16: 88,431,592	*SPIRE2*	5.8 (1.1)	1.5×10^{-7}	1.05	0.29
cg10527021	4: 104,863,920	-	-5.2 (1.0)	2.5×10^{-7}	0.64	0.52
cg03726525	16: 68,764,798	*CLEC18C*	5.1 (1.0)	2.8×10^{-7}	-6.98	2.9×10^{-12}
cg18074954	18: 43,041,490	-	-5.1 (1.0)	3.0×10^{-7}	1.29	0.20
cg25281562	12: 119,938,655	*C12orf43*	-5.1 (1.0)	3.5×10^{-7}	-1.81	0.07
cg12277524	1: 66,051,507	*PDE4B*	5.5 (1.1)	3.6×10^{-7}	N/A	N/A
cg03461559	17: 7,259,644	*NLGN2*	-5.1 (1.0)	4.0×10^{-7}	-1.83	0.07
cg14378231	1: 98,176,620	-	5.1 (1.0)	4.0×10^{-7}	2.34	0.02
cg10483660	13: 111,039,078	-	5.1 (1.0)	4.3×10^{-7}	-4.46	8.2×10^{-6}
cg02316596	11: 119,940,189	-	-5.0 (1.0)	5.9×10^{-7}	-1.38	0.17

5.4 DISCUSSION

Epigenetic factors have an important role in growth and development, and it has been suggested that they may provide an explanation for the association between adverse early life exposures and adult disease. While epigenetic modifications have consistently been associated with health outcomes, important questions regarding the mechanisms underlying these associations remain. In this chapter I have described a study which combines publicly available data and a new method for conducting EWAS to present evidence for a casual effect of DNA methylation on two important developmental traits, BMI and puberty timing.

5.4.1 Evidence bi-directional association between methylation and BMI

In the largest systematic assessment of causality between methylation and BMI to date, Wahl *et al*. performed bi-directional MR analyses on 187 MVPs identified in an EWAS for BMI and concluded that BMI influenced methylation while the reverse pathway was less likely. This conclusion was based on two primary lines of evidence: 1) a GRS for genetically predicted methylation was strongly correlated with BMI; and 2) single-SNP MRs identified only one MVP (at *NFAT2CIP*) which showed evidence for a causal effect on BMI, compared to three loci (*KLHL18*, *ABCG1* and *FTH1P20*) showing casual associations for BMI on methylation. With respect the first line of evidence, the strong correlation reported was likely an overestimate due to the artificial gap created on the plot as a result of not aligning the GRS to have the same direction of effect for all MVPs. This was confirmed here by repeating the analysis with all genetically predicted GRS effects on methylation aligned to positive values, resulting in a marked decrease in the R^2 value. Addressing the second point, we repeated the MR analyses using an independent and publicly available data source (BIOS) to identify *cis*-meQTLs associated with the BMI-MVPs. The reported association with *NFAT2CIP* was confirmed, in addition to identifying two further loci in/near *NT5DC2* and *GABBR1* which also show evidence for a causal effect of methylation on BMI. Unfortunately, performing the reverse MRs was not possible as the summary data for methylation was not available; therefore we are unable to provide a comparative estimate for the number of loci at which BMI causally influences methylation. However because of the preponderance of evidence to date which has suggested an overarching effect of BMI influencing methylation, these analyses certainly call into question these conclusions by presenting evidence for a bi-directional effect.

5.4.2 EWAS-FUSION identifies multiple loci with causal links to BMI

In the next stage of analysis, we greatly expanded the number of MVPs associated with BMI by predicting methylation levels using genetic variants which were analysed using EWAS-FUSION. This identified an additional 68 MVPs which were independently and causally associated with BMI. The broadly consistent directions of effect between this method and the single-SNP MRs provide reassurance for the validity of the results obtained from this model. Taken together with the results from the single-SNP MRs, these results reinforce the notion of bi-directionality in the relationship between BMI and methylation.

5.4.3 Genome-wide evidence for an effect of methylation on puberty timing

Previous studies have implicated epigenetic silencing in the control of puberty in females[225,232,233], while animal studies have demonstrated a link between methylation levels in pituitary-expressed genes and the initiation of puberty timing[234]. However, these studies have relied on assessing methylation in specific candidate genes, and large-scale genome-wide assessment studies assessing causality have not been conducted. The identification of 38 MVPs associated with puberty timing using EWAS-FUSION therefore provides valuable information on the role that DNA methylation has on puberty timing. Among these 38 MVPs are 9 loci which have not previously been associated with puberty timing in GWAS. Along with the novel associations found in the BMI EWAS-FUSION analysis, this highlights the utility of performing EWAS as a complementary approach to GWAS, extending the search space within the genome to identify variation which may act through different pathways to affect the same phenotype.

5.4.4 Overlap between puberty and BMI MVPs

A fundamental question arises as to the extent to which differential methylation acts through BMI in its influence of puberty timing. Unfortunately, the EWAS-FUSION pipeline does not allow for possibility of performing mediation analyses to assess this. However, the enrichment of puberty-related MVPs for associations with BMI suggests that this likely plays a significant role. Interestingly, He *at al*. highlighted methylation at the *SIM1* gene as being significantly associated with obesity in adolescents[235]. As discussed in Chapter 3, *SIM1* is notable among puberty loci as having discordant directions of effect in males and females. Further investigations involving sex-specific methylation may therefore shed light on this observation.

5.4.5 Strengths and Limitations

The primary strength of this study was the use of publicly available data, which provided large sample sizes for discovery of meQTL variants and in the BMI EWAS. Where the discovery sample for EWAS by Wahl *et al*. was comparatively limited (N=5,387), the use of EWAS-FUSION allowed for incorporation of GWAS summary statistics from the much larger GIANT and UK Biobank studies (maximum N=142,630), providing a distinct advantage in the power for discovery of MVPs casually associated with BMI. Furthermore, the re-analysis of existing data offered a significant improvement on previous work by corrected for methodological flaws which may have altered the conclusions drawn.

As in many epigenetic studies, methylation was measured in whole blood. While this is often used as it is relatively non-invasive and convenient to measure, there is considerable evidence showing that methylation is tissue specific[236–238]. Therefore, the effects of variable levels of methylation in tissues relevant to the traits considered (e.g. adipose and reproductive organs) is not necessarily captured with this study design. It is hoped that improvements to the variety of tissues for which high quality methylation data is available will support future work. Additionally, it was decided not to include *trans*-meQTLs, as these are more difficult to detect due to the propensity to produce false positives because of sequence similarity[239]. It may be that some or all of the non-GWAS associations are acting as *trans*-meQTLs. Finally, since this analysis was conducted an updated methylation microarray capturing ~850,000 CpG sites has become available[240]. Among other developments, this new array is enriched for enhancer regions, which can influence transcription levels despite not being located proximal to transcription start sites. Future analyses using this array may yield new insights into the role of methylation of enhancer regions on BMI and puberty timing.

5.4.5 Conclusion

The influence of epigenetic modifications on complex traits is an active area of research and debate. In the case of BMI, the potential of DNA methylation to causally impact this trait has largely been discounted in the literature, with explanations favouring causality in the opposite direction. The results presented in this chapter therefore represent an important contribution to this debate, with the identification of multiple genomic loci where variation in methylation causally impacts BMI. Taking this into account, it seems much more plausible that these traits have a complex dynamic in which each influences the other, which will have important and long-lasting implications for health. In the context of early life determinants of health, the observation that DNA methylation also influences puberty timing is of interest and suggests that methylation resulting from adverse conditions in early life may impact health through multiple pathways, both mediated by and independent of adiposity (**Figure 5.6**).

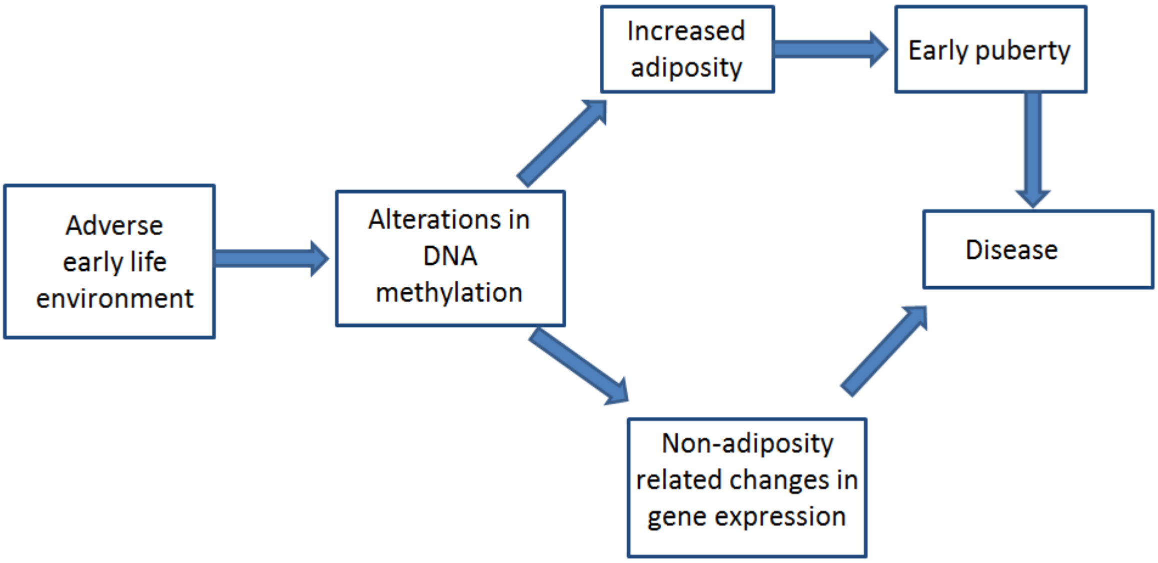

Adverse early life environment
Alterations in DNA methylation
Increased adiposity
Early puberty
Disease
Non-adiposity related changes in gene expression

Figure 5.6: Schematic representation of methylation-mediated pathways from early life exposures to disease.

Chapter 6: REPROWAS: A genetic atlas of reproductive health in the UK Biobank

Contributions and collaborations

This study was conceived in collaboration with colleagues at the MRC Epidemiology Unit, including Dr. John Perry (JP), Professor Ken Ong (KO), Dr. Alexander Busch (AB) and Dr. Felix Day (FD). The identification of UK Biobank reproductive phenotypes was performed by myself and AB. Curation of the phenotypes to produce the reproductive traits for GWAS was conducted by AB with input from KO. Data cleaning was performed by myself and AB; JP and FD implemented the GWAS models for all traits and performed the clumping of variants. All subsequent analyses were performed by me, while FD produced the figure for the associations at the *ESR1* locus.

SUMMARY

Reproductive health has important consequences at both the individual and societal level, and has been shown to be influenced by developmental exposures. While the heritable determinants of many common reproductive traits have been studied, for many others little is known regarding their genetic aetiology. Among the traits which have been studied, many of the identified variants have been shown to influence multiple reproductive conditions, indicating a large degree of shared genetic architecture underlying reproductive health. Investigating the links between various traits can have important benefits, particularly for drug discovery where medications can be developed which more specifically target the outcome of interest while reducing the possibility of any unwanted secondary effects. Phenome-wide association studies (PheWAS) have proven to be a useful tool for translational genetics. However adapting the methods which were developed for smaller studies and a limited numbers of traits to large-scale studies, with potentially thousands of phenotypes, has proven challenging. In this chapter I describe a study which aims to mitigate some of the common pitfalls of PheWAS, by conducting genetic discovery for traits related to reproduction which are available in the UK Biobank study. I conduct GWAS for 181 reproductive traits, the majority of which have never been studied using genome-wide genetic methods, and uncover novel associations that demonstrate the degree of genetic overlap underlying reproductive health at an unprecedented scale. This resource, called REPROWAS, can be used in future studies to gain insights into important questions related to the determinants of reproductive health and disease.

6.1 BACKGROUND

Aspects of reproductive health and fertility vary considerably in the population and have important implications for clinical outcomes and individual wellbeing. Thus far in this thesis, evidence has been presented as to how early life exposures can have a significant and lasting impact on health traits. This extends to reproductive health as well which, as discussed in Chapter 1, can have considerable physical, emotional and societal impacts. In the next two chapters of this thesis, I investigate the influence of early life exposures as they relate specifically to reproductive health outcomes.

6.1.1 Genetic determinants of reproductive health traits

For many of the reproductive traits which have been studied in-depth the aetiology appears to be complex, being influenced by a combination of environmental, social and genetic factors. The environmental determinants of reproductive health have been increasingly well characterised in recent years, with factors such as endocrine disruptors which mimic or inhibit hormone function receiving much attention[241,242]. With regard to the genetic determinants of reproductive health, the widespread use of GWAS over the last decade and more has increased the understanding of the role played by genetic variation for many reproductive traits. Consortium studies such as the Breast Cancer Association Consortium (BCAC), PRACTICAL and ReproGen have identified hundreds of variants associated with breast cancer[243], prostate cancer[143,244] and the timing of reproductive events[35,245], respectively. Large-scale genetic discovery has also been conducted for select other reproductive traits, identifying variants associated with PCOS[41], endometriosis[40,246,247], and benign prostate hyperplasia (BPH)[248,249] among other common conditions. Such studies have revealed complex polygenic architecture underlying these traits, as well as providing evidence for a substantial amount of overlap of the genetic determinants influencing different reproductive outcomes. Moreover, casual inferences have made use of the identified variants, demonstrating that reproductive traits can influence numerous health and social outcomes[250–253].

Despite this progress, genetic discovery for many other less common reproductive traits and diseases has not been conducted. Identification of genetic determinants from a broader range of phenotypes is necessary to gain a more complete understanding of the role played by genetic variation on overall reproductive health.

6.1.2 Phenome-wide association studies

As the field of statistical genetics progresses and GWAS summary statistics from ever larger studies and more traits become available, this presents the opportunity to leverage this data to gain a better understanding of complex trait genetics. In recent years, phenome-wide associations studies (PheWAS) have increasingly been employed to identify genomic loci which influence multiple traits[254–256]. Whereas GWAS examine the genome-wide effect of variants on a single trait of interest, PheWAS operates in reverse by examining the effect of a single or a small selection of genetic variants on a large number of traits. The PheWAS framework can be used to inform genetic casual inference methods[257], allowing pleiotropic genetic effects to be distinguished from other potential mechanisms for associated comorbidities. PheWAS can also be used to define subsets of diseases, which has important implications for precision medicine, particularly with regard to identifying therapeutic pharmaceutical targets[258,259].

Linkage of routine healthcare data to large cohort studies with genotypic data provides a valuable resource for identifying genetic variants with pleiotropic effects[260]. This is perhaps best exemplified by the UK Biobank study, which provides publicly available genotypic data on it's over 500,000 participants, along with extensive phenotypic data on health, socioeconomic and lifestyle traits[31]. Since the release of the full cohort's genotypic data in 2017, several studies have developed PheWAS methods to rapidly perform GWAS on the over 2,000 traits available in this dataset[261,262]. While certainly efficient, this approach highlights two of the main limitations of the PheWAS method as usually applied. Firstly, analysing millions of genetic variants over thousands of traits greatly increases the possibility of false positive associations, which must be accounted for by applying a highly restrictive correction for the multiple tests being run. Secondly, these rapid and automated approaches often maximise efficiency at the expense of fully cleaned and carefully considered phenotypes, and may not fully account for potential sources of bias and confounding. As such, caution must be taken when interpreting the results of such analyses, which can limit the ability to make meaningful inferences.

An alternative approach whereby only a subset of the available phenotypes are considered may therefore be advantageous. Limiting the number of traits, for example to only those related to reproductive health, can allow for a more nuanced approach involving careful curation of the phenotypes, mitigating the potential statistical and technical limitations of the PheWAS method.

6.1.3 Areas of opportunity

The UK Biobank study represents a valuable resource for application of a PheWAS framework, however many of the approaches employed to date have not been able to fully exploit this potential. In this chapter I describe a study which attempts to address some of the key limitations of PheWAS methods which have previously been employed, in order to enhance our understanding of the genetic determinants of reproductive health. The aims of this study are threefold:

1) To perform GWAS on expertly curated reproductive phenotypes in the UK Biobank study, in order to identify genetic determinants underlying reproductive health traits;
2) Applying a PheWAS framework to the GWAS results to allow inferences to be made at multiple levels of genomic resolution; and
3) To create a genetic atlas of reproductive health, to be called "REPROWAS", for use in future studies (including subsequent chapters of this thesis) to answer important questions about reproductive health and disease.

This resource can be used to investigate the shared genetic architecture underlying reproductive health traits, yielding novel insights which could impact the way they are diagnosed, treated and prevented.

6.2 METHODS

6.2.1 Identification of reproductive variables in UK Biobank study

All variables in UK Biobank which are related to reproductive health were considered for inclusion in this study. Reproductive variables were identified by manual screening of all variables listed in the online UK Biobank Data Showcase (http://biobank.ndph.ox.ac.uk/showcase/). Based on initial inspection, it was decided to focus on three sources of data:

- variables collected from the online touchscreen questionnaires
- ICD-10 coded diagnoses from Hospital Episode Statistics (HES) record linkage
- Self-reported illnesses from face-to-face interviews

Touchscreen Questionnaire

Participants were asked to complete an online touchscreen questionnaire during their baseline assessment at the UK Biobank testing centres (see Chapter 2). In addition, a subsample of 20,339 participants living within 35km of the UK Biobank Co-ordinating Centre in Stockport, UK completed a repeat questionnaire assessment between 2012 and 2013, while another 13,159 participants have completed a repeat assessment during imaging visits from 2014 onward. For each participant, the most recent assessment for which data was available for a given trait was used. Information on sexual history and reproductive health was collected, including questions on sexual behaviours, fertility and pregnancy outcomes, age of reproductive milestones (e.g. puberty and menopause) and routine health screenings. Responses of “Do not know” or “Prefer not to answer” were treated as missing. The full list of relevant variables from the touchscreen questionnaire is shown in **Supplementary Table 6.1**.

ICD-10 coded diagnoses from linked health records

UK Biobank electronic health record linkage provided further data on reproductive health outcomes in the form of International Classification of Disease, Version 10 (ICD-10) coded diagnoses obtained from Hospital Episode Statistics (HES) records. ICD-10 uses a hierarchical system to classify diseases, grouping them into increasingly specific categories in up to four levels (see **Supplementary Figure 6.1** for an example). Conditions related to reproductive health were selected by manual screening of the entire ICD-10 dictionary at the lowest level of classification (i.e. the specific diagnosis code), from UK Biobank data fields 41202 (main diagnosis) and 41204 (secondary diagnoses).

Self-reported illnesses

UK Biobank participants completed a verbal interview with a trained nurse as part of the baseline assessment, which included questions on past and current medical conditions. The interviewer was aware of answers to illness history questions from the touchscreen questionnaire, and for any discrepancies the interviewer repeated the question to confirm diagnoses and was able to amend touchscreen responses. Conditions were coded by UK Biobank researchers using a hierarchical classification system, divided into cancer and non-cancer illnesses (data fields 20001 and 20002, respectively). Similar to ICD-10 coding, the system uses four levels of increasing specificity (**Supplementary Figure 6.2**). Conditions related to reproduction were selected by manual screening at the lowest level of classification (i.e. the most specific).

6.2.2 Generation of phenotypes

Reproductive variables were curated to produce the Reproductive PheWAS ("REPROWAS") database, as described below and depicted in **Figure 6.1**.

Questionnaire Data

For continuous variables from the touchscreen responses, individuals were excluded based on a number of criteria (summarised in **Box 6.1**). Two variables, "Answered sexual history questions" and "Had menopause", were excluded as these do not allow for meaningful biological inference in the context of this study. Individuals with reported values of more than five standard deviations from the median value for that trait were set to missing. For participants who had follow-up information available, if the reported values differed by more than 10% of the median value for the trait between assessments they were also excluded from the analysis. Additional exclusion criteria were applied to specific traits. A variable for ever having been pregnant was derived based on reported numbers of live births, spontaneous miscarriages, stillbirths and terminations. Women who had never been pregnant were then excluded from analyses for pregnancy-related outcomes. For age at first sexual intercourse, individuals reporting ages of less than 12 years were set to missing. Finally, for the relative age of voice breaking and first facial hair variables, individuals who changed answers between relatively early and relatively late between follow-ups were excluded.

Disease variables

ICD-10 primary and secondary diagnosis codes were combined with self-reported illness data to derive phenotypes which maximised sample size and statistical power for genetic discovery while preserving disease-specific aetiology. This was performed with the assistance of clinically trained researchers with expertise in reproductive physiology. The list of variables, including the constituent ICD-10 and self-reported illness codes for derived variables, is shown in **Supplementary Table 6.2**. All individuals with any of the corresponding ICD-10 codes or self-reported illness codes for that trait were considered cases, with all other individuals considered controls after application of exclusion criteria (**Box 6.1)**. For sex-specific conditions, only individuals of the appropriate sex were considered as controls for that analysis. As described above, women who had never been pregnant were set to missing for pregnancy-related outcomes.

Box 6.1: Exclusion criteria for reproductive traits

For all traits:

- Women who have never been pregnant from pregnancy-related traits
- Mismatch between reported sex and genotypic sex
- Non-white European ancestry
- Opposite sex for sex-specific conditions

For questionnaire traits only:

- Reported values varied by more than 10% of median value between follow-ups
- Reported values more/less than 5 standard deviations from median
- Changing answers from relatively early to relatively late (or vice versa) for male puberty timing variables

6.2.3 Genome-wide association analyses

Genotyping and imputation methods for UK Biobank are described in Chapter 2. For all traits, analyses were limited to unrelated individuals of white European ancestry. GWAS for all traits was performed using BOLT-LMM (v2.3)[108]. For disease traits, only those with at least 100 cases were considered. All autosomal bi-allelic variants and insertions/deletions (indels) were analysed, and all models were adjusted for genotyping chip, age at initial assessment and the first 10 genetically determined PCs. BOLT-LMM can produce inflated test statistics for rare variants when the case-control ratio (μ) is unbalanced (<10%), therefore separate MAF cut-offs were used for binary traits: INFO>0.3 and MAF>0.01 if $\mu>0.1$; INFO>0.3 and MAF>0.10 if $\mu<0.1$. For binary/categorical traits, log odds ratios and standard errors are obtained for binary and categorical variables using the formulae 1 and 2 as described in Chapter 2.

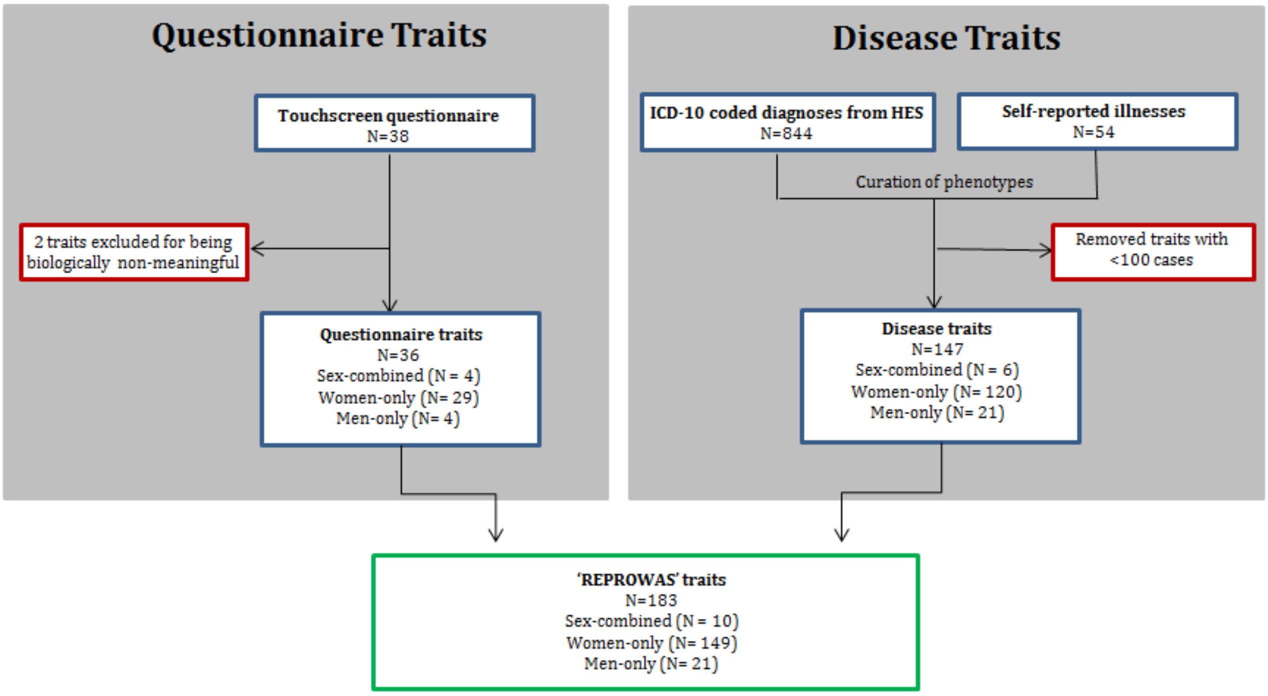


Figure 6.1: Flow chart of study design for reproductive PheWAS ("REPROWAS"). Reproductive phenotypes were identified via manual screening of the UK Biobank Data Showcase and expertly curated to maximise statistical power for genetic discovery while maintaining trait-specific aetiology.

6.3 RESULTS

6.3.1 Identification of variant-trait associations

After application of exclusion criteria, GWAS were run on a total of 181 traits (minimum N_{cases}=100) (**Table 6.1**). SNP-captured heritability estimates calculated using LD score regression ranged from 0 (for multiple traits) to 30.8% (for hair/balding pattern). Following distance-based clumping to identify lead variants within 1 MB windows, a total of 1,966 independent variant-trait associations were identified at a genome-wide level of significance ($P<5.0\times10^{-8}$). After correction for the number of tests, a total of 871 associations remained statistically significant ($P<2.7\times10^{-10}$, $=5.0\times10^{-8}/181$). Of the 181 traits, 132 (73%) had at least one associated variant at a genome-wide significance level and 58 (32%) had at least one associated variant after application of multiple testing correction. For many traits this represented the first identified genetic determinant.

Several of the novel associated variants were either non-synonymous mutations or were in strong LD with a non-synonymous variant ($r^2\geq0.8$, HaploReg v4.1). A rare variant (MAF=0.02) associated with acquired atrophy of the Fallopian tubes (rs16938754/G; OR=28.22, $P=3.7\times10^{-12}$), was in strong LD (r^2=0.86, D'=0.93) with a missense variant in *TMEM70*, a mitochondrial gene which encodes a protein involved in the oxidative phosphorylation process and has been linked to growth retardation and genital malformations in men[263]. Another missense variant (rs10929757/C) in the oestrogen-responsive *GREB1* gene, which has previously been associated with endometriosis[264], was associated with both leiomyoma (fibroids) of the uterus (OR=1.08, $P=8.1\times10^{-11}$) and risk of bilateral oophorectomy (removal of ovaries; OR=1.07 , $P=5.1\times10^{-10}$) at a genome-wide significant level. Additionally rs113008088/TGTC, a rare indel (MAF=0.01) which is associated with "absent, scanty or rare menstruation" (OR=97.47, $P=1.3\times10^{-10}$), is in high LD (r^2=0.81, D'=0.93) with a missense variant in the gene encoding the nebulin protein, a component of sarcomeres in skeletal muscle. Mutations in this gene have been observed with high frequency in Ashkenazi Jewish populations (1 in 108 individuals are carriers)[265], and it is possible that the high odds ratio observed here may reflect fine-level population stratification which is unaccounted for by BOLT-LMM. Finally, two missense variants in genes associated with the adaptive immune response, *BTAN3A2* and *HLA-B*, were also implicated in risk for prostate hyperplasia (rs9348716/A, OR=1.15, $P=8.6\times10^{-16}$) and phimosis (tight foreskin; rs1050529/T; OR=1.24, $P=2.1\times10^{-11}$), respectively.

It should be noted that while care was taken to account for the unbalanced case-control ratios which exist for many conditions, it is possible that BOLT is still not adequately calibrated to account for rare variants at these extremes. Therefore, the very high odds ratios observed for some conditions should be interpreted cautiously. Recently, novel methods (for example, SAIGE[266]) have been developed which address case-control imbalance in large genetic studies. Though beyond the scope of this analysis, we recommend comparing estimates from multiple methods if such variants are of interest to researchers.

