The impact of rare variation on extreme and unexpected phenotypes in the UK Biobank

Nikolas Alessandro Baya

Pembroke College
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy

Trinity 2025

Abstract

Human disease can be characterised by deviation from a "normal" or healthy state. Using exome sequencing data from the UK Biobank (UKB), I examined the effect of rare coding variants on deviations from normality to better understand disease. Biomarker deviations from reference ranges are commonly used to indicate disease. For the first time, I quantified the proportion of individuals in UKB with abnormal biomarker values across 32 blood and urine biomarkers. Disease risk increased with the number of abnormal biomarkers, highlighting their clinical relevance. By contrasting traditional linear association between continuous biomarker level with dichotomised (extreme vs. non-extreme value) biomarkers, I identified associations which would not otherwise have been found, including a *PNPLA3* missense variant associated with extreme serum triglyceride levels. Polygenic scores (PGS) predict complex traits and stratify disease risk but fail to fully capture individual-level variation. Misaligned individuals, whose phenotypes deviate from expectation based on PGS, allow us to identify risk factors beyond common-variant effects. I tested whether rare damaging variation in rare disorder genes can explain this deviation for ten traits: type 2 diabetes (T2D) cases carrying pathogenic variants have significantly lower PGS than non-carriers, and both coronary artery disease and T2D controls carrying protective variants have significantly higher PGS than non-carriers. Using this framework, I searched for novel protective or pathogenic genes, identifying *KANK1* as a therapeutic target for primary ovarian insufficiency. Finally, I led the largest-ever systematic functional characterisation of genes associated with fat accumulation, a disease mechanism in which abnormal deviations cause obesity and lipodystrophy. I identified *SLTM* as a novel therapeutic target for obesity, due to its effect on earlier onset of obesity, its replication in an independent dataset (All of Us), and the significant effect of *SLTM* knockdown on lipid accumulation in human fat cells. Taken together, this thesis advances understanding of the impact of rare variants on extreme and unexpected phenotypic deviations, and contributes meaningful insights into the genetic aetiology of human disease.

The impact of rare variation on extreme and unexpected phenotypes in the UK Biobank



Nikolas Alessandro Baya

Pembroke College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2025

Acknowledgements

Personal

I am grateful to Professor Cecilia Lindgren and Professor Simon Myers for their supervision, inspiring brilliance, and generosity. I have been fortunate to be a part of two interdisciplinary labs filled with wonderfully talented and kind people. My gratitude goes to Dr. Duncan Palmer for his mentorship and humour, and to my peers in the Lindgren lab, Fred, Samvida, and Barney, for their support and being gracious role models. I thank Saskia, Hannah, and Phil for their helpful discussions and Ilknur for her tireless lab work. From the Myers lab, I thank Lino, Hrushi, and Ipsita for their thoughtful feedback.

It has been a pleasure to share these past four years with my GMS cohort: Fred, Vassi, Piotrek, Annie, Jane, and Hrushi. This has included enduring Covid lockdowns, living together, learning to row, becoming DJs, going to Romania for Jane's wedding, an all-nighter at the Pembroke Ball, and becoming squash pros. A meaningful non-academic part of my life in Oxford has been the Pembroke College Boat Club. Thank you to James, Tom, Jakob, Irene, Kacper, Rob, Teddy, Simon, Sam, Niko, and Andy for helping me become a better rower and person. Special thanks go to the 2021 Summer Torpids M3 for sparking my passion and the 2025 Torpids W1 for making history.

My family and Vermont friends are my foundation. I thank Will, David, Thomas, Forrest, Lauren, Evan, Addie, and Emily for our friendship; long may it continue! I aspire to be as centred and supportive as my mother–my life has been guided by her wisdom. I am grateful for the love of my grandparents and hope I can continue to make them proud. I appreciate my *keluarga Indonesia* for their support from afar.

Institutional

I am grateful to the Wellcome Trust, whose generous research and living stipends have enabled pure academic pursuit without financial burden. To the Clarendon Scholarship, thank you for helping me, a non-EU student with non-EU tuition fees, to attend Oxford. I thank Pembroke College for the Oxford-Bendich Scholarship, which serendipitously brought me to Pembroke. Finally, thank you to the UK Biobank participants for their monumental contribution to science.

Statement of authorship

Collaboration is essential to good science. Some of the work presented in this thesis is the product of team-based collaborative efforts. Throughout this thesis, I indicate instances where data generation, data processing, quality control, or analysis was performed by a collaborator. Unless otherwise specified, all code and analyses to generate my thesis results, including text, figures, and tables, are my original work. Throughout my thesis I use the pronouns "we" and "I" to reflect both the collaborative nature of this work and and my individual contributions. Below I summarise the contributions of collaborators:

Resources used for multiple chapters

- Genetic ancestry classification – Duncan Palmer
- Variant consequence annotation – Frederik Lassen and Barney Hill

Chapter 2

- Blood and urine biomarker quality control – Samvida Venkatesh

Chapter 3

No additional areas to report.

Chapter 4

This chapter is based on an unpublished manuscript for which I share first authorship with Ilknur Sur Erdem, who was responsible for much of the *in vitro* work in this chapter. The shared first authorship reflects the complementary *in silico* and *in vitro* analyses necessary to make this work novel and high-impact. In the ordering of the two first authors, I am listed first, to reflect my leadership role in driving all aspects of the project.

***In vitro* analyses**

- Generation of wild type human white adipose tissue RNA-seq data – Richard Yan-Do and Saskia Reibe
- Generation of CRISPR knockdown cell lines – Ilknur Sur Erdem
- Generating RT-qPCR data – Ilknur Sur Erdem
- Quantifying BODIPY fluorescence with Opera Phenix High-Content Screening System – Ilknur Sur Erdem

***In silico* analyses**

- Longitudinal analysis of obesity age of onset – Samvida Venkatesh
- Validation of associated genes in All of Us – Duncan Palmer and Samvida Venkatesh
- Collation of OpenTargets druggability evidence – Duncan Palmer

Abstract

Human disease can be characterised by deviation from a "normal" or healthy state. Using exome sequencing data from the UK Biobank (UKB), I examined the effect of rare coding variants on deviations from normality to better understand disease. Biomarker deviations from reference ranges are commonly used to indicate disease. For the first time, I quantified the proportion of individuals in UKB with abnormal biomarker values across 32 blood and urine biomarkers. Disease risk increased with the number of abnormal biomarkers, highlighting their clinical relevance. By contrasting traditional linear association between continuous biomarker level with dichotomised (extreme vs. non-extreme value) biomarkers, I identified associations which would not otherwise have been found, including a *PNPLA3* missense variant associated with extreme serum triglyceride levels. Polygenic scores (PGS) predict complex traits and stratify disease risk but fail to fully capture individual-level variation. Misaligned individuals, whose phenotypes deviate from expectation based on PGS, allow us to identify risk factors beyond common-variant effects. I tested whether rare damaging variation in rare disorder genes can explain this deviation for ten traits: type 2 diabetes (T2D) cases carrying pathogenic variants have significantly lower PGS than non-carriers, and both coronary artery disease and T2D controls carrying protective variants have significantly higher PGS than non-carriers. Using this framework, I searched for novel protective or pathogenic genes, identifying *KANK1* as a therapeutic target for primary ovarian insufficiency. Finally, I led the largest-ever systematic functional characterisation of genes associated with fat accumulation, a disease mechanism in which abnormal deviations cause obesity and lipodystrophy. I identified *SLTM* as a novel therapeutic target for obesity, due to its effect on earlier onset of obesity, its replication in an independent dataset (All of Us), and the significant effect of *SLTM* knockdown on lipid accumulation in human fat cells. Taken together, this thesis advances understanding of the impact of rare variants on extreme and unexpected phenotypic deviations, and contributes meaningful insights into the genetic aetiology of human disease.

Contents

List of Figures **xiii**

List of Tables **xvii**

1 Introduction **1**
1.1 Preamble . . . 1
1.2 A brief history of (statistical) genetics and its application to the study of disease . . . 2
1.2.1 Early genetic theories . . . 2
1.2.2 The foundational theories in evolution and inheritance . . . 3
1.2.3 The beginning of statistical genetics . . . 4
1.2.4 The physical basis of inheritance . . . 5
1.2.5 The molecular basis of genetics . . . 6
1.2.6 Advancements in quantifying genetic variation . . . 6
1.2.7 The indirect approach . . . 7
1.2.8 The era of genome-wide association studies . . . 8
1.2.9 Rare variant association studies and sequencing . . . 9
1.3 Thesis aims . . . 12
1.3.1 Big questions, big data . . . 12
1.3.2 Aims . . . 12

2 The genetics of extreme blood and urine biomarkers **15**
2.1 Abstract . . . 16
2.2 Introduction . . . 17
2.3 Main . . . 20
2.3.1 Many individuals in UKB have biomarker levels outside of clinical reference ranges . . . 20
2.3.2 Using extreme biomarker levels to prioritise individuals with rare genetic disorders for genetic screening . . . 32
2.3.3 Discovering variants associated with extreme biomarker levels 38
2.3.4 Rare variant gene burden associated with multivariate biomarker extremes . . . 48

2.4 Methods . . . 53
2.4.1 Defining clinical reference ranges . . . 53
2.4.2 Quality control of UKB exome sequencing data . . . 53
2.4.3 Variant annotation . . . 55
2.4.4 Selection of blood and urine biomarkers . . . 55
2.4.5 Outlier overlap enrichment between pairs of biomarkers . . . 56
2.4.6 Enrichment of ICD-10 codes among individuals with outlying biomarkers . . . 56
2.4.7 Enrichment of predicted loss-of-function variants in GEL PanelApp rare disorder genes among individuals with extreme biomarkers . . . 57
2.4.8 Defining tail-status phenotypes . . . 57
2.4.9 Single-variant tests on biomarker continuous and tail-status phenotypes . . . 58
2.4.10 Testing odds of low- and high-tail associations having negative or positive effect on the continuous trait . . . 58
2.4.11 Defining significant non-monotonic allele frequency . . . 59
2.4.12 Simulating vQTLs . . . 59
2.4.13 Defining Mahalanobis distance on joint distribution of biomarkers . . . 60
2.4.14 Gene-based burden tests on Mahalanobis distance and single biomarkers . . . 60
2.5 Discussion . . . 61

3 Individuals with phenotypes misaligned with polygenic prediction are enriched for rare pathogenic variation 67
3.1 Abstract . . . 68
3.2 Introduction . . . 69
3.3 Main . . . 72
3.3.1 Identifying individuals with misaligned continuous phenotypes 72
3.3.2 Individuals with misaligned continuous phenotypes are enriched for predicted loss-of-function variants in monogenic disorder genes . . . 74
3.3.3 Individuals whose disease status is misaligned with genetically expected risk are enriched for predicted loss-of-function variants in monogenic genes . . . 78
3.3.4 Disease controls with high polygenic risk are enriched for rare protective predicted loss-of-function variants in monogenic disorder genes . . . 82

3.3.5 Individuals whose continuous phenotype misaligns with genetic expectation are enriched for predicted damaging variants in genes related to rare disorders . . . 84
3.3.6 Individuals with genetically unexpected phenotypes are nominally enriched for singleton predicted loss-of-function variants 88
3.3.7 Exome-wide scan for genes associated with misaligned status in continuous traits . . . 91
3.3.8 Exome-wide scan for predicted loss-of-function variant burden associated with protective misaligned disease status . . . 98
3.4 Methods . . . 102
3.4.1 Exome sequencing data QC . . . 102
3.4.2 Phenotype curation . . . 103
3.4.3 Monogenic and rare disorder gene curation . . . 104
3.4.4 Misaligned phenotype classification . . . 105
3.4.5 Enrichment testing of carrier status and misaligned status . 107
3.4.6 Hypergeometric test for probability of observing no carriers . 109
3.4.7 Association of carrier status with polygenic risk score for dichotomous traits . . . 109
3.4.8 Singleton burden in intolerant genes . . . 110
3.4.9 Exploratory burden testing on phenotype misalignment . . . 111
3.5 Discussion . . . 113

4 Combining evidence from human genetic and functional screens to identify pathways altering obesity and fat distribution 117
4.1 Abstract . . . 118
4.2 Introduction . . . 119
4.3 Main . . . 120
4.3.1 Discovery of genes associated with obesity and fat distribution through exome-wide association tests . . . 120
4.3.2 Convergence of rare and common variant evidence for genes associated with obesity and fat distribution . . . 124
4.3.3 Sex- and age-specific effects of genes associated with obesity and fat distribution . . . 128
4.3.4 Nominating target genes for *in vitro* functional characterisation through CRISPR knockdown . . . 129
4.3.5 Knockdown of obesity and fat distribution associated genes alters *in vitro* adipogenesis and lipid accumulation . . . 133
4.3.6 Transcriptome-wide effects of gene knockdowns . . . 134
4.4 Methods . . . 138

4.4.1 Imputed data QC . . . 138
4.4.2 Exome sequencing data QC . . . 139
4.4.3 Variant consequence annotation . . . 141
4.4.4 Phenotype curation . . . 142
4.4.5 Genetic association testing . . . 142
4.4.6 Common variant gene prioritisation scores . . . 146
4.4.7 AoU gene-level summary statistics . . . 147
4.4.8 Genebass gene-level summary statistics . . . 147
4.4.9 Genetic correlation estimates . . . 147
4.4.10 Selecting target genes . . . 148
4.4.11 Wet lab methods . . . 149
4.4.12 Differential expression analysis of CRISPR adipocyte RNA sequencing . . . 152
4.4.13 Statistical analysis of lipid accumulation assays . . . 152
4.4.14 Pathway enrichment in CRISPR RNA sequencing data . . . 153
4.4.15 Druggability evidence . . . 153
4.5 Discussion . . . 154

5 Conclusions and future directions 159
5.1 Conclusions . . . 160
5.1.1 Overall summary . . . 160
5.1.2 The genetics of extreme blood and urine biomarker levels . . . 160
5.1.3 Individuals with phenotypes misaligned with polygenic prediction are enriched for rare pathogenic variation . . . 162
5.1.4 Combining evidence from human genetic and functional screens to identify pathways altering obesity and fat distribution . . . 164
5.2 Limitations and next steps . . . 165
5.2.1 Overall limitations . . . 165
5.2.2 The genetics of extreme blood and urine biomarker levels . . . 168
5.2.3 Individuals with phenotypes misaligned with polygenic prediction are enriched for rare pathogenic variation . . . 170
5.2.4 Combining evidence from human genetic and functional screens to identify pathways altering obesity and fat distribution . . . 171
5.3 A vision for the future of genetics and health . . . 172
5.3.1 More data . . . 173
5.3.2 Better accessibility . . . 174
5.4 Final remarks . . . 177

S2 Supplement to Chapter 2 179
S2.1 Supplementary figures . . . 179
S2.2 Supplementary tables . . . 187

S3 Supplement to Chapter 3 **197**
S3.1 Supplementary notes 197
S3.1.1 Binarised continuous traits recapitulate significant loss-of-function variant enrichment observed for continuous traits . 197
S3.2 Supplementary figures 199
S3.3 Supplementary tables 203

S4 Supplement to Chapter 4 **213**
S4.1 Supplementary notes 213
S4.1.1 Calculating unweighted burden effect sizes 213
S4.1.2 Significant genes with low minor allele count 214
S4.2 Supplementary figures 214
S4.3 Supplementary tables 222

References **237**

List of Figures

2.1 Raw UKB blood and urine biomarker distributions relative to NHS reference ranges . . . 21
2.2 High variability in the number of individuals outside of clinical reference ranges for 32 UKB blood and urine biomarkers . . . 25
2.3 Percentage of individuals below, above and outside sex-specific and sex-agnostic reference ranges . . . 26
2.4 Many individuals are outliers for multiple biomarkers . . . 27
2.5 Comparing biomarker correlation with outlier overlap enrichment . 30
2.6 Disease prevalence increases with count of outlying biomarker measurements . . . 33
2.7 Enrichment of GEL PanelApp rare disorder pLoF variants among individuals with outlying biomarker levels . . . 36
2.8 Variants associated with tail-status phenotypes . . . 41
2.9 Effects on tail-status and continuous phenotypes for variants which are VOI for multiple biomarkers . . . 44
2.10 Effect size as a function of minor allele frequency for missense VOI 45
2.11 Non-monotonic allele frequency in either-tail associations . . . 46
2.12 Gene burden associations with Mahalanobis distance calculated on the joint distribution of biomarkers . . . 51

3.1 Classification of misaligned individuals for seven continuous traits . 73
3.2 Enrichment of pLoF and damaging variants in canonical monogenic disorder genes . . . 79
3.3 Pathogenic pLoFs are associated with lower PRS in disease cases . . 83
3.4 Protective pLoFs are associated with higher PRS in disease controls 85
3.5 Misaligned individuals are enriched for pLoF variants in GEL PanelApp rare disease genes . . . 87
3.6 Misaligned individuals in continuous traits are nominally enriched for singleton burden in intolerant genes . . . 91
3.7 Misaligned individuals are nominally enriched in some dichotomous traits for singleton burden in intolerant genes . . . 92

3.8 Strength of protective burden association, comparing burden test on original phenotype versus burden test of protective misaligned status 96
3.9 Strength of pathogenic burden association, comparing burden test on original phenotype versus burden test of pathogenic misaligned status 98
3.10 Gene-level rare pLoF burden association with PRS in disease controls 100

4.1 Consequences of rare missense variation on obesity or fat distribution related traits in UKB 124
4.2 Gene burden effects across all nine obesity and fat distribution traits for 19 genes with exome-wide significant burden associations 126
4.3 CRISPR knockdown strategy 133
4.4 Effect of CRISPR single-gene knockdown on lipid accumulation in human adipocytes 135
4.5 Effect of gene knockdown on differential gene expression in human adipocytes 137

5.1 Expected cost per human genome 174

S2.1 Heatmap of raw biomarker Pearson correlations 180
S2.2 Relative power of association for continuous trait versus tail status phenotype 181
S2.3 Allele frequency across phenotype distribution for coding non-synonymous either-tail VOI 182
S2.4 Allele frequency across phenotype distribution for coding non-synonymous low-tail-only VOI 183
S2.5 Allele frequency across phenotype distribution for coding non-synonymous high-tail-only VOI 184
S2.6 Marginal single-biomarker distributions for individuals with extreme Mahalanobis distance in the joint distribution of biomarkers 185
S2.7 Enrichment of ICD-10 codes in individuals with low Mahalanobis *P*-values 186

S3.1 Liability threshold model for balancing of common-variant risk and rare variant burden effects 200
S3.2 Schematic diagram of evaluating the observed test statistic against empirical null distribution of test statistics 201
S3.3 Power analysis of burden effects associated with lower PRS in disease controls 202

S4.1 MAD thresholds, split by tranche 215
S4.2 Gene burden effects across all nine obesity and fat distribution traits for 19 genes with exome-wide significant burden associations 216

S4.3 Ultra-rare variant allelic series for all FDR≤1% significant genes . . 217
S4.4 PheWAS of obesity and fat distribution associated genes using Genebass summary statistics 218
S4.5 Sex-differential and sex-specific analysis 219
S4.6 Longitudinal obesity age-at-onset analysis 220
S4.7 Longitudinal mRNA expression in wild type human white adipose tissue of genes selected for knockdown 221

List of Tables

2.1 Percentage of individuals outside of clinical reference ranges for blood and urine biomarkers . . . 23
2.2 Outside-range contingency tables for blood lipid biomarkers . . . 29
2.3 Contingency tables for low-density lipoprotein cholesterol (LDL-C) and apolipoprotein B (ApoB) . . . 29
2.4 Variant-level quality control (QC) median absolute deviation thresholds 54
2.5 Contingency table for testing effect direction of low- and high-tail associations . . . 59
2.6 Parameters for simulating vQTL . . . 60

3.1 Monogenic disorder genes for continuous phenotypes . . . 76
3.2 Canonical monogenic genes tested for predicted loss-of-function (pLoF) enrichment in disease traits . . . 81
3.3 Regression equations for testing singleton burden associated with continuous and dichotomous trait misalignment . . . 110

4.1 Exome-wide significant gene-level associations . . . 125

S2.1 Biomarker overview table . . . 188
S2.2 Correlation and odds of overlapping outliers . . . 189
S2.3 Odds of International Classification of Diseases, 10th revision (ICD-10) codes with outlying biomarkers . . . 190
S2.4 Curation of biomarkers with GEL disorders . . . 191
S2.5 Extreme biomarker GEL rare disorder enrichment tests . . . 192
S2.6 Significant tail-status and continuous biomarker associations . . . 193
S2.7 Odds of ICD-10 codes with Mahalanobis distance . . . 194
S2.8 Mahalanobis distance gene burden associations . . . 195
S2.9 Single-biomarker gene burdens associations . . . 196

S3.1 Phenotype/PGS overview . . . 204
S3.2 Canonical monogenic/GEL disorder gene enrichment – Continuous traits . . . 205
S3.3 Canonical monogenic gene enrichment – Dichotomous traits . . . 206

S3.4 GEL disorder gene enrichment – Dichotomous traits 207
S3.5 Singleton burden enrichment – Continuous traits 208
S3.6 Singleton burden enrichment – Dichotomous traits 209
S3.7 Exome-wide scan for protective genes – Continuous traits 210
S3.8 Exome-wide scan for pathogenic genes – Continuous traits 211
S3.9 Exome-wide scan for protective genes – Dichotomous traits 212

S4.1 Sample removal based on sample outliers defined by median absolute deviation (MAD) thresholds 223
S4.2 Sample sizes and phenotype standard deviations (SDs) for each adiposity-related trait (sex-combined and sex-stratified) 223
S4.3 Gene-level results for all false discovery rate (FDR)$\leq$ 1% significant genes 224
S4.4 Replication of UK Biobank (UKB)-significant genes in All of Us (AoU) 225
S4.5 Phenome-wide associations from Genebass 226
S4.6 Sex-specific and sex-differential gene-level results 227
S4.7 Obesity age-of-onset longitudinal analysis using Cox proportional hazards model 228
S4.8 Combined evidence 229
S4.9 Guide RNAs per gene target 230
S4.10RNA-seq differential expression to confirm knockdown 231
S4.11PCR primers 231
S4.12reverse transcription-quantitative polymerase chain reaction (RT-qPCR) differential expression to confirm knockdown 232
S4.13Effect of knockdown on lipid accumulation 233
S4.14Evidence of druggability for knockdown target genes 233
S4.15Genes differentially expressed between each knockdown and Cas9-empty cell line, with log FC and *P*-value 234
S4.16GSEA pathway analysis for RNA-seq of knockdown cell lines 235

List of abbreviations

AAC alternate allele count. 88

AAF alternate allele frequency. 40, 42, 43, 47

AAM age at menopause. 70, 72, 86, 89, 93, 94, 103, 104, 114, 163, 164

ADHD attention deficit hyperactivity disorder. 101

ALP alkaline phosphatase. 48, 52

ALT alanine aminotransferase. 48, 52

AoU All of Us. x, xviii, 9, 11, 115–117, 127, 133, 147, 165, 166, 174, 175, 225

ApoA apolipoprotein A. 28, 29, 32, 52

ApoB apolipoprotein B. xvii, 27–29, 32, 34, 36, 38, 52, 56, 62

AST aspartate aminotransferase. 42, 48, 52

BH Benjamini-Hochberg. 81, 84, 124, 145, 147

BMD bone mineral density. 70, 86, 89, 93, 101, 103, 105, 108, 113, 162, 169

BMI body mass index. 8, 70, 86, 87, 103, 104, 114, 119–122, 124–128, 133, 134, 142, 146, 147, 154, 156, 157, 165, 171

BODIPY boron dipyrromethene. 131, 135, 150, 152, 156, 157, 165

BUA broadband ultrasound attenuation. 101

CAD coronary artery disease. 69, 71, 78, 81–85, 88, 90, 91, 99, 101, 103, 105, 113, 114, 163

CADD Combined Annotation Dependent Depletion. 11, 121, 141, 142, 155

CDF cumulative distribution function. 107, 146

cDNA complementary DNA. 151

CI confidence interval. 22, 25, 26, 28, 30–34, 36, 38, 40, 42, 44, 46, 49, 59, 75, 77, 79, 81–88, 90–92, 99, 101, 107, 124, 125, 127, 130, 133, 157, 182–184, 186, 198, 199, 219

COAST Coding-Variant Allelic-Series Test. 171

CRISPR clustered regularly interspaced short palindromic repeats. ix, x, xiv, 117, 118, 120, 129, 133, 135, 148–150, 152, 153, 155–157, 165, 172

CRP C-reactive protein. 18, 40–42, 49

CVD cardiovascular disease. 18, 52, 53, 101, 176, 177

DMEM Dulbecco's Modified Eagle Medium. 149, 150

DNA deoxyribonucleic acid. 6, 7, 9

DXA dual-energy X-ray absorptiometry. 120, 171

ES exome sequencing. viii, x, 10, 11, 15, 22, 25, 39, 53–55, 58, 60, 64, 102, 111, 112, 117, 118, 120, 139–141, 146, 160, 165, 166, 173, 180

FBS foetal bovine serum. 149, 150

FC fold change. 118, 133–135, 137, 157

FDR false discovery rate. xviii, 68, 69, 75, 77, 78, 81, 82, 84, 87–90, 93, 96, 98–100, 108, 118, 124, 126–130, 133, 134, 145–148, 164, 198, 199, 214, 224

FH familial hypercholesterolaemia. 32, 34, 36, 37, 70, 75, 105, 115

FWER family-wise error rate. 135, 136

GATK Genome Analysis Toolkit. 54, 139

GEL Genomics England. viii, xiii, 15, 32, 36, 57, 63, 84, 86–88, 90, 91, 105, 109, 114, 170

GGT gamma glutamyltransferase. 40, 42

gnomAD Genome Aggregation Database. 53, 54, 110, 139, 175

GRM genetic relatedness matrix. 9, 108, 112, 142, 143

gRNA guide RNA. 130, 149

GWAS genome-wide association study. 8–11, 39, 47, 64, 98, 115, 119–121, 129, 138, 139, 144, 147, 148, 175

HbA1c glycated haemoglobin. 18, 33, 37, 38, 40, 41, 55, 64, 70, 72, 86, 93, 103, 104, 114, 161

HDL-C high-density lipoprotein cholesterol. 26, 28, 29, 32, 52

HLA human leukocyte antigen. 8

hWAT human white adipose tissue. 129, 130, 134, 135, 137, 148–150, 152, 153, 155–157, 229

ICD-10 International Classification of Diseases, 10th revision. viii, xiv, xvii, 15, 29, 31, 33, 49, 56, 57, 63, 103, 168, 170, 186, 190, 194

IOP intraocular pressure. 70, 103, 114, 169

IRNT inverse rank-normal transformation. 48, 60, 61, 65, 72, 111, 112, 114, 115, 143

LC low-confidence loss-of-function. 94, 141

LD linkage disequilibrium. 5, 7, 39, 108, 142, 144

LDL-C low-density lipoprotein cholesterol. xvii, 20, 27–29, 32, 36, 38, 56, 62, 68, 70, 72, 75, 77, 80, 86, 93, 96, 103–106, 113–116, 160, 162, 198, 199

LDSC LD score regression. 128, 147

LMIC low and middle-income countries. 177

LMM linear mixed model. 9

LOEUF loss-of-function observed/expected upper bound fraction. 88, 110

LoF loss-of-function. 70, 81, 82, 93, 99, 101, 104, 105, 120, 141, 155, 162, 163, 171

LOFTEE Loss-Of-Function Transcript Effect Estimator. 11, 57, 121, 141

MAC minor allele count. 50, 52, 54, 58, 96, 98, 100, 116, 118, 122, 124, 125, 127, 142, 143, 145, 148, 214

MAD median absolute deviation. xiv, xviii, 54, 139, 140, 215, 223

MAF minor allele frequency. 32, 39, 42, 47, 50, 57, 58, 60, 64, 68, 74, 77, 81, 84, 92, 107, 109, 111, 120, 124, 125, 129, 138, 139, 144–147, 167, 202, 213

MODY maturity-onset diabetes of the young. 81, 82, 104, 105

MRI magnetic resonance imaging. 120, 171

mRNA messenger ribonucleic acid. 129, 130, 133, 151, 153, 229

NHS National Health Service. xiii, 18, 20, 21, 24, 53, 57, 62, 63, 168

OFH Our Future Health. 11, 173

OP osteoporosis. 71, 78, 81, 83, 85, 90, 99, 101, 103–105, 113

OR odds ratio. 26–34, 36, 38, 40, 49, 56, 59, 75, 77–80, 82–86, 88, 91, 107, 186, 199

PC principal component. 54, 58, 106, 138, 143, 146

PGS polygenic scores. 12, 13, 68–72, 74, 75, 78, 79, 103, 104, 107–109, 113–115, 162, 197–199, 201

pLoF predicted loss-of-function. viii, ix, xiii, xiv, xvii, 16, 32, 36–38, 50–52, 55, 57, 60, 61, 63, 67–69, 74–90, 92, 93, 96, 98–100, 102, 104, 108–114, 116, 121, 125, 127, 129, 133, 134, 141, 144, 146, 147, 155, 162–164, 167, 171, 198, 199, 217, 218

POI primary ovarian insufficiency. 104, 105

PoPS Polygenic Priority Score. 124, 127, 146, 147, 164

PRS polygenic risk scores. xiii, xiv, 71, 78, 80–85, 88, 90–92, 98–104, 106, 109, 110, 112, 113, 116, 163, 176, 177, 197–200, 202

QC quality control. viii–x, xvii, 15, 25, 38, 53–58, 60, 64, 68, 102, 112, 117, 138–140, 143, 173, 182–184

RAP Research Analysis Platform. 60, 109, 111, 139, 144

REVEL Rare Exome Variant Ensemble Learner. 11, 121, 141, 142, 155

RF random forest. 53, 102, 138–140

RFLP restriction fragment length polymorphism. 6, 7, 12

RNA ribonucleic acid. x, 6, 118, 130, 133, 137, 150–153, 166

RT-qPCR reverse transcription-quantitative polymerase chain reaction. xviii, 130, 133, 151, 232

RVAS rare variant association study. 9–12

SAIGE Scalable and Accurate Implementation of GEneralized mixed model. 10, 50, 60, 61, 92, 99, 108, 111, 112, 125, 142–145, 167, 171, 202

SD standard deviation. xviii, 119, 124, 125, 223

SE standard error. 26, 46, 124, 125, 128, 129, 134, 145, 227

SEM standard error of the mean. 25, 124, 127, 133, 182–184

SES socioeconomic status. 176, 177

SHBG sex hormone-binding globulin. 33, 37, 55

SKAT-O optimal sequence kernel association test. 121, 124–126, 128, 129, 133, 134, 144, 145, 147, 214, 216, 218

SNP single nucleotide polymorphism. 7, 8, 138, 140, 144

SNV single nucleotide variant. 54

T1D type 1 diabetes. 101

T2D type 2 diabetes. 41, 42, 69, 71, 78, 81–83, 85, 88, 90, 91, 99, 102, 103, 105, 113, 114, 119, 163

UKB UK Biobank. vii, viii, xiii, xiv, xviii, 9, 11, 12, 15, 16, 19–25, 31–33, 38, 39, 43, 45, 47, 48, 53–56, 58, 61–63, 65, 66, 70, 71, 76, 78, 90, 102–106, 108, 114, 118, 120, 124, 127, 129, 138–140, 142–144, 146, 147, 154, 160, 161, 165, 166, 168, 169, 172–174, 225

VCF variant call format. 53, 139

VEP Variant Effect Predictor. 141

VOI variants of interest. xiii, xiv, 40–47, 182–184

vQTL variance quantitative trait loci. xvii, 38, 43, 45, 46, 59, 60

WGS whole-genome sequencing. 10–12, 166, 173, 174

WHR waist-to-hip ratio. 119

WHRadjBMI WHR adjusted for BMI. 119–122, 124, 125, 127–129, 146, 147, 154, 156, 157, 171, 219

1 Introduction

Contents

1.1 Preamble 1
1.2 A brief history of (statistical) genetics and its application to the study of disease 2
1.2.1 Early genetic theories 2
1.2.2 The foundational theories in evolution and inheritance . 3
1.2.3 The beginning of statistical genetics 4
1.2.4 The physical basis of inheritance 5
1.2.5 The molecular basis of genetics 6
1.2.6 Advancements in quantifying genetic variation 6
1.2.7 The indirect approach 7
1.2.8 The era of genome-wide association studies 8
1.2.9 Rare variant association studies and sequencing 9
1.3 Thesis aims 12
1.3.1 Big questions, big data 12
1.3.2 Aims 12

1.1 Preamble

My thesis was completed under the auspices of the Genomic Medicine and Statistics course. As such, my thesis lies at the intersection of three fields: genetics, medicine, and statistics. My purpose is to develop and apply statistical methods that use genetic data to better understand disease mechanisms–and ultimately

to improve medicine.

In this introductory chapter, I outline foundational theories in genetics, before explaining how modern statistics was born from the work of geneticists in the early 20th century. I go on to describe the developments in theory and technical methodology throughout the 20th century which, for the first time, enabled detailed study of the genetic underpinnings of human disease. Moving to the present day, I describe ever expanding collections of genetic data and statistical tools we have available to us for the study of human disease.

In subsequent chapters, I harness the latest data and statistical tools to study human disease. I also develop my own statistical methods and discover new insights, providing meaningful contributions to the intersection of genetics, medicine, and statistics.

1.2 A brief history of (statistical) genetics and its application to the study of disease

1.2.1 Early genetic theories

From the beginning, humans have sought to understand the mechanisms of our origin. The field of genetics (from the Greek *genesis*, meaning "origin") is the scientific lens through which we have been able to answer such questions of self. Genetics is rooted in the theory of inheritance: offspring inherit traits from their parents. The ancient Greek philosopher/physician Hippocrates believed that during the reproductive process each part of a parent's body contributed a particle and these particles would fuse together to form a child who exhibited a blend of the parental traits[1]. The assumption of blending inheritance implied that offspring would be similar to their parents. Within this framework, it would be difficult to imagine how species might evolve. Yet it was another ancient Greek philosopher, Anaximander, who could conceivably be credited with the earliest theory of evolution[2]: he believed that the origin of all living creatures was moisture and that the human species originated as embryos inside fish-like creatures[2].

The first person to propose a comprehensive scientific theory for evolution was Frenchman Jean-Baptiste Lamarck[3]. Lamarck theorised that acquired characteristics, due to use or disuse, in an organism could be inherited across generations[3,4]. Importantly, Lamarck believed that "environment affects the shape and organisation of animals", and therefore, changes in the natural environment were the reason for the evolution of species[5,6].

1.2.2 The foundational theories in evolution and inheritance

Fifty years later, building on Lamarck's theory that evolution was influenced by an organism's environment, Charles Darwin published his landmark book, *On the Origin of Species*. Darwin hypothesised that the same artificial selection of certain traits in domestic animals and plants could be replicated by selective pressures in nature[7]. Darwin's theory of so-called "natural selection" was used to explain how species evolved and now forms the backbone of modern evolutionary biology. But in Darwin's era there remained the question of genetic transmission: What are the principles by which genetic material is transmitted from parent to offspring?

Around this time, a friar named Gregor Mendel was growing peas in an abbey in what is now Brno, Czech Republic. By hybridising peas with various dichotomous characteristics (e.g. round or wrinkled seeds, yellow or green seed albumen), Mendel observed that the hybrid progeny was more likely to exhibit one form of a dichotomous trait more than the other, with observed "dominant" forms outnumbering the "recessive" forms with consistent ratios[8,9]. Mendel's experiments led to three laws of inheritance: 1) Dominance: Dominant forms override recessive forms of a trait, 2) Segregation: Maternal and paternal factors (now called "alleles") are randomly transmitted to progeny, and 3) Independent assortment: Genetic factors affecting different traits are inherited independently[8,9]. The principles from these experiments, proposed in 1865[8], have been likened to the missing piece of Darwin's puzzle of natural evolution[10], but the publication largely went unnoticed by the broader European scientific community.

It was not until 1900 that Mendel's work was rediscovered by Carl Correns and Hugo De Vries, who performed similar plant hybridisation experiments[10]. Mendelian genetics was popularised in the English-speaking world by William Bateson shortly after the rediscovery[10,11]. Bateson also happened to be the first to coin the term "genetics"[10].

1.2.3 The beginning of statistical genetics

In contrast to Mendel's study of dichotomous traits, a group of English scientists (who called themselves "biometricians") were attempting to explain the genetics underlying continuous traits in humans. These scientists used relatively large studies of thousands of related individuals to inform their theories on inheritance of continuous traits, with a particular focus on standing height[12]. One of the leading figures in the biometrician school of thought was Francis Galton (coincidentally Charles Darwin's half-cousin[13]). In his 1886 paper "Regression Towards Mediocrity in Hereditary Stature" Galton explained why height is an ideal phenotype for study: height can be easily measured, is frequently measured, and is a stable measurement through most of our adult life[12]. Importantly, Galton was aware that height was complex, the aggregate of many constituent parts. Galton identified that the sum of many elements is precisely why height as a continuous variable is distributed in a population with "beautiful regularity"[12]. Galton and his acolytes, including his protégé Karl Pearson (of correlation fame[14]), observed that adult offspring tended to have stature that approximated the sex-adjusted averaged stature of their parents[12] and that the correlation in quantitative traits between parent and offspring was around 0.5[15,16]. These observations appeared to support the theory of blending inheritance, which was at odds with Mendel's laws of discrete inheritance.

It took until 1918 for Ronald Aylmer Fisher to propose a model that reconciled Mendelian inheritance with complex traits[17]. Fisher hypothesised that multiple genetic factors (now called "genes") could contribute to the genetic component of a complex trait, with each gene inherited under Mendelian principles[18]. Just as Galton had once observed that variability in height was the aggregate of many

constituent parts, Fisher generalised this to any complex trait[18]. The research of Galton, Pearson, and Fisher brought not just new genetic theories but new statistical methods too. In their study of heredity, they concurrently developed or inspired many of the essential statistical tools we use today, including regression, correlation, and variance[12,15,17]. This marked the beginning of modern statistics[19,20], and with it, the field of statistical genetics.

1.2.4 The physical basis of inheritance

Meanwhile, across the pond, American researchers were studying the chromosomes of insects in the hopes of understanding the physical basis of inheritance. Chromosomes had first been observed by the German scientist Walther Flemming in 1882[21]. But it was unclear at the time exactly what purpose chromosomes fulfilled in the inheritance of genetic material. In 1902, an undergraduate student at Columbia University named Walter Sutton demonstrated that chromosomes obey Mendelian rules of inheritance[22,23]. In 1910, Thomas Hunt Morgan demonstrated that inheritance of eye colour in fruit flies was linked with a "sex factor 'X"', what we now call the X chromosome[24,25]. This was the first time that a gene was localised to a chromosome[24].

The following year, Morgan proposed that if genes were linearly arranged on chromosomes and homologous parental chromosomes split and exchange segments by "crossing over", then genes in close proximity on the same chromosome would be more likely to be inherited together[26]. This property has come to be known as linkage disequilibrium (LD). Applying this theory of LD, Alfred Sturtevant, one of Morgan's students, used the observed rate of crossovers in fruit fly offspring as an index for genetic distance and created the first linear map of genes (now called a "linkage map") in a chromosome[27]. Sturtevant's linkage analysis provided the theoretical framework to map the relative locations of any genes located on the same chromosome. However, in the 70 years following Sturtevant's work, linkage mapping was limited to biological markers that had a biochemical link to the trait being studied, such as the beta-globin gene/protein and sickle cell disease[28–31]. To map all human genes, scientists needed a method to quantify genetic differences

(determine the "genotype") at loci without easily-measured biological markers. As luck would have it, there would be major discoveries in the field of molecular biology during the middle of the century that would make this possible.

1.2.5 The molecular basis of genetics

In 1953, Watson and Crick were the first to propose an accurate model for the molecular structure of deoxyribonucleic acid (DNA)[32]. In a lecture four years later[33], Crick proposed a unifying "central dogma" of molecular biology in which he hypothesised the flow of encoded genetic information from DNA to ribonucleic acid (RNA) to protein[34]. The notion that proteins are composed of linear combinations of amino acids and, in particular, that the order was important to the function of the protein was first shown by Frederick Sanger a few years before[35–37]. Therefore, being able to determine the sequence of nucleotide bases in DNA, which held the information to make RNA and thus protein, was of great importance. However, early sequencing methods were time consuming[38]. In 1977, two sequencing methods were published which dramatically improved efficiency[39,40]. Of the two, Sanger's "chain-termination" approach (now "Sanger sequencing") has endured to this day as a common sequencing method[41].

1.2.6 Advancements in quantifying genetic variation

Although Sanger sequencing enabled highly efficient nucleotide-level resolution of DNA, efforts to map the human genome largely focused on another approach to quantify variation in the genome. In 1980, Botstein *et al.* demonstrated the method of restriction fragment length polymorphisms (RFLPs), which used restriction enzymes to target cleavage at specific sequences in human DNA[42]. Depending on variations in a person's DNA, the locations of the cleavages might be different for different individuals, resulting in a polymorphic marker that captures variation in the targeted region of DNA. Notably, RFLPs enabled systematic study of polymorphic loci which were not directly associated with an expressed product, such as loci in non-coding regions of the genome[43].

Three years later, a polymorphic DNA marker linked with Huntington's disease was identified[44], a discovery directly attributable to RFLPs. In the same year as the Huntington's disease marker discovery, a new method for isolating the location of a gene without requiring knowledge of its biochemical function was published[45], called positional cloning[46]. Positional cloning (also called chromosome walking and jumping) paired with linkage analysis enabled key discoveries, such as the cystic fibrosis gene in 1989[47] and the association of the Apolipoprotein E type 4 allele with increased Alzheimer's disease risk in 1993[48]. As useful as these strategies were to mapping the Mendelian mechanisms of rare disease (mapping more than 400 rare diseases by 1994[49–51]), there was some concern that linkage analysis was not suited to modelling more complex traits[50]. Nonetheless, positional cloning with RFLPs continued apace and led to the first genome-wide genetic maps[43]. This, along with high-throughput automated Sanger sequencing[38], enabled the Human Genome Project to sequence the whole human genome for the first time, releasing a draft sequence in 2001[50].

1.2.7 The indirect approach

Around this time, there was some debate on the best approach to discovering new genetic associations[52], especially for common complex diseases[45,52,53]. The direct approach involved testing functional variants in coding regions of the genome[52,53]. The indirect approach involved testing an array of variants spaced across the genome, not necessarily located in coding regions[45,52]. The key to the indirect approach was LD: rather than focusing on a specific causal variant, one could instead focus on the haplotype block of variants in which the causal variant was inherited. It was believed that the indirect approach would be more efficient at mapping the common human sequence variation expected to contribute to common disease[54]. Thus, the International HapMap Consortium was formed to characterise worldwide common sequence variation[54]. Understanding of common variation, coupled with technological advancements in large-scale assays for efficiently genotyping single

nucleotide polymorphisms (SNPs)[55,56], led to a new era of genetic association studies, dominated by the genome-wide association study (GWAS).

1.2.8 The era of genome-wide association studies

In 2005, Klein *et al.* tested 103,611 SNPs for association with age-related macular degeneration, using 96 cases and 50 controls[57], considered by some to be the first GWAS[58]. Two years later, the first large-scale validation of GWAS was performed by the Wellcome Trust Case Control Consortium, testing associations across seven common diseases in a British cohort of $\sim 17,000$ individuals[59].

In the years that followed there would be hundreds of GWAS. Most variants ($\sim 90\%$[60]) discovered by GWAS were in non-coding regions, a surprising finding to some researchers[61]. Linking non-coding GWAS loci to genes proved difficult. One instance which perhaps best exemplifies this challenge was the process uncovering the mechanism of association for the *FTO* gene with obesity. Despite being the strongest association in an early GWAS of body mass index (BMI), an obesity related trait, in 2007[62], it took eight years for researchers to implicate the *FTO* locus as a cis- for *IRX3* and *IRX5*, two developmental regulators with adipocyte-specific effects[63]. This enormous undertaking required a combination of *in silico* and *in vitro* approaches, until eventually demonstrating the causal link. And while there were other notable successes (e.g. human leukocyte antigen (HLA) locus with schizophrenia[64,65]) the challenge of linking non-coding loci to genes largely remained.

During this period there were several key developments in the field of statistical genetics. Statistical imputation of genotypes, which uses densely genotyped reference panels (such as the 1,000 Genomes Project[66]) to infer genotypes in a sparsely genotyped cohort, enabled GWAS to interrogate many more variants without the cost of densely genotyping all individuals[67,68]. Polygenic scoring enabled prediction of traits, especially disease, using common genetic variation[69,70]. Heritability estimation methods enabled the calculation of phenotypic variability attributable to GWAS loci[71,72].

Since then, the continued interest in GWAS as a tool for discovery has been fuelled by increases in sample size, mostly due[73] to the creation of large-scale biobanks, such as the UKB[74], FinnGen[75], Biobank Japan[76], Million Veteran Program[77], deCODE[78], Estonian Biobank[79], Lifelines Cohort[80], China Kadoorie Biobank[81], and AoU[82], to name but a few. Of these major biobanks, UKB has led the way in genetic discovery[73,74,83] due to its large size (half a million individuals) and deep phenotyping (including physical measurements, lifestyle questionnaires, electronic health records, blood and urine biomarkers[84,85], proteomic profiles[86], and imaging data[87,88]) paired with genetic data (including genotype array data and imputed genotypes).

To perform association tests on these large datasets, well-powered and well-calibrated computationally efficient methods were required. One solution was linear mixed models (LMMs). LMMs partition phenotype variance into components, using a random effect with the genetic relatedness matrix (GRM) and fixed effects with the genetic marker being tested, covariates, and residual environmental variance. Using the random effect with the GRM enables control of type I error in the presence of population stratification (including family structure) and cryptic relatedness, two common issues with GWAS[50,89,90]. Early methods employing mixed models were computationally intractable for biobank-scale GWAS[91], but several computationally optimised mixed model approaches have emerged for biobank-scale association testing since then[91–94].

1.2.9 Rare variant association studies and sequencing

Exome sequencing

Parallel to common variant GWAS efforts, rare variant association study (RVAS) were building momentum in the early 2000s. Unlike GWAS, RVAS required DNA sequencing in order to capture the much greater diversity in rare variants compared to common variants[95]. A key advancement in sequencing technology, so-called "next-generation sequencing"[96], enabled massively parallel sequencing and laid the path to large-scale RVAS.

Early RVAS favoured exome sequencing (ES) over whole-genome sequencing (WGS) because ES was a fifth of the cost[95], due to its specific targeting of protein-coding regions. Furthermore, protein-coding regions were expected to harbour variants with higher effect sizes and greater functional interpretability than those in non-coding regions, making ES a more efficient approach than WGS to increasing sample sizes of sequencing studies[97].

Collapsing tests

Sequencing aside, there are other important differences between RVAS and GWAS. Statistical power is low for single-variant association tests in rare variants, compared to common variants with the same effect size[98]. To boost power, rare variants are typically grouped together using "collapsing" tests[99]. Collapsing tests can broadly be separated into three categories: burden tests, variance-component tests, and combined tests. Burden tests were the earliest version[100–102] of a collapsing test and assume that grouped variants are causal and have the same direction of effect[99]. Variance-component tests have better power than burden tests when grouped variants have mixed directions of effect or when fewer variants are causal[99]. Combined (also called hybrid or optimal) tests are designed to combine the power of burden and variance-component tests[99]. Throughout this thesis I use burden and combined tests implemented by the mixed model association testing method Scalable and Accurate Implementation of GEneralized mixed model (SAIGE)[103], optimised to reduce computational cost in biobank-scale analyses. Although combined tests theoretically have optimal power when the underlying model of a gene is unknown[104], burden testing can yield better power when there is dense causality and unidirectional effects in the gene. Burden testing also yields an effect size, which can make the result more interpretable and enable comparison to variant-level association results.

Functional annotation

Another difference between RVAS and GWAS is the relative importance of functional annotation in RVAS. Grouping by annotation can refine signal-to-noise ratio and inform interpretation of results[99]. While there are continuing efforts to manually

curate the functional impact of variants using evidence from association tests (e.g. ClinVar[105]), this approach is not scalable to the large number of variants now available with ES and WGS data. Furthermore, classification of consequence can be inconsistent across curators. A 2015 study of ClinVar revealed that 17% of variants with multiple submitters had varying classifications[106]. Site-directed mutagenesis experiments offer an alternative approach[107], but are too expensive to predict variant consequences genome-wide[108].

For these reasons, computational predictors of variant consequence have become a common approach when annotating next-generation sequencing data[108]. In this thesis I use variant-level functional annotations generated by Frederik Lassen and Barney Hill, members of the Lindgren lab. Lassen and Hill used Rare Exome Variant Ensemble Learner (REVEL)[109], Combined Annotation Dependent Depletion (CADD)[110], and SpliceAI[111] to categorise damaging missense or protein-altering variants, with thresholds calibrated to the American College of Medical Genetics guidelines for pathogenic variants[112]. Lassen used Loss-Of-Function Transcript Effect Estimator (LOFTEE) guidelines[113] to annotate variants likely to cause complete loss of protein function. Throughout my thesis I use these annotations to group variants when performing collapsing tests.

Biobanks for rare variant association studies

Similar to GWAS, recent RVAS have been powered by increasingly large biobanks. Once again, UKB has led the way. In the last five years, UKB has been releasing sequenced exomes of participants in tranches. A final release in July 2022 brought the total to more than 450,000 exomes[114]. This ES dataset is the main source of genetic data for my thesis.

Despite releasing what was at the time the largest ever ES project, UKB soon followed with a release of WGS data for 500,000 participants the following year. There are several major WGS projects underway or planned around the world (AoU[82,115], China Kadoorie Biobank[81], PRECISE[116], Our Future Health[117]).

Already, large-scale RVAS are being performed with WGS data[118–120]. As sequencing costs continue to decrease[121], datasets will become even larger.

1.3 Thesis aims

1.3.1 Big questions, big data

As with any science, the field of genetics has been driven by big questions throughout its history: How do species evolve? What are the mechanisms of inheritance? How can we quantify heritability of traits? What is the physical basis of inheritance? How can we measure an individual's genetic material? What genetic markers are linked with disease? To answer these questions, scientists have used the data available to them, whether they be Darwin's observations of finches in the Galápagos Islands, Mendel's thousands of hybridised pea plants, measurements of Victorian era stature by Galton and Pearson, Morgan and Sturtevant's fruit fly experiments, RFLP genotyping for linkage studies, and now, biobanks containing rich phenotypic and genetic information for hundreds of thousands of individuals. With more data than ever, we can continue asking key questions to drive the field forward. In my thesis I aim to do just that.

1.3.2 Aims

The overall aim of my thesis is to enhance understanding of human disease mechanisms. I do this through the analysis of genetic and phenotypic data from the UKB, employing both established statistical approaches and novel methods developed in the course of this work. Specifically, I examine the genetic underpinnings of phenotypic deviation in various contexts: deviation of blood and urine biomarkers from clinical reference ranges, deviation of individuals from their expected phenotype predicted by polygenic scores (PGS), and the biological mechanism of fat accumulation, in which deviation from normality can cause disease:

- In Chapter 2, I begin at the phenotypic level, using extreme values of blood and urine biomarkers as a simplified model of disease to explore their associations

with clinical outcomes. I then integrate genetic data to examine the signals associated with these extreme biomarkers.

- In Chapter 3, I explore the interplay between rare and common variant components of disease-related phenotypes, testing whether rare deleterious variants can account for discrepancies in individuals whose observed phenotypes deviate from those predicted by their common-variant PGS.

- Finally, in Chapter 4, I narrow focus to a specific disease-related biological process: fat accumulation (adiposity), relevant to disorders such as obesity and lipodystrophy. I use common and rare variant genetic data to generate hypotheses for causal adiposity genes and validate these hypotheses through a systematic *in vitro* experiment.

Altogether, this thesis advances understanding of the impact of rare variants on extreme and unexpected phenotypic deviations, and contributes meaningful insights into the genetic aetiology of human disease.

2

The genetics of extreme blood and urine biomarkers

Contents

2.1 Abstract 16
2.2 Introduction 17
2.3 Main 20
2.3.1 Many individuals in UKB have biomarker levels outside of clinical reference ranges 20
2.3.2 Using extreme biomarker levels to prioritise individuals with rare genetic disorders for genetic screening 32
2.3.3 Discovering variants associated with extreme biomarker levels 38
2.3.4 Rare variant gene burden associated with multivariate biomarker extremes 48
2.4 Methods 53
2.4.1 Defining clinical reference ranges 53
2.4.2 Quality control of UKB exome sequencing data 53
2.4.3 Variant annotation 55
2.4.4 Selection of blood and urine biomarkers 55
2.4.5 Outlier overlap enrichment between pairs of biomarkers 56
2.4.6 Enrichment of ICD-10 codes among individuals with outlying biomarkers 56
2.4.7 Enrichment of predicted loss-of-function variants in Genomics England (GEL) PanelApp rare disorder genes among individuals with extreme biomarkers 57
2.4.8 Defining tail-status phenotypes 57
2.4.9 Single-variant tests on biomarker continuous and tail-status phenotypes 58

2.4.10 Testing odds of low- and high-tail associations having negative or positive effect on the continuous trait 58
2.4.11 Defining significant non-monotonic allele frequency . . . 59
2.4.12 Simulating vQTLs 59
2.4.13 Defining Mahalanobis distance on joint distribution of biomarkers 60
2.4.14 Gene-based burden tests on Mahalanobis distance and single biomarkers 60
2.5 Discussion 61

2.1 Abstract

Biomarkers indicate health status, but the prevalence and implications of abnormal (outside of the clinically-defined normal range) biomarker levels in the UKB remain underexplored. For the first time, I quantified the proportion of individuals with abnormal biomarker values across 32 blood and urine biomarkers in genetically European UKB participants and found that 96.4% had at least one abnormal value outside of a clinical reference range. Disease risk increased with the number of outlying biomarkers, highlighting their clinical relevance.

To investigate genetic drivers of abnormal biomarker levels, I tested for rare variant enrichment in individuals with extreme values and found significant enrichment of pLoF variants in genes linked to rare disorders, the first systematic validation of metabolic profiling for rare disorders in UKB. Contrasting traditional exome-wide association studies using continuous biomarkers versus using dichotomised (extreme vs. non-extreme value) phenotypes identified associations uniquely powered by the dichotomised approach, including a link between high serum triglyceride levels and a *PNPLA3* missense variant.

Recognising that some diseases disrupt multiple biomarkers, I used Mahalanobis distance to quantify how extreme each individual is in the joint distribution of biomarkers and performed gene burden testing. Many associated genes were linked to multiple individual biomarkers, validating this approach for detecting systemic genetic effects. This study demonstrates the clinical and genetic significance of

biomarker extremes and highlights novel phenotype definitions that enhance genetic discovery in large biobank cohorts.

2.2 Introduction

The term "biomarker" is the portmanteau of "biological marker" and is defined by the Biomarkers Definitions Working Group to be "a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention"[122]. Biomarkers have many roles. They can be an objective indicator for disease diagnosis, disease progression and severity, disease prognosis, and response to therapeutic intervention[122]. By definition, biomarkers can be simple measures such as pulse, body weight, or blood pressure, or involve sophisticated laboratory testing of patient samples, such as blood or urine, for specific molecular components. In this work, I focus on blood and urine biomarkers, which have become important clinical tools in the study of chronic disease.

The history of blood and urine biomarkers spans thousands of years. Possibly the first analysis of urine (urinalysis) was performed around 4000 BCE by Sumerian and Babylonian physicians who noted that urine could have different colours[123]. Around 400 BCE, Hippocrates (considered the "father of medicine") claimed that urine with bubbles on the surface was a marker of kidney disease[124]. His observation was prescient–now we know these bubbles are caused by high protein levels in urine, a symptom of kidney disease[125,126]. Another disease which features prominently in the history of urinalysis is diabetes mellitus, evident from its etymology: *mellitus* (Latin for "sweet") refers to the sweet taste of diabetic urine[127]. Ancient Ayurvedic medical textbooks also made reference to this symptom, characterising diabetes patients as having *madhumeha* (Sanskrit for "honey urine")[123]. In 1776, Liverpudlian physician Matthew Dobson proved that the sweetness was sugar and that the blood serum of diabetic patients was also sweet–the first reported observation of elevated blood glucose levels (hyperglycaemia)[128,129].

Around this time, quantification of biomarkers was becoming a science. In 1791, Antoine François Fourcroy was the first to propose a clinical laboratory for chemical analysis of patient samples[130]. At this stage, laboratories were simple and primarily used for research or teaching rather than diagnosis[130]. It took until the late 1800s for a full clinical laboratory, intended for patient diagnosis and with capabilities for chemical, physical and bacteriological analysis, to be set up in Munich[131].

The 20th century saw the establishment of key biomarkers as risk factors for common complex diseases and conditions, such as cholesterol for cardiovascular disease (CVD)[132] and glycated haemoglobin (HbA1c)[133,134] for diabetes mellitus, and C-reactive protein (CRP) for inflammation[135].

In the latter half of the 20th century there was a movement to standardise biomarker measurements. At a conference in 1966, Ralph Gräsbeck and Nils-Erik Saris proposed that better control values should be defined for biomarkers, replacing the existing notion of "normal values" with "reference values"[136]. An expert panel was formed and twenty years later the theory of reference values was codified by the International Federation of Clinical Chemistry[137]. Part of this theory included the definition of reference intervals, which delimit a range of reference values, typically such that a certain percentage (e.g. 95%[138]) of reference values fall within the interval[137]. Gräsbeck stressed the importance of not conflating reference intervals with decision limits, which define limits at which clinical decisions are made (e.g. HbA1c greater than 6.5%, equivalent to 48 mmol/mol, is the recommended decision limit for diagnosis of diabetes[139]). However, given that not all biomarkers have decision limits, in my work I refer to all biomarker limits (whether decision limits or reference intervals) as "reference ranges"; this is also the practice of clinical handbooks provided by the National Health Service (NHS)[140,141]. For simplicity, I refer to observed values which fall inside reference ranges as "normal" (i.e. healthy; not to be confused with Gaussian) and those outside as "abnormal" (i.e. not healthy), although the definition of "healthy" can be debated at length (see Gräsbeck[136]).

In the 21st century, blood and urine biomarkers can now be measured at scale with high-throughput screening assays. This has enabled the generation of massive

biomarker datasets, such as the UKB[74], for which biological samples of half a million individuals were screened for 34 blood and urine biomarkers[85,142]. The UKB also has genetic data–its combination of deep phenotyping and genetic data has made it a rich source for genetic association studies[73]. Using UKB, the rare and common variant genetic basis of biomarkers has already been explored in depth[85,143–145]. However, these studies in UKB focused on continuous variation. Given that extreme biomarker levels are generally associated with disease, I propose studying individuals in the extremes of biomarker distributions.

In this chapter, I systematically investigate extreme biomarker levels in the UKB, testing for association with disease and genetic factors using univariate and multivariate approaches. First, I evaluate the extent to which UKB individuals have normal or abnormal biomarker levels, as defined by clinical reference ranges. Second, I examine whether individuals who deviate in the multivariate distribution of biomarkers are enriched for disease, and if so, for which diseases. I quantify deviation in the multivariate distribution using Mahalanobis distance[146], a metric previously examined for multivariate analysis of biomarkers[147,148]. Third, given that metabolomic profiling has shown utility in aiding diagnosis of rare genetic disease[149], I test whether individuals with extreme biomarker levels are enriched for variants which cause rare genetic disorders. Fourth, I perform a genetic association study on single biomarkers, with continuous biomarker values dichotomised into extreme and non-extreme groups. This approach represents a departure from previous studies of biomarkers in UKB, which test biomarkers in their continuous form[85,143–145,150]. Finally, I perform a genetic association study on an individual's deviation in the multivariate distribution of biomarkers, calculated using Mahalanobis distance, and identify genes associated with widespread disruption to an individual's biomarker profile.

2.3 Main

2.3.1 Many individuals in UKB have biomarker levels outside of clinical reference ranges

I first sought to understand how biomarker levels of individuals in UKB compared to clinical reference ranges, collating clinical reference ranges for the 32 biomarkers provided by the NHS (Methods; Figure 2.1; Table 2.1; Table S2.1). Where available, I used sex- and age-specific reference ranges (Methods). Throughout, I defined individuals as reference range "outliers" for a biomarker if they had at least one measurement outside of the reference range.

For most biomarkers, the majority of individuals have normal biomarker levels

For nearly all (29/32) biomarkers, the majority of individuals were within reference range bounds (Figure 2.2, Table 2.1). This is expected, given that UKB is healthier on average than the general population[151]. However, there were three biomarkers where the majority of individuals lay outside of the reference range: LDL-C (outlying percentage=73.5%), total cholesterol (73.3%) and urine creatinine (52.0%) (Figure 2.2a).

For LDL-C, the vast majority of outlying individuals lie above ($n = 129,586$) rather than below ($n = 16$) the reference range ($1 - 3$ mmol/L). The percentage of outlying individuals was very similar to total cholesterol, which makes sense given that LDL-C is a component of total cholesterol and tightly correlated ($R = 0.952$; Figure S2.1; Table S2.2). Total cholesterol does not have a lower limit, so all outlying individuals for total cholesterol were those above the reference range. These high percentages of LDL-C and total cholesterol are unexpected, given that the United Kingdom is not among the European countries with highest levels of total cholesterol[152] and individuals in UKB tend to be healthier than the general population[151]. One factor that may influence this high percentage of outliers is that participants in the UKB were aged 40 to 69 at recruitment[74], meaning that female participants would be either entering menopause or be postmenopausal. LDL and

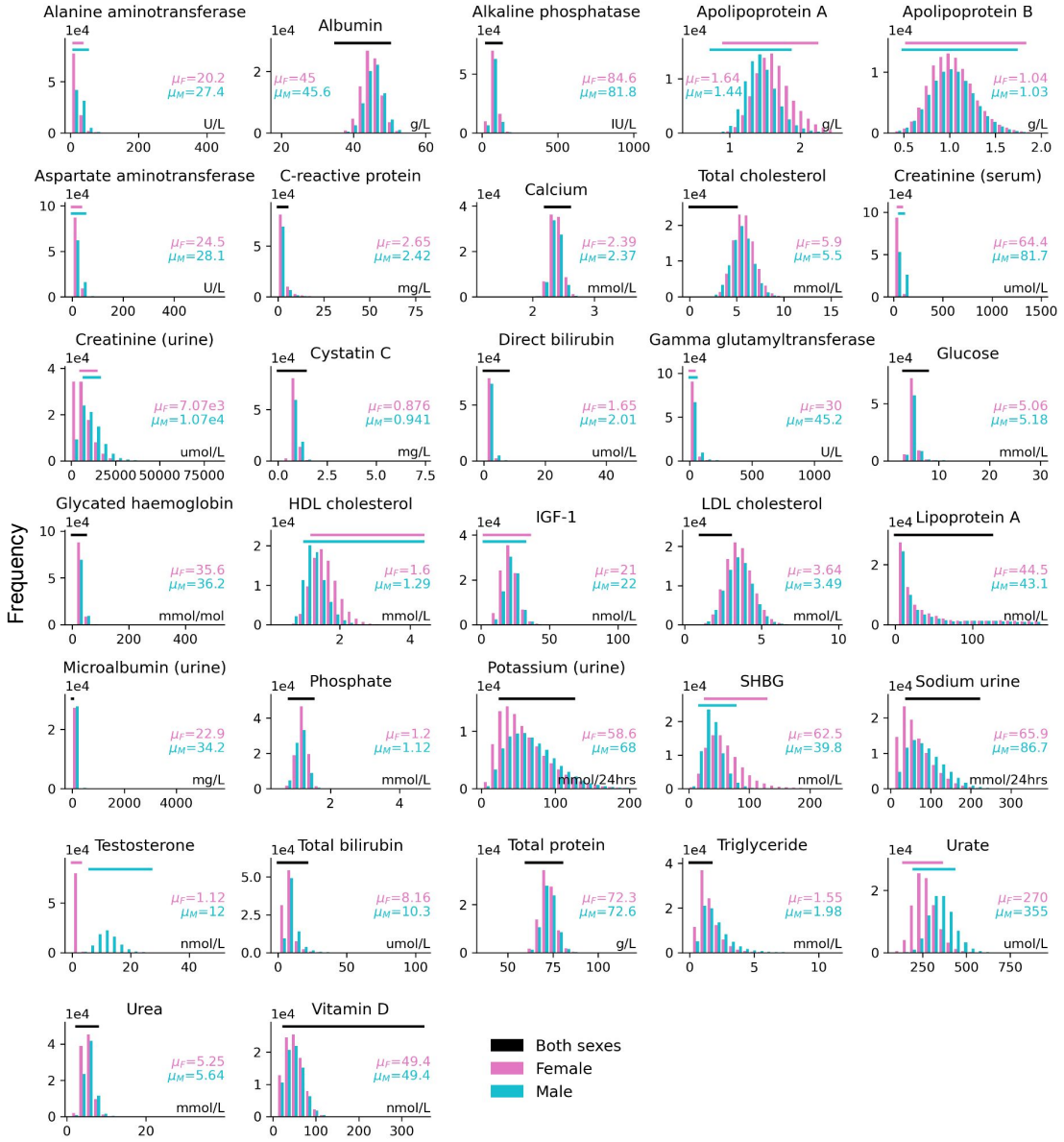

Figure 2.1: Raw UKB blood and urine biomarker distributions relative to NHS reference ranges. Sex-stratified histograms of the raw biomarkers are shown. Histogram bins were defined on the sex-combined biomarker distribution, using 20 equally spaced bins, and then sex-stratified counts in each bin were calculated. Histogram bars corresponding to female and male counts are plotted side by side in each bin, with the female bar (pink) always to the left of the male bar (cyan). Clinical reference ranges are indicated by the horizontal bars above the distributions. Where possible, sex-specific reference ranges were used ("Female", in pink; "Male", in cyan). Otherwise, sex-agnostic ("Both sexes", in black) reference ranges were used. Units of the raw biomarker are indicated in the bottom right of each plot. The sex-specific means (μ_M, μ_F) are shown to the left or right of each distribution. See Table S2.1 for an overview of phenotype-level statistics.

total cholesterol have been shown to increase during menopause and be higher after menopause[153–155]. Furthermore, the majority (55.1%) of the individuals in the ES 200k European subset were female, which may further upwardly bias the percentage of outliers.

For urine creatinine, there were outlying individuals above and below the reference range limits (Female 5200-14100 µmol/24hrs, Male: 7100-15900 µmol/24hrs[156]), with more than three times as many below the limit (39.2%) than above (12.8%) (Figure 2.2). Urine creatinine is a marker for muscle mass, with low creatinine levels indicating lower muscle mass[157–160]. There is a well-known association between aging and muscle loss[161,162], with decline potentially starting as early as 30 years of age[163]. Given that UKB participants were recruited at ages (40-69 years[74]) where muscle mass is already expected to be in decline, this may explain the high proportion of individuals with low urine creatinine.

Outlying percentage is high compared to using standard 95% reference ranges

Across biomarkers, the mean percentage of individuals below the reference interval was 5.07% (95% confidence interval (CI): [2.19, 7.96]), while the mean percentage above the reference range was 10.8% (95% CI: [4.42, 17.2]) (Figure 2.2b). Estimates of lower tail percentages were marginally higher than the 2.5% expected from a classic derivation of a 95% CI reference interval derived from healthy individuals[164], and estimates of upper tail percentages were significantly higher, with a 95% CI that did not include 2.5%. One explanation for this may be that 95% CI reference ranges are typically calculated using healthy individuals[164,165]. The percentage of outlying individuals in UKB may be inflated by the presence of disease and the late-adulthood age profile of participants. While age-specific reference ranges were used where available, they are not available for all biomarkers, and yet distributions of some biomarkers naturally shift with age (e.g. triglyceride[166], urine creatinine[167,168]). Additionally, not all reference ranges follow the 95% principle–for instance, thresholds for triglycerides[169] and vitamin D[170] are biologically-informed

Biomarker (abbreviation)	Units	Clinical reference range			% Outliers below/above/outside range (n)		
		Female	Male	Sex-agnostic	Below range	Above range	Outside range
Alanine aminotransferase (ALT)	U/L	10–35	10–50	-	3.3 (5,828)	6.59 (11,639)	9.89 (17,467)
Albumin (Alb)	g/L	-	-	35–50	0.0702 (114)	3.63 (5,902)	3.70 (6,016)
Alkaline phosphatase (ALP)	IU/L	-	-	30–130	0.125 (220)	3.75 (6,621)	3.87 (6,841)
Apolipoprotein A (ApoA)	g/L	0.91–2.24	0.73–1.86	-	0.0434 (70)	3.48 (5,622)	3.53 (5,692)
Apolipoprotein B (ApoB)	g/L	0.53–1.82	0.49–1.73	-	0.588 (1034)	0.338 (594)	0.926 (1,628)
Aspartate aminotransferase (AST)	U/L	0–35	0–50	-	0 (0)	4.47 (7,876)	4.47 (7,876)
C-reactive protein (CRP)	mg/L	-	-	0–5	0 (0)	11.2 (19,719)	11.2 (19,719)
Calcium (Ca)	mmol/L	-	-	2.2–2.6	1.86 (3,022)	1.7 (2,765)	3.56 (5,787)
Creatinine, serum (CRE)	µmol/L	45–84	59–104	-	1.65 (2914)	4.46 (7,878)	6.11 (10,792)
Creatinine, urine (UCRE)	µmol/24hrs	5,200–14,100	7,100–15,900	-	39.2 (70523)	12.8 (23,070)	52.0 (93,593)
Cystatin C (CysC)	mg/L	-	-	0–1.4	0 (0)	1.06 (1,866)	1.06 (1,866)
Direct bilirubin (DBil)	µmol/L	-	-	0–8	0 (0)	0.0633 (95)	0.0633 (95)
Gamma glutamyltransferase (GGT)	U/L	0–38	0–55	-	0 (0)	18.9 (33,433)	18.9 (33,433)
Glucose (Glu)	mmol/L	-	-	3–7.8	0.154 (250)	2.2 (3,566)	2.35 (3,816)
Glycated haemoglobin (HbA1c)	mmol/mol	-	-	0–48	0 (0)	3.24 (5,729)	3.24 (5,729)
HDL cholesterol (HDL-C)	mmol/L	1.2–inf	1–inf	-	14.6 (23,634)	0 (0)	14.6 (23,634)
Insulin-like growth factor 1 (IGF-1)	nmol/L	2.2–35.5	2–32.2	-	0.00114 (2)	2.35 (4,135)	2.35 (4,137)
LDL cholesterol (LDL-C)	mmol/L	-	-	1–3	0.00907 (16)	73.5 (129,586)	73.5 (129,602)
Lipoprotein A (Lp(a))	nmol/L	-	-	0–125	0 (0)	11.2 (15,749)	11.2 (15,749)
Microalbumin, urine (UMIC)	mg/L	-	-	0–20	0 (0)	23 (12,745)	23.0 (12,745)
Phosphate (PHO)	mmol/L	-	-	0.8–1.5	1.37 (2,220)	1.71 (2,772)	3.08 (4,992)
Potassium, urine (UPOT)	mmol/24hrs	-	-	25–125	9.95 (17,889)	5.48 (9,848)	15.4 (27,737)
Sex hormone binding globulin (SHBG)	nmol/L	27–128	18–77	-	6.41 (10,309)	3.33 (5,354)	9.73 (15,663)
Sodium urine (USOD)	mmol/24hrs	-	-	40–220	22.9 (41,185)	0.421 (756)	23.3 (41,941)
Testosterone (TES)	nmol/L	0–2.9	6–27	-	1.35 (2,156)	0.706 (1,127)	2.06 (3,283)
Total cholesterol (CHO)	mmol/L	-	-	0–5	0 (0)	73.3 (129,388)	73.3 (129,388)
Total bilirubin (TBil)	µmol/L	-	-	0–21	0 (0)	2.59 (4,560)	2.59 (4,560)
Total protein (TPro)	g/L	-	-	60–80	0.0869 (141)	3.64 (5,909)	3.73 (6,050)
Triglyceride (TG)	mmol/L	-	-	0–1.7	0 (0)	39.9 (70,475)	39.9 (70,475)
Urate (Urate)	µmol/L	140–360	200–430	-	0.756 (1,334)	11.3 (19,949)	12.1 (21,283)
Urea (Urea)	mmol/L	-	-	2.5–7.8	0.325 (573)	4.54 (8017)	4.87 (8,590)
Vitamin D (VitD)	nmol/L	-	-	25–inf	12 (20,324)	0 (0)	12.0 (20,324)

Table 2.1: Percentage of individuals outside of clinical reference ranges for blood and urine biomarkers. Reference ranges with no upper limit are indicated with infinity ("inf") in the upper limit. Where possible, sex-specific clinical reference ranges were used. Where possible, age-specific reference ranges relevant to the UKB participant age range were used (Methods; Table S2.1)

decision limits rather than statistically defined reference intervals, leading to higher-than-expected outlying rates (triglycerides: 40.0%; vitamin D: 12.0%; Figure 2.2a).

Sex differences in outlying percentages

Male and female percentages for above- and outside-range outliers was nearly a 1:1 correspondence (Figure 2.3). This suggests that biological differences between sexes in the UKB cohort, such as women undergoing menopause, generally do not skew percentages for above- and outside-range outliers. However, for below-range outliers, I observed a disproportionately higher proportion of female outliers, on average (Figure 2.3). Using sex-specific reference ranges, urine creatinine displayed the largest discrepancy between female and male below-range outliers (13.5 percentage points higher for females) and above-range outliers (8.7 percentage points higher for males) (Figure 2.3a). This may indicate that the urine creatinine sex-specific ranges are not well-calibrated for the age profile of UKB participants, who tend to be older. Previous studies have shown that urine creatinine distributions are a function of both age and sex[167,168], with higher levels in males and lower levels in older age groups. However, the available NHS reference ranges for urine creatinine only account for sex.

Among sex-agnostic biomarker ranges, urine sodium displayed the largest discrepancy between the sexes (15.9 percentage points higher for females) (Figure 2.3b). Point estimates for mean urine sodium excretion have been reported to be lower in British women compared to men[171]. This result bears out in UKB: I observe a significantly lower mean in females than males in UKB participants (two-sided t-test $t = -105$, $P < 1 \times 10^{-100}$) (Figure 2.1). This might support the implementation of sex-specific reference ranges for urine sodium, however, a more in-depth analysis linking disease risk in outlying individuals based on current or proposed thresholds would first be required.

Most individuals in UKB are outliers for at least one biomarker reference range

Nearly all (96.4%, 178,691/185,405) individuals were outliers for at least one of the 32 biomarkers included in this analysis (Figure 2.4c). The mean number of

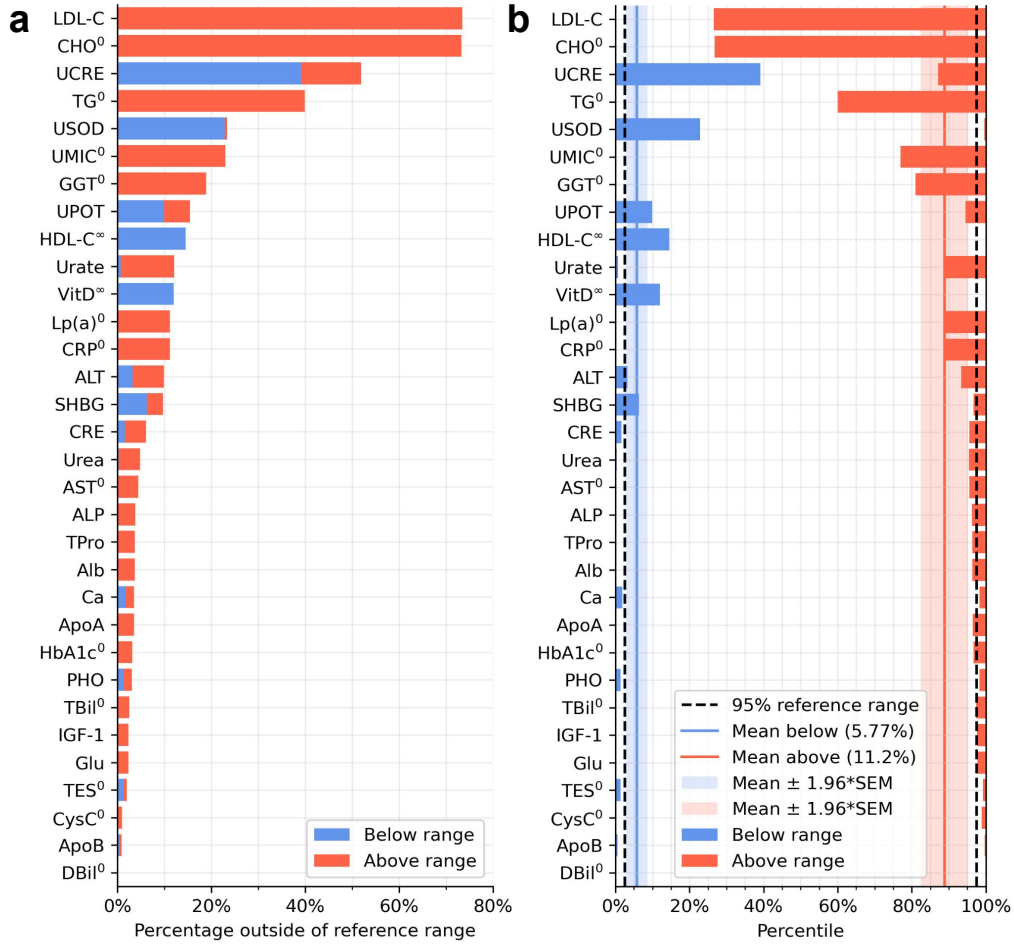

Figure 2.2: High variability in the number of individuals outside of clinical reference ranges for 32 UKB blood and urine biomarkers. **a**, Stacked percentages of individuals below and above reference ranges. **b**, Below and above outliers shown across the percentile range. Phenotypes are ordered by the total proportion of individuals above or below the reference range. Only individuals who were in the 200k ES tranche, passed individual-level QC, and were classified as genetically European were included ($n = 185,405$, Methods). Biomarker labels on the y-axis are annotated to indicate whether clinical reference ranges lack a lower bound (indicated with "^{0}"; i.e. only bounded by natural limit at zero) or lack an upper bound (indicated with "$^\infty$"). Testosterone (TES) has no lower bound in the female-specific reference range but does in male-specific reference range. Biomarkers without lower or upper thresholds were excluded from the calculation of mean percentage of lower and upper outliers, respectively. Sex-specific thresholds are used where possible, otherwise sex-agnostic thresholds are used. The mean percentage of below- and above-range outliers was calculated, with 95% CI calculated using $\pm 1.96 \times$ SEM. When calculating the mean percentages of below- and above-range outliers, biomarkers were excluded from the calculation if their reference range was unbounded for the left or right tail, respectively, of the distribution. SEM, standard error of mean

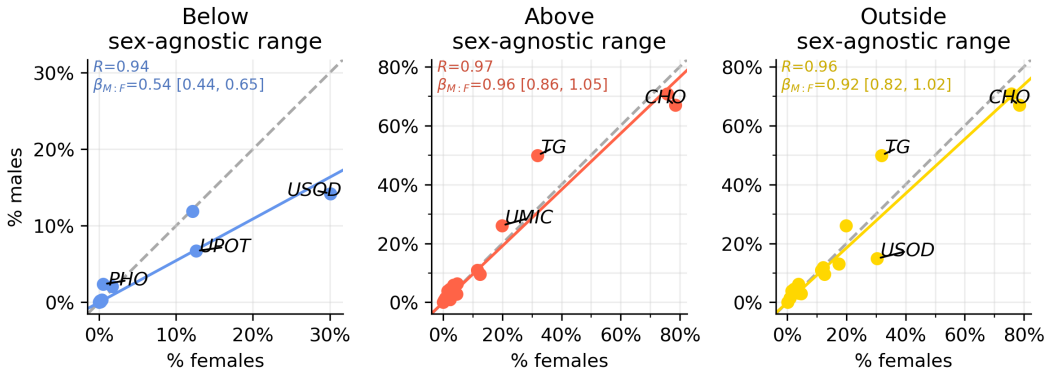

Figure 2.3: Percentage of individuals below, above and outside sex-specific and sex-agnostic reference ranges. **a**, Percentages of individuals below, above and outside sex-specific clinical reference ranges. **b**, Percentages of individuals below, above and outside sex-agnostic clinical reference ranges. Biomarkers with the largest difference in outlying percentages between sexes in each subplot are labelled. In the top left is reported the Pearson correlation coefficient between the male and female outlying percentages, as well as the regression coefficient, from the regression (without fitting an intercept) of male percentages against female percentages, with 95% CI calculated using 1.96 × SE.

biomarkers for which an individual was outside the reference range was 4.10 (95% CI = [4.09, 4.11]). On average, individuals tended to be above-range outliers for more biomarkers (mean [95% CI] = 3.00 [2.99, 3.01]) than below-range outliers (mean [95% CI] = 1.1 [1.09, 1.10]). This imbalance can be explained partly by the fact that more biomarkers had reference ranges with unbounded lower limits (12 biomarkers) than unbounded upper limits (2 biomarkers: high-density lipoprotein cholesterol (HDL-C) and Vitamin D).

Individuals were below-range outliers for up to nine biomarkers, and above-range outliers for as many as 17 (Figure 2.4). However, it was relatively uncommon to be both below-range or above-range for multiple biomarkers: individuals above the median for counts of below-range biomarkers (> 1 biomarkers) were significantly less likely to be above the median (> 3) for counts of above-range biomarkers (odds ratio (OR) [95% CI] = 0.888 [0.871, 0.905], two-sided Fisher's exact test $P = 8.45 \times 10^{-34}$) (Figure 2.4d).

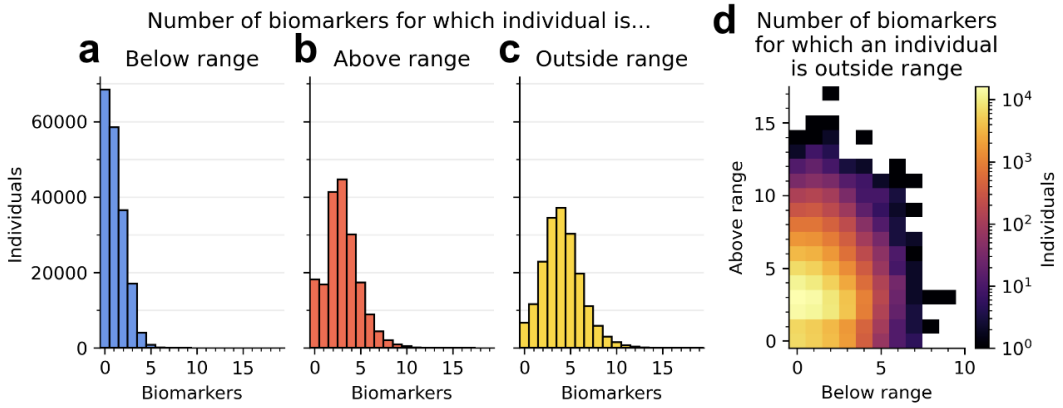

Figure 2.4: Many individuals are outliers for multiple biomarkers. Histogram of the distribution in number of biomarkers for which an individual is **a**, above, **b**, below or **c**, outside clinical reference ranges. **d**, Heatmap of the number of biomarkers for which an individual is above or below range. Heatmap cells which are not coloured indicate zero counts.

Significant enrichment of outlier status in pairs of correlated biomarkers

The high percentage of individuals with at least one biomarker outside of the reference range could be due to correlation present between biomarkers. For example, if an individual is an outlier for one biomarker, they may be more likely to be an outlier in a correlated biomarker. There is strong correlation in some biomarker pairs (e.g. R(ApoB, LDL-C) = 0.958, Pearson $P < 1 \times 10^{-100}$; R(Direct bilirubin, Total bilirubin) = 0.905, $P < 1 \times 10^{-100}$; Figure S2.1; Table S2.2), but it is not obvious to what degree this translates to overlap in outlying individuals between pairs of correlated biomarkers. To investigate this across all pairs of biomarkers, I tested whether outlying individuals in one biomarker were more likely to be outliers in another biomarker, using a series of Fisher's exact tests. For biomarkers with defined reference ranges but zeros in the contingency tables, I filled zeros with 0.5 to enable calculation of ORs. I tested enrichment for below- and above-range outlying individuals separately, as well as the combined outside-range enrichment (Figure 2.5; Table S2.2). As a point of comparison, I simulated pairs of normally distributed biomarkers across a range of correlation coefficients and defined outliers using three choices of reference range intervals: 99%, 95% and 90% (Methods). This represents the null, under the assumption that biomarkers are

normally distributed and reference ranges are defined using the same percentile thresholds across biomarkers. Biomarker pairs with outlier enrichment ORs above this null have higher overlap in outlying individuals than expected.

In the below-range enrichment results, the biomarker pairs with highest odds of sharing below-range outliers were nested pairs: apolipoprotein A (ApoA) is a component of HDL-C[172], serum albumin is a component of total protein[173], and ApoB is a component of LDL-C[174] (Figure 2.5a; Table S2.2).

Similarly, the highest ORs in the above-range enrichment results were for pairs of nested biomarkers. In addition to pairs of nested cholesterol-related serum biomarkers (ApoB, LDL-C, total cholesterol), there was strong overlap in above-range outliers for total bilirubin and direct bilirubin. Total bilirubin is the sum of direct and indirect bilirubin[175]), so this overlap is expected.

Three pairs of cholesterol-related biomarkers appeared in the outside-range results as being depleted for overlap (ApoA with HDL-C OR [95% CI] = 0.0699 [0.0552, 0.0886], two-sided Fisher's exact test $P = 1.17 \times 10^{-288}$; ApoB with total cholesterol OR [95% CI] = 0.224 [0.203, 0.248], $P = 1.36 \times 10^{-191}$; ApoB with LDL-C (OR [95% CI] = 0.204 [0.184, 0.226], $P = 1.53 \times 10^{-215}$) (Figure 2.5c; Table S2.2). On first appraisal this seems to contradict the biological link between these pairs, which manifests as a strong positive correlation ($R > 0.88$; Table S2.2). However, the depletion (OR < 1) is because the reference ranges for HDL-C and total cholesterol are unbounded for the upper and lower limits, respectively, whereas ApoA and ApoB are bounded for both upper and lower limits. This meant that in the outside-range contingency tables (Table 2.2) individuals with high HDL-C were definitionally unable to segregate with high-ApoA outliers, and likewise individuals with low total cholesterol were unable to segregate with low-ApoB outliers.

To explain the contradiction with LDL-C and ApoB, although there exist upper and lower limits for both biomarkers, LDL-C has so few below-range outliers compared to above-range outliers that the above-range outliers dominate the contingency table (Table 2.3). This results in a biologically contradictory

a.

	HDL-C outside limits	HDL-C inside limits
ApoA outside limits	70	5,618
ApoA inside limits	23,542	132,122

b.

	Total cholesterol outside limits	Total cholesterol outside limits
ApoA outside limits	627	1,000
ApoA inside limits	128,206	45,886

Table 2.2: Outside-range contingency tables for blood lipid biomarkers. **a**, Outside-range contingency table for HDL-C and ApoA. **b**, Outside-range contingency table for total cholesterol and ApoB.

a.

	LDL-C < lower limit	LDL-C ≥ lower limit
ApoB < lower limit	2	1,029
ApoB ≥ lower limit	1	174,424

b.

	LDL-C outside limits	LDL-C inside limits
ApoB outside limits	595	1,029
ApoB inside limits	128,529	45,303

Table 2.3: Contingency tables for LDL-C and ApoB. **a**, Below-range contingency table for LDL-C and ApoB. **b**, Outside-range contingency table for LDL-C and ApoB.

outside-range depleted OR simply because the LDL-C lower limit results in so few outlying individuals.

Apart from these three cholesterol-related aberrant pairs, the other 14 biomarker pairs with moderate correlation ($R^2 > 0.2$) were significantly enriched for overlapping outside-range outliers (OR > 1, two-sided Fisher's exact test $P < 1.24 \times 10^{-138}$) (Figure 2.5; Table S2.2).

Odds of disease increases with increasing number of biomarkers outside of reference ranges

Individuals who are outliers for multiple biomarkers may be more likely to develop disease, or conversely, individuals with disease may be more likely to be outliers for multiple biomarkers. To investigate this, I tested whether individuals with biomarkers outside of the reference range were more likely to have a disease within a broad category, using ICD-10 chapter codes.

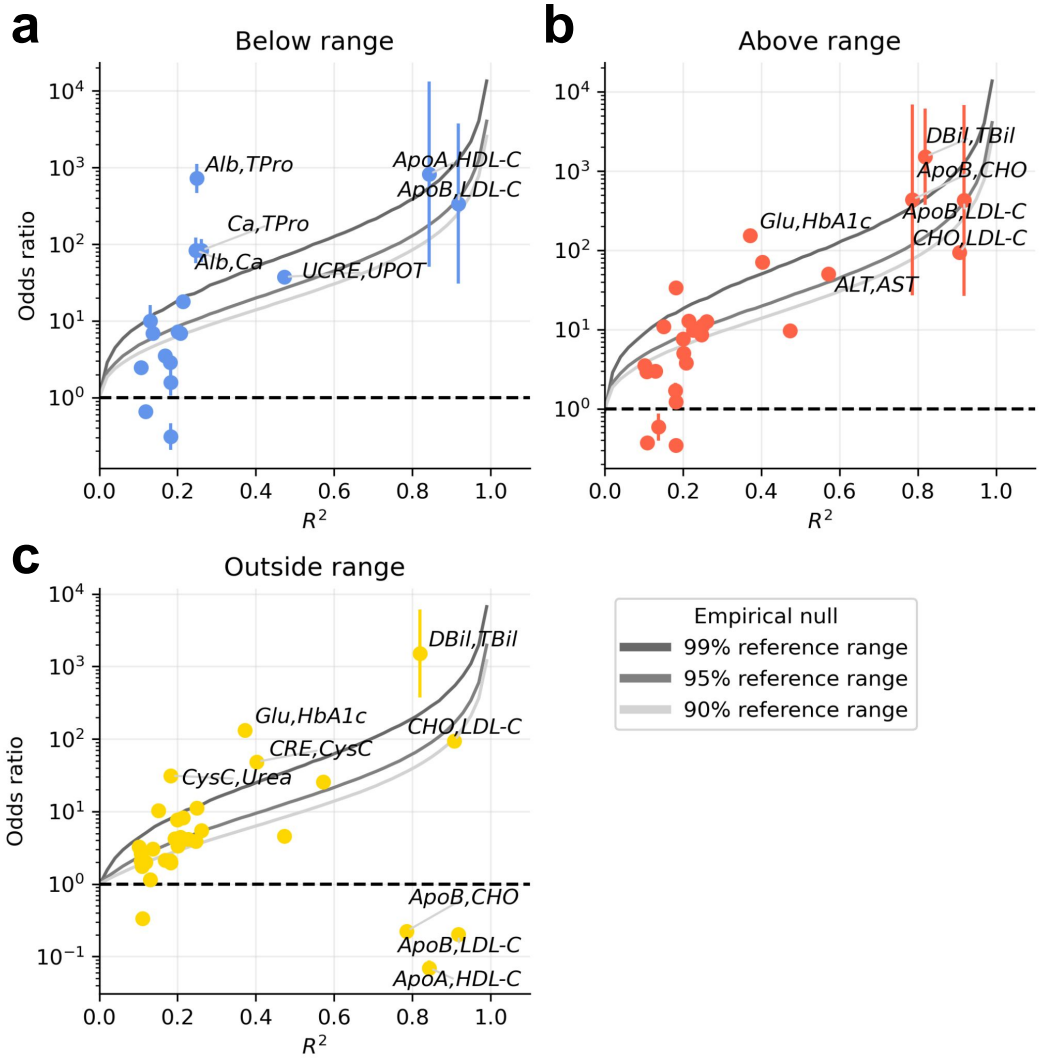

Figure 2.5: Comparing biomarker correlation with outlier overlap enrichment. Only biomarker pairs with $R^2 > 0.1$ are plotted. The five biomarker pairs with highest R^2 and the five pairs with highest OR are labelled. Error bars indicate 95% CI for OR. The grey bands for 99%, 95% and 90% reference ranges correspond to simulated empirical null ORs, assuming the two biomarkers are drawn from a bivariate normal distribution (Methods). The black dashed line indicates an OR of 1.

I selected the five disease-related ICD-10 chapters (chapters with "disease" in the name) with the highest counts, using the "Diagnoses - ICD10" entry (data field: 41270) in the UKB Data Showcase. As a negative control, I included ICD-10 codes related to transport accidents (ICD-10 codes V01-V99).

For all five disease-related ICD-10 chapter codes, odds of disease increased monotonically with the number of outlying biomarkers (Figure 2.6a; Table S2.3). All five were significantly enriched for all but the lowest bin of outlying biomarker counts (OR > 1, maximum one-sided Fisher's exact test $P = 1.09 \times 10^{-3}$). Of the selected chapters, ICD-10 codes from "Chapter IV Endocrine, nutritional and metabolic disease" showed the highest enrichment as the number of outlying biomarkers increased, with individuals who had 15 or more outlying biomarkers approximately ten times as likely to have at least one ICD-10 code from the chapter (OR [95% CI] = 10.6 [4.74, 23.8], one-sided Fisher's exact test $P = 4.86 \times 10^{-12}$). The negative control, ICD-10 codes related to transport accidents, was not significantly enriched for any of the outlying biomarker count bins, except nominally significant for individuals with four or fewer outlying biomarkers (OR [95% CI] = 1.08 [1.01, 1.15], $P = 0.0168$). A potential explanation could be that individuals with healthy biomarker levels are more likely to be regularly taking transport, and are therefore at elevated risk for accidents compared to individuals with more outlying biomarkers, who may be diseased and less active. However, this enrichment does not pass Bonferroni correction for 24 tests across 6 ICD-10 chapters and 4 bins ($P = 0.05/24 = 2.08 \times 10^{-3}$).

Specific ICD-10 codes within chapters may display varying degrees of association with count of outlying biomarkers. I hypothesised that by stratifying each chapter into its subchapters (called "blocks") I could increase the signal-to-noise ratio and find higher enrichment ORs. To limit multiple testing burden, I restricted analysis in each disease-related chapter to the three blocks with the most code counts in the UKB Data Showcase, a proxy for prevalence. The majority (9/15) of the ICD-10 blocks I tested showed significant disease enrichment (one-sided Fisher's exact test $P < 0.05/15$, Bonferroni correction) in bins of individuals with five or more outlying

biomarkers (Figure 2.6b; Table S2.3). The three blocks with highest ORs in the group of individuals with 15 or more outlying biomarkers were diabetes mellitus (OR [95% CI] = 23.8 [12.6, 44.8], $P = 1.23 \times 10^{-22}$), renal failure (OR [95% CI] = 16.5 [9.05, 30.1], $P = 8.63 \times 10^{-18}$), and hypertensive diseases (OR [95% CI] = 7.41 [3.66, 15.0], $P = 5.48 \times 10^{-10}$). Diabetes is known to be linked with many changes to an individual's metabolomic profile, which could result in diabetes disease cases having multiple outlying biomarkers[176]. Chronic kidney disease, a major disease classified under renal failure, is linked with many of the biomarkers selected for this analysis, including serum creatinine, cystatin C, albumin, microalbumin, ApoA, ApoB and HDL-C[177,178]. Essential hypertension, the primary form of hypertensive disease, is known to be linked with broad changes in lipid profile[179–181].

2.3.2 Using extreme biomarker levels to prioritise individuals with rare genetic disorders for genetic screening

Metabolomic profiling has been shown to aid diagnosis of multiple rare genetic disorders[149]. I hypothesised that even in UKB, a population cohort not ascertained for rare disease, individuals with extreme biomarkers would be enriched for pLoF variants in genes which cause rare disorders. To test this hypothesis, I estimated the enrichment of pLoF variants in genes known to cause rare genetic disorders among individuals with extreme biomarker levels, relative to individuals with biomarker levels within reference ranges.

Individuals with extreme biomarker levels are enriched for pLoF variants in genes which cause rare disorders

I tested enrichment of pLoFs in outlying individuals for 12 unique biomarkers and 10 unique GEL PanelApp[182,183] rare disorders, curated from evidence in literature (Methods; Table S2.4). Individuals with biomarker levels outside of clinical reference ranges were enriched (one-sided Fisher's exact test $P < 0.05$) for rare (minor allele frequency (MAF)$< 0.1\%$) pLoF variants in genes linked with familial hypercholesterolaemia (FH) for individuals with outlying LDL-C, ApoB, total cholesterol, severe hypertriglyceridaemia for individuals with outlying

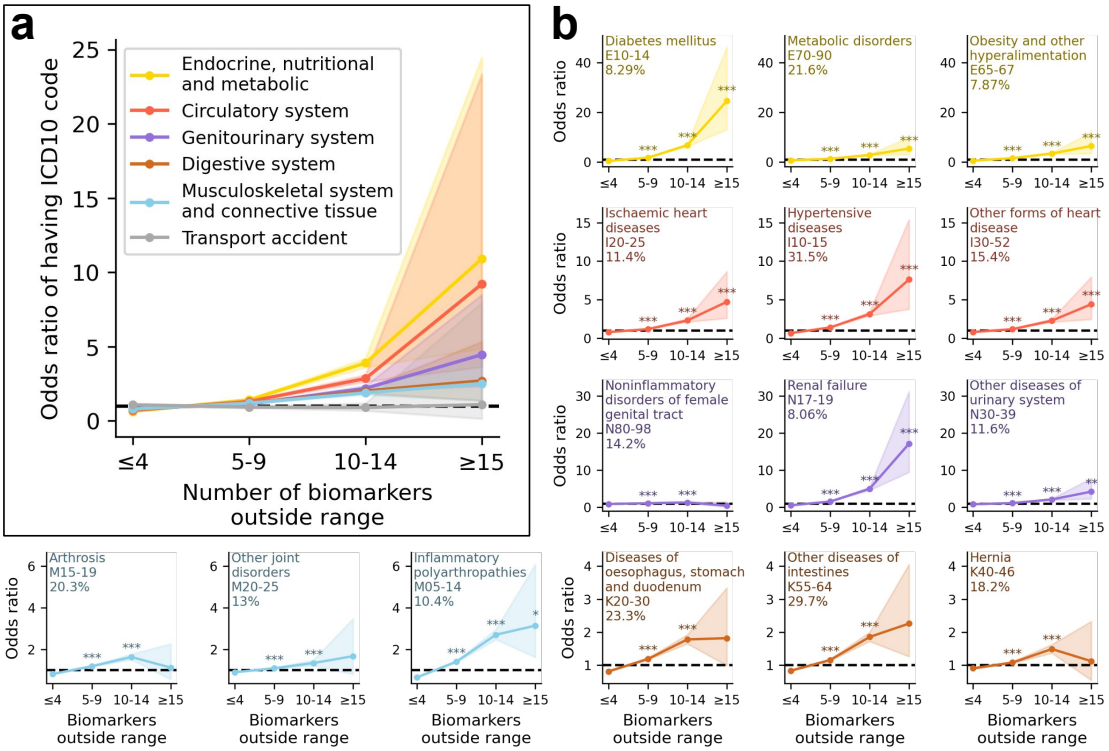

Figure 2.6: Disease prevalence increases with count of outlying biomarker measurements. **a**, Odds of having an ICD-10 code belonging to one of the top five most prevalent disease-related ICD-10 chapters (Methods), or an ICD-10 code relating to transport-related accidents, for four bins of biomarker outlier counts: ≤ 4, 5-9, 10-14, and ≥ 15. Each bin is compared to all other bins. **b**, Odds of having an ICD-10 code belonging to one of the top three most prevalent blocks under each of the most prevalent disease-related ICD-10 chapters, for each bin of number of outlying biomarkers. The title of each plot includes the name of the subcategory, the codes associated with that range, and the prevalence of the subcategory. $^{***}P < 1 \times 10^{-5}$, $^{**}P < 1 \times 10^{-4}$, $^{*}P < 0.05/15$ (Bonferroni correction for 15 blocks) for one-sided Fisher's exact test with an alternative hypothesis that the OR is greater than one. Shading around points indicates the 95% CI for each OR.

triglycerides, congenital hypothyroidism for individuals with outlying sex hormone-binding globulin (SHBG), and monogenic diabetes for individuals with outlying HbA1c (Figure 2.7a; Table S2.5). Using 95% reference ranges defined on percentiles of UKB individuals instead of clinical reference ranges, many of the same biomarker-disorder pairs were still significantly enriched, with urine microalbumin paired with proteinuric renal disease a new addition ($P = 0.0307$) (Figure 2.7b; Table S2.5). When using clinical or 95% reference ranges, enrichment was highest for blood

lipid biomarkers paired with FH, especially for individuals with ApoB outside the reference range (OR [90% CI] = 28.9 [14.4, 57.8]).

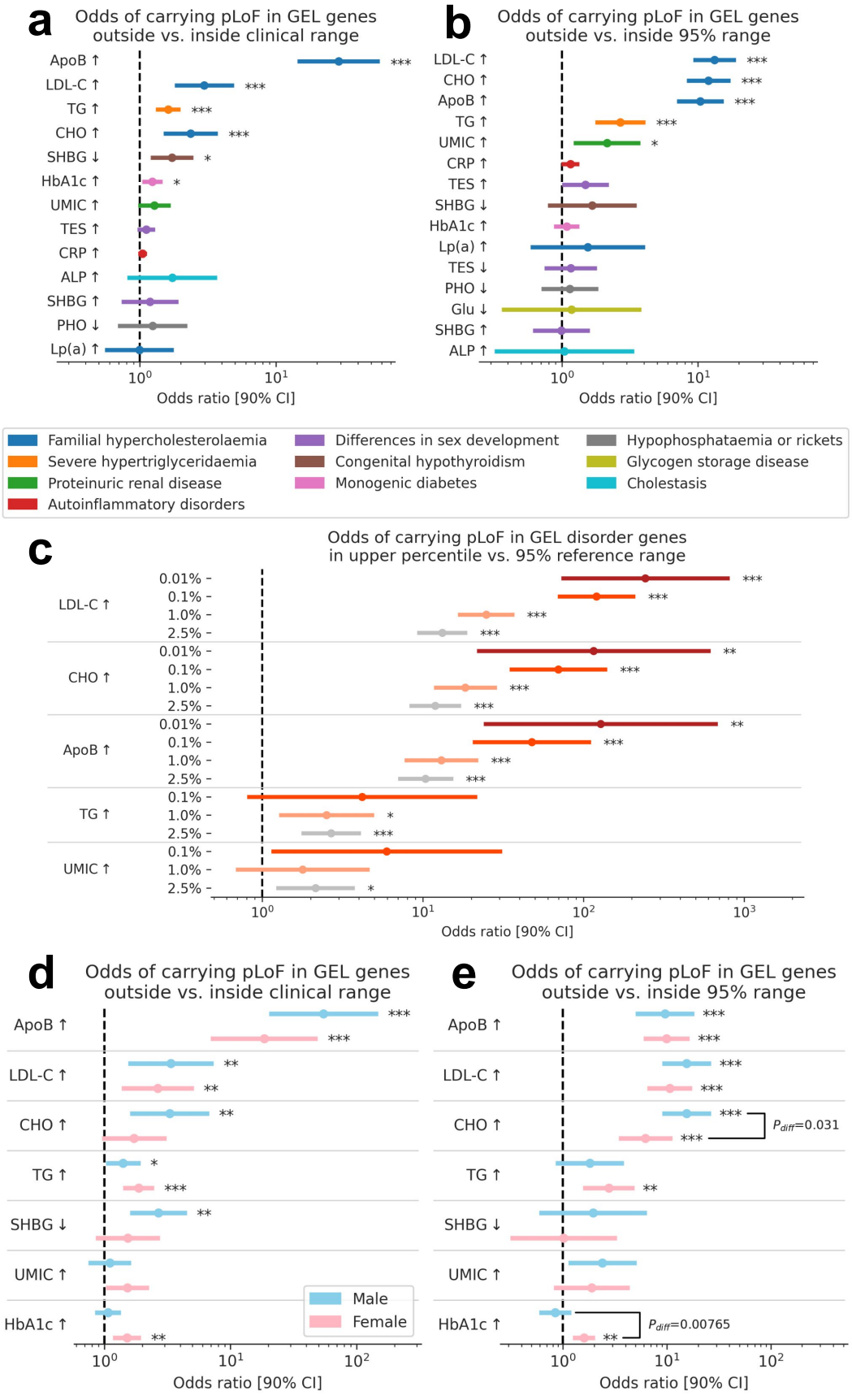
a
Odds of carrying pLoF in GEL genes outside vs. inside clinical range
ApoB ↑
LDL-C ↑
TG ↑
CHO ↑
SHBG ↓
HbA1c ↑
UMIC ↑
TES ↑
CRP ↑
ALP ↑
SHBG ↑
PHO ↓
Lp(a) ↑
Odds ratio [90% CI]
b
Odds of carrying pLoF in GEL genes outside vs. inside 95% range
LDL-C ↑
CHO ↑
ApoB ↑
TG ↑
UMIC ↑
CRP ↑
TES ↑
SHBG ↓
HbA1c ↑
Lp(a) ↑
TES ↓
PHO ↓
Glu ↓
SHBG ↑
ALP ↑
Odds ratio [90% CI]
Familial hypercholesterolaemia
Severe hypertriglyceridaemia
Proteinuric renal disease
Autoinflammatory disorders
Differences in sex development
Congenital hypothyroidism
Monogenic diabetes
Hypophosphataemia or rickets
Glycogen storage disease
Cholestasis
c
Odds of carrying pLoF in GEL disorder genes in upper percentile vs. 95% reference range
LDL-C ↑
CHO ↑
ApoB ↑
TG ↑
UMIC ↑
0.01%
0.1%
1.0%
2.5%
Odds ratio [90% CI]
d
Odds of carrying pLoF in GEL genes outside vs. inside clinical range
ApoB ↑
LDL-C ↑
CHO ↑
TG ↑
SHBG ↓
UMIC ↑
HbA1c ↑
Male
Female
Odds ratio [90% CI]
e
Odds of carrying pLoF in GEL genes outside vs. inside 95% range
ApoB ↑
LDL-C ↑
CHO ↑
TG ↑
SHBG ↓
UMIC ↑
HbA1c ↑
P_{diff}=0.031
P_{diff}=0.00765
Odds ratio [90% CI]

Figure 2.7 ***(previous page)*****: Enrichment of GEL PanelApp rare disorder pLoF variants among individuals with outlying biomarker levels**. **a**, Enrichment among individuals outside reference ranges. **b**, Enrichment among individuals outside a middle 95% percentile range. **c**, Enrichment among individuals in the extreme percentiles (2.5%, 1.0%, 0.1%, 0.01%) of the biomarker distribution compared to individuals in the middle 95% percentile range. **d**, Enrichment stratified by sex, using reference ranges (sex-specific ranges, when available) to define individuals with outlying biomarker levels. **e**, Enrichment stratified by sex (blue: male; pink: female), using sex-specific 95% percentile ranges to define inlying and outlying biomarker levels. P_{diff} is the P-value from a two-sided z-test testing for difference between the sex-specific enrichment results. Biomarkers are ordered from most to least significant. Error bars indicate the 90% CI, such that the lower limit corresponds to the lower limit of the one-sided 95% CI. Arrows next to the biomarker labels indicate whether the upper (↑) or lower tail (↓) of the biomarker distribution was tested for pLoF enrichment. One-sided Fisher's exact test with the alternate hypothesis that OR is greater than one: $^{***}P < 0.001$, $^{**}P < 0.01$, $^{*}P < 0.05$. pLoF, predicted loss-of-function; GEL, Genomics England. See Table 2.1 for biomarker abbreviations.

For biomarker-disorder pairs which had shown significant enrichment of pLoFs when using 95% reference ranges, I tested whether the enrichment of rare disorder pLoFs increased further into the biomarker extremes. I applied increasingly extreme thresholds, first classifying outliers as the top 2.5% in the biomarker distribution, then 1%, 0.1%, and finally 0.01%. As more extreme percentile thresholds were applied, enrichment of pLoFs increased for three of the blood lipid biomarkers linked with FH (LDL-C, ApoB, and total cholesterol; Figure 2.7c; Table S2.5). In the most extreme category, comparing individuals in the top 0.01% percentile of the biomarker distribution to individuals in the middle 95%, individuals with outlying LDL-C were estimated to be around 240-fold enriched for FH pLoF variants (Figure 2.7c; Table S2.5). Enrichment of pLoFs in the 0.01% extremes of the two other FH-paired biomarkers was also high, although about half the magnitude (ApoB OR [90% CI] = 128 [23.9, 686]; Total cholesterol OR [90% CI] = 116 [21.6, 618]). No other biomarkers demonstrated a monotonic trend in increasing OR point estimates with increasingly extreme percentile thresholds (Figure 2.7c; Table S2.5).

Sex differences in enrichment of pLoF variants in genes which cause rare disorders among individuals with extreme biomarkers

To assess the possibility of sex-specific effects in enrichment of pLoFs among individuals with extreme biomarkers, I stratified analysis by sex, testing for enrichment in extremes defined by clinical or 95% reference ranges (Figure 2.7d,e; Table S2.5). To control for sex differences in biomarker distribution, I calculated sex-stratified 95% reference ranges for each biomarker. To reduce multiple testing burden, I restricted my analysis to biomarker-disorder pairs which displayed significant enrichment in my sex-combined analysis using clinical or 95% reference ranges.

When biomarker extremes are defined using clinical reference ranges, three biomarker-disorder pairs demonstrated sex-specific enrichment: significant enrichment (one-sided Fisher's exact test $P < 0.05$) in one sex but not the other (Figure 2.7d; Table S2.5). I found male-specific enrichment for total cholesterol paired with FH ($P_{\text{male}} = 1.88 \times 10^{-3}$, $P_{female} = 0.0787$) and SHBG paired with congenital hypothyroidism ($P_{\text{male}} = 1.98 \times 10^{-3}$, $P_{\text{female}} = 0.156$), and female-specific enrichment for HbA1c paired with monogenic diabetes ($P_{\text{female}} = 7.26 \times 10^{-3}$, $P_{\text{male}} = 0.345$). Testing for significant sex-differential effects (two-sided z-test[184] $P<0.05$), none of the clinical reference range enrichment results were significantly different.

Using standardised 95% reference ranges, two biomarker-disorder pairs demonstrated sex-specific enrichment: female-specific enrichment for triglyceride paired with extreme hypertriglyceridaemia ($P_{\text{female}} = 7.56 \times 10^{-3}$, $P_{\text{male}} = 0.153$) and, once again, female-specific enrichment for HbA1c paired with monogenic diabetes ($P_{\text{female}} = 2.02 \times 10^{-3}$, $P_{\text{male}} = 0.813$) (Figure 2.7e; Table S2.5). Testing sex-differential effects, I observed significantly different sex-specific enrichment effects for HbA1c paired with monogenic diabetes ($P_{\text{diff}} = 7.65 \times 10^{-3}$) and total cholesterol paired with FH ($P_{\text{diff}} = 0.0310$) (Figure 2.7e; Table S2.5). Focusing on the more significant of the two sex-differential effects, the stronger enrichment of monogenic diabetes pLoF variants in females versus males has not been reported before, to the best of my knowledge. If true, this would suggest that high HbA1c (i.e. above a sex-specific 95% reference range) is more informative for pre-screening

for monogenic diabetes in females than males. Although the pLoF enrichment in females is moderate (OR [90% CI] = 1.59 [1.25, 2.03]), and therefore might not implicate high HbA1c as a crucial indicator for monogenic diabetes in females, it may still provide clues to the sex-specific biology of monogenic diabetes.

2.3.3 Discovering variants associated with extreme biomarker levels

Dichotomising continuous biomarkers into tail status

Having explored the purely phenotypic aspects of biomarker extremes, as well as validating genes known to be associated with rare disorders, I shifted focus to the discovery of novel genetic associations with extreme biomarker levels. I sought to interrogate whether there was genetic signal in the tails of single biomarker distributions. To standardise the analysis across phenotypes, I used percentile thresholds, similar to how standard reference ranges are designed[164]. Instead of using a centred 95% confidence interval defined by healthy individuals, I used an 80% confidence interval across everyone. This corresponds to 10% tails of upper and lower outliers, which approximately matches the observed 8.98% of above-range outliers when using clinical reference ranges. By including slightly more individuals in the outlying tails I also expected to attenuate issues of data sparsity, especially when performing a rare variant analysis.

Quality control was performed by Samvida Venkatesh on blood and urine biomarkers, following methods described by Sinnott-Armstrong *et al.*[85], with LDL-C, ApoB and total cholesterol excluded due to strong confounding effects from statins. I dichotomised 29 blood and urine biomarkers from the UKB into low-, high- and either-tail status, using the 80% confidence interval. Individuals with low-tail status were those with biomarker levels below the 10th percentile, high-tail were those with biomarker levels above the 90th percentile, and either-tail were those with biomarker levels either below or above the confidence interval. I included the either-tail status because I hypothesised that it could be powered to detect variants which increase trait variance – termed variance quantitative trait loci (vQTL).

Tail-status associations

Using UKB exome sequencing data across 160,151 genetically European individuals and 144,483 variants with MAF > 0.1%, I performed a GWAS on continuous and tail-status phenotypes for each blood and urine biomarker. These analyses identified 4,357 variant-biomarker pairs associated with tail status at exome-wide significance[185] ($P<5\times10^{-7}$; Table S2.6) across 27 unique biomarkers of 29 biomarkers tested, with urine potassium and urine sodium being the two biomarkers without significant tail status associations. This included 2,781 unique variants, 28.1% (781/2781) of which were missense variants (Table S2.6). These variants were not pruned for LD and therefore the number of independent loci associated with tail status may be fewer than the number of significantly associated variants.

There was significant positive correlation between the number of significant variants in the continuous trait and the number of significant low-tail ($R = 0.87$, $P = 1.34 \times 10^{-9}$) and high-tail ($R = 0.89$, $P = 1.79 \times 10^{-10}$) variants (Figure 2.8a), demonstrating that traits with more continuous phenotype associations tended to also have more tail-status associations. There was no significant correlation ($R = 0.12$, $P = 0.524$) between the number of significant variants for the continuous and either-tail phenotypes (Figure 2.8a).

