



THE UNIVERSITY *of* EDINBURGH

This thesis has been submitted in fulfilment of the requirements for a postgraduate degree (e. g. PhD, MPhil, DClinPsychol) at the University of Edinburgh. Please note the following terms and conditions of use:

- This work is protected by copyright and other intellectual property rights, which are retained by the thesis author, unless otherwise stated.
- A copy can be downloaded for personal non-commercial research or study, without prior permission or charge.
- This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the author.
- The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author.
- When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given.

A Multi-Omics Investigation of Major Depressive Disorder and Antidepressant Exposure

Ella Davyson



THE UNIVERSITY *of* EDINBURGH

Doctor of Philosophy

The University of Edinburgh

2025

Abstract

Major Depressive Disorder (MDD) is a complex psychiatric condition affecting more than 280 million people around the world and is estimated to be a leading cause of disability by 2030. The exact biological mechanisms of MDD are unknown and current pharmacological treatments (antidepressants) for MDD are limited. Most antidepressants take around 4-8 weeks for a clinical improvement in symptoms, and an estimated 40% of people are resistant to treatment. Furthermore, antidepressants are associated with a range of side effects, including gastrointestinal disturbances, nausea and sexual dysfunction. In sum, individuals suffering from MDD will undergo various antidepressant regimens over a period of months with a host of side effects for often sub-optimal changes to their symptoms. Therefore, there is a clinical need to identify more efficacious antidepressants with a lessened side effect profile. Furthermore, an increased understanding of how and when antidepressants exert their effects may also enable stratification of current treatments to those more likely to respond. Both goals require an increased understanding of the biological basis of MDD and the biological impact of currently prescribed antidepressants. This thesis aims to investigate both aspects using various multi-omics data in large population cohorts, primarily the UK Biobank and Generation Scotland.

Chapter 1 consists of a general introduction to MDD and antidepressants, followed by **Chapter 2** which describes the datasets used in this thesis. The next three chapters constitute the main analyses conducted and are each accompanied by in-depth methodologies. **Chapter 3** examines the associational and causal relationship between circulating metabolite measures and lifetime MDD in the UK Biobank. Our findings indicate

a potential causal role of omega-3 and omega-6 fatty acids measures in MDD, with a colocalised region in the *FADS* cluster on chromosome 11. Due to high linkage disequilibrium (LD) in this region, it is challenging to identify the potential causal gene underlying these observations. Therefore, **Chapter 4** aims to investigate this further using rare variants, which typically exhibit less LD and may provide further evidence of causality. This found evidence of disparate signals in the *FADS* region on omega-3 fatty acids and MDD. Specifically, our analysis indicated a nominally significant burden of rare pathogenic variants in *FADS1* in MDD, which was primarily driven by a singular missense variant in Exon 1. Next, **Chapter 5** examines the relationship between DNA methylation and both self-reported and prescription-derived antidepressant exposure in Generation Scotland. Our findings indicated robust associations between DNA methylation and antidepressant exposure, specifically in *DGUOK-AS1* and *KANK1*. Finally, **Chapter 6** discusses the findings of this work, its limitations and potential future directions for further elucidating on the biological basis of MDD and the effects of antidepressant exposure.

Overall, this thesis presents a series of analyses utilising multi-omics data to learn more about the biological differences in those with MDD and the potential impact of antidepressant medications. These findings contribute to a growing body of work to increase understanding of MDD and antidepressants, with the goal of developing new effective treatments.

Lay Summary

Depression is a serious condition that affects many people, yet current treatments don't work well for around 40% of individuals. Some people also experience unpleasant side effects like nausea or sexual dysfunction. We don't fully understand why some people don't respond to antidepressants or why some suffer more side effects than others. There is an urgent clinical need for more effective treatments for depression and better targeting of current treatments to those most likely to respond. To improve treatments, we ultimately need to learn more about what is happening in the body during depression and when taking antidepressants.

Genes play a big role in depression, with approximately 40% of a person's risk from their genetic code (DNA). Understanding more about which genes may predispose people to depression can help us to understand the underlying biological changes which may lead to a person becoming depressed. In turn, this could enable the identification of new targets for more effective treatments. Hundreds of genes have now been identified to be associated with depression. The next step is to understand how these genes affect the body.

One way to do this is through multi-omics, which looks at different biological molecules (like proteins or fats) to understand how certain genes influence bodily function. For example, if those carrying a certain gene have much higher cholesterol, it indicates that the gene may have a role in increasing cholesterol levels. This thesis used large-scale biological data in UK Biobank (~ 500,000 participants) and Generation Scotland (~ 20,000 participants) to explore what happens in the body during depression and when taking antidepressants.

The first part of this research found that people with depression tend to have lower omega-3 fatty acids (healthy fats found in fish) and a higher omega-6: omega-3 ratio, which could contribute to depression. A gene called *FADS1*, which helps to process omega-3 and omega-6 fats from the diet, may play a role in this. The second part of this research looked at how antidepressants affect the body's genes through a process called DNA methylation (where external factors like medication can change gene activity). This found strong evidence that taking antidepressants is linked to changes in certain genes, linked to important functions in the body. These include brain cell growth, immune response, energy production in cells and male fertility. Overall, this thesis shows how biological data can help us to understand depression better and could lead to improved treatments in the future.

Declaration of Originality

I declare that this thesis is my own composition and that it has not been submitted for any other degree of professional qualification at this university or any other institution. Parts of the work in this thesis have been previously published. The included publications are my own work.

The work presented in Chapter 3 has been published in *Biological Psychiatry* (2023):

Citation: Davyson, E., Shen, X., Gadd, D. A., Bernabeu, E., Hillary, R. F., McCartney, D. L., ... & McIntosh, A. M. (2023). Metabolomic investigation of major depressive disorder identifies a potentially causal association with polyunsaturated fatty acids. *Biological psychiatry*, 94(8), 630-639.

Author contributions are as follows: E.D., A.M.M., X.S., M.A., and R.E.M., conceived and designed the research. A.M.M, X.S., M.A., R.E.M., supervised E.D. in running the analysis. E.D. ran all analyses, including the MetWAS, GWAS, MR and colocalisation. D.L.M., D.A.G., R.F.H., and E.B., gave support on analysis debugging and manuscript drafting. E.D. drafted the article and is responsible for data and code open access (on GWAS catalogue and GitHub).

The work presented in **Chapter 5** has been published in *Nature Communications* (2025).

Citation: Davyson, E., Shen, X., Huider, F., Adams, M. J., Borges, K., McCartney, D. L., ... & McIntosh, A. M. (2025). Insights from a methylome-wide association study of antidepressant exposure. *Nature Communications*, 16(1), 1908.

Key author contributions are as follows: E.D., A.M.M., R.E.M., and N.R.W. were responsible for the conception and the design of the study. E.D. conducted the primary

data analyses, the MPS-meta-analysis and was primary author of the manuscript. X.S., M.J.A., D.L.M., and L.F.B., advised on methodology and data analysis. K.B. contributed to the derivation of prescription-derived antidepressant phenotypes. F.H., J.V.D., and D.B., ran the out-of-sample analysis in the Netherlands Twin Register. A.W., H.J.G., L.K., A.Teumer., and H.V., ran the out-of-sample analysis in Study of Health in Pomerania. T.Z., J.K., and M.O., ran the out-of-sample analysis in the Finnish Twin Cohort. F.S.D, S.M., F.S, A.J.F., U.D., and T.K., ran the out-of-sample analysis in FOR2107. A.Tapuc., D.C., E.B.B., and T.B ran the out-of-sample analysis in the Munich Antidepressant Response Study / UniPolar Depression Study. A.S.F.K., and P.Y., ran the out-of-sample analysis in the Avon Longitudinal Study of Parents and Children. C.C.Y.W., ran the out-of-sample analysis in the E-Risk Longitudinal Twin Study. L.A., H.L.F., and J.M., secured funding, oversaw data collection and the derivation of variables for the E-Risk Longitudinal Twin Study. S.R.C., P.R., and T.C.R., provided guidance and expertise on running the out-of-sample analysis in the Lothian Birth Cohorts. E.J.C.G.v.d.O, K.A.A., and B.W.J.H.P., ran the enrichment analysis in NESDA. A.M.M, R.E.M, N.R.W, X.S., and M.J.A. co-supervised the project. All authors read and approved the final manuscript.

Signed: Ella Davyson.

Date: 27/03/2025

Acknowledgements

I would like to thank my primary supervisors, Prof Andrew McIntosh and Prof Riccardo Marioni for their support, expertise and encouragement throughout both my master's degree and my PhD. I am also hugely indebted to Dr Mark Adams, my co-supervisor, for support on analytical techniques, data resources alongside novel research ideas. Also, I would like to thank Dr Xueyi Shen for the countless video calls, discussions and advice throughout this PhD. It was a privilege to learn from and with all my supervisory team. I would also like to thank the Centre of Doctoral Training and the Medical Research Council for entrusting me with this opportunity and providing an enjoyable learning experience. Especially, I would like to thank Dr Ian Simpson and Dr Diego Oyarzún for their encouragement in the early days when I was adjusting to the informatics field. A big thank you to Prof Naomi Wray for hosting me at the Institute of Molecular Biosciences (IMB) in Autumn 2022, an experience I will never forget. Also, I would like to thank everyone at the IMB for their welcome and friendship, with particular thanks to Alicia Walker and Lauren Barker.

Thank you to all the volunteers and participants in UK Biobank and Generation Scotland, who make all this research possible.

Thank you to several members of the Marioni Lab who provided counsel during this PhD, namely Dr Robert Hillary, Dr Danni Gadd and Dr Daniel McCartney for their guidance, support and friendship. I would also like to thank several members of the McIntosh group (past & present) including Dr Miruna Barbu, Dr Melisa Chuong, Dr Melissa Lewins, and (future) Dr Poppy Grimes for their friendship, support and guidance. I would also like to

thank those at Genomics England, including Dr Miruna Barbu for advising me about rare-variant analyses.

I would like to thank the Code First: Girls program in Bristol, for running free coding workshops which introduced me several coding concepts in a friendly, fun way, and kickstarted my trajectory into the field of data science. I would also like to thank Nicola Crombie from Theorise who provided regular support sessions throughout my PhD and helped me grow and establish healthy working habits.

Thank you to my parents, Carol & Everard, for always supporting me. Special thanks to my dad for listening to several practice presentations over the years and giving me endless tips from his radio days. A huge thank you to my sister Alice and her fiancée Oliver for being friendly faces in Edinburgh and for always hosting me at the Granton palace after I'd moved to Sheffield. I am also hugely indebted to my sibling Katie, a talented and patient teacher who regularly tutored me in Maths during my GCSEs and A-levels, and later took time to provide further support in linear algebra, statistical genetics and machine learning techniques throughout my Master's and PhD. I would also like to thank Peter and Susan for their support and the countless Sunday dinners. Thank you to all my lovely friends I am lucky to have, including Holly, Romy, Livvy, Annie, Molly, Ellie, Jazz, JoJo, Puck, Zoe, Tom and Sarah. I would also like to thank my Nanny June, for repeatedly saving my knitting failures, encouraging all things science and having the best stories.

Finally, thank you to my partner and best friend, Toby, for keeping me grounded, cheering me on and making me laugh every day.

This PhD has been a hugely rewarding experience and I am very grateful.

Contents

1	INTRODUCTION	27
1.1	MAJOR DEPRESSIVE DISORDER (MDD)	27
1.1.1	EPIDEMIOLOGY OF MDD	27
1.1.2	BIOLOGICAL THEORIES OF MDD	30
1.1.3	DIAGNOSIS OF MDD	35
1.2	ANTIDEPRESSANTS	37
1.2.1	FIRST GENERATION ANTIDEPRESSANTS	38
1.2.2	SECOND GENERATION ANTIDEPRESSANTS	39
1.2.3	ANTIDEPRESSANT TREATMENT RESISTANCE	44
1.2.4	RAPID ACTING ANTIDEPRESSANTS	46
1.3	GENOMIC STUDIES OF MDD	47
1.3.1	CANDIDATE GENE STUDIES	48
1.3.2	GENOME-WIDE ASSOCIATION STUDIES	49
1.3.3	CHALLENGES INTERPRETING GWAS ASSOCIATIONS	57
1.3.4	MISSING HERITABILITY	58
1.3.5	RARE GENETIC VARIATION	60
1.4	MULTI-OMICS	64
1.4.1	GENOTYPE-FIRST APPROACH	67
1.4.2	PHENOTYPE-FIRST APPROACH	68
1.4.3	METABOLOMICS	69
1.4.4	EPIGENOMICS AND ANTIDEPRESSANT EXPOSURE	72
1.5	AIMS AND OUTLINE OF THIS THESIS	76
2	DATASETS	78
2.1	UK BIOBANK (UKB)	78
2.1.1	ETHICS AND FUNDING	79
2.1.2	GENOTYPING DATA	79
2.1.3	WHOLE-GENOME SEQUENCING DATA	81
2.1.4	METABOLOMIC DATA	82
2.1.5	MDD PHENOTYPES	83
2.1.6	SOCIODEMOGRAPHIC FACTORS	87
2.2	GENERATION SCOTLAND	88
2.2.1	ETHICS AND FUNDING	88
2.2.2	DNA METHYLATION	89
2.2.3	PRESCRIPTION DATA	90
2.2.4	MDD PHENOTYPE	90
2.2.5	ANTIDEPRESSANT EXPOSURE PHENOTYPES	91
2.3	GENOTYPE TISSUE EXPRESSION (GTEx)	95
2.4	GENOME AGGREGATION DATABASE	97
3	CIRCULATING METABOLITES AND MAJOR DEPRESSIVE DISORDER	99

3.1	METHODS	100
3.1.1	METABOLOME-WIDE ASSOCIATION STUDY	100
3.1.2	GWAS	100
3.1.3	LD SCORE REGRESSION	102
3.1.4	MENDELIAN RANDOMISATION	104
3.1.5	COLOCALISATION	109
3.2	METABOLOMIC INVESTIGATION OF MAJOR DEPRESSIVE DISORDER IDENTIFIES A POTENTIALLY CAUSAL ASSOCIATION WITH POLYUNSATURATED FATTY ACIDS	112
3.3	CONCLUSION	127
4	<u>RARE PATHOGENIC <i>FADS</i> VARIANTS, MAJOR DEPRESSIVE DISORDER AND OMEGA-3 FATTY ACIDS</u>	129
4.1	IN-DEPTH METHODS	131
4.1.1	VARIANT ANNOTATION	131
4.1.2	AGGREGATE-VARIANT TESTING	140
4.2	INTRODUCTION	149
4.3	METHODS	156
4.3.1	POPULATION COHORT	156
4.3.2	WHOLE GENOME SEQUENCING DATA	156
4.3.3	ASSESSMENT OF MAJOR DEPRESSIVE DISORDER	157
4.3.4	VARIANT FILTERING AND EXTRACTION	158
4.3.5	VARIANT ANNOTATION	159
4.3.6	VARIANT PRIORITISATION	159
4.3.7	METABOLIC BIOMARKERS	161
4.3.8	SUB-COHORTS	161
4.3.9	STATISTICAL TESTS	162
4.3.10	AGGREGATE VARIANT TESTING	162
4.4	RESULTS	164
4.4.1	IDENTIFICATION OF RARE PATHOGENIC VARIANTS IN <i>FADS</i>	164
4.4.2	TESTING THE ASSOCIATION OF RARE-PRIORITISED VARIANTS WITH MDD AND METABOLITES	172
4.5	DISCUSSION	185
4.5.1	MINIMAL ASSOCIATION OF RARE VARIATION IN <i>FADS1/2/3</i> GENES WITH CIRCULATING LC-PUFAS	186
4.5.2	DISTINCT SIGNALS IN <i>FADS</i> FOR MDD AND METABOLITE TRAITS	188
4.5.3	DRIVER VARIANT FOR <i>FADS1</i> ASSOCIATION WITH MDD	189
4.5.4	STRENGTHS	191
4.5.5	LIMITATIONS	192
4.5.6	SUMMARY	193
4.6	CONCLUSION	194
5	<u>DNA METHYLATION AND ANTIDEPRESSANT EXPOSURE</u>	196
5.1	INTRODUCTION	196
5.2	METHODS	198
5.2.1	METHYLOME-WIDE ASSOCIATION STUDY	198
5.2.2	DIFFERENTIALLY METHYLATED REGIONS	200

5.2.3	ENRICHMENT (HYPERGEOMETRIC TESTS)	202
5.2.4	METHYLATION PROFILE SCORE (MPS) ANALYSIS	203
5.3	INSIGHTS FROM A METHYLOME-WIDE ASSOCIATION STUDY OF ANTIDEPRESSANT EXPOSURE	208
5.4	CONCLUSION	229
6	DISCUSSION	231
6.1	THE ROLE OF METABOLITES IN MDD	232
6.1.1	SUMMARY OF FINDINGS	232
6.1.2	OMEGA-3 AND OMEGA-6 FATTY ACIDS IN MDD	233
6.1.3	FATTY ACID DESATURASE (FADS) CLUSTER	237
6.1.4	POTENTIAL MECHANISMS	241
6.2	THE ASSOCIATION OF ANTIDEPRESSANTS WITH DNAM	244
6.2.1	SUMMARY OF FINDINGS	244
6.2.2	BIOLOGICAL IMPLICATIONS OF ASSOCIATIONS	244
6.2.3	THE POTENTIAL ROLE OF MITOCHONDRIAL DYSFUNCTION IN MDD	246
6.2.4	INSIGHTS INTO POTENTIAL SIDE EFFECTS	249
6.2.5	TREATMENT RESPONSE	250
6.3	LIMITATIONS & FUTURE DIRECTIONS	252
6.3.1	INTEGRATIVE ANALYSIS OF MULTI-OMICS	252
6.3.2	LONGITUDINAL/PROSPECTIVE CHANGES	255
6.3.3	MORE REPRESENTATIVE COHORTS	258
6.3.4	TISSUE SPECIFIC ANALYSES	261
6.3.5	TRANS-DIAGNOSTIC APPROACHES	265
6.3.6	TECHNOLOGICAL ADVANCES	267
7	RESPONSIBLE RESEARCH AND INNOVATION	273
8	CONCLUSION	276
9	REFERENCES	279
10	APPENDIX	333
10.1	SUPPLEMENTARY FIGURES	334
10.1.1	METHODS	337
10.1.2	RESULTS	338
10.2	SUPPLEMENTARY TABLES	383
10.2.1	METHODS	384
10.2.2	RESULTS	385

Abbreviations

5-HT	Serotonin
5-HTTLPR	Serotonin-transporter-linked promoter region
AA	Arachidonic acid
ACAT	Aggregated Cauchy Combination Test
ACAT-O	ACAT-Omnibus
ACAT-V	ACAT-Variant
AD	Anxiety Disorder
AF	Allele Frequency
ALA	Alpha-linoleic acid
ALSPAC	Avon Longitudinal Study of Parents and Children
AoU	All Of Us
ATC	Anatomical Therapeutic Classification
ATP	Adenosine triphosphate
AVT	Aggregate-variant-testing
BBB	Blood-brain barrier
BDNF	Brain-derived neurotropic factor
BH	Benjamini-Hochberg
BLUP	Best linear unbiased prediction
BMI	Body mass index
BNF	British National Formulary
BPD	Bipolar Disorder
CADD	Combined annotation dependent deletion
caQTLs	Chromatin accessibility quantitative trait loci
CDCV	Common-disease, Common Variant

CDS	Candidate Gene Studies
CHARGE	Cohorts for Heart and Aging Research in Genomic Epidemiology
CHI	Community Health Index
ChIP-Seq	Chromatin immunoprecipitation sequencing
CIDI	Clinical International Diagnostic Interview
CIDI-SF	CIDI-Short-Form
CONVERGE	China, Oxford, and Virginia Commonwealth University Experimental Research on Genetic Epidemiology
CpG	Cytosine-phosphate-Guanine
CRP	C-reactive Protein
CSF	Cerebrospinal fluid
CTC	Cell-type composition
CV	Coefficient of variation
DA	Dopamine
DEGs	Differentially expressed gene sets
DHA	Docoshaexanoic acid
DMR	Differentially methylated region
DNAm	DNA methylation
DNMT	DNA methyltransferase
DPA	Docosapentanoic acid
DSM	Diagnostic and Statistical Manual for Mental Disorders
E-Risk	Environmental Risk Longitudinal Twin Study
EEG	Electroencephalogram
ENCODE	Encyclopedia of DNA Elements

EPA	Eicosapentaenoic acid
eQTL	Expression Quantitative Trait Loci
ETC	Electron transport chain
ExAC	Exome Aggregation Consortium
FADS	Fatty Acid Desaturase
FDA	US Food and Drug Administration
FDR	False-discovery rate
FTC	Finn Twin Cohort
FUMA	Functional mapping and annotation
GBD	Global Burden of Disease
GC	Gas chromatography
GERP	Genomic and Evolutionary Profiling
GLA	Gamma-linoleic acid
GnomAD	Genome Aggregation Database
GP	General practice
GRM	Genomics relatedness matrix
GS	Generation Scotland
GTE _x	Genotype-Tissue-Expression Project
GWAS	Genome-wide association study
h^2_{SNP}	SNP-heritability
h^2_{SNV}	SNV-heritability
HapMap	Haplotype Map
HGMD	Human Gene Mutation Database
HGP	Human Genome Project

hisQTLs	Histone Quantitative Trait Loci
HMDB	Human Metabolome Database
HMM	Hidden Markov models
HPA	Hypothalamic Pituitary Axis
HWE	Hardy-Weinburg Equilibrium
IL-11	Interleukin-11
IL-6	Interleukin-6
InSIDE	Instrument Strength Independent of Direct Effect
IQR	Interquartile range
IV	Instrumental variable
IVW	Inverse-variance weighted
LA	Linoleic acid
LASSO	Least absolute shrinkage and selection operator
LBC1936	Lothian Birth Cohort 1936
LC	Liquid chromatography
LC-PUFAs	Long-chain polyunsaturated fatty acids
LD	Linkage disequilibrium
LDSC	LD-Score regression
lncRNA	Long non-coding RNA
LoF	Loss-of-function
LOFTEE	Loss-of-function transcript effect estimator
LOVO	Leave-one-variant-out analysis
LPCs	Lysophosphatidycholines

LRS	Long-read sequencing
LTP	Long-term potentiation
MAC	Minor allele count
MAF	Minor allele frequency
MAO	Monoamine oxidase
MAOIs	Monoamine oxidase inhibitors
MARS-UniDep	Munich Antidepressant Response Study-Unipolar Depression Study
MaxEntScan	Maximum Entropy Scan
MDD	Major Depressive Disorder
MDS	Multi-dimensional scaling
MetQTLs	Metabolite Quantitative Trait Loci
MetWAS	Metabolome-wide association study
MGB	Mass General Brigham Biobank
MIN_P	Minimum P-value
miRNA	microRNA
MOA	Mixed-linear-model Omics Association
molQTL	Molecular Quantitative Trait Loci
MPS	Methylation profile score
MR	Mendelian Randomisation
mRNA	Messenger RNA
MS	Mass spectrometry
mtDNA	Mitochondrial DNA
mTOR	Mammalian target of rapamycin

MVMR	Multi-variate mendelian randomisation
MVP	Millions Veterans Program
MWAS	Methylome-wide association study
NA	Noradrenaline
NESDA	Netherlands Study of Depression and Anxiety
NTR	Netherlands Twin Register
NGS	Next-generation sequencing
NGS	Next-generation sequencing
NHS	National Health Service
NIH	U.S. National Institute of Health
NMDA	N-methyl-D-aspartate
NMR	Nuclear Magnetic Resonance
OFH	Our Future Health
OLS	Ordinary-least squares
ONT	Oxford Nanopore Technologies
OPERA	Omics PIEiotRopic Association
ORM	Omics-relatedness matrix
OSCA	OmicS-data-based-Complex trait analysis
PacBio	Pacific Biosciences
PAT	Power-adjusted transferability
PCA	Principal Component Analysis
PCs	Principal Components
PGC	Psychiatric Genetics Consortium

PGC2	Second PGC GWAS of MDD
PGC3	Third PGC GWAS of MDD
PheWAS	Phenome-wide association studies
PHQ-9	Patient Health Questionnaire-9
PIS	Prescribing Information System
pLoF	Putative loss-of-function
PPIE	Patient and Public Involvement and Engagement
ProMENDA	Protein and Metabolite Network of Depression
PRS	Polygenic Risk Score
PSSD	Post-SSRI sexual dysfunction
PTVs	Protein truncating variants
pVCF	Aggregated variant call files
QC	Quality Control
RCT	Randomised clinical trials
RGC-ME	Regeneron Genetic Center Million Exomes
RNA-Seq	RNA sequencing
ROS	Reactive oxygen species
ROSMAP	Religious Orders Study/Memory and Aging Project
SBAT	Sparse burden association test
SCID	Structured Clinical Interview for DSM
SCZ	Schizophrenia
SEM	Structural equation modelling
SERT	Serotonin transporter protein
SHIP-Trend	Study of Health in Pomerania (Trend)