Table 6.1: Genetic associations with REPROWAS traits

Trait	N (Cases/Controls)	Known signals†	REPROWAS signals $P<5.0\times10^{-8}$ $(P<2.7\times10^{-11})$	Novel signals‡ $P<5.0\times10^{-8}$ $(P<2.7\times10^{-11})$	h^2 (%)§
Continuous Traits					
Age when periods started (menarche)	237,801	534	284 (179)	27 (6)	25.16
Age first had sexual intercourse	392,245	32	209 (84)	181 (63)	12.86
Age at menopause (last menstrual cycle)	142,883	190	97 (61)	19 (1)	8.41
Relative age of first facial hair	200,391	-	130 (60)	130 (60)	11.02
Birth weight of first child	194,476	505	73 (38)	7 (0)	8.88
Relative age voice broke	191,896	9	37 (16)	32 (12)	5.49
Age at first live birth	166,894	19	48 (13)	42 (10)	11.01
Years since last cervical smear test	189,529	-	10 (4)	10 (4)	2.97
Number of live births	244,959	3	12 (3)	9 (0)	6.50
Age started hormone-replacement therapy (HRT)	86,296	-	10 (3)	10 (3)	2.98
Lifetime number of sexual partners	367,437	118	17 (2)	2 (0)	3.78
Age at last live birth	166,685	0	7 (2)	7 (2)	6.04
Age at hysterectomy	455,34	-	5 (1)	5 (1)	1.72
Age started oral contraceptive pill	192,635	-	4 (1)	5 (1)	3.45
Length of menstrual cycle	41,774	7	4 (1)	0 (0)	1.02
Number of children fathered	204,950	-	1 (1)	1 (1)	3.57
Lifetime number of same-sex sexual partners	11,723	-	9 (0)	9 (0)	0.34
Age when last used oral contraceptive pill	175,786	-	1 (0)	1 (0)	2.26
Years since last breast cancer screening / mammogram	148,247	-	0 (0)	0 (0)	0.33
Time since last menstrual period	48,596	-	0 (0)	0 (0)	0.27
Number of spontaneous miscarriages	76,281	-	0 (0)	0 (0)	0.63
Number of pregnancy terminations	76,212	-	0 (0)	0 (0)	1.93
Age of primiparous women at birth of child	32,305	19	0 (0)	0 (0)	1.14

Table 6.1 (Continued)

Trait	N (Cases/Controls)	Known signals†	REPROWAS signals $P<5.0\times10^{-8}$ ($P<2.7\times10^{-11}$)	Novel signals‡ $P<5.0\times10^{-8}$ ($P<2.7\times10^{-11}$)	h^2 (%)§
Age at bilateral oophorectomy (both ovaries removed)	16,567	-	0 (0)	0 (0)	0.57
Number of stillbirths	76,171	1	0 (0)	(0/0)	0.16
Categorical/Binary Traits					
Hair/balding pattern	205,469*	1025	383 (268)	63 (19)	30.82
Malignant neoplasm of prostate	206,951 (7,169/199,782)	392	35 (22)	2 (1)	2.62
Hyperplasia of prostate	206,951 (16,479/190,472)	39	24 (18)	6 (2)	2.53
Leiomyoma of uterus (fibroids)	245,349 (16,619/228,730)	116	23 (15)	3 (0)	2.49
Malignant neoplasm of breast	245,349 (10,992/234,357)	670	18 (13)	0 (0)	2.21
Ever used hormone-replacement therapy (HRT)	244,329 (96,341/147,988)	-	11 (6)	11 (6)	4.75
Utero-vaginal prolapse	245,349 (7,931/237,418)	0	8 (4)	8 (4)	2.08
Malignant neoplasm of testis	206,951 (859/206,092)	34	6 (4)	1 (0)	0.86
Fistulae involving female genital tract	245,349 (201/245,148)	-	11 (3)	11 (3)	0.00
Polyhydramnios (excessive amniotic fluid)	208,121 (118/208,003)	-	9 (3)	9 (3)	0.04
Ever had hysterectomy (womb removed)	216,917 (18,447/198,478)	-	6 (3)	6 (3)	2.40
Endometriosis	245,349 (7,239/238,110)	38	4 (3)	0 (0)	1.29
Hydrocele and spermatocele	206,951 (1,480/205,471)	-	4 (2)	4 (3)	0.81
Absent, scanty or rare menstruation	245,349 (120/245,229)	-	13 (2)	13 (2)	0.13
Acquired atrophy of ovary and Fallopian tube	245,349 (129/245,220)	-	12 (2)	12 (2)	0.00
Bilateral oophorectomy (both ovaries removed)	241,930 (19,974/221,956)	-	8 (2)	8 (2)	2.08
Testicular hypofunction and male infertility	206,951 (197/206,754)	0	6 (2)	6 (2)	0.00
Polyp of corpus uteri	245,349 (9,691/235,658)	-	4 (2)	4 (2)	0.66
Carcinoma in situ of breast	245,349 (2,349/243,000)	-	4 (2)	4 (2)	0.31
Female urethrocele and cystocele	245,349 (5,339/240,010)	-	3 (2)	3 (2)	1.02
Redundant prepuce, phimosis and paraphimosis (foreskin disorders)	206,951 (2,290/204,661)	-	2 (2)	2 (2)	0.67
Non-inflammatory disorder of ovary, Fallopian tube and broad ligament, unspecified	245,349 (129/245,220)	-	14 (1)	14 (1)	0.08
Malignant neoplasm of cervix uteri	245,349 (279/245,070)	18	11 (1)	11 (1)	0.00
Single delivery by forceps and vacuum extractor	208,121 (176/207,945)	-	11 (1)	11 (1)	0.00
Oligohydramnios (deficient amniotic fluid)	208,121 (154/207,967)	-	11 (1)	11 (1)	0.00

Table 6.1 (Continued)

Trait	N (Cases/Controls)	Known signals†	REPROWAS signals $P<5.0\times10^{-8}$ ($P<2.7\times10^{-11}$)	Novel signals‡ $P<5.0\times10^{-8}$ ($P<2.7\times10^{-11}$)	h^2 (%)§
Intrapartum haemorrhage	208,121 (113/208,008)	-	11 (1)	11 (1)	0.28
Stricture and atresia of vagina	245,349 (193/245,156)	-	10 (1)	10 (1)	0.11
Other inflammatory disorders of penis	206,951 (135/206,816)	-	10 (1)	10 (1)	0.00
Complications of puerperium, not classified elsewhere	208,121 (215/207,906)	-	9 (1)	9 (1)	0.14
Female genital prolapse, unspecified	245,349 (283/245,066)	0	8 (1)	8 (1)	0.21
Benign neoplasm of male genital organs	206,951 (166/206,785)	-	7 (1)	7 (1)	0.04
Diseases of the digestive system complicating pregnancy, childbirth and the puerperium	208,121 (130/207,991)	-	7 (1)	7 (1)	0.00
Maternal care for known or suspected foetal abnormality and damage	415,968 (207,847/208,121)	-	6 (1)	6 (1)	0.16
Excessive and frequent menstruation with regular cycle	245,349 (9,477/235,872)	-	5 (1)	5 (1)	1.40
Carcinoma in situ of other and unspecified, prostate	206,951 (314/206,637)	-	5 (1)	5 (1)	0.02
Balanoposthitis (swollen penis)	206,951 (303/206,648)	-	5 (1)	5 (1)	0.00
Retained placenta and membranes, without haemorrhage	208,121 (246/207,875)	-	5 (1)	5 (1)	0.00
Vaginal and vulva dysplasia	245,349 (368/244,981)	-	4 (1)	4 (1)	0.09
Benign mammary dysplasia	245,349 (3,833/241,516)	-	2 (1)	2 (1)	0.75
Maternal care for other foetal problems	208,121 (1,615/206,506)	-	2 (1)	2 (1)	0.13
Enterocele and rectocele	245,349 (5,040/240,309)	-	1 (1)	1 (1)	1.17
Benign neoplasm of breast	245,349 (3,212/242,137)	-	1 (1)	1 (1)	0.49
Venous complications in the puerperium	208,121 (105/208,016)	-	19 (0)	19 (0)	0.00
Maternal care for tumour of corpus uteri	208,121 (108/208,013)	-	17 (0)	17 (0)	0.00
Malignant neoplasm and CIS of penis	206,951 (110/206,841)	-	17 (0)	17 (0)	0.02
Other disorders of amniotic fluid and membranes	208,121 (127/207,994)	-	13 (0)	13 (0)	0.00
Blighted ovum and non-hydatidiform mole	245,349 (119/245,230)	-	11 (0)	11 (0)	0.21
Congenital malformation of uterus and cervix	245,349 (213/245,136)	-	10 (0)	10 (0)	0.24
Gestational oedema and proteinuria	208,121 (167/207,954)	-	9 (0)	9 (0)	0.01
Neoplasm of uncertain or unknown behaviour, breast	245,349 (118/245,231)	-	8 (0)	8 (0)	0.00
Infections of genito-urinary tract in pregnancy	208,121 (253/207,868)	-	7 (0)	7 (0)	0.12

Table 6.1 (Continued)

Trait	N (Cases/Controls)	Known signals†	REPROWAS signals $P<5.0\times10^{-8}$ ($P<2.7\times10^{-11}$)	Novel signals‡ $P<5.0\times10^{-8}$ ($P<2.7\times10^{-11}$)	h^2 (%)§
Pyrexia during labour, not classified elsewhere	208,121 (144/207,977)	-	7 (0)	7 (0)	0.07
Puerperal infections	208,121 (251/207,870)	-	6 (0)	6 (0)	0.00
Other obstetric trauma	208,121 (230/207,891)	-	6 (0)	6 (0)	0.07
Failed induction of labour	208,121 (161/207,960)	-	6 (0)	6 (0)	0.07
Malignant neoplasm and CIS of vulva and vagina	245,349 (304/245,045)	-	5 (0)	5 (0)	0.00
Delayed delivery after spontaneous or unspecified rupture of membranes	208,121 (282/207,839)	-	5 (0)	5 (0)	0.41
Polyp of vagina and vulva	245,349 (240/245,109)	-	5 (0)	5 (0)	0.00
Maternal care for excessive foetal growth	208,121 (193/207,928)	-	5 (0)	5 (0)	0.00
Other non-inflammatory disorders of ovary, Fallopian tube and broad ligament	245,349 (561/244,788)	-	4 (0)	4 (0)	0.40
Pelvic inflammatory disease and peritonitis	245,349 (375/244,974)	-	4 (0)	4 (0)	0.00
Excessive vomiting in pregnancy	208,121 (240/207,881)	2	4 (0)	4 (0)	0.00
Impotence of organic origin	206,951 (303/206,648)	2	4 (0)	4 (0)	0.22
Sexual dysfunction (less desire)	452,300 (300/452,000)	0	4 (0)	4 (0)	0.20
Vagininits, vulvitis and vulvovaginitiis, infectious	245,349 (233/245,116)	-	4 (0)	4 (0)	0.00
Benign neoplasm of other and unspecified female genital organs	245,349 (234/245,115)	-	4 (0)	4 (0)	0.00
Congenital malformation of ovaries, Fallopian tubes and broad ligaments	245,349 (206/245,143)	-	4 (0)	4 (0)	0.20
Leukoplakia of penis	206,951 (149/206,802)	-	4 (0)	4 (0)	0.00
Other specified diseases of male genital organs	206,951 (2,335/204,616)	-	3 (0)	3 (0)	0.33
Placental disorders	208,121 (443/207,678)	-	3 (0)	3 (0)	0.13
Other benign neoplasms of uterus	245,349 (389/244,960)	-	3 (0)	3 (0)	0.02
Vaginal delivery following Caesarean section	208,121 (178/207,943)	-	3 (0)	3 (0)	0.14
Neoplasm of uncertain or unknown behaviour, ovary	145,349 (162/145,187)	-	3 (0)	3 (0)	0.00
Cystitis and other urinary tract infections	452,300 (17,565/434,735)	3	2 (0)	2 (0)	0.70
Other non-inflammatory disorders of uterus	245,349 (2,932/242,417)	-	2 (0)	2 (0)	0.11
Dysplasia of cervix uteri	245,349 (2,351/242,998)	-	2 (0)	2 (0)	0.09
Maternal care for other pregnancy conditions	208,121 (1,937/206,184)	-	2 (0)	2 (0)	0.03
Malignant neoplasm of corpus uteri	245,349 (1,407/243,942)	-	2 (0)	2 (0)	0.12

Table 6.1 (Continued)

Trait	N (Cases/Controls)	Known signals†	REPROWAS signals $P<5.0\times10^{-8}$ ($P<2.7\times10^{-11}$)	Novel signals‡ $P<5.0\times10^{-8}$ ($P<2.7\times10^{-11}$)	h^2 (%)§
Other disorders of prostate	206,951 (1,462/205,489)	-	2 (0)	2 (0)	0.08
Maternal care due to uterine scar from previous surgery	208,121 (1,305/206,816)	-	2 (0)	2 (0)	0.14
Other disorders of breast	245,349 (1,315/244,034)	-	2 (0)	2 (0)	0.28
Other conditions associated with female genital organs and menstrual cycle	245,349 (1,179/244,170)	-	2 (0)	2 (0)	0.00
Anaemia complicating pregnancy, childbirth and the puerperium	208,121 (943/207,178)	-	2 (0)	2 (0)	0.25
Ovulation bleeding	245,349 (1,035/244,314)	-	2 (0)	2 (0)	0.23
Carcinoma in situ of cervix uteri	245,349 (985/244,364)	-	2 (0)	2 (0)	0.22
Balanitis xerotica obliterans	206,951 (928/206,023)	-	2 (0)	2 (0)	0.28
Hypertrophy of breast	245,349 (697/244,652)	-	2 (0)	2 (0)	0.11
Labour and delivery complicated by umbilical cord problems	208,121 (686/207,435)	-	2 (0)	2 (0)	0.09
Single delivery by Caesarean section	208,121 (590/207,531)	-	2 (0)	2 (0)	0.00
Abnormalities of forces of labour	208,121 (484/207,637)	-	2 (0)	2 (0)	0.03
Undescended testicle and hypospadias	206,951 (416/206,535)	18	2 (0)	2 (0)	0.00
Acne	452,300 (399/451,901)	22	2 (0)	2 (0)	0.28
Other obstructed labour	208,121 (411/207,710)	-	2 (0)	2 (0)	0.72
Other complications of labour and delivery	208,121 (270/207,851)	-	2 (0)	2 (0)	0.66
Perineal laceration during delivery	208,121 (5,882/202,239)	-	1 (0)	1 (0)	0.39
Female infertility	245,349 (1,632/243,717)	-	1 (0)	1 (0)	0.00
Maternal care for known or suspected malpresentation of foetus	208,121 (1,172/206,949)	-	1 (0)	1 (0)	0.00
Inflammatory disease of cervix uteri	245,349 (1,274/244,075)	-	1 (0)	1 (0)	0.00
False labour	208,121 (1,152/206,969)	-	1 (0)	1 (0)	0.23
Erosion and ectropion of cervix uteri	245,349 (1,110/244,239)	-	1 (0)	1 (0)	0.23
Gestational hypertension and eclampsia	208,121 (789/207,332)	0	1 (0)	1 (0)	0.12
Polycystic ovarian syndrome	245,349 (700/244,649)	41	1 (0)	0 (0)	0.60
Cyst or abscess of Bartholin's gland	245,349 (776/244,573)	-	1 (0)	1 (0)	0.25
Ectopic pregnancy	208,121 (641/207,480)	-	1 (0)	1 (0)	0.23
Ever had same-sex intercourse	409,210 (13,953/395,257)	2	1 (0)	0 (0)	1.21

Table 6.1 (Continued)

Trait	N (Cases/Controls)	Known signals†	REPROWAS signals $P<5.0\times10^{-8}$ ($P<2.7\times10^{-11}$)	Novel signals‡ $P<5.0\times10^{-8}$ ($P<2.7\times10^{-11}$)	h^2 (%)§
Antepartum haemorrhage	208,121 (687/207,434)	-	1 (0)	1 (0)	0.00
Vascular disorders of male genital organs	206,951 (673/206,278)	-	1 (0)	1 (0)	0.12
Salpingitis and oophritis	245,349 (579/244,770)	-	1 (0)	1 (0)	0.28
Inflammatory disorders of breast	245,349 (511/244,838)	-	1 (0)	1 (0)	0.00
Obstructed labour due to malposition and malpresentation of foetus	208,121 (465/207,656)	-	1 (0)	1 (0)	0.23
Other specified disorders of penis	206,951 (474/206,477)	-	1 (0)	1 (0)	0.00
Multiple gestation	208,121 (292/207,829)	1	1 (0)	1 (0)	0.11
Excessive bleeding in the premenopausal period	245,349 (317/245,032)	-	1 (0)	1 (0)	0.00
Maternal care for poor foetal growth	208,121 (271/207,850)	-	1 (0)	1 (0)	0.00
Postmenopausal bleeding	245,349 (9,912/235,437)	-	0 (0)	0 (0)	0.30
Ovarian cysts	245,349 (8,253/237,096)	-	0 (0)	0 (0)	0.84
Other abnormal uterine and vaginal bleeding	245,349 (4,518/240,831)	-	0 (0)	0 (0)	0.34
Delivery complicated by foetal stress	208,121 (3,587/204,534)	-	0 (0)	0 (0)	0.10
Poly of cervix uteri	245,349 (3,693/241,656)	-	0 (0)	0 (0)	0.40
Excessive and frequent menstruation with irregular cycle	245,349 (2,976/242,373)	-	0 (0)	0 (0)	0.44
Ever had cervical smear test	244,634 (239,661/4,973)	-	0 (0)	0 (0)	0.53
Spontaneous abortion	208,121 (2,695/205,426)	-	0 (0)	0 (0)	0.18
Ever had stillbirth, spontaneous miscarriage or termination	205,106 (76,740/128,366)	-	0 (0)	0 (0)	1.60
Ever taken oral contraceptive pill	244,520 (200,828/43,692)	-	0 (0)	0 (0)	2.27
Gestational diabetes only	7,112 (863/6,249)	2	0 (0)	0 (0)	0.13
Pelvic peritoneal adhesions	245,349 (2,723/242,626)	-	0 (0)	0 (0)	0.16
Other menopausal and other perimenopausal disorders	245,349 (2,430/242,919)	-	0 (0)	0 (0)	0.00
Single spontaneous delivery	208,121 (2,211/205,910)	-	0 (0)	0 (0)	0.00
Long labour	208,121 (2,089/206,032)	-	0 (0)	0 (0)	0.35
Malignant neoplasm of ovary	245,349 (1,463/243,886)	115	0 (0)	0 (0)	0.00
Inflammatory diseases of prostate	206,951 (1,858/205,093)	-	0 (0)	0 (0)	0.80
Other non-inflammatory disorders of cervix uteri	245,349 (1,951/243,398)	-	0 (0)	0 (0)	0.11
Dysmenorrhoea (menstrual cramps)	245,349 (1,981/243,368)	5	0 (0)	0 (0)	0.20

Table 6.1 (Continued)

Trait	N (Cases/Controls)	Known signals†	REPROWAS signals $P<5.0\times10^{-8}$ $(P<2.7\times10^{-11})$	Novel signals‡ $P<5.0\times10^{-8}$ $(P<2.7\times10^{-11})$	h^2 (%)§
Irregular menstruation	245,349 (1,986/243,363)	-	0 (0)	0 (0)	0.24
Other non-inflammatory disorders of vulva and perineum	245,349 (1,789/243,560)	-	0 (0)	0 (0)	0.06
Unspecified lump in breast	245,349 (1,654/243,695)	-	0 (0)	0 (0)	0.38
Postpartum haemorrhage	208,121 (1,414/206,707)	-	0 (0)	0 (0)	0.56
Benign neoplasm of ovary	245,349 (1,609/243,740)	-	0 (0)	0 (0)	0.50
Premature rupture of membranes	208,121 (1,170/206,951)	0	0 (0)	0 (0)	0.16
Prolonged pregnancy	208,121 (1,293/206,828)	-	0 (0)	0 (0)	0.27
Missed abortion	208,121 (1,261/206,860)	-	0 (0)	0 (0)	0.00
Medical abortion	208,121 (1,156/206,965)	-	0 (0)	0 (0)	0.32
Endometrial hyperplasia	245,349 (1,219/244,130)	-	0 (0)	0 (0)	0.00
Hypertrophy of uterus	245,349 (1,239/244,110)	-	0 (0)	0 (0)	0.43
Dyspareunia (painful intercourse)	245,349 (1,131/244,218)	-	0 (0)	0 (0)	0.32
Inflammatory disorder including orchitis and epididymitis	206,951 (1,063/205,888)	-	0 (0)	0 (0)	0.19
Haemorrhage in early pregnancy	208,121 (988/207,133)	-	0 (0)	0 (0)	0.00
Other inflammation of vagina and vulva	245,349 (851/244,498)	-	0 (0)	0 (0)	0.06
Malposition of uterus	245,349 (835/244,514)	-	0 (0)	0 (0)	0.00
HIV	452,300 (346/451,954)	1	0 (0)	0 (0)	0.12
Unspecified maternal hypertension	208,121 (638/207,483)	-	0 (0)	0 (0)	0.03
Infections with a predominantly sexual mode of transmission	452,300 (618/451,682)	-	0 (0)	0 (0)	0.03
Diabetes mellitus in pregnancy	208,121 (385/207,736)	2	0 (0)	0 (0)	0.00
Other female genital prolapse	245,349 (492/244,857)	0	0 (0)	0 (0)	0.00
Inflammatory disease of uterus, except cervix	245,349 (310/245,039)	-	0 (0)	0 (0)	0.00
Other congenital malformations of male or female genital organs	452,300 (122/452,178)	-	0 (0)	0 (0)	0.06

† Known hits based on number of associations in EBI GWAS Catalog[267]

‡ Loci were considered novel if they were not in LD with any known GWAS hits ($r^2 > 0.05$)

§ SNP-heritability calculated using LDSC[113]

* For hair/balding pattern, male participants chose from one of 4 images most similar to their hair pattern; hence this is an ordered categorical variable

6.3.2 Individual SNPs are associated with multiple reproductive outcomes

As described above, one advantages of the PheWAS method is that it allows identification of genetic variants which have potentially pleiotropic effects. Of the 1,966 independent variants which were significantly associated with at least one reproductive trait at a genome-wide significant level, 40 were also the lead variant for at least one other trait. The most pleiotropic SNP was rs11031005, which maps near the *FSHB* gene and showed associations with eight different reproductive traits in women (**Figure 6.2**). *FSHB* encodes the beta subunit of follicle-stimulating hormone, and rising levels of this hormone have been hypothesised to accelerate the reproductive ageing process, particularly in women[268,269]. Another SNP (rs7124615/C) which is proximal to the gene encoding SCGB1C1, a protein found in high concentrations in prostate and uterus secretions, was associated with leiomyoma (fibroids) of the uterus (OR=1.15, P=1.6×10^{-18}), uterine polyps (OR=1.15, P=1.9×10^{-9}) and risk of bilateral oophorectomy (OR=1.21, P=2.8×10^{-10}) in women and with prostate hyperplasia (OR=1.13, P=1.9×10^{-11}) in men. Other variants of interest showed associations with a range of traits, including associations in both sexes. rs34811474/A, a missense variant within the *ANAPC4* gene which is involved in regulation of mitosis, was associated with balding pattern (β=0.02, P=4.6×10^{-11}), age at first sexual intercourse (β=0.05 years/allele, P=5.9×10^{-9}) and age at menarche (β=0.03 years/allele, P=4.6×10^{-9}). Similarly, a variant (rs72709458/T) mapping near the gene for telomerase reverse transcriptase (*TERT*) was associated with increased risk of leiomyoma of the uterus (OR=1.12, P=2.9×10^{-16}) and excessive menstruation (OR=1.11, P=6.1×10^{-9}) in women and showed a protective effect on prostate cancer risk (OR=0.85, P=2.7×10^{-14}) in men. Such examples highlight the shared genetic architecture of reproductive health, with multiple pathways and processes that individual genetic variants can influence.

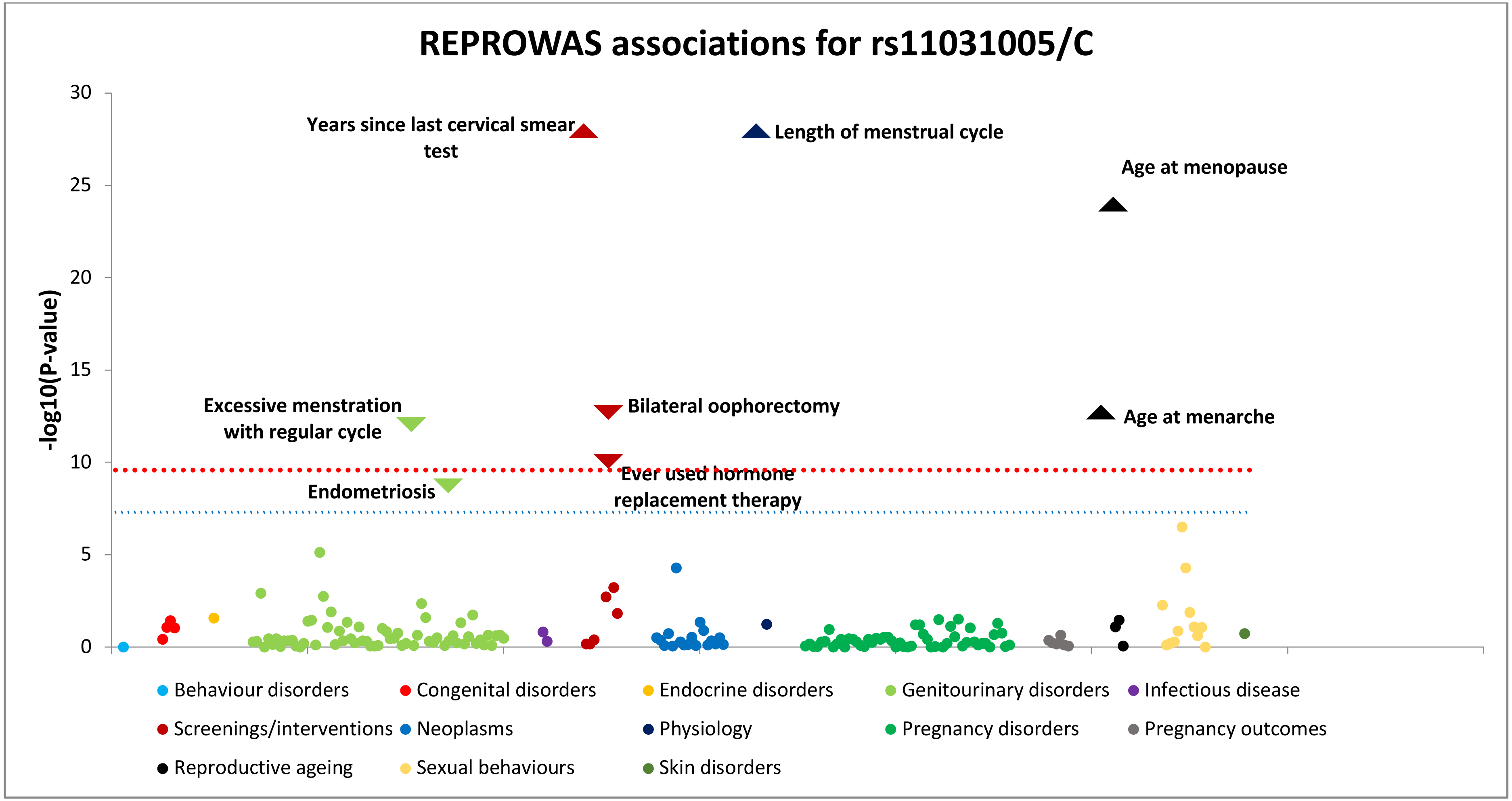


Figure 6.2: Phenotypic associations for the C-allele at rs11031005 near the *FSHB* gene. Phenotypes are grouped into 13 categories, coded by colour. The dashed blue line represents the genome-wide significance threshold ($P<5.0\times10^{-8}$) while the dashed red line denotes the Bonferroni-corrected significance level ($P<2.7\times10^{-10}$) which accounts for the number of tests. For significantly associated phenotypes, upward-pointing triangles represent positive associations while downward-pointing triangles represent negative associations. Non-significant associations are represented by filled circles.

6.3.3 Genomic regions are enriched for associations with reproductive traits

In addition to individual SNPs which are associated with multiple traits, the PheWAS framework also allows for identification of regions of the genome which show enrichment for associations with reproductive outcomes. A total of 95 genes (based on annotation of nearest gene for independently associated loci) were associated with multiple traits at a Bonferroni-corrected level of significance ($P<2.7\times10^{-10}$). Variants mapping to *WNT4*, which encodes a protein involved in embryonic development of the female reproductive system as well as androgen regulation in males, showed the most associations for an individual gene in this study. Variants at this locus were associated with multiple and diverse reproductive traits, including risk of bilateral oophorectomy (rs2235529/T; OR=1.12, $P=2.3\times10^{-14}$), leiomyoma of uterus (rs2235529/T; OR=1.15, $P=3.5\times10^{-19}$), age at hysterectomy (rs3820282/T; β=0.45 years/allele, $P=7.0\times10^{-11}$), enterocele and rectocele (rs3820282/C; OR=0.83, $P=8.2\times10^{-12}$), offspring birth weight (rs56318008/C; β=-0.03 kg/allele, $P=9.7\times10^{-9}$), uterovaginal prolapse (rs61768001/C; OR=0.85, $P=2.7\times10^{-13}$) and endometriosis (rs61768001/C; OR=1.17, $P=7.7\times10^{-12}$). In addition, an approximately 800kB region on chromosome 6 (cytoband 6q25.1-2) showed associations with 12 different reproductive traits in females (**Figure 6.3**). Variants at this locus mapped to three different genes: *SYNE1*, *CCDC170* and *ESR1*. The *ESR1* gene encodes oestrogen receptor alpha (ERα) transcription factor, a protein which is activated in response to binding by oestrogen promoting cell proliferation and differentiation. As well as being the primary sex hormone in women, oestrogen activity is essential for many cellular processes, and as such ERα is expressed in many different tissues throughout the body. eQTL analysis using GTEx data was performed and identified a number of variants associated with expression of this gene in a diverse range of tissues. It is notable that few appear to be related to any of the associated conditions however, which may reflect limitations of the GTEx data to distinguish different cell types within tissues. Because of its role in female sexual development and the ability to promote cell proliferation, the genetic variation at the *ESR1* locus has been implicated in a number of reproductive health outcomes including timing of menarche[35] and menopause[245] as well as breast cancer risk[270]. The results presented here are consistent with previous findings and clearly implicates oestrogen activity as a key mediator of many aspects of reproductive health in women.

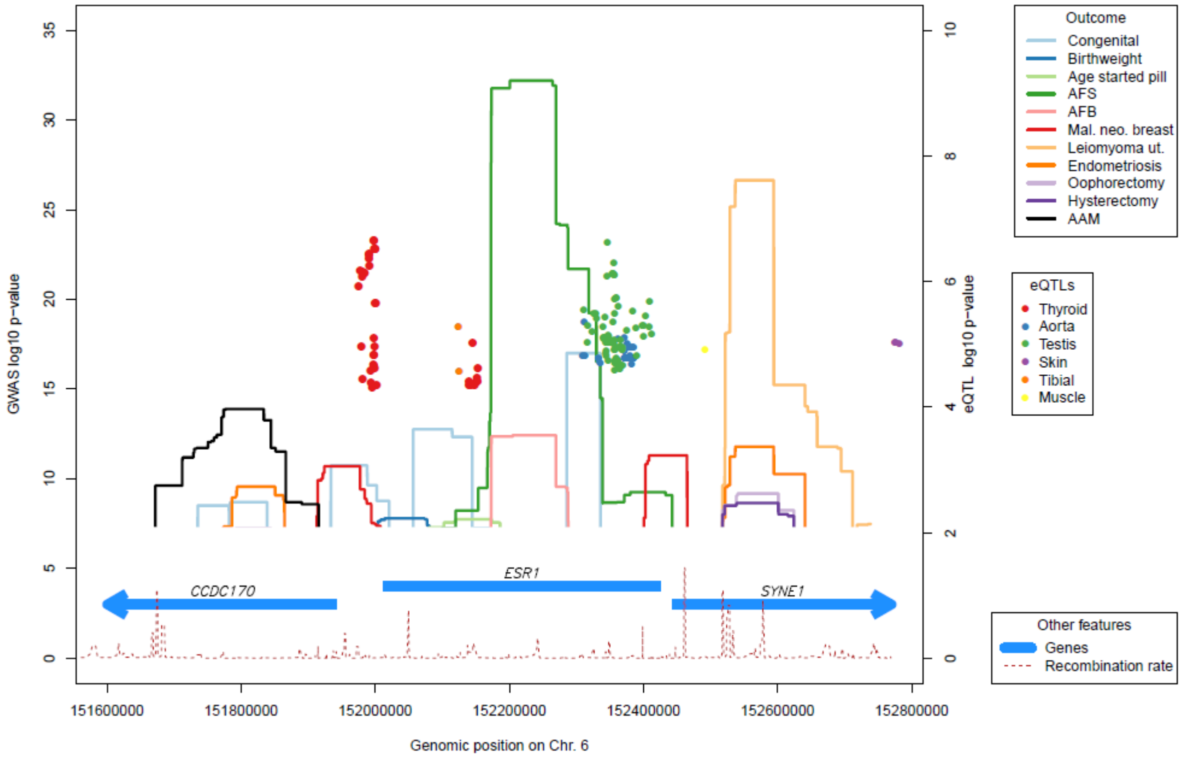


Figure 6.3: Regional association plots for select phenotypes at the ESR1/CCDC170/SYNE1 locus. $\log_{10}$ P-values are plotted for variants in this region, depicted by the joined lines. eQTLs for the three genes are also plotted, colour-coded by tissues in which they are expressed.

6.4 DISCUSSION

Reproductive health is an important component of overall health and wellbeing, and many aspects have been shown to be influenced by genetic factors. However while much progress has been made in resolving the genetic architecture for several traits, genetic discovery for many other traits has not been considered, limiting our insight into the influence of genetic factors on reproductive health as a whole. In this chapter, I have described a study to fill some of these gaps in knowledge by performing genetic discovery across a broad range of reproductive health phenotypes.

6.4.1 Limited evidence of polygenic architecture for many reproductive traits

GWAS were conducted for 181 traits which were broadly associated with reproductive health outcomes in the UK Biobank study. Estimates of the variance explained by SNPs in this sample were low to moderate for numerous traits, suggesting limited genetic influence and low heritability for many aspects of reproductive health. Heritability estimates were generally higher for continuous traits than for binary traits, likely the result of reduced statistical power due to the low number of cases for many disease traits. For traits which have previously been analysed in large-scale GWAS the proportion of variance explained by SNPs was lower than that reported in those studies. For example, SNP-based heritability for age at first sexual intercourse was estimated to be approximately 24% in a previous GWAS[271], compared to only 12.9% in this study. This may be due in part to differences in the methods used to estimate heritability. Here, LD score regression was used, which estimates heritability using common variants only. Simulations have shown that for non-polygenic traits, heritability may be underestimated compared to genomic relatedness residual maximum likelihood (GREML)-based methods[272]. For many of the disease traits in this study it appears that few genetic determinants of high penetrance may drive the genetic component of susceptibility. This may explain the preponderance of traits with very low heritability estimates which still have multiple associated variants. For example, "absent, scanty or rare menstruation" had a heritability estimate of 0.13% in this sample, despite having 13 independent genome-wide significant associated loci. Each of these loci are rare, with MAF<0.03 and odds ratios ranging from 12.9 to 97.4. Rare mutations affecting the hypothalamic-pituitary-gonadal axis which cause delayed puberty onset have been reported[273], though as mentioned above the inflated ORs seen here may be due to the extremely unbalanced case-control ratio as only 120 cases of this condition were reported. More generally, this highlights a limitation of UK Biobank for this analysis. Because data is only available for hospital admissions, under-ascertainment of cases will exist for many reproductive conditions as only the most severe will require hospitalisation. As a result, it is likely that heritability is underestimated for some traits.

