Longitudinal multi-dimensional investigation of metabolic and endocrine genetics

Samvida Sudheesh Venkatesh

University College
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy

Michaelmas 2023

Abstract

Genome-wide association studies (GWASs) in recent decades have revealed the genetic landscape and shared aetiology of common, complex traits across the spectrum of human phenotypes. In this work, I develop and apply statistical tools to interrogate the genetic basis of, and relationships between, metabolic and endocrine traits. I demonstrate that under-explored primary care electronic health records (EHRs), linked to massive biobank projects across the globe, are a valuable source of longitudinal and rare biomarker data for genetics studies. Using EHRs, I find a common missense variant in the *APOE* gene that is associated with weight-loss in adulthood, which replicates in three global biobanking cohorts of between 125,000 to 475,000 individuals each. While the heritability of weight-change is low ($< 3\%$), it has a distinct genetic architecture to baseline weight (genetic correlation, $r_G = 39\%$). Next, I perform GWASs for female and male infertility, as well as sex-specific reproductive hormones (with $> 700{,}000$ participants across seven global biobanks), to characterise the genetic contributions to these common but poorly understood phenotypes. I find 21 unique genetic loci for infertility, of which only six colocalise with reproductive hormone levels. While there is modest correlation between female infertility and heritable diseases of the reproductive tract, such as endometriosis ($r_G = 58\%$) and polycystic ovary syndrome (PCOS) ($r_G = 40\%$), I find no evidence for metabolic conditions such as obesity in the genetic aetiology of infertility. I explore these findings further through Mendelian Randomisation analyses to reveal heterogeneity in the genetically predicted causal effects of overall and central obesity on the risk of female reproductive conditions, including infertility, endometriosis, and PCOS, which may be partly genetically mediated by hormone levels. Through a range of genetics-based investigations, I outline the shared and distinct mechanisms of metabolic and endocrine disease in humans.

Longitudinal multi-dimensional investigation of metabolic and endocrine genetics



Samvida Sudheesh Venkatesh

University College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Michaelmas 2023

Acknowledgements

Personal

This thesis was made possible through the dedicated supervision of Professor Cecilia Lindgren and Professor Chris Holmes, who are both models of scientific passion and perseverance, and who have brought together teams of talented, generous, and kind scientists that have all inspired my academic journey. Thanks also to Professor Ben Neale for hosting me for a summer at the Broad Institute of Harvard and MIT in the United States. I am particularly grateful to senior mentors George, Laura, Saskia, Teresa, Duncan, Chris, and Hannah, who have guided this thesis in some form, and to fellow DPhil students Fred and Nik for their camaraderie.

It takes a village, and mine begins with my parents and sister, whose influence extends well beyond their critical feedback on colour schemes for this thesis. Thank you Amma, Appa, and Sampada, for your unwavering support. I also owe a huge debt of gratitude to extended family and friends across the globe, who have opened up their homes to me - it means the world.

My deepest thanks to friends in Oxford who have been a constant source of cheer: the Rhodes scholar community for engaging discussions; fellow DPhil students in the Genomic Medicine and Statistics program for the lunchtime coffees, chocolate, and chat; flatmates past and present, especially Anna, who has shared my trials and tribulations for over three years; and members of the Oxford University Badminton Club, particularly my (mixed doubles) partner Matthew. You have all enriched my Oxford experience, as I hope I have yours.

Institutional

I am blessed to have completed the DPhil under the auspices of several institutions, including the Rhodes Scholarship, Clarendon Fund at the University of Oxford, University College Oxford, and the Wellcome Trust. Beyond the communities of students nourished by each of these groups, I am thankful for the financial resources that enabled me to dedicate myself wholly to the pursuit of the DPhil and its dissemination at conferences and workshops around the world.

Previously published material

I feel deeply grateful for having had the chance to work with talented and dedicated collaborators on several projects during the course of my DPhil. Some of the work presented in this thesis has been developed with the input of such collaborators, and has resulted in peer-reviewed publications and article pre-prints. As scientific ideas and output are often generated collaboratively, I use the pronouns "I" and "we" inter-changeably in this thesis.

Throughout the thesis, any data processing or quality control steps that have been performed by others, or for which I amended scripts originally produced by others, are acknowledged as such. The work presented in this thesis, including all the code and analyses to generate text, figures, and tables are a result of my original work, except as specified in the text.

Chapters 2, 3, and 5 are adapted from sections of: my first-authored pre-print *The genetic architecture of changes in adiposity during adulthood*[1] and my first-authored journal publication *Obesity and risk of female reproductive conditions: A Mendelian Randomisation Study*[2]. I have received written permission from editors of the respective journals to reproduce published figures in part or in full in my thesis. I have also received written confirmation from all co-authors of each publication and my supervisor, Professor Cecilia Lindgren, to reuse parts of previously published text in my thesis.

Abstract

Genome-wide association studies (GWASs) in recent decades have revealed the genetic landscape and shared aetiology of common, complex traits across the spectrum of human phenotypes. In this work, I develop and apply statistical tools to interrogate the genetic basis of, and relationships between, metabolic and endocrine traits. I demonstrate that under-explored primary care electronic health records (EHRs), linked to massive biobank projects across the globe, are a valuable source of longitudinal and rare biomarker data for genetics studies. Using EHRs, I find a common missense variant in the *APOE* gene that is associated with weight-loss in adulthood, which replicates in three global biobanking cohorts of between 125,000 to 475,000 individuals each. While the heritability of weight-change is low ($< 3\%$), it has a distinct genetic architecture to baseline weight (genetic correlation, $r_G = 39\%$). Next, I perform GWASs for female and male infertility, as well as sex-specific reproductive hormones (with $> 700{,}000$ participants across seven global biobanks), to characterise the genetic contributions to these common but poorly understood phenotypes. I find 21 unique genetic loci for infertility, of which only six colocalise with reproductive hormone levels. While there is modest correlation between female infertility and heritable diseases of the reproductive tract, such as endometriosis ($r_G = 58\%$) and polycystic ovary syndrome (PCOS) ($r_G = 40\%$), I find no evidence for metabolic conditions such as obesity in the genetic aetiology of infertility. I explore these findings further through Mendelian Randomisation analyses to reveal heterogeneity in the genetically predicted causal effects of overall and central obesity on the risk of female reproductive conditions, including infertility, endometriosis, and PCOS, which may be partly genetically mediated by hormone levels. Through a range of genetics-based investigations, I outline the shared and distinct mechanisms of metabolic and endocrine disease in humans.

Contents

List of Figures **xi**

List of Tables **xv**

1 Introduction **1**
1.1 A brief history of human population genetics 1
1.2 Genome-wide association studies for genetic discovery 5
1.2.1 Statistical methods empowering GWASs and their interpretation 7
1.2.2 Phenotyping advances for improved GWASs 12
1.3 Metabolic and endocrine systems in humans 15
1.3.1 Genetic basis of obesity 18
1.3.2 Reproductive trait genetics 20
1.4 Thesis aims 23

2 Curation of a longitudinal resource for genetic studies in the UK Biobank **25**
2.1 Introduction to UK Biobank 26
2.1.1 Deep phenotyping 26
2.1.2 Follow-up assessments 27
2.1.3 Health-record linkage 27
2.1.4 Genetic data 28
2.2 QC of longitudinal phenotypes in UKBB-linked EHR 30
2.3 QC of genetic data 37
2.4 Genome-wide association studies 38
2.5 Post-GWAS 39
2.5.1 Lead variant identification 39
2.5.2 Lead variant classification 39
2.5.3 Genetic correlations and SNP-based heritability 40
2.5.4 Testing sex heterogeneity in genetic effects 41
2.6 Exome-wide association studies 41
2.6.1 Exome sequencing quality control 42
2.6.2 Variant annotation 42
2.6.3 Rare variant and gene-based association testing 43

3 The genetic architecture of changes in adiposity during adulthood 45
3.1 Introduction . . . 46
3.2 Methods . . . 47
3.2.1 Models to define change in adiposity over time . . . 48
3.2.2 Genome-wide association studies . . . 61
3.2.3 Sensitivity analyses for survival bias . . . 62
3.2.4 Replication of genetic associations in the UK Biobank . . . 64
3.2.5 Replication of genetic associations in the Million Veteran Program . . . 66
3.2.6 Replication of genetic associations in the Estonian Biobank . 67
3.2.7 Power calculations for replication sample sizes . . . 69
3.2.8 Statistical power comparison to published GWASs . . . 70
3.2.9 Longitudinal phenome-wide association studies . . . 71
3.2.10 Joint modelling of intra-individual mean and variance . . . 71
3.3 Results . . . 72
3.3.1 Characteristics of baseline adiposity and adiposity-change . . 72
3.3.2 Genetic architecture of adiposity-change in adulthood . . . 74
3.3.3 *APOE* variant rs429358 is associated with multiple adiposity-change phenotypes . . . 85
3.3.4 Genetic associations with adiposity-change are robust to a range of modelling choices . . . 91
3.3.5 Longitudinal data enable the identification of novel genetic signals for baseline adiposity . . . 96
3.3.6 No evidence for genetic basis of within-individual variability in weight . . . 103
3.4 Discussion . . . 107

4 The distinct genetic components of infertility and reproductive hormones 113
4.1 Introduction . . . 114
4.2 Methods . . . 115
4.2.1 GWASs of infertility and reproductive hormones in UKBB . 115
4.2.2 Meta-analysis of GWAS summary statistics . . . 116
4.2.3 Genetic correlations . . . 122
4.2.4 Colocalisation . . . 123
4.2.5 Overlaps with genetic regions under selection . . . 124
4.2.6 Tissue and cell-type prioritisation . . . 126
4.2.7 Whole exome sequencing analyses in UKBB . . . 127
4.3 Results . . . 128

4.3.1 Genome-wide meta-analyses identify novel genetic loci for female and male infertility . . . 128
4.3.2 Selection pressure may explain the persistence of some infertility-associated variants in the population . . . 135
4.3.3 Genetic determinants of reproductive hormone levels . . . 139
4.3.4 Non-overlapping genetics of hormones and infertility . . . 151
4.3.5 Phenome-wide genetic correlation landscape differs for infertility and reproductive hormones . . . 153
4.3.6 Heritability of testosterone is enriched in adrenal gland, liver, and ovarian cell types . . . 154
4.3.7 Rare variant contribution to reproductive hormone and infertility genetics . . . 157
4.4 Discussion . . . 165

5 Genetically causal relationships of obesity with female reproductive diseases 175
5.1 Introduction . . . 176
5.2 Methods . . . 177
5.2.1 Observational associations in UK Biobank . . . 177
5.2.2 Two-sample Mendelian Randomisation . . . 179
5.2.3 Non-linear Mendelian Randomisation . . . 185
5.2.4 Mediation Mendelian Randomisation . . . 186
5.2.5 Disease and variant clustering . . . 187
5.3 Results . . . 188
5.3.1 Obesity is associated with altered risk of female reproductive conditions in the UK Biobank . . . 188
5.3.2 Body fat distribution is genetically associated with altered risk of female reproductive conditions . . . 199
5.3.3 Hormones mediate the effects of genetically predicted obesity on female reproductive disorders . . . 207
5.3.4 Mechanisms by which different forms of obesity affect the risk of various female reproductive diseases . . . 216
5.4 Discussion . . . 218

6 Conclusions and Future Directions 227
6.1 Conclusions . . . 227
6.2 Future Directions . . . 229
6.2.1 Leveraging electronic health records for genetics research . . . 229
6.2.2 Interpretation of coding genetic variation . . . 233
6.2.3 Deciphering gene regulation by the non-coding genome . . . 237
6.2.4 The future of human health . . . 240

Appendices

A Supplement to Chapter 2 **245**

B Supplement to Chapter 3 **253**

C Supplement to Chapter 4 **267**

D Supplement to Chapter 5 **285**

Bibliography **295**

List of Figures

1.1 Where Do We Come From? What Are We? Where Are We Going? 2
1.2 Typical workflow for genome-wide association studies 8
1.3 Endocrine regulation by the hypothalamic-pituitary-target organ axis 16

3.1 Selection of discovery and validation datasets from UKBB-linked primary care and assessment centre records 48
3.2 QQ plots for the distribution of residuals and coefficients estimated from LME models of change in BMI over time 50
3.3 Comparison of hyperparameters for $AR(1)$ model as smoothing prior on spline coefficients 53
3.4 Comparison of cluster allocations across choice of $AR(1)$ model hyperparameters 54
3.5 Overview of clustering protocol 56
3.6 Cluster centroid trajectories for various combinations of filtering and initialisation parameters 58
3.7 Silhouette values for selection of K, L, and M clustering parameters 59
3.8 Comparison of cluster allocations across each of 10 random splits of training data 60
3.9 Workflow for carrying forward estimates from adiposity-change models to GWAS 62
3.10 Trajectories of adiposity-change modelled by linear and high-dimensional B-splines in a random sample of individuals. 75
3.11 Random sample of individuals with high probability of assignment to each adiposity-change cluster 76
3.12 Effect sizes of rs429358 on BMI-change slope in discovery (UKBB) and replication (MVP and EstBB) datasets 82
3.13 Heritability of, and genetic correlation between, adiposity traits . . 83
3.14 Comparison of effect sizes of GWS SNPs in main LME slope GWASs and sensitivity analyses unadjusted for LME intercept 84
3.15 Association of *APOE* SNP rs429358 with various adiposity-change phenotypes in UKBB 86
3.16 Phenome-wide associations of rs429358 with diseases in UKBB . . . 89

3.17 Comparison of rs429358 effects on adiposity-change phenotypes in held-out participants in UKBB and sensitivity analysis in held-out participants in UKBB without dementia diagnoses 90
3.18 Longitudinal phenome-wide effects of rs429358 on changes in biomarker levels in UKBB 92
3.19 Comparison of rs429358 associations with baseline adiposity and adiposity-change in main analyses and sensitivity analyses unadjusted for age 93
3.20 Correlation between follow-up metrics and intra-individual variance in adiposity 94
3.21 Comparison of effect sizes of GWS SNPs in main adiposity-change GWASs and sensitivity analyses that are unadjusted for follow-up metrics 95
3.22 Phenotypic and genetic correlations between adiposity-change metrics 97
3.23 Genome-wide novel and refined SNP associations with baseline adiposity estimated over the measurement time window for each individual 99
3.24 Variance in baseline adiposity explained by lead variants from GWASs 102
3.25 Comparison of effect sizes of GWS SNPs between average adiposity-trait and LME model intercept GWASs 104
3.26 Within-individual and between-individual variance in BMI and weight 105
3.27 Manhattan plots depicting SNPs associated with jointly modelled within-individual mean and variance in weight over multiple measurements 106

4.1 Manhattan plot depicting SNPs associated with all definitions of female infertility 130
4.2 Manhattan plot depicting SNPs associated with male infertility of all causes 132
4.3 Genetic correlations between female infertility and traits across the phenome 136
4.4 Trait-aligned Singleton Density Scores for directional selection at the *EBAG9* locus 139
4.5 Standardised BetaScan2 scores for balancing selection at infertility loci 140
4.6 Manhattan plot depicting SNPs associated with reproductive hormones in female-specific analyses 142
4.7 Manhattan plot depicting SNPs associated with reproductive hormones in male-specific analyses 143
4.8 Number of hormone-associated lead variants classified as novel or reported 144

4.9 Sex heterogeneity in the effects of genetic variants associated with testosterone . . . 149
4.10 Trait-aligned Singleton Density Scores for directional selection at testosterone-associated loci . . . 150
4.11 Comparison of effect sizes of SNPs associated with testosterone in main meta-analyses and sensitivity analyses without publicly available summary statistics . . . 152
4.12 Genetic correlations between female reproductive hormones and traits across the phenome . . . 155
4.13 Genetic correlations between testosterone in males and traits across the phenome . . . 156
4.14 Partitioned heritability LDSC regressions for female infertility and reproductive hormones across tissues and cell types . . . 158
4.15 Partitioned heritability LDSC regressions for female infertility and testosterone in females, in ovarian cell types . . . 159
4.16 Gene-based Manhattan plots for burden of rare variants associated with testosterone in UK Biobank . . . 160
4.17 Effect size versus allele frequency of genetic variants associated with reproductive hormones . . . 163
4.18 Effects of testosterone-associated rare variants on infertility . . . 166

5.1 Relationship between obesity traits and female reproductive conditions in UKBB . . . 194
5.2 Comparison of observational and genetically predicted causal effect sizes of obesity traits on female reproductive conditions . . . 195
5.3 Effect sizes of obesity traits on female reproductive conditions in UKBB with and without adjustment for smoking status. . . . 197
5.4 Effect of early-life body size on risk of female reproductive conditions in UKBB . . . 198
5.5 Genetically predicted causal effects of centralised obesity traits on risk of developing female reproductive conditions . . . 201
5.6 Comparison of MR results with different instrument selection strategies 205
5.7 Comparison of MR methods for estimating causal effect sizes . . . 206
5.8 Comparison of main MR results to sensitivity analysis with GIANT-only instruments . . . 208
5.9 Reciprocal MR for causal effect of female reproductive disease on levels of BMI, WHR, and WHRadjBMI . . . 210
5.10 Genetically predicted causal effects of hormones on female reproductive diseases . . . 211

5.11 Genetically mediated proportions for causal effects of obesity and hormones on female reproductive conditions 214
5.12 Principal component clustering of female reproductive conditions by shared obesity effects 217
5.13 Obesity-associated genetic variants annotated to substantial clusters 219
5.14 Heatmaps of obesity-associated SNP effect sizes on female reproductive conditions for variants belonging to "substantial" clusters . . . 220

6.1 Opportunities for EHR-based genetics research 230
6.2 Methods to assess the impact of genetic variation 234
6.3 Reinterpretation of "Where Do We Come From? What Are We? Where Are We Going?" by Paul Gauguin, generated by `DALL-E` . . 242

List of Tables

2.1 Characteristics of traits and participants in the curated UKBB-linked longitudinal research resource 32

3.1 Median residual variance from high-dimensional spline modelling of standardised obesity trait outcomes 52
3.2 Characterisation of adiposity-trait data in individuals assigned to their highest probability cluster in training and validation datasets . 60
3.3 Informed presence bias in UKBB-linked GP electronic health records 63
3.4 Comparison of cluster characteristics in UKBB, MVP, and EstBB within each analysis stratum 68
3.5 Characterisation of adiposity-trait data in longitudinal records across three biobanks 73
3.6 Lead SNPs identified from GWASs of adiposity-change 78
3.7 Sample sizes required for replication of adiposity-change lead SNPs, with effect sizes debiased for winner's curse 80
3.8 Replication of adiposity-change lead SNPs with linear slope change in BMI 81
3.9 Association of adiposity-change lead SNPs with self-reported weight change, independent of baseline BMI 87
3.10 Novel variants associated with obesity identified in GWASs for adiposity LME model intercepts in all strata 98
3.11 Sample sizes required for replication of novel baseline adiposity lead SNPs, with effect sizes debiased for winner's curse 100
3.12 Replication of novel lead SNPs for baseline adiposity 101
3.13 Comparison between LME model intercepts and average adiposity-trait across measurement window 103

4.1 Contributing cohort overview for infertility GWAS meta-analyses . 117
4.2 Contributing cohort overview for reproductive hormone GWAS meta-analyses 118
4.3 Publicly available summary statistics for reproductive hormone GWAS meta-analyses 121

4.4 List of lead variants associated with all definitions of female and male infertility . . . 131
4.5 Colocalisation of endometriosis and PCOS signals at lead variants for female infertility . . . 134
4.6 Heritability estimates for infertility and reproductive hormones . . . 137
4.7 List of infertility loci annotated with directional selection and balancing selection scores . . . 138
4.8 List of novel lead variants associated with reproductive hormones . 146
4.9 Colocalisation of female infertility and female reproductive hormone signals in loci for either set of traits . . . 153
4.10 Gene-based association test results for male testosterone levels in UKBB . . . 161
4.11 Gene-based association test results for female hormone levels in UKBB 162
4.12 Rare variants associated with altered reproductive hormone levels in UKBB . . . 164

5.1 Summary of genetic instruments for obesity-related traits, hormone mediators, and female reproductive disorders . . . 181
5.2 Summary of female reproductive disorders in UKBB . . . 189
5.3 Observational and genetically predicted causal effect of obesity traits on female reproductive conditions . . . 192
5.4 Comparison of linear and non-linear models testing the association of obesity traits with female reproductive conditions in UKBB . . . 196
5.5 Genetically predicted causal effects of centralised obesity traits on risk of developing female reproductive conditions . . . 202
5.6 Multivariable Mendelian Randomisation to estimate the effect of WHR on female reproductive disorders adjusted for BMI . . . 209
5.7 Genetically predicted causal effects of obesity traits on hormones and vice-versa . . . 213
5.8 Genetically mediated proportions for causal effects of obesity and hormones on female reproductive conditions . . . 215

A.1 List of quantitative phenotypes with corresponding CPRD codes in EHR and fields in UKBB assessment centre visits . . . 246
A.2 UKBB assessment centre minimum and maximum values of quantitative traits . . . 248
A.3 List of CPRD and ICD codes for bariatric surgery . . . 250

B.1 List of obesity-related traits in the GWAS Catalog . . . 254
B.2 List of read codes and ICD codes to determine history or diagnosis of dementia or Alzheimer's . . . 256

B.3 List of longevity-related traits in the GWAS Catalog 262
B.4 Refined variants associated with obesity identified in GWASs for adiposity LME model intercepts in all strata 263
B.5 Significant associations of longevity SNPs with adiposity-change phenotypes 265

C.1 Case definitions for the various female and male infertility analyses 268
C.2 List of hormone traits in the GWAS Catalog 270
C.3 Colocalisation of eQTLs from GTEx metabolic, endocrine, and reproductive tissues with female infertility loci 271
C.4 Lowest infertility-associated P-value for variants in genomic windows under directional selection 278
C.5 Heterogeneity in effect estimates across contributing cohorts for infertility lead variants 284

D.1 List of GWASs contributing to female reproductive condition summary statistics 286
D.2 Comparison of effect sizes of obesity traits on female reproductive conditions in UKBB with and without adjustment for smoking status 288
D.3 Power to detect causal effects using two-sample MR for each pair of obesity-related traits and female reproductive outcomes 289
D.4 Comparison of MR methods for estimating causal effect sizes 292

List of Abbreviations

3C chromosome conformation capture. 238, 239

ACTH adrenocorticotropic hormone. 15

AI artificial intelligence. 237, 239, 241

AIC Akaike information criterion. 57, 179, 191

ALSPAC Avon Longitudinal Study of Parents and Children. 13

AMH anti-Müllerian hormone. 122, 152, 153, 167, 186

AR(1) order-1 autoregressive. 51, 53, 54

BLUP best linear unbiased predictor. 49, 50, 65, 71, 74, 110

BMI body mass index, total body weight [kg]/standing height squared [m^2]. 18–20, 36, 37, 46–48, 50, 52, 61, 64–67, 69, 70, 72, 74–77, 79, 80, 82, 83, 85–87, 90, 91, 96, 99, 100, 102, 103, 105, 107–109, 154, 167, 169, 176–180, 185–188, 191, 194, 198–200, 203, 204, 207, 210, 212, 216–218, 221–225, 229, 230

bp base pair. 29

CDF cumulative distribution function. 41, 49, 55, 179

CI confidence interval. 77, 82, 85, 86, 88, 90, 92, 129, 151, 157, 164, 165, 188, 199, 200

CMS Composite of Multiple Signals. 125

COVID-19 coronavirus disease 2019. 239

CPRD Clinical Practice Research Datalink. 30, 63, 246, 247, 251

CRH corticotrophin-releasing hormone. 15

CRP C-reactive protein. 88, 110

CTV3 Clinical Terms Version 3. 28, 31

DEXA dual energy X-ray absorptiometry. 18, 19

DHEA dehydroepiandrosterone. 17, 21

DNA deoxyribonucleic acid. 3, 119, 236–239

EHR electronic health record. 13–15, 20, 23, 27, 30, 31, 36, 46–48, 64, 66, 69, 71, 72, 91, 107–109, 111, 112, 115, 116, 119, 120, 129, 141, 167, 171, 227–233, 241, 246, 247

eQTL expression quantitative trait locus. 11, 77, 124, 132, 238

EstBB Estonian Biobank. 68, 69, 72, 74, 79, 99, 100, 129, 135

FAI free androgen index. 36, 37

FAQ frequently asked question. 29

FDR false discovery rate. 126, 154, 157–159, 191, 199, 204, 207, 210, 212

FSH follicle stimulating hormone. 15, 21, 22, 36, 114–116, 120, 122, 124, 127, 141, 145, 154, 157, 159, 165, 171, 172, 186, 207, 210–212, 216, 228, 231

FWER family-wise error rate. 9, 64, 70, 79, 88–92, 94, 97, 122, 123, 128, 136, 152, 153, 155, 156, 160, 164, 165

GAM generalised additive model. 179

GIANT Genetic Investigation of ANthropometric Traits. 70, 103, 180, 185, 186, 200, 203, 204

GLP-1 glucagon-like-peptide-1. 69, 111

GnRH gonadotropin-releasing hormone. 15

GP general practitioner. 27, 48, 64, 72, 91, 227, 230, 231

GRM genetic relatedness matrix. 10, 43

GSA Global Screening Array. 69, 119, 120

GTEx Genotype-Tissue Expression. 11, 124, 126, 132, 154, 158, 170, 234

GWAS genome-wide association study. 5–7, 9–13, 18, 19, 21–23, 38–41, 46, 50, 57, 61–64, 66–70, 74, 79, 82, 85, 88, 91, 93, 96, 97, 99, 102, 103, 108, 110, 111, 115, 119–123, 125, 126, 128, 129, 135, 139, 141, 144, 145, 149, 151, 152, 157, 160, 163, 164, 168, 172, 177, 179, 180, 183–186, 200, 204, 205, 207, 211, 212, 216, 221, 225, 228–230, 233, 238–240

GWS genome-wide significant. 39, 70, 74, 82, 84, 93, 95, 96, 102–105, 125, 130, 139, 141–143, 149, 151, 172, 180, 185–187

HbA1c haemoglobin A1C / glycated haemoglobin. 27, 37

HC hip circumference. 64, 154, 169, 177, 180, 200, 203

HDL high-density lipoprotein. 77, 88, 110

HMB heavy menstrual bleeding. 122, 178, 183, 185, 191, 196, 198–200, 203, 222, 223

HRC Human Reference Consortium. 29

HWE Hardy-Weinberg equilibrium. 37

ICD International Classification of Diseases. 63, 116, 251

indels insertion, deletion, or insertion and deletion of nucleotides. 29, 30, 42, 43

IRB Institutional Review Board. 66

IVF *in vitro* fertilisation. 167

IVW inverse variance weighted. 121, 184, 186, 204

kb kilobase (one kilobaseis one thousand base pairs). 40, 64, 123, 125, 126, 134, 138–140, 150, 151, 153

LD linkage disequilibrium. 3–5, 9, 10, 12, 38–40, 64, 94, 96, 102, 122, 126–128, 136, 141, 144, 158, 159, 162, 180

LDL low-density lipoprotein. 6, 27

LH luteinising hormone. 15, 21, 36, 114–116, 120, 124, 127, 141, 145, 157, 165, 171, 172, 186, 228, 231

LME linear mixed-effects. 47, 61, 62, 64, 65, 67, 69–71, 74, 79, 80, 82, 85, 91, 94, 96, 102, 103, 108, 110

MAC minor allele count. 43

MAF minor allele frequency. 29, 30, 37, 39, 42, 43, 71, 74, 77, 85, 121, 123, 124, 128, 157, 159, 161, 164, 165, 184

MANE Matched Annotation from NCBI and EMBL-EBI. 42

Mb megabase (one megabase is one million base pairs). 39, 40, 64, 97, 124, 141, 144, 180, 185, 187, 271

MLE maximum likelihood estimate. 48, 52

MR Mendelian Randomisation. 12, 176, 177, 179, 180, 184–187, 200, 204, 207, 211, 212, 214, 221

mRNA messenger RNA. 237, 238

MVP Million Veteran Program. 12, 66, 67, 72, 74, 79

OR odds ratio. 77, 85, 88, 90, 129, 132, 151, 157, 164, 165, 184, 188, 191, 194, 196, 198–200, 203, 207, 210, 212

PAM Partitioning Around Medoids. 55

PC principal component. 38, 43, 61, 62, 65, 67, 69, 71, 84, 95, 104, 116, 185, 187, 216, 217

PCA principal component analysis. 28, 187, 216

PCOS polycystic ovary syndrome. 18, 21, 22, 114, 116, 122, 124, 128, 133, 135, 165, 167, 168, 176–178, 183, 188, 191, 194, 196, 198–200, 207, 210, 212, 216, 222–225, 228, 229, 232, 241

PheWAS phenome-wide association study. 11

pLoF putative loss-of-function. 30, 42, 43, 157, 162

PP posterior probability. 124, 132, 133, 151

pQTL protein quantitative trait locus. 11

QC quality control. 14, 28–31, 36, 37, 41, 42, 47, 64, 66, 68, 71, 108, 115, 119, 126, 127, 151, 177, 184

QTL quantitative trait locus. 11

SAT subcutaneous adipose tissue. 17

SD standard deviation. 31, 37, 51–54, 72, 77, 85, 87, 88, 109, 177, 188, 199, 200, 207

SDS Singleton Density Score. 125, 137, 139, 149

SE standard error. 56, 77, 79, 80, 85, 87, 88, 141, 149, 151, 207, 212

SHBG sex hormone binding globulin. 27, 77, 160

SLE systemic lupus erythematosus. 232

SNP single nucleotide polymorphism. 4, 7, 9, 10, 12, 18, 21, 22, 29, 37–40, 47, 64, 70, 74, 84, 85, 88, 93, 95, 96, 99, 100, 102–104, 107, 121, 125, 126, 128, 130, 132, 136, 142, 143, 145, 180, 187, 188, 216, 217, 233, 236

SNV single nucleotide variant. 30, 42, 160

T3 triiodthyronine. 15, 17

T4 thyroxine. 15, 17

TES transcription end site. 127, 128

TF transcription factor. 237

TRH thyrotropin-releasing hormone. 15

TSH thyroid-stimulating hormone. 15, 122, 152, 186

TSS transcription start site. 124, 127, 128

UK United Kingdom. 12, 26, 72, 103, 171, 204, 225, 228, 240

UKBB UK Biobank. 12, 14, 21, 26–31, 36–38, 43, 47–49, 62, 64–69, 71, 72, 74, 77, 79, 85, 88, 91, 97, 99, 100, 103, 106–111, 115, 116, 120, 122, 123, 127–129, 136, 153, 155–157, 161, 163, 165, 170–172, 177–180, 183, 185, 186, 188, 191, 195, 197–200, 203, 204, 207, 212, 221–223, 225, 227, 230, 246, 247, 249

USA United States of America. 12, 66, 72, 228

VA Veterans Affairs. 66

VAT visceral adipose tissue. 17, 18, 180, 200, 223, 224

VHA Veterans Health Administration. 66

WC waist circumference. 64, 65, 87, 154, 169, 177, 180, 200, 203

WCadjBMI WC adjusted for BMI. 64, 65, 87

WES whole-exome sequencing. 29, 30, 128, 157, 160, 161, 163, 165, 170, 171

WHR waist-to-hip ratio, waist circumference/hip circumference. 19, 36, 64, 65, 87, 177–180, 185–188, 191, 196, 198, 199, 203, 204, 207, 210, 212, 217, 218, 229

WHRadjBMI WHR adjusted for BMI. 19, 64, 65, 87, 177–180, 185–188, 191, 196, 199, 204, 207, 210, 212, 216–218, 221–224

1 Introduction

Contents

1.1 A brief history of human population genetics 1
1.2 Genome-wide association studies for genetic discovery 5
1.2.1 Statistical methods empowering GWASs and their interpretation 7
1.2.2 Phenotyping advances for improved GWASs 12
1.3 Metabolic and endocrine systems in humans 15
1.3.1 Genetic basis of obesity 18
1.3.2 Reproductive trait genetics 20
1.4 Thesis aims 23

1.1 A brief history of human population genetics

For millennia, our species has meditated upon a deeply existential question: where do we come from? Philosophers have long grappled with the concepts of time and reason to provide answers to our origins that range from the rational to the absurd. Keen observers quickly realised that like begets like, but initial attempts to apply the scientific method to unravel the question of inheritance had mixed success[3]. In the early 1700s, scientists speculated that fully pre-formed "homunculi" (little men) in sperm or ova unfolded into offspring[4]. In the early 1800s, Lamarck hypothesised that the movement of fluids generated simple organisms from non-living gelatinous



Figure 1.1: Where Do We Come From? What Are We? Where Are We Going? Reproduction of painting by Paul Gauguin, 1897, Original Title: D'ou Venons Nous / Que Sommes Nous / Où Allons Nous. The original painting can be found in the Museum of Fine Arts, Boston.

substances, with accelerating fluid motion generating more complex animals[5]. In the 1900s, Darwin and others built upon Hippocrates' theory of "pangenesis", which suggested that male and female seminal fluids were blended in equal proportions to produce offspring that somewhat resembled both parents[6]. Centuries later, geneticists now know where we came from: our parents, who came from theirs, who came from theirs, and so on, until we trace the lineage of humanity back to 60,000 years ago in Africa[7]. What we are, and where we are going, however, are open questions for the philosophers and geneticists of the future (Figure 1.1).

Our understanding of inheritance today is greatly shaped by Gregor Mendel, who performed hybridisation experiments with pea plants in the 1860s[8]. Remarkably, Mendel chose to study a set of dichotomous, un-linked, monogenic traits, such as tall or short plants, plants with purple or white flowers, whose crosses produced simple ratios of phenotypes in offspring[9]. Only thus could he mathematically define the three laws of inheritance, the mechanistic bases of which were illuminated a century later: "dominance", i.e. the dominance of one allele (or one form of a genetic variant) over another[10,11], "independent assortment", i.e. genes are un-linked and thus traits that are affected by independent genes are also un-linked[12], and "random segregation", arising from the separation of gene-carrying chromosomes during gametogenesis[13]. While aspects of these laws have since been shown incorrect[14,15],

it is nonetheless astonishing that the rules of inheritance, and indeed the molecular basis of life itself, encoded by deoxyribonucleic acid (DNA), are universal. So well do Mendel's laws generalise, that monogenic human diseases with dominant or recessive modes of inheritance are referred to as "Mendelian" traits, despite our 1.6 billion years of evolutionary distance to Mendel's peas[16].

However, the vast majority of human traits are not inherited following simple Mendelian laws, which is particularly obvious in the case of quantitative traits, such as height and weight, that vary continuously in the population[17]. Around the 1900s, biometricians began to recognise that some quantitative traits were correlated within families[18], and within a few decades, Ronald Fisher hypothesised the "infinitesimal" model to reconcile dichotomous Mendelian inheritance with quantitative trait variation[19]. Fisher proposed that an infinitely large number of discrete heritable differences, with infinitely small effects, could add up to produce the observed distribution of a quantitative trait in a population[19]. The term "gene", to describe a unit of heredity, had been coined in the intervening period between Mendel and Fisher, but Fisher made no effort to identify the causal genes that were the individual units of inheritance for any trait. He noted, however, that a statistical approach using family pedigrees could be applied to isolate these genetic differences[20]. Exactly such a method, called "linkage analysis", would become commonly employed in the following decades to identify trait-associated regions of the genome[21,22].

For a statistical geneticist, perhaps the most important deviation from Mendel's laws is that genes (and hence genetic variation) do not assort independently, but are physically linked on chromosomes[13]. There are over ten million common variant sites (with allele frequency $\geq 5\%$) in the human genome[23], and while it is now possible to sequence each of these[24,25], linkage analysis in family pedigrees was the only tractable method for genetic mapping through the 1990s and 2000s. This method exploits the phenomenon of linkage disequilibrium (LD), which arises due to co-segregation of blocks of the genome (due to evolutionary forces, such as drift, mutation, and selection), and is broken down by recombination[26]. LD decays with physical distance on the chromosome, and the larger the effective population size,

the weaker is the LD for a given distance due to more recombination events[27]. Thus, while it was (and still remains) challenging to pinpoint a specific mutation or gene altering disease risk, geneticists could exploit LD to identify a region of the genome (constituting a genomic "locus") that segregated with the disease[28].

Linkage mapping revolutionised medical genetics by successfully identifying genomic loci, and often even specific genes, where genetic variation caused a rare disorder with a Mendelian inheritance pattern[29,30]. However, the method proved more difficult for unpicking the genetic basis for common diseases and continuous traits[31,32]. For one, scientists needed to initially identify a genomic region of interest (or a set of "candidate genes") that could be further fine-mapped through linkage analysis, but the majority of candidate genes nominated through prior biological assumptions or assortment in family pedigrees have since failed to replicate[33–35]. Second, it is challenging to prove the causal effect of a genetic variant on a systemic trait: experimental evidence of a molecular phenotype, in cell lines or animal models, is not proof of systemic disease risk, and variants act in the context of their genetic backgrounds[36]. Thus, "causality" may pragmatically be defined as a variant or gene having a direct or indirect functional effect on the phenotype[36]. Finally, the penetrance of any single variant (i.e. effect size) is low, because many variants with small effects act in tandem to produce a phenotype; mapping the effect of a single nucleotide polymorphism (SNP) would thus need very large population studies[36,37].

Even so, it was not immediately obvious from Fisher's infinitesimal model that the majority of common, complex human traits would have mostly common polygenic architectures[36] - the first attempted genome-wide linkage scans for common diseases hypothesised that 15 loci may be causal for a trait[38]. While that number seemed high at the time[36,38], we now know that hundreds of genes are associated with anthropometric traits such as height (7,209 independent loci)[39] and body mass index (BMI) (941 independent loci)[40], and common diseases including type 2 diabetes (237 independent loci)[41] and coronary artery disease (262 independent loci)[42]. Moreover, theoretical models and empirical studies demonstrate that the majority of the heritability of complex diseases arises from common variation[43], so while rare variants

are individually important for informing disease risk, common variation is more informative of trait distribution at the population level. The limitations of linkage mapping in pedigrees under this scenario, where thousands of common variants have small effects, were soon overcome by the advent of genome-wide association testing, powered by large, population-based samples of unrelated individuals.

1.2 Genome-wide association studies for genetic discovery

Uncovering the polygenic architecture of common disease requires an unbiased scan of the complete genome, testing for differences in allele frequencies of genetic variants between disease cases and healthy controls in a large population[44]. This approach to associate genotypes with phenotypes at scale, termed a "genome-wide association study" (GWAS), was accelerated by three key advances: first, the sequencing of the entire human genome through the Human Genome Project in 2003, which produced a catalogue of common genetic variation[45]; second, the genome-wide quantification of LD, spearheaded by the International HapMap Project[46]; and third, the exponential decline in cost of whole-genome genotyping[47], which allowed for "typing" a small proportion of variants to then impute all tagged common variation through LD - indeed, in European populations, a selection of about 50,000 variants spread across the genome is sufficient to tag all ten million common genetic variants[46,48]. The first large-scale genome-wide association study (GWAS), published by the Wellcome Trust Case Control Consortium in 2007, associated 24 genes with seven diseases[49] and inspired a wave of large-scale population-based studies to determine the genetic architecture of common complex traits.

A quarter of a century later, GWASs have been conducted for thousands of human traits and diseases, in millions of individuals, revealing a genetic landscape that replicates across populations[50]. As of 8 November 2023, the GWAS Catalog (a curated database of published GWASs) reports nearly 300,000 genetic associations for approximately 30,000 unique phenotypes[50]. The public availability of summary statistics, i.e. variant allele frequencies, effect sizes, and significance P-values, has

enormously aided the validation of genetic findings and empowered meta-analyses to improve the precision of variant effect estimates[51,52]. To date, 542 studies have identified associations of the *FTO* gene with obesity-associated traits, and 333 studies report associations between the *TCF7L2* gene and type 2 diabetes[50]. This is a remarkable replication rate, and exemplifies the promise of unbiased genome-wide scans to deliver robust, reproducible associations between genotypes and phenotypes.

Beyond simply associating genes with traits, GWASs follow linkage mapping as the next revolution in medical genetics. Variant effect sizes estimated through GWASs are used to construct polygenic scores, which on the simplest level sum up the small effects of thousands of common trait-associated variants to produce a single genetic "risk" score, which can predict an individual's genetic risk for breast cancer[53], type 2 diabetes[41], hypertension[54], and rheumatoid arthritis[55], among other common diseases[56]. Clinical implementations of polygenic scores demonstrate that, when evaluated in combination with established clinical biomarkers, family history, and lifestyle factors, these scores stratify high-risk patients for whom early monitoring and treatment may vastly improve health outcomes[57]. Genes prioritised by GWASs are twice as likely to be approved drug targets[58] - the best-known example being *PCSK9*, the target of cholesterol-lowering statin medications, which was among the first genes to be associated with naturally lower levels of low-density lipoprotein (LDL) cholesterol in humans[59]. An additional 40 genes identified in GWASs are approved drug targets[60], and between 2-5% of drugs currently under clinical trial have GWAS evidence[61]. Finally, GWASs have enabled us to identify genetic correlations between phenotypes, pointing to shared aetiology between diseases[62], and empowered the use of genetic instrument variables for causal inference[63]. The latter, an approach termed "Mendelian Randomisation", has been especially valuable in identifying disease-causing traits, such as smoking, where intervention (behavioural, pharmacological, or otherwise) may be beneficial for patients[63].

A typical GWAS workflow begins with establishing a well-powered cohort with appropriate phenotyping, such as through questionnaires or linkage to medical records, and sample collection for genetic analysis (Figure 1.2)[64]. Following

genotyping, SNPs are imputed based on haplotype patterns in an ancestry-matched reference panel[48]; more recently, whole exome or whole genome sequencing can directly sequence genetic variants which are then aligned to a reference genome[24] (for details, see Chapter 2). Genetic association testing is usually performed with linear or logistic regression models in specially designed software for computational efficiency[65–67], and optionally, researchers will seek replication of their findings in an external cohort with no overlapping participants[64]. At this stage, summary statistics from multiple cohorts may be meta-analysed to allow narrowing down of effect sizes due to increased power from increased sample sizes[68]. Finally, findings are interpreted through a variety of downstream analyses, such as: quantifying trait heritability, fine-mapping of causal variants, gene prioritisation, identification of cell types and tissues of interest, and investigating shared genetic aetiology with other phenotypes[64]. Each of these steps will be discussed further in the following Sections.

1.2.1 Statistical methods empowering GWASs and their interpretation

The need to store and manipulate large-scale genomic data has inspired breakthroughs in computational and statistical methods in the past decade[70,71]. For genome-wide association studies, it is crucial to tractably perform hundreds of thousands of linear or logistic regressions in order to estimate the effect of each genetic variant on the trait of interest. Furthermore, GWASs need to account for ancestry and relatedness, i.e. population stratification, as these may induce false results[72]. As standard linear regressions assume independent and identically distributed random variables[73], the presence of related individuals who share genotypes can skew GWAS results and induce over-estimates of heritability[74]. Software packages to perform genome-wide association testing, including `BOLT-LMM`[66], `SAIGE`[67], and `REGENIE`[65], all implement linear mixed-effects models to account for genetic relatedness. The latter two implementations also enable association tests of rare variants with low minor allele counts, and binary traits with imbalanced case-to-control ratios, by using the saddle-point approximation to control for inflated false positive rates[75].

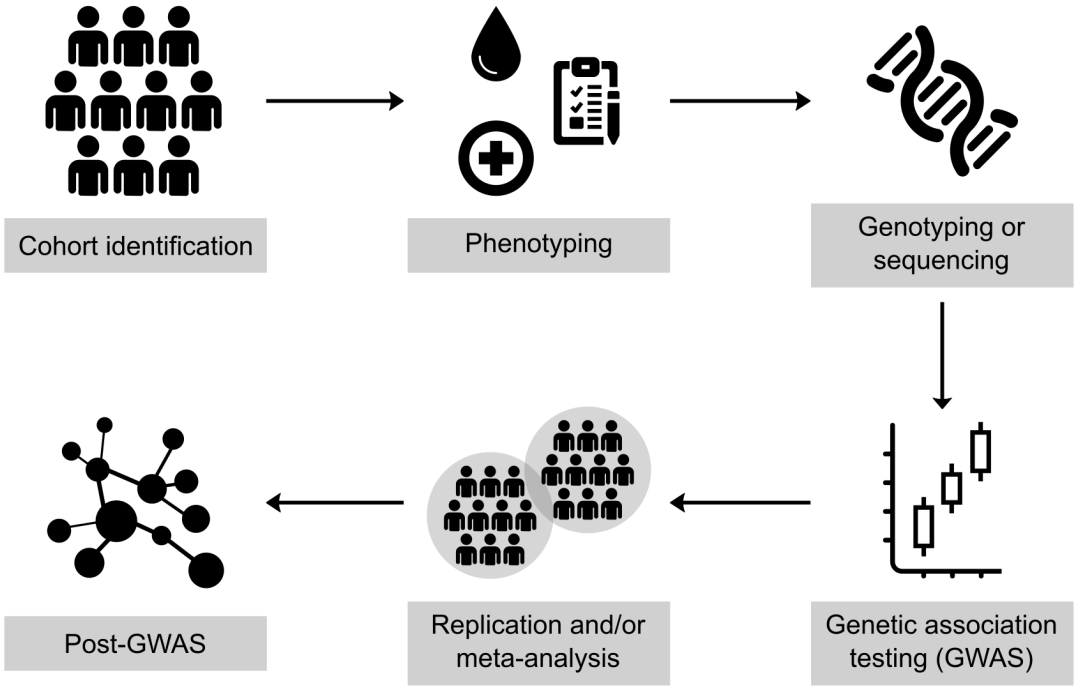


Figure 1.2: Typical workflow for genome-wide association studies (GWASs). (1) First, a suitable cohort must be identified for the disease or trait of interest, or if no such cohort exists, participants must be recruited into the study. (2) The cohort must be consented and phenotyped through biological sample collection and testing, questionnaire responses, and/or linkage to electronic medical records. Broad phenotyping allows for researchers to perform hypothesis-free scans and adjust for a range of confounders. (3) Genetic data is obtained through genotyping or whole genome sequencing, usually of blood samples, followed by alignment to the reference genome and imputation using an appropriate ancestry-matched reference panel (for genotyped samples). Quality control of variants and samples, detection of batch effects, and identification of genetic principal components to control for population stratification follow. (4) Associations between genotypes and phenotypes can then be estimated in a genome-wide association study (GWAS) with linear or logistic regression models, adjusted for confounders, and performed at scale for millions of variants simultaneously. (5) Replication of genetic findings may be sought in an external cohort with no participant overlap with the discovery dataset, and optionally, summary statistics from two or more cohorts may be meta-analysed to improve precision of genetic effect estimates. (6) Finally, *in silico* post-GWAS analyses include: nominating causal variants through fine-mapping, identifying effector genes through colocalisation and other gene prioritisation methods, and assessing the enrichment of functional pathways, cell types, and tissues for the trait. Further *in silico* downstream analyses may include unpicking the genetic architecture of the trait and shared aetiology with other traits. All icons obtained from the Noun Project database[69].

Accounting for false discovery is especially important in GWASs, where hundreds of thousands of independent variants are tested simultaneously - thus, a genome-wide significance threshold of $P < 5 \times 10^{-8}$ (family-wise error rate (FWER) controlled at 5% across 100,000 tests using the Bonferroni method) is typically used to establish variants with a high probability of trait association[76].

Despite the care taken to avoid spurious associations, it is inevitable that some genetic findings will be false. The term "winner's curse" is applied to describe the phenomenon wherein effect size estimates of SNPs that are close to the discovery threshold in the original GWAS are likely to be over-estimated[77]. It is therefore recommended to verify both the strength and significance of associations in an external cohort, with no participant overlap with the discovery cohort. Following validation, GWASs for the same trait may be meta-analysed across different cohorts. Meta-analysis empowers the discovery of variants that may not have reached statistical significance in any single study, reduces the chance of spurious findings by allowing researchers to investigate heterogeneity between studies, and enables higher precision in estimating variant effect sizes[68]. Meta-analysis software such as `METAL`[78] enable both fixed-effects (assuming variant effect sizes are constant across studies) and random-effects (allowing differences in variant effect sizes across studies) meta-analyses, and provide options to weight SNP effects by sample size or variance. While it is typically recommended that each cohort or study contributing to a meta-analysis follow the same analysis plan, from phentoype definition to genotyping, imputation, and association testing, publicly available summary statistics may also be incorporated into meta-analyses after stringent quality control[79].

Once genomic loci with high probabilities of association with a trait have been identified, several steps follow to interrogate the genetic architecture of the trait, molecular effects of individual variants, and systemic effects on human phenotypes. The heritability of the trait, which is the proportion of phenotypic variation that can be explained by genotypic variation in the population, can be estimated from SNP effects by methods such as LD score regression (`LDSC`)[62] and genomic relatedness restricted maximum likelihood (`GREML`)[80]. The former exploits LD patterns in an

ancestrally-matched reference population to separate polygenicity from population stratification and can estimate heritability from summary statistics alone[62]; the latter constructs a genetic relatedness matrix (GRM) from individual-level data, to estimate heritability under the assumption that in a population of unrelated individuals, those who are more genetically related are phenotypically more similar[80]. Heritability estimates from such genomic methods based on unrelated individuals may be compared to estimates from twin studies, which compare phenotypic differences between pairs of monozygotic (genetically identical) and dizygotic (who share, on average, 50% of their genetic material) twins to distinguish the genetic and environmental components of phenotypic variation[81]. To date, SNP-based estimates of heritability for most traits have not reached those estimated by twin studies[82,83], leading to much debate on the sources of "missing heritability", and whether twin studies produce inflated estimates of heritability[81]. Over-estimates of heritability in twin studies may arise from: (1) violations of the assumption that monozygotic and dizygotic twins have equally similar environments, (2) gene-environment interactions, or (3) non-additive genetic variance such as epistasis, i.e. interactions between genetic loci[82,83]. Modelling non-additive genetic variance may increase SNP-based estimates of heritability, but are unlikely to close the heritability gap[83].

A second set of GWAS follow-up analyses examines the molecular effects of genetic variants. While there was initially much hope that the sequencing of the human genome would pinpoint the exact genes responsible for each human trait[45], GWASs soon revealed that up to 90% of trait-associated variation is in non-coding regions of the genome and cannot be immediately assigned to effector genes[84,85]. Further, as the SNP with the lowest association P-value is not necessarily the causal variant for the trait, due to LD-tagging of several variants in a genomic locus, statistical and experimental methods may be employed to nominate causal variant(s)[86] (where a "causal" variant is pragmatically defined as described in Section 1.1). Single-SNP Bayesian approaches sequentially quantify the relative likelihood of models where each trait-associated variant in a locus is causal; scalable extensions to these models test a range of configurations,

with multiple causal variants, to produce credible sets of variants with posterior probabilities of causality (implemented in software such as `FINEMAP`[87]). Functional assays may also aid in assigning effector genes to non-coding regulatory regions of the genome. With the ability to perform transcriptome-wide sequencing and proteome characterisation at scale[88], genomic loci can be linked to variation in molecular traits such as gene expression levels (expression quantitative trait loci (eQTLs))[89] and protein adenylation, splicing, etc. (protein quantitative trait loci (pQTLs))[90,91]. Colocalisation of trait-associated genomic loci with molecular quantitative trait loci (QTLs) reveals effector genes and quantifies the likelihood of shared causal variants for molecular and systemic phenotypes at a genomic locus[92]. While QTLs may assist in gene prioritisation to some extent[93], theoretical and empirical work demonstrates that GWAS loci and molecular QTLs are systematically different in their genomic properties[94]. Instead, simply assigning variants to their nearest gene by physical location nominates the effector gene accurately for over a third of GWAS variants[95]. Regardless of prioritisation method, once genes have been identified, a range of downstream pathway, cell-type, and tissue enrichment methods are employed to nominate candidates for functional follow-up[96]. Enrichment analyses rely on the availability of robust single-cell and bulk functional data, such as epigenomic, transcriptomic, or proteomic sequencing in the tissue or cell types of interest for a trait, many of which are publicly available in databases such as the Encyclopaedia of DNA elements (ENCODE) project[84] and the Genotype-Tissue Expression (GTEx) project[97]; however, functional enrichment methods are less successful for traits whose heritability is enriched in cell types that have not been broadly sequenced[98].

Lastly, GWAS follow-up may include the interrogation of shared genetic relationships between different systemic phenotypes. Methods such as phenome-wide association studies (PheWASs), which examine the effect of a trait-associated variant on all traits for which genetic summary statistics are available, are popular in large biobanking cohort studies with linked electronic health records that can characterise a range of human phenotypes at scale[99] (and are described further in Section 1.2.2). Beyond examining individual variants with pleiotropic effects,

exploiting LD patterns allows the genome-wide characterisation of shared aetiology between pairs of traits, termed "genetic correlation"[62]. Finally, causal relationships between pairs of genetically correlated traits can be investigated using approaches such as Mendelian Randomisation (MR)[63], or the LD-based latent causal variable approach[100]. Assessing the genetic relationships among traits has been hugely beneficial in assessing the potential side-effects of pharmacological targets *in silico*[101], identifying clusters of related diseases with shared mechanistic bases[102–104], and conversely, distinct genetic subtypes of a single disease[105,106]. Advances in statistical methods to characterise phenotypes from heterogeneous health information is a promising direction for future GWASs[107].

1.2.2 Phenotyping advances for improved GWASs

The advent of biobanking projects, which gather biological samples such as blood, saliva, or urine, for tens of thousands of participants, linked to a range of demographic and health-related information, has revolutionised genetic association testing[108]. Biobanks may contain data modalities including anthropometric measurements, questionnaire responses, imaging scans, and biomarker test results[109]. With ever-falling sequencing costs, biobanks are also increasingly likely to have genetic data not only in the form of genotyped and imputed SNPs, but also whole-exome and whole-genome sequencing for each individual[110–112]. The extensive breadth and depth of data make large biobanks a cost-effective approach to large-scale population health research, particularly for genetics studies. The extraordinary success of the UK Biobank (UKBB)[113,114], a cohort of approximately 500,000 individuals living in the United Kingdom (UK), is mirrored across the globe in Japan[115], China[116], Korea[117], Taiwan[118], Qatar[119], Uganda[120], the United States of America (USA)[121–125], Canada[126], Iceland[127], Estonia[128], Finland[129], Norway[130], and the Netherlands[131], each of which boast at least one biobank of over 10,000 participants. Ancestral representation is only one of the axes of diversity among these biobanks - others include age (UKBB is ascertained for individuals who were aged 40 to 69 years at time of recruitment[114]), sex (Million Veteran Program (MVP), whose

recruitment population is military veterans, is about 90% male[132]), and health status (hospital-based biobanks such as Biobank Japan[115], the Michigan Genomics Initiative[121], and the Ontario Health Study[126] are naturally selected for less healthy individuals than the population at large).

The linkage of digitised health information, or electronic health records (EHRs), from primary care clinics, secondary care centres, prescription databases, and birth, cancer, and death registries, to biobank participant data, has exponentially increased the range of phenotypes available for study[107,133,134]. Prescription data, for example, has empowered the field of pharmacogenetics, wherein scientists examine the genetic determinants of drug response[135]. Statistical advances in electronic phenotyping have also enabled the extraction of refined disease subtypes, based on distinct clusters of symptoms or severity, some of which have been shown to have distinct genetic causes[136,137]. EHRs have additionally helped spearhead the study of co-morbidities and concomitant medication use for individuals with multiple disease diagnoses[138], the identification of rare symptoms and diseases[139], and studies of laboratory biomarkers that are not included in common blood test panels[140]. Perhaps one of the greatest advances to health research afforded by EHRs is their longitudinal aspect, which enables monitoring patterns of change in quantitative variables such as anthropometric traits or biomarker measurements over time, and identifying age at disease onset and diagnosis[107,133].

While cross-sectional snapshots of health information have proven valuable for GWASs, these are typically only adjusted for the effects of age as a covariate, and do not have the information to examine longitudinal changes in health status[64,79]. The enormous financial and logistical costs associated with regular follow-up of large groups of participants, over several decades, have prevented large-scale longitudinal genetic association studies[141]. However, where feasible, such studies have driven novel biological insights into human health, such as the discovery of distinct genetic components for growth in infancy, early childhood, and adolescence[142,143]. Longitudinal genetics studies have also capitalised on pregnancy-based cohorts, where maternal and foetal health are monitored regularly, such as the Avon

Longitudinal Study of Parents and Children (ALSPAC)[144] and the Norwegian Mother, Father, and Child Cohort Study (MoBa)[145], among others. Finally, large-scale pseudo-longitudinal genetic studies, which group cross-sectional individual data into age-bins to infer longitudinal changes across the lifespan, have revealed novel biology for change in brain structures[146]. However, as such studies do not follow individuals over time, they may be unable to separate cohort or bin-specific effects (for example, due to year of birth), from age-specific effects. Moreover, while pseudo-longitudinal studies attempt to estimate the age-specific effects of genetic variants in the population, this study design is inappropriate to study individual trajectories of change in biomarkers. Some of the first studies to utilise longitudinal EHR data linked to the UKBB have shown that within-individual variance in biomarkers, such as cholesterol and blood pressure, is associated with a different set of genetic markers than the time-averaged value of these biomarkers, further demonstrating the value of longitudinal health information for genetic studies[147].

However, as EHRs are not collected for research purposes, their biases must be examined thoroughly in both cross-sectional and longitudinal studies[148,149]. The data are ascertained for individuals who visit health providers, and may thus capture less healthy or more health-seeking subsets of the larger cohort, and inferences from such subsets may over-estimate effects in the general population[149]. Even within this group of individuals, the data will be sparse, and missingness patterns cannot be assumed random, as individuals are more likely to be tested for biomarkers relevant to their symptoms[150]. For example, patients with diabetes are more frequently monitored for glucose levels[151], women of menopausal age are more likely to be offered hormone tests[152,153], and individuals who appear overweight or obese on examination have their weight recorded more often[154,155]. Moreover, EHR data do not undergo the stringent standardisation and quality control (QC) that is typical of research data - different healthcare providers may use heterogeneous units of measurement, different protocols for biomarker testing, and use varying diagnostic thresholds - which can increase error rates in research studies based on these data[148].

Nonetheless, careful electronic phenotyping and interpretation of results can mitigate several of these biases and empower valuable EHR-based genetics research.

1.3 Metabolic and endocrine systems in humans

Studies of the intricately intertwined metabolic and endocrine systems in humans will benefit greatly from the deep and broad phenotyping available in EHR-linked biobanks. The human endocrine system consists of a network of glands and tissues that produce hormones, which are long-range chemical messengers with systemic physiological effects[156]. The hypothalamic-pituitary-target organ axes, which involve the thyroid gland, adrenal gland, and gonads (testes and ovaries), are central to endocrine regulation (Figure 1.3). Briefly, the hypothalamus secretes hormones such as thyrotropin-releasing hormone (TRH), corticotrophin-releasing hormone (CRH), and gonadotropin-releasing hormone (GnRH), which act on the anterior pituitary gland to release thyroid-stimulating hormone (TSH), adrenocorticotropic hormone (ACTH), and gonadotropins (follicle stimulating hormone (FSH) and luteinising hormone (LH)) respectively[156–159]. These pituitary hormones then target their respective organs to stimulate the production of triiodthyronine (T3) and thyroxine (T4) by the thyroid, corticosteroids by the adrenal gland, and sex steroids such as oestrogen, progesterone, and testosterone by the gonads[156–159]. These latter two categories of hormones, i.e. corticosteroids and sex steroids, are derivatives of cholesterol[160], and are closely physiologically linked to thyroid hormones[161]. Each of these hormones has wide-ranging effects on physiology, including: (1) regulation of metabolic rate, body temperature, and muscle contraction by thyroid hormones[162], (2) control of glucose uptake, blood pressure, and sleep/wake cycles by adrenal glucocorticoids such as cortisol[163], and (3) determination of foetal development, puberty, ovulation, menstruation, and pregnancy by sex steroids[164]. The hypothalamic-pituitary axis is under the control of complex negative feedback loops involving many of the hormones listed above, as well as other physiological factors[156] (Figure 1.3).

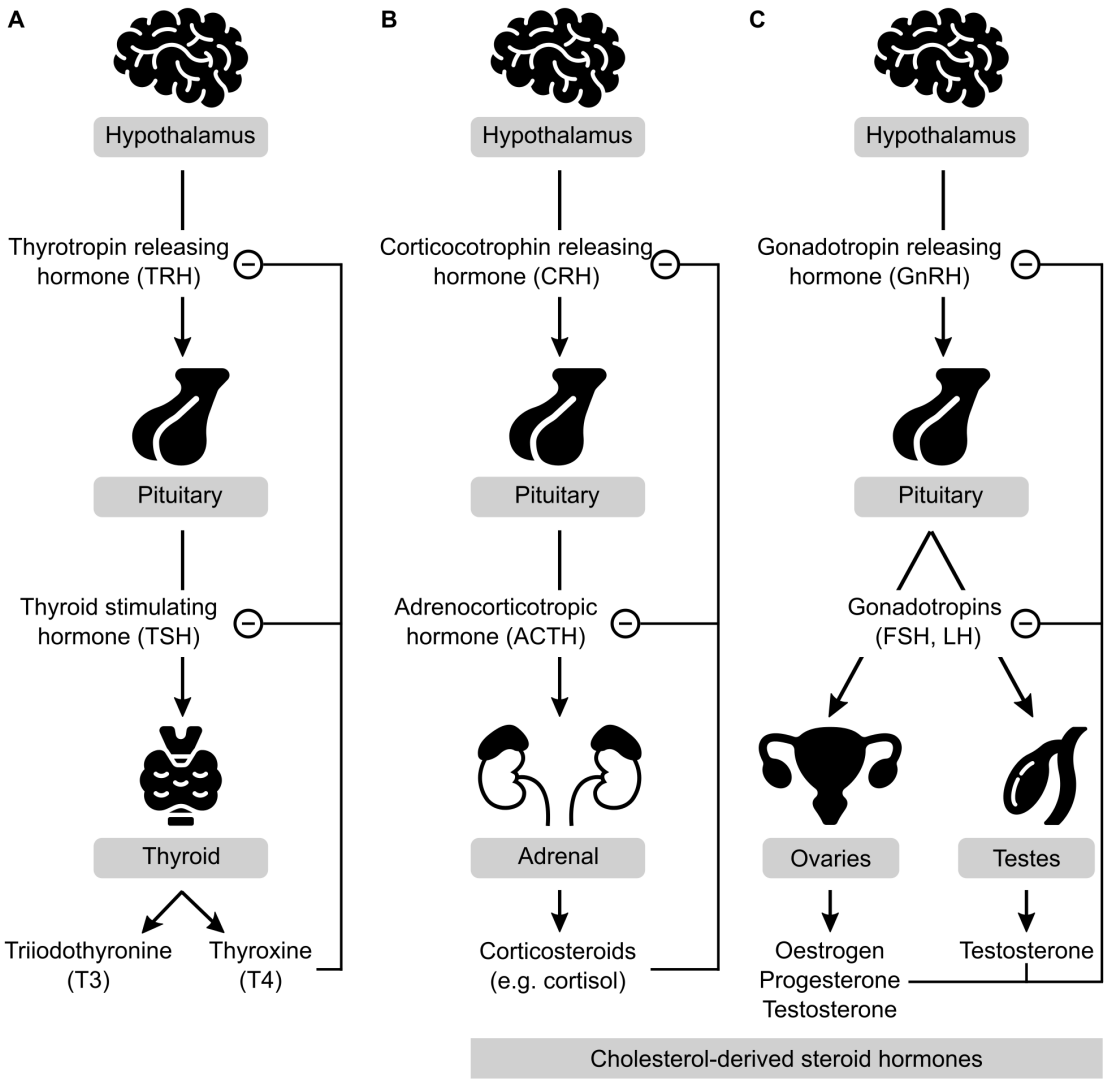


Figure 1.3: Endocrine regulation by the hypothalamic-pituitary-target organ axis. (A) Thyrotropin releasing hormone (TRH) secreted by the hypothalamus acts on the anterior pituitary gland to release thyroid stimulating hormone (TSH), which in turn stimulates the release of the thyroid hormones triiodthyronine (T3) and thyroxine (T4) from the thyroid gland. Thyroid hormones inhibit the production of TRH and TSH. (B) Corticoptrophin releasing hormone (CRH) secreted by the hypothalamus acts on the anterior pituitary gland to release adrenocorticotropic hormone (ACTH), which in turn stimulates the release of corticosteroids such as cortisol from the adrenal gland. Corticosteroids inhibit the production of CRH and ACTH. (C) Gonadotropin releasing hormone (GnRH) secreted by the hypothalamus acts on the anterior pituitary gland to release the gonadotropins - follicle stimulating hormone (FSH) and luteinising hormone (LH), which in turn stimulate the release of sex steroid hormones such as oestrogen, progesterone, and testosterone from the ovaries, and testosterone from the testes. The sex steroids inhibit the production of GnRH, FSH, and LH. Corticosteroids and sex steroid hormones belong to a class of cholesterol-derived steroid hormones. All icons obtained from the Noun Project database[69].

Adipose (fat) tissue is one such factor; while primarily recognised for its role in fat storage, it plays a critical role in the endocrine system through hormone secretion, conversion, and response[165]. Leptin and adiponectin, which are secreted by adipocytes (fat cells), modulate long-term energy balance through their actions on the hypothalamus and interactions with other hormones, including adrenal glucocorticoids[166]. Further, adipose tissue itself metabolises adrenal dehydroepiandrosterone (DHEA) and testosterone to oestrogen, using a series of steroidogenic enzymes including aromatase and hydroxysteroid dehydrogenases; adipocytes also convert thyroid T3 to T4[165,167]. In return, steroid and thyroid hormone activity determine the location and quantity of adipose deposits in the body through their metabolic activities, and adipose tissue responds to signals from glucocorticoids, androgens, and oestrogens[168]. Finally, adipose cells respond to insulin, which is a hormone produced by the pancreas to regulate glucose uptake; and interactions between cortisol, testosterone, and insulin regulate metabolism and other physiological processes[169,170].

The interplay between metabolic and endocrine systems varies by age and sex. There are distinct patterns of weight gain and loss through the lifespan[171,172], and both the quantity and distribution of adipose tissue shift with age[173–175]. While infants carry significant proportions of brown adipose tissue for thermogenesis[176], these deposits diminish in adults; instead, adults carry white adipose tissue primarily in subcutaneous deposits under the skin and muscles, and visceral deposits around internal organs, which differ in their metabolic and endocrine capacities[177]. Differential response to steroid hormones is hypothesised to be one reason why the distribution of body fat differs between the sexes: men tend to accumulate visceral adipose tissue (VAT) in "android" fat distribution patterns, whereas pre-menopausal women favour the accumulation of subcutaneous adipose tissue (SAT) in "gynoid" fat distribution patterns[177–179]. In post-menopausal women, however, adipose accumulation shifts from gynoid to android patterns, a transformation that may be driven by changing levels of oestrogen and other steroid hormones during menopause[180]. Understanding the processes that determine adipose tissue

accumulation and distribution are critical to understanding the increased risk posed by VAT deposits to development of cardiovascular disease[181,182], type 2 diabetes[174,183], and other common complex disorders[184,185].

Indeed, endocrine and metabolic dysregulation are involved in the aetiology of several rare and common diseases[156]. Congenital errors of metabolism such as familial hypercholesterolaemia and extreme early-onset obesity also predispose affected individuals to hormone deficiency and infertility[186–189]. Several studies report correlations between common forms of obesity and endocrine disorders such as type 2 diabetes[190], hypo- or hyper-thyroidism[191], reduced fertility[192–194], and polycystic ovary syndrome (PCOS) in women[195,196]. Several of these disorders are heritable, with underlying polygenic architectures[41,197], motivating a genetics-based investigation into the interactions between obesity and reproductive disease in humans.

1.3.1 Genetic basis of obesity

Obesity, which is often measured by body mass index, total body weight [kg]/standing height squared [m^2] (BMI), is among the most well-studied phenotypes in human genetics[198]. Not only is BMI heritable and correlated with several disease phenotypes, making it of great value for human genetics research[199–202], it is also cost-effective to obtain anthropometric measurements of height and weight; thus, over 300 GWASs for BMI, involving millions of participants, have been reported in the past two decades[50,198]. More recent genetic studies of obesity have also relied on advanced technologies to measure body fat mass, such as bioelectrical impedance[203,204] and dual energy X-ray absorptiometry (DEXA) scans[205,206], which are correlated with BMI but afford more precise measurements of fat distribution[207,208]. In all, there are nearly 50 unique obesity-associated phenotypes reported in the GWAS Catalog as of 8 November 2023[50].

GWASs for overall obesity and related traits have discovered over 1,000 independent obesity-associated variants across the genome, with a SNP-based heritability of 6%[40]. While this is far from saturating the expected heritability of 40-70% calculated

by twin studies[209], polygenic scores for BMI are powerful predictors of obesity, and estimate an average weight difference of 13 kg between individuals in the lowest and highest deciles of genetic risk[40]. Moreover, several genes implicated in rare, early-onset forms of obesity, such as *MC4R*, *LEPR*, and *SH2B1*, also harbour common variants associated with population variation in BMI[198]. Indeed, the penetrance of rare obesity-associated variants may depend in part on an individual's polygenic score for BMI[210]. While identifying the effector variants and genes for several robust and replicable obesity associations, such as the *FTO* and *TBX15-WARS2* loci, has proven challenging[211,212], it is evident that BMI-associated variants act through two core groups of tissues - one set of variants is enriched in the central nervous system and other neuronal cell types, to alter appetite regulation pathways, while the other set is enriched in adipose tissue and is likely to act via control of metabolism[213].

Given the differing health impacts of different adipose tissue deposits, it is natural that scientists have also attempted to isolate the genetic contributions to visceral and subcutaneous adipose mass. Bespoke scanning technologies can precisely identify specific components of adipose tissue accumulation, such as pericardial, subcutaneous, or visceral fat, but are expensive to obtain at scale[203]. Instead, the anthropometric measure of waist-to-hip ratio, waist circumference/hip circumference (WHR), which is a proxy for the ratio of visceral to subcutaneous adipose tissue mass and is highly correlated with bioelectrical impedance and DEXA scan estimates[214,215], has been leveraged for GWASs of central obesity independent of overall obesity. These studies have highlighted the differing genetic architecture of central obesity as compared to overall obesity[212,216,217]. The heritability of WHR adjusted for BMI (WHRadjBMI) is approximately 30-40%, and unlike BMI, is sex-specific[40,216]. Central obesity has higher heritability in women, and up to 30% of variants associated with WHRadjBMI in both sexes have higher effect sizes in women[216,218]. There is also limited evidence for the involvement of brain regions in the aetiology of central obesity; instead, heritability is primarily enriched in adipose tissue[218]. It is hypothesised that sex-specific mechanisms such as hormone regulation may be involved in the genetic determination of body fat distribution[204,219].

Most genetic studies of obesity in adults have relied on cross-sectional measurements of BMI or body fat distribution[198]. However, population trends in weight demonstrate that individuals tend to gain weight until the age of 50 years, but weight begins to plateau and then decline in late adulthood[220,221]. These patterns are also likely to be sex-specific, particularly around menopause in women, as shifting hormone levels may influence changes in total body fat mass and body fat distribution[222]. Cross-sectional studies of obesity have thus typically adjusted for linear and non-linear effects of age, as well as age $\times$ sex interactions[40,204,215,216], but there have been limited efforts to capture the genetic architecture of heterogeneity in weight-change patterns in individuals. Indeed, longitudinal studies in children have found that genes which have large effects on infant growth peter out later in childhood, whereas adult BMI-associated variants do not affect weight until children reach adolescence[142]. Pseudo-longitudinal studies of weight change in adults, in which cross-sectional individual data are binned into age groups[223], as well as longitudinal twin studies[221,224,225], indicate that adiposity-change in adults is heritable. However, the genetics of adiposity-change have been difficult to study at the population level due to the challenges in conducting large-scale longitudinal studies. Investigating the (potentially) sex-specific, long-term changes in overall adiposity and body fat distribution, is therefore a ripe problem for longitudinal EHR-based genetics research.

1.3.2 Reproductive trait genetics

Unlike obesity, which has proven straightforward to characterise via anthropometric measures, reproductive traits such as steroid hormone levels, diseases of the reproductive tract, and complications of fertility and pregnancy, are complex phenotypes to ascertain in EHRs. Sex hormone levels may be assayed through blood tests, but their accurate quantification in pre-menopausal women, in whom hormone levels fluctuate with the menstrual cycle[226], requires multiple repeat measurements over many weeks. It is also challenging to identify cases of female reproductive diseases in biobanks, as disorders such as endometriosis require invasive

diagnosis criteria including surgery[227,228], or syndromes such as PCOS present with heterogeneous combinations of symptoms including hyperandrogenism, obesity, anovulation, and polycystic ovaries[196]. Both male and female infertility remain under-diagnosed in population cohorts for several reasons - clinical, as a diagnosis may not be recorded if no treatment could be offered[229,230], psychological, due to anxiety and stress induced by receiving infertility diagnoses[231,232], or logistical, for reasons such as financial inability to access fertility treatment, or separation of fertility records from other health services[233–235]. Genetic studies of reproductive traits and diseases have therefore been at smaller scale than those for obesity.

Nonetheless, the UKBB release of testosterone levels for approximately 500,000 individuals has enabled GWASs for testosterone in men and women, revealing a polygenic trait architecture with SNP-based heritability of 10%[236]. Testosterone is phenotypically and genetically correlated with other hormone levels[156,237], obesity[238,239], cardiovascular disease[240,241], and immune-related traits[242,243]. While the phenotypic sexual dimorphism of testosterone has been established for several decades[244,245], recent GWASs demonstrate that the genetic correlation between testosterone in men and women is only about 3%, revealing a highly sex-specific genetic architecture[236]. Moreover, some genetic variants altering testosterone levels may affect other hormones, such as oestrogen, progesterone, DHEA, FSH, LH, etc. through their effects on the hypothalamic-pituitary axis, but these hormones have been not been characterised thoroughly in GWASs. Pott *et al.* (2019)[246] identify 15 loci associated with eight steroid hormones in up to 8,000 individuals, with some evidence for overlapping genetics, yet much of their underlying biology remains to be discovered. There has been limited genetic investigation into the tissues and cell types of interest for many steroid hormones, and beyond some rare monogenic endocrine disorders[156], little is known about rare variants with hormone-disrupting effects.

Given the systemic effects of steroid hormones on reproductive physiology, understanding the genetic drivers of hormone levels is also likely to inform the investigation of diseases of the reproductive tract[247–249]. The largest GWASs

to date of PCOS[250], endometriosis[251], and uterine fibroids[252] reveal that these conditions are heritable (with SNP-based heritability estimates between 3 and 8%), have polygenic architectures, and are genetically correlated with endocrine and metabolic conditions including type 2 diabetes and obesity. There is also evidence for shared genetic relationships between endometriosis and uterine fibroids, with genes such as *WNT4*, a key regulator of female reproductive tract development during embryogenesis, *GREB1*, an oestrogen receptor binding protein, and *FSHB*, encoding the beta-subunit of FSH, implicated in the aetiology of both conditions[252]. However, there has been no systematic investigation to date of the correlational and causal associations between various reproductive diseases and hormone levels. A genetics-based strategy may inform the treatment of reproductive diseases through hormonal and non-hormonal interventions, as well as guide our understanding of other metabolic and endocrine diseases.

Finally, heritable reproductive diseases in women are associated with reduced fertility. Endometriosis may cause difficulties in conceiving naturally[253], and up to 80% of female anovulatory infertility cases are attributed to women with PCOS[254,255]. The same may be true of male infertility, but there are few reports on common male reproductive diseases and their links to infertility[256]. Whole exome sequencing analyses in men with disorders of spermatogenesis have nominated candidate loci for male infertility[257,258], but there is no reported common variant study of male infertility of all causes. In general, there have been few studies to date that directly query the common genetic contributions to infertility in men and women. One of the difficulties in doing so arises from the natural selection pressure against fertility-reducing alleles, which keeps their frequency low in the population[259]. However, GWASs for proxy measures of fertility, such as childlessness[260], number of children ever born[261], or age at first sexual activity[262,263], indicate that there are likely to be shared genetic factors between infertility and endocrine or reproductive traits - for example, variants near *ESR1*, which is an oestrogen receptor, and the previously described *WNT4*, are associated with childlessness[260,261]. However, these proxy measures also capture behavioural and socio-economic effects on childbearing that

may confound biological findings[261,263]. Characterising the hormonal and non-hormonal determinants of infertility will thus require phenotyping of a complex and heterogeneous condition. The deep and broad phenotyping required to delineate reproductive phenotypes for GWASs are a suitable task for biobank-linked EHRs.

1.4 Thesis aims

In this thesis, I aim to conduct a comprehensive investigation into the genetic determinants of metabolic and endocrine traits using a longitudinal, multi-variable approach. I will first examine the genetic contribution to patterns of adiposity-change in adulthood by extracting longitudinal phenotypes from EHR data. Next, I will conduct a systematic search, across the allele frequency spectrum, for genetic variants associated with a range of reproductive hormones and categories of infertility, to identify shared and distinct genetic components for these traits. Finally, I will unpick the genetic relationships between adiposity and female reproductive diseases, potentially mediated by hormones, to identify causal associations between these closely linked metabolic and endocrine traits.

2

Curation of a longitudinal resource for genetic studies in the UK Biobank

Contents

2.1 Introduction to UK Biobank 26
2.1.1 Deep phenotyping 26
2.1.2 Follow-up assessments 27
2.1.3 Health-record linkage 27
2.1.4 Genetic data 28
2.2 QC of longitudinal phenotypes in UKBB-linked EHR 30
2.3 QC of genetic data 37
2.4 Genome-wide association studies 38
2.5 Post-GWAS 39
2.5.1 Lead variant identification 39
2.5.2 Lead variant classification 39
2.5.3 Genetic correlations and SNP-based heritability 40
2.5.4 Testing sex heterogeneity in genetic effects 41
2.6 Exome-wide association studies 41
2.6.1 Exome sequencing quality control 42
2.6.2 Variant annotation 42
2.6.3 Rare variant and gene-based association testing 43

Some of the material in this Chapter is adapted from my first-authored manuscript, titled "The genetic architecture of changes in adiposity during adulthood", available on medRxiv[1] and currently in second round of revisions at Nature Communications. I have obtained permission from my supervisor and all co-authors to re-use this material.

2.1 Introduction to UK Biobank

The UK Biobank (UKBB) is a prospective, deeply phenotyped, population-based cohort of approximately 500,000 individuals living in the UK[114]. The cohort was established to study the genetic and non-genetic determinants of diseases of middle and old age. Participants aged 40-69 years at recruitment had a range of demographic, medical, environmental, and genetic information recorded at initial assessment, and over multiple follow-up visits. Participants in UKBB are on average healthier and more likely to be female than the age-matched population of the UK[264], a bias whose effects I will discuss within each of the following Chapters. All analyses in this thesis were performed under UKBB application number 11867. The following description of the UKBB resource in Sections 2.1.1 to 2.1.4 is summarised from seminal papers by Sudlow *et al.* (2015)[114], Bycroft *et al.* (2018)[113], and Szustakowski *et al.* (2021)[265].

2.1.1 Deep phenotyping

Baseline assessments between 2006 and 2010 consisted of self-completed touch-screen questionnaires, interviews, physical and functional measures, as well as collection of biological samples (blood, urine, and saliva). In the questionnaire and interview, participants reported factors on: (1) family history and early-life events, such as family history of major diseases, birth weight, and childhood body size; (2) pre-existing health conditions, including medications, disability, and reproductive history; and (3) lifestyle factors such as smoking and alcohol consumption. A wide range of physical measures, including blood pressure, heart rate, spirometrics,

and anthropometrics such as height, weight, and hip and waist circumference were assessed at the baseline visit.

A collection of 34 biochemical markers (henceforth referred to as biomarkers), were assayed from blood, urine, and saliva samples collected at the baseline assessment centre visits[266]. Among these are established cardiovascular disease risk factors such as total cholesterol, LDL cholesterol, and triglycerides, the hormones testosterone and sex hormone binding globulin (SHBG), diagnostic measures for diabetes such as haemoglobin A1C / glycated haemoglobin (HbA1c) and glucose, as well as markers of renal and liver function. In subsequent extensions to the baseline phenotypes, subsets of participants were invited to multi-modal imaging assessments[267–269], and measurements of physical activity with accelerometry[270].

2.1.2 Follow-up assessments

Around 20,000 UKBB participants completed the first repeat assessment of all baseline measures between 2012 and 2013, and provided additional biological samples for biomarker assays. At the first imaging visit attended by 65,000 participants (in 2014 and beyond), several phenotypes measured at baseline, including height, weight, blood pressure, and lung function, were recorded again, creating a rich resource to study the longitudinal trajectories of phenotypes in middle and old age.

2.1.3 Health-record linkage

All participants of the UKBB provided consent for follow-up study through national EHR linkage to the death registry, cancer registry, and hospital in-patient records. Primary care data, i.e. data recorded by general practitioners (GPs), are available for approximately 45% (230,000 participants) of the UKBB cohort[271]. These contain data from three computer system suppliers across three countries: Egton Medical Information Systems (EMIS) Health[272] in Scotland and Wales, Vision Health[273] in Scotland, Wales, and England, and TPP[274] in England. This data, which shall henceforth be referred to as the UKBB-linked EHR, contains historical information

on clinical events such as diagnoses, history and symptoms of disease, laboratory test results, procedures, and prescriptions.

Clinical events in the primary care data are coded using a standard vocabulary, the Read v2 (used by the EMIS and Vision systems) and Clinical Terms Version 3 (CTV3) (used by the TPP system), whose thesaurus is available via the National Health Service Digital Technology Reference Data Update Distribution (NHS TRUD)[275] and mapped by the UKBB team to specific health outcome lists[271]. Data are available largely beginning in 1990, with extract dates, i.e. last expected date in the records, between 2016 and 2017. The data were minimally cleaned, as the UKBB team performed validation checks to flag events dated in the future or events dated prior to participants' birth date[271].

2.1.4 Genetic data

Genotyped and imputed data

The following description of the UKBB whole-genome genotyped and imputed data is adapted from Bycroft *et al.* (2018)[113].

Genotyping was performed using blood samples collected at the baseline assessment centre visit, for all participants, across two genotyping arrays. Around 50,000 participants were genotyped as part of the UK Biobank Lung Exome Variant Evaluation (UK BiLEVE) study using the Applied Biosystems UK BiLEVE Axiom Array by Affymetrix[276]. The remaining participants were genotyped with the Applied Biosystems UK Biobank Axiom Array, which shares 95% of markers with the UK BiLEVE Axiom Array. Genotype calling and QC was initially performed by Affymetrix[277]. The UKBB team performed further marker and sample QC, including inferring population structure via principal component analysis (PCA), array- and batch-based QC of markers, sample QC based on missing rate and heterozygosity, as well as sex chromosome QC to flag karyotypes distinct from from "XX" and "XY". Following the second round of QC, genotyping data was released for 488,377 participants and 805,426 markers from both genotyping arrays.

Additionally, kinship coefficients were estimated using `KING`[278] and reported for all pairs of individuals in the data who are third-degree relatives or closer.

Prior to phasing and imputation, markers were filtered to retain only those that were present in both genotyping arrays, failed QC in a maximum of one batch, displayed a missingness rate $< 5\%$, and with minor allele frequency (MAF) $> 0.01\%$ in unrelated participants. A subset of samples were removed prior to phasing based on missingness rate and heterozygosity; outliers on these axes were identified visually on a scatter-plot by the UKBB team[113]. Autosomal phasing was performed using `SHAPEIT3`[279] with the 1000 Genomes phase 3 dataset[280] used as a reference panel. The Human Reference Consortium (HRC) dataset[281] was used as the main imputation reference panel for imputation performed with `IMPUTE4`[282]; this was then combined with imputation performed using a merged UK10K[283] and 1000 Genomes phase 3[280] reference panel, with preference given to the HRC imputation for SNPs present in both. Additional QC was performed for markers on the X chromosome to exclude samples with sex chromosome aneuploidies; imputation was performed separately for the pseudo-autosomal region. The resultant set of 93,095,623 autosomal markers, including SNPs, short insertion, deletion, or insertion and deletion of nucleotides (indels) and large structural variants, and 3,963,705 markers on the X chromosome, in up to 487,442 individuals defined the initial UKBB release.

Exome sequencing

The following description of the UKBB exome sequencing data is adapted from Szustakowski *et al.* (2021)[265], Backman *et al.* (2021)[95], and the UKBB Final Exome Data Release frequently asked questions (FAQs) document[284].

The UKBB whole-exome sequencing (WES) data is available for 454,787 participants with a total targeted sequencing region of ∼39 million base pair (bp). Exomes were captured using the Integrated DNA Technologies (IDT) xGen Exome Research Panel v1.0. Multiplexed samples were sequenced by dual-indexed 75 × 75 bp paired-end reads on the Illumina NovaSeq 6000 platform using two lots of

flow cells (S2 and S4): the initial 50,000 sample release of UKBB WES data was sequenced with S2, while all other samples used the S4.

Sequencing reads were mapped using `BWA-MEM`[285] to the hg38 reference genome. Single-sample variants were called with `DeepVariant 0.0.10` and aggregated with `GLnexus 1.2.6`[286]. Any single nucleotide variants (SNVs) with read depth < 7 and indels with read depth < 10 were excluded, after which a variant-level allele balance filter was applied to ensure that the proportion of reads supporting the variant call exceeded a minimum threshold as follows - variants were retained if they had: (1) at least one homozygous carrier; or (2) at least one heterozygous carrier with allele balance ≥ 0.15 for SNVs and ≥ 0.20 for indels. Variants discordant with their genotyping chip counterpart were also excluded. Finally, similar to the sample-level QC applied to genotyping data, samples were excluded if there was a mismatch between genetically identified and reported sex, genetically identified sample duplication, and high heterozygosity or missingness (identified as visual outliers on a scatter-plot by the UKBB team[113]).

Following this variant-level and sample-level QC, the UKBB WES data contains 454,787 samples with 12.3 million variants, including SNVs, indels, and multi-allelic variants. Coverage is 20x or greater at 95.8% of targeted sites. While 15.5% of the covered SNVs are non-coding, 25.3% are synonymous, 53.8% missense, and 5.4% putative loss-of-function (pLoF) variants in coding genes. The majority of these are rare – 99% of missense and pLoF variants, and 98% of synonymous variants have MAF $< 1\%$ in this dataset.

2.2 Identification and QC of longitudinal quantitative phenotypes in primary care EHRs

Some of the QC scripts described in this section were written by Dr. Kayesha Coley while she was an internship student in the Lindgren group. I adapted and re-ran these scripts, and analysed their output.

We identified 55 quantitative phenotypes in the UKBB-linked EHRs; 29 from algorithmically-curated Clinical Practice Research Datalink (CPRD) Read v2

and CTV3 codes by Denaxas *et al.* (2020)[287] described below, 11 from the CALIBER EHR[288] resource, and the remaining via manual curation (Table A.1). Denaxas *et al.* (2020)[287] curated codes for blood biochemistry, blood count, and physical measure phenotypes with a semi-supervised phenotyping algorithm guided by previously defined CALIBER EHR[288] and NHS TRUD[275] mappings that were then verified by experts[287]. In addition to these and the CALIBER-curated phenotypes, we identified a further 15 metabolic and endocrine biomarkers from mapping tables provided by the UKBB team that were described in Section 2.1.3. We discarded any phenotypes for which fewer than 100 individuals in the UKBB-linked EHR had repeat measurements, resulting in 52 quantitative phenotypes carried forward to the longitudinal research resource (Table 2.1).

Where available, I integrated these EHR data with corresponding traits measured at UKBB baseline assessment, repeat visit, or imaging assessment (Table A.1). I performed both population-level and individual-level QC on data in the longitudinal research resource. Across the dataset, I only retained data points recorded in adulthood, i.e. between the ages of 20 - 80 years. To exclude any possible technical errors, I only retained values within 5 standard deviations (SDs) of the population mean trait value, and values from the EHR within ±10% of the maximum and minimum UKBB assessment centre measurement of the corresponding trait where available (Table A.2).

I subsequently performed longitudinal QC at the individual level. If a participant had multiple values of a trait recorded on the same day, I retained only the value closest to the individual's median value across all time-points. To exclude any extreme values for individual i with J_i data points represented as (value, age) pairs $(y_{i,j}, t_{i,j})$ for $j = 1, ..., J_i$ ordered chronologically, i.e. $t_{i,1} < ... < t_{i,J_i}$, a "jump" $P_{i,j}$ for $j = 1, ..., J_i - 1$ was defined as:

$$P_{i,j} = \log_2\left(\frac{|y_{i,j+1} - y_{i,j}|/y_{i,j}}{t_{i,j+1} - t_{i,j}}\right) \tag{2.1}$$

I removed any data points associated with extreme jumps, i.e. > 3 SDs away from the population mean jump, by excluding the value in the jump-pair farthest away from the individual's median value across all time-points.

Table 2.1: Characteristics of traits and participants in the curated UK Biobank-linked longitudinal research resource. S.D. = standard deviation.

Quantitative trait	Sex	n (%)	Mean baseline age, years (S.D.)	Mean baseline trait value (S.D.)	Mean # repeat measures (S.D.)	Mean follow-up length, years (S.D.)
Albumin, g/L	F	94606 (54.2%)	55.6 (8.43)	41.9 (3.57)	7.94 (8.62)	9.45 (4.45)
	M	79926 (45.8%)	56.2 (8.18)	42.6 (3.52)	8.41 (7.68)	9.5 (4.35)
Alkaline phosphatase (ALP), U/L	F	96376 (54.2%)	55.6 (8.47)	88.5 (47)	7.87 (8.59)	9.56 (4.5)
	M	81296 (45.8%)	56.1 (8.21)	87.7 (45.5)	8.37 (7.62)	9.6 (4.4)
Alanine aminotransferase (ALT), U/L	F	94768 (54.2%)	56.4 (8.47)	23.3 (14.3)	7.35 (8.07)	8.77 (4.37)
	M	80224 (45.8%)	56.9 (8.26)	31.3 (17.5)	7.93 (7.2)	8.88 (4.29)
Basophills, 10^9/L	F	96083 (55.3%)	54.2 (8.4)	0.0363 (0.0485)	7.59 (8.61)	8.43 (3.81)
	M	77800 (44.7%)	55.5 (8.08)	0.0354 (0.0468)	6.96 (7.29)	7.8 (3.61)
Bicarbonate, mmol/L	F	9565 (52.9%)	55.8 (8.43)	27.2 (2.58)	5.54 (5.77)	5.38 (4.19)
	M	8526 (47.1%)	56.9 (8.07)	27.5 (2.61)	5.98 (5.47)	5.39 (4.1)
Body mass index (BMI), kg/m2	F	103744 (54.5%)	48.4 (9.67)	25.6 (4.89)	7.91 (7.41)	13.5 (6.62)
	M	86595 (45.5%)	49.9 (9.59)	26.6 (4.05)	7.87 (7.18)	12.6 (6.53)
Calcium, mmol/L	F	56451 (57.3%)	58.1 (8.85)	2.36 (0.106)	4.36 (4.34)	7.35 (4.78)
	M	42134 (42.7%)	59 (8.76)	2.34 (0.0997)	4.32 (4.01)	7.32 (4.78)
Chloride, mmol/L	F	16094 (53.6%)	57.2 (8.83)	104 (2.78)	6.57 (7.88)	5.55 (3.92)
	M	13954 (46.4%)	58 (8.49)	103 (2.81)	6.94 (6.82)	5.52 (3.85)
Cholesterol, mmol/L	F	96965 (53.7%)	55 (8.39)	5.41 (1.39)	7.72 (7.24)	10.2 (5.57)
	M	83640 (46.3%)	54.6 (8.47)	5.29 (1.38)	9.19 (8.01)	11 (5.69)
Creatinine, umol/L	F	49864 (39.9%)	55.9 (8.32)	111 (400)	5.73 (6.63)	10.4 (4.4)
	M	75176 (60.1%)	55.8 (8.13)	110 (344)	8.38 (8)	10.2 (4.47)
C-reactive protein (CRP), mg/L	F	37897 (60%)	60.2 (8.44)	5.94 (7.92)	3.56 (5.11)	5.14 (3.43)
	M	25221 (40%)	61.4 (8.49)	6.9 (9.92)	3.27 (4.5)	4.84 (3.31)
Eosinophills, 10^9/L	F	96586 (55.2%)	54.1 (8.44)	0.175 (0.146)	7.7 (8.87)	8.55 (3.91)
	M	78336 (44.8%)	55.4 (8.1)	0.199 (0.151)	7.03 (7.42)	7.88 (3.69)

contd. Characteristics of traits and participants with the measured traits in the UKBB-linked EHR.

Quantitative trait	Sex	n (%)	Mean baseline age, years (S.D.)	Mean baseline trait value (S.D.)	Mean # repeat measures (S.D.)	Mean follow-up length, years (S.D.)
Free androgen index (FAI), %	F	24 (25.3%)	49.2 (6.49)	1.81 (2.38)	2.75 (1.48)	2.36 (2.33)
	M	71 (74.7%)	60.9 (7.48)	36.2 (27.6)	2.48 (0.86)	1.92 (2.4)
Ferritin, ug/L	F	26118 (66.7%)	56.2 (9.74)	62.2 (102)	3.88 (3.12)	5.22 (4.35)
	M	13020 (33.3%)	60.9 (8.54)	142 (211)	3.59 (2.83)	4.04 (3.62)
Forced expiratory volume (FEV1), L	F	4827 (49.9%)	60 (7.5)	1.89 (0.587)	3.99 (2.84)	4.69 (3.51)
	M	4843 (50.1%)	60.9 (7.45)	2.5 (0.815)	4.38 (3.05)	4.82 (3.54)
Fibrinogen, g/L	F	701 (63.2%)	58.1 (9.06)	3.82 (8.22)	2.69 (2.3)	3.05 (3.58)
	M	409 (36.8%)	60.8 (8.3)	3.35 (1.76)	3.19 (4.18)	2.15 (2.76)
Folate, ug/L	F	17366 (61.7%)	57.9 (9.28)	59.4 (137)	3.32 (2.44)	4.79 (3.95)
	M	10760 (38.3%)	61.1 (8.35)	47.3 (121)	3.33 (2.27)	4.09 (3.52)
Follicle stimulating hormone (FSH), IU/L	F	9326 (96.1%)	45.8 (5.14)	16.7 (24.8)	2.78 (1.4)	4.56 (3.84)
	M	377 (3.89%)	58.8 (8.32)	8.17 (10.3)	2.45 (0.997)	2.44 (3.26)
Forced vital capacity (FVC), L	F	3809 (50.3%)	60.1 (7.49)	2.65 (0.67)	3.48 (2.32)	4.33 (3.4)
	M	3758 (49.7%)	61.1 (7.49)	3.67 (0.929)	3.73 (2.53)	4.36 (3.38)
Glucose, mmol/L	F	66386 (54.2%)	56.9 (8.61)	5.25 (1.31)	4.31 (3.06)	8.36 (4.3)
	M	56202 (45.8%)	57.6 (8.29)	5.65 (1.65)	4.65 (3.4)	8.42 (4.25)
Haematocrit percentage, %	F	96434 (55.3%)	53.9 (8.51)	12.7 (18.1)	7.64 (8.81)	8.65 (4.03)
	M	78059 (44.7%)	55.4 (8.11)	16.5 (20.8)	6.96 (7.36)	7.95 (3.74)
Haemoglobin concentration, g/dL	F	91000 (55.3%)	53 (8.79)	13.3 (1.14)	7.66 (8.55)	8.71 (4.63)
	M	73659 (44.7%)	54.7 (8.24)	14.9 (1.1)	6.91 (7.29)	7.78 (4.18)
HbA1c, mmol/mol	F	11986 (51.9%)	63.3 (7.75)	39.6 (9.95)	4.36 (3.86)	2.53 (2.14)
	M	11127 (48.1%)	63.3 (7.81)	41 (11.9)	5.19 (4.6)	3.01 (2.37)
HbA1c standardised	F	27860 (48.9%)	63 (7.65)	41.7 (13.2)	4.66 (4.18)	2.73 (1.95)
	M	29141 (51.1%)	63.2 (7.55)	43.8 (15.4)	5.59 (4.88)	3.15 (2.08)

contd. Characteristics of traits and participants with the measured traits in the UKBB-linked EHR.

Quantitative trait	Sex	n (%)	Mean baseline age, years (S.D.)	Mean baseline trait value (S.D.)	Mean # repeat measures (S.D.)	Mean follow-up length, years (S.D.)
HDL cholesterol, mmol/L	F	91803 (53.4%)	56.9 (7.89)	1.67 (0.436)	6.11 (4.69)	8.43 (4.33)
	M	80208 (46.6%)	56.6 (7.96)	1.33 (0.355)	7.08 (5.15)	9.03 (4.35)
Height, cm	F	10084 (56.2%)	49 (9.5)	162 (6.48)	4.95 (4.12)	11.7 (5.91)
	M	7871 (43.8%)	50.4 (9.34)	176 (7)	5.17 (4.42)	11 (5.87)
Iron, umol/L	F	2341 (62.9%)	57 (9.52)	13.5 (8.45)	3.24 (2.46)	3.08 (3.03)
	M	1382 (37.1%)	61.4 (8.61)	17 (54.2)	2.92 (1.69)	2.63 (2.69)
LDL cholesterol, mmol/L	F	70460 (52.8%)	57.1 (7.87)	3.43 (0.975)	5.37 (4.18)	8.55 (4.23)
	M	62965 (47.2%)	57 (7.96)	3.31 (0.975)	6.01 (4.57)	9.03 (4.24)
Luteinising hormone (LH), IU/L	F	6407 (94.5%)	45.8 (5.12)	13.1 (15.9)	2.67 (1.31)	4.24 (3.58)
	M	370 (5.46%)	58.9 (8.09)	5.06 (5.38)	2.41 (0.907)	2.33 (3.02)
Lymphocytes, 10^9/L	F	96614 (55.2%)	54.1 (8.45)	1.99 (0.926)	7.72 (8.91)	8.56 (3.93)
	M	78373 (44.8%)	55.4 (8.11)	1.96 (1.07)	7.04 (7.44)	7.9 (3.71)
Mean corpuscular haemoglobin concentration (MCHb), pg	F	59113 (56.5%)	54.2 (8.94)	31.7 (6.91)	6.81 (8.06)	7.63 (4.55)
	M	45451 (43.5%)	56.3 (8.33)	32.1 (7.03)	6.35 (6.63)	7.01 (4.21)
Mean corpuscular volume (MCV), fL	F	97372 (55.2%)	53.5 (8.65)	90.3 (5.2)	7.93 (9.12)	9.08 (4.36)
	M	78960 (44.8%)	55 (8.19)	90.8 (4.58)	7.18 (7.59)	8.25 (4)
Monocytes, 10^9/L	F	96551 (55.2%)	54.1 (8.43)	0.438 (0.19)	7.69 (8.83)	8.52 (3.9)
	M	78335 (44.8%)	55.4 (8.1)	0.507 (0.213)	7.03 (7.43)	7.87 (3.69)
Neutrophills, 10^9/L	F	96631 (55.2%)	54.1 (8.44)	3.82 (1.57)	7.73 (8.9)	8.54 (3.9)
	M	78407 (44.8%)	55.4 (8.11)	3.92 (1.58)	7.07 (7.46)	7.89 (3.68)
Oestradiol, pmol/L	F	2478 (99.5%)	46.1 (4.85)	551 (458)	2.45 (1.4)	7.83 (4.66)
	M	13 (0.522%)	57.5 (9.05)	524 (796)	2.85 (1.95)	5.64 (3.54)
Phosphate, mmol/L	F	40502 (58.6%)	59.2 (8.7)	1.13 (0.171)	3.5 (3.26)	6.43 (4.47)
	M	28563 (41.4%)	60.3 (8.67)	1 (0.18)	3.42 (2.75)	6.31 (4.46)

contd. Characteristics of traits and participants with the measured traits in the UKBB-linked EHR.

Quantitative trait	Sex	n (%)	Mean baseline age, years (S.D.)	Mean baseline trait value (S.D.)	Mean # repeat measures (S.D.)	Mean follow-up length, years (S.D.)
Platelets, 10^9/L	F	97405 (55.2%)	53.5 (8.64)	272 (64.6)	7.96 (9.18)	9.08 (4.34)
	M	79021 (44.8%)	55 (8.18)	244 (59.8)	7.21 (7.64)	8.26 (4)
Potassium, mmol/L	F	85370 (53.8%)	55.1 (8.4)	4.38 (0.432)	8.5 (8.73)	8.74 (4.55)
	M	73185 (46.2%)	55.9 (8.1)	4.4 (0.45)	9.39 (8.4)	8.81 (4.47)
Progesterone, nmol/L	F	534 (99.3%)	42.9 (6.53)	13.3 (20.3)	2.6 (1.34)	2.26 (2.96)
	M	4 (0.743%)	63.1 (5.52)	15.2 (29)	2 (0)	1.81 (1.88)
Prolactin, mU/L	F	489 (58.2%)	47.4 (7.9)	544 (3290)	3.14 (2.73)	3.18 (3.41)
	M	351 (41.8%)	58.1 (8.2)	324 (1010)	2.75 (1.77)	2.37 (2.58)
Red blood cell count (RBC), 10^12/L	F	96929 (55.2%)	53.6 (8.59)	3.11 (2.04)	7.84 (8.97)	8.93 (4.26)
	M	78607 (44.8%)	55.1 (8.16)	3.14 (2.35)	7.12 (7.51)	8.16 (3.94)
Sodium, mmol/L	F	85610 (53.8%)	55.1 (8.41)	140 (2.43)	8.54 (8.76)	8.74 (4.54)
	M	73374 (46.2%)	55.9 (8.1)	140 (2.45)	9.44 (8.42)	8.81 (4.46)
Testosterone, nmol/L	F	1711 (16.9%)	49.8 (8.77)	1.4 (1.24)	2.26 (0.875)	8.84 (4.7)
	M	8388 (83.1%)	59.3 (8.05)	15 (6.26)	2.6 (2.05)	6.71 (4.33)
Total bilirubin, umol/L	F	95949 (54.2%)	55.6 (8.46)	9.84 (4.63)	7.7 (8.48)	9.54 (4.48)
	M	81116 (45.8%)	56.2 (8.2)	12.5 (5.77)	8.26 (7.53)	9.59 (4.38)
Triglycerides, mmol/L	F	89728 (53.3%)	56.4 (8.05)	1.35 (0.759)	6.22 (4.96)	9.13 (4.69)
	M	78716 (46.7%)	56.1 (8.15)	1.71 (0.985)	7.2 (5.45)	9.76 (4.79)
Urea, mmol/L	F	92715 (54.1%)	55.8 (8.32)	5.14 (1.34)	7.92 (7.93)	9.36 (4.18)
	M	78793 (45.9%)	56.3 (8.05)	5.59 (1.41)	8.75 (7.76)	9.48 (4.09)
Vitamin B12, ng/L	F	18328 (61.6%)	57.7 (9.31)	362 (238)	3.42 (2.57)	4.93 (4.06)
	M	11420 (38.4%)	61 (8.38)	320 (198)	3.42 (2.4)	4.23 (3.61)
Vitamin D, nmol/L	F	12822 (70.2%)	61.2 (8.46)	51.3 (27.8)	2.79 (1.4)	3.35 (2.3)
	M	5438 (29.8%)	62.5 (8.51)	48.1 (26.2)	2.65 (1.18)	3.24 (2.23)

contd. Characteristics of traits and participants with the measured traits in the UKBB-linked EHR.

Quantitative trait	Sex	n (%)	Mean baseline age, years (S.D.)	Mean baseline trait value (S.D.)	Mean # repeat measures (S.D.)	Mean follow-up length, years (S.D.)
White blood cell count (WBC), 10^9/L	F	97541 (55.2%)	53.4 (8.68)	6.54 (2.05)	8.05 (9.32)	9.16 (4.42)
	M	79089 (44.8%)	54.9 (8.21)	6.68 (2.11)	7.28 (7.72)	8.32 (4.06)
Waist circumference (WC), cm	F	27782 (52.5%)	57.1 (7.75)	85.6 (13.4)	2.68 (1.62)	4.12 (2.14)
	M	25153 (47.5%)	57.6 (7.91)	97.6 (12.5)	2.88 (1.73)	4.01 (2.24)
Weight, kg	F	113495 (54.6%)	48.2 (9.61)	67.7 (13.4)	8.47 (8.34)	13.8 (6.61)
	M	94252 (45.4%)	49.7 (9.55)	82.8 (13.6)	8.28 (7.85)	12.8 (6.55)
Waist-to-hip ratio (WHR)	F	1898 (52.3%)	58 (7.62)	0.826 (0.0743)	2.84 (1.83)	3.57 (1.73)
	M	1734 (47.7%)	58 (7.9)	0.937 (0.0637)	2.9 (1.93)	3.4 (1.92)

Finally, I performed additional phenotype-specific QC for the adiposity traits BMI, weight, and WHR. Participants with codes for "history of bariatric surgery", as identified previously by Kuan *et al.* (2019)[289] (category (1) in Table A.3), were excluded entirely, while only observations up to the date of bariatric surgery were retained for patients for whom this could be determined (categories (2) and (3) in Table A.3).

Characteristics of participants and phenotypes in the curated longitudinal research resource, following QC, are recorded in Table 2.1. Weight is the most commonly measured trait in this resource (207,747 individuals, or around 90% of participants, have repeated weight measures), whereas the hormone biomarkers free androgen index (FAI), progesterone, and prolactin each have repeat measurements in fewer than 1,000 participants in the dataset. Reflecting the participation bias in the general UKBB dataset[264], women are marginally more likely to be represented in the EHR data (approximately 55% female to 45% male for most phenotypes). However, some traits show larger sex imbalances, including ferritin (66.7% female), vitamin D (70.2% female), and the sex hormones FSH, LH, oestradiol, and progesterone (between 96.1 to 99.3% female), and FAI (74.7% male) and testosterone (83.1% male).

Mean age at first measurement of a trait varies from 42.9 years (SD = 6.53) for progesterone to 63.3 years (7.79) for HbA1c. Some traits are measured more frequently than others – sodium and potassium (mean 9.39 to 9.44 repeat measurements per participant) are the most frequently measured biomarkers, while FAI, oestradiol, progesterone, and testosterone (mean 2.45 to 2.60 repeat measurements per participant) are rarely measured across multiple visits. Finally, the length of follow-up, i.e. the number of years between first and last measurement of the trait, also varies greatly between traits. Weight and BMI are, on average, measured over 13.5 to 13.8 years per individual, whereas the sex hormones FAI (mean 1.92 years of follow-up) and progesterone (mean 2.15 years of follow-up) are only monitored briefly. This curated dataset therefore reflects a diverse range of biologically relevant traits with considerable heterogeneity in the availability of longitudinal information.

2.3 Quality control of genotyped and imputed data

In addition to the QC of genetic data performed by the UKBB team and described in Section 2.1.4, we performed post-imputation QC to retain only bi-allelic SNPs with MAF > 1%, imputation information score (a metric of uncertainty in imputed genotypes, with 0 being complete uncertainty and 1 being complete certainty) > 0.8, missingness rate < 5%, and Hardy-Weinberg equilibrium (HWE) exact test $P > 1 \times 10^{-6}$. The HWE states that the genotype frequencies AA, AB, and BB for a bi-allelic marker occur with relative proportions p^2, $2p(1-p)$, and $(1-p)^2$, where p is the allele frequency of A and $(1-p)$ that of B; significant deviations from HWE in large-scale populations indicate genotyping error[290]. We excluded individuals with sex chromosome aneuploidies, those whose genetically inferred sex did not match self-reported sex, and participants with an excess of third-degree relatives in UKBB, as reported by the UKBB team[113]. Finally, we excluded samples visually identified as heterozygosity and missingness outliers on a scatter-plot, those excluded from autosome phasing or kinship inference, and any other recommended exclusions[113].

2.4 Genome-wide association studies

I performed GWASs to test SNPs for associations with traits of interest, i.e. adiposity-related traits in Chapter 3 and reproductive hormones and infertility in Chapter 4, using the linear mixed model association testing software `REGENIE`[65], for whole genome regression modelling that accounts for relatedness and population stratification. Main analyses in both Chapters were restricted to individuals of genetically ascertained non-Finnish European ancestry, and additional analyses described in each Chapter were extended to other ancestral groups. The classification of ancestral groups was performed by Dr. Duncan Palmer, a post-doctoral researcher in the Lindgren group. In brief, UKBB samples were projected onto a principal component (PC) space, defined from unrelated samples in the 1000 Genomes Project database[280] after restricting to an LD-pruned collection of overlapping variants between UKBB and 1000 Genomes[291]. A random forest predictor was trained to identify broad ancestry population labels (European (EUR), African (AFR), East Asian (EAS), South Asian (SAS), and American (AMR)) using the first four PCs. Labels were only assigned where the random forest predictor exceeded 99% probability estimation in the ancestry labelling; samples of Finnish European ancestry were removed by this process. This reflects the unique population history of Finland, which is genetically isolated from most other European populations due to a recent bottleneck event[292].

All GWASs were performed separately in three strata: female-specific, male-specific, and sex-combined adjusted for sex. Each test was adjusted for the following covariates: 21 genetic PCs to account for population structure, genotyping array, and genotyping batch. Unpublished work in the Lindgren group has shown that much of the population stratification within individuals of European ancestry in the UK Biobank can be accounted for by adjustment for between 4 to 40 genetic PCs. Further trait-specific adjustments are described within the relevant Chapters.

2.5 Post-GWAS

2.5.1 Lead variant identification

Fine-mapping to identify lead variants

We defined genome-wide significant (GWS) loci by merging windows of 1 megabase (one megabase is one million base pairs) (Mb) around SNPs with $P < 5 \times 10^{-8}$, which accounts for the multiple testing burden of approximately one million independent variants across the genome[76]. Putative causal variants with posterior inclusion probability $> 95\%$ within each GWS locus were identified using the `FINEMAP` software[87], using LD information from European-ancestry individuals in the 1000 Genomes dataset[280]. As we performed fine-mapping with an LD-based method using a reference panel, we restricted the summary statistics to variants with MAF $\geq 1\%$ for which a high quality correlation matrix could be constructed[87]; however, low-frequency variants are *a priori* more likely to be causal[293] and may be missed by the implemented MAF cut-off.

Distance-based pruning to identify lead variants

As an alternative to fine-mapping, which may be miscalibrated for meta-analyses[294] and mismatched ancestral reference panels[87], I performed distance-based pruning to identify the SNP with the lowest P-value within each 1 Mb window at all loci with at least one GWS variant with $P < 5 \times 10^{-8}$.

2.5.2 Classification of lead variants as novel or reported

I adapted criteria developed by Benonisdottir *et al.* (2016)[295] to classify each autosomal lead variant for a given trait as novel, refined, or reported based on previously published GWASs for the trait of interest. I curated a list of all published SNPs associated with the trait in the GWAS Catalog[50] accessed on 27 March 2023. The list of traits, number of SNPs associated with each trait, and PubMed Identifiers for the association studies can be found in Tables B.1 and C.2. I then performed conditional analysis using `GCTA-COJO`[296], with LD information for European-ancestry individuals from the 1000 Genomes dataset[280], to identify the effect of the lead

variant conditional upon all other published trait-associated variants within ± 500 kilobases (one kilobase is one thousand base pairs) of the lead variant.

Reported SNPs in our study are those whose effects are fully accounted for, i.e. $P_{conditional} > 0.05$, by published trait-associated variants in the defined 1 Mb window. Refined SNPs fulfill all of the following criteria: (1) the SNP of interest is correlated with at least one published trait-associated variant in the 1 Mb window ($r^2 \geq 0.1$), (2) the SNP of interest accounts for the effect of all published trait-associated SNPs in the window in conditional analysis ($P_{conditional} > 0.05$ for all published variants), and (3) published trait-associated variants cannot fully account for the effect of the SNP of interest in conditional analysis ($P_{conditional} < 0.05$ for the refined variant of interest). Finally, I classified a lead variant in our study as novel if it was not in LD with ($r^2 < 0.1$), and conditionally independent of ($P_{conditional} < 0.05$), all published trait-associated variants in the defined 1 Mb window.

For lead variants on the X chromosome, for which genetic analysis is typically more difficult due to due to differences in recombination rates and selection pressures between the sexes[297–299], I did not use the above three-tiered, LD-based classification system. Instead, I classified an X-chromosome lead SNP as reported if it was within 1 Mb of a published obesity-associated variant and whose effect was not independent of the published variant ($P_{conditional} > 0.05$), and novel if not. I used the same distance-based classification system for lead variants from multi-ancestry GWAS meta-analyses, for which LD patterns may differ between populations, and where the identification of lead variants by fine-mapping may be mis-calibrated[294].

2.5.3 Genetic correlations and SNP-based heritability

Throughout this thesis, I estimate the SNP-based heritability (h^2_{G}) of a trait and the genetic correlation (r_{G}) between pairs of traits from GWAS summary statistics using LD-score regression as implemented in the `LDSC` software[62]. I use pre-computed LD-scores based on European-ancestry individuals in the 1000 Genomes dataset[280], restricted to HapMap3 SNPs[46].

For case-control traits, I convert the observed-scale heritability (h^2_{obs}) to liability-scale heritability (h^2_{liab}), which accounts for the disease prevalence in the sample (k) and population (K)[300]. Under the assumption that sample prevalence equals the population prevalence, I estimate h^2_{liab} as:

$$h^2_{\text{liab}} = h^2_{\text{obs}} \frac{k(1-k)}{\varphi(\Phi^{-1}(k))^2} \tag{2.2}$$

where the denominator is the squared probability density function of the standard Gaussian distribution evaluated at the k^{th} quantile of the inverse cumulative distribution function (CDF) of the standard Gaussian distribution.

2.5.4 Testing sex heterogeneity in genetic effects

I test for heterogeneity between female-specific and male-specific effects of a genetic variant on a trait by calculating Z-statistics and corresponding P-values from the sex-specific GWAS summary statistics for the variant:

$$Z_{\text{sex-het}} = \frac{\hat{\beta}_{\text{F}} - \hat{\beta}_{\text{M}}}{\sqrt{(\sigma^2_{\text{F}} + \sigma^2_{\text{M}})}} \tag{2.3}$$

where $\hat{\beta}_{\text{F}}$ and σ_{F} are the female-specific effect size and standard error of the variant on the trait respectively, and $\hat{\beta}_{\text{M}}$ and σ_{M} are the male-specific effect size and standard error respectively. A similar Z-statistic is calculated to test for heterogeneity between female-specific and male-specific estimates of trait h^2_{G} and r_{G} between traits.

2.6 Exome-wide association studies

The QC and analyses performed in this section were conducted by various members of the Lindgren group, including Nikolas Baya, Barney Hill, Frederik Heymann Lassen, and Dr. Duncan Palmer. I analysed their outputs and produced all the figures, tables, and text reported in this and following Chapters.

2.6.1 Exome sequencing quality control

Following protocols defined by Karczewski *et al.* (2020) for exome sequencing QC[301], we evaluated sample-level QC metrics, i.e. mean call rate and depth of coverage, on a set of high-quality variants (defined as those with MAF $> 0.1\%$ and call rate ≥ 0.99). The set of high-quality variants was only used for performing sample-level QC, exome testing was not restricted to this variant set. Sample-level filters were applied to remove individuals with metrics ≥ 4 times outside the median absolute deviation for each sample-level metric, and variant-level filters to exclude variants that did not pass previously defined thresholds for SNVs and indels, and variants with excess heterozygosity, as defined by Karczewski *et al.* (2020)[301]. Finally, genotype-level filters were applied to remove low quality genotypes with depth ≤ 10, genotype quality ≤ 20 and minor allele balance ≤ 0.2 for all alternate alleles for heterozygous genotypes.

