COMMON ‘INBORN ERRORS’ OF METABOLISM IN THE GENERAL POPULATION

Victoria P.W. Au Yeung

Ph.D. 2021

Common ‘Inborn Errors’ of Metabolism in the General Population

Victoria Pui Wa Au Yeung

Christ’s College

University of Cambridge

January 2021

This thesis is submitted for the degree of the Doctor of Philosophy

PREFACE

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text. It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my thesis has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. It does not exceed the prescribed word limit for the relevant Degree Committee.

This thesis consists of seven chapters structured around two main research projects. The first project, which is described in **Chapter 2**, characterised the metabolic changes associated with long-term weight gain and identified small molecules that are likely candidates that could mediate the association between weight gain and the risk of incident type 2 diabetes. The second project, which is described in **Chapters 3-6**, phenotypically characterised variation at genes known to cause rare (Mendelian) metabolic disorders, also known as 'Inborn Errors of Metabolism', or IEMs.

The work described in this thesis is my own and was carried out independently unless otherwise stated. I used results from the largest existing genetic association study (GWAS) of untargeted metabolites to identify and prioritise 'IEM gene-linked loci' and the collaborative nature of large-scale GWAS meant that several elements of my work had to be conducted in close collaboration with others (see details below). Also, for the first project, collaborators at Metabolon Inc. (Morrisville, North Carolina, USA) performed the structural characterisation of a metabolite with hitherto unknown identity, X-12063. This was found to be strongly associated with weight gain in my work **(Chapter 2, Section 2.4.3.5.)** and the collaboration was initiated by myself. For this work, Dr. Maik Pietzner helped to estimate associations of candidate mediators I had identified with 27 incident diseases in the EPIC-Norfolk cohort, as well as the contributions of genetic and non-genetic factors to the variance in candidate mediator levels **(Chapter 2, Section 2.4.3.3.)**. I have written up and submitted Chapter 2 for publication. During the process, I have received comments from all co-authors and these have been incorporated into a revised chapter version.

As mentioned above, all later chapters built on results from the Metabolon mGWAS consortium, for which primary GWAS analyses had been performed by Drs. Isobel Stewart (MRC Epidemiology Unit) and Praveen Surendran (Cardiovascular Epidemiology Unit, University of Cambridge) **(Chapter 3,**

Section 3.4.3.1.-3.4.3.7.). As a member of the central analytical team and writing group, I independently prioritised metabolite-associated variants at IEM genes and systematically characterised their metabolic and phenotypic effects. This large-scale effort included several other collaborators from the University of Cambridge in the UK, the Helmholtz Zentrum Munich (Germany), Dr. Karsten Suhre, a professor of physiology and biophysics at Weill Cornell Medicine-Qatar (Qatar), and Dr. Eric Fauman, an employee of Pfizer Inc. (USA). Dr. Eric Fauman performed likely causal gene annotation **(Chapter 3, Section 3.4.3.8.)**, provided an independent assessment of my mapping of metabolites to those known to be affected (or closely related) in the corresponding IEMs **(Chapter 3, Section 3.4.3.10.)** and helped to identify additional variants for phenotypic characterisation, as listed in **Appendix Ch5_ST2**.

During my PhD, I have made substantial contributions to other collaborative research projects not included in this thesis, including scientific papers that are under review or published, as summarised in the **Publications and Presentations** section.

Victoria Au Yeung
January 2021

ABSTRACT

Common ‘Inborn Errors’ of Metabolism in the General Population, by Victoria Pui Wa Au Yeung

Inborn errors of metabolism (IEMs) are a group of disorders characterised by the toxic accumulation or deficiency of circulating molecules (‘metabolites’) caused by rare genetic mutations. Previous studies have identified select examples where common variants at genes known to cause rare Mendelian diseases, including IEMs (e.g. *LPL*, *DBH, PPM1K)*, are linked to phenotypic consequences in the general population that also occur in patients with the corresponding rare disease. Advances in genetic and metabolic profiling at an epidemiological scale now provide an opportunity to systematically identify such examples in the population and characterise their downstream effects on health.

To assess the value of untargeted metabolomic profiling for the study of common complex diseases, I identified candidate mediators of the association between weight gain and future type 2 diabetes risk based on untargeted, large-scale metabolomic profiling of a large prospective cohort. Integration of metabolomics, genetic profiling and comprehensive longitudinal follow up for a range of diseases together with the application of Bayesian and genetic epidemiological methods enabled the identification of 20 candidate mediators. These reflected genetic susceptibility to adiposity and insulin resistance and explained most of the increased T2D risk associated with weight gain.

To systematically characterise the phenotypic effects of variation at IEM-causing genes, I identified sentinel variants at these genes associated with plasma metabolites affected in the corresponding IEM across the genome. Of the 202 ‘IEM familiar’ variants (IFVs) detected, 187 at 89 loci were not previously reported as pathogenic for the corresponding IEM in ClinVar and 51 of these were associated with extreme metabolite levels (<2.5th or >97.5th percentile) or had non-additive effects on metabolite levels. Phenome-wide assessment identified 1,553 IFV-phenotype associations at 108 loci. Of the detected associations, 703 at 54 loci were of particular interest as the phenotype related to a symptom of the corresponding IEM. At 24 of these 54 loci, genetic colocalisation detected shared genetic signals for IEM-related metabolites and phenotypes. For example, in line with norepinephrine deficiency causing dizziness on standing in severe cases of rare orthostatic hypotension (OMIM #223360), I identified a genetic signal at the dopamine beta hydroxylase (*DBH*) locus associated with decreased levels of the downstream catecholamine vanillylmandelate in the general population (IFV EAF=0.074). This signal was shared with that for lower risk of hypertension (based on 462,933 participants in UK Biobank) and other blood pressure-related phenotypes with high posterior probability of colocalisation ($PP_{colocalisation}$=0.94, with >99% of the probability explained by the IFV).

This work uses untargeted metabolomic profiling to identify underlying disease mechanisms and demonstrate the proof-of-principle that common variants can have similar health consequences to those caused by rare mutations at the same IEM gene.

PUBLICATIONS AND PRESENTATIONS

This list includes publications authored and co-authored during the PhD, including manuscripts under review or in preparation.

Publications

Auyeung VPW, Pietzner M, Stewart ID, Wittemans LBL, Williamson A, Day FR, Kerrison ND, Schaaf G, Sefler A, Adam KP, Miller L, Khaw KT, Lotta LA, Sharp SJ, Evans A, Newcombe PJ, Wareham NJ, Langenberg C. "Data-driven identification of metabolic pathways linking weight gain to type 2 diabetes: large-scale analysis of untargeted metabolomics in a prospective population-based cohort". (*Under review*)

Podmore C., Stewart ID, …, **Auyeung VPW**, .., Langenberg C. "Genetic regulation of iron metabolism, chronic iron overload and iron-tissue deposition in non-*HFE* carriers". (*Manuscript in preparation*)

Bowker N, Hansford R, Burgess S, Foley CN, **Auyeung VPW**, Erzurumluoglu MA, Stewart ID, Wheeler E, Pietzner M, Gribble FM, Reimann F, Bhatnagar P, Coghlan M, Wareham NJ, Langenberg C. "Genetically predicted glucose-dependent insulinotropic polypeptide (GIP) levels and cardiovascular disease risk are driven by independent mechanisms at the GIPR". (*Under review*)

Lotta LA, Pietzner M, Stewart ID, Wittemans LBL, Li C, Bonelli R, Raffler J, Biggs EK, Oliver-Williams C, **Auyeung VPW**, Luan J, Wheeler E, Paige E, Surendran P, Michelotti GA, Scott RA, Burgess S, Zuber V, Sanderson E, Koulman A, Imamura F, Forouhi NG, Khaw KT, MacTel Consortium, Griffin JL, Wood AM, Kastenmüller G, Danesh J, Butterworth AS, Gribble FM, Reimann F, Bahlo M, Fauman EB, Wareham NJ, Langenberg C. "A cross-platform approach identifies genetic regulators of human metabolism and health". ***Nature Genetics*** 2021; 53: 54-64.

Pietzner M., Wheeler E, Carrasco-Zanini J, Raffler J, Kerrison ND, Oerton E, **Auyeung VPW**, Luan J, Finan C, Casas JP, Ostroff R, Williams SA, Kastenmüller G, Ralser M, Gamazon ER, Wareham NJ, Hingorani AD, Langenberg C. "Genetic architecture of host proteins interacting with SARS-CoV-2". (*Under review*)

Presentations

Oral presentations

Auyeung VPW, Wittemans LBL, Stewart ID, Cai L, Li C, Day FR, Newcombe PJ, Sharp SJ, Khaw KT, Lotta LA, Wareham NJ, Langenberg C. "The metabolic consequences of weight gain on type 2 diabetes risk". (*Presented at the European Diabetes Epidemiology Group, 2018*)

Poster presentations

Auyeung VPW, metabolome Genetic Architecture Programme (mGAP) investigators. "Common 'inborn errors' of metabolism in the general population". (*Manuscript in preparation, Presented at the American Society of Human Genetics Annual Meeting, 2020*)

Auyeung VPW, Day FR, Wittemans LBL, Stewart ID, Cai L, Li C, Newcombe PJ, Khaw KT, Lotta LA, Wareham NJ, Langenberg C. "Understanding the metabolic pathways linking weight gain and type 2 diabetes risk." (*Presented at the 7th Annual Meeting of the EASD Study Group on the Genetics of Diabetes (EASD-SGGD), 2019*)

ACKNOWLEDGEMENTS

I would like to thank the people who have supported me throughout my PhD. First, I would like to thank my supervisor Dr. Claudia Langenberg and my adviser Dr. Felix Day for their invaluable expertise, support and advice. Claudia, thank you for giving me the confidence to become an independent researcher and for always encouraging me to take a step back and look at the 'bigger picture'. Your guidance and support have shaped the way I think about research questions and design analyses to address them, a process (though emphasised throughout my university courses) that I only learned how to do during my PhD project. Felix, thank you for being the office 'neighbour next door' and for all your guidance and support on everything from statistics and epidemiology related questions to PhD and career-related matters. Thank you also for your patience in teaching me how to evaluate study designs and to generate simple yet informative figures, both of which are invaluable skills to me going forward in my career.

I would also like to thank the members of the Diabetes Aetiology team at the MRC Epidemiology Unit, especially Dr. Maik Pietzner for his engaging and thought-provoking insights on a variety of research questions and to Dr. Isobel Stewart for her tireless and rigorous efforts to improve the quality of my work. You have both taught me so much and inspired me to stay curious, even when the methods get tedious! I am also immensely grateful for your continuous help and time spent addressing all my questions, both clever and silly. I would also like to thank Dr. Eric Fauman for his expertise in biochemistry and academic support in the IEMs work, which greatly contributed to the quality and findings of my research.

I would also like to thank the PhD students within the unit who shared my journey with me for their friendship and support, especially Hannah, Nick, Lina, Chunxiao, Alice, Julia, Chen, Laura and Eirini. I would also like to thank the PhD students of the Wellcome-Trust PhD in Metabolic and Cardiovascular Disease programme, especially Joy, Filip, Sam and Tom, for being a supportive and engaging community that I am proud to be a part of.

Finally, I would like to thank my parents and my sister for their support and encouragement, and my grandmother, who always had unwavering faith in me. I am also immensely grateful to my partner Radu, who has supported me in more ways than I can describe.

TABLE OF CONTENTS

ABBREVIATIONS AND TERMINOLOGY 12

INDEX OF FIGURES 14

INDEX OF TABLES 17

CHAPTER 1: OPPORTUNITIES TO STUDY COMPLEX DISEASE AETIOLOGY USING UNTARGETED METABOLOMIC PROFILING IN POPULATION COHORTS 18

1.1 Introduction to the Metabolome 18

1.1.1. The Metabolome in Health and Disease 18

1.1.2. Measurement of the Metabolome 18

1.2. Genetic Regulation of Metabolites 20

1.2.1. Genetic Regulation of Metabolite Levels in Rare Disease Patients 20

1.2.2. Genetic Architecture of Metabolite Levels in Population-based Studies 21

1.2.3. Metabolic Effects of Variation at IEM Genes May Translate into Health-related Effects in the General Population 23

1.3. The Phenome 24

1.3.1. Genetic Architecture of Complex Traits and Diseases 24

1.3.2. Approaches to Map Rare Disease Symptoms to Phenotypically Similar Common Diseases That Could Share the Same Aetiological Origins 24

1.4. Challenges and Opportunities of Assessing the Effects of Changing Metabolite Levels on Complex Disease 25

1.5. Thesis Overview 26

CHAPTER 2: IDENTIFICATION OF WEIGHT GAIN-ASSOCIATED METABOLIC PATHWAYS AND THEIR CUMULATIVE EFFECT ON INCIDENT TYPE 2 DIABETES RISK 29

2.1. Abstract 29

2.2. Introduction 30

2.3. Aim and Objectives 30

2.4. Methods 30

2.4.1. Study design and participants 30

2.4.2. Measurements 31

2.4.3. Statistical analysis 31

2.5. Results 35

2.5.1. Metabolites Associated With Weight Gain 35

2.5.2. Discovery of Metabolites Associated With T2D 36

2.5.3. Candidate Mediators Strongly Attenuated the Association of Weight Gain With Incident T2D 37

2.5.4. Genetic and Other Non-modifiable and Modifiable Factors Influenced Metabolite Levels 40

2.5.5. Candidate Mediators Reflected Specific and Non-specific Pathways in T2D and Cardiometabolic Disease 42

2.5.6. X-12063 Represented Steroid Modification and Clearance Pathways in Cardiometabolic Disease 43

2.5.7. Metabolic Network Visualisation Highlighted Dietary Origins of Candidate Mediators With Unidentified Chemical Identity 44

2.6. Discussion 44

2.6.1. Summary of Findings 44

2.6.2. Novelty of Findings 45

2.6.3. Study Strengths and Limitations 47

2.6.4. Conclusions 47

CHAPTER 3: SYSTEMATIC IDENTIFICATION OF THE METABOLIC EFFECTS OF VARIANTS AT IEM GENES 49

3.1. Abstract 49

3.2. Background 50

3.3. Aim and Objectives 51

3.4. Methods 52

3.4.1. Study design and participants 52

3.4.2. Phenotypic measurements 52

3.4.3. Statistical analysis 53

3.5. Results 60

3.5.1. A Total of 1,847 Locus-metabolite Associations Across 330 Loci and 646 Metabolites Were Identified in the Metabolon MGWAS 60

3.5.2. Metabolic loci in the Metabolon mGWAS are enriched for IEM genes 61

3.5.3. A Total of 202 of 241 Variants (84%) at Loci Linked to IEM Genes Were Associated With a Metabolite Implicated in IEM Aetiology 62

3.5.4. IFVs have large effect sizes on metabolite levels 63

3.6. Discussion 64

3.6.1. Summary of Findings 64

3.6.2. Novelty of Findings 65

3.6.3. Study Strengths and Limitations 65

3.6.4. Conclusions 67

CHAPTER 4: CHARACTERISATION OF METABOLITE-ASSOCIATED VARIANTS AT IEM GENES 68

4.1. Abstract 68

4.2. Background 69

4.3. Aim and Objectives 70

4.4. Methods 71

4.4.1. Study Design and Participants 71

4.4.2. Measurements 71

4.4.3. Statistical Analysis 72

4.5. Results 74

4.5.1. IFVs Accounted for Substantial Proportions of Variance in Metabolite Levels Explained by Genetics 74

4.5.2. IFVs May Alter Metabolite Levels Through Effects on Protein Activity or mRNA Splicing . 74

4.5.3. A Total of 48 of 187 IFVs Were Associated With Extreme Metabolite Levels 75

4.5.4. IFVs With Phenotypic Consequences That Could Plausibly Arise from Their Metabolic Effects 81

4.6. Discussion 85

4.6.1. Summary of Findings 85

4.6.2. Novelty of Findings 86

4.6.3. Study Strengths and Limitations 87

4.6.4. Conclusions 88

CHAPTER 5: PHENOTYPIC EFFECTS OF COMMON 'INBORN ERRORS' OF METABOLISM 89

5.1. Abstract 89

5.2. Background 90

5.3. Aim and Objectives 93

5.4. Methods 93

5.4.1. Studies and Measurements 93

5.4.2. Statistical Analysis 95

5.4.3. Annotation of Gene-drug and Variant-drug Interactions 98

5.5. Results 99

5.5.1. IFVs Are Associated With a Broad Range of Complex Phenotypes in the General Population 99

5.5.2. Colocalisation Analysis Validated Known Metabolic Contributions of Known IFVs to Complex Traits and Diseases 100

5.5.3. Variant-drug and Gene-drug Annotations of IFVs and Their Mapped IEM Genes 107

5.6. Discussion 107

CHAPTER 6: IEM FAMILIAR VARIANTS AND THEIR ASSOCIATION WITH IEM-RELATED DISEASE PROFILES 112

6.1. Abstract 112

6.2. Background 113

6.3. Aim and Objectives 113

6.4. Methods 114

6.4.1. Study Design and Participants 114

6.4.2. Measurements and Exclusions 114

6.4.3. Statistical Analysis 116

6.5. Results 118

6.5.1. Suitability of UK Biobank for PheRS Application 118

6.5.2. PheRS Analysis Replicated Known Associations for Genes Linked to Familial Forms of Dyslipidemia, Hyperbilirubinemias and Alzheimer's Disease 120

6.5.3. Eight of 12 Significant Variant-PheRS Associations Were Not Due to Participants Diagnosed With the Corresponding IEM 123

6.5.4. Significant Variant-PheRS Associations Implicated Metabolites in Disease and Highlight the Specific Effects of *UGT1A1* Variant Rs1976391 on Bilirubin Levels 125

6.6. Discussion 126

6.6.1. Summary of Study Findings 126

6.6.2. Novelty of Findings 126

6.6.3. Study Strengths and Limitations 129

6.6.4. Conclusions 130

CHAPTER 7: GENERAL DISCUSSION 131

7.1. Summary of Findings 131

7.2. Strengths 132

7.3. Limitations 133

7.3.1. Generalisability of Findings 133

7.3.2. Use of GWAS Summary Statistics to Assess Rare Variant Effects on Metabolites and Phenotypes 134

7.3.3. Suitability of Population-based Cohorts to Assess Variant Effects on IEM-related Conditions 134

7.3.4. Reverse Causality As a Possibility in the Absence of Formal Mediation and Causality Assessment 135

7.3.5. Relative Importance of Genetic and Non-genetic Influences on Metabolite Levels and T2D Risk 135

7.4. Future Work 135

7.4.1. Replication of Analyses in Populations of Different Ancestry and Assessment of Sex-specific Effects 135

7.4.2. The Need for Additional Studies and Genetic Sequencing Technologies to Characterise Rare Variant Effects 136

7.4.3. Replication of Analyses in Cohorts With Larger Genotypic, Metabolomic and Phenotypic Datasets 136

7.4.4. The Need for Formal Mediation or Causal Inference Assessment to Establish the Role of Metabolite Levels in Disease Aetiology 137

7.4.5. The Need for Integrated Genetic and Phenotypic Information to Assess the Relative Importance of Genetic and Non-genetic Influences on Metabolite Levels and T2D Risk 137

7.4.6. Extension of the IEMs Study to Phenotypically Characterise Variants at Genes Known to Cause Other Rare, Mendelian Disorders 138

7.5. Clinical Implications 138

7.5.1. Potential Applications of Candidate Mediators As Biomarkers in T2D Diagnosis, Prevention and Management 138

7.5.2. Future Potential Applications of Variant-metabolite-phenotype Maps Detected in the IEMs Study 140

APPENDIX 143

1. Supplementary Information 143

2. Supplementary Figures 144

3. Supplementary Tables 156

REFERENCES 192

ABBREVIATIONS AND TERMINOLOGY

95% CI: 95% Confidence Interval

adj.: Adjusted for

ADR: Adverse Drug Reaction

BCAAs: Branched-Chain Amino Acid(s)

BMI: Body Mass Index

BF: Bayes Factor

BVS: Bayesian Variable Selection

CDCV: Common Disease, Common Variant hypothesis

CDRV: Common Disease, Rare Variant hypothesis

EA: Effect Allele

EAF: Effect Allele Frequency

EHR: Electronic Health Record

EPIC-Norfolk: The European Prospective Investigation into Cancer Norfolk

FDR: False Discovery Rate calculated using the Benjamini-Hochberg method

GGM: Gaussian Graphical Model(ling)

GIM: Genetically Influenced Metabotype

GRS: Genetic Risk Score

GWAS: Genome-Wide Association Study

HPO: Human Phenotype Ontology

HR: Hazard Ratio

ICD-X: International Classification of Diseases (-version X, where X is a number)

IEM: Inborn Error of Metabolism

IEM gene: A gene at which rare mutations are known to cause an IEM

IEM gene-linked locus: A locus that contains a variant that is associated with metabolite levels and that is linked to an IEM gene

IEM gene-linked variant: A variant that is associated with metabolite levels and that is linked to an IEM gene. IEM gene-linked variants may or may not be pathogenic for the IEM they are linked to.

IFV: IEM Familiar Variant, i.e. a variant linked to an IEM that is associated with one or more metabolites affected in the corresponding IEM. IFVs may or may not be pathogenic for the IEM they are linked to

IR: Insulin Resistance

IRadjBMI: Insulin Resistance adjusted for BMI

LC-MS(/MS): Liquid Chromatography coupled with (tandem) Mass Spectrometry

LD: Linkage Disequilibrium

MAF: Minor Allele Frequency

mGWAS: Metabolome GWAS

Metabolon mGWAS: The GWAS of plasma metabolite levels comprising of a meta-analysis of participants in the EPIC-Norfolk and INTERVAL prospective cohorts, as described in **Chapter 3**.

MR: Mendelian Randomisation

NMR: Nuclear Magnetic Resonance spectroscopy

OMIM: Online Mendelian Inheritance in Man

OA: Other Allele

OR: Odds Ratio

PK: Pharmacokinetics

PP: Posterior Probability

$\mathrm{PP}_{\mathrm{alignment}}$: Posterior Probability of Alignment, derived using the ‘HyPrColoc’ method

$\mathrm{PP}_{\mathrm{coloc}}$: Posterior Probability of Colocalisation, derived using the ‘coloc’ method

$\mathrm{PP}_{\mathrm{explained}}$: Proportion of Posterior Probability of alignment Explained by a genetic variant, derived using the ‘HyPrColoc’ method

$\mathrm{PP}_{\mathrm{regional}}$: Posterior Probability of Regional alignment, derived using the ‘HyPrColoc’ method

PheRS: Phenotype Risk Score

PheWAS: Phenome-Wide Association Study

Research: The full body of research performed in this thesis.

SD: Standard Deviation

SE: Standard Error

SNP: Single Nucleotide Polymorphism

Study: The analyses performed within a chapter

T2D: Type 2 Diabetes

WHR: Waist-to-Hip Ratio

WHRadjBMI: Waist-to-Hip Ratio adjusted for BMI

INDEX OF FIGURES

Chapter 1

Figure 1: Overview of glycolysis and the tricarboxylic acid cycle. **P19**

Figure 2: Summary of Garrod's hypothesis. **P27**

Chapter 2

Figure 1: Study design. **P32**

Figure 2: Distributions of weight change (kg per year) across the subcohort, T2D case-cohort controls, and cases. **P36**

Figure 3: Metabolite associations with weight gain and with T2D (FDR<0.05). **P37**

Figure 4: Sex-combined, cumulative metabolite adjustment of a Prentice-weighted Cox regression model of the effects of weight gain on T2D risk. **P40**

Figure 5: Associations of genetic scores for BMI, waist-to-hip ratio (adjusted for BMI) and insulin resistance (adjusted and unadjusted for BMI) with candidate mediators. **P41**

Figure 6: Proportion of variance in metabolite levels explained by 49 biomarkers, anthropometric measures and prevalent diseases. **P42**

Figure 7: Significant associations of candidate mediators and X-12063 with 27 incident diseases and all-cause mortality in 11,966 EPIC-Norfolk cohort participants ($p \leq 2.4x10^{-3}$). **P43**

Figure 8: Structure of compound 233-023, which differs from the structure of X-12063 by only one stereoisomer. **P44**

Chapter 3

Figure 1: Overview of the study aim. **P51**

Figure 2: Example illustrating how GIMs are defined. **P56**

Figure 3: Schematic of terminology used in this thesis. **P60**

Figure 4: Flowchart summarising results from analyses performed in this Chapter. **P61**

Figure 5: Metabolite matching of 241 conditionally independent variants within GIMs at IEM gene-linked loci. **P63**

Figure 6: Absolute effect size (per allele per 1-SD relative concentration of metabolite) by minor allele frequency. **P64**

Chapter 4

Figure 1: Overview of the study aim. **P71**

Figure 2: Cumulative proportion of variance explained across 237 metabolites associated with at least one IFV. **P76**

Figure 3: Most severe variant consequence annotations by Variant Ensembl Predictor across 187 IFVs and 589 metabolite-associated variants at other genes. **P77**

Figure 4: Carrier means plotted onto non-carrier metabolite level distributions for the top three strongest variant-extreme metabolite level associations ($p \leq 4.8x10^{-5}$) by allele frequency category. **P78**

Figure 5: Strongest variant-metabolite associations displaying significant departure from linearity at Bonferroni significance ($p \leq 1.5x10^{-4}$). **P80**

Figure 6: Comparison of metabolic and phenotypic consequences caused by rare mutations and more common variants at the *CPS1* and *ACADS* genes. **P82**

Figure 7: Association of the *UGT1A1* variants rs1976391 and rs201829156 with ICD-10 codes in the UK Biobank, as calculated using the TreeWAS method. **P85**

Chapter 5

Figure 1: Overview of the study aim. **P90**

Figure 2: Summary of distinct, unique phenotypes associated with IFVs ($p \leq 1x10^{-5}$) across GWAS databases. **P99**

Figure 3: Summary of IFV associations ($p \leq 1x10^{-5}$) with 87 clinical outcomes. **P100**

Figure 4: Comparison of metabolic and phenotypic consequences caused by rare mutations and more common variants at the *DBH* and *TH* loci. **P102**

Figure 5: Comparison of metabolic and phenotypic consequences caused by rare mutations and more common variants at the *OPLAH* locus. **P104**

Figure 6: Comparison of metabolic and phenotypic consequences caused by rare mutations and more common variants at the *ARG1* locus. **P106**

Chapter 6

Figure 1: Summary of PheRS generation for a specific IEM or rare disease. **P116**

Figure 2: Study design. **P119**

Figure 3: Diagnostic plots of 125 PheRSs. **P120**

Figure 4: Summary of significant IFV-PheRS associations (FDR≤0.05). **P122**

Figure 5: Enrichment assessment (χ^2 p≤0.05) of phecodes contributing to the PheRS for 'Crigler-Najjar syndrome, Type 1' in heterozygote and homozygote carriers compared to homozygotic non-carriers of the G-allele of the *UGT1A1* variant rs1976391. **P123**

Figure 6: Association of 1-SD increases in weighted metabolite GRSs with the odds of having high PheRS for 'Crigler-Najjar Syndrome, Type I' and 'Gilbert Syndrome'. **P126**

Appendix, Supplementary Figures

Ch2_Fig1: Comparison of metabolite associations with weight gain, BMI and incident T2D. **P144**

Ch2_Fig2: Individual candidate mediator adjustment of a Prentice-weighted Cox regression model of the effects of weight gain on T2D risk. **P145**

Ch2_Fig3: Sex-stratified, cumulative adjustment of candidate mediators selected using weight gain and T2D in a Prentice-weighted Cox regression model of the effects of weight gain on T2D risk. **P146**

Ch2_Fig4: Sex-combined, cumulative adjustment of candidate mediators selected using BMI and T2D in a Prentice-weighted Cox regression model of the effects of BMI on T2D risk. **P147**

Ch2_Fig5: Sex-stratified, cumulative adjustment of candidate mediators selected using BMI and T2D in a Prentice-weighted Cox regression model of the effects of BMI on T2D risk. **P148**

Ch2_Fig6: Associations of genetic scores for body fat percentage, liver fat, hip and waist circumference with candidate mediators. **P149**

Ch2_Fig7: GGM of the metabolic network. **P150**

Ch5_Fig1: Summary of IFV associations ($p \leq 1x10^{-5}$) with 306 complex traits and phenotypes not classified as clinical outcomes. **P151**

Ch6_Fig1: Enrichment assessment (χ^2 p≤0.05) of phecodes contributing to the PheRS for 'Gilbert syndrome' in heterozygote and homozygote carriers compared to homozygotic non-carriers of the G-allele of the *UGT1A1* variant rs1976391. **P152**

Ch6_Fig2: Enrichment assessment (χ^2 p≤0.05) of phecodes contributing to the PheRS for 'Alzheimer Disease 2' in heterozygote and homozygote carriers compared to homozygotic non-carriers of the C-allele of the *APOE* variant rs429358. **P153**

Ch6_Fig3: Enrichment assessment (χ^2 p≤0.05) of phecodes contributing to the PheRS for 'Alzheimer Disease 4' in heterozygote and homozygote carriers compared to homozygotic non-carriers of the C-allele of the *APOE* variant rs429358. **P154**

Ch6_Fig4: Enrichment assessment (χ^2 p≤0.05) of phecodes contributing to the PheRS for 'Alzheimer Disease 4' in heterozygote and homozygote carriers compared to homozygotic non-carriers of the C-allele of the *APOE* variant rs204474. **P155**

INDEX OF TABLES

Chapter 2

Table 1: Summary of candidate mediators. **P39**

Chapter 3

Table 1: Summary and comparison of methods used to map genes to metabolic loci. **P58**

Table 2: Evidence of enrichment was robust against different methods of causal gene annotation and varying locus definitions. **P62**

Chapter 5

Table 1: (Non-exhaustive) summary of examples where the metabolic and phenotypic consequences of common variants mimic those of rare, IEM-causing variants at the same gene. **P92**

Table 2: Overview of databases recording variant-phenotype associations reported in GWAS or population-based studies. **P94**

Table 3: Druggable IEM genes assigned to loci with metabolite-phenotype clusters specific for the relevant IEM. **P107**

Chapter 6

Table 1: Types of linked electronic health records in UK Biobank. **P115**

Table 2: IEMs linked to significant variant-PheRS associations and mapped to ICD-10 codes, as reported in Orphanet or OMIM. **P124**

Appendix, Supplementary Tables

Ch2_ST1: Candidate mediators selected using different metabolite sets prioritised for BVS analysis. **P156-159**

Ch2_ST2: Pairwise partial correlations of X-12063 with metabolites that share genetic locus associations. **P160**

Ch3_ST1: IFVs reported to cause the corresponding IEM and common enough to be detected in this study. **P161**

Ch5_ST1: Summary of phenotypes and the cohorts included in the discovery efforts of GWAS studies used in colocalisation analysis. **P162**

Ch5_ST2: Additional 24 IFVs included in phenome-wide assessment. **P163-164**

Ch5_ST3: Summary of phenotypes that were associated with loci containing IFVs and that were tested for a shared genetic signal using colocalisation. **P165-185**

Ch5_ST4: Summary results for novel colocalising metabolite-phenotype clusters. **P186-189**

Ch5_ST5: Clinical annotations for nine IFVs in the PharmGKB database. **P190**

Ch6_ST1: Summary statistics for the top 20 tested variant-PheRS associations with the largest odds ratios and the numbers of high PheRS ‘cases’ and ‘controls’ by variant carrier status. **P191**

CHAPTER 1: OPPORTUNITIES TO STUDY COMPLEX DISEASE AETIOLOGY USING UNTARGETED METABOLOMIC PROFILING IN POPULATION COHORTS

1.1 Introduction to the Metabolome

1.1.1. The Metabolome in Health and Disease

The human metabolome is the complete collection of small molecules, or ‘metabolites’, present within blood, saliva, urine or other fluid or tissue samples. Metabolites within the metabolome represent the downstream products of genes as well as substrates from the environment and thus represent a readout of the joint influences of genes as well as modifiable and non-modifiable factors. In human blood serum, as many as 25,424 metabolites have been detected, or are expected based on known gene functions and environmental sources[1].

The responsiveness of metabolites to genetic and non-genetic influences makes them useful as clinical markers for the diagnosis, prediction and prognosis of disease. For example, glucose is used as a source of energy by the human body and is an obligate fuel for the central nervous system[1]. Glucose is derived mainly from dietary intake of carbohydrates[2], but under low blood glucose conditions can also be synthesised by the genetically-regulated breakdown of glycogen[3] or of non-carbohydrate precursors such as lactate, amino acids and glycerol[4]. These different methods of acquiring glucose demonstrate independent effects of genetics and dietary behaviours as well as potential interactions that in combination result in the regulation of glucose synthesis and catabolism. Glucose is also an example of a metabolite that is an established clinical marker of disease, in this case, type 2 diabetes. Thus, glucose and other metabolites provide the opportunity to study ‘intermediate’ molecules that link genes and health-related behaviours to disease.

1.1.2. Measurement of the Metabolome

Metabolites are small molecules that are processed or produced during a metabolic reaction. Often, these metabolic reactions do not act in isolation, but instead take the products from an upstream reaction and convert them into other metabolites that are then used as substrates for another reaction. Metabolic reactions string together in this way to form metabolic pathways, and products from one pathway may be used as substrates for other metabolic pathways. For example, the molecule glucose is converted in one reaction into glucose-6-phosphate, which is then fed into another reaction to produce fructose-6-phosphate[2]. These reactions represent the start of the glycolysis pathway that cleaves glucose (a six -carbon molecule) into pyruvate (a three-carbon molecule) **(Figure 1)**. Pyruvate is then fed into the tricarboxylic acid cycle with the aim of producing ATP to meet cellular energy requirements[3] **(Figure 1)**.

Metabolites can be measured in all biofluids, though saliva, blood, or urine are the most frequently measured. These media are easy to collect (as they are already sampled in clinical procedures) and cheap to sample compared to other bodily fluids such as cerebrospinal fluid. The blood metabolome, which forms the primary route of the circulatory system that connects organs and tissues, has the additional advantage of providing a comprehensive summary of the metabolic status of an individual.

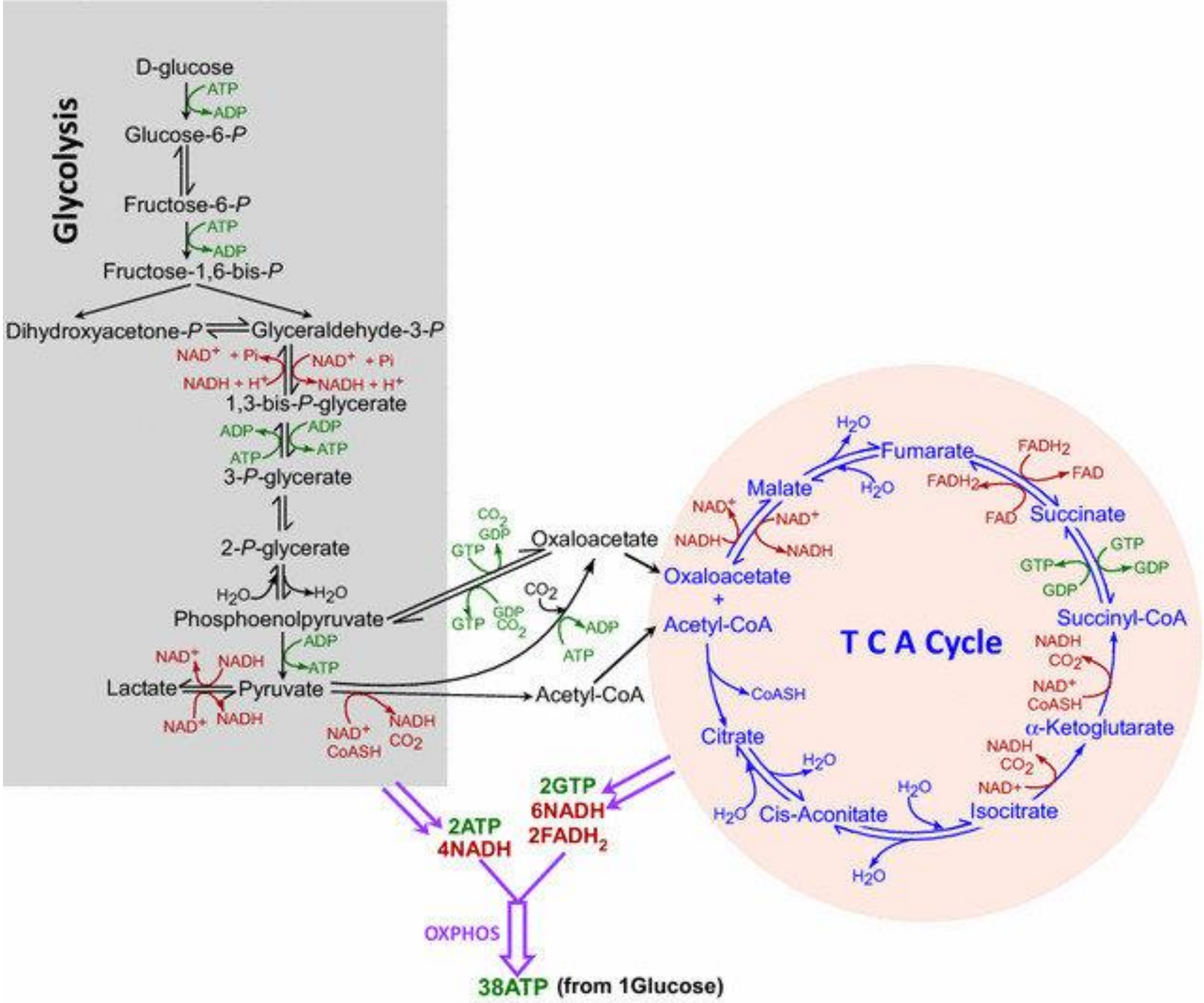


Figure 1: Overview of glycolysis and the tricarboxylic acid cycle. Production of ADP, ATP and cofactors including NADH and FADH are also shown. This figure was taken from a previous publication[4]. Abbreviations: Glucose-6-P, glucose 6-phosphate; Fructose-6-P, fructose 6-phosphate; Fructose -1,6-bis-P, fructose 1,6-bisphosphate; Dihydroxyacetone-P, dihydroxyacetone phosphate; Glyceraldehyde-3-P, glyceraldehyde 3-phosphate; 1,3-Bis-P-glycerate, 1,3-bisphosphoglycerate; 3-P-Glycerate, 3-phosphoglycerate; 2-P-Glycerate, 2-phosphoglycerate; OXPHOS, oxidative phosphorylation.

The large number of metabolites present in any given sample poses a challenge for measurement due to the widely differing chemical properties of metabolites (e.g. by polarity, molecular mass, or solubility). This hinders the development of biochemical assays to measure metabolites comprehensively and consistently across samples. Early studies thus aimed to measure defined sets of metabolites with known biological functions, such as amino acids or lipids. These ‘targeted’ profiling approaches enabled characterisation of the measured metabolites but neglected the potential biological relevance of other unmeasured metabolites and of between-metabolite correlations.

In recent years, high-throughput technologies leveraging the distinct chemical properties of metabolites have been developed to detect and identify as many metabolites within a sample as possible. In the field of biomedical research, gas or liquid phase chromatography coupled with mass spectrometry (GC-/LC-MS) is widely used for untargeted metabolomic profiling[5]. This method first separates molecules based on their solubility in a liquid or gas solvent called 'chromatography'[6]. Separation alone is insufficient when there are thousands of molecules in a sample, as molecules may have the same molecular weight by chance. Therefore, after initial separation by chromatography, mass spectrometry separates molecules further based on their mass and charge. First, molecules are fragmented into positively-charged ions by an ioniser and then propelled towards a negatively-charged plate at a speed that is proportional to the size of the ion[7]. Ions are detected as they pass through the detector, and the resulting spectrum can be analysed to identify composite molecules in a sample[7]. Other methods also exist to measure metabolite levels, though these are less relevant to the current thesis and are reviewed more comprehensively elsewhere[8].

Untargeted metabolomic profiling technologies have drastically improved metabolomic measurement compared to previous targeted approaches, with state-of-the-art techniques detecting over 1,000 metabolites in a single sample of blood plasma[9]. Though this is still far from a comprehensive measurement of the metabolome, these technologies enable the consistent detection of metabolites across samples, the assessment of between-metabolite correlations, and the implementation of untargeted metabolomic profiling in population-based studies to identify distinct metabolic pathways that contribute to complex disease.

1.2. Genetic Regulation of Metabolites

1.2.1. Genetic Regulation of Metabolite Levels in Rare Disease Patients

Much of the early understanding surrounding the genetic regulation of metabolite levels stems from patients with inborn errors of metabolism (IEMs) caused by rare mutations at a single gene resulting in the toxic accumulation or deficiency of metabolite levels that have severe phenotypic consequences[10]. The distinct and extreme metabolic and clinical consequences of IEMs have enabled their characterisation. This established IEM knowledge, combined with the burden imposed by IEMs on healthcare, enabled the development of newborn screening programmes to identify affected newborns and prevent disease onset[11,12]. One example of a gene known to cause an IEM is the phenylalanine hydroxylase (*PAH*) gene. *PAH* converts the amino acid phenylalanine into tyrosine[13], and mutations in this gene can cause a loss of function protein, which leads to a toxic build-up of phenylalanine. If untreated, this can lead to an IEM called phenylketonuria, which is characterised by intellectual disability, microcephaly and epilepsy[14]. Due to this inability to catabolise phenylalanine, individuals who are carriers of genetic risk variants for phenylketonuria are advised to avoid phenylalanine-rich foods such as milk, eggs, nuts and meat[15].

In recent years, rapid advances in genotyping and metabolomic profiling technologies have enabled the study of how genetic variation regulates metabolism in health and disease in population-based studies. These genome-wide associations studies (GWASs) of the metabolome[16–21] have identified associations between more common variation at IEM genes, including variants at the *PAH* gene[19–22], with metabolite levels known to be affected in the corresponding IEM. These findings suggest that more common variation at IEM genes may influence metabolite levels and have clinical consequences in the general population, though no systematic effort to date has been made to phenotypically characterise variants at IEM genes.

1.2.2. Genetic Architecture of Metabolite Levels in Population-based Studies

GWASs provide a robust and systematic method to characterise the genetic architecture underlying complex traits and diseases. In recent years, GWAS methodology, i.e. association analysis of genetic variants across the genome that are detected using genotyping arrays[23], has been used to study the genetic architecture influencing the human metabolome. As of January 2021, a total of 56 metabolome GWASs (mGWASs) had been published. A review of these studies is summarised here, and a comprehensive list of the mGWASs performed is available in previous publications and summaries[24,25].

Metabolome GWASs have provided novel insights into the regulation of metabolite levels by variation at IEM genes. Early studies, which performed targeted metabolomic profiling of amino acid and lipid species in cohorts with sample sizes of up to 5,000 participants, detected associations of variants at IEM genes known to affect amino acid and lipid metabolism, such as *LIPC*[16], *ACADS*[26], *ACADM*[26], *TAT*[27], and *PPM1K*[26,27]. These findings provided early evidence that common variants at IEM genes may influence levels of metabolites known to be affected in the corresponding IEMs. Furthermore, identification of these associations in cohorts of relatively modest sample size indicated that these variants could have large effect sizes on metabolite levels.

While several mGWASs have specifically measured targeted sets of amino acids and lipids[25], 16 have also successfully performed untargeted metabolomic profiling[19,20,35–40,22,28–34]. To date, untargeted metabolomic profiling methods have enabled the measurement of up to 644 unique metabolites[20]. In addition, increasing sample sizes, either from increased recruitment efforts or meta-analyses across cohorts or metabolomics measurement platforms, have enabled assessment of up to 80,000 participants[21,41]. Comprehensive measurement of the metabolome coupled with larger samples sizes have identified additional associations for variants at IEM genes and provided an opportunity to comprehensively assess the influence of variation at IEM genes on metabolic loci. In one study, 26 of 84 novel mQTLs were in the vicinity of an IEM gene, suggesting that the pool of genes influencing metabolite levels is enriched for those known to cause IEMs[19]. Whole genome and exome sequencing studies have also identified low-frequency and rare variants at IEM genes[20,22] that were associated with altered metabolite levels up to four standard deviations from the population mean.

Other study designs have shed insight into the effects of genetic variation on metabolite levels in the general population. For example, studies performed in twins-based cohorts showed that a large proportion of detected metabolites are heritable, though this varies by metabolite class[18,20,27]. GWAS analyses typically assume that genotypes have linear, dose-responsive effects on metabolite levels, though it has also been shown that some variants at IEM genes, such as the *ACADS* variant rs3916 and the *CPS1* variant rs715, display non-additive effects on metabolite levels that are also affected in the corresponding IEM[21].

GWASs of the metabolome have been conducted primarily in blood plasma samples, though emerging studies performed on urine and saliva samples have replicated many of the associations identified in blood[17,37,42] and identified biofluid-specific associations. These findings show that while blood is useful as a broad representative of metabolic processes, other biofluids may prove useful for research questions targeted at specific organ systems (e.g. the urinary metabolome, which is more reflective of kidney function[43]).

Variant-metabolite trait associations in GWASs often form clusters where genetic variants in high linkage disequilibrium (LD) are associated with metabolites that are functionally related, or which lie along the same metabolic pathway[44]. These clusters ('genetically influenced metabotypes'; GIMs) have three properties[24] in common: 1) the variance explained by common genetic variants is large, and variants often have large effect sizes; 2) GIMs can often be linked to an enzyme, transporter or metabolic regulator that is encoded at the genetic locus while the associated metabolites represent substrates or products of the encoded protein, and 3) GIMs are enriched for GWAS associations with clinical endpoints. These observations have been made in several metabolome GWASs[16,17,19] and further highlight the potential clinical relevance of genetic influences on metabolite levels.

While metabolome GWASs conducted to date provide much insight into the role of variation at IEM genes in influencing metabolite levels, a comprehensive identification of metabolite-associated variants at IEM genes has not been performed. In this thesis, I describe the results of a collaborative mGWAS effort performed in up to 20,000 participants with untargeted metabolomic profiling of 913 metabolites spanning eight metabolite classes, including structurally unidentified metabolites. This effort, which exceeds previous mGWASs in combining high metabolite coverage with large sample size, enables replication of previously reported variant-metabolite associations at IEM genes whilst providing an opportunity to identify additional, novel ones.

1.2.3. Metabolic Effects of Variation at IEM Genes May Translate into Health-related Effects in the General Population

GWASs of complex diseases and traits have identified specific instances where common variants in genes known to cause IEMs plausibly affect complex diseases and outcomes through smaller effects in the same metabolic pathway. For example, rare variants in the genes *APOB*, *LDLR* and *LPL*, which are involved in lipid metabolism, cholesterol biosynthesis and transport, can cause familial hypercholesterolemia, which is characterised by high circulating levels of cholesterol and other lipids[45–47]. GWAS studies have shown that common variants in these genes also lead to an increased risk of coronary artery disease, likely through milder effects on lipid metabolism[48]. The prevalence of variants at these genes, cost-effectiveness of genetic screening, and ability to treat have led NICE[49] to propose familial cascade screening[50] programmes to identify high-risk individuals for early prevention and management of these clinical outcomes.

GWASs of the metabolome have also identified additional common variants at IEM genes that are associated with the same metabolites and phenotypes as those known to be affected in the corresponding IEM. One example of an IEM gene for which common variants are associated with a complex, polygenic trait is the *PPM1K* gene. This gene encodes a serine/threonine phosphatase that activates the breakdown of the branched-chain amino acids (BCAAs) leucine, valine and isoleucine[51]. Rare variants in *PPM1K* are known to cause reduced activity or loss of function in the enzyme, resulting in elevated circulating concentrations of BCAAs, psychomotor retardation and metabolic decompensation[51]. By comparison, common variants near the *PPM1K* gene have been shown to induce mildly elevated levels of BCAAs (compared to what is seen for the IEM) that lead to an increased risk of insulin resistance and type 2 diabetes[52].

Another example is the IEM gene *GATM* that regulates creatine biosynthesis[53]. Rare variants in *GATM* cause Fanconi renotubular syndrome I, a disease that is characterised by increased levels of creatine, reduced levels of guanidinoacetate, renal tubular acidosis and osteomalacia[54]. Common variants in *GATM* have also been associated with reduced levels of guanidinoacetate, glomerular filtration rate and an increased risk of chronic kidney disease[55,56]. Thus, common variants in *GATM* may exert weaker effects on guanidinoacetate and creatine levels that impair renal function and metabolic health similarly to what is seen for rare patients with Fanconi renotubular syndrome I.

The above examples suggest that the metabolic effects of common variation across IEM genes may translate into health-related effects in the general population, though this has not been assessed systematically. Phenotypic characterisation of metabolite-associated variants at IEM genes may be conducted using different approaches, as described in the next section.

1.3. The Phenome

1.3.1. Genetic Architecture of Complex Traits and Diseases

While IEMs are caused by single or few mutations of large effect size at one gene, common diseases are driven by variants across multiple genes with small to moderate effect sizes[57,58]. Whether the majority of contributing variants are common (minor allele frequency (MAF)>0.05; the 'Common Disease, Common Variant' (CDCV) hypothesis)[59] or rare (MAF≤0.05; the 'Common Disease, Rare Variant' (CDRV) hypothesis[60,61]) is an ongoing subject of debate, with both hypotheses being supported by evidence from GWASs[62,63] and genetic linkage studies[64–66], respectively.

To date, GWASs provide a systematic and robust method for identifying genetic associations with disease. Despite this, an important caveat is that GWASs are more powered to identify common variant associations (due in part to their heavy multiple testing burden) and are therefore more likely to support the hypothesis that common diseases are driven by common variants. Whole genome and exome sequencing technologies, which are now being applied in population-based cohorts such as the UK Biobank[67], may identify additional rare variant associations in complex disease. Although early studies using these resources have found a limited number of rare variant effects[68,69], these technologies and methods will help to quantify the role of rare variation in complex disease in the future.

1.3.2. Approaches to Map Rare Disease Symptoms to Phenotypically Similar Common Diseases That Could Share the Same Aetiological Origins

A critical element of this thesis is the phenotypic characterisation of metabolite-associated variants at IEM genes. Although systematic phenotypic characterisation has not been performed to date, studies of rare and common diseases have shown that i) genetic loci associated with common diseases are significantly enriched for genes known to cause rare, Mendelian disorders[70,71], and ii) rare and common diseases associated with the same locus are significantly more likely to be phenotypically similar than not[71]. These findings suggest that variation at the same gene can contribute to a phenotype with varying degrees of severity.

One way of achieving systematic phenotypic characterisation is to conduct a phenome-wide association study ('PheWAS'). PheWAS can be performed within a single study, as demonstrated previously[72], though such an approach would require comprehensive measurement of multiple phenotypes and disease outcomes. Another simpler and faster method of performing phenome-wide assessment is to leverage results from GWASs conducted across thousands of complex phenotypes, as reported in databases of GWAS summary statistics[62,63,73,74]. Assessing the most powered GWASs for each phenotype maximises the power to detect known and novel phenotypic associations at a given genetic locus.

While phenome-wide assessment enables untargeted and comprehensive identification of phenotypic associations, it may also suffer from the detection of spurious associations. Specifically, ever-increasing sample sizes used in GWASs may also increase the likelihood of observing a significant variant association at a given significance threshold by chance. This could in turn lead to the detection of coincidental, shared associations of metabolites and phenotypes with the same genetic variant. To address this potential limitation, associated phenotypes could be selected using the well-documented clinical presentations of IEMs. This targeted approach is supported by previous studies[70,71] and enables the prioritisation of phenotypes that are likely to be driven by metabolite levels for downstream analyses.

The availability of standardised disease diagnoses and translation mappings across disease code systems (such as the Human Phenotype Ontology (HPO)[75] and International Classification of Disease Codes (ICD)[76]) enables an alternative approach to map rare disease symptoms to common diseases. In brief, phenotypes used to describe a rare disease are summarised into a single score known as a 'phenotype risk score' (PheRS)[77]. This approach has been used to identify the effects of rare variants at genes known to cause rare, Mendelian disorders on clinical endpoints in a hospital-based cohort[77] and is a strategy that I also explore in this thesis.

1.4. Challenges and Opportunities of Assessing the Effects of Changing Metabolite Levels on Complex Disease

In previous sections, I discussed how metabolites can be clinical biomarkers of prediction, diagnosis or prognosis of disease. Metabolomic profiling technologies have already provided opportunities to study the metabolome within large population-based cohorts. Challenges remain regarding the assessment of directions of association and causal effects, though these may be addressed by recently available datasets as well as methodological advances.

One challenge is that metabolite levels are regulated by genetic and environmental influences as well as other modifiable and non-modifiable risk factors. For example, high-calorie diets and sedentary lifestyles can lead to excess adiposity and to type 2 diabetes[78,79]. Environmental influences may also interact with genetic influences to modify disease risk. For example, genetic variation can alter where fat is stored in the body (for example, under the skin, around the waist area or around the gluteofemoral area)[80–82], the resulting differences of which have been shown to modulate the risk of type 2 diabetes[83]. The large number of potentially contributing factors makes it difficult to distinguish cause from effect in observational and prospective cohort studies, even when using statistical methods and adjusting for potential confounders. This challenge can now be addressed with the integration of genetics data with metabolomics and phenotypic data within the context of a single study, as genetics-based approaches have been shown to effectively assess directions of association as well as causal effects[84].

The importance of integrating genetic, metabolomic and phenotypic data is further emphasised when considering the strong correlations between genetic variation and metabolite levels. For example, mGWASs have shown that genetic loci may be specifically associated with one metabolite (e.g. *UMPS* locus with orotate levels[18]) or be highly pleiotropic (e.g. the *FADS1* locus, which has been associated with metabolite levels spanning diverse pathways of lipid metabolism[16]). The non-specificity displayed by loci such as the *FADS1* locus may hinder the mapping of metabolic changes, as well as downstream phenotypic consequences, to variant and gene function. Furthermore, metabolite levels are highly inter-correlated due to their connection via metabolic reactions and pathways, which limits the identification of representative metabolic pathways linking risk factors to disease. Yet these challenges can be countered with the systematic identification of genetic variants that independently co-regulate sets of correlated metabolites[44] as well as novel statistical methods that account for between-metabolite correlations[85,86]. Finally, the established metabolic and clinical sequelae of IEMs may help to map genetic variation to metabolic and phenotypic consequences with a high degree of confidence.

1.5. Thesis Overview

In this Chapter, I have summarised the opportunities presented by comprehensive metabolomic profiling in large-scale population-based cohorts or studies. Advances in untargeted metabolomic profiling technologies have enabled the successful characterisation of genetic associations across thousands of complex phenotypes as well as metabolic changes associated with disease. The research performed in this thesis aimed to link genetic, environmental, modifiable and non-modifiable risk factors to metabolic mechanisms contributing to complex disease.

In **Chapter 2**, I aimed to identify metabolic pathways linking weight gain to the development of type 2 diabetes. In this study, I combined genetic, metabolomic and phenotypic data with Bayesian statistical methods in a population cohort and case-cohort settings to establish directions of association and account for between-metabolite correlations. This integrative study design also enabled the assessment of genetic and environmental influences on levels of identified metabolic mediators of the association between weight gain and type 2 diabetes.

In subsequent chapters, I tested the hypothesis that variation at IEM genes have metabolic and health consequences mimicking those observed for rare mutations at the same genes **(Figure 2)**. Understanding of the metabolic and phenotypic consequences of variation at genes known to cause IEMs is unknown and therefore requires detailed systematic characterisation. Therefore, this study is divided across several chapters.

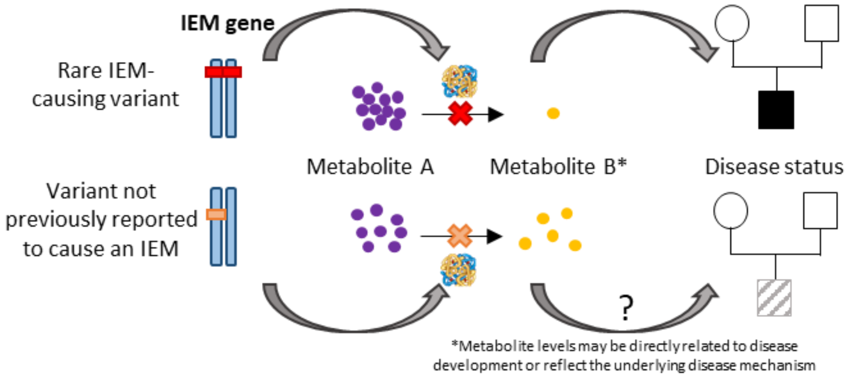

IEM gene
Rare IEM-causing variant
Variant not previously reported to cause an IEM
Metabolite A
Metabolite B*
Disease status
?
*Metabolite levels may be directly related to disease development or reflect the underlying disease mechanism

Figure 2: Summary of Garrod's hypothesis[2]**.** Rare, IEM-causing variants have large metabolic effects that cause the IEM, yet the clinical consequences of identified variation at IEM genes in metabolome GWAS studies remain unknown. This hypothesis is tested in this thesis.

In **Chapter 3**, I aimed to systematically identify and quantify the extent to which variation at IEM genes affect metabolite levels. For this study, I used the largest known mGWAS to date, which measured 913 metabolites spanning eight metabolite classes and structurally unidentified metabolites in up to 20,000 participants. Here, I used comprehensive causal gene annotation of metabolic loci and integrated IEM knowledge to systematically quantify the contribution of variation at IEM genes to metabolite levels. I also prioritised variants at these genes that were associated with metabolite levels known to be affected in the corresponding IEMs for in-depth metabolic and phenotypic characterisation.

Rare mutations are known to cause IEMs through extreme effects on metabolite levels. To test whether some of the variants prioritised in **Chapter 3** had large metabolic effects that could translate into observable phenotypes, I characterised them in terms of a) the proportions of variance on IEM-related metabolite levels explained, b) variant function, as predicted by the Ensembl Variant Effect Predictor (VEP)[87] annotation tool, c) effects on "extreme" metabolite levels (defined by reference guidelines[88–90] as the top or bottom 2.5th percentiles of the metabolite distribution), and d) non-additive effects on metabolite levels. These analyses, which were performed in **Chapter 4**, characterise the metabolic effects of prioritised variants according to characteristics that have previously been observed for rare mutations known to cause IEMs.

In **Chapter 5**, I further hypothesised that in-depth phenotypic assessment of the prioritised variants from **Chapter 3** could detect associated health effects in the general population. To test this, I performed a phenome-wide assessment using publicly available GWAS summary statistics to phenotypically characterise prioritised variants. Though this approach maximised the power to detect phenotypic associations, assessment of GWAS results across thousands of phenotypes could increase the likelihood of detecting a false positive association. To increase the specificity of the analysis, I prioritised phenotypes that were phenotypically similar to symptoms of the IEM that was implicated by the variant and its linked gene. The use of GWAS summary statistics for phenome-wide assessment also neglects the possibility that secondary causal signals could exist within the genetic region and drive phenotype associations independently of metabolic associations. Therefore, I used statistical colocalisation methods to test for shared genetic signals between IEM-related metabolic and phenotypic traits at genetic regions.

Rare disease symptoms can also be mapped to complex diseases using standardised disease codes and summarised in a score[77]. In **Chapter 6**, I therefore applied this approach in a population-based cohort setting and estimated the associations of prioritised variants from **Chapter 3** with the odds of having a high score for the corresponding IEM. To assess the role of IEM-related metabolite levels in significant variant-score associations, I also estimated the association between genetically-predicted levels of IEM-related metabolites with the corresponding scores.

Finally, in **Chapter 7** I discuss the strengths and limitations of the approaches I use in this thesis and discuss how my findings could translate into future clinical applications.

CHAPTER 2: IDENTIFICATION OF WEIGHT GAIN-ASSOCIATED METABOLIC PATHWAYS AND THEIR CUMULATIVE EFFECT ON INCIDENT TYPE 2 DIABETES RISK

2.1. Abstract

Background Weight gain and obesity lead to systemic metabolic changes that drive the global epidemic of type 2 diabetes (T2D). Here, I applied Bayesian methods to comprehensive profiling of the molecular blood signature of weight gain to systematically identify pathways that link weight gain to T2D development.

Methods I assessed 697 metabolites commonly-detected in baseline samples (1993-97) of a) randomly selected subcohort participants (n=11,972, 54% women) and b) a non-overlapping T2D case-cohort (n=1,503, 45% cases, 51% women) nested within the EPIC-Norfolk study. Bayesian variable selection identified independent candidate mediators significantly associated with weight gain before baseline and incident T2D. Prentice-weighted Cox regression was used to estimate the attenuation of candidate mediators to the weight gain-T2D association. I identified pathways indicated by candidate mediators using a data-driven metabolic network and test for genetic differences in adiposity-related traits to assess their functional role in T2D.

Results Average annual weight increased by 0.34 kg/year (standard deviation 0.27kg/year)between age 20 and baseline (mean age 60 years), with 93% subcohort participants gaining weight over time. Plasma levels of 529 metabolites (76%) were significantly associated with weight change. The strongest association was for an unknown molecule, X-12063 (beta±S.E.=0.37±0.01, $p<1x10^{-300}$), which we structurally identified as a steroid and named 'metabolonic lactone sulfate'. Of the 131 metabolites also significantly associated with incident diabetes, 20 were selected as candidate mediators (Bayes Factor>10) covering diverse biochemical pathways. Candidates reflected genetic susceptibility to adiposity and insulin resistance and individually accounted for little, but cumulatively for most of the increased risk of T2D associated with weight gain (T2D Hazard Ratio (95% CI) per 1-SD weight gain 2.79 (2.29;3.40) before versus 1.42 (1.13;1.79) after accounting for metabolite levels).

Conclusion Comprehensive untargeted metabolomic profiling identifies metabolic perturbations that may translate weight gain to T2D risk, including pathways affected by genetic susceptibility to adiposity and insulin resistance.

2.2. Introduction

Type 2 diabetes (T2D) is most commonly caused by weight gain and obesity and represents one of the largest health challenges globally[91]. Although over 90% of individuals with T2D are overweight or obese[92], the mechanisms through which weight gain and obesity lead to T2D are incompletely understood.

Recent advances in high-throughput profiling have enabled in-depth characterisation of the broad metabolic signature of obesity on small molecules circulating in human plasma, highlighting potential roles of lipids, aromatic and branched-chain amino acids[9,93–97]. Although these and other obesity-related metabolites have been related to T2D in independent studies[98–106], until now the systematic identification of metabolic pathways that are perturbed by weight gain and obesity and affect development of T2D has been difficult. This is because comprehensive untargeted profiling had not been performed at scale in prospective studies that assessed weight gain and were of sufficient size and duration to follow up for future T2D. Opportunities have arisen through the coming together of developments in a) high-throughput untargeted metabolomic technologies, b) statistical methods that account for metabolite correlation structures[86], c) the implementation/application of data-driven methods for pathway prioritisation[85], and d) genetic approaches to infer directions of effect and influence of T2D endophenotypes on identified molecular traits or signatures.

Here, I used untargeted profiling of the circulating plasma metabolome to identify pathways that link weight gain to future development of T2D. The context of a large-scale prospective population-based cohort enables assessment of specificity by integrating data on 27 incident diseases and to evaluate the role of genetic susceptibility to T2D endophenotypes on prioritised metabolic pathways.

2.3. Aim and Objectives

The aim of this study was to characterise the metabolic pathways mediating the association between weight gain and T2D. This aim was achieved through the following objectives:

1. To systematically characterise the metabolic profile of weight gain and of T2D;
2. To identify candidate mediators of the association between weight gain and T2D, and
3. To perform an *in silico* functional characterisation of candidate mediators.

2.4. Methods

2.4.1. Study design and participants

The European Prospective Investigation into Cancer and Nutrition (EPIC)-Norfolk study is a prospective cohort study of 25,639 individuals aged 40-79 years at baseline and recruited from 35 general practitioner practices across Norfolk in 1993-1997[107]. The study was approved by the Norfolk Research Ethics Committee (ref. 98CN01) and all participants gave signed informed consent.

The current study was designed to include two non-overlapping sets of EPIC-Norfolk participants: a) a subcohort of 11,972 participants drawn from EPIC-Norfolk participants with stored blood plasma sample who were not part of the T2D case-cohort study described below, and b) a T2D case-cohort nested within the EPIC-Norfolk cohort study, as previously been described in detail[107]. In brief, it includes 1,503 individuals (45% cases) ascertained using self-report, linkage to primary and secondary care, drug register, hospital admission, and mortality data[108].

2.4.2. Measurements

Weight at age 20 years ('initial weight') was based on self-reports at baseline. Height and weight were measured at baseline according to a standard protocol[107]. Weight change was calculated as the average annual difference between measured weight at recruitment (40-79 years of age) and self-reported weight at age 20 years and is referred to as weight gain for simplicity (see **Results**).

I selected non-fasted baseline plasma samples of 11,966 individuals (54% women) with a mean age of 60 years (SD: 6 years) for untargeted metabolomic measurement in a quasi-random design distributed across two equally sized measurement batches and of 1,503 participants from the T2D case-cohort (see **Appendix Supplementary Information**). Samples were selected and shipped for untargeted metabolite profiling using the Metabolon Inc. (Morrisville, North Carolina, USA) DiscoveryHD4^{TM} liquid chromatography tandem mass spectrometry (LC-MS/MS) platform, which measured up to 1,504 metabolites across 8 distinct metabolite classes in three batches. Of these, 697 metabolites that were commonly-detected (defined as being present in >85% of participants) were log-transformed, winsorised to five SDs to minimise the effect of outliers, and standardised (mean 0, SD 1) for analysis.

EPIC-Norfolk participants were genotyped using the Affymetrix UK Biobank Axiom^{TM} Array (ThermoFisher Scientific, Waltham, MA) and imputed to the Haplotype reference consortium v1.1[109], the 1000 Genomes project Phase 3[110] and UK10K[111] reference panels.

2.4.3. Statistical analysis

2.4.3.1. Exclusions

I included up to 10,666 and 1,344 (44% cases) participants of the subcohort and T2D case-cohort who had information on metabolite and covariate measurements, respectively **(Figure 1)**.

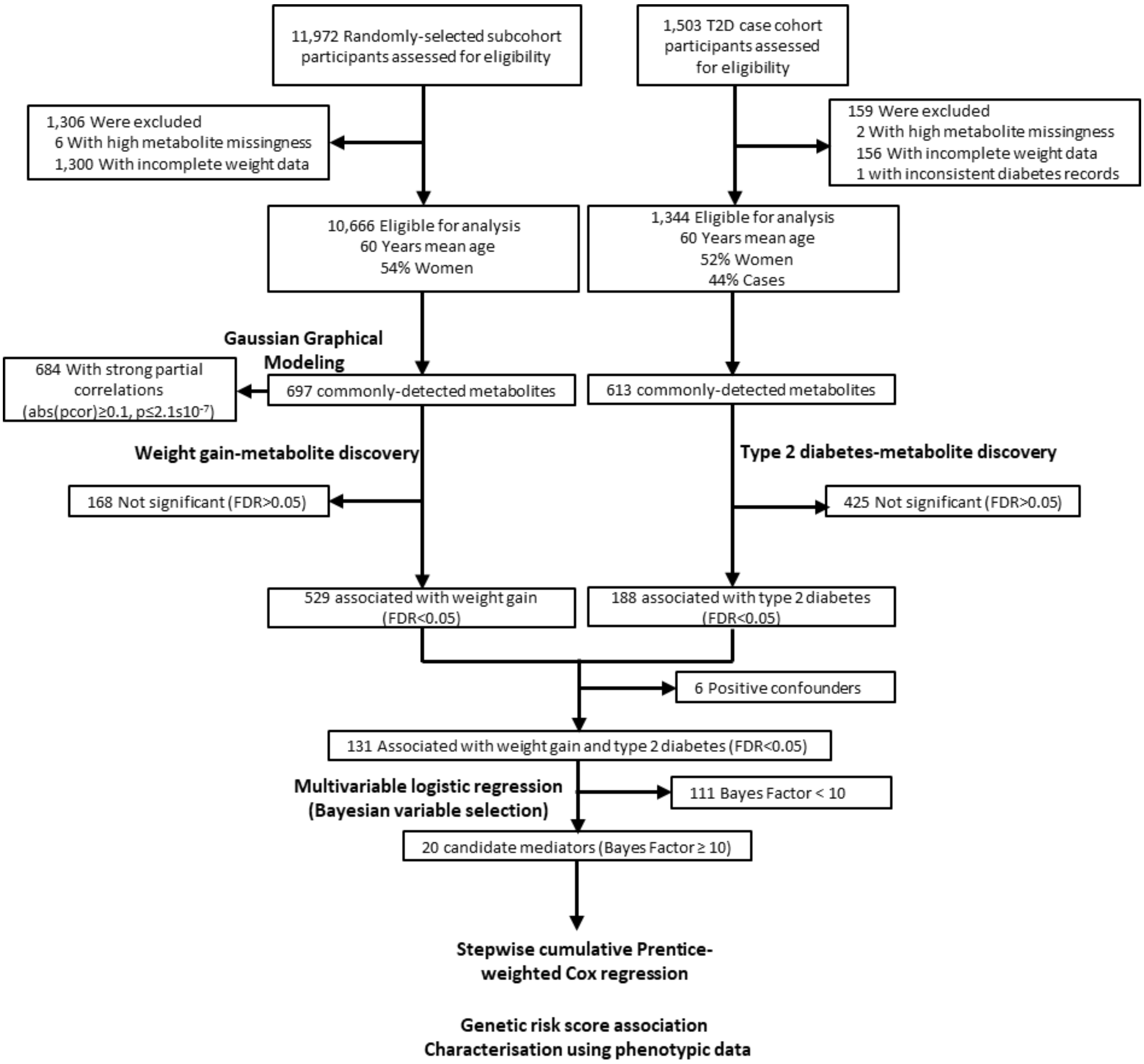

11,972 Randomly-selected subcohort participants assessed for eligibility
1,503 T2D case cohort participants assessed for eligibility
1,306 Were excluded
6 With high metabolite missingness
1,300 With incomplete weight data
159 Were excluded
2 With high metabolite missingness
156 With incomplete weight data
1 with inconsistent diabetes records
10,666 Eligible for analysis
60 Years mean age
54% Women
1,344 Eligible for analysis
60 Years mean age
52% Women
44% Cases
Gaussian Graphical Modeling
684 With strong partial correlations (abs(pcor)≥0.1, p≤2.1s10-7)
697 commonly-detected metabolites
613 commonly-detected metabolites
Weight gain-metabolite discovery
Type 2 diabetes-metabolite discovery
168 Not significant (FDR>0.05)
425 Not significant (FDR>0.05)
529 associated with weight gain (FDR<0.05)
188 associated with type 2 diabetes (FDR<0.05)
6 Positive confounders
131 Associated with weight gain and type 2 diabetes (FDR<0.05)
Multivariable logistic regression (Bayesian variable selection)
111 Bayes Factor < 10
20 candidate mediators (Bayes Factor ≥ 10)
Stepwise cumulative Prentice-weighted Cox regression
Genetic risk score association
Characterisation using phenotypic data

Figure 1: Study design. The number of individuals and commonly-detected metabolites present in each of the study cohorts, as well as analyses performed, are shown.

2.4.3.2. Association analysis of metabolites with weight gain and with T2D

To estimate the associations of weight gain (exposure) with plasma level of each metabolite (outcome) in subcohort participants, I used linear regression models adjusted for age, sex and weight at age 20 years. I used a linear regression model with the same covariates to estimate the association of baseline body mass index (BMI) with each metabolite and calculated the correlation between the estimated effect sizes.

In the non-overlapping EPIC-Norfolk T2D case-cohort, the association of metabolite levels (exposure) with incident T2D (outcome) was estimated using Cox regression models with age as the underlying scale and adjusting for sex, weight gain and initial weight.

I accounted for multiple testing by controlling the false discovery rate (FDR) at 5% using the Benjamini-Hochberg method[112].

2.4.3.3. Identification and assessment of candidate mediators

To identify a minimal set of informative metabolites out of 131 associated with weight gain, I used a Bayesian variable selection (BVS) procedure that combines multivariable logistic regression with a Prentice-weighted Cox regression[86] in the T2D case-cohort. The method does not allow for missing data and we therefore imputed missing values by replacement with a random number between zero and the lowest-measured intensity of the metabolite (zero-to-minimum value imputation). Imputed metabolite levels were then log-transformed, winsorised and standardised.

Multivariable logistic regression was performed using the R package R2BGLiMS v0.1-08-11-2019[86] with incident T2D as the outcome, adjusted for mean-centered age, sex, height, weight gain and initial weight. Metabolites with a Bayes Factor (BF) above a strong threshold of ten were defined as candidate mediators. Robustness to metabolite selection through BVS was tested by assessing metabolite sets that were prioritised using different methods: a) prioritisation using a nominal significance threshold with T2D, b) removing one metabolite from pairs of highly-correlated (R^2>0.8) metabolites, and c) using BMI- and T2D-associated metabolites.

Prentice-weighted Cox models were used to assess the separate and cumulative attenuation of BVS-selected candidate mediators on the weight gain-T2D association, with age as the underlying time scale and adjusting for sex, height, and weight at age 20 years. Metabolites were sequentially entered into the model in order of highest to lowest BF to assess the effect of individual and cumulative adjustment for these candidate mediators. Analyses were repeated separately in men (n=648, 53% cases) and women (n=695, 35% cases) to assess potential sex differences. To assess whether candidate mediators were most effective at attenuating the weight gain-T2D association, I compared the obtained HR from this model with the mean HR across 10,000 equivalent models adjusting for 22 randomly selected metabolites that were significantly associated with weight gain and with incident T2D risk. Comparison was performed using a one-sample t-test.