6.4.2 PheWAS framework allows for multiple levels of inference

We identified 1,966 genome-wide significant variant-trait associations, of which 871 remained significantly associated after application of a multiple test correction. Many of these associations were novel, as the majority of the traits had never before had genetic discovery conducted. By combining the GWAS summary results of all traits, inferences can be made at various levels of resolution of the genome in a PheWAS framework. Employing this approach revealed evidence of widespread pleiotropic effects amongst individual variants, with many having genome-wide significant associations with multiple, and often diverse, reproductive phenotypes. This was best exemplified by the SNP rs11031005 located proximal to the *FSHB* gene, a pituitary hormone which along with luteinising hormone is responsible for stimulating gamete production. This variant was associated with 8 different reproductive traits with seemingly diverse aetiology, including menstrual traits, ovarian cancer risk and reproductive timing. As mentioned, FSHB has been associated with accelerated reproductive ageing, and has been identified as a likely pharmacologic target for intervention in assisted reproductive technology[274]. The findings presented here may therefore have implications for the discovery of new drugs targeting this gene system, particularly as it relates to possible secondary effects.

This degree of pleiotropy was also apparent at the level of individual genes as well as chromosomal regions. This is highlighted by the *ESR1/SYNE1/CCDC170* locus on chromosome 6, where multiple variants within this region were independently associated with different aspects of reproductive health. The *ESR1* gene has frequently been implicated in genetic association studies of reproductive traits[40,270,275–277], and the role of oestrogen in reproductive function makes it a likely candidate for influencing disease risk.

6.4.3 Strengths, limitations and future directions

This study design has several key advantages compared to other similar efforts to conduct PheWAS in large cohort studies. Cortes *et al.* recently described a method to interrogate the hierarchical tree structure of ICD-10 coded diagnoses in UK Biobank using a Bayesian framework[261]. However for in their analysis, all patients who did not have the corresponding code for the disease were considered controls, with limited or no consideration of potential sources of confounding. In this study phenotypes were manually curated and considered individually in turn, with appropriate exclusions applied for each trait. This mitigates much of the potential for bias, and increases the robustness of the results. A second advantage of this study design is the use of a single data source for the analysis. While meta-analyses of multiple cohorts has the benefit of increasing the sample size and statistical power for discovery, a key limitation is the potential for bias due to differences in data collection or analysis between studies, giving rise to heterogeneity concerns. Here, the use of standardised ICD-10 codes and consistent methods of data collection used in UK Biobank mitigates this as a potential source of bias.

This study also has several important limitations which must be considered when interpreting the results. As discussed in earlier sections, the use of HES data for ascertainment of disease makes it likely that many cases with subclinical symptoms will not be included. With few cases for many conditions we are only powered to detect variants of large effect, which can bias heritability estimates downward and risks missing some of the common variation which may drive more mild cases of disease. Similarly, because UK Biobank is an older cohort focused on studying diseases of middle and old age conditions presenting in youth may be less well ascertained. This is particularly true for self-reported conditions, where recall of minor conditions which occurred decades previously may be limited. As an example, only ~400 participants reported ever having acne, which is likely to be a significant underestimate given the prevalence of the disease in the general population[278].

Some of these issues can be addressed in the future with the incorporation of primary care and prescription data from linkage to GP practice records, which have recently been released by UK Biobank. This may assist with ascertainment of more of the milder cases of disease, allowing for more robust genetic inference which may provide more insight into more common variation within the population. Further down the line, integration of exome and whole-genome sequencing will allow for a more fine-scale analysis and greater resolution of casual variants of reproductive disease.

6.4.4 Conclusions

This chapter describes the generation of the REPROWAS database, which expands the breadth of knowledge on the genetic determinants of reproductive health and disease. Combining GWAS summary results across multiple, expertly curated reproductive phenotypes reveals many novel variant-trait associations for conditions which have previously been understudied, and provides further evidence for substantial shared genetic architecture underlying many aspects of reproductive health. This data will be made available to other researchers, providing a valuable tool for examining genetic associations with reproductive health outcomes.

Chapter 7: Sex Hormones, BMI and Puberty Timing

Contributions and Collaborations

The genetic discovery for sex hormones was the result of a collaboration with colleagues including Katherine Ruth, Felix Day, Jessica Tyrrell, Deborah Thompson, Andrew Wood, Anubha Mahajan, Robin Beaumont, Laura Wittemans, Susan Martin, Mesut Erzurumluoglu, the 23andMe Research Team, Mark McCarthy, Claudia Langenberg, Douglas Easton, Nicholas Wareham, Stephen Burgess, Anna Murray, Ken Ong, Timothy Frayling and John Perry. All subsequent analyses were performed by me.

SUMMARY

While the association between puberty timing and health outcomes has been well established, comparatively less is known about the biological mechanisms that explain these associations. Due to the association between BMI and puberty, a common theory is that a genetic overlap between variants associated with BMI and puberty may provide a causal link to poor metabolic health. Alternatively, it has been suggested that sex hormone exposure may have a role. Sex steroid hormones are the key drivers of pubertal development and play a critical role in reproductive function. Sex steroids also have cell proliferative effects in many tissues, and circulating levels have been associated with risks for many diseases, and cancers in particular. However, to date few large-scale genetic studies have been conducted on variation in sex hormone levels within the population. In this chapter, I describe a study that uses data from a greatly expanded GWAS of sex hormone traits, including testosterone, SHBG and oestradiol. I conduct MR analyses using puberty timing associated variants for hundreds of reproductive outcomes in the REPROWAS data, to identify associations between puberty timing and general reproductive health. I then perform univariate and multivariate MR analyses using the newly generated sex hormone GWAS data, in combination with previously published data on BMI-related variants, to investigate the extent to which pubertal associations with reproductive outcomes are mediated by these potential explanatory factors.

7.1 BACKGROUND

7.1.1 Pathways from earlier puberty to reproductive disease

A substantial body of evidence exists in the literature linking puberty timing with reproductive health outcomes. Comparatively less is known about the mechanisms driving these associations, however. One of the most prevalent theories linking earlier puberty to adverse health outcomes is based on the bidirectional association between earlier puberty and higher BMI[279,280]. Genetic studies investigating population variation in puberty timing have provided insights which further implicate BMI-related factors as a potential explanatory mechanism. This is based on a strong and inverse genetic correlation between the two traits (rg= -0.35 for AAM in women[35]; rg= -0.32 for VB in men (Chapter 3)) and significant enrichment for known BMI signals among puberty-associated loci. This high degree of shared genetic architecture underlying regulation of puberty timing and of BMI suggests the possibility for pleiotropic effects of BMI-associated loci, which may mediate the associations between earlier puberty timing and adult disease.

An alternative theory has been proposed that sex hormones may be the primary driver of the associations between puberty timing and disease risk. Sex steroid hormones are the primary regulators of reproductive function, and their increased production in pre-pubertal children is an essential part of the initiation of puberty and the development of secondary sexual characteristics. The biological activity of sex hormones, primarily testosterone and oestrogens, are mediated through the binding to their receptors on the surfaces of cells. This has various effects on target tissues including promotion of cell proliferation and changes in cellular metabolism, among many other actions. Oestrogen and testosterone receptors are most highly expressed in reproductive tissues, but expression is also seen in non-reproductive tissues as well.

Due to the importance of sex hormone levels in the initiation of puberty and their diverse effects on cellular processes, it has been hypothesised that the increased duration and intensity of sex hormone exposure in individuals who go through puberty earlier may explain some of the reported adverse health associations[281–283]. This is supported by observed associations between sex hormone levels and disease risk for several reproductive conditions. For example, higher testosterone levels are observed in women with polycystic ovary syndrome (PCOS)[284], and testosterone has been positively associated with increased risk of cardiovascular disease, obesity and T2D[285–287]. Sex hormones are also known to promote growth and metastasis of certain tumours of reproductive organs. Breast cancer subtypes are routinely classified by their expression of receptors for oestrogen and progesterone (ER+/- and PR+/- tumours, respectively), and higher circulating levels of oestrogen have been associated with poorer health outcomes in

ER+ tumours[288]. Similarly, many ovarian cancer types are known to respond to oestrogen and progesterone[289], while in men there is some evidence for increased risk of prostate cancer in individuals with higher circulating levels of testosterone and oestrogen, though this evidence is not consistent across studies[290–293]. In totality however, there is a convincing body of evidence in support of the theory of increased endogenous sex hormone exposure as a mediator of disease risk.

7.1.2 Genetics of sex hormone levels

Despite the progress that has been made in determining the effect of sex hormone exposure on many conditions, for many other reproductive traits and diseases this association has not been characterised. Genetic methods many offer new insights, particularly as twin studies have demonstrated high heritability for circulating levels of testosterone ($h^2 \sim 65\%$)[294] and oestradiol ($h^2 \sim 30\text{-}45\%$)[295]. To date, however, GWAS have identified few genomic loci associated with these traits, including just two signals each for testosterone (*SHBG* and *FAM9B*) and oestradiol levels (*CYP19A1* and *FAM9B*) in men[296–298]. Another key regulator of sex hormone activity, sex hormone binding globulin (SHBG), has also been shown to have high heritability ($h^2 \sim 50\%$ in twin studies)[299] and has had 12 associated genetic loci discovered[300]. SHBG regulates bioactivity of sex hormones by binding to testosterone and oestrogens in the circulation, thereby preventing them from binding with their receptors and exerting their effects on target tissues. SHBG may also exert effects independently through binding with its own receptor. To date, genetic discovery for these sex hormone traits has been limited by relatively small sample sizes. Based on the high heritabilities of these traits it is likely that many more genes of small individual effect contribute to the genetic component of the population-level variation of their circulating levels, which will require larger cohorts to identify.

7.1.3 Areas of opportunity

While associations between puberty timing and several common reproductive health outcomes have been reported, many other traits have not been previously been considered. Furthermore, identifying associations is only the first step towards truly understanding the aetiological mechanisms which relate earlier sexual maturation to unfavourable health consequences later in life. Inferring casual associations using Mendelian randomisation analyses has the potential to increase our understanding of this important subject. In the previous chapter, I described the development of the REPROWAS database, which provides extensive data on the genetic determinants of a large number of reproductive traits. Here, I describe a study leveraging this data to systematically assess the effect of age at puberty onset on over 180 reproductive health outcomes, the majority of which have never been considered in such a context. The objectives of this study are as follows:

1) To identify reproductive health traits which are causally influenced by age at puberty;
2) To provide insight into the extent to which two proposed mechanisms, BMI-related factors and endogenous sex hormone exposure, mediate any identified associations between puberty timing and reproductive traits; and
3) To determine whether sex-specific genetic determinants of puberty timing have a differential effect on associated reproductive health outcomes.

In collaboration with colleagues who recently conducted an expanded GWAS for sex hormone levels in UK Biobank participants, this effort will provide valuable insights and pave the way for future, in-depth investigations into the biological consequences of early puberty.

7.2 METHODS

7.2.1 GWAS for sex hormone traits

Genetic discovery for sex hormone phenotypes were performed by colleagues at the University of Exeter and the MRC Epidemiology Unit, University of Cambridge (Ruth *et al*., unpublished). The following is a summary of the methods used in that study.

The study population comprised the ~500,000 participants in the UK Biobank cohort, for whom a panel of 34 biomarkers were produced from blood and urine samples collected at baseline assessment (see Chapter 2). These included three traits pertaining to sex hormones: SHBG, testosterone and oestradiol. These three phenotypes were used in the analysis described in this chapter. A fourth phenotype, which derived the amount of bioactive testosterone (Free-T) based on a previously validated method[301] was also included in the analysis. Genetic discovery was limited to individuals of white European ancestry, identified based on self-reported ancestry from online questionnaires and the first four genetically determined principle components. In total, 425,097 participants who had phenotype and genotype data which passed QC measures were included in the analyses. Association testing was performed using linear mixed models implemented in BOLT-LMM[108]. For each sex hormone phenotype, analyses were performed separately in men and women. Due to the older age of the UK Biobank cohort, oestradiol levels in women were often too low to detect as many participants had already reached menopause. As such, GWAS for oestradiol was only performed in men, where the trait was dichotomized to men at or below the lower limit of detection versus all others. All other traits were modelled as continuous variables based on measured concentrations of the hormone. Each model included age at baseline, genotyping chip and the first 10 genetically determined principle components as covariates. For SHBG, BMI was included as a covariate based on previous evidence that this reduces variance for the trait and therefore increases statistical power for detection of variant associations[300]. Effect estimates for SHBG were then compared to unadjusted models in order to account for the potential for collider bias[201], with variants showing large discrepancies discarded. Variants identified as being associated with SHBG in the adjusted models were used for the analyses described here, with the corresponding effect estimates and standard errors from the unadjusted models used to build weights in downstream analyses.

Variants with an imputation quality score <0.5 or MAF <0.01 were excluded from the analysis. Independent variants were identified by distance-based clumping of genome-wide significant variants ($P<5.0\times10^{-8}$), where the variant with the lowest P-value within 1MB was considered the lead variant at that locus. For cases where multiple lead variants were in LD with each other ($r^2>0.05$), they were excluded from the analysis.

7.2.2 Mendelian randomisation analyses for reproductive outcomes

In order to investigate the extent to which puberty timing influences reproductive health outcomes, Mendelian randomisation (MR) analyses were performed for all of the 181 phenotypes in the REPROWAS study (described in Chapter 6). Given the greater size of the discovery sample for the AAM GWAS compared to the voice breaking meta-analysis described in this thesis, as well as the strong genetic correlation between male and female puberty timing[35], the 389 variants identified in the most recent GWAS for AAM were used as the primary instrument for puberty timing in this analysis. These were applied to both male and female reproductive traits. Two-sample inverse variance weighted (IVW) MR analyses were performed for each of the 181 reproductive traits, using the AAM instrument as the exposure and the corresponding effect estimates from REPROWAS summary statistics as the outcome measure. Analyses were implemented using the 'TwoSampleMR' package in R. For all nominally significant associations identified by this method (P<0.05), further IVW MR analyses were conducted using instruments for BMI and the sex hormone phenotypes. For BMI, this comprised the 96 bi-allelic variants identified in the European sex-combined sample from the most recently published genetic discovery by the GIANT consortium[155]. For the sex hormone phenotypes, instruments were derived from the UK Biobank GWAS described above with sex-specific effect estimates and standard errors used for male and female reproductive outcomes. For all analyses, alleles were aligned such that the trait-increasing allele was considered the effect allele.

For all traits which were nominally associated with puberty, estimates for the degree of horizontal pleiotropy and heterogeneity were determined using the MR-Egger intercept and Cochrane's Q statistic, respectively. For both measures, P-values <0.05 are indicative of potential bias. To examine the impact such bias might have on the results, MR-Egger and weighted median (WM) analyses were conducted. These methods are robust to the effects of horizontal pleiotropy and provide a means of sensitivity analysis for the observed associations.

7.2.3 Multivariate MR Analyses

To further investigate the extent to which BMI and sex hormones mediate the associations between puberty timing and reproductive health outcomes, multivariate MR analyses were performed for all traits which were significantly associated with puberty timing. In each case, lookups for each of the AAM SNPs included in the instrument were performed for BMI and each of the sex hormone variables. These were then included as covariates in separate models using the method described by Burgess *et al.*[123] to examine the attenuation of the IVW effect estimate for AAM on each reproductive outcome.

7.2.4 Comparison of male and female-specific instruments for puberty timing

To assess the potential for sex-specific genetic effects on puberty timing influencing the results, a further sensitivity analysis was performed using the 76 male puberty timing variants identified meta-analysis in Chapter 3 as the instrumental variable. This analysis was run for all 181 REPROWAS traits, using the IVW estimate to compare with the estimates obtained from the AAM instrument.

7.3 RESULTS

7.3.1 Genome-wide associations for sex hormones and generation of genetic instruments

Genome-wide association studies for circulating levels of total testosterone (Total-T), bioactive testosterone (Free-T) and SHBG in both sexes, and total oestradiol in men only identified a total of 1,528 independently associated variants for all traits combined. Along with previously reported summary statistics for AAM, BMI and voice breaking, these were used to generate genetic instruments for MR analyses (**Table 7.1**).

Table 7.1: Genome-wide significant signals for sex hormones and other exposures

Exposure	Study	GWAS signals	Included in instrument for exposure#
AAM	Day et al.[35]	389	358
Voice breaking	Hollis et al. (Chapter 3)	76	76
BMI	Locke et al.[155]	97	96
Total-T	Ruth et al. (unpublished)		
Men		231	221
Women		254	248
Free-T	Ruth et al. (unpublished)		
Men		125	106
Women		180	176
SHBG	Ruth et al. (unpublished)		
Men		357	342
Women		359	351
Oestradiol	Ruth et al. (unpublished)	22	19

#– REPROWAS includes only autosomal variants. One BMI variant was not bi-allelic and was excluded from the analysis. In addition, 10 sex hormone variants were not bi-allelic and were also excluded (3 for Total-T in women; 1 for Free-T in men; 2 for Free-T in women; and 4 for SHBG in women).

7.3.2 Evidence for causal effects of puberty timing on reproductive health outcomes

Mendelian randomisation analyses were conducted to assess the effect of puberty timing (represented by AAM) on all reproductive health outcomes in the REPROWAS database (n=181 traits). This comprised 146 female traits, 25 male traits and 10 traits shared between the sexes. Using the IVW method, AAM showed nominally significant (P<0.05) associations with 36 traits (26 female-specific, 8 male-specific and 2 shared; **Supplementary Tables 7.1 and 7.2**). After correcting for the number of tests, this resulted in 12 significant associations in females ($P<3.2\times10^{-4}$, =0.05/156) and 4 significant associations in males ($P<1.4\times10^{-3}$, =0.05/35). Reassuringly, the most significant association for females was with age at menarche and for males with relative ages at voice breaking and first facial hair.

AAM showed further positive associations with several other traits related to reproductive ageing, including age at menopause (β=0.18 years/year later AAM, 95% CI=0.08-0.27), age at first live birth (β=0.28 years, 95% CI=0.20-0.37), age at last live birth (β=0.26 years, 95% CI=0.19-0.34), and age at first sexual intercourse (AFS; β=0.25 years, 95% CI=0.20-0.30). These were all in the expected direction of effect, indicating that later onset of puberty delays the onset of other reproductive events. Consistent with the hypothesis that earlier puberty confers increased health risks in later life, AAM was negatively associated with risks for leiomyoma of the uterus (OR=0.90, 95% CI=0.87-0.94), bilateral oophorectomy (OR=0.91, 95% CI=0.88-0.95), and cystitis/urinary tract infections (OR=0.93, 95% CI=0.90-0.95). Leiomyomas (fibroids) are non-cancerous growths on the uterus which can cause symptoms including heavy menstruation and pain, while bilateral oophorectomies (removal of both ovaries) are preventative measures for women at high risk of ovarian and breast cancer. The remaining associations (age started and finished taking oral contraceptive pill, age started HRT, and years since last cervical smear test) are all proxies for other reproductive timing traits and are therefore largely uninformative by themselves. However the directions of effect are consistent with expectations for each of these traits. In men, no further associations were observed.

7.3.2 MR sensitivity analyses

All of the associations showed evidence of significant heterogeneity. MR-Egger regressions were non-significant for all traits, however this is likely due to the low statistical power of this test. Conversely weighted median MR analyses, which relies on different assumptions to MR-Egger and is generally better powered to detect associations, were nominally significant for all traits ($P<0.05$). Furthermore, none of the traits showed evidence of horizontal pleiotropy (MR-Egger intercept P-value >0.05) with the exception of AAM ($P_{intercept}=1.9\times10^{-5}$). Pleiotropy would be expected in this instance given that known AAM variants were used to generate the instrument for exposure, and the trait was included as an outcome here merely as a positive control. Therefore, taken together the sensitivity analyses suggest that heterogeneity and horizontal pleiotropy are unlikely to significantly bias the interpretation of the results.

7.3.3 Associations for BMI and sex hormones among puberty-associated traits

For each of the traits which were significantly associated with AAM, further MR analyses were conducted using instruments for BMI, Total-T, Free-T and SHBG in both sexes and ostradiol in men. For the sex hormone traits, the instruments were derived using sex-specific effect estimates and applied to male and female-specific traits accordingly.

Associations with BMI

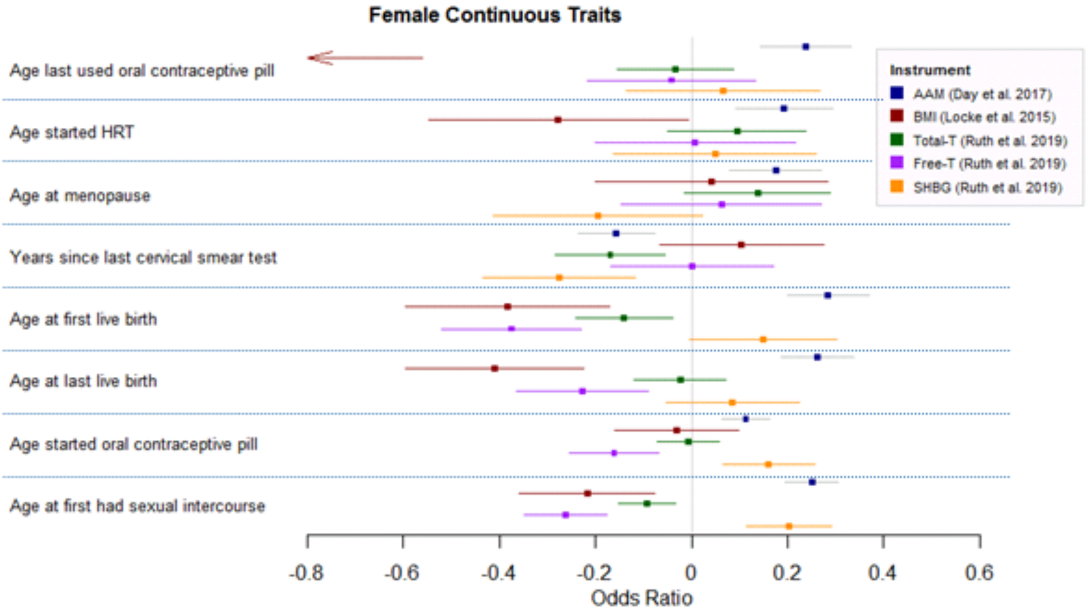
BMI was negatively associated with 6 of the 9 puberty-associated continuous phenotypes in women with at least nominal significance (**Figure 7.1 (a) and Supplementary Table 7.1**). The inverse relationship with these traits is consistent with the observation that higher BMI is associated with earlier onset of puberty. In contrast, among binary traits BMI was only associated with risk of leiomyoma of the uterus with nominal significance, where higher genetically predicted BMI showed evidence of increasing the risk for this trait (OR=1.13, 95%CI=1.01-1.27). No significant association with risk for either bilateral oophorectomy (OR=0.98, 95%CI=0.88-1.09) or cystitis/urinary tract infection (OR=1.06, 95%CI=0.96-1.07) was observed (**Figure 7.1 (b) and Supplementary Table 7.2**). In males, age at first sexual intercourse (β=-0.22, 95%CI=-0.36- -0.08) (**Figures 7.1 (c) and (d) and Supplementary Tables 7.1 and 5.2**).

Associations with sex hormone traits

Several continuous reproductive traits in women showed evidence of associations with testosterone and SHBG. A higher genetically predicted level of Total-T was significantly associated ($P<4.2\times10^{-3}$, =0.05/12 tests) with earlier sexual debut (AFS β=-0.09, 95%CI=-0.15- -0.03) and fewer years since last cervical smear test (β=-0.17, 95%CI=-0.28- -0.05) (**Figure 7.1 (a) and Supplementary Table 7.1)**. When considering bioactive testosterone (Free-T), AFS remained significantly associated (β=-0.26, 95%CI=-0.35 - -0.18) while negative associations were observed with additional traits related to reproductive ageing and behaviour including AAM (β=-0.11, 95%CI=-0.19 – 0.04) and ages at first (β=-0.37, 95%CI=-0.52 - -0.23) and last (β=-0.23, 95%CI=-0.36 - -0.09) live birth. SHBG showed only one significant association, where higher genetically predicted SHBG conferred later AFS (β=0,20, 95%CI=0.12-0.29). This is consistent with the observation for the effects of bioactive testosterone as SHBG is expected to be inversely correlated with Free-T, though this does not preclude a direct effect of SHBG.

Higher Total-T showed evidence of having a protective effect for leiomyoma of the uterus (OR=0.88, 95%CI=0.82-0.94) as well as a suggestive protective effect for risk of bilateral oophorectomy (OR=0.93, 95%CI=0.86-0.99). Bioactive testosterone showed null effects for both traits (**Figure 7.1 (b) and Supplementary Table 7.2**). Higher genetically predicted SHBG also demonstrated significantly protective effects for both traits (OR=0.78, 95%CI=0.71-0.85 for leiomyoma and OR=0.85, 95%CI=0.78-0.93 for bilateral oophorectomy). As no association was observed with Free-T, this suggests a potential causal role for a direct effect of SHBG binding (independent of testosterone) in the aetiology of these conditions.

In men, Total-T showed null effects on all puberty-related traits while higher SHBG was significantly associated with later ages of both puberty traits as well as AFS (β=0.13, 95%CI=0.04-0.22). (**Figure 7.1 (c) and Supplementary Table 7.1**).



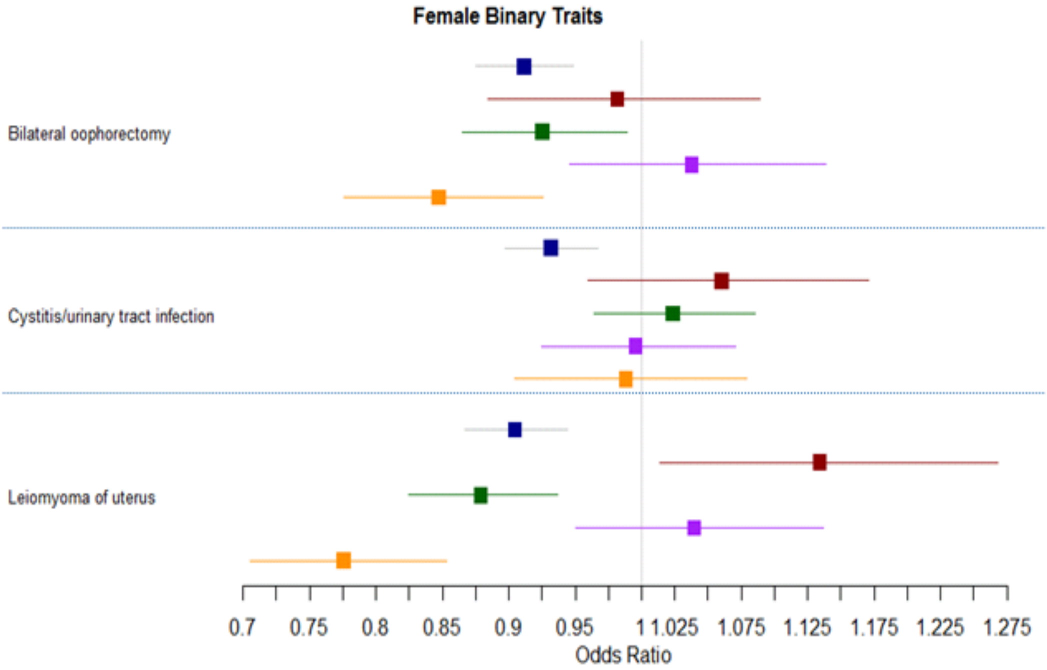


Figure 7.1: Forest plots for IVW MR effect estimates for AAM, BMI and sex hormones to female REPROWAS traits . Effect estimates for continuous traits are for years per year later AAM. Binary estimates are in odds ratios. Bars represent 95% confidence intervals.

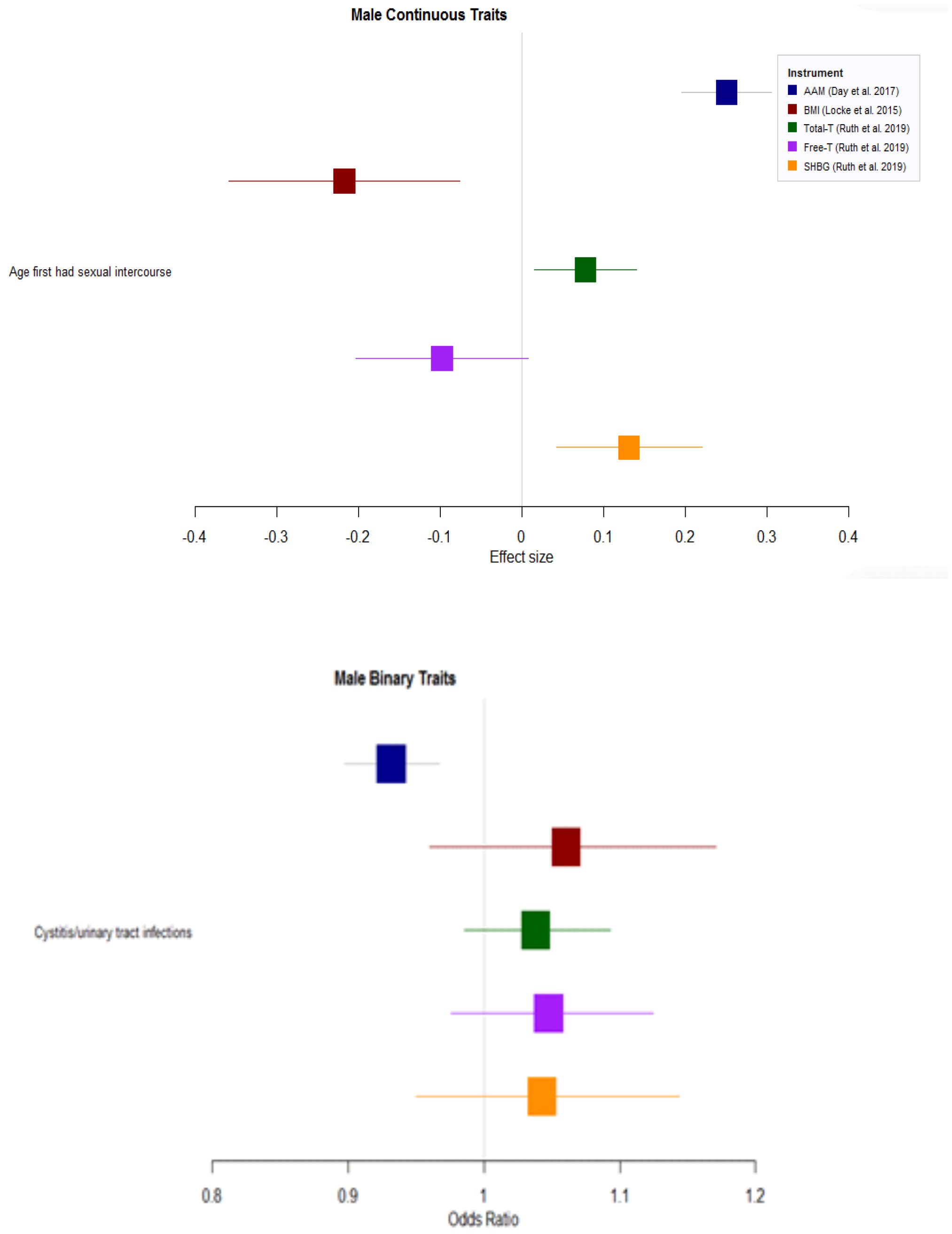


Figure 7.2: Forest plots for IVW MR effect estimates for AAM, BMI and sex hormones to male REPROWAS traits . Effect estimates for continuous traits are for years per year later AAM. Binary estimates are in odds ratios. Bars represent 95% confidence intervals.