The majority of variant-biomarker tail-status associations were with low- (71.6%, 3118/4357) or high-tail (61.8%, 2694/4357) phenotypes, with only 10.9% (474/4357) from the either-tail phenotype. This is likely because the either-tail model does not align with the additive genetic model which accounts for most statistical signal in complex trait GWAS. For example, none of the 24 variant-biomarker pair tail-status associations driven solely by either-tail status were associated with the continuous phenotype. This contrasts with tail-status associations driven solely by low- and high-tail status, of which the vast majority (low-tail: 94.5%, 1466/1551; high-tail: 93.5%, 1058/1132) were also associated with the continuous phenotype.

I hypothesised that the low-tail phenotype would have more power to detect variants which had a decreasing effect on the continuous version of the biomarker, while the high-tail phenotype would have more power to detect variants with an

increasing effect on the continuous trait. Pooling associations across all biomarkers, I found variants significantly associated with low-tail status ($P_{\text{low}} < 5 \times 10^{-7}$) were significantly (OR [90% CI] = 1,040 [974, 1,110], one-sided Fisher's exact test $P < 1 \times 10^{-100}$) more likely to have a negative effect ($\beta_{\text{cts}} < 0$, $P_{\text{cts}} < 5 \times 10^{-7}$) on the continuous trait (Methods). Similarly, significant high-tail variants ($P_{\text{high}} < 5 \times 10^{-7}$) were significantly more likely (OR [90% CI] = 928 [867, 993], one-sided Fisher's exact test $P < 1 \times 10^{-100}$) to have a positive effect ($\beta_{\text{cts}} > 0$, $P_{\text{cts}} < 5 \times 10^{-7}$) on the continuous trait (Methods).

204 variants associated with tail-status but not continuous biomarker measurement

A total of 206 variant-biomarker pairs (204 unique variants, 22 unique biomarkers) were significantly associated ($P < 5 \times 10^{-7}$) with a tail-status phenotype but not the continuous phenotype for the same biomarker (Table S2.6). I defined these markers as variants of interest (VOI). If I were to use a more stringent threshold, using Bonferroni correction to account for testing 29 biomarkers there would be a total of 100 significant ($P < 5 \times 10^{-5}/29 = 1.72 \times 10^{-8}$) variant-biomarker pairs (97 unique variants, 16 unique biomarkers).

There were two variants which were VOI in more than one phenotype. Neither were VOI under the more stringent significance threshold, adjusting for 29 traits tested ($P = 5 \times 10^{-7}/29$). The first variant (chr3:171014511:G:A, alternate allele frequency (AAF)=0.122) was a missense variant in *SLC2A2* (Solute Carrier Family 2 Member 2; encodes the protein GLUT2, glucose transporter protein type 2) that was not predicted to be damaging (Methods). This variant was significantly associated with high-tail urate (β_{high} [95% CI] = 0.094 [0.0590, 0.129], P_{high}=1.78×10^{-7}) and low-tail CRP (β_{low} [95% CI] = -0.101 [-0.138, -0.0636], $P_{\text{low}} = 8.78 \times 10^{-8}$) (Figure 2.9, Figure 2.10). In the continuous versions of both of these biomarkers, the variant had a positive effect, with an association signal bordering on exome-wide significance for CRP ($P_{\text{cts}} = 7.04 \times 10^{-7}$) but not urate ($P_{\text{cts}} = 2.59 \times 10^{-4}$). The variant was significantly associated with four continuous biomarker phenotypes: gamma glutamyltransferase (GGT), glucose, HbA1c and total bilirubin (Figure 2.9).

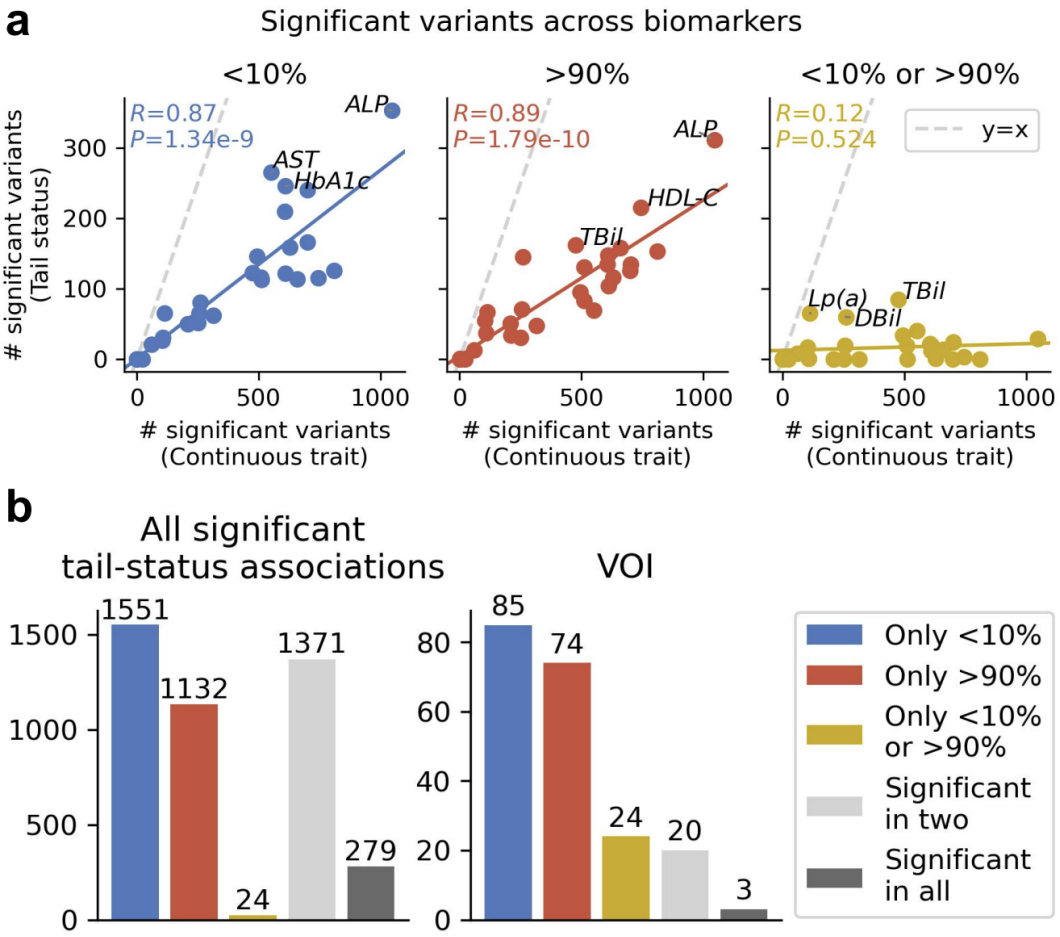

Figure 2.8: Variants associated with tail-status phenotypes. **a**, Number of significant variants per trait for each tail-status, compared to the number of significant variants in the continuous version of the biomarker. Pearson correlation coefficient and correlation *P*-value are in the top left. For each tail status, the top three biomarkers with the most tail-status associated variants are labelled. For single-variant comparison of tail-status versus continuous trait, see Figure S2.2. **b**, All significant tail-status variants and VOI stratified by the tail-status phenotype association.

Missense variants in *SLC2A2* are known to cause a rare autosomal recessive form of metabolic disease called Fanconi-Bickel syndrome, characterised by accumulation of glycogen in the liver, glucose intolerance, postprandial hyperglycaemia, fasting hypoglycaemia, and stunted growth[186,187]. This variant has previously been shown to be associated with triglycerides[188], higher risk of type 2 diabetes (T2D) with ketoacidosis[189], and lower fasting glucose and HbA1c[190]. An association with increased urate or CRP has not been reported. However, it is biologically plausible given that individuals with higher urate are at higher risk of T2D[191–193] and the

demonstrated link between T2D and higher CRP[194–196].

The second variant which was a VOI for more than one phenotype (assuming a significance threshold of $P = 5 \times 10^{-}7$) was a non-coding variant (chr8:10726702:C:T, AAF=0.607) mapped to *SOX7* (SRY-Box Transcription Factor 7). The two tail-status phenotypes which the VOI was associated with were low-tail aspartate aminotransferase (AST) (β_{low} [95% CI] = 0.0739 [0.0496, 0.0983], $P_{\text{low}} = 2.50 \times 10^{-9}$) and low-tail GGT ($\beta_{\text{low}}$ [95% CI] = 0.0661 [0.0414, 0.0908], $P_{\text{low}} = 1.51 \times 10^{-7}$) (Figure 2.9, Figure 2.10). In the continuous versions of these biomarkers the variant was close to exome-wide significance (AST $P_{\text{cts}} = 2.05 \times 10^{-6}$; GGT $P_{\text{cts}} = 7.60 \times 10^{-7}$). *SOX7* is closely related to *SOX18*. Both genes have been implicated in the vascular development of mice[197] and missense variants in *SOX18* are known to cause hypotrichosis-lymphedema-telangiectasia in humans[198], which does not have an obvious link with AST or GGT. Further work may be required to demonstrate whether this non-coding tail-status association is plausibly causal for AST and GGT levels.

Among the 64 coding non-synonymous variants involved in VOI associations, two were classified as damaging missense. None were classified as protein-truncating variants. Of the two predicted damaging missense VOI, one was a variant in *PLG* (Plasminogen) associated with low- and either-tail status for lipoprotein A (chr6:160714844:A:G, AAF=0.00180, P_{low}=3.90×10^{-10}, $\beta_{\text{low}} = -1.45$, P_{either}=5.73×10^{-11}, β_{either}=-0.95, P_{cts}=0.0189) (Figure 2.10). This was also a VOI under the more stringent threshold adjusting for 29 traits tested ($P = 5 \times 10^{-7}/29$). Nearby significant variants in *PLG* are also associated with lipoprotein A (Table S2.6). However, this VOI was among the rarer variants associated with tail status or continuous lipoprotein A in *PLG*. Therefore, one explanation for this VOI could be that the lower allele frequency may reduce statistical power to detect the continuous phenotype association, although this explanation assumes that effect sizes for nearby variants are similar. Rare variant (MAF$<$ 1%) burden in *PLG* is associated with lipoprotein A in Genebass[199] and in a recessive burden test (MAF$<$ 5%) by AstraZeneca[200], further evidence that this gene may be causally linked with

lipoprotein A levels in UKB. It could be that this VOI spuriously segregated with individuals in the lower tail of the phenotype distribution, making the low- and either-tail-status associations significant and the continuous trait association non-significant. Of the significant common variants in this gene (AAF > 0.27), all 11 are associated with continuous and all three tail-status phenotypes for lipoprotein A.

The other damaging missense variant was in *CAPN9* (Calpain 9) and was associated with either-tail status of glucose (chr1:230772053:C:T, AAF=0.00315, P_{either}=4.27×10^{-7}, β_{either}=0.398, $P_{\text{cts}} = 0.195$, $\beta_{\text{cts}} = 7.00 \times 10^{-3}$) (Figure 2.10). This was not a VOI under the more stringent threshold adjusting for 29 traits tested ($P = 5 \times 10^{-7}/29$). No other biomarkers associated with this variant or gene. *CAPN9* is known to be associated with gastric cancer and is highly expressed in the stomach[201]. Therefore there could be a plausible connection between *CAPN9* and glucose levels by way of the digestion of glucose. However, there is no further evidence from the literature that closely links *CAPN9* with blood glucose levels.

Either-tail-associated variants show non-monotonic allele frequency across the phenotype distribution

To investigate why VOI are significant for tail status and not the continuous trait, I plotted alternate allele frequency across the phenotype distribution for each missense VOI (Figure S2.3, Figure S2.4, Figure S2.5). Unlike VOI from other tail-status definitions, either-tail VOI showed evidence of non-monotonic allele frequencies across the phenotype distribution, with 7 of 17 missense either-tail VOI showing significant non-monotonicity (Methods). The non-monotonicity of allele frequency may be due to an atypical genetic model, such as vQTLs. As I stated before, transforming traits with either-tail status is theoretically a simple way to power vQTL discovery, because vQTLs are expected to have enriched or depleted allele frequencies in the tails of the phenotype distribution, relative to the middle of the distribution (Figure 2.11b).

Over half (10/17) of the missense either-tail VOI were lipoprotein A associations (Table S2.6). Using the more stringent threshold ($P = 5 \times 10^{-7}/29$), all but one variant (chr6:159813593:G:C in *PNLDC1*) remain VOI. Lipoprotein A has previously

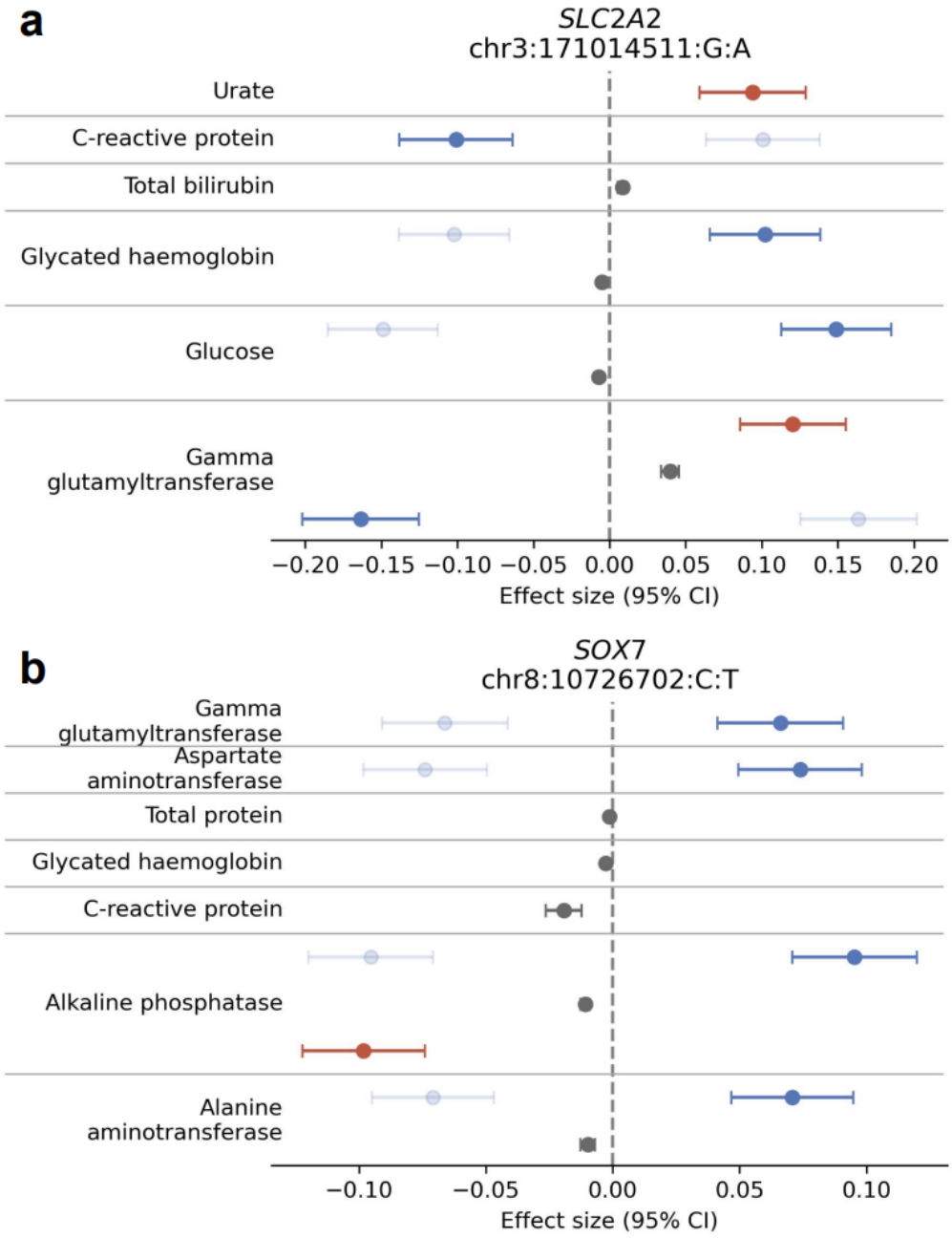

Figure 2.9: Effects on tail-status and continuous phenotypes for variants which are VOI for multiple biomarkers. **a**, Effect sizes for *SLC2A2*. **b**, Effect sizes for *SOX7*. Only tail-status and continuous phenotype associations which pass exome-wide significance ($P < 5 \times 10^{-7}$) are shown. Blue error bars: Low-tail associations. Red error bars: High-tail associations. Grey error bars: Continuous phenotype associations. For low-tail status associations, the flipped effect size is also shown as a faded error bar to orient to effect sizes from the high-tail and continuous trait phenotypes. All error bars indicate the 95% CI for effect size.

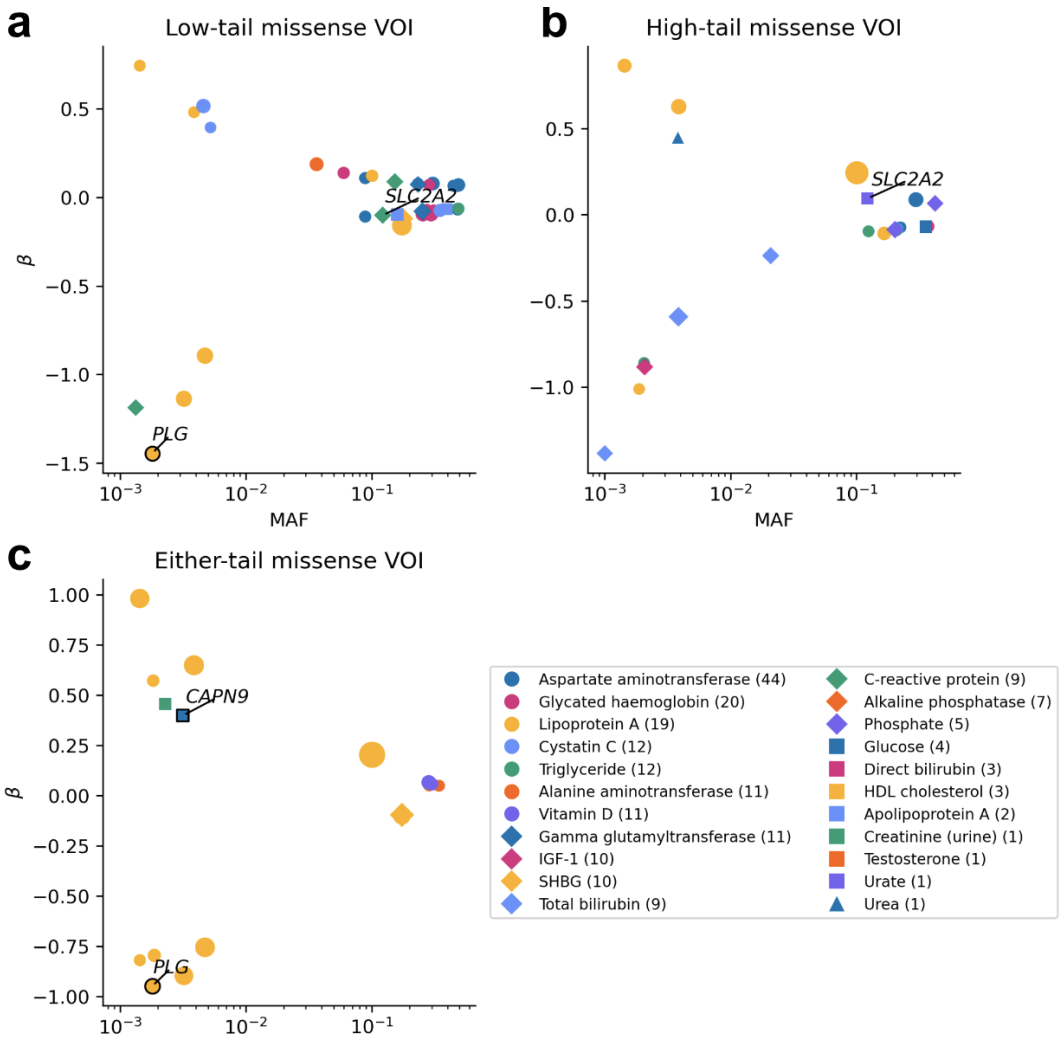

Figure 2.10: Effect size as a function of minor allele frequency for missense VOI. For VOI associated with **a**, low- **b**, high-, and **c**, either-tail dichotomised biomarker phenotypes. Outlined points are missense variants predicted to be damaging; all other variants are missense variants not predicted to be damaging. Labelled points are either damaging missense variants or variants which are VOI for multiple biomarkers. The number of VOI for each biomarker is indicated in parentheses in the legend. Point size is scaled by association significance.

been shown to be particularly enriched for vQTLs compared to other biomarkers[202]. All 10 lipoprotein A either-tail missense VOI lie in a $\sim$ 550kb-wide region of chromosome 6 (Figure 2.11). One of these missense VOI (chr6:160348796:C:G) in *SLC22A3* is within 11.2kb of a previously reported intronic lipoprotein A vQTL (rs405405, chr6:160360018:A:T) discovered in the Women's Genome Health Study[202], a dataset independent to UKB.

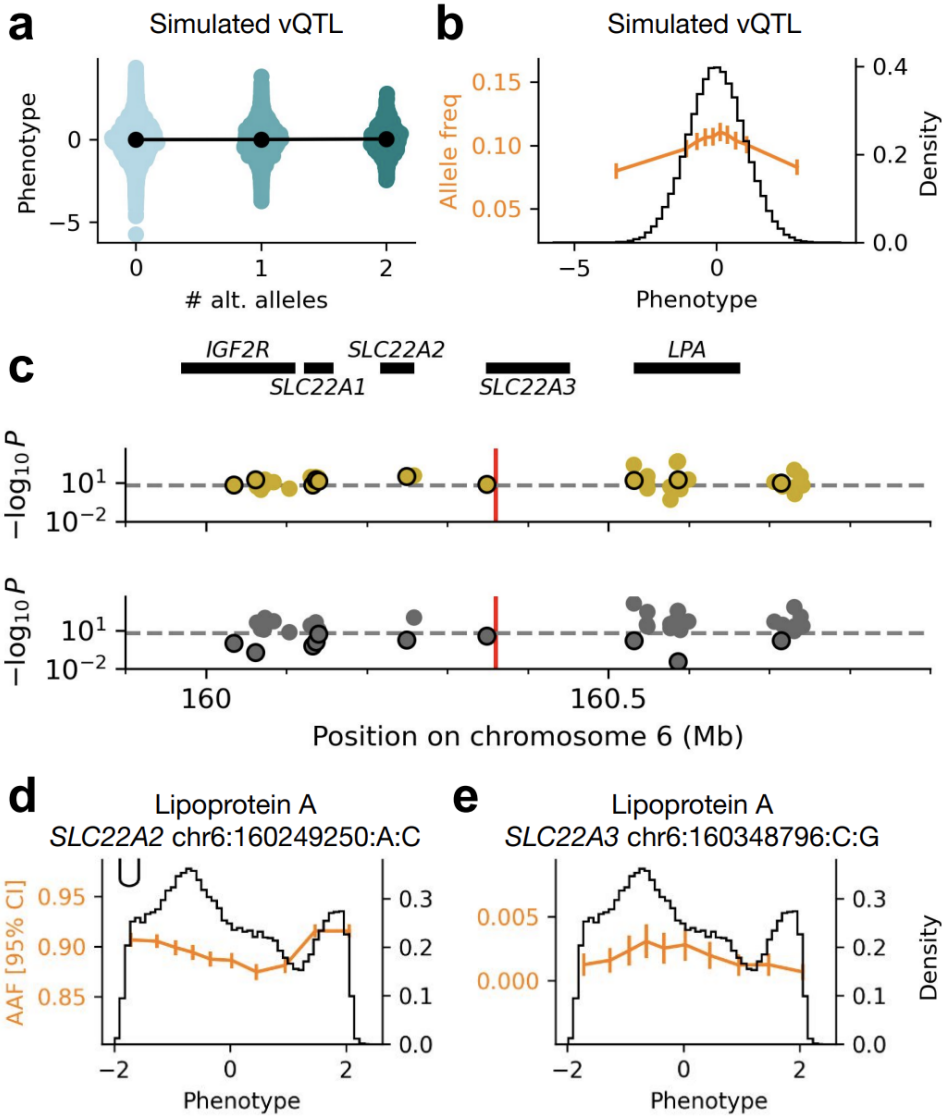

Figure 2.11: Non-monotonic allele frequency in either-tail associations. **a**, Genotype-stratified phenotype distributions and **b**, allele frequency across the phenotype distribution (orange) and the phenotype distribution (black) of a simulated variance-decreasing vQTL with scaling parameter $b = 1.1$, alternate allele frequency p=0.1, simulated for $n = 150,000$ individuals (Methods). **c**, Missense either-tail VOI for lipoprotein A are close to a previously reported vQTL. Red line: Location of previously reported lipoprotein A vQTL (rs405405, chr6:160360018:A:T)[202]. Dashed grey line indicates exome-wide significance ($P < 5 \times 10^{-7}$). Yellow points: either-tail status lipoprotein A associations. Grey points: continuous lipoprotein A associations. Circled points are VOI associations (significant for tail-status association but not the continuous traits). Gene regions from the MANE Select transcript[203] are shown in black boxes above. **d**, Alternate allele frequency across distribution (orange) and phenotype distribution (black) for missense either-tail lipoprotein A association which demonstrated significant monotonicity. **e**, Alternate allele frequency across distribution (orange) and phenotype distribution (black) for missense either-tail lipoprotein A association with non-monotonic allele frequency point estimates in *SLC22A3* close to previously published non-coding vQTL. Error bars for alternate allele frequency indicate the 95% CI calculated with $\pm 1.96 \times$ SE. Variant coordinates are in Genome Reference Consortium Human Build 38. vQTL, variance quantitative trait locus.

Tail-status transformation powers detection of liver disease risk variant

A high-tail status VOI (using the more stringent threshold of $P = 5 \times 10^{-7}/29$) for triglyceride (chr22:43928847:C:G, rs738409, AAF=0.206, P_{high}=9.95×10^{-9}, β_{high}=-0.0873) in *PNPLA3* (Patatin-like phospholipase domain-containing protein 3) is the "most robustly validated association" for risk of chronic liver disease[204] and has been shown to be associated with higher lipid accumulation in the liver[205,206], despite not being close to exome-wide significance in the continuous version of triglyceride (P_{cts}=1.36×10^{-3}, β_{cts}=-6.44×10^{-3}). The observed direction of effect of this variant in serum triglyceride ($\beta_{\text{high}} < 0$) seems to contradict its well-characterised link with increased lipids in the liver, especially given that serum triglyceride has been reported to be elevated in individuals with fatty livers[207]. One explanation for this is that this variant in *PNPLA3* has been shown to prevent lipid secretion by the liver[208,209], which would concurrently increase lipid accumulation in the liver while reducing the rate at which the liver releases lipids into the blood. Although the negative effect of this variant on serum triglyceride has been observed before in small cohorts ascertained for hepatic steatosis (n<600)[210–213], to the best of my knowledge it has not been observed in a large dataset not ascertained for hepatic steatosis. Early GWAS meta-analyses of more than 19,000 individuals did not observe an association in this variant for serum triglyceride[213,214]. Even in Genebass[199], which has more than double the number of UKB individuals used in my analysis, the variant was not exome-wide significant (Genebass P=2.90×10^{-3}). Gene-level burden of rare (MAF< 1%) missense variants among UKB individuals was also not associated with triglycerides (Genebass burden P=0.633)[199]. Thus, despite this variant being well characterised in the context of triglycerides and liver disease, it seems remarkably underpowered for detection using standard association methods. This is proof that by simply binarising a continuous trait into tail-status, it is possible to power detection of associations which cannot be found with the continuous trait. There may be variants in other traits where such associations may be powered for discovery by tail-status transformation.

2.3.4 Rare variant gene burden associated with multivariate biomarker extremes

Defining extremes on the joint distribution of biomarkers

Using combinations of biomarkers, rather than just a single biomarker, can improve diagnosis of disease. For example, high levels of alanine aminotransferase (ALT) and AST relative to alkaline phosphatase (ALP) and bilirubin can be indicative of liver disease[215], high urine sodium and low urate levels are associated with the syndrome of inappropriate secretion of antidiuretic hormone[216], and using combinations of blood lipids can improve prediction of coronary heart disease risk[217]. There have been previous attempts to jointly model biomarkers to improve disease diagnosis[218–221]. As I showed in a previous section, individuals with extreme levels of multiple biomarkers are enriched for disease. In this section I used Mahalanobis distance calculated on the joint distribution of biomarkers as a metric for overall biomarker extremeness, as has been done by previous studies[147,148]. However, to the best of my knowledge, Mahalanobis distance has never been used to study these blood and urine biomarkers in UKB, and has never been used for a genetic association study of biomarkers. I hypothesised that by using this metric I can boost power to detect genes associated with multiple biomarkers, which may be relevant to disease.

I subset to 28 biomarkers with at least 300,000 individuals and then subset to 199,933 individuals with no missing biomarker measurements. I performed inverse rank-normal transformation (IRNT) for each biomarker and calculated Mahalanobis distance on the joint distribution of transformed biomarkers. I converted Mahalanobis distance into P-values, using the assumption that Mahalanobis distance is expected to be distributed according to the chi-square distribution with k degrees of freedom, when calculated on a multivariate normal data with k variables[148].

Biomarker profile of individuals in the extremes of the joint biomarker distribution

Individuals in the extremes of the joint distribution of biomarkers ("outliers", Mahalanobis $P < 0.001$) have significantly different marginal distributions for each

biomarker (Kolmogorov-Smirnov test $P \leq 4.06 \times 10^{-14}$) compared to non-outlying individuals (Figure S2.6). The marginal distributions for outliers, relative to inliers, were particularly skewed for some biomarkers. For instance, outlying individuals had an excess of 25 percentage points relative to non-outliers in the upper half of the distribution for CRP, but only an excess of 2 percentage points in the lower half. The marginal distribution of albumin in outliers was skewed the other direction, with an excess of 27 percentage points relative to non-outliers in the lower half of the distribution, and an excess of 1 percentage point in the upper half of the distribution. The marginal distribution of urea in outliers was an example of balanced excess in both tails (lower tail: 12 percentage points, upper tail: 14 percentage points). There were also examples of biomarkers such as lipoprotein A and urine potassium, where the marginal distribution for joint-distribution outliers were less deviant and the total excess in the tails was only 11 percentage points.

Individuals in the extremes of the joint biomarker distribution are enriched for disease

Similar to the analysis of ICD-10 ORs as a function of number of outlying biomarkers, I stratified individuals by Mahalanobis P-value and tested the odds of having a given ICD-10 code in each stratum. As expected, I observed that individuals with smaller Mahalanobis P-values, who were farther out in the extremes of the joint biomarker distribution, tended to be significantly enriched for selected diseases (Figure S2.7; Table S2.7). The top two blocks with highest ORs (diabetes mellitus OR [95% CI] = 4.19 [3.66, 4.79], one-sided Fisher's exact test $P = 3.26 \times 10^{-76}$; renal failure OR [95% CI] = 3.89 [3.39, 4.46], P=1.67×10^{-65}) in the most extreme group of individuals (Mahalanobis $P < 0.001$) were the same top two blocks with the highest ORs from the previous analysis of number of outlying biomarkers. This makes sense: individuals in the extremes of the joint distribution are expected to overlap with individuals with outlying biomarkers.

Genes associated with the extremes of the joint biomarker distribution

The significant enrichment of multiple diseases in individuals with high Mahalanobis distance suggests that there is disease-relevant signal in those extreme individuals. To determine whether genetic variation can explain why individuals have high Mahalanobis distance I used a scalable mixed model framework that allows for the inclusion of related individuals (SAIGE[222]) to perform rare (MAF $< 1\%$) pLoF gene burden association testing using Mahalanobis distance as a phenotype. I hypothesised that using Mahalanobis distance as a phenotype might uniquely power associations which would not be discovered by combining summary statistics across biomarkers. Therefore, I also performed rare (MAF$< 1\%$) pLoF burden tests in each biomarker included in the joint distribution on which Mahalanobis distance was calculated. Then I combined the burden P-values across biomarkers using the Cauchy combination test, a well-powered test for combining P-values[223]. I also combined P-values by taking the minimum across biomarkers. I compared the P-values between the burden test of Mahalanobis distance and the combined P-values across biomarkers (Figure 2.12; Table S2.8).

In 14 genes, burden of rare (MAF $< 1\%$) pLoF variation was significantly associated ($P < 2.80 \times 10^{-6}$, Bonferroni correction for 17,873 genes tested) with Mahalanobis distance (Figure 2.12; Table S2.8). For all 14 genes, pLoF burden was positively correlated with increased Mahalanobis distance. All but one (*ZBTB2*) were significant and had lower P-values when combining P-values across biomarkers, whether using the Cauchy combination test or taking the minimum P-value (Figure 2.12a,b; Table S2.8).

The only gene significantly associated with Mahalanobis distance but where combining P-values across biomarkers did not result in a significant P-value was *ZBTB2* (Zinc Finger And BTB Domain Containing 2; burden $P = 4.12 \times 10^{-8}$) (Figure 2.12). Rare pLoF burden in *ZBTB2* was not significantly associated with any of the tested biomarkers when using Bonferroni correction for 17,873 genes tested, with the most significant single-biomarker burden association being with calcium ($P = 4.61 \times 10^{-3}$) (Figure 2.12c; Table S2.9). The minor allele count

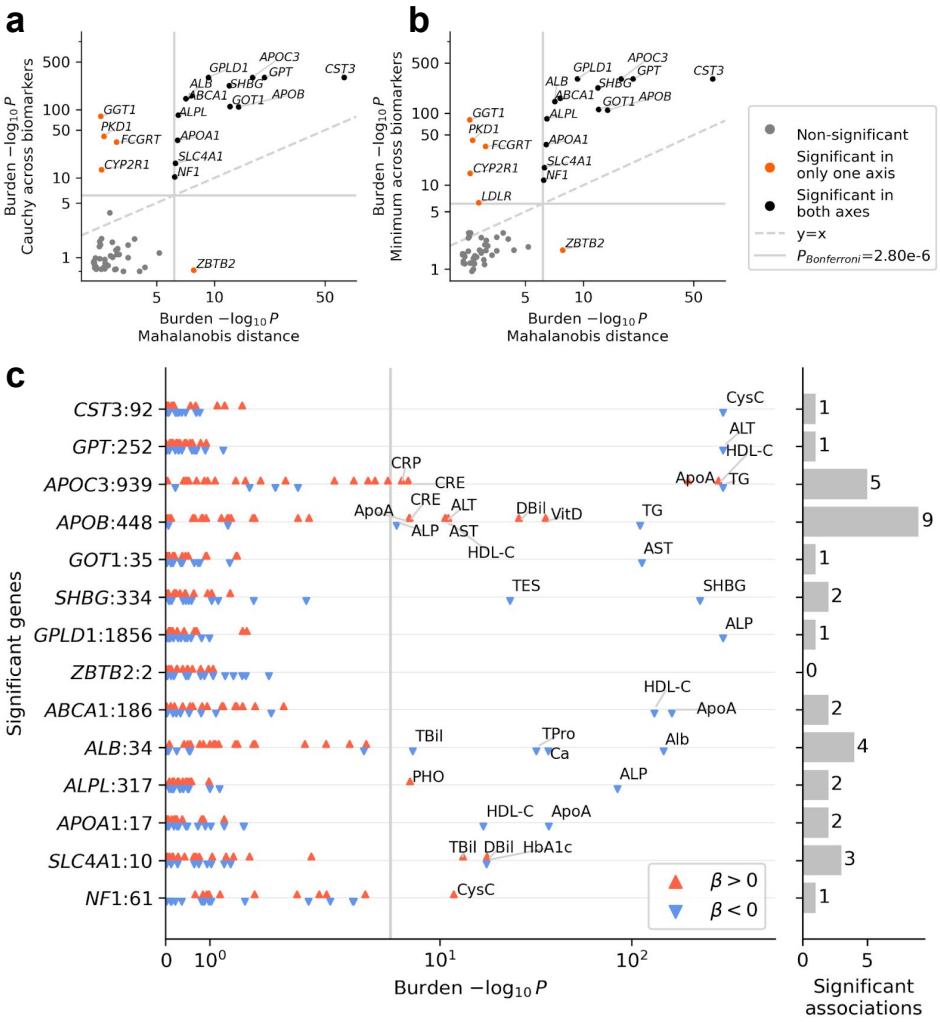

Figure 2.12: Gene burden associations with Mahalanobis distance calculated on the joint distribution of biomarkers. **a**, Comparison of association strength with Mahalanobis distance versus combining P-values across single biomarkers using the Cauchy combination test. **b**, Comparison of association strength with Mahalanobis distance versus combining P-values across single biomarkers by taking the minimum P-value. **c**, Single-biomarker associations for genes significantly associated with Mahalanobis distance. Genes are sorted on the y-axis in order of descending pLoF burden association strength with Mahalanobis distance, with the most significantly associated genes at the top. Gene labels include the gene symbol, followed by a colon and the minor allele count for the pLoF burden test of Mahalanobis distance. Upward orange and downward blue triangles indicate single-biomarker associations with positive and negative, respectively, pLoF burden effect size. The exome-wide significance threshold (Burden $P<2.80 \times 10^{-6}$) is shown as a vertical grey line. The number of significant single-biomarker burden associations for each gene is shown in the bar plot to the right of the main figure.

(MAC) of rare pLoF variants was very low in *ZBTB2* (MAC = 2), which makes it difficult to precisely estimate the burden effect size and makes it more likely that this is a spurious association. However, *ZBTB2* is a transcription factor and master regulator of a major tumour suppressor pathway[224–226], which may provide a plausible mechanism to the theory that pLoF burden in *ZBTB2* could affect multiple biological processes at once, driving individuals into the extremes of the joint distribution of biomarkers.

Among the 13 significant genes where combining *P*-values outperformed the *P*-value from the burden test of Mahalanobis distance, five genes (*GPT, CST3, GPLD1, GOT1, NF1*) were significantly associated with just one biomarker (Figure 2.12c; Table S2.9). This implies that the burden effect of these genes on Mahalanobis distance was largely driven in one dimension (i.e. a single biomarker) of the joint distribution of biomarkers. The eight genes in the remaining majority appeared to be driven by multiple biomarkers (Figure 2.12c; Table S2.9). The gene significantly associated with the most biomarkers was *APOB*, with nine: triglycerides, vitamin D, direct bilirubin, HDL-C, ALT, AST, ApoA, serum creatinine, and ALP. *APOB* codes for the protein ApoB, which helps to transport low density lipids through the blood[227]. The biomarker most strongly associated with *APOB* burden was triglyceride (burden P=1.22×10^{-111}), likely because transport of serum triglyceride in lipid particles is enabled by ApoB[227]. ApoB is also a biomarker for CVD, with higher levels of ApoB increasing risk of disease[228,229]. This strong link with CVD risk may result in *APOB* burden being correlated with other biomarkers for CVD risk. For example, vitamin D, bilirubin, HDL-C and ApoA (a component of HDL-C) are negatively correlated with CVD risk[230–233] and, correspondingly, *APOB* burden (lower ApoB, lower CVD risk) is positively correlated with these biomarkers. Of the liver- (ALT, AST, ALP) and kidney-related (serum creatinine) biomarkers, the link with *APOB* burden is less clear: increased ALT is associated with increased ApoB[234] and risk of stroke but not coronary heart disease[235], increased AST is occasionally associated with increased CVD risk[235,236], increased ALP is moderately

associated with increased CVD risk[235,237–241], and increased serum creatinine is associated with increased CVD risk[242].

2.4 Methods

2.4.1 Defining clinical reference ranges

I used reference ranges provided by the NHS to define the normal ranges of biomarker levels in UKB. I sought to keep the number of sources to a minimum, to maximise consistency across biomarkers. My sources were the East Sussex Healthcare NHS Trust Clinical Biochemistry Reference Ranges Handbook[140], North Bristol NHS Trust Age-Related Reference Ranges[141,243], Gloucestershire Hospitals NHS Trust Chemical Pathology[244], South Tees Hospitals NHS Trust Pathology Tests[245], NHS Health A to Z[246], and Royal United Hospitals Bath NHS Trust Pathology Services[247].

2.4.2 Quality control of UKB exome sequencing data

Ancestry labels were assigned by Duncan Palmer.

QC was separately performed for the 200k[248] and 450k[114] tranches of the UKB ES data, even though the 450k tranche is a superset of the individuals in the 200k tranche. For both the 200k and 450k tranches, I filtered to European ancestry using super-population ancestry labels assigned with a random forest (RF) classifier by Duncan Palmer. The method for assigning population labels is described in more detail in the "Sample QC and assigning population labels" section of Chapter 4 methods.

Exome sequencing 200k tranche

Where possible, my QC pipeline followed the approach of the Genome Aggregation Database (gnomAD) version 3.0[113]. Raw data was available in the form of multi-sample project-level variant call format (VCF) files, joint-called using DeepVariant[249] and GLnexus[250]. I split multiallelic sites and left-aligned indels using bcftools[251]. I calculated sample-level metrics using the `sample_qc()` function in Hail[252]. I

Metrics	MAD thresholds
call_rate n_snp	$[24, \infty)$
n_deletion n_insertion n_transition n_transversion r_insertion_deletion	[8, 8]
n_het n_hom_var n_non_ref	[16, 16]
n_singleton r_het_hom_var	[24, 24]

Table 2.4: Variant-level QC median absolute deviation thresholds.

removed 848 individuals who fell outside MAD thresholds defined for 12 sample-level principal component (PC)-regressed QC metrics also used by gnomAD v3.0[113]:

I set genotypes to missing if genotype quality (GQ) < 20, depth (DP) < 10, P-value from one-sided binomial test of alternate allele read depth relative to reference plus alternate allele read depth $< 1 \times 10^{-3}$ for heterozygous calls in single nucleotide variants (SNVs), alternate allele allelic balance < 0.3 for indel heterozygous calls, and alternate allele allelic balance < 0.7 for indel homozygous alternate calls. I kept all variants within the 50 bp-padded target capture region and not in a low complexity region[253]. I used Genome Analysis Toolkit (GATK)[254] to calculate the Phred-scaled P-value for the exact test of excess heterozygosity (`ExcessHet`), removing variants if `ExcessHet` ≤ 54.69, as per GATK-suggested best practices. Finally, I removed all invariant sites (MAC $= 0$). This resulted in a total of 12,962,630 curated variants in the analysis-ready dataset of 199,795 individuals.

Exome sequencing 450k tranche

A detailed description of the QC for the 450k tranche of the UKB ES data can be found in the "Exome sequencing data QC" section of Chapter 4 methods.

2.4.3 Variant annotation

All variant annotation was performed by Frederik Lassen and Barney Hill.

Annotation was performed separately for the 200k and 450k tranches of ES data. Briefly, for the 200k tranche, variants were annotated as protein-truncating variants ("PTV"), "damaging missense", "other missense" (missense variants not predicted to be damaging), synonymous variants, and other variants. The details of the annotation procedure can be found in Lassen *et al.*[255]. For the 450k tranche, coding variants were annotated as pLoF, "damaging missense", "other missense" and "synonymous". A detailed description of the variant annotation for the 450k tranche can be found in the "Variant consequence annotation" section of Chapter 4 methods.

2.4.4 Selection of blood and urine biomarkers

Raw biomarkers

The 32 biomarkers included in this analysis included 28 blood biochemistry biomarkers (UKB Category 17518: alanine aminotransferase [UKB field ID: 30620], albumin [30600], alkaline phosphatase [30610], apolipoprotein A [30630], apolipoprotein B [30640], aspartate aminotransferase [30650], C-reactive protein [30710], calcium [30680], cholesterol [30690], creatinine [30700], cystatin C [30720], direct bilirubin [30660], gamma glutamyltransferase [30730], glucose [30740], glycated haemoglobin [30750], HDL cholesterol [30760], IGF-1 [30770], LDL direct [30780], lipoprotein A [30790], phosphate [30810], SHBG [30830], testosterone [30850], total bilirubin [30840], total protein [30860], triglycerides [30870], urate [30880], urea [30670], vitamin D [30890]) and four urine assay biomarkers (UKB Category 100083: creatinine in urine [UKB field ID: 30510], microalbumin [30500], potassium [30520], sodium [30530]). Rheumatoid factor (UKB field ID: 30820) and oestradiol (30800) were excluded from all analysis because they were flagged for QC issues by Sinnott-Armstrong *et al.*[85].

Quality controlled biomarkers

All biomarker QC was performed by Samvida Venkatesh.

Biomarkers underwent a QC procedure intended to match that used by Sinnott-Armstrong *et al.*[85]. Total cholesterol, LDL-C and ApoB were excluded from the list of QCed biomarkers due to being strongly affected by statins[85]. These three biomarkers were not included in any downstream analysis involving QCed biomarkers.

2.4.5 Outlier overlap enrichment between pairs of biomarkers

Enrichment tests

I used the two-sided Fisher's exact test[256] to calculate the enrichment of overlap in outliers between pairs of biomarkers. I performed these overlap enrichment tests separately for below-, above-, or outside-range outliers.

Empirical null simulations

I simulated pairs of biomarkers for 200,000 individuals from bivariate normal distributions with covariance parametrised by the square root of a coefficient of determination (R^2) sampled at 50 intervals between 0 and 0.99, inclusive. For three reference ranges (99%, 95%, 90%), with each reference range covering the middle percentiles of the biomarkers, I defined outliers as individuals outside the reference range. I proceeded to calculate outlier overlap enrichment as I did for the non-simulated phenotypes. I performed this procedure fifty times for each coefficient of determination and reported the mean OR across replicates.

2.4.6 Enrichment of ICD-10 codes among individuals with outlying biomarkers

Using the UKB Data Showcase entry for ICD-10 codes (data field: 41270), I selected the five most prevalent (highest counts) ICD-10 chapters with "disease" in the name, as a heuristic for the chapter being disease related. I included ICD-10 codes related to

transport accidents (ICD-10 codes V01-V99), as a negative control, as these were not expected to strongly associate with an individual's number of outlying biomarkers.

I tested for enrichment of ICD-10 codes for a given range of outlying biomarker counts using Fisher's exact test[256]. For each set of individuals with a given range of outlying biomarker counts (e.g. five to nine outlying biomarkers), enrichment of ICD-10 codes was calculated relative to all other individuals not in that set.

2.4.7 Enrichment of predicted loss-of-function variants in GEL PanelApp rare disorder genes among individuals with extreme biomarkers

I curated pairs of biomarkers and related rare genetic disorders from the GEL PanelApp[182,183] using evidence from literature (Table S2.4). To limit multiple testing burden, I restricted further analysis to 12 "high confidence" pairs which I believed to have strong supporting evidence (Table S2.4) . The GEL PanelApp also provided an expertly-curated list of clinical-grade genes linked with each disorder; I only used genes where a single pathogenic allele was sufficient to result in the disorder (Mode of inheritance = Monoallelic). I defined extreme single biomarker levels using clinical reference ranges from the NHS (Table 2.1) and various percentile thresholds. Annotation of pLoF variants followed LOFTEE HC guidelines[113] (see "Variant consequence annotation" section of Chapter 4 methods. I used Fisher's exact test[256] to estimate enrichment of rare (MAF $< 0.1\%$) pLoF variants in GEL PanelApp rare disorder genes in individuals with extreme biomarker levels.

When testing enrichment using more extreme percentile thresholds than the typical 95% centred reference range I always use the same set of inlying individuals (middle 95% of biomarker distribution). The only change is to the set of outlying individuals, which is defined by the more extreme percentile threshold (1%, 0.1%, 0.01%).

2.4.8 Defining tail-status phenotypes

Dichotomisation of QCed continuous biomarkers into tail-status phenotypes was performed by thresholding the continuous phenotype on 10% and 90% percentile thresholds, corresponding to lower and upper tails containing 10% of the samples

for which a given biomarker was defined. The lower-tail status phenotype was dichotomised into individuals in the lower tail (cases, percentile<10%) and all other individuals (controls, percentile $\geq$ 10%). The upper-tail status phenotype was dichotomised into individuals in the upper tail (cases, percentile $>$ 90%) and all other individuals (controls, percentile$\leq$ 90%). The either-tail phenotype was dichotomised into individuals in the lower or upper tail (cases, percentile $<$ 10% or $>$ 90%) and all other individuals (controls, percentile $\geq$ 10% and $\leq$ 90%).

2.4.9 Single-variant tests on biomarker continuous and tail-status phenotypes

I used REGENIE v2.2.4[93] to perform single-variant association tests on continuous and dichotomised tail-status biomarker phenotypes in the ES data.

REGENIE involves two steps: First, a whole genome regression model is fit to a subset of variants. Second, marginal associations are tested for the larger set of variants, conditional on the whole genome regression model.

In the first step, I fit the whole genome model on genotype array data which had been subset using PLINK[257] to individuals with $<$ 10% missingness and variants with MAF $>$ 1%, MAC $\geq$ 100, missingness $<$ 1%, Hardy-Weinberg P>1×10^{-15}. The covariates used were sex, age, $sex \times age$, age^2, $sex \times age^2$, first 21 PCs (calculated by UKB), UKB assessment centre and UKB genotyping batch.

In the second step, I tested for marginal associations using the same covariates in the ES data subset to European individuals who passed QC (see Methods above) variants with MAF $>$ 0.1%. For the dichotomous tail-status phenotypes I used approximate Firth correction for associations where $P < 0.01$ (`--firth --approx --pThresh 0.01`).

2.4.10 Testing odds of low- and high-tail associations having negative or positive effect on the continuous trait

Using Fisher's exact test[256] I tested whether variants associated with the single-tail status were more likely to have an exome-wide significant association in the expected direction for the corresponding continuous trait (negative effect if low-tail,

a.

	$\beta_{cts} < 0$ and $P_{\text{cts}}<5 \times 10^{-7}$	Other
$P_{\text{low}} < 5 \times 10^{-7}$	1,616	1,502
$P_{\text{low}} \geq 5 \times 10^{-7}$	4,322	4,169,474

b.

	$\beta_{cts} > 0$ and $P_{\text{cts}}<5 \times 10^{-7}$	Other
$P_{\text{high}} < 5 \times 10^{-7}$	1,319	1,375
$P_{\text{high}} \geq 5 \times 10^{-7}$	4,311	4,169,909

Table 2.5: Contingency table for testing effect direction of low- and high-tail associations. **a**, Contingency table for testing whether low-tail associations are more likely to have a significant negative association with the continuous trait. **b**, Contingency table for testing whether high-tail associations are more likely to have a significant positive association with the continuous trait.

positive effect if high-tail). I used a one-sided test because of my prior hypothesis that there would be enrichment (OR > 1) rather than depletion. I reported the two-sided 90% CI so that the lower bound would correspond to the lower bound of the one-sided 95% CI. The *P*-values were too small for Python's numerical precision, so I reported "$P<1 \times 10^{-100}$" for simplicity.

2.4.11 Defining significant non-monotonic allele frequency

Significant non-monotonicity in allele frequency across the phenotype distribution was defined as having at least one 95% CI for allele frequency in any of the middle phenotype bins that was less than (or greater than) and non-overlapping with any of the 95% CIs to the left or right.

2.4.12 Simulating vQTLs

To simulate a vQTL with alternate allele frequency p and scaling parameter b for n individuals, I calculated the expected number of individuals for each diploid genotype under Hardy-Weinberg equilibrium[258] and simulated the phenotypes for individuals of each genotype using a normal distribution with variance parametrised by scaling parameter (Table 2.6). Following phenotype simulation for each genotype group, phenotypes were standardised to have mean = 0, variance = 1 across all individuals.

Genotype	Expected number of individuals	Variance of normal distribution
Homozygous reference allele (RR)	$n(1-p)^2$	$\|b\|^{sign(b)}$
Heterozygous (RA)	$2n(1-p)p$	$\|b\|^{2sign(b)}$
Homozygous alternate allele (AA)	np^2	$\|b\|^{3sign(b)}$

Table 2.6: Parameters for simulating vQTL.

2.4.13 Defining Mahalanobis distance on joint distribution of biomarkers

Calculation of Mahalanobis distance requires a fully defined dataset with no missing values. From the 29 QCed biomarkers I removed biomarkers with fewer than 300,000 samples. This only removed urine microalbumin ($n = 121,173$), leaving 28 biomarkers. Then I subset to individuals who had all 28 biomarkers defined, leaving 199,933 individuals. I performed IRNT on each biomarker and calculated Mahalanobis distance[146] on the joint distribution of IRNTed QCed biomarkers.

2.4.14 Gene-based burden tests on Mahalanobis distance and single biomarkers

I used SAIGE version 1.1.9[222] on the DNAnexus Research Analysis Platform (RAP) to test for gene-level rare (MAF $< 1\%$) variant burden associated with Mahalanobis distance, using individuals in the European-ancestry subset of the ES 450k tranche. My reason for not using REGENIE throughout this chapter was due to practicality. At the time of analysis I had already developed a pipeline for gene-level testing in SAIGE but not in REGENIE. Instead of creating a gene-level testing pipeline in REGENIE, I used the SAIGE pipeline described in the "Rare variant and gene-level testing in exome sequencing data" section of Chapter 4 methods, with the exception of not using IRNT on Mahalanobis distance (`--invNormalize=FALSE`), to preserve structure in the distance distribution. Although I tested multiple thresholds for maximum MAF (`--maxMAF_in_groupTest=0.0001,0.001,0.01`) and multiple variant consequence categories, I only used pLoF burden associations results for variants with maximum MAF of 0.01.

Using the same SAIGE pipeline, genes where the Mahalanobis distance burden association was $P < 0.001$ for at least one variant consequence category were also tested for association with single biomarkers. The single biomarkers were transformed using IRNT applied within SAIGE (`--invNormalize=TRUE`).

2.5 Discussion

In this chapter I collated reference ranges for blood and urine biomarkers in UKB and determined whether genetically European individuals in UKB had observed biomarker levels outside these reference ranges. For each biomarker I quantified the percentage of individuals with outlying values. I demonstrated that for some biomarkers there was a substantial percentage of individuals with outlying values and that, aggregating across biomarkers, nearly all individuals had one or more outlying biomarker values. I showed that disease risk across various disease categories was significantly higher for individuals with many outlying biomarker values, with risk increasing as a function of the number of outlying biomarkers.

Seeking a genetic explanation for extreme biomarker levels, I found that individuals with extreme levels of certain biomarkers, especially cholesterol-related biomarkers, were enriched for pLoF variants in genes linked with rare disorders. To further interrogate the genetics underlying biomarker extremes, I performed exome-wide single-variant association tests on biomarker tail-status phenotypes, where cases are individuals in the extremes of biomarker distributions and controls are individuals not in the extremes. I also performed single-variant association tests on the continuous biomarkers, and compared association strength of each variant with the tail-status and continuous phenotypes. Using the tail-status phenotype, I demonstrated for the first time in a large unascertained cohort an association between serum triglyceride upper-tail status and a missense variant in *PNPLA3* known to increase hepatic fat.

Finally, I sought to identify genes associated with widespread biomarker disruption, as this was likely to be linked with disease. I constructed a metric using Mahalanobis distance for an individual's extremeness in the joint distribution of

biomarkers and scanned exome-wide for gene burden associated with the metric. The majority of genes significantly associated with Mahalanobis distance were significantly associated with multiple single biomarkers, validating the Mahalanobis distance association test as an approach to identify genes which affect multiple biomarkers.

The choice of the blood and urine biomarkers was largely influenced by their high sample size in UKB ($n > 400,000$), which I expected to result in higher statistical power to detect associations, compared to other biomarkers in UKB with lower sample size. For example, the plasma NMR metabolic biomarkers[84] are more granular, with around 250 metabolites, but were available for fewer individuals ($n = 280,000$). It also would have been impractical in the framework of this analysis to analyse hundreds of biomarkers.

Defining what it means to be "normal" is complicated. I simplified by making the assumption that normal and abnormal (also called "outlying") individuals can be defined by whether they have observed biomarker values within or outside a reference range defined by the NHS. This approach does not account for within-individual variability of biomarker measurements, which can be influenced by many factors[259,260]. For example, if I had controlled for the use of cholesterol-lowering medications like statins, the adjusted percentage of above-range outlying individuals for cholesterol-related biomarkers (total cholesterol, LDL-C, ApoB) would likely be higher than what I have reported. Furthermore, for some biomarkers there were multiple cut points which could indicate abnormal levels. For example, vitamin D insufficiency is defined by values < 50 nmol/L while severe deficiency is defined as < 25 nmol/L[261].

Age-specific ranges were merged to match UKB participant enrolment age. This ensures individuals are not falsely classified as having abnormal biomarker levels, but could result in some individuals being classified as having normal biomarker levels even though they are outside their age-specific reference range. To improve specificity, age-specific reference ranges could be applied directly only to individuals in the matching age range.

Here I only focus on genetically European individuals. Some biomarkers can significantly differ between genetic ancestry groups[262]. Ideally, this work would be repeated for other genetic ancestry groups using ancestry-specific reference ranges, however these ranges are not readily available via the NHS.

When testing for disease enrichment among individuals with multiple outlying biomarkers, I used ICD-10 codes to define whether an individual was diseased. The use of ICD-10 codes is a standardised approach to defining disease[263]. The tree structure of ICD-10 classification enables nested testing of chapters and blocks within chapters. For the sake of simplicity, I only tested the five disease-related chapters with highest counts in the UKB Data Showcase and, within those chapters, only the three disease code blocks with the highest counts in the UKB Data Showcase. It is possible to perform enrichment tests in individuals with high counts of outlying biomarkers for all disease-related chapters, blocks within chapters, and single ICD-10 codes within blocks. This would add more breadth and depth to this particular enrichment analysis, but was not in the remit of this project.

When testing individuals with extreme biomarker levels for enrichment of pLoF variants in genes that cause rare disorders, I used the GEL PanelApp[182,183] to manually curate which disorders and genes were included. Similar options for curation include OMIM[264] and ClinVar[265], however the GEL PanelApp is most relevant to individuals in the UKB, as it uses the same gene panels as the NHS Genomic Medicine Service, and was most convenient to integrate into the analysis pipeline. Pairing biomarkers with rare disorders involved manual curation supported by literature. Given the large number of GEL panels ($n = 451$), it was not feasible to evaluate evidence in the literature for every biomarker-disorder pair. Taking a hypothesis-free approach by testing pLoF enrichment for every pair would likely be computationally tractable, but would necessitate a more stringent threshold for significance to account for multiple testing.

The dichotomisation of continuous biomarkers into tail-status phenotypes was intended to power discovery of variants enriched in the extremes of biomarker distributions. The dichotomisation resulted in a general loss in power, with far

fewer tail-status associations (1,233) than continuous trait associations (3,106) (Table S2.6). This is an expected phenomenon, as dichotomisation results in a loss of information about the continuous trait[266]. However, my results demonstrate that dichotomisation can also uniquely power associations which would be overlooked with traditional association tests of continuous traits.

Tail-status phenotypes for each biomarker were defined using the same percentile thresholds to standardise power loss due to dichotomisation. Choosing a different percentile threshold than the bottom or top 10% to define tail-status could affect the association test results. An alternative to percentiles could be to use the reference range limits, as these may be more biologically relevant to disease liability. For example, the upper limit of the HbA1c reference range (48 mmol/mol or 6.5%) corresponds directly to the WHO-recommended threshold for diagnosing diabetes[267]. However, association tests on tail-status defined by reference ranges might be very similar to previously published association tests on disease status, diminishing the novelty of the work.

When testing tail-status phenotypes, I only performed single-variant association tests and only used individuals in the ES 200k tranche. I could redo this analysis in the ES 450k tranche, more than doubling the sample size and increasing the statistical power. ES variants, although tending to have a stronger biological prior for being causal due to ES targeting coding variants, are relatively sparse genome-wide (144,483 variants with MAF>0.1%) relative to imputed genotypes (13,117,850 variants with MAF>0.1%; see Imputed data QC methods). Using a broader set of variants in the imputed dataset could result in additional associations, but those associated variants may be more difficult to trace back to a gene. Another potential improvement to the tail-status GWAS analysis could be to perform gene-level testing of rarer variants (MAF<0.1%). This would aggregate signal across variants, possibly providing power to detect associations in rare or ultrarare variants with large effect sizes. Such large-effect variants would be of particular interest in this analysis, as they would be expected to push individuals into the extremes of phenotype distributions.

To construct the Mahalanobis distance metric on the joint distribution of biomarkers, I first performed IRNT on the biomarkers. This enabled the conversion of Mahalanobis distance to P-values, quantifying the probability of observing certain individuals in the joint distribution. However, the rank-based transformation loses some structure in the phenotypes. Using minimally processed (e.g. mean-centred and variance-normalised) raw biomarkers would retain most of the structure in the phenotype distributions and possibly improve power for some associations. Mahalanobis distance requires a fully-defined dataset, and by subsetting to individuals with defined values for all selected biomarkers I substantially reduce the sample size. There is around a 2-fold difference in sample size between the Mahalanobis distance association tests ($n = 200,000$) and the single-biomarker association tests ($n \sim 400,000$), which may explain why combining P-values across single-biomarker association tests generally resulted in stronger associations (lower P-values). Restricting the single-biomarker association testing to the same set of individuals tested for Mahalanobis distance association (i.e. downsampling) might provide a better assessment of statistical power gained or lost by using Mahalanobis distance in a fully-defined dataset. Alternatively, imputing missing values in a dataset with missingness would have increased sample size for the association tests on Mahalanobis distance, possibly boosting power for discovery. If the method for imputation was trained on the same data used for analysis (e.g. by using the covariance structure of UKB biomarkers), one could argue it might not substantially increase the effective sample size. However, this could still be worthwhile, as any power gained is useful. Finally, I calculated Mahalanobis distance across all biomarkers in a phenotype-agnostic manner, only excluding biomarkers on the basis of low sample size. As implied by Solberg in 1995[148], calculating Mahalanobis distance on subsets of biomarkers grouped by biological pathways (e.g. liver function-related biomarkers, kidney function-related biomarkers, blood lipids) could refine the association signal to specific clinical subprofiles.

This work provides a systematic framework for understanding the prevalence, disease relevance, and genetic basis of extreme biomarker levels in UKB, offering

new insights into both individual and systemic health. By showing that nearly all individuals have at least one outlying biomarker and that disease risk increases with the number of outlying biomarkers, we have a better understanding of health and disease in UKB individuals. The validation that UKB individuals with extreme biomarker levels, particularly for cholesterol-related traits, are enriched for rare pathogenic variants underscores the role of biomarkers for rare disease diagnosis. Furthermore, by demonstrating that dichotomised tail-status phenotypes can enhance the detection of disease-linked genetic variants and that Mahalanobis distance can identify genes affecting multiple biomarkers, this work highlights the value of alternative phenotype definitions in genetic association studies. More broadly, these results illustrate how integrating phenotypic extremes with genetic analysis can reveal previously undetected biological relationships, offering a powerful approach for uncovering genetic factors underlying complex traits and disease.

3

Individuals with phenotypes misaligned with polygenic prediction are enriched for rare pathogenic variation

Contents

3.1 Abstract 68
3.2 Introduction 69
3.3 Main 72
3.3.1 Identifying individuals with misaligned continuous phenotypes 72
3.3.2 Individuals with misaligned continuous phenotypes are enriched for predicted loss-of-function variants in monogenic disorder genes 74
3.3.3 Individuals whose disease status is misaligned with genetically expected risk are enriched for predicted loss-of-function variants in monogenic genes 78
3.3.4 Disease controls with high polygenic risk are enriched for rare protective predicted loss-of-function variants in monogenic disorder genes 82
3.3.5 Individuals whose continuous phenotype misaligns with genetic expectation are enriched for predicted damaging variants in genes related to rare disorders 84
3.3.6 Individuals with genetically unexpected phenotypes are nominally enriched for singleton predicted loss-of-function variants 88
3.3.7 Exome-wide scan for genes associated with misaligned status in continuous traits 91

3.3.8 Exome-wide scan for predicted loss-of-function variant burden associated with protective misaligned disease status 98
3.4 Methods 102
3.4.1 Exome sequencing data QC 102
3.4.2 Phenotype curation 103
3.4.3 Monogenic and rare disorder gene curation 104
3.4.4 Misaligned phenotype classification 105
3.4.5 Enrichment testing of carrier status and misaligned status 107
3.4.6 Hypergeometric test for probability of observing no carriers 109
3.4.7 Association of carrier status with polygenic risk score for dichotomous traits 109
3.4.8 Singleton burden in intolerant genes 110
3.4.9 Exploratory burden testing on phenotype misalignment 111
3.5 Discussion 113

3.1 Abstract

Polygenic scores (PGS) predict complex traits and stratify disease risk but often fail to fully capture individual-level variation. Misaligned individuals, whose observed phenotypes deviate from their genetically expected values based on PGS, provide a powerful model for identifying factors beyond common-variant effects, including additional genetic factors. Here, I extend existing misalignment classification frameworks to seven continuous and three dichotomous traits, assessing whether misaligned individuals are enriched for rare (MAF $< 0.1\%$) genetic variation, including pLoF variants, damaging missense variants, and exome-wide singleton burden. I use empirical null simulations to mitigate spurious associations in traits where PGS explain less variance.

For the first time, I demonstrated that individuals with lower-than-expected LDL-C and higher-than-expected stature are enriched for predicted damaging missense variants in *PCSK9* (FDR-adjusted $P = 3.75 \times 10^{-3}$, empirical $P = 0.0230$) and *FBN1* (FDR-adjusted $P = 1.99 \times 10^{-2}$, empirical $P = 9.99 \times 10^{-3}$), respectively. Additionally, I identified significant enrichment of pLoF variants in *GORAB* among individuals misaligned for lower-than-expected bone mineral density (FDR-adjusted $P = 0.0155$, empirical $P = 1.00 \times 10^{-4}$). Using an individual's misalignment

classification as a phenotype, I performed an exome-wide scan across seven traits, to search for protective and pathogenic genes and identified *KANK1* as a potential therapeutic target for primary ovarian insufficiency.

For dichotomous disease status traits, I demonstrated that among individuals diagnosed with T2D, carriers of rare pathogenic pLoF variants had significantly lower polygenic risk than non-carriers (FDR-adjusted one-sided t-test $P = 6.01 \times 10^{-8}$). For the first time, I showed that disease controls carrying rare protective pLoF variants for T2D and coronary artery disease (CAD) had nominally higher polygenic risk (one-sided t-test $P < 0.05$) than non-carriers.

This study highlights the power of misalignment-based analyses in uncovering rare genetic contributors to complex traits and diseases, offering a pathway for improved molecular diagnoses and precision medicine approaches.

3.2 Introduction

Modelling how an individual develops disease is key to preventing disease. One approach to modelling genetic disease risk is the liability threshold model[268,269]. This model assumes that for any disease there underlies a continuous "liability", which may comprise of both genetic and environmental components[269], such that individuals develop disease if their liability exceeds a certain threshold. The genetic risk component could be further separated into risk contributed by common variants and rare variants, for a given heritable complex phenotype[270].

One way to quantify the common variant component of a phenotype is with PGS[70,271]. PGS have transformed the study of complex traits by enabling the prediction of phenotypes and disease risk based on the aggregate contribution of common genetic variation[272–275]. However, given that there are other components to phenotypic variability, assessing how an individual's observed phenotype diverges from their PGS could refine signal for other components and improve power in association studies. Indeed, adjusting for PGS has been shown to improve power in rare variant association analyses[276].

There are other applications for this notion of divergence from the genetically expected phenotype measured by PGS. Assuming a liability threshold model, individuals diagnosed with disease but who have low common variant risk would be expected to have other components of risk causing them to cross the liability threshold. It has recently been shown that diagnosed individuals with low common variant risk are more likely to carry rare pathogenic variants[277] or have a monogenic diagnosis[278]. Similarly, a recent paper by Hawkes *et al.*[279], demonstrated that individuals whose observed phenotypes misaligned with PGS were more likely to carry rare loss-of-function (LoF) variants in genes linked with monogenic disorders. There are many other recent examples of work involving the interplay of polygenic and monogenic risk, especially demonstrating how common variant polygenic risk can modulate the penetrance or expressivity of monogenic risk variants. Examples of disorders for which this has been demonstrated include breast cancer[280–282], Huntington's disease[283,284], glaucoma[285,286], familial hypercholesterolaemia (FH)[280,287], obesity[287], developmental disorders[288,289], chronic kidney disease[290], and schizophrenia[291].

In this chapter I extend the work by Hawkes *et al.*[279], with a primary focus on testing for enrichment of monogenic risk variants among individuals who deviate from their genetically expected phenotype based on PGS. Using seven continuous phenotypes in the UKB and PGS calculated by Genomics plc[292] (provided through UKB), I expand the set of continuous phenotypes beyond standing height and LDL-C to include BMI, HbA1c, bone mineral density (BMD), intraocular pressure (IOP), and age at menopause (AAM). For each phenotype, I initially restrict my investigation to two, potentially overlapping, gene sets: (1) a small set of "canonical" monogenic genes which are classically defined as being associated with each trait (or monogenic genes previously defined by Hawkes *et al.*[279]), and (2) a larger set of genes which are considered "diagnostic grade" for a related rare disease in the Genomics England gene panel[183]. With more than double the number of sequenced exomes available to Hawkes *et al.*, I expect to have greater statistical power.

I also expand on the work of Lu *et al.*[277], who demonstrated that prevalence of rare pathogenic variants was higher among disease cases with low polygenic risk

score (polygenic risk scores (PRS) are the equivalent of PGS for disease liability) compared to those with high PRS, for multiple diseases in UKB. However, Lu *et al.* were only able to perform this analysis for 44,550 individuals, which limited the power of the study, especially given the sparsity of rare pathogenic variants in their data. For T2D, CAD, and osteoporosis (OP), I followed Lu *et al.* in testing whether disease cases with rare pathogenic variants tended to have lower PRS, but equipped with a dataset nearly ten times larger. For the first time, I tested whether disease controls with rare protective variants tended to have higher PRS. This is expected under the liability threshold model, as rare protective variants could balance out high liability from common variant risk (Figure S3.1).

Assuming an inverse relationship between allele frequency and variance in effect size, as is typically observed between rare and common variants[270], singleton variants (variants which are only observed once in a dataset) associated with disease would be expected to have comparatively larger absolute effect sizes than their common variant counterparts. Therefore, I expect carriers of high-effect singleton variants to be enriched for large deviations between observed phenotypes and that predicted by their PGS. To test this assertion, I evaluated the effect of exome-wide singleton burden on deviation of observed phenotypes from PGS for a suite of continuous and dichotomous (disease status) traits.

Finally, I explore whether this framework of testing for deleterious variant enrichment in misaligned individuals can be reappropriated for discovery of novel pathogenic or protective genetic effects. To do this, I treat misalignment as the trait of interest and perform gene-level association testing of protective and pathogenic damaging rare variation.

This work is important because it provides a clear pathway to the validation and identification of novel genetic contributors to complex traits and diseases, paving the way for better molecular diagnoses and targeted therapeutic discovery.

3.3 Main

3.3.1 Identifying individuals with misaligned continuous phenotypes

To classify individuals who deviate from their genetically expected phenotype I used a method developed by Hawkes *et al.*[279] (Methods). Briefly, individuals are classified as deviating from their genetically expected phenotype if their observed phenotype, adjusted for covariates, is a significant outlier in the bivariate distribution of phenotype and PGS and if the observed phenotype does not align with a linear model for observed phenotype and PGS. I applied IRNT to all continuous phenotypes after adjusting for covariates, to generalise the method to phenotypes which are not normally distributed. This is the same approach used by Hawkes *et al.* to maximise normality for LDL-C[279].

For standing height I classified 0.14% (559/400116) of individuals as misaligned, a similar percentage to Hawkes *et al.* (0.15% [244/158951]). Of the 559 individuals with deviating phenotypes, 212 and 347 individuals were relatively short or tall, respectively, compared to their genetically expected phenotype.

For LDL-C I classified 0.08% (309/381945) of individuals as misaligned, marginally lower than the percentage classified by Hawkes *et al.* (0.12% [159/134979]). This may be because Hawkes *et al.* account for statins, whereas I do not. Of the 309 individuals I classified as misaligned, 189 and 120 individuals had lower or higher than expected LDL-C based on their LDL-C PGS.

I applied the misaligned classification to the five remaining continuous traits, with the percentage of misaligned individuals ranging from 0.07% (HbA1c) to 0.12% (AAM) (Figure 3.1). For all phenotypes except height there were more individuals with lower than expected phenotype than individuals with higher than expected phenotype (Figure 3.1).

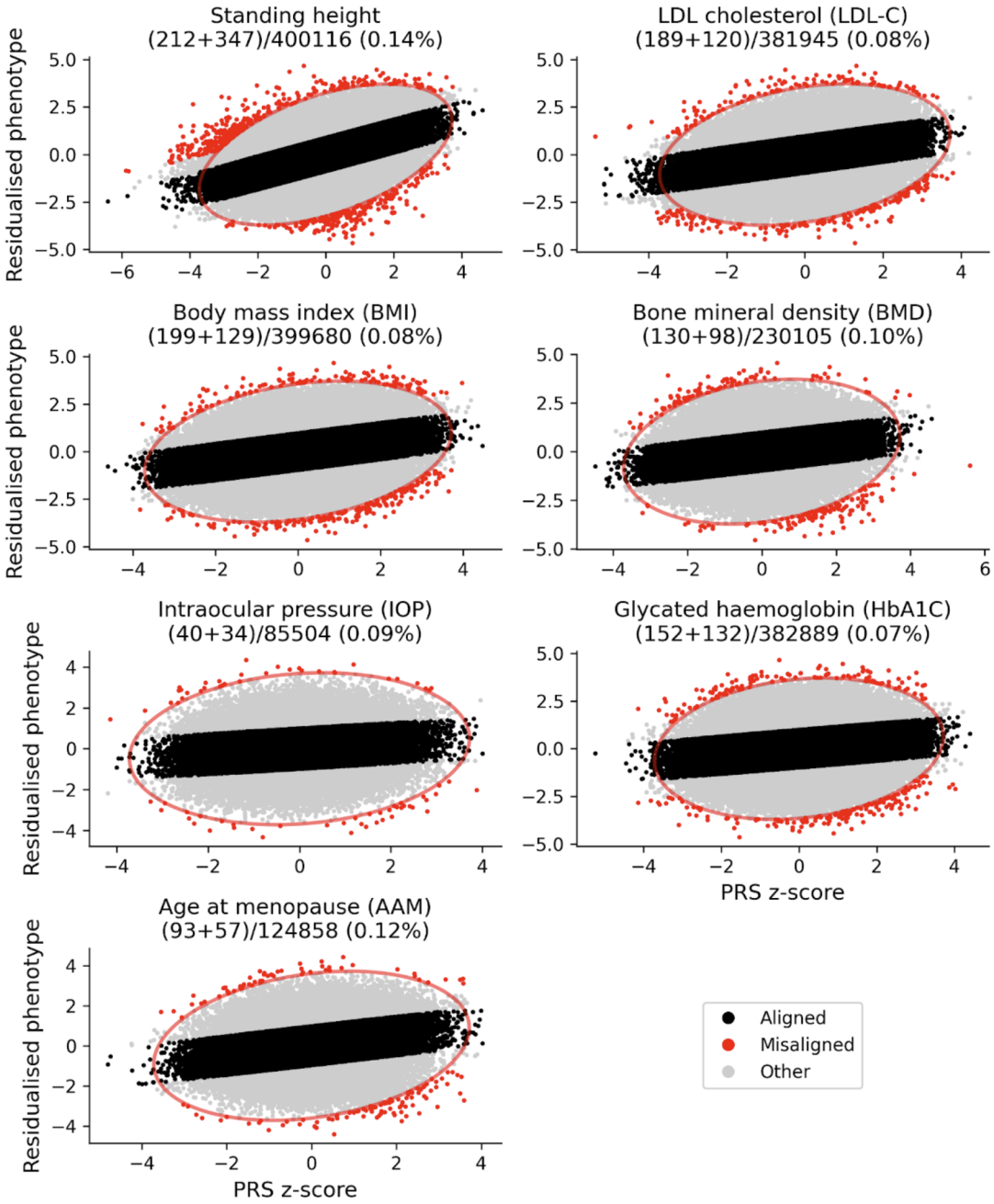

Figure 3.1: Classification of misaligned individuals for seven continuous traits. In the title of each plot is the name of the phenotype, the sum of low and high misaligned individuals divided by the total number of individuals with defined phenotype, and the total percentage of individuals who are classified as misaligned. The red oval delineates the Mahalanobis distance threshold.

3.3.2 Individuals with misaligned continuous phenotypes are enriched for predicted loss-of-function variants in monogenic disorder genes

Following categorisation into aligned and misaligned individuals, misaligned individuals can be further split into those whose residualised observed phenotype is higher than expected ("upper outliers") and lower than expected ("lower outliers"). In my enrichment tests, I test enrichment in one of the misaligned groups (either upper or lower outliers) against the aligned group, with the aligned group acting as a control group (see section 3.4.5 for enrichment test methods).

I tested misaligned individuals for enrichment of rare (MAF $< 0.1\%$) pLoF variants (see section 3.4.1 for annotation methods) in genes known to cause monogenic disorders relevant to the misaligned phenotype. The pairing of monogenic disorders with phenotypes was based on evidence in literature (Methods).

Establishing an empirical null for enrichment tests

In the extreme case, if a PGS explains no phenotypic variability, the slope of the phenotype-PGS regression will be zero. In this scenario, the expected value of phenotype-PGS residuals would be perfectly correlated with the covariate-residualised phenotype; individuals classified as misaligned would be outliers solely along the covariate-residualised phenotype axis, i.e. individuals with extreme phenotypes. Testing these extreme individuals for enrichment of pLoFs in genes known to be associated with the phenotype via monogenic disorders could result in spurious associations.

To be confident that the misalignment procedure is informed by PGS and that significant enrichment results are not false positives, I calculated an empirical P-value for each significant enrichment result by comparing the observed test statistic (Fisher's exact test P-value) against an empirical null distribution of test statistics. Spurious associations have the most statistical power when using truly random PGS. Therefore, the empirical null distribution was generated by simulating random PGS which explain no phenotypic variability, then running misalignment classification

and enrichment testing for each random PGS replicate (Methods). If the observed test statistic was not significantly higher than the null distribution generated by random PGS, then it was considered a spurious association.

Height and LDL-C

Using the sets of genes described in Hawkes *et al.*[279] within which rare damaging variation predisposes individuals to monogenic disorders of tall and short stature (Table 3.1), I observed significant enrichment (FDR-adjusted one-sided Fisher's exact test $P < 0.05$) of pLoF variants in individuals with higher than expected stature (OR [90% CI] = 24.1 [7.39, 78.9], $P_{adj} = 0.0153$) and individuals with lower than expected stature (OR [90% CI] = 87.9 [52.0, 149], $P_{adj} = 6.05 \times 10^{-17}$) (Figure 3.2; Table S3.2). I also tested single genes and observed significant enrichment in individuals with lower than expected stature for *ACAN* (OR [90% CI] = 406 [215, 767], $P_{adj} = 2.43 \times 10^{-19}$), *IGF1* (OR [90% CI] = 1300 [127, 13400], $P_{adj} = 8.12 \times 10^{-3}$), and *SHOX* (OR [90% CI] = 130 [23.1, 734], $P_{adj} = 0.0350$) (Figure 3.2; Table S3.2) which Hawkes *et al.*[279] were not powered to detect given their lower sample size ($n = 158,951$; approximately 40% of the 400,116 individuals available for my analysis of height).

Using the set of FH-related genes from Hawkes *et al.*[279] (Table 3.1), I found that individuals with high or low LDL-C were enriched for pLoF variants in genes where loss of function is known to cause high (*LDLR* OR [90% CI] = 207 [94.6, 453], $P_{adj} = 8.22 \times 10^{-10}$) or low LDL-C (*APOB* OR [90% CI] = 161 [115, 228], $P_{adj} = 6.57 \times 10^{-52}$; *PCSK9* OR [90% CI] = 52.4 [30.4, 90.2], $P_{adj} = 2.32 \times 10^{-13}$), respectively (Figure 3.2; Table S3.2). Individuals with lower than genetically expected LDL-C were also enriched for carrying the combined burden of *PCSK9* and *APOB* pLoF variants (OR [90% CI] = 113 [82.9, 153], $P_{adj} = 2.74 \times 10^{-59}$) (Figure 3.2; Table S3.2).

Of the nine significant trait-gene pairs tested (including single-gene and grouped-gene tests), four demonstrated significant empirical P-value results ($P_{emp} < 0.05$): *ACAN* ($P_{emp} = 3.00 \times 10^{-4}$), *IGF1* ($P_{emp} = 0.0150$) and grouped short stature genes ($P_{emp} = 2.00 \times 10^{-3}$) for height, and *PCSK9* ($P_{emp} = 7.99 \times 10^{-3}$) for LDL-C (Figure 3.2; Table S3.2).

Trait	Disorder	Effect of LoF on trait	Genes
Standing height	Growth disorder	Trait-decreasing (pathogenic)	*EZH2, FBN1, NSD1*
		Trait-increasing (pathogenic)	*ACAN, FGFR3, IGF1, IGF1R, NPR2, SHOX*
LDL cholesterol (LDL-C)	Familial hypercholesterolemia	Trait-decreasing (protective)	*APOB, PCSK9*
		Trait-increasing (pathogenic)	*LDLR*
Body mass index (BMI)	Monogenic obesity	Trait-increasing (pathogenic)	*LEP, LEPR, MC4R, PCSK1, POMC*
Intraocular pressure (IOP)	Primary open-angle glaucoma	Trait-increasing (pathogenic)	*MYOC*
Bone mineral density (BMD)	Osteogenesis imperfecta	Trait-decreasing (pathogenic)	*COL1A1, COL1A2*
Glycated haemoglobin (HbA1C)	Maturity-onset diabetes of the young (MODY)	Trait-increasing (pathogenic)	*GCK, HNF1A*
Age at menopause (AAM)	N/A	Trait-increasing (protective)	*CHEK2*

Table 3.1: Monogenic disorder genes for continuous phenotypes. See Methods for a description of how genes were selected. Genes for each disorder are listed in alphabetical order.

Additional continuous phenotypes

For other continuous phenotypes, no misaligned individuals carry pLoF variants in genes causing monogenic disorders (Table 3.1), likely due to the extremely low frequency of these variants in a population cohort such as UKB. In each trait-gene pair there was a relatively high probability (minimum of hypergeometric test $P(\text{carriers} = 0) = 0.80$; Methods; Table S3.2) of observing zero pLoF carriers in the set of misaligned individuals, given the number of carriers in the group of misaligned and aligned individuals tested.

Missense variants which are deleterious but not predicted to result in loss of function ("damaging missense" variants; Methods) are by definition expected to have less of an effect on gene product than pLoFs. Assuming a causal link between

gene product and phenotype, this implies that damaging missense variants are also expected to have attenuated effect sizes on the phenotype compared to pLoF variants. Although effect size is generally attenuated for damaging missense variants, they tend to occur in higher frequencies. For example, the aggregated pLoF allele count across genes tested for enrichment was 3,609; for damaging missense alleles the count was 20,959, almost six-fold higher. I hypothesised that testing damaging missense variants instead might boost power in enrichment testing by increasing the number of carriers of putatively damaging variation, albeit at smaller effect size, among misaligned individuals.

Individuals with misaligned continuous phenotypes are enriched for predicted damaging missense variants in monogenic disorder genes

I tested for enrichment of rare (MAF$<$ 0.1%) damaging missense variants in misaligned individuals with the same methods as for pLoF variants, for the same set of continuous traits and corresponding monogenic disorder genes. This is an extension beyond Hawkes *et al.*[279], who only considered pLoF variants.

I observed a significant enrichment (Fisher's exact test FDR-adjusted $P < 0.05$) of damaging missense variants among individuals with higher than genetically expected LDL-C in *LDLR* (OR [90% CI] = 61.0 [40.3, 92.4], $P_{adj} = 1.94 \times 10^{-25}$) (Figure 3.2; Table S3.2). Among individuals with lower than genetically expected LDL-C I observed a significant enrichment of damaging missense variants in *APOB* (OR [90% CI] = 13.9 [7.65, 25.3], $P_{adj} = 1.42 \times 10^{-6}$), *PCSK9* (OR [90% CI] = 20.2 [7.70, 53.0], $P_{adj} = 3.00 \times 10^{-3}$) and grouped *APOB* and *PCSK9* variants (OR [90% CI]=15.5 [9.28, 25.9], $P_{adj} = 3.15 \times 10^{-9}$) (Figure 3.2; Table S3.2). Each significant damaging missense enrichment result recapitulated pLoF enrichment results observed in the same single or grouped genes. For height, individuals with taller than expected stature were enriched for damaging missense variants in *FBN1* (OR [90% CI]=4.27 [2.17, 8.42], $P_{adj} = 0.0153$), a gene where pLoF enrichment was nominally significant (pLoF OR [90% CI] = 66.2 [11.92, 368], P=0.0163, $P_{adj} = 0.0590$). Of the significant damaging missense enrichment results

($P_{adj} < 0.05$), empirical P-values were significant for *PCSK9* ($P_{emp} = 0.0230$) and *FBN1* ($P_{emp} = 9.99 \times 10^{-3}$).

As expected, given that I am testing monogenic genes where damaging missense variants are predicted (Methods) to have a less deleterious effect than pLoF variants, the enrichment of damaging missense variants consistently had lower OR point estimates than pLoF variants (7/7 single- and grouped-gene tests with FDR $\leq 5\%$ significant pLoF and damaging missense enrichment, Figure 3.2).

3.3.3 Individuals whose disease status is misaligned with genetically expected risk are enriched for predicted loss-of-function variants in monogenic genes

Lu *et al.*[277] have shown in a set of common complex diseases in UKB that rare pathogenic variants occur at higher frequency in cases with low PRS, with low PRS defined by a percentile threshold. However, the authors only reported point estimates for variant frequency and did not appear to perform any statistical tests. Their work was performed on an earlier tranche of the exome-sequencing data with fewer than 50,000 individuals, so they may have been underpowered to make any statistical claims. Given that the number of sequenced exomes in UKB has increased more than ninefold[114] since the work by Lu *et al.*, this may provide the power increase necessary to observe an enrichment of rare pathogenic variants in low-PRS cases in UKB, if such an enrichment exists.

Note, in this section, I refer to PGS as PRS when higher scores are intended to positively correlate with disease case status. When referring to a continuous trait, I continue to use the term PGS.

I examined three common complex diseases: T2D, CAD, and OP. These common complex diseases have been included in previous UKB studies involving PRS and penetrance of pathogenic variants[277,280]. I tested the hypotheses that (1) disease cases with low PRS are enriched for pathogenic rare pLoF variants and (2) disease controls with high PRS are enriched for protective rare pLoF variants.

For each hypothesis, I took a two-step approach. First, within a given disease status group (cases or controls), I used linear regression to test the association of

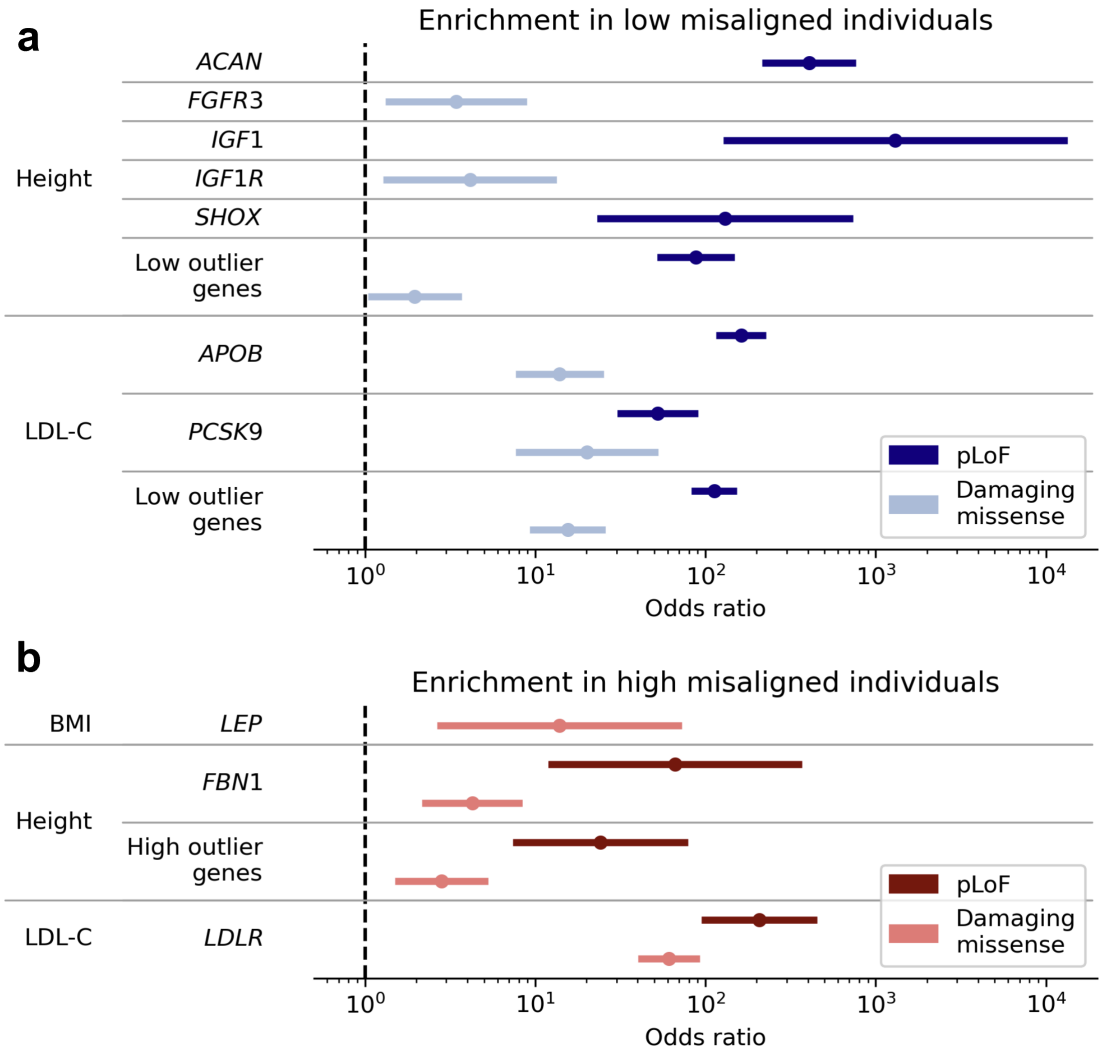

Figure 3.2: Enrichment of pLoF and damaging variants in canonical monogenic disorder genes. Enrichment in **a**, low and **b**, high misaligned individuals. "Low misaligned" individuals are individuals with phenotypes lower than genetically expected by PGS. "High misaligned" individuals are individuals with phenotypes higher than genetically expected by PGS. Error bars indicate two-sided 90% CI for OR, such that the lower bound indicates the lower bound for a one-sided 95% CI.

PRS with (pathogenic or protective) pLoF carrier status. Carrier status is a binary variable, defined as whether or not an individual carries at least one pLoF in the gene(s) tested. Second, inspired by the preliminary work of Lu *et al.*[277], I defined low and high PRS groups and tested enrichment of pLoFs between the groups. Specifically, I used Fisher's exact test to test for enrichment of predicted pathogenic variants in cases with low PRS versus cases with high PRS and for enrichment of predicted protective variants in controls with high PRS versus controls with low PRS. I used quantile thresholds to define low and high PRS groups, such that individuals below a certain threshold were designated as having low PRS and individuals above a threshold were designated as having high PRS. The quantile thresholds were defined on the cohort-wide PRS distribution. I used a quantile threshold at 0.5, following the work of Lu *et al.*[277] as well as two additional pairs of thresholds: the lower and upper quartiles (quantiles: [0.25, 0.75]) and the lower and upper deciles (quantiles: [0.1, 0.9]). The inclusion of additional thresholds was to observe whether the OR of enrichment increased when comparing more extreme pairs of quantile thresholds.

The first step is best powered to detect enrichment because it uses all samples available in the disease status group (unlike the high/low testing when quantile thresholds are not equal to 0.5) and is testing a continuous dependent variable (PRS), whereas the second step is only testing dichotomous variables (carrier status, high/low PRS status). The value of the second step is that it provides more granular insight into the trends in pLoF enrichment as I compare groups of ever more misaligned (low PRS cases, high PRS controls) and aligned individuals (high PRS cases, low PRS controls). To validate my approach, I applied it to binarised versions of height and LDL-C, continuous traits for which we have already observed an enrichment of rare damaging variants in individuals whose phenotypes deviate from genetic expectation. I replicated the strongest single- and grouped-gene pLoF associations for misaligned height and all single- and grouped-gene pLoF associations for misaligned LDL-C (subsection S3.1.1). This demonstrated that my pLoF enrichment testing approach for dichotomous case-control traits was

Disease	Associated disorder	Effect of LoF	Genes
Type 2 diabetes (T2D)	Maturity-onset diabetes of the young (MODY)[294]	Pathogenic	*GCK*, *HNF1A*, *HNF4A*, *HNF1B*
	n/a	Protective[295]	*SLC30A8*
Coronary artery disease (CAD)	Familial hypercholesterolaemia	Protective[296–298]	*APOB*, *PCSK9*
		Pathogenic[299]	*LDLR*
	Hereditary combined hypolipidemia[300]	Protective	*ANGPTL3*
Osteoporosis (OP)	Osteogenesis imperfecta[301]	Pathogenic	*COL1A1,COL1A2*

Table 3.2: Canonical monogenic genes tested for pLoF enrichment in disease traits.

sufficiently powered to detect the same pLoF enrichment in misaligned individuals observed for continuous traits, provided that the enrichment is strong enough.

Disease cases with low polygenic risk are enriched for rare pathogenic predicted loss-of-function variants in monogenic disorder genes

Among T2D, CAD, and OP cases I tested for association of rare (MAF<0.1%) pathogenic pLoF variants with lower PRS. I also tested for enrichment of pLoF variants in low-PRS cases versus high-PRS cases. As with the enrichment testing of misaligned individuals in continuous traits, I tested for pLoF enrichment in a curated set of canonical monogenic disorder genes (Table 3.2; Methods). I performed the Benjamini-Hochberg (BH)[293] procedure to control FDR at 5% separately for the regression association tests and Fisher's exact tests.

In T2D cases, pLoF carrier status in four single- or grouped-gene masks (*HNF1A*, *GCK*, *HNF4A*, and grouped maturity-onset diabetes of the young (MODY) genes) was significantly associated with lower PRS (FDR-adjusted one-sided *t*-test P<0.05) (Figure 3.3a; Table S3.3). The association with lowest P-value was observed for the grouped MODY genes (β [90% CI] = -0.734 [-0.943, -0.524], $P_{\text{adj}} = 6.01 \times 10^{-8}$).

When testing pLoF enrichment in low vs. high-PRS T2D cases for these single- or grouped-gene masks, the enrichment with the lowest P-value was always when testing bottom 25% vs. top 25% PRS. Within these enrichment tests, three gene masks were significant (FDR-adjusted one-sided Fisher's exact test P<0.05):

HNF1A (OR [90% CI]=21.1 [5.75, 77.6], P_{adj}=1.93 × 10^{-4}) and *GCK* (OR [90% CI] = 6.16 [2.47, 15.4], P_{adj} = 0.0110), and grouped MODY genes (OR [90% CI] = 9.00 [4.67, 17.4], P_{adj} = 1.89 × 10^{-6}) (Figure 3.3b; Table S3.3).

I observed a monotonic trend in increasing OR point estimates when comparing ever more extreme PRS quantiles (Figure 3.3b; Table S3.3). This suggests that enrichment of pLoFs increases as we compare increasingly misaligned-with-genetic-expectation individuals (low-PRS cases) to increasingly aligned-with-genetic-expectation individuals (high-PRS cases).

3.3.4 Disease controls with high polygenic risk are enriched for rare protective predicted loss-of-function variants in monogenic disorder genes

I tested for enrichment of rare pLoF variants in *APOB*, *PCSK9*, and *ANGPTL3* for CAD controls, and *SLC30A8* for T2D controls; these are genes where LoF is expected to be protective for the corresponding disease (Table 3.2). Although there were no FDR-controlled significant associations for protective pLoF carrier status with high PRS, three genes demonstrated nominal associations close to FDR-adjusted significance: *ANGPTL3* (β [90% CI] = 0.0742 [0.0118, 0.0137], P = 0.0253, P_{adj}=0.0602) and *APOB* (β [90% CI] = 0.0685 [0.0106, 0.126], P = 0.0258, P_{adj}=0.0602) for CAD controls and *SLC30A8* (β [90% CI] = 0.116 [0.0113, 0.221], P = 0.0342, P_{adj} = 0.0684) for T2D controls (Figure 3.4a; Table S3.3).

Comparing disease controls with high versus low PRS, I observed significant (FDR-adjusted one-sided Fisher's exact test $P < 0.05$) pLoF enrichment in *APOB* (top vs. bottom half of PRS OR [90% CI] = 1.28 [1.08, 1.51], P_{adj}=0.0373) (Figure 3.4b; Table S3.3). The protective pLoF enrichment in *SLC30A8* among T2D controls in the top versus bottom half of the PRS distribution achieved nominal significance, but was not significant after FDR control (OR [90% CI] = 1.34 [1.08, 1.66], P = 0.0152, P_{adj} = 0.0636).

Similar to the pathogenic pLoF testing in PRS tails, some gene masks (*ANGPTL3*, grouped CAD genes) exhibited monotonically increasing trends in OR point estimates, although all tests for these gene masks were non-significant, except for

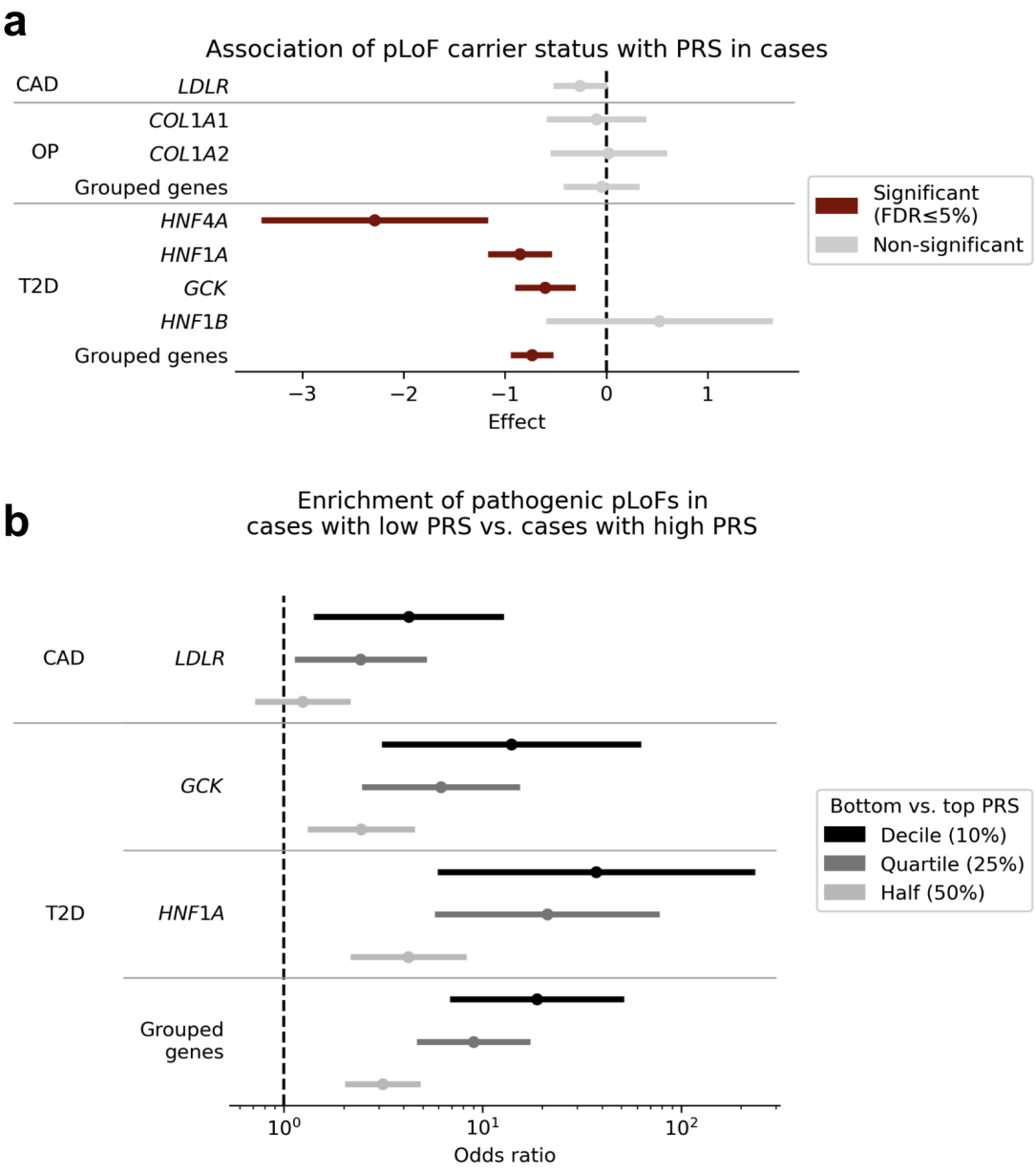

Figure 3.3: Pathogenic pLoFs are associated with lower PRS in disease cases. **a**, Effect of pLoF variant carrier status on PRS among disease cases. **b**, Comparing pathogenic pLoF enrichment in extremes of PRS for disease cases. Error bars indicate two-sided 90% CI for OR, such that the lower bound indicates the lower bound for a one-sided 95% CI. pLoF, predicted loss-of-function; PRS, polygenic risk score; CAD, coronary artery disease; OP, osteoporosis; T2D, type 2 diabetes.

one nominally significant enrichment for grouped CAD genes (top vs. bottom PRS quartiles OR [90% CI] = 1.12 [1.01, 1.25], $P = 0.0393$, $P_{adj} = 0.115$) (Figure 3.4b).

3.3.5 Individuals whose continuous phenotype misaligns with genetic expectation are enriched for predicted damaging variants in genes related to rare disorders

Having demonstrated that misaligned individuals are enriched for both pLoF and damaging missense variation in monogenic genes for continuous traits and pLoF variants in monogenic disorder genes for dichotomous traits, I expanded the set of genes analysed to include genes which cause rare disorders related to the traits analysed. To do this, I extracted genes from the expertly-reviewed GEL PanelApp gene sets[182,183] for rare diseases relevant to the phenotypes included in my analysis (Methods). I restricted analysis to genes which qualified for the highest-confidence category ("Green"), designated to be of "diagnostic-grade"[182,183].

I tested all continuous traits for enrichment of rare (MAF$<0.1\%$) pLoF and damaging missense variants in GEL rare disease genes in misaligned individuals. I used the same enrichment testing framework as described above for the canonical monogenic genes. I excluded 7.69% (38/494) of single- and grouped-gene masks because they had already been included in the enrichment testing of canonical monogenic disorder genes. I performed the BH procedure[293] to control FDR at 5% across all pLoF and damaging missense enrichment tests for continuous traits.

In dichotomous traits I tested for association of PRS with rare (MAF$<0.1\%$) pLoF carrier status in disease cases, to determine if cases with low PRS are carrying pathogenic pLoF variants in GEL genes. I also tested for rare (MAF$<0.1\%$) pLoF enrichment in low versus high PRS cases. I applied the BH procedure[293] to control FDR at 5% across all pLoF enrichment tests for dichotomous traits.

Significant enrichment of predicted loss-of-function variants in GEL rare disorder genes for individuals with misaligned continuous traits

There were three gene masks for which misaligned individuals were significantly enriched (Fisher's exact test FDR-adjusted $P<0.05$) for pLoF variants after FDR

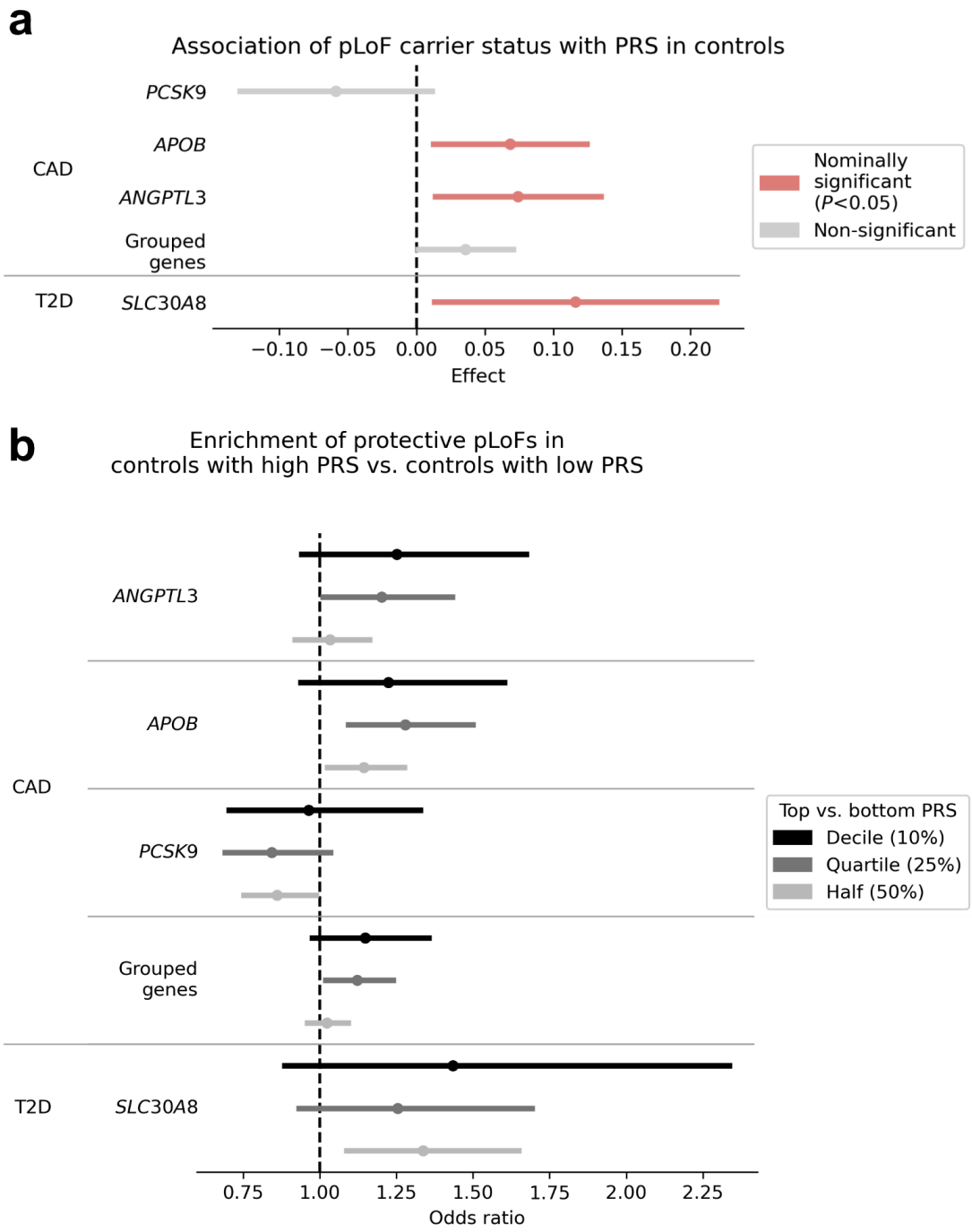

Figure 3.4: Protective pLoFs are associated with higher PRS in disease controls. **a**, Effect of pLoF variant carrier status on PRS among disease controls. **b**, Comparing pathogenic pLoF enrichment in extremes of PRS for disease controls. Error bars indicate two-sided 90% CI for OR, such that the lower bound indicates the lower bound for a one-sided 95% CI. pLoF, predicted loss-of-function; PRS, polygenic risk score; CAD, coronary artery disease; OP, osteoporosis; T2D, type 2 diabetes.

correction (Figure 3.5; Table S3.2). Individuals with higher than genetically expected LDL-C were enriched for pLoF variants in the grouped GEL PanelApp set of hypercholesterolaemia genes (OR [90% CI]=16.6 [8.74, 31.6], P_{adj}=9.36 × 10^{-5}). Individuals with lower than genetically expected BMD were enriched for pLoF variants in *GORAB* (OR [90% CI]=145 [42.0, 500], P_{adj}=0.0174), a gene linked with osteogenesis imperfecta in the GEL PanelApp. Individuals with lower than genetically expected height were enriched for pLoF variants in the GEL PanelApp set of monogenic short stature genes (OR [90% CI]=2.48 [1.70, 3.62], P_{adj} = 0.0354). Of these enrichment results, two had significant empirical P-values: hypercholesterolaemia genes (P_{emp} = 0.0360) and *GORAB* (P_{emp} = 1.00 × 10^{-4}).

There were eight other nominally significant (Fisher's exact test P<0.05) single- or grouped-gene pLoF enrichment results (Figure 3.5; Table S3.2). Four of these results were for BMD: *COPB2* (OR [90% CI]=1,220 [119, 12600], P=1.65 × 10^{-3}), grouped osteogenesis imperfecta genes (OR [90% CI] = 3.49 [1.84, 6.62], P = 5.24 × 10^{-3}), *SPARC* (OR [90% CI] = 204 [34.3, 1210], P = 5.75 × 10^{-3}), and *PPIB* (OR [90% CI] = 42.1 [7.86, 226], P = 0.0244). All but *SPARC* had significant empirical P-values (*COPB2* P_{emp} = 1.00 × 10^{-4}, grouped osteogenesis imperfecta genes P_{emp}=1.10 × 10^{-3}, *PPIB* P_{emp}=0.0482). However, the sparsity of pLoFs in these genes (each single-gene result was driven by only one pLoF carrier in the misaligned set of individuals) means that more simulation replicates may be needed to meaningfully explore the empirical null distribution of test statistics.

The remaining trait-gene pairs with nominally significant misalignment enrichment were LDL-C with *ABCG5* (OR [90% CI] = 16.7 [5.16, 54.3], P = 6.83 × 10^{-3}), AAM with *AARS2* (OR [90% CI] = 15.9 [4.86, 52.0], P=7.67 × 10^{-3}), BMI with *CPE* (OR [90% CI] = 134 [24.4, 732], P7.98 × 10^{-3}), and HbA1c with *ZNF808* (OR [90% CI] = 27.8 [5.27, 146], P = 0.0361). The enrichment empirical P-values were significant for *ABCG5* (P_{emp}=5.99 × 10^{-3}) and *AARS2* (P_{emp} = 0.0140).

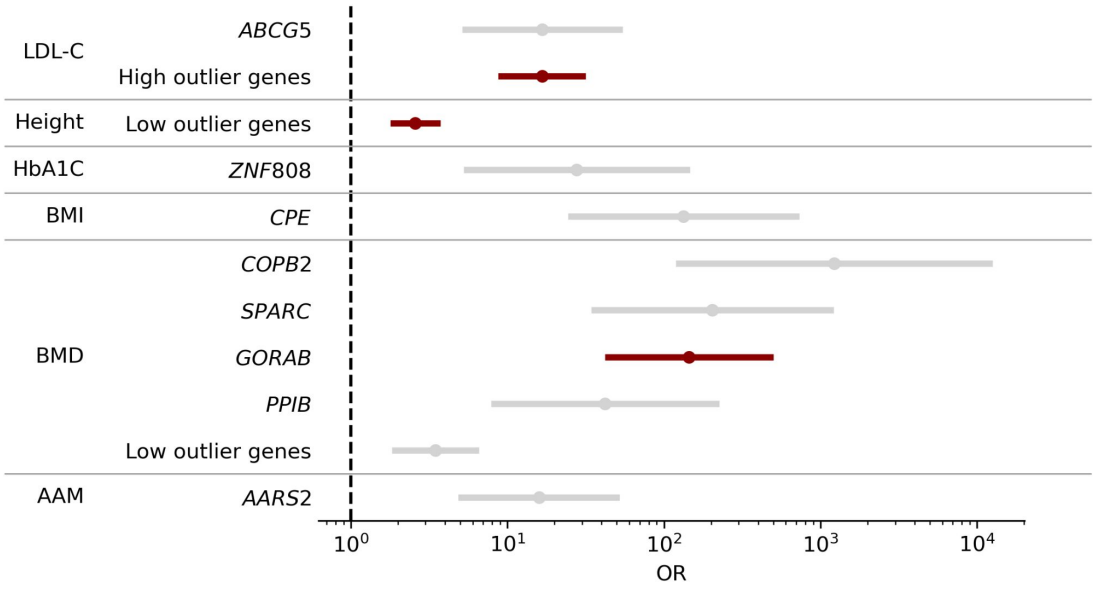

Figure 3.5: Misaligned individuals are enriched for pLoF variants in GEL PanelApp rare disease genes. Error bars show 90% CI intervals, such that the lower error bar is the lower bound for the 95% CI of the odds ratio from the one-sided Fisher's exact test. Red: significant after FDR-correction ($P_{\text{adj}} < 0.05$). Grey: nominally significant ($P < 0.05$).

Nominally significant enrichment of predicted damaging missense variants in GEL rare disorder genes for individuals with misaligned continuous traits

After FDR≤5% correction, there was no significant enrichment of damaging missense variants in GEL rare disorder genes for individuals with misaligned continuous traits. However, there were five nominally significant (Fisher's exact test $P < 0.05$) single- or grouped-gene enrichment results: three for misaligned height (*AKT2*, *IGF2*, grouped tall stature genes) and two for misaligned BMI (*BBS10*, *MKS1*) (Table S3.2). All had significant empirical P-values (Table S3.2). The lack of FDR-corrected significant results may be explained by the less deleterious effect damaging missense variants are predicted to have compared to pLoF variants. A large effect size is key to detecting enrichment in misaligned individuals because small effects are unlikely to push individuals into the misaligned class.