SKAT	Sequence kernel association test
SKAT-O	SKAT-Omnibus
SNP	Single nucleotide polymorphism
SNRI	Serotonin-noradrenaline reuptake inhibitor
SNV	Single nucleotide variant
sQTLs	Alternative splicing Quantitative Trait Loci
SSGAC	Social Science Genetic Association Consortium
SSRIs	Selective serotonin reuptake inhibitors
STAR*D	Sequenced Treatment Alternatives to Relieve Depression
STRADL	Stratifying Depression and Resilience Longitudinally
T2D	Type-2 Diabetes
TCAs	Tricyclic antidepressants
TDI	Townsend deprivation index
TF	Transcription factor
TFBS	Transcription factor binding site
TGP	1000 Genomes Project
TNF	Tumor necrosis factor
TOPMed	Trans-Omics for Precision Medicine
TRD	Treatment resistant depression
UKB	UK Biobank
UTR	Untranslated region
VEP	Variant effect predictor
WES	Whole-exome sequencing

WGBS	Whole-genome bi-sulphite sequencing
WGS	Whole-genome sequencing
WHO	World Health Organisation
WR	Wald Ratio
YLD	Years Lived with Disability

Figures

Figure 1.1 The increasing global burden of MDD	29
Figure 1.2 Hypotheses of MDD aetiopathogenesis	30
Figure 1.3 Rising rates of antidepressant prescriptions from 1998-2018	41
Figure 1.4 Mechanisms, side effects and example drug names of commonly prescribed antidepressants.	44
Figure 1.5 Discovery of genomic loci associated with MDD.	56
Figure 1.6 Complex interactions between environmental exposures (exposome) and multi-omic layers.....	65
Figure 2.1 Estimation of treatment periods from prescription records.....	95
Figure 2.2 The Genome Aggregation Database (GnomAD).	98
Figure 3.1 Principles and assumptions of Mendelian Randomisation.	107
Figure 3.2 Colocalisation framework and hypotheses.	111
Figure 4.1 Enrichment of error in pLoF variants in functional regions.	133
Figure 4.2 Different aggregate variant tests (AVT) employed.....	142
Figure 4.3 Beta distribution with shape (MAF, 1, 25) for assigning weights to rare variants in aggregated variant testing (AVT).	143
Figure 4.4 Schematic of the analysis presented in Chapter 4.	155
Figure 4.5 Prioritised variants in each gene in the FADS cluster.....	167
Figure 4.6 Summary of the prioritised variants per gene present in each cohort.....	168
Figure 4.7 Summary of the prioritised variants and participant carriers in each cohort.	170
Figure 4.8 Variant masks per gene for variant testing.....	172

Figure 4.9 Counts of MDD status in carriers/non-carriers of prioritised variants in the FADS cluster.	175
Figure 4.10 Aggregate-variant-testing of prioritised variants within each gene with MDD and omega-3 related measures.	178
Figure 4.11 Leave one variant out analysis for prioritised variants in <i>FADS1</i> (Mask 5 AF < 0.01) with MDD and all metabolomic traits.	183
Figure 5.1 Random effects meta-analysis.	205
Figure 6.1 Metabolic pathways of omega-3 and omega-6 fatty acids from dietary precursors.	235
Figure 6.2 Impact of dietary omega-3 and omega-6 levels and <i>FADS</i> haplotypes on lipid species and physiological processes.	240

Tables

Table 1.1 The diagnostic criteria for Major Depressive Disorder (MDD) from the Diagnostic and Statistical Manual for Mental Disorders (V).....	37
Table 2.1 Assessment of lifetime MDD in the UKB from the CIDI-SF.	85
Table 2.2 Antidepressant medications as defined by the British National Formulary (BNF).....	92
Table 4.1 The Variant Effect Predictor (VEP) consequence terms and associated “IMPACT” term.	131
Table 4.2 Putative loss of function variants filtered out by LOFTEE.	134
Table 4.3 Individual prediction tools which are included in the rare ensemble exome learner (REVEL).	136
Table 4.4 The criteria for each mask created per gene.	160
Table 4.5 The demographics of the MDD cohort and metabolite-cohort.	173
Table 4.6 The demographics of the carriers and non-carriers of variant chr11_61816814_G_C in the MDD cohort.	181

Introduction to Thesis

The prevalence and impact of Major Depressive Disorder (MDD) is steadily increasing, whilst current treatments are ineffective in a substantial proportion of individuals and accompanied by unpleasant side effects. There is a clinical need for novel therapeutic agents and for paradigms which could predict the most effective medication for an individual. Both these avenues are hindered by the lack of in-depth knowledge around the biological mechanisms underlying MDD. Notably, MDD is influenced by an interplay of genetic, environmental and psychological factors. Understanding the biological factors which may genetically predispose individuals to MDD can illuminate on potential pathological mechanisms.

Advances in genotyping technologies and the reference genome have enabled an explosion of genome-wide association studies (GWAS), which can identify genetic loci associated with a disease. Despite a slow start, GWAS in MDD have now uncovered over 500 genetic loci associated with MDD. However, translating GWAS hits into meaningful biological and clinical insight has proved challenging. Firstly, GWAS hits represent genomic loci with many highly correlated variants due to linkage disequilibrium (LD), therefore making it difficult to ascertain the causal variant and gene underlying an association. Secondly, many GWAS loci associated with MDD are in noncoding regions with unknown function. Both rare-variant analysis and multi-omics approaches can aid in the biological interpretation of GWAS results. Rare variants are typically not under selective pressure and thus do not show high levels of LD, making their associations easier to interpret. Multi-omics data can assess the impact of genetic variants on a range of molecular phenotypes and therefore giving insight into the functional impact of

genetic variants which in turn can aid their interpretation from GWAS. Also, multi-omics data downstream of genomics can give insight into environmental impact and provide a more holistic view of perturbed biological pathways.

In this thesis, I demonstrate how multi-omics analysis can expand our understanding of MDD and antidepressant exposure.

Chapter 1 introduces core concepts in the thesis, namely, MDD, antidepressants, genetic studies and multi-omics analysis. This chapter also provides the aims of this thesis. In **Chapter 2**, I detail the two population cohorts used in this thesis, alongside reference datasets. **Chapters 3 – 5** denote main analyses chapters, focused on metabolomic links to MDD (**Chapter 3 & 4**) and epigenetic analysis of antidepressant exposure (**Chapter 5**). In **Chapter 6**, I discuss the findings of this work, alongside a discussion of its limitations and potential future directions to build upon this research. Finally, **Chapter 7** describes ethical aspects to this work before **Chapter 8** concludes this thesis, summarising the findings, limitations and future directions.

1 Introduction

In this chapter, I introduce major depressive disorder (MDD), detailing its epidemiology, diagnosis and various theories surrounding its pathophysiology. I then provide historical context for the development of antidepressant medications which are highly prescribed today and how they are intertwined with various hypotheses of MDD. Furthermore, I introduce treatment-resistant depression (TRD) as this is a core component of the high global burden of MDD and the motivation for new therapeutic targets. Following this, I move to introducing candidate gene studies (CGS) and various genome-wide association studies (GWAS) of MDD. Then I introduce core limitations of GWAS including the notion of missing heritability and lack of biological insight from GWAS loci. The role of rare variants and multi-omics analysis in aiding GWAS interpretations are then discussed. Finally, the aims of this thesis and the thesis analysis chapters are outlined.

1.1 Major Depressive Disorder (MDD)

Major Depressive Disorder (MDD), commonly known as depression, is a complex psychiatric disorder which primarily manifests with a persistent low mood, lack of interest in previously enjoyed activities, and an inability to feel pleasure (anhedonia).

1.1.1 *Epidemiology of MDD*

Typically emerging in early adulthood (20-25 years)¹ and more common in females², MDD is often a chronic condition consisting of multiple depressive episodes which can be severely debilitating to an individual's daily life. This is exacerbated by the fact that MDD has high rates of comorbidities with both psychiatric and physical conditions,

including (but not restricted to) anxiety³, substance use disorders⁴, type 2 diabetes mellitus^{5,6}, cancer^{7,8}, autoimmune⁹ and cardiovascular diseases¹⁰. Additionally, MDD is associated with many negative psychosocial¹¹ and employment outcomes¹².

Furthermore, those with MDD also have an increased mortality, primarily through high rates of suicide and comorbidities relating to cardiovascular health^{10,13,14}. Therefore, on an individual and community level, MDD is a condition with far-reaching consequences. Currently, the prevalence of lifetime MDD is estimated to be between 10 and 15%, although estimates have been shown to differ in retrospective or prospective studies¹⁵. Nonetheless, epidemiological observations have shown the incidence of depression to be steadily increasing^{16–18}, and the World Health Organisation (WHO) has predicted MDD to be the leading cause of disability by 2030¹⁹. The most recent study on the Global Burden of Disease (GBD) found that over 280 million people suffer from MDD, and that MDD was the second leading cause of disability, accounting for 6.2% of all years lived with disability (YLDs) (Figure 1.1)¹⁶. MDD also has a substantial economic burden, estimated at \$333.7 billion in the United States alone²⁰. This is due to both direct healthcare costs and to indirect work-related costs such as unemployment and absenteeism, that contribute to lost productivity. Current treatments for MDD are limited in terms of their efficacy and non-response, which also contributes to the growing societal and economic impact of MDD. MDD is being increasingly recognized as a high priority global health concern, illustrated by the recent partnership of the Lancet and the World Psychiatric Association to establish a clinical commission of experts with the broad aim of reducing the world-wide burden of MDD²¹. Overall, MDD is a complex disorder, which is a critical global health priority, due to its increasing prevalence, suboptimal treatment, high comorbidities and increased mortality.

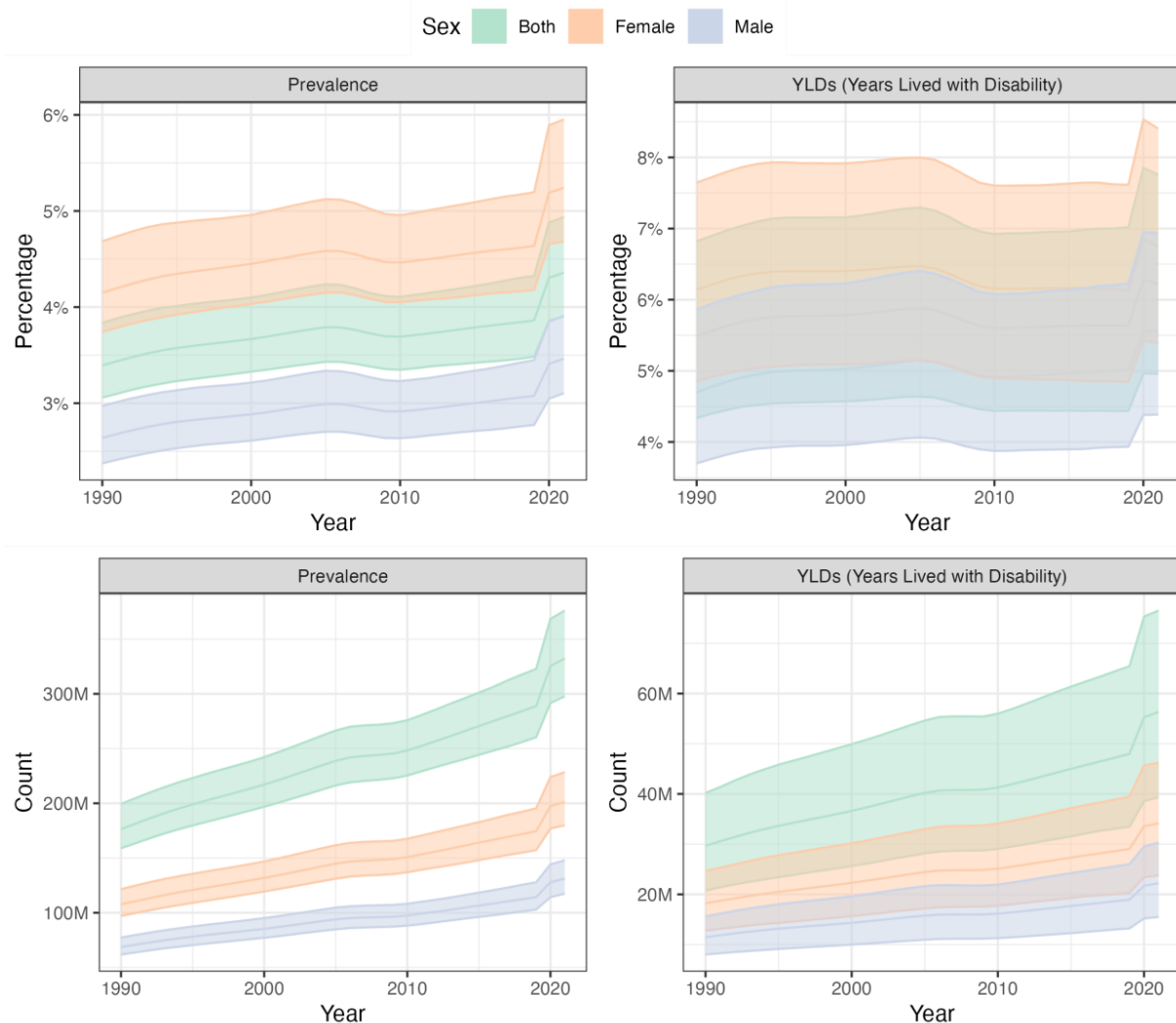


Figure 1.1 The increasing global burden of MDD

The global prevalence (left) and years lived with disability (right) in terms of % (top) and raw counts (bottom) separated by biological sex (color). Data obtained from the Global Burden of Incident Disease Study¹⁶ (<https://vizhub.healthdata.org/gbd-results/>) on 11/02/2025.

1.1.2 Biological theories of MDD

The explicit biological basis of MDD is unknown although there are various hypotheses (Figure 1.2), which are not necessarily mutually exclusive²². Three of the most wellstudied hypotheses are outlined below.

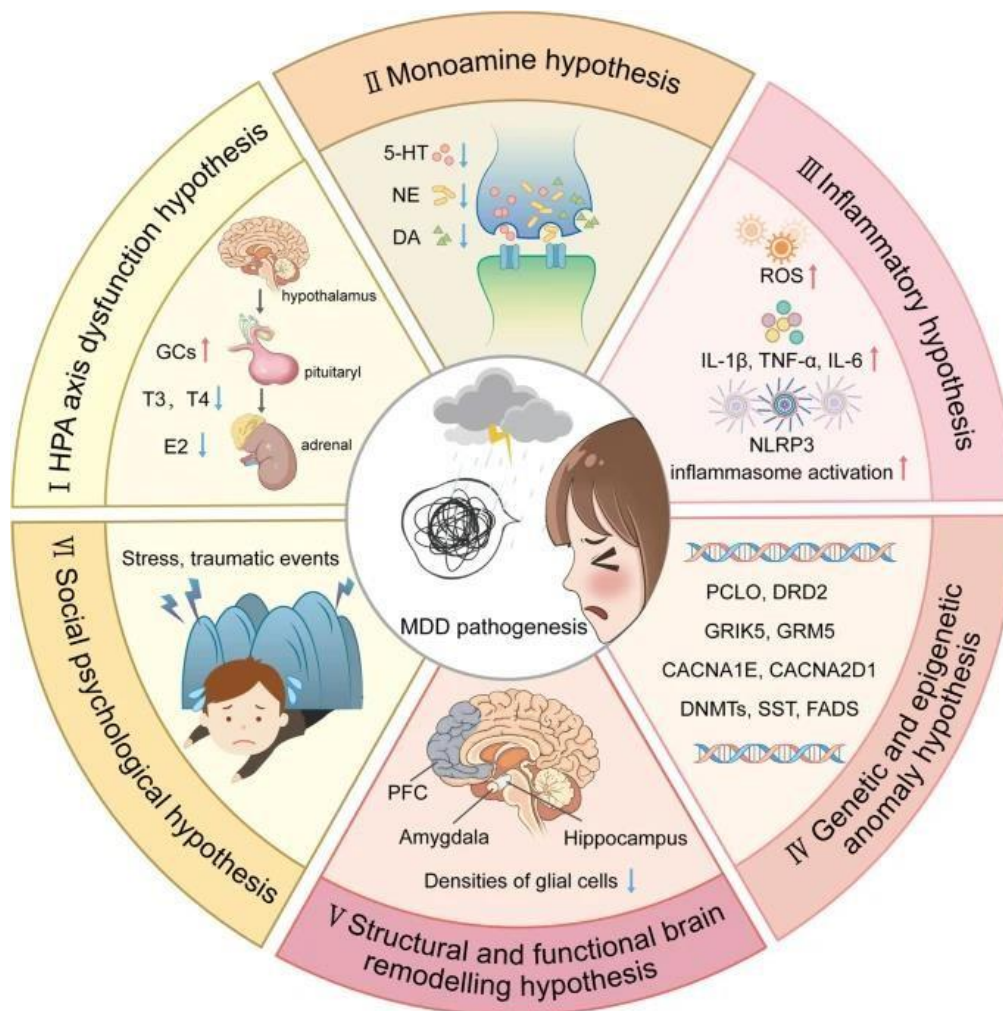


Figure 1.2 Hypotheses of MDD aetiopathogenesis

A schematic of six prominent hypotheses around the aetiopathogenesis of MDD in the literature. I) *Hypothalamic-pituitary-axis (HPA) dysfunction hypothesis*: Hormonal

dysregulation of glucocorticoids and estrogen is core to the pathogenesis of MDD. *II) The monoamine hypothesis*: Depletion of monoamines such as serotonin (5-HT) and noradrenaline (NA) is central to MDD. *III) Inflammatory hypothesis*: Increased neuroinflammation is core to MDD. *IV) Genetic and epigenetic hypothesis*: Genetic variation and epigenetic mechanisms can increase susceptibility to MDD. *V) The structural and functional brain remodelling hypothesis or the neurotrophic/synaptic plasticity hypothesis*: Reduced brain volume and synaptic plasticity is core to MDD. *VI) The social psychological hypothesis*: Traumatic or stressful life events are central to MDD pathogenesis. This figure is taken from Cui et al, 2024²². Copyright: *Signal Transduction and Targeted Therapy*

1.1.2.1 Monoamine hypothesis

The first long-standing hypothesis postulated that MDD was a consequence of decreased monoamines, such as serotonin (5-HT), noradrenaline (NA) and dopamine (DA) in the brain^{23,24}. This largely stemmed from the discovery of first-generation antidepressants which appeared to increase monoamine concentrations^{25,26}. The monoamine hypothesis is the basis of almost all antidepressant drugs which are widely prescribed today, which increase 5-HT and/or NA concentrations in the brain through various mechanisms. However, although antidepressants cause an immediate increase in monoamine concentrations, any symptomatic benefit is usually not seen until 4-8 weeks of treatment, indicating that increasing monoamine concentrations may not be the primary causal mechanism by which antidepressants improve symptoms^{27,28}. Also, antidepressants have a sub-optimal efficacy, and many people do not respond to antidepressant therapy, further casting doubt on the monoamine hypothesis²⁷. A recent

controversial umbrella review claimed that there was no evidence that 5-HT had a causal role in MDD²⁹. However, this has led to several rebuttals arguing that this review is methodologically flawed and that 5-HT concentrations are a key factor^{30–32}. More nuanced hypotheses state that rather than being a sole causal factor in MDD, reduced monoamine concentrations may be a marker of the general neural dysfunction which occurs in MDD, secondary to its pathology^{31,33}.

1.1.2.2 Neuroinflammatory hypothesis

It has also been proposed that MDD is caused by dysregulated immune function^{34,35}.

This was primarily due to the observations that immune challenges can lead to ‘sickness behaviour’³⁶ which presents very similarly to MDD³⁷, alongside observations of the disproportionate prevalence of MDD in those with autoimmune disorders^{38,39}. Many cross-sectional studies have observed an association of increased inflammatory markers, such as interleukin-6 (IL-6)⁴⁰, C-reactive protein (CRP)⁴¹ and several proinflammatory cytokines⁴² with MDD. Also, longitudinal studies have shown the association of baseline inflammatory markers with incident MDD. Specifically, the Avon Longitudinal Study of Parents and Children (ALSPAC) found a dose-response relationship between IL-6 and CRP in childhood and depressive symptoms in early adulthood⁴³. Also, longitudinal analysis in the Whitehall II cohort found baseline CRP predicted cognitive depressive symptoms at follow-up, independent of baseline depressive symptoms and medication use⁴⁴. Furthermore, it was observed that interferon-alpha therapy, introduced for cancer and Hepatitis-C and which has proinflammatory effects, often led to depressive behaviors as a side-effect^{45–47}. Additionally, a meta-analysis of clinical trials of anti-cytokines therapy for a range of chronic inflammatory conditions found that therapy significantly

improved depressive symptoms⁴⁸. Importantly, this effect was seen independently of any improvement in the primary physical illness, indicating that the reduction in depressive symptoms was a causal consequence of therapy rather than an artefact of improved physical health⁴⁸. This led to a flurry of small clinical trials investigating the role of anti-inflammatory treatments for MDD, which had mixed findings^{49–52}. A meta-analysis of these studies found a significant moderate antidepressant effect of anti-inflammatory therapies on depressive symptoms⁵³. Notably, one study investigating the antidepressant effect of an anti-cytokine antibody targeting Tumor Necrosis Factor (TNF) found that antidepressant benefit was dependent on baseline levels of TNF⁴⁹. Therefore, anti-inflammatory therapies may be more effective in a subset of MDD patients who exhibit high inflammation. This is indication that the current nosology of MDD diagnosis encompasses a range of different biological subtypes, each of which may be more responsive to different treatment strategies.

1.1.2.3 The neurotrophic/synaptogenic hypothesis

Another prominent theory of depression, the neurotrophic hypothesis, postulates that MDD results from decreased neurotrophic support resulting in neuronal atrophy⁵⁴. Support for the neurotrophic hypothesis largely comes from brain-imaging studies which found that patients with MDD have been found to have reduced brain volume of limbic brain regions⁵⁵ which may be reversed with antidepressant treatment⁵⁶. Neurotrophic factors represent a family of proteins which promote neuronal outgrowth and survival⁵⁷, and the most studied with respect to MDD is brain-derived neurotrophic factor (BDNF)^{58–60}. Several pre-clinical studies found that chronic stress, a well-known risk factor for MDD, and antidepressant therapy have opposing actions on BDNF signaling⁵⁴.

However, knocking out BDNF in rodent models is not sufficient to induce depressive behaviors, indicating that the role of BDNF in causing / alleviating MDD is more complex than a simple deficiency⁶¹. Instead, several studies have indicated that lower BDNF levels may represent a vulnerability to MDD, which may develop upon traumatic or stressful environmental exposures⁶². BDNF influences synaptic transmission and plays an active role in promoting synaptic plasticity and neurogenesis⁶³. The neurotrophic hypothesis was refined to the synaptogenic hypothesis of MDD which argues that MDD is caused by a disruption of homeostatic mechanisms controlling synaptic plasticity, resulting in destabilization of brain networks underlying mood and emotion^{59,64}. Functional imaging and electroencephalogram (EEG) studies support the notion of dysregulated brain networks in those with MDD^{55,65,66}. Synaptic plasticity relates to the strengthening or weakening of synapses dependent on experience and is the core mechanism for learning and memory^{67,68} and is crucial for cognitive function⁶⁹. It is well documented that those with MDD commonly exhibit cognitive impairments, usually presenting as a negative affective bias leading to pessimistic judgement and catastrophizing⁷⁰. Further evidence for this hypothesis arose following the introduction of Ketamine, a psychomimetic agent with effects on synaptic plasticity which exhibits rapid antidepressant effects (< 2 hours)^{71,72}. Ketamine is a non-selective N-methyl-D-aspartate (NMDA) receptor antagonist and has been demonstrated to rapidly increase synaptogenesis and reverse neuronal atrophy caused by stress⁷¹. Furthermore, the antidepressant effects of ketamine are dependent on the mammalian target of rapamycin (mTOR) signaling pathway⁷³, a key mechanism of synaptic plasticity⁷⁴. Overall, while traditional antidepressants chemically increase monoamine concentrations, their delayed-onset and sub-optimal efficacy indicate that MDD is not

simply the result of neurotransmitter imbalance. Various studies have identified the role of immune dysfunction, however the utility of using anti-inflammatory agents for the treatment of MDD is unclear and may be more effective in certain biological subtypes of MDD which exhibit higher inflammation. Finally, studies have shown that MDD is associated with impaired synaptic plasticity and holistic network dysfunction, emphasized by the rapid antidepressant actions of Ketamine. Together, these findings illustrate the biological complexity of MDD.