7.3.4 Multivariate MR analyses

Multivariate MR analyses, in which genetically predicted effects on BMI and sex hormone traits were included as model covariates, were conducted to further test the mediating effects of these factors in the association between puberty timing and reproductive health outcomes. Among continuous traits, adjustment for genetically predicted BMI significantly attenuated the association estimate for age last using the contraceptive pill (from 0.23 to 0.1 years/year later puberty), which became statistically insignificant (P=0.06) (**Table 7.2**). Attenuation of effects was also observed for age starting HRT and age at last birth in women and AFS in both sexes, however these all remained significantly associated. Adjustment for BMI showed no effect on any of the binary traits in men or women. Similarly, adjustment for genetically predicted Total-T, SHBG and Free-T had no effect on any of the traits.

Table 7.2: Multivariate MR analyses for puberty-associated traits

	Unadjusted		Adj. BMI		Adj. Testosterone		Adj. Free-T		Adj. SHBG	
Trait	Beta (SE)	P-value	Beta (SE)	P-value	Beta (SE)	P-value	Beta (SE)	P-value	Beta (SE)	P-value
Female Continuous Traits										
Age last used oral contraceptive	0.23 (0.14, 0.32)	2.3×10^{-6}	0.10 (0.00, 0.21)	0.06	0.23 (0.13, 0.32)	3.4×10^{-6}	0.20 (0.10, 0.29)	4.6×10^{-5}	0.21 (0.11, 0.30)	2.5×10^{-5}
Age started HRT	0.19 (0.09, 0.29)	3.1×10^{-4}	0.15 (0.03, 0.26)	0.02	0.19 (0.09, 0.29)	2.2×10^{-4}	0.19 (0.09, 0.29)	3.5×10^{-4}	0.19 (0.08, 0.29)	4.5×10^{-4}
Age at menopause	0.18 (0.08, 0.27)	3.3×10^{-4}	0.22 (0.11, 0.33)	1.2×10^{-4}	0.18 (0.09, 0.28)	1.5×10^{-4}	0.20 (0.10, 0.30)	6.3×10^{-5}	0.19 (0.10, 0.29)	1.2×10^{-4}
Years since last cervical smear	-0.16 (-0.24, -0.08)	7.7×10^{-5}	-0.18 (-0.27, -0.09)	1.7×10^{-4}	-0.17 (-0.25, -0.09)	4.7×10^{-5}	-0.17 (-0.25, -0.09)	4.7×10^{-5}	-0.17 (-0.25, -0.09)	4.7×10^{-5}
Age at last birth	0.26 (0.18, 0.33)	4.5×10^{-11}	0.20 (0.11, 0.28)	9.3×10^{-6}	0.26 (0.19, 0.34)	4.3×10^{-11}	0.25 (0.18, 0.33)	2.5×10^{-10}	0.25 (0.18, 0.33)	3.1×10^{-10}
Age started oral contraceptive	0.11 (0.06, 0.16)	1.5×10^{-5}	0.11 (0.05, 0.16)	3.6×10^{-4}	0.11 (0.06, 0.16)	1.3×10^{-5}	0.11 (0.06, 0.16)	4.8×10^{-5}	0.10 (0.05, 0.15)	6.9×10^{-5}
Age at first sexual intercourse (females)	0.25 (0.20, 0.31)	$<2\times10^{-16}$	0.19 (0.13, 0.25)	4.7×10^{-9}	0.25 (0.20, 0.31)	$<2\times10^{-16}$	0.25 (0.19, 0.30)	$<2\times10^{-16}$	0.25 (0.19, 0.30)	$<2\times10^{-16}$
Female Binary Traits										
Bilateral oophorectomy	0.91 (0.87, 0.95)	4.3×10^{-6}	0.90 (0.86, 0.95)	3.6×10^{-5}	0.90 (0.87, 0.94)	1.7×10^{-6}	0.90 (0.87, 0.94)	1.8×10^{-6}	0.90 (0.87, 0.94)	2.0×10^{-6}
Cystitis/UTI (females)	0.93 (0.90, 0.97)	2.5×10^{-4}	0.93 (0.89, 0.97)	8.8×10^{-4}	0.93 (0.90, 0.97)	2.7×10^{-4}	0.93 (0.89, 0.96)	1.3×10^{-4}	0.92 (0.89, 0.96)	3.4×10^{-5}
Leiomyoma of uterus	0.90 (0.86, 0.94)	2.1×10^{-6}	0.91 (0.87, 0.96)	2.8×10^{-4}	0.90 (0.86, 0.94)	1.9×10^{-6}	0.90 (0.86, 0.94)	3.8×10^{-6}	0.90 (0.86, 0.94)	4.8×10^{-6}
Male Continuous Traits										
Age at first sexual intercourse (males)	0.25 (0.20, 0.31)	$<2\times10^{-16}$	0.19 (0.13, 0.25)	4.7×10^{-9}	0.25 (0.20, 0.31)	$<2\times10^{-16}$	0.25 (0.20, 0.31)	$<2\times10^{-16}$	0.25 (0.19, 0.30)	$<2\times10^{-16}$
Male Binary Traits										
Cystitis/UTI (males)	0.93 (0.90, 0.97)	2.5×10^{-4}	0.93 (0.89, 0.97)	8.8×10^{-4}	0.93 (0.90, 0.97)	2.2×10^{-4}	0.93 (0.90, 0.97)	2.9×10^{-4}	0.93 (0.89, 0.96)	1.4×10^{-4}

7.3.4 Comparison of AAM and voice breaking as instruments for puberty timing

As described above, for the primary analyses in this study it was decided to use AAM-associated variants to generate the instrument for puberty timing. This is based on the larger sample size for AAM compared to voice breaking and strong genetic correlation (r_g=0.74) between these phenotypes[35]. However, this has the potential to miss any sex-specific effects of puberty variants in men, which may have differential effects that are not adequately captured by an instrument composed solely of AAM variants. We therefore repeated the IVW MR analyses across all 181 REPROWAS phenotypes, using the 76 variants identified in the voice breaking GWAS meta-analysis described in Chapter 3.

Using the voice breaking instrument, we identified a total of 26 nominally significant associations with reproductive outcomes (**Supplementary Table 7.3**). Of the 37 traits which showed nominally significant associations with puberty based on the AAM-derived instrument, 15 also showed nominally significant associations when the voice breaking instrument was used. All 15 traits were directionally concordant with the AAM instrument. In addition, a further 11 associations were identified which showed null association with the AAM instrument. Of these, 8 were specific to females while prostate cancer was the only male-specific association. The presence of multiple female-specific traits may reflect the effects underlying pathways which are present in both sexes but which have greater influence on the exposure (that is, puberty timing) in males. One plausible example is testosterone level, which influences reproductive physiology in both sexes but is present in much higher concentrations in men. We therefore repeated the IVW MR analyses using the instrumental exposures for Total-T, Free-T and SHBG for all the traits associated with the male-specific score (**Figure 7.3**). Higher genetically predicted Total-T was significantly associated with increased risk for two traits: breast cancer (OR=1.13, 95%CI=1.05-1.22) and 'other disorders of the breast' (OR=1.25, 95%CI=1.03-1.50), a phenotype which includes fat necrosis and atrophy of breast tissue as well as galactorrhoea (milk discharge unrelated to breastfeeding). The effect in breast cancer was slightly higher when considering biologically active testosterone (Free-T; OR=1.23, 95%CI=1.11-1.37) but showed null effects on the 'other disorders of the breast' phenotype (OR=1.23, 95%CI=0.95-1.61). Higher SHBG was associated with increased risk for both false labour (commonly known as Braxton Hicks contractions; OR=1.72, 95%CI=1.36-2.36) and carcinomas of the cervix (OR=1.61, 95%CI=1.15-2.27). Neither of these traits showed any evidence for association with Total-T or Free-T, again suggesting a mechanism involving testosterone-independent effects of SHBG receptor binding. These results further highlight the importance of sex-specific analyses, as these important associations would not have been discovered otherwise.

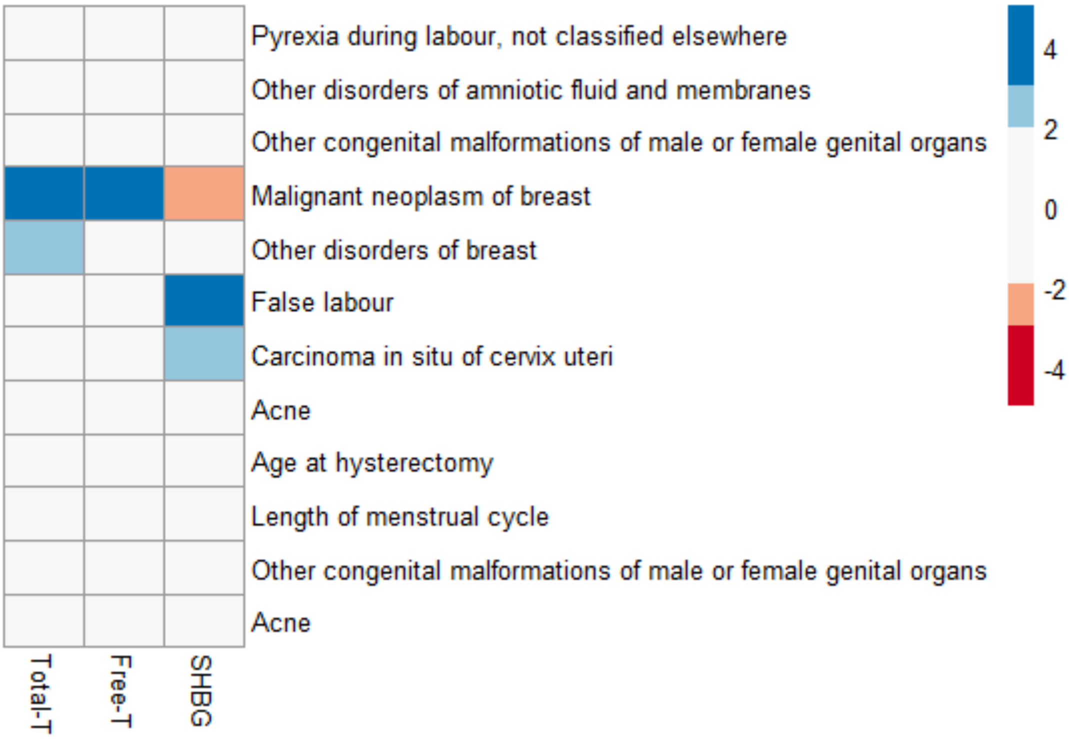


Figure 7.3: Associations with sex hormone traits for puberty-association phenotypes identified using voice breaking instrument. Association statistics represented by Z scores, with male and female-specific weights applied for sex-specific traits.

7.4 DISCUSSION

Puberty timing is an established risk factor for many reproductive health outcomes. However the relative importance of two of the hypothesised potential mediating mechanisms, BMI-related effects and increased endogenous sex-hormone exposure, is unknown for many traits. Here, I leverage the data obtained from our REPROWAS study to provide valuable insights into the relative importance of these two potential mediators on several important reproductive traits and health outcomes.

7.4.1 Casual effect of puberty timing on several reproductive traits

Mendelian randomisation analyses found evidence for association with 37 of the 181 traits analysed with at least nominal significance. The majority of the associations were female-specific , which is expected as these made up more than 80% of the total number of traits analysed. Of the 15 traits that remained associated after correction for multiple testing 12 were for continuous outcomes. This is due to greater powered to detect associations than in the binary ICD-10 coded disease traits, many of which had low case-control ratios. Many of these traits were related to timing of reproductive events and sexual behaviours which have had similar associations reported in previous studies. For example, earlier AFS and AFB have both been associated with puberty in large-scale genetic studies[271]. Similarly, age at menopause has repeatedly been shown to be associated with puberty timing[302]. This relationship has been suggested to be more U-shaped with both relatively earlier and relatively later puberty onset associated with earlier age at menarche[128,245], an association that would not be captured with the current study design. Of more interest were the three associations with disease traits: leiomyoma of the uterus, bilateral oophorectomy and cystitis/UTI. Both leiomyomas and oophorectomies have been associated with earlier puberty[39], but this is the first time to our knowledge that either has been assessed using an MR framework to infer causality. Other traits previously reported to be associated with puberty timing showed suggestive evidence of causal association here, including endometriosis, hysterectomy and PCOS[303]. These traits did not survive the multiple testing correction, but were all directionally consistent with observational reports in which later puberty reduces the risk of these conditions.

7.4.2 Effects of BMI and sex hormone levels on reproductive milestones and behaviours

BMI was inversely associated with several of the reproductive ageing and behaviour phenotypes. Given the well-established relationship between higher BMI and earlier puberty it is highly likely that much of this effect is acting through BMI-mediated pathways which influence puberty timing. However, univariate MRs cannot rule out a direct effect of BMI on these phenotypes which acts independently of the effect on puberty. The multivariate MRs, in which the effect of genetically predicted BMI was adjusted for, showed evidence of attenuation for several of the continuous phenotypes suggesting that many of the associations operate through BMI pathways which are independent of puberty.

Testosterone levels showed associations with several reproductive milestones in both men and women in univariate analyses. Most studies on the effects of sex hormones on reproductive health outcomes in women have naturally focused on oestrogen exposure given that this is the predominant mediator of sexual function in women. Thus, the influence of androgens on female reproductive traits remains understudied. Here, we observed that Free-T appeared to be a stronger predictor than Total-T for most traits, with higher Free-T causally associated with reduced ages of menarche, first sexual intercourse, and first and last birth. Previous studies on the effects of androgen levels on menarche have largely focused on menstrual disorders, for example higher testosterone in elite athletes or patients with PCOS[304,305]. However, to our knowledge the effect of testosterone concentrations on population-level variation in AAM has not previously been studied using genetic methods. This study therefore represents the first robust evidence that higher Free-T is associated with earlier AAM.

For AFS, previous analyses have suggested a role for oestrogen activity influencing this trait. This is based on a GWAS signal in the *ESR1* locus which appears to be independent of other variants in that region which are associated with other reproductive traits, suggesting a role for oestrogen activity on neuro-behavioural determinants of reproduction[271]. Here, we found that both higher Free-T and Total-T predicted lower AFS in women, suggesting a more general effect of sex hormones on sexual behaviours. Interestingly, in men we observed a nominally significant effect of Total-T on AFS which was directionally opposite to the effect in women, whereby higher Total-T predicted older AFS. This suggests the possibility of differential effects of sex hormones on behaviour in men and women, though further investigation will be required to confirm this.

7.4.3 BMI and sex hormones influence risk for reproductive disease

Earlier AAM was associated with increased risk for leiomyomas of the uterus, bilateral oophorectomy and cystitis/UTI. Leiomyomas have previously been associated with both BMI and endogenous sex hormones including androgens and oestrogens[306,307]. Here we demonstrate evidence for a nominally significant casual effect of higher BMI on increased uterine leiomyoma risk which is directionally consistent with previous studies. With regard to sex hormones, we report an inverse association between Total-T and leiomyoma risk. Previous studies have demonstrated that higher Total-T increases the risk of incident fibroids but reduces the risk of recurrence[307]. It is therefore possible that these effects are consistent, as HES record linkage only goes back to 1996 and incident cases may have been diagnosed previous to this. For bilateral oophorectomies, no association was found with BMI while significant inverse associations were observed for both Total-T and SHBG. Androgen levels have been associated with ovarian cancer, however oestrogens are a stronger predictor of risk[308]. It has thus been hypothesised that the association may be due, at least in part, to the fact that testosterone is an intermediate on the oestrogen synthesis pathway[309]. This is circumstantially supported by the null association observed for Free-T Furthermore in the multivariate analyses, all sex hormone variables had no effect on the disease traits when included as covariates. Without accurate measures of oestrogen levels it is impossible to decipher the true underlying biology of this relationship

7.4.5 Strengths, limitations and future directions

The main strength of this study was the systematic application of robust MR techniques to the expertly curated REPROWAS data, which allowed us to infer casual associations between puberty timing and many reproductive traits and health outcomes that had not previously been considered. Together with the expanded GWAS for sex hormones, this allows valuable insights into the mechanisms which underlie these associations.

This study also has several key limitations which must also be mentioned. Firstly, while the sex hormone GWAS represents by far the largest of its kind yet conducted, the measurements were obtained at a single time point and largely in individuals of middle- and older-age. A more accurate depiction of lifetime exposure to sex hormones many be obtained by taking measurements over several time points during the life course of participants, though this is perhaps an unrealistic for cohorts of this scale. Secondly, the lack of data on oestrogen levels in women was unfortunate as the majority of traits considered were specific to females. While valuable insights into the role played by androgen hormones in female reproductive health were generated, the full breadth of sex hormone exposure cannot be fully appreciated without the measurement of oestrogen level. Related to this point, the molecular pathways and intermediates products of sex hormone synthesis is complex, and while capturing the effects of the key endpoints of these pathways is a valuable starting point a more in-depth picture will likely be gained in future studies by considering a broader spectrum of sex hormones and their intermediates. Finally, while this study aimed to shed light on the relative importance of BMI and sex hormones as explanatory mechanisms for puberty-disease associations, the scale of the data made mediation analyses untenable. Careful consideration of individual phenotypes will be required to fully disentangle the effects of these mediators independent of each other.

7.4.6 Conclusion

Puberty timing is a key early life exposure and predisposes to many adverse health outcomes in later life. We used MR analyses to identify causal relationships between puberty timing and over 50 reproductive traits, and applied genetic instruments for BMI and various sex hormone traits while implicates both of these as potential mediators of the associations. Identifying the mechanisms which link puberty to later life health is an important part of understanding the aetiology and may inform intervention strategies aimed at mitigating the effects of earlier puberty.

Chapter 8: Concluding discussion

SUMMARY

In this thesis I have endeavoured to contribute to the collective understanding of the effect of early life exposures on health outcomes, with particular regard to traits related to reproductive health. To do so I decided to focus on two key early life exposures, birth weight and puberty timing. While both of these traits have often been considered in both genetic and observational studies, important questions remain regarding the aetiological pathways through which they exert their effects on health. In this concluding chapter I summarise the findings of this thesis, putting them into context of what was already known and the implications that they have for the field of research. I discuss some of the general limitations of the methods and data sources, and suggest potential future avenues that are now open as a result of the work presented here.

8.1 Summary of findings

8.1.1 The importance of sex-specific study designs in population genetics

Puberty represents perhaps the final period of development, culminating with the transition into adulthood. In 2005, the journal *Science* published a list of 125 "big questions which face inquiry over the next quarter century"[310]. Among these was the question of "what triggers puberty?" Much insight has been gained on this topic since, with the role of common genetic variation receiving considerable attention since the advent of GWAS. Despite this progress, the vast majority of this work has been completed in women. This highlights a fundamental issue in genetic epidemiology, where research in the age of large-scale biobanks must strike a balance between maximising statistical power on the one hand and maintaining the accuracy of the phenotype as a measure of the outcome on the other. Puberty is not a specific time point but rather a period of transition and physiological change which can take years to complete. While studies such as ALSPAC[77] have been able to characterise this in a (relatively) small cohort of individuals with longitudinal follow-up, this level of accuracy is not practical at the scale of study the size of UK Biobank where proxy measures must be used. AAM has traditionally been considered to be a more accurate marker than those used for men, due to the greater accuracy and reduced ambiguity of its recall. Because of the significant amount of genetic overlap that has been observed with male traits, it is often assumed that application of weights and scores derived from AAM represent a preferable option to a potentially noisier phenotype derived from male-specific traits. This may have merit, however the work presented in Chapter 3 of this thesis highlights the importance of sex-specificity with the identification of multiple puberty loci that appear to be specific to men. By and large the results of this male-specific study provide validation for what is already known or assumed, as downstream analyses using these results showed broadly similar sizes and directions of effect compared with AAM which confirm puberty timing as an important determinant of health and social outcomes. Yet there are differences that are captured as well. This is underscored in Chapter 7, where the application of male and female-specific genetic instruments for puberty identifies several distinct associations with reproductive outcomes. A more widespread application of sex-specific genetic discovery has been advocated, and examples of sexual dimorphism of genetic effects for many traits have been reported[311–313]. This is of heightened importance in the field of reproductive genetics, and the work presented here reinforces the importance of sex-specific approaches.

8.1.2 Mechanisms for the developmental origins of disease

One of the aims of this thesis was to investigate the potential mechanisms which may explain the associations between early life exposures and adult disease. In Chapters 4 and 5 I explored the roles of birth weight and DNA methylation in determining aspects of adiposity and its distribution throughout the body. Recent genetic studies have called into question fundamental elements of the Developmental Origins of Health and Disease (DOHaD) hypothesis, which is predicated on the idea of phenotypic plasticity during critical windows of development in response to environmental triggers. While a contribution of inherited factors is not incompatible with this concept, the potential for the pleiotropic effects of genes which influence both birth weight and the health outcome of interest must be taken into account when examining potential developmental causes[35,314,315]. In Chapter 4 I used multivariate MR analyses to demonstrate a causal effect of maternal genetically-predicted birth weight on body fat distribution in offspring. This approach theoretically reduces the probability that pleiotropic effects of the genes are influencing the results, as adjusting for the foetal effects implies that the maternal effects represent the developmental environment. The effects observed are in line with predictions of the DOHaD hypothesis, with lower genetically-predicted birth weight conferring increased central fat deposition and decreased peripheral fat – a metabolically unhealthy distribution which may explain the associations with poor cardio-metabolic health. This does not necessarily imply that these effects are entirely a representation of the intra-uterine environment as post-natal influences may also arise, and may further depend to some extent on the paternal genotype as well[316]. Overall, these results lend credence to the idea of developmental environment having an influence on adiposity and by extension the associated comorbidities. While this certainly does not invalidate a role for shared genetic effects, it does suggest that multiple mechanisms may be involved.

Following on from this, the role of epigenetic factors in the aetiology of reproductive and cardio-metabolic disease were then examined. Epigenetics are a logical mechanism for DOHaD as this is highly compatible with the underlying idea of phenotypic plasticity, which by definition describes differential health outcomes in individuals with similar genotypes. While a role for epigenetic mechanisms in DOHaD has been reasonably well-established, a fundamental question regarding the directionality of the association remains. Epigenetic studies investigating causality in the BMI-methylation relationship seem to have arrived at the overall consensus that BMI is primarily the cause and methylation the consequence[221]. However this is based on tenuous evidence. In Chapter 5 I re-analysed existing data as well as employing a novel method to impute methylation and simultaneously infer causality, and found widespread evidence for a causal relationship in the opposite direction in which methylation confers changes in BMI. This suggests a level of bi-directionality which has previously been largely dismissed, and which should have implications for future research on this subject.

The role of DNA methylation in the aetiology of puberty timing was also explored, with multiple casual loci identified. Of particular interest were the 9 loci which had not previously been identified in GWAS for puberty timing, which indicates that these loci are only associated with this trait through variable levels of methylation. This highlights the utility of performing genetic discovery across multiple platforms.

8.1.3 Genome-wide genetic determinants of reproductive health

In Chapter 6 I described a PheWAS study conducted for all reproductive health traits in UK Biobank, which were expertly curated to maximise sample size and preserve specificity of aetiology for each trait. This revealed evidence of widespread pleiotropy, with many SNPs, genes and regions associated with multiple reproductive health outcomes. PheWAS studies have become more common in recent years as the translation of GWAS results to therapeutic targets becomes of prioritised[255,317,318]. In this context, understanding the genetic links between diseases is of critical importance as this can help to reduce the likelihood of adverse unintended effects. This is highlighted here by the the *FSHB* gene, which is among the most pleiotropic loci discovered in our study. Drugs targeting FSHB have been proposed and developed for the treatment several conditions including obesity, cancer and infertility[274,319,320]. The multiple outcomes associated with this hormone, as highlighted here, may therefore have individual-level implications for the prescribing of these drugs, and potentially inform research for more targeted pharmacological approaches.

A further utility of the PheWAS framework is demonstrated in Chapter 7, which uses the results of the REPROWAS study to investigate the influence of puberty timing on reproductive health outcomes. Using MR analyses, I provided robust evidence for the effect of puberty timing in both men and women on multiple reproductive traits, and demonstrated a role for both BMI and sex hormones in the aetiology of many of these. While this study design did not allow for a definitive answer for the partitioning of the influence of these two potential explanatory factors, this does provide initial evidence and motivation for more detailed follow-up of specific associations of interest, and provides a framework with which to do so.

8.2 Limitations and caveats

Strengths and limitations specific to each study design have been discussed within those chapters. Here, I focus on some of the common limitations and considerations of genetic epidemiological studies and discuss how they may impact the findings of this thesis and the generalisability of the results.

8.2.1 Selection bias

For all population-based cohort studies, as were used in this thesis, the potential for a selection bias exists. In the UK Biobank study, which was designed to recruit a representative sample of he population on the UK, invited over 9 million participants in order to achieve their ultimate sample size of 500,000 (a response rate of 5.5%). A recent study which compared sociodemographic and health-related outcomes of UK Biobank participants compared to the general population found that study participants were generally healthier and from better socioeconomic positions. In addition, participants were more likely to be older, female and less likely to be smokers or heavy drinkers[321]. Similarly for the 23andMe cohort, participants are self-selected based on their ability and willingness to pay for personalised genetic testing. This is also likely to lead to a non-representative study sample, particularly with respect to those from lower socioeconomic backgrounds. This is indicative of a so-called 'healthy volunteer' effect, and while this does not affect the internal validity of these analyses, it must be taken into consideration if the findings are to be extended to the general population.

8.2.2 Self-reporting and ascertainment bias

Several of the traits considered in this thesis were based on self-reported data. This is commonly used in large cohorts and the limitations and potential for bias are well-documented. A particular consideration in this thesis is that the UK Biobank cohort, on which many of the chapters are based, was designed to capture the diseases of middle and older-age. Therefore using this cohort to examine exposures in early life (such as puberty), which for most participants will have occurred many decades earlier, means that recall bias may be more likely than in other studies.

8.2.3 Collider bias

Collider bias may occur in genetic discovery when the trait of interest is adjusted for a heritable covariate. The potential for collider bias was demonstrated by Day *et al.*, who showed that spurious (and biologically implausible) autosomal GWAS associations with sex arose when sex was adjusted for standing height[201]. In general, conditioning on a collider can induce spurious and directionally opposite associations between variants and phenotypes[322]. This may be of particular importance in the body composition GWAS, in which fat mass variables were adjusted for total body fat while lean mass was adjusted for height. Both of these covariates are highly heritable, allowing for the potential of collider bias. In these analyses, collider bias was formally assessed and accounted for by removal of variants which showed associations with the covariate but not with the outcome in the unadjusted model (as proposed by Aschard *et al.*[323]), thus reducing the potential of collider bias affecting the results.

8.2.4 Reverse causality

As discussed in previous chapters, Mendelian randomisation methods provide a means of making causal inferences regarding the association between an exposure and an outcome of interest. These methods are used extensively throughout this thesis, and the inferences made from these analyses are robust provided that the underlying assumptions are not violated (see Chapter 2). Sensitivity analyses including MR-Egger, MR-PRESSO and weighted-median MR have been used to test for the possibility of pleiotropy and the heterogeneous effects of SNPs influencing the conclusions. However, these methods are not able to fully mitigate the potential of reverse causality. While bi-directional MRs can provide evidence for or against a causal effect in the opposite direction, these were not feasible in many cases where genetic instruments were not available for a trait. Recently proposed extensions to the MR framework may be able to address this limitation. For example MR-Steiger[324], which employs a mediation-based approach to assign directionality to SNP-phenotype associations, should be used in future work to further test the causal effect of observed associations.

8.3 Implications and future directions

The findings from this thesis have built upon previous results, contributing to the understanding and discussion of the influence of early life exposures on later life health outcomes. Ultimately, however, the goal for all genetic epidemiological studies should be the translation of findings into clinical and public health benefits. After years of relatively slow progress, the potential of population-based genetic studies to transform clinical practice is beginning to be realised.

The utility of genetic data in clinical and public health settings are in two principle areas: 1) identifying targets for pharmaceutical intervention (i.e. drug discovery); and 2) risk stratification to identify groups and individuals at higher risk for poor health outcomes. For the exposures considered in this thesis, pharmaceutical interventions should not be considered a high priority. As such, the findings presented here are more likely to contribute towards the second of these goals. The use of polygenic risk scores in clinical medicine has frequently been advocated[325–327], with individuals at higher genetic risk informed of their increased susceptibility and possible risk-reduction strategies. Such an approach can easily be envisaged for puberty timing, where interventions to reduce obesity in childhood could be implemented. Similarly for birth weight, identifying mothers and children who are genetically predisposed to lower birth weight can inform pre- and post-natal care to reduce the potential for downstream health consequences. Identifying which individuals should be targeted, and when, must be carefully considered. Given the importance of early life exposures in determining health outcomes, the earlier genetic information can be obtained the better the chances for successful intervention. While higher risk individuals may be identified based on factors including clinical indicators, socio-demographics and family history, in the future routine genetic screening may bring the benefits of precision medicine to all individuals.

Several general avenues of future research may be recommended in order to increase the potential for clinical utility. UK Biobank has recently announced an initiative to complete whole genome sequencing (WGS) on the full cohort, providing an unparalleled resource of genome-wide coverage. As WGS becomes more regularly applied to larger genetic databases, fine-mapping and pinpointing of casual variants will become more feasible. Once available, these resources can be applied to all of the phenotypes discussed here, increasing the power for discovery and yielding new insights into biological processes. Additionally, a key limitation of population genetic studies to date is the lack of ethnic diversity in discovery samples. It has been suggested that this 'Eurocentric' bias may lead to an exacerbation of health inequalities which disadvantage non-white populations, as different allele frequencies and LD structure limits the generalisability of polygenic risk scores[328]. It is therefore imperative that efforts be made to make genetic studies more inclusive, with multi- and trans-ethnic studies prioritised in order to capture the full spectrum of human genetic diversity.

8.4 Conclusion

The work presented in this thesis expands our knowledge of the influence of early life exposures on general and reproductive health. The results provide evidence of the importance of developmental processes in influencing health throughout the life course, which may have implications for clinical interventions and public health strategies to combat the increasing burden of obesity of reproductive disease on our society.

REFERENCES

1. World Health Organization. Overweight and Obese. http://www.who.int/mediacentre/factsheets/fs311/en/ (2016).
2. PHE.
3. Bentham, J. *et al.* Worldwide trends in body-mass index, underweight, overweight, and obesity from 1975 to 2016: a pooled analysis of 2416 population-based measurement studies in 128·9 million children, adolescents, and adults. *Lancet* **390**, 2627–2642 (2017).
4. van Jaarsveld, C. H. M. & Gulliford, M. C. Childhood obesity trends from primary care electronic health records in England between 1994 and 2013: population-based cohort study. *Arch. Dis. Child.* **0**, 1–6 (2015).
5. Mayer-Davis, E. J. *et al.* Incidence Trends of Type 1 and Type 2 Diabetes among Youths, 2002–2012. *N. Engl. J. Med.* **376**, 1419–1429 (2017).
6. Zhou, B. *et al.* Worldwide trends in blood pressure from 1975 to 2015: a pooled analysis of 1479 population-based measurement studies with 19·1 million participants. *Lancet* **389**, 37–55 (2017).
7. Kelsey, M. M., Zaepfel, A., Bjornstad, P. & Nadeau, K. J. Age-related consequences of childhood obesity. *Gerontology* **60**, 222–228 (2014).
8. Kim, D. D. & Basu, A. Estimating the Medical Care Costs of Obesity in the United States: Systematic Review, Meta-Analysis, and Empirical Analysis. *Value Heal.* **19**, 602–613 (2016).
9. Statistics on Obesity, Physical Activity and Diet, England, 2019. https://digital.nhs.uk/data-and-information/publications/statistical/statistics-on-obesity-physical-activity-and-diet/statistics-on-obesity-physical-activity-and-diet-england-2019.
10. World Health Organization. Sexual and reproductive health. https://www.who.int/reproductivehealth/en/.
11. World Health Organization. https://www.who.int/hiv/data/en/.
12. Ding, T. *et al.* The prevalence of polycystic ovary syndrome in reproductiveaged women of different ethnicity: A systematic review and meta-analysis. *Oncotarget* **8**, 96351–96358 (2017).
13. Ding, T., Hardiman, P. J., Petersen, I. & Baio, G. Incidence and prevalence of diabetes and cost of illness analysis of polycystic ovary syndrome: A Bayesian modelling study. *Hum. Reprod.* **33**, 1299–1306 (2018).
14. Rogers, P. A. W. *et al.* Priorities for endometriosis research: Recommendations from an international consensus workshop. *Reprod. Sci.* **16**, 335–346 (2009).
15. Simoens, S. *et al.* The burden of endometriosis: Costs and quality of life of women with endometriosis and treated in referral centres. *Hum. Reprod.* **27**, 1292–1299 (2012).