Limited evidence for enrichment of predicted loss-of-function variants in GEL genes for disease cases with misaligned disease status

There were no GEL genes where pLoF carrier status associated significantly (FDR-adjusted one-sided t-test $P < 0.05$) with lower PRS in cases (Table S3.4). There were two nominal associations: pLoF carriers in *LIPA* had lower CAD PRS (β [90% CI] = -0.479 [-0.930, -0.0288], P=0.0401, $P_{\text{adj}} = 0.945$) and carriers of grouped GEL T2D gene pLoFs had lower T2D PRS (β [90% CI]=-0.124 [-1.42×10^{-3}, 0.0373], $P = 0.0462$, $P_{\text{adj}} = 0.945$) among CAD and T2D cases, respectively.

I also tested enrichment of pLoFs in low PRS versus high PRS cases. There were no significant results (FDR-adjusted one-sided Fisher's exact test $P < 0.05$), however there were six nominally significant results (Table S3.4). Five of these were for T2D single- or grouped-gene tests, for a variety of low and high PRS thresholds: grouped GEL T2D genes (OR [90% CI] = 1.89 [1.35, 2.64], $P = 2.24 \times 10^{-3}$), *PLIN1* (OR [90% CI] = 4.66 [1.89, 11.51], $P = 0.0112$), *PPARG* (OR [90% CI] = 15.82 [2.37, 105.74], $P = 0.0143$), *TRMT10A* (OR [90% CI] = 1.84 [1.10, 3.08], $P = 0.0384$), and *ABCC8* (OR [90% CI] = 4.66 [1.45, 14.92], $P = 0.0492$). The other nominally significant enrichment was in *LIPA* (OR [90% CI]=3.41 [1.34, 8.72], $P = 0.0268$) for CAD cases in the bottom versus top half of the PRS distribution.

3.3.6 Individuals with genetically unexpected phenotypes are nominally enriched for singleton predicted loss-of-function variants

I tested whether the burden of singleton variants (alternate allele count (AAC) = 1) is associated with misaligned status. I aggregated singleton burden across all genes predicted to be intolerant to damaging variants, using the gene intolerance threshold (loss-of-function observed/expected upper bound fraction (LOEUF) < 0.6) proposed by Karczewski *et al.*[113]. I initially subset to intolerant genes because I was only interested in testing pLoF burden, where I had the strongest prior for singleton burden effect on phenotype. I expanded to test missense and synonymous variants, to test an additional hypothesis of whether there existed a gradient of

singleton burden effects from most deleterious to least deleterious variant categories. Subsetting to intolerant genes is expected to increase signal-to-noise ratio when testing variant effects, because variants in tolerant genes are not expected to affect the function of the genes, and by extension, not expected to affect the trait. This filtering is particularly important for exome-wide aggregated burden because there are many more variants being aggregated together than in the previous single- or grouped-gene tests.

I stratified the analysis across four protein-coding variant consequence categories: pLoF, damaging missense, other missense and synonymous (see Methods for variant consequence definitions), and tested both higher and lower than expected misaligned groups for enrichment relative to aligned individuals. I tested for burden by regressing misaligned status (misaligned vs. aligned) against the number of singletons carried by an individual for a given variant consequence, controlled for total singleton count across all protein-coding variants[302]. I calculated one-sided P-values for the singleton burden term, given my hypothesis that ultra-rare variants are more likely to cause individuals to be misaligned rather than aligned.

Nominal singleton enrichment in continuous traits

There were no significant (FDR-adjusted one-sided t-test $P<0.05$) enrichment results (Figure 3.6; Table S3.5) after FDR correction. Seven trait-consequence pairs were nominally significant (one-sided t-test $P < 0.05$) for singleton enrichment in misaligned individuals. However, three of these nominally significant enrichment results were for synonymous variants (lower than expected height $P = 0.0345$; higher than expected BMD $P = 0.0446$; higher than expected AAM $P = 0.0473$), which contradicts the expectation that synonymous variants have null effect. These synonymous singleton associations deserve further investigation, perhaps by investigating whether the singleton burden association signal is driven by a subset of individuals or genes.

Approaching the results qualitatively, one notable observation was the monotonic trend across functional categories in enrichment effect sizes for height (Figure 3.6),

a trait that showed strong enrichment when testing deleterious burden in monogenic and rare disorder genes. Individuals with lower than expected height were increasingly enriched for more deleterious singleton burden, while individuals with higher than expected height were increasingly depleted for more deleterious singleton burden. This suggests that, in general, deleterious singletons are more likely to decrease rather than increase height. This is consistent with the work of Backman *et al.*[114], who found that, among genes where pLoF or likely deleterious missense singleton burden was significantly associated with standing height, the burden effects in most (8/10) genes decreased height (see Backman *et al.*[114] Supplementary Data 2). This also aligns with a recent analysis of UKB whole-genome sequencing data by Hawkes *et al.*[119], who observed that among four genes where pLoF singletons were significantly associated with height, three genes were associated with decreased height (see Hawkes *et al.*[119] Supplementary Data 6).

Nominal association of missense singleton burden with misaligned disease status

I performed one-sided *t*-tests for association of singleton burden with PRS for CAD, OP, and T2D. In cases I tested whether higher singleton burden was associated with lower PRS, and in controls I tested whether higher singleton burden was associated with higher PRS. This was based on the same hypotheses used for the canonical and GEL tests: Association of variant burden with lower or higher PRS could indicate effects which are pathogenic or protective, respectively, against common-variant risk scores.

After FDR correction, there were no significant associations (FDR-adjusted one-sided *t*-test $P < 0.05$) (Figure 3.7; Table S3.6). Only two associations were nominally significant: damaging missense singletons were associated with higher PRS among OP controls (β [90% CI] = 5.36×10^{-3} [7.33×10^{-4}, 9.99×10^{-3}], one-sided *t*-test $P = 0.0284$), suggestive of a protective effect; synonymous singletons were associated with higher PRS among CAD controls (β [90% CI] = 4.07×10^{-3} [2.94×10^{-4}, 7.85×10^{-3}], one-sided *t*-test $P = 0.0381$). Notably, among cases for CAD and T2D the point estimates for effect of singleton burden on PRS monotonically decreases

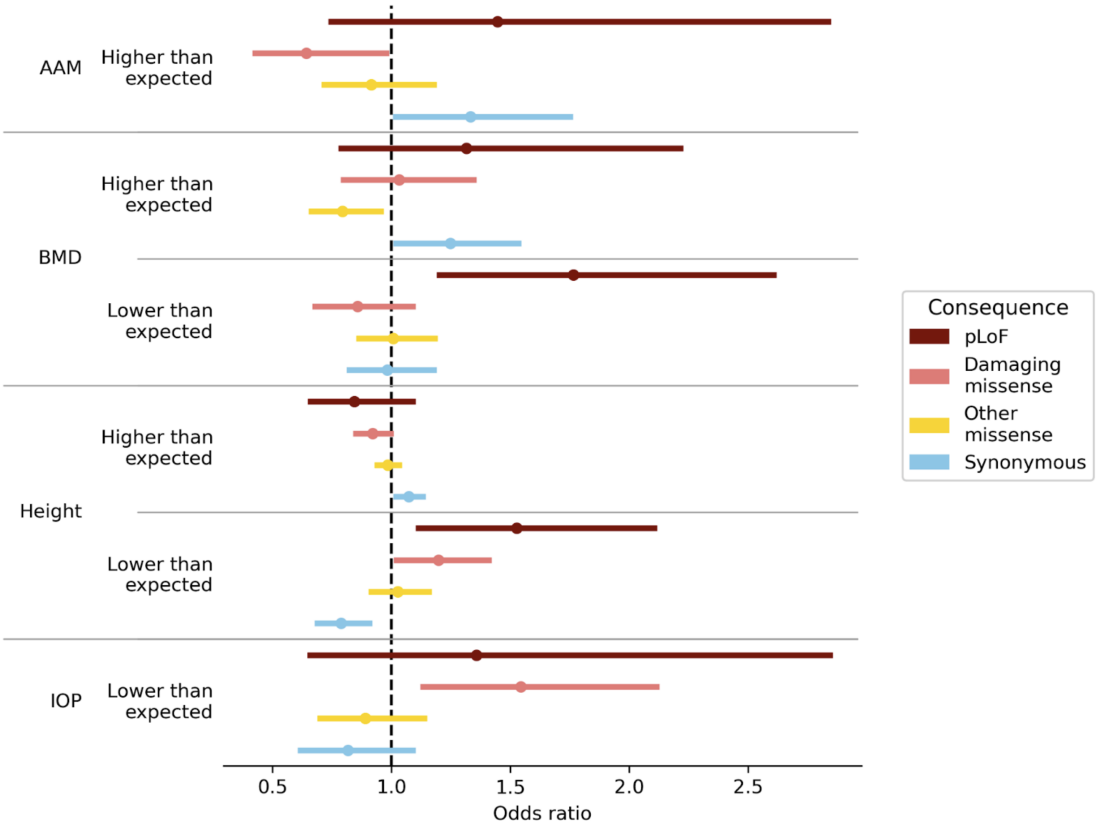

Figure 3.6: Misaligned individuals in continuous traits are nominally enriched for singleton burden in intolerant genes. Error bars are two-sided 90% CIs for OR, such that the lower error bar indicates the lower bound of the one-sided 95% CI of the OR from logistic regression of misaligned status against number of singletons, with the total number of singletons as a covariate (Methods).

with increasing consequence severity. This observation suggests that, for CAD and T2D, damaging singleton burden in intolerant genes may be pathogenic and cause disease cases to misalign with their PRS. However, given the overlap in CIs, more precise estimates of effect size may be needed to strengthen this claim.

3.3.7 Exome-wide scan for genes associated with misaligned status in continuous traits

Hypothesising that there may be genes associated with misaligned status that were not included in the canonical or GEL gene sets, I scanned all genes for missense variation associated with misaligned status. Genes where missense variation causes individuals to have lower than genetically expected disease risk, or lower than

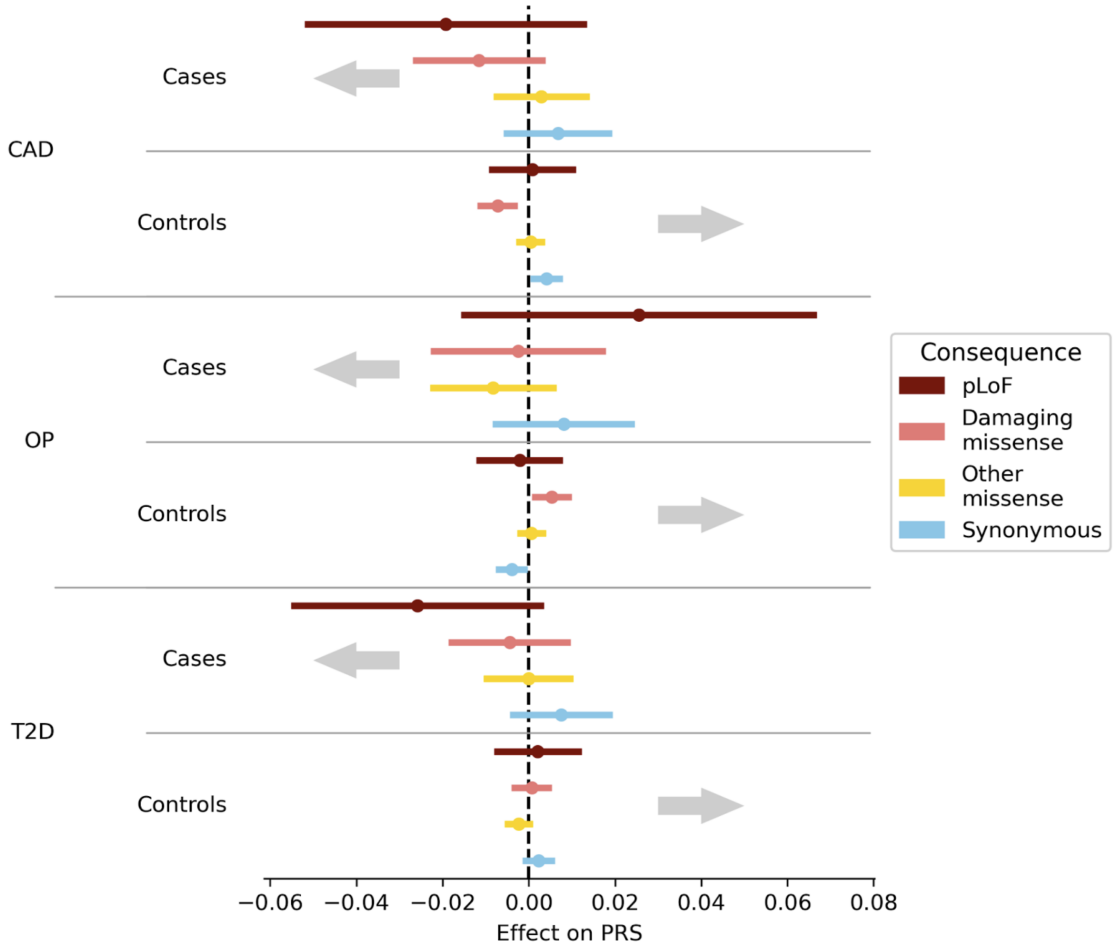

Figure 3.7: Misaligned individuals are nominally enriched in some dichotomous traits for singleton burden in intolerant genes. Error bars are two-sided 90% CIs for the effect of singleton burden on PRS, such that the upper bound (when testing cases) and lower bound (when testing controls) correspond to the upper and lower bounds, respectively, of the one-sided 95% CI. Grey arrows indicate the hypothesised direction of effect for the alternative hypothesis of the one-sided tests.

expected liability phenotype, would be ideal targets for therapeutic deactivation. Conversely, a gene where missense variation is pathogenic could be therapeutically targeted for increased expression, assuming druggability of the gene.

For misaligned status of each continuous trait, I performed gene burden testing of rare (MAF $< 0.1\%$) pLoF and damaging missense variants with SAIGE[103], testing separately for protective and pathogenic effects on phenotype misalignment (Methods). As a comparator, I also performed association tests on the original continuous traits.

Protective effects were defined as those which are trait decreasing, except for BMD and AAM where protective effects are trait increasing. Pathogenic effects are defined as trait-increasing, except for BMD and AAM where they are trait-decreasing. For height, where trait-increasing and trait-decreasing effects can both be pathogenic, I arbitrarily chose to represent trait-increasing effects as pathogenic and trait-decreasing effects as protective. For continuous phenotypes, these definitions are oriented by the diseases of interest for which the continuous phenotypes act as liability phenotypes. When testing misaligned status for protective or pathogenic effects, I always test for enrichment in the misaligned group corresponding to the effect on the original trait (lower than expected misaligned when effect is negative, higher than expected misaligned when effect is positive).

Protective effect enrichment tests

In the burden tests of protective effect, seven genes across four traits harboured significant enrichment (FDR-adjusted one-sided burden $P < 0.05$) of pLoF or damaging missense variants among individuals whose phenotype misalignment was protective (Figure 3.8; Table S3.7). Four of the significant burden associations have already been highlighted in previous sections: pLoF burden in *ACAN* (P_{adj}=4.20×10^{-13}) for individuals with shorter than genetically expected stature and, for individuals with lower than genetically expected LDL-C, pLoF burden in *PCSK9* (P_{adj}=4.14×10^{-10}) and burden in *APOB* (pLoF P_{adj}=4.46×10^{-47}; damaging missense P_{adj}=2.12×10^{-3}).

Individuals with lower than expected HbA1c were enriched for damaging variation in three genes: *EPB41* (pLoF $P_{adj} = 5.93 \times 10^{-6}$), *PIEZO1* (damaging missense $P_{adj} = 0.0111$), and *SLC4A1* (damaging missense $P_{adj} = 0.0158$; pLoF $P_{adj} = 0.0181$). Damaging variants in these genes are implicated in haemolysis[303–305], which can lead to low HbA1c[306].

Individuals with higher than expected AAM were enriched for damaging missense variants in *KANK1* (P_{adj}=0.0267). *KANK1* is a uterine fibroid risk gene[307], much like *CHEK2*, a tumour suppressor gene[308] where LoF also increases AAM[309]. Uterine

fibroids become less common after menopause[310], so if damaging variation in a gene was protective for primary ovarian insufficiency, increasing AAM, it might extend the window of time for uterine fibroids to appear, increasing the odds of developing uterine fibroids. When assessing the original trait, association of damaging missense burden in *KANK1* with AAM was null in our study (two-sided burden $P = 0.964$) and Genebass[199] (missense|low-confidence loss-of-function (LC) burden $P = 0.879$).

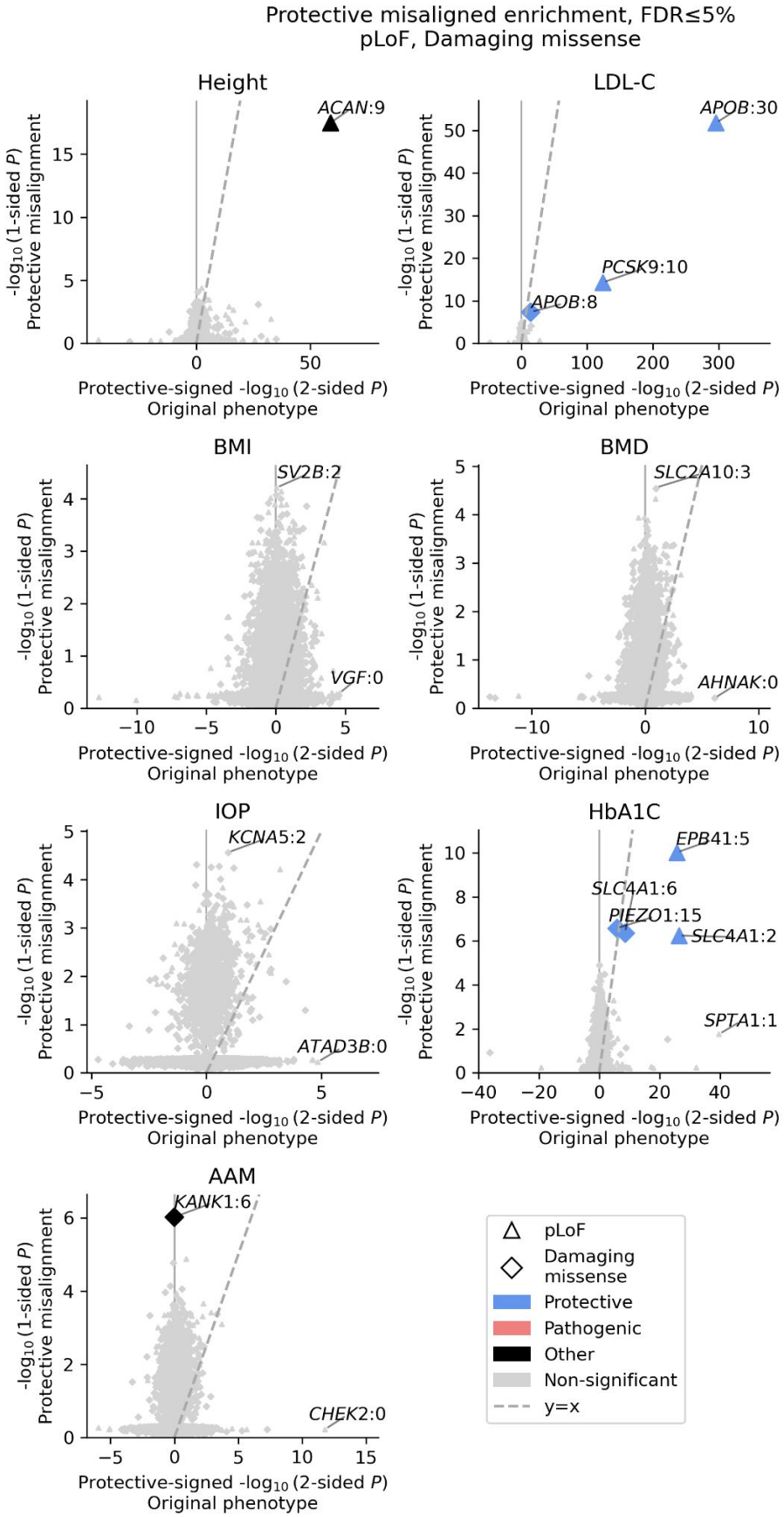
Protective misaligned enrichment, FDR≤5%
pLoF, Damaging missense
Height
ACAN:9
LDL-C
APOB:30
PCSK9:10
APOB:8
BMI
SV2B:2
VGF:0
BMD
SLC2A10:3
AHNAK:0
IOP
KCNA5:2
ATAD3B:0
HbA1C
EPB41:5
SLC4A1:6
PIEZO1:15
SLC4A1:2
SPTA1:1
AAM
KANK1:6
CHEK2:0
-log10 (1-sided P) Protective misalignment
Protective-signed -log10 (2-sided P) Original phenotype
pLoF
Damaging missense
Protective
Pathogenic
Other
Non-significant
y=x

Figure 3.8 ***(previous page)*****: Strength of protective burden association, comparing burden test on original phenotype versus burden test of protective misaligned status**. Genes are labelled (*gene*:MAC in misaligned individuals) if they are significant in the misaligned status burden enrichment tests after FDR correction (FDR-adjusted $P < 0.05$), have the strongest burden association for protective misaligned status, or have the strongest significant (FDR-adjusted $P < 0.05$) protective burden association in the original trait. Strength of protective burden association is measured by $-\log_{10} P_{\text{burden}}$, and is signed according to direction of effect: protective burden is positive, pathogenic burden is negative.

Pathogenic effect enrichment tests

In the burden tests of pathogenic effect, three genes across two traits, height and LDL-C, were enriched for pLoF or damaging missense burden in misaligned individuals (Figure 3.9; Table S3.8). *LDLR*, a gene highlighted in my analysis of genes known to cause monogenic disorders related to LDL-C, was significantly enriched (FDR-adjusted one-sided burden $P < 0.05$), for both pLoF (P_{adj}=8.68×10^{-7}) and damaging missense (P_{adj}=4.10×10^{-21}) variants among individuals with higher than genetically expected LDL-C (Figure 3.9; Table S3.8). Among individuals with taller than genetically expected stature, there was significant enrichment of damaging missense burden in *SLC25A39* (P_{adj}=9.52×10^{-3}) and *EIF3D* (P_{adj}=9.52×10^{-3}; note that the adjusted P-values are identical because the unadjusted P-values were similar). Both genes show no evidence of burden association with the original trait (*SLC25A39* two-sided burden $P = 0.265$; *EIF3D* $P = 0.366$). Further analysis of these genes may be required as there is no literature suggesting a plausible link between these two genes and tall stature.

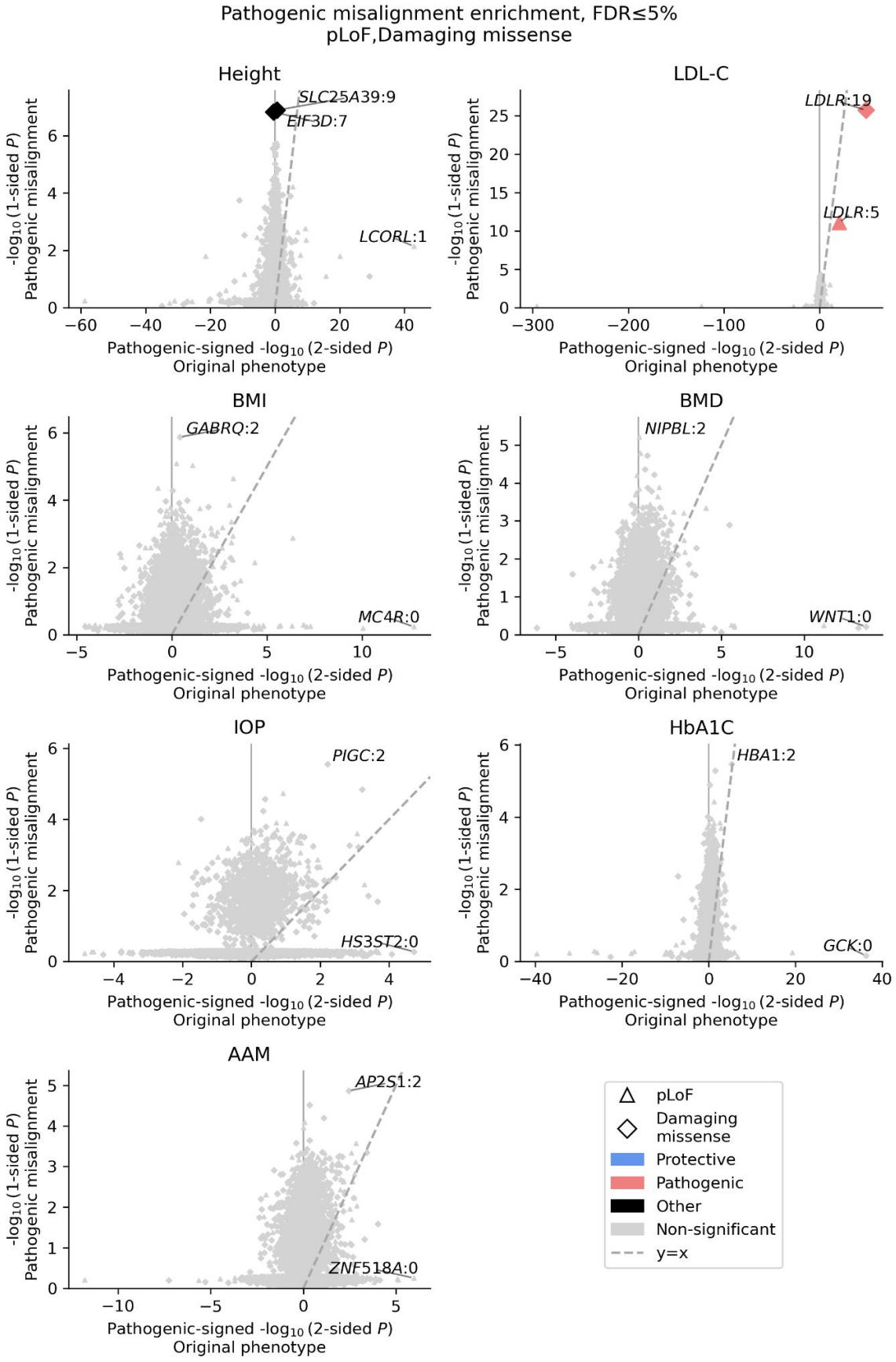
Pathogenic misalignment enrichment, FDR≤5%
pLoF,Damaging missense
Height
SLC25A39:9
EIF3D:7
LCORL:1
LDL-C
LDLR:19
LDLR:5
BMI
GABRQ:2
MC4R:0
BMD
NIPBL:2
WNT1:0
IOP
PIGC:2
HS3ST2:0
HbA1C
HBA1:2
GCK:0
AAM
AP2S1:2
ZNF518A:0
-log10(1-sided P) Pathogenic misalignment
Pathogenic-signed -log10(2-sided P) Original phenotype
pLoF
Damaging missense
Protective
Pathogenic
Other
Non-significant
y=x

Figure 3.9 *(previous page)*: Strength of pathogenic burden association, comparing burden test on original phenotype versus burden test of pathogenic misaligned status. Genes are labelled (*gene*:MAC in misaligned individuals) if they are significant in the misaligned status burden enrichment tests after FDR correction (FDR-adjusted $P < 0.05$), have the strongest burden association for pathogenic misaligned status, or have the strongest pathogenic burden association in the original trait. Strength of pathogenic burden association is measured by $-\log_{10} P_{\text{burden}}$, and is signed according to direction of effect: pathogenic burden is positive, protective burden is negative.

3.3.8 Exome-wide scan for predicted loss-of-function variant burden associated with protective misaligned disease status

Given the nominally significant association of protective pLoF burden with low PRS among disease controls, I hypothesised that I could discover novel protective genes by testing each gene for pLoF burden associated with higher PRS in controls. In other words, I am searching for pLoF burden which suppresses high common-variant polygenic risk.

Previous work has used PRS to scan genome-wide for protective genes. Hou *et al.*[311] performed a GWAS on Alzheimer's disease case-control status in individuals with high PRS (80th or 90th percentile) to identify genes which modulate disease status conditional on having high genetically predicted risk. Their approach tests disease status misalignment along the axis of case-control status: Controls with high PRS are the misaligned individuals and cases with high PRS are the aligned individuals. In my work, I also defined controls with high PRS to be misaligned, but my aligned individuals are controls with low PRS. I hypothesised that by testing a continuous variable (PRS), rather than a dichotomous variable (case-control status), I may have increased power to detect pLoF associations which suppress common-variant risk.

There has also been previous work searching for rare variants which suppress genetic risk. Chen *et al.*[312] scanned for genes which prevent penetrance of Mendelian disease variants. My work is distinct from Chen *et al.* because I scan for associations protective for aggregate common-variant risk, rather than Mendelian risk from highly penetrant rare variants.

To perform my genome-wide scan for protective pLoF burden, I subset to disease controls for T2D, CAD, and OP, and performed a SAIGE burden test on PRS as the phenotype. As a point of comparison, I also performed traditional burden tests on disease status.

Protective predicted loss-of-function burden

Among controls, there were no genes in which pLoF burden was significantly associated with increased PRS after FDR correction[293] (FDR-adjusted one-sided burden $P < 0.05$) (Figure 3.10; Table S3.9). The canonical protective genes were not among the top associations, suggesting that this method may not be well powered to detect genes similar to the canonical protective genes. To assess the biological plausibility of the three top nominally associated (one-sided burden $P < 0.05$) genes in each disease, I used the Open Targets platform[313] to gather evidence from literature, such as references of variants associated with increased or decreased expression of genes.

In T2D, pLoF burden in *ZFP36L2* was associated with higher PRS among disease controls (β_{PRS} [90% CI] = 0.0244 [0.0135, 0.0353], one-sided burden $P_{\text{PRS}} = 1.18 \times 10^{-4}$, $\text{OR}_{\text{disease}}$ [95% CI] = 1.04 [1.00, 1.08], two-sided burden $P_{\text{disease}} = 0.0503$). Intergenic variants inferred to decrease expression of *ZFP36L2* have been shown to be protective for T2D[75,314,315]. In *VWA7* (β_{PRS} [90% CI] = 6.75×10^{-3} [3.80×10^{-3}, 9.71×10^{-3}], one-sided burden P_{PRS}=8.53×10^{-5}, $\text{OR}_{\text{disease}}$ [95% CI] = 0.996 [0.985, 1.01], two-sided burden $P_{\text{disease}} = 0.475$), protein-truncating variants have previously been associated[200] with decreased fat body mass, a trait with high common-variant genetic correlation with T2D[316] and aetiological links with T2D[317]. In *SUN5* (β_{PRS} [90% CI] = 0.0154 [8.62×10^{-3}, 0.0223], one-sided burden $P_{\text{PRS}} = 9.80 \times 10^{-5}$, $\text{OR}_{\text{disease}}$ [95% CI] = 0.996 [0.970, 1.02], two-sided burden $P_{\text{disease}} = 0.784$), damaging variants have been linked with risk of male infertility[318], which has been suggested as a risk factor for T2D[319] – however, assuming that the effect direction of LoF on infertility carries over to T2D, LoF would not be protective for T2D.

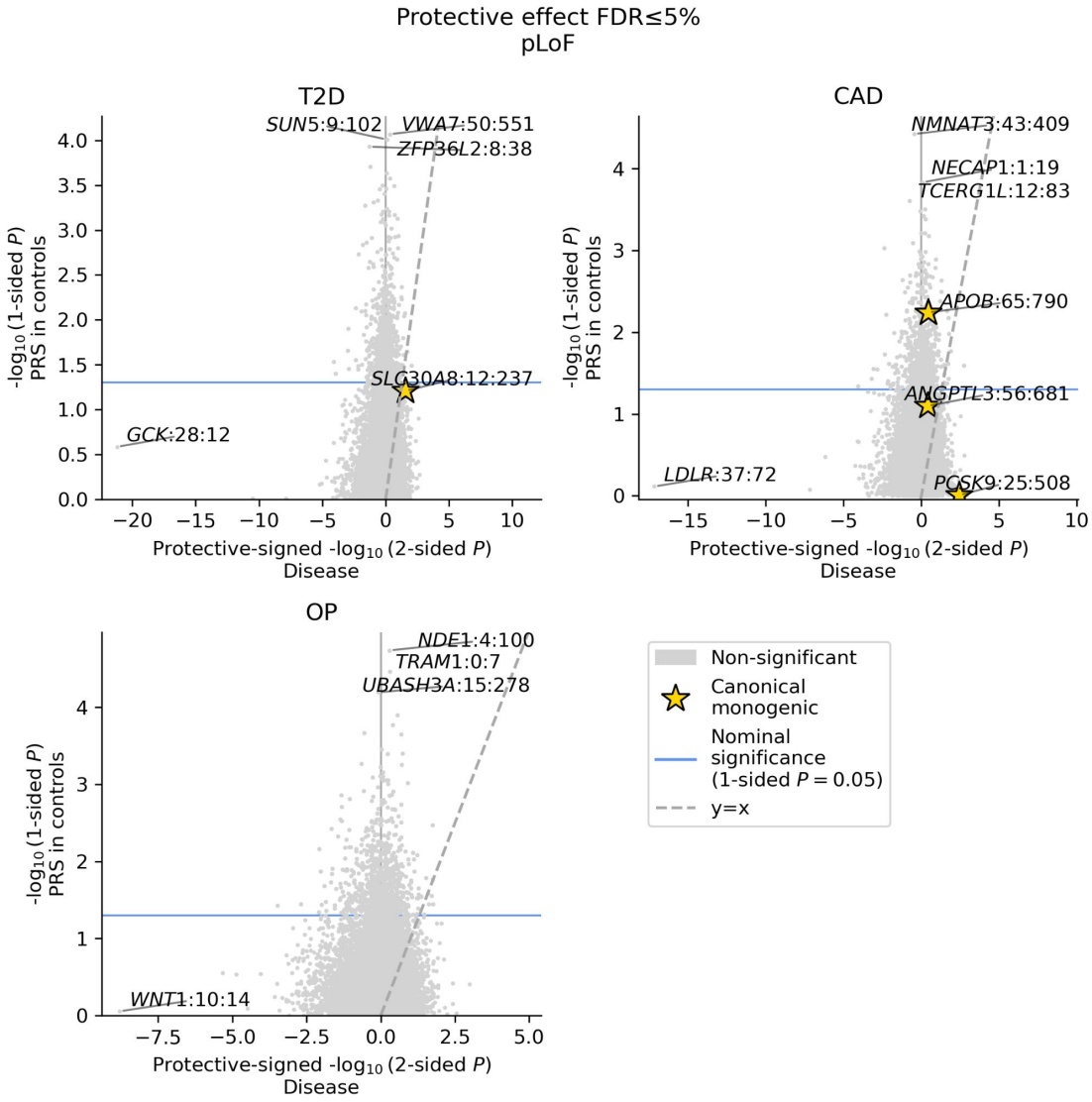

Figure 3.10: Gene-level rare pLoF burden association with PRS in disease controls. On the x-axis is the pLoF burden association with case-control disease status, quantified as $-\log_{10} P_{\text{disease}}$, where P_{disease} is the P-value for gene burden associated with disease status. This axis is signed using the negative sign of the burden effect on disease risk, such that positive values correspond to protective (lower odds of disease) effect. On the y-axis is the gene burden association with PRS in controls, where the burden P-values are calculated with the alternative hypothesis that gene burden has a positive effect on PRS. Only genes which have MAC ≥ 5 are shown. Top three genes with lowest one-sided P-value for the burden association test of PRS and the gene with lowest P_{disease} are labelled (*gene*:MAC in disease cases:MAC in disease controls). Gold star: canonical protective monogenic disorder genes. Light grey: Genes where FDR-adjusted one-sided $P_{\text{PRS}} > 0.05$, i.e. not significant after FDR correction. Blue line: Nominal significance threshold (one-sided burden $P_{\text{PRS}} = 0.05$). MAC, minor allele count.

Among top nominally protective genes for CAD, *NMNAT3* (β_{PRS} [90% CI]=8.38$\times 10^{-3}$ [4.90 $\times 10^{-3}$, 0.0119], one-sided burden $P_{\text{PRS}} = 3.78 \times 10^{-5}$, $\text{OR}_{\text{disease}}$ [95% CI] = 1.01 [0.993, 1.02], two-sided burden $P_{\text{disease}} = 0.343$) has been linked with attention deficit hyperactivity disorder (ADHD) via a nearby intergenic variant associated with decreased ADHD risk[320]. There is some evidence that ADHD is a risk factor for CVD, of which CAD is a subcategory[321]. The gene *TCERG1L* (β_{PRS} [90% CI] = 0.0164 [8.64×10^{-3}, 0.0241], one-sided burden $P_{\text{PRS}} = 2.49 \times 10^{-4}$, $\text{OR}_{\text{disease}}$ [95% CI] = 1.02 [0.991, 1.05], two-sided burden P_{disease}=0.171) has been linked with adiponectin[322], an inflammatory regulator proposed as a CAD and obesity biomarker[323]. Lastly, intergenic variants mapped to *NECAP1* (β_{PRS} [90% CI] = 0.0346 [0.0189, 0.0504], one-sided burden $P_{\text{PRS}} = 1.47 \times 10^{-4}$, $\text{OR}_{\text{disease}}$ [95% CI] = 0.992 [0.92, 1.07], two-sided burden P_{disease}=0.826) have been shown to associate with neutrophil count, a marker for CAD[324–326].

Among top nominally protective genes associated with higher OP PRS, *TRAM1* (β_{PRS} [90% CI] = 0.0601 [0.0352, 0.0849], one-sided burden P_{PRS}=3.42×10^{-5}, $\text{OR}_{\text{disease}}$ [95% CI] = 0.955 [0.838, 1.09], two-sided burden $P_{\text{disease}} = 0.491$) has previously been implicated with measures of bone fragility. Intergenic variants mapped to *TRAM1* which are inferred to increase gene product are also associated with decreased heel BMD[327] and heel broadband ultrasound attenuation (BUA)[328], an indicator of fracture risk[329,330]. This suggests that decreased gene product (e.g. caused by LoF variants) could be associated with increased BMD and BUA, reducing risk of OP. For *UBASH3A* (β_{PRS} [90% CI] = 9.76×10^{-3} [5.57×10^{-3}, 0.014], one-sided burden $P_{\text{PRS}} = 6.41 \times 10^{-5}$, $\text{OR}_{\text{disease}}$ [95% CI] = 1 [0.981, 1.02], two-sided burden P_{disease}=0.811), the disease with highest association score in Open Targets was type 1 diabetes (T1D). Intronic variants in *UBASH3A* inferred to increase gene product also associated with increased risk of T1D[331–333]. T1D is known to increase bone fragility[334], so LoF protective for T1D could theoretically decrease risk of bone fragility and OP. There was no evidence in Open Targets linking *NDE1* with decreased OP risk.

Note that although some of the top three associations per trait described above have relatively small effect sizes (e.g. *VWA7* burden association with higher PRS in T2D controls has an effect size point estimate of $\beta_{\text{PRS}} = 6.75 \times 10^{-3}$), the majority (5/9) had effect sizes expected to be sufficient for discovery (Figure S3.3).

Overall, there was modest evidence supporting the biological plausibility that genes highlighted by this association method have protective effects. Given that the canonical protective genes are not among the top genes associated with PRS, careful interpretation of genes more significantly associated with PRS than canonical genes may be necessary to avoid nominating false positives.

3.4 Methods

3.4.1 Exome sequencing data QC

QC and assigning population labels

I performed sample-, variant- and genotype-level QC for the UKB 450k ES dataset following the approach of Karczewski *et al.*[199]. Broad continental genetic ancestry labels (Africans, Admixed Americans, East Asians, Europeans, South Asians) were assigned by Duncan Palmer using a RF classifier trained on the 1000 Genomes dataset[66]. For this analysis, I subset to individuals assigned to the European super-population because this was the super-population with the largest sample size in UKB. For a complete description of sequencing data QC and population label assignment, see section 4.4.

Variant consequence annotation

All variant annotation was performed by Frederik Lassen and Barney Hill.

Coding variants were annotated as pLoF, "damaging missense", "other missense" and "synonymous". A detailed description of the variant annotation methods can be found in subsection 4.4.3.

3.4.2 Phenotype curation

Continuous phenotypes

I selected phenotypes from the set of phenotypes corresponding to the "Standard PRS" scores available in UKB[292]. Standing height (UKB field ID: 50) and LDL-C (UKB field ID: 30780), the two phenotypes analysed by Hawkes *et al.*[279], were chosen to benchmark my approach against the original Hawkes method.

The five other continuous traits (body mass index (BMI) [UKB field ID: 21001], estimated bone mineral density (BMD) [UKB field ID: 78], intraocular pressure (IOP) [UKB field IDs: 5262, 5254], glycated haemoglobin (HbA1c) [UKB field ID: 30750], and age at menopause (AAM) [UKB field ID: 3581]) were prioritised for inclusion based on sample size in UKB, heritability, how closely the trait available in UKB matched the trait used to create the PGS, and whether there was prior evidence of monogenic disorders or genes with strong effect contributing to trait variability (Table 3.1). IOP was calculated as the mean between left (UKB field ID: 5262) and right (UKB field ID: 5262) corneal-compensated measurements. I chose the corneal-compensated measurement method instead of the Goldman-correlated measurement method because it has been reported to be a better indicator for glaucomatous eyes[335]. For AAM, I excluded individuals who responded "Do not know" or "Prefer not to answer".

Dichotomous phenotypes

I included three dichotomous disease-status phenotypes: type 2 diabetes (T2D) (ICD-10 code: E11), coronary artery disease (CAD) (ICD-10 code: I25.1), and osteoporosis (OP) (ICD-10 codes: M80, M81). Case status was defined by ICD-10 diagnoses (UKB field ID: 41270). For each disease, individuals with the disease ICD-10 code(s) were defined as cases, while everyone lacking the disease ICD-10 code, but for whom UKB field ID 41270 was still defined, was assigned control status.

Polygenic scores

I use PGS calculated by Genomics plc[292] and provided through the UKB (data field IDs in Table S3.1). I chose to use the "Standard PRS" (UKB category 301) scores instead of the "Enhanced PRS" (UKB category 302) because the Standard PRS category of scores are trained solely using non-UKB data[292], making them less likely to be overfit to UKB, and because, as a result they are available for more than three times as many UKB samples compared to the Enhanced PRS.

3.4.3 Monogenic and rare disorder gene curation

Canonical monogenic gene curation

I defined "canonical" monogenic genes as genes which are well-known for causing the most common monogenic disorders associated with the phenotype being studied.

For standing height and LDL-C I chose the same monogenic genes as in Hawkes *et al.*[279], for consistency and to check for replicability of associations (Table 3.1). For BMI, I selected the first five monogenic obesity genes discovered[336]: *MC4R*, *PCSK1*, *LEP*, *POMC*, *LEPR*. For bone mineral density I chose the two genes (*COL1A1*, *COL1A2*) where LoF variation is the most common cause of monogenic OP[337]. For intraocular pressure I chose *MYOC*, in which mutations are the most common single-gene cause of primary open-angle glaucoma via increased intraocular pressure[338]. For glycated haemoglobin I chose the genes commonly responsible for MODY, the most common monogenic form of diabetes[339]. Within the subtypes of MODY, *GCK*-MODY (MODY2) and *HNF1A*-MODY (MODY3), caused by genetic defects in *GCK* and *HNF1A*, respectively, are the two most common subtypes, accounting for around 80% of MODY cases in the UK[340,341]. Patients in both subtypes can present with hyperglycemia[339,342], which can be indicated by higher than normal HbA1c levels[139]. For AAM there are no common monogenic disorders. As a proxy, I selected *CHEK2*, the gene with highest strength of association in the burden test for pLoF variants for AAM in Genebass[199] (pLoF burden P=3.75 × 10^{-32}). Early menopause can be caused by primary ovarian insufficiency (POI), a partly

monogenic disorder[343]. However, there are more than 100 putative monogenic POI genes and in UKB most POI cases are not caused by monogenic mechanisms[343].

For T2D I selected the four MODY genes responsible for the most common subtypes of MODY in the UK: *HNF1A*, *GCK*, *HNF4A*, *HNF1B*[340]. For CAD I selected the same FH genes (*PCSK9*, *APOB*, *LDLR*) used for LDL-C, because these monogenic risk genes have been shown to interact with CAD polygenic risk to modulate overall disease risk[280]. For OP I selected the same osteogenesis imperfecta (i.e. monogenic OP) genes used for BMD (*COL1A1*, *COL1A2*).

Rare disorder gene curation

To broaden the scan for enrichment to more genes and objectively define gene sets causing rare disorders, I used the evidence-based, expertly-curated and frequently reviewed GEL PanelApp resource[182,183].

When selecting rare disorder gene sets within the GEL PanelApp, I selected disorders that cause extreme, early-onset or monogenic forms of the phenotypes I include for analysis. I only included genes rated as "Green" (diagnostic-grade, conservatively defined) in the GEL PanelApp and excluded genes where the mode of pathogenicity was "Other" or started with "Loss-of-function variants (as defined in pop up message) DO NOT cause this phenotype". This resulted in a final list of diagnostic-grade genes for each selected GEL rare disorder, such that LoF variants in each gene were expected to cause the disorder.

3.4.4 Misaligned phenotype classification

Continuous traits

To determine individuals whose observed phenotypes are aligned or misaligned with their genetically predicted phenotype, we use a method developed by Hawkes *et al.*[279] for use with continuous phenotypes, such as standing height.

The outline of the method is as follows: First, covariates are residualised from the observed phenotype. Mahalanobis distance is calculated for each individual in the bivariate distribution formed by the residualised observed phenotype and

the genetically predicted phenotype. Mahalanobis distances are translated into P-values, using the property that Mahalanobis distance is expected to be distributed according to the chi-square distribution with k degrees of freedom, when calculated on a multivariate normal data with k variables[344]. The residualised observed phenotypes are regressed against the genetically predicted phenotypes and residuals are calculated for each individual, measuring the difference between the residualised observed phenotype and the expected phenotype given an individual's genetically predicted phenotype. Individuals are then classified as misaligned or aligned if they satisfied the following conditions, with thresholds derived from Hawkes *et al.*[279]:

Misaligned individuals

- Mahalanobis P-value < 0.001, and
- Absolute residual between residualised observed and expected phenotypes > 2

Aligned individuals

- Absolute residual between residualised observed and expected phenotypes < 1

In the first residualisation step, the covariates used were sex, age, age^2, $sex \times age$, $sex \times age^2$, the first 21 PCs (calculated on UKB genotype array data[74] and provided by UKB), and UKB assessment centre. Unlike Hawkes *et al.*[279], I did not account for statins when adjusting LDL-C for covariates. This simplifying omission is conservative, in that it is more likely to result in false negatives than false positives, as I expect cholesterol-lowering medication to attenuate the misalignment effects of deleterious variants associated with hypercholesterolaemia.

Case-control dichotomous traits

To contrast individuals with aligned versus misaligned disease status I set percentile thresholds on covariate-residualised PRS distributions within cases and controls separately. Percentile thresholds were chosen to compare the top versus bottom halves (threshold: 50%), quartiles (thresholds: 25%, 75%), and deciles (thresholds: 10%, 90%). This followed the approach of Lu *et al.*[277], who compared the top and bottom halves of the PRS distribution using point estimates for pathogenic variant

	Non-carrier	Carrier
Aligned	a	b
Upper/lower misaligned	c	d

prevalence. I included the more extreme thresholds (quartiles, deciles) to assess whether enrichment increased in the extremes of the distribution.

3.4.5 Enrichment testing of carrier status and misaligned status

Exact tests on contingency tables

To test whether individuals misaligned with genetic expectation are enriched for a binary variable (e.g. carrier status – whether or not an individual carries one or more rare [MAF$<$ 0.1%] damaging variants in a gene), I used Fisher's exact test[256], applied to a 2×2 contingency table, to calculate the probability P of observing the counts in the contingency table.

$$P = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!b!c!d!n!} \tag{3.1}$$

where n is the total number of observations $(a+b+c+d)$. This is the same test used by Hawkes *et al.*[279] to evaluate enrichment in misaligned individuals versus aligned individuals. I calculated ORs and 90% CIs[345] for OR with the following equations:

$$\text{OR} = \frac{a/b}{c/d}; \qquad z_q = \Phi^{-1}(q) \tag{3.2}$$

$$\text{OR}_q = e^{\ln(\text{OR}) + z_q \sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}} \tag{3.3}$$

$$\text{OR 90\% CI} = [\text{OR}_{q=0.05}, \text{OR}_{q=0.95}] \tag{3.4}$$

where OR_q is the OR CI limit corresponding to quantile q and Φ^{-1} is the inverse of the normal distribution cumulative distribution function (CDF).

Calculating empirical P-values with simulated polygenic scores

To establish an empirical null distribution for enrichment test statistics, I tested the misalignment enrichment method using simulated PGS. The simulated PGS

were generated independently of phenotypes, which means the expected phenotypic variance explained is zero. The PGS were simulated from a multivariate normal distribution parametrised to have mean zero and covariance structure based on the GRM, such that individuals who were estimated to share a proportion G of their genetic material, according to the GRM, shared the same proportion of variance explained (R^2=G) across simulated PGS replicates. The sparse GRM was calculated with SAIGE[103], using 5,000 randomly selected markers from LD-pruned UKB genotype array data. Only related pairs with relatedness greater than 0.05 were defined in the sparse matrix.

For each real (using non-simulated PGS) enrichment test in which there were carriers of deleterious variants, I estimated the empirical null by generating replicates of enrichment test statistics based on simulated PGS rather than real PGS (Figure S3.2). This involved running the misalignment classification procedure using simulated PGS instead of the real PGS, and then testing for enrichment of deleterious variants in the misaligned individuals. The empirical P-value was calculated by the following equation:

$$P_{\text{emp}} = \frac{r+1}{n+1} \tag{3.5}$$

where n is the number of test statistic replicates drawn from the null and r is the number of those replicate test statistics which are greater than or equal to the observed test statistic[346,347]. I used $-\log_{10}$(Fisher's exact test P) as my test statistic when calculating the empirical P-value. I did not exclude replicates where misaligned individuals had none of the variants I was testing (replicate Fisher's exact test P=1).

I used 1,000 replicates for most trait-gene-consequence combinations. In combinations with significant enrichment in misaligned individuals (FDR-adjusted one-sided Fisher's exact test $P < 0.05$) for which the observed test statistic was more significant than all 1,000 empirical null replicates (empirical $P = \frac{1}{1001}$=9.99×10^{-4}), I increased the replicates to 10,000. This was the case for two combinations: pLoF enrichment in *GORAB* for individuals with misaligned BMD and pLoF enrichment in *ACAN* for individuals with misaligned height. I did not test more than 10,000

replicates per combination due to the high computational and financial cost when running replicates on the DNAnexus RAP.

My simulation method differs from the simulation method employed by Hawkes *et al.*[279]. Instead of both simulating a phenotype and a PGS (from that simulated phenotype), I only focused on creating an empirical null for enrichment test statistics. I employed a simpler approach, only simulating PGS, that more directly answered whether PGS sufficiently explained phenotypic variance in order for the enrichment of rare, damaging variants in misaligned individuals to remain significant.

3.4.6 Hypergeometric test for probability of observing no carriers

When testing misaligned individuals for enrichment of rare (MAF $< 0.1\%$) pLoF variants in monogenic disorder genes, there were many instances where, due to the rarity of being misaligned compounded with the rarity of carrying rare pLoF variants, there were no misaligned individuals carrying rare pLoF variants. To quantify the high likelihood of observing no overlap in misaligned status and pLoF carrier status, I used the hypergeometric distribution to calculate the probability of observing no carriers (k=0) in a sample of n individuals (misaligned individuals) from a larger group of N individuals (aligned and misaligned individuals) with K carriers.

$$P(k=0;n,N,K) = \frac{\binom{K}{0}\binom{N-K}{n}}{\binom{N}{n}} = \frac{\binom{N-K}{n}}{\binom{N}{n}} \tag{3.6}$$

3.4.7 Association of carrier status with polygenic risk score for dichotomous traits

In case-control dichotomous traits I tested for the effect of pLoF carrier status (for pLoFs in pathogenic/protective monogenic or GEL rare disorder genes) on covariate-residualised PRS (PRS′) using linear regression, in cases and controls separately:

$$\text{PRS}'_{\text{cases}} \sim \beta_{\text{cases}} * \mathbf{1}_{\{\text{pathogenic}>0\}} + 1 \tag{3.7}$$

$$\text{PRS}'_{\text{controls}} \sim \beta_{\text{controls}} * \mathbf{1}_{\{\text{protective}>0\}} + 1 \tag{3.8}$$

where $\mathbf{1}_{\{\text{pathogenic}>0\}}$ and $\mathbf{1}_{\{\text{protective}>0\}}$ are indicators of whether an individual carries a pathogenic or protective pLoF, respectively. When evaluating the P-value associated with the effect, I used one-sided hypotheses. The alternative hypothesis for β_{cases} was that it was negative, because cases carrying pathogenic pLoFs were expected to have lower PRS than other cases. The alternative hypothesis for β_{controls} was that it was positive, because controls carrying protective pLoFs were expected to have higher PRS than other controls.

3.4.8 Singleton burden in intolerant genes

I defined intolerant genes using the gene intolerance threshold (LOEUF<0.6) proposed by Karczewski *et al.*[113], using gnomAD v4.1 constraint metrics downloaded from the gnomAD website. Of the 18,623 canonical transcripts available, 3,900 were defined as intolerant.

Singleton variants were defined as variants with an alternate allele count of 1 and alternate allele frequency $< 0.01\%$. I used PLINK[257] to aggregate singleton counts for each chromosome and each variant consequence category.

Following the approach of Genovese *et al.*[302], I tested for the effect of consequence-specific singleton burden (s_c, for consequence category c) using regression, controlling for overall singleton burden by including as a covariate the total singleton burden (s_{total}) across consequence categories. I used logistic regression when testing misaligned status in continuous traits, and linear regression when testing for effect of singleton burden on covariate-residualised PRS (PRS′) in cases and controls for dichotomous case-control traits:

Continuous traits		Misaligned status $\sim s_c + s_{\text{total}} + 1$
Dichotomous traits	Cases	$\text{PRS}'_{\text{cases}} \sim s_c + s_{\text{total}} + 1$
	Controls	$\text{PRS}'_{\text{controls}} \sim s_c + s_{\text{total}} + 1$

Table 3.3: Regression equations for testing singleton burden associated with continuous and dichotomous trait misalignment.

3.4.9 Exploratory burden testing on phenotype misalignment

I used SAIGE version 1.1.9[222] on the DNAnexus RAP to test gene-level rare variant (MAF $< 0.1\%$) burden association of continuous and dichotomous phenotype misalignment, using individuals with European ancestry in the ES 450k tranche. For a detailed description of the pipeline, see "Rare variant and gene-level testing in exome sequencing data" in Chapter 4 methods.

Continuous phenotypes

When performing burden tests for misaligned continuous phenotypes, I treated misaligned status as a dichotomous phenotype, with misaligned individuals as cases and aligned individuals as controls. I tested both lower misaligned (phenotype lower than genetically expected) versus aligned individuals and higher misaligned (phenotype higher than genetically expected) versus aligned individuals. I tested separately for the burden of pLoF variants and predicted damaging missense variants. The only covariate included when fitting the SAIGE null model was sequencing tranche, the tranche in which an individual was exome sequenced, because all other covariates had already been regressed out of the continuous phenotype before misaligned status classification.

To compare the power in testing misaligned individuals for gene enrichment relative to traditional approaches, I also ran gene burden tests using the original phenotypes used as input to the misaligned phenotype classification method. I applied IRNT within SAIGE to all continuous phenotypes using the `--invNormalize=TRUE` flag. Height failed to converge during the fitting of the null model in SAIGE, requiring the use of REGENIE[93] instead of SAIGE for height. Just as with all continuous traits analysed in SAIGE, I applied IRNT in REGENIE using the `--apply-rint` flag.

Case-control dichotomous phenotypes

When performing burden tests for misaligned disease status, I tested for association of pLoF gene burden with covariate-residualised PRS in disease controls. I did not test for association of damaging missense burden. I did not apply IRNT within SAIGE to PRS (`--invNormalize=FALSE`) because PRS have already been normalised. Similar to the analysis for continuous phenotypes, the only covariate included when fitting the SAIGE null model was *sequencing tranche*, because all other covariates had already been regressed out of the PRS.

When fitting the null model in SAIGE, estimation of variance components consistently failed for association tests on covariate-residualised PRS. This was the same issue described above when performing association tests on height. Given that this issue seemed to occur only in phenotypes with a large genetic component (height, PRS), I hypothesised that there may be issues caused by using highly related individuals, who would also have highly similar phenotypes. Using relatedness estimates (G) from the GRM created by SAIGE, I filtered to a maximal unrelated set of individuals. I created a greedy algorithm which iterated through a list of individuals (sorted by the number of relatives, with fewest relatives first) with at least one relative ($G > 2^{-3.5}$; corresponding to relatives 3rd degree or closer), adding their relatives to the list of individuals for removal. Individuals who had already been added to the list for removal were skipped if encountered later in the iteration. There were 114,184 related individuals in the GRM of Europeans who passed ES QC ($n = 402,375$). Using my greedy algorithm, I flagged 57,646 individuals for removal, resulting in a maximal unrelated set of 344,729 individuals. Using these unrelated individuals I was able to run SAIGE without encountering the previous error.

To compare the power in testing for misaligned disease status relative to traditional approaches, I also ran gene burden tests on disease status (case/control) for each disease included in this study. I did not filter to unrelated individuals because I did not encounter errors when fitting the null model in SAIGE.

3.5 Discussion

In this chapter, I demonstrated that the Hawkes *et al.*[279] misalignment classification framework can be generalised and extended to additional continuous traits. In particular, I demonstrated that even when PGS explains a relatively small amount of phenotypic variability, testing for enrichment of damaging rare variation in misaligned individuals is possible while controlling for false discoveries by comparing against an empirical null distribution. For dichotomous disease status traits, I built on work by Lu *et al.*[277] and developed a framework for testing enrichment of pathogenic and protective variants in disease cases and controls misaligned with their PRS.

Using misaligned classification in continuous traits, I replicated many of the pLoF enrichment results observed by Hawkes *et al.*[279] for misaligned height and LDL-C. I also observed that individuals with lower than genetically expected LDL-C and higher than genetically expected stature were enriched for rare damaging missense variants in *PCSK9* and *FBN1*, respectively. This suggests that phenotype misalignment may be influenced by variants expected to have more moderate effects on gene function. Another novel observation was that in BMD, a trait not investigated by Hawkes *et al.*, misaligned individuals were enriched for rare pLoF variants in *GORAB*, a gene linked with a monogenic form of OP. These results further validate the use of misalignment classification for prioritising individuals when screening of rare genetic disorders related to height, LDL-C, and BMD.

Using misaligned classification in dichotomous traits, I found that T2D disease cases carrying pathogenic rare pLoFs tended to have significantly lower PRS. For the first time, I observed that T2D and CAD disease controls carrying protective rare pLoFs tended to have higher PRS. As with the continuous traits, these results support the use of PRS when prioritising individuals for rare disorder screening. These results also provide strong evidence of a liability threshold model where disease status is determined by a balance of pathogenic and protective effects (Figure S3.1).

For the first time, I tested misaligned individuals for singleton burden across a spectrum of variant deleteriousness categories, observing monotonic trends in effect

size as a function of variant consequence that were suggestive of deleterious singleton burden effects on CAD and T2D disease status misalignment. Although evidence was modest, it could point to singleton burden contributions from binned by putative deleteriousness as an additional genetic contributor to phenotype misalignment.

Moving from validation to discovery, I tested exome-wide for genes associated with protective or pathogenic misaligned phenotype status and identified a gene, *KANK1*, where damaging variation may be protective for primary ovarian insufficiency. This demonstrated that misalignment classification could improve power in gene discovery, especially for self-reported phenotypes like AAM, which tend to be noisier than well-measured phenotypes[348].

There were some limitations in this study. First, testing for enrichment of rare damaging variants among misaligned individuals has limited statistical power due to the low or absent variant carriers among misaligned individuals. For example, when testing continuous traits, there were zero variant carriers in misaligned individuals for more than half (26/46) of the canonical monogenic gene enrichment tests and 79.5% (380/478) of GEL PanelApp gene enrichment tests, across all pLoF and damaging missense enrichment tests. This made it impossible to assess relative enrichment of damaging variants in misaligned individuals versus aligned individuals in these instances. In addition to the double jeopardy in statistical power when testing for enrichment of rare variants in a rare group of individuals, there are also expected to be fewer pathogenic variants in UKB because, on average, UKB participants are healthier than the general population[151], a limitation also cited by Hawkes *et al.*[279]. Second, there was inconsistency in sample size across phenotypes. While height, LDL-C, BMI and HbA1c have sample sizes greater than 380,000 (Table S3.1), AAM ($n = 124,858$) and IOP ($n = 85,504$) had lower sample sizes. This, in addition to differences in phenotypic variability explained by PGS, makes it difficult to compare the statistical power between phenotypes when testing enrichment of misaligned individuals. Third, using IRNT on covariate-residualised phenotypes to enable the extension of the Hawkes *et al.* framework to traits with non-normal distributions may result in false-positive or false-negative enrichment

results. The empirical null simulations are expected to temper the false positive rate, while an elevated false negative rate may be the price to pay for using IRNT. Fourth, I used variant consequence categories based on computational prediction (see Methods) rather than strictly clinical evidence. Using computational prediction may lead to some miscategorised variants, but enabled me to scale the enrichment tests to many more variants than would be feasible through manual curation of single variants informed by clinical case studies. Fifth, unlike Hawkes *et al.*, I did not restrict my analysis to unrelated individuals. This was a choice intended to maximise sample size. The potential downside of this choice would be inflated type I error rates caused by related individuals being classified as misaligned, while also being carriers of the variants being tested for enrichment. Future iterations of this work may benefit from restricting to unrelated samples. Finally, this study was performed solely in individuals classified as having European genetic ancestry (Methods). As noted by Hawkes *et al.*[279], whose study suffers the same limitation, identifying misaligned individuals in non-European populations is difficult, as these populations are generally less represented in GWAS with sufficient sample size to generate informative PGS. However, as increasingly diverse large biobanks are created, such as the AoU dataset[115], misalignment classification will become feasible for more ancestry groups.

There are several areas of opportunity to build off of this work. As Hawkes *et al.* did for height, I could investigate whether non-genetic factors, such as socioeconomic status, diet, or medication, could explain phenotype misalignment. For instance, adjusting for statin use would likely have resulted in more individuals with extreme LDL-C, assuming that individuals with high pre-medication LDL-C are more likely to be prescribed statins. Given that individuals with FH are more likely to have extreme LDL-C and thus prescribed statins to bring their levels down to normal, correcting for statins could also increase the number of FH variant carriers among the individuals classified as misaligned. This would boost the signal for enrichment, demonstrating how correcting for such non-genetic factors could improve power for detecting rare genetic effects on misalignment. Although my work has replicated

results from Hawkes *et al.*, I used a superset of the individuals included in their study, so the datasets are not independent. Replication of results in an independent cohort, such as the AoU dataset[115], would further increase confidence in results. Stratifying analysis by sex and investigating sex-specific enrichment, as Hawkes *et al.* did for LDL-C, could provide more granularity to the results. As observed for pLoF enrichment tests in individuals with extreme biomarker levels (see "Individuals with extreme biomarker levels are enriched for pLoF variants in genes which cause rare disorders"), there may be sex-differential results. However, stratification would increase the multiple testing burden and also decrease the sample size, making it less likely to have fully non-zero contingency tables. Performing a variant-level exome-wide scan for association could highlight the specific variants driving gene-level associations. However, this higher-resolution approach may only be practical for rare but not ultrarare variants, as statistical power to detect very rare variant associations (e.g. MAC $\leq$ 10) is attenuated compared to associations in more common variants with the same effect size[349]. Lastly, when performing exome-wide scans for novel protective genetic effects associated with PRS in disease controls, I only tested pLoF burden. This was because I expected pLoFs to have a much stronger effect than other missense variants. However, as observed when testing gene burden associated with misaligned status in continuous traits, there may be value in expanding to damaging missense variants as well.

In this chapter I demonstrated that for a range of heritable continuous traits and complex diseases, misaligned individuals are enriched for rare damaging variants. This work highlights the value in considering both common and rare genetic components of trait variability, as it may prioritise certain individuals for rare disease diagnosis and could be leveraged to find new pathogenic or protective associations.

4

Combining evidence from human genetic and functional screens to identify pathways altering obesity and fat distribution

Contents

4.1 Abstract 118
4.2 Introduction 119
4.3 Main 120
4.3.1 Discovery of genes associated with obesity and fat distribution through exome-wide association tests 120
4.3.2 Convergence of rare and common variant evidence for genes associated with obesity and fat distribution 124
4.3.3 Sex- and age-specific effects of genes associated with obesity and fat distribution 128
4.3.4 Nominating target genes for *in vitro* functional characterisation through clustered regularly interspaced short palindromic repeats (CRISPR) knockdown 129
4.3.5 Knockdown of obesity and fat distribution associated genes alters *in vitro* adipogenesis and lipid accumulation 133
4.3.6 Transcriptome-wide effects of gene knockdowns 134
4.4 Methods 138
4.4.1 Imputed data QC 138
4.4.2 Exome sequencing data QC 139
4.4.3 Variant consequence annotation 141
4.4.4 Phenotype curation 142
4.4.5 Genetic association testing 142
4.4.6 Common variant gene prioritisation scores 146
4.4.7 AoU gene-level summary statistics 147

4.4.8 Genebass gene-level summary statistics 147
4.4.9 Genetic correlation estimates 147
4.4.10 Selecting target genes 148
4.4.11 Wet lab methods 149
4.4.12 Differential expression analysis of CRISPR adipocyte RNA sequencing 152
4.4.13 Statistical analysis of lipid accumulation assays 152
4.4.14 Pathway enrichment in CRISPR RNA sequencing data . 153
4.4.15 Druggability evidence 153
4.5 Discussion 154

4.1 Abstract

Overall adiposity and body fat distribution are heritable traits associated with altered risk of cardiometabolic disease and mortality. Performing rare variant (minor allele frequency<1%) burden association testing using exome sequencing data from 402,375 participants in the UK Biobank for nine overall and tissue-specific fat distribution traits, we identified 19 genes where putatively damaging rare variation associated with at least one trait (Bonferroni-adjusted $P<1.58\times10^{-7}$) and 50 additional genes at FDR≤1% ($P \leq 4.37\times10^{-5}$). These 69 genes exhibited significantly higher (one-sided t-test $P = 3.58\times10^{-18}$) common variant prioritisation scores than genes not significantly enriched for rare putatively damaging variation, with evidence of monotonic allelic series (dose-response relationships) among ultra-rare variants (MAC $\leq$ 10) in 22 genes. Combining rare and common variation evidence, allelic series and longitudinal analysis, we selected 14 genes for CRISPR knockdown in human white adipose tissue cell lines. In three previously uncharacterised target genes, knockdown increased (two-sided t-test $P<0.05$) lipid accumulation, a cellular phenotype relevant for fat mass traits, compared to Cas9-empty negative controls: *COL5A3* (fold change (FC)=1.72, P=0.0028), *EXOC7* (FC=1.35, $P = 0.0096$), and *TRIP10* (FC=1.39, P=0.0157); furthermore, knockdown of *PPARG* (FC=0.25, $P = 5.52\times10^{-7}$) and *SLTM* (FC=0.51, $P = 1.91\times10^{-4}$) resulted in reduced lipid accumulation. Integrating across population-

based genetic and *in vitro* functional evidence, we highlight therapeutic avenues for altering obesity and body fat distribution by modulating lipid accumulation.

4.2 Introduction

One in four adults globally are either overweight or obese[350,351]. While higher overall adiposity increases morbidity and mortality,[350,352] disease risk is also informed by the location and distribution of excess fat within particular depots[353,354]. Abnormal distribution of fat is often attributed to lipodystrophy syndromes, which can cause generalised or selective fat mass loss and depot-specific fat growth[355]. Independent of overall BMI, individuals with higher central adiposity have increased risk of cardiometabolic diseases, including T2D and stroke[356,357]; in contrast, individuals with higher hip circumference, an indicator for gluteal adiposity, have lower risk of such outcomes. For example, a SD increase in hip circumference has been shown to reduce risk of T2D by $\sim 40\%$[358] or myocardial infarction by $\sim 10\%$[359]. Previous studies indicate that fat distribution, as assessed by waist-to-hip ratio (WHR), has a strong heritable component independent of BMI, with narrow sense heritability of up to 56% in women and 32% in men[356,360].

BMI-associated genes are enriched in tissues of the central nervous system, notably the hypothalamus which is involved in appetite regulation[361]. Indeed, blockbuster GLP1-receptor agonists that are prescribed for weight loss[362,363] act primarily through the dorsomedial hypothalamus to control food intake[364]. In contrast, GWASs for WHR adjusted for BMI (WHRadjBMI) indicate enrichment of genes associated with fat distribution in adipose tissue[365]. However, there are currently few therapeutic avenues to modify both obesity and fat distribution[366].

Understanding the genetic aetiology of body fat and fat distribution may drive new therapies for obesity. Mendelian genetic studies have identified rare variants in genes such as *PPARG*, a master regulator of adipocyte differentiation, and *INSR*, the insulin receptor, associated with lipodystrophies and extreme forms of central body fat distribution[367,368]. In the general population, rare protein-coding variants with large effects can point to genetic and molecular mechanisms underpinning fat

distribution. For example, previous work has demonstrated that rare LoF alleles in *GPR75* are protective against obesity[369], and protein-truncating variants in *INHBE* are associated with favourable fat distribution[370].

Indeed, while there have been several common and rare-variant association studies of obesity[365,369,371,372], mechanistic studies of the pathways by which the identified genes affect adipogenesis have typically been limited to characterising just one or two targets[373–381]. The difficulty of manipulating human adipocytes in high-throughput experimental assays has prevented genome-wide functional studies of the genetic drivers of adipogenesis.

Here, we integrated ES and common variant GWASs for nine obesity-related and fat distribution traits to nominate genes associated with overall obesity or central adiposity in up to 402,375 participants in UKB. We then designed an *in vitro* functional assay for lipid accumulation in human white adipocytes to biologically validate a considerable selection of the identified genes using CRISPR-Cas9 knockdown experiments. Taken together, we demonstrate that converging multi-modal evidence from GWAS, ES, and *in vitro* knockdown studies in a relevant context can generate putative therapeutic targets for obesity or lipodystrophy.

4.3 Main

4.3.1 Discovery of genes associated with obesity and fat distribution through exome-wide association tests

We performed gene-level discovery testing across 1,827,504 rare variants (MAF $< 1\%$) in 18,788 genes for association with three obesity or fat distribution related traits, i.e. overall obesity (BMI and body fat percentage) and central fat distribution (WHRadjBMI) in up to 402,375 participants of European ancestry in the UKB (Table S4.1, Table S4.2). We additionally assessed the effects of these genes on tissue-specific fat component phenotypes derived from dual-energy X-ray absorptiometry (DXA) and magnetic resonance imaging (MRI) scans (such as android tissue fat percentage, abdominal fat ratio, and visceral adipose tissue volume), in up to 39,671 participants of European ancestry in the UKB (Table S4.2). We defined

damaging rare variants as those annotated to be high-confidence pLoF or damaging missense using a combination of existing effect prediction tools (LOFTEE[113], CADD[110], REVEL[109], and SpliceAI[111]; Methods). The reported gene-level effect sizes were estimated by burden testing, while P-values were estimated using the optimal sequence kernel association test (SKAT-O) for improved power (Methods). Finally, to account for potential confounding due to nearby common variants, we conditioned gene-level associations on fine-mapped GWAS loci on the same chromosome (Methods).

We identified 19 unique genes carrying rare damaging variation associated with BMI, body fat percentage, or WHRadjBMI at exome-wide significance (SKAT-O $P < 1.58 \times 10^{-7}$, Bonferroni adjustment for 315,996 unique tests) (Table 4.1 and Figure 4.1a). Four genes associated with body fat percentage are also exome-wide significant for BMI with the same direction of effect (*PTPRG* and *GPR151*) or WHRadjBMI with opposing directions of effect (*PDE3B* and *PLIN4*) (Figure 4.2, Figure S4.2). *PLIN4* is also negatively associated with gynoid fat percentage (SKAT-O $P = 2.16 \times 10^{-6}$), suggesting that the association of *PLIN4* with lower body fat percentage may be driven by reduced gynoid fat deposits, which in turn results in a positive association with WHRadjBMI (an anthropometric proxy for the ratio of android to gynoid fat accumulation). We see a qualitatively similar effect of burden of rare variation in *PDE3B*, which is associated with higher overall obesity as measured by body fat percentage, but reduced WHRadjBMI (exome-wide significant), android to gynoid ratio, abdominal fat ratio, and visceral adipose tissue volume (all $P < 0.05$), indicating that the increased body fat percentage in these individuals may be driven by higher subcutaneous rather than visceral fat accumulation (Figure 4.2, Figure S4.2).

While the 19 exome-wide significant genes were not significantly associated (all $P > 4.38 \times 10^{-4}$, Bonferroni adjustment for 19 unique genes $\times$ 6 phenotypes) with any other tissue-specific fat components, we observed that nominally significant associations ($P < 0.05$) generally shared the expected direction of effect on the primary and supplemental phenotypes (Figure 4.2, Figure S4.2). Genes that are

positively associated with BMI (*MC4R*) and body fat percentage (*RIF1*) are also positively associated with android fat percentage, gynoid fat percentage, android/gynoid ratio, and total tissue fat percentage; and *PPARG* is negatively associated with body fat percentage as well as gynoid and total tissue fat percentage (Figure 4.2, Figure S4.2). Similarly, *COL5A3* is associated with increased WHRadjBMI and higher android to gynoid ratio, while *PLIN1* is associated with lower WHRadjBMI and higher gynoid fat percentage as expected.

Finally, we found evidence of monotonic allelic series, that is, increasingly large effects of increasingly damaging variants in a dose-response relationship, among ultra-rare variants (MAC $\leq$ 10) in *COL5A3, DIDO1, INSR, PLIN1, PTPRG, PPARG* and *SLC12A5* (Figure 4.1b).

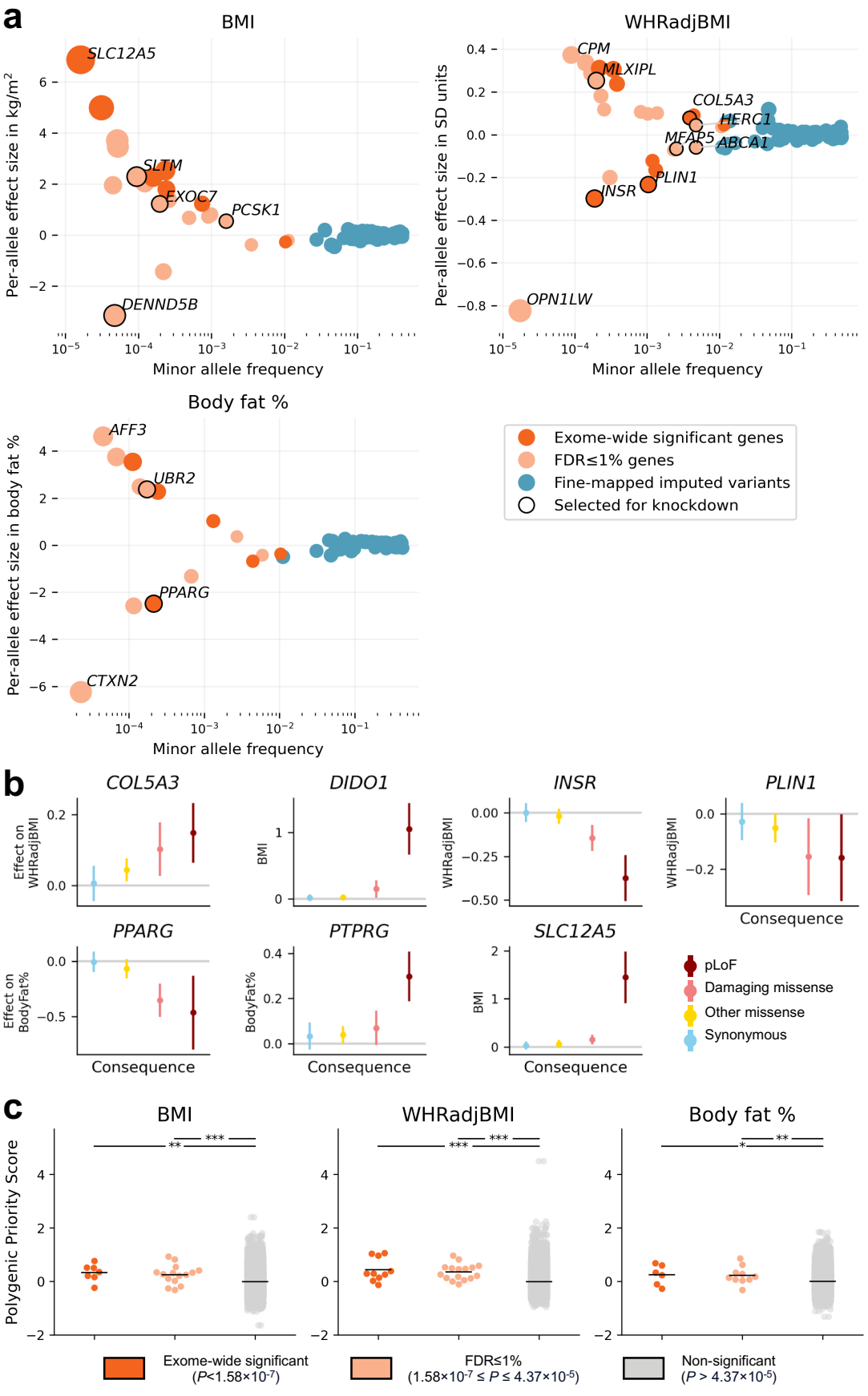
a
BMI
WHRadjBMI
Body fat %
Per-allele effect size in kg/m²
Per-allele effect size in SD units
Per-allele effect size in body fat %
Minor allele frequency
SLC12A5
SLTM
EXOC7
PCSK1
DENND5B
CPM
MLXIPL
COL5A3
HERC1
MFAP5
ABCA1
PLIN1
INSR
OPN1LW
AFF3
UBR2
PPARG
CTXN2
Exome-wide significant genes
FDR≤1% genes
Fine-mapped imputed variants
Selected for knockdown
b
COL5A3
DIDO1
INSR
PLIN1
PPARG
PTPRG
SLC12A5
Effect on WHRadjBMI
BMI
WHRadjBMI
Effect on BodyFat%
BodyFat%
Consequence
pLoF
Damaging missense
Other missense
Synonymous
c
BMI
WHRadjBMI
Body fat %
Polygenic Priority Score
Exome-wide significant (P<1.58×10⁻⁷)
FDR≤1% (1.58×10⁻⁷ ≤ P ≤ 4.37×10⁻⁵)
Non-significant (P > 4.37×10⁻⁵)

Figure 4.1 ***(previous page)*****: Consequences of rare missense variation on obesity or fat distribution related traits in UKB.** **a**, Gene- and variant-level effects as a function of aggregated minor allele frequency. Only gene-level results which are exome-wide or FDR≤ 1% significant are shown. Only fine-mapped common (MAF > 1%) variants with posterior inclusion probability ≥ 0.9 are shown. Effect sizes for BMI and body fat percentage are converted to the kg/m^2 and fat percentage scale, respectively, by multiplying the effect size in SD units by the empirical SD of each trait (4.75 kg/m^2 for BMI, 8.51% for body fat percentage). Genes selected for knockdown are outlined in black and labelled. Genes with highest and lowest effect size for each trait are also labelled. Point sizes are scaled by effect size. **b**, Allelic series for exome-wide significant genes with monotonic relationships between effect size and consequence severity in ultra-rare (MAC≤10) variant burden tests. 95% CIs for effect size were defined as ±1.96 standard errors (SEs). For allelic series of all FDR≤ 1% genes, see Figure S4.3. **c**, Enrichment of Polygenic Priority Score (PoPS)[382] among exome-wide and FDR ≤ 1% significant genes compared to other genes. Significance of one-sided t-test with alternate hypothesis that a significant gene category has higher average PoPS than the non-significant gene category: $^{***}P < 0.001$, $^{**}P < 0.01$, $^{*}P < 0.05$. Results of t-test for BMI PoPS: exome-wide significant gene PoPS vs. non-significant gene PoPS t-statistic (P) = 2.79 (2.65×10^{-3}), FDR ≤ 1% vs. non-significant = 3.17 (7.64×10^{-4}). Results of t-test for WHRadjBMI PoPS: exome-wide significant vs. non-significant = 4.96 (3.60×10^{-7}), FDR ≤ 1% vs. non-significant = 5.19 (1.06×10^{-7}). Results of t-test for body fat percentage PoPS: exome-wide significant vs. non-significant = 2.04 (2.07×10^{-2}), FDR≤ 1% vs. non-significant = 2.46 (7.02×10^{-3}). The mean PoPS for each gene category are indicated by black bars. For BMI: Exome-wide significant genes PoPS mean [95% CI from ±1.96 standard error of the mean (SEM)] (number of genes) = 0.324 [0.086, 0.562] (7), FDR ≤ 1% = 0.250 [0.070, 0.430] (15), non-significant = -0.009 [-0.012, -0.006] (34613). For WHRadjBMI: Exome-wide significant=0.430 [0.161, 0.698] (10), FDR≤ 1% = 0.353 [0.207, 0.500] (16), non-significant = -0.014 [-0.017, -0.011] (34609). For body fat percentage: Exome-wide significant = 0.242 [-0.060, 0.544] (6), FDR≤ 1% = 0.226 [0.023, 0.429] (10), non-significant = -0.006 [-0.009, -0.003] (34619).

4.3.2 Convergence of rare and common variant evidence for genes associated with obesity and fat distribution

Genes associated with obesity and fat distribution in rare variant gene-level tests overlapped with those highlighted by common variant associations, demonstrated by a significant enrichment of Polygenic Priority Score (PoPS)[382] (one-sided t-test $P < 0.05$, Figure 4.1c) among genes harbouring exome-wide significant damaging rare variant association signals. Moreover, 50 rare-variant gene-level associations which do not reach exome-wide significance but pass a less stringent significance threshold (SKAT-O $P \leq 4.37 \times 10^{-5}$, using the BH procedure[293] for FDR≤ 1% on minimum SKAT-O P-value across traits and two variant masks; Methods)

Gene (genomic location)	Phenotype	Consequence	Variants in mask	Cumulative MAF (MAC)	SKAT-O *P*-value	Burden *P*-value	Effect size [95% CI]	Effect size, unweighted
ANKRD12 (18:9136228)	WHRadjBMI	pLoF	127	0.00038 (308)	1.71×10^{-9}	6.67×10^{-8}	0.010 [0.006, 0.013]	0.238
APBA1 (9:69427532)	BMI	pLoF	59	0.00016 (128)	6.25×10^{-8}	5.18×10^{-8}	0.019 [0.012, 0.026]	0.475
BLTP1 (4:122152331)	BMI	pLoF	282	0.00075 (600)	2.15×10^{-10}	9.19×10^{-11}	0.011 [0.007, 0.014]	0.258
COL5A3 (19:9959561)	WHRadjBMI	pLoF plus damaging missense	362	0.00388 (3,111)	5.23×10^{-10}	2.98×10^{-9}	0.003 [0.002, 0.004]	0.078
DIDO1 (20:62877738)	BMI	pLoF	22	0.00003 (25)	1.14×10^{-7}	1.14×10^{-7}	0.042 [0.026, 0.057]	1.050
GPR151 (5:146513144)	Body fat %	pLoF	37	0.01039 (8,209)	6.60×10^{-8}	2.51×10^{-7}	-0.002 [-0.003, -0.001]	-0.043
	BMI	pLoF	37	0.01036 (8,313)	1.35×10^{-7}	2.02×10^{-7}	-0.003 [-0.004, -0.002]	-0.058
INHBE (12:57452323)	WHRadjBMI	pLoF	23	0.00119 (954)	3.27×10^{-9}	1.37×10^{-7}	-0.005 [-0.007, -0.003]	-0.123
INSR (19:7112255)	WHRadjBMI	pLoF	55	0.00019 (151)	6.38×10^{-8}	4.20×10^{-7}	-0.012 [-0.017, -0.007]	-0.299
KEAP1 (19:10486125)	WHRadjBMI	pLoF plus damaging missense	88	0.00034 (273)	1.76×10^{-11}	2.52×10^{-12}	0.012 [0.009, 0.016]	0.305
MC4R (18:60371062)	BMI	pLoF	23	0.00023 (187)	9.30×10^{-13}	1.87×10^{-13}	0.021 [0.016, 0.027]	0.532
PDE3B (11:14643804)	WHRadjBMI	pLoF	68	0.00132 (1,057)	5.91×10^{-10}	7.09×10^{-10}	-0.007 [-0.009, -0.005]	-0.166
	Body fat %	pLoF	66	0.00132 (1,046)	1.34×10^{-7}	2.09×10^{-7}	0.005 [0.003, 0.007]	0.102
PKD1 (16:2088708)	WHRadjBMI	pLoF	97	0.00022 (176)	7.68×10^{-11}	8.50×10^{-7}	0.012 [0.007, 0.017]	0.311
PLD1 (3:171600404)	WHRadjBMI	pLoF plus damaging missense	240	0.01163 (9,324)	4.69×10^{-10}	2.88×10^{-9}	0.002 [0.001, 0.003]	0.044
PLIN1 (15:89664367)	WHRadjBMI	pLoF	48	0.00104 (836)	8.94×10^{-23}	8.98×10^{-21}	-0.009 [-0.011, -0.007]	-0.233
PLIN4 (19:4502192)	WHRadjBMI	pLoF	121	0.00444 (3,560)	1.00×10^{-13}	3.25×10^{-14}	0.004 [0.003, 0.005]	0.091
	Body fat %	pLoF	120	0.00444 (3,506)	1.23×10^{-10}	2.11×10^{-10}	-0.003 [-0.004, -0.002]	-0.080
PPARG (3:12287368)	Body fat %	pLoF plus damaging missense	66	0.00021 (168)	2.90×10^{-8}	4.34×10^{-7}	-0.012 [-0.016, -0.007]	-0.292
PTPRG (3:61561569)	BMI	pLoF	120	0.00024 (194)	1.37×10^{-8}	8.31×10^{-8}	0.015 [0.01, 0.021]	0.377
	Body fat %	pLoF	120	0.00024 (192)	1.35×10^{-7}	8.25×10^{-7}	0.011 [0.006, 0.015]	0.267
RIF1 (2:151409883)	Body fat %	pLoF	66	0.00011 (89)	1.41×10^{-7}	1.41×10^{-7}	0.017 [0.01, 0.023]	0.416
SLC12A5 (20:46021690)	BMI	pLoF	13	0.00002 (13)	1.21×10^{-7}	1.21×10^{-7}	0.058 [0.036, 0.079]	1.446

Table 4.1: Exome-wide significant gene-level associations. Gene-level exome-wide significance threshold: SKAT-O $P < 1.58 \times 10^{-7}$. Genomic location indicates the chromosome and base pair coordinates for the start of the gene in Genome Reference Consortium Human Build 38. Variants in the mask are those which are included in the SKAT-O test for a given gene. They include all variants in the gene with MAF $< 1\%$ and consequence annotation matching the specific mask. Total MAF is the sum of MAF for all variants in the mask. In parentheses, MAC is the total minor allele count for the gene among individuals included in the association test. The effect size of a gene is estimated using burden testing and is in units of the phenotype's SD. The 95% CI is defined by an interval centred on the effect size point estimate, ± 1.96 SEs. The unweighted effect size is a version of the burden effect size that does not weight variants by Beta(MAF; 1, 25), the default weighting used by SAIGE. Using the unweighted effect sizes puts the gene-level effect on the same scale as variant-level effects. See subsection S4.1.1 for calculation of unweighted effect size. WHRadjBMI, waist-to-hip ratio adjusted for BMI; BMI, body mass index; pLoF, predicted loss-of-function.

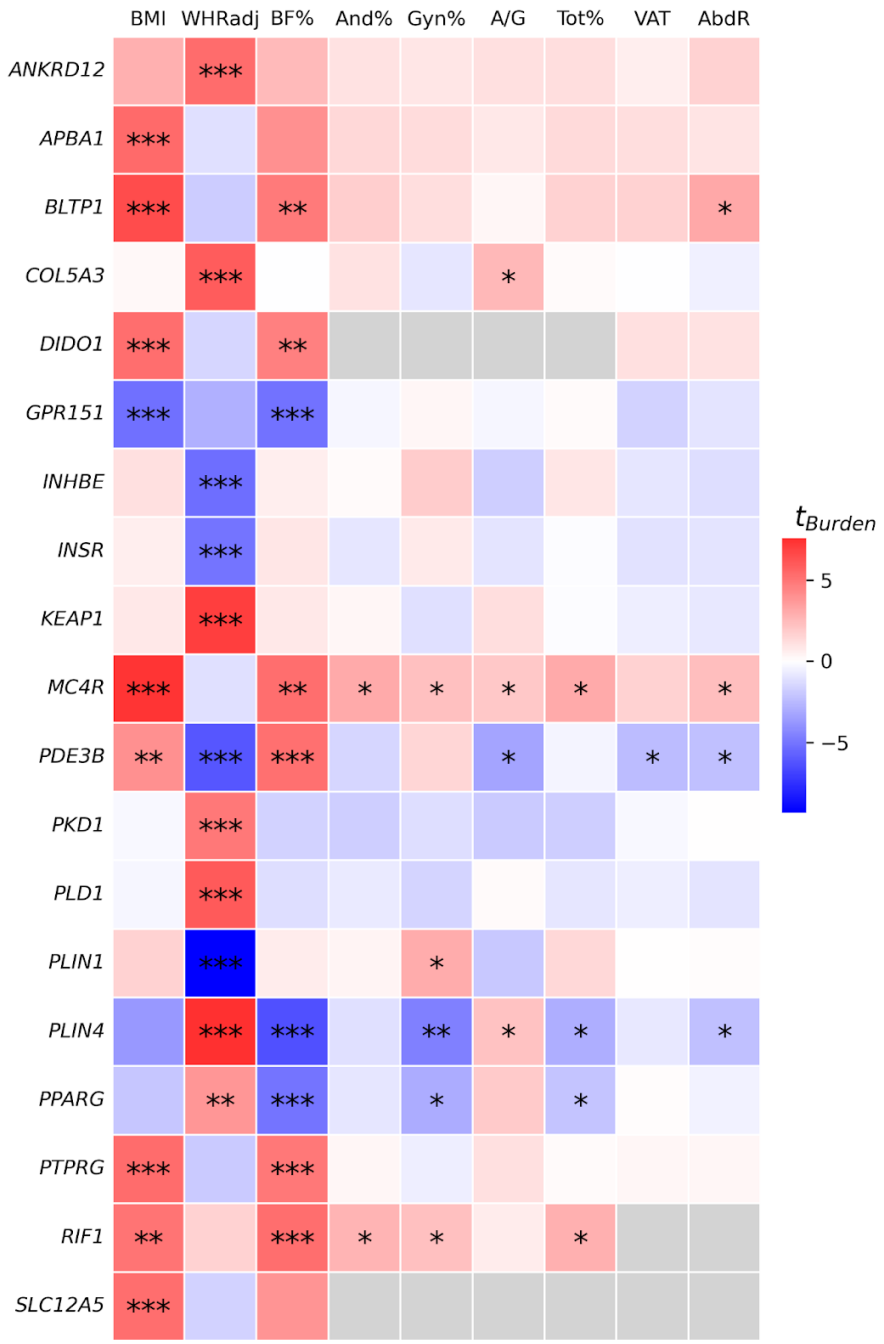

Figure 4.2: Gene burden effects across all nine obesity and fat distribution traits for 19 genes with exome-wide significant burden associations. Heatmap is coloured by the burden test t-statistic. Only the result of the consequence mask with the lowest SKAT-O P is shown for each trait-gene pair. Significance of SKAT-O test: ***Exome-wide significant ($P < 1.58 \times 10^{-7}$), **FDR≤1% significant ($P \leq 4.37 \times 10^{-5}$), *Nominal significance ($P < 0.05$). Grey squares indicate that the gene-level test could not be performed due to insufficient allele counts. WHRadj, waist-to-hip ratio adjusted for BMI. BF%, body fat percentage. And%, android fat percentage. Gyn%, gynoid fat percentage. A/G, android-gynoid fat percentage ratio. Tot%, total fat percentage. VAT, visceral adipose tissue. AbdR, abdominal fat ratio.

also have significantly higher PoPS compared to non-significant genes (one-sided t-test $P < 0.01$, Figure 4.1c), indicating value in examining these additional genes as potential therapeutic targets for obesity and lipodystrophy (Table S4.3). Seven additional gene-level associations also passed the FDR≤1% significance threshold, but were supported by minor allele counts less than ten (subsection S4.1.2). Combining PoPS across all gene-trait pairs (for traits with PoPS available: BMI, WHRadjBMI, body fat percentage) which are exome-wide or FDR≤1% significant and comparing to PoPS of non-significant gene-trait pairs, we observe strong enrichment of scores (one-sided t-test: $t = 8.47$, $P = 1.26 \times 10^{-17}$, mean [95% CI using $\pm 1.96 \times$ SEM] PoPS of 64 exome-wide or FDR≤ 1% significant gene-trait pairs = 0.307 [0.223, 0.392], mean of 103,841 non-significant gene-trait pairs = -0.010 [-0.012, -0.008]), demonstrating again the convergence of rare and common variant evidence when FDR≤1% significant genes are included.

Of the 31 genes known to cause severe monogenic obesity or lipodystrophy[336,355], three (*PPARG*, *PLIN1*, and *MC4R*) are associated with an obesity or fat distribution trait at the exome-wide level ($P < 1.58 \times 10^{-7}$) in our rare-variant gene-level association testing, representing a 125-fold enrichment over random chance (Fisher's exact test $P = 3.87 \times 10^{-6}$). Of the 69 FDR-significant genes, 23 have previously been reported to have rare variant gene-level associations with traits related to obesity or fat distribution in UKB participants, across a variety of variant annotation and gene-level association testing methods (Table S4.8)[369,370,372,383]. We looked up the effects of pLoF variant burden in FDR-significant genes with MAC≥10 in the multi-ancestry meta-analysed All of Us (AoU) tables[384] for available phenotypes: BMI and WHRadjBMI. Accounting for the multiple testing burden of 69 genes and two phenotypes (Bonferroni adjustment, $P < 3.62 \times 10^{-4}$), we replicated, with matching effect directions, the associations of *SLTM*, *MC4R*, and *KIF1B* with BMI, and *PKD1* and *PLIN4* with WHRadjBMI (Table S4.4). We also found that *HECTD4* is positively associated with body fat percentage in UKB and BMI in AoU; and an additional 18 of our FDR-significant genes are nominally associated ($P < 0.05$) with an obesity-related trait in AoU (Table S4.4).

Lastly, to evaluate the phenome-wide effects of genes related to obesity and fat distribution, we scanned 4,529 phenotypes in Genebass[199] summary statistics for associations with the 69 FDR-significant genes. We found 549 significant associations (Genebass SKAT-O $P \leq 9.98 \times 10^{-6}$, FDR$\leq 1\%$) across 211 traits for 41/69 genes (Figure S4.4, Table S4.5). The most significant Genebass associations were among blood biochemistry lipid phenotypes (Figure S4.4a, Table S4.5). We also observed shared associations across FDR-significant genes for phenotypes significantly genetically correlated to BMI and body fat percentage, such as fat mass and distribution traits. For example, eight FDR-significant genes (*BLTP1, DIDO1, GPR151, MC4R, PDE3B, PLIN4, RIF1, UBR2*) are associated with arm fat mass, leg fat mass, arm fat percentage, whole body fat mass, and hip circumference (all traits have common variant LD score regression (LDSC)[71] $r_g > 0.80$, $P \leq 1 \times 10^{-308}$ with BMI and body fat percentage, Table S4.5)[316]. We see some evidence for pleiotropy, as defined by associations with phenotypes genetically uncorrelated with BMI or body fat percentage, among results with common variant genetic correlations available: *PDE3B* and *STAB1* are associated with platelet distribution width (BMI r_g=0.025, P=0.274; body fat percentage $r_g = 9.7 \times 10^{-4}$, $P = 0.958$) and *ANKRD12* is associated with neutrophil percentage (BMI r_g=−0.029, P=0.117; body fat percentage $r_g = -8.8 \times 10^{-3}$, $P = 0.639$) (Table S4.5).

4.3.3 Sex- and age-specific effects of genes associated with obesity and fat distribution

We performed all association tests in sex-specific strata, identifying three genes (*INSR, PDE3B, PLIN4*) with significant differences (sex-difference two-sided heterogeneity test of effect sizes $P_{\text{diff}} < 2.67 \times 10^{-6}$, Bonferroni adjusted for 18,737 genes tested for sex-differential effects; Methods) in burden effect sizes. We observed opposing sex-specific effects for *INSR* on WHRadjBMI (burden β_{female} (SE) = -0.00458 (9.44×10^{-4}), $\beta_{\text{male}} = 0.00195$ (0.00100), $P_{\text{diff}} = 1.66 \times 10^{-6}$). We found that female-specific effects may be driving the reported sex-combined associations between WHRadjBMI and *PDE3B* ($\beta_{\text{female}} = -0.0169$ (0.00169), $\beta_{\text{male}} = -0.00288$

(0.00179), sex-difference $P = 8.02 \times 10^{-9}$) and *PLIN4* (β_{female}=0.00822 (9.15×10^{-4}), $\beta_{\text{male}} = 0.00108$ (9.86×10^{-4}), $P_{\text{diff}} = 8.19 \times 10^{-8}$) (Figure S4.5a, Table S4.6). Within the sex-specific analysis, we find two genes with female-specific associations ($P < 2.67 \times 10^{-6}$, Bonferroni adjusted for 18,737 genes tested for sex-specific effects) of pLoF variation on obesity and fat distribution (*CASQ1* with body fat percentage, female SKAT-O $P = 1.12 \times 10^{-6}$; *PLXND1* with WHRadjBMI, female $P = 9.69 \times 10^{-10}$), and two genes with male-specific associations, both for the ratio of android to gynoid tissue fat percentage (*ZNF841* male $P = 1.04 \times 10^{-6}$; *TINF2* male $P = 4.84 \times 10^{-7}$) (Figure S4.5b, Table S4.6), which do not reach significance at the FDR$\leq$ 1% threshold ($P \leq 4.37 \times 10^{-5}$) in sex-combined analyses.

We leveraged the age at diagnosis of obesity in longitudinal health records linked to UKB participants to identify that a burden of rare missense variation (MAF $< 1\%$) in five genes was associated ($P < 2.42 \times 10^{-4}$, Bonferroni adjusted for 207 gene-consequence mask pairs) with elevated lifetime risk of obesity (Figure S4.6, Table S4.7). Individuals carrying rare missense variants in *MC4R* (Cox proportional hazard ratio [HR] (SE)=1.46 (0.0843), $P = 7.75 \times 10^{-6}$), men with pLoF variants in *SLTM* (HR=5.37 (0.354), $P = 2.05 \times 10 - 6$) and damaging missense variants in *PCSK1* (HR=1.86 (0.165), $P = 1.63 \times 10^{-4}$), and women with pLoF variants in *DIDO1* (HR=11.2 (0.578), $P = 2.85 \times 10^{-5}$) and *SLC12A5* (HR=14.8 (0.708), $P = 1.45 \times 10^{-4}$) were at risk of earlier age at onset of obesity (Figure S4.6, Table S4.7).

4.3.4 Nominating target genes for *in vitro* functional characterisation through CRISPR knockdown

As adipose tissue is enriched for the expression of fat distribution genes identified from common variant GWASs[365], we assessed the messenger ribonucleic acid (mRNA) counts of overall and tissue-specific fat distribution-associated genes from our exome-wide gene-level association testing in differentiated human white adipose tissue (hWAT) derived pre-adipocytes *in vitro* (Figure 4.3a). Genes with FDR-significant results in gene-level tests had higher expression in hWAT 8 and 24 days after differentiation compared to non-significant genes (one-sided *t*-test $P = 1.55 \times 10^{-43}$,

comparing mRNA normalised counts; Figure 4.3a). From the 56 of 69 FDR-significant genes sufficiently expressed in differentiated hWAT (mean normalised mRNA count > 10 at 8 and 24 days after differentiation), we selected for functional characterisation 14 genes (*ABCA1, COL5A3, DENND5B, EXOC7, HERC1, INSR, MFAP5, MLXIPL, PCSK1, PLIN1, PPARG, SLTM, TRIP10, UBR2*; Figure 4.3b; Figure S4.7) which: (1) demonstrated monotonic allelic series indicating a dose-response relationship with gene dosage (*ABCA1, COL5A3, DENND5B, INSR, PCSK1, PLIN1, PPARG*) (Figure S4.3), (2) associated with obesity age-of-onset (*SLTM, PCSK1*) (Figure S4.6, Table S4.7), or (3) are implicated in lipid or glucose metabolism pathways by mouse whole-body or adipocyte-specific gene perturbation (Table S4.8).

We selected four guide RNAs (gRNAs) per gene target to generate 14 knockdown hWAT cell lines (Methods, Figure 4.3c, and Table S4.9). We additionally generated a negative control cell line with Cas9 empty vector. We confirmed gene knockdown through a two-step process: (1) RNA-seq expression of the target gene in knockdown cell lines relative to expression in Cas9-empty cell line and observed significantly decreased ($P < 0.05$, one-sided Wald test) expression for *COL5A3, DENND5B, EXOC7, HERC1, INSR, MFAP5, MLXIPL, SLTM, TRIP10*, and *UBR2* (Figure 4.3d, Table S4.10) and (2) RT-qPCR measurements of RNA for genes whose knockdown could not be confirmed via RNA-seq and observed significantly decreased ($P < 0.05$, one-sided t-test) for *ABCA1* and *PPARG* (Figure 4.3e, Table S4.12). Expression of *PCSK1* quantified by RT-qPCR was not significantly decreased compared to the control (RT-qPCR fold change [95% CI] = 0.624 [0.373, 1.045], one-sided t-test $P = 0.0665$), but was suggestive of a knockdown. We retained the *PCSK1* possible knockdown in downstream analysis, with a note of caution in the interpretation of results. The only gene knockdown which could not be confirmed and was not close to having significantly decreased expression was *PLIN1* (RT-qPCR fold change [95% CI]=0.933 [0.390, 2.234], one-sided t-test $P = 0.450$).

Abnormal adipogenesis and lipid accumulation are hallmarks of adipose tissue dysfunction[385]. As the process of adipocyte differentiation is considered a

necessary precursor to lipid accumulation[386], we use boron dipyrromethene (BODIPY) staining[387] (Methods) to measure both lipid accumulation and adipogenesis more generally. The BODIPY assay was conducted for six replicates of each knockdown cell line.

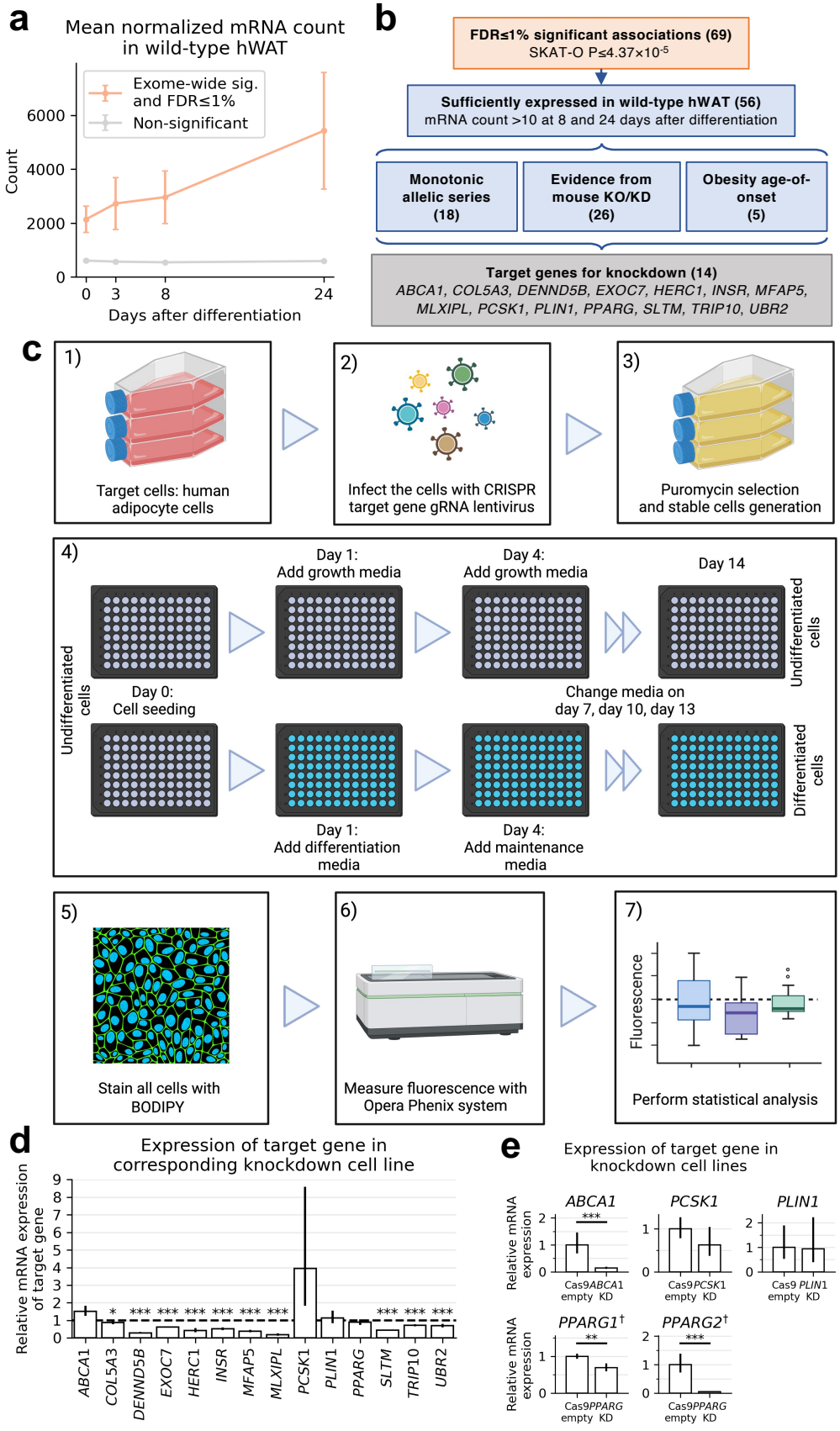
a
Mean normalized mRNA count in wild-type hWAT
Exome-wide sig. and FDR≤1%
Non-significant
Count
Days after differentiation
b
FDR≤1% significant associations (69)
SKAT-O P≤4.37×10-5
Sufficiently expressed in wild-type hWAT (56)
mRNA count >10 at 8 and 24 days after differentiation
Monotonic allelic series (18)
Evidence from mouse KO/KD (26)
Obesity age-of-onset (5)
Target genes for knockdown (14)
ABCA1, COL5A3, DENND5B, EXOC7, HERC1, INSR, MFAP5, MLXIPL, PCSK1, PLIN1, PPARG, SLTM, TRIP10, UBR2
c
1)
Target cells: human adipocyte cells
2)
Infect the cells with CRISPR target gene gRNA lentivirus
3)
Puromycin selection and stable cells generation
4)
Undifferentiated cells
Day 0: Cell seeding
Day 1: Add growth media
Day 4: Add growth media
Day 14
Undifferentiated cells
Change media on day 7, day 10, day 13
Day 1: Add differentiation media
Day 4: Add maintenance media
Differentiated cells
5)
Stain all cells with BODIPY
6)
Measure fluorescence with Opera Phenix system
7)
Fluorescence
Perform statistical analysis
d
Expression of target gene in corresponding knockdown cell line
Relative mRNA expression of target gene
ABCA1 COL5A3 DENND5B EXOC7 HERC1 INSR MFAP5 MLXIPL PCSK1 PLIN1 PPARG SLTM TRIP10 UBR2
e
Expression of target gene in knockdown cell lines
ABCA1
PCSK1
PLIN1
PPARG1†
PPARG2†
Relative mRNA expression
Cas9 empty KD

Figure 4.3 ***(previous page)*: CRISPR knockdown strategy**. **a**, Mean normalised mRNA counts across four time points (days after differentiation: 0, 3, 8, 24) in human white adipocyte tissue for 69 genes passing FDR≤1% significance (SKAT-O $P \leq 4.37 \times 10^{-5}$) versus all non-significant genes (n=60,605). Error bars indicate 95% CIs as defined by $\pm 1.96 \times$ SEM normalised mRNA count. **b**, Selection strategy for genes to include in conditional knockdown experiment. Numbers in parentheses indicate the number of genes in a given group. **c**, Experimental design for CRISPR knockdown, adipocyte cell culture and imaging. **d**, Confirmation of knockdown efficacy using relative mRNA expression in differentiated adipocytes. Error bars indicate 95% CIs defined by $\pm 1.96\times$ SEM. One-sided Wald test comparing mRNA count in knockdown cell line relative to count in Cas9 empty, with the alternative hypothesis that the count is lower in the knockdown: $^{***}P < 0.001$, $^{**}P < 0.01$, $^{*}P < 0.05$. **e**, Confirmation of knockdown efficacy using relative RNA expression quantified with RT-qPCR in undifferentiated adipocytes. Error bars indicate 95% CIs defined by $\pm 1.96 \times$ SEM. One-sided *t*-test comparing mRNA count in knockdown cell line relative to count in Cas9 empty, with the alternative hypothesis that the count is lower in the knockdown: $^{***}P < 0.001$, $^{**}P < 0.01$, $^{*}P < 0.05$. Knockdown confirmation for *PPARG* was performed by checking two isoforms, *PPARG1* and *PPARG2*. †*PPARG* knockdown confirmation was performed in differentiated adipocytes because expression of *PPARG* isoforms was too low in undifferentiated adipocytes.

4.3.5 Knockdown of obesity and fat distribution associated genes alters *in vitro* adipogenesis and lipid accumulation

We observed significantly reduced ($P < 0.05$ in two-sided *t*-test of mean FC in fluorescence between knockdown and control Cas9-empty cell line) lipid accumulation in the *PPARG* knockdown cell line (Figure 4.4, Table S4.13). Inactivation of *PPARG* through knockout has previously been shown to lower lipid accumulation in human and mouse pre-adipocytes[374,388,389], providing validation of the assay used here.

We also noted reduced lipid accumulation in knockdown of the novel target *SLTM* (FC=0.514, $P = 1.91 \times 10^{-4}$) (Figure 4.4a, Table S4.13). *SLTM* encodes the SAFB-like transcription modulator, which regulates the GLI family of transcription factors in mice[390]. These GLI transcription factors in turn control expression of lipid metabolic genes, including critical adipogenesis transcription factors *PPARG* and *C/EBP* (α, β, γ, and δ)[391]. Notably, the *SLTM* pLoF burden association was replicated in a multi-ancestry meta-analysis conducted in AoU[384] as the top association for BMI (AoU meta-analysed $P = 2.75 \times 10^{-7}$). While this suggests a novel therapeutic target for altering fat deposition, more work is

needed to understand why individuals carrying rare pLoF variants in *SLTM* have higher BMI (burden β (SE) = 0.0191 (0.00455), SKAT-O $P = 2.65 \times 10^{-6}$), and the target tractability of *SLTM*, which may be amenable to small molecule binding (Table S4.14).

We observed significantly increased lipid accumulation relative to Cas9-empty cells in knockdowns of *EXOC7* (FC=1.35, P=0.0096), *TRIP10* (FC=1.39, P = 0.0157), and *COL5A3* (FC=1.72, P = 0.0028) (Figure 4.4a, Table S4.13). The products of these genes are all variously involved in lipid uptake, synthesis, and adipogenesis. *EXOC7* is a component of the exocyst complex, which regulates the uptake of free fatty acids by adipocytes[392]. *TRIP10* is an interactor of the thyroid hormone receptor (TR-β1)[393], which regulates *de novo* fatty acid synthesis[394]. TR-β1 has been proposed as a target to treat dyslipidaemia[395], and the TR-β1 agonist resmetirom is currently in Phase III clinical trials for treating non-alcoholic fatty liver disease[396]. However, there are no known drugs directly targeting *TRIP10* (Table S4.14). We also observed increased lipid accumulation in knockdowns of *COL5A3* (FC=1.72, $P = 2.78 \times 10^{-3}$) (Figure 4.4a, Table S4.13) – as a ubiquitous component of the extracellular matrix, type V collagen informs the proper differentiation and development of adipocytes[397]. A Phase II clinical trial of collagenase histolyticum, a hydrolytic enzyme that degrades COL5A3, has proven successful in dissolving lipomas, i.e. benign lumps of fatty tissue underneath the skin[398] (Table S4.14). The *PCSK1* knockdown cell line also demonstrated significantly increased lipid accumulation, however we could not draw any conclusions because the knockdown was not confirmed.