1.1.3 *Diagnosis of MDD*

Although conceptualized as a single disorder, MDD often has varied presentations and is increasingly thought to act as an umbrella term for different disorders with distinct etiologies⁷⁵. The diagnostic criteria for MDD as stated by the Diagnostic and Statistical Manual of Mental Disorders (DSM)⁷⁶ is shown in **Table 1.1**. Essentially, the diagnosis of MDD is based on the presence of at least 5 symptoms over a minimum of 2 weeks from 9 broad, often dichotomous options including weight changes, fatigue, altered sleeping and recurrent thoughts of death. Therefore, often completely different symptomatic presentations can result in the same MDD diagnosis and severity scores⁷⁷, although they may reflect biologically distinct mechanisms. Attempts to disentangle this heterogeneity include subtyping MDD according to the age-of-onset (early-onset MDD)⁸¹, contextual life events (post-partum MDD)⁸² and number of episodes (recurrent MDD)⁸³. Broad clusters of specific symptoms have also been used to subtype MDD to an extent, such as atypical MDD, which is currently defined as a diagnosis of MDD with mood reactivity, increased weight, hypersomnia and interpersonal rejection hypersensitivity⁷⁸. Different MDD subtypes have been shown to share a genetic component with other

psychiatric disorders, indicating shared biological pathways⁷⁹. A study found that typical and atypical MDD shared genetic effects with SCZ and body-mass-index (BMI) respectively⁷⁹ whilst post-partum MDD showed important shared genetic etiology with BPD⁸⁰. The symptom overlaps, high co-morbidities and shared genetic effects across psychiatric disorders indicate that their current nosology as separate entities in the DSM does not reflect underlying pathophysiological processes⁸¹. However, increased understanding of causal molecular mechanisms which may underlie MDD subtypes could enable more objective, biologically informed subtyping. In turn, this could also benefit the implementation of precision medicine approaches for patient stratification to certain medications⁸².

Table 1.1 The diagnostic criteria for Major Depressive Disorder (MDD) from the Diagnostic and Statistical Manual for Mental Disorders (V).

“A” criteria represent core symptoms of MDD, of which one must be met for a diagnosis of MDD. “B” criteria represent other symptoms of MDD that can contribute to a diagnosis (following the presence of 5 or more symptoms).

DSM-5 Major Depressive Disorder	Symptoms
“A” criteria	Depressed Mood
	Markedly diminished interest or pleasure in almost all activities
“B” criteria	Significant weight loss/gain or decrease/increase in appetite
	Psychomotor agitation or retardation
	Fatigue or loss of energy
	Feelings of worthlessness or excessive/inappropriate guilt
	Fatigue or loss of energy
	Diminished concentration or indecisiveness
	Recurrent thoughts of death, suicidal ideation, plans or an attempt

1.2 Antidepressants

Pharmacological treatments for MDD, known as antidepressants, are the most common first-line treatment for MDD and are widely prescribed. However, the exact mechanisms of how they lead to therapeutic effects are unknown. Rational drug design usually begins with a detailed understanding of pathology underlying the disease of interest, enabling the identification of logical drug targets and effective treatment⁸³. However, modern-day pharmaceutical antidepressants did not follow this trajectory, and their development was instead heavily influenced by serendipitous observations in the 1950s²⁵. Here I describe

the evolution of antidepressant therapies from these discoveries to modern-day, illustrating how these have contributed to the biological theories of MDD.

1.2.1 *First generation antidepressants*

The first serendipitous observation was that Reserpine, an anti-hypertensive drug, both depleted 5-HT levels in the brain and often resulted in clinical depression at high doses^{84,85}. Furthermore, Iproniazid, an antimycobacterial agent, was found to improve mood in patients in a tuberculosis trial⁸⁶. Further study found that Iproniazid inhibited monoamine oxidase (MAO), a mitochondrial enzyme which prevents the degradation of monoamines in the presynaptic terminal⁸⁷. This quickly led to the development MAO Inhibitors (MAOIs) for the treatment of MDD²⁵. However, MAOIs had significant adverse side effects, including a potentially lethal interaction with tyramine-rich foods, requiring strict dietary control during treatment⁸⁸. Around the same time, tricyclic antidepressants (TCAs) were developed following observations of improved mood in schizophrenia (SCZ) trials^{89,90}. TCAs also increase the levels of monoamines in the brain primarily through blocking the re-uptake of NA and 5-HT⁸⁷. However, TCAs are pharmacologically non-selective and also interact with cholinergic and histaminergic receptors resulting in various atropine-like side effects such as dizziness, sedation and blurred vision⁹¹. Additionally, in overdose, TCAs were linked to seizures and cardiotoxicity⁹². These first-generation antidepressants (MAOIs and TCAs) showed significant efficacy in treating MDD compared to placebo in trials and represented a significant milestone in the treatment of MDD²⁵. Both treatments primarily increase the levels of monoamines in the brain, and influenced the subsequent monoamine hypothesis of MDD, which postulated MDD to be caused by a deficit of monoamines in the brain²³ (1.1.2.1). However, their

substantial side effects limited their use at a large scale, and subsequent development was focused on improving the safety and side effects of these monoamine-increasing drugs, whilst maintaining their efficacy²⁵.

1.2.2 *Second generation antidepressants*

Second-generation antidepressants emerged in the 1980s with selective serotonin reuptake inhibitors (SSRIs) which were designed to selectively target levels of 5-HT by inhibiting the serotonin reuptake transporter⁹³. With this selectivity, SSRIs maintained the efficacy seen with first-generation antidepressants but with much improved tolerability, ultimately leading to improved compliance and outcomes^{93,94}. Most importantly, SSRIs were much safer in potential overdose than previous medications, with a low risk of seizures and toxicity⁸⁷. The selective nature of these drugs also reduced the muscarinic side effects which accompany TCA medication⁹¹. Furthermore, SSRIs were shown to be effective and better-tolerated treatments for many other psychiatric conditions, such as anxiety disorders⁹⁵, panic disorders⁹⁶, obsessive compulsive disorder⁹⁷, bulimia nervosa⁹⁸ and chronic pain⁹⁹. Despite these improvements, SSRI treatment is still accompanied by non-negligible side effects, including gastrointestinal issues, restlessness and sexual dysfunction⁹¹. Perhaps most concerning was the association of increased suicidal ideation in the initial period of treatment¹⁰⁰, especially in younger individuals¹⁰¹. However, it has been shown that SSRIs are not associated with increased risk of completed suicide, and with appropriate management of these side effects, are associated with a reduction in suicide rates¹⁰². Therefore, SSRIs are overall a major improvement on previous antidepressants and are internationally recommended as the

first line of treatment for MDD¹⁰³⁻¹⁰⁵. The introduction of SSRIs led to an rapid increase of antidepressant prescriptions, which continue to rise today^{106,107} (Figure 1.3).

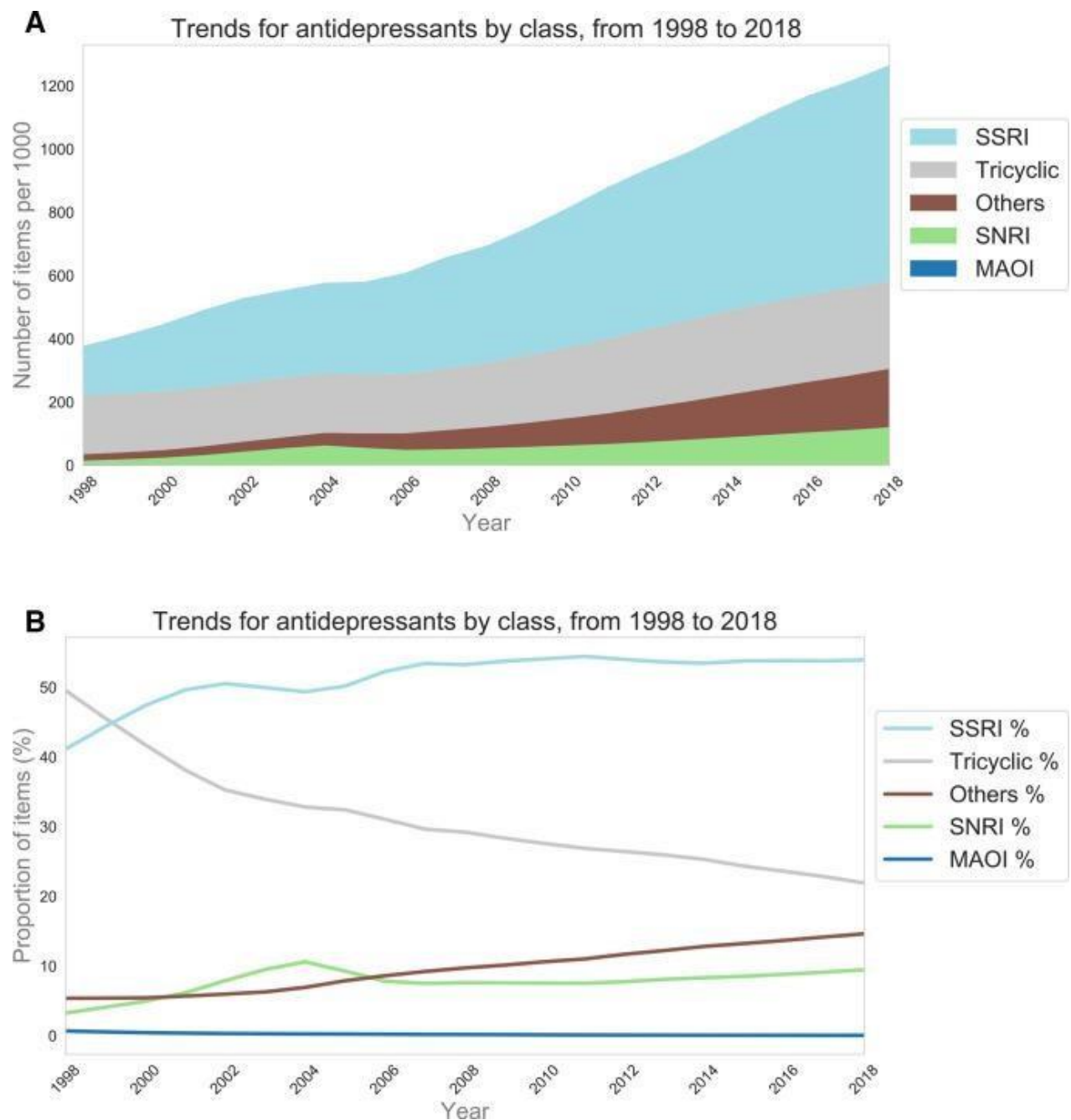


Figure 1.3 Rising rates of antidepressant prescriptions from 1998-2018

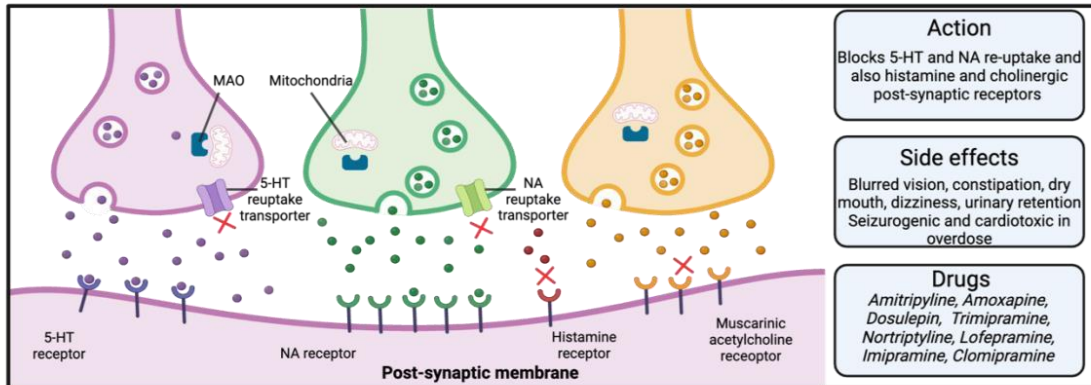
The population standardized number of antidepressant items per 1000 (A, top) and proportion of antidepressant items (B, bottom) separated by class (color) from general practices (GP) in England from 1998-2018. SSRI = selective-serotonin reuptake inhibitor, SNRI= serotonin-noradrenaline reuptake inhibitor, MAOI = monoamine oxidase reuptake

inhibitor. Practice level summaries of prescribing activity are freely available from NHS Digital and compiled by OpenPrescribing.net. This Figure is taken from Bogowicz et al, 2021¹⁰⁶. Copyright: *BJGP Open*.

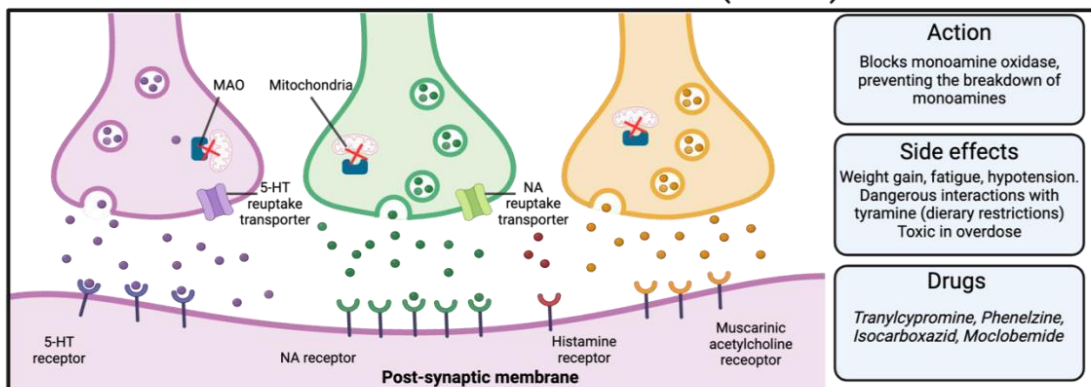
Alternative treatments have also been developed, all focusing on different mechanisms to increase 5-HT and NA transmission in the brain through slightly different mechanisms. These include serotonin-noradrenaline re-uptake inhibitors (SNRIs), which inhibit both 5-HT and NA reuptake and are thought to be more efficacious than SSRIs for severe MDD cases¹⁰⁸. These are accompanied by similar side effects profile to SSRIs, such as headaches, fatigue and sexual dysfunction and in rare cases, hypertension^{91,109}. Another class of antidepressants are termed 'Atypical', as they do not fit into the neat categories of 5-HT and/or NA inhibitors, although all still ultimately result in increased monoamine concentrations. For example, Mirtazapine is a common atypical antidepressant which has a sophisticated dual mechanism at several 5-HT postsynaptic receptors, alongside a presynaptic NA receptor¹¹⁰. This treatment is less prone to sexual dysfunction side effects, although it is associated with weight gain and sedative effects¹¹⁰. There is some evidence that Mirtazapine has a quicker onset of action^{111,112}, yet the overall efficacy is comparable to SSRIs¹¹¹. Ultimately, there are a host of antidepressant agents now available which are relatively safe but each having a non-negligible side effect profile (Figure 1.4).

● Serotonin/5-HT ● Noradrenaline/NA ● Acetylcholine ● Histamine

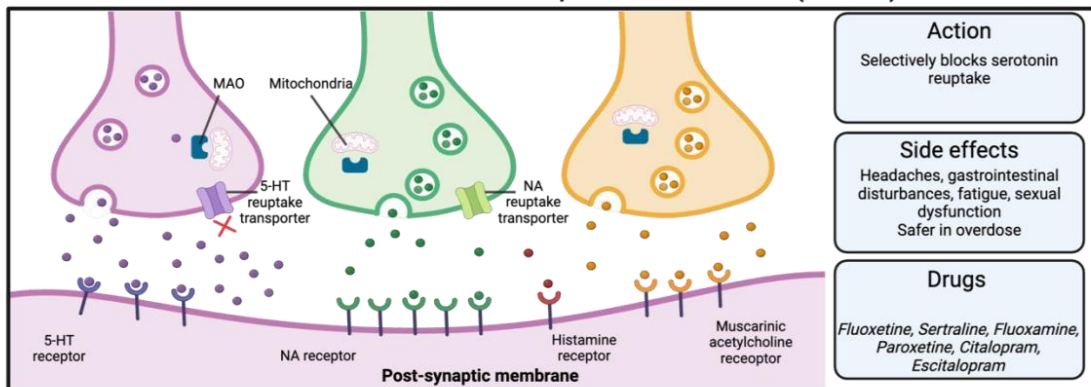
Tricyclic Antidepressants (TCAs)



Monoamine Oxidase Inhibitors (MAOIs)



Selective Serotonin Re-uptake Inhibitors (SSRIs)



Serotonin-Noradrenaline Re-uptake Inhibitors (SNRIs)

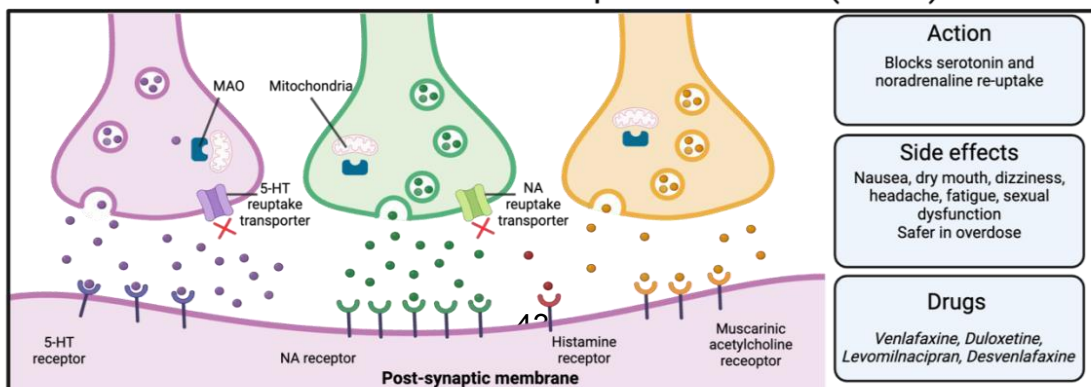


Figure 1.4 Mechanisms, side effects and example drug names of commonly prescribed antidepressants.

Each panel represents a schematic of a synapse with presynaptic terminals (top) from serotonergic (purple), noradrenergic (green) and histaminergic (yellow) neurons with a post-synaptic membrane (bottom). The red crosses mark the biochemical target of each antidepressant class, with a description of the action, side effects and drugs for each antidepressant class on the right. Different neurotransmitters are represented by differently colored circles. This Figure was made using Biorender.

1.2.3 *Antidepressant treatment resistance*

Although safer second-generation antidepressants have been transformative in the international treatment of MDD compared to previous alternatives, there are two core challenges to their effective treatment of MDD. Firstly, all antidepressants discussed hitherto have a delayed onset of action and can take between 4-8 weeks to see symptomatic improvement⁸⁷. Moreover, despite the efficacy of antidepressants over placebo observed in clinical trials¹¹³, many individuals do not respond to treatment in clinical practice¹¹⁴. It is estimated that only 45% of patients respond to treatment (~50% reduction in symptoms) and only 35% of patients achieve remission (reduction of symptoms below MDD diagnostic threshold)¹¹⁵. Additionally, more successive antidepressant treatment trials were associated with reduced likelihood of remission and increased likelihood of future relapse^{116,117}. Importantly, remission is the overarching aim of antidepressant treatment, as residual MDD symptoms above a diagnostic threshold are associated with a significant reduction in quality of life and higher rates of relapse^{118,119}. Treatment-resistant depression (TRD) is broadly defined as MDD showing

an inadequate response to at least two courses of antidepressant therapy^{114,120,121}.

Estimates of the prevalence of TRD range from 20 to 60% of those with MDD, due to heterogeneous definitions and different studied populations¹²¹. Nevertheless, TRD and residual symptoms are associated with a host of negative outcomes, such as impaired psychosocial functioning, higher rates of relapse, increased risk of suicide, increased comorbidities and reduced quality of life¹²². Also, TRD is estimated to present an additional societal cost between \$29-48 billion annually¹²³, due increased use of medical resources¹²⁴ alongside higher rates of unemployment¹²³.

Treatment for those with TRD usually consists of trialing various courses of antidepressant therapies, which given the delayed onset of action and range of side effects, is a bleak experience for those already suffering a debilitating condition. Attempts have been made to establish the best plan of action for those who do not respond to an initial course of an SSRI including different switching and augmentation strategies. This includes the Sequenced Treatment Alternatives to Relieve Depression (STAR*D) trial, which studied a large naturalistic population and found that there was no 'clear winner' in terms of treatment strategies following a non-response to Citalopram^{115,116,125}. However, given the complex and varied presentations of MDD⁷⁵, it is perhaps unreasonable to expect there to be a 'one size fits all' approach in its treatment. Finding evidence-based personalized strategies for MDD treatment is an active area of research with so far mixed results¹²⁶. A major challenge in developing this type of precision psychiatry is predominantly due to lack of knowledge for potential biological subtypes of MDD, which could inform potential treatment.

1.2.4 *Rapid acting antidepressants*

A new era in antidepressant treatment may have begun with the approval of the intranasal spray esketamine for the treatment of TRD by the US Food and Drug Administration (FDA)¹²⁷. Esketamine, an isomer of ketamine, acts as a high-affinity antagonist at the NMDA receptor, which is critical for synaptic plasticity, learning and memory. The antidepressant effects of ketamine were first reported in 2000, with a rapid-onset (within 4 hours) and with sustained effects (up to 7 days)⁷². Subsequent studies validated these findings, also showing that ketamine was a promising treatment for TRD^{128,129}. Ketamine also provides relief from suicidal ideation and presents important progress for the emergency care from those at risk of suicide¹³⁰. However, ketamine is not a miracle drug and there are limitations to its widespread and long-term use¹³¹. Firstly, as it can be associated with transient hypertension and dissociative effects immediately after administration, patients require a minimum of 2-hour monitoring within a health-care setting, which is impractical on a large-scale and likely to contribute to non-adherence¹³¹. Additionally, the efficacy of ketamine relies upon its repeated administration, raising concerns about the potential harmful consequences of prolonged ketamine use such as liver injury and neurocognitive impairment¹³². A recent systematic review of sustained ketamine uses over 1 year found minimal evidence for these adverse effects postulating that the low doses currently approved are not sufficient for such side effects¹³². Finally, as ketamine has historically been an illicit drug used for recreational purposes, there have been concerns about the potential for abuse and addiction¹³³, yet there is little evidence to support this outcome in practice^{131,133}.

In summary, commonly prescribed antidepressants today are inherently still based upon chance observations dating back to the 1950s. Whilst these medications represent an important step in the history and treatment of MDD, they are limited in terms of their delayed onset of action, side effect profiles, and sub-optimal action in a substantial proportion of patients. Increased understanding regarding the pathophysiology of MDD and the molecular mechanisms of antidepressants themselves may aid in identifying novel treatment targets, as well as potentially stratifying patients to different personalized treatment options.

1.3 Genomic studies of MDD

Genetics can provide meaningful insights into disease etiology and be impactful in the development of novel therapeutics. It is estimated that putative drug targets which are supported by genetic evidence are 2.6 times more likely to be successful in clinical trials^{134,135}. It has long been observed that MDD often runs in families, indicating a genetic influence¹³⁶. In 2000, Sullivan et al, (2000) performed a meta-analysis of family based studies and found empirical significance of familial aggregation in MDD, with strong association of the diagnosis of MDD between first-degree relatives¹³⁷.

Furthermore, using various twin studies, Sullivan et al, (2000) estimated the heritability (the proportion of variation in liability due to additive genetic factors) of MDD to be between 31 and 42%¹³⁷. Therefore, there is a substantial genetic component to MDD. A key research objective in better classifying and treating MDD is to understand more about the genetic factors which may predispose individuals to the development of MDD.

1.3.1 *Candidate gene studies*

First attempts to investigate this genetic basis of MDD were focused on candidate gene studies (CGS), which hypothesized that a major component of genetic influence on a phenotype was due a functional mutation on a biologically relevant gene¹³⁸. This involved identifying an allele or set of alleles within candidate gene hypothesized to have a role in the disease and testing the difference in allele frequencies¹³⁹. Critically, the success of CGS relies on robust biological knowledge of etiological mechanisms for the identification of plausible candidate genes. At the time, candidate genes for MDD were selected largely based on the monoamine hypothesis, with most of the candidate gene investigations focusing on the serotonin-transporter-linked promoter region (5HTTLPR) within *SLC6A4*, a gene which encodes the serotonin transporter protein (SERT)¹⁴⁰. Despite a flurry of studies investigating the association of candidate genes with MDD, there was a worrying lack of replication which questioned the validity of CGS in investigating genetics of MDD^{141,142}. Of particular note, a prospective longitudinal study (n = 837) found evidence of 5-HTTLPR allele moderating the effect of stressful life events on developing MDD¹⁴³. This was a highly influential study in psychiatric genetics, indicating a gene-environment interaction in MDD¹⁴³. However, subsequent meta analyses, including a pre-registered report on nearly 39,000 individuals¹⁴⁴ and another prospective longitudinal study in a similar population found no evidence of the moderation effect of 5-HTTLPR¹⁴⁵. Many have attributed the reproducibility crisis in CGS of MDD to individual studies being severely underpowered to detect (what we now know to be) individual small genetic effects¹⁴², inadequate control of population stratification¹⁴⁶ and strong publication bias¹⁴⁷. Additionally, CGS are inherently limited as they do not generate new biological insights beyond validating the existing hypotheses they are built

upon. In contrast, a comprehensive genome-wide approach could identify genetic factors which are not yet considered and shed light on previously unknown biological pathways underlying MDD.