16. CRUK Cancer Report.

17. Kershaw, E. E. & Flier, J. S. Adipose tissue as an endocrine organ. *J. Clin. Endocrinol. Metab.* **89**, 2548–2556 (2004).

18. Gesink Law, D. C., Maclehose, R. F. & Longnecker, M. P. Obesity and time to pregnancy. *Hum. Reprod.* **22**, 414–420 (2007).

19. Wise, L. A. *et al.* An internet-based prospective study of body size and time-to-pregnancy. *Hum. Reprod.* **25**, 253–264 (2010).

20. Beydoun, H. A. *et al.* Polycystic ovary syndrome, body mass index and outcomes of assisted reproductive technologies. *Reprod. Biomed. Online* **18**, 856–863 (2009).

21. Kakoly, N. S., Earnest, A., Moran, L. J., Teede, H. J. & Joham, A. E. Group-based developmental BMI trajectories, polycystic ovary syndrome, and gestational diabetes: A community-based longitudinal study. *BMC Med.* **15**, 1–9 (2017).

22. Liu, Y. & Zhang, W. Association between body mass index and endometriosis risk: A meta-analysis. *Oncotarget* **8**, 46928–46936 (2017).

23. Leitzmann, M. F. *et al.* Body mass index and risk of ovarian cancer. *Cancer* **115**, 812–822 (2009).

24. Poorolajal, J., Jenabi, E. & Masoumi, S. Z. Body mass index effects on risk of Ovarian cancer: A meta-analysis. *Asian Pacific J. Cancer Prev.* **15**, 7665–7671 (2014).

25. Kawachi, A. *et al.* Association of BMI and height with the risk of endometrial cancer, overall and by histological subtype: A population-based prospective cohort study in Japan. *Eur. J. Cancer Prev.* **28**, 196–202 (2019).

26. Jenabi, E. & Poorolajal, J. The effect of body mass index on endometrial cancer: A meta-analysis. *Public Health* **129**, 872–880 (2015).

27. Liu, K. *et al.* Association between body mass index and breast cancer risk: Evidence based on a dose–response meta-analysis. *Cancer Manag. Res.* **10**, 143–151 (2018).

28. Schoemaker, M. J. *et al.* Association of Body Mass Index and Age With Subsequent Breast Cancer Risk in Premenopausal Women. *JAMA Oncol.* **4**, e181771 (2018).

29. Dixon, S. C. *et al.* Adult body mass index and risk of ovarian cancer by subtype: A Mendelian randomization study. *Int. J. Epidemiol.* **45**, 884–895 (2016).

30. Visscher, P. M. *et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am. J. Hum. Genet.* **101**, 5–22 (2017).

31. 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.* **12**, 1–10 (2015).

32. Pickrell, J. K. *et al.* Detection and interpretation of shared genetic influences on 42 human traits. *Nat. Genet.* **48**, 709–717 (2016).

33. Hammerschlag, A. R., Leeuw, C. A. De & Benjamins, J. Affiliations : **5**, (2018).

34. Karlsson Linnér, R. *et al.* Genome-wide association analyses of risk tolerance and risky

behaviors in over 1 million individuals identify hundreds of loci and shared genetic influences. *Nat. Genet.* **51**, 245–257 (2019).

35. Day, F. R. *et al.* Genomic analyses identify hundreds of variants associated with age at menarche and support a role for puberty timing in cancer risk. *Nat. Genet.* **49**, 834–841 (2017).
36. Perry, J. R. B. *et al.* Parent-of-origin-specific allelic associations among 106 genomic loci for age at menarche. *Nature* **514**, 92–97 (2014).
37. Pharoah, P. *et al.* Commonly studied single-nucleotide polymorphisms and breast cancer: Results from the Breast Cancer Association Consortium. *J. Natl. Cancer Inst.* **98**, 1382–1396 (2006).
38. Wu, L. *et al.* Analysis of Over 140,000 European Descendants Identifies Genetically Predicted Blood Protein Biomarkers Associated with Prostate Cancer Risk. *Cancer Res.* **79**, 4592–4598 (2019).
39. Gayther, S. A. *et al.* Tagging single nucleotide polymorphisms in cell cycle control genes and susceptibility to invasive epithelial ovarian cancer. *Cancer Res.* **67**, 3027–3035 (2007).
40. Sapkota, Y. *et al.* Meta-analysis identifies five novel loci associated with endometriosis highlighting key genes involved in hormone metabolism. *Nat. Commun.* **8**, (2017).
41. Day, F. *et al.* Large-scale genome-wide meta-analysis of polycystic ovary syndrome suggests shared genetic architecture for different diagnosis criteria. *PLoS Genet.* **14**, 1–20 (2018).
42. Michailidou, K. *et al.* Association analysis identifies 65 new breast cancer risk loci. *Nature* **551**, 92–94 (2017).
43. Schumacher, F. R. *et al.* Association analyses of more than 140,000 men identify 63 new prostate cancer susceptibility loci. *Nat. Genet.* **50**, 928–936 (2018).
44. Fehringer, G. *et al.* Cross-cancer genome-wide analysis of lung, ovary, breast, prostate, and colorectal cancer reveals novel pleiotropic associations. *Cancer Res.* **76**, 5103–5114 (2016).
45. Smith, G. D. & Ebrahim, S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *Int. J. Epidemiol.* **32**, 1–22 (2003).
46. Barker, D. J. P. & Osmond, C. Infant Mortality, Childhood Nutrition, and Ischaemic Heart Disease in England and Wales. *Lancet* **327**, 1077–1081 (1986).
47. Barker, D. J. P. *et al.* Fetal nutrition and cardiovascular disease in adult life. *Lancet* **341**, 938–941 (1993).
48. Barker, D. J. P., Winter, P. D., Osmond, C., Margetts, B. & Simmonds, S. J. Weight in infacny and death from ischaemic heart disease. *The Lacet* **334**, 577–580 (1989).

49. Lumey, L. H. *et al.* Cohort profile: The Dutch Hunger Winter families study. *Int. J. Epidemiol.* **36**, 1196–1204 (2007).

50. Schulz, L. C. The Dutch hunger winter and the developmental origins of health and disease. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 16757–16758 (2010).

51. Ravelli, G.-P., Stein, Z. A. & Susser, M. W. Obesity in young men after famine exposure in utero and early infancy. *N. Engl. J. Med.* **295**, 349–53 (1976).

52. Stein, A. D., Zybert, P. A., Van Der Pal-De Bruin, K. & Lumey, L. H. Exposure to famine during gestation, size at birth, and blood pressure at age 59 y: Evidence from the dutch famine. *Eur. J. Epidemiol.* **21**, 759–765 (2006).

53. Martin-Gronert, M. S. & Ozanne, S. E. Mechanisms underlying the developmental origins of disease. *Rev. Endocr. Metab. Disord.* **13**, 85–92 (2012).

54. Wadhwa, P. D., Buss, C., Entrigner, S. & Swanson, J. M. Developmental Origins of Health and Disease: Brief History of the Approach and Current Focus on Epigenetic Mechanisms. *Semin. Reprod. Med.* **27**, 358–368 (2009).

55. Barker, D. J. P. The origins of the developmental origins theory. *J. Intern. Med.* **261**, 412–417 (2007).

56. Kuh, D., Ben-Shlomo, Y., Lynch, J., Hallqvist, J. & Power, C. Life course epidemiology. *J. Epidemiol. Community Health* **57**, 778–883 (2003).

57. Goyal, D., Limesand, S. W. & Goyal, R. Epigenetic responses and the developmental origins of health and disease. *J. Endocrinol.* **242**, T105–T119 (2019).

58. Soubry, A. Epigenetics as a Driver of Developmental Origins of Health and Disease: Did We Forget the Fathers? *BioEssays* **40**, 1–10 (2018).

59. Bianco-Miotto, T., Craig, J. M., Gasser, Y. P., Van Dijk, S. J. & Ozanne, S. E. Epigenetics and DOHaD: From basics to birth and beyond. *J. Dev. Orig. Health Dis.* **8**, 513–519 (2017).

60. Janssen, P. A. *et al.* Standards for the measurement of birth weight, length and head circumference at term in neonates of European, Chinese and South Asian ancestry. *Open Med.* **1**, e74-88 (2007).

61. Norris, T. *et al.* Updated birth weight centiles for England and Wales. *Arch. Dis. Child. Fetal Neonatal Ed.* **103**, F577–F582 (2018).

62. Abu-Saad, K. & Fraser, D. Maternal nutrition and birth outcomes. *Epidemiol. Rev.* **32**, 5–25 (2010).

63. Abubakari, A. & Jahn, A. Maternal dietary patterns and practices and birth weight in Northern Ghana. *PLoS One* **11**, 1–17 (2016).

64. Abubakari, A., Kynast-Wolf, G. & Jahn, A. Maternal determinants of birth weight in Northern Ghana. *PLoS One* **10**, 1–15 (2015).

65. Ahmed, S., Hassen, K. & Wakayo, T. A health facility based case-control study on determinants of low birth weight in Dassie town, Northeast Ethiopia: The role of

nutritional factors. *Nutr. J.* **17**, 1–10 (2018).

66. Tersigni, C. *et al.* Celiac disease and reproductive disorders: Meta-analysis of epidemiologic associations and potential pathogenic mechanisms. *Hum. Reprod. Update* **20**, 582–593 (2014).

67. Bernstein, I. M., Mongeon, J. A., Badger, G. J. & Solomon, L. Maternal Smoking and Its Association With Birth Weight. *Obstet. Gynecol.* **106**, 986–991 (2005).

68. Witt, W. P. *et al.* Maternal stressful life events prior to conception and the impact on infant birth weight in the United States. *Am. J. Public Health* **104**, 81–89 (2014).

69. Clausson, B., Lichtenstein, P. & Cnattingius, S. Genetic influence on birthweight and gestational length determined by studies in offspring of twins. *BJOG An Int. J. Obstet. Gynaecol.* **107**, 375–381 (2000).

70. Mook-Kanamori, D. O. *et al.* Heritability estimates of body size in fetal life and early childhood. *PLoS One* **7**, (2012).

71. Jahanfar, S. Birth weight and anthropometric measurements of twins. *Ann. Hum. Biol.* **45**, 395–400 (2018).

72. Warrington, Nicole M., Beaumont, Robin N., Horikoshi, Momoko, Day, Felix R., Helgeland, Oyvind, Laurin, C. Maternal and fetal genetic effects on birth weight and their relevance to cardio-metabolic risk factors. *Nat. Genet.* **51**, 804–814 (2019).

73. Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747–753 (2009).

74. Mendle, J., Beltz, A. M., Carter, R. & Dorn, L. D. Understanding Puberty and Its Measurement: Ideas for Research in a New Generation. *J. Res. Adolesc.* **29**, 82–95 (2019).

75. Marshall, W. A. & Tanner, J. M. Variations in Pattern of Pubertal Changes in Girls. (1969).

76. Marshall, W. A. & Tanner, J. M. Variations in the Pattern of Pubertal Changes in Boys. (1970).

77. 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.* **42**, 111–127 (2013).

78. Day, F. R. *et al.* Shared genetic aetiology of puberty timing between sexes and with health-related outcomes. *Nat. Commun.* **6**, 8842 (2015).

79. Herman-Giddens, M. E. *et al.* Secondary sexual characteristics in boys: Data from the pediatric research in office settings network. *Pediatrics* **130**, (2012).

80. Biro, F. M. *et al.* Pubertal assessment method and baseline characteristics in a mixed longitudinal study of girls. *Pediatrics* **126**, (2010).

81. Morris, D. H., Jones, M. E., Schoemaker, M. J., Ashworth, A. & Swerdlow, A. J. Secular trends in age at menarche in women in the UK born 1908-93: Results from the breakthrough generations study. *Paediatr. Perinat. Epidemiol.* **25**, 394–400 (2011).

82. Cabanes, A. *et al.* Decline in age at menarche among Spanish women born from 1925 to

1962. *BMC Public Health* **9**, 1–7 (2009).

83. Euling, S. Y. *et al.* Examination of US puberty-timing data from 1940 to 1994 for secular trends: Panel findings. *Pediatrics* **121**, (2008).

84. Silva, H. P. & Padez, C. Secular trends in age at Menarche among Caboclo populations from Pará, Amazonia, Brazil: 1930-1980. *Am. J. Hum. Biol.* **18**, 83–92 (2006).

85. Ahmed, M. L., Ong, K. K. & Dunger, D. B. Childhood obesity and the timing of puberty. *Trends Endocrinol. Metab.* **20**, 237–242 (2009).

86. Mumby, H. S. *et al.* Mendelian randomisation study of childhood BMI and early menarche. *J. Obes.* **2011**, (2011).

87. Day, F. R., Elks, C. E., Murray, A., Ong, K. K. & Perry, J. R. Puberty timing associated with diabetes, cardiovascular disease and also diverse health outcomes in men and women: the UK Biobank study. *Sci. Rep.* **5**, 11208 (2015).

88. Day, F. R., Perry, J. R. B. & Ong, K. K. Genetic Regulation of Puberty Timing in Humans. *Neuroendocrinology* **102**, 247–255 (2015).

89. Consequences, E. *et al.* Genetic consequences of social stratification in Great Britain. *bioRxiv* 457515 (2018).

90. UK Biobank Coordinating Centre. UK Biobank: Protocol for a large-scale prospective epidemiological resource UK Biobank Coordinating Centre Stockport. *UKBB-PROT-09-06 (Main Phase)* **06**, 1–112 (2007).

91. Elliott, P. & Peakman, T. C. The UK Biobank sample handling and storage protocol for the collection, processing and archiving of human blood and urine. *Int. J. Epidemiol.* **37**, 234–244 (2008).

92. Welsh, S., Peakman, T., Sheard, S. & Almond, R. Comparison of DNA quantification methodology used in the DNA extraction protocol for the UK Biobank cohort. *BMC Genomics* **18**, 1–7 (2017).

93. 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.* **3**, 769–781 (2015).

94. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).

95. Connell, J. O. *et al.* Europe PMC Funders Group Haplotype estimation for biobank scale datasets. **48**, 817–820 (2016).

96. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

97. McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279–1283 (2016).

98. Walter, K. *et al.* The UK10K project identifies rare variants in health and disease. *Nature* **526**, 82–90 (2015).

99. Bycroft, C. *et al.* Genome-wide genetic data on ~500,000 UK Biobank participants. *bioRxiv* 166298 (2017) doi:10.1101/166298.

100. Rolfe, E. *et al.* Association between birth weight and visceral fat in adults. *Am. J. Clin. Nutr.* **92**, 347–352 (2010).

101. Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* **5**, (2009).

102. Riboli, E. *et al.* European Prospective Investigation into Cancer and Nutrition (EPIC): study populations and data collection. *Public Health Nutr.* **5**, 1113–1124 (2002).

103. Day, N. *et al.* EPIC-Norfolk: Study design and characteristics of the cohort. *Br. J. Cancer* **80**, 95–103 (1999).

104. Hayat, S. A. *et al.* Cohort profile: A prospective cohort study of objective physical and cognitive capability and visual health in an ageing population of men and women in Norfolk (EPIC-Norfolk 3). *Int. J. Epidemiol.* **43**, 1063–1072 (2014).

105. Eriksson, N. *et al.* Web-based, participant-driven studies yield novel genetic associations for common traits. *PLoS Genet.* **6**, 1–20 (2010).

106. Li, Q. S., Tian, C., Seabrook, G. R., Drevets, W. C. & Narayan, V. A. Analysis of 23andMe antidepressant efficacy survey data: implication of circadian rhythm and neuroplasticity in bupropion response. *Transl. Psychiatry* **6**, e889 (2016).

107. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).

108. Loh, P. R. *et al.* Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.* **47**, 284–290 (2015).

109. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: Fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).

110. Machiela, M. J. & Chanock, S. J. LDlink: A web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. *Bioinformatics* **31**, 3555–3557 (2015).

111. Pruim, R. J. *et al.* LocusZoom: Regional visualization of genome-wide association scan results. *Bioinformatics* **27**, 2336–2337 (2011).

112. Serge, A. V. *et al.* Common inherited variation in mitochondrial genes is not enriched for associations with type 2 diabetes or related glycemic traits. *PLoS Genet.* **6**, e1001058 (2010).

113. Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet* **advance on**, 291–295 (2015).

114. 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* **33**, 272–279 (2017).

115. Gibbs, R. A. *et al.* The International HapMap Project. *Nature* **426**, 789–796 (2003).

116. Katan M.B. Apolipoprotein E isoforms, serum cholesterol, and cancer. *Lancet* **1**, 507–508 (1986).

117. Pierce, B. L. & Burgess, S. Efficient design for mendelian randomization studies: Subsample and 2-sample instrumental variable estimators. *Am. J. Epidemiol.* **178**, 1177–1184 (2013).

118. Bowden, J. *et al.* A framework for the investigation of pleiotropy in two-sample summary data Mendelian randomization. *Stat. Med.* **36**, 1783–1802 (2017).

119. Bowden, J., Smith, G. D. & Burgess, S. Mendelian randomization with invalid instruments: Effect estimation and bias detection through Egger regression. *Int. J. Epidemiol.* **44**, 512–525 (2015).

120. Burgess, S. & Thompson, S. G. Erratum to: Interpreting findings from Mendelian randomization using the MR-Egger method (Eur J Epidemiol, 10.1007/s10654-017-0255-x). *Eur. J. Epidemiol.* **32**, 391–392 (2017).

121. Bowden, J., Davey Smith, G., Haycock, P. C. & Burgess, S. Consistent Estimation in Mendelian Randomization with Some Invalid Instruments Using a Weighted Median Estimator. *Genet. Epidemiol.* **40**, 304–314 (2016).

122. Verbanck, M., Chen, C.-Y., Neale, B. & Do, R. Widespread pleiotropy confounds causal relationships between complex traits and diseases inferred from Mendelian randomization. *bioRxiv* 157552 (2017) doi:10.1101/157552.

123. Burgess, S. *et al.* Dissecting Causal Pathways Using Mendelian Randomization with Summarized Genetic Data: Application to Age at Menarche and Risk of Breast Cancer. *Genetics* genetics.300191.2017 (2017) doi:10.1534/genetics.117.300191.

124. Burgess, S. & Thompson, S. G. Multivariable Mendelian randomization: The use of pleiotropic genetic variants to estimate causal effects. *Am. J. Epidemiol.* **181**, 251–260 (2015).

125. Harries, M. L. L., Walker, J. M., Williams, D. M., Hawkins, S. & Hughes, I. A. Changes in the male voice at puberty. 445–447 (1997).

126. Hillman, J. B. & Biro, F. M. HHS Public Access. **47**, 322–323 (2017).

127. Vizmanos, B., Martí-Henneberg, C., Clivillé, R., Moreno, A. & Fernández-Ballart, J. Age of pubertal onset affects the intensity and duration of pubertal growth peak but not final height. *Am. J. Hum. Biol.* **13**, 409–416 (2001).

128. Perry, J. R. B., Murray, A., Day, F. R. & Ong, K. K. Molecular insights into the aetiology of female reproductive ageing. *Nat. Rev. Endocrinol.* **11**, 725–734 (2015).

129. Villamor, E. & Jansen, E. C. Nutritional Determinants of the Timing of Puberty. *Annu. Rev.*

Public Health **37**, 33–46 (2016).

130. Buck Louis, G. M. *et al.* Environmental factors and puberty timing: Expert panel research needs. *Pediatrics* **121**, (2008).
131. Sørensen, K., Mouritsen, A., Aksglaede, L. & Hagen, C. P. HOR MON E RE SE ARCH I N Recent Secular Trends in Pubertal Timing : Implications for Evaluation and Diagnosis of Precocious Puberty. 137–145 (2012) doi:10.1159/000336325.
132. Özen, S. Pubertal Development. **3**, 1–6 (2011).
133. How Does Childhood Socioeconomic Hardship Affect Reproductive Strategy ? Pathways of Development. **363**, 356–363 (2016).
134. Karaolis-danckert, N., Kroke, A., Remer, T. & Buyken, A. E. Dietary Protein Intake throughout Childhood Is Associated with the Timing of Puberty 1 – 3. 565–571 (2010) doi:10.3945/jn.109.114934.Methods.
135. Lee, J. M. *et al.* Timing of Puberty in Overweight Versus Obese Boys. *Pediatrics* **137**, e20150164–e20150164 (2016).
136. Cousminer, D. L. *et al.* Genome-wide association study of sexual maturation in males and females highlights a role for body mass and menarche loci in male puberty. *Hum. Mol. Genet.* **23**, 4452–4464 (2014).
137. Wagner, I. V. *et al.* Effects of obesity on human sexual development. *Nat. Rev. Endocrinol.* **8**, 246–254 (2012).
138. Hou, H. *et al.* Gene expression profiling of puberty-associated genes reveals abundant tissue and sex-specific changes across postnatal development. *Hum. Mol. Genet.* **26**, 3585–3599 (2017).
139. Lunetta, K. L. *et al.* Rare coding variants and X-linked loci associated with age at menarche. *Nat. Commun.* **6**, 1–8 (2015).
140. Turley, P. *et al.* Multi-trait analysis of genome-wide association summary statistics using MTAG. *Nat. Genet.* 1–9 (2018) doi:10.1038/s41588-017-0009-4.
141. Frysz, M., Howe, L. D., Tobias, J. H. & Paternoster, L. Using SITAR (SuperImposition by Translation and Rotation) to estimate age at peak height velocity in Avon Longitudinal Study of Parents and Children. *Wellcome Open Res.* **3**, 90 (2018).
142. Consortium, T. Gte. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).
143. Schumacher, F. R. *et al.* Association analyses of more than 140,000 men identify 63 new prostate cancer susceptibility loci. *Nat. Genet.* 1 (2018) doi:10.1038/s41588-018-0142-8.
144. Timmers, P. R. H. J. *et al.* Genomics of 1 million parent lifespans implicates novel pathways and common diseases and distinguishes survival chances. *Elife* **8**, 1–40 (2019).
145. Hysi, P. G. *et al.* Genome-wide association meta-analysis of individuals of European ancestry identifies new loci explaining a substantial fraction of hair color variation and

heritability. *Nat. Genet.* **50**, (2018).

146. Kayserili, H. *et al.* ALX4 dysfunction disrupts craniofacial and epidermal development. *Hum. Mol. Genet.* **18**, 4357–4366 (2009).

147. Hu Lei Yeh, Shuyuan Cui, Yun Li, Xin Chang, Hong-Chiang Jin, Jie Chang, Chawnshang, S. L. Infiltrating T cells promote prostate cancer metastasis via modulation of FGF11→ miRNA-541→ androgen receptor (AR)→ MMP9 signaling. *Mol Oncol* **9**, 44–57 (2015).

148. Jackstadt, R. *et al.* AP4 is a mediator of epithelial–mesenchymal transition and metastasis in colorectal cancer. *J. Exp. Med.* **210**, 1331–1350 (2013).

149. Simons, N. *et al.* A common gene variant in glucokinase regulatory protein interacts with glucose metabolism on diabetic dyslipidemia: The combined CODAM and Hoorn studies. *Diabetes Care* **39**, 1811–1817 (2016).

150. Donnelly, M. P. *et al.* A global view of the OCA2-HERC2 region and pigmentation. *Hum. Genet.* **131**, 683–696 (2012).

151. Praetorius, C. *et al.* XA polymorphism in IRF4 affects human pigmentation through a tyrosinase-dependent MITF/TFAP2A pathway. *Cell* **155**, 1022 (2013).

152. Jacobs, L. C. *et al.* A genome-wide association study identifies the skin color genes IRF4, MC1R, ASIP, and BNC2 influencing facial pigmented spots. *J. Invest. Dermatol.* **135**, 1735–1742 (2015).

153. Sitek, A., Zadzińska, E., Rosset, I. & Antoszewski, B. Is increased constitutive skin and hair pigmentation an early sign of puberty? *HOMO- J. Comp. Hum. Biol.* **64**, 205–214 (2013).

154. Kukla-Bartoszek, M. *et al.* Investigating the impact of age-depended hair colour darkening during childhood on DNA-based hair colour prediction with the HIrisPlex system. *Forensic Sci. Int. Genet.* **36**, 26–33 (2018).

155. Locke, A. E. *et al.* Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518**, 197–206 (2015).

156. Slominski, A., Tobin, D.J., Shibahara, S., Wortsman, J. Melanin Pigmentation in Mammalian Skin and Its Hormonal Regulation. *Physiol. Rev.* **84**, 1155–1228 (2004).

157. Amoah, D. *et al.* Birth Weight was longitudinally associated with Cardiometabolic Risk Markers in Mid-Adulthood. *Ann. Epidemiol.* **26**, 643–647 (2016).

158. Jornayvaz, F. R. *et al.* Low birth weight leads to obesity, diabetes and increased leptin levels in adults: the CoLaus study. *Cardiovasc. Diabetol.* **15**, 1–10 (2016).

159. Gennser, G., Rymark, P. E. R. & Isberg, P. E. R. E. Low birth weight and risk of high blood in adulthood. *Br. Med. J.* **296**, 1498–1500 (1988).

160. Launer, L. J., Hofmnan, A. & Grobbee, D. E. Relation between birth weight and blood pressure: longitudinal study of infants and children. *Br. Med. J.* **307**, 1451–1454 (1993).

161. Mi, D., Fang, H., Zhao, Y. & Zhong, L. Birth weight and type 2 diabetes: A meta-analysis. *Exp. Ther. Med.* **14**, 5313–5320 (2017).

162. Hernandez, C. D. *et al.* Association between abdominal fat distribution, adipocytokines and metabolic alterations in obese low-birth-weight children. *Pediatr. Obes.* **11**, 285–291 (2015).

163. Ibáñez, L. *et al.* Early Development of Visceral Fat Excess after Spontaneous Catch-Up Growth in Children with Low Birth Weight. *J. Clin. Endocrinol. Metab.* **93**, 925–928 (2008).

164. Johnsson, I. W., Haglund, B., Ahlsson, F. & Gustafsson, J. A high birth weight is associated with increased risk of type 2 diabetes and obesity. *Pediatr. Obes.* **10**, 77–83 (2014).

165. Skilton, M. R. *et al.* High birth weight is associated with obesity and increased carotid wall thickness in young adults: The cardiovascular risk in young finns study. *Arterioscler. Thromb. Vasc. Biol.* **34**, 1064–1068 (2014).

166. Barker, D. J. P. The fetal and infant origins of adult disease. *Br. Med. J.* **301**, 1111 (1990).

167. Barker DJ, Hales CN, Fall CF, Osmond C, Phipps K, C. P. Type 2 (non-insulin-dependent) diabetes mellitus, hypertension and hyperlipidaemia (syndrome X): relation to reduced fetal growth. *Diabetologia* **36**, 62–67 (1993).

168. Horikoshi, M. *et al.* Genome-wide associations for birth weight and correlations with adult disease. *Nature* **538**, 248–252 (2016).

169. Jelenkovic, A. *et al.* Association between birthweight and later body mass index: an individual-based pooled analysis of 27 twin cohorts participating in the CODATwins project. *Int. J. Epidemiol.* 1488–1498 (2017) doi:10.1093/ije/dyx031.

170. Rasmussen F, J. M. The relation of weight, length and ponderal index at birth to body mass index and overweight among 18-year-old males in Sweden. *Eur. J. Epidemiol.* **14**, 373–380 (1998).

171. Kelly, L. A. *et al.* Birth Weight and Body Composition in Overweight Latino Youth: A Longitudinal Analysis. *Obesity* **16**, 2524–2528 (2008).

172. Yuan, Z. *et al.* Possible role of birth weight on general and central obesity in Chinese children and adolescents: a cross-sectional study. *Ann. Epidemiol.* **25**, 748–752 (2015).

173. te Velde, S. J., Twisk, J. W. R., van Mechelen, W. & Kemper, H. C. G. Birth Weight , Adult Body Composition , and Subcutaneous Fat Distribution. *Obes. Res.* **11**, 202–208 (2003).

174. Samuel Klein, David Allison, Steven Heymsfield, David Kelley, Rudolph Leibel, Cathy Nonas, R. K. Waist Circumference and Cardiometabolic. *Diabetes Care* **30**, 1647–1652 (2007).

175. Despre, J. *et al.* Abdominal Obesity and the Metabolic Syndrome: Contribution to Global Cardiometabolic Risk. *Atheroscler. Thromb. Vasc. Biol.* **28**, 1039–1049 (2008).

176. Fonseca, M. J., Severo, M., Correia, S. & Santos, A. C. Effect of birth weight and weight change during the first 96 h of life on childhood body composition — path analysis. *Int. J. Obes.* **39**, 579–585 (2015).

177. Lindberg, J. *et al.* Overweight, Obesity, and Body Composition in 3.5- and 7-Year-Old

Swedish Children Born with Marginally Low Birth Weight. *J. Pediatr.* **167**, 1246–1252 (2015).

178. Rillamas-Sun, E., Sowers, M. R., Harlow, S. D. & Randolph Jr., J. F. The Relationship of Birth Weight With Longitudinal Changes in Body Composition in Adult Women. *Obesity* **20**, 463–465 (2012).

179. Kensara, O. A. *et al.* Fetal programming of body composition: relation between birth weight and body composition measured with dual-energy X-ray absorptiometry and anthropometric methods in older Englishmen. *Am. J. Clin. Nutr.* **82**, 980–987 (2005).

180. Barker, M., Robinson, S., Osmond, C. & Barker, D. J. P. Birth weight and body fat distribution in adolescent girls. *Arch. Dis. Child.* **77**, 381–383 (1997).

181. Pereira-Freire, J. A., Lemos, J. O., de Sousa, A. F., Meneses, C. C. & Rondó, P. H. C. Association between weight at birth and body composition in childhood: A Brazilian cohort study. *Early Hum. Dev.* **91**, 445–449 (2015).

182. Labayen, I. *et al.* Small Birth Weight and Later Body Composition and Fat Distribution in Adolescents : The AVENA Study. *Obesity* **16**, 1680–1686 (2008).

183. De Koning, L., Merchant, A. T., Pogue, J. & Anand, S. S. Waist circumference and waist-to-hip ratio as predictors of cardiovascular events: Meta-regression analysis of prospective studies. *Eur. Heart J.* **28**, 850–856 (2007).

184. Fox, C. S. *et al.* Abdominal visceral and subcutaneous adipose tissue compartments: Association with metabolic risk factors in the framingham heart study. *Circulation* **116**, 39–48 (2007).

185. Vega, G. L. *et al.* Influence of body fat content and distribution on variation in metabolic risk. *J. Clin. Endocrinol. Metab.* **91**, 4459–4466 (2006).

186. Araújo de França, G. V *et al.* Associations of birth weight, linear growth and relative weight gain throughout life with abdominal fat depots in adulthood: the 1982 Pelotas (Brazil) birth cohort study. *Int. J. Obes.* **40**, 14–21 (2016).

187. Sokolovic, N., Kuriyan, R., Kurpad, A. V. & Thomas, T. Sleep and birthweight predict visceral adiposity in overweight/obese children. *Pediatr. Obes.* **8**, 41–44 (2013).

188. Albanese, C. V, Diessel, E. & Genant, H. K. Clinical Applications of Body Composition Measurements Using DXA. *J. Clin. Densitom.* **6**, 75–85 (2003).

189. Borga, M. *et al.* Advanced body composition assessment: from body mass index to body composition profiling. *J. Investig. Med.* **66**, 887–895 (2018).

190. Stults-Kolehmainen, M. A. *et al.* DXA estimates of fat in abdominal, trunk and hip regions varies by ethnicity in men. *Nutr. Diabetes* **3**, 4–9 (2013).