4.3.6 Transcriptome-wide effects of gene knockdowns

Obesity-gene knockdowns had wide-ranging effects on the hWAT transcriptome, with between eight and 1,904 genes differentially expressed ($|\log_2(\text{FC})| > 1$, FDR-adjusted $P < 0.05$) in knockdown cell lines as compared to Cas9-empty negative controls (Figure 4.5 and Table S4.15). We examined the molecular effects of these knockdowns through pathway enrichment analyses for pathways in

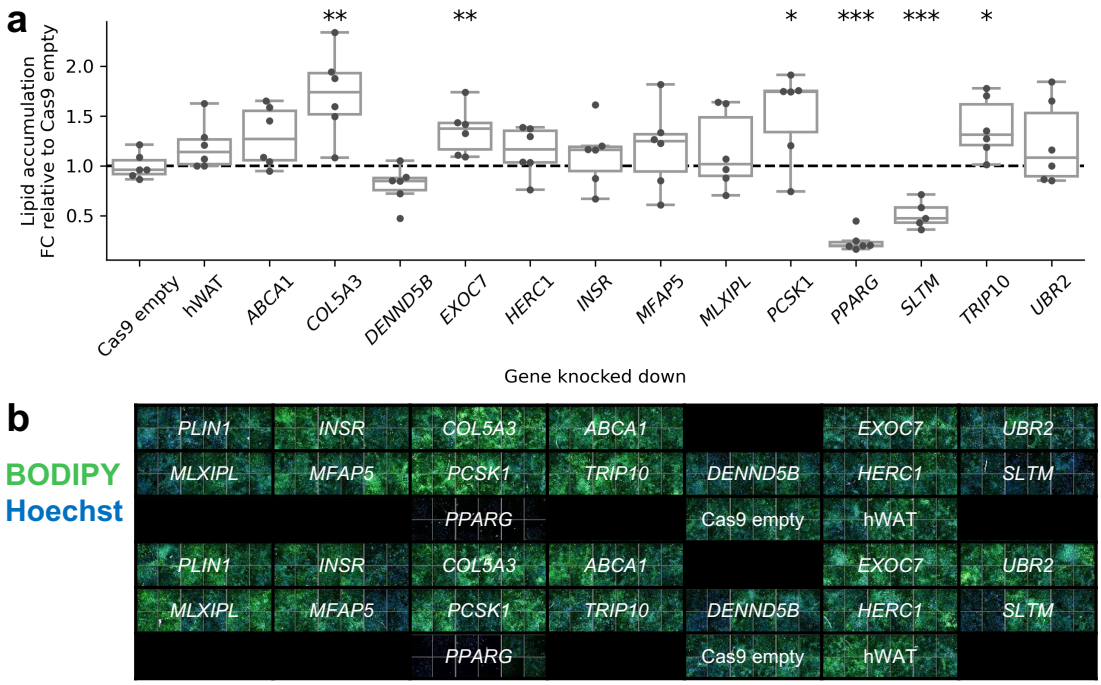

Figure 4.4: Effect of CRISPR single-gene knockdown on lipid accumulation in human adipocytes. **a**, Effect of single-gene knockdown on lipid accumulation measured by whole-well mean cytoplasm intensity of BODIPY dye fluorescence. Points represent measurements from separately-cultured cells (biological replicates) from the same knockdown cell line. Each knockdown cell line has six replicates, except *SLTM*, which had one outlying replicate removed (Methods). **b**, Differentiated human adipocyte knockdown cell lines stained with BODIPY (green) and Hoechst (blue), imaged at 10x magnification. The elements of the box plots are median (horizontal line in box), upper (Q3) and lower (Q1) quartiles marking the ends of the box, with whiskers extending to the most extreme points within the range [Q1-IQR, Q3+IQR], where IQR = Q3-Q1. All points are shown. Two-sided *t*-test of mean fold change relative to Cas9-empty cells: $^{***}P < 0.001$, $^{**}P < 0.01$, $^{*}P < 0.05$. Summary statistics are available in Table S4.13. FC, fold change. hWAT, human white adipose tissue.

the KEGG[399], REACTOME[400], and GO:Biological Process and GO:Molecular Function[401] databases.

Knockdown of *PPARG*, a master regulator of adipogenesis, significantly alters 95 pathways (family-wise error rate (FWER) P<0.05), including: (1) transcriptional regulation through PRC2-driven DNA methylation (number of genes (n) = 59, $P < 0.0001$), histone acetylation ($n = 138$, $P < 0.0001$), and small RNAs ($n = 102$, $P = 0.0001$), (2) protein synthesis by modulating rRNA expression through SIRT1 (n=63, $P < 0.0001$), ERCC6 (CSB) and EHMT2 (G9a) ($n = 71$, $P < 0.0001$)

and the B-WICH chromatin remodelling complex ($n = 86$, $P = 0.0001$), and (3) functions critical to cell division such as assembly of the ORC complex at the origin of replication ($n = 64$, $P < 0.0001$), condensation of prophase chromosomes ($n = 69$, $P < 0.0001$), and cell cycle checkpoints ($n = 286$, $P < 0.0001$) (Table S4.16). As expected, *PPARG* knockdown also alters pathways important for adipogenesis, such as the transcription of androgen-receptor regulated genes ($n = 62$, $P < 0.0001$), RUNX1 regulation[402] ($n = 89$, $P = 0.0001$), formation of the beta-catenin TCF trans-activating complex[403] (n=87, $P = 0.0001$), and pre-NOTCH expression and processing[404] ($n = 104$, $P = 0.0003$) (Table S4.16).

We observed enrichment of several mitotic, transcription, and translation pathways in knockdowns of lipid accumulation-altering genes *EXOC7* and *TRIP10* (Table S4.16), but did not find mechanisms directly implicated in lipid accumulation. On the other hand, while *MLXIPL* knockdown did not affect lipid accumulation in our assays, we nevertheless observed enrichment (FWER $P < 0.05$) of pathways involved in adipogenesis, such as interferon α/β signalling[405] ($n = 66$, $P = 0.0016$) and lipopolysaccharide binding[406] ($n = 32$, $P = 0.0284$). *MLXIPL* encodes the carbohydrate-responsive element-binding protein (ChREBP), which is thought to regulate gene expression in response to glucose and reduces adipose tissue mass in mouse knockout models[407]. While this may be a potential therapeutic target for fat distribution, mice deficient in *MLXIPL*/ChREBP do not tolerate fructose in their diet and develop severe diarrhoea and irritable bowel syndrome[408].

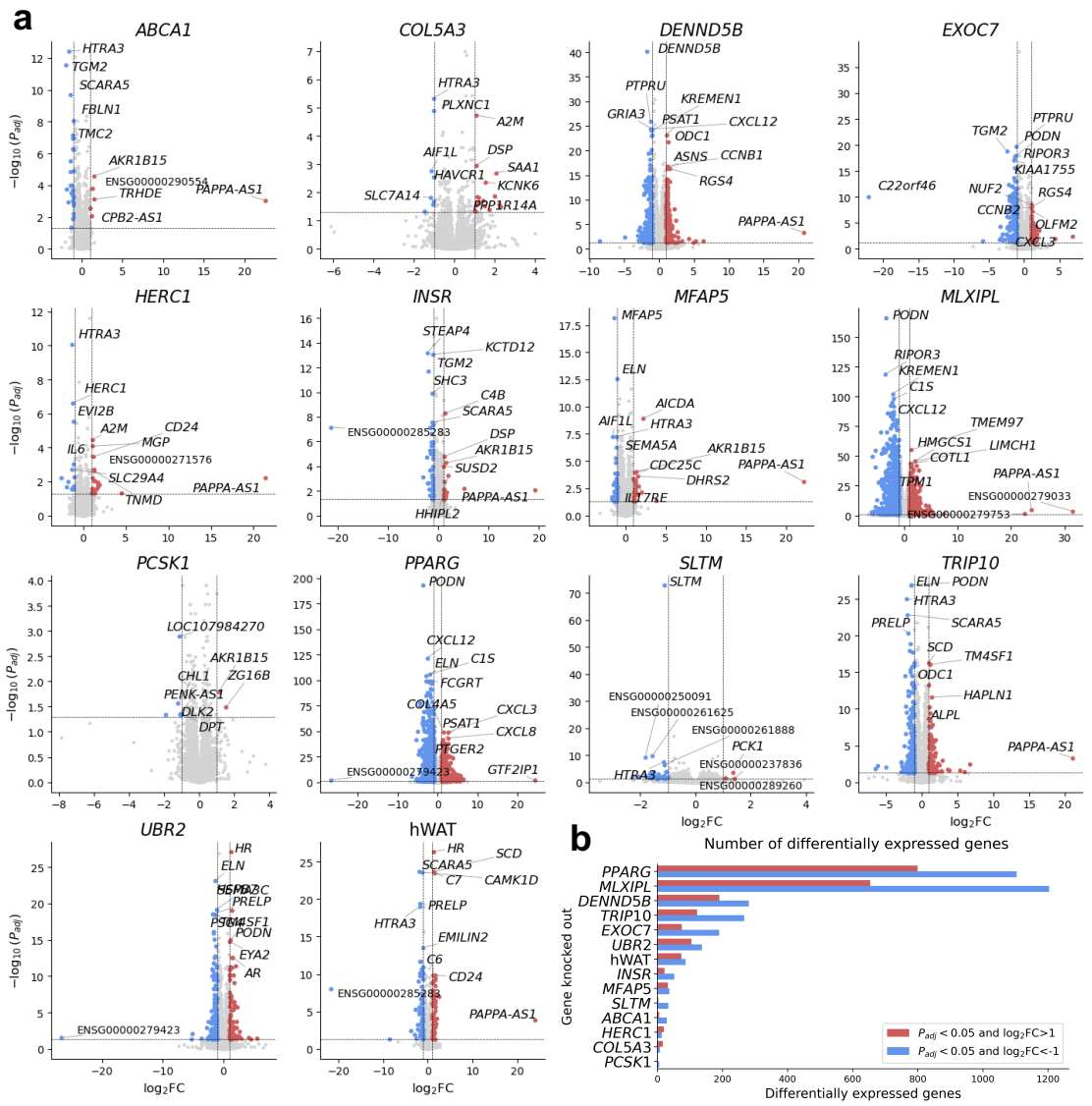

Figure 4.5: Effect of gene knockdown on differential gene expression in human adipocytes. **a**, Gene differential expression results from RNA sequencing data performed on hWAT wild-type and single-gene knockdowns. Gene differential expression analysis was performed with DESeq2[409], comparing each cell line (wild-type hWAT or knockdown) to the Cas9-empty cell line. All significance thresholds use P_{adj}, a P-value corrected for multiple testing calculated in DESeq2. Blue points indicate genes which are significantly downregulated ($P_{\text{adj}} < 0.05$ and $\log_2 \text{FC} < 1$). Red points indicate genes which are significantly upregulated ($P_{\text{adj}} < 0.05$ and $\log_2 \text{FC} > 1$). Grey points are all other genes. Genes are labelled if they are in the top five most significant (lowest P_{adj}) red or top five most significant blue points, in the top five most significant regardless of point colour, or have $|\log_2 \text{FC}| > 10$. The horizontal and vertical dashed lines indicate the significance cutoffs of $P_{\text{adj}} < 0.05$ and $|\log_2 \text{FC}| > 1$, respectively. **b**, Counts of significantly downregulated (blue, $P_{\text{adj}} < 0.05$ and $\log_2 \text{FC} < -1$) and upregulated (red, $P_{\text{adj}} < 0.05$ and $\log_2 \text{FC} > 1$) genes for each single-gene knockdown and hWAT. The y-axis is sorted by the total number of differentially expressed genes for each knockdown condition. FC, fold change.

4.4 Methods

4.4.1 Imputed data QC

Sample QC and assigning population labels

Assignment of population labels was performed by Duncan Palmer.

Beginning with the 487,409 individuals with phased (UKB field ID: 22438) and imputed (imputation performed by Wellcome Trust Centre for Human Genetics, UKB field ID: 22828) genotype data, we restricted to unrelated individuals with low autosomal missingness rates used for PCs by Bycroft *et al.*[74]. We then used genotyping array data, subset to LD-pruned autosomal variants, from these samples to project into the PC space defined by the 1000 Genomes dataset[66], ensuring that we correctly account for shrinkage bias in the projection[410]. Next, we used the "super-population" labels (AFR=Africans, AMR=Admixed Americans, EAS=East Asians, Europeans=EUR, South Asians=SAS) of the 1000 Genomes dataset to train a RF classifier, using the randomForest (4.6) library in R[411], and predicted the super-population for each of the UKB samples. Samples with classification probability >0.99 for the European super-population, were retained for downstream analysis.

Following our ancestry based filtering regime, we remove samples who withdrew from UKB participation as well as those individuals who were omitted from phasing and imputation. For GWAS, we further subset individuals passing imputed data QC to those who also passed exome sequencing QC (described below), resulting in a set of 397,315 individuals.

Variant QC

Over 92 million imputed variants across the autosomes and chromosome X are available for analysis. As a starting point for our initial collection of GWAS, we subset to variants with MAF $> 0.1\%$ in the subset of individuals defined in the sample QC procedure and an INFO score >0.8 from the UKB SNP manifest file. Following this collection of filtering steps, 16.7 million variants were retained for common variant GWAS. After performing GWAS, two additional filters were

performed, retaining variants with Hardy-Weinberg $P > 1 \times 10^{-10}$, calculated on the whole QCed subset of individuals, and retaining variant-level summary statistics if MAF $> 0.1\%$, calculated for each GWAS. This results in 13,117,850 variants for downstream common variant analysis.

4.4.2 Exome sequencing data QC

Exome sequencing data QC summary

I wrote all code for exome sequencing data QC and generated an initial version of QCed data. Following an update to variant consequence annotations, Barney Hill reran my code to generate the final set of QCed data.

Using the UKB RAP we accessed the gnomAD VCF files containing GATK-called genotypes for 454,671 individuals. We followed an identical sample QC protocol to Karczewski *et al.*[199], filtering individuals on covariate-regressed individual-level metrics using MADs. We removed variants flagged by the gnomAD QC team for failing one or more of their filters (`AC0`, `RF`, `MonoAllelic`, `InbreedingCoeff`). Genotype calls were set to missing if they failed filters for genotype quality, depth and allele balance (described in Karczewski *et al.*[199]). We filter to European ancestry using the super-population ancestry labels assigned with an RF classifier, described above for imputed data QC.

Variant and sample-level QC

We defined "high quality" variants as those MAF $> 0.1\%$ and call rate ≥ 0.99 falling within the UKB capture intervals plus 50 bp padding. These variants were used to evaluate sample-level metrics of mean call rate and depth and retained samples satisfying all of the following:

- Genetic sex inferred as XX or XY (specifically, genetic sex is defined).
- Mean call rate ≥ 0.99 among high quality variants.
- Mean coverage $\geq 20\times$ among high quality variants.
- Not withdrawn.

Next, I removed variants satisfying at least one of the following criteria:

- The variant lies outside the UKB capture plus 50 bp padding.
- The variant lies within a low complexity region.
- The variant lies within a segmental duplication.

In this (sample, variant) set, I ran Hail's `sample_qc()`[252] to remove samples lying outside the median±4 MADs within each super-population (see above section on imputed data QC). The QC protocol was split by UKB ES tranche (50k, 200k, 450k) to guard against batch effects, as tranches were sequenced in separate runs. The following metrics were used for QC:

- Number of deletions (`n_deletion`).
- Number of insertions (`n_insertion`).
- Number of SNPs (`n_snp`).
- Ratio of insertions to deletions (`r_insertion_deletion`).
- Ratio of transitions to transversions (`r_ti_tv`).
- Ratio of heterozygous variants to homozygous alternate variants (`r_het_hom_var`).

Following MAD filtering (Figure S4.1, Table S4.1), 402,375 European samples were retained for analysis. For each sample, we excluded non-passing sites as described in Karzcewski *et al.*[199]. Briefly, an RF classifier was trained to distinguish true positives from false positive variants using a collection of allele and site annotations. Variants were assigned "PASS" to maximise sensitivity and specificity across a series of readouts in trio data and precision-recall in two truth samples, after which samples with excess heterozygosity (defined as inbreeding coefficient < -0.3) were removed. Next, we removed low quality genotypes by filtering to the subset of genotypes with depth ≥ 10 (5 among haploid calls), genotype quality ≥ 20, and minor allele balance > 0.2 for all alternate alleles for heterozygous genotypes. Following this filter, we remove variants that were not called as "high quality" among any sample. The resulting high quality European call set consisted of 402,375 samples and 25,229,669 variants.

4.4.3 Variant consequence annotation

Variant annotation was performed by Frederik Lassen and Barney Hill.

We annotated ES variants using Variant Effect Predictor (VEP) v105 (corresponding to GENCODE v39)[412] with the LOFTEE v1.04_GRCh38[113] and dbNSFP[413] plugins, annotating variants with CADD v1.6[110], and REVEL using dbNSFP4.3[109] and LoF confidence using LOFTEE. We provide code and instructions for this step in our VEP_105_LOFTEE repository[414], which contains a Docker/Singularity container for reproducibility of annotations. Next, we ran SpliceAI v1.3[111] using the GENCODE v39 gene annotation file to ensure alignment between VEP and SpliceAI transcript annotations. For variant-specific annotations we use "canonical" transcripts. We separated variants by transcript using `bcftools +split-vep` and filtered to MANE Select[203] protein-coding transcripts. If the gene lacked a MANE Select transcript we selected the canonical transcript defined by GENCODE v39. Using this collection of missense, pLoF, and splice metrics, and annotations of variant consequence on the canonical transcript, we then determine a set of variant categories for gene-based testing.

Variant consequence categories

1. High confidence pLoF: High-confidence LoF variants, as defined by LOFTEE[113] (LOFTEE HC).
2. Damaging missense/protein-altering: At least one of:
   (a) Variant annotated as missense/start-loss/stop-loss/in-frame indel and (REVEL≥0.773 or CADD≥28.1 or both).
   (b) Any variant with SpliceAI delta score (DS)≥0.2 where SpliceAI DS is the maximum of the set {`DS_AG`, `DS_AL`, `DS_DG`, `DS_DL`} for each annotated variant (where `DS_AG`, `DS_AL`, `DS_DG` and `DS_DL` are delta score (acceptor gain), delta score (acceptor loss), delta score (donor gain), and delta score (donor loss), respectively).
   (c) Low-confidence LoF variants, as defined by LOFTEE (LOFTEE LC)

3. Other missense/protein-altering: Missense/start-loss/stop-loss/in-frame indel not categorised in (2).
4. Synonymous: Synonymous variants with SpliceAI DS<0.2 in the gene.

REVEL and CADD score cut-offs are chosen to reflect the supporting level for pathogenicity (PP3) from the American College of Medical Genetics and Genomics and the Association for Molecular Pathology (ACMG/AMP) criteria[112]

4.4.4 Phenotype curation

We used the following nine phenotypes, either directly measured by UKB or derived from UKB phenotypes: body mass index (UKB field ID: 21001), waist-to-hip ratio (derived from waist circumference [UKB field ID: 48] and hip circumference [UKB field ID: 49]) with BMI regressed out, body fat percentage (UKB field ID: 23099), android tissue fat percentage (UKB field ID: 23247), gynoid tissue fat percentage (UKB field ID: 23264), ratio of android to gynoid tissue fat percentage (derived from android and gynoid tissue fat percentage phenotypes), total tissue fat percentage (UKB field ID: 23281), visceral adipose tissue volume (UKB field ID: 22407), abdominal fat ratio (UKB field ID: 22434).

4.4.5 Genetic association testing

All variant- and gene-level associations were performed in the European-ancestry subset of the UKB using SAIGE[94], a mixed model framework that accounts for sample relatedness.

In SAIGE step 0, we constructed a GRM using the UKB genotyping array data. The genotyped data is LD pruned using PLINK (`--indep-pairwise 50 5 0.05`)[415], and the sparse GRM is calculated using the `createSparseGRM.R` function within SAIGE, using 5,000 randomly selected markers, with relatedness cutoff of 0.05,

To generate a variance ratio file for subsequent steps in SAIGE, we selected 2,000 variants from the genotyping array data to define a PLINK dataset. For testing common variants in imputed data, we extracted 2000 variants with MAC $\geq$ 20. For

testing the exome-sequencing data, we extracted two sets of 1,000 variants with $10 \leq$ MAC < 20 and MAC ≥ 20, and combined these sets of markers.

In SAIGE step 1 for each trait, the null model is fit using the curated phenotype data and sparse GRM, with no genetic contribution. Default parameters were used in SAIGE, except `--relatednessCutoff 0.05`, `--useSparseGRMtoFitNULL TRUE` and `--isCateVarianceRatio TRUE`. In the sex-combined analyses, we account for *age*, *sex*, age^2, *age* $\times$ *sex*, $age^2 \times sex$, the first 21 PCs, UKB assessment centre, and exome-sequencing tranche (50k[416], 150k[248] or 250k[114] tranches) as fixed effects. For sex-specific analyses, we exclude sex-dependent terms (*sex*, *age* $\times$ *sex*, $age^2 \times$ *sex*) when accounting for fixed effects. All continuous traits were IRNTed using the `--invNormalize` flag in SAIGE. For SAIGE step 2, we always use the flag `--LOCO FALSE` and `--is_fastTest TRUE`.

Common variant association testing in imputed data

We performed single-variant genetic association testing of the UKB imputed data using SAIGE version 1.1.6.3 on the Oxford Biomedical Research Computing cluster. For consistency across analyses, only individuals who passed sample-level QC in both the imputed and exome-sequencing data were included.

Following null model fitting, we carried out variant testing in SAIGE step 2 with default parameters, except for `--is_Firth_beta TRUE` and `--pCutoffforFirth=0.1`.

Finemapping

Using summary statistics from the common variant association tests, finemapping loci are identified for each trait. Starting with the most significant variant in the pool of genome-wide significant variants, a 1 Mb window centred around the variant is created. All genome-wide variants falling in this window are considered part of the locus defined by the most significant variant and removed from the pool of genome-wide significant variants. Then we proceed to the next most significant variant in the pool of genome-wide significant variants and repeat until the pool of genome-wide significant variants is empty. Loci with overlapping windows are merged.

LD matrices were calculated with LDstore v2.0[417], using the same set of individuals used in the GWAS for each trait. Finemapping was performed with FINEMAP v1.4[417], using shotgun stochastic search and allowing a maximum of 10 causal SNPs per locus.

Rare variant and gene-level testing in exome sequencing data

We carried out rare variant and gene-level genetic association testing in the European-ancestry subset of the UKB exome sequencing data using SAIGE. All analyses involving the exome sequencing data was carried out on the UKB RAP using SAIGE version wzhou88/saige:1.1.9[94].

Following null model fitting, we carried out variant and gene-level testing in SAIGE step 2 using the variant categories described above, with the `--is_single_in_groupTest=TRUE` flag. All other parameters were set to default, except `--maxMAF_in_groupTest=0.0001,0.001,0.01`, although we only report results for maximum MAF of 0.01. We included the following collection of group tests, using the annotations defined above (see subsection 4.4.3):

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

For genes with no damaging missense/protein altering variants, only the pLoF variant consequence mask was considered when performing Bonferroni multiple-testing correction. Variant-level and gene burden tests are always two-sided.

Obesity and fat distribution trait gene-level association tests

For each gene and for each obesity and fat distribution trait we tested the SKAT-O association of rare variant (MAF $< 1\%$) variation. Exome-wide statistical significance ($P < 1.58 \times 10^{-7}$) for gene-level tests was defined using Bonferroni correction for 315,996 unique phenotype-gene-consequence mask combinations. To estimate the strength and direction of effect of damaging variation in genes on phenotypes, we also performed gene burden tests, however, the P-values from the burden analysis were not used to determine significance.

To expand the set of obesity and fat distribution associated genes to consider for functional screening, we followed an FDR-control process similar to Zhou *et al.*[418]. We selected the gene-level association result which had the lowest SKAT-O P-value (across all phenotypes and both variant masks) and applied the BH procedure to identify 83 significant genes when controlling FDR to 1% (corresponding to a P-value of 4.37×10^{-5}). If we had performed the FDR-controlled selection using the full set of results instead of taking the result with the lowest P-value this would be equivalent to an FDR of 12.0%.

We re-ran the gene-level association tests for the 83 FDR-significant genes, conditional on the top finemapped common (MAF $> 0.1\%$) variants for finemapped loci (described above). For each trait, we identified finemapped common variants for each trait which have the highest Bayes factor for being a causal variant (i.e. strongest evidence for causality) in their respective loci. Variants tied for the highest Bayes factor are all selected. There are a total of 915 finemapped trait-variant pairs (846 unique variants) on which to condition. We used the `--condition` flag in SAIGE to condition on the selected common variants when performing gene-level tests for the corresponding phenotype. We perform the gene-level association tests for each chromosome independently and only conditioned on common variants from the same chromosome as the genes being tested.

After conditioning, seven genes (*ACVR1C, PRPH, PYGM, SEC16B, TNFRSF6B, TRIM40, TUBE1*) were no longer significant at the predefined SKAT-O $P \leq 4.37 \times 10^{-5}$ threshold, leaving 76 genes. Of these 76, seven genes (*DEFB112, CHMP4B, FEZF2, GLP1R, PCBD2, TM4SF20, VGF*) were flagged as having very low minor allele count (MAC<10), leaving 69 genes to be considered for functional screening.

Sex difference analyses

We use sex-specific estimated burden effect sizes (β) and SEs to calculate z-scores corresponding to sex-differential effects using the heterogeneity test described by Martin *et al.*[184]:

$$z = \frac{\beta_{\text{female}} - \beta_{\text{male}}}{\sqrt{SE^2_{\text{female}} + SE^2_{\text{male}}}} \tag{4.1}$$

The sex-differential P (P_{diff}) is estimated under a two-sided hypothesis:

$$P_{\text{diff}} = 2 \times \Phi(-|z|); \quad \Phi = \text{CDF}_{\text{Normal}} \tag{4.2}$$

Genes were considered to exhibit significant sex-differential effects if $P_{\text{diff}} < 2.67 \times 10^{-6}$, Bonferroni adjusted for 18,737 genes tested for sex-differential effects.

Age at diagnosis longitudinal analyses

This analysis was performed by Samvida Venkatesh. Plots of survival curves (Figure S4.6) were generated by me, using code written by Samvida Venkatesh.

We curated age at diagnosis of obesity from the UKB linked primary care and hospital record data using mapping tables generated by Kuan *et al.*[419]. Any codes related to "history of obesity", for which accurate age at diagnosis could not be extracted, were excluded. We left-truncated observations at the age of the first record (of any code) in either primary care or hospital data, and right-censored at the age of the last record. For each of the 69 FDR-significant genes identified from ES analyses, we performed Cox-proportional hazards modelling[420] to estimate differences in lifetime risk of developing obesity between carriers of pLoF, damaging missense, other missense, synonymous, or non-coding variants (at MAF < 1%) and wild-type individuals (reference group). All effects were adjusted for sex (in sex-combined analyses), the first 10 genetic PCs, birth cohort (in ten-year intervals from 1930 to 1970), and UKB assessment centre. Cox modelling was performed using the R package survival v3.3.1[421]. We visualised age at onset probabilities using Kaplan-Meier survival curves with the R package survminer v0.4.9[422].

4.4.6 Common variant gene prioritisation scores

We used Polygenic Priority Score (PoPS) calculated by Weeks *et al.*[382] for UKB BMI, WHRadjBMI and body fat percentage. Scores were downloaded from `https:`

//www.finucanelab.org/data. There were 34,635 gene-level PoPS available each for BMI, WHRadjBMI and body fat percentage, totalling 103,905 gene-trait PoPS.

4.4.7 AoU gene-level summary statistics

Validation of associated genes in AoU was performed by Duncan Palmer and Samvida Venkatesh.

We downloaded gene-level meta-analysed summary statistics from the All by All browser (https://allbyall.researchallofus.org/), for the two adiposity-related traits which were also available in the AoU[115] dataset: BMI and body fat percentage. Only meta-analysed MAF $< 1\%$ pLoF results corresponding to our FDR$\leq 1\%$ genes were used. The replication significance threshold (SKAT-O $P < 3.62 \times 10^{-4}$) was Bonferroni corrected for 138 tests (two traits, 69 genes).

4.4.8 Genebass gene-level summary statistics

Gene-level association test summary statistics were calculated by Karczewski *et al.*[199] and downloaded as a Hail Matrix Table from https://app.genebass.org/downloads. Only MAF$< 1\%$ pLoF-associated results corresponding to our 69 FDR $\leq 1\%$ genes were used. Associations were significant if Genebass SKAT-O $P \leq 9.98 \times 10^{-6}$, controlled for FDR$\leq 1\%$ using the BH procedure[293].

4.4.9 Genetic correlation estimates

Genetic correlation estimates were accessed from the Neale lab's UKB genetic correlation browser: https://ukbb-rg.hail.is/rg_browser/[316]. These genetic correlations were estimated using LDSC[423] v1.0.1, using summary statistics from common (MAF $> 0.1\%$) imputed variant GWAS. Genetic correlation P-values less than or equal to 1×10^{-308} are reported in Table S4.5 as zero due to numerical precision limits in Microsoft Excel.

4.4.10 Selecting target genes

Wild-type hWAT RNA-seq data was generated by Richard Yan-Do and Saskia Reibe.

To choose target genes for CRISPR-Cas9 knockdown we consider all MAC ≥ 10 gene association results with FDR < 0.01 which remain significant ($P \leq 4.37 \times 10^{-5}$) after conditioning on common variation significantly associated with the trait ("Rare variant and gene-level testing in exome sequencing data" methods). We then compiled multiple lines of evidence for each gene:

- Sufficiently expressed in wild-type hWAT cells (RNA-seq transcript counts > 10 at eight and 24 days after differentiation).
- Monotonic allelic series indicating a dose-response relationship with gene dosage and a obesity or fat distribution trait.
- Significant obesity age-of-onset association.
- Published functional work involving gene knockout or knockdown implicates the role of the gene in adiposity.

Using these lines of evidence we separated the genes into categories: positive controls (genes previously implicated in obesity pathways by functional work in human adipocytes), plausible candidates (genes which have not been implicated in adiposity pathways by functional work in human adipocytes, but have suggestive evidence from other functional work or from GWAS/burden association results), and impossible candidates (RNA-seq counts are too low or gene seems to be associated with a pathway that we are not testing). From these categories we chose 14 target genes for knockdown (*ABCA1, COL5A3, DENND5B, EXOC7, HERC1, INSR, MFAP5, MLXIPL, PCSK1, PLIN1, PPARG, SLTM, TRIP10, UBR2*).

4.4.11 Wet lab methods

All wet lab work involving data generation was performed by Ilknur Sur Erdem. Val Millar contributed expert help in the use of the PerkinElmer Opera Phenix/High Content Imaging. We are grateful to Dr. Elena Navarro-Guerrero and the Ebner group for cell culture-related lab resources. We also thank Professor Benedikt Kessler and Dr. Adan Pinto Fernandez for wet lab resources.

Cell culture and reagents

Human white adipose tissue cells were cultured in Dulbecco's Modified Eagle Medium (DMEM) (Sigma Aldrich, Cat #D6546) with 10% foetal bovine serum (FBS) (ThermoFisher Cat #A5256801) in a 37° C humidified incubator with 5% CO_2[424]. Passage 40 was used for the experiments.

Cell line generation

For each gene target, four gRNAs were commercially synthesised and cloned into the lentiviral vector pLentiCRISPRv2 (GenScript, UK). gRNA sequences are listed in Table S4.9. Lentivirus for individual gRNAs was produced. Briefly, CRISPR plasmids were cotransfected in HEK293T with packaging vectors pMD2.G (Addgene, 12259) and psPAX2 (Addgene, 12260) using FuGENE® HD Transfection Reagent (Promega Cat #E2311). Viral supernatant was collected at 48 hours and 72 hours post-transfection for each gRNA. The lentivirus from the four gRNAs were pooled. hWAT cells were transduced with the pooled virus and with 2 µg/ml polybrene. The selection was performed for 3 days in 1 µg/ml puromycin with media changes as required.

Differentiation of human white adipose tissue cells

HWAT adipocytes were derived from an established preadipocyte immortalised cell line[424] generated from human neck fat samples. The wild-type hWAT and single-gene knockdown (including the negative control cell line with the empty Cas9 vector) cells were cultured in DMEM containing 10% FBS. To differentiate into adipocytes via addition of differentiation mixture, the cells were induced with 0.5mM

3-isobutyl-1-methylxanthine (IBMX, Sigma Cat #5879) and 0.1μM dexamethasone (Sigma Cat #1756), 33μM Biotin (Sigma Cat # 4639), 17μM Pantothenate (Sigma Cat #5155), 0.5μM insulin (Sigma Cat #2643), 2nM 3,3',5-Triiodo-L-thyronine (T3, Sigma, Cat #6397), and 1μM Rosiglitazone (APE Cat #A4304) for 3 days. Next, culture medium was replaced to DMEM supplemented with 10% FBS and 33μM Biotin, 17μM Pantothenate, 0.5μM insulin, 2 nM 3,3',5-Triiodo-L-thyronine, and 1μM Rosiglitazone for 3 days. Medium was subsequently changed every 3 days for the following days. The cells were scrupulously maintained below 70–80% confluence at all assay stages to preserve their differentiation capacity.

Lipid staining with BODIPY

Undifferentiated and differentiated hWAT and single-gene knockdown cells were stained with 2.9 µM BODIPY 505/515 (ThermoFisher Cat #4639) and Hoechst (Hoechst 33342, Trihydrochloride, Trihydrate - 10 mg/mL solution in water, ThermoFisher Cat #H3570) for 15 min at 37° C in complete medium, washed with phosphate buffered saline. Cells were analysed for fluorescence intensity using an Opera Phenix™ High-Content Screening System (Revvity, UK) and Harmony® high-content analysis software v4.9 (Revvity, UK) provided by the Cellular High Throughput Screening Group (cmd.ox.ac.uk). For each knockdown condition, there were six replicates of undifferentiated and differentiated cells.

RNA sequencing of human white adipose tissue cells

I performed read alignment, transcript quantification and calculation of normalised transcript counts.

Wild-type hWAT and single-gene knockdown cells were seeded in six-well plates and collected after either (1) zero (undifferentiated), 3, 8 or 24 days of differentiation when assessing sufficient expression in wild-type hWAT or (2) 14 days of differentiation when performing RNA-seq for cell lines involved in the CRISPR knockdown experiment (wild-type hWAT, and single-gene knockdown cells, including Cas9-empty cells). Following collection, RNA was isolated using TRIzol (Thermo Fisher

Scientific, Carlsbad, CA, USA). Subsequently, RNA samples were purified using the Direct-zol™ RNA Miniprep protocol as per manufacturer's (Zymo Research, Cat # R2050) recommendations. Quantification of RNA was performed using the Nanodrop 1000 Spectrophotometer. All RNA samples were sequenced using Illumina 2x150bp paired-end sequencing. Read alignment was performed with STAR[425] v2.7.10b. Transcript quantification estimation was performed with RSEM[426] v1.3.2. Normalised transcript counts were calculated with DESeq2[409]v1.42.0.

Complementary DNA (cDNA) generation and quantitative RT-PCR

Data was generated by Ilknur Sur Erdem. Using the RT-qPCR data, I performed the knockdown confirmation analysis.

For RT-qPCR, total RNA was extracted for gene expression analysis using Direct-zol RNA Miniprep, (Cat #R2050, Zymo Research). After measurement of the RNA (Nanodrop, ThermoScientific, USA), the cDNA synthesis was performed using using High-Capacity cDNA Reverse Transcription Kit (Applied Biosystems™, Cat #4368814) according to the manufacturer's instructions. RT-qPCR was carried out with 1 µL of the 1:5 diluted reverse transcription sample with 10 µL of 2× SYBR Green Master Mix (iQ SYBR Green Supermix, Cat #1708882, Biorad) and 100 nM specific gene primer pairs in a 10-µL total volume in 96-well microtiter plates. The primer sequences for this study are listed in Table S4.11. LightCycler 96 (Roche, USA) was used to measure CT values of each sample. Each experiment was performed at least in triplicate. β-Actin mRNA was employed as an internal standard.

The expression level of each gene was determined by RT-qPCR and normalised against β-Actin mRNA level. Knockdown was confirmed by one-sided t-test (α=0.05) using the 2-ΔΔCT method[427] for amplification performed in separate wells, with the alternative hypothesis that knockdown decreased gene expression.

4.4.12 Differential expression analysis of CRISPR adipocyte RNA sequencing

Differential expression analysis was performed using DESeq2[409] v1.42.0 on RNA sequencing data from CRISPR adipocytes. Genes with raw count less than 10 were removed. The Cas9-empty cell line was used as the baseline condition. Each knockdown or wild-type hWAT was then separately compared to Cas9 empty.

4.4.13 Statistical analysis of lipid accumulation assays

Quantification of BODIPY fluorescence was performed by Ilknur Sur Erdem. I performed all downstream statistical analysis.

We quantify lipid accumulation using whole-well mean cytoplasm fluorescence of the BODIPY stain, as measured by the Opera Phenix™ High-Content Screening System. For each knockdown and the wild-type hWAT, we combined replicates with the baseline negative control, Cas9 empty, and regressed readout (mean cytoplasm fluorescence) against knockdown status:

$$readout \sim is_knockdown + 1$$

For wild-type hWAT replicates, *is_knockdown* is True to separate Cas9 empty and the wild-type replicates. After regression, we calculated Cook's distance[428] for each observation. Outliers were removed if Cook's distance exceeded the median of the F distribution with degrees of freedom n and $n - p$, where n is the number of observations included in the regression (lipid accumulation n=12) and p is the number of predictors, including the intercept (p=2)[429]. Only one observation was flagged as an outlier, for the *SLTM* knockdown: cytoplasm mean fold change relative to Cas9 empty=1.618, Cook's distance=0.865. We removed this outlier and reran the regression for *SLTM* knockdown lipid accumulation. We consider the knockdown to have a significant effect on the readout if P<0.05 for the *is_knockdown* term from the linear regression. Note that this is the equivalent of a two-sided t-test of mean readout between observations from Cas9 empty and a single-gene knockdown or the wild-type hWAT. In Table S4.13 we report the fold change of mean fluorescence

of each single-gene knockdown relative to the Cas9 empty cell line and the P-value from the two-sided t-test, after removing outliers identified using Cook's distance.

4.4.14 Pathway enrichment in CRISPR RNA sequencing data

We used GSEA[430] v4.3.3 to test for pathway enrichment in CRISPR-Cas9 knockdown adipocyte RNA sequencing data. Regularised log transformed mRNA counts calculated by DESeq2 are provided to GSEA. The pathway enrichment for each knockdown or wild-type hWAT is compared against the baseline control, the empty Cas9 vector, such that pathways with more positive normalised enrichment scores (NES) are those enriched in Cas9 empty and those with more negative NES are those enriched in the knockdown or hWAT.

We used pathway sets from KEGG Legacy, KEGG Medicus, REACTOME, GO:Biological Process, and GO:Molecular Function, downloaded from GSEA's Molecular Signatures Database v2023.2.Hs[430]. We always used 10,000 permutations when running GSEA.

4.4.15 Druggability evidence

Druggability evidence was collated by Duncan Palmer.

Druggability evidence was collated for each knockdown target gene using OpenTargets profiles[431], including both known drugs and drug tractability. Here we provide a summary of terms within the "Tractability" columns:

Small molecule; Structure with Ligand: Target has been co-crystallised with a small molecule[432], High-Quality Ligand: Target with ligand(s) (PFI $\leq$ 7, SMART hits $\leq$ 2, scaffolds $\geq$ 2)[433], High-Quality Pocket: Target has a DrugEBIlity[434] score of $\geq$ 0.7, Med-Quality Pocket: Target has a DrugEBIlity score between 0 and 0.7, Druggable Family: Target is considered druggable as per the Finan *et al.*[435] druggable genome pipeline.

Antibody; UniProt loc high conf: High confidence that the subcellular location of the target is either plasma membrane, extracellular region/matrix, or secretion[436],

GO CC (Gene Ontology Cellular Component) high conf: High confidence that the subcellular location of the target is either plasma membrane, extracellular region/matrix, or secretion[437], UniProt loc med conf: Medium confidence that the subcellular location of the target is either plasma membrane, extracellular region/matrix, or secretion[436], UniProt SigP or TMHMM (transmembrane hidden Markov model helix prediction)[438]: Target has a predicted signal peptide or trans-membrane regions, and not destined to organelles (source: Uniprot SigP, TMHMM[438]), GO CC med conf: Medium confidence that the subcellular location of the target is either plasma membrane, extracellular region/matrix, or secretion[437], Human Protein Atlas loc: High confidence that the target is located in the plasma membrane[439].

PROTAC; Literature: Target mentioned in a set of PROTAC-related publications curated by OpenTargets[431], UniProt Ubiquitination: Target tagged with the Uniprot keyword "Ubl conjugation [KW-0832]", which indicates that the protein has a ubiquitination site, based on evidence from the literature[436], Database Ubiquitination: Target has reported ubiquitination sites in PhosphoSitePlus[440], mUbiSiDa[441], or Kim *et al.*[442], Half-life Data: Target has available half-life data[443]) Small Molecule Binder: Target has a reported small-molecule ligand in ChEMBL with a measured activity of at least 10 µM in a target-based assay[433].

4.5 Discussion

Through a series of population-level genetic and *in vitro* functional genomics investigations, we demonstrate a model to map the biological mechanisms of obesity and body fat distribution, from variant to molecular function to systemic phenotype. We utilised complementary rare variant gene-based analyses and common variant prioritisation to nominate 69 genes with robust associations to BMI, WHRadjBMI, body fat percentage, and six tissue-derived fat components in up to 402,375 participants in UKB. By evaluating the sex-specific and phenome-wide associations of these genes, we built a comprehensive understanding of their systemic effects. Combining multiple lines of genomic, transcriptomic, and prior functional evidence,

we selected 14 genes for functional characterisation by CRISPR knockdown in hWAT cell lines. Given the difficulty of manipulating adipocytes in high-throughput experiments, we believe this is one of the first studies to comprehensively identify and target multiple genes for functional characterisation in adipogenesis. We observed reduced lipid accumulation in *PPARG* and *SLTM* knockdown, and increased lipid accumulation in knockdowns of *COL5A3*, *EXOC7* and *TRIP10*. While some of these genes have previously been identified through large-scale genetic analyses, we show their functional effects on adipogenesis for the first time, and highlight potential therapeutic avenues to regulate body fat mass and/or distribution. Taken together, our population-based and *in vitro* genetics investigations highlight molecular mechanisms of, and therapeutic avenues to alter, fat distribution.

The recent growth in sample sizes of exome-sequenced participants in biobanks has accelerated the development of gene-level analyses[93,222]. By annotating rare variants with their putative functional consequences and then collapsing these across each gene, gene-level testing has enabled the discovery of novel genes associated with complex human traits[199,200]. While gene-level testing is a powerful strategy for gene discovery, it remains limited by the requirement of high-quality functional annotations of rare variants. Here, we combined evidence from CADD[110], REVEL[109], and SpliceAI[111] annotations to generate high-confidence sets of variants, and integrated evidence from common variant studies to prioritise putative risk genes for functional follow up. Using these annotation categories, we built allelic series that showed the effect of each category of variants within a given gene. In assessing the monotonicity of these allelic series, we made the simplifying assumption that missense variants not predicted to be LoF or highly damaging would act in the same direction as pLoFs or damaging variants. However, missense variants may also result in gain of function[444]. Categorising missense variants into gain- and loss-of-function before assessing monotonicity could result in identifying more genes with monotonic allelic series. Furthermore, our analysis was also only focused on coding variation. In the future, improved rare variant masks, such as those that better

stratify missense variants or incorporate non-coding rare variation outside exonic regions, may help identify additional therapeutic targets to alter fat distribution.

Genome-wide common variant and exome-wide rare variant analyses can provide complementary lines of evidence to nominate genes associated with human traits[200,445]. We found that genes implicated for obesity and body fat distribution by a burden of rare damaging variation were: (1) associated with rare Mendelian forms of obesity and lipodystrophies, and (2) also prioritised by common variant prioritisation scores[382] for BMI, WHRadjBMI and body fat percentage. Converging evidence from across the allele frequency and phenotypic severity spectra can therefore identify high-confidence genes for a range of human traits. Assessing the sex-specific and phenome-wide consequences of these genes can provide evidence for therapeutic potential, for example by revealing possible side-effects on platelet distribution or neutrophil percentage.

We measured the effects of CRISPR gene knockdowns on lipid accumulation in hWAT cell lines using the BODIPY assay, which we performed in six technical replicates per knockdown cell line to ensure replicability. As the process of adipocyte differentiation is considered a necessary precursor to lipid accumulation[386], we interpreted the BODIPY readout as a measure of both lipid accumulation and adipogenesis more generally. However, we acknowledge that lipid accumulation is only one of many pathways involved in adipocyte behaviour, alongside glucose uptake, stem cell dynamics, adipocyte lifespan, depot-specific effects, etc. Thus, a failure to find alterations in lipid accumulation and adipogenesis due to gene knockdown *in vitro* does not imply that the gene is not relevant for obesity nor that it is an unattractive drug target. Genetic variation associated with obesity will likely affect other components of the adipose tissue niche, including innervation, immune cells, or stromal components in a depot-specific manner[446–448]. Indeed, BMI-associated genetic variation is known to alter the expression of genes in the central nervous system[361], and blockbuster GLP1-receptor agonists that are prescribed for weight loss[362,363] act primarily through the dorsomedial hypothalamus to control

food intake[364]. Therapeutic targets against obesity will thus need to span the range of mechanisms through which excess body fat may accumulate.

Our confirmed knockdown of *INSR* did not significantly affect lipid accumulation as measured by the BODIPY assay, despite this gene having a generally well-characterised role in body fat accumulation[449,450]. However, *INSR* was still partly expressed (RNA-seq expression FC [95% CI] = 0.521 [0.452, 0.601]), which may explain why the knockdown did not result in significant changes to lipid accumulation.

Of the two genes carrying rare missense variation associated with increased BMI and with confirmed knockdowns, *EXOC7* knockdown increased lipid accumulation in hWAT cell lines while *SLTM* knockdown decreased lipid accumulation. Rare missense variation in *PPARG* is associated with decreased body fat percentage and the disruption of *PPARG* in hWAT cell lines resulted in decreased lipid accumulation. Similarly, of the two genes affecting central obesity (WHRadjBMI or ratio of android to gynoid tissue fat percentage) with confirmed knockdowns significantly increasing lipid accumulation, *COL5A3* was correspondingly associated with increased central obesity, whereas *TRIP10* was associated with decreased central obesity. Our results demonstrate the importance of systematically characterising the functional effects of gene perturbation at the level of cell lines and model organisms, which may not follow the expected direction of effect based on missense variation in humans.

In summary, we demonstrate that the convergence of evidence across rare and common genetic variation can help identify high-confidence target genes for overall adiposity and body fat distribution. We provided in-depth functional characterisation through CRISPR knockdowns in human adipocytes, allowing us to nominate candidate genes for therapeutic targets. Through functional readouts and transcriptomic analyses, we also highlighted several novel molecular mechanisms by which genetic variation may impact adipogenesis. Our results provide a model by which future work can integrate genetic and functional evidence to identify, design, and validate potential drug targets to alter overall and tissue-specific fat distribution.

5

Conclusions and future directions

Contents

5.1 Conclusions 160
5.1.1 Overall summary 160
5.1.2 The genetics of extreme blood and urine biomarker levels 160
5.1.3 Individuals with phenotypes misaligned with polygenic prediction are enriched for rare pathogenic variation . . 162
5.1.4 Combining evidence from human genetic and functional screens to identify pathways altering obesity and fat distribution 164
5.2 Limitations and next steps 165
5.2.1 Overall limitations 165
5.2.2 The genetics of extreme blood and urine biomarker levels 168
5.2.3 Individuals with phenotypes misaligned with polygenic prediction are enriched for rare pathogenic variation . . 170
5.2.4 Combining evidence from human genetic and functional screens to identify pathways altering obesity and fat distribution 171
5.3 A vision for the future of genetics and health 172
5.3.1 More data 173
5.3.2 Better accessibility 174
5.4 Final remarks 177

5.1 Conclusions

5.1.1 Overall summary

In this thesis, I used exome sequencing data from the UK Biobank (UKB) to examine the effect of rare coding variants on deviation from normality across several contexts, in an effort to better understand disease. My work has implications across a wide range of fields, from clinical biomarkers for disease, to improving diagnosis of rare disease, to target discovery for common complex diseases such as obesity.

5.1.2 The genetics of extreme blood and urine biomarker levels

In Chapter 2 I studied 32 blood and urine biomarkers in UKB. Given that abnormal biomarker levels are typically linked with disease, I focused on individuals in the tails of biomarker distributions. Although there have been several initial papers published on UKB biomarkers[85,143–145], my work marks the first comprehensive study of biomarker extremes in UKB.

Many individuals in UKB have biomarker levels outside of reference ranges

I collated reference ranges for each biomarker, which I used to define biomarker values as normal and abnormal (or "outlying"). I quantified the proportion of individuals in UKB with biomarker values outside of references. For several biomarkers (LDL-C, total cholesterol, and urine creatinine), a majority of individuals had values outside the reference range. Nearly all individuals (96.4%) had abnormal values for at least one biomarker. This raises an interesting philosophical question of what it means to be normal and healthy, providing further validation that a "multiple univariate" approach to measuring health (i.e. evaluating multiple single biomarkers independently) is inadequate[148].

Disease odds increase with increasing number of biomarkers outside of reference ranges

I found that individuals with many abnormal biomarkers were enriched for disease, with higher odds ratios in complex diseases that are correlated with multiple biomarkers. This is further evidence that complex diseases affect multiple pathways and supports the use of a high dimensional multivariate biomarker distribution when stratifying individuals by disease risk.

Individuals with extreme biomarkers are enriched for rare genetic disease

Biomarkers can be valuable for rare disease screening[149]. To test this in UKB, I determined whether individuals with extreme biomarker levels were more likely to carry variants expected to cause rare genetic disorders. I found significant enrichment in some biomarker-disorder pairs, mostly among blood lipid biomarkers. I also found interesting sex-specific results, including enrichment of monogenic diabetes variants among females with high HbA1c, implying that HbA1c is more informative in women when screening for monogenic diabetes.

Dichotomising continuous biomarkers can uniquely power discovery of genetic associations

Next, I focussed on the discovery of novel genetic associations with extreme biomarker levels. I dichotomised continuous biomarkers into extreme and non-extreme "tail-status" phenotypes and performed an exome-wide scan for associated variants, hoping to find variants uniquely associated with tail-status and not the continuous version of the biomarker. As expected, statistical power was depleted in the tail-status approach compared to using continuous phenotypes. However, there were some instances where the tail-status approach outperformed continuous phenotypes, such as detecting a missense variant in *PNPLA3* known to increase hepatic fat. This suggests that although the tail-status method is unlikely to replace testing continuous biomarkers, it may be useful as a secondary analysis.

Rare variant gene burden associated with multivariate biomarker extremes

Given my evidence that individuals with higher numbers of abnormal biomarkers are enriched for disease, I sought to identify genes which affect multiple biomarkers and thus may be relevant to disease. I quantified the deviation of each individual in the multivariate distribution of biomarkers using Mahalanobis distance as a metric, and performed an exome-wide scan for deleterious gene burden associated with Mahalanobis distance. The majority of resulting associated genes were also significantly associated with multiple single biomarkers, validating this approach for identifying genes which affect multiple biomarkers.

5.1.3 Individuals with phenotypes misaligned with polygenic prediction are enriched for rare pathogenic variation

In Chapter 3, I examined phenotypic deviation from expected phenotype, as estimated with common-variant PGS, as a means for prioritising individuals for rare variant screening. I built on work by Hawkes *et al.*[279] and Lu *et al.*[277], developing and applying methods for testing enrichment of rare pathogenic or protective variant burden in individuals who misalign with their expected phenotypes.

Individuals with misaligned continuous phenotypes are enriched for pLoF and damaging missense burden

I extended an existing misalignment classification method developed by Hawkes *et al.*[279] to additional phenotypes and, for the first time, tested whether misalignment could be explained by missense variants predicted to be less deleterious than fully LoF variants. I replicated the pLoF enrichment observed by Hawkes *et al.* for height and LDL-C, and observed significant pLoF enrichment in individuals with misaligned BMD, a trait not studied by Hawkes *et al.*. I demonstrated that misalignment classification can be a useful prioritisation tool when screening for rare genetic disorders related to height, LDL-C, and BMD.

Individuals with misaligned dichotomous phenotypes are enriched for pLoF burden

Hypothesising that misalignment classification could be extended to dichotomous traits and building on work by Lu *et al.*[277], I developed a framework for testing enrichment of pathogenic and protective variants in individuals whose polygenic risk scores deviate from their disease status. I demonstrated that disease cases carrying pathogenic rare pLoFs tended to have lower PRS, and, for the first time, that disease controls carrying protective rare pLoFs tended to have higher PRS. These results provide strong statistical evidence for the prioritisation of low-PRS disease cases in rare variant screening. Furthermore, these results support a liability threshold model for disease where disease liability is defined by the balance of rare and common variant risk/protection. This framework could also be used as further validation of pathogenic or protective variants identified through traditional association studies.

Individuals with misaligned phenotypes are nominally enriched for deleterious singleton burden

In search of other possible genetic factors influencing phenotype misalignment, I tested whether singleton burden across non-disease-specific LoF-intolerant genes was associated with phenotype misalignment, for both continuous and dichotomous traits. The results suggested that deleterious singleton variant burden might contribute to lower than expected stature and be pathogenic for T2D and CAD. This provided another potential genetic explanation for phenotype misalignment, in addition to previous gene-level tests of pLoF and damaging missense burden.

Exome-wide scan for pathogenic and protective misaligned effects nominates potential novel protective gene for age at menopause

Lastly, I turned from gene validation to discovery and tested exome-wide for rare variant burden associated with misaligned status. I identified a novel association between the burden of rare damaging missense variation in *KANK1* and later AAM. This finding demonstrates that the use of misaligned status can increase association signals in gene discovery, particularly for self-reported phenotypes (e.g.

AAM) which may be noisier than well-measured phenotypes[348]. I did not find compelling evidence for protective pLoF gene burden associated with misaligned disease status, which suggests that testing on misaligned status of continuous phenotypes may be the better-powered approach.

5.1.4 Combining evidence from human genetic and functional screens to identify pathways altering obesity and fat distribution

In Chapter 4, I examined the genetics of fat accumulation (adiposity), a biological mechanism in which deviation from normality can cause obesity and lipodystrophy. I led the largest-ever systematic evaluation of adiposity genes in human fat cells. We nominated *SLTM* as a novel adiposity gene and potential therapeutic target for obesity.

Discovery of genes associated with obesity and fat distribution through exome-wide association tests

Using gene-level association tests conditioned on nearby common variant signals, I scanned the exome for associations between damaging classes of rare variation and nine traits related to obesity and fat distribution. I identified 19 exome-wide significant genes ($P < 1.58 \times 10^{-7}$) and 50 additional genes which passed an FDR-controlled significance threshold ($P \leq 4.37 \times 10^{-5}$). These genes overlapped with those highlighted by common variant associations, demonstrated by a significant enrichment of PoPS[382]. This demonstrated a convergence of rare and common variant effects on fat accumulation.

Multiple lines of evidence used to select genes for functional characterisation

To prioritise a subset of genes for functional characterisation, I used evidence from gene expression data, allelic series of ultrarare variants, longitudinal analysis and publications of knockdown/knockout experiments. Genes were selected for functional characterisation if they were sufficiently expressed in wild type human adipocytes and demonstrated monotonicity in allelic series, had missense burden

associated with earlier onset of obesity, or had been shown to affect lipid metabolism pathways in gene perturbation experiments. We selected 14 genes for our knockdown experiments in human fat cells (adipocytes).

Effect of single-gene knockdown on adipogenesis and lipid accumulation in human adipocytes

Using CRISPR-Cas9, single-gene knockdown was performed in human adipocytes for each of the 14 genes selected for functional characterisation. We measured adipogenesis and lipid accumulation using BODIPY fluorescent dye to indicate the presence of lipids in adipocytes. Five genes demonstrated significant changes in lipid accumulation: two which resulted in decreased lipid accumulation (*PPARG*, *SLTM*) and three which resulted in increased lipid accumulation (*COL5A3*, *EXOC7*, *TRIP10*). Inactivation of *PPARG* through knockout has been shown to lower lipid accumulation in human and mouse pre-adipocytes[374,388,389], providing validation of our assay. Of the remaining genes, *SLTM* is the most exciting: its gene-level association with BMI has been replicated in AoU, an independent dataset to UKB. The evidence suggests that *SLTM* is a novel adiposity gene and deserves further evaluation as a therapeutic target for obesity.

5.2 Limitations and next steps

5.2.1 Overall limitations

At the intersection of my three thesis projects was the genomic dataset used (UKB ES data subset to Europeans), a focus on protein-coding variants, variant consequence annotation, and variant/gene-level association testing. Here I identify aspects which would benefit or expand on each of the shared components.

Replication in an independent cohort with diverse ancestry

It would have benefitted all of my projects if I had replicated analyses in an independent cohort. If I had to choose a replication cohort it would be the AoU dataset[115]. Similar to UKB, AoU is deeply phenotyped, with many of the same

phenotypes available, and also contains sequencing data for over 400,000 individuals at the time of writing. Although there are fewer individuals of European ancestry in AoU, which would diminish the statistical power of an ancestry-matched European-only replication study, subsets of AoU represent the largest sample sizes for the underserved African American and Latinx populations[451]. Replicating in AoU for multiple ancestry groups could demonstrate the generalisability of my results.

Expanding analysis to non-coding regions

My analysis primarily was primarily focussed on the coding regions of the genome. With the arrival of large-scale WGS data in UKB[118,120] comes the opportunity for exploring non-coding regions. Indeed, several recent publications have described associations between the non-coding genome[118–120] and traits of interest in the UKB. That said, while there is no doubt that non-coding regions harbour valuable signal, such as the discovery of variants in a region of non-coding RNA *RNU4-2* being able to explain 0.4% of individuals with neurodevelopmental disorder[452], some researchers question whether the discovery yield with WGS compared to imputed data or ES data is worth the substantial costs to sequence, process, and analyse WGS data[453].

Advancing methods for functional annotations

An important aspect of coding (and non-coding) variant analysis is annotation of consequence. In this thesis I use variant-level functional annotations predicted *in silico*. Computational predictors of variant consequence have become a common approach when annotating next-generation sequencing data[108], because it is generally impractical to manually curate variants using evidence from association tests or mutagenesis experiments. Manual curation of variant consequence using clinical evidence could make some of my results more clinically translatable, especially when testing for enrichment of variants which cause rare disorders. However, computational prediction of variant consequence is an area of constant development and expected to continue to improve. Recent methods have used machine learning, especially deep learning[111,454,455]. As the genetic datasets which inform these

models continue to grow and statistical methods develop, manual curation may not be necessary.

Improved methods for association testing

Performing association tests on genomic datasets typically requires specialised software packages which are optimised to have maximal statistical power (even in the presence of sparse data, such as rare variants or low prevalence dichotomous traits) and be computationally tractable for terabyte-scale high-dimensional data. I used SAIGE[103] and REGENIE[93] for variant- and gene-level association tests, which are both designed for computational efficiency and controlled type I error rates when testing rare variants in unbalanced case-control traits. Association testing continues to be an area of development, with focuses on improving statistical power and computational efficiency. Some notable recent examples yet to be broadly adopted include Quickdraws[456], which achieves a boost in efficiency by using graphical processing units, and gVAMP[457], which seems to achieve the holy grail of association testing, jointly modelling all variants, while remaining computationally tractable at biobank scale.

Consistency in definition of rare variants when performing burden tests

Throughout the thesis chapters I use two different maximum MAF thresholds (0.1% and 1%) when defining the rare variants included in burden tests. The choice of MAF threshold was systematic and depended on the type of enrichment test being performed. For all misaligned phenotype testing, I used the threshold of MAF$<$ 0.1% to match Hawkes *et al.*[279]. I used the same Fisher's exact test framework when testing individuals with outlying biomarker levels for enrichment of pLoF variants in known rare disorder genes - therefore, I used MAF$<$0.1% to define rare variants. For all exploratory gene burden testing in SAIGE and REGENIE, I used a threshold of MAF$<$ 1%, in line with several recent large-scale gene burden studies[199,369,370,418]. In future work, for consistency across studies it may be preferable to use a single MAF threshold to define "rare" variants.

5.2.2 The genetics of extreme blood and urine biomarker levels

Reference ranges are imperfect and can be improved

I used reference ranges defined by the NHS. Ideally, we would have ranges that are perfectly tailored to an individual's unique biological profile, but this is not the case. The UKB would be an ideal dataset to estimate reference intervals which are sex-specific, age-specific and matched to the disease profile of an individual. To the best of my knowledge, this has not been done. An online reference interval browser parametrised by sex, age and disease could be a high-impact resource for academics and clinicians.

Broaden scan for diseases associated with higher number of abnormal biomarkers

When testing whether individuals with high numbers of abnormal biomarkers were enriched for disease, I restricted my analysis to only the ICD-10 chapters and blocks within those chapters which had the highest counts in the UKB Data Showcase. A broader association scan could identify more diseases associated with widespread biomarker disruption. This would further support the use of multiple biomarkers to inform clinical decisions in complex disease and could prioritise which diseases are best suited to this multivariate approach.

Scan for novel biomarkers which can prioritise individuals for rare disorder screening

When testing whether individuals with abnormal biomarkers were enriched for variants which cause rare disorders, I tested only the biomarker-disorder pairs with the strongest evidence in the literature. I could expand to lower confidence biomarker-disorder pairs, or take an all-by-all approach and test every biomarker against every disorder.

Expand tail-status association tests

When defining tail-status phenotypes, I used percentile thresholds to standardise the power lost by dichotomisation. I could test whether dichotomising with reference range limits improves power. However, association tests on tail-status defined by reference ranges may be similar to association tests on disease status, diminishing the novelty of such an analysis.

I performed the tail-status association study only for blood and urine biomarkers, given that extreme biomarker levels are expected to be associated with disease. However, this method could naturally be applied to other biomarkers, such as those used in the Chapter 3 (e.g. BMD and IOP).

Refine gene burden test on Mahalanobis distance

To achieve a fully-defined dataset of biomarkers, as required to construct Mahalanobis distance, I had to subset to individuals with all selected biomarkers defined. This limited my sample size and statistical power compared to testing single biomarkers. Ideally all individuals would have defined biomarker values. If that is not a realistic expectation, another option would be to impute missing values. However, ideally the imputation method would not be directly trained on the covariance structure in UKB biomarkers because it could exaggerate the Mahalanobis distance, which is also informed by the covariance.

Another limitation was that I used all biomarkers when calculating Mahalanobis distance. While this phenotype-agnostic approach is desirable for its simplicity, it could be refined by calculating Mahalanobis distance on subsets of biomarkers, grouped by their relevance to specific diseases. This approach could be better powered to identify genes with subtle effects across multiple biomarkers than standard single-biomarker association tests. Furthermore, by using sets of fewer biomarkers there would be more individuals with fully defined biomarker values in each set, which also improves the power.

5.2.3 Individuals with phenotypes misaligned with polygenic prediction are enriched for rare pathogenic variation

More samples are needed to improve power

Testing for enrichment of rare damaging variants among misaligned individuals has limited statistical power due to very few or no variant carriers among misaligned individuals. With larger datasets, this issue might be less likely to occur. Alternatively, these tests might be better powered in a dataset specifically ascertained for individuals with rare disease (e.g. GEL 100,000 Genomes Project[458]), who are expected to carry more of the pathogenic variants I am testing.

Some phenotypes had much lower sample size than others, which could explain why I did not observe as many significant results in those phenotypes. Higher sample sizes would improve power for those phenotypes and would enable a fair comparison of significant results across phenotypes.

Incorporate clinical evidence in curation of pathogenic variants and individuals with rare disorders

For practical reasons, I used variant annotations predicted by computational methods. Using clinical evidence to manually curate pathogenic and protective variants could refine the results and improve the therapeutic translatability of this work. There may even be specific variants which are more likely to be enriched in misaligned individuals, which would provide further granularity into the genetic causes of misalignment.

I assume that individuals have a rare disorder if they carry deleterious variants in genes which cause the rare disorder. Another approach could be to use ICD-10 codes to determine whether an individual has been diagnosed with that rare disorder. There may be instances where an individual carries the variants expected to cause the disease but has not been diagnosed, or has been diagnosed but does not carry the expected variants. In the first case, it would be interesting to see whether these undiagnosed individuals are classified as misaligned because it would provide direct evidence of how misaligned classification can improve prioritisation of individuals for

rare disorder diagnosis. In the second case, we could examine whether misaligned classification could help prevent the misdiagnosis of individuals.

Expand misaligned disease status association tests

I restricted analysis of misaligned disease status to pLoFs and in the discovery analysis I only scanned for protective effects. To extend this work, I could test damaging missense variants, as was done for the continuous traits. I could also scan for novel pathogenic effects, which, based on the higher odds ratios in the tests of canonical genes, may be better powered than the scan for protective LoF effects.

5.2.4 Combining evidence from human genetic and functional screens to identify pathways altering obesity and fat distribution

More samples needed for DXA and MRI phenotypes

The DXA and MRI phenotypes had sample sizes about ten times smaller than the three main phenotypes (BMI, WHRadjBMI, body fat percentage), leading to lower power. It would be ideal to have higher sample sizes for the DXA/MRI traits, as they can provide more granular insights into depot-specific fat accumulation.

Use COAST to evaluate allelic series

During the course of this project, the Coding-Variant Allelic-Series Test (COAST) was developed by McCaw *et al.*[459] to perform gene-based tests based on a monotonic allelic series prior. A recent update[460] has adapted this method to take summary statistics as input. We could use this as an alternative method to quantitatively assess evidence of monotonicity in allelic series. In future work, we could use COAST as our primary association test, rather than using SAIGE.

Expand functional characterisation to other cellular phenotypes

In our work, we used one assay to test lipid accumulation and adipogenesis in human adipocytes. We could test other cellular phenotypes relevant to adiposity, such as the number of lipid droplets in adipocytes or glucose uptake mediated by insulin[461].

Expand functional characterisation to whole-organism gene perturbation

We tested the effect of single-gene perturbation at the cellular level. This is a relatively simple model to observe effects of gene perturbation. The next step would be to observe effects *in vivo* at the organism level, such as using tissue-specific or whole-body knockdown/knockout in mice. This would demonstrate the phenome-wide effects of a single-gene perturbation and could inform whether a gene is a viable drug target.

Global CRISPR screen

For practical reasons, I had to choose a subset of genes to functionally characterise. Given that there is no standard approach to selecting genes for functional characterisation, the selection procedure can be biased by choice of criteria. It would be ideal to perturb every single gene in the genome through a global CRISPR screen[462] as it would eliminate the need for establishing criteria. However, it would also introduce the question of how to weight rare variant association evidence with global CRISPR screen evidence.

Precise variant-level perturbation

I assumed allelic heterogeneity of damaging variants in a gene, both when conducting gene-level association tests and when perturbing the genes with CRISPR. In reality, there may be slightly different effects of damaging variants in a gene. Variant-level perturbation would therefore be ideal, replicating the exact missense variants observed in UKB.

5.3 A vision for the future of genetics and health

My ultimate aim as a statistical geneticist is to improve public health. Here is my broad vision for how I believe the field of genetics will evolve and how it is best placed to benefit worldwide public health.

5.3.1 More data

Health-related datasets will continue to grow in size.

It will soon be commonplace for biobanks to number in the millions of participants. I predict that the dataset which will succeed UKB as the gold standard for biobanks worldwide will be Our Future Health (OFH)[117]. Although UKB has dominated the field of genetics for the past decade, largely due to its deep phenotyping, high coverage of genetic data, and relative ease of accessibility, its major shortcoming is its lack of genetic diversity and restricted age range. In comparison, OFH is predicted to be the largest multi-ethnic cohort in the world[117] and will include adult participants of all ages. The tenfold increase in sample size, from half a million in UKB to five million in OFH, will also enable more granular studies that could lead to improvements in personalised medicine. Furthermore, compared to biobanks in other countries, OFH has the advantage of pre-existing infrastructure from UKB, which may aid its success.

The number of whole genomes sequenced will continue to increase, with sequencing costs per genome becoming much cheaper in recent years (Figure 5.1). Assuming the trend in cost continues, we would expect to see a $100 genome by 2030. However, unless there is a paradigm shift in the hardware and software involved in storage, processing and analysis, we might find that imputed and ES data are sufficient for most association studies. For instance, a recent study by Gaynor *et al.*[453] suggested that to maximise the number of genetic association discoveries, larger datasets of imputed and ES data should be prioritised over WGS data.

Meta-analysis across datasets can contribute to discovery of genetic associations that are not powered for discovery in a single dataset[465]. As more datasets become available, consortium-led federated approaches should be taken to standardise quality control and analysis procedures across datasets. This will help to reduce confounding heterogeneity between datasets. Nascent consortia like the Global Biobank Meta-analysis Initiative[466] and the Biobank Rare Variant Analysis consortium[467] will continue to grow in importance.

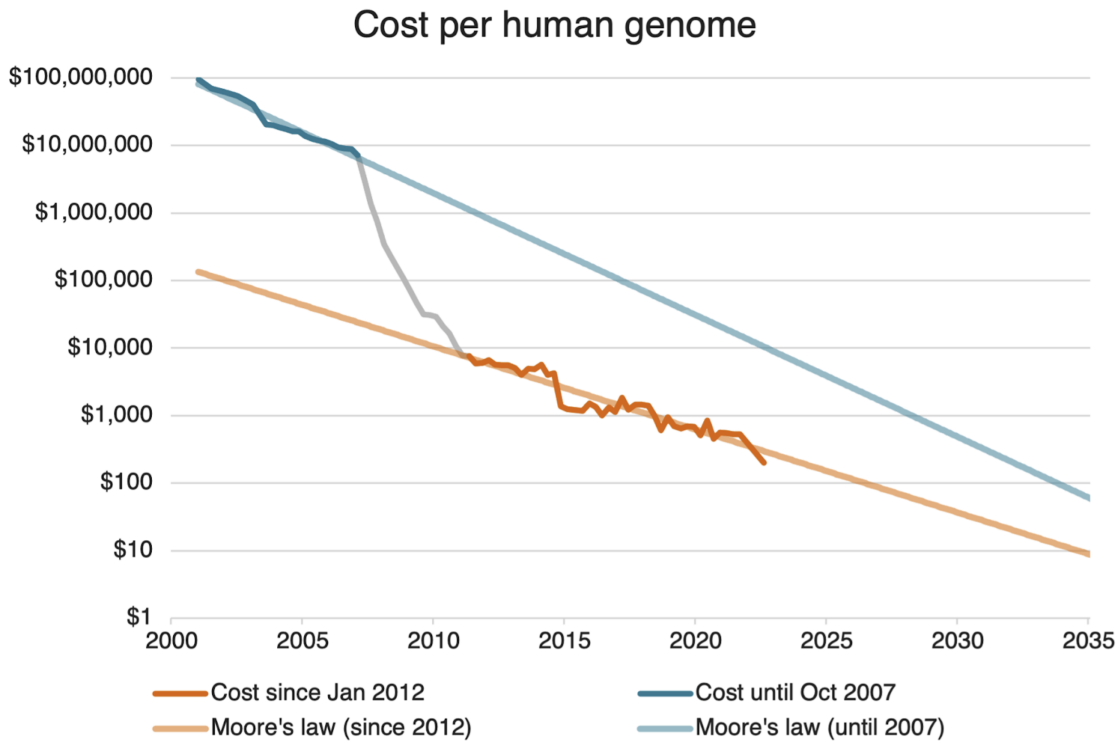

Figure 5.1: Expected cost per human genome. Based on "production" costs provided by the National Human Genome Research Institute until May 2023[121] and the $200 per genome cost estimated by Illumina in April 2023[463]. Costs are not adjusted for inflation. Moore's law is fit with exponential models on data points approximately before (until October 2007) and after (since January 2012) next-generation WGS was introduced in 2008[464].

5.3.2 Better accessibility

Analysis platforms

Data is only useful to researchers if it can be accessed and analysed with ease. The future of genomic analyses should be on the cloud, as it enables scalable analysis and equitable access across institutions. Given the sensitivity of health-related data, it is unlikely that biobanks will allow full freedom to analyse datasets on generic cloud platforms like Google Cloud Platform or Amazon Web Services. Biobanks like UKB and AoU are leading the charge with their bespoke cloud-based analysis platforms, which I used for my analysis of UKB. However, I believe more effort is needed to make these bespoke platforms as functional and efficient as generic commercial platforms. This can be as "simple" as improving the documentation, or

as complex as optimising tools like Hail[252] to run natively on the platform.

Browsers for results

Improving accessibility to results is as important as access to input data. I am grateful to researchers who make their summary statistics highly accessible (e.g. gnomAD[113], Genebass[199], AoU[115]) or aggregate data across sources (e.g. GWAS Catalog[468], Open Targets[431]), and make the data available as online resources. Key features which make these resources successful are that they are intuitively designed, have low-latency browsing, include data visualisation (except GWAS Catalog), and allow users to easily download data in delimited table form. This should be the standard for any publication of large datasets. One idea to encourage the creation of accessible resources could be for publishing companies to fund browser engineers who would work with authors to make their data accessible. This could involve making a single browser for each publication or harmonising data before funnelling into an aggregated dataset. Aside from the positive impact on open science, it may even be in the financial interest of publishing companies to host good browsers as it could drive more traffic to their publications or provide an edge over competitors.

Genetics for disease target discovery

As of 2023, there have been at least 40 germline genetic observations which have directly led to approved therapies across 36 rare and 4 common conditions[469]. Drugs with genetic support are 2.6 times more likely to succeed[470]. As genomic data grows, opportunities for genetic discovery of therapeutic targets will increase. Even if large association studies of case-control status for complex disease reach saturation in heritability explained (like that observed for association studies of height in populations of European ancestry[471]), there will always be opportunities to better characterise disease in subsets of the populations (e.g. stratifying by sex, age, ancestry groups or clinical subtypes).

Genetics for personalised medicine

There are many applications for genomics in personalised medicine[472]. Below I focus on two areas which I believe to be most relevant to my work: diagnosis of rare genetic disease and PRS.

Diagnosis of rare genetic disease

The use of sequencing data has been shown to yield diagnostic rates of 20% or more[458,473–475], a rate higher than most diagnostic tools used in medicine[476]. Experts predict that by 2030 the clinical significance of practically all observed genetic variants will be known[477]. In the next ten years I expect that for most high-income countries sequencing will become a first-line genetic test for children with undiagnosed rare disease.

Polygenic risk scores

For common diseases, PRS have shown promise in risk prediction[275]. Accurate prediction can prioritise individuals with high risk who could benefit from preventative intervention. For example, individuals classified as having high genetic risk for coronary heart disease had a 44% reduction in relative risk following statin therapy, while other individuals had a 24% reduction[274]. The cost effectiveness of PRS is still under evaluation, with mixed results for cancer and CVD screening[478,479]. Until costs are significantly reduced, I predict that PRS are unlikely to be adopted widely by public healthcare providers, but will be touted loudly by private personalised medicine healthcare providers.

Equitability and pragmatism in worldwide public health

The future of genetics and its clinical applications is exciting. But as I have alluded, due to high cost many of the recent advances in our field are more likely to benefit individuals with high socioeconomic status (SES) in high-income countries than other individuals around the world. This is symptomatic of worldwide health inequity[480]. Part of this inequity means that in many countries there

are substantial improvements needed in basic healthcare services before genomic medicine is considered. For example, a recent study estimated that in low and middle-income countries (LMIC) only 8% of individuals eligible for statin therapy for primary prevention of CVD used statins, compared to 50% in high-income countries[481]. For LMIC, the costs associated with PRS would probably be better spent on improving accessibility to healthcare.

So what can we, as statistical geneticists, do to improve public health? Most likely, we will continue to make scientific contributions to common and rare diseases, even though deeply ingrained societal structures mean that those contributions are most likely to benefit individuals with high SES in high-income countries. However, if given the choice, there are ways we can improve the equitability of our research. We can focus on diseases with a high public health burden in LMIC, make existing therapies cheaper and more accessible, and actively contribute to ongoing efforts[482] to create new genetic datasets and increase research capabilities in LMIC.

5.4 Final remarks

The study of genetics has improved understanding of common and rare diseases and improved the lives of countless individuals. With an explosion of data and constant development of new statistical methods, this is a particularly exciting time to be a statistical geneticist. However, the future of statistical genetics is not just in advancing methods and discovering new genetic associations, but in ensuring that ultimately those advances contribute to a healthier, more equitable world.

S2
Supplement to Chapter 2

S2.1 Supplementary figures

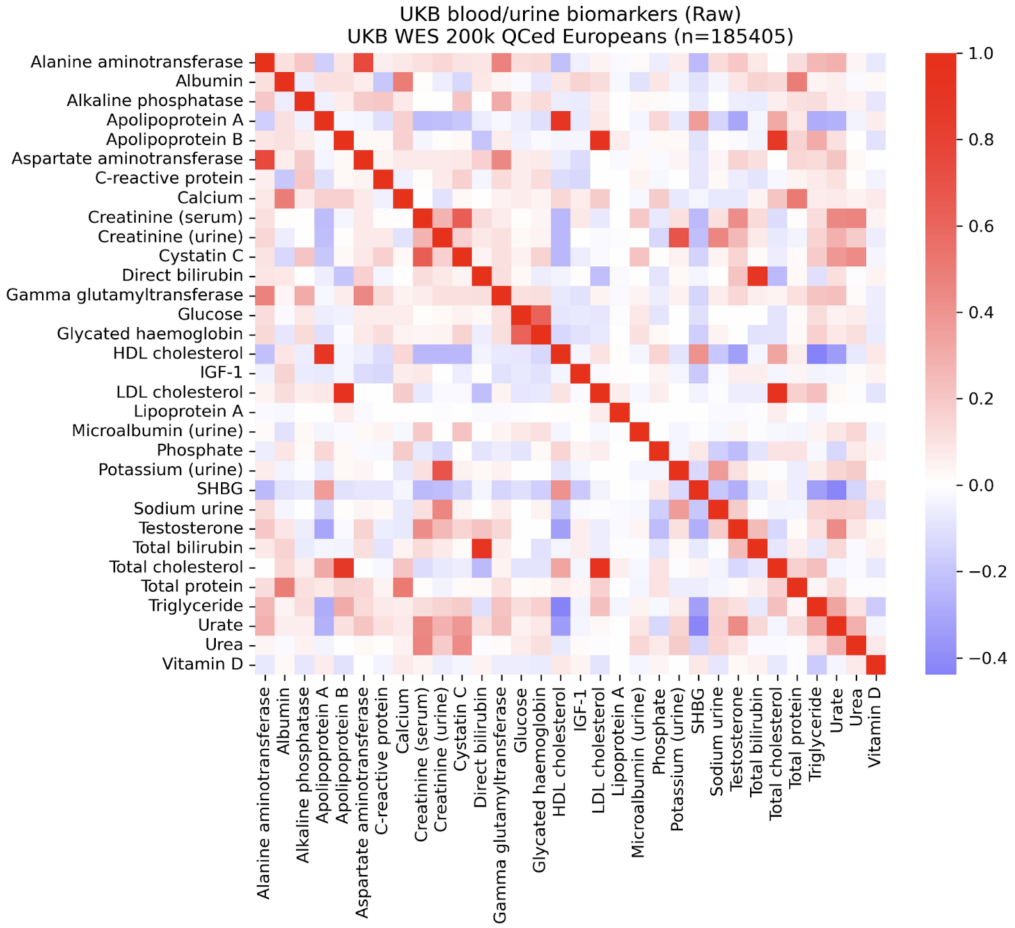

Figure S2.1: Heatmap of raw biomarker Pearson correlations. Only individuals who passed the ES 200k tranche QC and have European ancestry were included. Biomarkers are sorted in alphabetical order.

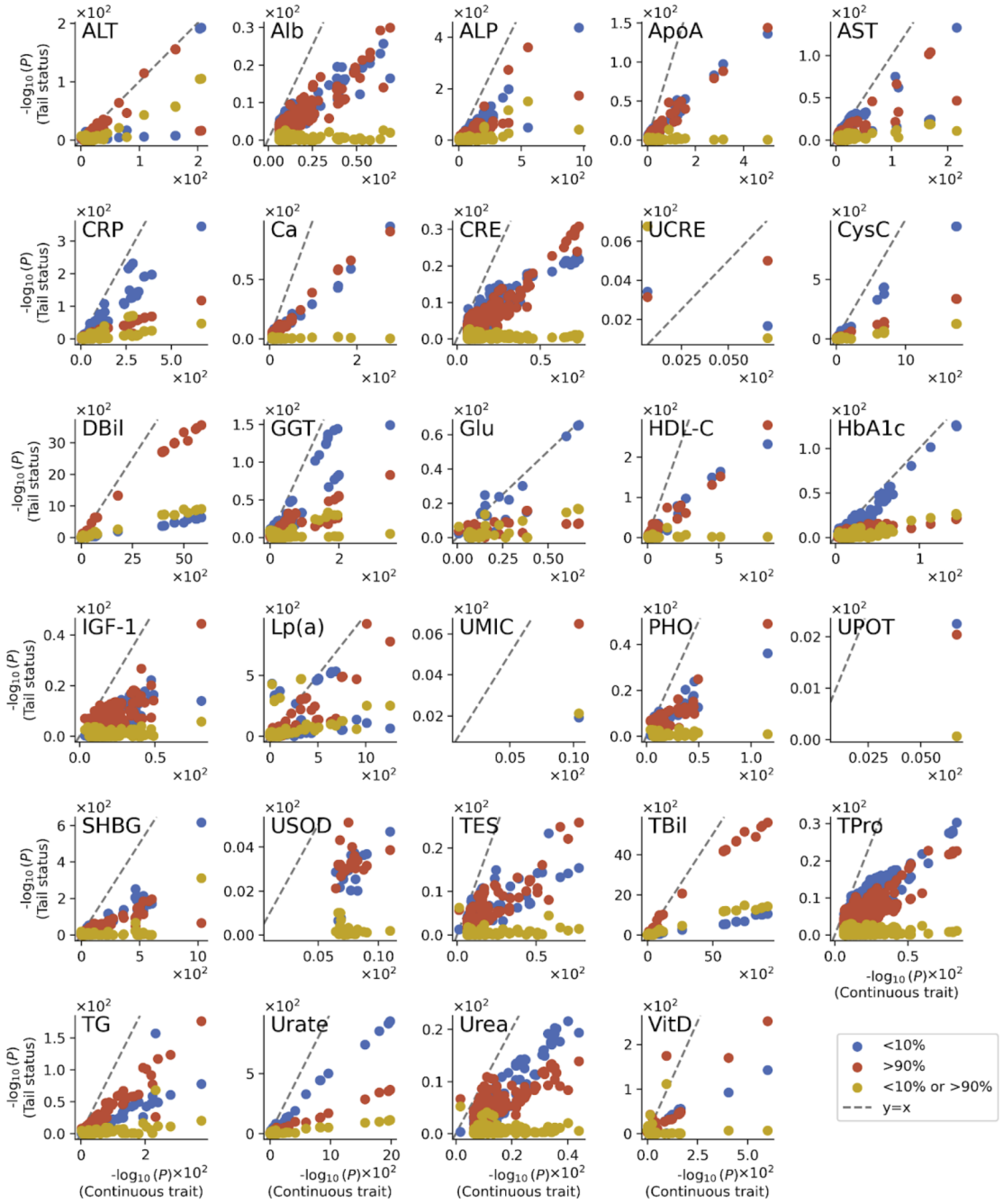

Figure S2.2: Relative power of association for continuous trait versus tail status phenotype. Only variants with which passed exome-wide significance ($P < 5 \times 10^{-7}$ for association with the continuous trait or any tail status are plotted.

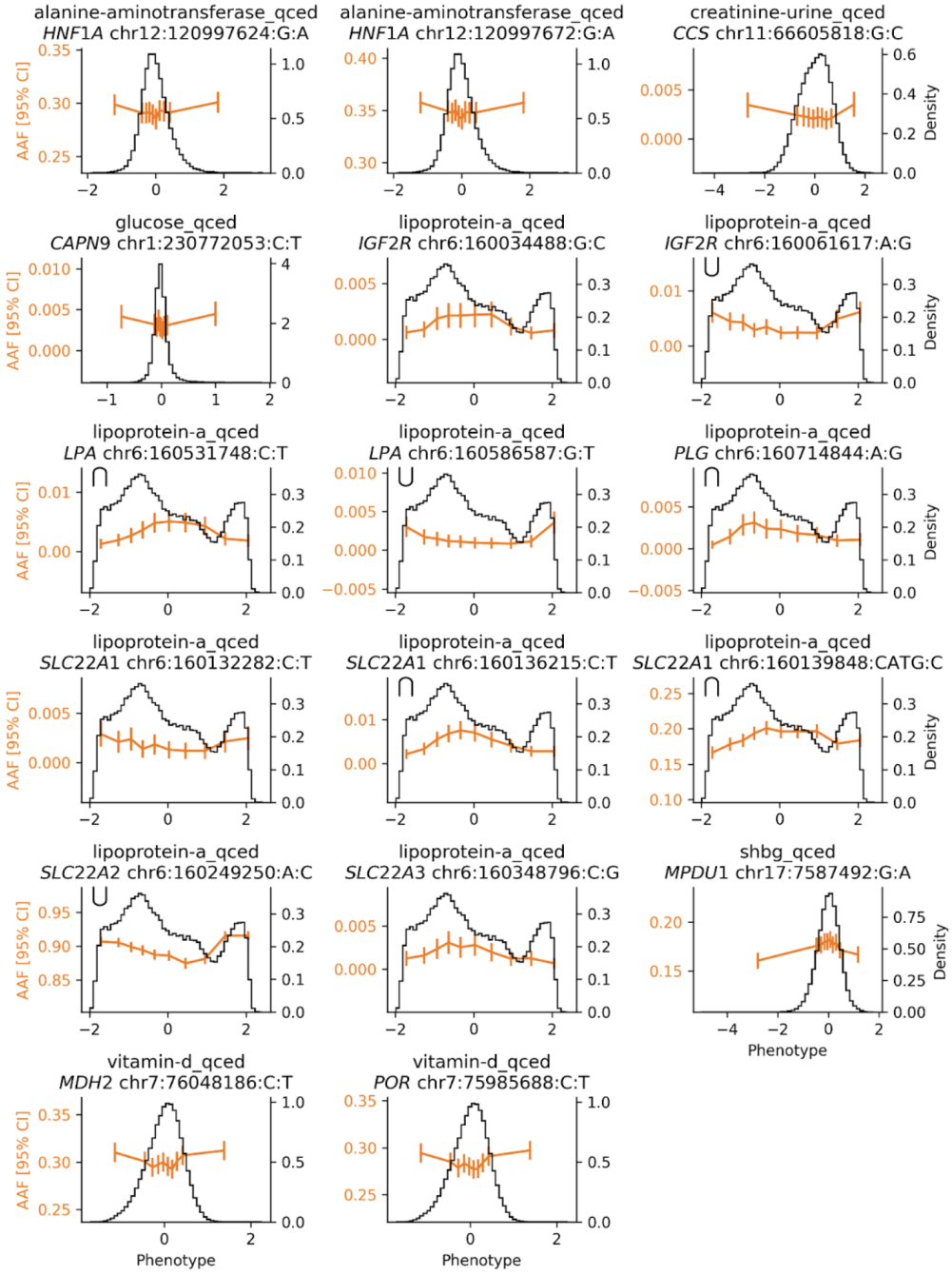

Figure S2.3: Allele frequency across phenotype distribution for coding non-synonymous either-tail VOI. Orange: Mean alternate allele frequency, with 95% CI calculated as $\pm 1.96 \times$ SEM. Black: Histogram of QCed biomarker distribution. Non-monotonic allele frequency distributions are highlighted with the $\cup$ or $\cap$ in the upper left of the plot, depending on whether the non-monotonicity was concave up or concave down, respectively. Non-monotonicity was defined as having at least one CI in the middle that was less than (or greater than) and non-overlapping with any of the CIs to the left or right. VOI, variant-of-interest (see Methods). CI, confidence interval. SEM, standard error of mean.

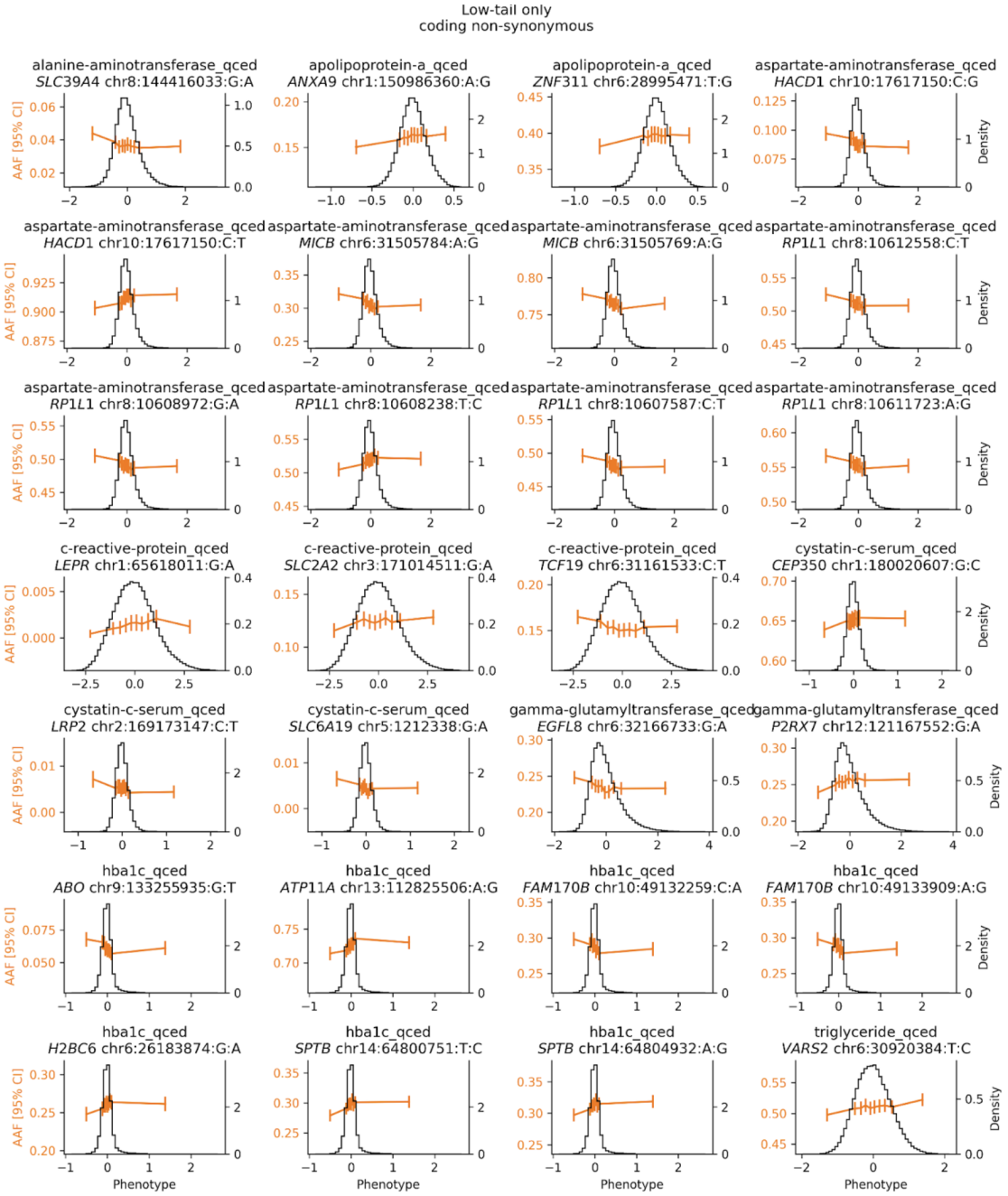

Figure S2.4: Allele frequency across phenotype distribution for coding non-synonymous low-tail-only VOI. Orange: Mean alternate allele frequency, with 95% CI calculated as $\pm 1.96 \times$ SEM. Black: Histogram of QCed biomarker distribution. Non-monotonic allele frequency distributions are highlighted with the $\cup$ or $\cap$ in the upper left of the plots. Non-monotonicity was defined as having at least one CI in the middle that was less than (or greater than) and non-overlapping with any of the CIs to the left or right. This plot only includes variants which were significant for low-tail status alone. VOI, variant-of-interest (see Methods). CI, confidence interval. SEM, standard error of mean.

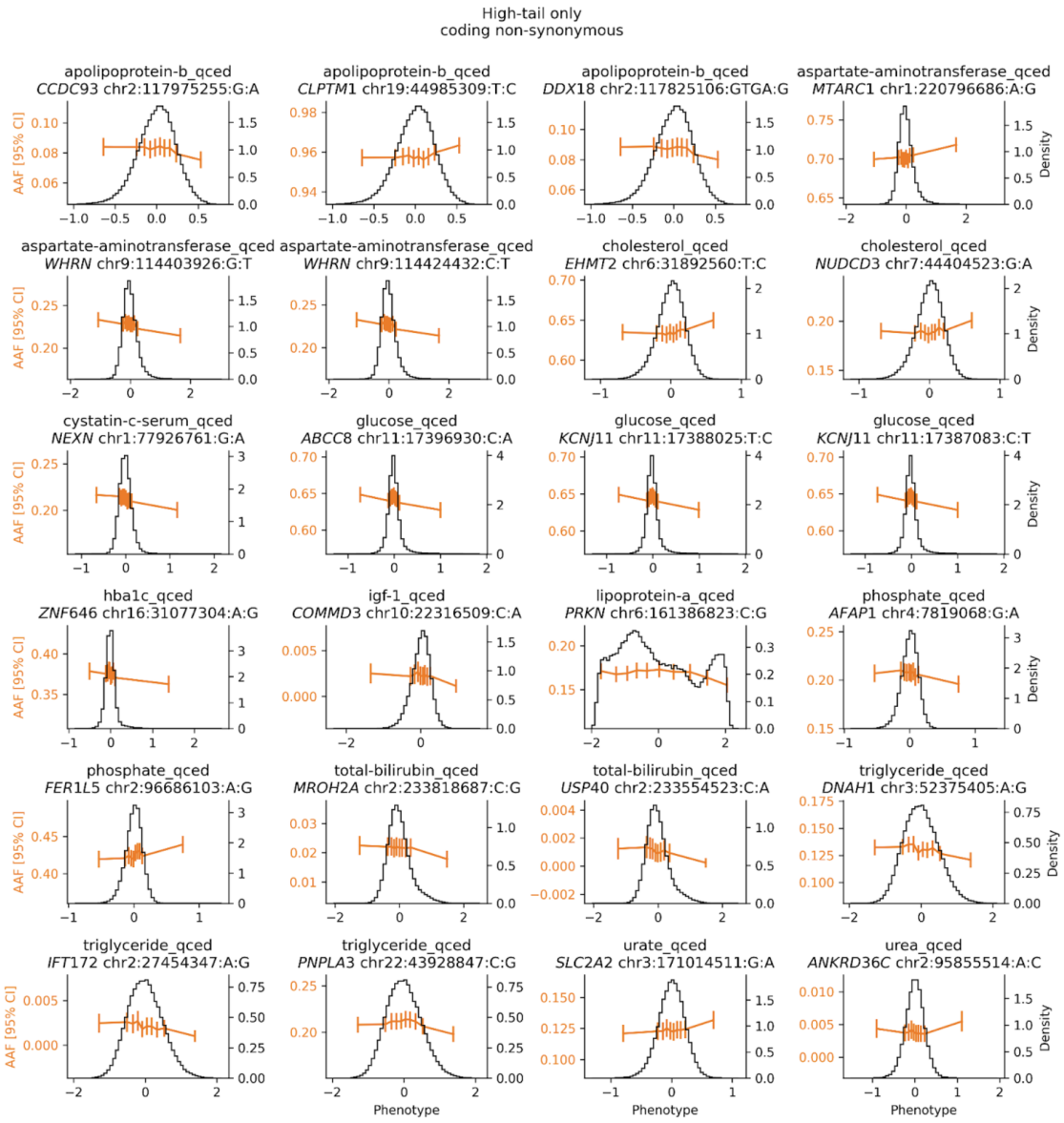

Figure S2.5: Allele frequency across phenotype distribution for coding non-synonymous high-tail-only VOI. Orange: Mean alternate allele frequency, with 95% CI calculated as $\pm 1.96 \times$ SEM. Black: Histogram of QCed biomarker distribution. Non-monotonic allele frequency distributions are highlighted with the $\cup$ or $\cap$ in the upper left of the plots. Non-monotonicity was defined as having at least one CI in the middle that was less than (or greater than) and non-overlapping with any of the CIs to the left or right. This plot only includes variants which were significant for high-tail status alone. VOI, variant-of-interest (see Methods). CI, confidence interval. SEM, standard error of mean.

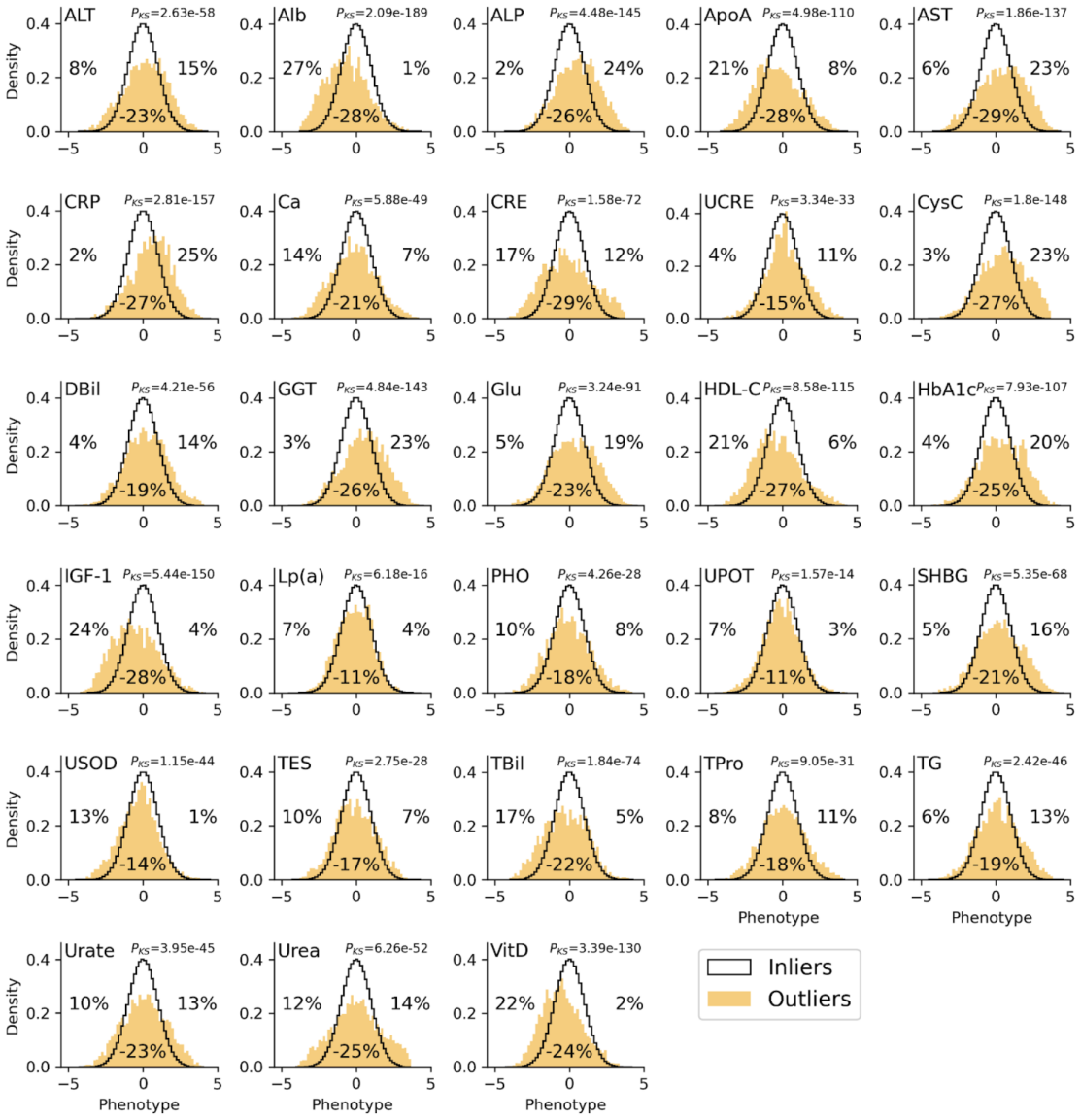

Figure S2.6: Marginal single-biomarker distributions for individuals with extreme Mahalanobis distance in the joint distribution of biomarkers. Outliers are individuals with Mahalanobis distance $P < 0.001$, inliers are everyone else. Percentages to the left and right of the inlier distribution indicate the relative excess of outliers in that tail. Graphically, the excess is the percent area of the outlier distribution histogram that is outside of the inlier distribution histogram. The negative percentage inside the inlier distribution indicates the percentage point difference between the inlier distribution and the outlier distribution. Graphically, the deficit is the percent of white space within the black inlier histogram. In the upper right corner is shown the P-value of the two-sided Kolmogorov-Smirnov test (P_{KS}) comparing the marginal biomarker distributions of inliers to the marginal distribution for outliers.

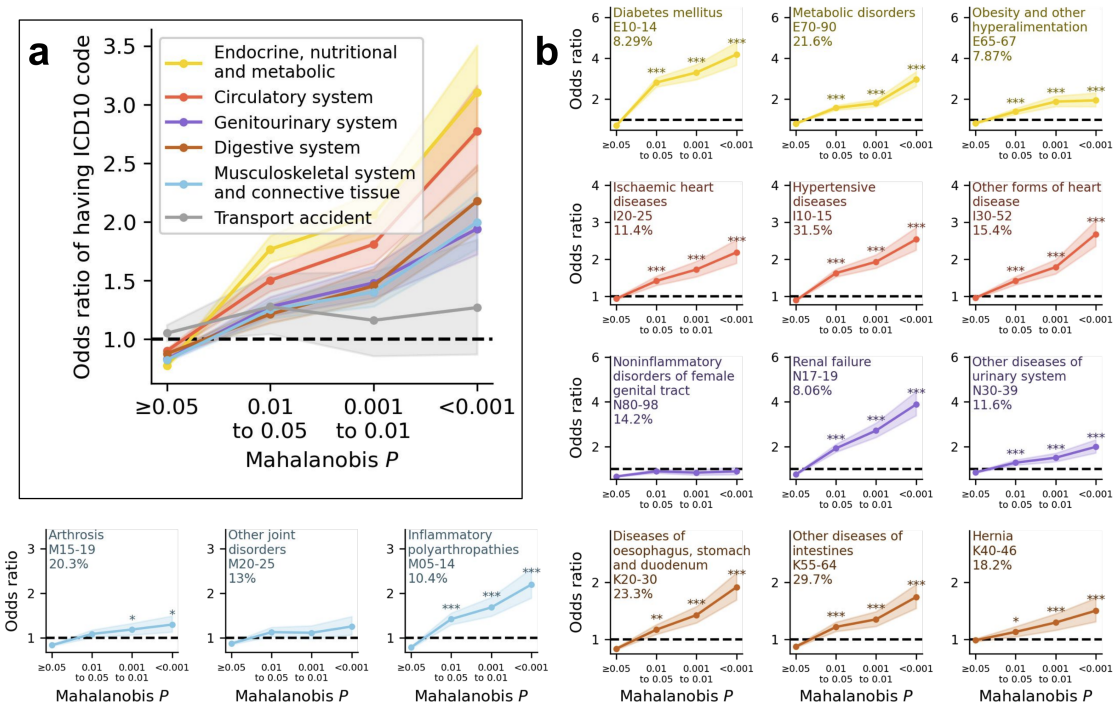

Figure S2.7: Enrichment of ICD-10 codes in individuals with low Mahalanobis P-values. **a**, Odds of having an ICD-10 code belonging to one of the top five most prevalent disease-related ICD-10 chapters (Methods), or an ICD-10 code relating to transport-related accidents, for each subgroup of individuals defined by a range of Mahalanobis distance P-values (low P-values correspond to high Mahalanobis distance). **b**, Odds of having an ICD-10 code belonging to one of the top three most prevalent subcategories under each of the most prevalent disease-related ICD-10 chapters, for each subgroup of individuals defined by a range of Mahalanobis distance P-values.. The title of each plot includes the name of the subcategory, the codes associated with that range, and the prevalence of the subcategory. $^{***}P < 1 \times 10^{-5}$, $^{**}P < 1 \times 10^{-4}$, $^{*}P < 0.05/15$ (Bonferroni correction for 15 subcategories) for one-sided Fisher's exact test with an alternative hypothesis that the odds ratio is greater than one. Shading around points indicates the 95% CI for each OR.