2.6.2 Variant annotation

Variants were annotated using `VEP v105`[302], `LOFTEE v1.04-GRCh38`[301], `CADD v1.6`[303], and `REVEL` using `dbNSFP4.3`[304]. We then ran `SpliceAI v1.3`[305] to annotate transcripts, defining the "canonical" transcripts for variant-specific annotations as follows: Matched Annotation from NCBI and EMBL-EBI (MANE) Select[306] where available, and canonical where no MANE Select transcript was available (for protein-coding genes only). The collection of missense, pLoF, splice metrics, and annotations of variant consequences on the canonical transcript were used to determine a set of variant categories for gene-based testing as follows:

1. High-confidence pLoF variants: high-confidence loss of function variants defined by `LOFTEE`

2. Damaging missense or protein-altering variants: at least one of the following -

    - variants annotated as missense, start-loss, stop-loss, or in-frame indels, with `REVEL` score ≥ 0.773 and/or `CADD` score ≥ 28.1[307]

- variants with `SpliceAI` delta score ≥ 0.2 for any of the acceptor gain, acceptor loss, donor gain, or donor loss scores
- low-confidence loss of function variants defined by `LOFTEE`

3. Other missense or protein-altering variants: missense, start-loss, stop-loss, or in-frame indels not categorised in (2) above

4. Synonymous variants: synonymous variants with `SpliceAI` delta score < 0.2

2.6.3 Rare variant and gene-based association testing

We performed rare variant genetic association testing in individuals of genetically ascertained non-Finnish European ancestry, classified as described in Section 2.4, using `SAIGE v1.1.19`[67]. All association tests were performed separately in three strata: female-specific, male-specific, and sex-combined adjusted for sex. Each test was adjusted for the following covariates: age, age^2, age $\times$ sex, age$^2\times$ sex, and the first 10 genetic PCs to account for population structure. A sparse GRM, as required by `SAIGE` in step 0, was constructed using 5,000 randomly selected markers in the UKBB genotyping array data. We generated a variance ratio file for subsequent steps in `SAIGE` by randomly selecting 1000 variants each with minor allele count (MAC) < 20 and MAC > 20. The null model in step 1 was fitted with the generated sparse GRM, variance ratio, and default values for all parameters except relatedness cut-off (which was set to 0.05). In step 2, variant and gene-based testing were carried out together in group-based test calls, with default values for all parameters except maximum MAF categories, which were set to (0.0001, 0.001, 0.01). Using variant annotations described in Section 2.6.2 above, we included the following collections of group tests:

- High-confidence pLoF
- Damaging missense or protein-altering
- High-confidence pLoF or damaging missense or protein-altering

- Other missense or protein-altering
- Synonymous
- All of the above

The Cauchy combination test was performed to collapse significance values across all annotations for each gene[308].

3

The genetic architecture of changes in adiposity during adulthood

Contents

3.1 Introduction 46
3.2 Methods 47
3.2.1 Models to define change in adiposity over time 48
3.2.2 Genome-wide association studies 61
3.2.3 Sensitivity analyses for survival bias 62
3.2.4 Replication of genetic associations in the UK Biobank . 64
3.2.5 Replication of genetic associations in the Million Veteran Program 66
3.2.6 Replication of genetic associations in the Estonian Biobank 67
3.2.7 Power calculations for replication sample sizes 69
3.2.8 Statistical power comparison to published GWASs . . . 70
3.2.9 Longitudinal phenome-wide association studies 71
3.2.10 Joint modelling of intra-individual mean and variance . 71
3.3 Results 72
3.3.1 Characteristics of baseline adiposity and adiposity-change 72
3.3.2 Genetic architecture of adiposity-change in adulthood . 74
3.3.3 *APOE* variant rs429358 is associated with multiple adiposity-change phenotypes 85
3.3.4 Genetic associations with adiposity-change are robust to a range of modelling choices 91
3.3.5 Longitudinal data enable the identification of novel genetic signals for baseline adiposity 96
3.3.6 No evidence for genetic basis of within-individual variability in weight 103
3.4 Discussion 107

Some of the material in this Chapter is adapted from my first-authored manuscript, titled "The genetic architecture of changes in adiposity during adulthood", available on medRxiv[1] and currently in second round of revisions at Nature Communications. I have obtained permission from my supervisor and all co-authors to re-use this material.

3.1 Introduction

Cross-sectional genome-wide association studies have identified thousands of genetic variants associated with obesity[198], but have provided limited insight into the mechanisms by which individuals gain and lose weight over a lifetime. Twin studies indicate that these dynamic processes of weight change may be heritable[224,225,309], and recent work has illuminated the age-specific genetic landscape of infant, childhood, and adolescent BMI[142,143,310]. An EHR-based study of about 20,000 adult participants also provides evidence that the genetic basis of weight-change is likely to be distinct from that of baseline obesity[311], however less is known about specific genetic variants that contribute to patterns of weight change. The paucity of GWASs for traits that change over the adult lifespan may be partly attributed to challenges in building and maintaining a genetics-linked cohort to regularly follow participants over their lifetime[141].

Longitudinal data are a key feature of EHRs, which are increasingly being adopted in the clinic and integrated into biobanks[107,134]. Although not designed for research, and thus affected by ascertainment bias, non-random missingness, sparsity, and high rates of technical and data entry errors[148,149], EHR-based genetics research has proven to robustly replicate findings from purpose-built cohorts, across a range of diseases[312–314]. The longitudinal data in EHR also provide the opportunity to average repeated quantitative trait measurements from an individual over time, which, as expected[315,316], is better powered for GWAS than a randomly sampled cross-sectional value of the same trait[317,318]. Finally, repeat measurements of

quantitative traits in the EHR enable researchers to derive longitudinal metrics such as linear rates of change[319], trajectory-based clusters[320], and within-individual variability over time[147], all of which may provide complementary information to cross-sectional values and help uncover novel biology underpinning human trait variation.

Efficiently utilising the longitudinal data in biobank-linked EHRs requires models that can scale to millions of observations across hundreds of thousands of individuals. Simple linear mixed-effects (LME) models target the linear rate in change of a trait over time[147,321,322], whereas more complex semi-parametric regression models generate flexible longitudinal patterns from combinations of basis functions, such as B-splines, regularised to induce temporal smoothness[323]. Although it is theoretically possible to construct full joint models that incorporate repeat trait observations and genetic data simultaneously[147], no existing methods can fit such customisable models at scale. Two-stage approaches that carry forward summary metrics from a longitudinal sub-model to a genetic sub-model – for example, by identifying sub-groups of individuals with similar trait trajectories via clustering[324–327], with sub-group membership tested for association with genetic variation – are thus preferred for their tractability[319].

In this Chapter, I leverage linked EHRs in the UK Biobank, Million Veteran Program, USA, and the Estonian Biobank, to study the genetic architecture of adiposity-change in adults. I apply a two-stage approach to first derive linear and non-linear metrics of change in adiposity, and then perform genome-wide testing to identify SNPs associated with adiposity-change, independent of baseline obesity. I report specific genetic variants associated with changes in BMI and weight over time, compare the distinct genetic components of adiposity-change and baseline adiposity, and demonstrate the power of leveraging multiple EHR measurements for the discovery of novel variants associated with obesity.

3.2 Methods

The data for discovery analyses in this Chapter were derived from the UKBB-linked EHR resource described in Chapter 2. Identification and QC of adiposity traits –

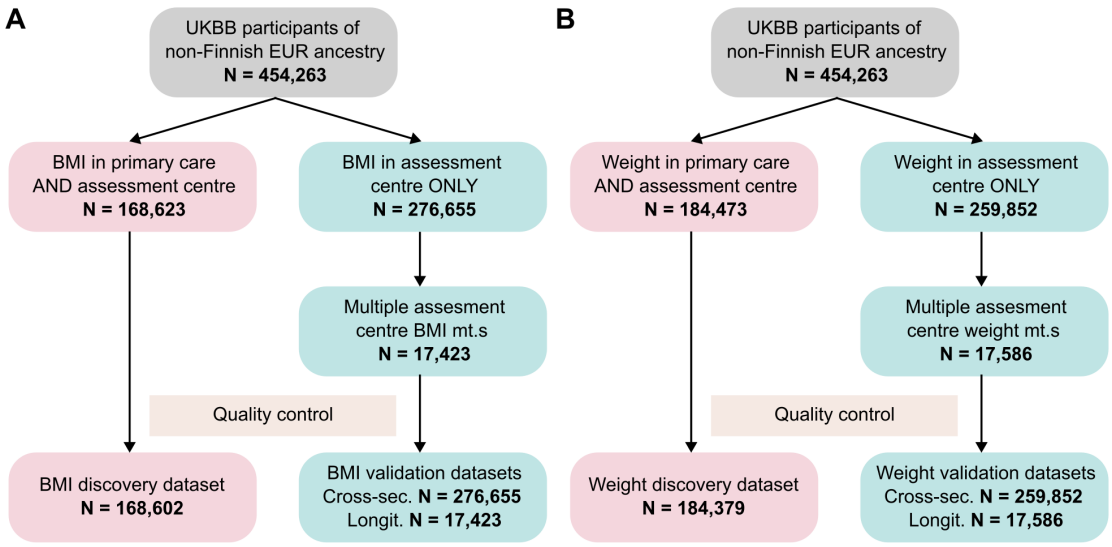


Figure 3.1: Selection of discovery and validation datasets from UK Biobank-linked primary care and assessment centre records. (A) Body mass index (BMI) data. (B) Weight data. mt.s = measurements

BMI and weight – are described in Section 2.2 of the above Chapter. For the main analyses in UKBB, we restrict our dataset to individuals of genetically ascertained European ancestry with at least one measurement in the UKBB-linked GP EHR; for genetic validation, we use participants without any GP records (for detailed Methods, see Section 3.2.4 and Figure 3.1). We proceed with 168,623 participants with multiple BMI measurements (1.33 million measurements in total) and 184,473 participants with multiple weight measurements (1.54 million measurements in total).

3.2.1 Models to define change in adiposity over time

Linear mixed-effects models

Each model below was constructed in sex-specific and sex-combined analysis strata. I calculated the rate of change of each adiposity-trait over time using linear mixed-effects models with random-slope and random-intercept terms calculated using the maximum likelihood estimate (MLE) procedure in the `lme4` package[328] in `R`[329]. The linear rate of change in trait y over the measurement window for individual i can

be approximated as the sum of the fixed and random effects $(\beta_1 + u_{i,1})$ below:

$$\begin{aligned} y_{i,j} &= x_i^T\gamma + (\beta_0 + u_{i,0}) + (\beta_1 + u_{i,1}) \cdot (t_{i,j} - t_{i,1}) + \varepsilon_{i,j} \\ u_{i,k} &\sim \mathcal{N}(0, \sigma^2_{u,k}), \quad k = 1, 2 \\ \varepsilon_{i,j} &\sim \mathcal{N}(0, \sigma^2_{\varepsilon}), \end{aligned} \tag{3.1}$$

where individual-specific covariates x_i in Equation (3.1) are: baseline age, (baseline age)^2, a one-hot encoding of data provider, year of birth, and sex (in sex-combined analysis). The model estimates variance parameters $\sigma^2_{u,k}$ and σ^2_{ε} and outputs the fixed effect model estimates $\hat{\gamma}$, $\hat{\beta}_0$, $\hat{\beta}_1$ and best linear unbiased predictors (BLUPs) of the random effects $\hat{u}_{i,0}$ and $\hat{u}_{i,1}$.

To test for an association between genetic variants and the approximated rate of change in trait y over time estimated by (3.1), I additionally linearly adjusted and rank-based inverse normally-transformed the BLUP as follows:

$$\hat{u}_{i,1} = x_{1,i}^T\gamma_1 + \varepsilon_{1,i} \tag{3.2}$$

where the slope-adjusting covariates $x_{1,i}$ in (3.2) are: the intercept BLUP $\hat{u}_{i,0}$, baseline age, (baseline age)^2, sex (in sex-combined strata), year of birth, UKBB assessment centre, number of follow-ups, and total length of follow-up (in years). The residual from Equation (3.2) provides an estimation of the slope change in trait y over time, adjusted for baseline value of y and possible confounders. Finally, I apply a deterministic rank-based inverse normal transformation[330] to the residuals from fitting model (3.2) as follows:

$$\tilde{u}_{i,1} = \Phi^{-1}\left(\frac{r(\hat{u}_{i,1} - x_{1,i}^T\hat{\gamma}_1) - c}{N - 2c + 1}\right) \tag{3.3}$$

where $r(\hat{u}_{i,1} - x_{1,i}^T\hat{\gamma}_1)$ is the rank of the ith residual among all N residuals, the offset c is 0.5, and $\Phi(\cdot)$ is the CDF of the standard Gaussian distribution.

The residuals, and intercept and slope BLUPs deviate from Gaussian distributions (Figure 3.2); such model mis-specification may lead to miscalibration of confidence intervals and hypothesis tests, although these are likely to be mitigated by large sample size, due to the central limit theorem. However, the mis-specification

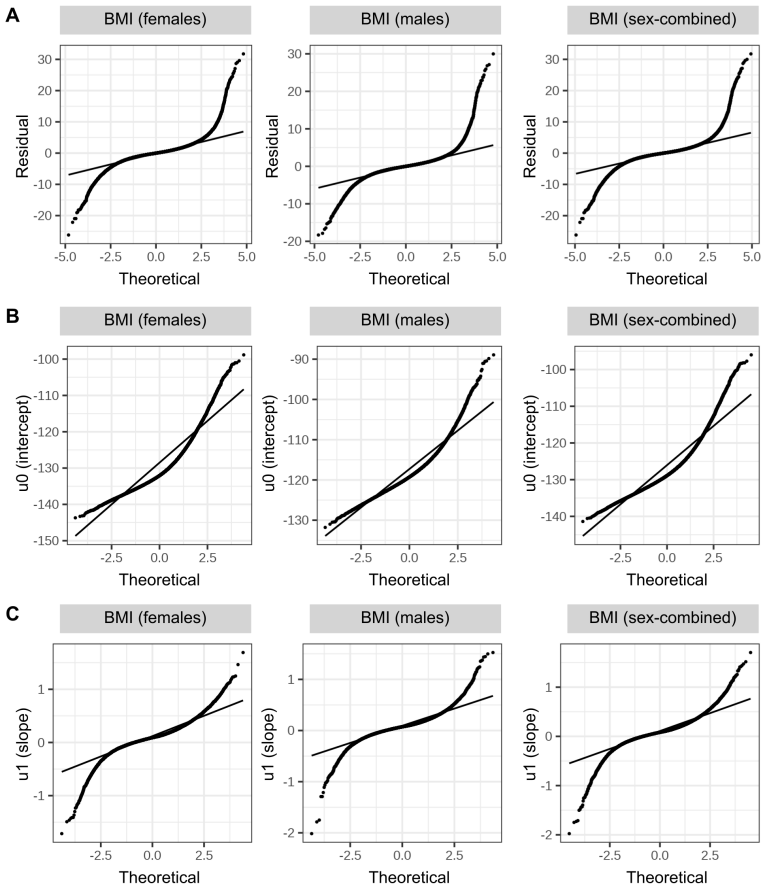


Figure 3.2: Quantile-quantile (QQ) plots for the distribution of residuals and coefficients estimated from linear mixed-effects (LME) models of change in body mass index (BMI) over time. (A) Residuals from models in each stratum (i.e. females, males, and sex-combined) plotted against theoretical quantiles of the normal distribution. (B) QQ-plots for the best linear unbiased predictor (BLUP) estimates of u0, i.e. intercept term, from models in each stratum. (C) QQ-plots for the BLUP estimates of u1, i.e. slope term, from models in each stratum.

does not affect the calibration of GWAS P-values for normally-transformed intercept and slope BLUPs.

Modelling non-linear trajectories

The models described in this section were developed in conjunction with Dr. George Nicholson and Dr. Habib Ganjgahi, who are both post-doctoral researchers in the Holmes group. I wrote, adapted, and ran scripts for these models, and analysed their output.

We model non-linear changes in BMI and weight with a regularised cubic B-spline

model (i.e. basis of degree 3) with $n_{\text{df}} = 100$ degrees of freedom. For semi-parametric regression, we use regularised splines with a relatively large number of knots to enable functional flexibility without over-fitting[323,331]. We incorporate $n_{\text{df}} - 4$ knots (as there are three degrees of freedom and one intercept) that are spaced evenly over each individual's first $T = 7500$ days (about 20.5 years) from baseline trait measurement. Conditional on the spline coefficients $\boldsymbol{b}_i$, the likelihood of measurements $\boldsymbol{y}_i$, i.e. individual i's J_i-vector of measurements from days $t_{i,1}, \ldots, t_{i,J_i}$ is:

$$p(\boldsymbol{y}_i \mid \boldsymbol{b}_i, \sigma^2) = \text{MVN}(\boldsymbol{y}_i \mid \boldsymbol{Z}_i \boldsymbol{X}_B \boldsymbol{b}_i,\ \boldsymbol{I}\sigma^2) \tag{3.4}$$

where the n_{df}-vector $\boldsymbol{b}_i$ contains the spline basis coefficients for individual i; $\boldsymbol{X}_B$ is the $(T+1) \times n_{\text{df}}$ matrix of spline basis functions evaluated at days $0, \ldots, T$ post-baseline; and $\boldsymbol{Z_i}$ is a $J_i \times (T+1)$ matrix whose j^{th} row extracts day $t_{i,j} - t_{i,1}$ post-baseline, i.e.,

$$[\boldsymbol{Z}_i]_{j,k} = \begin{cases} 1 & \text{if } k = t_{i,j} - t_{i,1} + 1 \\ 0 & \text{otherwise.} \end{cases}$$

The smoothing prior on spline coefficients $\boldsymbol{b}_i$ is specified with an order-1 autoregressive (AR(1)) model to allow these coefficients to vary smoothly around an individual-specific mean value μ_i. We specify a non-informative prior on μ_i, $\text{N}(\mu_i \mid 0, \sigma_\mu^2)$ with a large SD σ_μ. The resulting prior for $\boldsymbol{b}_i$ is:

$$\begin{aligned} p(\boldsymbol{b}_i) &= \text{MVN}(\boldsymbol{b}_i \mid 0, \boldsymbol{\Sigma}_B) \\ \boldsymbol{\Sigma}_B &:= \boldsymbol{\Sigma}_{\text{AR}(1)} + \sigma_\mu^2 \vec{\boldsymbol{1}} \\ [\boldsymbol{\Sigma}_{\text{AR}(1)}]_{k,k'} &:= \sigma^2_{AR(1)} \phi^{|k=k'|}, \end{aligned} \tag{3.5}$$

where $\boldsymbol{\Sigma}_{\text{AR}(1)}$ is the $n_{\text{df}} \times n_{\text{df}}$ autocovariance matrix for an AR(1) model with lag-1 autocorrelation $\phi \in [0, 1)$ and scale parameter $\sigma^2_{AR(1)} > 0$; $\vec{\boldsymbol{1}}$ is an $n_{\text{df}} \times n_{\text{df}}$ matrix of ones.

The closed-form posterior for the prior in Equation (3.5) and the likelihood in Equation (3.4), which are a specific case of the Bayes linear model[332], is:

$$\begin{aligned} p(\boldsymbol{b}_i \mid \boldsymbol{y}_i, \boldsymbol{\Sigma}_B, \sigma^2) &= \text{MVN}(\boldsymbol{b}_i \mid \boldsymbol{m}_i, \sigma^2 \boldsymbol{V}_i) \\ \boldsymbol{V}_i &:= (\boldsymbol{X}_B^T \boldsymbol{Z}_i^T \boldsymbol{Z}_i \boldsymbol{X}_B + \boldsymbol{\Sigma}_B^{-1})^{-1} \\ \boldsymbol{m}_i &:= \boldsymbol{V}_i \boldsymbol{X}_B^T \boldsymbol{Z}_i^T \boldsymbol{y}_i. \end{aligned} \tag{3.6}$$

Table 3.1: Median residual variance from high-dimensional spline modelling of standardised obesity trait outcomes.

Obesity trait	Sex strata	Median residual variance
Standardised body mass index (BMI), kg/m2	F	0.814
	M	0.594
	sex-combined	0.719
Standardised weight, kg	F	5.12
	M	5.15
	sex-combined	5.17

The $(\boldsymbol{y}_i, \boldsymbol{b}_i)$ are conditionally independent across individuals i given the hyper-parameters $\sigma^2_{AR(1)}$, ϕ, σ_μ, and σ^2. The posterior in Equation (3.6) can therefore be evaluated in parallel across individuals. We chose values of the hyperparameters in the smoothing prior by visualising randomly selected samples of individual adiposity trajectories: $\sigma^2_{AR(1)} := 2.5$, $\phi := 0.99$, $\sigma_\mu := 100$ (Figure 3.3). We additionally checked sensitivity to choice of parameters by comparing cluster allocations for 5,000 randomly selected individuals across the following settings of hyperparameters: (i) $\sigma^2_{AR(1)} := 0.5$, $\phi := 0.900$, $\sigma_\mu := 10$, (ii) $\sigma^2_{AR(1)} := 2.5$, $\phi := 0.990$, $\sigma_\mu := 100$, and (iii) $\sigma^2_{AR(1)} := 10$, $\phi := 0.999$, $\sigma_\mu := 500$ (Figure 3.4, see Section 3.2.1 for detailed methods).

For each adiposity-trait, in each analysis stratum (for example, BMI-female-specific, or weight-male-specific), we set σ^2 to the median of the individual-specific MLEs, i.e. $\sigma^2 := \text{median}\left\{\frac{1}{J_i}||\boldsymbol{y}_i - \boldsymbol{Z}_i\boldsymbol{X}_B\boldsymbol{m}_i||_2^2 : i = 1, \ldots, n\right\}$. Each MLE is calculated from Equation (3.4) after substituting the maximum *a posteriori* estimate for $\boldsymbol{b}_i$ with $\boldsymbol{m}_i$ from Equation (3.6) (Table 3.1).

Prior to inputting measurements $\boldsymbol{y}_i$ to the likelihood for the regularised spline model in Equation (3.4), I pre-processed $\boldsymbol{y}_i$ by taking the standardised residual from a linear regression model $\boldsymbol{y}_{i,j} = x_i^T\gamma + \epsilon_{i,j}$ for all $i = 1, \ldots, N$ individuals and $j = 1, \ldots, J_i$ time-points, with covariates for: baseline age, (baseline age)^2, a one-hot encoding of data provider, year of birth, and sex. I standardised the residuals by subtracting the mean and dividing by the SD of all residuals across all individuals and time-points.

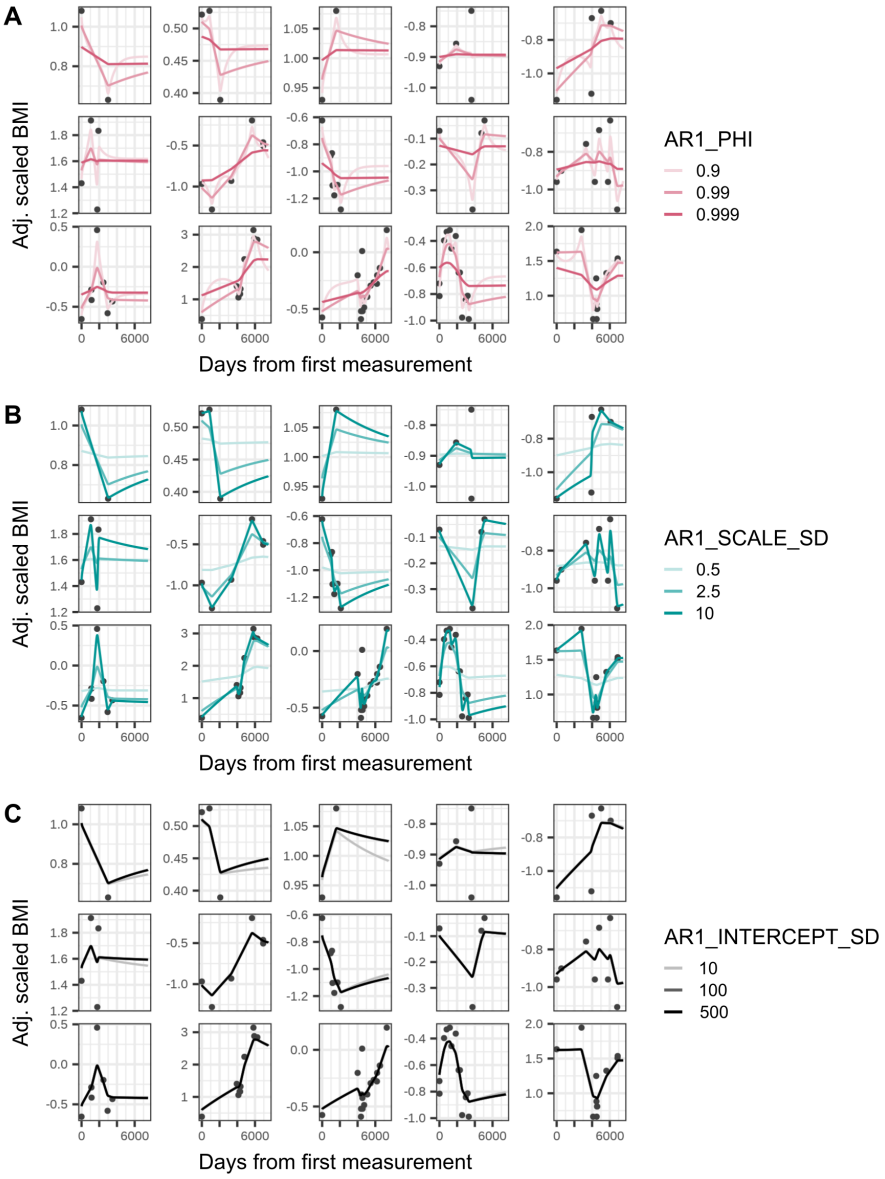


Figure 3.3: Comparison of hyperparameters for $AR(1)$ model as smoothing prior on spline coefficients for adiposity trajectories in randomly sampled individuals. Each sub-panel displays in black points the observed body mass index (BMI) values, adjusted for covariates and normally transformed (see Section 2.1.1) for a randomly selected individual. (A) Order 1-autoregressive ($AR(1)$) scale standard deviation (SD) := 2.5 and intercept SD := 100, with pink lines displaying high-dimensional B-spline model fits at different values of AR(1) autocorrelation ϕ. (B) AR(1) autocorrelation := 0.99 and intercept SD := 100, with green lines displaying model fits at different values of AR(1) scale SD. (C) AR(1) autocorrelation := 0.99 and scale SD := 2.5, with black lines displaying model fits at different values of AR(1) intercept SD.

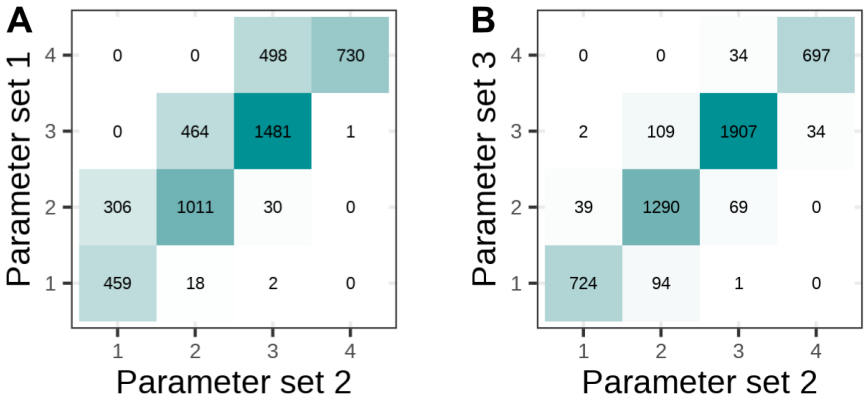


Figure 3.4: Comparison of cluster allocations across choice of $AR(1)$ model hyperparameters in the BMI sex-combined analysis. Five thousand randomly selected individuals were assigned to clusters (see Section 3.2.1) following high-dimensional B-spline modelling with different values of order 1-autoregressive ($AR(1)$) model hyperparameters for smoothing of B-spline coefficients. Parameter set 1: AR(1) autocorrelation := 0.900, AR(1) scale standard deviation (SD) := 0.5, AR(1) intercept SD := 10; parameter set 2: AR(1) autocorrelation := 0.990, AR(1) scale SD := 2.5, AR(1) intercept SD = 100; parameter set 3: AR(1) autocorrelation := 0.999, AR(1) scale SD := 10, AR(1) intercept SD = 500. The plots display the number of individuals assigned to each of the four clusters following: (A) modelling with hyperparameter sets 1 and 2; and (B) hyperparameter sets 2 and 3 respectively.

An individual i's posterior change from baseline is defined as:

$$\begin{aligned}\tilde{\boldsymbol{b}}_i &:= (0, u_{i,2} - u_{i,1}, u_{i,3} - u_{i,1}, \ldots)^T \\ &\equiv \boldsymbol{D}\boldsymbol{b}\end{aligned}$$

where the j^{th} row of $\boldsymbol{D}$ is $(e_j - e_1)^T$ and e_k is the k^{th} basis vector with n_{df} entries of zeroes everywhere except the k^{th} entry, which is one. We linearly transform the posterior in Equation (3.6) to calculate the posterior for $\tilde{\boldsymbol{b}}_i$ so that

$$p(\tilde{\boldsymbol{b}}_i \mid \boldsymbol{y}_i, \Sigma_B, \sigma^2) = \text{MVN}(\tilde{\boldsymbol{b}}_i \mid \boldsymbol{D}\boldsymbol{m}_i, \sigma^2 \boldsymbol{D}\boldsymbol{V}_i\boldsymbol{D}^T) \tag{3.7}$$

with $\boldsymbol{m}_i$ and $\boldsymbol{V}_i$ defined in Equation (3.6).

Soft clustering of individuals by non-linear trajectory patterns

The methods described in this section were developed in conjunction with Dr. George Nicholson and Dr. Habib Ganjgahi, who are both post-doctoral researchers in

the Holmes group. I wrote, adapted, and ran scripts for these methods, and analysed their output.

Figure 3.5 provides a graphical overview of the clustering protocol. Any two individuals typically have different numbers of adiposity measurements, taken at disparate ages, leading to distinct adiposity-change profiles. The precision with which any particular spline coefficient can be estimated therefore varies across individuals, which induces heteroscedasticity that must be incorporated into clustering. We define the scaled Euclidean distance between each pair of individuals (i, i') in the space of baselined spline basis coefficients as:

$$d(i,i') = \sqrt{\sum_{k=1}^{n_{\mathrm{df}}} \frac{([\boldsymbol{D}\boldsymbol{m}_i]_k - [\boldsymbol{D}\boldsymbol{m}'_i]_k)^2}{([\boldsymbol{D}\boldsymbol{V}_i\boldsymbol{D}^T]_{k,k} + [\boldsymbol{D}\boldsymbol{V}'_i\boldsymbol{D}^T]_{k,k})\sigma^2}} \tag{3.8}$$

where $\boldsymbol{m}_i$ and $\sigma^2\boldsymbol{V}_i$ are the posterior mean and covariance of individual i's spline coefficients b_i taken from Equation (3.6). For each spline coefficient k in Equation (3.8), the squared difference between individual i's and individual i''s mean coefficients is standardised by the sum of the corresponding variances.

We perform k-medoids clustering using the Partitioning Around Medoids (PAM) algorithm[333,334] implemented in the `cluster` package[335] in `R`[329]. Cluster centroids are learned from training on a randomly selected subset of 80% of individuals in each analysis stratum. For a fixed number of clusters $K := 4$, we initialise cluster membership by empirical quantiles of the estimated fold-change in obesity trait between baseline and year $M := 2$, i.e. according to bins $\mathcal{B}_{1:K}$:

$$\begin{aligned}
\mathcal{B}_k &:= \left[\hat{F}^{-1}\left(\frac{k-1}{K}\right), \hat{F}^{-1}\left(\frac{k}{K}\right)\right) \qquad k = 1, \ldots, K \\
\hat{F}(\cdot) &:= \text{empirical CDF of } \left\{\frac{[\boldsymbol{X}_B\boldsymbol{D}\boldsymbol{m}_i]_{M+1}}{[\boldsymbol{X}_B\boldsymbol{D}\boldsymbol{m}_i]_1} : \; i = 1, \ldots, N\right\} \\
\text{individual } i \text{ in bin } k &\iff \frac{[\boldsymbol{X}_B\boldsymbol{D}\boldsymbol{m}_i]_{M+1}}{[\boldsymbol{X}_B\boldsymbol{D}\boldsymbol{m}_i]_1} \in \mathcal{B}_k \; .
\end{aligned} \tag{3.9}$$

We run the clustering algorithm $S = 10$ times for robustness, each on a random sub-sample of 5,000 individuals (without replacement) in the training set. For

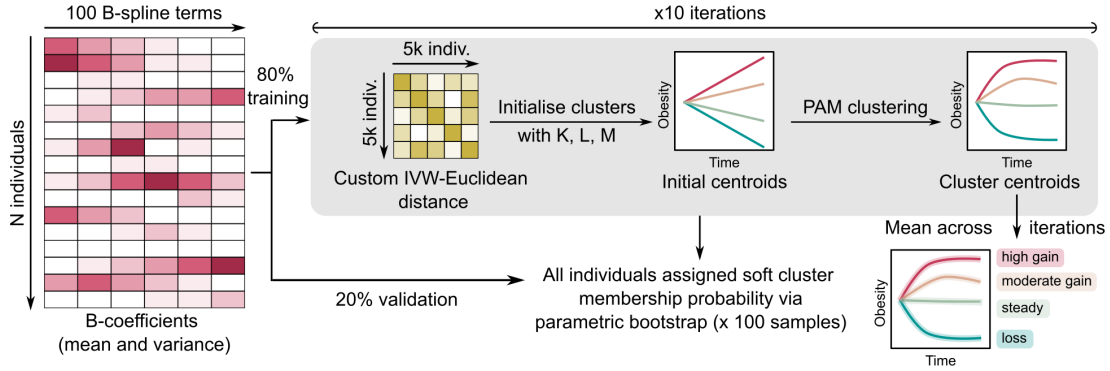


Figure 3.5: Overview of clustering protocol. In each analysis stratum, individuals are randomly assigned to training (80%) or validation (20%) sets. In each of $S = 10$ iterations, 5,000 individuals are randomly sub-sampled from the training set for the construction of a custom inverse-variance weighted (IVW) Euclidean distance matrix. Cluster centroids are determined by partitioning around medoids (PAM) clustering of the 5,000 individuals. As PAM clustering can be sensitive to starting centroid, we tested combinations of K, L, and M parameters (see Methods) to choose values of $K = 4$ (number of clusters), $L = 2$ (all 5,000 individuals used for initialisation of centroids), and $M = 2$ (centroids initialised by empirical quartile of the estimated fold-change in obesity trait between baseline and 2 years post-baseline) for the main analysis. Finally, centroids were determined by averaging across all $S = 10$ iterations after ordering of the non-overlapping and monotonic centroids. All individuals (across training and validation sets) are then assigned soft cluster membership probabilities for each cluster via a parametric bootstrap across 100 Monte Carlo samples from the B-coefficient posterior.

each clustering output $s = 1, \ldots, S$, we calculate the point-wise mean of each cluster's constituent individuals:

$$\boldsymbol{c}_{k,s} := \frac{1}{|\mathcal{C}_k^{(s)}|} \sum_{i \in \mathcal{C}_k^{(s)}} \boldsymbol{D}\boldsymbol{m}_i \tag{3.10}$$

For each clustering s, we observe that all trajectories $\boldsymbol{c}_{s,1:K}$ are monotonic and non-overlapping (Figure 3.6). We can therefore define the ordered cluster means $\boldsymbol{c}_{k,s}$,

$$k < k' \iff [\boldsymbol{c}_{(k),s}]_j > [\boldsymbol{c}_{(k'),s}]_j \quad \forall j = 1, \ldots, n_{\text{df}} \ ,$$

and average the k^{th} ordered mean across all S clusterings. The highest-adiposity cluster mean is given by $\boldsymbol{c}_{(1)}$ and the lowest-adiposity cluster mean is given by $\boldsymbol{c}_{(K)}$:

$$\boldsymbol{c}_{(k)} := \frac{1}{S} \sum_{s=1}^{S} \boldsymbol{c}_{(k),s} \ ,$$

with corresponding point-wise standard errors (SEs). We calculate silhouette values, which evaluate the similarity between cluster members (cohesion) versus non-cluster

members (separation)[336], to investigate the sensitivity of the resulting clusters to: number of clusters K, filter parameter L (which filters to individuals with a minimum of L measurements), and cluster initialisation parameter M (Figure 3.7). We test values of $K \in (2, \ldots, 8)$, $L \in \{2, 5, 10\}$, and $M \in \{1, 2, 5, 10, \text{random}\}$ (where random is random initialisation). As we implemented the `pam` function using the `cluster.only=T` option for tractability, we could not calculate values such as Akaike information criterion (AIC)based on the total pairwise dissimilarity between cluster members. Instead, by visual inspection of the silhouette plots, we chose values of $K = 4$, $L = 2$, and $M = 2$ to produce dense and separable clusters for the main analysis. To check sensitivity to an outlier training split, we compared cluster allocations over each of the $S = 10$ random training splits for a set of 5,000 randomly sampled individuals that were held out of the training sets (Figure 3.8).

Once cluster centroids were calculated, we defined individual i's soft cluster membership probability of belonging to cluster k as the posterior probability of being closest (as measured by Euclidean distance) to cluster k's centroid:

$$\pi_{i,(k)} := \int \mathbb{I}\left(k = \underset{k'}{\text{argmin}}\ ||\tilde{\boldsymbol{b}}_i - \boldsymbol{c}_{(k')}||_2\right) \text{MVN}(\tilde{\boldsymbol{b}}_i \mid \boldsymbol{D}\boldsymbol{m}_i,\ \sigma^2 \boldsymbol{D}\boldsymbol{V}_i\boldsymbol{D}^T) d\tilde{\boldsymbol{b}}_i \quad (3.11)$$

where the second term in the integrand is the posterior from Equation (3.6). We approximate the integral in Equation (3.11) by taking 100 Monte Carlo samples from the posterior.

We validate the clustering by comparing cluster properties of the 80% training set used to define cluster centroids with the held-out 20% validation set. For the purpose of validation, we assign each individual to the cluster k for which they have the highest membership probability $\pi_{i,(k)}$ and compare properties of individuals assigned to each cluster – distributions of sex, baseline age, number of follow-up measures, and length of follow-up time – as well as proportion of individuals in the training and validation sets assigned to the cluster (Table 3.2).

Finally, we create a phenotype for GWAS by taking the bounded logit-transformed cumulative cluster probabilities for each individual i, defined as: p(k1) = bounded logit($\pi_{i,(1)}$), p(k1) + p(k2) = bounded logit($\pi_{i,(1)} + \pi_{i,(2)}$), and p(k1) + p(k2) +

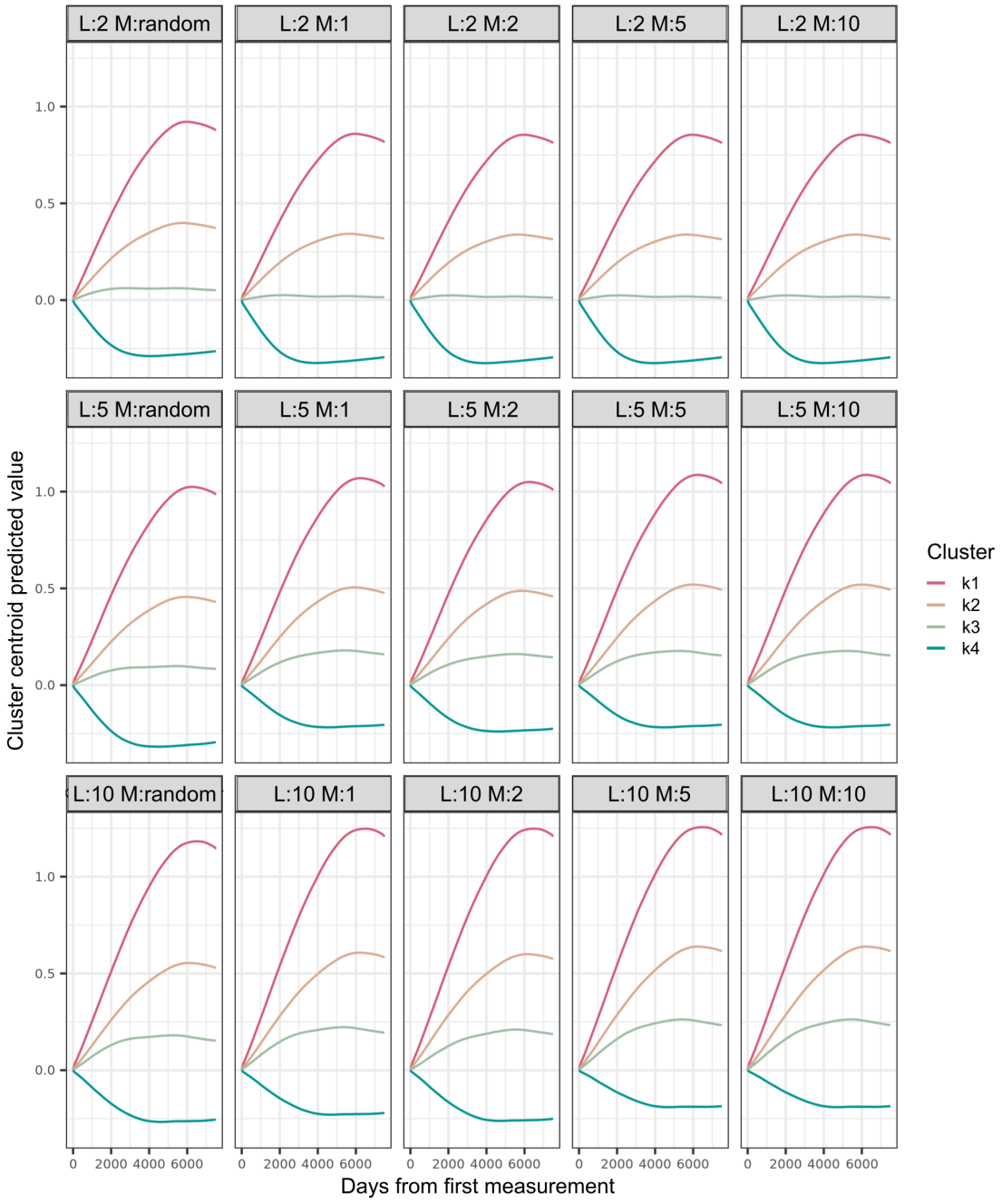


Figure 3.6: Cluster centroid trajectories for various combinations of L (filtering) and M (initialisation) parameters. Centroids are averaged over $S = 10$ iterations, with clustering parameters $K := 4$, $L =$ filter parameter (filtered to individuals with minimum number of measurements) taking values 2, 5, or 10, and $M :=$ initialisation parameter for empirical quartile of the estimated fold change in obesity trait between baseline and year M with values 1, 2, 5, or 10 years, or random initialisation. Displayed here are trajectories for cluster centroids calculated from B-spline coefficients from models of body mass index (BMI) in the sex-combined analysis; the y-axis is the predicted standardised residualised BMI.

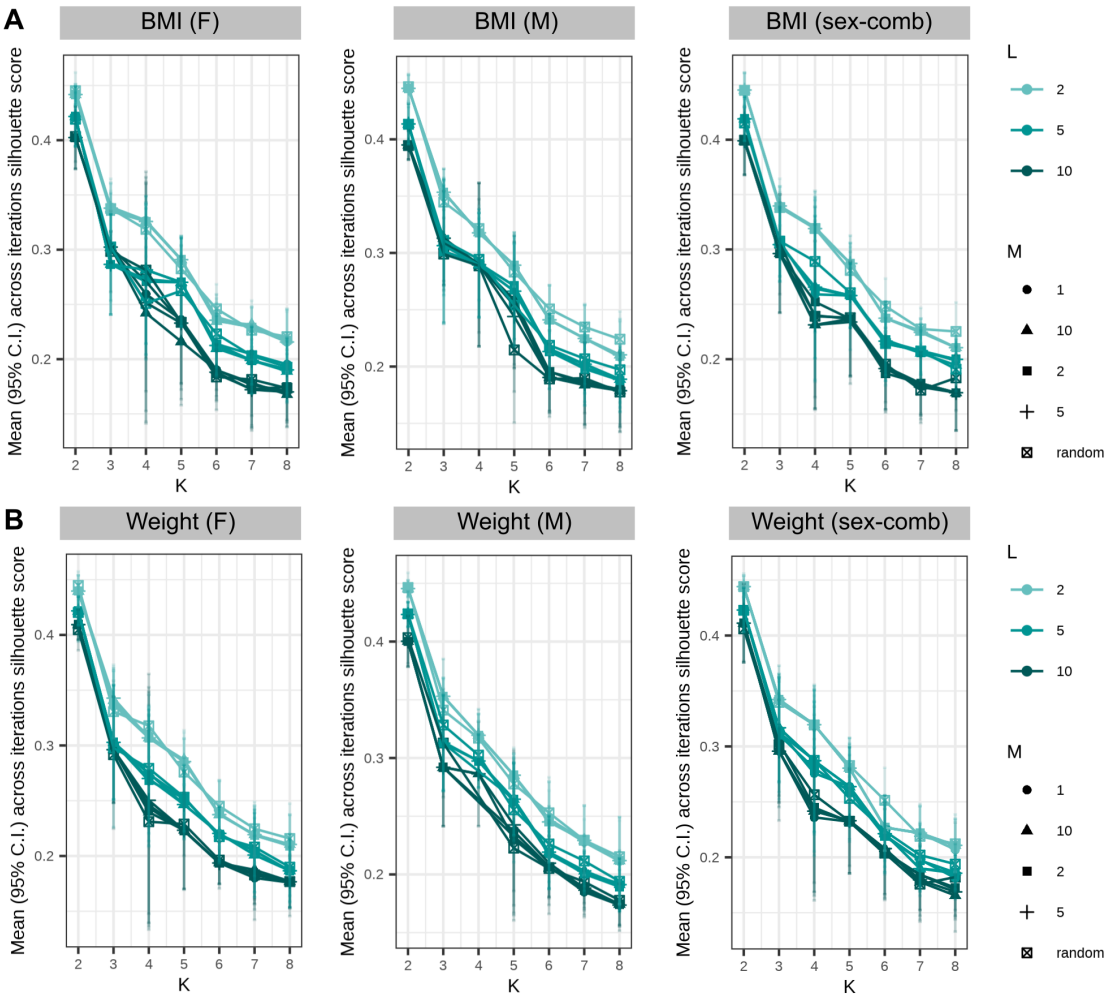


Figure 3.7: Silhouette values for selection of K (number of clusters), L (filtering), and M (initialisation) clustering parameters. $K \in (2, \ldots, 8)$, training set filtered to individuals with minimum $L \in \{2, 5, 10\}$ measurements, and $M :=$ initialisation parameter for empirical k-tile of the estimated fold-change in obesity trait between baseline and year M with values 1, 2, 5, or 10 years, or random initialisation. For each combination of parameters, partitioning around medoids (PAM) clustering is run $S = 10$ times, taking a random sub-sample of 5,000 individuals (subject to filtering parameter) at each iteration. The mean and 95% confidence interval (C.I.) silhouette value across the 10 iterations of clustering with each combination of parameters is plotted. (A) Clustering of B-spline coefficients from models of body mass index (BMI) over time in each of the three analysis strata: female (F), males (M), and sex-combined (sex-comb). (B) Same as (A), but with weight as the outcome in B-spline models.

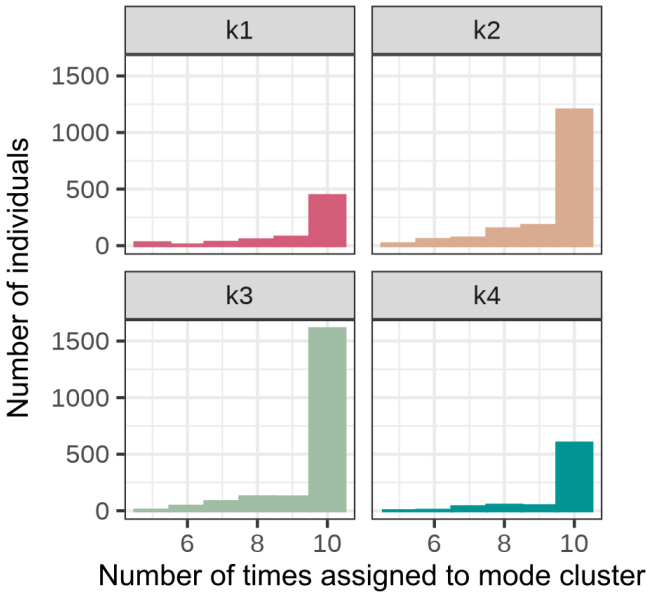


Figure 3.8: Comparison of cluster allocations across each of 10 random splits of training data (BMI sex-combined analysis). A set of 5,000 randomly selected individuals were held out of cluster centroid training and assigned to the centroids generated in each of 10 random splits of the training data. The histograms display the number of times each individual was assigned to their modal cluster - 10 indicates that the individual was assigned to the same cluster in each random split.

Table 3.2: Characterisation of adiposity-trait data in individuals assigned to their highest probability cluster in training and validation datasets. S.D. = standard deviation.

Cluster	Set	Number of individuals (%)	Mean number of repeat measurements (S.D.)	Mean length of follow-up, years (S.D.)	Mean age at first measurement, years (S.D.)	Median and I.Q.R. of trait value at first measurement
1	Training	60,154 (46.4%)	8.52 (7.52)	15.3 (6.25)	46.6 (9.31)	24.9 (22.6, 27.7)
1	Validation	15,001 (46.2%)	8.61 (8.00)	15.3 (6.23)	46.5 (9.35)	24.8 (22.5, 27.6)
2	Training	6 (0%)	35.0 (17.5)	20.4 (3.20)	51.1 (4.62)	31.0 (28.6, 36.5)
2	Validation	1 (0%)	19.0 (NA)	22.1 (NA)	41.6 (NA)	31.0 (NA)
3	Training	0	NA	NA	NA	NA
3	Validation	0	NA	NA	NA	NA
4	Training	69,606 (53.6%)	7.31 (6.85)	11.5 (6.41)	51.6 (9.35)	25.9 (23.3, 28.9)
4	Validation	17,440 (53.8%)	7.34 (6.85)	11.6 (6.43)	51.5 (9.38)	25.9 (23.3, 29.0)

p(k3) = bounded logit$(\pi_{i,(1)} + \pi_{i,(2)} + \pi_{i,(3)})$. These represent the bounded log odds of being in the highest adiposity-change cluster (k1), highest two clusters (k1 or k2), and highest three clusters (k1, k2, or k3), respectively. These outcomes were chosen as cluster centroids were ordered and monotonic, meaning that individuals with high p(k1) would generally follow p(k2) > p(k3) > p(k4); similarly, individuals with high p(k4) would generally follow p(k3) > p(k2) > p(k1). This means that adiposity-change trajectories of individuals with high p(k1) are more similar to individuals in

the moderate-gain cluster (k2) than to those in the loss cluster (k4). To prevent infinite log odds at $\pi \in \{0, 1\}$, I defined the bounded logit transform[337] as:

$$\text{bounded logit}(\pi) \equiv \text{logit}\left(\frac{(S-1)\pi + 0.5}{S}\right) \qquad \pi \in [0, 1] \ ,$$

where $S = 100$, which is the number of Monte Carlo samples from the posterior in approximating Equation (3.11).

3.2.2 Genome-wide association studies

I performed GWASs on five phenotypes for each adiposity trait (i.e. BMI and weight), in each analysis stratum (i.e. female-specific, male-specific, and sex-combined, as some adiposity traits are known to display sex-specific genetic architecture[216]), resulting in 30 GWASs in the main analysis. The five phenotypes were: LME intercept $\tilde{u}_{i,0}$ and slope $\tilde{u}_{i,1}$, and the bounded, logit-transformed soft clustering probabilities p(k1), p(k1) + p(k2), and p(k1) + p(k2) + p(k3) described above (Figure 3.9). As the clustering-derived traits were not previously adjusted for confounders, I included covariates for baseline trait value, baseline age, (baseline age^2), sex, year of birth, and assessment centre during genome-wide testing. All the genetic association tests included covariates for the first 21 genetic PCs, genotyping batch, and genotyping array, as described in Section 2.4 of Chapter 2.

I additionally conducted a sensitivity analysis to compare the baseline adiposity trait, estimated using LME intercept $\tilde{u}_{i,0}$, to the average adiposity-trait value over time for each individual. I adjusted the average trait for the covariates described in Equation (3.2) and rank-based inverse normally transformed the residual as in Equation (3.3) prior to performing GWAS.

Full details of genome-wide association testing, identification and classification of lead variants, and post-GWAS steps to estimate heritability, genetic correlations, and to test for sex-heterogeneity, are described in the relevant sections of Chapter 2.

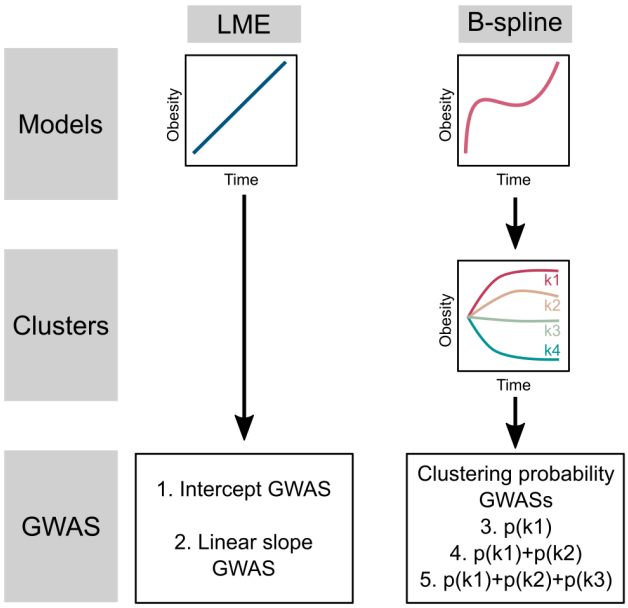


Figure 3.9: Workflow for carrying forward estimates from adiposity-change models to genome-wide association studies (GWASs). This workflow was processed in parallel for the two adiposty traits, body mass index (BMI) and weight.

3.2.3 Sensitivity analyses for survival bias

Testing the effects of covariates in LME models and GWASs

As adiposity-change associated variants may be correlated with age in UKBB due to survival bias[338], we included baseline age and (baseline age)^2 as covariates in GWASs for adiposity-change traits. To assess the impact of the age adjustment, we additionally tested for the association of lead variants with adiposity-change uncorrected for the effects of age. We directly included each lead variant in LME models for adiposity-change as explanatory variables with fixed effects for additive dosage and dosage $\times$ time interactions. We adjusted each set of models for the covariates described previously, i.e. baseline trait value, a one-hot encoding of data provider, year of birth, 21 genetic PCs, and sex (in sex-combined analyses).

We observe that follow-up metrics – length of follow-up and number of measurements of an adiposity trait – are correlated with each other (Table 3.3). As we adjust for both covariates in our LME models and GWASs, we tested whether lead variants in our study were associated with either follow-up metric. We also tested the sensitivity of our results to these adjustments by conducting GWASs

Table 3.3: Informed presence bias in UK Biobank (UKBB)-linked general practitioner (GP) electronic health records. S.D. = standard deviation.

Obesity trait	Sex	Number of measurements in GP	Number of individuals	Mean (S.D.) obesity value (BMI, kg/m2) or weight (kg) in GP records	Mean (S.D.) BMI (kg/m2) in UKBB assessment centre records
Body mass index (BMI)	F	0	15,919	-	27.0 (5.15)
		1 to 3	32,431	25.3 (4.31)	26.1 (4.39)
		4 to 9	37,116	26.1 (4.56)	26.9 (4.80)
		10+	18,506	28.9 (5.82)	29.8 (6.17)
	M	0	13,281	-	27.6 (4.10)
		1 to 3	29,616	26.1 (3.58)	26.9 (3.63)
		4 to 9	27,952	27.2 (3.83)	27.9 (4.00)
		10+	16,250	29.2 (4.71)	30.0 (4.99)
Weight	F	0	7,542	-	27.0 (5.14)
		1 to 3	33,324	67.0 (11.7)	26.0 (4.28)
		4 to 9	40,399	68.8 (12.4)	26.8 (4.70)
		10+	22,707	75.3 (16.0)	29.7 (6.21)
	M	0	6,780	-	27.5 (4.06)
		1 to 3	30,849	81.8 (12.2)	26.8 (3.57)
		4 to 9	30,306	84.4 (13.2)	27.8 (3.94)
		10+	19,164	90.0 (16.0)	30.0 (4.99)

for all of the adiposity-change traits without including length of follow-up and number of measurements as covariates.

Excluding participants with diagnosed Alzheimer's disease from data

One of the identified adiposity-change lead variants, rs429358 in the *APOE* locus, is a well-established risk factor for Alzheimer's disease and dementia[339], which are themselves associated with weight loss[340]; I therefore tested the sensitivity of all lead variants associated with adiposity-change to the inclusion of patients with diagnoses of Alzheimer's disease or dementia in our sample. I identified 272 participants with a history or diagnosis of Alzheimer's disease or dementia in primary-care records, secondary-care hospital records, or self-report, using CPRD and International Classification of Diseases (ICD) codes previously identified by Kuan *et al.* (2019)[289] (Table B.2).

Longevity SNPs associated with adiposity-change traits

To mitigate concerns that variants associated with longitudinal weight change are confounded by their effects on lifespan, I curated a list of 138 independent variants associated with longevity in the GWAS Catalog[50], accessed on 27 March 2023 (Table

B.3). I identified independent SNPs that passed genotyping and imputation QC filters in UKBB by pairwise pruning variants in LD ($r^2 > 0.1$) in a 1 Mb window to retain the variant with lower genotyping missingness. One of the lead variants identified in this study, i.e. rs429358 in the *APOE* locus, was pruned out in favour of rs4420638, which is 11kb away from the lead variant and in LD with rs429358 with $r^2 = 0.69$. We looked up the effects of these variants in the various adiposity-change GWAS summary statistics and established significance at $P = 3.60 \times 10^{-4}$ (FWER controlled at 5% across 138 tests using the Bonferroni method).

3.2.4 Replication of genetic associations in the UK Biobank

Identification and QC of adiposity phenotypes for validation

Participants with BMI and weight measurements in UKBB assessment centre visits who were not included in the GP data release, and thus had no records in the UKBB-linked EHR, were held out of our discovery analyses (Figure 3.1). There were 276,655 such individuals of European ancestry with at least one measurement of BMI and 259,852 with at least one measurement of weight, who constituted an independent set for the validation of cross-sectional results arising from the LME intercept $\tilde{u}_{i,0}$ GWAS.

For the replication of longitudinal results, i.e. LME slope $\tilde{u}_{i,1}$, we used a subset of individuals in the held-out set with multiple observations of BMI ($\mathrm{N}_{\mathrm{individuals}} = 17,423$) and weight ($\mathrm{N}_{\mathrm{individuals}} = 17,586$). I additionally curated abdominal adiposity-change phenotypes for up to 28,323 participants of European ancestry, with multiple waist circumference (WC) and hip circumference (HC) records in UKBB assessment centre visits, who were held out of discovery analyses. I derived WHR at each visit by taking the ratio of WC measured in cm to HC measured in cm as described in Section 2.2. I additionally calculated WC adjusted for BMI (WCadjBMI) and WHRadjBMI values at each visit for which WC, HC, and BMI were recorded concurrently by taking the residual of WC and WHR from linear regression models with BMI and an intercept term as predictors.

Adiposity intercept

I derived cross-sectional adiposity phenotypes for the 276,655 individuals of European ancestry with BMI measurements and 259,852 individuals with weight measurements in the held-out set of UKBB, by retaining the observed trait value closest to the individual's median trait value over time (if multiple measurements were present). I regressed the obesity trait on age, age^2, year of birth, a one-hot encoding of data provider, and sex, and applied deterministic rank-based inverse normal transformation to the residuals, as in Equation (3.3). The transformed trait was tested for association with genetic variants as described in Section 2.4 of Chapter 2.

Adiposity-change linear slope

I derived adiposity-change phenotypes using the slope BLUPs from LME models, as described in Section 3.2.1, for individuals with multiple measurements of BMI ($N_{individuals} = 17,423$), weight ($N_{individuals} = 17,586$), WC and WCadjBMI ($N_{individuals} = 29,207$), and WHR and WHRadjBMI ($N_{individuals} = 28,323$). I tested for the association of these slope traits, adjusted for the first 21 genetic PCs, genotyping batch, and genotyping array, with lead genetic variants associated with adiposity-change in our discovery analyses via the linear regression framework implemented in `PLINK`[341]. As `PLINK` does not account for family structure, I restricted the dataset to unrelated individuals by excluding the sample with higher rate of genotyping missingness in all pairs of 2nd degree or closer relatives (kinship coefficient > 0.0884).

Self-reported weight change

At each UKBB assessment centre visit, participants were asked the question "Compared with one year ago, has your weight changed?", with options to self-report: "No - weigh about the same", "Yes - gained weight", "Yes - lost weight", "Do not know", or "Prefer not to answer". I encoded the 1-year self-reported weight change response at the baseline assessment centre visit for each individual as an ordinal categorical variable with levels: "loss", "no change", and "gain", excluding

participants who did not respond or responded with "Do not know" or "Prefer not to answer". I retained 252,948 individuals of European ancestry who were not included in any of the discovery analyses, but provided self-reported 1-year weight change.

3.2.5 Replication of genetic associations in the Million Veteran Program

The analyses described in this section were performed by Gregory Linchangco Jr., under the supervision of Dr. Yan Sun, who are both associated with the Million Veteran Program and Emory University, USA. I provided the scripts for these analyses, and interpreted and produced all plots and text from their output.

QC, modelling of adiposity-change, and GWASs in the MVP cohort were all performed exactly as in the UKBB discovery analyses, with any exceptions noted below.

Cohort description

The MVP mega-biobank, with approximately 950,000 participants enrolled to date, is actively recruiting participants from the 6.9 million individuals aged 18 and over who make use of the services provided by the Veterans Health Administration (VHA), from around 50 Veterans Affairs (VA) facilities across the USA[132]. The VA Central Institutional Review Board (IRB) 10–02 protocol gained approval from the VA Central IRB in 2010, and the enrollment of study participants commenced in early 2011.

Identification and QC of longitudinal adiposity phenotypes

Weight, height, and other covariate records were compiled from the MVP Baseline Survey, which collected information on demographics, health status, lifestyle habits, military experience, and physical traits, and supplemented with EHRs. A survey cleaning algorithm was used to process self-reported data, ensuring quality through expert-defined rules, full details of which have been described previously[342]. Following population-level and individual-level QC of repeat BMI

measurements as described in Section 2.2, we retained 404,503 male European-ancestry participants with 20.6 million observations of BMI and 33,200 female European-ancestry participants with 1.94 million observation of BMI.

For each participant, we calculated linear rates of change in BMI over time with LME models as described in Section 3.2.1; we also calculated each individual's soft cluster membership probability of belonging to clusters whose centroids were defined in the UKBB discovery data (Table 3.4).

Quality control of genotyped and imputed data

Genetic data for this study was obtained from the MVP custom-genotyped dataset with imputation to the 1000 Genomes project[280] on genome build hg19, and filtered to markers with imputation information score > 0.30 with minor allele count > 30. Full characteristics of the MVP cohort[132] and associated genetic data[343] have been described previously.

GWAS

All analyses were performed in sex-specific and sex-combined strata. Genetic association analysis was performed using `REGENIE v2.2.4`[65]. All GWASs were adjusted for baseline age, (baseline age)^2, the first 10 genetic PCs, and sex (in sex-combined analyses). For the purpose of this thesis, I have only received permission to share variants reported in our associated manuscript, Venkatesh *et al.* (2023): "The genetic architecture of changes in adiposity during adulthood"[1].

3.2.6 Replication of genetic associations in the Estonian Biobank

The analyses described in this section were performed by Dr. Kadri Arumae and Dr. Uku Vainik, who are both associated with the Estonian Biobank and the University of Tartu, Estonia. I provided the scripts for these analyses, and interpreted and produced all plots and text from their output.

Table 3.4: Comparison of cluster characteristics in UK Biobank (UKBB), Million Veteran Program (MVP), and Estonian Biobank (EstBB) within each analysis stratum. Hard cluster assignments were decided as the maximum soft clustering probability of an individual belonging to each cluster. BMI = body mass index, S.D. = standard deviation, I.Q.R. = interquartile range.

Cluster	Cohort	Number of individuals (%)	Mean number of repeat measurements (S.D.)	Mean length of follow-up, years (S.D.)	Mean age at first measurement, years (S.D.)	Median and I.Q.R. of trait value at first measurement
			Females			
k1	UKBB	41,255 (46.8%)	8.62 (7.84)	15.7 (6.17)	46.0 (9.15)	24.0 (22.0, 27.1)
	MVP	17,726 (53.4%)	44.4 (42.8)	12.4 (5.7)	41.8 (12.2)	27.6 (23.9, 31.9)
	EstBB	29,763 (36.3%)	3.65 (1.97)	6.15 (4.68)	41.7 (13.8)	24.5 (21.6, 28.5)
k2	UKBB	6 (0%)	24.2 (21.0)	21.0 (3.90)	43.7 (3.72)	26.8 (24.5, 31.6)
	MVP	506 (1.52%)	99.2 (75.6)	19 (1.5)	45.4 (8.81)	28.6 (24.3, 33.1)
	EstBB	0	NA	NA	NA	NA
k3	UKBB	0	NA	NA	NA	NA
	MVP	71 (0.214%)	93.9 (59.5)	18.8 (1.51)	46.5 (8.44)	28.6 (24, 34.9)
	EstBB	0	NA	NA	NA	NA
k4	UKBB	46,982 (53.2%)	7.28 (6.85)	12.0 (6.54)	50.8 (9.55)	25.0 (22.6, 28.5)
	MVP	14,897 (44.9%)	43.6 (47.4)	11.6 (6.05)	47.2 (12.6)	29.1 (24.9, 34.1)
	EstBB	52,271 (63.7%)	3.31 (1.69)	4.34 (3.76)	43.4 (14.5)	24.4 (21.6, 28.7)
			Males			
k1	UKBB	34,570 (46.8%)	8.42 (7.41)	14.8 (6.34)	47.4 (9.46)	25.6 (23.5, 28.1)
	MVP	180,337 (44.6%)	38.3 (40.4)	11.9 (5.55)	53.4 (13.2)	28.8 (25.8, 32.3)
	EstBB	17,194 (37.6%)	3.77 (2.03)	6.08 (4.47)	40.1 (14.3)	26 (23.5, 29.1)
k2	UKBB	3 (0%)	36.0 (21.7)	24.0 (1.52)	48.8 (2.88)	29.7 (29.1, 34.8)
	MVP	4,324 (1.07%)	89.9 (63.7)	18.6 (1.6)	53.4 (13.2)	28.9 (25.8, 32.7)
	EstBB	0	NA	NA	NA	NA
k3	UKBB	2 (0%)	28.5 (6.36)	19.8 (2.33)	55.6 (5.09)	29.7 (NA)
	MVP	497 (0.123%)	97.5 (63.8)	18.3 (1.71)	52.1 (7.45)	29.1 (25.9, 32.7)
	EstBB	0	NA	NA	NA	NA
k4	UKBB	39,390 (53.2%)	7.38 (6.86)	11.0 (6.21)	52.5 (9.08)	26.6 (24.4, 29.2)
	MVP	219,345 (54.2%)	37.4 (41.2)	11.3 (5.55)	58.1 (11.3)	29.5 (26.5, 33.2)
	EstBB	28,541 (62.4%)	3.48 (1.82)	4.45 (3.66)	43.1 (14.5)	26.6 (24.1, 29.7)
			Sex-combined			
k1	UKBB	75,155 (46.3%)	8.52 (7.52)	15.3 (6.25)	46.6 (9.31)	24.9 (22.6, 27.7)
	MVP	194,097 (44.3%)	39 (40.8)	12 (5.57)	52.2 (13.5)	28.7 (25.7, 32.3)
	EstBB	46,588 (36.5%)	3.69 (1.99)	6.13 (4.61)	41.1 (13.9)	25.1 (22.2, 28.8)
k2	UKBB	7 (0%)	35.0 (17.5)	20.4 (3.20)	51.1 (4.62)	31.0 (28.6, 36.5)
	MVP	4,758 (1.09%)	89.4 (62.6)	18.7 (1.63)	51.6 (7.78)	28.8 (25.6, 32.6)
	EstBB	0	NA	NA	NA	NA
k3	UKBB	0	NA	NA	NA	NA
	MVP	613 (0.14%)	105 (72)	18.2 (1.83)	52.2 (7.21)	29.3 (26.3, 33.8)
	EstBB	0	NA	NA	NA	NA
k4	UKBB	87,046 (53.7%)	7.31 (6.85)	11.5 (6.41)	51.6 (9.35)	25.9 (23.3, 28.9)
	MVP	238,235 (54.4%)	37.7 (41.6)	11.3 (5.58)	57.4 (11.7)	29.4 (26.4, 33.2)
	EstBB	81,181 (63.5%)	3.37 (1.74)	4.38 (3.73)	43.3 (14.6)	25.4 (22.4, 29.2)

QC, modelling of adiposity-change, and GWASs in the Estonian Biobank (EstBB) cohort were all performed exactly as in the UKBB discovery analyses, with any exceptions noted below.

Cohort description

EstBB is a volunteer-based sample of Estonian residents comprising about 20% of the Estonian adult population (N > 210,000), recruited by medical personnel and

through media campaigns. Various health and demographic data have been collected from the participants, both by medical workers and via self-reports, since 2002. The cohort has previously been described in detail by Leitsalu *et al.* (2015)[128].

Identification and QC of longitudinal adiposity phenotypes

Repeat BMI measurements, collected by doctors in EHRs, were available for 193,528 participants. Individuals with records of bariatric surgery or use of glucagon-like-peptide-1 (GLP-1) inhibitors (blood glucose lowering drugs that typically also result in weight loss, drug codes A10BJ*), such as semaglutide, were excluded.

For each participant, we calculated linear rates of change in BMI over time with LME models as described in Section 3.2.1; we also calculated each individual's soft cluster membership probability of belonging to clusters whose centroids were defined in the UKBB discovery data (Table 3.4).

Quality control of genotyped and imputed data

Genetic data for this study was obtained from genotyping with the Illumina Global Screening Array (GSA) microchip at 700,000 markers and imputed to the hg19 reference genome using a custom panel[344]. Full characteristics of the EstBB cohort[128] and associated genetic data[344] have been described previously.

GWAS

All analyses were performed in sex-specific and sex-combined strata. Genetic association analysis was performed using `REGENIE v3.2` software[65]. All GWASs were adjusted for baseline age, (baseline age)^2, the first 20 genetic PCs, and sex (in sex-combined analyses).

3.2.7 Power calculations for replication sample sizes

I corrected the observed variant effect sizes from discovery GWASs for winner's curse through an implementation first described by Palmer *et al.* (2017)[345]. Briefly, I solve for the bias using the following maximum likelihood model,

$$\beta_{obs} = \beta_{true} + s\frac{\phi(\frac{\beta_{true}}{s} - c) - \phi(\frac{-\beta_{true}}{s} - c)}{\psi(\frac{\beta_{true}}{s} - c) + \psi(\frac{-\beta_{true}}{s} - c)}$$

where β_{obs} is the effect size in the discovery GWAS, β_{true} is the (assumed true) effect size in the source population, and $c = 5.33$ is the test statistic corresponding to a discovery $\alpha = 5 \times 10^{-8}$. The above method is valid for a single variant, whereas when testing hundreds of variants, it may be more appropriate to implement an empirical Bayes correction as suggested by Ferguson *et al.* (2013) across all variants in a genome-wide scan[346]. The sample size required to replicate the (assumed true) unbiased effect size is then calculated for $\alpha = \frac{0.05}{M_{var}}$) (FWER controlled at 5% across the number of independent variants tested, M_{var}, using the Bonferroni method), as follows:

$$\text{power}(\alpha, \text{ncp}) = 1 - \chi^2_1((\chi^2_1)^{-1}(1-\alpha), \text{ncp})$$

under the alternative distribution which is non-central χ^2_1 with non-centrality parameter (ncp) per variant estimated for a normalised trait with variance 1 as:

$$\text{ncp} \approx N\frac{2\beta^2_{true}\text{AF}(1-\text{AF})}{1 - 2\beta^2_{true}\text{AF}(1-\text{AF})}$$

where AF is the variant allele frequency.

3.2.8 Statistical power comparison to published GWASs

I accessed publicly available summary statistics for sex-specific and sex-combined analyses of cross-sectional BMI from the 2019 Genetic Investigation of ANthropometric Traits (GIANT) Consortium meta-analysis[216], which included 434,793 women and 374,755 men. Variants that reached GWS in either the GIANT meta-analysis or our in-house BMI LME model intercept GWAS, and were reported in both studies, were carried forward for power comparison: resulting in 3,620 SNPs in the female-specific analysis, 1,818 SNPs in the male-specific analysis, and 10,447 SNPs in the sex-combined analysis. I calculated the χ^2 statistic as β^2/σ^2 for each SNP and obtained the ratio of $\chi^2_{\text{in-house}}$ to χ^2_{GIANT}. By comparing the median

$\chi^2_{\text{in-house}}/\chi^2_{\text{GIANT}}$ to the median sample-size ratio $N_{\text{in-house}}/N_{\text{GIANT}}$, we quantified the boost in power afforded by our longitudinally-derived phenotype over that expected from sample size differences between the two studies.

3.2.9 Longitudinal phenome-wide association studies

I used a two-stage approach to test the identified lead variants for association with change over time in any of 45 quantitative phenotypes in the longitudinal research resource (Table 2.1, all non-adiposity traits with > 1,000 participants with repeat measurements were included). In the first stage, I calculated the linear rate of change in each trait using the LME approach outlined in Section 3.2.1, adjusted for the covariates described in Equation (3.2) and rank-based inverse normally transformed the residual as in Equation (3.3). In the second-stage, I tested for the additive association of number of copies of the minor allele (0, 1, or 2) of each lead genetic variant with the appropriately adjusted and transformed slope BLUPs $\tilde{u}_{i,1}$ for all quantitative traits, with a linear regression model adjusted for the first 21 genetic PCs, genotyping array, genotyping batch, and sex (in sex-combined analyses).

3.2.10 Joint modelling of intra-individual mean and variance

The following analyses were performed using the `TrajGWAS` package[147] in `Julia`[347], for 177,472 unrelated individuals of European ancestry with multiple measurements of weight in the UKBB-linked EHR resource used for discovery analyses above. Briefly, TrajGWAS analysis is conducted in two stages to test for genetic effects on longitudinal trajectory mean, intra-individual variance, and a joint effect on either mean or variance in a LME model framework[147]. In the first stage, I fit a null model for weight with fixed effects for the intercept, age, age^2, sex, and 21 genetic PCs and include random effects for intercept and age. In the second stage, I perform score testing with the saddle-point approximation under the full model, i.e. including genome-wide effects for all variants with MAF > 1% in the genotyped and imputed UKBB data that passed QC as described in Section 2.3.

3.3 Results

3.3.1 Characteristics of baseline adiposity and adiposity-change

Data description

We obtained 2.95 million adiposity-trait (BMI and weight) records for up to 184,379 individuals by combining UKBB assessment centre records with longitudinal data from GP-linked EHR (Table 3.5). Participants in this dataset have, on average, 7.82 (SD = 7.20) observations of BMI and 8.30 (8.02) observations of weight. The mean age at first measurement is between 48.3 to 50.1 years for the different phenotypes in men and women, with an average follow-up length of between 12.7 to 13.9 years. As with the general UKBB dataset, which has an over-representation of healthy, female participants compared to the population of the UK[264,348], our discovery dataset has an over-representation of women (54.3% female); and the average BMI at baseline (25.6 kg/m^2 for women and 26.6 kg/m^2 for men) is lower than the national average, which was 27.6 kg/m^2 for both men and women in 2019[349].

We sought replication of genetic findings in two additional cohorts with longitudinal EHR data – MVP (N = 437,703) based in the USA and EstBB (N = 127,760) based in Estonia – whose demographic and obesity trait characteristics are distinct from UKBB (Table 3.5). Individuals in MVP are predominantly male (92.4%), and are on average 3.0 units of BMI heavier than male participants in the UKBB (mean UKBB = 26.6 kg/m^2, mean MVP = 29.6 kg/m^2), but are followed up for similar lengths of time (mean UKBB = 12.7 years, mean MVP = 14.0 years). On the other hand, participants in the EstBB population are more likely to be female (64.2% female) and have similar BMI to female participants in the UKBB (mean UKBB = 25.6 kg/m^2, mean EstBB = 24.5 kg/m^2), but are approximately six years younger at baseline (mean UKBB = 48.6 years, mean EstBB = 42.8 years) and have fewer repeat BMI records (mean UKBB = 7.82, mean EstBB = 3.83).

Table 3.5: Characterisation of adiposity-trait data in longitudinal records curated from UK Biobank, Million Veteran Program, and Estonian Biobank. S.D. = standard deviation.

Obesity trait	Sex	Number of individuals (%)	Number of measurements	Mean number of repeat measurements (S.D.)	Mean length of follow-up, years (S.D.)	Mean age at first measurement, years (S.D.)	Median and I.Q.R. of trait value at first measurement
Discovery (UK Biobank)							
Body mass index (BMI), kg/m^2	F	91,556 (54.3%)	740,772	7.82 (7.26)	13.6 (6.64)	48.6 (9.64)	25.6 (4.85)
	M	77,046 (45.7%)	625,630	7.82 (7.14)	12.7 (6.54)	50.1 (9.57)	26.6 (4.05)
Weight, kg	F	100,343 (54.4%)	865,247	8.37 (8.23)	13.9 (6.62)	48.3 (9.58)	67.7 (13.3)
	M	84,036 (45.6%)	715,083	8.23 (7.82)	12.8 (6.57)	49.9 (9.53)	83 (13.6)
Replication (Million Veterans Program)							
Body mass index (BMI), kg/m^2	F	33,200 (7.6%)	1,976,676	87.33 (87.9)	15.08 (5.06)	45.3 (11.7)	29.2 (25.0, 34.1)
	M	404,503 (92.4%)	22,046,815	77.14 (97.2)	14.02 (5.01)	55.8 (10.7)	29.6 (26.4, 33.4)
Replication (Estonian Biobank)							
Body mass index (BMI), kg/m^2	F	82,032 (64.2%)	281,412	3.43 (1.80)	4.7 (4.00)	42.8 (14.29)	24.5 (21.6, 28.7)
	M	45,728 (35.8%)	164,135	3.59 (1.91)	4.6 (3.84)	42.0 (14.48)	26.4 (23.9, 29.6)

Adiposity-change phenotypes for GWAS

For each individual, I estimated change in BMI or weight over time using: (1) LME models with both fixed and random effects of intercept and age, which allowed for the estimation of linear rate of change in adiposity, and (2) regularised high-dimensional B-spline models to capture non-linear patterns of temporal change (Figure 3.10, see Section 3.2.1 for detailed methods). To identify groups of individuals with similar patterns of change in adiposity over time, we performed k-medoids clustering on the posterior high-dimensional B-spline coefficients to define four clusters of adiposity-change: high gain (k1), moderate gain (k2), stable (k3), and loss (k4); we then estimated each individual's posterior probability of belonging to a cluster (Figure 3.11). Clusters were validated through a series of sensitivity analyses (see Section 3.2.1 for details), and cluster characteristics are consistent across UKBB, MVP, and EstBB (Table 3.4). Each set of models was constructed in six analysis "strata" – female-specific BMI and female-specific weight, male-specific BMI and male-specific weight, and sex-combined BMI and sex-combined weight – and adjusted for confounding variables of baseline age, (baseline age)^2, year of birth, data provider, and sex (in sex-combined analyses).

Per analysis stratum, I carried five sets of phenotypes forward for GWAS: (i) intercept BLUPs from LME models, which represent baseline adiposity, (ii) slope BLUPs from LME models, which represent linear change in adiposity over time, and logit-transformed probabilities of belonging to (iii) the high gain cluster (k1), (iv) the high or moderate gain clusters (k1 or k2), and (v) all but the loss cluster (k1, k2, or k3) (Figure 3.9).

3.3.2 Genetic architecture of adiposity-change in adulthood

Genetic variants have effects on change in weight and BMI over time, independent of baseline obesity

We identified nine unique lead variants associated with at least one adiposity-change phenotype at GWS ($P < 5 \times 10^{-8}$) (Table 3.6). Four of the nine SNPs have no previously reported cross-sectional obesity associations, including: rs747792 (MAF

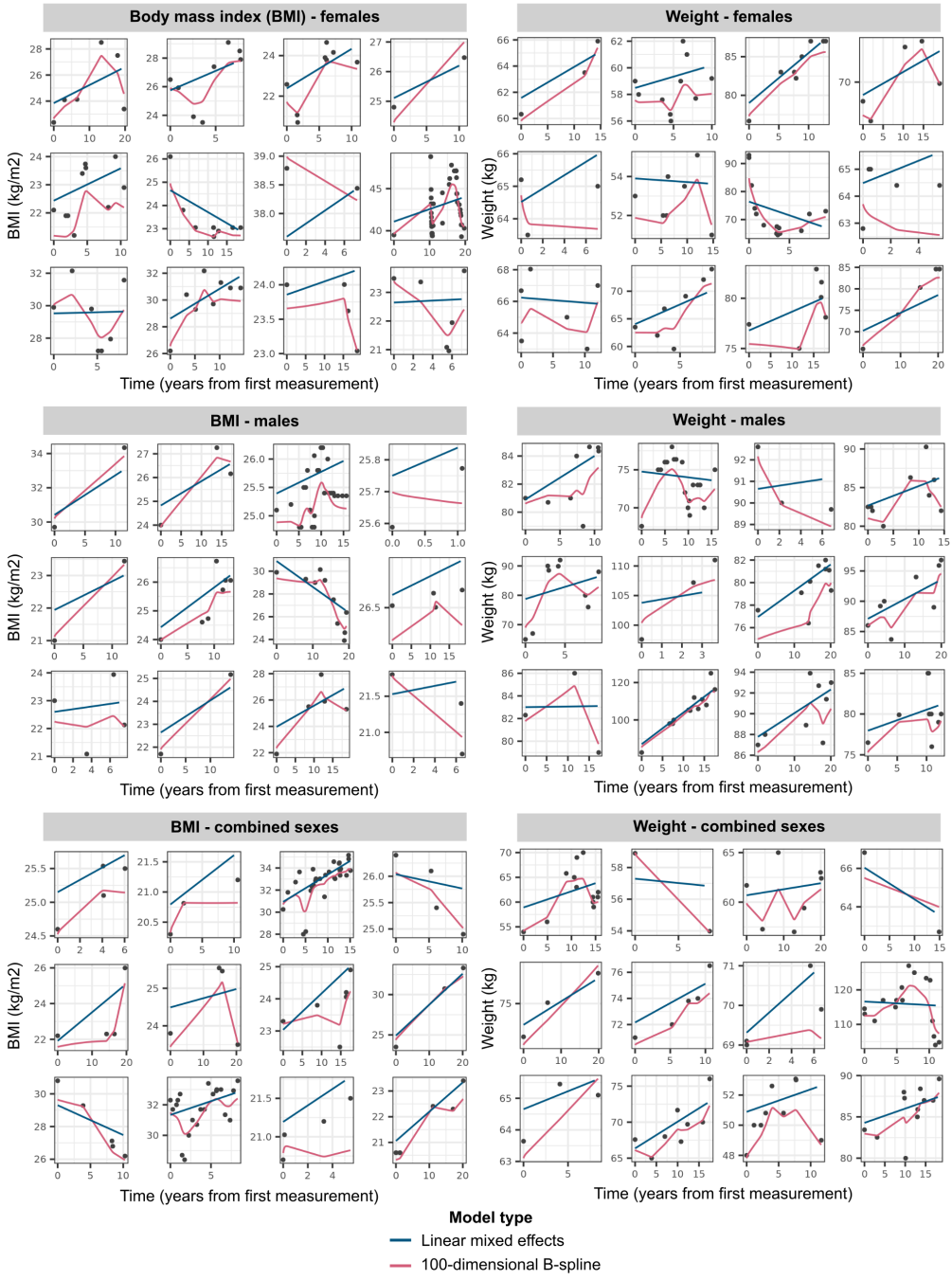


Figure 3.10: Trajectories of BMI and weight change modelled by linear and high-dimensional B-splines in a random sample of individuals. Black points display observed records of BMI or weight, with blue and pink lines representing fits from linear mixed-effects models and regularised 100-dimensional B-spline models respectively. Top row: female-specific models, middle row: male-specific models, bottom row: combined-sexes models. All models included covariates for baseline age, (baseline age)^2, data provider, and year of birth. Combined-sexes models additionally included sex as a covariate.

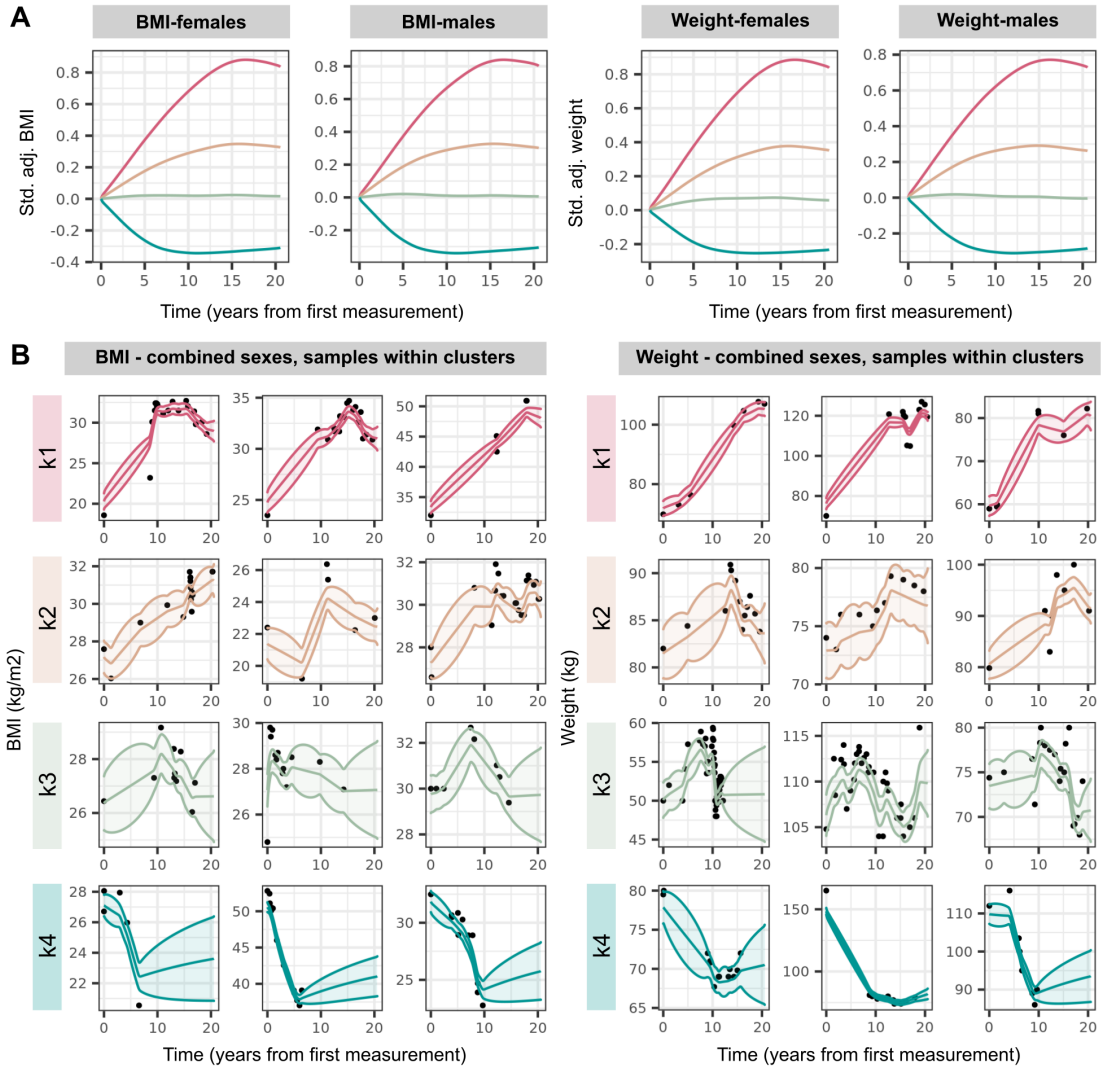


Figure 3.11: Random sample of individuals with high probability of assignment to each adiposity-change cluster. (A) Trajectories of cluster centroids in the sex-specific stratum are plotted as standardised (std.) and covariate-adjusted (adj.) body mass index (BMI) or weight over time, as years from first measurement. The four clusters k1 (high gain, red), k2 (moderate gain, orange), k3 (stable, green), and k4 (loss, blue) were determined via partitioning around medoids (PAM) clustering with a customised Euclidean distance matrix constructed from high-dimensional B-spline coefficients for change in adiposity over time. (B) BMI and weight trajectories over time are plotted for a random sample of individuals in the 99th percentile of probability of belonging to each cluster, as determined by parametric bootstrap. The lines display predicted fits and ribbons represent 95% confidence intervals around the mean predicted fit.

= 30.8%), whose nearest gene *POMC* harbours rare variants that cause a severe monogenic form of obesity due to pro-opiomelanocortin (POMC) deficiency[350], and which is an eQTL for *ADCY3*[351], loss of which causes severe monogenic obesity[352], and rs186269301 (MAF = 2.78%) which is reported to associate with high-density lipoprotein (HDL) cholesterol and SHBG in the UKBB[353], which are both correlated with obesity. I also find a female specific variant, rs9373305 (MAF = 39.6%) in the *GJE1* gap-junction protein locus (P-value for heterogeneity between female-specific and male-specific effects $P_{\text{sex-het}} = 0.00321$), which is associated with lower odds of belonging to the high gain, moderate gain, or stable weight cluster in women (odds ratio (OR) (95% confidence interval (CI)) = 0.940 (0.920 - 0.961), $P = 4.07 \times 10^{-8}$).

Variants previously reported to associate with cross-sectional obesity display concordant directions of effect on adiposity-change phenotypes in our study, independent of their effects on baseline trait value. For example, the minor A allele of rs7206122 (MAF = 39.4%) in the *FTO* locus and the minor A allele of rs11661876 (MAF = 26.5%) in the *MC4R* locus, which are both reported to associate with lower weight in cross-sectional analyses of the UKBB[354], are associated with lower odds of belonging to the high gain, moderate gain, or stable weight clusters (OR (95% CI) for rs7206122 = 0.952 (0.937 - 0.968), $P = 6.28 \times 10^{-9}$; OR for rs11661876 = 0.945 (0.927 - 0.962), $P = 1.30 \times 10^{-9}$), independent of baseline weight. Moreover, rs35519679 (MAF = 25.4%) in the *TCF7L2* locus, whose minor allele A is associated with increased risk of type 2 diabetes in both men and women[355], has male-specific effects on BMI loss in our study (β (SE) = 0.0327 (0.0059) SD decrease in expected BMI slope over time, $P = 3.01 \times 10^{-8}$, $P_{\text{sex-het}} = 0.00311$). Finally, a missense variant in the *APOE* locus, rs429358 (MAF = 15.5%), is associated with all BMI and weight change metrics in all strata defined here, and is discussed in detail in Section 3.3.3.

Multi-cohort replication of variants associated with adiposity-change

We sought replication of all adiposity-change lead variants by testing their association with linear slope of BMI over time and BMI-change cluster belonging, in individuals held out of discovery analyses. In external cohorts, we calculated non-linear

Table 3.6: Lead SNPs identified from genome-wide association studies (GWASs) of linear mixed-effects (LME) model slopes of adiposity or posterior probability of membership in adiposity-change clusters (high gain (k1), high or moderate gain (k1 or k2), or all but loss (k1, k2, or k3)), independent of baseline obesity. MAF = minor allele frequency, TSS = transcription start site, SE = standard error, OR = odds ratio, CI = confidence interval

SNP	chr:pos (hg19)	MAF	Nearest TSS	Adiposity trait	Sex-combined		Female-specific		Male-specific		Sex heterogeneity P-value
					Beta (SE) or OR (95% CI)	P-value	Beta (SE) or OR (95% CI)	P-value	Beta (SE) or OR (95% CI)	P-value	
Linear slope (u1)											
rs811041	6:26226032	0.282	H3C6	Weight	-0.0205 (0.00365)	2.07E-08	-0.0158 (0.00493)	0.00134	-0.0244 (0.00542)	6.52E-06	0.24
rs35519679	10:114818754	0.254	TCF7L2	BMI	-0.0185 (0.00399)	3.57E-06	-0.00895 (0.00544)	0.0996	-0.0327 (0.0059)	3.01E-08	0.00311
rs429358	19:45411941	0.155	APOE	BMI	-0.0591 (0.00475)	1.49E-35	-0.0661 (0.00645)	1.22E-24	-0.0493 (0.00704)	2.51E-12	0.0775
rs429358	19:45411941	0.155	APOE	Weight	-0.0614 (0.00454)	1.4E-41	-0.0659 (0.00616)	1.05E-26	-0.0538 (0.00673)	1.32E-15	0.183
p(k1) High gain cluster											
rs186269301	12:123917276	0.0278	RILPL2	BMI	1.12 (1.08 - 1.17)	2.65E-09	1.12 (1.06 - 1.18)	8.15E-05	1.07 (1.01 - 1.14)	0.0139	0.325
rs429358	19:45411941	0.155	APOE	BMI	0.913 (0.897 - 0.929)	9.48E-25	0.902 (0.88 - 0.925)	4.03E-16	0.922 (0.899 - 0.946)	7.6E-10	0.217
p(k1) + p(k2) High or moderate gain clusters											
rs714827	14:103113552	0.0814	RCOR1	Weight	1.12 (1.08 - 1.16)	8.07E-09	1.15 (1.09 - 1.21)	2.48E-07	1.10 (1.04 - 1.16)	9.80E-04	0.275
rs429358	19:45411941	0.155	APOE	BMI	0.900 (0.884 - 0.916)	1.55E-31	0.883 (0.861 - 0.906)	1.57E-21	0.923 (0.899 - 0.948)	2.8E-09	0.0172
rs429358	19:45411941	0.155	APOE	Weight	0.844 (0.82 - 0.869)	4.71E-31	0.833 (0.801 - 0.867)	1.1E-19	0.850 (0.815 - 0.887)	6.86E-14	0.494
p(k1) + p(k2) + p(k3) All but loss cluster											
rs747792	2:25473547	0.308	POMC	BMI	0.963 (0.95 - 0.976)	2.33E-08	0.957 (0.940 - 0.975)	3.48E-06	0.972 (0.953 - 0.991)	0.00464	0.257
rs9373305	6:141683547	0.396	GJE1	Weight	0.961 (0.945 - 0.977)	2.43E-06	0.94 (0.92 - 0.961)	4.07E-08	0.989 (0.964 - 1.01)	0.37	0.00321
rs7206122	16:53822440	0.394	RPGRIP1L/FTO	Weight	0.952 (0.937 - 0.968)	6.28E-09	0.946 (0.925 - 0.967)	6.65E-07	0.96 (0.937 - 0.985)	0.00153	0.361
rs11661876	18:57889652	0.265	MC4R	Weight	0.945 (0.927 - 0.962)	1.3E-09	0.958 (0.935 - 0.982)	0.000587	0.93 (0.904 - 0.956)	3.85E-07	0.12
rs429358	19:45411941	0.155	APOE	BMI	0.905 (0.89 - 0.921)	6.92E-31	0.892 (0.871 - 0.913)	1.37E-21	0.920 (0.897 - 0.943)	5.66E-11	0.0809
rs429358	19:45411941	0.155	APOE	Weight	0.853 (0.834 - 0.872)	1.8E-44	0.847 (0.822 - 0.873)	1.51E-27	0.864 (0.835 - 0.894)	2.6E-17	0.398

trajectories of BMI-change with high-dimensional B-spline models as described previously; individuals were then assigned to clusters based on the centroids previously determined in the UKBB discovery analyses. We were adequately powered (80% power), at $P < 5.55 \times 10^{-3}$ (FWER controlled at 5% across 9 variants using the Bonferroni method), to replicate the effects of: (1) eight of nine variants for adiposity-change in a held-out sample of up to 39,388 individuals of European ancestry in the UKBB who were not included in the discovery analyses, (2) all variants in 125,209 participants in the EstBB cohort, and (3) all variants in 437,703 participants in the MVP cohort (Table 3.7). Four of the nine variants replicate at $P < 5.55 \times 10^{-3}$ and have consistent directions of effect between discovery and replication analyses in UKBB, i.e. alleles which lower an individual's posterior probability of belonging to a BMI or weight gain cluster in the discovery GWASs are associated with decrease in expected BMI or weight slope over time in the replication analyses (Table 3.8). Two of these, i.e. rs7206122 in the *FTO* locus and rs429358 in the *APOE* locus, also replicate with consistent direction of effect on BMI-change in the EstBB cohort (Table 3.8 and Figure 3.12). Finally, the minor allele of rs429358 is also associated with BMI loss in the MVP dataset (Figure 3.12). Although we were powered to seek replication of the other variants in MVP, data-sharing permissions were not secured for publication of this thesis, hence are not reported.

Adiposity-change and baseline adiposity have different magnitudes of heritability and are modestly genetically correlated

The heritability estimates of all adiposity-change phenotypes, i.e. LME slopes and posterior probabilities of belonging to various combinations of adiposity-change clusters, ranged between 0.998% (SE = 0.47) and 3.47% (0.52), as compared to baseline trait heritability estimates of between 20.6% (1.14) and 26.7% (1.35) (Figure 3.13). The ratio of SNP-based heritability (h^2_G) of adiposity LME intercepts to their slope counterparts was between 7.6 to 17.0 (h^2_G of female-specific BMI intercept = 20.6% (1.14) vs BMI slope = 2.70% (0.50); h^2_G of male-specific BMI intercept = 21.2% (1.32) vs BMI slope = 1.25% (0.54)). Further, I observed that female-specific h^2_G of BMI-change was significantly higher than the male-specific

Table 3.7: Sample sizes required for replication of adiposity-change lead SNPs, with effect sizes debiased for winner's curse. BMI = body mass index, SE = standard error

SNP	chr:pos (hg19)	MAF	Adiposity trait	Beta (SE) or OR (95% CI)	P-value	Minimum replication sample size (80% power)
Linear slope (u1)						
rs811041	6:26226032	0.282	Weight	-0.0205 (0.00365)	2.07E-08	84,900
rs35519679	10:114818754	0.254	BMI	-0.0327 (0.0059)	3.01E-08	35,800
rs429358	19:45411941	0.155	BMI	-0.0591 (0.00475)	1.49E-35	14,300
rs429358	19:45411941	0.155	Weight	-0.0614 (0.00454)	1.4E-41	13,200
p(k1) High gain cluster						
rs186269301	12:123917276	0.0278	BMI	1.12 (1.08 - 1.17)	2.65E-09	19,700
rs429358	19:45411941	0.155	BMI	0.913 (0.897 - 0.929)	9.48E-25	6,000
p(k1) + p(k2) High or moderate gain clusters						
rs714827	14:103113552	0.0814	Weight	1.12 (1.08 - 1.16)	8.07E-09	7,800
rs429358	19:45411941	0.155	BMI	0.900 (0.884 - 0.916)	1.55E-31	4,400
rs429358	19:45411941	0.155	Weight	0.844 (0.82 - 0.869)	4.71E-31	1,700
p(k1) + p(k2) + p(k3) All but loss cluster						
rs747792	2:25473547	0.308	BMI	0.963 (0.95 - 0.976)	2.33E-08	19,600
rs9373305	6:141683547	0.396	Weight	0.94 (0.92 - 0.961)	4.07E-08	7,900
rs7206122	16:53822440	0.394	Weight	0.952 (0.937 - 0.968)	6.28E-09	12,600
rs11661876	18:57889652	0.265	Weight	0.945 (0.927 - 0.962)	1.3E-09	11,200
rs429358	19:45411941	0.155	BMI	0.905 (0.89 - 0.921)	6.92E-31	5,000
rs429358	19:45411941	0.155	Weight	0.853 (0.834 - 0.872)	1.8E-44	2,000

h^2_{G} in all strata (sex-heterogeneity P-value $P_{\text{sex-het}} < 0.05$), but saw no such sex difference in the heritability estimates of baseline BMI (Figure 3.13). The higher heritability of BMI-change in women replicates in the Estonian Biobank (h^2_{G} = 2.15% (0.56) in women vs 1.80% (0.98) in men).

In line with previously reported evidence for distinct genetic drivers of baseline adiposity and change in weight over time[309,311], I found that the genome-wide genetic correlations r_{G} between LME intercepts and LME slopes were modest, between 40.0% to 60.0% (Figure 3.13). However, the point estimates of r_{G} between baseline adiposity and adiposity-change clusters varied widely for different pairs of traits, from -27.2% (SE = 6.36%) for the relationship between baseline weight and posterior probability of belonging to any weight cluster but the loss cluster, to 73.0% (SE = 10.8%) for the relationship between male-specific baseline weight and posterior

Table 3.8: Replication of adiposity-change lead SNPs with linear slope change in body mass index (BMI) in: (1) individuals of European ancestry in the UK Biobank who were held out of discovery analyses, or (2) the Estonian biobank cohort. TSS = transcription start site, SE = standard error

SNP	chr:pos (hg19)	Nearest TSS	DISCOVERY Beta (SE)	REPLICATION			
				UK Biobank, N = 39,388		EstBB, N = 125,209	
				Beta (SE)	P-value	Beta (SE)	P-value
rs747792	2:25473547	POMC	-0.0377 (0.00684)	0.0119 (0.00777)	0.125	-0.000722 (0.00475)	0.879
rs811041	6:26226032	H3C6	-0.0205 (0.00365)	-0.018 (0.00802)	0.0249	-0.00313 (0.00411)	0.446
rs9373305	6:141683547	GJE1	-0.0617 (0.0113)	0.0245 (0.00734)	0.000837	0.00204 (0.00407)	0.617
rs35519679	10:114818754	TCF7L2	-0.0327 (0.0059)	-0.00301 (0.00823)	0.715	-0.00593 (0.00516)	0.25
rs186269301	12:123917276	RILPL2	0.116 (0.0194)	0.023 (0.0213)	0.282	0.00613 (0.0176)	0.727
rs714827	14:103113552	RCOR1	0.111 (0.0193)	-0.0269 (0.0131)	0.0393	0.00652 (0.00749)	0.384
rs7206122	16:53822440	RPGRIP1L/FTO	-0.0489 (0.00841)	-0.0282 (0.00737)	0.000127	-0.0113 (0.00399)	0.00463
rs11661876	18:57889652	MC4R	-0.057 (0.00939)	-0.0127 (0.00816)	0.12	-0.0105 (0.00491)	0.0322
rs429358	19:45411941	APOE	-0.169 (0.0146)	-0.0653 (0.00996)	5.4E-11	-0.0256 (0.00593)	1.61E-05

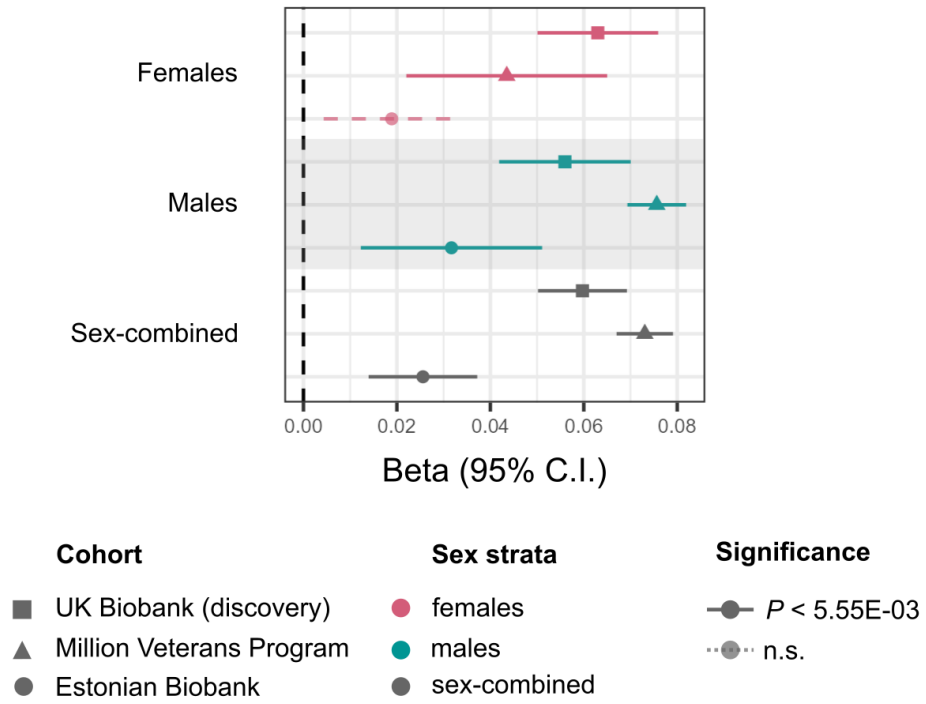


Figure 3.12: Effect sizes of rs429358 on BMI-change slope in discovery (UK Biobank) and replication (Million Veteran Program and Estonian Biobank) datasets. Effect size (beta) and 95% CI for associations of rs429358 with BMI linear slope change over time estimated from linear mixed-effects models in all analysis strata. Estimates of slope change are adjusted for baseline BMI, and P-values for significance are controlled at 5% across number of tests performed via the Bonferroni method. n.s. = non-significant

probability of belonging to the high weight gain cluster. Finally, as adiposity-change and baseline adiposity are partially genetically correlated, adjusting for the latter in GWASs for adiposity-change may induce collider bias due to adjustments along a causal pathway (wherein high baseline weight is a result of increased rate of weight gain, or vice-versa)[356]. A comparison of the results of two sets of GWASs for the LME slope trait – one that included the LME intercept as a covariate and one that did not – shows that 98.2% of associations reach GWS in both sets; and while the effect sizes of GWS variants identified in the LME-intercept-adjusted GWASs are significantly smaller than those in the unadjusted GWASs (Bland-Altman ratio[357] $P < 0.00383$), the average effect estimate difference is less than 9.69×10^{-4} units (Figure 3.14).

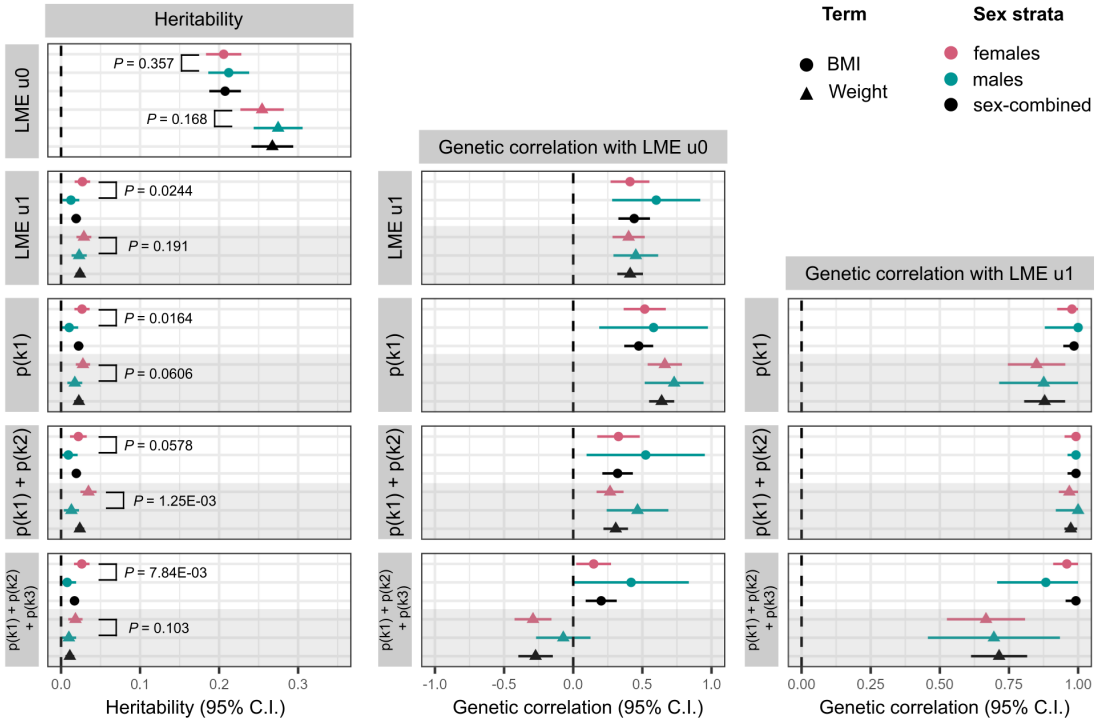


Figure 3.13: SNP-based heritability of, and genetic correlation between, baseline adiposity and adiposity-change phenotypes. Left column: heritability (h^2_G) estimates and 95% confidence intervals (CI), calculated using the LDSC software[62] on a subset of 1 million HapMap3 SNPs[46] for the following traits: baseline body mass index (BMI) and weight, estimated from intercepts of linear mixed-effects (LME) models of adiposity traits over time (u0); and adiposity-change estimated using: LME models for linear slope change over time (u1), or posterior probability of membership in a high gain BMI or weight cluster p(k1), high or moderate gain clusters (p(k1) + p(k2)) or any cluster but loss (p(k1) + p(k2) + p(k3)), all of which are adjusted for baseline trait value. P-values display the level of significance of heterogeneity between the female- and male-specific estimates in each panel. Middle column: Genetic correlation, r_G, and 95% CI between baseline BMI or weight (u0) and each adiposity-change phenotype. Right column: Genetic correlation, r_G, and 95% CI between BMI-slope or weight-slope (u1) and each adiposity-change phenotype. In all panels, circles represent BMI, triangles represent weight; points are coloured by analysis stratum (pink: female-specific, green: male-specific, grey: sex-combined).

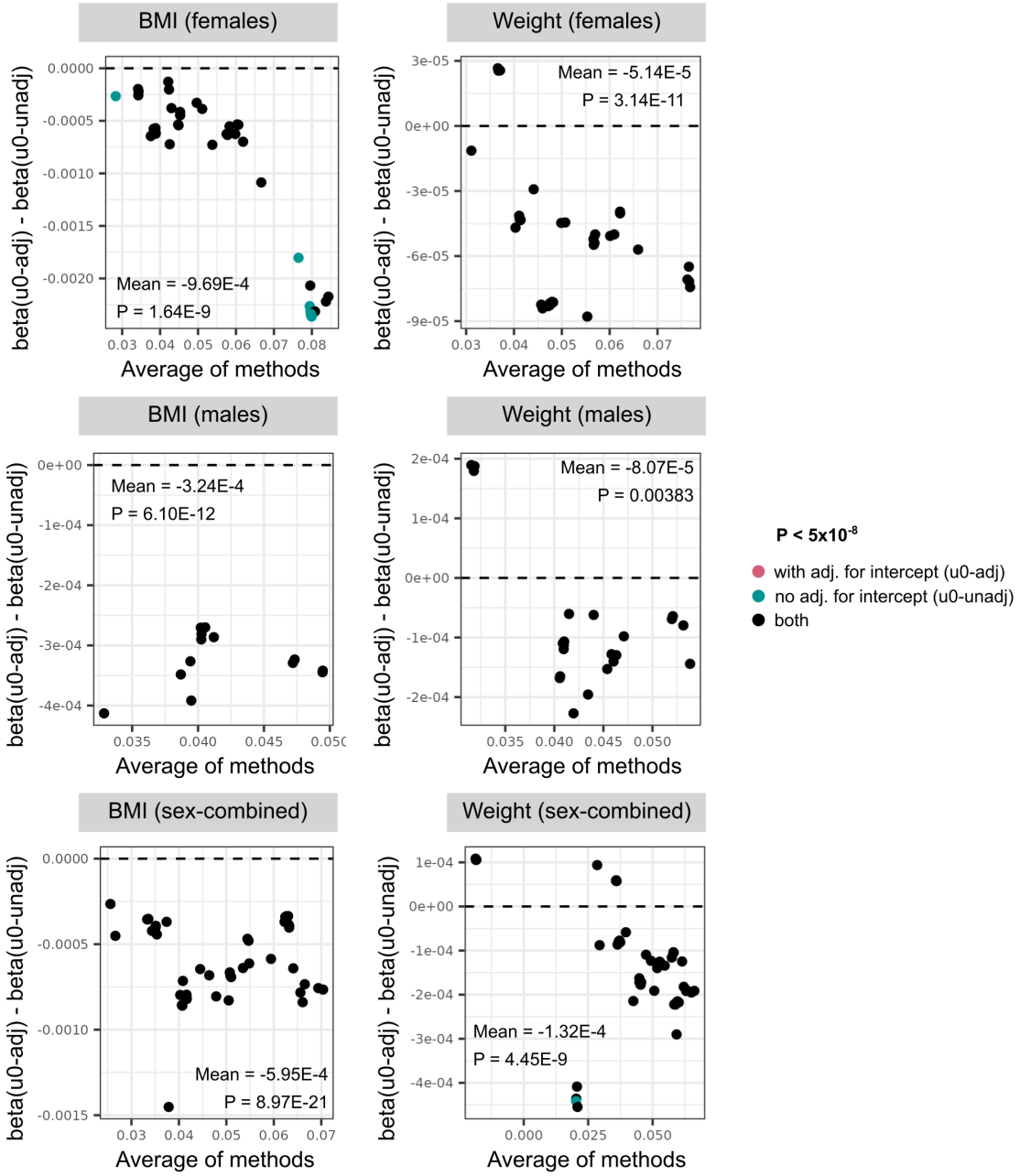


Figure 3.14: Bland-Altman plot[357] to compare effect sizes of genome-wide significant (GWS) SNPs in main linear mixed-effects (LME) model slope (u1) genome-wide association studies (GWASs) and sensitivity analyses unadjusted for LME intercept (u0) values. Each panel represents a different analysis stratum. SNPs that reach genome-wide significance ($P < 5 \times 10^{-8}$) in GWASs for either the u0-unadjusted or u0-adjusted trait are plotted, coloured by whether SNP is significant in both GWASs or only one. All variants that reach GWS in the u0-adj GWAS are also significant in the u0-unadj GWAS. For each GWS SNP, the difference between effect sizes estimated in the u0-adj regression and u0-unadj regression is plotted against the mean estimate from both regressions. The mean difference and one-sample t-test P-value (with null hypothesis of mean difference = 0) are displayed in each panel. All GWASs were adjusted for the first 21 genetic PCs, genotyping batch, and genotyping array.

3.3.3 *APOE* variant rs429358 is associated with multiple adiposity-change phenotypes

Independent of baseline obesity, rs429358 is associated with weight loss over time

A common missense variant in *APOE*, rs429358 (MAF = 15.5%), is associated with negative rates of change in weight and BMI over time, and lower posterior probabilities of belonging to a weight or BMI gain cluster, in both men and women (Table 3.6). Each additional copy of the minor C allele of rs429358 is associated with 0.0591 (SE = 0.00475, $P = 1.49 \times 10^{-35}$) SD decrease in expected BMI slope over time, and 0.0614 (SE = 0.00454, $P = 1.40 \times 10^{-41}$) SD decrease in expected weight slope over time (Figure 3.15A). Independent of baseline obesity, carriers of the minor C allele of rs429358 are at lower odds of membership in the high gain BMI cluster (OR (95% CI) = 0.913 (0.897 - 0.929), $P = 9.48 \times 10^{-25}$), which lowers the membership posterior probability from 40% for a homozygous dominant participant to 36.5% for a heterozygous individual. Although the minor allele of rs429358 is also associated with lower baseline BMI (β (SE) = 0.014 (0.00460) SD lower BMI LME model intercept) and weight (β (SE) = 0.00963 (0.00433)), these associations do not reach genome-wide significance ($P > 2.33 \times 10^{-3}$).