1.3.2 *Genome-wide association studies*

1.3.2.1 GWAS era

The completion of the Human Genome Project (HGP), an international effort to sequence the entire human genome, introduced a new era in investigating the role of genetics in disease^{148,149}. It was found that any two human genomes were roughly 99.9% identical, with the 0.01% variation being responsible for all heritable differences between individuals. The most common genetic variation presents as a singular base pair substitution between chromosomes known as a single nucleotide polymorphism (SNP)¹⁴⁹. However, directly testing the association of all SNPs by directly sequencing was not feasible due to the limitations and cost of sequencing technologies^{150,151}. An indirect approach was proposed, whereby genomic regions tagged by index SNPs were tested for an association¹⁵¹. This approach is based upon the principle of linkage disequilibrium (LD) in the population, defined as the non-random association between alleles at different genomic loci due to evolutionary forces¹⁵². Generally, the LD between alleles is stronger for those which are physically closer together on a chromosome, and results in sets of alleles often being inherited together, known as a haplotype¹⁵³. The LD structure in the human genome was extensively catalogued in the Haplotype Map (HapMap) project and resulted in the identification of a subset of SNPs which, due to LD, tagged most of the common genomic variation in several human populations^{153,154}. At the

same time, commercial companies developed specialised genotype arrays to genotype every index SNP in a single assay which could then be used to impute ungenotyped SNPs¹⁵⁰. This provided an automated and affordable way to capture most of the common genetic variation within individuals in large population studies, enabling large-scale genome-wide association studies (GWAS)¹⁵⁰.

In 2007, the first large-scale GWAS using SNP array was conducted for seven common diseases (2000 cases each and 3000 shared controls)¹⁵⁵ and illustrated the validity of GWAS in identifying novel variant- disease associations. Furthermore, GWAS data could be used to estimate the proportion of the total variance in liability to a disease explained by common genetic SNPs (SNP-heritability, h^2_{SNP}) from GWAS summary statistics^{156,157}. It became evident that complex traits were highly polygenic (influenced by many genetic variants), and that each individual genetic variant had a small effect. Methods were also developed to provide an aggregated measure of genetic risk for polygenic diseases (polygenic risk scores; PRS), weighted using GWAS summary statistics, which may be clinically useful in identifying individuals at high risk of a particular disorder¹⁵⁸. As sample sizes and the subsequent power to detect associations of a small effect increased, the number of variant-trait associations identified by GWAS has exponentially increased. To date (21/03/2025), there are 786,898 variant-trait associations uploaded to the GWAS catalogue, a resource of published GWAS studies¹⁵⁹.

1.3.2.2 Initial GWAS

Initial attempts to identify genetic variants associated with MDD using GWAS were disappointing, with numerous studies reporting no significantly associated variants¹⁶⁰⁻¹⁶⁵.

However, most of these studies were focused on individual cohorts with small sample sizes ($N < 3000$). With increasing recognition of the large samples required to identify genetic associations with complex traits (typically of small effect), the Psychiatric Genetics Consortium (PGC) was established with the aim of combining studies to increase the sample size and the power to identify meaningful genetic associations relevant to the etiology of MDD¹⁶⁶. However, the first PGC GWAS (PGC1) in 18,000 individuals (9,000 clinically ascertained MDD cases) from 9 primary cohorts did not identify any significant associations¹⁶⁶, despite having a h^2_{SNP} of ~ 0.2 ¹⁶⁷. Furthermore, another GWAS of depressive symptoms in over 30,000 individuals in the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) consortium also failed to identify any significant associations¹⁶⁸.

The distinct lack of significant associations for MDD at this stage compared to other complex psychiatric traits such as SCZ¹⁶⁹ was due to various reasons. One is that MDD has a relatively high prevalence ($\sim 15\%$) in the population, especially compared to other traits with GWAS success such as SCZ ($\sim 1\%$). This means that differences in liability between cases and controls are less pronounced and therefore the power to detect differences between them is reduced¹⁷⁰. Using power calculations, Wray et al, (2012) estimated that a MDD GWAS would require a sample size 2.4-fold greater than SCZ to identify a variant with the same effect size¹⁶⁵. Another factor is that the heritability of MDD (~ 0.37) is considerably lower than SCZ (~ 0.81). Heritability of a trait is dependent on the number of risk alleles, risk allele frequencies and/or risk allele effect size. Assuming roughly equivalent number of risk alleles for MDD and SCZ, the lowered heritability of MDD means that each risk allele is of smaller effect and therefore require greater power to identify¹⁶⁵. Wray et al, (2012) also estimated that an MDD GWAS would require a sample

sizes 4-5 times greater than that for SCZ to detect variants explaining an equal portion of genetic variance¹⁶⁵. Finally, the heterogeneity of MDD also limits the power of MDD GWAS. A current MDD diagnosis may encompass biologically distinct subtypes, each with its own genetic architecture. When these subtypes are amalgamated, their unique signals may become diluted and reduce the ability to detect significant associations.

1.3.2.3 Increasing power in GWAS

Increasing the power in MDD GWAS relied upon increased sample size and/or concerted efforts to reduce the heterogeneity in MDD cases^{165,170}. The China, Oxford, and Virginia Commonwealth University Experimental Research on Genetic Epidemiology (CONVERGE) consortium took the latter option and analysed a more homogenous group of severe, recurrent MDD in Han Chinese women¹⁷¹. Although this study had just 5,000 cases, it was successful in identifying two genome-wide significant associations, illustrating the utility of reducing heterogeneity in MDD GWAS. However, obtaining detailed clinical diagnoses in cohort studies is expensive and time-consuming. Other approaches focused on increasing sample size by using different measures of MDD¹⁷² and/or relying on self-report metrics¹⁷³, although this was often at the expense of increasing MDD heterogeneity. The Social Science Genetic Association Consortium (SSGAC) consortium increased sample size by meta-analysing GWAS of genetically correlated measures of depression. Specifically, they combined clinical MDD GWAS with a GWAS on measure of 'depressive symptoms' in the UK Biobank (UKB, N ~ 180,000) and found 2 genome-wide significant associations with depressive symptoms. Another way of drastically increasing sample size is through minimal phenotyping, demonstrated by the efforts from the genetic testing company 23andMe (Mountain View,

CA). Specifically, they defined MDD cases as those who self-reported measures of previous clinical diagnosis of MDD ($N = 75,000$) and controls who self-reported no history of MDD ($N \sim 230,000$), and identified 17 significant genetic associations, 2 of which replicated in an independent sample¹⁷³. However, both these approaches had considerably lower h^2_{SNP} estimate of MDD (SSGAC ~ 0.04 and 23andMe ~ 0.06) than that from the PGC based on exclusively clinically ascertained samples ($h^2_{\text{SNP}} \sim 0.2$), indicating a potential dilution of genetic signal through combining different MDD definitions and relying on self-report. A thorough comparison of genetic architecture of different MDD definitions in UKB found that self-report measures captured non-specific effects on MDD compared to strictly defined MDD phenotype¹⁷⁴. Nonetheless, the increased power in discovery gained through these approaches were seen to outweigh the potential increase in heterogeneity, especially with minimal phenotyping which is more cost-effective to implement in large population cohorts. Crucially, the increased power allows for the detection of genetic associations which can then be further examined regarding their relevance to subtypes of MDD, rather than missing associations entirely due to insufficient sample size.

As such, sample sizes and numbers of identified genome-wide significant associations steadily increased. The second PGC (PGC2) study utilised data from international cohorts which ascertained MDD in various ways including formal diagnostic structured clinical interviews, electronic health records and self-report of clinical diagnosis (23andMe) or clinical cut-off of depressive symptoms (UKB, first release), resulting in $\sim 135,000$ cases and $\sim 344,000$ controls¹⁷⁵. This study identified 44 genetic risk variants for MDD, alongside a h^2_{SNP} of 0.08. Importantly, the analysis established the validity of meta-analysing these cohorts by estimating the common variant genetic correlations

alongside out-of-sample prediction of MDD using PRS¹⁷⁵. Furthermore, a GWAS of three MDD definitions in UKB identified 17 concordant significant genetic loci in all definitions alongside a strong genetic correlation between clinically diagnosed MDD and a broader self-declared phenotype (“Have you ever seen a psychiatrist for nerves, anxiety, tension or depression?”)¹⁷⁶. A meta-analysis of this broad MDD/nerves phenotype in MDD with 23andMe self-report GWAS and PGC2 (without 23andMe and UKB samples), totalling ~ 800,000 individuals (~240,000 cases), identified 102 genetic risk variants¹⁷⁷.

1.3.2.4 Multi-ancestry GWAS

All the large, well-powered GWAS discussed thus far, apart from the CONVERGE consortium, were conducted in predominantly European populations. Different ancestral groups typically have different allele frequencies meaning that identified associations may not be generalisable to another¹⁷⁸. Accordingly, improving the genetic diversity in MDD GWAS is a major priority, to ensure that any biological insights are of equal benefit and do not contribute to ongoing health disparities¹⁷⁸. Furthermore, including different ancestral groups with different LD patterns can boost power and aid in fine-mapping GWAS associations to help to discover putative causal variants¹⁷⁹. In 2021, Levey et al,¹⁸⁰ combined the GWAS from Howard et al, (2019) (predominantly European ancestry) with an analysis of ~310,000 participants from the U.S Department of Veterans Affairs Millions Veterans Program (MVP), resulting in an MDD GWAS in 1.2 million participants and identifying 183 genetic risk loci (223 significant SNPs). The MVP had a large sample of African American participants (~ 60,000), enabling a well powered bi-ancestral GWAS of MDD. Notably, including individuals of African ancestry increased the number of

significant SNPs (+10) and genetic risk loci (+5) compared to the European-only analysis. A recent multi-ancestry GWAS of MDD of almost 1 million participants (MDD cases ~ 88,000) of African, East Asian, South Asian and Hispanic/Latin American ancestry from 21 different cohorts illustrated the importance of genetic diversity in MDD GWAS in investigating the generalisability of identified associations¹⁸¹. This study used power-adjusted transferability (PAT) ratios, an approach to account for differences in sample size, LD and minor allele frequencies (MAF) between populations which may affect the power to detect associations, to assess the transferability of loci across populations¹⁸². It found that genetic loci associated with MDD in European populations had a low transferability low in African (PAT = 0.27), East Asian and South Asian (PAT = 0.29) populations, and a slightly higher transferability in the Hispanic/Latin American (PAT = 0.63) group. Furthermore, this study demonstrated the benefit of including multiple ancestries in the fine-mapping, with an additional 22 loci resolved to <10 putatively causal variants compared to European-only analysis¹⁸¹.

1.3.2.5 PGC3

In early 2025, the PGC published the third GWAS (PGC3) using 688,808 cases and 4.3 million controls across 109 cohorts of multiple ancestries¹⁸³. This study estimated the h^2_{SNP} to be 8.4% and identified 635 genome-wide significant loci, confirming the high polygenicity of MDD. A high proportion of the associated loci were novel (46%), and the inclusion of ancestrally diverse samples (23% of this study) was influential in identifying 100 of these associations. Using structural equation modelling (SEM), this study also empirically investigated the impact of using different criteria used for MDD phenotyping, from either a clinical interview, electronic health records (EHR), questionnaires or self-

report. Using clinical diagnosis as a reference (loading = 1), the other MDD phenotyping methods (EHR, questionnaire, self-report) loaded strongly onto a common factor, although self-report was slightly lower (0.85 vs 0.92-0.95). Furthermore, only one locus showed evidence of significant heterogeneity between phenotypes, supporting the use of multiple phenotypes.

Despite a slow start, through novel phenotyping methods and vast sample sizes, GWAS has now been successful in identifying genetic variants associated with MDD (Figure 1.5). However, translating these associations into meaningful biological insight has proved challenging.

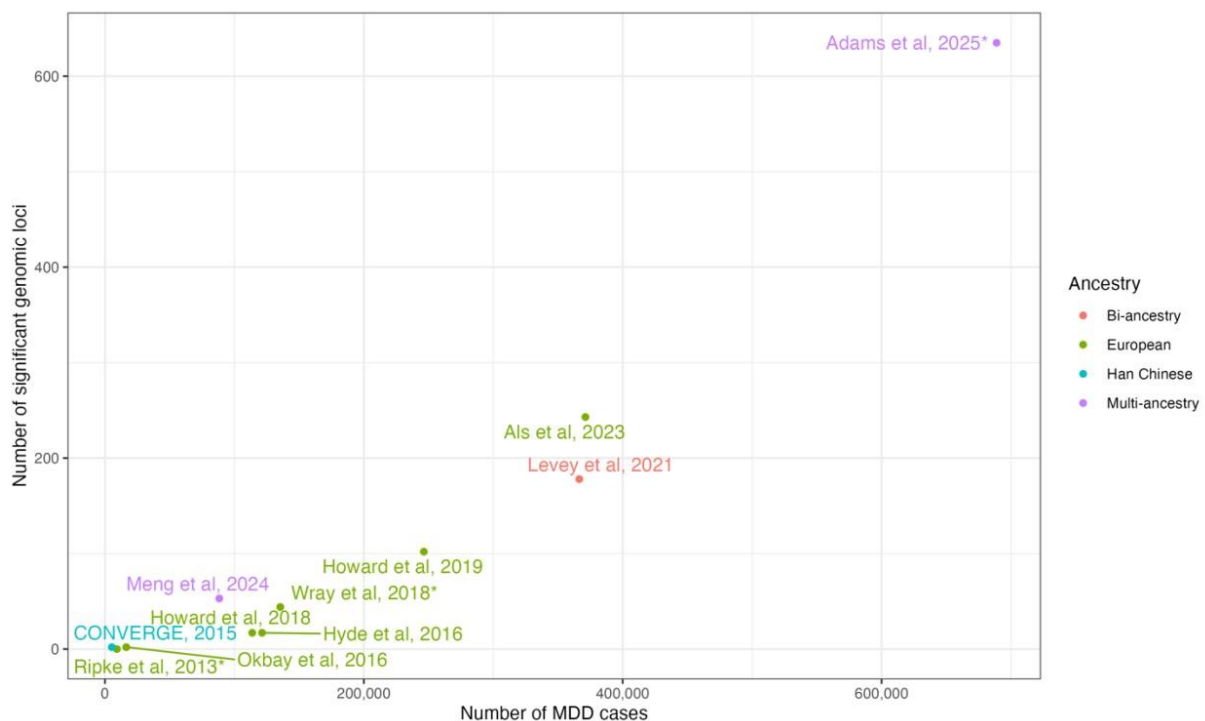


Figure 1.5 Discovery of genomic loci associated with MDD.

Each individual GWAS is shown as a dot, with the number of MDD cases (the limiting factor for effective sample size) on the x-axis, and number of genome-wide significant

loci on the y-axis. The ancestry group(s) used within each GWAS are represented by colour. Each of the PGC analyses (Ripke et al, 2013¹⁶⁶, Wray et al, 2018¹⁷⁵ and Adams et al, 2025¹⁸³) are denoted with an * in their label. This graph shows the overall trend that increasing sample size is correlated with increased discovery of associated genomic loci.

1.3.3 *Challenges interpreting GWAS associations*

The LD structure within the genome was key for enabling large-scale GWAS through the identification of “tag” SNPs and creation of cost-effective micro-arrays. However, by this design, a GWAS association simply represents the association of a genomic region with high LD, and does not pinpoint the variant, or gene which may be driving the signal¹⁸⁴.

Extensive simulations have shown that the probability of tag SNP being the true causal variant decreases with small effect sizes^{185,186}, as seen in MDD. The patterns of LD can be complex, making it challenging to pinpoint the underlying causal variant within an associated genomic locus- a process known as fine-mapping¹⁸⁷. Furthermore, > 90% of GWAS associations are in non-coding (intronic or intergenic) regions of the genome without a clear mechanistic impact on biological function¹⁸⁸. Subsequently, interpreting the biological significance of GWAS associations has proved a major challenge. It was observed that most GWAS associations were concentrated in regulatory elements of the genome^{189,190} and systematically perturbed transcription factor recognition sequences¹⁹¹, indicating that associated loci act by regulating gene expression.

Intuitively, it was assumed that a regulatory variant would be having effect on the closest proximal gene¹⁹². However, this is not always the case as there are often long-range regulatory interactions between enhancers and genes which are far apart¹⁹³. A prime

example of this is the strong genetic association of SNPs within the intronic region of *FTO* and obesity. Despite being located within the *FTO* gene, these variants did not influence the expression of *FTO*, but instead influenced the expression of distant *IRX3* gene, which resulted in altered body mass and composition^{194,195}. Therefore, understanding the impact of a GWAS association on regulatory function relevant to disease aetiology often requires follow-up functional assays. However, the sheer volume of GWAS associations in large-cohort studies makes individual functional follow-up studies for each genomic locus unachievable¹⁸⁴. Multi-omics can bridge this gap by using high-throughput molecular assays to systematically evaluate the impact of genetic variation on molecular phenotypes, in turn aiding the interpretation of GWAS associations.

1.3.4 *Missing heritability*

Due to their focus on common variants by design, GWAS are inherently based on the ‘common-disease, common-variant’ (CDCV) hypothesis, which posits that common diseases are largely attributable to common allelic variants present in more than 1-5% of the population¹⁵⁰. As a result, many expected GWAS to identify the genetic variants responsible for the heritability estimates obtained from twin studies¹⁹⁶. However, across many complex traits the h^2_{SNP} estimates obtained from GWAS were substantially lower than the heritability estimated using family studies. This gap became known as the “missing heritability” problem¹⁹⁷ and was extensively investigated using height, a model trait in population genetics. Twin studies estimated the heritability of height to be 80%^{198,199}, however initial GWAS found that genome-wide significant variants had a h^2_{SNP} of 5%²⁰⁰. The causes of missing heritability have been widely debated^{201–203} and

include the role of gene-environment interactions²⁰⁴, gene-gene regulation (epistasis)²⁰⁵ and the potential overestimation of heritability in twin studies due to underlying assumptions²⁰⁶. However, the latter explanation has mostly been discarded as assumption-free methods- such as allele-by-descent analysis- have produced heritability estimates concordant to those from traditional family studies²⁰⁷.

Most notably, missing heritability reinvigorated the CDCV debate, with many arguing that the discrepancy between heritability estimates was due to the contribution of rare genetic variants with large effects to common disease, which GWAS failed to capture due to the low coverage of low frequency variants²⁰⁸. Others believed that the issue arose because GWAS lacked statistical power to detect all relevant common variants of small effect²⁰³. GWAS typically employ stringent p-value thresholds ($p < 5 \times 10^{-8}$)^{209,210} to account for the vast number of association tests (per SNP), which can lead to false negatives - particularly for variants with small effect sizes - and bias h^2_{SNP} estimates downward²¹¹. To address this, new statistical methods were developed to provide unbiased estimates h^2_{SNP} by fitting all SNPs simultaneously as random effects, rather than testing each SNP individually²¹¹. This method predicted that when modelled together, common genetic variants could account for approximately 45% of the variance human height, establishing a theoretical upper bound for h^2_{SNP} achievable in a GWAS with a sufficiently large sample²¹². This prediction was validated in a landmark GWAS of human height in 5.4 million individuals, which identified 7,209 SNPs that together explained the entirety of the estimated h^2_{SNP} in European populations ($\sim 45\%$)²¹³. This supports the notion that missing heritability was in part due to the insufficient power of early GWAS to detect all common variants with small effect. However, the discrepancy of the saturated h^2_{SNP} (45%) and heritability from family studies (80%) in height also

indicates a role of additional genetic factors which are not captured using GWAS, such as rare genetic variants.

1.3.5 Rare genetic variation

1.3.5.1 Next-generation sequencing

The allele frequency (AF) of genetic variants in a population falls along a continuous spectrum usually defined by the minor allele²¹⁴. The binary categorisation of “common” and “rare” variants is debated, but typically a threshold of $MAF > 1\%$ and $MAF < 1\%$ is used to separate the two respectively²¹⁵. In contrast to common genetic variants, rare genetic variants are usually more recent mutational events which exhibit minimal LD with neighbouring variants²¹⁶. As GWAS inherently rely on patterns of LD for accurate imputation on non-genotyped variants, the imputation and representation of rare genetic variants within GWAS is poor^{151,217}. Customised SNP arrays can be used to detect rare variants in theorised genes of interest for different diseases^{218,219}, but these are limited to discoveries based upon a priori functional hypotheses of relevant genes. Hypothesis-free analysis of rare genetic variants and their role in disease would require direct sequencing of the whole genome. Sanger sequencing, the traditional sequencing method was limited in its utility to population studies due to its limited throughput and high cost, illustrated by the sequencing of the first genome in the HGP, which took over a decade to complete and cost approximately \$3 billion¹⁴⁸. However, the HGP was key to developing next-generation sequencing (NGS) technologies which have drastically reduced the time (~ 1 day) and expense (< \$1000) to sequence a human genome²²⁰. In

summary, NGS works by breaking up DNA strands into short fragments which can then be sequenced in parallel, creating hundreds of thousands of DNA sequences, called “short-reads”. Various algorithms have been developed to then re-assemble the shortreads using a reference genome (“sequence alignment”), first provided by the HGP, and to identify differences between the genome and the reference (“variant calling”)²²⁰⁻²²².

Next-generation sequencing (NGS) techniques can be used to measure the entire genome (whole-genome sequencing; WGS) or be limited to the protein-coding regions of the genome (whole-exome sequencing; WES)²²². Both WGS and WES provide a substantially larger number of genetic variants (both common and rare) than array genotyping and imputation, ultimately improving the resolution of genetic studies²²³. The reduced cost of these technologies alongside the observation of that rare genetic variants may play a hitherto unexplored role in common complex disorders provided a strong rationale for the use of NGS in large population studies^{208,224}. Examples include the UK Biobank (UKB, N ~ 500,000), which has now released WES²²⁵ and WGS²²⁶ data on all participants, and All Of Us (AoU, N_{expected} ~ 1 million)²²⁷, which has released WGS data on ~ 240,000 participants²²⁸. These rich resources provide exciting opportunities to investigate the role of rare genetic variants in health and disease in a hypothesis-free manner.

1.3.5.2 Unique features of rare variants

Although rare genetic variants are individually rare (MAF < 1%), they are collectively common and make up most genetic variations within humans²¹⁶. Importantly, rare genetic variants have distinct features than that of common variants which can be used to gain additional biological insight into complex diseases, such as MDD²²⁹. As mentioned, rare

genetic variants typically do not display high levels of LD, making their associations easier to map underlying causal genes^{215,229}. It has also been shown that individual rare variants are more likely to have functional effects and have larger effects on complex traits which can be informative regarding disease aetiology²³⁰. Given the inherent rarity of rare variants, WGS and WES are generally underpowered to test individual associations in the same way as a GWAS. Instead, a common approach is a gene-based test whereby rare variants are aggregated together to evaluate the overall burden within individual genes^{231,232}. These methods are powerful when there is a collection of alleles in the same gene which result in a spectrum of phenotypic severity (an allelic series)^{233,234} and can be indicative of causality²¹⁵. Also, the identification of genes with allelic series are of significant therapeutic interest, as pharmaceuticals can be developed to mimic the protective alleles in the series which are linked with less severe phenotypes and subsequent clinical benefit^{233,235}. A prime example of this is the successful development of evolocumab for cardiovascular disease following the discovery of an allelic series in *PCSK9* which affect LDL cholesterol levels^{236,237}. Evolocumab, a monoclonal antibody inhibiting *PCSK9*, was shown to reduce LDL cholesterol levels and the risk of various cardiovascular events in a randomised double-blind placebo-controlled clinical trials²³⁸ and was subsequently approved by the FDA. Overall, the systematic examination of the contribution of rare genetic variants to complex traits can be beneficial for gaining biological insight beyond GWAS and can often be easier to translate into clinical impact²²⁹.

1.3.5.3 Rare variants in MDD

Many studies are beginning to demonstrate that rare genetic variants have an independent contribution to the risk on common complex diseases^{230,231,239-245}, although it is generally observed that they explain less variance than common genetic variants²⁴⁰. Nevertheless, rare genetic variants may play a larger role in psychiatric and neurodevelopmental disorders²⁴⁵⁻²⁴⁷. Ganna et al, (2018) investigated the burden of rare protein-truncating variants (PTVs) in genes intolerant to this variation across the phenotypic spectrum and found a significantly higher burden in psychiatric compared with somatic complex traits²⁴⁷. In 2018, researchers used WES to examine the relative contributions of rare and common genetic variants in MDD. Specifically, multidimensional scaling (MDS) on the Hamming Distance (a normalised metric of genetic similarity)^{248,249} revealed an increased separation of MDD cases and controls for rare variants, highlighting their significant role in MDD²⁵⁰. However, this MDS approach does not inform about potentially important biological processes which may be mediating this role. More recently, investigations using large-scale WES data have enabled more specific and thorough analyses²⁵¹⁻²⁵³, uncovering novel genes associated with MDD and demonstrating the cumulative effects of common and rare genetic variants in MDD. Cheng et al, (2022) used WES in UKB (N = 120,033) and found an increased exome-wide burden of rare PTV in loss-of-function intolerant genes across various MDD phenotypes and identified two novel risk genes for MDD, *SLC2A1* and *NOG*²⁵². Interestingly, this study observed that incorporation of rare variant burden improved genetic prediction by PRS in EHR-defined depression only, indicating that different phenotypic definitions of MDD have different contributions of rare and common variants. Another recent study used gene-based analysis on WES in UKB (N = 296,199) and identified 6 novel genes associated with depressive symptoms²⁵³.