191. Miller, K. L. *et al.* Multimodal population brain imaging in the UK Biobank prospective epidemiological study. *Nat. Neurosci.* **19**, 1523–1535 (2016).

192. Winkler T W, Day F R, Croteau-Chonka D C, Wood A R, Locke A E, Magi R, Ferreira T, Fall

T, G. M. Quality control and conduct of genome-wide association meta-analyses. *Nat. Protoc.* **9**, 1192–1212 (2014).

193. Blackmore, D. G. *et al.* Growth hormone responsive neural precursor cells reside within the adult mammalian brain. *Sci. Rep.* **2**, 1–10 (2012).

194. 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.* **8**, 1–19 (2017).

195. Ng, M. C. Y. *et al.* Meta-Analysis of Genome-Wide Association Studies in African Americans Provides Insights into the Genetic Architecture of Type 2 Diabetes. *PLoS Genet.* **10**, (2014).

196. Shi, Y. *et al.* Genome-wide association study identifies eight new risk loci for polycystic ovary syndrome. *Nat. Genet.* **44**, 1020–1025 (2012).

197. Szklarczyk, D. *et al.* STRING v11: protein–protein association networks with increased coverage , supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **47**, 607–613 (2019).

198. Warrington, N. M. *et al.* Maternal and fetal genetic effects on birth weight and their relevance to cardio-metabolic risk factors. *Nat. Genet.* **51**, 804–814 (2019).

199. Hattersley, A. T. & Tooke, J. E. The fetal insulin hypothesis: An alternative explanation of the association of low birthweight with diabetes and vascular disease. *Lancet* **353**, 1789–1792 (1999).

200. Loos, R. J. F., Beunen, G., Fagard, R., Derom, C. & Vlietinck, R. Birth weight and body composition in young adult men — a prospective twin study. *Int. J. Obes.* **25**, 1537–1545 (2001).

201. Day, F. R., Loh, P. R., Scott, R. A., Ong, K. K. & Perry, J. R. B. A Robust Example of Collider Bias in a Genetic Association Study. *Am. J. Hum. Genet.* **98**, 392–393 (2016).

202. Smith, Z. D. & Meissner, A. DNA methylation: roles in mammalian development. *Nat. Rev. Genet.* **14**, 204–220 (2013).

203. Mohn, F. & Schubeler, D. Genetics and epigenetics: stability and plasticity during cellular differentiation. *Trends Genet.* **25**, 129–136 (2009).

204. Gendrel, A.-V. & Heard, E. Noncoding RNAs and Epigenetic Mechanisms During X-Chromosome Inactivation. *Annu. Rev. Cell Dev. Biol.* **30**, 561–580 (2014).

205. Adalsteinsson, B. T. & Ferguson-Smith, A. C. Epigenetic Control of the Genome—Lessons from Genomic Imprinting. *Genes (Basel).* **5**, 635–655 (2014).

206. Nikolova, Y. S. & Hariri, A. R. Can we observe epigenetic effects on human brain function? *Trends Cogn. Sci.* **19**, 366–373 (2015).

207. Probst, A. V., Dunleavy, E. & Almouzni, G. Epigenetic inheritance during the cell cycle. *Nat. Rev. Mol. Cell Biol.* **10**, 192–206 (2009).

208. Chan, S. W.-L. *et al.* RNA Silencing Genes Control de Novo DNA Methylation. *Science (80-.).*

303, 1336–1336 (2004).

209. Huang, T. *et al.* Meta-analyses of gene methylation and smoking behavior in non-small cell lung cancer patients. *Sci. Rep.* **5**, 1–8 (2015).

210. Lim, U. & Song, M. Chapter 23 Dietary and Lifestyle Factors of DNA Methylation. **863**, 359–376 (2012).

211. Richmond, R. C. *et al.* Prenatal exposure to maternal smoking and offspring DNA methylation across the lifecourse: findings from the Avon Longitudinal Study of Parents and Children (ALSPAC). *Human* **24**, 2201–17 (2015).

212. Koukoura, O., Sifakis, S. & Spandidos, D. A. DNA methylation in the human placenta and fetal growth (Review). *Mol. Med. Rep.* **5**, 883–889 (2012).

213. Horvath, S. DNA methylation age of human tissues and cell types DNA methylation age of human tissues and cell types. *Genome Biol.* **14**, R115 (2013).

214. Weisenberger, D. J. Characterizing DNA methylation alterations from The Cancer Genome Atlas. *J. Clin. Invest.* **124**, 17–23 (2014).

215. Li, Y., Gorelik, G., Strickland, F. M. & Richardson, B. C. Oxidative Stress, T Cell DNA Methylation, and Lupus. *Arthritis Rheumatol.* **66**, 1574–82 (2014).

216. Brasa, S. *et al.* Reciprocal changes in DNA methylation and hydroxymethylation and a broad repressive epigenetic switch characterize FMR1 transcriptional silencing in fragile X syndrome. *Clin. Epigenetics* **8**, 15 (2016).

217. Waterland, R. A. & Michels, K. B. Epigenetic Epidemiology of the Developmental Origins Hypothesis. *Annu. Rev. Nutr.* **27**, 363–88 (2007).

218. Tobi, E. W. *et al.* DNA methylation signatures link prenatal famine exposure to growth and metabolism. *Nature* **5**, 1–13 (2014).

219. Feinberg, A. P. *et al.* Personalized Epigenomic Signatures That Are Stable Over Time and Covary with Body Mass Index. *Sci. Transl. Med.* **2**, 49ra67 (2010).

220. Dick, K. J. *et al.* DNA methylation and body-mass index : a genome-wide analysis. *Lancet* **383**, 1990–1998 (2014).

221. Wahl, S. *et al.* Epigenome-wide association study of body mass index, and the adverse outcomes of adiposity. *Nature* **541**, 81–86 (2016).

222. Mendelson, M. M. *et al.* Association of Body Mass Index with DNA Methylation and Gene Expression in Blood Cells and Relations to Cardiometabolic Disease : A Mendelian Randomization Approach. 1–30 (2017) doi:10.1371/journal.pmed.1002215.

223. Li, E., Beard, C. & Jaenisch, R. Role for DNA methylation in genomic imprinting. *Trends Genet.* **10**, 78 (1994).

224. Richmond, R. C. *et al.* DNA Methylation and BMI : Investigating Identified Methylation Sites at HIF3A in a Causal Framework. **65**, 1231–1244 (2016).

225. Thompson, E. E. *et al.* Global DNA methylation changes spanning puberty are near

predicted estrogen- responsive genes and enriched for genes involved in endocrine and immune processes. 1–10 (2018).

226. Bonder, M. J. *et al.* Disease variants alter transcription factor levels and methylation of their binding sites. *Nat. Genet.* **49**, 131–138 (2017).

227. Gusev, A. *et al.* Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.* **48**, 245–52 (2016).

228. Mancuso, N. *et al.* Integrating Gene Expression with Summary Association Statistics to Identify Genes Associated with 30 Complex Traits. *Am. J. Hum. Genet.* **100**, 473–487 (2017).

229. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: A tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).

230. Larder, R. *et al.* Obesity-associated gene TMEM18 has a role in the central control of appetite and body weight regulation. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 9421–9426 (2017).

231. Marinov, G. K. *et al.* The landscape of genomic imprinting across diverse adult human tissues. *Genome Res.* **25**, 927–936 (2015).

232. Lomniczi, A. *et al.* Epigenetic control of female puberty. *Nat. Publ. Gr.* **16**, 281–289 (2013).

233. Patterns, M. *et al.* Genetics & Epigenetics. 51–62 (2013) doi:10.4137/GEG.S12897.

234. Yuan, X. *et al.* Genome-wide DNA methylation analysis of pituitaries during the initiation of puberty in gilts. 1–16 (2019).

235. He, F. *et al.* Association between DNA methylation in obesity-related genes and body mass index percentile in adolescents. *Sci. Rep.* 1–8 (2019) doi:10.1038/s41598-019-38587-7.

236. Dmitrijeva, M., Ossowski, S., Serrano, L. & Schaefer, M. H. Tissue-specific DNA methylation loss during ageing and carcinogenesis is linked to chromosome structure, replication timing and cell division rates. *Nucleic Acids Res.* **46**, 7022–7039 (2018).

237. Slieker, R. C., Relton, C. L., Gaunt, T. R., Slagboom, P. E. & Heijmans, B. T. Age-related DNA methylation changes are tissue-specific with ELOVL2 promoter methylation as exception. *Epigenetics and Chromatin* **11**, 1–11 (2018).

238. Lokk, K. *et al.* DNA methylome profiling of human tissues identifies global and tissue-specific methylation patterns. *Genome Biol.* **15**, (2014).

239. Battle, A. False positives in trans-eQTL and co-expression analyses arising from RNA-sequencing alignment errors [version 2 ; peer review : 3 approved] Ashis Saha. 1–28 (2019).

240. Moran, S., Arribas, C. & Esteller, M. Validation of a DNA methylation microarray for 850,000 CpG sites of the human genome enriched in enhancer sequences. *Epigenomics* **8**, 389–399 (2016).

241. Mimoto, M. S., Nadal, A. & Sargis, R. M. Polluted Pathways: Mechanisms of Metabolic Disruption by Endocrine Disrupting Chemicals. *Curr. Environ. Heal. reports* **4**, 208–222

(2017).

242. Heindel, J. J. *et al.* Metabolism Disrupting Chemicals and Metabolic Disorders. *Reprod. Toxicol.* **68**, 3–33 (2017).

243. Ferreira, M. A. *et al.* Genome-wide association and transcriptome studies identify target genes and risk loci for breast cancer. *Nat. Commun.* **10**, 1–18 (2019).

244. Eeles, R. A. *et al.* Identification of 23 new prostate cancer susceptibility loci using the iCOGS custom genotyping array. *Nat. Genet.* **45**, 385–391 (2013).

245. Perry, J. R. B. *et al.* Parent-of-origin-specific allelic associations among 106 genomic loci for age at menarche. *Nature* **514**, 92–97 (2014).

246. Nyholt, D. R. *et al.* Genome-wide association meta-analysis identifies new endometriosis risk loci. *Nat. Genet.* **44**, 1355–1359 (2012).

247. Albertsen, H. M., Chettier, R., Farrington, P. & Ward, K. Genome-Wide Association Study Link Novel Loci to Endometriosis. *PLoS One* **8**, (2013).

248. Gudmundsson, J. *et al.* Genome-wide associations for benign prostatic hyperplasia reveal a genetic correlation with serum levels of PSA. *Nat. Commun.* **9**, 1–8 (2018).

249. Hellwege, J. N. *et al.* Heritability and genome-wide association study of benign prostatic hyperplasia (BPH) in the eMERGE network. *Sci. Rep.* **9**, 1–10 (2019).

250. Zhang, Q., Greenbaum, J., Zhang, W. D., Sun, C. Q. & Deng, H. W. Age at menarche and osteoporosis: A Mendelian randomization study. *Bone* **117**, 91–97 (2018).

251. Sequeira, M. E., Lewis, S. J., Bonilla, C., Smith, G. D. & Joinson, C. Association of timing of menarche with depressive symptoms and depression in adolescence: Mendelian randomisation study. *Br. J. Psychiatry* **210**, 39–46 (2017).

252. Gill, D. *et al.* Age at menarche and lung function: a Mendelian randomization study. *Eur. J. Epidemiol.* **32**, 701–710 (2017).

253. Gill, D. *et al.* Age at Menarche and Time Spent in Education: A Mendelian Randomization Study. *Behav. Genet.* **47**, 480–485 (2017).

254. Verma, A. *et al.* PheWAS and Beyond: The Landscape of Associations with Medical Diagnoses and Clinical Measures across 38 , 662 Individuals from Geisinger. *Am. J. Hum. Genet.* **102**, 592–608 (2018).

255. Diogo, D. *et al.* Phenome-wide association studies across large population cohorts support drug target validation. *Nat. Commun.* **9**, 1–13 (2018).

256. Hebbring, S. J. The challenges, advantages and future of phenome-wide association studies. *Immunology* **141**, 157–165 (2014).

257. Millard, L. A. C. *et al.* MR-PheWAS: Hypothesis prioritization among potential causal effects of body mass index on many outcomes, using Mendelian randomization. *Sci. Rep.* **5**, 1–17 (2015).

258. Barzel, B. & Barabási, A. L. Response to letter of correspondence - Bastiaens et al. *Nat.*

Biotechnol. **33**, 339–342 (2015).

259. Khosravi, A., Jayaram, B., Goliaei, B. & Masoudi-Nejad, A. Active repurposing of drug candidates for melanoma based on GWAS, PheWAS and a wide range of omics data. *Mol. Med.* **25**, 1–11 (2019).

260. Robinson, J. R., Wei, W.-Q., Roden, D. M. & Denny, J. C. Defining Phenotypes from Clinical Data to Drive Genomic Research. *Annu. Rev. Biomed. Data Sci.* **1**, 69–92 (2018).

261. Cortes, A. *et al.* Bayesian analysis of genetic association across tree- structured routine healthcare data in the UK Biobank. *Nat. Genet.* **49**, 1311–1318 (2017).

262. Millard, L. A. C., Davies, N. M., Gaunt, T. R., Smith, G. D. & Tilling, K. Software application profile: PHESANT: A tool for performing automated phenome scans in UK Biobank. *Int. J. Epidemiol.* **47**, 29–35 (2018).

263. Cizkova, A. *et al.* TMEM70 mutations cause isolated ATP synthase deficiency and neonatal mitochondrial encephalocardiomyopathy. *Nat. Genet.* **40**, 1288–1290 (2008).

264. Matalliotaki, C. *et al.* Role of FN1 and GREB1 gene polymorphisms in endometriosis. *Mol. Med. Rep.* **20**, 111–116 (2019).

265. Anderson, S. L. *et al.* Nemaline myopathy in the Ashkenazi Jewish population is caused by a deletion in the nebulin gene. *Hum. Genet.* **115**, 185–190 (2004).

266. Zhou, W. *et al.* Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat. Genet.* **50**, 1335–1341 (2018).

267. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**, D1005–D1012 (2019).

268. Klein, N. A. *et al.* Reproductive aging: Accelerated ovarian follicular development associated with a monotropic follicle-stimulating hormone rise in normal older women. *J. Clin. Endocrinol. Metab.* **81**, 1038–1045 (1996).

269. McTavish, K. J. *et al.* Rising follicle-stimulating hormone levels with age accelerate female reproductive failure. *Endocrinology* **148**, 4432–4439 (2007).

270. Lei, J. T., Gou, X., Seker, S. & Ellis, M. J. ESR1 alterations and metastasis in estrogen receptor positive breast cancer . *J. Cancer Metastasis Treat.* **2019**, (2019).

271. Day, F. R. *et al.* Physical and neuro-behavioural determinants of reproductive onset and success. *Nat. Genet.* **48**, 617–623 (2017).

272. Evans, L. M. *et al.* Comparison of methods that use whole genome data to estimate the heritability and genetic architecture of complex traits. *Nat. Genet.* **50**, 737–745 (2018).

273. Day, F. R., Perry, J. R. B. & Ong, K. K. Genetic Regulation of Puberty Timing in Humans. *Neuroendocrinology* **102**, 247–255 (2015).

274. Conforti, A. *et al.* Pharmacogenetics of FSH action in the female. *Front. Endocrinol. (Lausanne).* **10**, (2019).

275. Eriksson, N. *et al.* Genetic variants associated with breast size also influence breast cancer risk. *BMC Med. Genet.* **13**, (2012).

276. Lindström, S. *et al.* Genome-wide association study identifies multiple loci associated with both mammographic density and breast cancer risk. *Nat. Commun.* **5**, 1–7 (2014).

277. Dunning, A. M. *et al.* Association of ESR1 gene tagging SNPs with breast cancer risk. *Hum. Mol. Genet.* **18**, 1131–1139 (2009).

278. Tan, J. K. L. & Bhate, K. A global perspective on the epidemiology of acne. *Br. J. Dermatol.* **172**, 3–12 (2015).

279. Kaplowitz, P. B. Link between body fat and the timing of puberty. *Pediatrics* **121**, (2008).

280. Li, W. *et al.* Association between obesity and puberty timing: A systematic review and meta-analysis. *Int. J. Environ. Res. Public Health* **14**, (2017).

281. Sutcliffe, S. & Colditz, G. A. Prostate cancer: Is it time to expand the research focus to early-life exposures? *Nat. Rev. Cancer* **13**, 208–218 (2013).

282. Salonia, A. *et al.* Circulating sex steroids and prostate cancer: Introducing the time-dependency theory. *World J. Urol.* **31**, 267–273 (2013).

283. Chavez-MacGregor, M. *et al.* Lifetime cumulative number of menstrual cycles and serum sex hormone levels in postmenopausal women. *Breast Cancer Res. Treat.* **108**, 101–112 (2008).

284. DeVane, G. W., Czekala, N. M., Judd, H. L. & Yen, S. S. C. Circulating gonadotropins, estrogens, and androgens in polycystic ovarian disease. *Am. J. Obstet. Gynecol.* **121**, 496–500 (1975).

285. Ding, E. L., Song, Y., Malik, V. S. & Liu, S. Sex Differences of Endogenous Sex Hormones and Risk of Type 2 Diabetes. *Jama* **295**, 1288 (2006).

286. Kim, C. & Halter, J. B. Endogenous sex hormones, metabolic syndrome, and diabetes in men and women. *Curr. Cardiol. Rep.* **16**, 1–20 (2014).

287. Zhao, D. *et al.* Endogenous Sex Hormones and Incident Cardiovascular Disease in Post-Menopausal Women. *J. Am. Coll. Cardiol.* **71**, 2555–2566 (2018).

288. Tyanova, S. *et al.* Proteomic maps of breast cancer subtypes. *Nat. Commun.* **7**, 1–11 (2016).

289. Ho, S. M. Estrogen, progesterone and epithelial ovarian cancer. *Reprod. Biol. Endocrinol.* **1**, 1–8 (2003).

290. Nelles, J. L., Hu, W. Y. & Prins, G. S. Estrogen action and prostate cancer. *Expert Rev. Endocrinol. Metab.* **6**, 437–451 (2011).

291. Watts, E. L. *et al.* Low Free Testosterone and Prostate Cancer Risk: A Collaborative Analysis of 20 Prospective Studies. *Eur. Urol.* **74**, 585–594 (2018).

292. Klap, J., Schmid, M. & Loughlin, K. R. The relationship between total testosterone levels and prostate cancer: A review of the continuing controversy. *J. Urol.* **193**, 403–414

(2015).

293. Roddam, A. W., Allen, N. E., Appleby, P. & Key, T. J. Endogenous sex hormones and prostate cancer: A collaborative analysis of 18 prospective studies. *J. Natl. Cancer Inst.* **100**, 170–183 (2008).

294. Bogaert, V. *et al.* Heritability of blood concentrations of sex-steroids in relation to body composition in young adult male siblings. *Clin. Endocrinol. (Oxf).* **69**, 129–135 (2008).

295. Travison, T. G. *et al.* The heritability of circulating testosterone, oestradiol, oestrone and sex hormone binding globulin concentrations in men: The Framingham Heart Study. *Clin. Endocrinol. (Oxf).* **80**, 277–282 (2014).

296. Ohlsson, C. *et al.* Genetic determinants of serum testosterone concentrations in men. *PLoS Genet.* **7**, 1–11 (2011).

297. Eriksson, A. L. *et al.* Genetic determinants of circulating estrogen levels and evidence of a causal effect of estradiol on bone density in men. *J. Clin. Endocrinol. Metab.* **103**, 991–1004 (2018).

298. Ruth, K. S. *et al.* Genome-wide association study with 1000 genomes imputation identifies signals for nine sex hormone-related phenotypes. *Eur. J. Hum. Genet.* **24**, 284–290 (2016).

299. Coviello, A. D. *et al.* Circulating testosterone and SHBG concentrations are heritable in women: The Framingham Heart Study. *J. Clin. Endocrinol. Metab.* **96**, 1491–1495 (2011).

300. Coviello, A. D. *et al.* A genome-wide association meta-analysis of circulating sex hormone-binding globulin reveals multiple loci implicated in sex steroid hormone regulation. *PLoS Genet.* **8**, (2012).

301. Vermeulen, A., Verdonck, L. & Kaufman, J. M. A critical evaluation of simple methods for the estimation of free testosterone in serum. *J. Clin. Endocrinol. Metab.* **84**, 3666–3672 (1999).

302. Forman, M., Mangini, L., Thelus-Jean, R. & Hayward. Life-course origins of the ages at menarche and menopause. *Adolesc. Health. Med. Ther.* **4**, 1–21 (2013).

303. Nnoaham, K. E., Webster, P., Kumbang, J., Kennedy, S. H. & Zondervan, K. T. Is early age at menarche a risk factor for endometriosis? A systematic review and meta-analysis of case-control studies. *Fertil. Steril.* **98**, 702-712.e6 (2012).

304. Lagowska, K. & Kapczuk, K. Testosterone concentrations in female athletes and ballet dancers with menstrual disorders. *Eur. J. Sport Sci.* **16**, 490–497 (2016).

305. Lerchbaum, E., Schwetz, V., Rabe, T., Giuliani, A. & Obermayer-Pietsch, B. Hyperandrogenemia in polycystic ovary syndrome: Exploration of the role of free testosterone and androstenedione in metabolic phenotype. *PLoS One* **9**, (2014).

306. Reis, F. M., Bloise, E. & Ortiga-Carvalho, T. M. Hormones and pathogenesis of uterine fibroids. *Best Pract. Res. Clin. Obstet. Gynaecol.* **34**, 13–24 (2016).

307. Wong, J. Y. Y., Gold, E. B., Johnson, W. O. & Lee, J. S. Circulating sex hormones and risk of

uterine fibroids: Study of women's health across the nation (swan). *J. Clin. Endocrinol. Metab.* **101**, 123–130 (2016).

308. Zhu, H. *et al.* The role of the androgen receptor in ovarian cancer carcinogenesis and its clinical implications. *Oncotarget* **8**, 29395–29405 (2017).

309. Ose, J. *et al.* Androgens are differentially associated with ovarian cancer subtypes in the Ovarian Cancer Cohort Consortium. *Cancer Res.* **77**, 3951–3960 (2017).

310. Science 125th Anniversary Edition. *Science (80-.).*

311. Traglia, M. *et al.* Genetic mechanisms leading to sex differences across common diseases and anthropometric traits. *Genetics* **205**, 979–992 (2017).

312. Khramtsova, E. A., Davis, L. K. & Stranger, B. E. The role of sex in the genomics of human complex traits. *Nat. Rev. Genet.* **20**, 173–190 (2019).

313. Zeng, Y. *et al.* Sex Differences in Genetic Associations With Longevity. *JAMA Netw. Open* **1**, e181670 (2018).

314. Shields, B. M. *et al.* Mutations in the Glucokinase Gene of the Fetus Result in Reduced Placental Weight. *Diabetes Care* **31**, 753–757 (2008).

315. Nyirenda, M. J. & Byass, P. Pregnancy, programming, and predisposition. *Lancet Glob. Heal.* **7**, e404–e405 (2019).

316. Sharp, G. C., Lawlor, D. A. & Richardson, S. S. It's the mother!: How assumptions about the causal primacy of maternal effects influence research on the developmental origins of health and disease. *Soc. Sci. Med.* **213**, 20–27 (2018).

317. Gao, M., Quan, Y., Zhou, X. H. & Zhang, H. Y. PheWAS-based systems genetics methods for anti-breast cancer drug discovery. *Genes (Basel).* **10**, (2019).

318. Robinson, J. R., Denny, J. C., Roden, D. M. & Van Driest, S. L. Genome-wide and Phenome-wide Approaches to Understand Variable Drug Actions in Electronic Health Records. *Clin. Transl. Sci.* **11**, 112–122 (2018).

319. Modi, D. A. *et al.* Targeting of follicle stimulating hormone peptide-conjugated dendrimers to ovarian cancer cells. *Nanoscale* **6**, 2812–2820 (2014).

320. Thangeswaran, P. *et al.* HHS Public Access. **546**, 107–112 (2018).

321. Fry, A. *et al.* Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants with Those of the General Population. *Am. J. Epidemiol.* **186**, 1026–1034 (2017).

322. Cole, S. R. *et al.* Illustrating bias due to conditioning on a collider. *Int. J. Epidemiol.* **39**, 417–420 (2010).

323. Aschard, H., Vilhjálmsson, B. J., Joshi, A. D., Price, A. L. & Kraft, P. Adjusting for heritable covariates can bias effect estimates in genome-wide association studies. *Am. J. Hum. Genet.* **96**, 329–339 (2015).

324. Hemani, G., Tilling, K. & Smith, G. D. Orienting the causal relationship between imprecisely

measured traits using genetic instruments. *bioRxiv* 117101 (2017) doi:10.1101/117101.

325. Lambert, S. A., Abraham, G. & Inouye, M. Towards clinical utility of polygenic risk scores. *Hum. Mol. Genet.* **28**, R133–R142 (2019).

326. Khera, A. V. *et al.* Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat. Genet.* **50**, 1219–1224 (2018).

327. Torkamani, A., Wineinger, N. E. & Topol, E. J. The personal and clinical utility of polygenic risk scores. *Nat. Rev. Genet.* **19**, 581–590 (2018).

328. Martin, A. R. *et al.* Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* **51**, 584–591 (2019).

SUPPLEMENTARY TABLES AND FIGURES

Supplementary Table 4.1: KEGG pathways from STRING analysis showing enrichment for genes in the two primary clusters of birth weight variants

Pathway	Observed gene count (Cluster 1)	Observed gene count (Cluster 2)	Background gene count	FDR (Cluster 1)	FDR (Cluster 2)	Assigned Cluster
Endocytosis	9	-	242	3.50E-04	-	1
Cell cycle	6	-	123	1.20E-03	-	1
Hedgehog signaling pathway	4	-	46	1.60E-03	-	1
Insulin signaling pathway	6	-	134	1.60E-03	-	1
Neomycin, kanamycin and gentamicin biosynthesis	2	-	5	3.60E-03	-	1
B cell receptor signaling pathway	4	-	71	5.30E-03	-	1
Aldosterone-regulated sodium reabsorption	3	-	37	7.20E-03	-	1
Carbohydrate digestion and absorption	3	-	42	9.80E-03	-	1
Endocrine and other factor-regulated calcium reabsorption	3	-	47	1.26E-02	-	1
Viral myocarditis	3	-	56	1.92E-02	-	1
Antigen processing and presentation	3	-	66	2.75E-02	-	1
Phagosome	4	-	145	3.95E-02	-	1
Non-alcoholic fatty liver disease (NAFLD)	4	-	149	4.26E-02	-	1
Oxytocin signaling pathway	4	-	149	4.26E-02	-	1
Galactose metabolism	2	-	31	4.48E-02	-	1
Fc gamma R-mediated phagocytosis	3	-	89	4.77E-02	-	1
Fructose and mannose metabolism	2	-	33	4.77E-02	-	1
Starch and sucrose metabolism	2	-	33	4.77E-02	-	1
Allograft rejection	2	-	35	4.99E-02	-	1
Cortisol synthesis and secretion	6	-	63	-	3.20E-04	2
Long-term depression	5	-	60	-	1.70E-03	2
Chagas disease (American trypanosomiasis)	6	-	101	-	1.80E-03	2
Renin secretion	5	-	63	-	1.80E-03	2

Supplementary Table 4.1 (Continued)

Pathway	Observed gene count (Cluster 1)	Observed gene count (Cluster 2)	Background gene count	FDR (Cluster 1)	FDR (Cluster 2)	Assigned Cluster
Th17 cell differentiation	6	-	102	-	1.90E-03	2
TNF signaling pathway	6	-	108	-	2.10E-03	2
Leukocyte transendothelial migration	6	-	112	-	2.40E-03	2
Cocaine addiction	4	-	49	-	4.40E-03	2
Aldosterone synthesis and secretion	5	-	93	-	5.50E-03	2
Phosphatidylinositol signaling system	5	-	97	-	6.10E-03	2
Melanogenesis	5	-	98	-	6.20E-03	2
Glucagon signaling pathway	5	-	100	-	6.40E-03	2
Toll-like receptor signaling pathway	5	-	102	-	6.80E-03	2
Inflammatory bowel disease (IBD)	4	-	62	-	7.50E-03	2
Antifolate resistance	3	-	31	-	9.00E-03	2
Leishmaniasis	4	-	70	-	1.02E-02	2
Sphingolipid signaling pathway	5	-	116	-	1.02E-02	2
Vascular smooth muscle contraction	5	-	119	-	1.08E-02	2
Thyroid hormone synthesis	4	-	73	-	1.10E-02	2
Graft-versus-host disease	3	-	36	-	1.15E-02	2
Calcium signaling pathway	6	-	179	-	1.20E-02	2
Natural killer cell mediated cytotoxicity	5	-	124	-	1.20E-02	2
RNA degradation	4	-	77	-	1.20E-02	2
Thyroid cancer	3	-	37	-	1.20E-02	2
Dopaminergic synapse	5	-	128	-	1.27E-02	2
Hepatitis C	5	-	131	-	1.38E-02	2
Fluid shear stress and atherosclerosis	5	-	133	-	1.45E-02	2
Salivary secretion	4	-	86	-	1.58E-02	2

Supplementary Table 4.1 (Continued)

Pathway	Observed gene count (Cluster 1)	Observed gene count (Cluster 2)	Background gene count	FDR (Cluster 1)	FDR (Cluster 2)	Assigned Cluster
Th1 and Th2 cell differentiation	4	-	88	-	1.67E-02	2
Morphine addiction	4	-	91	-	1.77E-02	2
Amoebiasis	4	-	94	-	1.93E-02	2
Pancreatic secretion	4	-	95	-	1.98E-02	2
NOD-like receptor signaling pathway	5	-	166	-	2.89E-02	2
Serotonergic synapse	4	-	112	-	3.20E-02	2
Tuberculosis	5	-	172	-	3.20E-02	2
Purine metabolism	5	-	173	-	3.23E-02	2
Oocyte meiosis	4	-	116	-	3.48E-02	2
Renin-angiotensin system	2	-	23	-	3.52E-02	2
Amphetamine addiction	3	-	65	-	3.63E-02	2
Herpes simplex infection	5	-	181	-	3.67E-02	2
Epithelial cell signaling in Helicobacter pylori infection	3	-	66	-	3.70E-02	2
Renal cell carcinoma	3	-	68	-	3.92E-02	2
Bile secretion	3	-	71	-	4.34E-02	2
Inositol phosphate metabolism	3	-	73	-	4.57E-02	2
Hippo signaling pathway - multiple species	2	-	28	-	4.60E-02	2
Pertussis	3	-	74	-	4.65E-02	2
Pathways in cancer	19	22	515	3.06E-07	1.26E-07	Both
Endocrine resistance	9	8	95	2.37E-06	5.98E-05	Both
FoxO signaling pathway	10	9	130	2.37E-06	5.98E-05	Both
Breast cancer	10	10	147	3.64E-06	3.11E-05	Both
Proteoglycans in cancer	11	10	195	3.86E-06	1.40E-04	Both

Supplementary Table 4.1 (Continued)