Further, I found that carriers of each additional copy of the minor C allele of rs429358 are at 0.951 (95% CI = 0.934 - 0.978, $P = 4.89 \times 10^{-12}$) lower odds of being in a higher ordinal self-reported weight gain category, independent of their BMI (Table 3.9). This translates to a 2.6 percentage point difference in the proportion of participants reporting weight gain between individuals with TT (28.0%) and CC (25.4%) genotypes at this SNP (Figure 3.15C). These analyses were based on 252,948 participants of European ancestry in the UKBB, who reported weight change in the last year as “gained weight”, “weigh about the same”, or “lost weight” and were held out of the discovery GWASs. The adiposity-change lead variants rs7206122 (in the *FTO* locus) (OR = 0.975 (0.965 - -.986), $P = 2.04 \times 10^{-6}$) and rs11661876 (in the *MC4R* locus) (OR = 0.974 (0.962 - 0.985), $P = 4.58 \times 10^{-6}$) are also associated with self-reported weight loss in the past year (Table 3.9).

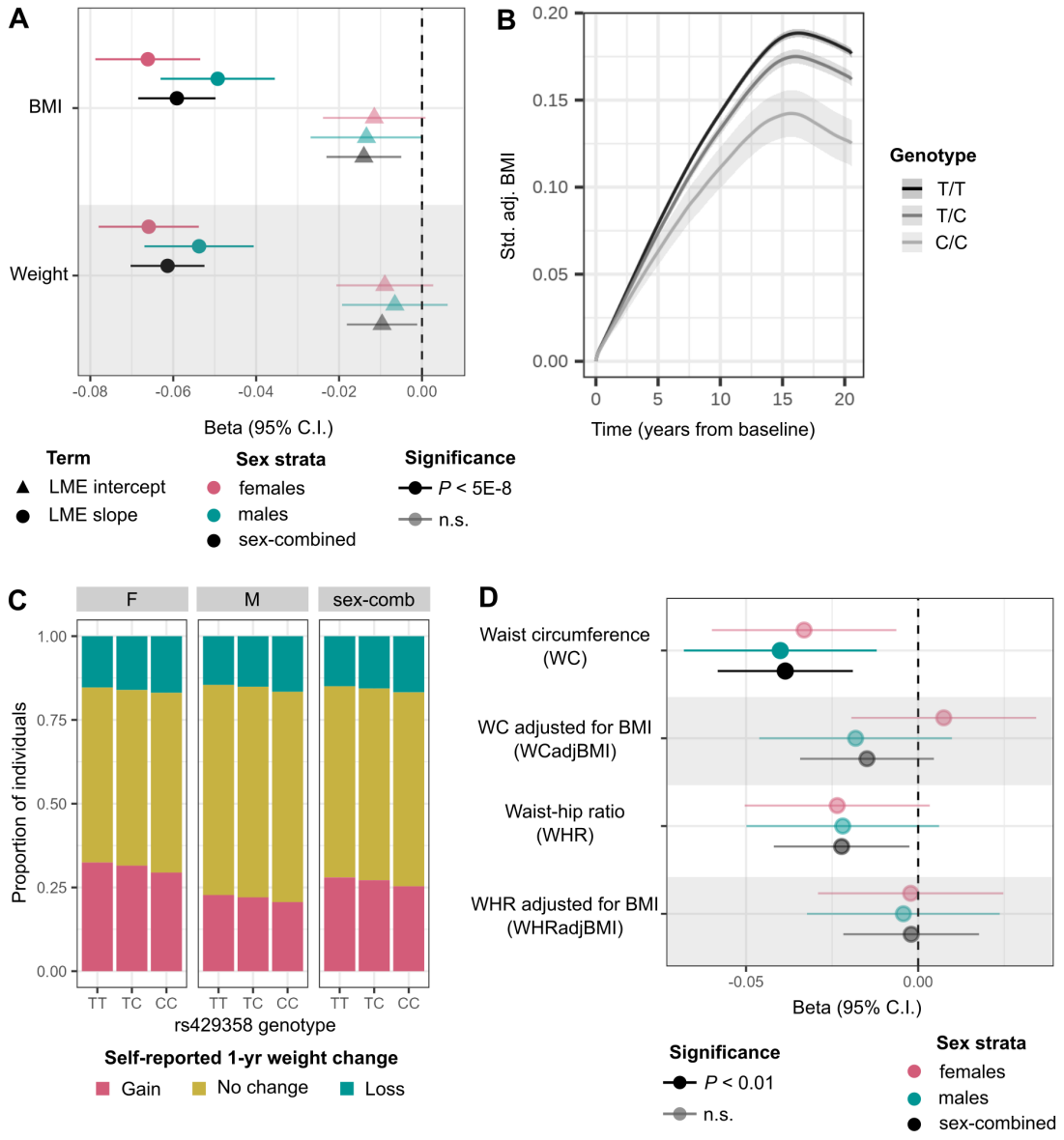


Figure 3.15: Association of *APOE* SNP rs429358 with various adiposity-change phenotypes in the UK Biobank (UKBB). (A) Effect size (beta) and 95% CI for associations of rs429358 with BMI and weight intercepts or linear slope change over time estimated from linear mixed-effects (LME) models in all analysis strata. (B) Modelled trajectories of standardised (std.) covariate-adjusted (adj.) BMI in carriers of the different rs429358 genotypes. (C) Proportion of individuals in the UKBB held-out set who self-report weight gain, weight loss, or no change in weight over the past year for carriers of each rs429358 genotype. (D) Effect size and 95% CI for associations of rs429358 with slopes over time of waist circumference (WC) and waist-to-hip ratio (WHR), adjusted for BMI (-adjBMI), estimated from LME models in the UKBB held-out set. Across all panels, estimates of trait change are adjusted for baseline trait values. n.s. = non-significant.

Table 3.9: Association of adiposity-change lead SNPs with self-reported weight change in the past 1 year, independent of baseline body mass index (BMI). TSS = transcription start site, SE = standard error, OR = odds ratio, CI = confidence interval.

SNP	chr:pos (hg19)	Nearest TSS	Sex strata	Sample size	OR (95% CI)	P-value
rs747792	2:25473547	POMC	F	136519	1.01 (0.997 - 1.03)	0.943
			M	116429	1.01 (0.992 - 1.03)	0.851
			sex_comb	273782	1.01 (1 - 1.02)	0.974
rs811041	6:26226032	H3C6	F	136519	0.997 (0.981 - 1.01)	0.357
			M	116429	1.01 (0.991 - 1.03)	0.831
			sex_comb	273782	0.998 (0.987 - 1.01)	0.392
rs9373305	6:141683547	GJE1	F	136519	1.01 (0.992 - 1.02)	0.829
			M	116429	0.998 (0.981 - 1.01)	0.399
			sex_comb	273782	1 (0.993 - 1.01)	0.774
rs35519679	10:114818754	TCF7L2	F	136519	1.01 (0.996 - 1.03)	0.934
			M	116429	1.01 (0.993 - 1.03)	0.886
			sex_comb	273782	1.01 (1 - 1.02)	0.971
rs186269301	12:123917276	RILPL2	F	136519	0.995 (0.953 - 1.04)	0.415
			M	116429	1.03 (1.02 - 1.03)	1
			sex_comb	273782	1.01 (0.976 - 1.04)	0.66
rs714827	14:103113552	RCOR1	F	136519	0.997 (0.97 - 1.03)	0.422
			M	116429	1.02 (0.989 - 1.05)	0.896
			sex_comb	273782	1 (0.982 - 1.02)	0.54
rs7206122	16:53822440	RPGRIP1L/FTO	F	136519	0.983 (0.969 - 0.998)	0.0121
			M	116429	0.971 (0.955 - 0.988)	0.000334
			sex_comb	273782	0.975 (0.965 - 0.986)	2.04E-06
rs11661876	18:57889652	MC4R	F	136519	0.978 (0.962 - 0.994)	0.00326
			M	116429	0.97 (0.952 - 0.988)	0.000599
			sex_comb	273782	0.974 (0.962 - 0.985)	4.58E-06
rs429358	19:45411941	APOE	F	136519	0.952 (0.933 - 0.971)	7.62E-07
			M	116429	0.956 (0.934 - 0.978)	4.42E-05
			sex_comb	273782	0.951 (0.938 - 0.965)	4.89E-12

Finally, I attempted to unpick the effects of rs429358 on changes in overall obesity and abdominal obesity by assessing the association of this variant with linear slope change in WC (β (SE) = 0.0386 (0.0100) SD decrease in expected WC slope over time, $P = 1.16 \times 10^{-4}$), WHR (β (SE) = 0.0222 (0.0100), $P = 0.0268$), and their BMI-adjusted counterparts, WCadjBMI (β (SE) = 0.0150 (0.00991), $P = 0.133$) and WHRadjBMI (β (SE) = 0.00204 (0.0100), $P = 0.839$) in up to 39,535 participants of European ancestry held out of the discovery analyses (Figure 3.15D). The significant association of rs429358 with lower WC slope is attenuated upon adjustment for BMI, indicating that any changes in abdominal adiposity associated with this variant may be attributed to changes in overall obesity.

Phenome- and longitudinal phenome-wide consequences of rs429358

I assessed pleiotropy in the phenotype-wide effects of the rs429358 variant in *APOE* across 290 disease end-points defined in primary care, secondary care, and self-reported records in approximately 450,000 individuals of European ancestry in the

UKBB, allowing for sample overlap with the discovery GWASs. Phenome-wide significance was established at $P < 1.72 \times 10^{-4}$ (FWER controlled at 5% across 290 phenotypes using the Bonferroni method). While the strongest association of rs429358 is with increased risk of dementia (OR (95% CI) = 2.78 (2.62 - 2.95), $P = 2.29 \times 10^{-248}$), the variant is also associated with altered risk of cardio-vascular, metabolic, endocrine, digestive, and musculo-skeletal diseases (Figure 3.16). Male carriers of the minor C allele of rs429358 are at significantly higher risk of circulatory disorders such as coronary heart disease (OR = 1.09 (1.07 - 1.12), $P = 2.75 \times 10^{-12}$), angina (OR = 1.09 (1.06 - 1.12), $P = 1.65 \times 10^{-8}$), and myocardial infarction (OR = 1.06 - 1.14, $P = 1.95 \times 10^{-7}$), but both men and women who carry the minor allele are protected against development of type 2 diabetes (OR = 0.913 (0.892 - 0.934), $P = 6.94 \times 10^{-15}$) (Figure 3.16).

Given the strong association of rs429358 with risk of developing dementia, which is itself often accompanied by weight loss[340], I tested the association of this SNP with adiposity-change after excluding 272 participants with hospital or self-reported records of Alzheimer's disease or dementia (Table B.2), but find no significant difference in the effect estimates between these sets of analyses (all heterogeneity P-values $P_{\text{het}} > 0.05$, Figure 3.17). While there may be similar concerns that the association of rs429358 with weight loss is a consequence of cancer or its treatment, there is no association between this variant and risk of developing any of the 34 cancers tested in phenotype-wide analyses ($P > 1.72 \times 10^{-4}$).

Finally, I tested the association of rs429358 with linear rates of change in the levels of 45 quantitative traits, including cardio-metabolic, endocrine, immune, and cancer-related biomarkers. Reflecting the pleiotropy seen in the binary disease associations, each additional copy of the minor C allele of rs429358 is associated with increase in levels of total cholesterol (β (SE) = 0.0306 (0.00431) per SD increase in expected slope of total cholesterol over time, $P = 1.30 \times 10^{-12}$), HDL cholesterol ($\beta = 0.0201$ (0.00477), $P = 2.59 \times 10^{-5}$), and C-reactive protein (CRP) ($\beta = 0.0259$ (0.00519), $P = 6.08 \times 10^{-7}$), but decrease in levels of triglycerides (β = -0.024 (0.00511), $P = 1.99 \times 10^{-6}$), potassium (β = -0.0233 (0.00491), $P = 1.97 \times 10^{-6}$),

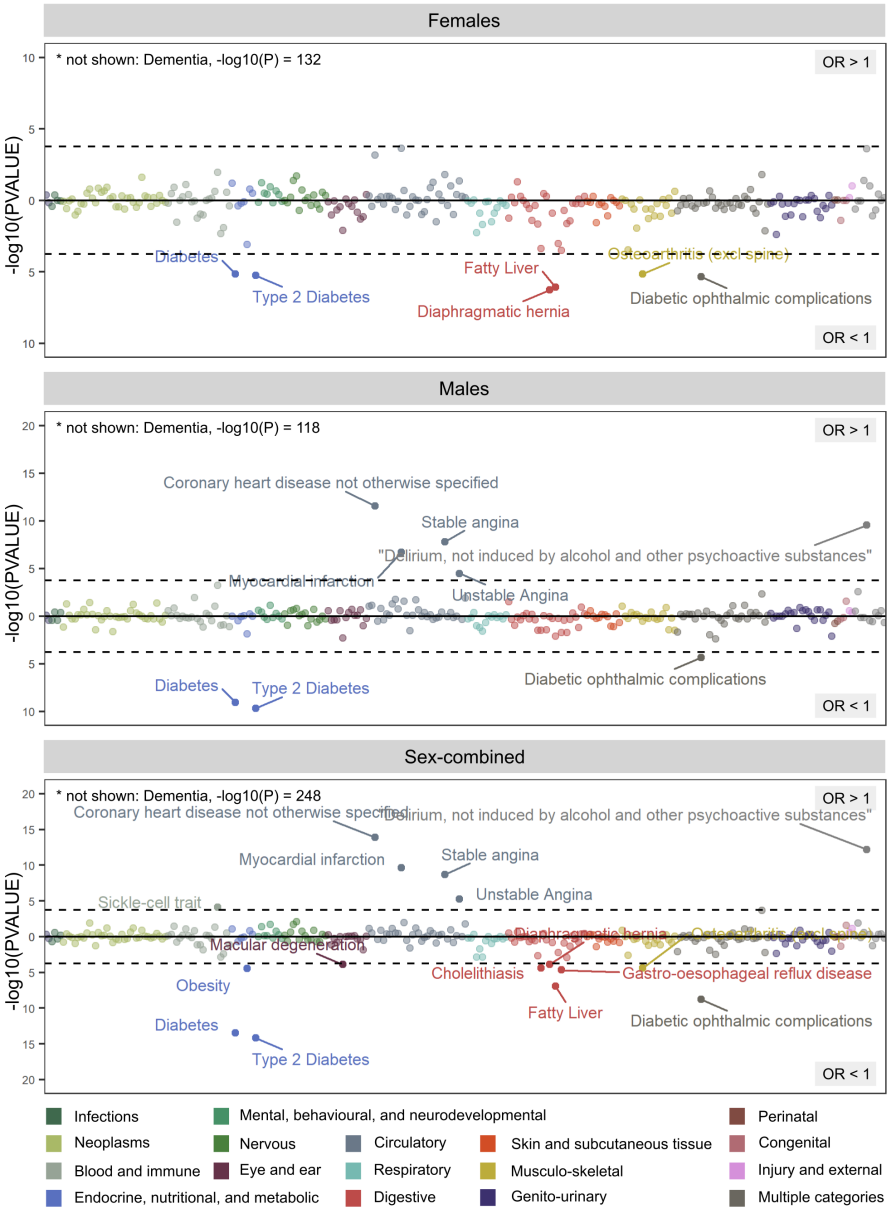


Figure 3.16: Phenome-wide associations of rs429358 with diseases in UK Biobank. Logistic regression models were fitted to assess the association of rs429358 with 290 disease phenotypes, curated from primary care, secondary care, and self-report records in individuals of genetically ascertained European ancestry in the UK Biobank (N approx. 450,000, allowing for sample overlap with the adiposity-change discovery analyses). All associations in the sex-combined analysis were adjusted for sex. The Miami plots display $-\log_{10}(P)$ for significance of the association, with deleterious effects of the minor C allele (odds ratio OR > 1) shown above the horizontal 0 line, and protective effects (OR < 1) shown below the horizontal 0 line. Associations are coloured by International Classification of Diseases (ICD) chapter, and those significant at $P < 0.05/290$, (FWER controlled at 5% across 290 phenotypes using the Bonferroni method) are labelled.

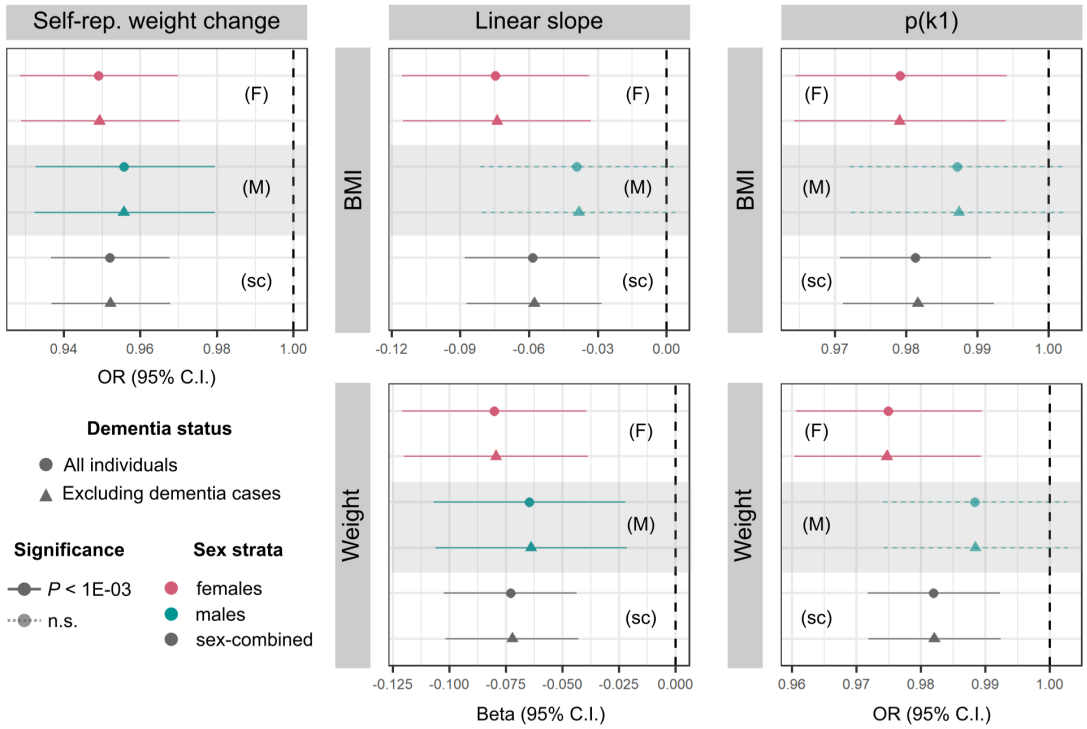


Figure 3.17: Comparison of rs429358 effects on adiposity-change phenotypes in held-out participants in UK Biobank(UKBB) and sensitivity analysis in held-out participants in UKBB without dementia diagnoses. Left column: Odds ratios (OR) and 95% confidence intervals (CIs) from proportional odds logistic regression models for association of minor C allele of rs429358 with self-reported weight change in the past year, coded as an ordinal categorical variable with levels "Loss", "No change", and "Gain". Middle column: Effect size (beta) and 95% CI for associations of rs429358 genotype with body mass index (BMI) and weight linear slope change over time, estimated using linear mixed-effects (LME) models. Right column: OR and 95% CI for association of rs429358 genotype with posterior probability of membership in the BMI and weight high gain (k1) clusters. All estimates are adjusted for BMI (self-reported weight change) or baseline adiposity-trait value. Circles represent associations in all participants in the held-out set of UK Biobank, while estimates represented by triangles are estimated in the subset of participants with dementia or Alzheimer's disease diagnoses. Points are coloured by sex stratum. Solid lines are significant at $P < 1 \times 10^{-3}$ (FWER controlled at 5% across 15 tests using the Bonferroni method), dashed lines are non-significant (n.s.).

lymphocyte count (β = -0.0186 (0.00468), $P = 6.80 \times 10^{-5}$), and eosinophill count (β = -0.0161 (0.00463), $P = 5.13 \times 10^{-4}$). These analyses were based on up to 153,844 individuals of European ancestry in the UK Biobank, allowing for sample overlap with the adiposity-change discovery analyses. Linear slope for change in biomarker level was calculated with LME models and adjusted for baseline trait value; statistical significance was established at $P < 1.11 \times 10^{-3}$ after FWER control at 5% across 45 biomarkers using the Bonferroni method.

3.3.4 Genetic associations with adiposity-change are robust to a range of modelling choices

Reported GWS associations are unlikely to be driven solely by survival bias

The rs429358 variant in *APOE* is known to be associated with lifespan and is correlated with participant age in the UKBB[338]. To ensure that our results were not induced by the adjustments for age and its related covariates, i.e. follow-up metrics and year of birth, we constructed two sets of LME models for adiposity-change. Both included rs429358 dosage and dosage × time interactions as fixed effects, but with separate sets of covariates: (i) only data provider, 21 genetic PCs, genotyping batch, genotyping array, and sex, and (ii) all of the above and baseline age, (baseline age)^2, number of follow-up measurements, length of follow-up, and year of birth. The latter set of results are those presented in the main analyses; however, we did not observe any significant differences (heterogeneity P-value $P_{\text{het}} < 0.05$) between the estimates from each of these sets of models (Figure 3.19).

To further mitigate against survival bias, we adjust for length of follow-up and number of observations in GWASs, which are positively correlated with an individual's BMI in the UKBB-linked GP EHR - a phenomenon termed informed presence or ascertainment bias[149] (Table 3.3). There is also weak positive correlation between intra-individual variance in BMI and follow-up metrics, i.e. length of follow-up ($r^2 = 0.111$) and number of observations in the longitudinal record ($r^2 = 0.0581$), but there is considerable heteroskedasticity in these relationships (Figure 3.20). The adjustment for follow-up metrics in GWASs for adiposity-change may induce

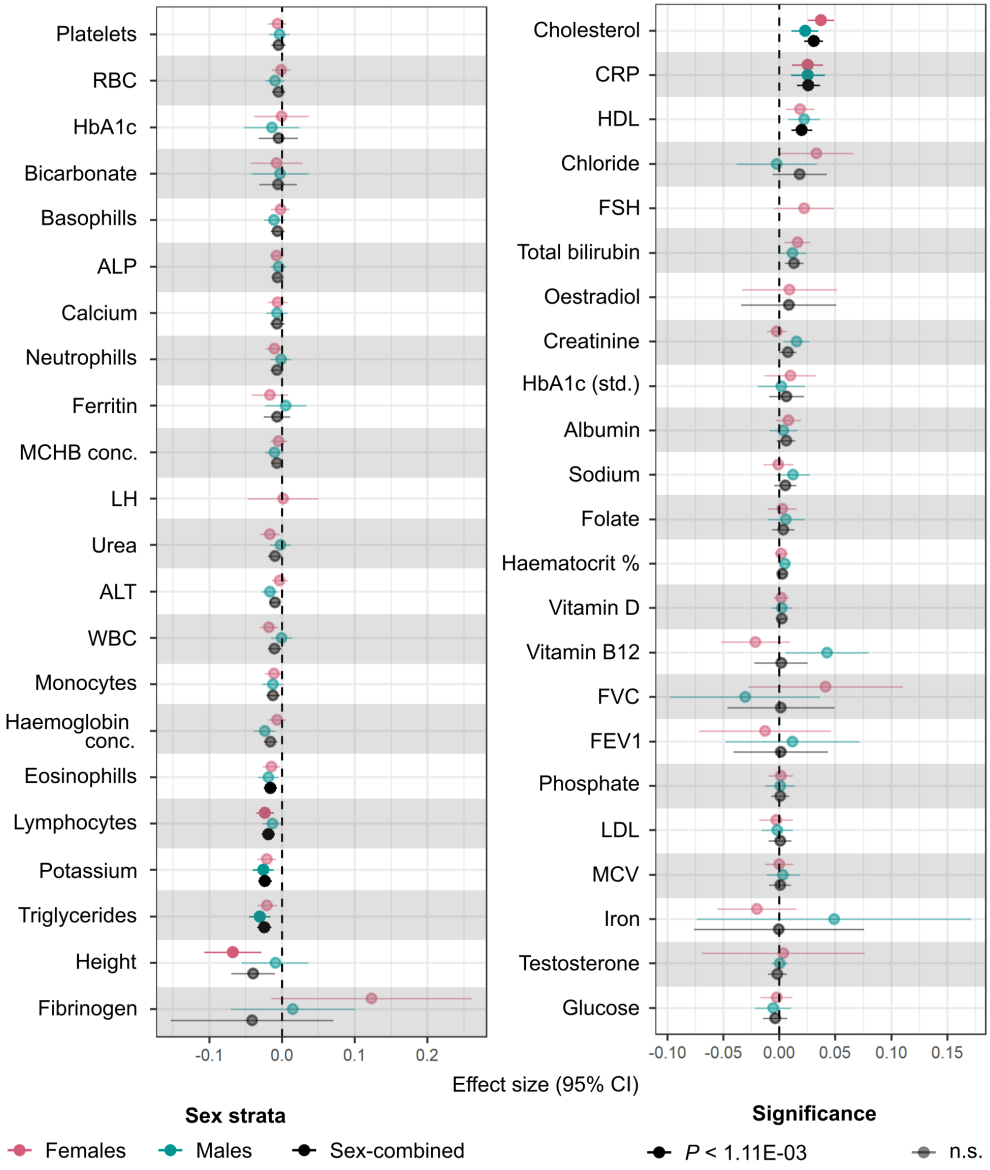


Figure 3.18: Longitudinal phenome-wide effects of rs429358 on changes in 45 biomarker levels in UK Biobank. Effect size and 95% CI for associations of rs429358 with linear slope change in quantitative biomarkers over time, estimated from linear mixed-effects models. Estimates of trait change are adjusted for baseline trait values, and P-values for significance are FWER controlled at 5% across 45 tests via the Bonferroni method. n.s. = non-significant. ALP = alkaline phosphatase, ALT = alanine aminotransferase, CRP = C-reactive protein, FEV1 = forced expiratory volume, FSH = follicle stimulating hormone, FVC = forced vital capacity, HbA1c (std.) = haemoglobin A1C (standardised), HDL = high density lipoprotein, LDL = low density lipoprotein, LH = luteinising hormone, MCHB conc. = mean corpuscular haemoglobin concentration, MCV = mean corpuscular volume, RBC = red blood cell count, WBC = white blood cell count.

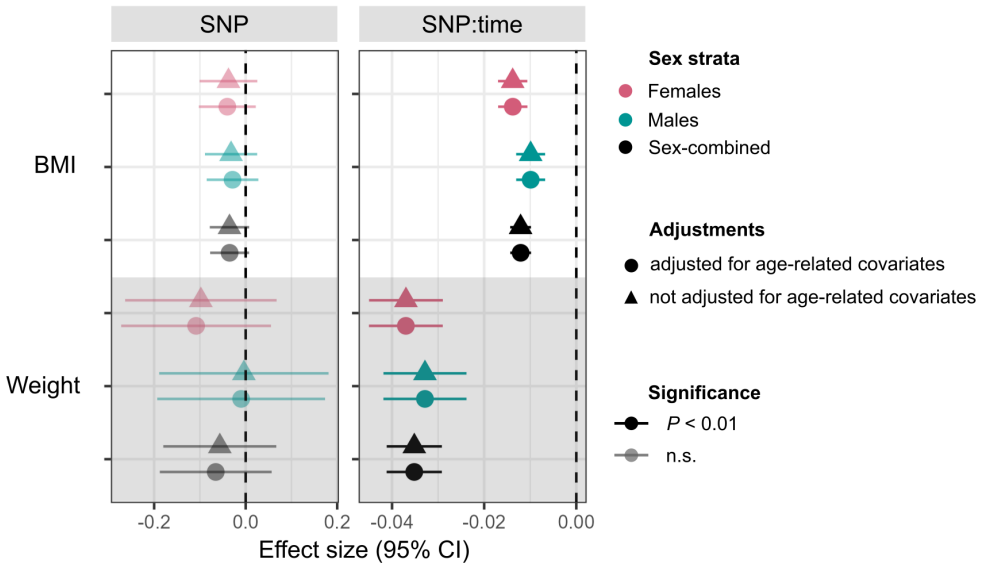


Figure 3.19: Comparison of rs429358 associations with baseline adiposity and adiposity-change in main analyses and sensitivity analyses unadjusted for age. Displayed are the effect size (β) and 95% confidence intervals for the effect of each additional copy of the minor C allele of rs429358 on body mass index (BMI) and weight. The SNP and SNP×time interaction were included as fixed effects in a linear mixed-effects (LME) model for adiposity with fixed and random effects of time, with additional adjustments for data provider, 21 genetic principal components, genotyping batch, genotyping array, and sex (in sex-combined analyses). Models were either adjusted for age-related covariates: baseline age, (baseline age)^2, year of birth, number of follow-up measurements, and length of follow-up (circles), or not adjusted for these additional covariates (triangles).

collider bias[356], so I additionally report results from GWASs unadjusted for length of follow-up and number of observations. The difference between effect size estimates for GWS variants reported in at least one the two sets of analyses – (i) including length of follow-up and number of follow-up observations as covariates for GWAS, and (ii) without these covariates – is significant in 5 of 6 strata ($P < 1.07 \times 10^{-8}$), but small (the absolute value of the average difference between estimates is less than 5.47×10^{-4}) (Figure 3.21).

More generally, as ageing is itself associated with reduction in body weight[358,359], survival bias may confound our reported genetic associations; we therefore tested whether published longevity-associated SNPs were also associated with adiposity-change. We find that 135 of 138 independent variants associated with lifespan in the GWAS Catalog[50] are not associated with any of the adiposity-change traits we

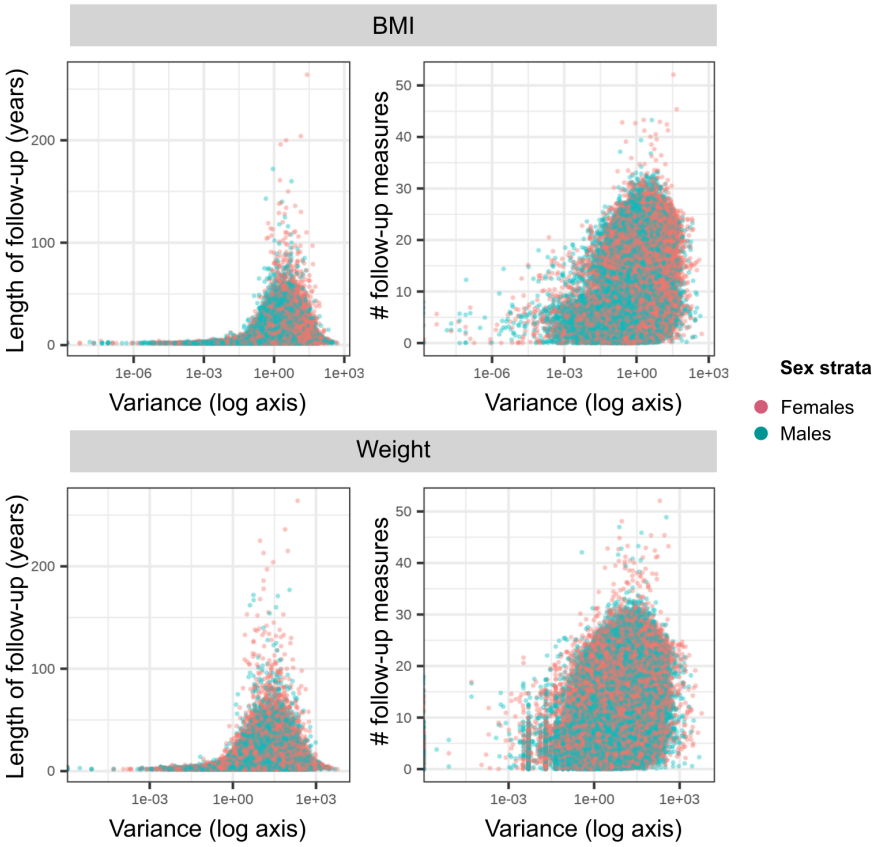


Figure 3.20: Correlation between follow-up metrics and intra-individual variance in body mass index (BMI) and weight. Intra-individual variance in BMI and weight across multiple measures is plotted on the log-x-axis vs length of follow-up in years (left) and number of follow-up measures (right) on the y-axis. Each point represents an individual (females = pink, males = blue).

analyse in this study ($P > 3.62 \times 10^{-4}$, FWER controlled at 5% across 138 tests using the Bonferroni method). The three variants associated with both lifespan and change in adiposity are: rs4420638 (in LD with rs429358, $r^2 = 0.690$), rs7412, another missense variant in *APOE*, and rs10852521 in the known obesity-associated *FTO* locus (Table B.5), suggesting that these associations may be pleiotropic or part of shared pathways between longevity and adiposity.

Different definitions of adiposity-change are strongly genetically and phenotypically correlated

The phenotypes captured by the LME models for linear slope change (u1) and the posterior probabilities of belonging to the different adiposity-change clusters

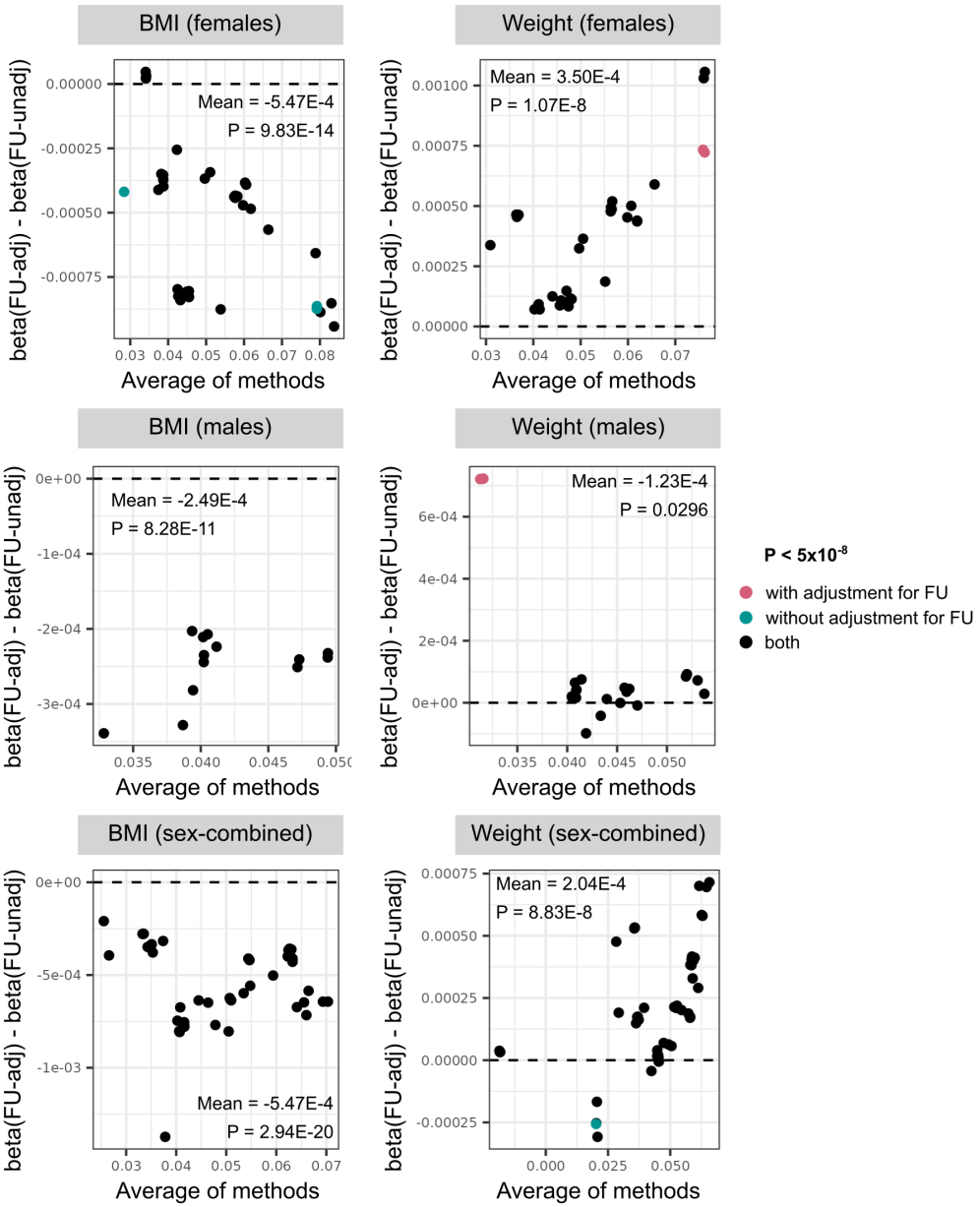


Figure 3.21: Bland-Altman plot[357] to compare effect sizes of genome-wide significant (GWS) SNPs in main linear mixed-effects (LME) model slope (u1) genome-wide association studies (GWASs) sensitivity analyses that are unadjusted for length of follow-up (FU) and number of FU measurements. Each panel represents a different analysis stratum. SNPs that reach genome-wide significance ($P < 5 \times 10^{-8}$) in GWASs for either the FU-unadjusted or FU-adjusted trait are plotted, coloured by whether SNP is significant in both GWASs or only one. For each GWS SNP, the difference between effect sizes estimated in the FU-adj regression and FU-unadj regression is plotted against the mean estimate from both regressions. The mean difference and one-sample t-test P-value (with null hypothesis of mean difference = 0) are displayed in each panel. All GWASs were adjusted for the first 21 genetic PCs, genotyping batch, and genotyping array.

are strongly positively correlated (point estimates of r^2 between BMI u1 and BMI-change clusters = 0.758 to 0.829, and between weight u1 and weight-change clusters = 0.637 to 0.760) (Figure 3.22A). The corresponding genetic correlations are even stronger, with point estimates of r^2 between 0.714 and 0.992 across the various strata (Figure 3.22B). Moreover, we report results for two adiposity traits –BMI and weight – throughout this study, which we expect to track closely in adults as height does not change significantly over time. Indeed, we observe strong correlations between the slopes for weight and BMI change ($r^2 = 0.877$) and between the posterior probabilities of membership in the various BMI and weight clusters (point estimates of r^2 between 0.713 and 0.798) (Figure 3.22). Once again, the corresponding genetic correlations are higher than the phenotypic correlations (r_G between BMI and weight slopes = 0.996, r_G between BMI and weight posterior probability of cluster membership between 0.892 and 1), indicating that the genetic architecture highlighted here is robust to the metric of adiposity used to define trajectories.

3.3.5 Longitudinal data enable the identification of novel genetic signals for baseline adiposity

Discovery of novel genetic variants for obesity in UK Biobank

We hypothesised that, by sharing information across multiple observations over time, the individual-level random intercept terms from LME models would provide higher statistical power for GWAS than a single cross-sectional measurement. Indeed, despite our GWAS having 4 times fewer samples than the largest published European-ancestry GWAS for BMI to date[216], we identify 20 novel autosomal loci and three novel loci on the X chromosome, and refine 67 previously described autosomal signals for obesity among the 471 uniquely fine-mapped lead variants that reach GWS across all strata (Figure 3.23, Table 3.10, and Table B.4; see Section 2.5.2 in Chapter 2 for detailed description of lead variant classification as published, refined, or novel). The 23 novel SNPs for obesity are not in LD with ($r^2 < 0.1$), and conditionally independent of (conditional P-value $P_{cond} < 0.05$),

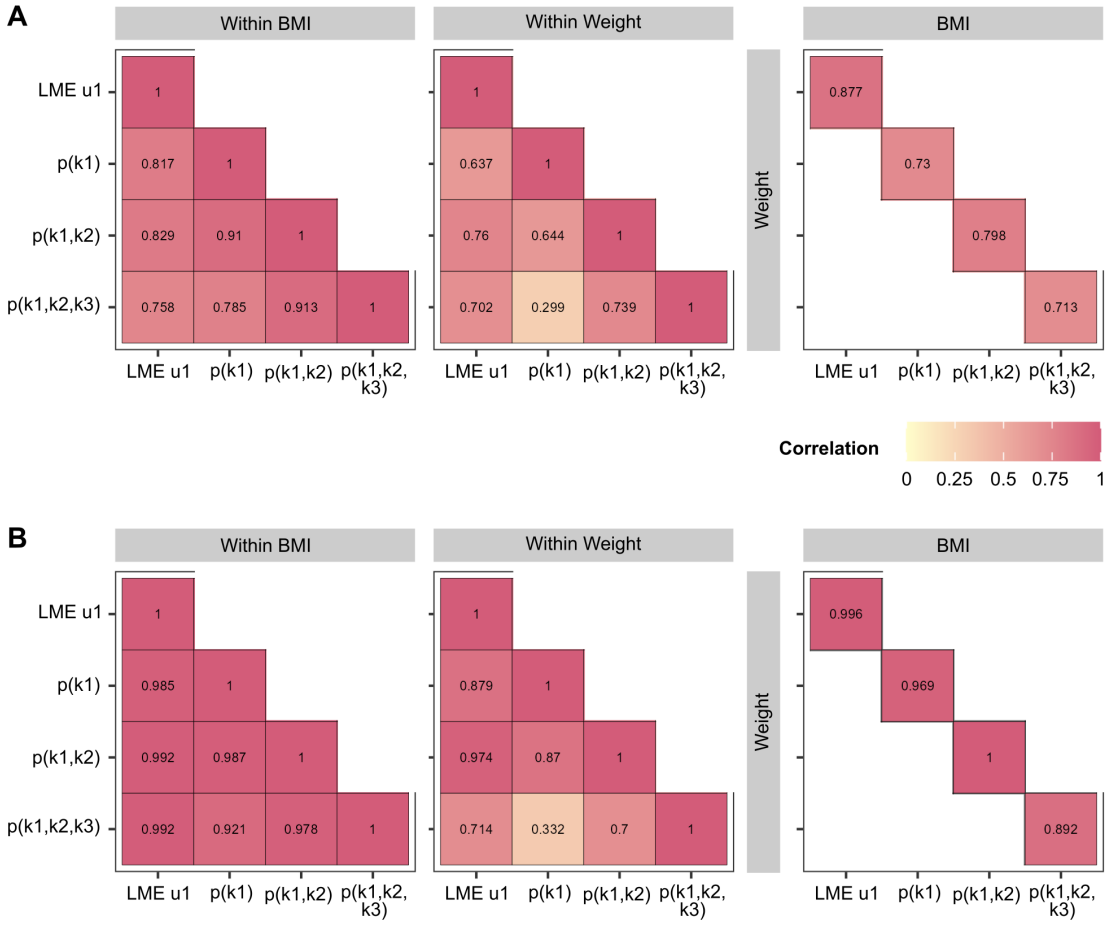


Figure 3.22: (A) Phenotypic and (B) genetic correlations between adiposity-change phenotypes defined through linear mixed-effects (LME) model slopes (u1) for body mass index (BMI) and weight, or posterior probability of belonging to a BMI or weight change cluster. Genetic correlations were calculated from GWAS summary statistics for each pair of traits, using the LDSC software[62] on a subset of 1 million HapMap3 SNPs[46]. p(k1) = posterior probability of belonging to high gain cluster, p(k1,k2) = posterior probability of belonging to high or moderate gain clusters, p(k1,k2,k3) = posterior probability of belonging to high or moderate gain or steady clusters.

any variants within a 1 Mb window of the lead variant that are associated with 44 obesity-related traits in the GWAS Catalog[50].

Novel SNPs for obesity replicate in UKBB held-out data

We were adequately powered (80% power) to replicate the sex-specific effects of all 20 novel autosomal variants for baseline obesity, at $P < 2 \times 10^{-3}$ (FWER controlled at 5% across 20 variants using the Bonferroni method), in: (1) cross-sectional UKBB

Table 3.10: Novel variants associated with obesity identified in genome-wide association studies (GWASs) for body mass index (BMI) and weight linear mixed-effects (LME) model intercepts in sex-specific and sex-combined strata; alongside the nearest published obesity-associated SNP in the GWAS Catalog. MAF = minor allele frequency, TSS = transcription start site.

Novel SNP from LME intercept GWAS				Nearest published obesity-associated SNP from GWAS Catalog				
SNP	chr:pos (hg19)	MAF	Nearest TSS	Distance to novel SNP (kb)	SNP	chr:pos (hg19)	MAF	Nearest TSS
BMI (males)								
rs79586444	8:33688233	0.138	DUSP26	57.3	rs7012923	8:33630937	0.318	DUSP26
rs294382	11:30607515	0.185	MPPED2	115	rs11031105	11:30492070	0.144	MPPED2
rs17794645	16:13813090	0.0601	ERCC4	163	rs2113334	16:13976356	0.49	ERCC4
BMI (sex-combined)								
rs11161573	1:85724581	0.246	C1orf52	44.2	rs11161548	1:85680408	0.398	SYDE2
rs7569339	2:133469717	0.149	NCKAP5	53.9	rs13033310	2:133523605	0.259	NCKAP5
rs2279620	17:7792465	0.153	CHD3	37.5	rs12939056	17:7754993	0.187	KDM6B
rs112469935	23:76550037	0.017	FGF16	794	rs151280158	23:77344410	0.0485	PGK1
rs55669243	23:79680770	0.0109	TENT5D	1040	rs11314467	23:78643198	0.012	ITM2A
rs111356468	23:80307658	0.0135	HMGN5	1660	rs11314467	23:78643198	0.012	ITM2A
Weight (females)								
rs2012385	2:242422405	0.296	FARP2	60.2	chr2:241422788	2:242362203	0.0737	ANKMY1
rs7962636	12:116450006	0.401	MED13L	254	rs12810788	12:116196322	0.155	MED13L
Weight (sex-combined)								
rs11161573	1:85724581	0.246	C1orf52	44.2	rs11161548	1:85680408	0.398	SYDE2
rs12755723	1:90007695	0.227	LRRC8B	123	chr1:89419588	1:89885147	0.0289	CCBL2
rs12470032	2:145380620	0.245	ZEB2	225	rs10538039	2:145605197	0.0795	ZEB2
rs7598349	2:201247479	0.482	SPATS2L	275	rs751854756	2:201522012	<0.01	AOX1
rs183675110	6:616564	0.0291	HUS1B	180	rs555334473	6:796747	<0.01	EXOC2
rs7090503	10:18674640	0.383	CACNB2	89.8	rs7912454	10:18584792	0.135	CACNB2
rs3213229	11:2155876	0.0735	IGF2	3.34	rs3168310	11:2152541	0.361	IGF2
rs11217530	11:119708875	0.354	NECTIN1	110	rs10892501	11:119818622	0.408	NECTIN1
rs2682805	12:77556056	0.426	E2F7	103	rs310796	12:77453226	0.329	E2F7
rs7962636	12:116450006	0.401	MED13L	254	rs12810788	12:116196322	0.155	MED13L
rs7332078	13:96554544	0.177	UGGT2	54.9	rs9556514	13:96609478	0.221	UGGT2
rs11156978	14:21924673	0.483	CHD8	0.448	rs78484822	14:21924225	<0.01	CHD8
rs1057035	14:95554142	0.357	DICER1	144	rs10133840	14:95410529	0.309	GSC
rs9905062	17:75721715	0.329	TNRC6C	20	rs9906155	17:75701697	0.0477	TNRC6C

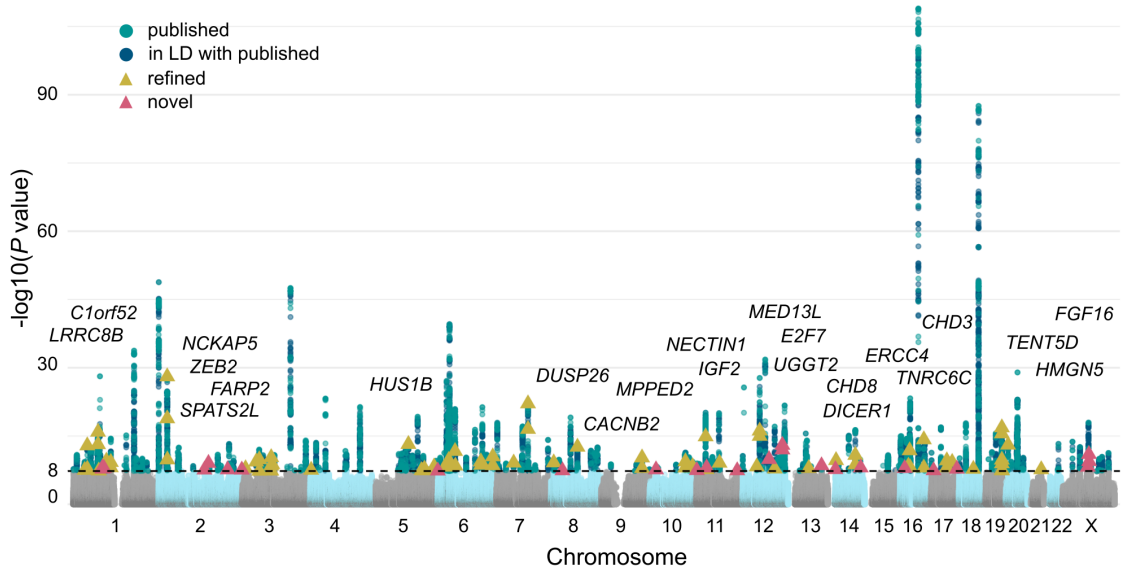


Figure 3.23: Genome-wide novel and refined SNP associations with baseline adiposity estimated over the measurement time window for each individual. Combined Manhattan plot displaying genome-wide SNP associations with body mass index (BMI) or weight linear mixed-effects (LME) model intercepts in sex-specific and sex-combined strata. Each point represents a single SNP, with genome-wide significant (GWS) SNPs ($P < 5 \times 10^{-8}$) coloured in: green for previously published associations, blue for SNPs in linkage disequilibrium (LD) ($r^2 > 0.1$) with published associations, yellow for "refined" SNPs that represent conditionally independent (conditional P-value $P_{\text{cond}} < 0.05$) associations with obesity, with larger effect sizes (P-value for difference, $P_{\text{diff}} < 0.05$), than published variants in the region, and pink for novel associations. Novel loci are annotated with nearest gene.

assessment centre measurements of BMI or weight in up to 276,655 individuals held out of the discovery GWASs, and (2) estimated baseline BMI from linear mixed-effects models in 125,209 participants in the EstBB cohort (Table 3.11).

Six novel SNPs for baseline adiposity replicate in the UKBB (Table 3.12). Some of the replicated variants are in loci that harbour independent obesity-associated SNPs, such as rs7598349 in the *SPATS2L* locus, which contains common variants associated with appendicular lean mass in the UKBB[206], rs7962636 in the *MED13L* locus, whose encoded protein is part of the PPAR-α pathway for activating genes involved in fatty acid oxidation, transport, synthesis, and storage[360], and rs11156978 in the *CHD8* locus, in which rare pathogenic variants confer risk of being overweight compared to the general population[361]. We also discover at least one gene that has not previously been genetically linked to any obesity phenotype, *LRRC8B*, whose

Table 3.11: Sample sizes required for replication of novel baseline adiposity lead SNPs, with effect sizes debiased for winner's curse. BMI = body mass index, SE = standard error.

SNP	chr:pos (hg19)	MAF	Adiposity trait	Beta (SE)	P-value	Minimum replication sample size (80% power)
rs11161573	1:85724581	0.289	BMI_sex_comb	0.0212 (0.00369)	9.52E-09	92,000
rs12755723	1:90007695	0.226	Weight_sex_comb	-0.0214 (0.00375)	1.19E-08	105,000
rs7569339	2:133469717	0.08	BMI_sex_comb	0.0341 (0.00614)	2.69E-08	98,000
rs12470032	2:145380620	0.369	Weight_sex_comb	-0.02 (0.00327)	9.36E-10	89,000
rs7598349	2:201247479	0.462	Weight_sex_comb	0.0176 (0.00316)	2.47E-08	109,000
rs2012385	2:242422405	0.343	Weight_F	0.0252 (0.00454)	2.77E-08	59,000
rs183675110	6:616564	0.014	Weight_sex_comb	0.082 (0.015)	4.47E-08	94,000
rs79586444	8:33688233	0.168	BMI_M	-0.0371 (0.00676)	3.95E-08	44,000
rs7090503	10:18674640	0.494	Weight_sex_comb	-0.0176 (0.00315)	2.42E-08	109,000
rs3213229	11:2155876	0.074	Weight_sex_comb	0.0334 (0.00605)	3.54E-08	111,000
rs294382	11:30607515	0.278	BMI_M	0.0319 (0.00555)	9.14E-09	41,000
rs11217530	11:119708875	0.318	Weight_sex_comb	-0.0185 (0.00337)	3.92E-08	114,000
rs2682805	12:77556056	0.436	Weight_sex_comb	0.0203 (0.00317)	1.32E-10	80,000
rs7962636	12:116450006	0.351	Weight_F	0.0324 (0.00454)	1.02E-12	33,000
rs7332078	13:96554544	0.154	Weight_sex_comb	0.0257 (0.00434)	3.13E-09	97,000
rs11156978	14:21924673	0.485	Weight_sex_comb	-0.0176 (0.00315)	2.17E-08	109,000
rs1057035	14:95554142	0.374	Weight_sex_comb	0.0186 (0.00324)	9.15E-09	104,000
rs17794645	16:13813090	0.161	BMI_M	0.0392 (0.00688)	1.27E-08	41,000
rs2279620	17:7792465	0.113	BMI_sex_comb	0.0289 (0.00526)	3.85E-08	101,000
rs9905062	17:75721715	0.433	Weight_sex_comb	-0.0179 (0.00318)	1.76E-08	107,000
rs112469935	23:76550037	0.017	BMI_sex_comb	-0.0286 (0.00492)	6.32E-09	614,000
rs55669243	23:79680770	0.0109	BMI_sex_comb	-0.0371 (0.00546)	1.02E-11	534,000
rs111356468	23:80307658	0.0135	BMI_sex_comb	-0.0346 (0.00552)	3.71E-10	514,000

encoded protein forms a complex with LRRC8A/SWELL1 to regulate adipocyte size, insulin signalling, and glucose homeostasis[362]. None of these variants replicate in the Estonian Biobank at $P < 2.17 \times 10^{-3}$, but we observe nominal replication of rs7962636 with BMI-change in women ($P = 5.84 \times 10^{-3}$) (Table 3.12). The lack of replication in the Estonian dataset may indicate cohort-specific effects such as the healthy volunteer bias in UKBB, or age-specific effects, as the EstBB population is on average 6 to 8 years younger than the UKBB (Table 3.5).

Longitudinally-derived phenotypes help refine known genetic signals for obesity

The 67 refined SNPs are conditionally independent of published obesity-associated variants in this population (conditional P-value $P_{\text{cond}} < 0.05$), and represent stronger associations with baseline adiposity than do published variants (P-values for difference in effect sizes, $P_{\text{diff}} < 0.05$). Together, the refined and novel SNPs explain 0.416% of variance in baseline BMI (in addition to the 2.92% explained

Table 3.12: Replication of novel lead SNPs for baseline adiposity in: (1) held-out set of cross-sectional adiposity measurements in UK Biobank, and (2) BMI intercept (u1) in Estonian Biobank; significance established at $P < 2.17 \times 10^{-3}$ after Bonferroni-correction at 5% across 23 independent SNPs tested. *blank fields indicate that the SNP was missing in the Estonian Biobank dataset.

SNP	chr:pos (hg19)	Nearest TSS	Strata	Discovery (UK Biobank)		Replication (UK Biobank)			Replication (Estonian Biobank, BMI)		
				Beta (SE)	P-value	Sample size	Beta (SE)	P-value	Sample size	Beta (SE)	P-value
rs11161573	1:85724581	C1orf52	BMI_sex_comb	0.0212 (0.00369)	9.52E-09	276655	0.00395 (0.00298)	0.19	125209	0.00215 (0.00141)	0.126
			Weight_sex_comb	0.0196 (0.00347)	1.66E-08	259852	0.00646 (0.00303)	0.033			
rs12755723	1:90007695	LRRC8B	Weight_sex_comb	-0.0214 (0.00375)	1.19E-08	259852	-0.0125 (0.00328)	0.00014	125209	-0.00191 (0.00154)	0.215
rs7569339	2:133469717	NCKAP5	BMI_sex_comb	0.0341 (0.00614)	2.69E-08	276655	0.00679 (0.00502)	0.18	125209	0.00126 (0.00227)	0.579
rs12470032	2:145380620	ZEB2	Weight_sex_comb	-0.02 (0.00327)	9.36E-10	259852	-0.00923 (0.00287)	0.0013	125209	0.00141 (0.00137)	0.303
rs7598349	2:201247479	SPATS2L	Weight_sex_comb	0.0176 (0.00316)	2.47E-08	259852	0.00954 (0.00277)	0.00056	125209	0.000827 (0.00129)	0.522
rs2012385	2:242422405	FARP2	Weight_F	0.0252 (0.00454)	2.77E-08	141359	0.00774 (0.00404)	0.055			
rs183675110	6:616564	HUS1B	Weight_sex_comb	0.082 (0.015)	4.47E-08	259852	-0.0118 (0.0133)	0.37	125209	-0.00426 (0.0376)	0.91
rs79586444	8:33688233	DUSP26	BMI_M	-0.0371 (0.00676)	3.95E-08	126431	-0.00754 (0.00551)	0.17			
rs7090503	10:18674640	CACNB2	Weight_sex_comb	-0.0176 (0.00315)	2.42E-08	259852	-0.00437 (0.00278)	0.12	125209	-0.00151 (0.0013)	0.243
rs3213229	11:2155876	IGF2	Weight_sex_comb	0.0334 (0.00605)	3.54E-08	259852	0.0133 (0.00533)	0.013	125209	0.00168 (0.0024)	0.483
rs294382	11:30607515	MPPED2	BMI_M	0.0319 (0.00555)	9.14E-09	126431	0.00677 (0.00455)	0.14	44816	0.0028 (0.00256)	0.275
rs11217530	11:119708875	NECTIN1	Weight_sex_comb	-0.0185 (0.00337)	3.92E-08	259852	-0.00163 (0.00296)	0.58	125209	-0.000293 (0.00137)	0.831
rs2682805	12:77556056	E2F7	Weight_sex_comb	0.0203 (0.00317)	1.32E-10	259852	0.0065 (0.00278)	0.019	125209	0.000837 (0.00133)	0.529
rs7962636	12:116450006	MED13L	Weight_F	0.0324 (0.00454)	1.02E-12	141359	0.00671 (0.00403)	0.096	80393	0.00504 (0.00183)	0.00584
			Weight_sex_comb	0.0245 (0.0033)	9.88E-14	259852	0.00944 (0.0029)	0.0011	125209	0.00258 (0.00147)	0.0791
rs7332078	13:96554544	UGGT2	Weight_sex_comb	0.0257 (0.00434)	3.13E-09	259852	0.00868 (0.00382)	0.023	125209	0.00206 (0.00182)	0.258
rs11156978	14:21924673	CHD8	Weight_sex_comb	-0.0176 (0.00315)	2.17E-08	259852	-0.0121 (0.00278)	1.40E-05	125209	2.73E-05 (0.00136)	0.984
rs1057035	14:95554142	DICER1	Weight_sex_comb	0.0186 (0.00324)	9.15E-09	259852	0.015 (0.00286)	1.60E-07	125209	0.00208 (0.00135)	0.124
rs17794645	16:13813090	ERCC4	BMI_M	0.0392 (0.00688)	1.27E-08	126431	-0.00202 (0.00559)	0.72	44816	-0.00178 (0.00361)	0.622
rs2279620	17:7792465	CHD3	BMI_sex_comb	0.0289 (0.00526)	3.85E-08	276655	0.00572 (0.00429)	0.18	125209	0.00333 (0.00208)	0.109
rs9905062	17:75721715	TNRC6C	Weight_sex_comb	-0.0179 (0.00318)	1.76E-08	259852	-0.0062 (0.0028)	0.027	125209	0.00203 (0.00133)	0.128
rs112469935	23:76550037	FGF16	BMI_sex_comb	-0.0286 (0.00492)	6.32E-09	276655	-0.0028 (0.0022)	0.203			
rs55669243	23:79680770	TENT5D	BMI_sex_comb	-0.0371 (0.00546)	1.02E-11	276655	-0.0018 (0.0025)	0.472			
rs111356468	23:80307658	HMGN5	BMI_sex_comb	-0.0346 (0.00552)	3.71E-10	276655	-0.0015 (0.0025)	0.549			

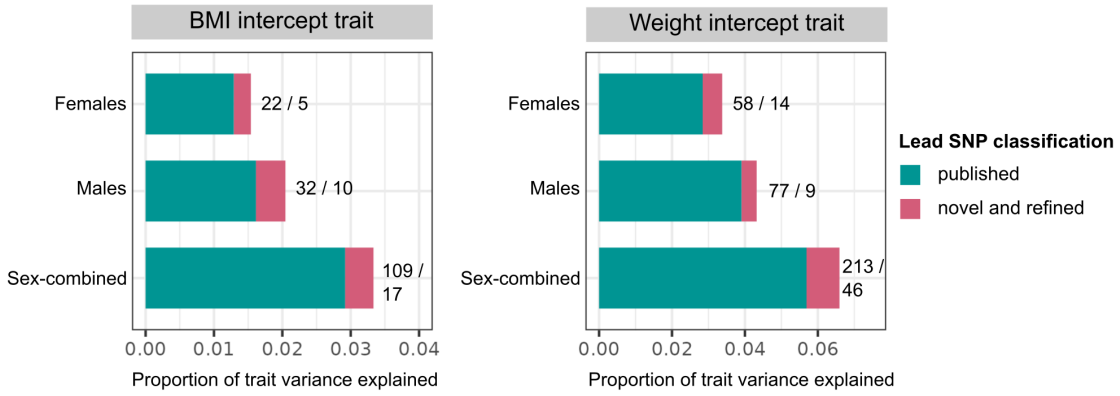


Figure 3.24: Proportion of variance in baseline body mass index (BMI) or weight, estimated using linear mixed-effects (LME) model intercepts, that can be explained by lead variants from genome-wide association studies (GWASs). In green is the proportion of variance explained by previously published obesity-associated SNPs and SNPs in LD with these; in pink is the proportion of variance explained by novel and refined SNPs as defined above. The numbers specify the number of lead variants in each of these categories (published / refined & novel).

by published variants), and 0.904% of variance in baseline weight (in addition to the 5.69% explained by published variants) (Figure 3.24). Finally, I estimated the power gained from utilising baseline BMI from repeat longitudinal data by comparing the GWS SNPs from our in-house BMI GWAS of LME model intercepts to the largest published European-ancestry BMI meta-analysis to date[216]. I observe an increase in median chi-squared statistics of between 5.90% (in sex-combined analyses) and 14.3% (in female-specific analyses) over what would be expected from a cross-sectional GWAS of equivalent sample size.

Modelled baseline adiposity compared to time-averaged adiposity value

The adiposity intercept from LME models can be interpreted as an estimate of an individual's adiposity-trait at the age of first measurement in the dataset, and is strongly correlated with the average adiposity value for an individual across their measurement time window (r^2 between 0.840 and 0.958, Table 3.13). The traits are also genetically similar, as up to 80% of the genetic loci discovered in GWASs for the LME intercept adiposity-trait also reach genome-wide significance in the corresponding GWAS for average-trait, and the effect sizes of GWS variants

Table 3.13: Comparison between linear mixed-effects (LME) model intercepts (u0) and average adiposity-trait across measurement window.

Obesity trait	Sex	Correlation between u0 and average	Number of lead SNPs associated with:		
			Both u0 and average	Only u0	Only average
Body mass index (BMI)	F	0.953	21	6	14
	M	0.952	20	8	9
	combined sexes	0.95	82	23	34
Weight	F	0.956	53	19	16
	M	0.958	57	19	10
	combined sexes	0.84	158	87	35

are nearly identical in both studies (absolute value of mean difference between effect estimates < 0.00172, Figure 3.25).

Informed presence bias in adiposity observations does not affect genetic associations with baseline adiposity

As described in Section 3.3.4, there is ascertainment bias in our discovery cohort – on average, women with ten or more weight measurements are 8.3 kg (3.6 units of BMI) heavier than their counterparts with 1 to 3 measurements, with a corresponding difference in men of 8.2 kg (3.1 units of BMI) (Table 3.3) – which may suggest that our longitudinally-derived metrics fail to reflect the population of the UKBB or the UK more broadly. However, the BMI LME model intercept metric from our in-house GWAS is genetically perfectly correlated with the unascertained cross-sectional BMI GWAS in the GIANT Consortium[216] ($r_G = 1$ and $P < 1 \times 10^{-16}$ in all sex strata). Furthermore, 95.7% of the GWS associations identified in our studies have either been previously published, or are correlated with published SNPs for obesity in the GWAS Catalog[50], suggesting that we capture a genetically similar phenotype to other studies of obesity.

3.3.6 No evidence for genetic basis of within-individual variability in weight

Within-individual (intra-individual) variances across multiple measurements of BMI (mean $\sigma^2 = 3.02$) and weight (mean $\sigma^2 = 21.9$) are between 10 to 14 times lower

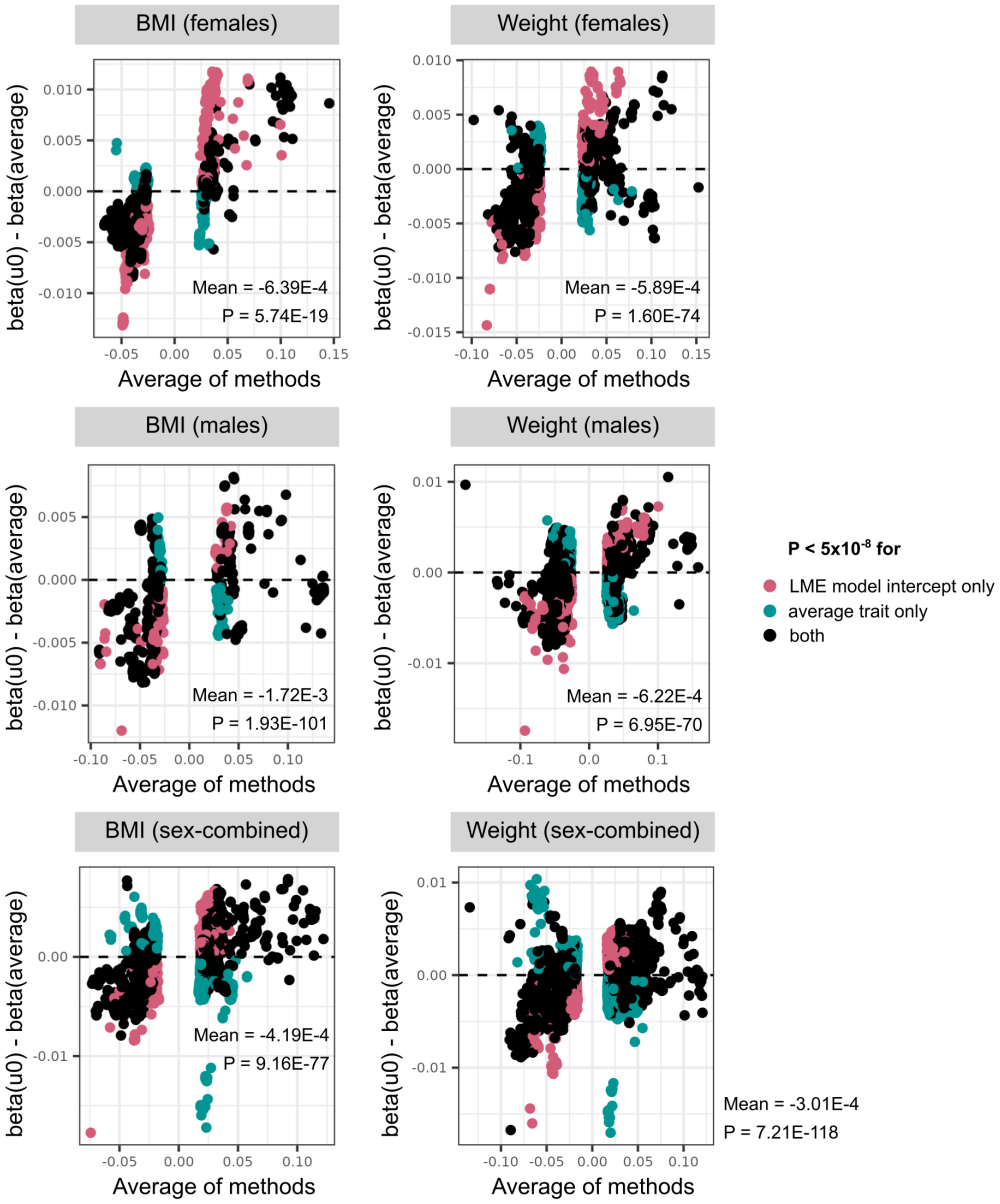


Figure 3.25: Bland-Altman plot[357] to compare effect sizes of genome-wide significant (GWS) SNPs between average adiposity-trait and linear mixed-effects (LME) model intercept (u0) genome-wide association studies (GWASs). Each panel represents a different analysis stratum. SNPs that reach genome-wide significance ($P < 5 \times 10^{-8}$) in GWASs for either the average-trait or LME-u0 GWASs are plotted, coloured by whether SNP is significant in both GWASs or only one. For each GWS SNP, the difference between effect sizes estimated in the average-trait regression and LME-u0 regression is plotted against the mean estimate from both regressions. The mean difference and one-sample t-test P-value (with null hypothesis of mean difference = 0) are displayed in each panel. All GWASs were adjusted for the first 21 genetic PCs, genotyping batch, and genotyping array.

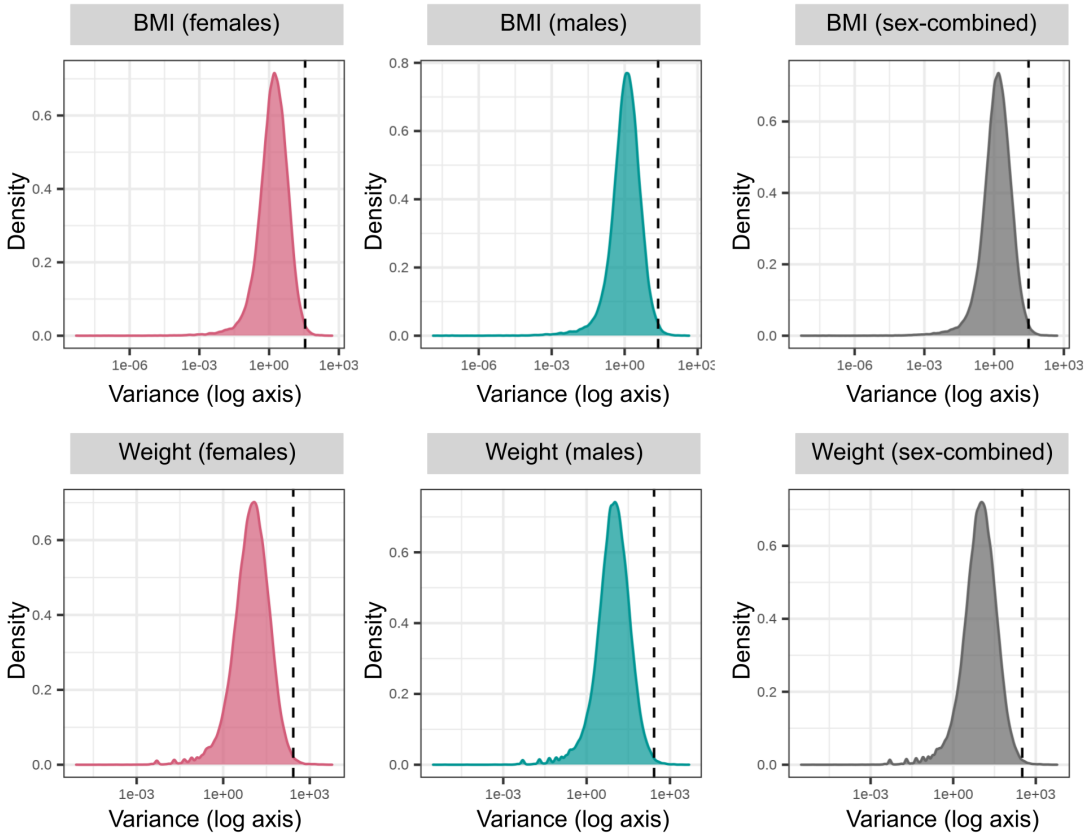


Figure 3.26: Within-individual and between-individual variance in body mass index (BMI) and weight. The distribution of within-individual (intra-individual) variance in BMI and weight across multiple measures is plotted on the log-axis; each panel represents a different stratum as labelled. In dashed lines are the between-individual (inter-individual) variances in each stratum.

than the corresponding between-individual (inter-individual) variances ($\sigma^2_{\text{BMI}} = 30.3$ and $\sigma^2_{\text{weight}} = 322$). However, there is a large spread in the values of within-individual variance (Figure 3.26). We therefore tested whether there was a genetic component to this variance by jointly modelling genome-wide effects on the mean and within-individual variability in weight for 177,472 unrelated individuals of European ancestry with multiple measurements of weight using the `TrajGWAS` software. While we replicate 138 known obesity-related genetic variants for mean weight at GWS ($P < 5\times10^{-8}$), we find no evidence for significant genetic associations with within-individual variability in weight (Figure 3.27).

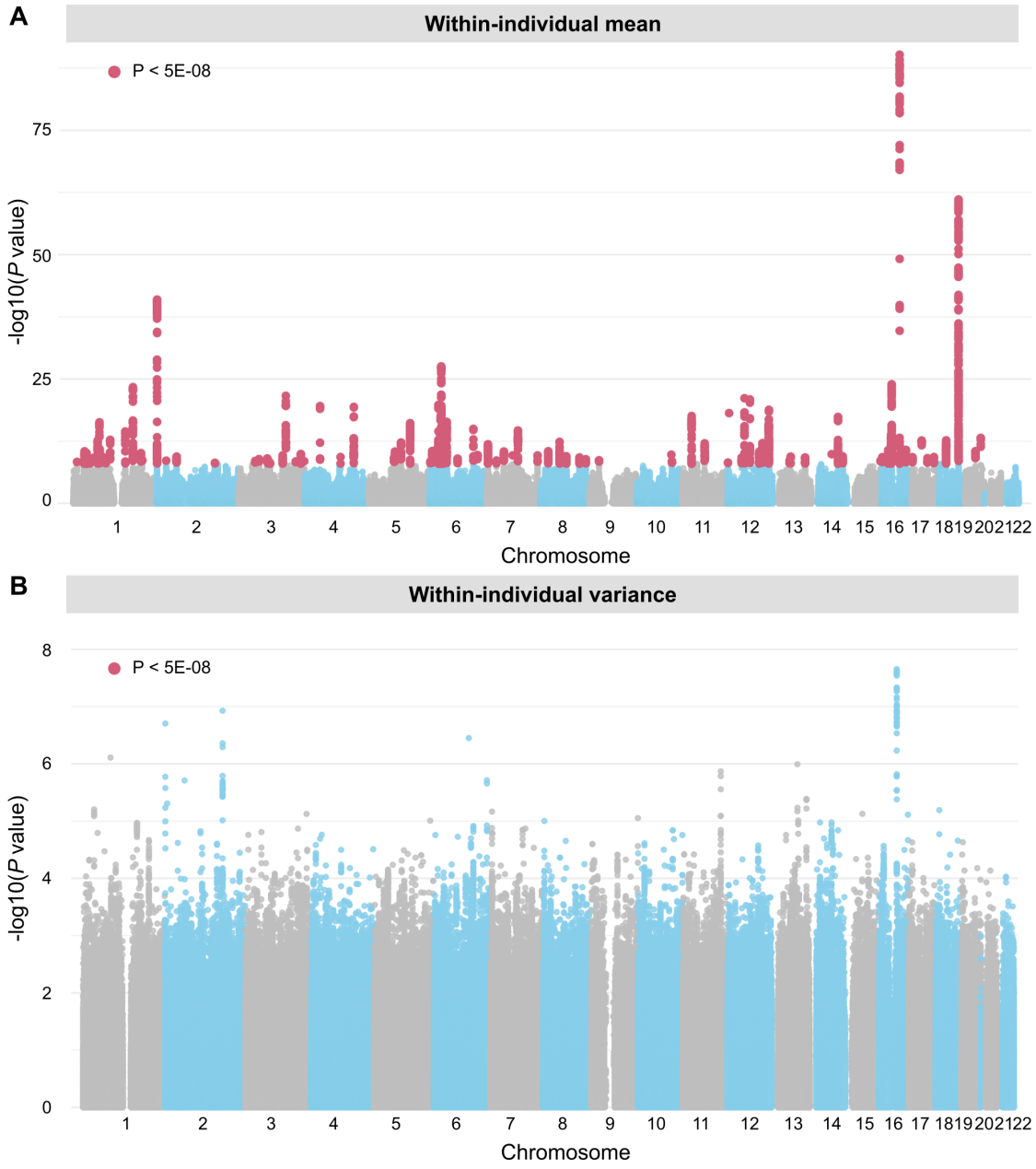


Figure 3.27: Manhattan plots depicting SNPs associated with jointly modelled within-individual (A) mean and (B) variance in weight over multiple measurements. Genome-wide association studies (GWASs) for within-individual mean and variance were performed using the `TrajGWAS` software for 177,472 unrelated male and female participants of European ancestry with multiple measurements of weight in the UK Biobank (UKBB) and UKBB-linked electronic health records. Each point is a genetic variant, represented along the hg19 chromosome and position x-axis, with $-\log_{10}(P)$ for significance of association with the phenotype on the y-axis. Variants that cross the genome-wide significance threshold of $P < 5 \times 10^{-8}$ are highlighted in pink.

3.4 Discussion

By combining genetic and EHR data modalities in the UK Biobank for 2.95 million observations of BMI and weight across 184,379 participants, I find nine genetic loci associated with change in adiposity over adulthood. Five of these variants replicate in held-out participants of the UKBB, two in the Estonian Biobank, and one (rs429358 in *APOE*) in the Million Veteran Program, USA, representing validation across three global cohorts of European-ancestry individuals. I show that the heritability of adiposity-change, h^2_G between <1% to 3.47%, is approximately 10x lower than that of baseline adiposity, and is sex-specific with larger effects in women. We demonstrate the boost in power for genetic associations from modelling multiple measurements of adiposity over time, and discover 23 novel lead variants associated with weight or BMI, including one near *LRRC8A*, a gene that has not previously been linked to obesity in humans. While previous efforts to understand the genetic basis of weight-change have relied primarily on twin studies[224,225,309], or have assessed the effects of cross-sectional obesity risk scores[311] on change in adiposity, I show here for the first time that combining longitudinal EHR data with genome-wide imputation of genetic variants enables the discovery of SNPs associated with adiposity-change in adults, independent of their effects on baseline obesity.

Understanding the genetic basis of adiposity-change may help inform personalised treatment for obesity, which is associated with a range of co-morbidities[363] and is a growing global health burden[364]. Several studies have investigated the interactions between genetic variants for cross-sectional obesity and weight loss guided by surgical[365,366], medical[367], dietary[368], or behavioural[367,369–371] interventions, but these are limited by small sample sizes (fewer than 1,000 participants) and short follow-up times (6 months to 5 years), with inconsistent results among studies. Given the evidence in this Chapter for distinct genetic determinants of, and modest genetic correlations between, baseline obesity and weight change, I propose that genetic variants associated with adiposity-change will provide a distinct and complementary avenue to cross-sectional genetics for the discovery of weight-loss therapies.

Data from EHRs provide a rich resource for longitudinal health research, but as they are not designed for research, studies must be cautious to mitigate against error-prone records, informative patient sampling, sparsity, and non-random missingness, among other biases[148,149]. The robustness of our results leads us to believe that our curated UKBB-linked EHR dataset guards against some of these biases, as: (1) while the UKBB-linked EHR cohort has an over-representation of sick patients with higher BMI than a random sampling of UKBB participants, the UKBB is itself skewed towards healthier individuals[264] with lower BMI than the national average; (2) data in EHRs is of lower quality than observations collected for research purposes[148], but we append the more reliable UKBB assessment centre measurement records to EHR data for this study, and (3) we thoroughly investigate outlier data entries in the EHR through stringent population-level and individual-level QC, which controls the rates of technical error.

Throughout this Chapter, I highlight that the choice of longitudinal metric for genetic studies must be phenotype-aware. While within-individual variance is a biologically meaningful phenotype for traits such as blood pressure or triglyceride levels, whose fluctuations are indicative of risk for heart disease[372,373], most individuals follow steady patterns of long-term changes in weight[374,375] and we find no evidence for a genetic basis to intra-individual variance in weight. Rather, metrics such as LME slopes and cluster-belonging probabilities, whether calculated via observations of BMI or weight, provide robust, replicable phenotypes for GWAS. Indeed, three of nine genetic variants associated with our derived adiposity-change metrics are also associated with self-reported weight-change over a one year period, despite the unreliability of self-reported assessments of weight[376]. That the other six variants are not associated with weight change over a one year period may be due to the shorter length of follow-up, demonstrating the value of regular follow up over longer terms in longitudinal studies. Studies of adiposity-change should consider the baseline age profiles of participants, as the average healthy individual gains weight throughout early and mid-adulthood and begins to gradually lose weight beyond age 60 years[358,359]. Here, we adjust for linear and quadratic effects

of baseline age to replicate genetic associations with BMI change in two cohorts with age profiles that are distinctly different from the UK Biobank (mean baseline age = 50.1 years, (SD = 9.57)), i.e. the Million Veteran Program (mean baseline age = 55.8 years (10.7)) and the Estonian Biobank (mean baseline age = 42.0 years (14.5)). The differences in replication success across cohorts indicate that it will be important to further interrogate the age-specific effects of genetic variants on weight and BMI across adulthood.

It is also crucial to interrogate the sex-specificity of age-related changes in adiposity, which we show here to have modestly higher heritability in women. The gynoid-like fat distribution in pre-menopausal women, which favours gluteo-femoral fat accumulation, shifts to android-like fat distribution favouring abdominal adiposity deposits after menopause[180,222], and is accompanied by higher risk of cardio-metabolic disease in post-menopausal women[181,377]. Body fat distribution has been shown to have a sex-specific genetic basis, with higher heritability in women[216], but the mechanisms by which genetic variants cause changes in body fat distribution over time remain unknown. It is hypothesised that bi-directional causal relationships between sex hormones, and adipose tissue, may be involved in the menopause-associated shift in fat distribution[378,379]. We were under-powered to find genetic variants associated with changes in abdominal obesity, for which the UKBB-linked EHR had 10-15 times fewer repeat observations than for overall obesity, but understanding the mechanisms of age-related and sex-specific changes in fat distribution remains an open question with pressing health implications.

The association of the minor C allele of rs429358, a missense variant in *APOE*, with weight loss independent of baseline obesity, replicates across a range of adiposity-change metrics in individuals of European ancestry in three cohorts across the globe. *APOE* codes for apolipoprotein E, which is a core component of plasma lipoproteins and is essential for cholesterol transport and homesotasis[380] in several tissues across the body, including the central nervous system, muscle, heart, liver, and adipose tissue[381]. Our study and previous work have shown associations of the pleiotropic rs429358 variant with hundreds of biomarkers and diseases[382]. We find associations

between rs429358 and linear rates of change in several traits associated with cross-sectional obesity, such as total and HDL cholesterol[383,384], triglycerides[384], and the chronic inflammation marker CRP[385]; however, the longitudinal and causal nature of these relationships remains to be established. As rs429358 is also reported to be associated with longevity[338], we ensured the robustness of our results to survival bias through three independent approaches, including by demonstrating that other longevity-associated variants have no effects on weight change. Furthermore, as rs429358 is strongly associated with dementia[339], which is itself often accompanied by weight loss[340], we validated the effect of this variant on weight loss in a subset of individuals without diagnoses of dementia in the UKBB. Given the replicability of our results across a range of analyses and cohorts, I hypothesise that the *APOE* effect on weight loss may act through cholesterol regulation pathways that determine the body's metabolic response to environmental and dietary factors, as seen in mouse models[386,387]. It has recently been suggested that the onset and severity of Alzheimer's disease in individuals with the rs429358 variant may be mediated by *APOE*-linked cholesterol dysregulation[388], which suggests that systemic, ageing-associated aberrations in cholesterol homeostasis may have far-ranging consequences from weight loss to cognitive decline.

Our two-stage analytic approach, wherein we calculate longitudinal summary statistics in the first stage, for GWAS in the second stage, aims for computational efficiency and convenience by approximating the fitting of a full joint model that would incorporate raw adiposity measurements and genome-wide imputed variants in the same model. Our approach is composable and scales well to millions of observations across hundreds of thousands of individuals, which enables us to construct longitudinal models efficiently using a range of models suited for the data, separately from the genetic analysis which we can perform efficiently using software designed specifically to conduct large-scale GWAS. However, this approximation of a full joint model means that we cannot propagate posterior uncertainty from the longitudinal sub-model, as we only take forward a point estimate (LME model BLUPs or clustering probabilities) to the genetic sub-model. The growing scale of,

and interest in, longitudinal data for genetics research demands the development of methods such as Markov melding[389] combined with mixed-model GWAS to allow for genetic association testing in the presence of population stratification.

The UKBB resource has enabled thousands of health-related scientific discoveries in individuals of European ancestry[390], but lower sample sizes for other ancestral groups limits the generalisability of these findings, widening existing racial health inequities[391]. Large sample sizes are particularly relevant for the study of longitudinal traits, as we show here that the heritability of weight change is an order of magnitude lower than cross-sectional weight. The lower heritability may be in part due to time-varying environmental covariates that are associated with weight change, such as smoking status[392] or medication use[393], which we did not account for in our study. For example, the association of a variant in *TCF7L2*, which is the strongest known genetic risk factor for type 2 diabetes[394], with weight loss in our study may arise from weight loss as a consequence of treatments for diabetes[393,395]. Such considerations will become even more important in the future, with rising rates of prescription and self-medication of GLP-1 receptor agonists as a treatment for obesity[396,397]. However, it is not straightforward to ascertain the direction of causality by which non-genetic covariates are associated with weight change. For example, the use of statins to lower blood pressure is connected to weight gain[398,399], but high blood pressure may itself be a consequence of weight gain[400], and inappropriate adjustments along this causal pathway may lead to unexpected collider biases[356]. In general, despite their longitudinal nature, it is challenging to determine causality from EHRs alone without a prospective study design. Recent statistical developments in emulating randomised control trials from observational EHR data have been successful at addressing these concerns[401,402], and the creative integration of these methods with genetic data offers a promising avenue for multi-modal causal inference.

In this Chapter, I describe the largest study to date of the genome-wide architecture of adiposity-change in adulthood, and the first to identify replicable genetic variants associated with weight change, independent of baseline obesity. I

conducted a detailed investigation of the genetic effects of rs429358 in the *APOE* locus, which may lead to weight loss through systemic ageing-associated cholesterol dysregulation; more such biological pathways are likely to emerge from large-scale longitudinal genetic studies. Finally, I add evidence to support the growing use of EHR data in genetics research[107,134], particularly through the added power of longitudinal EHR data for the discovery of novel biology. Integrating multiple correlated biomarkers in EHRs, with information on rare genetic variants in biobanks, may provide novel insights into ageing, with broad implications for human health.

4

The distinct genetic components of infertility and reproductive hormones

Contents

4.1 Introduction 114
4.2 Methods 115
4.2.1 GWASs of infertility and reproductive hormones in UKBB 115
4.2.2 Meta-analysis of GWAS summary statistics 116
4.2.3 Genetic correlations 122
4.2.4 Colocalisation 123
4.2.5 Overlaps with genetic regions under selection 124
4.2.6 Tissue and cell-type prioritisation 126
4.2.7 Whole exome sequencing analyses in UKBB 127
4.3 Results 128
4.3.1 Genome-wide meta-analyses identify novel genetic loci for female and male infertility 128
4.3.2 Selection pressure may explain the persistence of some infertility-associated variants in the population 135
4.3.3 Genetic determinants of reproductive hormone levels . . 139
4.3.4 Non-overlapping genetics of hormones and infertility . . 151
4.3.5 Phenome-wide genetic correlation landscape differs for infertility and reproductive hormones 153
4.3.6 Heritability of testosterone is enriched in adrenal gland, liver, and ovarian cell types 154
4.3.7 Rare variant contribution to reproductive hormone and infertility genetics 157
4.4 Discussion 165

4.1 Introduction

Infertility, which is defined as the inability to achieve pregnancy within 12 months of regular unprotected sexual intercourse, affects one in six couples across the globe[403]. A range of environmental and genetic factors may drive infertility, including the decline of sperm and oocyte quality and quantity with parental age[404,405], sexually and non-sexually transmitted infectious diseases such as genital tuberculosis[406], chlamydia[407], and gonorrhoea[408], and rare Mendelian disorders such as cystic fibrosis[409,410]. However, the exact cause of infertility cannot be diagnosed in up to 28% of couples, and up to 40% of women, with infertility[411,412].

Heritable diseases of the female reproductive tract, particularly endometriosis[253] and PCOS[413], are thought to be responsible for a large proportion of female infertility cases, with PCOS in particular accounting for up to 80% of cases of anovulatory infertility[254]. It is hypothesised that sex hormone dysregulation[414,415] and obesity[416], which often accompany reproductive diseases, may also be involved in the aetiology of infertility. The hypothalamic-pituitary axis, which also includes the thyroid gland[417], adrenal gland[418], and gonads (ovaries and testes)[419], is central to this regulation. The gonadotropins follicle stimulating hormone (FSH) and luteinising hormone (LH), which are produced by the anterior pituitary in response to gonadotropin-releasing hormone, induce steroid hormone production in the gonads[419]. The hypothalamic-pituitary axis is tightly regulated by complex feedback loops, and disruptions anywhere along the axis can cause infertility through conditions such as hypogonadotropic hypogonadism[420], hypothyroidism[421], and primary ovarian insufficiency[422]. Furthermore, hormonally active adipose tissue interacts closely with the hypothalamic-pituitary axis through the production and metabolism of leptin and steroid hormones, which are thought to mediate the association between obesity and infertility[423,424].

However, little is known about the genetics of reproductive hormones and infertility, which are notoriously difficult phenotypes to estimate in large-scale

populations. Hormone levels fluctuate widely, particularly as they cycle in menstruating women[226], so robust repeated measurements may be necessary to outline their underlying genetic basis. Infertility has historically been poorly diagnosed and under-reported[403], and negative selection against infertility naturally limits the frequency of risk alleles in the population[259]. Genome-wide association studies (GWASs) have thus typically queried proxy measures for infertility such as childlessness[260,261] or age at first sexual intercourse[262,263], but these are likely to be confounded by socio-economic and behavioural factors.

We aggregated data from a range of sources, including primary care and hospital EHRs and self-report, across seven global cohorts with over 1 million participants, to perform the first reported GWAS meta-analyses for male infertility and five categories of female infertility; we also report results from the largest sex-specific GWASs to-date for five reproductive hormones. Through these meta-analyses and whole-exome sequencing analyses in the UK Biobank, we comprehensively catalogue the common and rare genetic contributions to infertility and reproductive hormone levels, quantify the extent of shared genetic architecture between these traits, and prioritise genes and cell types for further functional investigation of the hormonal and non-hormonal drivers of infertility.

4.2 Methods

4.2.1 GWASs of infertility and reproductive hormones in UKBB

Identification of infertility cases were performed by Dr. Laura Wittemans and Dr. Teresa Ferreira, both post-doctoral researchers in the Lindgren group.

The data for the UKBB-based GWASs in this study were derived from the UKBB-linked EHR resource described in Chapter 2. Identification and QC of hormones – FSH, LH, oestradiol, progesterone, and testosterone – are described in Section 2.2 of the above Chapter. For the European-ancestry analyses, we restrict our dataset to individuals of genetically ascertained non-Finnish European ancestry; other ancestral groups are genetically ascertained as described in Section 2.4.

We defined five categories of female infertility - all causes (F-ALL, 3,746 cases / 260,413 controls), anovulatory (F-ANOV, 53 cases / 264,106 controls), anatomical (F-ANAT, 34 cases / 264,125 controls), idiopathic infertility by exclusion of known causes (i.e. anatomical and anovulatory infertility, PCOS, endometriosis, and uterine leiomyoma) (F-EXCL, 3,012 cases / 261,147 controls), and idiopathic infertility by inclusion of a diagnosis code for idiopathic infertility (F-INCL, 78 cases / 264,125 controls) - and male infertility of all causes (M-ALL, 650 cases / 222,393 controls). Cases were identified through self-report, primary care read codes, and secondary care ICD codes amongst European-ancestry participants in the UKBB (Tables C.1 and 4.1). Sex-matched controls in each category were defined as individuals not identified as cases for that category. For traits with ≥ 50 cases, i.e. all but F-ANAT, we performed genetic association testing, adjusted for the first 21 genetic PCs, genotyping batch, and genotyping array, as described in Section 2.4 of Chapter 2.

I derived cross-sectional hormone phenotypes by retaining the observed trait value closest to the individual's median trait value over time (if multiple measurements were present) for FSH, LH, oestradiol, progesterone, and testosterone in women and men of genetically ascertained European, African, East Asian, and South Asian ancestries in the UKBB-linked EHR (Tables A.1 and 4.2). Within each sex- and ancestry-group, I regressed the derived cross-sectional hormone trait on age, age^2, year of birth, a one-hot encoding of data provider, and applied deterministic rank-based inverse normal transformation to the residuals, as in Equation (3.3). Across sex-specific and sex-combined strata with ≥ 100 samples, the transformed trait was tested for genetic association, adjusted for the first 21 genetic PCs, genotyping batch, and genotyping array, as described in Section 2.4 of Chapter 2.

4.2.2 Meta-analysis of GWAS summary statistics

Contributing cohorts

All participating studies were approved by their respective institutional review boards. We provided standardised protocols to identify infertility cases and controls (scripted by Dr. Laura Wittemans), and quantitative hormone trait values (scripted by me),

Table 4.1: Contributing cohort overview for infertility genome-wide association study (GWAS) meta-analyses.

Cohort	Ancestry	Infertility category	Sample size		
			Cases	Controls	Prevalence
UK Biobank	European	Female infertility of all causes (F_ALL)	3,746	260,413	1.438%
		Anovulatory female infertility (F_ANOV)	53	264,106	0.02%
		Anatomical female infertility (F_ANAT)	34	264,125	0.013%
		Idiopathic female infertility by exclusion (F_EXCL)	3,012	261,147	1.153%
		Idiopathic female infertility by inclusion (F_INCL)	78	264,125	0.03%
		Male infertility of all causes (M_ALL)	650	222,393	0.292%
Estonian Biobank	European	Female infertility of all causes (F_ALL)	12,028	111,407	10.796%
		Anovulatory female infertility (F_ANOV)	3,384	111,407	3.038%
		Anatomical female infertility (F_ANAT)	2,299	111,407	2.064%
		Idiopathic female infertility by exclusion (F_EXCL)	5,558	111,407	4.989%
		Idiopathic female infertility by inclusion (F_INCL)	5,067	111,407	4.548%
		Male infertility of all causes (M_ALL)	1,294	18,551	6.975%
FinnGen	European	Female infertility of all causes (F_ALL)	14,759	111,583	13.227%
		Anovulatory female infertility (F_ANOV)	2,792	111,583	2.502%
		Anatomical female infertility (F_ANAT)	1,795	228,515	0.786%
		Idiopathic female infertility by exclusion (F_EXCL)	11,303	219,007	5.161%
		Idiopathic female infertility by inclusion (F_INCL)	2,420	227,890	1.062%
		Male infertility of all causes (M_ALL)	1,429	130,139	1.098%
Danish Biobank	European	Female infertility of all causes (F_ALL)	8,361	63,434	13.181%
		Anovulatory female infertility (F_ANOV)	545	71,250	0.765%
		Anatomical female infertility (F_ANAT)	1,233	70,562	1.747%
		Idiopathic female infertility by exclusion (F_EXCL)	3,352	68,443	4.898%
		Idiopathic female infertility by inclusion (F_INCL)	2,714	69,081	3.929%
		Male infertility of all causes (M_ALL)	3,512	39,495	8.892%
deCode	European	Female infertility of all causes (F_ALL)	1,130	118,821	0.951%
		Anatomical female infertility (F_ANAT)	910	109,114	0.834%
		Idiopathic female infertility by exclusion (F_EXCL)	133	116,306	0.114%
Genes and Health	South Asian	Female infertility of all causes (F_ALL)	1,176	21,347	5.509%
		Anovulatory female infertility (F_ANOV)	585	21,938	2.667%
		Anatomical female infertility (F_ANAT)	594	21,930	2.709%
		Idiopathic female infertility by exclusion (F_EXCL)	924	21,599	4.278%
		Idiopathic female infertility by inclusion (F_INCL)	593	21,930	2.704%
		Male infertility of all causes (M_ALL)	213	17,786	1.198%

Table 4.2: Contributing cohort overview for reproductive hormone genome-wide association study (GWAS) meta-analyses.

Cohort	Ancestry	Hormone	Sex	Sample size
UK Biobank	European	Follicle stimulating hormone (FSH)	F	20,800
			M	2,114
		Luteinising hormone (LH)	F	16,391
			M	2,237
		Oestradiol	F	54,609
			M	17,491
		Progesterone	F	1,755
		Testosterone	F	197,038
			M	197,340
deCode	European	Follicle stimulating hormone (FSH)	F	23,038
			M	3,981
		Luteinising hormone (LH)	F	16,391
			M	18,804
		Oestradiol	F	33,013
			M	17,483
		Progesterone	F	10,614
		Testosterone	F	29,503
			M	28,093
Estonian Biobank	European	Follicle stimulating hormone (FSH)	F	4,457
		Luteinising hormone (LH)	F	3,353
		Oestradiol	F	2,740
		Progesterone	F	2,683
		Testosterone	F	3,306
			M	1,113
Avon Longitudinal Study of Parents and Children (ALSPAC)	European	Follicle stimulating hormone (FSH)	F	3,101
		Luteinising hormone (LH)	F	3,102
		Testosterone	F	2,380
UK Biobank	African	Oestradiol	F	1,544
		Testosterone	F	3,681
			M	6,867
UK Biobank	East Asian	Testosterone	F	1,079
UK Biobank	South Asian	Oestradiol	F	1,039
		Testosterone	F	2,880
			M	4,139
Genes and Health	South Asian	Follicle stimulating hormone (FSH)	F	6,494
		Luteinising hormone (LH)	F	6,336
		Oestradiol	F	2,335
		Progesterone	F	2,055
		Testosterone	F	3,643
			M	1,530

from self-report, primary care, and secondary care EHRs. We additionally provided guidelines for QC of genetic data and GWASs, jointly scripted by Dr. Laura Wittemans, Dr. Teresa Ferreira, and myself. Analysts from each cohort performed phenotype extraction, QC of phenotypic and genetic data, and GWASs following these protocols.

Avon Longitudinal Study of Parents and Children (ALSPAC). ALSPAC is a longitudinal population-based study that recruited 13,761 pregnant women resident in the South West of England, with expected delivery dates between 1 April 1991 and 31 December 1992, who have continued to be followed up over the last 32 years[144]. In this study, we performed GWASs on reproductive hormone values measured in follow-up assessments (i.e. not during pregnancy) from approximately 3,000 mothers with linked genetic data. Genome-wide genotyping was performed using the Illumina 660 quad array at 467,540 markers and imputed to the hg19 reference genome[144].

Copenhagen Hospital Biobank (Danish Blood Donor Study). The Copenhagen Biobank includes samples drawn for blood type testing or antibody screening from 423,000 hospitalised patients and outpatients in Danish hospitals, with genome-wide genotyping and imputed data available for 154,000 participants[425]. Genotyping was performed by deCODE Genetics using the Illumina GSA microchip at 660,000 markers and imputed to the hg38 reference genome using a custom panel[425].

deCODE Genetics. deCODE Genetics has collected DNA samples for 160,112 Icelanders, linked to deeply phenotyped records from questionnaires and healthcare centres[127]. Variants were identified by whole-genome sequencing and mapped to the hg38 reference genome using a custom panel[111].

Estonian Biobank. The Estonian Biobank is a population-based biobank of Estonia, reflecting the age, sex, and geographic distribution of the population, with a cohort size of 200,000 individuals (approximately 20% of the nation) with EHR linkage[128]. Individuals were genotyped using the Illumina GSA microchip at 700,000 markers and imputed to the hg19 reference genome using a custom panel[344].

FinnGen. The FinnGen research project includes a collection of 500,000 samples from research cohorts in Finnish hospitals and prospective Finnish biobanks with EHR linkage[129]. Individuals were genotyped using ThermoFisher Axiom custom arrays at 664,510 markers and imputed to the hg38 reference genome using a custom panel[129].

Genes and Health. The Genes and Health study is a longitudinal prospective cohort of participants of British Bangladeshi and Pakistani origin recruited from East London, Bradford, and Manchester, with EHR linkage[426]. Participants were genotyped using the Illumina GSA Multi-Disease 24v3-0-EA chip and imputed to the hg38 reference genome using a multi-ancestral reference panel from the 1000 Genomes Project[280,426].

Publicly available GWAS summary statistics for reproductive hormones

I accessed publicly available GWAS summary statistics, without UKBB data, for FSH and LH[427–429]), progesterone[246], oestradiol[246], and testosterone[246,427,430]) from the GWAS Catalog[50] on 27 March 2023 (Table 4.3). As the extraction of hormone values, covariate adjustments, and methods for genetic association analyses differed from the protocols set out in this Chapter for cohort GWASs, I report sensitivity results from GWAS meta-analyses of hormones excluding the publicly available data.

Quality control of GWAS summary statistics

The downloaded summary statistics from Suhre *et al.* (2017)[428] were missing non-effect alleles, which I merged in by matching chromosome and genome position information from Ensembl hg19 variant call files[431]. These and the summary statistics from Prins *et al.* (2017)[430] were missing allele frequencies, which I merged in from chromosome, position, and allele-matched frequencies derived from European-ancestry individuals in the 1000 Genomes dataset[280]. All summary statistics on genome build hg19 were lifted over to hg38 using the `UCSC liftOver` tool[432] - between 97.3% and 99.7% of variants were successfully lifted over to the hg38 genome assembly. Summary statistics from all studies were filtered to retain variants that met the following criteria: (1) on the autosomes or X chromosome, (2)

Table 4.3: Publicly available summary statistics for reproductive hormone genome-wide association study (GWAS) meta-analyses.

Publicly available summary statistics				
Cohort	**Ancestry**	**Hormone**	**Sex**	**Sample size**
Dennis et al. 2021	European and African	Follicle stimulating hormone (FSH)	sex-combined	3,484
		Luteinising hormone (LH)	sex-combined	2,133
		Testosterone	sex-combined	1,588
Suhre et al. 2017	European	Follicle stimulating hormone (FSH)	sex-combined	997
		Luteinising hormone (LH)	sex-combined	997
Thareja et al. 2023	Middle Eastern	Follicle stimulating hormone (FSH)	sex-combined	2,935
		Luteinising hormone (LH)	sex-combined	2,935
Pott et al. 2019	European	Oestradiol	F	2,607
		Oestradiol	M	4,191
		Progesterone	F	1,261
		Testosterone	F	3,352
		Testosterone	M	4,291
Prins et al. 2017	European	Testosterone	M	4,259

with imputation information score > 0.8 (where available), (3) bi-allelic variants with A, C, G, T alleles, (4) with plausible standard errors (< 10) and P-values ($0 < P < 1$), and (5) without duplicate entries.

Meta-analysis

Fixed-effects inverse variance weighted (IVW) meta-analysis was performed using `METAL`[78]. We report results from European-ancestry and all-ancestry meta-analyses for each trait. Genome-wide significance was established at 5×10^{-8}, and lead variants are reported if they have MAF $> 1\%$ or are present in at least two studies. Hormone-associated variants were classified based on conditional analysis as - (1) previously reported for the hormone of interest, (2) previously reported for any of 28 reproductive hormones, or (3) novel - based on SNP associations published in the GWAS Catalog as of 27 March 2023[50] (Table C.2). Full details of the identification of lead variants using distance-based pruning, classification by reported and novel status, and post-GWAS steps to estimate heritability, genetic correlations, and test for sex-heterogeneity, are described in the relevant sections of Chapter 2.

4.2.3 Genetic correlations

Infertility and reproductive hormones

Genetic correlations among the three categories of female infertility with significant heritability[62] ($z > 4$), i.e. F-ALL, F-ANOV and F-INCL, as well as between female infertility and heritable female reproductive hormones (FSH and testosterone) were performed as detailed in Chapter 2. All analyses in this section were restricted to individuals of European-ancestry, as LD-score regression relies on population-specific LD patterns. I additionally tested for genetic correlation between phenotypes in our study and previously characterised reproductive hormones; I obtained summary statistics from GWASs of TSH[433] (sex-combined analysis, N = 247,107 participants) and anti-Müllerian hormone (AMH) (N = 7,049 pre-menopausal participants)[434] from the largest publicly available European-ancestry studies to date. Significant r_G after multiple testing was established at 2.5×10^{-3} (FWER controlled at 5% across 20 tests using the Bonferroni method).

Female reproductive conditions

An overview of the studies contributing to female reproductive outcomes in this Chapter, including case and control definitions and sample sizes, can be found in Table D.1 with detailed methods described in Section 5.2.2. In brief, I obtained European-ancestry GWAS summary statistics for four female reproductive conditions - (1) endometriosis from Rahmioglu *et al.* (2023)[251], with 57,248 cases and 698,764 controls, (2) heavy menstrual bleeding (HMB) by meta-analysing GWASs from Gallagher *et al.* (2019)[252] and FinnGen data freeze 9[129], totalling 31,309 cases and 318,510 controls, (3) PCOS by meta-analysing GWASs from Tyrmi *et al.* (2022)[250], Day *et al.* (2018)[197], and an in-house UKBB-based GWAS, totalling 14,467 cases and 430,267 controls, and (4) uterine leimoyomas by meta-analysing GWAS summary statistics generated by the Neale lab[354] and FinnGen data freeze 9[129], totalling 42,446 cases and 588,955 controls. I additionally obtained summary statistics from a GWAS of spontaneous dizygotic twinning (8,265 cases and 264,567 controls) from the largest European-ancestry study of female fecundity to date[435]. I tested for

genetic correlation between phenotypes in our study and the five reproductive conditions detailed above. Significant r_G after multiple testing was established at 2.00×10^{-3} (FWER controlled at 5% across 25 tests using the Bonferroni method).

Phenome-wide outcomes

I obtained summary statistics for 703 outcomes with significant heritability in the UKBB ($z > 4$) generated by the Neale lab[354]. As many of these phenotypes are correlated, I use the number of effectively independent phenotypes ($M_{eff} = 340$) estimated by the Neale lab[354] to establish significant r_G after multiple testing.

4.2.4 Colocalisation

All genetic colocalisation analyses were performed using the Bayesian framework implemented in the `coloc v5.1.0` package[436] under a single causal variant assumption[437], with summary statistics from European-ancestry GWASs. Only common variants (MAF > 1%) within windows of ±50 kb around each lead variant for an infertility or reproductive hormone trait were retained. For each pair of traits tested for colocalisation, we set the prior probability of variants in a locus being causally associated with trait 1 (p_1) and trait 2 (p_2) := 1×10^{-4}, and the prior for joint association $p_{12} := 1 \times 10^{-6}$, as recommended by the developers of `coloc`[437]. The priors for p_1 and p_2 indicate 99% confidence in a true association when each trait is associated with the locus at $P < 5 \times 10^{-8}$; the prior for p_{12} assumes that the probabilities of shared and non-shared causal variants for each trait are equally likely. I tested five hypotheses: H_0 = no association with either trait in region, H_1 = association with trait 1 in region, but not trait 2, H_2 = association with trait 2 in region, but not trait 1, H_3 = association with both traits in region, but different causal variants, and H_4 = association with both traits in region, and a shared causal variant.

Infertility and reproductive hormones

I tested for colocalisation between each female infertility category and each female-specific hormone (FSH, LH, oestradiol, and testosterone) at all genetic loci associated with at least one of the pair of traits tested. The single male infertility locus with common variants (MAF $> 1\%$) in the European-ancestry analysis did not contain enough significant associations (only 12 common variants with $P < 1 \times 10^{-6}$) for colocalisation analyses.

Female reproductive conditions

As we noticed that some lead variants for female infertility had previously been reported as associated with endometriosis and PCOS, we estimated the posterior probability (PP) of colocalisation of genetic signals between each category of female infertility and the two reproductive disorders. European-ancestry summary statistics for endometriosis and PCOS were obtained as described in Section 4.2.3 above.

eQTLs from the GTEx Project

The analyses in this section were performed by Dr. Laura Wittemans, a post-doctoral researcher in the Lindgren group. I produced the tables, figures, and text in this Chapter based on the results of these analyses.

We assessed colocalisation of genetic signals for female infertility with eQTLs for all genes whose transcription start sites (TSSs) are within 1Mb of an infertility lead variant. We downloaded publicly available eQTL data derived from 12 metabolic, endocrine, or reproductive tissues and cell types in the GTEx project[97]: Adipose Subcutaneous, Adipose Visceral Omentum, Adrenal Gland, Liver, Ovary, Pancreas, Pituitary, Prostate, Testis, Thyroid, Uterus, and Vagina.

4.2.5 Overlaps with genetic regions under selection

All selection scans described below were reported on the hg19 reference genome build, so genomic positions were lifted over to hg38 using the `UCSC liftOver` tool[432] to match the infertility and reproductive hormone summary statistics. To

avoid confounding by population stratification, I report results in this section from European-ancestry meta-analyses alone.

Directional selection

Following guidelines described by Mathieson *et al.* (2023)[260], I identified 54 genomic regions under directional selection from three previously reported genome-wide scans: (1) 39 regions from the Composite of Multiple Signals (CMS) test, which infers historical selection on the order of the past 50,000 years[438], (2) 12 regions from an ancient DNA scan that infers allele frequency from ancient genomes to determine selection over the past 10,000 years[439], and (3) three regions from Singleton Density Scores (SDSs) which use the pattern of singleton variants to identify recent selection in the past 2,000 to 3,000 years[440]. For each genomic window under directional selection, I report the variant with the lowest P-value associated with each category of infertility trait within the window (Tables C.4).

Singleton Density Scores

I downloaded publicly available SDSs for SNPs in the UK10K dataset[440] to report the highest (positive selection of derived allele over ancestral allele in the past 2,000 to 3,000 years) and lowest (negative selection) SDS within the ±10kb window around each infertility or hormone lead SNP. To calculate trait-SDS for each phenotype, I aligned each SDS to the trait-increasing allele rather than the derived allele[440]. For each lead variant window containing variants with extreme SDSs (top 97.5$^{\text{th}}$ or bottom 2.5$^{\text{th}}$ percentiles), I report the direction of selection with respect to the trait-increasing allele. Percentiles of SDSs were evaluated only on a subset of variants within ±10 kb of any variant reported in the GWAS Catalog to account for genomic context. Further, as variants that are sub-GWS for a trait may nonetheless be under selection, I evaluated genome-wide trait-SDS by calculating the mean trait-SDS in each bin of 1000 variants, ranked by P-value for the trait association, following the protocol outlined by Field *et al.* (2016)[440].

Balancing selection

I accessed publicly available standardised BetaScan2 scores, which detect balancing selection using polymorphism and substitution data, for all SNPs in the 1000 Genomes dataset[441]. For the ±10kb window around each infertility or hormone lead variant, I report the highest standardised BetaScan2 score and its percentile. Percentiles of BetaScan2 scores were evaluated only on a subset of variants within ±10 kb of any variant reported in the GWAS Catalog to account for genomic context.

4.2.6 Tissue and cell-type prioritisation

I estimated the polygenic contributions of genes with tissue-specific expression profiles to the heritability of infertility and hormones using stratified LD-score regression[62]. Briefly, this method assesses whether SNPs in high LD with a category of variants have higher GWAS χ^2 statistics than those in low LD with the category. Categories of variants may be stratified, for instance, by functional annotation or by vicinity to genes with cell-type specific expression profiles. Analyses for partitioned heritability were performed for the three traits with significant heritability in European-ancestry analyses ($z > 7$ for F-ALL, Testosterone-F, and Testosterone-M), restricted to 1 million HapMap3 variants[46] as recommended by Finucane *et al.* (2015)[442].

Gene sets and LD scores for 205 tissues and cell-types from the GTEx Project database[97] and the Franke lab single-cell database[443] were downloaded from Finucane *et al.* (2018)[444]. We established tissue-wide significance at $-\log_{10}(P) > 2.75$, which corresponds to false discovery rate (FDR) threshold $< 5\%$[444].

Ovarian cell types

The identification and QC of single-cell gene expression datasets for ovarian cell types was performed by Melody Parker, a research assistant in the Lindgren group. I wrote the scripts to produce LD scores and perform partitioned heritability analyses, and analysed their outputs.

We identified two publicly available single-cell gene expression datasets for ovarian cell types: (1) from Fan *et al.* (2019)[445] who performed single-cell RNA sequencing on ovarian tissue from five adult women undergoing fertility preservation procedures with 20,676 cells across 19 identified cell types, and (2) from Jin *et al.* (2022)[446] who performed single-nucleus RNA sequencing on autopsy samples from four women (aged 49 - 54 years, with normal ovarian histology) with 42,568 cells across 8 identified cell types. The datasets were aligned and filtered using the QC pipelines provided by the authors of each study, and clustered with identical parameters to replicate the results of each individual study. Gene sets for each cluster were identified as recommended by Finucane *et al.* (2018)[444] - briefly, we identified differential expression between the cells in each cluster and all other clusters by using the Wilcoxon rank sum test implemented in `Seurat v3.0`[447–449], and returned the top 10% of genes that are specifically expressed in each cluster (i.e. with positive average log-fold-change values), ranked by differential expression P-value.

I computed annotation-specific LD scores for these gene sets using hg38 coordinates for gene TSSs and transcription end sites (TESs) obtained from Ensembl[431], across 1 million HapMap3 variants[46] with LD information from European-ancestry individuals in the 1000 Genomes phase 3 dataset[280]. Partitioned heritability analyses were then performed using LD-score regression[62] as described above.

4.2.7 Whole exome sequencing analyses in UKBB

The QC and analyses performed in this section were conducted by various members of the Lindgren group, including Nikolas Baya, Barney Hill, Frederik Heymann Lassen, and Dr. Duncan Palmer. I wrote the scripts to produce the tables, figures, and text in this Chapter based on the results of these analyses.

We assessed the effects of rare variants and the burden of rare variants in a gene on binary traits with at least 500 cases, i.e. female infertility (F-ALL and F-INCL), male infertility (M-ALL), and reproductive hormones with at least 5,000 samples, i.e. FSH-F, LH-F, Oestradiol-F, Oestradiol-M, Testosterone-F, and Testosterone-M, using the whole-exome sequencing data available in the UKBB.

Sample sizes for each can be found in Tables 4.1 and 4.2. For a detailed description of variant annotation and exome-wide association tests, see Section 2.6 of Chapter 2.

I report Cauchy combined test P-values for gene-based associations, and burden test effect sizes and standard errors for the combination of variants in the gene (based on MAF thresholds and consequence annotation) with the lowest burden test P-value. Gene-wide significance was established at $P < 5 \times 10^{-6}$ (FWER controlled at 5% using the Bonferroni method, across 10,000 effectively independent genes). For variant-based tests where multiple variants in a gene were associated with the trait at exome-wide significance ($P < 1 \times 10^{-7}$, FWER controlled at 5% using the Bonferroni method, across 200,000 effectively independent variants after accounting for LD), I report results for the variant with the lowest P-value. I annotated genes from WES analyses as - (1) previously reported for the hormone of interest, (2) previously reported for any of 28 reproductive hormones, or (3) novel - based on the reported genes for SNP associations published in the GWAS Catalog as of 27 March 2023[50] (Table C.2).

To examine concordance between rare variant signals (from UKBB WES analyses, with MAF < 1%) and common variant signals (from GWAS meta-analyses reported in this Chapter, with MAF > 1%) in the same gene, I annotated common variant associations within transcribed gene regions based on hg38 coordinates for gene TSSs and TESs obtained from Ensembl[431].