Furthermore, two novel genes, *TRIM2* and *UBD* were associated with structural brain changes relevant to MDD alongside a series of proteins involved in cytokine inflammatory cascades²⁵². So, WES and WGS studies have begun to elucidate the role of rare genetic variants complex diseases including MDD and more biological insights can be expected as the number of datasets with NGS increase²²⁴.

1.4 Multi-omics

The term 'omics' simply refers to the comprehensive and systematic study of a set of molecules, and multi-omics refers to the study of multiple omics either in parallel or integrated analyses²⁵⁴ (Figure 1.6).

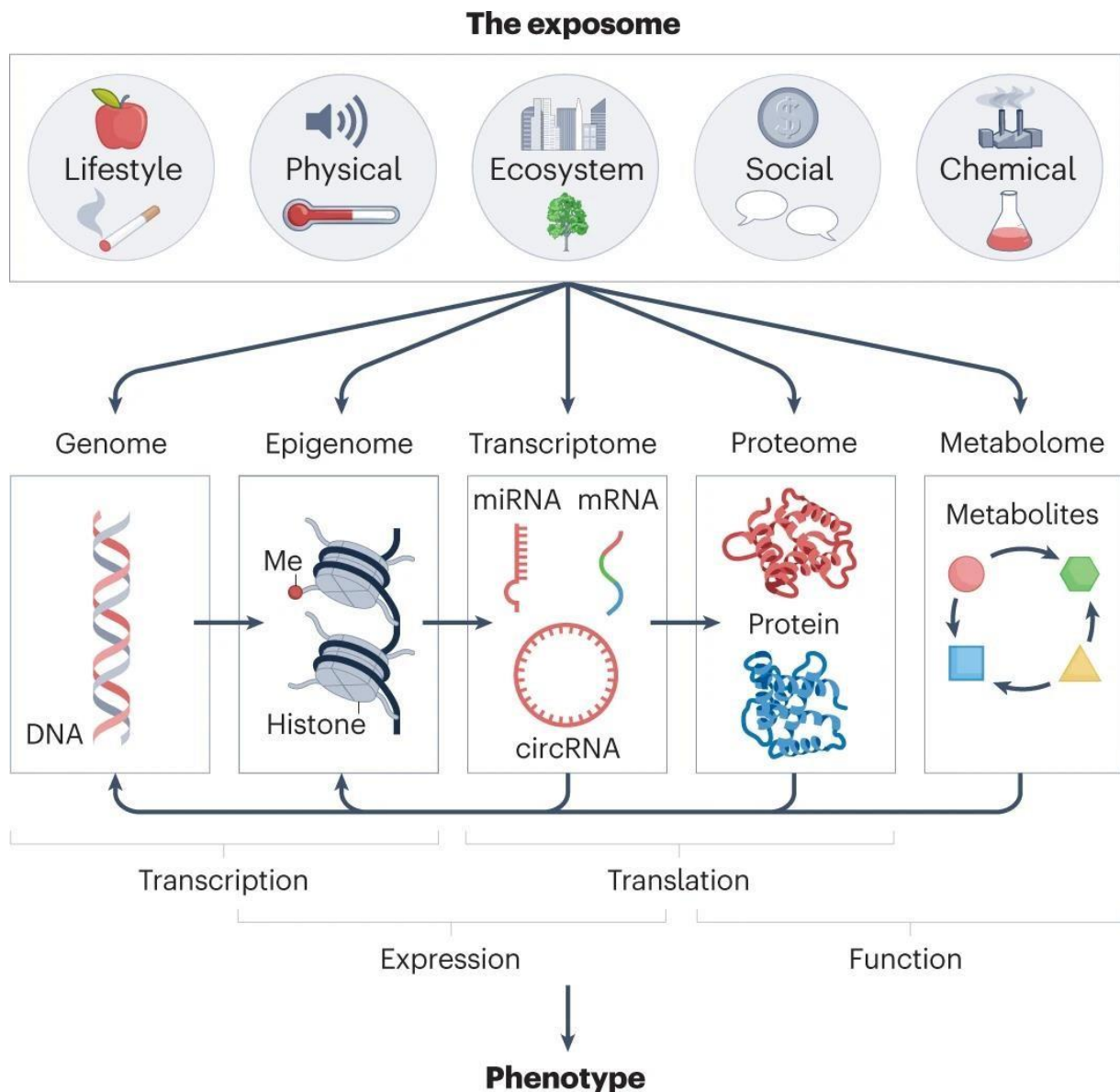


Figure 1.6 Complex interactions between environmental exposures (exposome) and multi-omic layers.

Schematic showing the downflow of biological variance from genomics, epigenomics, transcriptomics, proteomics and metabolomics, alongside the layers of regulation between them. This Figure was taken from Wu et al, 2023²⁵⁵. Copyright: *Nature Reviews Genetics*.

Genomics, the examination of genome-wide genetic variation, is the most-well studied omics layer through the extensive applications of GWAS. However, translation of genetic associations to meaningful clinical impact has been challenging, in part due to many associations in non-coding regions of the genome with unknown function^{184,188}. In 1958, Francis Crick proposed the central dogma of molecular biology, outlining a linear process by which genetic code has a functional impact on cellular function through the transcription into messenger RNA (mRNA) and subsequent translation in protein products²⁵⁶. The study of mRNA expression and protein levels, referred to as transcriptomics and proteomics respectively, in conjunction with genomics can be informative in understanding the functional impact of genetic variants within the central dogma. The Encyclopedia of DNA Elements (ENCODE) project used various multi-omics assays on immortalised human cell lines to delineate all the functional elements encoded in the human genome. This project found that whilst > 80% of the genome was transcribed, only 3% was translated into functional proteins, and indicated a strong regulatory role for non-coding DNA which had previously been discarded as “junk”²⁵⁷. This implies that whilst the central dogma is a useful concept, it is likely to be an oversimplified model capturing only part of complex, multi-faceted system. There are various other omics which lie outside the traditional central dogma outlined by Crick, such as epigenomics and metabolomics, which are also influential in regulating gene expression, biological processes and disease development^{258–262}. There are two main ways in which multi-omics can be used to gain biological insights into a disease²⁵⁴. Firstly, the genome-first approach uses multi-omics data to characterise the functional impact of genetic variants on molecular phenotypes and primarily aids in interpreting GWAS associations. Secondly, the phenotype-first approach uses multi-omics data to

directly examine associations of molecular phenotypes with a trait of interest, which may provide insights into relevant biological processes to disease that are not captured by genomics. This thesis employs both the genotype-first and phenotype-first approach integrating metabolomics and epigenomics to investigate MDD and antidepressant exposure respectively. The unique insights gained from each approach are discussed, followed by a comprehensive review of metabolomic and epigenomic analyses related to MDD and antidepressant exposure.

1.4.1 *Genotype-first approach*

Technological advances have enabled large-scale measurement of multi-omics data in conjunction with genotyping, which in turn allows for genetic drivers of molecular endophenotypes to be identified using GWAS²⁶³. Genetic loci associated with changes in a molecular phenotype are referred to as molecular quantitative trait loci (molQTL)²⁶³. The mapping of molQTLs not only provides insight into genetic regulatory mechanisms themselves but also may help to predict functional impact of disease associated genomic-loci^{188,254,263}. The most well studied molecular phenotype is transcriptomics, the comprehensive study of gene expression within a certain cell or tissue and has led to the identification of genetic loci associated with changes in gene expression known as expression quantitative trait loci (eQTLs)^{264,265}. Using blood samples from 31,684 individuals, the international consortia eQTLGen were successful in mapping blood-based eQTLs for 88% of autosomal genes²⁶⁶. Notably, eQTLs were observed both near (< 1Mb, *cis*) and far (> 1Mb, *trans*) from the associated gene, highlighting the complexity of regulation in the human genome²⁶⁶. In contrast to genomics, which is identical across all cells and tissues, multi-omic profiles differ depending on cellular and tissue context²⁶⁷⁻

²⁶⁹. Therefore, efforts have been established to identify tissue-specific expression profiles and eQTLs. A key example is the Genotype-Expression Project (GTEx) which uses post-mortem donors to establish tissue-specific expression profiles and perform eQTL mapping^{267,270}. Due to the nature of data collection²⁷¹, the number of samples (N = 838) in GTEx are an order magnitude smaller than blood-based markers such as eQTLGen. However, the resolution offered by tissue-specific approaches make it a valuable scientific resource to enrich GWAS associations and inform on potential tissue specific biological effects relevant to a disease²⁷⁰. In particular, molQTLs can be used as instrumental variables in summary-based Mendelian Randomisation (MR) analyses to assess the causal effect of a molecular phenotype on a complex trait (subject to core assumptions, detailed in Section 3.1.4.2) which would be indicative of a causal role in its aetiology¹⁹². Using brain-based eQTLs and MDD GWAS summary statistics, this approach has identified potentially causal associations with the expression of immune related genes in various brain regions and MDD²⁷².

1.4.2 *Phenotype-first approach*

Emerging evidence is showing a host of regulatory processes that can affect gene expression and protein levels independent of base genetic code^{273,274}. For example, for the same gene, alternative splicing and post-translational modifications can result in various transcripts and proteins which may each play a specific role in disease pathology^{275,276}. Additionally, environmental exposures such as pollutants, diet, and medications can regulate gene expression through epigenomic mechanisms²⁷⁷. Therefore, looking at genetics alone discards this complexity and instead acts as a reductionist snapshot of biological variance^{254,278}. The phenotype first approach instead tests the association of a phenotype (here, MDD) with multi-omics data directly, often in

an approach analogous to a GWAS, and enables the detection of relevant biological processes which may not be captured with genomics alone. In this approach, rather than predicting the functional impact of genetic associations, functional gene products are tested for an association with a disease directly²⁵⁴. This can be especially informative for omics downstream of most regulatory processes (such as proteomics and metabolomics) as these molecules directly influence cellular behaviour and represent biological layers 'closer' to the phenotype^{279,280}. The phenotype-first approach also offers promise in potentially identifying biomarkers, an active research area in MDD²⁸¹.

1.4.3 *Metabolomics*

Metabolomics is the study of small low-molecular weight compounds collectively known as metabolites, including peptides, oligonucleotides, sugars, organic acids, ketones, aldehydes, amino acids, lipids and alkaloids²⁸². Metabolites are produced in the body from endogenous chemical reactions and from the break-down of exogenous compounds from the diet and pharmacological agents^{282,283}. Importantly, metabolomics falls downstream of the central dogma and environmental regulation and therefore provides an integrated functional measure of cellular activity which is more directly linked to phenotype than other omics layers²⁸⁴. In addition to classical roles in regulating cellular homeostasis, emerging evidence is showing metabolites to be important signalling molecules for a range of functions such as immune activation, nutrient sensing and embryonic development²⁸⁵. However, metabolomics encompasses a range of very chemically diverse molecules and there is no one single instrument which can measure the full metabolome^{284,286}. In fact, the full metabolome is not yet defined and there are likely multiple metabolite compounds which are yet to be identified^{287,288}.

Generally, to gain a broad assessment of a metabolomic profile, researchers employ a suite of different methods involving liquid or gas chromatography (LC/GC) combined with mass spectrometry (MS) or ^1H -nuclear magnetic resonance (NMR) spectroscopy. Both MS and NMR methods separate metabolites by their resonance frequencies in a magnetic field or by their unique mass-to-charge ratio respectively, resulting in a spectral profile of separation which can then be used to identify and quantify metabolites in parallel^{289,290}. Each approach has its distinct advantages and disadvantages^{282,291}. In short, NMR is advantageous as it is highly reproducible, requires minimal sample preparation which can produce technical artefacts, and is nondestructive to the sample meaning it can be reanalysed using other techniques^{291,292}. Conversely, MS is much more sensitive than NMR and through different combinations of LC/GC methods can measure a much broader spectrum of metabolites. However, MS is generally less reproducible, more time consuming and requires more complex data analysis than NMR²⁸⁶. Until recently, metabolomics was not routinely applied in large-scale epidemiological studies, due to limited capacity of analytical techniques²⁷⁹. Recent advances in sample preparation alongside automated laboratory and data processing have enabled the measurement of hundreds to thousands of metabolites in a high throughput manner²⁹⁰. Thus, metabolomics is rapidly being added to the repertoire of omics data collected in large-scale population cohorts^{293–296}.

1.4.3.1 Metabolomics and MDD

The application of metabolomics has already shown substantial promise in identifying unique metabolomic profiles and biomarkers associated with cardiovascular disease^{290,297}, diabetes²⁹⁸, response to infection²⁹⁹, COVID-19³⁰⁰ and more³⁰¹.

Metabolomics was of particular interest in MDD due to the high comorbidity between MDD and metabolic disorders and traits linked to metabolic dysfunction, such as obesity³⁰². It was found metabolically “unhealthy” obese individuals (characterised by high triglycerides, low high-density lipoprotein cholesterol and elevated inflammatory markers) had 23% higher odds of depressive symptoms than obese individuals without this metabolomic signature³⁰³. A plethora of studies have been performed characterising metabolomic differences between MDD and healthy controls in human and animal models, across various metabolomic platforms, and across different biological samples (including blood, urine and cerebrospinal fluid)^{280,304–314}. As there is not one standardised methodological approach, these discrepancies and the use of different commercial technological platforms have led to inconsistent findings²⁸⁰. The Protein and Metabolite Network of Depression Database (ProMENDA) was developed and aims to provide a standardised resource for all knowledge in metabolomic and proteomic research in MDD, containing data from 1,370 studies so far^{280,304}. However, the systematic evaluation of this catalogue using a vote-counting approach found that the majority of reported metabolite-MDD associations were not replicated across different studies³¹⁵. Yet, of those that did replicate, it was found that two metabolites (glutamic acid and phosphatidylcholine) were consistently upregulated and nine metabolites (tryptophan, kynurenic acid, kynurenine, acetylcarnitine, 5-HT, creatinine, inosine, phenylalanine and valine) were consistently downregulated in those with MDD³¹⁵. Notably, most of these studies have small sample sizes (mean N = 137), which in combination with MDD as a heterogeneous phenotype may be an explanation for small rates of replication so far.

A few large, concerted efforts have been made to combine metabolomics data from multiple studies to increase power for detecting true signals and warrant mentioning. Bot

et al, (2020)³¹⁶ combined NMR metabolomics data from nine Dutch cohorts resulting in 10,145 controls and 5,283 MDD cases and found an immuno-metabolomic signature (characterised by less high-density cholesterol, and more very-low density lipoprotein and triglycerides) which was significantly associated with MDD. Subsequent analyses have found that this signature was disproportionately associated with those with atypical depressive symptoms, identifying a more biologically homogenous group of MDD patients which could be used to identify those at higher cardiometabolic risk³¹⁷. Another large study (N = 23, 972) detected significantly reduced B-hydroxylate, leucine and tryptophan levels in participants with depressive symptoms and indicated a role in prefrontal mitochondrial biogenesis in chronic stress and MDD³¹⁸. Overall, metabolomics is beginning to uncover molecular pathways associated with MDD alongside potential biomarkers of different MDD symptom profiles. During my PhD, the UKB released three tranches of metabolomic data, now in up ~275,000 participants, enabling an unprecedented opportunity to investigate metabolomic links to MDD in a large population dataset and to build upon these findings³¹⁹.

1.4.4 *Epigenomics and antidepressant exposure*

Epigenomics is the study of dynamic reversible modifications to DNA that regulate gene expression without changing underlying DNA sequence^{258,320}. Importantly, epigenetic factors are influenced by both genetic and environmental exposures and can be studied to assess the impact of environmental factors on gene expression, molecular phenotypes and disease progression³²¹. DNA methylation (DNAm) was the first epigenetic factor to be discovered and is the most widely studied³²². Specifically, DNAm involves the addition of a methyl group to the pyrimidine ring of the cytosine (“C”) base in DNA, usually at

cytosine-phosphate-guanine dinucleotide (CpG) sites, which can alter levels of gene expression³²³. Several studies have identified DNAm signatures associated with smoking³²⁴, age³²⁵, BMI³²⁶ and alcohol consumption³²⁷, which are common risk factors for a variety of diseases, including MDD. Additionally, DNAm changes are crucial for nervous system development³²⁸ and have been associated with response to traumatic early-life experience and biological trajectories leading to psychiatric disorders³²⁹. Studying the methylome has been facilitated by commercial microarrays which can measure the methylation levels at up to 850,000 CpG sites throughout the genome³³⁰. First, DNA is treated with bisulphite which converts unmethylated sites to uracil (“U”) and then applied to a microarray of oligonucleotide probes for hybridisation³³¹. For each CpG probe, the ratio of methylated (“C”) to unmethylated (“U”) DNA strands in a sample is then computed as a Beta value with 0 and 1 representing no and complete methylation respectively^{331,332}. The Beta values are often logarithmically transformed into M-values ($\log_2\text{Beta}$), which are more statistically appropriate for conventional linear regression models³³¹. The reduced cost and high throughput of this technology alongside new bioinformatics tools led to many methylome-wide association studies (MWAS), which test the association of methylation levels at each CpG probe with a phenotype individually, analogous to a GWAS³³¹. It should be noted that whilst microarrays are cheaper and more straightforward to analyse, the largest array only covers an estimated 2-4% of the epigenome^{330,333} and other sequencing-based technologies have been proposed for higher resolution³³⁴. However, epigenomic profiling has prohibitive cost compared to genotyping, meaning most MWAS are typically an order of magnitude smaller than GWAS^{254,335}.

1.4.4.1 DNA methylation and antidepressant exposure

There is growing evidence of the role of DNAm in MDD³³⁵⁻³³⁹, yet, specific associations were inconsistent, likely reflecting the small sample sizes. To gain power and biological insight, combining data from multiple studies is the way forward^{335,339}. Recently, the PGC conducted the largest-scale MWAS of lifetime MDD across 18 studies in European (N = 24,754) and East Asian (N = 2,089) populations which identified 15 CpGs significantly associated with MDD³³⁵. Furthermore, a methylation profile score (MPS) was significantly predictive of MDD status in an independent sample, indicating these findings are generalisable across cohorts³³⁵. Given the epigenetic signature of MDD, it has been hypothesised that antidepressants may in part exert their effects through epigenetic mechanisms to create a “window of plasticity” and reverse MDD-associated changes³²³. A variety of in-vitro and animal models have shown interaction between antidepressants and DNAm machinery alongside relevant cellular and behavioural changes.

Antidepressants have been shown to interact in vitro with DNA methyltransferase I (DNMT1), a key enzyme for DNA methylation³⁴⁰. Animal models have also shown that chronic exposure to paroxetine reversed stress-induced epigenetic changes in a rodent model and was accompanied by an increase in 5-HT receptor (*5-HT1A*) expression³⁴¹. Furthermore, DNMT inhibitors themselves have been shown to exhibit a dose dependent antidepressant-like effect alongside decreased DNAm and increased BDNF levels in the hippocampus³⁴².

Whilst these experimental studies provide a strong knowledge base, large-scale epidemiological studies of DNAm and antidepressant use are important for understanding their relationship on a population level. The relationship between DNA methylation and

antidepressant exposure is less well studied than that of MDD³²³. Findings have been inconclusive, although most have had small sample sizes ($N < 500$) and a prior focus on particular genes based on theories of MDD, such as *BDNF*, *MAOA*, *SLC6A4* and *IL-11*^{322,323}. A recent MWAS of self-reported antidepressant use was conducted across two cohorts ($N = 8,877$) found altered methylation in 10 CpGs with roles in the innate immune system³⁴³. Although this represents a promising start to understanding more about antidepressant mechanisms, this study did not robustly evaluate potential confounding by indication or use robust measures of antidepressant exposure from EHRs. As sample sizes continue to grow, there is an opportunity to perform larger analyses, incorporate EHR information and disentangle potential confounding effects by MDD to learn more about the associations between antidepressants and DNAm.

1.5 Aims and outline of this thesis

The overarching aim of this thesis was to comprehensively examine multi-omics data in large population datasets to provide biological insights regarding the pathophysiology of MDD and antidepressant mechanisms. Specifically, I focused on an in-depth investigation of metabolites and MDD within the UK Biobank (**Chapter 3 & 4**) and DNA methylation and antidepressant exposure in Generation Scotland (**Chapter 5**). First, **Chapter 2** will introduce and detail the datasets and population cohorts used within my analyses. Each of the analysis chapters (**Chapters 3-5**) follows a consistent structure. Each chapter begins with a broad introduction, followed by a methods section that outlines the rationale, strengths and limitations of the chosen methodologies. The main analysis is then presented, culminating in a concluding section. **Chapter 3** presents a hypothesis-free comprehensive study of the associational and potentially causal relationship between 249 circulating metabolites and MDD in the UK Biobank (N = 29,757). Following a metabolome-wide association study (MetWAS), I performed a GWAS on each metabolite (N = 88,329) to assess the genetic correlations with MDD and to identify genetic instruments for MR analyses. Potentially causal relationships were assessed with complementary MR and colocalisation techniques. **Chapter 4** aimed to investigate the association between rare genetic variants within a colocalised region in the Fatty Acid Desaturase (*FADS*) cluster identified in **Chapter 3** with MDD (N = 360,202) and omega-3 related metabolomic traits (N = 206,483) using WGS data. It was hypothesised that a significant burden of rare pathogenic variants within a gene in MDD was indicative of a causal functional role in the aetiology of MDD. Furthermore, consistent burden between metabolite traits and MDD would be indicative of a shared biological pathway. I first identified rare genetic variants likely to be pathogenic based on

several in-silico annotations and curated databases and then tested the association of these variants using different aggregate-variant-testing (AVT) methods to account for different genetic architectures. **Chapter 5** aims to assess associations of DNAm with antidepressant exposure in a large-scale MWAS in Generation Scotland. This included comparison of self-reported (N = 16,531) and EHR-based (N = 7,951) antidepressant exposure, alongside in-depth functional annotation. Furthermore, I restricted my analyses to those with MDD to directly examine the effects of confounding by indication. A MPS was trained using Generation Scotland and applied to eight external datasets, to assess the generalisability of the findings. In **Chapter 6**, I will discuss my findings and future directions for research within this field and in **Chapter 7** I will discuss some ethical considerations. Finally, in **Chapter 8** I summarise the rationale, findings, strengths and limitations of this thesis.

2 Datasets

In this chapter, I will outline the datasets used in this thesis. Firstly, I will describe the two main cohort populations used in this thesis, UKB³⁴⁴ (**Chapter 3 & Chapter 4**) and Generation Scotland (GS)³⁴⁵ (**Chapter 5**). Then I will provide a summary of the GTEx Project²⁶⁷, which was used to identify eQTLs and perform tissue enrichment analysis (**Chapter 3 & Chapter 5**). Finally, I will describe the Genome Aggregation Database (GnomAD) which was used to annotate variant allele frequencies in **Chapter 4**. Here, I describe analysis on single nucleotide variants (SNVs) which are analogous to SNPs but include rare genetic variants that occur in < 1% of the population.

2.1 UK Biobank (UKB)

The UK Biobank (UKB) is a large population cohort (N~500,000) of participants aged between 40 and 69 years old at the time of recruitment (2006-2010)³⁴⁶. Baseline assessments were undertaken at 22 different centres across the UK and involved a host of touchscreen questionnaires, face-to-face interviews, anthropometric measurements and collection of biological samples (urine, blood and saliva). Various other measures have been taken in subsets of UKB participants, including 24-hour activity monitoring (N ~ 100,000), repeated assessment measures (N ~ 20,000) and whole-body magnetic resonance imaging (MRI) (N ~ 50,000). Additionally, various online questionnaires have been sent to participants to investigate particular aspects of health, such as the Thoughts and Feelings questionnaire focusing on mental illness.

Various multi-omics data are also now available for baseline blood samples and some repeated visit blood samples, including genome-wide genotyping, WGS, WES,

metabolomics and proteomics alongside a range of clinical urine and blood biomarkers.

Overall, the deep phenotypic and multi-omics data available in this large cohort make it a rich resource for investigating the role of multi-omics in health and disease.

2.1.1 *Ethics and funding*

Every participant provided fully informed consent for their data to be used in biomedical research. The UKB has approval from the North-west Multi-centre Research Ethics Committee (MREC) (11/NW/0382) as a Research Tissue Bank (RTB) compliant with the Human Tissue Act (HTA). The ethical approval is overviewed and renewed every 5 years (next renewal due in 2026). The UKB has also established an Ethics Advisory Committee (EAC) to monitor and advise on ethical issues that may arise from research in UKB. The UKB was initially funded by the Wellcome Trust, Medical Research Council, Department of Health, Scottish Government and the Northwest Regional Development Agency. The UKB has also received funding from the British Heart Foundation, Cancer Research UK, Diabetes UK, Dementias Platform UK, National Institute of Health and Research (NIHR) and the Welsh government alongside philanthropic donations from Eric Schmidt (former CEO of Google) and Ken Griffin (CEO of Citadel). The WGS data (**Chapter 4**) was funded from government (MRC, UK Research & Innovation Industrial Strategy Challenge Fund), the Wellcome charity, and four industry partners (Amgen, AstraZeneca, GlaxoSmithKline and Johnson & Johnson).