S2.2 Supplementary tables

Some supplementary tables have been cropped due to size. The full supplementary tables can be downloaded here: https://doi.org/10.5281/zenodo.15275847

Table S2.1: Biomarker overview table.

pheno_fmt	abbrev	units	n_both_sex	mean_both_	stdev_both_	n_female	mean_femal	stdev_femal	n_male	mean_male	stdev_male	ref_range_fe	ref_range_m	ref_range_b	link
Alanine aminotra	ALT	U/L	176588	23.4412408	13.89882162	97191	20.24797605	12.08674026	79397	27.35016172	14.93514682	(10, 35)	(10, 50)		https://www.esht.nhs.uk/wp-
Albumin	Alb	g/L	162481	45.29366055	2.611886513	88764	45.03988813	2.585027197	73717	45.59923261	2.611359657			(35, 50)	https://www.esht.nhs.uk/wp-
Alkaline phospha	ALP	IU/L	176646	83.36179534	25.68674453	97199	84.62626262	26.87663815	79447	81.81478973	24.06132619			(30, 130)	https://www.esht.nhs.uk/wp-
Apolipoprotein A	ApoA	g/L	161416	1.547166978	0.2700892692	87855	1.640886973	0.2660382426	73561	1.435235791	0.2291811722	(0.91, 2.24)	(0.73, 1.86)		https://www.gloshospitals.nh
Apolipoprotein B	ApoB	g/L	175774	1.034017431	0.2369092253	96862	1.038839535	0.2364663087	78912	1.028098451	0.2373193505	(0.53, 1.82)	(0.49, 1.73)		https://www.nbt.nhs.uk/seve
Aspartate amino	AST	U/L	176029	26.15150004	10.25545891	96874	24.52195532	9.336008025	79155	28.14582149	10.9541908	(0, 35)	(0, 50)		https://www.nbt.nhs.uk/sites
C-reactive protei	CRP	mg/L	176241	2.545189371	4.29226208	96990	2.651215795	4.288857765	79251	2.41543072	4.292889752			(0, 5)	https://www.esht.nhs.uk/wp-
Calcium	Ca	mmol/L	162392	2.38182735	0.09420719401	88711	2.388548906	0.09730869613	73681	2.373734681	0.08966645488			(2.2, 2.6)	https://www.esht.nhs.uk/wp-
Creatinine (seru	CRE	umol/L	176539	72.21560732	17.88925318	97134	64.43325921	13.36960129	79405	81.73554436	18.11447043	(45, 84)	(59, 104)		https://www.esht.nhs.uk/wp-
Creatinine (urine	UCRE	umol/L	180103	8721.518848	5691.713274	98848	7066.826188	4834.969811	81255	10734.47879	5998.703683	(5200, 14100)	(7100, 15900)		https://www.southtees.nhs.u
Cystatin C	CysC	mg/L	176624	0.9049951705	0.1738972861	97182	0.8758132576	0.1649417365	79442	0.9406936255	0.1778668362			(0, 1.4)	https://www.nbt.nhs.uk/seve
Direct bilirubin	DBil	umol/L	150050	1.827398867	0.8300322984	75897	1.649551102	0.6855830453	74153	2.009429423	0.920614869			(0, 8)	https://www.southtees.nhs.u
Gamma glutamy	GGT	U/L	176557	36.79240868	41.32300917	97153	29.96142373	33.78188569	79404	45.15030855	47.68271497	(0, 38)	(0, 55)		https://www.nbt.nhs.uk/sites
Glucose	Glu	mmol/L	162265	5.114158555	1.176510907	88641	5.061429011	1.01684338	73624	5.177643282	1.34106622			(3, 7.8)	https://www.nbt.nhs.uk/seve
Glycated haemo	HbA1c	mmol/mol	176891	35.85044858	6.406428052	97378	35.57959498	5.648606532	79513	36.18215763	7.213260695			(0, 48)	https://www.nbt.nhs.uk/seve
HDL cholesterol	HDL-C	mmol/L	162392	1.46079257	0.3833414683	88714	1.604168395	0.3780675233	73678	1.288157008	0.3118208948	(1.2, inf)	(1, inf)		https://www.nhs.uk/conditio
Insulin-like grow	IGF-1	nmol/L	175703	21.46711881	5.661032989	96666	21.01100648	5.73967192	79037	22.02496583	5.512308597	(2.2, 35.5)	(2, 32.2)		https://www.esht.nhs.uk/wp-
LDL cholesterol	LDL-C	mmol/L	176345	3.569914537	0.8646040332	97051	3.635570628	0.8650516145	79294	3.489555502	0.857244569			(1, 3)	https://www.esht.nhs.uk/wp-
Lipoprotein A	Lp(a)	nmol/L	140639	43.91805779	49.25090605	77655	44.5433766	49.41230614	62984	43.14708212	49.04059698			(0, 125)	https://www.gloshospitals.nh
Microalbumin (ur	UMIC	mg/L	55485	28.61318374	117.590763	27324	22.86596399	94.13020223	28161	34.18958489	136.3245645			(0, 20)	https://www.ruh.nhs.uk/path
Phosphate	PHO	mmol/L	162156	1.166173882	0.1610284942	88585	1.200731851	0.1512741921	73571	1.124563496	0.1625701893			(0.8, 1.5)	https://www.esht.nhs.uk/wp-
Potassium (urine	UPOT	mmol/24hrs	179722	62.85608334	33.64536609	98638	58.5984506	33.05488213	81084	68.03545755	33.63121127			(25, 125)	https://www.southtees.nhs.u
Sex hormone bir	SHBG	nmol/L	160907	52.14999379	27.79947571	87731	62.47620989	30.86483915	73176	39.76985186	16.61665729	(27, 128)	(18, 77)		https://www.esht.nhs.uk/wp-
Sodium urine	USOD	mmol/24hrs	179725	75.30086799	43.09782846	98553	65.92217081	39.35991122	81172	86.68778396	44.66772145			(40, 220)	https://www.esht.nhs.uk/wp-
Testosterone	TES	nmol/L	159676	6.474466864	6.033738079	80966	1.123485908	0.7071843773	78710	11.97881859	3.686348987	(0, 2.9)	(6, 27)		https://www.nbt.nhs.uk/sites
Total bilirubin	TBil	umol/L	175934	9.134628724	4.414348687	96806	8.155362168	3.698568945	79128	10.33267339	4.897203973			(0, 21)	https://www.esht.nhs.uk/wp-
Total cholesterol	CHO	mmol/L	176636	5.719173702	1.137530557	97183	5.895559583	1.120115089	79453	5.503427171	1.121374905			(0, 5)	https://www.nhs.uk/conditio
Total protein	TPro	g/L	162265	72.4446151	4.02698413	88658	72.34402208	4.021121656	73607	72.56577717	4.030729761			(60, 80)	https://www.esht.nhs.uk/wp-
Triglyceride	TG	mmol/L	176492	1.742268958	1.014242066	97124	1.550538302	0.8505682372	79368	1.976893093	1.141126727			(0, 1.7)	https://www.esht.nhs.uk/wp-
Urate	Urate	umol/L	176447	308.0576706	79.96656568	97081	270.0829678	65.28666739	79366	354.5085704	71.28316695	(140, 360)	(200, 430)		https://www.nbt.nhs.uk/sites
Urea	Urea	mmol/L	176513	5.424622096	1.381752403	97123	5.251393079	1.312581572	79390	5.636544527	1.433749925			(2.5, 7.8)	https://www.nbt.nhs.uk/sites
Vitamin D	VitD	nmol/L	168925	49.41768122	20.87736762	91826	49.41033585	20.84902185	77099	49.42642966	20.91120958			(25, inf)	https://www.gloshospitals.nh

Table S2.2: Correlation and odds of overlapping outliers.

pheno1	pheno2	pheno_label	r	pearson_pv	r2	odds_ratio_	pval_fisher_	pval_fisher_	or_95ci_low	or_95ci_upp	or_90ci_low	or_90ci_upp	odds_ratio_	pval_fisher_	pval_fisher_	or_95ci_low	or_95ci_upp	or_90ci_low	or_90ci_upp	odds_ratio_	pval_fisher_	pval_fisher_
direct-bilirubin	total-bilirubin	DBil,TBil	0.9051861794	0	0.8193620193	1515.597646	1.23E-138	1.23E-138	373.3519459	6152.468867	467.6743755	4911.614456		1.00E+00	1.00E+00	2465.385031	36467498.26	5334.568563	16853551.19	1515.597646	1.23E-138	1.23E-138
glucose	hba1c	Glu,HbA1c	0.6099140116	0	0.3719951015	131.8245015	0.00E+00	0.00E+00	121.6411857	142.8603239	123.223668	141.0256606		1	1	13.03614526	33248.24294	24.49042593	17697.89247	154.4506459	0.00E+00	0.00E+00
cholesterol	ldl-cholesterol	CHO,LDL-C	0.9520741533	0	0.9064451935	93.4956909	0	0	90.31897617	96.78413759	90.8223287	96.24774372		1	1	212.0461063	572447.702	400.1859873	303322.2304	93.73521439	0.00E+00	0.00E+00
creatinine-serum	cystatin-c-serum	CRE,CysC	0.6343569559	0	0.4024087475	48.5417875	0	0	43.70944594	53.90837341	44.45258316	53.00715877		1	1	1.181596281	3002.090927	2.219128423	1598.492199	70.70351647	0	0
cystatin-c-serum	urea	CysC,Urea	0.4268974342	0	0.1822414193	31.2036012	0	0	28.37791404	34.31065181	28.81431089	33.7910121		1	1	6.082833172	15475.94496	11.42528864	8239.405966	33.82274873	0	0
alanine-aminotra	aspartate-amino	ALT,AST	0.7560778045	0	0.5716536464	25.61321341	0	0	24.35511658	26.93629894	24.55313491	26.71906067		1	1	0.5810610855	1476.058957	1.091262568	785.9523865	50.65985064	0	0
cystatin-c-serum	direct-bilirubin	CysC,DBil	0.0418402427	4.37E-59	0.001750605909	15.25819471	2.41E-11	2.41E-11	8.481324779	27.45001658	9.321102677	24.97692751		1	1	2465.812588	36473822.57	5335.493705	16856473.98	15.25819471	2.41E-11	2.41E-11
alkaline-phospha	direct-bilirubin	ALP,DBil	-0.00586594058	0.02308523685	3.44E-05	12.34544911	2.55E-20	2.55E-20	8.003106844	19.04386842	8.580724492	17.76191671		1	1	15.94223548	40704.8013	29.95261225	21665.07289	12.80205246	9.70E-21	9.70E-21
aspartate-amino	direct-bilirubin	AST,DBil	0.1639885013	0	0.02689222855	11.33135774	9.10E-21	9.10E-21	7.413800473	17.3190078	7.937102367	16.17714655		1	1	2467.128148	36493281.97	5338.340294	16865467.2	11.33135774	9.10E-21	9.10E-21
albumin	total-protein	Alb,TPro	0.4993559873	0	0.2493564021	11.16014253	0	0	10.45366442	11.91436573	10.56415377	11.78975467	721.3135436	1.90E-83	1.90E-83	464.7607081	1119.486263	498.79285	1043.104824	11.2966806	0	0
cystatin-c-serum	urate	CysC,Urate	0.3877936942	0	0.1503839493	10.36546764	0	0	9.444806811	11.37587264	9.587108853	11.20701987		1	1	2.603735171	6617.968259	4.890171033	3523.687945	10.87134552	0	0
creatinine-serum	urea	CRE,Urea	0.4627620521	0	0.2141487169	8.255553954	0.00E+00	0	7.842990674	8.689819219	7.907901381	8.618490268	17.89874232	2.10E-102	2.10E-102	14.65253297	21.8641362	15.13162275	21.17188499	12.88371367	0	0
albumin	direct-bilirubin	Alb,DBil	0.08068410749	2.59E-198	0.006509925201	8.017957555	4.06E-12	4.06E-12	4.945224933	12.99994322	5.344767679	12.02814551		1	1	38.88476615	100152.8545	73.1087628	53268.85831	2.77013345	0.008078071735	0.008078071735
aspartate-amino	gamma-glutamy	AST,GGT	0.4472598123	0	0.2000413397	7.678571113	0	0	7.325214534	8.048973045	7.380907883	7.988238747		1.00E+00	1.00E+00	2893.122796	42794478.73	6260.101664	19777583.27	7.678571113	0	0
apolipoprotein-b	direct-bilirubin	ApoB,DBil	-0.2012662524	0	0.04050810435	7.30754661	0.000257609443	0.000257609443	3.188836884	16.74599216	3.643620819	14.65581632		1	1	3.009777596	7652.06397	5.652895989	4074.196792	0	1	1
calcium	direct-bilirubin	Ca,DBil	-0.00036029838	0.8934694284	1.30E-07	6.650229849	9.32E-09	9.32E-09	3.914348429	11.29831895	4.262514618	10.37546167		1	1	1.109857154	2820.017626	2.08440842	1501.537177	2.006912874	0.1955346246	0.1955346246
direct-bilirubin	total-protein	DBil,TPro	0.02726868828	3.94E-24	0.000743581360	6.142423995	2.93E-08	2.93E-08	3.610380887	10.45024714	3.932397655	9.594495735		1	1	26.44667052	67812.50716	49.70551661	36080.80463	5.373135537	8.04E-07	8.04E-07
cystatin-c-serum	hba1c	CysC,HbA1c	0.1591148528	0	0.02531753639	6.023231153	1.24E-110	1.24E-110	5.288026944	6.86065217	5.399867853	6.718555809		1	1	2773.538427	41025619.33	6001.346649	18960099.85	6.023231153	1.24E-110	1.24E-110
calcium	total-protein	Ca,TPro	0.5106566988	0	0.2607702641	5.523457297	0.00E-02	0.00E-02	5.119241897	5.959589551	5.182174567	5.887215902	82.85037827	1.19E-107	1.19E-107	58.85720469	116.624383	62.1832874	110.3863348	12.68561359	0	0
cystatin-c-serum	microalbumin-ur	CysC,UMIC	0.2171394189	0	0.04714952724	5.193783532	1.20E-159	1.20E-159	4.606875667	5.855462426	4.696552184	5.743657542		1	1	868.2298788	12842814.43	1878.666177	5935336.118	5.193783532	1.20E-159	1.20E-159
alkaline-phospha	aspartate-amino	ALP,AST	0.1904160858	0	0.03625828575	4.657567902	0.00E-02	0.00E-02	4.347973954	4.989206236	4.396323105	4.934336773		1	1	15.81788103	40354.40027	29.71702614	21479.97917	4.83273043	0	0
creatinine-urine	potassium-urine	UCRE,UPOT	0.6880406512	0	0.4733999377	4.556198444	0.00E+00	0	4.41663397	4.700173118	4.438780359	4.676722563	37.4699253	0	0	35.03170406	40.07784775	35.41272263	39.64663538	9.658482834	0.00E+00	0.00E+00
creatinine-urine	sodium-urine	UCRE,USOD	0.4555124417	0	0.2074915845	4.448152181	0.00E+00	0	4.336773571	4.562391258	4.354490314	4.543828646	6.926043033	0	0	6.756195708	7.100160235	6.783218969	7.07187433	3.814438779	2.40E-58	2.40E-58
c-reactive-protei	cystatin-c-serum	CRP,CysC	0.1702468653	0	0.0289839951	4.328186842	1.73E-159	1.73E-159	3.929047449	4.767873531	3.990642039	4.694282562		1	1	2895.984138	42836802.93	6266.292997	19797143.52	4.328186842	1.73E-159	1.73E-159
hdl-cholesterol	triglyceride	HDL-C,TG	-0.4388160272	0	0.1925595057	4.253207209	0	0	4.128219427	4.381979175	4.148063504	4.36101606		1	1	0.1164837772	295.8643983	0.2187603045	157.5395624		1	1
creatinine-serum	direct-bilirubin	CRE,DBil	0.1201983984	0	0.01444765497	4.173322962	8.01E-07	8.01E-07	2.547227849	6.837481991	2.757654956	6.315737401		1	1	1.243143054	3158.718192	2.334733128	1681.878983	2.531498082	0.009695375097	0.009695375097
alanine-aminotra	gamma-glutamy	ALT,GGT	0.4757424913	0	0.226330918	4.145496138	0	0	4.012008398	4.28342529	4.033176036	4.260944248		1	1	0.5813868138	1476.885863	1.091874271	786.3927094	9.901910213	0	0
hba1c	hdl-cholesterol	HbA1c,HDL-C	-0.1407488919	0	0.01981025057	4.131682431	0	0	3.896731553	4.380799519	3.933583629	4.339757666		1	1	0.1164850521	295.8683025	0.2187627385	157.5416128		1	1
direct-bilirubin	glucose	DBil,Glu	0.02653056171	6.23E-23	0.000703870704	4.08870674	0.001252790999	0.001252790999	1.975127894	8.464020409	2.220231059	7.529631988		1	1	11.96988852	30532.72198	22.48752956	16252.2645	4.400195695	0.000794051655	0.000794051655
alkaline-phospha	cystatin-c-serum	ALP,CysC	0.2020414972	0	0.0408207666	4.06382252	2.05E-68	2.05E-68	3.553612663	4.647285747	3.631094086	4.548120508		1	1	15.86467723	40473.78554	29.80494204	21543.526	4.20998714	3.22E-71	3.22E-71
direct-bilirubin	urea	DBil,Urea	-0.00826413285	0.001374438309	6.83E-05	4.063787645	1.03E-05	1.03E-05	2.373258668	6.958520893	2.587614128	6.382083729		1	1	5.740169631	14606.91342	10.78183373	7776.614157	2.269729058	0.03828731115	0.02453822328
cystatin-c-serum	glucose	CysC,Glu	0.05216329476	4.28E-98	0.00272100932	4.029078475	1.04E-40	1.04E-40	3.391583086	4.786400023	3.486814529	4.655674464		1	1	12.82609254	32704.32174	24.09532428	17408.71598	4.097526227	1.13E-39	1.13E-39
direct-bilirubin	gamma-glutamy	DBil,GGT	0.1145085057	0	0.01311219787	3.996544404	2.34E-10	2.34E-10	2.648739464	6.030176767	2.829832281	5.644280505		1	1	2466.240145	36480146.87	5336.418846	16859396.78	3.996544404	2.34E-10	2.34E-10
cystatin-c-serum	hdl-cholesterol	CysC,HDL-C	-0.2458865708	0	0.06046020571	3.995196952	9.77E-147	9.77E-147	3.622982793	4.405651255	3.680396702	4.33692343		1	1	0.1165016376	295.9097615	0.2187938469	157.5637171		1	1
alkaline-phospha	gamma-glutamy	ALP,GGT	0.3104693097	0	0.09639119227	3.983336809	0	0	3.792194175	4.184113841	3.822294285	4.151164446		1	1	15.86575714	40476.54058	29.80697087	21544.99247	4.179353875	0	0
albumin	calcium	Alb,Ca	0.4967904964	0	0.2468007973	3.914431971	9.02E-170	9.02E-170	3.600257389	4.256022833	3.64901198	4.199157947	83.40950679	1.35E-87	1.35E-87	56.98631359	122.0845039	60.58564574	114.8315866	8.567574688	7.94E-307	7.94E-307
direct-bilirubin	testosterone	DBil,TES	0.2176290833	0	0.04736241789	3.910006021	0.003040495109	0.003040495109	1.807499621	8.458174434	2.046224607	7.471392453		1	1	1.422639425	3615.175048	2.67186538	1924.906319	6.55362291	0.003918918097	0.003918918097
shbg	testosterone	SHBG,TES	-0.2720400053	0	0.0740057645	3.589917035	1.95E-162	1.95E-162	3.304031727	3.900538912	3.348409151	3.848843954	6.814590777	2.26E-238	2.26E-238	6.176406056	7.518716716	6.274823351	7.400789609	2.087857566	2.15E-07	1.62E-07
glucose	testosterone	Glu,TES	0.0067184120424	0.01002806901	4.51E-05	3.567815435	3.58E-54	3.58E-54	3.107770985	4.095960428	3.177517301	4.006054341	0.3122204859	0.3858627362	0.9589807818	0.04377354987	2.226952854	0.06003323337	1.623794461	0.9892301548	1	0.5492777945
alkaline-phospha	c-reactive-protei	ALP,CRP	0.1981514984	0	0.03926401632	3.530373369	0	0	3.343121084	3.728113882	3.372542104	3.695590963		1	1	15.90847984	40586.35972	29.8872828	21603.41205	3.640244745	0	0
urate	urea	Urate,Urea	0.2871255979	0	0.08244110898	3.508516083	0	0	3.342666809	3.682594111	3.368792164	3.654035187	7.935799801	5.19E-18	5.19E-18	5.532601541	11.38287621	5.86295266	10.74150213	3.928807678	0	0
cystatin-c-serum	igf-1	CysC,IGF-1	-0.0533891296	5.28E-111	0.00285039916	3.500879898	2.97E-33	2.97E-33	2.940117108	4.168595881	3.02380156	4.053228963		1	1	1373.28409	5611836.045	2679.567638	2876077.86	3.502675559	2.83E-33	2.83E-33
hba1c	testosterone	HbA1c,TES	-0.0048556259	0.05795425044	2.36E-05	3.499902023	5.71E-71	5.71E-71	3.106581339	3.943020586	3.1666968	3.868167666		1	1	1.444478474	3670.507131	2.712871605	1954.375035	1.226545996	0.1971142485	0.1127420077
direct-bilirubin	shbg	DBil,SHBG	-0.0383650267	8.89E-46	0.001471888776	3.495398068	3.29E-06	3.29E-06	2.171837891	5.625561513	2.344522211	5.211214291		1	1	0.3120672083	792.7013635	0.5860760307	422.088754	8.859784077	5.33E-12	5.33E-12
cystatin-c-serum	testosterone	CysC,TES	0.156707047	0	0.02455709857	3.467861681	6.14E-26	6.14E-26	2.846511788	4.224842732	2.938320209	4.09283665		1	1	1.448419024	3680.446372	2.720267951	1959.670385	1.748371733	0.02424906423	0.01457764724
glucose	hdl-cholesterol	Glu,HDL-C	-0.08104695623	1.86E-234	0.006568609115	3.441964222	5.58E-243	5.40E-243	3.216797363	3.682892135	3.251978432	3.643049288	0.8930185681	0.5909772685	0.7530593391	0.6190190897	1.28829979	0.6565868619	1.214587451		1	1
creatinine-serum	urate	CRE,Urate	0.4482490262	0	0.2009271895	3.373243042	0.00E+00	0.00E+00	3.227186056	3.525910322	3.250234196	3.5009073	7.176710974	4.01E-65	4.01E-65	5.993496302	8.593511668	6.169642538	8.348162813	5.018617933	0	0
apolipoprotein-a	cholesterol	ApoA,CHO	0.3187621183	0	0.101609288	3.269241649	1.58E-214	7.65E-215	3.000164283	3.562451902	3.041880922	3.513596104		1	1	45.40349065	116942.5721	85.36486356	62198.90431	3.524254448	6.48E-232	3.65E-232
apolipoprotein-b	hba1c	ApoB,HbA1c	-0.02164884085	7.29E-19	0.000468672310	3.221229988	9.00E-31	9.00E-31	2.711933268	3.826171816	2.788018513	3.721755286		1	1	3.346120972	8507.082966	6.284602487	4529.439815	1.340401846	0.1515643043	0.1024771719
potassium-urine	sodium-urine	UPOT,USOD	0.369735806	0	0.1367045663	3.041196881	0	0	2.960236921	3.124371028	2.973106223	3.110846965	6.866851892	0	0	6.645977742	7.09506663	6.681003104	7.057870527	0.5887641092	0.006354345405	0.9981231927
c-reactive-protei	direct-bilirubin	CRP,DBil	-0.03399280102	1.55E-39	0.001155510521	3.005560632	1.97E-05	1.97E-05	1.889945678	4.779711298	2.036298119	4.436184775		1	1	2462.260577	36421282.19	5327.807912	16832192.3	3.005560632	1.97E-05	1.97E-05

Table S2.3: Odds of ICD-10 codes with outlying biomarkers.

chapter	desc	is_chapter	cases	prevalence	num_biomai	num_biomai	odds_ratio	pval_fisher	pval_fisher_	or_95ci_low	or_95ci_upp	or_90ci_low	or_90ci_upp
Endocrine, nutrit	Diabetes mellitus	FALSE	15356	0.08283793844	15	19	23.76969109	1.23E-22	1.23E-22	12.60149575	44.83580568	13.95505389	40.48699627
Genitourinary sy	Renal failure	FALSE	14937	0.08057764304	15	19	16.50866996	8.63E-18	8.63E-18	9.049273055	30.11691461	9.967618489	27.34215642
Endocrine, nutrit	Endocrine, nutritional and metabolic	TRUE	61584	0.3322148737	15	19	10.63060054	4.86E-12	4.86E-12	4.73906732	23.84639428	5.396381878	20.94174771
Circulatory syste	Circulatory system	TRUE	86188	0.464941146	15	19	8.9799278	6.51E-09	5.67E-09	3.539307689	22.7838635	4.11081403	19.61633455
Circulatory syste	Hypertensive diseases	FALSE	58333	0.314677355	15	19	7.408452975	5.48E-10	5.48E-10	3.660058146	14.99571135	4.099427422	13.38849791
Endocrine, nutrit	Diabetes mellitus	FALSE	15356	0.08283793844	10	14	6.807014742	0	0	6.2922955	7.363838793	6.372342901	7.271336526
Endocrine, nutrit	Obesity and other hyperalimentation	FALSE	14584	0.07867338462	15	19	6.062506361	6.87E-07	6.87E-07	3.249648737	11.31014037	3.592331725	10.23123314
Endocrine, nutrit	Metabolic disorders	FALSE	40067	0.2161414222	15	19	5.241175018	7.88E-08	7.88E-08	2.873341647	9.560267778	3.164869581	8.67963588
Genitourinary sy	Renal failure	FALSE	14937	0.08057764304	10	14	5.04933945	6.21E-251	6.21E-251	4.64830746	5.484970411	4.710565022	5.412477858
Circulatory syste	Ischaemic heart diseases	FALSE	21105	0.1138509176	15	19	4.903814	2.93E-06	2.93E-06	2.672494938	8.998105628	2.946453195	8.16147081
Genitourinary sy	Genitourinary system	TRUE	65970	0.3558751497	15	19	4.317651396	3.91E-06	2.68E-06	2.259219307	8.25157324	2.507161153	7.435546596
Circulatory syste	Other forms of heart disease	FALSE	28478	0.1536245644	15	19	4.189253312	9.73E-06	9.73E-06	2.306809192	7.607843498	2.539055007	6.911958685
Genitourinary sy	Other diseases of urinary system	FALSE	21471	0.1158253045	15	19	3.950522641	7.61E-05	7.61E-05	2.117716946	7.369553881	2.341009847	6.666622595
Endocrine, nutrit	Endocrine, nutritional and metabolic	TRUE	61584	0.3322148737	10	14	3.899289897	1.15E-270	9.32E-271	3.605184156	4.217388362	3.650926505	4.164548829
Endocrine, nutrit	Obesity and other hyperalimentation	FALSE	14584	0.07867338462	10	14	3.408245123	1.31E-122	1.31E-122	3.111998769	3.732692615	3.157828861	3.678519429
Musculoskeletal	Inflammatory polyarthropathies	FALSE	19287	0.104043717	15	19	3.230642023	0.001391120764	0.001391120764	1.663730434	6.273280617	1.851052948	5.638438325
Circulatory syste	Hypertensive diseases	FALSE	58333	0.314677355	10	14	3.157688277	8.73E-197	6.38E-197	2.927218991	3.406303145	2.963104179	3.365050518
Digestive systen	Digestive system	TRUE	98966	0.533872064	15	19	2.969243521	0.001340549273	0.000940697927	1.46693283	6.010095967	1.643027413	5.365952519
Circulatory syste	Circulatory system	TRUE	86188	0.464941146	10	14	2.875766377	1.10E-155	7.35E-156	2.649008656	3.12193478	2.684220605	3.08098084
Endocrine, nutrit	Metabolic disorders	FALSE	40067	0.2161414222	10	14	2.86948221	1.44E-152	1.16E-152	2.661015939	3.094279906	2.693480179	3.056984869
Musculoskeletal	Inflammatory polyarthropathies	FALSE	19287	0.104043717	10	14	2.689001461	7.96E-89	6.46E-89	2.460819767	2.938341504	2.496154105	2.896747777
Musculoskeletal	Musculoskeletal system and connective t	TRUE	77557	0.4183812185	15	19	2.688282943	0.001874514924	0.001093110938	1.441216371	5.014420683	1.593154693	4.536198032
Digestive systen	Other diseases of intestines	FALSE	55118	0.2973340382	15	19	2.363765065	0.004749367788	0.00374848707	1.308943594	4.268621893	1.439424295	3.881680545
Circulatory syste	Ischaemic heart diseases	FALSE	21105	0.1138509176	10	14	2.305438366	5.09E-64	4.08E-64	2.107633537	2.521807498	2.138250504	2.485698494
Circulatory syste	Other forms of heart disease	FALSE	28478	0.1536245644	10	14	2.295095481	9.17E-77	7.41E-77	2.113420604	2.49238758	2.141627884	2.459560461
Genitourinary sy	Genitourinary system	TRUE	65970	0.3558751497	10	14	2.209185904	2.47E-95	1.52E-95	2.049708374	2.38107158	2.074548903	2.352560767
Genitourinary sy	Other diseases of urinary system	FALSE	21471	0.1158253045	10	14	2.129176002	1.00E-51	7.55E-52	1.943752228	2.332288233	1.97243554	2.298371915
Digestive systen	Digestive system	TRUE	98966	0.533872064	10	14	2.009901542	9.61E-69	5.42E-69	1.853745784	2.179211542	1.878007334	2.151058804
Musculoskeletal	Musculoskeletal system and connective t	TRUE	77557	0.4183812185	10	14	1.896651202	5.15E-63	3.18E-63	1.758898423	2.045192454	1.780350719	2.020548953
Digestive systen	Diseases of oesophagus, stomach and d	FALSE	43188	0.232977656	15	19	1.881616392	0.04895337028	0.03505319777	1.017959923	3.478015358	1.123634661	3.150917615
Digestive systen	Other diseases of intestines	FALSE	55118	0.2973340382	10	14	1.865214942	4.25E-57	2.65E-57	1.730046952	2.010943561	1.75109828	1.986768428
Digestive systen	Diseases of oesophagus, stomach and d	FALSE	43188	0.232977656	10	14	1.789455398	5.39E-45	2.98E-45	1.654640924	1.935254093	1.6756095	1.911036325
Endocrine, nutrit	Diabetes mellitus	FALSE	15356	0.08283793844	5	9	1.787619613	1.70E-256	8.99E-257	1.729347461	1.847855305	1.738586288	1.838035823
Musculoskeletal	Other joint disorders	FALSE	24148	0.1302663804	15	19	1.717101549	0.1731823524	0.1111783788	0.8253014268	3.572558623	0.9284704079	3.175586108
Musculoskeletal	Arthrosis	FALSE	37692	0.2033294853	10	14	1.613528009	4.10E-28	2.40E-28	1.48601651	1.751980962	1.505815457	1.728945351
Endocrine, nutrit	Obesity and other hyperalimentation	FALSE	14584	0.07867338462	5	9	1.56447183	1.95E-145	1.06E-145	1.512250576	1.618496396	1.52052723	1.609686469
Genitourinary sy	Renal failure	FALSE	14937	0.08057764304	5	9	1.560322083	7.63E-147	4.31E-147	1.508788276	1.613616066	1.516957231	1.604926595
Digestive systen	Hernia	FALSE	33822	0.1824527712	10	14	1.484484413	5.33E-18	2.96E-18	1.361429611	1.618661703	1.3805023	1.596298661
Endocrine, nutrit	Endocrine, nutritional and metabolic	TRUE	61584	0.3322148737	5	9	1.423265168	3.24E-267	1.88E-267	1.395408184	1.45167827	1.399849785	1.447072221

Table S2.4: Curation of biomarkers with GEL disorders.

pheno	Low biomarker GEL panel or gene	Confi	High biomarker GEL panel or g	Confid	Conf	Sources
alkaline-phosphatase			Cholestasis	H	H	Scharschmidt et al. 1983 https://www.nejm.org/doi/full/10.1056/NEJM198306233082507
apolipoprotein-b	APOB (LoF decreases ApoB)		Familial hypercholesterolaemia	H	H	Tremblay et al. 2004 https://doi.org/10.1194/jlr.M300448-JLR200
c-reactive-protein	HNF1A (but only when comparing between diabetes patients)		Autoinflammatory disorders (H)	H	H	Sproston & Ashworth 2018 https://doi.org/10.3389/fimmu.2018.00754 McDonald et al. 2011 https://doi.org/10.2337/dc11-0323
ldl-cholesterol			Familial hypercholesterolaemia (H)	H	H	Defesche et al. 2017 https://www.nature.com/articles/nrdp201793
cholesterol			Familial hypercholesterolaemia (H)	H	H	Defesche et al. 2017 https://www.nature.com/articles/nrdp201793 Civeira et al. 2022 https://doi.org/10.1016/j.jacl.2022.09.006
lipoprotein-a			Familial hypercholesterolaemia (H)	H	H	Li et al. 2017 https://doi.org/10.1016/j.atherosclerosis.2017.03.021 Tromp et al. 2023 https://doi.org/10.1016/j.atherosclerosis.2022.11.020
microalbumin-urine			Proteinuric renal disease (H) Renal superpanel (Microalbuminuria	H	H	Chugh & Bakris 2007 https://doi.org/10.1111/j.1524-6175.2007.06445.x Urine albumin to creatinine ratio = microalbumin; https://www.nhs.uk/conditions/acr-test/ Barton 2008 https://www.nature.com/articles/ncpneph0891
triglyceride			Severe hypertriglyceridaemia (H) Familial chylomicronaemia syndrome	H	H	Baass et al. 2020 https://doi.org/10.1111/joim.13016
phosphate	Hypophosphataemia or rickets	H			H	Amanzadeh & Reilly 2006 https://www.nature.com/articles/ncpneph0124
oestradiol (excluded)	Primary ovarian insufficiency Differences in sex development – Females	H	Differences in sex development - Ma	H	HH	Sopiarz & Sparzak 2023 https://www.ncbi.nlm.nih.gov/books/NBK589674/
shbg	Congenital hypothyroidism (H) Hypogonadotropic hypogonadism (M)	H	Differences in sex development (H)	H	HH	Dumoulin et al. 1995 https://doi.org/10.1530/eje.0.1320594 Leger et al. https://doi.org/10.1210/jcem-71-5-1147 Anderson 1974 https://doi.org/10.1111/j.1365-2265.1974.tb03298.x Winters 2020 https://doi.org/10.1186/s40842-020-00106-3
testosterone	Differences in sex development (H) – low in men Hypogonadotropic hypogonadism (M)	H	Differences in sex development (H) -	H	HH	Zolla 2023 https://doi.org/10.3390/metabo13060681
glucose	Glycogen storage disease (H) Congenital hyperinsulinism (L) - typically newborns Ketotic hypoglycaemia (L) (in children)	H	GCK (mild hyperglycemia)	M	HM	Fukuda et al. 2023 https://pubmed.ncbi.nlm.nih.gov/37114839/ Steele et al. 2013 https://doi.org/10.1371/journal.pone.0065326 Elkamel et al. 2018 https://ojrd.biomedcentral.com/articles/10.1186/s13023-018-0867-6

Table S2.5: Extreme biomarker GEL rare disorder enrichment tests.

pheno_desc	phen	csq_	max_	sex	outlier_ty	outlier	disorder	n_carrier_in	n_carrier_in	n_alt	n_alt_in_sul	n_alt_in_infl	n_alt_in_out	min_pheno	max_pheno	inlier_min_p	inlier_max_	low_outlier_	high_outlier	odds_ratio	or_90ci_low	or_90ci_upp	pval_fisher
LDL cholesterol (LDL-C)	raw	pLoF	0.001	bothsexes	high_outlier	95ci_1p	Familial hypercholesterolaemia	81	21	111	103	82	21	0.266	9.797	2.003	5.387	1.764	5.787	24.75810307	16.53495767	37.0707733	3.25E-21
LDL cholesterol (LDL-C)	raw	pLoF	0.001	bothsexes	high_outlier	95ci_2.5p	Familial hypercholesterolaemia	81	28	111	110	82	28	0.266	9.797	2.003	5.387	2.002	5.388	13.18218994	9.187732743	18.9132767	2.87E-20
LDL cholesterol (LDL-C)	raw	pLoF	0.001	male	high_outlier	95ci_0.1p	Familial hypercholesterolaemia	32	9	111	42	33	9	0.266	9.797	1.894	5.241	1.25	6.465	280.4668787	148.8455293	528.4785539	4.50E-19
Cholesterol	raw	pLoF	0.001	bothsexes	high_outlier	95ci_2.5p	Familial hypercholesterolaemia	83	26	111	110	84	26	0.601	15.46	3.598	8.085	3.597	8.086	11.93434094	8.242943137	17.27883976	4.76E-18
LDL cholesterol (LDL-C)	raw	pLoF	0.001	bothsexes	high_outlier	95ci_0.1p	Familial hypercholesterolaemia	81	10	111	92	82	10	0.266	9.797	2.003	5.387	1.323	6.741	120.4875385	68.97044322	210.4850462	8.69E-18
Cholesterol	raw	pLoF	0.001	bothsexes	high_outlier	95ci_1p	Familial hypercholesterolaemia	83	16	111	100	84	16	0.601	15.46	3.598	8.085	3.25	8.616	18.39511263	11.73043199	28.84635188	9.27E-15
Apolipoprotein B (ApoB)	raw	pLoF	0.001	bothsexes	high_outlier	95ci_2.5p	Familial hypercholesterolaemia	81	22	105	104	82	22	0.4	2	0.613	1.553	0.612	1.554	10.39024834	6.993964686	15.43577434	1.82E-14
LDL cholesterol (LDL-C)	raw	pLoF	0.001	male	high_outlier	95ci_1p	Familial hypercholesterolaemia	32	12	111	45	33	12	0.266	9.797	1.894	5.241	1.664	5.612	35.97386233	20.58565967	62.86506197	2.36E-14
Cholesterol	raw	pLoF	0.001	male	high_outlier	95ci_1p	Familial hypercholesterolaemia	32	10	111	43	33	10	0.601	15.46	3.384	7.799	3.057	8.315	29.90377724	16.45725044	54.33689523	1.57E-11
LDL cholesterol (LDL-C)	raw	pLoF	0.001	male	high_outlier	95ci_2.5p	Familial hypercholesterolaemia	32	13	111	46	33	13	0.266	9.797	1.894	5.241	1.893	5.242	15.50294311	9.020097708	26.64508223	6.78E-11
Cholesterol	raw	pLoF	0.001	male	high_outlier	95ci_2.5p	Familial hypercholesterolaemia	32	13	111	46	33	13	0.601	15.46	3.384	7.799	3.383	7.8	15.48603101	9.010274492	26.61596565	6.87E-11
Cholesterol	raw	pLoF	0.001	male	high_outlier	95ci_0.1p	Familial hypercholesterolaemia	32	5	111	38	33	5	0.601	15.46	3.384	7.799	2.453	9.475	152.4917763	68.45501553	339.6937633	5.19E-10
Cholesterol	raw	pLoF	0.001	bothsexes	high_outlier	95ci_0.1p	Familial hypercholesterolaemia	83	6	111	90	84	6	0.601	15.46	3.598	8.085	2.586	9.908	69.7481354	34.61894347	140.5242883	7.02E-10
LDL cholesterol (LDL-C)	raw	pLoF	0.001	female	high_outlier	95ci_2.5p	Familial hypercholesterolaemia	50	14	111	64	50	14	0.266	9.797	2.115	5.488	2.114	5.489	10.66862654	6.484483654	17.55260687	7.47E-10
Apolipoprotein B (ApoB)	raw	pLoF	0.001	bothsexes	high_outlier	95ci_1p	Familial hypercholesterolaemia	81	11	105	93	82	11	0.4	2	0.613	1.553	0.544	1.669	13.02899799	7.674852486	22.11831288	3.56E-09
Apolipoprotein B (ApoB)	raw	pLoF	0.001	female	high_outlier	95ci_2.5p	Familial hypercholesterolaemia	50	13	105	63	50	13	0.4	2	0.627	1.566	0.626	1.567	9.926563408	5.945330749	16.57378963	6.27E-09
LDL cholesterol (LDL-C)	raw	pLoF	0.001	male	high_outlier	95ci_0.01	Familial hypercholesterolaemia	32	3	111	36	33	3	0.266	9.797	1.894	5.241	0.955	7.282	1040.84375	353.4034027	3065.493155	6.92E-09
Apolipoprotein B (ApoB)	raw	pLoF	0.001	bothsexes	high_outlier	clinical	Familial hypercholesterolaemia	96	6	105	103	97	6	0.4	2	0.53	1.82	0.529	1.821	28.85391198	14.40650748	57.78973408	1.18E-07
LDL cholesterol (LDL-C)	raw	pLoF	0.001	female	high_outlier	95ci_1p	Familial hypercholesterolaemia	50	8	111	58	50	8	0.266	9.797	2.115	5.488	1.884	5.907	15.25266409	8.144640939	28.56402924	1.65E-07
Apolipoprotein B (ApoB)	raw	pLoF	0.001	bothsexes	high_outlier	95ci_0.1p	Familial hypercholesterolaemia	81	4	105	86	82	4	0.4	2	0.613	1.553	0.432	1.883	47.68058839	20.44567442	111.194107	2.23E-06
Cholesterol	raw	pLoF	0.001	female	high_outlier	95ci_1p	Familial hypercholesterolaemia	55	7	111	62	55	7	0.601	15.46	3.876	8.264	3.553	8.802	12.14006312	6.267885528	23.51369244	3.91E-06
Apolipoprotein B (ApoB)	raw	pLoF	0.001	male	high_outlier	95ci_2.5p	Familial hypercholesterolaemia	32	8	105	41	33	8	0.4	2	0.599	1.537	0.598	1.538	9.577171904	4.996330609	18.35791681	6.55E-06
Apolipoprotein B (ApoB)	raw	pLoF	0.001	female	high_outlier	95ci_0.1p	Familial hypercholesterolaemia	50	3	105	53	50	3	0.4	2	0.627	1.566	0.435	1.902	58.28423645	21.77584264	156.0009536	2.52E-05
LDL cholesterol (LDL-C)	raw	pLoF	0.001	female	high_outlier	95ci_0.1p	Familial hypercholesterolaemia	50	3	111	53	50	3	0.266	9.797	2.115	5.488	1.456	6.887	57.81131707	21.6005797	154.7249392	2.58E-05
Apolipoprotein B (ApoB)	raw	pLoF	0.001	male	high_outlier	clinical	Familial hypercholesterolaemia	36	3	105	40	37	3	0.4	2	0.53	1.73	0.529	1.821	54.7076972	20.25400493	147.7698926	3.16E-05
Cholesterol	raw	pLoF	0.001	female	high_outlier	95ci_2.5p	Familial hypercholesterolaemia	55	9	111	64	55	9	0.601	15.46	3.876	8.264	3.875	8.265	6.231442821	3.447712105	11.26279644	3.65E-05
LDL cholesterol (LDL-C)	raw	pLoF	0.001	bothsexes	high_outlier	95ci_0.01	Familial hypercholesterolaemia	81	2	111	84	82	2	0.266	9.797	2.003	5.387	1.006	7.674	242.9289289	72.58457866	813.0441149	3.78E-05
Apolipoprotein B (ApoB)	raw	pLoF	0.001	male	high_outlier	95ci_1p	Familial hypercholesterolaemia	32	5	105	38	33	5	0.4	2	0.599	1.537	0.531	1.648	14.98337189	6.786649896	33.07986071	3.91E-05
LDL cholesterol (LDL-C)	raw	pLoF	0.001	bothsexes	high_outlier	clinical	Familial hypercholesterolaemia	12	98	111	111	12	99	0.266	9.797	1.001	3	0.986	3.001	2.976397125	1.799695567	4.922465781	4.06E-05
Triglyceride (TG)	raw	pLoF	0.001	bothsexes	high_outlier	clinical	Severe hypertriglyceridaemia	118	129	247	247	118	129	0.231	11.278	0.231	1.7		1.701	1.614972553	1.30959913	1.991553206	0.000109467632
Triglyceride (TG)	raw	pLoF	0.001	female	high_outlier	clinical	Severe hypertriglyceridaemia	72	64	247	136	72	64	0.231	11.278	0.231	1.7		1.701	1.872399164	1.411334629	2.484087445	0.000215611761
Apolipoprotein B (ApoB)	raw	pLoF	0.001	female	high_outlier	95ci_1p	Familial hypercholesterolaemia	50	5	105	55	50	5	0.4	2	0.627	1.566	0.558	1.684	9.521728634	4.398293401	20.61329427	0.000275458097
Cholesterol	raw	pLoF	0.001	bothsexes	high_outlier	clinical	Familial hypercholesterolaemia	15	95	111	111	16	95	0.601	15.46	0.601	5		5.001	2.358575874	1.493345687	3.725112143	5.43E-04
Apolipoprotein B (ApoB)	raw	pLoF	0.001	male	high_outlier	95ci_0.1p	Familial hypercholesterolaemia	32	2	105	35	33	2	0.4	2	0.599	1.537	0.43	1.857	60.95588235	18.2586996	203.4985884	0.000585832586
Triglyceride (TG)	raw	pLoF	0.001	bothsexes	high_outlier	95ci_2.5p	Severe hypertriglyceridaemia	227	16	247	243	227	16	0.231	11.278	0.607	4.439	0.606	4.44	2.681711277	1.751820922	4.1052007	0.000617362190
Apolipoprotein B (ApoB)	raw	pLoF	0.001	female	high_outlier	clinical	Familial hypercholesterolaemia	60	3	105	63	60	3	0.4	2	0.53	1.82	0.529	1.821	18.49271583	6.971041005	49.05731273	0.000696716683
Cholesterol	raw	pLoF	0.001	female	high_outlier	95ci_0.1p	Familial hypercholesterolaemia	55	2	111	57	55	2	0.601	15.46	3.876	8.264	2.93	10.127	34.75292051	10.57725541	114.1851489	0.001701724767
Cholesterol	raw	pLoF	0.001	male	high_outlier	clinical	Familial hypercholesterolaemia	6	39	111	46	7	39	0.601	15.46	0.601	5		5.001	3.290521884	1.599485954	6.769383775	0.001875871679
Sex hormone-binding globulin	raw	pLoF	0.001	male	low_outlier	clinical	Congenital hypothyroidism	23	18	122	41	23	18	0.39	241.92	27	77	26.99	128.01	2.682300805	1.598314617	4.501452677	0.001984666424
Glycated haemoglobin (HbA1	raw	pLoF	0.001	female	high_outlier	95ci_2.5p	Monogenic diabetes	1159	48	2491	1315	1265	50	15	515.2	27.9	47.4	27.8	47.5	1.590263839	1.246706412	2.028496087	0.002016188285
Triglyceride (TG)	raw	pLoF	0.001	female	high_outlier	95ci_1p	Severe hypertriglyceridaemia	124	6	247	130	124	6	0.231	11.278	0.586	3.779	0.518	4.647	4.609553051	2.315464878	9.176550046	0.002522980141
Testosterone	raw	pLoF	0.001	male	high_outlier	95ci_2.5p	Differences in sex development	240	15	525	271	252	19	0.35	54.342	5.87	20.193	5.869	20.194	2.37962696	1.53E+00	3.689329065	2.70E-03
LDL cholesterol (LDL-C)	raw	pLoF	0.001	male	high_outlier	clinical	Familial hypercholesterolaemia	5	40	111	46	5	41	0.266	9.797	1.001	3	0.986	3.001	3.370517864	1.544648507	7.354676884	0.002986447931
Cholesterol	raw	pLoF	0.001	male	high_outlier	95ci_0.01	Familial hypercholesterolaemia	32	1	111	34	33	1	0.601	15.46	3.384	7.799	1.995	10.627	306.7775735	55.07978248	1708.657431	0.003552436632
Apolipoprotein B (ApoB)	raw	pLoF	0.001	male	high_outlier	95ci_0.01	Familial hypercholesterolaemia	32	1	105	34	33	1	0.4	2	0.599	1.537	0.404	1.966	304.7794118	54.72102473	1697.528332	0.003575681538
Testosterone	raw	pLoF	0.001	bothsexes	high_outlier	95ci_1p	Differences in sex development	461	12	525	505	489	16	0.35	54.342	0.45	18.502	0.396	20.751	2.477607589	1.530364159	4.011162525	0.004615856084
Testosterone	raw	pLoF	0.001	male	high_outlier	95ci_1p	Differences in sex development	240	8	525	264	252	12	0.35	54.342	5.87	20.193	4.847	22.414	3.177232487	1.76E+00	5.745980824	4.84E-03

Table S2.6: Significant tail-status and continuous biomarker associations.

pheno	gene_symbo	csq	iro	pos	varid	effe	effe	is_voi	is_sig	is_sig	is_sig	is_sig	nlog10pval_cts	nlog10pval_low	nlog10pval_high	nlog10pval_eith	beta_cts
lipoprotein-a_qced	SLC22A2	other_missense	6	160,249,250	chr6:160249250:A:C	C	0.8991	True	False	True	True	True	1.95	7.17	25.01	31.89	0.0187538
lipoprotein-a_qced	IGF2R	other_missense	6	160,061,617	chr6:160061617:A:G	G	0.0038	True	False	True	True	True	0.19	6.48	11.03	18.71	0.0165152
lipoprotein-a_qced	SLC22A1	other_missense	6	160,136,215	chr6:160136215:C:T	T	0.0047	True	False	True	False	True	1.30	12.54	5.48	18.17	-0.0637867
lipoprotein-a_qced	LPA	other_missense	6	160,531,748	chr6:160531748:C:T	T	0.0032	True	False	True	False	True	1.56	12.29	4.57	16.39	0.0866778
lipoprotein-a_qced	LPA	other_missense	6	160,586,587	chr6:160586587:G:T	T	0.0014	True	False	True	True	True	0.04	6.52	9.12	17.50	-0.0063536
lipoprotein-a_qced	SLC22A1	other_missense	6	160,139,848	chr6:160139848:CATG	C	0.1738	True	False	True	False	True	5.02	18.28	0.91	13.86	0.0259817
shbg_qced	MPDU1	other_missense	17	7587492	chr17:7587492:G:A	A	0.1740	True	False	True	False	True	2.06	12.07	2.88	13.82	0.0055447
lipoprotein-a_qced	PLG	damaging_missense	6	160,714,844	chr6:160714844:A:G	G	0.0018	True	False	True	False	True	1.72	9.41	2.16	10.24	-0.1233460
vitamin-d_qced	POR	other_missense	7	75985688	chr7:75985688:C:T	T	0.2840	True	False	False	False	True	0.60	3.58	5.95	10.04	0.0018652
lipoprotein-a_qced	SLC22A3	other_missense	6	160,348,796	chr6:160348796:C:G	G	0.0018	True	False	False	True	True	3.94	2.29	6.30	8.23	-0.1986590
hba1c_qced	H2BC6	other_missense	6	26183874	chr6:26183874:G:A	A	0.2554	True	False	True	False	False	4.26	10.23	0.02	6.04	0.0022867
hba1c_qced	SPTB	other_missense	14	64800751	chr14:64800751:T:C	C	0.2958	True	False	True	False	False	6.05	10.77	1.36	3.28	0.0026615
vitamin-d_qced	MDH2	other_missense	7	76048186	chr7:76048186:C:T	T	0.3000	True	False	False	False	True	0.38	3.42	3.91	7.97	0.0012883
aspartate-aminotransferase_qce	MTARC1	other_missense	1	220796686	chr1:220796686:A:G	G	0.7032	True	False	False	True	False	6.04	0.98	10.42	3.76	0.0048488
cystatin-c-serum_qced	SLC6A19	other_missense	5	1212338	chr5:1212338:G:A	A	0.0046	True	False	True	False	False	5.20	9.77	0.18	5.14	-0.0161877
lipoprotein-a_qced	SLC22A1	other_missense	6	160,132,282	chr6:160132282:C:T	T	0.0018	True	False	False	False	True	0.67	3.30	3.63	7.23	-0.0644203
total-bilirubin_qced		other_missense	2	233713766	chr2:233713766:G:C	C	0.0038	True	False	False	True	False	5.91	0.10	9.28	4.43	-0.0359503
lipoprotein-a_qced	IGF2R	other_missense	6	160,034,488	chr6:160034488:G:C	C	0.0014	True	False	False	False	True	0.98	3.95	2.91	6.89	-0.0951677
alanine-aminotransferase_qced	SLC39A4	other_missense	8	144416033	chr8:144416033:G:A	A	0.0364	True	False	True	False	False	6.24	9.11	0.38	4.19	-0.0194737
aspartate-aminotransferase_qce	MICB	other_missense	6	31505784	chr6:31505784:A:G	G	0.3065	True	False	True	False	False	6.26	8.79	0.13	4.74	-0.0049106
creatinine-urine_qced	CCS	other_missense	11	66605818	chr11:66605818:G:C	C	0.0022	True	False	False	False	True	0.70	3.43	3.14	6.77	-0.0298058
triglyceride_qced	PNPLA3	other_missense	22	43928847	chr22:43928847:C:G	G	0.2057	True	False	False	True	False	2.87	0.06	8.00	5.02	-0.0064417
alanine-aminotransferase_qced	HNF1A	other_missense	12	120997672	chr12:120997672:G:A	A	0.3428	True	False	False	False	True	0.25	2.62	3.54	6.61	0.0009106
glucose_qced	CAPN9	damaging_missense	1	230772053	chr1:230772053:C:T	T	0.0031	True	False	False	False	True	0.71	2.36	3.97	6.37	0.0070039
lipoprotein-a_qced	PRKN	other_missense	6	161,386,823	chr6:161386823:C:G	G	0.1660	True	False	False	True	False	5.10	0.90	8.60	2.92	-0.0267699
c-reactive-protein_qced	LEPR	other_missense	1	65618011	chr1:65618011:G:A	A	0.0013	True	False	True	False	False	4.23	6.65	0.57	5.15	0.1942190
alanine-aminotransferase_qced	HNF1A	other_missense	12	120997624	chr12:120997624:G:A	A	0.2864	True	False	False	False	True	1.29	1.58	4.43	6.33	0.0031372
aspartate-aminotransferase_qce	RP1L1	other_missense	8	10612558	chr8:10612558:C:T	T	0.5107	True	False	True	False	False	4.38	8.36	0.49	3.38	-0.0037062
urea_qced	ANKRD36C	other_missense	2	95855514	chr2:95855514:A:C	C	0.0038	True	False	False	True	False	1.62	0.34	6.64	5.23	0.0145203
cystatin-c-serum_qced	CEP350	other_missense	1	180020607	chr1:180020607:G:C	C	0.6509	True	False	True	False	False	5.22	8.20	0.51	3.42	0.0022899
aspartate-aminotransferase_qce	RP1L1	other_missense	8	10608972	chr8:10608972:G:A	A	0.4902	True	False	True	False	False	4.83	7.85	0.15	4.06	-0.0039206
c-reactive-protein_qced	TCF19	other_missense	6	31161533	chr6:31161533:C:T	T	0.1527	True	False	True	False	False	5.15	6.97	0.20	4.78	-0.0221192
hba1c_qced	ABO	other_missense	9	133255935	chr9:133255935:G:T	T	0.0599	True	False	True	False	False	5.17	7.14	0.08	4.36	-0.0046853
hba1c_qced	SPTB	other_missense	14	64804932	chr14:64804932:A:G	G	0.3102	True	False	True	False	False	5.48	7.87	1.71	1.92	0.0024822
aspartate-aminotransferase_qce	RP1L1	other_missense	8	10608238	chr8:10608238:T:C	C	0.5177	True	False	True	False	False	4.58	7.59	0.24	3.63	0.0038042
aspartate-aminotransferase_qce	RP1L1	other_missense	8	10607587	chr8:10607587:C:T	T	0.4816	True	False	True	False	False	4.44	7.57	0.22	3.66	-0.0037397
gamma-glutamyltransferase_qce	P2RX7	other_missense	12	121167552	chr12:121167552:G:A	A	0.2529	True	False	True	False	False	3.74	7.64	0.48	3.19	0.0084011
cystatin-c-serum_qced	NEXN	other_missense	1	77926761	chr1:77926761:G:A	A	0.2107	True	False	False	True	False	6.13	1.62	7.57	2.05	-0.0029297
apolipoprotein-a_qced	ANXA9	other_missense	1	150986360	chr1:150986360:A:G	G	0.1594	True	False	True	False	False	5.43	7.52	1.20	2.19	0.0033959
glucose_qced	KCNJ11	other_missense	11	17388025	chr11:17388025:T:C	C	0.6435	True	False	False	True	False	5.78	1.74	7.33	1.77	-0.0030262
aspartate-aminotransferase_qce	RP1L1	other_missense	8	10611723	chr8:10611723:A:G	G	0.5524	True	False	True	False	False	4.15	7.10	0.18	3.51	-0.0036129
phosphate_qced	AFAP1	other_missense	4	7819068	chr4:7819068:G:A	A	0.2025	True	False	False	True	False	5.60	0.73	7.26	2.80	-0.0027525
triglyceride_qced	VARS2	other_missense	6	30920384	chr6:30920384:T:C	C	0.5106	True	False	True	False	False	5.90	6.85	3.23	0.67	0.0078312
glucose_qced	KCNJ11	other_missense	11	17387083	chr11:17387083:C:T	T	0.6434	True	False	False	True	False	5.81	1.79	7.22	1.69	-0.0030348
glucose_qced	ABCC8	other_missense	11	17396930	chr11:17396930:C:A	A	0.6439	True	False	False	True	False	6.02	1.84	7.05	1.62	-0.0030940
c-reactive-protein_qced	SLC2A2	other_missense	3	171014511	chr3:171014511:G:A	A	0.1223	True	False	True	False	False	6.15	7.06	1.92	1.41	0.0267721

Table S2.7: Odds of ICD-10 codes with Mahalanobis distance.

chapter	desc	is_chapter	n	cases	prevalence	mahal_pval_	mahal_pval_	odds_ratio	pval_fisher	pval_fisher_	or_95ci_low	or_95ci_upp	or_90ci_low	or_90ci_upp
Endocrine, nutrit	Diabetes mellitus	FALSE	10143	15368	0.08285752797	0.01	0.05	2.814465492	1.51E-115	1.27E-115	2.597456398	3.049604996	2.63118175	3.010516475
Endocrine, nutrit	Diabetes mellitus	FALSE	4765	15368	0.08285752797	0.001	0.01	3.3016444	1.20E-79	1.20E-79	2.955779181	3.68798042	3.008835821	3.622948009
Endocrine, nutrit	Endocrine, nutritional and metabol	TRUE	2942	61621	0.3322334546	0	0.001	3.102622862	3.46E-76	1.86E-76	2.747350301	3.503837359	2.801594554	3.435996336
Endocrine, nutrit	Diabetes mellitus	FALSE	2942	15368	0.08285752797	0	0.001	4.186380201	3.26E-76	3.26E-76	3.660333802	4.788027578	3.740215968	4.685766634
Endocrine, nutrit	Endocrine, nutritional and metabol	TRUE	10143	61621	0.3322334546	0.01	0.05	1.766952612	1.16E-67	6.17E-68	1.658472499	1.882528373	1.675452887	1.863449315
Endocrine, nutrit	Metabolic disorders	FALSE	2942	40093	0.2161639035	0	0.001	2.965743782	1.18E-65	1.04E-65	2.63122092	3.342796537	2.682339406	3.279091438
Genitourinary sy	Renal failure	FALSE	2942	14943	0.08056611403	0	0.001	3.885017696	1.67E-65	1.67E-65	3.38546312	4.458285901	3.461212887	4.360714868
Circulatory syste	Circulatory system	TRUE	2942	86231	0.4649198005	0	0.001	2.775214209	1.55E-58	9.34E-59	2.43606285	3.161582594	2.487651626	3.096017876
Endocrine, nutrit	Endocrine, nutritional and metabol	TRUE	4765	61621	0.3322334546	0.001	0.01	2.058978127	2.46E-52	1.55E-52	1.877861881	2.257562694	1.905867362	2.22438928
Circulatory syste	Hypertensive diseases	FALSE	2942	58364	0.3146731365	0	0.001	2.53701235	4.05E-52	2.39E-52	2.25184152	2.858296912	2.295426767	2.804023964
Genitourinary sy	Renal failure	FALSE	4765	14943	0.08056611403	0.001	0.01	2.714289452	1.32E-50	9.96E-51	2.412306693	3.05407569	2.458487064	2.996707745
Circulatory syste	Hypertensive diseases	FALSE	10143	58364	0.3146731365	0.01	0.05	1.629631104	3.67E-49	2.31E-49	1.528714655	1.737209444	1.544507299	1.719446413
Circulatory syste	Other forms of heart disease	FALSE	2942	28488	0.1535948241	0	0.001	2.679231706	1.20E-45	8.71E-46	2.359429768	3.042380252	2.408142995	2.980837328
Circulatory syste	Hypertensive diseases	FALSE	4765	58364	0.3146731365	0.001	0.01	1.934510525	2.42E-43	1.58E-43	1.764005261	2.12149649	1.79036784	2.09025815
Genitourinary sy	Renal failure	FALSE	10143	14943	0.08056611403	0.01	0.05	1.920723771	5.83E-39	4.18E-39	1.753491876	2.103904702	1.779361317	2.073316851
Circulatory syste	Circulatory system	TRUE	10143	86231	0.4649198005	0.01	0.05	1.502159573	2.20E-36	1.34E-36	1.409550227	1.600853478	1.424044679	1.584559401
Endocrine, nutrit	Metabolic disorders	FALSE	10143	40093	0.2161639035	0.01	0.05	1.578109333	3.85E-36	2.09E-36	1.472912563	1.690819354	1.489339683	1.672169952
Circulatory syste	Circulatory system	TRUE	4765	86231	0.4649198005	0.001	0.01	1.81010571	6.64E-36	4.29E-36	1.64721107	1.989109192	1.67237506	1.959179348
Digestive systen	Digestive system	TRUE	2942	99017	0.5338563149	0	0.001	2.178285984	6.66E-34	4.28E-34	1.910236926	2.483948335	1.950993287	2.432058511
Endocrine, nutrit	Metabolic disorders	FALSE	4765	40093	0.2161639035	0.001	0.01	1.80303282	7.28E-30	4.17E-30	1.634803273	1.988574041	1.660750933	1.957504455
Musculoskeletal	Musculoskeletal system and conn	TRUE	2942	77594	0.4183528777	0	0.001	1.997994773	9.24E-30	5.12E-30	1.770684803	2.254485443	1.805403703	2.211130456
Genitourinary sy	Genitourinary system	TRUE	2942	66008	0.355886238	0	0.001	1.94111027	1.65E-27	1.09E-27	1.723399639	2.18632347	1.756678297	2.144905579
Circulatory syste	Other forms of heart disease	FALSE	4765	28488	0.1535948241	0.001	0.01	1.789629703	5.45E-24	3.82E-24	1.607441994	1.992466592	1.63542954	1.958368976
Digestive systen	Diseases of oesophagus, stomach	FALSE	2942	43207	0.2329532282	0	0.001	1.913784625	2.37E-23	1.70E-23	1.691529522	2.165242487	1.725437414	2.122691649
Circulatory syste	Ischaemic heart diseases	FALSE	2942	21114	0.1138374444	0	0.001	2.189469986	4.57E-23	3.71E-23	1.89558851	2.528913209	1.940026562	2.470986179
Musculoskeletal	Inflammatory polyarthropathies	FALSE	2942	19293	0.1040194096	0	0.001	2.194210508	8.66E-22	6.36E-22	1.891697049	2.545100844	1.937357177	2.485117258
Digestive systen	Other diseases of intestines	FALSE	2942	55145	0.2973176978	0	0.001	1.740379138	8.93E-19	5.72E-19	1.543034307	1.96296319	1.573181895	1.925346048
Endocrine, nutrit	Obesity and other hyperalimentati	FALSE	4765	14590	0.07866289257	0.001	0.01	1.882120018	9.11E-18	5.79E-18	1.646032737	2.152068839	1.681887375	2.106190826
Genitourinary sy	Other diseases of urinary system	FALSE	2942	21485	0.115837714	0	0.001	1.986595085	1.44E-17	1.01E-17	1.714497359	2.301875831	1.755585204	2.24800256
Circulatory syste	Other forms of heart disease	FALSE	10143	28488	0.1535948241	0.01	0.05	1.422043089	2.18E-17	1.17E-17	1.314601969	1.538265264	1.331311327	1.518958417
Circulatory syste	Ischaemic heart diseases	FALSE	4765	21114	0.1138374444	0.001	0.01	1.726877724	3.10E-17	2.16E-17	1.531784469	1.946818716	1.561594115	1.909655424
Genitourinary sy	Genitourinary system	TRUE	4765	66008	0.355886238	0.001	0.01	1.480783345	1.77E-16	1.07E-16	1.349909994	1.624344827	1.370142609	1.60035846
Digestive systen	Digestive system	TRUE	4765	99017	0.5338563149	0.001	0.01	1.455037709	4.85E-15	2.78E-15	1.323225071	1.59998082	1.343581782	1.575739387
Musculoskeletal	Inflammatory polyarthropathies	FALSE	4765	19293	0.1040194096	0.001	0.01	1.682760884	1.38E-14	8.08E-15	1.485067226	1.906771723	1.51520818	1.868841676
Circulatory syste	Ischaemic heart diseases	FALSE	10143	21114	0.1138374444	0.01	0.05	1.418182294	6.59E-14	3.89E-14	1.29880039	1.54853743	1.317292732	1.526798843
Genitourinary sy	Genitourinary system	TRUE	10143	66008	0.355886238	0.01	0.05	1.276702183	1.30E-13	7.36E-14	1.197374629	1.361285286	1.209787604	1.347317876
Musculoskeletal	Inflammatory polyarthropathies	FALSE	10143	19293	0.1040194096	0.01	0.05	1.420984102	3.60E-13	2.17E-13	1.297270984	1.556495013	1.316408467	1.533867237
Musculoskeletal	Musculoskeletal system and conn	TRUE	4765	77594	0.4183528777	0.001	0.01	1.402701776	7.07E-13	3.78E-13	1.279377	1.537914369	1.298446752	1.515327655
Endocrine, nutrit	Obesity and other hyperalimentati	FALSE	2942	14590	0.07866289257	0	0.001	1.943391347	1.64E-12	1.14E-12	1.638304374	2.305291977	1.683908422	2.242859457
Musculoskeletal	Musculoskeletal system and conn	TRUE	10143	77594	0.4183528777	0.01	0.05	1.252539995	3.43E-12	1.90E-12	1.175797442	1.33429142	1.187810627	1.320796769

Table S2.8: Mahalanobis distance gene burden associations.

chrom	gene_symb	max_	Grou	Pvalue_Bur	BETA_Burd	SE_Burden	MAC	Number_rar	Number_ulti	Pvalue_Burden_cauchy	Pvalue_Burden_minpva
20	CST3	0.01	pLoF	1.18E-66	0.06436952604	0.003732275144	92	3	10	2.80E-299	1.00E-300
8	GPT	0.01	pLoF	2.13E-21	0.02150027067	0.002263496244	252	4	38	2.80E-299	1.00E-300
11	APOC3	0.01	pLoF	4.28E-18	0.01067265388	0.001230836321	939	2	5	2.80E-299	1.00E-300
2	APOB	0.01	pLoF	6.14E-15	0.01328812594	0.001703383553	448	6	93	3.41E-110	1.22E-111
10	GOT1	0.01	pLoF	3.28E-13	0.04404147959	0.006047766268	35	0	15	3.22E-112	1.15E-113
17	SHBG	0.01	pLoF	4.29E-13	0.01440887435	0.001988468919	334	2	22	4.98E-226	1.78E-227
6	GPLD1	0.01	pLoF	6.95E-10	0.005559834348	0.000901516835	1856	5	47	2.80E-299	1.00E-300
6	ZBTB2	0.01	pLoF	4.12E-08	0.1386536881	0.02527635066	2	0	2	0.8332746337	0.004609543756
9	ABCA1	0.01	pLoF	7.44E-08	0.01413085358	0.002626470486	186	2	73	1.77E-161	6.34E-163
4	ALB	0.01	pLoF	2.60E-07	0.03152821043	0.006121384346	34	1	16	5.51E-146	1.97E-147
1	ALPL	0.01	pLoF	1.31E-06	0.009792020332	0.002023914859	317	3	31	1.34E-83	4.79E-85
11	APOA1	0.01	pLoF	1.49E-06	0.04174026661	0.008673017129	17	0	11	3.15E-36	1.12E-37
17	SLC4A1	0.01	pLoF	2.20E-06	0.05350225699	0.0113015859	10	0	9	3.39E-17	2.19E-18
17	NF1	0.01	pLoF	2.54E-06	0.02161643647	0.004594376613	61	0	52	3.97E-11	1.42E-12
19	ELAVL1	0.01	pLoF	8.29E-06	0.1589087277	0.03565026852	1	0	1	0.04492876735	0.002787790312
10	VAX1	0.01	pLoF	2.35E-05	0.06751980512	0.01596656202	5	0	3	0.5297518475	0.004773946226
10	C10orf105	0.01	pLoF	3.52E-05	0.1046068157	0.02528733715	2	0	2	0.9859972919	0.0007985253742
6	TNFAIP3	0.01	pLoF	5.11E-05	0.08353924343	0.02062350722	3	0	4	0.004354646733	0.000300081361
X	AVPR2	0.01	pLoF	8.16E-05	0.04692190979	0.01191037609	9	0	2	0.0991561957	0.005287563015
22	SF3A1	0.01	pLoF	8.19E-05	0.07035299051	0.01786133208	4	0	4	0.01216493552	0.0005841732801
3	TBL1XR1	0.01	pLoF	0.000114976909	0.1372216681	0.03558103173	1	0	1	0.998171531	0.0103735364
7	TRA2A	0.01	pLoF	0.000155517027	0.02258461381	0.005971463667	36	0	23	0.05835844808	0.002323316894
1	MYOG	0.01	pLoF	0.000174042895	0.0948797609	0.02527438815	2	0	2	0.02236402996	0.0009800192575
12	MFAP5	0.01	pLoF	0.000179938034	0.01637888996	0.004372793893	67	2	16	0.05966002491	0.004844869232
19	FCGRT	0.01	pLoF	0.000184815304	0.04230992192	0.01131609254	10	0	9	3.43E-34	1.22E-35
2	TMEM131	0.01	pLoF	0.000194270088	0.0169645861	0.004552597389	62	1	26	0.02838308609	0.002003308652
6	CRYBG1	0.01	pLoF	0.000198929887	-0.00458213796	0.001231634497	857	13	100	0.682584908	0.04774241075
14	CKB	0.01	pLoF	0.000210913235	0.04187305361	0.01130002075	10	0	7	0.7543026539	0.02276727019
4	LEF1	0.01	pLoF	0.000251515181	0.05346683915	0.01460558402	6	0	5	0.09028289789	0.005766429956
8	REEP4	0.01	pLoF	0.000252057558	0.02352663361	0.006427760529	31	0	14	0.0966281979	0.007544038602
3	IGF2BP2	0.01	pLoF	0.000253835363	0.05844757822	0.01597644705	5	0	5	0.007394635487	0.0006235624102
22	TANGO2	0.01	pLoF	0.000290105391	0.0254680149	0.007027639034	26	0	15	0.0762602009	0.004623154158
19	LDLR	0.01	pLoF	0.000305468528	0.01992926645	0.005519627825	42	1	23	5.29E-05	1.89E-06
13	GRK1	0.01	pLoF	0.000352636895	0.01529805872	0.004281325508	70	1	17	0.376324718	0.01370766114
3	DVL3	0.01	pLoF	0.000398720456	0.02987356356	0.008436645108	18	0	12	0.3246352996	0.01145842605
2	FHL2	0.01	pLoF	0.000410184197	0.03374366629	0.009549797253	14	0	7	0.3917353729	0.04194740218
11	FEN1	0.01	pLoF	0.000433731557	0.04196420884	0.01192618902	9	0	7	0.2490588332	0.02205157632
3	TSC22D2	0.01	pLoF	0.00044309416	0.02959114074	0.008423348992	18	1	3	0.1390534558	0.01806149639
19	ANKLE1	0.01	pLoF	0.000447428692	-0.01111993711	0.003167709912	128	4	30	0.6701116711	0.1001754416
1	GJA5	0.01	pLoF	0.000461192351	0.03229066562	0.009219742941	15	1	3	0.4915452455	0.03994841589

Table S2.9: Single-biomarker gene burdens associations. Only genes with Mahalanobis distance burden association $P < 0.001$ for at least one annotation category were tested for association with single biomarkers.

pheno	chrom	gene_symb	max_MA	Grou	Pvalue_Bur	nlog10Pvalu	BETA_Burd	SE_Burden	MAC	Number_ran	Number_ult
alanine_aminotransferase_	8	GPT	0.01	pLoF	1.00E-300	300	-0.06757838845	0	540	6	52
alkaline_phosphatase_qcec	6	GPLD1	0.01	pLoF	1.00E-300	300	-0.03730594509	0	3555	10	55
cystatin_c_serum_qced	20	CST3	0.01	pLoF	1.00E-300	300	-0.1375280873	0	178	4	14
triglyceride_qced	11	APOC3	0.01	pLoF	1.00E-300	300	-0.0527533801	0	1775	4	4
hdl_cholesterol_qced	11	APOC3	0.01	pLoF	1.74E-283	282.7597152	0.03758957311	0.001044744427	1624	4	4
shbg_qced	17	SHBG	0.01	pLoF	1.78E-227	226.750307	-0.05756130488	0.001787629951	635	4	30
apolipoprotein_a_qced	11	APOC3	0.01	pLoF	1.12E-197	196.9513691	0.03175697355	0.001058719505	1577	4	4
apolipoprotein_a_qced	9	ABCA1	0.01	pLoF	6.34E-163	162.1982105	-0.06005689387	0.002207900052	332	3	103
albumin_qced	4	ALB	0.01	pLoF	1.97E-147	146.7057043	-0.102536613	0.003965346959	102	3	33
hdl_cholesterol_qced	9	ABCA1	0.01	pLoF	7.15E-132	131.14561	-0.05387454581	0.002204752164	333	3	104
aspartate_aminotransferase	10	GOT1	0.01	pLoF	1.15E-113	112.9386525	-0.104827163	0.004626405599	75	1	21
triglyceride_qced	2	APOB	0.01	pLoF	1.22E-111	110.9141782	-0.03170511351	0.001412117413	815	10	139
alkaline_phosphatase_qcec	1	ALPL	0.01	pLoF	4.79E-85	84.31943884	-0.03184167737	0.001629367945	612	5	40
gamma_glutamyltransferas	22	GGT1	0.01	pLoF	3.61E-82	81.44201121	-0.06306941885	0.003284654393	149	3	23
hdl_cholesterol_qced	16	CETP	0.01	pLoF	6.83E-78	77.16581469	0.04237978995	0.002268378181	312	10	35
lipoprotein_a_qced	6	LPA	0.01	pLoF	3.57E-63	62.44784307	-0.01675141285	0.000998445578	1700	13	112
urate_qced	11	SLC22A12	0.01	pLoF	3.42E-44	43.46565514	-0.0651858905	0.004674837188	76	1	34
cystatin_c_serum_qced	16	PKD1	0.01	pLoF	3.65E-43	42.43719773	0.0509289392	0.003697470179	158	2	87
apolipoprotein_a_qced	16	CETP	0.01	pLoF	1.57E-39	38.80437558	0.03063581562	0.002328632659	296	10	35
alkaline_phosphatase_qcec	17	ASGR1	0.01	pLoF	9.90E-39	38.00433388	0.04424498758	0.003399231058	139	3	18
apolipoprotein_a_qced	11	APOA1	0.01	pLoF	1.12E-37	36.94949136	-0.09707761207	0.007566850513	28	0	15
calcium_qced	4	ALB	0.01	pLoF	1.82E-37	36.73970217	-0.05072624696	0.003965513993	102	3	33
vitamin_d_qced	2	APOB	0.01	pLoF	2.78E-36	35.55670406	0.01797755874	0.001429238102	795	10	139
albumin_qced	19	FCGRT	0.01	pLoF	1.22E-35	34.91228431	-0.09605890237	0.00770898503	27	0	17
creatinine_serum_qced	16	PKD1	0.01	pLoF	1.22E-34	33.91239891	0.04556949075	0.003712174447	157	2	87
total_protein_qced	4	ALB	0.01	pLoF	1.69E-32	31.77309954	-0.04706540881	0.003964931674	102	3	33
urea_qced	16	PKD1	0.01	pLoF	3.72E-30	29.42988486	0.04218824196	0.003697396604	158	2	87
total_protein_qced	19	FCGRT	0.01	pLoF	2.03E-28	27.69257069	-0.08526556147	0.007711517749	27	0	17
calcium_qced	3	CASR	0.01	pLoF	9.54E-28	27.02025125	0.07611646439	0.006972191972	33	0	28
direct_bilirubin_qced	2	APOB	0.01	pLoF	1.30E-26	25.88466595	0.01588278195	0.001487569587	734	9	132
hdl_cholesterol_qced	16	LCAT	0.01	pLoF	4.86E-26	25.31301827	-0.05809422182	0.005504433176	53	1	11
testosterone_qced	17	SHBG	0.01	pLoF	6.14E-24	23.21154854	-0.0232464475	0.002304008475	616	3	30
apolipoprotein_a_qced	16	LCAT	0.01	pLoF	2.36E-20	19.62748822	-0.05087466665	0.005503085	53	1	11
hba1c_qced	7	GIGYF1	0.01	pLoF	2.01E-19	18.69758297	0.03213471956	0.003565405656	127	1	73
direct_bilirubin_qced	17	SLC4A1	0.01	pLoF	2.19E-18	17.6594434	0.07301816073	0.008347773624	23	0	21
hba1c_qced	17	SLC4A1	0.01	pLoF	2.70E-18	17.56896104	-0.08463359083	0.009701823775	17	0	16
hdl_cholesterol_qced	11	APOA1	0.01	pLoF	1.50E-17	16.82533225	-0.06452077524	0.007566176396	28	0	15
vitamin_d_qced	11	CYP2R1	0.01	pLoF	1.95E-15	14.70907596	-0.0226508011	0.00285122653	233	6	34
total_bilirubin_qced	17	SLC4A1	0.01	pLoF	6.30E-14	13.20037687	0.05888437689	0.007849570234	26	0	23
c_reactive_protein_qced	1	CRP	0.01	pLoF	4.78E-13	12.3203912	-0.04135392561	0.005718696796	49	2	1
glucose_qced	7	GIGYF1	0.01	pLoF	1.32E-12	11.88007305	0.02587755793	0.003648630763	121	1	72
cystatin_c_serum_qced	17	NF1	0.01	pLoF	1.42E-12	11.84801768	0.0257632084	0.003637749042	122	0	94
hdl_cholesterol_qced	2	APOB	0.01	pLoF	8.63E-12	11.06374245	0.01005400187	0.00147255361	750	9	132
urate_qced	16	PKD1	0.01	pLoF	1.85E-11	10.7335542	0.02525224834	0.003759095092	158	2	87
alanine_aminotransferase_	2	APOB	0.01	pLoF	1.85E-11	10.73231631	0.009478716485	0.001411106055	816	10	140
aspartate_aminotransferase	2	APOB	0.01	pLoF	2.14E-11	10.66987598	0.009483661675	0.001416267685	810	10	139
total_bilirubin_qced	12	SLCO1B1	0.01	pLoF	9.92E-11	10.00353531	0.01786024368	0.002761246492	211	5	48
calcium_qced	19	FCGRT	0.01	pLoF	1.68E-10	9.775630387	-0.04923597724	0.00770709771	27	0	17
cystatin_c_serum_qced	7	GIGYF1	0.01	pLoF	4.96E-08	7.304614291	0.02028284879	0.003719727785	129	1	75
total_bilirubin_qced	4	ALB	0.01	pLoF	5.73E-08	7.241829302	-0.0218143896	0.004019587587	113	3	36
phosphate_qced	1	ALPL	0.01	pLoF	9.84E-08	7.007115619	0.009051540142	0.001698320631	556	5	39
apolipoprotein_a_qced	2	APOB	0.01	pLoF	1.11E-07	6.955464546	0.007783886591	0.001466428102	742	9	132
creatinine_serum_qced	2	APOB	0.01	pLoF	1.24E-07	6.90646194	0.007504449211	0.001419294585	815	10	139
creatinine_serum_qced	11	APOC3	0.01	pLoF	1.42E-07	6.847016184	0.005303929137	0.001007901717	1775	4	4
hdl_cholesterol_qced	7	GIGYF1	0.01	pLoF	1.51E-07	6.822243369	-0.0190335283	0.003624154543	121	1	72
c_reactive_protein_qced	11	APOC3	0.01	pLoF	4.68E-07	6.329313112	0.005057335445	0.001003679831	1773	4	4
alkaline_phosphatase_qcec	19	FCGRT	0.01	pLoF	6.28E-07	6.202378513	0.03772190073	0.007570799211	28	0	17
alkaline_phosphatase_qcec	2	APOB	0.01	pLoF	1.13E-06	5.946248242	-0.00685962675	0.00140935024	816	10	140

S3

Supplement to Chapter 3

S3.1 Supplementary notes

S3.1.1 Binarised continuous traits recapitulate significant loss-of-function variant enrichment observed for continuous traits

The continuous traits were binarised using quantile thresholds to define low and high threshold simulated disease cases, with the continuous trait acting as disease liability. Low threshold cases are intended to simulate a disease where cases are at the lower tail of the continuous trait (e.g. short stature disorders), while high threshold cases are intended to simulate a disease where cases are at the upper tail of the continuous trait (e.g. tall stature disorders, hypercholesterolemia). I simulated a range of disease prevalences by choosing quantile thresholds at 0.01, 0.05, 0.10 and 0.20 for the low threshold cases and 0.99, 0.95, 0.90 and 0.80 for the high threshold cases. These correspond to case prevalences of 1%, 5%, 10% and 20% for the respective quantile thresholds of low and high threshold cases. I used the PGS of the continuous trait as the PRS of the binarised trait, reversing the order of the PGS for the low threshold cases, such that the PRS was positively correlated with case status. I tested for pathogenic and protective enrichment, as described

above, with pathogenic/protective effects oriented with respect to simulated disease status, not the continuous trait from which they were derived.

In binarised height, I observed significant positive association (FDR-adjusted one-sided t-test $P < 0.05$) of pLoF carrier with lower PRS in simulated cases for both low and high case thresholds. In low threshold cases (i.e. case status defined by being in the lower quantiles of height), pLoF carrier status in two (*ACAN*, *IGF1R*) of six monogenic short stature genes was associated with lower PRS (with PRS oriented such that low PRS corresponds to lower "risk" of shorter stature; lower PRS = higher height PGS) (Table S3.3). The strongest single-gene pLoF carrier association with lower PRS was observed in *ACAN* for 1% low threshold cases (β [90% CI] = -0.890 [-1.17, -0.614], $P_{\text{adj}} = 1.74 \times 10^{-6}$), which replicated the strongest single-gene pLoF association in the continuous trait misalignment testing. When pLoFs were grouped across short stature genes, the association was significant (maximum $P_{\text{adj}} = 1.57 \times 10^{-4}$) for all four simulated low threshold case prevalences (1%, 5%, 10%, 20%), replicating the association observed when testing grouped short stature genes for continuous height misalignment. In high threshold cases, pLoF carrier status in two (*FBN1*, *NSD1*) of three monogenic tall stature genes was associated with lower PRS (lower PRS = lower height PGS). The association was strongest when grouping pLoFs across tall stature genes for 10% high threshold case status (β [90% CI] = -0.592 [-0.879, -0.304], $P_{\text{adj}} = 2.10 \times 10^{-3}$), which replicated the significant grouped tall stature gene association observed in the continuous trait misalignment testing.

In binarised LDL-C, I replicated all of the same single-gene (*APOB, PCSK9, LDLR*) and grouped-gene associations observed when testing misaligned status in continuous LDL-C. In low threshold cases, pLoF carrier status was associated with lower PRS (lower PRS = higher LDL-C PGS) for pLoFs in *APOB* (minimum $P_{\text{adj}} = 1.37 \times 10^{-5}$) and *PCSK9* ($\beta$ [90% CI] = -0.419 [-0.689, -0.149], $P_{\text{adj}} = 0.0206$). Testing association of grouped *APOB* and *PCSK9* pLoFs was also significant (minimum $P_{\text{adj}} = 1.74 \times 10^{-6}$). In high threshold cases, pLoF carrier status was

associated with lower PRS (lower PRS = higher LDL-C PGS) for pLoFs in *LDLR* (minimum $P_{\text{adj}} = 1.37 \times 10^{-5}$).

Next, I tested individuals in PRS tails for enrichment of pLoFs. For binarised height, low threshold cases in the lower PRS tail were enriched for pLoFs across multiple case prevalences and PRS quantile pairs for pLoFs in grouped short stature genes and *ACAN* (Table S3.3). I observed enrichment of pLoFs in bottom versus top PRS deciles for 10% prevalence high-threshold (height quantile≥0.90) cases in *NSD1* (OR [90% CI] = 51.1 [9.83, 265], $P_{\text{adj}} = 0.0247$) and grouped tall stature genes (OR [90% CI] = 34.0 [7.56, 153], $P_{\text{adj}} = 0.0322$) (Table S3.3). This replicated two (*ACAN*, grouped short stature genes) of the three enrichment results observed for the continuous trait with significant FDR-adjusted Fisher's exact test P-value and empirical P-value. The remaining gene which was not replicated at FDR $\leq 5\%$ was *IGF1*. However, pLoF enrichment in *IGF1* was nominally significant (OR [90% CI] = 57.3 [5.57, 590], ($P = 0.0342$, $P_{\text{adj}} = 0.162$) in bottom versus top PRS deciles for 5% low-threshold (height quantile < 0.05) binarised height cases. When testing enrichment of pLoFs for binarised LDL-C, all genes (*LDLR, APOB, PCSK9*) shown to associate ($P_{\text{adj}} < 0.05$) with misaligned status for non-binarised LDL-C were replicated (binarised LDL-C $P_{\text{adj}} < 0.05$) when testing for enrichment in low-PRS cases compared to high-PRS cases, for multiple case prevalences and pairs of PRS quantile thresholds (Table S3.3). The most significant enrichment was observed in *LDLR*, where 20% high-threshold (LDL-C quantile ≥ 0.8) cases in the bottom PRS decile compared to those in the top PRS decile were enriched for pLoFs (OR [90% CI] = 14.54 [5.41, 39.11], $P_{\text{adj}} = 1.37 \times 10^{-5}$). In *PCSK9*, the only gene with significant empirical P-value ($P_{emp} < 0.05$) among the three associated with LDL-C misaligned status, 1% low-threshold (LDL-C quantile < 0.01) cases in the bottom PRS decile were significantly enriched for pLoFs compared to those in the top PRS decile (OR [90% CI] = 4.97 [2.24, 11.0], $P_{\text{adj}} = 0.0206$).

S3.2 Supplementary figures

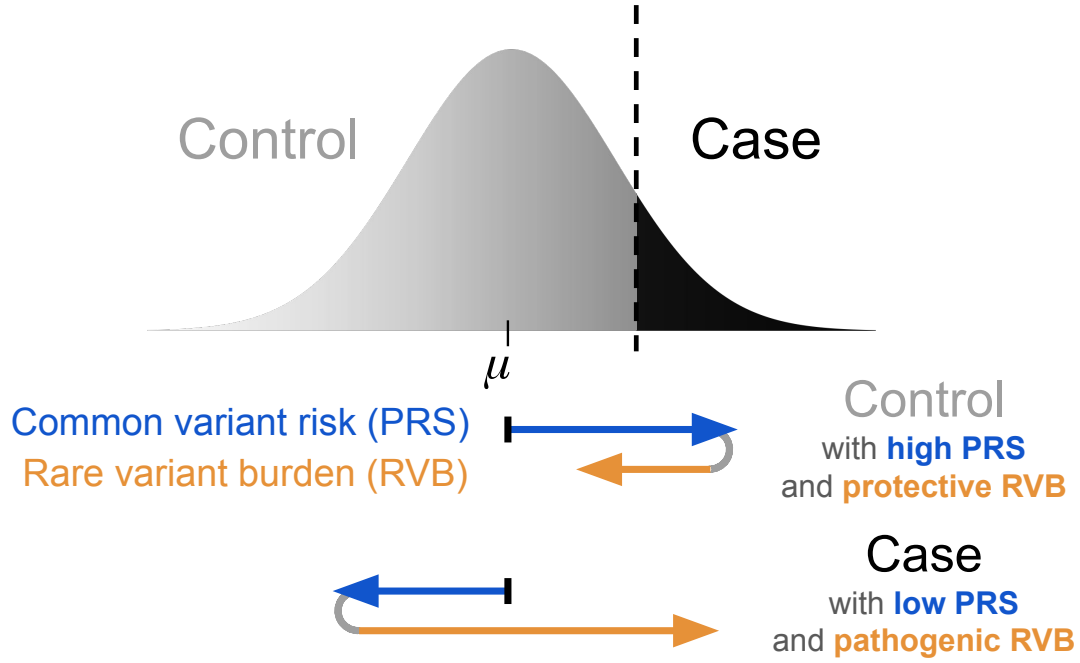

Figure S3.1: Liability threshold model for balancing of common-variant risk and rare variant burden effects. Controls with high PRS are expected to carry protective rare variant burden. Cases with low PRS are expected to carry pathogenic rare variant burden. μ indicates the centre of the genetic liability distribution. The dashed line indicates the liabilty threshold, such that cases are those with total liability greater than the threshold and controls are those with total liability less than the threshold. Total genetic liability is defined here as the net sum of rare and common variant genetic effects.

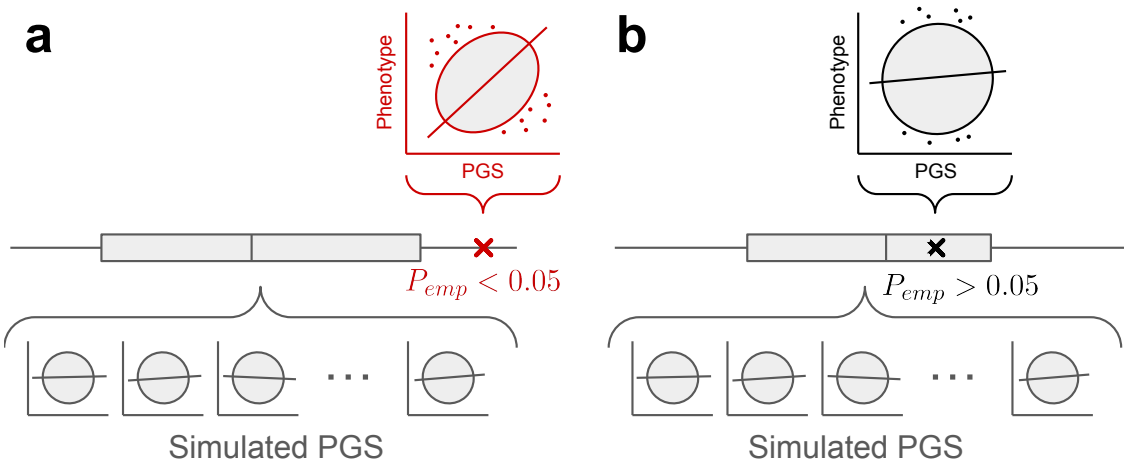

Figure S3.2: Schematic diagram of evaluating the observed test statistic against empirical null distribution of test statistics. The observed test statistic is calculated from an enrichment test using misalignment classification informed by the real PGS. The empirical null distribution of test statistics is calculated using replicates of simulated random PGS. **a**, Theoretical example of enrichment with significant empirical P-value. **b**, Theoretical example of enrichment with non-significant empirical P-value, i.e. the real PGS is no different from random. PGS, polygenic score.

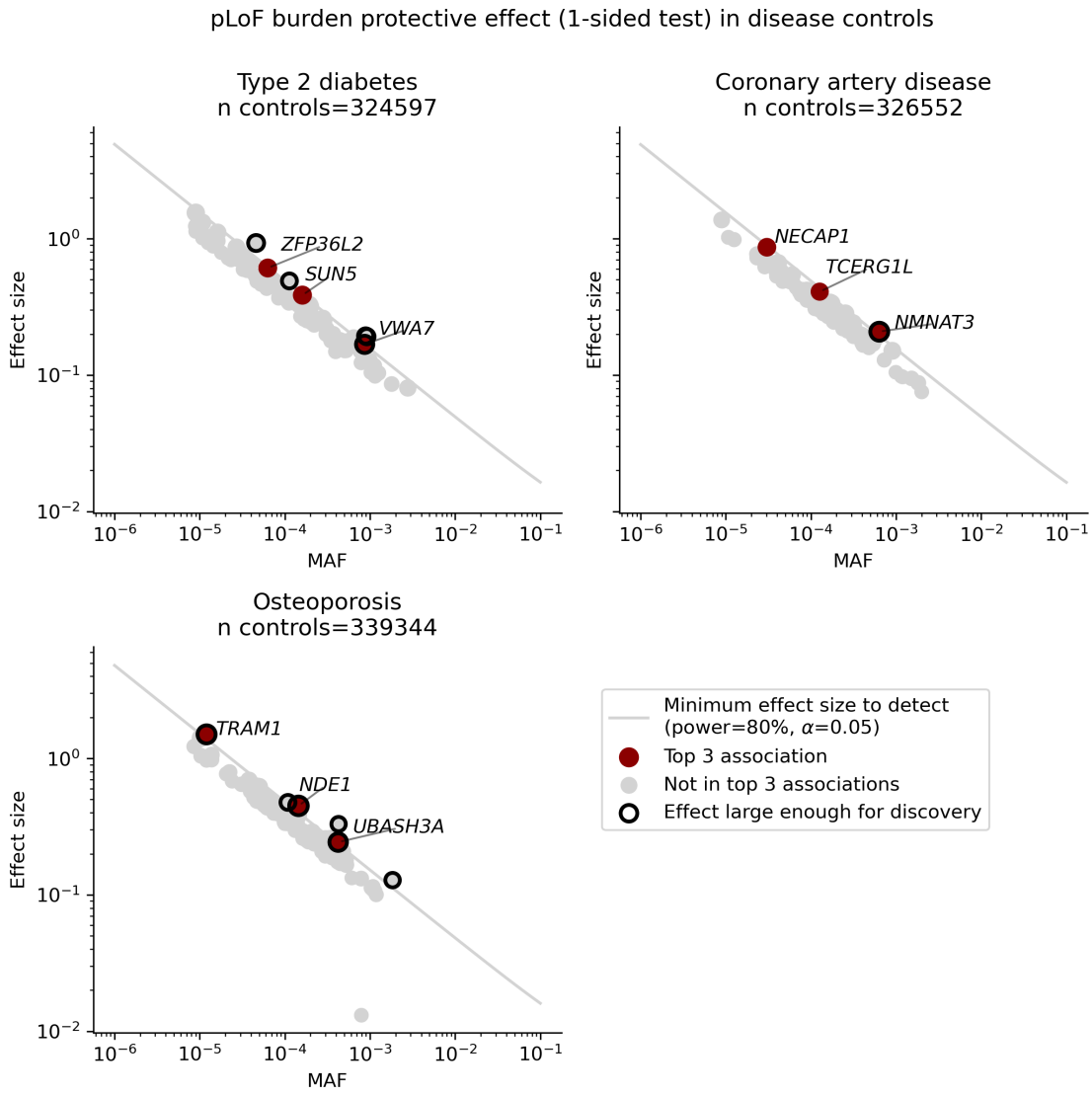

Figure S3.3: Power analysis of burden effects associated with lower PRS in disease controls. The grey line indicates the minimum burden effect size expected to be necessary for discovery of a gene's association, given the cumulative burden MAF in that gene. The top three gene associations per trait are shown in red and labelled with gene symbols. Other associations are shown in grey. Genes where effect size was expected to be sufficient for discovery are highlighted with a black outline. For computational practicality, only genes in the top 100 most significant per trait were included in the power analysis. Power analysis was performed assuming 80% power and $\alpha = 0.05$ using the `solve_power` method in Python's statsmodels[483]. Burden effect sizes have been adjusted from a weighted scale (the default Beta(MAF,1,25) in SAIGE) to an unweighted scale, the correct scale for such power calculations.

S3.3 Supplementary tables

Some supplementary tables have been cropped due to size. The full supplementary tables can be downloaded here: https://doi.org/10.5281/zenodo.15275847

Table S3.1: Phenotype/PGS overview.

Phenotype (abbreviation)	pheno_id	**Original phenotype**	**UKB data field**	**ICD10 code(s)**	n total	n cases	Case prevalence	**PGS**	**PGS UKB data field**	**R^2 ('PRS_only', from Thompson et al. 2024 Supp Table 4)**	**AUC ('PRS_only' from Thompson et al. 2024 Supp Table 3)**	**GEL disorder (panel app reviewed 'green' genes)**
Standing height	height	Standing height	50		400116			Standard PRS for height (HEIGHT)	26240	0.089		Monogenic short stature (66) Beckwith-Wiedemann syndrome and other congenital overgrowth disorders (24)
LDL cholesterol (LDL-C)	ldl	LDL direct	30780		381945			Standard PRS for low density lipoprotein cholesterol (LDL_SF)	26250	0.162		Familial hypercholesterolemia (9)
Body mass index (BMI)	bmi	BMI	21001		399680			Standard PRS for body mass index (BMI)	26216	0.076		Severe early onset obesity (33)
Type 2 diabetes (T2D)	t2d		41270	E11	357120	32523	0.0911	Standard PRS for type 2 diabetes (T2D)	26285		0.704	Monogenic diabetes (35) Neonatal diabetes (34)
Coronary artery disease (CAD)	cad		41270	I25.1	357120	30568	0.0856	Standard PRS for coronary artery disease (CAD)	26227		0.665	Familial hypercholesterolemia (9) Extreme early-onset hypertension (15) Pulmonary arterial hypertension (11)
Osteoporosis (OP)	osteoporosis		41270	M80+M81	357120	17776	0.0498	Standard PRS for osteoporosis (OP)	26258		0.619	Osteogenesis imperfecta (34)
Bone mineral density (BMD)	bmd	Heel BMD t-score automated	78		230105			Standard PRS for estimated bone mineral density t-score (EBMDT)	26234	0.043		Osteogenesis imperfecta (34)
Intraocular pressure (IOP)	iopccmean	Intra-ocular pressure, corneal-compensated	5263		85504			Standard PRS for intraocular pressure (IOP)	26246	0.014		Glaucoma, developmental (18)
Glycated haemoglobin (HbA1C)	hba1c	Glycated haemoglobin	30750		382889			Standard PRS for glycated haemoglobin (HBA1C_DF)	26238	0.077		Monogenic diabetes (35) Neonatal diabetes (34)
Age at menopause (AAM)	aam	Age at menopause (last menstrual period)	3581		124858			Standard PRS for age at menopause (AAM)	26202	0.102		Primary ovarian insufficency (28)

Table S3.2: Canonical monogenic/GEL disorder gene enrichment – Continuous traits.

gene_set_family	pheno_id	pheno_desc	grou	std_resid_ty	csq_subs	max_maf	outlier_type	modified_ge	gene_symbc	all	pval_n_alt_r	degrees_fre	odds_ra	or_90ci	or_90ci	pval_l	qval_l	qval_l	qval_hyperg	qval_hyperg	pval_empiri	n_empirical
canonical_genes	ldl	LDL cholesterol	round0	bothsexes_raw_	pLoF	0.001	low_outlier	ldl-low_outlier_c	grouped:APOB-l	1469	1.72E-215	381943	112.59	82.90	152.93	4.73e-61	1.37e-59		**2.74e-59**		5.89e-02	1000
canonical_genes	ldl	LDL cholesterol	round0	bothsexes_raw_	pLoF	0.001	low_outlier	APOB	APOB	1567	1.04E-116	381943	161.81	115.09	227.50	2.26e-53	3.28e-52		**6.57e-52**		1.61e-01	1000
canonical_genes	ldl	LDL cholesterol	round0	bothsexes_raw_	damaging_mis	0.001	high_outlier	LDLR	LDLR	5009	6.18E-49	381943	61.01	40.27	92.43	1.01e-26		2.92e-25	**1.94e-25**		3.31e-01	1000
canonical_genes	ldl	LDL cholesterol	round0	bothsexes_raw_	damaging_mis	0.001	high_outlier	ldl-high_outlier_c	grouped:LDLR	5009	6.18E-49	381943	61.01	40.27	92.43	1.01e-26					3.31e-01	1000
gel_genes	ldl	LDL cholesterol	round0	bothsexes_raw_	damaging_mis	0.001	high_outlier	LDLR	LDLR	5009	6.18E-49	381943	61.01	40.27	92.43	1.01e-26					3.31e-01	1000
canonical_genes	height	Standing height	round0	bothsexes_raw_	pLoF	0.001	low_outlier	ACAN	ACAN	9055	2.02E-41	400114	406.17	215.09	767.00	1.68e-20	1.62e-19		**2.43e-19**		9.99e-04	1000
gel_genes	height	Standing height	round0	bothsexes_raw_	pLoF	0.001	low_outlier	ACAN	ACAN	9055	2.02E-41	400114	406.17	215.09	767.00	1.68e-20					3.00e-04	10000
canonical_genes	height	Standing height	round0	bothsexes_raw_	pLoF	0.001	low_outlier	height-low_outlie	grouped:SHOX-l	3388	4.32E-67	400114	87.91	52.02	148.56	5.21e-18	3.78e-17		**6.05e-17**		2.00e-03	1000
canonical_genes	ldl	LDL cholesterol	round0	bothsexes_raw_	pLoF	0.001	low_outlier	PCSK9	PCSK9	0476	8.49E-124	381943	52.38	30.42	90.20	2.40e-14	1.39e-13		**2.32e-13**		7.99e-03	1000
gel_genes	ldl	LDL cholesterol	round0	bothsexes_raw_	damaging_mis	0.001	high_outlier	ldl-high_outlier_c	grouped:ABCG5	1657	7.32E-25	381943	11.61	7.69	17.54	9.15e-14				**4.17e-11**	5.76e-01	1000
canonical_genes	ldl	LDL cholesterol	round0	bothsexes_raw_	pLoF	0.001	high_outlier	LDLR	LDLR	9934	3.83E-21	381943	206.97	94.55	453.07	9.92e-11	4.79e-10		**8.22e-10**		1.57e-01	1000
canonical_genes	ldl	LDL cholesterol	round0	bothsexes_raw_	pLoF	0.001	high_outlier	ldl-high_outlier_c	grouped:LDLR	9934	3.83E-21	381943	206.97	94.55	453.07	9.92e-11					1.57e-01	1000
gel_genes	ldl	LDL cholesterol	round0	bothsexes_raw_	pLoF	0.001	high_outlier	LDLR	LDLR	9934	3.83E-21	381943	206.97	94.55	453.07	9.92e-11					1.57e-01	1000
canonical_genes	ldl	LDL cholesterol	round0	bothsexes_raw_	damaging_mis	0.001	low_outlier	ldl-low_outlier_c	grouped:APOB-l	2022	8.21E-27	381943	15.51	9.28	25.93	4.35e-10		6.30e-09	**3.15e-09**		6.19e-02	1000
canonical_genes	ldl	LDL cholesterol	round0	bothsexes_raw_	damaging_mis	0.001	low_outlier	APOB	APOB	7707	7.12E-16	381943	13.90	7.65	25.25	2.21e-07		2.14e-06	**1.42e-06**		2.10e-01	1000
gel_genes	ldl	LDL cholesterol	round0	bothsexes_raw_	pLoF	0.001	high_outlier	ldl-high_outlier_c	grouped:ABCG5	1876	2.59E-15	381943	16.63	8.74	31.62	4.11e-07				**9.36e-05**	3.60e-02	1000
gel_genes	bmd	Bone mineral de	round1	bothsexes_raw_	pLoF	0.001	low_outlier	GORAB	GORAB	9544	0.4280241602	230103	144.87	42.00	499.71	1.14e-04				**1.74e-02**	1.00e-04	10000
gel_genes	height	Standing height	round0	bothsexes_raw_	pLoF	0.001	low_outlier	height-low_outlie	grouped:ACAN-/	2187	4.18E-06	400114	2.48	1.70	3.62	3.11e-04				**3.54e-02**	1.61e-01	10000
canonical_genes	ldl	LDL cholesterol	round0	bothsexes_raw_	damaging_mis	0.001	low_outlier	PCSK9	PCSK9	3835	1.62E-15	381943	20.19	7.70	52.95	5.17e-04		3.75e-03	**3.00e-03**		2.30e-02	1000
canonical_genes	height	Standing height	round0	bothsexes_raw_	pLoF	0.001	low_outlier	IGF1	IGF1	5041	3.76E-06	400114	1302.69	126.88	13374.69	1.54e-03	6.38e-03		**8.12e-03**		1.50e-02	1000
gel_genes	height	Standing height	round0	bothsexes_raw_	pLoF	0.001	low_outlier	IGF1	IGF1	5041	3.76E-06	400114	1302.69	126.88	13374.69	1.54e-03					1.67e-02	10000
gel_genes	bmd	Bone mineral de	round1	bothsexes_raw_	pLoF	0.001	low_outlier	COPB2	COPB2	2614	3.80E-05	230103	1221.99	118.81	12568.13	1.65e-03				1.50e-01	1.00e-04	10000
canonical_genes	height	Standing height	round0	bothsexes_raw_	pLoF	0.001	high_outlier	height-high_outli	grouped:FBN1-E	5564	4.17E-05	400114	24.14	7.39	78.86	3.42e-03	1.24e-02		**1.53e-02**		4.43e-01	1000
canonical_genes	height	Standing height	round0	bothsexes_raw_	damaging_mis	0.001	high_outlier	FBN1	FBN1	4737	3.43E-06	400114	4.27	2.17	8.42	3.43e-03		1.99e-02	**1.53e-02**		9.99e-03	1000
gel_genes	bmd	Bone mineral de	round1	bothsexes_raw_	pLoF	0.001	low_outlier	bmd-low_outlier_	grouped:ALPL-B	3362	0.01521798238	230103	3.49	1.84	6.62	5.24e-03				3.31e-01	1.10e-03	10000
gel_genes	bmd	Bone mineral de	round1	bothsexes_raw_	pLoF	0.001	low_outlier	SPARC	SPARC	3622	0.004795634153	230103	203.66	34.26	1210.74	5.75e-03				3.31e-01	4.01e-01	10000
gel_genes	ldl	LDL cholesterol	round0	bothsexes_raw_	pLoF	0.001	high_outlier	ABCG5	ABCG5	5368	3.23E-12	381943	16.73	5.16	54.30	6.83e-03				3.31e-01	5.99e-03	1000
gel_genes	height	Standing height	round0	bothsexes_raw_	damaging_mis	0.001	high_outlier	AKT2	AKT2	5233	0.2866050764	400114	4.33	2.06	9.09	6.93e-03				3.31e-01	9.99e-04	1000
gel_genes	aam	Age at menopau	round1	bothsexes_raw_	pLoF	0.001	low_outlier	AARS2	AARS2	5985	0.7540599461	124856	15.90	4.86	52.04	7.67e-03				3.31e-01	1.40e-02	1000
gel_genes	bmi	Body mass inde	round0	bothsexes_raw_	pLoF	0.001	high_outlier	CPE	CPE	4694	0.1733244099	399678	133.56	24.36	732.37	7.98e-03				3.31e-01	1.60e-02	1000
canonical_genes	height	Standing height	round0	bothsexes_raw_	pLoF	0.001	low_outlier	SHOX	SHOX	5526	7.80E-17	400114	130.26	23.12	733.94	8.45e-03	2.72e-02		**3.50e-02**		3.05e-01	1000
gel_genes	height	Standing height	round0	bothsexes_raw_	pLoF	0.001	low_outlier	SHOX	SHOX	5526	7.80E-17	400114	130.26	23.12	733.94	8.45e-03					3.15e-01	10000
canonical_genes	height	Standing height	round0	bothsexes_raw_	damaging_mis	0.001	high_outlier	height-high_outli	grouped:FBN1-E	3851	0.001627505708	400114	2.81	1.50	5.28	1.45e-02		7.02e-02	5.61e-02		2.40e-02	1000
canonical_genes	height	Standing height	round0	bothsexes_raw_	pLoF	0.001	high_outlier	FBN1	FBN1	4331	6.81E-07	400114	66.20	11.92	367.59	1.63e-02	4.72e-02		5.90e-02		1.00e+00	1000
gel_genes	bmi	Body mass inde	round0	bothsexes_raw_	damaging_mis	0.001	high_outlier	BBS10	BBS10	2705	0.6576075418	399678	9.21	2.84	29.81	2.10e-02				7.96e-01	2.00e-02	1000
gel_genes	height	Standing height	round0	bothsexes_raw_	damaging_mis	0.001	high_outlier	height-high_outli	grouped:AKT2-A	2357	0.3274115734	400114	1.63	1.13	2.34	2.37e-02				7.96e-01	6.99e-03	1000
gel_genes	bmd	Bone mineral de	round1	bothsexes_raw_	pLoF	0.001	low_outlier	PPIB	PPIB	4261	0.5627878529	230103	42.13	7.86	225.88	2.44e-02				7.96e-01	4.82e-02	10000
gel_genes	bmi	Body mass inde	round0	bothsexes_raw_	damaging_mis	0.001	high_outlier	MKS1	MKS1	9107	0.2314386546	399678	4.58	1.75	11.98	3.00e-02				9.11e-01	2.20e-02	1000
gel_genes	hba1c	Glycated haemo	round1	bothsexes_raw_	pLoF	0.001	high_outlier	ZNF808	ZNF808	4128	0.3710384745	382887	27.77	5.27	146.42	3.61e-02				1.00e+00	8.14e-01	1000
gel_genes	height	Standing height	round0	bothsexes_raw_	damaging_mis	0.001	low_outlier	IGF2	IGF2	3437	0.8581640118	400114	21.01	3.99	110.69	4.74e-02				1.00e+00	4.50e-02	1000

Table S3.3: Canonical monogenic gene enrichment – Dichotomous traits.

pheno_id	pheno_desc	group	csq	max_m	outlier_ty	gene_symbc	q_lo	q_hi	model_desc	odds_	or_90	or_90	pval_fisl	beta_is_	beta_is_	beta_is_	se_is_carrie	tscore_is_ca	pval_is_carr	pval_is_carr	degrees_fre	qval_fisher_	dedup_qval_
t2d	Type 2 diabetes	round0b	pLoF	0.001	high_outlier	t2d-high_outlier_	0.25	0.75	enrichment_in_cases_w_low_prs	9.00	4.67	17.36	4.50E-08	-7.34e-01	-9.43e-01	-5.24e-01	1.27e-01	-5.76e+00	8.59E-09	4.30E-09	32521	**1.89E-06**	**6.01E-08**
t2d	Type 2 diabetes	round0b	pLoF	0.001	high_outlier	t2d-high_outlier_	0.1	0.9	enrichment_in_cases_w_low_prs	18.74	6.83	51.40	2.68E-06	-7.34e-01	-9.43e-01	-5.24e-01	1.27e-01	-5.76e+00	8.59E-09	4.30E-09	32521	**5.63E-05**	**6.01E-08**
t2d	Type 2 diabetes	round0b	pLoF	0.001	high_outlier	HNF1A	0.25	0.75	enrichment_in_cases_w_low_prs	21.13	5.75	77.60	1.38E-05	-8.53e-01	-1.17e+00	-5.37e-01	1.92e-01	-4.44e+00	9.21E-06	4.60E-06	32521	**0.000193147624**	**3.22E-05**
t2d	Type 2 diabetes	round0b	pLoF	0.001	high_outlier	t2d-high_outlier_	0.5	0.5	enrichment_in_cases_w_low_prs	3.14	2.03	4.86	1.90E-05	-7.34e-01	-9.43e-01	-5.24e-01	1.27e-01	-5.76e+00	8.59E-09	4.30E-09	32521	**0.000199214904**	**6.01E-08**
t2d	Type 2 diabetes	round0b	pLoF	0.001	high_outlier	HNF1A	0.5	0.5	enrichment_in_cases_w_low_prs	4.23	2.16	8.29	3.71e-04	-8.53e-01	-1.17e+00	-5.37e-01	1.92e-01	-4.44e+00	9.21E-06	4.60E-06	32521	**0.0028528844**	**3.22E-05**
t2d	Type 2 diabetes	round0b	pLoF	0.001	high_outlier	HNF1A	0.1	0.9	enrichment_in_cases_w_low_prs	37.32	5.93	234.96	4.09e-04	-8.53e-01	-1.17e+00	-5.37e-01	1.92e-01	-4.44e+00	9.21E-06	4.60E-06	32521	**0.0028528844**	**3.22E-05**
t2d	Type 2 diabetes	round0b	pLoF	0.001	high_outlier	GCK	0.25	0.75	enrichment_in_cases_w_low_prs	6.16	2.47	15.39	1.83e-03	-6.04e-01	-9.03e-01	-3.05e-01	1.82e-01	-3.32e+00	8.90e-04	4.45e-04	32521	**0.01100313942**	**0.001556940024**
t2d	Type 2 diabetes	round0b	pLoF	0.001	high_outlier	GCK	0.1	0.9	enrichment_in_cases_w_low_prs	13.98	3.11	62.81	7.88e-03	-6.04e-01	-9.03e-01	-3.05e-01	1.82e-01	-3.32e+00	8.90e-04	4.45e-04	32521	**0.03734001188**	**0.001556940024**
cad	Coronary artery	round0b	pLoF	0.001	low_outlier	APOB	0.25	0.75	enrichment_in_controls_w_high_prs	1.28	1.08	1.51	8.00e-03	6.85e-02	1.06e-02	1.26e-01	3.52e-02	1.95e+00	5.16e-02	2.58e-02	326550	**0.03734001188**	0.06024212507
t2d	Type 2 diabetes	round0b	pLoF	0.001	low_outlier	SLC30A8	0.5	0.5	enrichment_in_controls_w_high_prs	1.34	1.08	1.66	1.52e-02	1.16e-01	1.13e-02	2.21e-01	6.37e-02	1.82e+00	6.84e-02	3.42e-02	324595	0.06355766188	0.06844669515
t2d	Type 2 diabetes	round0b	pLoF	0.001	high_outlier	GCK	0.5	0.5	enrichment_in_cases_w_low_prs	2.44	1.31	4.56	1.66e-02	-6.04e-01	-9.03e-01	-3.05e-01	1.82e-01	-3.32e+00	8.90e-04	4.45e-04	32521	0.06355766188	**0.001556940024**
t2d	Type 2 diabetes	round0b	pLoF	0.001	high_outlier	HNF4A	0.25	0.75	enrichment_in_cases_w_low_prs	inf		1.56	284.23	2.54e-02	-2.29e+00	-3.41e+00	-1.17e+00	6.80e-01	-3.37e+00	7.62e-04	3.81e-04	32521	0.08903012061	**0.001556940024**
cad	Coronary artery	round0b	pLoF	0.001	low_outlier	APOB	0.5	0.5	enrichment_in_controls_w_high_prs	1.14	1.02	1.29	3.27e-02	6.85e-02	1.06e-02	1.26e-01	3.52e-02	1.95e+00	5.16e-02	2.58e-02	326550	0.105612882	0.06024212507
cad	Coronary artery	round0b	pLoF	0.001	low_outlier	cad-low_outlier_	0.25	0.75	enrichment_in_controls_w_high_prs	1.12	1.01	1.25	3.93e-02	3.59e-02	-8.26e-04	7.26e-02	2.23e-02	1.61e+00	1.08e-01	5.39e-02	326550	0.1145087441	0.08389349093
cad	Coronary artery	round0b	pLoF	0.001	high_outlier	LDLR	0.1	0.9	enrichment_in_cases_w_low_prs	4.25	1.41	12.82	4.09e-02	-2.60e-01	-5.23e-01	4.05e-03	1.60e-01	-1.62e+00	1.05e-01	5.27e-02	30566	0.1145087441	0.08389349093
cad	Coronary artery	round0b	pLoF	0.001	high_outlier	LDLR	0.25	0.75	enrichment_in_cases_w_low_prs	2.43	1.13	5.22	5.08e-02	-2.60e-01	-5.23e-01	4.05e-03	1.60e-01	-1.62e+00	1.05e-01	5.27e-02	30566	0.1325032144	0.08389349093
cad	Coronary artery	round0b	pLoF	0.001	low_outlier	ANGPTL3	0.25	0.75	enrichment_in_controls_w_high_prs	1.20	1.00	1.44	5.36e-02	7.42e-02	1.18e-02	1.37e-01	3.79e-02	1.96e+00	5.05e-02	2.53e-02	326550	0.1325032144	0.06024212507
t2d	Type 2 diabetes	round0b	pLoF	0.001	high_outlier	HNF4A	0.5	0.5	enrichment_in_cases_w_low_prs	inf		0.84	151.93	6.86e-02	-2.29e+00	-3.41e+00	-1.17e+00	6.80e-01	-3.37e+00	7.62e-04	3.81e-04	32521	0.160041725	**0.001556940024**
t2d	Type 2 diabetes	round0b	pLoF	0.001	high_outlier	HNF4A	0.1	0.9	enrichment_in_cases_w_low_prs	inf		1.08	321.34	9.72e-02	-2.29e+00	-3.41e+00	-1.17e+00	6.80e-01	-3.37e+00	7.62e-04	3.81e-04	32521	0.2129583112	**0.001556940024**
cad	Coronary artery	round0b	pLoF	0.001	low_outlier	cad-low_outlier_	0.1	0.9	enrichment_in_controls_w_high_prs	1.15	0.97	1.36	1.01e-01	3.59e-02	-8.26e-04	7.26e-02	2.23e-02	1.61e+00	1.08e-01	5.39e-02	326550	0.2129583112	0.08389349093
cad	Coronary artery	round0b	pLoF	0.001	low_outlier	ANGPTL3	0.1	0.9	enrichment_in_controls_w_high_prs	1.25	0.93	1.68	1.23e-01	7.42e-02	1.18e-02	1.37e-01	3.79e-02	1.96e+00	5.05e-02	2.53e-02	326550	0.2384491391	0.06024212507
t2d	Type 2 diabetes	round0b	pLoF	0.001	low_outlier	SLC30A8	0.25	0.75	enrichment_in_controls_w_high_prs	1.25	0.92	1.70	1.30e-01	1.16e-01	1.13e-02	2.21e-01	6.37e-02	1.82e+00	6.84e-02	3.42e-02	324595	0.2384491391	0.06844669515
cad	Coronary artery	round0b	pLoF	0.001	low_outlier	APOB	0.1	0.9	enrichment_in_controls_w_high_prs	1.22	0.93	1.61	1.31e-01	6.85e-02	1.06e-02	1.26e-01	3.52e-02	1.95e+00	5.16e-02	2.58e-02	326550	0.2384491391	0.06024212507
t2d	Type 2 diabetes	round0b	pLoF	0.001	low_outlier	SLC30A8	0.1	0.9	enrichment_in_controls_w_high_prs	1.43	0.88	2.35	1.44e-01	1.16e-01	1.13e-02	2.21e-01	6.37e-02	1.82e+00	6.84e-02	3.42e-02	324595	0.2519017834	0.06844669515
cad	Coronary artery	round0b	pLoF	0.001	high_outlier	LDLR	0.5	0.5	enrichment_in_cases_w_low_prs	1.24	0.71	2.16	3.12e-01	-2.60e-01	-5.23e-01	4.05e-03	1.60e-01	-1.62e+00	1.05e-01	5.27e-02	30566	0.5047441735	0.08389349093
cad	Coronary artery	round0b	pLoF	0.001	low_outlier	cad-low_outlier_	0.5	0.5	enrichment_in_controls_w_high_prs	1.02	0.95	1.10	3.12e-01	3.59e-02	-8.26e-04	7.26e-02	2.23e-02	1.61e+00	1.08e-01	5.39e-02	326550	0.5047441735	0.08389349093
cad	Coronary artery	round0b	pLoF	0.001	low_outlier	ANGPTL3	0.5	0.5	enrichment_in_controls_w_high_prs	1.03	0.91	1.17	3.50e-01	7.42e-02	1.18e-02	1.37e-01	3.79e-02	1.96e+00	5.05e-02	2.53e-02	326550	0.5395832937	0.06024212507
osteoporosis	Osteoporosis	round0b	pLoF	0.001	high_outlier	COL1A1	0.1	0.9	enrichment_in_cases_w_low_prs	2.04	0.45	9.17	3.60e-01	-9.88e-02	-5.90e-01	3.92e-01	2.98e-01	-3.31e-01	7.40e-01	3.70e-01	17774	0.5395832937	0.5183342363
osteoporosis	Osteoporosis	round0b	pLoF	0.001	high_outlier	osteoporosis-hig	0.1	0.9	enrichment_in_cases_w_low_prs	1.53	0.37	6.37	4.58e-01	-4.81e-02	-4.22e-01	3.26e-01	2.27e-01	-2.12e-01	8.32e-01	4.16e-01	17774	0.6629074721	0.529676614
cad	Coronary artery	round0b	pLoF	0.001	low_outlier	PCSK9	0.1	0.9	enrichment_in_controls_w_high_prs	0.96	0.69	1.34	6.12e-01	-5.86e-02	-1.31e-01	1.36e-02	4.39e-02	-1.34e+00	1.82e-01	9.09e-01	326550	0.8314745507	0.9091091375
osteoporosis	Osteoporosis	round0b	pLoF	0.001	high_outlier	osteoporosis-hig	0.5	0.5	enrichment_in_cases_w_low_prs	0.95	0.44	2.09	6.24e-01	-4.81e-02	-4.22e-01	3.26e-01	2.27e-01	-2.12e-01	8.32e-01	4.16e-01	17774	0.8314745507	0.529676614
osteoporosis	Osteoporosis	round0b	pLoF	0.001	high_outlier	COL1A2	0.5	0.5	enrichment_in_cases_w_low_prs	0.98	0.30	3.26	6.40e-01	2.18e-02	-5.54e-01	5.97e-01	3.50e-01	6.22e-02	9.50e-01	5.25e-01	17774	0.8314745507	0.6122752446
osteoporosis	Osteoporosis	round0b	pLoF	0.001	high_outlier	COL1A1	0.5	0.5	enrichment_in_cases_w_low_prs	0.93	0.33	2.62	6.53e-01	-9.88e-02	-5.90e-01	3.92e-01	2.98e-01	-3.31e-01	7.40e-01	3.70e-01	17774	0.8314745507	0.5183342363
osteoporosis	Osteoporosis	round0b	pLoF	0.001	high_outlier	COL1A1	0.25	0.75	enrichment_in_cases_w_low_prs	0.91	0.23	3.61	6.81e-01	-9.88e-02	-5.90e-01	3.92e-01	2.98e-01	-3.31e-01	7.40e-01	3.70e-01	17774	0.8415086011	0.5183342363
osteoporosis	Osteoporosis	round0b	pLoF	0.001	high_outlier	osteoporosis-hig	0.25	0.75	enrichment_in_cases_w_low_prs	0.65	0.17	2.43	8.13e-01	-4.81e-02	-4.22e-01	3.26e-01	2.27e-01	-2.12e-01	8.32e-01	4.16e-01	17774	0.9755107878	0.529676614
cad	Coronary artery	round0b	pLoF	0.001	low_outlier	PCSK9	0.25	0.75	enrichment_in_controls_w_high_prs	0.84	0.68	1.04	9.17e-01	-5.86e-02	-1.31e-01	1.36e-02	4.39e-02	-1.34e+00	1.82e-01	9.09e-01	326550	1	0.9091091375
cad	Coronary artery	round0b	pLoF	0.001	low_outlier	PCSK9	0.5	0.5	enrichment_in_controls_w_high_prs	0.86	0.74	1.00	9.58e-01	-5.86e-02	-1.31e-01	1.36e-02	4.39e-02	-1.34e+00	1.82e-01	9.09e-01	326550	1	0.9091091375
t2d	Type 2 diabetes	round0b	pLoF	0.001	high_outlier	HNF1B	0.25	0.75	enrichment_in_cases_w_low_prs	0.00	0.10	17.75	1.00e+00	5.23e-01	-5.95e-01	1.64e+00	6.80e-01	7.70e-01	4.42e-01	7.79e-01	32521	1	0.8391715189
t2d	Type 2 diabetes	round0b	pLoF	0.001	high_outlier	HNF1B	0.1	0.9	enrichment_in_cases_w_low_prs	0.00	0.27	80.23	1.00e+00	5.23e-01	-5.95e-01	1.64e+00	6.80e-01	7.70e-01	4.42e-01	7.79e-01	32521	1	0.8391715189
t2d	Type 2 diabetes	round0b	pLoF	0.001	high_outlier	HNF1B	0.5	0.5	enrichment_in_cases_w_low_prs	0.00	0.05	9.49	1.00e+00	5.23e-01	-5.95e-01	1.64e+00	6.80e-01	7.70e-01	4.42e-01	7.79e-01	32521	1	0.8391715189
osteoporosis	Osteoporosis	round0b	pLoF	0.001	high_outlier	COL1A2	0.25	0.75	enrichment_in_cases_w_low_prs	0.00	0.04	7.67	1.00e+00	2.18e-02	-5.54e-01	5.97e-01	3.50e-01	6.22e-02	9.50e-01	5.25e-01	17774	1	0.6122752446
osteoporosis	Osteoporosis	round0b	pLoF	0.001	high_outlier	COL1A2	0.1	0.9	enrichment_in_cases_w_low_prs	0.00	0.09	26.41	1.00e+00	2.18e-02	-5.54e-01	5.97e-01	3.50e-01	6.22e-02	9.50e-01	5.25e-01	17774	1	0.6122752446