4.3 Results

4.3.1 Genome-wide meta-analyses identify novel genetic loci for female and male infertility

Prevalence of infertility varies by cohort

We identified female infertility cases attributed to - all causes (F-ALL), anatomical causes (F-ANAT), anovulation (F-ANOV), idiopathic infertility (unknown causes) defined by exclusion of known causes (such as anatomical, anovulatory, PCOS, endometriosis, or uterine leiomyomas) (F-EXCL), or idiopathic infertility defined via inclusion of diagnostic codes for idiopathic infertility (F-INCL) - as well as male

infertility of all causes (M-ALL), through self-report, primary care, and secondary care EHR in six global cohorts (Table C.1). The prevalence of F-ALL ranges from 0.951% in the Icelandic deCODE population to 13.2% in the Copenhagen Hospital (Danish) Biobank, whereas the prevalence of male infertility is between 0.292% in the UK Biobank and 8.89% in the Danish Biobank. Anatomical infertility is the least prevalent cause of female infertility in three of six cohorts (prevalence in UKBB = 0.013%, EstBB = 2.06%, FinnGen = 0.786%). Five cohorts primarily represent individuals of European ancestry (totalling 40,024 cases / 665,658 controls for F-ALL and 6,885 cases / 410,578 controls for M-ALL), whereas the Genes and Health study population consists of British individuals of Bangladeshi and Pakistani descent[426] (1,176 cases / 21,347 controls for F-ALL and 213 cases / 17,786 controls for M-ALL) (Table 4.1).

Novel genetic loci associated with female and male infertility

I identified 19 unique loci associated with at least one category of female infertility, ten of which have not previously been reported to associate with a reproductive condition, and two loci associated with male infertility of all causes, both of which have not previously been reported, at a genome-wide significant threshold of $P < 5 \times 10^{-8}$ (Figures 4.1 and 4.2 and Table 4.4). Only five of the 21 unique loci have previously been reported in GWASs for infertility proxy variables such as childlessness[260,261] or age at first sexual intercourse[262,263].

Novel associations in this study include an intronic variant in *EBAG9*, rs1964514, which is associated with three categories of female infertility (OR (95% CI) for F-ALL = 1.13 (1.09 - 1.16), F-EXCL = 1.13 (1.09 - 1.17), F-INCL = 1.18 (1.11 - 1.25)). Transcripts of *EBAG9* are highly expressed in megakaryocytes[93,450], and knockout mice show abnormal T-cell physiology and decreased susceptibility to bacterial infection[451], supporting a potential role for *EBAG9* in maternal immunity during early pregnancy[452]. Another gene with a potential role in pregnancy is *AKAP12*, where an intronic variant rs6938404 is associated with F-ALL (0.958 (0.943 - 0.973)); the expression of *AKAP12* is highest in umbilical endothelial

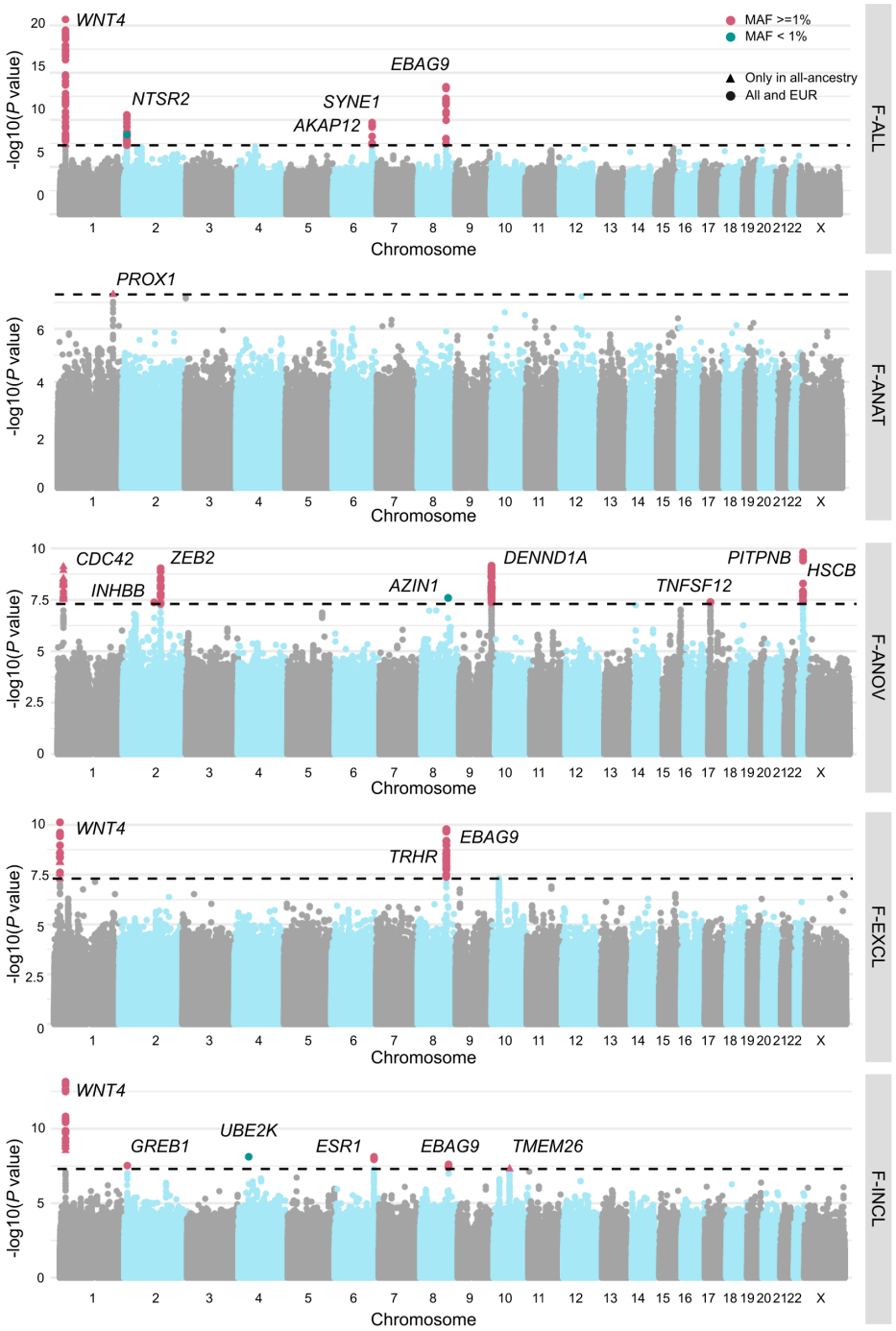


Figure 4.1: Manhattan plot depicting SNPs associated with all definitions of female infertility. Each panel displays a different category of female infertility (all cause = F-ALL, anatomical = F-ANAT, anovulatory = F-ANOV, idiopathic exclusion definition = F-EXCL, idiopathic inclusion definition = F-INCL). Each point depicts a single SNP, with genome-wide significant (GWS) SNPs ($P < 5 \times 10^{-8}$) coloured in pink for common variants with minor allele frequency (MAF) $\geq 1\%$ and green for those with MAF $< 1\%$; triangles represent SNPs that only reach GWS in the all-ancestry meta-analysis. SNPs are annotated with nearest gene.

Table 4.4: List of lead variants associated with all definitions of female and male infertility. CI = confidence interval, EUR = European, MAF = minor allele frequency, OR = odds ratio, TSS = transcription start site.

RSID	chr:pos (hg38)	Nearest TSS	All ancestries			EUR only		
			Average MAF	OR (95% CI)	P-value	Average MAF	OR (95% CI)	P-value
Female infertility of all causes (F-ALL)								
rs61768001	chr1:22139327	WNT4	0.166	0.909 (0.891 - 0.927)	2.25E-21	0.163	0.911 (0.893 - 0.93)	1.24E-19
rs10200851	chr2:11581956	NTSR2	0.458	0.951 (0.937 - 0.965)	2.90E-11	0.456	0.951 (0.936 - 0.965)	5.84E-11
rs6938404	chr6:151222906	AKAP12	0.453	0.958 (0.943 - 0.973)	3.88E-08	0.453	0.958 (0.943 - 0.973)	3.88E-08
rs17803970	chr6:152232583	SYNE1	0.0836	1.09 (1.06 - 1.12)	1.91E-10	0.0836	1.1 (1.07 - 1.13)	7.50E-11
rs1964514	chr8:109463457	EBAG9	0.0595	1.13 (1.09 - 1.16)	3.01E-14	0.0597	1.13 (1.09 - 1.16)	6.68E-14
Anatomical female infertility (F-ANAT)								
rs340879	chr1:213983171	PROX1	0.418	0.906 (0.874 - 0.939)	4.95E-08	0.418	0.902 (0.869 - 0.936)	5.06E-08
Anovulatory female infertility (F-ANOV)								
rs72665317	chr1:22040580	CDC42	0.19	0.875 (0.839 - 0.913)	7.76E-10	0.18	0.886 (0.847 - 0.927)	1.45E-07
rs72827480	chr2:120388925	INHBB	0.401	0.905 (0.873 - 0.938)	4.20E-08	0.401	0.905 (0.873 - 0.938)	4.20E-08
rs1852684	chr2:145068818	ZEB2	0.367	1.12 (1.08 - 1.16)	9.25E-10	0.35	1.12 (1.08 - 1.17)	3.44E-10
rs552953683	chr8:102898586	AZIN1	0.0024	0.341 (0.234 - 0.498)	2.54E-08	0.0024	0.341 (0.234 - 0.498)	2.54E-08
rs9696009	chr9:123856954	DENND1A	0.0777	1.21 (1.14 - 1.29)	6.87E-10	0.0695	1.24 (1.16 - 1.32)	2.40E-10
rs9902027	chr17:7537667	TNFSF12	0.255	0.895 (0.86 - 0.931)	4.06E-08	0.255	0.895 (0.86 - 0.931)	4.06E-08
rs143459581	chr22:28068862	PITPNB	0.0419	1.3 (1.19 - 1.43)	1.21E-08	0.0419	1.3 (1.19 - 1.43)	1.21E-08
rs17879961	chr22:28725099	HSCB	0.0389	0.739 (0.673 - 0.811)	1.55E-10	0.0389	0.739 (0.673 - 0.811)	1.55E-10
Idiopathic female infertility, exclusion definition (F-EXCL)								
rs61768001	chr1:22139327	WNT4	0.165	0.923 (0.902 - 0.946)	7.49E-11	0.162	0.928 (0.906 - 0.951)	2.48E-09
rs111597692	chr8:109039973	TRHR	0.0323	1.16 (1.1 - 1.22)	1.51E-08	0.0323	1.16 (1.1 - 1.22)	1.51E-08
rs17378154	chr8:109568721	EBAG9	0.059	1.13 (1.09 - 1.17)	1.64E-10	0.0593	1.13 (1.09 - 1.17)	3.36E-10
Idiopathic female infertility, inclusion definition (F-INCL)								
rs61768001	chr1:22139327	WNT4	0.17	0.87 (0.839 - 0.903)	6.87E-14	0.165	0.872 (0.84 - 0.905)	8.96E-13
rs11692588	chr2:11544358	GREB1	0.358	0.919 (0.892 - 0.947)	2.98E-08	0.358	0.919 (0.892 - 0.947)	2.98E-08
rs190290095	chr4:39786858	UBE2K	0.0022	0.227 (0.137 - 0.375)	7.60E-09	0.0022	0.227 (0.137 - 0.375)	7.60E-09
rs851982	chr6:151703850	ESR1	0.428	0.921 (0.895 - 0.947)	7.60E-09	0.437	0.922 (0.896 - 0.949)	2.86E-08
rs17378154	chr8:109568721	EBAG9	0.0565	1.18 (1.11 - 1.25)	2.47E-08	0.0569	1.18 (1.11 - 1.25)	4.97E-08
rs74156208	chr10:61509370	TMEM26	0.184	1.10 (1.06 - 1.14)	4.96E-08	0.187	1.10 (1.07 - 1.15)	5.44E-08
Male infertility of all causes (F-ALL)								
rs150639836	chr10:53879806	MBL2	0.0109	0.505 (0.402 - 0.636)	5.72E-09	0.0109	0.505 (0.402 - 0.636)	5.72E-09
rs139862664	chr10:116879589	HSPA12A	0.0072	0.388 (0.277 - 0.543)	3.29E-08	0.0072	0.388 (0.277 - 0.543)	3.29E-08

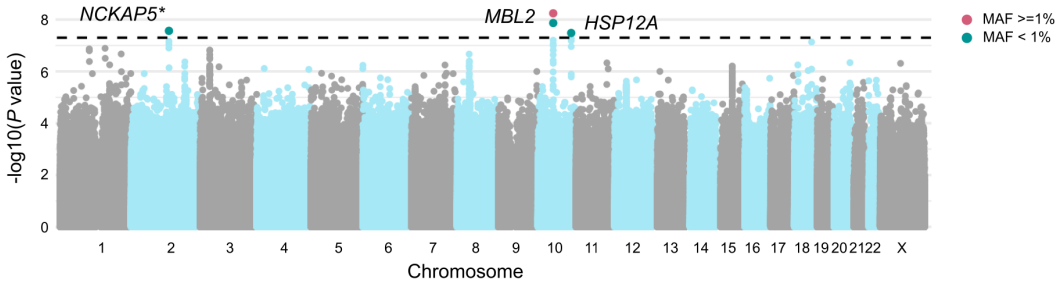


Figure 4.2: Manhattan plot depicting SNPs associated with male infertility of all causes. Each point depicts a single SNP, with genome-wide significant (GWS) SNPs ($P < 5 \times 10^{-8}$) coloured in pink for common variants with minor allele frequency (MAF) $\geq 1\%$ and green for those with MAF $< 1\%$. Loci are annotated with nearest gene. *lead variant is only present in one study.

cells[93,450], and knockouts are associated with delayed fertility or infertility in mice[453], and reduced litter size in goats[454]. Two signals may exert their effects on fertility through thyroid gland disruption - rs111597692 (nearest gene *TRHR*, encoding the thyrotropin hormone receptor)[455] associated with F-EXCL (OR = 1.16 (1.1 - 1.22)) and rs143459581 (nearest gene *PITPNB*[93,450], associated with thyroid cancer[129]), a signal for F-ANOV (OR = 1.3 (1.19 - 1.43)). Finally, rs139862664 associated with male infertility (OR = 0.388 (0.277 - 0.543)), is near *ENO4*, which is expressed in the testis and may play a role in sperm motility[456]. Male mice with *ENO4* knockouts display infertility, abnormal sperm morphology and physiology, and decreased testis weight, among other altered male reproductive tract phenotypes[457].

Prioritised genes at infertility loci

We assessed the genetic colocalisation of signals for female infertility with eQTLs for genes in metabolic, endocrine, and reproductive tissues from the GTEx Project[97]. All but one of the loci are associated with multiple eQTLs in the above tissues (colocalisation posterior probability (PP) $> 95\%$ for shared genetic loci with different causal variants) (Table C.3). The F-ANOV associated variant rs1852684 (OR = 1.12 (1.08 - 1.17), nearest gene *ZEB2*) colocalises with an adipose-eQTL for *GTDC1*, whose differential splicing is associated with human oocyte maturation deficiency[458]. I hypothesise that the putative causal gene for rs10200851 (associated with F-ALL, OR = 0.951 (0.936 - 0.965), nearest gene *NTSR2*) is *LPIN1*, an eQTL across

several metabolic and endocrine tissues, which is associated with age at menarche[459], ovarian cysts[460], and ovarian endometriosis[129] in women, and knockouts of which induce delayed female fertility in mice[461].

Genetic relationships between female infertility and reproductive conditions

Five lead variants for female infertility are in endometriosis-associated loci, including those in the vicinity of genes: *NTSR2 / LPIN1* (rs10200851 associated with F-ALL = 0.951 (0.937 - 0.965)), *SYNE1* (rs17803970 associated with F-ALL = 1.09 (1.06 - 1.12)), *CDC42* (rs72665317 associated with F-ANOV = 0.875 (0.839 - 0.913)), *ESR1* (rs851982 associated with F-INCL = 0.921 (0.895 - 0.947)), and *WNT4* (rs61768001 associated with three categories of female infertility (F-ALL = 0.909 (0.891 - 0.927), F-EXCL = 0.923 (0.902 - 0.946), F-INCL = 0.87 (0.839 - 0.903)) (Table 4.4). The latter locus is associated with several female reproductive phenotypes, including uterine fibroids[462,463], endometriosis[251,464], female genital prolapse[129], and bilateral oophorectomy[129], which reflects the role of *WNT4* as a key regulator of female reproductive tract development in embryogenesis[465–467].

The causal variant at the *WNT4* locus is likely to be shared by the three associated female infertility categories and endometriosis (colocalisation PP > 93.6%), as is the variant at the *ESR1* locus shared by F-INCL and endometriosis (PP > 94%) (Table 4.5). However, there is evidence that the other three loci with signals for both female infertility and endometriosis, *NTSR2 / LPIN1*, *SYNE1*, and *CDC42*, may be driven by different causal variants (PP > 76.5%). Additionally, I find that three loci for F-ANOV that have previously been reported to associate with PCOS (nearest genes *INHBB*, *PITPNB*, and *HSCB*) are likely to share the same causal variant for both disorders (PP > 89.2%), but three other F-ANOV loci that contain genetic signals for PCOS, i.e. nearest genes *ZEB2*, *DENND1A*, and *TNFSF12*, may have different causal variants for the two traits (Table 4.5).

Genome-wide, I observe positive genetic correlation between endometriosis and F-ALL ($r_G = 0.585$, $P = 8.98 \times 10^{-14}$) and F-INCL ($r_G = 0.710$, $P = 5.94 \times 10^{-10}$), suggesting that while there are shared genetic relationships between endometriosis

Table 4.5: Colocalisation of endometriosis and polycystic ovary syndrome (PCOS) signals form recent European-ancestry GWASs at lead variants identified in female infertility GWASs. Windows of ±50 kb around the EUR-ancestry lead variant are used to define loci for colocalisation; the RSID of the all-ancestry lead variant in the locus is noted if different from the EUR-ancestry variant. Under the assumption of a single causal variant at the locus, both-same-causal = both traits are associated and share a single causal variant, both-diff-causal = both traits are associated, but with different causal variants, only-infert = only infertility has a genetic association in the region.

RSID	chr:pos (hg38)	Endometriosis			PCOS		
		N SNPs	Most likely hypothesis (H)	Posterior probability of most likely H	N SNPs	Most likely hypothesis (H)	Posterior probability of most likely H
Female infertility of all causes (F-ALL)							
rs61768001	chr1:22089327-22189327	360	both_same_causal	96.3%	377	only_infert	96%
rs10165819 (EUR), rs10200851 (all-anc)	chr2:11531956-11631956	280	both_diff_causal	92.9%	307	only_infert	97.7%
rs6938404	chr6:151172906-151272906	299	only_infert	94.2%	320	only_infert	93.4%
rs17803970	chr6:152182583-152282583	372	both_diff_causal	93.3%	392	only_infert	92%
rs1964514	chr8:109413457-109513457	248	only_infert	99.2%	271	only_infert	98.3%
Anovulatory infertility (F-ANOV)							
rs72827480	chr2:120338925-120438925	351	only_infert	96.1%	367	both_same_causal	89.2%
rs1852684	chr2:145018818-145118818	243	only_infert	66%	255	both_diff_causal	85.3%
rs10818878 (EUR), rs9696009 (all-anc)	chr9:123786073-123886073	235	only_infert	98.2%	246	both_diff_causal	52.9%
rs12602989 (EUR), rs9902027 (all-anc)	chr17:7503172-7603172	281	only_infert	82.4%	316	both_diff_causal	58.5%
rs143459581	chr22:28018862-28118862	182	only_infert	27.3%	204	both_same_causal	92.1%
rs17879961	chr22:28675099-28775099	235	only_infert	75.9%	265	both_same_causal	99.7%
Idiopathic infertility, exclusion definition (F-EXCL)							
rs61768001	chr1:22089327-22189327	360	both_same_causal	95.5%	377	only_infert	96.1%
rs72676844 (EUR), rs111597692 (all-anc)	chr8:108908927-109008927	307	only_infert	95.3%	360	only_infert	96.3%
rs12549027 (EUR), rs17378154 (all-anc)	chr8:109425283-109525283	244	only_infert	99.1%	267	only_infert	98.4%
Idiopathic infertility, inclusion definition (F-INCL)							
rs61768001	chr1:22093958-22193958	364	both_same_causal	93.6%	379	only_infert	95.7%
rs11692588	chr2:11494358-11594358	205	both_diff_causal	76.5%	220	only_infert	97.3%
rs851982	chr6:151653850-151753850	318	both_same_causal	94%	345	only_infert	97.6%
rs17368310 (EUR), rs17378154 (all-anc)	chr8:109409837-109509837	254	only_infert	98.5%	259	only_infert	97.7%

and female infertility, they nonetheless have some distinct genetic aetiology (Figure 4.3C). Similarly, there is positive but modest genetic correlation between F-ANOV and PCOS ($r_G = 0.403$, $P = 2.20 \times 10^{-3}$), indicating substantial non-shared genetics of anovulatory infertility and PCOS (Figure 4.3C). Finally, we observe negative correlation ($r_G = -0.740$, $P = 4.93 \times 10^{-5}$) between anovulatory infertility and spontaneous dizygotic twinning, during which two oocytes are released and fertilised in a single menstrual cycle[435].

No evidence for genetic heterogeneity amongst participating cohorts

Despite considerable heterogeneity in diagnosis criteria and prevalence of infertility across cohorts, all lead variants from the infertility GWAS meta-analyses were present in at least two of the six cohorts studied (Table C.5). Among these are three variants for F-ANOV, two variants for F-INCL, and one variant for M-ALL that are shared only between FinnGen and EstBB (Table C.5); as the Finnish population underwent a recent bottleneck event and is genetically connected to Estonians, common variants in these two populations may not be represented in other European-ancestry cohorts[292]. Eighteen of 21 lead variants show no evidence for heterogeneity in effect estimates across cohorts (all $P_{het} > 0.05$), and the remaining have consistent effect directions in all of the studies in which they are present (Table C.5).

4.3.2 Selection pressure may explain the persistence of some infertility-associated variants in the population

The genome-wide heritability estimates (on the liability scale, accounting for disease prevalence[300]) for all categories of infertility are less than 10% (lowest for M-ALL at 1.12% and highest for F-ANOV at 9.54%) (Table 4.6). This is lower than heritability estimates of two-thirds of all heritable binary phenotypes in the UK Biobank with population prevalence similar to that of infertility (20/64 phenotypes with $Z > 4$ and prevalence $< 5\%$ have heritability estimates $< 10\%$)[354]. We tested whether there was evidence for: (i) variants associated with infertility in loci under historical or recent directional selection as identified by three genome-wide selection scans[438–440],

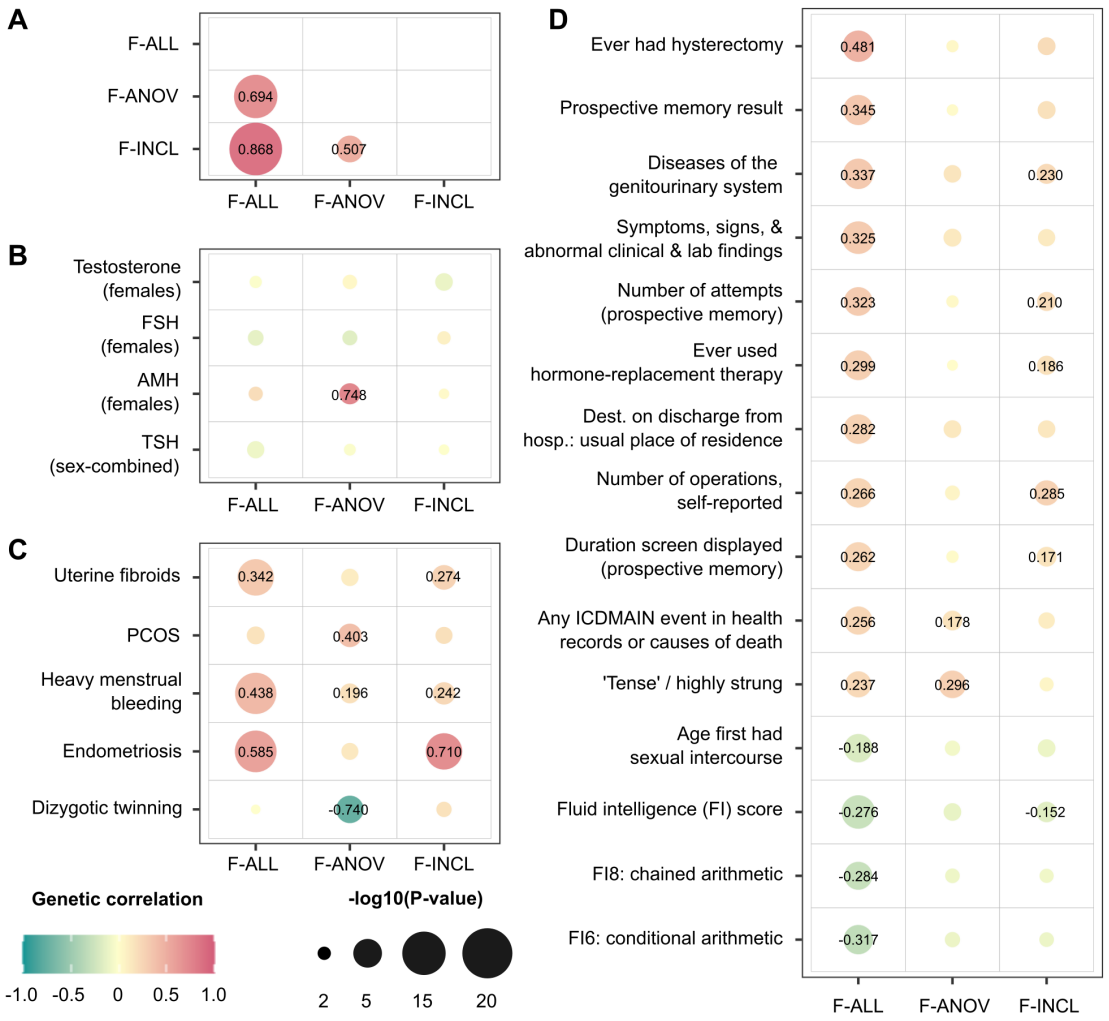


Figure 4.3: SNP-based genetic correlations (r_G) between female infertility and traits across the phenome. Only traits with significant SNP-based heritability ($z > 4$) as estimated by LD-score regression were retained for genetic correlation calculations, performed using the LDSC software[62] on a subset of 1 million HapMap3 SNPs[46]. Points are coloured by r_G estimate, sized by $-\log_{10}(P)$, and labelled with r_G estimate if nominally significant ($P < 0.05$). (A) r_G amongst three definitions of female infertility (all cause = F-ALL, anovulatory = F-ANOV, and idiopathic infertility defined by inclusion = F-INCL). (B) r_G between female infertility traits and reproductive hormones: testosterone, follicle stimulating hormone (FSH), and anti-Müllerian hormone (AMH) in female-specific analyses, and thyroid stimulating hormone (TSH) from sex-combined analysis. (C) r_G between female infertility traits and female reproductive conditions, with summary statistics generated from the largest available European-ancestry studies for each trait (see Methods). PCOS = polycystic ovary syndrome. (D) r_G between female infertility traits and all heritable phenotypes ($z > 4$) in the UK Biobank as generated by the Neale lab[354]. Only phenotypes significantly associated with at least one female infertility trait at $P < 4.90 \times 10^{-5}$ (FWER controlled at 5% across 1,020 tests using the Bonferroni method, accounting for 340 effectively independent UKBB phenotypes and 3 female infertility traits) are displayed.

Table 4.6: SNP-based heritability estimates for infertility and reproductive hormones.

Trait	Sex	Observed-scale heritability % (SE)	Liability-scale heritability %
Infertility			
Female infertility all cause (F-ALL)	F	0.66% (0.08)	2.32% (0.28)
Anatomical infertility (F-ANAT)	F	0.22% (0.08)	1.56% (0.57)
Anovulatory infertility (F-ANOV)	F	0.75% (0.17)	9.54% (2.16)
Idiopathic female infertility by exclusion (F-EXCL)	F	0.22% (0.06)	1.15% (0.31)
Idiopathic female infertility by inclusion (F-INCL)	F	0.48% (0.1)	3.43% (0.71)
Male infertility all cause (M-ALL)	M	0.12% (0.1)	1.12% (0.93)
Reproductive hormones			
Follicle stimulating hormone (FSH)	F	4.59% (1.06)	
	M	14.88% (7.49)	
Luteinising hormone (LH)	F	5.27% (1.33)	
	M	5.36% (6.18)	
Oestradiol	F	1.06% (0.51)	
	M	2.78% (1.51)	
Progesterone	F	-0.7% (2.35)	
Testosterone	F	9.87% (1)	
	M	15.67% (1.79)	

or (ii) recent directional selection (over the last 2,000 to 3,000 years) measured by SDSs[440] and balancing selection measured by BetaScan2 scores[441] at infertility loci.

At the *EBAG9* locus associated with F-ALL, F-EXCL, and F-INCL, we observe extreme SDSs (in the lowest 0.25$^{\text{th}}$ percentile and highest 99.75$^{\text{th}}$ percentile) for several variants, indicating that this infertility-associated genomic region may be under recent directional selection (Table 4.7). When the SDSs are aligned to the infertility-risk increasing alleles, I note that 32 of 33 variants display positive selection of the infertility risk allele, and variants with high probabilities of association with infertility also have high scores for directional selection (Figure 4.4). The *EBAG9* gene is associated with infectious response phenotypes[451,452], suggesting that the locus may be under selection for its effects on the immune system.

I additionally observe signatures of balancing selection, which maintains multiple alleles in the population through mechanisms such as heterozygote advantage or time-varying fitness[468,469], at the *NTSR2 / LPIN1* locus for F-ALL (standardised BetaScan2 (StdB2) scores in the 99.6$^{\text{th}}$ percentile) (Table 4.7 and Figure 4.5). I also report high scores for balancing selection (StdB2 $>$ 99$^{\text{th}}$ percentile) at the *INHBB* locus for F-ANOV, the *GREB1* locus for F-INCL, and the *MBL2* locus for M-ALL (Table 4.7 and Figure 4.5). However, there appears to be no relationship

Table 4.7: List of infertility loci (European-ancestry, common variants) annotated with minimum and maximum Singleton Density Score (for recent directional selection; score > 0 indicates positive selection of derived allele) and maximum Standardised BetaScan2 score (for balancing selection). N SNPs indicates the number of variants in the 10kb window around the lead SNP for which selection scores are available. Variant score percentiles were calculated against a background of all variants within ±10 kb of variants reported in the GWAS Catalog.

Trait	Lead variant	chr:pos (hg38) of 10kb window around lead variant	Singleton Density Score			Standardised BetaScan2 Score	
			N SNPs	Min (%le)	Max (%le)	N SNPs	Max (%le)
F-ALL, F-EXCL, F-INCL	rs61768001	chr1:22129327-22149327	26	-0.407 (35%)	1.19 (88.5%)	76	2.23 (87.1%)
F-ALL	rs10165819	chr2:11571976-11591976	39	-1.35 (9.18%)	1.56 (94%)	91	7.55 (99.6%)
F-ALL	rs6938404	chr6:151212906-151232906	29	-1.34 (9.35%)	1.26 (89.9%)	78	1.27 (76.6%)
F-ALL	rs17803970	chr6:152222583-152242583	53	-1.27 (10.6%)	1.17 (88.1%)	103	3.34 (93.2%)
F-ALL, F-EXCL, F-INCL	rs1964514	chr8:109453457-109473457	33	-3.12 (0.103%)	2.94 (99.8%)	80	2.92 (91.4%)
F-ANOV	rs72827480	chr2:120378925-120398925	48	-1.82 (3.67%)	1.71 (95.5%)	120	6.45 (99.2%)
F-ANOV	rs1852684	chr2:145058818-145078818	40	-1.08 (14.6%)	0.95 (83.5%)	87	3.68 (94.4%)
F-ANOV	rs10818878	chr9:123826073-123846073	11	-1.35 (9.3%)	1.78 (96%)	36	2.99 (91.7%)
F-ANOV	rs12602989	chr17:7543172-7563172	43	-1.54 (6.65%)	1.38 (91.8%)	95	2.43 (88.5%)
F-ANOV	rs143459581	chr22:28058862-28078862	27	-1.86 (3.32%)	1.66 (95.1%)	54	5.14 (97.9%)
F-ANOV	rs17879961	chr22:28715099-28735099	15	-1.11 (13.7%)	2.26 (98.7%)	65	2.07 (85.8%)
F-EXCL	rs72676844	chr8:108948927-108968927	31	-1.96 (2.67%)	3.81 (100%)	98	4.86 (97.4%)
F-INCL	rs55938609	chr1:22133958-22153958	27	-1.14 (12.9%)	1.62 (94.7%)	76	2.35 (88%)
F-INCL	rs11692588	chr2:11534358-11554358	51	-1.98 (2.59%)	1.74 (95.7%)	110	6.25 (99%)
F-INCL	rs851982	chr6:151693850-151713850	18	-0.395 (35.5%)	0.567 (72.4%)	65	2.32 (87.8%)
M-ALL	rs150639836	chr10:53869806-53889806	53	-2.67 (0.502%)	1.41 (92.1%)	93	6.29 (99.1%)

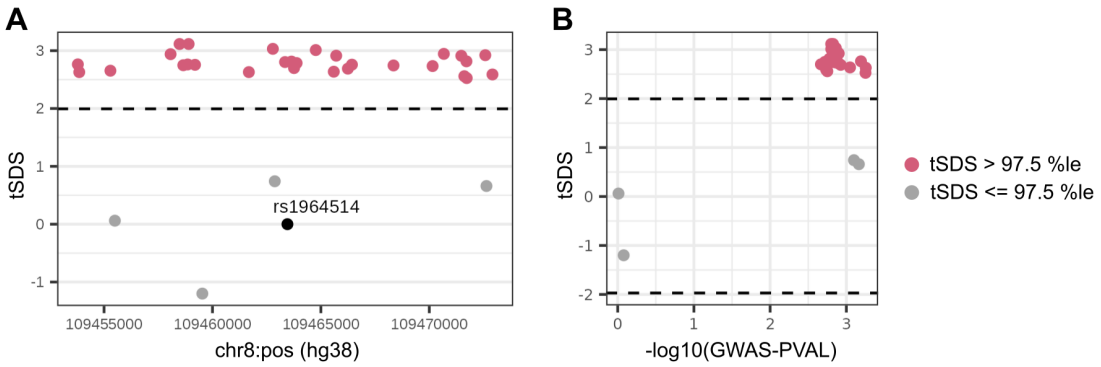


Figure 4.4: Trait-aligned Singleton Density Scores (tSDSs) at the *EBAG9* locus. Variants in the window of ±10kb around the lead female infertility-associated (F-ALL) variant are displayed. The tSDSs are aligned to the infertility-risk increasing allele, wherein a positive tSDS indicates positive selection for infertility-risk increasing allele at the locus. In both plots, dashed lines indicate 2.5$^{\text{th}}$ percentile (%le) and 97.5$^{\text{th}}$ %le of SDS, and variants below or above this threshold respectively are coloured in pink. (A) Locus plot depicting genomic position on the x-axis and tSDS on the y-axis. The lead variant rs1964514 is not present in the tSDS dataset and thus assigned tSDS of 0. (B) Scatter plot depicting relationship between GWAS p-value (-log10) for the variant association with F-ALL on the x-axis and tSDS on the y-axis.

between the balancing selection score for a variant and probability of association with infertility, as measured by $-log10(P)$ from GWAS , at any of these loci. These are all gene-dense regions with pleiotropic effects ranging from immune response[470,471] to metabolism[472,473], which may explain the signals of balancing selection and persistence of infertility risk alleles in these genomic loci.

There is no global signature of directional selection against infertility when SDSs are aligned to the infertility-risk increasing allele, i.e. the mean genome-wide trait-SDS is not significantly different from 0 for any category of infertility (all $P > 0.05$). I also did not observe any GWS variants for infertility in 54 genomic regions under historical (50,000 years or 10,000 years) or recent (2,000 to 3,000 years) directional selection (Table C.4)

4.3.3 Genetic determinants of reproductive hormone levels

Classification of novel hormone-associated genetic loci

As hormone dysregulation is hypothesised to play a substantial role in the aetiology of infertility[414,415], I conducted sex-specific meta-analyses of five reproductive hormones

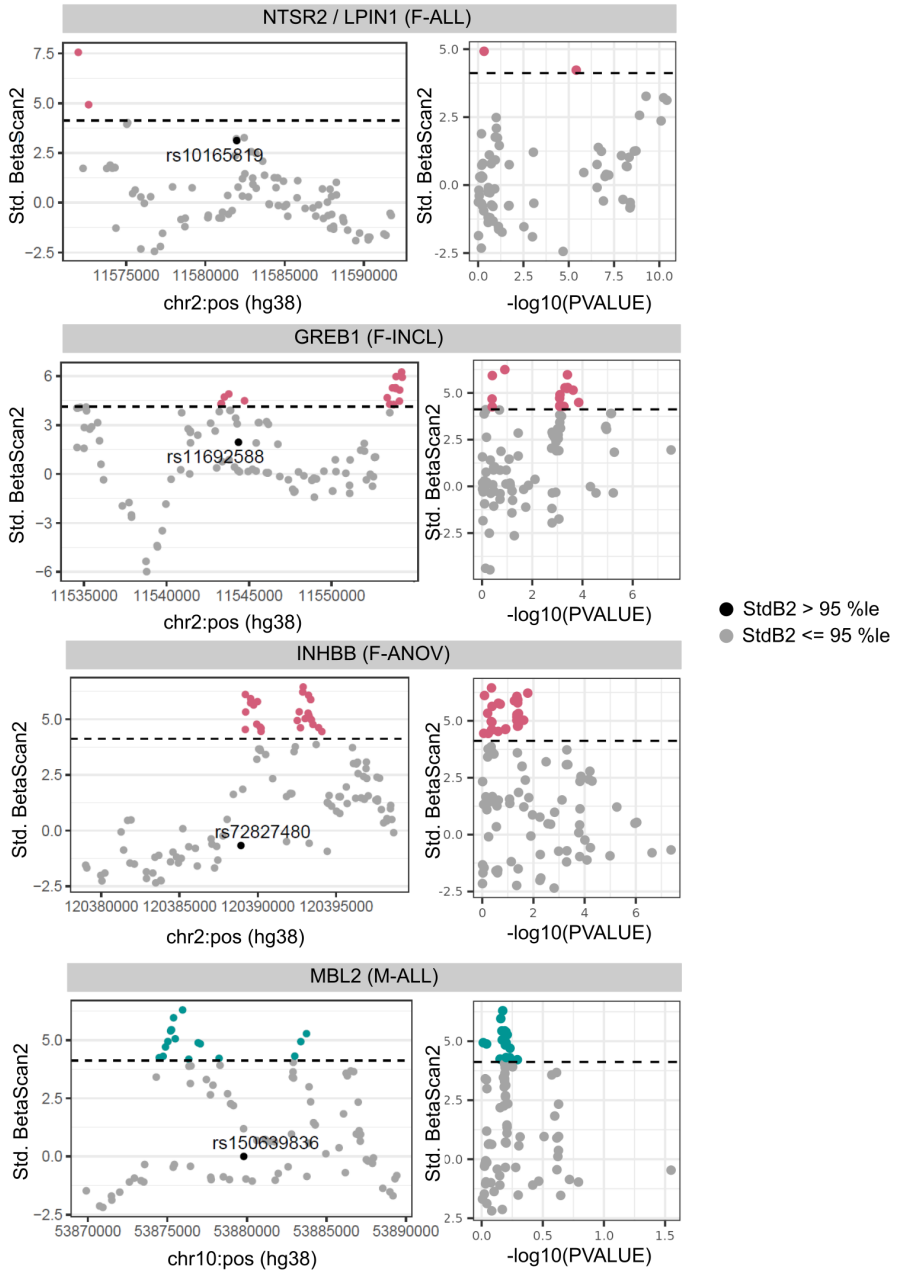


Figure 4.5: Standardised BetaScan2 scores (StdB2) for balancing selection at infertility loci. Each panel displays windows of ±10kb around a lead infertility-associated variant, annotated with nearest gene: rs10165819 (female infertility of all causes, F-ALL), rs11692588 (female idiopathic infertility by inclusion, F-INCL), rs72827480 (anovulatory infertility, F-ANOV), and rs150639836 (male infertility of all causes, M-ALL). TDashed lines indicate 95th percentile of StdB2, and variants crossing this threshold are coloured in pink (for female infertility loci) or green (for the male infertility locus). Left: Locus plot depicting genomic position on the x-axis and StdB2 on the y-axis. he lead variant rs150639836 is not present in the StdB2 dataset and thus assigned StdB2 of 0. Right: Scatter plot depicting relationship between GWAS p-value (-log10) for the variant association with infertility on the x-axis and StdB2 on the y-axis.

- FSH ($N_{\text{female}} = 57,890$, $N_{\text{male}} = 6,095$), LH ($N_{\text{female}} = 47,986$, $N_{\text{male}} = 6,769$), oestradiol ($N_{\text{female}} = 97,887$, $N_{\text{male}} = 39,165$), progesterone ($N_{\text{female}} = 18,368$), and testosterone ($N_{\text{female}} = 246,862$, $N_{\text{male}} = 243,951$) - collected at assessment centre visits or identified through primary care and secondary care EHRs, in six global cohorts and publicly available summary statistics (Tables 4.2 and 4.3). At a GWS threshold of $P < 5 \times 10^{-8}$, I identified two independent genetic loci associated with oestradiol in females (Oestradiol-F), six with LH-F, 11 with FSH-F, and 153 with Testosterone-F. Additionally, I find one locus each for FSH and LH in males (FSH-M and LH-M), six loci for Oestradiol-M, and 261 loci for Testosterone-M (Figures 4.6 and 4.7). We replicate known signals near genes encoding the hormones themselves, such as rs11031005 (nearest gene *FSHB*, which encodes the FSH beta subunit) associated with FSH-F (β (SE) = -0.104 (0.0092), $P = 1.08 \times 10^{-29}$) and FSH-M (β = -0.206 (0.0263), $P = 3.75 \times 10^{-15}$), and rs753307 (nearest gene *LHB*, which encodes the LH beta subunit) associated with LH-F (β = 0.0605 (0.0072), $P = 4.53 \times 10^{-17}$). We also replicate associations near genes encoding enzymes for steroid hormone metabolism, such as rs45446698 (nearest gene *CYP3A7*, whose encoded enzyme metabolises a precursor of oestrogen) associated with Oestradiol-F (β = -0.0782 (0.0114), $P = 8.26 \times 10^{-12}$), and rs13133311 (nearest gene *HSD17B13*, whose encoded protein catalyses the oxidation of oestradiol[474]) associated with Testosterone-M (β = -0.0307 (0.00270), $P = 2.84 \times 10^{-29}$).

I defined a set of published hormone-associated variants, based on 28 reproductive hormones in the GWAS Catalog[50], to identify novel hormone variants that are not in LD ($r^2 < 0.1$) with, and statistically independent of (conditional analysis P-value $P_{\text{cond}} < 0.05$) any published variants within a 1 Mb window of lead variants in our study; I also identify variants novel to each hormone that have only been reported previously for another of the 28 reproductive hormones; all other variants are classified as previously reported (see Methods). Across men and women, we replicate two previously published loci for FSH, four for LH, five for oestradiol, and 323 loci for testosterone (Figure 4.8).

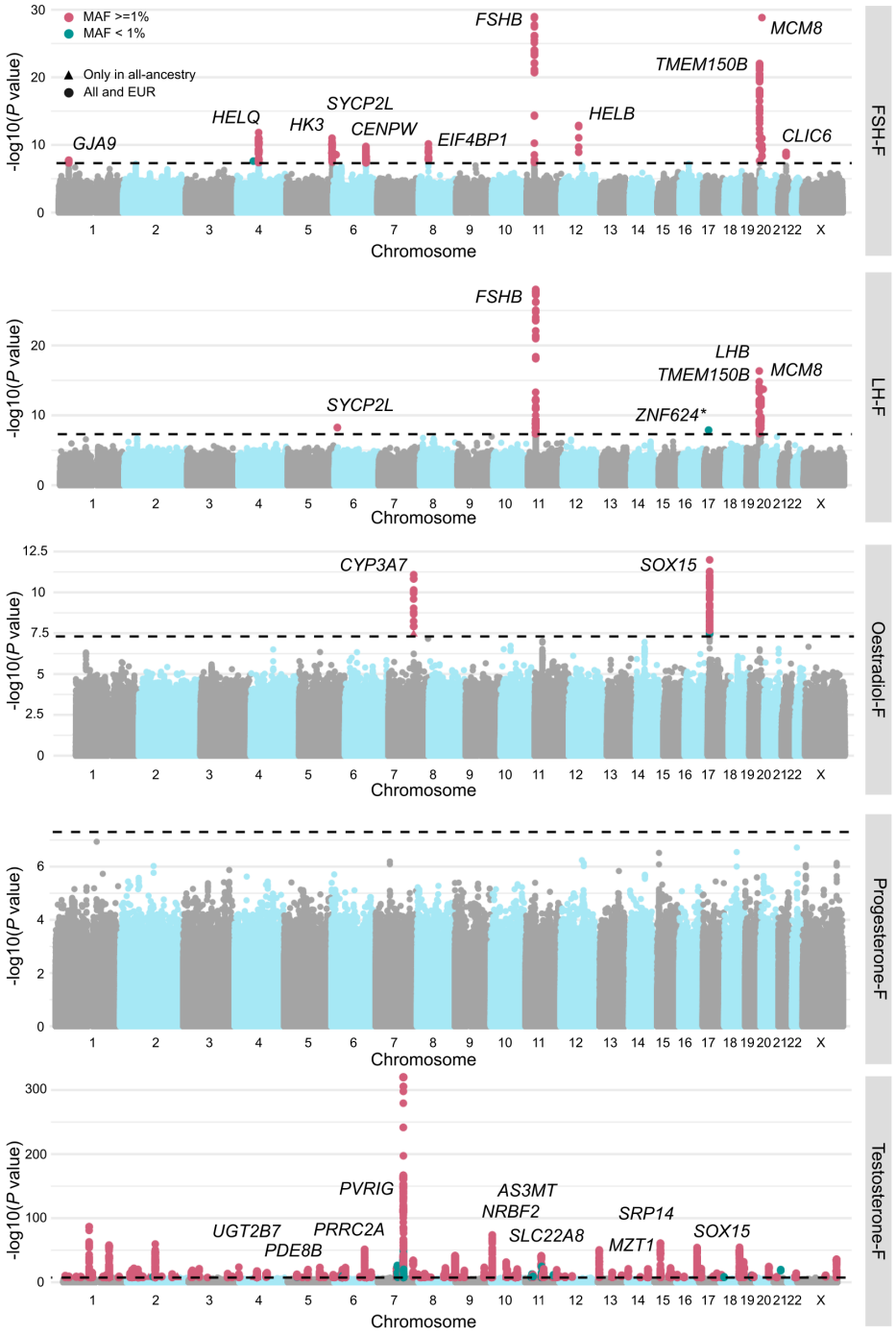


Figure 4.6: Manhattan plot depicting SNPs associated with reproductive hormones in female-specific analyses. Each panel displays a different reproductive hormone (FSH = follicle stimulating hormone, LH = luteinising hormone). Each point depicts a single SNP, with genome-wide significant (GWS) SNPs ($P < 5 \times 10^{-8}$) coloured in pink for common variants with minor allele frequency (MAF) $\geq 1\%$ and green for those with MAF $< 1\%$; triangles represent SNPs that only reach GWS in the all-ancestry meta-analysis. All loci are annotated with nearest gene; for testosterone, only the ten most significant loci with novel lead variants are annotated. *lead variant is only present in one study.

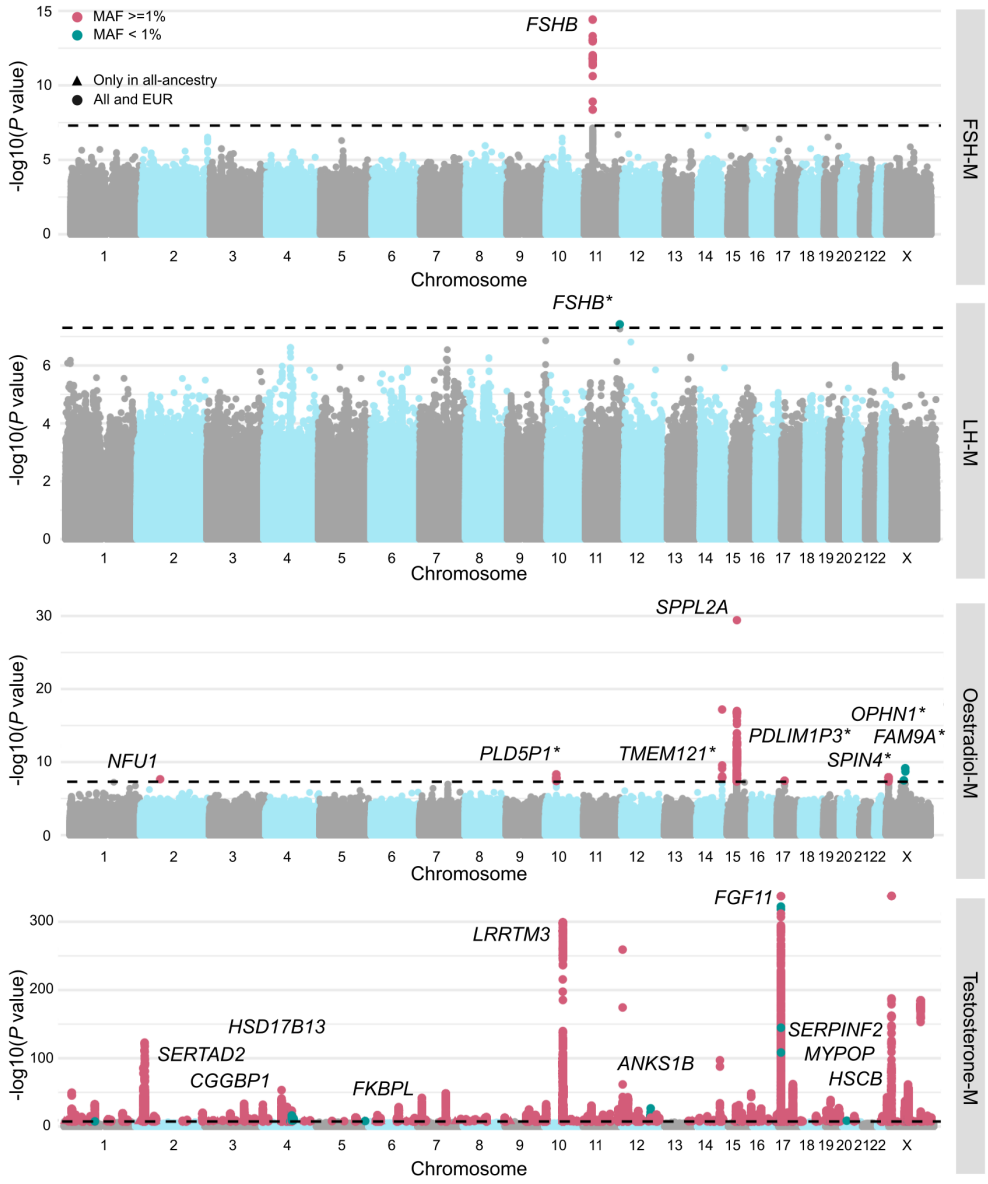


Figure 4.7: Manhattan plot depicting SNPs associated with reproductive hormones in male-specific analyses. Each panel displays a different reproductive hormone (FSH = follicle stimulating hormone, LH = luteinising hormone). Each point depicts a single SNP, with genome-wide significant (GWS) SNPs ($P < 5 \times 10^{-8}$) coloured in pink for common variants with minor allele frequency (MAF) $\geq$ 1% and green for those with MAF $<$ 1%; triangles represent SNPs that only reach GWS in the all-ancestry meta-analysis. All loci are annotated with nearest gene; for testosterone, only the ten most significant loci with novel lead variants are annotated. *lead variants are only present in one study.

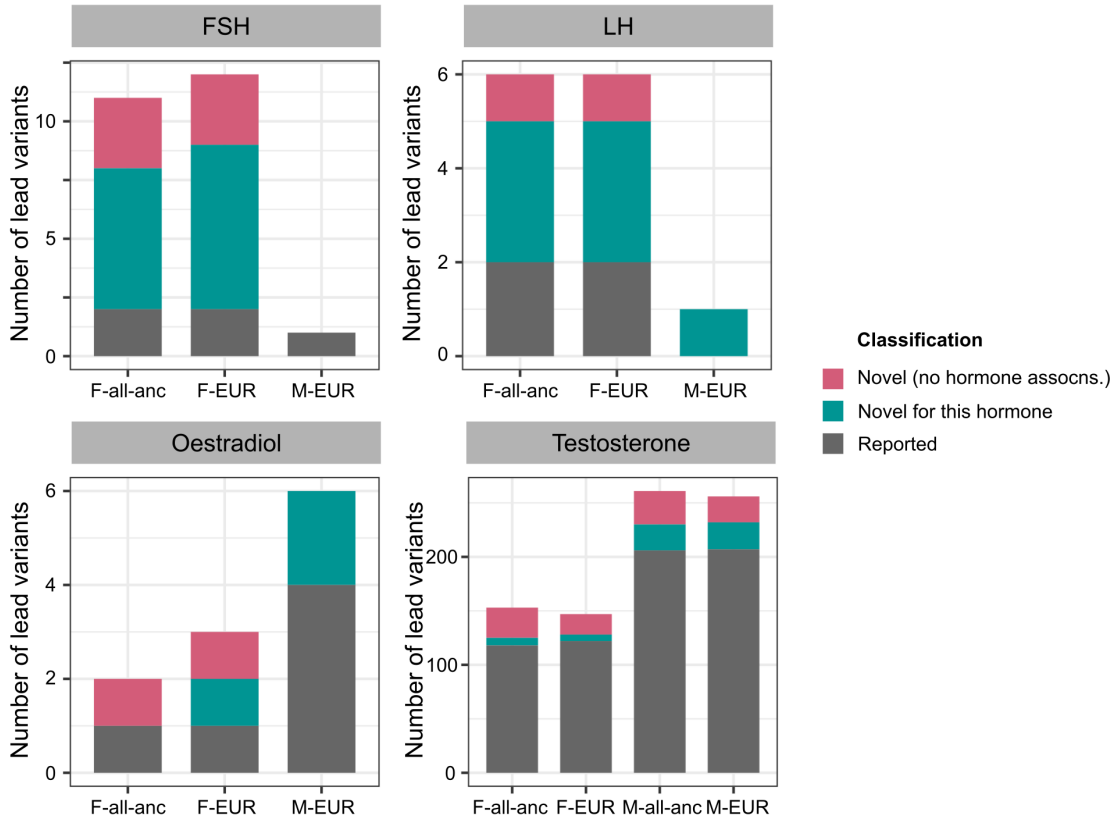


Figure 4.8: Number of hormone-associated lead variants classified as novel or reported. Each panel displays a different hormone (FSH = follicle stimulating hormone, LH = luteinising hormone). Lead variants in each analysis stratum (F = female-specific, M = male-specific, all-anc = all ancestries included in meta-analysis, EUR = European-ancestry meta-analysis) are classified as: (1) novel (no hormone associations) if they are not in LD ($r^2 < 0.1$) with, and conditionally independent of (conditional P-value $P_{\text{cond}} < 0.05$), any variants within a 1 Mb window of the lead variant that are associated with 28 reproductive hormones in the GWAS Catalog[50], plotted in pink, (2) novel for this hormone if they are not in LD ($r^2 < 0.1$) with, and conditionally independent of (conditional P-value $P_{\text{cond}} < 0.05$), the respective hormone-associated variants within a 1 Mb window of the lead variant, plotted in green, and (3) reported otherwise, plotted in grey. assocns. = associations

I find 39 novel variants for testosterone in men, including near genes *SPOCK1* (rs1073917: β = -0.0160 (0.0029), $P = 4.69 \times 10^{-8}$), which is a target for the androgen receptor[475], *NR4A3* (rs10988865: β = -0.0161 (0.0029), $P = 4.33 \times 10^{-8}$), which coordinates cellular response to corticotropin-hormone and thyrotropin-hormone releasing stimulus[476,477] and regulates adipogenesis[478], and obesity-associated genes *ANKS1B* (rs144998814: $\beta = 0.133$ (0.0162), $P = 2.34 \times 10^{-16}$) and *ANO10* (rs6809522: $\beta = 0.016$ (0.0029), $P = 3.00 \times 10^{-8}$)[216] (Table 4.8). The 28 novel

reproductive hormone variants associated with testosterone in women include those near *LAMTOR4* (rs17250196: β = -0.131 (0.0067), $P = 4.02 \times 10^{-86}$), associated with hyperthyroidism[479] and age at menarche and menopause[459], obesity-associated *CCDC146* (rs138240474: β = -0.116 (0.0207), $P = 2.03 \times 10^{-8}$)[216], which is also expressed in the fallopian tubes and endometrium[93,450], and *SLC8A1* (rs12611602: $\beta = 0.0163$ (0.003), $P = 3.79 \times 10^{-8}$), which causes increased pancreatic beta cell proliferation and insulin secretion in knockout mouse models[480]. Finally, I report lead variants independent of previously published hormone SNPs in the *HELQ* locus for FSH-F (rs4235062: β = -0.046 (0.0065), $P = 1.50 \times 10^{-12}$), *TMEM150B* locus for FSH-F (rs28875253: β = -0.0599 (0.0061), $P = 9.90 \times 10^{-23}$) and LH-F (rs11668309: $\beta = 0.0519$ (0.0071), $P = 3.91 \times 10{-}13$), and in the *SOX15-SAT2* locus[236,246,430] for Oestradiol-F (rs3933469: $\beta = 0.0363$ (0.0051), $P = 1.02 \times 10^{-12}$) (Table 4.8).

Female reproductive hormone levels vary with menopause status, and are even used to diagnose menopause in some health systems[153], and may therefore affect the genetic studies of hormone levels reported here. While we were able to adjust for age and age^2, which is expected to account for some of the effect of menopause, we were not able to directly account for menopause in our analyses as not all of the cohorts included in this study reported data on menopause status. Even within cohorts such as the UK Biobank where age at menopause could be ascertained, over 30% of women with reproductive health conditions reported being unsure of their menopause status (see Chapter 5 for details); we therefore chose not to adjust hormone levels in any cohort for menopause.

Genetic variants associated with testosterone are sex-specific and some may be under selection

Of the 153 independent loci associated with testosterone in female-specific GWAS and 261 in male-specific GWAS , only 15 (i.e. 9.8% of Testosterone-F loci and 5.75% of Testosterone-M loci) are shared between the sexes. 39.7% of the unique lead variants identified across both sexes have sex-differential effects (sex heterogeneity $P < 0.05/399$), and 45.9% have opposing directions of effect in males and females (Figure 4.9). Genome-wide, the SNP-based heritability of testosterone is 1.5x higher

Table 4.8: List of novel lead variants associated with reproductive hormones. CI = confidence interval, EUR = European, MAF = minor allele frequency, OR = odds ratio, TSS = transcription start site.

RSID	chr:pos (hg38)	Nearest TSS	All ancestries			EUR only		
			Average MAF	Beta (SE)	P-value	Average MAF	Beta (SE)	P-value
Follicle stimulating hormone (females)								
rs4235062	chr4:83430000	HELQ	0.475	-0.046 (0.0065)	1.50E-12	0.475	-0.046 (0.0065)	1.50E-12
rs28875253	chr19:55317846	TMEM150B	0.409	-0.0599 (0.0061)	9.09E-23	0.401	-0.0656 (0.0066)	2.36E-23
rs2834747	chr21:35072824	CLIC6	0.293	-0.0432 (0.0071)	1.27E-09	0.293	-0.0432 (0.0071)	1.27E-09
Luteinising hormone (females)								
rs11668309	chr19:55322092	TMEM150B	0.377	0.0519 (0.0071)	3.91E-13	0.364	0.0577 (0.0078)	1.68E-13
Oestradiol (females)								
rs3933469	chr17:7596184	SOX15	0.271	0.0363 (0.0051)	1.02E-12	0.27	0.0354 (0.0052)	1.18E-11

contd. List of novel lead variants associated with reproductive hormones.

RSID	chr:pos (hg38)	Nearest TSS	All ancestries			EUR only		
			Average MAF	Beta (SE)	P-value	Average MAF	Beta (SE)	P-value
Testosterone (females)								
rs4076000	chr1:68371486	RPE65	0.408	0.0177 (0.0029)	1.54E-09	0.395	-0.0185 (0.003)	1.01E-09
rs3790611	chr1:112525170	ST7L	0.256	0.0265 (0.0033)	7.44E-16	0.26	0.0261 (0.0034)	7.87E-15
rs12611602	chr2:40349985	SLC8A1	0.448	0.0163 (0.003)	3.79E-08	0.445	0.0158 (0.003)	1.21E-07
rs2349101	chr2:48695564	LHCGR	0.423	0.0217 (0.003)	6.10E-13	0.424	0.0217 (0.003)	5.38E-13
rs11683263	chr2:62311546	B3GNT2	0.485	0.0185 (0.003)	4.79E-10	0.485	0.0185 (0.003)	4.79E-10
rs2002155	chr4:56879874	REST	0.469	-0.0162 (0.003)	4.92E-08	0.466	-0.0158 (0.003)	1.89E-07
rs6422324	chr4:69087385	UGT2B7	0.46	0.026 (0.003)	1.04E-17	0.453	-0.0263 (0.0031)	1.65E-17
rs58282071	chr5:35267641	PRLR	0.0483	0.0492 (0.0068)	5.30E-13	0.0484	0.0518 (0.0069)	8.89E-14
rs6453294	chr5:77192830	PDE8B	0.35	0.0272 (0.0031)	8.93E-19	0.351	0.0266 (0.0031)	7.91E-18
rs7766042	chr6:1366483	FOXF2	0.106	-0.0312 (0.0048)	1.00E-10	0.107	-0.0314 (0.0048)	8.92E-11
rs9348696	chr6:25781710	SLC17A4	0.296	0.0223 (0.0032)	2.16E-12	0.291	0.0231 (0.0033)	1.33E-12
rs116325030	chr6:31046656	HCG22	0.0134	0.0941 (0.0131)	6.09E-13	0.0123	0.101 (0.0134)	6.15E-14
rs2261033	chr6:31635814	PRRC2A	0.431	-0.0268 (0.0029)	1.65E-20	0.434	-0.0267 (0.003)	3.15E-19
rs2284828	chr6:33580716	BAK1	0.32	0.0173 (0.0031)	3.09E-08	0.325	0.0176 (0.0032)	2.69E-08
rs4568468	chr6:126863321	RSPO3	0.432	0.0199 (0.003)	2.25E-11	0.382	0.0202 (0.003)	3.47E-11
rs546825690	chr7:76545933	SPDYE16	0.0075	0.118 (0.0199)	3.53E-09	0.0075	0.118 (0.0199)	3.53E-09
rs138240474	chr7:77108986	CCDC146	0.0063	-0.116 (0.0207)	2.03E-08	0.0063	-0.116 (0.0207)	2.03E-08
rs17250196	chr7:100219573	PVRIG	0.0513	-0.131 (0.0067)	4.02E-86	0.0519	-0.132 (0.0067)	6.80E-86
rs2280861	chr8:23547272	SLC25A37	0.255	-0.0253 (0.0033)	3.29E-14	0.243	-0.0249 (0.0034)	5.03E-13
rs1660319	chr8:100325302	RNF19A	0.396	0.021 (0.003)	1.56E-12	0.39	-0.0213 (0.003)	1.94E-12
rs2242271	chr10:997529	GTPBP4	0.491	0.0205 (0.0029)	2.01E-12	0.489	0.0207 (0.003)	2.54E-12
rs10822145	chr10:63174788	NRBF2	0.481	0.0264 (0.0029)	3.62E-20	0.482	0.0267 (0.003)	1.50E-19
rs3740393	chr10:102876898	AS3MT	0.146	0.0387 (0.0041)	2.12E-21	0.14	0.0394 (0.0042)	1.81E-20
rs377204392	chr11:63075350	SLC22A8	0.0679	0.0797 (0.0059)	4.02E-42	0.0685	0.0794 (0.0059)	2.39E-41
rs7130050	chr11:110316369	RDX	0.231	0.0203 (0.0035)	5.68E-09	0.24	-0.0193 (0.0035)	2.20E-08
rs1337985	chr13:72560774	MZT1	0.373	-0.0248 (0.003)	3.14E-16	0.375	-0.0242 (0.0031)	2.90E-15
rs28587891	chr15:40072905	SRP14	0.353	0.0507 (0.0031)	1.29E-61	0.354	0.0511 (0.0031)	1.56E-61
rs3933469	chr17:7596184	SOX15	0.269	0.0315 (0.0033)	3.56E-22	0.269	0.0314 (0.0033)	5.27E-21

contd. List of novel lead variants associated with reproductive hormones.

RSID	chr:pos (hg38)	Nearest TSS	All ancestries			EUR only		
			Average MAF	Beta (SE)	P-value	Average MAF	Beta (SE)	P-value
Testosterone (males)								
rs1211184	chr1:35422546	NCDN	0.124	0.0232 (0.0041)	1.36E-08	0.0983	0.0248 (0.0043)	7.24E-09
rs1223788	chr1:214134645	SMYD2	0.178	0.0244 (0.0037)	4.33E-11	0.179	0.0252 (0.0037)	1.33E-11
rs62140396	chr2:28023974	RBKS	0.117	-0.0276 (0.0039)	1.95E-12	0.117	-0.0277 (0.0039)	1.94E-12
rs34642907	chr2:64668597	SERTAD2	0.207	-0.0274 (0.0034)	3.95E-16	0.208	-0.0275 (0.0034)	6.67E-16
rs6809522	chr3:43926407	ANO10	0.403	0.016 (0.0029)	3.00E-08	0.406	0.0146 (0.003)	8.37E-07
rs6805957	chr3:88209625	CGGBP1	0.155	0.032 (0.0035)	1.30E-19	0.155	0.0322 (0.0036)	1.18E-19
rs4894419	chr3:172433046	GHSR	0.441	-0.0186 (0.0026)	1.42E-12	0.446	-0.0189 (0.0027)	1.34E-12
rs6856650	chr4:759920	PCGF3	0.0632	0.0305 (0.006)	4.34E-07	0.0598	0.0339 (0.0062)	4.08E-08
rs56388229	chr4:69100940	UGT2B7	0.459	-0.0204 (0.0028)	4.88E-13	0.454	-0.0209 (0.0029)	2.30E-13
rs13133311	chr4:87296071	HSD17B13	0.424	-0.0307 (0.0027)	2.48E-29	0.431	-0.0315 (0.0028)	7.74E-30
rs147039802	chr4:104406043	CXXC4	0.0057	0.154 (0.0202)	2.36E-14	0.0057	0.154 (0.0202)	2.57E-14
rs2676118	chr4:168167754	ANXA10	0.464	0.0158 (0.0026)	1.28E-09	0.459	0.0166 (0.0028)	3.70E-09
rs1073917	chr5:137652050	SPOCK1	0.307	-0.016 (0.0029)	4.69E-08	0.303	-0.0166 (0.003)	3.60E-08
rs3096698	chr6:32139362	FKBPL	0.358	0.0225 (0.0029)	8.93E-15	0.359	0.0223 (0.003)	6.82E-14
rs4840067	chr6:99657153	PRDM13	0.465	-0.0221 (0.003)	1.31E-13	0.452	-0.023 (0.0031)	5.93E-14
rs2248551	chr6:131603549	MED23	0.171	-0.0203 (0.0036)	1.74E-08	0.17	-0.0203 (0.0037)	2.81E-08
rs4646450	chr7:99668695	CYP3A5	0.166	0.0248 (0.0035)	8.32E-13	0.162	0.0262 (0.0035)	8.91E-14
rs10988865	chr9:99580668	NR4A3	0.274	-0.0161 (0.0029)	4.33E-08	0.276	-0.0159 (0.0029)	6.79E-08
rs2274162	chr9:114425289	AKNA	0.229	0.020 (0.0034)	3.23E-09	0.228	0.0213 (0.0035)	8.47E-10
rs62576340	chr9:134373515	RXRA	0.178	0.0258 (0.0039)	4.43E-11	0.178	0.0261 (0.004)	4.16E-11
rs10867084	chr9:137880269	CACNA1B	0.489	-0.0157 (0.0026)	2.24E-09	0.486	-0.0159 (0.0027)	2.23E-09
rs10995311	chr10:62805174	ADO	0.44	-0.0232 (0.0027)	4.64E-18	0.44	-0.0234 (0.0027)	2.14E-18
rs10822156	chr10:63320967	NRBF2	0.48	-0.0994 (0.0027)	7.03E-300	0.48	-0.0998 (0.0028)	7.14E-288
rs7092153	chr10:65494502	LRRTM3	0.415	0.0444 (0.0028)	1.11E-58	0.412	0.0442 (0.0028)	1.84E-56
rs11827024	chr11:10243526	SBF2	0.486	0.0163 (0.0027)	2.38E-09	0.491	0.0183 (0.0028)	5.50E-11
rs9704692	chr11:10344482	AMPD3	0.465	-0.0171 (0.0028)	1.57E-09	0.466	-0.0173 (0.0028)	1.19E-09
rs6487369	chr12:24017085	SOX5	0.413	-0.0163 (0.0027)	1.83E-09	0.413	-0.0163 (0.0027)	1.83E-09
rs144998814	chr12:99826727	ANKS1B	0.0074	0.133 (0.0162)	2.34E-16	0.0074	0.133 (0.0162)	2.34E-16
rs2393776	chr12:120986603	HNF1A	0.382	0.0239 (0.0028)	5.85E-18	0.376	0.0239 (0.0028)	3.62E-17
rs17764243	chr14:69497596	PLEKHD1	0.263	-0.0169 (0.003)	1.16E-08	0.263	-0.0171 (0.003)	9.63E-09
rs7143025	chr14:100330863	SLC25A47	0.268	0.0252 (0.0034)	8.78E-14	0.269	0.0249 (0.0034)	2.81E-13
rs3976	chr17:1749797	SERPINF2	0.232	0.0297 (0.0032)	1.70E-20	0.228	0.0292 (0.0033)	2.73E-19
rs4151121	chr17:7438975	FGF11	0.369	-0.108 (0.0027)	<1E-300	0.373	-0.1104 (0.0028)	<1E-300
rs8079769	chr17:9685138	USP43	0.25	-0.022 (0.0032)	5.22E-12	0.241	-0.0214 (0.0033)	4.71E-11
rs72827832	chr17:48329746	SKAP1	0.232	-0.0184 (0.0032)	1.44E-08	0.235	-0.0188 (0.0033)	8.92E-09
rs7242706	chr18:52809677	DCC	0.208	-0.0172 (0.0032)	4.76E-08	0.209	-0.0168 (0.0032)	1.49E-07
rs8103923	chr19:45899125	MYPOP	0.126	-0.0445 (0.0041)	2.06E-27	0.127	-0.044 (0.0041)	1.97E-26
rs1800961	chr20:44413724	HNF4A	0.0336	-0.0465 (0.0071)	5.64E-11	0.0338	-0.0482 (0.0072)	2.11E-11
rs1807609	chr22:28735141	HSCB	0.306	0.0238 (0.003)	7.38E-16	0.306	0.024 (0.003)	9.84E-16

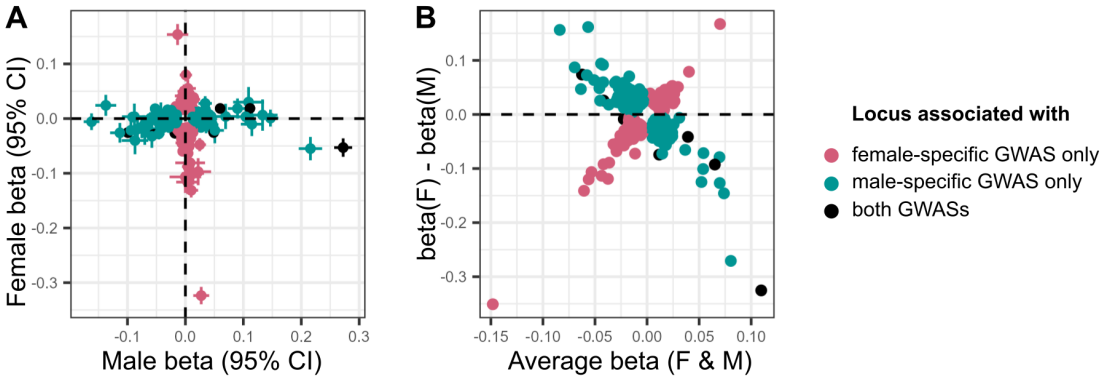


Figure 4.9: Sex heterogeneity in the effects of genetic variants associated with testosterone. Lead variants in either female-specific testosterone GWAS (coloured in pink), male-specific testosterone GWAS (coloured in green), or both (coloured in black) are displayed. Results are from all-ancestry meta-analyses. (A) Scatter plot comparing effect sizes and 95% confidence intervals (CIs), plotted as error bars, of GWS variants. (B) Bland-Altman plot displaying the difference between effect sizes estimated in the female-specific and male-specific GWASs for each variant, plotted against the mean estimate from both GWASs.

in men (15.67%, SE = 1.79%) than in women (9.87%, SE = 1.00%) (Table 4.6), and there is no significant genetic correlation between testosterone in men and women (r_G = 3.61%, P = 0.399).

I observe negative selection of Testosterone-M-increasing alleles at the *DNAL1* locus, and variants with higher probability of association with Testosterone-M also have more extreme directional selection scores (Figure 4.10). We observe positive selection of testosterone-increasing alleles at three loci: female-specific associations in the *CXCR4* and *SLC17A1* loci, and male-specific associations at the *SRD5A2* locus, as measured by extreme SDSs (in the lowest 0.25^{th} percentile and highest 99.75^{th} percentile) for several variants in these regions, but there is no consistent relationship between $-log10(P)$ of variant association with Testosterone and tSDS at these loci (Figure 4.10). However, there is no global signature of directional selection for testosterone, in men or women, when SDSs are aligned to the testosterone-increasing allele, i.e. the mean genome-wide trait-SDS is not significantly different from 0 (both $P > 0.05$).

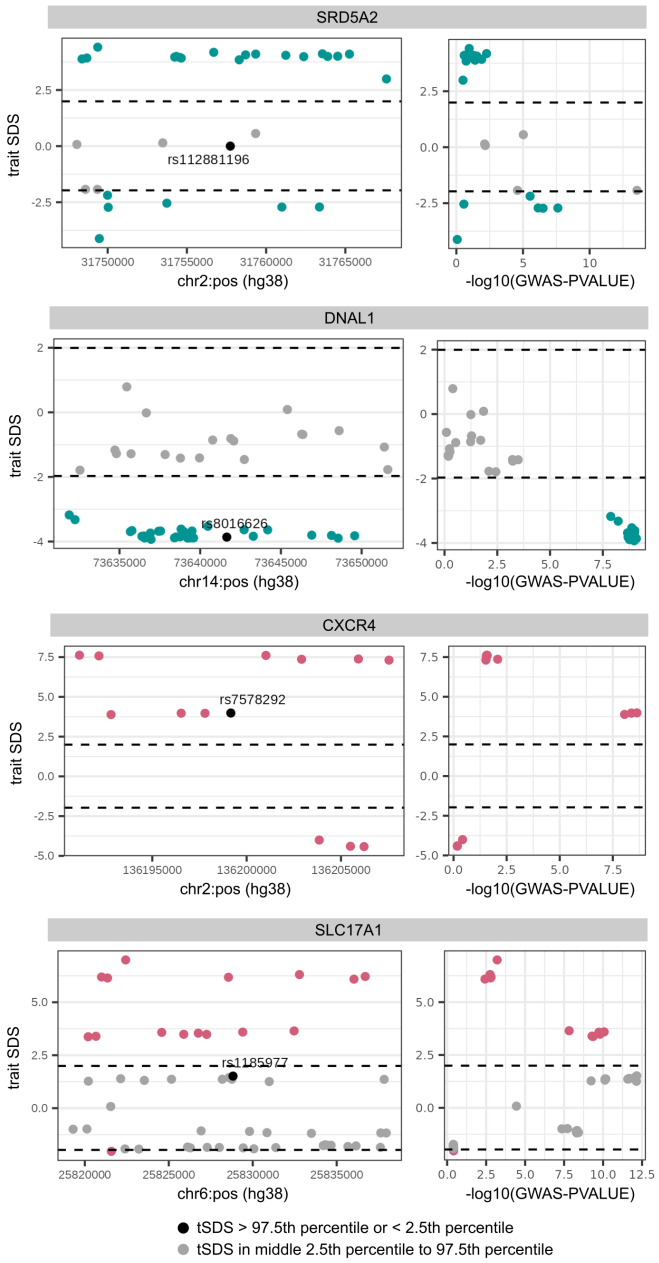


Figure 4.10: Trait-aligned Singleton Density Scores (tSDSs) at testosterone-associated loci with extreme values of tSDS. Variants in windows of ±10kb around each lead testosterone-associated variant (in black) are displayed. The tSDSs are aligned to the testosterone-increasing allele, wherein a positive tSDS indicates positive selection for the testosterone-increasing allele at the locus. In both plots, dashed lines indicate 2.5^{th} percentile (%le) and 97.5^{th} %le of SDS, and variants below or above this threshold respectively are coloured in pink (female-specific analyses) or green (male-specific analyses). Left: Locus plots depicting genomic position on the x-axis and tSDS on the y-axis. The lead variant rs112881196 is not present in the tSDS dataset and thus assigned tSDS of 0. Right: Scatter plots depicting relationship between -log10 of the GWAS p-value for the variant association with testosterone on the x-axis and tSDS on the y-axis.

No evidence for genetic heterogeneity across cohorts

I assessed the robustness of the reproductive hormone GWAS meta-analyses to the inclusion of summary statistics from publicly available datasets, which have different biomarker assays, genotyping and QC filters, and GWAS protocols to the cohort studies included in this Chapter. A large proportion of lead variants associated with Oestradiol-F (2/2), Oestradiol-M (2/4), Testosterone-F (146/153), and Testosterone-M (230/243), are recovered at GWS in the meta-analyses excluding publicly available datasets, and the remaining are still nominally associated at $P < 5 \times 10^{-5}$. There is also no heterogeneity between effect size estimates from meta-analyses with and without public summary statistics at all lead variants across all strata (heterogeneity $P > 0.01$) (Figure 4.11). Finally, as with the infertility meta-analyses in Section 4.3.1, all reported lead variants from the reproductive hormone GWAS meta-analyses are present in at least two of the cohorts examined and the majority (81%) show no evidence for heterogeneity in effect estimates across cohorts ($P_{\text{het}} > 0.05$).

4.3.4 Non-overlapping genetics of hormones and infertility

The only female infertility locus in our study with genetic signals for a female reproductive hormone is the *TNFSF12* locus, where the lead variant rs9902027 is protective against risk of F-ANOV (OR (95% CI) = 0.895 (0.86 - 0.931), $P = 4.06 \times 10^{-8}$), and is associated with higher Oestradiol-F (β (SE) = 0.0317 (0.0054), $P = 3.81 \times 10^{-9}$) and Testosterone-F (β (SE) = 0.0191 (0.0035), $P = 4.87 \times 10^{-8}$). Seven other female infertility loci are within 500 kb of a male testosterone variant ($P < 5 \times 10^{-8}$), but have no GWS associations with female reproductive hormones.

As we may have missed identifying sub-GWS signals for hormones in infertility loci and vice-versa, we assessed the statistical colocalisation of genetic signals between each pair of female infertility and reproductive hormone traits at all loci that reach GWS for at least one trait. In addition to the colocalisation of F-ANOV with Oestradiol-F and Testosterone-F at the *TNFSF12* locus on chr17 described above (PP > 95% for association with both traits, but with different

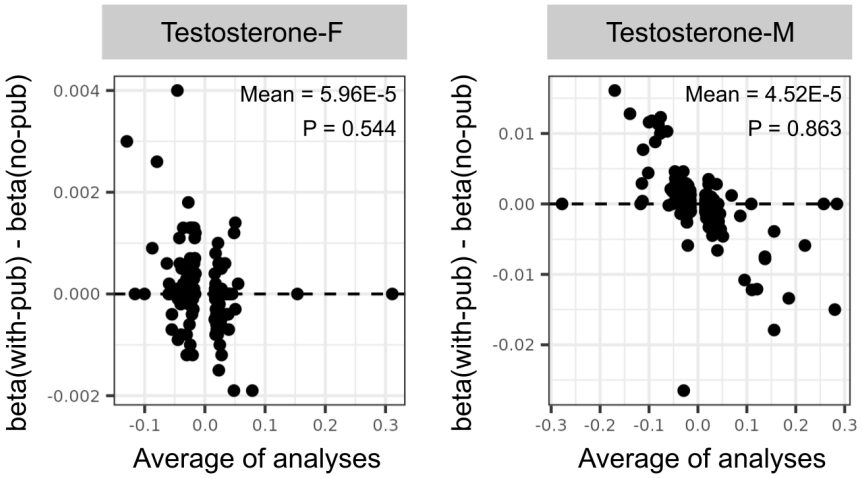


Figure 4.11: Comparison of effect sizes of SNPs associated with testosterone in main meta-analyses and sensitivity analyses without publicly available summary statistics. Variants that are genome-wide significant (GWS $P < 5 \times 10^{-8}$) in the main GWAS with public data are plotted. The Bland-Altman plot displays the difference between effect sizes estimated in the with-public data (with-pub) and no-public data (no-pub) GWASs for each variant, plotted against the mean estimate from the two sets of analyses. Results are from all-ancestry meta-analyses. The mean difference and one-sample t-test P-value (with null hypothesis of mean difference = 0) are displayed for each of the strata.

causal variants), I find evidence for shared genetic associations, with different causal variants, for F-ANOV and Testosterone-F at the *CDC42* locus on chr1 (Table 4.9). I also observe colocalisation between four categories of female infertility (all except F-ANAT) and Testosterone-F at the *RAP1GAP* and *WNT4* loci on chr1, with the latter carrying causal variants for several female reproductive conditions (Table 4.9). Finally, there is colocalisation of genetic signals between F-ALL and Testosterone-F at the *UGT2B11* (chr4) and *CYP2C8* (chr10) loci - CYP2C8 is an enzyme involved in steroid hormone metabolism, but has not previously been shown to have effects on fertility.

Despite the evidence for colocalisation of female infertility signals with testosterone and oestradiol at some loci, there is no genome-wide genetic correlation between any category of female infertility and any reproductive hormone in this study, or AMH (with key roles in folliculogenesis[481]) and TSH (all $P > 2.5 \times 10^{-3}$ after FWER control at 5% across 20 tests, Figure 4.3B). We observe nominally

Table 4.9: Posterior probabilities (PPs) of colocalisation of female infertility and female reproductive hormone signals, with different causal variants, in loci for either set of traits (for pairs of traits with PP > 95%). Windows of ±50 kb around the EUR-ancestry lead variant were used to define loci for colocalisation. Under the assumption of a single causal variant at the locus, both traits are associated, but with different causal variants, at each of these loci.

Lead variant	chr:pos (hg38) of 50kb window around lead variant	Reproductive hormone	Infertility category	N SNPs	Posterior probability of H3 (shared locus, different causal SNPs)
rs12130039	chr1:21675577-21775577	Testosterone_F	F-ALL	4320	100%
			F-ANOV	4041	99.30%
			F-EXCL	4307	99.90%
			F-INCL	4073	100%
rs72665317	chr1:21990580-22090580	Testosterone_F	F-ANOV	3843	99.30%
rs61768001	chr1:22089327-22189327	Testosterone_F	F-ALL	3996	100%
			F-ANOV	3774	99.30%
			F-EXCL	3986	99.90%
			F-INCL	3800	100%
rs114488830	chr4:69140408-69240408	Testosterone_F	F-ALL	5598	96.80%
rs542798166	chr10:95031567-95131567	Testosterone_F	F-ALL	3546	96.50%
rs9902027	chr17:7487667-7587667	Oestradiol_F	F-ANOV	2782	96.90%
		Testosterone_F	F-ANOV	2757	99.80%
rs3933469	chr17:7546184-7646184	Oestradiol_F	F-ANOV	2782	96.90%
		Testosterone_F	F-ANOV	2758	99.80%
rs62623385	chr17:7894519-7994519	Oestradiol_F	F-ANOV	3014	96.90%
		Testosterone_F	F-ANOV	2979	99.80%

significant correlation ($r_G = 0.748$, $P = 0.0130$) between F-ANOV and AMH, but this finding requires validation in better-powered studies.

4.3.5 Phenome-wide genetic correlation landscape differs for infertility and reproductive hormones

I attempted to unpick the genetic aetiology of infertility through a phenome-wide genetic correlation scan with 703 heritable phenotypes in the UKBB. At a phenome-wide significant threshold of $P < 4.90 \times 10^{-5}$ (FWER controlled at 5% across 1,020 tests using the Bonferroni method, accounting for 340 effectively independent UKBB phenotypes and 3 female infertility traits), I identified 11 phenotypes that are positively correlated with F-ALL, including female reproductive conditions (such as hysterectomy, $r_G = 0.481$ and diseases of the genito-urinary system, $r_G = 0.337$), general illness (abnormal clinical findings, $r_G = 0.325$, hospitalisation, $r_G = 0.282$, and number of operations, $r_G = 0.266$), and prospective memory (number of attempts, $r_G = 0.323$, duration screen displayed, $r_G = 0.262$, and overall result, $r_G = 0.345$) (Figure 4.3D). Four phenotypes were negatively correlated with female

infertility: age at first sexual intercourse ($r_G = -0.188$), and three measures of fluid intelligence (overall $r_G = -0.276$) (Figure 4.3D).

A similar phenome-wide scan for heritable reproductive hormones (FSH-F, Testosterone-F, and Testosterone-M) reveals a different phenotypic landscape. FSH is nearly perfectly genetically correlated with menopause-related phenotypes such as age at menopause ($r_G = -0.990$), and modestly correlated with the adiposity metric of waist circumference (WC) ($r_G = -0.189$) (Figure 4.12). Testosterone-F is genetically correlated with 24 adiposity-related phenotypes, including body fat percentage ($r_G = 0.172$), BMI ($r_G = 0.170$), WC ($r_G = 0.163$), and hip circumference (HC) ($r_G = 0.148$); and while Testosterone-M is also genetically correlated with these same phenotypes, the direction of association is negative (body fat percentage $r_G = -0.320$, BMI $r_G = -0.302$, WC $r_G = -0.344$, and HC $r_G = -0.263$) (Figure 4.13). Instead, testosterone in males is positively correlated with general well-being, as indicated by: no medication for cholesterol, blood pressure, diabetes, etc. ($r_G = 0.302$), no illnesses of siblings or parents ($r_G > 0.205$), recent strenuous physical activity ($r_G = 0.190$), and forced vital capacity ($r_G = 0.148$). None of the intelligence- or memory-related traits that are genetically correlated with female infertility are associated with hormones (all $P > 0.05$); and none of the 24 obesity-related traits associated with female reproductive hormones are also associated with any category of female infertility (all $P > 0.05$).

4.3.6 Heritability of testosterone is enriched in adrenal gland, liver, and ovarian cell types

I find significant enrichment (FDR $< 5\%$) of the heritability of Testosterone-F in the adrenal gland ($P = 1.03 \times 10^{-3}$) and hepatocytes ($P = 9.36 \times 10^{-4}$), and of Testosterone-M in the liver ($P = 1.16 \times 10^{-6}$), particularly hepatocytes ($P = 3.61 \times 10^{-4}$), but no other enrichment across 205 tissues and cell-types from the GTEx Project database[97] and the Franke lab single-cell database[443]; nor did I find evidence for tissue-specific enrichment of the heritability of female infertility of all causes in these samples (Figure 4.14). As these datasets do not comprehensively

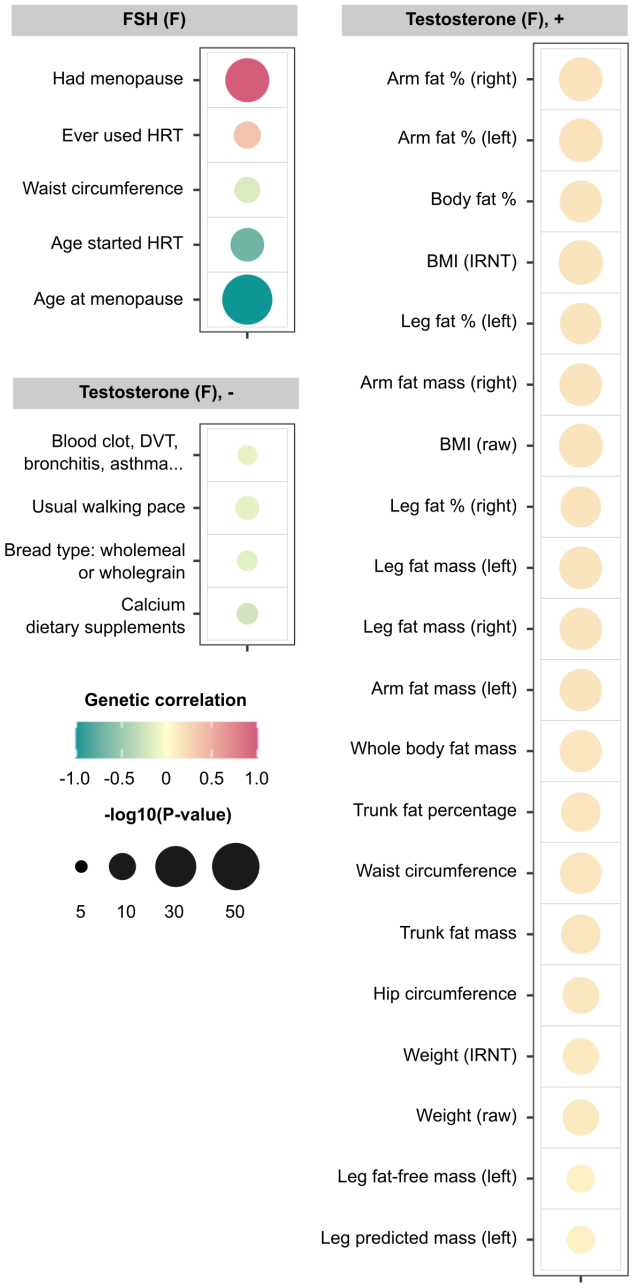


Figure 4.12: SNP-based genetic correlations (r_G) between follicle stimulating hormone in females (FSH-F) or Testosterone-F and traits across the phenome. Trait summary statistics were generated by the Neale lab[354] and genetic correlation calculations were performed using the LDSC software[62] on a subset of 1 million HapMap3 SNPs[46]. Points are coloured by r_G estimate and sized by $-\log_{10}(P)$. Only phenotypes associated at $P < 4.90 \times 10^{-5}$ (FWER controlled at 5% across 1,020 tests using the Bonferroni method, accounting for 340 effectively independent UKBB phenotypes and 3 hormone strata, FSH-F Testosterone-F, and Testosterone-M) are displayed, up to a maximum of 20 phenotypes. HRT = hormone replacement therapy, BMI = body mass index, IRNT = inverse-rank normally transformed, DVT = deep vein thrombosis.



Figure 4.13: SNP-based genetic correlations (r_G) between testosterone in males and traits across the phenome. Trait summary statistics were generated by the Neale lab[354] and genetic correlation calculations were performed using the LDSC software[62] on a subset of 1 million HapMap3 SNPs[46]. Points are coloured by r_G estimate and sized by $-\log_{10}(P)$. Only phenotypes associated at $P < 4.90 \times 10^{-5}$ (FWER controlled at 5% across 1,020 tests using the Bonferroni method, accounting for 340 effectively independent UKBB phenotypes and 3 hormone strata, FSH-F Testosterone-F, and Testosterone-M) are displayed, up to a maximum of 20 phenotypes each that are: left, negatively correlated with testosterone (-), and right, positively correlated with testosterone (+). IRNT = inverse-rank normally transformed.

represent female reproductive tissues, particularly the ovary, we curated gene sets for ovarian cell types from publicly available single-cell gene expression databases[445,446]. At FDR < 5%, I find enrichment of F-ALL heritability in ovarian stromal cells ($P = 2.52 \times 10^{-3}$) and Testosterone-F across a range of ovarian endothelial and immune cell types, as well as granulosa cell progenitors ($P = 0.0106$), granulosa cumulus cells ($P = 0.0113$), and theca cells ($P = 0.00176$) (Figure 4.14).

4.3.7 Rare variant contribution to reproductive hormone and infertility genetics

Having outlined the genetic architecture of infertility and reproductive hormones through GWASs, I next characterised the contribution of rare coding variants (MAF < 1%) to risk of F-ALL (3,746 cases / 260,413 controls), F-EXCL (3,012 cases / 261,147 controls), M-ALL (650 cases / 222,393 controls), and levels of FSH-F ($N = 20,800$), LH-F ($N = 16,391$), Oestradiol-F ($N = 54,609$), and testosterone ($N_{\text{female}} = 197,038$, $N_{\text{male}} = 197,340$) - with WES analyses in the UKBB. Gene-burden analyses, which aggregate the effects of rare variants in a gene across the population, implicate *PLEKHG4*, which is highly expressed in the testis and ovary, for F-EXCL (burden test OR (95%CI) = 1.04 (1.02 - 1.06), Cauchy $P = 1.37 \times 10^{-8}$ when aggregated across all variants with MAF < 1%). Rare variants in *PLEKHG4* are known to cause cerebellar ataxia[482], which is itself associated with steroid hormone deficiency and hypogonadism[483,484]. No other genes or variants are associated with any category of infertility in the UKBB at exome-wide significant thresholds.

WES analyses implicate novel genes associated with testosterone

Gene-based analyses identify two novel genes for testosterone in males, *ELFN2* (Cauchy $P = 2.33 \times 10^{-6}$) and *HSD17B2* (burden test $P = 1.33 \times 10^{-11}$ when aggregated across pLoF variants with MAF < 0.1%) (Figure 4.16 and Table 4.10). The latter encodes the hydroxysteroid 17-beta dehydrogenase 2 enzyme that catalyzes the oxidation of oestradiol, testosterone, and dihydrotestosterone to their less active forms and thus regulates the biological potency of steroid hormones[485,486].

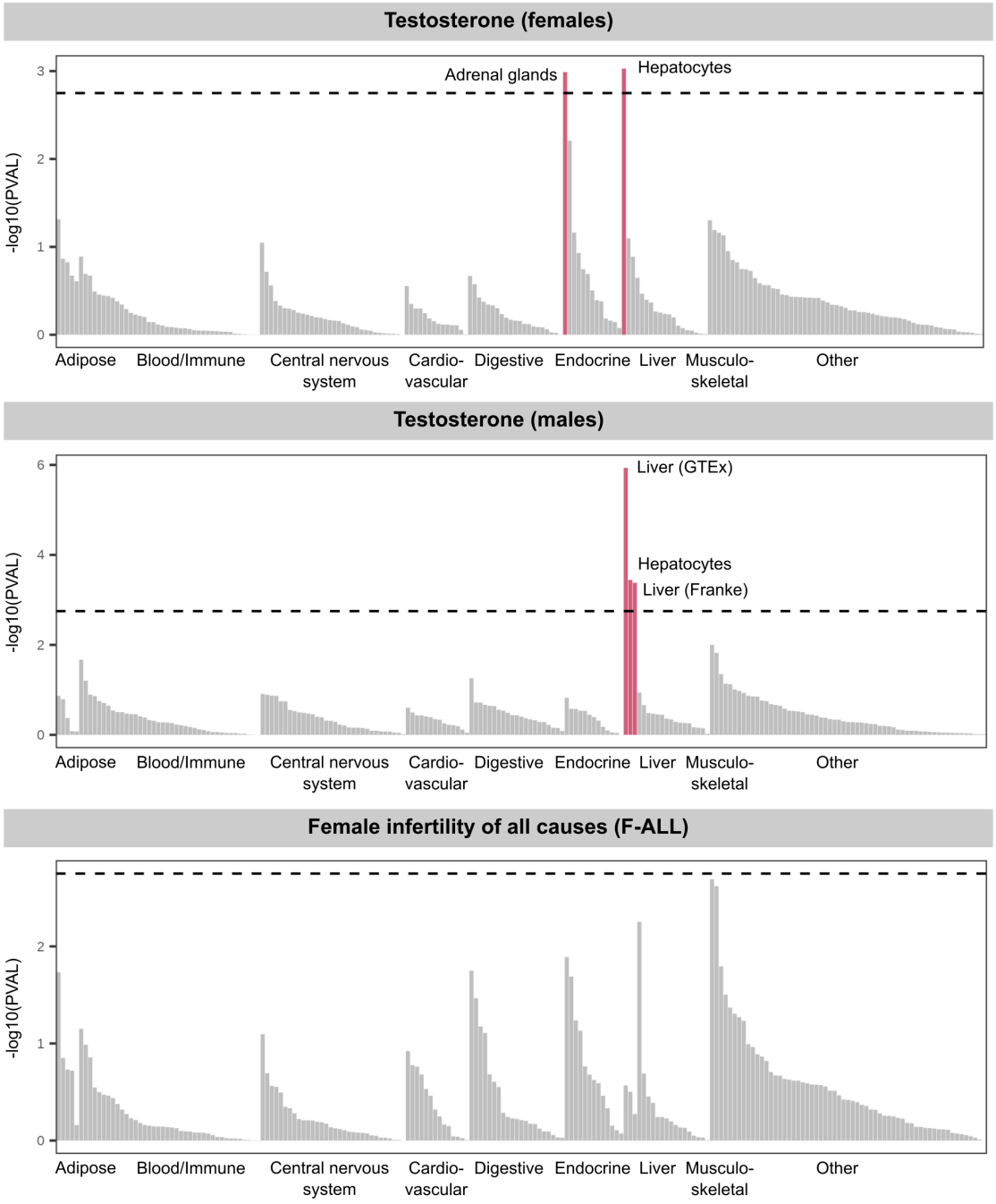


Figure 4.14: Partitioned heritability LD-score regressions for female infertility and reproductive hormones across tissues and cell types. Each panel displays a different trait, whose heritability across 205 tissues and cell-types from the GTEx Project database[97] and the Franke lab single-cell database[443] was assessed using partitioned LD-score regression[62]. Tissues and cell types are broadly grouped by organ system, and those that reach significance (FDR $< 5\%$, dashed line), are annotated and coloured in pink.

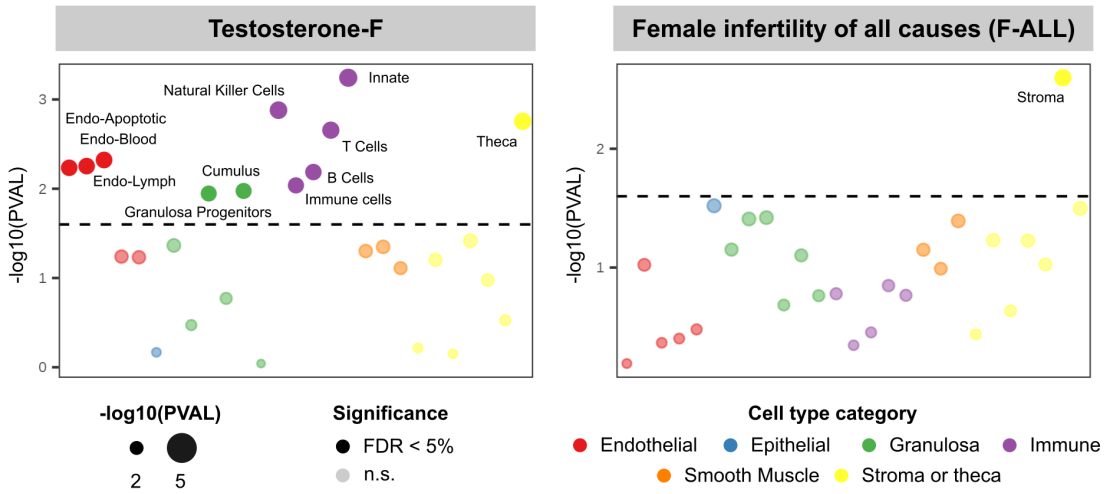


Figure 4.15: Partitioned heritability LD-score regressions for female infertility and testosterone in females, in ovarian cell types. Each panel displays a different trait, whose heritability across 27 ovarian cell types, identified from two single-cell sequencing datasets of adult human ovaries (Fan *et al.* (2019)[445] and Jin *et al.* (2022)[446]) (see Methods) was assessed using partitioned LD-score regression[62]. Cell types are grouped by tissue, and those that reach significance (FDR < 5%, dashed line), are annotated and coloured in bold. n.s. = non-significant.

There are no reported disease associations for rare variants in *HSD17B2*. I also identify, for the first time, the related *HSD11B1* gene for Testosterone-F (burden test $P = 1.93 \times 10^{-6}$ when aggregated across missense variants with MAF < 0.01%); pathogenic variants in this gene are reported to cause hyperandrogenism in women due to cortisone reductase deficiency[487,488]. We replicate known Testosterone-F associations with steroid hormone metabolic enzyme, *HSD11B2*, which catalyzes the conversion of hydroxyglucocorticoids such as cortisol to inactive forms such as cortisone[236] (Table 4.11).

Including *HSD11B1*, I find nine novel genes associated with Testosterone-F at exome-wide significance ($P < 5 \times 10^{-6}$), including those expressed in the ovaries (*ANAPC2*) and adrenal glands (*GPC2*), and metabolism (*PDE3B*, *TAP2*, *NOTCH4*, and *ZKSCAN1*) (Figure 4.16 and Table 4.11). Finally, across the reproductive hormones studied here, we replicate genes that have previously been reported to carry either rare or common variants associated with: FSH-F (genes *CHEK2* and *DCLRE1A*), both of which are associated with age at menopause, Oestradiol-F (gene *SHBG*), which encodes the sex-hormone binding globulin protein,

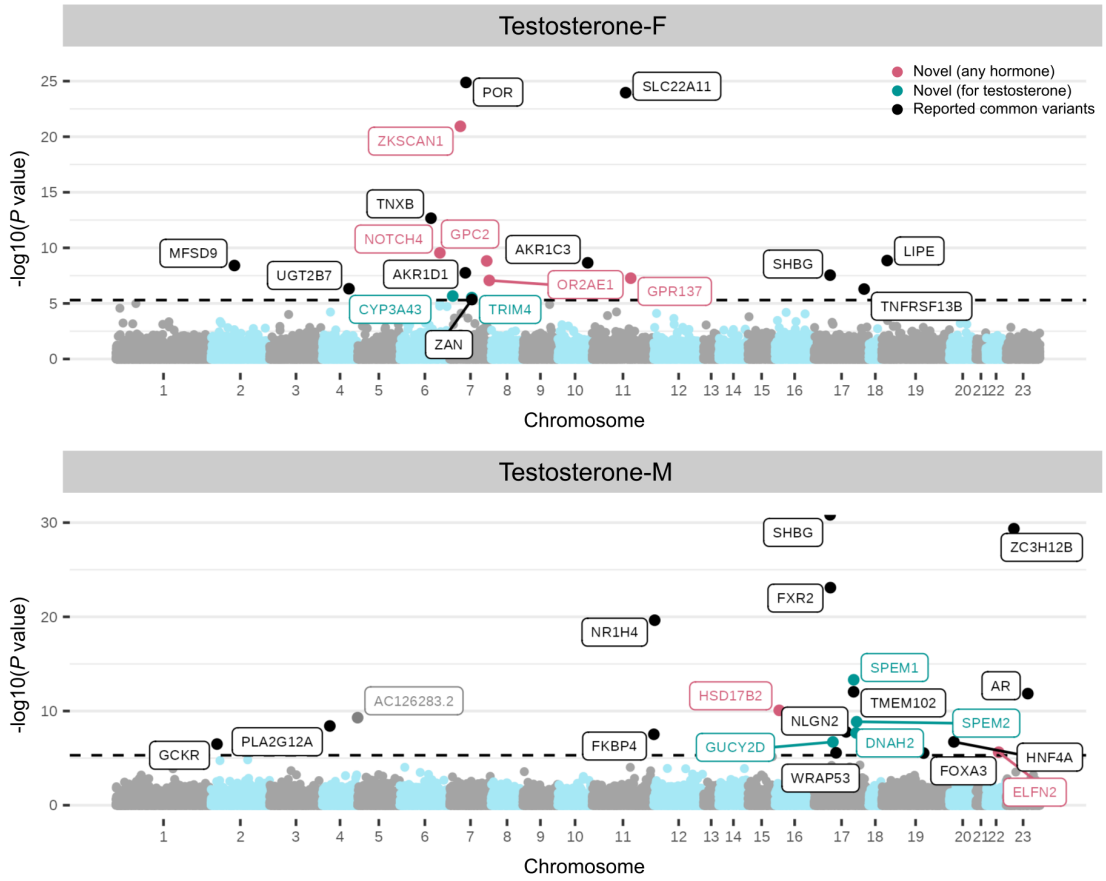


Figure 4.16: Gene-based Manhattan plots for burden of rare variants associated with testosterone in females and males in UK Biobank. Significance levels estimated using the Cauchy combined Pvalue are displayed, with significant genes (exome-wide significant at $P < 5 \times 10^{-6}$, FWER controlled at 5% using the Bonferroni method, across 10,000 effectively independent genes) coloured in: pink if no previous common variant associations with the gene have been reported for any of 28 reproductive hormones in the GWAS Catalog[50], green if no previous common variant associations with the gene have been reported for testosterone, and black otherwise.

Testosterone-F (18 genes, including two that are novel for testosterone but have known estrone associations), and Testosterone-M (22 genes, including eight that are novel for testosterone but have known SHBG associations) (Figure 4.16 and Tables 4.11 and 4.10).

Concordance between rare and common genetic architecture for testosterone

We integrated the associations of rare SNVs from WES analyses with genotyped and imputed common variants from GWAS meta-analyses to discover hormone-

Table 4.10: Gene-based association test results for male testosterone levels in UK Biobank. Genes that reach exome-wide significance ($P < 5 \times 10^{-6}$) for at least one test (Cauchy combined, Burden, or SKAT) are displayed, along with classification based on previously reported common variants previously for the gene.

ENSG symbol	HGNC Gene ID	Classification	P-value		
			Cauchy combined	Burden	SKAT
Testosterone (males)					
ENSG00000086696	HSD17B2	novel any hormone	8.49E-11	1.69E-11	5.58E-08
ENSG00000166897	ELFN2	novel any hormone	2.33E-06	7.95E-06	1.02E-05
ENSG00000198146	ZNF770	novel Testosterone	7.10E-06	3.34E-04	4.63E-06
ENSG00000131482	G6PC	novel Testosterone	5.92E-05	1.10E-05	6.88E-03
ENSG00000132518	GUCY2D	novel Testosterone	1.92E-07	5.36E-02	6.37E-08
ENSG00000181323	SPEM1	novel Testosterone	4.85E-14	3.62E-07	6.66E-15
ENSG00000183914	DNAH2	novel Testosterone	2.08E-08	2.26E-04	4.75E-08
ENSG00000184560	SPEM2	novel Testosterone	1.37E-09	3.29E-03	1.96E-10
ENSG00000115234	SNX17	novel Testosterone	1.83E-05	7.90E-05	2.60E-05
ENSG00000157992	KRTCAP3	novel Testosterone	1.51E-05	8.67E-03	1.10E-05
ENSG00000004478	FKBP4	reported	2.97E-08	1.19E-06	5.75E-08
ENSG00000012504	NR1H4	reported	2.30E-20	5.29E-21	1.44E-20
ENSG00000129214	SHBG	reported	0.00E+00	4.72E-263	2.88E-305
ENSG00000129245	FXR2	reported	8.02E-24	5.30E-09	1.15E-24
ENSG00000141499	WRAP53	reported	2.73E-06	7.86E-03	1.28E-06
ENSG00000169992	NLGN2	reported	1.64E-08	6.59E-05	8.80E-08
ENSG00000181284	TMEM102	reported	9.08E-13	1.04E-04	1.30E-13
ENSG00000170608	FOXA3	reported	2.92E-06	1.05E-05	8.87E-06
ENSG00000084734	GCKR	reported	3.23E-07	5.74E-07	2.19E-04
ENSG00000101076	HNF4A	reported	1.89E-07	1.48E-07	1.55E-05
ENSG00000123739	PLA2G12A	reported	3.88E-09	7.45E-09	4.20E-08
ENSG00000102053	ZC3H12B	reported	4.41E-30	1.39E-08	6.30E-31
ENSG00000131080	EDA2R	reported	1.56E-05	2.79E-05	4.02E-05
ENSG00000169083	AR	reported	1.42E-12	1.21E-07	2.04E-13

associated genetic variants across the allele frequency spectrum (Figure 4.17). Of the 29 unique genes carrying rare variants associated with testosterone in men or women in the UK Biobank, 18 also contain testosterone-associated common variants in the meta-analyses reported in this Chapter. In 15 of these 18 genes, the rare variant has a larger absolute effect size on the trait than any associated common variant (Table 4.12). For example, a rare missense variant (chr17:7631360:C:T, MAF = 0.612%) in *SHBG* is associated with large effects on Testosterone-M ($\beta = -0.743$, $P = 1.69 \times 10^{-291}$), Testosterone-F ($\beta = -0.121$, $P = 7.83 \times 10^{-9}$), and Oestradiol-F ($\beta = -0.259$, $P = 1.03 \times 10^{-9}$) in the UKBB WES analyses; while there are also several common variants (MAF > 1%) in *SHBG*, their effect sizes are over 7.5 times smaller (largest absolute effect on Testosterone-M: $\beta = -0.197$, $P = 3.46 \times 10^{-82}$, Testosterone-F: $\beta = -0.0324$, $P = 6.55 \times 10^{-22}$, and Oestradiol-F: $\beta = -0.035$,

Table 4.11: Gene-based association test results for female hormone levels in UK Biobank. Genes that reach exome-wide significance ($P < 5 \times 10^{-6}$) for at least one test (Cauchy combined, Burden, or SKAT) are displayed, along with classification based on previously reported common variants previously for the gene.

ENSG symbol	HGNC Gene ID	Classification	P-value		
			Cauchy combined	Burden	SKAT
Follicle stimulating hormone (females)					
ENSG00000198924	DCLRE1A	novel any hormone	6.16E-05	9.92E-04	6.40E-05
ENSG00000183765	CHEK2	novel FSH	4.16E-09	1.40E-08	1.19E-06
Oestradiol (females)					
ENSG00000131507	NDFIP1	novel any hormone	4.66E-07	4.67E-07	4.66E-07
ENSG00000129214	SHBG	reported	3.59E-09	4.04E-08	6.21E-09
Testosterone (females)					
ENSG00000117594	HSD11B1	novel any hormone	1.02E-05	9.07E-05	9.31E-06
ENSG00000152270	PDE3B	novel any hormone	5.78E-05	4.48E-05	1.59E-04
ENSG00000173264	GPR137	novel any hormone	5.36E-08	9.96E-07	3.71E-08
ENSG00000204267	TAP2	novel any hormone	1.75E-05	3.56E-03	6.70E-06
ENSG00000204301	NOTCH4	novel any hormone	2.80E-10	5.65E-02	3.67E-11
ENSG00000106261	ZKSCAN1	novel any hormone	1.16E-21	5.44E-15	2.15E-21
ENSG00000213420	GPC2	novel any hormone	1.51E-09	1.97E-01	1.55E-09
ENSG00000244623	OR2AE1	novel any hormone	8.52E-08	7.67E-05	7.60E-08
ENSG00000176248	ANAPC2	novel any hormone	1.25E-05	2.00E-01	5.76E-06
ENSG00000103449	SALL1	novel Testosterone	6.43E-05	5.19E-05	8.93E-03
ENSG00000021461	CYP3A43	novel Testosterone	2.16E-06	4.17E-03	1.93E-06
ENSG00000146833	TRIM4	novel Testosterone	3.00E-06	3.94E-03	2.07E-06
ENSG00000057252	SOAT1	reported	2.68E-05	1.97E-05	3.92E-03
ENSG00000196139	AKR1C3	reported	2.21E-09	4.15E-09	3.80E-05
ENSG00000168065	SLC22A11	reported	1.09E-24	9.69E-21	2.08E-24
ENSG00000176387	HSD11B2	reported	8.68E-05	2.82E-05	1.41E-04
ENSG00000129214	SHBG	reported	2.85E-08	8.62E-07	2.95E-07
ENSG00000240505	TNFRSF13B	reported	5.10E-07	4.40E-07	1.42E-05
ENSG00000079435	LIPE	reported	1.39E-09	5.14E-09	3.22E-10
ENSG00000135953	MFSD9	reported	3.89E-09	1.33E-07	2.06E-07
ENSG00000127419	TMEM175	reported	5.99E-05	5.03E-05	4.36E-02
ENSG00000171234	UGT2B7	reported	4.77E-07	3.37E-06	2.38E-06
ENSG00000168477	TNXB	reported	2.16E-13	1.76E-05	4.13E-14
ENSG00000204296	TSBP1	reported	9.60E-06	3.09E-02	3.87E-06
ENSG00000122787	AKR1D1	reported	1.76E-08	1.65E-08	1.98E-03
ENSG00000127948	POR	reported	1.32E-25	3.09E-14	1.06E-22
ENSG00000146839	ZAN	reported	4.51E-06	1.21E-02	1.04E-06

$P = 2.67 \times 10^{-11}$) (Table 4.12). Of note, we did not condition rare variants on nearby common variants, so some of the associations reported in Table 4.12 may be marginally significant due to LD with common causal variants.

The remaining 11 genes do not carry any common variants with reproductive hormone effects, such as the steroid metabolism gene *HSD17B2* associated with Testosterone-M (chr16:82090933:T:G, pLoF variant $\beta = -0.333$, $P = 1.99 \times 10^{-8}$) (Table 4.12). Some of the novel missense variants associated with Testosterone-F

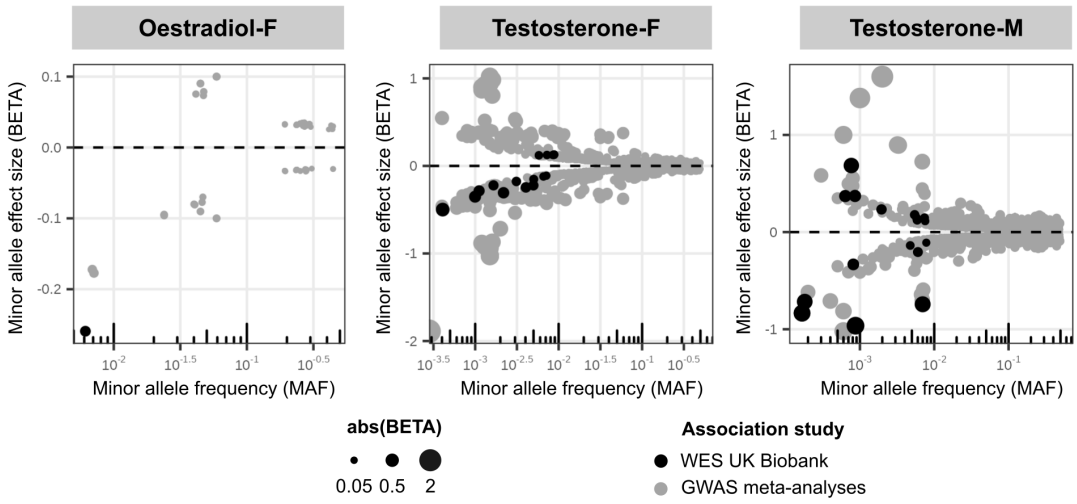


Figure 4.17: Effect size versus allele frequency of genetic variants associated with reproductive hormones. Each panel displays a reproductive hormone with at least one exome-wide significant variant ($P < 1 \times 10^{-7}$) in the UK Biobank whole exome sequencing (WES) analyses. Variants discovered at genome-wide significance ($P < 5 \times 10^{-8}$) in GWAS meta-analyses (coloured in grey) and exome-wide significance in the UK Biobank WES analyses (coloured in black) are plotted, sized by the absolute value of their effect size. Effect sizes are aligned to the minor allele, plotted against minor allele frequency on the log x-axis.

are in genes that are highly expressed in the testis, but whose functions in the female reproductive tract are not known, such as the testis expressed basic protein 1 gene *TSBP1* (chr6:32293376:C:T, $\beta = 0.125$, $P = 4.46 \times 10^{-10}$), *GPC2* (chr7:100171569:G:A, $\beta = -0.351$, $P = 6.65 \times 10^{-11}$), and *STAG3* (chr7:100204708:C:T, $\beta = -0.284$, $P = 2.31 \times 10^{-8}$). The latter gene is associated with premature menopause, primary ovarian insufficiency, and hypergonadotrophic hypogonadism in women[489,490], and induces female infertility through the absence of oocytes in knockout mouse models[479].