2.4.3.4. Characterisation of candidate mediators

To investigate the specificity of candidate mediators and X-12063 for T2D, metabolite associations with the risk of 26 incident diseases (in addition to T2D) and all-cause mortality (defined in the **Appendix Supplementary Information**) were estimated using Cox regression models with age as the underlying time scale adjusting for sex in the EPIC-Norfolk subcohort. Significance in these analyses were assessed using a Bonferroni-corrected threshold adjusting for 21 assessed metabolites

($p<2.4x10^{-3}$) to minimise the chance of false negatives as we were interested in any possible overlap. The proportion of variance explained for each metabolite by 49 prevalent diseases, anthropometric, lifestyle or biomarker measurements was also calculated using data from EPIC-Norfolk subcohort, as previously described[113].

2.4.3.5. Genetic risk score association analyses

I constructed genetic risk scores (GRS) from genetic variants (or their proxies $R^2>0.8$) known to be associated with T2D endophenotypes or risk factors, including insulin resistance (adjusted and unadjusted for BMI)[114], BMI[115], waist-to-hip ratio[116], body fat percentage[117], liver fat[118], waist and hip circumference[81] for 8,787 unrelated participants with available genotype data. Of the 2,323 variants considered across eight scores, one was not captured and no proxy ($R^2\geq0.8$) could be identified. Linear regression models were used to estimate the association with plasma levels of candidate mediators as well as X-12063 adjusted for age, sex and the first four principal components (Bonferroni $p=2.4x10^{-3}$). The zero-to-minimum value imputed dataset used to identify candidate mediators was also used for this analysis.

2.4.3.6. Structural and functional assessment of X-12063

To achieve full structural characterisation of this unknown compound, X-12063 was isolated from 40 litres of human plasma. Approximately 25-50 µg of purified compound was obtained and analysed by LC-MS/MS and nuclear magnetic resonance spectroscopy (NMR) to generate a candidate structure. A synthetic route was devised to produce enough of the candidate compound (211-023) for full analysis by LC-MS/MS and NMR. Detailed description of the extraction, purification, analysis and structural elucidation of X-12063 are available on request. To gain additional structural insights, we further assessed the correlations of X-12063 with other compounds that shared associations with genetic regions encoding metabolic enzymes[19,20] and on location in data-driven metabolic networks[29]. These methods were performed by Metabolon Inc. (Morrisville, North Carolina, USA).

Results from published GWAS studies of the metabolome were used to identify the metabolic reactions that X-12063 was involved in[19,20]. Metabolites sharing variant associations with X-12063 in these studies were also tested for correlation with X-12063.

2.4.3.7. Gaussian graphical model (GGM) construction

To calculate a data-driven network, I used the R package mice version 3.6.0 to impute missing values for 697 metabolites present in the random subcohort with the multiple imputation by chained equations method[119] to preserve cross-metabolite dependencies. Mean partial correlations were calculated across each of 20 imputed datasets using the R package GeneNet v1.2.14[120–122], transformed using Fisher's z transformation using the R package psych v1.9.12.31[123], pooled using Rubin's rules[124], meta-analysed using the R package meta v4.12-0[125] and then back-transformed. P-values were calculated using the Fisher's z transformed partial correlations. We declared significant edges connecting metabolites as those passing a stringent Bonferroni-corrected p-value ($p \leq 2.1x10^{-7}$) and having an absolute partial correlation >0.1. We visualised the network using Cytoscape v3.6.1[126]. We computed node degree and node centrality using the R package igraph (v1.2.6)[127] and compared these metrics between candidate mediators and other metabolites in the network using a Kruskal-Wallis test.

All statistical analyses and graphics were performed and produced using R version 3.5.3.[128] and STATA version 14.2[129].

2.5. Results

2.5.1. Metabolites Associated With Weight Gain

A total of 9,899 subcohort participants (93%) gained weight over time (mean annual change ± SD=0.34±0.27 kg/ year). Weight change was similar in the cohort portion of the T2D case-cohort (0.34±0.28 kg/year, p=0.59) **(Figure 2)**, but higher in incident cases, i.e. individuals who later developed T2D (0.53±0.35 kg/year, $p=3x10^{-33}$).

Of 697 metabolites commonly-detected in the subcohort, plasma levels of 529 metabolites across all eight biochemical classes were significantly associated with weight gain, including 306 positive and 223 inverse associations **(Figure 3)**. The strongest association with weight gain by far was identified for a hitherto unknown metabolite X-12063 (beta±SE=0.37±0.01, $p<1x10^{-300}$). Associations observed between weight gain and metabolites were strongly correlated with those observed for baseline BMI **(Appendix Ch2_Fig1)**, indicating that attained BMI strongly reflects weight gain ($R^2=0.99$, $p<2.2x10^{-16}$).

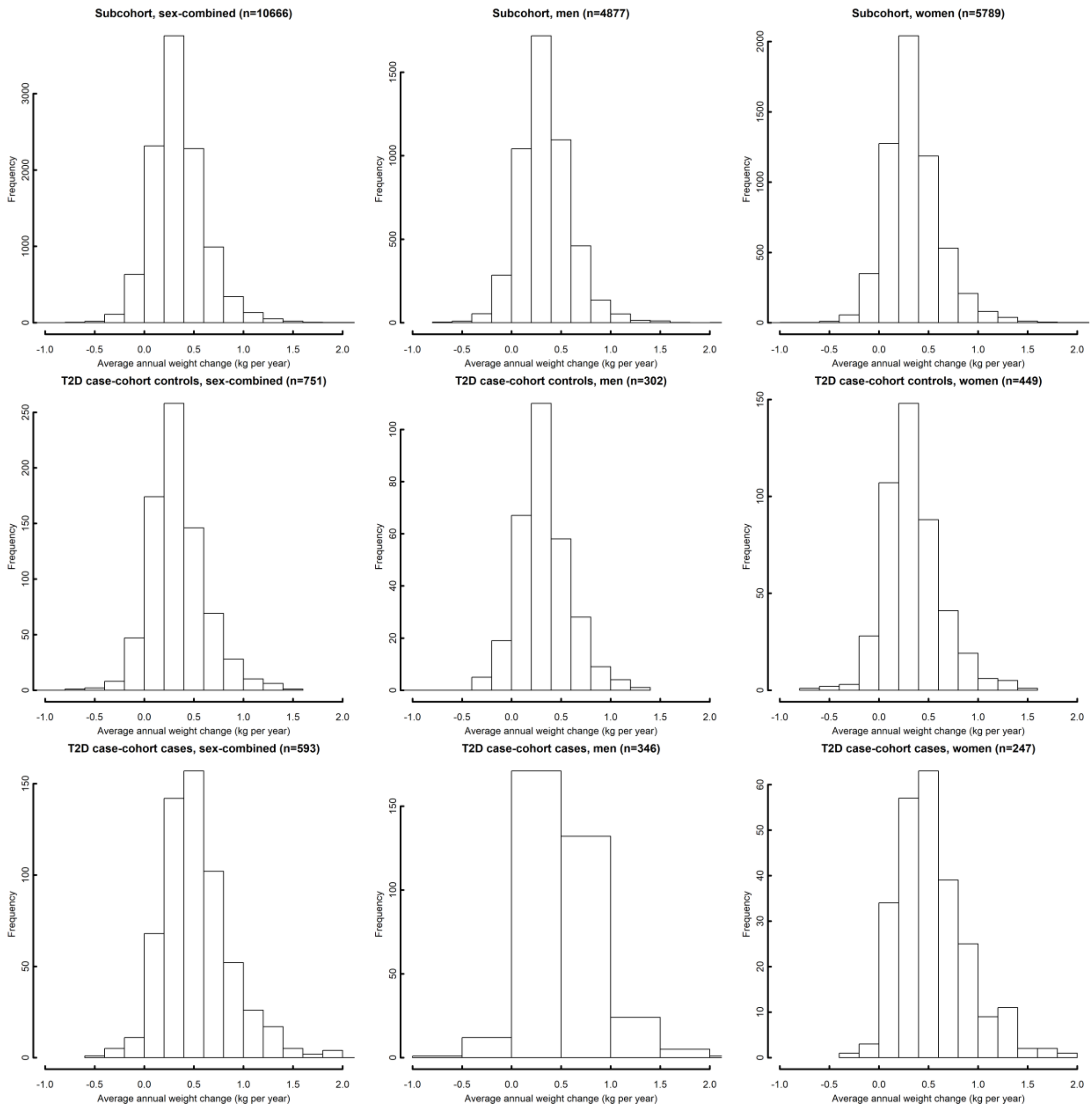


Figure 2: Distributions of weight change (kg per year) across the subcohort, T2D case-cohort controls, and cases. Sex-combined and sex-stratified distributions of weight change are also shown by cohort.

2.5.2. Discovery of Metabolites Associated With T2D

A total of 188 metabolites were associated with incident T2D (FDR<0.05) **(Figure 3)**; of these, mannose (HR (95% CI)= 3.21 (2.57;4.01), p<1x10^{-300}) and (random) glucose (HR (95% CI)= 2.37 (1.93;2.91), p=2.2x10^{-16}) were most strongly-associated. Of the 188 metabolites, 131 had concordant directions of association between weight gain and T2D and were assessed as potential mediators **(Figure 3)**.

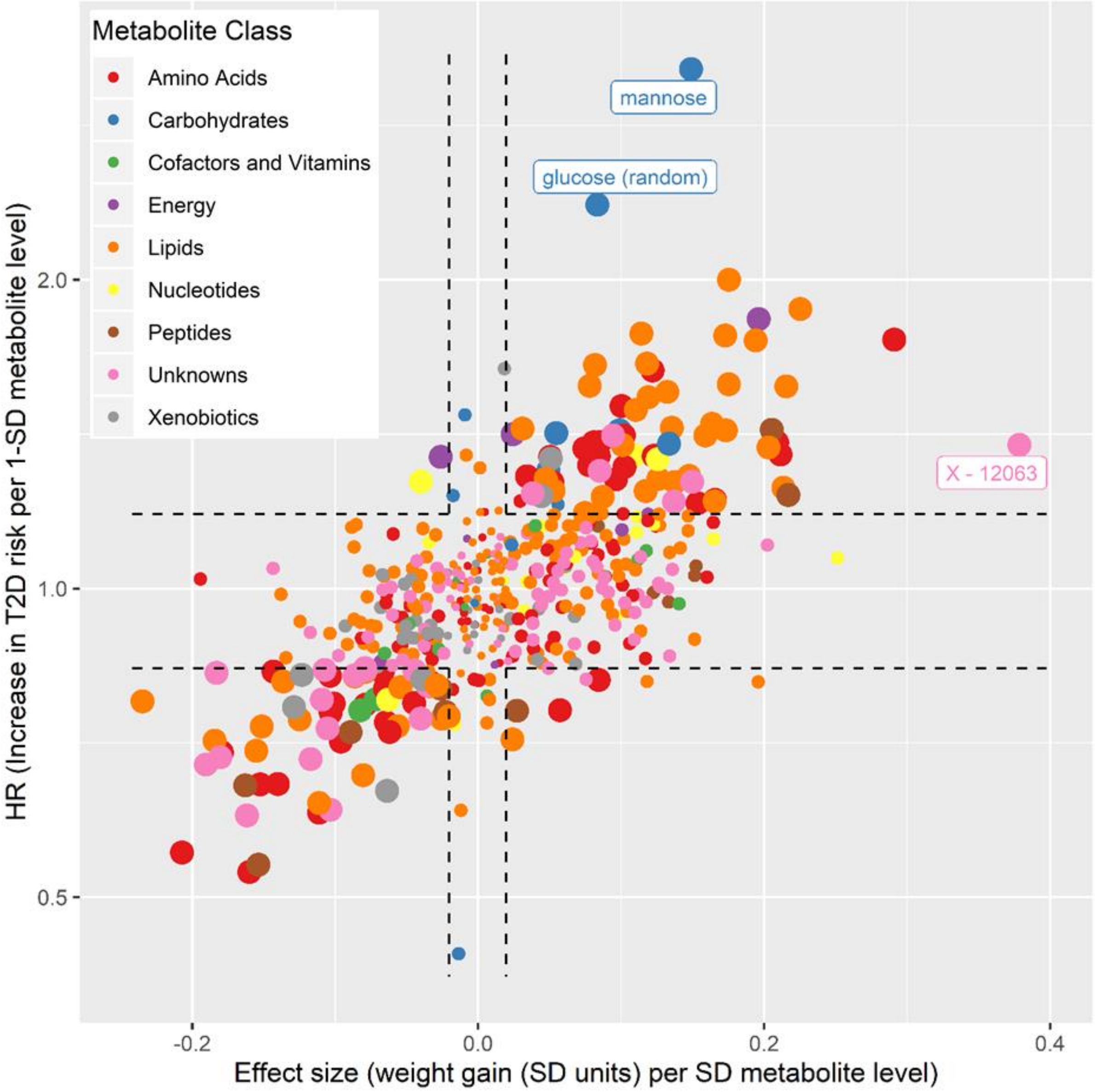


Figure 3: Metabolite associations with weight gain and with T2D (FDR<0.05). Large points in the upper right and lower left quadrants represent metabolites associated with both measures that have concordant directions of effect while smaller points represent those associated with only one measure.

2.5.3. Candidate Mediators Strongly Attenuated the Association of Weight Gain With Incident T2D

Of 131 considered metabolites, 20 were prioritised as candidate mediators of weight gain and T2D (BF≥10) **(Table 1).** These originate from diverse pathways including carbohydrate metabolism (mannose, erythronate* and lactate), proteinogenic amino acids (glutamate, histidine), *N*-acetylated amino acids (*N*-acetylglycine, *N*-acetylaspartate, *N*-acetylmethionine, *N*-trimethyl-5-aminovalerate), vitamin C catabolism (threonate), (lyso)phospholipid metabolism (2-linoleoyl-GPC (18:2)*, 1-palmitoyl-2-linoleoyl-GPE (16:0/18:2)), triglyceride degradation (1-palmitoylglycerol (16:0)), fatty acid metabolism (2-hydroxystearate, 2-hydroxypalmitate), steroid metabolism (pregn steroid monosulfate*) and purine metabolism (xanthine), as well as one unknown compound, X-21258, with less obvious references to biological pathways. Results were robust to changes in metabolite sets **(Appendix Ch2_ST1)**.

Weight gain was associated with an increased risk of T2D over the follow up time (Hazard Ratio (HR) 95% Confidence Interval (95% CI) per SD weight change/year 2.79 (2.29;3.40), p=1.7x10^{-24}). Individually, 20 candidate mediators accounted for little **(Appendix Ch2_Fig2)**, but cumulatively for most of this increased risk (HR (95% CI) 1.42 (1.13;1.79), p=2.5x10^{-3} after adjustment for the strongest 18 metabolites) **(Figure 4)**. The cumulative attenuated HR was significantly lower than the mean HR obtained across 10,000 models adjusting for randomly selected sets of 20 T2D- and weight gain-associated metabolites (mean HR (95% CI) 2.06 (2.056;2.063), p≤2.2x10^{-16}). In women, the association between weight gain and T2D was not significant after accounting for all candidate mediators (HR (95% CI) 2.41 (1.89;3.07), p=1.2x10^{-12} before versus 1.28 (0.96;1.69), p=0.09 after adjustment). In men, corresponding associations were 3.57 (2.78;4.57), p=1x10^{-23} before versus 1.60 (1.15;2.23), p=5.5x10^{-3}) after adjustment for all candidate mediators **(Appendix Ch2_Fig3)**. Similar results were obtained for BMI instead of weight gain **(Appendix Ch2_ST1d; Ch2_Fig4-5)**.

Table 1: Summary of candidate mediators. *metabolite assignment was made with high confidence but is not definite.

Metabolite	Pathway	Class	Bayes Factor
Mannose	Fructose, Mannose and Galactose Metabolism	Carbohydrate	Inf
N-acetylglycine	Glycine, Serine and Threonine Metabolism	Amino Acid	43536
1-palmitoylglycerol (16:0)	Monoacylglycerol	Lipid	8842
2-hydroxystearate	Fatty Acid, Monohydroxy	Lipid	1239
Glutamate	Glutamate Metabolism	Amino Acid	675
N-acetylaspartate (NAA)	Alanine and Aspartate Metabolism	Amino Acid	106
N-delta-acetylornithine	Urea cycle; Arginine and Proline Metabolism	Amino Acid	68
X - 21258	Unknown	Unknown	66
N-acetylmethionine	Methionine, Cysteine, SAM and Taurine Metabolism	Amino Acid	54
Lactate	Glycolysis, Gluconeogenesis, and Pyruvate Metabolism	Carbohydrate	37
Pregn steroid monosulfate*	Steroid	Lipid	36
Pyroglutamine*	Glutamate Metabolism	Amino Acid	26
Erythronate*	Aminosugar Metabolism	Carbohydrate	21
1-palmitoyl-2-linoleoyl-GPE (16:0/18:2)	Phospholipid Metabolism	Lipid	16
Threonate	Ascorbate and Aldarate Metabolism	Cofactors and Vitamins	15
2-linoleoyl-GPC (18:2)*	Lysolipid	Lipid	15
Xanthine	Purine Metabolism, (Hypo)Xanthine/Inosine containing	Nucleotide	15
2-hydroxypalmitate	Fatty Acid, Monohydroxy	Lipid	13
N-trimethyl 5-aminovalerate	Lysine Metabolism	Amino Acid	11
Histidine	Histidine Metabolism	Amino Acid	10

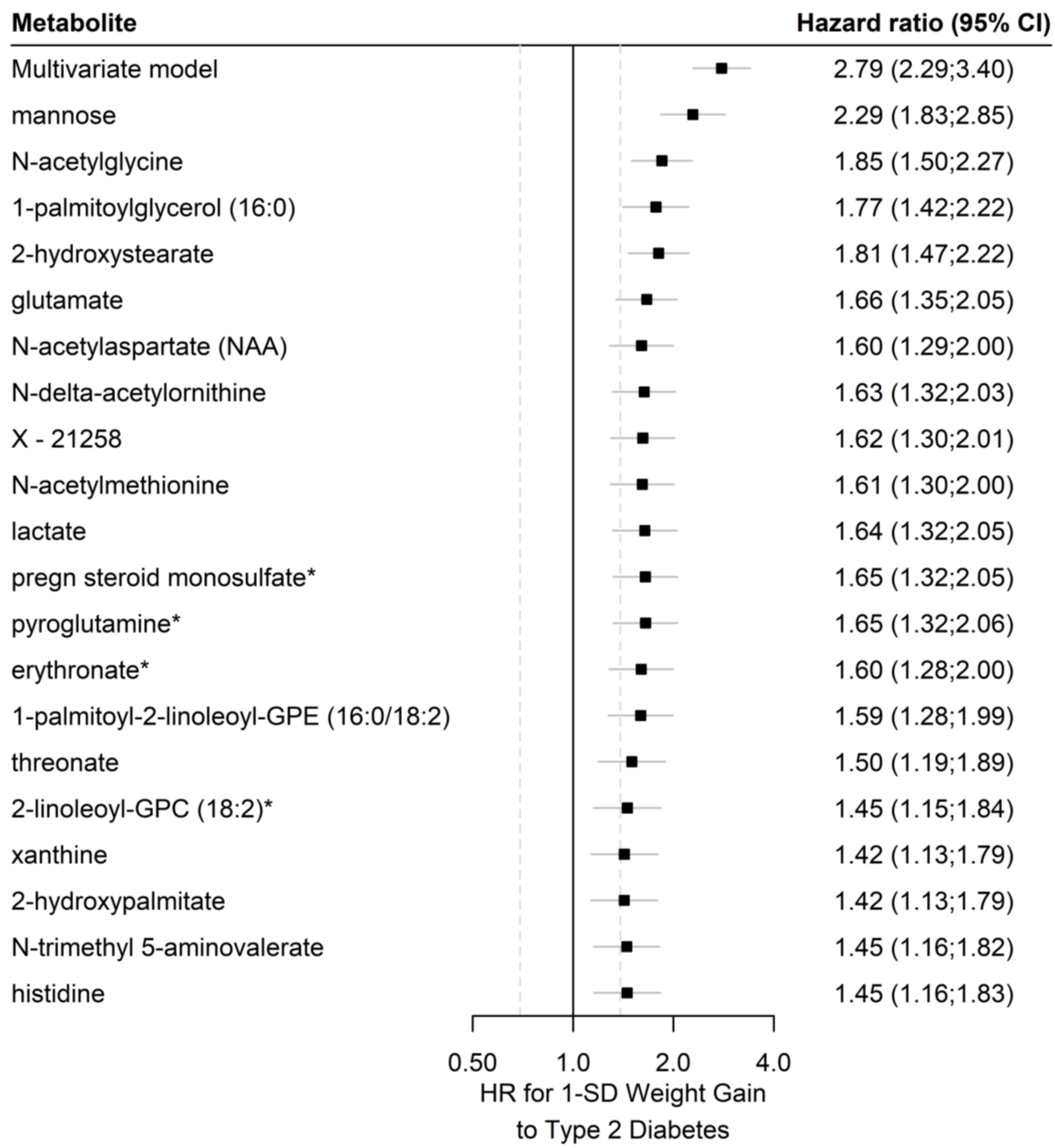


Figure 4: Sex-combined, cumulative metabolite adjustment of a Prentice-weighted Cox regression model of the effects of weight gain on T2D risk.

2.5.4. Genetic and Other Non-modifiable and Modifiable Factors Influenced Metabolite Levels

Genetic susceptibility to T2D endophenotypes and risk factors had strong influences on metabolite levels. For example, GRSs for BMI and insulin resistance were strongly associated with seven candidate mediators, including inverse associations with amino acids such as *N*-acetylglycine and a positive association with mannose **(Figure 5; Appendix Ch2_Fig6)**. Plasma levels of X-12063 were significantly higher in individuals with increased genetic risk for higher BMI, WHR and insulin resistance **(Figure 5)**.

Plasma levels of glutamate were also positively associated with increased genetic risk for higher waist and hip circumference **(Appendix Ch2_Fig6)**.

I found that biomarkers of lipid metabolism, liver and renal function explained a significant proportion of the variance in metabolite levels (p-values≤2.4x10^{-3}) **(Figure 6)**. Many of these pointed to a role for specific lifestyle behaviours. For example, 58.7% of threonate was explained by plasma vitamin C levels, highlighting the strong influence of modifiable behaviours. Notably, apart from mannose, plasma levels of none of the candidate mediators were considerably explained by markers of glucose metabolism, highlighting the diverse and complementary nature of the identified pathways.

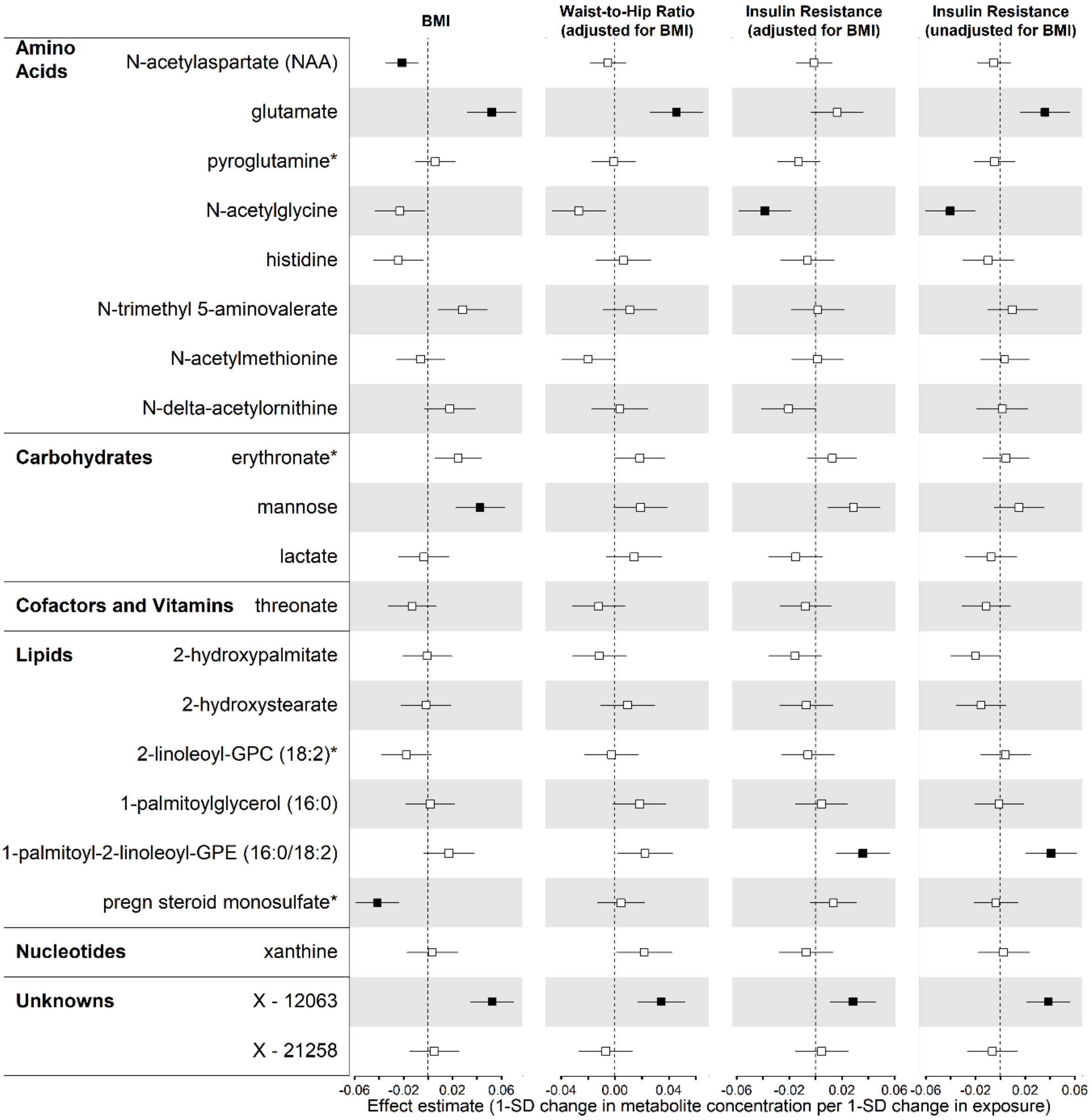


Figure 5: Associations of genetic scores for BMI, waist-to-hip ratio (adjusted for BMI) and insulin resistance (adjusted and unadjusted for BMI) with candidate mediators. X-12063 is included as a metabolite of interest. Significant associations (p≤2.4x10^{-3}) are filled in while non-significant associations are empty.

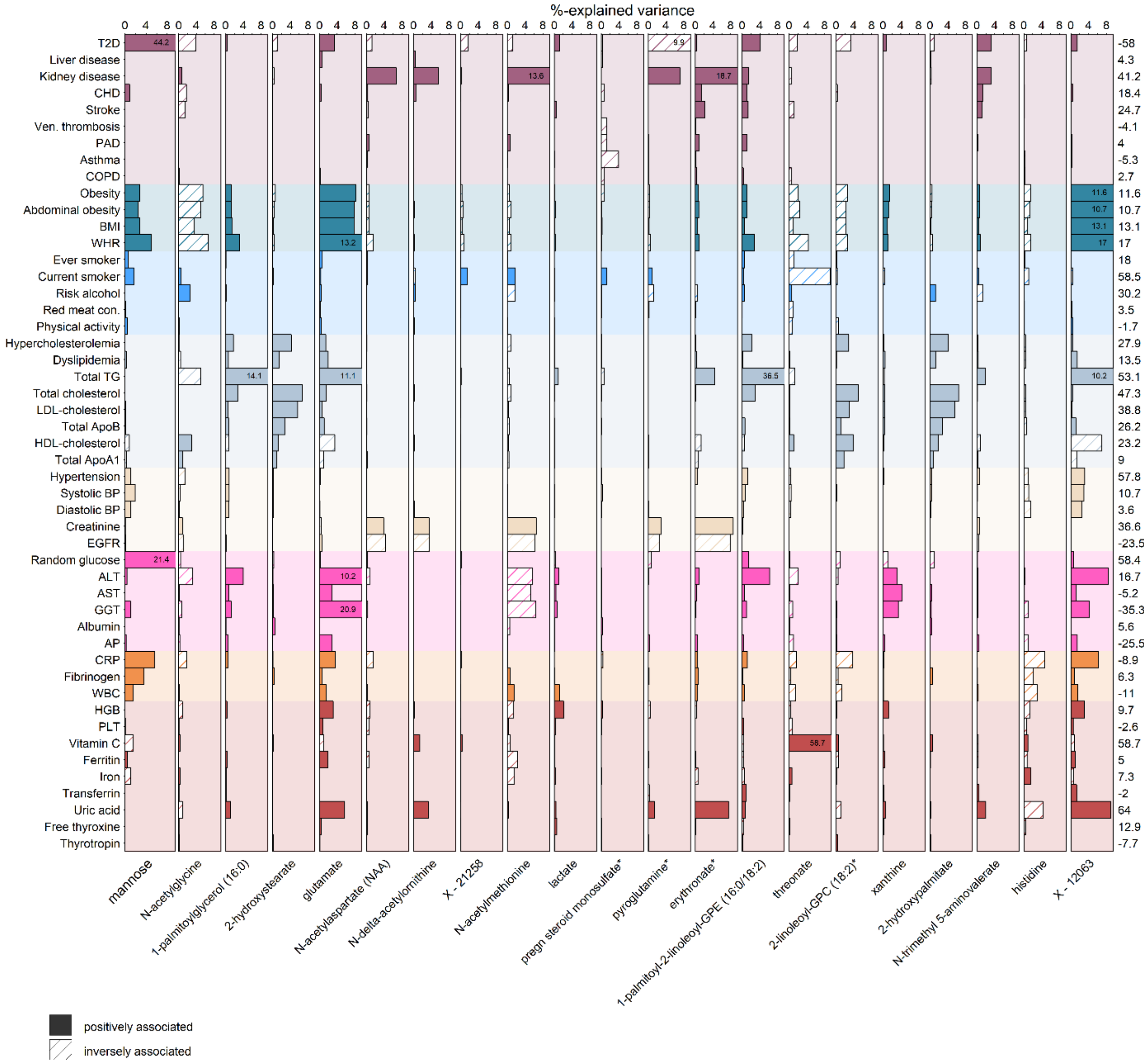


Figure 6: Proportion of variance in metabolite levels explained by 49 biomarkers, anthropometric measures and prevalent diseases. The number at the end of each row represents the maximum proportion of variance explained for any of the assessed metabolites, with the sign indicating direction of effect. X-12063 is included as a metabolite of interest. CHD=coronary heart disease; PAD=peripheral artery disease; COPD=chronic obstructive pulmonary disorder; TG=triglycerides; SBP/DBP=systolic/diastolic blood pressure; ALT=alanine aminotransferase; AST=aspartate aminotransferase; GGT=gamma-glutamyltransferase; AP=alkaline phosphatase; CRP=C-reactive protein; WBC=white blood cell count; HGB=haemoglobin; PLT=platelets.

2.5.5. Candidate Mediators Reflected Specific and Non-specific Pathways in T2D and Cardiometabolic Disease

Considering 26 incident diseases in addition to T2D and all-cause mortality, higher levels of candidate mediators were significantly associated with an average of 5.1 diseases (range: 0-10 diseases) ($p \leq 2.4 \times 10^{-3}$) **(Figure 7)**. I found plasma levels of 2-hydroxystearate, lactate, X-21258, and xanthine to be most specific for T2D, whereas other candidate mediators, and X-12063, were associated with several other cardiometabolic outcomes such as liver disease, heart failure and coronary heart disease

(Figure 7). Threonate was the most pleiotropic candidate mediator and was inversely associated with the risk of ten diseases.

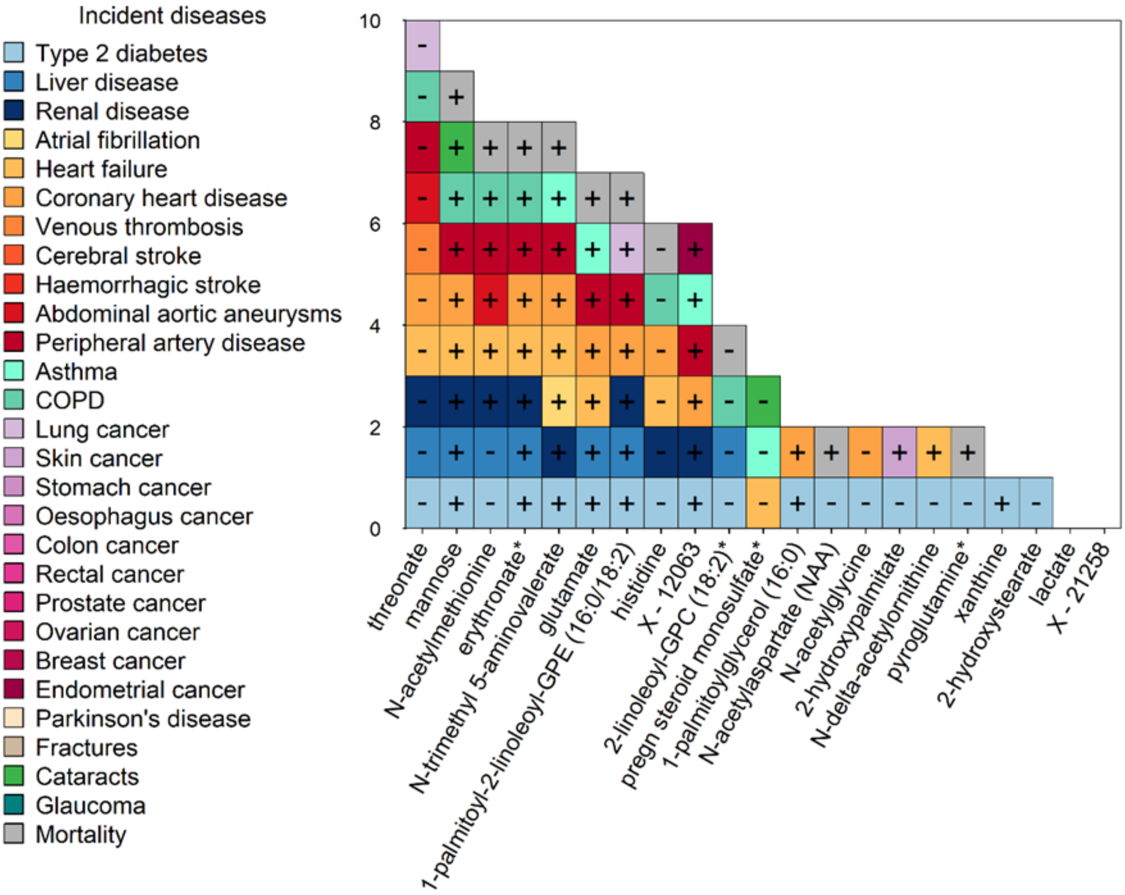


Figure 7: Significant associations of candidate mediators and X-12063 with 27 incident diseases and all-cause mortality in 11,966 EPIC-Norfolk cohort participants ($p \leq 2.4 \times 10^{-3}$). The y-axis represents the number of diseases associated with each metabolite. Directions of effect are indicated by the sign in each box.

2.5.6. X-12063 Represented Steroid Modification and Clearance Pathways in Cardiometabolic Disease

The top weight gain-associated metabolite, X-12063, was also associated with T2D (HR (95% CI)=1.38 (1.16-1.65), $p=3.4 \times 10^{-4}$), renal disease (beta±SE=0.12±0.036, $p<8.2 \times 10^{-4}$), coronary heart disease (beta±SE=0.097±0.024, $p<4 \times 10^{-5}$), peripheral artery disease (beta±SE=0.13±0.034, $p<1.4 \times 10^{-4}$), asthma (beta±SE=0.14±0.045, $p<1.3 \times 10^{-3}$) and endometrial cancer (beta±SE=0.49±0.13, $p<2.3 \times 10^{-3}$) **(Figure 8)**.

To identify the molecular identity of X-12063, structural profiling was performed in collaboration with Metabolon Inc. Analysis of the accurate mass data showed that X-12063 had a molecular formula of $C_{22}H_{36}O_6S$, with a negative ion accurate mass of 427.2161 m/z ± 5ppm and a characteristic neutral loss of a sulfate moiety, but little other structural information that allowed a match to a known compound. The low signal intensity of this sulfated compound in healthy individuals indicated that X-12063 was normally present at very low levels in plasma. Though the stereochemistry of endogenous X-12063

(Figure 8) could not be synthesised, a similar compound differing by one stereoisomer was produced to permit confirmation of proposed candidate structure. Given its structural identity, X-12063 was named as 'metabolonic lactone sulfate', and is hereafter referred to as such.

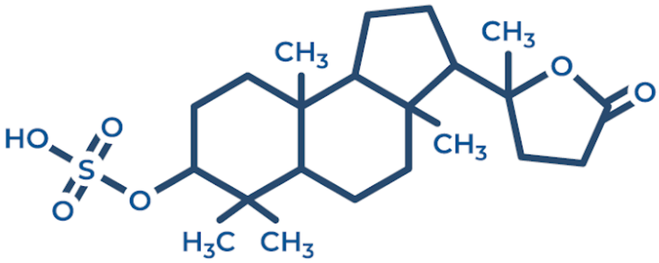


Figure 8: Structure of compound 233-023, which differs from the structure of X-12063 by only one stereoisomer.

I obtained support for this structural profiling from genome-wide metabolomic profiling association studies[19,20]. In these studies, genetic variants near genes known to be involved in sulfation or glucuronidation of steroid and bile acid species (*SLCO1B1*, *CYP3A5*, *CYP3A7*) were associated with plasma levels of metabolonic lactone sulfate. Accordingly, the highest correlation coefficients of metabolonic lactone sulfate were seen for steroid products of these genes **(Appendix Ch2_ST2)**[29].

2.5.7. Metabolic Network Visualisation Highlighted Dietary Origins of Candidate Mediators With Unidentified Chemical Identity

I created a data-driven metabolic network including 684 metabolites connected by 1,769 edges **(Appendix Ch2_Fig7)**. Candidate mediators were distributed widely across the network and did not represent hotspots within the network as neither node degrees (p=0.44) nor centrality measures (p=0.54) differed from the distribution across all metabolites in the network. The network helped to narrow down potential pathways that candidate metabolites may represent. For example, X-21258 belonged to a cluster of metabolites including 4-allylphenol sulfate, of likely exogenous origin. Metabolonic lactone sulfate also clustered with steroid metabolites such as glycocholenate sulfate* and 16α-hydroxy DHEA 3-sulfate **(Appendix Ch2_Fig7)**, corroborating its structural elucidation.

2.6. Discussion

2.6.1. Summary of Findings

Here I presented a systematic investigation into pathways that could potentially mediate the effects of weight gain on incident T2D risk. Integration of genetic, metabolomic and incident disease data within the context of a single large and prospectively designed cohort is a clear strength of this work, together with the opportunity to test specificity of associations across 27 incident diseases. Using BVS, I prioritised 20 “independent” candidate mediators that cumulatively accounted for most of the increased T2D risk associated with weight gain. I distinguished associations that are specific for T2D from those that show associations with other diseases and may be responsible for the broader detrimental health effects of weight gain that lead to obesity-associated multimorbidity. Candidate mediators were significantly more effective at attenuating the association between weight gain and incident T2D risk than randomly selected sets of metabolites but did not represent connection hubs in the metabolic network. I thus propose that identified candidate mediators represent specific yet diverse metabolic pathways that best capture the underlying pathology of T2D risk induced by weight gain.

2.6.2. Novelty of Findings

By integrating multiple layers of data, I identified pathways and certain domains of health at the intersection between weight gain and T2D onset. Amongst the 20 candidate mediators highlighted, I identified some that were well known (such as mannose, *N*-acetylglycine and glutamate[95,101,130]), but also others that were novel and some of which were specifically associated with incident T2D.

Candidate mediators represented metabolic pathways reflecting the diverse mechanisms underlying the effects of weight gain on T2D pathogenesis. Increased plasma levels of *N*-acetylated amino acids emerged as a hallmark of multiple cardiometabolic outcomes, and I identified *N*-acetylmethionine to exhibit an unexpected pattern, being inversely associated with T2D but positively with cardiovascular outcomes such as coronary heart disease. While the latter association may well relate to the accumulation of small molecules with impaired kidney function, the relation of *N*-acetylmethionine with T2D might be of hepatic origin, where it can be used as an alternative methyl donor for the generation of gluthathione. This action mimics the clinical application of *N*-acetylmethionine during liver toxicity caused by poisoning with acetaminophen or bromobenzene[131,132] and is in line with the inverse association seen with incident liver disease. Few metabolites or pathways they represent were specific for T2D with one of the exceptions being 2-hydroxystearate, which has been suggested as a proxy for *de novo* lipogenesis[133].

Several candidate mediators, such as *N*-acetylaspartate, *N*-delta-acetylornithine and erythronate, have been associated with glomerular filtration rate and kidney disease in other studies[134,135] and thus represent markers of kidney function. These associations are corroborated by the findings, which showed that variance in levels of these metabolites is explained in part by prevalent kidney disease, glomerular filtration rate and urate, another molecular marker of kidney function. Previous studies have suggested that elevated erythronate levels may be driven by the breakdown of glycated proteins or by vitamin C intake[134,135], though vitamin C did not explain a significant proportion of variance in erythronate levels in the current study. All three metabolites associated with kidney function were associated with cardiometabolic outcomes in this study, supporting previous findings that declining kidney function could be an early marker of cardiometabolic disease[136,137].

Strong genetic and other non-modifiable and modifiable determinants of candidate mediators were also identified. These determinants explained a large proportion of variance in candidate mediator levels and provided insight into directions of effect and potential strategies for behavioural and pharmacological interventions targeting these pathways. Specifically, several candidate mediators could be attributed to health-related behaviours that have been associated with incident T2D, including proxies for fruit consumption such as vitamin C intake[138] and vegetable and fibre intake[139,140]. Inclusion of dietary behaviours in lifestyle guidelines and T2D prevention programmes remains under discussion[141]. This evidence reinforces the narrative that while some metabolic pathways (e.g. amino acid metabolism, bile acid metabolism) are in part regulated by genetic variation, many of the effects of weight gain on T2D are still modifiable through changes in lifestyle and behaviour.

Shifts in the microbial composition of the gut microbiome in response to weight gain are also well-documented[142]; accordingly, I detected candidate mediators that may represent such changes in circulation. For instance, the metabolite *N*-trimethyl 5-aminovalerate represents the obligate synthesis of trimethylated amino acids by the gut microbiota[143]. Elevated levels of *N*-trimethyl 5-aminovalerate have been associated with increased risk of liver steatosis[143] and with other incident diseases in this study. *N*-acetylglycine, which is a known candidate mediator, has also been associated with fibre intake and is produced by gut microbiota[144].

In this study, a repeatedly observed but hitherto unknown metabolic surrogate of adiposity, namely, metabolonic lactone sulfate, was structurally identified. Steroid metabolites are synthesised from cholesterol by multiple enzymes, including those encoded by genes linked to this metabolite in previous GWASs[19,20]. The effect of overall adiposity on metabolonic lactone sulfate levels may also be explained by steroid sulfatase and oestrogen sulfotransferase, which have been shown to affect adipocyte turnover and regulate energy and glucose homeostasis[145,146]. Although metabolonic lactone sulfate was not selected as a candidate mediator, it may well relate to other T2D-associated comorbidities including kidney disease and endometrial cancer. The identification of metabolonic lactone sulfate as a sulfated steroid derivative reiterates previous research highlighting shifts in steroid and closely related bile acid metabolism as a hallmark of obesity-induced deteriorations of metabolism with systemic consequences for multiple endocrine signalling systems. The identification of pregn steroid monosulfate* as a candidate mediator also supports the role of steroid sulfation and bile acid signalling in T2D pathogenesis, as reviewed elsewhere[145–147].

2.6.3. Study Strengths and Limitations

Untargeted metabolomic profiling in a large population cohort setting enabled the systematic characterisation of shared metabolomic changes between weight gain and T2D. Bayesian feature selection methods robustly accounted for between-metabolite correlations, identifying a set of candidate mediators that best explained the effects of weight gain on incident T2D risk. These methods, coupled with the prospective nature of the population cohort study design, provide preliminary insight into causal directions of effect that should be validated with formal statistical frameworks such as mediation analysis[148] and Mendelian randomisation[149]. Additional incorporation of multiple layers of data enabled the comprehensive study of complex genetic and lifestyle behaviours that candidate mediators likely represent.

This study had some limitations. Weight gain was calculated based on self-reported data before baseline, which might be affected by reporting bias. However, associations of weight gain on metabolites were highly-correlated with those observed cross-sectionally for measured BMI, suggesting that the effect of current BMI on metabolites likely captures previous weight trajectories.

In this study, I showed that the 20 candidate mediators identified explained more of the association between weight gain and incident T2D risk than randomly selected "control" sets of metabolites. Nevertheless, a residual "unexplained" association remained. This may be due to several reasons. One possible reason is the presence of measurement and quantification errors during metabolomic profiling. Such errors could occur from differences in chemical properties between metabolites, despite the adaptability of LC-MS technologies to measure different classes of metabolites at scale. Furthermore, weight at age 20 years was self-reported and not explicitly measured. Although the calculated measure of average annual weight change replicated several metabolite associations for BMI and weight change[9,96,97], potential reporting bias may also account for the residual association. Another possible reason is that non-fasted plasma samples were not immediately analysed on collection but instead stored at -175°C until shipping for analysis, when they were then temporarily stored at -70°C **(Appendix Supplementary Information)**. Storage at this low temperature would reduce, but not eliminate, the degradation of specific metabolites in the sample[150]. Another potential factor is that blood represents systemic changes in metabolism and not tissue-specific effects[151,152].

2.6.4. Conclusions

By integrating comprehensive metabolomic profiling with data-driven Bayesian methods I identified metabolic perturbations that link weight gain and T2D risk and have not previously been described. Identified mediators were driven by genetic susceptibility to endophenotypes known to cause T2D, such as adiposity and insulin resistance, as well as modifiable risk factors, opening direct opportunities for prevention.

CHAPTER 3: SYSTEMATIC IDENTIFICATION OF THE METABOLIC EFFECTS OF VARIANTS AT IEM GENES

3.1. Abstract

Background Previous metabolome GWASs have identified examples of variants at IEM genes that influence metabolite levels in the general population. Systematic metabolomic characterisation of these variants will enable causal assessment of their effects on metabolic pathways and clinical outcomes in the general population.

Methods A metabolome GWAS was conducted in the EPIC-Norfolk and INTERVAL prospective cohorts for 17 million genotyped and imputed variants and 913 metabolites detected in plasma (discovery: 14,296 participants; validation: 5,698 participants). Metabolic loci were defined and conditionally independent signals were identified. To test for enrichment of metabolic loci for IEM genes, metabolic loci were assigned to genes based on the biochemical literature, physical proximity (±5kb) and varying locus intervals and tested for enrichment for IEM genes (as identified by the database Orphanet) using a binomial two-tailed test. Conditionally independent variants at metabolic loci linked to IEM genes were prioritised if they were associated with a metabolite affected in IEM pathology or a related one (as identified by the databases IEMBase or OMIM).

Results Metabolic loci demonstrated significant and robust enrichment for IEM genes (fold-change enrichment: 8.44, binomial $p=3.98\times10^{-61}$). I identified 241 variants at 108 loci that were linked to likely causal IEM genes annotated by the biochemical literature and/or physical proximity (±5kb), of 791 conditionally independent variants at 320 loci. Of the 241 variants, 202 at 90 loci were associated with a metabolite implicated in IEM aetiology or a related one. Absolute per allele effect sizes and minor allele frequencies for the 187 variants not reported pathogenic for the IEM ranged from 0.063-2.75 per 1-SD metabolite difference and 0.0009-0.495.

Conclusions Untargeted metabolomic profiling identified variants at IEM genes that cause differences in plasma metabolites in the general population with effects that are specific to metabolites affected in the corresponding IEMs.

3.2. Background

The clinical importance of studying disease mechanisms is well illustrated using IEMs. IEMs, such as phenylketonuria and maple syrup urine disease, are caused by rare and heritable genetic mutations that lead to the toxic accumulation of molecules along a single metabolic pathway[10]. IEMs are individually rare but collectively common; although the prevalence of a given IEM ranges from one in 2,500 to one in 100,000, as many as one in every 784 newborns are affected by an IEM[153,154]. The molecular mechanisms of IEMs are well understood due to their extreme metabolic and phenotypic consequences. This enables the development of cost-effective prevention screening programmes for select IEMs in newborns[155], with opportunities of expansion to cover additional IEMs under constant review[11,156].

Evidence based on rare disease patients suggests that more common variation at IEM genes may also influence metabolite levels in the general population. For example, common variants (MAF>0.05) at the genes *APOB*, *LPL* and *LDLR* are associated with elevated cholesterol levels, reflecting extreme cases of hypercholesterolemia and hyperlipidemia caused by rare mutations at the same genes[45–47]. In this particular example, elevated levels of cholesterol caused by common variants also predispose carriers to coronary artery disease, necessitating the development of a genetic screening programme[157] to identify high-risk individuals and prevent or manage disease risk.

Additional metabolic effects of variation at IEM genes have been detected in population-based studies by GWASs of the metabolome[16,18–21]. One such identified example is the *CPS1* gene, which catalyses the first step of ammonia catabolism in the urea cycle[158]. Rare mutations in *CPS1* cause CPS1 deficiency, an IEM characterised by toxic accumulation of ammonia, cerebral damage, developmental delay and coma[159]. In recent GWASs of the metabolome (mGWASs), a detected *CPS1* variant was associated with higher glycine levels[18,19]. Glycine is an important source of ammonia through the ammonia-glycine cleavage complex, suggesting that the *CPS1* variant reduces the flux of ammonia catabolism through the urea cycle, leading to increased flux via the ammonia-glycine cleavage complex instead[160]. Metabolome GWASs performed using larger cohorts have also shown that metabolic loci are also more likely to be IEM genes than expected by chance[19,21]. Metabolomic profiling combined with whole genome and exome sequencing data[20,22] have also identified low-frequency variants at IEM genes that are associated with metabolite levels up to four standard deviations from the population mean. However, no study to date has adopted a systematic approach to characterise the metabolic and phenotypic effects of variants across IEM genes.

Here, I used findings from the largest mGWAS (hereafter referred to as the ‘Metabolon mGWAS’[161]) to date, which measured 913 metabolites across eight metabolite classes as well as chemically unidentified metabolites in up to 20,000 cohort participants. The breadth of metabolite levels measured and large sample size used in this mGWAS enables replication of previous associations as well as the potential identification of novel ones. To assess the enrichment of metabolic loci for IEM genes, different approaches were used to assign likely causal genes to variant-metabolite associations. Previous mGWASs have assigned candidate genes using one or more of the following approaches: a.) physical location of the variant relative to genes in the region[16]; b.) manual assessment of the biochemical literature[19], c.) integration of gene expression and epigenomic data[20], or d.) a combination of these methods[21]. Each approach has advantages and disadvantages, as explored in this Chapter. Finally, the metabolic consequences of IEMs have been systematically recorded in the database IEMBase[162], enabling an assessment of how specific metabolic associations for variants at IEM genes are for metabolite levels affected in the corresponding IEM.

3.3. Aim and Objectives

The aim of this study was to systematically identify variants at IEM genes that influence levels of corresponding IEM-affected metabolites in the general population **(Figure 1)**. The focus of this study was to identify variants not previously reported to cause the relevant IEM, although the role of detected pathogenic variants on metabolite levels in the general population was also considered. The objectives of the study were to:

1. Assess metabolic loci for enrichment of genes known to cause IEMs.
2. Identify variants at IEM genes that are associated with a metabolite affected in the corresponding IEM.

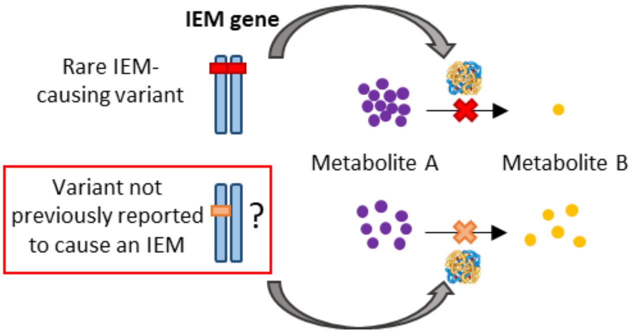


Figure 1: Overview of the study aim. The aim of this study (highlighted in the red box) was to systematically identify variants at IEM genes that influence levels of corresponding IEM-affected metabolites in the general population.

3.4. Methods

3.4.1. Study design and participants

The European Prospective Investigation into Cancer and Nutrition (EPIC)-Norfolk[107] study is a prospective cohort study of 25,639 individuals aged 40-79 years at baseline and recruited from 35 GP practices across Norfolk in 1993-1997. The study was approved by the Norfolk Research Ethics Committee (ref. 98CN01) and all participants gave signed informed consent. This study comprises of data measured in two non-overlapping sets of participants:

Type 2 diabetes (T2D) case-cohort - the design of the EPIC-Norfolk nested T2D case-cohort study, including ascertainment and verification of incident T2D cases, has previously been described in detail[108]. In brief, it includes 1,503 individuals (45% cases) ascertained using self-report, linkage to primary and secondary care, drug register, hospital admission, and mortality data[108].

Subcohort – a subcohort was drawn from all EPIC-Norfolk participants who were not part of the T2D case-cohort study. For metabolite measurements (see **Section 3.4.2.2.** below), participants were measured in two sets, each comprising of approximately 6,000 participants.

Participants were also included from the INTERVAL[163,164] study, which aims to determine the optimal frequency at which individuals can donate blood and comprises of 45,265 whole blood donors recruited between 2012 and 2014.

3.4.2. Phenotypic measurements

3.4.2.1. Genetic profiling

Genetic processing protocols were similar between studies. Participant genomes were genotyped using the UK Biobank Affymetrix Axiom Array and imputed using the Haplotype Reference Consortium[109] as well as the combined UK10K[111]/1000 Genomes[110] imputation panels in EPIC-Norfolk and to the UK10K/1000 Genomes imputation panel in INTERVAL. Genetic variants were excluded if the imputation quality INFO score was <0.4 in EPIC-Norfolk or <0.3 in INTERVAL. Where possible, genotyped values of SNPs were used instead of imputed dosages, though the latter were tested to ensure they met standard quality control parameters (e.g. no deviation from Hardy-Weinberg equilibrium).

3.4.2.2. Metabolomic profiling

In both studies, metabolomic profiling was performed on non-fasted citrate plasma samples obtained at baseline using LC-MS/MS in the untargeted Metabolon HD4 Discovery^{TM} platform. In EPIC-Norfolk, samples for metabolomic profiling were selected in the order in which they were stored at baseline (quasi-random selection). Prior to GWAS, metabolites yielding measurements in less than 100 samples were excluded in EPIC-Norfolk and INTERVAL. Median-normalised and natural log-transformed values of metabolite levels within each batch were used, and values that were >5 standard deviations from the mean metabolite value were winsorised to avoid variation due to extreme values. Metabolite residuals after adjusting for batch, age and gender were standardised to a mean of zero and a standard deviation of one. In INTERVAL, metabolite levels were also regressed by recruitment centre, plate number, appointment month, lag time between the blood donation appointment and sample processing, and the first five ancestry principal components.

3.4.3. Statistical analysis

3.4.3.1. Exclusions

Due to differences between studies, GWAS analyses were performed separately within each study and then meta-analysed. Participants who were not part of the subcohort of the T2D case-cohort of EPIC-Norfolk were excluded. In INTERVAL, participants with measurements for fewer than 300 metabolites were also excluded. Additionally, participants who had one or more of high rates of missing metabolite measurements or incomplete genotype data were excluded. This left 5,841 and 8,455 participants from EPIC-Norfolk and INTERVAL respectively for the GWAS discovery analysis and a non-overlapping set of 5,698 participants from EPIC-Norfolk for validation analysis.

3.4.3.2. Genome-wide association analysis

The Metabolon mGWAS, overseen by researchers at the University of Cambridge, comprises of the largest known metabolome GWAS with untargeted metabolomic profiling to date with meta-analyses of genetic and metabolomic profiling data performed in 14,296 participants across the EPIC-Norfolk and INTERVAL cohorts.

After data processing, linear mixed models (LMMs) were used to perform univariate linear regression of the standardised residuals. LMMs can be used to jointly model all genotyped markers and are useful for GWAS studies because they effectively account for potential confounding due to population stratification[165,166] and hidden relatedness[167]. Therefore, all individuals of EPIC-Norfolk were included, and the model accounted for population structure based on autosome-wide genotyped SNPs. The software BOLT-LMM v2.2 was used to enable rapid computation of LMMs and retention of statistical power[168]. In some cases, BOLT was unable to compute LMMs because heritability estimates were extremely high or low. In these cases, an alternative software that uses Bayesian and frequentist methods to calculate associations called SNPTEST v2.5.2[169] was used to analyse individuals who were no more than second-degree or third-degree related to each other. In SNPTEST models, the first four principal components were included in the EPIC-Norfolk cohort to account for variation due to population sub-structure. No principal components were included in models for the INTERVAL cohort, as these had already been adjusted for when calculating residuals of metabolite levels as described in **Section 3.4.2.2.**

3.4.3.3. Meta-analysis of GWAS results

Prior to meta-analysis, duplicated variants, copy number and insertion or deletion variants for which alleles were not known were excluded. Variants that had implausible parameters were also excluded (specifically, if one or more of the following was true: a.) where absolute effect size was greater than 10; b.) standard error was less than zero or greater than 10; c.) imputation quality INFO score was less than 0.3 based on the full dataset analysed and complete phenotype data, or d.) Hardy-Weinberg equilibrium $p \leq 1x10^{-6}$ based on unrelated individuals). A total of 913 metabolites that were available in both studies and detectable in at least 100 individuals within each study were taken forward for meta-analysis.

Meta-analysis between studies was performed using a fixed effects model that calculates the mean of effect sizes across studies, inversely weighting study contribution based on the estimated standard error[170,171]. A z-score statistic of the combined mean divided by the estimated standard error was then calculated to derive p-values that determined the statistical significance of the overall measure. In this study, the software, METAL[172], was used to perform inverse-variance weighted meta-analysis.

3.4.3.4. Locus definition

Variants that had a.) a meta-analysed p-value less than $5x10^{-8}$; b.) a minor allele count greater than ten in both studies; c.) same directions of effect for the reference allele, and d.) a p-value less than $1x10^{-2}$ in both studies prior to meta-analysis were considered. For each metabolite, the strongest-associated variant was selected. Variants that were associated with the same metabolite and in LD ($R^2 > 0.1$) with this variant were clumped. Variants across different metabolites were also clumped if they existed in LD ($R^2 \geq 0.6$) with each other. Subsequently, variants within a 500kb window flanking the variant with the strongest association for any metabolite were distance pruned in order of smallest to largest p-value, and the locus coordinates were expanded to -250kb from the variant at the start of the locus and +250kb from the variant at the end of the locus.

3.4.3.5. Validation and conditional analysis

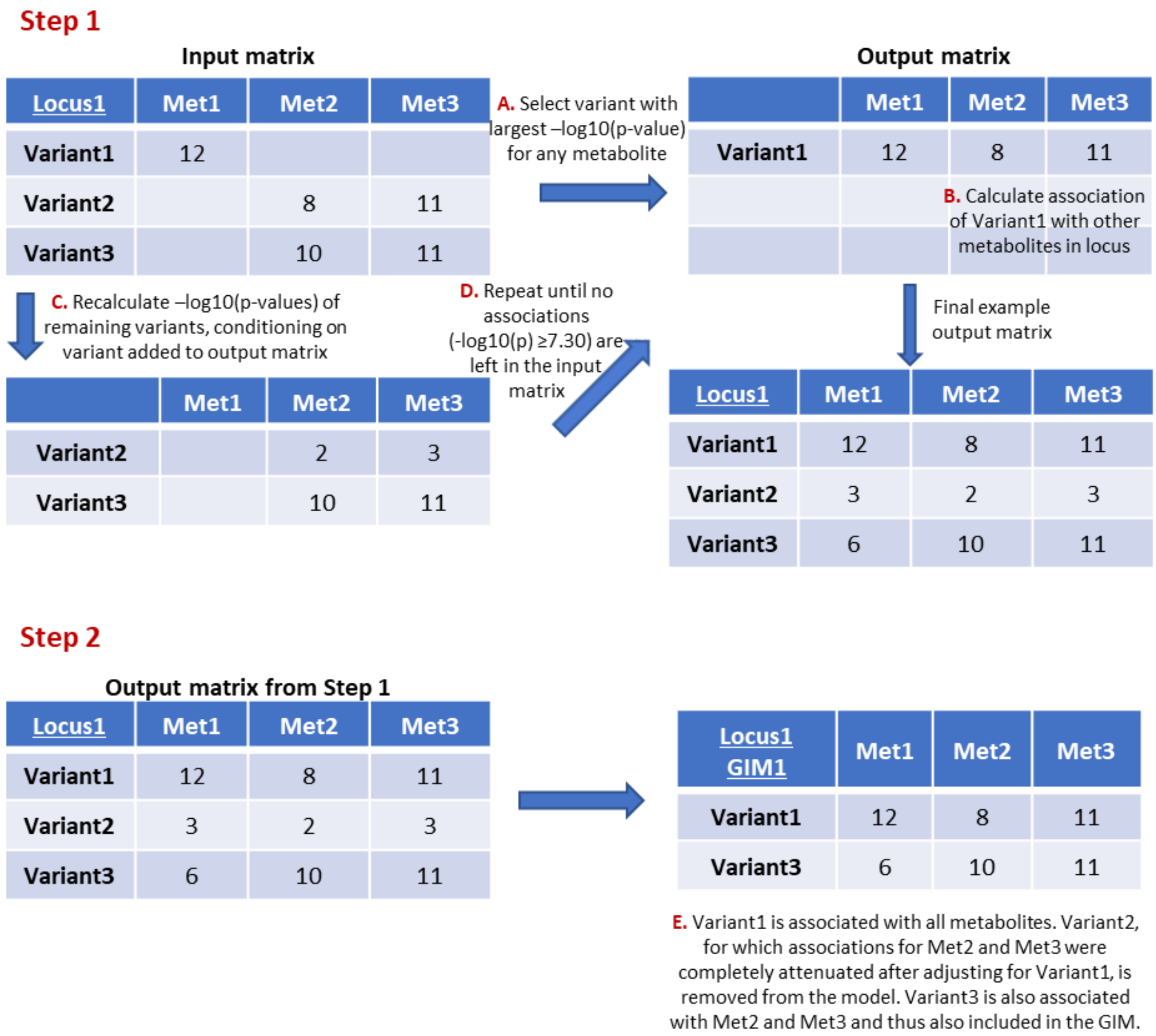
Validation was performed by meta-analysing discovery results with GWAS results from a non-overlapping subset (n=5,698) of the EPIC-Norfolk cohort. The strongest variant-metabolite associations for each locus were taken forward for analysis if they were either significant in the combined discovery and validation set or if the metabolite was only present in the discovery set but was significant. Genetic associations were significant at Bonferroni-corrected threshold ($p \leq 5.48 \times 10^{-11}$, i.e. 5×10^{-8} correcting for 913 metabolites tested).

To identify additional metabolite-associated variants within loci, conditional analysis was performed in SNPTEST 2.5.1 using combined genetic and phenotypic data from the discovery set (n=14,296), residuals for each metabolite in the locus, study and the first five principal components of genetic ancestry as fixed effects. In this step, variants with p-values meeting a threshold of 1.25×10^{-8} (correcting for the maximum number of metabolites and variants across tested regions) were considered. The variant with the smallest p-value for any metabolite within a locus was conditioned on, and the next most strongly associated variant was retained if it was still significant with a p-value smaller than 1.25×10^{-8}. This step was repeated, leaving a final fitted model that included all selected variants and excluding any variants that were no longer significant in that model.

3.4.3.6. Definition of genetically influenced metabotypes

Conditional analysis was performed by selecting variants using association p-values for a reference metabolite at each locus. This accounted for LD between variants associated with the same metabolite, but not for LD between variants associated with different metabolites in the same locus. To effectively account for the latter, genetically influenced metabotypes (GIMs), which are minimal variant sets that best represent signals across metabolites within a locus[24,44], were defined. GIM definition was performed as outlined in **Figure 2**.

It should be noted that initial locus definition **(Section 3.4.3.4.)** identified 330 loci. However, at ten loci it was subsequently found that the comprising GIMs were not independent of those at neighbouring loci. These GIMs and loci were therefore dropped, leaving 320 loci (which is reported in the **Results**).



Step 1

Input matrix

Locus1	Met1	Met2	Met3
Variant1	12		
Variant2		8	11
Variant3		10	11

A. Select variant with largest –log10(p-value) for any metabolite

Output matrix

	Met1	Met2	Met3
Variant1	12	8	11

B. Calculate association of Variant1 with other metabolites in locus

C. Recalculate –log10(p-values) of remaining variants, conditioning on variant added to output matrix

	Met1	Met2	Met3
Variant2		2	3
Variant3		10	11

D. Repeat until no associations (-log10(p) ≥7.30) are left in the input matrix

Final example output matrix

Locus1	Met1	Met2	Met3
Variant1	12	8	11
Variant2	3	2	3
Variant3	6	10	11

Step 2

Output matrix from Step 1

Locus1	Met1	Met2	Met3
Variant1	12	8	11
Variant2	3	2	3
Variant3	6	10	11

Locus1 GIM1	Met1	Met2	Met3
Variant1	12	8	11
Variant3	6	10	11

E. Variant1 is associated with all metabolites. Variant2, for which associations for Met2 and Met3 were completely attenuated after adjusting for Variant1, is removed from the model. Variant3 is also associated with Met2 and Met3 and thus also included in the GIM.

Figure 2: Example illustrating how GIMs are defined. Numbers in boxes represent -log$_{10}$(P-values), and a -log$_{10}$(P-value) of 7.30 is considered significant in this example. In this example, Variant1 is in high LD with Variant2; as Variant1 has the strongest association in the locus, it is selected to represent the metabolite associations of Variant2, which is subsequently dropped from the GIM. Variant3 is not in high LD with either Variant1 or Variant2 but is also associated with Met2 and Met3; therefore, it is included in the same GIM as Variant1.

3.4.3.7. Definition of novelty

At the time of writing, the definition of what constitutes a ‘novel’ GIM or variant-metabolite association was still being revised for the Metabolon mGWAS manuscript (in preparation). For pragmatic purposes, the novelty of associations was therefore defined in this thesis by comparing associations against reported associations in the two largest published mGWASs[19,20]. All lead variants or proxies ($R^2 \geq 0.1$) and their metabolite associations were queried at study specific p-value thresholds ($1.03x10^{-10}$ and $1.9x10^{-11}$, respectively[19,20]) and considered novel if they had not previously been reported.

3.4.3.8. Likely causal gene annotation

In the Metabolon mGWAS, likely causal genes were assigned to 791 conditionally independent variants using the biochemical literature. The automated step leverages the rich history of biochemistry available in the literature and biochemical databases. Briefly, this pipeline works by exploiting overlaps in the names of the metabolites and the names of the proteins acting on them, as well as the genes encoding those proteins and rare diseases caused by those genes. For example, the enzyme phenylalanine-4-hydroxylase encoded by the phenylalanine hydroxylase (*PAH*) gene is the genetic cause of hyperphenylalaninemia, and the gene overlaps the strongest GWAS signal for circulating phenylalanine levels. Therefore, a fuzzy text similarity metric (pair coefficient) encoded in the ruby gem 'fuzzy_match' software was used to compare metabolite names to gene and protein names across resources from the Human Metabolome Database (HMDB)[1], Online Mendelian Inheritance in Man (OMIM)[173], UniProt[174], Ensembl[175], Gene Ontology[176,177] and the Kyoto Encyclopaedia of Genes and Genomes[178]. A score >0.5 was considered a match, and all automated hits were manually reviewed for plausibility.

The 20 genes closest to each of the variants with the strongest metabolite associations within each GIM were manually reviewed as the likely causal gene using the biochemical literature. Likely causal gene annotation was then reviewed based on other associated metabolites with the variant under consideration and within the GIM. Other genes were also considered if the Entrez gene or UniProt description of the gene suggested it could potentially be related to the metabolite. If experimental evidence could be found linking one of the 20 closest genes to the metabolite, or more (as was commonly observed for paralogs with similar molecular functions), all genes were selected as the biologically most likely causal genes. If no existing evidence was found, no causal gene was manually selected. For each manually selected causal gene, the earliest experimental evidence linking the gene (preferably the human gene) to the metabolite was identified. The median publication year for the identified experimental evidence was 2000.

To complement the above methods of causal gene annotation, I used Ensembl VEP[87] and SNiPA[179] annotations to identify the closest genes (±5kb) of variants or their proxies (R^2>0.6). These methods of likely causal gene annotation provide a conservative list of genes mapping to metabolic loci and assume that the gene driving the signal is one near the lead SNP. To test the robustness of enrichment to the number of genes mapped, I also mapped variants to genes using more relaxed locus definitions: i.) gene location within a wider 500kb window, and ii.) the Metabolon mGWAS locus (as defined in **Section 3.4.3.4.**). Genes located within a 500kb window of the conditionally independent variant were identified using a list of protein-coding genes obtained from the human reference annotation database, GENCODE[180]. **Table 1** provides a summary and comparison of the different locus definitions.

Table 1: Summary and comparison of methods used to map genes to metabolic loci.

Method	Description	Advantages	Disadvantages
Biochemical literature	The 20 protein-coding genes closest to a lead variant were assessed using the biochemical literature, previous GWAS annotations and a semi-automated pipeline.	This method is highly specific. Likely causal gene annotations using this method are more certain than those from other methods.	Much of the biochemical literature is established from known IEM associations. This method is therefore likely to be biased towards detecting IEM gene enrichment.
Closest gene(s)	Loci were annotated with all genes within 5kb of the lead SNP or a linked one ($R^2 > 0.6$) using the VEP and SNiPA software.	This method provides a conservative list of genes mapped to metabolic loci.	This method assumes that the gene driving the signal is the one closest to the lead or proxy variant. The 5kb distance threshold may be overly conservative.
500kb window	Genes within 500kb window of the lead SNP were identified using GENCODE.	This method provides a wider distance interval to assess potential causal genes. This method is easy to implement.	This method does not consider gene annotations for variants in LD and may not identify the causal gene as accurately.
Metabolon mGWAS Locus	Genes within the defined loci **(Section 3.4.3.4)** of each lead SNP were identified using GENCODE.	This method accounts for LD between variants associated with the same metabolite and more likely to identify the likely causal gene more accurately than the 500kb window method.	This method is sensitive (as some loci such as the *FADS1* locus are very large), but less specific and thus more likely to produce a conservative estimate of enrichment.

3.4.3.9. Enrichment assessment

Enrichment of metabolic loci detected in the Metabolon mGWAS for IEM genes was assessed using a two-tailed binomial test. As IEM genes are usually protein-coding genes, estimates of the number of protein-coding genes (19,817)[180] and IEM genes (785; 4%)[181] across the genome were used to calculate the expected proportion of IEM genes in a test sampling of genes.

3.4.3.10. Identification of variants at IEM genes associated with corresponding IEM-related metabolite(s)

In the rest of this chapter, the term ‘IEM gene-linked variant’ is used to refer to variants at metabolic loci that are mapped to IEM genes. These variants may or may not be associated with the same metabolites affected in the relevant IEM or be pathogenic for the IEM they are linked to **(Abbreviations and Terminology; Figure 2)**.

To prioritise a set of variants at IEM genes for clinical follow up in subsequent chapters, metabolites associated with IEM gene-linked variants were annotated based on their relevance to the corresponding IEMs. Databases of rare Mendelian disorders can be leveraged to achieve this purpose. For example, the Online Inheritance in Man (OMIM) database[173] curates symptoms based on case reports and molecular studies in the literature. More recently, the IEMBase database[162] integrates expert-compiled information across 530 IEMs to highlight clinical symptoms and metabolites affected in IEMs, making it an ideal resource for systematic assessment of the metabolic and phenotypic consequences of variation at IEM genes.

To prioritise IEM gene-linked variants that were associated with metabolites affected in the corresponding IEMs, associated metabolites were compared to those reported as affected in the corresponding IEMs by OMIM and IEMBase. Variants for which one or more associated metabolites in the GIM were present in the biochemical testing panel for the relevant IEM, or reported as being altered in case reports, were prioritised. Identified variants were termed ‘IEM familiar variants’ (IFVs) based on their link to the IEM gene and their similar metabolite associations mimicking those of the corresponding IEMs **(Abbreviations; Figure 3)**. Where possible, HMDB IDs were used to compare across non-standardised metabolite names.