Table S3.4: GEL disorder gene enrichment – Dichotomous traits.

phenc	pheno_desc	group	csq	max_maf	outlier_type	gene_symbc	q_lov	q_hig	model_desc	odds_ri	or_90ci	or_90ci	pval_fis	beta_is_carr	beta_is_carr	beta_is_carr	se_is_carrie	tscore_is_ca	pval_is_carr	pval_is_carr	degrees_fre	qval_fisher_	dedup_qval_
t2d	Type 2 diabetes	round0b	pLoF	0.001	high_outlier	HNF1A	0.25	0.75	enrichment_in_cases_w_low_prs	21.13	5.75	77.60	1.38e-05	-0.8531612031	-1.169538764	-0.5367836421	0.1923384206	-4.435729483	9.21E-06	4.60E-06	32521		
t2d	Type 2 diabetes	round0b	pLoF	0.001	high_outlier	HNF1A	0.5	0.5	enrichment_in_cases_w_low_prs	4.23	2.16	8.29	3.71e-04	-0.8531612031	-1.169538764	-0.5367836421	0.1923384206	-4.435729483	9.21E-06	4.60E-06	32521		
t2d	Type 2 diabetes	round0b	pLoF	0.001	high_outlier	HNF1A	0.1	0.9	enrichment_in_cases_w_low_prs	37.32	5.93	234.96	4.09e-04	-0.8531612031	-1.169538764	-0.5367836421	0.1923384206	-4.435729483	9.21E-06	4.60E-06	32521		
t2d	Type 2 diabetes	round0b	pLoF	0.001	high_outlier	GCK	0.25	0.75	enrichment_in_cases_w_low_prs	6.16	2.47	15.39	1.83e-03	-0.6041444942	-0.903146638	-0.3051423504	0.1817752179	-3.323580085	0.000889680013	0.000444840007	32521		
t2d	Type 2 diabetes	round0b	pLoF	0.001	high_outlier	t2d-high_outlier_	0.1	0.9	enrichment_in_cases_w_low_prs	1.89	1.35	2.64	2.24e-03	-0.06271625581	-0.1240154799	-0.00141703168	0.03726622052	-1.682925044	0.09239920154	0.04619960077	32521	0.5784289907	0.9446324256
t2d	Type 2 diabetes	round0b	pLoF	0.001	high_outlier	GCK	0.1	0.9	enrichment_in_cases_w_low_prs	13.98	3.11	62.81	7.88e-03	-0.6041444942	-0.903146638	-0.3051423504	0.1817752179	-3.323580085	0.000889680013	0.000444840007	32521		
t2d	Type 2 diabetes	round0b	pLoF	0.001	high_outlier	PLIN1	0.1	0.9	enrichment_in_cases_w_low_prs	4.66	1.89	11.51	1.12e-02	-0.1446990078	-0.3621434209	0.07274540533	0.1321930508	-1.09460374	0.2736983833	0.1368491916	32521	1	0.9446324256
t2d	Type 2 diabetes	round0b	pLoF	0.001	high_outlier	PPARG	0.25	0.75	enrichment_in_cases_w_low_prs	15.82	2.37	105.74	1.43e-02	-0.6186456684	-1.264430579	0.02713924217	0.3925981645	-1.575773206	0.1150878064	0.05754390322	32521	1	0.9446324256
t2d	Type 2 diabetes	round0b	pLoF	0.001	high_outlier	GCK	0.5	0.5	enrichment_in_cases_w_low_prs	2.44	1.31	4.56	1.66e-02	-0.6041444942	-0.903146638	-0.3051423504	0.1817752179	-3.323580085	0.000889680013	0.000444840007	32521		
t2d	Type 2 diabetes	round0b	pLoF	0.001	high_outlier	HNF4A	0.25	0.75	enrichment_in_cases_w_low_prs	inf	1.56	284.23	2.54e-02	-2.28872868	-3.407040026	-1.170417335	0.6798656556	-3.366442563	0.000762325814	0.000381162907	32521		
cad	Coronary artery	round0b	pLoF	0.001	high_outlier	LIPA	0.5	0.5	enrichment_in_cases_w_low_prs	3.41	1.34	8.72	2.68e-02	-0.479225414	-0.929697840	-0.02875298785	0.2738594887	-1.749895233	0.08014642734	0.04007321367	30566	1	0.9446324256
t2d	Type 2 diabetes	round0b	pLoF	0.001	high_outlier	TRMT10A	0.5	0.5	enrichment_in_cases_w_low_prs	1.84	1.10	3.08	3.84e-02	-0.216389750	-0.457757129	0.02497762946	0.1467367671	-1.474679822	0.1403083235	0.07015416177	32521	1	0.9446324256
cad	Coronary artery	round0b	pLoF	0.001	high_outlier	LDLR	0.1	0.9	enrichment_in_cases_w_low_prs	4.25	1.41	12.82	4.09e-02	-0.2595371317	-0.5231271441	0.004052880681	0.160246492	-1.61961194	0.1053259796	0.05266298981	30566		
t2d	Type 2 diabetes	round0b	pLoF	0.001	high_outlier	ABCC8	0.1	0.9	enrichment_in_cases_w_low_prs	4.66	1.45	14.92	4.92e-02	-0.182722418	-0.4817706932	0.1163258571	0.181803263	-1.005055768	0.3148774752	0.1574387376	32521	1	0.9446324256
cad	Coronary artery	round0b	pLoF	0.001	high_outlier	LDLR	0.25	0.75	enrichment_in_cases_w_low_prs	2.43	1.13	5.22	5.08e-02	-0.2595371317	-0.5231271441	0.004052880681	0.160246492	-1.61961194	0.1053259796	0.05266298981	30566		
t2d	Type 2 diabetes	round0b	pLoF	0.001	high_outlier	t2d-high_outlier_	0.25	0.75	enrichment_in_cases_w_low_prs	1.25	1.01	1.55	5.24e-02	-0.06271625581	-0.1240154799	-0.00141703168	0.03726622052	-1.682925044	0.09239920154	0.04619960077	32521	1	0.9446324256
t2d	Type 2 diabetes	round0b	pLoF	0.001	high_outlier	HNF4A	0.5	0.5	enrichment_in_cases_w_low_prs	inf	0.84	151.93	6.86e-02	-2.28872868	-3.407040026	-1.170417335	0.6798656556	-3.366442563	0.000762325814	0.000381162907	32521		
t2d	Type 2 diabetes	round0b	pLoF	0.001	high_outlier	CEL	0.1	0.9	enrichment_in_cases_w_low_prs	2.87	1.12	7.35	7.58e-02	0.01355745358	-0.2250629697	0.2521778768	0.1450667841	0.09345663562	0.9255413743	0.5372293128	32521	1	0.9729727947
cad	Coronary artery	round0b	pLoF	0.001	high_outlier	ABCG5	0.25	0.75	enrichment_in_cases_w_low_prs	2.00	1.00	4.01	7.99e-02	-0.1856202275	-0.4492157649	0.07797531	0.1602498509	-1.15831763	0.2467434895	0.1233717448	30566	1	0.9446324256
t2d	Type 2 diabetes	round0b	pLoF	0.001	high_outlier	ABCC8	0.5	0.5	enrichment_in_cases_w_low_prs	1.82	0.97	3.45	9.01e-02	-0.182722418	-0.4817706932	0.1163258571	0.181803263	-1.005055768	0.3148774752	0.1574387376	32521	1	0.9446324256
osteoporo	Osteoporosis	round0b	pLoF	0.001	high_outlier	BMP1	0.25	0.75	enrichment_in_cases_w_low_prs	inf	0.68	122.91	9.30e-02	-0.6895413527	-1.417500538	0.03841783238	0.4425446931	-1.55812817	0.1192206724	0.05961033622	17774	1	0.9446324256
t2d	Type 2 diabetes	round0b	pLoF	0.001	high_outlier	HNF4A	0.1	0.9	enrichment_in_cases_w_low_prs	inf	1.08	321.34	9.72e-02	-2.28872868	-3.407040026	-1.170417335	0.6798656556	-3.366442563	0.000762325814	0.000381162907	32521		
t2d	Type 2 diabetes	round0b	pLoF	0.001	high_outlier	AKT2	0.5	0.5	enrichment_in_cases_w_low_prs	4.23	0.94	18.98	1.16e-01	-0.453053084	-1.160479069	0.2543728998	0.4300722089	-1.053434923	0.2921495318	0.1460747659	32521	1	0.9446324256
t2d	Type 2 diabetes	round0b	pLoF	0.001	high_outlier	AKT2	0.25	0.75	enrichment_in_cases_w_low_prs	5.27	1.02	27.32	1.22e-01	-0.4530530844	-1.160479069	0.2543728998	0.4300722089	-1.053434923	0.2921495318	0.1460747659	32521	1	0.9446324256
t2d	Type 2 diabetes	round0b	pLoF	0.001	high_outlier	WFS1	0.25	0.75	enrichment_in_cases_w_low_prs	1.51	0.83	2.72	1.70e-01	-0.07788840902	-0.2572009784	0.1014241603	0.1090111962	-0.7144991681	0.4749236988	0.2374618494	32521	1	0.9729727947
osteoporo	Osteoporosis	round0b	pLoF	0.001	high_outlier	CRTAP	0.5	0.5	enrichment_in_cases_w_low_prs	1.91	0.76	4.77	1.84e-01	-0.152619534	-0.604209219	0.2989701522	0.2745327253	-0.5559247389	0.5782693032	0.2891346516	17774	1	0.9729727947
cad	Coronary artery	round0b	pLoF	0.001	high_outlier	cad-high_outlier_	0.25	0.75	enrichment_in_cases_w_low_prs	1.28	0.85	1.92	1.88e-01	-0.07170139608	-0.2037470994	0.06034430728	0.08027565444	-0.8931898044	0.3717625692	0.1858812846	30566	1	0.9665826798
t2d	Type 2 diabetes	round0b	pLoF	0.001	high_outlier	PPARG	0.5	0.5	enrichment_in_cases_w_low_prs	2.82	0.74	10.80	1.89e-01	-0.618645668	-1.264430579	0.02713924217	0.3925981645	-1.575773206	0.1150878064	0.05754390322	32521	1	0.9446324256
cad	Coronary artery	round0b	pLoF	0.001	high_outlier	ABCG5	0.5	0.5	enrichment_in_cases_w_low_prs	1.39	0.81	2.40	2.03e-01	-0.185620227	-0.449215765	0.07797531	0.1602498509	-1.15831763	0.2467434895	0.1233717448	30566	1	0.9446324256
t2d	Type 2 diabetes	round0b	pLoF	0.001	high_outlier	PLIN1	0.25	0.75	enrichment_in_cases_w_low_prs	1.54	0.76	3.12	2.15e-01	-0.1446990078	-0.3621434209	0.07274540533	0.1321930508	-1.09460374	0.2736983833	0.1368491916	32521	1	0.9446324256
t2d	Type 2 diabetes	round0b	pLoF	0.001	high_outlier	CEL	0.25	0.75	enrichment_in_cases_w_low_prs	1.54	0.76	3.12	2.15e-01	0.01355745358	-0.2250629697	0.2521778768	0.1450667841	0.09345663562	0.9255413743	0.5372293128	32521	1	0.9729727947
cad	Coronary artery	round0b	pLoF	0.001	high_outlier	cad-high_outlier_	0.1	0.9	enrichment_in_cases_w_low_prs	1.46	0.76	2.83	2.25e-01	-0.07170139608	-0.2037470994	0.06034430728	0.08027565444	-0.8931898044	0.3717625692	0.1858812846	30566	1	0.9665826798
t2d	Type 2 diabetes	round0b	pLoF	0.001	high_outlier	APPL1	0.1	0.9	enrichment_in_cases_w_low_prs	1.99	0.70	5.69	2.25e-01	0.0835979033	-0.1380650708	0.3052608774	0.1347576807	0.6203572431	0.5350269625	0.7324865187	32521	1	0.9729727947
t2d	Type 2 diabetes	round0b	pLoF	0.001	high_outlier	ABCC8	0.25	0.75	enrichment_in_cases_w_low_prs	1.76	0.68	4.55	2.44e-01	-0.182722418	-0.4817706932	0.1163258571	0.181803263	-1.005055768	0.3148774752	0.1574387376	32521	1	0.9446324256
osteoporo	Osteoporosis	round0b	pLoF	0.001	high_outlier	BMP1	0.1	0.9	enrichment_in_cases_w_low_prs	inf	0.35	105.78	2.46e-01	-0.6895413527	-1.417500538	0.03841783238	0.4425446931	-1.55812817	0.1192206724	0.05961033622	17774	1	0.9446324256
osteoporo	Osteoporosis	round0b	pLoF	0.001	high_outlier	UNC45A	0.1	0.9	enrichment_in_cases_w_low_prs	inf	0.35	105.78	2.46e-01	-0.2123533867	-0.5317844613	0.1070776878	0.1941901823	-1.093533073	0.2741746362	0.1370873181	17774	1	0.9446324256
osteoporo	Osteoporosis	round0b	pLoF	0.001	high_outlier	BMP1	0.5	0.5	enrichment_in_cases_w_low_prs	2.45	0.55	11.01	2.83e-01	-0.689541353	-1.417500538	0.03841783238	0.4425446931	-1.55812817	0.1192206724	0.05961033622	17774	1	0.9446324256
osteoporo	Osteoporosis	round0b	pLoF	0.001	high_outlier	PPIB	0.5	0.5	enrichment_in_cases_w_low_prs	2.45	0.55	11.01	2.83e-01	-0.242487025	-0.970489777	0.4855157266	0.4425711784	-0.5479051436	0.5837639173	0.2918819586	17774	1	0.9729727947
osteoporo	Osteoporosis	round0b	pLoF	0.001	high_outlier	UNC45A	0.25	0.75	enrichment_in_cases_w_low_prs	1.63	0.62	4.27	2.90e-01	-0.2123533867	-0.5317844613	0.1070776878	0.1941901823	-1.093533073	0.2741746362	0.1370873181	17774	1	0.9446324256
cad	Coronary artery	round0b	pLoF	0.001	high_outlier	LDLRAP1	0.1	0.9	enrichment_in_cases_w_low_prs	5.30	0.52	54.30	2.92e-01	0.01514453196	-0.5262476083	0.5565366723	0.3291330749	0.04601340041	0.9632998698	0.5183500651	30566	1	0.9729727947
t2d	Type 2 diabetes	round0b	pLoF	0.001	high_outlier	GATA6	0.25	0.75	enrichment_in_cases_w_low_prs	5.27	0.51	53.98	2.94e-01	-0.4701735241	-1.588671486	0.6483244376	0.679979107	-0.6914528979	0.4892859071	0.2446429536	32521	1	0.9729727947
t2d	Type 2 diabetes	round0b	pLoF	0.001	high_outlier	STAT3	0.25	0.75	enrichment_in_cases_w_low_prs	5.27	0.51	53.98	2.94e-01	0.2005916943	-0.9179129926	1.319096381	0.6799831955	0.2949950758	0.7679995245	0.6160002378	32521	1	0.9729727947
t2d	Type 2 diabetes	round0b	pLoF	0.001	high_outlier	INSR	0.5	0.5	enrichment_in_cases_w_low_prs	1.57	0.63	3.92	2.94e-01	-0.164648008	-0.587478960	0.2581829431	0.2570556431	-0.6405150513	0.52184232	0.26092116	32521	1	0.9729727947
cad	Coronary artery	round0b	pLoF	0.001	high_outlier	LDLR	0.5	0.5	enrichment_in_cases_w_low_prs	1.24	0.71	2.16	3.12e-01	-0.2595371317	-0.5231271441	0.004052880681	0.160246492	-1.61961194	0.1053259796	0.05266298981	30566		
t2d	Type 2 diabetes	round0b	pLoF	0.001	high_outlier	t2d-high_outlier_	0.5	0.5	enrichment_in_cases_w_low_prs	1.05	0.91	1.21	3.15e-01	-0.062716256	-0.124015480	-0.00141703168	0.03726622052	-1.682925044	0.09239920154	0.04619960077	32521	1	0.9446324256
osteoporo	Osteoporosis	round0b	pLoF	0.001	high_outlier	WNT1	0.5	0.5	enrichment_in_cases_w_low_prs	1.64	0.58	4.63	3.16e-01	-0.283383601	-0.798224191	0.23145699	0.3129845407	-0.9054236351	0.3652532721	0.182626636	17774	1	0.9665826798
osteoporo	Osteoporosis	round0b	pLoF	0.001	high_outlier	SERPINF1	0.5	0.5	enrichment_in_cases_w_low_prs	3.27	0.44	24.53	3.23e-01	-0.115258496	-1.055059986	0.824542993	0.5713289568	-0.2017375368	0.8401242245	0.4200621122	17774	1	0.9729727947
cad	Coronary artery	round0b	pLoF	0.001	high_outlier	LIPA	0.25	0.75	enrichment_in_cases_w_low_prs	2.23	0.50	9.99	3.26e-01	-0.4792254139	-0.9296978399	-0.02875298785	0.2738594887	-1.749895233	0.08014642734	0.04007321367	30566	1	0.9446324256

Table S3.5: Singleton burden enrichment – Continuous traits.

pheno_id	pheno_desc	group	std_resid_ty	csq_subset	outlier_type	gene_symbc	carrier_in	n_alt	n_alt_in_sul	n_alt_in_inli	n_alt_in_oul	beta_ur_bur	se_ur_burde	beta_ur_bur	beta_ur_bur	pval_ur_bur	pval_emp	beta_ur_bur	se_ur_burde	qval_ur_bur
bmd	Bone mineral de	round1	bothsexes_raw_	pLoF	low_outlier	ur_max_ac_1	17	18105	12319	12301	18	0.5680055081	0.2398334133	0.09793201798	1.038078998	0.00893480943	0.05	0.000512908781	0.03461665051	0.308468
iopccmean	Intraocular press	round1	bothsexes_raw_	damaging_missense	low_outlier	ur_max_ac_1	19	39209	26978	26945	33	0.4337788011	0.1946548761	0.05225524402	0.8153023582	0.01292713221	0.031	0.01894588198	0.05056552249	0.308468
height	Standing height	round0	bothsexes_raw_	pLoF	low_outlier	ur_max_ac_1	25	31673	21220	21194	26	0.4232133632	0.1985532988	0.03404889747	0.8123778288	0.01652504842	0.4386	0.02527810389	0.02528952167	0.308468
height	Standing height	round0	bothsexes_raw_	synonymous	high_outlier	ur_max_ac_1	275	313118	212178	211014	1164	0.07020511342	0.0386123494	-0.00547509139	0.1458853182	0.0345170518	0.0599	0.2353423848	0.01487731322	0.310112
height	Standing height	round0	bothsexes_raw_	damaging_missense	low_outlier	ur_max_ac_1	83	174313	117726	117613	113	0.1800349024	0.1036528191	-0.02312462292	0.3831944278	0.04120269805	0.6164	0.008506776447	0.02925490867	0.310112
bmd	Bone mineral de	round1	bothsexes_raw_	synonymous	high_outlier	ur_max_ac_1	51	175041	119942	119855	87	0.2212298576	0.1301524868	-0.03386901654	0.4763287318	0.04458767558	0.1858	-0.07237003686	0.06024095653	0.310112
aam	Age at menopau	round1	bothsexes_raw_	synonymous	high_outlier	ur_max_ac_1	29	96834	67660	67608	52	0.2859845483	0.1711252596	-0.04942096066	0.6213900572	0.0473428951	0.2897	-0.1119331735	0.08101551515	0.310112
bmi	Body mass inde:	round0	bothsexes_raw_	pLoF	low_outlier	ur_max_ac_1	22	31630	21280	21258	22	0.3536102239	0.2155887668	-0.06894375901	0.7761642069	0.05048163225		-0.00337464519	0.02781318358	
bmi	Body mass inde:	round0	bothsexes_raw_	damaging_missense	high_outlier	ur_max_ac_1	49	174106	118316	118247	69	0.2101372564	0.132849067	-0.05024691489	0.4705214276	0.05685111562		-0.01188237044	0.03823115879	
ldl	LDL cholesterol	round0	bothsexes_raw_	damaging_missense	low_outlier	ur_max_ac_1	75	166396	114609	114512	97	0.1727462565	0.1115494703	-0.04589070527	0.3913832183	0.06073857157		-0.01705433111	0.03216371845	
ldl	LDL cholesterol	round0	bothsexes_raw_	pLoF	low_outlier	ur_max_ac_1	19	30250	20620	20599	21	0.3418437565	0.220951566	-0.09122131287	0.7749088258	0.0609149394		0.000588921226	0.02784276232	
hba1c	Glycated haemo	round1	bothsexes_raw_	synonymous	high_outlier	ur_max_ac_1	73	299527	205902	205772	130	0.1155788305	0.1044664809	-0.08917547197	0.320333133	0.1342831922		0.008622912606	0.044913175	
aam	Age at menopau	round1	bothsexes_raw_	damaging_missense	low_outlier	ur_max_ac_1	35	53158	37034	36992	42	0.1843887561	0.1703484321	-0.1494941709	0.5182716831	0.1395342157		-0.08142797901	0.05479097172	
ldl	LDL cholesterol	round0	bothsexes_raw_	pLoF	high_outlier	ur_max_ac_1	12	30250	20611	20599	12	0.2877042689	0.2924768075	-0.2855502737	0.8609588115	0.1626363182		-0.03579176575	0.03958069258	
iopccmean	Intraocular press	round1	bothsexes_raw_	other_missense	high_outlier	ur_max_ac_1	26	143796	97792	97725	67	0.1693993054	0.1881807531	-0.1994349707	0.5382335816	0.1840101949		-0.06625145867	0.1228677484	
aam	Age at menopau	round1	bothsexes_raw_	pLoF	high_outlier	ur_max_ac_1	5	9579	6481	6475	6	0.3683000674	0.4124201582	-0.4400434427	1.176643577	0.1859240755		-0.02489263933	0.05527659242	
bmd	Bone mineral de	round1	bothsexes_raw_	pLoF	high_outlier	ur_max_ac_1	10	18105	12311	12301	10	0.2739423584	0.3202114642	-0.3536721114	0.9015568283	0.1961365984		-0.00553414944	0.04107833549	
bmi	Body mass inde:	round0	bothsexes_raw_	pLoF	high_outlier	ur_max_ac_1	11	31630	21271	21258	13	0.237706528	0.2798683654	-0.3108354682	0.7862485242	0.1978433249		0.01347361719	0.03246778957	
bmi	Body mass inde:	round0	bothsexes_raw_	damaging_missense	low_outlier	ur_max_ac_1	78	174106	118340	118247	93	0.08372788908	0.1133281327	-0.138395251	0.3058510292	0.23001169		-0.00769561242	0.03146932944	
iopccmean	Intraocular press	round1	bothsexes_raw_	pLoF	low_outlier	ur_max_ac_1	4	6910	4709	4704	5	0.3059289304	0.4516496358	-0.5793043557	1.191162216	0.2490904548		0.07514945715	0.03855320735	
hba1c	Glycated haemo	round1	bothsexes_raw_	damaging_missense	low_outlier	ur_max_ac_1	50	166833	113883	113816	67	0.07101450406	0.1335643531	-0.190771628	0.3328006361	0.2974714323		-0.03288194064	0.0387778664	
height	Standing height	round0	bothsexes_raw_	other_missense	low_outlier	ur_max_ac_1	163	626857	423798	423440	358	0.02636902804	0.07838303742	-0.1272617253	0.1799997814	0.3682800672		0.01900216974	0.0498001635	
ldl	LDL cholesterol	round0	bothsexes_raw_	damaging_missense	high_outlier	ur_max_ac_1	39	166396	114563	114512	51	0.03418776847	0.1523626045	-0.2644429363	0.3328184732	0.4112292506		-0.03435188289	0.04405233557	
aam	Age at menopau	round1	bothsexes_raw_	pLoF	low_outlier	ur_max_ac_1	7	9579	6482	6475	7	0.07753863855	0.3813990602	-0.6700035194	0.8250807966	0.4194503026		-0.05548342982	0.04765277871	
bmd	Bone mineral de	round1	bothsexes_raw_	damaging_missense	high_outlier	ur_max_ac_1	33	98372	67379	67336	43	0.03200594	0.1664530312	-0.2942420011	0.3582538811	0.4237607394		-0.00438498175	0.04615885473	
bmd	Bone mineral de	round1	bothsexes_raw_	other_missense	low_outlier	ur_max_ac_1	97	352447	241512	241306	206	0.007656681486	0.1030677638	-0.1943561355	0.2096694985	0.4703907236		0.009007863674	0.06553512772	
hba1c	Glycated haemo	round1	bothsexes_raw_	other_missense	low_outlier	ur_max_ac_1	110	599770	412018	411791	227	0.001721402126	0.09780477094	-0.1899759489	0.1934187532	0.4929788294		-0.02402522396	0.06316429634	
bmd	Bone mineral de	round1	bothsexes_raw_	synonymous	low_outlier	ur_max_ac_1	68	175041	119956	119855	101	-0.0181166426	0.1169391054	-0.2473172891	0.2110840039	0.5615591788		0.01869562239	0.04838156014	
ldl	LDL cholesterol	round0	bothsexes_raw_	synonymous	low_outlier	ur_max_ac_1	89	299007	206427	206277	150	-0.01641328729	0.09572220226	-0.2040288037	0.1712022291	0.5680720088		0.01216184575	0.03981168815	
ldl	LDL cholesterol	round0	bothsexes_raw_	other_missense	high_outlier	ur_max_ac_1	86	598427	411842	411667	175	-0.01944416261	0.1105009641	-0.2360260522	0.197137727	0.5698387163		-0.01911088089	0.07108770156	
hba1c	Glycated haemo	round1	bothsexes_raw_	synonymous	low_outlier	ur_max_ac_1	72	299527	205883	205772	111	-0.02012814314	0.1109176561	-0.2375267491	0.1972704628	0.5720002936		-0.01704555127	0.0471865636	
iopccmean	Intraocular press	round1	bothsexes_raw_	damaging_missense	high_outlier	ur_max_ac_1	11	39209	26961	26945	16	-0.06242404009	0.269403478	-0.5904548569	0.4656067768	0.5916187697		0.03749159459	0.06087768995	
hba1c	Glycated haemo	round1	bothsexes_raw_	damaging_missense	high_outlier	ur_max_ac_1	44	166833	113878	113816	62	-0.03400320359	0.1373737848	-0.3032558218	0.2352494146	0.597748247		0.05090644115	0.03182234498	
aam	Age at menopau	round1	bothsexes_raw_	other_missense	low_outlier	ur_max_ac_1	69	193589	134766	134639	127	-0.03843827254	0.1297356935	-0.2927202318	0.2158436868	0.6164920479		-0.03305865739	0.08425187118	
bmi	Body mass inde:	round0	bothsexes_raw_	other_missense	high_outlier	ur_max_ac_1	96	626148	426957	426750	207	-0.03277432434	0.1012178712	-0.2311613519	0.1656127032	0.6269550646		0.03542699724	0.06303942909	
ldl	LDL cholesterol	round0	bothsexes_raw_	synonymous	high_outlier	ur_max_ac_1	57	299007	206362	206277	85	-0.04450503198	0.1265974233	-0.2926359816	0.2036259176	0.6374111088		-0.01647450394	0.05349412305	
iopccmean	Intraocular press	round1	bothsexes_raw_	synonymous	high_outlier	ur_max_ac_1	15	72943	49961	49931	30	-0.09251058409	0.2124048468	-0.5088240838	0.3238029156	0.6684135587		0.05856682274	0.08255849107	
bmi	Body mass inde:	round0	bothsexes_raw_	synonymous	low_outlier	ur_max_ac_1	106	312749	213458	213306	152	-0.04138947175	0.09460657424	-0.2268183573	0.1440394138	0.6691219734		0.01613385553	0.03909848178	
height	Standing height	round0	bothsexes_raw_	other_missense	high_outlier	ur_max_ac_1	315	626857	425364	423440	1924	-0.01663686343	0.03546748863	-0.08615314115	0.05287941429	0.6804913085		0.2683716507	0.02046088916	
aam	Age at menopau	round1	bothsexes_raw_	other_missense	high_outlier	ur_max_ac_1	42	193589	134721	134639	82	-0.08835513297	0.1592639185	-0.4005124132	0.2238021472	0.7104739067		0.03017040949	0.09970952716	
hba1c	Glycated haemo	round1	bothsexes_raw_	pLoF	high_outlier	ur_max_ac_1	8	30250	20352	20343	9	-0.2044803893	0.3353510491	-0.8617684455	0.4528076669	0.728986108		0.04929153254	0.02744121454	
hba1c	Glycated haemo	round1	bothsexes_raw_	other_missense	high_outlier	ur_max_ac_1	98	599770	412020	411791	229	-0.05928803075	0.09550255268	-0.246473034	0.1278969725	0.7326342771		0.07873156433	0.05762092765	

Table S3.6: Singleton burden enrichment – Dichotomous traits.

pheno_id	pheno_desc	group	ob	prs_fielc	csq	outlier_typ	gene_sy	q_l	q_h	odds_ra	or_90ci_low	or_90ci_up	pval_fisher	beta_n_alt	beta_n_alt_9(	beta_n_alt_90ci	se_n_alt	tscore_n_alt	pval_n_alt_2sid	pval_n_alt_1sid
osteoporosis	Osteoporosis	round0b	osteo	p26258_zsc	damaging_m	low_outlier	ur_max_ac_	0.5	0.5	1.021	1.009	1.033	0.002072968099	0.005363783368	0.0007334630857	0.009994103651	0.002815027165	1.905410873	0.05672752721	0.02836376361
osteoporosis	Osteoporosis	round0b	osteo	p26258_zsc	damaging_m	low_outlier	ur_max_ac_	0.25	0.75	1.024	1.007	1.041	0.01053328661	0.005363783368	0.0007334630857	0.009994103651	0.002815027165	1.905410873	0.05672752721	0.02836376361
osteoporosis	Osteoporosis	round0b	osteo	p26258_zsc	damaging_m	low_outlier	ur_max_ac_	0.1	0.9	1.031	1.004	1.059	0.03030900326	0.005363783368	0.0007334630857	0.009994103651	0.002815027165	1.905410873	0.05672752721	0.02836376361
cad	Coronary artery disea	round0b	cad	p26227_zsc	synonymous	low_outlier	ur_max_ac_	0.5	0.5	0.995	0.983	1.006	0.7712684998	0.00407386553	0.0002943655245	0.007853365536	0.002297766331	1.772967719	0.07623498812	0.03811749406
cad	Coronary artery disea	round0b	cad	p26227_zsc	synonymous	low_outlier	ur_max_ac_	0.1	0.9	0.988	0.963	1.014	0.7792151477	0.00407386553	0.0002943655245	0.007853365536	0.002297766331	1.772967719	0.07623498812	0.03811749406
cad	Coronary artery disea	round0b	cad	p26227_zsc	synonymous	low_outlier	ur_max_ac_	0.25	0.75	0.992	0.976	1.008	0.8003554651	0.00407386553	0.0002943655245	0.007853365536	0.002297766331	1.772967719	0.07623498812	0.03811749406
t2d	Type 2 diabetes	round0b	t2d	p26285_zsc	pLoF	high_outlier	ur_max_ac_	0.5	0.5	1.060	0.986	1.140	0.09777036968	-0.02585191455	-0.05523515914	0.00353133004	0.01786323541	-1.44721345	0.1478467899	0.07392339494
t2d	Type 2 diabetes	round0b	t2d	p26285_zsc	pLoF	high_outlier	ur_max_ac_	0.25	0.75	1.056	0.940	1.185	0.2309427499	-0.02585191455	-0.05523515914	0.00353133004	0.01786323541	-1.44721345	0.1478467899	0.07392339494
t2d	Type 2 diabetes	round0b	t2d	p26285_zsc	pLoF	high_outlier	ur_max_ac_	0.1	0.9	0.947	0.757	1.183	0.6772500561	-0.02585191455	-0.05523515914	0.00353133004	0.01786323541	-1.44721345	0.1478467899	0.07392339494
cad	Coronary artery disea	round0b	cad	p26227_zsc	damaging_m	high_outlier	ur_max_ac_	0.5	0.5	1.035	0.992	1.080	0.09579928192	-0.0115666063	-0.027000753	0.003867540402	0.009383010537	-1.232718034	0.2176904626	0.1088452313
cad	Coronary artery disea	round0b	cad	p26227_zsc	damaging_m	high_outlier	ur_max_ac_	0.25	0.75	1.030	0.966	1.098	0.229085681	-0.0115666063	-0.027000753	0.003867540402	0.009383010537	-1.232718034	0.2176904626	0.1088452313
cad	Coronary artery disea	round0b	cad	p26227_zsc	damaging_m	high_outlier	ur_max_ac_	0.1	0.9	1.013	0.907	1.131	0.4357254973	-0.0115666063	-0.027000753	0.003867540402	0.009383010537	-1.232718034	0.2176904626	0.1088452313
t2d	Type 2 diabetes	round0b	t2d	p26285_zsc	synonymous	low_outlier	ur_max_ac_	0.25	0.75	1.002	0.985	1.018	0.4357777129	0.002286980853	-0.001475283471	0.006049245177	0.002287287772	0.9998658155	0.317376195	0.1586880975
t2d	Type 2 diabetes	round0b	t2d	p26285_zsc	synonymous	low_outlier	ur_max_ac_	0.5	0.5	0.993	0.981	1.004	0.8513028589	0.002286980853	-0.001475283471	0.006049245177	0.002287287772	0.9998658155	0.317376195	0.1586880975
t2d	Type 2 diabetes	round0b	t2d	p26285_zsc	synonymous	low_outlier	ur_max_ac_	0.1	0.9	0.981	0.956	1.008	0.8807358022	0.002286980853	-0.001475283471	0.006049245177	0.002287287772	0.9998658155	0.317376195	0.1586880975
cad	Coronary artery disea	round0b	cad	p26227_zsc	pLoF	high_outlier	ur_max_ac_	0.5	0.5	1.039	0.965	1.119	0.2031369971	-0.01927898784	-0.05205120391	0.01349322824	0.01992348879	-0.9676512001	0.333226254	0.166613127
cad	Coronary artery disea	round0b	cad	p26227_zsc	pLoF	high_outlier	ur_max_ac_	0.25	0.75	1.059	0.947	1.185	0.2100005522	-0.01927898784	-0.05205120391	0.01349322824	0.01992348879	-0.9676512001	0.333226254	0.166613127
cad	Coronary artery disea	round0b	cad	p26227_zsc	pLoF	high_outlier	ur_max_ac_	0.1	0.9	1.060	0.878	1.281	0.32313962	-0.01927898784	-0.05205120391	0.01349322824	0.01992348879	-0.9676512001	0.333226254	0.166613127
osteoporosis	Osteoporosis	round0b	osteo	p26258_zsc	other_misser	high_outlier	ur_max_ac_	0.5	0.5	1.015	0.959	1.075	0.3379818435	-0.008259319225	-0.02287577849	0.006357140044	0.008885713095	-0.9295055036	0.3526397983	0.1763198992
osteoporosis	Osteoporosis	round0b	osteo	p26258_zsc	other_misser	high_outlier	ur_max_ac_	0.1	0.9	0.943	0.823	1.080	0.7746224549	-0.008259319225	-0.02287577849	0.006357140044	0.008885713095	-0.9295055036	0.3526397983	0.1763198992
osteoporosis	Osteoporosis	round0b	osteo	p26258_zsc	other_misser	high_outlier	ur_max_ac_	0.25	0.75	0.934	0.859	1.014	0.916818567	-0.008259319225	-0.02287577849	0.006357140044	0.008885713095	-0.9295055036	0.3526397983	0.1763198992
t2d	Type 2 diabetes	round0b	t2d	p26285_zsc	damaging_m	high_outlier	ur_max_ac_	0.5	0.5	1.016	0.972	1.061	0.2819328017	-0.004427229098	-0.0186176357	0.009763177503	0.008626908882	-0.5131883457	0.6078230636	0.3039115318
t2d	Type 2 diabetes	round0b	t2d	p26285_zsc	damaging_m	high_outlier	ur_max_ac_	0.25	0.75	1.022	0.954	1.096	0.3089484505	-0.004427229098	-0.0186176357	0.009763177503	0.008626908882	-0.5131883457	0.6078230636	0.3039115318
t2d	Type 2 diabetes	round0b	t2d	p26285_zsc	damaging_m	high_outlier	ur_max_ac_	0.1	0.9	1.023	0.904	1.158	0.393889723	-0.004427229098	-0.0186176357	0.009763177503	0.008626908882	-0.5131883457	0.6078230636	0.3039115318
t2d	Type 2 diabetes	round0b	t2d	p26285_zsc	pLoF	low_outlier	ur_max_ac_	0.5	0.5	1.012	0.990	1.034	0.1932841513	0.002052508999	-0.008118195257	0.01222321325	0.00618333149	0.3319422551	0.739933077	0.3699665385
t2d	Type 2 diabetes	round0b	t2d	p26285_zsc	pLoF	low_outlier	ur_max_ac_	0.1	0.9	1.003	0.953	1.054	0.4727488809	0.002052508999	-0.008118195257	0.01222321325	0.00618333149	0.3319422551	0.739933077	0.3699665385
t2d	Type 2 diabetes	round0b	t2d	p26285_zsc	pLoF	low_outlier	ur_max_ac_	0.25	0.75	0.997	0.966	1.029	0.5648454319	0.002052508999	-0.008118195257	0.01222321325	0.00618333149	0.3319422551	0.739933077	0.3699665385
osteoporosis	Osteoporosis	round0b	osteo	p26258_zsc	other_misser	low_outlier	ur_max_ac_	0.5	0.5	1.017	1.004	1.030	0.01416935007	0.0006108900586	-0.002702782377	0.003924562495	0.002014564296	0.3032368139	0.7617095814	0.3808547907
osteoporosis	Osteoporosis	round0b	osteo	p26258_zsc	other_misser	low_outlier	ur_max_ac_	0.1	0.9	1.038	1.009	1.068	0.0160879145	0.0006108900586	-0.002702782377	0.003924562495	0.002014564296	0.3032368139	0.7617095814	0.3808547907
osteoporosis	Osteoporosis	round0b	osteo	p26258_zsc	other_misser	low_outlier	ur_max_ac_	0.25	0.75	1.008	0.990	1.027	0.2227179681	0.0006108900586	-0.002702782377	0.003924562495	0.002014564296	0.3032368139	0.7617095814	0.3808547907
t2d	Type 2 diabetes	round0b	t2d	p26285_zsc	damaging_m	low_outlier	ur_max_ac_	0.5	0.5	1.002	0.990	1.015	0.3795546771	0.0006346481065	-0.004037678796	0.005306975009	0.002840564954	0.2234231981	0.8232062917	0.4116031459
t2d	Type 2 diabetes	round0b	t2d	p26285_zsc	damaging_m	low_outlier	ur_max_ac_	0.25	0.75	0.997	0.980	1.015	0.6056013278	0.0006346481065	-0.004037678796	0.005306975009	0.002840564954	0.2234231981	0.8232062917	0.4116031459
t2d	Type 2 diabetes	round0b	t2d	p26285_zsc	damaging_m	low_outlier	ur_max_ac_	0.1	0.9	0.973	0.946	1.001	0.94583476	0.0006346481065	-0.004037678796	0.005306975009	0.002840564954	0.2234231981	0.8232062917	0.4116031459
cad	Coronary artery disea	round0b	cad	p26227_zsc	other_misser	low_outlier	ur_max_ac_	0.25	0.75	0.984	0.966	1.002	0.9302728815	0.0004224536481	-0.002925169891	0.003770077187	0.002035204828	0.2075730376	0.8355624882	0.4177812441
cad	Coronary artery disea	round0b	cad	p26227_zsc	other_misser	low_outlier	ur_max_ac_	0.1	0.9	0.964	0.936	0.992	0.981244281	0.0004224536481	-0.002925169891	0.003770077187	0.002035204828	0.2075730376	0.8355624882	0.4177812441
cad	Coronary artery disea	round0b	cad	p26227_zsc	other_misser	low_outlier	ur_max_ac_	0.5	0.5	0.982	0.970	0.995	0.9877275568	0.0004224536481	-0.002925169891	0.003770077187	0.002035204828	0.2075730376	0.8355624882	0.4177812441
osteoporosis	Osteoporosis	round0b	osteo	p26258_zsc	damaging_m	high_outlier	ur_max_ac_	0.5	0.5	1.026	0.973	1.083	0.2174276262	-0.002520913464	-0.02284240865	0.01780058173	0.01235394787	-0.2040573175	0.8383110446	0.4191555223
osteoporosis	Osteoporosis	round0b	osteo	p26258_zsc	damaging_m	high_outlier	ur_max_ac_	0.25	0.75	0.968	0.895	1.048	0.756239428	-0.002520913464	-0.02284240865	0.01780058173	0.01235394787	-0.2040573175	0.8383110446	0.4191555223
osteoporosis	Osteoporosis	round0b	osteo	p26258_zsc	damaging_m	high_outlier	ur_max_ac_	0.1	0.9	0.911	0.799	1.039	0.8861090068	-0.002520913464	-0.02284240865	0.01780058173	0.01235394787	-0.2040573175	0.8383110446	0.4191555223
cad	Coronary artery disea	round0b	cad	p26227_zsc	pLoF	low_outlier	ur_max_ac_	0.25	0.75	0.995	0.965	1.027	0.5991000997	0.0008279060289	-0.009325161672	0.01098097373	0.006172609362	0.1341257773	0.8933031991	0.4466515995
cad	Coronary artery disea	round0b	cad	p26227_zsc	pLoF	low_outlier	ur_max_ac_	0.5	0.5	0.993	0.972	1.015	0.7014809566	0.0008279060289	-0.009325161672	0.01098097373	0.006172609362	0.1341257773	0.8933031991	0.4466515995
cad	Coronary artery disea	round0b	cad	p26227_zsc	pLoF	low_outlier	ur_max_ac_	0.1	0.9	0.968	0.922	1.017	0.8613458606	0.0008279060289	-0.009325161672	0.01098097373	0.006172609362	0.1341257773	0.8933031991	0.4466515995
t2d	Type 2 diabetes	round0b	t2d	p26285_zsc	other_misser	high_outlier	ur_max_ac_	0.25	0.75	1.041	0.965	1.122	0.1971981119	-6.74E-05	-0.01044746066	0.01031268339	0.006310455935	-0.01067888522	0.9914797098	0.4957398549
t2d	Type 2 diabetes	round0b	t2d	p26285_zsc	other_misser	high_outlier	ur_max_ac_	0.5	0.5	1.008	0.961	1.056	0.4017328467	-6.74E-05	-0.01044746066	0.01031268339	0.006310455935	-0.01067888522	0.9914797098	0.4957398549
t2d	Type 2 diabetes	round0b	t2d	p26285_zsc	other_misser	high_outlier	ur_max_ac_	0.1	0.9	0.865	0.759	0.985	0.9688132539	-6.74E-05	-0.01044746066	0.01031268339	0.006310455935	-0.01067888522	0.9914797098	0.4957398549
osteoporosis	Osteoporosis	round0b	osteo	p26258_zsc	pLoF	low_outlier	ur_max_ac_	0.25	0.75	1.001	0.972	1.032	0.4726186421	-0.002132846326	-0.01219795959	0.00793226694	0.006119137669	-0.3485534141	0.7274248232	0.6362875884
osteoporosis	Osteoporosis	round0b	osteo	p26258_zsc	pLoF	low_outlier	ur_max_ac_	0.5	0.5	0.996	0.975	1.017	0.6265276075	-0.002132846326	-0.01219795959	0.00793226694	0.006119137669	-0.3485534141	0.7274248232	0.6362875884
osteoporosis	Osteoporosis	round0b	osteo	p26258_zsc	pLoF	low_outlier	ur_max_ac_	0.1	0.9	0.987	0.941	1.035	0.6856877047	-0.002132846326	-0.01219795959	0.00793226694	0.006119137669	-0.3485534141	0.7274248232	0.6362875884

Table S3.7: Exome-wide scan for protective genes – Continuous traits.

pheno_id	chrom	gene_symbo	gene_id	max_MAF	Group	MAC_ori	Number_rare	Number_ulti	BETA_Burde	SE_Burden_	Pvalue_Burd	qval_Burder	BETA_Burde	SE_Burden_	Pvalue_Burd	qval_Burder
ldl	2	APOB	ENSG00000084	0.001	pLoF	906	10	147	-0.04814663659	0.001308220913	1.64E-296	3.59E-291	0.3116416675	0.02050186074	1.75E-52	3.82E-47
ldl	1	PCSK9	ENSG00000169	0.001	pLoF	570	7	55	-0.03905166559	0.001645382244	1.60E-124	1.75E-119	0.2458130594	0.03174904765	4.88E-15	5.32E-10
hba1c	1	EPB41	ENSG00000159	0.001	pLoF	148	4	45	-0.03353930269	0.003150438849	1.82E-26	4.42E-22	0.32810996	0.05148892987	9.30E-11	6.76E-06
ldl	2	APOB	ENSG00000084	0.001	damaging_misse	1315	20	133	-0.00838171625	0.001083040051	1.00E-14	1.37E-10	0.1558048805	0.02906657438	4.15E-08	0.00226538716
hba1c	16	PIEZO1	ENSG00000103	0.001	damaging_misse	7302	117	490	-0.00218300126	0.000453455891	1.48E-06	0.005837484472	0.08643481459	0.01722819127	2.62E-07	0.01144360521
hba1c	17	SLC4A1	ENSG00000004	0.001	damaging_misse	915	18	128	-0.00752012345	0.001268154393	3.03E-09	2.13E-05	0.1726044284	0.03508461505	4.33E-07	0.01575678431
hba1c	17	SLC4A1	ENSG00000004	0.001	pLoF	21	0	17	-0.09080591156	0.008416681917	3.89E-27	1.06E-22	0.4742335972	0.09745847048	5.69E-07	0.01774033879
aam	9	KANK1	ENSG00000107	0.001	damaging_misse	931	13	92	-6.49E-05	0.001421170052	0.9635676744	0.99858809	0.1652978304	0.0346877351	9.43E-07	0.02570372795
ldl	15	ATP10A	ENSG00000206	0.001	pLoF	371	10	97	-0.00025180880	0.002056961458	0.9025681021	0.9977052449	0.2012354569	0.04606486045	6.25E-06	0.1515675226
hba1c	6	PRR3	ENSG00000204	0.001	damaging_misse	38	0	14	-0.00056221263	0.00620960597	0.9278587606	0.9981979038	0.3578857537	0.08498836186	1.27E-05	0.2614731226
aam	11	FKBP2	ENSG00000173	0.001	pLoF	34	1	8	0.01013886327	0.006583820439	0.1235685063	0.9238226712	0.3568028715	0.08489846131	1.32E-05	0.2614731226
aam	22	TMPRSS6	ENSG00000187	0.001	damaging_misse	513	9	96	-0.00037698894	0.001723185001	0.8268257102	0.9953108732	0.1854801634	0.04470036467	1.67E-05	0.3029187354
ldl	14	NRXN3	ENSG00000021	0.001	damaging_misse	1434	28	162	0.000942312368	0.001042464586	0.366033866	0.9720523846	0.1298251551	0.03154250532	1.93E-05	0.3235506029
bmd	1	IFFO2	ENSG00000169	0.001	damaging_misse	423	9	39	0.003006389461	0.001884504859	0.1106405567	0.9205109253	0.1699198617	0.04225867827	2.90E-05	0.386518569
bmd	20	SLC2A10	ENSG00000197	0.001	pLoF	171	3	14	0.004708760579	0.002970376699	0.112911768	0.9210049486	0.2230045291	0.05533745668	2.79E-05	0.386518569
iopccmean	12	KCNA5	ENSG00000130	0.001	damaging_misse	48	1	19	-0.00906823330	0.005711556401	0.1123543828	0.9210049486	0.3316371354	0.08218760435	2.73E-05	0.386518569
hba1c	1	HFM1	ENSG00000162	0.001	pLoF	662	13	95	-0.00134357165	0.001491296709	0.3676192237	0.9721667842	0.1705981154	0.04280488936	3.37E-05	0.386518569
hba1c	11	LRRC55	ENSG00000183	0.001	damaging_misse	73	3	6	-0.01040911249	0.004477108181	0.02007391937	0.8066255074	0.3226850899	0.08087666888	3.31E-05	0.386518569
hba1c	12	KICS2	ENSG00000174	0.001	pLoF	55	1	23	-0.00135879025	0.005174418279	0.792860685	0.9945927287	0.322653729	0.08085549022	3.30E-05	0.386518569
ldl	13	SOX21	ENSG00000125	0.001	damaging_misse	55	0	37	-0.00362056089	0.005282832075	0.4931264393	0.9810030431	0.3176714946	0.0803272834	3.83E-05	0.4146492263
hba1c	22	TUBA8	ENSG00000183	0.001	pLoF	68	1	22	-5.99E-05	inf	1	1	0.3155456568	0.07998922001	3.99E-05	0.4146492263
ldl	7	CD36	ENSG00000135	0.001	damaging_misse	1655	34	108	-0.00113110075	0.000962712125	0.2400304519	0.9581605926	0.1186282591	0.03028846228	4.49E-05	0.4426320059
bmi	15	SV2B	ENSG00000185	0.001	pLoF	61	0	31	-0.00076057373	0.005070771734	0.8807711541	0.996980523	0.3009895209	0.07833090747	6.09E-05	0.4426320059
bmd	3	PRKAR2A	ENSG00000114	0.001	pLoF	42	1	14	0.009265141216	0.005956062012	0.1198075872	0.923211553	0.3077099632	0.0788783976	4.79E-05	0.4426320059
iopccmean	8	PHF20L1	ENSG00000129	0.001	damaging_misse	65	0	40	-0.00911365660	0.004933628571	0.06471058614	0.8933824509	0.3052744538	0.07913517787	5.72E-05	0.4426320059
iopccmean	9	PRSS3	ENSG00000010	0.001	damaging_misse	63	1	16	-0.00541526916	0.004986572509	0.2774921752	0.9641679253	0.306350511	0.07906971567	5.34E-05	0.4426320059
iopccmean	17	DRC3	ENSG00000171	0.001	damaging_misse	65	2	13	-0.00237330474	0.00490630123	0.6285804428	0.9863495285	0.3058135426	0.07910357309	5.53E-05	0.4426320059
iopccmean	19	SIGLEC14	ENSG00000254	0.001	pLoF	8	0	5	-0.04254620038	0.01245492969	0.000635453888	0.369903398	0.2967151965	0.07718784616	6.05E-05	0.4426320059
iopccmean	21	URB1	ENSG00000142	0.001	damaging_misse	605	10	83	0.001516946863	0.001663713328	0.3618825392	0.9711791612	0.1661064098	0.04263722057	4.89E-05	0.4426320059
hba1c	19	APLP1	ENSG00000105	0.001	pLoF	88	2	32	-0.00575318085	0.004081272133	0.1586419269	0.9371445581	0.3019017258	0.07833017516	5.80E-05	0.4426320059
ldl	1	PCSK9	ENSG00000169	0.001	damaging_misse	320	6	40	-0.01732557555	0.002188600605	2.45E-15	3.56E-11	0.2011002915	0.05322667976	7.90E-05	0.4728387903
ldl	2	RESP18	ENSG00000182	0.001	pLoF	347	9	22	-0.00084628678	0.002104426057	0.6875764872	0.9894370666	0.1918534192	0.05213579552	0.000116672323	0.4728387903
ldl	2	NOTO	ENSG00000214	0.001	pLoF	84	2	15	-0.00372826401	0.004272446231	0.3828648779	0.9735782772	0.2737701287	0.07492888821	0.000129227669	0.4728387903
ldl	5	ZNF354B	ENSG00000178	0.001	pLoF	92	2	22	0.000961071384	0.004098234505	0.8145901293	0.9953108732	0.2763387852	0.07526377072	0.000120516330	0.4728387903
ldl	6	OR5V1	ENSG00000243	0.001	damaging_misse	317	3	16	-0.00240044996	0.002207675312	0.2768953241	0.9641679253	0.2005030213	0.05322132633	8.25E-05	0.4728387903
ldl	10	DMBT1	ENSG00000187	0.001	pLoF	1611	26	121	-0.00018263743	0.000929561601	0.8442368639	0.9956862677	0.1046558871	0.02857121515	0.000124647345	0.4728387903
ldl	12	TMEM19	ENSG00000139	0.001	pLoF	344	4	18	-0.00264041075	0.002116278849	0.2121531232	0.9515788189	0.1907467128	0.05205531411	0.000124004017	0.4728387903
ldl	14	G2E3	ENSG00000092	0.001	damaging_misse	371	5	29	0.000163305377	0.002117740186	0.9385336217	0.9983321714	0.190905173	0.05209619415	0.000123923483	0.4728387903
ldl	16	C16orf78	ENSG00000166	0.001	damaging_misse	76	2	10	-0.00342158612	0.004492742577	0.4463102851	0.9784493551	0.2765670515	0.07528709499	0.000119625910	0.4728387903
ldl	19	MAP2K2	ENSG00000126	0.001	damaging_misse	731	14	76	-0.00094288675	0.00145275062	0.5163154187	0.9827786377	0.1504339881	0.04081906009	0.000114178189	0.4728387903
bmi	1	C1orf109	ENSG00000116	0.001	pLoF	802	5	15	-0.00114971953	0.001419647314	0.4180190404	0.9774159535	0.1502671083	0.04112329397	0.000129061426	0.4728387903
bmi	1	NPL	ENSG00000135	0.001	damaging_misse	1328	14	67	0.001141840867	0.001091192984	0.2953694023	0.9649683467	0.1251391939	0.0340796983	0.000120349995	0.4728387903
bmi	6	CDYL	ENSG00000153	0.001	damaging_misse	302	4	33	-0.00017399504	0.002086609306	0.933544224	0.9983321714	0.1996920951	0.0530594038	8.38E-05	0.4728387903
bmi	9	ASB6	ENSG00000148	0.001	pLoF	84	1	27	0.003450281785	0.00432205884	0.4246988057	0.9774159535	0.2846161522	0.07633844041	9.64E-05	0.4728387903
bmi	12	FAM234B	ENSG00000084	0.001	pLoF	70	1	30	-0.00402726861	0.004732226586	0.3947524737	0.9744958724	0.2945557237	0.07752382111	7.25E-05	0.4728387903

Table S3.8: Exome-wide scan for pathogenic genes – Continuous traits.

pheno_id	chro	gene_symb	gene_id	max_MAF	Group	MAC_origin	Number_ran	Number_ulti	BETA_Burd	SE_Burden_	Pvalue_Bur	qval_Burder	BETA_Burd	SE_Burden_	Pvalue_Bur	qval_Burder
ldl	19	LDLR	ENSG00000130	0.001	damaging_misse	1426	36	152	0.01530487566	0.00104117865	6.49E-49	4.72E-44	0.2424673531	0.02292136512	1.88E-26	4.10E-21
ldl	19	LDLR	ENSG00000130	0.001	pLoF	105	1	52	0.03591552595	0.00381877072	5.20E-21	1.03E-16	0.3662360581	0.05434402773	7.96E-12	8.68E-07
bmi	X	GABRQ	ENSG00000268	0.001	damaging_misse	27	0	18	0.00664457500	0.00761363517	0.3828154765	0.9735438896	0.4422864604	0.09418871506	1.33E-06	0.09656069455
iopccmean	1	PIGC	ENSG00000135	0.001	damaging_misse	19	0	7	0.02501132176	0.00912177519	0.00610782826	0.6717902667	0.4154608975	0.09142447999	2.76E-06	0.1471293987
hba1c	16	HBA1	ENSG00000206	0.001	damaging_misse	32	0	16	0.0307177873	0.00677323497	5.76E-06	0.01793353416	0.4073868992	0.09049910753	3.37E-06	0.1471293987
hba1c	3	RAP2B	ENSG00000181	0.001	damaging_misse	29	0	18	0.01540389097	0.00711280294	0.03033741274	0.8431539317	0.39137915	0.08875086638	5.17E-06	0.1880011455
bmd	5	NIPBL	ENSG00000164	0.001	pLoF	17	1	2	-0.00147879522	0.00938715182	0.8748447154	0.9969775614	0.3845729073	0.08795680822	6.15E-06	0.191542439
bmi	X	MAGEA4	ENSG00000147	0.001	pLoF	31	1	1	0.00400882568	0.00710190320	0.5724329339	0.9851206716	0.373659064	0.08670417001	8.18E-06	0.2229606425
bmi	4	TRIM2	ENSG00000109	0.001	pLoF	37	0	21	0.01138691557	0.00650597024	0.08007939597	0.9021465248	0.3687428987	0.08617844374	9.39E-06	0.227653301
hba1c	6	SLC22A7	ENSG00000137	0.001	damaging_misse	609	11	52	0.000915343152	0.00155276207	0.5555302702	0.9846074122	0.1877271816	0.04459675921	1.28E-05	0.2633101233
aam	19	AP2S1	ENSG00000042	0.001	damaging_misse	24	0	16	-0.02282571706	0.00784139410	0.00360359486	0.6055233166	0.3557372774	0.08481404791	1.37E-05	0.2633101233
iopccmean	4	KLHL2	ENSG00000109	0.001	damaging_misse	33	0	22	0.02368834594	0.00690327520	0.00060033101	0.3667686173	0.3540340116	0.08466910487	1.45E-05	0.2633101233
bmd	1	GORAB	ENSG00000120	0.001	pLoF	28	0	14	-0.00211686876	0.00731574256	0.7723074188	0.994209482	0.3486510538	0.08389120165	1.62E-05	0.271695051
iopccmean	6	SMPD2	ENSG00000135	0.001	pLoF	36	1	9	0.01024689917	0.00658402334	0.1196301425	0.9231036303	0.344405533	0.08345079048	1.84E-05	0.273100813
bmd	11	TOMT	ENSG00000284	0.001	damaging_misse	149	5	23	-0.00329528229	0.00317124621	0.2987523136	0.9655908397	0.2332863097	0.05659579798	1.88E-05	0.273100813
bmi	12	WASHC4	ENSG00000136	0.001	pLoF	260	5	78	0.00846536046	0.00246514940	0.00059471091	0.3662691683	0.2293150818	0.05622512043	2.27E-05	0.3089211974
iopccmean	2	CFAP210	ENSG00000154	0.001	damaging_misse	46	1	15	0.00489188089	0.00582750640	0.4012186989	0.9752611562	0.3304021828	0.08182292533	2.70E-05	0.345798736
aam	5	KDM3B	ENSG00000120	0.001	damaging_misse	148	3	41	-0.00237488969	0.00316415186	0.4529167649	0.9784767982	0.223500712	0.05575568127	3.05E-05	0.3700959941
hba1c	7	ZNF786	ENSG00000197	0.001	pLoF	339	9	64	0.00393966549	0.00208068304	0.05829829577	0.8882245425	0.2187645667	0.05522154919	3.72E-05	0.4273492656
bmd	X	MOSPD2	ENSG00000130	0.001	damaging_misse	40	0	16	-0.00522761922	0.00612501020	0.3933890555	0.9744634448	0.3125451825	0.07963306688	4.34E-05	0.4554944558
bmi	12	ITGA5	ENSG00000161	0.001	pLoF	67	1	39	-0.00636702727	0.00483065357	0.1874879967	0.9446670566	0.3114688565	0.07941003041	4.39E-05	0.4554944558
ldl	3	TPRA1	ENSG00000163	0.001	damaging_misse	431	13	74	0.00332638877	0.00188826106	0.07813439042	0.901964647	0.2123222584	0.0545457821	4.96E-05	0.4917097892
bmi	11	NPAS4	ENSG00000174	0.001	damaging_misse	82	3	21	0.00042576678	0.00432436455	0.9215688958	0.9979972421	0.3060144603	0.07883483418	5.19E-05	0.4917926027
iopccmean	6	PPP1R10	ENSG00000204	0.001	damaging_misse	61	1	29	0.00375424753	0.00507495960	0.4594461817	0.9793648213	0.304401783	0.07890504342	5.72E-05	0.5147401941
bmd	5	G3BP1	ENSG00000145	0.001	damaging_misse	481	1	33	-0.00284951591	0.00179760396	0.1129266997	0.9208903133	0.1622043502	0.04219384568	6.05E-05	0.5147401941
bmd	11	SPDYC	ENSG00000204	0.001	pLoF	46	1	15	0.00143822003	0.00572281580	0.8015718514	0.9950076842	0.2984080389	0.07786465448	6.35E-05	0.5147401941
aam	6	RNF8	ENSG00000112	0.001	damaging_misse	44	0	21	-0.01017223939	0.00580279371	0.07960372266	0.9021465248	0.3000986913	0.07832713453	6.37E-05	0.5147401941
aam	10	FOXI2	ENSG00000186	0.001	pLoF	51	1	10	-0.00097383844	0.00537662742	0.8562699283	0.9960604303	0.2911026698	0.07713716392	8.04E-05	0.5250599549
ldl	10	NT5C2	ENSG00000076	0.001	damaging_misse	494	3	80	0.00014403607	0.00182098077	0.9369546143	0.9983312231	0.2007804683	0.05330891915	8.28E-05	0.5250599549
hba1c	1	ZNF691	ENSG00000164	0.001	damaging_misse	115	3	10	-0.00153652977	0.00356684087	0.6666271073	0.9882616065	0.2836930856	0.07612309693	9.70E-05	0.5250599549
iopccmean	21	PFKL	ENSG00000141	0.001	damaging_misse	346	5	76	-0.00455683882	0.00215065552	0.03410619724	0.8536281372	0.1965050941	0.05277287991	9.82E-05	0.5250599549
ldl	5	MACIR	ENSG00000181	0.001	damaging_misse	518	3	8	-0.00015537217	0.00168861792	0.9266889472	0.9981573317	0.1969006231	0.05300256689	0.000101631514	0.5250599549
bmi	3	PAK2	ENSG00000180	0.001	damaging_misse	390	4	36	0.00286532614	0.00201410560	0.1548435744	0.9363035064	0.1944226957	0.05234233954	0.000101825428	0.5250599549
ldl	12	PHLDA1	ENSG00000139	0.001	pLoF	144	4	26	0.00496989668	0.00347774749	0.1529880761	0.9353846771	0.2809910713	0.07585577884	0.000105995851	0.5250599549
bmi	1	GPATCH3	ENSG00000198	0.001	damaging_misse	1791	17	55	-0.00068029276	0.00094642079	0.4722605826	0.9803587005	0.1244675194	0.03363711806	0.000107671217	0.5250599549
aam	16	ACSM2B	ENSG00000066	0.001	pLoF	222	2	45	-6.21E-05	0.00664050486	0.9925417039	1	0.1940791577	0.05251640115	0.000109687793	0.5250599549
hba1c	6	C6orf118	ENSG00000112	0.001	pLoF	124	3	22	0.00411781380	0.00343573502	0.2307127553	0.9572804969	0.2782798757	0.07544672582	0.000112821577	0.5250599549
bmi	2	PXDN	ENSG00000130	0.001	damaging_misse	1029	19	120	0.00099264647	0.0012382776	0.4227642133	0.9773870393	0.1474998796	0.0402258643	0.000122806183	0.5250599549
iopccmean	5	NDUFS6	ENSG00000145	0.001	pLoF	2	0	2	0.03660570229	0.02786675549	0.188981667	0.9450183475	0.5371846585	0.1466373323	0.0001244664	0.5250599549
ldl	2	GAD1	ENSG00000128	0.001	damaging_misse	145	3	48	-0.00092401282	0.00324520656	0.7758497862	0.9942857226	0.2757053843	0.07531489562	0.000125762613	0.5250599549
iopccmean	17	PPP1R9B	ENSG00000108	0.001	pLoF	3	0	3	0.03766218104	0.02275761357	0.09793937901	0.9153940334	0.5348192665	0.1463153017	0.00012846485	0.5250599549
bmd	21	BACH1	ENSG00000156	0.001	pLoF	63	1	22	-0.00930163674	0.00487843192	0.05656129066	0.8851995144	0.2728342368	0.07469488877	0.000129774274	0.5250599549
bmd	7	SEMA3C	ENSG00000075	0.001	damaging_misse	579	8	78	-0.00177089641	0.00161937724	0.2741452292	0.963871503	0.1462529168	0.04013545866	0.000134225738	0.5250599549
bmi	X	TSPAN7	ENSG00000156	0.001	pLoF	2	0	1	0.09161969691	0.02773388024	0.000954758573	0.4311377396	0.5295917192	0.1455893708	0.0001376105	0.5250599549
hba1c	3	KPNA4	ENSG00000186	0.001	pLoF	1	0	1	0.1139282491	0.03819788961	0.00285830243	0.5723720332	0.5268529389	0.1452140595	0.00014275045	0.5250599549

Table S3.9: Exome-wide scan for protective genes – Dichotomous traits.

pheno_id	chror	gene_sym	gene_id	MAC_ca	MAC_cont	Numbe	Number_t	BETA_Burd	SE_Burden_	Pvalue_Bur	qval_Burder	BETA_Burd	SE_Burden_	Pvalue_Bur	qval_Burder
osteoporosis	16	NDE1	ENSG00000072	4	100	2	26	-0.01195711823	0.01802615988	0.5071254221	0.9810421235	0.0179835288	0.00435669088	1.83E-05	0.6836608905
osteoporosis	8	TRAM1	ENSG00000067	0	7	0	7	-0.04582587044	0.06650667199	0.4907970253	0.9810421235	0.06005275468	0.01508307493	3.42E-05	0.6836608905
cad	3	NMNAT3	ENSG00000163	43	409	4	10	0.006689208434	0.007048971854	0.3426397945	0.9810421235	0.008383735671	0.002118092775	3.78E-05	0.6836608905
osteoporosis	21	UBASH3A	ENSG00000160	15	278	6	42	0.002663810973	0.01113825195	0.8109824483	0.9862095333	0.009760355005	0.002548513549	6.41E-05	0.7467842253
t2d	6	VWA7	ENSG00000204	50	551	10	78	-0.00410454862	0.005744359552	0.4748961346	0.9810421235	0.006750949305	0.001795952935	8.53E-05	0.7467842253
t2d	20	SUN5	ENSG00000167	9	102	3	29	-0.0037162146	0.0135879619	0.7844741516	0.9852644369	0.01544124629	0.004146281755	9.80E-05	0.7467842253
t2d	2	ZFP36L2	ENSG00000152	8	38	0	30	0.04058650271	0.02073899573	0.05034583534	0.9678389124	0.02441662476	0.00663859209	0.000117543095	0.7467842253
osteoporosis	6	NDUFAF4	ENSG00000123	2	88	1	9	-0.02246482292	0.02067185048	0.2771539522	0.9810421235	0.01746765308	0.004772466328	0.000126063665	0.7467842253
cad	12	NECAP1	ENSG00000089	1	19	0	12	-0.00835195767	0.03802623071	0.8261540689	0.9864423984	0.03464168458	0.009569444374	0.000147283903	0.7467842253
cad	X	SLC16A2	ENSG00000147	0	1	0	2	-0.04208036389	0.1845446934	0.8196286581	0.9862095333	0.1426988239	0.03950709512	0.000151934203	0.7467842253
osteoporosis	18	ATP8B1	ENSG00000081	8	223	4	72	-0.01079589527	0.01289574357	0.4024985162	0.9810421235	0.01027944413	0.002869636108	0.000170394203	0.7467842253
t2d	12	RFX4	ENSG00000111	2	5	0	6	0.1031370654	0.05452609919	0.05855531044	0.9731828288	0.06194344849	0.01747462841	0.000196481749	0.7467842253
osteoporosis	6	GNMT	ENSG00000124	8	96	2	19	0.01927920425	0.01783382128	0.2796760764	0.9810421235	0.01542376248	0.00438236121	0.000216172670	0.7467842253
osteoporosis	X	ZNF449	ENSG00000173	0	11	0	7	-0.04860850168	0.03869767006	0.2090763594	0.9810421235	0.05665487997	0.01614281243	0.000224387212	0.7467842253
t2d	9	NOL8	ENSG00000198	47	475	15	103	-0.00130360196	0.006165680904	0.8325527552	0.9864423984	0.006717668509	0.00191822106	0.000230863447	0.7467842253
cad	10	TCERG1L	ENSG00000176	12	83	2	34	0.02047263247	0.01494827956	0.1708228992	0.9810421235	0.01637936084	0.004704011996	0.000248843610	0.7467842253
t2d	17	CLUH	ENSG00000132	9	131	2	64	-0.0095513242	0.01251250403	0.4452592532	0.9810421235	0.01311145827	0.003785182311	0.000266211276	0.7467842253
cad	17	ERBB2	ENSG00000141	12	162	2	70	-0.00632642705	0.01149873144	0.58219264	0.9810421235	0.01137533514	0.003327515126	0.000314758481	0.7467842253
t2d	15	LDHAL6B	ENSG00000171	2	13	1	0	0.02244933163	0.03688902589	0.5428135	0.9810421235	0.03719645116	0.01089113861	0.000318564846	0.7467842253
t2d	8	ZNF250	ENSG00000196	0	17	0	13	-0.04481173311	0.03601820945	0.2134476533	0.9810421235	0.03453489624	0.0101211647	0.000322267373	0.7467842253
cad	5	EPB41L4A	ENSG00000129	56	592	7	72	0.001062160805	0.005803304274	0.8547769101	0.9864423984	0.006054544047	0.00177732761	0.000328955172	0.7467842253
t2d	10	NKX1-2	ENSG00000229	1	10	0	2	-0.00608832516	0.04006028469	0.8792034366	0.9886305341	0.04442223213	0.01308802174	0.000344255671	0.7467842253
t2d	14	AHNAK2	ENSG00000185	17	135	5	9	0.01104411949	0.01149307581	0.3365839463	0.9810421235	0.01270761354	0.003747440974	0.000348176539	0.7467842253
osteoporosis	1	OR6P1	ENSG00000186	10	201	2	10	-0.00108395013	0.01322695187	0.934686388	0.9933926413	0.01031602986	0.003042492821	0.000348634770	0.7467842253
cad	11	CD82	ENSG00000085	10	118	3	23	-0.00464937411	0.01299574594	0.7205220033	0.9825448267	0.01347430869	0.003997864985	0.000375328239	0.7467842253
t2d	11	LARGE2	ENSG00000165	42	392	9	39	0.001605278901	0.006626744987	0.8085923087	0.9862095333	0.007116115083	0.002120695844	0.000396027134	0.7467842253
osteoporosis	4	PDLIM3	ENSG00000154	2	79	2	21	-0.02584518931	0.01935651405	0.1818046127	0.9810421235	0.01673579426	0.004991069918	0.000394903887	0.7467842253
osteoporosis	18	ZNF236	ENSG00000130	1	32	0	25	-0.01736214117	0.03265660106	0.5949629293	0.9810421235	0.02484145737	0.007418646017	0.000406227456	0.7467842253
osteoporosis	11	OR52H1	ENSG00000181	0	5	0	3	-0.04082544507	0.1405752328	0.7714971979	0.9851279129	0.06687182988	0.01997475995	0.000407254695	0.7467842253
cad	11	KRTAP5-10	ENSG00000204	9	113	1	3	-0.00818779069	0.01300407174	0.5289348286	0.9810421235	0.01367749407	0.004089854302	0.000412527192	0.7467842253
t2d	3	NR1I2	ENSG00000144	23	180	5	33	0.01179552763	0.009964095423	0.236490992	0.9810421235	0.01042634561	0.003148266904	0.000463533536	0.8120509446
t2d	X	KIF4A	ENSG00000090	2	7	0	4	0.0787048251	0.04957608313	0.1123866974	0.9810421235	0.05264262783	0.01597468344	0.000491446037	0.8163064048
t2d	9	TLE1	ENSG00000196	5	16	0	20	0.06837576341	0.03033883309	0.02421269181	0.8910353234	0.03217779175	0.009826049304	0.000528789629	0.8163064048
osteoporosis	7	FGL2	ENSG00000127	3	31	0	23	0.04170469906	0.0337906808	0.2171257325	0.9810421235	0.02507749107	0.007669817688	0.000538434988	0.8163064048
osteoporosis	6	HEY2	ENSG00000135	0	26	0	9	-0.04440990079	0.03466578023	0.2001629029	0.9810421235	0.02769074363	0.008498311835	0.000560247946	0.8163064048
t2d	11	FAR1	ENSG00000197	3	37	0	26	-0.00527309808	0.02300747285	0.8187207251	0.9862095333	0.02129368336	0.006550080525	0.000575193358	0.8163064048
osteoporosis	9	ACER2	ENSG00000177	4	93	2	17	0.000720451086	0.02471282454	0.9767426268	0.9970470452	0.01447137146	0.004465050478	0.000595489079	0.8163064048
cad	19	POLR1G	ENSG00000117	2	23	1	4	-0.00048232819	inf	1	1	0.02662491776	0.008235421854	0.000612540167	0.8163064048
cad	14	POMT2	ENSG00000009	18	197	3	61	-0.00501379612	0.009662509037	0.6038362384	0.9810421235	0.009701292957	0.003001953948	0.000615368104	0.8163064048
t2d	19	RGL3	ENSG00000205	185	1781	20	53	0.001684122437	0.003204493257	0.5992006865	0.9810421235	0.003243586113	0.001004124097	0.000618379394	0.8163064048
t2d	16	MEIOB	ENSG00000162	29	332	3	32	-0.00445905452	0.007523641335	0.5534004193	0.9810421235	0.00749788519	0.002321822658	0.000620446468	0.8163064048
t2d	16	CES3	ENSG00000172	60	513	6	36	0.004088708932	0.005775895022	0.4790124299	0.9810421235	0.006039822762	0.001882170508	0.000666064256	0.8163064048
cad	4	CCDC96	ENSG00000173	2	43	1	13	-0.02114679847	0.02202939184	0.3370876048	0.9810421235	0.02173093653	0.0067746285	0.000669018449	0.8163064048
t2d	19	COX7A1	ENSG00000161	12	184	5	8	-0.01334531423	0.01026714722	0.1936669787	0.9810421235	0.009978437379	0.003115027655	0.000679259383	0.8163064048
t2d	16		ENSG00000289	27	313	3	23	-0.00479189818	0.007765555419	0.5371879694	0.9810421235	0.007649374662	0.002389150747	0.000683065510	0.8163064048
osteoporosis	11	LUZP2	ENSG00000187	8	141	1	22	-0.00266746741	0.01459852555	0.8550163879	0.9864423984	0.01158006817	0.003620801748	0.000691428419	0.8163064048

S4

Supplement to Chapter 4

S4.1 Supplementary notes

S4.1.1 Calculating unweighted burden effect sizes

First, summary statistics for variant-level results are recalculated relative to the minor allele: the sign of the t-statistic is flipped if the original effect allele is not the minor allele and the frequency of the effect allele is recalculated accordingly. Then, for all single-variant and grouped ultra-rare variant burden results in a given gene which satisfy the maximum minor allele frequency and variant consequence of the desired result to unweight, the following summary statistics are summed: t-statistic, variance of t-statistic, and MAF. The unweighted t-statistic is calculated by dividing the summed t-statistics by the summed variance of t-statistics. The unweighted MAF is the summed MAF across variants and grouped ultra-rare variants in the gene.

$$t_{unweighted,gene} = \frac{\sum_j t_{weighted,j}}{\sum_j \mathrm{Var}(t_{weighted,j})}, \text{for variant } j \text{ in } gene \tag{S4.1}$$

$$\mathrm{MAF}_{unweighted,gene} = \sum_j \mathrm{MAF}_{weighted,j}, \text{for variant } j \text{ in } gene \tag{S4.2}$$

S4.1.2 Significant genes with low minor allele count

In sex-combined and sex differential gene-level associations we sought to avoid spurious associations by excluding putatively significant genes if they were supported by a total MAC < 10 across both sexes.

Among the sex-combined gene-level associations which were significant at FDR $\leq 1\%$ (SKAT-O $P \leq 4.37 \times 10^{-5}$), there were seven genes which were excluded for having MAC < 5: *DEFB112* (MAC=1), *CHMP4B* (MAC $= 2$), *FEZF2* (MAC $= 3$), *GLP1R* (MAC $= 3$), *PCBD2* (MAC $= 4$), *VGF* (MAC $= 8$), and *TM4SF20* (MAC $= 9$) (Table S4.3).

In the sex differential analysis, we excluded a significant (sex-difference $P < 2.67 \times 10^{-6}$) differential effect at *SENP5* for visceral adipose tissue volume (β_{female} (SE) $= 0.08930$ (0.028055), $\beta_{\text{male}} = -0.09967$ (0.028675), $P_{\text{diff}} = 2.33 \times 10^{-6}$) due to combined MAC < 10 ($\text{MAC}_{\text{female}} = 2$, $\text{MAC}_{\text{male}} = 2$) (Table S4.5).

S4.2 Supplementary figures

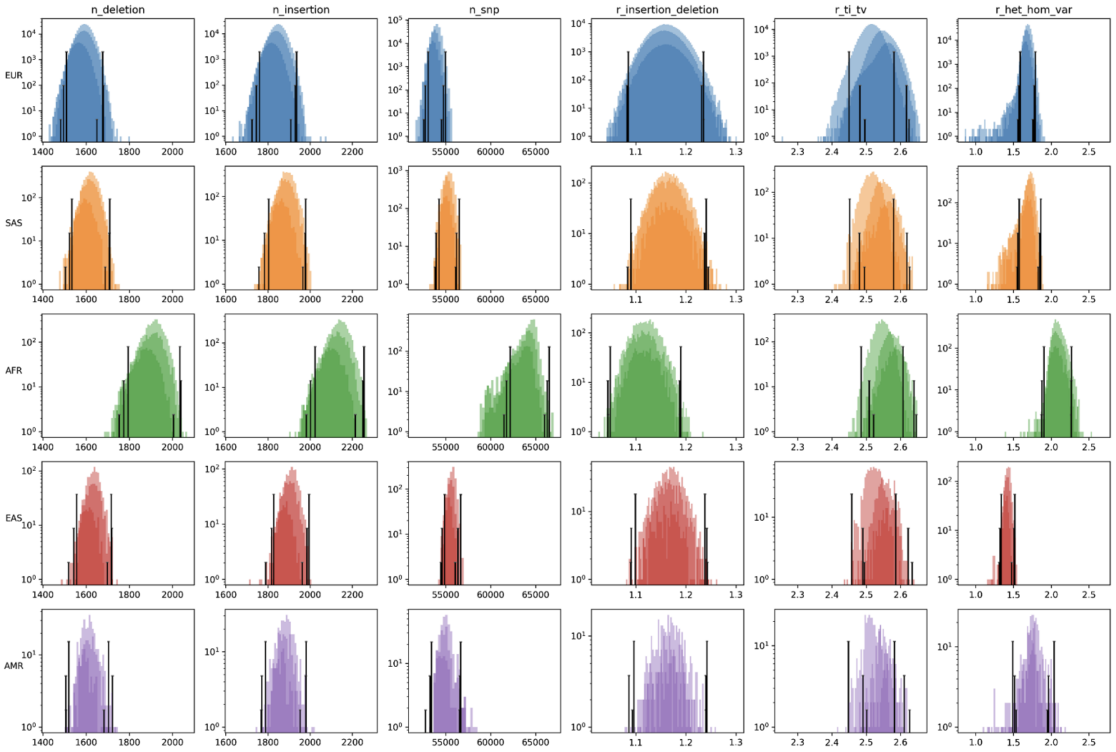

Figure S4.1: MAD thresholds, split by tranche. Samples with any of `n_deletion`, `n_insertion`, `n_snp`, `r_insertion_deletion`, `r_ti_tv`, and `r_het_hom_var` exceeding four MADs from the median are removed. MAD thresholds are displayed as vertical lines, conditional on tranche size (50k, 200k, 250k).

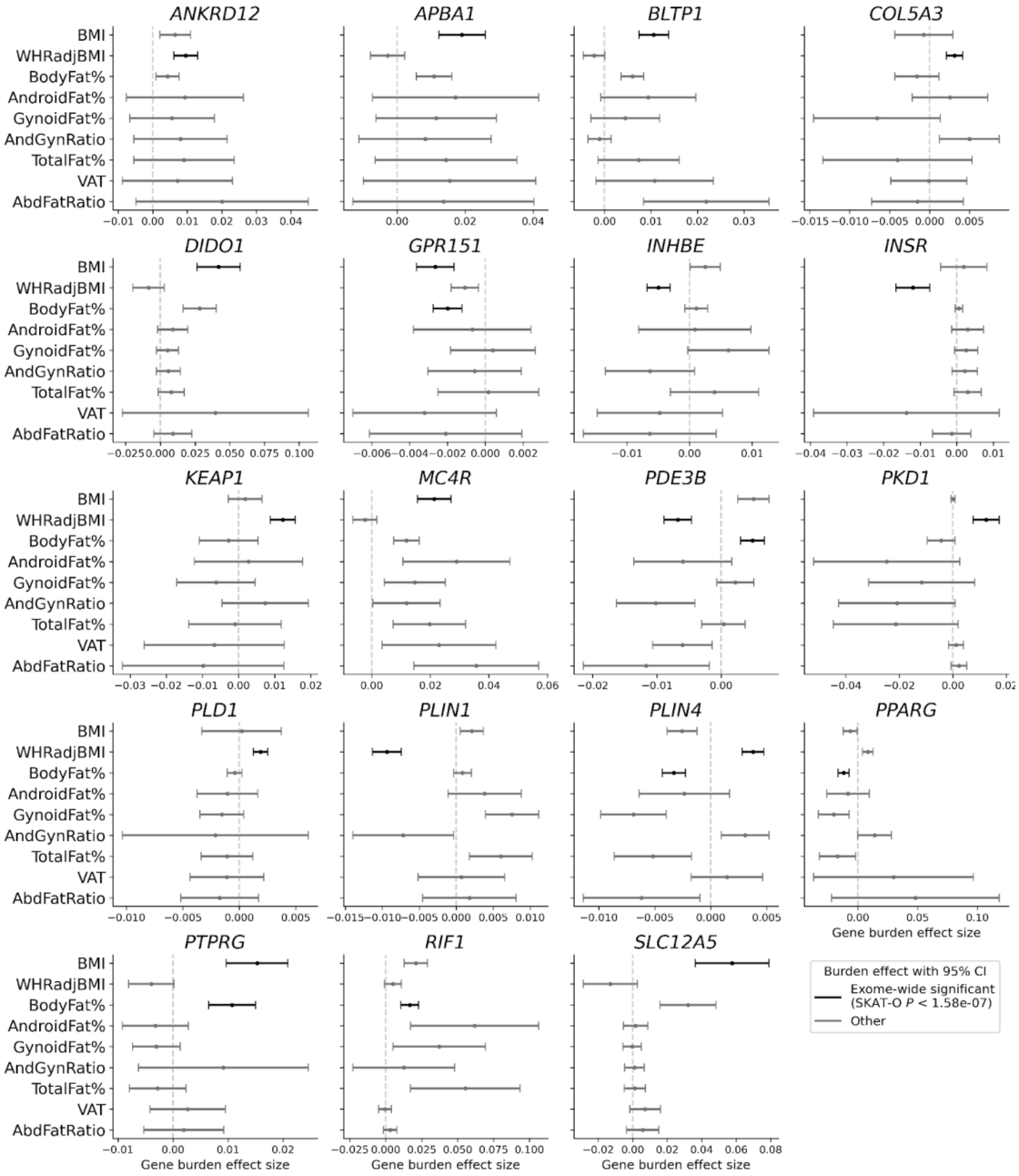

Figure S4.2: Gene burden effects across all nine obesity and fat distribution traits for 19 genes with exome-wide significant burden associations. Confidence intervals for effect size defined as ±1.96 standard errors. Only the result of the consequence mask with the lowest SKAT-O *P*-value is shown for each trait-gene pair. The vertical dotted line marks an effect size of zero. VAT, visceral adipose tissue.

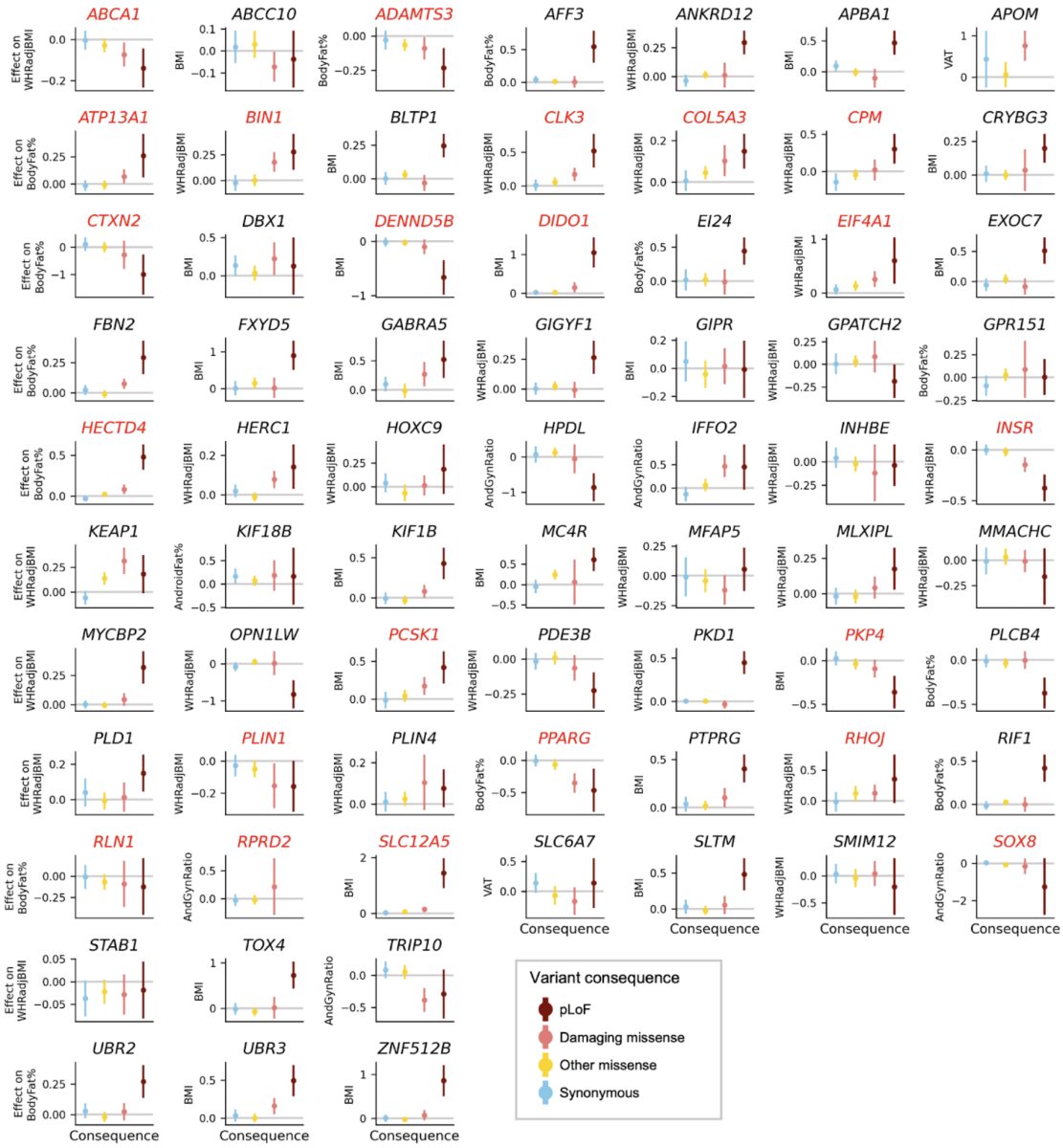

Figure S4.3: Ultra-rare variant allelic series for all FDR≤1% significant genes. Monotonic allelic series are highlighted with red gene symbols. Allelic series are defined as monotonic if they have strictly increasing or strictly decreasing point estimates of effect size across increasingly deleterious variant consequences (least to most deleterious: synonymous, other missense, damaging missense, pLoF). Effect sizes come from ultra-rare variant burden for variants with minor allele count ≤ 10 in each gene, grouped by variant consequence. Confidence intervals for effect size defined as ± 1.96 standard errors. pLoF, predicted loss-of-function.

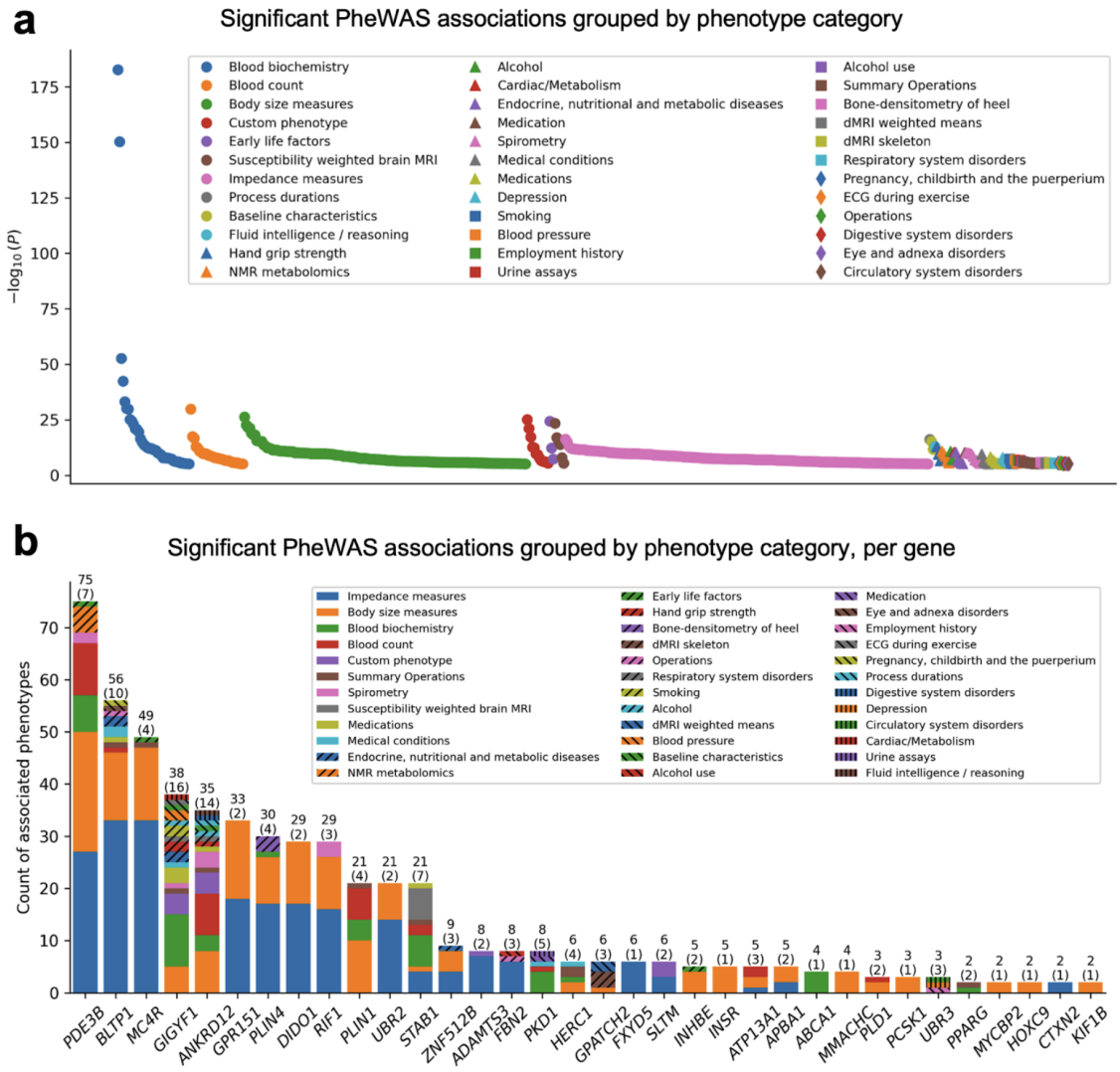

Figure S4.4: PheWAS of obesity and fat distribution associated genes using Genebass summary statistics. Only the pLoF variant mask results from Genebass are used. Significant associations are controlled for FDR≤ 1% (SKAT-O $P \leq 9.98 \times 10^{-6}$), resulting in 549 significant associations across 211 phenotypes and 41/69 obesity and fat distribution associated genes. **a**, Significant associations grouped by phenotype category, with phenotype groups ordered from left to right by the lowest P-value in the category. Significance of association is measured on the y-axis as $-\log 10$(Genebass SKAT-O P-value). **b**, Significant associations per gene, grouped by phenotype category. The total number of significant phenotype associations is shown at the top of each bar, with the number of phenotype categories shown in parentheses. Only genes with at least two phenotype associations are shown.

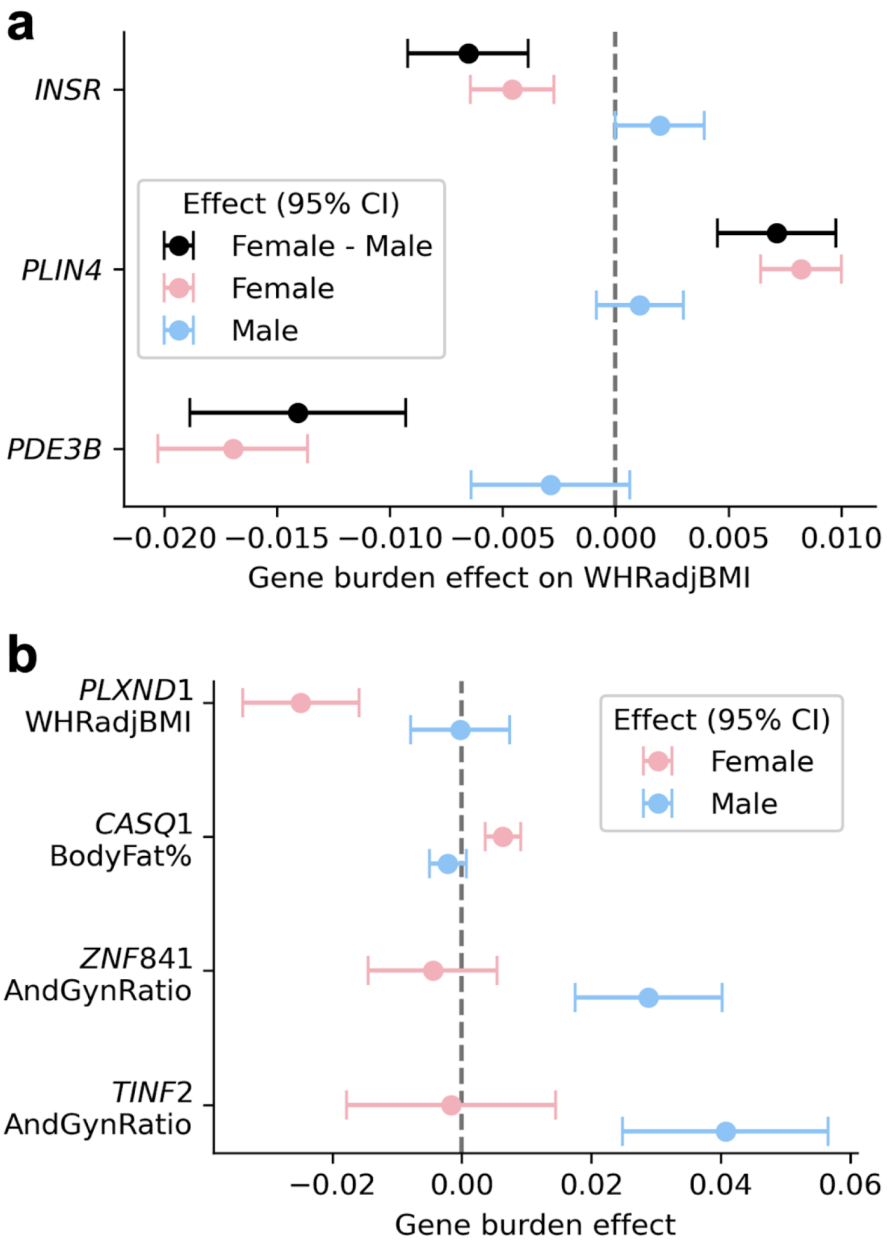

Figure S4.5: Sex-differential and sex-specific analysis. **a**, Genes with significant sex-differential effects (sex-difference P$<2.67 \times 10^{-6}$, Bonferroni adjusted for 18,737 genes tested for sex-differential effects). All three significant sex-differential gene burden effects are on WHRadjBMI. **b**, Female- (*PLXND1*, *CASQ1*) and male-specific (*ZNF841*, *TINF2*) gene-level significant associations ($P<2.67 \times 10^{-6}$, Bonferroni adjusted for 18,737 genes tested for sex-specific effects). CIs for effect size defined as ±1.96 standard errors.

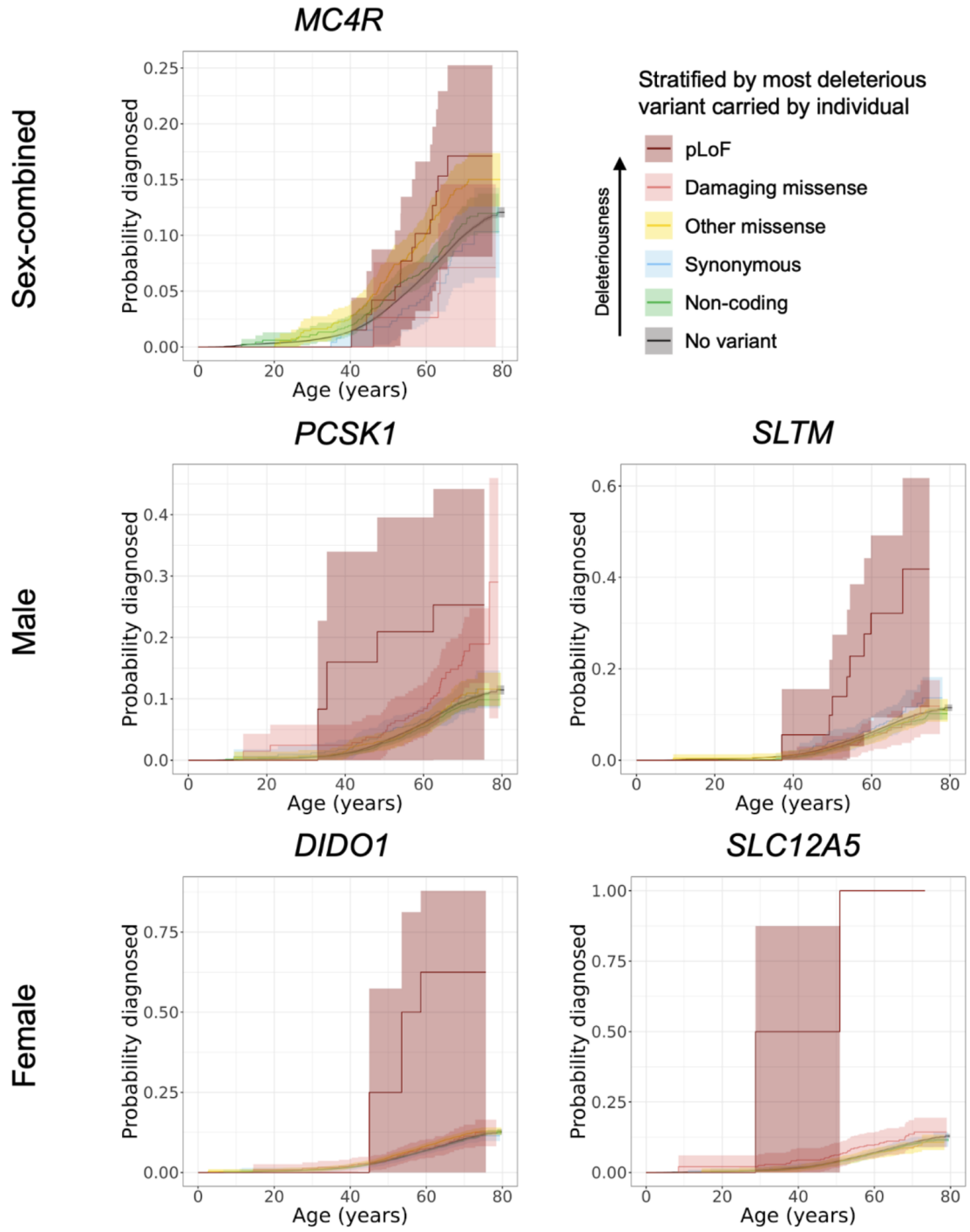

Figure S4.6: Longitudinal obesity age-at-onset analysis. Longitudinal analysis was performed using Cox-proportional hazards modelling (see "namere-fch:4:methods:longitudinal" methods). Individuals are stratified by the most deleterious carried in the gene. Only genes with significant age-of-onset associations in both sexes (*MC4R*), males only (*PCSK1*, *SLTM*), or females only (*DIDO1*, *SLC12A5*) are shown. 95% confidence intervals are indicated by shading.

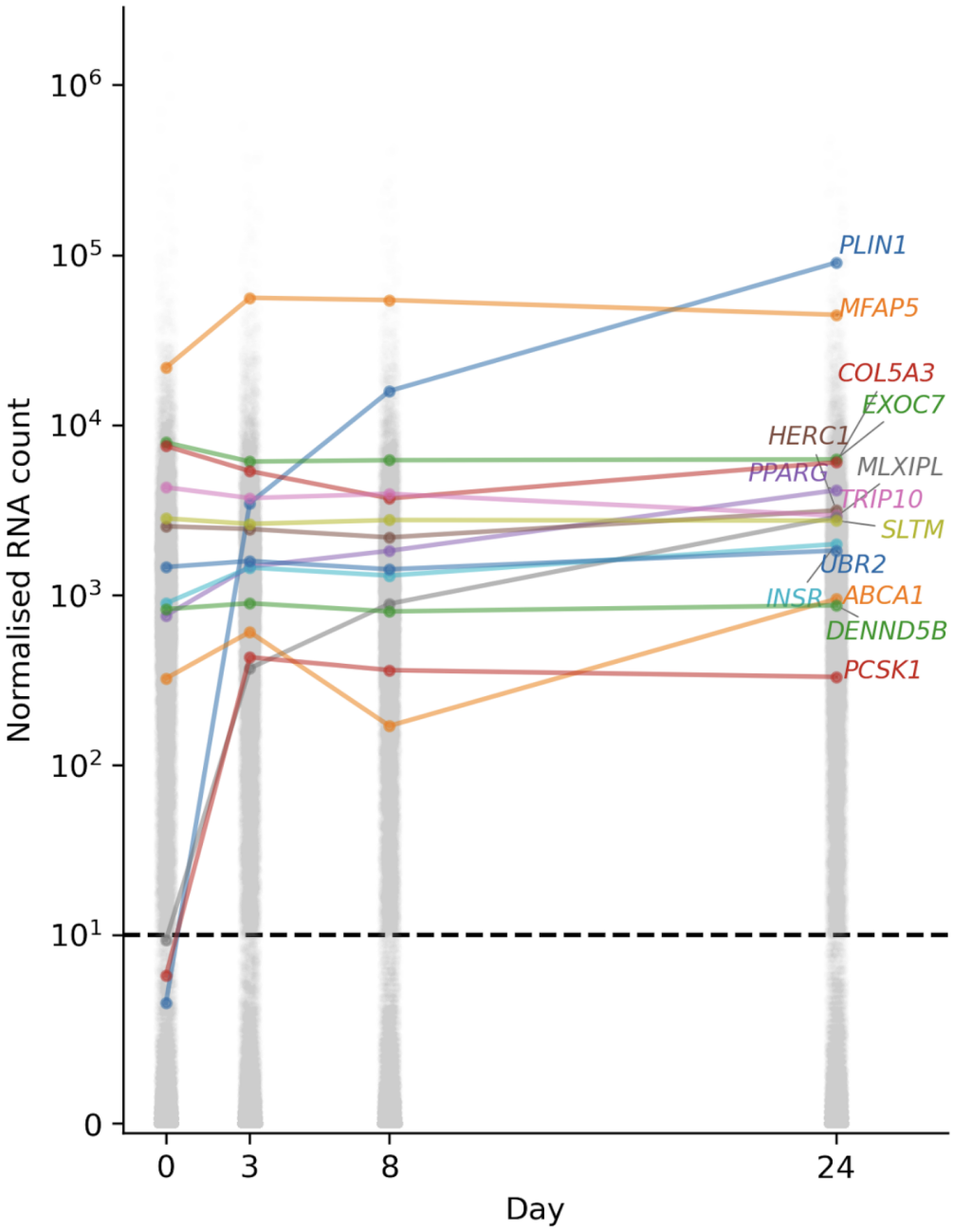

Figure S4.7: Longitudinal mRNA expression in wild type human white adipose tissue of genes selected for knockdown. Normalised mRNA count was measured at four time points across 24 days of differentiation. Day 0 corresponds to the undifferentiated state. Coloured traces are genes selected for knockdown. All other genes are indicated with grey points. The horizontal dashed line indicates the minimum threshold needed on days 8 and 24 for a gene to be selected as a knockdown target. The y-axis uses the 'symlog' scale, such that the scale is linear between 0 and 10 and logarithmic for values greater than 10.

S4.3 Supplementary tables

[h]

Some supplementary tables have been cropped due to size. The full supplementary tables can be downloaded here: https://doi.org/10.5281/zenodo.15275847

Table S4.1: Sample removal based on sample outliers defined by MAD thresholds. Threshold = median ±4 MADs, split by sequencing tranche.

QC Metric	# Samples in UK Biobank Whole Exome Sequencing Tranche		
	250k	**150k**	**50k**
Total before QC	**237,091**	**139,745**	**46,034**
Number of deletions	1,510	825	344
Number of insertions	1573	1,118	236
Number of SNPs	1,856	1,013	362
Ratio of insertions to deletions	1,741	954	339
Ratio of transitions to transversions	2,890	1,598	1781
Ratio of heterozygous variants to homozygous alternate variants	2,301	1,496	478
Total failing	10,861	6,396	3,238
Total passed QC	**226,230**	**133,349**	**42,796**
Grand total			**402,375**

Table S4.2: Sample sizes and phenotype SDs for each adiposity-related trait (sex-combined and sex-stratified).

Phenotype	UKB Field ID	Both sexes		Female		Male	
		N	Std. dev.	N	Std. dev.	N	Std. dev.
Body mass index (BMI)	21001	401059	4.74783	217716	5.11381	183343	4.22661
Waist-to-hip ratio, adjusted for BMI (WHRadjBMI)	n/a	400948	0.08064	217655	0.06263	183293	0.05214
Body fat percentage (BF%)	23099	395147	8.50612	214846	6.86752	180301	5.80123
Android tissue fat percentage	23247	39671	0.11148	20456	0.11721	19215	0.10322
Gynoid tissue fat percentage	23264	39671	0.09288	20456	0.06664	19215	0.06159
Android-gynoid tissue fat percentage ratio	n/a	39671	0.08305	20456	0.07372	19215	0.06631
Total tissue fat percentage	23281	39671	0.28970	20456	0.20838	19215	0.24515
Visceral adipose tissue (VAT) volume	22407	21253	2.28788	11000	1.52205	10253	2.35928
Abdominal fat ratio	22434	20687	0.11222	10892	0.10543	9795	0.09907

Table S4.3: Gene-level results for all FDR$\leq 1\%$ significant genes.