Increased risk of infertility in individuals carrying rare testosterone-lowering variants

Ten of the 29 unique genes with rare testosterone-associated variants, in men or women, have some previously reported association with infertility in humans or animal models. I therefore examined whether individuals carrying these variants were at higher risk of infertility in the UKBB, assessing significance at $P < 1.72 \times 10^{-3}$

Table 4.12: Rare variants associated with altered reproductive hormone levels in UK Biobank. Only the variant with the lowest P-value in each gene with at least one variant reaching exome-wide significance $P < 1 \times 10^{-7}$ is displayed. Genes are annotated with the common-variant (MAF > 1%) association of largest effect from the European-ancestry GWAS meta-analyses described in this Chapter.

MarkerID (hg38)	Gene symbol	UK Biobank Whole Exome Sequencing			Highest effect GWAS common variant (MAF > 1%) in gene	
		MAF	Beta (SE)	P-value	Beta (SE)	P-value
Oestradiol (females)						
chr17:7631360:C:T	SHBG	0.00612	-0.259 (0.0425)	1.03E-09	0.035 (0.0053)	3.08E-11
Testosterone (females)						
chr2:102723774:C:A	MFSD9	0.00713	-0.111 (0.0202)	3.26E-08		
chr6:32069045:C:T	TNXB	0.00881	0.129 (0.0182)	1.13E-12	-0.0584 (0.0102)	9.03E-09
chr6:32201219:C:G	NOTCH4	0.00854	0.125 (0.0185)	1.08E-11	-0.0884 (0.0155)	1.24E-08
chr6:32293376:C:T	TSBP1	0.00724	0.125 (0.02)	4.46E-10	-0.0375 (0.0063)	3.39E-09
chr7:100033558:G:A	ZKSCAN1	0.00218	-0.306 (0.0364)	3.57E-17		
chr7:100171569:G:A	GPC2	0.000996	-0.351 (0.0538)	6.65E-11	-0.0666 (0.0047)	4.52E-45
chr7:100204708:C:T	STAG3	0.00111	-0.284 (0.0509)	2.31E-08		
chr7:100594531:C:T	FBXO24	0.000408	-0.499 (0.084)	2.90E-09	-0.122 (0.0123)	3.10E-23
chr7:100766559:A:G	ZAN	0.00502	-0.152 (0.024)	2.13E-10		
chr7:75981558:C:T	POR	0.00405	-0.245 (0.0267)	5.12E-20	-0.0202 (0.0033)	6.94E-10
chr7:99876535:G:T	OR2AE1	0.00166	-0.222 (0.0416)	9.49E-08	-0.366 (0.0502)	2.92E-13
chr11:64286653:C:A	GPR137	3.12E-03	-0.178 (0.0305)	5.54E-09	-0.0426 (0.0047)	1.53E-19
chr11:64556141:C:T	SLC22A11	0.00498	-0.222 (0.0241)	2.75E-20		
chr17:7631360:C:T	SHBG	0.00658	-0.121 (0.021)	7.83E-09	-0.0494 (0.009)	4.53E-08
chr19:42410516:G:A	LIPE	0.00579	0.121 (0.0224)	6.05E-08		
Testosterone (males)						
chr4:109717668:C:T	PLA2G12A	0.00754	0.113 (0.0196)	7.44E-09		
chr12:100532530:T:C	NR1H4	0.00603	-0.205 (0.0219)	6.34E-21	-0.087 (0.0117)	8.67E-14
chr16:82090933:T:G	HSD17B2	0.000815	-0.333 (0.0593)	1.99E-08		
chr17:7285104:G:A	SLC2A4	0.00195	0.233 (0.0385)	1.31E-09		
chr17:7417551:G:A	NLGN2	0.00584	0.122 (0.0222)	4.16E-08	-0.128 (0.0077)	8.31E-62
chr17:7421083:G:A	SPEM1	0.00739	0.146 (0.0198)	1.57E-13	-0.0796 (0.0053)	2.47E-51
chr17:7427170:C:T	SPEM2	0.0058	0.144 (0.0223)	1.19E-10	-0.107 (0.003)	1.73E-272
chr17:7437401:C:T	TMEM102	0.00593	0.121 (0.0221)	4.60E-08	-0.0657 (0.0052)	5.65E-37
chr17:7560782:G:A	TNFSF13	0.000166	-0.835 (0.132)	2.34E-10		
chr17:7593093:G:A	FXR2	0.00545	0.178 (0.023)	1.07E-14	-0.212 (0.0109)	5.86E-84
chr17:7631360:C:T	SHBG	0.00698	-0.743 (0.0204)	1.69E-291	-0.197 (0.0103)	3.46E-82
chr17:8013934:T:C	GUCY2D	0.000858	0.372 (0.058)	1.46E-10	-0.0979 (0.0071)	7.32E-43
chr20:44413714:C:T	HNF4A	0.000181	-0.716 (0.126)	1.31E-08	-0.0482 (0.0072)	2.11E-11
chrX:65488953:T:C	ZC3H12B	0.000766	0.685 (0.0613)	4.76E-29		
chrX:67721909:C:G	AR	0.000634	0.371 (0.0673)	3.62E-08	0.114 (0.008)	8.28E-46

after FWER control at 5% across 29 genes using the Bonferroni method. A Testosterone-F-lowering missense variant chr7:100171569:G:A in *GPC2* is associated with higher risk of F-EXCL (OR (95% CI) = 2.63 (1.40 - 4.92), $P = 1.25 \times 10^{-3}$); the *GPC2* gene has not previously been reported to associate with infertility, but reduces adrenal gland size in knockout mouse models[479] (Figure 4.18). I also report nominal associations between the Testosterone-F-lowering missense variant chr7:100766559:A:G in *ZAN* (zonadhesin) and higher risk of F-ALL (OR (95% CI) = 1.37 (0.949 - 1.98), $P = 0.0463$) (Figure 4.18); *ZAN* is expressed in the testis and knockout mice display impaired sperm binding to the zona pellucida[491], but its role

in female infertility is unclear. In men, the testosterone-lowering damaging missense variant chr20:44413714:C:T in *HNF4A* has a nominal association with increased risk of infertility (OR (95% CI) = 17.6 (2.26 - 137.07), $P = 3.08 \times 10^{-3}$) (Figure 4.18).

Finally, I assessed whether any of the genes associated with reproductive hormones also had effects on infertility in the UKBB WES gene-based analyses, at a threshold of $P < 1 \times 10^{-3}$ after FWER control at 5% across 50 unique genes using the Bonferroni method. Two Testosterone-F associated genes are also associated with risk of female infertility: the testosterone 6-beta-hydroxylase enzyme-coding gene *CYP3A43* (associated with F-EXCL, burden test OR (95% CI) = 1.02 (1.01 - 1.03), $P = 4.84 \times 10^{-4}$ across all variants with MAF < 1%) and the ovary-expressed *TRIM4* gene (associated with F-ALL, burden test OR (95% CI) = 1.03 (1.01 - 1.05), $P = 4.05 \times 10^{-4}$ across all variants with MAF < 0.1%), indicating a potential role for hormonal infertility driven by a burden of rare variants in these genes.

4.4 Discussion

In this comprehensive genetic investigation of infertility and related reproductive phenotypes in over 1 million individuals, we identify 19 genetic loci associated with female infertility, two with male infertility, and novel variants for female reproductive hormones FSH (3 novel variants), LH (1), oestradiol (1), and testosterone (28) and male testosterone (39). I find that half of female infertility loci overlap with genes for endometriosis and PCOS, and just one (*TNFSF12*) contains variants associated with testosterone in women. The modest genetic correlations between F-ALL and these phenotypes (max r_G (SE) = 0.585(0.0785) with endometriosis) indicates that a substantial proportion of female infertility is not driven by known reproductive conditions or reproductive hormone regulation. There is evidence at some non-hormonal infertility loci for recent directional selection (*EBAG9*) and balancing selection (*LPIN1*, *MBL2*), which leads me to hypothesise that pleiotropy may drive selection at these loci. Finally, although we demonstrate non-shared genetic architectures of infertility and reproductive hormones at the population level, genes containing rare protein-coding variants associated with testosterone are also

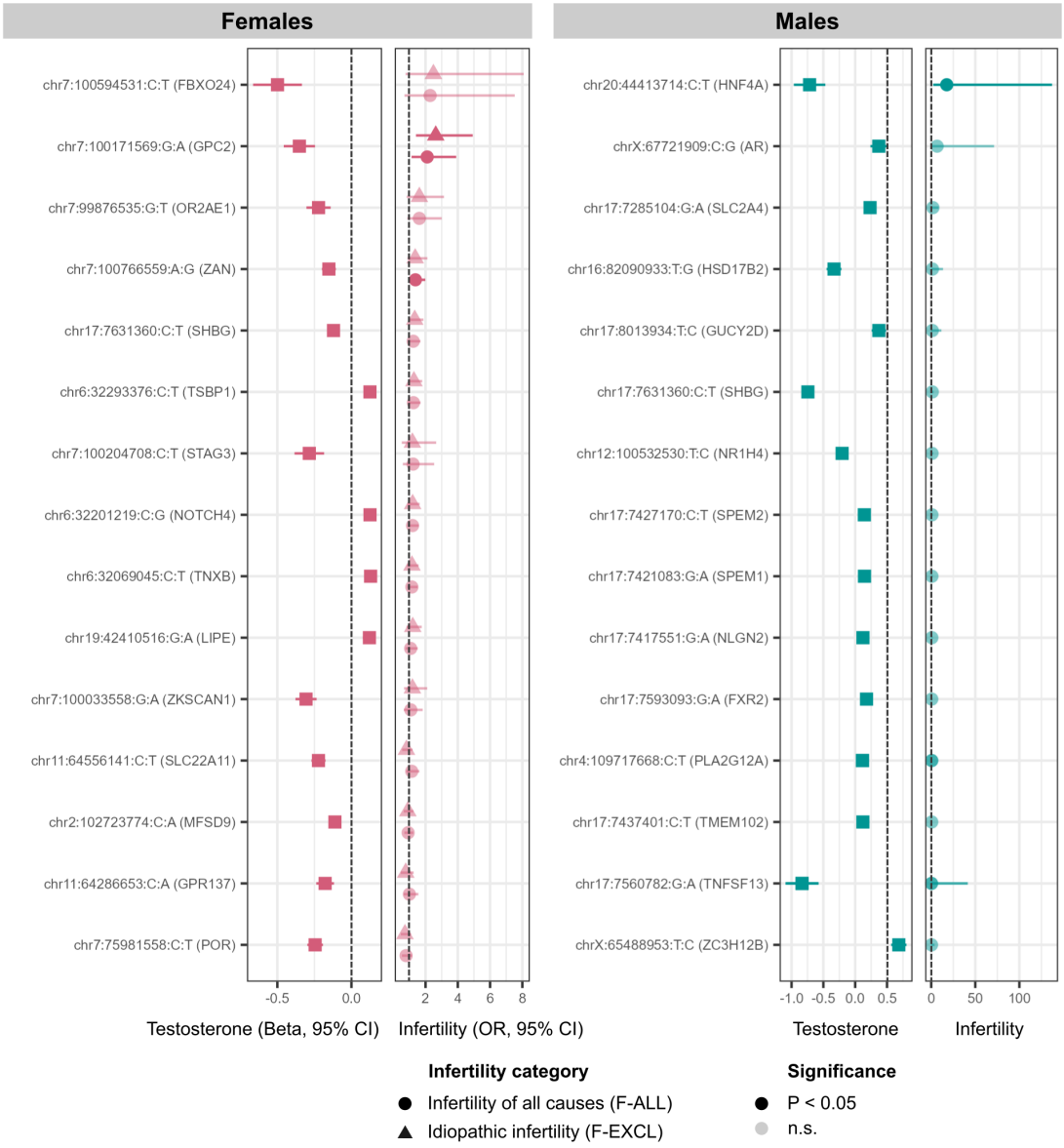


Figure 4.18: Effects of testosterone-associated rare variants on infertility in females (left) and males (right). Per gene, the variant with lowest P-value of all variants that reach exome-wide significance ($P < 1 \times 10^{-7}$) in the UK Biobank whole exome sequencing (WES) analyses for testosterone is displayed. Effect sizes (beta and 95% confidence intervals (CIs) for the variant effect on testosterone are to the left of each plot, and effect sizes (odds ratios (ORs) and 95% CIs) for the variant effect on infertility are to the right of each plot. Variants that reach nominal significance (P < 0.05) are coloured in solid shapes.

associated with higher risk of male infertility (*HNF4A*) and female infertility (*GPC2*, *CYP3A43*, *TRIM4*) in the UK Biobank and may inform individual diagnoses.

Although highly prevalent, the complex and heterogeneous causes of infertility remain poorly understood[403]. In this Chapter, I show that infertility recorded as idiopathic in EHRs (F-INCL) is strongly genetically correlated with endometriosis ($r_G = 0.710(0.115)$), which is itself under-diagnosed and often remains untreated for up to 10 years after symptom-onset[227,228]. Some proportion of idiopathic infertility cases may thus be owed to under-diagnosis of endometriosis, whose early recognition and treatment could not only greatly benefit the quality of life for affected women, but may also prevent future infertility[492,493]. While weight loss is often recommended as beneficial for improved fertility[494,495], and BMI thresholds determine access to public funding for fertility treatment[496], I do not find substantial genetic evidence indicating a role for obesity in driving infertility risk. This adds genetic support to evidence from randomised control trials demonstrating no fertility benefits from short-term weight loss in overweight and obese women[497]. Given that current fertility treatments such as *in vitro* fertilisation (IVF) pose an outsize physical, emotional, and financial burden on couples and healthcare systems[498–501], a richer understanding of the genetic and non-genetic contributions to infertility is urgently necessary.

PCOS is a heritable cause of up to 80% of anovulatory infertility cases that may be treated through induced ovulation[197,250,254]. Reassuringly, we observe strong negative genetic correlation ($r_G = -0.740(0.182)$) between F-ANOV and spontaneous dizygotic twinning, a heritable metric of female fecundity that captures the propensity for multiple ovulation[435]. These results indicate value in studying the genetic basis of ovulation rate to develop treatments for anovulatory infertility. Higher levels of anti-Müllerian hormone are associated with increased ovarian reserve in healthy women[502,503], but women with PCOS who have high levels of AMH respond poorly to ovulation treatment[504]. Indeed in our study, there is nominally significant positive correlation between levels of AMH and risk of F-ANOV ($r_G = 0.748(0.301)$), which may reflect ovulation issues in high-AMH PCOS cases.

However, there is overall only modest genetic correlation between PCOS and F-ANOV ($r_G = 0.403(0.131)$). I suggest that hypothalamic-pituitary-ovarian disorders of Group I (failures of hypothalamic pituitary function), non-PCOS endocrinopathies in Group II (hypothyroidism, prolactinaemias, etc.) and Group III disorders (ovarian failure) may also contribute significantly to the genetic aetiology of anovulatory infertility and will require different treatments to PCOS-associated infertility[505].

Previous efforts to catalogue the genome-wide architecture of infertility have relied on proxy measures such as childlessness, number of children ever born, age at first sexual intercourse and age at first birth[260–263], which may be confounded by behavioural, socio-economic, and lifestyle factors. While I do find modest genetic correlation between female infertility and age at first sexual intercourse ($r_G = -0.188(0.044)$), indicating that the latter does capture some shared biology with fertility, our meta-analyses do not replicate the associations of infertility proxy variables with putative behavioural loci such as *CADM2* for risk-taking[262,506], or *AFF3* and *CAMKV* for educational attainment[262,263,507,508]. We suggest that the clinically defined or self-reported infertility cases in our cohorts may be less confounded by behavioural factors than were previous GWASs for proxy measures. Nonetheless, I observe genetic correlations between F-ALL and requiring more time and attempts at prospective memory tests (max $r_G = 0.345(0.0736)$), as well as lower fluid intelligence scores (min $r_G = -0.312(0.0691)$). As these traits are also correlated with educational achievement[509,510], which is itself associated with delayed age at first birth[511], there may still be residual confounding of age in the female infertility of all causes phenotype studied in this Chapter.

The heritability of all categories of infertility is low ($< 10\%$). Non-genetic factors such as infection[406–408] and delayed reproduction[405], and other environmental and lifestyle factors will likely explain some population variance in infertility. I show in this Chapter that balancing selection and recent (2,000 to 3,000 years) positive selection for infertility risk alleles at pleiotropic loci may explain the persistence of genetic factors for infertility. For example, the *EBAG9* locus associated with female infertility is under recent directional selection, perhaps because *EBAG9*, which is

highly expressed in CD34- / CD41+ / CD42+ megakaryocytes[93,450], plays a role in T-cell mediated cytotoxicity as part of the adaptive immune memory response to infection[451]. *EBAG9* may also play a complementary role in the placenta during early pregnancy, where lower *EBAG9* expression is associated with inappropriate activation of the maternal immune system and results in foetal rejection[452]. Other pleiotropic loci with putative metabolic and immune-related fitness advantages are the *LPIN1* locus[459] associated with female infertility and the *MBL2* locus[471] associated with male infertility, which are both under balancing selection.

Testosterone is among the most sexually dimorphic phenotypes in humans, and has a corresponding dimorphic genetic architecture. Previous studies[236] and ours estimate the genetic correlation between female-specific and male-specific testosterone to be negligible (between 0 to 3%), highlighting the importance of dis-aggregating both genetic and phenotypic analyses by sex. I also report opposing systemic associations of testosterone with adiposity in the two sexes; testosterone in women is genetically correlated with higher overall and central obesity as measured by BMI, WC and HC, body fat percentage, fat mass, etc., all of which are negatively correlated with testosterone in men. These sexually dimorphic genetic correlations reflect reported observational and causal associations in the literature[512,513], and may be related to the hormonal activity of adipose tissue, which metabolises steroid hormones, and whose distribution differs in men and women[514]. As testosterone is thought to partly mediate the deleterious cardiovascular consequences of increased visceral adipose tissue mass in men and post-menopausal women[181,515], understanding the shared genetic aetiology of testosterone and adiposity could lead to novel insights and treatments for cardiovascular disease. I additionally note that the sex-specific tissue enrichment of testosterone reflects underlying metabolism. The liver, which produces sex-hormone binding globulin that regulates the bio-availability of testosterone, is enriched for testosterone heritability in both men and women, but the adrenal is only significantly enriched in women, in whom adrenal glands produces a relatively large proportion of total testosterone[516].

Unexpectedly, the testes and ovaries are not significantly enriched for the heritability of testosterone, which is surprising as the gonads are major sites for testosterone production[517,518]. However, neither organ is dis-aggregated into tissues or cell types in the GTEx database, upon which these analyses were based, so gene expression profiles may not capture cell-type specific effects. We therefore collated single-cell gene expression profiles for ovary cell types from publicly available databases, where we do indeed observe enrichment for testosterone heritability in the androgen-secreting thecal cells and androgen-responsive granulosa cells[519–521]. The heritability of female infertility of all causes is enriched in ovarian stromal cells, which are similar to fibroblasts and form the stroma connective tissue. Although there are several causal roles hypothesised for stromal dysfunction in infertility, such as impaired folliculogenesis[522], restricted blood flow[523], and ovarian scarring[524], more work is needed to robustly replicate these findings. In general, there is a critical need for more functional studies of gonadal cell types, in both men and women, to enable a mechanistic understanding of the genetic variation associated with reproductive hormones and infertility.

Rare protein-coding variants that are captured in whole exome sequencing studies are valuable in revealing the underlying biology of phenotypes as they directly implicate effector genes[525]. Indeed, in our UKBB-based WES analyses, we find that several enzymes involved in steroid hormone synthesis and metabolism, such as those in the hydroxysteroid dehydrogenase (HSD) family[486], carry rare variants that drastically alter reproductive hormone levels. We also find an association between female infertility and a burden of rare variants in *PLEKHG4*, which has no previous links to infertility but is associated with spinocerebellar ataxia, which frequently co-occurs with hypogonadism[483,484]. Finally, we display concordance between the testosterone-altering effects of common and rare variation, reflecting the inverse relationship between allele frequency and effect size that has been observed across several other phenotypes[526]. As we were limited to the UKBB dataset for WES analyses, we were under-powered to test exome-wide associations for most categories of infertility (with fewer than 1,000 cases) and all male reproductive hormones but

testosterone (with fewer than 1,000 samples). Although it is theoretically possible to meta-analyse rare variants from WES analyses or high-confidence imputed rare variants across cohorts, this remains analytically challenging and requires the development of novel methodology[527]. Doing so is likely to prove immensely fruitful for the discovery of novel biology.

We employed a broad search strategy to maximise sample sizes for cases of infertility and reproductive hormone levels in our meta-analyses. Procedures for the diagnosis of infertility vary by country and have changed over time[403], which we observe as a wide spread in the prevalence of infertility across cohorts. The case proportion of male and female infertility is between 0.3% to 1.4% in the UKBB, which may reflect: (1) the biases of UKBB for healthy participants[264], (2) participants who were born between 1950-70 and thus reached reproductive age when infertility treatments were not widely available[528], and (3) the separation of fertility treatment records from other medical databases in the UK[528]. On the other hand, the Estonian and Danish data more closely reflect the population prevalence of infertility (between 7.0% to 13.2%) - the Estonian biobank is currently recruiting participants and is reflective of the current age distribution in Estonia[128], with recent infertility diagnoses included in the dataset, while the Copenhagen Hospital Biobank consists of patients with blood draws in Danish hospitals[425] and is thus more likely to capture individuals who are unwell and have interacted with the healthcare system, which is genetically correlated with infertility in our study (max $r_G = 0.325$). We also capture reproductive hormone values that were assayed using different methodologies, in primary care or hospital EHRs, and at different ages and stages of the menstrual cycle in women. There is likely to be ascertainment bias in these values, for example because measuring levels of FSH and LH was recommended to diagnose menopause in the UK until the late 1990s[153]. Although we are able to adjust for some of these covariates, such as age (which is expected to account for much of the effect of menopause on hormone levels), we do not have the granularity to account for hormonal fluctuations during the menstrual cycle and pregnancy, among other events that can drastically alter hormone levels[226,529] and

for which the data may be ascertained[530]. The noisiness of the phenotype is likely to contribute to the low number of GWS signals for all hormones but testosterone in our study. However, it is reassuring that the associations we do observe are biologically robust for each hormone, such as the *FSHB* and *LHB* associations with FSH and LH respectively, even when we include summary statistics from publicly available GWASs based on proteomic phenotypes, adjusted for different covariates, and processed with different pipelines than our cohort GWASs. In the future, longitudinal GWASs for steroid hormones that can incorporate mean and variance of hormone levels over the menstrual cycle, or phenotypes that calculate ratios between various hormones over time, will likely reveal fundamental biology that is missed by the broad-stroke assessments in this study.

Indeed, despite our best efforts to encompass heterogeneity in the causes and effects of infertility across the globe, we were under-powered to perform granular GWASs for different categories of male infertility, whose diagnosis and treatment lags behind female infertility[531]. Further, while the all-ancestry GWAS meta-analyses included one cohort of South Asian ancestry (Genes and Health) for infertility meta-analyses, and individuals of African, East Asian, and South Asian ancestries in the UKBB for hormone meta-analyses, over 90% of participants in our study were nonetheless of European ancestry. Even within the European-ancestry analyses, we demonstrate the importance of diverse population representation for biological insights in genetic studies. We identify six infertility-associated variants that are common in the Finnish and Estonian populations, but not present in any of the other European cohorts, which may have immediate implications for fertility - such as a male infertility variant near *ENO4*, a gene expressed in the testis and involved in sperm motility[456,457]. Increasing the ancestral diversity in GWASs will not only be of great benefit to populations with ancestry-specific variants that affect fertility and reproductive hormone levels, but may also reveal shared novel biology.

In this first, large-scale investigation of the genetic determinants of infertility and reproductive hormones, I nominate several genes associated with infertility through their effects on reproductive disease and selection pressures among others. I

also provide evidence in this Chapter against the prevailing notion that reproductive hormone dysregulation and obesity are strongly correlated with infertility at the population level. Instead, I suggest that other genetic and non-genetic avenues must be explored to treat complex and heterogeneous fertility disorders that impact the physical, emotional, and financial well-being of millions of individuals across the globe.

5

Genetically causal relationships of obesity with female reproductive diseases

Contents

5.1 Introduction 176
5.2 Methods 177
5.2.1 Observational associations in UK Biobank 177
5.2.2 Two-sample Mendelian Randomisation 179
5.2.3 Non-linear Mendelian Randomisation 185
5.2.4 Mediation Mendelian Randomisation 186
5.2.5 Disease and variant clustering 187
5.3 Results 188
5.3.1 Obesity is associated with altered risk of female reproductive conditions in the UK Biobank 188
5.3.2 Body fat distribution is genetically associated with altered risk of female reproductive conditions 199
5.3.3 Hormones mediate the effects of genetically predicted obesity on female reproductive disorders 207
5.3.4 Mechanisms by which different forms of obesity affect the risk of various female reproductive diseases 216
5.4 Discussion 218

Some of the material in this Chapter is adapted from my first-authored publication, titled "Obesity and risk of female reproductive conditions: A Mendelian randomisation study", published in PLOS Medicine[2]. I have obtained permission from the journal, my supervisor, and all co-authors to re-use this material.

5.1 Introduction

Increased body fat mass has been associated with altered risk of a range of female reproductive conditions, including PCOS[195,196], irregular and heavy menstrual bleeding (HMB)[532,533], endometriosis[534,535], uterine fibroids[252,536], disorders of pregnancy such as preeclampsia or eclampsia[537,538], and issues with fertility[416,539]. There are close links between the metabolic and endocrine systems that may contribute to these associations. Adipose tissue is an androgen-modifying tissue that metabolises testosterone to oestrogen, and is the main producer of oestrogen in post-menopausal women, whose ovaries no longer produce the hormone[379]. Visceral and subcutaneous adipose depots accumulate in different parts of the human body, and have distinct metabolic and endocrine functions[540]. There is also evidence that changes in body fat distribution may induce, or be induced by, fluctuations in metabolic and reproductive hormones, including insulin, adipokines, oestrogen, and testosterone[423,424,514]. Such changes in fat distribution and hormone levels have been linked to the risk of cardio-metabolic disease in post-menopausal women[181,377], but less is known about the intricate relationships between body fat distribution, hormones, and female reproductive conditions of the fertile years.

Previous studies of the associations between obesity and reproductive disorders have primarily relied on observational designs including case-control settings[535,536], cross-sectional studies in groups of women randomly sampled from the population[532,538], or cross-sectional studies only in overweight and obese women[533]. While effect directions have largely been consistent across these studies, there is considerable heterogeneity in effect sizes, likely leading from different definitions of cases and study populations, and small sample sizes (fewer than 1,000 participants) in most studies. Mendelian Randomisation is an instrumental variable approach to estimating causal associations between sets of traits using genetic variants that are randomly assigned and fixed at conception[63], and is therefore less susceptible to the issues of confounding and reverse causation that bias observational studies[541,542]. MR has been broadly applied to study the effects of BMI on the spectrum of common complex diseases, including female reproductive disorders such as ovarian

cancer[543], endometrial cancer[544], and PCOS[545]. However, the aetiological roles of different forms of obesity, and in particular their interplay with hormones, in several female reproductive disorders remain unknown.

In this Chapter, I apply a range of observational and causal models – including logistic regressions, generalised additive models, two-sample MR, mediation analyses, and clustering – to dissect the heterogeneous and hormonally-mediated roles of overall and central obesity in the development of female reproductive disorders. Doing so in the large-scale population-based UKBB cohort, and by using summary statistics from large-scale GWAS meta-analyses, enables the systematic comparison of effect sizes across various forms of obesity and several female reproductive conditions.

5.2 Methods

5.2.1 Observational associations in UK Biobank

Analyses in this section were restricted to women (defined here as individuals with "XX" sex chromosomes who self-reported sex as "female") of genetically ascertained non-Finnish European ancestry in the UKBB who passed all QC steps previously described (Section 2.3). We further restricted the dataset to unrelated women by excluding the individual with higher rate of genotyping missingness in all pairs of 2nd degree or closer relatives (kinship coefficient > 0.0884), resulting in 225,388 participants carried forward for analysis.

Obesity-related phenotypes

I obtained values of BMI, WC, and HC from baseline UKBB assessment centre measurements and calculated WHR as the ratio of WC to HC. I performed QC to exclude any individuals missing values of BMI or WHR, or whose values of BMI or WHR were > 5 SDs away from the mean, retaining 224,502 individuals. I further calculated WHRadjBMI for each individual by taking the residual of WHR linearly regressed on BMI. To estimate body size from a time-point prior to the development of most female reproductive conditions, we obtained categorical values

for the UKBB self-reported questionnaire response on comparative body size at age 10 years (with presented options: "thinner", "about average", or "plumper").

Reproductive condition phenotypes

Cases of six female reproductive conditions were defined from primary care records, hospital in-patient records, and self-reported illness codes, with controls defined as participants without a diagnosis of that condition (Table 5.2). The following case numbers were carried forward for regression analyses: endometriosis (7,596 cases; 216,906 controls), HMB (31,069 cases; 193,433 controls), female infertility of all causes (F-ALL, 3,509 cases; 220,993 controls), PCOS (2,058 cases; 222,444 controls), preeclampsia (2,276 cases; 222,226 controls; this includes cases of eclampsia in order to capture individuals with preeclampsia whose condition developed into eclampsia), and uterine fibroids (20,395 cases; 204,107 controls).

Linear models

I performed logistic regression in two stages to estimate the confounder-adjusted and normally-transformed associations of BMI, WHR, and WHRadjBMI with each of the six female reproductive disorders. In the first stage, I linearly adjusted and normally-transformed each obesity trait y for individual $i = 1, \ldots, N$ as follows:

$$\hat{y}_i = x_i^T \beta + \varepsilon_i \tag{5.1}$$

where x_i are the trait-adjusting covariates: age, $(\text{age})^2$, UKBB assessment centre, and smoking status (coded as categorical variable: "Never", "Previous", or "Current"). We adjust for both linear and quadratic effects of age to account for the non-linear association of BMI and other adiposity phenotypes with age (see Chapter 3). The residual from Equation (5.1) provides an estimation of the obesity trait y adjusted for possible confounders. I then apply a deterministic rank-based inverse normal transformation[330] to the residuals from fitting model Equation (5.1) as follows:

$$\tilde{y}_i = \Phi^{-1}\left(\frac{r(\hat{y}_i - x_i^T \hat{\beta}) - c}{N - 2c + 1}\right) \tag{5.2}$$

where $r(\hat{y}_i - x_i^T\hat{\beta})$ is the rank of the ith residual among all N residuals, the offset c is 0.5, and $\Phi(\cdot)$ is the CDF of the standard Gaussian distribution.

The maximum sample size for regressions after excluding participants missing covariate values was 224,278 individuals. In the second stage, I performed logistic regressions to test for the association of each female reproductive disease with the appropriately transformed and normally-adjusted obesity trait values. All logistic regression models were implemented using the programming language `R`[329].

To test the sensitivity of our results to adjustment factors, we ran a set of logistic regression models excluding smoking status as a covariate.

Non-linear models

To test for non-linearity in the relationship between obesity and female reproductive conditions, I fit fractional polynomial models with the closed test procedure as implemented in the `mfp v1.5.4` package[546] in `R`. We tested for the:

1. evidence of overall association,
2. likelihood of non-linearity, and
3. best-fitting fractional polynomial function.

Additionally, to provide flexibility to model curves that cannot be represented by n^{th} degree polynomials, we fit generalised additive models (GAMs) allowing for thin-plate regression spline smoothing of each obesity trait using the `mgcv v1.8-38` package[547] in `R`.

As above, all models were adjusted for age, (age)^2, UKBB assessment centre, and smoking status. Model fits were compared with AIC[548].

5.2.2 Two-sample Mendelian Randomisation

Constructing instruments from obesity-related exposures

I constructed instruments for MR from lead variants for BMI, WHR, and WHRadjBMI from a recent large-scale GWAS meta-analysis, which combined summary

statistics from the GIANT Consortium and UKBB to achieve a maximum sample size of 806,801 individuals[216]. The reported lead variants were obtained by LD clumping of variants with $r^2 > 0.05$ in GWS loci (established at $P < 5 \times 10^{-9}$ to account for denser imputation data) to identify independent variants within windows of ±5Mb[216]. I also derived instruments for predicted VAT mass from a recent female-specific, European-ancestry GWAS in 161,118 individuals in the UKBB[549]; WC and HC instruments from European-ancestry GWASs of 462,166 and 462,117 individuals respectively, released by the Neale lab[354] and curated in the `ieugwasr` package[550]; and WHR instruments with specific effects on waist circumference or hip circumference alone from European-ancestry GWASs in 18,330 individuals[551].

I tested three weighting strategies for the BMI, WHR, and WHRadjBMI instruments, for which both female-specific and sex-combined lead variants and summary statistics were available: (i) sex-combined lead variants weighted by effect sizes in the sex-combined GWAS, (ii) female-specific lead variants weighted by effect sizes in the sex-combined GWAS, and (iii) female-specific lead variants weighted by effect sizes in the female-specific GWAS. The latter strategy produced the strongest instruments for all three obesity-related traits, as evaluated by the mean F-statistic across all SNPs in the instrument, where $F = \frac{\hat{\beta}^2}{\sigma^2}$ for a variant with estimated effect size $\hat{\beta}$ and standard error σ (Table 5.1). For robustness, I report results derived from all the weighting strategies tested. I also constructed instruments for BMI, WHR, and WHRadjBMI using lead variants and summary statistics from a previous release of the GIANT Consortium without UKBB participants[552] to test the sensitivity of results to UKBB sample overlap between the exposure and outcome summary statistics.

Finally, we constructed a joint WHR and BMI instrument using lead variants from both GWASs in Pulit *et al.* (2019)[216] to perform multivariable MR.

Outcomes

An overview of the studies contributing to female reproductive outcomes in this Chapter, including case and control definitions and sample sizes, can be found in

Table 5.1: Summary of genetic instruments for obesity-related traits, hormone mediators, and female reproductive disorders, with number of variants per instrument and instrument strength. GWAS = genome-wide association study.

Exposure	Number of SNPs in instrument	Effect size GWAS sample size (cases/controls)	Lead SNPs from GWAS in:	Effect size from GWAS in:	F-statistic	Data source
Obesity-related traits						
Body mass index (BMI)	670	806801	Combined sexes	Combined sexes	66.6	UKBB European-ancestry individuals (Pulit *et al.* 2019)
	670	434785	Combined sexes	Females	38.1	
	281	434785	Females	Females	61.3	
Waist-hip ratio (WHR)	316	697702	Combined sexes	Combined sexes	59.8	Meta-analysis of UKBB and GIANT European-ancestry individuals (Pulit *et al.* 2019)
	316	381147	Combined sexes	Females	49.4	
	203	381147	Females	Females	74.8	
WHRadjBMI	346	694602	Combined sexes	Combined sexes	71.9	Meta-analysis of UKBB and GIANT European-ancestry individuals (Pulit *et al.* 2019)
	346	379486	Combined sexes	Females	64.1	
	266	379486	Females	Females	84.7	
Predicted visceral adipose tissue mass	22	161168	Combined sexes	Females	58.6	UKBB European-ancestry individuals (Karlsson *et al.* 2019)
Waist circumference (WC)	597	462166	Combined sexes	Combined sexes	45.1	UKBB European-ancestry individuals, MRC-IEU GWAS (ukb-b-9405)
Hip circumference (HC)	648	462117	Combined sexes	Combined sexes	49.5	UKBB European-ancestry individuals, MRC-IEU GWAS (ukb-b-15590)
WC-specific WHR	36	18330	Combined sexes	Combined sexes	65.7	European-ancestry individuals (Lotta *et al*. 2018)
HC-specific WHR	22	18330	Combined sexes	Combined sexes	115	

contd. Summary of genetic instruments for obesity-related traits, hormone mediators, and female reproductive disorders, with number of variants per instrument and instrument strength.

Exposure	Number of SNPs in instrument	Effect size GWAS sample size (cases/controls)	Lead SNPs from GWAS in:	Effect size from GWAS in:	F-statistic	Data source
Hormone-related traits for mediation						
Insulin (fasting)	38	151013	Combined sexes	Combined sexes	51.5	European-ancestry individuals (Chen *et al.* 2021)
Follicle stimulating hormone	5	30596	Females	Females	38.7	In-house GWAS meta-analyses of European-ancestry individuals from 4 cohorts and publicly available GWASs,
Testosterone	17	38541	Females	Females	66.1	
Female reproductive disorders for reciprocal MR						
Endometriosis	20	57248/698764	Females	Females	50.2	Meta-analyses of GWASs in European-ancestry individuals, see Appendix C for details
Heavy menstrual bleeding	33	31309/318510	Females	Females	46.4	
Infertility (all cause)	5	38894/546837	Females	Females	51	
Infertility (anovulatory)	6	6774/558346	Females	Females	36.2	
PCOS	25	14467/430267	Females	Females	49.7	
Preeclampsia or eclampsia	5	7531/420551	Females	Females	34.3	
Uterine fibroids	100	42446/588955	Females	Females	71.2	

Table D.1. I obtained GWAS summary statistics for three female reproductive conditions – endometriosis[251], HMB[252], and PCOS[197,250] – from the largest European-ancestry studies with publicly available GWAS summary statistics, downloaded from the GWAS Catalog[50] on 21 Aug 2023. As the published HMB GWAS[252] did not contain samples from FinnGen, I meta-analysed this with matched phenotypes in FinnGen data freeze 9[129]. For PCOS, I additionally included an in-house GWAS of PCOS in the UKBB defined by a combination of hospital in-patient records, primary care records, and self-report, performed by Dr. Laura Wittemans, a post-doctoral researcher in the Lindgren group. For female reproductive disorders without published GWAS summary statistics, i.e. preeclampsia and uterine fibroids, I meta-analysed matched phenotype definitions from FinnGen data freeze 9[129] and UKBB summary statistics generated by the Neale lab[354]. In all, maximum sample sizes for each condition were: endometriosis (57,248 cases, 698,764 controls), HMB (31,309 cases, 318,510 controls), PCOS (14,467 cases, 430,267 controls), preeclampsia or eclampsia (7,531 cases, 420,551 controls), and uterine fibroids (42,446 cases, 588,955 controls).

All summary statistics were aligned to hg19 or GRCh37 genome builds, with the exception of those downloaded from FinnGen data freeze 9[129], which I lifted over from hg38 to hg19 using the `UCSC liftOver` tool[432] – 95.5% of variants were successfully lifted over to the hg19 genome assembly. As the downloaded PCOS summary statistics from Tyrmi *et al.* (2022)[250] were missing allele frequencies, I added frequencies for matching alleles from European-ancestry individuals in the 1000 Genomes dataset[280] and discarded any variants not present in the 1000 Genomes project (retaining 10.3 million variants). Effect sizes and standard errors for the HMB dataset[252] and the datasets released by the Neale lab[354] for preeclampsia and uterine fibroids were estimated using linear regression models (rather than logistic regression models), so these were transformed to the odds ratio scale using the approximation:

$$\log(\mathrm{OR}) = \frac{\beta}{\mu(1-\mu)} \tag{5.3}$$

where μ is the case fraction in case-control analyses[553]. Prior to IVW fixed-effects meta-analysis using `METAL`[78], I performed QC of the individual studies to retain only bi-allelic variants present on the autosomes or X chromosome, with MAF $\geq$ 1% and imputation information score > 0.8 (where available). I further retained only variants with plausible standard errors (< 10) and P-values ($0 < P < 1$).

In addition to the five female reproductive conditions above, I included European-ancestry GWAS summary statistics for the five female infertility categories defined in Chapter 4, i.e. F-ALL, F-ANAT, F-ANOV, F-INCL, and F-EXCL; detailed descriptions of case definitions and meta-analysis strategies can be found in the relevant sections of Chapter 4. I lifted over the female infertility summary statistics from genome build hg38 to hg19 using the `UCSC liftOver` tool[432] – 96.5% of variants were successfully lifted over to the hg19 genome assembly.

Power calculation

I calculated the power to detect causal effects with two-sample MR using two methods, one designed for general 2-sample[554] and the other for MR performed on binary outcomes[555]. I calculated the power afforded to investigate an effect size of OR = 1.5 or more extreme, at a significance level of 5%, based on the proportion of cases in case–control GWASs and the variance explained by genetic instruments for the exposure to each method.

Methods for Mendelian Randomisation

All MR analyses described in this section were performed using the `TwoSampleMR v0.5.7` package[556]. Summary statistics used for MR were harmonised to ensure effect allele consistency across exposure and outcome datasets. I used Rücker's framework, which assesses the trade-off between power and pleiotropy using the Cochran Q-statistic, to choose between the IVW (better-powered) and MR-Egger (robust to pleiotropy) methods for each exposure $\times$ outcome pair tested[557]. I additionally report results from the weighted median method, which protects against outlier variants.

Reciprocal MR

I performed reciprocal MR to regress BMI, WHR, and WHRadjBMI as outcomes on female reproductive disorders. I constructed genetic instruments for each reproductive condition from the GWAS summary statistics described in Section 5.2.2 by identifying lead variants in each 1 Mb locus that reached GWS ($P < 5 \times 10^{-8}$) through distance-based pruning (see detailed Methods in Section 2.5.1). I calculated instrument strength using F-statistics as above (Table 5.1). Summary statistics for the obesity-related traits as outcomes were obtained from a meta-analysis of the GIANT Consortium and UKBB[216].

5.2.3 Non-linear Mendelian Randomisation

I conducted non-linear MR to test for evidence of non-linearity in the genetic relationship between BMI and female reproductive conditions with $> 5\%$ prevalence in the UKBB, i.e. HMB and uterine fibroids, to maintain sufficient sample sizes in each BMI quantile to allow for the estimation of localised average causal effects.

A BMI instrument was constructed from 96 lead variants, identified in female-specific GWAS meta-analysis of BMI in a previous release of the GIANT Consortium without UKBB participants[552], for which genotyping or imputation data were available in UKBB. I calculated a BMI genetic risk score based on these instrument variants with the linear scoring function in `PLINK v2.00` for the 224,502 female participants with observations of BMI in the UKBB as described in Section 5.2.1. The variance in BMI explained by the instrument was calculated after adjusting for: age, $(\text{age})^2$, smoking status, UKBB assessment centre, genotyping array, genotyping batch, and the first 21 genetic PCs.

I implemented two binomial non-linear MR models using the `nlmr v2.0` package[558] in R: (i) fractional polynomial models with 100 BMI quantiles, and (ii) piecewise linear models with 10 BMI quantiles. I tested for evidence of non-linearity with the fractional polynomial non-linearity statistic and Cochran's Q test, and checked for heterogeneity of the instrumental variable with the Cochran's Q and trend tests. These tests have all been described previously by Staley *et al.* (2017)[558].

After the work in this Chapter was published, the developers of the above non-linear Mendelian Randomisation approach noted that violations of the strong parametric assumptions for residual stratification, i.e. the effect of the genetic instrument is linear and constant in the cohort, may result in biased effect estimates[559]. They instead proposed an improved implementation that uses doubly-ranked stratification rather than the residual stratification method, and recommend that any future work compares estimates from both methods to produce robust estimates[560].

5.2.4 Mediation Mendelian Randomisation

To test whether the genetic effect of obesity on female reproductive conditions may be mediated by metabolic or endocrine hormones, I curated female-specific lead variants and GWAS summary statistics for the five reproductive hormones described in Chapter 4, i.e. FSH, LH, progesterone, oestradiol, and testosterone levels. I additionally obtained GWAS summary statistics from the GWAS Catalog for levels of leptin (combined sexes)[561], TSH (combined sexes)[433], AMH[434], and fasting insulin[562]. Of these, FSH, testosterone, and fasting insulin were the only hormones with at least five independent GWS loci and instrument F-statistics > 10, and therefore powered for genetic mediation analyses[563]. I performed two-step Mendelian randomisation to estimate the total and indirect effects of obesity on female reproductive disorders and calculated the proportion of effect mediated by the above hormones. All analyses were implemented using the `TwoSampleMR v0.5.7` package[556] in `R`[329].

Step 1: Genetic effect of obesity on hormones

In the first step of mediation MR, I regressed the proposed mediators on BMI, WHR, and WHRadjBMI using the IVW method. Instruments for obesity-related traits were obtained from female-specific and female-weighted GWAS in the meta-analysis of the GIANT Consortium and UKBB[216]. I also assessed whether the tested hormones – FSH, testosterone, and fasting insulin – had a causal effect on obesity by constructing genetic instruments from lead variants for these hormones for reciprocal MR. Hormone lead variants were obtained by distance-based pruning

of each 1 Mb GWS locus ($P < 5 \times 10^{-8}$, see detailed Methods in Section 2.5.1). I calculated instrument strength with F-statistics as above (Table 5.1).

Step 2: Obesity-adjusted genetic effect of hormones on female reproductive traits

In the second step of mediation MR, I performed multivariable MR using a combined genetic instrument for each exposure × mediator combination to estimate the exposure-adjusted effect of the mediator on each female reproductive outcome. I restricted analyses to those exposure × mediator combinations where the obesity-related exposure had a nominally significant effect ($P < 0.05$) on the hormone mediator (or vice-versa if the hormone had a genetically predicted causal effect on the obesity-related trait). Female reproductive outcomes for each analysis were also restricted to those where the exposure had a nominally significant effect ($P < 0.05$) on the outcome.

Finally, the proportion of mediated effect of exposure on outcome was calculated as the ratio of the indirect to the total effect, where the indirect effect is the estimate of the exposure-adjusted mediator on the outcome, and the total effect is the estimate of the unadjusted exposure on the outcome. Standard errors for the proportion of mediated effect were calculated with the delta method[564].

5.2.5 Disease and variant clustering

I clustered female reproductive conditions by the estimated genetic effects of BMI, WHR, and WHRadjBMI on each condition using PCA. For each obesity-related trait, the input for dimension reduction by PCA was a matrix of single-variant genetic effects, estimated with the Wald ratio[556], across all female-specific, female-weighted instrument SNPs and the ten female reproductive outcomes. SNP effect sizes across each disease were scaled to have variance 1, and the first two PCs were plotted to visualise the separation between diseases.

To cluster instrument variants with similar effects on the outcomes, I applied heterogeneity clustering implemented in `MRClust v0.1.0`[565], which identifies clusters of genetic variants with similar magnitude, direction, and precision of causal

effect estimates on the outcome. I ran MRClust separately for each combination of obesity trait × female reproductive condition to identify SNPs belonging to distinct biological pathways (i.e. variants assigned ≥80% probability of belonging to a "substantial" cluster, as opposed to "null" or "junk" clusters).

Variants were annotated to functionally prioritised genes using the `v2g` score in the Open Targets Genetics platform[93,450]. Gene expression profiles and mouse phenotypes for each gene were also obtained from the Open Targets Genetics platform accessed on 30 August 2023. I performed gene over-representation analysis to test for the enrichment of genes in "substantial" clusters versus all genes annotated to obesity-associated variants using the `WebGestalt` platform[566] with the KEGG pathway database[567]. Nominally significant pathways ($P < 0.05$) are reported.

5.3 Results

5.3.1 Obesity is associated with altered risk of female reproductive conditions in the UK Biobank

The observed effect of centralised obesity on reproductive disease is partly driven by overall obesity

Overall obesity, as measured by BMI, and centralised obesity, as measured by WHR and WHRadjBMI, are associated with altered risk of all six female reproductive conditions tested – endometriosis, heavy menstrual bleeding, female infertility of all causes, PCOS, preeclampsia, and uterine fibroids, identified from a combination of health records and self-report – in a sample of 224,278 women in the UKBB. Women were aged 57.2 years on average (SD = 7.82) at the time of assessment, with mean BMI 26.9 kg/m^2 (5.07). The prevalence of reproductive conditions varies from <1% for PCOS to 13.8% for heavy menstrual bleeding, with 76.4% of women having none of the diseases studied here (Table 5.2).

BMI is associated with increased risk of five of the six reproductive conditions tested, with the strongest association observed between BMI and PCOS (OR (95% CI) per SD increase in BMI = 1.30 (1.26 - 1.35), $P = 2.01 \times 10^{-32}$) (Table 5.3 and Figures 5.1 and 5.2). Infertility is the only reproductive condition with an

Table 5.2: Summary of female reproductive disorders in the UK Biobank. S.D. = standard deviation, BMI = body mass index, WHR = waist-hip ratio, PCOS = polycystic ovary syndrome.

Diagnosis	Number of cases (prevalence)	Mean age, years (S.D.)	Mean BMI, kg/m2 (S.D.)	Mean WHR (S.D.)	Smoking status (prevalence)	Had menopause (prevalence)	Diagnosis definition
None	171477 (76.4%)	57.2 (7.82)	26.9 (5.07)	0.816 (0.0697)	Never - 58.1%; Previous - 32.8%; Current - 8.78%	No - 21.1%; Yes - 66.1%; Not sure - 12.7%	
Endometriosis	7596 (3.38%)	53.8 (7.84)	27.5 (5.27)	0.817 (0.0701)	Never - 59.2%; Previous - 30.9%; Current - 9.47%	No - 22.6%; Yes - 36.2%; Not sure - 41.1%	ICD10 primary and secondary diagnosis and cause of death: N80[0-9] ICD9 primary and secondary diagnosis: 617[0-9] Self-reported non-cancer illness code: 1402 (endometriosis) Primary care records
Heavy menstrual bleeding	31069 (13.8%)	53.4 (7.41)	27.5 (5.45)	0.816 (0.0698)	Never - 59.2%; Previous - 30.4%; Current - 10%	No - 32.5%; Yes - 44.4%; Not sure - 23.1%	ICD10 primary and secondary diagnosis and cause of death: N92[0-6] ICD9 primary and secondary diagnosis: 626[2|3|4|5|6|8|9] Primary care records
Infertility	3509 (1.56%)	51.5 (7.64)	26.5 (5.01)	0.805 (0.069)	Never - 61.6%; Previous - 29.6%; Current - 8.46%	No - 43.3%; Yes - 42.3%; Not sure - 14.3%	ICD10 primary and secondary diagnosis and cause of death: N97[0|1|2|3|4|8|9] ICD9 primary and secondary diagnosis: 628[0|1|2|3|4|8|9] Self-reported non-cancer illness code: 1403 (female infertility) Primary care records

contd. Summary of female reproductive disorders in the UK Biobank.

Diagnosis	Number of cases (prevalence)	Mean age, years (S.D.)	Mean BMI, kg/m2 (S.D.)	Mean WHR (S.D.)	Smoking status (prevalence)	Had menopause (prevalence)	Diagnosis definition
PCOS	2058 (0.917%)	51.6 (8.14)	28.5 (6.67)	0.824 (0.0749)	Never - 60%; Previous - 30.9%; Current - 8.75%	No - 42.5%; Yes - 36.4%; Not sure - 21.1%	ICD10 primary and secondary diagnosis and cause of death: E282 ICD9 primary and secondary diagnosis: 2564 Self-reported non-cancer illness code: 1350 (polycystic ovaries / polycystic ovarian syndrome) Primary care records
Preeclampsia or eclampsia	2276 (1.01%)	54.4 (8.58)	28 (5.59)	0.821 (0.0705)	Never - 65%; Previous - 28%; Current - 6.77%	No - 34.7%; Yes - 50.2%; Not sure - 15.1%	ICD10 primary and secondary diagnosis and cause of death: O14[0|1|2|9], O15[0|1|2|9] ICD9 primary and secondary diagnosis: 642[4|5|6|7] Self-reported non-cancer illness code: 1073 (gestational hypertension / pre-eclampsia) Primary care records
Uterine fibroids	20395 (9.08%)	56.7 (7.49)	27.5 (5.23)	0.82 (0.0698)	Never - 60.8%; Previous - 31.8%; Current - 7.09%	No - 18.2%; Yes - 47%; Not sure - 34.8%	ICD10 primary and secondary diagnosis and cause of death: O14[0|1|2|9], D25[0|1|2|9] ICD9 primary and secondary diagnosis: 218[9] Self-reported non-cancer illness code: 1351/1352 (uterine fibroids / uterine polyps) Primary care records

inverse association with BMI (OR = 0.933 (0.899 - 0.966), $P = 4.52 \times 10^{-5}$). Central obesity, i.e. WHR, is also associated with increased risk of the same five reproductive conditions, with the strongest association once again with PCOS (OR = 1.23 (1.19 - 1.28), $P = 3.07 \times 10^{-21}$); however, these associations are attenuated by adjustment for BMI. There is no significant association between WHRadjBMI and endometriosis, preeclampsia, or uterine fibroids (FDR-adjusted P-value > 0.05), and the ORs for the associations of WHR with PCOS (WHR OR = 1.23 vs WHRadjBMI OR = 1.08, heterogeneity P-value $P_{\text{het}} = 8.08 \times 10^{-7}$) and heavy menstrual bleeding (WHR OR = 1.08 vs WHRadjBMI OR = 1.03, $P_{\text{het}} = 2.14 \times 10^{-8}$) are attenuated after adjustment for BMI, indicating that these relationships may be mostly driven by overall obesity.

Non-linear effects of BMI on risk of developing female reproductive conditions

I additionally found evidence of non-linear associations between BMI and five of the six female reproductive conditions studied (all but infertility), as indicated by lower AIC[548] for generalised additive models with BMI splines as compared to logistic regression models (Table 5.4). There is an inverted-U shaped relationship between BMI and endometriosis, i.e. disease prevalence is lower at both ends of the BMI spectrum; whereas the J-shaped relationship between BMI and PCOS indicates that both low and high BMI are associated with higher predicted probability of disease (Figure 5.1). The increased prevalence of HMB, preeclampsia, and uterine fibroids appears to plateau with increasing BMI. Upon adjustment for BMI, we also see evidence of a U-shaped association between WHR and PCOS, but no other non-linear effects of centralised obesity on reproductive conditions. The best fitting fractional polynomial function of obesity for each of these associations is listed in Table 5.4.

Adjustment for potential confounders in regression models

All regression models above were adjusted for age, $(\text{age})^2$, UKBB assessment centre, and self-reported smoking status (as a categorical covariate with values "Never", "Previous", or "Current"). The latter adjustment for smoking status was included

Table 5.3: Observational and genetically predicted causal effect of obesity traits on female reproductive conditions. Heterogeneity P-value is calculated for the difference between logistic regression and MR effect sizes. BMI = body mass index, CI = confidence interval, OR = odds ratio, PCOS = polycystic ovary syndrome, S.D. = standard deviation, WHR = waist-hip ratio, WHRadjBMI = WHR adjusted for BMI.

		Logistic regression		Mendelian Randomisation			
Diagnosis	**Obesity trait**	**OR (95% CI) per 1 S.D. higher obesity trait**	**P-value**	**Number of SNPs**	**OR (95% CI) per 1 S.D. higher genetically predicted obesity trait**	**P-value**	**Heterogeneity P-value**
Endometriosis	BMI	1.12 (1.1 - 1.15)	1.46E-23	274	1.04 (0.946 - 1.13)	4.50E-01	8.30E-02
	WHR	1.07 (1.04 - 1.09)	4.93E-08	195	1.1 (0.99 - 1.23)	7.53E-02	5.34E-01
	WHRadjBMI	1.01 (0.991 - 1.04)	2.30E-01	254	1.11 (1.02 - 1.2)	1.61E-02	4.52E-02
Heavy menstrual bleeding	BMI	1.11 (1.09 - 1.12)	1.51E-60	275	1.25 (1.16 - 1.35)	9.24E-09	1.89E-03
	WHR	1.08 (1.07 - 1.09)	9.47E-37	196	1.21 (1.11 - 1.32)	1.52E-05	1.17E-02
	WHRadjBMI	1.03 (1.02 - 1.04)	2.35E-07	254	1.12 (1.04 - 1.2)	2.14E-03	3.06E-02
Infertility (all causes)	BMI	0.933 (0.899 - 0.966)	4.52E-05	275	0.955 (0.902 - 1.01)	1.19E-01	4.86E-01
	WHR	0.966 (0.933 - 0.999)	4.23E-02	196	1.07 (0.999 - 1.14)	5.30E-02	8.71E-03
	WHRadjBMI	0.997 (0.963 - 1.03)	8.43E-01	254	1.1 (1.05 - 1.16)	1.71E-04	1.11E-03
Infertility (anatomical)	BMI			274	1.06 (0.923 - 1.21)	4.14E-01	
	WHR			196	1.09 (0.964 - 1.24)	1.62E-01	
	WHRadjBMI			254	1.13 (1.01 - 1.25)	2.90E-02	
Infertility (anovulatory)	BMI			274	1.05 (0.927 - 1.19)	4.44E-01	
	WHR			195	1.32 (1.15 - 1.52)	7.95E-05	
	WHRadjBMI			251	1.29 (1.16 - 1.45)	4.66E-06	

contd. Observational and genetically predicted causal effect of obesity traits on female reproductive conditions.

		Logistic regression		Mendelian Randomisation			
Diagnosis	**Obesity trait**	**OR (95% CI) per 1 S.D. higher obesity trait**	**P-value**	**Number of SNPs**	**OR (95% CI) per 1 S.D. higher genetically predicted obesity trait**	**P-value**	**Heterogeneity P-value**
Infertility (idiopathic, exclusion)	BMI			275	0.928 (0.869 - 0.99)	2.35E-02	
	WHR			195	0.968 (0.901 - 1.04)	3.65E-01	
	WHRadjBMI			253	1.03 (0.971 - 1.09)	3.21E-01	
Infertility (idiopathic, inclusion)	BMI			274	0.866 (0.781 - 0.962)	7.05E-03	
	WHR			195	1.08 (0.956 - 1.22)	2.12E-01	
	WHRadjBMI			253	1.13 (1.03 - 1.24)	1.05E-02	
PCOS	BMI	1.3 (1.26 - 1.35)	2.01E-32	275	1.92 (1.59 - 2.32)	1.77E-11	8.19E-05
	WHR	1.23 (1.19 - 1.28)	3.07E-21	196	1.54 (1.25 - 1.89)	4.67E-05	4.05E-02
	WHRadjBMI	1.08 (1.03 - 1.12)	7.10E-04	254	1.26 (1.07 - 1.48)	5.47E-03	7.00E-02
Preeclampsia or eclampsia	BMI	1.24 (1.2 - 1.29)	4.68E-25	275	1.67 (1.48 - 1.89)	1.56E-16	4.88E-06
	WHR	1.13 (1.09 - 1.17)	4.33E-09	196	1.39 (1.21 - 1.6)	3.38E-06	5.03E-03
	WHRadjBMI	1.02 (0.981 - 1.06)	3.05E-01	254	1.26 (1.13 - 1.41)	3.71E-05	4.32E-04
Uterine fibroids	BMI	1.11 (1.1 - 1.13)	2.25E-46	275	1.08 (0.999 - 1.17)	5.23E-02	5.33E-01
	WHR	1.06 (1.05 - 1.07)	3.65E-15	196	1.19 (1.1 - 1.3)	2.91E-05	5.63E-03
	WHRadjBMI	1.01 (0.999 - 1.03)	6.28E-02	254	1.16 (1.08 - 1.24)	2.28E-05	1.73E-04

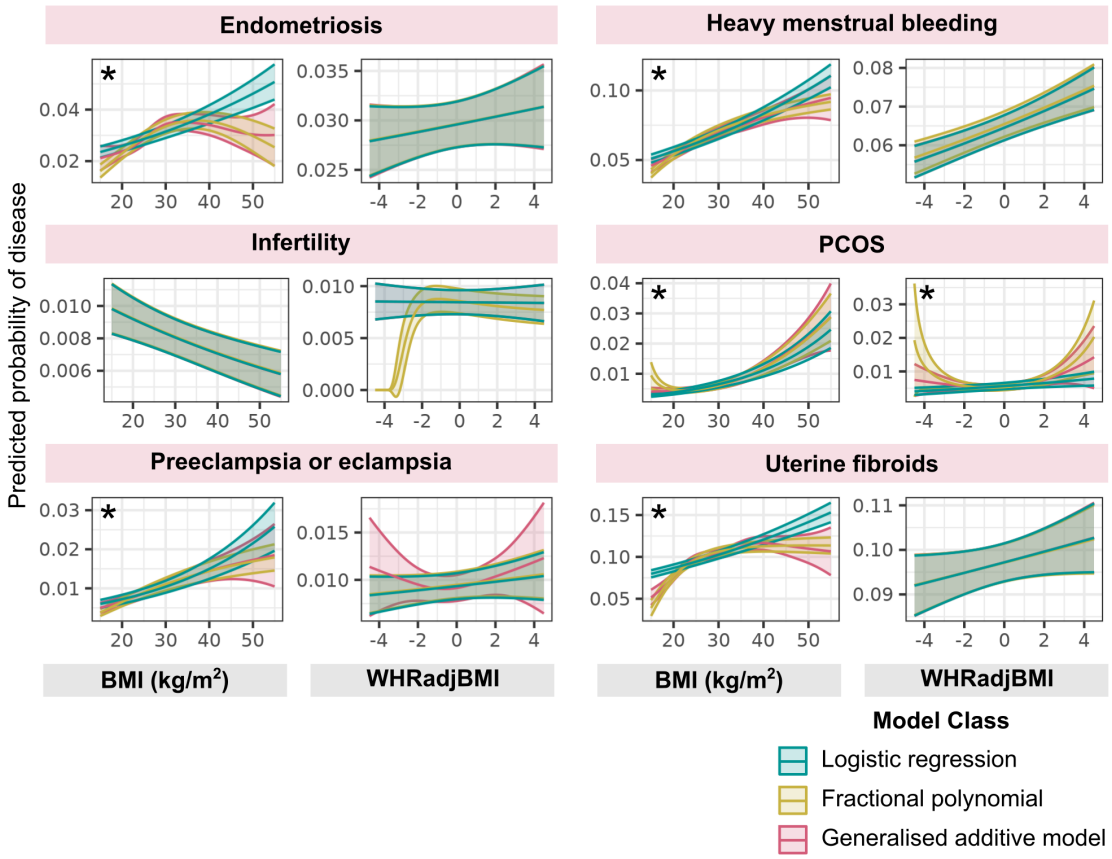


Figure 5.1: Relationship between obesity traits and female reproductive conditions in UK Biobank. Logistic regression, fractional polynomial, and generalised additive models were fitted to estimate the probability of developing female reproductive disorders as a function of obesity-related traits in UK Biobank. Mean predicted fits estimated by logistic and best-fitting non-linear models, and 95% confidence intervals about the mean are displayed. Asterisks indicate that non-linear models fit the data better than linear models (as evaluated with Akaike information criterion[548]). All models were adjusted for age, age-squared, assessment centre, and smoking status. BMI = body mass index, PCOS = polycystic ovary syndrome, WHRadjBMI, waist-to-hip ratio adjusted for BMI.

to protect against confounding of the association between obesity and female reproductive disorders; however, we find that there is no significant difference in the direction or strength of effects estimated with and without adjustment for smoking status for the associations of BMI with five of six reproductive outcomes (heterogeneity P-value $P_{\text{het}} > 0.05$) (Figure 5.3 and Table D.2). The association between BMI and PCOS is strengthened upon adjusting for smoking status (smoking-unadjusted OR = 1.24, $P = 1.16 \times 10^{-21}$, smoking-adjusted OR = 1.30, $P = 2.01 \times 10^{-32}$; $P_{\text{het}} = 0.0376$), indicating that smoking may be a confounder of this

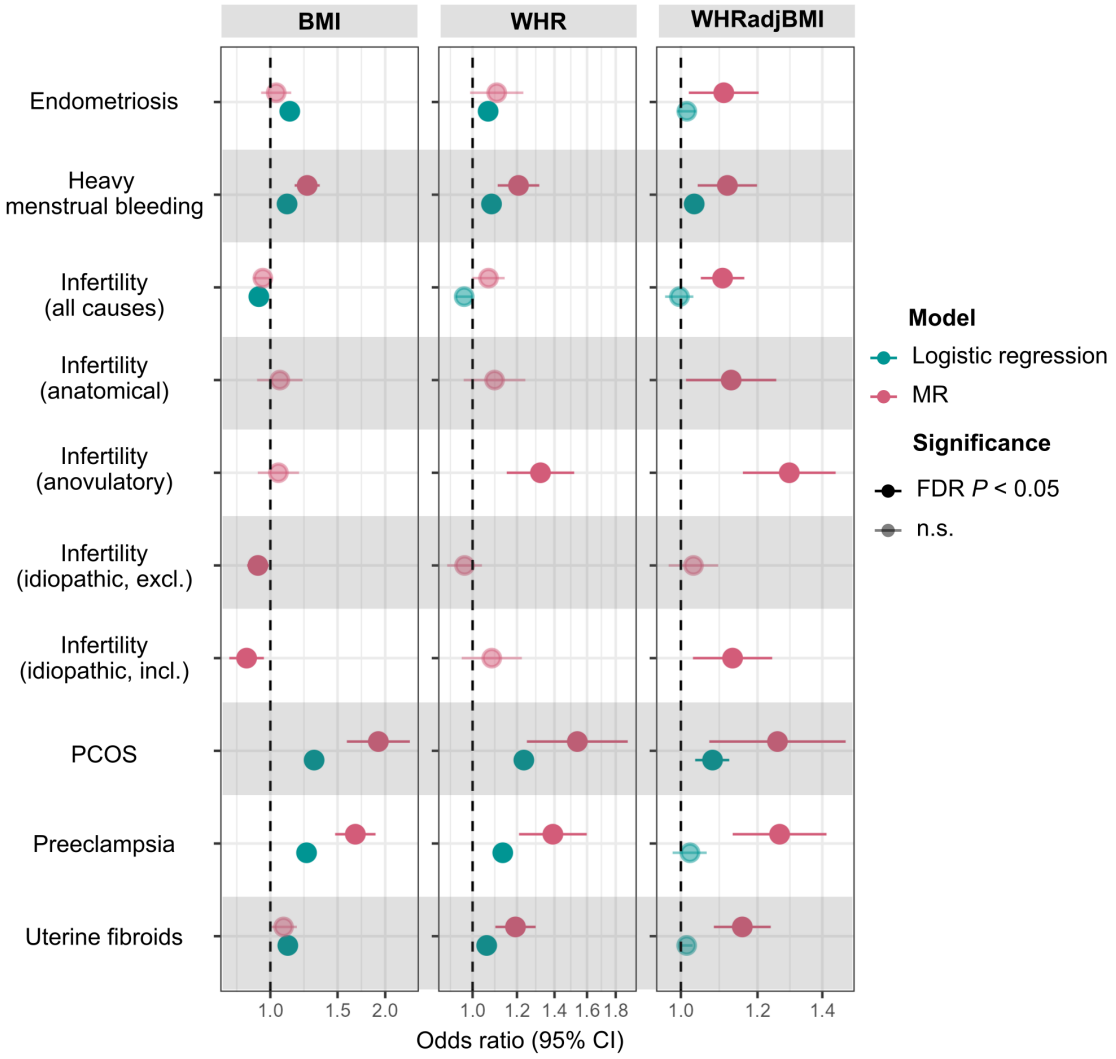


Figure 5.2: Comparison of observational and genetically predicted causal effect sizes of obesity traits on female reproductive conditions. Observational effects were estimated by fitting logistic regression models in the UK Biobank (UKBB) adjusted for age, $(\text{age})^2$, assessment centre, and smoking status, while causal effects were estimated using two-sample Mendelian Randomisation (MR) with the inverse variance weighted method, from meta-analysed genome-wide association summary statistics for obesity traits and female reproductive conditions. Odds ratio and 95% confidence intervals (CIs) are displayed. Solid-coloured points represent associations with false discovery rate (FDR)-adjusted P-value < 0.05, while faded points represent non-significant (n.s.) P-values > 0.05 after FDR adjustment for 18 tests (observational) and 30 tests (MR); the discrepancy in number of tests performed is due to the lack of sufficient sample size for logistic regressions with different categories of infertility in UKBB. BMI = body mass index, Infertility idiopathic excl. = idiopathic infertility defined by exclusion, Infertility idiopathic incl. = idiopathic infertility defined by inclusion, PCOS = polycystic ovary syndrome, WHR = waist-to-hip ratio, WHRadjBMI, waist-to-hip ratio adjusted for BMI.

Table 5.4: Comparison of linear and non-linear models testing the association of obesity traits with female reproductive conditions in UK Biobank. AIC = Akaike information criterion[548], BMI = body mass index, PCOS = polycystic ovary syndrome, WHR = waist-hip ratio, WHRadjBMI = WHR adjusted for BMI.

Diagnosis	Obesity trait	Logistic regression	Generalised additive model	Fractional Polynomial	
		AIC	AIC	AIC	Terms in best-fitting model
Endometriosis	BMI	64684.76	64643.59	64673.00	$x(1 + \log(x))$
	WHR	64739.28	64739.29	64763	x
	WHRadjBMI	64767.88	64768.00	64792	x
Heavy menstrual bleeding	BMI	165182.40	165165.40	165974.00	$\frac{1}{x}$
	WHR	165318.00	165313.70	166142.00	$\frac{1}{x^2}$
	WHRadjBMI	165465.90	165465.90	166304.00	x
Infertility (all causes)	BMI	33552.70	33552.76	33556.00	x
	WHR	33564.25	33564.25	33567.00	x
	WHRadjBMI	33568.13	33568.14	33563.00	$\frac{1 + \log(x)}{x^2}$
PCOS	BMI	21640.67	21632.36	21643.00	$\frac{1 + \log(x)}{x}$
	WHR	21739.61	21739.61	21754.00	x
	WHRadjBMI	21833.38	21825.26	21835.00	$x^2 + x^{0.5}$
Preeclampsia or eclampsia	BMI	24827.77	24822.94	24865.00	$\frac{1}{x}$
	WHR	24888.34	24888.34	24933.00	x
	WHRadjBMI	24922.26	24920.01	24967.00	x
Uterine fibroids	BMI	135028.70	134970.00	135159.00	$\frac{1 + \log(x)}{x^2}$
	WHR	135139.60	135139.60	135336.00	x
	WHRadjBMI	135198.60	135198.60	135397.00	x

association. While less is known about the causal relationship between smoking and abdominal adiposity, the associations between WHR or WHRadjBMI and endometriosis, HMB, PCOS, and preeclampsia all significantly increase in magnitude when adjusted for smoking status (Figure 5.3 and Table D.2). Finally, adjustment for smoking attenuates the effect of WHR with infertility (smoking-unadjusted OR = 0.841, $P = 2.53 \times 10^{-24}$, smoking-adjusted OR = 0.966, $P = 0.0423$; $P_{\text{het}} = 2.48 \times 10^{-7}$) and WHRadjBMI with infertility (smoking-unadjusted OR = 0.872, $P = 8.80 \times 10^{-16}$, smoking-adjusted OR = 0.997, $P = 0.843$; $P_{\text{het}} = 2.73 \times 10^{-7}$), suggesting that smoking may mediate or confound the associations of centralised obesity with infertility.

Menopause status may also be a potential confounder of the association between obesity and reproductive disease, as there are distinct changes in body fat distribution around menopause[180,222], whose timing may be affected by the history or

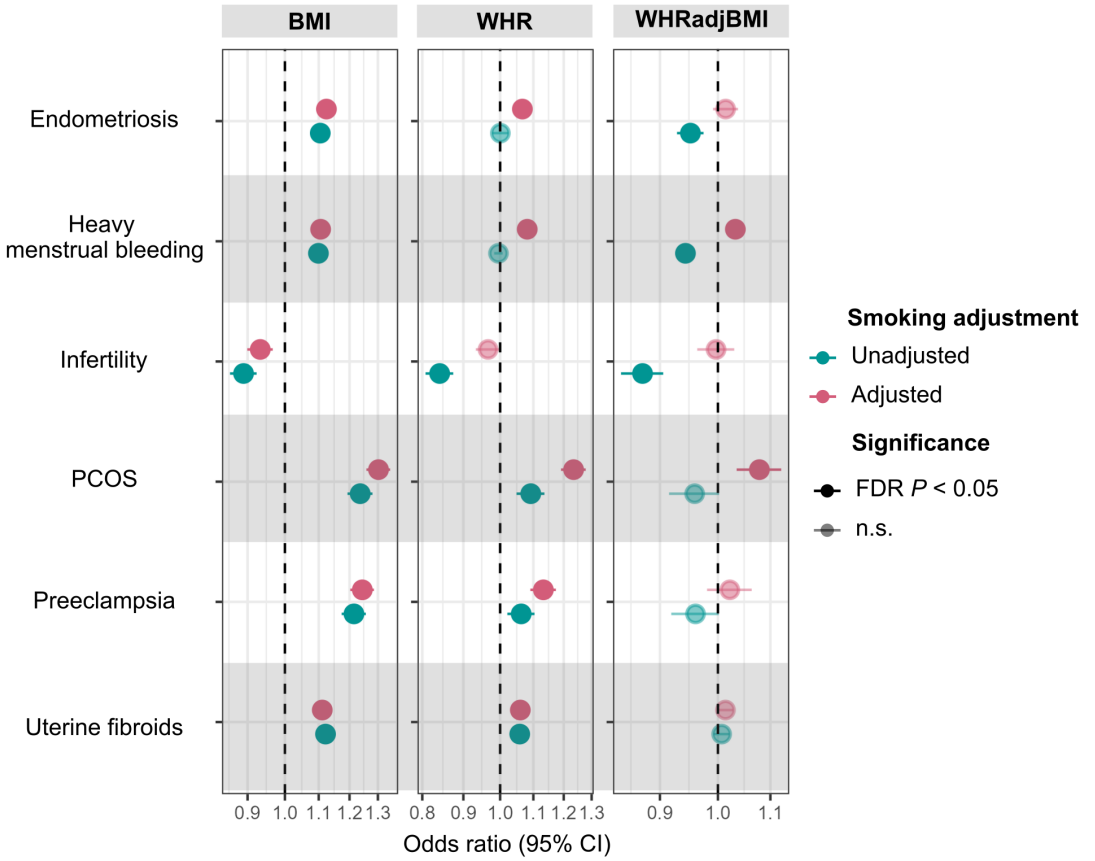


Figure 5.3: Effect sizes of obesity traits on female reproductive conditions in UK Biobank (UKBB) with and without adjustment for smoking status. Logistic regression models were fitted to estimate the probability of developing six female reproductive disorders as a function of body mass index (BMI), waist-to-hip ratio (WHR), and WHR adjusted for BMI (WHRadjBMI). Two sets of models were fitted, one including smoking status as a covariate ("adjusted") and the other without ("unadjusted"). All models included age, $(\text{age})^2$, and UKBB assessment centre as covariates. Odds ratio and 95% confidence intervals (CIs) are plotted. Solid-coloured points represent associations with false discovery rate (FDR)-adjusted P-value < 0.05, while faded points represent non-significant (n.s.) P-values > 0.05 after FDR adjustment for 18 tests. PCOS = polycystic ovary syndrome.

presence of a female reproductive condition[568–570]. However, we did not include menopause status as a covariate in our analyses as: (1) timing of menopause is correlated with age, which we already adjust for in our models; and (2) women with diagnoses of reproductive diseases in the UKBB are up to 3 times more uncertain of having reached menopause (for example, due to hysterectomy) than women without these diseases (proportion of women reporting "Not sure" of menopause status = 12.4% for women without a female reproductive condition, 41.1% for women with

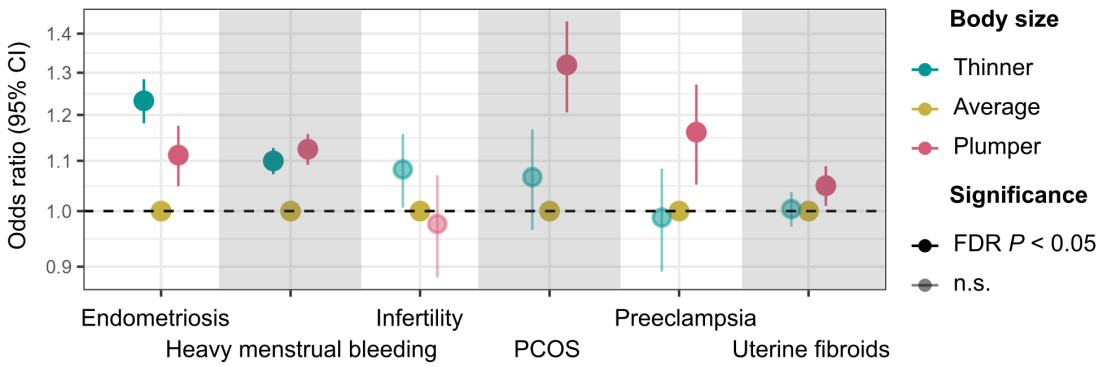


Figure 5.4: Effect of early-life body size on risk of female reproductive conditions in UK Biobank Logistic regression models were fitted to estimate the probability of developing six female reproductive disorders as a function of self-reported body size at age 10 years. Body size was treated as a categorical variable with levels: "average" (reference category), "thinner" and "plumper", as reported by participants. Odds ratio and 95% confidence intervals are displayed. Solid-coloured points represent associations with false discovery rate (FDR)-adjusted P-value < 0.05, while faded points represent non-significant (n.s.) P-values > 0.05 after FDR adjustment for six tests. PCOS = polycystic ovary syndrome.

endometriosis, and 34.8% for women with uterine fibroids) (Table 5.2).

Association of childhood body size with development of female reproductive conditions later in life

BMI and WHR were measured at UKBB assessment centres at recruitment (age 40-69 years), often several decades after the diagnosis of many of the female reproductive disorders studied here. We therefore estimated the effect of comparative childhood body size, which was self-reported by participants in the UKBB as “thinner”, “average”, or “plumper” at age 10 years, on the development of reproductive conditions later in life. We find that both being thinner than average and plumper than average in childhood increases risk of endometriosis (thinner OR $= 1.23$, $P = 1.70 \times 10^{-15}$, plumper OR $= 1.11$, $P = 1.08 \times 10^{-3}$) and HMB (thinner OR $= 1.10$, $P = 1.04 \times 10^{-11}$, plumper OR $= 1.12$, $P = 2.33 \times 10^{-12}$) (Figure 5.4). However, only being plumper than average in childhood poses significant risk for the development of PCOS (OR $= 1.32$, $P = 1.65 \times 10^{-6}$), preeclampsia or eclampsia (OR $= 1.16$, $P = 7.55 \times 10^{-3}$), and uterine fibroids (OR $= 1.05$, $P = 1.67 \times 10^{-2}$) later in life.

5.3.2 Body fat distribution is genetically associated with altered risk of female reproductive conditions

Heterogeneous causal effects of overall and centralised obesity on risk of developing female reproductive conditions

Centralised obesity, adjusted for BMI, is a causal risk factor for the development of nine out of ten reproductive outcomes analysed here (all but F-EXCL), with genetically predicted causal effect sizes larger than observational effects estimated in the UKBB (all heterogeneity P-values $P_{\text{het}} < 0.05$) (Figure 5.2 and Table 5.3). WHRadjBMI is most strongly associated with F-ANOV (OR (95% CI) per SD increase in genetically predicted WHRadjBMI = 1.29 (1.16 - 1.45), $P = 4.66 \times 10^{-6}$) and PCOS (OR = 1.26 (1.07 - 1.48), $P = 0.00547$). Without adjustment for overall obesity, WHR is still associated with increased risk of five out of these nine conditions – PCOS (OR = 1.54 (1.25 - 1.89), $P = 4.67 \times 10^{-5}$), preeclampsia or eclampsia (OR = 1.39 (1.21 - 1.60), $P = 3.38 \times 10^{-6}$), F-ANOV (OR = 1.32 (1.15 - 1.52), $P = 7.95 \times 10^{-5}$), HMB (OR = 1.21 (1.11 - 1.32), $P = 1.52 \times 10^{-5}$), and uterine fibroids (OR = 1.19 (1.10 - 1.30), $P = 2.91 \times 10^{-5}$) – once again with causal effect sizes larger than their observationally predicted counterparts in UKBB (all heterogeneity P-values $P_{\text{het}} < 0.05$). There is no genetically predicted protective effect of increased WHR or WHRadjBMI on any of the female reproductive conditions studied here.

However, genetically predicted BMI is protective against the development of idiopathic infertility (F-INCL OR = 0.866 (0.781 - 0.962), $P = 0.00705$, F-EXCL OR = 0.928 (0.869 - 0.990), $P = 0.0235$); and in the UKBB, BMI is observationally inversely associated with infertility of all causes (OR = 0.933 (0.899 - 0.966), $P = 4.52 \times 10^{-5}$). Unlike with centralised obesity, we find no genetically predicted causal effect of BMI on endometriosis and uterine fibroids (all FDR-adjusted P-values > 0.05); but we do observe that higher BMI increases risk of PCOS (OR = 1.92 (1.59 - 2.32), $P = 1.77 \times 10^{-11}$), preeclampsia or eclampsia (OR = 1.67 (1.48 - 1.89), $P = 1.56 \times 10^{-16}$), and heavy menstrual bleeding ((OR = 1.25 (1.16 - 1.35), $P = 9.24 \times 10^{-9}$). As with WHR and WHRadjBMI, the causal effect of

BMI on these reproductive conditions is significantly larger than the observational association (all heterogeneity P-values $P_{het} < 0.05$) (Figure 5.2 and Table 5.3).

All genetically predicted causal effects were estimated using two-sample MR with female-specific obesity instruments derived from the meta-analysis released by the GIANT Consortium[216] of up to 806,801 individuals of European ancestry (Table 5.1) and female reproductive outcome summary statistics derived from meta-analyses of European-ancestry GWASs in UKBB[354], FinnGen[129], and publicly available summary statistics (Table D.1, see Methods Section 5.2.2 for details).

Dissecting the causal effect of centralised obesity on risk of female reproductive conditions

Given the consistent associations between centralised body fat and most female reproductive conditions, I attempted to dissect these further by using genetic instruments for waist circumference and hip circumference[354], and instruments for predicted VAT mass[549]. The latter is significantly associated with increased risk of PCOS (OR (95% CI) per 1 kg increase in predicted VAT mass = 2.51 (1.76 - 3.57), $P = 3.34 \times 10^{-7}$) and preeclampsia or eclampsia (OR = 1.87 (1.32 - 2.66), $P = 4.65 \times 10^{-4}$), but is protective against the development of endometriosis (OR = 0.771 (0.631 - 0.943), $P = 0.0112$), F-EXCL (OR = 0.808 (0.684 - 0.956), $P = 0.0127$), and F-ALL (OR = 0.829 (0.717 - 0.960), $P = 0.0123$). The direction of effects of centralised obesity traits on each female reproductive disorder is consistent, i.e. WC, HC, and predicted VAT mass are all associated with increased risk of HMB, PCOS, and preeclampsia, decreased risk of endometriosis, F-ALL, and F-INCL, and have no effect on uterine fibroids, F-ANAT, and F-ANOV (Table 5.5 and Figure 5.5).

Increased waist circumference poses a significantly stronger risk than does hip circumference for the risk of developing PCOS (WC OR = 2.05 (1.69 - 2.48), HC OR = 1.51 (1.29 - 1.77), heterogeneity P-value $P_{het} = 0.0153$) and HMB (WC OR = 1.27 (1.18 - 1.37), HC OR = 1.12 (1.05 - 1.19), $P_{het} = 0.00980$) (Table 5.5 and Figure 5.5). Furthermore, we only observed a protective effect of increased hip circumference, but not waist circumference, on the development of infertility of all causes (OR (95% CI) per 1 SD increase in HC = 0.933 (0.889 - 0.979), $P = 0.00454$)

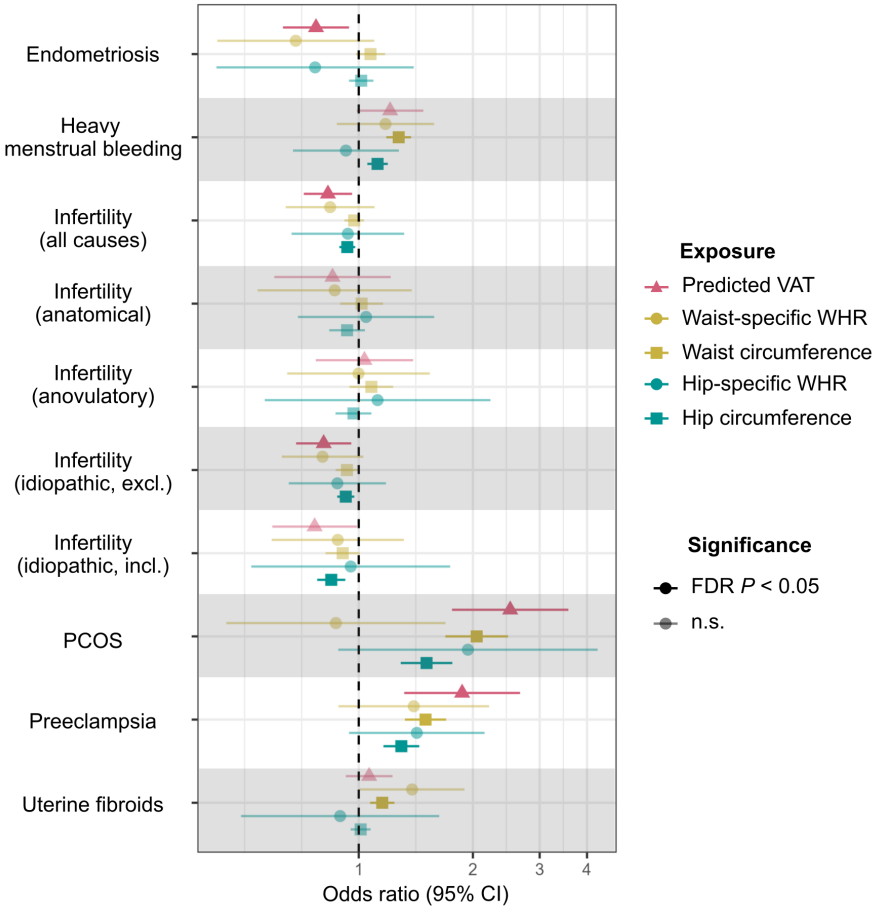


Figure 5.5: Genetically predicted causal effects of centralised obesity traits on risk of developing female reproductive conditions. Genetic instruments for predicted visceral adipose tissue (VAT) mass, waist circumference, hip circumference, waist-specific waist-to-hip ratio (WHR), and hip-specific WHR were derived from recent European-ancestry genome-wide association studies. Odds ratios and 95% confidence intervals (CIs) for causal effects, estimated using two-sample Mendelian Randomisation (MR) with the inverse variance weighted method, are displayed. Solid-coloured points represent associations with false discovery rate (FDR)-adjusted P-value < 0.05, while faded points represent non-significant (n.s.) P-values > 0.05 after FDR adjustment for 50 tests. Infertility idiopathic excl. = idiopathic infertility defined by exclusion, Infertility idiopathic incl. = idiopathic infertility defined by inclusion, PCOS = polycystic ovary syndrome.

Table 5.5: Genetically predicted causal effects of centralised obesity traits on risk of developing female reproductive conditions. HC = hip circumference, PCOS = polycystic ovary syndrome, WC = waist circumference.

Diagnosis	Obesity trait	Number of SNPs	OR (95% CI) per 1 S.D. higher genetically predicted obesity trait	P-value
Endometriosis	HC	643	1.01 (0.943 - 1.09)	6.93E-01
	Hip-specific WHR	22	0.767 (0.422 - 1.4)	3.86E-01
	WC	593	1.07 (0.983 - 1.17)	1.15E-01
	Waist-specific WHR	36	0.682 (0.424 - 1.1)	1.15E-01
	Predicted VAT	22	0.771 (0.631 - 0.943)	1.12E-02
Heavy menstrual bleeding	HC	648	1.12 (1.05 - 1.19)	3.04E-04
	Hip-specific WHR	22	0.925 (0.671 - 1.28)	6.35E-01
	WC	595	1.27 (1.18 - 1.37)	2.45E-10
	Waist-specific WHR	36	1.18 (0.876 - 1.58)	2.81E-01
	Predicted VAT	22	1.21 (0.989 - 1.48)	6.37E-02
Infertility (all causes)	HC	645	0.933 (0.889 - 0.979)	4.54E-03
	Hip-specific WHR	22	0.936 (0.665 - 1.32)	7.05E-01
	WC	594	0.973 (0.916 - 1.03)	3.64E-01
	Waist-specific WHR	36	0.841 (0.642 - 1.1)	2.07E-01
	Predicted VAT	22	0.829 (0.717 - 0.96)	1.23E-02
Infertility (anatomical)	HC	644	0.932 (0.836 - 1.04)	2.00E-01
	Hip-specific WHR	22	1.05 (0.691 - 1.58)	8.33E-01
	WC	592	1.02 (0.894 - 1.16)	7.93E-01
	Waist-specific WHR	36	0.864 (0.541 - 1.38)	5.41E-01
	Predicted VAT	22	0.853 (0.599 - 1.21)	3.77E-01
Infertility (anovulatory)	HC	642	0.968 (0.868 - 1.08)	5.64E-01
	Hip-specific WHR	22	1.12 (0.565 - 2.23)	7.43E-01
	WC	587	1.08 (0.945 - 1.23)	2.58E-01
	Waist-specific WHR	35	0.999 (0.648 - 1.54)	9.95E-01
	Predicted VAT	22	1.04 (0.77 - 1.39)	8.18E-01
Infertility (idiopathic, exclusion)	HC	645	0.924 (0.877 - 0.974)	2.98E-03
	Hip-specific WHR	22	0.878 (0.654 - 1.18)	3.89E-01
	WC	594	0.931 (0.87 - 0.996)	3.87E-02
	Waist-specific WHR	36	0.803 (0.627 - 1.03)	8.15E-02
	Predicted VAT	22	0.808 (0.684 - 0.956)	1.27E-02
Infertility (idiopathic, inclusion)	HC	642	0.846 (0.777 - 0.922)	1.25E-04
	Hip-specific WHR	22	0.952 (0.521 - 1.74)	8.74E-01
	WC	589	0.907 (0.817 - 1.01)	6.64E-02
	Waist-specific WHR	35	0.88 (0.589 - 1.31)	5.34E-01
	Predicted VAT	22	0.765 (0.592 - 0.989)	4.07E-02

contd. Genetically predicted causal effects of centralised obesity traits on risk of developing female reproductive conditions.

Diagnosis	Obesity trait	Number of SNPs	OR (95% CI) per 1 S.D. higher genetically predicted obesity trait	P-value
PCOS	Hip-specific WHR	22	1.94 (0.883 - 4.26)	9.88E-02
	WC	594	2.05 (1.69 - 2.48)	1.57E-13
	Waist-specific WHR	36	0.87 (0.447 - 1.69)	6.83E-01
	Predicted VAT	22	2.51 (1.76 - 3.57)	3.34E-07
Preeclampsia or eclampsia	HC	648	1.3 (1.16 - 1.44)	3.15E-06
	Hip-specific WHR	22	1.42 (0.943 - 2.15)	9.29E-02
	WC	595	1.5 (1.32 - 1.7)	2.18E-10
	Waist-specific WHR	36	1.4 (0.884 - 2.21)	1.52E-01
	Predicted VAT	22	1.87 (1.32 - 2.66)	4.65E-04
Uterine fibroids	HC	648	1.01 (0.952 - 1.07)	7.09E-01
	Hip-specific WHR	22	0.893 (0.489 - 1.63)	7.13E-01
	WC	595	1.15 (1.07 - 1.24)	1.54E-04
	Waist-specific WHR	36	1.38 (1.01 - 1.9)	4.56E-02
	Predicted VAT	22	1.07 (0.924 - 1.23)	3.83E-01

and idiopathic infertility (F-INCL OR = 0.846 (0.777 - 0.922), $P = 1.25 \times 10^{-4}$; F-EXCL OR = 0.924 (0.877 - 0.974), $P = 0.00298$). Although we attempted to examine the heterogeneity between HC and WC effects further, using instruments with waist-specific and hip-specific effects on WHR[551], the power to detect an effect size (OR) of > 1.5 at significance level 5% with these instruments was only between 7% to 20%, due to the 10x lower proportion of variance explained by waist-specific and hip-specific WHR instruments ($R^2 = 0.4\%$) as compared to WHR itself ($R^2 = 4\%$) (Table D.3).

Limited power for non-linear MR

Five of the six female reproductive disorders tested in UKBB, i.e. all but infertility, displayed evidence for non-linear associations with BMI (Figure 5.1 and Table 5.4), so I tested whether there was genetic evidence for non-linearity in the causal effect of BMI on these conditions in a sample of 224,278 unrelated women of European ancestry in UKBB. I constructed a BMI instrument from the 2015 release of GIANT[552] with 97 variants that together explain 1.69% of variance in BMI. However, despite restricting the outcomes to HMB and uterine fibroids, which occur

at > 5% prevalence in this population (Table 5.2), we were still under-powered to detect localised average causal effects in each BMI quantile (FDR-adjusted P-value > 0.05 for both sets of associations).

Reported results are robust to a range of modelling choices

Genetic instruments from female-specific GWASs with female-specific weights were found to be the strongest instruments, as measured by F-statistics, for WHR (F_{female} = 74.8, $F_{sex\text{-}combined}$ = 59.8), and WHRadjBMI (F_{female} = 84.7, $F_{sex\text{-}combined}$ = 71.9) (Table 5.1). However, due to concerns of sex-differential participation bias in the UKBB, which has an over-representation of healthy female participants compared to the general population of the UK[348,356], I report results from analyses with sex-combined instruments (weighted by sex-combined effects, as well as female-specific effects) in Figure 5.6. We find significant associations (FDR-adjusted P-value for 30 tests < 0.05) between obesity trait × reproductive conditions for 19/30 pairs tested when using female-specific, female-weighted instruments, of which 18/19 are also significant when using another weighting strategy (Figure 5.6).

I prioritised two-sample MR results from the IVW method over the less-powered methods of MR-Egger, which is robust to horizontal pleiotropy[563], and weighted median, which is robust to outlier variants[571], as there was no evidence of directional pleiotropy for 23/34 tests for which we found significant associations at FDR-adjusted P-value < 0.05 (Table D.4). For the 11 exposure × outcome pairs with significant pleiotropy (MR-Egger intercept P-value < 0.05), we nevertheless observed directional consistency between all methods tested (Figure 5.7).

As sample overlap of UKBB participants between the exposure and outcome GWAS sources may bias MR results[572], I compared the main results reported here to an analysis using female-specific, female-weighted instruments for BMI, WHR, and WHRadjBMI from a previous release of the GIANT Consortium without UKBB participants[552]. Twelve of the 19 exposure × outcome pairs that reached significance in the main analysis also reached significance when using genetic instruments derived

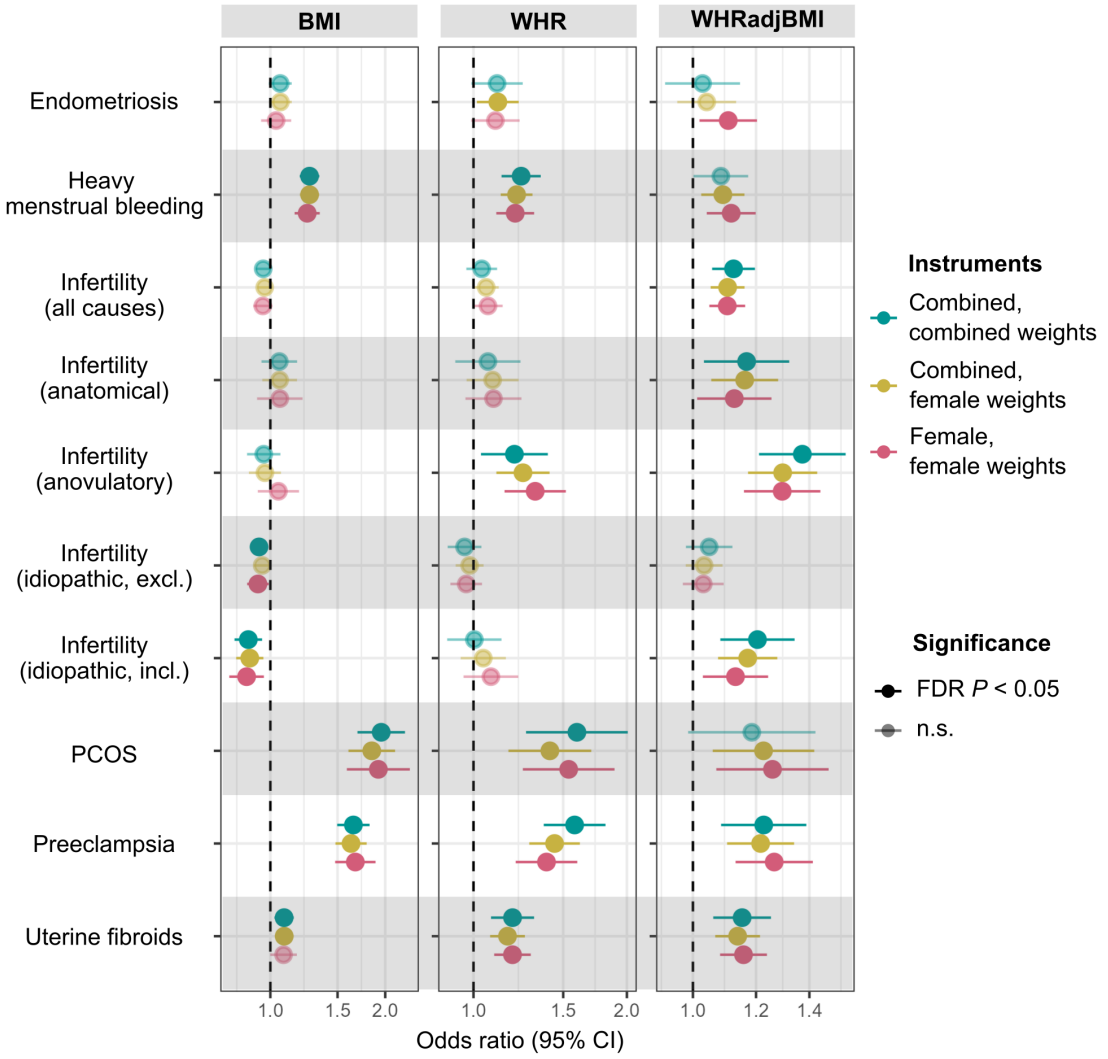


Figure 5.6: Comparison of MR results with different instrument selection strategies. Genetic instruments for body mass index (BMI), waist-to-hip ratio (WHR), and WHR adjusted for BMI (WHRadjBMI) were constructed using index variants from: (i) combined-sexes GWAS, weighted by combined-sex effect size, (ii) combined-sexes GWAS, weighted by female effect size, or (iii) female-specific GWAS, weighted by female effect size, all reported by Pulit *et al.* (2019)[216]. Causal effects were estimated using two-sample Mendelian Randomisation (MR) with the inverse variance weighted method. Odds ratio and 95% confidence intervals (CIs) are displayed. Solid-coloured points represent associations with false discovery rate (FDR)-adjusted P-value < 0.05, while faded points represent non-significant (n.s.) P-values > 0.05 after FDR adjustment for 30 tests. Infertility idiopathic excl. = idiopathic infertility defined by exclusion, Infertility idiopathic incl. = idiopathic infertility defined by inclusion, PCOS = polycystic ovary syndrome.

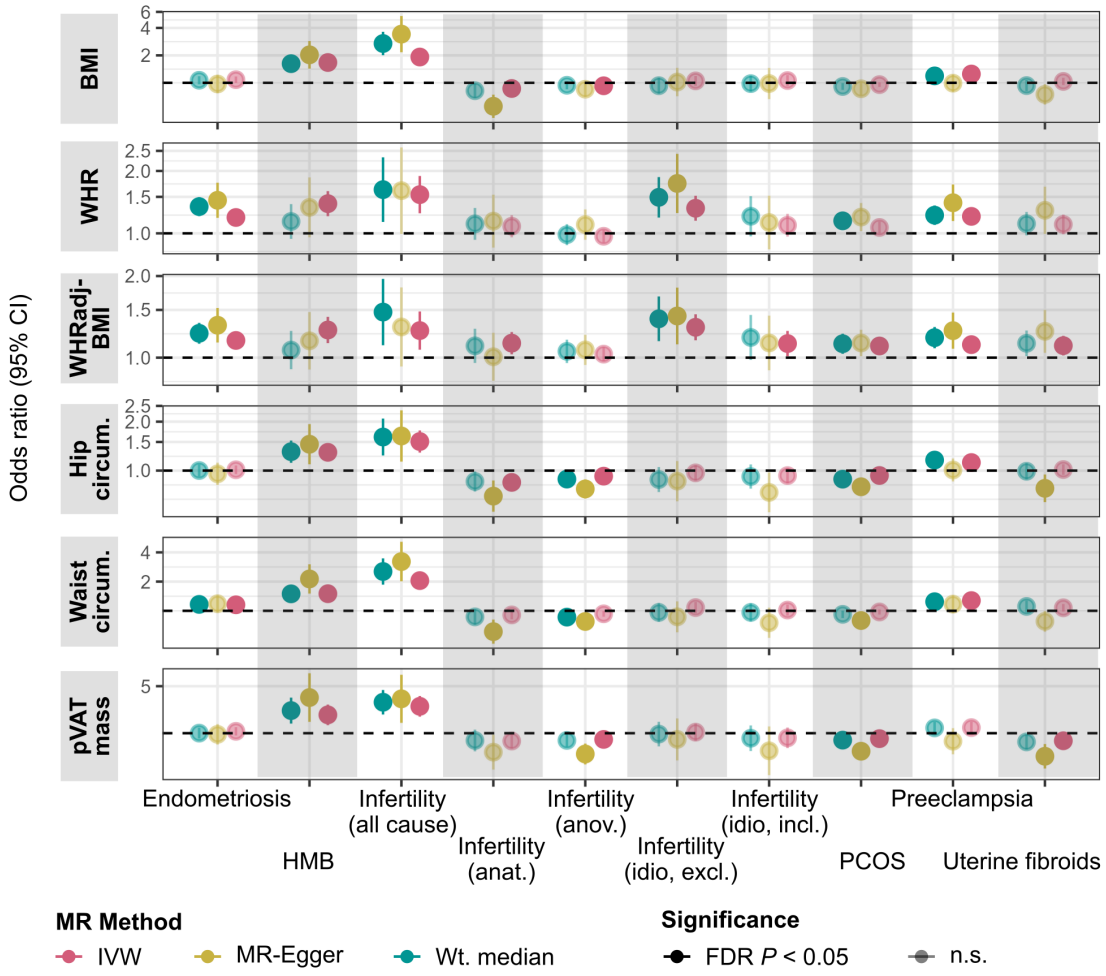


Figure 5.7: Comparison of inverse variance weighted (IVW), MR-Egger and weighted median methods for estimating causal effect sizes. Causal effects were estimated using three different two-sample MR methods from meta-analysed genome-wide association summary statistics for obesity traits and female reproductive conditions. Odds ratio and 95% confidence intervals (CIs) are displayed. Solid-coloured points represent associations with false discovery rate (FDR)-adjusted P-value < 0.05, while faded points represent non-significant (n.s.) P-values > 0.05 after FDR adjustment for 30 tests. BMI, body mass index, circum = circumference, Infertility anat. = anatomical infertility, Infertility anov. = anovulatory infertility, Infertility idio, excl. = idiopathic infertility defined by exclusion, Infertility idio, incl. = idiopathic infertility defined by inclusion, PCOS = polycystic ovary syndrome, pVAT = predicted visceral adipose tissue, WHR = waist-to-hip ratio, WHRadjBMI, waist-to-hip ratio adjusted for BMI, Wt. median = weighted median.

from GWASs without UKBB participants, and the other seven show consistent effect directions with the main results (Figure 5.8).

Finally, there are potential concerns of collider bias in the WHRadjBMI instrument, arising from the adjustment of the WHR GWAS for the heritable covariate of BMI, which may share residual confounders with WHR[573]. We therefore performed multivariable MR using a joint instrument for WHR and BMI to estimate the BMI-adjusted effect of WHR on female reproductive conditions. We find no difference between causal effects estimated using a WHRadjBMI instrument and one from a multivariable instrument (all heterogeneity P-values $P_{\text{het}} > 0.05$, Table 5.6).

Limited evidence for reciprocal causal effects of female reproductive conditions on obesity

We established the direction of genetic causality between obesity traits and seven female reproductive disorders through reciprocal MR by constructing genetic instruments for reproductive traits (Table 5.1; instruments for F-ANAT, F-INCL, and F-EXCL were not well-powered for analysis). We found no evidence for reciprocal effects of six out of seven reproductive conditions on increased obesity (FDR-adjusted P-value for 21 tests > 0.05), but some indication that incidence of anovulatory infertility is genetically causal for increased WHR (β (SE) = 0.0430 (0.0133), $P = 0.00119$) and WHRadjBMI ($\beta = 0.0547$ (0.0133), $P = 3.74 \times 10^{-5}$) (Figure 5.9).

5.3.3 Hormones mediate the effects of genetically predicted obesity on female reproductive disorders

Reproductive and metabolic hormones are genetically related to the development of female reproductive conditions

We observe protective effects of FSH levels against the development of PCOS (OR per 1 SD increased FSH = 0.281 (0.116 - 0.680), $P = 0.00488$), preeclampsia (OR = 0.662 (0.492 - 0.893), $P = 0.00685$), and infertility (F-ALL OR = 0.776 (0.678 - 0.888), $P = 2.15 \times 10^{-4}$; F-EXCL OR = 0.716 (0.604 - 0.850), $P = 1.26 \times 10^{-4}$); further, increased levels of testosterone are protective against the development of F-ANAT (OR = 0.843 (0.740 - 0.959), $P = 0.00938$) (Figure 5.10). On the other hand,

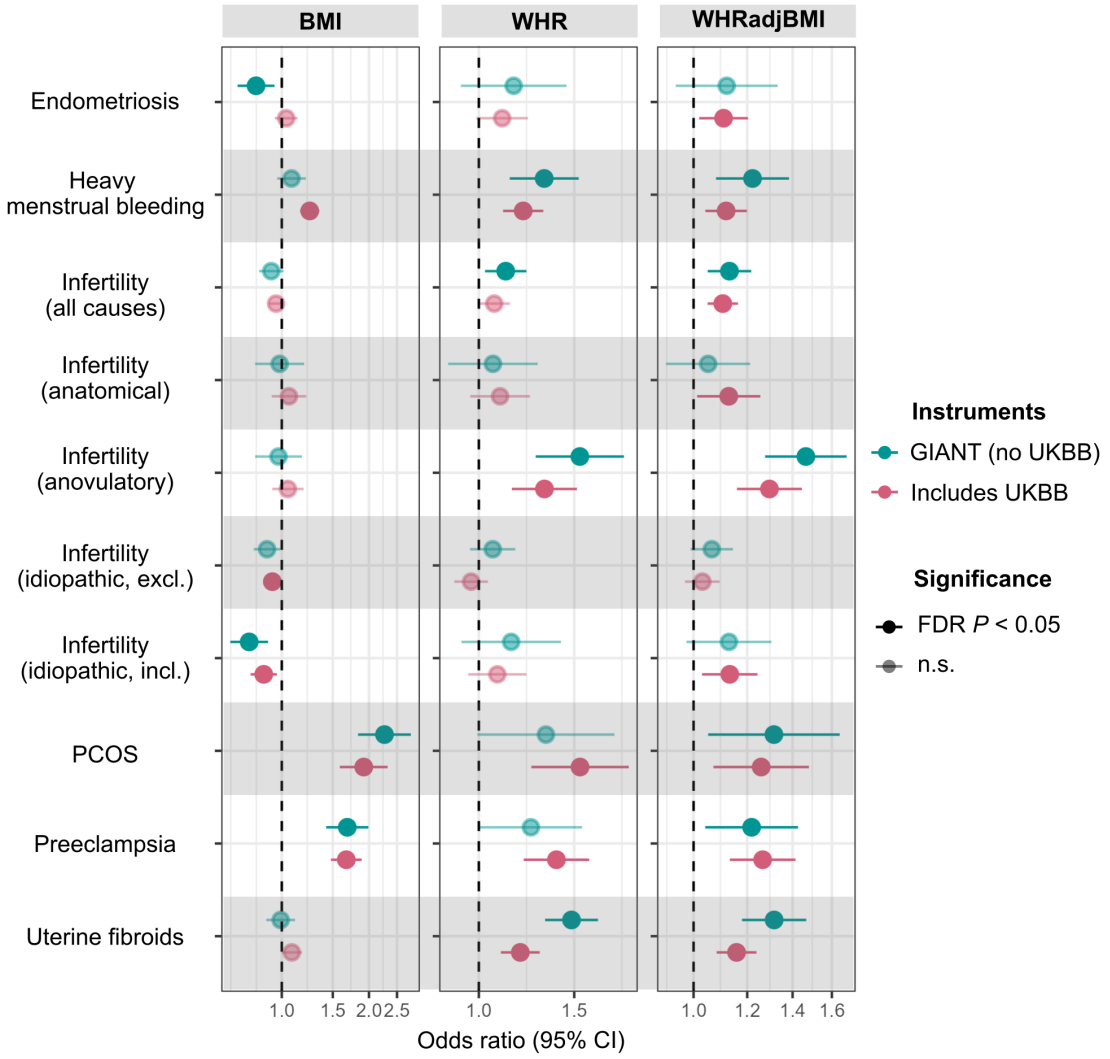


Figure 5.8: Comparison of main MR results to sensitivity analysis with GIANT-only instruments. Causal effects were estimated using inverse variance weighted two-sample MR with two sets of instruments: (i) from the GIANT-UKBB meta-analysis reported by Pulit *et al.* (2019)[216], and (ii) from the GIANT Consortium without UKBB participants reported by Locke *et al.* (2015)[552]. Odds ratio and 95% confidence intervals (CIs) from both sets of results are displayed. Solid-coloured points represent associations with false discovery rate (FDR)-adjusted P-value < 0.05, while faded points represent non-significant (n.s.) P-values > 0.05 after FDR adjustment for 30 tests. BMI = body mass index, Infertility idio, excl. = idiopathic infertility defined by exclusion, Infertility idio, incl. = idiopathic infertility defined by inclusion, PCOS = polycystic ovary syndrome, WHR = waist-to-hip ratio, WHRadjBMI = WHR adjusted for BMI.

Table 5.6: Multivariable Mendelian Randomisation to estimate the effect of WHR on female reproductive disorders adjusted for BMI. Heterogeneity P-value is generated by comparing the multivariable MR effect sizes to those generated using instruments for WHRadjBMI with two-sample MR using the inverse variance weighted method. BMI = body mass index, PCOS = polycystic ovary syndrome, WHR = waist-to-hip ratio, WHRadjBMI = WHR adjusted for BMI.

Diagnosis	Two-sample MR		Multivariable MR		Heterogeneity
	OR (95% CI)	P-value	OR (95% CI)	P-value	P-value
Endometriosis	1.11 (1.02 - 1.2)	1.61E-02	1.14 (1.03 - 1.25)	9.01E-03	0.68
Heavy menstrual bleeding	1.12 (1.04 - 1.2)	2.14E-03	1.2 (1.11 - 1.3)	3.46E-06	0.17
Infertility (all causes)	1.1 (1.05 - 1.16)	1.71E-04	1.09 (1.03 - 1.16)	4.10E-03	0.75
Infertility (anatomical)	1.29 (1.16 - 1.45)	4.66E-06	1.35 (1.18 - 1.53)	5.89E-06	0.65
Infertility (anovulatory)	1.13 (1.01 - 1.25)	2.90E-02	1.11 (0.97 - 1.26)	1.34E-01	0.82
Infertility (idiopathic, exclusion)	1.03 (0.971 - 1.09)	3.21E-01	0.985 (0.922 - 1.05)	6.62E-01	0.32
Infertility (idiopathic, inclusion)	1.13 (1.03 - 1.24)	1.05E-02	1.13 (1.01 - 1.26)	2.72E-02	1.00
PCOS	1.26 (1.07 - 1.48)	5.47E-03	1.37 (1.13 - 1.66)	1.51E-03	0.51
Preeclampsia or eclampsia	1.26 (1.13 - 1.41)	3.71E-05	1.31 (1.15 - 1.48)	2.88E-05	0.71
Uterine fibroids	1.16 (1.08 - 1.24)	2.28E-05	1.22 (1.12 - 1.32)	1.89E-06	0.36

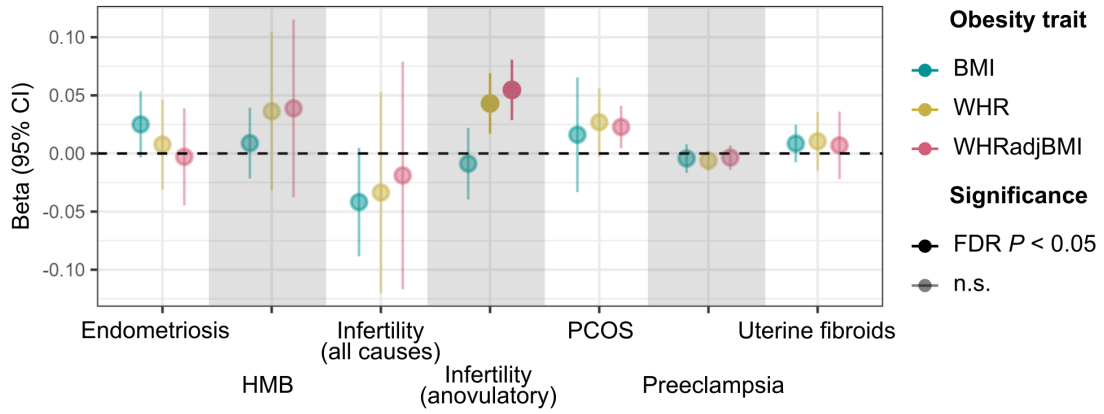


Figure 5.9: Reciprocal MR for causal effect of female reproductive disease on levels of BMI, WHR, and WHRadjBMI. Causal effects were estimated using inverse variance weighted two-sample MR with genetic instruments for reproductive conditions constructed from lead variants, defined as independent variants in a 1Mb genome-wide significant locus, in meta-analyses for each disorder. Outcomes were derived from the GIANT-UKBB meta-analysis reported by Pulit *et al.* (2019)[216]. Effect estimates are plotted as β and 95% confidence intervals (CIs). Solid-coloured points represent associations with false discovery rate (FDR)-adjusted P-value < 0.05, while faded points represent non-significant (n.s.) P-values > 0.05 after FDR adjustment for 21 tests. BMI = body mass index, HMB = Heavy menstrual bleeding, PCOS = polycystic ovary syndrome, WHR = waist-to-hip ratio, WHRadjBMI = WHR adjusted for BMI.

testosterone is a risk factor for PCOS (OR = 1.73 (1.47 - 2.03), $P = 5.43 \times 10^{-11}$), and fasting insulin is a risk factor for F-ANOV (OR = 2.12 (1.39 - 3.23), $P = 5.16 \times 10^{-4}$). Many of these associations remain significant after adjustment for obesity-related traits (FDR-adjusted P-value for 120 tests < 0.05) (Figure 5.10).