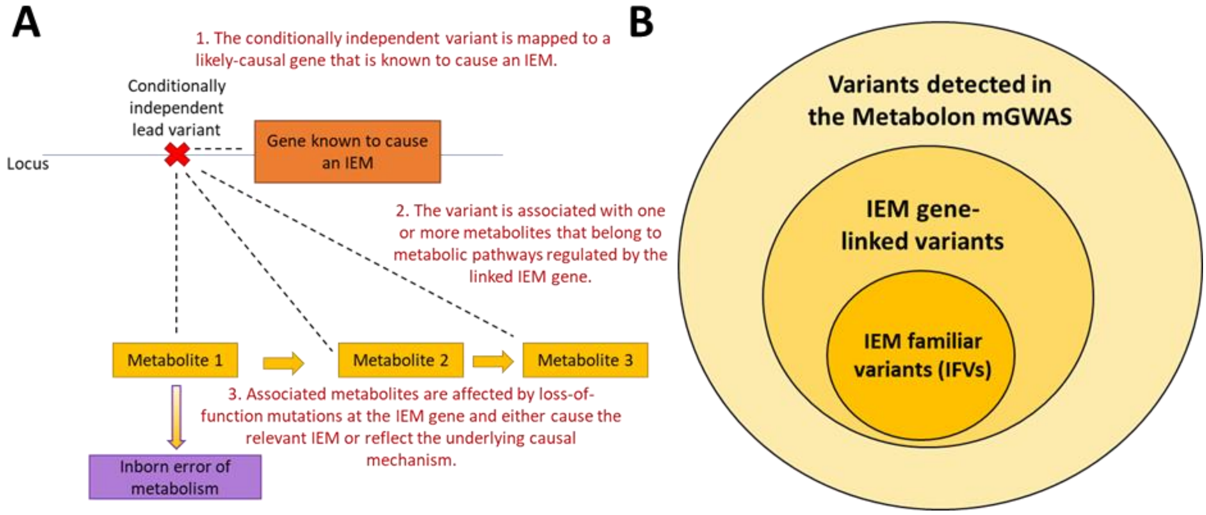

A
1. The conditionally independent variant is mapped to a likely-causal gene that is known to cause an IEM.
Conditionally independent lead variant
Locus
Gene known to cause an IEM
2. The variant is associated with one or more metabolites that belong to metabolic pathways regulated by the linked IEM gene.
Metabolite 1
Metabolite 2
Metabolite 3
3. Associated metabolites are affected by loss-of-function mutations at the IEM gene and either cause the relevant IEM or reflect the underlying causal mechanism.
Inborn error of metabolism
B
Variants detected in the Metabolon mGWAS
IEM gene-linked variants
IEM familiar variants (IFVs)

Figure 3: Schematic of terminology used in this thesis. A.) The three criteria defining IFVs: 1.) the variant is conditionally independent and one where the annotated likely causal gene is known to cause an IEM; 2.) the variant is associated with one or more metabolites that belong to metabolic pathways regulated by the linked IEM gene, where one or more of the associated metabolites is 3.) dysregulated by loss of function mutations in the gene and either causes the IEM or reflects the underlying metabolic mechanism causing it. **B.)** Venn diagram demonstrating relationships between terminologies used in the thesis. IEM gene-linked variants form a subset of all variants detected in the Metabolon mGWAS and are located at loci for which the likely causal gene is known to cause an IEM. IFVs are the subset of IEM gene-linked variants which are also associated with metabolites known to be dysregulated in IEM aetiology. IEM gene-linked variants satisfy criterion 1 of **Figure 3A** while IFVs and their associated metabolites satisfy all criteria in **Figure 3A**.

3.4.3.11. ClinVar annotation of IFVs

IFVs reported pathogenic for the corresponding IEM were identified using the ClinVar[182] database (https://www.ncbi.nlm.nih.gov/clinvar/, last downloaded in October 2019). Evidence for variants annotated as 'likely-pathogenic' or 'pathogenic' for the corresponding IEM or those with conflicting interpretations of pathogenicity were manually curated to confirm or review their likely effect.

All statistical analyses and graphics were performed and produced using R version 3.5.3.[128], Excel (Microsoft Office 16) and STATA version 14.2[129].

3.5. Results

3.5.1. A Total of 1,847 Locus-metabolite Associations Across 330 Loci and 646 Metabolites Were Identified in the Metabolon MGWAS

The Metabolon mGWAS covered 913 plasma metabolites across eight broad classes including lipids (n=301), amino acids (n=153), xenobiotics (n=92), nucleotides (n=23), peptides (n=20), carbohydrates (n=19), cofactors or vitamins (n=17) and energy-related metabolites (n=7). In addition, 281 detected metabolites of unidentified chemical structure were assessed. Discovery and validation identified

1,834 regional associations across 320 loci and 646 metabolites ($p \leq 5.48 \times 10^{-11}$) **(Figure 4, Box 1)**. The 320 loci contained 423 GIMs, 304 of which had not been identified previously in the two largest published mGWAS studies[19,20].

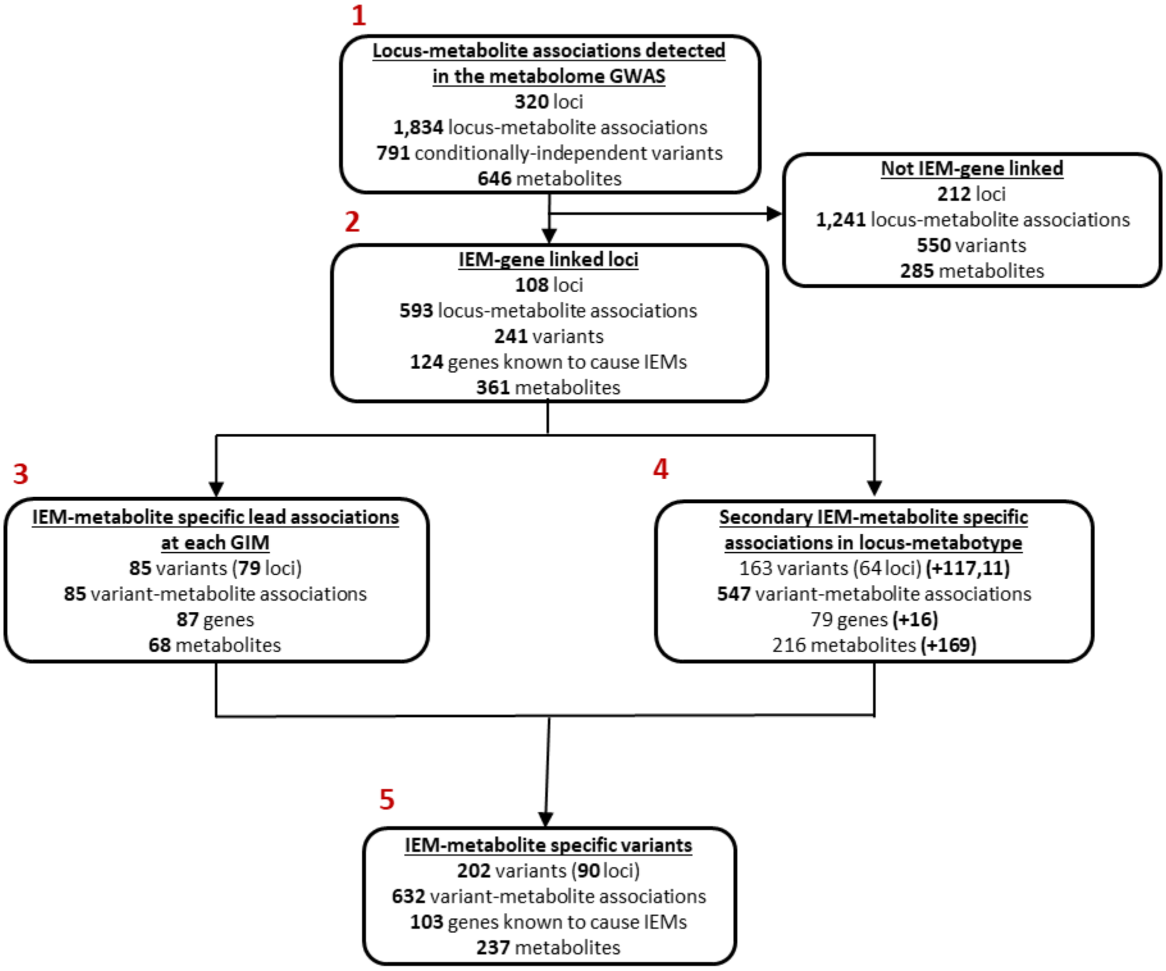

1
Locus-metabolite associations detected in the metabolome GWAS
320 loci
1,834 locus-metabolite associations
791 conditionally-independent variants
646 metabolites
Not IEM-gene linked
212 loci
1,241 locus-metabolite associations
550 variants
285 metabolites
2
IEM-gene linked loci
108 loci
593 locus-metabolite associations
241 variants
124 genes known to cause IEMs
361 metabolites
3
IEM-metabolite specific lead associations at each GIM
85 variants (79 loci)
85 variant-metabolite associations
87 genes
68 metabolites
4
Secondary IEM-metabolite specific associations in locus-metabotype
163 variants (64 loci) (+117,11)
547 variant-metabolite associations
79 genes (+16)
216 metabolites (+169)
5
IEM-metabolite specific variants
202 variants (90 loci)
632 variant-metabolite associations
103 genes known to cause IEMs
237 metabolites

Figure 4: Flowchart summarising results from analyses performed in this Chapter. In **Box 2**, the number of IEM genes corresponds to the number identified based on the biochemical literature and closest gene(s) methods. Numbers in red signpost relevant boxes in the main text.

3.5.2. Metabolic loci in the Metabolon mGWAS are enriched for IEM genes

Of the 320 loci detected, 216 were annotated with 253 individual likely causal genes or gene sets based on the biochemical literature. These gene sets comprised of 290 unique genes that were used as the sample size for enrichment analysis. Of these 290 genes, 97 have been reported to harbour mutations that cause an IEM, representing significant enrichment of metabolic loci for IEM genes (fold-enrichment: 8.44, binomial $p=3.98 \times 10^{-61}$). Evidence of enrichment was still present in sensitivity analyses using other locus definitions **(Table 2)**.

Table 2: Evidence of enrichment was robust against different methods of causal gene annotation and varying locus definitions. Analysis was performed against a background of 19,817 protein-coding genes (4% IEM).

Method	Total number of genes	Number known to cause an IEM	Fold-change enrichment	Two-tailed binomial p-value
Biochemical literature	290	97	8.44	3.98×10^{-61}
Closest gene(s)	964	117	3.06	1.67×10^{-26}
500kb window	3441	195	1.43	7.84×10^{-7}
Metabolon mGWAS locus	5187	261	1.27	7.94×10^{-5}

3.5.3. A Total of 202 of 241 Variants (84%) at Loci Linked to IEM Genes Were Associated With a Metabolite Implicated in IEM Aetiology

A total of 241 variants at 108 loci and 100 GIMs were linked to an IEM gene by the biochemical literature and/or closest gene(s) methods **(Figure 4, Box 2)**. Of these variants, 85 were associated with a metabolite implicated in IEM aetiology where the association represented the strongest association at a GIM within that locus **(Figure 4, Box 3)**.

When considering additional associations within GIMs, a further 117 variants had metabolite associations that mimicked metabolic consequences of the IEM **(Figure 4, Box 4)**. The strongest variant-metabolite association was also one reflecting metabolic consequences of an IEM at all but ten of these GIMs, for which seven were due to the metabolite in the strongest association being chemically unidentified. This resulted in a total of 202 IFVs located within GIMs and conditionally independent of one another.

Altogether, 202 of 241 (84%) conditionally independent variants at 90 IEM gene-linked loci were associated with the same or related metabolite implicated in IEM aetiology **(Figure 4, Box 5; Figure 5)**. Of these IFVs, 55 (27%) were flagged only by the biochemical literature (which is expected given that likely causal gene annotation was based on metabolite association) and 144 (71%) were identified using both the closest gene(s) and biochemical literature methods. Of the 632 IEM-specific variant-metabolite associations for the 202 IFVs, 520 (82.3%) were not reported in the two largest published mGWASs to date[19,20].

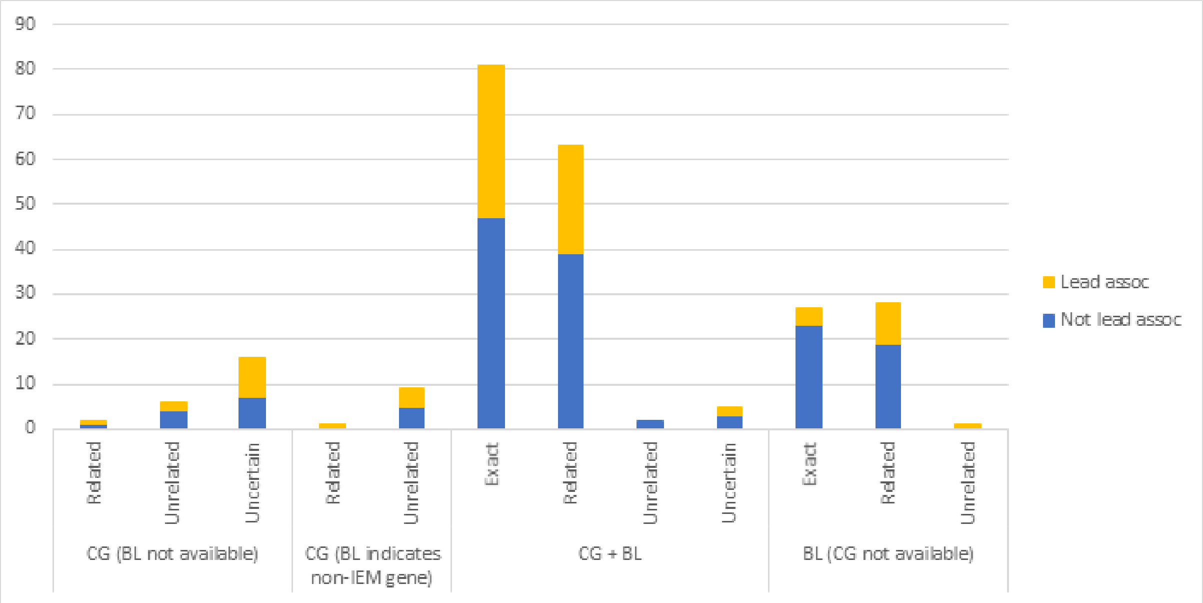


Figure 5: Metabolite matching of 241 conditionally independent variants within GIMs at IEM gene-linked loci. Yellow bars represent variants with the strongest metabolite association at the GIM while blue bars represent secondary associations. Assessment of these variants was performed based on likely causal IEM gene annotation by the closest gene(s) (CG) and/or biochemical literature (BL) annotation methods.

3.5.4. IFVs have large effect sizes on metabolite levels

Of the 202 conditionally independent variants specifically associated with corresponding IEM metabolites, 15 variants at 13 IEM genes were reported directly or by a linked variant ($R^2 \geq 0.6$) to be likely-pathogenic or pathogenic for the corresponding IEM in ClinVar ($0.0009 \leq MAF \leq 0.079$) **(Appendix Ch3_ST1)**. The clinical significance of these variants was also supported by VEP[87], which annotated ten of these as missense, two as splice donor/acceptor variants, two as intronic and one as a non-coding transcript exon variant. Of the 10,581 unrelated participants of European ancestry that were present in the EPIC-Norfolk cohort and had metabolomic data, 3,063 (29%) were heterozygote carriers and 88 (0.83%) were homozygote carriers for at least one pathogenic-predicted IFV.

Meta-data from the EPIC-Norfolk cohort indicated that three participants with metabolomic data were diagnosed with disease codes roughly corresponding to IEMs caused by *PAH* (E70.1 – 'Other hyperphenylalaninemias'), *FMO3* (E72.5 – 'Disorders of glycine metabolism') and *LIPC* (E78.4 – 'Other hyperlipidemia'). Although it is possible that these three individuals may have been carriers of the corresponding IFVs, the corresponding genotype data could not be retrieved due to low numbers of affected individuals and risk of de-anonymisation.

Of the remaining 187 IFVs, per allele absolute effect sizes and minor allele frequencies ranged from 0.063-2.75 per 1-SD metabolite difference and 0.0009-0.495. A total of 68 of all 202 strongest IFV-metabolite associations exceeded the large average effect size of *FTO* variants on BMI (0.35 kg/m^2 per allele)[183] **(Figure 6)**. This is similar to what was observed in terms of frequencies and effect sizes for all identified metabolite-associated variants, i.e. including those at other genes.

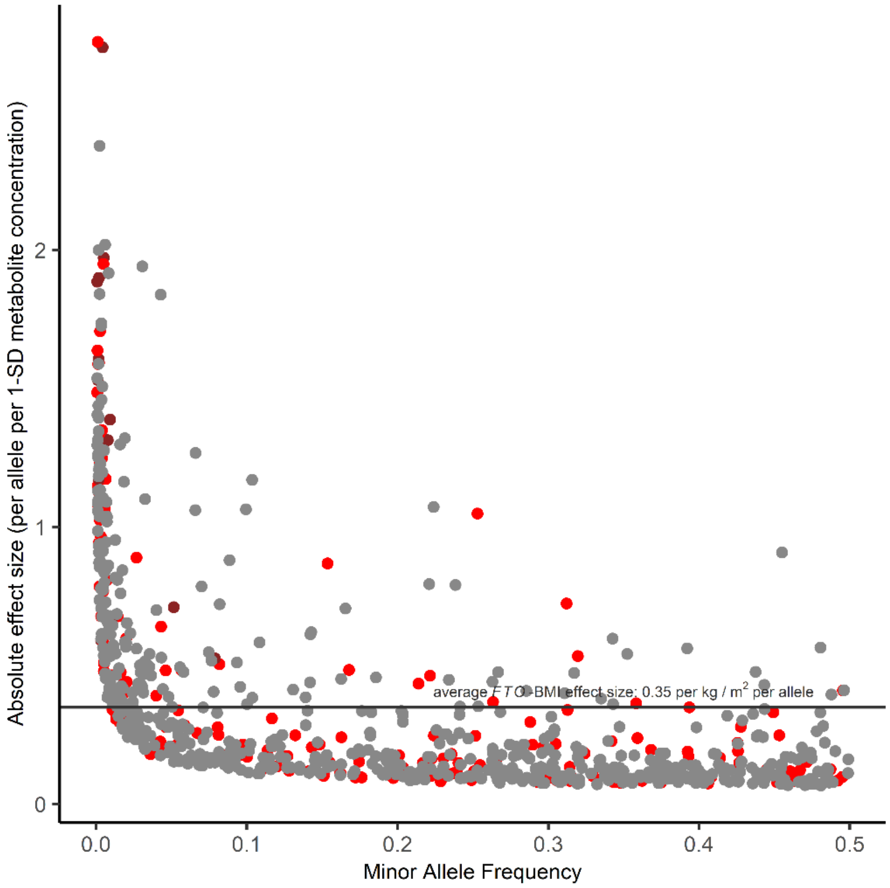


Figure 6: Absolute effect size (per allele per 1-SD relative concentration of metabolite) by minor allele frequency. Data for the most IEM-specific metabolite associations of each variant are shown. Pathogenic IFVs (n=15): dark red; IFVs not reported pathogenic for the IEM (n=187): red; variants at other genes (n=589): black. The horizontal black line represents the average reported effect size (per allele) of variants at the *FTO* gene and BMI.

3.6. Discussion

3.6.1. Summary of Findings

Systematic quantification and identification of IFVs in this study revealed two key findings, the first of which is that variation at IEM genes contributed substantially to metabolite levels in the general population. This was supported by the increase in identified locus-metabolite associations discovered in the Metabolon mGWAS that, combined with known loci, still showed enrichment for IEM genes[19,21]. The second key finding is that 84% of conditionally independent variants at IEM gene-linked loci were associated with a metabolite affected in the corresponding IEM or with a related one. For the strongest metabolite association, 68 of the 202 IFVs had effect sizes surpassing that of the large average effect size of *FTO* variants on BMI (0.35 kg/m^2 per allele)[183]. However, only 15 of these were previously reported to be pathogenic for the relevant IEM.

3.6.2. Novelty of Findings

These findings, which built on results from previous mGWASs[19,20], provide evidence to support the speculation that IEMs represent 'extreme examples of metabolic variation, probably everywhere present in minor degrees'[10]. Already, previous mGWASs have identified select examples where common variants at IEM genes are linked with metabolic effects also observed in the rare disease. One example is the identification of variants at genes known to cause familial dyslipidemia; in the general population, common variants at these genes were also associated with cholesterol and other lipid species[21]. Another example is the association of a *CPS1* variant, rs1047891, with higher levels of glycine, as previously discussed **(Chapter 3, Section 3.2.)**[18,19,158–160]. In addition to replicating 112 examples previously reported in the two largest mGWAS studies published[19,20] at the time of writing, an additional 520 previously unreported metabolite associations for IFVs were identified. These novel associations increase the likelihood of detecting previously unreported genotype-metabolite-phenotype links in downstream variant characterisation **(see Chapters 4-6)**.

In this study, I identified 15 IFVs that were common enough to be detected in the general population and reported to be potentially pathogenic for the corresponding IEM in ClinVar. Almost 30% of unrelated participants of European ancestry in the EPIC-Norfolk cohort were carriers of at least one of these IFVs, yet only three participants were diagnosed with corresponding IEMs. Due to data privacy concerns, it is unclear whether these three participants are indeed carriers of the corresponding pathogenic-predicted IFVs or whether their IEMs are caused by other mutations.

3.6.3. Study Strengths and Limitations

Strengths of this study included the use of untargeted metabolomic profiling and the definition of GIMs, which together enabled the detection of metabolic consequences of variation at IEM genes with high confidence. Another strength of this study was the comprehensive annotation of likely causal genes for independent variants, which enabled the systematic prioritisation of variants linked to IEM genes at scale.

Yet another strength of this study was the systematic comparison between metabolites associated with variants linked to IEM genes and those known to be affected in the corresponding IEM. Previous studies[16–21] have indicated select examples where variants at IEM genes are associated with metabolite levels that reflect more extreme metabolic sequelae linked to rare mutations at the same genes. However, this study is the first, to my knowledge, to use genotypic and untargeted metabolomic profiling data to formally quantify and demonstrate the striking similarities of metabolic consequences between common polymorphisms and rare, IEM-causing mutations at the same gene.

The novel associations detected in this study could contribute to a more systemic understanding of how genetic variation affects metabolic pathways. For example, cholesterol and triglyceride levels are known to be associated with variants at genes known to cause familial dyslipidemia. Yet in this study, associations for variants at these genes were also observed for other lipid species such as sphingomyelins and plasmalogens. While the genetic basis of familial dyslipidemia is relatively well-established in screening programmes, further assessment of the metabolic effects of IFVs at these genes may help to characterise the role of diverse lipid species in cholesterol metabolism and disease aetiology.

There were also limitations to this study. One was that the study was highly reliant on likely causal gene annotations from the biochemical literature. The biochemical literature is biased towards mapping loci to IEM genes, as much of the knowledge regarding biochemical genetics is derived from the study of IEMs. This limitation was addressed by testing for enrichment using more sensitive locus definitions, the results of which still indicated robust evidence of enrichment of metabolic loci for IEM genes.

During likely causal gene annotation, some variants were annotated with multiple paralogs of a gene or with multiple genes using methods based on the biochemical literature or on the closest gene. In these cases, variants were considered as linked to an IEM gene if at least one of the causal gene annotations was known to cause an IEM. Consideration of all potential likely causal gene annotations may have biased the estimate of enrichment or increased the likelihood of finding a variant at an IEM gene that was associated with IEM-related metabolite levels. However, variants at metabolic loci were more likely to be annotated with non-IEM genes in this study, which would have led to an under-estimation of enrichment. The significant enrichment detected despite this limitation therefore speaks to the growing body of evidence demonstrating that variation at genes known to cause IEMs can also influence metabolite levels in the general population.

Another limitation was that metabolites could be referred to using different names in the literature, making comparison difficult. Although efforts have been made to document different metabolite names and provide links across databases[1,184,185], these efforts remain limited for metabolites that are only recently detectable due to advances in LC-MS/MS technologies. I therefore used HMDB IDs[1] to identify metabolites across databases where possible and performed careful review of relevant literature sources and molecular structures to address this limitation.

In this study, up to 30% participants in the EPIC-Norfolk cohort were identified as carriers of at least one of the 15 pathogenic IFVs detected. However, the numbers of heterozygote and/or homozygote carriers for each IFV would likely be too low to detect biologically meaningful effects. Cohorts of rare disease patients with whole genome sequencing data, such as Genomics England[186], may have larger numbers of carriers and thus be better powered to assess the clinical consequences of these IFVs.

3.6.4. Conclusions

Metabolic loci are enriched for IEM genes and variation at these genes contributes to inter-individual differences in metabolite levels also affected in the corresponding IEM. The large effects of IFVs on metabolite levels may indicate that these variants could also have health or health-related effects. Therefore, further characterisation of these variants as well as phenotypic assessment, which are performed in the following two chapters, may help to identify genetic subgroups of disease for targeted screening and risk management.

CHAPTER 4: CHARACTERISATION OF METABOLITE-ASSOCIATED VARIANTS AT IEM GENES

4.1. Abstract

Background Rare mutations known to cause IEMs have large effects on metabolite levels, but the effects of more frequent, and even common, variation at such genes is poorly described. This study aimed to systematically assess whether metabolite-associated variants at IEM genes demonstrate similarly strong metabolic effects to those described in the corresponding IEM.

Methods Characterisation of the associations of IEM-affected metabolites with 202 variants at IEM genes in the Metabolon mGWAS was performed based on a) proportion of variance in metabolite levels explained, b) variant function, c) likelihood of carriers being in the highest or lowest 2.5th percentile of the population distribution of the tested metabolite, and d) presence of non-additive effects on metabolite levels. Genetic variants causing an IEM according to the ClinVar database were excluded. Analyses were performed in up to 10,581 participants of the EPIC-Norfolk cohort and adjusted for age, sex, four principal components and measurement batch. To prioritise variants with metabolic effects most likely to translate into downstream clinical consequences, a subset of variants that were either associated with extreme metabolite levels or displayed non-additive effects were queried in databases for previously reported phenotypic associations.

Results Most of the identified IFVs (187 of 202) were not included in the ClinVar database as pathogenic for the corresponding IEM. These 187 IFVs varied widely in terms of the variance they explained, from the extreme (38.7%) to very small ($8.5x10^{-5}$%) genetic contributions. A total of 48 of the 187 IFVs were significantly associated with extreme metabolite levels (Bonferroni-corrected $p \leq 4.9x10^{-5}$) and seven had additional, significant (Bonferroni-corrected $p \leq 1.5x10^{-4}$) non-additive effects on associated metabolite levels. Variants consistently highlighted in these analyses such as the *CPS1* variant rs1047891, the *ACADS* variant rs2014355 and the *UGT1A1* variant rs1976391, were associated with clinical outcomes similar to those observed in the corresponding IEM.

Conclusion Variants mapping to known IEM genes in the general population demonstrated strong effects on metabolite levels that could be linked to clinical phenotypes, justifying in-depth and systematic phenotypic characterisation across metabolite-associated variants at IEM genes.

4.2. Background

In **Chapter 3** of this thesis, I leveraged knowledge of the metabolic effects of IEMs to show that common variants at the corresponding, causative genes can affect similar changes in levels of IEM-affected metabolites. Assessment of the absolute effect sizes of detected variants at IEM genes showed that many had large metabolic effects. Further systematic, in-depth characterisation of metabolite-associated variants at these genes could identify those that have metabolic effects large enough to potentially translate into phenotypic consequences.

Systematic metabolic characterisation may be guided by knowledge of how rare mutations known to cause IEMs affect metabolite levels. For example, rare mutations known to cause IEMs also affect metabolite levels with specific patterns of association that can be used to guide the systematic characterisation of IFVs prioritised in **Chapter 3**. For example, extreme metabolic perturbations caused by rare, IEM-causing mutations have been used to infer that rare mutations explain large proportions of variance in metabolite levels. Indeed, sequencing studies of IEMs have also identified more common variants at IEM genes that exhibit intermediate yet tangible effects on protein activity and metabolite levels. One example is the DBH variant rs1611115 (GnomAD[187] MAF=0.19). While rare mutations at *DBH* can cause extremely low levels of DBH activity, leading to sympathetic noradrenergic dysfunction (OMIM #223360)[188,189], rs1611115 has been shown to account for substantial variance (35-52%) in DBH activity levels in a group of healthy individuals[190].

Rare mutations are also often located in protein-coding regions of the genome and thus affect protein function. Variant annotation tools such as VEP[87] enable functional prediction of genetic variants based on aligned mRNA sequences, genomic location and homology across species. These tools can be used to test whether more common variation at IEM genes also affect protein function by affecting the amino acid sequence of the protein or instead regulate protein activity.

The extreme effects of rare mutations on metabolite levels also suggests that common variants at the same gene may also predispose carriers to more extreme levels of metabolites affected in the corresponding IEMs. For IEMs, the definition of 'extreme' metabolite levels is typically based on reference values measured in children, as IEMs often manifest in early life[191]. Despite this challenge, access to individual-level metabolomic data enables the measurement of population distributions for metabolite levels. This data can be used to define 'extreme' metabolite levels, the thresholds of which are recommended to be the extreme 2.5th percentiles of the population distribution[88–90].

Rare, IEM-causing mutations are also known to have non-additive effects on metabolite levels, which in part account for autosomal dominant or recessive modes of inheritance of IEMs. Findings from a previous mGWAS have also shown that more common variants at IEM genes may also display non-additive effects on metabolite levels (such as the *ACADS* variant rs3916 on butyrylcarnitine levels and the *CPS1* variant rs715 on glycine levels)[21].

In this study, I used individual-level genetic and metabolomic data from the EPIC-Norfolk population cohort to systematically characterise the metabolic consequences of variants at IEM genes identified in the previous chapter based on a) variant function, b) the proportions of variance of metabolite levels explained, c) association with extreme levels of IEM-affected metabolites, and d) evidence for non-additive effects on metabolite levels. If demonstrated, these characteristics may highlight common variants that contribute to downstream phenotypic consequences.

One important caveat is that metabolic effects may not manifest as clinical outcomes. At the variant level, this could be due to the inhibitory effects of a genetic variant on another ('epistasis') that may be detected indirectly by modelling genotypes in dominant or recessive models[192]. At the metabolite level, other metabolic pathways may be up- or down-regulated to compensate for potentially deleterious perturbations along one pathway ('metabolic canalisation'[193]). Despite this, phenotypic assessment of genetic variants can be performed using publicly available GWAS summary statistics to test whether variants displaying similar characteristics to IEM-causing mutations at the same genes have detectable clinical consequences.

4.3. Aim and Objectives

The aim of this study was to identify and describe characteristics of the identified variant-metabolite associations for IFVs. The objectives of this study were to:

1. Estimate the proportions of variance in metabolite levels explained by associated genetic variants, individually and cumulatively;
2. Characterise variant function;
3. Estimate the association of IFVs with extreme metabolite levels;
4. Estimate non-additive effects of IFVs on metabolite levels, and
5. Assess the clinical consequences of IFVs with large metabolic effects, as characterised in the previous objectives.

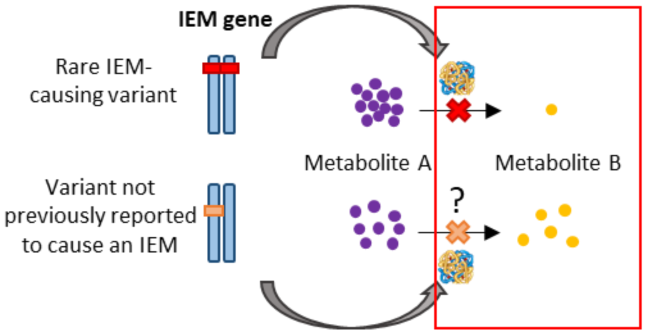


Figure 1: Overview of the study aim. With the identification of IFVs in the previous chapter, the aim of this study (highlighted in the red box) was to systematically characterise their metabolic associations.

4.4. Methods

4.4.1. Study Design and Participants

The European Prospective Investigation into Cancer and Nutrition (EPIC)-Norfolk[107] study is a prospective cohort study of 25,639 individuals aged 40-79 years at baseline and recruited from 35 practices across Norfolk in 1993-1997. The study was approved by the Norfolk Research Ethics Committee (ref. 98CN01) and all participants gave signed informed consent. The EPIC-Norfolk comprises of two non-overlapping sets of participants (total n=13,475):

Type 2 diabetes (T2D) case-cohort - the design of the EPIC-Norfolk nested T2D case-cohort study, including ascertainment and verification of incident T2D cases has previously been described in detail[108]. In brief, it includes 1,503 individuals (45% cases) ascertained using self-report, linkage to primary and secondary care, drug register, hospital admission, and mortality data[108].

Subcohort – a subcohort of 11,972 participants who were not part of the T2D case-cohort was drawn from all EPIC-Norfolk participants.

4.4.2. Measurements

4.4.2.1. Genetic profiling

Participant genomes were genotyped using the UK Biobank Affymetrix Axiom Array and imputed using the Haplotype Reference Consortium[109] as well as the combined UK10K[111]/1000 Genomes[110] imputation panels. Genotypes were used where possible; otherwise, imputed dosages that were converted to hard-calls (controls: dosage ≤ 0.2, heterozygotes: 0.9 ≤ dosage ≤ 1.1, homozygotes: dosage ≥ 1.8) were used. In **Section 4.4.3.2.**, the 202 IFVs identified in **Chapter 3, Section 3.5.4.** were used while in **Sections 4.4.3.3.** and **Sections 4.4.3.5.** only the 187 IFVs that were not previously reported to cause an IEM in ClinVar were included in analysis **(Chapter 3, Section 3.5.5.)**.

4.4.2.2. Metabolomic profiling

Metabolomic profiling was performed on non-fasted citrate plasma samples obtained at baseline using LC-MS/MS in the untargeted Metabolon HD4 Discovery^{TM} platform. Samples for metabolic profiling were selected in the order in which they were stored at baseline (quasi-random selection). Measurement was performed in three stages, one for the T2D case-cohort and for two for roughly equal sets of 6,000 participants in the subcohort. Metabolite levels were log-transformed, winsorised to 5 SDs of the mean and standardised. Residuals of metabolite levels adjusted for age, sex and the first four principal components were then calculated.

4.4.2.3. Exclusions

Eight participants with high rates of missing metabolite measurements and 645 participants who were T2D cases not in the subcohort fraction of the T2D case-cohort study were excluded. In addition, 1,286 participants without genotypic data and 955 participants who were related to other participants in EPIC-Norfolk were excluded, leaving 10,581 participants in the EPIC-Norfolk subcohort and the subcohort fraction of the T2D case-cohort for statistical analysis. Levels of 646 metabolites that were associated with at least one conditionally independent lead variant in the Metabolon mGWAS **(Chapter 3)** were included in analysis.

4.4.3. Statistical Analysis

4.4.3.1. Estimate of proportions of variance explained by genetics on plasma metabolite levels

Variants explaining large proportions of variance on metabolite levels were hypothesised to be more likely to contribute to more distal complex phenotypes or diseases. To estimate the proportion of variance explained by genetics for plasma levels of 237 IEM-related metabolites, linear regression and ANOVA models were used, and the proportions explained by the 202 IFVs individually and cumulatively were quantified. Analyses were adjusted for age, sex, measurement batch and the first four principal components of genetic ancestry. A binomial test for enrichment was used to test for differences in metabolite class membership between metabolites for which over half of the genetics-explained variance was driven by IFVs and those explained by other variants (Bonferroni significance p=0.05/8 metabolite classes=0.00625).

4.4.3.2. Functional variant annotation and enrichment assessment

The most severe predicted functional effect for each IFV and variants in LD ($R^2 \geq 0.95$) was identified using Ensembl Variant Effect Predictor (VEP)[87] Build 37 (downloaded May 2020). VEP annotates functional consequences by integrating information from aligned mRNA sequences from the NCBI

Reference Sequence Database RefSeq, genomic location and homology across species. Functional VEP categories were tested for enrichment for IFVs compared to metabolite-associated variants at other genes (detected in the Metabolon mGWAS) using a Fisher's test (significant p≤0.05).

4.4.3.3. Association with extreme metabolite levels (metabolite extremes assessment)

One-tailed logistic models were used to test whether carriers of candidate IFVs **(Chapter 3, Section 3.5.5.)** were more likely to be associated with extremes of the population distribution of metabolite levels compared to non-carriers, as would be expected from the IEM. The outcome was 'extreme metabolite level', which was defined as being below the 2.5th or above the 97.5th percentiles of the population distribution, as recommended by previous publications[88–90]. The direction of effect assessed was based on the direction of effect of the associated variant. Thus, an one-tailed logistic regression test estimating the likelihood of carriers having extreme metabolite levels compared to controls was performed for variant-metabolite pairs with at least five individuals in all strata. Models were performed separately for each measurement batch and adjusted for age, sex and the first four principal components of genetic ancestry, then meta-analysed using a fixed effects model in the R package metafor (v2.4-0). Significant associations were identified based on consistency with mGWAS direction of effect and on a Bonferroni-corrected threshold (0.025/512 independent tests).

4.4.3.4. Estimate of non-additive effects of IFVs on associated metabolite levels

Previous studies have shown that rare variants causing IEMs display non-additive effects on metabolite levels[194]. Therefore, it was hypothesised that some of the IFVs identified in **Chapter 3** could have non-additive effects on metabolite levels. For any association where the metabolite was implicated in IEM aetiology, non-additive effects were estimated in a linear regression model using metabolite residuals as the outcome and a recessively coded genotype as the exposure, adjusting for effects due to an additively coded genotype. Only variants with at least five homozygous carriers of the minor allele were considered. Models were performed separately for each measurement batch and adjusted for age, sex and the first four principal components of genetic ancestry, then meta-analysed using a fixed effects model in the R package metafor (v2.4-0). Significance was assessed at Bonferroni-corrected threshold (p≤0.05/334 independent tests).

4.4.3.5. Phenome-wide assessment of variants with large metabolic effects

To assess whether IFVs with large metabolic effects could be linked to phenotypic associations, phenotypic assessment was performed on the subset of variants that either i.) were significantly associated with extreme metabolite levels in a direction consistent with that observed in the Metabolon mGWAS **(Section 4.4.2.3.)** or ii.) had non-additive effects on metabolite levels **(Section 4.4.2.4.)**. Queries were performed in search engines such as Google Scholar (https://scholar.google.com/) and NCBI PubMed (https://pubmed.ncbi.nlm.nih.gov/) as well as in the publicly available databases GWAS Catalog[62], PhenoScanner[73], OpenGWAS[63], and OpenTargets[74]. Variants in LD ($R^2 \geq 0.8$) were also queried, and associations that met a relaxed significance threshold of $p=1x10^{-5}$ were assessed for relevance to corresponding IEM symptoms. A more comprehensive description of the phenome-wide assessment of these variants, and others, is available in **Chapter 5**.

All statistical analyses and graphics were performed and produced using R version 3.5.3.[128] and STATA version 14.2[129].

4.5. Results

4.5.1. IFVs Accounted for Substantial Proportions of Variance in Metabolite Levels Explained by Genetics

The 202 IFVs were associated with 237 unique metabolites known to be affected in the corresponding IEMs. Individual variants accounted for a median of 0.46% (range: $8.5x10^{-5}$;38.7%) proportion of variance in corresponding metabolite levels. Common variants explained a median of 0.48% variance (range $8.5x10^{-7}$;38.7%) compared to low-frequency (0.42% (0.023;23.4%)) and rare (0.35% ($6.3x10^{-6}$;4.8%)) variants.

IFVs cumulatively explained over half of the total variance explained by genetics in 127 of 237 (53.6%) associated metabolite levels **(Figure 2)**. The cumulative proportion of variance explained by the 791 independent mGWAS detected variants on associated metabolite levels had a median of 2.23% with a range of 0.082% (docosahexanoate) to 46.8% (ethylmalonate). In the most notable example, 96% of the variance in ethylmalonate levels explained by genetics (total percentage variance explained = 46.8%) could be attributed to IFVs **(Figure 2)**. When considering individual proportions of variance explained for ethylmalonate, 82.7% of the variance attributable to genetic variation could be explained by the *ACADS* IFV rs2014355 (MAF=0.25) alone. Metabolites for which over half of the proportion of variance explained by genetics was accounted for by IFVs were significantly enriched for all classes ($p \leq 0.00625$) compared to metabolites driven by other variants except for energy metabolites, which were depleted, and lipids and xenobiotics, which were not significant ($p>0.00625$). However, these results may be unreliable due to the small numbers of metabolites present in many of the metabolite classes.

4.5.2. IFVs May Alter Metabolite Levels Through Effects on Protein Activity or mRNA Splicing

Of the 187 IFVs not previously reported to cause an IEM, 41.7% were predicted to be intronic **(Figure 3)**. A further 10.6% had intergenic, up- or downstream gene, regulatory region, transcription factor binding site or non-coding transcript exon annotations, all of which are predicted to have regulatory or modifier effects in Ensembl **(Figure 3)**. In addition, 27.2% IFVs were missense, suggesting that they alter the protein-coding sequence, and 12.9% were loss of function (including splice region, 5'/3' untranslated region, frameshift or stop gained/lost variants). No evidence of enrichment of IFVs within any VEP category was found compared to metabolite-associated variants at other genes (Fisher's p>0.05 for all categories tested) **(Figure 3)**.

4.5.3. A Total of 48 of 187 IFVs Were Associated With Extreme Metabolite Levels

Of the strongest metabolite associations for the 187 variants, 11 had an absolute effect size ≥0.5 (per allele per 1-SD metabolite difference). The number of associations with absolute effect size ≥0.5 increased to 75 when also considering additional conditionally independent associations.

A total of 104 of 512 assessed variant-metabolite pairs were significantly associated ($p \leq 4.9 \times 10^{-5}$) with higher or lower odds for extreme metabolite levels. Of these, 101 (representing 48 IFVs) had consistent directions of effect with the corresponding Metabolon mGWAS association. Of the 48 IFVs associated with extreme metabolite levels, 19 also had an absolute effect size ≥0.5.

The median OR of having extreme metabolite levels was 2.62 (range: 1.71-76). The maximum observed odds ratio was for the rare missense *DMGDH* variant rs145258663 (MAF=0.0050); despite wide confidence intervals, carriers had a 76-fold higher likelihood of having dimethylglycine levels in the 90th percentile of the distribution (OR (95% CI) = 76 (48.3;119), $p=1.1 \times 10^{-78}$) **(Figure 4)**. Other rare, conditionally independent variants at *DMGDH* had large ORs in this analysis: rs142181836 (MAF=0.0011) and rs184410852 (MAF=0.0014) with extremely high dimethylglycine levels (OR (95% CI) = 41.3 (15.3;112), $p=2.5 \times 10^{-13}$ and OR (95% CI) = 41 (11.6;142), $p=6.7 \times 10^{-9}$, respectively). No ClinVar annotations were found for rs145258663, rs142181836, rs184410852 or their proxies.

Several variants at the common end of the frequency spectrum also conferred risk of having extreme metabolite levels. Examples included the *ACADS* variant rs2014355 (MAF=0.25) with extremely high levels of ethylmalonate (OR (95% CI) = 23.5 (11;50.5), $p=5.3 \times 10^{-16}$) and butyrylcarnitine (OR (95% CI) = 23.8 (11;51), $p=4.1 \times 10^{-16}$) and the *UGT1A1* variant rs1976391 (MAF=0.31) with extremely high bilirubin levels (Z,Z) (OR (95% CI) = 18.5 (8.2;42.1), $p=3.2 \times 10^{-12}$) **(Figure 4)**. Both variants were predicted directly or through a linked variant ($R^2 \geq 0.6$) to be benign for the corresponding IEMs in ClinVar.

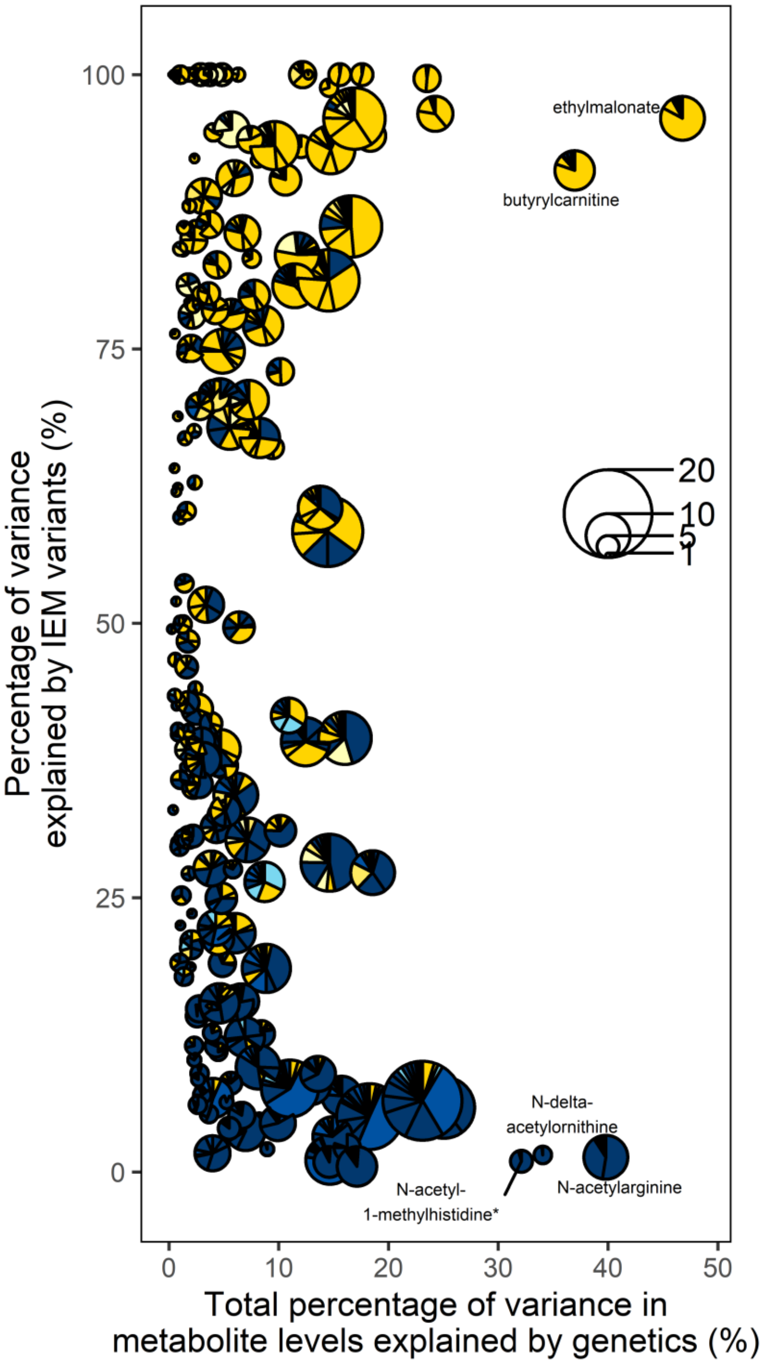


Figure 2: **Cumulative proportion of variance explained across 237 metabolites associated with at least one IFV.** Each pie chart represents one metabolite, with size corresponding to the number of associated variants. Metabolite levels for which >30% proportion of variance were explained by genetics are labelled. Yellow represents variance explained by IFVs while blue represents variance explained by variants at other genes. Lighter shadings represent low-frequency or rare variants.

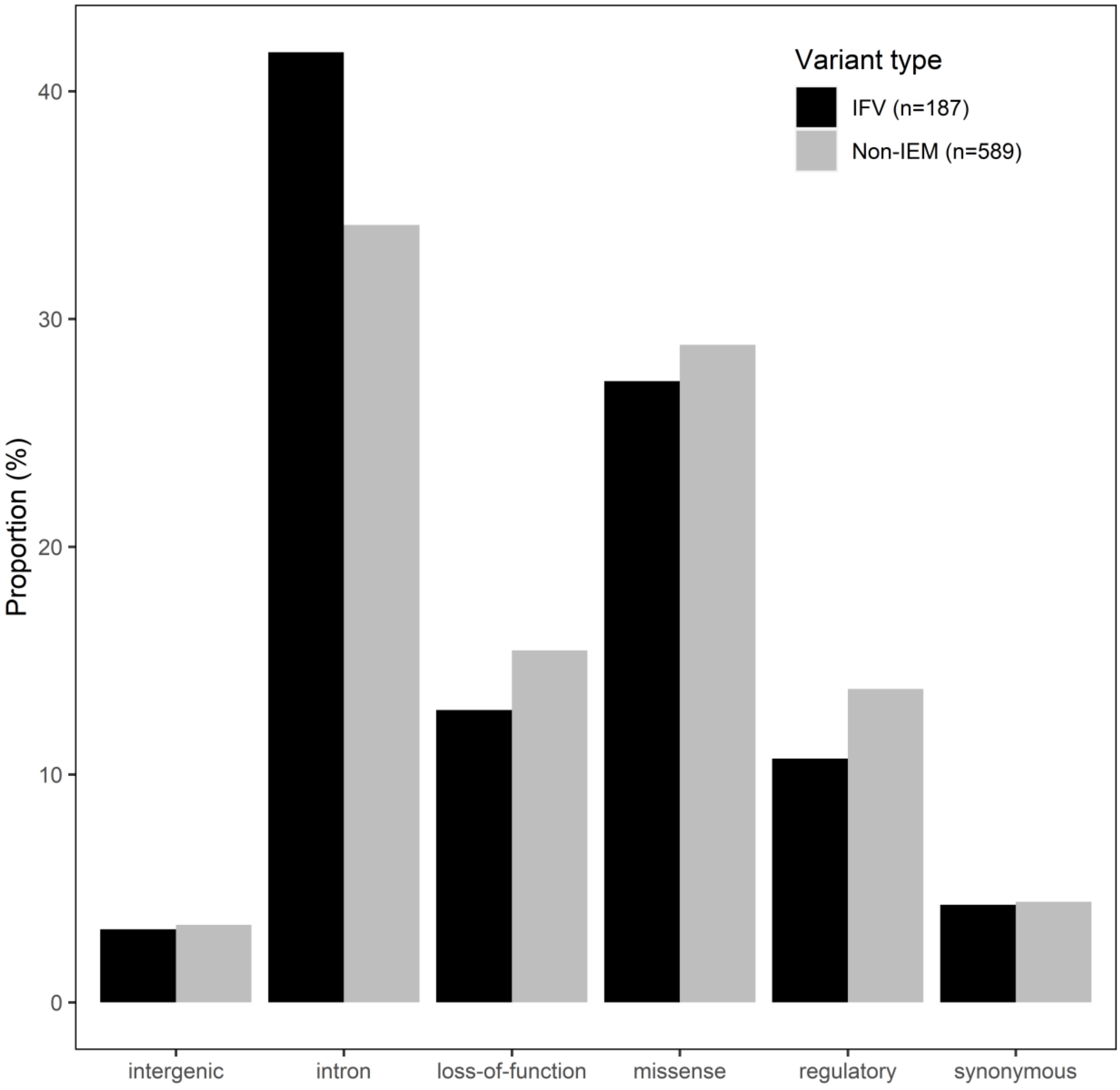


Figure 3: Most severe variant consequence annotations by Variant Ensembl Predictor across 187 IFVs and 589 metabolite-associated variants at other genes. Fifteen IFVs annotated as likely-pathogenic or pathogenic for the IEM in ClinVar were excluded from this comparison. The ‘loss of function’ category includes splice region, donor or acceptor variants, 5’/3’ UTR variants, frameshift variants and stop gained or lost variants. The ‘regulatory’ category includes up- or downstream gene, regulatory region, transcription factor binding site or non-coding transcript exon variants.

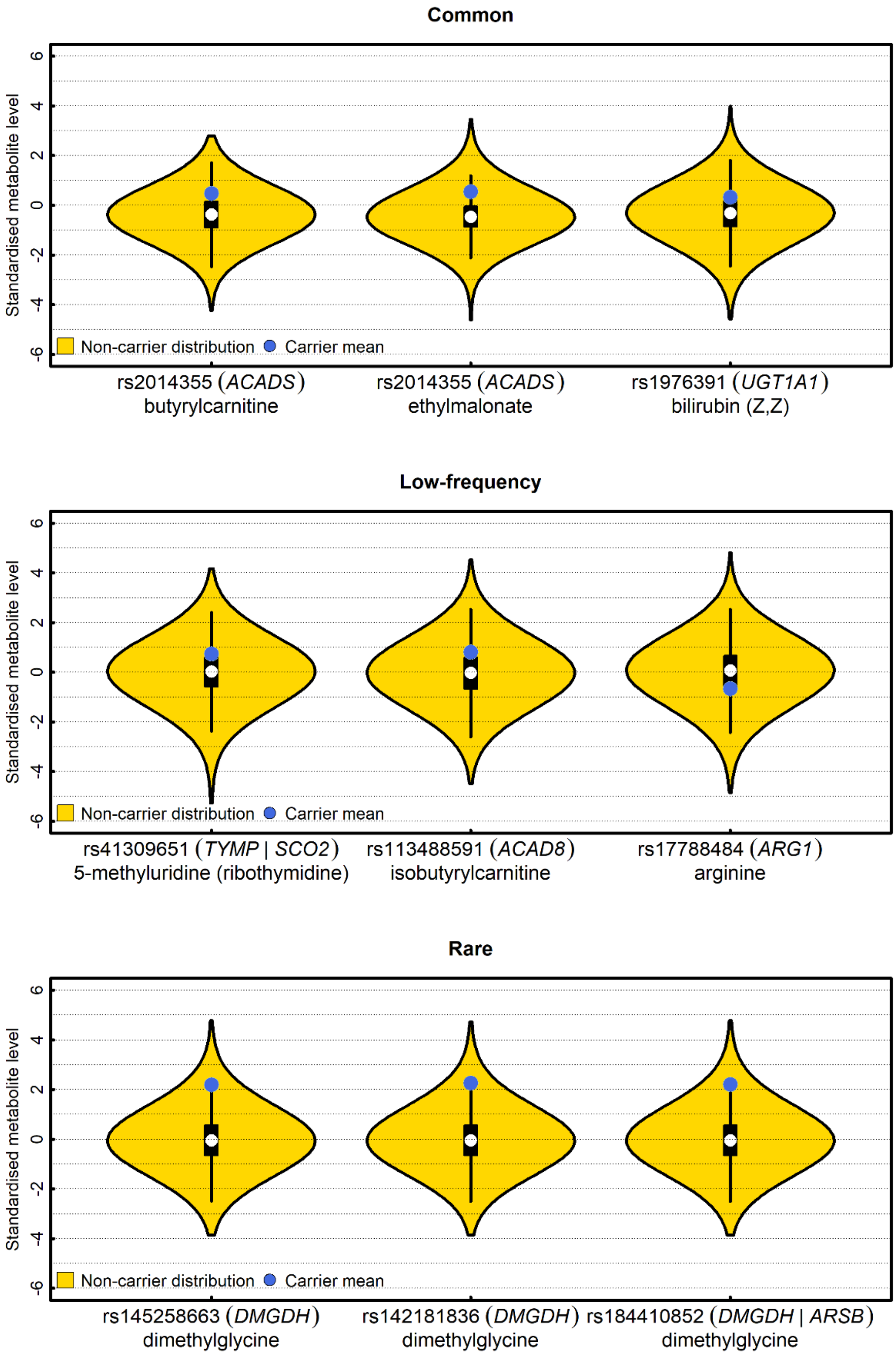


Figure 4: Carrier means plotted onto non-carrier metabolite level distributions for the top three strongest variant-extreme metabolite level associations (p≤4.8x10^{-5}) by allele frequency category. Carriers include homozygotes and heterozygotes. Common: MAF≥0.05; Low-frequency: 0.01≤MAF<0.05; Rare: MAF<0.01.

The association of several IFVs with extreme metabolite levels suggested the presence of non-additive effects on metabolite levels. Broad assessment identified 20 associations for which 7 IFVs displayed significant non-additive effects across 17 metabolites ($p \leq 1.5x10^{-4}$). In 18 (90%) of these associations, the IFV was also significantly associated ($p \leq 4.8x10^{-5}$) with extreme levels for that metabolite, although associations for the *UGT1A1* variant rs201829156 were directionally inconsistent to what was expected from the Metabolon mGWAS **(Figure 5)**. Of the two that were not associated with extreme metabolite levels, the *GCDH* variant rs8012 (MAF=0.45) was associated with extremely low glutarylcarnitine levels at nominal significance but did not reach Bonferroni-corrected significance (OR (95% CI) = 2.0 (1.4;3), $p=2.8x10^{-4}$). The other (*LCT* variant rs4988235 (MAF=0.30)) could not be tested due to a lack of power (i.e. <5 controls with metabolite levels in the extreme 97.5th percentile of the population in any tested subcohort).

Of the seven variants with non-additive effects on metabolite levels, five were associated with an even greater increase in metabolite levels compared to what would be expected under an additive model **(Figure 5)**. In contrast, two variants (the *LCT* variant rs4988235 and the *CPS1* variant rs1047891) were associated with a lower increase compared to that expected under an additive genetic model **(Figure 5)**. Accounting for non-additive effects in the computation of metabolite variance explained an additional median variance of 1.22% (range: 0.21;13.7%) across the 17 assessed metabolites.

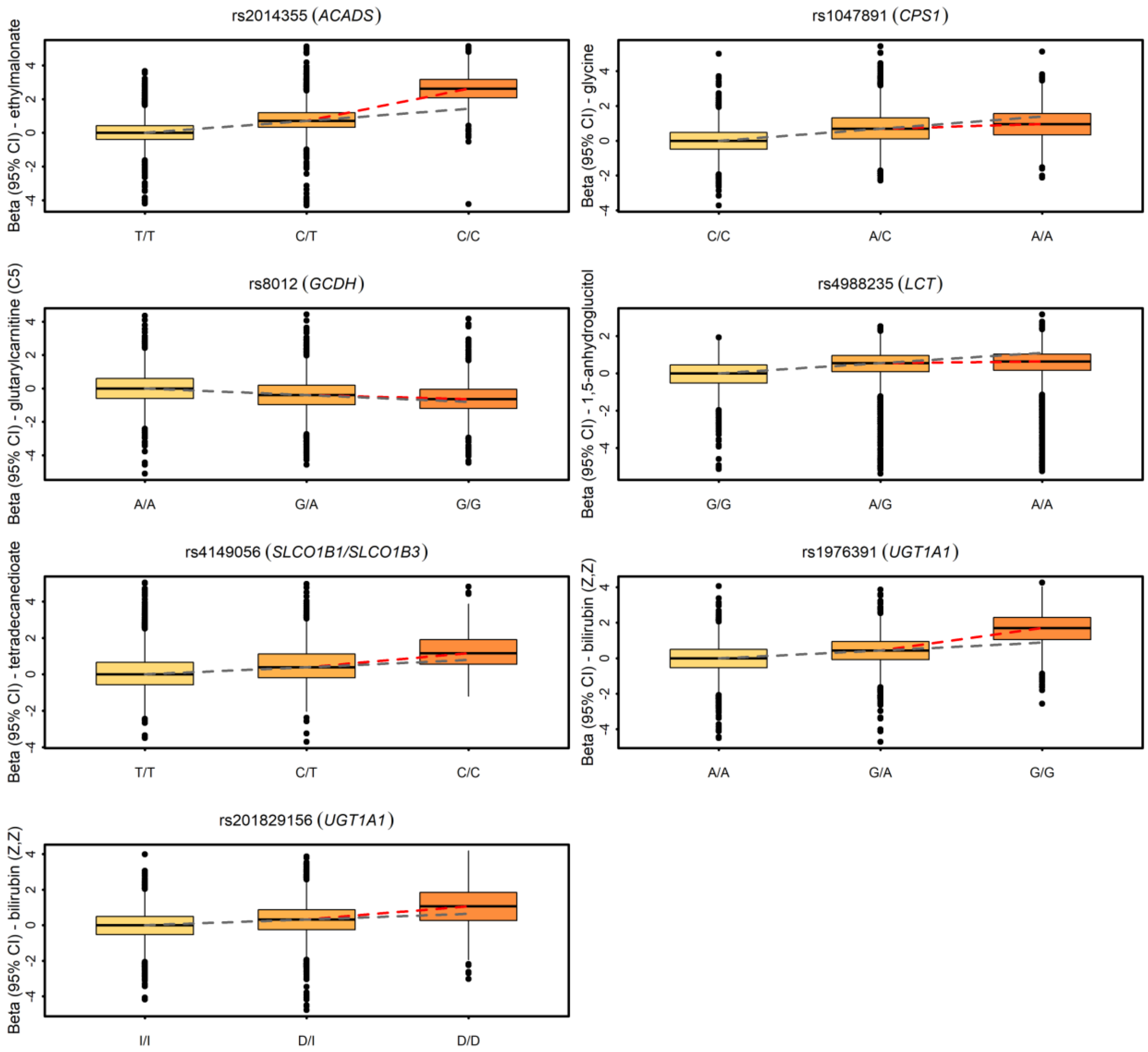


Figure 5: Strongest variant-metabolite associations displaying significant departure from linearity at Bonferroni significance ($p \leq 1.5 \times 10^{-4}$). The dashed grey line represents the expected trend under a linear model while the dashed red line represents the observed trend. *UGT1A1* and other *UGT* paralogs are mapped to the locus containing variants rs1976391 and rs201829156 but here labelled as '*UGT1A1*' for brevity.

4.5.4. IFVs With Phenotypic Consequences That Could Plausibly Arise from Their Metabolic Effects

Of the 187 IFVs assessed, 51 were either significantly associated with an increased likelihood of having extreme metabolite levels or had non-additive effects on metabolite levels. For 21 of the 51 assessed variants, phenotype associations in GWAS studies ($p \leq 1x10^{-5}$) that were related to one or more symptoms of the corresponding IEM were identified, indicating a need for further testing of a shared genetic signal at the corresponding loci. Three of the most notable examples are highlighted here, and the rest are summarised in the full phenome-wide assessment in **Chapter 5 (Appendix Ch5_ST4)**.

4.5.4.1. The CPS1 variant rs1047891 indicates a role of ammonia and glycine metabolism in chronic kidney disease

The association of the *CPS1* variant rs1047891 with glycine levels serves as a prime example for which health consequences in the general population have been identified. The *CPS1* gene encodes carbamoyl phosphate synthetase, which catalyses the first step of ammonia catabolism in the urea cycle[158]. Rare *CPS1* mutations can lead to CPS1 deficiency (OMIM #237300) and hyperammonemia in the blood, which in turn results in vomiting, muscle weakness and psychomotor delay if left untreated[159]. Although ammonia was not detected in the Metabolon mGWAS study, an association was detected between the A-allele of the *CPS1* variant rs1047891 (effect allele frequency (EAF)=0.32) with elevated plasma levels of glycine (beta±S.E.=0.53±0.013, p=6.9x10^{-385}), which is interconverted with ammonia through several pathways through the glycine cleavage complex[160]. The association of rs1047891 with glycine has also been identified in previous mGWASs[18,19]. Metabolic characterisation of rs1047891 effects on glycine showed that rs1047891 is associated with extremely high glycine levels (OR (95% CI): 14.6 (8.5; 25.0), p=3.3x10^{-22}) and had non-additive effects on glycine that reduced the expected effects under an additive model (beta±S.E.=-0.51±0.057, p=7.47x10^{-9}) **(Figure 5)**.

In an independent GWAS study, the same glycine-elevating allele of rs1047891 was associated with an increase in chronic kidney disease risk, as measured by a decrease in estimated glomerular filtration rate[56,195]. There is a plausible link between *CPS1* gene function, glycine and chronic kidney disease. The variant rs1047891 could reduce the efficiency of CPS1 in breaking down ammonia, which in turn could drive the conversion of ammonia to glycine as an alternative route of catabolism **(Figure 6A)**. Chronic exposure to elevated ammonia levels in the blood may be linked to reduced ammonia excretion, which in turn could lead to reduced kidney function and chronic kidney disease[196,197]. This example shows how severe metabolic dysregulation and mild, longer-term changes in the same metabolic pathway may lead to impaired organ function. While current evidence suggests that rs1047891 is the likely causal variant for glycine and chronic kidney disease associations within the locus, this has not been experimentally validated.

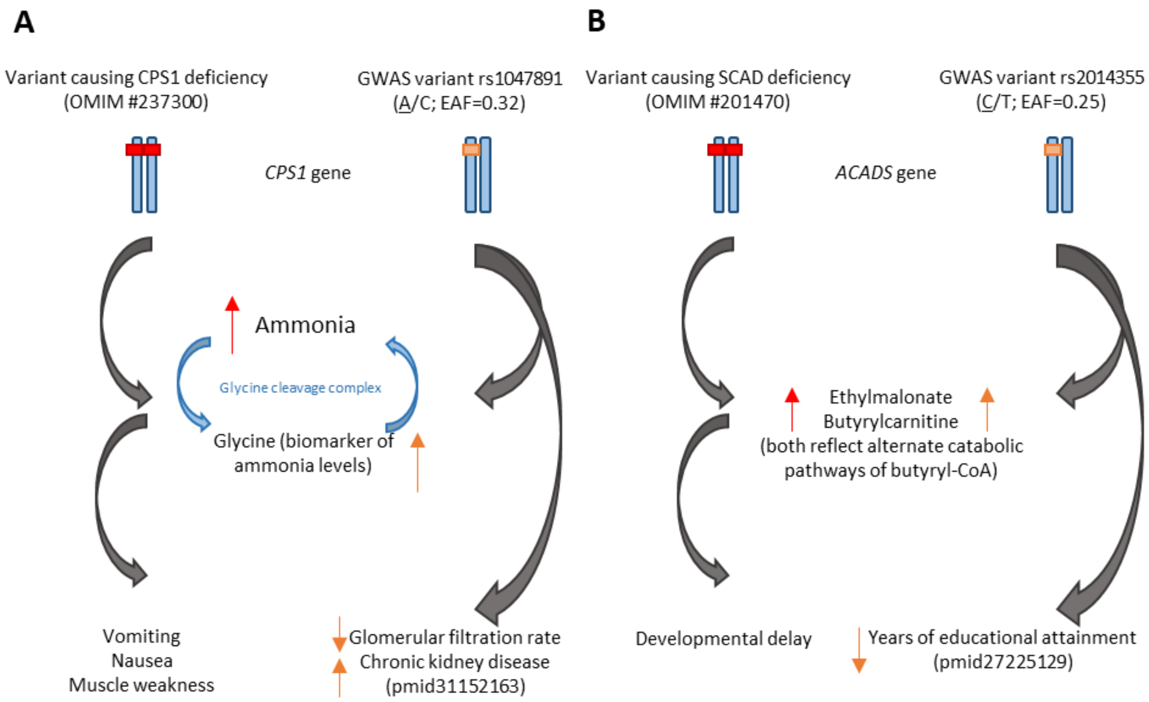

A
Variant causing CPS1 deficiency (OMIM #237300)
GWAS variant rs1047891 (A/C; EAF=0.32)
CPS1 gene
Ammonia
Glycine cleavage complex
Glycine (biomarker of ammonia levels)
Vomiting
Nausea
Muscle weakness
Glomerular filtration rate
Chronic kidney disease
(pmid31152163)
B
Variant causing SCAD deficiency (OMIM #201470)
GWAS variant rs2014355 (C/T; EAF=0.25)
ACADS gene
Ethylmalonate
Butyrylcarnitine
(both reflect alternate catabolic pathways of butyryl-CoA)
Developmental delay
Years of educational attainment
(pmid27225129)

Figure 6: Comparison of metabolic and phenotypic consequences caused by rare mutations and more common variants at the *CPS1* and *ACADS* genes. (A) Identification of associations of the *CPS1* variant rs1047891 with glycine levels and chronic kidney disease that could potentially reflect long-term effects of dysregulated ammonia catabolism in CPS1 deficiency (OMIM #237300). **(B)** Identification of associations of the *ACADS* variant rs2014355 with butyrylcarnitine and ethylmalonate levels as well as with years of educational attainment, which could potentially reflect dysregulated short-chain acylcarnitine catabolism and developmental delay in SCAD deficiency (OMIM #201470). In both examples, a causal pathway from variant to metabolite to trait or disease is plausible but has not been experimentally validated or tested using causal inference methods.

4.5.4.2. The ACADS variant rs2014355 is associated with ethylmalonate and with educational attainment, which may reflect symptoms of developmental delay in SCAD deficiency

Metabolite associations of the missense *ACADS* variant rs2014355 demonstrate how IFVs may affect complex phenotypes that reflect specific symptoms of the corresponding IEM. The *ACADS* gene encodes mitochondrial short-chain acyl-CoA dehydrogenase, which breaks down short-chain fatty acids in fatty acid oxidation[198]. The ACADS enzyme preferentially takes butyryl-CoA as its substrate, therefore, rare mutations causing protein inactivity result in elevated levels of metabolic by-products from alternative routes of butyryl-CoA catabolism including butyrylcarnitine and ethylmalonate[199] **(Figure 6B)**. Accordingly, the C-allele of the *ACADS* variant rs2014355 (EAF=0.25) was significantly associated with elevated levels of these by-products in the Metabolon mGWAS. Furthermore, rs2014355 was associated with extremely high levels of ethylmalonate (OR (95% CI):23.5 (11.0;50.5), p=$5.3x10^{-16}$) and butyrylcarnitine (OR (95% CI): 23.8 (11.1;51.0), p=$4.1x10^{-16}$) **(Figure 4)**. Significant non-additive effects were also observed for these associations, suggesting that the large metabolic effects observed at the *ACADS* locus could translate into downstream health consequences **(Figure 5)**.

In the IEM, SCAD deficiency (OMIM #201470), common symptoms include vomiting, failure to grow at the expected rate, muscle wasting, seizures and developmental delay[198]. In GWAS studies, the C-allele of rs2014355 (also the butyrylcarnitine/ethylmalonate raising allele) is significantly associated with a decrease in years of educational attainment (p=4.99x10^{-6}, N=293,723)[200], a cognitive phenotype that may in part reflect delayed speech and language development[201]. The potential relevance of these phenotypes and their link to metabolite levels warrants further phenotypic assessment.

4.5.4.3. Strong effects of UGT1A1 variants on bilirubin metabolism are reflected in IEM-related outcomes and drug response phenotypes

The conditionally independent variants rs1976391 and rs201829156 are located at a locus containing the IEM gene *UGT1A1* (as well as other paralogs) and represent another notable example where phenotypic associations in the literature mimic those of the IEM. The *UGT1A1* gene belongs to a family of UDP-glucuronosyltransferases and encodes the only enzyme in the family that can glucuronidate bilirubin[202]. Rare mutations in *UGT1A1* can result in two IEMs, Gilbert syndrome (OMIM #143500) and Crigler-Najjar syndrome type I (OMIM #218800). These IEMs are distinct in clinical presentation, though symptoms of hyperbilirubinemia, jaundice and liver dysfunction are observed in both[203–205]. In the Metabolon mGWAS, the G-allele of rs1976391 (EAF=0.31) was positively associated with biliverdin and the 'E' and 'Z' isomers of bilirubin while the deletion allele of rs201829156 (EAF=0.35) was negatively associated with these metabolites. Both variants were associated with extreme metabolite levels **(Figure 4)** and had non-additive effects on metabolite levels **(Figure 5)**. In VEP, the most severe consequence predicted based on linked variants was non-coding transcript exon variant for rs1976391 and missense variant for rs201829156. However, the most severe consequence based on the exact variant was 'intron', indicating that these variants may have regulatory effects on mRNA processing. Variance in bilirubin (Z,Z) levels was predominantly explained by rs1976391 and rs201829156, which cumulatively accounted for 80-90% of the 25% estimated variance explained by genetics (i.e. ~2% total variance in bilirubin (Z,Z) levels explained). Although only rs1976391 was directionally consistent with the mGWAS direction of effect, associations for both this variant and rs201829156 showed significant departure from linearity.

Though not previously reported to cause Gilbert syndrome or Crigler-Najjar syndrome type I, the bilirubin-raising allele of rs1976391 is in strong LD (R^2>0.8) with the T-allele of the lead SNP in the region, rs887829, which is associated with increased self-report of biliary or pancreas problems in UK Biobank (p=1.7x10^{-9}, n=337,159). In another study, rs1976391 also had a high posterior probability of risk association with the ICD-10 code E80 (‘Disorders of porphyrin and bilirubin metabolism’) (PP_{risk}=0.80), K80 (‘Cholelithiasis’) (PP_{risk}=0.83), R17 (‘Unspecified jaundice’) (PP_{risk}=0.32) and all daughter categories[206] **(Figure 7)**. This same study showed that rs201829156 had a strong posterior probability of risk (PP_{risk}=0.90) association with the ICD-10 code E80 (‘Disorders of porphyrin and bilirubin metabolism’)[206] **(Figure 7)**. All of these phenotypes closely mimic symptoms of the IEMs, with cholelithiasis being a direct consequence of hyperbilirubinemia.

The metabolic effects of IFVs may be large enough to cause differential efficacy or toxicity responses to drugs in carriers compared to non-carriers. For example, atazanavir is a drug used to treat human immunodeficiency virus (HIV) infections, and previous studies have shown that atazanavir inhibits UGT1A1 activity in a concentration-dependent manner, increasing total bilirubin levels as a result. HIV patients who carry one or both copies of the T-allele of the rs887829 variant (which is in LD with the bilirubin-raising allele of rs1976391) are at an increased risk of developing hyperbilirubinemia when taking atazanavir[207]. It has also been shown that pre-screening for rs887829 genotype could markedly reduce the rate of bilirubin-related side effects as well as the rate at which patients discontinue atazanavir treatment[208].

The *UGT1A1* example highlights two key messages about common variation at IEM genes: 1.) that common variation can mimic phenotypic consequences resulting from rare mutations in the same gene, and 2.) that in some cases such as this one, the metabolic consequences of genetic variation can be large enough to impact response phenotypes to medications.