Gene symbol	Ensembl ID	Phenotype (for gene test with minimal p-value)	Consequence mask (mask with minimal p-value for trait)	Total minor allele count	Rare variants in mask (MAF<0.01, MAC>10)	Ultra-rare variants in mask (MAC≤10)	Total MAF	Common finemapped variants exist for conditioning	Significance category	SKAT-O P (conditioned)	Burden P (conditioned)	Burden effect size (conditioned)	Burden standard error (conditioned)	SKAT P (conditioned)	Unweighted burden effect size (conditioned)	SKAT-O P (unconditioned)	Burden P (unconditioned)	Burden effect size (unconditioned)	Burden standard error (unconditioned)	SKAT P (unconditioned)	Unweighted burden effect size (unconditioned)
ABCA1	ENSG00000165029	whr_adj_bmi	pLoF;damaging_missense	3825	32	362	0.004771106	TRUE	bh_sig_fdr_0.01_after_c	6.50E-07	3.07E-07	-0.002436154674	0.000475923084	7.39E-05	-0.05976051318	3.62E-07	1.80E-07	-0.002473999577	0.000474062101	4.91E-05	-0.0607089850
ABCC10	ENSG00000124574	bmi	pLoF;damaging_missense	9115	77	435	0.011364311	TRUE	bh_sig_fdr_0.01_after_c	2.06E-05	2.36E-05	-0.001855819322	0.000438921997	0.0003162002761	-0.04428010353	2.93E-05	3.04E-05	-0.001830433485	0.000438905181 3	0.0004510512741	-0.0436405023
ADAMTS3	ENSG00000156140	body_fat_percentage	pLoF;damaging_missense	4705	22	185	0.005951816	TRUE	bh_sig_fdr_0.01_after_c	5.07E-06	4.30E-06	-0.002157873895	0.000469473401	0.0006700957338	-0.05009559626	8.01E-06	7.13E-06	-0.002107341014	0.000469374504	0.0007292916 71	-0.0489743683
AFF3	ENSG00000144218	body_fat_percentage	pLoF	36	0	27	4.56E-05	TRUE	bh_sig_fdr_0.01_after_c	1.22E-05	1.22E-05	0.021764575943	0.004974986174	1.22E-05	0.5435096336	1.03E-05	1.03E-05	0.02193899849	0.004974664348	1.03E-05	0.547874762
ANKRD12	ENSG00000101745	whr_adj_bmi	pLoF	308	5	122	0.000383691	TRUE	exome_wide	1.71E-09	6.67E-08	0.009540854221	0.001766865559	1.30E-08	0.2377890298	1.74E-09	1.24E-08	0.009480547895	0.001664982217	1.59E-08	0.236285990
APBA1	ENSG00000107282	bmi	pLoF	128	2	57	0.00015893	TRUE	exome_wide	6.25E-08	5.12E-08	0.019059990448	0.003499139614	1.72E-06	0.4753837798	6.25E-08	5.18E-08	0.019053320652	0.003499122719	1.69E-06	0.475217405
APOM	ENSG00000204444	visceral_adipose_tissue_v	pLoF;damaging_missense	21	0	9	0.000494048	FALSE	bh_sig_fdr_0.01_after_c	4.37E-05	4.37E-05	0.03083239993	0.007544498851	4.37E-05	0.7617232826	4.37E-05	4.37E-05	0.03083239993	0.007544498851	4.37E-05	0.761723282
ATP13A1	ENSG00000105726	body_fat_percentage	pLoF	134	3	40	0.0001690636	TRUE	bh_sig_fdr_0.01_after_c	2.85E-05	1.45E-05	0.01120844556	0.00258476118	0.0009374454287	0.2798584019	3.05E-05	1.57E-05	0.01116459135	0.00258483302	0.001017304307	0.278762976
BIN1	ENSG00000136717	whr_adj_bmi	pLoF;damaging_missense	656	18	129	0.0008189259	TRUE	bh_sig_fdr_0.01_after_c	3.94E-06	0.0002118925194	0.00427057387	0.001152838525	4.31E-06	0.1066809532	4.15E-06	0.0002917528952	0.004176080203	0.001152812499	4.62E-06	0.104383366
BLTP1	ENSG00000138688	bmi	pLoF	600	7	275	0.000747103	TRUE	exome_wide	2.15E-10	1.72E-10	0.01044231736	0.001635544831	5.51E-08	0.2576612275	2.15E-10	9.19E-11	0.01059763181	0.001635492504	4.53E-08	0.261495815
CHMP4B	ENSG00000101421	body_fat_percentage	pLoF	2	0	1	2.53E-06	TRUE	bh_sig_fdr_0.01_after_c	3.88E-05	3.88E-05	-0.08673684377	0.02108012016	3.88E-05	-2.16768016	3.51E-05	3.51E-05	-0.08719996291	0.02107395004	3.51E-05	-2.1798653
CLK3	ENSG00000179335	whr_adj_bmi	pLoF;damaging_missense	1107	13	121	0.0013800314	TRUE	bh_sig_fdr_0.01_after_c	2.88E-06	3.32E-06	0.004059858565	0.000873072491	7.01E-06	0.100929561	3.66E-06	4.11E-06	0.00402122932	0.000873028403	7.89E-06	0.0999754439
COL5A3	ENSG00000080573	whr_adj_bmi	pLoF;damaging_missense	3111	58	304	0.003883862	TRUE	exome_wide	5.23E-10	2.98E-09	0.003124218761	0.000526599593 8	6.00E-06	0.07808121691	4.92E-10	2.89E-09	0.003126834517	0.000526590378 5	6.82E-06	0.0781562508
CPM	ENSG00000135678	whr_adj_bmi	pLoF	72	1	26	8.94E-05	TRUE	bh_sig_fdr_0.01_after_c	1.45E-05	1.21E-05	0.01492598581	0.003410829736	0.000246432292	0.3726466552	1.45E-05	1.21E-05	0.01492343443	0.003410713199	0.000246248366	0.372583577
CRYBG3	ENSG00000080200	bmi	pLoF	730	13	148	0.000908904	TRUE	bh_sig_fdr_0.01_after_c	2.61E-05	3.08E-05	0.006133808436	0.001471907645	0.0001319365054	0.1526598253	1.82E-05	2.11E-05	0.006260167834	0.001471803739	0.000118268715	0.155791966
CTXN2	ENSG00000233932	body_fat_percentage	pLoF	18	1	4	2.31E-05	TRUE	bh_sig_fdr_0.01_after_c	3.40E-05	2.94E-05	-0.0293588596	0.007026759778	0.000382375953	-0.7336896325	2.85E-05	2.65E-05	-0.02952599955	0.007026729353	0.000352012786	-0.737869244
DBX1	ENSG00000109851	bmi	pLoF;damaging_missense	397	7	56	0.0004956439	TRUE	bh_sig_fdr_0.01_after_c	2.29E-05	0.004428666811	0.005654716449	0.001986972335	1.07E-05	0.1412768243	2.11E-05	0.004286471339	0.005675181209	0.001986918722	9.59E-06	0.141787722
DEFB112	ENSG00000180872	tissuefatp_androidgynoidra	pLoF	1	0	1	1.26E-05	TRUE	bh_sig_fdr_0.01_after_c	3.20E-06	3.20E-06	0.1461730912	0.03138531749	3.20E-06	3.650312062	2.65E-06	2.65E-06	0.1472643288	0.03135934741	2.65E-06	3.68049361
DENND5B	ENSG00000170456	bmi	pLoF	38	0	32	4.74E-05	TRUE	bh_sig_fdr_0.01_after_c	3.35E-05	3.35E-05	-0.02662894327	0.006419517333	3.35E-05	-0.6649612096	2.79E-05	2.79E-05	-0.02689707127	0.006419343293	2.79E-05	-0.671662923
DIDO1	ENSG00000101191	bmi	pLoF	25	0	22	3.12E-05	TRUE	exome_wide	1.10E-07	1.10E-07	0.0420305057	0.007916791942	1.10E-07	1.049923541	1.14E-07	1.14E-07	0.04197333497	0.007916235879	1.14E-07	1.04854889
EI24	ENSG00000149547	body_fat_percentage	pLoF	54	0	33	6.83E-05	TRUE	bh_sig_fdr_0.01_after_c	1.49E-05	1.49E-05	0.01765913128	0.004077564947	1.49E-05	0.4407540674	1.40E-05	1.40E-05	0.01771361095	0.004077534376	1.40E-05	0.442115064
EIF4A1	ENSG00000161960	whr_adj_bmi	pLoF;damaging_missense	204	5	61	0.0002549395	TRUE	bh_sig_fdr_0.01_after_c	2.13E-05	0.02016976556	0.004700281827	0.002023214907	1.82E-05	0.1174938213	1.75E-05	0.01824507791	0.004776005287	0.002023211092	1.45E-05	0.119384385
EXOC7	ENSG00000182473	bmi	pLoF	157	2	46	0.0001962426	TRUE	bh_sig_fdr_0.01_after_c	1.97E-05	0.001186319614	0.01025576103	0.003163260197	1.20E-05	0.2560678942	2.06E-05	0.001197231715	0.01024734722	0.003163211746	1.27E-05	0.255856604
FBN2	ENSG00000138829	body_fat_percentage	pLoF	111	0	81	0.000140454	TRUE	bh_sig_fdr_0.01_after_c	3.75E-05	3.75E-05	0.01173022825	0.002845360819	3.75E-05	0.2922674842	4.14E-05	4.14E-05	0.01166502847	0.002845330418	4.14E-05	0.290643721
FEZF2	ENSG00000153266	abdominal_fat_ratio	pLoF;damaging_missense	2	0	2	4.83E-05	FALSE	bh_sig_fdr_0.01_after_c	2.74E-05	2.74E-05	-0.1061483647	0.02530985207	2.74E-05	-2.65062652	2.74E-05	2.74E-05	-0.1061483647	0.02530985207	2.74E-05	-2.6506265
FXYD5	ENSG00000089327	bmi	pLoF	36	1	7	4.49E-05	TRUE	bh_sig_fdr_0.01_after_c	5.89E-06	0.01280731691	0.01642248148	0.006597804669	4.79E-06	0.4104343666	6.01E-06	0.01298019294	0.01639095413	0.006597783359	5.01E-06	0.409644332
GABRA5	ENSG00000186297	bmi	pLoF;damaging_missense	207	5	77	0.0002603439	TRUE	bh_sig_fdr_0.01_after_c	2.45E-05	2.90E-05	0.011489610413	0.002749594387	0.0001736134299	0.2903894654	2.38E-05	2.70E-05	0.01154208481	0.002749550854	0.000151887545 13	0.291483371
GIGYF1	ENSG00000146830	whr_adj_bmi	pLoF	141	1	84	0.000176187	TRUE	bh_sig_fdr_0.01_after_c	3.32E-06	2.60E-06	0.01144511379	0.002435027885	4.92E-05	0.2853374942	3.29E-06	2.57E-06	0.01144985803	0.002434967899	4.80E-05	0.285455779
GIPR	ENSG00000010310	bmi	pLoF;damaging_missense	2835	22	139	0.0035321911	TRUE	bh_sig_fdr_0.01_after_c	2.32E-07	1.34E-05	-0.003319038569	0.000762327615	3.01E-07	-0.08183021639	2.46E-07	1.27E-05	-0.003327698243	0.000762332369	2.67E-07	-0.082047662
GLP1R	ENSG00000112164	abdominal_fat_ratio	pLoF	3	0	3	7.25E-05	FALSE	bh_sig_fdr_0.01_after_c	1.68E-06	1.68E-06	-0.09898641589	0.02067157806	1.68E-06	-2.470358256	1.68E-06	1.68E-06	-0.09898641589	0.02067157806	1.68E-06	-2.47035825
GPATCH2	ENSG00000092978	whr_adj_bmi	pLoF	246	4	30	0.0003065756	TRUE	bh_sig_fdr_0.01_after_c	3.71E-06	1.35E-05	-0.008031838959	0.00184531348	6.66E-06	-0.2003451759	5.65E-06	1.81E-05	-0.007911787643	0.001845345818	1.04E-05	-0.19734747
GPR151	ENSG00000173250	bmi	pLoF	8313	15	22	0.0103648152	TRUE	exome_wide	6.60E-08	8.80E-08	-0.002676161107	0.000500234455 4	2.28E-07	-0.0581131238	1.35E-07	2.02E-07	-0.002639589237	0.000507848213	3.10E-07	-0.0573468755
GPR151	ENSG00000173250	body_fat_percentage	pLoF	8209	15	22	0.0103868597	TRUE	exome_wide	6.60E-08	2.51E-07	-0.001990823695	0.000386067821 5	1.55E-07	-0.0434708934	6.61E-08	2.53E-07	-0.001990418158	0.000386063462 5	1.57E-07	-0.0434597026
HECTD4	ENSG00000173064	body_fat_percentage	pLoF;damaging_missense	2153	37	382	0.002726338	TRUE	bh_sig_fdr_0.01_after_c	4.36E-06	0.009691677998	0.001714149607	0.000626935262	2.52E-06	0.0430174486	3.14E-06	0.008677691303	0.00173921021	0.000662683103	2.19E-06	0.0436402990
HERC1	ENSG00000103657	whr_adj_bmi	pLoF;damaging_missense	3810	66	580	0.004756249	TRUE	bh_sig_fdr_0.01_after_c	3.46E-05	0.0002170822364	0.001750689765	0.004733819784	8.63E-05	0.04385842316	3.68E-05	0.0002235317394	0.001747131168	0.004473720083	9.95E-05	0.043766322
HOXC9	ENSG00000180806	whr_adj_bmi	pLoF	114	5	14	0.0001411643	TRUE	bh_sig_fdr_0.01_after_c	3.52E-05	2.31E-05	0.01347756375	0.003184355752	0.002724758552	0.3367096545	4.09E-05	2.63E-05	0.01338453962	0.003184255472	0.00296824781	0.334386543
HPDL	ENSG00000186603	tissuefatp_androidgynoidra	pLoF	15	0	6	0.000189055	TRUE	bh_sig_fdr_0.01_after_c	2.08E-05	2.08E-05	-0.03470516736	0.008153680388	2.08E-05	-0.8636777246	2.22E-05	2.22E-05	-0.03458655772	0.00815333699	2.22E-05	-0.860749591
IFFO2	ENSG00000169991	tissuefatp_androidgynoidra	pLoF;damaging_missense	78	2	22	0.000982391	TRUE	bh_sig_fdr_0.01_after_c	2.87E-05	0.001079534287	0.01177750259	0.003602857113	2.44E-05	0.0857105192	2.87E-05	0.001175256796	0.01168039162	0.003602787993	2.42E-05	0.290188785
INHBE	ENSG00000139269	whr_adj_bmi	pLoF	954	7	16	0.0011891288	TRUE	exome_wide	3.27E-09	1.37E-07	-0.004955141013	0.000940963467	2.07E-07	-0.1228934469	1.46E-08	1.39E-07	-0.004984235127	0.000946401497 8	1.72E-07	-0.123814776
INSR	ENSG00000171105	whr_adj_bmi	pLoF	151	2	53	0.000188163	TRUE	exome_wide	6.38E-08	4.20E-07	-0.01197157795	0.002366031914	8.15E-08	-0.2986388617	6.58E-08	5.59E-07	-0.01180797572	0.002359217604	8.62E-08	-0.294561626
KEAP1	ENSG00000079999	whr_adj_bmi	pLoF;damaging_missense	273	4	84	0.0003407624	TRUE	exome_wide	1.76E-11	2.52E-12	0.01226552267	0.001751679263	7.43E-08	0.3054322307	2.84E-11	4.05E-12	0.01214844811	0.001751670505	7.76E-08	0.302521303
KIF18B	ENSG00000186185	android_tissue_fatp	pLoF;damaging_missense	932	3	26	0.011746165	FALSE	bh_sig_fdr_0.01_after_c	1.59E-05	2.00E-05	-0.006563483449	0.001538894268	4.51E-05	-0.1393643276	1.59E-05	2.00E-05	-0.006563483449	0.001538894268	4.51E-05	-0.139364327
KIF1B	ENSG00000054523	bmi	pLoF	98	0	68	0.000122177	TRUE	bh_sig_fdr_0.01_after_c	1.85E-05	1.85E-05	0.01715217678	0.004005713897	1.85E-05	0.4275478056	1.53E-05	1.53E-05	0.01732195588	0.004005621369	1.53E-05	0.431780750
MC4R	ENSG00000166603	bmi	pLoF	187	5	18	0.0002343482	TRUE	exome_wide	1.05E-12	2.03E-13	0.02129722972	0.002898912275	6.16E-09	0.5317709455	9.30E-13	1.87E-13	0.02132873889	0.002898803217	5.43E-09	0.532557276
MFAP5	ENSG00000197614	whr_adj_bmi	pLoF;damaging_missense	2042	14	56	0.0025481857	TRUE	bh_sig_fdr_0.01_after_c	7.42E-06	4.10E-05	-0.002669990051	0.006508946296	1.29E-05	-0.06588517092	3.98E-06	2.78E-05	-0.002727441079	0.006507992224	7.71E-06	-0.0673111220
MLXIPL	ENSG00000009950	whr_adj_bmi	pLoF	159	3	41	0.00019823 4	TRUE	bh_sig_fdr_0.01_after_c	1.17E-05	9.52E-06	0.01013993553	0.002290004237	0.002300995521	0.2529458908	1.17E-05	9.32E-06	0.01015010568	0.002289977492	0.00227531357	0.253199723
MMACHC	ENSG00000132763	whr_adj_bmi	pLoF	819	4	13	0.0010214232	TRUE	bh_sig_fdr_0.01_after_c	1.84E-05	0.000119101599	0.00396480116	0.001030362385	1.87E-05	0.09741590583	2.19E-05	0.000135771783 4	0.003931524449	0.001030342122	2.15E-05	0.0965999057
MYCBP2	ENSG00000005810	whr_adj_bmi	pLoF	137	2	77	0.000170199	TRUE	bh_sig_fdr_0.01_after_c	2.05E-06	3.42E-06	0.011483911593	0.002472874029	7.75E-06	0.2864688063	3.17E-06	3.79E-06	0.01143075184	0.002472895082	8.34E-06	0.285143424
OPN1LW	ENSG00000102076	whr_adj_bmi	pLoF	14	0	7	1.75E-05	TRUE	bh_sig_fdr_0.01_after_c	2.38E-05	2.38E-05	-0.02395528973	0.007798002714	2.38E-05	-0.8233367831	2.51E-05	2.51E-05	-0.02385378668	0.007797129645	2.51E-05	-0.821001598
PCBD2	ENSG00000132570	gynoid_tissue_fatp	pLoF;damaging_missense	4	0	4	5.04E-05	FALSE	bh_sig_fdr_0.01_after_c	1.25E-05	1.25E-05	-0.06219051652	0.01423691875	1.25E-05	-1.552549056	1.25E-05	1.25E-05	-0.06219051652	0.01423691875	1.25E-05	-1.55294805
PCSK1	ENSG00000175426	bmi	pLoF;damaging_missense	1281	22	155	0.001598591	TRUE	bh_sig_fdr_0.01_after_c	1.44E-05	3.96E-05	0.004569197442	0.001111770986	2.81E-05	0.1136116967	1.99E-05	5.47E-05	0.004484645221	0.001111605319	3.58E-05	0.111508622
PDE3B	ENSG00000152270	whr_adj_bmi	pLoF	1057	8	60	0.001318139	TRUE	exome_wide	5.91E-10	7.09E-10	-0.006755271716	0.001095904367	7.38E-08	-0.1660829845	2.86E-17	5.09E-18	-0.007841328126	0.000906373179 7	4.75E-16	-0.19242940
PDE3B	ENSG00000152270	body_fat_percentage	pLoF	1046	8	58	0.001323577	TRUE	exome_wide	1.34E-07	2.09E-07	0.004901635221	0.000944178038 3	3.27E-07	0.1203964977	1.34E-07	2.11E-07	0.004899350731	0.000944088144	3.35E-07	0.120339752
PKD1	ENSG00000008710	whr_adj_bmi	pLoF	176	2	95	0.000218657	TRUE	exome_wide	7.68E-11	8.50E-07	0.0124505851	0.002528795553	1.12E-11	0.3106894993	7.23E-11	8.17E-07	0.01246999187	0.002528810889	1.06E-11	0.311172882
PKP4	ENSG00000144283	bmi	pLoF	176	2	77	0.000220137	TRUE	bh_sig_fdr_0.01_after_c	4.20E-05	4.85E-05	-0.012149565 66	0.002990640659	7.35E-05	-0.3030428995	4.28E-05	4.98E-05	-0.01213054376	0.002990493866	7.79E-05	-0.302566422
PLCB4	ENSG00000101333	body_fat_percentage	pLoF	92	1	48	0.0001161053	TRUE	bh_sig_fdr_0.01_after_c	2.85E-05	0.000105249444	-0.012141966 32	0.003130857413	2.82E-05	-0.3030742788	3.40E-05	0.000118850661	-0.01204897277	0.003130835469	3.11E-05	-0.300755757
PLD1	ENSG00000075651	whr_adj_bmi	pLoF;damaging_missense	9324	41	199	0.011632087	TRUE	exome_wide	4.69E-10	2.88E-09	0.001887202574	0.000317803064 2	2.66E-05	0.04406445691	1.03E-07	5.57E-08	0.001842300073	0.000339145829 5	4.39E-05	0.043002067 8
PLIN1	ENSG00000166819	whr_adj_bmi	pLoF	836	12	36	0.001042011	TRUE	exome_wide	8.94E-23	8.98E-21	-0.009371330447	0.001002557993	2.66E-17	-0.2334085689	4.36E-23	5.92E-21	-0.009415287963	0.001002539077	1.18E-17	-0.23450775
PLIN4	ENSG00000167676	whr_adj_bmi	pLoF	3560	19	102	0.00443755	TRUE	exome_wide	1.00E-13	3.25E-14	0.003776329467	0.000497682254 1	4.36E-11	0.09130005671	1.45E-13	4.31E-14	0.00375417304 2	0.000497157722 7	7.38E-11	0.090749532
PLIN4	ENSG00000167676	body_fat_percentage	pLoF	3506	19	101	0.004435092	TRUE	exome_wide	1.23E-10	2.11E-10	-0.003301576271	0.000519703198 4	1.28E-09	-0.0800915789	1.83E-10	2.98E-10	-0.003277163314	0.000519686246	1.85E-09	-0.079493977
PPARG	ENSG00000132170	body_fat_percentage	pLoF;damaging_missense	168	4	62	0.0002139847	TRUE	exome_wide	2.90E-08	4.34E-07	0.011730631165	0.002321309053	9.37E-08	-0.2923395165	2.90E-08	2.94E-07	-0.011847096897	0.002310660912	4.23E-08	-0.29556758
PTPRG	ENSG00000144724	bmi	pLoF	194	1	119	0.000242159	TRUE	exome_wide	2.90E-08	1.15E-07	0.01515621282	0.002858979803	3.13E-08	0.3771484675	1.37E-08	8.31E-08	0.01528245049	0.002851022423	2.39E-08	0.380291959
PTPRG	ENSG00000144724	body_fat_percentage	pLoF	192	1	119	0.000243233	TRUE	exome_wide	1.35E-07	8.25E-07	0.01073035686	0.002176862644	1.46E-07	0.2670474882	6.53E-08	5.09E-07	0.01091957858	0.002173930635	8.52E-08	0.271759958
RHOJ	ENSG00000126785	whr_adj_bmi	pLoF;damaging_missense	8561	11	69	0.0106763866	TRUE	bh_sig_fdr_0.01_after_c	3.23E-06	3.62E-06	0.001775156643	0.000383223679 5	0.0001052310138	0.03553327592	4.67E-06	4.06E-06	0.00176621411	0.000383265885 2	0.000117497726 3	0.035341492
RIF1	ENSG00000080345	body_fat_percentage	pLoF	89	0	66	0.000112616	TRUE	exome_wide	1.41E-07	1.41E-07	0.01668790568	0.00317017555	1.41E-07	0.4160680706	1.82E-07	1.82E-07	0.01653590076	0.003169972048	1.82E-07	0.412282696
RLN1	ENSG00000107018	body_fat_percentage	pLoF;damaging_missense	534	8	22	0.0006757369	TRUE	bh_sig_fdr_0.01_after_c	1.38E-06	1.41E-06	-0.006250266792	0.001295656893	0.000116231024	-0.1556433538	2.21E-06	1.79E-06	-0.006185008241	0.001294949483	0.000121729287 2	-0.154022130
RPRD2	ENSG00000163125	tissuefatp_androidgynoidra	pLoF;damaging_missense	288	2	7	0.003630433	TRUE	bh_sig_fdr_0.01_after_c	1.28E-05	6.61E-06	0.008684174316	0.001927321149	4.23E-05	0.20869415	1.59E-05	8.19E-06	0.008594273268	0.001926903115	5.55E-05	0.206222999
SLC12A5	ENSG00000124140	bmi	pLoF	13	0	13	1.62E-05	TRUE	exome_wide	1.11E-07	1.11E-07	0.05786782956	0.01090122219	1.11E-07	1.446113428	1.21E-07	1.21E-07	0.05768101435	0.01090087835	1.21E-07	1.44146283
SLC6A7	ENSG00000011083	visceral_adipose_tissue_vi	pLoF;damaging_missense	351	4	35	0.008257777	FALSE	bh_sig_fdr_0.01_after_c	1.80E-05	1.11E-05	-0.008575697312	0.001951113865	0.000401023945 2	-0.2011382407	1.80E-05	1.11E-05	-0.008575697312	0.001951113865	0.000401023945 2	-0.201138240
SLTM	ENSG00000137776	bmi	pLoF	76	0	51	9.47E-05	TRUE	bh_sig_fdr_0.01_after_c	2.24E-05	2.24E-05	0.01930245953	0.004552922555	2.24E-05	0.4814592708	2.65E-05	2.65E-05	0.01912871388	0.004552801935	2.65E-05	0.477131777
SMIM12	ENSG00000163866	whr_adj_bmi	pLoF;damaging_missense	183	3	21	0.0002298691	TRUE	bh_sig_fdr_0.01_after_c	3.08E-05	0.000727195126 8	0.007232842706	0.002108140395 3	2.62E-05	0.1809963309	2.52E-05	0.000579415005	0.007356580542	0.002140381772	2.19E-05	0.184289565
SOX8	ENSG00000005513	tissuefatp_androidgynoidra	pLoF;damaging_missense	26	1	12	0.0003280547	FALSE	bh_sig_fdr_0.01_after_c	2.39E-05	3.77E-05	-0.02546412981	0.006179134931	2.15E-05	-0.6336512293	2.39E-05	3.77E-05	-0.02546412981	0.006179134931	2.15E-05	-0.633651229
STAB1	ENSG00000010327	whr_adj_bmi	pLoF	1866	30	211	0.002327263	TRUE	bh_sig_fdr_0.01_after_c	2.82E-05	1.46E-05	-0.002910936713	0.000671613267 8	0.04054473086	-0.07215964066	2.41E-05	1.28E-05	-0.002930754094	0.000671568257 3	0.031204083 87	-0.072663994
TM4SF20	ENSG00000168955	tissuefatp_androidgynoidra	pLoF	9	0	8	0.000113433	FALSE	bh_sig_fdr_0.01_after_c	3.05E-05	3.05E-05	-0.0436834662	0.01047687508	3.05E-05	-1.089115869	3.05E-05	3.05E-05	-0.0436834662	0.01047687508	3.05E-05	-1.08911586
TOX4	ENSG00000092203	bmi	pLoF	42	0	30	5.24E-05	TRUE	bh_sig_fdr_0.01_after_c	1.77E-06	1.77E-06	0.02914954438	0.008100735812	1.77E-06	0.7277970868	1.93E-06	1.93E-06	0.02904037385	0.008100426588	1.93E-06	0.725097096
TRIP10	ENSG00000125733	tissuefatp_androidgynoidra	pLoF;damaging_missense	159	4	50	0.002003312	FALSE	bh_sig_fdr_0.01_after_c	1.48E-05	0.005675829596	-0.006974956096	0.002521727236	1.28E-05	-0.1734557157	1.48E-05	0.005675829596	-0.006974956096	0.002521727236	1.28E-05	-0.173455715
UBR2	ENSG00000024048	body_fat_percentage	pLoF	138	1	71	0.0001074312	TRUE	bh_sig_fdr_0.01_after_c	1.27E-05	1.18E-05	0.01118232597	0.002552081787	6.55E-05	0.2786577168	1.05E-05	1.03E-05	0.01125942577	0.002552196106	5.86E-05	0.280578135
UBR3	ENSG00000144357	bmi	pLoF;damaging_missense	790	15	227	0.000987187	TRUE	bh_sig_fdr_0.01_after_c	8.99E-07	1.73E-06	0.00676213404	0.001413724026	2.21E-06	0.1674282328	8.98E-07	1.59E-06	0.006785373255	0.001413913893	2.43E-06	0.167979096
VGF	ENSG00000128564	bmi	pLoF	8	0	8	9.97E-06	TRUE	bh_sig_fdr_0.01_after_c	3.87E-05	3.87E-05	-0.05718067825	0.01389582943	3.87E-05	-1.429127994	2.69E-05	2.69E-05	-0.0583330163	0.01389499756	2.69E-05	-1.45797652
ZNF512B	ENSG00000196700	bmi	pLoF	41	1	25	5.14E-05	TRUE	exome_wide	3.16E-07	4.98E-07	0.03121621 32	0.006209738543	1.23E-06	0.7798556768	3.10E-07	4.43E-07	0.03133610415	0.006205735655	1.11E-06	0.782850773

Table S4.4: Replication of UKB-significant genes in AoU.

gene_id	gene_symbc	Pvalue_SKA	Pvalue_Burc	Pvalue_SKA	BETA_Burde	annotation	EUR_or_met	phenotype	Direction
ENSG00000166	MC4R	7.40E-15	8.96E-08	7.14E-13	5.3466	pLoF;missenseL	META	BMI	+
ENSG00000166	MC4R	2.07E-12	5.48E-06	8.53E-12	-4.5454	pLoF;missenseL	META	WHRadjBMI	-
ENSG00000137	SLTM	2.75E-07	5.74E-08	2.75E-07	5.4266	pLoF	META	BMI	+
ENSG00000166	MC4R	1.53E-06	3.20E-06	1.21E-05	0.0064564	pLoF;missenseL	EUR	BMI	+
ENSG00000166	MC4R	1.65E-06	1.89E-06	3.83E-06	-4.7647	pLoF	META	WHRadjBMI	-
ENSG00000137	SLTM	2.45E-06	5.61E-07	2.45E-06	-5.0041	pLoF	META	WHRadjBMI	-
ENSG00000166	MC4R	6.28E-06	3.28E-06	1.12E-05	0.021425	pLoF	EUR	BMI	+
ENSG00000010	GIPR	7.24E-06	8.65E-08	0.008373	5.3531	pLoF;missenseL	META	WHRadjBMI	+
ENSG00000166	MC4R	8.10E-06	3.97E-06	2.35E-05	-0.021906	pLoF	EUR	WHRadjBMI	-
ENSG00000008	PKD1	8.60E-06	8.60E-06	8.60E-06	0.022502	pLoF	EUR	WHRadjBMI	+
ENSG00000166	MC4R	1.42E-05	1.15E-05	0.0001393	-0.006059	pLoF;missenseL	EUR	WHRadjBMI	-
ENSG00000173	HECTD4	1.65E-05	2.32E-06	1.44E-05	4.7232	pLoF	META	BMI	+
ENSG00000173	HECTD4	1.72E-05	3.17E-05	1.43E-05	0.016465	pLoF	EUR	BMI	+
ENSG00000008	PKD1	7.95E-05	2.73E-05	7.95E-05	4.1945	pLoF	META	WHRadjBMI	+
ENSG00000166	MC4R	7.97E-05	1.63E-06	0.00011547	4.795	pLoF	META	BMI	+
ENSG00000010	GIPR	0.00012815	6.08E-06	0.01902	-4.5236	pLoF;missenseL	META	BMI	-
ENSG00000173	HECTD4	0.00017125	0.00060071	0.00011071	-0.013348	pLoF	EUR	WHRadjBMI	-
ENSG00000010	GIPR	0.00022992	0.00014742	0.0025596	0.0029237	pLoF;missenseL	EUR	WHRadjBMI	+
ENSG00000167	PLIN4	0.00025388	0.032061	0.02166	-2.1437	pLoF	META	BMI	-
ENSG00000101	DIDO1	0.00028511	0.00028511	0.00028511	0.041353	pLoF	EUR	BMI	+
ENSG00000054	KIF1B	0.00029211	7.08E-05	0.00029211	3.9737	pLoF	META	BMI	+
ENSG00000137	SLTM	0.0002932	0.0002932	0.0002932	0.023424	pLoF	EUR	BMI	+
ENSG00000092	TOX4	0.00030118	0.00030118	0.00030118	0.045052	pLoF	EUR	BMI	+
ENSG00000167	PLIN4	0.00031559	0.027092	0.012077	2.2102	pLoF	META	WHRadjBMI	+
ENSG00000092	TOX4	0.00041449	0.00041449	0.00041449	-0.044199	pLoF	EUR	WHRadjBMI	-
ENSG00000101	PLCB4	0.00098164	0.00054265	0.0019407	-3.4588	pLoF	META	BMI	-
ENSG00000173	HECTD4	0.0010971	0.0001429	0.0008292	-3.8031	pLoF	META	WHRadjBMI	-
ENSG00000166	PLIN1	0.0018104	0.0011765	0.0083134	-0.010569	pLoF	EUR	WHRadjBMI	-
ENSG00000010	GIPR	0.0022786	0.0013489	0.022363	-0.0024798	pLoF;missenseL	EUR	BMI	-
ENSG00000146	GIGYF1	0.0023005	0.0017361	0.0035299	0.01324	pLoF	EUR	BMI	+
ENSG00000079	KEAP1	0.0023922	0.00045619	0.0023922	3.5052	pLoF	META	WHRadjBMI	+
ENSG00000166	PLIN1	0.0025892	0.012053	0.0033899	-2.5106	pLoF	META	WHRadjBMI	-
ENSG00000101	ANKRD12	0.00263	0.00093389	0.0071512	3.3097	pLoF	META	BMI	+
ENSG00000101	DIDO1	0.002737	0.00056461	0.002737	3.4481	pLoF	META	BMI	+
ENSG00000092	GPATCH2	0.0028288	0.51592	0.013815	-0.64964	pLoF;missenseL	META	BMI	-
ENSG00000092	TOX4	0.0029042	0.028068	0.0029042	2.1963	pLoF	META	BMI	+
ENSG00000169	IFFO2	0.0030078	0.90815	0.0027227	0.11538	pLoF;missenseL	META	WHRadjBMI	+
ENSG00000101	ANKRD12	0.0034128	0.0034128	0.0034128	0.013358	pLoF	EUR	BMI	+
ENSG00000024	UBR2	0.0037231	0.00085835	0.0037231	3.3333	pLoF	META	BMI	+
ENSG00000175	PCSK1	0.003998	0.003998	0.003998	0.018959	pLoF	EUR	BMI	+
ENSG00000101	DIDO1	0.0040107	0.0040107	0.0040107	-0.034267	pLoF	EUR	WHRadjBMI	-
ENSG00000054	KIF1B	0.0040489	0.00087856	0.0040489	-3.3268	pLoF	META	WHRadjBMI	-
ENSG00000175	PCSK1	0.0041431	0.004402	0.0040848	0.0030849	pLoF;missenseL	EUR	BMI	+
ENSG00000137	SLTM	0.0050448	0.0050448	0.0050448	-0.018221	pLoF	EUR	WHRadjBMI	-
ENSG00000107	APBA1	0.005286	0.005286	0.005286	0.018716	pLoF	EUR	BMI	+
ENSG00000146	GIGYF1	0.0068503	0.0011467	0.0090641	3.2518	pLoF	META	BMI	+
ENSG00000101	DIDO1	0.0069891	0.001891	0.0069891	-3.1068	pLoF	META	WHRadjBMI	-
ENSG00000101	PLCB4	0.0070801	0.005628	0.014916	-0.015333	pLoF	EUR	BMI	-
ENSG00000105	ATP13A1	0.0071693	0.0013068	0.0071693	3.2145	pLoF	META	BMI	+
ENSG00000196	ZNF512B	0.0079381	0.89542	0.0045845	8.63E-05	pLoF;missenseL	EUR	WHRadjBMI	+
ENSG00000101	ANKRD12	0.0079431	0.17704	0.0074052	1.3499	pLoF;missenseL	META	WHRadjBMI	+
ENSG00000169	IFFO2	0.00814	0.61095	0.0045091	-0.00050674	pLoF;missenseL	EUR	WHRadjBMI	-
ENSG00000144	UBR3	0.0088912	0.0088912	0.0088912	0.014708	pLoF	EUR	BMI	+
ENSG00000197	MFAP5	0.0093856	0.0093856	0.0093856	0.017115	pLoF	EUR	BMI	+
ENSG00000105	ATP13A1	0.0095588	0.0095588	0.0095588	0.018684	pLoF	EUR	BMI	+
ENSG00000144	UBR3	0.01121	0.00081587	0.01121	3.3474	pLoF	META	BMI	+
ENSG00000175	PCSK1	0.011291	0.0096888	0.011291	2.5867	pLoF	META	BMI	+
ENSG00000196	ZNF512B	0.013058	0.13374	0.0075505	0.0011133	pLoF;missenseL	EUR	BMI	+
ENSG00000101	PLCB4	0.013391	0.011506	0.026899	2.527	pLoF	META	WHRadjBMI	+
ENSG00000107	APBA1	0.014668	0.0046776	0.014668	-2.8284	pLoF	META	WHRadjBMI	-
ENSG00000196	ZNF512B	0.015039	0.015039	0.015039	0.029096	pLoF	EUR	BMI	+
ENSG00000080	COL5A3	0.016894	0.0033974	0.013595	2.9293	pLoF	META	WHRadjBMI	+
ENSG00000092	TOX4	0.017002	0.012166	0.017002	-2.5073	pLoF	META	WHRadjBMI	-
ENSG00000101	PLCB4	0.02008	0.24349	0.025874	1.1663	pLoF;missenseL	META	WHRadjBMI	+
ENSG00000132	PPARG	0.020132	0.26387	0.020132	1.1173	pLoF	META	WHRadjBMI	+
ENSG00000079	KEAP1	0.020712	0.020712	0.020712	0.024125	pLoF	EUR	WHRadjBMI	+
ENSG00000008	PKD1	0.021105	0.021105	0.021105	-0.01172	pLoF	EUR	BMI	-

Table S4.5: Phenome-wide associations from Genebass.

gene_symb	annotation	description	description	category	Pvalue_SK/	Pvalue_Bur	BETA_Burd	SE_Burden	Pvalue_SK/	total_variar	n_cases	n_controls	heritability	saige_versi	inv_normal	trait_type	phenocode	pheno_sex	coding	modifier	n_cases_de	n_cases_bo	n_cases_fe	n_cases_m	coding_des	gene_id	bmi rg	bmi rg pval	body_fat_pe	body_fat_p
ABCA1	pLoF	Apolipoprotein /	Measured by im	Blood biochemi	2.25E-183	3.22E-184	-0.0543259	0.00187685	9.23E-32	107	343018		0	SAIGE_0.44.5	FALSE	continuous	30630	both_sexes		imt	423583	427532	227111	196472		ENSG00000165029				
ABCA1	pLoF	HDL cholestero	Measured by er	Blood biochemi	6.12E-151	8.74E-152	-0.0480374	0.00183053	3.28E-27	109	344938		0	SAIGE_0.44.5	FALSE	continuous	30760	both_sexes		imt	425927	429894	229155	196772		ENSG00000165029				
PDE3B	pLoF	Ratio of circulat	Triglycerides m	Blood biochemi	2.54E-53	3.62E-54	-0.0197457	0.00127414	2.58E-46	64	343470		0	SAIGE_0.44.5	FALSE	continuous	ratio_TG_HDL	both_sexes		custom	423734	427677	228091	195643		ENSG00000152270				
PDE3B	pLoF	Triglycerides	Measured by G	Blood biochemi	4.91E-43	7.01E-44	-0.0164616	0.0011849	7.82E-35	68	376509		0	SAIGE_0.44.5	FALSE	continuous	30870	both_sexes		imt	464971	469240	252065	212906		ENSG00000152270				
PKD1	pLoF	Cystatin C	Measured by la	Blood biochemi	8.56E-34	1.22E-34	0.0220065	0.00179268	9.14E-06	90	376784		0	SAIGE_0.44.5	FALSE	continuous	30720	both_sexes		imt	465312	469584	252198	213114		ENSG00000008710				
PDE3B	pLoF	HDL cholestero	Measured by er	Blood biochemi	9.62E-31	1.37E-31	0.0135415	0.00115803	2.98E-26	65	344938		0	SAIGE_0.44.5	FALSE	continuous	30760	both_sexes		imt	425927	429894	229155	196772		ENSG00000152270				
BLTP1/KIAA11(	pLoF	Mean platelet (t	Result of "Mear	Blood count	1.83E-30	6.55E-24	-0.0163586	0.00162235	1.28E-05	261	383461		0.0978906034	SAIGE_0.44.5	FALSE	continuous	30100	both_sexes		imt	473354	478192	256116	217238		ENSG00000131	0.04768	0.0005156	0.04982	0.0005469
PKD1	pLoF	Creatinine	Measured by er	Blood biochemi	1.85E-30	2.64E-31	0.0185395	0.001593	1.29E-05	90	376624		0	SAIGE_0.44.5	FALSE	continuous	30700	both_sexes		imt	465112	469383	252094	213018		ENSG00000008710				
PDE3B	pLoF	Waist-to-hip rati	Waist circumfer	Body size meas	6.42E-27	9.17E-28	-0.0194766	0.00178345	5.85E-21	51	212912		0.0400681377	SAIGE_0.44.5	FALSE	continuous	WHRadjBMI_fe	females		custom	263681	271844	263542	139		ENSG00000152270				
PKD1	pLoF	Urea	Measured by G	Blood biochemi	8.51E-26	1.22E-26	0.0203996	0.00190944	0.000160096	90	376551		0	SAIGE_0.44.5	FALSE	continuous	30670	both_sexes		imt	465031	469297	252046	212985		ENSG00000008710				
SLTM	pLoF	Corneal_resistance_factor_left_	Custom phenot		1.01E-25	0.0100466	0.0617224	0.0239771	1.44E-26	16	83193		0.0368175617	SAIGE_0.44.5	FALSE	continuous	Corneal_resista	both_sexes		custom	108123	111928	58239	49884		ENSG00000137776				
ABCA1	pLoF	Ratio of circulat	Triglycerides m	Blood biochemi	4.75E-25	6.79E-26	0.0211594	0.00201083	5.61E-05	109	343470		0	SAIGE_0.44.5	FALSE	continuous	ratio_TG_HDL	both_sexes		custom	423734	427677	228091	195643		ENSG00000165029				
MC4R	pLoF	Comparative bc	ACE touchscree	Early life factor	5.17E-25	7.39E-26	0.0212604	0.00202196	1.23E-12	21	388179		0	SAIGE_0.44.5	FALSE	continuous	1687	both_sexes			477982	491484	260083	217899		ENSG00000165	0.5636	3.89E-223	0.3718	9.48E-71
STAB1	pLoF	Median T2star (	Median T2star (	Susceptibility w	4.58E-24	8.47E-24	0.0482629	0.00479847	1.20E-07	39	16739		0.0925639851	SAIGE_0.44.5	FALSE	continuous	25031	both_sexes		imt	19176	19664	10091	9085		ENSG00000010327				
PLIN1	pLoF	HDL cholestero	Measured by er	Blood biochemi	1.27E-23	2.22E-21	0.0122144	0.00128652	4.40E-18	43	344938		0	SAIGE_0.44.5	FALSE	continuous	30760	both_sexes		imt	425927	429894	229155	196772		ENSG00000166819				
PLIN1	pLoF	Waist-to-hip rati	Waist circumfer	Body size meas	2.70E-23	1.64E-21	-0.00940011	0.000986795	3.33E-17	44	393454		0	SAIGE_0.44.5	FALSE	continuous	WHRadjBMI_cu	both_sexes		custom	486148	499256	263770	222378		ENSG00000166819				
PDE3B	pLoF	Hip circumferen	Hip circumferen	Body size meas	3.60E-22	5.14E-23	0.0179497	0.00181699	1.70E-18	51	212928		0.0468852855	SAIGE_0.44.5	FALSE	continuous	HCadjBMI_fem	females		custom	263702	271866	263563	139		ENSG00000152270				
PDE3B	pLoF	Hip circumferen	Hip circumferen	Body size meas	5.82E-22	8.31E-23	0.0116874	0.00118888	2.00E-19	69	393490		0	SAIGE_0.44.5	FALSE	continuous	HCadjBMI_cust	both_sexes		custom	486191	499300	263791	222400		ENSG00000152270				
ANKRD12	pLoF	Touchscreen_	duration_custom	Custom phenot	9.99E-22	1.65E-21	173.821	18.2494	5.44E-07	124	394476		0.0110974293	SAIGE_0.44.5	FALSE	continuous	Touchscreen_d	both_sexes		custom	487020	500891	264180	222840		ENSG00000101745				
ABCA1	pLoF	Cholesterol	Measured by Cl	Blood biochemi	1.38E-21	1.97E-22	-0.0184954	0.00189826	2.42E-05	110	376808		0	SAIGE_0.44.5	FALSE	continuous	30690	both_sexes		imt	465341	469615	252205	213136		ENSG00000165029				
ANKRD12	pLoF	Total protein	Measured by bi	Blood biochemi	2.52E-21	1.68E-21	-0.0225832	0.00237134	6.67E-06	115	344688		0	SAIGE_0.44.5	FALSE	continuous	30860	both_sexes		imt	425665	429627	229023	196642		ENSG00000101745				
GIGYF1	pLoF	Glycated haem	Measured by Hl	Blood biochemi	3.51E-20	2.19E-20	0.0274693	0.00296886	4.70E-10	85	376931		0	SAIGE_0.44.5	FALSE	continuous	30750	both_sexes		imt	463892	466530	251549	212343		ENSG00000146830				
PDE3B	pLoF	Waist-to-hip rati	Waist circumfer	Body size meas	1.34E-19	1.91E-20	-0.00815984	0.000880517	1.05E-16	69	393454		0	SAIGE_0.44.5	FALSE	continuous	WHRadjBMI_cu	both_sexes		custom	486148	499256	263770	222378		ENSG00000152270				
PLIN4	pLoF	Waist-to-hip rati	Waist circumfer	Body size meas	6.63E-19	1.50E-18	0.00795223	0.000904735	6.34E-15	76	212912		0.0400681377	SAIGE_0.44.5	FALSE	continuous	WHRadjBMI_fe	females		custom	263681	271844	263542	139		ENSG00000167676				
PLIN1	pLoF	Waist-to-hip rati	Waist circumfer	Body size meas	7.97E-19	3.59E-16	-0.0160544	0.00196947	1.06E-15	37	212912		0.0400681377	SAIGE_0.44.5	FALSE	continuous	WHRadjBMI_fe	females		custom	263681	271844	263542	139		ENSG00000166819				
ANKRD12	pLoF	Neutrophill perc	Result of "Neut	Blood count	4.42E-18	9.44E-18	0.0192889	0.00224799	3.97E-07	124	382770		0	SAIGE_0.44.5	FALSE	continuous	30200	both_sexes		imt	472490	477317	255661	216829		ENSG00000101	-0.02897	0.1168	-0.00878	0.6389
SLTM	pLoF	IOP_Goldmann_	left_custom	Custom phenot	6.77E-18	0.202592	0.0476452	0.0373922	9.67E-19	16	83193		0.0391377144	SAIGE_0.44.5	FALSE	continuous	IOP_Goldmann	both_sexes		custom	108123	111928	58239	49884		ENSG00000137776				
STAB1	pLoF	Median T2star (	Median T2star (	Susceptibility w	1.31E-17	4.07E-18	0.0413239	0.0047626	2.95E-05	39	16739		0.0795067858	SAIGE_0.44.5	FALSE	continuous	25030	both_sexes		imt	19176	19664	10091	9085		ENSG00000010327				
ANKRD12	pLoF	Lymphocyte pe	Result of "Lymp	Blood count	2.08E-17	3.29E-17	-0.0187087	0.00221773	1.09E-06	124	382770		0	SAIGE_0.44.5	FALSE	continuous	30180	both_sexes		imt	472490	477317	255661	216829		ENSG00000101	0.03823	0.02894	0.01405	0.4235
PLIN1	pLoF	Ratio of circulat	Triglycerides m	Blood biochemi	4.31E-17	1.28E-15	-0.0113163	0.00141515	1.50E-13	43	343470		0	SAIGE_0.44.5	FALSE	continuous	ratio_TG_HDL	both_sexes		custom	423734	427677	228091	195643		ENSG00000166819				
FBN2	pLoF	Body fat percer	Body compositi	Impedance mea	6.06E-17	1.35E-15	0.0180735	0.00226197	2.40E-05	79	387669		0	SAIGE_0.44.5	FALSE	continuous	BFPadjBMI_cu	both_sexes		custom	478991	491844	260272	218719		ENSG00000138829				
ANKRD12	pLoF	Touchscreen	duration	Process duratio	6.98E-17	3.77E-17	0.0173952	0.00206599	2.98E-05	124	394656		0	SAIGE_0.44.5	FALSE	continuous	630	both_sexes		imt	487776	501683	264538	223238		ENSG00000101745				
PDE3B	pLoF	Hip circumferen	Hip circumferen	Body size meas	2.93E-16	1.15E-16	0.00995448	0.00120098	1.88E-15	69	394124		0	SAIGE_0.44.5	FALSE	continuous	49	both_sexes		imt	487159	500317	264208	222951		ENSG00000152	0.8454	0.00E+00	0.8093	0.00E+00
PDE3B	pLoF	Hip circumferen	Hip circumferen	Body size meas	3.77E-16	1.53E-16	0.00991234	0.00120093	2.22E-15	69	394124		0	SAIGE_0.44.5	FALSE	continuous	hip_circ_custom	both_sexes		custom	487134	500285	264193	222941		ENSG00000152270				
STAB1	pLoF	Median T2star (	Median T2star (	Susceptibility w	4.37E-16	1.84E-16	0.039734	0.00482655	2.34E-05	39	16739		0.0483809409	SAIGE_0.44.5	FALSE	continuous	25028	both_sexes		imt	19176	19664	10091	9085		ENSG00000010327				
ANKRD12	pLoF	Townsend depr	Townsend depri	Baseline charac	6.35E-16	4.35E-16	0.0181871	0.00223746	1.75E-05	124	394375		0	SAIGE_0.44.5	FALSE	continuous	189	both_sexes		imt	487689	501913	264497	223192		ENSG00000010	0.3532	1.37E-46	0.336	1.01E-43
BLTP1/KIAA11(	pLoF	Weight	Weight was me	Body size meas	6.37E-16	9.13E-14	0.0104598	0.00140346	7.24E-05	266	393709		0	SAIGE_0.44.5	FALSE	continuous	21002	both_sexes		imt	486628	499762	263954	222674		ENSG00000131	0.8188	0.00E+00	0.7541	0.00E+00
BLTP1/KIAA11(	pLoF	Weight	Weight was me	Body size meas	6.62E-16	9.50E-14	0.0104526	0.00140346	7.22E-05	266	393709		0	SAIGE_0.44.5	FALSE	continuous	weight_custom	both_sexes		custom	486603	499730	263939	222664		ENSG00000138688				
BLTP1/KIAA11(	pLoF	Weight	Weight of perso	Impedance mea	8.67E-16	1.16E-13	0.010465	0.00141014	9.89E-05	265	387978		0	SAIGE_0.44.5	FALSE	continuous	23098	both_sexes		imt	479533	492407	260352	219181		ENSG00000131	0.8184	0.00E+00	0.7542	0.00E+00
GIGYF1	pLoF	Cholesterol	Measured by Cl	Blood biochemi	1.41E-15	2.57E-16	-0.0253395	0.0030933	1.98E-06	82	376808		0	SAIGE_0.44.5	FALSE	continuous	30690	both_sexes		imt	465341	469615	252205	213136		ENSG00000146830				
STAB1	pLoF	Median T2star (	Median T2star (	Susceptibility w	1.50E-14	3.37E-15	0.0380616	0.00483231	0.00016544	39	16739		0.0697781376	SAIGE_0.44.5	FALSE	continuous	25029	both_sexes		imt	19176	19664	10091	9085		ENSG00000010327				
PLIN1	pLoF	Apolipoprotein /	Measured by im	Blood biochemi	2.08E-14	2.06E-13	0.00966664	0.00131612	1.86E-11	43	343018		0	SAIGE_0.44.5	FALSE	continuous	30630	both_sexes		imt	423583	427532	227111	196472		ENSG00000166819				
PDE3B	pLoF	Waist-to-hip rati	Waist circumfer	Body size meas	2.21E-14	3.16E-15	-0.0128198	0.00162595	1.15E-09	51	213209		0	SAIGE_0.44.5	FALSE	continuous	WHR_female_c	females		custom	264079	272256	263939	140		ENSG00000152270				
PDE3B	pLoF	Hip circumferen	Hip circumferen	Body size meas	6.56E-14	2.61E-13	0.0133043	0.00181917	3.16E-14	51	213227		0.0394668906	SAIGE_0.44.5	FALSE	continuous	hip_circ_female	females		custom	264102	272280	263962	140		ENSG00000152270				
GIGYF1	pLoF	Glucose	Measured by he	Blood biochemi	8.52E-14	1.22E-14	0.0253388	0.0032847	0.000158355	80	344692		0	SAIGE_0.44.5	FALSE	continuous	30740	both_sexes		imt	425625	429590	228982	196643		ENSG00000146830				
BLTP1/KIAA11(	pLoF	Basal metabolic	Body compositi	Impedance mea	1.02E-13	1.35E-12	0.00736827	0.00103943	0.00108761	265	387962		0	SAIGE_0.44.5	FALSE	continuous	23105	both_sexes		imt	479514	492388	260343	219171		ENSG00000131	0.6274	0.00E+00	0.4643	1.45E-142
ANKRD12	pLoF	Fluid intelligenc	This is a simple	Fluid intelligenc	1.15E-13	5.28E-14	-0.0285407	0.00379288	4.99E-05	53	128302		0.0352956276	SAIGE_0.44.5	FALSE	continuous	20016	both_sexes		imt	160451	165471	87009	73442		ENSG00000010	-0.199	1.26E-23	-0.2046	5.64E-24
ANKRD12	pLoF	Mean corpuscu	Result of "Mear	Blood count	1.52E-13	3.93E-12	-0.0149056	0.00214791	2.52E-07	124	383466		0	SAIGE_0.44.5	FALSE	continuous	30050	both_sexes		imt	473359	478197	256119	217240		ENSG00000010	-0.09462	1.69E-08	-0.09693	1.55E-09
GIGYF1	pLoF	Hand grip stren	Left grip strengt	Hand grip stren	1.73E-13	2.77E-14	-0.017061	0.00224236	1.23E-05	83	393208		0	SAIGE_0.44.5	FALSE	continuous	46	both_sexes		imt	486004	499126	263455	222549		ENSG00000144	-0.0755	0.0001071	-0.1771	3.28E-20
GIGYF1	pLoF	Hand grip stren	Hand grip stren	Custom phenot	2.41E-13	8.09E-14	-0.0226369	0.00303084	6.96E-07	82	392236		0	SAIGE_0.44.5	FALSE	continuous	HGSadjHtWt_c	both_sexes		custom	484264	497276	262661	221603		ENSG00000146830				
PDE3B	pLoF	Apolipoprotein /	Measured by im	Blood biochemi	3.91E-13	8.97E-14	0.00882736	0.00118404	1.61E-12	65	343018		0	SAIGE_0.44.5	FALSE	continuous	30630	both_sexes		imt	423583	427532	227111	196472		ENSG00000152270				
PLIN4	pLoF	Waist-to-hip rati	Waist circumfer	Body size meas	4.54E-13	2.18E-13	0.00336867	0.000459113	7.83E-11	98	393454		0	SAIGE_0.44.5	FALSE	continuous	WHRadjBMI_cu	both_sexes		custom	486148	499256	263770	222378		ENSG00000167676				
SLTM	pLoF	Corneal_hysteresis_left_custom		Custom phenot	5.54E-13	0.017328	0.0551043	0.0231562	7.92E-14	16	83193		0.0260622074	SAIGE_0.44.5	FALSE	continuous	Corneal_hyster	both_sexes		custom	108123	111928	58239	49884		ENSG00000137776				
FBN2	pLoF	Fat free mass ir	Fat free mass ir	Impedance mea	6.11E-13	1.96E-12	-0.0156188	0.00221933	0.000263679	79	387569		0	SAIGE_0.44.5	FALSE	continuous	FFMadjBMI_cu	both_sexes		custom	478858	491710	260069	218789		ENSG00000138829				
PLIN1	pLoF	Waist-to-hip rati	Waist circumfer	Body size meas	6.27E-13	8.32E-13	-0.00736729	0.00102955	1.86E-09	44	394086		0	SAIGE_0.44.5	FALSE	continuous	WHR_custom	both_sexes		custom	487089	500239	264170	222919		ENSG00000166819				
ANKRD12	pLoF	Mean corpuscu	Result of "Mear	Blood count	6.42E-13	1.77E-10	-0.0137583	0.00215655	5.95E-08	124	383467		0	SAIGE_0.44.5	FALSE	continuous	30040	both_sexes		imt	473360	478198	256120	217240		ENSG00000010	-0.09592	1.44E-08	-0.09241	2.34E-08
PDE3B	pLoF	Comparative he	ACE touchscree	Early life factor	6.47E-13	9.24E-14	0.00610983	0.000819964	6.56E-10	70	388487		0	SAIGE_0.44.5	FALSE	continuous	1697	both_sexes			478606	492134	259250	219356		ENSG00000152	0.005015	0.796	0.07846	8.27E-06
PLIN1	pLoF	Triglycerides	Measured by G	Blood biochemi	8.44E-13	1.71E-12	-0.0093764	0.00132881	1.19E-09	43	376509		0	SAIGE_0.44.5	FALSE	continuous	30870	both_sexes		imt	464971	469240	252065	212906		ENSG00000166819				
STAB1	pLoF	Albumin	Measured by Bl	Blood biochemi	8.71E-13	8.70E-13	0.00681824	0.000953641	1.62E-05	211	345108		0	SAIGE_0.44.5	FALSE	continuous	30600	both_sexes		imt	426128	430098	229239	196889		ENSG00000010327				
PDE3B	pLoF	Sitting height	Distance from n	Body size meas	1.14E-12	1.62E-13	0.00658506	0.000892687	1.73E-08	69	393937		0	SAIGE_0.44.5	FALSE	continuous	20015	both_sexes		imt	486411	496254	263915	222496		ENSG00000152	-0.05848	0.0001836	-0.03566	0.02759
BLTP1/KIAA11(	pLoF	Arm fat mass (r	Body compositi	Impedance mea	1.36E-12	2.00E-11	0.0105738	0.00157684	0.000476787	265	387864		0	SAIGE_0.44.5	FALSE	continuous	23120	both_sexes		imt	479389	492260	260268	219121		ENSG00000131	0.9483	0.00E+00	0.9229	0.00E+00
BLTP1/KIAA11(	pLoF	Arm fat-free ma	Body compositi	Impedance mea	1.39E-12	1.03E-11	0.00685336	0.00100746	0.00215086	265	387788		0	SAIGE_0.44.5	FALSE	continuous	23125	both_sexes		imt	479297	492167	260229	219068		ENSG00000131	0.6115	4.78E-289	0.4206	9.73E-107
BLTP1/KIAA11(	pLoF	Whole body wa	Body compositi	Impedance mea	1.96E-12	1.13E-11	0.00674021	0.000992773	0.00361294	265	387974		0	SAIGE_0.44.5	FALSE	continuous	23102	both_sexes		imt	479528	492402	260350	219178		ENSG00000131	0.5461	1.74E-200	0.3491	3.23E-68
STAB1	pLoF	Total protein	Measured by bi	Blood biochemi	2.10E-12	5.05E-13	0.00696296	0.000963886	0.000770059	211	344688		0	SAIGE_0.44.5	FALSE	continuous	30860	both_sexes		imt	425665	429627	229023	196642		ENSG00000010327				
BLTP1/KIAA11(	pLoF	Leg fat mass (ri	Body compositi	Impedance mea	2.20E-12	2.05E-10	0.0081602	0.00128358	6.80E-05	265	387940		0	SAIGE_0.44.5	FALSE	continuous	23112	both_sexes		imt	479486	492359	260327	219159		ENSG00000131	0.9375	0.00E+00	0.9392	0.00E+00
GIGYF1	pLoF	Townsend depr	Townsend depri	Baseline charac	2.23E-12	3.18E-13	0.0228988	0.00314265	7.40E-05	85	394375		0	SAIGE_0.44.5	FALSE	continuous	189	both_sexes		imt	487689	501913	264497	223192		ENSG00000146	0.3532	1.37E-46	0.336	1.01E-43
BLTP1/KIAA11(	pLoF	Whole body fat	Body compositi	Impedance mea	2.97E-12	1.55E-11	0.00667904	0.000990468	0.00409377	265	387945		0	SAIGE_0.44.5	FALSE	continuous	23101	both_sexes		imt	479486	492359	260351	219135		ENSG00000131	0.5423	2.91E-196	0.3486	2.10E-68
BLTP1/KIAA11(	pLoF	Arm fat mass (l	Body compositi	Impedance mea	3.14E-12	3.35E-11	0.0104503	0.00157613	0.000679218	265	387798		0	SAIGE_0.44.5	FALSE	continuous	23124	both_sexes		imt	479312	492183	260232	219080		ENSG00000131	0.9481	0.00E+00	0.9222	0.00E+00
BLTP1/KIAA11(	pLoF	Whole body fat	Body compositi	Impedance mea	3.18E-12	1.65E-11	0.00666982	0.000990467	0.00404165	265	387945		0	SAIGE_0.44.5	FALSE	continuous	WBfatfreemass	both_sexes		custom	479461	492328	260336	219125		ENSG00000138688				
BLTP1/KIAA11(	pLoF	Arm predicted r	Body compositi	Impedance mea	3.49E-12	2.09E-11	0.00672235	0.00100341	0.00276432	265	387773		0	SAIGE_0.44.5	FALSE	continuous	23126	both_sexes		imt	479279	492148	260221	219058		ENSG00000131	0.612	1.93E-290	0.4219	1.45E-107

Table S4.6: Sex-specific and sex-differential gene-level results. These are genes that only reach significance in sex-specific strata and genes with significant heterogeneity in effect sizes between sexes (phenotype, male-specific β & SE & P, female-specific β & SE & P, sex-heterogeneity P-value)

pheno	gene_symbol	gene_id	Group	Pvalue_Burden_sexdiff	BETA_Burden_sexdiff	Pvalue_SKATO_f	Pvalue_SKATO_m	Pvalue_Burden_f	Pvalue_Burden_m	Pvalue_SKAT_f	Pvalue_SKAT_m	BETA_Burden_f	BETA_Burden_m	SE_Burden_f	SE_Burden_m	MAC_f	MAC_m
bmi	BLTP1	ENSG00000138688	pLoF	0.9162425777	0.0003440534838	1.08E-06	2.17E-05	8.10E-07	1.97E-05	1.23E-06	7.74E-05	0.01093647081	0.01059241733	0.002217012307	0.002481555641	333	267
bmi	PCSK1	ENSG00000175426	pLoF;damaging	0.004024509481	-0.006366190717	0.007967923239	1.97E-08	0.2291827159	5.29E-07	0.004712884641	5.89E-07	0.001855917008	0.008222107725	0.00154342472	0.001639325068	675	606
bmi	DIDO1	ENSG00000101191	pLoF	0.03138924604	0.03388756027	3.29E-07	0.04627650829	3.29E-07	0.04627650829	3.29E-07	0.04627650829	0.05684137645	0.02295381618	0.01113195955	0.01151801314	13	12
bmi	SLC12A5	ENSG00000124140	pLoF	0.003320579989	0.07584209847	2.67E-08	0.7854123965	2.67E-08	0.7854123965	2.67E-08	0.7854123965	0.06957362507	-0.006268473398	0.01250850351	0.02302278163	10	3
whr_adj_bmi	ACVR1C	ENSG00000123612	pLoF;damaging	0.3330307593	-0.0015700226	6.57E-09	8.17E-05	1.23E-08	7.27E-05	9.32E-08	0.0002998755922	-0.006340065986	-0.004770043386	0.001113234557	0.001202357354	1399	1137
whr_adj_bmi	PLXND1	ENSG00000004399	pLoF	3.73E-05	-0.02460243957	9.69E-10	0.3301265389	5.73E-08	0.9391702864	6.91E-09	0.2379804105	-0.02490065897	-0.0002982194063	0.004588156413	0.003907858462	73	83
whr_adj_bmi	PLD1	ENSG00000075651	pLoF;damaging	0.01851149166	0.002073794993	6.37E-09	0.05990697422	5.82E-08	0.03437935101	1.90E-06	0.5042132037	0.003406722762	0.001332927769	0.0006280399027	0.0006300489402	5019	4306
whr_adj_bmi	PPARG	ENSG00000132170	pLoF;damaging	0.002078232242	0.01842975111	1.11E-06	0.9391228749	1.33E-06	0.9391228749	2.10E-06	0.9391228749	0.01878366937	0.0003539182577	0.00388451136	0.004634115681	104	71
whr_adj_bmi	MLXIPL	ENSG00000009950	pLoF	0.01379007416	0.01524244934	2.12E-06	0.4327402264	2.41E-06	0.3207047816	0.0003230675971	0.5364824077	0.01983676256	0.004594313217	0.004206446803	0.004626650661	88	71
whr_adj_bmi	GIGYF1	ENSG00000146830	pLoF	3.42E-05	-0.02808265544	0.8495007182	2.23E-10	0.7480916943	2.31E-10	1	4.98E-08	-0.001747153557	0.02633550189	0.005440191406	0.004154241128	53	88
whr_adj_bmi	PDE3B	ENSG00000152270	pLoF	8.02E-09	-0.01407496424	9.36E-23	0.07097836121	1.34E-23	0.1081966777	3.08E-20	0.05118200131	-0.01695353059	-0.002878566353	0.001693159906	0.001791990212	568	489
whr_adj_bmi	PDE3B	ENSG00000152270	pLoF;damaging	0.01417916221	-0.0030555312206	1.52E-14	0.1457138638	9.93E-08	0.08825983871	2.25E-15	0.315885124	-0.004612529352	-0.001557217146	0.0008657005879	0.0009135117039	2219	1947
whr_adj_bmi	PLIN1	ENSG00000166819	pLoF;damaging	0.002892650688	-0.004665872696	1.75E-20	1.47E-05	4.76E-19	1.10E-05	4.29E-16	0.0009377702324	-0.009713505178	-0.005047632482	0.00108924013	0.001147918566	1347	1179
whr_adj_bmi	PKD1	ENSG00000008710	pLoF	0.3981165073	-0.005731434464	0.000238298117	9.95E-07	0.00526407746	0.000108085037	0.000163438697	7.28E-07	0.01331370236	0.01904513683	0.004771229307	0.004919080671	90	86
whr_adj_bmi	ANKRD12	ENSG00000101745	pLoF	0.3103654387	-0.004500336978	0.0006089137618	1.18E-06	0.0005863233326	1.40E-06	0.001365285232	1.74E-05	0.01085175526	0.01535209224	0.003156540533	0.003181964737	158	150
whr_adj_bmi	KEAP1	ENSG00000079999	pLoF;damaging	0.005564276939	0.01334743864	1.70E-11	0.0443975679	4.80E-12	0.03350411241	1.99E-11	0.1085040763	0.02133617032	0.007988731683	0.003087121816	0.00375765295	165	108
whr_adj_bmi	COL5A3	ENSG00000080573	pLoF	0.02920650887	0.00763693029	1.09E-06	0.1261401166	1.09E-06	0.09088461175	0.0001424923196	0.1052413532	0.01195019376	0.004313263468	0.002451721281	0.002551095167	263	237
whr_adj_bmi	COL5A3	ENSG00000080573	pLoF;damaging	0.0006282647971	0.004838602881	8.81E-11	0.2138228699	5.58E-11	0.125753757	1.50E-07	0.4202115114	0.006430898077	0.001592295197	0.0009811345911	0.001039993075	1654	1432
whr_adj_bmi	PLIN4	ENSG00000167676	pLoF	8.19E-08	0.007142038833	4.78E-19	0.4013371763	2.74E-19	0.274650919	5.79E-15	0.3387413327	0.008219205591	0.001077166758	0.0009154262601	0.0009860420954	1991	1557
whr_adj_bmi	PLIN4	ENSG00000167676	pLoF;damaging	0.0087555854	0.002307314199	3.93E-09	0.1578060797	3.65E-08	0.09711505368	3.35E-07	0.176107217	0.003376002572	0.001068688372	0.0006130251262	0.0006441748107	5369	4337
whr_adj_bmi	INSR	ENSG00000171105	pLoF	5.99E-06	-0.02868301072	4.75E-12	0.898306035	8.93E-11	0.898306035	6.79E-13	0.898306035	-0.02926325004	-0.0005802393227	0.004513102505	0.004540159248	77	74
whr_adj_bmi	INSR	ENSG00000171105	pLoF;damaging	1.66E-06	-0.006534124678	6.61E-07	0.02966573191	1.21E-06	0.05169632197	3.14E-06	0.01964990931	-0.004580962087	0.001953162591	0.0009437632478	0.001003859187	1906	1642
body_fat_percentage	CASQ1	ENSG00000143318	pLoF	1.62E-05	0.008513639038	1.12E-06	0.09076499428	4.05E-06	0.1411862158	1.46E-06	0.07709412635	0.006348771895	-0.002164867143	0.001377589133	0.001471309497	873	764
body_fat_percentage	CASQ1	ENSG00000143318	pLoF;damaging	0.006143292705	0.004392359344	2.20E-06	0.141364963	0.001084211379	0.5318156252	1.87E-06	0.09055925758	0.003643307393	-0.0007490519514	0.001114941851	0.001198029133	1304	1118
body_fat_percentage	GPR151	ENSG00000173250	pLoF	0.2430869168	-0.001161376836	1.96E-06	0.009220867059	5.28E-06	0.008680884057	2.77E-06	0.009192276902	-0.003125967214	-0.001964590379	0.0006864981037	0.0007485943916	4477	3732
body_fat_percentage	GPR151	ENSG00000173250	pLoF;damaging	0.2097942641	-0.001242669634	1.97E-06	0.01083537376	5.64E-06	0.01269431485	2.78E-06	0.009080441838	-0.003101649163	-0.001858979528	0.0006832899995	0.0007459095505	4509	3753
body_fat_percentage	PLIN4	ENSG00000167676	pLoF	0.9665677178	-5.56E-05	1.32E-06	2.70E-05	2.15E-06	1.75E-05	2.26E-06	6.12E-05	-0.004340733491	-0.004285123764	0.0009159247413	0.0009978818214	1964	1530
tissuefatp_androidgynoidratio	TINF2	ENSG00000092330	pLoF;damaging	0.0002355226919	-0.0423425020	0.8414987126	4.84E-07	0.8414987126	4.84E-07	0.8414987126	4.84E-07	-0.001645129133	0.0406973729	0.008226599267	0.008086767346	24	25
tissuefatp_androidgynoidratio	ZNF841	ENSG00000197608	pLoF;damaging	1.41E-05	-0.03332042803	0.2689555456	1.04E-06	0.3748188261	6.19E-07	0.1909202832	1.07E-05	-0.00449546915	0.02882495888	0.005065413439	0.0057822912	64	49
visceral_adipose_tissue_volum	SENP5	ENSG00000119231	pLoF	2.33E-06	0.1889763726	0.001456640398	0.0005090883077	0.001456640398	0.0005090883077	0.001456640398	0.0005090883077	0.08930459507	-0.09967177753	0.0280550985	0.02867486552	2	2

Table S4.7: Obesity age-of-onset longitudinal analysis using Cox proportional hazards model.

Gene	Max alternate allele frequency	Sex	Consequence group	Number of individuals in group	Effect size	Standard error	P-value
SLTM	0.01	male_only	pLoF	32	5.372929194	0.3540849785	2.05E-06
MC4R	0.01	sex_combined	other_missense	1864	1.457897071	0.08430368913	7.75E-06
DIDO1	0.01	female_only	pLoF	10	11.23207718	0.5779158958	2.85E-05
SLTM	0.01	sex_combined	pLoF	64	3.212363668	0.3016838505	0.000109591752
SLC12A5	0.01	female_only	pLoF	8	14.76128355	0.7083700326	0.000144529955
PCSK1	0.01	male_only	damaging_missense	421	1.862206767	0.1649328618	0.000163385102
EI24	0.01	female_only	pLoF	28	4.453560126	0.4085762274	0.000256305768
GABRA5	0.01	female_only	pLoF	16	5.683840315	0.5004393106	0.000516204913
RIF1	0.01	male_only	pLoF	33	3.726880486	0.4086621537	0.001285418280
MC4R	0.01	female_only	other_missense	1051	1.42892169	0.112318967	0.001484295083
UBR3	0.01	sex_combined	pLoF	76	2.73116892	0.3163883924	0.001495160370
EI24	0.01	sex_combined	pLoF	45	3.320392448	0.3781474707	0.001505690230
MC4R	0.01	male_only	other_missense	813	1.493264245	0.1276105682	0.001677437617
GABRA5	0.01	sex_combined	pLoF	30	3.963230329	0.4473853008	0.002083832288
SLC12A5	0.01	sex_combined	pLoF	11	8.577467752	0.7074707973	0.002383300054
DIDO1	0.01	sex_combined	pLoF	19	5.7593419	0.5776306013	0.002437105039
UBR3	0.01	female_only	pLoF	43	3.138535397	0.3782246946	0.002494460172
PLCB4	0.01	male_only	synonymous	1211	1.387526207	0.1105130153	0.003040088760
AFF3	0.01	female_only	pLoF	15	4.180003468	0.5002504906	0.004247219683
ATP13A1	0.01	male_only	damaging_missense	1662	1.317058649	0.09802943775	0.004963857327
DIDO1	0.01	female_only	other_missense	12223	1.110505533	0.03845081176	0.006411490817
APOM	0.01	male_only	synonymous	383	0.3570436564	0.3781886416	0.006464570740
APBA1	0.01	female_only	pLoF	59	2.778201998	0.3782250685	0.006901160335
ATP13A1	0.01	male_only	pLoF	37	3.266670173	0.4475991974	0.008176036334
BIN1	0.01	female_only	synonymous	2223	0.7605932036	0.1037625825	0.008356043525
GIGYF1	0.01	sex_combined	pLoF	114	2.11336113	0.2888703576	0.00958742478
APBA1	0.01	sex_combined	damaging_missense	437	0.4608411297	0.3016404365	0.01021998562
AFF3	0.01	sex_combined	pLoF	30	3.146698827	0.4473395646	0.01038906321
KEAP1	0.01	female_only	pLoF	23	3.566934785	0.5003576074	0.01103478088
PLCB4	0.01	sex_combined	synonymous	2693	1.214562812	0.07759553738	0.01224185357
APOM	0.01	sex_combined	synonymous	902	0.6376050242	0.1797996597	0.012314995
PCSK1	0.01	sex_combined	damaging_missense	910	1.372295313	0.1272866146	0.01290447949
ABCA1	0.01	female_only	damaging_missense	1504	1.276650076	0.09979522149	0.01438882729
PCSK1	0.01	sex_combined	pLoF	92	2.239573438	0.3334961589	0.01562000913
KIF1B	0.01	female_only	synonymous	5021	1.148391065	0.05766675325	0.01642485727
KIF1B	0.01	sex_combined	synonymous	9160	1.110104979	0.04423643779	0.01821200651
HECTD4	0.01	sex_combined	non_coding	49087	1.050406924	0.02145179825	0.02187845002
RHOJ	0.01	male_only	non_coding	3767	0.8334561674	0.07952016194	0.0219681593
HPDL	0.01	female_only	other_missense	1569	1.249447555	0.09775577737	0.02271814563
APBA1	0.01	male_only	damaging_missense	206	0.2704226527	0.5775308846	0.02354864851
HOXC9	0.01	female_only	synonymous	1476	1.257026284	0.1010866294	0.02364191213
MC4R	0.01	sex_combined	pLoF	141	1.86814149	0.2775416043	0.02434040481
HERC1	0.01	female_only	synonymous	8725	1.106743763	0.04516241007	0.02472190643
RPRD2	0.01	sex_combined	damaging_missense	2364	0.7996158964	0.09985612732	0.02512600558
ATP13A1	0.01	sex_combined	pLoF	105	1.964308832	0.3016900468	0.02523010929
CPM	0.01	sex_combined	non_coding	6703	0.8818237009	0.05655734142	0.02617272044
EIF4A1	0.01	male_only	synonymous	468	1.450289729	0.1672341073	0.02621502227
PLD1	0.01	female_only	damaging_missense	3797	0.8454162102	0.07598775064	0.02711129362
KEAP1	0.01	sex_combined	pLoF	41	2.684775509	0.4473908073	0.02728196919
ZNF512B	0.01	sex_combined	non_coding	29535	0.9424139366	0.02710337207	0.02864676681
HOXC9	0.01	sex_combined	synonymous	2685	1.184905924	0.07757676955	0.02873971284
RPRD2	0.01	female_only	damaging_missense	1304	0.7419830073	0.1365460543	0.02884816415
PLIN4	0.01	sex_combined	non_coding	40260	0.9492622914	0.0239218681	0.02950482747
HECTD4	0.01	male_only	synonymous	9453	1.105768786	0.04635864795	0.0301008342
GIGYF1	0.01	female_only	synonymous	2719	1.183219484	0.07781976288	0.03062533636
MYCBP2	0.01	female_only	other_missense	9055	1.101342805	0.04471226494	0.03085643621
AFF3	0.01	male_only	synonymous	5170	0.8631731947	0.06858666615	0.03192776879
GPR151	0.01	sex_combined	synonymous	402	0.5091553485	0.3163385721	0.03285945092
INHBE	0.01	sex_combined	damaging_missense	94	2.025626945	0.3334726121	0.03428078982
FXYD5	0.01	female_only	pLoF	18	3.389704377	0.5776394961	0.0345725569
DENND5B	0.01	male_only	synonymous	4157	0.8526272039	0.07547468305	0.03465198736
RIF1	0.01	sex_combined	pLoF	72	2.108274979	0.3536994673	0.03496462444
SLC12A5	0.01	sex_combined	non_coding	63444	0.9600252936	0.01952866096	0.03670645735
FXYD5	0.01	sex_combined	pLoF	26	2.824777021	0.5001664112	0.03787842808
ANKRD12	0.01	female_only	synonymous	7708	1.103116446	0.04763647656	0.03938215791
MYCBP2	0.01	sex_combined	other_missense	16465	1.072763813	0.03415021834	0.03971067642
DIDO1	0.01	sex_combined	other_missense	22173	1.062407846	0.02958727072	0.04074862583
APBA1	0.01	sex_combined	pLoF	106	1.959839962	0.3334903029	0.04362894182
GIPR	0.01	female_only	non_coding	8018	0.9031465648	0.05055146715	0.04388553447
SLC12A5	0.01	female_only	non_coding	34795	0.9494477101	0.02591425811	0.04530766032
ZNF512B	0.01	female_only	non_coding	16284	0.93052491	0.03601089677	0.04554634441
PCSK1	0.01	male_only	pLoF	35	2.711198074	0.5002877792	0.04619208773
FXYD5	0.01	female_only	other_missense	597	1.354816775	0.1529087979	0.04704105707
PLIN4	0.01	female_only	pLoF	1499	0.7817493423	0.1246732642	0.04827595194
INHBE	0.01	male_only	pLoF	332	1.458598992	0.1929932355	0.05047637901

Table S4.8: Combined evidence. Includes gene-level association results, hWAT mRNA transcript counts, evidence of allelic series, longitudinal analysis, and gene perturbation evidence from literature.

Gene symbol	Ensembl ID	SKAT-O p-value (conditioned)	Significance	hWAT D8 mean mRNA count	hWAT D24 mean mRNA count	hWAT mRNA count >10 on days 8 and 24	Monotonic ultra-rare variant allelic series	Obesity age-of-onset association	Selected for knockdown	Fat formation evidence from literature	Glucose uptake relative to wild-type	Insulin-stimulated glucose uptake relative to wild-type	Reference links
ABCA1	ENSG00000165029	6.50E-07	bh_sig_fdr_0.01_after_cond	169.3194378	945.621995	yes	yes	no	yes	Increased adipose triglyceride, increased body weight on high-fat, high-cholesterol diet (mouse adipocyte-specific KO – de Haan et al., 2014) "significantly lower body weight, epididymal fat pad weight, and adipocyte size" (mouse adipocyte-specific KO – Cuffe et al., 2018)	No change (mouse adipocyte-specific KO – Cuffe et al., 2018)	No change (mouse adipocyte-specific KO – Cuffe et al., 2018)	de Haan et al., 2014 https://doi.org/10.1194/jlr.m045294 Cuffe et al., 2018 https://doi.org/10.1161%2FATVBAHA.117.309880
COL5A3	ENSG00000080573	5.23E-10	exome_wide	3692.857194	6020.588685	yes	yes	no	yes	Reduced dermal fat in females but not males (mouse KO – Huang et al., 2011)	Decreased ("defective") in WAT (mouse KO – Huang et al., 2011)	Insulin resistance conferred by KO (mouse KO – Huang et al., 2011)	Huang et al., 2011 https://doi.org/10.1172%2FJCI45096
DENND5B	ENSG00000170456	2.79E-05	bh_sig_fdr_0.01_after_cond	798.0044751	867.2555085	yes	yes	no	yes	Lower fat mass, decreased rate of weight gain, higher lipid accumulation in intestinal cells (mouse KO – Gordon et al., 2019) Decreased lipid accumulation in liver (mouse KO – Mobilia et al., 2022)			Gordon et al., 2019 https://doi.org/10.1038/s41598-019-40296-0 Mobilia et al., 2022 https://doi.org/10.1016/j.jlr.2022.100296
EXOC7	ENSG00000182473	2.06E-05	bh_sig_fdr_0.01_after_cond	6197.674884	6264.105104	yes	no	no	yes			Inhibited GLUT4 exocytosis (mouse adipocyte KO – Wang et al., 2019)	Wang et al., 2019 https://doi.org/10.1074%2Fjbc.RA119.010821
HERC1	ENSG00000103657	3.46E-05	bh_sig_fdr_0.01_after_cond	2177.203646	3138.381076	yes	no	no	yes	Higher body fat percentage and body fat weight (mouse KO – Powell et al., 2021)			Powell et al., 2021 https://doi.org/10.2147%2FDMSO.S322083
INSR	ENSG00000171105	6.38E-08	exome_wide	1297.281289	1990.281378	yes	yes	no	yes	Reduced fat mass (mouse fat-specific KO – Blüher et al., 2002) "striking delipidation" (mouse 3T3L1 adipocytes – Groeneveld et al., 2016) Decreased subcutaneous WAT (mouse adipose-specific KO – Sakaguchi et al., 2017) Reduced number of oil droplets measured by oil-red O staining, reduced ability to differentiate into adipocytes (mouse iPSC-derived adipocytes - Gupta et al., 2018)	Unchanged (mouse fat-specific KO adipocytes – Blüher et al., 2002)	Reduced (mouse fat-specific KO adipocytes – Blüher et al., 2002)	Blüher et al., 2002 https://doi.org/10.1016/s1534-5807(02)00199-5 Groeneveld et al., 2016 https://doi.org/10.1038%2Fsrep21105 Sakaguchi et al., 2017 https://doi.org/10.1016/j.cmet.2016.12.008 Gupta et al., 2018 https://doi.org/10.1016%2Fj.molmet.2018.09.003
MFAP5	ENSG00000197614	7.42E-06	bh_sig_fdr_0.01_after_cond	54237.5874	44426.02507	yes	no	no	yes			No significant difference (human SGBS adipocyte knockdown – Vaittinen et al., 2015)	Vaittinen et al., 2015 https://doi.org/10.1002/oby.21103
MLXIPL	ENSG00000009950	1.17E-05	bh_sig_fdr_0.01_after_cond	885.8208737	2870.313861	yes	yes	no	yes	Reduced adipose tissue and decreased lipogenesis in liver (mouse KO – Iizuka et al., 2004) "DNL normalizes in adipose tissue after deletion of ChREBP in AG4OX mice", "Glucose incorporation into newly synthesized fatty acids is also decreased in wild-type and AG4OX adipose tissue after ChREBP knockout" (mouse wild-type and Glut4 over-expressed ChREBP KO – Herman et al., 2012)	Reduced glucose transport over insulin dose range (mouse adipose-specific KO – Vijayakumar et al., 2017)	Unaffected?: "This suggests that the lower glucose transport in [adipose-specific] KO adipocytes does not result from impaired insulin signaling" (mouse adipose-specific KO – Vijayakumar et al., 2017)	Iizuka et al., 2004 https://doi.org/10.1073/pnas.0401516101 Herman et al., 2012 https://doi.org/10.1038/nature10986 Vijayajumar et al., 2017 https://doi.org/10.1016/j.celrep.2017.09.091
PCSK1	ENSG00000175426	1.99E-05	bh_sig_fdr_0.01_after_cond	360.4219153	328.8183345	yes	yes	yes	yes	Increased adipocyte hypertrophy (mouse PC1 N222D mutation – Lloyd et al., 2006)			Lloyd et al., 2006 https://doi.org/10.1093/hmg/ddl111
PLIN1	ENSG00000166819	8.94E-23	exome_wide	15819.68088	90124.62444	yes	yes	no	yes	Less lipid content in whole body (mouse KO – Martinez-Botas et al., 2000) Decreased adipose tissue mass (mouse KO – Tansey et al., 2001) Decreased adipocyte size (mouse KO – Lyu et al., 2015) Increased ectopic fat, possibly due to less fat storage in adipose tissue (mouse KO – Liu et al., 2015) When fed normal chow, no difference in body weight, increased inguinal and epididymal WAT, not increased liver or BAT, smaller lipid droplets (mouse KO – Sohn et al., 2018)	Glucose intolerance (mouse KO – Tansey et al., 2001) Higher fasting glucose implies lower glucose uptake (mouse KO – Sohn et al., 2018)	Peripheral insullin resistance (mouse KO – Tansey et al., 2001) Less sensitive to insulin (mouse KO – Sohn et al., 2018)	Martinez-Botas et al., 2000 https://doi.org/10.1038/82630 Tansey et al., 2001 https://doi.org/10.1073/pnas.101042998 Lyu et al., 2015 https://doi.org/10.1371/journal.pone.0117536 Liu et al., 2015 https://doi.org/10.1093/cvr/cvu214 Sohn et al., 2018 https://doi.org/10.1074%2Fjbc.RA118.003541
PPARG	ENSG00000132170	2.90E-08	exome_wide	1814.93178	4126.392825	yes	no	no	yes	Required for differentiation of adipose tissue (chimeric mouse embryonic stem cells – Rosen et al., 1999) Decreased fat formation (mouse adipose-specific KO – Jones et al., 2005) Decreased lipid accumulation (human adipocyte CRISPR KO stained with BODIPY – Kamble et al., 2020) Decreased lipid accumulation (human adipocyte KO stained with BODIPY – Laber et al., 2023)		Insulin resistance in adipocytes indicated by higher plasma free fatty acids (mouse KO – He et al., 2003) Reduced (mouse 3T3L1 adipocytes – Liao et al., 2007)	Rosen et al., 1999 https://doi.org/10.1016/s1097-2765(00)80211-7 He et al., 2003 https://doi.org/10.1073/pnas.2536828100 Jones et al., 2005 https://doi.org/10.1073/pnas.0306743102 Liao et al., 2007 https://doi.org/10.1152/ajpendo.00695.2006 Kamble et al., 2020 https://doi.org/10.1038/s41598-020-67293-y Laber et al., 2023 https://doi.org/10.1016/j.xgen.2023.100346
SLTM	ENSG00000137776	2.65E-05	bh_sig_fdr_0.01_after_cond	2754.30271	2734.091609	yes	no	yes	yes				
TRIP10	ENSG00000125733	1.48E-05	bh_sig_fdr_0.01_after_cond	3929.906497	2922.478267	yes	no	no	yes			"Overexpression blocks insulin-stim glucose transport" (mouse 3T3L1 adipocytes mutated – Chang, Adams, Saltiel, 2002) Increased (rat myoblast knockdown – Hartig et al., 2009) Increased (mouse KO adipocytes – Feng et al., 2010)	Chang, Adams, Saltiel 2002 https://doi.org/10.1073/pnas.202495599 Hartig et al., 2009 https://doi.org/10.1242/jcs.041343 Feng et al., 2010 https://doi.org/10.1074/jbc.m109.041038
UBR2	ENSG00000024048	1.27E-05	bh_sig_fdr_0.01_after_cond	1413.611345	1820.618191	yes	no	no	yes	"reduced the size and number of lipid droplets" in liver cancer cell line (HepG2 knockdown – Xu et al., 2023)			Xu et al., 2023 https://doi.org/10.2337/db22-0362

Table S4.9: Guide RNAs per gene target.

Gene	Sequence	Target site	Guide score	Strand
ABCA1	TGGCTTGTTGGCCTCAGCTG	chr9:107665940-107665959	1.292	-
ABCA1	GAGTCGGGTAACGGAAACAG	chr9:107646751-107646770	2.216	+
ABCA1	CCAGAAAGACACCAGCATGA	chr9:107645374-107645393	1.298	-
ABCA1	GAAGCTTCAAGATTTCCTGG	chr9:107624058-107624077	3.024	-
COL5A3	AAGGCCCTGGGTGTGCAGGG	chr19:10116871-10116890	2.47	-
COL5A3	TAGCTGGACTCACTGTGCTG	chr19:10116267-10116286	1.452	-
COL5A3	AAGAAAGGGCGAGGTCGCAA	chr19:10114311-10114330	1.961	-
COL5A3	TCACCTAGGAGACCCAGCGC	chr19:10116517-10116536	2.049	+
DENND5B	GATCGACGCGGACAGCGGGC	chr12:31743661-31743680	1.511	-
DENND5B	GTCAAGGCAAATATGAGCAG	chr12:31595765-31595784	2.702	-
DENND5B	GGAGCTCAAATGCCTCCCGA	chr12:31632595-31632614	3.414	+
DENND5B	GTGTCAAAGACCCGAAGCAA	chr12:31600561-31600580	1.451	+
EXOC7	AGCAGACGGAGAATCTGCAG	chr17:74097858-74097877	1.881	-
EXOC7	AAGCATGGCCAAGATTCAGA	chr17:74097408-74097427	1.032	-
EXOC7	TGAGCGCGGGAAGGAGGCCC	chr17:74094069-74094088	1.485	-
EXOC7	AGGAGGCATCCGCTCGACGG	chr17:74099741-74099760	2.321	-
HERC1	AGTGCATCAGTGCTTGACCG	chr15:64067353-64067372	2.751	+
HERC1	GAACCAACCAGAACATCGGA	chr15:64056341-64056360	1.698	-
HERC1	AGAAGGAGAAGTCTTCAGTT	chr15:64048725-64048744	1.251	-
HERC1	AATGTAGGAGAGGTTTCTTG	chr15:64046775-64046794	1.041	-
INSR	AACCTGATGAACATCACCCG	chr19:7267586-7267605	3.45	-
INSR	ACAAATGCAAGAACTCGCGG	chr19:7184410-7184429	3.852	-
INSR	GAGGTGGCACACCTTGGGAC	chr19:7174694-7174713	1.742	+
INSR	GGAAAGTGACACCAGAGCGT	chr19:7172348-7172367	2.616	+
MFAP5	TATGTCGCTCTTGGGACCCA	chr12:8814682-8814701	1.055	-
MFAP5	GAACTCACCCAAGTCATCTG	chr12:8807025-8807044	1.06	+
MFAP5	GTCGGAAGTAATTGGAGCGA	chr12:8800734-8800753	1.69	+
MFAP5	GAGCGACGGAGTCTCCTAGG	chr12:8800748-8800767	2.018	+
MLXIPL	GGACTCGGACACAGACTCGG	chr7:73038735-73038754	1.755	-
MLXIPL	CCTGAACAACGCCATCTGGA	chr7:73030419-73030438	0.933	-
MLXIPL	TGATGCGCACCGGAAGCCGG	chr7:73021922-73021941	2.232	-
MLXIPL	CTGGAAGCGGCGCATCGAGG	chr7:73021708-73021727	1.292	-
PCSK1	AGCAGCCTCGGCCATCGCCG	chr5:95768596-95768615	2.973	-
PCSK1	GAAGTAAACGTTCAGCTCTA	chr5:95761581-95761600	1.149	-
PCSK1	ATGAAGGTCCAGCTTGGGCA	chr5:95759119-95759138	1.025	+
PCSK1	AATGCTGGATGGCATTGTGA	chr5:95748168-95748187	1.061	-
PLIN1	GCGAGGATGGCAGTCAACAA	chr15:90220707-90220726	1.973	-
PLIN1	GAAGACCTACACCAGCACTA	chr15:90216558-90216577	3.053	-
PLIN1	CCGAGGCTTGGACCACCTGG	chr15:90214751-90214770	0.934	-
PLIN1	CTTGTCTGAAGTGCTCGCGA	chr15:90213386-90213405	1.757	+
PPARG	ATTCACAAGAACAGATCCAG	chr3:12421366-12421385	1.553	+
PPARG	AGAGCCTTCCAACTCCCTCA	chr3:12422894-12422913	1.344	+
PPARG	GTTTCAGAAATGCCTTGCAG	chr3:12434214-12434233	1.895	+
PPARG	GCTGACCAAAGCAAAGGCGA	chr3:12447526-12447545	3.041	+
SLTM	TGAAGTCCGAGCTGAAGCGG	chr15:59225657-59225676	2.418	-
SLTM	CAAGAGGCACATGAGCAAGA	chr15:59205858-59205877	1.651	-
SLTM	TAAGAAGGAAGACTGCGTGA	chr15:59191789-59191808	3.514	-
SLTM	GACGAGCAATGGAACTTCGA	chr15:59185098-59185117	0.893	-
TRIP10	CTTCGGCCATGTGACTCCGA	chr19:6743739-6743758	1.174	-
TRIP10	TGCTCGAGCGCCACACGCAG	chr19:6741034-6741053	0.851	+
TRIP10	CAAGACATACACGGACACTG	chr19:6743062-6743081	1.581	-
TRIP10	ATTTGAGCGGGACTGCCGGG	chr19:6743516-6743535	2.732	+
UBR2	TTCGGCCGAGGAGATTGCGG	chr6:42532114-42532133	3.19	+
UBR2	ATGCAGGATCTTCACCACAA	chr6:42541624-42541643	1.474	-
UBR2	TCCAAGATGACAACATCAGG	chr6:42561923-42561942	0.865	+
UBR2	TTGAGACATCGAGATGCCCA	chr6:42600353-42600372	1.808	+

Table S4.10: RNA-seq differential expression to confirm knockdown.

knockdown_ce	target_gene	ensembl_gene_	log2fc	log2fc_se	zscore	pval_1sided	fc	fc_95ci_lower	fc_95ci_upper
ABCA1	ABCA1	ENSG00000165	0.596	0.139	4.283	1.00e+00	1.511	1.251	1.825
COL5A3	COL5A3	ENSG00000080	-0.197	0.088	-2.246	1.24e-02	0.872	0.774	0.983
DENND5B	DENND5B	ENSG00000170	-1.789	0.127	-14.094	2.08e-45	0.289	0.243	0.344
EXOC7	EXOC7	ENSG00000182	-0.698	0.051	-13.742	2.83e-43	0.617	0.575	0.661
HERC1	HERC1	ENSG00000103	-1.216	0.186	-6.541	3.07e-11	0.431	0.335	0.554
INSR	INSR	ENSG00000171	-0.940	0.105	-8.940	1.94e-19	0.521	0.452	0.601
MFAP5	MFAP5	ENSG00000197	-1.364	0.138	-9.882	2.48e-23	0.388	0.322	0.469
MLXIPL	MLXIPL	ENSG00000009	-2.455	0.295	-8.322	4.32e-17	0.182	0.122	0.272
PCSK1	PCSK1	ENSG00000175	1.985	0.569	3.487	1.00e+00	3.959	1.827	8.580
PLIN1	PLIN1	ENSG00000166	0.195	0.224	0.872	8.08e-01	1.145	0.845	1.553
PPARG	PPARG	ENSG00000132	-0.165	0.128	-1.293	9.81e-02	0.892	0.750	1.061
SLTM	SLTM	ENSG00000137	-1.139	0.061	-18.692	2.89e-78	0.454	0.418	0.493
TRIP10	TRIP10	ENSG00000125	-0.481	0.063	-7.668	8.74e-15	0.717	0.658	0.780
UBR2	UBR2	ENSG00000024	-0.521	0.096	-5.441	2.65e-08	0.697	0.612	0.794

Table S4.11: PCR primers.

Gene/isoform	Forward/Rev	Sequence
PPARG1	Forward	AAAGAAGCCGACACTAAACC
PPARG1	Reverse	CTTCCATTACGGAGAGATCC
PPARG2	Forward	TCCATGCTGTTATGGGTGAA
PPARG2	Reverse	TCAAAGGAGTGGGAGTTGTC
ACTIN	Forward	CCAACCGCGAGAAGATGA
ACTIN	Reverse	CCAGAGGCGTACAGGGATAG
PLIN1	Forward	TGTGCAATGCCTATGAGAAGG
PLIN1	Reverse	AGGGCGGGGATCTTTTCCT
INSR	Forward	AAAACGAGGCCCGAAGATTTC
INSR	Reverse	GAGCCCATAGACCCGGAAG
COL5A3	Forward	TGACCGGGCATTCAGAATTGG
COL5A3	Reverse	CGGGCACCCCTTTCATCAT
EXOC7	Forward	AGCCCGGAACTCAACAAAGTG
EXOC7	Reverse	ATCAGGCTGCGAAATTCGGA
PCSK1	Forward	CTGGATGGCATTGTGACGGAT
PCSK1	Reverse	GCCCCAGCTTGCACTGTAAA
ABCA1	Forward	ACCCACCCTATGAACAACATGA
ABCA1	Reverse	GAGTCGGGTAACGGAAACAGG

Table S4.12: RT-qPCR differential expression to confirm knockdown.

knockdown_ce	qpcr_target	beta_actin_ct_r	beta_actin_ct_s	target_gene_ct	target_gene_ct	delta_ct	delta_ct_sem	delta_delta_ct	delta_delta_ct_	ddct_95ci_lowe	ddct_95ci_uppe	fc	fc_95ci_lower	fc_95ci_upper	delta_delta_ct_pval_1	sided
Cas9	ABCA1	16.773	0.084	26.103	0.268	9.330	0.281	0.000	0.281	-0.551	0.551	1.000	0.683	1.465	5.00e-01	
ABCA1	ABCA1	14.550	0.026	26.673	0.195	12.123	0.196	2.793	0.196	2.408	3.178	0.144	0.110	0.188	1.55e-05	
Cas9	PCSK1	18.613	0.050	27.633	0.171	9.020	0.178	0.000	0.178	-0.350	0.350	1.000	0.785	1.274	5.00e-01	
PCSK1	PCSK1	17.767	0.093	27.467	0.368	9.700	0.379	0.680	0.379	-0.063	1.423	0.624	0.373	1.045	6.65e-02	
Cas9	PPARG1	15.737	0.017	23.400	0.049	7.663	0.052	0.000	0.052	-0.102	0.102	1.000	0.932	1.073	5.00e-01	
PPARG	PPARG1	15.303	0.098	23.503	0.061	8.200	0.116	0.537	0.116	0.310	0.763	0.689	0.589	0.807	2.82e-03	
Cas9	PPARG2	15.737	0.017	27.173	0.239	11.437	0.239	0.000	0.239	-0.469	0.469	1.000	0.722	1.384	5.00e-01	
PPARG	PPARG2	15.303	0.098	31.067	0.032	15.763	0.103	4.327	0.103	4.124	4.529	0.050	0.043	0.057	7.29e-08	
Cas9	PLIN1	16.463	0.049	35.967	0.467	19.503	0.469	0.000	0.469	-0.920	0.920	1.000	0.528	1.892	5.00e-01	
PLIN1	PLIN1	16.300	0.035	35.903	0.642	19.603	0.643	0.100	0.643	-1.159	1.359	0.933	0.390	2.234	4.41e-01	

Table S4.13: Effect of knockdown on lipid accumulation.

	BODIPY lipid accumulation 10x, differentiated		
	Checked for Cook's distance outliers (**one outlier removed for *SLTM***)		
	Two-sided tests (control vs. knockdown)		
Knockdown cell line	**Fold change** (mean cytoplasm fluorescence, relative to Cas9 empty)	***t*-test sample size**	**Two-sided *t*-test pval**
Cas9 empty	1		
hWAT	1.199	12	1.02E-01
INSR	1.115	12	4.32E-01
PPARG	0.245	12	5.52E-07
ABCA1	1.296	12	5.23E-02
COL5A3	1.723	12	2.78E-03
DENND5B	0.808	12	6.98E-02
EXOC7	1.354	12	9.63E-03
HERC1	1.149	12	2.18E-01
MFAP5	1.185	12	3.23E-01
MLXIPL	1.149	12	3.99E-01
SLTM	0.514	11	1.91E-04
TRIP10	1.386	12	1.57E-02
UBR2	1.23	12	2.29E-01

Table S4.14: Evidence of druggability for knockdown target genes.

Gene	**Effect of gene knockdown** in hWAT cells on lipid accumulation (fold change; P-value)	Known drugs	Tractability		
			Small molecule	**Antibody**	**PROTAC**
ABCA1	Non-significant 1.30 (0.052)	PROBUCOL - small molecule inhibitor	Structure with ligand Druggable family	UniProt loc high conf GO CC high conf UniProt SigP or TMHMM	Half-life data
COL5A3	Increases lipid accumulation* 1.72 (0.0028)	COLLAGENASE CLOSTRIDIUM HISTOLYTICUM enzyme, acting through collagen hydrolytic enzyme OCRIPLASMIN - protein, acting through collagen h	-	UniProt loc high conf GO CC high conf UniProt loc med conf UniProt SigP or TMHMM Human Protein Atlas loc	-
DENND5B	Non-significant 0.81 (0.070)	-	-	UniProt SigP or TMHMM	Database ubiquitination Half-life data
EXOC7	Increases lipid accumulation* 1.35 (0096)	-	-	UniProt loc high conf GO CC high conf UniProt loc med conf Human protein atlas loc	Database ubiquitination Half-life data
HERC1	Non-significant 1.15 (0.22)	-	Structure with ligand	UniProt loc med conf	Database ubiquitination Half-life data
INSR	Non-significant 1.12 (0.43)	INSULIN - protein, insulin receptor agonist LINSITINIB, BMS-754807, KW-2450 FREE BASE,	Structure with ligand High-quality ligand High-quality pocket Druggable family	GO CC high conf UniProt loc med conf UniProt SigP or TMHMM Human protein atlas loc	Database ubiquitination Half-life data Small molecule binder
MFAP5	Non-significant 1.19 (0.32)	-	-	GO CC high conf UniProt loc med conf UniProt SigP or TMHMM	-
MLXIPL	Non-significant 1.15 (0.40)	-	Structure with ligand	-	-
PCSK1	Increases lipid accumulation* 1.52 (0.022)	-	High-quality ligand	UniProt SigP or TMHMM	Small molecule binder
PLIN1	Non-significant 1.10 (0.46)	-	High-quality ligand	-	Half-life data Small molecule binder
PPARG	Reduces lipid accumulation** 0.25 (5.5 × 10-7)	PIOGLITAZONE HYDROCHLORIDE, MESALAMIN	Structure with ligand High-quality ligand High-quality pocket Druggable family	-	UniProt Ubiquitination Database Ubiquitination Small molecule binder
SLTM	Reduces lipid accumulation** 0.51 (1.9 × 10-4)	-	-	-	UniProt Ubiquitination Database Ubiquitination Half-life data Small molecule binder
TRIP10	Increases lipid accumulation* 1.39 (0.016)	-	-	UniProt loc med conf	Database ubiquitination Half-life data
UBR2	Non-significant 1.23 (0.23)	-	-	Human protein atlas loc	Database ubiquitination Half-life data

Table S4.15: Genes differentially expressed between each knockdown and Cas9-empty cell line, with log FC **and** P**-value.**

knockdown	ensembl	symbol	baseMean	log2fc	log2fc_se	zscore	pvalue	padj	entrezid	uniprot
PPARG	ENSG0000017	*PODN*	7609.738533	-3.712199109	0.1234844996	-30.06206546	1.52E-198	3.29E-194	127435	Q7Z5L7
MLXIPL	ENSG0000017	*PODN*	7609.738533	-3.436997904	0.1233300306	-27.86829687	6.47E-171	1.65E-166	127435	Q7Z5L7
PPARG	ENSG0000010	*CXCL12*	29205.47967	-2.583621	0.1080221175	-23.91751857	2.01E-126	2.18E-122	6387	P48061
MLXIPL	ENSG0000004	*RIPOR3*	1561.942611	-3.591255436	0.1520582957	-23.61762256	2.54E-123	3.24E-119	140876	B7Z3F0
PPARG	ENSG0000018	*C1S*	69802.5854	-2.038567749	0.09124460587	-22.34178919	1.45E-110	1.05E-106	716	F8WCZ6
PPARG	ENSG0000004	*ELN*	1988.237674	-2.830991326	0.1280203415	-22.11360548	2.34E-108	1.26E-104	2006	P15502
MLXIPL	ENSG0000018	*KREMEN1*	1141.597862	-2.211632007	0.1006394706	-21.97579134	4.91E-107	4.17E-103	83999	Q96MU8
PPARG	ENSG0000010	*FCGRT*	802.0495779	-2.992470564	0.1386213494	-21.58737148	2.36E-103	9.34E-100	2217	A0A024QZI2
PPARG	ENSG0000008	*CDC14B*	784.6084108	-2.116307947	0.09805411745	-21.58306048	2.59E-103	9.34E-100	8555	O60729
PPARG	ENSG0000006	*PTPRU*	1154.92183	-2.436606012	0.1130109591	-21.56079402	4.19E-103	1.30E-99	10076	Q92729
PPARG	ENSG0000008	*MMP2*	103840.2386	-1.613920573	0.07492264597	-21.54115825	6.41E-103	1.73E-99	4313	P08253
PPARG	ENSG0000018	*KREMEN1*	1141.597862	-2.132719794	0.09912983989	-21.51440773	1.14E-102	2.74E-99	83999	Q96MU8
MLXIPL	ENSG0000018	*C1S*	69802.5854	-1.964904025	0.09125701552	-21.5315394	7.89E-103	5.03E-99	716	F8WCZ6
PPARG	ENSG0000013	*ANGPTL2*	14639.80916	-1.670093447	0.07785559934	-21.45116679	4.45E-102	9.63E-99	23452	A0A024R868
PPARG	ENSG0000019	*SLC2A10*	2989.808888	-1.639187538	0.07680023708	-21.34352184	4.48E-101	8.81E-98	81031	O95528
MLXIPL	ENSG0000010	*CXCL12*	29205.47967	-2.289146409	0.1079977597	-21.19623976	1.03E-99	5.28E-96	6387	P48061
MLXIPL	ENSG0000016	*GPRC5B*	1118.397742	-2.575908505	0.1232438328	-20.90091202	5.25E-97	2.23E-93	51704	B7Z831
MLXIPL	ENSG0000008	*CDC14B*	784.6084108	-2.06535892	0.09932840556	-20.79323541	4.98E-96	1.82E-92	8555	O60729
PPARG	ENSG0000016	*PLPP3*	76319.58354	-1.718034995	0.08290712057	-20.72240579	2.18E-95	3.92E-92	8613	O14495
PPARG	ENSG0000007	*CYBRD1*	36591.06515	-1.648272681	0.07994223147	-20.61829712	1.88E-94	3.13E-91	79901	Q53TN4
PPARG	ENSG0000017	*APLN*	2130.107625	-2.563536472	0.1259498824	-20.3536234	4.31E-92	6.66E-89	8862	Q9ULZ1
MLXIPL	ENSG0000012	*XPNPEP2*	2818.231218	-2.035201277	0.1001391298	-20.32373639	7.93E-92	2.53E-88	7512	O43895
PPARG	ENSG0000015	*C1R*	70354.22677	-2.173588594	0.1073973068	-20.23876257	4.46E-91	6.44E-88	715	P00736
PPARG	ENSG0000004	*CELF2*	1710.030015	-2.110098513	0.104662379	-20.16100278	2.15E-90	2.91E-87	10659	O95319
MLXIPL	ENSG0000004	*ELN*	1988.237674	-2.577104243	0.1277762464	-20.16888361	1.84E-90	5.21E-87	2006	P15502
MLXIPL	ENSG0000012	*C3*	50763.10943	-3.196588491	0.1586293465	-20.15130593	2.62E-90	6.68E-87	718	B4DR57
PPARG	ENSG0000012	*XPNPEP2*	2818.231218	-2.002720775	0.09961147146	-20.10532266	6.63E-90	8.44E-87	7512	O43895
MLXIPL	ENSG0000007	*SEMA3C*	7430.77732	-2.068723586	0.1034877305	-19.99003723	6.72E-89	1.56E-85	10512	Q99985
PPARG	ENSG0000004	*RIPOR3*	1561.942611	-2.936521482	0.1471676261	-19.95358327	1.40E-88	1.68E-85	140876	B7Z3F0
PPARG	ENSG0000016	*FBLN2*	14300.71261	-1.731389163	0.0874124355	-19.80712645	2.58E-87	2.94E-84	2199	P98095
MLXIPL	ENSG0000016	*ARHGEF3*	410.5567991	-2.595942282	0.1309791065	-19.81951435	2.02E-87	4.29E-84	50650	Q9NR81
MLXIPL	ENSG0000004	*CELF2*	1710.030015	-2.072071464	0.1053125304	-19.67545035	3.50E-86	6.87E-83	10659	O95319
MLXIPL	ENSG0000013	*GPNMB*	38387.69614	-2.090810593	0.1069651222	-19.54665735	4.40E-85	8.02E-82	10457	Q14956
MLXIPL	ENSG0000008	*MMP2*	103840.2386	-1.45696104	0.07492508056	-19.44557188	3.18E-84	5.40E-81	4313	P08253
PPARG	ENSG0000012	*SCPEP1*	4729.754587	-1.471830756	0.07646498118	-19.24842893	1.46E-82	1.57E-79	59342	Q9H2J9
PPARG	ENSG0000010	*PTGDS*	904.6965451	-4.482053378	0.2338415874	-19.16704992	6.98E-82	7.19E-79	5730	A0A024R8G3
MLXIPL	ENSG0000013	*ANGPTL2*	14639.80916	-1.493992709	0.07786444865	-19.18709674	4.74E-82	7.56E-79	23452	A0A024R868
MLXIPL	ENSG0000016	*PLPP3*	76319.58354	-1.589777956	0.08291282077	-19.17409082	6.09E-82	9.14E-79	8613	O14495
PPARG	ENSG0000017	*LRRN4CL*	2945.30918	-2.650914618	0.1385105277	-19.13872297	1.20E-81	1.18E-78	221091	Q8ND94
MLXIPL	ENSG0000015	*C1R*	70354.22677	-2.046868411	0.1074035458	-19.0577359	5.67E-81	8.03E-78	715	P00736
MLXIPL	ENSG0000006	*PTPRU*	1154.92183	-2.128959874	0.1122831265	-18.96063941	3.61E-80	4.84E-77	10076	Q92729
MLXIPL	ENSG0000007	*CYBRD1*	36591.06515	-1.514298005	0.07995453269	-18.93948916	5.39E-80	6.87E-77	79901	Q53TN4
MLXIPL	ENSG0000013	*SLC39A8*	1805.980268	-2.34893238	0.1242084196	-18.91121703	9.22E-80	1.12E-76	64116	A0A024RDG0
PPARG	ENSG0000018	*PRELP*	220.9330264	-4.364720095	0.2322406399	-18.79395482	8.46E-79	7.96E-76	5549	P51888
PPARG	ENSG0000014	*ADAMTSL4*	1980.148994	-2.237733572	0.1193224384	-18.75366949	1.81E-78	1.63E-75	54507	B7ZMJ3
PPARG	ENSG0000007	*FBLN1*	58569.51236	-2.744265173	0.1467171841	-18.7044564	4.55E-78	3.94E-75	2192	P23142
MLXIPL	ENSG0000018	*THBS2*	20567.81656	-1.540950739	0.08237790712	-18.70587385	4.43E-78	5.14E-75	7058	P35442
MLXIPL	ENSG0000017	*LRRN4CL*	2945.30918	-2.583073606	0.1387907012	-18.61128724	2.60E-77	2.89E-74	221091	Q8ND94
PPARG	ENSG0000012	*ZNF436*	2374.305426	-1.27708089	0.06872987127	-18.58116226	4.57E-77	3.80E-74	80818	A0A024RAF3
SLTM	ENSG0000013	*SLTM*	1991.5509	-1.138953462	0.06093356118	-18.69172653	5.78E-78	9.16E-74	79811	Q9NWH9
MLXIPL	ENSG0000003	*TIMP2*	45111.05927	-1.472501554	0.07986925663	-18.43639989	6.71E-76	7.05E-73	7077	A0A140VK57
PPARG	ENSG0000007	*P4HA2*	3270.033978	-1.000493626	0.05460694075	-18.32173002	5.55E-75	4.45E-72	8974	O15460
PPARG	ENSG0000018	*THBS2*	20567.81656	-1.495533285	0.08231255033	-18.16895818	9.09E-74	7.02E-71	7058	P35442
MLXIPL	ENSG0000018	*PRELP*	220.9330264	-4.351080802	0.2394249687	-18.1730453	8.44E-74	8.28E-71	5549	P51888
PPARG	ENSG0000019	*MFAP5*	29786.07311	-2.507548121	0.1382410489	-18.13895468	1.57E-73	1.17E-70	8076	B3KW70
MLXIPL	ENSG0000016	*SCARA5*	3909.42154	-3.330048553	0.183475569	-18.14982001	1.29E-73	1.22E-70	286133	Q6ZMJ2
PPARG	ENSG0000013	*IRF4*	689.0491764	-2.348293317	0.129574113	-18.12316721	2.09E-73	1.51E-70	3662	Q15306
PPARG	ENSG0000010	*FAM20A*	241.0101065	-3.541658935	0.1959618376	-18.07320741	5.18E-73	3.62E-70	54757	B7Z4Y3

Table S4.16: GSEA pathway analysis for RNA-seq of knockdown cell lines.

knockdown	source	term_name	pval_fwer	fdr_qval	es	nes	pval_nomin	size	rank_at_ma	leading_ed	ref_cond
HWAT	KEGG	REFERENCE_TRANSLATION_INITIATION	0	0	-0.73776484	-2.3193924	0	80	5534	tags=96%, list=	CAS9
MLXIPL	GO:MF	STRUCTURAL_CONSTITUENT_OF_RIBOSOME	0	0	-0.6967887	-2.5461855	0	161	6531	tags=89%, list=	CAS9
MLXIPL	REAC	TRANSLATION	0	0	-0.639043	-2.510293	0	293	7491	tags=81%, list=	CAS9
MLXIPL	REAC	EUKARYOTIC_TRANSLATION_INITIATION	0	0	-0.678907	-2.385718	0	118	6853	tags=88%, list=	CAS9
MLXIPL	REAC	EUKARYOTIC_TRANSLATION_ELONGATION	0	0	-0.69273406	-2.3670323	0	92	6531	tags=92%, list=	CAS9
PPARG	GO:MF	STRUCTURAL_CONSTITUENT_OF_CHROMATIN	0	0	-0.7698365	-2.4808223	0	82	2631	tags=55%, list=	CAS9
PPARG	REAC	MEIOTIC_RECOMBINATION	0	0	-0.77423364	-2.4663687	0	81	2988	tags=53%, list=	CAS9
PPARG	REAC	HDACS_DEACETYLATE_HISTONES	0	0	-0.75490135	-2.464103	0	90	2631	tags=42%, list=	CAS9
PPARG	REAC	ASSEMBLY_OF_THE_ORC_COMPLEX_AT_THE_ORIGIN_OF_R	0	0	-0.7883182	-2.4238055	0	64	2552	tags=53%, list=	CAS9
PPARG	REAC	SIRT1_NEGATIVELY_REGULATES_RRNA_EXPRESSION	0	0	-0.7877483	-2.4170198	0	63	2552	tags=52%, list=	CAS9
PPARG	REAC	ACTIVATED_PKN1_STIMULATES_TRANSCRIPTION_OF_AR_AI	0	0	-0.7866811	-2.4151099	0	62	2552	tags=55%, list=	CAS9
PPARG	REAC	PRC2_METHYLATES_HISTONES_AND_DNA	0	0	-0.77198243	-2.4142282	0	68	2552	tags=50%, list=	CAS9
PPARG	REAC	MEIOSIS	0	0	-0.7141444	-2.4115667	0	111	2988	tags=42%, list=	CAS9
PPARG	REAC	DNA_METHYLATION	0	0	-0.79296345	-2.4091816	0	59	2552	tags=58%, list=	CAS9
PPARG	REAC	HATS_ACETYLATE_HISTONES	0	0	-0.69032586	-2.3985481	0	138	2631	tags=28%, list=	CAS9
PPARG	REAC	ERCC6_CSB_AND_EHMT2_G9A_POSITIVELY_REGULATE_RR	0	0	-0.7583654	-2.3794625	0	71	2552	tags=45%, list=	CAS9
PPARG	REAC	CONDENSATION_OF_PROPHASE_CHROMOSOMES	0	0	-0.76128596	-2.3784404	0	69	2673	tags=54%, list=	CAS9
PPARG	REAC	CELL_CYCLE_CHECKPOINTS	0	0	-0.62783736	-2.3727796	0	286	3531	tags=35%, list=	CAS9
PPARG	REAC	HCMV_LATE_EVENTS	0	0	-0.7034629	-2.3719025	0	112	2799	tags=36%, list=	CAS9
PPARG	REAC	DNA_REPLICATION	0	0	-0.65857834	-2.3685408	0	183	3294	tags=37%, list=	CAS9
TRIP10	GO:MF	STRUCTURAL_CONSTITUENT_OF_RIBOSOME	0	0	-0.67637557	-2.4374495	0	161	6657	tags=88%, list=	CAS9
TRIP10	REAC	TRANSLATION	0	0	-0.63166666	-2.4307377	0	293	7133	tags=79%, list=	CAS9
UBR2	KEGG	REFERENCE_TRANSLATION_INITIATION	0	0	-0.7635457	-2.2848167	0	80	5569	tags=98%, list=	CAS9
UBR2	KEGG	RIBOSOME	0	0	-0.7538681	-2.284199	0	87	5569	tags=94%, list=	CAS9
PPARG	REAC	POSITIVE_EPIGENETIC_REGULATION_OF_RRNA_EXPRESSI	0.0001	3.16E-06	-0.6810356	-2.255813	0	101	2552	tags=32%, list=	CAS9
PPARG	REAC	SENESCENCE_ASSOCIATED_SECRETORY_PHENOTYPE_SAS	0.0001	3.26E-06	-0.67774266	-2.2733104	0	106	2552	tags=35%, list=	CAS9
PPARG	REAC	MITOTIC_SPINDLE_CHECKPOINT	0.0001	3.35E-06	-0.680116	-2.2765875	0	110	3682	tags=41%, list=	CAS9
PPARG	REAC	TRANSCRIPTIONAL_REGULATION_BY_SMALL_RNAS	0.0001	3.46E-06	-0.68469936	-2.2770076	0	102	2552	tags=32%, list=	CAS9
PPARG	REAC	PKMTS_METHYLATE_HISTONE_LYSINES	0.0001	3.57E-06	-0.7325468	-2.2890484	0	68	2320	tags=29%, list=	CAS9
PPARG	REAC	NEGATIVE_EPIGENETIC_REGULATION_OF_RRNA_EXPRESSI	0.0001	3.69E-06	-0.6868068	-2.293121	0	104	2552	tags=33%, list=	CAS9
PPARG	REAC	M_PHASE	0.0001	3.82E-06	-0.58967656	-2.3005917	0	409	6351	tags=48%, list=	CAS9
PPARG	REAC	RNA_POLYMERASE_I_PROMOTER_ESCAPE	0.0001	3.95E-06	-0.7146818	-2.3119805	0	86	2552	tags=37%, list=	CAS9
PPARG	REAC	RHO_GTPASES_ACTIVATE_PKNS	0.0001	4.10E-06	-0.711437	-2.3120077	0	89	2552	tags=39%, list=	CAS9
PPARG	REAC	CHROMOSOME_MAINTENANCE	0.0001	4.26E-06	-0.67034733	-2.3184755	0	135	5441	tags=61%, list=	CAS9
PPARG	REAC	RUNX1_REGULATES_GENES_INVOLVED_IN_MEGAKARYOCY	0.0001	4.43E-06	-0.71112597	-2.3201935	0	89	2552	tags=38%, list=	CAS9
PPARG	REAC	B_WICH_COMPLEX_POSITIVELY_REGULATES_RRNA_EXPRE	0.0001	4.61E-06	-0.7140344	-2.321661	0	86	2552	tags=37%, list=	CAS9
PPARG	REAC	FORMATION_OF_THE_BETA_CATENIN_TCF_TRANSACTIVATI	0.0001	4.81E-06	-0.71523786	-2.3266206	0	87	2552	tags=40%, list=	CAS9
PPARG	REAC	CHROMATIN_MODIFICATIONS_DURING_THE_MATERNAL_TO_	0.0001	5.03E-06	-0.75435656	-2.3302453	0	65	2552	tags=51%, list=	CAS9
PPARG	REAC	DNA_REPLICATION_PRE_INITIATION	0.0001	5.27E-06	-0.6630819	-2.3348327	0	155	3294	tags=35%, list=	CAS9
PPARG	REAC	DEPOSITION_OF_NEW_CENPA_CONTAINING_NUCLEOSOMES	0.0001	5.54E-06	-0.75066143	-2.342062	0	68	3154	tags=59%, list=	CAS9
PPARG	REAC	DISEASES_OF_PROGRAMMED_CELL_DEATH	0.0001	5.83E-06	-0.7052885	-2.344768	0	99	2698	tags=41%, list=	CAS9
PPARG	GO:BP	NUCLEAR_CHROMOSOME_SEGREGATION	0.0002	6.12E-06	-0.59468496	-2.2549517	0	300	3886	tags=36%, list=	CAS9
PPARG	REAC	HCMV_EARLY_EVENTS	0.0001	6.15E-06	-0.67819524	-2.3520193	0	132	2799	tags=32%, list=	CAS9
PPARG	REAC	DNA_DAMAGE_TELOMERE_STRESS_INDUCED_SENESCENC	0.0001	6.51E-06	-0.7446326	-2.3527577	0	74	2552	tags=46%, list=	CAS9
PPARG	REAC	RMTS_METHYLATE_HISTONE_ARGININES	0.0001	6.92E-06	-0.7406531	-2.3544955	0	74	2631	tags=38%, list=	CAS9
PPARG	GO:BP	MITOTIC_SISTER_CHROMATID_SEGREGATION	0.0003	8.28E-06	-0.62492484	-2.2442813	0	186	3864	tags=40%, list=	CAS9
PPARG	REAC	HCMV_INFECTION	0.0003	8.49E-06	-0.6383467	-2.2490819	0	156	2799	tags=27%, list=	CAS9
PPARG	REAC	PRE_NOTCH_EXPRESSION_AND_PROCESSING	0.0003	8.71E-06	-0.67438376	-2.250143	0	104	2559	tags=34%, list=	CAS9
PPARG	REAC	RESOLUTION_OF_SISTER_CHROMATID_COHESION	0.0003	8.95E-06	-0.65540546	-2.2523198	0	123	3682	tags=40%, list=	CAS9
PPARG	REAC	MATERNAL_TO_ZYGOTIC_TRANSITION_MZT	0.0004	1.01E-05	-0.6639556	-2.2360573	0	112	2552	tags=31%, list=	CAS9
PPARG	REAC	RECOGNITION_AND_ASSOCIATION_OF_DNA_GLYCOSYLASE	0.0004	1.03E-05	-0.75546706	-2.2389398	0	51	2552	tags=51%, list=	CAS9
PPARG	REAC	MITOTIC_PROPHASE	0.0004	1.05E-05	-0.64697194	-2.2398145	0	138	5249	tags=49%, list=	CAS9
PPARG	REAC	RNA_POLYMERASE_I_TRANSCRIPTION	0.0004	1.08E-05	-0.6734025	-2.2416832	0	106	2552	tags=30%, list=	CAS9
PPARG	REAC	HOMOLOGY_DIRECTED_REPAIR	0.0005	1.13E-05	-0.6373751	-2.2112637	0	134	3676	tags=42%, list=	CAS9
PPARG	REAC	ACTIVATION_OF_ANTERIOR_HOX_GENES_IN_HINDBRAIN_DE	0.0005	1.15E-05	-0.65323895	-2.2140524	0	117	2552	tags=30%, list=	CAS9
PPARG	GO:BP	SISTER_CHROMATID_SEGREGATION	0.0005	1.17E-05	-0.6041861	-2.2169898	0	225	3864	tags=36%, list=	CAS9
PPARG	GO:BP	CHROMOSOME_SEPARATION	0.0005	1.20E-05	-0.6985851	-2.2189248	0	79	3886	tags=51%, list=	CAS9
PPARG	GO:BP	REGULATION OF CHROMOSOME SEGREGATION	0.0005	1.23E-05	-0.65215725	-2.2298179	0	125	3739	tags=40%, list=	CAS9

References

1. Poczai, P. & Santiago-Blay, J. A. Chip off the old block: Generation, development, and ancestral concepts of heredity. en. *Front. Genet.* **13,** 814436. doi:10.3389/fgene.2022.814436. (9, 2022).
2. Kočandrle, R. & Kleisner, K. Evolution born of moisture: Analogies and parallels between Anaximander's ideas on origin of life and man and later pre-Darwinian and Darwinian evolutionary concepts. *J. Hist. Biol.* **46,** 103–124. doi:10.1007/s10739-012-9334-8. (1 2013).
3. Burkhardt Jr, R. W. Lamarck, evolution, and the inheritance of acquired characters. en. *Genetics* **194,** 793–805. doi:10.1534/genetics.113.151852. (4 2013).
4. Lamarck, J.-B. d. M. *Recherches sur l'organisation des corps vivans et particulièrement sur son origine, sur la cause de ses développemens et des progrès de sa composition, et sur celle qui, tendant continuellement à la détruire dans chaque individu, amène nécessairement sa mort; précédé du discours d'ouverture du course de zoologie, donné dans le Muséum national d'Histoire* fr. (Maillard, Paris, 1802).
5. Mayr, E. Lamarck revisited. *J. Hist. Biol.* **5,** 55–94. doi:10.1007/BF02113486. (1 1972).
6. Lamarck, J.-B. d. M. *Philosophie zoologique, ou Exposition des considérations relatives à l'histoire naturelle des animaux* fr (Muséum national d'histoire naturelle, 1809).
7. Darwin, C. *On the origin of species by means of natural selection, or, The preservation of favoured races in the struggle for life /* doi:10.5962/bhl.title.68064. (John Murray, London : 1859).
8. Mendel, G., Punnett, R. C. & Burndy Library. *Versuche über Pflanzen-Hybriden /* doi:10.5962/bhl.title.61004. (Im Verlage des Vereines, Brünn : 1866).
9. Stenseth, N. C., Andersson, L. & Hoekstra, H. E. Gregor Johann Mendel and the development of modern evolutionary biology. en. *Proc. Natl. Acad. Sci. U. S. A.* **119,** e2201327119. doi:10.1073/pnas.2201327119. (30 26, 2022).
10. Berry, A. & Browne, J. Mendel and Darwin. en. *Proc. Natl. Acad. Sci. U. S. A.* **119,** e2122144119. doi:10.1073/pnas.2122144119. (30 26, 2022).
11. Bateson, W. & Mendel, G. *Cambridge library collection - Darwin, evolution and genetics: Mendel's principles of heredity: A defence, with a translation of Mendel's original papers on hybridisation: A defence, with a translation of Mendel's original papers on hybridisation* doi:10.1017/cbo9780511694462. (Cambridge University Press, Cambridge, England, 7, 2010).