Insulin and testosterone genetically mediate the effects of obesity on female reproductive conditions

I therefore performed a series of mediation analyses to estimate: (i) whether FSH, testosterone, and insulin mediate the genetic effects of obesity on female reproductive conditions, and (ii) if so, what proportion of the total genetic effect of obesity can be attributed to each of these hormones. In the first step of mediation analysis, I found evidence that BMI and WHRadjBMI had causal effects, in opposite directions, on insulin (β_{BMI} = -0.0318 (0.0105), $P = 0.00243$; $\beta_{\text{WHRadjBMI}} = 0.164$ (0.0105), $P = 1.25 \times 10^{-54}$) and testosterone ($\beta_{\text{BMI}} = 0.166$ (0.0355), $P = 2.84 \times 10^{-6}$; $\beta_{\text{WHRadjBMI}}$ = -0.0877 (0.0318), $P = 0.00589$); however, increased WHR was only

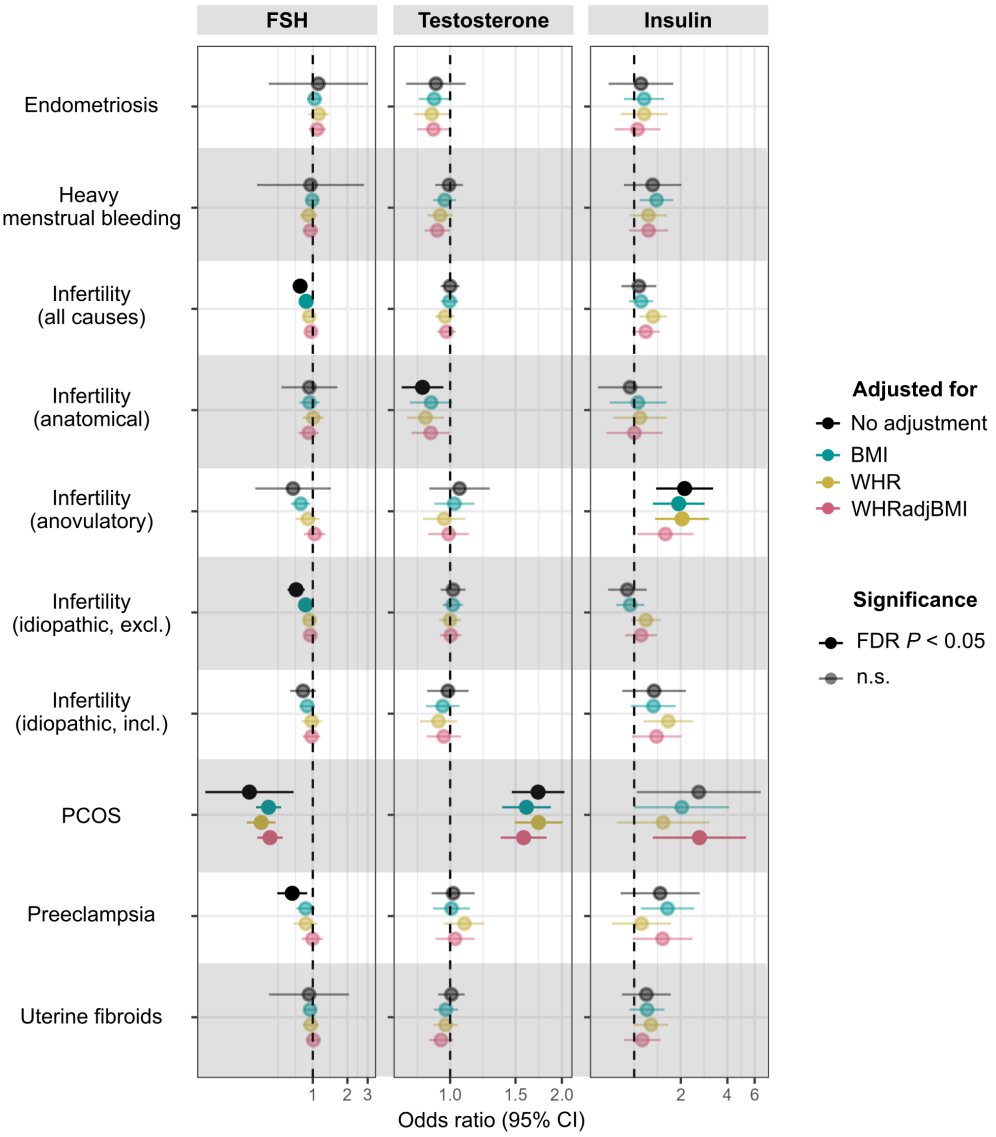


Figure 5.10: Genetically predicted causal effects of hormones on female reproductive diseases. Causal effects for unadjusted hormones were estimated using inverse variance weighted two-sample MR with genetic instruments constructed from lead variants, defined as independent variants in a 1Mb genome-wide significant locus, in meta-analyses for FSH and testosterone described in Chapter 4, and lead variants identified by Chen *et al.* (2021)[562] in a large-scale insulin GWAS. Effect estimates for obesity-adjusted hormones were calculated using multivariable MR with a joint instrument for each obesity × hormone pair, and are plotted as odds ratios and 95% confidence intervals (CIs). Solid-coloured points represent associations with false discovery rate (FDR)-adjusted P-value < 0.05, while faded points represent non-significant (n.s.) P-values > 0.05 after FDR adjustment for 120 tests. BMI = body mass index, FSH = follicle stimulating hormone, PCOS = polycystic ovary syndrome, WHR = waist-to-hip ratio, WHRadjBMI = WHR adjusted for BMI.

associated with increased levels of insulin (β = 0.152 (0.0154), $P = 5.76 \times 10^{-23}$) (Table 5.7). No obesity trait had a significant genetic effect on levels of FSH (FDR-adjusted P-value for 9 tests > 0.05).

In the second step, using multivariable MR to estimate the obesity-adjusted effects of hormones, I find that: independent of the effects of BMI or WHR, increased levels of insulin are a risk factor for the development of F-ANOV (OR for BMI-adjusted insulin = 1.94 (1.32 - 2.85), $P = 0.00905$; OR for WHR-adjusted insulin = 2.04 (1.36 - 3.05), $P = 0.00772$); and independent of any obesity trait, testosterone is a risk factor for PCOS (OR for BMI-adjusted testosterone = 1.60 (1.38 - 1.87), $P = 1.62 \times 10^{-8}$; OR for WHR-adjusted testosterone = 1.73 (1.49 - 2.01), $P = 2.43 \times 10^{-11}$; OR for WHRadjBMI-adjusted testosterone = 1.58 (1.37 - 1.82), $P = 1.04 \times 10^{-8}$) (Figure 5.10). For all obesity × female reproductive outcome combinations with a nominally significant causal association (MR P-value < 0.05), I estimated the proportion of total effect of obesity that is mediated by insulin or testosterone. Insulin mediates 38.8% (SE = 15.4) of the total effect of WHR on anovulatory infertility, and testosterone mediates 12.1% (3.70) of the total effect of BMI on PCOS (Table 5.8 and Figure 5.11).

Evidence for bi-directional effects between obesity traits and hormones

I further dissected the causal pathways between obesity, hormones, and reproductive conditions by testing for reciprocal effects of reproductive and metabolic hormones on BMI, WHR, and WHRadjBMI, which may mediate the effects of these hormones on reproductive conditions. Indeed, I found that increased levels of fasting insulin are associated with lower BMI (β = -0.247 (0.0870), $P = 0.00444$) but higher WHR (β = 0.893 (0.218), $P = 4.32 \times 10^{-5}$); and increased levels of testosterone are genetically causal for higher BMI (β = 0.0731 (0.0320), $P = 0.0227$) (Table 5.7). As above, I performed two-step mediation analysis to estimate that 7.88% (SE = 3.90) of the total effect of testosterone on PCOS is mediated by BMI (Figure 5.11 and Table 5.8).

To limit bias from sample overlap, I constructed hormone instruments from female-specific GWASs without UKBB participants; lead variants and summary

Table 5.7: Genetically predicted causal effects of obesity traits on hormones and vice-versa. Effect sizes were estimated using two-sample Mendelian Randomisation with the inverse variance weighted method. BMI = body mass index, FSH = follicle stimulating hormone, S.D. = standard deviation, SE = standard error, WHR = waist-to-hip ratio, WHRadjBMI = WHR adjusted for BMI.

		Effect of obesity trait on hormone			Effect of hormone on obesity trait		
Obesity trait	**Hormone**	**Number of SNPs**	**Beta (SE) per 1 S.D. higher genetically predicted obesity trait**	**P-value**	**Number of SNPs**	**Beta (SE) per 1 S.D. higher genetically predicted hormone level**	**P-value**
BMI	FSH	275	-0.0527 (0.0335)	1.15E-01	5	0.0147 (0.0326)	6.51E-01
	Testosterone	275	0.166 (0.0355)	2.84E-06	12	0.0731 (0.0321)	2.27E-02
	Insulin	274	-0.0318 (0.0105)	2.43E-03	36	-0.247 (0.087)	4.45E-03
WHR	FSH	196	-0.0385 (0.0325)	2.36E-01	5	0.035 (0.0312)	2.62E-01
	Testosterone	196	-0.0583 (0.0437)	1.82E-01	12	0.0319 (0.028)	2.55E-01
	Insulin	195	0.152 (0.0154)	5.76E-23	36	0.893 (0.218)	4.32E-05
WHRadjBMI	FSH	254	-0.024 (0.0263)	3.61E-01	5	0.0343 (0.0359)	3.39E-01
	Testosterone	254	-0.0876 (0.0318)	5.89E-03	12	-0.00102 (0.024)	9.66E-01
	Insulin	254	0.164 (0.0105)	1.25E-54	36	1.12 (0.264)	2.31E-05

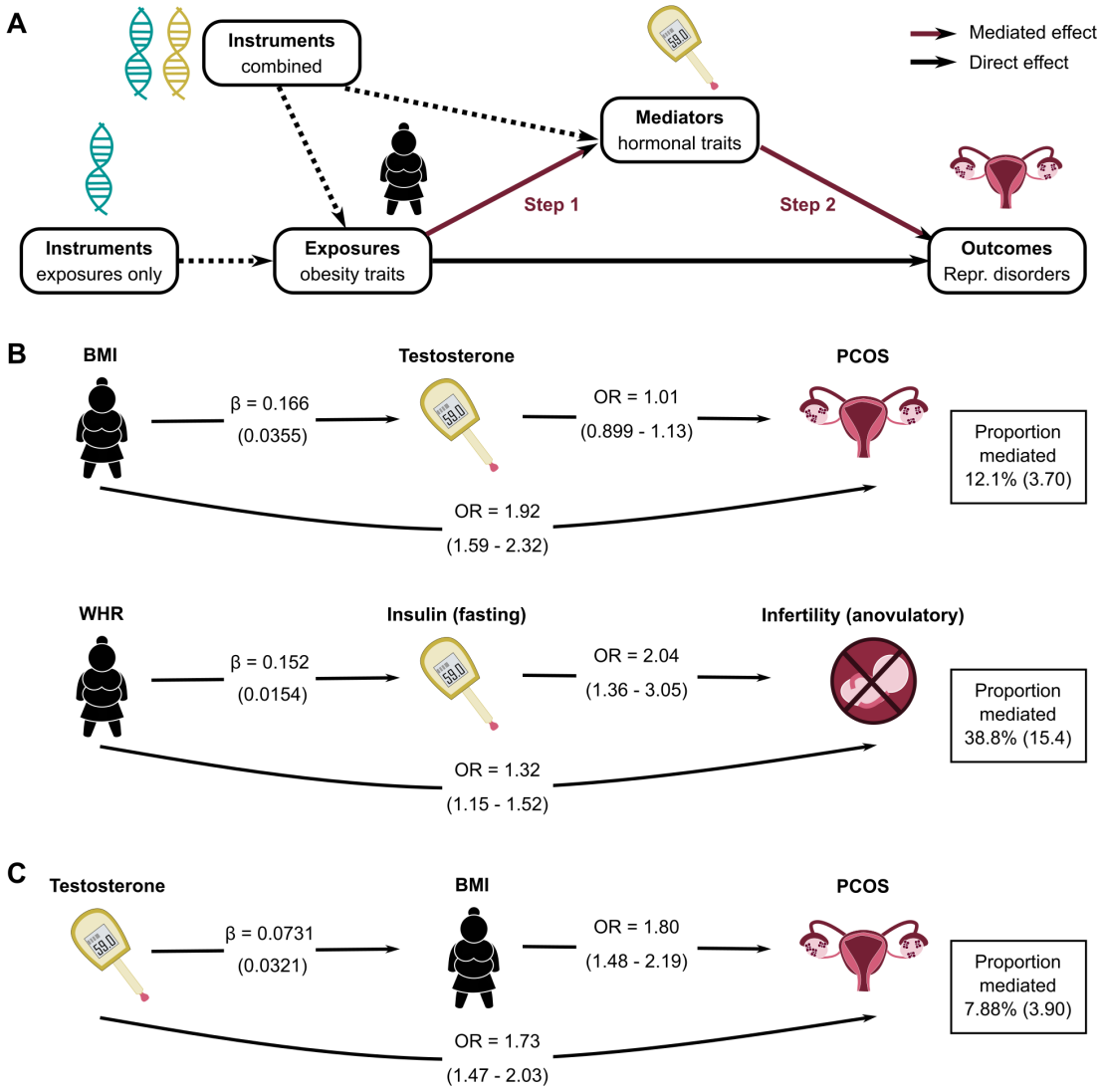


Figure 5.11: Genetically mediated proportions for causal effects of obesity and hormones on female reproductive conditions. (A) Methods overview for genetic mediation analysis using Mendelian Randomisation (MR). In the first step, the effect of the exposure on the mediator is estimated with genetic instruments for the exposure using two-sample MR; in the second step, the exposure-adjusted effect of the mediator on the outcome is estimated with joint instruments using multivariable MR. The proportion of mediated effect is calculated as the product of step 1 and step 2 estimates (indirect effect) to the total effect of the exposure on the outcome. (B) Mediation results for obesity × hormone × reproductive outcome sets with a significant proportion of mediation (95% confidence interval does not include 1). (C) Reciprocal mediation results for hormone × obesity × reproductive outcome sets with a significant proportion of mediation. BMI = body mass index, PCOS = polycystic ovary syndrome, WHR = waist-to-hip ratio.

Table 5.8: Genetically mediated proportions for causal effects of obesity and hormones on female reproductive conditions. Total effect sizes were estimated using two-sample MR with the inverse variance weighted method; adjusted effect sizes were estimated using multivariable MR. BMI = body mass index, FSH = follicle stimulating hormone, WHR = waist-to-hip ratio, WHRadjBMI = WHR adjusted for BMI.

Exposure	Mediator	Outcome	Proportion mediated	Exposure x outcome		Exposure x mediator		Mediator (adj.) x outcome	
			(SE)	OR (95% CI)	P-value	Beta (SE)	P-value	OR (95% CI)	P-value
BMI	Insulin	HMB	-4.76% (2.54)	1.25 (1.16 - 1.35)	9.24E-09	-0.0318 (0.0105)	2.43E-03	1.39 (1.09 - 1.79)	9.18E-03
		PCOS	-3.46% (2.15)	1.92 (1.59 - 2.32)	1.77E-11			2.03 (1.00 - 4.11)	4.85E-02
		Preeclampsia	-3.07% (1.65)	1.67 (1.48 - 1.89)	1.56E-16			1.64 (1.11 - 2.44)	1.41E-02
	Testosterone	PCOS	12.1% (3.70)	1.92 (1.59 - 2.32)	1.77E-11	0.166 (0.0355)	2.84E-06	1.01 (0.899 - 1.13)	8.87E-01
WHR	Insulin	Infertility (anovulatory)	38.8% (15.4)	1.32 (1.15 - 1.52)	7.95E-05	0.152 (0.0154)	5.76E-23	2.04 (1.36 - 3.05)	5.15E-04
WHRadjBMI	Insulin	Infertility (anovulatory)	29.5% (15.1)	1.29 (1.16 - 1.45)	4.66E-06	0.164 (0.0105)	1.25E-54	1.59 (1.05 - 2.41)	2.90E-02
		PCOS	69.3% (35.8)	1.26 (1.07 - 1.48)	5.47E-03			2.64 (1.32 - 5.29)	6.09E-03
	Testosterone	Endometriosis	8.92% (6.59)	1.11 (1.02 - 1.2)	1.61E-02	-0.0876 (0.0318)	5.89E-03	0.901 (0.816 - 0.996)	4.17E-02
		HMB	6.34% (4.42)	1.12 (1.04 - 1.2)	2.14E-03			0.923 (0.854 - 0.998)	4.52E-02
		Infertility (anatomical)	8.87% (6.81)	1.13 (1.01 - 1.25)	2.90E-02			0.886 (0.787 - 0.997)	4.42E-02
		PCOS	-17.4% (9.31)	1.26 (1.07 - 1.48)	5.47E-03			1.58 (1.37 - 1.82)	3.46E-10
Insulin	BMI	PCOS	-16.9% (10.5)	2.61 (1.04 - 6.58)	4.14E-02	-0.247 (0.087)	4.45E-03	1.93 (1.58 - 2.35)	9.92E-11
	WHR	PCOS	32.4% (20.8)	2.61 (1.04 - 6.58)	4.14E-02	0.893 (0.218)	4.32E-05	1.42 (1.12 - 1.78)	3.12E-03
Testosterone	BMI	PCOS	7.88% (3.90)	1.73 (1.47 - 2.03)	5.43E-11	0.0731 (0.0321)	2.27E-02	1.80 (1.48 - 2.19)	3.41E-09

statistics for FSH and testosterone were obtained from the GWAS meta-analyses described in Chapter 4, while lead variants and summary statistics for fasting insulin were published by Chen *et al.* (2021)[562]. All instruments had at least 5 variants with F-statistics > 30 to protect against weak instrument bias[574] (Table 5.1).

5.3.4 Mechanisms by which different forms of obesity affect the risk of various female reproductive diseases

Diseases cluster by shared genetic relationships with overall and centralised obesity

I applied dimension reduction with PCA to the matrices of single-variant effect estimates of the BMI and WHRadjBMI genetic instruments on each female reproductive condition (281 BMI SNPs × 10 reproductive conditions and 266 WHRadjBMI SNPs × 10 reproductive conditions) to identify reproductive conditions with shared genetic relationships with overall and centralised obesity. The first two PCs explain 39% of variance in BMI SNP effects and 40.7% of variance in WHRadjBMI SNP effects (Figure 5.12). Visual inspection of the first two PCs reveals that endometriosis, heavy menstrual bleeding, and uterine fibroids form a distinct cluster in both the BMI and WHRadjBMI effect plots. While the five types of infertility cluster together, particularly along PC1 of WHRadjBMI effects, PC2 groups anovulatory infertility with PCOS and preeclampsia in both the BMI and WHRadjBMI analyses. The clustering of diseases with common biological mechanisms, such as endometriosis and uterine fibroids[252,575], and PCOS and anovulatory infertility[196,576], suggests that clustering identifies shared genetic components of the aetiological role of obesity in the development of different reproductive conditions.

Obesity-associated genetic variants with similar effects on female reproductive conditions

We next clustered SNPs in each obesity genetic instrument with shared effects on female reproductive conditions using `MRClust`[565], which identifies variants with similar scaled effect sizes on the outcome ("substantial" clusters), as opposed to those with no effect ("null" clusters) or those which cannot be assigned to a

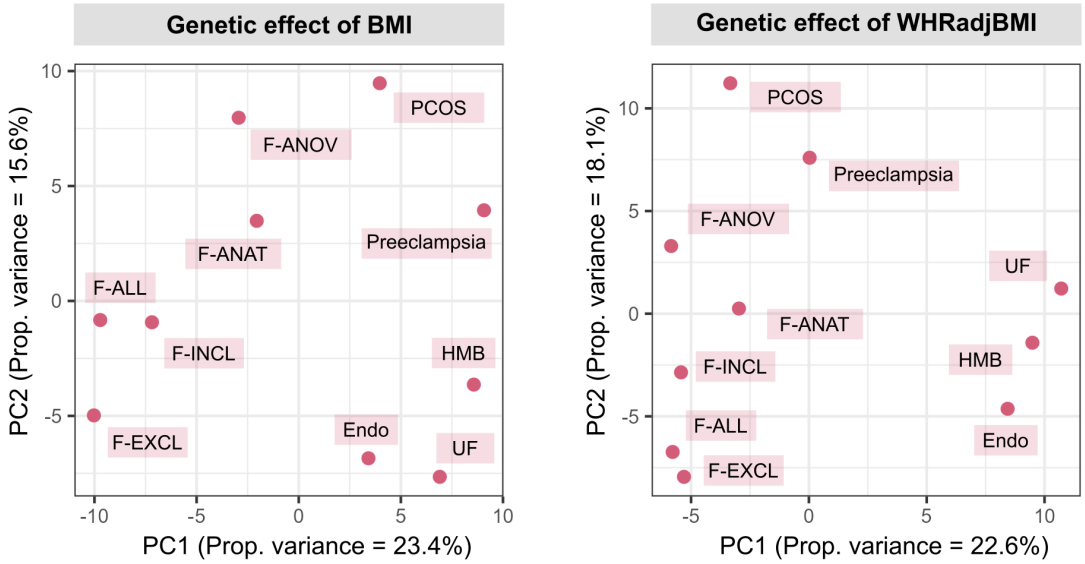


Figure 5.12: Principal component clustering of female reproductive conditions by shared obesity effects. Causal effects of each BMI and WHRadjBMI instrument variant were estimated using the Wald ratio, and the matrix of single-SNP × disease effects was dimensionally reduced using principal component analysis (PCA). The first two PCs are plotted, labelled with the proportion of variance (prop. variance) explained. Each point represents a female reproductive condition; Endo = endometriosis, HMB = heavy menstrual bleeding, F-ALL = female infertility of all causes, F-ANAT = anatomical female infertility, F-ANOV = anovulatory female infertility, F-EXCL = idiopathic female infertility by exclusion definition, F-INCL = idiopathic female infertility by inclusion definition, PCOS = polycystic ovary syndrome, UF = uterine fibroids.

cluster ("junk" clusters); we only performed this analysis for the 19/30 obesity trait × reproductive outcome pairs with a nominally significant total effect of obesity on the outcome ($P < 0.05$).

Eleven of the 19 pairs tested have a single substantial cluster, with the majority of SNPs in the cluster belonging uniquely to each obesity × reproductive disease pair (Figure 5.13). However, there are four WHR- and four WHRadjBMI-associated genetic variants, annotated to the same genes by Open Targets prioritisation[93,450], in the substantial clusters for both preeclampsia and anovulatory infertility. Three of these genes - *IQCN*, *ZNRF3*, and *THADA* – are highly expressed in female reproductive tissues including the ovary, uterus, and endometrium. Similarly, the four WHR-associated genetic variants shared by anovulatory infertility and uterine fibroids are all mapped to genes expressed in the uterus or testis (*HMGA1*,

ZC3H11B, *VEGFA*, and *RSPO3*), and associated with angiogenesis in endothelial cells and abnormal reproductive phenotypes in mice[577]. However, while several BMI-associated variants are shared by two diseases in this analysis, none are annotated to genes with known expression or functions in the reproductive system, perhaps indicating different mechanisms of effect for overall and centralised obesity on these traits.

Distinct clusters of variants drive the association between centralised obesity and some reproductive conditions

Genetic variants associated with WHR and WHRadjBMI form two sets of "substantial" clusters each when clustered by their effects on endometriosis, heavy menstrual bleeding, and infertility of all causes. WHRadjBMI variants form an additional third cluster in their associations with uterine fibroids (Figure 5.14). Each of these clusters may represent distinct mechanisms by which centralised obesity-associated genetic variants impact risk of female reproductive conditions. I annotated the variants in each cluster to the gene with the highest predicted functional score in the Open Targets Catalog[93,450] to perform gene-set enrichment analysis[566] against all 354 genes annotated to obesity-associated variants. Of the two clusters of WHRadjBMI variants with substantial causal effects on heavy menstrual bleeding, one is enriched for extracellular matrix organisation and degradation, while the other for Rho GTPase signalling, perhaps indicating two distinct pathways by which these variants mediate the effect of centralised obesity on menstrual bleeding. We could not perform enrichment analyses for the other obesity × reproductive outcome pairs as there were fewer than five genes annotated to variants in each substantial cluster.

5.4 Discussion

Using a combination of observational and genetic approaches, I find here that both overall and centralised obesity alter the risk of developing female reproductive conditions to heterogeneous extents. Some of these effects are mediated by the metabolic and reproductive hormones – insulin and testosterone – with potential

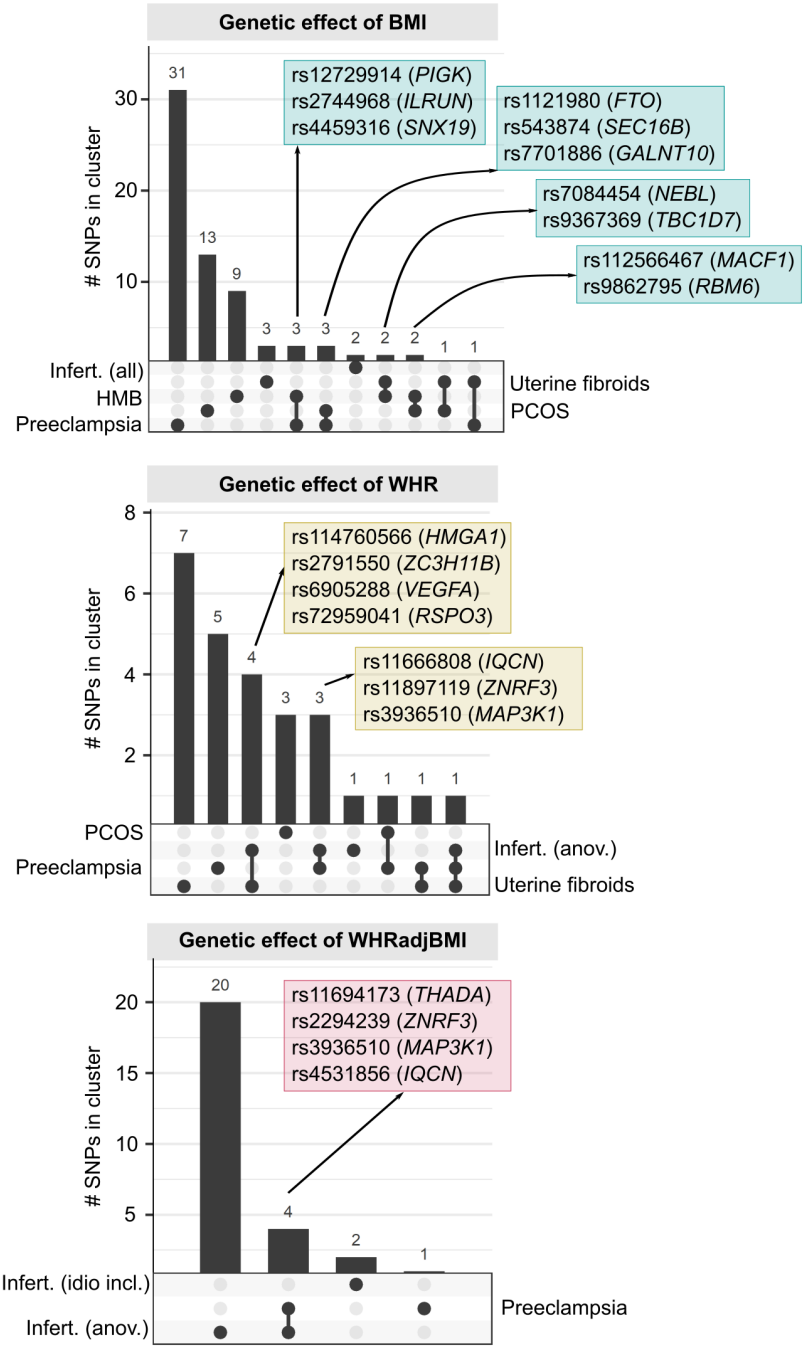


Figure 5.13: Obesity-associated genetic variants annotated to substantial clusters by MRClust. For each obesity × female reproductive condition pair with one substantial cluster, the overlap between variants assigned to that cluster with probability ≥80% is plotted. Variants are annotated to the gene with the highest functional score as classified by the Open Targets Catalog[93,450]. BMI = body mass index, HMB = heavy menstrual bleeding, Infert. = infertility, anov. = anovulatory, idio incl. = idiopathic inclusion definition, PCOS = polycystic ovary syndrome, WHR = waist-to-hip ratio, WHRadjBMI = WHR adjusted for BMI.

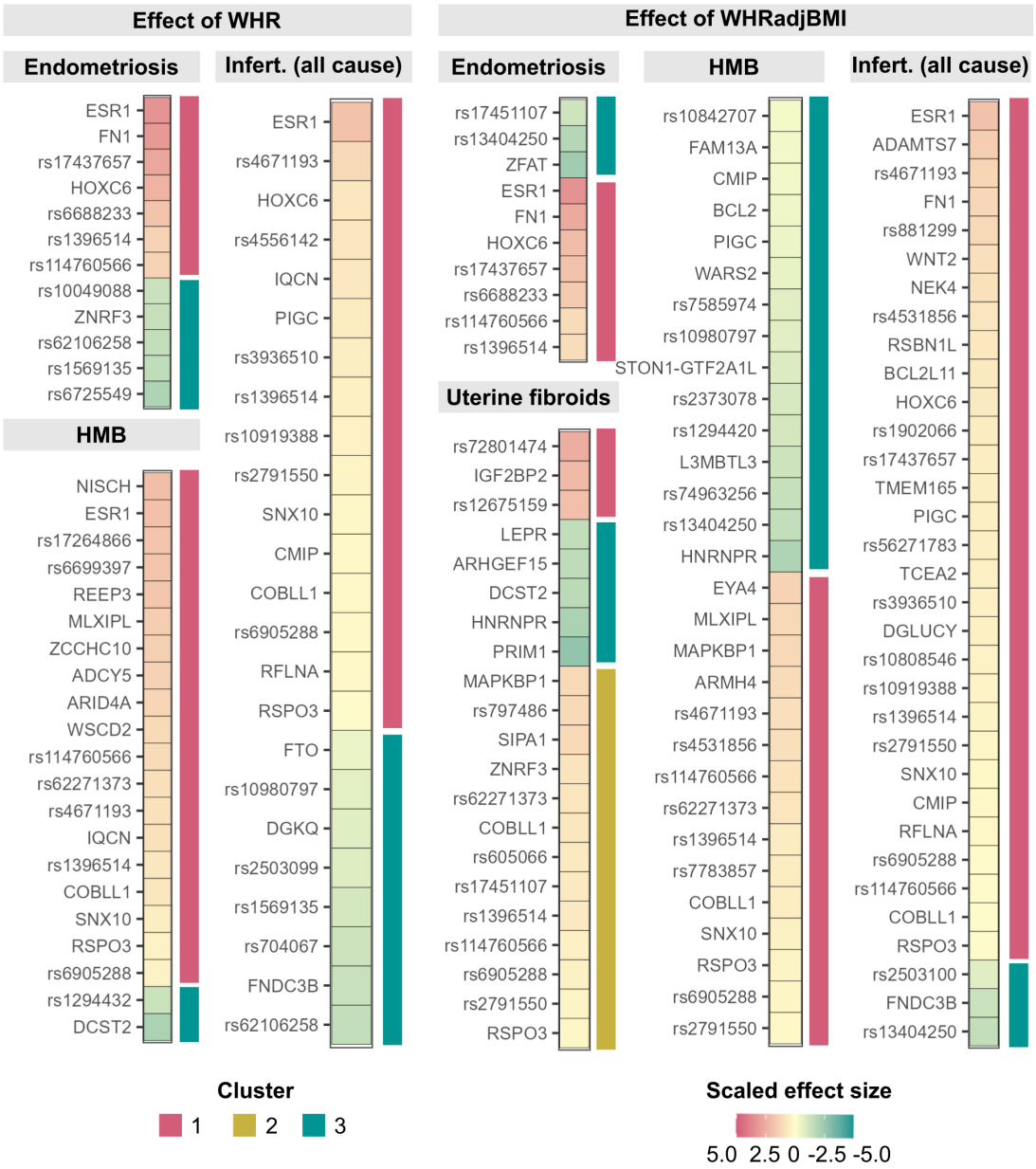


Figure 5.14: Heatmaps of obesity-associated SNP effect sizes for variants belonging to "substantial" clusters. For each pair of obesity × reproductive outcome associations with at least two "substantial" clusters identified by `MRClust`, scaled effect sizes of each variant in a "substantial" cluster are plotted. Only variants with >80% probability of belonging to the cluster are displayed; if > 30 variants meet this criteria, only the top 30 variants by probability are displayed. Variants are annotated to the gene with the highest functional score as classified by the Open Targets Catalog[93,450]; if no gene could be assigned, the variant ID is listed. HMB = heavy menstrual bleeding, Infert. = infertility, WHR = waist-to-hip ratio, WHRadjBMI = WHR adjusted for body mass index.

bi-directional effects between hormones and obesity-related traits. My analyses also point to shared genetic factors through which obesity affects reproductive conditions, likely through common biological mechanisms.

Many of the results I present are based on the Mendelian Randomisation approach, using genetic instrumental variables[63,578], that is less likely to be affected by reverse causation and confounding by unobserved factors that often bias observational studies[541,542]. Two-sample MR[579], which relies on GWAS summary statistics rather than individual-level data, may be sensitive to the choice and strength of instruments[574], the method used to estimate causal effect sizes[563,571], and sources of genetic associations for exposures and outcomes, particularly the level of similarity in participant populations and participant overlap[572,580]. When the exposure and outcome data sources used for two-sample MR are completely non-overlapping, effect estimates may be biased towards the null, whereas when the datasets are the same, the causal effect estimate is biased towards the observational association[572]. Sample overlap between the exposure and outcome datasets may modify other biases, such as weak instrument bias which biases the causal effect estimate towards the observational correlation, particularly when combined with sample overlap[581]. Genetic instruments may also be biased by winner's curse, wherein variant effect estimates are greater in the discovery study, but it is often difficult to find a third independent dataset to select genetic instruments. The results we present here are robust to a range of sensitivity analyses that investigate these biases, and the concordance of our results across different approaches increases our confidence in the findings that obesity traits are causally associated with risk of various female reproductive disorders. However, it will still be important to triangulate these results with other sources of evidence for causality, such as randomised control trials and longitudinal cohort studies.

We found that the genetically predicted causal effects of both BMI, i.e. overall obesity, and WHRadjBMI, i.e. centralised obesity independent of overall obesity, were over twice as strong as their observational counterparts in the UKBB for risk of preeclampsia and heavy menstrual bleeding; as was the effect of BMI on risk of

PCOS. Similarly, WHRadjBMI had only a causal association, but no observational association, with endometriosis, infertility, and uterine fibroids. The substantial difference between observational and causal estimates may be attributed to several factors. First, the age at recruitment in UKBB (40 - 69 years), upon which the observations of BMI and WHRadjBMI are based, is many years after the diagnosis of most of the conditions studied here - so the observed associations may capture downstream effects of the reproductive conditions on obesity, such as weight change due to treatment[393,582]. Second, although the observational models are adjusted for linear and quadratic effects of age, we may miss residual effects of menopause, during which there are significant changes in body fat distribution[180,222], and whose timing may itself be influenced by the presence or treatment of reproductive conditions[569,570]. While we considered adjusting for menopause as a covariate in our models, it was impractical to do so in the UKBB population where up to 41% of women with a reproductive disorder report being unsure of their menopause status (as opposed to 12% of those without one of the conditions studied here). Finally, there may be confounding of the observational associations by unmeasured variables, or referral bias wherein obesity status itself influences the likelihood of receiving a diagnosis[583].

BMI was positively associated with five of six reproductive conditions studied in the UKBB (all but infertility); however, centralised obesity independent of BMI was only observationally associated with the development of HMB and PCOS. While these observational associations may have indicated a stronger role for overall obesity, rather than abdominal adiposity, in driving the development of reproductive conditions, we noted that WHRadjBMI posed a significant genetically predicted risk for the development of nine out of ten female reproductive conditions, while genetically predicted overall obesity was only associated with increased risk of heavy menstrual bleeding, PCOS, and preeclampsia. Moreover, there are considerable differences in the observational effect sizes of the two forms of obesity on risk of reproductive disease: BMI has a significantly stronger effect than does WHRadjBMI on risk of HMB, PCOS, and preeclampsia, while the opposite is observed for endometriosis and uterine fibroids. Some of these differences may be attributed to

smoking status. Adjustment for smoking significantly increases the magnitude of associations between BMI and PCOS, as well as WHRadjBMI and endometriosis, HMB, PCOS, and preeclampsia. Future work to dissect the independent causal effects of tobacco use and adiposity on risk of reproductive disease may utilise multivariable Mendelian Randomisation with combined genetic instruments for smoking and obesity. Rather, these differences highlight that the aetiological role of obesity in risk of developing female reproductive conditions is heterogeneous in effect size and direction, and may be driven primarily by centralised obesity (endometriosis and uterine fibroids) or by both overall and centralised obesity (HMB, PCOS, and preeclampsia). The clustering of reproductive diseases by the genetic effect of obesity traits strengthens evidence for the notion that some disorders are disproportionately affected by one or the other form of obesity.

Abdominal (centralised) adiposity is a known risk factor for cardio-metabolic disease[515,584,585] and may be linked to female reproductive conditions through similar pathways of hyperinsulinaemia[424,586] and hyperandrogenism[424,512]. Indeed, we find that genetically predicted VAT mass is associated with increased risk of PCOS and preeclampsia, replicating observational associations in the literature[195,538,587]. Supporting the stronger effect of abdominal fat (primarily visceral adipose tissue) over gluteo-femoral fat (primarily subcutaneous adipose tissue)[588] for these diseases, we find that increased waist circumference poses a higher risk for the development of both these conditions than does increased hip circumference. While metrics of visceral adiposity have also been previously linked to the development of uterine fibroids[536,589,590], we failed to replicate these in our genetic analyses, suggesting that either: (i) genetically predicted VAT mass may not capture the phenotype derived from imaging scans, or (ii) that the previously reported observational associations may be biased by small sample sizes (fewer than 100 patients per study) or confounding.

The only female reproductive conditions for which we observed protective effects of obesity were endometriosis and infertility (idiopathic and all-cause). While increased BMI is positively associated with endometriosis in UKBB, genetically

predicted VAT mass is protective against the disease. Low BMI is a known risk factor for endometriosis[534,591], and we also observe that being "thinner" than average in childhood confers higher risk of disease (odds increase by 23%) than does being "plumper" (odds increase by 11%). The observational association between adult BMI and endometriosis may therefore be biased by reverse causation, such as weight gain due to hormonal treatments[582], pain and inflammation[592,593], or early onset of menopause[594]. Higher genetically predicted overall obesity and VAT mass are protective against the development of idiopathic infertility, with the latter also protecting against infertility of all causes, which strengthens the few observational studies that associate low BMI with reductions in fertility[539,595]. Moreover, by categorising infertility based on pathology (i.e. all causes, anatomical, anovulatory, or idiopathic), we were able to identify the heterogeneous, and opposing effects, of overall and centralised obesity on each type. Unlike the protective effects offered by increased BMI and predicted VAT mass, increased WHRadjBMI is associated with increased risk of all types of infertility examined here. Our findings add genetic evidence to a previous study demonstrating that increased abdominal fat is linked with anovulation, driven by the accumulation of subcutaneous rather than visceral adipose tissue[255]. However, there is still much work to be done to unpick the mechanisms underlying relationships between adipose tissue, hormone levels, and different pathologies of female infertility.

We applied a genetics-based mediation analysis to investigate the metabolic and reproductive hormones through which obesity may affect reproductive disease risk. As has been seen previously *in vitro*[596], in animal models[597], and in epidemiological and clinical studies in humans[378,513,598], we find bi-directional relationships between obesity and testosterone and insulin. Genetically predicted BMI and testosterone are both causal for increased risk of PCOS, and they also each mediate between 8-12% of the other's effect, bolstering previous reports on the relationships between obesity, testosterone, and PCOS[196]. Increased insulin levels are also a risk factor for PCOS[599,600] and other disorders of ovulation[416]; here, we find that 39% of the effect of centralised obesity on anovulatory infertility is genetically mediated by levels of

fasting insulin. There are already safe and inexpensive treatments for PCOS, such as metformin[601], that target insulin levels to also treat anovulation[602]. We hope that larger-scale GWASs for metabolic and reproductive hormones will help identify more such interventional targets to help women with a range of reproductive conditions and fertility issues, a significant proportion of which remain untreated[603,604].

In this systematic, large-scale characterisation of both the observational and genetic relationships between obesity, hormones, and female reproductive disorders, we provide substantial evidence for the aetiological roles of obesity in the risk of developing reproductive conditions. Female reproductive diseases remain under-reported and under-diagnosed, both in the UK and abroad[605–607]. Despite using three sets of sources to identify reproductive diagnoses – primary care records dating back to 1938, hospital in-patient records dating back to 1981, and self-report – the prevalence of all of the disorders we studied in UKBB were lower than estimates of their population prevalence. We may have failed to capture reproductive disease diagnoses due to oversight in the electronic health records, if the diagnosis was made in a setting without record linkage to the UKBB (such as maternity records), or participants' failure to recall a diagnosis during assessment. The lower prevalence of reproductive conditions in the UKBB limited our ability to investigate non-linearity in the genetic effects of obesity on the development of reproductive conditions. Given the observational evidence for U-shaped, inverted-U-shaped, and plateauing relationships of BMI with different reproductive diseases in our study and elsewhere[534,539], assessing genetic non-linearity is a promising avenue for future research. The low prevalence of reproductive diseases also limited our analyses to women of genetically ascertained European ancestry in UKBB, and lack of GWAS summary statistics in individuals of non-European ancestry also hindered causal analyses in these groups. The prevalence of some female reproductive conditions is higher in women of non-European ancestries, notably the higher rates of uterine fibroids in women of African ancestry[608,609], so it is crucial that future studies are attentive to equitable ancestral representation. In general, I make an appeal for the

better characterisation, diagnosis, and treatment of female reproductive conditions, which will vastly improve the mental and physical well-being of millions of women.

In conclusion, I have shown in this Chapter that obesity plays a substantial, but heterogeneous, role in the development of a range of female reproductive outcomes, and identified hormonal pathways as potential targets for the treatment of these conditions.

6

Conclusions and Future Directions

Contents

6.1 Conclusions 227
6.2 Future Directions 229
6.2.1 Leveraging electronic health records for genetics research 229
6.2.2 Interpretation of coding genetic variation 233
6.2.3 Deciphering gene regulation by the non-coding genome . 237
6.2.4 The future of human health 240

6.1 Conclusions

This thesis uses tools from statistical genetics to build upon decades of research on the drivers of common, complex metabolic and endocrine traits. In particular, I have demonstrated the utility of EHRs, linked to large biobanking projects, as a powerful and cost-effective data source for genetic studies. Within each Chapter, I also highlight the benefits of validation studies and meta-analyses across cohorts as a means to derive robust biological findings with implications for global health.

In Chapter 2, I lay out methods to extract cross-sectional and longitudinal quantitative phenotypic information from primary care EHRs linked to the UKBB for genetic association studies. For example, repeated measures of biomarkers measured in GP clinics can be distilled into phenotypes such as within-individual

average or rate of change, with different genetic drivers. In Chapter 3, I exploit the longitudinal aspect of EHR data to study the genetic factors responsible for baseline adiposity and adiposity-change. We validate the effects of rs429358, a missense variant in *APOE*, on weight-loss in individuals across three cohorts based in the UK, USA, and Estonia, hypothesising that systemic cholesterol dysregulation induced by this pleiotropic variant may drive weight loss in adults. Moreover, while there is some genetic correlation between obesity and rates of weight gain, the non-shared genetic aetiology of these phenotypes suggests value in studying longitudinal change in adiposity as a means to understand the rising rates of obesity across the globe[364]. Finally, I point to evidence of sex-heterogeneity in the genetic drivers of weight-change, outlining a case for further study into the sex-specific causes and consequences of obesity.

In Chapter 4, we utilise EHRs linked to seven biobanking projects across the globe, to extract measurements for reproductive hormones such as FSH and LH, whose genetic basis has not previously been outlined at scale. I show that despite the biases of EHR-derived phenotypes, we replicate several known genes for reproductive hormones and discover novel loci for steroid hormone biology through both common and rare variant analyses. With this blueprint for the genetic architecture of reproductive hormones, I interrogate the genetic basis of infertility, which is observationally associated with metabolic and endocrine disease[187,419]. In this first reported large-scale GWAS for infertility, I identify 21 loci associated with male and female infertility of different causes, many of which are not associated with reproductive diseases. While there is overall genetic correlation of 40% to 60% between various categories of female infertility and female reproductive disorders such as endometriosis and PCOS, I do not find substantial overlap in the genetic causes of infertility and reproductive hormones or adiposity-related traits.

I examined these surprising findings further in Chapter 5, where I interrogate the genetically predicted causal relationships between different types of obesity (overall or central) and risk of developing female reproductive disorders, which may be mediated by hormones. I find heterogeneity in the causal effects of

obesity on these conditions; while both overall and central obesity are genetically associated with higher risk of heavy menstrual bleeding, preeclampsia, and PCOS, only central adiposity increases risk of endometriosis, four categories of infertility, and uterine fibroids. Furthermore, I find that insulin may partly mediate the increased risk of anovulatory infertility posed by increased WHR; and that there are bi-directional relationships between BMI × testosterone and the risk they confer on the development of PCOS. The varying effects of overall and central obesity on each condition may point to differing causal roles for visceral and subcutaneous adipose depots on risk of reproductive disease, as well as to the genetic heterogeneity among these diseases.

Through GWASs for a range of metabolic and endocrine phenotypes, and *in silico* analyses to predict the molecular and systemic effects of genetic variants, this thesis makes a small contribution to answering one of the fundamental questions posed in Chapter 1: "what are we?". We can now turn to the question of where we are going next.

6.2 Future Directions

6.2.1 Leveraging electronic health records for genetics research

My thesis adds to growing evidence on the utility of EHRs for scalable and cost-effective genetics research[107,133,134] (Figure 6.1). Here, we apply a two-stage approach for genetic analysis, wherein we extract longitudinal or cross-sectional phenotypes, such as lifetime-averaged values, linear rates of change, or trajectory clusters for change in BMI and weight in the first stage, for genetic association testing in the second stage. While computationally efficient, two-stage approaches are limited in their ability to propagate error, as only the point estimate of the phenotype from the first stage is carried forward for GWAS[319]. In the future, it will be important to fit joint models, in a framework that accounts for relatedness and population stratification, to estimate the time-invariant and time-varying effects of each genetic variant[147]. Software packages such as `BOLT-LMM`[66] and `SAIGE`[67] for whole-genome

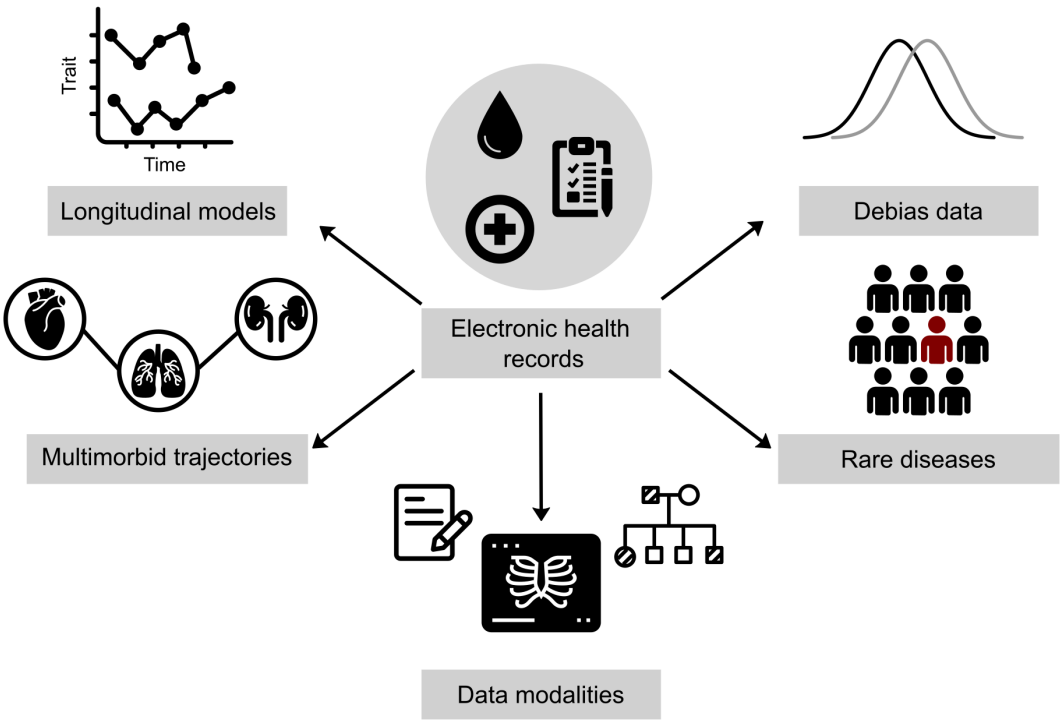


Figure 6.1: Opportunities for electronic health record (EHR)-based genetics research. Advances in deriving phenotypes from EHRs may inform powerful genome-wide association scans in the future. Some avenues for EHR phenotyping include: (1) incorporation of longitudinal metrics into joint models that can estimate the time-invariant and time-varying effects of each genetic variant, (2) the study of multimorbid, temporal disease trajectories to reveal shared or distinct genetic mechanisms among different diseases or within disease sub-types, (3) integration of data modalities such as free text, imaging scans, family history, etc., (4) identification of rare diseases through unusual symptom clusters in large EHR populations, and (5) debiasing measurements in EHR data to allow generalisable inferences from ascertained EHRs to the population at large. All icons obtained from the Noun Project database[69].

association testing have empowered the global adoption of GWASs in the past two decades; developing similarly easy-to-use and scalable software packages will be a necessary step to further longitudinal genetics analyses.

Alongside computational advances for scalability, statistical advances in debiasing EHR-derived phenotypes will be required for robust genetic studies that generalise to the population[610]. For example, I note in this thesis that individuals with more frequent measurements of BMI in UKBB-linked GP records are more likely to be overweight than the population of the UKBB; this may arise either because obesity is

often accompanied by co-morbidities that require regular monitoring[364], or because health practitioners are more likely to record weight measurements for patients who look overweight on examination[611,612]. Here, we demonstrate the robustness of our longitudinal weight-change metrics to such bias, but regardless, it will be important to develop a diversity of methods that can properly account for the longitudinal ascertainment bias in EHR data. "Front-door" adjustment for confounding ascertainment in longitudinal EHRs performs well for predicting prescriptions[613], and "doubly robust" estimators that circumvent model mis-specification may improve causal inference on the effect of certain drug classes on cancer risk[614]. Even for cross-sectional analyses, non-random missingness in EHR may skew inference - here, for example, I report genetic associations with sex hormones (such as FSH and LH, which are measured in GP clinics to diagnose menopause[153]), that overlap with reported genetic associations for age at menopause[615]. While these shared variants may truly be associated with both traits, they may also be confounded by the informed measurement of FSH and LH in menopausal women. Estimating genetic pleiotropy among common diseases in debiased EHR data might prove valuable for unpicking such effects[616], but these methods need further validation and expansion to a range of quantitative and binary phenotypes. Given the enormous value of health information in primary care EHRs, particularly for under-studied traits that are not commonly measured in biomarker panels, developing methods to debias EHR data will likely yield outsize benefits for genetics research[107,133,134].

Although not directly explored in this thesis, the broad phenotyping available in EHRs makes them well-suited for the study of multimorbidity, which is defined as the co-occurrence of two or more chronic health conditions in an individual[617]. As EHRs are longitudinal, they enable the study of temporal disease progression patterns and time-ordering of multimorbidity trajectories. A recent large-scale, temporal disease network based on over seven million patients in the Danish health registry suggests that there may be distinct trajectories of, for example, hypertension that progresses to heart failure versus hypertension that progresses to chronic kidney disease[618]. These may point to distinct genetic and/or environmental mechanisms

underlying disease progression, as heterogeneous comorbidity profiles can reveal disease subtypes with differential genetic risk profiles[105]. Early disease diagnosis and management may prevent a cascade of future health complications, making progress in elucidating multimorbidity trajectories hugely beneficial for personalised medicine and will alleviate the burden on healthcare systems.

In addition to informing population-based research on common, complex disorders, EHRs may be valuable for the discovery of rare symptoms of common disease, or rare diseases themselves. The genetics of rare disease have typically been studied in purpose-built cohorts, such as the Genomics England 100,000 Genomes Project[619] or the Deciphering Developmental Disorders study[620], as patients and families who visit specialised clinics can be referred for whole genome sequencing to inform their diagnosis. Although individually rare, together rare diseases have a prevalence of up to 10% in the population[621], so there may be much to learn from scanning EHR data for rare combinations of symptoms that could indicate the presence of a genetic disorder. For example, when doctors at Stanford were presented with a patient showing unusual complications of paediatric systemic lupus erythematosus (SLE), with an unclear course of treatment, they searched the entire Stanford EHR database to identify ten paediatric SLE patients with similarly unusual symptoms; examining the treatment trajectories for these additional cases enabled doctors to decide on the best treatment for their patient[622]. Following this success, several reports have used EHR data as a source of "collective memory" to inform rare disease treatment[139]. As population-based biobanks now increasingly have whole genome sequencing data linked to EHRs, rare disease scans are an exciting avenue for genetic discovery. In a similar vein, EHRs may allow the discovery of rare clusters of symptoms associated with common syndromes. For example, about 5% of patients with PCOS in the Danish disease trajectory browser[618] are also affected by sleep disorders, which are not commonly recognised complications of PCOS, and may thus point to novel disease mechanisms. Population-based biobanks, which have a vast array of genetic and environmental information linked to EHRs, will be valuable to unpick such heterogeneous components of disease aetiology.

Finally, there are emerging opportunities to extract a wealth of phenotypic information from EHRs that will require advances in epidemiological, statistical, and computational tools. For example, the first simulated randomised control trials based on longitudinal EHR data have proven that synthetic trials can replicate results from experimental data, at much lower cost to participants and healthcare systems[401,402]. Integrating polygenic risk scores as covariates in these trials will be crucial for accurate inference, particularly for heritable conditions such as cardiovascular disease and type 2 diabetes. However, whether EHR-simulated randomised control trials will generalise across a range of interventions and diseases remains to be seen. There is also a plethora of phenotypic information in EHR free-text fields, linked imaging records from ultrasound or magnetic resonance imaging scans, family history, and more, that are extremely valuable for genetic discovery if correctly extracted and debiased. As an under-appreciated source of diverse data, with representation along axes of ancestry, socioeconomic status, gender identity, and sexual orientation, primary care EHRs may also help expand genetic studies to under-served populations[623,624].

While many have questioned whether the era of GWAS is drawing to a close[625,626], I propose that powerful phenotyping, driven by EHRs, will continue to transform our understanding of the human phenotypic and genetic landscape.

6.2.2 Interpretation of coding genetic variation

A GWAS pipeline begins with phenotyping and ends with the interpretation of genetic variation (Figure 1.2) - and even if the GWAS era is drawing to a close, the era of variant interpretation is only just underway[625,627]. In this thesis, I identify hundreds of trait-associated SNPs across a range of phenotypes, but few are in coding genes (such as variants in *APOE* associated with adiposity-change and in *FSHB* for follicle stimulating hormone), and even fewer are missense variants with the potential to alter protein structure and function through changes to amino acids. Indeed, across all published GWAS variants, about 85% are in the non-coding regulatory genome, and cannot be directly linked to their effector genes[628,629]. Identifying

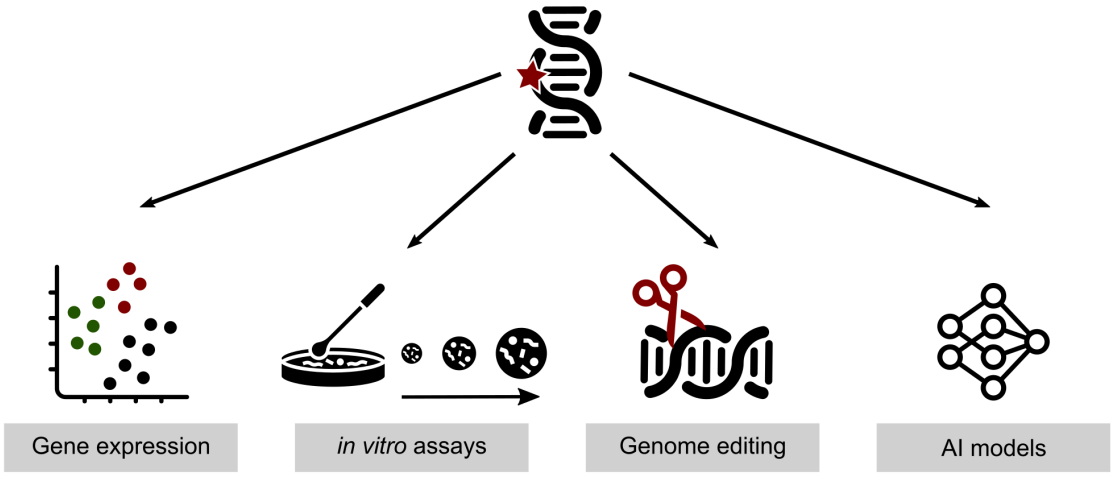


Figure 6.2: Methods to assess the impact of genetic variation. Genetic variation may affect protein structure and function (missense or splicing variants), or gene expression (non-coding variants) to alter downstream systemic phenotypes. The effects of genetic variation may be assayed through means such as: (1) bulk and single-cell gene expression databases to identify tissues and cell types of action, (2) longitudinal studies of cell growth and development *in vitro* to assess context-specific variant effects, (3) modelling of variants in cell lines or model organisms through CRISPR gene editing technologies, and (4) prediction of systemic effects, such as pathogenicity, with large artificial intelligence (AI) models. All icons obtained from the Noun Project database[69].

variant mechanisms remains challenging due to heterogeneity in variant penetrance, diversity in the cell types and states in which a variant may exert its effect, and prediction of systemic effects from molecular changes. In the following two sections, I will discuss various statistical and functional methods for the interpretation of coding and non-coding genetic variation (Figure 6.2).

If a variant can be assigned to a gene, as is often the case for protein-coding variants, large-scale databases such as the GTEx project[97], a compendium of gene expression for dozens of primary human tissues, can reveal tissue-specific gene expression profiles affected by the variant. Researchers have identified tissue-specific effects for a spectrum of genes associated with immune-related diseases[630], cancers[631] and metabolic diseases[632], among many others. Bulk gene expression profiles in tissues may mask the effects of individual cell populations; for example, variants associated with some rare muscular disorders disease may act only through their effects on fast-twitch myocytes, but not other cell types within muscles[633].

The advent of single-cell -omics technology has accelerated the integration of single-cell expression datasets from different sources, properly accounting for data normalisation and batch-correction to reduce confounding, to build cell-type databases[634,635]. Variants associated with context-specific changes in gene expression will emerge when these single-cell atlases are queried. Still, the utility of gene expression databases is limited for the study of inaccessible cell types; for example, a comprehensive catalogue of gene expression in the ovaries and testes is necessary to assess the gonadal effects of reproductive trait-associated variants that I discover in Chapter 4. Unpicking cell-type heterogeneity in these tissues is crucial due to the diverse functions of each cell type; follicle cells in the ovary produce oestradiol, and mature into the corpus luteum to begin secreting progesterone, whereas granulosa cells promote follicular growth, and richly vascularised thecal/interstitial cells support capillary growth[156,636]. Retrieving ovarian and testicular tissues from healthy requires invasive surgery[637,638], but it is not straightforward to study gonadal gene expression in model organisms, whose reproductive cycles do not mirror those of humans[639]. Moreover, genetic variants associated with infertility or reproductive hormones in women may act on menstrual cycle-dependent changes in physiology, but these effects are difficult to recapitulate in mice, which have oestrous cycles rather than menstrual cycles[640]. Devising methods to model state-dependent effects, systematically across all human tissues, will be enormously powerful in revealing novel biology.

Longitudinal protein-coding variant effects may be studied *in vitro*, such as through stem cells that are induced to develop into mature cell types of interest[641–643]. To study adipogenesis, for example, pre-adipocytes are cultured using a series of factors to develop into mature adipocytes, and the chemical cocktails to stimulate adipogenesis may be modified to simulate obesogenic or starvation environments[644]. Trajectories of gene expression, chromatin accessibility, epigenetic modifications, proteomics, and more, are then evaluated by sequencing cells harvested at regular time-points along the adipogenic course[645,646]. Multi-modal longitudinal analyses can be designed to evaluate the effects of genetic variants that act at specific

stages during adipogenesis; for example, cells with engineered allele substitutions in the *PPARGC1A* gene demonstrate alterations in lipogenesis and mitochondrial respiration, but not proliferation of pre-adipocytes[647]. It is also important to characterise the sources of stem cells used for functional enquiry - adipogenesis is commonly studied in mouse cell lines (such as 3T3-L1)[648], human cell lines (derived from adult white adipose or brown adipose tissue[649], or patients with severe lipodystrophies whose cells readily accumulate fat[650]), or directly in primary pre-adipocytes extracted from patients[651] - which all display different developmental trajectories. Patient-derived cell lines additionally enable researchers to examine differences in age, sex, and polygenic disease risk that may explain some variation in multi-modal -omic trajectories[645]. More broadly, identifying state-specific effects of coding genetic variants will be a massive undertaking that requires large-scale single-cell sequencing using model systems that can capture the diversity of age, sex, genetic liability, and cell type heterogeneity across the population spectrum.

Precise genome editing, based on the CRISPR-Cas9 technology for genetic manipulation at the base-pair level[652,653], has revolutionised our ability to introduce a variant of interest into a cell line or model organism through "knock-in" and "knock-out" studies[654]. Once appropriate read-outs, such as lipid accumulation by adipose cells[655], pathogen response by dendritic cells[656], or imaging-based phenotypes that can be applied broadly across cell types[657], have been established, precise genome editing enables scientists to examine the phenotypic effects of a molecular genetic change. In the ten years since the discovery of CRISPR-Cas9 systems, the technology has been: (1) engineered to target a range of different types of genetic modifications, including gene inhibition[658], precise allele substitutions[659,660], and prime editing without needing to introduce double-stranded DNA breaks[661], (2) expanded for delivery to thousands of model systems, from cells to tissues to organisms[662], and (3) scaled to conduct massively parallel assays in thousands or millions of cells simultaneously for CRISPR screens[663]. In the future, such large-scale studies will enable us to move beyond statistical association of SNPs with functional read-outs, to pinpoint causal molecular phenotypic effects induced by genetic variants. Still,

functional screens will be limited in their ability to extrapolate whole-organism systemic effects from causal molecular changes in model systems.

Developments in artificial intelligence (AI) for biology in the past few years, which have been powered by the availability of large-scale biological training data and growing capabilities of machine learning architectures, have transformed our ability to predict the systemic effects of coding variation in humans. The `AlphaFold` algorithm, which predicted the three-dimensional structure of 200 million proteins from amino-acid sequence alone[664], has recently been extended to the `AlphaMissense` model, which predicts the pathogenicity of missense variants by examining predicted changes to protein structure[665]. Another deep learning model, `PrimateAI`, employs an orthogonal approach to variant classification by comparing missense variants in humans to those in our evolutionary cousins, primates, by reasoning that benign variation in primates is unlikely to be pathogenic in humans[666]. Beyond the prediction of pathogenicity from coding missense variants, AI models may be trained to predict specific molecular changes caused by coding or non-coding variants - such as `SpliceAI`, which can model messenger RNA (mRNA) sequence from DNA sequence to identify cryptic splice sites that give rise to alternative splicing of transcripts in patients with rare genetic disorders[305].

6.2.3 Deciphering gene regulation by the non-coding genome

While we are rapidly progressing towards understanding the molecular and systemic impacts of missense (and more broadly, coding) genetic variants, deciphering the non-coding genome has been challenging[625,627]. Unlike protein-coding DNA, which can be read universally and linearly as a three-letter code for amino acids, the non-coding regulatory genome is cell-type specific and non-linear, guided by three-dimensional genome interactions. Transcriptional regulation is partly orchestrated by a series of DNA-binding proteins, such as transcription factors (TFs), to alter mRNA abundance and stability[667,668]. TFs bind to enhancers that may be thousands of kilobases away from their effector gene in linear sequence, and must be physically brought together by three-dimensional chromatin architecture to

exert their influence[669–671]. Moreover, multiple cell-type and cell-state specific DNA-binding elements in the same region can influence one another's binding dynamics to provide tight control of gene regulation[669,670]. A single enhancer may interact with over four gene promoters in the same context, and a promoter may interact with multiple enhancers, so many-to-many mappings are required to determine the effects of genetic variation in non-coding regulatory elements of the genome[669].

Statistical methods such as fine-mapping, described in Chapters 1 and 2, can nominate putative causal variants and effector genes, but do not account for cell-type specificity as they rely only on linkage patterns in linear genomic sequence[672]. Colocalisation of non-coding GWAS signals with eQTLs, as described in Chapter 4, may point to shared genetic aetiology between a GWAS trait and transcriptional changes, but are often computationally intractable for genome-wide analysis and can thus only identify proximal genes[673,674]. And while cell-type enrichment and gene prioritisation methods based on single-cell mRNA and protein expression have been broadly successful at identifying cell types of importance for a range of traits[444,675], the limited resolution and breadth of -omics data across cell types means that the mechanisms of action of most causal variants remain unknown[676,677]. All of these methods have not had much success in unpicking complex loci of high gene density, such as the 11q12.2 region associated with autoimmune responses[678] and the *TBX15-WARS2* locus associated with adiposity[212].

Functional efforts to decipher non-coding genetic variation have typically relied on a variant-by-variant approach that combines deep and broad sequencing, such as chromosome conformation capture (3C), epigenome, and transcriptome sequencing, with deep and broad phenotyping on the molecular and systemic levels. For example, an intronic variant in *COBLL1* (identified through GWASs to be associated with higher risk of type 2 diabetes but lower risk of obesity), has recently been shown to have differential enhancer activity in subcutaneous adipocytes, and is a causal variant for altered *COBLL1* expression, with validated effects in cell lines across differentiation time-points, and in mice[679]. While this metabolic variant lies within the intron of its effector gene, other work has demonstrated the ability of systematic

screens to unpick causal variants at gene-dense loci. A gain-of-function allele at the 3p21.31 locus, associated with risk of respiratory failure in coronavirus disease 2019 (COVID-19), was identified as the causal variant for altered *LZTFL1*-induced viral response in pulmonary epithelial cells, with its effects validated through genome editing[680]. Both the *COBLL1* and *LZTFL1* discoveries were enabled by leaps in experimental and computational methodologies, such as high-throughput 3C sequencing that can build short-range maps of chromatin interactions in a cell population[681–684], and AI models to predict chromatin architecture[685]. Novel methods to improve the range and resolution of three-dimensional genome maps will transform our ability to predict the genetic effects of distal enhancers on their effector genes[686].

While the above examples prove that the mechanisms of regulatory genetic variants can be resolved, deciphering all trait-associated GWAS variants will require an unprecedented scale of effort. One promising approach is to train AI models to predict emergent properties of the genome, such as chromatin accessibility, enhancer-promoter interactions, and gene expression, from linear DNA sequence alone[685,687]. Deep neural networks such as `deepC`, trained on cell-type specific data including 3C, open chromatin, and epigenome sequencing, can be generalised through transfer learning methods to cell-type agnostic predictions of genome architecture[685]. Transformer architectures, such as the `Enformer` algorithm, may also be trained on cell-type specific epigenomic data, to predict the effects of genetic variation on proximal gene expression[688]. As the quality and breadth of training and validation data improve, so will the ability of AI models to predict gene expression from sequence alone, revolutionising our ability to annotate GWAS loci with causal effects. Finally, examining large biobanks of individuals with deep whole genome sequencing will enable researchers to extrapolate causal molecular effects to systemic trait or disease phenotypes.

6.2.4 The future of human health

The enormous undertakings described above will eventually inform a better understanding of human health and the ways by which we prevent, treat, and cure disease. This vision of personalised precision medicine will partly be realised through genetics, for example, through polygenic risk score stratification to detect and treat disease at early stages[56,57], personalised interventions guided by a patient's predicted response to drug targets[101,135,689], and genome editing to erase severe monogenic disease[690,691]. Newborn blood spot ("heel prick") screening for rare genetic diseases, which is already a widespread practice in the UK for nine genetic disorders[692], will become commonplace for hundreds of rare disease diagnoses through whole genome sequencing[693]. Soon, newborn screening may even guide the prevention of common disease through early polygenic risk scoring, meaning that babies born in the future may be able to live their entire lives free from heritable illnesses. Armed with genetic information, parents can choose not to implant embryos, or selectively abort foetuses, with rare and severe monogenic disorders that greatly impact quality of life[694]. Some companies in the United States already provide services for parents to select embryos based on genetic scores, not just for rare diseases, but for heritable common traits such as height and risk of cardiovascular disease[695] - it is only a matter of years before such embryo selection becomes *de jure*. Finally, the era of genome editing is well upon us[662], and both scientists and the public must tackle thorny ethical questions regarding the modification of the human germline before embryo selection and editing become widely available to customise a child's genetic profile.

In delivering upon our ambitious goal to prevent all human disease, it is important to centre principles of equity in healthcare research and practice. Biases in polygenic score utility across ancestral groups will worsen existing inequalities in health outcomes between individuals of White ancestry and those of other ancestral backgrounds, but may be addressed through the inclusion of diverse participants in genetic studies[696]. While an admirable goal, our efforts cannot end with equitable ancestral representation in GWASs. Equitable research will require conscious commitments across a variety of actors in the health research ecosystem. For

example, there is evidence of systematic bias against African American patients in AI algorithms for healthcare prioritisation, which are trained on historical EHR data that perpetuate decades of conscious or unconscious racism in favour of wealthy, White American patients[697]. Developing fair algorithms, which can be interrogated by stakeholders across the spectrum of patients, healthcare providers, insurance companies, and legal representation, will be crucial to the future of algorithm-guided decision-making in hospitals[698]. We must also reflect upon whether research resources are fairly allocated to under-served populations. Women have historically been under-represented in medical research, from cell lines[699,700] to clinical trials[701,702], leading to worse healthcare outcomes for women across a range of disorders[703,704]. As we begin to recognise sex-specific symptoms of diseases[705,706], and fund the study of female-specific disorders such as endometriosis and PCOS, we must aim for equitable health outcomes across the sexes. And while we are spending billions of dollars researching neuro-degenerative disorders and cancers, which primarily affect the ageing populations of the western world, the global burden of disease is tilted towards developing countries in Africa and Asia[707]. Unsafe sanitation[708], malnutrition[709], and more recently, the effects of air pollution and extreme temperatures[710,711], are major drivers of healthcare needs in the developing world, but are still mostly overlooked by medical research[712]. Solving issues at the intersection of climate and health will be a challenge for the coming decades.

Nonetheless, the future of healthcare (and humanity) looks bright. Paul Gauguin painted "Where Do We Come From? What Are We? Where Are We Going?" in 1897, at a moment of deep personal crisis (Figure 1.1). His daughter had just died of pneumonia, an infectious respiratory illness that killed two million children a year in the 1900s[713]. In the following century, medical advances have reduced the childhood pneumonia mortality rate by over ten times, a rate that is still in steep decline[714]. And indeed, Gauguin's own mental health would likely have been addressed better today, through the panoply of behavioural, therapeutic, and pharmaceutical interventions for the treatment of mental illness[715]. As we continue to make astonishing leaps in science and medicine, I hope that Gauguin and his

ilk can rest in comfort, knowing that we are heading, partly guided by statistical genetics, towards a promising future (Figure 6.3).



Figure 6.3: Reinterpretation of "Where Do We Come From? What Are We? Where Are We Going?" by Paul Gauguin, generated by DALL-E. The text-to-image model DALL-E was prompted to reinterpret the original Paul Gauguin painting through a "more hopeful" lens, while maintaining his distinctive post-impressionist style.

Appendices

A

Supplement to Chapter 2

Table A.1: List of quantitative phenotypes with corresponding CPRD codes in EHR and fields in UKBB assessment centre visits. CPRD = Clinical Practice Research Datalink, EHR = electronic health records, UKBB = UK Biobank.

Phenotype	Read v2	CTV3	UKBB Field Code	Code-list source
Alanine aminotransferase (ALT)	44G3., 44GB., 44G..	44G3., XaLJx	30620	Denaxas *et al.* (2020)
Albumin	44M4.	XE2eA	30600	Denaxas *et al.* (2020)
Alkaline phosphatase (ALP)	44F.., 44F3.	44F3., XE2px	30610	Denaxas *et al.* (2020)
Basophills	42L..	42L..	30160	Denaxas *et al.* (2020)
Bicarbonate	44I7.	XE2q2		CALIBER
Body mass index (BMI)	22K..	22K..	21001	Manually curated
Calcium	44I8.	XE2q3, Xabpk	30680	Denaxas *et al.* (2020)
Chloride	44I6.	XE2q1		CALIBER
Cholesterol	44P.., 44PJ., 44PH.	XE2eD, XaJe9	30690	Denaxas *et al.* (2020)
C-reactive Protein (CRP)	44CS.	XaINL	30710	Denaxas *et al.* (2020)
Creatinine	44J3., 44J3z	44J3z, XE2q5	30510	Denaxas *et al.* (2020)
Eosinophills	42K..	42K..	30150	Denaxas *et al.* (2020)
Ferritin	42R4.	XE24r		CALIBER
Fibrinogen	42jA., 42Q9., 42Qn.	42Q9., Xalqo, X76wQ		CALIBER
Folate	42U.., 42U4., 42U5.	42U4., 42U5.		CALIBER
Follicle stimulating hormone (FSH)	4434., 443h.	4434., XM0lx, XE25J		Manually curated
Forced expiratory volume (FEV1)	339O.	X77Qu	20150	Denaxas *et al.* (2020)
Forced vital capacity (FVC)	3396.	3396.	20151	Denaxas *et al.* (2020)
Free androgen index (FAI)	4474.	4474., XE2pm		Manually curated
Glucose	44g.., 44g0., 44U.., 44TJ.	44g0., XM0ly	30740	Denaxas *et al.* (2020)
Haematocrit percentage	4258., 4257., 425..	X76tb, X76tc, XE2Zq	30030	Denaxas *et al.* (2020)
Haemoglobin concentration	423..	Xa96v, XE2m6	30020	Denaxas *et al.* (2020)
HbA1c	42W4., 42W.., 44TB.	XaERp		Denaxas *et al.* (2020)
HbA1cStd	42W5.	XaPbt	30750	Denaxas *et al.* (2020)
Height	229..	229..	50	Denaxas *et al.* (2020)
High density lipoprotein (HDL)	44P5., 44PB.	44P5., 44PB.	30760	Denaxas *et al.* (2020)
Iron	42R7.	X76tH		CALIBER
Low density lipoprotein (LDL)	44P6.	44P6.	30780	Denaxas *et al.* (2020)
Luteinising hormone (LH)	4433., 443e.	4433., XM0lv		Manually curated
Lymphocytes	42M..	42M..	30120	Denaxas *et al.* (2020)
Mean corpuscular haemoglobin concentration (MCHb)	429..	429..	30050	Denaxas *et al.* (2020)
Mean corpuscular volume (MCV)	42A..	42A..	30040	Denaxas *et al.* (2020)
Monocytes	42N..	42N..	30130	Denaxas *et al.* (2020)
Neutrophills	42J..	42J..	30140	Denaxas *et al.* (2020)
Oestradiol		4465.	30800	Manually curated
Phosphate	44I9.	XE2q4	30810	CALIBER
Platelets	42P..	42P..	30080	Denaxas *et al.* (2020)
Potassium	44I4.	XE2pz	30520	CALIBER
Progesterone	4463.	4463., X7727, X80GL, XE2dq, 5784.		Manually curated
Prolactin		XaELX		Manually curated
Red blood cells (RBC)	426..	426..	30010	Denaxas *et al.* (2020)
Sodium	44I5.	XE2q0	30530	CALIBER
Testosterone	4473.	4473., XE2dr	30850	Manually curated
Total Bilirubin	44EC., 44E.., 44E3.	44E.., XaERu, XE2qu	30840	Denaxas *et al.* (2020)
Triglycerides	44Q.., 44Q4.	44Q4., XE2q9	30870	Denaxas *et al.* (2020)
Urea	44J9., 44J8., 44J..	XM0lt, XaDvl	30670	Denaxas *et al.* (2020)
Vitamin B12	42T..	XE2pf		CALIBER
Vitamin D	44LA.	XE2e7, XaEQa, Xabo0	30890	CALIBER

contd. List of quantitative phenotypes with corresponding CPRD codes in EHR and fields in UKBB assessment centre visits.

Phenotype	Read v2	CTV3	UKBB Field Code	Code-list source
Waist circumference (WC)	22N0.	Xa041	48	Manually curated
Waist-to-hip ratio (WHR)		X76CX	Combination	Manually curated
Weight	22A..	22A..	21002	Manually curated
White blood cells (WBC)	42H.., 42H7.	42H.., XaIdY	30000	Denaxas *et al.* (2020)

Table A.2: UK Biobank assessment centre minimum and maximum values of quantitative traits. UKBB = UK Biobank.

Phenotype, units	UKBB reported range	Retained minimum value	Retained maximum value
Alanine aminotransferase (ALT), U/L	[3.01 - 495.19]	2.709	544.709
Albumin, g/L	[17.38 - 59.8]	15.642	65.78
Alkaline phosphatase (ALP), U/L	[8 - 1416.7]	7.2	1558.37
Basophills, 10^9/L	[0 - 3.03]	0	3.333
Body mass index (BMI), kg/m2	[12.12 - 74.68]	10.908	82.148
Calcium, mmol/L	[1.05 - 3.611]	0.945	3.9721
Cholesterol, mmol/L	[0.601 - 15.46]	0.5409	17.006
C-reactive Protein (CRP), mg/L	[0.08 - 79.96]	0.072	87.956
Creatinine, umol/L	[88 - 88160]	79.2	96976
Eosinophills, 10^9/L	[0 - 9.6]	0	10.56
Forced expiratory volume (FEV1), L	[0.01 - 22.33]	0.009	24.563
Forced vital capacity (FVC), L	[0.01 - 32.77]	0.009	36.047
Glucose, mmol/L	[0.995 - 36.813]	0.8955	40.4943
Haemoglobin concentration, g/dL	[0.09 - 22.27]	0.081	24.497
Haematocrit percentage, %	[0.05 - 72.48]	0.045	79.728
HbA1c, mmol/mol	[15 - 515.2]	13.5	566.72
Height, cm	[75 - 209]	67.5	229.9
High density lipoprotein (HDL), mmol/L	[0.219 - 4.401]	0.1971	4.8411
Low density lipoprotein (LDL), mmol/L	[0.266 - 9.797]	0.2394	10.7767
Lymphocytes, 10^9/L	[0 - 196.41]	0	216.051
Mean corpuscular haemoglobin (MCHb), pg	[0 - 95.67]	0	105.237
Mean corpuscular volume (MCV), fL	[52.1 - 160.3]	46.89	176.33
Monocytes, 10^9/L	[0 - 113.39]	0	124.729
Neutrophills, 10^9/L	[0 - 76.42]	0	84.062
Oestradiol, pmol/L	[175 - 14588]	157.5	16046.8
Phosphate, mmol/L	[0.369 - 4.702]	0.3321	5.1722
Platelets, 10^9/L	[0.3 - 1821]	0.27	2003.1
Potassium, mmol/L	[2.4 - 200]	2.16	220
Red blood cells, 10^12/L	[0.006 - 7.911]	0.0054	8.7021
Sodium, mmol/L	[10 - 380.7]	9	418.77
Testosterone, nmol/L	[0.35 - 54.342]	0.315	59.7762
Total bilirubin, umol/L	[1.08 - 144.52]	0.972	158.972
Triglycerides, mmol/L	[0.231 - 11.278]	0.2079	12.4058

contd. UKBB assessment centre minimum and maximum values of quantitative traits

Phenotype, units	UKBB reported range	Retained minimum value	Retained maximum value
Urea, mmol/L	[0.81 - 41.83]	0.729	46.013
Vitamin D	[10 - 362]	9	398.2
Waist circumference (WC), cm	[20 - 197]	18	216.7
Waist-to-hip ratio (WHR)	[0.56 - 1.27]	0.504	1.397
Weight, kg	[30 - 197.7]	27	217.47
White blood cells, 10^9/L	[0 - 389.7]	0	428.67

Table A.3: List of CPRD and ICD codes for bariatric surgery. CPRD = Clinical Practice Research Datalink, ICD = International Classification of Diseases.

Category	read_code	read_term	CPRD_MEDCODE
History of bariatric surgery (1)	76136	Maintenance of gastric band	97691
History of bariatric surgery (1)	761A5	Removal of gastric band	90454
History of bariatric surgery (1)	ZV45P	[V]Presence of gastric bypass	97014
Possible bariatric surgery (2)	7613	Plastic operations on stomach	22741
Possible bariatric surgery (2)	76130	Gastroplasty NEC	14910
Possible bariatric surgery (2)	7613y	Other specified plastic operation on stomach	62471
Possible bariatric surgery (2)	7613z	Plastic operation on stomach NOS	32581
Bariatric surgery (3)	76114	Sleeve gastrectomy and duodenal switch	93378
Bariatric surgery (3)	76115	Sleeve gastrectomy NEC	89259
Bariatric surgery (3)	76116	Laparoscopic sleeve gastrectomy	89148
Bariatric surgery (3)	76131	Partitioning of stomach	34119
Bariatric surgery (3)	76131	Mason vertical banded gastroplasty	59622
Bariatric surgery (3)	76132	Laparoscopic adjustable gastric banding	18863
Bariatric surgery (3)	76133	Partitioning of stomach using band	88474
Bariatric surgery (3)	76134	Partitioning of stomach using staples	56926
Bariatric surgery (3)	76135	Partitioning of stomach NEC	90600
Bariatric surgery (3)	76141	Bypass of stomach by anastomosis of stomach to duodenum	40977
Bariatric surgery (3)	76150	Bypass stomach by anastomosis stomach to transposed jejunum	39797
Bariatric surgery (3)	76160	Bypass of stomach by anastomosis of stomach to jejunum NEC	7082
Bariatric surgery (3)	76160	Mason high gastric bypass	92957
Bariatric surgery (3)	76160	Printer high gastric bypass	48417
Bariatric surgery (3)	76166	Laparoscopic gastric bypass	95929
Bariatric surgery (3)	761A.	Other open operations on stomach	54923
Bariatric surgery (3)	7633	Bypass of jejunum	39123

contd. List of CPRD and ICD codes for bariatric surgery.

Category	read_code	read_term	CPRD_MEDCODE
Bariatric surgery (3)	76330	Bypass of jejunum by anastomosis of jejunum to jejunum	56107
Bariatric surgery (3)	76331	Bypass of jejunum by anastomosis of jejunum to ileum	19929
Bariatric surgery (3)	76332	Bypass of jejunum by anastomosis of jejunum to colon	62848
Bariatric surgery (3)	7633y	Other specified bypass of jejunum	64123
Bariatric surgery (3)	7633z	Bypass of jejunum NOS	60421
Bariatric surgery (3)	76425	Duodenal switch	90517

B

Supplement to Chapter 3

Table B.1: List of obesity-related traits in the GWAS Catalog, accessed on 27 March 2023, with number of associations per trait and PubMed Identifiers (PMIDs) of reporting studies.

Obesity trait	Number of unique SNPs in GWASCatalog	Study PMIDs
adipose tissue measurement	632	22589738, 22044751, 29178545, 22589742, 31501611, 32349335, 30363675, 31900429, 34239500, 34128465, 27918534, 26480920, 30942860, 31513605, 35773277
anthropometric measurement, body mass index	1	19260139
anthropometric measurement, body weight	2	19260139
anthropometric measurement, waist circumference	2	19260139
BMI-adjusted hip circumference	3989	28552196, 25673412, 30576415, 34165540, 34021172
BMI-adjusted waist circumference	4325	28448500, 28552196, 28443625, 25673412, 26785701, 30576415, 31669095, 34165540, 33766948, 34021172
BMI-adjusted waist-hip ratio	4216	28448500, 28552196, 28443625, 25673412, 28270201, 30575882, 26785701, 31669095, 30389748, 34165540, 33766948, 34021172, 34226706, 34054925, 26426971, 28430825, 30778226, 27195708, 30239722, 35399580, 35168826, 35538205, 35999217, 35793639
body fat distribution	139	19197348, 22744181, 21897333, 30664634, 34239500
body fat percentage	429	31453325, 30593698, 26833246, 31752434, 33619380, 33541420, 30026463, 34662886, 33910581, 33980691, 28270201, 35051171, 36276968, 36329257
body mass index	7573	17903300, 19396169, 19260139, 22013104, 19197348, 19079260, 20397748, 19079261, 19851299, 21037115, 18454148, 23517042, 22982992, 22446040, 20966902, 22589738, 23251661, 22417934, 23583978, 23563607, 23669352, 22344219, 28892062, 28552196, 28448500, 28443625, 27416945, 27455349, 25811787, 25673413, 22344221, 24064335, 22581228, 23192594, 20935630, 21701570, 25133637, 24861553, 26278006, 28296344, 28270201, 27790247, 30239722, 30595370, 31453325, 31754094, 26865924, 25953783, 30108283, 31675503, 29381148, 32363570, 31498392, 30271922, 29615537, 32700739, 31752434, 33539483, 32653958, 30677029, 33048379, 32324875, 30327367, 32902719, 31669095, 31575865, 29868124, 31978080, 33654129, 31888951, 33209044, 33619380, 30026463, 33824429, 33045005, 33766948, 31903547, 34184762, 33888571, 34226706, 26426971, 28371326, 28430825, 24903457, 24322204, 24827717, 17434869, 17658951, 24281739, 25044758, 24348519, 25551457, 26604143, 27296613, 27918534, 28598419, 34684477, 34594039, 34210852, 30108127, 30166351, 30045251, 31217584, 29273807, 29551738, 26920376, 31513605, 26961502, 35299963, 35945198, 35399580, 35538205, 35051171, 35315439, 36581621, 36376304, 36276968, 35794696, 36329257, 35793639
body mass index, body weights and measures	1	30166351
body mass index, obesity	3	22013104

contd. List of obesity-related traits in the GWAS Catalog, accessed on 27 March 2023, with number of associations per trait and PMIDs of reporting studies.

Obesity trait	Number of unique SNPs in GWASCatalog	Study PMIDs
body mass index, visceral:subcutaneous adipose tissue ratio	4	27918534
body weight	1102	19260139, 19197348, 19079260, 19851299, 20966902, 23251661, 23633212, 28552196, 23643386, 30108283, 31647587, 31902109, 32902719, 34662886, 34309184, 34523824, 34074324, 34054925, 28548082, 17658951, 27918534, 34594039, 33625064, 30166351, 31560688
body weight loss	16	23633212, 23643386, 34523824
body weight, body weights and measures	2	30166351
body weights and measures	34	34662886, 30166351
fat body mass	196	22013104, 23575227, 28552196, 24064335, 32239398, 32719433, 33491310, 30026463, 34662886, 34239500, 28224759, 32963334, 30944420
fat body mass, lean body mass	9	32719433, 34239500
fat body mass, obesity	2	22013104
height-adjusted body mass index	2	25044758
hip circumference	4237	19260139, 23251661, 28552196, 25271088, 25673412, 31453325, 26675016, 30576415, 30108283, 31675503, 34165540, 30026463, 34662886, 34021172, 36276968, 35794696, 36329257
hip circumference, anthropometric measurement	2	19260139
lean body mass	295	23251661, 19268274, 28724990, 28552196, 28743860, 23108985, 30593698, 32193455, 32719433, 32970232, 32929176, 34662886, 34239500
lean mass-adjusted fat body mass	22	32239398, 32719433, 30944420
liver fat measurement	44	32766472, 32247823, 34128465, 34957434, 36513072
longitudinal BMI measurement	82	25760438, 25133637, 25378290, 26323598, 25953783, 31902109, 32126890, 31840077, 28044437, 35658861, 35794696
metabolically healthy obesity	14	30120429
obese body mass index status	26	30677029
obesity	260	19151714, 22013104, 18159244, 23563609, 19557161, 21544081, 23505185, 21706003, 19553259, 21552555, 19714249, 23563607, 22484627, 23818313, 20421936, 21935397, 31504550, 25963547, 31852448, 31118946, 31888951, 34836030, 24465431, 21708048, 30120429, 36187959
overweight body mass index status	3	25133637
pericardial adipose tissue measurement	4	22589742, 27918534
pericardial adipose tissue measurement, body weight, body height	3	27918534
renal sinus adipose tissue measurement	8	22044751
subcutaneous adipose tissue measurement	62	22589738, 29178545, 30363675, 34128465, 27918534, 26480920, 31513605
underweight body mass index status	2	25133637
visceral adipose tissue measurement	405	22589738, 31501611, 30363675, 31900429, 34239500, 34128465, 27918534, 26480920, 30942860, 35773277
visceral adipose tissue measurement, body mass index	45	22589738, 27918534
visceral:subcutaneous adipose tissue ratio	48	22589738, 30363675, 27918534, 26480920
Obesity trait	**Number of unique SNPs in GWASCatalog**	**Study PMIDs**
visceral:total adipose tissue ratio	1	30363675
waist circumference	4647	17903300, 19260139, 19197348, 20838585, 20966902, 23251661, 18454146, 19557197, 23754948, 28448500, 28552196, 28443625, 29212154, 23966867, 25673412, 31453325, 31170924, 26785701, 30576415, 30108283, 31647587, 31902109, 31928498, 31669095, 32902719, 33209044, 34165540, 30026463, 33766948, 34074324, 34021172, 31513605, 35051171, 36276968, 35794696, 36329257
waist height ratio	2	23251661
waist-hip ratio	4927	19396169, 20694148, 20935629, 23563607, 23754948, 28448500, 28552196, 28443625, 25673412, 23966867, 23192594, 28270201, 30575882, 31453325, 30239722, 26785701, 30576415, 30108283, 31669095, 31217584, 29868124, 33209044, 33619380, 30389748, 34165540, 33541420, 30026463, 33766948, 30595370, 34021172, 34226706, 34054925, 26426971, 28548082, 28430825, 17658951, 30778226, 27195708, 26920376, 35399580, 35168826, 35538205, 35051171, 35999217, 36276968, 35794696, 36329257, 35793639
waist-hip ratio, body mass index	8	29868124

Table B.2: List of Clinical Terms Version 2 (V2) and Version 3 (CTV3) read codes and International Classification of Diseases (ICD) codes to determine history or diagnosis of dementia or Alzheimer's.

Category	Description	Read V2 code	CTV3 code	ICD9 code	ICD10 code
History of Dementia	H/O: dementia	1461	1461		
History of Dementia	Dementia monitoring	66h..	XaMJC		
History of Dementia	Dementia annual review	6AB..	XaMGF		
History of Dementia	Dementia care plan	8CMZ.	XaaBZ		
History of Dementia	Dementia monitoring administration	9Ou..	XaMFy		
History of Dementia	Dementia monitoring first letter	9Ou1.	XaMG0		
History of Dementia	Dementia monitoring second letter	9Ou2.	XaMGG		
History of Dementia	Dementia monitoring third letter	9Ou3.	XaMGI		
History of Dementia	Dementia monitoring verbal invite	9Ou4.	XaMGJ		
History of Dementia	Dementia monitoring telephone invite	9Ou5.	XaMGK		
History of Dementia	Exception reporting: dementia quality indicators	9hD..	XaLFf		
History of Dementia	Excepted from dementia quality indicators: Patient unsuitabl	9hD0.	XaLFo		
History of Dementia	Excepted from dementia quality indicators: Informed dissent	9hD1.	XaLFp		
Diagnosis of Dementia	Senile and presenile organic psychotic conditions	E00..	X002w		
Diagnosis of Dementia	Senile and presenile organic psychotic conditions	E00..	X00R2		
Diagnosis of Dementia	Senile and presenile organic psychotic conditions	E00..	XE1Xr		
Diagnosis of Dementia	Senile/presenile dementia	E00..	X002w		
Diagnosis of Dementia	Senile/presenile dementia	E00..	X00R2		
Diagnosis of Dementia	Senile/presenile dementia	E00..	XE1Xr		
Diagnosis of Dementia	Senile dementia	E00..	X002w		
Diagnosis of Dementia	Senile dementia	E00..	X00R2		

contd. List of read codes and ICD codes to determine history or diagnosis of dementia or Alzheimer's.

Category	Description	Read V2 code	CTV3 code	ICD9 code	ICD10 code
Diagnosis of Dementia	Senile dementia	E00..	XE1Xr		
Diagnosis of Dementia	Uncomplicated senile dementia	E000.	E000.		
Diagnosis of Dementia	Presenile dementia	E001.	E001.		
Diagnosis of Dementia	Uncomplicated presenile dementia	E0010	E0010		
Diagnosis of Dementia	Presenile dementia with delirium	E0011	E0011		
Diagnosis of Dementia	Presenile dementia with paranoia	E0012	E0012		
Diagnosis of Dementia	Presenile dementia with depression	E0013	E0013		
Diagnosis of Dementia	Presenile dementia NOS	E001z	E001z		
Diagnosis of Dementia	Senile dementia with depressive or paranoid features	E002.	E002.		
Diagnosis of Dementia	Senile dementia with paranoia	E0020	E0020		
Diagnosis of Dementia	Senile dementia with depression	E0021	E0021		
Diagnosis of Dementia	Senile dementia with depressive or paranoid features NOS	E002z	E002z		
Diagnosis of Dementia	Senile dementia with delirium	E003.	E003.		
Diagnosis of Dementia	Multi infarct dementia	E004.	Xa0lH		
Diagnosis of Dementia	Multi infarct dementia	E004.	XE1Xs		
Diagnosis of Dementia	Arteriosclerotic dementia	E004.	Xa0lH		
Diagnosis of Dementia	Arteriosclerotic dementia	E004.	XE1Xs		
Diagnosis of Dementia	Uncomplicated arteriosclerotic dementia	E0040	E0040		
Diagnosis of Dementia	Arteriosclerotic dementia with delirium	E0041	E0041		
Diagnosis of Dementia	Arteriosclerotic dementia with paranoia	E0042	E0042		
Diagnosis of Dementia	Arteriosclerotic dementia with depression	E0043	E0043		
Diagnosis of Dementia	Arteriosclerotic dementia NOS	E004z	E004z		

contd. List of read codes and ICD codes to determine history or diagnosis of dementia or Alzheimer's.

Category	Description	Read V2 code	CTV3 code	ICD9 code	ICD10 code
Diagnosis of Dementia	Arteriosclerotic dementia NOS	E004z	E004z		
Diagnosis of Dementia	Other senile and presenile organic psychoses	E00y.	XE1Xt		
Diagnosis of Dementia	Other senile and presenile organic psychoses	E00y.	X00R0		
Diagnosis of Dementia	Presbyophrenic psychosis	E00y.	XE1Xt		
Diagnosis of Dementia	Presbyophrenic psychosis	E00y.	X00R0		
Diagnosis of Dementia	Senile or presenile psychoses NOS	E00z.	E00z.		
Diagnosis of Dementia	Dementia in conditions EC	E041.	E041.		
Diagnosis of Dementia	[X]Dementia in Alzheimer's disease	Eu00.	Eu00.		
Diagnosis of Dementia	[X]Dementia in Alzheimer's disease with early onset	Eu000	X002x		
Diagnosis of Dementia	[X]Presenile dementia,Alzheimer's type	Eu000	X002x		
Diagnosis of Dementia	[X]Primary degen dementia, Alzheimer's type, presenile onset	Eu000	X002x		
Diagnosis of Dementia	[X]Alzheimer's disease type 2	Eu000	X002x		
Diagnosis of Dementia	[X]Dementia in Alzheimer's disease with late onset	Eu001	Eu00.		
Diagnosis of Dementia	[X]Dementia in Alzheimer's disease with late onset	Eu001	X0030		
Diagnosis of Dementia	[X]Alzheimer's disease type 1	Eu001	Eu00.		
Diagnosis of Dementia	[X]Alzheimer's disease type 1	Eu001	X0030		
Diagnosis of Dementia	[X]Senile dementia,Alzheimer's type	Eu001	Eu00.		
Diagnosis of Dementia	[X]Senile dementia,Alzheimer's type	Eu001	X0030		

contd. List of read codes and ICD codes to determine history or diagnosis of dementia or Alzheimer's.

Category	Description	Read V2 code	CTV3 code	ICD9 code	ICD10 code
Diagnosis of Dementia	[X]Primary degen dementia of Alzheimer's type, senile onset	Eu001	Eu00.		
Diagnosis of Dementia	[X]Primary degen dementia of Alzheimer's type, senile onset	Eu001	X0030		
Diagnosis of Dementia	[X]Dementia in Alzheimer's dis, atypical or mixed type	Eu002	Eu002		
Diagnosis of Dementia	[X]Dementia in Alzheimer's disease, unspecified	Eu00z	Eu00z		
Diagnosis of Dementia	[X]Alzheimer's dementia unspec	Eu00z	Eu00z		
Diagnosis of Dementia	[X]Vascular dementia	Eu01.	XE1Xs		
Diagnosis of Dementia	[X]Arteriosclerotic dementia	Eu01.	XE1Xs		
Diagnosis of Dementia	[X]Vascular dementia of acute onset	Eu010	X003R		
Diagnosis of Dementia	[X]Multi-infarct dementia	Eu011	Xa0lH		
Diagnosis of Dementia	[X]Predominantly cortical dementia	Eu011	Xa0lH		
Diagnosis of Dementia	[X]Subcortical vascular dementia	Eu012	X003T		
Diagnosis of Dementia	[X]Mixed cortical and subcortical vascular dementia	Eu013	X003V		
Diagnosis of Dementia	[X]Other vascular dementia	Eu01y	Eu01y		
Diagnosis of Dementia	[X]Vascular dementia, unspecified	Eu01z	Eu01z		
Diagnosis of Dementia	[X] Presenile dementia NOS	Eu02z	Eu02z		
Diagnosis of Dementia	[X] Presenile dementia NOS	Eu02z	E00z.		
Diagnosis of Dementia	[X] Presenile dementia NOS	Eu02z	XE1Xr		
Diagnosis of Dementia	[X] Presenile dementia NOS	Eu02z	XE1Z6		
Diagnosis of Dementia	[X] Presenile dementia NOS	Eu02z	X00R2		
Diagnosis of Dementia	[X] Presenile dementia NOS	Eu02z	Xa1GB		
Diagnosis of Dementia	[X] Presenile psychosis NOS	Eu02z	Eu02z		
Diagnosis of Dementia	[X] Presenile psychosis NOS	Eu02z	E00z.		
Diagnosis of Dementia	[X] Presenile psychosis NOS	Eu02z	XE1Xr		
Diagnosis of Dementia	[X] Presenile psychosis NOS	Eu02z	XE1Z6		

contd. List of read codes and ICD codes to determine history or diagnosis of dementia or Alzheimer's.

Category	Description	Read V2 code	CTV3 code	ICD9 code	ICD10 code
Diagnosis of Dementia	[X] Presenile psychosis NOS	Eu02z	X00R2		
Diagnosis of Dementia	[X] Presenile psychosis NOS	Eu02z	Xa1GB		
Diagnosis of Dementia	[X] Primary degenerative dementia NOS	Eu02z	Eu02z		
Diagnosis of Dementia	[X] Primary degenerative dementia NOS	Eu02z	E00z.		
Diagnosis of Dementia	[X] Primary degenerative dementia NOS	Eu02z	XE1Xr		
Diagnosis of Dementia	[X] Primary degenerative dementia NOS	Eu02z	XE1Z6		
Diagnosis of Dementia	[X] Primary degenerative dementia NOS	Eu02z	X00R2		
Diagnosis of Dementia	[X] Primary degenerative dementia NOS	Eu02z	Xa1GB		
Diagnosis of Dementia	[X] Unspecified dementia	Eu02z	Eu02z		
Diagnosis of Dementia	[X] Unspecified dementia	Eu02z	E00z.		
Diagnosis of Dementia	[X] Unspecified dementia	Eu02z	XE1Xr		
Diagnosis of Dementia	[X] Unspecified dementia	Eu02z	XE1Z6		
Diagnosis of Dementia	[X] Unspecified dementia	Eu02z	X00R2		
Diagnosis of Dementia	[X] Unspecified dementia	Eu02z	Xa1GB		
Diagnosis of Dementia	[X] Senile psychosis NOS	Eu02z	Eu02z		
Diagnosis of Dementia	[X] Senile psychosis NOS	Eu02z	E00z.		
Diagnosis of Dementia	[X] Senile psychosis NOS	Eu02z	XE1Xr		
Diagnosis of Dementia	[X] Senile psychosis NOS	Eu02z	XE1Z6		
Diagnosis of Dementia	[X] Senile psychosis NOS	Eu02z	X00R2		
Diagnosis of Dementia	[X] Senile psychosis NOS	Eu02z	Xa1GB		
Diagnosis of Dementia	[X] Senile dementia, depressed or paranoid type	Eu02z	Eu02z		
Diagnosis of Dementia	[X] Senile dementia, depressed or paranoid type	Eu02z	E00z.		
Diagnosis of Dementia	[X] Senile dementia, depressed or paranoid type	Eu02z	XE1Xr		
Diagnosis of Dementia	[X] Senile dementia, depressed or paranoid type	Eu02z	XE1Z6		

contd. List of read codes and ICD codes to determine history or diagnosis of dementia or Alzheimer's.

Category	Description	Read V2 code	CTV3 code	ICD9 code	ICD10 code
Diagnosis of Dementia	[X] Senile dementia, depressed or paranoid type	Eu02z	X00R2		
Diagnosis of Dementia	[X] Senile dementia, depressed or paranoid type	Eu02z	Xa1GB		
Diagnosis of Dementia	[X] Senile dementia NOS	Eu02z	Eu02z		
Diagnosis of Dementia	[X] Senile dementia NOS	Eu02z	E00z.		
Diagnosis of Dementia	[X] Senile dementia NOS	Eu02z	XE1Xr		
Diagnosis of Dementia	[X] Senile dementia NOS	Eu02z	XE1Z6		
Diagnosis of Dementia	[X] Senile dementia NOS	Eu02z	X00R2		
Diagnosis of Dementia	[X] Senile dementia NOS	Eu02z	Xa1GB		
History of Dementia	[X]Delirium superimposed on dementia	Eu041	Eu041		
Diagnosis of Dementia	Alzheimer's disease	F110.	F110.		
Diagnosis of Dementia	Alzheimer's disease with early onset	F1100	XaIKB		
Diagnosis of Dementia	Alzheimer's disease with late onset	F1101	XaIKC		
Diagnosis of Dementia	[X]Other Alzheimer's disease	Fyu30	Fyu30		
History of Dementia	Language disorder of dementia	ZS7C5			
Diagnosis of Dementia	Dementia in Alzheimer's disease				F000A
Diagnosis of Dementia	Dementia in Alzheimer's disease				F001A
Diagnosis of Dementia	Dementia in Alzheimer's disease				F002A
Diagnosis of Dementia	Dementia in Alzheimer's disease				F009A
Diagnosis of Dementia	Vascular dementia			2904	F010
Diagnosis of Dementia	Vascular dementia			2904	F011
Diagnosis of Dementia	Vascular dementia			2904	F012
Diagnosis of Dementia	Vascular dementia			2904	F013
Diagnosis of Dementia	Vascular dementia			2904	F018
Diagnosis of Dementia	Vascular dementia			2908	F018
Category	**Description**	**Read V2 code**	**CTV3 code**	**ICD9 code**	**ICD10 code**
Diagnosis of Dementia	Vascular dementia			2904	F019
Diagnosis of Dementia	Unspecified dementia				F03X
Diagnosis of Dementia	Delirium superimposed on dementia			2903	F051
Diagnosis of Dementia	Alzheimer's disease				G300
Diagnosis of Dementia	Alzheimer's disease				G301
Diagnosis of Dementia	Alzheimer's disease				G308
Diagnosis of Dementia	Alzheimer's disease				G309

Table B.3: List of longevity-related traits in the GWAS Catalog, accessed on 27 March 2023, with number of associations per trait and PubMed Identifiers (PMIDs) of reporting studies.

Longevity trait	Number of unique SNPs in GWASCatalog	Study PMIDs
age at death	2	17903295
aging	3	17903295
disease free survival	5	17903295
longevity	204	21418511, 20304771, 20834067, 21740922, 21612516, 25918517, 31484785, 26912274, 31413261, 33948810,
longevity, aging	3	21782286
longevity, healthspan, parental longevity	24	32678081
parental longevity	363	29030599, 27015805, 31484785, 29227965, 30642433,
telomere length	8	24478790

Table B.4: Refined variants associated with obesity identified in genome-wide association studies (GWASs) for body mass index (BMI) and weight linear mixed-effects (LME) model intercepts in sex-specific and sex-combined strata; alongside the published obesity-associated SNP in highest linkage disequilibrium (LD) in the GWAS Catalog. MAF = minor allele frequency, TSS = transcription start site.

Refined SNP from LME intercept GWAS			Published obesity-associated SNP from GWAS Catalog in highest LD			
SNP	chr:pos (hg37)	Nearest TSS	r2 with refined SNP	SNP	chr:pos (hg37)	Nearest TSS
BMI (females)						
rs6746013	2:25143649	ADCY3	0.219	rs62140608	2:25070992	ADCY3
rs382210	3:84966018	CADM2	0.227	rs113639185	3:84769088	CADM2
rs73116335	12:50285780	FAIM2	0.254	rs1248264	12:50271045	FAIM2
rs6499244	16:69735271	NFAT5	0.215	rs4783581	16:69552785	NFAT5
rs28590228	19:47581242	ZC3H4	0.232	rs12973221	19:47614660	ZC3H4
BMI (males)						
rs116377258	1:110131521	GNAI3	0.211	rs62623713	1:110019439	SYPL2
rs2526389	3:50192826	SEMA3F	0.228	rs2014830	3:50172397	SEMA3F
rs72834677	6:26106604	H1-6	0.209	rs1977200	6:26466392	BTN2A1
rs77630251	10:104338370	SUFU	0.213	rs7893954	10:104318966	SUFU
rs10150482	14:79891882	NRXN3	0.267	rs7141420	14:79899454	NRXN3
rs8059619	16:29925445	ASPHD1	0.336	rs72798148	16:29926552	ASPHD1
rs62136859	19:47582356	ZC3H4	0.236	rs12973221	19:47614660	ZC3H4
BMI (sex-combined)						
rs6746013	2:25143649	ADCY3	0.219	rs62140608	2:25070992	ADCY3
rs382210	3:84966018	CADM2	0.227	rs113639185	3:84769088	LINC00971
rs10012797	4:3385176	RGS12	0.311	rs2051559	4:3298800	RGS12
rs7728485	5:139077144	CXXC5	0.346	rs356480	5:139106981	PSD2
rs4867646	5:170580152	RANBP17	0.388	rs74402020	5:170712474	RANBP17
rs2980423	8:8142575	PRAG1	0.201	rs200658291	8:8330299	LINC02950
rs11818043	10:104391627	SUFU	0.21	rs7893954	10:104318966	SUFU
rs2466826	16:28348344	NPIPB6	0.43	rs2726040	16:28330790	SBK1
rs6499244	16:69735271	NFAT5	0.215	rs4783581	16:69552785	NFAT5
rs28590228	19:47581242	ZC3H4	0.232	rs12973221	19:47614660	ZC3H4
rs6038557	20:6594349	BMP2	0.207	rs6054364	20:6575061	BMP2
Weight (females)						
rs12742446	1:33877057	PHC2	0.863	rs55745410	1:33802817	PHC2
rs61774713	1:41722396	SCMH1	0.33	rs3157	1:41493111	SCMH1
rs10218727	1:75012637	LRRC53	0.802	rs1514175	1:74991644	TNNI3K
rs2140090	3:41254646	CTNNB1	0.413	rs9867802	3:41343555	ULK4
rs396843	6:160740313	SLC22A3	0.261	rs394487	6:160778639	SLC22A3
rs116301700	6:33519068	BAK1	0.225	rs169738	6:33537546	BAK1
rs3798526	6:50809787	TFAP2B	0.35	rs4141973	6:50946521	TFAP2B
rs11030121	11:27736207	BDNF	0.216	rs4514364	11:27456059	LGR4
rs12367809	12:50256063	BCDIN3D	0.285	rs706798	12:50266164	FAIM2
rs1396869	17:46506929	SKAP1	0.38	rs3764400	17:46123932	NFE2L1
rs309190	19:47629264	SAE1	0.202	rs3810291	19:47569003	ZC3H4

contd. Refined variants associated with obesity identified in GWASs for BMI and weight LME model intercepts in sex-specific and sex-combined strata

Refined SNP from LME intercept GWAS			Published obesity-associated SNP from GWAS Catalog in highest LD			
SNP	chr:pos (hg37)	Nearest TSS	r2 with refined SNP	SNP	chr:pos (hg37)	Nearest TSS
Weight (males)						
rs116377258	1:110131521	GNAI3	0.211	rs62623713	1:110019439	SYPL2
rs2815753	1:72812324	NEGR1	0.203	rs2821248	1:72575560	NEGR1
rs13100903	3:50174572	SEMA3F	0.215	rs55754265	3:49687486	BSN
rs6954290	7:92283098	CDK6	0.2	rs11772579	7:92176640	RBM48
rs1323438	9:119115531	PAPPA	0.501	rs7869550	9:119134796	PAPPA
rs72841270	10:104642237	AS3MT	0.323	rs56119742	10:104579256	CYP17A1
rs11107120	12:93985482	SOCS2	0.284	rs10745657	12:93973501	SOCS2
rs10150482	14:79891882	NRXN3	0.267	rs7141420	14:79899454	NRXN3
rs28590228	19:47581242	ZC3H4	0.232	rs12973221	19:47614660	ZC3H4
Weight (sex-combined)						
rs11555079	1:41493749	SLFNL1	0.337	rs3157	1:41493111	SLFNL1
rs6699744	1:72825144	NEGR1	0.204	rs1620977	1:72729142	NEGR1
rs273866	1:97263793	PTBP2	0.255	rs3002271	1:97442229	PTBP2
rs754209	2:25502769	DNMT3A	0.202	rs7581217	2:25524944	DNMT3A
rs6792811	3:10142452	FANCD2	0.558	rs62245551	3:10174836	VHL
rs4115674	3:84979990	CADM2	0.248	rs113639185	3:84769088	CADM2
rs2915293	3:87995951	ZNF654	0.394	rs56398417	3:88024986	ZNF654
rs6894631	5:133864436	JADE2	0.369	rs11948454	5:133850622	JADE2
rs864354	5:142634708	NR3C1	0.515	rs7701443	5:142792650	NR3C1
rs144489757	5:95694609	PCSK1	0.221	rs12517140	5:95639072	PCSK1
rs10457469	6:126083658	HEY2	0.229	rs9375417	6:126248374	NCOA7
rs8180649	6:126992202	CENPW	0.203	rs10457481	6:126813721	CENPW
rs520829	6:160767905	SLC22A3	0.253	rs3127596	6:160953035	LPA
rs2517609	6:30170287	TRIM26	0.203	rs407238	6:29806901	HLA-G
rs7742908	6:33519062	BAK1	0.225	rs169738	6:33537546	BAK1
rs9395747	6:51799902	PKHD1	0.218	rs2499468	6:51809081	PKHD1
rs11575375	7:50571779	DDC	0.306	rs4947581	7:50565967	DDC
rs2282978	7:92264410	CDK6	0.2	rs11772579	7:92176640	RBM48
rs201519328	8:77223088	ZFHX4	0.211	rs1459180	8:76982267	ZFHX4
rs1323438	9:119115531	PAPPA	0.501	rs7869550	9:119134796	PAPPA
rs705150	10:125230186	LINC02641	0.325	rs7090367	10:125234492	LINC02641
rs567056	11:30435051	MPPED2	0.274	rs607987	11:30223574	FSHB
rs880610	11:68191947	LRP5	0.26	rs4988321	11:68174189	LRP5
rs1379765	12:41850195	PDZRN4	0.285	rs1405552	12:41746673	PDZRN4
rs9538141	13:59178258	PCDH17	0.218	rs4886088	13:59128526	PCDH17
rs45528934	14:23793305	PABPN1	0.363	rs1057119	14:23746269	BCL2L2
rs12708665	16:24728227	TNRC6A	0.355	rs7186893	16:24806420	TNRC6A
rs112875011	17:46420326	SKAP1	0.401	rs3764400	17:46123932	NFE2L1
rs3815358	17:62010270	CD79B	0.474	rs72845891	17:61696759	MAP3K3
rs9783858	18:42534584	SETBP1	0.307	rs9967367	18:42606091	SETBP1
rs28590228	19:47581242	ZC3H4	0.232	rs12973221	19:47614660	ZC3H4
rs8133137	21:40290503	ETS2	0.586	rs9983113	21:40315316	ETS2

Table B.5: Significant associations of longevity SNPs, curated from the GWAS Catalog, with adiposity-change phenotypes; significance established at $P < 0.05/138$ after Bonferroni-correction at 5% across 138 independent SNPs tested.

SNP	chr:pos (hg19)	MAF	Adiposity trait strata	Linear slope (u1)		p (k1) High gain cluster		p(k1) + p(k2) High or moderate gain clusters		p(k1) + p(k2) + p(k3) All but loss cluster	
				Beta (SE)	P-value	OR (95% CI)	P-value	OR (95% CI)	P-value	OR (95% CI)	P-value
rs10852521	16:53804965	0.497	BMI_M	-0.00419 (0.00522)	0.42	0.99 (0.985 - 0.996)	0.00069	0.99 (0.985 - 0.996)	6.00E-04	0.989 (0.984 - 0.995)	0.00016
rs7412	19:45412079	0.0811	BMI_F	-0.0317 (0.00871)	0.00027	0.984 (0.975 - 0.993)	0.00086	0.986 (0.976 - 0.995)	0.0022	0.984 (0.975 - 0.994)	0.00095
rs7412	19:45412079	0.0811	BMI_sex_comb	-0.0275 (0.00644)	2.00E-05	0.986 (0.979 - 0.993)	4.80E-05	0.986 (0.979 - 0.993)	6.00E-05	0.987 (0.98 - 0.994)	2.00E-04
rs7412	19:45412079	0.0811	Weight_sex_comb	-0.0256 (0.00618)	3.30E-05	0.986 (0.98 - 0.992)	1.70E-05	0.985 (0.979 - 0.992)	4.10E-06	0.986 (0.98 - 0.992)	1.70E-05
rs4420638	19:45422946	0.19	BMI_F	0.054 (0.00607)	6.30E-19	1.02 (1.01 - 1.03)	1.70E-10	1.02 (1.01 - 1.03)	1.60E-09	1.02 (1.01 - 1.03)	1.10E-08
rs4420638	19:45422946	0.19	BMI_M	0.0483 (0.00663)	3.30E-13	1.02 (1.01 - 1.02)	2.10E-06	1.02 (1.01 - 1.02)	2.80E-06	1.02 (1.01 - 1.02)	1.40E-05
rs4420638	19:45422946	0.19	BMI_sex_comb	0.0514 (0.00448)	1.60E-30	1.02 (1.01 - 1.02)	4.10E-15	1.02 (1.01 - 1.02)	7.60E-14	1.02 (1.01 - 1.02)	5.60E-12
rs4420638	19:45422946	0.19	Weight_F	0.056 (0.0058)	4.30E-22	1.02 (1.01 - 1.03)	1.30E-09	1.02 (1.01 - 1.02)	2.10E-09	1.02 (1.01 - 1.02)	5.40E-09
rs4420638	19:45422946	0.19	Weight_M	0.053 (0.00635)	7.00E-17	1.02 (1.01 - 1.02)	6.70E-07	1.02 (1.01 - 1.02)	1.40E-06	1.01 (1.01 - 1.02)	7.40E-06
rs4420638	19:45422946	0.19	Weight_sex_comb	0.0546 (0.00428)	2.80E-37	1.02 (1.01 - 1.02)	3.60E-16	1.02 (1.01 - 1.02)	7.50E-16	1.02 (1.01 - 1.02)	2.90E-15

C

Supplement to Chapter 4

Table C.1: Case definitions (ICD9 and ICD10 codes and primary care Read v2 and Read v3 codes) for the various female and male infertility analyses.