Pathway	Observed gene count (Cluster 1)	Observed gene count (Cluster 2)	Background gene count	FDR (Cluster 1)	FDR (Cluster 2)	Assigned Cluster
Longevity regulating pathway - multiple species	7	5	61	9.11E-06	1.70E-03	Both
Prostate cancer	8	7	97	1.13E-05	3.30E-04	Both
Signaling pathways regulating pluripotency of stem cells	9	8	138	1.17E-05	3.50E-04	Both
Glioma	7	7	68	1.20E-05	6.10E-05	Both
PI3K-Akt signaling pathway	13	9	348	1.23E-05	9.20E-03	Both
Melanoma	7	7	72	1.40E-05	7.88E-05	Both
Rap1 signaling pathway	10	11	203	2.05E-05	5.98E-05	Both
Longevity regulating pathway	7	7	88	4.15E-05	2.20E-04	Both
Viral carcinogenesis	9	7	183	6.18E-05	4.60E-03	Both
HIF-1 signaling pathway	7	7	98	7.02E-05	3.30E-04	Both
mTOR signaling pathway	8	6	148	9.73E-05	6.40E-03	Both
Prolactin signaling pathway	6	6	69	1.00E-04	3.70E-04	Both
Cellular senescence	8	11	156	1.20E-04	1.01E-05	Both
Type II diabetes mellitus	5	3	46	2.00E-04	1.74E-02	Both
Estrogen signaling pathway	7	6	133	3.50E-04	4.60E-03	Both
Maturity onset diabetes of the young	4	3	26	3.60E-04	6.40E-03	Both
Small cell lung cancer	6	4	92	3.60E-04	1.82E-02	Both
Epstein-Barr virus infection	8	6	194	4.20E-04	1.54E-02	Both
Human papillomavirus infection	10	10	317	4.20E-04	2.20E-03	Both
AGE-RAGE signaling pathway in diabetic complications	6	10	98	4.40E-04	2.27E-06	Both
Non-small cell lung cancer	5	4	66	7.00E-04	9.00E-03	Both
Hepatocellular carcinoma	7	9	163	8.40E-04	2.20E-04	Both
MAPK signaling pathway	9	10	293	1.00E-03	1.70E-03	Both

Supplementary Table 4.1 (Continued)

Pathway	Observed gene count (Cluster 1)	Observed gene count (Cluster 2)	Background gene count	FDR (Cluster 1)	FDR (Cluster 2)	Assigned Cluster
AMPK signaling pathway	6	5	120	1.10E-03	1.10E-02	Both
Chronic myeloid leukemia	5	4	76	1.10E-03	1.20E-02	Both
Kaposi's sarcoma-associated herpesvirus infection	7	9	183	1.40E-03	3.60E-04	Both
Relaxin signaling pathway	6	6	130	1.50E-03	4.40E-03	Both
HTLV-I infection	8	9	250	1.60E-03	1.90E-03	Both
Measles	6	5	133	1.60E-03	1.45E-02	Both
Focal adhesion	7	7	197	1.80E-03	6.10E-03	Both
Ovarian steroidogenesis	4	5	49	1.90E-03	7.40E-04	Both
Progesterone-mediated oocyte maturation	5	5	94	2.20E-03	5.70E-03	Both
MicroRNAs in cancer	6	7	149	2.40E-03	1.90E-03	Both
Regulation of lipolysis in adipocytes	4	4	53	2.40E-03	5.50E-03	Both
cGMP-PKG signaling pathway	6	7	160	3.30E-03	2.50E-03	Both
Insulin resistance	5	6	107	3.50E-03	2.10E-03	Both
Ras signaling pathway	7	7	228	3.60E-03	1.03E-02	Both
Central carbon metabolism in cancer	4	5	65	4.30E-03	1.90E-03	Both
Acute myeloid leukemia	4	4	66	4.40E-03	9.00E-03	Both
Neurotrophin signaling pathway	5	4	116	4.40E-03	3.48E-02	Both
Thyroid hormone signaling pathway	5	6	115	4.40E-03	2.60E-03	Both
p53 signaling pathway	4	6	68	4.60E-03	3.60E-04	Both
Adherens junction	4	3	71	5.30E-03	4.34E-02	Both
Osteoclast differentiation	5	5	124	5.30E-03	1.20E-02	Both
Platelet activation	5	5	123	5.30E-03	1.18E-02	Both
Pancreatic cancer	4	5	74	5.70E-03	2.60E-03	Both
cAMP signaling pathway	6	7	195	6.70E-03	5.90E-03	Both

Supplementary Table 4.1 (Continued)

Pathway	Observed gene count (Cluster 1)	Observed gene count (Cluster 2)	Background gene count	FDR (Cluster 1)	FDR (Cluster 2)	Assigned Cluster
EGFR tyrosine kinase inhibitor resistance	4	6	78	6.70E-03	6.50E-04	Both
Phospholipase D signaling pathway	5	6	145	9.40E-03	6.10E-03	Both
Gastric cancer	5	7	147	9.80E-03	1.90E-03	Both
Cushing's syndrome	5	8	153	1.12E-02	5.70E-04	Both
Influenza A	5	6	168	1.60E-02	1.02E-02	Both
Transcriptional misregulation in cancer	5	11	169	1.61E-02	1.62E-05	Both
Cholinergic synapse	4	4	111	1.92E-02	3.14E-02	Both
Endometrial cancer	3	5	58	2.04E-02	1.50E-03	Both
VEGF signaling pathway	3	3	59	2.10E-02	3.05E-02	Both
Basal cell carcinoma	3	4	63	2.47E-02	7.80E-03	Both
Fc epsilon RI signaling pathway	3	3	67	2.82E-02	3.81E-02	Both
Apelin signaling pathway	4	5	133	3.18E-02	1.45E-02	Both
Gastric acid secretion	3	4	72	3.31E-02	1.08E-02	Both
Adrenergic signaling in cardiomyocytes	4	6	139	3.58E-02	5.50E-03	Both
Alcoholism	4	5	142	3.79E-02	1.70E-02	Both
Hepatitis B	4	7	142	3.79E-02	1.80E-03	Both
Hippo signaling pathway	4	5	152	4.43E-02	2.10E-02	Both
Insulin secretion	3	5	84	4.44E-02	4.20E-03	Both
Colorectal cancer	3	5	85	4.48E-02	4.30E-03	Both
Gap junction	3	4	87	4.69E-02	1.62E-02	Both
Dilated cardiomyopathy (DCM)	3	4	88	4.77E-02	1.67E-02	Both
GnRH signaling pathway	3	4	88	4.77E-02	1.67E-02	Both
Jak-STAT signaling pathway	4	5	160	4.77E-02	2.54E-02	Both
Inflammatory mediator regulation of TRP channels	3	8	92	4.99E-02	5.98E-05	Both

Supplementary Table 5.1: Sentinel meQTLs for BMI-MVPs from BIOS data

CpG	Chr: Pos	Nearest Gene	Sentinel cis-meQTL	SNP Chr: pos	Distance from CpG	Association P-value	BMI Beta	BMI P-value
cg00094412	6: 29592854	*GABBR1*	rs2747429	6: 29648377	55523	1.38E-74	-0.0187	1.01E-05
cg00108715	3: 52565015	*NT5DC2*	rs6778735	3: 52565100	85	8.60E-28	0.0151	9.42E-07
cg00138407	3: 47386505	*KLHL18*	rs295458	3: 47385585	-920	2.52E-14	-0.0037	1.60E-01
cg00238353	10: 129785537	*PTPRE*	rs61873722	10: 129789935	4398	1.34E-29	0.0006	9.12E-01
cg00244001	10: 126336805	*FAM53B*	rs11245333	10: 126383088	46283	6.60E-17	0.0013	5.87E-01
cg00431050	10: 103985730	*ELOVL3*	rs61873698	10: 104207431	221701	3.32E-19	-0.0173	1.54E-02
cg00673344	3: 156807691	*LINC00880*	rs3113	3: 156814949	7258	4.54E-30	0.0036	3.57E-01
cg00863378	16: 56549757	*BBS2*	rs72814488	16: 56498493	-51264	2.04E-236	-0.0089	3.34E-02
cg00973118	16: 374570	*AXIN1*	rs3848365	16: 374617	47	2.81E-223	-0.0011	7.70E-01
cg01243823	16: 50732212	*NOD2*	rs6500328	16: 50736656	4444	5.78E-30	0.0034	1.99E-01
cg01511901	13: 31004719	*UBE2L5P*	rs1028937	13: 30945840	-58879	5.22E-11	-0.001	7.14E-01
cg02286155	5: 176826262	*SLC34A1*	rs7708314	5: 176809618	-16644	2.07E-09	0.0042	2.88E-01
cg02650017	17: 47301614	*PHOSPHO1*	rs850523	17: 47313931	12317	4.31E-14	-0.0075	5.08E-02
cg02716826	9: 33447032	*AQP3*	rs591810	9: 33447424	392	1.59E-15	0.0029	3.35E-01
cg03050965	1: 101705237	*S1PR1*	rs6680048	1: 101620370	-84867	9.28E-09	-0.0041	1.91E-01
cg03159676	16: 85600536	*GSE1*	rs9921154	16: 85597734	-2802	2.81E-31	0.0032	3.85E-01
cg03433986	11: 62477624	*BSCL2*	rs516580	11: 62492124	14500	1.79E-12	0.0049	1.09E-01
cg03523676	14: 24540235	*CPNE6*	rs7146599	14: 24527672	-12563	5.90E-73	0.0101	6.38E-03
cg03725309	1: 109757585	*SARS*	rs4970829	1: 109757295	-290	4.32E-08	-0.0082	2.47E-01
cg03885055	1: 16723232	*SPATA21*	rs149561	1: 16797485	74253	6.51E-06	-0.0051	1.76E-01
cg04011474	2: 28904455	*RNA5SP89*	rs55720575	2: 28889033	-15422	1.91E-14	0.0064	7.84E-02
cg04126866	10: 85932763	*C10orf99*	rs68097915	10: 85937538	4775	3.42E-30	-0.0005	8.88E-01
cg04232128	5: 138861241	*TMEM173*	rs9716069	5: 138842818	-18423	8.56E-15	-0.0056	7.15E-02
cg04577162	7: 73667397	*RFC2*	rs3135660	7: 73664710	-2687	2.19E-33	0.0093	1.79E-01
cg05063895	16: 2073518	*SLC9A3R2*	rs28698483	16: 2080571	7053	5.07E-06	0.0117	2.51E-03
cg05720226	7: 116786597	*ST7*	rs10270156	7: 116789934	3337	4.02E-230	-0.0003	9.50E-01
cg05845030	12: 91573247	*DCN*	rs2639662	12: 91683609	110362	8.22E-10	0	9.99E-01
cg06012428	6: 157477204	*ARID1B*	rs74854649	6: 157373244	-103960	1.11E-06	0.0013	8.44E-01
cg06192883	15: 52554171	*MYO5C*	rs71472932	15: 52541976	-12195	2.43E-05	-0.0119	4.19E-02
cg06559575	12: 53490352	*IGFBP6*	rs10876407	12: 53502438	12086	1.43E-30	-0.0125	5.28E-04
cg06898549	12: 41083590	*CNTN1*	rs4768288	12: 40922091	-161499	3.95E-12	-0.0032	3.43E-01
cg07021906	16: 87866833	*SLC7A5*	rs4262954	16: 87814595	-52238	1.39E-18	0.004	2.86E-01
cg07037944	15: 64290807	*DAPK2*	rs28660107	15: 64281243	-9564	5.05E-05	0.007	1.29E-01
cg07136133	11: 36422377	*PRR5L*	rs59943655	11: 36422532	155	1.04E-20	-0.008	2.12E-01
cg07471614	8: 125855152	*LINC00964*	rs11777682	8: 125860396	5244	2.62E-12	-0.0004	9.12E-01
cg07504977	10: 102131012	*LINC00263*	rs145744215	10: 102129774	-1238	6.60E-49	0.001	8.39E-01

Supplementary Table 5.1 (Continued)

CpG	Chr: Pos	Nearest Gene	Sentinel cis-meQTL	SNP Chr: pos	Distance from CpG	Association P-value	BMI Beta	BMI P-value
cg07682160	19: 18959935	*UPF1*	rs7250622	19: 18984843	24908	1.41E-06	-0.0116	2.10E-03
cg07728579	15: 83475013	*FSD2*	rs12594415	15: 83527562	52549	1.25E-26	0.0017	6.75E-01
cg07769588	19: 10655622	*ATG4D*	rs62128793	19: 10642931	-12691	1.85E-10	0.0003	9.50E-01
cg08305942	16: 79692354	*MAF*	rs4889009	16: 79700447	8093	1.03E-27	-0.0003	9.08E-01
cg08309687	21: 35320596	*LINC00649*	rs8128167	21: 35307200	-13396	1.92E-68	-0.0038	1.54E-01
cg08443038	16: 89006877	*CBFA2T3*	rs491224	16: 88995480	-11397	9.73E-30	0.0079	3.42E-02
cg08548559	22: 31686097	*PIK3IP1*	rs739427	22: 31659101	-26996	3.27E-310	0.0049	5.77E-02
cg08726900	16: 89550474	*ANKRD11*	rs2019604	16: 89615765	65291	8.48E-06	0.0081	1.35E-02
cg08857797	17: 40927699	*VPS25*	rs55999482	17: 40919453	-8246	2.39E-07	0.002	6.26E-01
cg09152259	2: 128156114	*MAP3K2*	rs2276683	2: 128146762	-9352	3.27E-310	0.0009	8.37E-01
cg09554443	1: 167487762	*CD247*	rs58946921	1: 167506299	18537	5.65E-14	0.0035	5.19E-01
cg09664445	17: 2612406	*CLUH*	rs12051903	17: 2555511	-56895	4.55E-11	0.0011	7.78E-01
cg09777883	11: 112093696	*BCO2*	rs67245191	11: 112072161	-21535	1.43E-106	0.0079	4.23E-02
cg10179300	5: 14147618	*TRIO*	rs352657	5: 14170410	22792	2.18E-45	-0.0046	2.12E-01
cg10438589	4: 14531493	*LINC00504*	rs16890352	4: 14385522	-145971	2.64E-31	0.0127	3.96E-02
cg10513161	3: 183705727	*ABCC5*	rs6773408	3: 183716609	10882	6.93E-68	0.0031	3.99E-01
cg10717869	1: 205780912	*SLC41A1*	rs4396169	1: 205768309	-12603	3.74E-12	0.0065	1.21E-02
cg10919522	14: 74227441	*ELMSAN1*	rs11624967	14: 74227246	-195	1.95E-40	0.0041	1.20E-01
cg10922280	16: 68034227	*DUS2L*	rs8059305	16: 68036666	2439	1.77E-21	0.0048	1.61E-01
cg10927968	11: 1807333	*CTSD*	rs55877559	11: 1778176	-29157	1.07E-04	-0.0041	5.27E-01
cg11024682	17: 17730094	*SREBF1*	rs8070432	17: 17480474	-249620	4.83E-19	-0.0034	5.35E-01
cg11080651	5: 10445523	*ROPN1L*	rs62364292	5: 10445835	312	6.10E-06	-0.0128	8.15E-02
cg11202345	17: 76976057	*LGALS3BP*	rs8071100	17: 76783078	-192979	4.25E-12	0.005	2.94E-01
cg11376147	11: 57261198	*SLC43A1*	rs2511984	11: 57278733	17535	1.76E-07	-0.0052	2.11E-01
cg11614585	20: 897050	*ANGPT4*	rs1014898	20: 895531	-1519	3.34E-12	0.0045	2.93E-01
cg11650298	13: 44690989	*SMIM2-AS1*	rs10507524	13: 44684600	-6389	3.43E-07	0.0018	6.77E-01
cg11832534	1: 3563998	*WRAP73*	rs3765706	1: 3597224	33226	7.32E-35	-0.0005	9.07E-01
cg11927233	5: 170816542	*NPM1*	rs78276384	5: 170815521	-1021	9.48E-15	0.0092	6.73E-02
cg12593793	1: 156074135	*LMNA*	rs915179	1: 156078249	4114	2.69E-18	0.0061	2.22E-02
cg12992827	3: 101901234	*ZPLD1*	rs13087130	3: 101901067	-167	3.48E-73	0.0015	6.01E-01
cg13591783	9: 75768868	*ANXA1*	rs2795112	9: 75769950	1082	5.65E-37	-0.0011	7.82E-01
cg13781414	9: 138951648	*NACC2*	rs4842069	9: 138891645	-60003	5.51E-07	-0.0021	5.94E-01
cg13922488	19: 14545201	*PKN1*	rs3826758	19: 14572497	27296	2.95E-06	-0.0061	1.07E-01
cg14020176	17: 72764985	*SLC9A3R1*	rs12601504	17: 72754829	-10156	2.02E-41	0.003	2.77E-01
cg14264316	9: 134280803	*PRRC2B*	rs34159422	9: 134386659	105856	7.11E-07	0.0068	2.74E-01
cg14476101	1: 120255992	*PHGDH*	rs11583993	1: 120255370	-622	3.85E-228	0.0109	7.48E-02

Supplementary Table 5.1 (Continued)

CpG	Chr: Pos	Nearest Gene	Sentinel cis-meQTL	SNP Chr: pos	Distance from CpG	Association P-value	BMI Beta	BMI P-value
cg15323828	1: 226053673	*TMEM63A*	rs9919303	1: 226044766	-8907	3.50E-216	0.0048	1.22E-01
cg15681239	3: 38080203	*DLEC1*	rs4389435	3: 38080101	-102	1.54E-219	0.0008	7.82E-01
cg16163382	2: 37938640	*CDC42EP3*	rs9808547	2: 37944134	5494	5.45E-16	-0.0058	4.11E-02
cg16578636	10: 92987457	*PCGF5*	rs2648718	10: 93011770	24313	1.10E-33	0.0077	9.80E-03
cg16594806	1: 59473943	*PHBP3*	rs2716121	1: 59474554	611	2.44E-144	-0.0003	9.00E-01
cg16815882	1: 35908609	*KIAA0319L*	rs1188633	1: 35903912	-4697	4.51E-09	-0.0047	3.11E-01
cg16846518	3: 128062608	*EEFSEC*	rs2687729	3: 127895226	-167382	1.51E-11	0.0056	5.55E-02
cg17260706	11: 118782879	*BCL9L*	rs523604	11: 118755738	-27141	8.69E-05	0.0018	5.04E-01
cg17501210	6: 166970252	*RPS6KA2*	rs9355572	6: 166732650	-237602	1.09E-08	-0.0016	7.08E-01
cg17901584	1: 55353706	*DHCR24*	rs687565	1: 55364663	10957	6.22E-33	0.0064	8.61E-02
cg17971578	1: 36852463	*STK40*	rs72663467	1: 36837953	-14510	1.29E-09	-0.0093	9.52E-02
cg18098839	3: 167742700	*GOLIM4*	rs73174980	3: 167747239	4539	4.97E-45	0.0111	1.25E-01
cg18120259	6: 43894639	*C6orf223*	rs7745517	6: 43895095	456	1.12E-173	0.0074	8.98E-02
cg18181703	17: 76354621	*SOCS3*	rs62080378	17: 76282256	-72365	1.34E-05	-0.0029	4.30E-01
cg18219562	17: 41773643	*MEOX1*	rs1107747	17: 41774270	627	1.43E-05	-0.0052	5.29E-02
cg18513344	3: 195531298	*MUC4*	rs2688530	3: 195535466	4168	2.03E-62	0.0021	6.40E-01
cg19217955	17: 7123994	*ACADVL*	rs62062765	17: 6894691	-229303	1.70E-05	0.0022	7.48E-01
cg19373099	2: 210008092	*CRYGFP*	rs6707503	2: 210068146	60054	3.56E-27	0.0035	3.59E-01
cg19566658	7: 100466241	*TRIP6*	rs13306969	7: 100525805	59564	1.12E-124	-0.0029	3.90E-01
cg19589396	8: 103937374	*RPL5P24*	rs613049	8: 103931309	-6065	1.80E-67	0.0018	5.08E-01
cg21108085	11: 44591098	*CD82*	rs10769059	11: 44590729	-369	7.46E-06	0.0008	8.31E-01
cg21429551	7: 30635762	*GARS*	rs3779250	7: 30694260	58498	2.89E-13	0.0001	9.85E-01
cg21486834	17: 74477542	*RHBDF2*	rs3826288	17: 74477340	-202	5.63E-30	0.0066	1.66E-01
cg22012981	3: 58522689	*ACOX2*	rs4681863	3: 58521124	-1565	3.85E-26	-0.0045	4.19E-01
cg22103219	7: 101934892	*SH2B2*	rs803092	7: 101932722	-2170	1.35E-47	0.0026	4.77E-01
cg22488164	12: 14716910	*PLBD1*	rs746690	12: 14783798	66888	1.93E-17	-0.0048	9.00E-02
cg22534374	1: 201511610	*RPS10P7*	rs11582434	1: 201507911	-3699	1.91E-266	0.0026	4.76E-01
cg23032421	3: 3152038	*IL5RA*	rs168025	3: 3150530	-1508	2.65E-21	-0.0005	8.88E-01
cg23232188	3: 121556543	*EAF2*	rs9854539	3: 121503395	-53148	8.09E-30	0.0061	2.22E-02
cg24403644	20: 42574624	*TOX2*	rs4812770	20: 42575630	1006	2.53E-15	-0.0053	2.52E-01
cg24469729	7: 27160520	*HOXA-AS2*	rs3807592	7: 27136729	-23791	4.22E-135	0.0036	2.15E-01
cg24679890	19: 17246356	*MYO9B*	rs111366154	19: 17309577	63221	7.20E-13	0.0164	3.44E-02
cg25001190	1: 61668835	*NFIA*	rs78378256	1: 61662952	-5883	1.94E-17	0	9.95E-01
cg25197194	3: 128758787	*EFCC1*	rs2341295	3: 128751102	-7685	2.72E-54	0.004	1.32E-01
cg25217710	1: 156609523	*BCAN*	rs6666910	1: 156566111	-43412	4.46E-06	0.0006	8.82E-01
cg25435714	7: 157083381	*RN7SL142P*	rs2527874	7: 157073637	-9744	7.38E-49	0.0048	3.25E-01

Supplementary Table 5.1 (Continued)

CpG	Chr: Pos	Nearest Gene	Sentinel cis-meQTL	SNP Chr: pos	Distance from CpG	Association P-value	BMI Beta	BMI P-value
cg25570328	2: 108903952	*SULT1C2*	rs2305484	2: 108905030	1078	1.29E-07	0.0001	9.79E-01
cg25649826	17: 20938740	*USP22*	rs7226229	17: 20924077	-14663	3.34E-127	0.0045	1.59E-01
cg26033520	10: 74004071	*ANAPC16*	rs4746108	10: 74016892	12821	3.59E-12	-0.0022	4.28E-01
cg26253134	2: 70751721	*TGFA*	rs72912111	2: 70773571	21850	1.83E-12	-0.0083	1.40E-01
cg26357885	14: 65006204	*HSPA2*	rs45526332	14: 65006583	379	2.70E-24	-0.0129	7.40E-02
cg26361535	8: 144576604	*ZC3H3*	rs7844860	8: 144636949	60345	1.95E-44	0.003	4.12E-01
cg26403843	5: 158634085	*RNF145*	rs6556405	5: 158635102	1017	3.27E-310	0.0066	3.95E-02
cg26542660	4: 56813860	*CEP135*	rs878956	4: 56814115	255	1.04E-19	-0.0026	4.92E-01
cg26663590	16: 28959310	*NFATC2IP*	rs115616784	16: 29000446	41136	4.75E-69	0.0197	2.36E-07
cg26687842	13: 41055491	*LINC00598*	rs2721069	13: 41143720	88229	1.34E-10	0.0047	9.99E-02
cg26804423	7: 8201134	*ICA1*	rs10085429	7: 8209069	7935	5.06E-44	-0.0012	7.32E-01
cg26878209	10: 112375475	*SMC3*	rs11195232	10: 112375282	-193	3.70E-42	-0.0027	4.13E-01
cg26894079	11: 122954435	*CLMP*	rs34817879	11: 123023729	69294	2.95E-06	0.0009	8.47E-01
cg26952928	8: 142230233	*SLC45A4*	rs3824235	8: 142222261	-7972	7.92E-34	0.0045	2.27E-01
cg27050612	17: 46133198	*NFE2L1*	rs2325750	17: 46022330	-110868	2.26E-06	-0.0052	6.64E-02
cg27087650	19: 45255796	*BCL3*	rs62117162	19: 45239536	-16260	7.81E-21	0.0023	7.25E-01
cg27115863	22: 37921640	*CARD10*	rs6000762	22: 37918472	-3168	1.01E-09	0.0014	6.69E-01
cg27117792	12: 102330180	*DRAM1*	rs12298720	12: 102318856	-11324	6.23E-12	0.0024	5.62E-01
cg27184903	15: 29285727	*APBA2*	rs11852567	15: 29292973	7246	1.49E-41	0.0064	2.83E-01
cg27547344	1: 43765617	*TIE1*	rs3120124	1: 43764165	-1452	7.96E-91	0.0014	7.16E-01
cg27614723	15: 92399897	*SLCO3A1*	rs7165398	15: 92394095	-5802	2.32E-07	0.0085	4.94E-03

Supplementary Table 6.1: Reproductive variables from online touchscreen questionnaire

	Touchscreen Question	# non-missing (before cleaning)
1	Answered sexual history questions*	501,713
2	"What was your age when you first had sexual intercourse? (Sexual intercourse includes vaginal, oral or anal intercourse)"	435,505
3	"About how many sexual partners have you had in your lifetime?"	405,179
4	"Have you ever had sexual intercourse with someone of the same sex?"	448,912
5	"When did you start to grow facial hair?"	217,846
6	"When did your voice break?"	207,852
7	"Which of the following best describes your hair/balding patterns?"	224,889
8	"How many children have you fathered?"	224,898
9	"Have you ever been for breast cancer screening (a mammogram)?"	272,479
10	"How many years ago was your last screen?"	160,705
11	"Have you ever had a cervical smear test?"	272,209
12	"How many years ago was your last cervical smear test?"	208,494
13	"How old were you when your periods started?"	264,605
14	"Have you had your menopause (periods stopped)?"*	272,453
15	"How many children have you given birth to? (Please include live births only)"	272,636
16	"What was the birth weight of your first child in pounds? (do not include twins)"	216,756
17	"How old were you when you had your FIRST child?"	184,636
18	"How old were you when you had your LAST child?"	184,260
19	"Have you ever had any stillbirths, spontaneous miscarriages or terminations?"	268,386
20	"Have you ever taken the contraceptive pill? (include the 'mini-pill')?"	272,046
21	"About how old were you when you first went on the contraceptive pill?"	212,309
22	"How old were you when you last used the contraceptive pill?"	192,752
23	"Have you ever used hormone replacement therapy (HRT)?"	271,891
24	"How old were you when you had your hysterectomy?"	49,870
25	"Have you had BOTH ovaries removed?"	268,861
26	"How old were you when you first used HRT?"	93,253
27	"How old were you when you last used HRT?"	77,344
28	"How old were you when your periods stopped?"	154,662
29	"Have you ever had a hysterectomy (womb removed)?"	241,272
30	"How many sexual partners of the same sex have you had in your lifetime?"	13,358
31	"How many days since your last menstrual period?"	56,244
32	"How many days is your usual menstrual cycle? (The number of days between each menstrual period)"	48,386
33	"How many stillbirths? (enter 0 if none)"	86,997
34	"How many spontaneous miscarriages? (enter 0 if none)"	86,806
35	"How many terminations? (enter 0 if none)"	86,367
36	"How old were you when you had your first child?"	36,416
37	"How old were you when you had BOTH ovaries removed?"	21,206
38	"Did you only have diabetes during pregnancy?"†	9,551

*Not included in GWAS.
†Only asked to women who indicated they had been diagnosed with diabetes by a doctor.

Supplementary Table 6.2: REPROWAS traits with constituent UK Biobank variables

Phenotype	Cases (N)		Constituent variables	
	ICD-10	Self-reports	ICD-10 codes	Self-reported illness codes
Hyperplasia of prostate	15,530	8,974	N40	Enlarged prostate; BPH/benign prostatic hypertrophy
"Cystitis & other urinary tract infections"	20,910	3,083	N30.0, N30.1, N30.2, N30.3, N30.4, N30.8, N30.9, N39.0	Urinary tract infection/kidney infection; cystitis
Leiomyoma of uterus	15,102	8,681	D25.0, D25.1, D25.2, D25.9	Uterine fibroids
Malignant neoplasm of breast	17,354	-	C50.0, C50.1, C50.2, C50.3, C50.4, C50.5, C50.6, C50.8, C50.9	-
Malignant neoplasm of prostate	9,013	4,159	C61	Prostate cancer
Polyp of corpus uteri	10,166	1,592	N84.0	Uterine polyps
Excessive and frequent menstruation with regular cycle	11,746	-	N92.0	-
Postmenopausal bleeding	11,553	-	N95.0	-
Endometriosis	6,198	4,374	N80.0, N80.1, N80.2, N80.3, N80.4, N80.5, N80.6, N80.8, N80.9	Endometriosis
"Ovarian cysts"	6,188	4,347	N83.0, N83.1, N83.2	Ovarian cyst or cysts
Uterovaginal prolapse	6,721	3,038	N81.2, N81.3, N81.4	Vaginal prolapse/uterine prolapse
Perineal laceration during delivery	7,643	-	O70.0, O70.1, O70.2, O70.3, O70.9	-
"Enterocele & Rectocele"	6,205	-	N81.5, N81.6	-
"Female urethrocele & cystocele"	6,047	-	N81.0, N81.1	-
Other abnormal uterine and vaginal bleeding	5,298	-	N93.0, N93.8, N93.9	-
Labour and delivery complicated by foetal stress [distress]	4,714	-	O68.0, O68.1, O68.2, O68.3, O68.8, O68.9	-
Benign mammary dysplasia	2,436	2,245	N60.0, N60.1, N60.2, N60.3, N60.4, N60.8, N60.9	Fibrocystic disease; breast cysts
Polyp of cervix uteri	3,984	228	N84.1	Cervical polyps

Supplementary Table 6.2 (Continued)

Phenotype	Cases (N)		Constituent variables	
	ICD-10	Self-reports	ICD-10 codes	Self-reported illness codes
Benign neoplasm of breast	1,817	2,118	D24	Breast fibroadenoma; benign breast lump
Excessive and frequent menstruation with irregular cycle	3,424	-	N92.1	-
Spontaneous abortion	1,943	1,480	O03.0, O03.1, O03.3, O03.4, O03.5, O03.6, O03.8, O03.9	Miscarriage
"Other noninflammatory disorders of uterus"	3,379	-	N85.3, N85.5, N85.6, N85.7, N85.8, N85.9	-
Female pelvic peritoneal adhesions	3,278	-	N73.6	-
Dysplasia of cervix uteri	2,197	804	N87.0, N87.1, N87.2, N87.9	Abnormal smear (cervix)
"Other menopausal and other perimenopausal disorders"	2,272	550	N95.1, N95.2, N95.3, N95.8, N95.9	Menopausal symptoms/menopause
Carcinoma in situ of breast	2,807	-	D05.0, D05.1, D05.7, D05.9	-
Single spontaneous delivery	2,699	-	O80.0, O80.1, O80.8, O80.9	-
Other specified disorders of male genital organs	2,631	-	N50.8	-
Maternal care for other conditions predominantly related to pregnancy	2,624	-	O26.0, O26.1, O26.2, O26.3, O26.4, O26.5, O26.6, O26.7, O26.8, O26.9	-
Redundant prepuce, phimosis and paraphimosis	2,549	-	N47	-
Long labour	2,539	-	O63.0, O63.1, O63.9	-
Malignant neoplasm of ovary	1,593	904	C56	Ovarian cancer
Inflammatory diseases of prostate	2,203	246	N41.0, N41.1, N41.2, N41.3, N41.8, N41.9	Prostatitis
Other noninflammatory disorders of cervix uteri	2,336	-	N88.0, N88.1, N88.2, N88.3, N88.4, N88.8, N88.9	-
"Dysmenorrhoea"	1,947	378	N94.4, N94.5, N94.6	Dysmenorrhoea
"Irregular menstruation"	2,248	-	N92.5, N92.6	-

Supplementary Table 6.2 (Continued)