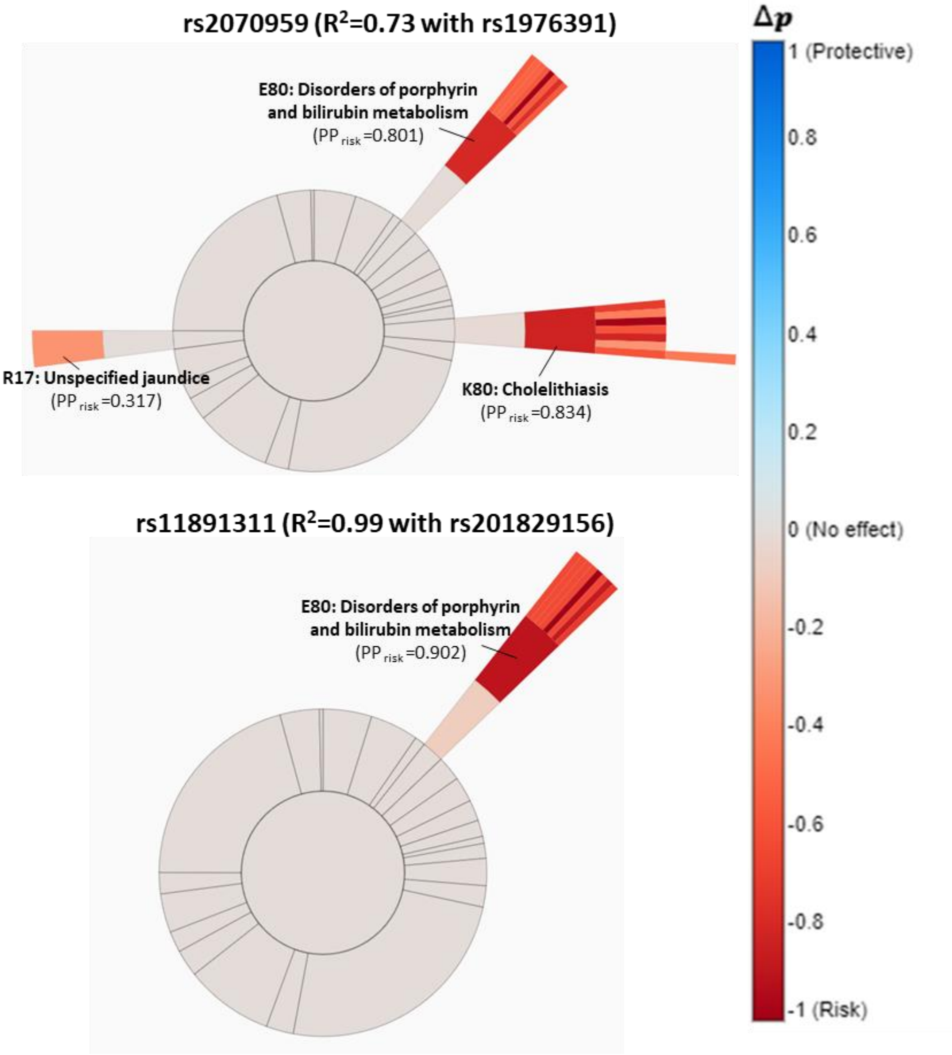


Figure 7: Association of the *UGT1A1* variants rs1976391 and rs201829156 with ICD-10 codes in the UK Biobank, as calculated using the TreeWAS method[206]**.** 'Δp' refers to the difference in posterior probability of risk and posterior probability of protection.

4.6. Discussion

4.6.1. Summary of Findings

Systematic and comprehensive characterisation revealed the large effects of some IFVs on metabolite levels, reflecting patterns of association seen for rare, IEM-causing mutations. Specifically, IFVs individually and cumulatively explained large proportions of variance, and 51 of the 187 assessed IFVs were associated with increased odds of having extreme metabolite levels (<2.5th or >97.5th percentile of the population metabolite distribution) and/or had non-additive effects on metabolite levels. These results showed that several of the detected IFVs have large metabolic effects, warranting further phenotypic assessment.

4.6.2. Novelty of Findings

Here, I observed wide variation in the proportions of variance explained by genetics on plasma metabolite levels. The proportions of variance estimated for 153 of the 237 assessed metabolites were similar to those calculated in a previous study[20] and provide confidence in the estimates of other previously unassessed metabolites presented here. For many metabolites such as ethylmalonate, estimates of proportion of variance explained are considered large compared to those for more distal complex phenotypes (for example, 3,290 near-independent SNPs associated with the highly-heritable height trait cumulatively explain 24.6% variance)[115].

I also observed that 41% of the 187 assessed IFVs were predicted to be intronic variants. Intronic variants are known to regulate gene expression by altering mRNA splicing[209]. An additional 10.6% were predicted to have variant effects relating to the regulation of gene expression. These observations contrasted with the knowledge that IEM-causing mutations often directly affect the protein-coding sequence, instead aligning with previous observations that complex trait-associated variants in GWAS studies are often located in non-coding regions and have regulatory functions[210].

In this study, I also assessed IFVs for association with 'extreme' metabolite thresholds, which were defined using individual-level data. This analysis showed that homozygote and heterozygote carriers of 48 IFVs were significantly more likely than controls to have extreme metabolite levels. Of these IFVs, 19 also had large absolute effect sizes (≥0.5 per 1-SD allele per 1-SD change in metabolite level) for one or more corresponding IEM-related metabolites. Another variant at the *UGT1A1* gene, rs201829156, was also associated with extreme levels of bilirubin, but with contrasting direction of effect to what was observed in the Metabolon mGWAS.

In GWAS studies, it is common practice to model genetic effects using an additive model, ignoring potential effects such as dominance or epistasis. Despite this, many disease-causing variants display non-additive genetic effects, making the detection of such effects useful for identifying variants with potential health consequences. In this study, non-additive effects were detected for seven variants, five of which were also associated with extreme metabolite levels. The *UGT1A1* variant rs201829156 was also one of the five variants, suggesting that non-additive effects could account for the different direction of effect observed in the additive model assessed in the Metabolon mGWAS. These results show that modelling non-additive genetic effects could be useful to identify variants with potential health consequences as a downstream analysis to initial GWAS discovery efforts.

For the other two variants with non-additive effects (the *LCT* variant rs4988235 and the *CPS1* variant rs1047891), the degree of increase with two copies of the metabolite-raising allele was lower than that expected under an additive genetic model. This could be accounted for by two explanations. One explanation is that epistatic effects (i.e. interactions where the effects of a genetic variant are suppressed by another) may exist to suppress the metabolic effects of these variants. Another potential explanation is metabolic canalisation, whereby the effects of one gene are buffered by the effects of other genes with similar function or by alternative routes of metabolism[193]. However, these explanations would suggest a lack of phenotype resulting from these variants, an expectation that is countered in the case of the *CPS1* variant rs1047891 **(Chapter 4, Section 4.5.4.1.)**.

4.6.3. Study Strengths and Limitations

A strength of this study was the use of specific characteristics of variant-metabolite associations (such as the explanation of large proportions of variance in metabolite levels, association with extreme metabolite levels and display of non-additive genetic effects) to prioritise variants with the greatest likelihood of leading to potential downstream health effects. Another strength of this study was the access to individual-level data in a cohort of up to 10,581 participants. This enabled the development of a systematic definition for 'extreme' metabolite level thresholds that i) follows recommended guidelines to define extreme thresholds as being in the lowest or highest 2.5th percentiles of the population distribution[88–90], and ii) can be used instead of reference ranges that are typically measured in children (as IEMs often manifest in early life[191]). The resulting analysis was both comprehensive and robust, enabling the assessment of multiple variants at scale and consistently identifying variants of interest. Indeed, 21 of the 51 variants that were either significantly associated with extreme metabolite levels or displayed non-additive effects were also associated with phenotypes that were similar to those observed in the corresponding IEM. These findings provide evidence to support further assessment of whether these metabolite levels could translate into observable health consequences in the general population.

Despite the systematic and rigorous approach used, there were also some caveats to this study. One was that the proportions of variance in metabolite levels estimated do not account for environmental factors, which are also known to affect metabolite levels. This is especially true given that non-fasted plasma samples were used for metabolite measurement that systematically increase the effects of external factors such as diet on metabolite levels. This could lead to greater uncertainty in estimates of association between genetic variants and metabolite levels. Despite this, the similarity of metabolic consequences of genetic variants to those known to cause IEMs at the same gene, as well as detection of IEM-related phenotypic consequences for some of these variants, suggests that these associations are not wholly confounded by non-genetic factors.

Another limitation is that the use of only one of the two cohorts analysed in the Metabolon mGWAS greatly reduced the power to detect effects for low-frequency and rare variants (MAF≤0.05). Despite this, some of the strongest associations with extreme metabolite levels were observed for rare IFVs, in line with previous observations that rare variants exert larger effects than common ones. Furthermore, the rare IFVs detected in the Metabolon mGWAS were validated using sequencing data, suggesting that their presence, as well as their metabolic effects, are real.

Finally, the metabolite levels used in this study were log-transformed and normalised by Metabolon. These measurements therefore do not represent raw measurements and cannot be used to infer potential clinical effects. Nevertheless, standardisation of metabolite levels was necessary to enable direct comparison across individuals and to define 'extreme' metabolite levels based on the population distribution.

4.6.4. Conclusions

This study showed that several of the tested IFVs have extreme and non-additive metabolic effects and are associated with clinical consequences in GWAS summary statistics. These findings necessitate systematic phenotypic characterisation across IFVs as well as assessment of the shared genetic signals between metabolic and phenotypic traits associated at the same locus.

CHAPTER 5: PHENOTYPIC EFFECTS OF COMMON 'INBORN ERRORS' OF METABOLISM

5.1. Abstract

Background IFVs have been shown to exert strong effects on metabolite levels in the general population and are associated with several complex traits and phenotypes in other studies. To systematically test whether IFVs are also associated with phenotypes similar to those seen in patients with the respective IEM, I developed a pipeline leveraging large-scale data repositories.

Methods Published associations ($p \leq 1x10^{-5}$) were obtained for loci harbouring IFVs and their proxies ($R^2 \geq 0.8$) from the publicly available databases GWAS Catalog, OpenGWAS, OpenTargets and PhenoScanner. This list was pruned to include phenotypes mapping to the respective symptoms of the IEM, as reported by the rare disease databases IEMBase, Orphanet and OMIM. Significant ($p \leq 1x10^{-5}$) phenotype associations in this list were then tested for evidence of a shared genetic signal with IEM-related metabolite associations using statistical colocalisation.

Results A total of 108 loci harbouring 132 IFVs were associated with 1,553 distinct phenotypes with system-wide effects ranging from cancer to neurological, cognitive and psychosocial outcomes. Of these, 45 loci had IEM-related phenotypic associations and were tested for evidence of a shared genetic signal. At 24 loci, metabolic and phenotypic traits had high posterior probabilities of a shared genetic signal (regional posterior probability ≥ 0.7, alignment posterior probability ≥ 0.7). One example was for a signal at *DBH* (rs6271 T/C, EAF=0.074) driving increased 3-methoxytyrosine and dopamine sulfate (2) levels and higher pulse rate ($PP_{regional}$=0.97, $PP_{alignment}$=0.97, >99% likely explained by rs6271), in line with effects on catecholamine biosynthesis and autonomic dysfunction seen in patients with orthostatic hypotension (OMIM #223360).

Conclusion Variation at IEM genes can have metabolic and health consequences mimicking those caused by rare mutations at the same gene. This study identified genetic subgroups underlying complex traits and clinically manifest outcomes and highlighted metabolites as potential biomarkers in disease pathogenesis.

5.2. Background

The identification of genetic variation at IEM genes with large effects on metabolites **(Chapters 3 and 4)** gives rise to the question of whether these same variants contribute to complex traits and diseases in the general population. Understanding the phenotypic consequences of these variants may help to prioritise biomarkers along mechanistic pathways that aid the identification of subgroups among certain complex diseases, i.e. those which have a similar aetiology to the IEM phenotype, and hence tailor diagnosis and possibly treatment.

The distinctive clinical presentations of IEMs could be used to estimate the complex phenotypes resulting from less severe genetic variation **(Figure 1)**. Support for this approach is provided by a previous study, which showed that GWAS gene sets for complex traits are more likely to contain signals for phenotypically similar rare diseases than for phenotypically unmatched ones[71]. Mendelian loci have also been shown to be enriched for complex disease genome-wide association signals[70]. Indeed, several independent GWAS studies have identified examples where the metabolic and phenotypic consequences of common variants mimic those induced by rare, IEM-causing variants at the same gene **(Table 1)**. These findings strengthen the rationale for using IEM knowledge as a framework to map IEM-affected metabolites to complex disease traits.

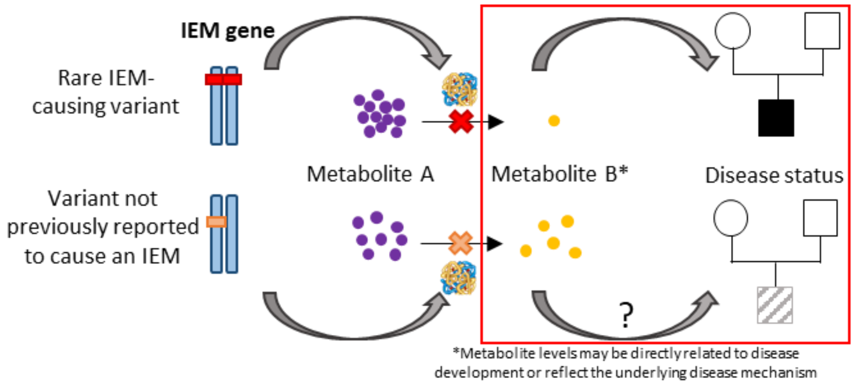

IEM gene
Rare IEM-causing variant
Variant not previously reported to cause an IEM
Metabolite A
Metabolite B*
Disease status
?
*Metabolite levels may be directly related to disease development or reflect the underlying disease mechanism

Figure 1: Overview of the study aim. The aim of this study (highlighted in the red box) was to assess whether the strong metabolic effects of IFVs lead to clinical outcomes, and if so, to what extent they reflect symptoms observed in the corresponding IEM.

Genome-wide association signals mapping to IEM genes are known to affect disease endpoints by disrupting intermediary metabolic pathways that are known to cause the toxic accumulation or deficiency of metabolites **(Table 1)**. In the case of genes known to cause familial dyslipidemia, common variants at these genes are known to result in hyperlipidemia, which is a major risk factor of coronary artery disease[211]. Over time, these effects have been demonstrated as frequent and widespread enough to necessitate a family-targeted screening programme[212,213]. The efficacy of this intervention is further facilitated by the known mechanisms of action of these genes as well as the accessibility to statins and other cholesterol lowering medications.

To investigate the clinical consequences of more common variation at IEM genes, I leveraged the widespread availability of summary statistics across GWAS studies. Summary statistics have been made publicly available in databases and record billions of variant-phenotype associations across thousands of traits and diseases[62,63,73,74]. Different databases enable the systematic identification and assessment of robust findings as part of this 'phenome-wide' approach.

The unprecedented wealth of available GWAS summary statistics enables the identification of phenotypic associations as well as their linkage to metabolites associated with the same variant. However, large sample sizes of GWASs increase the likelihood of observing a significant variant-phenotype association at any given threshold by chance. This can increase the chance of coincidental overlap of genetic signals between metabolic and phenotypic traits, leading to the identification and interpretation of spurious variant-metabolite-phenotype associations. The importance of assessing the statistical likelihood of true 'sharedness' of a genetic signal is exemplified by findings from a phenotypic assessment of protein quantitative trait loci, where only 10% of the phenotype associations identified had an estimated posterior probability indicating that the queried variant was also the same signal for protein levels[214]. Testing for a shared genetic signal between two traits can be performed using statistical colocalisation methods. Colocalisation methods are usually implemented in a Bayesian framework to yield a posterior probability of whether the traits at a given locus are caused by a shared genetic signal or driven by two distinct signals for each trait within the region[215]. A recent extension of this method is colocalisation across more than two traits ('multi-trait colocalisation'), which can identify multiple trait clusters that are driven by different causal variants in a region[216,217]. Colocalisation methods can also perform multi-trait fine-mapping to identify a candidate causal SNP that explains the largest proportion of posterior probability of colocalisation[215,217].

In this study, I performed phenome-wide assessment followed by multi-trait colocalisation to systematically test for shared genetic signals between IEM-affected metabolic and phenotypic traits across multiple loci. Findings from this research may help to identify genetic subgroups underlying complex disease outcomes and provide preliminary evidence to assess their clinical applicability.

Table 1: (Non-exhaustive) summary of examples where the metabolic and phenotypic consequences of common variants mimic those of rare, IEM-causing variants at the same gene. The associated metabolite is either directly related to disease development or reflects the underlying disease mechanism. DOE: Direction of Effect

IEM gene	IEM (OMIM#)	Affected metabolites (DOE)	IEM symptoms (DOE)	Example of common variant (EA/OA, EAF)	Link to IEM gene (Pubmed ID)	Associated metabolites (DOE) (Pubmed ID)	Associated complex traits/diseases (Pubmed ID)
APOB	Familial dyslipidemias (144010, 615558)	Cholesterol (+) Triglycerides (+)	Hypercholesterolemia (+) Xanthomas (+) Coronary artery disease (+)	rs934197 (A/G, 0.33)	**APOB* is the closest gene to rs934197 and encodes apolipoprotein B which is involved in LDL cholesterol metabolism[218].	Cholesterol (+) (Metabolon mGWAS)	Coronary artery disease (+)[219]
HAL	Histidinemia (235800)	Histidine (+) *Trans*-urocanate (-)	No clinical presentation aside from elevated histidine levels, though *trans*-urocanate is thought to be involved in skin response to UV light[220]	rs3213737 (G/A, 0.48)	rs3213737 was linked to the *HAL* gene based on annotation by the software SNiPA[37,179].	*Trans*-urocanate (+)[37]	Skin cancer (-)[37]
DBH	Orthostatic hypotension (223360)	Dopamine (+) Norepinephrine (-)	DBH protein activity (-) Sympathetic noradrenergic function (-) Orthostatic hypotension (+)	rs6271 (T/C, 0.074)	*The *DBH* gene encodes the enzyme dopamine beta hydroxylase, which converts dopamine to norepinephrine. Norepinephrine is the precursor for vanillylmandelate.	Vanillylmandelate (-) (Metabolon mGWAS)	Blood pressure (-)[221]

*Annotation is based on likely causal gene annotation using the biochemical literature in the Metabolon mGWAS **(Chapter 3, Section 3.4.3.8.)**.

5.3. Aim and Objectives

The aim of this study was to systematically characterise the phenotypic consequences of IFVs. This was achieved through the following objectives:

1. Perform a phenome-wide analysis of IFVs using summary statistics from UK Biobank, GWAS consortia and publicly available databases; and
2. Test whether associations of IEM-related metabolites and phenotypes at a locus are driven by the same genetic signal and whether the IFV is the likely causal variant.

5.4. Methods

5.4.1. Studies and Measurements

5.4.1.1. Metabolites

Summary statistics for 182 metabolites were obtained from the Metabolon mGWAS (n=14,296) that was performed in the EPIC-Norfolk and INTERVAL cohorts. A description of these prospective cohorts, as well as protocols for metabolite measurement and processing and GWAS analysis, have been described previously in **Chapter 3, Sections 3.4.1.-3.4.3**.

5.4.1.2. Phenotypes

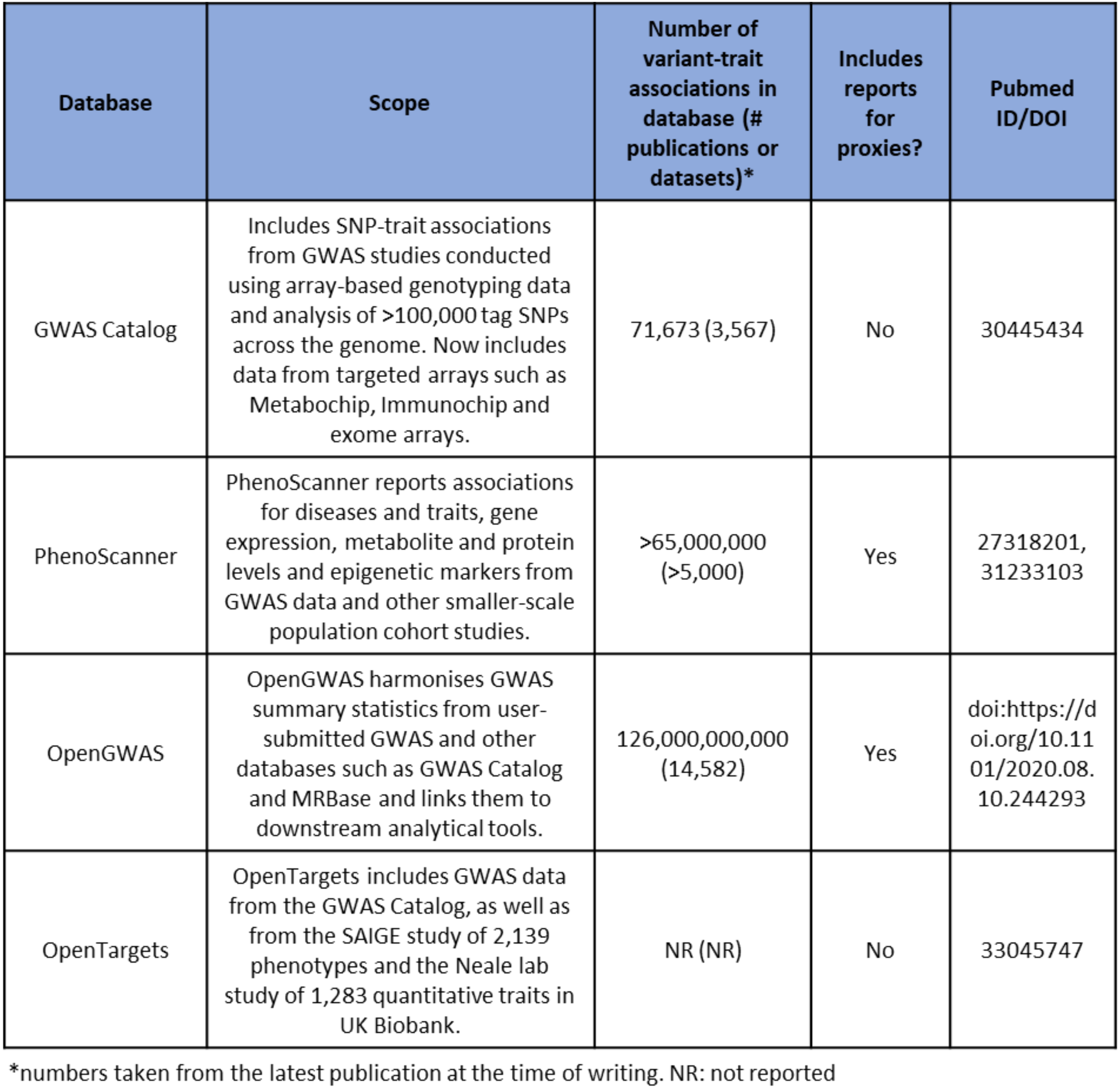
GWAS summary statistics taken from the UK Biobank and from other GWAS consortia were used in this study. The UK Biobank is a prospective population cohort of 500,000 participants with measured genetic and phenotypic data, as described previously[67]. Briefly, participants underwent interviews and filled in questionnaires pertaining to socio-demographic and lifestyle factors at baseline and undertook a physical check to take anthropometric measurements such as blood pressure, hand grip strength and cardiorespiratory fitness. Additional phenotypic measurements were taken at follow up, either during in-person health checks or by web-based questionnaires. Electronic medical records from the death and cancer registries, hospital inpatient data and primary care data were also linked to participants' records.

In this study, 29 other phenotypes measured in different cohorts were used. These measurements are listed in the **Appendix Ch5_ST1**.

5.4.1.3. Access to GWAS summary statistics

GWAS summary statistics for 128 phenotypes (82.3% continuous traits, 17.2% clinical outcomes) were taken from the latest release of UK Biobank from the UK Biobank website (http://www.nealelab.is/ukbiobank) and from the MRC IEU OpenGWAS database[63], which harmonises GWAS summary statistics from UK Biobank and other GWAS consortia for rapid access in a high performance computing environment. In this study, the databases GWAS Catalog, PhenoScanner, OpenGWAS and OpenTargets were queried. A brief overview of these databases is provided in **Table 2**.

Table 2: Overview of databases recording variant-phenotype associations reported in GWAS or population-based studies.



Database	Scope	Number of variant-trait associations in database (# publications or datasets)*	Includes reports for proxies?	Pubmed ID/DOI
GWAS Catalog	Includes SNP-trait associations from GWAS studies conducted using array-based genotyping data and analysis of >100,000 tag SNPs across the genome. Now includes data from targeted arrays such as Metabochip, Immunochip and exome arrays.	71,673 (3,567)	No	30445434
PhenoScanner	PhenoScanner reports associations for diseases and traits, gene expression, metabolite and protein levels and epigenetic markers from GWAS data and other smaller-scale population cohort studies.	>65,000,000 (>5,000)	Yes	27318201, 31233103
OpenGWAS	OpenGWAS harmonises GWAS summary statistics from user-submitted GWAS and other databases such as GWAS Catalog and MRBase and links them to downstream analytical tools.	126,000,000,000 (14,582)	Yes	doi:https://doi.org/10.1101/2020.08.10.244293
OpenTargets	OpenTargets includes GWAS data from the GWAS Catalog, as well as from the SAIGE study of 2,139 phenotypes and the Neale lab study of 1,283 quantitative traits in UK Biobank.	NR (NR)	No	33045747

*numbers taken from the latest publication at the time of writing. NR: not reported

UK Biobank is powered to assess continuous traits at scale. However, participants in the UK Biobank have a lower prevalence and incidence of disease compared to what is observed in the general population[222], making this resource potentially underpowered to detect associations for disease outcomes or traits that are not easily measured in a large population cohort. Therefore, GWAS summary statistics for 29 phenotypes (79% continuous traits, 21% clinical outcomes) were also obtained from GWAS case-control studies using consortia websites and from the databases GWAS

Catalog and OpenGWAS. These phenotypes are listed in the **Appendix Ch5_ST1**, with further details of their measurement in comprising cohorts available in the references. The cohorts in which these phenotypes were measured were predominantly of European ancestry.

5.4.2. Statistical Analysis

5.4.2.1. Inclusions

For this study, 187 IFVs not previously reported as pathogenic for the corresponding IEM in ClinVar **(Chapter 3)** were assessed as well as 24 IFVs identified in other GWASs that were identified using the same protocol described previously **(Chapter 3, Section 3.4.3.8.)**. The 24 additional IFVs are summarised in the **Appendix Ch5_ST2**.

5.4.2.2. Phenome-wide assessment

The databases GWAS Catalog, PhenoScanner, OpenGWAS and OpenTargets were queried for phenotypic associations (significant $p \leq 1x10^{-5}$) of the 187 IFVs and their proxies ($R^2 \geq 0.8$). Simultaneous query across these databases enabled comprehensive detection of associations that may only be present in one of these databases. Associated phenotypes recorded in the GWAS databases range from intermediate molecular phenotypes, such as 'Eosinophil percentage of granulocytes'[223], to self-reported (e.g. 'Self-reported: Asthma') and clinically diagnosed (e.g. 'Diagnoses - main ICD-10: I20 Angina pectoris') outcomes in UK Biobank[67]. All associations were aligned to the metabolite-raising effect allele of the corresponding IFV and the association with the strongest p-value was reported.

To summarise the associations obtained across databases and assess their potential clinical relevance, associations with clinical outcomes or diseases that could be mapped to ICD-10 codes were reported separately from other phenotypes. Phenotypes that differed by small differences in description (e.g. 'Atherosclerosis' and 'Coronary atherosclerosis'), study design or data collection (e.g. 'Self-reported: Asthma' and 'Diagnosed by doctor: Asthma') were mapped to common terms to reduce redundancy (e.g. 'Atherosclerosis' and 'Asthma', respectively). Phenotypes were then assigned to categories that were developed on an *ad hoc* basis. This resulted in 15 categories: 'Anthropometry', 'Medication', 'Lifestyle', 'Bone', 'Hematological', 'Respiratory', 'Renal', 'Eye', 'Reproductive and urinary', 'Gastrointestinal (GIT)', 'Inflammatory', 'Endocrine and Metabolism', 'Cancer', 'Cardiovascular', 'Neurological, cognitive or behavioural'. An additional 'Miscellaneous' category was used for 31 phenotypes that did not fall into the above categories, making a total of 16 categories.

5.4.2.3. Prioritisation of IEM-related phenotypes

To assess whether IEM-related metabolic effects could lead to IEM-related health effects at the same genetic locus, associated phenotypes identified in phenome-wide analysis were pruned to include only IEM-related phenotypes. Phenotypes were assessed based on their phenotypic similarity with symptoms observed in patients with the IEM corresponding to the associated IFV, as reported in IEMBase[162] and other relevant literature. Evidence supporting the prioritisation of phenotypes for variants at a given IEM gene was given for all prioritised phenotypes. Evidence statements for phenotypes that were followed up in colocalisation analysis are provided in **Appendix Ch5_ST3**.

Some late-onset clinical outcomes such as cancer are unlikely to share symptoms with IEMs due to their complex clinical presentations, though this study provided an opportunity to investigate their underlying metabolic mechanisms. In the absence of similar clinical presentation, these outcomes were also prioritised if they had been associated with at least one corresponding IEM-affected metabolite. This step was only applied in exceptional cases to minimise the risk of detecting spurious associations and only ten associations prioritised using this method were tested for colocalisation **(Appendix Ch5_ST3)**.

5.4.2.4. Multi-trait colocalisation

In the UK Biobank, GWAS summary statistics with fewer than 100 cases were excluded. Amongst the 210 prioritised phenotypes, 62 for which summary statistics had i) fewer than 1,000 SNPs in the region of interest; ii) incomplete information on chromosome, position, effect allele, other allele, effect sizes, standard errors or p-values; iii) position coordinates not reported in The Genome Reference Consortium Human Build 37, or iv) primarily non-European ancestry, were excluded. A further six were excluded due to poor (e.g. 'NET100 0056', with no additional information in the corresponding publication) or non-specific (e.g. 'Heart function tests') phenotype descriptions. Phenotypes relating to measurements of levels of gene expression, metabolites or proteins were also excluded, leaving 142 distal complex traits and disease outcomes with available and complete GWAS summary statistics for analysis. Each prioritised phenotype was represented by the largest GWAS study (number of cases for binary traits or total number of participants otherwise) with a significant association in the locus.

Multi-trait colocalisation analysis (referred to hereafter as 'colocalisation'), was performed to identify metabolic and phenotypic traits that share a common causal variant and to identify clusters of traits driven by the lead variant at the locus. I used the method HyPrColoc[217], which iterates over scenarios of colocalisation using a Bayesian approach, using the R package 'hyprcoloc' (v1.0). Loci defined in the

Metabolon mGWAS study **(Chapter 3, Section 3.4.3.4.)** were used for IFVs; otherwise, 500Kbp regions surrounding the lead variant were used.

HyPrColoc was run using the following settings: 1.) prior probability that a variant is associated with a single trait (prior1) = 1x10^{-4}, and 2.) the prior probability that a variant is associated with one trait, given it is already associated with another (prior2) = 0.98. A regional threshold of 0.5 was used, denoting the prior probability that all traits share an association with one or more variants in the region. To estimate the prior probability that all traits align and share a single causal variant, regional and alignment prior probabilities of 0.5 were used. These settings, while less stringent than those recommended[217], are compensated for by the high prior probability based on IEM knowledge. Cluster stability was assessed by using more stringent prior2 values (0.99, 0.999) and regional and alignment threshold values (0.6, 0.7, 0.8, 0.9). Only variants present in all included traits were considered for a given locus and any variants with a standard error of zero were removed. To provide HyPrColoc with sufficient information to assess colocalisation while allowing for the inclusion of multiple traits for testing, colocalisation was performed if there were more than 1,000 variants present across all included traits at the locus. For loci with fewer than 1,000 variants in a region after including all prioritised phenotypes, colocalisation was performed using only clinical outcomes and IEM-affected metabolites. Exceptions to this were as follows:

- At the *APOE* locus, the clinical outcome 'Myocardial infarction' was also excluded to obtain ≥1,000 variants in the region.
- At the *DDC* locus, only associations for body fat measurements were present; therefore, the representative measures of 'Comparative height at age 10', 'Height', 'Weight' and 'Body mass index' were tested.

If multiple signals were present on visual inspection of regional plots, colocalisation was repeated by focusing on two traits (one metabolite, one phenotype that was preferably a disease outcome) in the cluster and applying a 'masking' approach. Briefly, this approach identifies lead SNPs in the region, and for each lead SNP in turn, calculates the residual association of variants in LD while ignoring variants that are in LD with other lead SNPs[224]. Masking colocalisation was implemented using the R package 'coloc'[215] (v4.0-4), which calculates a posterior probability of colocalisation (PP_{coloc}), at which a threshold of 0.70 was considered strong evidence of colocalisation. At some loci (e.g. *OPLAH* – **Chapter 5, Results 5.5.3.2.**), colocalisation was first assessed using HyPrColoc and then on a narrowed region excluding other lead SNPs using coloc.

For practical implementation in colocalisation analysis, measurements that were continuous were specified as 'continuous' while phenotypes that were binary, categorical or ordinal were specified as 'binary'.

Colocalising clusters were retained if they i.) contained both metabolite and phenotypic traits, where ii.) there was strong evidence of a signal in the region across traits (regional posterior probability ($PP_{regional}$) ≥0.7)) and lead signals in the region were aligned (alignment posterior probability ($PP_{alignment}$) ≥0.7)). Novel, colocalising clusters were annotated based on novelty of the variant-to-phenotype association, metabolite-to-phenotype association, or on novel synthesis of the link between the IEM gene with the variant-metabolite-phenotype association.

5.4.3. Annotation of Gene-drug and Variant-drug Interactions

To determine the therapeutic potential of variants at IEM genes, IEM genes assigned to IFVs were annotated for druggability. At the variant level, variant-gene-drug annotations were downloaded from the database PharmGKB[225] in March 2020. Downloaded annotations contained 6,971 variant-drug pairs reported in the literature (53% significant), as well as a ranked assessment of the evidence supporting clinical relevance. Evidence for clinical relevance was categorised as tier 1 (high: based on guidelines from the Clinical Pharmacogenetics Implementation Consortium or known clinical implementation), tier 2 (moderate: variant is in a (pharmaco)gene known to be associated with drug toxicity, metabolism, pharmacokinetics or efficacy), tier 3 (low: based on a single significant and unvalidated study, or annotation for a variant-drug combination evaluated in multiple studies but lacking clear evidence of association) and tier 4 (preliminary: based on a single case report, non-significant study or *in vitro*, molecular or functional assay evidence only).

Druggable genes, as defined by Finan *et al.* (2017)[226], were also annotated in this study. Briefly, druggable genes were identified by their association with licensed first-in-class drugs from 2005, drug targets in late phase clinical development, pre-clinical phase small molecules with protein-binding measurements and genes that encode potential protein drug targets[226]. Based on these criteria, a total of 4,479 (22%) of roughly 20,300 protein-coding genes were estimated to be druggable[226].

All statistical analyses and graphics were performed and produced using R version 3.5.3.[128] and STATA version 14.2[129].

5.5. Results

5.5.1. IFVs Are Associated With a Broad Range of Complex Phenotypes in the General Population

Phenome-wide assessment identified 1,553 associations between 108 loci harbouring 132 IFVs and 393 phenotypes ($p \leq 1x10^{-5}$) **(Figure 2A)**. Phenotypes spanned across 16 categories including 'Lifestyle' (16%), 'Neurological, cognitive or behavioural' (14%) and 'Hematological' (13%). When considering only clinical outcomes (i.e. phenotypes that could be mapped to a disease code from the International Classification of Diseases, 10th Revision (ICD-10)), 274 associations between 48 loci harbouring 75 IFVs and 87 clinical outcomes were found **(Figure 2B)**. Associated clinical outcomes predominantly belonged to 'Cardiovascular' (23%), 'Cancer' (17%) and 'Inflammatory' (15%) categories.

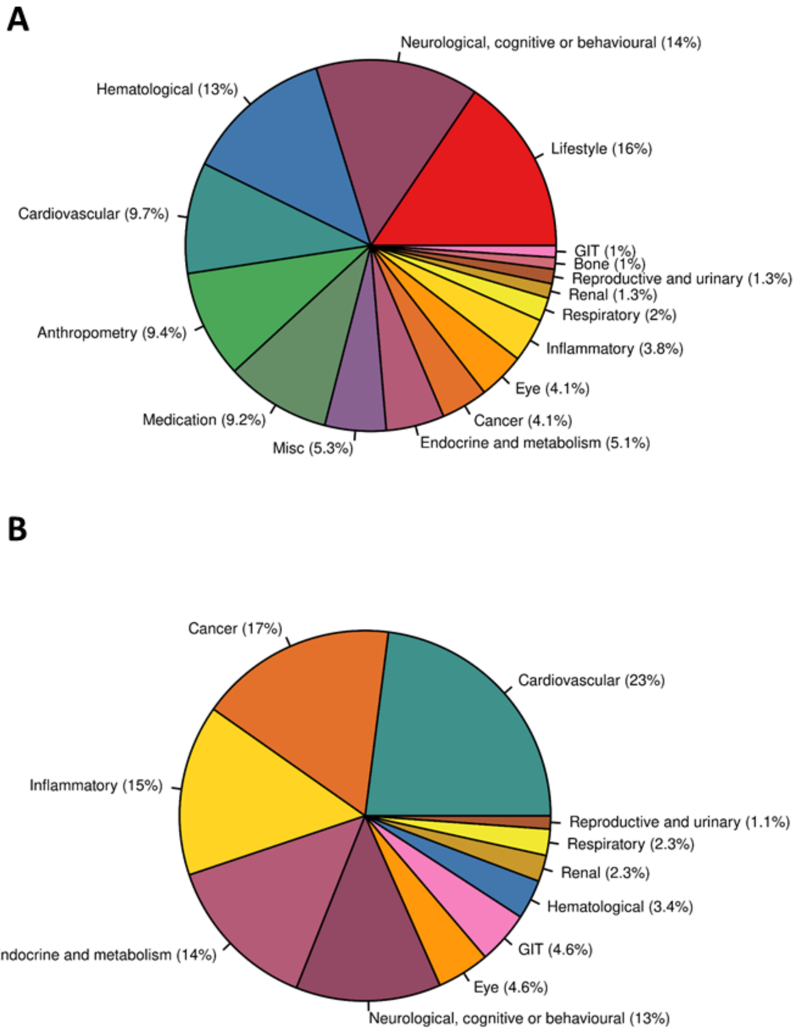


Figure 2: Summary of distinct, unique phenotypes associated with IFVs ($p \leq 1x10^{-5}$) across GWAS databases. A.) All distinct traits and outcomes (n=393); **B.)** Subset of distinct clinical outcomes (n=87).

Of the 1,553 associations detected, 703 were identified as being reflective of symptoms observed in the corresponding implicated IEM **(Figure 2)**. These prioritised associations represented 212 phenotypes (28% clinical outcomes) at 54 IEM gene-linked loci **(Figure 3; Appendix Ch5_Fig1)**.

Of the 703 associations at 54 IEM gene-linked loci that were prioritised for colocalisation, 137 could not be tested either due to missing or unavailable GWAS summary statistics or to low SNP coverage in the region (number of variants < 1,000). This left 566 associations at 45 loci (containing 62 IFVs) for colocalisation analysis.

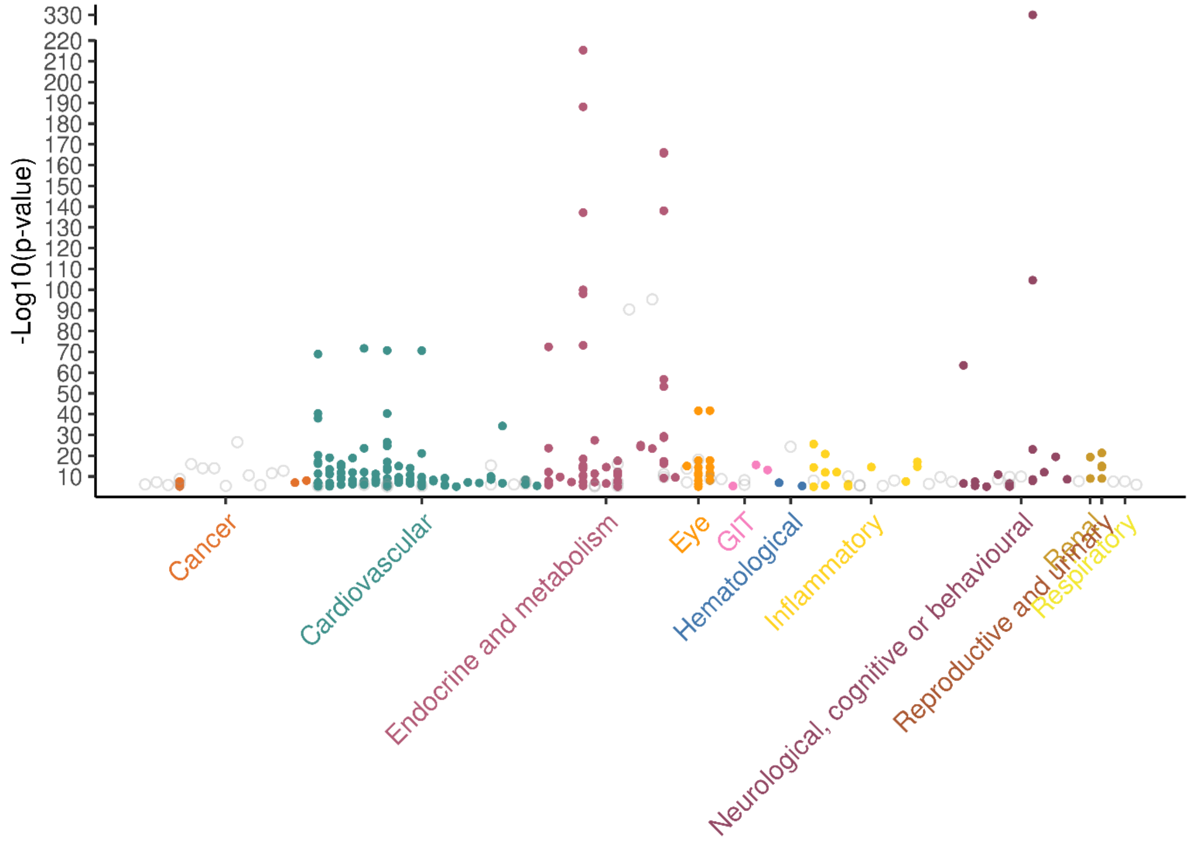


Figure 3: Summary of IFV associations ($p \leq 1x10^{-5}$) with 87 clinical outcomes. Filled in, coloured circles represent prioritised phenotypes while empty, grey circles represent associated but non-prioritised phenotypes. Strongest p-value associations for each distinct phenotype are shown.

5.5.2. Colocalisation Analysis Validated Known Metabolic Contributions of Known IFVs to Complex Traits and Diseases

At 24 of the 45 tested loci, colocalising clusters containing metabolites and phenotypes ($PP_{regional} \geq 0.7$, $PP_{alignment} \geq 0.7$) were detected. All 25 clusters detected at these loci were expected based on IEM knowledge and included 11 clusters at 11 loci that were novel. Amongst the 24 loci were seven that contained nine of 51 IFVs previously identified as having large and/or non-additive effects on metabolite levels **(Chapter 4, Section 4.5.4)**. Of the nine IFVs at the seven loci, four (the *OPLAH* variant rs3935209, the *CPS1* variant rs1047891, the *APOE*/*APOC2* variant rs7412 and the *PCSK9* variant rs11591147) were identified as the candidate causal variant driving signals at the locus, either directly or by a proxy variant (R^2>0.8).

Novel examples of interest at the *DBH*, *TH*, *OPLAH* and *ARG1* loci are covered in subsequent sub-sections, and other novel examples are briefly summarised in **Appendix Ch5_ST4**.

5.5.2.1. Variation across two IEM gene-linked loci implicated dopamine metabolism in blood pressure regulation

The *DBH* gene converts dopamine to norepinephrine and regulates sympathetic noradrenergic function[227]. Rare mutations at the *DBH* gene are known to cause blood pressure dysregulation, resulting in dizziness upon standing ('orthostatic hypotension') (OMIM #223360)[227]. In this study, I found that the T-allele of the *DBH* variant rs6271 (EAF=0.074) was the shared signal between decreased levels of a downstream catecholamine, vanillylmandelate (beta±SE: -0.16±0.023, p=8x10^{-13}) and decreased self-report of hypertension in 462,933 participants in UK Biobank[63] (beta±SE: -0.013±0.0017, p=1x10^{-14}), amongst other blood pressure and blood pressure-related phenotypes **(Figure 4, Appendix Ch5_ST3-ST4)**. Strong evidence of colocalisation was detected for these phenotypes ($PP_{regional}$=0.97, $PP_{alignment}$=0.97), with the variant rs6271 predicted to explain >99% of the colocalisation probability.

The role of dopamine and catecholamine metabolism in autonomic function was supported by another identified example at the tyrosine hydroxylase (*TH*) locus. The *TH* gene encodes an enzyme that lies upstream of the *DBH* gene and catalyses the rate-limiting step of converting tyrosine to levodopa (DOPA), which in turn is converted by other enzymes into other catecholamines[228]. Rare mutations in the gene can lead to the IEM Segawa syndrome (OMIM #605407); in severe cases, autonomic dysfunction (symptoms of which include orthostatic hypotension, dysregulation of pulse rate and sweating abnormalities) may occur[229]. Hypotonia, which decreases muscle tone, is also observed in less severe cases of Segawa syndrome[229].

In this study, the A-allele of the *TH* variant rs10840516 (EAF=0.24) was significantly associated with increases in levels of the catecholamines 3-methoxytyrosine (beta±SE: 0.18±0.014, p=4.8x10^{-36}) and dopamine sulfate (2) (beta±SE: 0.094±0.014, p=2.4x10^{-11}) and with increased pulse rate measured in 436,424 participants in UK Biobank[63] (beta±SE: 0.012±0.025, p=1.1x10^{-6}), reflecting autonomic dysfunction in the more severe form of Segawa syndrome **(Figure 4A)**. Colocalisation analysis showed that levels of 3-methoxytyrosine and dopamine sulfate (2) shared a genetic signal with pulse rate in the region ($PP_{regional}$>0.99, $PP_{alignment}$=0.79). A variant in high R^2 with rs10840516, rs11564705 (MAF=0.24, R^2=0.98), was predicted to explain 36% of the colocalisation probability. The variant rs10840516 was also significantly associated with measures of muscle mass in UK Biobank **(Appendix Ch5_ST3)**, reflecting symptoms of hypotonia observed in Segawa syndrome. However, these phenotypes did not colocalise with catecholamine levels and were driven by a variant independently of rs10840516 (rs35506085, MAF=0.19, R^2=-0.10) **(Appendix Ch5_ST4)**.

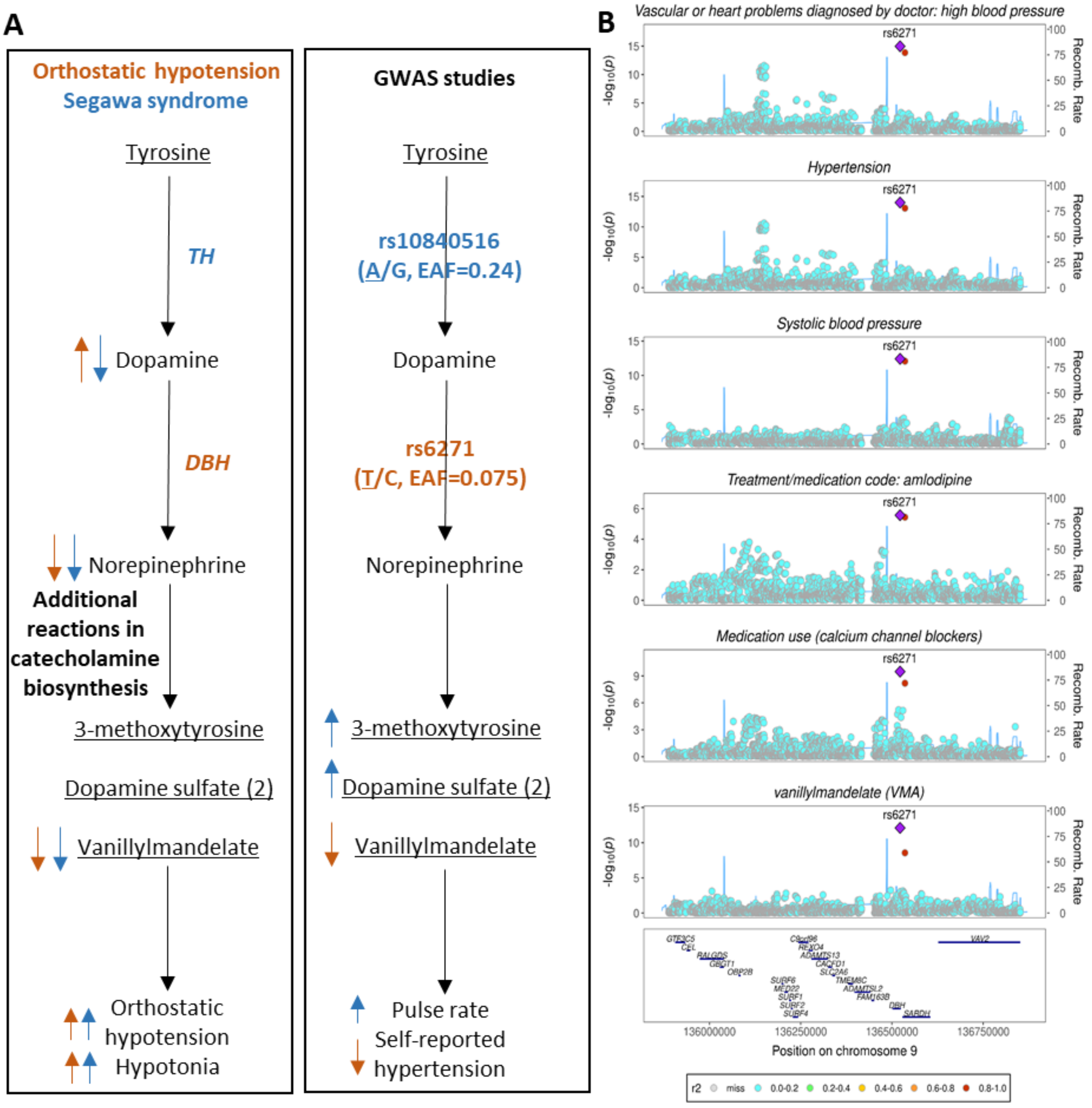


Figure 4: Comparison of metabolic and phenotypic consequences caused by rare mutations and more common variants at the *DBH* and *TH* loci. A) Summary of the metabolic and phenotypic consequences of the *DBH* variant rs6271 (EAF=0.074) and the *TH* variant rs10840516 (EAF=0.24). Arrows representing observed directions of effect are colour-coded based on metabolic and phenotypic consequences specifically related to orthostatic hypotension (OMIM #233360) or Segawa syndrome (OMIM #605407). Metabolites that are underlined were measured in the Metabolon mGWAS, while those not underlined were not detected. **B)** Stacked regional plots demonstrating

alignment of select metabolites and blood pressure and blood pressure-related phenotypes at the *DBH* locus.

5.5.2.2. The OPLAH locus suggests that homeostasis of 5-oxoproline levels is linked to cognitive performance

The 5-oxoprolinase (*OPLAH*) gene encodes an enzyme that converts 5-oxoproline (also known as pyroglutamate) to L-glutamate[230]. Mutations at the *OPLAH* gene can cause 5-oxoprolinase deficiency (OMIM #260005), which is characterised by toxic accumulation of 5-oxoproline levels. Elevated 5-oxoproline levels are thought to induce oxidative stress in the cerebral cortex and inhibit synaptic cleft transmission, leading to psychomotor and mental retardation as well as vomiting and nausea[231,232].

In the Metabolon mGWAS, the G-allele of the *OPLAH* variant rs3935209 (EAF=0.082) was associated with decreased levels of 5-oxoproline (beta±SE: -0.36±0.022, $p=6.6\times10^{-63}$), increased levels of the downstream metabolite 6-oxopiperidine-2-carboxylic acid (beta±SE: 0.50±0.022, $p=8.1\times10^{-116}$) and with decreased cognitive performance (beta±SE: -0.024±0.0053, $p=5.4\times10^{-6}$)[233] and performance in intelligence tests (beta±SE: -0.021±0.0047, $p=1.0\times10^{-5}$)[234]. Cognitive performance and performance in intelligence tests could reflect symptoms of mental retardation[234] in 5-oxoprolinase deficiency **(Figure 5A)**. The directions of effect observed in this study contrasted with those expected from IEM knowledge, suggesting that homeostasis, rather than direct levels of 5-oxoproline, is important. Specifically, toxic accumulation of 5-oxoproline levels observed in 5-oxoprolinase deficiency and other IEMs (e.g. glutathione synthetase deficiency (OMIM #266130)) has been shown to cause oxidative stress and impair neurological function[235,236]. Conversely, low levels of 5-oxoproline have been observed in individuals with age-related cognitive decline and impairment[237,238], indicating that 5-oxoproline is involved in maintaining neurological function.

Despite the regional plot strongly indicating a shared causal signal across all traits **(Figure 5B)**, HyPrColoc identified two colocalising clusters due to a second lead signal in the region for cognitive performance and intelligence at rs2721173 ($R^2=0.033$ with rs3935209). To account for multiple signals in the region, I focused on the traits 'Cognitive performance' and '5-oxoproline' and used a 'masking' approach[224] that assumes all lead SNPs are in linkage equilibrium **(Chapter 5, Methods 5.4.3.)**. This approach did not identify evidence for colocalisation ($PP_{coloc}=0.039$). However, narrowing the region to exclude rs2721173, as well as the region depleted of SNPs **(Figure 5B)**, identified strong evidence of rs3935209 as the candidate causal SNP in HyPrColoc ($PP_{regional}=0.95$, $PP_{alignment}=0.90$, $PP_{explained}>0.99$) and in the masking approach ($PP_{coloc}=0.87$). By contrast, narrowing the region to include only rs2721173 led HyPrColoc to detect two independent signals, one driving metabolite levels and another driving phenotypes.

While cognitive performance is a broad phenotype that is affected by genetic as well as environmental factors, the results of this study are supported by a Mendelian randomisation study that demonstrates a positive, causal effect of 5-oxoproline levels on performance in intelligence tests[239]. These results would benefit from further validation, as plasma levels of 5-oxoproline are also reflective of dietary and drug intake[240,241] and may not reflect 5-oxoproline levels in cerebrospinal fluid that are elevated in OPLAH deficiency.

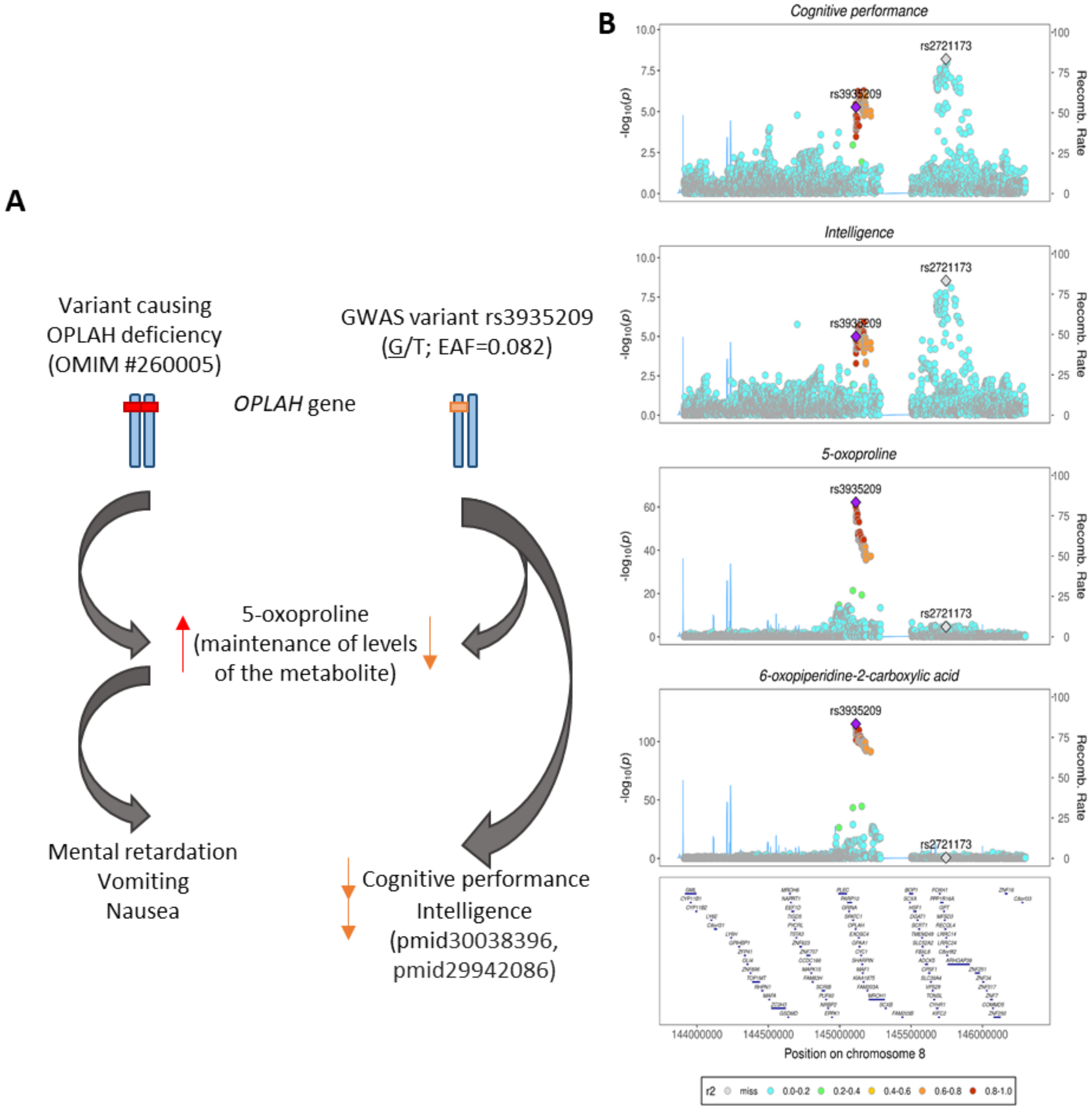


Figure 5: Comparison of metabolic and phenotypic consequences caused by rare mutations and more common variants at the *OPLAH* locus. A) Summary of the metabolic and phenotypic consequences of the *OPLAH* variant rs3935209 (EAF=0.082). Levels of 5-oxoproline are thought to directly impact both IEM and complex trait phenotypes through reduced synaptic cleft transmission and impaired synthesis of essential lipids in the brain. The variant rs3935209 was also associated with increased levels of 6-oxopiperidine-2-carboxylic acid, which corroborates the direction of association

as a downstream metabolite of 5-oxoproline. **B)** Stacked regional plots demonstrating alignment of metabolites and phenotypes at the *OPLAH* locus.

5.5.2.3. A novel polymorphism at the ARG1 locus links arginine levels with type 2 diabetes risk

The arginase (*ARG1*) gene encodes an enzyme that catalyses the conversion of arginine to ornithine in the final step of the urea cycle[242]. *ARG1* mutations can cause argininemia (OMIM #207800), which is characterised by high arginine levels that result in feeding difficulties, seizures, spastic paraplegia and severe mental retardation[243,244]. In the Metabolon mGWAS, an *ARG1* variant, rs71753454, was detected at which the insertion allele (EAF=0.22) was significantly associated with an increase in arginine levels (beta±SE: 0085±0.014, $p=2.97x10^{-9}$) and with 12 body fat composition measures (one of which is shown in **Figure 6B**), which could reflect changes in body fat composition resulting from feeding difficulties in ARG1 deficiency **(Appendix Ch5_ST3)**. The *ARG1* variant rs71753454 was also associated with an increase in type 2 diabetes risk (aligned association for proxy variant rs3756784 (R^2=0.82), beta±SE: 0.051±0.01, $p=2.6x10^{-8}$)[245], which could be a long-term consequence of altered body fat composition.

At the *ARG1* locus, strong colocalisation between arginine levels, type 2 diabetes and other body fat composition phenotypes was observed ($PP_{regional}$>0.99, $PP_{alignment}$=0.95) **(Figure 6B; Appendix Ch5_ST4)**. Sensitivity analysis demonstrated that arginine and type 2 diabetes colocalised across all tested configurations in a highly stable cluster. At this locus, the variant rs2781668 (MAF=0.17), which is in moderate LD with rs71753454 (R^2=0.6), was highlighted as the candidate causal variant ($PP_{explained}$=0.94). The direction of association observed at this locus contrasts with that reported in observational studies, which suggest that arginine supplementation can reduce the risk of type 2 diabetes by improving glucose clearance and reducing oxidative stress[246–249]. This discrepancy suggests that a) arginine levels are driven by genetic and environmental factors, and b) the effects of rs71753454 on arginine levels and type 2 diabetes may be locus-specific.

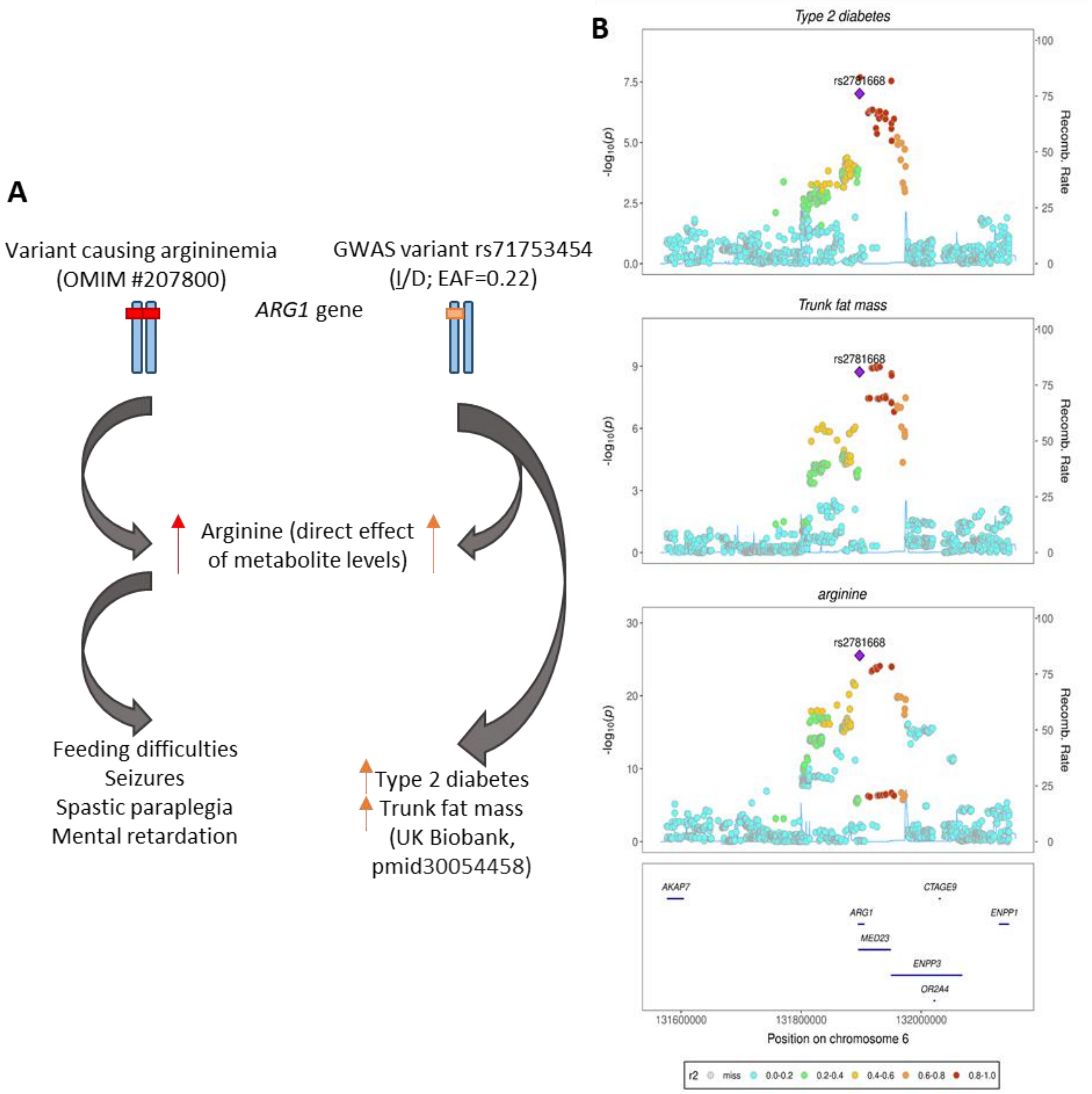


Figure 6: Comparison of metabolic and phenotypic consequences caused by rare mutations and more common variants at the *ARG1* locus. A) Summary of the metabolic and phenotypic associations of the *ARG1* variant rs71753454 (EAF=0.22). **B)** Stacked regional plots demonstrating alignment of arginine, type 2 diabetes and trunk fat mass (representative of other body fat composition phenotypes) at the *ARG1* locus. The candidate causal variant rs2781668 (MAF=0.17), which is in moderate LD with the IEM metabolite-associated variant rs71753454 (R^2=0.6), is labelled. Variant rs71753454 is not present in the dataset.

5.5.3. Variant-drug and Gene-drug Annotations of IFVs and Their Mapped IEM Genes

In PharmGKB, eleven IFVs at the *CTH*, *DPYD*, *UGT1A1*, *SLCO1B1/SLCO1B3*, *APOC1*, *APOE* and *MTHFR* genes were associated with drug responses in PharmGKB. Drug responses primarily represented toxicity responses to warfarin, statins and agents such as mycophenolate mofetil **(Appendix Ch5_ST5)**. Of the 11 variants with detailed drug annotations, only three (*SLCO1B1/SLCO1B3* variant rs4149056 with simvastatin and *DPYD* variants rs3918290 and rs67376798 with toxicity of capecitabine, fluorouracil and pyrimidine analogues, which are used to treat neoplasms) were annotated in medical society-endorsed guidelines (PharmGKB tier 1) and one (the variant rs7412) which was located at a known, important, pharmacogene *APOE2* (PharmGKB tier 2) **(Appendix Ch5_ST5)**. The rest were associated in lab-based assays or unreplicated studies (PharmGKB tiers 3 and 4) **(Appendix Ch5_ST5)**. A total of 56 of 103 IEM genes (containing 111 of 211 assessed IFVs) were classified as druggable according to Finan *et al.* (2017)[226] **(Table 3)**.

Table 3: Druggable IEM genes assigned to loci with metabolite-phenotype clusters specific for the relevant IEM. Categories as defined according to Finan *et al.* (2017)[226].

Nature of drug-gene association	Number of IEM genes
Tier 1: Gene encodes an efficacy target of approved small molecules, biotherapeutic drugs and clinical phase drug candidates	28
Tier 2: Gene encodes targets with known bioactive drug-like small molecule binding partners, including those with ≥50% identity (over ≥75% of the sequence)	14
Tier 3: Gene encodes members of key druggable gene families not already included in Tiers 1 or 2 or demonstrate some similarity with approved drug targets	14

5.6. Discussion

5.6.1. Summary of findings

A phenome-wide approach was used to systematically characterise the phenotypic associations of IFVs. Careful and systematic comparison of the phenotypes with each IFV showed that 212 of the 393 (54%) complex traits and disease outcomes were related to one or more manifest symptoms of the corresponding IEMs and prioritised. Colocalisation analysis, which was performed on IEM-related metabolite and phenotype associations, detected shared genetic signals for clusters of these traits at 24 of the 45 assessed loci. This systematic, data-driven approach successfully demonstrated the proof-of-principle that common variation at IEM genes contributes to phenotypic consequences in the general population and identified genetic subgroups affecting complex metabolic diseases.

At some loci, genetic influences on metabolite levels were found be large enough to cause inter-individual differences in drug response. Using the PharmGKB database, 11 of the 187 IFVs at seven IEM genes were reportedly associated with differential efficacy or toxicity responses to drugs that treat several conditions. A total of 56 of the 103 IEM genes harbouring IFVs were encoded enzymes or solute transporters that could be targetable by drugs, highlighting metabolic pathways that could potentially be assessed in drug safety evaluations.

5.6.2. Novelty of findings

Previous studies have identified select examples of variants at IEM genes with IEM-related metabolic and phenotypic consequences[19,21,37,46,250], and one study[21] has used a similar approach to that described in this Chapter. However, none of these prior studies have achieved the same extent of systematic characterisation across variants at IEM genes as this study has. One key characteristic that determined the success of the approach highlighted in this study was the integration of GWAS summary statistics from case-control studies. Many of the GWAS summary statistics are derived from the population-based UK Biobank cohort. Although this cohort has a large sample size, it is over-representative of healthy volunteers[222], and thus the prevalence and incidence of diseases in the UK Biobank are lower than those observed in the general population. The inclusion of summary statistics from better powered case-control studies of disease outcomes therefore maximises the detection of phenotypic associations. The inclusion of results from a mGWAS that used untargeted metabolomic profiling also enabled the detection of additional variant-metabolite associations and increased the number of IFVs and associated metabolic pathways identified for downstream analysis. As a result, this approach successfully identified independent loci, such as the *TH* and *DBH* loci, that linked the same metabolic pathway to the same phenotypic outcomes.

In **Chapter 4**, I identified 51 IFVs with large metabolic effects, yet only four of these IFVs were identified in this chapter as candidate causal signals for metabolic and phenotypic associations within the locus. This could be due to biological reasons such as the mode of inheritance; IEMs are usually inherited in an autosomal recessive manner[251,252], therefore, carriers of common variants at IEM genes may not have observable phenotypic effects. This could also lead to a 'threshold' effect, where metabolite levels have undetectable phenotypes within a physiologically normal range and a 'full' phenotype beyond that range. While some metabolites are thought to have linear effects on traits (e.g. LDL cholesterol with blood pressure[253]), threshold effects have also been observed in other cases (e.g. iron levels on hepatic inflammation[254]). Lack of observable phenotype could also be due to metabolic canalisation, where potentially deleterious metabolic perturbations are buffered by the effects of other genes with similar function or by alternative routes of metabolism[193].

The effects of drugs on metabolic pathways linked to IFVs indicated common pathways of drug metabolism and excretion rather than phenotypic effects relating to those seen in the corresponding IEM. In this study, drug response annotations were identified for 11 of the IFVs, four of which were annotated in medical society-endorsed guidelines or had moderate evidence for association **(Appendix Ch5_ST5)**. In this study, I also used a literature and database-curated set of 'druggable genes'[226] to identify 56 IEM genes encoding proteins that can or have been used as drug targets. This preliminary evidence suggested that the metabolic effects of these IFVs were large enough to affect the efficacy or toxicity of drugs, which was expected since IEM knowledge facilitates the development and study of drugs that target affected metabolic pathways. Additional evidence would be required to validate these associations and test the importance of IFVs in predicting drug response for pharmacogenetic screening.

5.6.3. Study strengths and limitations

A primary strength of this study was the data-driven approach used that consisted of many elements contributing to its success: a) phenome-wide assessment, which used multiple databases to systematically incorporate findings from UK Biobank and other GWAS consortia and thus maximise the number of phenotypic associations detected; b) the use of IEM knowledge to select phenotypes that were likely to share a genetic signal with IEM-related metabolites, increasing the study's power to detect biologically relevant associations, and c) colocalisation analysis, which was used to account for potential secondary signals in a given genetic region. The systematic nature of the approach also enabled the detection of loci that independently implicated the same genetic, metabolic, and phenotypic traits (e.g. the *DBH* and *TH* loci), increasing confidence in the accuracy of the findings and of the study approach.

Another strength is that the current study approach leverages data that is publicly available (excepting GWAS metabolite summary statistics taken from the Metabolon mGWAS, though this data is in preparation for publication and other summary statistics for metabolite levels also exist). Therefore, this approach can be adapted to fit broader research purposes. For example, the prioritisation of phenotypes based on previously demonstrated metabolite-phenotype links instead of on IEM knowledge enabled the detection of unanticipated variant-metabolite-phenotype links.

Another strength of this study is the flexibility of the approach used to phenotypically characterise genetic variants. For example, genes known to cause rare, Mendelian disorders not classified as IEMs have also been shown to be enriched in GWAS loci for complex traits and diseases[70,71]. Although the phenotypic consequences of rare mutations are also well described for these diseases, the metabolic mechanisms underpinning them are less well known. This gap in knowledge could also benefit from the application of the current study's approach. The molecular mechanisms underlying detected variant-phenotype associataions could be investigated by also including GWASs of gene expression and epigenetic data that are reported in many of the GWAS databases used in this study **(Table 2)**.

This study also had some limitations. One was the inability to characterise phenotypic associations for low-frequency and rare IFVs. These variants were detected using advanced genotyping and imputation methods in the Metabolon mGWAS and validated using sequencing data. Yet despite these efforts, phenotypic characterisation of these variants was limited by GWASs of complex traits and diseases, many of which were performed using older genotyping and imputation methods and therefore limited in terms of variant coverage.

In this study, I maximised the robustness of colocalisation results by only including summary statistics from GWASs of large sample size that had high-quality genotyping within the assessed region and by performing sensitivity analyses. Despite these rigorous efforts, the results of coloc and HyPrColoc algorithms may still have been affected by factors such as small cohort sample size, SNP density, LD structure, and the presence of multiple causal variants in the region, especially when these variants explained similar proportions of trait variation[215,217]. Colocalisation using conditional summary statistics may help to identify additional colocalising signals at these loci, though current methods to achieve this are either time-consuming, require full genome-wide summary statistics (which are not accessible from databases like OpenGWAS[63,255]), or are undergoing refinement[224].

Systematic integration of GWAS summary statistics detected 1,553 associations for consideration, yet only half of these were prioritised for colocalisation. Phenotype prioritisation was necessary to reduce the number of traits considered in colocalisation analysis and increase the power to detect biologically relevant variant-metabolite-phenotype associations, but at the cost of missing less obvious yet biologically relevant associations. A previous study has shown that in some cases, genes may cause rare and common diseases that are phenotypically dissimilar[71], which could speak to an incomplete understanding of rare disease biology. In the future, the use of improved genotyping methods in GWASs of complex phenotypes and refinement of colocalisation methods to handle multiple causal signals within a region could enable assessment of these associations as well.