Infertility category	Case inclusions		Case exclusions	
	ICD codes	Read codes	ICD codes	Read codes
Female infertility of all causes (F_ALL)	ICD9 - 6280; 6283; 6284; 6288; 6289; ICD10 - N97; N970; N971; N972; N973; N974; N978; N979	Read v2 - K5B..; K5B0.; K5B00; K5B01; K5B0z; K5B1.; K5B1z; K5B2.; K5B20; K5B21; K5B23; K5B2z; K5B3.; K5B30; K5B31; K5B3z; K5B4.; K5B40; K5B41; K5B5.; K5B51; K5B6.; K5B7.; K5By.; K5Byz; K5Bz.; Kyu9G; K26y4; K5B10; K5B11; K5By0; K5By1; K5Byz; Read v3 - K5B..; XE0ex; K5B00; K5B01; K5B0z; K5B1.; K5B1z; K5B2.; K5B20; K5B21; X4099; K5B2z; K5B3.; K5B30; K5B31; K5B3z; K5B4.; K5B40; K5B41; K5B5.; K5B51; X409B; XaZ6K; K5By.; XE0ey; K5Bz.; Kyu9G; K26y4; K5B10; K5B11; K5By0; K5By1; K5Byz		
Anovulatory female infertility (F_ANOV)	ICD9 - 6283; 6284; ICD10 - N971; N972; N973	Read v2 - K5B2.; K5B20; K5B21; K5B23; K5B2z; K5B3.; K5B30; K5B31; K5B3z; K5B4.; K5B40; K5B41; K5B5.; K5B51; Read v3 - K5B2.; K5B20; K5B21; X4099; K5B2z; K5B3.; K5B30; K5B31; K5B3z; K5B4.; K5B40; K5B41; K5B5.; K5B51		
Anatomical female infertility (F_ANAT)	ICD9 - 6280; ICD10 - N970	Read v2 - K5B0.; K5B00; K5B01; K5B0z; Read v3 - XE0ex; K5B00; K5B01; K5B0z		

contd. Case definitions for the various female and male infertility analyses

Infertility category	Case inclusions		Case exclusions	
	ICD codes	Read codes	ICD codes	Read codes
Idiopathic female infertility by exclusion (F_EXCL)	ICD9 - 6284; 6288; 6289; ICD10 - N97; N974; N978; N979	Read v2 - K5B..; K5B6.; K5B7.; K5By.; K5Byz; K5Bz.; Kyu9G; K26y4; K5B10; K5B11; K5By0; K5By1; K5Byz; Read v3 - K5B..; X409B; XaZ6K; K5By.; XE0ey; K5Bz.; Kyu9G; K26y4; K5B10; K5B11; K5By0; K5By1; K5Byz	ICD9 - 6280; 6283; 6284; 2564; 218; 617; 6170; 6171; 6172; 6173; 6174; 6175; 6176; 6178; 6179; ICD10 - N970; N971; N972; N973; E28.2; D25; D25.0; D25.1; D25.2; D25.9; N80; N80.0; N80.1; N80.2; N80.3; N80.4; N80.5; N80.6; N80.8; N80.9	Read v2 - K5B0.; K5B00; K5B01; K5B0z; K5B1.; K5B1z; K5B2.; K5B20; K5B21; K5B23; K5B2z; K5B3.; K5B30; K5B31; K5B3z; K5B4.; K5B40; K5B41; K5B5.; K5B51; Read v3 - XE0ex; K5B00; K5B01; K5B0z; K5B1.; K5B1z; K5B2.; K5B20; K5B21; X4099; K5B2z; K5B3.; K5B30; K5B31; K5B3z; K5B4.; K5B40; K5B41; K5B5.; K5B51
Idiopathic female infertility by inclusion (F_INCL)	ICD9 - 6284; 6288; ICD10 - N978	Read v2 - K5By.; K5Byz; Kyu9G; Read v3 - K5By.; K5Byz; Kyu9G		
Male infertility of all causes (M_ALL)	ICD9 - 606; ICD10 - N46	Read v2 - 4914; 4915; 4916; 4917; K26..; K260.; K261.; K2610; K262.; K26y.; K26y2; K26yz; K26z.; K5B6.; K26y4; K5B10; K5B11; K5By0; K5By1; K5Byz; Read v3 - 4914; 4915; 4916; 4917; K26..; XE0eA; K261.; XaXbT; XaXUx; K26y.; K26yz; K26z.; X409B; K26y4; K5B10; K5B11; K5By0; K5By1; K5Byz		

Table C.2: List of hormone traits in the GWAS Catalog, accessed on 27 March 2023, with number of associations per trait and PubMed Identifiers (PMIDs) of reporting studies.

Hormone trait	Number of unique SNPs in GWASCatalog	Study PMIDs
17-hydroxyprogesterone measurement	27	31169883, 34822396
androstenedione measurement, estrone measurement	14	28429243
anti-Mullerian hormone measurement	13	30649302, 35274129
dehydroepiandrosterone sulphate	63	26014426, 31169883, 28887542, 34748635, 35347128
dihydrotestosterone measurement	2	22936694
estradiol measurement	89	22675492, 23518928, 23251661, 29325096, 24049095, 26014426, 31169883, 34255042, 34822396, 35192695, 32042192
estrogen measurement, response to aromatase inhibitor	3	32701512
estrone conjugate measurement	4	28429243, 33495599
estrone conjugate measurement, estrone measurement	1	28429243
estrone measurement	21	29325096, 28429243
follicle stimulating hormone measurement	31	26284813, 24049095, 26014426, 32185793, 35305013, 34239129
hormone measurement, cortisol:DHEAS ratio measurement	1	23696881
hormone measurement, dehydroepiandrosterone measurement	1	23696881
hormone measurement, dehydroepiandrosterone sulphate measurement	3	26014426
hormone measurement, estradiol measurement	1	26014426
hormone measurement, follicle stimulating hormone measurement	2	26014426
hormone measurement, free androgen index	1	26014426
hormone measurement, progesterone measurement	3	26014426
hormone measurement, sex hormone-binding globulin measurement	1	26014426
hormone measurement, testosterone measurement	2	26014426
human Chorionic Gonadotropin measurement	17	28240269, 35870639, 29875488, 34814699, 36168886, 34239129
Inhibin B measurement	1	32341457
luteinizing hormone measurement	12	26284813, 26014426, 32185793, 35870639, 29875488, 34814699, 36168886, 34239129
luteinizing hormone measurement, hormone measurement	2	26014426
oxytocin measurement	1	31361301
progesterone measurement	41	26014426, 31169883, 34822396
sex hormone-binding globulin measurement	2031	22675492, 18464913, 26284813, 22829776, 24049095, 26014426, 32042192, 34662886, 35078996, 34226706, 33980691, 35192695, 34814699, 36168886, 34321204, 33462484, 35264221
testosterone measurement	1600	22675492, 23251661, 26284813, 24049095, 21998597, 26014426, 31169883, 33407712, 34822396, 34337532, 22936694, 35192695, 32042192, 33587031, 35347128, 34321204, 33462484

Table C.3: Colocalisation of eQTLs from GTEx metabolic, endocrine, and reproductive tissues with female infertility loci. Genes with transcription start sites within 1Mb of the EUR-ancestry lead variant were assessed. Under the assumption of a single causal variant at the locus, all of the loci presented here have posterior probabilities (PPs) > 95% for association of both the eQTL and infertility trait, but with different causal variants, in the region.

RSID	chr:pos (hg38)	Gene name	Tissues	Posterior probability of H3 (shared locus, different causal SNPs)
Female infertility of all causes (F-ALL)				
rs61768001	chr1:22139327	EIF4G3	Testis	99.50%
		NBPF3	Adipose Visceral Omentum, Adrenal Gland, Prostate, Pituitary, Thyroid, Liver, Pancreas, Vagina, Testis, Uterus, Ovary, Adipose Subcutaneous	100.00%
		USP48	Adipose Visceral Omentum, Pancreas, Adipose Subcutaneous, Pituitary, Thyroid	100.00%
		LDLRAD2	Adipose Visceral Omentum	97.10%
		HSPG2	Pituitary	98.90%
		CELA3B	Pancreas	100.00%
		CELA3A	Pancreas	100.00%
		CDC42	Testis	100.00%
		WNT4	Thyroid	100.00%
		ZBTB40	Thyroid	99.80%
		EPHB2	Adrenal Gland, Thyroid	100.00%
		LACTBL1	Testis	100.00%
		TEX46	Testis	100.00%
rs10200851	chr2:11581956	NOL10	Pituitary	98.30%
		ATP6V1C2	Testis	100.00%
		PDIA6	Testis	100.00%
		C2orf50	Pituitary	100.00%
		SLC66A3	Thyroid	97.30%
		E2F6	Testis	100.00%
		LPIN1	Liver	99.70%
rs6938404	chr6:151222906	PLEKHG1	Testis	96.20%
		MTHFD1L	Testis	96.30%
		AKAP12	Thyroid	99.50%
		RMND1	Thyroid, Adipose Visceral Omentum, Pancreas	100.00%
		CCDC170	Liver, Pancreas	99.90%
		ESR1	Thyroid	99.90%
		SYNE1	Liver	96.30%
rs17803970	chr6:152232583	AKAP12	Thyroid	99.50%
		RMND1	Thyroid, Adipose Visceral Omentum, Pancreas	100.00%
		CCDC170	Liver, Pancreas	99.90%
		ESR1	Thyroid	99.90%
		SYNE1	Liver	96.30%
		VIP	Testis, Thyroid, Adipose Subcutaneous	100.00%
		MTRF1L	Pituitary, Prostate, Adrenal Gland, Thyroid, Uterus, Pancreas, Ovary, Vagina, Adipose Visceral Omentum, Testis, Adipose Subcutaneous, Liver	100.00%
		RGS17	Thyroid, Prostate, Adipose Visceral Omentum,	100.00%
rs1964514	chr8:109463457	TMEM74	Adipose Subcutaneous	99.20%
		NUDCD1	Thyroid, Testis	100.00%
		ENY2	Prostate	96.80%
		PKHD1L1	Adipose Visceral Omentum, Pancreas, Thyroid	100.00%
		EBAG9	Thyroid	98.10%

contd. Colocalisation of eQTLs from GTEx metabolic, endocrine, and reproductive tissues with female infertility loci

RSID	chr:pos (hg38)	Gene name	Tissues	Posterior probability of H3 (shared locus, different causal SNPs)
Anovulatory female infertility (F-ANOV)				
rs72827480	chr2:120388925	TMEM37	Thyroid, Adrenal Gland	97.10%
		SCTR	Adipose Visceral Omentum, Testis, Thyroid, Pancreas, Adipose Subcutaneous	97.10%
		CFAP221	Thyroid, Pancreas	97.10%
		TMEM177	Liver, Thyroid, Pituitary	97.10%
		EPB41L5	Adrenal Gland	97.10%
		TMEM185B	Liver, Adipose Subcutaneous, Testis	97.10%
		RALB	Pituitary, Testis, Adipose Subcutaneous, Thyroid	97.10%
		INHBB	Adipose Subcutaneous	97.10%
		TFCP2L1	Pancreas	97.10%
rs1852684	chr2:145068818	GTDC1	Adipose Visceral Omentum	95.20%
rs9696009	chr9:123856954	RABGAP1	Thyroid, Testis	100%
		STRBP	Testis, Thyroid	100%
		CRB2	Thyroid	99.90%
		DENND1A	Adipose Subcutaneous	99.90%
		NEK6	Pancreas, Prostate, Thyroid	100%
		PSMB7	Pancreas	99.80%
		ADGRD2	Pituitary	100%
		OLFML2A	Pancreas, Adipose Subcutaneous	100%

contd. Colocalisation of eQTLs from GTEx metabolic, endocrine, and reproductive tissues with female infertility loci

RSID	chr:pos (hg38)	Gene name	Tissues	Posterior probability of H3 (shared locus, different causal SNPs)
Anovulatory female infertility (F-ANOV)				
rs9902027	chr17:7537667	PITPNM3	Testis, Thyroid	99.70%
		KIAA0753	Pancreas, Pituitary, Prostate, Testis, Thyroid, Adipose Subcutaneous, Ovary, Adipose Visceral Omentum	99.70%
		MED31	Pituitary, Testis, Thyroid	99.80%
		C17orf100	Testis, Thyroid, Pituitary	99.80%
		SLC13A5	Adipose Subcutaneous, Adipose Visceral Omentum	99.80%
		XAF1	Ovary	99.80%
		FBXO39	Adipose Visceral Omentum, Thyroid, Adipose Subcutaneous, Uterus	99.80%
		TEKT1	Testis	99.80%
		ALOX12	Thyroid, Adipose Subcutaneous, Adrenal Gland	99.80%
		RNASEK	Adipose Visceral Omentum, Adipose Subcutaneous, Pancreas, Testis, Thyroid	99.80%
		SLC16A13	Adipose Subcutaneous	96.20%
		SLC16A11	Thyroid	99.80%
		CLEC10A	Adipose Subcutaneous	97.20%
		ASGR2	Thyroid, Adipose Subcutaneous, Adipose Visceral Omentum	99.80%
		DLG4	Testis	98.10%
		ACADVL	Pituitary, Adipose Subcutaneous	99.80%
		CTDNEP1	Adipose Visceral Omentum, Adipose Subcutaneous, Testis, Pancreas, Adrenal Gland	99.80%
		ELP5	Testis, Adrenal Gland, Adipose Subcutaneous, Adipose Visceral Omentum	99.80%
		CLDN7	Liver, Thyroid, Adipose Visceral Omentum, Adrenal Gland, Adipose Subcutaneous	99.80%
		SLC2A4	Liver	99.80%
		YBX2	Thyroid, Liver	99.80%
		EIF5A	Pancreas, Liver, Adrenal Gland, Adipose Visceral Omentum, Pituitary, Thyroid, Adipose Subcutaneous, Prostate, Testis	99.80%
		GPS2	Thyroid	99.70%
		NLGN2	Thyroid	99.80%
		SPEM2	Testis	99.80%
		SPEM3	Testis	99.70%
		TMEM102	Testis	99.80%
		FGF11	Thyroid	99.80%
		CHRNB1	Pituitary, Prostate, Testis, Ovary, Adipose Visceral Omentum, Thyroid, Adipose Subcutaneous	99.80%
		ZBTB4	Thyroid	99%
		SLC35G6	Testis	99.80%
		TNFSF12	Adipose Subcutaneous, Adipose Visceral Omentum	99.80%
		TNFSF13	Adipose Subcutaneous	99.80%
		SENP3	Testis	99.80%

contd. Colocalisation of eQTLs from GTEx metabolic, endocrine, and reproductive tissues with female infertility loci

		FXR2	Testis	99.70%
		SHBG	Adrenal Gland	99.70%
		SAT2	Adipose Visceral Omentum, Ovary, Pancreas, Thyroid	99.80%
		ATP1B2	Pituitary	99.70%
		EFNB3	Thyroid, Adipose Visceral Omentum, Pancreas, Adipose Subcutaneous, Testis, Adrenal Gland	99.80%
		DNAH2	Pituitary	96.60%
		KDM6B	Testis, Adipose Subcutaneous	99.80%
		NAA38	Adipose Subcutaneous, Thyroid, Prostate, Pituitary	99.80%
		KCNAB3	Adipose Subcutaneous, Thyroid, Pancreas	99.80%
		CNTROB	Testis, Thyroid, Adipose Subcutaneous, Pancreas	99.80%
		GUCY2D	Testis	99.80%
		HES7	Testis	99.80%
		VAMP2	Testis	99.80%
		CTC1	Adipose Subcutaneous, Thyroid, Adipose Visceral Omentum	99.80%
		PFAS	Adipose Subcutaneous	95.20%
		SLC25A35	Adipose Subcutaneous	99.70%
		ODF4	Testis	99.80%
		KRBA2	Adipose Subcutaneous, Testis, Adipose Visceral Omentum, Ovary, Thyroid	99.80%
		NDEL1	Thyroid	99.80%

contd. Colocalisation of eQTLs from GTEx metabolic, endocrine, and reproductive tissues with female infertility loci

RSID	chr:pos (hg38)	Gene name	Tissues	Posterior probability of H3 (shared locus, different causal SNPs)
Anovulatory female infertility (F-ANOV)				
rs143459581	chr22:28068862	PITPNB	Adrenal Gland	98.90%
		TTC28	Thyroid, Testis	99.70%
		CHEK2	Liver, Testis	99.70%
		HSCB	Thyroid	99.70%
		CCDC117	Thyroid	99.50%
		XBP1	Pancreas	99.70%
		C22orf31	Thyroid	99.70%
rs17879961	chr22:28725099	PITPNB	Adrenal Gland	98.90%
		TTC28	Thyroid, Testis	99.70%
		CHEK2	Liver, Testis	99.70%
		HSCB	Thyroid	99.70%
		CCDC117	Thyroid	99.50%
		XBP1	Pancreas	99.70%
		C22orf31	Thyroid	99.70%
		KREMEN1	Adrenal Gland, Pancreas, Liver, Thyroid, Adipose Visceral Omentum, Adipose Subcutaneous, Ovary	99.70%
		EMID1	Pancreas, Adrenal Gland, Testis	99.70%
		RHBDD3	Thyroid, Pancreas, Adipose Visceral Omentum	99.70%
		GAS2L1	Adipose Subcutaneous, Thyroid	99.70%
		RASL10A	Adipose Subcutaneous	99.70%
		AP1B1	Adipose Subcutaneous, Adipose Visceral Omentum	99.70%
		RFPL1	Prostate	96.50%
		NEFH	Testis	97.60%
		THOC5	Adrenal Gland, Thyroid, Adipose Subcutaneous, Adipose Visceral Omentum, Ovary, Pituitary, Testis, Uterus, Prostate, Vagina	99.30%
		NIPSNAP1	Adrenal Gland, Thyroid, Testis, Adipose Visceral Omentum, Pancreas, Adipose Subcutaneous	99.30%
		NF2	Testis	99.10%

contd. Colocalisation of eQTLs from GTEx metabolic, endocrine, and reproductive tissues with female infertility loci

RSID	chr:pos (hg38)	Gene name	Tissues	Posterior probability of H3 (shared locus, different causal SNPs)
Idiopathic female infertility, exclusion definition (F-EXCL)				
rs61768001	chr1:22139327	EIF4G3	Pancreas	99.80%
		NBPF3	Adipose Visceral Omentum, Testis, Liver, Prostate, Uterus, Ovary, Pancreas, Adrenal Gland, Vagina, Thyroid, Pituitary, Adipose Subcutaneous	99.90%
		USP48	Pituitary, Pancreas, Thyroid, Adipose Visceral Omentum, Adipose Subcutaneous	99.90%
		LDLRAD2	Adipose Subcutaneous	99.90%
		HSPG2	Prostate, Testis, Pancreas, Adipose Subcutaneous, Thyroid, Adipose Visceral Omentum	99.90%
		CELA3B	Pancreas	99.90%
		CELA3A	Pancreas	99.90%
		CDC42	Testis	99.90%
		WNT4	Thyroid	99.90%
		ZBTB40	Prostate	99.80%
		EPHB2	Adrenal Gland, Thyroid	99.90%
		LACTBL1	Testis	99.90%
		TEX46	Testis	99.80%
rs17378154	chr8:109568721	TMEM74	Adipose Subcutaneous	99.20%
		NUDCD1	Testis, Thyroid	100%
		ENY2	Prostate	96.80%
		PKHD1L1	Thyroid, Pancreas, Adipose Visceral Omentum	100%
		EBAG9	Thyroid	98.10%

contd. Colocalisation of eQTLs from GTEx metabolic, endocrine, and reproductive tissues with female infertility loci

RSID	chr:pos (hg38)	Gene name	Tissues	Posterior probability of H3 (shared locus, different causal SNPs)
Idiopathic female infertility, inclusion definition (F-INCL)				
rs61768001	chr1:22139327	EIF4G3	Testis	99.50%
		NBPF3	Adipose Visceral Omentum, Pancreas, Thyroid, Ovary, Prostate, Uterus, Adipose Subcutaneous, Pituitary, Testis, Vagina, Liver, Adrenal Gland	100%
		USP48	Adipose Subcutaneous, Pituitary, Thyroid, Adipose Visceral Omentum, Pancreas	100%
		LDLRAD2	Adipose Visceral Omentum	97.10%
		HSPG2	Pituitary	98.90%
		CELA3B	Pancreas	100%
		CELA3A	Pancreas	100%
		CDC42	Testis	100%
		WNT4	Thyroid	100%
		ZBTB40	Thyroid	99.80%
		EPHB2	Thyroid, Adrenal Gland	100%
		LACTBL1	Testis	100%
		TEX46	Testis	100%
rs11692588	chr2:11544358	NOL10	Testis, Thyroid	98.70%
		ATP6V1C2	Testis	98.70%
		PDIA6	Testis	98.70%
		C2orf50	Pituitary	98.70%
		SLC66A3	Pituitary, Pancreas	98.70%
		E2F6	Testis	98.70%
		LPIN1	Pituitary, Prostate, Adipose Subcutaneous, Ovary, Thyroid, Pancreas, Adipose Visceral Omentum	98.70%
rs851982	chr6:151703850	MTHFD1L	Testis, Thyroid	99.30%
		AKAP12	Thyroid	98.90%
		RMND1	Pancreas, Thyroid	99.40%
		CCDC170	Thyroid, Testis	99.40%
		ESR1	Testis	99.40%
		SYNE1	Adrenal Gland	99.40%
rs17378154	chr8:109568721	TMEM74	Adipose Subcutaneous	98.70%
		NUDCD1	Testis, Thyroid	99.50%
		ENY2	Prostate	96.40%
		PKHD1L1	Thyroid, Adipose Visceral Omentum, Pancreas	99.50%
		EBAG9	Thyroid	97.70%

Table C.4: Lowest infertility-associated P-value for variants in genomic windows under directional selection. F-ALL = female infertility of all causes, F-ANAT = anatomical infertility, F-ANOV = anovulatory infertility, F-EXCL = idipathic female infertility by exclusion, F-INCL = idiopathic female infertility by inclusion, M-ALL = male infertility of all causes, CMS = Composite of Multiple Signals.

Method	Region (hg19)	LENGTH (kb)	F-ALL		F-ANAT		F-ANOV		F-EXCL		F-INCL		M-ALL	
			N SNPs	Min PVAL	N SNPs	Min PVAL	N SNPs	Min PVAL	N SNPs	Min PVAL	N SNPs	Min PVAL	N SNPs	Min PVAL
CMS (Selection over 50,000 years)														
CMS	chr1:1490074-1509034	18.96	74	0.01893	73	0.002276	66	0.001507	72	0.01994	66	0.001088	69	0.002774
CMS	chr1:30707291-30724056	16.765	91	0.03639	91	0.0194	92	0.009314	91	0.01516	92	0.01107	92	0.0335
CMS	chr1:35494810-35612236	117.426	219	0.01313	224	0.005897	211	0.01116	220	0.01392	217	0.005089	223	0.02021
CMS	chr1:151369475-151403916	34.441	69	0.07816	69	0.007803	65	0.08033	69	0.02221	67	0.1733	67	0.1126
CMS	chr2:18365580-18390599	25.019	127	0.008531	127	0.03288	120	0.02099	125	0.005863	126	0.01396	128	0.0242
CMS	chr2:69208834-69222144	13.31	97	0.0003951	96	0.06615	95	0.006994	97	0.0137	95	0.07083	96	0.01933
CMS	chr2:74531420-74690039	158.619	255	0.01429	256	0.001484	245	0.06182	254	0.01475	251	0.01322	251	0.01797
CMS	chr2:136407479-136685228	277.749	593	5.00E-04	590	0.01024	588	0.006292	592	0.0009196	590	0.001522	591	0.003286
CMS	chr2:136906692-136915698	9.006	19	0.1482	19	0.004954	19	0.007159	18	0.1455	19	0.1701	19	0.1622
CMS	chr2:163520526-163556004	35.478	106	0.00211	107	0.04135	105	0.02711	106	0.0179	105	0.04199	106	0.03844

contd. Lowest infertility-associated P-value for variants in genomic windows under directional selection

Method	Region (hg19)	LENGTH (kb)	F-ALL		F-ANAT		F-ANOV		F-EXCL		F-INCL		M-ALL	
			N SNPs	Min PVAL	N SNPs	Min PVAL	N SNPs	Min PVAL	N SNPs	Min PVAL	N SNPs	Min PVAL	N SNPs	Min PVAL
CMS (Selection over 50,000 years)														
CMS	chr2:178565129-178573882	8.753	37	0.08732	35	0.03997	35	0.0772	37	0.009949	35	0.1778	36	0.018
CMS	chr2:206204079-206248179	44.1	250	0.01495	247	0.009945	227	0.005986	249	0.02092	229	0.001424	243	0.01967
CMS	chr3:3450253-3450683	0.43	4	0.7316	4	0.01203	4	0.3852	4	0.1157	4	0.3215	4	0.9547
CMS	chr3:110537075-110552967	15.892	52	0.03859	52	0.007128	52	0.09316	52	0.1996	53	0.2552	52	0.01347
CMS	chr3:123291091-123417587	126.496	392	4.91E-05	379	0.008101	375	0.001516	392	6.07E-05	364	0.02219	391	0.01164
CMS	chr3:130209258-130220079	10.821	85	0.00113	81	0.01926	78	0.007444	85	0.007559	77	0.0136	82	0.0209
CMS	chr4:5332008-5342822	10.814	80	0.06806	83	0.0165	78	0.001009	77	0.06133	78	0.06246	80	0.01775
CMS	chr4:61387484-61451788	64.304	475	0.00111	465	0.003174	453	0.005213	473	0.0002904	453	0.00571	473	0.05036
CMS	chr4:72147944-72392006	244.062	562	0.005935	561	0.006215	534	0.001807	559	0.001545	545	0.02875	562	0.008731
CMS	chr4:176370319-176371655	1.336	5	0.4205	5	0.5676	5	0.5292	5	0.219	5	0.5427	5	0.7573

contd. Lowest infertility-associated P-value for variants in genomic windows under directional selection

Method	Region (hg19)	LENGTH (kb)	F-ALL		F-ANAT		F-ANOV		F-EXCL		F-INCL		M-ALL	
			N SNPs	Min PVAL	N SNPs	Min PVAL	N SNPs	Min PVAL	N SNPs	Min PVAL	N SNPs	Min PVAL	N SNPs	Min PVAL
CMS (Selection over 50,000 years)														
CMS	chr8:16230615-16261374	30.759	145	0.01573	143	0.03613	138	0.06728	145	0.04825	140	0.06374	144	0.02128
CMS	chr9:135170185-135194829	24.644	117	0.08662	116	0.01825	115	0.04077	116	0.01854	114	0.01372	116	0.0005726
CMS	chr10:83611881-83688159	76.278	213	0.006222	215	0.006572	211	0.002899	212	0.006825	212	0.007501	214	0.02436
CMS	chr10:118143363-118148935	5.572	19	0.01563	19	0.03633	19	0.08087	18	0.0003345	19	0.02674	19	0.01357
CMS	chr12:66084230-66099730	15.5	39	0.005743	40	0.01601	40	0.01451	39	0.00857	40	0.004465	39	0.001398
CMS	chr12:73081929-73111465	29.536	120	0.08847	118	0.02333	107	0.03315	122	0.05753	107	0.07944	116	0.02099
CMS	chr12:80193361-80354208	160.847	323	0.02098	329	9.01E-05	268	0.01262	289	0.02032	274	0.01176	317	0.00657
CMS	chr15:28365618-28516084	150.466	253	0.000474	247	0.0006093	216	0.04895	249	0.02299	214	0.03559	232	0.0001018
CMS	chr15:45145764-45258860	113.096	438	0.0002668	476	0.04204	359	0.001939	439	0.0008286	423	0.01884	431	0.003912
CMS	chr15:48539587-48633153	93.566	208	0.0006331	208	0.008309	201	0.05186	207	0.01058	201	0.07656	204	0.0303

contd. Lowest infertility-associated P-value for variants in genomic windows under directional selection

Method	Region (hg19)	LENGTH (kb)	F-ALL		F-ANAT		F-ANOV		F-EXCL		F-INCL		M-ALL	
			N SNPs	Min PVAL	N SNPs	Min PVAL	N SNPs	Min PVAL	N SNPs	Min PVAL	N SNPs	Min PVAL	N SNPs	Min PVAL
CMS (Selection over 50,000 years)														
CMS	chr18:7491752-7651318	159.566	546	0.0002087	551	0.0005739	531	0.0008525	544	0.00273	536	0.01058	545	0.001367
CMS	chr19:33533258-33626009	92.751	366	0.001254	366	0.05123	355	0.009693	368	0.007229	356	0.004249	368	0.005255
CMS	chr20:53342055-53393426	51.371	145	0.00917	146	0.0262	141	0.01339	145	0.006652	143	0.02331	144	0.01672

contd. Lowest infertility-associated P-value for variants in genomic windows under directional selection. aDNA = ancient DNA

Method	Region (hg19)	LENGTH (kb)	F-ALL		F-ANAT		F-ANOV		F-EXCL		F-INCL		M-ALL	
			N SNPs	Min PVAL	N SNPs	Min PVAL	N SNPs	Min PVAL	N SNPs	Min PVAL	N SNPs	Min PVAL	N SNPs	Min PVAL
aDNA (Selection over 10,000 years)														
aDNA	chr2:135285279-135602156	316.878	821	0.003987	813	0.006027	794	0.00319	822	0.006249	799	0.004387	811	0.002241
aDNA	chr2:135631400-137015991	1384.591	2875	0.0003642	2859	0.003785	2850	0.0002166	2872	8.78E-05	2855	0.000466	2867	0.00189
aDNA	chr4:38736550-38846885	110.335	633	0.07918	645	6.78E-05	587	0.009985	615	0.01464	600	0.08249	614	0.002139
aDNA	chr5:33832958-33966133	133.175	320	0.02195	319	0.01531	311	0.00162	320	0.01905	310	0.008227	320	0.00448
aDNA	chr5:131358999-131804347	445.349	1466	0.0007991	1471	0.009665	1412	0.0002558	1461	0.002743	1430	0.0007584	1464	0.004494
aDNA	chr6:28309590-28700352	390.762	1203	0.005644	1212	0.001775	1165	0.0075	1194	0.004575	1192	0.01968	1248	0.001569
aDNA	chr6:29918099-30103923	185.824	3650	0.001568	3637	0.0003276	3440	0.00164	3647	0.001266	3499	0.001055	3529	0.0003816
aDNA	chr6:30763893-31127562	363.669	3447	0.0008072	3406	0.0004721	3327	0.003989	3453	0.0005519	3343	0.0004393	3453	0.008341
aDNA	chr6:31140047-33056731	1916.684	40501	0.00018	39541	3.49E-05	31409	4.45E-05	40426	0.0002073	31592	5.45E-05	34834	0.0006702
aDNA	chr11:61548874-61616012	67.138	132	0.002242	131	0.003913	128	0.001921	133	0.02765	128	0.01163	131	0.007135
aDNA	chr11:88513636-88911696	398.06	1343	0.003112	1351	0.001268	1310	0.0006118	1341	0.002366	1305	0.0002719	1329	0.01223
aDNA	chr12:111884608-112610714	726.106	1148	0.001693	1142	0.00476	1074	0.0139	1133	0.01027	1092	0.01284	1127	0.000586

contd. Lowest infertility-associated P-value for variants in genomic windows under directional selection. SDS = Singleton Density Score

Method	Region (hg19)	LENGTH (kb)	F-ALL		F-ANAT		F-ANOV		F-EXCL		F-INCL		M-ALL	
			N SNPs	Min PVAL	N SNPs	Min PVAL	N SNPs	Min PVAL	N SNPs	Min PVAL	N SNPs	Min PVAL	N SNPs	Min PVAL
SDS (Selection over 2000-3000 years)														
SDS	chr2:136000000-137000000	1000	1950	5.00E-04	1943	0.003785	1939	0.002795	1948	0.0009196	1942	0.0007814	1948	0.00189
SDS	chr6:25000000-28600000	3632.451	12840	2.25E-05	12568	0.0001734	12144	0.000583	12816	0.000245	12131	0.0007083	12597	9.37E-05
SDS	chr10:49700000-50500000	799.998	3396	0.0007485	3363	0.003526	3265	0.0002978	3397	0.0002809	3277	0.0009304	3360	0.0004081

Table C.5: Lead variants associated with female or male infertility in all-ancestry GWASs, along with total number of cases and controls, effect direction in each contributing cohort (+, -, or ? for missing, where the order of cohorts is displayed in the title row), and P-value for heterogeneity in effect estimates between cohorts.

RSID	chr:pos (hg38)	Nearest TSS	Average MAF	OR (95% CI)	P-value	Number of cases	Number of controls	Direction	Heterogeneity P-value
Female infertility of all causes (F-ALL) - FinnGen, UKBB, EstBB, Danish, deCODE, Genes&Health									
rs61768001	chr1:22139327	WNT4	0.166	0.909 (0.891 - 0.927)	2.25E-21	40900	672171	------	0.711
rs10200851	chr2:11581956	NTSR2	0.458	0.951 (0.937 - 0.965)	2.90E-11	40900	672171	------	0.476
rs6938404	chr6:151222906	AKAP12	0.453	0.958 (0.943 - 0.973)	3.88E-08	39724	650824	-----?	0.0173
rs17803970	chr6:152232583	SYNE1	0.0836	1.09 (1.06 - 1.12)	1.91E-10	40900	672171	+++++-	0.26
rs1964514	chr8:109463457	EBAG9	0.0595	1.13 (1.09 - 1.16)	3.01E-14	40900	672171	++++-+	0.151
Anatomical female infertility (F-ANAT) - FinnGen, EstBB, Danish, deCODE, Genes&Health									
rs340879	chr1:213983171	PROX1	0.418	0.906 (0.874 - 0.939)	4.95E-08	6822	541524	-----	0.589
Anovulatory female infertility (F-ANOV) - FinnGen, EstBB, Danish, Genes&Health									
rs72665317	chr1:22040580	CDC42	0.19	0.875 (0.839 - 0.913)	7.76E-10	7315	316158	----	0.141
rs72827480	chr2:120388925	INHBB	0.401	0.905 (0.873 - 0.938)	4.20E-08	6721	294228	---?	0.931
rs1852684	chr2:145068818	ZEB2	0.367	1.12 (1.08 - 1.16)	9.25E-10	7315	316158	++++	0.46
rs552953683	chr8:102898586	AZIN1	0.0024	0.341 (0.234 - 0.498)	2.54E-08	6176	222978	--??	0.225
rs9696009	chr9:123856954	DENND1A	0.0777	1.21 (1.14 - 1.29)	6.87E-10	7315	316158	++++	0.00127
rs9902027	chr17:7537667	TNFSF12	0.255	0.895 (0.86 - 0.931)	4.06E-08	6721	294228	---?	0.063
rs143459581	chr22:28068862	PITPNB	0.0419	1.3 (1.19 - 1.43)	1.21E-08	6176	222978	++??	0.21
rs17879961	chr22:28725099	HSCB	0.0389	0.739 (0.673 - 0.811)	1.55E-10	6176	222978	--??	0.112
Idiopathic female infertility, exclusion definition (F-EXCL) - FinnGen, UKBB, EstBB, Danish, deCODE, Genes&Health									
rs61768001	chr1:22139327	WNT4	0.165	0.923 (0.902 - 0.946)	7.49E-11	24046	783011	------	0.415
rs111597692	chr8:109039973	TRHR	0.0323	1.16 (1.1 - 1.22)	1.51E-08	23122	761412	+++++?	0.182
rs17378154	chr8:109568721	EBAG9	0.059	1.13 (1.09 - 1.17)	1.64E-10	24046	783011	++++++	0.609
Idiopathic female infertility, inclusion definition (F-INCL) - FinnGen, EstBB, Danish, Genes&Health									
rs61768001	chr1:22139327	WNT4	0.17	0.87 (0.839 - 0.903)	6.87E-14	10794	430296	----	0.655
rs11692588	chr2:11544358	GREB1	0.358	0.919 (0.892 - 0.947)	2.98E-08	10201	408366	---?	0.704
rs190290095	chr4:39786858	UBE2K	0.0022	0.227 (0.137 - 0.375)	7.60E-09	7487	339285	--??	0.437
rs851982	chr6:151703850	ESR1	0.428	0.921 (0.895 - 0.947)	7.60E-09	10794	430296	----	0.0349
rs17378154	chr8:109568721	EBAG9	0.0565	1.18 (1.11 - 1.25)	2.47E-08	10794	430296	++++	0.843
rs74156208	chr10:61509370	TMEM26	0.184	1.10 (1.06 - 1.14)	4.96E-08	7487	339285	--??	0.195
Male infertility of all causes (F-ALL) - FinnGen, UKBB, EstBB, Danish, Genes&Health									
rs150639836	chr10:53879806	MBL2	0.0109	0.505 (0.402 - 0.636)	5.72E-09	2723	148681	-?-??	0.528
rs139862664	chr10:116879589	HSPA12A	0.0072	0.388 (0.277 - 0.543)	3.29E-08	3313	357824	---??	0.46

D

Supplement to Chapter 5

Table D.1: List of genome-wide association studies contributing to female reproductive condition summary statistics, with sample sizes and case/control definitions.

Phenotype		Source	Number of cases	Number of controls	Cohorts	Phenotype definition
Endometriosis		Rahmioglu *et al.* (2023)	57248	698764	Meta-analysis of 22 European-ancestry cohorts, including FinnGen and UK Biobank	Cases: mixed surgically confirmed, medical records, and self-reported; Controls: mixed male and female
Heavy menstrual bleeding		Gallagher *et al.* (2019)	9813	210946	UK Biobank (European-ancestry)	Cases: hospital-linked records; Controls: female
		FinnGen Data Freeze 9	21496	107564	FinnGen	Cases: ICD8 code 626[67], ICD9 code 6262A|6262B|6266A, ICD10 code N92.0 in medical records; Controls: female
Infertility	**All cause**	In-house	38894	546837	Meta-analysis of 5 European-ancestry cohorts, including FinnGen and UK Biobank	Cases: mixed medical records, self-reported; Controls: female
	Anovulatory	In-house	6774	558346	Meta-analysis of 5 European-ancestry cohorts, including FinnGen and UK Biobank	Cases: infertility caused by anovulation in medical records; Controls: female
	Anatomical	In-house	5361	674609	Meta-analysis of 5 European-ancestry cohorts, including FinnGen and UK Biobank	Cases: infertility caused by tubal, cervical, or uterine abnormalities in medical records; Controls: female
	Idiopathic, exclusion	In-house	23225	660004	Meta-analysis of 5 European-ancestry cohorts, including FinnGen and UK Biobank	Cases: all female infertility cases without known cause: anovulation, anatomical, PCOS, endometriosis; Controls: female
	Idiopathic, inclusion	In-house	10279	672503	Meta-analysis of 5 European-ancestry cohorts, including FinnGen and UK Biobank	Cases: ICD10 code N97.8 (or equivalent) in medical records; Controls: female
		Tyrmi *et al.* (2021)	3609	229818	Meta-analysis of FinnGen and Estonian Biobank	Cases: ICD8 code 256.90, ICD9 code 256.4, ICD10 code E28.2; Controls: female
		Day *et al.* (2018)	10074	103164	Meta-analysis of 7 European-ancestry cohorts	Cases: mixed NIH or Rotterdam Criteria, self-reported; Controls: mixed male and female

contd. List of genome-wide association studies contributing to female reproductive condition summary statistics, with sample sizes and case/control definitions.

Phenotype	Source	Number of cases	Number of controls	Cohorts	Phenotype definition
PCOS	Day *et al.* (2018)	10074	103164	Meta-analysis of 7 European-ancestry cohorts	Cases: mixed NIH or Rotterdam Criteria, self-reported; Controls: mixed male and female
	In-house	784	97285	UK Biobank (European-ancestry)	Cases: mixed medical records, self-reported, and symptom-based; Controls: female without menstrual disorders (heavy menstruation, irregular menstruation, or no menstruation), endometriosis, fibroids, miscarriage, or pre-eclampsia
Preeclampsia or eclampsia	FinnGen Data Freeze 9	7212	194266	FinnGen	Cases: ICD8 code 6370[349]|63710|63799|66120, ICD9 code 642[4-7], ICD10 code O11, O14, O15 in medical records or death registry; Controls: female
	UK Biobank Neale Lab	319	226285	UK Biobank (European-ancestry)	Cases: PheCode 642.1; Controls: female
Uterine fibroids	FinnGen Data Freeze 9	31661	179209	FinnGen	Cases: ICD8 code 21899, ICD9 code 218, ICD10 code D25 in medical records or death registry; Controls: female
	UK Biobank Neale Lab	10785	409746	UK Biobank (European-ancestry)	Cases: PheCode D25; Controls: mixed male and female

Table D.2: Comparison of effect sizes of obesity traits on female reproductive conditions in the UK Biobank with and without adjustment for smoking status. Heterogeneity P-value is calculated for the difference between the adjusted-model and unadjusted-model effect estimates. CI = confidence interval, OR = odds ratio, PCOS = polycystic ovary syndrome, S.D. = standard deviation.

Diagnosis	Obesity trait	Adjusted for smoking status		Unadjusted for smoking status		Heterogeneity P-value
		OR (95% CI) per 1 S.D. higher obesity trait	P-value	OR (95% CI) per 1 S.D. higher obesity trait	P-value	
Endometriosis	BMI	1.12 (1.1 - 1.15)	1.46E-23	1.1 (1.08 - 1.13)	1.38E-17	2.42E-01
	WHR	1.07 (1.04 - 1.09)	4.93E-08	1 (0.978 - 1.02)	9.55E-01	8.10E-05
	WHRadjBMI	1.01 (0.991 - 1.04)	2.30E-01	0.951 (0.928 - 0.974)	1.91E-05	1.47E-04
Heavy menstrual bleeding	BMI	1.11 (1.09 - 1.12)	1.51E-60	1.1 (1.09 - 1.11)	1.01E-53	4.28E-01
	WHR	1.08 (1.07 - 1.09)	9.47E-37	0.995 (0.983 - 1.01)	4.37E-01	6.60E-23
	WHRadjBMI	1.03 (1.02 - 1.04)	2.35E-07	0.943 (0.931 - 0.955)	5.59E-22	6.76E-25
Infertility (all causes)	BMI	0.933 (0.899 - 0.966)	4.52E-05	0.89 (0.856 - 0.923)	7.20E-12	7.52E-02
	WHR	0.966 (0.933 - 0.999)	4.23E-02	0.841 (0.807 - 0.874)	2.53E-24	2.48E-07
	WHRadjBMI	0.997 (0.963 - 1.03)	8.43E-01	0.872 (0.839 - 0.905)	8.88E-16	2.73E-07
PCOS	BMI	1.3 (1.26 - 1.35)	2.01E-32	1.24 (1.19 - 1.28)	1.16E-21	3.76E-02
	WHR	1.23 (1.19 - 1.28)	3.07E-21	1.09 (1.05 - 1.14)	7.62E-05	6.39E-06
	WHRadjBMI	1.08 (1.03 - 1.12)	7.10E-04	0.959 (0.915 - 1)	5.70E-02	1.52E-04
Preeclampsia or eclampsia	BMI	1.24 (1.2 - 1.29)	4.68E-25	1.22 (1.17 - 1.26)	2.61E-20	3.31E-01
	WHR	1.13 (1.09 - 1.17)	4.33E-09	1.06 (1.02 - 1.1)	4.15E-03	1.98E-02
	WHRadjBMI	1.02 (0.981 - 1.06)	3.05E-01	0.96 (0.919 - 1)	5.35E-02	3.88E-02
Uterine fibroids	BMI	1.11 (1.1 - 1.13)	2.25E-46	1.12 (1.11 - 1.14)	1.25E-54	3.26E-01
	WHR	1.06 (1.05 - 1.07)	3.65E-15	1.06 (1.04 - 1.07)	1.94E-14	8.66E-01
	WHRadjBMI	1.01 (0.999 - 1.03)	6.28E-02	1.01 (0.992 - 1.02)	3.61E-01	4.97E-01

Table D.3: Power to detect causal effects using two-sample MR for each pair of obesity-related traits and female reproductive outcomes. Power was estimated to detect odds ratio (OR) > 1.5 at a significance level (alpha) = 5%. BMI = body mass index, HC = hip circumference, PCOS = polycystic ovary syndrome, WC = waist circumference, WHR = waist-to-hip ratio, WHRadjBMI = WHR adjusted for BMI.

Exposure	Outcome	Exposure sample size, variance explained by instrument	Outcome case proportion	Power to detect OR > 1.5 at alpha = 5%
BMI	Endometriosis	434785 (6.0%)	7.57%	100%
	Heavy menstrual bleeding		8.95%	100%
	Infertility (all causes)		6.64%	100%
	Infertility (anatomical)		1.20%	100%
	Infertility (anovulatory)		0.79%	100%
	Infertility (idiopathic, exclusion)		3.40%	100%
	Infertility (idiopathic, inclusion)		1.50%	100%
	PCOS		3.25%	100%
	Preeclampsia		1.76%	100%
	Uterine fibroids		6.72%	100%
WHR	Endometriosis	381147 (4.0%)	7.57%	100%
	Heavy menstrual bleeding		8.95%	100%
	Infertility (all causes)		6.64%	100%
	Infertility (anatomical)		1.20%	100%
	Infertility (anovulatory)		0.79%	100%
	Infertility (idiopathic, exclusion)		3.40%	100%
	Infertility (idiopathic, inclusion)		1.50%	100%
	PCOS		3.25%	100%
	Preeclampsia		1.76%	100%
	Uterine fibroids		6.72%	100%
WHRadjBMI	Endometriosis	379486 (3.6%)	7.57%	100%
	Heavy menstrual bleeding		8.95%	100%
	Infertility (all causes)		6.64%	100%
	Infertility (anatomical)		1.20%	100%
	Infertility (anovulatory)		0.79%	100%
	Infertility (idiopathic, exclusion)		3.40%	100%
	Infertility (idiopathic, inclusion)		1.50%	100%
	PCOS		3.25%	100%
	Preeclampsia		1.76%	100%
	Uterine fibroids		6.72%	100%

contd. Power to detect causal effects using two-sample MR for each pair of obesity-related traits and female reproductive outcomes.

Exposure	Outcome	Exposure sample size, variance explained by instrument	Outcome case proportion	Power to detect OR > 1.5 at alpha = 5%
Predicted visceral adipose tissue mass	Endometriosis	161168 (3.7%)	7.57%	100%
	Heavy menstrual bleeding		8.95%	100%
	Infertility (all causes)		6.64%	100%
	Infertility (anatomical)		1.20%	99%
	Infertility (anovulatory)		0.79%	93%
	Infertility (idiopathic, exclusion)		3.40%	100%
	Infertility (idiopathic, inclusion)		1.50%	100%
	PCOS		3.25%	100%
	Preeclampsia		1.76%	100%
	Uterine fibroids		6.72%	100%
Waist circumference	Endometriosis	462166 (3.2%)	7.57%	100%
	Heavy menstrual bleeding		8.95%	100%
	Infertility (all causes)		6.64%	100%
	Infertility (anatomical)		1.20%	100%
	Infertility (anovulatory)		0.79%	100%
	Infertility (idiopathic, exclusion)		3.40%	100%
	Infertility (idiopathic, inclusion)		1.50%	100%
	PCOS		3.25%	100%
	Preeclampsia		1.76%	100%
	Uterine fibroids		6.72%	100%
Hip circumference	Endometriosis	462117 (3.9%)	7.57%	100%
	Heavy menstrual bleeding		8.95%	100%
	Infertility (all causes)		6.64%	100%
	Infertility (anatomical)		1.20%	100%
	Infertility (anovulatory)		0.79%	100%
	Infertility (idiopathic, exclusion)		3.40%	100%
	Infertility (idiopathic, inclusion)		1.50%	100%
	PCOS		3.25%	100%
	Preeclampsia		1.76%	100%
	Uterine fibroids		6.72%	100%

contd. Power to detect causal effects using two-sample MR for each pair of obesity-related traits and female reproductive outcomes.

Exposure	Outcome	Exposure sample size, variance explained by instrument	Outcome case proportion	Power to detect OR > 1.5 at alpha = 5%
WC-specific WHR	Endometriosis	18330 (0.4%)	7.57%	20%
	Heavy menstrual bleeding		8.95%	22%
	Infertility (all causes)		6.64%	18%
	Infertility (anatomical)		1.20%	8%
	Infertility (anovulatory)		0.79%	7%
	Infertility (idiopathic, exclusion)		3.40%	12%
	Infertility (idiopathic, inclusion)		1.50%	8%
	PCOS		3.25%	12%
	Preeclampsia		1.76%	9%
	Uterine fibroids		6.72%	18%
HC-specific WHR	Endometriosis	18330 (0.4%)	7.57%	20%
	Heavy menstrual bleeding		8.95%	22%
	Infertility (all causes)		6.64%	18%
	Infertility (anatomical)		1.20%	8%
	Infertility (anovulatory)		0.79%	7%
	Infertility (idiopathic, exclusion)		3.40%	12%
	Infertility (idiopathic, inclusion)		1.50%	8%
	PCOS		3.25%	12%
	Preeclampsia		1.76%	9%
	Uterine fibroids		6.72%	18%

Table D.4: Comparison of inverse variance weighted (IVW), MR-Egger and weighted median methods for estimating causal effect sizes using Mendelian Randomisation. Pleiotropy P-value is calculated to test for horizontal pleiotropy as the difference of the MR-Egger intercept from 0. CI = confidence interval, OR = odds ratio, PCOS = polycystic ovary syndrome.

		Inverse-variance weighted		MR-Egger		Weighted median		Egger intercept
Diagnosis	**Obesity trait**	**OR (95% CI)**	**P-value**	**OR (95% CI)**	**P-value**	**OR (95% CI)**	**P-value**	**pleiotropy P-value**
Endometriosis	BMI	1.04 (0.946 - 1.13)	4.50E-01	0.747 (0.577 - 0.967)	2.76E-02	0.932 (0.808 - 1.08)	3.36E-01	8.66E-03
	WHR	1.1 (0.99 - 1.23)	7.53E-02	1.29 (0.993 - 1.68)	5.76E-02	1.11 (0.977 - 1.27)	1.06E-01	2.00E-01
	WHRadjBMI	1.11 (1.02 - 1.2)	1.61E-02	1.25 (1.04 - 1.5)	1.69E-02	1.13 (1.01 - 1.26)	2.85E-02	1.43E-01
	Hip circumference	1.01 (0.943 - 1.09)	6.93E-01	0.778 (0.64 - 0.946)	1.20E-02	0.991 (0.882 - 1.11)	8.72E-01	4.11E-03
	Waist circumference	1.07 (0.983 - 1.17)	1.15E-01	0.78 (0.608 - 0.999)	4.92E-02	1.11 (0.967 - 1.27)	1.40E-01	6.83E-03
	Predicted VAT mass	0.771 (0.631 - 0.943)	1.12E-02	0.454 (0.299 - 0.69)	1.43E-03	0.742 (0.573 - 0.962)	2.43E-02	1.27E-02
Heavy menstrual bleeding	BMI	1.25 (1.16 - 1.35)	9.24E-09	0.99 (0.798 - 1.23)	9.30E-01	1.19 (1.07 - 1.32)	9.38E-04	2.56E-02
	WHR	1.21 (1.11 - 1.32)	1.52E-05	1.4 (1.15 - 1.72)	1.20E-03	1.22 (1.1 - 1.36)	2.55E-04	1.08E-01
	WHRadjBMI	1.12 (1.04 - 1.2)	2.14E-03	1.26 (1.08 - 1.47)	3.69E-03	1.19 (1.08 - 1.3)	2.27E-04	8.90E-02
	Hip circumference	1.12 (1.05 - 1.19)	3.04E-04	1.01 (0.858 - 1.19)	9.16E-01	1.16 (1.06 - 1.27)	8.45E-04	1.68E-01
	Waist circumference	1.27 (1.18 - 1.37)	2.45E-10	1.17 (0.951 - 1.44)	1.38E-01	1.24 (1.12 - 1.37)	3.04E-05	3.99E-01
	Predicted VAT mass	1.21 (0.989 - 1.48)	6.37E-02	0.76 (0.487 - 1.19)	2.41E-01	1.2 (0.978 - 1.47)	8.04E-02	3.61E-02
Infertility (all causes)	BMI	0.955 (0.902 - 1.01)	1.19E-01	0.861 (0.731 - 1.01)	7.30E-02	0.912 (0.838 - 0.992)	3.14E-02	1.82E-01
	WHR	1.07 (0.999 - 1.14)	5.30E-02	1.2 (1.02 - 1.4)	2.45E-02	1.15 (1.04 - 1.27)	6.60E-03	1.12E-01
	WHRadjBMI	1.1 (1.05 - 1.16)	1.71E-04	1.13 (1.01 - 1.27)	3.03E-02	1.13 (1.03 - 1.23)	6.61E-03	6.15E-01
	Hip circumference	0.933 (0.889 - 0.979)	4.54E-03	0.795 (0.702 - 0.901)	3.35E-04	0.886 (0.824 - 0.953)	1.12E-03	6.71E-03
	Waist circumference	0.973 (0.916 - 1.03)	3.64E-01	0.794 (0.674 - 0.936)	6.06E-03	0.92 (0.843 - 1)	5.90E-02	9.55E-03
	Predicted VAT mass	0.829 (0.717 - 0.96)	1.23E-02	0.539 (0.403 - 0.722)	5.02E-04	0.789 (0.661 - 0.941)	8.27E-03	4.66E-03
Infertility (anatomical)	BMI	1.06 (0.923 - 1.21)	4.14E-01	0.982 (0.663 - 1.46)	9.29E-01	0.98 (0.801 - 1.2)	8.42E-01	6.90E-01
	WHR	1.09 (0.964 - 1.24)	1.62E-01	1.13 (0.837 - 1.52)	4.30E-01	1.21 (0.967 - 1.51)	9.54E-02	8.27E-01
	WHRadjBMI	1.13 (1.01 - 1.25)	2.90E-02	1.13 (0.899 - 1.43)	2.92E-01	1.19 (0.981 - 1.44)	7.84E-02	9.59E-01
	Hip circumference	0.932 (0.836 - 1.04)	2.00E-01	0.735 (0.554 - 0.973)	3.20E-02	0.919 (0.775 - 1.09)	3.35E-01	7.34E-02
	Waist circumference	1.02 (0.894 - 1.16)	7.93E-01	0.753 (0.527 - 1.08)	1.21E-01	0.963 (0.776 - 1.19)	7.31E-01	7.76E-02
	Predicted VAT mass	0.853 (0.599 - 1.21)	3.77E-01	0.547 (0.238 - 1.26)	1.71E-01	0.842 (0.543 - 1.3)	4.40E-01	2.63E-01

contd. Comparison of various MR methods for estimating causal effect sizes.

		Inverse-variance weighted		MR-Egger		Weighted median		Egger intercept
Infertility (anovulatory)	BMI	1.05 (0.927 - 1.19)	4.44E-01	1.02 (0.714 - 1.46)	9.14E-01	0.93 (0.76 - 1.14)	4.81E-01	8.66E-01
	WHR	1.32 (1.15 - 1.52)	7.95E-05	1.74 (1.25 - 2.42)	1.12E-03	1.49 (1.19 - 1.87)	4.86E-04	7.16E-02
	WHRadjBMI	1.29 (1.16 - 1.45)	4.66E-06	1.42 (1.12 - 1.81)	4.31E-03	1.39 (1.15 - 1.68)	6.44E-04	3.80E-01
	Hip circumference	0.968 (0.868 - 1.08)	5.64E-01	0.861 (0.648 - 1.14)	3.04E-01	0.881 (0.738 - 1.05)	1.60E-01	3.82E-01
	Waist circumference	1.08 (0.945 - 1.23)	2.58E-01	0.867 (0.601 - 1.25)	4.44E-01	0.963 (0.772 - 1.2)	7.38E-01	2.08E-01
	Predicted VAT mass	1.04 (0.77 - 1.39)	8.18E-01	0.808 (0.394 - 1.66)	5.66E-01	0.97 (0.638 - 1.48)	8.87E-01	4.65E-01
Infertility (idiopathic, exclusion)	BMI	0.928 (0.869 - 0.99)	2.35E-02	0.85 (0.706 - 1.02)	8.87E-02	0.94 (0.85 - 1.04)	2.23E-01	3.29E-01
	WHR	0.968 (0.901 - 1.04)	3.65E-01	1.1 (0.931 - 1.3)	2.61E-01	0.985 (0.877 - 1.11)	8.02E-01	9.76E-02
	WHRadjBMI	1.03 (0.971 - 1.09)	3.21E-01	1.07 (0.94 - 1.21)	3.13E-01	1.05 (0.956 - 1.16)	2.89E-01	5.34E-01
	Hip circumference	0.924 (0.877 - 0.974)	2.98E-03	0.771 (0.673 - 0.883)	1.89E-04	0.887 (0.817 - 0.963)	4.18E-03	4.76E-03
	Waist circumference	0.931 (0.87 - 0.996)	3.87E-02	0.776 (0.644 - 0.936)	8.22E-03	0.861 (0.779 - 0.952)	3.34E-03	4.18E-02
	Predicted VAT mass	0.808 (0.684 - 0.956)	1.27E-02	0.491 (0.344 - 0.7)	8.30E-04	0.781 (0.63 - 0.969)	2.44E-02	6.54E-03
Infertility (idiopathic, inclusion)	BMI	0.866 (0.781 - 0.962)	7.05E-03	0.553 (0.412 - 0.741)	9.51E-05	0.82 (0.684 - 0.982)	3.12E-02	1.51E-03
	WHR	1.08 (0.956 - 1.22)	2.12E-01	1.15 (0.854 - 1.54)	3.67E-01	1.11 (0.931 - 1.33)	2.44E-01	6.77E-01
	WHRadjBMI	1.13 (1.03 - 1.24)	1.05E-02	1.01 (0.822 - 1.24)	9.43E-01	1.11 (0.956 - 1.28)	1.75E-01	2.12E-01
	Hip circumference	0.846 (0.777 - 0.922)	1.25E-04	0.697 (0.559 - 0.869)	1.43E-03	0.857 (0.743 - 0.989)	3.47E-02	6.24E-02
	Waist circumference	0.907 (0.817 - 1.01)	6.64E-02	0.61 (0.458 - 0.812)	7.36E-04	0.866 (0.718 - 1.04)	1.31E-01	3.60E-03
	Predicted VAT mass	0.765 (0.592 - 0.989)	4.07E-02	0.522 (0.288 - 0.948)	4.52E-02	0.78 (0.538 - 1.13)	1.88E-01	1.82E-01
PCOS	BMI	1.92 (1.59 - 2.32)	1.77E-11	3.42 (2.17 - 5.41)	2.74E-07	2.7 (2.01 - 3.63)	4.62E-11	6.95E-03
	WHR	1.54 (1.25 - 1.89)	4.67E-05	1.61 (0.994 - 2.6)	5.43E-02	1.63 (1.14 - 2.33)	7.97E-03	8.42E-01
	WHRadjBMI	1.26 (1.07 - 1.48)	5.47E-03	1.3 (0.928 - 1.82)	1.29E-01	1.47 (1.11 - 1.95)	7.07E-03	8.38E-01
	Hip circumference	1.51 (1.29 - 1.77)	2.54E-07	1.63 (1.14 - 2.35)	8.28E-03	1.61 (1.24 - 2.09)	3.57E-04	6.37E-01
	Waist circumference	2.05 (1.69 - 2.48)	1.57E-13	3.22 (2.01 - 5.14)	1.33E-06	2.54 (1.86 - 3.47)	4.33E-09	3.92E-02
	Predicted VAT mass	2.51 (1.76 - 3.57)	3.34E-07	3.24 (1.42 - 7.38)	1.10E-02	2.89 (1.9 - 4.39)	7.30E-07	5.04E-01

contd. Comparison of various MR methods for estimating causal effect sizes.

		Inverse-variance weighted		MR-Egger		Weighted median		Egger intercept
Preeclampsia or eclampsia	BMI	1.67 (1.48 - 1.89)	1.56E-16	2.03 (1.43 - 2.87)	8.92E-05	1.62 (1.35 - 1.94)	1.57E-07	2.45E-01
	WHR	1.39 (1.21 - 1.6)	3.38E-06	1.33 (0.957 - 1.86)	9.04E-02	1.14 (0.941 - 1.38)	1.81E-01	7.90E-01
	WHRadjBMI	1.26 (1.13 - 1.41)	3.71E-05	1.15 (0.904 - 1.47)	2.51E-01	1.07 (0.907 - 1.25)	4.38E-01	4.05E-01
	Hip circumference	1.3 (1.16 - 1.44)	3.15E-06	1.45 (1.09 - 1.93)	1.02E-02	1.31 (1.12 - 1.53)	8.47E-04	3.92E-01
	Waist circumference	1.5 (1.32 - 1.7)	2.18E-10	2.13 (1.51 - 3.01)	2.26E-05	1.5 (1.24 - 1.8)	1.96E-05	3.42E-02
	Predicted VAT mass	1.87 (1.32 - 2.66)	4.65E-04	3.39 (1.48 - 7.79)	9.33E-03	2.17 (1.39 - 3.38)	6.43E-04	1.41E-01
Uterine fibroids	BMI	1.08 (0.999 - 1.17)	5.23E-02	0.971 (0.771 - 1.22)	8.00E-01	1.07 (0.963 - 1.18)	2.15E-01	3.21E-01
	WHR	1.19 (1.1 - 1.3)	2.91E-05	1.44 (1.19 - 1.76)	2.90E-04	1.35 (1.21 - 1.5)	3.59E-08	3.56E-02
	WHRadjBMI	1.16 (1.08 - 1.24)	2.28E-05	1.32 (1.14 - 1.52)	2.84E-04	1.23 (1.13 - 1.34)	4.55E-06	5.30E-02
	Hip circumference	1.01 (0.952 - 1.07)	7.09E-01	0.955 (0.815 - 1.12)	5.68E-01	1 (0.919 - 1.09)	9.80E-01	4.40E-01
	Waist circumference	1.15 (1.07 - 1.24)	1.54E-04	1.18 (0.963 - 1.45)	1.11E-01	1.16 (1.06 - 1.28)	2.34E-03	7.99E-01
	Predicted VAT mass	1.07 (0.924 - 1.23)	3.83E-01	0.961 (0.677 - 1.36)	8.25E-01	1 (0.835 - 1.2)	9.79E-01	5.33E-01

Bibliography

[1] Venkatesh, S. S. *et al.* The genetic architecture of changes in adiposity during adulthood. *medRxiv* 2023.01.09.23284364 (2023). URL `https://www.medrxiv.org/content/medrxiv/early/2023/01/11/2023.01.09.23284364.full.pdf`.

[2] Venkatesh, S. S. *et al.* Obesity and risk of female reproductive conditions: A Mendelian randomisation study. *PLoS Med* **19**, e1003679 (2022). URL `https://www.ncbi.nlm.nih.gov/pubmed/35104295`.

[3] Mayr, E. *The growth of biological thought: Diversity, evolution, and inheritance* (Harvard University Press, 1982).

[4] Grafton, A. & Siraisi, N. G. *Natural Particulars: Nature and the Disciplines in Renaissance Europe* (MIT Press, 1999).

[5] de Lamarck, J.-B. d. M. *Philosophie zoologique ou exposition des considérations relatives à l'histoire naturelle des animaux*, vol. 1 (Savy, 1873).

[6] Stubbe, H. *History of genetics: from prehistoric times to the rediscovery of Mendel's laws* (The MIT Press, 1972).

[7] Ramachandran, S. *et al.* Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proc Natl Acad Sci U S A* **102**, 15942–7 (2005). URL `https://www.ncbi.nlm.nih.gov/pubmed/16243969`.

[8] Mendel, G. *Versuche über Pflanzen-Hybriden* (Im Verlage des Vereines, Brünn, 1866). URL `https://www.biodiversitylibrary.org/item/124139`.

[9] Ellis, T. H. N., Hofer, J. M. I., Swain, M. T. & van Dijk, P. J. Mendel's pea crosses: varieties, traits and statistics. *Hereditas* **156**, 33 (2019). URL `https://www.ncbi.nlm.nih.gov/pubmed/31695583`.

[10] Wright, S. Physiological and evolutionary theories of dominance. *The American Naturalist* **68**, 24–53 (1934).

[11] Kacser, H. & Burns, J. A. The molecular basis of dominance. *Genetics* **97**, 639–66 (1981). URL `https://www.ncbi.nlm.nih.gov/pubmed/7297851`.

[12] Morgan, T. H. Sex limited inheritance in Drosophila. *Science* **32**, 120–2 (1910). URL `https://www.ncbi.nlm.nih.gov/pubmed/17759620`.

[13] Sutton, W. S. On the morphology of the chromoso group in Brachystola magna. *The Biological Bulletin* **4**, 24–39 (1902).

[14] Schacherer, J. Beyond the simplicity of Mendelian inheritance. *C R Biol* **339**, 284–8 (2016). URL `https://www.ncbi.nlm.nih.gov/pubmed/27344551`.

[15] Badano, J. L. & Katsanis, N. Beyond Mendel: an evolving view of human genetic disease transmission. *Nat Rev Genet* **3**, 779–89 (2002). URL https://www.ncbi.nlm.nih.gov/pubmed/12360236.

[16] Hedges, S. B. & Kumar, S. *The timetree of life* (OUP Oxford, 2009).

[17] Galton, F. Regression towards mediocrity in hereditary stature. *The Journal of the Anthropological Institute of Great Britain and Ireland* **15**, 246–263 (1886).

[18] Pearson, K. Mathematical contributions to the theory of evolution. VIII. On the correlation of characters not quantitatively measurable. *Proceedings of the Royal Society of London* **66**, 241–244 (1900).

[19] Fisher, R. A. XV.—The correlation between relatives on the supposition of Mendelian inheritance. *Earth and Environmental Science Transactions of the Royal Society of Edinburgh* **52**, 399–433 (1919).

[20] Fisher, R. A. On the mathematical foundations of theoretical statistics. *Philosophical transactions of the Royal Society of London. Series A, containing papers of a mathematical or physical character* **222**, 309–368 (1922).

[21] Smith, C. A. Some simple methods for linkage analysis. *Ann Hum Genet* **53**, 277–83 (1989). URL https://www.ncbi.nlm.nih.gov/pubmed/2596834.

[22] Taylor, E. W., Xu, J., Jabs, E. W. & Meyers, D. A. Linkage analysis of genetic disorders. *Methods Mol Biol* **68**, 11–25 (1997). URL https://www.ncbi.nlm.nih.gov/pubmed/9055247.

[23] Liao, W. W. *et al.* A draft human pangenome reference. *Nature* **617**, 312–324 (2023). URL https://www.ncbi.nlm.nih.gov/pubmed/37165242.

[24] Shendure, J., Mitra, R. D., Varma, C. & Church, G. M. Advanced sequencing technologies: methods and goals. *Nat Rev Genet* **5**, 335–44 (2004). URL https://www.ncbi.nlm.nih.gov/pubmed/15143316.

[25] Ng, P. C. & Kirkness, E. F. Whole genome sequencing. *Methods Mol Biol* **628**, 215–26 (2010). URL https://www.ncbi.nlm.nih.gov/pubmed/20238084.

[26] Hartl, D. L., Clark, A. G. & Clark, A. G. *Principles of population genetics*, vol. 116 (Sinauer associates Sunderland, MA, 1997).

[27] Hill, W. G. & Robertson, A. The effects of inbreeding at loci with heterozygote advantage. *Genetics* **60**, 615–28 (1968). URL https://www.ncbi.nlm.nih.gov/pubmed/5728744.

[28] Bodmer, W. F. Human genetics: the molecular challenge. *Cold Spring Harb Symp Quant Biol* **51 Pt 1**, 1–13 (1986). URL https://www.ncbi.nlm.nih.gov/pubmed/3472706.

[29] Botstein, D. & Risch, N. Discovering genotypes underlying human phenotypes: past successes for Mendelian disease, future approaches for complex disease. *Nat Genet* **33 Suppl**, 228–37 (2003). URL https://www.ncbi.nlm.nih.gov/pubmed/12610532.

[30] Lander, E. S. & Schork, N. J. Genetic dissection of complex traits. *Science* **265**, 2037–48 (1994). URL https://www.ncbi.nlm.nih.gov/pubmed/8091226.

[31] Lander, E. S. & Botstein, D. Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* **121**, 185–99 (1989). URL https://www.ncbi.nlm.nih.gov/pubmed/2563713.

[32] Terwilliger, J. D. & Weiss, K. M. Linkage disequilibrium mapping of complex disease: fantasy or reality? *Curr Opin Biotechnol* **9**, 578–94 (1998). URL https://www.ncbi.nlm.nih.gov/pubmed/9889136.

[33] Border, R. *et al.* No support for historical candidate gene or candidate gene-by-interaction hypotheses for major depression across multiple large samples. *Am J Psychiatry* **176**, 376–387 (2019). URL https://www.ncbi.nlm.nih.gov/pubmed/30845820.

[34] Rankinen, T. *et al.* The human obesity gene map: the 2005 update. *Obesity (Silver Spring)* **14**, 529–644 (2006). URL https://www.ncbi.nlm.nih.gov/pubmed/16741264.

[35] Hirschhorn, J. N., Lohmueller, K., Byrne, E. & Hirschhorn, K. A comprehensive review of genetic association studies. *Genet Med* **4**, 45–61 (2002). URL https://www.ncbi.nlm.nih.gov/pubmed/11882781.

[36] Visscher, P. M., Brown, M. A., McCarthy, M. I. & Yang, J. Five years of GWAS discovery. *Am J Hum Genet* **90**, 7–24 (2012). URL https://www.ncbi.nlm.nih.gov/pubmed/22243964.

[37] Goldstein, D. B. Common genetic variation and human traits. *N Engl J Med* **360**, 1696–8 (2009). URL https://www.ncbi.nlm.nih.gov/pubmed/19369660.

[38] Risch, N. *et al.* A genomic screen of autism: evidence for a multilocus etiology. *Am J Hum Genet* **65**, 493–507 (1999). URL https://www.ncbi.nlm.nih.gov/pubmed/10417292.

[39] Yengo, L. *et al.* A saturated map of common genetic variants associated with human height. *Nature* **610**, 704–712 (2022). URL https://www.ncbi.nlm.nih.gov/pubmed/36224396.

[40] Yengo, L. *et al.* Meta-analysis of genome-wide association studies for height and body mass index in approximately 700000 individuals of European ancestry. *Hum Mol Genet* **27**, 3641–3649 (2018). URL https://www.ncbi.nlm.nih.gov/pubmed/30124842.

[41] Mahajan, A. *et al.* Multi-ancestry genetic study of type 2 diabetes highlights the power of diverse populations for discovery and translation. *Nat Genet* **54**, 560–572 (2022). URL https://www.ncbi.nlm.nih.gov/pubmed/35551307.

[42] Tcheandjieu, C. *et al.* Large-scale genome-wide association study of coronary artery disease in genetically diverse populations. *Nat Med* **28**, 1679–1692 (2022). URL https://www.ncbi.nlm.nih.gov/pubmed/35915156.

[43] Hill, W. G., Goddard, M. E. & Visscher, P. M. Data and theory point to mainly additive genetic variance for complex traits. *PLoS Genet* **4**, e1000008 (2008). URL https://www.ncbi.nlm.nih.gov/pubmed/18454194.

[44] Risch, N. & Merikangas, K. The future of genetic studies of complex human diseases. *Science* **273**, 1516–7 (1996). URL https://www.ncbi.nlm.nih.gov/pubmed/8801636.

[45] Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001). URL https://www.ncbi.nlm.nih.gov/pubmed/11237011.

[46] International HapMap, C. *et al.* Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52–8 (2010). URL https://www.ncbi.nlm.nih.gov/pubmed/20811451.

[47] Jenkins, S. & Gibson, N. High-throughput SNP genotyping. *Comp Funct Genomics* **3**, 57–66 (2002). URL https://www.ncbi.nlm.nih.gov/pubmed/18628885.

[48] Marchini, J. & Howie, B. Comparing algorithms for genotype imputation. *Am J Hum Genet* **83**, 535–9; author reply 539–40 (2008). URL https://www.ncbi.nlm.nih.gov/pubmed/18940314.

[49] Wellcome Trust Case Control, C. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–78 (2007). URL https://www.ncbi.nlm.nih.gov/pubmed/17554300.

[50] Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res* **42**, D1001–6 (2014). URL https://www.ncbi.nlm.nih.gov/pubmed/24316577.

[51] MacArthur, J. A. L. *et al.* Workshop proceedings: GWAS summary statistics standards and sharing. *Cell Genom* **1** (2021). URL https://www.ncbi.nlm.nih.gov/pubmed/36082306.

[52] Thelwall, M. *et al.* Is useful research data usually shared? an investigation of genome-wide association study summary statistics. *PLoS One* **15**, e0229578 (2020). URL https://www.ncbi.nlm.nih.gov/pubmed/32084240.

[53] Mavaddat, N. *et al.* Polygenic risk scores for prediction of breast cancer and breast cancer subtypes. *Am J Hum Genet* **104**, 21–34 (2019). URL https://www.ncbi.nlm.nih.gov/pubmed/30554720.

[54] Kurniansyah, N. *et al.* A multi-ethnic polygenic risk score is associated with hypertension prevalence and progression throughout adulthood. *Nat Commun* **13**, 3549 (2022). URL https://www.ncbi.nlm.nih.gov/pubmed/35729114.

[55] Lim, A. J. W. *et al.* Robust SNP-based prediction of rheumatoid arthritis through machine-learning-optimized polygenic risk score. *J Transl Med* **21**, 92 (2023). URL https://www.ncbi.nlm.nih.gov/pubmed/36750873.

[56] Hingorani, A. D. *et al.* Performance of polygenic risk scores in screening, prediction, and risk stratification: secondary analysis of data in the Polygenic Score Catalog. *BMJ Med* **2**, e000554 (2023). URL https://www.ncbi.nlm.nih.gov/pubmed/37859783.

[57] Lambert, S. A., Abraham, G. & Inouye, M. Towards clinical utility of polygenic risk scores. *Hum Mol Genet* **28**, R133–R142 (2019). URL https://www.ncbi.nlm.nih.gov/pubmed/31363735.

[58] Nelson, M. R. *et al.* The support of human genetic evidence for approved drug indications. *Nat Genet* **47**, 856–60 (2015). URL https://www.ncbi.nlm.nih.gov/pubmed/26121088.

[59] Willer, C. J. *et al.* Newly identified loci that influence lipid concentrations and risk of coronary artery disease. *Nat Genet* **40**, 161–9 (2008). URL https://www.ncbi.nlm.nih.gov/pubmed/18193043.

[60] Trajanoska, K. *et al.* From target discovery to clinical drug development with human genetics. *Nature* **620**, 737–745 (2023). URL https://www.ncbi.nlm.nih.gov/pubmed/37612393.

[61] King, E. A., Davis, J. W. & Degner, J. F. Are drug targets with genetic support twice as likely to be approved? Revised estimates of the impact of genetic support for drug mechanisms on the probability of drug approval. *PLoS Genet* **15**, e1008489 (2019). URL https://www.ncbi.nlm.nih.gov/pubmed/31830040.

[62] Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet* **47**, 291–5 (2015). URL https://www.ncbi.nlm.nih.gov/pubmed/25642630.

[63] Smith, G. D. & Ebrahim, S. 'Mendelian randomization': Can genetic epidemiology contribute to understanding environmental determinants of disease? *Int J Epidemiol* **32**, 1–22 (2003). URL https://www.ncbi.nlm.nih.gov/pubmed/12689998.

[64] Uffelmann, E. *et al.* Genome-wide association studies. *Nature Reviews Methods Primers* **1**, 59 (2021). URL https://doi.org/10.1038/s43586-021-00056-9.

[65] Mbatchou, J. *et al.* Computationally efficient whole-genome regression for quantitative and binary traits. *Nat Genet* **53**, 1097–1103 (2021). URL https://www.ncbi.nlm.nih.gov/pubmed/34017140.

[66] Loh, P. R., Kichaev, G., Gazal, S., Schoech, A. P. & Price, A. L. Mixed-model association for biobank-scale datasets. *Nat Genet* **50**, 906–908 (2018). URL https://www.ncbi.nlm.nih.gov/pubmed/29892013.

[67] 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). URL https://www.ncbi.nlm.nih.gov/pubmed/30104761.

[68] Evangelou, E. & Ioannidis, J. P. Meta-analysis methods for genome-wide association studies and beyond. *Nat Rev Genet* **14**, 379–89 (2013). URL https://www.ncbi.nlm.nih.gov/pubmed/23657481.

[69] The Noun Project. https://thenounproject.com/. Accessed: 2020-11-03.

[70] Stephens, Z. D. *et al.* Big data: Astronomical or genomical? *PLoS Biol* **13**, e1002195 (2015). URL https://www.ncbi.nlm.nih.gov/pubmed/26151137.

[71] Langmead, B. & Nellore, A. Cloud computing for genomic data analysis and collaboration. *Nat Rev Genet* **19**, 208–219 (2018). URL https://www.ncbi.nlm.nih.gov/pubmed/29379135.

[72] Sul, J. H., Martin, L. S. & Eskin, E. Population structure in genetic studies: Confounding factors and mixed models. *PLoS Genet* **14**, e1007309 (2018). URL https://www.ncbi.nlm.nih.gov/pubmed/30589851.

[73] Allen, R. G. D. The assumptions of linear regression. *Economica* **6**, 191–201 (1939). URL http://www.jstor.org/stable/2548931.

[74] Young, A. I. *et al.* Relatedness disequilibrium regression estimates heritability without environmental bias. *Nat Genet* **50**, 1304–1310 (2018). URL https://www.ncbi.nlm.nih.gov/pubmed/30104764.

[75] Kuonen, D. Saddlepoint approximations for distributions of quadratic forms in normal variables. *Biometrika* **86**, 929–935 (1999).

[76] Pe'er, I., Yelensky, R., Altshuler, D. & Daly, M. J. Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. *Genet Epidemiol* **32**, 381–5 (2008). URL https://www.ncbi.nlm.nih.gov/pubmed/18348202.

[77] Palmer, C. & Pe'er, I. Statistical correction of the winner's curse explains replication variability in quantitative trait genome-wide association studies. *PLoS Genet* **13**, e1006916 (2017). URL https://www.ncbi.nlm.nih.gov/pubmed/28715421.

[78] Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–1 (2010). URL https://www.ncbi.nlm.nih.gov/pubmed/20616382.

[79] Winkler, T. W. *et al.* Quality control and conduct of genome-wide association meta-analyses. *Nat Protoc* **9**, 1192–212 (2014). URL https://www.ncbi.nlm.nih.gov/pubmed/24762786.

[80] Yang, J., Lee, S. H., Wray, N. R., Goddard, M. E. & Visscher, P. M. GCTA-GREML accounts for linkage disequilibrium when estimating genetic variance from genome-wide SNPs. *Proc Natl Acad Sci U S A* **113**, E4579–80 (2016). URL https://www.ncbi.nlm.nih.gov/pubmed/27457963.

[81] Polderman, T. J. *et al.* Meta-analysis of the heritability of human traits based on fifty years of twin studies. *Nat Genet* **47**, 702–9 (2015). URL https://www.ncbi.nlm.nih.gov/pubmed/25985137.

[82] Young, A. I. Solving the missing heritability problem. *PLoS Genet* **15**, e1008222 (2019). URL https://www.ncbi.nlm.nih.gov/pubmed/31233496.

[83] Matthews, L. J. & Turkheimer, E. Three legs of the missing heritability problem. *Stud Hist Philos Sci* **93**, 183–191 (2022). URL https://www.ncbi.nlm.nih.gov/pubmed/35533541.

[84] Consortium, E. P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012). URL https://www.ncbi.nlm.nih.gov/pubmed/22955616.

[85] Maurano, M. T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190–5 (2012). URL https://www.ncbi.nlm.nih.gov/pubmed/22955828.

[86] Spain, S. L. & Barrett, J. C. Strategies for fine-mapping complex traits. *Hum Mol Genet* **24**, R111–9 (2015). URL https://www.ncbi.nlm.nih.gov/pubmed/26157023.

[87] Benner, C. *et al.* FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* **32**, 1493–501 (2016). URL https://www.ncbi.nlm.nih.gov/pubmed/26773131.

[88] Miao, Z., Humphreys, B. D., McMahon, A. P. & Kim, J. Multi-omics integration in the age of million single-cell data. *Nat Rev Nephrol* **17**, 710–724 (2021). URL https://www.ncbi.nlm.nih.gov/pubmed/34417589.

[89] Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506–11 (2013). URL https://www.ncbi.nlm.nih.gov/pubmed/24037378.

[90] Li, L. *et al.* An atlas of alternative polyadenylation quantitative trait loci contributing to complex trait and disease heritability. *Nat Genet* **53**, 994–1005 (2021). URL https://www.ncbi.nlm.nih.gov/pubmed/33986536.

[91] Garrido-Martin, D., Borsari, B., Calvo, M., Reverter, F. & Guigo, R. Identification and analysis of splicing quantitative trait loci across multiple tissues in the human genome. *Nat Commun* **12**, 727 (2021). URL https://www.ncbi.nlm.nih.gov/pubmed/33526779.

[92] Ye, Y., Zhang, Z., Liu, Y., Diao, L. & Han, L. A multi-omics perspective of quantitative trait loci in precision medicine. *Trends Genet* **36**, 318–336 (2020). URL https://www.ncbi.nlm.nih.gov/pubmed/32294413.

[93] Mountjoy, E. *et al.* An open approach to systematically prioritize causal variants and genes at all published human GWAS trait-associated loci. *Nat Genet* **53**, 1527–1533 (2021). URL https://www.ncbi.nlm.nih.gov/pubmed/34711957.

[94] Mostafavi, H., Spence, J. P., Naqvi, S. & Pritchard, J. K. Systematic differences in discovery of genetic effects on gene expression and complex traits. *Nat Genet* **55**, 1866–1875 (2023). URL https://www.ncbi.nlm.nih.gov/pubmed/37857933.

[95] Backman, J. D. *et al.* Exome sequencing and analysis of 454,787 UK Biobank participants. *Nature* **599**, 628–634 (2021). URL https://www.ncbi.nlm.nih.gov/pubmed/34662886.

[96] Cano-Gamez, E. & Trynka, G. From GWAS to Function: Using functional genomics to identify the mechanisms underlying complex diseases. *Front Genet* **11**, 424 (2020). URL https://www.ncbi.nlm.nih.gov/pubmed/32477401.

[97] Consortium, G. T. The Genotype-Tissue Expression (GTEx) project. *Nat Genet* **45**, 580–5 (2013). URL https://www.ncbi.nlm.nih.gov/pubmed/23715323.

[98] Wang, J. *et al.* Imputing gene expression in uncollected tissues within and beyond GTEx. *Am J Hum Genet* **98**, 697–708 (2016). URL https://www.ncbi.nlm.nih.gov/pubmed/27040689.

[99] Pendergrass, S. A. & Ritchie, M. D. Phenome-wide association studies: Leveraging comprehensive phenotypic and genotypic data for discovery. *Curr Genet Med Rep* **3**, 92–100 (2015). URL https://www.ncbi.nlm.nih.gov/pubmed/26146598.

[100] O'Connor, L. J. & Price, A. L. Distinguishing genetic correlation from causation across 52 diseases and complex traits. *Nat Genet* **50**, 1728–1734 (2018). URL https://www.ncbi.nlm.nih.gov/pubmed/30374074.

[101] Micaglio, E. *et al.* Role of pharmacogenetics in adverse drug reactions: An update towards personalized medicine. *Front Pharmacol* **12**, 651720 (2021). URL https://www.ncbi.nlm.nih.gov/pubmed/33995067.

[102] Wang, K., Gaitsch, H., Poon, H., Cox, N. J. & Rzhetsky, A. Classification of common human diseases derived from shared genetic and environmental determinants. *Nat Genet* **49**, 1319–1325 (2017). URL https://www.ncbi.nlm.nih.gov/pubmed/28783162.

[103] Turley, P. *et al.* Multi-trait analysis of genome-wide association summary statistics using mtag. *Nat Genet* **50**, 229–237 (2018). URL https://www.ncbi.nlm.nih.gov/pubmed/29292387.

[104] Cortes, A., Albers, P. K., Dendrou, C. A., Fugger, L. & McVean, G. Identifying cross-disease components of genetic risk across hospital data in the UK Biobank. *Nat Genet* **52**, 126–134 (2020). URL https://www.ncbi.nlm.nih.gov/pubmed/31873298.

[105] Jiang, X. *et al.* Age-dependent topic modeling of comorbidities in UK Biobank identifies disease subtypes with differential genetic risk. *Nat Genet* **55**, 1854–1865 (2023). URL https://www.ncbi.nlm.nih.gov/pubmed/37814053.

[106] Dahl, A. & Zaitlen, N. Genetic influences on disease subtypes. *Annu Rev Genomics Hum Genet* **21**, 413–435 (2020). URL https://www.ncbi.nlm.nih.gov/pubmed/32873077.

[107] Wolford, B. N., Willer, C. J. & Surakka, I. Electronic health records: the next wave of complex disease genetics. *Hum Mol Genet* **27**, R14–R21 (2018). URL https://www.ncbi.nlm.nih.gov/pubmed/29547983.

[108] Abdellaoui, A., Yengo, L., Verweij, K. J. H. & Visscher, P. M. 15 years of GWAS discovery: Realizing the promise. *Am J Hum Genet* **110**, 179–194 (2023). URL https://www.ncbi.nlm.nih.gov/pubmed/36634672.

[109] Olson, J. E. *et al.* Biobanks and personalized medicine. *Clin Genet* **86**, 50–5 (2014). URL https://www.ncbi.nlm.nih.gov/pubmed/24588254.

[110] Halldorsson, B. V. *et al.* The sequences of 150,119 genomes in the UK Biobank. *Nature* **607**, 732–740 (2022). URL https://www.ncbi.nlm.nih.gov/pubmed/35859178.

[111] Gudbjartsson, D. F. *et al.* Large-scale whole-genome sequencing of the Icelandic population. *Nat Genet* **47**, 435–44 (2015). URL https://www.ncbi.nlm.nih.gov/pubmed/25807286.

[112] Venner, E. *et al.* Whole-genome sequencing as an investigational device for return of hereditary disease risk and pharmacogenomic results as part of the All of Us Research Program. *Genome Med* **14**, 34 (2022). URL https://www.ncbi.nlm.nih.gov/pubmed/35346344.

[113] Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018). URL https://www.ncbi.nlm.nih.gov/pubmed/30305743.

[114] 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**, e1001779 (2015). URL https://www.ncbi.nlm.nih.gov/pubmed/25826379.

[115] Nagai, A. *et al.* Overview of the BioBank Japan Project: Study design and profile. *J Epidemiol* **27**, S2–S8 (2017). URL https://www.ncbi.nlm.nih.gov/pubmed/28189464.

[116] Chen, Z. *et al.* China Kadoorie Biobank of 0.5 million people: survey methods, baseline characteristics and long-term follow-up. *Int J Epidemiol* **40**, 1652–66 (2011). URL https://www.ncbi.nlm.nih.gov/pubmed/22158673.

[117] Cho, S. Y. *et al.* Opening of the national biobank of Korea as the infrastructure of future biomedical science in Korea. *Osong Public Health Res Perspect* **3**, 177–84 (2012). URL https://www.ncbi.nlm.nih.gov/pubmed/24159511.

[118] Fan, C. T., Lin, J. C. & Lee, C. H. Taiwan Biobank: a project aiming to aid Taiwan's transition into a biomedical island. *Pharmacogenomics* **9**, 235–46 (2008). URL https://www.ncbi.nlm.nih.gov/pubmed/18370851.

[119] Al Thani, A. *et al.* Qatar Biobank Cohort Study: Study design and first results. *Am J Epidemiol* **188**, 1420–1433 (2019). URL https://www.ncbi.nlm.nih.gov/pubmed/30927351.

[120] Nsubuga, G. *et al.* Biobanking in East and Central Africa: A case of the Integrated Biorepository of H3Africa Uganda. *Open Res Afr* **5**, 33 (2022). URL https://www.ncbi.nlm.nih.gov/pubmed/37224314.

[121] Zawistowski, M. *et al.* The Michigan Genomics Initiative: A biobank linking genotypes and electronic clinical records in Michigan Medicine patients. *Cell Genom* **3**, 100257 (2023). URL https://www.ncbi.nlm.nih.gov/pubmed/36819667.

[122] Boutin, N. T. *et al.* The evolution of a large biobank at Mass General Brigham. *J Pers Med* **12** (2022). URL https://www.ncbi.nlm.nih.gov/pubmed/36013271.

[123] Wiley, L. K. *et al.* Building a vertically-integrated genomic learning health system: The Colorado Center for Personalized Medicine Biobank. *medRxiv* 2022.06.09.22276222 (2022). URL https://www.medrxiv.org/content/medrxiv/early/2022/10/24/2022.06.09.22276222.full.pdf.

[124] Johnson, R. *et al.* Leveraging genomic diversity for discovery in an electronic health record linked biobank: the UCLA ATLAS Community Health Initiative. *Genome Med* **14**, 104 (2022). URL https://www.ncbi.nlm.nih.gov/pubmed/36085083.

[125] Verma, A. *et al.* The Penn Medicine BioBank: Towards a genomics-enabled learning healthcare system to accelerate precision medicine in a diverse population. *J Pers Med* **12** (2022). URL https://www.ncbi.nlm.nih.gov/pubmed/36556195.

[126] Kirsh, V. A. *et al.* Cohort Profile: The Ontario Health Study (OHS). *Int J Epidemiol* **52**, e137–e151 (2023). URL https://www.ncbi.nlm.nih.gov/pubmed/35962976.

[127] Hakonarson, H., Gulcher, J. R. & Stefansson, K. deCODE genetics, Inc. *Pharmacogenomics* **4**, 209–15 (2003). URL https://www.ncbi.nlm.nih.gov/pubmed/12605555.

[128] Leitsalu, L. *et al.* Cohort Profile: Estonian Biobank of the Estonian Genome Center, University of Tartu. *Int J Epidemiol* **44**, 1137–47 (2015). URL https://www.ncbi.nlm.nih.gov/pubmed/24518929.

[129] Kurki, M. I. *et al.* FinnGen provides genetic insights from a well-phenotyped isolated population. *Nature* **613**, 508–518 (2023). URL https://www.ncbi.nlm.nih.gov/pubmed/36653562.

[130] Asvold, B. O. *et al.* Cohort Profile Update: The HUNT Study, Norway. *Int J Epidemiol* **52**, e80–e91 (2023). URL https://www.ncbi.nlm.nih.gov/pubmed/35578897.

[131] Sijtsma, A. *et al.* Cohort Profile Update: Lifelines, a three-generation cohort study and biobank. *Int J Epidemiol* **51**, e295–e302 (2022). URL https://www.ncbi.nlm.nih.gov/pubmed/34897450.

[132] Gaziano, J. M. *et al.* Million Veteran Program: A mega-biobank to study genetic influences on health and disease. *J Clin Epidemiol* **70**, 214–23 (2016). URL https://www.ncbi.nlm.nih.gov/pubmed/26441289.

[133] Beesley, L. J. *et al.* The emerging landscape of health research based on biobanks linked to electronic health records: Existing resources, statistical challenges, and potential opportunities. *Stat Med* **39**, 773–800 (2020). URL https://www.ncbi.nlm.nih.gov/pubmed/31859414.

[134] Wei, W. Q. & Denny, J. C. Extracting research-quality phenotypes from electronic health records to support precision medicine. *Genome Med* **7**, 41 (2015). URL https://www.ncbi.nlm.nih.gov/pubmed/25937834.

[135] Tasa, T. *et al.* Genetic variation in the Estonian population: pharmacogenomics study of adverse drug effects using electronic health records. *Eur J Hum Genet* **27**, 442–454 (2019). URL https://www.ncbi.nlm.nih.gov/pubmed/30420678.

[136] Wang, Y. *et al.* Unsupervised machine learning for the discovery of latent disease clusters and patient subgroups using electronic health records. *J Biomed Inform* **102**, 103364 (2020). URL https://www.ncbi.nlm.nih.gov/pubmed/31891765.

[137] Landi, I. *et al.* Deep representation learning of electronic health records to unlock patient stratification at scale. *NPJ Digit Med* **3**, 96 (2020). URL https://www.ncbi.nlm.nih.gov/pubmed/32699826.

[138] Sarwar, T. *et al.* The secondary use of electronic health records for data mining: Data characteristics and challenges. *ACM Comput. Surv.* **55**, Article 33 (2022). URL https://doi.org/10.1145/3490234.

[139] Garcelon, N., Burgun, A., Salomon, R. & Neuraz, A. Electronic health records for the diagnosis of rare diseases. *Kidney Int* **97**, 676–686 (2020). URL https://www.ncbi.nlm.nih.gov/pubmed/32111372.

[140] Singhal, P. *et al.* Opportunities and challenges for biomarker discovery using electronic health record data. *Trends Mol Med* **29**, 765–776 (2023). URL https://www.ncbi.nlm.nih.gov/pubmed/37474378.

[141] Shilo, S., Rossman, H. & Segal, E. Axes of a revolution: challenges and promises of big data in healthcare. *Nat Med* **26**, 29–38 (2020). URL https://www.ncbi.nlm.nih.gov/pubmed/31932803.

[142] Helgeland, O. *et al.* Characterization of the genetic architecture of infant and early childhood body mass index. *Nat Metab* **4**, 344–358 (2022). URL https://www.ncbi.nlm.nih.gov/pubmed/35315439.

[143] Couto Alves, A. *et al.* GWAS on longitudinal growth traits reveals different genetic factors influencing infant, child, and adult BMI. *Sci Adv* **5**, eaaw3095 (2019). URL https://www.ncbi.nlm.nih.gov/pubmed/31840077.

[144] Fraser, A. *et al.* Cohort Profile: the Avon Longitudinal Study of Parents and Children: ALSPAC mothers cohort. *Int J Epidemiol* **42**, 97–110 (2013). URL https://www.ncbi.nlm.nih.gov/pubmed/22507742.

[145] Magnus, P. *et al.* Cohort Profile Update: The Norwegian Mother and Child Cohort Study (MoBa). *Int J Epidemiol* **45**, 382–8 (2016). URL https://www.ncbi.nlm.nih.gov/pubmed/27063603.

[146] Brouwer, R. M. *et al.* Genetic variants associated with longitudinal changes in brain structure across the lifespan. *Nat Neurosci* **25**, 421–432 (2022). URL https://www.ncbi.nlm.nih.gov/pubmed/35383335.

[147] Ko, S. *et al.* GWAS of longitudinal trajectories at biobank scale. *Am J Hum Genet* **109**, 433–445 (2022). URL https://www.ncbi.nlm.nih.gov/pubmed/35196515.

[148] Bowman, S. Impact of electronic health record systems on information integrity: quality and safety implications. *Perspect Health Inf Manag* **10**, 1c (2013). URL https://www.ncbi.nlm.nih.gov/pubmed/24159271.

[149] Goldstein, B. A., Bhavsar, N. A., Phelan, M. & Pencina, M. J. Controlling for informed presence bias due to the number of health encounters in an electronic health record. *Am J Epidemiol* **184**, 847–855 (2016). URL https://www.ncbi.nlm.nih.gov/pubmed/27852603.

[150] McGee, G., Haneuse, S., Coull, B. A., Weisskopf, M. G. & Rotem, R. S. On the nature of informative presence bias in analyses of electronic health records. *Epidemiology* **33**, 105–113 (2022). URL https://www.ncbi.nlm.nih.gov/pubmed/34711733.

[151] Zheng, T. *et al.* A machine learning-based framework to identify type 2 diabetes through electronic health records. *Int J Med Inform* **97**, 120–127 (2017). URL https://www.ncbi.nlm.nih.gov/pubmed/27919371.

[152] Pham, A. N. Q. *et al.* Development and validation of a case definition for problematic menopause in primary care electronic medical records. *BMC Med Inform Decis Mak* **23**, 202 (2023). URL https://www.ncbi.nlm.nih.gov/pubmed/37798700.

[153] for Health, N. I. & Care Excellence, U. Diagnosis of menopause and perimenopause. https://cks.nice.org.uk/topics/menopause/diagnosis/diagnosis-of-menopause-perimenopause/ (2022). Accessed: 2022-09-14.

[154] Funk, L. M., Shan, Y., Voils, C. I., Kloke, J. & Hanrahan, L. P. Electronic health record data versus the National Health and Nutrition Examination Survey (NHANES): A comparison of overweight and obesity rates. *Med Care* **55**, 598–605 (2017). URL https://www.ncbi.nlm.nih.gov/pubmed/28079710.

[155] Baer, H. J., Karson, A. S., Soukup, J. R., Williams, D. H. & Bates, D. W. Documentation and diagnosis of overweight and obesity in electronic health records of adult primary care patients. *JAMA Intern Med* **173**, 1648–52 (2013). URL https://www.ncbi.nlm.nih.gov/pubmed/23835808.

[156] Melmed, S., Polonsky, K. S., Larsen, P. R. & Kronenberg, H. M. Williams textbook of. *Endocrinology 13th* (2015).

[157] Papadimitriou, A. & Priftis, K. N. Regulation of the hypothalamic-pituitary-adrenal axis. *Neuroimmunomodulation* **16**, 265–71 (2009). URL https://www.ncbi.nlm.nih.gov/pubmed/19571587.

[158] Ortiga-Carvalho, T. M., Chiamolera, M. I., Pazos-Moura, C. C. & Wondisford, F. E. Hypothalamus-pituitary-thyroid axis. *Comprehensive Physiology* **6**, 1387–1428 (2011).

[159] Plant, T. M. 60 years of neuroendocrinology: The hypothalamo-pituitary-gonadal axis. *J Endocrinol* **226**, T41–54 (2015). URL https://www.ncbi.nlm.nih.gov/pubmed/25901041.

[160] Miller, W. L. Molecular biology of steroid hormone synthesis. *Endocrine reviews* **9**, 295–318 (1988).

[161] Duarte-Guterman, P., Navarro-Martin, L. & Trudeau, V. L. Mechanisms of crosstalk between endocrine systems: regulation of sex steroid hormone synthesis and action by thyroid hormones. *Gen Comp Endocrinol* **203**, 69–85 (2014). URL https://www.ncbi.nlm.nih.gov/pubmed/24685768.

[162] Hulbert, A. J. Thyroid hormones and their effects: a new perspective. *Biol Rev Camb Philos Soc* **75**, 519–631 (2000). URL https://www.ncbi.nlm.nih.gov/pubmed/11117200.

[163] Baxter, J. D. & Rousseau, G. G. Glucocorticoid hormone action: an overview. *Monogr Endocrinol* **12**, 1–24 (1979). URL https://www.ncbi.nlm.nih.gov/pubmed/386083.

[164] Chan, L. & O'Malley, B. W. Mechanism of action of the sex steroid hormones (first of three parts). *N Engl J Med* **294**, 1322–8 (1976). URL https://www.ncbi.nlm.nih.gov/pubmed/177871.

[165] Ronti, T., Lupattelli, G. & Mannarino, E. The endocrine function of adipose tissue: an update. *Clin Endocrinol (Oxf)* **64**, 355–65 (2006). URL https://www.ncbi.nlm.nih.gov/pubmed/16584505.

[166] Papargyri, P. *et al.* Links between HPA axis and adipokines: clinical implications in paradigms of stress-related disorders. *Expert Rev Endocrinol Metab* **13**, 317–332 (2018). URL https://www.ncbi.nlm.nih.gov/pubmed/30422016.

[167] Li, J., Papadopoulos, V. & Vihma, V. Steroid biosynthesis in adipose tissue. *Steroids* **103**, 89–104 (2015). URL https://www.ncbi.nlm.nih.gov/pubmed/25846979.

[168] Infante, M., Armani, A., Mammi, C., Fabbri, A. & Caprio, M. Impact of adrenal steroids on regulation of adipose tissue. *Compr Physiol* **7**, 1425–1447 (2017). URL https://www.ncbi.nlm.nih.gov/pubmed/28915330.

[169] Jazet, I. M., Pijl, H. & Meinders, A. Adipose tissue as an endocrine organ: im pact on insulin resistance. *The Netherlands journal of medicine* **61**, 194–212 (2003).

[170] Bjorntorp, P. Growth hormone, insulin-like growth factor-I and lipid metabolism: interactions with sex steroids. *Horm Res* **46**, 188–91 (1996). URL https://www.ncbi.nlm.nih.gov/pubmed/8950619.

[171] Malhotra, R., Ostbye, T., Riley, C. M. & Finkelstein, E. A. Young adult weight trajectories through midlife by body mass category. *Obesity (Silver Spring)* **21**, 1923–34 (2013). URL https://www.ncbi.nlm.nih.gov/pubmed/23408493.

[172] Botoseneanu, A. & Liang, J. Latent heterogeneity in long-term trajectories of body mass index in older adults. *J Aging Health* **25**, 342–63 (2013). URL https://www.ncbi.nlm.nih.gov/pubmed/23264442.

[173] Marinac, C. R. *et al.* Elucidating under-studied aspects of the link between obesity and multiple myeloma: Weight pattern, body shape trajectory, and body fat distribution. *JNCI Cancer Spectr* **3**, pkz044 (2019). URL https://www.ncbi.nlm.nih.gov/pubmed/31448358.

[174] Kuwahara, K. *et al.* Body mass index trajectory patterns and changes in visceral fat and glucose metabolism before the onset of type 2 diabetes. *Sci Rep* **7**, 43521 (2017). URL https://www.ncbi.nlm.nih.gov/pubmed/28266592.

[175] Kim, M. N. *et al.* Weight gain during early adulthood, trajectory of body shape and the risk of nonalcoholic fatty liver disease: A prospective cohort study among women. *Metabolism* **113**, 154398 (2020). URL https://www.ncbi.nlm.nih.gov/pubmed/33058854.

[176] Gilsanz, V., Hu, H. H. & Kajimura, S. Relevance of brown adipose tissue in infancy and adolescence. *Pediatr Res* **73**, 3–9 (2013). URL https://www.ncbi.nlm.nih.gov/pubmed/23090604.

[177] Frank, A. P., de Souza Santos, R., Palmer, B. F. & Clegg, D. J. Determinants of body fat distribution in humans may provide insight about obesity-related health risks. *J Lipid Res* **60**, 1710–1719 (2019). URL https://www.ncbi.nlm.nih.gov/pubmed/30097511.

[178] Lumish, H. S., O'Reilly, M. & Reilly, M. P. Sex differences in genomic drivers of adipose distribution and related cardiometabolic disorders: Opportunities for precision medicine. *Arterioscler Thromb Vasc Biol* **40**, 45–60 (2020). URL https://www.ncbi.nlm.nih.gov/pubmed/31747800.

[179] Gavin, K. M. & Bessesen, D. H. Sex differences in adipose tissue function. *Endocrinol Metab Clin North Am* **49**, 215–228 (2020). URL https://www.ncbi.nlm.nih.gov/pubmed/32418585.

[180] Ambikairajah, A., Walsh, E., Tabatabaei-Jafari, H. & Cherbuin, N. Fat mass changes during menopause: a metaanalysis. *Am J Obstet Gynecol* **221**, 393–409 e50 (2019). URL https://www.ncbi.nlm.nih.gov/pubmed/31034807.

[181] Goossens, G. H., Jocken, J. W. E. & Blaak, E. E. Sexual dimorphism in cardiometabolic health: the role of adipose tissue, muscle and liver. *Nat Rev Endocrinol* **17**, 47–66 (2021). URL https://www.ncbi.nlm.nih.gov/pubmed/33173188.

[182] Tchernof, A. & Despres, J. P. Pathophysiology of human visceral obesity: an update. *Physiol Rev* **93**, 359–404 (2013). URL https://www.ncbi.nlm.nih.gov/pubmed/23303913.

[183] Al-Sofiani, M. E., Ganji, S. S. & Kalyani, R. R. Body composition changes in diabetes and aging. *J Diabetes Complications* **33**, 451–459 (2019). URL https://www.ncbi.nlm.nih.gov/pubmed/31003924.

[184] Ballin, M., Nordstrom, P., Niklasson, J. & Nordstrom, A. Associations of visceral adipose tissue and skeletal muscle density with incident stroke, myocardial infarction, and all-cause mortality in community-dwelling 70-year-old individuals: A prospective cohort study. *J Am Heart Assoc* **10**, e020065 (2021). URL https://www.ncbi.nlm.nih.gov/pubmed/33870709.

[185] Yang, Y. *et al.* Visceral adiposity and high intramuscular fat deposition independently predict critical illness in patients with SARS-CoV-2. *Obesity (Silver Spring)* **28**, 2040–2048 (2020). URL https://www.ncbi.nlm.nih.gov/pubmed/32677752.

[186] He, Y., Tian, J., Oddy, W. H., Dwyer, T. & Venn, A. J. Association of childhood obesity with female infertility in adulthood: a 25-year follow-up study. *Fertil Steril* **110**, 596–604 e1 (2018). URL https://www.ncbi.nlm.nih.gov/pubmed/30196944.

[187] Thorogood, M., Seed, M., De Mott, K. & Guideline Development, G. Management of fertility in women with familial hypercholesterolaemia: summary of NICE guidance. *BJOG* **116**, 478–9 (2009). URL https://www.ncbi.nlm.nih.gov/pubmed/19250359.

[188] Marais, A. D. Familial hypercholesterolaemia. *Clin Biochem Rev* **25**, 49–68 (2004). URL https://www.ncbi.nlm.nih.gov/pubmed/18516203.

[189] Krude, H. *et al.* Severe early-onset obesity, adrenal insufficiency and red hair pigmentation caused by pomc mutations in humans. *Nat Genet* **19**, 155–7 (1998). URL https://www.ncbi.nlm.nih.gov/pubmed/9620771.

[190] Pillon, N. J., Loos, R. J. F., Marshall, S. M. & Zierath, J. R. Metabolic consequences of obesity and type 2 diabetes: Balancing genes and environment for personalized care. *Cell* **184**, 1530–1544 (2021). URL https://www.ncbi.nlm.nih.gov/pubmed/33675692.

[191] Alwan, H. *et al.* A systematic review and meta-analysis investigating the relationship between metabolic syndrome and the incidence of thyroid diseases. *Endocrine* (2023). URL https://www.ncbi.nlm.nih.gov/pubmed/37688711.

[192] Kim, N. *et al.* Effect of lipid metabolism on male fertility. *Biochem Biophys Res Commun* **485**, 686–692 (2017). URL https://www.ncbi.nlm.nih.gov/pubmed/28235483.

[193] Pugh, S. J. *et al.* Preconception maternal lipoprotein levels in relation to fecundability. *Hum Reprod* **32**, 1055–1063 (2017). URL https://www.ncbi.nlm.nih.gov/pubmed/28333301.

[194] Wang, S., Wang, J., Jiang, Y. & Jiang, W. Association between blood lipid level and embryo quality during in vitro fertilization. *Medicine (Baltimore)* **99**, e19665 (2020). URL https://www.ncbi.nlm.nih.gov/pubmed/32221094.

[195] Zhu, S. *et al.* Imaging-based body fat distribution in polycystic ovary syndrome: A systematic review and meta-analysis. *Front Endocrinol (Lausanne)* **12**, 697223 (2021). URL https://www.ncbi.nlm.nih.gov/pubmed/34566888.

[196] Azziz, R. Polycystic ovary syndrome. *Obstet Gynecol* **132**, 321–336 (2018). URL https://www.ncbi.nlm.nih.gov/pubmed/29995717.

[197] 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**, e1007813 (2018). URL https://www.ncbi.nlm.nih.gov/pubmed/30566500.

[198] Loos, R. J. F. & Yeo, G. S. H. The genetics of obesity: from discovery to biology. *Nat Rev Genet* **23**, 120–133 (2022). URL https://www.ncbi.nlm.nih.gov/pubmed/34556834.

[199] Schulze, M. B. Metabolic health in normal-weight and obese individuals. *Diabetologia* **62**, 558–566 (2019). URL https://www.ncbi.nlm.nih.gov/pubmed/30569272.

[200] Katta, N., Loethen, T., Lavie, C. J. & Alpert, M. A. Obesity and coronary heart disease: Epidemiology, pathology, and coronary artery imaging. *Curr Probl Cardiol* **46**, 100655 (2021). URL https://www.ncbi.nlm.nih.gov/pubmed/32843206.

[201] Pugliese, G. *et al.* Obesity and infectious diseases: pathophysiology and epidemiology of a double pandemic condition. *Int J Obes (Lond)* **46**, 449–465 (2022). URL https://www.ncbi.nlm.nih.gov/pubmed/35058571.

[202] Hou, J. *et al.* Obesity and bone health: A complex link. *Front Cell Dev Biol* **8**, 600181 (2020). URL https://www.ncbi.nlm.nih.gov/pubmed/33409277.

[203] Ward, L. C. Bioelectrical impedance analysis for body composition assessment: reflections on accuracy, clinical utility, and standardisation. *Eur J Clin Nutr* **73**, 194–199 (2019). URL https://www.ncbi.nlm.nih.gov/pubmed/30297760.

[204] Rask-Andersen, M., Karlsson, T., Ek, W. E. & Johansson, A. Genome-wide association study of body fat distribution identifies adiposity loci and sex-specific genetic effects. *Nat Commun* **10**, 339 (2019). URL https://www.ncbi.nlm.nih.gov/pubmed/30664634.

[205] Feng, G. J. *et al.* Identification of pleiotropic loci underlying hip bone mineral density and trunk lean mass. *J Hum Genet* **66**, 251–260 (2021). URL https://www.ncbi.nlm.nih.gov/pubmed/32929176.

[206] Pei, Y. F. *et al.* The genetic architecture of appendicular lean mass characterized by association analysis in the UK Biobank study. *Commun Biol* **3**, 608 (2020). URL https://www.ncbi.nlm.nih.gov/pubmed/33097823.

[207] Bracco, D. *et al.* Segmental body composition assessed by bioelectrical impedance analysis and DEXA in humans. *J Appl Physiol (1985)* **81**, 2580–7 (1996). URL https://www.ncbi.nlm.nih.gov/pubmed/9018509.

[208] Rashmi, R. & Snekhalatha, U. Evaluation of body composition parameters using various diagnostic methods: A meta analysis study. *Obesity Medicine* **16**, 100150 (2019).

[209] Min, J., Chiu, D. T. & Wang, Y. Variation in the heritability of body mass index based on diverse twin studies: a systematic review. *Obes Rev* **14**, 871–82 (2013). URL https://www.ncbi.nlm.nih.gov/pubmed/23980914.

[210] Chami, N., Preuss, M., Walker, R. W., Moscati, A. & Loos, R. J. F. The role of polygenic susceptibility to obesity among carriers of pathogenic mutations in MC4R in the UK Biobank population. *PLoS Med* **17**, e1003196 (2020). URL https://www.ncbi.nlm.nih.gov/pubmed/32692746.

[211] Loos, R. J. & Yeo, G. S. The bigger picture of FTO: the first GWAS-identified obesity gene. *Nat Rev Endocrinol* **10**, 51–61 (2014). URL https://www.ncbi.nlm.nih.gov/pubmed/24247219.

[212] Pan, D. Z. *et al.* Identification of TBX15 as an adipose master trans regulator of abdominal obesity genes. *Genome Med* **13**, 123 (2021). URL https://www.ncbi.nlm.nih.gov/pubmed/34340684.

[213] Leyden, G. M. *et al.* Harnessing tissue-specific genetic variation to dissect putative causal pathways between body mass index and cardiometabolic phenotypes. *Am J Hum Genet* **109**, 240–252 (2022). URL https://www.ncbi.nlm.nih.gov/pubmed/35090585.

[214] Baumgartner, R. N., Chumlea, W. C. & Roche, A. F. Associations between bioelectric impedance and anthropometric variables. *Hum Biol* **59**, 235–44 (1987). URL https://www.ncbi.nlm.nih.gov/pubmed/3596564.

[215] Agrawal, S. *et al.* Inherited basis of visceral, abdominal subcutaneous and gluteofemoral fat depots. *Nat Commun* **13**, 3771 (2022). URL https://www.ncbi.nlm.nih.gov/pubmed/35773277.

[216] Pulit, S. L. *et al.* Meta-analysis of genome-wide association studies for body fat distribution in 694 649 individuals of European ancestry. *Hum Mol Genet* **28**, 166–174 (2019).

[217] Wen, W. *et al.* Genome-wide association studies in East Asians identify new loci for waist-hip ratio and waist circumference. *Sci Rep* **6**, 17958 (2016). URL https://www.ncbi.nlm.nih.gov/pubmed/26785701.

[218] Hansen, G. T. *et al.* Genetics of sexually dimorphic adipose distribution in humans. *Nat Genet* **55**, 461–470 (2023). URL https://www.ncbi.nlm.nih.gov/pubmed/36797366.

[219] Winkler, T. W. *et al.* A joint view on genetic variants for adiposity differentiates subtypes with distinct metabolic implications. *Nat Commun* **9**, 1946 (2018). URL https://www.ncbi.nlm.nih.gov/pubmed/29769528.

[220] Block, J. P., Subramanian, S. V., Christakis, N. A. & O'Malley, A. J. Population trends and variation in body mass index from 1971 to 2008 in the Framingham Heart Study Offspring Cohort. *PLoS One* **8**, e63217 (2013). URL https://www.ncbi.nlm.nih.gov/pubmed/23675464.

[221] Kaidesoja, M. *et al.* FinnTwin16: A longitudinal study from age 16 of a population-based Finnish twin cohort. *Twin Res Hum Genet* **22**, 530–539 (2019). URL https://www.ncbi.nlm.nih.gov/pubmed/31796134.

[222] Toth, M. J., Tchernof, A., Sites, C. K. & Poehlman, E. T. Menopause-related changes in body fat distribution. *Ann N Y Acad Sci* **904**, 502–6 (2000). URL https://www.ncbi.nlm.nih.gov/pubmed/10865795.

[223] Winkler, T. W. *et al.* The influence of age and sex on genetic associations with adult body size and shape: A large-scale genome-wide interaction study. *PLoS Genet* **11**, e1005378 (2015). URL https://www.ncbi.nlm.nih.gov/pubmed/26426971.

[224] Hjelmborg, J. *et al.* Genetic influences on growth traits of BMI: a longitudinal study of adult twins. *Obesity (Silver Spring)* **16**, 847–52 (2008). URL https://www.ncbi.nlm.nih.gov/pubmed/18239571.

[225] Austin, M. A. *et al.* Genetic influences on changes in body mass index: a longitudinal analysis of women twins. *Obes Res* **5**, 326–31 (1997). URL https://www.ncbi.nlm.nih.gov/pubmed/9285839.

[226] Mihm, M., Gangooly, S. & Muttukrishna, S. The normal menstrual cycle in women. *Anim Reprod Sci* **124**, 229–36 (2011). URL https://www.ncbi.nlm.nih.gov/pubmed/20869180.

[227] Kiesel, L. & Sourouni, M. Diagnosis of endometriosis in the 21st century. *Climacteric* **22**, 296–302 (2019). URL https://www.ncbi.nlm.nih.gov/pubmed/30905186.

[228] Agarwal, S. K. *et al.* Clinical diagnosis of endometriosis: a call to action. *Am J Obstet Gynecol* **220**, 354 e1–354 e12 (2019). URL https://www.ncbi.nlm.nih.gov/pubmed/30625295.

[229] Carte, E. *'Dirty' and Infertile: the Underdiagnosis and Impact of Chlamydia*. Master's thesis, 'Georgetown University' (2014).

[230] Oakley, L., Doyle, P. & Maconochie, N. Lifetime prevalence of infertility and infertility treatment in the UK: results from a population-based survey of reproduction. *Hum Reprod* **23**, 447–50 (2008). URL https://www.ncbi.nlm.nih.gov/pubmed/18033808.

[231] Simionescu, G. *et al.* The complex relationship between infertility and psychological distress (review). *Exp Ther Med* **21**, 306 (2021). URL https://www.ncbi.nlm.nih.gov/pubmed/33717249.

[232] Wu, W., La, J., Schubach, K. M., Lantsberg, D. & Katz, D. J. Psychological, social, and sexual challenges affecting men receiving male infertility treatment: a systematic review and implications for clinical care. *Asian J Androl* **25**, 448–453 (2023). URL https://www.ncbi.nlm.nih.gov/pubmed/36412462.

[233] Schmidt, L., Munster, K. & Helm, P. Infertility and the seeking of infertility treatment in a representative population. *Br J Obstet Gynaecol* **102**, 978–84 (1995). URL https://www.ncbi.nlm.nih.gov/pubmed/8652489.

[234] Kessler, L. M., Craig, B. M., Plosker, S. M., Reed, D. R. & Quinn, G. P. Infertility evaluation and treatment among women in the United States. *Fertil Steril* **100**, 1025–32 (2013). URL https://www.ncbi.nlm.nih.gov/pubmed/23849845.