12. Galton, F. Regression towards mediocrity in hereditary stature. *J. Anthropol. Inst. G. B. Irel.* **15,** 246. doi:10.2307/2841583. (1886).

13. Fancher, R. E. Scientific cousins: the relationship between Charles Darwin and Francis Galton. en. *Am. Psychol.* **64,** 84–92. doi:10.1037/a0013339. (2 2009).

14. Pearson, K. Notes on the history of correlation. *Biometrika* **13,** 25–45. doi:10.1093/biomet/13.1.25. (1 1920).

15. Pearson, K. & Lee, A. On the laws of inheritance in man: I. Inheritance of physical characters. en. *Biometrika* **2,** 357–462. doi:10.1093/biomet/2.4.357. (4 1, 1903).

16. Visscher, P. M. Commentary: Height and Mendel's theory: the long and the short of it. en. *Int. J. Epidemiol.* **42,** 944–945. doi:10.1093/ije/dyt069. (4 2013).

17. Fisher, R. A. XV.—the correlation between relatives on the supposition of Mendelian inheritance. en. *Trans. R. Soc. Edinb.* **52,** 399–433. doi:10.1017/s0080456800012163. (2 1919).

18. Visscher, P. M. & Goddard, M. E. From R.a. fisher's 1918 paper to GWAS a century later. en. *Genetics* **211,** 1125–1130. doi:10.1534/genetics.118.301594. (4 2019).

19. Rao, C. R. R. A. Fisher: The founder of modern statistics. *Stat. Sci.* **7,** 34–48. doi:10.1214/ss/1177011442. (1 1, 1992).

20. Varberg, D. E. The development of modern statistics. *Math. Teach.* **56,** 252–257. (4 1963).

21. Flemming, W. *Zellsubstanz, Kern und Zelltheilung /* doi:10.5962/bhl.title.168645. (F.C.W. Vogel, Leipzig : 1882).

22. Crow, E. W. & Crow, J. F. 100 years ago: Walter Sutton and the chromosome theory of heredity. en. *Genetics* **160,** 1–4. doi:10.1093/genetics/160.1.1. (1 2002).

23. Sutton, W. S. On the morphology of the chromoso group in Brachystola Magna. *Biol. Bull.* **4,** 24–39. doi:10.2307/1535510. (1 1902).

24. Green, M. M. 2010: A century of Drosophila genetics through the prism of the white gene. en. *Genetics* **184,** 3–7. doi:10.1534/genetics.109.110015. (1 2010).

25. Morgan, T. H. Sex limited inheritance in Drosophila. en. *Science* **32,** 120–122. doi:10.1126/science.32.812.120. (812 22, 1910).

26. Morgan, T. Random segregation versus coupling in Mendelian inheritance. *Science* **34,** 384. doi:10.1126/SCIENCE.34.873.384. (873 22, 1911).

27. Sturtevant, A. H. The linear arrangement of six sex-linked factors in Drosophila, as shown by their mode of association. en. *J. Exp. Zool.* **14,** 43–59. doi:10.1002/jez.1400140104. (1 20, 1913).

28. Altshuler, D., Daly, M. J. & Lander, E. S. Genetic mapping in human disease. en. *Science* **322,** 881–888. doi:10.1126/science.1156409. (5903 7, 2008).

29. Ingram, V. M. A specific chemical difference between the globins of normal human and sickle-cell anæmia hæmoglobin. en. *Nature* **178,** 792–794. doi:10.1038/178792a0. (4537 13, 1956).

30. Jeffreys, A. J. DNA sequence variants in the G gamma-, A gamma-, delta- and beta-globin genes of man. en. *Cell* **18,** 1–10. doi:10.1016/0092-8674(79)90348-9. (1 1979).

31. Kan, Y. W. & Dozy, A. M. Polymorphism of DNA sequence adjacent to human beta-globin structural gene: relationship to sickle mutation. en. *Proc. Natl. Acad. Sci. U. S. A.* **75,** 5631–5635. doi:10.1073/pnas.75.11.5631. (11 1978).

32. Watson, J. D. & Crick, F. H. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. en. *Nature* **171,** 737–738. doi:10.1038/171737a0. (4356 25, 1953).

33. Cobb, M. 60 years ago, Francis Crick changed the logic of biology. en. *PLoS Biol.* **15,** e2003243. doi:10.1371/journal.pbio.2003243. (9 2017).

34. Crick, F. H. On protein synthesis. en. *Symp. Soc. Exp. Biol.* **12,** 138–163. (1958).

35. Hutchison 3rd, C. A. DNA sequencing: bench to bedside and beyond. en. *Nucleic Acids Res.* **35,** 6227–6237. doi:10.1093/nar/gkm688. (18 12, 2007).

36. Sanger, F. The terminal peptides of insulin. en. *Biochem. J.* **45,** 563–574. doi:10.1042/bj0450563. (5 1949).

37. Sanger, F. & Tuppy, H. The amino-acid sequence in the phenylalanyl chain of insulin. 2. The investigation of peptides from enzymic hydrolysates. en. *Biochem. J.* **49,** 481–490. doi:10.1042/bj0490481. (4 1951).

38. Shendure, J. *et al.* DNA sequencing at 40: past, present and future. en. *Nature* **550,** 345–353. doi:10.1038/nature24286. (7676 19, 2017).

39. Maxam, A. M. & Gilbert, W. A new method for sequencing DNA. en. *Proc. Natl. Acad. Sci. U. S. A.* **74,** 560–564. doi:10.1073/pnas.74.2.560. (2 1977).

40. Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. en. *Proc. Natl. Acad. Sci. U. S. A.* **74,** 5463–5467. doi:10.1073/pnas.74.12.5463. (12 1977).

41. Heather, J. M. & Chain, B. The sequence of sequencers: The history of sequencing DNA. en. *Genomics* **107,** 1–8. doi:10.1016/j.ygeno.2015.11.003. (1 2016).

42. Botstein, D., White, R. L., Skolnick, M. & Davis, R. W. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. en. *Am. J. Hum. Genet.* **32,** 314–331. (3 1980).

43. De Souza, N. Get out the map: RFLP concept; RFLP realization. en. *Nat. Rev. Genet.* **8,** S10–S10. doi:10.1038/nrg2260. (S1 2007).

44. Gusella, J. F. *et al.* A polymorphic DNA marker genetically linked to Huntington's disease. en. *Nature* **306,** 234–238. doi:10.1038/306234a0. (5940 1983).

45. Collins, F. S., Guyer, M. S. & Charkravarti, A. Variations on a theme: cataloging human DNA sequence variation. en. *Science* **278,** 1580–1581. doi:10.1126/science.278.5343.1580. (5343 28, 1997).

46. Bender, W. *et al.* Molecular Genetics of the Bithorax Complex in Drosophila melanogaster. en. *Science* **221,** 23–29. doi:10.1126/science.221.4605.23. (4605 1, 1983).

47. Rommens, J. M. *et al.* Identification of the cystic fibrosis gene: chromosome walking and jumping. en. *Science* **245,** 1059–1065. doi:10.1126/science.2772657. (4922 8, 1989).

48. Corder, E. H. *et al.* Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. en. *Science* **261,** 921–923. doi:10.1126/science.8346443. (5123 13, 1993).

49. Cooper, D. N. & Schmidtke, J. Molecular genetic approaches to the analysis and diagnosis of human inherited disease: an overview. en. *Ann. Med.* **24,** 29–42. doi:10.3109/07853899209164142. (1 1992).

50. Lander, E. S. & Schork, N. J. Genetic dissection of complex traits. en. *Science* **265,** 2037–2048. doi:10.1126/science.8091226. (5181 30, 1994).

51. McKusick, V., Francomano, C. A. & Antonarakis, S. E. *Mendelian inheritance in man: Vol 1* 10th ed. en. 2560 pp. (Johns Hopkins University Press, Baltimore, MD, 1, 1992).

52. Risch, N. J. Searching for genetic determinants in the new millennium. en. *Nature* **405,** 847–856. doi:10.1038/35015718. (6788 15, 2000).

53. Risch, N. & Merikangas, K. The future of genetic studies of complex human diseases. en. *Science* **273,** 1516–1517. doi:10.1126/science.273.5281.1516. (5281 13, 1996).

54. International HapMap Consortium. The International HapMap Project. en. *Nature* **426,** 789–796. doi:10.1038/nature02168. (6968 18, 2003).

55. Kennedy, G. C. *et al.* Large-scale genotyping of complex DNA. en. *Nat. Biotechnol.* **21,** 1233–1237. doi:10.1038/nbt869. (10 2003).

56. Wang, D. G. *et al.* Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. en. *Science* **280,** 1077–1082. doi:10.1126/science.280.5366.1077. (5366 15, 1998).

57. Klein, R. J. *et al.* Complement factor H polymorphism in age-related macular degeneration. en. *Science* **308,** 385–389. doi:10.1126/science.1109557. (5720 15, 2005).

58. Das, S., Abecasis, G. R. & Browning, B. L. Genotype imputation from large reference panels. en. *Annu. Rev. Genomics Hum. Genet.* **19,** 73–96. doi:10.1146/annurev-genom-083117-021602. (31, 2018).

59. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. en. *Nature* **447,** 661–678. doi:10.1038/nature05911. (7145 7, 2007).

60. Hindorff, L. A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. en. *Proc. Natl. Acad. Sci. U. S. A.* **106,** 9362–9367. doi:10.1073/pnas.0903103106. (23 9, 2009).

61. Manolio, T. A. Genomewide association studies and assessment of the risk of disease. en. *N. Engl. J. Med.* **363,** 166–176. doi:10.1056/NEJMra0905980. (2 8, 2010).

62. Frayling, T. M. *et al.* A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. en. *Science* **316,** 889–894. doi:10.1126/science.1141634. (5826 11, 2007).

63. Claussnitzer, M. *et al.* FTO obesity variant circuitry and adipocyte browning in humans. en. *N. Engl. J. Med.* **373,** 895–907. doi:10.1056/NEJMoa1502214. (10 3, 2015).

64. Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. en. *Nature* **511,** 421–427. doi:10.1038/nature13595. (7510 24, 2014).

65. Sekar, A. *et al.* Schizophrenia risk from complex variation of complement component 4. en. *Nature* **530,** 177–183. doi:10.1038/nature16549. (7589 11, 2016).

66. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. en. *Nature* **526,** 68–74. doi:10.1038/nature15393. (7571 1, 2015).

67. Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. A new multipoint method for genome-wide association studies by imputation of genotypes. en. *Nat. Genet.* **39,** 906–913. doi:10.1038/ng2088. (7 2007).

68. Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. en. *Nat. Rev. Genet.* **11,** 499–511. doi:10.1038/nrg2796. (7 2010).

69. International Schizophrenia Consortium *et al.* Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. en. *Nature* **460,** 748–752. doi:10.1038/nature08185. (7256 6, 2009).

70. Wray, N. R., Goddard, M. E. & Visscher, P. M. Prediction of individual genetic risk to disease from genome-wide association studies. en. *Genome Res* **17,** 1520–1528. doi:10.1101/gr.6665407. (10 2007).

71. Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. en. *Nat. Genet.* **47,** 291–295. doi:10.1038/ng.3211. (3 2015).

72. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. en. *Nat. Genet.* **42,** 565–569. doi:10.1038/ng.608. (7 2010).

73. Abdellaoui, A., Yengo, L., Verweij, K. J. H. & Visscher, P. M. 15 years of GWAS discovery: Realizing the promise. en. *Am. J. Hum. Genet.* **110,** 179–194. doi:10.1016/j.ajhg.2022.12.011. (2 2, 2023).

74. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. en. *Nature* **562,** 203–209. doi:10.1038/s41586-018-0579-z. (7726 2018).

75. Kurki, M. I. *et al.* FinnGen provides genetic insights from a well-phenotyped isolated population. en. *Nature* **613,** 508–518. doi:10.1038/s41586-022-05473-8. (7944 2023).

76. Nagai, A. *et al.* Overview of the BioBank Japan Project: Study design and profile. en. *J. Epidemiol.* **27,** S2–S8. doi:10.1016/j.je.2016.12.005. (3S 2017).

77. Gaziano, J. M. *et al.* Million Veteran Program: A mega-biobank to study genetic influences on health and disease. en. *J. Clin. Epidemiol.* **70,** 214–223. doi:10.1016/j.jclinepi.2015.09.016. (2016).

78. Stefansson, K. 2017 William Allan award. en. *Am. J. Hum. Genet.* **102,** 351–353. doi:10.1016/j.ajhg.2018.01.012. (3 1, 2018).

79. Leitsalu, L. & Metspalu, A. in *Genomic and Precision Medicine* (eds Ginsburg, G. S. & Willard, H. F.) 119–129 (Elsevier, 2017). doi:10.1016/b978-0-12-800681-8.00008-6.

80. Sijtsma, A. *et al.* Cohort Profile Update: Lifelines, a three-generation cohort study and biobank. en. *Int. J. Epidemiol.* **51,** e295–e302. doi:10.1093/ije/dyab257. (5 13, 2022).

81. Walters, R. G. *et al.* Genotyping and population characteristics of the China Kadoorie Biobank. en. *Cell Genom.* **3,** 100361. doi:10.1016/j.xgen.2023.100361. (8 9, 2023).

82. All of Us Research Program Investigators *et al.* The "All of Us" Research Program. en. *N. Engl. J. Med.* **381,** 668–676. doi:10.1056/NEJMsr1809937. (7 15, 2019).

83. 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. en. *PLoS Med.* **12,** e1001779. doi:10.1371/journal.pmed.1001779. (3 2015).

84. Julkunen, H. *et al.* Atlas of plasma NMR biomarkers for health and disease in 118,461 individuals from the UK Biobank. en. *Nat. Commun.* **14,** 604. doi:10.1038/s41467-023-36231-7. (1 3, 2023).

85. Sinnott-Armstrong, N. *et al.* Genetics of 35 blood and urine biomarkers in the UK Biobank. en. *Nat. Genet.* **53,** 185–194. doi:10.1038/s41588-020-00757-z. (2 2021).

86. Sun, B. B. *et al.* Plasma proteomic associations with genetics and health in the UK Biobank. en. *Nature* **622,** 329–338. doi:10.1038/s41586-023-06592-6. (7982 2023).

87. Elliott, L. T. *et al.* Genome-wide association studies of brain imaging phenotypes in UK Biobank. en. *Nature* **562,** 210–216. doi:10.1038/s41586-018-0571-7. (7726 2018).

88. Littlejohns, T. J. *et al.* The UK Biobank imaging enhancement of 100,000 participants: rationale, data collection, management and future directions. en. *Nat. Commun.* **11,** 2624. doi:10.1038/s41467-020-15948-9. (1 26, 2020).

89. Kang, H. M. *et al.* Efficient control of population structure in model organism association mapping. en. *Genetics* **178,** 1709–1723. doi:10.1534/genetics.107.080101. (3 2008).

90. Voight, B. F. & Pritchard, J. K. Confounding from cryptic relatedness in case-control association studies. en. *PLoS Genet.* **1,** e32. doi:10.1371/journal.pgen.0010032. (3 2005).

91. Loh, P.-R. *et al.* Efficient Bayesian mixed-model analysis increases association power in large cohorts. en. *Nat. Genet.* **47,** 284–290. doi:10.1038/ng.3190. (3 2015).

92. Jiang, L. *et al.* A resource-efficient tool for mixed model association analysis of large-scale data. en. *Nat. Genet.* **51,** 1749–1755. doi:10.1038/s41588-019-0530-8. (12 2019).

93. Mbatchou, J. *et al.* Computationally efficient whole-genome regression for quantitative and binary traits. en. *Nat. Genet.* **53,** 1097–1103. doi:10.1038/s41588-021-00870-7. (7 2021).

94. Zhou, W. *et al.* Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. en. *Nat. Genet.* **50,** 1335–1341. doi:10.1038/s41588-018-0184-y. (9 2018).

95. Kiezun, A. *et al.* Exome sequencing and the genetic basis of complex traits. en. *Nat. Genet.* **44,** 623–630. doi:10.1038/ng.2303. (6 29, 2012).

96. Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. en. *Nature* **437,** 376–380. doi:10.1038/nature03959. (7057 15, 2005).

97. Zuk, O. *et al.* Searching for missing heritability: designing rare variant association studies. en. *Proc. Natl. Acad. Sci. U. S. A.* **111,** E455–64. doi:10.1073/pnas.1322563111. (4 28, 2014).

98. Asimit, J. & Zeggini, E. Rare variant association analysis methods for complex traits. en. *Annu. Rev. Genet.* **44,** 293–308. doi:10.1146/annurev-genet-102209-163421. (1 2010).

99. Lee, S., Abecasis, G. R., Boehnke, M. & Lin, X. Rare-variant association analysis: study designs and statistical tests. en. *Am. J. Hum. Genet.* **95,** 5–23. doi:10.1016/j.ajhg.2014.06.009. (1 3, 2014).

100. Dering, C. *et al.* A comprehensive evaluation of collapsing methods using simulated and real data: excellent annotation of functionality and large sample sizes required. en. *Front. Genet.* **5,** 323. doi:10.3389/fgene.2014.00323. (15, 2014).

101. Fitze, G. *et al.* Association between c135G/A genotype and RET proto-oncogene germline mutations and phenotype of Hirschsprung's disease. en. *Lancet* **359,** 1200–1205. doi:10.1016/S0140-6736(02)08218-1. (9313 6, 2002).

102. Morgenthaler, S. & Thilly, W. G. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). en. *Mutat. Res.* **615,** 28–56. doi:10.1016/j.mrfmmm.2006.09.003. (1-2 3, 2007).

103. Zhou, W. *et al.* Scalable generalized linear mixed model for region-based association tests in large biobanks and cohorts. en. *Nat. Genet.* **52,** 634–639. doi:10.1038/s41588-020-0621-6. (6 2020).

104. Lee, S. *et al.* Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. en. *Am. J. Hum. Genet.* **91,** 224–237. doi:10.1016/j.ajhg.2012.06.007. (2 10, 2012).

105. Landrum, M. J. *et al.* ClinVar: public archive of relationships among sequence variation and human phenotype. en. *Nucleic Acids Res.* **42,** D980–5. doi:10.1093/nar/gkt1113. (Database issue 2014).

106. Rehm, H. L. *et al.* ClinGen–the clinical genome resource. en. *N. Engl. J. Med.* **372,** 2235–2242. doi:10.1056/NEJMsr1406261. (23 4, 2015).

107. Tabet, D., Parikh, V., Mali, P., Roth, F. P. & Claussnitzer, M. Scalable functional assays for the interpretation of human genetic variation. en. *Annu. Rev. Genet.* **56,** 441–465. doi:10.1146/annurev-genet-072920-032107. (1 30, 2022).

108. Livesey, B. J. & Marsh, J. A. Interpreting protein variant effects with computational predictors and deep mutational scanning. en. *Dis. Model. Mech.* **15,** dmm049510. doi:10.1242/dmm.049510. (6 1, 2022).

109. Ioannidis, N. M. *et al.* REVEL: An ensemble method for predicting the pathogenicity of rare missense variants. en. *Am. J. Hum. Genet.* **99,** 877–885. doi:10.1016/j.ajhg.2016.08.016. (4 6, 2016).

110. Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. CADD: predicting the deleteriousness of variants throughout the human genome. en. *Nucleic Acids Res.* **47,** D886–D894. doi:10.1093/nar/gky1016. (D1 8, 2019).

111. Jaganathan, K. *et al.* Predicting Splicing from Primary Sequence with Deep Learning. en. *Cell* **176,** 535–548.e24. doi:10.1016/j.cell.2018.12.015. (3 24, 2019).

112. 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. en. *Genet. Med.* **17,** 405–424. doi:10.1038/gim.2015.30. (5 2015).

113. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. en. *Nature* **581,** 434–443. doi:10.1038/s41586-020-2308-7. (7809 2020).

114. Backman, J. D. *et al.* Exome sequencing and analysis of 454,787 UK Biobank participants. en. *Nature* **599,** 628–634. doi:10.1038/s41586-021-04103-z. (2025) (7886 18, 2021).

115. All of Us Research Program Genomics Investigators. Genomic data in the All of Us Research Program. en. *Nature* **627,** 340–346. doi:10.1038/s41586-023-06957-x. (8003 2024).

116. Wong, E. *et al.* The Singapore National Precision Medicine strategy. en. *Nat. Genet.* **55,** 178–186. doi:10.1038/s41588-022-01274-x. (2 2023).

117. Cook, M. B. *et al.* Our Future Health: a unique global resource for discovery and translational research. en. *Nat. Med.* **31,** 728–730. doi:10.1038/s41591-024-03438-0. (3 2025).

118. Halldorsson, B. V. *et al.* The sequences of 150,119 genomes in the UK Biobank. en. *Nature* **607,** 732–740. doi:10.1038/s41586-022-04965-x. (7920 2022).

119. Hawkes, G. *et al.* Whole-genome sequencing in 333,100 individuals reveals rare non-coding single variant and aggregate associations with height. en. *Nat. Commun.* **15,** 8549. doi:10.1038/s41467-024-52579-w. (1 3, 2024).

120. Li, S., Carss, K. J., Halldorsson, B. V. & Cortes, A. Whole-genome sequencing of half-a-million UK Biobank participants. *Genetic and Genomic Medicine.* (8, 2023).

121. Wetterstrand, K. A. *DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP)* en. https://www.genome.gov/sequencingcostsdata (2025).

122. Biomarkers Definitions Working Group. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. en. *Clin. Pharmacol. Ther.* **69,** 89–95. doi:10.1067/mcp.2001.113989. (3 2001).

123. Burroughs Wellcome & Company. *The evolution of urine analysis: an historical sketch of the clinical examination of urine* 1 online resource (276 pages illustrations, portraits). (Burroughs, Wellcome, London, 1911).

124. Hippocrates. *The aphorisms: with a translation into Latin, and English* (eds Coar, T. (& Howitt, W.) 1 online resource (viii, 314 pages). (Longman, London, England, 1822).

125. Crnogorac-Jurcevic, T. A brief history of urine examination - From ancient uroscopy to 21st century urinomics. en. *Adv. Biomark. Sci. Technol.* **6,** 83–90. doi:10.1016/j.abst.2024.06.001. (2024).

126. Diskin, C. J., Stokes, T. J., Dansby, L. M., Carter, T. B. & Radcliff, L. Surface tension, proteinuria, and the urine bubbles of Hippocrates. en. *Lancet* **355,** 901–902. doi:10.1016/S0140-6736(00)01097-7. (9207 11, 2000).

127. Eknoyan, G. & Nagy, J. A history of diabetes mellitus or how a disease of the kidneys evolved into a kidney disease. en. *Adv. Chronic Kidney Dis.* **12,** 223–229. doi:10.1053/j.ackd.2005.01.002. (2 2005).

128. MacFarlane, I. A. Mathew Dobson of Liverpool (1735–1784) and the history of diabetes. en. *Pract. Diabetes Int.* **7,** 246–248. doi:10.1002/pdi.1960070603. (6 1990).

129. Osler, W. *The principles and practice of medicine* (ed McCrae, T.) 1 online resource (xxiv, 1168 pages frontispiece (port.) illustrations, diagrams). (D. Appleton and company, London, 1920).

130. Büttner, J. The origin of clinical laboratories. en. *Eur. J. Clin. Chem. Clin. Biochem.* **30,** 585–593. doi:10.1515/cclm.1992.30.10.585. (10 1992).

131. Sunderman, F. Evolution of clinical science. An overview. *Ann. Clin. Lab. Sci.* **23,** 231–248. (4 1993).

132. Dawber, T. R., Moore, F. E. & Mann, G. V. Coronary heart disease in the Framingham study. en. *Am. J. Public Health Nations. Health* **47,** 4–24. doi:10.2105/ajph.47.4_pt_2.4. (4 Pt 2 1957).

133. The Expert Committee on the Diagnosis and Classification of Diabetes Mellitus. Report of the expert committee on the diagnosis and classification of diabetes mellitus. en. *Diabetes Care* **20,** 1183–1197. doi:10.2337/diacare.20.7.1183. (7 1997).

134. John, W. G., Mosca, A., Weykamp, C. & Goodall, I. HbA1c standardisation: history, science and politics. en. *Clin. Biochem. Rev.* **28,** 163–168. (4 2007).

135. Gabay, C. & Kushner, I. Acute-phase proteins and other systemic responses to inflammation. en. *N. Engl. J. Med.* **340,** 448–454. doi:10.1056/NEJM199902113400607. (6 11, 1999).

136. Gräsbeck, R. The evolution of the reference value concept. en. *Clin. Chem. Lab. Med.* **42,** 692–697. doi:10.1515/CCLM.2004.118. (7 2004).

137. Solberg, H. E. International Federation of Clinical Chemistry. Scientific committee, Clinical Section. Expert Panel on Theory of Reference Values and International Committee for Standardization in Haematology Standing Committee on Reference Values. Approved recommendation (1986) on the theory of reference values. Part 1. The concept of reference values. en. *Clin. Chim. Acta* **165,** 111–118. doi:10.1016/0009-8981(87)90224-5. (1 29, 1987).

138. Whyte, M. B. & Kelly, P. The normal range: it is not normal and it is not a range. en. *Postgrad. Med. J.* **94,** 613–616. doi:10.1136/postgradmedj-2018-135983. (1117 2018).

139. International Expert Committee. International Expert Committee report on the role of the A1C assay in the diagnosis of diabetes. en. *Diabetes Care* **32,** 1327–1334. doi:10.2337/dc09-9033. (7 2009).

140. Fuggle, S. *Clinical Biochemistry Reference Ranges Handbook* https://www.esht.nhs.uk/wp-content/uploads/2017/08/Clinical-Biochemistry-reference-ranges-handbook.pdf (2025).

141. North Bristol NHS Trust. *Age-related reference ranges: Biochemistry* https://www.nbt.nhs.uk/sites/default/files/document/Biochemistry%20Age%20Related%20Reference%20Ranges.pdf (2025).

142. UK Biobank. *Biomarker data* en. https://www.ukbiobank.ac.uk/enable-your-research/about-our-data/biomarker-data (2025).

143. Flynn, E. *et al.* Sex-specific genetic effects across biomarkers. en. *Eur. J. Hum. Genet.* **29,** 154–163. doi:10.1038/s41431-020-00712-w. (1 2021).

144. Nag, A. *et al.* Effects of protein-coding variants on blood metabolite measurements and clinical biomarkers in the UK Biobank. en. *Am. J. Hum. Genet.* **110,** 487–498. doi:10.1016/j.ajhg.2023.02.002. (3 2, 2023).

145. Smith, C. J. *et al.* Integrative analysis of metabolite GWAS illuminates the molecular basis of pleiotropy and genetic correlation. en. *Elife* **11.** doi:10.7554/eLife.79348. (8, 2022).

146. Mahalanobis, P. C. On the Generalised Distance in Statistics. en. *Proceedings of the National Institute of Sciences of India* **2,** 49–55. doi:10.1007/s13171-019-00164-5. (1936).

147. Boyd, J. C. & Lacher, D. A. The multivariate reference range: an alternative interpretation of multi-test profiles. en. *Clin. Chem.* **28,** 259–265. doi:10.1093/clinchem/28.2.259. (2 1982).

148. Solberg, H. E. Multivariate reference regions. en. *Scand. J. Clin. Lab. Invest. Suppl.* **222,** 3–5. doi:10.3109/00365519509088443. (1995).

149. Kernohan, K. D. & Boycott, K. M. The expanding diagnostic toolbox for rare genetic diseases. en. *Nat. Rev. Genet.* **25,** 401–415. doi:10.1038/s41576-023-00683-w. (6 2024).

150. Sun, Q. *et al.* Analyses of biomarker traits in diverse UK biobank participants identify associations missed by European-centric analysis strategies. en. *J. Hum. Genet.* **67,** 87–93. doi:10.1038/s10038-021-00968-0. (2 2022).

151. Fry, A. *et al.* Comparison of sociodemographic and health-related characteristics of UK Biobank participants with those of the general population. en. *Am. J. Epidemiol.* **186,** 1026–1034. doi:10.1093/aje/kwx246. (9 1, 2017).

152. Martin, S. S. *et al.* Lipid distributions in the Global Diagnostics Network across five continents. en. *Eur. Heart J.* **44,** 2305–2318. doi:10.1093/eurheartj/ehad371. (25 1, 2023).

153. Ambikairajah, A., Walsh, E. & Cherbuin, N. Lipid profile differences during menopause: a review with meta-analysis. en. *Menopause* **26,** 1327–1333. doi:10.1097/GME.0000000000001403. (11 2019).

154. Inaraja, V., Thuissard, I., Andreu-Vazquez, C. & Jodar, E. Lipid profile changes during the menopausal transition. en. *Menopause* **27,** 780–787. doi:10.1097/gme.0000000000001532. (7 2020).

155. Van Beresteijn, E. C. *et al.* Perimenopausal increase in serum cholesterol: a 10-year longitudinal study. en. *Am. J. Epidemiol.* **137,** 383–392. doi:10.1093/oxfordjournals.aje.a116686. (4 15, 1993).

156. South Tees Hospitals NHS Foundation Trust. *Creatinine* en. https://www.southtees.nhs.uk/services/pathology/tests/creatinine/ (2025).

157. Oterdoom, L. H. *et al.* Urinary creatinine excretion, an indirect measure of muscle mass, is an independent predictor of cardiovascular disease and mortality in the general population. en. *Atherosclerosis* **207,** 534–540. doi:10.1016/j.atherosclerosis.2009.05.010. (2 2009).

158. Heymsfield, S. B., Arteaga, C., McManus, C., Smith, J. & Moffitt, S. Measurement of muscle mass in humans: validity of the 24-hour urinary creatinine method. en. *Am. J. Clin. Nutr.* **37,** 478–494. doi:10.1093/ajcn/37.3.478. (3 1983).

159. Proctor, D. N., O'Brien, P. C., Atkinson, E. J. & Nair, K. S. Comparison of techniques to estimate total body skeletal muscle mass in people of different age groups. en. *Am. J. Physiol. Endocrinol. Metab.* **277,** E489–E495. doi:10.1152/ajpendo.1999.277.3.E489. (3 1, 1999).

160. Wang, Z. M., Gallagher, D., Nelson, M. E., Matthews, D. E. & Heymsfield, S. B. Total-body skeletal muscle mass: evaluation of 24-h urinary creatinine excretion by computerized axial tomography. en. *Am. J. Clin. Nutr.* **63,** 863–869. doi:10.1093/ajcn/63.6.863. (6 1996).

161. Holloszy, J. O. The biology of aging. en. *Mayo Clin. Proc.* **75,** S3–S9. doi:10.1016/s0025-6196(19)30634-2. (1 2000).

162. Metter, E. J., Conwit, R., Tobin, J. & Fozard, J. L. Age-associated loss of power and strength in the upper extremities in women and men. en. *J. Gerontol. A Biol. Sci. Med. Sci.* **52,** B267–76. doi:10.1093/gerona/52a.5.b267. (5 1997).

163. Gava, P., Kern, H. & Carraro, U. Age-associated power decline from running, jumping, and throwing male masters world records. en. *Exp. Aging Res.* **41,** 115–135. doi:10.1080/0361073X.2015.1001648. (2 2015).

164. Jones, G. & Barker, A. Reference intervals. en. *Clin. Biochem. Rev.* **29 Suppl 1,** S93–7. (2008).

165. Horn, P. S. & Pesce, A. J. Reference intervals: an update. en. *Clin. Chim. Acta* **334,** 5–23. doi:10.1016/s0009-8981(03)00133-5. (1-2 2003).

166. Greenfield, M. S., Kraemer, F., Tobey, T. & Reaven, G. Effect of age on plasma triglyceride concentrations in man. en. *Metabolism* **29,** 1095–1099. doi:10.1016/0026-0495(80)90221-8. (11 1980).

167. Barr, D. B. *et al.* Urinary creatinine concentrations in the U.S. population: implications for urinary biologic monitoring measurements. en. *Environ. Health Perspect.* **113,** 192–200. doi:10.1289/ehp.7337. (2 2005).

168. Forni Ogna, V. *et al.* New anthropometry-based age- and sex-specific reference values for urinary 24-hour creatinine excretion based on the adult Swiss population. en. *BMC Med.* **13,** 40. doi:10.1186/s12916-015-0275-x. (1 27, 2015).

169. Parhofer, K. G. & Laufs, U. The diagnosis and treatment of hypertriglyceridemia. en. *Dtsch. Arztebl. Int.* **116,** 825–832. doi:10.3238/arztebl.2019.0825. (49 6, 2019).

170. Amrein, K. *et al.* Vitamin D deficiency 2.0: an update on the current status worldwide. en. *Eur. J. Clin. Nutr.* **74,** 1498–1513. doi:10.1038/s41430-020-0558-y. (11 2020).

171. Khaw, K.-T. *et al.* Blood pressure and urinary sodium in men and women: the Norfolk Cohort of the European Prospective Investigation into Cancer (EPIC-Norfolk). en. *Am. J. Clin. Nutr.* **80,** 1397–1403. doi:10.1093/ajcn/80.5.1397. (5 2004).

172. Georgila, K., Vyrla, D. & Drakos, E. Apolipoprotein A-I (ApoA-I), immunity, inflammation and cancer. en. *Cancers (Basel)* **11,** 1097. doi:10.3390/cancers11081097. (8 1, 2019).

173. Busher, J. T. en. in *Clinical Methods: The History, Physical, and Laboratory Examinations* (Butterworths, Boston, 1990).

174. Xiao, L. *et al.* The LDL-C/ApoB ratio predicts cardiovascular and all-cause mortality in the general population. en. *Lipids Health Dis.* **22,** 104. doi:10.1186/s12944-023-01869-1. (1 21, 2023).

175. Ball, M., Miller, I. & Cotten, S. W. Direct bilirubin higher than total bilirubin? en. *Clin. Chem.* **61,** 889. doi:10.1373/clinchem.2014.237040. (6 2015).

176. McKillop, A. M. & Flatt, P. R. Emerging applications of metabolomic and genomic profiling in diabetic clinical medicine. en. *Diabetes Care* **34,** 2624–2630. doi:10.2337/dc11-0837. (12 2011).

177. Lopez-Giacoman, S. & Madero, M. Biomarkers in chronic kidney disease, from kidney function to kidney damage. en. *World J. Nephrol.* **4,** 57–73. doi:10.5527/wjn.v4.i1.57. (1 6, 2015).

178. Rosenstein, K. & Tannock, L. R. en. in *Endotext* (MDText.com, Inc., South Dartmouth (MA), 2000).

179. Chen, S. & Cheng, W. Relationship between lipid profiles and hypertension: A cross-sectional study of 62,957 Chinese adult males. en. *Front. Public Health* **10,** 895499. doi:10.3389/fpubh.2022.895499. (18, 2022).

180. Chruściel, P. *et al.* Associations between the lipid profile and the development of hypertension in young individuals - the preliminary study. en. *Arch. Med. Sci.* **18,** 25–35. doi:10.5114/aoms.2019.86197. (1 2022).

181. Osuji, C. U., Omejua, E. G., Onwubuya, E. I. & Ahaneku, G. I. Serum lipid profile of newly diagnosed hypertensive patients in nnewi, South-East Nigeria. en. *Int. J. Hypertens.* **2012,** 710486. doi:10.1155/2012/710486. (4, 2012).

182. Stark, Z. *et al.* Scaling national and international improvement in virtual gene panel curation via a collaborative approach to discordance resolution. en. *Am. J. Hum. Genet.* **108,** 1551–1557. doi:10.1016/j.ajhg.2021.06.020. (9 2, 2021).

183. Martin, A. R. *et al.* PanelApp crowdsources expert knowledge to establish consensus diagnostic gene panels. en. *Nature Genetics* **51,** 1560–1565. doi:10.1038/s41588-019-0528-2. (2024) (11 1, 2019).

184. Martin, J. *et al.* Examining sex-differentiated genetic effects across neuropsychiatric and behavioral traits. en. *Biol. Psychiatry* **89,** 1127–1137. doi:10.1016/j.biopsych.2020.12.024. (12 15, 2021).

185. Fadista, J., Manning, A. K., Florez, J. C. & Groop, L. The (in)famous GWAS P-value threshold revisited and updated for low-frequency variants. en. *Eur. J. Hum. Genet.* **24,** 1202–1205. doi:10.1038/ejhg.2015.269. (8 2016).

186. Kehar, M. *et al.* Fanconi-Bickel syndrome - mutation in SLC2A2 gene. en. *Indian J. Pediatr.* **81,** 1237–1239. doi:10.1007/s12098-014-1487-3. (11 2014).

187. Santer, R., Steinmann, B. & Schaub, J. Fanconi-Bickel syndrome–a congenital defect of facilitative glucose transport. en. *Curr. Mol. Med.* **2,** 213–227. doi:10.2174/1566524024605743. (2 2002).

188. Graham, S. E. *et al.* The power of genetic diversity in genome-wide association studies of lipids. en. *Nature* **600,** 675–679. doi:10.1038/s41586-021-04064-3. (7890 2021).

189. Verma, A. *et al.* Diversity and scale: Genetic architecture of 2068 traits in the VA Million Veteran Program. en. *Science* **385,** eadj1182. doi:10.1126/science.adj1182. (6706 19, 2024).

190. Willems, S. M. *et al.* Large-scale exome array summary statistics resources for glycemic traits to aid effector gene prioritization. en. *Wellcome Open Res.* **8,** 483. doi:10.12688/wellcomeopenres.18754.1. (20, 2023).

191. Dehghan, A., van Hoek, M., Sijbrands, E. J. G., Hofman, A. & Witteman, J. C. M. High serum uric acid as a novel risk factor for type 2 diabetes. en. *Diabetes Care* **31,** 361–362. doi:10.2337/dc07-1276. (2 2008).

192. Bhole, V., Choi, J. W. J., Kim, S. W., de Vera, M. & Choi, H. Serum uric acid levels and the risk of type 2 diabetes: a prospective study. en. *Am. J. Med.* **123,** 957–961. doi:10.1016/j.amjmed.2010.03.027. (10 2010).

193. Niskanen, L. *et al.* Serum uric acid as a harbinger of metabolic outcome in subjects with impaired glucose tolerance: the Finnish Diabetes Prevention Study. en. *Diabetes Care* **29,** 709–711. doi:10.2337/diacare.29.03.06.dc05-1465. (3 2006).

194. Crook, M. A., Tutt, P., Simpson, H. & Pickup, J. C. Serum sialic acid and acute phase proteins in type 1 and type 2 diabetes mellitus. en. *Clin. Chim. Acta* **219,** 131–138. doi:10.1016/0009-8981(93)90204-h. (1-2 15, 1993).

195. Rodríguez-Morán, M. & Guerrero-Romero, F. Increased levels of C-reactive protein in noncontrolled type II diabetic subjects. en. *J. Diabetes Complications* **13,** 211–215. doi:10.1016/s1056-8727(99)00047-1. (4 1999).

196. Thorand, B. *et al.* C-reactive protein as a predictor for incident diabetes mellitus among middle-aged men: results from the MONICA Augsburg cohort study, 1984-1998. en. *Arch. Intern. Med.* **163,** 93–99. doi:10.1001/archinte.163.1.93. (1 13, 2003).

197. Cermenati, S. *et al.* Sox18 and Sox7 play redundant roles in vascular development. en. *Blood* **111,** 2657–2666. doi:10.1182/blood-2007-07-100412. (5 1, 2008).

198. Irrthum, A. *et al.* Mutations in the transcription factor gene SOX18 underlie recessive and dominant forms of hypotrichosis-lymphedema-telangiectasia. en. *Am. J. Hum. Genet.* **72,** 1470–1478. doi:10.1086/375614. (6 2003).

199. Karczewski, K. J. *et al.* Systematic single-variant and gene-based association testing of thousands of phenotypes in 394,841 UK Biobank exomes. en. *Cell Genom.* **2,** 100168. doi:10.1016/j.xgen.2022.100168. (9 14, 2022).

200. Wang, Q. *et al.* Rare variant contribution to human disease in 281,104 UK Biobank exomes. en. *Nature* **597,** 527–532. doi:10.1038/s41586-021-03855-y. (7877 2021).

201. Peng, P. *et al.* Decreased expression of Calpain-9 predicts unfavorable prognosis in patients with gastric cancer. en. *Sci. Rep.* **6,** 29604. doi:10.1038/srep29604. (1 12, 2016).

202. Westerman, K. E. *et al.* Variance-quantitative trait loci enable systematic discovery of gene-environment interactions for cardiometabolic serum biomarkers. en. *Nat. Commun.* **13,** 3993. doi:10.1038/s41467-022-31625-5. (1 9, 2022).

203. Morales, J. *et al.* A joint NCBI and EMBL-EBI transcript set for clinical genomics and research. en. *Nature* **604,** 310–315. doi:10.1038/s41586-022-04558-8. (7905 2022).

204. Abul-Husn, N. S. *et al.* A protein-truncating HSD17B13 variant and protection from chronic liver disease. en. *N. Engl. J. Med.* **378,** 1096–1106. doi:10.1056/NEJMoa1712191. (12 22, 2018).

205. Sookoian, S. & Pirola, C. J. Meta-analysis of the influence of I148M variant of patatin-like phospholipase domain containing 3 gene (PNPLA3) on the susceptibility and histological severity of nonalcoholic fatty liver disease. en. *Hepatology* **53,** 1883–1894. doi:10.1002/hep.24283. (6 2011).

206. Romeo, S. *et al.* Genetic variation in PNPLA3 confers susceptibility to nonalcoholic fatty liver disease. en. *Nat. Genet.* **40,** 1461–1465. doi:10.1038/ng.257. (12 2008).

207. Flores, Y. N. *et al.* Serum lipids are associated with nonalcoholic fatty liver disease: a pilot case-control study in Mexico. en. *Lipids Health Dis.* **20,** 136. doi:10.1186/s12944-021-01526-5. (1 10, 2021).

208. He, S. *et al.* A sequence variation (I148M) in PNPLA3 associated with nonalcoholic fatty liver disease disrupts triglyceride hydrolysis. en. *J. Biol. Chem.* **285,** 6706–6715. doi:10.1074/jbc.M109.064501. (9 26, 2010).

209. Johnson, S. M. *et al.* PNPLA3 is a triglyceride lipase that mobilizes polyunsaturated fatty acids to facilitate hepatic secretion of large-sized very low-density lipoprotein. en. *Nat. Commun.* **15,** 4847. doi:10.1038/s41467-024-49224-x. (1 6, 2024).

210. Krarup, N. T. *et al.* The PNPLA3 rs738409 G-allele associates with reduced fasting serum triglyceride and serum cholesterol in Danes with impaired glucose regulation. en. *PLoS One* **7,** e40376. doi:10.1371/journal.pone.0040376. (7 5, 2012).

211. Zain, S. M. *et al.* A multi-ethnic study of a PNPLA3 gene variant and its association with disease severity in non-alcoholic fatty liver disease. en. *Hum. Genet.* **131,** 1145–1152. doi:10.1007/s00439-012-1141-y. (7 2012).

212. Hotta, K. *et al.* Association of the rs738409 polymorphism in PNPLA3 with liver damage and the development of nonalcoholic fatty liver disease. en. *BMC Med. Genet.* **11,** 172. doi:10.1186/1471-2350-11-172. (22, 2010).

213. Speliotes, E. K. *et al.* PNPLA3 variants specifically confer increased risk for histologic nonalcoholic fatty liver disease but not metabolic disease. en. *Hepatology* **52,** 904–912. doi:10.1002/hep.23768. (3 2010).

214. Kathiresan, S. *et al.* Common variants at 30 loci contribute to polygenic dyslipidemia. en. *Nat. Genet.* **41,** 56–65. doi:10.1038/ng.291. (1 2009).

215. Lala, V., Zubair, M. & Minter, D. A. en. in *StatPearls* (StatPearls Publishing, Treasure Island (FL), 2025).

216. Decaux, G. & Musch, W. Clinical laboratory evaluation of the syndrome of inappropriate secretion of antidiuretic hormone. en. *Clin. J. Am. Soc. Nephrol.* **3,** 1175–1184. doi:10.2215/CJN.04431007. (4 2008).

217. Talmud, P. J., Hawe, E., Miller, G. J. & Humphries, S. E. Nonfasting apolipoprotein B and triglyceride levels as a useful predictor of coronary heart disease risk in middle-aged UK men. en. *Arterioscler. Thromb. Vasc. Biol.* **22,** 1918–1923. doi:10.1161/01.atv.0000035521.22199.c7. (11 1, 2002).

218. Yang, J., Zhang, W., Albert, P. S., Liu, A. & Chen, Z. Combining biomarkers to improve diagnostic accuracy in detecting diseases with group-tested data. en. *Stat. Med.* **43,** 5182–5192. doi:10.1002/sim.10230. (27 30, 2024).

219. Nickolas, T. L. *et al.* Diagnostic and prognostic stratification in the emergency department using urinary biomarkers of nephron damage: a multicenter prospective cohort study. en. *J. Am. Coll. Cardiol.* **59,** 246–255. doi:10.1016/j.jacc.2011.10.854. (3 17, 2012).

220. Parikh, C. R. *et al.* Postoperative biomarkers predict acute kidney injury and poor outcomes after adult cardiac surgery. en. *J. Am. Soc. Nephrol.* **22,** 1748–1757. doi:10.1681/ASN.2010121302. (9 2011).

221. Meisner, A., Parikh, C. R. & Kerr, K. F. Biomarker combinations for diagnosis and prognosis in multicenter studies: Principles and methods. en. *Stat. Methods Med. Res.* **28,** 969–985. doi:10.1177/0962280217740392. (4 2019).

222. Zhou, W. *et al.* SAIGE-GENE+ improves the efficiency and accuracy of set-based rare variant association tests. en. *Nat. Genet.* **54,** 1466–1469. doi:10.1038/s41588-022-01178-w. (10 2022).

223. Liu, Y. & Xie, J. Cauchy combination test: a powerful test with analytic p-value calculation under arbitrary dependency structures. en. *J. Am. Stat. Assoc.* **115,** 393–402. doi:10.1080/01621459.2018.1554485. (529 2020).

224. Jeon, B.-N. *et al.* ZBTB2, a novel master regulator of the p53 pathway. en. *J. Biol. Chem.* **284,** 17935–17946. doi:10.1074/jbc.M809559200. (27 3, 2009).

225. Lahalle, A. *et al.* The p53 pathway and metabolism: The tree that hides the Forest. en. *Cancers (Basel)* **13,** 133. doi:10.3390/cancers13010133. (1 4, 2021).

226. Harris, S. L. & Levine, A. J. The p53 pathway: positive and negative feedback loops. en. *Oncogene* **24,** 2899–2908. doi:10.1038/sj.onc.1208615. (17 18, 2005).

227. Devaraj, S., Semaan, J. R. & Jialal, I. en. in *StatPearls* (StatPearls Publishing, Treasure Island (FL), 2025).

228. Behbodikhah, J. *et al.* Apolipoprotein B and cardiovascular disease: Biomarker and potential therapeutic target. en. *Metabolites* **11,** 690. doi:10.3390/metabo11100690. (10 8, 2021).

229. Mach, F. *et al.* 2019 ESC/EAS Guidelines for the management of dyslipidaemias: lipid modification to reduce cardiovascular risk. en. *Eur. Heart J.* **41,** 111–188. doi:10.1093/eurheartj/ehz455. (1 1, 2020).

230. Lin, J.-P. *et al.* Association between the UGT1A1*28 allele, bilirubin levels, and coronary heart disease in the Framingham Heart Study. en. *Circulation* **114,** 1476–1481. doi:10.1161/CIRCULATIONAHA.106.633206. (14 3, 2006).

231. Karjalainen, M. K. *et al.* Apolipoprotein A-I concentrations and risk of coronary artery disease: A Mendelian randomization study. en. *Atherosclerosis* **299,** 56–63. doi:10.1016/j.atherosclerosis.2020.02.002. (2020).

232. Gordon, D. J. *et al.* High-density lipoprotein cholesterol and cardiovascular disease. Four prospective American studies. en. *Circulation* **79,** 8–15. doi:10.1161/01.cir.79.1.8. (1 1989).

233. Danik, J. S. & Manson, J. E. Vitamin d and cardiovascular disease. en. *Curr. Treat. Options Cardiovasc. Med.* **14,** 414–424. doi:10.1007/s11936-012-0183-8. (4 2012).

234. Siddiqui, M. S. *et al.* Association between high-normal levels of alanine aminotransferase and risk factors for atherogenesis. en. *Gastroenterology* **145,** 1271–9.e1–3. doi:10.1053/j.gastro.2013.08.036. (6 2013).

235. Kunutsor, S. K., Apekey, T. A. & Khan, H. Liver enzymes and risk of cardiovascular disease in the general population: a meta-analysis of prospective cohort studies. en. *Atherosclerosis* **236,** 7–17. doi:10.1016/j.atherosclerosis.2014.06.006. (1 2014).

236. Ndrepepa, G. Aspartate aminotransferase and cardiovascular disease—a narrative review. *J. Lab. Precis. Med.* **6,** 6–6. doi:10.21037/jlpm-20-93. (0 2021).

237. Kunutsor, S. K. *et al.* Serum alkaline phosphatase and risk of incident cardiovascular disease: Interrelationship with high sensitivity C-reactive protein. en. *PLoS One* **10,** e0132822. doi:10.1371/journal.pone.0132822. (7 13, 2015).

238. Li, J.-W., Xu, C., Fan, Y., Wang, Y. & Xiao, Y.-B. Can serum levels of alkaline phosphatase and phosphate predict cardiovascular diseases and total mortality in individuals with preserved renal function? A systemic review and meta-analysis. en. *PLoS One* **9,** e102276. doi:10.1371/journal.pone.0102276. (7 17, 2014).

239. Iqbal, M. P., Mehboobali, N., Azam, I. & Tareen, A. K. Association of alkaline phosphatase with acute myocardial infarction in a population with high prevalence of hypovitaminosis D. en. *Clin. Chim. Acta* **425,** 192–195. doi:10.1016/j.cca.2013.08.009. (21, 2013).

240. Rahmani, J. *et al.* Elevated liver enzymes and cardiovascular mortality: a systematic review and dose-response meta-analysis of more than one million participants. en. *Eur. J. Gastroenterol. Hepatol.* **31,** 555–562. doi:10.1097/MEG.0000000000001353. (5 2019).

241. Adamidis, P. S. *et al.* Association of alkaline phosphatase with cardiovascular disease in patients with dyslipidemia: A 6-year retrospective study. en. *J. Cardiovasc. Dev. Dis.* **11,** 60. doi:10.3390/jcdd11020060. (2 15, 2024).

242. Wannamethee, S. G., Shaper, A. G. & Perry, I. J. Serum creatinine concentration and risk of cardiovascular disease: a possible marker for increased risk of stroke: A possible marker for increased risk of stroke. en. *Stroke* **28,** 557–563. doi:10.1161/01.str.28.3.557. (3 1997).

243. North Bristol NHS Foundation Trust. *Apolipoprotein B* en. https://www.nbt.nhs.uk/severn-pathology/requesting/test-information/apolipoprotein-b (2025).

244. Gloucestershire Hospitals NHS Foundation Trust. *Tests and investigations* en. https://www.gloshospitals.nhs.uk/our-services/services-we-offer/pathology/tests-and-investigations/ (2025).

245. South Tees Hospitals NHS Foundation Trust. *Pathology tests* en. https://www.southtees.nhs.uk/services/pathology/tests/ (2025).

246. NHS. *High cholesterol - Cholesterol levels* en. https://www.nhs.uk/conditions/high-cholesterol/cholesterol-levels/ (2025).

247. Information Services Department & Royal United Hospital. *Royal United Hospital Bath NHS Trust* en. https://www.ruh.nhs.uk/pathology/blood_test_info/tubelisting.asp?id=228&listing=a?id=228&listing=a (2025).

248. Szustakowski, J. D. *et al.* Advancing human genetics research and drug discovery through exome sequencing of the UK Biobank. en. *Nat. Genet.* **53,** 942–948. doi:10.1038/s41588-021-00885-0. (7 2021).

249. Poplin, R. *et al.* A universal SNP and small-indel variant caller using deep neural networks. en. *Nat. Biotechnol.* **36,** 983–987. doi:10.1038/nbt.4235. (10 2018).

250. Lin, M. F. *et al.* GLnexus: joint variant calling for large cohort sequencing. *Bioinformatics.* (11, 2018).

251. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. en. *Gigascience* **10.** doi:10.1093/gigascience/giab008. (2 16, 2021).

252. (2025). hail: Cloud-native genomic dataframes and batch computing. URL: https://github.com/hail-is/hail.

253. Li, H. Toward better understanding of artifacts in variant calling from high-coverage samples. en. *Bioinformatics* **30,** 2843–2851. doi:10.1093/bioinformatics/btu356. (20 15, 2014).

254. Van der Auwera, G. A. & O'Connor, B. D. *Genomics in the Cloud: Using Docker, GATK, and WDL in Terra* en. 496 pp. ("O'Reilly Media, Inc.", 2, 2020).

255. Lassen, F. H. *et al.* Exome-wide evidence of compound heterozygous effects across common phenotypes in the UK Biobank. en. *Cell Genom.* **4,** 100602. doi:10.1016/j.xgen.2024.100602. (7 10, 2024).

256. Fisher, R. A. The Logic of Inductive Inference. en. *Journal of the Royal Statistical Society* **98,** 39–82. doi:https://doi.org/10.2307/2342435. (1935).

257. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. en. *Am. J. Hum. Genet.* **81,** 559–575. doi:10.1086/519795. (3 2007).

258. Hardy, G. H. Mendelian proportions in a mixed population. en. *Science* **28,** 49–50. doi:10.1126/science.28.706.49. (706 10, 1908).

259. Spaan, S. *et al.* Variability of biomarkers in volunteer studies: The biological component. en. *Toxicol. Lett.* **198,** 144–151. doi:10.1016/j.toxlet.2010.06.008. (2 5, 2010).

260. Aylward, L. L. *et al.* Sources of variability in biomarker concentrations. en. *J. Toxicol. Environ. Health B Crit. Rev.* **17,** 45–61. doi:10.1080/10937404.2013.864250. (1 2014).

261. North Bristol NHS Trust. *Vitamin D* en. https://www.nbt.nhs.uk/severn-pathology/requesting/test-information/vitamin-d (2025).

262. Tahmasebi, H., Trajcevski, K., Higgins, V. & Adeli, K. Influence of ethnicity on population reference values for biochemical markers. en. *Crit. Rev. Clin. Lab. Sci.* **55,** 359–375. doi:10.1080/10408363.2018.1476455. (5 2018).

263. Hirsch, J. A. *et al.* ICD-10: History and context. en. *AJNR Am. J. Neuroradiol.* **37,** 596–599. doi:10.3174/ajnr.A4696. (4 2016).

264. Amberger, J. S., Bocchini, C. A., Scott, A. F. & Hamosh, A. OMIM.org: leveraging knowledge across phenotype-gene relationships. en. *Nucleic Acids Res.* **47,** D1038–D1043. doi:10.1093/nar/gky1151. (D1 8, 2019).

265. Landrum, M. J. *et al.* ClinVar: improving access to variant interpretations and supporting evidence. en. *Nucleic Acids Res.* **46,** D1062–D1067. doi:10.1093/nar/gkx1153. (D1 4, 2018).

266. Fedorov, V., Mannino, F. & Zhang, R. Consequences of dichotomization. *Pharmaceutical Statistics* **8,** 50–61. doi:10.1002/pst.331 (2008).

267. *Glycated haemoglobin (HbA1c) for the diagnosis of diabetes* en. (2025) (World Health Organization, 2011).

268. Falconer, D. S. The inheritance of liability to certain diseases, estimated from the incidence among relatives. en. *Ann. Hum. Genet.* **29,** 51–76. doi:10.1111/j.1469-1809.1965.tb00500.x. (1 1965).

269. Zuk, O., Hechter, E., Sunyaev, S. R. & Lander, E. S. The mystery of missing heritability: Genetic interactions create phantom heritability. en. *Proc. Natl. Acad. Sci. U. S. A.* **109,** 1193–1198. doi:10.1073/pnas.1119675109. (4 24, 2012).

270. Gibson, G. Rare and common variants: twenty arguments. en. *Nat. Rev. Genet.* **13,** 135–145. doi:10.1038/nrg3118. (2 18, 2012).

271. Ma, Y. & Zhou, X. Genetic prediction of complex traits with polygenic scores: a statistical review. en. *Trends Genet.* **37,** 995–1011. doi:10.1016/j.tig.2021.06.004. (11 2021).

272. Khera, A. V. *et al.* Genetic risk, adherence to a healthy lifestyle, and coronary disease. en. *N. Engl. J. Med.* **375,** 2349–2358. doi:10.1056/NEJMoa1605086. (24 15, 2016).

273. Maas, P. *et al.* Breast cancer risk from modifiable and nonmodifiable risk factors among white women in the United States. en. *JAMA Oncol.* **2,** 1295–1302. doi:10.1001/jamaoncol.2016.1025. (10 1, 2016).

274. Natarajan, P. *et al.* Polygenic risk score identifies subgroup with higher burden of atherosclerosis and greater relative benefit from statin therapy in the primary prevention setting. en. *Circulation* **135,** 2091–2101. doi:10.1161/CIRCULATIONAHA.116.024436. (22 30, 2017).

275. Torkamani, A., Wineinger, N. E. & Topol, E. J. The personal and clinical utility of polygenic risk scores. en. *Nat. Rev. Genet.* **19,** 581–590. doi:10.1038/s41576-018-0018-x. (9 2018).

276. Jurgens, S. J. *et al.* Adjusting for common variant polygenic scores improves yield in rare variant association analyses. *Nature genetics* **55.** doi:10.1038/s41588-023-01342-w. (2024) (4 2023).

277. Lu, T. *et al.* Individuals with common diseases but with a low polygenic risk score could be prioritized for rare variant screening. en. *Genet. Med.* **23,** 508–515. doi:10.1038/s41436-020-01007-7. (3 2021).

278. Huang, Q. Q. *et al.* Examining the role of common variants in rare neurodevelopmental conditions. en. *Nature* **636,** 404–411. doi:10.1038/s41586-024-08217-y. (8042 2024).

279. Hawkes, G. *et al.* Identification and analysis of individuals who deviate from their genetically-predicted phenotype. en. *PLoS Genet* **19,** e1010934. doi:10.1371/journal.pgen.1010934. (9 2023).

280. Fahed, A. C. *et al.* Polygenic background modifies penetrance of monogenic variants for tier 1 genomic conditions. en. *Nature Communications* **11,** 1–9. doi:10.1038/s41467-020-17374-3. (2024) (1 20, 2020).

281. Kuchenbaecker, K. B. *et al.* Evaluation of Polygenic Risk Scores for Breast and Ovarian Cancer Risk Prediction in BRCA1 and BRCA2 Mutation Carriers. en. *J Natl Cancer Inst* **109,** djw302. doi:10.1093/jnci/djw302. (2024) (7 9, 2017).

282. Mars, N. *et al.* The role of polygenic risk and susceptibility genes in breast cancer over the course of life. en. *Nature Communications* **11,** 1–9. doi:10.1038/s41467-020-19966-5. (2024) (1 14, 2020).

283. Identification of Genetic Factors that Modify Clinical Onset of Huntington's Disease. *Cell* **162,** 516–526. doi:10.1016/j.cell.2015.07.003. (2024) (3 30, 2015).

284. Moss, D. J. H. *et al.* Identification of genetic variants associated with Huntington's disease progression: a genome-wide association study. en. *Lancet Neurol.* **16,** 701–711. doi:10.1016/S1474-4422(17)30161-8. (9 2017).

285. Craig, J. E. *et al.* Multitrait analysis of glaucoma identifies new risk loci and enables polygenic prediction of disease susceptibility and progression. en. *Nature Genetics* **52,** 160–166. doi:10.1038/s41588-019-0556-y. (2024) (2 20, 2020).

286. Zebardast, N. *et al.* Characteristics of p.Gln368Ter myocilin variant and influence of polygenic risk on glaucoma penetrance in the UK Biobank. en. *Ophthalmology* **128,** 1300–1311. doi:10.1016/j.ophtha.2021.03.007. (9 2021).

287. Oetjens, M. T., Kelly, M. A., Sturm, A. C., Martin, C. L. & Ledbetter, D. H. Quantifying the polygenic contribution to variable expressivity in eleven rare genetic disorders. en. *Nature Communications* **10,** 1–10. doi:10.1038/s41467-019-12869-0. (2024) (1 25, 2019).

288. Kingdom, R., Beaumont, R. N., Wood, A. R., Weedon, M. N. & Wright, C. F. Genetic modifiers of rare variants in monogenic developmental disorder loci. en. *Nature Genetics* **56,** 861–868. doi:10.1038/s41588-024-01710-0. (2024) (5 18, 2024).

289. Niemi, M. E. K. *et al.* Common genetic variants contribute to risk of rare severe neurodevelopmental disorders. en. *Nature* **562,** 268–271. doi:10.1038/s41586-018-0566-4. (2024) (7726 26, 2018).

290. Khan, A. *et al.* Polygenic risk alters the penetrance of monogenic kidney disease. en. *Nature Communications* **14,** 1–10. doi:10.1038/s41467-023-43878-9. (2024) (1 14, 2023).

291. Tansey, K. E. *et al.* Common alleles contribute to schizophrenia in CNV carriers. en. *Molecular Psychiatry* **21,** 1085–1089. doi:10.1038/mp.2015.143. (2024) (8 22, 2015).

292. Thompson, D. J. *et al.* A systematic evaluation of the performance and properties of the UK Biobank Polygenic Risk Score (PRS) Release. *PLOS ONE* **19,** e0307270. doi:10.1371/journal.pone.0307270. (2024) (9 18, 2024).

293. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Series B Stat. Methodol.* **57,** 289–300. (1 1995).

294. Bonnefond, A. *et al.* Monogenic diabetes. en. *Nat. Rev. Dis. Primers* **9,** 12. doi:10.1038/s41572-023-00421-w. (1 9, 2023).

295. Flannick, J. *et al.* Loss-of-function mutations in SLC30A8 protect against type 2 diabetes. en. *Nat. Genet.* **46,** 357–363. doi:10.1038/ng.2915. (4 2014).

296. Cohen, J. C., Boerwinkle, E., Mosley Jr, T. H. & Hobbs, H. H. Sequence variations in PCSK9, low LDL, and protection against coronary heart disease. en. *N Engl J Med* **354,** 1264–1272. doi:10.1056/NEJMoa054013. (12 23, 2006).

297. Ference, B. A. *et al.* Variation in PCSK9 and HMGCR and risk of cardiovascular disease and diabetes. en. *N. Engl. J. Med.* **375,** 2144–2153. doi:10.1056/NEJMoa1604304. (22 1, 2016).

298. Peloso, G. M. *et al.* Rare Protein-Truncating Variants in APOB, Lower Low-Density Lipoprotein Cholesterol, and Protection Against Coronary Heart Disease. en. *Circulation: Genomic and Precision Medicine.* doi:10.1161/CIRCGEN.118.002376. (2025) (2019).

299. Soutar, A. K. & Naoumova, R. P. Mechanisms of Disease: genetic causes of familial hypercholesterolemia. en. *Nature Clinical Practice Cardiovascular Medicine* **4,** 214–225. doi:10.1038/ncpcardio0836. (2025) (4 2007).

300. Stitziel, N. O. *et al.* ANGPTL3 deficiency and protection against coronary artery disease. en. *J. Am. Coll. Cardiol.* **69,** 2054–2063. doi:10.1016/j.jacc.2017.02.030. (16 25, 2017).

301. Rauch, F. & Glorieux, F. H. Osteogenesis imperfecta. en. *Lancet* **363,** 1377–1385. doi:10.1016/S0140-6736(04)16051-0. (9418 24, 2004).

302. Genovese, G. *et al.* Increased burden of ultra-rare protein-altering variants among 4,877 individuals with schizophrenia. en. *Nature Neuroscience* **19,** 1433–1441. doi:10.1038/nn.4402. (2025) (11 3, 2016).

303. Conboy, J., Mohandas, N., Tchernia, G. & Kan, Y. W. Molecular basis of hereditary elliptocytosis due to protein 4.1 deficiency. *The New England journal of medicine* **315.** doi:10.1056/NEJM198609113151105. (2025) (11 11, 1986).

304. Eber, S. & Lux, S. E. Hereditary spherocytosis–defects in proteins that connect the membrane skeleton to the lipid bilayer. *Seminars in hematology* **41.** doi:10.1053/j.seminhematol.2004.01.002. (2025) (2 2004).

305. Karamatic, C. V. *et al.* Missense mutations in PIEZO1, which encodes the Piezo1 mechanosensor protein, define Er red blood cell antigens. *Blood* **141.** doi:10.1182/blood.2022016504. (2025) (2 12, 2023).

306. Rushakoff, R. J., MacMaster, H. W. & Shah, A. D. Hereditary spherocytosis and other factors that alter HBA1C levels. en. *AACE Clin. Case Rep.* **1,** e212–e213. doi:10.4158/ep14508.cor. (3 2015).

307. Gallagher, C. S. *et al.* Genome-wide association and epidemiological analyses reveal common genetic origins between uterine leiomyomata and endometriosis. en. *Nat. Commun.* **10,** 4857. doi:10.1038/s41467-019-12536-4. (1 24, 2019).

308. Cybulski, C. *et al.* CHEK2 is a multiorgan cancer susceptibility gene. en. *Am. J. Hum. Genet.* **75,** 1131–1135. doi:10.1086/426403. (6 2004).

309. Ruth, K. S. *et al.* Genetic insights into biological mechanisms governing human ovarian ageing. en. *Nature* **596,** 393–397. doi:10.1038/s41586-021-03779-7. (7872 2021).

310. Ulin, M., Ali, M., Chaudhry, Z. T., Al-Hendy, A. & Yang, Q. Uterine Fibroids in Menopause and Perimenopause. *Menopause (New York, N.Y.)* **27,** 238. doi:10.1097/GME.0000000000001438. (2025) (2 2020).

311. Hou, J. *et al.* Polygenic resilience scores capture protective genetic effects for Alzheimer's disease. en. *Transl. Psychiatry* **12,** 296. doi:10.1038/s41398-022-02055-0. (1 25, 2022).

312. Chen, R. *et al.* Analysis of 589,306 genomes identifies individuals resilient to severe Mendelian childhood diseases. en. *Nat. Biotechnol.* **34,** 531–538. doi:10.1038/nbt.3514. (5 2016).

313. Ochoa, D. *et al.* The next-generation Open Targets Platform: reimagined, redesigned, rebuilt. en. *Nucleic Acids Res.* **51,** D1353–D1359. doi:10.1093/nar/gkac1046. (D1 6, 2023).

314. Bonàs-Guarch, S. *et al.* Re-analysis of public genetic data reveals a rare X-chromosomal variant associated with type 2 diabetes. en. *Nat. Commun.* **9,** 321. doi:10.1038/s41467-017-02380-9. (1 22, 2018).

315. Xue, A. *et al.* Genome-wide association analyses identify 143 risk variants and putative regulatory mechanisms for type 2 diabetes. en. *Nat. Commun.* **9,** 2941. doi:10.1038/s41467-018-04951-w. (1 27, 2018).

316. *UKBB Genetic Correlation* https://ukbb-rg.hail.is/rg_browser/ (2024).

317. Klein, S., Gastaldelli, A., Yki-Järvinen, H. & Scherer, P. E. Why does obesity cause diabetes? en. *Cell Metab.* **34,** 11–20. doi:10.1016/j.cmet.2021.12.012. (1 4, 2022).

318. Zhu, F. *et al.* Biallelic SUN5 mutations cause autosomal-recessive acephalic spermatozoa syndrome. en. *Am. J. Hum. Genet.* **99,** 942–949. doi:10.1016/j.ajhg.2016.08.004. (4 6, 2016).

319. Graziani, A., Scafa, R., Grande, G. & Ferlin, A. Diabetes and male fertility disorders. en. *Mol. Aspects Med.* **99,** 101303. doi:10.1016/j.mam.2024.101303. (101303 2024).

320. Karlsson Linnér, R. *et al.* Multivariate analysis of 1.5 million people identifies genetic associations with traits related to self-regulation and addiction. en. *Nat. Neurosci.* **24,** 1367–1376. doi:10.1038/s41593-021-00908-3. (10 2021).

321. Li, L. *et al.* Attention-deficit/hyperactivity disorder is associated with increased risk of cardiovascular diseases: A systematic review and meta-analysis. en. *JCPP Adv.* **3,** e12158. doi:10.1002/jcv2.12158. (3 2023).

322. Chen, G. *et al.* Genome-wide association study identifies novel loci association with fasting insulin and insulin resistance in African Americans. en. *Hum. Mol. Genet.* **21,** 4530–4536. doi:10.1093/hmg/dds282. (20 15, 2012).

323. Ouchi, N. *et al.* Novel modulator for endothelial adhesion molecules: adipocyte-derived plasma protein adiponectin: Adipocyte-derived plasma protein adiponectin. en. *Circulation* **100,** 2473–2476. doi:10.1161/01.cir.100.25.2473. (25 1999).

324. Gillum, R. F., Mussolino, M. E. & Madans, J. H. Counts of neutrophils, lymphocytes, and monocytes, cause-specific mortality and coronary heart disease: the NHANES-I epidemiologic follow-up study. en. *Ann. Epidemiol.* **15,** 266–271. doi:10.1016/j.annepidem.2004.08.009. (4 2005).

325. Luo, J., Thomassen, J. Q., Nordestgaard, B. G., Tybjærg-Hansen, A. & Frikke-Schmidt, R. Neutrophil counts and cardiovascular disease. en. *Eur. Heart J.* **44,** 4953–4964. doi:10.1093/eurheartj/ehad649. (47 14, 2023).

326. Wheeler, J. G., Mussolino, M. E., Gillum, R. F. & Danesh, J. Associations between differential leucocyte count and incident coronary heart disease: 1764 incident cases from seven prospective studies of 30,374 individuals. en. *Eur. Heart J.* **25,** 1287–1292. doi:10.1016/j.ehj.2004.05.002. (15 2004).

327. Kemp, J. P. *et al.* Identification of 153 new loci associated with heel bone mineral density and functional involvement of GPC6 in osteoporosis. en. *Nat. Genet.* **49,** 1468–1475. doi:10.1038/ng.3949. (10 2017).

328. *UK Biobank Neale Lab GWAS Round 2* http://www.nealelab.is/uk-biobank/ (2025).

329. Bauer, D. C. *et al.* Broadband ultrasound attenuation predicts fractures strongly and independently of densitometry in older women. A prospective study. Study of Osteoporotic Fractures Research Group. en. *Arch. Intern. Med.* **157,** 629–634. doi:10.1001/archinte.157.6.629. (6 24, 1997).

330. Welch, A. *et al.* Broadband ultrasound attenuation (BUA) of the heel bone and its correlates in men and women in the EPIC-Norfolk cohort: a cross-sectional population-based study. en. *Osteoporos. Int.* **15,** 217–225. doi:10.1007/s00198-003-1410-7. (3 2004).

331. Chiou, J. *et al.* Interpreting type 1 diabetes risk with genetics and single-cell epigenomics. en. *Nature* **594,** 398–402. doi:10.1038/s41586-021-03552-w. (7863 2021).

332. Inshaw, J. R. J. *et al.* Analysis of overlapping genetic association in type 1 and type 2 diabetes. en. *Diabetologia* **64,** 1342–1347. doi:10.1007/s00125-021-05428-0. (6 2021).

333. Onengut-Gumuscu, S. *et al.* Fine mapping of type 1 diabetes susceptibility loci and evidence for colocalization of causal variants with lymphoid gene enhancers. en. *Nat. Genet.* **47,** 381–386. doi:10.1038/ng.3245. (4 2015).

334. Napoli, N. *et al.* Mechanisms of diabetes mellitus-induced bone fragility. en. *Nat. Rev. Endocrinol.* **13,** 208–219. doi:10.1038/nrendo.2016.153. (4 2017).

335. Ehrlich, J. R., Radcliffe, N. M. & Shimmyo, M. Goldmann applanation tonometry compared with corneal-compensated intraocular pressure in the evaluation of primary open-angle Glaucoma. *BMC ophthalmology* **12.** doi:10.1186/1471-2415-12-52. (2025) (25, 2012).

336. Loos, R. J. F. & Yeo, G. S. H. The genetics of obesity: from discovery to biology. en. *Nat. Rev. Genet.* **23,** 120–133. doi:10.1038/s41576-021-00414-z. (2 2022).

337. Mäkitie, R. E., Costantini, A., Kämpe, A., Alm, J. J. & Mäkitie, O. New Insights Into Monogenic Causes of Osteoporosis. en. *Front. Endocrinol.* **10,** 432844. doi:10.3389/fendo.2019.00070. (2025) (25, 2019).

338. Kwon, Y. H., Fingert, J. H., Kuehn, M. H. & Alward, W. L. M. Primary open-angle glaucoma. en. *N Engl J Med* **360,** 1113–1124. doi:10.1056/NEJMra0804630. (11 12, 2009).

339. Naylor, R. N. *et al.* Cost-effectiveness of MODY genetic testing: translating genomic advances into practical health applications. en. *Diabetes Care* **37,** 202–209. doi:10.2337/dc13-0410. (1 2014).

340. Frayling, T. M. *et al.* beta-cell genes and diabetes: molecular and clinical characterization of mutations in transcription factors. *Diabetes* **50 Suppl 1.** doi:10.2337/diabetes.50.2007.s94. (2025) (2001).

341. Shields, B. M. *et al.* Maturity-onset diabetes of the young (MODY): how many cases are we missing? *Diabetologia* **53.** doi:10.1007/s00125-010-1799-4. (2025) (12 2010).

342. Steele, A. M. *et al.* Increased all-cause and cardiovascular mortality in monogenic diabetes as a result of mutations in the HNF1A gene. en. *Diabet Med* **27,** 157–161. doi:10.1111/j.1464-5491.2009.02913.x. (2 2010).

343. Shekari, S. *et al.* Penetrance of pathogenic genetic variants associated with premature ovarian insufficiency. en. *Nature Medicine* **29,** 1692–1699. doi:10.1038/s41591-023-02405-5. (2025) (7 22, 2023).

344. Brereton, R. G. The chi squared and multinormal distributions: The chi squared and multinormal distributions. en. *J. Chemom.* **29,** 9–12. doi:10.1002/cem.2680. (1 2015).

345. Tenny, S. & Hoffman, M. R. in *StatPearls [Internet]* (StatPearls Publishing, 22, 2023). (2025).

346. Davison, A. C. & Hinkley, D. V. *Bootstrap Methods and Their Application* en. 606 pp. (Cambridge University Press, 28, 1997).

347. North, B. V., Curtis, D. & Sham, P. C. A Note on the Calculation of Empirical P Values from Monte Carlo Procedures. *American Journal of Human Genetics* **71,** 439. doi:10.1086/341527. (2025) (2 2002).

348. Den Tonkelaar, I. Validity and reproducibility of self-reported age at menopause in women participating in the DOM-project. en. *Maturitas* **27,** 117–123. doi:10.1016/s0378-5122(97)01122-5. (2 1997).

349. Liu, Y. *et al.* ACAT: A fast and powerful p value combination method for rare-variant analysis in sequencing studies. en. *Am. J. Hum. Genet.* **104,** 410–421. doi:10.1016/j.ajhg.2019.01.002. (3 7, 2019).

350. GBD 2015 Obesity Collaborators *et al.* Health Effects of Overweight and Obesity in 195 Countries over 25 Years. en. *N. Engl. J. Med.* **377.** doi: 10.1056/NEJMoa1614362, 13–27. doi:10.1056/NEJMoa1614362. (1 12, 2017).

351. *WHO | Obesity and overweight* http://www.who.int/mediacentre/factsheets/fs311/en/ (2018).

352. Heymsfield, S. B. & Wadden, T. A. Mechanisms, Pathophysiology, and Management of Obesity. *N. Engl. J. Med.* **376.** PMID: 28099824, 254–266. doi:10.1056/NEJMra1514009. (3 2017).

353. Després, J.-P. & Lemieux, I. Abdominal obesity and metabolic syndrome. en. *Nature* **444,** 881–887. doi:10.1038/nature05488. (7121 14, 2006).

354. Jayedi, A., Soltani, S., Zargar, M. S., Khan, T. A. & Shab-Bidar, S. Central fatness and risk of all cause mortality: systematic review and dose-response meta-analysis of 72 prospective cohort studies. en. *BMJ* **370,** m3324. doi:10.1136/bmj.m3324. (23, 2020).

355. Zammouri, J. *et al.* Molecular and Cellular Bases of Lipodystrophy Syndromes. en. *Front. Endocrinol.* **12,** 803189. doi:10.3389/fendo.2021.803189. (2021).

356. Shungin, D. *et al.* New genetic loci link adipose and insulin biology to body fat distribution. en. *Nature* **518,** 187–196. doi:10.1038/nature14132. (7538 12, 2015).

357. Wang, Y., Rimm, E. B., Stampfer, M. J., Willett, W. C. & Hu, F. B. Comparison of abdominal adiposity and overall obesity in predicting risk of type 2 diabetes among men1–3. *Am. J. Clin. Nutr.* **81,** 555–563 (2005).

358. Snijder, M. B. *et al.* Associations of hip and thigh circumferences independent of waist circumference with the incidence of type 2 diabetes: the Hoorn Study. en. *Am. J. Clin. Nutr.* **77,** 1192–1197. doi:10.1093/ajcn/77.5.1192. (5 2003).

359. Yusuf, S. *et al.* Obesity and the risk of myocardial infarction in 27,000 participants from 52 countries: a case-control study. en. *Lancet* **366,** 1640–1649. doi:10.1016/S0140-6736(05)67663-5. (9497 5, 2005).

360. Rose, K. M., Newman, B., Mayer-Davis, E. J. & Selby, J. V. Genetic and behavioral determinants of waist-hip ratio and waist circumference in women twins. en. *Obes. Res.* **6,** 383–392. (6 1998).

361. Locke, A. E. *et al.* Genetic studies of body mass index yield new insights for obesity biology. en. *Nature* **518,** 197–206. doi:10.1038/nature14177. (7538 12, 2015).

362. Garvey, W. T. *et al.* Two-year effects of semaglutide in adults with overweight or obesity: the STEP 5 trial. en. *Nat. Med.* **28,** 2083–2091. doi:10.1038/s41591-022-02026-4. (2024) (10 10, 2022).

363. Wadden, T. A. *et al.* Tirzepatide after intensive lifestyle intervention in adults with overweight or obesity: the SURMOUNT-3 phase 3 trial. en. *Nat. Med.* **29,** 2909–2918. doi:10.1038/s41591-023-02597-w. (2024) (11 15, 2023).

364. Kim, K. S. *et al.* GLP-1 increases preingestive satiation via hypothalamic circuits in mice and humans. en. *Science,* eadj2537. doi:10.1126/science.adj2537. (27, 2024).

365. Pulit, S. L. *et al.* Meta-analysis of genome-wide association studies for body fat distribution in 694 649 individuals of European ancestry. en. *Hum. Mol. Genet.* **28,** 166–174. doi:10.1093/hmg/ddy327. (1 1, 2019).

366. Chakhtoura, M. *et al.* Pharmacotherapy of obesity: an update on the available medications and drugs under investigation. en. *EClinicalMedicine* **58,** 101882. doi:10.1016/j.eclinm.2023.101882. (2023).

367. Garg, A. Acquired and inherited lipodystrophies. en. *N. Engl. J. Med.* **350,** 1220–1234. doi:10.1056/NEJMra025261. (12 18, 2004).

368. Semple, R. K., Savage, D. B., Cochran, E. K., Gorden, P. & O'Rahilly, S. Genetic syndromes of severe insulin resistance. en. *Endocr. Rev.* **32,** 498–514. doi:10.1210/er.2010-0020. (4 2011).

369. Akbari, P. *et al.* Sequencing of 640,000 exomes identifies *GPR75* variants associated with protection from obesity. en. *Science* **373.** doi:10.1126/science.abf8683. (6550 2, 2021).

370. Akbari, P. *et al.* Multiancestry exome sequencing reveals INHBE mutations associated with favorable fat distribution and protection from diabetes. en. *Nat. Commun.* **13,** 4844. doi:10.1038/s41467-022-32398-7. (1 23, 2022).

371. Yengo, L. *et al.* Meta-analysis of genome-wide association studies for height and body mass index in 700000 individuals of European ancestry. en. *Hum. Mol. Genet.* **27,** 3641–3649. doi:10.1093/hmg/ddy271. (20 15, 2018).

372. Zhao, Y. *et al.* Protein-truncating variants in BSN are associated with severe adult-onset obesity, type 2 diabetes and fatty liver disease. en. *Nat. Genet.* **56,** 579–584. doi:10.1038/s41588-024-01694-x. (4 2024).

373. Insulin receptor-mediated signaling regulates pluripotency markers and lineage differentiation. *Molecular Metabolism* **18,** 153–163. doi:10.1016/j.molmet.2018.09.003. (2024) (1, 2018).

374. Kamble, P. G. *et al.* Proof-of-concept for CRISPR/Cas9 gene editing in human preadipocytes: Deletion of FKBP5 and PPARG and effects on adipocyte differentiation and metabolism. en. *Sci. Rep.* **10,** 10565. doi:10.1038/s41598-020-67293-y. (1 29, 2020).

375. Liao, W. *et al.* Suppression of PPAR-γ attenuates insulin-stimulated glucose uptake by affecting both GLUT1 and GLUT4 in 3T3-L1 adipocytes. en. *American Journal of Physiology-Endocrinology and Metabolism.* doi:10.1152/ajpendo.00695.2006. (2024) (1, 2007).

376. Liu, S. *et al.* Development of hypertrophic cardiomyopathy in perilipin-1 null mice with adipose tissue dysfunction. en. *Cardiovasc. Res.* **105,** 20–30. doi:10.1093/cvr/cvu214. (2024) (1 21, 2014).

377. Lyu, Y. *et al.* Defective differentiation of adipose precursor cells from lipodystrophic mice lacking perilipin 1. en. *PLoS One* **10,** e0117536. doi:10.1371/journal.pone.0117536. (2 19, 2015).

378. Martinez-Botas, J. *et al.* Absence of perilipin results in leanness and reverses obesity in Leprdb/db mice. en. *Nat. Genet.* **26,** 474–479. doi:10.1038/82630. (2024) (4 2000).

379. Sakaguchi, M. *et al.* Adipocyte Dynamics and Reversible Metabolic Syndrome in Mice with an Inducible Adipocyte-Specific Deletion of the Insulin Receptor. en. *Cell Metab.* **25,** 448–462. doi:10.1016/j.cmet.2016.12.008. (2 7, 2017).

380. Sohn, J. H. *et al.* Perilipin 1 (Plin1) deficiency promotes inflammatory responses in lean adipose tissue through lipid dysregulation. en. *J. Biol. Chem.* **293,** 13974–13988. doi:10.1074/jbc.RA118.003541. (36 7, 2018).

381. Tansey, J. T. *et al.* Perilipin ablation results in a lean mouse with aberrant adipocyte lipolysis, enhanced leptin production, and resistance to diet-induced obesity. *Proc. Natl. Acad. Sci. U. S. A.* **98.** doi:10.1073/pnas.101042998. (2024) (11 22, 2001).

382. Weeks, E. M. *et al.* Leveraging polygenic enrichments of gene features to predict genes underlying complex traits and diseases. en. *Nat. Genet.* **55,** 1267–1276. doi:10.1038/s41588-023-01443-6. (8 2023).

383. Kaisinger, L. R. *et al.* Large-scale exome sequence analysis identifies sex- and age-specific determinants of obesity. en. *Cell Genom* **3,** 100362. doi:10.1016/j.xgen.2023.100362. (8 9, 2023).

384. *All by All* https://allbyall.researchallofus.org/ (2025).

385. Richard, A. J., White, U., Elks, C. M. & Stephens, J. M. *Adipose Tissue: Physiology to Metabolic Dysfunction* (2024) (MDText.com, Inc., 4, 2020).

386. Kunitomi, H. *et al.* The insulin-PI3K-Rac1 axis contributes to terminal adipocyte differentiation through regulation of actin cytoskeleton dynamics. en. *Genes Cells* **25,** 165–174. doi:10.1111/gtc.12747. (3 2020).

387. Qiu, B. & Simon, M. C. BODIPY 493/503 Staining of Neutral Lipid Droplets for Microscopy and Quantification by Flow Cytometry. en. *Bio Protoc* **6.** doi:10.21769/BioProtoc.1912. (17 5, 2016).

388. Jones, J. R. *et al.* Deletion of PPARγ in adipose tissues of mice protects against high fat diet-induced obesity and insulin resistance. en. *Proceedings of the National Academy of Sciences* **102,** 6207–6212. doi:10.1073/pnas.0306743102. (2024) (17 26, 2005).

389. Laber, S. *et al.* Discovering cellular programs of intrinsic and extrinsic drivers of metabolic traits using LipocyteProfiler. en. *Cell Genom* **3,** 100346. doi:10.1016/j.xgen.2023.100346. (7 12, 2023).

390. Zhang, Z., Zhan, X., Kim, B. & Wu, J. A proteomic approach identifies SAFB-like transcription modulator (SLTM) as a bidirectional regulator of GLI family zinc finger transcription factors. en. *J. Biol. Chem.* **294,** 5549–5561. doi:10.1074/jbc.RA118.007018. (14 5, 2019).

391. Pospisilik, J. A. *et al.* Drosophila genome-wide obesity screen reveals hedgehog as a determinant of brown versus white adipose cell fate. en. *Cell* **140,** 148–160. doi:10.1016/j.cell.2009.12.027. (1 8, 2010).

392. Inoue, M., Akama, T., Jiang, Y. & Chun, T.-H. The Exocyst Complex Regulates Free Fatty Acid Uptake by Adipocytes. *PLoS One* **10,** e0120289. doi:10.1371/journal.pone.0120289. (2024) (3 13, 2015).

393. Lee, J. W., Choi, H. S., Gyuris, J., Brent, R. & Moore, D. D. Two classes of proteins dependent on either the presence or absence of thyroid hormone for interaction with the thyroid hormone receptor. *Mol. Endocrinol.* **9.** doi:10.1210/mend.9.2.7776974. (2024) (2 1995).

394. Ma, Y. *et al.* Adipocyte Thyroid Hormone β Receptor–Mediated Hormone Action Fine-tunes Intracellular Glucose and Lipid Metabolism and Systemic Homeostasis. en. *Diabetes* **72,** 562–574. doi:10.2337/db22-0656. (2024) (5 1, 2023).

395. Pramfalk, C., Pedrelli, M. & Parini, P. Role of thyroid receptor β in lipid metabolism. *Biochim. Biophys. Acta* **1812.** doi:10.1016/j.bbadis.2010.12.019. (2024) (8 2011).

396. Harrison, S. A. *et al.* Resmetirom for nonalcoholic fatty liver disease: a randomized, double-blind, placebo-controlled phase 3 trial. en. *Nat. Med.* **29,** 2919–2928. doi:10.1038/s41591-023-02603-1. (2024) (11 16, 2023).

397. Huang, G. *et al.* α3(V) collagen is critical for glucose homeostasis in mice due to effects in pancreatic islets and peripheral tissues. *J. Clin. Invest.* **121.** doi:10.1172/JCI45096. (2024) (2 2011).

398. Gerut, Z. E. Abstract: The enzymatic dissolution of human fat. en. *Plast. Reconstr. Surg. Glob. Open* **5,** 66–67. doi:10.1097/01.gox.0000526258.02970.67. (9S 2017).

399. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. en. *Nucleic Acids Res.* **28,** 27–30. doi:10.1093/nar/28.1.27. (2024) (1 1, 2000).

400. Jassal, B. *et al.* The reactome pathway knowledgebase. *Nucleic Acids Res.* **48.** doi:10.1093/nar/gkz1031. (2024) (D1 8, 2020).

401. Aleksander, S. A. *et al.* The Gene Ontology knowledgebase in 2023. en. *Genetics* **224,** iyad031. doi:10.1093/genetics/iyad031. (2024) (1 3, 2023).

402. Luo, Y. *et al.* Runx1 regulates osteogenic differentiation of BMSCs by inhibiting adipogenesis through Wnt/β-catenin pathway. *Arch. Oral Biol.* **97.** doi:10.1016/j.archoralbio.2018.10.028. (2024) (2019).

403. De Winter, T. J. J. & Nusse, R. Running Against the Wnt: How Wnt/β-Catenin Suppresses Adipogenesis. en. *Frontiers in Cell and Developmental Biology* **9.** doi:10.3389/fcell.2021.627429. (2024) (2021).

404. Nueda, M.-L. *et al.* DLK proteins modulate NOTCH signaling to influence a brown or white 3T3-L1 adipocyte fate. en. *Sci. Rep.* **8,** 1–16. doi:10.1038/s41598-018-35252-3. (2024) (1 16, 2018).

405. Huang, L.-Y., Chiu, C.-J., Hsing, C.-H. & Hsu, Y.-H. Interferon Family Cytokines in Obesity and Insulin Sensitivity. en. *Cells* **11.** doi:10.3390/cells11244041. (2024) (24 2022).

406. Zhao, M. & Chen, X. Endocrine and Metabolic Dysfunction during Aging and Senescence: Effect of lipopolysaccharides on adipogenic potential and premature senescence of adipocyte progenitors. en. *American Journal of Physiology - Endocrinology and Metabolism* **309,** E334. doi:10.1152/ajpendo.00601.2014. (2024) (4 8, 2015).

407. Iizuka, K., Bruick, R. K., Liang, G., Horton, J. D. & Uyeda, K. Deficiency of carbohydrate response element-binding protein (ChREBP) reduces lipogenesis as well as glycolysis. en. *Proc. Natl. Acad. Sci. U. S. A.* **101,** 7281–7286. doi:10.1073/pnas.0401516101. (2024) (19 26, 2004).

408. Oh, A. R. *et al.* ChREBP deficiency leads to diarrhea-predominant irritable bowel syndrome. en. *Metabolism* **85,** 286–297. doi:10.1016/j.metabol.2018.04.006. (2024) (15, 2018).

409. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. en. *Genome Biol.* **15,** 1–21. doi:10.1186/s13059-014-0550-8. (2024) (12 5, 2014).

410. Asymptotic properties of principal component analysis and shrinkage-bias adjustment under the generalized spiked population model. *J. Multivar. Anal.* **173,** 145–164. doi:10.1016/j.jmva.2019.02.007. (2024) (1, 2019).

411. (2022). randomForest: Breiman and Cutlers Random Forests for Classification and Regression. URL: hhttps://cran.r-project.org/web/packages/randomForest/randomForest.pdf.

412. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. en. *Genome Biol.* **17,** 122. doi:10.1186/s13059-016-0974-4. (1 6, 2016).

413. Liu, X., Li, C., Mou, C., Dong, Y. & Tu, Y. dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs. en. *Genome Med.* **12,** 103. doi:10.1186/s13073-020-00803-9. (1 2, 2020).

414. *VEP version 105 with LOFTEE v1.04_GRCh38 containers* en. https://github.com/BRaVa-genetics/vep105_loftee (2024).

415. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. en. *Gigascience* **4,** 7. doi:10.1186/s13742-015-0047-8. (25, 2015).

416. Van Hout, C. V. *et al.* Exome sequencing and characterization of 49,960 individuals in the UK Biobank. en. *Nature* **586,** 749–756. doi:10.1038/s41586-020-2853-0. (2024) (7831 21, 2020).

417. Benner, C. *et al.* Prospects of Fine-Mapping Trait-Associated Genomic Regions by Using Summary Statistics from Genome-wide Association Studies. en. *Am. J. Hum. Genet.* **101,** 539–551. doi:10.1016/j.ajhg.2017.08.012. (4 5, 2017).

418. Zhou, S. *et al.* Converging evidence from exome sequencing and common variants implicates target genes for osteoporosis. en. *Nat. Genet.* **55,** 1277–1287. doi:10.1038/s41588-023-01444-5. (2024) (8 9, 2023).

419. Kuan, V. *et al.* A chronological map of 308 physical and mental health conditions from 4 million individuals in the English National Health Service. en. *Lancet Digit Health* **1,** e63–e77. doi:10.1016/S2589-7500(19)30012-3. (2 2019).

420. Kaplan, E. L. & Meier, P. Nonparametric Estimation from Incomplete Observations. *J. Am. Stat. Assoc.* **53,** 457–481. doi:10.1080/01621459.1958.10501452. (282 1, 1958).

421. Therneau, T. M. *survival: Survival Analysis* (The R Foundation, 22, 2001). doi:10.32614/cran.package.survival. https://cran.r-project.org/package=survival.

422. Kassambara, A., Kosinski, M. & Biecek, P. Drawing Survival Curves using 'ggplot2' [R package survminer version 0.4.8]. (25, 2020).

423. Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. en. *Nat. Genet.* **47,** 1236–1241. doi:10.1038/ng.3406. (2024) (11 28, 2015).

424. 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.** doi:10.1038/nm.3881. (2024) (7 2015).

425. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. en. *Bioinformatics* **29,** 15. doi:10.1093/bioinformatics/bts635. (2024) (1 2013).

426. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. en. *BMC Bioinformatics* **12,** 1–16. doi:10.1186/1471-2105-12-323. (2024) (1 4, 2011).

427. Livak, K. J. & Schmittgen, T. D. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. en. *Methods* **25,** 402–408. doi:10.1006/meth.2001.1262. (4 2001).

428. Cook, R. D. Detection of influential observation in linear regression. *Technometrics* **19,** 15–18. doi:10.1080/00401706.1977.10489493. (1 1977).

429. Fox, J. & Scott Long, J. *Modern Methods of Data Analysis* en. 468 pp. (SAGE Publications, Incorporated, 1990).

430. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. en. *Proc. Natl. Acad. Sci. U. S. A.* **102,** 15545–15550. doi:10.1073/pnas.0506580102. (43 25, 2005).

431. Buniello, A. *et al.* Open Targets Platform: facilitating therapeutic hypotheses building in drug discovery. en. *Nucleic Acids Res.* **53,** D1467–D1475. doi:10.1093/nar/gkae1128. (D1 6, 2025).

432. Berman, H. M. The Protein Data Bank. *Nucleic Acids Res.* **28,** 235–242. doi:10.1093/nar/28.1.235. (1 1, 2000).

433. Zdrazil, B. *et al.* The ChEMBL Database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. en. *Nucleic Acids Res.* **52,** D1180–D1192. doi:10.1093/nar/gkad1004. (D1 5, 2024).

434. *DrugEBIlity - Structure-based component* https://chembl.github.io/drugebility-structure-based-component/ (2025).

435. Finan, C. *et al.* The druggable genome and support for target identification and validation in drug development. en. *Sci. Transl. Med.* **9.** doi:10.1126/scitranslmed.aag1166. (383 29, 2017).

436. UniProt Consortium. UniProt: The universal protein knowledgebase in 2025. en. *Nucleic Acids Res.* **53,** D609–D617. doi:10.1093/nar/gkae1010. (D1 6, 2025).

437. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. en. *Nat. Genet.* **25,** 25–29. doi:10.1038/75556. (1 2000).

438. Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E. L. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. en. *J. Mol. Biol.* **305,** 567–580. doi:10.1006/jmbi.2000.4315. (3 19, 2001).

439. Thul, P. J. *et al.* A subcellular map of the human proteome. en. *Science* **356.** doi:10.1126/science.aal3321. (6340 26, 2017).

440. Hornbeck, P. V. *et al.* PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. en. *Nucleic Acids Res.* **43,** D512–20. doi:10.1093/nar/gku1267. (Database issue 2015).

441. Chen, T. *et al.* mUbiSiDa: a comprehensive database for protein ubiquitination sites in mammals. en. *PLoS One* **9,** e85744. doi:10.1371/journal.pone.0085744. (1 17, 2014).

442. Kim, W. *et al.* Systematic and quantitative assessment of the ubiquitin-modified proteome. en. *Mol. Cell* **44,** 325–340. doi:10.1016/j.molcel.2011.08.025. (2 21, 2011).

443. Mathieson, T. *et al.* Systematic analysis of protein turnover in primary cells. en. *Nat. Commun.* **9,** 689. doi:10.1038/s41467-018-03106-1. (1 15, 2018).

444. Gerasimavicius, L., Livesey, B. J. & Marsh, J. A. Loss-of-function, gain-of-function and dominant-negative mutations have profoundly different effects on protein structure. en. *Nat. Commun.* **13,** 3895. doi:10.1038/s41467-022-31686-6. (1 6, 2022).

445. Cirulli, E. T. & Goldstein, D. B. Uncovering the roles of rare variants in common disease through whole-genome sequencing. en. *Nat. Rev. Genet.* **11,** 415–425. doi:10.1038/nrg2779. (6 2010).

446. Man, K., Kallies, A. & Vasanthakumar, A. Resident and migratory adipose immune cells control systemic metabolism and thermogenesis. en. *Cell. Mol. Immunol.* **19,** 421–431. doi:10.1038/s41423-021-00804-7. (3 2022).

447. Wang, Y. *et al.* The role of somatosensory innervation of adipose tissues. en. *Nature* **609,** 569–574. doi:10.1038/s41586-022-05137-7. (7927 2022).

448. Zhang, T. *et al.* CXCL1 mediates obesity-associated adipose stromal cell trafficking and function in the tumour microenvironment. en. *Nat. Commun.* **7,** 11674. doi:10.1038/ncomms11674. (31, 2016).

449. Koprulu, M. *et al.* Identification of rare loss-of-function genetic variation regulating body fat distribution. en. *J. Clin. Endocrinol. Metab.* **107,** 1065–1077. doi:10.1210/clinem/dgab877. (4 24, 2022).

450. Softic, S. *et al.* Lipodystrophy due to adipose tissue-specific insulin receptor knockout results in progressive NAFLD. en. *Diabetes* **65,** 2187–2200. doi:10.2337/db16-0213. (8 2016).

451. Ramirez, A. H., Gebo, K. A. & Harris, P. A. Progress with the all of us research program: Opening access for researchers: Opening access for researchers. en. *JAMA* **325,** 2441–2442. doi:10.1001/jama.2021.7702. (24 22, 2021).

452. Chen, Y. *et al.* De novo variants in the RNU4-2 snRNA cause a frequent neurodevelopmental syndrome. en. *Nature* **632,** 832–840. doi:10.1038/s41586-024-07773-7. (8026 2024).

453. Gaynor, S. M. *et al.* Yield of genetic association signals from genomes, exomes and imputation in the UK Biobank. en. *Nat. Genet.* **56,** 2345–2351. doi:10.1038/s41588-024-01930-4. (11 2024).

454. Cheng, J. *et al.* Accurate proteome-wide missense variant effect prediction with AlphaMissense. en. *Science* **381,** eadg7492. doi:10.1126/science.adg7492. (6664 22, 2023).

455. Gao, H. *et al.* The landscape of tolerated genetic variation in humans and primates. en. *Science* **380,** eabn8153. doi:10.1126/science.abn8197. (6648 2, 2023).

456. Loya, H., Kalantzis, G., Cooper, F. & Palamara, P. F. A scalable variational inference approach for increased mixed-model association power. en. *Nat. Genet.* **57,** 461–468. doi:10.1038/s41588-024-02044-7. (2 2025).

457. Depope, A., Bajzik, J., Mondelli, M. & Robinson, M. R. Joint modelling of whole genome sequence data for human height via approximate message passing. *Genetics.* (17, 2023).

458. 100,000 Genomes Project Pilot Investigators *et al.* 100,000 genomes pilot on rare-disease diagnosis in health care - preliminary report. en. *N. Engl. J. Med.* **385,** 1868–1880. doi:10.1056/NEJMoa2035790. (20 11, 2021).

459. McCaw, Z. R. *et al.* An allelic-series rare-variant association test for candidate-gene discovery. en. *Am. J. Hum. Genet.* **110,** 1330–1342. doi:10.1016/j.ajhg.2023.07.001. (8 3, 2023).

460. McCaw, Z. R. *et al.* A Scalable Framework for Identifying Allelic Series from Summary Statistics. *Genetics.* (3, 2024).

461. Morigny, P., Boucher, J., Arner, P. & Langin, D. Lipid and glucose metabolism in white adipocytes: pathways, dysfunction and therapeutics. en. *Nat. Rev. Endocrinol.* **17,** 276–295. doi:10.1038/s41574-021-00471-8. (5 2021).

462. Bock, C. *et al.* High-content CRISPR screening. en. *Nat. Rev. Methods Primers* **2,** 1–23. doi:10.1038/s43586-021-00093-4. (1 10, 2022).

463. *Illumina's revolutionary NovaSeq X exceeds 200th order milestone in first quarter 2023* en. https://emea.illumina.com/company/news-center/press-releases/2023/4b48580b-42d4-419b-8512-5adcbb069836.html (2025).

464. Wheeler, D. A. *et al.* The complete genome of an individual by massively parallel DNA sequencing. en. *Nature* **452,** 872–876. doi:10.1038/nature06884. (7189 17, 2008).

465. Evangelou, E. & Ioannidis, J. P. A. Meta-analysis methods for genome-wide association studies and beyond. en. *Nat. Rev. Genet.* **14,** 379–389. doi:10.1038/nrg3472. (6 2013).

466. Zhou, W. *et al.* Global Biobank Meta-analysis Initiative: Powering genetic discovery across human disease. en. *Cell Genom.* **2,** 100192. doi:10.1016/j.xgen.2022.100192. (10 12, 2022).

467. *BRaVa: Biobank Rare Variant Analysis* https://brava-genetics.github.io/BRaVa/index.html (2025).

468. Cerezo, M. *et al.* The NHGRI-EBI GWAS Catalog: standards for reusability, sustainability and diversity. en. *Nucleic Acids Res.* **53,** D998–D1005. doi:10.1093/nar/gkae1070. (D1 6, 2025).

469. Trajanoska, K. *et al.* From target discovery to clinical drug development with human genetics. en. *Nature* **620,** 737–745. doi:10.1038/s41586-023-06388-8. (7975 2023).

470. Minikel, E. V., Painter, J. L., Dong, C. C. & Nelson, M. R. Refining the impact of genetic evidence on clinical success. en. *Nature* **629,** 624–629. doi:10.1038/s41586-024-07316-0. (8012 2024).

471. Yengo, L. *et al.* A saturated map of common genetic variants associated with human height. en. *Nature* **610,** 704–712. doi:10.1038/s41586-022-05275-y. (7933 2022).

472. Rehm, H. L. Evolving health care through personal genomics. en. *Nat. Rev. Genet.* **18,** 259–267. doi:10.1038/nrg.2016.162. (4 2017).

473. Bertoli-Avella, A. M. *et al.* Successful application of genome sequencing in a diagnostic setting: 1007 index cases from a clinically heterogeneous cohort. en. *Eur. J. Hum. Genet.* **29,** 141–153. doi:10.1038/s41431-020-00713-9. (1 2021).

474. Costain, G. *et al.* Genome sequencing as a diagnostic test in children with unexplained medical complexity. en. *JAMA Netw. Open* **3,** e2018109. doi:10.1001/jamanetworkopen.2020.18109. (9 1, 2020).

475. Stranneheim, H. *et al.* Integration of whole genome sequencing into a healthcare setting: high diagnostic rates across multiple clinical entities in 3219 rare disease patients. en. *Genome Med.* **13,** 40. doi:10.1186/s13073-021-00855-5. (1 17, 2021).

476. Rehm, H. L. Time to make rare disease diagnosis accessible to all. en. *Nat. Med.* **28,** 241–242. doi:10.1038/s41591-021-01657-3. (2 2022).

477. Fowler, D. M. & Rehm, H. L. Will variants of uncertain significance still exist in 2030? en. *Am. J. Hum. Genet.* **111,** 5–10. doi:10.1016/j.ajhg.2023.11.005. (1 4, 2024).

478. Dixon, P., Keeney, E., Taylor, J. C., Wordsworth, S. & Martin, R. M. Can polygenic risk scores contribute to cost-effective cancer screening? A systematic review. en. *Genet. Med.* **24,** 1604–1617. doi:10.1016/j.gim.2022.04.020. (8 2022).

479. Kiflen, M. *et al.* Cost-effectiveness of polygenic risk scores to guide statin therapy for cardiovascular disease prevention. en. *Circ. Genom. Precis. Med.* **15,** e003423. doi:10.1161/CIRCGEN.121.003423. (5 2022).

480. Ruger, J. P. & Kim, H.-J. Global health inequalities: an international comparison. en. *J. Epidemiol. Community Health* **60,** 928–936. doi:10.1136/jech.2005.041954. (11 2006).

481. Marcus, M. E. *et al.* Use of statins for the prevention of cardiovascular disease in 41 low-income and middle-income countries: a cross-sectional study of nationally representative, individual-level data. en. *Lancet Glob. Health* **10,** e369–e379. doi:10.1016/S2214-109X(21)00551-9. (3 2022).

482. Hetu, M., Koutouki, K. & Joly, Y. Genomics for all: International open science genomics projects and capacity building in the developing world. en. *Front. Genet.* **10,** 95. doi:10.3389/fgene.2019.00095. (15, 2019).

483. Seabold, S. & Perktold, J. *statsmodels: Econometric and statistical modeling with python* doi:10.25080/Majora-92bf1922-011 (SciPy Proceedings, 2010).