2.1.2 *Genotyping data*

Genome-wide genotyping was conducted for 488,377 participants, assayed using two similar genotyping arrays. An initial sample of 49,950 participants were genotyped at 807,411 markers using the Applied Biosystems UK BiLEVE Axiom Array by Affymetrix.

The remainder, 438,427 participants were genotyped using the Applied Biosystems UK Biobank Axiom Array at 825,927 markers. The BiLEVE and Axiom array are closely related and share 95% of the marker content³⁴⁶. DNA was extracted from blood samples stored from baseline recruitment and genotyping was carried out in 106 sequential batches (~4,700 samples for each batch). Quality control (QC) steps performed by UKB have been discussed in depth elsewhere³⁴⁶. In brief, various statistical tests were employed to check for consistency of genotyping markers across different experimental variables, such as array and batch (removing ~ 0.97% of markers). Imputation was performed using the Haplotype Reference Consortium³⁴⁷ and the merged UK10K and 1000 Genomes phase 3 reference panels³⁴⁸, using the IMPUTE4 program (<https://jmarchini.org/software/>), resulting in 93,095,623 autosomal SNPs, short indels and large structural variants in 487,442 individuals. Various variant-level and sample-level QC were applied to this data. Variants with a MAF < 0.01, significant departure from the Hardy-Weinburg Equilibrium (HWE, $p < 5 \times 10^{-6}$) and > 2% missingness across all samples were removed. Samples were removed which were marked as failing in-house Affymetrix or UKBiLEVE QC, marked as having poor heterozygosity by UKB, had a missingness of > 2% in the genotype array and those which had a mismatch between genetically inferred and self-reported sex³⁴⁶. Following QC, there was a total of 19,435,808 autosomal variants in 487,410 participants. Participants of European ancestry (N = 457, 546) were identified based on k-means clustering on genetic principal components¹⁷⁷. Unrelated participants were identified in a two-step process as previously described¹⁷⁷. Firstly, any related participants identified by UKB were removed based on kinship coefficients (> 0.044) calculated using the KING toolset (N = 131,790). Then, using a genomic relatedness matrix (GRM), one individual from each related

group with a genetic relatedness < 0.025 with all other participants was selected to be reinstated (N = 55,745) into the sample. The analysis in this thesis is restricted to QC'd, unrelated participants of European ancestry (N = 371,437).

2.1.3 *Whole-genome sequencing data*

Whole genome sequencing (WGS) was performed for 490,549 individuals were sequenced as part of the main analysis or the pilot phase, known as the Vanguard Project (N = 49,934). Sequencing was performed at the Wellcome Sanger Institute in Cambridge, UK (N = 243,027) and the deCODE facility in Reykjavik, Iceland (N = 247,522), using Illumina NovaSeq 6000 instruments with paired end sequencing performed using S4 flowcells (v1.0 chemistry)²²⁶. Sequencing processing was done using three commensurate bioinformatics pipelines, which have been detailed elsewhere²²⁶. In brief, sequences were originally processed and joint-called with BWA-mem and GATK (Category 270), and reprocessed and joint-called using DRAGEN 3.7.8 (Category 185). All pipelines were designed to comply with principles of functional equivalence and the Broad Institute Best Practices Workflows. The data used in this thesis was from the Illumina DRAGEN Bio-IT Platform Germline Pipeline v3.7.8³⁴⁹, which is an upcoming standard across the genetics community. Sequence reads were aligned to the GRCh38 graph genome reference and SNVs and indels were called at a sample level²²⁶. The full command is described elsewhere²²⁶, and relevant BED and reference files are available to download from Illumina (<https://developer.illumina.com/dragen/dragen-popgen>). DRAGEN Machine Learning Recalibration was performed on each sample to refine variant and genotype quality, filtering out samples with a QUAL < 3.0 . Aggregated VCF files were created using the DRAGEN Iterative gVCF Genotyper to aggregate samples

per batch of 1000 and perform genotyping across all 490,451 individuals across all autosomes, sex chromosomes, mitochondria, and 3341 ALT contigs of GRCh38.

2.1.4 *Metabolomic data*

Metabolic biomarker profiling of EDTA plasma samples from all participants, released in three randomised tranches, was performed using ^1H -NMR by Nightingale Health Inc, Finland. Throughout this PhD, Phase I (N ~120,000) and Phase II (N ~ 275,000) were released, used in **Chapter 3** and **Chapter 4** respectively. Specifically, 249 metabolites were measured (168 in absolute concentrations and 81 in derived ratios), including detailed measures of cholesterol metabolism, fatty acid compositions, amino acids, ketone bodies, glycolysis-related metabolites and two inflammatory protein measures (albumin and glycoprotein acetyls)³⁵⁰. Most biomarkers are measured in absolute concentrations (mmol/L), with the exceptions of apolipoproteins (g/L), diameter measures (nm) and the degree of unsaturation (total number of pi bonds and rings within a molecule). Details of the Nightingale Health profiling platform and experimentation have been described in depth elsewhere and are known for high repeatability and an absence of batch effects^{295,350,351}. In brief, a minimum of 90ul of plasma was plated onto 96 well plates and shipped in dry ice to Nightingale Health Laboratories in Finland in batches of 5,000-20,000 samples. Samples were thawed, centrifuged and mixed in a 1:1 ratio with phosphate buffer. The samples were profiled using nine 500MHz spectrometers (Bruker ADVANCE IIHD), using two NMR spectra and absolute concentrations were quantified using proprietary software (Nightingale Health, quantification library 2020). Integrated quality procedures at Nightingale Health verify the sample quality using signs of degradation or contamination issues. To assess variability between spectrometers, pre specified coefficient of variation (CV) targets were used for

two control samples (Nightingale Health) in each 96-well plate, alongside blind duplicate samples provided by UKB. Targets were met for each consecutive batch, and for most metabolites there was a CV < 5%. However, as minor differences in biomarker concentrations may arise from different spectrometers, I included spectrometer as a covariate for analysis of metabolic data, as recommended³⁵⁰.

2.1.5 *MDD phenotypes*

I used two different MDD phenotypes from the UKB within this thesis. Initially, I focused on clinical measures of lifetime MDD captured through the clinical international diagnostic interview- short form (CIDI-SF) to assess metabolomic links to MDD in **Chapter 3**. In **Chapter 4**, I used an agglomerative measure of lifetime MDD used in the most recent PGC GWAS of MDD¹⁸³. This measure ascertains lifetime MDD from various sources, including self-reported MDD, CIDI-SF, GP diagnoses, antidepressant prescriptions and inpatient hospital records. This was primarily to boost sample size in order to gain power required for robust rare-variant analyses.

2.1.5.1 Clinical International Diagnostic Interview Short Form (CIDI-SF)

The Clinical International Diagnostic Interview (CIDI) is a World Health Organisation (WHO) instrument for mental health surveys, assessing different psychiatric disorders according to the DSM-IV³⁵². The CIDI-Short Form (CIDI-SF) is an adapted version of this questionnaire which can be entirely self-administered in a fraction of the time whilst still showing reliability, consistency and validity in evaluating psychiatric disorders^{353,354}. The CIDI-SF uses a stem-branch logic where initial screening questions are used alongside skip rules to ensure only relevant detailed questions are asked of each participant³⁵³.

This was the base of the online questionnaire sent out to participants by the UKB, titled “Thoughts and Feelings”. The CIDI-SF used was modified to include lifetime history, as done in other genetic studies in the PGC³⁵⁵. This questionnaire was sent by initial email invitation to 339,229 participants, of whom 157,366 completed the questionnaire by July 2017. Participants were classed as cases for lifetime MDD if they reported at least one core symptom for most of the day on most or all of the days within a two-week period, combined with at least four secondary depressive symptoms with some/ a lot of impairment³⁵⁵. If participants did not respond to initial core symptoms questions (Table 2.1 Assessment of lifetime MDD in the UKB from the CIDI-SF.) then they were classed as missing (NA). After excluding participants which had missing data, there were 126,077 participants with case/control status for lifetime MDD using the CIDI-SF³⁵⁵.

Table 2.1 Assessment of lifetime MDD in the UKB from the CIDI-SF.

The verbatim questions, data fields and responses required to contribute to a lifetime MDD status are shown for the questions in the CIDI-SF relevant to lifetime MDD. To meet criteria for lifetime MDD, participants had to report at least one core symptom present for most/all day for most/all of two-week period, combined with at least four additional secondary depressive symptoms. Participants who reported “No” to both core symptoms were classed as controls.

Core symptoms	Question	Data Field	Lifetime MDD
Persistent sadness	Have you ever had a time in your life when you felt sad, blue, depressed for two weeks or more in a row?	20446	“Yes”
Loss of interest	Have you ever had a time in your life lasting two weeks or more when you lost interest in most things like hobbies, work, or activities that usually give you pleasure?	20441	“Yes”
Core symptoms frequency & impairment			
Fraction of day affected	How much of the day did these feelings usually last?	20436	“Most of the day” or “All day long”
Frequency of depressed days	Did you feel this way: (day options)	20439	“Every day” or “Almost every day”
Impairment	How much did these problems interfere with your life or activities?	20440	“Somewhat” or “A lot”
Secondary depressive symptoms			
Tired or low energy	Did you feel more tired out or low on energy than is usual for you?	20449	“Yes”

Gain or loss of weight	Did you gain or lose weight without trying, or did you	20536	“Gained weight” or “Lost weight”
	stay about the same weight?		or “Both gained and lost some weight during the episode”
Sleep change	Did your sleep change?	20532	“Yes”
Trouble concentrating	Did you have a lot more trouble concentrating that usual?	20435	“Yes”
Feeling worthless	People sometimes feel down on themselves, no good, worthless. Did you feel this way?	20450	“Yes”
Thinking about death	Did you think a lot about death – either your own, someone else’s or death in general?	20437	“Yes”

2.1.5.2 Combined measure of lifetime MDD

The CIDI-SF is only available for ~ 1/4 of participants in the UKB, and the integration of primary care data such as GP diagnoses, prescriptions and hospital records can inform on a potential lifetime MDD and boost sample sizes for analyses. As there is no single centralised system for collecting and sharing primary care data in the UKB, linkage to EHRs has been an arduous task involving many different data suppliers. Nonetheless, as of 2019, the UKB have made available primary care records for 45% of UKB cohort, all of whom provided consent for EHR linkage. Recently, the GWAS of MDD conducted by the PGC adopted an aggregate measure of MDD based upon both questionnaire and primary care data¹⁸³. Participants were classed as cases for lifetime MDD if they met any two of the following criteria: self-reported or CIDI-SF determined lifetime MDD, self-

reported use of antidepressant medications, one to two GP diagnoses of MDD, three or more episodes of GP antidepressant prescribing. Participants were also classed as cases if they met either of the following criteria: three or more GP diagnoses of MDD or an inpatient diagnosis of MDD. Controls were ascertained as those who did not self-report lifetime MDD or any antidepressant treatment or had two or fewer GP antidepressant prescriptions alongside no GP or hospital inpatient records of a mood, psychotic or personality disorder.

2.1.6 Sociodemographic factors

The UKB has measured an in-depth range of sociodemographic factors of participants which were collected at baseline assessment at the same time as biological samples were taken. The primary factors I have adjusted for in my analyses include basic objective demographic variables such as age (Data-Field: 21003), biological sex (Data-Field: 31) and BMI (Data-Field: 21001). Social factors include smoking status (Data-Field: 20116), socioeconomic status (Data-Field: 189), educational qualifications (Data-Field: 6138) and ethnic background (Data-Field: 21000). Smoking status was measured as a categorical variable for current/previous/never smokers. Socioeconomic status was measured using Townsend Deprivation Index (TDI), a continuous measure based on four broad indicators of deprivation: unemployment, non-home ownership, non-car ownership and household overcrowding³⁵⁶. Educational qualifications were collapsed into a binary variable as to whether participants had a university/college degree or not. Self-reported ethnic background was collapsed into a categorical variable or White/Mixed/Other.

2.2 Generation Scotland

Generation Scotland (GS) is a family-based population study (N ~ 24, 000) comprising of volunteers across Scotland aged between 18-98 years, with baseline data collection between 2006-2011. Potential participants (“probands”) were randomly identified and invited to participate using patient lists from collaborating general medical practices (Glasgow and Tayside, Ayrshire, Arran and Northeast Scotland). Following a pre-clinical questionnaire (PCQ), a total of 21,476 participants attended an extensive in-person research clinic where in-depth questionnaires were administered, alongside various physical measurements and biological sampling. A further 2,484 participants who could not attend in-person assessment completed the PCQ and sent a saliva sample for DNA extraction. In total, there were 7,953 probands recruited alongside 16,007 family members. The sample was 59% female, with a wide range of ages and sociodemographic backgrounds, although participants were broadly healthier and wealthier than the general Scottish population³⁴⁵.

2.2.1 *Ethics and funding*

All components of GS received ethical approval from the NHS Tayside Committee of Medical Research Ethics (05/S1401/89) and has also received Research Tissue Bank status by the East of Scotland Research Ethics Service (20-ES-0021). All participants provided informed consent for the use of their health data for biomedical research. GS is supported by the Chief Scientist Office of the Scottish Government Health Directorates (CZD/16/6) and the Scottish Funding Council (HR03006).

2.2.2 DNA methylation

DNA methylation (DNAm) was profiled from baseline blood samples using the Illumina Methylation EPIC array³³⁰, for 19,390 individuals, in 4 distinct sets between 2017 and 2021. DNAm profiling took place at the Edinburgh Clinical Research Facility, Edinburgh, Scotland, by the Genetics Core Laboratory. Sets 1, 2 and 4 contained related individuals within and between each set, whilst set 3 contained individuals unrelated to each other and unrelated to those in set 1 (relatedness defined as GRM threshold < 0.05). Quality control (QC) and normalisation steps were applied separately for each set using the R packages *meffil*³⁵⁷, *wateRmelon*³⁵⁸, and *shinyMethyl*³⁵⁹ and have been described in depth elsewhere. Broadly, CpG-level QC included removing CpG probes with low bead count (< 3) or a high detection p-value (> 0.05) in 5% of samples. Furthermore, nonspecific, cross-hybridising, and SNP associated probes (with MAF $> 5\%$) were removed^{360,361}. Sample level QC involved removing samples with a mismatch between predicted and recorded sex, those with $> 1\%$ of CpGs with a high detection p-value or samples with outlying median methylated signal intensity. For each set, the proportion of white-blood-cells was predicted using the Houseman method, implemented in the *minfi*³⁶² R package, using the reference data from Reinus et al, (2012)³⁶³. Following QC, the sets were combined (N = 18,869) and dasen normalisation was performed across all individuals³⁵⁸. Beta values were transformed to M-values using the 'beta2M' function in the *lumi* R package³⁶⁴ and were standardized using the "std-probe" flag in the OSCA software³⁶⁵.

2.2.3 *Prescription data*

Consent was obtained for 23,063 participants for their data in GS to be linked to the National Health Service (NHS) through their Community Health Index (CHI) number, routinely used in NHS Scotland. Detailed procedures for record linkage in GS have been described elsewhere³⁶⁶. Prescription data was obtained from the Scottish National Prescribing Information System (PIS), which has individual prescribing and dispensing events in a community setting since 2009. Detailed information about the PIS infrastructure and processing has been detailed elsewhere³⁶⁷. For these analyses, I used a PIS data release between April 2009 and January 2017, consisting of over 2 million prescriptions for 21,318 individuals in GS. Each prescription data-entry had extensive information including the medication name, formulation, quantity, strength, British National Formulary (BNF) codes, dispense data, paid date and free-text instructions from the prescriber which usually denote the required dosage (e.g “1 tablet a day”).

2.2.4 *MDD phenotype*

Participants were first screened for lifetime MDD using two questions (“Have you ever seen anybody for emotional or psychiatric problems?” and “Was there ever a time when you, or someone else, thought you should see someone because of the way you were feeling or acting?”). Participants who responded positively to either question were asked to complete the Structured Clinical Interview for DSM-IV (SCID)³⁶⁸, a clinician administered, semi-structured interview designed for reliable psychiatric diagnoses according to the DSM-IV. Lifetime MDD was defined as the presence of either depressed mood or anhedonia, with a minimum of four further symptoms with clinical significance (symptoms lasting almost all day/all day for most/all days) within a two-week period³⁶⁹. Participants

who answered negatively and participants with no MDD ascertained from SCID were classed as controls. Participants who were screened to take part in the SCID but did not do so were set to missing (N = 1,169). In total, there were 2,639 cases and 16,160 controls of lifetime MDD.

2.2.5 *Antidepressant exposure phenotypes*

There are different classification systems used for medications worldwide. For this thesis, I defined antidepressant medications using the BNF, specifically Chapter “0403” (Table 2.2). This is roughly equivalent to the subclass “NO6A” in the Anatomical Therapeutic Classification (ATC) subclass. One medication, flupentixol, was excluded due to its primary indication as an anti-psychotic in the literature and according to ATC (subclass NO5A). Self-reported and prescription-derived antidepressant exposure were obtained from questionnaire and prescription data respectively, described in detail below.

Table 2.2 Antidepressant medications as defined by the British National Formulary (BNF).

BNF Chapter codes, Names and Drugs for each sub-class. All antidepressant drugs, except for flupentixol (bold), are also present in the anatomical therapeutic classification (ATC) antidepressant subclass (“N06A”).

BNF Chapter	BNF Name	Drugs
4.3.1	Tricyclic and related antidepressant drugs	Amitriptyline, Amoxapine, Clomipramine, Dolsulepin, Doxepin, Imipramine, Lofepramine, Maprotiline, Mianserin, Nortriptyline, Trazodone, Trimipramine
4.3.2	Monoamine oxidase inhibitors (MAOIs)	Isocarboxazid, Moclobemide, Phenelzine, Tranylcypromine
4.3.3	Selective serotonin reuptake inhibitors	Citalopram, Duloxetine, Escitalopram, Fluoxetine, Fluvoxamine, Paroxetine, Sertraline
4.3.4	Other antidepressant drugs	Agomelatine, Duloxetine, Flupentixol , Mirtazapine, Nefazodone, Oxitriptan, Reboxetine, Tryptophan, Venlafaxine, Vortioxetine

2.2.5.1 Self-reported

In GS, two different questionnaires were used to measure medication use throughout recruitment. The first (N = 9,924) asked participants to free-text medications they were currently taking, allowing specific antidepressants to be named and detected. The second

questionnaire (N = 13,675) specifically asked “Are you regularly taking any of the following medications?” with an “Antidepressants” option and a “Yes/No” checkbox. Self-reported antidepressant exposure was therefore a combined measure of participants who named an antidepressant medication in the first version, and individuals who checked the “Yes” on the antidepressants tick box (N = 1,508). Unexposed participants were classed as those who did not name any antidepressants or checked “No” on the antidepressants box (N = 15,023).

2.2.5.2 Prescription-derived

Prescription-derived antidepressant exposure at the blood draw was estimated by mapping theoretical periods of treatment (calculated using consistent prescription events) to the research visit where blood was taken (from which DNAm was measured). First, antidepressant prescriptions were selected from PIS database using BNF code (“0403”), resulting in 174,454 prescriptions from 7,544 individuals. Most prescriptions were for antidepressants in tablet form (N = 173,818), with some liquid formulations (N = 636). Then, free-text prescribing instructions were manually parsed using a series of rules and regex expressions to ascertain the appropriate daily dosage per prescription (e.g. “1 tablet a day” ~ 1, and “2 mane” ~ 2). For ambiguous prescriptions (e.g. “as directed/advised”) (N = 9,216), I imputed dosages to be the same as the most frequent informative prescription given within the last 12 months (for the same medication) to the same individual. In the absence of such data, tablet prescriptions were imputed to a default of 1 tablet a day, whilst liquid prescriptions were discarded. To avoid complications with titrating dosages, prescriptions with changing dosages over the prescription (e.g. “take 2 a day for a week, then 1, then stop”) were discarded (N = 5,478), leaving 168,970

prescriptions from 7,373 individuals. For instances where > 1 antidepressant prescription was recorded on the same day per participant ($N = 2,604$), continuous treatment collected ahead of time or simultaneous treatment was assumed based on whether the medications differed and if the combined dosage was over/under the maximum recommended dosage using a decision tree strategy (Section 5.3 - Supplementary material).

A theoretical prescription duration was estimated as the fraction of total dosage divided by instructed daily dosage, extended by 10% to account for non-perfect adherence.

Next a theoretical end date for each prescription was calculated by adding the estimated duration onto the dispense date. Using these metrics, all antidepressant medications per individual were assigned to a treatment period. The first prescription is assigned to treatment period 1. Each following prescription is assigned to the same or new treatment period if the dispense date is before/after the predicted end date of the previous prescription respectively. Each treatment period is defined from the dispense date of the first prescription and the theoretical end date of the last prescription within the period (Figure 2.1). Participants who are within an active period of treatment at the time of their in-clinic assessment and blood sample are classed as antidepressant exposed ($N = 861$). Participants who have no evidence of antidepressant medications on record and were recruited after GS linkage (April 2009) were classed as unexposed to antidepressants ($N = 7,090$).

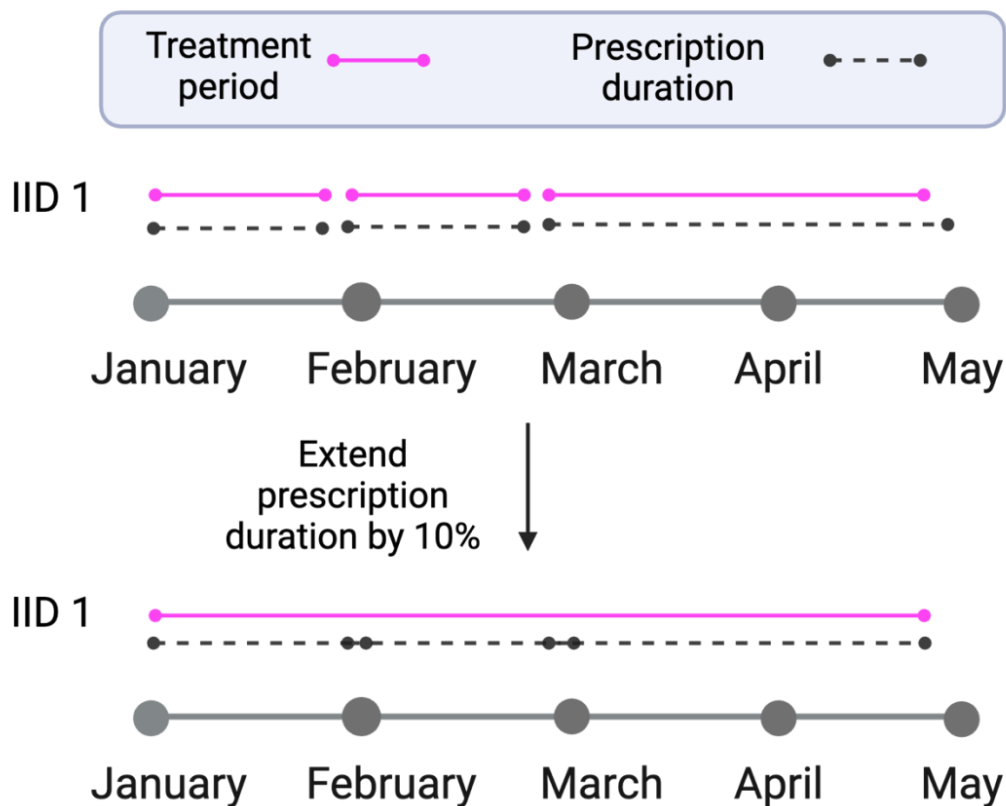


Figure 2.1 Estimation of treatment periods from prescription records.

Each prescription record had a theoretical duration (black dotted line) annotated to it from the dosage and dosage instructions. Consecutive prescriptions were combined into a treatment period (pink line) if they were dispensed around the time the previous prescription was due to end (with a 10% leeway).

2.3 Genotype Tissue Expression (GTEx)

The GTEx project, funded by the U.S. National Institute of Health (NIH), aims to determine how genetics influences gene expression in across a wide range of different tissues and how this relationship is altered in the development of disease²⁶⁷. The project aimed to collected data from ~ 50 tissues in up to ~ 1000 post-mortem donors with tissue collection

occurring < 24 hours of death. The details on data procurement, consent procedures and eligibility criteria have been described in depth elsewhere²⁷¹.