[235] Barratt, C. L. R. *et al.* The diagnosis of male infertility: an analysis of the evidence to support the development of global WHO guidance-challenges and future research opportunities. *Hum Reprod Update* **23**, 660–680 (2017). URL https://www.ncbi.nlm.nih.gov/pubmed/28981651.

[236] Ruth, K. S. *et al.* Using human genetics to understand the disease impacts of testosterone in men and women. *Nat Med* **26**, 252–258 (2020). URL https://www.ncbi.nlm.nih.gov/pubmed/32042192.

[237] Oduwole, O. O., Huhtaniemi, I. T. & Misrahi, M. The roles of luteinizing hormone, follicle-stimulating hormone and testosterone in spermatogenesis and folliculogenesis revisited. *Int J Mol Sci* **22** (2021). URL https://www.ncbi.nlm.nih.gov/pubmed/34884539.

[238] Nokoff, N. *et al.* Sex differences in effects of obesity on reproductive hormones and glucose metabolism in early puberty. *J Clin Endocrinol Metab* **104**, 4390–4397 (2019). URL https://www.ncbi.nlm.nih.gov/pubmed/30985874.

[239] Gambineri, A. & Pelusi, C. Sex hormones, obesity and type 2 diabetes: is there a link? *Endocr Connect* **8**, R1–R9 (2019). URL https://www.ncbi.nlm.nih.gov/pubmed/30533003.

[240] Goodale, T., Sadhu, A., Petak, S. & Robbins, R. Testosterone and the heart. *Methodist Debakey Cardiovasc J* **13**, 68–72 (2017). URL https://www.ncbi.nlm.nih.gov/pubmed/28740585.

[241] Schafer, S. *et al.* Low testosterone concentrations and prediction of future heart failure in men and in women: evidence from the large FINRISK97 study. *ESC Heart Fail* **8**, 2485–2491 (2021). URL https://www.ncbi.nlm.nih.gov/pubmed/33934533.

[242] Buendia-Gonzalez, F. O. & Legorreta-Herrera, M. The similarities and differences between the effects of testosterone and DHEA on the innate and adaptive immune response. *Biomolecules* **12** (2022). URL https://www.ncbi.nlm.nih.gov/pubmed/36551196.

[243] Bianchi, V. E. The anti-inflammatory effects of testosterone. *J Endocr Soc* **3**, 91–107 (2019). URL https://www.ncbi.nlm.nih.gov/pubmed/30582096.

[244] Kroon, J., Pereira, A. M. & Meijer, O. C. Glucocorticoid sexual dimorphism in metabolism: Dissecting the role of sex hormones. *Trends Endocrinol Metab* **31**, 357–367 (2020). URL https://www.ncbi.nlm.nih.gov/pubmed/32037025.

[245] Lipshutz, S. E., George, E. M., Bentz, A. B. & Rosvall, K. A. Evaluating testosterone as a phenotypic integrator: From tissues to individuals to species. *Mol Cell Endocrinol* **496**, 110531 (2019). URL https://www.ncbi.nlm.nih.gov/pubmed/31376416.

[246] Pott, J. *et al.* Genetic association study of eight steroid hormones and implications for sexual dimorphism of coronary artery disease. *J Clin Endocrinol Metab* **104**, 5008–5023 (2019). URL https://www.ncbi.nlm.nih.gov/pubmed/31169883.

[247] Tang, Z. R., Zhang, R., Lian, Z. X., Deng, S. L. & Yu, K. Estrogen-receptor expression and function in female reproductive disease. *Cells* **8** (2019). URL https://www.ncbi.nlm.nih.gov/pubmed/31546660.

[248] Cangiano, B., Swee, D. S., Quinton, R. & Bonomi, M. Genetics of congenital hypogonadotropic hypogonadism: peculiarities and phenotype of an oligogenic disease. *Hum Genet* **140**, 77–111 (2021). URL https://www.ncbi.nlm.nih.gov/pubmed/32200437.

[249] Krassas, G. E., Poppe, K. & Glinoer, D. Thyroid function and human reproductive health. *Endocr Rev* **31**, 702–55 (2010). URL https://www.ncbi.nlm.nih.gov/pubmed/20573783.

[250] Tyrmi, J. S. *et al.* Leveraging Northern European population history: novel low-frequency variants for polycystic ovary syndrome. *Hum Reprod* **37**, 352–365 (2022). URL https://www.ncbi.nlm.nih.gov/pubmed/34791234.

[251] Rahmioglu, N. *et al.* The genetic basis of endometriosis and comorbidity with other pain and inflammatory conditions. *Nat Genet* **55**, 423–436 (2023). URL https://www.ncbi.nlm.nih.gov/pubmed/36914876.

[252] Gallagher, C. S. *et al.* Genome-wide association and epidemiological analyses reveal common genetic origins between uterine leiomyomata and endometriosis. *Nat Commun* **10**, 4857 (2019). URL https://www.ncbi.nlm.nih.gov/pubmed/31649266.

[253] Bonavina, G. & Taylor, H. S. Endometriosis-associated infertility: From pathophysiology to tailored treatment. *Front Endocrinol (Lausanne)* **13**, 1020827 (2022). URL https://www.ncbi.nlm.nih.gov/pubmed/36387918.

[254] Thessaloniki, E. A.-S. P. C. W. G. Consensus on infertility treatment related to polycystic ovary syndrome. *Hum Reprod* **23**, 462–77 (2008). URL https://www.ncbi.nlm.nih.gov/pubmed/18308833.

[255] Kuchenbecker, W. K. *et al.* The subcutaneous abdominal fat and not the intraabdominal fat compartment is associated with anovulation in women with obesity and infertility. *J Clin Endocrinol Metab* **95**, 2107–12 (2010). URL https://www.ncbi.nlm.nih.gov/pubmed/20200335.

[256] Eisenberg, M. L. *et al.* Male infertility. *Nat Rev Dis Primers* **9**, 49 (2023). URL https://www.ncbi.nlm.nih.gov/pubmed/37709866.

[257] Oud, M. S. *et al.* A de novo paradigm for male infertility. *Nat Commun* **13**, 154 (2022). URL https://www.ncbi.nlm.nih.gov/pubmed/35013161.

[258] Li, Y. *et al.* Whole-exome sequencing of a cohort of infertile men reveals novel causative genes in teratozoospermia that are chiefly related to sperm head defects. *Hum Reprod* **37**, 152–177 (2021). URL https://www.ncbi.nlm.nih.gov/pubmed/34791246.

[259] Gardner, E. J. *et al.* Reduced reproductive success is associated with selective constraint on human genes. *Nature* **603**, 858–863 (2022). URL https://www.ncbi.nlm.nih.gov/pubmed/35322230.

[260] Mathieson, I. *et al.* Genome-wide analysis identifies genetic effects on reproductive success and ongoing natural selection at the FADS locus. *Nat Hum Behav* **7**, 790–801 (2023). URL https://www.ncbi.nlm.nih.gov/pubmed/36864135.

[261] Barban, N. *et al.* Genome-wide analysis identifies 12 loci influencing human reproductive behavior. *Nat Genet* **48**, 1462–1472 (2016). URL https://www.ncbi.nlm.nih.gov/pubmed/27798627.

[262] Day, F. R. *et al.* Physical and neurobehavioral determinants of reproductive onset and success. *Nat Genet* **48**, 617–623 (2016). URL https://www.ncbi.nlm.nih.gov/pubmed/27089180.

[263] Mills, M. C. *et al.* Identification of 371 genetic variants for age at first sex and birth linked to externalising behaviour. *Nat Hum Behav* **5**, 1717–1730 (2021). URL https://www.ncbi.nlm.nih.gov/pubmed/34211149.

[264] 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). URL https://www.ncbi.nlm.nih.gov/pubmed/28641372.

[265] Szustakowski, J. D. *et al.* Advancing human genetics research and drug discovery through exome sequencing of the UK Biobank. *Nat Genet* **53**, 942–948 (2021). URL https://www.ncbi.nlm.nih.gov/pubmed/34183854.

[266] Elliott, P., Peakman, T. C. & Biobank, U. K. The UK Biobank sample handling and storage protocol for the collection, processing and archiving of human blood and urine. *Int J Epidemiol* **37**, 234–44 (2008). URL https://www.ncbi.nlm.nih.gov/pubmed/18381398.

[267] Miller, K. L. *et al.* Multimodal population brain imaging in the UK Biobank prospective epidemiological study. *Nat Neurosci* **19**, 1523–1536 (2016). URL https://www.ncbi.nlm.nih.gov/pubmed/27643430.

[268] Petersen, S. E. *et al.* Imaging in population science: cardiovascular magnetic resonance in 100,000 participants of UK Biobank - rationale, challenges and approaches. *J Cardiovasc Magn Reson* **15**, 46 (2013). URL https://www.ncbi.nlm.nih.gov/pubmed/23714095.

[269] Coffey, S. *et al.* Protocol and quality assurance for carotid imaging in 100,000 participants of UK Biobank: development and assessment. *Eur J Prev Cardiol* **24**, 1799–1806 (2017). URL https://www.ncbi.nlm.nih.gov/pubmed/28925747.

[270] Doherty, A. *et al.* Large scale population assessment of physical activity using wrist worn accelerometers: The UK Biobank study. *PLoS One* **12**, e0169649 (2017). URL https://www.ncbi.nlm.nih.gov/pubmed/28146576.

[271] Team, U. B. UK Biobank Primary Care Linked Data. https://biobank.ndph.ox.ac.uk/showcase/showcase/docs/primary_care_data.pdf (2019). Accessed: 2020-01-03.

[272] EMIS Health. https://www.emishealth.com/. Accessed: 2023-07-31.

[273] Vision Health. https://www.cegedim-healthcare.co.uk/. Accessed: 2023-07-31.

[274] TPP UK. https://tpp-uk.com/. Accessed: 2023-07-31.

[275] NHS UK Read Codes. https://isd.digital.nhs.uk/trud/user/guest/group/0/pack/9. Accessed: 2020-09-25.

[276] 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**, 769–81 (2015). URL https://www.ncbi.nlm.nih.gov/pubmed/26423011.

[277] Affymetrix. Axiom Genotyping Solution Data Analysis Guide. http://tools.thermofisher.com/content/sfs/manuals/axiom_genotyping_solution_analysis_guide.pdf (2017). Accessed: 2020-02-14.

[278] Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–73 (2010). URL https://www.ncbi.nlm.nih.gov/pubmed/20926424.

[279] O'Connell, J. *et al.* Haplotype estimation for biobank-scale data sets. *Nat Genet* **48**, 817–20 (2016). URL https://www.ncbi.nlm.nih.gov/pubmed/27270105.

[280] Genomes Project, C. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015). URL https://www.ncbi.nlm.nih.gov/pubmed/26432245.

[281] McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet* **48**, 1279–83 (2016). URL https://www.ncbi.nlm.nih.gov/pubmed/27548312.

[282] Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G. R. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat Genet* **44**, 955–9 (2012). URL https://www.ncbi.nlm.nih.gov/pubmed/22820512.

[283] Huang, J. *et al.* Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. *Nat Commun* **6**, 8111 (2015). URL https://www.ncbi.nlm.nih.gov/pubmed/26368830.

[284] Team, U. B. UK Biobank Final Exome Data Release FAQs / July 2022. https://www.ukbiobank.ac.uk/media/najcnoaz/access_064-uk-biobank-exome-release-faq_v11-1_final-002.pdf (2022). Accessed: 2023-02-12.

[285] Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM (2013). 1303.3997.

[286] Yun, T. *et al.* Accurate, scalable cohort variant calls using DeepVariant and GLnexus. *Bioinformatics* **36**, 5582–5589 (2021). URL https://www.ncbi.nlm.nih.gov/pubmed/33399819.

[287] Denaxas, S. *et al.* A semi-supervised approach for rapidly creating clinical biomarker phenotypes in the UK Biobank using different primary care EHR and clinical terminology systems. *JAMIA Open* **3**, 545–556 (2020). URL https://www.ncbi.nlm.nih.gov/pubmed/33619467.

[288] Ioannidis, J. P. A. & Bossuyt, P. M. M. Waste, leaks, and failures in the biomarker pipeline. *Clin Chem* **63**, 963–972 (2017). URL https://www.ncbi.nlm.nih.gov/pubmed/28270433.

[289] Kuan, V. *et al.* A chronological map of 308 physical and mental health conditions from 4 million individuals in the English National Health Service. *Lancet Digit Health* **1**, e63–e77 (2019). URL https://www.ncbi.nlm.nih.gov/pubmed/31650125.

[290] Weir, B. S. Genetic data analysis II. *(No Title)* (1996).

[291] Prive, F., Luu, K., Blum, M. G. B., McGrath, J. J. & Vilhjalmsson, B. J. Efficient toolkit implementing best practices for principal component analysis of population genetic data. *Bioinformatics* **36**, 4449–4457 (2020). URL https://www.ncbi.nlm.nih.gov/pubmed/32415959.

[292] Kivisild, T. *et al.* Patterns of genetic connectedness between modern and medieval Estonian genomes reveal the origins of a major ancestry component of the Finnish population. *Am J Hum Genet* **108**, 1792–1806 (2021). URL https://www.ncbi.nlm.nih.gov/pubmed/34411538.

[293] Whiffin, N., Ware, J. S. & O'Donnell-Luria, A. Improving the understanding of genetic variants in rare disease with large-scale reference populations. *JAMA* **322**, 1305–1306 (2019). URL https://www.ncbi.nlm.nih.gov/pubmed/31469401.

[294] Kanai, M. *et al.* Meta-analysis fine-mapping is often miscalibrated at single-variant resolution. *Cell Genom* **2** (2022). URL https://www.ncbi.nlm.nih.gov/pubmed/36643910.

[295] Benonisdottir, S. *et al.* Epigenetic and genetic components of height regulation. *Nat Commun* **7**, 13490 (2016). URL https://www.ncbi.nlm.nih.gov/pubmed/27848971.

[296] 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). URL https://www.ncbi.nlm.nih.gov/pubmed/21167468.

[297] Schaffner, S. F. The X chromosome in population genetics. *Nat Rev Genet* **5**, 43–51 (2004). URL https://www.ncbi.nlm.nih.gov/pubmed/14708015.

[298] Veeramah, K. R., Gutenkunst, R. N., Woerner, A. E., Watkins, J. C. & Hammer, M. F. Evidence for increased levels of positive and negative selection on the X chromosome versus autosomes in humans. *Mol Biol Evol* **31**, 2267–82 (2014). URL https://www.ncbi.nlm.nih.gov/pubmed/24830675.

[299] Hinch, A. G., Altemose, N., Noor, N., Donnelly, P. & Myers, S. R. Recombination in the human Pseudoautosomal region PAR1. *PLoS Genet* **10**, e1004503 (2014). URL https://www.ncbi.nlm.nih.gov/pubmed/25033397.

[300] Dempster, E. R. & Lerner, I. M. Heritability of threshold characters. *Genetics* **35**, 212 (1950).

[301] Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020). URL https://www.ncbi.nlm.nih.gov/pubmed/32461654.

[302] McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol* **17**, 122 (2016). URL https://www.ncbi.nlm.nih.gov/pubmed/27268795.

[303] Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res* **47**, D886–D894 (2019). URL https://www.ncbi.nlm.nih.gov/pubmed/30371827.

[304] Ioannidis, N. M. *et al.* REVEL: An ensemble method for predicting the pathogenicity of rare missense variants. *Am J Hum Genet* **99**, 877–885 (2016). URL https://www.ncbi.nlm.nih.gov/pubmed/27666373.

[305] Jaganathan, K. *et al.* Predicting splicing from primary sequence with deep learning. *Cell* **176**, 535–548 e24 (2019). URL https://www.ncbi.nlm.nih.gov/pubmed/30661751.

[306] Morales, J. *et al.* A joint NCBI and EMBL-EBI transcript set for clinical genomics and research. *Nature* **604**, 310–315 (2022). URL https://www.ncbi.nlm.nih.gov/pubmed/35388217.

[307] Richards, S. *et al.* Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med* **17**, 405–24 (2015). URL https://www.ncbi.nlm.nih.gov/pubmed/25741868.

[308] Liu, Y. & Xie, J. Cauchy combination test: a powerful test with analytic p-value calculation under arbitrary dependency structures. *J Am Stat Assoc* **115**, 393–402 (2020). URL https://www.ncbi.nlm.nih.gov/pubmed/33012899.

[309] Fabsitz, R. R., Sholinsky, P. & Carmelli, D. Genetic influences on adult weight gain and maximum body mass index in male twins. *Am J Epidemiol* **140**, 711–20 (1994). URL https://www.ncbi.nlm.nih.gov/pubmed/7942773.

[310] Silventoinen, K. *et al.* Changing genetic architecture of body mass index from infancy to early adulthood: an individual based pooled analysis of 25 twin cohorts. *Int J Obes (Lond)* **46**, 1901–1909 (2022). URL https://www.ncbi.nlm.nih.gov/pubmed/35945263.

[311] Xu, J. *et al.* Exploring the clinical and genetic associations of adult weight trajectories using electronic health records in a racially diverse biobank: a phenome-wide and polygenic risk study. *Lancet Digit Health* **4**, e604–e614 (2022). URL https://www.ncbi.nlm.nih.gov/pubmed/35780037.

[312] Gottesman, O. *et al.* The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future. *Genet Med* **15**, 761–71 (2013). URL https://www.ncbi.nlm.nih.gov/pubmed/23743551.

[313] Monda, K. L. *et al.* A meta-analysis identifies new loci associated with body mass index in individuals of African ancestry. *Nat Genet* **45**, 690–6 (2013). URL https://www.ncbi.nlm.nih.gov/pubmed/23583978.

[314] Postmus, I. *et al.* Pharmacogenetic meta-analysis of genome-wide association studies of LDL cholesterol response to statins. *Nat Commun* **5**, 5068 (2014). URL https://www.ncbi.nlm.nih.gov/pubmed/25350695.

[315] Fan, R. *et al.* Longitudinal association analysis of quantitative traits. *Genet Epidemiol* **36**, 856–69 (2012). URL https://www.ncbi.nlm.nih.gov/pubmed/22965819.

[316] Chiu, Y. F., Justice, A. E. & Melton, P. E. Longitudinal analytical approaches to genetic data. *BMC Genet* **17 Suppl 2**, 4 (2016). URL https://www.ncbi.nlm.nih.gov/pubmed/26866891.

[317] Furlotte, N. A., Eskin, E. & Eyheramendy, S. Genome-wide association mapping with longitudinal data. *Genet Epidemiol* **36**, 463–71 (2012). URL https://www.ncbi.nlm.nih.gov/pubmed/22581622.

[318] Goldstein, J. A. *et al.* LabWAS: Novel findings and study design recommendations from a meta-analysis of clinical labs in two independent biobanks. *PLoS Genet* **16**, e1009077 (2020). URL https://www.ncbi.nlm.nih.gov/pubmed/33175840.

[319] Gauderman, W. J. *et al.* Longitudinal data analysis in pedigree studies. *Genet Epidemiol* **25 Suppl 1**, S18–28 (2003). URL https://www.ncbi.nlm.nih.gov/pubmed/14635165.

[320] Justice, A. E. *et al.* Genome-wide association of trajectories of systolic blood pressure change. *BMC Proc* **10**, 321–327 (2016). URL https://www.ncbi.nlm.nih.gov/pubmed/27980656.

[321] Laird, N. M. & Ware, J. H. Random-effects models for longitudinal data. *Biometrics* **38**, 963–74 (1982). URL https://www.ncbi.nlm.nih.gov/pubmed/7168798.

[322] Xu, H. *et al.* High-throughput and efficient multilocus genome-wide association study on longitudinal outcomes. *Bioinformatics* **36**, 3004–3010 (2020). URL https://www.ncbi.nlm.nih.gov/pubmed/32096821.

[323] Ruppert, D., Wand, M. P. & Carroll, R. J. *Semiparametric regression* (Cambridge University Press, 2003).

[324] Luong, D. T. A. & Chandola, V. A k-means approach to clustering disease progressions. In *2017 IEEE International conference on healthcare informatics (ICHI)*, 268–274 (IEEE, 2017).

[325] Hedman, A. K. *et al.* Identification of novel pheno-groups in heart failure with preserved ejection fraction using machine learning. *Heart* **106**, 342–349 (2020). URL https://www.ncbi.nlm.nih.gov/pubmed/31911501.

[326] Lee, C. & Van Der Schaar, M. Temporal phenotyping using deep predictive clustering of disease progression. In *International Conference on Machine Learning*, 5767–5777 (PMLR, 2020).

[327] Mullin, S. *et al.* Longitudinal K-means approaches to clustering and analyzing EHR opioid use trajectories for clinical subtypes. *J Biomed Inform* **122**, 103889 (2021). URL https://www.ncbi.nlm.nih.gov/pubmed/34411708.

[328] Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* **67**, 1 – 48 (2015). URL https://www.jstatsoft.org/index.php/jss/article/view/v067i01.

[329] Team, R. D. C. R: A language and environment for statistical computing (2010). URL https://cir.nii.ac.jp/crid/1370294721063650048.

[330] Beasley, T. M., Erickson, S. & Allison, D. B. Rank-based inverse normal transformations are increasingly used, but are they merited? *Behavior Genetics* **39**, 580–595 (2009). URL https://doi.org/10.1007/s10519-009-9281-0.

[331] Eilers, P. H. & Marx, B. D. Flexible smoothing with B-splines and penalties. *Statistical science* **11**, 89–121 (1996).

[332] O'Hagan, A. *Kendall's Advanced Theory of Statistic 2B* (John Wiley & Sons, 2010).

[333] Reynolds, A. P., Richards, G., de la Iglesia, B. & Rayward-Smith, V. J. Clustering rules: a comparison of partitioning and hierarchical clustering algorithms. *Journal of Mathematical Modelling and Algorithms* **5**, 475–504 (2006).

[334] Schubert, E. & Rousseeuw, P. J. Faster k-medoids clustering: improving the PAM, CLARA, and CLARANS algorithms. In *Similarity Search and Applications: 12th International Conference, SISAP 2019, Newark, NJ, USA, October 2–4, 2019, Proceedings 12*, 171–187 (Springer, 2019).

[335] Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M. & Hornik, K. *cluster: Cluster Analysis Basics and Extensions* (2022). URL `https://CRAN.R-project.org/package=cluster`. R package version 2.1.4 — For new features, see the 'Changelog' file (in the package source).

[336] Rousseeuw, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics* **20**, 53–65 (1987).

[337] Smithson, M. & Verkuilen, J. A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. *Psychological methods* **11**, 54 (2006).

[338] Pilling, L. C. *et al.* Human longevity: 25 genetic loci associated in 389,166 UK Biobank participants. *Aging (Albany NY)* **9**, 2504–2520 (2017). URL `https://www.ncbi.nlm.nih.gov/pubmed/29227965`.

[339] Serrano-Pozo, A., Das, S. & Hyman, B. T. APOE and Alzheimer's disease: advances in genetics, pathophysiology, and therapeutic approaches. *Lancet Neurol* **20**, 68–80 (2021). URL `https://www.ncbi.nlm.nih.gov/pubmed/33340485`.

[340] Gillette Guyonnet, S. *et al.* IANA (International Academy on Nutrition and Aging) Expert Group: weight loss and Alzheimer's disease. *J Nutr Health Aging* **11**, 38–48 (2007). URL `https://www.ncbi.nlm.nih.gov/pubmed/17315079`.

[341] Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015). URL `https://www.ncbi.nlm.nih.gov/pubmed/25722852`.

[342] Nguyen, X. T. *et al.* Baseline characterization and annual trends of body mass index for a mega-biobank cohort of US Veterans 2011-2017. *J Health Res Rev Dev Ctries* **5**, 98–107 (2018). URL `https://www.ncbi.nlm.nih.gov/pubmed/33117892`.

[343] Hunter-Zinck, H. *et al.* Genotyping array design and data quality control in the Million Veteran Program. *Am J Hum Genet* **106**, 535–548 (2020). URL `https://www.ncbi.nlm.nih.gov/pubmed/32243820`.

[344] Reisberg, S. *et al.* Translating genotype data of 44,000 biobank participants into clinical pharmacogenetic recommendations: challenges and solutions. *Genet Med* **21**, 1345–1354 (2019). URL `https://www.ncbi.nlm.nih.gov/pubmed/30327539`.

[345] Palmer, C. & Pe'er, I. Statistical correction of the winner's curse explains replication variability in quantitative trait genome-wide association studies. *PLoS Genet* **13**, e1006916 (2017). URL `https://www.ncbi.nlm.nih.gov/pubmed/28715421`.

[346] Ferguson, J. P., Cho, J. H., Yang, C. & Zhao, H. Empirical Bayes correction for the winner's curse in genetic association studies. *Genet Epidemiol* **37**, 60–8 (2013). URL `https://www.ncbi.nlm.nih.gov/pubmed/23012258`.

[347] Bezanson, J., Edelman, A., Karpinski, S. & Shah, V. B. Julia: A fresh approach to numerical computing. *SIAM Review* **59**, 65–98 (2017). URL https://doi.org/10.1137/141000671.

[348] Pirastu, N. *et al.* Genetic analyses identify widespread sex-differential participation bias. *Nat Genet* **53**, 663–671 (2021). URL https://www.ncbi.nlm.nih.gov/pubmed/33888908.

[349] Health Survey for England. http://healthsurvey.hscic.gov.uk/data-visualisation/data-visualisation/explore-the-trends/weight/adult/bmi.aspx. Accessed: 2023-06-09.

[350] Krude, H. & Gruters, A. Implications of proopiomelanocortin (POMC) mutations in humans: the POMC deficiency syndrome. *Trends Endocrinol Metab* **11**, 15–22 (2000). URL https://www.ncbi.nlm.nih.gov/pubmed/10652501.

[351] Kerimov, N. *et al.* A compendium of uniformly processed human gene expression and splicing quantitative trait loci. *Nat Genet* **53**, 1290–1299 (2021). URL https://www.ncbi.nlm.nih.gov/pubmed/34493866.

[352] Saeed, S. *et al.* Loss-of-function mutations in ADCY3 cause monogenic severe obesity. *Nat Genet* **50**, 175–179 (2018). URL https://www.ncbi.nlm.nih.gov/pubmed/29311637.

[353] Barton, A. R., Sherman, M. A., Mukamel, R. E. & Loh, P. R. Whole-exome imputation within UK Biobank powers rare coding variant association and fine-mapping analyses. *Nat Genet* **53**, 1260–1269 (2021). URL https://www.ncbi.nlm.nih.gov/pubmed/34226706.

[354] Neale Lab UK Biobank GWAS summary statistics. http://www.nealelab.is/uk-biobank/. Accessed: 2020-08-13.

[355] Donertas, H. M., Fabian, D. K., Valenzuela, M. F., Partridge, L. & Thornton, J. M. Common genetic associations between age-related diseases. *Nat Aging* **1**, 400–412 (2021). URL https://www.ncbi.nlm.nih.gov/pubmed/33959723.

[356] Munafo, M. R., Tilling, K., Taylor, A. E., Evans, D. M. & Davey Smith, G. Collider scope: when selection bias can substantially influence observed associations. *Int J Epidemiol* **47**, 226–235 (2018). URL https://www.ncbi.nlm.nih.gov/pubmed/29040562.

[357] Bland, J. M. & Altman, D. G. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* **1**, 307–10 (1986). URL https://www.ncbi.nlm.nih.gov/pubmed/2868172.

[358] Seidell, J. C. & Visscher, T. L. Body weight and weight change and their health implications for the elderly. *Eur J Clin Nutr* **54 Suppl 3**, S33–9 (2000). URL https://www.ncbi.nlm.nih.gov/pubmed/11041073.

[359] Cao, X. *et al.* Weight change across adulthood and accelerated biological aging in middle-aged and older adults. *Am J Clin Nutr* **117**, 1–11 (2023). URL https://www.ncbi.nlm.nih.gov/pubmed/36789928.

[360] Stienstra, R., Duval, C., Muller, M. & Kersten, S. PPARs, obesity, and inflammation. *PPAR Res* **2007**, 95974 (2007). URL https://www.ncbi.nlm.nih.gov/pubmed/17389767.

[361] Beighley, J. S. *et al.* Clinical phenotypes of carriers of mutations in CHD8 or its conserved target genes. *Biol Psychiatry* **87**, 123–131 (2020). URL https://www.ncbi.nlm.nih.gov/pubmed/31526516.

[362] Zhang, Y. *et al.* SWELL1 is a regulator of adipocyte size, insulin signalling and glucose homeostasis. *Nat Cell Biol* **19**, 504–517 (2017). URL https://www.ncbi.nlm.nih.gov/pubmed/28436964.

[363] Bluher, M. Obesity: global epidemiology and pathogenesis. *Nat Rev Endocrinol* **15**, 288–298 (2019). URL https://www.ncbi.nlm.nih.gov/pubmed/30814686.

[364] Collaborators, G. B. D. O. *et al.* Health effects of overweight and obesity in 195 countries over 25 years. *N Engl J Med* **377**, 13–27 (2017). URL https://www.ncbi.nlm.nih.gov/pubmed/28604169.

[365] Liou, T. H. *et al.* ESR1, FTO, and UCP2 genes interact with bariatric surgery affecting weight loss and glycemic control in severely obese patients. *Obes Surg* **21**, 1758–65 (2011). URL https://www.ncbi.nlm.nih.gov/pubmed/21720911.

[366] Sarzynski, M. A. *et al.* Associations of markers in 11 obesity candidate genes with maximal weight loss and weight regain in the SOS bariatric surgery cases. *Int J Obes (Lond)* **35**, 676–83 (2011). URL https://www.ncbi.nlm.nih.gov/pubmed/20733583.

[367] Delahanty, L. M. *et al.* Genetic predictors of weight loss and weight regain after intensive lifestyle modification, metformin treatment, or standard care in the Diabetes Prevention Program. *Diabetes Care* **35**, 363–6 (2012). URL https://www.ncbi.nlm.nih.gov/pubmed/22179955.

[368] Zhang, X. *et al.* FTO genotype and 2-year change in body composition and fat distribution in response to weight-loss diets: the POUNDS LOST Trial. *Diabetes* **61**, 3005–11 (2012). URL https://www.ncbi.nlm.nih.gov/pubmed/22891219.

[369] McCaffery, J. M. *et al.* Genetic predictors of change in waist circumference and waist-to-hip ratio with lifestyle intervention: The Trans-NIH Consortium for genetics of weight loss response to lifestyle intervention. *Diabetes* **71**, 669–676 (2022). URL https://www.ncbi.nlm.nih.gov/pubmed/35043141.

[370] Papandonatos, G. D. *et al.* Genetic predisposition to weight loss and regain with lifestyle intervention: Analyses from the Diabetes Prevention Program and the Look AHEAD randomized controlled trials. *Diabetes* **64**, 4312–21 (2015). URL https://www.ncbi.nlm.nih.gov/pubmed/26253612.

[371] Holzapfel, C. *et al.* Association between single nucleotide polymorphisms and weight reduction in behavioural interventions-a pooled analysis. *Nutrients* **13** (2021). URL https://www.ncbi.nlm.nih.gov/pubmed/33801339.

[372] Li, H. *et al.* Triglyceride-glucose index variability and incident cardiovascular disease: a prospective cohort study. *Cardiovasc Diabetol* **21**, 105 (2022). URL https://www.ncbi.nlm.nih.gov/pubmed/35689232.

[373] Nuyujukian, D. S. *et al.* Blood pressure variability and risk of heart failure in ACCORD and the VADT. *Diabetes Care* **43**, 1471–1478 (2020). URL https://www.ncbi.nlm.nih.gov/pubmed/32327422.

[374] Speakman, J. R. *et al.* Set points, settling points and some alternative models: theoretical options to understand how genes and environments combine to regulate body adiposity. *Dis Model Mech* **4**, 733–45 (2011). URL https://www.ncbi.nlm.nih.gov/pubmed/22065844.

[375] Muller, M. J., Geisler, C., Heymsfield, S. B. & Bosy-Westphal, A. Recent advances in understanding body weight homeostasis in humans. *F1000Res* **7** (2018). URL https://www.ncbi.nlm.nih.gov/pubmed/30026913.

[376] Nawaz, H., Chan, W., Abdulrahman, M., Larson, D. & Katz, D. L. Self-reported weight and height: implications for obesity research. *Am J Prev Med* **20**, 294–8 (2001). URL https://www.ncbi.nlm.nih.gov/pubmed/11331120.

[377] Stefanska, A., Bergmann, K. & Sypniewska, G. Metabolic syndrome and menopause: Pathophysiology, clinical and diagnostic significance. *Adv Clin Chem* **72**, 1–75 (2015). URL https://www.ncbi.nlm.nih.gov/pubmed/26471080.

[378] Kelly, D. M. & Jones, T. H. Testosterone and obesity. *Obes Rev* **16**, 581–606 (2015). URL https://www.ncbi.nlm.nih.gov/pubmed/25982085.

[379] Nelson, L. R. & Bulun, S. E. Estrogen production and action. *J Am Acad Dermatol* **45**, S116–24 (2001). URL https://www.ncbi.nlm.nih.gov/pubmed/11511861.

[380] Kowal, R. C., Herz, J., Goldstein, J. L., Esser, V. & Brown, M. S. Low density lipoprotein receptor-related protein mediates uptake of cholesteryl esters derived from apoprotein E-enriched lipoproteins. *Proc Natl Acad Sci U S A* **86**, 5810–4 (1989). URL https://www.ncbi.nlm.nih.gov/pubmed/2762297.

[381] Kockx, M., Traini, M. & Kritharides, L. Cell-specific production, secretion, and function of apolipoprotein E. *J Mol Med (Berl)* **96**, 361–371 (2018). URL https://www.ncbi.nlm.nih.gov/pubmed/29516132.

[382] Lumsden, A. L., Mulugeta, A., Zhou, A. & Hypponen, E. Apolipoprotein E (APOE) genotype-associated disease risks: a phenome-wide, registry-based, case-control study utilising the UK Biobank. *EBioMedicine* **59**, 102954 (2020). URL https://www.ncbi.nlm.nih.gov/pubmed/32818802.

[383] Garrison, R. J. *et al.* Obesity and lipoprotein cholesterol in the Framingham offspring study. *Metabolism* **29**, 1053–60 (1980). URL https://www.ncbi.nlm.nih.gov/pubmed/7432169.

[384] Albrink, M. J. *et al.* Intercorrelations among plasma high density lipoprotein, obesity and triglycerides in a normal population. *Lipids* **15**, 668–76 (1980). URL https://www.ncbi.nlm.nih.gov/pubmed/7421421.

[385] Panagiotakos, D. B., Pitsavos, C., Yannakoulia, M., Chrysohoou, C. & Stefanadis, C. The implication of obesity and central fat on markers of chronic inflammation: The ATTICA study. *Atherosclerosis* **183**, 308–15 (2005). URL https://www.ncbi.nlm.nih.gov/pubmed/16285994.

[386] von Hardenberg, S., Gnewuch, C., Schmitz, G. & Borlak, J. ApoE is a major determinant of hepatic bile acid homeostasis in mice. *J Nutr Biochem* **52**, 82–91 (2018). URL https://www.ncbi.nlm.nih.gov/pubmed/29175670.

[387] Wang, J. *et al.* ApoE and the role of very low density lipoproteins in adipose tissue inflammation. *Atherosclerosis* **223**, 342–9 (2012). URL `https://www.ncbi.nlm.nih.gov/pubmed/22770993`.

[388] Blanchard, J. W. *et al.* APOE4 impairs myelination via cholesterol dysregulation in oligodendrocytes. *Nature* **611**, 769–779 (2022). URL `https://www.ncbi.nlm.nih.gov/pubmed/36385529`.

[389] Goudie, R. J. B., Presanis, A. M., Lunn, D., De Angelis, D. & Wernisch, L. Joining and splitting models with Markov melding. *Bayesian Anal* **14**, 81–109 (2019). URL `https://www.ncbi.nlm.nih.gov/pubmed/30631389`.

[390] Conroy, M. *et al.* The advantages of UK Biobank's open-access strategy for health research. *J Intern Med* **286**, 389–397 (2019). URL `https://www.ncbi.nlm.nih.gov/pubmed/31283063`.

[391] Martin, A. R. *et al.* Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat Genet* **51**, 584–591 (2019). URL `https://www.ncbi.nlm.nih.gov/pubmed/30926966`.

[392] Carreras-Torres, R. *et al.* Role of obesity in smoking behaviour: Mendelian randomisation study in UK Biobank. *BMJ* **361**, k1767 (2018). URL `https://www.ncbi.nlm.nih.gov/pubmed/29769355`.

[393] Yerevanian, A. & Soukas, A. A. Metformin: Mechanisms in human obesity and weight loss. *Curr Obes Rep* **8**, 156–164 (2019). URL `https://www.ncbi.nlm.nih.gov/pubmed/30874963`.

[394] Del Bosque-Plata, L., Martinez-Martinez, E., Espinoza-Camacho, M. A. & Gragnoli, C. The role of TCF7L2 in type 2 diabetes. *Diabetes* **70**, 1220–1228 (2021). URL `https://www.ncbi.nlm.nih.gov/pubmed/34016596`.

[395] Lazzaroni, E. *et al.* Anti-diabetic drugs and weight loss in patients with type 2 diabetes. *Pharmacol Res* **171**, 105782 (2021). URL `https://www.ncbi.nlm.nih.gov/pubmed/34302978`.

[396] Wang, J. Y. *et al.* GLP-1 receptor agonists for the treatment of obesity: Role as a promising approach. *Front Endocrinol (Lausanne)* **14**, 1085799 (2023). URL `https://www.ncbi.nlm.nih.gov/pubmed/36843578`.

[397] Eberly, L. A. *et al.* Racial, ethnic, and socioeconomic inequities in glucagon-like peptide-1 receptor agonist use among patients with diabetes in the US. *JAMA Health Forum* **2**, e214182 (2021). URL `https://www.ncbi.nlm.nih.gov/pubmed/35977298`.

[398] Borghi, C., Veronesi, M., Prandin, M. G., Dormi, A. & Ambrosioni, E. Statins and blood pressure regulation. *Curr Hypertens Rep* **3**, 281–8 (2001). URL `https://www.ncbi.nlm.nih.gov/pubmed/11470010`.

[399] Singh, P. *et al.* Statins decrease leptin expression in human white adipocytes. *Physiol Rep* **6** (2018). URL `https://www.ncbi.nlm.nih.gov/pubmed/29372612`.

[400] McCarron, D. A. & Reusser, M. E. Body weight and blood pressure regulation. *Am J Clin Nutr* **63**, 423S–425S (1996). URL `https://www.ncbi.nlm.nih.gov/pubmed/8615333`.

[401] Kutcher, S. A., Brophy, J. M., Banack, H. R., Kaufman, J. S. & Samuel, M. Emulating a randomised controlled trial with observational data: An introduction to the target trial framework. *Can J Cardiol* **37**, 1365–1377 (2021). URL https://www.ncbi.nlm.nih.gov/pubmed/34090982.

[402] Shortreed, S. M., Rutter, C. M., Cook, A. J. & Simon, G. E. Improving pragmatic clinical trial design using real-world data. *Clin Trials* **16**, 273–282 (2019). URL https://www.ncbi.nlm.nih.gov/pubmed/30866672.

[403] Team, W. Infertility Prevalence Estimates, 1990–2021. https://www.who.int/publications/i/item/978920068315 (2023). Accessed: 2023-11-19.

[404] Brandt, J. S., Cruz Ithier, M. A., Rosen, T. & Ashkinadze, E. Advanced paternal age, infertility, and reproductive risks: A review of the literature. *Prenat Diagn* **39**, 81–87 (2019). URL https://www.ncbi.nlm.nih.gov/pubmed/30520056.

[405] Crawford, N. M. & Steiner, A. Z. Age-related infertility. *Obstet Gynecol Clin North Am* **42**, 15–25 (2015). URL https://www.ncbi.nlm.nih.gov/pubmed/25681837.

[406] Tjahyadi, D. *et al.* Female genital tuberculosis: Clinical presentation, current diagnosis, and treatment. *Infect Dis Obstet Gynecol* **2022**, 3548190 (2022). URL https://www.ncbi.nlm.nih.gov/pubmed/36438172.

[407] Passos, L. G. *et al.* The correlation between chlamydia trachomatis and female infertility: A systematic review. *Rev Bras Ginecol Obstet* **44**, 614–620 (2022). URL https://www.ncbi.nlm.nih.gov/pubmed/35576969.

[408] Tsevat, D. G., Wiesenfeld, H. C., Parks, C. & Peipert, J. F. Sexually transmitted diseases and infertility. *Am J Obstet Gynecol* **216**, 1–9 (2017). URL https://www.ncbi.nlm.nih.gov/pubmed/28007229.

[409] Shteinberg, M., Taylor-Cousar, J. L., Durieu, I. & Cohen-Cymberknoh, M. Fertility and pregnancy in cystic fibrosis. *Chest* **160**, 2051–2060 (2021). URL https://www.ncbi.nlm.nih.gov/pubmed/34284004.

[410] Jarzabek, K. *et al.* Cystic fibrosis as a cause of infertility. *Reprod Biol* **4**, 119–29 (2004). URL https://www.ncbi.nlm.nih.gov/pubmed/15297887.

[411] Krausz, C. & Riera-Escamilla, A. Genetics of male infertility. *Nat Rev Urol* **15**, 369–384 (2018). URL https://www.ncbi.nlm.nih.gov/pubmed/29622783.

[412] Matzuk, M. M. & Lamb, D. J. The biology of infertility: research advances and clinical challenges. *Nat Med* **14**, 1197–213 (2008). URL https://www.ncbi.nlm.nih.gov/pubmed/18989307.

[413] Hoeger, K. M., Dokras, A. & Piltonen, T. Update on PCOS: Consequences, challenges, and guiding treatment. *J Clin Endocrinol Metab* **106**, e1071–e1083 (2021). URL https://www.ncbi.nlm.nih.gov/pubmed/33211867.

[414] Vannuccini, S. *et al.* Infertility and reproductive disorders: impact of hormonal and inflammatory mechanisms on pregnancy outcome. *Hum Reprod Update* **22**, 104–15 (2016). URL https://www.ncbi.nlm.nih.gov/pubmed/26395640.

[415] Concepcion-Zavaleta, M. *et al.* Assessment of hormonal status in male infertility. an update. *Diabetes Metab Syndr* **16**, 102447 (2022). URL https://www.ncbi.nlm.nih.gov/pubmed/35272174.

[416] Silvestris, E., de Pergola, G., Rosania, R. & Loverro, G. Obesity as disruptor of the female fertility. *Reprod Biol Endocrinol* **16**, 22 (2018). URL https://www.ncbi.nlm.nih.gov/pubmed/29523133.

[417] Mariotti, S. & Beck-Peccoz, P. Physiology of the hypothalamic-pituitary-thyroid axis. In Feingold, K. R. *et al.* (eds.) *Endotext* (MDText, South Dartmouth (MA), 2000). URL https://www.ncbi.nlm.nih.gov/pubmed/25905193.

[418] Alexandraki, K. I., Sanpawithayakul, K. & Grossman, A. Adrenal insufficiency. In Feingold, K. R. *et al.* (eds.) *Endotext* (MDText, South Dartmouth (MA), 2000). URL https://www.ncbi.nlm.nih.gov/pubmed/25905309.

[419] Achermann, J. C. & Jameson, J. L. Fertility and infertility: genetic contributions from the hypothalamic-pituitary-gonadal axis. *Mol Endocrinol* **13**, 812–8 (1999). URL https://www.ncbi.nlm.nih.gov/pubmed/10379880.

[420] Layman, L. C. Hypogonadotropic hypogonadism. *Endocrinol Metab Clin North Am* **36**, 283–96 (2007). URL https://www.ncbi.nlm.nih.gov/pubmed/17543719.

[421] Gaitonde, D. Y., Rowley, K. D. & Sweeney, L. B. Hypothyroidism: an update. *Am Fam Physician* **86**, 244–51 (2012). URL https://www.ncbi.nlm.nih.gov/pubmed/22962987.

[422] De Vos, M., Devroey, P. & Fauser, B. C. Primary ovarian insufficiency. *Lancet* **376**, 911–21 (2010). URL https://www.ncbi.nlm.nih.gov/pubmed/20708256.

[423] Leeners, B., Geary, N., Tobler, P. N. & Asarian, L. Ovarian hormones and obesity. *Hum Reprod Update* **23**, 300–321 (2017). URL https://www.ncbi.nlm.nih.gov/pubmed/28333235.

[424] Kopelman, P. G. Hormones and obesity. *Baillieres Clin Endocrinol Metab* **8**, 549–75 (1994). URL https://www.ncbi.nlm.nih.gov/pubmed/7980347.

[425] Sorensen, E. *et al.* Data Resource Profile: The Copenhagen Hospital Biobank (CHB). *Int J Epidemiol* **50**, 719–720e (2021). URL https://www.ncbi.nlm.nih.gov/pubmed/33169150.

[426] Finer, S. *et al.* Cohort Profile: East London Genes & Health (ELGH), a community-based population genomics and health study in British Bangladeshi and British Pakistani people. *Int J Epidemiol* **49**, 20–21i (2020). URL https://www.ncbi.nlm.nih.gov/pubmed/31504546.

[427] Dennis, J. K. *et al.* Clinical laboratory test-wide association scan of polygenic scores identifies biomarkers of complex disease. *Genome Med* **13**, 6 (2021). URL https://www.ncbi.nlm.nih.gov/pubmed/33441150.

[428] Suhre, K. *et al.* Connecting genetic risk to disease end points through the human blood plasma proteome. *Nat Commun* **8**, 14357 (2017). URL https://www.ncbi.nlm.nih.gov/pubmed/28240269.

[429] Thareja, G. *et al.* Differences and commonalities in the genetic architecture of protein quantitative trait loci in European and Arab populations. *Hum Mol Genet* **32**, 907–916 (2023). URL https://www.ncbi.nlm.nih.gov/pubmed/36168886.

[430] Prins, B. P. *et al.* Genome-wide analysis of health-related biomarkers in the UK Household Longitudinal Study reveals novel associations. *Sci Rep* **7**, 11008 (2017). URL https://www.ncbi.nlm.nih.gov/pubmed/28887542.

[431] Martin, F. J. *et al.* Ensembl 2023. *Nucleic Acids Res* **51**, D933–D941 (2023). URL https://www.ncbi.nlm.nih.gov/pubmed/36318249.

[432] Kuhn, R. M., Haussler, D. & Kent, W. J. The UCSC genome browser and associated tools. *Brief Bioinform* **14**, 144–61 (2013). URL https://www.ncbi.nlm.nih.gov/pubmed/22908213.

[433] Williams, A. T. *et al.* Genome-wide association study of thyroid-stimulating hormone highlights new genes, pathways and associations with thyroid disease. *Nat Commun* **14**, 6713 (2023). URL https://www.ncbi.nlm.nih.gov/pubmed/37872160.

[434] Verdiesen, R. M. G. *et al.* Genome-wide association study meta-analysis identifies three novel loci for circulating anti-Mullerian hormone levels in women. *Hum Reprod* **37**, 1069–1082 (2022). URL https://www.ncbi.nlm.nih.gov/pubmed/35274129.

[435] Mbarek, H. *et al.* Genome-wide association study meta-analysis of dizygotic twinning illuminates genetic regulation of female fecundity. *Hum Reprod* (2023). URL https://www.ncbi.nlm.nih.gov/pubmed/38052102.

[436] Giambartolomei, C. & et. al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLOS Genetics* (2014). URL https://doi.org/10.1371/journal.pgen.1004383.

[437] Wallace, C. Eliciting priors and relaxing the single causal variant assumption in colocalisation analyses. *PLOS Genetics* (2020). URL https://doi.org/10.1371/journal.pgen.1008720.

[438] Grossman, S. R. *et al.* Identifying recent adaptations in large-scale genomic data. *Cell* **152**, 703–13 (2013). URL https://www.ncbi.nlm.nih.gov/pubmed/23415221.

[439] Mathieson, I. *et al.* Genome-wide patterns of selection in 230 ancient Eurasians. *Nature* **528**, 499–503 (2015). URL https://www.ncbi.nlm.nih.gov/pubmed/26595274.

[440] Field, Y. *et al.* Detection of human adaptation during the past 2000 years. *Science* **354**, 760–764 (2016). URL https://www.ncbi.nlm.nih.gov/pubmed/27738015.

[441] Siewert, K. M. & Voight, B. F. BetaScan2: Standardized statistics to detect balancing selection utilizing substitution data. *Genome Biol Evol* **12**, 3873–3877 (2020). URL https://www.ncbi.nlm.nih.gov/pubmed/32011695.

[442] Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat Genet* **47**, 1228–35 (2015). URL https://www.ncbi.nlm.nih.gov/pubmed/26414678.

[443] Pers, T. H. *et al.* Biological interpretation of genome-wide association studies using predicted gene functions. *Nat Commun* **6**, 5890 (2015). URL https://www.ncbi.nlm.nih.gov/pubmed/25597830.

[444] Finucane, H. K. *et al.* Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nat Genet* **50**, 621–629 (2018). URL https://www.ncbi.nlm.nih.gov/pubmed/29632380.

[445] Fan, X. *et al.* Single-cell reconstruction of follicular remodeling in the human adult ovary. *Nat Commun* **10**, 3164 (2019). URL https://www.ncbi.nlm.nih.gov/pubmed/31320652.

[446] Jin, C. *et al.* The regulatory landscapes of human ovarian ageing. *bioRxiv* 2022.05.18.492547 (2022). URL https://www.biorxiv.org/content/biorxiv/early/2022/05/19/2022.05.18.492547.full.pdf.

[447] Stuart, T. *et al.* Comprehensive integration of single-cell data. *Cell* **177**, 1888–1902 (2019). URL https://doi.org/10.1016/j.cell.2019.05.031.

[448] Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology* **36**, 411–420 (2018). URL https://doi.org/10.1038/nbt.4096.

[449] Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nature Biotechnology* **33**, 495–502 (2015). URL https://doi.org/10.1038/nbt.3192.

[450] Ghoussaini, M. *et al.* Open Targets Genetics: systematic identification of trait-associated genes using large-scale genetics and functional genomics. *Nucleic Acids Res* **49**, D1311–D1320 (2021). URL https://www.ncbi.nlm.nih.gov/pubmed/33045747.

[451] Ruder, C. *et al.* The tumor-associated antigen EBAG9 negatively regulates the cytolytic capacity of mouse CD8+ T cells. *J Clin Invest* **119**, 2184–203 (2009). URL https://www.ncbi.nlm.nih.gov/pubmed/19620783.

[452] Basta, P. *et al.* The immunohistochemical analysis of RCAS1, HLA-G, and B7H4-positive macrophages in partial and complete hydatidiform mole in both applied therapeutic surgery and surgery followed by chemotherapy. *Am J Reprod Immunol* **65**, 164–72 (2011). URL https://www.ncbi.nlm.nih.gov/pubmed/20645939.

[453] Akakura, S., Huang, C., Nelson, P. J., Foster, B. & Gelman, I. H. Loss of the SSeCKS/Gravin/AKAP12 gene results in prostatic hyperplasia. *Cancer Res* **68**, 5096–103 (2008). URL https://www.ncbi.nlm.nih.gov/pubmed/18593908.

[454] Kang, Z., Bai, Y., Lan, X. & Zhao, H. Goat AKAP12: Indel mutation detection, association analysis with litter size and alternative splicing variant expression. *Front Genet* **12**, 648256 (2021). URL https://www.ncbi.nlm.nih.gov/pubmed/34093646.

[455] Wilber, J. F. & Xu, A. H. The thyrotropin-releasing hormone gene 1998: cloning, characterization, and transcriptional regulation in the central nervous system, heart, and testis. *Thyroid* **8**, 897–901 (1998). URL https://www.ncbi.nlm.nih.gov/pubmed/9827656.

[456] Nawaz, S. *et al.* A variant in sperm-specific glycolytic enzyme enolase 4 (ENO4) causes human male infertility. *J Gene Med* e3583 (2023). URL https://www.ncbi.nlm.nih.gov/pubmed/37640479.

[457] Nakamura, N. *et al.* Disruption of a spermatogenic cell-specific mouse enolase 4 (eno4) gene causes sperm structural defects and male infertility. *Biol Reprod* **88**, 90 (2013). URL https://www.ncbi.nlm.nih.gov/pubmed/23446454.

[458] Li, J. *et al.* Aberrant spliceosome expression and altered alternative splicing events correlate with maturation deficiency in human oocytes. *Cell Cycle* **19**, 2182–2194 (2020). URL https://www.ncbi.nlm.nih.gov/pubmed/32779509.

[459] Kichaev, G. *et al.* Leveraging polygenic functional enrichment to improve GWAS power. *Am J Hum Genet* **104**, 65–75 (2019). URL https://www.ncbi.nlm.nih.gov/pubmed/30595370.

[460] Sakaue, S. *et al.* A cross-population atlas of genetic associations for 220 human phenotypes. *Nat Genet* **53**, 1415–1424 (2021). URL https://www.ncbi.nlm.nih.gov/pubmed/34594039.

[461] Langner, C. A., Birkenmeier, E. H., Roth, K. A., Bronson, R. T. & Gordon, J. I. Characterization of the peripheral neuropathy in neonatal and adult mice that are homozygous for the fatty liver dystrophy (fld) mutation. *J Biol Chem* **266**, 11955–64 (1991). URL https://www.ncbi.nlm.nih.gov/pubmed/2050689.

[462] Valimaki, N. *et al.* Genetic predisposition to uterine leiomyoma is determined by loci for genitourinary development and genome stability. *Elife* **7** (2018). URL https://www.ncbi.nlm.nih.gov/pubmed/30226466.

[463] Rafnar, T. *et al.* Variants associating with uterine leiomyoma highlight genetic background shared by various cancers and hormone-related traits. *Nat Commun* **9**, 3636 (2018). URL https://www.ncbi.nlm.nih.gov/pubmed/30194396.

[464] Luong, H. T. *et al.* Fine mapping of variants associated with endometriosis in the WNT4 region on chromosome 1p36. *Int J Mol Epidemiol Genet* **4**, 193–206 (2013). URL https://www.ncbi.nlm.nih.gov/pubmed/24319535.

[465] Pellegrino, M., Maiorino, R. & Schonauer, S. WNT4 signaling in female gonadal development. *Endocr Metab Immune Disord Drug Targets* **10**, 168–74 (2010). URL https://www.ncbi.nlm.nih.gov/pubmed/20350288.

[466] Bernard, P. & Harley, V. R. Wnt4 action in gonadal development and sex determination. *Int J Biochem Cell Biol* **39**, 31–43 (2007). URL https://www.ncbi.nlm.nih.gov/pubmed/16905353.

[467] Pitzer, L. M., Moroney, M. R., Nokoff, N. J. & Sikora, M. J. WNT4 balances development vs disease in gynecologic tissues and women's health. *Endocrinology* **162** (2021). URL https://www.ncbi.nlm.nih.gov/pubmed/33963381.

[468] Lewontin, R. C. Polymorphism and heterosis: old wine in new bottles and vice versa. *Journal of the History of Biology* 337–349 (1987).

[469] Dobzhansky, T. A review of some fundamental concepts and problems of population genetics. *Cold Spring Harb Symp Quant Biol* **20**, 1–15 (1955). URL https://www.ncbi.nlm.nih.gov/pubmed/13433550.

[470] Abbaci, A. *et al.* Neurotensin receptor type 2 protects B-cell chronic lymphocytic leukemia cells from apoptosis. *Oncogene* **37**, 756–767 (2018). URL https://www.ncbi.nlm.nih.gov/pubmed/29059151.

[471] Eisen, D. P. & Osthoff, M. If there is an evolutionary selection pressure for the high frequency of MBL2 polymorphisms, what is it? *Clin Exp Immunol* **176**, 165–71 (2014). URL https://www.ncbi.nlm.nih.gov/pubmed/24255984.

[472] Li, J. *et al.* Neurotensin is an anti-thermogenic peptide produced by lymphatic endothelial cells. *Cell Metab* **33**, 1449–1465 e6 (2021). URL https://www.ncbi.nlm.nih.gov/pubmed/34038712.

[473] Maurizi, G., Petaisto, T., Maurizi, A. & Della Guardia, L. Key-genes regulating the liposecretion process of mature adipocytes. *J Cell Physiol* **233**, 3784–3793 (2018). URL https://www.ncbi.nlm.nih.gov/pubmed/28926092.

[474] Su, W. *et al.* Role of HSD17B13 in the liver physiology and pathophysiology. *Mol Cell Endocrinol* **489**, 119–125 (2019). URL https://www.ncbi.nlm.nih.gov/pubmed/30365983.

[475] Rytinki, M. *et al.* Dynamic sumoylation is linked to the activity cycles of androgen receptor in the cell nucleus. *Mol Cell Biol* **32**, 4195–205 (2012). URL https://www.ncbi.nlm.nih.gov/pubmed/22890844.

[476] Terashima, R., Tani, T., Sakakibara, K., Kurusu, S. & Kawaminami, M. Thyrotropin-releasing hormone stimulates NR4A3 expression in the pituitary thyrotrophs of proestrus rats. *Endocr J* **70**, 805–814 (2023). URL https://www.ncbi.nlm.nih.gov/pubmed/37211401.

[477] Fernandez, P. M. *et al.* Nuclear receptors Nor1 and NGFI-B/Nur77 play similar, albeit distinct, roles in the hypothalamo-pituitary-adrenal axis. *Endocrinology* **141**, 2392–400 (2000). URL https://www.ncbi.nlm.nih.gov/pubmed/10875239.

[478] Zhao, Y. & Bruemmer, D. NR4A orphan nuclear receptors: transcriptional regulators of gene expression in metabolism and vascular biology. *Arterioscler Thromb Vasc Biol* **30**, 1535–41 (2010). URL https://www.ncbi.nlm.nih.gov/pubmed/20631354.

[479] Groza, T. *et al.* The International Mouse Phenotyping Consortium: comprehensive knockout phenotyping underpinning the study of human disease. *Nucleic Acids Res* **51**, D1038–D1045 (2023). URL https://www.ncbi.nlm.nih.gov/pubmed/36305825.

[480] Nguidjoe, E. *et al.* Heterozygous inactivation of the Na/Ca exchanger increases glucose-induced insulin release, beta-cell proliferation, and mass. *Diabetes* **60**, 2076–85 (2011). URL https://www.ncbi.nlm.nih.gov/pubmed/21659499.

[481] Rzeszowska, M. *et al.* Anti-Mullerian hormone: structure, properties and appliance. *Ginekol Pol* **87**, 669–674 (2016). URL https://www.ncbi.nlm.nih.gov/pubmed/27723076.

[482] Ishikawa, K. *et al.* An autosomal dominant cerebellar ataxia linked to chromosome 16q22.1 is associated with a single-nucleotide substitution in the 5' untranslated region of the gene encoding a protein with spectrin repeat and Rho guanine-nucleotide exchange-factor domains. *Am J Hum Genet* **77**, 280–96 (2005). URL https://www.ncbi.nlm.nih.gov/pubmed/16001362.

[483] Seminara, S. B., Acierno, J., J. S., Abdulwahid, N. A., Crowley, J., W. F. & Margolin, D. H. Hypogonadotropic hypogonadism and cerebellar ataxia: detailed phenotypic characterization of a large, extended kindred. *J Clin Endocrinol Metab* **87**, 1607–12 (2002). URL https://www.ncbi.nlm.nih.gov/pubmed/11932290.

[484] Holmes, G. A form of familial degeneration of the cerebellum. *Brain* **30**, 466–489 (1908).

[485] Durocher, F., Morissette, J., Labrie, Y., Labrie, F. & Simard, J. Mapping of the HSD17B2 gene encoding type II 17 beta-hydroxysteroid dehydrogenase close to D16S422 on chromosome 16q24.1-q24.2. *Genomics* **25**, 724–6 (1995). URL https://www.ncbi.nlm.nih.gov/pubmed/7759109.

[486] Saloniemi, T., Jokela, H., Strauss, L., Pakarinen, P. & Poutanen, M. The diversity of sex steroid action: novel functions of hydroxysteroid (17beta) dehydrogenases as revealed by genetically modified mouse models. *J Endocrinol* **212**, 27–40 (2012). URL https://www.ncbi.nlm.nih.gov/pubmed/22045753.

[487] Lawson, A. J. *et al.* Cortisone-reductase deficiency associated with heterozygous mutations in 11beta-hydroxysteroid dehydrogenase type 1. *Proc Natl Acad Sci U S A* **108**, 4111–6 (2011). URL https://www.ncbi.nlm.nih.gov/pubmed/21325058.

[488] Draper, N. *et al.* Mutations in the genes encoding 11beta-hydroxysteroid dehydrogenase type 1 and hexose-6-phosphate dehydrogenase interact to cause cortisone reductase deficiency. *Nat Genet* **34**, 434–9 (2003). URL https://www.ncbi.nlm.nih.gov/pubmed/12858176.

[489] Le Quesne Stabej, P. *et al.* STAG3 truncating variant as the cause of primary ovarian insufficiency. *Eur J Hum Genet* **24**, 135–8 (2016). URL https://www.ncbi.nlm.nih.gov/pubmed/26059840.

[490] Caburet, S. *et al.* Mutant cohesin in premature ovarian failure. *N Engl J Med* **370**, 943–949 (2014). URL https://www.ncbi.nlm.nih.gov/pubmed/24597867.

[491] Tardif, S. *et al.* Zonadhesin is essential for species specificity of sperm adhesion to the egg zona pellucida. *J Biol Chem* **285**, 24863–70 (2010). URL https://www.ncbi.nlm.nih.gov/pubmed/20529856.

[492] Bonavina, G. & Taylor, H. S. Endometriosis-associated infertility: From pathophysiology to tailored treatment. *Front Endocrinol (Lausanne)* **13**, 1020827 (2022). URL https://www.ncbi.nlm.nih.gov/pubmed/36387918.

[493] de Ziegler, D., Borghese, B. & Chapron, C. Endometriosis and infertility: pathophysiology and management. *Lancet* **376**, 730–8 (2010). URL https://www.ncbi.nlm.nih.gov/pubmed/20801404.

[494] Gambineri, A. *et al.* Female infertility: which role for obesity? *Int J Obes Suppl* **9**, 65–72 (2019). URL https://www.ncbi.nlm.nih.gov/pubmed/31391925.

[495] Maheshwari, A., Stofberg, L. & Bhattacharya, S. Effect of overweight and obesity on assisted reproductive technology–a systematic review. *Hum Reprod Update* **13**, 433–44 (2007). URL https://www.ncbi.nlm.nih.gov/pubmed/17584821.

[496] Calhaz-Jorge, C. *et al.* Survey on ART and IUI: legislation, regulation, funding and registries in European countries: The European IVF-monitoring Consortium (EIM) for the European Society of Human Reproduction and Embryology (ESHRE). *Hum Reprod Open* **2020**, hoz044 (2020). URL https://www.ncbi.nlm.nih.gov/pubmed/32042927.

[497] Gaskins, A. J. Recent advances in understanding the relationship between long- and short-term weight change and fertility. *F1000Res* **7** (2018). URL https://www.ncbi.nlm.nih.gov/pubmed/30416711.

[498] Payne, N., Seenan, S. & van den Akker, O. Experiences of involuntary childlessness and treatment in the UK: what has changed in 20 years? *Hum Fertil (Camb)* **24**, 333–340 (2021). URL https://www.ncbi.nlm.nih.gov/pubmed/31707869.

[499] Kwan, I., Wang, R., Pearce, E. & Bhattacharya, S. Pain relief for women undergoing oocyte retrieval for assisted reproduction. *Cochrane Database Syst Rev* **5**, CD004829 (2018). URL https://www.ncbi.nlm.nih.gov/pubmed/29761478.

[500] Katz, P. *et al.* Costs of infertility treatment: results from an 18-month prospective cohort study. *Fertil Steril* **95**, 915–21 (2011). URL https://www.ncbi.nlm.nih.gov/pubmed/21130988.

[501] Lai, J. D. *et al.* Unmet financial burden of infertility care and the impact of state insurance mandates in the United States: analysis from a popular crowdfunding platform. *Fertil Steril* **116**, 1119–1125 (2021). URL https://www.ncbi.nlm.nih.gov/pubmed/34246467.

[502] Steiner, A. Z., Stanczyk, F. Z., Patel, S. & Edelman, A. Antimullerian hormone and obesity: insights in oral contraceptive users. *Contraception* **81**, 245–8 (2010). URL https://www.ncbi.nlm.nih.gov/pubmed/20159182.

[503] Bedenk, J., Vrtacnik-Bokal, E. & Virant-Klun, I. The role of anti-Mullerian hormone (AMH) in ovarian disease and infertility. *J Assist Reprod Genet* **37**, 89–100 (2020). URL https://www.ncbi.nlm.nih.gov/pubmed/31755000.

[504] Tal, R. *et al.* Characterization of women with elevated antimullerian hormone levels (AMH): correlation of AMH with polycystic ovarian syndrome phenotypes and assisted reproductive technology outcomes. *Am J Obstet Gynecol* **211**, 59 e1–8 (2014). URL https://www.ncbi.nlm.nih.gov/pubmed/24593938.

[505] Crosignani, P. G., Bianchedi, D., Riccaboni, A. & Vegetti, W. Management of anovulatory infertility. *Hum Reprod* **14 Suppl 1**, 108–19 (1999). URL https://www.ncbi.nlm.nih.gov/pubmed/10573028.

[506] Pasman, J. A. *et al.* The CADM2 gene and behavior: A phenome-wide scan in UK-Biobank. *Behav Genet* **52**, 306–314 (2022). URL https://www.ncbi.nlm.nih.gov/pubmed/35867259.

[507] Gunnarsson, B. *et al.* A sequence variant associating with educational attainment also affects childhood cognition. *Sci Rep* **6**, 36189 (2016). URL https://www.ncbi.nlm.nih.gov/pubmed/27811963.

[508] Okbay, A. *et al.* Genome-wide association study identifies 74 loci associated with educational attainment. *Nature* **533**, 539–42 (2016). URL https://www.ncbi.nlm.nih.gov/pubmed/27225129.

[509] Nisbett, R. E. *et al.* Intelligence: new findings and theoretical developments. *Am Psychol* **67**, 130–59 (2012). URL https://www.ncbi.nlm.nih.gov/pubmed/22233090.

[510] Rask-Andersen, M., Karlsson, T., Ek, W. E. & Johansson, A. Modification of heritability for educational attainment and fluid intelligence by socioeconomic deprivation in the UK Biobank. *Am J Psychiatry* **178**, 625–634 (2021). URL https://www.ncbi.nlm.nih.gov/pubmed/33900812.

[511] Balbo, N., Billari, F. C. & Mills, M. Fertility in advanced societies: A review of research: La fecondite dans les societes avancees: un examen des recherches. *Eur J Popul* **29**, 1–38 (2013). URL https://www.ncbi.nlm.nih.gov/pubmed/23440941.

[512] Tchernof, A. *et al.* Androgens and the regulation of adiposity and body fat distribution in humans. *Compr Physiol* **8**, 1253–1290 (2018). URL https://www.ncbi.nlm.nih.gov/pubmed/30215860.

[513] Loh, N. Y. *et al.* Sex hormones, adiposity, and metabolic traits in men and women: a Mendelian randomisation study. *Eur J Endocrinol* **186**, 407–416 (2022). URL https://www.ncbi.nlm.nih.gov/pubmed/35049520.

[514] Palmer, B. F. & Clegg, D. J. The sexual dimorphism of obesity. *Mol Cell Endocrinol* **402**, 113–9 (2015). URL https://www.ncbi.nlm.nih.gov/pubmed/25578600.

[515] Schulze, M. B. & Stefan, N. Genetic predisposition to abdominal adiposity and cardiometabolic risk. *JAMA* **317**, 2334 (2017). URL https://www.ncbi.nlm.nih.gov/pubmed/28609526.

[516] Rehman, K. S. & Carr, B. R. Sex differences in adrenal androgens. *Semin Reprod Med* **22**, 349–60 (2004). URL https://www.ncbi.nlm.nih.gov/pubmed/15635502.

[517] Zirkin, B. R. & Papadopoulos, V. Leydig cells: formation, function, and regulation. *Biol Reprod* **99**, 101–111 (2018). URL https://www.ncbi.nlm.nih.gov/pubmed/29566165.

[518] Burger, H. G. Androgen production in women. *Fertil Steril* **77 Suppl 4**, S3–5 (2002). URL https://www.ncbi.nlm.nih.gov/pubmed/12007895.

[519] Fortune, J. E. & Armstrong, D. T. Androgen production by theca and granulosa isolated from proestrous rat follicles. *Endocrinology* **100**, 1341–7 (1977). URL https://www.ncbi.nlm.nih.gov/pubmed/557406.

[520] Liu, Y. X. & Hsueh, A. J. Synergism between granulosa and theca-interstitial cells in estrogen biosynthesis by gonadotropin-treated rat ovaries: studies on the two-cell, two-gonadotropin hypothesis using steroid antisera. *Biol Reprod* **35**, 27–36 (1986). URL https://www.ncbi.nlm.nih.gov/pubmed/3091103.

[521] Wen, X., Li, D., Tozer, A. J., Docherty, S. M. & Iles, R. K. Estradiol, progesterone, testosterone profiles in human follicular fluid and cultured granulosa cells from luteinized pre-ovulatory follicles. *Reprod Biol Endocrinol* **8**, 117 (2010). URL https://www.ncbi.nlm.nih.gov/pubmed/20937107.

[522] Kinnear, H. M. *et al.* The ovarian stroma as a new frontier. *Reproduction* **160**, R25–R39 (2020). URL https://www.ncbi.nlm.nih.gov/pubmed/32716007.

[523] Samimi, M., Pourhanifeh, M. H., Mehdizadehkashi, A., Eftekhar, T. & Asemi, Z. The role of inflammation, oxidative stress, angiogenesis, and apoptosis in the pathophysiology of endometriosis: Basic science and new insights based on gene expression. *J Cell Physiol* **234**, 19384–19392 (2019). URL https://www.ncbi.nlm.nih.gov/pubmed/31004368.

[524] Zhou, Y., Ding, X. & Wei, H. Reproductive immune microenvironment. *J Reprod Immunol* **152**, 103654 (2022). URL https://www.ncbi.nlm.nih.gov/pubmed/35728349.

[525] Petersen, B. S., Fredrich, B., Hoeppner, M. P., Ellinghaus, D. & Franke, A. Opportunities and challenges of whole-genome and -exome sequencing. *BMC Genet* **18**, 14 (2017). URL https://www.ncbi.nlm.nih.gov/pubmed/28193154.

[526] Park, J. H. *et al.* Distribution of allele frequencies and effect sizes and their interrelationships for common genetic susceptibility variants. *Proc Natl Acad Sci U S A* **108**, 18026–31 (2011). URL https://www.ncbi.nlm.nih.gov/pubmed/22003128.

[527] Li, X. *et al.* Powerful, scalable and resource-efficient meta-analysis of rare variant associations in large whole genome sequencing studies. *Nat Genet* **55**, 154–164 (2023). URL https://www.ncbi.nlm.nih.gov/pubmed/36564505.

[528] Fertilisation, H. & Embryology Authority, U. Fertility treatment 2021: preliminary trends and figures (2023). URL https://www.hfea.gov.uk/about-us/publications/research-and-data/fertility-treatment-2021-preliminary-trends-and-figures/.

[529] Schock, H. *et al.* Hormone concentrations throughout uncomplicated pregnancies: a longitudinal study. *BMC Pregnancy Childbirth* **16**, 146 (2016). URL https://www.ncbi.nlm.nih.gov/pubmed/27377060.

[530] for Health, N. I. & Care Excellence, U. Fertility problems: assessment and treatment. https://www.nice.org.uk/guidance/cg156/ifp/chapter/tests-for-women (2017). Accessed: 2021-04-23.

[531] Agarwal, A., Mulgund, A., Hamada, A. & Chyatte, M. R. A unique view on male infertility around the globe. *Reprod Biol Endocrinol* **13**, 37 (2015). URL https://www.ncbi.nlm.nih.gov/pubmed/25928197.

[532] Wei, S., Schmidt, M. D., Dwyer, T., Norman, R. J. & Venn, A. J. Obesity and menstrual irregularity: associations with SHBG, testosterone, and insulin. *Obesity (Silver Spring)* **17**, 1070–6 (2009). URL https://www.ncbi.nlm.nih.gov/pubmed/19180069.

[533] Douchi, T. *et al.* Relationship of upper body obesity to menstrual disorders. *Acta Obstet Gynecol Scand* **81**, 147–50 (2002). URL https://www.ncbi.nlm.nih.gov/pubmed/11942905.

[534] Shah, D. K., Correia, K. F., Vitonis, A. F. & Missmer, S. A. Body size and endometriosis: results from 20 years of follow-up within the Nurses' Health Study II prospective cohort. *Hum Reprod* **28**, 1783–92 (2013). URL https://www.ncbi.nlm.nih.gov/pubmed/23674552.

[535] McCann, S. E., Freudenheim, J. L., Darrow, S. L., Batt, R. E. & Zielezny, M. A. Endometriosis and body fat distribution. *Obstet Gynecol* **82**, 545–9 (1993). URL https://www.ncbi.nlm.nih.gov/pubmed/8377980.

[536] Sun, K., Xie, Y., Zhao, N. & Li, Z. A case-control study of the relationship between visceral fat and development of uterine fibroids. *Exp Ther Med* **18**, 404–410 (2019). URL https://www.ncbi.nlm.nih.gov/pubmed/31258679.

[537] Spradley, F. T. Metabolic abnormalities and obesity's impact on the risk for developing preeclampsia. *Am J Physiol Regul Integr Comp Physiol* **312**, R5–R12 (2017). URL https://www.ncbi.nlm.nih.gov/pubmed/27903516.

[538] Ray, J. G. *et al.* Preeclampsia and preterm birth associated with visceral adiposity in early pregnancy. *J Obstet Gynaecol Can* **39**, 78–81 (2017). URL https://www.ncbi.nlm.nih.gov/pubmed/27998690.

[539] Grodstein, F., Goldman, M. B. & Cramer, D. W. Body mass index and ovulatory infertility. *Epidemiology* **5**, 247–50 (1994). URL https://www.ncbi.nlm.nih.gov/pubmed/8173001.

[540] Ibrahim, M. M. Subcutaneous and visceral adipose tissue: structural and functional differences. *Obes Rev* **11**, 11–8 (2010). URL https://www.ncbi.nlm.nih.gov/pubmed/19656312.

[541] Angrist, J. D., Imbens, G. W. & Rubin, D. B. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association* **91**, 444–455 (1996). URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1996.10476902.

[542] Richmond, R. C. & Davey Smith, G. Mendelian randomization: Concepts and scope. *Cold Spring Harb Perspect Med* **12** (2022). URL https://www.ncbi.nlm.nih.gov/pubmed/34426474.

[543] Dixon, S. C. *et al.* Adult body mass index and risk of ovarian cancer by subtype: a Mendelian randomization study. *Int J Epidemiol* **45**, 884–95 (2016). URL https://www.ncbi.nlm.nih.gov/pubmed/27401727.

[544] Painter, J. N. *et al.* (genetic risk score) Mendelian randomization shows that obesity measured as body mass index, but not waist:hip ratio, is causal for endometrial cancer. *Cancer Epidemiol Biomarkers Prev* **25**, 1503–1510 (2016). URL https://www.ncbi.nlm.nih.gov/pubmed/27550749.

[545] Brower, M. A. *et al.* Bidirectional Mendelian randomization to explore the causal relationships between body mass index and polycystic ovary syndrome. *Hum Reprod* **34**, 127–136 (2019). URL https://www.ncbi.nlm.nih.gov/pubmed/30496407.

[546] Ambler, G. & Benner, A. *mfp: Multivariable Fractional Polynomials* (2023). URL https://CRAN.R-project.org/package=mfp. R package version 1.5.4.

[547] Wood, S. N. Thin-plate regression splines. *Journal of the Royal Statistical Society (B)* **65**, 95–114 (2003).

[548] Akaike, H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **19**, 716–723 (1974).

[549] Karlsson, T. *et al.* Contribution of genetics to visceral adiposity and its relation to cardiovascular and metabolic disease. *Nat Med* **25**, 1390–1395 (2019). URL https://www.ncbi.nlm.nih.gov/pubmed/31501611.

[550] Hemani, G. *ieugwasr: R Interface to the OpenGWAS Database API* (2023). URL `https://github.com/MRCIEU/ieugwasr`. R package version 0.1.5.

[551] Lotta, L. A. *et al.* Association of genetic variants related to gluteofemoral vs abdominal fat distribution with type 2 diabetes, coronary disease, and cardiovascular risk factors. *JAMA* **320**, 2553–2563 (2018). URL `https://www.ncbi.nlm.nih.gov/pubmed/30575882`.

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

[553] Lloyd-Jones, L. R., Robinson, M. R., Yang, J. & Visscher, P. M. Transformation of summary statistics from linear mixed model association on all-or-none traits to odds ratio. *Genetics* **208**, 1397–1408 (2018). URL `https://www.ncbi.nlm.nih.gov/pubmed/29429966`.

[554] Deng, L., Zhang, H. & Yu, K. Power calculation for the general two-sample Mendelian randomization analysis. *Genet Epidemiol* **44**, 290–299 (2020). URL `https://www.ncbi.nlm.nih.gov/pubmed/32048336`.

[555] Brion, M. J., Shakhbazov, K. & Visscher, P. M. Calculating statistical power in Mendelian randomization studies. *Int J Epidemiol* **42**, 1497–501 (2013). URL `https://www.ncbi.nlm.nih.gov/pubmed/24159078`.

[556] Hemani, G. *et al.* The MR-Base platform supports systematic causal inference across the human phenome. *Elife* **7** (2018). URL `https://www.ncbi.nlm.nih.gov/pubmed/29846171`.

[557] Bowden, J. *et al.* A framework for the investigation of pleiotropy in two-sample summary data Mendelian randomization. *Stat Med* **36**, 1783–1802 (2017). URL `https://www.ncbi.nlm.nih.gov/pubmed/28114746`.

[558] Staley, J. R. & Burgess, S. Semiparametric methods for estimation of a nonlinear exposure-outcome relationship using instrumental variables with application to Mendelian randomization. *Genet Epidemiol* **41**, 341–352 (2017). URL `https://www.ncbi.nlm.nih.gov/pubmed/28317167`.

[559] Burgess, S. Violation of the constant genetic effect assumption can result in biased estimates for Non-Linear Mendelian Randomization. *Hum Hered* **88**, 79–90 (2023). URL `https://www.ncbi.nlm.nih.gov/pubmed/37651993`.

[560] Tian, H., Mason, A. M., Liu, C. & Burgess, S. Relaxing parametric assumptions for non-linear Mendelian randomization using a doubly-ranked stratification method. *PLoS Genet* **19**, e1010823 (2023). URL `https://www.ncbi.nlm.nih.gov/pubmed/37390109`.

[561] Kilpelainen, T. O. *et al.* Genome-wide meta-analysis uncovers novel loci influencing circulating leptin levels. *Nat Commun* **7**, 10494 (2016). URL `https://www.ncbi.nlm.nih.gov/pubmed/26833098`.

[562] Chen, J. *et al.* The trans-ancestral genomic architecture of glycemic traits. *Nat Genet* **53**, 840–860 (2021). URL `https://www.ncbi.nlm.nih.gov/pubmed/34059833`.

[563] Bowden, J. *et al.* Assessing the suitability of summary data for two-sample Mendelian randomization analyses using MR-Egger regression: the role of the I2 statistic. *Int J Epidemiol* **45**, 1961–1974 (2016). URL https://www.ncbi.nlm.nih.gov/pubmed/27616674.

[564] Doob, J. L. The limiting distributions of certain statistics. *The Annals of Mathematical Statistics* **6**, 160–169 (1935).

[565] Foley, C. N. *mrclust: Identifying Clustered Heterogeneity in Mendelian Randomization Analyses* (2023). R package version 0.1.0.

[566] Liao, Y., Wang, J., Jaehnig, E. J., Shi, Z. & Zhang, B. WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. *Nucleic Acids Res* **47**, W199–W205 (2019). URL https://www.ncbi.nlm.nih.gov/pubmed/31114916.

[567] Kanehisa, M., Furumichi, M., Sato, Y., Kawashima, M. & Ishiguro-Watanabe, M. KEGG for taxonomy-based analysis of pathways and genomes. *Nucleic Acids Res* **51**, D587–D592 (2023). URL https://www.ncbi.nlm.nih.gov/pubmed/36300620.

[568] Ottolina, J. *et al.* Does endometriosis influence the age of menopause? *Minerva Ginecol* **70**, 171–177 (2018). URL https://www.ncbi.nlm.nih.gov/pubmed/28891281.

[569] Sullivan, S. D., Sarrel, P. M. & Nelson, L. M. Hormone replacement therapy in young women with primary ovarian insufficiency and early menopause. *Fertil Steril* **106**, 1588–1599 (2016). URL https://www.ncbi.nlm.nih.gov/pubmed/27912889.

[570] Millan-de Meer, M., Luque-Ramirez, M., Nattero-Chavez, L. & Escobar-Morreale, H. F. PCOS during the menopausal transition and after menopause: a systematic review and meta-analysis. *Hum Reprod Update* (2023). URL https://www.ncbi.nlm.nih.gov/pubmed/37353908.

[571] 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**, 304–14 (2016). URL https://www.ncbi.nlm.nih.gov/pubmed/27061298.

[572] Burgess, S., Davies, N. M. & Thompson, S. G. Bias due to participant overlap in two-sample Mendelian randomization. *Genet Epidemiol* **40**, 597–608 (2016). URL https://www.ncbi.nlm.nih.gov/pubmed/27625185.

[573] Hartwig, F. P., Tilling, K., Davey Smith, G., Lawlor, D. A. & Borges, M. C. Bias in two-sample Mendelian randomization when using heritable covariable-adjusted summary associations. *Int J Epidemiol* **50**, 1639–1650 (2021). URL https://www.ncbi.nlm.nih.gov/pubmed/33619569.

[574] Burgess, S. & Thompson, S. G. Bias in causal estimates from Mendelian randomization studies with weak instruments. *Stat Med* **30**, 1312–23 (2011). URL https://www.ncbi.nlm.nih.gov/pubmed/21432888.

[575] McGrath, I. M., Montgomery, G. W. & Mortlock, S. Insights from Mendelian randomization and genetic correlation analyses into the relationship between endometriosis and its comorbidities. *Hum Reprod Update* (2023). URL https://www.ncbi.nlm.nih.gov/pubmed/37159502.

[576] Thurston, L., Abbara, A. & Dhillo, W. S. Investigation and management of subfertility. *J Clin Pathol* **72**, 579–587 (2019). URL https://www.ncbi.nlm.nih.gov/pubmed/31296604.

[577] Liu, J., Schiltz, J. F., Ashar, H. R. & Chada, K. K. Hmga1 is required for normal sperm development. *Mol Reprod Dev* **66**, 81–9 (2003). URL https://www.ncbi.nlm.nih.gov/pubmed/12874803.

[578] Sanderson, E. *et al.* Mendelian randomization. *Nat Rev Methods Primers* **2** (2022). URL https://www.ncbi.nlm.nih.gov/pubmed/37325194.

[579] Burgess, S., Butterworth, A. & Thompson, S. G. Mendelian randomization analysis with multiple genetic variants using summarized data. *Genet Epidemiol* **37**, 658–65 (2013). URL https://www.ncbi.nlm.nih.gov/pubmed/24114802.

[580] Zhao, Q., Wang, J., Spiller, W., Bowden, J. & Small, D. S. Two-sample instrumental variable analyses using heterogeneous samples. *Statistical Science* **34**, 317–333 (2019).

[581] Mounier, N. & Kutalik, Z. Bias correction for inverse variance weighting Mendelian randomization. *Genet Epidemiol* **47**, 314–331 (2023). URL https://www.ncbi.nlm.nih.gov/pubmed/37036286.

[582] Cho, B. *et al.* Safety and effectiveness of dienogest (Visanne(R)) for treatment of endometriosis: A large prospective cohort study. *Reprod Sci* **27**, 905–915 (2020). URL https://www.ncbi.nlm.nih.gov/pubmed/32052358.

[583] Puhl, R. M., Luedicke, J. & Grilo, C. M. Obesity bias in training: attitudes, beliefs, and observations among advanced trainees in professional health disciplines. *Obesity (Silver Spring)* **22**, 1008–15 (2014). URL https://www.ncbi.nlm.nih.gov/pubmed/24124078.

[584] Despres, J. P. & Lemieux, I. Abdominal obesity and metabolic syndrome. *Nature* **444**, 881–7 (2006). URL https://www.ncbi.nlm.nih.gov/pubmed/17167477.

[585] Misra, A. & Vikram, N. K. Clinical and pathophysiological consequences of abdominal adiposity and abdominal adipose tissue depots. *Nutrition* **19**, 457–66 (2003). URL https://www.ncbi.nlm.nih.gov/pubmed/12714101.

[586] Stoll, B. A. Upper abdominal obesity, insulin resistance and breast cancer risk. *Int J Obes Relat Metab Disord* **26**, 747–53 (2002). URL https://www.ncbi.nlm.nih.gov/pubmed/12037643.

[587] Chandrasekaran, S., Hunt, H., Melhorn, S., Gammill, H. S. & Schur, E. A. Adipokine profiles in preeclampsia. *J Matern Fetal Neonatal Med* **33**, 2812–2817 (2020). URL https://www.ncbi.nlm.nih.gov/pubmed/30572749.

[588] Hamdy, O., Porramatikul, S. & Al-Ozairi, E. Metabolic obesity: the paradox between visceral and subcutaneous fat. *Curr Diabetes Rev* **2**, 367–73 (2006). URL https://www.ncbi.nlm.nih.gov/pubmed/18220642.

[589] Ciavattini, A. *et al.* The association between ultrasound-estimated visceral fat deposition and uterine fibroids: an observational study. *Gynecol Endocrinol* **33**, 634–637 (2017). URL https://www.ncbi.nlm.nih.gov/pubmed/28332865.

[590] Vignini, A. *et al.* Preperitoneal fat thicknesses, lipid profile, and oxidative status in women with uterine fibroids. *Reprod Sci* **24**, 1419–1425 (2017). URL https://www.ncbi.nlm.nih.gov/pubmed/28122481.

[591] Backonja, U., Buck Louis, G. M. & Lauver, D. R. Overall adiposity, adipose tissue distribution, and endometriosis: A systematic review. *Nurs Res* **65**, 151–66 (2016). URL https://www.ncbi.nlm.nih.gov/pubmed/26938364.

[592] Okifuji, A. & Hare, B. D. The association between chronic pain and obesity. *J Pain Res* **8**, 399–408 (2015). URL https://www.ncbi.nlm.nih.gov/pubmed/26203274.

[593] Engstrom, G. *et al.* Inflammation-sensitive plasma proteins are associated with future weight gain. *Diabetes* **52**, 2097–101 (2003). URL https://www.ncbi.nlm.nih.gov/pubmed/12882928.

[594] Yasui, T. *et al.* Association of endometriosis-related infertility with age at menopause. *Maturitas* **69**, 279–83 (2011). URL https://www.ncbi.nlm.nih.gov/pubmed/21605953.

[595] Oliveira, J. B. A. Does low BMI affect ART outcomes? *JBRA Assist Reprod* **22**, 1 (2018). URL https://www.ncbi.nlm.nih.gov/pubmed/29516713.

[596] Erion, K. A. & Corkey, B. E. Hyperinsulinemia: a cause of obesity? *Curr Obes Rep* **6**, 178–186 (2017). URL https://www.ncbi.nlm.nih.gov/pubmed/28466412.

[597] Templeman, N. M., Skovso, S., Page, M. M., Lim, G. E. & Johnson, J. D. A causal role for hyperinsulinemia in obesity. *J Endocrinol* **232**, R173–R183 (2017). URL https://www.ncbi.nlm.nih.gov/pubmed/28052999.

[598] Xu, C. *et al.* Bidirectional temporal relationship between obesity and hyperinsulinemia: longitudinal observation from a Chinese cohort. *BMJ Open Diabetes Res Care* **9** (2021). URL https://www.ncbi.nlm.nih.gov/pubmed/33632707.

[599] Xu, Y. & Qiao, J. Association of insulin resistance and elevated androgen levels with polycystic ovarian syndrome (PCOS): A review of literature. *J Healthc Eng* **2022**, 9240569 (2022). URL https://www.ncbi.nlm.nih.gov/pubmed/35356614.

[600] Dunaif, A. Insulin resistance and the polycystic ovary syndrome: mechanism and implications for pathogenesis. *Endocr Rev* **18**, 774–800 (1997). URL https://www.ncbi.nlm.nih.gov/pubmed/9408743.

[601] Xu, Y., Wu, Y. & Huang, Q. Comparison of the effect between pioglitazone and metformin in treating patients with PCOS:a meta-analysis. *Arch Gynecol Obstet* **296**, 661–677 (2017). URL https://www.ncbi.nlm.nih.gov/pubmed/28770353.

[602] Morley, L. C., Tang, T., Yasmin, E., Norman, R. J. & Balen, A. H. Insulin-sensitising drugs (metformin, rosiglitazone, pioglitazone, d-chiro-inositol) for women with polycystic ovary syndrome, oligo amenorrhoea and subfertility. *Cochrane Database Syst Rev* **11**, CD003053 (2017). URL https://www.ncbi.nlm.nih.gov/pubmed/29183107.

[603] Campbell, P. F. Relieving endometriosis pain: why is it so tough? *Obstet Gynecol Clin North Am* **30**, 209–20 (2003). URL https://www.ncbi.nlm.nih.gov/pubmed/12699267.

[604] Carson, S. A. & Kallen, A. N. Diagnosis and management of infertility: A review. *JAMA* **326**, 65–76 (2021). URL https://www.ncbi.nlm.nih.gov/pubmed/34228062.

[605] Shea, A. A., Wever, F., Ventola, C., Thornburg, J. & Vitzthum, V. J. More than blood: app-tracking reveals variability in heavy menstrual bleeding construct. *BMC Womens Health* **23**, 170 (2023). URL https://www.ncbi.nlm.nih.gov/pubmed/37041503.

[606] Copp, T. *et al.* Clinicians' perspectives on diagnosing polycystic ovary syndrome in Australia: a qualitative study. *Hum Reprod* **35**, 660–668 (2020). URL https://www.ncbi.nlm.nih.gov/pubmed/32101283.

[607] Laberge, P. Y., Vilos, G. A., Vilos, A. G. & Janiszewski, P. M. Burden of symptomatic uterine fibroids in Canadian women: a cohort study. *Curr Med Res Opin* **32**, 165–75 (2016). URL https://www.ncbi.nlm.nih.gov/pubmed/26455364.

[608] De La Cruz, M. S. & Buchanan, E. M. Uterine fibroids: Diagnosis and treatment. *Am Fam Physician* **95**, 100–107 (2017). URL https://www.ncbi.nlm.nih.gov/pubmed/28084714.

[609] Pavone, D., Clemenza, S., Sorbi, F., Fambrini, M. & Petraglia, F. Epidemiology and risk factors of uterine fibroids. *Best Pract Res Clin Obstet Gynaecol* **46**, 3–11 (2018). URL https://www.ncbi.nlm.nih.gov/pubmed/29054502.

[610] Gianfrancesco, M. A., Tamang, S., Yazdany, J. & Schmajuk, G. Potential biases in machine learning algorithms using electronic health record data. *JAMA Intern Med* **178**, 1544–1547 (2018). URL https://www.ncbi.nlm.nih.gov/pubmed/30128552.

[611] Verberne, L. D. M., Nielen, M. M. J., Leemrijse, C. J., Verheij, R. A. & Friele, R. D. Recording of weight in electronic health records: an observational study in general practice. *BMC Fam Pract* **19**, 174 (2018). URL https://www.ncbi.nlm.nih.gov/pubmed/30447691.

[612] Rea, S., Bailey, K. R., Pathak, J. & Haug, P. J. Bias in recording of body mass index data in the electronic health record. *AMIA Jt Summits Transl Sci Proc* **2013**, 214–8 (2013). URL https://www.ncbi.nlm.nih.gov/pubmed/24303267.

[613] Sun, H. *et al.* Debiased, longitudinal and coordinated drug recommendation through multi-visit clinic records. *Advances in Neural Information Processing Systems* **35**, 27837–27849 (2022).

[614] Rao, S. *et al.* Targeted-BEHRT: deep learning for observational causal inference on longitudinal electronic health records. *IEEE Transactions on Neural Networks and Learning Systems* (2022).

[615] Ojavee, S. E. *et al.* Genetic insights into the age-specific biological mechanisms governing human ovarian aging. *Am J Hum Genet* **110**, 1549–1563 (2023). URL https://www.ncbi.nlm.nih.gov/pubmed/37543033.

[616] Li, R. *et al.* A regression framework to uncover pleiotropy in large-scale electronic health record data. *J Am Med Inform Assoc* **26**, 1083–1090 (2019). URL https://www.ncbi.nlm.nih.gov/pubmed/31529123.

[617] Skou, S. T. *et al.* Multimorbidity. *Nat Rev Dis Primers* **8**, 48 (2022). URL https://www.ncbi.nlm.nih.gov/pubmed/35835758.

[618] Siggaard, T. *et al.* Disease trajectory browser for exploring temporal, population-wide disease progression patterns in 7.2 million Danish patients. *Nat Commun* **11**, 4952 (2020). URL https://www.ncbi.nlm.nih.gov/pubmed/33009368.

[619] Investigators, G. P. P. *et al.* 100,000 Genomes pilot on rare-disease diagnosis in health care - preliminary report. *N Engl J Med* **385**, 1868–1880 (2021). URL https://www.ncbi.nlm.nih.gov/pubmed/34758253.

[620] Firth, H. V., Wright, C. F. & Study, D. D. D. The Deciphering Developmental Disorders (DDD) study. *Dev Med Child Neurol* **53**, 702–3 (2011). URL https://www.ncbi.nlm.nih.gov/pubmed/21679367.

[621] Nguengang Wakap, S. *et al.* Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet database. *Eur J Hum Genet* **28**, 165–173 (2020). URL https://www.ncbi.nlm.nih.gov/pubmed/31527858.

[622] Frankovich, J., Longhurst, C. A. & Sutherland, S. M. Evidence-based medicine in the EMR era. *N Engl J Med* **365**, 1758–9 (2011). URL https://www.ncbi.nlm.nih.gov/pubmed/22047518.

[623] Pineda-Moncusí, M. *et al.* Digital ethnicity data in population-wide electronic health records in England: a description of completeness, coverage, and granularity of diversity. *medRxiv* 2022.11.11.22282217 (2022). URL https://www.medrxiv.org/content/medrxiv/early/2022/11/11/2022.11.11.22282217.full.pdf.

[624] Dalton, J. E. *et al.* Evaluating and modeling neighborhood diversity and health using electronic health records. *Med Decis Making* **42**, 1027–1040 (2022). URL https://www.ncbi.nlm.nih.gov/pubmed/36255188.

[625] Gallagher, M. D. & Chen-Plotkin, A. S. The Post-GWAS Era: From association to function. *Am J Hum Genet* **102**, 717–730 (2018). URL https://www.ncbi.nlm.nih.gov/pubmed/29727686.

[626] Huang, Q. Genetic study of complex diseases in the post-GWAS era. *J Genet Genomics* **42**, 87–98 (2015). URL https://www.ncbi.nlm.nih.gov/pubmed/25819085.

[627] Lappalainen, T. & MacArthur, D. G. From variant to function in human disease genetics. *Science* **373**, 1464–1468 (2021). URL https://www.ncbi.nlm.nih.gov/pubmed/34554789.

[628] Maurano, M. T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190–5 (2012). URL https://www.ncbi.nlm.nih.gov/pubmed/22955828.

[629] Watanabe, K. *et al.* A global overview of pleiotropy and genetic architecture in complex traits. *Nat Genet* **51**, 1339–1348 (2019). URL https://www.ncbi.nlm.nih.gov/pubmed/31427789.

[630] Guo, H. *et al.* Integration of disease association and eQTL data using a Bayesian colocalisation approach highlights six candidate causal genes in immune-mediated diseases. *Hum Mol Genet* **24**, 3305–13 (2015). URL https://www.ncbi.nlm.nih.gov/pubmed/25743184.

[631] Gong, J. *et al.* PancanQTL: systematic identification of cis-eQTLs and trans-eQTLs in 33 cancer types. *Nucleic Acids Res* **46**, D971–D976 (2018). URL https://www.ncbi.nlm.nih.gov/pubmed/29036324.

[632] Caliskan, M. *et al.* Genetic and epigenetic fine mapping of complex trait associated loci in the human liver. *Am J Hum Genet* **105**, 89–107 (2019). URL https://www.ncbi.nlm.nih.gov/pubmed/31204013.

[633] Eraslan, G. *et al.* Single-nucleus cross-tissue molecular reference maps toward understanding disease gene function. *Science* **376**, eabl4290 (2022). URL https://www.ncbi.nlm.nih.gov/pubmed/35549429.

[634] Consortia, C. Z. I. S.-C. C.-. *et al.* Single cell profiling of COVID-19 patients: an international data resource from multiple tissues. *medRxiv* 2020.11.20.20227355 (2020). URL https://www.medrxiv.org/content/medrxiv/early/2020/11/23/2020.11.20.20227355.full.pdf.

[635] Rozenblatt-Rosen, O., Stubbington, M. J. T., Regev, A. & Teichmann, S. A. The Human Cell Atlas: from vision to reality. *Nature* **550**, 451–453 (2017). URL https://www.ncbi.nlm.nih.gov/pubmed/29072289.

[636] Williams, C. J. & Erickson, G. F. Morphology and physiology of the ovary. In Feingold, K. R. *et al.* (eds.) *Endotext* (MDText, South Dartmouth (MA), 2000). URL https://www.ncbi.nlm.nih.gov/pubmed/25905186.

[637] Beckmann, M. W. *et al.* Operative techniques and complications of extraction and transplantation of ovarian tissue: the Erlangen experience. *Arch Gynecol Obstet* **295**, 1033–1039 (2017). URL https://www.ncbi.nlm.nih.gov/pubmed/28197717.

[638] Dardashti, K., Williams, R. H. & Goldstein, M. Microsurgical testis biopsy: a novel technique for retrieval of testicular tissue. *J Urol* **163**, 1206–7 (2000). URL https://www.ncbi.nlm.nih.gov/pubmed/10737497.

[639] Strassmann, B. I. The evolution of endometrial cycles and menstruation. *Q Rev Biol* **71**, 181–220 (1996). URL https://www.ncbi.nlm.nih.gov/pubmed/8693059.

[640] Thompson, C. & Wessel, G. M. Do I menstruate or "estruate"?-Is this just a potato, potatoe thing? *Mol Reprod Dev* **87**, 737–738 (2020). URL https://www.ncbi.nlm.nih.gov/pubmed/32592190.

[641] Vailhe, B., Vittet, D. & Feige, J. J. In vitro models of vasculogenesis and angiogenesis. *Lab Invest* **81**, 439–52 (2001). URL https://www.ncbi.nlm.nih.gov/pubmed/11304563.

[642] Miller, A. J. & Spence, J. R. In vitro models to study human lung development, disease and homeostasis. *Physiology (Bethesda)* **32**, 246–260 (2017). URL https://www.ncbi.nlm.nih.gov/pubmed/28404740.

[643] Telfer, E. E. In vitro models for oocyte development. *Theriogenology* **49**, 451–60 (1998). URL https://www.ncbi.nlm.nih.gov/pubmed/10732026.

[644] Cristancho, A. G. & Lazar, M. A. Forming functional fat: a growing understanding of adipocyte differentiation. *Nat Rev Mol Cell Biol* **12**, 722–34 (2011). URL https://www.ncbi.nlm.nih.gov/pubmed/21952300.

[645] Shan, B. *et al.* Multilayered omics reveal sex- and depot-dependent adipose progenitor cell heterogeneity. *Cell Metab* **34**, 783–799 e7 (2022). URL https://www.ncbi.nlm.nih.gov/pubmed/35447091.

[646] Rampler, E. *et al.* The power of LC-MS based multiomics: Exploring adipogenic differentiation of human mesenchymal stem/stromal cells. *Molecules* **24** (2019). URL https://www.ncbi.nlm.nih.gov/pubmed/31597247.

[647] Huang, M. *et al.* Engineered allele substitution at PPARGC1A rs8192678 alters human white adipocyte differentiation, lipogenesis, and PGC-1alpha content and turnover. *Diabetologia* **66**, 1289–1305 (2023). URL https://www.ncbi.nlm.nih.gov/pubmed/37171500.

[648] Zebisch, K., Voigt, V., Wabitsch, M. & Brandsch, M. Protocol for effective differentiation of 3T3-L1 cells to adipocytes. *Anal Biochem* **425**, 88–90 (2012). URL https://www.ncbi.nlm.nih.gov/pubmed/22425542.

[649] Xue, R. *et al.* Clonal analyses and gene profiling identify genetic biomarkers of the thermogenic potential of human brown and white preadipocytes. *Nat Med* **21**, 760–8 (2015). URL https://www.ncbi.nlm.nih.gov/pubmed/26076036.

[650] Wabitsch, M. *et al.* Characterization of a human preadipocyte cell strain with high capacity for adipose differentiation. *Int J Obes Relat Metab Disord* **25**, 8–15 (2001). URL https://www.ncbi.nlm.nih.gov/pubmed/11244452.

[651] Skurk, T. & Hauner, H. Primary culture of human adipocyte precursor cells: expansion and differentiation. *Methods Mol Biol* **806**, 215–26 (2012). URL https://www.ncbi.nlm.nih.gov/pubmed/22057455.

[652] Doudna, J. A. & Charpentier, E. Genome editing. the new frontier of genome engineering with CRISPR-Cas9. *Science* **346**, 1258096 (2014). URL https://www.ncbi.nlm.nih.gov/pubmed/25430774.

[653] Zhang, F., Wen, Y. & Guo, X. CRISPR/Cas9 for genome editing: progress, implications and challenges. *Hum Mol Genet* **23**, R40–6 (2014). URL https://www.ncbi.nlm.nih.gov/pubmed/24651067.

[654] Sternberg, S. H. & Doudna, J. A. Expanding the biologist's toolkit with CRISPR-Cas9. *Mol Cell* **58**, 568–74 (2015). URL https://www.ncbi.nlm.nih.gov/pubmed/26000842.

[655] Laber, S. *et al.* Discovering cellular programs of intrinsic and extrinsic drivers of metabolic traits using LipocyteProfiler. *Cell Genom* **3**, 100346 (2023). URL https://www.ncbi.nlm.nih.gov/pubmed/37492099.

[656] Parnas, O. *et al.* A genome-wide CRISPR screen in primary immune cells to dissect regulatory networks. *Cell* **162**, 675–86 (2015). URL https://www.ncbi.nlm.nih.gov/pubmed/26189680.

[657] Yan, X. *et al.* High-content imaging-based pooled CRISPR screens in mammalian cells. *J Cell Biol* **220** (2021). URL https://www.ncbi.nlm.nih.gov/pubmed/33465779.

[658] Qi, L. S. *et al.* Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. *Cell* **152**, 1173–83 (2013). URL https://www.ncbi.nlm.nih.gov/pubmed/23452860.

[659] Komor, A. C., Kim, Y. B., Packer, M. S., Zuris, J. A. & Liu, D. R. Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. *Nature* **533**, 420–4 (2016). URL https://www.ncbi.nlm.nih.gov/pubmed/27096365.

[660] Gaudelli, N. M. *et al.* Programmable base editing of A*T to G*C in genomic DNA without DNA cleavage. *Nature* **551**, 464–471 (2017). URL https://www.ncbi.nlm.nih.gov/pubmed/29160308.

[661] Anzalone, A. V. *et al.* Search-and-replace genome editing without double-strand breaks or donor DNA. *Nature* **576**, 149–157 (2019). URL https://www.ncbi.nlm.nih.gov/pubmed/31634902.

[662] Wang, J. Y. & Doudna, J. A. CRISPR technology: A decade of genome editing is only the beginning. *Science* **379**, eadd8643 (2023). URL https://www.ncbi.nlm.nih.gov/pubmed/36656942.

[663] Bock, C. *et al.* High-content CRISPR screening. *Nat Rev Methods Primers* **2** (2022). URL https://www.ncbi.nlm.nih.gov/pubmed/37214176.

[664] Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021). URL https://www.ncbi.nlm.nih.gov/pubmed/34265844.

[665] Cheng, J. *et al.* Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science* **381**, eadg7492 (2023). URL https://www.ncbi.nlm.nih.gov/pubmed/37733863.

[666] Gao, H. *et al.* The landscape of tolerated genetic variation in humans and primates. *Science* **380**, eabn8153 (2023). URL https://www.ncbi.nlm.nih.gov/pubmed/37262156.

[667] Deplancke, B., Alpern, D. & Gardeux, V. The genetics of transcription factor DNA binding variation. *Cell* **166**, 538–554 (2016). URL https://www.ncbi.nlm.nih.gov/pubmed/27471964.

[668] Lambert, S. A. *et al.* The human transcription factors. *Cell* **172**, 650–665 (2018). URL https://www.ncbi.nlm.nih.gov/pubmed/29425488.

[669] Hamamoto, K. & Fukaya, T. Molecular architecture of enhancer-promoter interaction. *Curr Opin Cell Biol* **74**, 62–70 (2022). URL https://www.ncbi.nlm.nih.gov/pubmed/35168174.

[670] Stadhouders, R. *et al.* Transcription regulation by distal enhancers: who's in the loop? *Transcription* **3**, 181–6 (2012). URL https://www.ncbi.nlm.nih.gov/pubmed/22771987.

[671] Yao, L., Berman, B. P. & Farnham, P. J. Demystifying the secret mission of enhancers: linking distal regulatory elements to target genes. *Crit Rev Biochem Mol Biol* **50**, 550–73 (2015). URL https://www.ncbi.nlm.nih.gov/pubmed/26446758.

[672] Schaid, D. J., Chen, W. & Larson, N. B. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat Rev Genet* **19**, 491–504 (2018). URL https://www.ncbi.nlm.nih.gov/pubmed/29844615.

[673] Nica, A. C. & Dermitzakis, E. T. Expression quantitative trait loci: present and future. *Philos Trans R Soc Lond B Biol Sci* **368**, 20120362 (2013). URL https://www.ncbi.nlm.nih.gov/pubmed/23650636.

[674] Kang, J. B., Raveane, A., Nathan, A., Soranzo, N. & Raychaudhuri, S. Methods and insights from single-cell expression quantitative trait loci. *Annu Rev Genomics Hum Genet* **24**, 277–303 (2023). URL https://www.ncbi.nlm.nih.gov/pubmed/37196361.

[675] Ulirsch, J. C. *et al.* Interrogation of human hematopoiesis at single-cell and single-variant resolution. *Nat Genet* **51**, 683–693 (2019). URL https://www.ncbi.nlm.nih.gov/pubmed/30858613.

[676] Lahnemann, D. *et al.* Eleven grand challenges in single-cell data science. *Genome Biol* **21**, 31 (2020). URL https://www.ncbi.nlm.nih.gov/pubmed/32033589.

[677] Trapnell, C. Defining cell types and states with single-cell genomics. *Genome Res* **25**, 1491–8 (2015). URL https://www.ncbi.nlm.nih.gov/pubmed/26430159.

[678] Ni, J. *et al.* All human genes of the uteroglobin family are localized on chromosome 11q12.2 and form a dense cluster. *Ann N Y Acad Sci* **923**, 25–42 (2000). URL https://www.ncbi.nlm.nih.gov/pubmed/11193762.

[679] Glunk, V. *et al.* A non-coding variant linked to metabolic obesity with normal weight affects actin remodelling in subcutaneous adipocytes. *Nat Metab* **5**, 861–879 (2023). URL https://www.ncbi.nlm.nih.gov/pubmed/37253881.

[680] Downes, D. J. *et al.* Identification of LZTFL1 as a candidate effector gene at a COVID-19 risk locus. *Nat Genet* **53**, 1606–1615 (2021). URL https://www.ncbi.nlm.nih.gov/pubmed/34737427.

[681] Belton, J. M. *et al.* Hi-C: a comprehensive technique to capture the conformation of genomes. *Methods* **58**, 268–76 (2012). URL https://www.ncbi.nlm.nih.gov/pubmed/22652625.

[682] Han, J., Zhang, Z. & Wang, K. 3C and 3C-based techniques: the powerful tools for spatial genome organization deciphering. *Mol Cytogenet* **11**, 21 (2018). URL https://www.ncbi.nlm.nih.gov/pubmed/29541161.

[683] Downes, D. J. *et al.* Capture-C: a modular and flexible approach for high-resolution chromosome conformation capture. *Nat Protoc* **17**, 445–475 (2022). URL https://www.ncbi.nlm.nih.gov/pubmed/35121852.

[684] Hamley, J. C., Li, H., Denny, N., Downes, D. & Davies, J. O. J. Determining chromatin architecture with Micro Capture-C. *Nat Protoc* **18**, 1687–1711 (2023). URL https://www.ncbi.nlm.nih.gov/pubmed/36991220.

[685] Schwessinger, R. *et al.* DeepC: predicting 3D genome folding using megabase-scale transfer learning. *Nat Methods* **17**, 1118–1124 (2020). URL https://www.ncbi.nlm.nih.gov/pubmed/33046896.

[686] Karollus, A., Mauermeier, T. & Gagneur, J. Current sequence-based models capture gene expression determinants in promoters but mostly ignore distal enhancers. *Genome Biol* **24**, 56 (2023). URL https://www.ncbi.nlm.nih.gov/pubmed/36973806.

[687] Agarwal, V. & Shendure, J. Predicting mRNA abundance directly from genomic sequence using deep convolutional neural networks. *Cell Rep* **31**, 107663 (2020). URL https://www.ncbi.nlm.nih.gov/pubmed/32433972.

[688] Avsec, Z. *et al.* Effective gene expression prediction from sequence by integrating long-range interactions. *Nat Methods* **18**, 1196–1203 (2021). URL https://www.ncbi.nlm.nih.gov/pubmed/34608324.

[689] Chan, S. L., Jin, S., Loh, M. & Brunham, L. R. Progress in understanding the genomic basis for adverse drug reactions: a comprehensive review and focus on the role of ethnicity. *Pharmacogenomics* **16**, 1161–78 (2015). URL https://www.ncbi.nlm.nih.gov/pubmed/25978008.

[690] Long, C., Amoasii, L., Bassel-Duby, R. & Olson, E. N. Genome editing of monogenic neuromuscular diseases: A systematic review. *JAMA Neurol* **73**, 1349–1355 (2016). URL https://www.ncbi.nlm.nih.gov/pubmed/27668807.

[691] Frangoul, H. *et al.* CRISPR-Cas9 gene editing for sickle cell disease and beta-thalassemia. *N Engl J Med* **384**, 252–260 (2021). URL https://www.ncbi.nlm.nih.gov/pubmed/33283989.

[692] Therrell, B. L. *et al.* Current status of newborn screening worldwide: 2015. *Semin Perinatol* **39**, 171–87 (2015). URL https://www.ncbi.nlm.nih.gov/pubmed/25979780.

[693] Stark, Z. & Scott, R. H. Genomic newborn screening for rare diseases. *Nat Rev Genet* **24**, 755–766 (2023). URL https://www.ncbi.nlm.nih.gov/pubmed/37386126.

[694] Bayefsky, M. J. & Berkman, B. E. Implementing expanded prenatal genetic testing: Should parents have access to any and all fetal genetic information? *Am J Bioeth* **22**, 4–22 (2022). URL https://www.ncbi.nlm.nih.gov/pubmed/33459580.

[695] Orchid Health. https://www.orchidhealth.com/. Accessed: 2023-12-05.

[696] Wang, Y. *et al.* Theoretical and empirical quantification of the accuracy of polygenic scores in ancestry divergent populations. *Nat Commun* **11**, 3865 (2020). URL https://www.ncbi.nlm.nih.gov/pubmed/32737319.

[697] Obermeyer, Z., Powers, B., Vogeli, C. & Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* **366**, 447–453 (2019). URL https://www.ncbi.nlm.nih.gov/pubmed/31649194.

[698] Nazer, L. H. *et al.* Bias in artificial intelligence algorithms and recommendations for mitigation. *PLOS Digit Health* **2**, e0000278 (2023). URL https://www.ncbi.nlm.nih.gov/pubmed/37347721.

[699] Beery, A. K. & Zucker, I. Sex bias in neuroscience and biomedical research. *Neurosci Biobehav Rev* **35**, 565–72 (2011). URL https://www.ncbi.nlm.nih.gov/pubmed/20620164.

[700] Shah, K., McCormack, C. E. & Bradbury, N. A. Do you know the sex of your cells? *Am J Physiol Cell Physiol* **306**, C3–18 (2014). URL https://www.ncbi.nlm.nih.gov/pubmed/24196532.

[701] Cottingham, M. D. & Fisher, J. A. Gendered logics of biomedical research: Women in U.S. Phase I clinical trials. *Soc Probl* **69**, 492–509 (2022). URL https://www.ncbi.nlm.nih.gov/pubmed/35449716.

[702] Bierer, B. E., Meloney, L. G., Ahmed, H. R. & White, S. A. Advancing the inclusion of underrepresented women in clinical research. *Cell Rep Med* **3**, 100553 (2022). URL https://www.ncbi.nlm.nih.gov/pubmed/35492242.

[703] De Freitas, S. *et al.* Comprehensive framework of factors accounting for worse aortic aneurysm outcomes in females: A scoping review. *Semin Vasc Surg* **36**, 508–516 (2023). URL https://www.ncbi.nlm.nih.gov/pubmed/38030325.

[704] Gaudino, M. *et al.* Sex differences in outcomes after coronary artery bypass grafting: a pooled analysis of individual patient data. *Eur Heart J* **43**, 18–28 (2021). URL https://www.ncbi.nlm.nih.gov/pubmed/34338767.

[705] Volgman, A. S. *et al.* Sex differences in cardiovascular disease and cognitive impairment: Another health disparity for women? *J Am Heart Assoc* **8**, e013154 (2019). URL https://www.ncbi.nlm.nih.gov/pubmed/31549581.

[706] Ortona, E. *et al.* Sex-based differences in autoimmune diseases. *Ann Ist Super Sanita* **52**, 205–12 (2016). URL https://www.ncbi.nlm.nih.gov/pubmed/27364395.

[707] Diseases, G. B. D. & Injuries, C. Global burden of 369 diseases and injuries in 204 countries and territories, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019. *Lancet* **396**, 1204–1222 (2020). URL https://www.ncbi.nlm.nih.gov/pubmed/33069326.

[708] Clasen, T. *et al.* Estimating the impact of unsafe water, sanitation and hygiene on the global burden of disease: evolving and alternative methods. *Trop Med Int Health* **19**, 884–93 (2014). URL https://www.ncbi.nlm.nih.gov/pubmed/24909205.

[709] Chong, B. *et al.* Trends and predictions of malnutrition and obesity in 204 countries and territories: an analysis of the Global Burden of Disease Study 2019. *EClinicalMedicine* **57**, 101850 (2023). URL https://www.ncbi.nlm.nih.gov/pubmed/36864983.

[710] Carlson, C. J., Alam, M. S., North, M. A., Onyango, E. & Stewart-Ibarra, A. M. The health burden of climate change: A call for global scientific action. *PLOS Climate* **2**, e0000126 (2023). URL https://doi.org/10.1371/journal.pclm.0000126.

[711] Shaffer, R. M. *et al.* Improving and expanding estimates of the global burden of disease due to environmental health risk factors. *Environ Health Perspect* **127**, 105001 (2019). URL https://www.ncbi.nlm.nih.gov/pubmed/31626566.

[712] Hotez, P. J. *Forgotten people, forgotten diseases: the neglected tropical diseases and their impact on global health and development* (John Wiley & Sons, 2021).

[713] Mulholland, K. Perspectives on the burden of pneumonia in children. *Vaccine* **25**, 2394–7 (2007). URL https://www.ncbi.nlm.nih.gov/pubmed/17064827.

[714] Whitney, C. G. Measuring progress on preventing pneumonia deaths: are we there yet? *Lancet Infect Dis* **17**, 1100–1101 (2017). URL `https://www.ncbi.nlm.nih.gov/pubmed/28843577`.

[715] Team, W. World mental health report: Transforming mental health for all. `https://iris.who.int/bitstream/handle/10665/356119/9789240049338-eng.pdf?sequence=1` (2022). Accessed: 2023-12-19.