Phenotype	Cases (N)		Constituent variables	
	ICD-10	Self-reports	ICD-10 codes	Self-reported illness codes
Other noninflammatory disorders of vulva and perineum	2,106	-	N90.5, N90.6, N90.7, N90.8, N90.9	-
"Maternal care for other foetal problems"	2,082	-	O36.0, O36.1, O36.3, O36.7, O36.8, O36.9	-
Female infertility	1,438	627	N97.0, N97.1, N97.2, N97.3, N97.4, N97.8, N97.9	Female infertility
Unspecified lump in breast	1,952	-	N63	-
Malignant neoplasm of corpus uteri	1,918	-	C54.0, C54.1, C54.2, C54.3, C54.8, C54.9, C55	-
Postpartum haemorrhage	1,778	-	O72.0, O72.1, O72.2, O72.3	-
Benign neoplasm of ovary	1,769	-	D27	-
Other disorders of prostate	1,730	-	N42.0, N42.1, N42.2, N42.3, N42.8, N42.9	-
Hydrocele and spermatocele	1,702	-	N43.0, N43.1, N43.2, N43.3, N43.4	-
Premature rupture of membranes	1,662	-	O42.0, O42.1, O42.2, O42.9	-
Maternal care due to uterine scar from previous surgery	1,639	-	O34.2	-
Prolonged pregnancy	1,569	-	O48	-
Other disorders of breast	1,558	-	N64.0, N64.1, N64.2, N64.3, N64.4, N64.5, N64.8, N64.9	-
Missed abortion	1,546	-	O02.1	-
Medical abortion	1,514	-	O04.0, O04.1, O04.3, O04.4, O04.5, O04.6, O04.8, O04.9	-
Maternal care for known or suspected malpresentation of foetus	1,485	-	O32.0, O32.1, O32.2, O32.3, O32.4, O32.5, O32.6, O32.8, O32.9	-
"Endometrial hyperplasia"	1,483	-	N85.0, N85.1	-
Inflammatory disease of cervix uteri	1,422	-	N72	-

Supplementary Table 6.2 (Continued)

Phenotype	Cases (N)		Constituent variables	
	ICD-10	Self-reports	ICD-10 codes	Self-reported illness codes
False labour	1,421	-	O47.0, O47.1, O47.9	-
Hypertrophy of uterus	1,416	-	N85.2	-
"Other conditions associated with female genital organs and menstrual cycle"	1,393	-	N94.8, N94.9, N94.0, N94.2, N94.3	-
Malignant neoplasm of testis	387	921	C62.0, C62.1, C62.9	Testicular cancer
Erosion and ectropion of cervix uteri	1,145	120	N86	Cervical erosion
Dyspareunia	1,262	-	N94.1	-
"Inflammatory disorder inlcuding orchitis and epididymitis"	1,248	-	N45.0, N45.9, N49.2, N49.8, N49.0, N49.1	-
Haemorrhage in early pregnancy	1,202	-	O20.0, O20.8, O20.9	-
Anaemia complicating pregnancy, childbirth and the puerperium	1,181	-	O99.0	-
"Gestational hypertension & Eclampsia"	1,172	-	O13, O14.0, O14.1, O14.9, O15.0, O15.1, O15.2, O15.9	-
Ovulation bleeding	1,171	-	N92.3	-
Carcinoma in situ of cervix uteri	723	411	D06.0, D06.1, D06.7, D06.9	Cervical intra-epithelial neoplasia (cin)/pre-cancerous cells (cervix)
Balanitis xerotica obliterans	1,041	-	N48.6	-
Hypertrophy of breast	1,017	-	N62	-
Other inflammation of vagina and vulva	988	-	N76.0, N76.1, N76.2, N76.3, N76.4, N76.5, N76.6, N76.8	-
Polycystic ovarian syndrome	300	660	E28.2	Polycystic ovaries/polycystic ovarian syndrome
"Cyst or abscess of Bartholin's gland"	960	-	N75.0, N75.1	-
Malposition of uterus	927	-	N85.4	-

Supplementary Table 6.2 (Continued)

Phenotype	Cases (N)		Constituent variables	
	ICD-10	Self-reports	ICD-10 codes	Self-reported illness codes
Human immunodeficiency virus [HIV] disease	365	519	B20.0, B20.1, B20.2, B20.3, B20.4, B20.6, B20.7, B20.8, B21.0, B21.1, B21.2, B21.3, B21.7, B21.8, B22.0, B22.1, B22.2, B22.7, B23.0, B23.2, B23.8, B24	HIV/AIDS
Ectopic pregnancy	384	493	O00.1, O00.2, O00.8, O00.9	Ectopic pregnancy
Unspecified maternal hypertension	853	-	O16	-
"Antepartum haemorrhage"	850	-	O46.0, O46.8, O46.9	-
Labour and delivery complicated by umbilical cord complications	831	-	O69.0, O69.1, O69.2, O69.3, O69.4, O69.5, O69.8, O69.9	-
Vascular disorders of male genital organs	774	-	N50.1	-
Single delivery by Caesarean section	755	-	O82.0, O82.1, O82.8, O82.9	-
Salpingitis and oophoritis	737	-	N70.0, N70.1, N70.9	-
"Infections with a predominantly sexual mode of transmission"	717	8	A51.4, A52.1, A52.3, A52.7, A52.8, A53.0, A53.9, A54.4, A54.8, A54.9, A56.0, A56.8, A58.0, A59.0, A59.9, A60.0, A60.1, A63.0, A63.8, A64, B08.1, B37.3	Chlamydia
Diabetes mellitus in pregnancy	408	306	O24.0, O24.1, O24.3, O24.4, O24.9	Gestational diabetes
Other noninflammatory disorders of ovary, Fallopian tube and broad ligament	628	-	N83.8	-
Inflammatory disorders of breast	613	-	N61	-
Abnormalities of forces of labour	601	-	O62.0, O62.1, O62.2, O62.3, O62.4, O62.8, O62.9	-
Placental disorders	557	-	O43.0, O43.1, O43.8, O43.9, O44.0, O44.1	-
Obstructed labour due to malposition and malpresentation of foetus	556	-	O64.0, O64.1, O64.2, O64.3, O64.4, O64.5, O64.8, O64.9	-

Supplementary Table 6.2 (Continued)

Phenotype	Cases (N)		Constituent variables	
	ICD-10	Self-reports	ICD-10 codes	Self-reported illness codes
Other specified disorders of penis	543	-	N48.8	-
Other female genital prolapse	527	-	N81.8	-
"Undescended testicle & Hypospadias"	270	227	Q53.0, Q53.1, Q53.2, Q53.9, Q54.0, Q54.1, Q54.4, Q54.8, Q54.9	Undescended testicle
Acne	83	401	L70.0, L70.8, L70.9	Acne/acne vulgaris
"Other obstructed labour"	481	-	O65.4, O65.5, O65.8, O65.9, O66.0, O66.1, O66.2, O66.4, O66.5, O66.8, O66.9	-
"Pelvic inflammatory disease and peritonitis"	464	-	N73.0, N73.1, N73.2, N73.3, N73.5, N73.8, N73.9	-
Other benign neoplasms of uterus	455	-	D26.0, D26.1, D26.7, D26.9	-
"Vaginal and vulva dysplasia"	446	-	N89.0, N89.1, N89.2, N89.3, N89.4, N90.0, N90.1, N90.2, N90.3, N90.4	-
"Malignant neoplasm & CIS of vulva & vagina"	396	39	C51.0, C51.1, C51.2, C51.8, C51.9, C52, D07.1, D07.2	Vaginal cancer
Multiple gestation	390	-	O30.0, O30.1, O30.2, O30.8	-
Malignant neoplasm of cervix uteri	388	-	C53.0, C53.1, C53.8, C53.9	-
Excessive vomiting in pregnancy	382	-	O21.0, O21.1, O21.2, O21.8, O21.9	-
Impotence of organic origin	376	-	N48.4	-
"Sexual dysfunction (less desire)"	375	-	F52.0, F52.1, F52.2, F52.6	-
Maternal care for known or suspected foetal abnormality and damage	372	-	O35.0, O35.1, O35.2, O35.3, O35.4, O35.5, O35.7, O35.8, O35.9	-
Inflammatory disease of uterus, except cervix	370	-	N71.0, N71.1, N71.9	-
Excessive bleeding in the premenopausal period	370	-	N92.4	-

Supplementary Table 6.2 (Continued)

Phenotype	Cases (N)		Constituent variables	
	ICD-10	Self-reports	ICD-10 codes	Self-reported illness codes
"Alopecia"	206	162	L63.0, L63.1, L63.8, L63.9, L64.0, L64.8, L64.9, L65.8, L65.9, L66.0, L66.1, L66.2, L66.3, L66.8, L66.9	Alopecia/hair loss
Maternal care for poor foetal growth	368	-	O36.5	-
"Carcinoma in situ of other and unspecified, prostate"	361	-	D07.5	-
"Puerperal infections"	338	-	O85, O86.0, O86.1, O86.2, O86.3, O86.4, O86.8	-
Infections of genito-urinary tract in pregnancy	335	-	O23.0, O23.1, O23.3, O23.4, O23.5, O23.9	-
Delayed delivery after spontaneous or unspecified rupture of membranes	333	-	O75.6	-
"Testicular hypofunction & Male Infertility"	292	36	E29.1, N64	Male infertility
Balanoposthitis	328	-	N48.1	-
Female genital prolapse, unspecified	314	-	N81.9	-
"Other complications of labour and delivery"	312	-	O75.0, O75.1, O75.3, O75.4, O75.5, O75.8, O75.9	-
Retained placenta and membranes, without haemorrhage	291	-	O73.0, O73.1	-
Complications of the puerperium, not elsewhere classified	287	-	O90.0 , O90.1 , O90.2 , O90.3 , O90.4 , O90.8 , O90.9	-
Fistulae involving female genital tract	282	-	N82.0 , N82.1 , N82.2 , N82.3 , N82.4 , N82.5 , N82.8 , N82.9	-
Other obstetric trauma	272	-	O71.1, O71.2, O71.3, O71.4, O71.5, O71.6, O71.7, O71.8, O71.9	-
Vaginitis, vulvitis and vulvovaginitis in infectious and parasitic diseases classified elsewhere	270	-	N77.1	-
"Polyp of vagina & vulva"	266	-	N84.2, N84.3	- -
Benign neoplasm of other and unspecified female genital organs	256	-	D28.0 , D28.1 , D28.2 , D28.7 , D28.9	-

Supplementary Table 6.2 (Continued)

Phenotype	Cases (N)		Constituent variables	
	ICD-10	Self-reports	ICD-10 codes	Self-reported illness codes
Congenital malformations of uterus and cervix	246	-	Q51.0, Q51.1, Q51.2, Q51.3, Q51.4, Q51.8, Q51.9	-
Congenital malformations of ovaries, Fallopian tubes and broad ligaments	238	-	Q50.0, Q50.1, Q50.3, Q50.4, Q50.5, Q50.6	-
Maternal care for excessive foetal growth	231	-	O36.6	-
Stricture and atresia of vagina	226	-	N89.5	-
Single delivery by forceps and vacuum extractor	216	-	O81.0 , O81.1 , O81.2, O81.3 , O81.4 , O81.5	-
Failed induction of labour	214	-	O61.0 , O61.1 , O61.8 , O61.9	-
Vaginal delivery following previous Caesarean section	211	-	O75.7	-
"Gestational oedema & proteinuria"	195	-	O12.0 , O12.1 , O12.2	-
Oligohydramnios	193	-	O41.0	-
Maternal care for tumour of corpus uteri	186	-	O34.1	-
"Malignant neoplasm & CIS of penis"	141	43	C60.0 , C60.1 , C60.2 , C60.8 , C60.9 , D07.4	Penis cancer
Pyrexia during labour, not elsewhere classified	184	-	O75.2	-
"Neoplasm of uncertain or unknown behaviour, ovary"	183	-	D39.1	-
Benign neoplasm of male genital organs	182	-	D29.0, D29.1, D29.2, D29.3, D29.4, D29.9	-
Leukoplakia of penis	172	-	N48.0	-
Hyperprolactinaemia	91	78	E22.1	Hypoprolactinaemia
Diseases of the digestive system complicating pregnancy, childbirth and the puerperium	169	-	O99.6	-
Polyhydramnios	162	-	O40	-

Phenotype	Cases (N)		Constituent variables	
	ICD-10	Self-reports	ICD-10 codes	Self-reported illness codes
"Other disorders of amniotic fluid and membranes"	152	-	O41.1, O41.8, O41.9	-
Absent, scanty and rare menstruation	151	-	N91.0, N91.1, N91.2, N91.4, N91.5	-
Noninflammatory disorder of ovary, Fallopian tube and broad ligament, unspecified	149	-	N83.9	-
"Other congenital malformations of male or female genital organs"	145	-	Q52.0, Q52.1, Q52.2, Q52.4, Q52.5, Q52.6, Q52.7, Q52.8, Q55.0, Q55.1, Q55.2, Q55.4, Q55.5, Q55.6, Q55.8	-
Other inflammatory disorders of penis	144	-	N48.2	-
Acquired atrophy of ovary and Fallopian tube	140	-	N83.3	-
"Intrapartum haemorrhage"	139	-	O67.8, O67.9	-
"Neoplasm of uncertain or unknown behaviour, breast"	138	-	D48.6	-
Hydatidiform mole	28	102	O01.0, O01.1, O01.9	-
Pre-existing hypertension complicating pregnancy, childbirth and the puerperium	122	-	O10.0, O10.1, O10.2, O10.9	-
Venous complications in the puerperium	120	-	O87.0, O87.1, O87.2, O87.8, O87.9	-
Maternal care for abnormality of vulva and perineum	119	-	O34.7	-
Malignant neoplasm of other and unspecified female genital organs	118	-	C57.0, C57.1, C57.4, C57.7, C57.8, C57.9	-
Endocrine, nutritional and metabolic diseases complicating pregnancy, childbirth and the puerperium	113	-	O99.2	-
Venous complications in pregnancy	109	-	O22.0, O22.1, O22.2, O22.3, O22.4, O22.5, O22.8, O22.9	-
Mental disorders and diseases of the nervous system complicating pregnancy, childbirth and the puerperium	105	-	O99.3	-
Atrophy of testis	100	-	N50.0	-
"Maternal care for intra-uterine death & hydrops fetalis"	99	-	O36.2, O36.4	-

Supplementary Table 7.1: Nominally significant IVW MR associations between AAM and REPROWAS phenotypes (continuous traits)

Trait	AAM		BMI		Total-T		Free-T		SHBG	
	Beta (95% CI)	P-value	Beta (95% CI)	P-value	Beta (95% CI)	P-value	Beta (95% CI)	P-value	Beta (95% CI)	P-value
Female Traits										
Age when periods started (menarche)	0.96 (0.94, 0.98)	5.0×10^{-8}	-0.71 (-0.83, -0.6)	1.8×10^{-34}	-0.04 (-0.09, 0.00)	0.06	-0.11 (-0.19, -0.04)	2.2×10^{-3}	0.01 (-0.06, 0.08)	0.76
Age first had sexual intercourse	0.25 (0.20, 0.30)	2.3×10^{-19}	-0.22 (-0.36, -0.08)	2.6×10^{-3}	-0.09 (-0.15, -0.03)	2.1×10^{-3}	-0.26 (-0.35, -0.18)	1.5×10^{-9}	0.2 0 (0.12, 0.29)	5.6×10^{-6}
Age at last live birth	0.26 (0.19, 0.34)	7.8×10^{-12}	-0.41 (-0.60, -0.22)	1.5×10^{-5}	-0.02 (-0.12, 0.07)	0.63	-0.23 (-0.36, -0.09)	1.2×10^{-3}	0.09 (-0.05, 0.22)	0.23
Age at first live birth	0.28 (0.20, 0.37)	4.7×10^{-11}	-0.38 (-0.60, -0.17)	4.1×10^{-4}	-0.14 (-0.24, -0.04)	6.6×10^{-3}	-0.37 (-0.52, -0.23)	4.2×10^{-7}	0.15 (0.00, 0.30)	0.06
Age when last used oral contraceptive pill	0.24 (0.14, 0.33)	5.8×10^{-7}	-0.81 (-1.06, -0.56)	2.1×10^{-10}	-0.03 (-0.16, 0.09)	0.58	-0.04 (-0.22, 0.13)	0.63	0.06 (-0.14, 0.27)	0.53
Age started oral contraceptive pill	0.11 (0.06, 0.16)	8.0×10^{-6}	-0.03 (-0.16, 0.10)	0.64	-0.01 (-0.07, 0.06)	0.84	-0.16 (-0.25, -0.07)	5.4×10^{-4}	0.16 (0.07, 0.26)	1.0×10^{-3}
Years since last cervical smear test	-0.16 (-0.24, -0.08)	1.1×10^{-4}	0.10 (-0.07, 0.27)	0.23	-0.17 (-0.28, -0.05)	3.8×10^{-3}	0.00 (-0.17, 0.17)	0.99	-0.28 (-0.43, -0.12)	7.1×10^{-4}
Age started hormone-replacement therapy (HRT)	0.19 (0.09, 0.29)	1.6×10^{-4}	-0.28 (-0.55, -0.01)	0.04	0.09 (-0.05, 0.24)	0.20	0.01 (-0.2, 0.22)	0.95	0.05 (-0.16, 0.26)	0.65
Age at menopause (last menstrual cycle)	0.18 (0.08, 0.27)	3.0×10^{-4}	0.04 (-0.20, 0.28)	0.74	0.14 (-0.01, 0.29)	0.08	0.06 (-0.15, 0.27)	0.56	-0.20 (-0.41, 0.02)	0.08
Male Traits										
Age first had sexual intercourse	0.25 (0.20, 0.30)	2.3×10^{-19}	-0.22 (-0.36, -0.08)	2.6×10^{-3}	0.08 (0.02, 0.14)	0.01	-0.10 (-0.20, 0.01)	0.7	0.13 (0.04, 0.22)	4.0×10^{-3}
Number of children fathered	-0.02 (-0.03, 0.00)	8.6×10^{-3}	0.03 (-0.01, 0.07)	0.16	0.00 (-0.02, 0.02)	0.73	0.05 (0.02, 0.08)	1.6×10^{-3}	-0.02 (-0.06, 0.01)	0.24

AAM = age at menarche; BMI = body mass index; Total-T = total testosterone; Free-T = free (biologically active) testosterone; SHBG = sex hormone binding globulin

Supplementary Table 7.2: Nominally significant IVW MR associations between AAM and REPROWAS phenotypes (categorical traits)

Trait	AAM		BMI		Total-T		Free-T		SHBG	
	OR (95% CI)	P-value	OR (95% CI)	P-value	OR (95% CI)	P-value	OR (95% CI)	P-value	OR (95% CI)	P-value
Female Traits										
Leiomyoma of uterus	0.90 (0.87, 0.94)	3.5×10^{-6}	1.13 (1.01, 1.27)	0.03	0.88 (0.82, 0.94)	7.6×10^{-5}	1.04 (0.95, 1.14)	0.39	0.78 (0.71, 0.85)	1.6×10^{-7}
Bilateral oophorectomy (both ovaries removed)	0.91 (0.88, 0.95)	6.8×10^{-6}	0.98 (0.88, 1.09)	0.73	0.93 (0.86, 0.99)	0.02	1.04 (0.95, 1.14)	0.43	0.85 (0.78, 0.93)	2.3×10^{-4}
Cystitis and other urinary tract infections	0.93 (0.90, 0.97)	2.1×10^{-4}	1.06 (0.96, 1.17)	0.25	1.02 (0.96, 1.09)	0.45	1.00 (0.92, 1.07)	0.90	0.99 (0.90, 1.08)	0.79
Ever taken oral contraceptive pill	1.05 (1.02, 1.08)	4.8×10^{-4}	0.85 (0.80, 0.92)	2.1×10^{-5}	1.00 (0.96, 1.03)	0.91	1.04 (0.98, 1.09)	0.18	0.92 (0.87, 0.98)	9.9×10^{-3}
Maternal care for tumour of corpus uteri	0.45 (0.28, 0.73)	1.0×10^{-3}	0.88 (0.26, 2.91)	0.83	0.90 (0.46, 1.76)	0.75	0.63 (0.24, 1.61)	.033	1.08 (0.37, 3.14)	0.89
Congenital malformation of uterus and cervix	0.59 (0.43, 0.82)	1.7×10^{-3}	4.41 (1.93, 10.09)	4.3×10^{-4}	0.90 (0.56, 1.44)	0.65	1.05 (0.54, 2.05)	0.88	0.50 (0.23, 1.07)	0.07
Ever had hysterectomy (womb removed)	0.94 (0.89, 0.98)	4.2×10^{-3}	0.92 (0.83, 1.02)	0.12	0.93 (0.87, 0.99)	0.02	1.00 (0.92, 1.09)	0.99	0.86 (0.78, 0.95)	1.9×10^{-3}
Enterocele and rectocele	0.91 (0.85, 0.98)	9.3×10^{-3}	1.08 (0.90, 1.29)	0.40	1.03 (0.93, 1.14)	0.60	1.19 (1.03, 1.39)	0.02	0.77 (0.66, 0.91)	1.7×10^{-3}
Excessive and frequent menstruation with regular cycle	0.93 (0.88, 0.98)	9.9×10^{-3}	1.21 (1.08, 1.37)	1.5×10^{-3}	0.93 (0.85, 1.01)	0.08	1.00 (0.89, 1.14)	0.94	0.83 (0.73, 0.93)	1.3×10^{-3}
Prolonged pregnancy	1.19 (1.04, 1.36)	0.01	0.74 (0.53, 1.04)	0.08	0.97 (0.80, 1.17)	0.73	0.92 (0.70, 1.19)	0.51	1.70 (1.26, 2.28)	4.5×10^{-4}
Polycystic ovarian syndrome	0.79 (0.66, 0.96)	0.01	3.53 (2.19, 5.69)	2.4×10^{-7}	1.63 (1.25, 2.13)	2.9×10^{-4}	2.25 (1.51, 3.35)	6.7×10^{-5}	0.49 (0.32, 0.74)	7.8×10^{-4}
Irregular menstruation	0.88 (0.79, 0.98)	0.02	1.31 (1.00, 1.72)	0.05	1.03 (0.88, 1.21)	0.68	1.24 (1.00, 1.55)	0.06	0.85 (0.66, 1.10)	0.22
Labour and delivery complicated by umbilical cord problems	1.25 (1.03, 1.51)	0.02	0.60 (0.38, 0.95)	0.03	0.93 (0.72, 1.20)	0.57	0.69 (0.48, 1.00)	0.05	1.05 (0.70, 1.59)	0.80

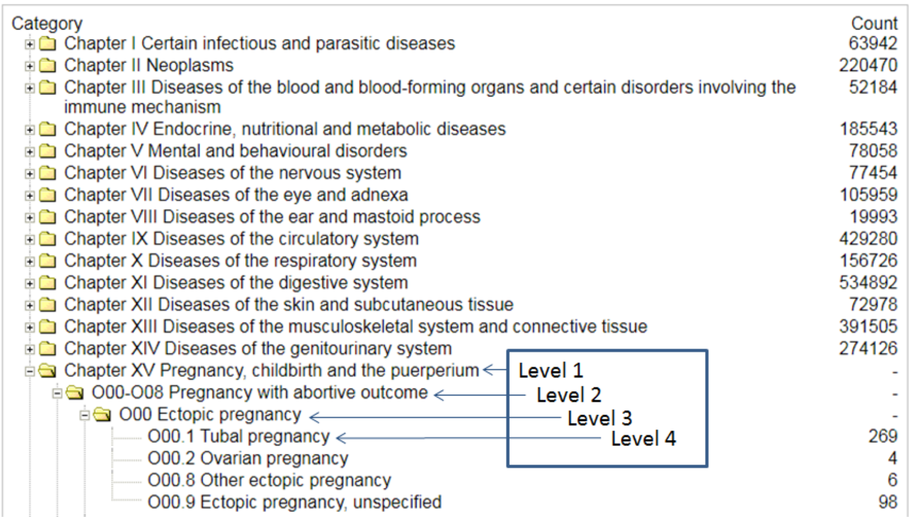
Supplementary Table 7.2 (Continued)

Trait	AAM		BMI		Total-T		Free-T		SHBG	
	OR (95% CI)	P-value	OR (95% CI)	P-value	OR (95% CI)	P-value	OR (95% CI)	P-value	OR (95% CI)	P-value
Maternal care for poor foetal growth	1.41 (1.05, 1.90)	0.02	0.64 (0.31, 1.32)	0.23	1.13 (0.75, 1.71)	0.56	1.00 (0.54, 1.83)	0.99	1.31 (0.68, 2.51)	0.42
Maternal care for other foetal problems	1.14 (1.02, 1.28)	0.02	1.22 (0.88, 1.69)	0.23	1.04 (0.88, 1.24)	0.65	0.80 (0.63, 1.01)	0.06	1.34 (1.02, 1.77)	0.04
Long labour	1.13 (1.01, 1.25)	0.03	0.92 (0.71, 1.19)	0.53	0.82 (0.70, 0.96)	0.01	0.72 (0.58, 0.90)	3.5×10^{-3}	1.37 (1.07, 1.76)	0.01
Endometriosis	0.93 (0.87, 1.00)	0.04	1.06 (0.90, 1.26)	0.47	0.86 (0.78, 0.95)	2.4×10^{-3}	0.84 (0.74, 0.96)	0.01	1.02 (0.89, 1.17)	0.79
Female urethrocele and cystocele	0.93 (0.88, 1.00)	0,04	1.08 (0.91, 1.29)	0.37	1.02 (0.92, 1.13)	0.66	1.05 (0.91, 1.20)	0.52	0.84 (0.72, 0.99)	0.04
Excessive vomiting in pregnancy	1.36 (1.01, 1.85)	0.05	0.73 (0.33, 1.63)	0.44	1.43 (0.92, 2.22)	0.12	1.23 (0.64, 2.35)	0.54	1.32 (0.63, 2.76)	0.47
Male Traits										
Relative age voice broke	0.08 (0.07, 0.08)	8.5×10^{-164}	-0.06 (-0.08, -0.05)	1.4×10^{-17}	0.01 (0.00, 0.02)	0.04	-0.03 (-0.05, -0.01)	1.9×10^{-3}	0.02 (0.01, 0.03)	1.3×10^{-4}
Relative age of first facial hair	0.11 (0.1, 0.12)	2.4×10^{-123}	-0.09 (-0.11, -0.07)	4.7×10^{-17}	0.01 (-0.01, 0.02)	0.32	-0.05 (-0.08, -0.02)	6.1×10^{-4}	0.03 (0.01, 0.05)	1.4×10^{-3}
Cystitis and other urinary tract infections	0.93 (0.90, 0.97)	2.1×10^{-4}	1.06 (0.96, 1.17)	0.25	1.04 (0.99, 1.09)	0.16	1.05 (0.98, 1.12)	0.20	1.04 (0.95, 1.14)	0.38
Impotence of organic origin	0.69 (0.53, 0.9)	7.3×10^{-3}	1.34 (0.67, 2.67)	0.41	0.99 (0.69, 1.42)	0.95	1.12 (0.65, 1.92)	0.69	0.95 (0.51, 1.79)	0.88
Inflammatory diseases of prostate	0.87 (0.77, 0.97)	0.01	0.95 (0.7, 1.28)	0.47	1.01 (0.87, 1.17)	0.93	1.13 (0.89, 1.44)	0.31	0.92 (0.69, 1.22)	0.55
Redundant prepuce, phimosis and paraphimosis	0.89 (0.80, 0.98)	0.02	1.69 (1.31, 2.17)	4.5×10^{-5}	0.94 (0.81, 1.08)	0.37	0.99 (0.81, 1.2)	0.91	1.04 (0.80, 1.34)	0.79
Balanitis xerotica obliterans	0.84 (0.72, 0.99)	0.03	1.64 (1.04, 2.59)	0.03	0.91 (0.74, 1.11)	0.35	0.96 (0.7, 1.3)	0.77	0.85 (0.59, 1.21)	0.37
Inflammatory disorder including orchitis and epididymitis	0.85 (0.74, 0.99)	0.03	1.63 (1.13, 2.36)	9.5×10^{-3}	1.04 (0.86, 1.26)	0.71	0.93 (0.70, 1.23)	0.61	1.22 (0.87, 1.72)	0.25

Supplementary Table 7.2 (Continued)

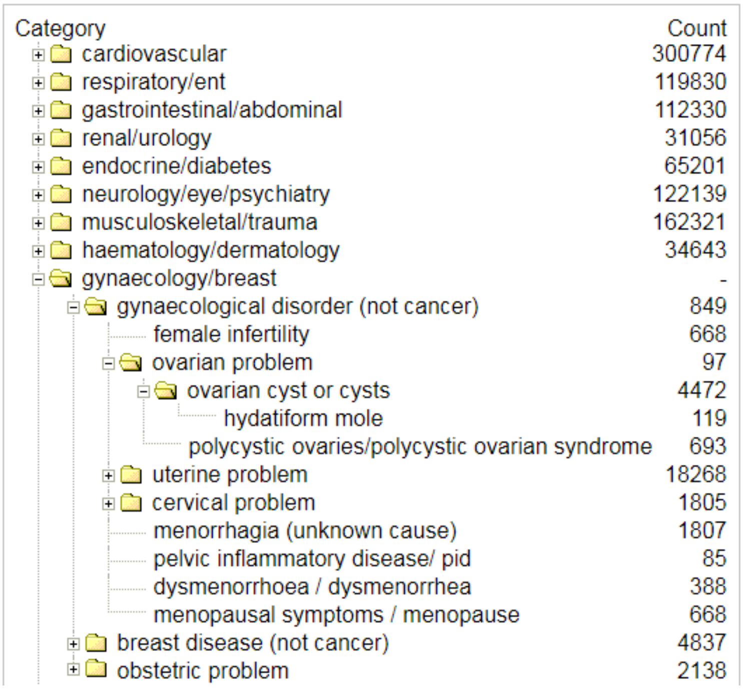
Trait	AAM		BMI		Total-T		Free-T		SHBG	
	OR (95% CI)	P-value	OR (95% CI)	P-value	OR (95% CI)	P-value	OR (95% CI)	P-value	OR (95% CI)	P-value
Other congenital malformations of male or female genital organs	0.71 (0.46, 1.10)	0.12	0.62 (0.19, 2.04)	0.4	0.92 (0.52, 1.64)	0.78	0.80 (0.32, 2.00)	0.63	0.81 (0.29, 2.24)	0.68

OR = odds ratio; AAM = age at menarche; BMI = body mass index; Total-T = total testosterone; Free-T = free (biologically active) testosterone; SHBG = sex hormone binding globulin



Category	Count
Chapter I Certain infectious and parasitic diseases	63942
Chapter II Neoplasms	220470
Chapter III Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism	52184
Chapter IV Endocrine, nutritional and metabolic diseases	185543
Chapter V Mental and behavioural disorders	78058
Chapter VI Diseases of the nervous system	77454
Chapter VII Diseases of the eye and adnexa	105959
Chapter VIII Diseases of the ear and mastoid process	19993
Chapter IX Diseases of the circulatory system	429280
Chapter X Diseases of the respiratory system	156726
Chapter XI Diseases of the digestive system	534892
Chapter XII Diseases of the skin and subcutaneous tissue	72978
Chapter XIII Diseases of the musculoskeletal system and connective tissue	391505
Chapter XIV Diseases of the genitourinary system	274126
Chapter XV Pregnancy, childbirth and the puerperium	-
O00-O08 Pregnancy with abortive outcome	-
O00 Ectopic pregnancy	-
O00.1 Tubal pregnancy	269
O00.2 Ovarian pregnancy	4
O00.8 Other ectopic pregnancy	6
O00.9 Ectopic pregnancy, unspecified	98

Supplementary Figure 6.1: Hierarchical structure of ICD-10 coded diagnoses for disease categorisation. Example for ectopic pregnancies, showing four levels of increasing specificity.



Category	Count
cardiovascular	300774
respiratory/ent	119830
gastrointestinal/abdominal	112330
renal/urology	31056
endocrine/diabetes	65201
neurology/eye/psychiatry	122139
musculoskeletal/trauma	162321
haematology/dermatology	34643
gynaecology/breast	-
gynaecological disorder (not cancer)	849
female infertility	668
ovarian problem	97
ovarian cyst or cysts	4472
hydatiform mole	119
polycystic ovaries/polycystic ovarian syndrome	693
uterine problem	18268
cervical problem	1805
menorrhagia (unknown cause)	1807
pelvic inflammatory disease/ pid	85
dysmenorrhoea / dysmenorrhea	388
menopausal symptoms / menopause	668
breast disease (not cancer)	4837
obstetric problem	2138

Supplementary Figure 6.2: Self-reported illness classifications in UK Biobank Data Showcase. Example showing polycystic ovarian syndrome.