Eligible donors were aged between 21 and 70 years, with a BMI between 18.5 and 35. Potential donors who had recently received a blood transfusion (< 48 hours), chemotherapy or radiotherapy (< 2 years), been diagnosed with metastatic cancer, or had any history of intravenous drug abuse or exposure to HIV/AIDs were not included in the study²⁷¹. Informed consent was obtained from next-of-kin consent for the collection and banking of de-identified tissue samples²⁷¹. The data used in this thesis was version 8³⁷⁰, which encompasses 17,382 RNA-sequencing samples from 52 tissues of 838 postmortem donors (66% Male and 85% European American). Whole-genome sequencing data was used to identify *cis*-eQTLs and *trans*-eQTLs, defined as variants within 1Mb of the genes transcription start site (TSS) or variants on a different chromosome respectively. Detailed methodology surrounding sequencing, quality control, and eQTL mapping has been described elsewhere³⁷⁰⁻³⁷². Identified eQTLs showed a high replication rate in other datasets with RNA expression data, including Twins UK (subcutaneous adipose tissue, EBV-transformed lymphocytes, skin and whole-blood) and eQTLGen (whole-blood). Specifically in this thesis, GTEx data was used in **Chapter 3** to identify eQTLs for genes (*FADS1*, *FADS2*, *FADS3*, *TMEM258*, *MYRF*, *FEN1*) located within a colocalised region (PP.H4 > 0.8) between docosahexaenoic acid (DHA) and MDD. The eQTLs were then used in subsequent MR and colocalisation analyses to assess potential causal relationships between gene expression and MDD. Additionally, GTEx was used in **Chapter 5** to test for tissue enrichment of genes annotated to top CpGs associated with antidepressant exposure, using a hypergeometric test.

2.4 Genome Aggregation Database

The genome aggregation database (gnomAD), evolved from the Exome Aggregation Consortium (ExAC), is a reference population database consisting of exome and genome sequences from population cohorts worldwide³⁷³. All raw data from each contributing cohort has been reprocessed through gnomAD pipelines and jointly variant called, resulting in the largest harmonised sequencing dataset in the world. This publicly available resource (<https://gnomad.broadinstitute.org/>) enables comprehensive estimates of allele frequencies for genetic variants in the human genome across various different ancestries³⁷⁴. The most recent release of gnomAD (v4) consists of 730,947 exomes (including UKB samples) and 76,215 genomes from > 100 studies around the world (Figure 2.2). In **Chapter 5**, I used the allele frequencies from the gnomAD v4 genomes as an additional filter for identifying rare ($AF < 0.01$) genetic variants in each gene in the *FADS* cluster. Specifically, rare variants in UKB had to also be rare or absent in gnomAD to be taken forward for variant annotation.

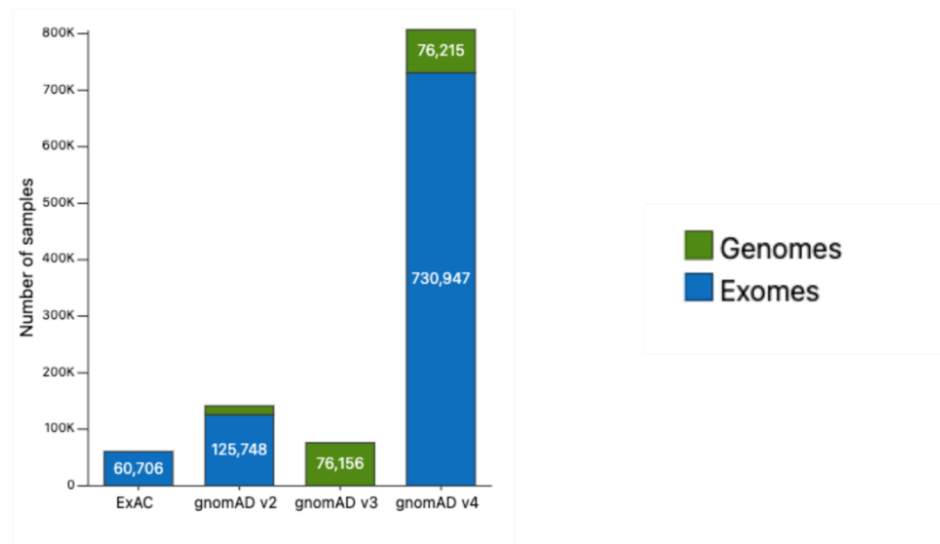
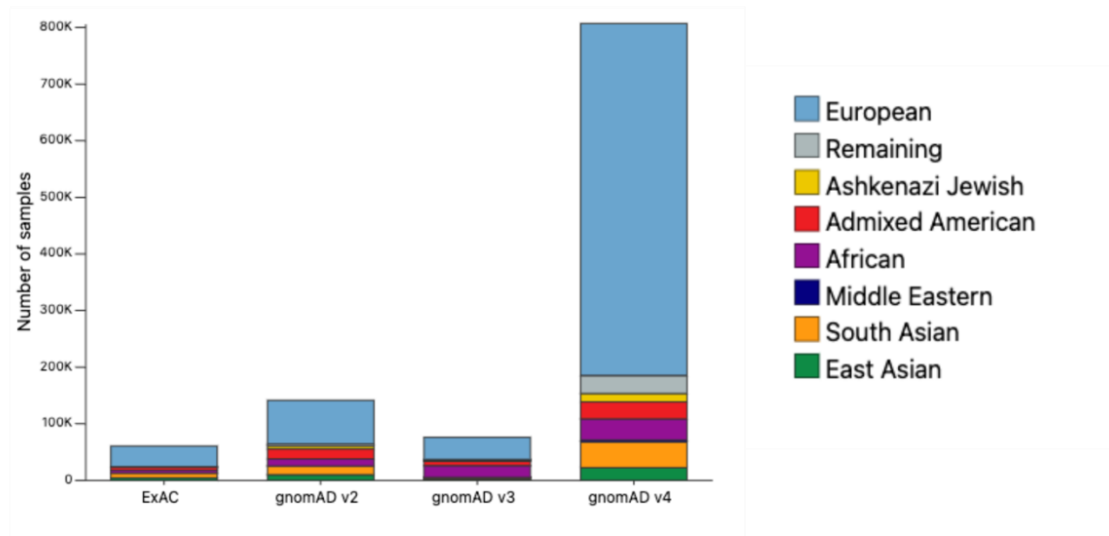
A**B**

Figure 2.2 The Genome Aggregation Database (GnomAD).

A) The sample sizes for genomes (green) and exomes (blue) in main release of the dataset. B) The per genetic ancestry group count of samples in gnomAD releases (across exomes and genomes). Both subplots of this Figure were taken from the gnomAD browser (<https://gnomad.broadinstitute.org/>).

3 Circulating metabolites and Major Depressive Disorder

In this chapter, I focus on investigating the links between circulating metabolites and lifetime MDD in the UK Biobank. This analysis begins looking at associations between metabolites and lifetime MDD in a metabolome-wide association study (MetWAS), and the impact of including common confounders into the model (e.g. BMI). I then performed multiple GWAS, to enable assessment of the SNV-heritability of each metabolite and the genetic correlation with MDD. However, as associational studies are limited in their interpretation, I also investigated the potential causal links between the two using complementary Mendelian Randomisation (MR) and colocalisation. Each of these methodologies is described in detail in Section 3.1.

This study was published in *Biological Psychiatry* in 2023 and is included in full in Section 3.2. My role was to conduct all analyses and drafting the manuscript. Code for this work is available at the GitHub repository:

https://github.com/Elladavyson/Metabolomics_MDD. The summary statistics for each metabolite GWAS were also uploaded to the GWAS catalogue (<https://www.ebi.ac.uk/gwas/publications/36764567>). The supplementary information, figures and tables associated with this publication are available online ([https://www.biologicalpsychiatryjournal.com/article/S0006-3223\(23\)00055-0/fulltext](https://www.biologicalpsychiatryjournal.com/article/S0006-3223(23)00055-0/fulltext)).

3.1 Methods

3.1.1 *Metabolome-wide association study*

Prior to all analyses, each metabolite marker was transformed to the standard normal distribution using rank-based inverse normalisation. The metabolome-wide association study (MetWAS) tested the association of each metabolite (Y) with lifetime MDD (X) using a linear ordinary-least-squares (OLS) regression model ($n = 29,757$), implemented with the “*lm()*” function in the “*stats*” R package (Equation 3.1). This model assumes a linear relationship between metabolite and MDD status, normal distribution and equal variance (homoscedasticity) of residuals (ε) and no multi-collinearity between lifetime MDD and covariates (W).

$$Y \sim W\alpha + Xb + \varepsilon \quad \text{Equation 3.1}$$

An initial “basic” model was first implemented, which covaried for age, sex, assessment centre and NMR spectrometer only. Then, an “adjusted” model was used which covaried for common confounds between metabolite levels and lifetime MDD, such as socioeconomic status (using TDI), smoking status, ethnicity, educational attainment and BMI. For each model, the false-discovery rate (FDR) was corrected for using the Benjamini-Hochberg (BH) correction, and significance was defined as associations with an FDR-corrected p value ($P_{\text{FDR}} < 0.05$).

3.1.2 *GWAS*

A GWAS was performed on each metabolite measure in unrelated participants of

European ancestry (N = 88, 329), using PLINK 2.0. Prior to analysis, single nucleotide variants (SNVs) and small indels with a MAF < 0.001, a minor allele count (MAC) < 20 and an INFO score < 0.1 were removed. The GWAS procedure is also based on a linear regression model, which is shown in Equation 3.2.

$$Y = W\alpha + X_s\beta_s + \varepsilon \quad \text{Equation 3.2}$$

Here, Y represents a vector of phenotype (metabolite) measures, W is a matrix of covariates (age, sex, assessment centre, genotyping array and top ten genetic principal components), α is their corresponding effect sizes, X_s is the vector of genotype values at SNV s , β_s is the corresponding fixed effect size, and ε is the random effect of residual errors. The genotypes at each SNV were coded as 0,1,2 depending on the number of copies of the alternate allele, in an additive genetic model³⁷⁵. To adjust for multiple corrections (249 GWAS in total), significance was defined as the genome-wide threshold (5×10^{-8}) divided by the number of principal components (PCs) which account for 99% of the variance within the normalised metabolomic data. Principal component analysis (PCA), conducted using the “*prcomp()*” function from the “*stats*” R package, found 42 PCs accounted for 99% of the variance, making the threshold for significance in the GWAS 1×10^{-9} ($5 \times 10^{-8}/42$). For significantly associated SNVs, an F statistic was calculated as a measure of associational strength (Equation 3.3).

$$F = \left(\frac{(N - k - 1)}{k} \right) \times \left(\frac{r^2}{(1 - r^2)} \right) \quad \text{Equation 3.3}$$

Here, N = sample size, k = number of SNVs and r^2 = the variance explained in metabolite levels by the SNV, calculated as a function of MAF and SNV effect size (β) (Equation 3.4).

$$r^2 = 2 \times \text{MAF} \times (1 - \text{MAF}) \times \beta^2 \quad \text{Equation 3.4}$$

3.1.3 *LD Score regression*

LD-score regression (LDSC) and its extensions are powerful tools which can estimate heritability and genetic correlation of polygenic traits from GWAS summary statistics alone^{376,377}. The SNV-heritability (h^2_{SNV}) of each metabolite was calculated using LDSC, which provides estimates of heritability and potential confounding biases, such as cryptic relatedness or population stratification³⁷⁶. This technique was first developed as a tool to partition genomic inflation in a GWAS due to true polygenic effects or confounding effects and relies upon an intuitive relationship between LD and GWAS effect sizes. Due to the nature of genotyping arrays, GWAS effect size at a given SNV incorporates the effects of all other variants it tags by virtue of LD. Therefore, SNV markers in large block of LD will have elevated GWAS effect estimates as they are more likely to tag a causal variant (by tagging more variants in total). Bulik-Sullivan et al, (2015)³⁷⁶ showed that for true polygenic signal, there will be a genome-wide positive correlation between SNV LD-score (a measure of LD) and GWAS effect estimates (as χ^2 statistics). LD-scores at each GWAS SNV j (ℓ_j) can be estimated using a matching reference panel as the sum of the squared correlation (r^2) between variant j and all other nearby M variants (within 1 centimorgan). Under the polygenic model, the expected χ^2 of variant j is shown in Equation 3.5, where N is the GWAS sample size, M is the number of SNVs and α is the

contribution of confounding biases. Fitting a regression between the GWAS effect estimates and the LD-score of SNVs can therefore detect inflation due to confounding (by the intercept, $N_\alpha + 1$) alongside estimation of h^2_{SNV} from polygenic signal (by the slope).

$$E[\chi^2 | \ell_j] = 1 + N_\alpha + \frac{Nh^2\ell_j}{M} \quad \text{Equation 3.5}$$

Genetic correlation between each metabolite trait and MDD was also calculated using LDSC³⁷⁷. I used the GWAS summary statistics generated for each metabolite, and the MDD summary statistics from the MDD GWAS from Howard et al, (2019)¹⁷⁷. The calculation of genetic correlation using LDSC relies on the same principles as heritability estimates, simply extended to two traits. In this model, the LD score is instead regressed upon the product of z scores from GWAS of separate traits (z_1z_2). The expected value for z_1z_2 for SNV j under a polygenic model is shown in **Equation 3.6**, where N_1 and N_2 are the sample sizes of each trait GWAS, and ρ_g is the genetic covariance between them.

$$E[z_{1j}z_{2j}\ell_j] = \frac{\sqrt{N_1N_2} \rho_g}{M} \ell_j + \frac{\rho_g N_s}{\sqrt{N_1N_2}} \quad \text{Equation 3.6}$$

Genetic covariance can be estimated by the slope of the regression line between z_1z_2 and LD score. Estimates of genetic correlation (r_g), ranging from -1 (complete negative correlation) and +1 (complete positive correlation) are then obtained by normalising the genetic covariance by the estimated h^2_{SNV} of each trait (h_1h_2) (Equation 3.7).

$$r_g := \frac{\varrho_g}{\sqrt{h_1 h_2}}$$

Equation 3.7

3.1.4 *Mendelian Randomisation*

In this thesis, I used bidirectional two-sample MR to estimate the causal effects of each metabolite on MDD and vice versa. The principles of MR and its assumptions are discussed. Then, the specific methods I used (inverse-variance weighted, weighted median and MR Egger) are described.

3.1.4.1 Principles

Statistical association (identified through MetWAS and LDSC analyses) alone is not enough to infer a causal relationship. Observed associations between metabolites and MDD could be due to unobserved confounders (e.g. unmeasured lifestyle factors which influence both MDD and metabolite levels separately) or reverse causation (i.e. MDD or processes leading to MDD have an influence on metabolite levels). Randomised clinical trials (RCT) are the gold-standard for assessing causality between exposures and outcomes. An RCT randomly assigns participants to different exposure arms (e.g. treatment vs placebo), resulting in any known and unknown confounders to be evenly distributed, so any difference in outcomes observed between groups can be directly attributed to the causal effects of the exposure. Mendelian Randomisation (MR) is a causal inference method which utilises the random segregation of alleles at conception to estimate causal effects of modifiable exposures (e.g. metabolite levels) on an outcome (e.g. MDD) analogous to an RCT (Figure 3.1)^{378–380}. The underlying principles of MR rely on Mendel's laws of

segregation and independent assortment, whereby genetic alleles are randomly inherited from each parent at conception, similar to the randomisation of participants in a RCT to different treatment arms³⁸¹. Therefore, a genetic variant robustly associated with an exposure can be used as an instrumental variable (IV) to identify randomised populations with different average levels of exposure, which can then be used to estimate the causal effect of the exposure on an outcome³⁷⁹. The effect estimates from MR are less prone to confounding due to the randomisation of alleles and reverse causation, as genetic variation is fixed from conception.

3.1.4.2 Assumptions

Importantly, the reliability of unbiased MR estimates depends on three core assumptions (Figure 3.1)³⁷⁸. The relevance assumption assumes that the genetic variant used as the IV is robustly associated with the exposure. This can be explicitly tested by GWAS, followed by measures of associational strength by both the r^2 statistic (Equation 3.4) and F -statistic (Equation 3.3). The independence assumption assumes that the genetic variant used as the IV is not associated with any known or unknown confounders. Possible violations of this assumption can be tested by systematically assessing the association of the IV with known confounders, however the association with possible unknown confounders cannot be explicitly tested. The exclusion-restriction assumption assumes that the IV is associated with the outcome only through its association with the exposure (vertical pleiotropy). Horizontal pleiotropy, whereby a genetic variant influences multiple traits through different independent pathways, is a direct violation of the exclusion-restriction criteria. This assumption is not possible to prove empirically, but

various sensitivity analyses can be applied to identify potential violations and provide unbiased causal estimates when this assumption is violated (e.g. MR-Egger)³⁸².

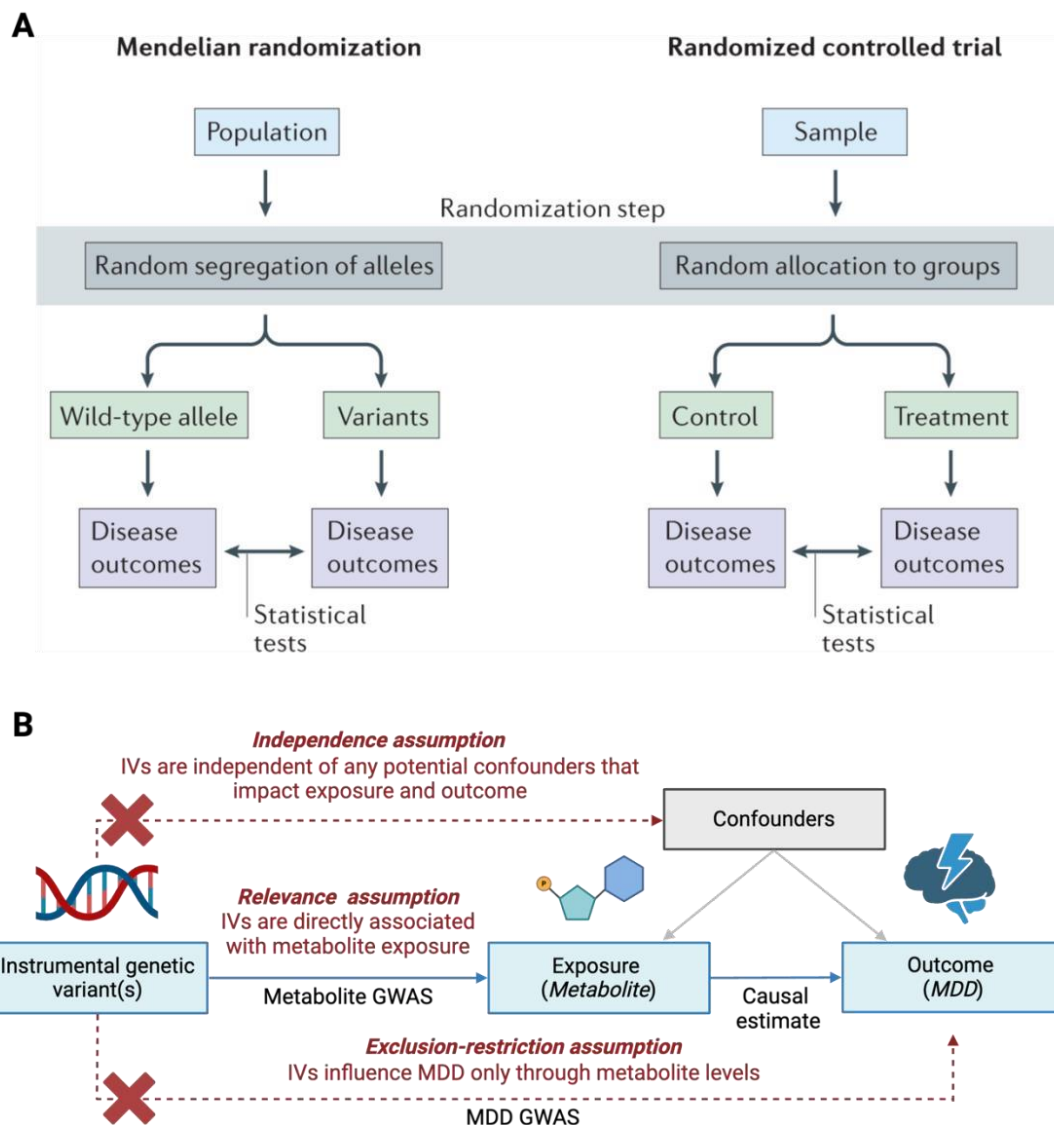


Figure 3.1 Principles and assumptions of Mendelian Randomisation.

A) A schematic showing the analogy of MR (left) to randomised controlled trials (RCTs) (right) via random segregation of alleles at conception. B) Instrumental variable (IVs) framework for MR for assessing causal effect of an exposure on an outcome via vertical pleiotropy (blue boxes). The three core assumptions of MR are shown in red. Part A of this Figure was taken from Sanderson et al (2022)³⁷⁹, whilst part B was made using

3.1.4.3 MR-methods

This thesis specifically used two-sample MR, whereby the estimates of an IVs association on exposure and outcome are taken from GWAS in separate samples³⁸³. In the simplest instance where there is one IV, the causal effect of an exposure on an outcome can be calculated using the Wald Ratio (WR), calculated by the IV's effect on the outcome divided by the IV's effect on the exposure (Equation 3.8).

$$\text{Wald's Ratio} = \frac{\beta_{IV-outcome}}{\beta_{IV-exposure}} \quad \text{Equation 3.8}$$

Where there are multiple IVs for an exposure, meta-analytical methods can be used to combine individual Wald Ratio estimates and detect potential violations to core assumptions. The inverse-variance weighted (IVW) method calculates causal estimate as the mean WR from all IVs, each weighted by the inverse of its variance in the outcome (i.e. giving higher weights to more precise estimates)³⁸⁴. This is equivalent to a weighted linear regression, with the intercept fixed at 0. The weighted median MR method uses the same weighting approach, but causal estimates are obtained from the median WR (instead of the mean) making it more robust to the inclusion of invalid IVs exhibiting horizontal pleiotropy³⁸⁵. The MR-Egger method extends the weighted linear regression (IVW) method to detect the presence of horizontal pleiotropy by allowing the intercept to deviate from 0³⁸². The regression parameters can be used to estimate the presence of horizontal pleiotropy (the intercept) and the causal effect adjusted for any pleiotropy (the slope). However, this method depends upon the Instrument Strength

Independent of Direct Effect (InSIDE) assumption, which assumes that the IV's association with the exposure is not correlated to its pleiotropic effects³⁸². Overall, MR is a powerful tool that can be used to estimate causal effects of an exposure on an outcome unaffected by issues which typically plague observational associations, such as confounding and reverse causation. However, the reliability of MR estimates depends upon three core assumptions, two of which are impossible to prove empirically. Furthermore, due to the LD between neighbouring variants, two traits may be causally linked to distinct variants within a locus which are highly correlated. This possibility would violate the exclusion-restriction assumption of MR by providing a path from the genetic IV to the outcome not through the exposure (but through LD with the outcome causal variant), therefore biasing MR estimates. Colocalisation is a complementary approach which can assess the probability of shared or distinct causal variants for two traits at a genetic locus. This can therefore provide confidence that LD is not biasing MR estimates and also increase confidence in causality between two traits³⁸⁶.

3.1.5 *Colocalisation*

The central motivation behind colocalisation was to identify traits with a shared causal variant at a genetic locus, as if two traits share a causal variant, they are more likely to be linked by a shared causal mechanism³⁸⁷. Colocalisation approaches assess the probability of a shared causal variant or distinct causal variants at a genetic locus between two traits based off GWAS summary statistics³⁸⁷. In this analysis, I used colocalisation from the “*coloc*” R package, which uses Bayesian statistics to assess the probability of five possible hypotheses underlying GWAS signals from two traits at the same locus (Figure 3.2):

H₀: No associations with either trait

H₁: Association with trait 1, not with trait 2

H₂: Association with trait 2, not with trait 1

H₃: Association with trait 1 and trait 2, two independent SNVs

H₄: Association with trait 1 and trait 2, one shared SNV

Using a set of predefined prior probabilities for each hypothesis, colocalisation uses a Bayesian framework to calculate the probability of each hypothesis (H_h) given the observed GWAS summary statistics (D) known as the posterior probability ($P(H_h|D)$). For each hypothesis, this is calculated over all possible configurations (S) of GWAS summary statistics (aligning with all five hypotheses) (Equation 3.9).

$$P(H_h|D) \propto \sum_{S \in S_h} P(D|S)P(S) \quad \text{Equation 3.9}$$

A high posterior probability of distinct causal variants for each trait (H₃), indicates potential violations to the exclusion-restriction assumption of MR, and a high posterior probability of a shared causal variant (H₄) is further evidence for a causal relationship between two traits. However, colocalisation is also subject to the assumptions that the shared causal variant has been directly genotyped or imputed in the GWAS, and that there is at most only one causal variant per trait at the genetic locus³⁸⁷.