The use of IEM knowledge and systematic assessment of shared genetic signals across loci increased confidence in the results of this study, as evidenced by the corroboration of findings at the *OPLAH* locus[239] and by variant-metabolite-phenotype links implicated by independent loci. Yet despite the high degree of confidence in the findings based on IEM knowledge, colocalising trait clusters only provided evidence of a shared genetic signal, but not of a causal relationship between metabolite levels and complex phenotypes. This was well illustrated by the *DBH* example: although vanillylmandelate levels colocalised with hypertension in the current study, it is more likely that vanillylmandelate is a biomarker of DBH protein activity, the loss of function which causes orthostatic hypotension, rather than the causal metabolite (*Dr. Eric Fauman, personal comms.*). In recent years, Mendelian randomisation (MR) has become a popular method for assessing causality between an exposure and an outcome variable of interest[256,257]. MR uses independent genetic variants that are associated with both the exposure and outcome as instrumental variables to test whether a genetically-predicted increase in the exposure significantly affects the outcome. Recent efforts have made it possible to perform MR using GWAS summary statistics[40,239,255], making it a feasible approach to test for causal relationships at loci where colocalising metabolite-phenotype clusters were detected.

5.6.4. Conclusions

This study presented a rigorous approach that utilised large-scale GWAS summary statistics and IEM knowledge to show that the metabolic effects of common variants at IEM genes can translate into similar health effects as those seen in patients with the corresponding IEM. The approach outlined here can be extended and adapted to assess the metabolic and phenotypic effects of variants at genes known to cause other rare, Mendelian disorders to enable further identification of genetic subgroups of complex disease.

CHAPTER 6: IEM FAMILIAR VARIANTS AND THEIR ASSOCIATION WITH IEM-RELATED DISEASE PROFILES

6.1. Abstract

Background Mendelian disorder-related symptoms combined into a disorder-specific score (‘phenotypic risk score’, or ‘PheRS’) have been used to identify rare variant effects in a hospital patient cohort with whole exome sequencing data. The aim of this chapter was to assess the utility of PheRSs in a healthy volunteer cohort setting and test for association of IFVs with IEM-specific disease profiles.

Methods Genotypic and electronic health record data from 351,987 European, unrelated participants in the UK Biobank cohort were used. PheRSs were constructed for each participant by summing the observed number of disease-related symptoms, each weighted inversely by their frequency in the cohort. Carriers of IFVs were then compared with homozygous non-carriers for association with a high corresponding rare disease PheRS value (defined as a score above the median PheRS value or above the 95th percentile). Associations reaching significance (FDR≤0.05) using both PheRS definitions were prioritised for downstream assessment. Genetically-predicted levels of 50 IEM-related metabolites linked to PheRS-associated IFVs were also tested for association with the corresponding high PheRS using logistic regression. The contribution of the IFV to the corresponding metabolite-PheRS association was also assessed by the IFV’s inclusion or exclusion in the genetic risk score. Models were adjusted for age, sex and the first ten principal components of genetic ancestry.

Results Nine variants at seven IEM genes were associated with high PheRS for nine rare diseases (FDR≤0.05), replicating known associations for variants at genes known to cause familial forms of dyslipidemias, Alzheimer’s disease and hyperbilirubinemias. Genetically-predicted levels for IEM-related metabolites were significantly associated with high PheRS in 125 of 167 tested cases (Bonferroni-corrected $p\leq3x10^{-4}$). The *UGT1A1* variant rs1976391 was solely responsible for the association between genetically-predicted levels of bilirubin and bilirubin derivatives with high PheRSs for ‘Crigler-Najjar syndrome, type I’ and ‘Gilbert syndrome’ that are characterised by hyperbilirubinemia and jaundice.

Conclusion This study shows that PheRSs can be used to replicate known variant-PheRS associations and highlight specific variant effects on IEM-related disease profiles. PheRS application in population cohorts with higher incidences of rare disease may enable the additional identification of novel variant associations.

6.2. Background

Integration of untargeted metabolomic profiling and GWAS summary statistics with IEM knowledge in **Chapter 5** showed that in some cases, variants at IEM genes may exert metabolic and phenotypic consequences mimicking those caused by rare, IEM-causing variants. Yet despite highlighting novel insights into the potential phenotypic consequences of variants at IEM genes, this framework cannot be used to test whether variation at IEM genes may lead to the increased incidence of phenotypes that together represent the clinical presentation of an IEM.

In a previous study[77], researchers developed a method that combines the observed incidence of multiple phenotypes relating to a rare disease into a summary score. These rare disease-specific 'phenotype risk scores' (PheRSs) were then applied in the BioVU[258] (hospital-based) cohort to detect the effects of rare variants at genes known to cause those rare diseases.

PheRS methodology presents a novel approach by which the effects of IFVs on disease profiles can be studied. This approach, which relies on disease codes, can now be adopted in population cohorts with electronic health record (EHR) data using published mappings across different disease code systems. To date, disease mappings have been developed for the Human Phenotype Ontology (HPO)[259], which contains terms to describe rare diseases and rare disease symptoms, the International Classification of Diseases version 10 (ICD-10) codes, which is used to assign billing codes in hospitals, and iii) phecodes, which were created by aggregating ICD-9 codes[72] and have been shown to describe clinical phenotypes more accurately for research purposes compared to codes from other systems[260].

Here, I tested for effects of IFVs and genetically-predicted levels of corresponding IEM-related metabolites on IEM-related disease profiles in the UK Biobank[67]. The use of PheRSs in a population-based cohort with genotypic and phenotypic data enables large-scale assessment of variant effects on disease profiles mimicking those observed in the corresponding IEM or rare disease.

6.3. Aim and Objectives

In this study, the PheRS approach was applied as a complementary method to the bespoke framework with the aim of systematically estimating the effects of IFVs on IEM-related disease profiles. The objectives of this study were to:

1. Estimate the association of IFVs with PheRSs for linked IEMs or rare diseases reported in Orphanet and OMIM;
2. Test whether the IEM-related metabolic effects of IFVs confer additional risk, beyond a polygenic background, of having a high PheRS for the corresponding IEM.

6.4. Methods

6.4.1. Study Design and Participants

The UK Biobank[67] is a prospective cohort of 500,000 participants aged between 40 and 69 years who were recruited between the years 2006-2010. Recruitment was performed at 22 assessment centres that were designed for the purpose and located across the United Kingdom[261]. Participants provided electronic signed consent at recruitment and ethics approval for the UK Biobank study was obtained from the North West Centre for Research Ethics Committee (11/NW/0382).

6.4.2. Measurements and Exclusions

6.4.2.1. Genetic profiling

Genetic profiling conducted in the UK Biobank has been described previously[67]. Briefly, blood samples were collected from participants at recruitment, and DNA was extracted within 18 months of collection. Samples were then genotyped using the Affymetrix GeneTitan Multi-Channel Instrument and genotypes were called from the array intensity data. Poor quality markers that were affected by batch effects, plate effects, departures from Hardy-Weinberg equilibrium, sex effects, array effects and discordance across control replicates were set to missing in the measurement batch. If there was evidence that a marker was not reliable across batches, genotype calls were excluded from data altogether. Genotypes of all 500,000 samples were imputed to the Haplotype Reference Consortium[109] reference panel using an updated version of IMPUTE2[67,262].

6.4.2.2. Electronic health record data

EHR data containing disease codes from the International Classification of Diseases version 10 (ICD-10) from the UK Biobank cohort were used. This data comprises of disease codes from the death registry, cancer registry and hospital inpatient episodes data across the UK **(Table 1)**. ICD-10 codes are coded with an alphabetical character followed by digits, with longer strings of following digits corresponding to increasing specificity of the phenotype. For example, the ICD-10 code ‘E78.0’ corresponds to ‘Pure hypercholesterolemia’ while ‘E78.01’ corresponds to the more specific ‘Hypertriglyceridemia, Familial’. In UK Biobank, ICD-10 codes are specified to three digits (e.g. ‘E78.0’).

Table 1: Types of linked electronic health records in UK Biobank[263].

Type of data	External provider	Region	Period of data available
Deaths	HSCIC ISD	England and Wales Scotland	April 2006 onwards
Cancer registrations	HSCIC ISD	England and Wales Scotland	Since inception – 1980s Since inception – 1950s
Hospital inpatient episodes	HES (HSCIC) PEDW (SAIL) SMR	England Wales Scotland	Since inception – 1997 Since inception – 1999 Since inception – 1981

HES: Hospital Episode Statistics; HSCIC: Health and Social Care Information Centre; ISD: Information Services Department; PEDW: Patient Episode Data for Wales; SAIL: Secure Anonymised Information Linkage; SMR: Scottish Morbidity Records

6.4.2.3. Identification of ICD-10 codes corresponding to IEMs

ICD-10 codes for IEMs linked to significant variant-PheRS associations were identified using the rare disease database Orphanet[181]. In cases where Orphanet did not have a code or provided a more specific code than that available in UK Biobank, an approximate ICD-10 code was identified in OMIM[173] or using the website 'www.icd10data.com'. For example, 'Hypertriglyceridemia, Familial' was coded as 'E78.01' in Orphanet, but due to UK Biobank only specifying ICD-10 codes to three digits was instead approximated as 'E78.0' ('Pure hypercholesterolemia'). No ICD-10 code was found for 'Hyperlipoproteinemia, Type V' amongst sources; therefore, the ICD-10 code 'E78.5' ('Hyperlipidemia, unspecified') was approximated.

6.4.2.4. Exclusions

Of 486,954 UK Biobank participants in the dataset, 134,937 that were non-European or related individuals were excluded. An additional 30 participants that subsequently withdrew from the study were excluded, leaving 351,987 participants for analysis.

For this study, 187 IFVs not previously reported as pathogenic for the corresponding IEM in ClinVar **(Chapter 3)** were assessed as well as 24 IFVs identified in other GWASs that were identified using the same protocol described previously **(Chapter 3, Section 3.4.3.8.)**. The 24 additional IFVs are summarised in the **Appendix Ch5_ST2**.

6.4.2.5. Construction of PheRSs

The HPO database (https://hpo.jax.org/app/, last accessed August 12, 2020) was used to identify HPO terms that describe the IEMs and rare diseases linked to 124 IEM genes that harboured IFVs. The HPO database is linked to disease IDs recorded in the rare disease databases Orphanet[181] and OMIM[173] and may thus contain multiple disease IDs for the same rare disease, which increases the multiple testing burden. To minimise the multiple testing burden, I queried the HPO database for Orphanet IDs corresponding to IEMs and rare diseases, and in the absence of results for the Orphanet ID queried the corresponding OMIM ID instead. HPO terms and individually diagnosed ICD-10 codes in UK Biobank corresponding to the disease IDs detected were then converted into phecodes using previously published disease code mappings[77,264] to generate PheRSs.

PheRS generation was performed as previously described[77] in UK Biobank. Briefly, a PheRS is a summary measure of all disease codes that are related to an IEM or rare disease and are observed in an individual[77]. Disease codes frequently observed in the population are less likely to be specific to one disease; therefore, each code was weighted by the log-inverse of its observed frequency in the population **(Figure 1A)**. Weighted disease codes were then summed to give the PheRS of an individual **(Figure 1B)**. Diseases which had fewer than three describing phecodes present in at least one participant in UK Biobank, or for which less than half of all descriptive HPO terms could be translated into phecodes observed in UK Biobank, were excluded.

A

$$w_p = \log\left(\frac{N}{n_p}\right)$$

B

$$PheRS_i = \sum_{p=1}^{m} w_p x_{i,p}$$

$$\text{Where } x_{i,p} = \begin{cases} 1 \text{ if individual}_i \text{ has phenotype}_p \\ 0 \text{ otherwise} \end{cases}$$

Figure 1: Summary of PheRS generation for a specific IEM or rare disease. A.) The weight of a phenotype, w_p, is the log of the phenotype's observed frequency in the population of *N* individuals divided by the number of individuals with phenotype *p*, n_p. **B.)** For an individual *i*, the PheRS for an IEM or rare disease, as defined by *m* phecodes, is calculated as the sum of the observed weighted phecodes. Equations are taken from the original publication[77].

6.4.3. Statistical Analysis

6.4.3.1. Association testing of IFVs with PheRSs for linked IEMs and rare diseases

In this study, I tested whether carriers of IFVs were more likely to have a high PheRS value for the corresponding IEM or rare disease compared to non-carriers. To assess this systematically, I performed logistic regression with variant carrier status as the exposure and a 'high' PheRS (defined as a value above the median PheRS of participants carrying at least one relevant disease code) as the outcome, adjusting for age, sex, the first ten principal components of genetic ancestry and recruitment centre. This study design took three considerations into account:

1. Participants often had discrete numbers of phecodes across the distribution, creating a non-normally distributed PheRS outcome. The logistic regression model and 'high PheRS' definition based on the median account for this observation.

2. For each PheRS, most UK Biobank participants did not have corresponding phecodes **(Figure 3B)**, often reducing the median to zero. To increase the statistical power to detect associations, the median value was calculated using only participants with non-zero value PheRSs.
3. PheRSs describe a single IEM or rare disease; therefore, IFVs were only tested for association with PheRSs for the corresponding IEM or rare disease to reduce multiple testing burden.

Imputed dosages were used and hard-call genotypes coding for the metabolite-raising effect allele were generated with controls having a dosage ≤ 0.2, heterozygotes having a dosage between 0.9-1.1, and homozygotes having a dosage ≥ 1.8. Associations with fewer than five individuals in any stratum were excluded. Significance was assessed at FDR threshold of p=0.05. Phecode compositions of controls, heterozygotes and homozygotes were qualitatively assessed and frequency information of phecodes was obtained from Orphanet (downloaded December 2018).

To test the robustness of associations, analyses were repeated using an extreme PheRS (defined as having a score in the 95th percentile when excluding participants with no observed PheRS-related phecodes). Analyses were also repeated excluding participants in UK Biobank that were diagnosed with the corresponding IEMs and assessed at p≤0.05.

6.4.3.2. Phecode enrichment assessment

A chi-squared test was performed to identify which phecodes were responsible for the observed association between the *UGT1A1* variant rs1976391 with high PheRSs for 'Crigler-Najjar Syndrome Type 1' and 'Gilbert syndrome' as well as between the *APOE* variants rs429358 and rs204474 with high PheRSs for 'Alzheimer Disease 2' and 'Alzheimer Disease 4'. For each of the phecodes contributing to these PheRSs, the proportion of heterozygotic or homozygotic carriers with the phecode was compared with that of non-carriers (significant p≤0.05).

6.4.3.3. Weighted metabolite GRS-PheRS association analysis

A total of 52 metabolites were significantly associated with IFVs after conditional analysis ($p \leq 5x10^{-8}$), of which two were excluded due to association only with the *CYP7A1* IFV rs4738684. Standardised weighted GRSs for each metabolite were generated using results from the Metabolon mGWAS by summing the observed numbers of alleles for each variant weighted by that variant's effect size on the metabolite. Of the 178 conditionally independent variants considered across 50 metabolite scores, nine were not captured and no proxy ($R^2 \geq 0.8$) could be identified. Logistic regression was then performed with weighted metabolite GRS as the exposure and the corresponding PheRS as the outcome adjusting for age, sex, the first ten principal components of genetic ancestry and recruitment centre. To estimate the contribution of IFVs to the corresponding metabolite GRSs, models using genetic risk scores excluding the corresponding IFV of interest as the exposure and high PheRS as the outcome were tested. Significance was assessed at a Bonferroni-corrected threshold (p=0.05/167 independent tests=3x10^{-4}).

All statistical analyses and graphics were performed and produced using R version 3.5.3.[128] and STATA version 14.2[129].

6.5. Results

6.5.1. Suitability of UK Biobank for PheRS Application

Of 162 rare diseases linked to 124 IEM genes and 211 IFVs, 161 were successfully queried in the HPO database **(Figure 2, Boxes 1-2)**. In UK Biobank, PheRSs were constructed for 125 rare diseases, enabling association testing for 254 variant-disease pairs **(Figure 2, Box 3)**.

A total of 371 phecodes were required to describe the 125 PheRSs; of these, 336 (91%) were present in at least one participant of UK Biobank **(Figure 3A)**. The maximum PheRS that could be theoretically reached, based on the relevant, observed phecodes and weights in UK Biobank, was moderately and positively correlated (R^2=0.50, p=2.3x10^{-9}) with the number of participants with at least one of the relevant phecodes **(Figure 3B)**. However, many of the phecodes contributing to a PheRS were infrequently observed across participants in UK Biobank, as the maximum theoretical weighted PheRS was much larger than its corresponding unweighted value **(Figure 3C)**. This was irrespective of the specificity of the code for the rare disease. For example, the IEM 'Crigler-Najjar syndrome, type I' is characterised primarily by hyperbilirubinemia. The corresponding phecode for this symptom in the PheRS, 'Disorders of bilirubin excretion' (phecode #277.4), had a weighting of 2.95. However, weightings for five other symptoms that were less specific to the IEM had stronger weightings (e.g. 'Mental retardation' (phecode #315.3) with a weighting of 3.77) in the PheRS, indicating the low observed frequency of IEM-related symptoms in the UK Biobank cohort.

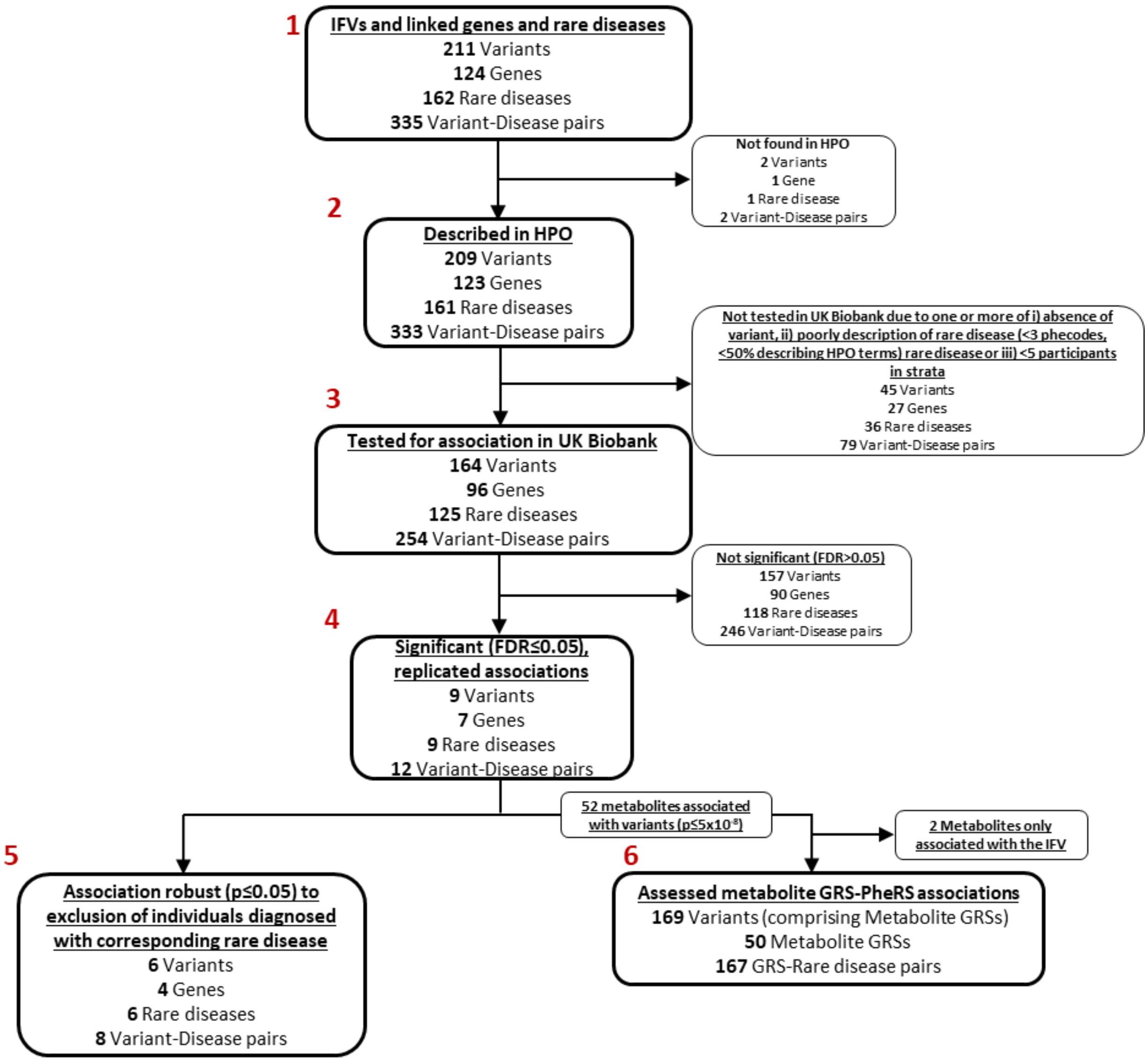

1
IFVs and linked genes and rare diseases
211 Variants
124 Genes
162 Rare diseases
335 Variant-Disease pairs
Not found in HPO
2 Variants
1 Gene
1 Rare disease
2 Variant-Disease pairs
2
Described in HPO
209 Variants
123 Genes
161 Rare diseases
333 Variant-Disease pairs
Not tested in UK Biobank due to one or more of i) absence of variant, ii) poorly description of rare disease (<3 phecodes, <50% describing HPO terms) rare disease or iii) <5 participants in strata
45 Variants
27 Genes
36 Rare diseases
79 Variant-Disease pairs
3
Tested for association in UK Biobank
164 Variants
96 Genes
125 Rare diseases
254 Variant-Disease pairs
Not significant (FDR>0.05)
157 Variants
90 Genes
118 Rare diseases
246 Variant-Disease pairs
4
Significant (FDR≤0.05), replicated associations
9 Variants
7 Genes
9 Rare diseases
12 Variant-Disease pairs
52 metabolites associated with variants (p≤5x10⁻⁸)
2 Metabolites only associated with the IFV
5
Association robust (p≤0.05) to exclusion of individuals diagnosed with corresponding rare disease
6 Variants
4 Genes
6 Rare diseases
8 Variant-Disease pairs
6
Assessed metabolite GRS-PheRS associations
169 Variants (comprising Metabolite GRSs)
50 Metabolite GRSs
167 GRS-Rare disease pairs

Figure 2: Study design. Rare diseases includes IEMs. Numbers in red reference boxes in the main text. Unless specified otherwise, ‘Variants’ refers to the number of corresponding IFVs.

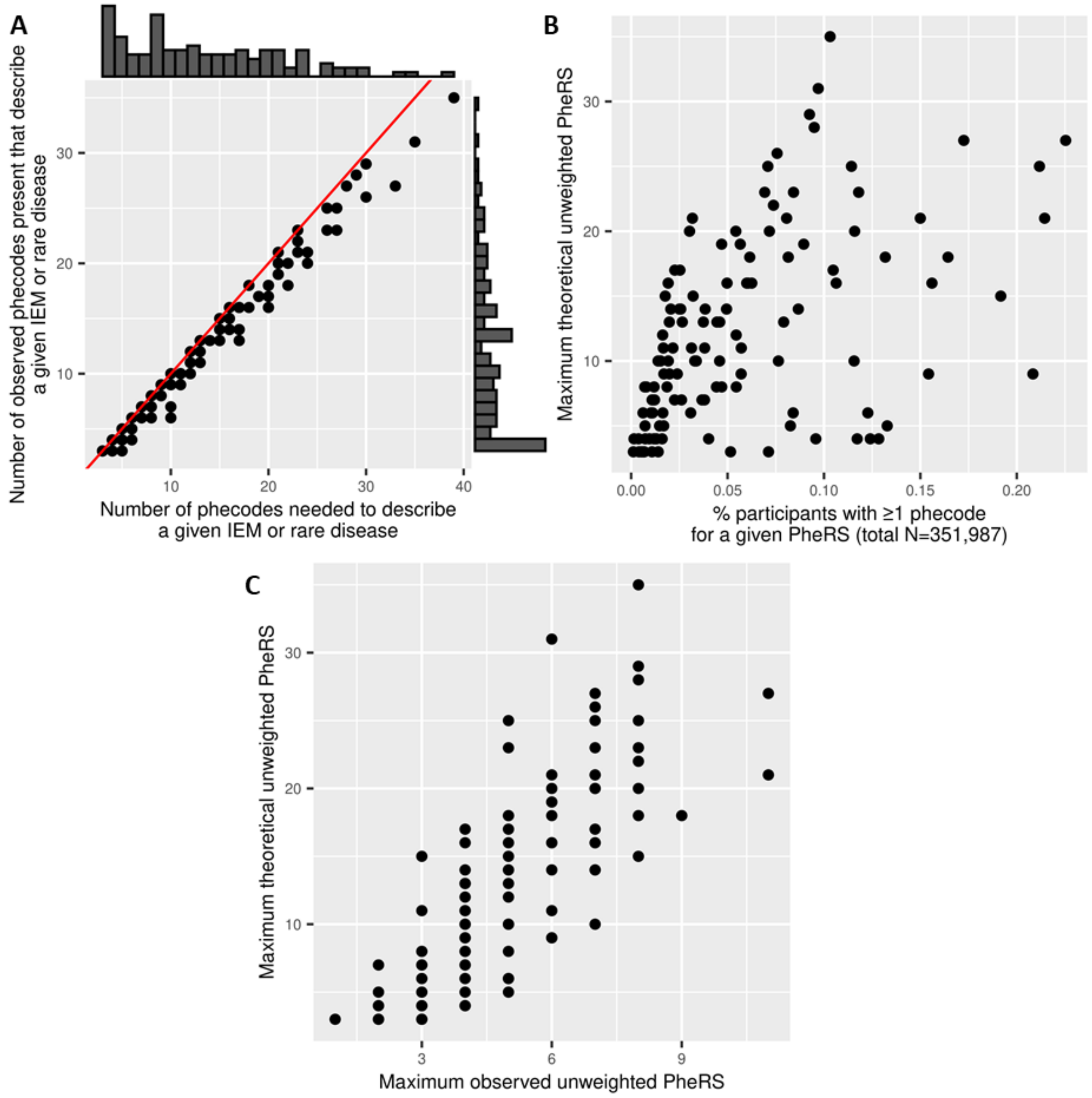


Figure 3: Diagnostic plots of 125 PheRSs. A.) The number of phecodes used to describe a disease compared to the number available in UK Biobank for PheRS generation. **B.)** The proportion of UK Biobank participants with at least one phecode for a given PheRS increases with the number of phecodes used to describe it. **C.)** The range of observed PheRSs in UK Biobank compared to the theoretical unweighted maximum PheRS. The maximum theoretical unweighted PheRS refers to the maximum PheRS attainable through diagnosis with all contributing phecodes observed in UK Biobank. PheRSs that could be described by at least three phecodes and that were observed in participants in UK Biobank were included for analysis.

6.5.2. PheRS Analysis Replicated Known Associations for Genes Linked to Familial Forms of Dyslipidemia, Hyperbilirubinemias and Alzheimer's Disease

Of the 254 tested variant-PheRS associations, 12 significant associations (FDR≤0.05) were identified across nine variants (seven IEM genes) with high PheRSs for nine IEMs. In addition to identifying associations for variants at genes known to cause familial dyslipidemias with related PheRSs, an association for the G-allele of the *UGT1A1* variant rs1976391 (MAF=0.31) with high PheRS for 'Crigler-Najjar syndrome type I', an IEM characterised by hyperbilirubinemia and jaundice[203–205], was also replicated (OR (95% CI): 1.10 (1.06, 1.15), FDR-adjusted p=3x10^{-4}) **(Figure 4)**. For this association, the phecode 'Disorders of bilirubin metabolism' was significantly enriched in carriers compared to controls (fold-enrichment: 42.73, χ^2=3.16, p=8.8x10^{-72}) **(Figure 5)**. Association analysis using a more extreme PheRS (defined as having values in the 95th percentile) as an outcome replicated 11 of the 12 associations. The 20 associations with the largest log-transformed odds ratios (regardless of significance) had large effect sizes but also had large confidence intervals due to low numbers of participants in one or more strata **(Appendix Ch6_ST1)**.

Mutations at the *UGT1A1* gene are also known to cause a milder and commonly observed form of inherited hyperbilirubinemia called Gilbert syndrome (OMIM #143500); however, this IEM was only recorded in the OMIM database (and therefore missed in initial analysis using Orphanet database IDs). Association analysis of IFVs at the *UGT1A1* gene with high PheRS for Gilbert syndrome identified a significant association for the variant rs1976391 (OR (95% CI): 1.23 (1.16, 1.30), p=1.4x10^{-11}) that was primarily driven by enrichment of the phecodes 'Disorders of bilirubin excretion' and 'Jaundice (not of newborn)' in carriers compared to non-carriers **(Appendix Ch6_Fig1)**.

Associations of the *APOE* variants rs429358 and rs204474 with high PheRSs for 'Alzheimer Disease 2' and 'Alzheimer Disease 4' were also observed **(Figure 4)**. These associations were primarily driven by enrichment of the phecodes 'Alzheimers disease' and 'Dementias', as well as other phecodes (such as 'Amyloidosis' and 'Mild cognitive impairment'), in carriers compared to non-carriers **(Appendix Ch6_Fig2-4)**.

Gene	rsID	Homozygote non-carriers / Heterozygotes / Homozygote carriers	Most specific IEM-related metabolite	IEM
UGT1A1	rs1976391	164885/ 151891/ 35169	bilirubin (Z,Z)	Crigler-najjar Syndrome Type 1
LPL	rs1441764	164507/ 151520/ 35275	1-oleoyl- 2-linoleoyl-glycerol (18:1/18:2)	Hyperlipidemia, Familial Combined, 3
LPL	rs15285	28689/ 143383/ 179915	1-oleoyl- 2-linoleoyl-glycerol (18:1/18:2)	Hyperlipidemia, Familial Combined, 3
LDLR	rs118068660	2681/ 56886/ 278899	cholesterol	Homozygous Familial Hypercholesterolemia
CYP7A1	rs4738684	154959/ 156638/ 39737	taurocholenate sulfate	Hypercholesterolemia due to cholesterol 7alpha-hydroxylase deficiency
APOE	rs429358	251141/ 92349/ 8497	cholesterol	Alzheimer Disease 2
APOE	rs204474	146669/ 159879/ 43838	1-(1-enyl-stearoyl)- 2-linoleoyl-GPE (P-18:0/18:2)*	Alzheimer Disease 4
APOE	rs429358	251141/ 92349/ 8497	cholesterol	Alzheimer Disease 4
APOE	rs429358	251141/ 92349/ 8497	cholesterol	Dysbetalipoproteinemia
APOB	rs934197	154276/ 157245/ 40004	cholesterol	Homozygous Familial Hypercholesterolemia
APOA5	rs964184	264862/ 80893/ 6232	1-oleoyl- 2-linoleoyl-glycerol (18:1/18:2)	Hyperlipoproteinemia, Type V
APOA5	rs964184	264862/ 80893/ 6232	1-oleoyl- 2-linoleoyl-glycerol (18:1/18:2)	Hypertriglyceridemia, Familial

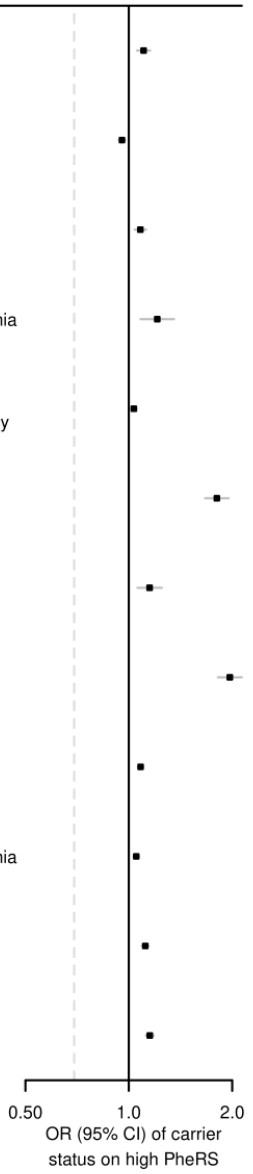


Figure 4: Summary of significant IFV-PheRS associations (FDR≤0.05). Odds ratios represent the per allele change in risk of having a PheRS where coded alleles represent alleles associated with increasing levels of the most specific IEM-related metabolite.

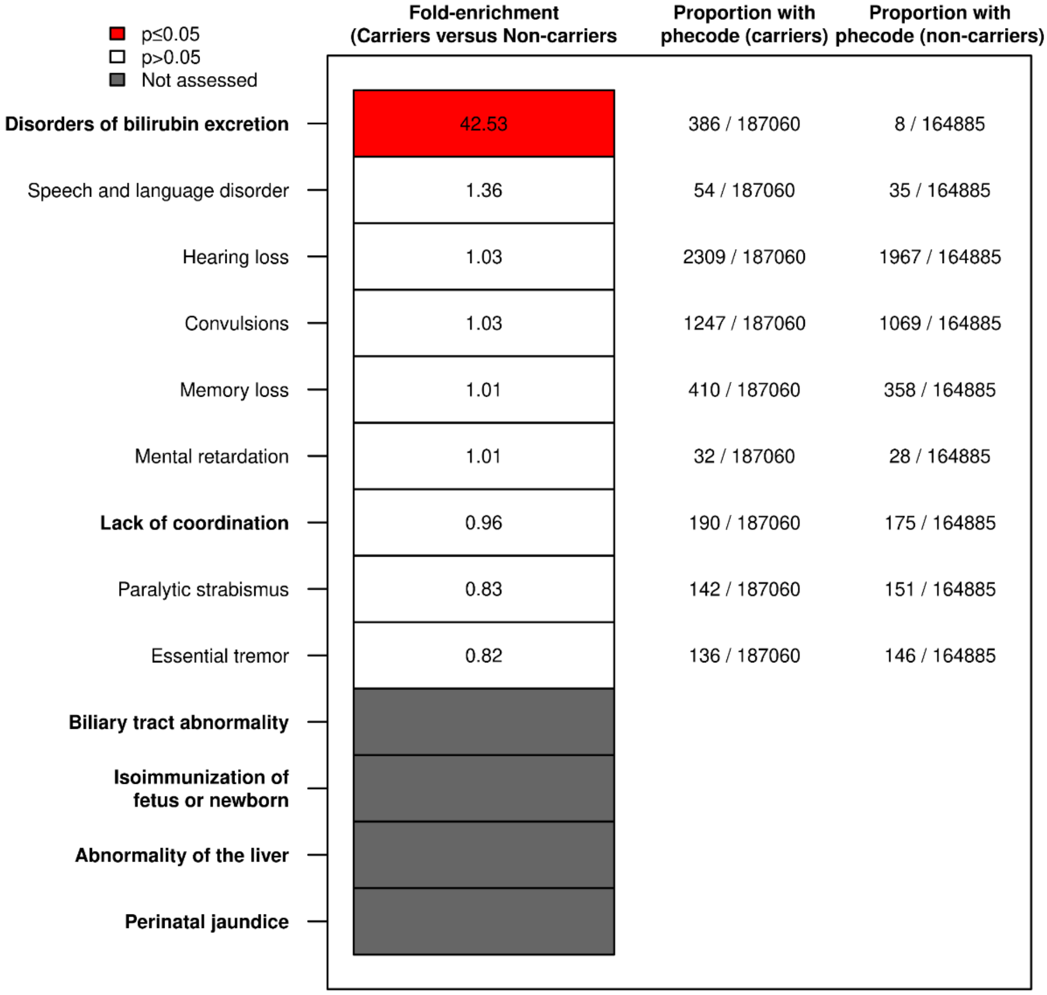


Figure 5: Enrichment assessment (χ^2 p≤0.05) of phecodes contributing to the PheRS for 'Crigler-Najjar syndrome, Type 1' in heterozygote and homozygote carriers compared to homozygotic non-carriers of the G-allele of the *UGT1A1* variant rs1976391. Phecodes in bold are observed in 80-99% observed IEM cases while other phecodes are observed in 5-29% IEM cases, as reported in the Orphanet database[181].

6.5.3. Eight of 12 Significant Variant-PheRS Associations Were Not Due to Participants Diagnosed With the Corresponding IEM

The nine IEMs and rare diseases identified were linked to ICD-10 codes in Orphanet and other databases. 'Hypertriglyceridemia, Familial' was assigned to disease code E78.0 ('Pure hypercholesterolemia') **(Table 2)**. As no ICD-10 code was found for 'Hyperlipoproteinemia, Type V' in Orphanet or OMIM, the disease code E78.5 ('Hyperlipidemia, unspecified') was assigned instead. IEMs or rare diseases and their mapped ICD-10 codes are listed in **Table 2**.

Table 2: IEMs linked to significant variant-PheRS associations and mapped to ICD-10 codes, as reported in Orphanet or OMIM. Farthest right column shows the number of UK Biobank participants (total N=351,987) that were diagnosed with the IEM.

IEM gene	IEM or rare disease	ICD-10 diagnosis code for the IEM or rare disease	Notes	Number of UK Biobank participants with ICD-10 code
UGT1A1	Crigler-Najjar syndrome type I	E80.5	N/A	0
UGT1A1	Gilbert syndrome	E80.4	N/A	372
LPL	Hyperlipidemia, Familial Combined, 3	E78.3	N/A	9
LDLR	Homozygous familial hypercholesterolemia	E78.0	N/A	33,308
CYP7A1	Hypercholesterolemia due to cholesterol 7alpha-hydroxylase deficiency	E78.0	N/A	33,308
APOE	Dysbetalipoproteinemia	E78.2	N/A	143
APOE	Alzheimer Disease 4	G30.0	N/A	80
APOE	Alzheimer Disease 2	G30.0	N/A	80
APOB	Homozygous Familial Hypercholesterolemia	E78.01	UK Biobank ICD-10 codes are not as specific; therefore, this IEM was assigned 'E78.0' ('Pure hypercholesterolemia') instead	33,308
APOA5	Hypertriglyceridemia, Familial	E78.01	UK Biobank ICD-10 codes are not as specific; therefore, this IEM was assigned 'E78.0' ('Pure hypercholesterolemia') instead	33,308
APOA5	Hyperlipoproteinemia, Type V	E78.5	No ICD-10 code was found in Orphanet or OMIM, so the ICD-10 diagnosis was approximated to be E78.5 ('Hyperlipidemia, unspecified')	5,116

A total of 36,138 of 351,987 participants in UK Biobank were diagnosed with at least one IEM corresponding to a PheRS significantly associated with an IFV **(Table 2)**. Eight of the 12 tested associations remained significant (p≤0.05) after excluding participants who were diagnosed with the ICD-10 code for the corresponding IEM **(Figure 2, Box 5)**. The four non-significant associations (p>0.05) were those for which the closest corresponding ICD-10 code was E78.0 ('Pure hypercholesterolemia').

No diagnoses were identified for ‘Crigler-Najjar syndrome type I’. However, complete attenuation was observed after the exclusion of 372 participants diagnosed with Gilbert syndrome **(Table 2)** for associations of the IFV rs1976391 with high PheRSs for ‘Crigler-Najjar syndrome type I’ (OR (95% CI): 1.10 (1.06, 1.15), p=3x10^{-4} before vs 1.02 (0.97; 1.06), p=0.43 after exclusion) and for ‘Gilbert syndrome’ (OR (95% CI): 1.23 (1.16, 1.30), p=1.4x10^{-11} before vs 1.06 (1.00, 1.13), p=0.068 after exclusion).

The remaining associations retained similar odds ratios and p-values as in the main analysis **(Figure 2, Box 5)**. However, the association of the *APOE* variant rs429358 with high PheRS for ‘Dysbetalipoproteinemia’ was almost completely attenuated after exclusion of participants with the ICD-10 code E78.2 (‘Mixed hyperlipidemia’; n=143) (OR (95% CI): 1.08 (1.06-1.10), FDR-adjusted p=3x10^{-14} vs 1.02 (1.00, 1.05), p=0.048 before and after exclusion).

6.5.4. Significant Variant-PheRS Associations Implicated Metabolites in Disease and Highlight the Specific Effects of *UGT1A1* Variant Rs1976391 on Bilirubin Levels

The nine IFVs highlighted in significant variant-PheRS associations were associated with 50 metabolites in the Metabolon mGWAS for which GRSs could be constructed **(Figure 2, Box 6)**. Aside from three bilirubin derivatives, which are cofactors or vitamins, all other metabolites were lipid species. Metabolite GRSs comprised of a median of 17 variants (range: 3; 36), and 167 weighted metabolite GRS-PheRS pairs available for testing.

In 125 of 167 tested cases, genetic susceptibility for increased metabolite levels was significantly associated with high PheRS (Bonferroni p≤3x10^{-4}). Association analysis repeated after excluding the corresponding IFV resulted in complete attenuation in 45 of the 125 cases. Of these 45 associations, 30 1-SD increases in GRSs for lipid species with high PheRS for ‘Alzheimer Disease 4’ and ‘Alzheimer Disease 2’ were completely attenuated after excluding the *APOE* variant rs429358 from the GRS. Lipid species in these associations comprised of 18 of the 31 distinct lipid species that rs429358 was associated with. In three other cases, 1-SD increases in GRSs for bilirubin (Z,Z), bilirubin (E,E) and biliverdin with high PheRS for ‘Crigler-Najjar syndrome type I’ were completely attenuated after excluding the *UGT1A1* variant rs1976391 **(Figure 6)**. The *UGT1A1* variant rs1976391 was also solely responsible for the association between genetically-predicted levels of bilirubin metabolites and high PheRS for Gilbert syndrome **(Figure 6)**.

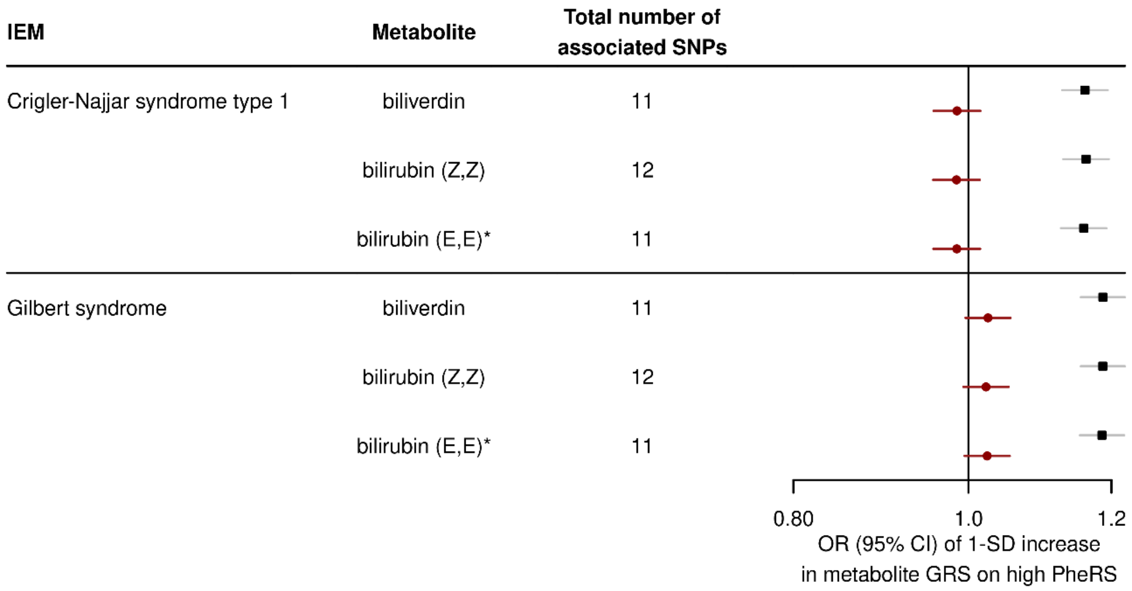


Figure 6: Association of 1-SD increases in weighted metabolite GRSs with the odds of having high PheRS for 'Crigler-Najjar Syndrome, Type I' and 'Gilbert Syndrome'. Black bars represent the score including all associated, conditionally independent variants. Red bars represent the association when the GRS does not include the *UGT1A1* variant rs1976391.

6.6. Discussion

6.6.1. Summary of Study Findings

I presented a novel application of the PheRS in a large population cohort to characterise variant effects on rare disease-related profiles. Systematic construction of PheRSs across 125 rare diseases in the UK Biobank showed that PheRSs can detect common variant effects on disease, as evidenced by the replicated associations at genes known to cause familial forms of dyslipidemias, Alzheimer's disease and hyperbilirubinemias. I also demonstrated that IEM-related metabolite levels associated with these variants were also associated with corresponding PheRSs. Notably, associations of bilirubin levels with PheRS for Crigler-Najjar syndrome I and for a less severe form of hyperbilirubinemia, Gilbert syndrome were driven entirely by the *UGT1A1* IFV rs1976391, whereas associations of many lipid species with PheRSs for familial dyslipidemias and Alzheimer's disease were driven by variants at several independent loci. These findings suggest that PheRSs can be used to detect common variant associations with IEM-related disease profiles and assess the relative contributions of individual variants to metabolite and disease profile associations.

6.6.2. Novelty of Findings

PheRSs complemented the phenome-wide approach outlined in **Chapter 5** by enabling assessment of common variant associations with disease diagnoses that in combination relate to a specific, rare, Mendelian disorder. In contrast to previous studies, which have only tested PheRSs for up to 16 Mendelian disorders to date[77,250], I developed PheRSs for 125 IEMs and rare diseases and assessed their utility in the UK Biobank population cohort. Specifically, the large sample size of UK Biobank provided good coverage (91%) of the phecodes required across rare diseases. Furthermore, the replication of known variant-disease associations in this study demonstrated the proof-of-concept that PheRSs are also sensitive to variant effects in a population cohort setting. Of the 12 associations detected, four were driven by participants who had been diagnosed for the IEM 'Pure hypercholesterolemia'. This attenuation may be conservative as the corresponding ICD-10 code was highly non-specific, though used in the absence of more specific ICD-10 codes for the corresponding IEMs. Associations of the *UGT1A1* variant rs1976391 with high PheRS for 'Crigler-Najjar syndrome, Type I' and 'Gilbert syndrome' were also attenuated after excluding 372 participants diagnosed with Gilbert syndrome. Gilbert syndrome is a less extreme form of hyperbilirubinemia than Crigler-Najjar syndrome[265], suggesting that PheRS for the latter captures information about the former as expected.

Association of the *APOE* variants rs429358 and rs204474 with high PheRSs for 'Alzheimer Disease 2' and 'Alzheimer Disease 4' were detected. Common variants at the *APOE* gene, including rs429358, are known to predispose carriers to Alzheimer's disease and dementia[266]. The association of rs429358 with high PheRS for 'Alzheimer Disease 4' is notable since Alzheimer Disease 4 is usually attributed to mutations in the *PSEN2* gene[267]. *PSEN2* encodes a subunit protein of gamma-secretase that cleaves amyloid precursor proteins, the products of which are known to contribute to Alzheimer's disease development[268]. The variant-PheRS associations detected, as well as the enrichment of the phecodes 'Amyloidosis' in the PheRS for 'Alzheimer Disease 4' in carriers compared to non-carriers of the variant rs429358 **(Appendix Ch6_Fig3)**, suggests that *APOE* variants may either affect risk of 'Alzheimer Disease 4' alone or alongside variation at the *PSEN2* gene, as has been suggested by a previous study[268].

Despite successfully replicating known associations between genetic variants and PheRSs for rare diseases, I was unable to identify additional novel examples. I expected to identify more novel associations, given several new variant-metabolite-phenotype links were identified using the phenome-wide approach in **Chapter 5** and previous studies had also reported large overlap between genetic loci associated with rare, Mendelian disorders and complex diseases[70,71].

The lack of detected associations in this study suggests a lack of study power, one potential reason for which is the choice of cohort used. In a previous study, five PheRSs were developed and applied within a hospital-based cohort to identify rare variant associations with PheRSs as well as clinical endpoints likely resulting from these rare variants[77]. This is in contrast with my approach, which developed and applied PheRSs for 125 IEMs and rare diseases in the population-based cohort UK Biobank. Despite the large sample size, the UK Biobank is comprised of participants who are shown to represent a healthier demographic[222,269]. This means that there is a lower prevalence and incidence of disease in

UK Biobank compared to what is observed in the general population. This was supported by **Figure 3**, which showed that up to 80% UK Biobank participants do not carry corresponding phecodes used to construct a given PheRS.

Lack of study power may also arise from the fact that symptoms frequently observed in the clinical presentation of a rare, Mendelian disorder often occur in early life. According to the Orphanet database, six conditions are observed in 80-99% patients suffering from Crigler-Najjar syndrome type I: 'Disorders of bilirubin excretion', 'Biliary tract abnormality', 'Isoimmunization of fetus or newborn', 'Abnormality of the liver' and 'Perinatal jaundice' **(Figure 5)**. However, in this study diagnoses were only observed for 'Disorders of bilirubin excretion', the enrichment of which appeared to drive the observed variant-PheRS association **(Figure 5)**. Similarly, associations between *APOE* variants rs429358 and rs204474 with high PheRSs for familial forms of Alzheimer's disease were primarily driven by phecodes for 'Alzheimers disease' and 'Dementias' **(Appendix Ch6_Fig2-4)**. These examples suggest that the UK Biobank may be unsuitable for constructing PheRSs for diseases characterised by extreme, early-onset symptoms.

Another difference between the current study and previous ones is that this study used genotyping data to assess the effects of common variants while previous studies focus on the detection of rare variant effects using whole exome sequencing[77]. It is possible that the associations detected in the current study tag the effects of multiple rare variants (i.e. 'synthetic associations'). This was not tested due to the lack of sequencing data at the time of analysis, though the recent availability of whole exome sequencing[68] in the UK Biobank can be combined with previously published methods[270] to test for synthetic associations.

Using metabolomic data from the Metabolon mGWAS, I also showed that IFVs could be used to infer associations between corresponding metabolite levels and PheRSs. Notably, I showed that variants driving metabolite-PheRS associations could be attributed to the effects of one variant, such as the variant rs1976391 at the *UGT1A1* gene on bilirubin levels. I also found that metabolite-PheRS associations could also be driven by variants at many loci, as evidenced by association analyses of lipid species with PheRSs for familial dyslipidemias. The *APOE* variant rs429358 was the primary genetic signal driving the associations of several lipid species with PheRSs for Alzheimer's disease but not for others, suggesting that this variant could have pleiotropic effects. These results, which have not been similarly assessed or replicated elsewhere to my knowledge, suggest that PheRSs capture the metabolic effects of genetic variation that lead to clinically manifest outcomes.

6.6.3. Study Strengths and Limitations

A strength of this study was the novel adaptation of the PheRS approach to assess variant effects on disease in a population-based cohort at scale. Testing of this approach enabled the characterisation of PheRSs at scale in the UK Biobank and assessment of its feasibility. In this study, PheRSs were easy to develop at scale and were successfully used to replicate known variant-disease associations, thus demonstrating their ability to complement the assessment of individual phenotypes in **Chapter 5**. PheRS construction was also facilitated using standardised disease code systems as well as access to mappings across these systems, enabling replication of the current study's results in any cohort.

Another strength of this study was the integration of results from the Metabolon mGWAS to estimate the associations of IEM-related metabolites with corresponding PheRSs. This effort revealed that lipid species driven by genome-wide polygenic effects can predispose individuals to increased morbidity of familial dyslipidemia-related symptoms, in contrast to the rs1976391-specific effects on bilirubin levels and PheRS for Crigler-Najjar syndrome type I. Thus, I showed that the integration of metabolomics data can be used to assess the relative contributions of individual genetic variants and to assess the genetic architecture underlying disease profiles related to rare, Mendelian disorders.

In addition to its strengths, this study also had limitations. During this study, it became apparent that the UK Biobank cohort does not capture diagnoses related to extreme, early-onset symptoms of IEMs because of the cohort's over-representation of healthy participants. This remains the likeliest reason for the lack of novel associations identified despite measures taken to increase study power (such as the testing of IFVs with PheRSs for diseases known to be caused by the corresponding IEM gene). Therefore, this study could be replicated in other population-based cohorts with genotypic and EHR data such as the Precision Medicine Initiative 'All of Us' research programme[271,272]. For common diseases that are usually followed up in secondary care, hospital-based cohorts such as the BioVU cohort[258] may also be useful for testing for additional associations.

Other limitations of this study were related to the method by which PheRSs are constructed. For example, PheRSs were developed and deployed within the same cohort. This could have caused the analysis to detect cohort-specific effects, however, the lack of novel associations detected suggests that this was not a problem with the current study. Another limitation was that the set of disease codes used to describe rare diseases, while completely derived from the HPO database, could differ depending on whether the Orphanet ID or the OMIM ID for the disease was queried. To address this limitation, I queried Orphanet IDs first and only used OMIM IDs if a search result was missing for the former to obtain one PheRS per rare disease and only constructed PheRSs for which enough descriptive phecodes were available. These approaches helped to reduce the multiple testing burden of the study. However, a comparison of PheRSs derived from different database IDs for the same disease would be required to facilitate robust PheRS construction and assessment.

Other limitations of the PheRS methodology included i) translation across HPO terms, phecodes and ICD-10 codes, which enabled PheRS construction in any cohort at the cost of translation inaccuracies and loss of information, and ii) phecode weights calculated based on population frequency and not on their specificity to a given disease. Yet despite these methodological limitations, previous studies have shown that PheRSs developed for select rare diseases consistently distinguish Mendelian disorder-diagnosed cases from controls[77,250]. Work to incorporate disease specificity of the symptom into phecode weights is ongoing (Dr. Stefanie Müller, *personal comms*).

6.6.4. Conclusions

This work showed that IFVs may affect IEM-related disease through their effects on IEM-related metabolites. These findings suggested that knowledge of rare, Mendelian disorders can be used to identify clinical endpoints that result from the metabolic effects of common variants at the relevant genes.

CHAPTER 7: GENERAL DISCUSSION

7.1. Summary of Findings

In this thesis, application of Bayesian methods successfully identified 20 candidate metabolite mediators of the association between weight gain and T2D **(Chapter 2)**. Candidate mediators individually accounted for little, but cumulatively for much of the association between weight gain and T2D. They were also shown to represent the combined contribution of genetic, modifiable and non-modifiable factors (such as genetic risk for increased visceral adiposity, vitamin C intake, and measures of kidney function) on metabolite levels contributing to T2D risk. Whilst highlighting potential biomarkers of disease risk linked to obesity and weight gain, this study demonstrates that individual contributions of multiple risk factors can now be assessed within the context of a single study.

Here, I also applied an IEM-centric approach to results from large-scale genotypic, metabolomic and phenotypic datasets to prioritise and phenotypically characterise variants associated with metabolite levels in the general population **(Chapters 3-6)**. In the largest GWAS metabolome to date, I showed that variants at IEM genes make a large contribution to metabolite levels despite IEM genes only accounting for 4% of all protein-coding genes in the human genome and identified variants at IEM genes that were specifically associated with metabolites linked to the corresponding IEM **(Chapter 3)**. In-depth assessment of IFV characteristics showed that many explained large proportions of variance in metabolite levels, were associated with 'extreme' metabolite levels or had non-additive effects on metabolite levels **(Chapter 4)**. Rigorous colocalisation analysis demonstrated shared genetic signals for metabolic and phenotypic consequences linked to the corresponding IEM at a large proportion of assessed loci **(Chapter 5)**. In **Chapter 6**, I also replicated established associations for IFVs at genes known to cause forms of familial dyslipidemias, Alzheimer's disease and hyperbilirubinemias with high corresponding values of PheRSs. Furthermore, integration of findings from the Metabolon mGWAS showed that PheRSs may reflect the metabolic effects of variation at one gene or at multiple genes. These results highlight genetic subgroups that may benefit from targeted disease prevention and management strategies and show that rare disease knowledge can be used to guide the identification of potential metabolic and health consequences of variation at IEM genes.

The discussions of individual chapters have already outlined the findings, strengths, and limitations of specific studies. Here, I elaborate on the strengths and limitations that are important to the overall research conducted in this thesis and consider the clinical implications of my findings.

7.2. Strengths

The timeliness of the research performed in this thesis is highlighted by the utilisation of large-scale datasets and methods that have only recently been made available. Chief among the opportunities leveraged in this thesis were the availability of untargeted metabolomic profiling data from the largest mGWAS to date as well as the recent availability of large-scale GWAS summary statistics from population-based and case-control consortia that enabled phenome-wide assessment of variants of interest. Another opportunity was the development of methods such as BVS in the weight gain study and colocalisation in the IEMs study. These unprecedented opportunities enabled the integration of findings from large-scale genetic, metabolomic and phenotypic datasets into a single study. Thus, the set of related investigations conducted and outlined in this thesis have important strengths that extend beyond earlier investigation in terms of linking genetics, modifiable and non-modifiable risk factors to metabolite levels and disease pathogenesis.

One specific advantage was the study design used in the study of weight gain and T2D ('weight gain study'; **Chapter 2**). The EPIC-Norfolk cohort is a large population-based cohort with long-term follow up, which enabled the calculation of a measure of weight change that could be assessed and compared with the more widely used measure of BMI. This study design also enabled the sequential measurement of weight change, metabolite levels and incident T2D, which follows the study's purpose to identifying potential mediators of weight change and incident T2D risk. The integration of untargeted metabolomics profiling and Bayesian methods enabled comprehensive identification of candidate mediators while accounting for between-metabolite correlations. The additional integration of data-driven approaches with genetic and phenotypic data enabled in-depth characterisation of candidate mediators and the genetic and environmental risk factors they represent, all within the context of a single study.

The primary advantage of the study of genetic overlap between IEMs and complex traits and diseases (the 'IEMs' study; **Chapters 3-6**) was the approach used. The use of findings from the largest metabolome GWAS (the 'Metabolon mGWAS') to date, which also used untargeted metabolomic profiling, enabled the prioritisation of 202 IFVs, for which over 80% metabolite associations had not been reported in the previous two largest GWAS efforts[19,20]. Furthermore, advanced chip-based genotyping technologies and imputation methods in the Metabolon mGWAS enabled the detection of low-frequency and rare variants that were also validated using sequencing data. Many of these rare variants had large metabolic effects that were detected in this work, enabling their detection for phenotypic characterisation. Access to individual-level data within one of the cohorts included in the Metabolon mGWAS, as well as GWAS summary statistics from population-based and case-control studies, enabled powered and systematic assessment of the metabolic and phenotypic consequences of variation at IEM genes at a scale that no other study has previously achieved.

Associations in GWAS summary statistics are reported without the context of other variants in LD, neglecting the possibility that traits associated with the same region may be driven by distinct causal variants or by secondary signals. Therefore, the use of colocalisation methods in the IEMs study to test for shared genetic signals was useful for implicating the phenotype association as a potential consequence of variant influences on metabolite levels. Furthermore, knowledge of the metabolic and phenotypic consequences of IEMs was used to guide the selection of genetic variants as well as IEM-related locus-metabolite and locus-phenotype associations. This reduced the likelihood of detecting false positives, which was a risk based on the large number of associations reaching significance within GWASs conducted across multiple phenotypes and cohorts.

Another advantage of the IEMs study was the adaptation of a method within the UK Biobank cohort to test for variant associations with clusters of symptoms and conditions that collectively describe a rare, Mendelian disorder. This application differs from that of previous studies by using a population-based cohort instead of a hospital-based one, thus enabling the assessment of whether such a method could be useful for identifying variant effects on disease profiles.

Finally, the findings of this thesis were replicated by many studies in other cohorts, indicating their generalisability. For example, candidate mediators identified in the weight gain study of weight gain study have been associated with BMI and with T2D in previous observational studies **(Chapter 2, Section 2.5.3.)**. In the IEMs study, I also replicated select examples highlighted in previous GWASs. Integration of multiple layers of data in the IEMs study also enabled the detection of novel examples, including loci that independently validated the association of metabolic and phenotypic consequences (e.g. the *DBH* and *TH* loci). In addition, the novel example detected at the *OPLAH* locus was supported by findings from a Mendelian randomisation study, which used an instrumental variable containing a variant in LD with the corresponding IFV to show that 5-oxoproline levels were causal for performance in intelligence tests. In the IEMs study, most of the resources used are publicly available, enabling adaptation of the approaches outlined for use in other studies with similar research purposes.

7.3. Limitations

7.3.1. Generalisability of Findings

Despite the replication of findings of this thesis in other studies, the research described in this thesis was performed primarily within cohorts of European ancestry to avoid spurious findings based on population differences. Although some associations were replicated in populations of ancestries, such as that of the *UGT1A1* variant rs1976391 with bilirubin levels and cholelithiasis[273,274], further work would be required to demonstrate the generalisability of this study's findings to other ancestries. Furthermore, genetic variants may have sex-specific effects on metabolite levels. Although examples of sex-specific metabolic effects were detected in the Metabolon mGWAS, their phenotypic assessment was beyond the scope of this research, which aimed to systematically test whether the metabolic effects of variants at IEM genes have observable phenotypic effects. Furthermore, systematic phenotypic assessment of sex-specific associations are difficult to perform at scale as GWAS summary statistics usually report results for sex-combined analyses.

7.3.2. Use of GWAS Summary Statistics to Assess Rare Variant Effects on Metabolites and Phenotypes

The large sample size, use of chip-based genotype sequencing, and integration of sequencing data in the Metabolon mGWAS enabled the rigorous detection and validation of low-frequency and rare variants. Despite these efforts, the IEMs study used GWAS summary statistics that may have been performed using older genotyping and imputation methods. Many of these phenotypic datasets were also based on genotyping rather than sequencing data. Thus, despite the rigorous detection and follow up of rare variants and their metabolic effects in the Metabolon mGWAS, this study lacked the information and power to detect and assess phenotypic associations for low-frequency and rare variants. The advent of larger cohort sample sizes, improved genotyping quality in GWASs, and availability of whole genome and exome sequencing data (such as in the UK Biobank[68]), will facilitate the detection of rare variant associations with complex traits and diseases.

7.3.3. Suitability of Population-based Cohorts to Assess Variant Effects on IEM-related Conditions

Although EPIC-Norfolk and UK Biobank provide outstanding and large-scale research resources with very detailed phenotyping, their study designs presented some limitations in the context of the research described in this thesis. For example, the EPIC-Norfolk cohort is representative of an older demographic of the population[67,107]. The prevalent diseases present within this demographic could therefore mask genetic effects on metabolite levels, though the prioritisation of genetic effects based on the known metabolic effects of corresponding IEMs reduced the likelihood that these associations were purely driven by prevalent disease.

In the IEMs study, phenotypic assessment was primarily performed using summary statistics and phenotypic data from the UK Biobank. Previous assessment has shown that the UK Biobank suffers from the 'healthy volunteer'[222] bias, therefore reducing the prevalence and incidence of diseases in this cohort compared to that observed in the general population as well as the power to detect phenotypic associations for disease outcomes. This limitation was addressed by supplementing GWAS summary statistics from UK Biobank with more powered findings from GWASs performed in case-control studies in **Chapter 5** and by only testing IFVs for association with phenotypic scores for the corresponding IEMs to reduce multiple testing burden in **Chapter 6**.

7.3.4. Reverse Causality As a Possibility in the Absence of Formal Mediation and Causality Assessment

Despite the rigorous study design and approaches used across studies, no formal assessment of mediation (in the weight gain study) or causality (in the IEMs study) could be performed. In the weight gain study, formal mediation analysis using survival data was not performed due to the complexity of integrating results across a case-cohort and a randomly selected subcohort nested within EPIC-Norfolk. To mitigate this, events of weight gain, metabolite level measurements and T2D incidence occurred sequentially, and regression models were performed according to the mediation guidelines set out by Baron and Kenny[148]. In the IEMs study, low numbers of genetic variants that were commonly associated with traits of interest across loci would have provided unreliable results in causal assessment using Mendelian randomisation[256]. Therefore, I used prior IEM knowledge to prioritise phenotypes that were likely to be a cause of genetic effects on metabolite levels. In addition, statistical colocalisation methods were used to rule out the possibility that multiple associations at a locus were merely due to linkage.

7.3.5. Relative Importance of Genetic and Non-genetic Influences on Metabolite Levels and T2D Risk

In the weight gain study, evidence suggested that genetic and other modifiable and non-modifiable risk factors contributed to events of weight gain, changing metabolite levels and altered T2D risk. This finding was achieved by integrating layers of genetic and phenotypic data to perform comprehensive characterisation of candidate mediators that is difficult to achieve in a single study. Despite these advantages, the risk factors were assessed separately, as described in the genetic risk score approach and the variance decomposition analysis **(Chapter 2, Section 2.5.4.)**. This analysis therefore precluded a comparison of the relative importance of each risk factor in contributing to changes in metabolite levels, which could be used to identify the most important risk factors to address in T2D prevention programmes.

7.4. Future Work

7.4.1. Replication of Analyses in Populations of Different Ancestry and Assessment of Sex-specific Effects

To increase the generalisability of this study's findings to other cohorts, the analyses performed in this thesis may be replicated within populations of different ancestries. GWASs of several complex phenotypes and disease outcomes have already been performed in Asian and African populations as well as in multi-ethnic cohorts[275–277]. As these efforts continue, greater coverage of phenotypes assessed within these populations will facilitate phenome-wide assessment and enable the identification and assessment of ancestry-wide and ancestry-specific effects of variation at IEM genes. In addition, sex-specific effects of genetic variants on metabolite levels, which have been identified in the Metabolon mGWAS, may also be characterised for well-powered, health-related, complex phenotypes where sex-specific differences in genetic effects have been reported, such as BMI[278] (which is a major risk factor of T2D), kidney function[279] (which is used to diagnose chronic kidney disease), and glycaemic traits[280].

7.4.2. The Need for Additional Studies and Genetic Sequencing Technologies to Characterise Rare Variant Effects

This work identified phenotypic effects for more common IFVs based on the use of GWAS summary statistics, however, emerging opportunities may help to phenotypically characterise low-frequency and rare IFVs detected in the Metabolon mGWAS. For example, new GWASs of phenotypes using improved genotyping and imputation panels may enable greater discovery of phenotypic associations for low-frequency and rare variants. Furthermore, the recent release of whole exome sequencing data for up to 200,000 participants of the UK Biobank cohort provides new opportunities to test for health-related consequences of such variants. As the power to detect phenotypic effects decreases with MAF, rare variants within genes are often collapsed and then tested for association with a phenotype of interest using burden or SKAT tests[281–283]. These methods have been used in a whole exome-sequencing study to identify rare variant effects and demonstrate the significant contribution of singletons (variants present in only one individual in the test population) in complex disease[283]. The PheRS approach, which has previously been used to detect rare variant effects on clinical endpoints in a hospital-based cohort[77], could also be used to test for rare variant effects on disease profiles linked to the corresponding IEM.

7.4.3. Replication of Analyses in Cohorts With Larger Genotypic, Metabolomic and Phenotypic Datasets

As discussed in **Chapter 5**, the metabolic effects of IFVs may not have translated into phenotypic effects for biological reasons as well as other analytical reasons. Although little can be done to address the biological reasons, the methodological limitations of this research can be addressed with better powered GWASs of complex phenotypes. Increasing study power can be achieved by applying better genotyping and imputation methods, leveraging newly available whole genome and exome sequencing data resources, and increasing the number of disease cases analysed for disease outcomes, either by increasing recruitment of cases or by performing meta-analyses across cohorts.

Analyses performed in **Chapter 6** also highlight the potential of cohort-based studies with genotypic and EHR data to systematically assess the phenotypic effects of genetic variation. Different cohort study designs can be used to phenotypically assess different complex phenotypes. For example, the UK Biobank is a population-based cohort containing EHR data from primary and secondary care sources for healthy volunteers and may be useful for studying later-onset, common diseases such as coronary artery disease. Alternatively, EHR data in hospital-based cohorts such as BioVU[258] could be used to study clinical endpoints that are treated in secondary care (e.g. severe cases of cholelithiasis that require surgery to treat).

As mentioned in **Chapter 4** and **Chapter 5**, metabolite levels associated with IFVs may reflect the underlying molecular mechanism but not be directly causal for the corresponding IEM. This could be due in part to lack of detection of metabolites that are believed to be causal, as observed in the *DBH* example **(Chapter 5, Section 5.5.3.1.)**. This work could therefore be updated to identify additional variant-metabolite-phenotype links as well as biological insights with ongoing efforts to structurally identify hitherto unknown metabolites and improvements in untargeted metabolomic profiling technologies.

7.4.4. The Need for Formal Mediation or Causal Inference Assessment to Establish the Role of Metabolite Levels in Disease Aetiology

While replication of the findings using larger genotypic, metabolomic and phenotypic datasets would increase confidence in the accuracy and generalisability of this study’s findings, formal mediation or causality assessments would be required to assess the causal directions of effect implied by findings in this thesis. Recent Mendelian randomisation studies assessing the causal effects of metabolites on one to a few select disease outcomes have been met with early success[40,284–287]. Therefore, systematic mediation and causality assessment will be facilitated with GWAS summary statistics based on larger genotypic, metabolomic and phenotypic datasets. Other study designs such as randomised trials may be useful for assessing causal effects, though these are more time-consuming and expensive to perform than *in silico* approaches and may only be feasible for metabolites that could be clinically relevant to disease outcomes.

7.4.5. The Need for Integrated Genetic and Phenotypic Information to Assess the Relative Importance of Genetic and Non-genetic Influences on Metabolite Levels and T2D Risk

It has been shown that genetic variation and environmental risk factors studied independently of each other have effects on metabolite levels that are observable and may translate into complex metabolic disease. One question arising from this work is whether these factors contribute to metabolite levels differentially. This could be investigated by expanding the variance decomposition approach applied in the weight gain study **(Chapter 2, Section 2.5.4.)** to include both genetic and non-genetic factors, thus enabling a comparison of the relative contributions of multiple associated risk factors to altered metabolite levels and incident T2D risk.

7.4.6. Extension of the IEMs Study to Phenotypically Characterise Variants at Genes Known to Cause Other Rare, Mendelian Disorders

In this thesis, I focused on variation at IEM genes due to their known metabolic consequences. However, the Metabolon mGWAS also detected variant-metabolite associations at genes known to cause other rare, Mendelian disorders. Phenotypic characterisation of these variants would also be beneficial, as other studies have shown that variants at these genes also affect complex traits and diseases[62,70]. The approach highlighted in the IEMs study could be extended to include variants at genes known to cause other rare, Mendelian disorders to systematically identify further examples of variant-metabolite-phenotype consequences. This could also help to characterise lesser-known metabolic mechanisms that affect rare, non-IEM Mendelian disorders. However, a stricter interpretation of the results would be required, as the metabolite associations in this extended study could not be used to identify variant- or gene-specific phenotypic consequences.

7.5. Clinical Implications

7.5.1. Potential Applications of Candidate Mediators As Biomarkers in T2D Diagnosis, Prevention and Management

7.5.1.1. Aetiological understanding of T2D

The 20 candidate mediators identified in the weight gain study represent metabolites that potentially lie along the causal pathway between weight gain and T2D. This is supported by the near-complete attenuation of candidate mediators of the association between weight gain and incident T2D risk. However, individual candidate mediators did not account for much of the association between weight gain and T2D risk and were not more likely to be centrally-connected metabolites in the data-driven metabolic network compared to other assessed metabolites **(Chapter 2, Section 2.5.7.)**. These findings suggest that T2D is a heterogeneous disease caused by the dysregulation of multiple metabolic processes.

Some of the candidate mediators selected in the weight gain study represented genetic risk for known T2D endophenotypes (such as metabolites belonging to amino acid and bile acid metabolism), while others represented markers of general wellbeing. One example of the latter is threonate, a metabolite derived from vitamin C. Previous studies provide some evidence to suggest that dietary vitamin C intake can lower the risk of T2D by acting as an antioxidant[288] and by lowering blood glucose levels[138,289]. The antioxidative properties of vitamin C have been suggested to be beneficial for other diseases including cardiovascular disease[290] and asthma[291]. The estimation of an inverse association of threonate with the incidence of ten other diseases corroborated the role of vitamin C as a marker of a healthy lifestyle linked to obesity and weight gain **(Chapter 2, Section 2.5.5.)**.

7.5.1.2. Candidate mediators as biomarkers for T2D prevention and prediction

Evidence in the weight gain study suggested that candidate mediators may capture additional information beyond traditional measures such as blood glucose and total cholesterol. Furthermore, their close link to genetic and behavioural influences may enable the development of targeted lifestyle interventions and public health guidelines. In addition to threonate, which represented vitamin C intake as a proxy for fruit consumption, other identified candidate mediators could be attributed to specific lifestyle behaviours associated with incident T2D risk including vegetable and fibre intake[138–140]. Inclusion of these behaviours in lifestyle guidelines and T2D prevention programmes remains under discussion[141].

Although candidate mediators highlight potential risk factors of T2D, their predictive value beyond that provided by traditional clinical measures was not measured in this study. For example, some candidate mediators such as 2-hydroxystearate and 2-hydroxypalmitate may be difficult to measure in blood samples and, in a clinical setting, not be informative beyond traditional measures like total cholesterol. However, previous evidence suggests that the inclusion of metabolite levels could increase a model’s predictive value[292].

Candidate mediators may also represent the effects of additional risk factors of T2D. However, it is unclear whether direct measurement of these metabolites is more clinically informative than directly measuring the risk factors themselves. Using diet as an example factor, food frequency questionnaires or diaries are simpler and cheaper to implement than metabolite measurement. Furthermore, metabolite levels are often sensitive to a range of risk factors, which could confuse their interpretation. However, one could also argue that food frequency questionnaires introduce reporting biases and are thus less accurate than metabolite measurements. Close consideration of these points is required prior to any recommendation that candidate mediators should be used as biomarkers to predict or prevent incident T2D.

7.5.2. Future Potential Applications of Variant-metabolite-phenotype Maps Detected in the IEMs Study

7.5.2.1. Aetiological understanding of complex diseases

In demonstrating a shared genetic basis between IEMs and complex phenotypes, I identified metabolic pathways that potentially influence complex trait development, disease pathogenesis and even health-related behaviours in the general population. For example, I detected a shared genetic signal between 5-oxoproline levels and cognitive performance at the *OPLAH* locus **(Chapter 5, Section 5.5.2.2.)**. Although cognitive performance is influenced in part by non-genetic factors such as socioeconomic environment[293,294], evidence from the current study suggests that cognitive performance can also be influenced by genetic factors. In this example, cognitive performance plausibly reflects the symptom of intellectual disability observed in OPLAH deficiency. Several other loci contained IEM-related phenotypes that could plausibly result from metabolic effects of the corresponding IFV **(Appendix Ch5_ST3)**, suggesting that knowledge of the clinical sequelae caused by rare mutations can be used to estimate the metabolic and phenotypic consequences of common variants at the same gene.

7.5.2.2. Prevention and management strategies to target genetic subtypes of disease

Newborn screening programmes exist to detect and prevent IEMs in affected neonates[11,12,156]. In the UK, newborn screening programmes test for rare Mendelian disorders such as cystic fibrosis and sickle cell disease as well as IEMs such as phenylketonuria, glutaric aciduria type I and isovaleric acidemia[295]. Familial cascade screening[50] is recommended by NICE[49] to identify individuals with a high genetic risk of developing hypercholesterolemia and coronary artery disease.

Systematic identification or screening of the population to identify “subtler” manifestations of variation at IEM genes could be an useful and cost-effective strategy. This is especially true for examples identified in this study **(Chapter 5, Section 5.5.2.)** that independently implicate the role of the same metabolic pathway in complex disease. However, demonstration of a genetic mechanism underlying disease is insufficient to predict the success of a genetic screening programme. In 1968, Wilson and Jungner[296] proposed additional criteria that should be used to assess the feasibility and practicality of implementing a genetic screening programme. Examples of these criteria include the importance of the health problem, the availability of suitable tests and treatments and the economic feasibility of identifying, diagnosing and treating high-risk populations.

Applying these criteria to the familial cascade screening programme mentioned above highlights the challenges of implementing an effective genetic screening programme. Already, the cascade screening programme achieves many of Wilson and Jungner’s criteria due to i) the prevalence of risk variants at genes known to cause hypercholesterolemia[46], ii) the socioeconomic and health burdens imposed by coronary artery disease[297,298], iii) the availability of genetic screening and biochemical tests to identify risk variants[299], and iv) the availability of statins such as simvastatin to treat disease[157]. Nevertheless, several challenges remain. For example, known mutations at the genes *APOB*, *LDLR* and *PCSK9*, while accounting substantially for familial hypercholesterolemia, do not account for all observed cases of the disease. Indeed, it is estimated that ~15% patients with familial hypercholesterolemia do not have a known mutation at any of these genes[300].

7.5.2.3. Drug and pathway discovery

The mapping of metabolite levels and metabolic pathways to clinical phenotypes in the current study highlights potential metabolic biomarkers of complex disease, some of which may capture additional information compared to current clinical measures that are specific to the corresponding disease. For example, mutations at the *CPS1* gene are known to cause hyperammonemia in the IEM CPS1 deficiency[159]. In this study and in previous ones[18,19], the variant rs1047891 at the *CPS1* gene was robustly associated with increased levels of glycine and with an elevated risk of chronic kidney disease. This association can be explained by the breakdown of ammonia via the ammonia-glycine cleavage complex in lieu of its usual breakdown via the urea cycle[160]. While a previous study[19] suggested a link between glycine levels and chronic kidney disease via this mechanism, the IEMs study was the first to demonstrate a shared genetic signal for these traits. Based on scientific evidence alone, glycine could be an useful biomarker for chronic kidney disease. Patients with chronic kidney disease who have elevated levels of glycine could also be referred for screening for variants at the *CPS1* gene and benefit from recommended treatments used to manage CPS1 deficiency and other urea cycle-related disorders[301]. However, the feasibility and efficacy of these strategies must also be rigorously assessed using the criteria discussed in the previous section.

In some cases, IFVs that had large metabolic effects were also associated with specific reactions to drug intake. One such example was the *UGT1A1* variant rs1976391, which exerted large effects on bilirubin levels **(Chapter 4)**. This variant was in LD with another variant that has been associated with adverse side reactions to atazanavir[207], which is used to treat HIV infections. Initial trials assessing the impact of pharmacogenetic screening of these variants in high LD[33,34] showed notable reductions in adverse reactions as well as improvements in clinical endpoints and improved perception of the drug in treatment compared to control groups. However, the rate of drug adherence remained unaltered between groups[33,34]. These results reflect growing public understanding and acceptance of diagnosis and treatment based on genetic screening methods but also demonstrate the existence of other complex factors that may limit the efficacy of such screening programmes, requiring further assessment.

7.6. Conclusions

Metabolic processes in the human body are strongly influenced by variation in genetic and non-genetic factors that contribute to health and disease. Here, I use data-driven approaches that combine comprehensive genetic and metabolomic profiling with phenotypic data to investigate the shared genetic basis underlying rare, Mendelian and complex polygenic diseases. Application of Bayesian methods to a prospective cohort study design enabled the identification of metabolites that mediate the known association between weight gain and T2D and assess the contributing modifiable and non-modifiable risk factors. Future work building on these results may contribute towards an improved understanding of genetic subgroups and the relative contribution of genetic and health-related behaviours to metabolite levels in complex metabolic disease.

APPENDIX

1. Supplementary Information

Metabolomics measurement

Samples for metabolomic profiling were selected in the order in which they were stored at baseline (quasi-random selection). Individuals were selected and profiled in two equally sized batches of ~6,000 participants and all analyses performed separately in each batch and meta-analysed. Samples were also taken from 1,503 (45% cases) participants from the non-overlapping T2D case-cohort. Citrate plasma samples were collected at the baseline visit without requesting an overnight fast and stored at -175^{o}C in the gas phase of liquid nitrogen until shipping for analysis, when they were stored in short-term storage at -70^{o}C.

Model specification for multivariable logistic regression

Multivariable logistic regression was performed using the R package R2BGLiMS v0.1-08-11-2019 [86] with incident T2D as the outcome, adjusted for mean-centered age, sex, height, weight gain and initial weight. For each metabolite, the prior odds of association was set to the inverse of the number of metabolites included (n=131), and 20 million Reversible Jump MCMC iterations were used to calculate the posterior probability (i.e. the proportion of models containing a given metabolite).

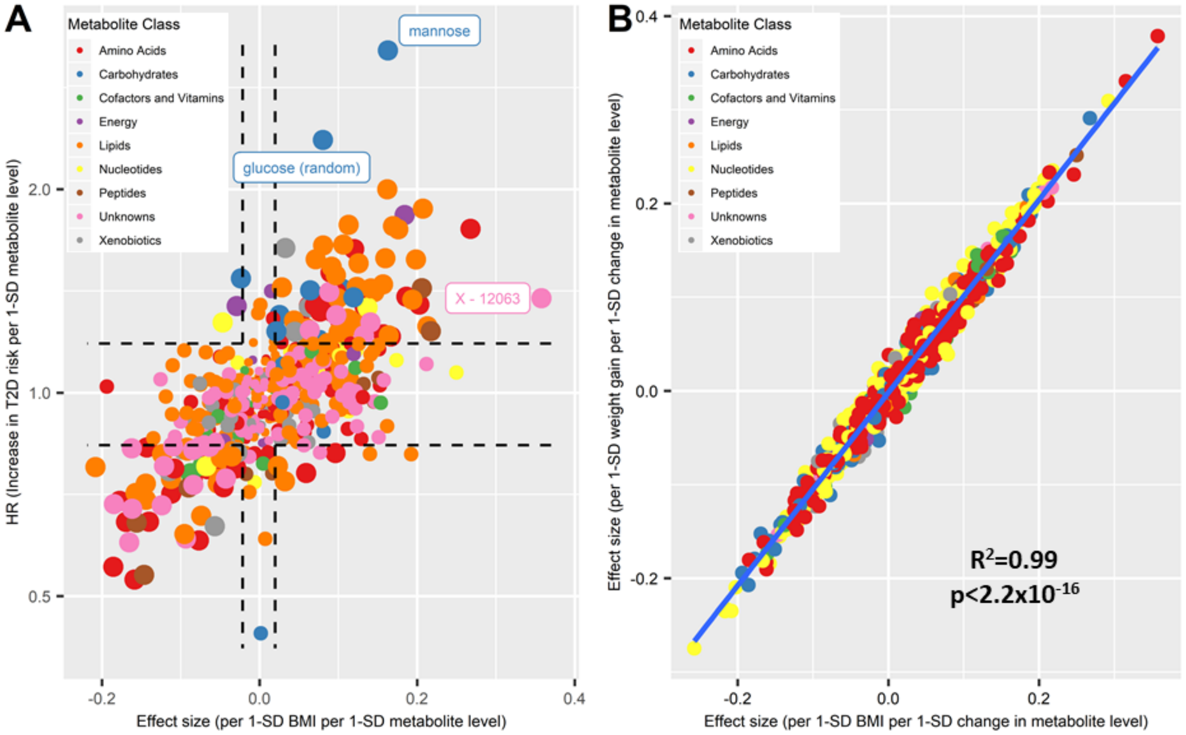
Evaluation of candidate mediators

To assess whether candidate mediators were most effective at attenuating the weight gain-T2D association, we compared the obtained HR from this model with the mean HR across 10,000 equivalent models adjusting for 22 metabolites randomly drawn from the set of metabolites significantly associated with weight gain and with incident T2D risk. Comparison was performed using a t-test.

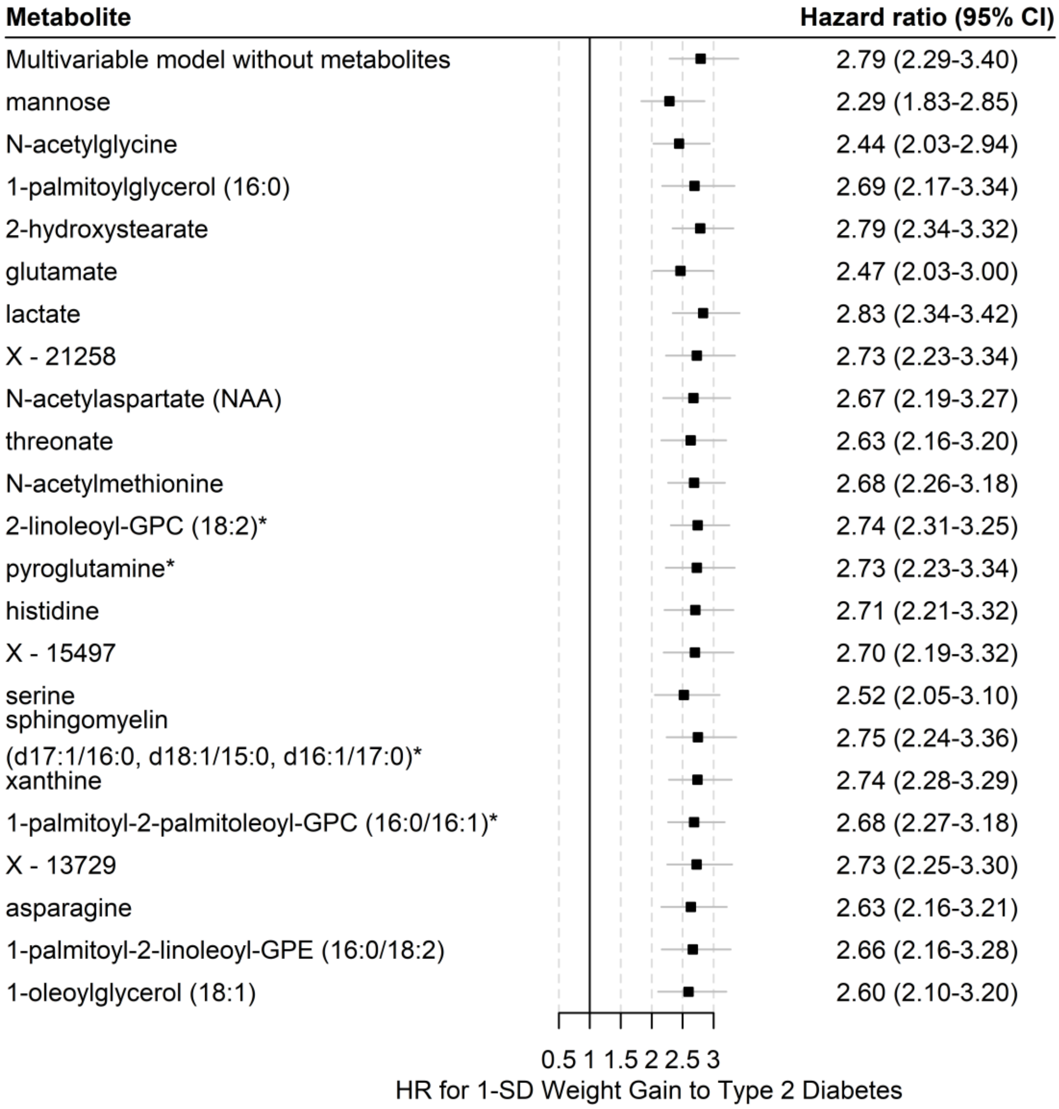
Disease specificity of candidate mediators

Censoring dates were defined for each disease separately based on electronic health records or death of participants. All available codes from hospital admission data were used to define disease onset; thus, such data may represent the more severe spectrum of each included disease. Causes of death were recorded according to ICD-9 and ICD-10 diagnostic codes and aligned afterwards, whereas disease codes obtained from hospital records were already coded based on the ICD-10 system. This study contains follow up information up to the 31st of March 2016. To investigate the specificity of candidate mediators and X-12063 for T2D, metabolite associations with the risk of 27 incident diseases and all-cause mortality using Cox regression models with age as the underlying time scale adjusted for sex in the EPIC-Norfolk subcohort. Further details and overviews of the 27 incident diseases assessed have been described elsewhere[113].

2. Supplementary Figures

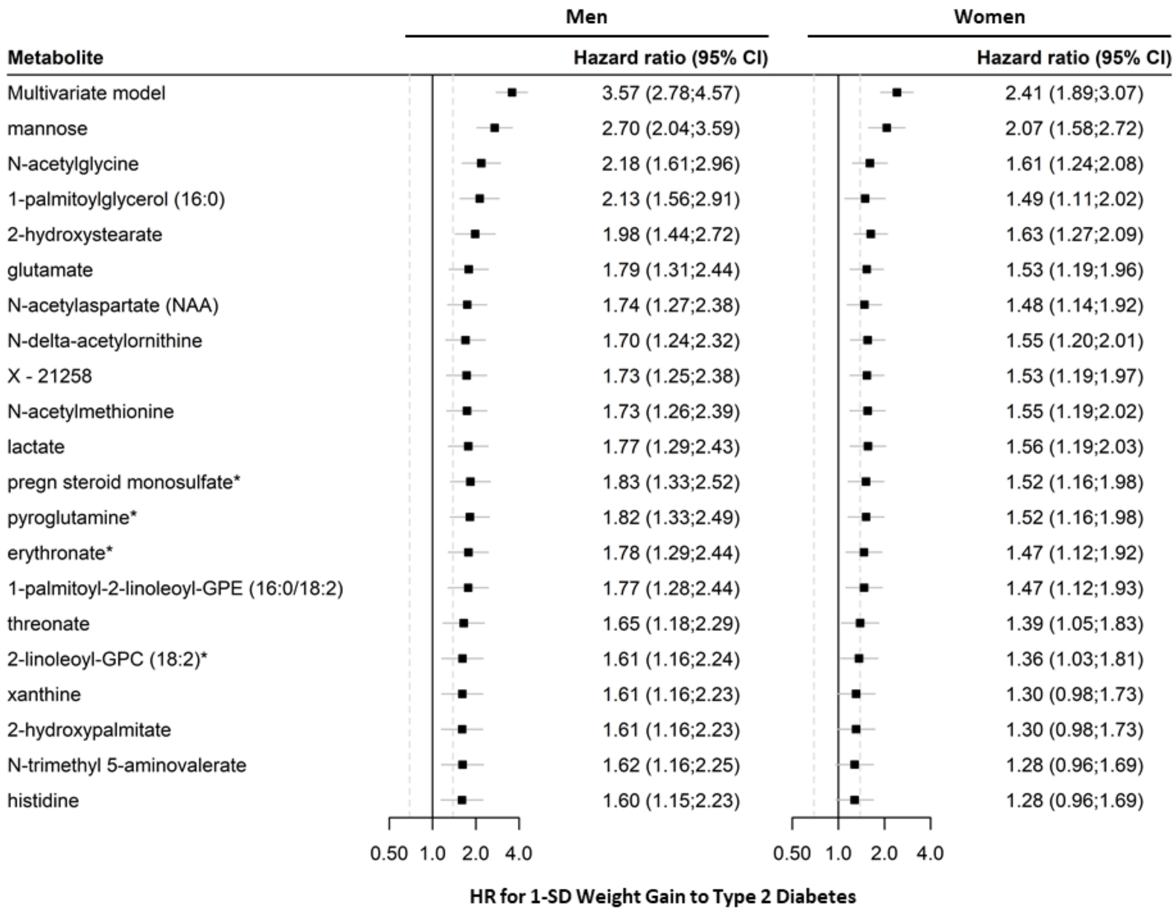


Ch2_Fig1: Comparison of metabolite associations with weight gain, BMI and incident T2D. A.) Overlap of significant metabolites associated with BMI and with T2D (FDR<0.05). Large points in the upper right and lower left quadrants represent metabolites associated with both measures while medium-sized points within the lines represent those associated with only one measure. Points in the lower right and upper left quadrants represent metabolites with discordant directions of effect between associations with weight gain and T2D. **B.)** Comparison of the effect sizes of weight gain on metabolites with effect sizes of BMI on metabolites (R^2=0.99, p<2.2x10^{-16}).



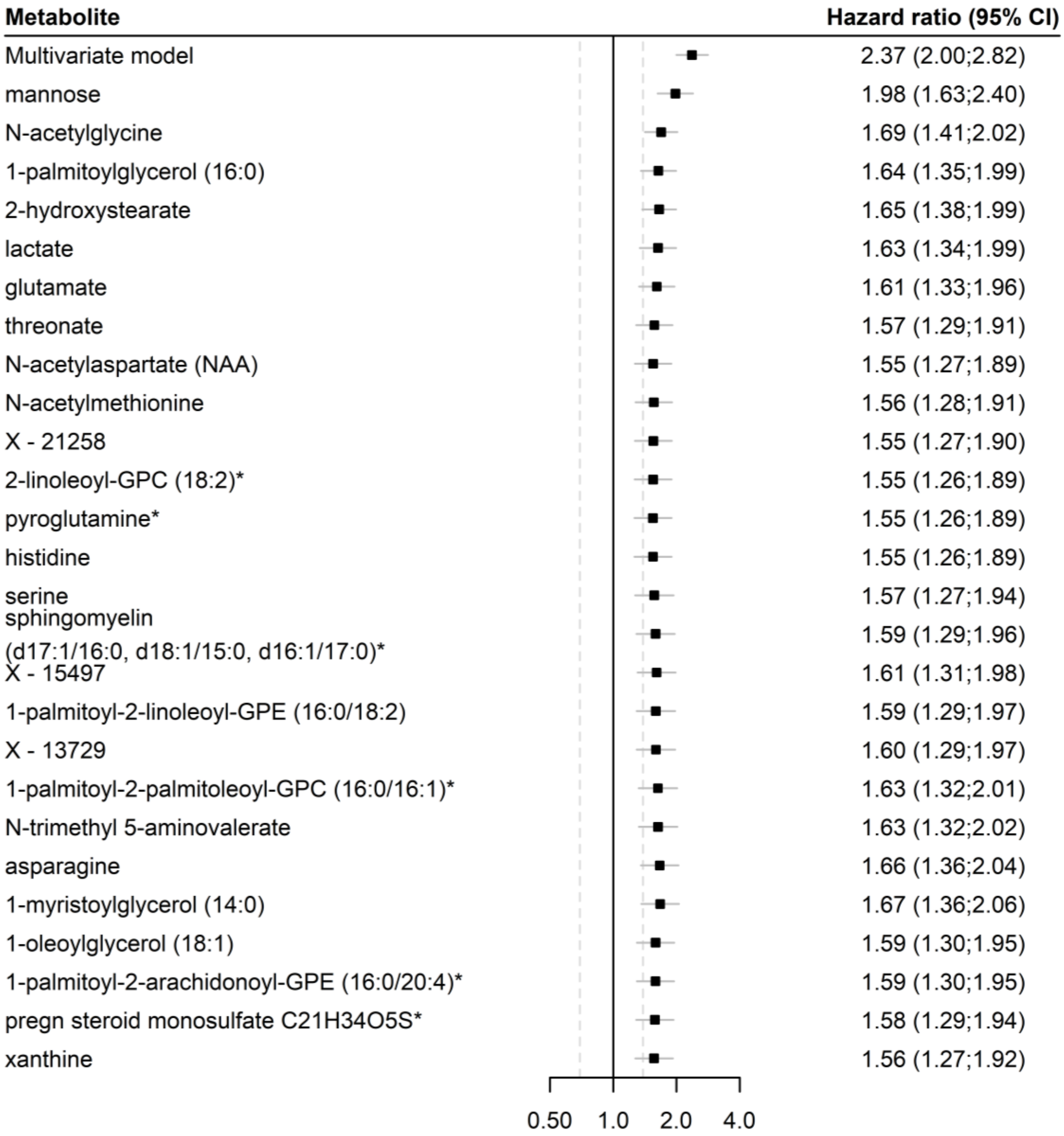
Metabolite	Hazard ratio (95% CI)
Multivariable model without metabolites	2.79 (2.29-3.40)
mannose	2.29 (1.83-2.85)
N-acetylglycine	2.44 (2.03-2.94)
1-palmitoylglycerol (16:0)	2.69 (2.17-3.34)
2-hydroxystearate	2.79 (2.34-3.32)
glutamate	2.47 (2.03-3.00)
lactate	2.83 (2.34-3.42)
X - 21258	2.73 (2.23-3.34)
N-acetylaspartate (NAA)	2.67 (2.19-3.27)
threonate	2.63 (2.16-3.20)
N-acetylmethionine	2.68 (2.26-3.18)
2-linoleoyl-GPC (18:2)*	2.74 (2.31-3.25)
pyroglutamine*	2.73 (2.23-3.34)
histidine	2.71 (2.21-3.32)
X - 15497	2.70 (2.19-3.32)
serine	2.52 (2.05-3.10)
sphingomyelin (d17:1/16:0, d18:1/15:0, d16:1/17:0)*	2.75 (2.24-3.36)
xanthine	2.74 (2.28-3.29)
1-palmitoyl-2-palmitoleoyl-GPC (16:0/16:1)*	2.68 (2.27-3.18)
X - 13729	2.73 (2.25-3.30)
asparagine	2.63 (2.16-3.21)
1-palmitoyl-2-linoleoyl-GPE (16:0/18:2)	2.66 (2.16-3.28)
1-oleoylglycerol (18:1)	2.60 (2.10-3.20)

Ch2_Fig2: Individual candidate mediator adjustment of a Prentice-weighted Cox regression model of the effects of weight gain on T2D risk. The multivariate model is adjusted for sex, height and weight at age 20 years with age as the underlying scale.



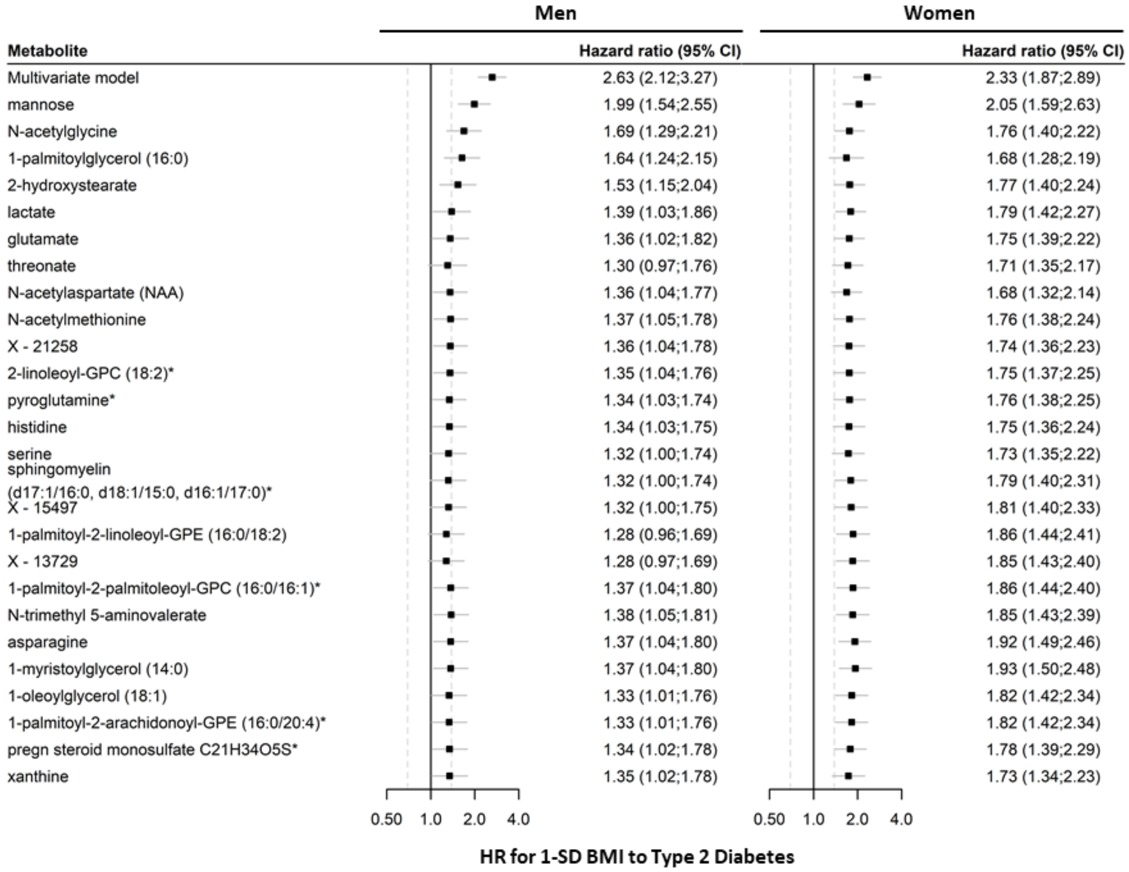
	Men	Women
Metabolite	**Hazard ratio (95% CI)**	**Hazard ratio (95% CI)**
Multivariate model	3.57 (2.78;4.57)	2.41 (1.89;3.07)
mannose	2.70 (2.04;3.59)	2.07 (1.58;2.72)
N-acetylglycine	2.18 (1.61;2.96)	1.61 (1.24;2.08)
1-palmitoylglycerol (16:0)	2.13 (1.56;2.91)	1.49 (1.11;2.02)
2-hydroxystearate	1.98 (1.44;2.72)	1.63 (1.27;2.09)
glutamate	1.79 (1.31;2.44)	1.53 (1.19;1.96)
N-acetylaspartate (NAA)	1.74 (1.27;2.38)	1.48 (1.14;1.92)
N-delta-acetylornithine	1.70 (1.24;2.32)	1.55 (1.20;2.01)
X - 21258	1.73 (1.25;2.38)	1.53 (1.19;1.97)
N-acetylmethionine	1.73 (1.26;2.39)	1.55 (1.19;2.02)
lactate	1.77 (1.29;2.43)	1.56 (1.19;2.03)
pregn steroid monosulfate*	1.83 (1.33;2.52)	1.52 (1.16;1.98)
pyroglutamine*	1.82 (1.33;2.49)	1.52 (1.16;1.98)
erythronate*	1.78 (1.29;2.44)	1.47 (1.12;1.92)
1-palmitoyl-2-linoleoyl-GPE (16:0/18:2)	1.77 (1.28;2.44)	1.47 (1.12;1.93)
threonate	1.65 (1.18;2.29)	1.39 (1.05;1.83)
2-linoleoyl-GPC (18:2)*	1.61 (1.16;2.24)	1.36 (1.03;1.81)
xanthine	1.61 (1.16;2.23)	1.30 (0.98;1.73)
2-hydroxypalmitate	1.61 (1.16;2.23)	1.30 (0.98;1.73)
N-trimethyl 5-aminovalerate	1.62 (1.16;2.25)	1.28 (0.96;1.69)
histidine	1.60 (1.15;2.23)	1.28 (0.96;1.69)

Ch2_Fig3: Sex-stratified, cumulative adjustment of candidate mediators selected using weight gain and T2D in a Prentice-weighted Cox regression model of the effects of weight gain on T2D risk. The multivariate model is adjusted for sex, height and weight at age 20 years with age as the underlying scale.

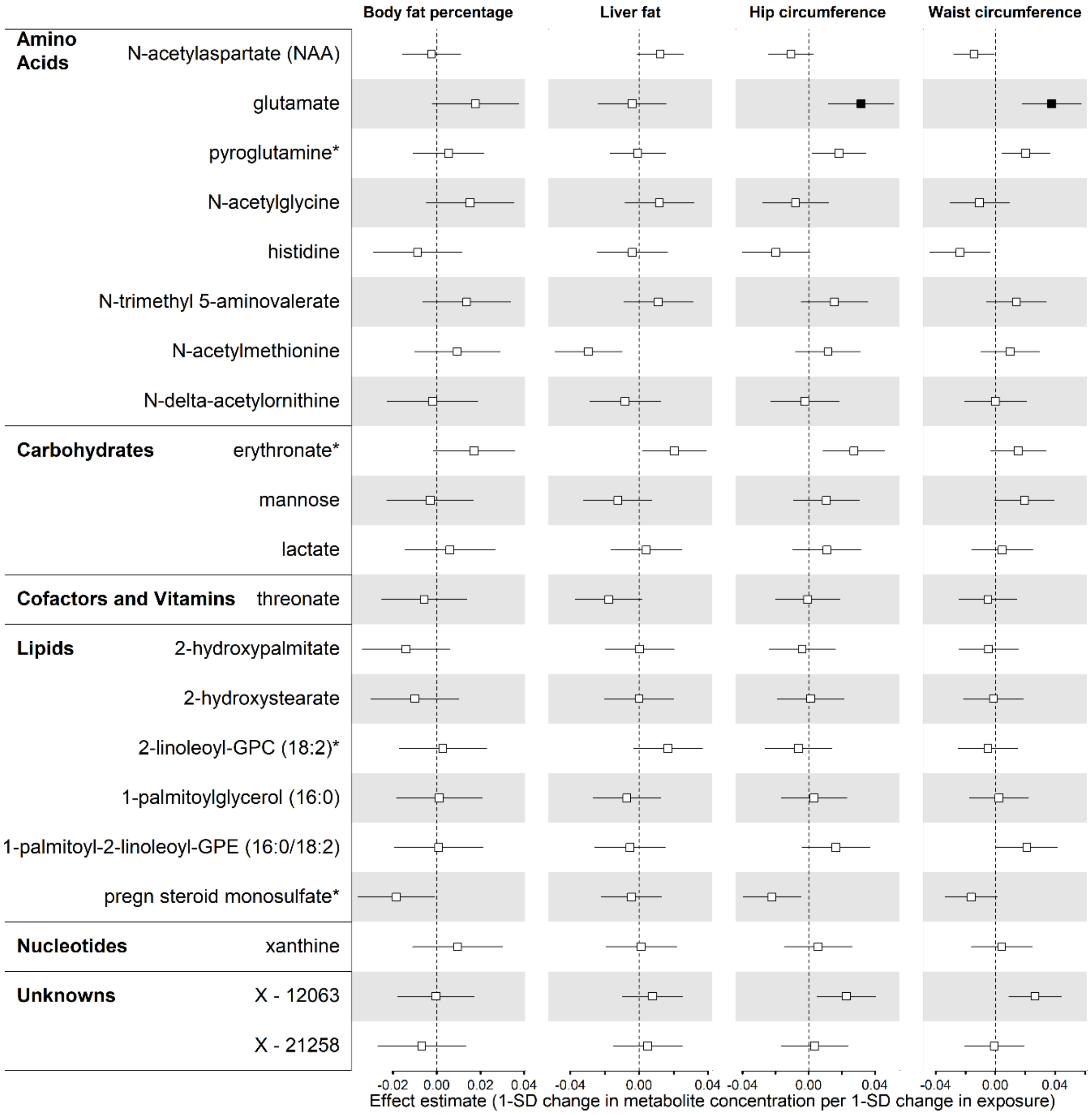


Metabolite	Hazard ratio (95% CI)
Multivariate model	2.37 (2.00;2.82)
mannose	1.98 (1.63;2.40)
N-acetylglycine	1.69 (1.41;2.02)
1-palmitoylglycerol (16:0)	1.64 (1.35;1.99)
2-hydroxystearate	1.65 (1.38;1.99)
lactate	1.63 (1.34;1.99)
glutamate	1.61 (1.33;1.96)
threonate	1.57 (1.29;1.91)
N-acetylaspartate (NAA)	1.55 (1.27;1.89)
N-acetylmethionine	1.56 (1.28;1.91)
X - 21258	1.55 (1.27;1.90)
2-linoleoyl-GPC (18:2)*	1.55 (1.26;1.89)
pyroglutamine*	1.55 (1.26;1.89)
histidine	1.55 (1.26;1.89)
serine	1.57 (1.27;1.94)
sphingomyelin (d17:1/16:0, d18:1/15:0, d16:1/17:0)*	1.59 (1.29;1.96)
X - 15497	1.61 (1.31;1.98)
1-palmitoyl-2-linoleoyl-GPE (16:0/18:2)	1.59 (1.29;1.97)
X - 13729	1.60 (1.29;1.97)
1-palmitoyl-2-palmitoleoyl-GPC (16:0/16:1)*	1.63 (1.32;2.01)
N-trimethyl 5-aminovalerate	1.63 (1.32;2.02)
asparagine	1.66 (1.36;2.04)
1-myristoylglycerol (14:0)	1.67 (1.36;2.06)
1-oleoylglycerol (18:1)	1.59 (1.30;1.95)
1-palmitoyl-2-arachidonoyl-GPE (16:0/20:4)*	1.59 (1.30;1.95)
pregn steroid monosulfate C21H34O5S*	1.58 (1.29;1.94)
xanthine	1.56 (1.27;1.92)

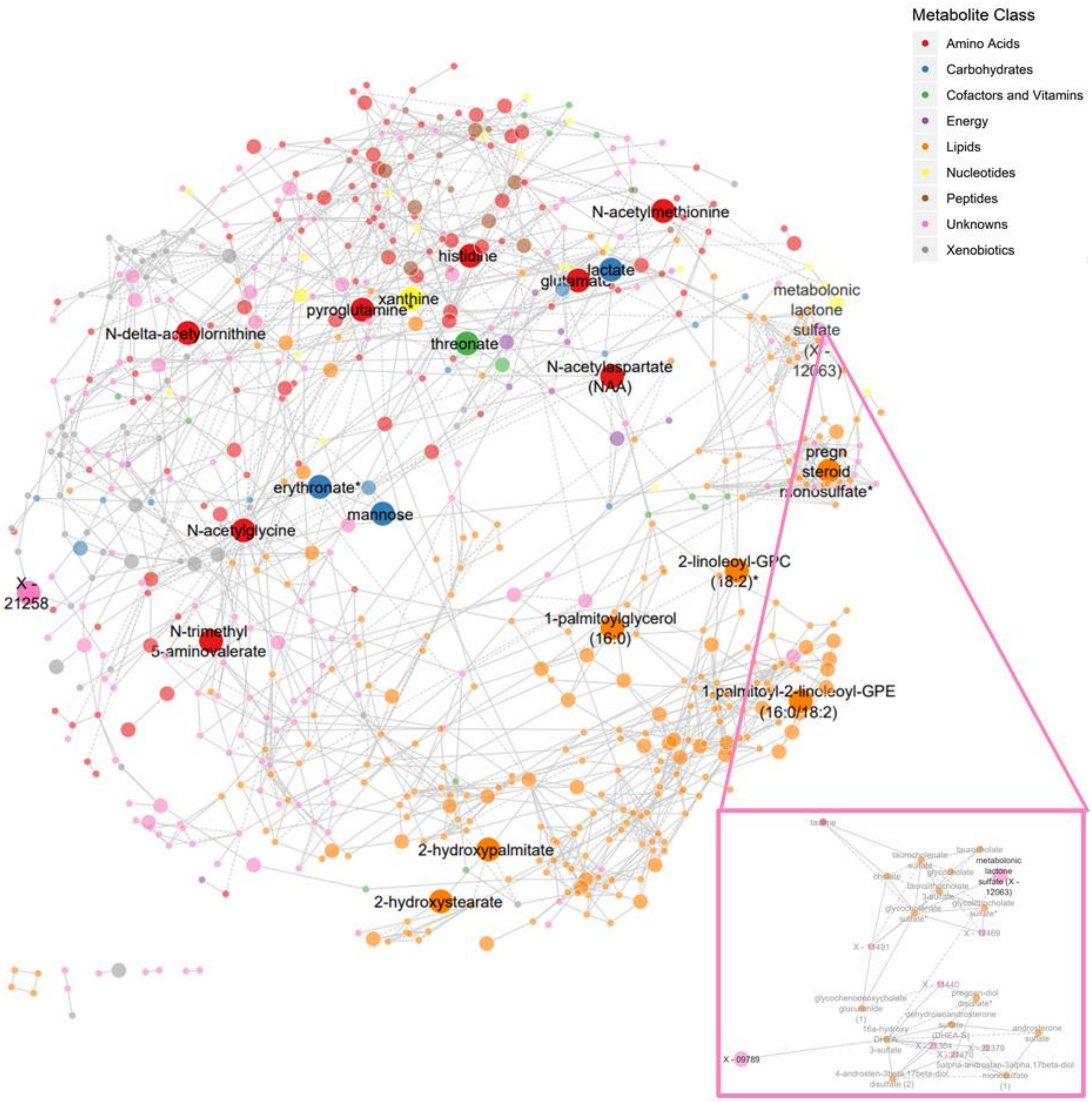
Ch2_Fig4: Sex-combined, cumulative adjustment of candidate mediators selected using BMI and T2D in a Prentice-weighted Cox regression model of the effects of BMI on T2D risk. The multivariate model is adjusted for sex, height and weight at age 20 years with age as the underlying scale.



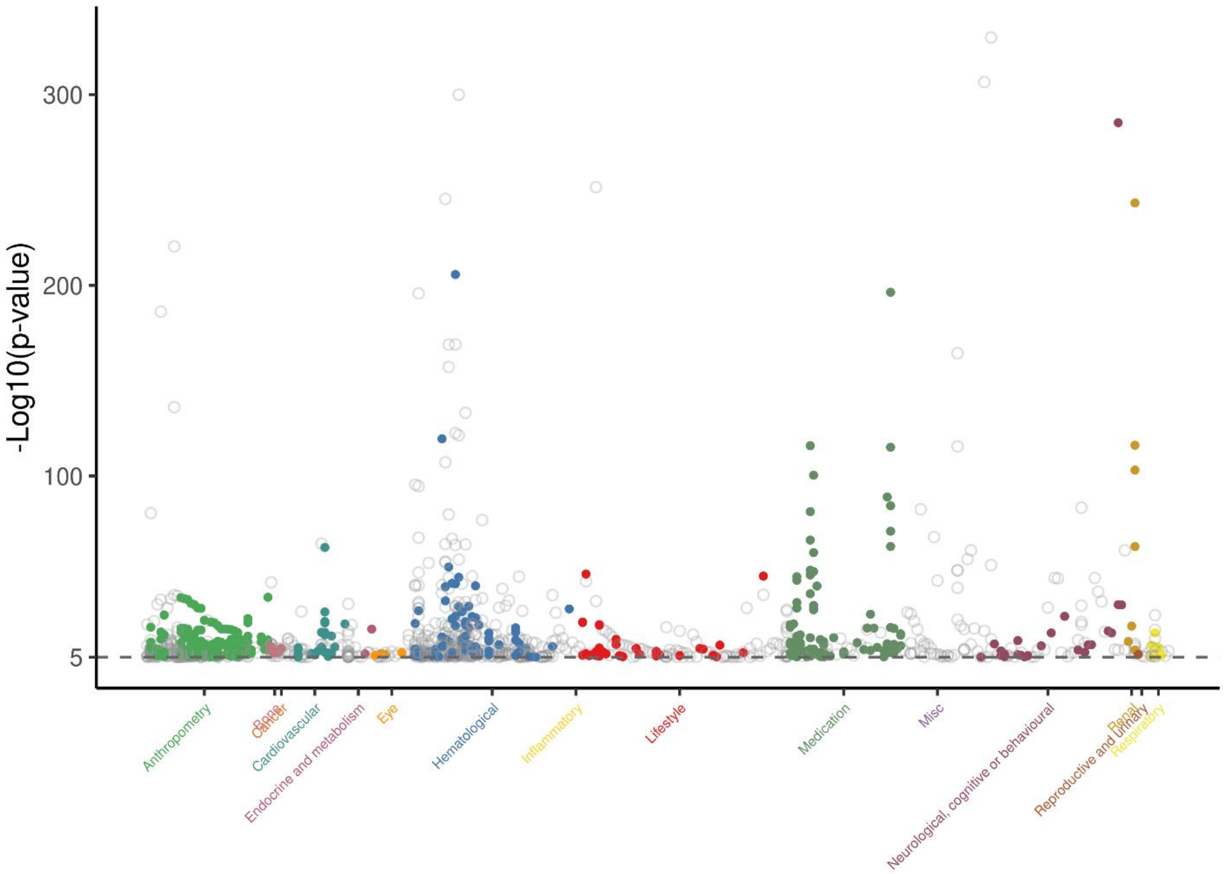
Ch2_Fig5: Sex-stratified, cumulative adjustment of candidate mediators selected using BMI and T2D in a Prentice-weighted Cox regression model of the effects of BMI on T2D risk. The multivariate model was adjusted for height, with age as the underlying timescale.



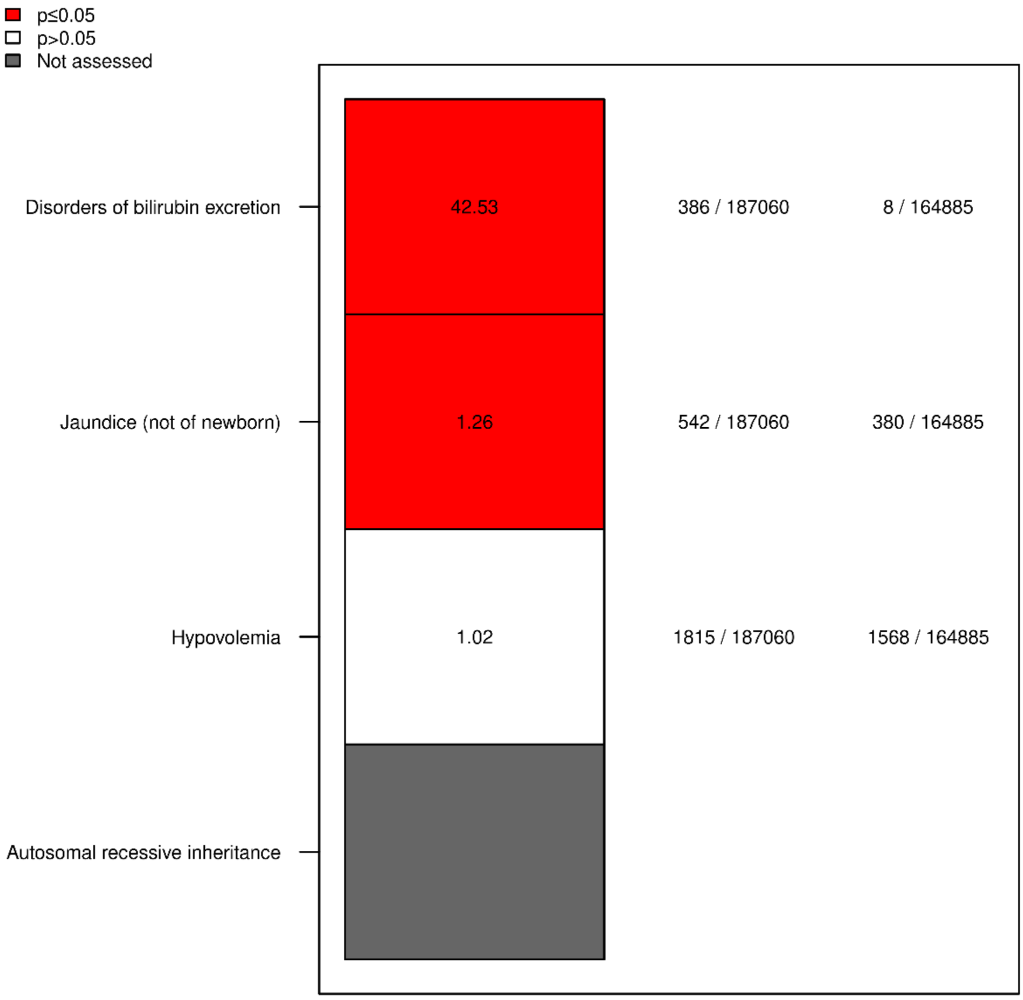
Ch2_Fig6: Associations of genetic scores for body fat percentage, liver fat, hip and waist circumference with candidate mediators. X-12063 is included as a metabolite of interest. Significant associations ($p \leq 2.4 \times 10^{-3}$) are filled in while non-significant associations are empty.



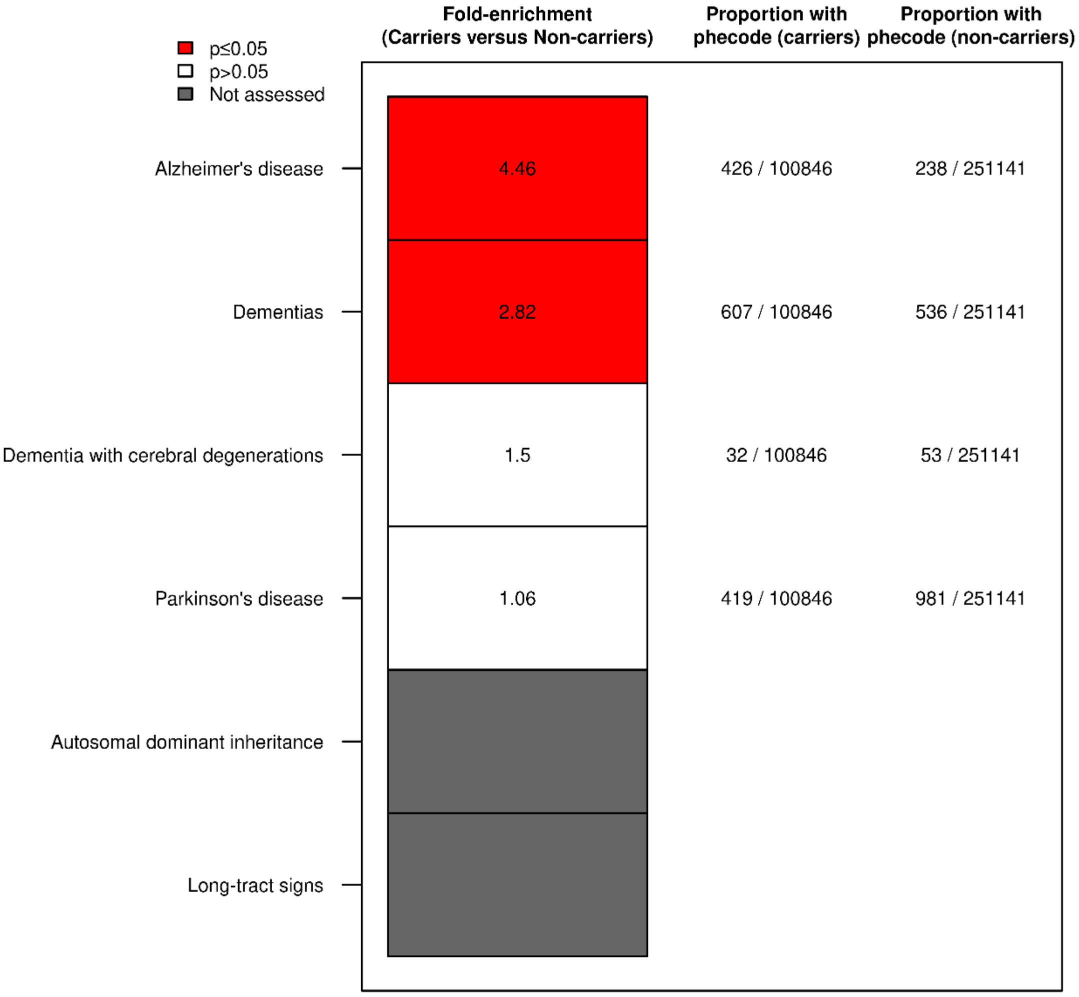
Ch2_Fig7: GGM of the metabolic network. The network comprises of 684 metabolites with abs(pcor) ≥0.1 that represent 1,769 connections. Large circles represent candidate mediators, medium circles those significantly associated with weight gain and T2D (FDR<0.05) and small circles other metabolites. Dashed lines represent negative partial correlations. Inset: first and second order partial correlations for metabolonic lactone sulfate (X-12063).



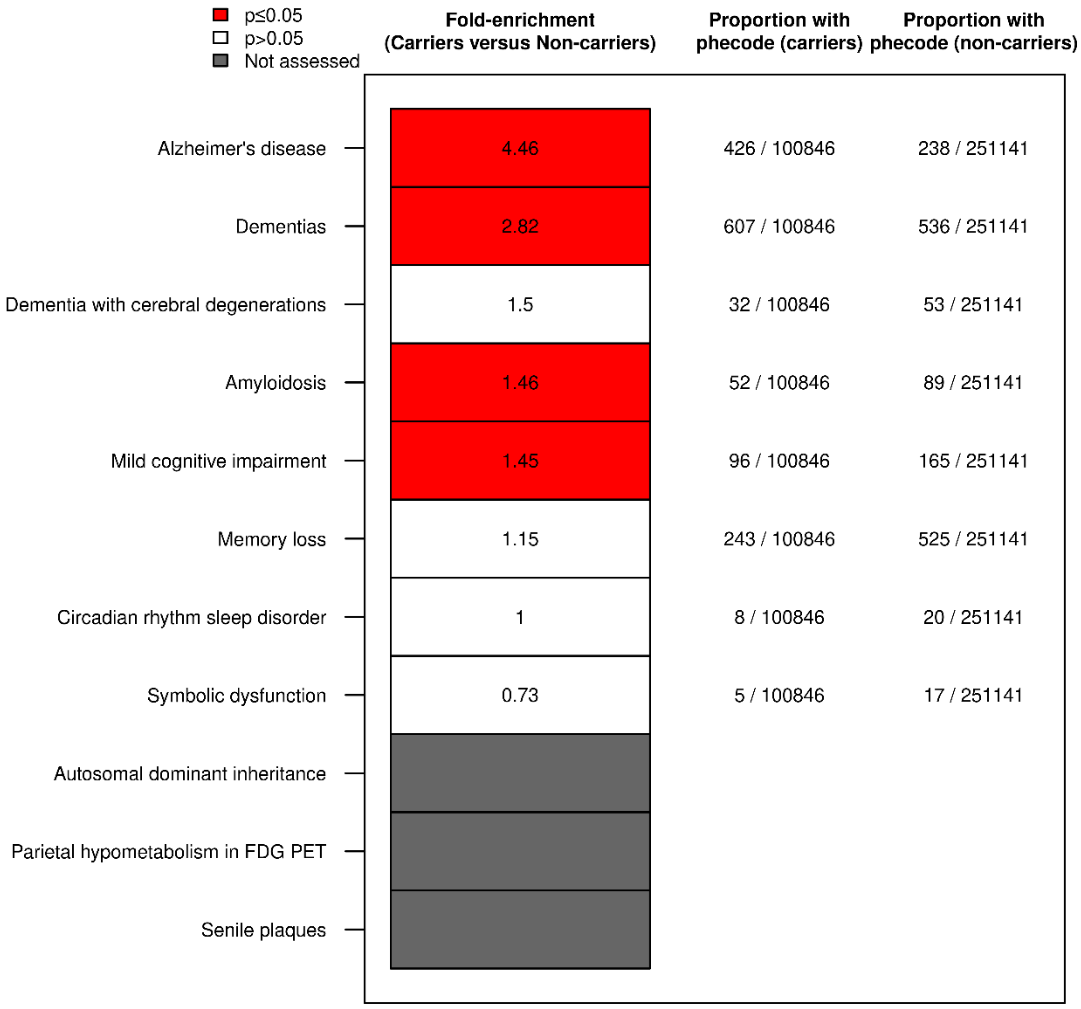
Ch5_Fig1: Summary of IFV associations (p≤1x10^{-5}) with 306 complex traits and phenotypes not classified as clinical outcomes. Filled in, coloured circles represent prioritised phenotypes while empty, grey circles represent associated but de-prioritised phenotypes. Strongest p-value associations for each distinct phenotype are shown.



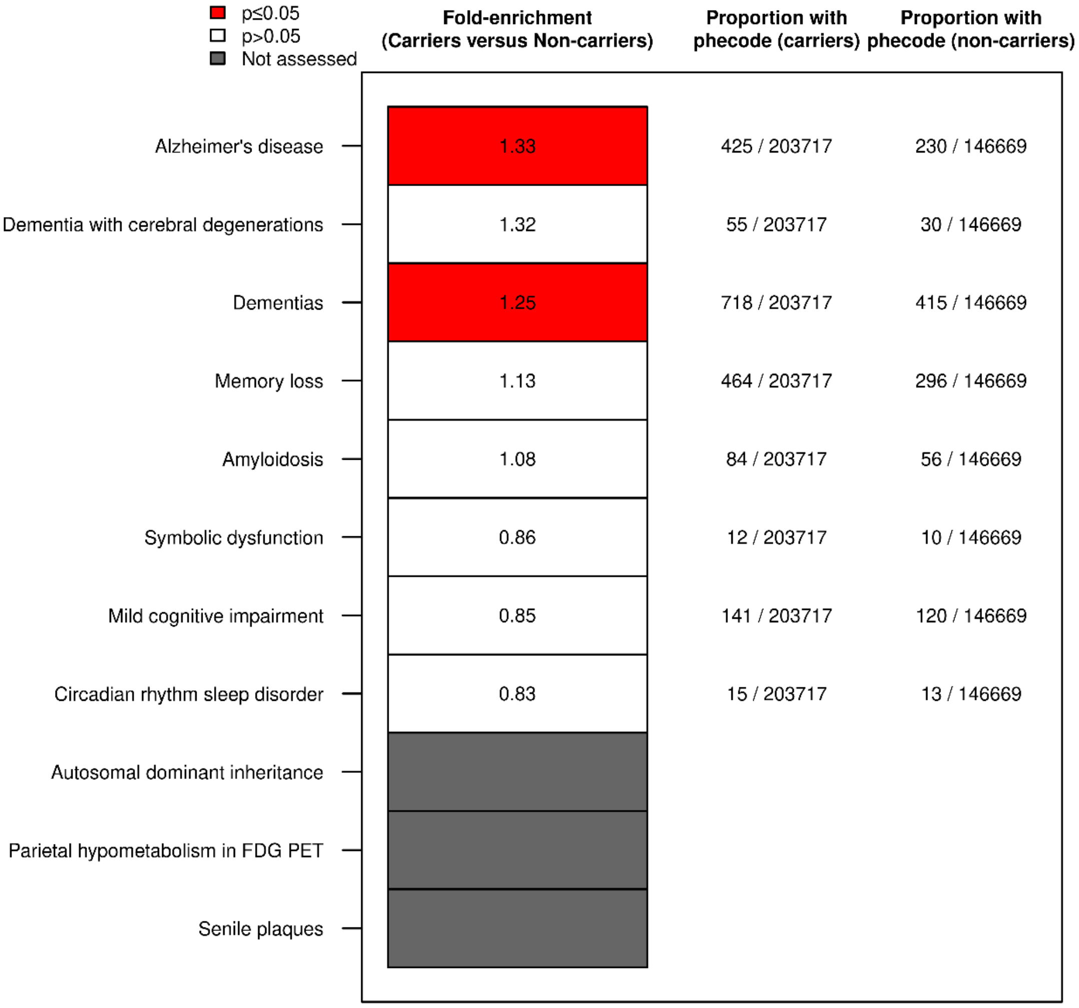
Ch6_Fig1: Enrichment assessment (χ^2 p≤0.05) of phecodes contributing to the PheRS for 'Gilbert syndrome' in heterozygote and homozygote carriers compared to homozygotic non-carriers of the G-allele of the *UGT1A1* variant rs1976391. This condition was only reported in the OMIM database, therefore, frequencies of these phecodes in reported cases of Gilbert syndrome were not available in Orphanet database.



Ch6_Fig2: Enrichment assessment (χ^2 p≤0.05) of phecodes contributing to the PheRS for 'Alzheimer Disease 2' in heterozygote and homozygote carriers compared to homozygotic non-carriers of the C-allele of the *APOE* variant rs429358. This condition was only reported in the OMIM database, therefore, frequencies of these phecodes in reported cases of Alzheimer Disease 2 were not available in Orphanet database.



Ch6_Fig3: Enrichment assessment (χ^2 p≤0.05) of phecodes contributing to the PheRS for 'Alzheimer Disease 4' in heterozygote and homozygote carriers compared to homozygotic non-carriers of the C-allele of the *APOE* variant rs429358. This condition was only reported in the OMIM database, therefore, frequencies of these phecodes in reported cases of Alzheimer Disease 4 were not available in Orphanet database.



Ch6_Fig4: Enrichment assessment (χ^2 p≤0.05) of phecodes contributing to the PheRS for 'Alzheimer Disease 4' in heterozygote and homozygote carriers compared to homozygotic non-carriers of the C-allele of the *APOE* variant rs204474. This condition was only reported in the OMIM database, therefore, frequencies of these phecodes in reported cases of Alzheimer Disease 4 were not available in Orphanet database.

3. Supplementary Tables

Ch2_ST1: Candidate mediators selected using different metabolite sets prioritised for BVS analysis. Metabolite names followed by a '*' are assignments made with high confidence but are not definite. Inf = Infinity

a. Candidate mediators selected from 164 metabolites significantly associated with weight gain (FDR<0.05) and nominally associated with T2D (p<0.05).

Metabolite	Pathway	Class	Posterior probability	Bayes Factor
mannose	Fructose, Mannose and Galactose Metabolism	Carbohydrate	1	Inf
N-acetylglycine	Glycine, Serine and Threonine Metabolism	Amino Acid	0.9984	102336
1-palmitoylglycerol (16:0)	Monoacylglycerol	Lipid	0.9982	90947
2-hydroxystearate	Fatty Acid, Monohydroxy	Lipid	0.9738	6096
glutamate	Glutamate Metabolism	Amino Acid	0.782	588
3-hydroxyoctanoate	Fatty Acid, Monohydroxy	Lipid	0.7054	393
N-acetylaspartate (NAA)	Alanine and Aspartate Metabolism	Amino Acid	0.4756	149
X - 21258			0.2934	68
N-delta-acetylornithine	Urea cycle; Arginine and Proline Metabolism	Amino Acid	0.2696	61
lactate	Glycolysis, Gluconeogenesis, and Pyruvate Metabolism	Carbohydrate	0.2286	49
N-acetylmethionine	Methionine, Cysteine, SAM and Taurine Metabolism	Amino Acid	0.1992	41
cysteine	Methionine, Cysteine, SAM and Taurine Metabolism	Amino Acid	0.1612	32
pregn steroid monosulfate*	Steroid	Lipid	0.1318	25
xanthine	Purine Metabolism, (Hypo)Xanthine/Inosine containing	Nucleotide	0.1266	24
threonate	Ascorbate and Aldarate Metabolism	Cofactors and Vitamins	0.1	18
1-oleoylglycerol (18:1)	Monoacylglycerol	Lipid	0.0992	18
erythronate*	Aminosugar Metabolism	Carbohydrate	0.0926	17
pyroglutamine*	Glutamate Metabolism	Amino Acid	0.085	15
N-trimethyl-5-aminovalerate	Lysine Metabolism	Amino Acid	0.0708	12
1-palmitoyl-2-linoleoyl-GPE (16:0/18:2)	Phospholipid Metabolism	Lipid	0.0664	12
DSGEGDFXAEGGGVR*	Fibrinogen Cleavage Peptide	Peptide	0.0656	12
histidine	Histidine Metabolism	Amino Acid	0.0588	10

b. Candidate mediators selected from 529 metabolites significantly associated with weight gain (FDR<0.05).

Metabolite	Pathway	Class	Posterior probability	Bayes Factor
mannose	Fructose, Mannose and Galactose Metabolism	Carbohydrate	1	Inf
N-acetylglycine	Glycine, Serine and Threonine Metabolism	Amino Acid	1	Inf
1-palmitoylglycerol (16:0)	Monoacylglycerol	Lipid	0.9998	2644471
2-hydroxystearate	Fatty Acid, Monohydroxy	Lipid	0.9934	79623
N-acetylaspartate (NAA)	Alanine and Aspartate Metabolism	Amino Acid	0.9082	5234
malate	TCA Cycle	Energy	0.8772	3779
glutamate	Glutamate Metabolism	Amino Acid	0.7194	1356
3-hydroxyoctanoate	Fatty Acid, Monohydroxy	Lipid	0.6292	898
4-androsten-3beta,17beta-diol monosulfate (1)	Steroid	Lipid	0.2684	194
glycerol 3-phosphate	Glycerolipid Metabolism	Lipid	0.2598	186
dehydroisoandrosterone sulfate (DHEA-S)	Steroid	Lipid	0.1526	95
X - 21258			0.08	46
X - 11315			0.0758	43
N-delta-acetylornithine	Urea cycle; Arginine and Proline Metabolism	Amino Acid	0.0756	43
etiocholanolone glucuronide	Steroid	Lipid	0.0608	34
1-methylhistidine	Histidine Metabolism	Amino Acid	0.0596	34
N-acetylmethionine	Methionine, Cysteine, SAM and Taurine Metabolism	Amino Acid	0.0512	29
xanthine	Purine Metabolism, (Hypo)Xanthine/Inosine containing	Nucleotide	0.0488	27
cysteine	Methionine, Cysteine, SAM and Taurine Metabolism	Amino Acid	0.0472	26
threonate	Ascorbate and Aldarate Metabolism	Cofactors and Vitamins	0.0408	23
erythronate*	Aminosugar Metabolism	Carbohydrate	0.0382	21
X - 17337			0.0332	18
1-palmitoyl-2-linoleoyl-GPE (16:0/18:2)	Phospholipid Metabolism	Lipid	0.0284	15
pyroglutamine*	Glutamate Metabolism	Amino Acid	0.026	14
histidine	Histidine Metabolism	Amino Acid	0.0252	14
X - 24422			0.0246	13
3-ureidopropionate	Pyrimidine Metabolism, Uracil containing	Nucleotide	0.0242	13
1-oleoylglycerol (18:1)	Monoacylglycerol	Lipid	0.0234	13
serine	Glycine, Serine and Threonine Metabolism	Amino Acid	0.0226	12
pregn steroid monosulfate*	Steroid	Lipid	0.0224	12
lactate	Glycolysis, Gluconeogenesis, and Pyruvate Metabolism	Carbohydrate	0.021	11

c. Candidate mediators selected from 122 metabolites significantly associated with weight gain (FDR<0.05) and with T2D (FDR<0.05) after excluding one metabolite from pairs of highly-correlated metabolite pairs (R2>0.8).

Metabolite	Pathway	Class	Posterior probability	Bayes Factor
mannose	Fructose, Mannose and Galactose Metabolism	Carbohydrate	1	Inf
N-acetylglycine	Glycine, Serine and Threonine Metabolism	Amino Acid	0.9984	76128
1-palmitoylglycerol (16:0)	Monoacylglycerol	Lipid	0.9884	10395
2-hydroxystearate	Fatty Acid, Monohydroxy	Lipid	0.8686	806
glutamate	Glutamate Metabolism	Amino Acid	0.8312	601
N-acetylaspartate (NAA)	Alanine and Aspartate Metabolism	Amino Acid	0.4632	105
N-delta-acetylornithine	Urea cycle; Arginine and Proline Metabolism	Amino Acid	0.3508	66
X - 21258			0.3394	63
N-acetylmethionine	Methionine, Cysteine, SAM and Taurine Metabolism	Amino Acid	0.313	56
lactate	Glycolysis, Gluconeogenesis, and Pyruvate Metabolism	Carbohydrate	0.2394	38
pregn steroid monosulfate*	Steroid	Lipid	0.2232	35
pyroglutamine*	Glutamate Metabolism	Amino Acid	0.16	23
erythronate*	Aminosugar Metabolism	Carbohydrate	0.1368	19
2-hydroxypalmitate	Fatty Acid, Monohydroxy	Lipid	0.1318	19
2-linoleoyl-GPC (18:2)*	Lysolipid	Lipid	0.1136	16
1-palmitoyl-2-linoleoyl-GPE (16:0/18:2)	Phospholipid Metabolism	Lipid	0.1088	15
threonate	Ascorbate and Aldarate Metabolism	Cofactors and Vitamins	0.106	14
xanthine	Purine Metabolism, (Hypo)Xanthine/Inosine containing	Nucleotide	0.1044	14
N-trimethyl-5-aminovalerate	Lysine Metabolism	Amino Acid	0.0882	12
histidine	Histidine Metabolism	Amino Acid	0.0758	10

d. Candidate mediators selected from 129 metabolites significantly associated with BMI (FDR<0.05) and with T2D (FDR<0.05).

Metabolite	Pathway	Class	Posterior probability	Bayes Factor
mannose	Fructose, Mannose and Galactose Metabolism	Carbohydrate	1	Inf
N-acetylglycine	Glycine, Serine and Threonine Metabolism	Amino Acid	0.9956	29189
1-palmitoylglycerol (16:0)	Monoacylglycerol	Lipid	0.987	9794
2-hydroxystearate	Fatty Acid, Monohydroxy	Lipid	0.9244	1577
lactate	Glycolysis, Gluconeogenesis, and Pyruvate Metabolism	Carbohydrate	0.6968	296
glutamate	Glutamate Metabolism	Amino Acid	0.6312	221
threonate	Ascorbate and Aldarate Metabolism	Cofactors and Vitamins	0.3922	83
N-acetylaspartate (NAA)	Alanine and Aspartate Metabolism	Amino Acid	0.3798	79
N-acetylmethionine	Methionine, Cysteine, SAM and Taurine Metabolism	Amino Acid	0.3202	61
X - 21258	Unknown	Unknown	0.3014	56
2-linoleoyl-GPC (18:2)*	Lysophospholipid	Lipid	0.235	40
pyroglutamine*	Glutamate Metabolism	Amino Acid	0.1348	20
histidine	Histidine Metabolism	Amino Acid	0.129	19
serine	Glycine, Serine and Threonine Metabolism	Amino Acid	0.1054	15
sphingomyelin (d17:1/16:0, d18:1/15:0, d16:1/17:0)*	Sphingolipid Metabolism	Lipid	0.1004	14
X - 15497	Unknown	Unknown	0.0984	14
1-palmitoyl-2-linoleoyl-GPE (16:0/18:2)	Phosphatidylethanolamine (PE)	Lipid	0.0972	14
X - 13729	Unknown	Unknown	0.0946	13
1-palmitoyl-2-palmitoleoyl-GPC (16:0/16:1)*	Phosphatidylcholine (PC)	Lipid	0.092	13
N-trimethyl 5-aminovalerate	Lysine Metabolism	Amino Acid	0.0876	12
asparagine	Alanine and Aspartate Metabolism	Amino Acid	0.0852	12
1-myristoylglycerol (14:0)	Monoacylglycerol	Lipid	0.0824	12
1-oleoylglycerol (18:1)	Monoacylglycerol	Lipid	0.082	12
1-palmitoyl-2-arachidonoyl-GPE (16:0/20:4)*	Phosphatidylethanolamine (PE)	Lipid	0.0816	11
pregn steroid monosulfate C21H34O5S*	Progestin Steroids	Lipid	0.0762	11
xanthine	Purine Metabolism, (Hypo)Xanthine/Inosine containing	Nucleotide	0.0734	10

Ch2_ST2: Pairwise partial correlations of X-12063 with metabolites that share genetic locus associations. Metabolites with shared genetic locus associations to X-12063 were reported in previous studies[19,20] and partial correlations were calculated using the GGM method. Metabolites with partial correlations that were significant at Bonferroni-corrected threshold ($p \leq 2.1 \times 10^{-7}$) are in bold.

Metabolite	Partial correlation	p-value
androsterone sulfate	0.017	0.13
dehydroisoandrosterone sulfate (DHEA-S)	-0.035	0.0012
taurocholenate sulfate	0.0069	0.51
epiandrosterone sulfate	0.024	0.027
tetradecanedioate	0.020	0.060
hexadecanedioate	-0.012	0.26
5alpha-androstan-3alpha,17beta-diol monosulfate (1)	**0.072**	**4.53×10^{-10}**
5alpha-androstan-3beta,17beta-diol disulfate	0.035	0.0013
4-androsten-3alpha,17alpha-diol monosulfate (3)	0.0020	0.85
16a-hydroxy DHEA 3-sulfate	**-0.12**	**4.61×10^{-28}**

Ch3_ST1: IFVs reported to cause the corresponding IEM and common enough to be detected in this study.

Variant	Gene	IEM	Minor allele frequency	Associated metabolite(s) with known structural identity
rs28941785	*CTH*	Cystathionase deficiency	0.0093	cystathionine
rs77931234	*ACADM*	Medium-chain acyl-CoA dehydrogenase deficiency	0.0078	hexanoylcarnitine, octanoylcarnitine, cis-4-decenoyl carnitine, decanoylcarnitine, hexanoylglycine
rs72549326	*FMO3*	Trimethylaminuria	0.0033	S-methylcysteine
rs121912698	*ACY1*	Aminoacylase 1 deficiency	0.0044	acetylglutamate, *N*-acetylthreonine, *N*-acetylleucine, *N*-acetylisoleucine, *N*-acetylglycine, *N*-acetylhistidine
rs121434346	*SLC6A19*	Hartnup disorder	0.005	methionine sulfone, 3-methoxytyrosine, *N*-delta-acetylornithine, 1-methylhistidine
rs77010315	*SLC36A2*	Iminoglycinuria	0.011	pyroglutamine*, carnitine, acetylcarnitine
rs5030858	*PAH*	Phenylketonuria	0.0012	phenylalanine, gamma-glutamylphenylalanine
rs5030861	*PAH*	Phenylketonuria	0.0017	phenylalanine, gamma-glutamylphenylalanine, phenylpyruvate
rs75193786	*PAH*	Phenylketonuria	0.0009	phenylalanine, gamma-glutamylphenylalanine, phenylpyruvate
rs1800556	*ACADS*	Short-chain acyl-CoA dehydrogenase deficiency	0.052	ethylmalonate, butyrylcarnitine
rs113298164	*LIPC*	Hepatic lipase deficiency	0.002	1-palmitoyl-2-docosahexaenoyl-GPE (16:0/22:6), 1-palmitoyl-2-arachidonoyl-GPE (16:0/20:4), 1-stearoyl-2-arachidonoyl-GPE (18:0/20:4), 1-stearoyl-2-docosahexaenoyl-GPE (18:0/22:6)
rs148686510	*ACSF3*	Combined malonic and methylmalonic aciduria	0.0082	ethylmalonate
rs534908188	*ACADVL*	Very long-chain acyl-CoA dehydrogenase deficiency	0.0019	myristoleoylcarnitine, laurylcarnitine
rs117935223	*PRODH*	Type 1 hyperprolinemia	0.079	proline
rs143493067	*UPB1*	Beta-ureidopropionase deficiency	0.002	3-ureidopropionate

***metabolite assignment was made with high confidence but is not definite**

Ch5_ST1: Summary of phenotypes and the cohorts included in the discovery efforts of GWAS studies used in colocalisation analysis.

Phenotype$^{\text{ref of GWAS study/studies used}}$	Cohorts included in GWAS at discovery stage$^{\text{refs}}$
Alzheimer's disease[302]	Psychiatric Genomics Consortium[303], Alzheimer's Disease Sequencing Project[304], International Genomics of Alzheimer's Project[305], UK Biobank[67]
Body mass index[115]	GIANT consortium[80], UK Biobank[67]
Breast cancer[306,307]	African American Breast Cancer Consortium[308], Triple-Negative Breast Cancer Consortium[308], NCI Breast and Prostate Cancer Cohort Consortium[309], Breast Cancer Association Consortium[310], Non-Hispanic White women discovery set[311]
Chronic kidney disease[195,312]	See Pattaro *et al*. (2016)[312] and Wuttke *et al.* (2019)[195] for full description of 49 and 85 cohorts included in meta-analysis, respectively
Cognitive performance[200]	COGENT consortium[313], UK Biobank[67]
Glomerular filtration rate[195,314]	UK Household Longitudinal Study[314]; see Wuttke *et al.* (2019)[195] for full description of 85 cohorts included in meta-analysis
Granulocyte count[223]	UK Biobank[67], UK BiLEVE study[315], INTERVAL[163]
Granulocyte percentage of myeloid white cells[223]	UK Biobank[67], UK BiLEVE study[315], INTERVAL[163]
Hippocampal volume[316]	Alzheimer's Disease Neuroimaging Initiative[317]
Intelligence[233]	COGENT consortium[313], UK Biobank[67], Rotterdam study[318], Generation R study[319], Swedish Twin Registry[320], Spit for Science[321], High-IQ/Health and Retirement study[322], Twins Early Development study[323], Danish Twin Registry[324], IMAGEN[325], Brisbane Longitudinal Twin Study[326], Netherlands study of Cognition, Environment and Genes[327], Genes for Good[233], Swedish Twin Studies of Aging[328]
Ischaemic stroke[329]	MEGASTROKE consortium[329]
Mean corpuscular haemoglobin[223]	UK Biobank[67], UK BiLEVE study[315], INTERVAL[163]
Mean corpuscular volume[223]	UK Biobank[67], UK BiLEVE study[315], INTERVAL[163]
Mean platelet volume[223]	UK Biobank[67], UK BiLEVE study[315], INTERVAL[163]
Multiple sclerosis[330]	International Multiple Sclerosis Genetics Consortium[330]
Myeloid white cell count[223]	UK Biobank[67], UK BiLEVE study[315], INTERVAL[163]
Myocardial infarction[331]	CARDIoGRAMplusC4D Consortium[331]
Neutrophil percentage of white cells[223]	UK Biobank[67], UK BiLEVE study[315], INTERVAL[163]
Optic disc area[332]	International Glaucoma Genetics Consortium[332]
Platelet count[223]	UK Biobank[67], UK BiLEVE study[315], INTERVAL[163]
Platelet distribution width[223]	UK Biobank[67], UK BiLEVE study[315], INTERVAL[163]
Posterior cortical atrophy[333]	Patients who were diagnosed to have posterior cortical atrophy by their doctors, had multidomain cognitive impairment fulfilling criteria for Alzheimer's disease and dementia and who fulfilled at least one criterion in previously published posterior cortical atrophy diagnosis guidelines[333]
Pulse pressure[334]	International Consortium for Blood Pressure[335], UK Biobank[67]
Pulse rate[336]	CARDIoGRAMplusC4D Consortium[331], International Stroke Genetics Consortium[329], UK Biobank[67]
QRS duration[337]	CHARGE EKG exome-chip consortium[338]
Red blood cell count[223]	UK Biobank[67], UK BiLEVE study[315], INTERVAL[163]
Red cell distribution width[223]	UK Biobank[67], UK BiLEVE study[315], INTERVAL[163]
Reticulocyte count[223]	UK Biobank[67], UK BiLEVE study[315], INTERVAL[163]
Reticulocyte fraction of red cells[223]	UK Biobank[67], UK BiLEVE study[315], INTERVAL[163]

Ch5_ST2: Additional 24 IFVs included in phenome-wide assessment, as identified in the literature. Chr = Chromosome, EAF = Effect allele frequency.

Source	IEM gene	IEM	IFV	Chr	Position	EA/OA	EAF	IEM-associated metabolite
pmid23824729	*MTHFR*	Homocystinuria due to methylene tetrahydrofolate reductase deficiency	rs1801133	1	11856378	A/G	0.34	homo-cysteine
biorxiv_660506 v1	*ALPL*	Hypophosphatasia	rs3820293	1	21806621	G/T	0.45	phosphate
biorxiv_660506 v1	*LDLRAP1*	Homozygous familial hypercholesterolemia	rs586178	1	25747230	C/G	0.44	cholesterol
pmid23824729	*MMACHC*	Methylmalonic acidemia with homocystinuria, type cblC	rs4660306	1	45978675	T/A/C	0.33	homo-cysteine
pmid26025379	*SLC30A10*	Cirrhosis-dystonia-polycythemia-hypermanganesemia syndrome	rs1776029	1	220080028	A/G	0.18	manganese
pmid23824729	*MTR*	Methylcobalamin deficiency type cblG	rs2275565	1	237048676	T/G	0.21	homo-cysteine
biorxiv_660506 v1	*ABCB11*	Benign recurrent intrahepatic cholestasis type 2 | Progressive familial intrahepatic cholestasis type 2	rs2287623	2	169830155	G/A	0.40	cholesterol
pmid26068415	*AGPS*	Rhizomelic chondrodysplasia punctata type 3	rs7582179	2	178370631	A/G	0.17	PC ae C44:5
pmid31959995	*HIBCH*	Neurodegeneration due to 3-hydroxyisobutyryl-CoA hydrolase deficiency	rs291468	2	191188119	A/G	0.4	methyl-malonate
pmid26352407	*AGXT*	Primary hyperoxaluria type 1	rs6748734	2	241837452	A/G	0.29	alpha-keto isovalerate
pmid28263315	*D2HGDH*	D-2-hydroxyglutaric aciduria	rs6707874	2	242709363	A/G	0.64	2-hydroxy glutaric acid
pmid25352340	*TF*	Congenital atransferrinemia	rs8177240	3	133477701	T/G	0.67	transferrin
biorxiv_660506 v1	*SLC2A2*	Glycogen storage disease due to GLUT2 deficiency	rs5400	3	170732300	A/G	0.12	glucose
pmid23754956	*MMAA*	Vitamin B12-responsive methylmalonic acidemia type cblA	rs2270655	4	146576418	G/C	0.94	vitamin B12
pmid30275531	*LCAT*	Familial LCAT deficiency | Fish-eye disease | LCAT deficiency	rs56070533	16	67942320	A/G	0.15	HDL cholesterol
pmid19185284	*BCO1*	Hereditary hypercarotenemia and vitamin A deficiency	rs6564851	16	81264597	G/T	0.64	carotenoid, tocopherol beta-carotene
pmid27005778	*SLC13A5*	Amelocerebrohypohidrotic syndrome | Pyridoxine-dependent epilepsy | Undetermined early-onset epileptic encephalopathy	rs172642	17	6595398	C/A	0.48	citrate
pmid23754956	*CD320*	Methylmalonic aciduria due to transcobalamin receptor defect	rs2336573	19	8367709	T/C	0.031	vitamin B12

pmid26401656	*HLCS*	Holocarboxylase synthetase deficiency	rs1571700	21	38336834	A/G	0.41	2-hydroxy isovaleryl carnitine
pmid24651765	*CBS*	Classic homocystinuria	rs234714	21	44488033	T/C	0.2	homo-cysteine post-methionine load test
pmid27005778	*SLC25A1*	D,L-2-hydroxyglutaric aciduria	rs2040771	22	19161935	T/C	0.48	citrate
pmid23754956	*TCN2*	Transcobalamin deficiency	rs1131603	22	31018975	C/T	0.055	vitamin B12
pmid28588231	*SLC5A1*	Glucose-galactose malabsorption	rs117086479	22	32389342	G/A	0.06	1,5-anhydro glucitol ea
pmid29403010	*SLC6A8*	X-linked creatine transporter deficiency	rs5987107	23	152875584	A/G	0.33	creatinine

Ch5_ST3: Summary of phenotypes that were associated with loci containing IFVs and that were tested for a shared genetic signal using colocalisation.

Gene (Locus chr:start-end)	IFV/s (EA/OA; EAF)	Trait	Category	Evidence sentence	Evidence link or Pubmed ID
ABCA1 (9:107264123-107915739)	rs2575876 (A/G; 0.25)	Hip circumference	Anthropometry	ABCA1 deficiency is associated with hypercholesterolemia, which can increase lipid deposits in adipose tissue and increase body size	pmid195100, pmid30881070
		Waist-to-hip ratio	Anthropometry		
		Medication use (HMG CoA reductase inhibitors)	Medication	ABCA1 deficiency is associated with hypercholesterolemia, which could require treatment	pmid195100, pmid30881070
		Cholesterol lowering medication | medication for cholesterol, blood pressure, diabetes, or take exogenous hormones	Medication		
		Treatment with atorvastatin	Medication	ABCA1 deficiency is associated with hypercholesterolemia, which may be treated with statins	pmid195100, pmid30881070
		Hypercholesterolemia	Cardiovascular	ABCA1 deficiency is associated with hypercholesterolemia	pmid195100, pmid30881070
		Self-reported high cholesterol	Endocrine and metabolism		
ACADS (12:119943590-122716919)	rs2014355 (C/T; 0.25)	Time to complete round	Neurological, cognitive or behavioural	ACADS deficiency is associated with developmental delay, which may affect educational attainment and 'time to complete round' of a cognitive function experiment	Roe, C. R., Ding, J. Mitochondrial fatty acid oxidation disorders.In: Scriver, C. R.; Beaudet, A. L.; Sly, W. S.; Valle, D. (eds.) : The Metabolic and Molecular Bases of Inherited Disease. Vol. II. (7th ed.) New York: McGraw-Hill (pub.) 2001. Pp. 2297-2326.
ACADS (12:119943590-122716919)	rs2014355 (C/T; 0.25)	Red cell distribution width	Hematological	ACADS deficiency is associated with exercise intolerance, which could plausibly be explained by a lack of red blood cells to carry haemoglobin, or a reduced capacity of red blood cells to carry haemoglobin	pmid30477112, pmid3891376, pmid10647532
		Mean sphered cell volume	Hematological		
		Mean reticulocyte volume	Hematological		
		Reticulocyte count	Hematological		
		Mean corpuscular volume	Hematological		
		Mean corpuscular haemoglobin	Hematological		
		Red blood cell count	Hematological		
ACSF3 (16:88668096-90424092)	rs36099289 (A/C; 0.13)	Leg fat-free mass	Anthropometry	ACSF3 deficiency is associated with muscular-axial hypotonia and feeding difficulties, which may affect body fat and muscle composition	pmid21841779
	rs72817435 (A/G; 0.04)	Arm fat mass	Anthropometry		
		Hip circumference	Anthropometry		
		Weight	Anthropometry		
		Trunk fat mass	Anthropometry		
		Leg fat mass	Anthropometry		
		Body fat percentage	Anthropometry		
		Arm fat percentage	Anthropometry		
		Body mass index	Anthropometry		
		Trunk fat percentage	Anthropometry		
		Whole body fat mass	Anthropometry		

ADSL (22:40507228-41007228)	rs8135371 (C/A; 0.15)	Comparative body size at age 10	Anthropometry	ADSL deficiency is associated with developmental delay and may also reduce growth	pmid18830228
ALPL (1:21535330-22206759)	rs1531829 (A/G; 0.35)	Heel quantitative ultrasound index (QUI), direct entry	Bone	Increased alkaline phosphatase levels are linked to osteomalacia, a disease characterised by softening of the bone. Osteomalacia/rickets are also accompanied by vitamin D and calcium deficiency	pmid26823351
		Heel bone mineral density	Bone		
		Heel bone mineral density (BMD) T-score, automated	Bone		
		Heel broadband ultrasound attenuation	Bone		
APOA5/APOA4/ APOA1/APOC3 (11:116274569-117470429)	rs530885291 (A/G; 0.0026)	Mean corpuscular haemoglobin	Hematological	APOA1 deficiency is associated with visceral amyloidosis, which is also accompanied by coagulation factor abnormalities	https://omim.org/entry/105200, pmid16731289
		Platelet distribution width	Hematological		
		Mean sphered cell volume	Hematological		
		Red cell distribution width	Hematological	APOA1 deficiency is associated with visceral amyloidosis, which is also accompanied by red blood cell abnormalities	https://omim.org/entry/105200, pmid16731289
		Cholesterol lowering medication | medication for cholesterol, blood pressure, diabetes, or take exogenous hormones	Medication	Apolipoprotein gene deficiencies are associated with hypertriglyceridemia	pmid610428, pmid29396262
		Treatment with ferrous salt product	Medication	Ferrous salt product is used to treat iron deficiency anemia; this may reflect abnormalities in red blood cell production	pmid16731289
	rs964184 (C/G; 0.87)	Medication for pain relief, constipation, heartburn: aspirin	Medication	APOA1 deficiency is associated with coronary artery disease, which may induce heartburn and require aspirin	pmid3089658, pmid29396262
		Treatment with aspirin	Medication		
		Coronary artery disease	Cardiovascular	APOA1 deficiency is associated with coronary artery disease	pmid3089658, pmid29396262
		Medication use (agents acting on the renin-angiotensin system)	Medication	APOA1 deficiency is associated with hypercholesterolemia and may cause hypertension, which can be treated with renin-angiotensin targeting drugs	pmid29396262
		Medication use (HMG CoA reductase inhibitors)	Medication	APOA1 deficiency is associated with hypercholesterolemia, which can be treated with cholesterol lowering medications	pmid29396262
		Medication use (antithrombotic agents)	Medication		
		Treatment with simvastatin	Medication		
		Fenofibrate | treatment/medication code	Medication		
		Cholesterol lowering medication | medication for cholesterol, blood pressure, diabetes, or take exogenous hormones	Medication		
		Rosuvastatin | treatment/medication code	Medication		

		Lipitor 10mg tablet | treatment/medication code	Medication		
		Treatment with atorvastatin	Medication		
		Treatment with ezetimibe	Medication		
		Treatment with bezafibrate	Medication		
		Cheese intake	Lifestyle	APOA1 deficiency is associated with hypercholesterolemia, which may affect intake of high fat foods	pmid3089658, pmid29396262
		Metabolic syndrome	Endocrine and metabolism	APOA1 deficiency is associated with hypercholesterolemia, which may lead to metabolic syndrome	pmid29396262
		Self-reported high cholesterol	Endocrine and metabolism	APOA1 deficiency is associated with hypercholesterolemia	pmid29396262
		Hypercholesterolemia	Cardiovascular		
		Reticulocyte fraction of red cells	Hematological	APOA1 deficiency is associated with visceral amyloidosis, which is also accompanied by red blood cell and coagulation factor abnormalities	https://omim.org/entry/105200, pmid16731289
		Platelet distribution width	Hematological		
		Mean platelet volume	Hematological		
		Platelet crit	Hematological		
		Reticulocyte percentage	Hematological		
		Reticulocyte count	Hematological		
		Red cell distribution width	Hematological		
		Mean reticulocyte volume	Hematological		
		Platelet count	Hematological		
APOB (2:20893329-21546682)	rs934197 (A/G; 0.33)	Coronary artery disease	Cardiovascular	APOB deficiency is known to cause premature hyperlipidemia and cardiovascular disease	pmid19200547, pmid7883971
		Treatment with ezetimibe	Medication	APOB deficiency is known to cause premature hyperlipidemia, which can be treated by statins or other cholesterol lowering medications	pmid19200547, pmid18702965
		Medication use (HMG CoA reductase inhibitors)	Medication		
		Lipitor 10mg tablet | treatment/medication code	Medication		
		Treatment with atorvastatin	Medication		
		Rosuvastatin | treatment/medication code	Medication		
		Treatment with simvastatin	Medication		
		Cholesterol lowering medication | medication for cholesterol, blood pressure, diabetes, or take exogenous hormones	Medication		
		Yes, because of other reasons | major dietary changes in the last 5 years	Lifestyle	APOB deficiency is known to cause premature hyperlipidemia, which may affect the consumption of high fat foods	pmid19200547
		Hypercholesterolemia	Cardiovascular	APOB deficiency is known to cause premature hyperlipidemia	pmid19200547
		Self-reported high cholesterol	Endocrine and metabolism		

APOE/APOC/ APOC2/APOC4 (19:45081103-45922478)	rs204474 (T/C; 0.65)	Alzheimer's disease	Neurological, cognitive or behavioural	APOE deficiency is associated with dementia, which may link to Alzheimer's disease	pmid4194379
		Leg predicted mass	Anthropometry	APOE deficiency is associated with hyperlipidemia, which may increase body fat composition and size	pmid27481046
		Weight	Anthropometry		
		Hip circumference	Anthropometry		
		Arm predicted mass	Anthropometry		
	rs429358 (C/T; 0.15)	Pulse rate	Hematological	APOE deficiency can lead to elevated cholesterol levels, which may cause hypertension	pmid27481046
		Diastolic blood pressure	Hematological		
		Hippocampal volume	Neurological, cognitive or behavioural	APOE deficiency is associated with dementia and has been associated with hippocampal atrophy	pmid15956166
		Time to complete round	Neurological, cognitive or behavioural	APOE deficiency is associated with dementia, which may affect cognitive performance	pmid4194379
		Posterior cortical atrophy	Neurological, cognitive or behavioural	APOE deficiency is associated with dementia, which may link to Alzheimer's disease and posterior cortical atrophy	pmid4194379
		Dementias	Neurological, cognitive or behavioural	APOE deficiency is associated with dementia, which may link to Alzheimer's disease	pmid4194379
		Alzheimer's disease	Neurological, cognitive or behavioural		
		Cause of death: unspecified dementia	Neurological, cognitive or behavioural	APOE deficiency is associated with dementia	pmid4194379
		Year ended full time education	Neurological, cognitive or behavioural	APOE deficiency is associated with dementia; earlier-onset effects may impact educational ability	pmid4194379
		Medication for pain relief, constipation, heartburn: aspirin	Medication	APOE deficiency is associated with elevated cholesterol levels and coronary artery disease, which may cause hypertension and heartburn and require aspirin	pmid27481046
		Treatment with aspirin	Medication		
		Salt added to food	Lifestyle	APOE deficiency is associated with elevated cholesterol levels that are responsive to cholesterol intake from diet; this may affect intake of high fat foods	pmid7868975
		Butter/spreadable butter | spread type	Lifestyle	APOE deficiency is associated with elevated cholesterol levels that are responsive to cholesterol intake from diet; this may affect intake of high fat foods	pmid7868975
		Cheese intake	Lifestyle	APOE deficiency is associated with elevated cholesterol levels that are responsive to cholesterol intake from diet; this may affect intake of high fat foods	pmid7868975
		Other type of spread/margarine | spread type	Lifestyle		
		Non-butter spread type details: Flora Pro-Active or Benecol	Lifestyle		
		Oily fish intake	Lifestyle		
		Rosuvastatin | treatment/medication code	Medication	APOE deficiency is associated with elevated cholesterol levels, which may be treated by cholesterol lowering medications	pmid27481046
		Treatment with atorvastatin	Medication		
		Treatment with simvastatin	Medication		

		Treatment with ezetrol 10mg tablet	Medication		
		Cholesterol lowering medication | medication for cholesterol, blood pressure, diabetes, or take exogenous hormones	Medication		
		Treatment with ezetimibe	Medication		
		Lipitor 10mg tablet | treatment/medication code	Medication		
		Type 2 diabetes	Endocrine and metabolism	APOE deficiency is associated with elevated cholesterol levels, which may increase risk of type 2 diabetes	pmid27481046
		Atherosclerosis	Cardiovascular	APOE deficiency is associated with elevated cholesterol levels, which may lead to adverse cardiovascular outcomes	pmid27481046
		Major coronary heart disease event	Cardiovascular		
		Angina pectoris	Cardiovascular		
		Chronic ischaemic heart disease	Cardiovascular		
		Ischaemic heart disease	Cardiovascular		
		Myocardial infarction	Cardiovascular		
		Coronary artery disease	Cardiovascular		
		Hypercholesterolemia	Cardiovascular		
		Body fat percentage	Anthropometry	APOE deficiency is associated with elevated cholesterol levels, which may lead to obesity and increased fat mass	pmid6261329
		Trunk predicted mass	Anthropometry		
		Trunk fat percentage	Anthropometry		
		Weight change compared with 1 year ago	Anthropometry		
		Leg fat percentage	Anthropometry		
		Weight	Anthropometry		
		Body mass index	Anthropometry		
		Arm fat percentage	Anthropometry		
		Whole body fat mass	Anthropometry		
		Leg fat mass	Anthropometry		
		Arm fat mass	Anthropometry		
		Waist circumference	Anthropometry		
		Arm predicted mass	Anthropometry		
		Leg predicted mass	Anthropometry		
		Waist-to-hip ratio	Anthropometry		
		Hip circumference	Anthropometry		
		Trunk fat mass	Anthropometry		
		Self-reported high cholesterol	Endocrine and metabolism	APOE deficiency is associated with elevated cholesterol levels	pmid27481046
		Metabolic syndrome	Endocrine and metabolism	APOE deficiency is associated with hyperlipidemia, which increases risk of metabolic syndrome	pmid6261329
		Monocyte count	Hematological	APOE deficiency is associated with splenomegaly, which can reduce the number of platelets and white blood cells in the bloodstream	pmid4242937
		Sum basophil neutrophil counts	Hematological		
		Mean platelet volume	Hematological		
		Platelet crit	Hematological		
		Platelet count	Hematological		
	rs5112 (G/C; 0.54)	Treatment with atorvastatin	Medication	APOE deficiency is associated with elevated cholesterol levels, which can be treated with cholesterol lowering medications	pmid27481046
		Cholesterol lowering medication | medication for cholesterol, blood pressure, diabetes, or take exogenous hormones	Medication		
		Arm predicted mass	Anthropometry	APOE deficiency is associated with elevated cholesterol	pmid6261329
		Trunk predicted mass	Anthropometry		

		Leg predicted mass	Anthropometry	levels, which may lead to obesity and increased fat mass	
		Hypercholesterolemia	Cardiovascular	APOE deficiency is associated with elevated cholesterol levels	pmid27481046
		Self-reported high cholesterol	Endocrine and metabolism		
		Granulocyte count	Hematological	APOE deficiency is associated with splenomegaly, which can reduce the number of white blood cells in the bloodstream	pmid4242937
		Monocyte count	Hematological		
		White blood cell count	Hematological		
		Sum neutrophil eosinophil counts	Hematological		
		Myeloid white cell count	Hematological		
	rs7412 (T/C; 0.081)	Pulse rate	Hematological	APOE deficiency can lead to elevated cholesterol levels, which may cause hypertension	pmid27481046
		Pulse pressure	Hematological		
		Systolic blood pressure	Hematological		
		Alzheimer's disease	Neurological, cognitive or behavioural	APOE deficiency is associated with dementia, which may link to Alzheimer's disease	pmid4194379
		Blood pressure medication | medication for cholesterol, blood pressure, diabetes, or take exogenous hormones	Medication	APOE deficiency is associated with elevated cholesterol levels and coronary artery disease, which may cause hypertension and heartburn and require aspirin	pmid27481046
		Medication for pain relief, constipation, heartburn: aspirin	Medication		
		Treatment with aspirin	Medication		
		Full cream | milk type used	Lifestyle	APOE deficiency is associated with elevated cholesterol levels that are responsive to cholesterol intake from diet; this may affect intake of high fat foods	pmid7868975
		White | bread type	Lifestyle		
		Other type of spread/margarine | spread type	Lifestyle		
		Yes, because of other reasons | major dietary changes in the last 5 years	Lifestyle		
		Butter/spreadable butter | spread type	Lifestyle		
		Non-butter spread type details: Flora Pro-Active or Benecol	Lifestyle		
		Treatment with ezetrol 10mg tablet	Medication	APOE deficiency is associated with elevated cholesterol levels, which can be treated with cholesterol lowering medications	pmid27481046
		Treatment with simvastatin	Medication		
		Lipitor 10mg tablet | treatment/medication code	Medication		
		Cholesterol lowering medication | medication for cholesterol, blood pressure, diabetes, or take exogenous hormones	Medication		
		Treatment with atorvastatin	Medication		
		Medication use (HMG CoA reductase inhibitors)	Medication		
		Rosuvastatin | treatment/medication code	Medication		
		Treatment with ezetimibe	Medication		

		Medication use (agents acting on the renin-angiotensin system)	Medication	APOE deficiency is associated with elevated cholesterol levels, which can lead to hypertension that is treated with aspirin/renin-angiotensin targeting drugs. APOE deficiency is associated with elevated cholesterol levels, which may lead to adverse cardiovascular outcomes	pmid27481046 pmid27481046
		Medication use (antithrombotic agents)	Medication		
		Medication use (calcium channel blockers)	Medication		
		Ischaemic heart disease	Cardiovascular		
		Myocardial infarction	Cardiovascular		
		Hypercholesterolemia	Cardiovascular		
		Coronary artery disease	Cardiovascular		
		Angina pectoris	Cardiovascular		
		Hypertension	Cardiovascular		
		Diagnoses - secondary ICD10: Z95.5 Presence of coronary angioplasty implant and graft	Cardiovascular		
		Chronic ischaemic heart disease	Cardiovascular		
		Vascular or heart problems diagnosed by doctor: high blood pressure	Cardiovascular		
		Atherosclerosis	Cardiovascular		
		Weight	Anthropometry	APOE deficiency is associated with elevated cholesterol levels, which may lead to obesity and increased fat mass	pmid6261329
		Trunk fat mass	Anthropometry		
		Leg fat mass	Anthropometry		
		Body fat percentage	Anthropometry		
		Arm fat percentage	Anthropometry		
		Whole body fat mass	Anthropometry		
		Leg predicted mass	Anthropometry		
		Trunk fat percentage	Anthropometry		
		Hip circumference	Anthropometry		
		Waist circumference	Anthropometry		
		Arm predicted mass	Anthropometry		
		Arm fat mass	Anthropometry		
		Body mass index	Anthropometry		
		Leg fat percentage	Anthropometry		
		Self-reported high cholesterol	Endocrine and metabolism	APOE deficiency is associated with elevated cholesterol levels	pmid27481046
		Deep venous thrombosis	Hematological	APOE deficiency is associated with femoral bruits, which are indicative of high blood pressure in the leg. This may relate to deep venous thrombosis, which includes sudden hypertension, leg pain/swelling/redness/warmth and leg blood clotting	http://iembase.org/app/#!/disorder/507
		Eosinophil count	Hematological	APOE deficiency is associated with splenomegaly, which can reduce the number of red blood cells, platelets and white blood cells in the bloodstream	pmid4242937
		Reticulocyte count	Hematological		
		Platelet crit	Hematological		
		Platelet count	Hematological		
		Sum basophil neutrophil counts	Hematological		
		Red blood cell count	Hematological		
ARG1 (6:131575856-132147278)	rs71753454 (I/D; 0.22)	Type 2 diabetes	Endocrine and metabolism	ARG1 deficiency is associated with feeding/protein aversion and vomiting, which could alter body fat composition and impact risk of type 2 diabetes	pmid27549856, pmid9762606
		Leg fat mass	Anthropometry	ARG1 deficiency is associated with feeding/protein aversion and vomiting, which could alter	pmid9762606, pmid2246859
		Leg predicted mass	Anthropometry		
		Arm fat mass	Anthropometry		
		Trunk fat mass	Anthropometry		
		Body fat percentage	Anthropometry		

		Waist circumference	Anthropometry	body fat composition and body size	
		Arm fat percentage	Anthropometry		
		Hip circumference	Anthropometry		
		Weight	Anthropometry		
		Body mass index	Anthropometry		
		Whole body fat mass	Anthropometry		
		Leg fat-free mass	Anthropometry	ARG1 deficiency is associated with spastic diplegia, a symptom of which is hypotonia. Low muscle tone could affect muscle mass in the body	pmid9762606, pmid2246859
CETP (16:56689969-57265091)	rs12149545 (A/G; 0.31)	Treatment with simvastatin	Medication	CETP deficiency is associated with altered cholesterol levels, which may be treated with cholesterol lowering medications	pmid168823, pmid3937535
		Medication use (HMG CoA reductase inhibitors)	Medication		
		Cholesterol lowering medication | medication for cholesterol, blood pressure, diabetes, or take exogenous hormones	Medication		
		Metabolic syndrome	Endocrine and metabolism	CETP deficiency is associated with altered cholesterol levels, which may lead to metabolic syndrome	pmid168823, pmid3937535
		Self-reported high cholesterol	Endocrine and metabolism	CETP deficiency is associated with altered cholesterol levels	pmid168823, pmid3937535
		Coronary artery disease	Cardiovascular	CETP deficiency leads to increased HDL cholesterol and low levels of LDL cholesterol, which may lead to cardiovascular disease	pmid168823, pmid3937535
		Myocardial infarction	Cardiovascular		
CPS1 (2:209335210-212467075)	rs1047891 (A/C; 0.32)	Operation code: cholecystectomy/gall bladder removal	Endocrine and metabolism	A case report of CPS1 deficiency reports severe gallbladder wall oedema, which could indicate gallbladder injury over time that leads to gallbladder removal	pmid27834067
		Systolic blood pressure	Hematological	CPS1 deficiency is associated with an increased risk of neonatal pulmonary hypertension, which affects systolic blood pressure	pmid11407344, pmid17188582
		Height	Anthropometry	CPS1 deficiency is associated with developmental delay, which includes stunted growth	https://ghr.nlm.nih.gov/condition/carbamoyl-phosphate-synthetase-i-deficiency
		Sitting height	Anthropometry		
		Pain type experienced in last month: headache	Neurological, cognitive or behavioural	CPS1 deficiency is associated with encephalopathy, a common symptom of which is headaches	pmid19793055
		Glomerular filtration rate	Renal	CPS1 deficiency is associated with inefficient clearance of ammonia; this could plausibly accumulate in the kidneys over time and dysregulate renal function	pmid20383146
		Chronic kidney disease	Renal		
		Hypertension	Cardiovascular	CPS1 deficiency is associated with neonatal susceptibility to pulmonary hypertension, symptoms of which include hypertension and high blood pressure	pmid11407344, pmid17188582
		Vascular or heart problems diagnosed by doctor: high blood pressure	Cardiovascular		
		Leg fat mass	Anthropometry		

		Whole body fat mass	Anthropometry	CPS1 deficiency is associated with poor feeding, which could affect body fat composition	https://ghr.nlm.nih.gov/condition/carbamoyl-phosphate-synthetase-i-deficiency
		Arm fat-free mass	Anthropometry		
		Whole body fat-free mass	Anthropometry		
		Weight	Anthropometry		
		Hip circumference	Anthropometry		
		Leg predicted mass	Anthropometry		
		Body mass index	Anthropometry		
		Trunk fat-free mass	Anthropometry		
		Arm predicted mass	Anthropometry		
		Trunk predicted mass	Anthropometry		
		Leg fat-free mass	Anthropometry		
		Arm fat mass	Anthropometry		
	rs13411696 (A/G; 0.44)	Pain type experienced in last month: headache	Neurological, cognitive or behavioural	CPS1 deficiency is associated with encephalopathy, a common symptom of which is headaches	pmid19793055
		Weight	Anthropometry	CPS1 deficiency is associated with poor feeding, which could affect body fat composition	https://ghr.nlm.nih.gov/condition/carbamoyl-phosphate-synthetase-i-deficiency
		Leg fat mass	Anthropometry		
		Arm predicted mass	Anthropometry		
		Arm fat-free mass	Anthropometry		
CPT2 (1:53193901-54025740)	rs35316080 (G/A; 0.36)	Systolic blood pressure	Hematological	CPT2 deficiency is associated with ventricular hypertrophy and cardiomegaly, both of which are a result of high blood pressure	pmid18550408, pmid1961225
CYP7A1 (8:59061697-59661042)	rs4738684 (G/A; 0.66)	Disorders of gallbladder, biliary tract and pancreas	Endocrine and metabolism	CYP7A1 deficiency is a metabolic disorder of gallbladder, biliary tract and pancreas	NA
		Cholelithiasis and cholecystitis	Endocrine and metabolism	CYP7A1 deficiency is associated with an increased risk of gallstone formation	pmid1682550
		Operation code: cholecystectomy/gall bladder removal	Endocrine and metabolism		
		Hypercholesterolemia	Cardiovascular	CYP7A1 deficiency is associated with hypercholesterolemia and may increase risk of cardiovascular disease	pmid29529257
		Self-reported high cholesterol	Endocrine and metabolism		
		Medication use (HMG CoA reductase inhibitors)	Medication	CYP7A1 deficiency is associated with hypercholesterolemia and may require cholesterol lowering treatment	pmid29529257
		Treatment with atorvastatin	Medication		
		Cholesterol lowering medication | medication for cholesterol, blood pressure, diabetes, or take exogenous hormones	Medication		
DBH (9:135888765-136851769)	rs6271 (T/C; 0.074)	Medication use (calcium channel blockers)	Medication	DBH deficiency is associated with orthostatic hypotension, which is treatable by calcium channel blockers	pmid3010116, pmid2300263, pmid1677640
		Pulse rate	Hematological	DBH deficiency is associated with orthostatic hypotension	pmid3010116, pmid2300263, pmid1677640
		Diastolic blood pressure	Hematological		
		Systolic blood pressure	Hematological		
		Hypertension	Cardiovascular		
		Blood pressure medication | medication for cholesterol, blood pressure, diabetes, or take exogenous hormones	Medication		

		Vascular or heart problems diagnosed by doctor: high blood pressure	Cardiovascular		
		Treatment/medication code: amlodipine	Medication		
DDC (7:50029382-50891329)	rs11771818 (A/G; 0.11)	Leg fat-free mass	Anthropometry	DDC deficiency is associated with both hyper- and hypotonia in different parts of the body; this can affect muscle mass	pmid12891654, pmid9309516, pmid1700191, pmid1357595
		Waist circumference	Anthropometry		
		Arm fat-free mass	Anthropometry		
		Leg predicted mass	Anthropometry		
		Trunk predicted mass	Anthropometry		
		Arm predicted mass	Anthropometry		
		Weight	Anthropometry		
		Whole body fat-free mass	Anthropometry		
		Trunk fat-free mass	Anthropometry		
		Comparative height size at age 10	Anthropometry	DDC deficiency is associated with developmental delay, which may affect comparative height size	pmid20505134
		Sitting height	Anthropometry		
		Height	Anthropometry		
		Whole body fat mass	Anthropometry	DDC deficiency is associated with poor feeding and swallowing difficulties, which may affect body fat composition and body size	pmid12891654, pmid20505134
		Trunk fat mass	Anthropometry		
		Arm fat mass	Anthropometry		
		Leg fat mass	Anthropometry		
		Body mass index	Anthropometry		
DPYD (1:97031198-98484512)	rs60392383 (G/T; 0.20)	Schizophrenia	Neurological, cognitive or behavioural	Although there is an unclear link between the IEM and outcome, DPYD deficiency can cause attention deficit disorder, which is highly comorbid with schizophrenia and may contribute to the latter's clinical presentation	pmid27575859, pmid24459374
		Breast cancer	Cancer	DPYD metabolism may affect fluoropyrimidines and drug therapies used in breast cancer; therefore, this link may be due to effects along the same metabolic pathway rather than a direct effect of pyrimidine metabolism on breast cancer	pmid29340111
	rs72977723 (A/G; 0.11)	Schizophrenia	Neurological, cognitive or behavioural	Although there is an unclear link between the IEM and outcome, DPYD deficiency can cause attention deficit disorder, which is highly comorbid with schizophrenia and may contribute to the latter's clinical presentation	pmid27575859, pmid24459374
ETFA (15:76247217-77203256)	rs2291449 (G/A; 0.093)	Alcohol intake frequency	Lifestyle	ETFA deficiency leads to fatty infiltration of the liver, which may affect alcohol intake frequency	pmid514320, pmid16434667
	rs2959850 (T/C; 0.46)	Qualifications: None of the above	Neurological, cognitive or behavioural	ETFA deficiency is associated with congenital brain anomalies, which may have some impact on educational attainment	pmid3754423
		Diverticulosis and diverticulitis	GIT	ETFA deficiency is associated with hepatomegaly, which presents as an abdominal mass and may increase risk of diverticulosis, as has been reported in one case report	pmid6862997, https://www.journalagent.com/z4/download_fulltext.asp?pdir=vtd&plng=tur&un=VTD-26535
ETFDH (4:158816749-159898596)	rs67481496 (T/A; 0.27)	Breast cancer	Cancer	Although the link between the IEM and outcome is not clear, the associated metabolite carnitine and its	pmid29445084, pmid32641979

				derivatives are important in breast cancer by providing an increased supply of energy to cancer cells	
FMO3 (1:170170833-172080696)	rs714839 (T/C; 0.41)	Mean platelet volume	Hematological	FMO3 deficiency is associated with defective membrane function in platelets, which may also indicate an effect on mean platelet volume	https://omim.org/entry/602079#11
GATM (15:44859591-46199886)	rs1145091 (C/T; 0.26)	Chronic kidney disease	Renal	GATM deficiency can cause decreased solute and water reabsorption in the kidney, which can result in renal insufficiency	pmid11090339
		Glomerular filtration rate	Renal		
		Arm predicted mass	Anthropometry	GATM deficiency is associated with failure to thrive, which may affect weight gain	pmid20682460
		Whole body fat-free mass	Anthropometry		
		Arm fat-free mass	Anthropometry		
		Trunk fat-free mass	Anthropometry		
		Trunk predicted mass	Anthropometry		
	rs2486274 (T/G; 0.38)	Glomerular filtration rate	Renal	GATM deficiency is associated with renal acidosis and insufficiency	pmid11090339
		Chronic kidney disease	Renal		
GCDH (19:12246385-13469265)	rs56397034 (C/G; 0.39)	Breast cancer	Cancer	Although the link between the IEM and outcome is unclear, the associated metabolite glutarylcarnitine has been associated with mouse breast cancer tissues and has been listed as one in a panel of biomarkers selected to diagnose breast cancer in a patent	pmid32641979, https://patents.google.com/patent/WO2016038157A1/en (PCT WO2016038157 A1)
		Type 2 diabetes	Endocrine and metabolism	Although there is an unclear link between the IEM and outcome, glutarylcarnitine levels are known to be elevated in type 2 diabetes patients compared to controls	pmid30423132, pmid23296094
		Metformin | treatment/medication code	Medication		
		Height	Anthropometry	GCDH deficiency is associated with developmental delay, which may affect height	pmid7564239
	rs8012 (G/A; 0.55)	Metformin | treatment/medication code	Medication	Although there is an unclear link between the IEM and outcome, glutarylcarnitine levels are known to be elevated in type 2 diabetes patients compared to controls	pmid30423132, pmid23296094
		Type 2 diabetes	Endocrine and metabolism	Although there is an unclear link between the IEM and outcome, levels of the associated metabolite glutarylcarnitine are shown to be elevated in obese versus lean subjects and in type 2 diabetes patients	pmid13990765, pmid15230623, pmid23296094
		Height	Anthropometry	GCDH deficiency is associated with developmental delay, which may affect height	pmid7564239
		Leg fat-free mass	Anthropometry	GCDH deficiency is associated with dystonia, which may affect muscle mass	pmid16602100
		Trunk predicted mass	Anthropometry		
		Leg predicted mass	Anthropometry		
		Whole body fat-free mass	Anthropometry		
		Trunk fat-free mass	Anthropometry		

HAL (12:95953054-96683079)	rs3213737 (A/G; 0.57)	Use of sun/uv protection	Lifestyle	HAL deficiency leads to a decrease in levels of trans-urocanate, which modulate skin response to UV light	pmid1943682
		Childhood sunburn occasions	Lifestyle		
		Ease of skin tanning	Lifestyle		
	rs61937878 (T/C; 0.0066)	Childhood sunburn occasions	Lifestyle		
HPD (12:119943590-122716919)	rs11043222 (T/C; 0.12)	Sitting height	Anthropometry	HPD deficiency is associated with failure to thrive, which can impact height and weight gain. pmid1130176, pmid858207, pmid7278885	
		Trunk fat-free mass	Anthropometry		
		Whole body fat-free mass	Anthropometry		
		Leg fat-free mass	Anthropometry		
		Trunk predicted mass	Anthropometry		
		Leg predicted mass	Anthropometry		
		Arm fat-free mass	Anthropometry		
		Height	Anthropometry		
		Weight	Anthropometry		
		Arm predicted mass	Anthropometry		
		Fluid intelligence score	Neurological, cognitive or behavioural	HPD deficiency is associated with mild mental retardation	pmid11073718
LCT (2:135169570-136883771)	rs4988235 (A/G; 0.70)	Bread intake	Lifestyle	LCT deficiency is associated with inability to digest lactose, which could reduce dairy consumption or products that are normally consumed with dairy (eg bread and butter)	pmid22826639
		Cereal intake	Lifestyle		
		Cheese intake	Lifestyle		
		Milk added to cereal	Lifestyle		
		Never/rarely have milk | milk type used	Lifestyle		
		Leg fat mass	Anthropometry	LCT deficiency is associated with metabolic acidosis, a symptom of which is vomiting and poor feeding. These symptoms can affect weight	pmid13565838
		Hip circumference	Anthropometry		
		Weight	Anthropometry		
		Arm fat mass	Anthropometry		
		Waist circumference	Anthropometry		
		Whole body fat mass	Anthropometry		
		Body fat percentage	Anthropometry		
		Arm fat percentage	Anthropometry		
		Trunk fat mass	Anthropometry		
		Body mass index	Anthropometry		
		Trunk fat percentage	Anthropometry		
		Leg fat percentage	Anthropometry		
		Forced vital capacity (fvc)	Respiratory	LCT deficiency is associated with metabolic acidosis, which causes rapid and shallow breathing. Metabolic acidosis may also lead to respiratory depression	pmid26215149
		Forced expiratory volume in 1-second (fev1)	Respiratory		
LDLR (19:10939298-11586444)	rs118068660 (T/C; 0.097)	Ischaemic stroke	Cardiovascular	LDLR deficiency is associated with hypercholesterolemia and adverse cardiovascular outcomes	pmid1301956
		Chronic ischaemic heart disease	Cardiovascular		
		Stroke	Cardiovascular		
		Ischaemic heart disease	Cardiovascular		
		Angina pectoris	Cardiovascular		
		Hypercholesterolemia	Cardiovascular		
		Major coronary heart disease event	Cardiovascular		
		Coronary artery disease	Cardiovascular		
		Self-reported high cholesterol	Endocrine and metabolism		
		Diseases of the circulatory system	Cardiovascular		
		Myocardial infarction	Cardiovascular		
		Atherosclerosis	Cardiovascular		
		Diagnoses - secondary ICD10: Z95.5 Presence of coronary angioplasty implant and graft	Cardiovascular		
		Weight	Anthropometry	LDLR deficiency is associated with hypercholesterolemia and may lead to increased weight	pmid1301956

		Non-butter spread type details: Flora Pro-Active or Benecol	Lifestyle	LDLR deficiency is associated with hypercholesterolemia, which may alter intake of high fat foods like butter	pmid1301956
		Lipitor 10mg tablet | treatment/medication code	Medication	LDLR deficiency is associated with hypercholesterolemia, which may require cholesterol lowering treatments	pmid1301956
		Treatment with aspirin	Medication		
		Rosuvastatin | treatment/medication code	Medication		
		Medication for pain relief, constipation, heartburn: aspirin	Medication		
		Medication use (HMG CoA reductase inhibitors)	Medication		
		Medication use (agents acting on the renin-angiotensin system)	Medication		
		Treatment with atorvastatin	Medication		
		Number of treatments/medications taken	Medication		
		Treatment with simvastatin	Medication		
		Treatment with ezetimibe	Medication		
		Cholesterol lowering medication | medication for cholesterol, blood pressure, diabetes, or take exogenous hormones	Medication		
		Medication use (antithrombotic agents)	Medication		
LIPC (15:58140540-59642157)	rs1077835 (G/A; 0.21)	Medication use (HMG CoA reductase inhibitors)	Medication	LIPC deficiency is associated with elevated cholesterol levels, which can be treated by HMG CoA reductase inhibitors	pmid1883393
		Coronary artery disease	Cardiovascular	LIPC deficiency is associated with elevated cholesterol levels, which may lead to coronary artery disease	pmid1883393
		Metabolic syndrome	Endocrine and metabolism	LIPC deficiency is associated with elevated cholesterol levels, which may lead to metabolic syndrome	pmid1883393
		Self-reported high cholesterol	Endocrine and metabolism	LIPC deficiency is associated with elevated cholesterol levels	pmid1883393
	rs12708454 (C/A; 0.31)	Self-reported high cholesterol	Endocrine and metabolism	LIPC deficiency is associated with hypercholesterolemia	pmid1883393
	rs1601935 (T/G; 0.65)	Self-reported high cholesterol	Endocrine and metabolism	LIPC deficiency is associated with elevated cholesterol levels	pmid1883393
		Cholesterol lowering medication | medication for cholesterol, blood pressure, diabetes, or take exogenous hormones	Medication	LIPC deficiency is associated with hypercholesterolemia, which may be treated using cholesterol lowering medication	pmid1883393
		Medication use (HMG CoA reductase inhibitors)	Medication		
		Treatment with atorvastatin	Medication		

		Metabolic syndrome	Endocrine and metabolism	LIPC deficiency is associated with hypercholesterolemia, which may increase risk of metabolic syndrome	pmid1883393
		Atherosclerosis	Cardiovascular	LIPC deficiency is associated with hypercholesterolemia, which may lead to adverse cardiovascular outcomes	pmid1883393
		Hypercholesterolemia	Cardiovascular	LIPC deficiency is associated with hypercholesterolemia	pmid1883393
	rs35853021 (T/G; 0.36)	Medication use (HMG CoA reductase inhibitors)	Medication	LIPC deficiency is associated with elevated cholesterol levels, which may be treated using cholesterol lowering medication	pmid1883393
		Treatment with atorvastatin	Medication		
		Cholesterol lowering medication | medication for cholesterol, blood pressure, diabetes, or take exogenous hormones	Medication		
		Metabolic syndrome	Endocrine and metabolism	LIPC deficiency is associated with elevated cholesterol levels, which may lead to metabolic syndrome	pmid1883393
		Self-reported high cholesterol	Endocrine and metabolism	LIPC deficiency is associated with elevated cholesterol levels	pmid1883393
		Atherosclerosis	Cardiovascular	LIPC deficiency is associated with hypercholesterolemia, which may lead to adverse cardiovascular outcomes	pmid1883393
		Hypercholesterolemia	Cardiovascular	LIPC deficiency is associated with hypercholesterolemia	pmid1883393
LPL (8:19565852-20190058)	rs1441764 (T/C; 0.32)	Type 2 diabetes	Endocrine and metabolism	LPL deficiency is associated with hypercholesterolemia, which may lead to an increased risk of type 2 diabetes	pmid20301485
		Self-reported high cholesterol	Endocrine and metabolism	LPL deficiency is associated with hypercholesterolemia	pmid20301485
		Waist-to-hip ratio	Anthropometry	LPL deficiency is associated with lipid accumulation and increased lipid storage in adipose tissue, which could lead to an increase in waist-to-hip ratio	pmid20301485
		Metabolic syndrome	Endocrine and metabolism	LPL deficiency is associated with lipid accumulation and increased lipid storage in adipose tissue, which could lead to metabolic syndrome	pmid20301485
		Angina pectoris	Cardiovascular	LPL deficiency is associated with lipid accumulation and increased lipid storage in adipose tissue, which in turn has been associated with increased risk of adverse cardiovascular outcomes	pmid20301485
		Ischaemic heart disease	Cardiovascular		
		Coronary artery disease	Cardiovascular		
		Myocardial infarction	Cardiovascular		
		Fenofibrate | treatment/medication code	Medication		
		Diseases of the circulatory system	Cardiovascular		
		Major coronary heart disease event	Cardiovascular		
		Treatment with aspirin	Medication	LPL deficiency is associated with lipid accumulation and increased lipid storage in adipose tissue, which in turn has been linked to hypercholesterolemia and need to treat	pmid20301485
		Cholesterol lowering medication | medication for cholesterol, blood pressure, diabetes, or take exogenous hormones	Medication		

		Treatment with atorvastatin	Medication		
		Medication use (HMG CoA reductase inhibitors)	Medication		
		Treatment with simvastatin	Medication		
		Medication for pain relief, constipation, heartburn: aspirin	Medication		
	rs15285 (T/C; 0.29)	Self-reported high cholesterol	Endocrine and metabolism	LPL deficiency is associated with hypercholesterolemia	pmid20301485
		Waist-to-hip ratio	Anthropometry	LPL deficiency is associated with lipid accumulation and increased lipid storage in adipose tissue, which could lead to an increase in waist-to-hip ratio	pmid20301485
		Metabolic syndrome	Endocrine and metabolism	LPL deficiency is associated with lipid accumulation and increased lipid storage in adipose tissue, which could lead to metabolic syndrome	pmid20301485
		Ischaemic heart disease	Cardiovascular	LPL deficiency is associated with lipid accumulation and increased lipid storage in adipose tissue, which in turn has been associated with increased risk of adverse cardiovascular outcomes	pmid20301485
		Myocardial infarction	Cardiovascular		
		Angina pectoris	Cardiovascular		
		Major coronary heart disease event	Cardiovascular		
		Coronary artery disease	Cardiovascular		
		Atherosclerosis	Cardiovascular		
		Diseases of the circulatory system	Cardiovascular		
		Treatment with aspirin	Medication	LPL deficiency is associated with lipid accumulation and increased lipid storage in adipose tissue, which in turn has been linked to hypercholesterolemia and need to treat	pmid20301485
		Cholesterol lowering medication | medication for cholesterol, blood pressure, diabetes, or take exogenous hormones	Medication		
		Treatment with simvastatin	Medication		
		Medication use (HMG CoA reductase inhibitors)	Medication		
		Treatment with atorvastatin	Medication		
		Fenofibrate | treatment/medication code	Medication		
		Medication for pain relief, constipation, heartburn: aspirin	Medication		
		Type 2 diabetes	Endocrine and metabolism	LPL deficiency is associated with lipid accumulation, which is linked to increased risk of type 2 diabetes	pmid3552532, pmid18985010
MTHFR (1:11356378-12356378)	rs1801133 (A/G; 0.34)	Multiple sclerosis	Inflammatory	MTHFR deficiency is associated with seizures and neuropathy, which may be reflected in multiple sclerosis	pmid25024447
NAGS (17:41860044-42705022)	rs860354 (A/T; 0.64)	Comparative height size at age 10	Anthropometry	NAGS deficiency is associated with developmental delay, which could affect relative height in childhood	pmid1405478
NT5C3A (7:32640237-33645764)	rs4316067 (G/A; 0.30)	Comparative body size at age 10	Anthropometry	NT5C3A deficiency is associated with developmental delay, which may affect comparative body size at age 10	http://iembase.org/app/#!/disorder/156

		Mean sphered cell volume	Hematological	NT5C3A deficiency is associated with hemolytic anemia with abnormal (non-spherocytic) red blood cell shapes. This can affect the count, volume and shape of red blood cells	pmid3352512, pmid4372252
		Red cell distribution width	Hematological		
		Reticulocyte fraction of red cells	Hematological		
		Haematocrit percentage	Hematological		
		Reticulocyte count	Hematological		
		Red blood cell count	Hematological		
		Reticulocyte percentage	Hematological		
		Mean corpuscular volume	Hematological		
		Mean corpuscular haemoglobin	Hematological		
		Mean reticulocyte volume	Hematological		
OPLAH/CYC1 (8:143909290-146458377)	rs3935209 (G/T; 0.082)	Cognitive performance	Neurological, cognitive or behavioural	OPLAH deficiency is associated with psychomotor retardation, which may impair cognitive function	pmid7542714
		Intelligence	Neurological, cognitive or behavioural		
PCCB (3:135676622-136542036)	rs645040 (T/G; 0.77)	Schizophrenia	Neurological, cognitive or behavioural	Late-onset PCCB deficiency is associated with psychosis, which is a symptom of schizophrenia	pmid18174561
		Sleep duration	Lifestyle	PCCB deficiency is associated with anemia, which may affect sleep duration	pmid30879957
		Medication for pain relief, constipation, heartburn: aspirin	Medication	PCCB deficiency is associated with cardiomyopathy and prolonged QT interval, which are often accompanied with hypertension and may require blood pressure lowering medication	pmid30879957
		Treatment with aspirin	Medication		
		Blood pressure medication | medication for cholesterol, blood pressure, diabetes, or take exogenous hormones	Medication		
		Medication use (agents acting on the renin-angiotensin system)	Medication		
		Coronary artery disease	Cardiovascular	PCCB deficiency is associated with cardiomyopathy, which is directly related to coronary artery disease	pmid30879957
		Ischaemic heart disease	Cardiovascular		
		Major coronary heart disease event	Cardiovascular		
		Myocardial infarction	Cardiovascular		
		Relative age of first facial hair	Reproductive and urinary	PCCB deficiency is associated with developmental retardation and may delay facial hair development in men	pmid30879957
		Metabolic syndrome	Endocrine and metabolism	PCCB deficiency is associated with hyperglycinemia, anemia and metabolic decompensation, all of which may contribute to metabolic syndrome	pmid30879957
		Time to complete round	Neurological, cognitive or behavioural	PCCB deficiency is associated with intellectual disability, which may affect 'time to complete round' in a cognitive function experiment	pmid30879957
		Lymphocyte count	Hematological	PCCB deficiency is associated with leukopenia	pmid6026548
		Monocyte count	Hematological		
		Pork intake	Lifestyle	PCCB deficiency is associated with metabolic acidosis and feeding difficulties, for which	pmid25205257
		Variation in diet	Lifestyle		
		Lamb/mutton intake	Lifestyle		

		Beef intake	Lifestyle	treatment includes low protein intake and specific food formulas	
		Forced expiratory volume in 1-second (fev1), predicted	Respiratory	PCCB deficiency is associated with metabolic acidosis, which can be caused by a build-up of carbon dioxide in the blood due to poor lung function	pmid30879957
		Wheeze or whistling in the chest in last year	Respiratory		
		Neutrophil percentage	Hematological	PCCB deficiency is associated with neutropenia	pmid6026548
		Neutrophil percentage of white cells	Hematological		
		Vascular or heart problems diagnosed by doctor: high blood pressure	Cardiovascular	PCCB deficiency is associated with QT interval prolongation, which leads to a slower heart rate. A reduced risk of hypertension would be expected as a milder phenotype based on this symptom of the IEM	pmid30879957
		Hypertension	Cardiovascular		
		Platelet crit	Hematological	PCCB deficiency is associated with thrombocytopenia, which is directly related to platelet crit (loss of total platelet mass)	pmid6026548, pmid13693094
		Platelet count	Hematological		
		Self-reported high cholesterol	Endocrine and metabolism	PCCB is involved in the breakdown of cholesterol, so deficiency would be linked to hypercholesterolemia	https://rarediseases.org/rare-diseases/propionic-acidemia/
		Cholesterol lowering medication | medication for cholesterol, blood pressure, diabetes, or take exogenous hormones	Medication		
		Medication use (HMG CoA reductase inhibitors)	Medication		
		Treatment with simvastatin	Medication		
		Body fat percentage	Anthropometry	Symptoms of PCCB deficiency include feeding difficulties and metabolic/ketoacidosis, which could affect body fat composition and body size	pmid30879957
		Body mass index	Anthropometry		
		Leg fat percentage	Anthropometry		
		Trunk fat mass	Anthropometry		
		Hip circumference	Anthropometry		
		Arm fat percentage	Anthropometry		
		Waist circumference	Anthropometry		
		Whole body fat mass	Anthropometry		
		Arm fat mass	Anthropometry		
		Weight	Anthropometry		
		Waist-to-hip ratio	Anthropometry		
		Arm predicted mass	Anthropometry		
		Leg fat mass	Anthropometry		
PCSK9 (1:55255647-55755647)	rs11591147 (T/G; 0.018)	Treatment with atorvastatin	Medication	High cholesterol caused by PCSK9 mutations are often treated with statins	pmid16424354
		Medication use (HMG CoA reductase inhibitors)	Medication		
		Treatment with simvastatin	Medication		
		Cholesterol lowering medication | medication for cholesterol, blood pressure, diabetes, or take exogenous hormones	Medication		
		Lipitor 10mg tablet | treatment/medication code	Medication	PCSK9 deficiency is known to cause early-onset hypercholesterolemia and coronary artery disease, requiring treatment	pmid16424354
		Treatment with ezetimibe	Medication		

		Medication for pain relief, constipation, heartburn: aspirin	Medication		
		Chronic ischaemic heart disease	Cardiovascular	PCSK9 deficiency is known to cause early-onset hypercholesterolemia and coronary artery disease	pmid17435765
		Angina pectoris	Cardiovascular		
		Coronary artery disease	Cardiovascular		
		Self-reported high cholesterol	Endocrine and metabolism		
		Ischaemic heart disease	Cardiovascular		
		Myocardial infarction	Cardiovascular		
		Diseases of the circulatory system	Cardiovascular		
		Atherosclerosis	Cardiovascular		
		Hypercholesterolemia	Cardiovascular		
		Butter/spreadable butter | spread type	Lifestyle	PCSK9 deficiency is known to cause early-onset hypercholesterolemia and may reduce high fat food intake	pmid17435765
		Non-butter spread type details: Flora Pro-Active or Benecol	Lifestyle		
PHGDH (1:119218284-120594880)	rs561931 (G/A; 0.59)	Haematocrit percentage	Hematological	Megaloblastic anaemia has been reported in some patients with PHGDH deficiency. This symptom causes the formation of larger than normal red blood cells and fewer red blood cells in total	pmid21113737
		Red blood cell count	Hematological		
		Multiple sclerosis	Inflammatory	PHGDH deficiency is associated with microcephaly, a condition in which myelinating abnormalities have been observed. This may affect multiple sclerosis risk	pmid8758134, pmid28007986, pmid28413018
PPM1K (4:88255945-89499444)	rs10018448 (G/A; 0.53)	Type 2 diabetes	Endocrine and metabolism	PPM1K deficiency is associated with metabolic derangement and feeding difficulties, which may affect body fat deposition and type 2 diabetes risk	pmid14567968
		Waist circumference	Anthropometry	PPM1K deficiency is associated with metabolic derangement and feeding difficulties, which may affect waist circumference	pmid14567968
		Leg fat percentage	Anthropometry		
PSPH (7:55625223-56388322)	rs4470984 (G/A; 0.76)	Height	Anthropometry	PSPH deficiency is associated with growth retardation	pmid9222972
		Sitting height	Anthropometry		
RBP4 (10:95110964-95610964)	rs10882283 (C/A; 0.38)	Optic disc area	Eye	RBP4 deficiency is associated with abnormal eye development, a symptom of which includes optic pit, a congenital protrusion or depression of the optic disc. This symptom may be reflected by variation in optic disc area	pmid24168988
SLC16A12 (10:90972995-91893687)	rs7081788 (A/G; 0.24)	Breast cancer	Cancer	Although the link between the IEM and outcome is not clear, levels of the associated metabolite creatine are elevated in tumourous breast cancer tissue compared to adjacent non-cancerous tissue	pmid23613877
SLC22A5 (5:129311501-132923264)	rs538021413 (D/I; 0.63)	Blood pressure medication | medication for cholesterol, blood pressure, diabetes, or	Medication	Impaired fatty acid oxidation in heart muscle due to SLC22A5 deficiency may cause an accumulation of	pmid12210323

		take exogenous hormones		lipids in blood vessels that may lead to hypertension and a need to treat	
		Diastolic blood pressure	Hematological		
		Vascular or heart problems diagnosed by doctor: high blood pressure	Cardiovascular	Impaired fatty acid oxidation in heart muscle due to SLC22A5 deficiency may cause an accumulation of lipids in blood vessels that may lead to hypertension and cardiomyopathy	pmid12210323
		Hypertension	Cardiovascular		
		Forced expiratory volume in 1-second (fev1), predicted	Respiratory	SLC22A5 deficiency can lead to cardiovascular abnormalities and irregular respiration	pmid9634512
		Forced expiratory volume in 1-second (fev1)	Respiratory		
		Glomerular filtration rate	Renal	SLC22A5 deficiency can lead to increased urinary carnitine excretion, which may indicate defects in renal carnitine reabsorption	pmid7131143
		Weight	Anthropometry	SLC22A5 deficiency can reduce lipid stores in muscle tissue, which leads to known symptoms of skeletal myopathy	pmid4687787, pmid4414743, pmid123043
		Leg predicted mass	Anthropometry		
		Arm predicted mass	Anthropometry		
		Leg fat-free mass	Anthropometry		
		Trunk predicted mass	Anthropometry		
		Trunk fat-free mass	Anthropometry		
		Whole body fat-free mass	Anthropometry		
		Arm fat-free mass	Anthropometry		
SLC25A19 (17:71599741-74891603)	rs7222784 (A/T; 0.68)	Height	Anthropometry	SLC25A19 deficiency can cause developmental delay, which may affect height	pmid20583149
		Comparative height size at age 10	Anthropometry		
SLC5A1 (22:31889342-32889342)	rs117086479 (G/A; 0.06)	Type 2 diabetes	Endocrine and metabolism	SLC5A1 deficiency is associated with glucose/galactose malabsorption, which may contribute to type 2 diabetes	pmid20486940
SLC7A9 (19:32932750-34004548)	rs7247977 (C/T; 0.39)	Chronic kidney disease	Renal	SLC7A9 deficiency is associated with the formation of renal calculi as a result of low renal reabsorption ability of cystine, which can lead to chronic kidney disease	pmid21255007
		Glomerular filtration rate	Renal		
SLCO1A2/ *SLCO1B1/* *SLCO1B3* (12:20367806-23013636)	rs4149056 (C/T; 0.15)	Daytime dozing / sleeping (narcolepsy)	Neurological, cognitive or behavioural	Although the link between the IEM and outcome is not clear, hyperbilirubinemia has been shown to affect sleep-wake cycles of affected newborns	pmid28072860
TF (3:132977701-133977701)	rs8177240 (T/G; 0.67)	Red blood cell count	Hematological	TF deficiency is associated with anemia, which may affect the volume and size of red blood cells	pmid11110675
		Mean corpuscular volume	Hematological		
		Mean corpuscular haemoglobin	Hematological	TF deficiency is associated with anemia, which may be defined as reduced haemoglobin content	pmid11110675, pmid21694802
TH (11:1735287-2479759)	rs10840516 (A/G; 0.24)	Years of schooling	Neurological, cognitive or behavioural	Case reports have identified siblings with severe intellectual disability as well as other symptoms compatible with TH deficiency	pmid21937992
		Pulse rate	Hematological	TH deficiency is associated with autonomic dysfunction, which may affect pulse rate	pmid22815559
		Arm fat-free mass	Anthropometry		pmid22815559
		Weight	Anthropometry		

		Arm predicted mass	Anthropometry	TH deficiency is associated with hypotonia, which decreases muscle tone	
		Leg fat-free mass	Anthropometry		
		Trunk predicted mass	Anthropometry		
		Trunk fat-free mass	Anthropometry		
		Leg predicted mass	Anthropometry		
		Whole body fat-free mass	Anthropometry		
		Body mass index	Anthropometry		
TKT (3:52712681-54087252)	rs4687717 (C/T; 0.57)	Pulse rate	Hematological	TKT deficiency is linked to congenital heart defects, which may alter pulse rate	pmid27259054
		Spherical power	Eye	TKT deficiency is linked to uveitis and conjunctivitis, both of which may lead to reduced visual acuity and affect spherical power	pmid27259054
	rs4687718 (G/A; 0.87)	QRS duration	Cardiovascular	TKT deficiency is associated with congenital heart defects, which may affect QRS duration	pmid27259054
		Pulse rate	Hematological	TKT deficiency is linked to congenital heart defects, which may alter pulse rate	pmid27259054
		3mm weak meridian	Eye	TKT deficiency is linked to uveitis and conjunctivitis, both of which may lead to reduced visual acuity and affect spherical power	pmid27259054
		6mm weak meridian	Eye		
UGT1A1/UGT1A3/UGT1A4/UGT1A5/UGT1A6/UGT1A7/UGT1A8/UGT1A9/UGT1A10 (2:233892619-235187605)	rs1976391 (G/A; 0.31)	Skin colour	Anthropometry	UGT1A1 deficiency can cause neonatal jaundice, which is often treated by phototherapy. This may affect skin colour and ease of skin tanning, though it is important to note that these phenotypes may be a side effect rather than a direct cause of the metabolic alteration	pmid12983120, pmid23950218
		Ease of skin tanning	Lifestyle		
		Disorders of gallbladder, biliary tract and pancreas	Endocrine and metabolism	UGT1A1 deficiency is a disorder of the gallbladder and biliary tract that leads to hyperbilirubinemia and cholelithiasis	pmid12983120
		Operation code: cholecystectomy/gall bladder removal	Endocrine and metabolism	UGT1A1 deficiency is associated with hyperbilirubinemia, which can lead to the formation of gallstones and may require surgery to remove	pmid14013759
		Self-reported liver or biliary/pancreas problem	Endocrine and metabolism	UGT1A1 deficiency is associated with jaundice and is a known risk factor of gallstone formation	pmid8596320
		Cholelithiasis and cholecystitis	Endocrine and metabolism		
	rs201829156 (D/I; 0.35)	Skin colour	Anthropometry	UGT1A1 deficiency can cause neonatal jaundice, which is often treated by phototherapy. This may affect skin colour and ease of skin tanning, though it is important to note that these phenotypes may be a side effect rather than a direct cause of the metabolic alteration	pmid12983120, pmid23950218
	rs201829156 (D/I; 0.35)	Ease of skin tanning	Lifestyle		
	rs201829156 (D/I; 0.35)	Self-reported liver or biliary/pancreas problem	Endocrine and metabolism	UGT1A1 deficiency is associated with jaundice and	pmid8596320

		Cholelithiasis and cholecystitis	Endocrine and metabolism	is a known risk factor of gallstone formation	
UMPS (3:124028592-125111128)	rs2242247 (T/A; 0.17)	Heel bone mineral density	Bone	Patients with urolithiasis have increased disintegration and lower bone mass	pmid25568567, pmid18359393
		Mean platelet volume	Hematological	UMPS deficiency is associated with megaloblastic anemia, a condition in which larger than usual red blood cells are produced. Reduced platelet counts with varying sizes are also recorded in patients with UMPS deficiency	https://pdfs.semanticscholar.org/477a/e0651623c694a7cc8f8833178d18ba09b6d5.pdf, https://www.intechopen.com/books/current-topics-in-anemia/megaloblastic anemia
		Platelet distribution width	Hematological		

Ch5_ST4: Summary results for novel colocalising metabolite-phenotype clusters. PP_R = regional posterior probability; PP_A = alignment posterior probability; PP_E = explained posterior probability.

Locus	IFVs	Colocalising traits	PP_R/PP_A	Candidate causal variant (R^2 with IFV)	PP_E
ABCA1 (9:107264123-107915739)	rs2575876	Self-reported high cholesterol, Cholesterol lowering medication | medication for cholesterol, blood pressure, diabetes, or take exogenous hormones, Medication use (HMG CoA reductase inhibitors), sphingomyelin (d18:1/20:0, d16:1/22:0)*, sphingomyelin (d18:1/14:0, d16:1/16:0)*	0.99/0.97	rs2740488	0.51
ACSF3 (16:88668096-90424092)	rs36099289, rs72817435	Body mass index, ethylmalonate	0.99/0.70	rs72817435 (-)	0.11
APOA5/ APOA4/ APOA1/ APOC3 (11:116274569-117470429)	rs530885291, rs964184	Self-reported high cholesterol, Hypercholesterolemia, Coronary artery disease, Metabolic syndrome, 1-palmitoyl-2-docosahexaenoyl-GPE (16:0/22:6)*, 1-oleoyl-2-linoleoyl-glycerol (18:1/18:2), 1-oleoyl-3-linoleoyl-glycerol (18:1/18:2), 1-stearoyl-2-arachidonoyl-GPE (18:0/20:4), 1-stearoyl-2-docosahexaenoyl-GPE (18:0/22:6)*, 1-palmitoyl-2-linoleoyl-glycerol (16:0/18:2)*, 1-palmitoyl-3-linoleoyl-glycerol (16:0/18:2)*, 1-stearoyl-2-arachidonoyl-GPI (18:0/20:4), 1-stearoyl-2-linoleoyl-GPE (18:0/18:2)*, 1-stearoyl-2-oleoyl-GPE (18:0/18:1), 1-stearoyl-2-linoleoyl-GPI (18:0/18:2), 1-palmitoyl-2-stearoyl-GPC (16:0/18:0), 1-palmitoyl-2-linoleoyl-GPC (16:0/18:2), 1-stearoyl-2-linoleoyl-GPC (18:0/18:2)*, N-palmitoyl-sphingosine (d18:1/16:0), 1-margaroyl-2-linoleoyl-GPC (17:0/18:2)*, 1-palmitoyl-2-linoleoyl-GPE (16:0/18:2), 1-oleoyl-2-docosahexaenoyl-GPC (18:1/22:6)*, phosphatidylcholine (16:0/22:5n3, 18:1/20:4)*, 1-stearoyl-GPE (18:0), 1-stearoyl-GPC (18:0)	1/>0.99	rs964184 (-)	1
APOB (2:20893329-21546682)	rs934197	Self-reported high cholesterol, Treatment with simvastatin, Treatment with atorvastatin, Hypercholesterolemia, Medication use (HMG CoA reductase inhibitors), cholesterol, palmitoyl sphingomyelin (d18:1/16:0), 1-palmitoyl-2-stearoyl-GPC (16:0/18:0), palmitoyl dihydrosphingomyelin (d18:0/16:0)*	1/0.79	rs1367117	0.52
APOE/ APOC1/ APOC2/ APOC4 (19:45081103-45922478)	rs7412, rs5112, rs429358, rs204474	Deep venous thrombosis, Coronary artery disease, Self-reported high cholesterol, Hypercholesterolemia, Angina pectoris, Myocardial infarction, Vascular or heart problems diagnosed by doctor: high blood pressure, Hypertension, Diagnoses -	>0.99/>0.99	rs7412 (-)	1

		secondary ICD10: Z95.5 Presence of coronary angioplasty implant and graft, 1-(1-enyl-stearoyl)-2-linoleoyl-GPE (P-18:0/18:2)*, cholesterol, N-palmitoyl-sphingosine (d18:1/16:0), 1-stearoyl-GPE (18:0), 1-palmitoyl-2-stearoyl-GPC (16:0/18:0), 1-nonadecanoyl-GPC (19:0), 1-stearoyl-2-arachidonoyl-GPE (18:0/20:4), 1-palmitoyl-2-docosahexaenoyl-GPE (16:0/22:6)*, 1-oleoyl-2-linoleoyl-glycerol (18:1/18:2), lactosyl-N-palmitoyl-sphingosine, sphingomyelin (d18:1/20:1, d18:2/20:0)*			
		Metabolic syndrome, 1-oleoylglycerol (18:1), palmitoyl sphingomyelin (d18:1/16:0)	1/0.95	rs483082 (<0.1 with all IFVs)	0.62
ARG1 (6:131575856-132147278)	rs71753454	Type 2 diabetes, Hip circumference, Weight, Body mass index, Leg fat-free mass, Leg predicted mass, Arm fat mass, Whole body fat mass, Trunk fat mass, Leg fat mass, Waist circumference, Arm fat percentage, Body fat percentage, arginine	>0.99/0.95	rs2781668 (0.64)	0.85
CETP (16:56689969-57265091)	rs12149545	Treatment with simvastatin, Coronary artery disease, Cholesterol lowering medication | medication for cholesterol, blood pressure, diabetes, or take exogenous hormones, Self-reported high cholesterol, Metabolic syndrome, Medication use (HMG CoA reductase inhibitors), 1-(1-enyl-palmitoyl)-2-palmitoleoyl-GPC (P-16:0/16:1)*, 1-(1-enyl-palmitoyl)-2-linoleoyl-GPC (P-16:0/18:2)*, 1-(1-enyl-palmitoyl)-2-palmitoyl-GPC (P-16:0/16:0)*, 1-(1-enyl-palmitoyl)-2-oleoyl-GPC (P-16:0/18:1)*, 1-(1-enyl-palmitoyl)-2-arachidonoyl-GPC (P-16:0/20:4)*, 1-palmitoyl-2-linoleoyl-glycerol (16:0/18:2)*, 1-palmitoyl-2-arachidonoyl-GPC (16:0/20:4), 1-(1-enyl-stearoyl)-2-oleoyl-GPC (P-18:0/18:1), 1-stearoyl-2-linoleoyl-GPI (18:0/18:2), 1-palmitoyl-2-linoleoyl-GPC (16:0/18:2), 1,2-dipalmitoyl-GPC (16:0/16:0), 1-palmityl-2-arachidonoyl-GPC (O-16:0/20:4)*, 1-(1-enyl-palmitoyl)-2-myristoyl-GPC (P-16:0/14:0)*, 1-stearoyl-2-arachidonoyl-GPC (18:0/20:4), 1-(1-enyl-stearoyl)-2-linoleoyl-GPC (P-18:0/18:2)*, 1-myristoyl-2-linoleoyl-GPC (14:0/18:2)*, 1-palmitoyl-2-palmitoleoyl-GPC (16:0/16:1)*	0.97/0.89	rs183130	0.9391
CPS1 (2:209335210-212467075)	rs1047891, rs13411696	Operation code: cholecystectomy/gall bladder removal, Vascular or heart problems diagnosed by doctor: high blood pressure, Hypertension, Chronic kidney disease, glycine, N-acetylglycine, gamma-glutamylglycine, propionylglycine, 3-methylglutarylcarnitine (2), homoarginine, isobutyrylglycine, hexanoylglycine, tigloylglycine, isovalerylglycine, cinnamoylglycine, creatine, pyroglutamine*, citrulline	0.99/0.99	rs1047891 (-)	1
CPT2 (1:53193901-54025740)	rs35316080	Systolic blood pressure, succinylcarnitine	0.96/0.83	rs35316080 (-)	0.49
CYP7A1 (8:59061697-59661042)	rs4738684	Operation code: cholecystectomy/gall bladder removal, Disorders of gallbladder, biliary tract and pancreas, Self-reported high cholesterol, Cholelithiasis and cholecystitis, Cholesterol lowering medication | medication for cholesterol, blood pressure, diabetes, or take exogenous hormones, Treatment with	0.87/0.75	rs10107182 (1)	0.51

		atorvastatin, Medication use (HMG CoA reductase inhibitors), taurocholenate sulfate, glycodeoxycholate sulfate, 7-alpha-hydroxy-3-oxo-4-cholestenoate (7-Hoca), glycochenodeoxycholate glucuronide (1), glycodeoxycholate, deoxycholate			
DBH (9:135888765-136851769)	rs6271	Diastolic blood pressure, Vascular or heart problems diagnosed by doctor: high blood pressure, Hypertension, Systolic blood pressure, Blood pressure medication | medication for cholesterol, blood pressure, diabetes, or take exogenous hormones, Treatment/medication code: amlodipine, Medication use (calcium channel blockers), vanillylmandelate (VMA)	0.97/0.97	rs6271	1
ETFA (15:76247217-77203256)	rs2291449, rs2959850	Alcohol intake frequency, isovalerylcarnitine, butyrylcarnitine, ethylmalonate, dimethylglycine	>0.99/0.72	rs78185702 (0.98 with rs2291449, 0.11 with rs2959850)	0.98
GATM (15:44859591-46199886)	rs1145091, rs2486274	Chronic kidney disease, homoarginine	1/0.95	rs1049518	0.16
HAL (12:95953054-96683079)	rs3213737, rs61937878	Childhood sunburn occasions, Ease of skin tanning, histidine	>0.99/0.98	rs3213737	0.66
LCT (2:135169570-136883771)	rs4988235	Bread intake, Cereal intake, Milk added to cereal, galactonate, 1,5-anhydroglucitol (1,5-AG)	0.96/0.92	rs182549	0.68
LDLR (19:10939298-11586444)	rs118068660	Treatment with ezetimibe, Stroke, Rosuvastatin | treatment/medication code, Ischaemic heart disease, Atherosclerosis, Diseases of the circulatory system, Non-butter spread type details: Flora Pro-Active or Benecol, Ischaemic stroke, Self-reported high cholesterol, Cholesterol lowering medication | medication for cholesterol, blood pressure, diabetes, or take exogenous hormones, Treatment with simvastatin, Hypercholesterolemia, Treatment with atorvastatin, Coronary artery disease, Angina pectoris, Medication for pain relief, constipation, heartburn: aspirin, Treatment with aspirin, Chronic ischaemic heart disease, Major coronary heart disease event, Number of treatments/medications taken, Weight, Medication use (agents acting on the renin-angiotensin system), Medication use (antithrombotic agents), Medication use (HMG CoA reductase inhibitors), cholesterol, palmitoyl dihydrosphingomyelin (d18:0/16:0)*, palmitoyl sphingomyelin (d18:1/16:0), glycosyl-N-palmitoyl-sphingosine, lactosyl-N-palmitoyl-sphingosine, N-palmitoyl-sphingosine (d18:1/16:0), stearoyl sphingomyelin (d18:1/18:0), sphingomyelin (d18:1/15:0, d16:1/17:0)*, sphingomyelin (d18:1/17:0, d17:1/18:0, d19:1/16:0), sphingomyelin (d18:1/14:0, d16:1/16:0)*, sphingomyelin (d18:1/20:0, d16:1/22:0)*, 1,2-dipalmitoyl-GPC (16:0/16:0), sphingomyelin (d18:1/22:1, d18:2/22:0, d16:1/24:1)*	0.99/0.76	rs138294113	0.35
LPL (8:19565852-20190058)	rs1441764, rs15285	Coronary artery disease, Self-reported high cholesterol, Ischaemic heart disease, Angina pectoris, Myocardial infarction, Type 2 diabetes, Diseases of the circulatory system, Metabolic syndrome,	0.99/0.53	rs15285 (-)	0.97

		1-oleoyl-2-linoleoyl-glycerol (18:1/18:2), 1-palmitoyl-2-linoleoyl-glycerol (16:0/18:2)*, 1-oleoyl-3-linoleoyl-glycerol (18:1/18:2), 1-palmitoyl-3-linoleoyl-glycerol (16:0/18:2)*, 1-(1-enyl-palmitoyl)-2-oleoyl-GPC (P-16:0/18:1)*, 1-(1-enyl-palmitoyl)-2-linoleoyl-GPC (P-16:0/18:2)*			
OPLAH/CYC1 (8:143909290-146458377)	rs3935209	Cognitive performance, Intelligence, 5-oxoproline, 6-oxopiperidine-2-carboxylic acid	0.94/0.9	rs3935209 (-)	1
PCSK9 (1:55255647-55755647)	rs11591147	Treatment with ezetimibe, Butter/spreadable butter | spread type, Lipitor 10mg tablet | treatment/medication code, Treatment with atorvastatin, Chronic ischaemic heart disease, Angina pectoris, Coronary artery disease, Self-reported high cholesterol, Treatment with simvastatin, Cholesterol lowering medication | medication for cholesterol, blood pressure, diabetes, or take exogenous hormones, Hypercholesterolemia, Ischaemic heart disease, Atherosclerosis, Diseases of the circulatory system, Non-butter spread type details: Flora Pro-Active or Benecol, Medication for pain relief, constipation, heartburn: aspirin, Myocardial infarction, Medication use (HMG CoA reductase inhibitors), cholesterol, palmitoyl dihydrosphingomyelin (d18:0/16:0)*, palmitoyl sphingomyelin (d18:1/16:0), stearoyl sphingomyelin (d18:1/18:0), sphingomyelin (d18:1/17:0, d17:1/18:0, d19:1/16:0), lactosyl-N-palmitoyl-sphingosine, sphingomyelin (d18:1/15:0, d16:1/17:0)*, sphingomyelin (d18:2/16:0, d18:1/16:1)*	0.95/0.95	rs11591147 (-)	1
PPM1K (4:88255945-89499444)	rs10018448	Waist circumference, Leg fat percentage, valine, alpha-hydroxyisovalerate, 3-methyl-2-oxobutyrate, 4-methyl-2-oxopentanoate, 3-methyl-2-oxovalerate	>0.99/0.71	rs1129043 (0.50)	>0.99
RBP4 (10:95110964-95610964)	rs10882283	Optic disc area, retinol (Vitamin A)	>0.99/>0.99	rs10882283 (-)	>0.99
SLC7A9 (19:32932750-34004548)	rs7247977	Chronic kidney disease, homocitrulline	0.99/0.95	rs7247977 (-)	0.38
TH (11:1735287-2479759)	rs10840516	Pulse rate, 3-methoxytyrosine, dopamine sulfate (2)	0.99/0.79	rs11564705 (0.97)	0.3644
UGT1A1/UGT1A3/UGT1A4/UGT1A5/UGT1A6/UGT1A7/UGT1A8/UGT1A9/UGT1A10 (2:233892619-235187605)	rs1976391, rs201829156	Cholelithiasis and cholecystitis, Skin colour, Disorders of gallbladder, biliary tract and pancreas, Operation code: cholecystectomy/gall bladder removal, Ease of skin tanning, p-cresol-glucuronide*, bilirubin (Z,Z), bilirubin (E,E)*	0.98/0.88	rs887829	1

Ch5_ST5: Clinical annotations for nine IFVs in the PharmGKB database. Chr = Chromosome, Pos = Position, EA = Effect Allele, OA = Other Allele, EAF = EA frequency, ADR = Adverse drug reaction, PK = Pharmacokinetics. Tiers are based on PharmGKB's tiers of evidence[225].

Annotated IEM genes at locus	IFV (Chr:Pos:EA:OA)	EAF	Chemical	Phenotypes	Type of response	Tier
APOE/APOC2	rs429358 (19: 45411941:C:T)	0.15	Simvastatin	Myocardial Infarction	Efficacy,	3
			Acenocoumarol	Haemorrhage	Toxicity/ADR	3
			Warfarin	Venous thromboembolism	Efficacy	3
			HmG CoA reductase inhibitors	Alzheimer's disease	Efficacy	3
			Antivirals for treatment of HIV infections/combinations/ritonavir	HIV infections, hyperlipidemias and hypertriglyceridemia	Toxicity/ADR	3
APOE/APOC2	rs7412 (19: 45412079:T:C)	0.081	Atorvastatin	Coronary disease, Hyperlipidemias	Efficacy	2
			Warfarin	Hyperlipidemias	Efficacy	NA
			Fenofibrate	Venous thromboembolism	Efficacy	3
			Fluvastatin	Hypertriglyceridemia	Efficacy	3
			Antivirals for treatment of HIV infections/combinations/ritonavir	HIV infections, hyperlipidemias and hypertriglyceridemia	Toxicity/ADR	3
			Pravastatin	Coronary disease, Hyperlipidemias	Efficacy	3
CTH	rs1021737 (1:70904800:T:G)	0.29	Busulfan and cyclophosphamide	Hemopoietic stem cell transplant	Toxicity/ADR	3
DPYD	rs3918290 (1: 97915614:T:C)	0.0039	Capecitabine, fluorouracil and pyrimidine analogues	Neoplasms	Toxicity/ADR	1A
			Tegafur	Neoplasms	Toxicity/ADR	3
			Fluorouracil	Neoplasms	Efficacy	3
DPYD	rs67376798 (1: 97547947:A:T)	0.0072	Capecitabine, fluorouracil and pyrimidine analogues	Neoplasms	Toxicity/ADR	1A
MTHFR	rs1801133 (1: 11856378:A:G)	0.34	Tegafur	Neoplasms	Toxicity/ADR	3
SLCO1B1/SLCO1B3	rs2291075 (12:21331625:T:C)	0.39	Cytarabine, daunorubicin, etoposide, mitoxantrone	Acute myeloid leukaemia	Efficacy	3
SLCO1B1/SLCO1B3	rs4149056 (12:21331549:C:T)	0.15	Simvastatin	Muscular diseases, Central core myopathy	Toxicity/ADR	1A
UGT1A1	rs1976391 (2: 234665983:G:A)	0.31	Risperidone	Hyperprolactinemia	Toxicity/ADR	3
			Oxazepam	Treatment of anxiety and insomnia, Control of alcohol withdrawal symptoms	Metabolism/ PK	4
UGT1A1	rs3732218 (2: 234627304:A:G)	0.084	Anastrozole	Breast cancer	Other	4
UGT1A1	rs72551330 (2: 234580678:C:T)	0.014	Mycophenolate mofetil	Kidney transplantation	Toxicity/ADR	3
			Sulfinpyrazone	Gout	Metabolism/ PK	4

Ch6_ST1: Summary statistics for the top 20 tested variant-PheRS associations with the largest odds ratios and the numbers of high PheRS 'cases' and 'controls' by variant carrier status. EA = Effect allele, OA = Other allele, EAF = EA frequency in UK Biobank.

IEM gene	IFV (EA/OA; EAF)	PheRS	OR (95% CI)	FDR-adjusted p-value	Carriers of effect allele (% homozygotes)		Homozygotes for other allele	
					High PheRS	Controls	High PheRS	Controls
PCSK9	rs11591147 (G/T; 0.98)	Homozygous Familial Hypercholesterolemia	2 (1.03; 3.88)	0.37	57816 (97.12)	10	294055 (96.44)	106
APOE	rs429358 (C/T; 0.16)	Alzheimer Disease 4	1.97 (1.82; 2.13)	1.5x10^{-59}	1086 (19.8)	1397	99760 (8.3)	249744
APOE	rs429358 (C/T; 0.16)	Alzheimer Disease 2	1.8 (1.67; 1.95)	1.8x10^{-46}	1074 (18.25)	1510	99772 (8.32)	249631
ACY1	rs189171677 (T/C; 0.00018)	Neurological Conditions Associated With Aminoacylase 1 Deficiency	1.74 (0.88; 3.43)	0.59	9 (0)	16134	116 (0)	335577
DDC	rs930707 (G/A; 0.98)	Aromatic L-amino Acid Decarboxylase Deficiency	1.72 (0.75; 3.93)	0.61	36837 (96.78)	6	314728 (96.89)	91
APOE	rs7412 (C/T; 0.92)	Alzheimer Disease 2	1.61 (0.86; 3.01)	0.6	2574 (87.84)	10	347164 (85.16)	2239
SLC5A1	rs117086479 (A/G; 0.92)	Glucose/galactose Malabsorption	1.57 (0.91; 2.72)	0.57	3576 (85.82)	13	343247 (86.02)	1936
LDLR	rs118068660 (C/T; 0.91)	Homozygous Familial Hypercholesterolemia	1.21 (1.08; 1.35)	0.02	55398 (84.5)	383	280387 (82.77)	2298
LIPC	rs121912502 (T/C; 0.0014)	Hyperlipidemia Due To Hepatic Triacylglycerol Lipase Deficiency	1.21 (0.96; 1.51)	0.56	90 (0)	24988	917 (0.11)	325992
APOC2	rs5112 (G/C; 0.54)	Apolipoprotein C-ii Deficiency	1.2 (1.03; 1.4)	0.18	922 (36.98)	205	232136 (35.83)	61934
GGT1	rs2236626 (T/C; 0.78)	Glutathionuria	0.84 (0.54; 1.3)	0.82	343 (64.14)	21	307959 (63.54)	15801
SCO2	rs74479613 (C/T; 0.91)	Cardioencephalomyopathy, Fatal Infantile, Due To Cytochrome C Oxidase Deficiency 1	0.81 (0.59; 1.1)	0.6	4220 (83.27)	42	344866 (83.76)	2738
GATM	rs1145091 (T/C; 0.74)	Cerebral Creatine Deficiency Syndrome 3	0.8 (0.59; 1.1)	0.6	478 (61.09)	42	328497 (59.24)	22951
UGT1A1	rs72551330 (C/T; 0.013)	Crigler-najjar Syndrome Type 1	0.8 (0.69; 0.93)	0.07	168 (1.19)	8246	8562 (0.64)	334516
ACY1	rs187813956 (T/C; 0.0013)	Neurological Conditions Associated With Aminoacylase 1 Deficiency	0.78 (0.55; 1.11)	0.6	33 (0)	16053	887 (0)	333759
ACY1	rs187092353 (T/C; 0.0015)	Neurological Conditions Associated With Aminoacylase 1 Deficiency	0.78 (0.56; 1.08)	0.6	37 (0)	16087	996 (0)	334210
APOE	rs7412 (C/T; 0.92)	Lipoprotein Glomerulopathy	0.71 (0.43; 1.17)	0.6	1826 (86.36)	16	347912 (85.18)	2233
APOC2	rs7412 (C/T; 0.92)	Apolipoprotein C-ii Deficiency	0.7 (0.4; 1.24)	0.64	1344 (85.19)	12	348394 (85.18)	2237
ACY1	rs545740325 (A/G; 0.0011)	Neurological Conditions Associated With Aminoacylase 1 Deficiency	0.66 (0.43; 0.99)	0.38	23 (0)	16112	745 (0.27)	334825
GGT1	rs186765281 (G/A; 0.016)	Glutathionuria	0.63 (0.31; 1.27)	0.61	8 (0)	394	10928 (0.75)	340657

REFERENCES

1. Wishart, D. S. *et al.* HMDB 4.0: The human metabolome database for 2018. *Nucleic Acids Res.* **46**, D608–D617 (2018).

2. Berg, J. M., Tymoczko, J. L. & Stryer, L. Glycolysis Is an Energy-Conversion Pathway in Many Organisms. in *Biochemistry* (W H Freeman, 2002).

3. Scheffler, I. E. *Mitochondria*. (1999).

4. Alam, M. M., Lal, S., FitzGerald, K. E. & Zhang, L. A holistic view of cancer bioenergetics: mitochondrial function and respiration play fundamental roles in the development and progression of diverse tumors. *Clin. Transl. Med.* **5**, (2016).

5. Wishart, D. S. Emerging applications of metabolomics in drug discovery and precision medicine. *Nature Reviews Drug Discovery* vol. 15 473–484 (2016).

6. Miller, J. M. Chromatography. in *Encyclopedia of Applied Spectroscopy* 1055–1102 (Wiley-VCH Verlag GmbH & Co. KGaA, 2009). doi:10.1002/3527600434.eap064.pub2.

7. de Hoffmann, E. Mass Spectrometry. in *Kirk-Othmer Encyclopedia of Chemical Technology* (John Wiley & Sons, Inc., 2005). doi:10.1002/0471238961.1301191913151518.a01.pub2.

8. National Research Council (US) Committee. Isolating, Identifying, Imaging, and Measuring Substances and Structures. in *Beyond the Molecular Frontier: Challenges for Chemistry and Chemical Engineering* (National Academies Press (US), 2003).

9. Cirulli, E. T. *et al.* Profound Perturbation of the Metabolome in Obesity Is Associated with Health Risk. *Cell Metab.* **29**, 488-500.e2 (2019).

10. Garrod, A. E. The incidence of alkaptonuria: a study in chemical individuality. *Lancet* **ii**, 1616–1620 (1902).

11. Vallejo-Torres, L. *et al.* Cost-Effectiveness Analysis of a National Newborn Screening Program for Biotinidase Deficiency. *Pediatrics* **136**, e424–e432 (2015).

12. Jacob, H. Update on expanded newborn screening. *Arch. Dis. Child. Educ. Pract. Ed.* **101**, 139 (2016).

13. Lichter-Konecki, U., Hipke, C. M. & Konecki, D. S. Human phenylalanine hydroxylase gene expression in kidney and other nonhepatic tissues. *Mol. Genet. Metab.* **67**, 308–316 (1999).

14. Zurflüh, M. R. *et al.* Molecular genetics of tetrahydrobiopterin-responsive phenylalanine hydroxylase deficiency. *Hum. Mutat.* **29**, 167–175 (2008).

15. MacLeod, E. L. & Ney, D. M. Nutritional management of phenylketonuria. *Annales Nestle* vol. 68 58–69 (2010).

16. Gieger, C. *et al.* Genetics Meets Metabolomics: A Genome-Wide Association Study of Metabolite Profiles in Human Serum. *PLoS Genet.* **4**, e1000282 (2008).

17. Suhre, K. *et al.* A genome-wide association study of metabolic traits in human urine. *Nat. Genet.* **43**, 565 (2011).

18. Rhee, E. P. *et al.* A genome-wide association study of the human metabolome in a community-based cohort. *Cell Metab.* **18**, 130–143 (2013).

19. Shin, S. Y. *et al.* An atlas of genetic influences on human blood metabolites. *Nat. Genet.* **46**, 543–550 (2014).

20. Long, T. *et al.* Whole-genome sequencing identifies common-to-rare variants associated with human blood metabolites. *Nat. Genet.* **49**, 568–578 (2017).

21. Lotta, L. A. *et al.* A cross-platform approach identifies genetic regulators of human metabolism and health. *Nat. Genet.* **53**, 54–64 (2021).

22. Rhee, E. P. *et al.* An exome array study of the plasma metabolome. *Nat. Commun.* **7**, 12360 (2016).

23. Lamy, P., Grove, J. & Wiuf, C. A review of software for microarray genotyping. *Human Genomics* vol. 5 304–309 (2011).

24. Kastenmüller, G., Raffler, J., Gieger, C. & Suhre, K. Genetics of human metabolism: an update. *Hum. Mol. Genet.* **24**, R93–R101 (2015).

25. Suhre, K. A Table of all published GWAS with metabolomics – Human Metabolic Individuality. http://www.metabolomix.com/list-of-all-published-gwas-with-metabolomics/ (2015).

26. Illig, T. *et al.* A genome-wide perspective of genetic variation in human metabolism. *Nat. Genet.* **42**, 137–141 (2010).

27. Kettunen, J. *et al.* Genome-wide association study identifies multiple loci influencing human serum metabolite levels. *Nat. Genet.* **44**, 269–276 (2012).

28. Suhre, K. *et al.* Human metabolic individuality in biomedical and pharmaceutical research.

Nature **477**, 54–62 (2011).

29. Krumsiek, J. *et al.* Mining the Unknown: A Systems Approach to Metabolite Identification Combining Genetic and Metabolic Information. *PLoS Genet.* **8**, e1003005 (2012).

30. Yu, B. *et al.* Genetic Determinants Influencing Human Serum Metabolome among African Americans. *PLoS Genet.* **10**, e1004212 (2014).

31. Yazdani, A., Yazdani, A., Liu, X. & Boerwinkle, E. Identification of Rare Variants in Metabolites of the Carnitine Pathway by Whole Genome Sequencing Analysis. *Genet. Epidemiol.* **40**, 486–491 (2016).

32. Yu, B. *et al.* Loss-of-function variants influence the human serum metabolome. *Sci. Adv.* **2**, e1600800 (2016).

33. Yousri, N. A. *et al.* Whole-exome sequencing identifies common and rare variant metabolic QTLs in a Middle Eastern population. *Nat. Commun.* **9**, 1–13 (2018).

34. Feofanova, E. V. *et al.* Sequence-based analysis of lipid-related metabolites in a multiethnic study. *Genetics* **209**, 607–616 (2018).

35. Yazdani, A. *et al.* Genome analysis and pleiotropy assessment using causal networks with loss of function mutation and metabolomics. *BMC Genomics* **20**, (2019).

36. Al-Khelaifi, F. *et al.* Metabolic profiling of elite athletes with different cardiovascular demand. *Scand. J. Med. Sci. Sports* **29**, sms.13425 (2019).

37. Schlosser, P. *et al.* Genetic studies of urinary metabolites illuminate mechanisms of detoxification and excretion in humans. *Nat. Genet.* **52**, 167–176 (2020).

38. Nag, A. *et al.* Genome-wide scan identifies novel genetic loci regulating salivary metabolite levels. *Hum. Mol. Genet.* **29**, 864–875 (2020).

39. Panyard, D. *et al.* Cerebrospinal fluid metabolomics identifies 19 brain-related phenotype associations. *bioRxiv* 2020.02.14.948398 (2020) doi:10.1101/2020.02.14.948398.

40. Qin, Y. *et al.* Genome-wide association and Mendelian randomization analysis prioritizes bioactive metabolites with putative causal effects on common diseases. (2020) doi:10.1101/2020.08.01.20166413.

41. Harshfield, E. L. *et al.* Genome-wide analysis of blood lipid metabolites in over 5,000 South Asians 1 reveals biological insights at cardiometabolic disease loci 2 3. *Syed Nadeem Hasan*

Rizvi **10**, 2020.10.16.20213520 (2020).

42. Liebsch, C. *et al.* The Saliva Metabolome in Association to Oral Health Status. *J. Dent. Res.* **98**, 642–651 (2019).

43. Köttgen, A., Raffler, J., Sekula, P. & Kastenmüller, G. Genome-Wide Association Studies of Metabolite Concentrations (mGWAS): Relevance for Nephrology. *Seminars in Nephrology* vol. 38 151–174 (2018).

44. Suhre, K. & Gieger, C. Genetic variation in metabolic phenotypes: Study designs and applications. *Nature Reviews Genetics* vol. 13 759–769 (2012).

45. Hobbs, H. H., Brown, M. S. & Goldstein, J. L. Molecular genetics of the LDL receptor gene in familial hypercholesterolemia. *Hum. Mutat.* **1**, 445–466 (1992).

46. Humphries, S. E. *et al.* Genetic causes of familial hypercholesterolaemia in patients in the UK: relation to plasma lipid levels and coronary heart disease risk. *J. Med. Genet.* **43**, 943–9 (2006).

47. Civeira, F. *et al.* Comparison of Genetic Versus Clinical Diagnosis in Familial Hypercholesterolemia. *Am. J. Cardiol.* **102**, 1187-1193.e1 (2008).

48. Christiansen, M. K. *et al.* Coronary artery disease-associated genetic variants and biomarkers of inflammation. *PLoS One* **12**, e0180365 (2017).

49. *Familial hypercholesterolaemia: identification and management Clinical guideline*. www.nice.org.uk/guidance/cg71 (2008).

50. Hadfield, S. G. & Humphries, S. E. Implementation of cascade testing for the detection of familial hypercholesterolaemia. *Current Opinion in Lipidology* vol. 16 428–433 (2005).

51. Oyarzabal, A. *et al.* A Novel Regulatory Defect in the Branched-Chain α-Keto Acid Dehydrogenase Complex Due to a Mutation in the PPM1K Gene Causes a Mild Variant Phenotype of Maple Syrup Urine Disease. *Hum. Mutat.* **34**, 355–362 (2013).

52. Lotta, L. A. *et al.* Genetic Predisposition to an Impaired Metabolism of the Branched-Chain Amino Acids and Risk of Type 2 Diabetes: A Mendelian Randomisation Analysis. *PLOS Med.* **13**, e1002179 (2016).

53. Humm, A., Fritsche, E., Mann, K., Göhl, M. & Huber, R. Recombinant expression and isolation of human L-arginine:glycine amidinotransferase and identification of its active-site cysteine residue. *Biochem. J.* **322 (Pt 3)**, 771–6 (1997).

54. Lichter-Konecki, U., Broman, K. W., Blau, E. B. & Konecki, D. S. Genetic and physical mapping of the locus for autosomal dominant renal Fanconi syndrome, on chromosome 15q15.3. *Am. J. Hum. Genet.* **68**, 264–268 (2001).

55. Köttgen, A. *et al.* Multiple loci associated with indices of renal function and chronic kidney disease. *Nat. Genet.* **41**, 712–717 (2009).

56. Köttgen, A. *et al.* New loci associated with kidney function and chronic kidney disease. *Nat. Genet.* **42**, 376–384 (2010).

57. Hirschhorn, J. N. & Daly, M. J. Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics* vol. 6 95–108 (2005).

58. Altmüller, J., Palmer, L. J., Fischer, G., Scherb, H. & Wjst, M. Genomewide scans of complex human diseases: True linkage is hard to find. *American Journal of Human Genetics* vol. 69 936–950 (2001).

59. Reich, D. E. & Lander, E. S. On the allelic spectrum of human disease. *Trends Genet.* **17**, 502–510 (2001).

60. Pritchard, J. K. Are rare variants responsible for susceptibility to complex diseases? *Am. J. Hum. Genet.* **69**, 124–137 (2001).

61. Gorlov, I. P., Gorlova, O. Y., Sunyaev, S. R., Spitz, M. R. & Amos, C. I. Shifting Paradigm of Association Studies: Value of Rare Single-Nucleotide Polymorphisms. *Am. J. Hum. Genet.* **82**, 100–112 (2008).

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

63. Elsworth, B. *et al.* The MRC IEU OpenGWAS data infrastructure. *bioRxiv* 2020.08.10.244293 (2020) doi:10.1101/2020.08.10.244293.

64. Nickerson, K. P. & McDonald, C. Crohn's disease-associated adherent-invasive Escherichia coli adhesion is enhanced by exposure to the ubiquitous dietary polysaccharide maltodextrin. *PLoS One* **7**, e52132 (2012).

65. Bobadilla, J. L., Macek, M., Fine, J. P. & Farrell, P. M. Cystic fibrosis: A worldwide analysis of CFTR mutations - Correlation with incidence data and application to screening. *Human Mutation* vol. 19 575–606 (2002).

66. Szabo, C., Masiello, A., Ryan, J. F. & Brody, L. C. The Breast Cancer Information Core: Database design, structure, and scope. *Human Mutation* vol. 16 123–131 (2000).

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

68. Van Hout, C. V. *et al.* Exome sequencing and characterization of 49,960 individuals in the UK Biobank. *Nature* **586**, 749–756 (2020).

69. Jurgens, S. J. *et al.* Rare Genetic Variation Underlying Human Diseases and Traits: Results from 200,000 Individuals in the UK Biobank. *bioRxiv* 2020.11.29.402495 (2020) doi:10.1101/2020.11.29.402495.

70. Blair, D. R. *et al.* A Nondegenerate Code of Deleterious Variants in Mendelian Loci Contributes to Complex Disease Risk. *Cell* **155**, 70–80 (2013).

71. Freund, M. K. *et al.* Phenotype-specific enrichment of Mendelian disorder genes near GWAS regions across 62 complex traits. *Am. J. Hum. Genet.* **103**, 535–552 (2018).

72. Denny, J. C. *et al.* PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics* **26**, 1205–1210 (2010).

73. Kamat, M. A. *et al.* PhenoScanner V2: An expanded tool for searching human genotype-phenotype associations. *Bioinformatics* **35**, 4851–4853 (2019).

74. Ghoussaini, M. *et al.* Open Targets Genetics: systematic identification of trait-associated genes using large-scale genetics and functional genomics. *Nucleic Acids Res.* **49**, (2020).

75. Köhler, S. *et al.* Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. *Nucleic Acids Res.* **47**, D1018–D1027 (2019).

76. WHO | List of Official ICD-10 Updates. *WHO* (2017).

77. Bastarache, L. *et al.* Phenotype risk scores identify patients with unrecognized mendelian disease patterns. *Science (80-.).* **359**, 1233–1239 (2018).

78. Hill, J. O. & Peters, J. C. Environmental contributions to the obesity epidemic. *Science* vol. 280 1371–1374 (1998).

79. Hu, F. B. Sedentary lifestyle and risk of obesity and type 2 diabetes. in *Lipids* vol. 38 103–108 (Lipids, 2003).

80. Locke, A. E. *et al.* Genetic studies of body mass index yield new insights for obesity biology.

Nature **518**, 197–206 (2015).

81. Shungin, D. *et al.* New genetic loci link adipose and insulin biology to body fat distribution. *Nature* **518**, 187–196 (2015).

82. Pulit, S. L. *et al.* Meta-analysis of genome-wide association studies for body fat distribution in 694,649 individuals of European ancestry. *bioRxiv* 304030 (2018) doi:10.1101/304030.

83. Loos, R. J. F. & Kilpeläinen, T. O. Genes that make you fat, but keep you healthy. *J. Intern. Med.* **284**, 450–463 (2018).

84. Pingault, J. B. *et al.* Using genetic data to strengthen causal inference in observational research. *Nature Reviews Genetics* vol. 19 566–580 (2018).

85. Krumsiek, J., Suhre, K., Illig, T., Adamski, J. & Theis, F. J. Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data. *BMC Syst. Biol.* **5**, 21 (2011).

86. Newcombe, P. J., Connolly, S., Seaman, S., Richardson, S. & Sharp, S. J. A two-step method for variable selection in the analysis of a case-cohort study. *Int. J. Epidemiol.* **47**, 597–604 (2018).

87. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, (2016).

88. Guidi, G. C. & Salvagno, G. L. Reference intervals as a tool for total quality management. *Biochemia Medica* vol. 20 165–172 (2010).

89. Horn, P. S. & Pesce, A. J. Reference intervals: An update. *Clinica Chimica Acta* vol. 334 5–23 (2003).

90. Koerbin, G. *et al.* Effect of population selection on 99th percentile values for a high sensitivity cardiac troponin I and T assays. *Clin. Biochem.* **46**, 1636–1643 (2013).

91. Obesity and overweight. https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight.

92. Health England, P. *Adult obesity and type 2 diabetes*. https://www.gov.uk/government/publications/adult-obesity-and-type-2-diabetes (2014).

93. Morris, C. *et al.* The relationship between BMI and metabolomic profiles: A focus on amino acids. in *Proceedings of the Nutrition Society* vol. 71 634–638 (Proc Nutr Soc, 2012).

94. Chen, H.-H. *et al.* The metabolome profiling and pathway analysis in metabolic healthy and abnormal obesity. *Int. J. Obes.* **39**, 1241–1248 (2015).

95. Zhao, H. *et al.* Metabolomics-identified metabolites associated with body mass index and prospective weight gain among Mexican American women. *Obes. Sci. Pract.* **2**, 309–317 (2016).

96. Würtz, P. *et al.* Metabolic Signatures of Adiposity in Young Adults: Mendelian Randomization Analysis and Effects of Weight Change. *PLoS Med.* **11**, (2014).

97. Wahl, S. *et al.* Multi-omic signature of body weight change: Results from a population-based cohort study. *BMC Med.* **13**, (2015).

98. Park, S., Sadanala, K. C. & Kim, E. K. A metabolomic approach to understanding the metabolic link between obesity and diabetes. *Molecules and Cells* vol. 38 587–596 (2014).

99. Xu, F. *et al.* Metabolic signature shift in type 2 diabetes mellitus revealed by mass spectrometry-based metabolomics. *J. Clin. Endocrinol. Metab.* **98**, (2013).

100. Suhre, K. *et al.* Metabolic footprint of diabetes: A multiplatform metabolomics study in an epidemiological setting. *PLoS One* **5**, (2010).

101. Menni, C. *et al.* Biomarkers for type 2 diabetes and impaired fasting glucose using a nontargeted metabolomics approach. *Diabetes* **62**, 4270–6 (2013).

102. Drogan, D. *et al.* Untargeted metabolic profiling identifies altered serum metabolites of type 2 diabetes mellitus in a prospective, nested case control study. *Clin. Chem.* **61**, 487–497 (2015).

103. Crawford, S. O. *et al.* Association of blood lactate with type 2 diabetes: The atherosclerosis risk in communities carotid MRI study. *Int. J. Epidemiol.* **39**, 1647–1655 (2010).

104. Zhang, A. H. *et al.* Metabolomics study of type 2 diabetes using ultra-performance LC-ESI/quadrupole-TOF high-definition MS coupled with pattern recognition methods. *J. Physiol. Biochem.* **70**, 117–128 (2014).

105. Libert, D. M., Nowacki, A. S. & Natowicz, M. R. Metabolomic analysis of obesity, metabolic syndrome, and type 2 diabetes: Amino acid and acylcarnitine levels change along a spectrum of metabolic wellness. *PeerJ* **2018**, (2018).

106. Rebholz, C. M. *et al.* Serum metabolomic profile of incident diabetes. *Diabetologia* **61**, 1046–1054 (2018).

107. Day, N. *et al. EPIC-Norfolk: study design and characteristics of the cohort*. *BHIISh Journalof Cancer* http://www.srl.cam.ac.uk/epic/publications/day_bjc_1999_clean.pdf (1999).

108. Langenberg, C. *et al.* Gene-Lifestyle Interaction and Type 2 Diabetes: The EPIC InterAct Case-Cohort Study. *PLoS Med.* **11**, e1001647 (2014).

109. McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279–83 (2016).

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

111. Walter, K. *et al.* The UK10K project identifies rare variants in health and disease. *Nature* **526**, 82–89 (2015).

112. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B* **57**, 289–300 (1995).

113. Pietzner, M. *et al.* Small molecule profiling identifies shared and distinct pathways to non-communicable disease multimorbidity. *Accept. Nat. Med. (or Press.* (2020).

114. Scott, R. A. *et al.* Large-scale association analyses identify new loci influencing glycemic traits and provide insight into the underlying biological pathways. *Nat. Genet. 2012 449* **44**, 991 (2012).

115. Yengo, L. *et al.* Meta-analysis of genome-wide association studies for height and body mass index in ∼700000 individuals of European ancestry. *Hum. Mol. Genet.* (2018) doi:10.1093/hmg/ddy271.

116. Pulit, S. L. *et al.* Meta-Analysis of genome-wide association studies for body fat distribution in 694 649 individuals of European ancestry. *Hum. Mol. Genet.* **28**, 166–174 (2019).

117. Lu, Y. *et al.* New loci for body fat percentage reveal link between adiposity and cardiometabolic disease risk. *Nat. Commun.* **7**, 10495 (2016).

118. Speliotes, E. K. *et al.* Genome-Wide Association Analysis Identifies Variants Associated with Nonalcoholic Fatty Liver Disease That Have Distinct Effects on Metabolic Traits. *PLoS Genet.* **7**, e1001324 (2011).

119. van Buuren, S. & Groothuis-Oudshoorn, K. mice: Multivariate imputation by chained equations in R. *J. Stat. Softw.* **45**, 1–67 (2011).

120. Opgen-Rhein, R. & Strimmer, K. Inferring Gene Dependency Networks from Genomic Longitudinal Data: A Functional Data Approach. *REVSTAT* **4**, 53–65 (2006).

121. Schäfer, J. & Strimmer, K. A Shrinkage Approach to Large-Scale Covariance Matrix Estimation

and Implications for Functional Genomics A Shrinkage Approach to Large-Scale Covariance Matrix Estimation and Implications for Functional Genomics *. *Stat. Appl. Genet. Mol. Biol.* **4**, (2005).

122. Opgen-Rhein, R. & Strimmer, K. Using Regularized Dynamic Correlation to Infer Gene Dependency Networks from Time-series Microarray Data. in *4th International Workshop on Computational Systems Biology (WCSB 2006)* (2006).

123. Revelle, W. R. psych: Procedures for Personality and Psychological Research. (2017).

124. Rubin, D. B. *Multiple Imputation for Nonresponse in Surveys*. (John Wiley & Sons, 2004).

125. Schwarzer, G. meta: An R Package for Meta-Analysis. *R News* **7**, 40–45 (2007).

126. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–504 (2003).

127. Csárdi, G. & Nepusz, T. The igraph software package for complex network research. *InterJournal* **Complex Systems**, 1695 (2006).

128. Eddelbuettel, D. & François, R. **Rcpp** : Seamless *R* and *C++* Integration. *J. Stat. Softw.* **40**, 1–18 (2011).

129. StataCorp. Stata Statistical Software: Release 14. (2015).

130. Mardinoglu, A. *et al.* Plasma Mannose Levels Are Associated with Incident Type 2 Diabetes and Cardiovascular Disease. *Cell Metab.* **26**, 281–283 (2017).

131. Skoglund, L. A., Ingebrigtsen, K. & Nafstad, I. Effects of N-acetyl-DL-methionine on the liver, GSH synthesis and plasma ALAT level in male Bom:NMRI mice. *Gen. Pharmacol.* **17**, 647–9 (1986).

132. Lertratanangkoon, K., Scimeca, J. M. & Wei, J. Inhibition of Glutathione Synthesis with Propargylglycine Enhances N-Acetylmethionine Protection and Methylation in Bromobenzene-Treated Syrian Hamsters. *J. Nutr.* **129**, 649–656 (1999).

133. Hammarstedt, A. *et al.* Adipose tissue dysfunction is associated with low levels of the novel Palmitic Acid Hydroxystearic Acids. *Sci. Rep.* **8**, (2018).

134. Sekula, P. *et al.* A metabolome-wide association study of kidney function and disease in the general population. *J. Am. Soc. Nephrol.* **27**, 1175–1188 (2016).

135. Schleicher, E. D., Wagner, E. & Nerlich, A. G. Increased accumulation of the glycoxidation

product N(ε)- (carboxymethyl)lysine in human tissues in diabetes and aging. *J. Clin. Invest.* **99**, 457–468 (1997).

136. Matsushita, K. *et al.* Incorporating kidney disease measures into cardiovascular risk prediction: Development and validation in 9 million adults from 72 datasets. *EClinicalMedicine* **27**, 100552–100552 (2020).

137. Lu, J. *et al.* Reduced Kidney Function Is Associated With Cardiometabolic Risk Factors, Prevalent and Predicted Risk of Cardiovascular Disease in Chinese Adults: Results From the REACTION Study. *J. Am. Heart Assoc.* **5**, (2016).

138. Dakhale, G. N., Chaudhari, H. V. & Shrivastava, M. Supplementation of vitamin C reduces blood glucose and improves glycosylated hemoglobin in type 2 diabetes mellitus: A randomized, double-blind study. *Adv. Pharmacol. Pharm. Sci.* (2011) doi:10.1155/2011/195271.

139. Li, M., Fan, Y., Zhang, X., Hou, W. & Tang, Z. Fruit and vegetable intake and risk of type 2 diabetes mellitus: meta-analysis of prospective cohort studies. *BMJ Open* **4**, e005497 (2014).

140. McRae, M. P. Dietary Fiber Intake and Type 2 Diabetes Mellitus: An Umbrella Review of Meta-analyses. *Journal of Chiropractic Medicine* vol. 17 44–53 (2018).

141. Forouhi, N. G., Misra, A., Mohan, V., Taylor, R. & Yancy, W. Dietary and nutritional approaches for prevention and management of type 2 diabetes. *BMJ* **361**, (2018).

142. Angelakis, E., Armougom, F., Million, M. & Raoult, D. The relationship between gut microbiota and weight gain in humans. *Future Microbiology* vol. 7 91–109 (2012).

143. Zhao, M. *et al.* TMAVA, a Metabolite of Intestinal Microbes, Is Increased in Plasma From Patients With Liver Steatosis, Inhibits γ-Butyrobetaine Hydroxylase, and Exacerbates Fatty Liver in Mice. *Gastroenterology* **158**, 2266-2281.e27 (2020).

144. Lustgarten, M. S., Price, L. L., Chalé, A. & Fielding, R. A. Metabolites related to gut bacterial metabolism, peroxisome proliferator-activated receptor-alpha activation, and insulin sensitivity are associated with physical function in functionally-limited older adults. *Aging Cell* **13**, 918–25 (2014).

145. Gao, J. *et al.* Sex-specific effect of estrogen sulfotransferase on mouse models of type 2 diabetes. *Diabetes* **61**, 1543–1551 (2012).

146. Jiang, M. *et al.* Hepatic overexpression of steroid sulfatase ameliorates mouse models of

obesity and type 2 diabetes through sex-specific mechanisms. *J. Biol. Chem.* **289**, 8086–8097 (2014).

147. Ferrell, J. M. & Chiang, J. Y. L. Understanding bile acid signaling in diabetes: From pathophysiology to therapeutic targets. *Diabetes and Metabolism Journal* vol. 43 257–272 (2019).

148. Baron, R. M. & Kenny, D. A. *The Moderator-Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations*. vol. 51 (1986).

149. Smith, G. D. & Ebrahim, S. Mendelian Randomization: Genetic Variants as Instruments for Strengthening Causal Inference in Observational Studies. (2008).

150. Stevens, V. L., Hoover, E., Wang, Y. & Zanetti, K. A. Pre-analytical factors that affect metabolite stability in human urine, plasma, and serum: A review. *Metabolites* vol. 9 (2019).

151. Berg, J. M., Tymoczko, J. L. & Stryer, L. Each Organ Has a Unique Metabolic Profile. in *Biochemistry* (W H Freeman, 2002).

152. Ly-Verdú, S. *et al.* The impact of blood on liver metabolite profiling - A combined metabolomic and proteomic approach. *Biomed. Chromatogr.* **28**, 231–240 (2014).

153. Sanderson, S., Green, A., Preece, M. A. & Burton, H. The incidence of inherited metabolic disorders in the West Midlands, UK. *Arch. Dis. Child.* **91**, 896–9 (2006).

154. Campeau, P. M., Scriver, C. R. & Mitchell, J. J. A 25-year longitudinal analysis of treatment efficacy in inborn errors of metabolism. *Mol. Genet. Metab.* **95**, 11–16 (2008).

155. Pollitt, R. J. *et al.* Neonatal screening for inborn errors of metabolism: cost, yield and outcome. *Health technology assessment (Winchester, England)* vol. 1 i–iv, 1 (1997).

156. Cipriano, L. E., Rupar, C. A. & Zaric, G. S. The cost-effectiveness of expanding newborn screening for up to 21 inherited metabolic disorders using tandem mass spectrometry: Results from a decision-analytic model. *Value Heal.* **10**, 83–97 (2007).

157. Familial hypercholesterolaemia: identification and management. *NICE Guid.* (2008).

158. Häberle, J. *et al.* Molecular defects in human carbamoy phosphate synthetase I: mutational spectrum, diagnostic and protein structure considerations. *Hum. Mutat.* **32**, 579–589 (2011).

159. Klaus, V. *et al.* Highly variable clinical phenotype of carbamylphosphate synthetase 1 deficiency in one family: an effect of allelic variation in gene expression? *Clin. Genet.* **76**, 263–

269 (2009).

160. Kikuchi, G., Motokawa, Y., Yoshida, T. & Hiraga, K. Glycine cleavage system: Reaction mechanism, physiological significance, and hyperglycinemia. *Proceedings of the Japan Academy Series B: Physical and Biological Sciences* vol. 84 246–263 (2008).

161. Surendran, P. & Stewart, I. D. Large-scale GWAS of human plasma metabolome. (Abstract 91/Program 32). Presented at the 68th Annual Meeting of The American Society of Human Genetics. in (2018).

162. Lee, J. J. Y., Wasserman, W. W., Hoffmann, G. F., Van Karnebeek, C. D. M. & Blau, N. Knowledge base and mini-expert platform for the diagnosis of inborn errors of metabolism. *Genet. Med.* **20**, 151–158 (2018).

163. Moore, C. *et al.* The INTERVAL trial to determine whether intervals between blood donations can be safely and acceptably decreased to optimise blood supply: study protocol for a randomised controlled trial. *Trials* **15**, 363 (2014).

164. Moore, C. *et al.* Recruitment and representativeness of blood donors in the INTERVAL randomised trial assessing varying inter-donation intervals. *Trials* **17**, 458 (2016).

165. Listgarten, J. *et al.* Improved linear mixed models for genome-wide association studies. *Nature Methods* vol. 9 525–526 (2012).

166. Zhou, X. & Stephens, M. Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.* **44**, 821–824 (2012).

167. Voight, B. F. & Pritchard, J. K. Confounding from cryptic relatedness in case-control association studies. *PLoS Genet.* **1**, e32–e32 (2005).

168. Loh, P. R., Kichaev, G., Gazal, S., Schoech, A. P. & Price, A. L. Mixed-model association for biobank-scale datasets. *Nature Genetics* vol. 50 906–908 (2018).

169. Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat. Genet.* **39**, 906–913 (2007).

170. Fleiss, J. L. The statistical basis of meta-analysis. *Statistical methods in medical research* vol. 2 121–145 (1993).

171. de Bakker, P. I. W. *et al.* Practical aspects of imputation-driven meta-analysis of genome-wide association studies. *Hum. Mol. Genet.* **17**, (2008).

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

173. Amberger, J. S., Bocchini, C. A., Schiettecatte, F., Scott, A. F. & Hamosh, A. OMIM.org: Online Mendelian Inheritance in Man (OMIM(R)), an online catalog of human genes and genetic disorders. *Nucleic Acids Res.* **43**, D789–D798 (2015).

174. Bateman, A. UniProt: A worldwide hub of protein knowledge. *Nucleic Acids Res.* **47**, D506–D515 (2019).

175. Yates, A. D. *et al.* Ensembl 2020. *Nucleic Acids Res.* **48**, D682–D688 (2020).

176. Ashburner, M. *et al.* Gene ontology: Tool for the unification of biology. *Nature Genetics* vol. 25 25–29 (2000).

177. Carbon, S. *et al.* The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res.* **47**, D330–D338 (2019).

178. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* **44**, D457–D462 (2016).

179. Arnold, M., Raffler, J., Pfeufer, A., Suhre, K. & Kastenmüller, G. SNiPA: An interactive, genetic variant-centered annotation browser. *Bioinformatics* **31**, 1334–1336 (2015).

180. Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* **22**, 1760–74 (2012).

181. Rath, A. *et al.* Representation of rare diseases in health information systems: The orphanet approach to serve a wide range of end users. *Hum. Mutat.* **33**, 803–808 (2012).

182. Landrum, M. J. *et al.* ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res.* **44**, D862–D868 (2016).

183. Qi, Q. *et al.* FTO genetic variants, dietary intake and body mass index: insights from 177,330 individuals. *Hum. Mol. Genet.* **23**, 6961–6972 (2014).

184. Gaulton, A. *et al.* The ChEMBL database in 2017. *Nucleic Acids Res.* **45**, D945–D954 (2017).

185. Kim, S. *et al.* PubChem in 2021: new data content and improved web interfaces. *Nucleic Acids Res.* **49**, D1388–D1395 (2021).

186. Peplow, M. The 100 000 genomes project. *BMJ* **353**, (2016).

187. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in

141,456 humans. *Nature* **581**, 434–443 (2020).

188. Biaggioni, I., Goldstein, D. S., Atkinson, T. & Robertson, D. Dopamine-beta-hydroxylase deficiency in humans. *Neurology* **40**, 370–3 (1990).

189. Dunnette, J. & Weinshilboum, R. Human serum dopamine β hydroxylase: correlation of enzymatic activity with immunoreactive protein in genetically defined samples. *Am. J. Hum. Genet.* **28**, 155–166 (1976).

190. Zabetian, C. P. *et al.* A quantitative-trait analysis of human plasma-dopamine β-hydroxylase activity: Evidence for a major functional polymorphism at the DBH locus. *Am. J. Hum. Genet.* **68**, 515–522 (2001).

191. Brunetti-Pierri, N., Parenti, G. & Andria, G. Inborn Errors of Metabolism and Newborns. in *Neonatology* 1805–1832 (Springer International Publishing, 2018). doi:10.1007/978-3-319-29489-6_258.

192. Cordell, H. J. Epistasis: What it means, what it doesn't mean, and statistical methods to detect it in humans. *Human Molecular Genetics* vol. 11 2463–2468 (2002).

193. Waddington, C. H. Canalization of development and genetic assimilation of acquired characters. *Nature* **183**, 1654–1655 (1959).

194. Pozarickij, A., Williams, C. & Guggenheim, J. A. Non-additive (dominance) effects of genetic variants associated with refractive error and myopia. *Mol. Genet. Genomics* **295**, 843–853 (2020).

195. Wuttke, M. *et al.* A catalog of genetic loci associated with kidney function from analyses of a million individuals. *Nat. Genet.* **51**, 957–972 (2019).

196. Pagonas, N. *et al.* Volatile Organic Compounds in Uremia. *PLoS One* **7**, (2012).

197. Vallet, M. *et al.* Urinary ammonia and long-term outcomes in chronic kidney disease. *Kidney Int.* **88**, 137–145 (2015).

198. Jethva, R., Bennett, M. J. & Vockley, J. Short-chain acyl-coenzyme A dehydrogenase deficiency. *Molecular Genetics and Metabolism* vol. 95 195–200 (2008).

199. Corydon, M. J. *et al.* Ethylmalonic aciduria is associated with an amino acid variant of short chain acyl-coenzyme A dehydrogenase. *Pediatr. Res.* **39**, 1059–1066 (1996).

200. Lee, J. J. *et al.* Gene discovery and polygenic prediction from a genome-wide association

study of educational attainment in 1.1 million individuals. *Nat. Genet.* **50**, 1112–1121 (2018).

201. Roulstone, S., Law, J., Rush, R., Clegg, J. & Peters, T. *Investigating the role of language in children's early educational outcomes*.

202. Ritter, J. K. *et al.* A novel complex locus UGT1 encodes human bilirubin, phenol, and other UDP- glucuronosyltransferase isozymes with identical carboxyl termini. *J. Biol. Chem.* **267**, 3257–3261 (1992).

203. Crigler, J. F. & Najjar, V. A. Congenital Familial Nonhemolytic Jaundice with Kernicterus. *Pediatrics* **10**, (1952).

204. Bosma, P. J. *et al.* The Genetic Basis of the Reduced Expression of Bilirubin UDP-Glucuronosyltransferase 1 in Gilbert's Syndrome. *N. Engl. J. Med.* **333**, 1171–1175 (1995).

205. Koiwai, O. *et al.* Gilbert's syndrome is caused by a heterozygous missense mutation in the gene for bilirubin UDP-glucuronosyltransferase. *Hum. Mol. Genet.* **4**, 1183–1186 (1995).

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

207. Johnson, D. H. *et al.* Genomewide association study of atazanavir pharmacokinetics and hyperbilirubinemia in AIDS Clinical Trials Group protocol A5202. *Pharmacogenet. Genomics* **24**, 195–203 (2014).

208. Vardhanabhuti, S. *et al.* Screening for UGT1A1 Genotype in Study A5257 Would Have Markedly Reduced Premature Discontinuation of Atazanavir for Hyperbilirubinemia. *Open Forum Infect. Dis.* **2**, ofv085 (2015).

209. Sharp, P. A. Split genes and RNA splicing. *Cell* **77**, 805–815 (1994).

210. Maurano, M. T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science (80-.).* **337**, 1190–1195 (2012).

211. Sun, Y. V. *et al.* Effects of Genetic Variants Associated with Familial Hypercholesterolemia on Low-Density Lipoprotein-Cholesterol Levels and Cardiovascular Outcomes in the Million Veteran Program. *Circ. Genomic Precis. Med.* **11**, (2018).

212. Ademi, Z. *et al.* Cascade screening based on genetic testing is cost-effective: Evidence for the implementation of models of care for familial hypercholesterolemia. *J. Clin. Lipidol.* **8**, 390–400 (2014).

213. Kerr, M. *et al.* Cost effectiveness of cascade testing for familial hypercholesterolaemia, based on data from familial hypercholesterolaemia services in the UK. *Eur. Heart J.* **38**, 1832–1839 (2017).

214. Pietzner, M. *et al.* Genetic architecture of host proteins interacting with SARS-CoV-2. *bioRxiv Prepr. Serv. Biol.* 2020.07.01.182709 (2020) doi:10.1101/2020.07.01.182709.

215. Giambartolomei, C. *et al.* Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. *PLoS Genet.* **10**, e1004383 (2014).

216. Giambartolomei, C. *et al.* A Bayesian framework for multiple trait colocalization from summary association statistics. *Bioinformatics* **34**, 2538–2545 (2018).

217. Foley, C. *et al.* A fast and efficient colocalization algorithm for identifying shared genetic risk factors across multiple traits. *bioRxiv* 592238 (2019) doi:10.1101/592238.

218. Higgins, M. J. P., Lecamwasam, D. S. & Galton, D. J. A NEW TYPE OF FAMILIAL HYPERCHOLESTEROLÆMIA. *Lancet* **306**, 737–740 (1975).

219. Richardson, T. G. *et al.* Evaluating the relationship between circulating lipoprotein lipids and apolipoproteins with risk of coronary heart disease: A multivariable Mendelian randomisation analysis. *PLoS Med.* **17**, e1003062 (2020).

220. Taylor, R. G., Levy, H. L. & McInnes, R. R. Histidase and histidinemia. Clinical and molecular considerations. *Mol. Biol. Med.* 101–16 (1991).

221. Wain, L. V. *et al.* Novel Blood Pressure Locus and Gene Discovery Using Genome-Wide Association Study and Expression Data Sets from Blood and the Kidney. *Hypertension* **70**, e4–e19 (2017).

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

223. Astle, W. J. *et al.* The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. *Cell* **167**, 1415-1429.e19 (2016).

224. Wallace, C. Eliciting priors and relaxing the single causal variant assumption in colocalisation analyses. *PLoS Genet.* **16**, (2020).

225. Hewett, M. *et al.* PharmGKB: The pharmacogenetics knowledge base. *Nucleic Acids Res.* **30**, 163–165 (2002).

226. Finan, C. *et al.* The druggable genome and support for target identification and validation in drug development. *Sci. Transl. Med.* **9**, (2017).

227. Kim, C. H. *et al.* Mutations in the dopamine β-hydroxylase gene are associated with human norepinephrine deficiency. *Am. J. Med. Genet.* **108**, 140–147 (2002).

228. Tolleson, C. & Claassen, D. The Function of Tyrosine Hydroxylase in the Normal and Parkinsonian Brain. *CNS Neurol. Disord. - Drug Targets* **11**, 381–386 (2012).

229. Stamelou, M. *et al.* Myoclonus-dystonia syndrome due to tyrosine hydroxylase deficiency. *Neurology* **79**, 435–441 (2012).

230. Watanabe, T., Abe, K., Ishikawa, H. & Iijima, Y. Bovine 5-oxo-L-prolinase: Simple assay method, purification, cDNA cloning, and detection of mRNA in the coronary artery. *Biol. Pharm. Bull.* **27**, 288–294 (2004).

231. Pederzolli, C. D. *et al.* Acute administration of 5-oxoproline induces oxidative damage to lipids and proteins and impairs antioxidant defenses in cerebral cortex and cerebellum of young rats. *Metab. Brain Dis.* **25**, 145–154 (2010).

232. Silva, A. R. *et al.* L-pyroglutamic acid inhibits energy production and lipid synthesis in cerebral cortex of young rats in vitro. *Neurochem. Res.* **26**, 1277–1283 (2001).

233. Savage, J. E. *et al.* Genome-wide association meta-analysis in 269,867 individuals identifies new genetic and functional links to intelligence. *Nat. Genet.* **50**, 912–919 (2018).

234. Robertson, P. L., Buchanan, D. N. & Muenzer, J. 5-Oxoprolinuria in an adolescent with chronic metabolic acidosis, mental retardation, and psychosis. *J. Pediatr.* **118**, 92–95 (1991).

235. Calpena, E. *et al.* 5-oxoprolinuria in heterozygous patients for 5-oxoprolinase (OPLAH) missense changes. in *JIMD Reports* vol. 7 123–128 (Springer, 2013).

236. Marstein, S., Jellum, E., Halpern, B., Eldjarn, L. & Perry, T. L. Biochemical Studies of Erythrocytes in a Patient with Pyroglutamic Acidemia (5-Oxoprolinemia). *N. Engl. J. Med.* **295**, 406–412 (1976).

237. Grioli, S., Lomeo, C., Quattropani, M., Spignoli, G. & Villardita, C. Pyroglutamic acid improves the age associated memory impairment. *Fundam. Clin. Pharmacol.* **4**, 169–173 (1990).

238. Yi, L. *et al.* Serum Metabolic Profiling Reveals Altered Metabolic Pathways in Patients with Post-traumatic Cognitive Impairments. *Sci. Rep.* **6**, (2016).

239. J, Y. *et al.* Causal relationships between genetically determined metabolites and human intelligence: A Mendelian randomization study. (2020) doi:10.21203/RS.3.RS-29322/V3.

240. Maeso, N. *et al.* Capillary electrophoresis for caffeine and pyroglutamate determination in coffees. Study of the in vivo effect on learning and locomotor activity in mice. *J. Pharm. Biomed. Anal.* **41**, 1095–1100 (2006).

241. Dempsey, G. A., Lyall, H. J., Corke, C. F. & Scheinkestel, C. D. Pyroglutamic acidemia: A cause of high anion gap metabolic acidosis. *Crit. Care Med.* **28**, 1803–1807 (2000).

242. Wu, G. & Morris, S. M. Arginine metabolism: Nitric oxide and beyond. *Biochemical Journal* vol. 336 1–17 (1998).

243. Cederbaum, S. D., Shaw, K. N. F. & Valente, M. Hyperargininemia. *J. Pediatr.* **90**, 569–573 (1977).

244. Scaglia, F. *et al.* Clinical Consequences of Urea Cycle Enzyme Deficiencies and Potential Links to Arginine and Nitric Oxide Metabolism. *J. Nutr.* **134**, 2775S-2782S (2004).

245. Xue, A. *et al.* Genome-wide association analyses identify 143 risk variants and putative regulatory mechanisms for type 2 diabetes. *Nat. Commun.* **9**, 2941 (2018).

246. Piatti, P. *et al.* Long-Term Oral L-Arginine Administration Improves Peripheral and Hepatic Insulin Sensitivity in Type 2 Diabetic Patients. *Diabetes Care* **24**, 875–880 (2001).

247. Gannon, M. C., Nuttall, J. A. & Nuttall, F. Q. Oral arginine does not stimulate an increase in insulin concentration but delays glucose disposal. *Am. J. Clin. Nutr.* **76**, 1016–1022 (2002).

248. Lucotti, P. *et al.* Beneficial effects of a long-term oral L-arginine treatment added to a hypocaloric diet and exercise training program in obese, insulin-resistant type 2 diabetic patients. *Am. J. Physiol. - Endocrinol. Metab.* **291**, (2006).

249. Monti, L. D. *et al.* Decreased diabetes risk over 9 year after 18-month oral l-arginine treatment in middle-aged subjects with impaired glucose tolerance and metabolic syndrome (extension evaluation of l-arginine study). *Eur. J. Nutr.* **57**, 2805–2817 (2018).

250. Bastarache, L. *et al.* Improving the phenotype risk score as a scalable approach to identifying patients with Mendelian disease. *J. Am. Med. Informatics Assoc.* **26**, 1437–1447 (2019).

251. Agana, M., Frueh, J., Kamboj, M., Patel, D. R. & Kanungo, S. Common metabolic disorder (inborn errors of metabolism) concerns in primary care practice. *Ann. Transl. Med.* **6**, 469–469 (2018).

252. Ezgu, F. Inborn Errors of Metabolism. in *Advances in Clinical Chemistry* vol. 73 195–250 (Academic Press Inc., 2016).

253. Borghi, C. *et al.* The association between blood pressure and lipid levels in Europe: European study on cardiovascular risk prevention and management in usual daily practice. *J. Hypertens.* **34**, 2155–2163 (2016).

254. Beinker, N. K. *et al.* Threshold effect of liver iron content on hepatic inflammation and fibrosis in hepatitis B and C. *J. Hepatol.* **25**, 633–638 (1996).

255. Hemani, G. *et al.* The MR-base platform supports systematic causal inference across the human phenome. *Elife* **7**, (2018).

256. Burgess, S. *et al.* Guidelines for performing Mendelian randomization investigations. *Wellcome Open Res.* **4**, (2020).

257. Smith, G. D. & Ebrahim, S. 'Mendelian randomization': Can genetic epidemiology contribute to understanding environmental determinants of disease? *International Journal of Epidemiology* vol. 32 1–22 (2003).

258. Roden, D. *et al.* Development of a Large-Scale De-Identified DNA Biobank to Enable Personalized Medicine. *Clin. Pharmacol. Ther.* **84**, 362–369 (2008).

259. Köhler, S. *et al.* The Human Phenotype Ontology in 2017. *Nucleic Acids Res.* **45**, D865–D876 (2017).

260. Wei, W.-Q. *et al.* Evaluating phecodes, clinical classification software, and ICD-9-CM codes for phenome-wide association studies in the electronic health record. *PLoS One* **12**, e0175508 (2017).

261. *UK Biobank Showcase User Guide*. http://www.ukbiobank.ac.uk (2017).

262. Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G. R. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat. Genet.* **44**, 955–959 (2012).

263. *UK Biobank Integrating electronic health records into the UK Biobank Resource*. http://www.ukbiobank.ac.uk/ (2014).

264. Wu, P. *et al.* Developing and Evaluating Mappings of ICD-10 and ICD-10-CM codes to Phecodes. *bioRxiv* 462077 (2018) doi:10.1101/462077.

265. Sampietro, M. & Iolascon, A. Molecular pathology of Crigler-Najjar type I and II and Gilbert's syndromes. *Haematologica* **84**, 150–157 (1999).

266. Kulminski, A. M. *et al.* Genetic and regulatory architecture of Alzheimer's disease in the APOE region. *Alzheimer's Dement.* **12**, (2020).

267. Cai, Y., An, S. S. A. & Kim, S. Mutations in presenilin 2 and its implications in Alzheimer's disease and other dementia-associated disorders. *Clin. Interv. Aging* **10**, 1163–1172 (2015).

268. Levy-Lahad, E. *et al.* Genomic structure and expression of STM2, the chromosome 1 familial Alzheimer disease gene. *Genomics* **34**, 198–204 (1996).

269. Dixon, P., Davey Smith, G. & Hollingworth, W. The Association Between Adiposity and Inpatient Hospital Costs in the UK Biobank Cohort. *Appl. Health Econ. Health Policy* **17**, 359–370 (2019).

270. Fuchsberger, C. *et al.* The genetic architecture of type 2 diabetes. *Nature* **536**, 41–47 (2016).

271. Collins, F. S. & Varmus, H. A New Initiative on Precision Medicine. *N. Engl. J. Med.* **372**, 793–795 (2015).

272. Sankar, P. L. & Parker, L. S. The Precision Medicine Initiative's All of Us Research Program: An agenda for research on its ethical, legal, and social issues. *Genetics in Medicine* vol. 19 743–750 (2017).

273. Milton, J. N. *et al.* A genome-wide association study of total bilirubin and cholelithiasis risk in sickle cell anemia. *PLoS One* **7**, 34741 (2012).

274. Sanna, S. *et al.* Common variants in the SLCO1B3 locus are associated with bilirubin levels and unconjugated hyperbilirubinemia. *Hum. Mol. Genet.* **18**, 2711–2718 (2009).

275. Spracklen, C. N. *et al.* Identification of type 2 diabetes loci in 433,540 East Asian individuals. *Nature* **582**, 240–245 (2020).

276. Keene, K. L. *et al.* Genome-Wide Association Study Meta-Analysis of Stroke in 22 000 Individuals of African Descent Identifies Novel Associations with Stroke. *Stroke* **51**, 2454–2463 (2020).

277. Mahajan, A. *et al.* Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. *Nat. Genet.* **46**, 234–244 (2014).

278. Randall, J. C. *et al.* Sex-stratified Genome-wide Association Studies Including 270,000

Individuals Show Sexual Dimorphism in Genetic Loci for Anthropometric Traits. *PLoS Genet.* **9**, (2013).

279. Graham, S. E. *et al.* Sex-specific and pleiotropic effects underlying kidney function identified from GWAS meta-analysis. *Nat. Commun.* **10**, (2019).

280. Lagou, V. *et al.* Sex-dimorphic genetic effects and novel loci for fasting glucose and insulin variability. *Nat. Commun.* **12**, 1–18 (2021).

281. Wu, M. C. *et al.* Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.* **89**, 82–93 (2011).

282. Guo, M. H., Plummer, L., Chan, Y. M., Hirschhorn, J. N. & Lippincott, M. F. Burden Testing of Rare Variants Identified through Exome Sequencing via Publicly Available Control Data. *Am. J. Hum. Genet.* **103**, 522–534 (2018).

283. Cirulli, E. T. *et al.* Genome-wide rare variant analysis for thousands of phenotypes in over 70,000 exomes from two cohorts. *Nat. Commun.* **11**, (2020).

284. Lord, J. *et al.* Deciphering the causal relationship between blood metabolites and Alzheimer's Disease: a Mendelian Randomization study. (2020) doi:10.1101/2020.04.28.20083253.

285. Liu, J. *et al.* A Mendelian Randomization Study of Metabolite Profiles, Fasting Glucose, and Type 2 Diabetes. *Diabetes* **66**, (2017).

286. Liu, L. *et al.* Assessing the Associations of Blood Metabolites With Osteoporosis: A Mendelian Randomization Study. *J. Clin. Endocrinol. Metab.* **103**, 1850–1855 (2018).

287. Carvalho, C. M. *et al.* Investigating Causality Between Blood Metabolites and Emotional and Behavioral Responses to Traumatic Stress: a Mendelian Randomization Study. *Mol. Neurobiol.* **57**, 1542–1552 (2020).

288. Balbi, M. E. *et al.* Antioxidant effects of vitamins in type 2 diabetes: A meta-analysis of randomized controlled trials. *Diabetol. Metab. Syndr.* **10**, 18 (2018).

289. Afkhami-Ardekani, M. & Shojaoddiny-Ardekani, A. Effect of vitamin C on blood glucose, serum lipids & serum insulin in type 2 diabetes patients. *Indian J. Med. Res.* **126**, 471–474 (2007).

290. Moser, M. A. & Chun, O. K. Vitamin C and heart health: A review based on findings from epidemiologic studies. *International Journal of Molecular Sciences* vol. 17 (2016).

291. Tecklenburg, S. L., Mickleborough, T. D., Fly, A. D., Bai, Y. & Stager, J. M. Ascorbic acid supplementation attenuates exercise-induced bronchoconstriction in patients with asthma. *Respir. Med.* **101**, 1770–1778 (2007).

292. Liu, J. *et al.* Metabolomics based markers predict type 2 diabetes in a 14-year follow-up study. *Metabolomics* **13**, 104 (2017).

293. Tucker-Drob, E. M., Briley, D. A. & Harden, K. P. Genetic and Environmental Influences on Cognition Across Development and Context. *Curr. Dir. Psychol. Sci.* **22**, 349–355 (2013).

294. Taylor, L., Watkins, S. L., Marshall, H., Dascombe, B. J. & Foster, J. The impact of different environmental conditions on cognitive function: A focused review. *Frontiers in Physiology* vol. 6 372 (2016).

295. Newborn blood spot test - NHS. https://www.nhs.uk/conditions/baby/newborn-screening/blood-spot-test/ (2018).

296. Wilson, J. & Jungner, G. Principles and practice of screening for disease. *J. R. Coll. Gen. Pract.* **16**, 318 (1968).

297. Fox, K. M. *et al.* Clinical and economic burden associated with cardiovascular events among patients with hyperlipidemia: A retrospective cohort study. *BMC Cardiovasc. Disord.* **16**, 13 (2016).

298. Patel, P. *et al.* Hidden burden of electronic health record-identified familial hypercholesterolemia: Clinical outcomes and cost of medical care. *J. Am. Heart Assoc.* **8**, (2019).

299. Gidding, S. S. *et al.* The Agenda for Familial Hypercholesterolemia: A Scientific Statement from the American Heart Association. *Circulation* **132**, 2167–2192 (2015).

300. Varret, M., Abifadel, M., Rabès, J. P. & Boileau, C. Genetic heterogeneity of autosomal dominant hypercholesterolemia. *Clinical Genetics* vol. 73 1–13 (2008).

301. Häberle, J. *et al.* Suggested guidelines for the diagnosis and management of urea cycle disorders. *Orphanet Journal of Rare Diseases* vol. 7 (2012).

302. Jansen, I. E. *et al.* Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. *Nat. Genet.* **51**, 404–413 (2019).

303. Sklar, P. Psychiatric Genomics Consortium: Past And Present. *Eur. Neuropsychopharmacol.* **27**, S359 (2017).

304. Beecham, G. W. *et al.* The Alzheimer's Disease Sequencing Project: Study design and sample selection. *Neurol. Genet.* **5**, (2017).

305. Sims, R. *et al.* Rare coding variants in PLCG2, ABI3, and TREM2 implicate microglial-mediated innate immunity in Alzheimer's disease. *Nat. Genet.* **49**, 1373–1384 (2017).

306. Michailidou, K. *et al.* Large-scale genotyping identifies 41 new loci associated with breast cancer risk. *Nat. Genet.* **45**, 353–361 (2013).

307. Michailidou, K. *et al.* Genome-wide association analysis of more than 120,000 individuals identifies 15 new susceptibility loci for breast cancer. *Nat. Genet.* **47**, 373–380 (2015).

308. Haiman, C. A. *et al.* A common variant at the TERT-CLPTM1L locus is associated with estrogen receptor-negative breast cancer. *Nat. Genet.* **43**, 1210–1214 (2011).

309. Hunter, D. & Chanock, S. J. A candidate gene approach to searching for low-penetrance breast and prostate cancer genes. *Nature Reviews Cancer* vol. 5 977–985 (2005).

310. Pharoah, P. *et al.* Commonly studied single-nucleotide polymorphisms and breast cancer: Results from the Breast Cancer Association Consortium. *J. Natl. Cancer Inst.* **98**, 1382–1396 (2006).

311. Ahsan, H. *et al.* A Genome-wide Association Study of Early-Onset Breast Cancer Identifies PFKM as a Novel Breast Cancer Gene and Supports a Common Genetic Spectrum for Breast Cancer at Any Age. *Cancer Epidemiol. Biomarkers Prev.* **23**, 658–669 (2014).

312. Pattaro, C. *et al.* Genetic associations at 53 loci highlight cell types and biological pathways relevant for kidney function. *Nat. Commun.* **7**, (2016).

313. Trampush, J. W. *et al.* GWAS meta-analysis reveals novel loci and genetic correlates for general cognitive function: A report from the COGENT consortium. *Mol. Psychiatry* **22**, 336–345 (2017).

314. Prins, B. P. *et al.* Genome-wide analysis of health-related biomarkers in the UK Household Longitudinal Study reveals novel associations. *Sci. Rep.* **7**, (2017).

315. Wain, L. V *et al.* Novel insights into the genetics of smoking behaviour, lung function, and chronic obstructive pulmonary disease (UK BiLEVE): a genetic association study in UK Biobank. *Lancet Respir. Med.* **3**, 769–781 (2015).

316. Scelsi, M. A. *et al.* Genetic study of multimodal imaging Alzheimer's disease progression score implicates novel loci. *Brain* **141**, 2167–2180 (2018).

317. Petersen, R. C. *et al.* Alzheimer's Disease Neuroimaging Initiative (ADNI): Clinical characterization. *Neurology* **74**, 201–209 (2010).

318. Hofman, A. *et al.* The Rotterdam Study: 2016 objectives and design update. *Eur. J. Epidemiol.* **30**, 661–708 (2015).

319. Kooijman, M. N. *et al.* The Generation R Study: design and cohort update 2017. *Eur. J. Epidemiol.* **31**, 1243–1264 (2016).

320. Lichtenstein, P. *et al.* The Swedish Twin Registry: A unique resource for clinical, epidemiological and genetic studies. *Journal of Internal Medicine* vol. 252 184–205 (2002).

321. Dick, D. M. *et al.* Spit for Science: Launching a longitudinal study of genetic and environmental influences on substance use and emotional health at a large US university. *Front. Genet.* **5**, (2014).

322. Zabaneh, D. *et al.* A genome-wide association study for extremely high intelligence. *Mol. Psychiatry* **23**, 1226–1232 (2018).

323. Trouton, A., Spinath, F. M. & Plomin, R. Twins Early Development Study (TEDS): A multivariate, longitudinal genetic investigation of language, cognition and behavior problems in childhood. *Twin Res.* **5**, 444–448 (2002).

324. Skytthe, A. *et al.* The Danish Twin Registry: Linking surveys, national registers, and biological information. *Twin Res. Hum. Genet.* **16**, 104–111 (2013).

325. Schumann, G. *et al.* The IMAGEN study: Reinforcement-related behaviour in normal brain function and psychopathology. *Molecular Psychiatry* vol. 15 1128–1139 (2010).

326. Gillespie, N. A. *et al.* The brisbane longitudinal twin study: Pathways to cannabis use, abuse, and dependence project - Current status, preliminary results, and future directions. *Twin Res. Hum. Genet.* **16**, 21–33 (2013).

327. Polderman, T. J. C. *et al.* Attentional switching forms a genetic link between attention problems and autistic traits in adults. *Psychol. Med.* **43**, 1985–1996 (2013).

328. Finkel, D. & Pedersen, N. L. Processing speed and longitudinal trajectories of change for cognitive abilities: The Swedish Adoption/Twin Study of Aging. *Aging, Neuropsychology, and Cognition* vol. 11 325–345 (2004).

329. Malik, R. *et al.* Multiancestry genome-wide association study of 520,000 subjects identifies 32 loci associated with stroke and stroke subtypes. *Nat. Genet.* **50**, 524–537 (2018).

330. Beecham, A. H. *et al.* Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis. *Nat. Genet.* **45**, 1353–1362 (2013).

331. Nikpay, M. *et al.* A comprehensive 1000 Genomes-based genome-wide association meta-analysis of coronary artery disease. *Nat. Genet.* **47**, 1121–1130 (2015).

332. Bonnemaijer, P. W. M. *et al.* Multi-trait genome-wide association study identifies new loci associated with optic disc parameters. *Commun. Biol.* **2**, (2019).

333. Schott, J. M. *et al.* Genetic risk factors for the posterior cortical atrophy variant of Alzheimer's disease. *Alzheimer's Dement.* **12**, 862–871 (2016).

334. Evangelou, E. *et al.* Genetic analysis of over 1 million people identifies 535 new loci associated with blood pressure traits. *Nat. Genet.* **50**, 1412–1425 (2018).

335. Ehret, G. B. *et al.* The genetics of blood pressure regulation and its target organs from association studies in 342,415 individuals. *Nat. Genet.* **48**, 1171–1184 (2016).

336. Zhu, Z. *et al.* Genetic overlap of chronic obstructive pulmonary disease and cardiovascular disease-related traits: A large-scale genome-wide cross-trait analysis. *Respir. Res.* **20**, 64–64 (2019).

337. Prins, B. P. *et al.* Exome-chip meta-analysis identifies novel loci associated with cardiac conduction, including ADAMTS6. *Genome Biol.* **19**, (2018).

338. Grove, M. L. *et al.* Best Practices and Joint Calling of the HumanExome BeadChip: The CHARGE Consortium. *PLoS One* **8**, 24 (2013).