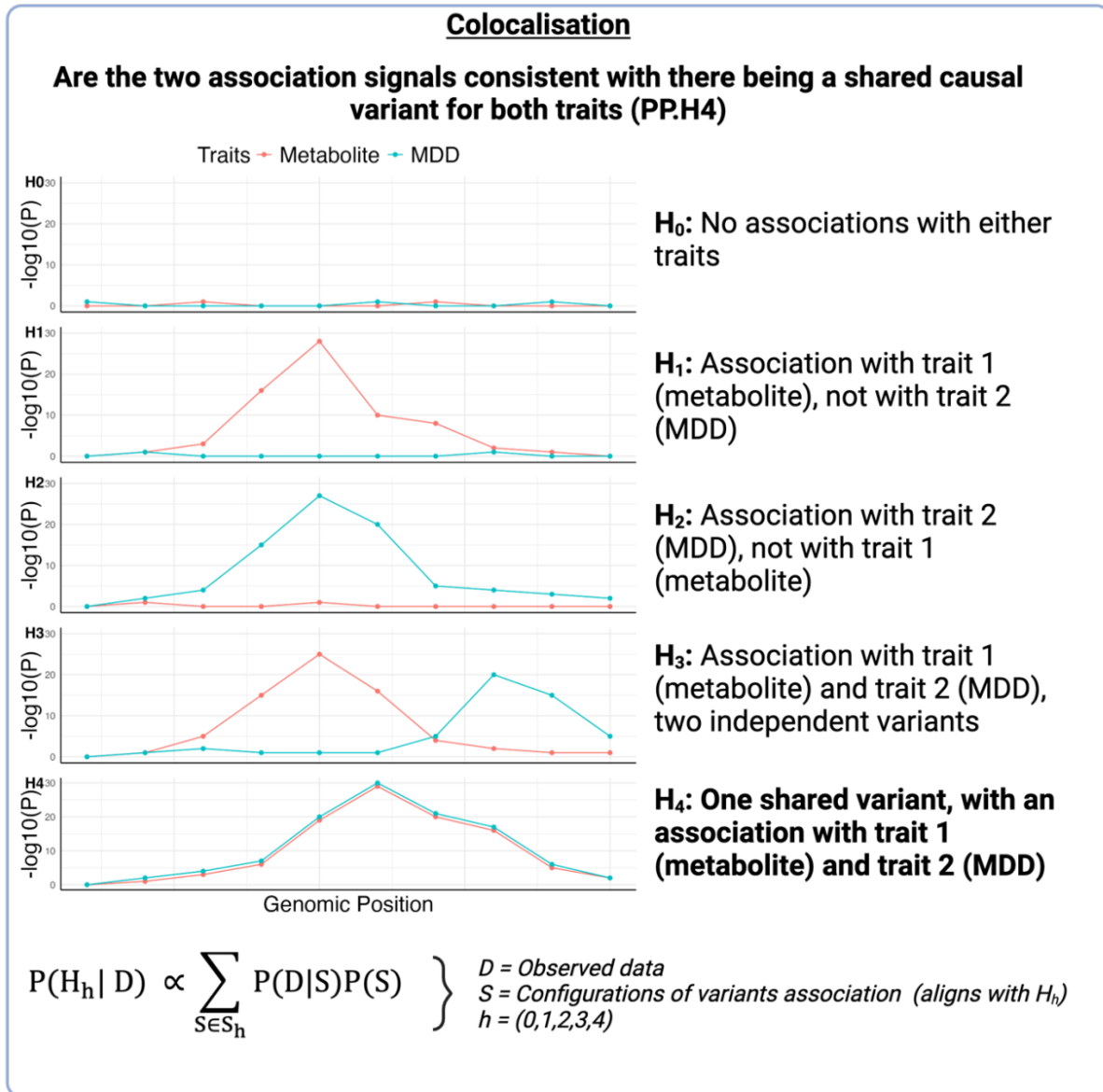


Figure 3.2 Colocalisation framework and hypotheses.

For two traits (here a metabolite and lifetime MDD status), colocalisation uses Bayesian statistics to compute the probability of five possible hypotheses (H₀-H₄) given the GWAS summary statistics at a genetic locus. A high posterior probability of H₄ (shared causal variant) is additional evidence for a causal relationship between the two traits. A higher posterior probability of H₃ (distinct causal variants) indicates potential violations of the exclusion restriction assumption in MR.

3.2 Metabolomic investigation of major depressive disorder identifies a potentially causal association with polyunsaturated fatty acids

Metabolomic Investigation of Major Depressive Disorder Identifies a Potentially Causal Association With Polyunsaturated Fatty Acids

Eleanor Davyson, Xueyi Shen, Danni A. Gadd, Elena Bernabeu, Robert F. Hillary, Daniel L. McCartney, Mark Adams, Riccardo Marioni, and Andrew M. McIntosh

ABSTRACT

BACKGROUND: Metabolic differences have been reported between individuals with and without major depressive disorder (MDD), but their consistency and causal relevance have been unclear.

METHODS: We conducted a metabolome-wide association study of MDD with 249 metabolomic measures available in the UK Biobank ($n = 29,757$). We then applied two-sample bidirectional Mendelian randomization and colocalization analysis to identify potentially causal relationships between each metabolite and MDD.

RESULTS: A total of 191 metabolites tested were significantly associated with MDD (false discovery rate–corrected $p < .05$), which decreased to 129 after adjustment for likely confounders. Lower abundance of omega-3 fatty acid measures and a higher omega-6 to omega-3 ratio showed potentially causal effects on liability to MDD. There was no evidence of a causal effect of MDD on metabolite levels. Furthermore, genetic signals associated with docosahexaenoic acid colocalized with loci associated with MDD within the fatty acid desaturase gene cluster. Post hoc Mendelian randomization of gene-transcript abundance within the fatty acid desaturase cluster demonstrated a potentially causal association with MDD. In contrast, colocalization analysis did not suggest a single causal variant for both transcript abundance and MDD liability, but rather the likely existence of two variants in linkage disequilibrium with one another.

CONCLUSIONS: Our findings suggest that decreased docosahexaenoic acid and increased omega-6 to omega-3 fatty acids ratio may be causally related to MDD. These findings provide further support for the causal involvement of fatty acids in MDD.

<https://doi.org/10.1016/j.biopsych.2023.01.027>

Major depressive disorder (MDD) is a debilitating condition estimated to affect 322 million people, leading to a global total of 50 million years lived with disability (1). This is due to both the high prevalence of MDD and its frequent comorbidity with other conditions (2–4), particularly cardiovascular disease (5). Most antidepressants target monoamine-related pathways (6) but are ineffective in 40% of cases (7). Better understanding of the molecular mechanisms of MDD may aid in the development of more effective treatments.

Twin studies have estimated the heritability of MDD as ~37% (8), and investigating this genetic component of MDD may help to identify its pathophysiological basis. Recently, genome-wide association studies (GWASs) have been increasingly successful in identifying genetic variants associated with MDD (9–11). Despite this success, the high polygenicity of MDD presents significant challenges to researchers who wish to identify actionable mechanisms lying between genetic variants and clinical phenotype (12). One means of translating genetic findings into biomarkers and biological mechanisms is through the incorporation of molecular phenotype data, such as metabolomics. These data may help

to clarify the downstream functional impact of identified risk single nucleotide variants (SNVs) on molecular traits. This information can then be utilized to stratify risk, provide therapeutic targets, or enable lifestyle interventions (13).

There have been several recent studies investigating the links between genetic risk for MDD and molecular phenotypes, including DNA methylation (14,15), messenger RNA (mRNA) levels (16,17), and protein markers (18,19). Metabolomics considers the role of circulating metabolites (20), including lipids, amino acids, and other small molecules that have been implicated in a range of disorders and some hypothesis-driven studies of MDD (21,22). The metabolome has not been characterized to the same extent as other molecular phenotypes (23), but the recent availability of reproducible high-throughput metabolomics makes studies feasible for many thousands of individuals (24). Studying the association between metabolite levels and MDD may offer potential biomarkers and interventional targets for MDD and provides a means to investigate MDD's comorbidity with cardiometabolic disorders (2,3). Most previous metabolic studies of MDD have concentrated on small numbers of metabolites (25), often used sample sizes of

<100 (25–29), and implemented a variety of methodological approaches all leading to somewhat heterogeneous findings (30). However, larger and more standardized studies are beginning to emerge; a recent meta-analysis of 230 metabolite levels in 15,428 individuals (5283 cases and 10,145 controls) found evidence of a distinct metabolomic profile associated with MDD (31).

The UK Biobank (UKB) recently released a large metabolomic dataset (32), in which we conducted a comprehensive and agnostic examination of metabolite associations with MDD. Association analysis between metabolites and MDD is limited because of the susceptibility of these analyses to confounding and reverse causation. We therefore also assessed potentially causal associations between MDD and metabolite levels, using two-sample Mendelian randomization (MR). Finally, we sought to identify potentially causal gene targets using MR and colocalization analysis of expression quantitative trait loci (eQTL) (Figure S1 in Supplement 1).

METHODS AND MATERIALS

Study Population

The UKB is a large prospective study ($N = 502,492$) of participants recruited from around the United Kingdom between the ages of 40 and 69 (33). Baseline data collection, including genotyping and extensive phenotyping for a range of health outcomes, occurred between 2006 and 2010. UKB data acquisition was approved by the North West National Health Service Research Ethics Committee (11/NW/0382). The analysis and data acquisition for the present study were conducted under application #4844. Written informed consent was obtained from all participants.

Metabolic Biomarkers

A total of 249 metabolite biomarkers were quantified in non-fasting baseline EDTA plasma samples ($n = 118,021$) (33), collected at baseline assessment between April 2007 and December 2010 (34), and measured between June 2019 and April 2020 (31). Measurement was performed using the high-throughput nuclear magnetic resonance (NMR) spectroscopy profiling platform, developed by Nightingale Health Ltd. Detailed methodology of the Nightingale NMR pipeline has been described elsewhere (35–37). These biomarkers include detailed cholesterol measures, fatty acid compositions, and low-weight metabolites such as amino acids and ketones (Table S1 in Supplement 2). Most biomarkers were measured in absolute concentration (mmol/L), excluding apolipoproteins A and B (measured in g/L), diameter measures (diameter of very-low-density lipoprotein/low-density lipoprotein/high-density lipoprotein particles in nm), and the degree of unsaturation measure (the total number of pi bonds and rings within a molecule). Prior to analysis, the levels of each metabolite biomarker were transformed to a standard normal distribution using rank-based inverse normalization. Therefore, all reported beta coefficients from regression analyses represent the standardized change in ranked metabolite levels between MDD cases and controls.

Assessment of MDD Status

The MDD phenotype used in this study was derived from the online Thoughts and Feelings Questionnaire, fully completed by 126,077 participants in July 2017 (38). The questionnaire included the Composite International Diagnostic Interview Short Form, from which MDD lifetime diagnoses were derived (38) (Supplemental Methods and Materials in Supplement 1). This sample (referred to as the Mental Health Questionnaire sample) comprised 37,430 cases and 88,647 controls of whom 29,757 (8840 cases and 20,917 controls) also had metabolite levels derived from their blood samples (Table 1).

Metabolome-wide Association Study

Linear regression models were used to test the association of each metabolite level with the MDD phenotype ($n = 29,757$), referred to as a metabolome-wide association study (MetWAS). In the basic MetWAS, only age, sex, assessment center, and NMR spectrometer were included as covariates. MetWAS analysis was also conducted with adjustment for socioeconomic status, smoking status, ethnicity, educational attainment, and body mass index (Supplemental Methods and Materials in Supplement 1) due to the strong and potentially confounding association between these factors with both MDD and metabolite levels (39–42). Significance was defined by false discovery rate (FDR)-corrected p (p_{FDR}) values $< .05$.

Metabolite GWASs

For each metabolite, a GWAS was performed on unrelated individuals of European ancestry ($n = 88,329$) using PLINK 2.0 (43). Unrelated individuals were identified as previously reported (44). In brief, related individuals were initially excluded based on a shared relatedness of up to the third degree using kinship coefficients (>0.044) calculated using the KING toolset (45). One member of each group of related individuals was then subsequently added back in, by creating a genomic relationship matrix and selecting individuals with a genetic relatedness of <0.025 with any other participant. Association testing was performed on the imputed genotypes using an additive model adjusted for age, sex, genotyping array, assessment center, spectrometer, and 10 genetic principal components. Variants with minor allele frequency (MAF) < 0.001 , minor allele count < 20 , and an INFO score < 0.1 were removed. Insertion-deletions were included in the analysis. To account for multiple comparisons, significance was defined as the

Table 1. Demographics of MDD Cases and Controls

Demographic Characteristics	Controls, $n = 20,917$	MDD Cases, $n = 8840$
Age, Years, Mean (SD)	56.67 (7.67)	54.12 (7.55)
BMI, Mean (SD)	26.48 (4.18)	27.26 (5.12)
Female, %	49.95%	68.82%
SES, Mean	−1.92	−1.37
Never Smoked, %	59.72%	52.58%
Ethnicity (Mixed/Other/White), %	0.45%/2.2%/97%	0.7%/2%/97%
Attended University/College, %	46.36%	45.67%

BMI, body mass index; MDD, major depressive disorder; SES, socioeconomic status.

genome-wide threshold divided by the number of principal components, which account for 99% of the variance within the metabolomic data ($5 \times 10^{-8}/42 = 1.19 \times 10^{-9}$), as described previously (46,47). An F statistic for the strength of the association between the sentinel SNVs and metabolite levels was calculated using the method: $F = ([N - k - 1] / k) \times (r^2 / [1 - r^2])$, in which N = sample size, k = number of SNVs, and r^2 = variance explained in metabolite levels by the genetic instruments. The r^2 statistic was calculated as follows: $r^2 = 2 \times \text{MAF} \times (1 - \text{MAF}) \times \text{beta}^2$, in which beta = effect size of the SNV and MAF = the SNV effect allele frequency. SNV heritability for each metabolite and the genetic correlation of the metabolite with MDD was performed using linkage disequilibrium (LD) score regression as implemented in the LDSC package (<https://github.com/bulik/ldsc>) (48,49). Precomputed LD scores based on the 1000 Genomes Project data for European ancestry (50) were used as the reference panel for regression weights in the analysis (<https://data.broadinstitute.org/alkesgroup/LDSCORE>).

Metabolite Bidirectional MR Analysis

Bidirectional MR was performed using the TwoSampleMR R package, version 0.5.6 (51,52) to assess causality between the metabolite (exposure) and MDD (outcome) for all metabolites, and vice versa. The MR analysis used SNV effect sizes from the UKB metabolite GWASs conducted in this study and the MDD meta-analysis GWAS, performed by the Psychiatric Genomics Consortium (9), with UKB participants removed using 3 MR methods (inverse variance weighted, MR Egger, and weighted median). Genetic instruments were identified by clumping genome-wide significant SNVs ($p < 1.19 \times 10^{-9}$ and $\text{MAF} > 0.005$) using the `clump_data` function in the R package TwoSampleMR ($r^2 = 0.001$, within a 10-Mb window). Metabolites with <15 genetic instruments after harmonization with MDD summary statistics were omitted from the MR analysis.

Colocalization

Colocalization analysis was performed between MDD and metabolite genetic signals that showed potentially causal associations using MR, with a window of 1 Mb. Lead SNVs were identified using the default clumping method ($r^2 = 0.1$) on the SNP2GENE tool within FUMA software (53). To avoid running multiple colocalization tests of overlapping regions, lead SNVs were ordered by significance and iteratively filtered so that each lead SNV was >1 Mb away from any other lead SNV. For each of the final lead SNVs, all SNVs within a 1-Mb window from the metabolite GWAS and the MDD GWAS were extracted from the respective summary statistics. Colocalization analysis was then performed using the R `coloc` package (version 5.1.1) (54). The prior probabilities were set to default values: P1 (SNV is associated with the metabolite) and P2 (SNV is associated with MDD) were both set at 1×10^{-4} , and P12 (SNV is associated with both MDD and the metabolite) was set at 1×10^{-5} . An arbitrary posterior probability > 0.8 was set for evidence of each of the 4 hypotheses: PP.H0—no causal variant, PP.H1 and PP.H2—causal variant for one of the traits, PP.H3—distinct causal variant for each trait, and PP.H4—shared causal variant for both traits.

MR and Colocalization With eQTL Data

To probe possible mechanistic pathways between exposure (metabolite) and outcome (MDD), colocalized MDD and metabolite genetic signals were investigated to identify whether they localized to any causal genes within the colocalized region. For genes lying within the same LD window as colocalized MDD-metabolite genetic signals, we first identified any cis-eQTLs from whole blood gene expression data ($n = 948$) from the Genotype-Tissue Expression (GTEx) Project (48). Second, MR and complementary colocalization analyses were performed to assess the causality of gene expression levels (mRNA) with MDD. The MR analysis was performed using a Wald ratio test (55), using a single instrumental SNV defined as the most significant eQTL present in the dataset. For this post hoc analysis, significance was defined as $p < .05$. The instrumental SNVs were also used as lead SNVs in the colocalization analysis. For each lead SNV, MDD GWAS (9) and GTEx (56) summary statistics for variants within ± 1 Mb were extracted as inputs for the colocalization analysis.

RESULTS

MetWAS MDD Association Analysis

The demographics of the Mental Health Questionnaire sample (metabolomic and MDD data available; $n = 29,757$) are presented in Table 1. The basic MetWAS, without adjustment for common confounders, found that 191 (~76%) metabolites were significantly associated with MDD ($p_{\text{FDR}} < .05$) (Figure S2 in Supplement 1; Table S2 in Supplement 2). The most significant positive metabolite-MDD association was for monounsaturated fatty acids to total fatty acids (%), ($\beta = 0.177$, $p_{\text{FDR}} = 1.52 \times 10^{-43}$), and the most significant negative association was for polyunsaturated fatty acids to monounsaturated fatty acids ratio ($\beta = -0.176$, $p_{\text{FDR}} = 1.85 \times 10^{-43}$). After adjustment for possible confounders, 133 (~53%) metabolites were significantly associated with MDD, 129 of which (~52%) were shared in the unadjusted model (Figure 1; Table S3 in Supplement 2). The most significant positive metabolite-MDD association remained monounsaturated fatty acids to total fatty acids ($\beta = 0.069$, $p_{\text{FDR}} = 2.22 \times 10^{-7}$), while the most significant negative metabolite-MDD association changed to free cholesterol to total lipids in large low-density lipoprotein percentage ($B = -0.070$, $p_{\text{FDR}} = 2.82 \times 10^{-7}$). Across most metabolites ($n = 224$, unadjusted β range: -0.176 to 0.177), inclusion of confounders into the analysis attenuated the absolute effect sizes by an average of 61% (adjusted β range: -0.070 to 0.069) (Figure S3 in Supplement 1).

Metabolite GWASs

Manhattan plots for all 249 GWASs can be found in Supplement 3. All metabolites had multiple independent ($r^2 = 0.001$, within a 10-Mb window) genome-wide significant SNV associations ($p < 1.19 \times 10^{-9}$), all with F statistics > 10 (Table S4 in Supplement 2). The number of significant independent SNVs ranged from 4 (phenylalanine and acetoacetate) to 60 (cholesteryl esters in large high-density lipoprotein) (Table S5 in Supplement 2). SNV heritability of metabolites varied from 2.41% (histidine) to 19.9% (triglycerides to phosphoglycerides ratio). A total of 144 metabolites (57%) showed a significant

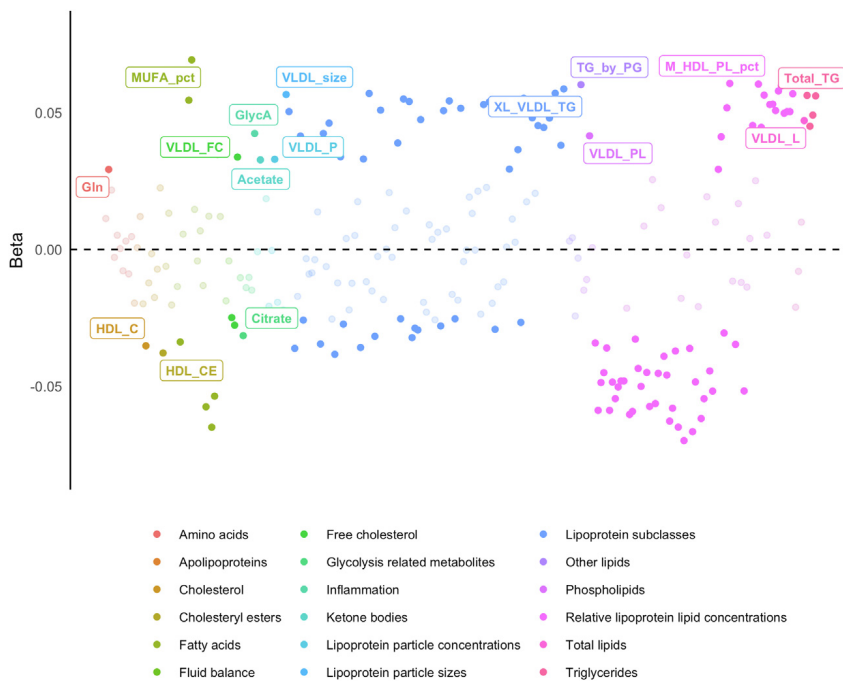


Figure 1. Adjusted metabolome-wide association study. Standardized effect sizes for metabolite–major depressive disorder + covariates. Significant associations (false discovery rate–corrected $p < .05$) are shaded in, and the most significant association for each metabolite group is labeled. C, cholesterol; CE, cholesteryl esters; FC, free cholesterol; Gln, glutamine; GlycA, glycoprotein acetyls; HDL, high-density lipoprotein; L, large; LDL, low-density lipoprotein; M, medium; MUFA, monounsaturated fatty acids; P, particles; pct, percentage; PG, phosphoglycerides; PL, phospholipid; TG, triglyceride; VLDL, very-low-density lipoprotein; XL, extra-large.

genetic correlation with MDD (Table S6 in Supplement 2), with absolute correlations ranging from (–) 0.073 (omega-3 fatty acids to total fatty acids [%]) to (+) 0.180 (phospholipids to total lipids in large high-density lipoprotein [%]). Of the metabolites that showed a significant genetic correlation with MDD, 119 (~83%) also had significant associations with the MDD phenotype in both the unadjusted and adjusted MetWAS analyses.

Mendelian Randomization

In the metabolite (exposure) to MDD (outcome) MR analysis, 21 metabolites were excluded for having too few (<15) genetic instruments (Figure S4 in Supplement 1). MR analysis on the remaining 228 metabolites (Figure 2; Figure S5 in Supplement 1) found 5 with significant evidence for a causal relationship with MDD (Table 2; Table S7 in Supplement 2), all relating to long-chain polyunsaturated fatty acids (LC-PUFAs). Specifically, omega-3 fatty acid levels (omega-3 fatty acids, omega-3 to total fatty acids [%], and docosahexaenoic acid [DHA]) and the degree of unsaturation had a negative causal effect on MDD, while the omega-6 to omega-3 ratio had a positive causal effect on lifetime MDD (Figure S6 in Supplement 1). These metabolites also had strong correlations with the other metabolites that significantly associated with MDD in the MetWAS phenotypic analysis, ranging from omega-3 fatty acids (significantly correlated with 79% of associated metabolites) to the degree of unsaturation (significantly correlated with 100% of the other MDD-associated metabolites) (Figure S7 in Supplement 1). Of the 5 putatively causal metabolites, only the degree of unsaturation showed consistency in previous observational analyses, maintaining a significant association and genetic correlation with MDD (Figure S8 in Supplement 1). In the MR analysis in

which one SNV was left out, 3 of the significant metabolites to MDD MR results—degree of unsaturation, omega-3 fatty acids, and DHA—were dependent on rs174528, a single SNV in the *FADS* gene cluster on chromosome 11 (Figure S9 in Supplement 1; Table S8 in Supplement 2).

In the MDD (exposure) to metabolite (outcome) MR analysis, all metabolite measures were included. In this analysis, there was no significant evidence that MDD was significantly causal to a change in any of the metabolite levels tested (β range: –0.28 to 0.26, p_{FDR} range: .91 to .99) (Figure S10 in Supplement 1; Table S9 in Supplement 2).

Colocalization

The number of lead SNVs identified across the genome for the 5 metabolites with MR $p_{FDR} < .05$ ranged from 20 (omega-3 fatty acids to total fatty acids [%]) to 30 (DHA) (Table S10 in Supplement 2). DHA had strong evidence for colocalization (PP.H4 = 0.89) with MDD at SNV rs2727271 (Table 2), which is an intronic variant for the fatty acid desaturase 2 (*FADS2*) gene (Figure 3). The rest of the metabolites showed weak evidence of colocalization with MDD (PP.H4 < 0.2) (Figure S11 in Supplement 1; Table S11 in Supplement 2).

MR and Colocalization With eQTL Data

Downstream eQTL–MDD MR analysis was performed for the genes myelin regulatory factor (*MYRF*), flap endonuclease 1 (*FEN1*), transmembrane protein 258 (*TMEM258*), and *FADS1/2/3* (Figure 3D) because these were within the DHA–MDD colocalized region on chromosome 11 and had eQTL data available from GTEx. The analysis found that the mRNA levels of *MYRF*, *TMEM258*, *FADS1*, *FADS2*, and *FADS3* had potentially positive causal effects on lifetime MDD status [odds ratio

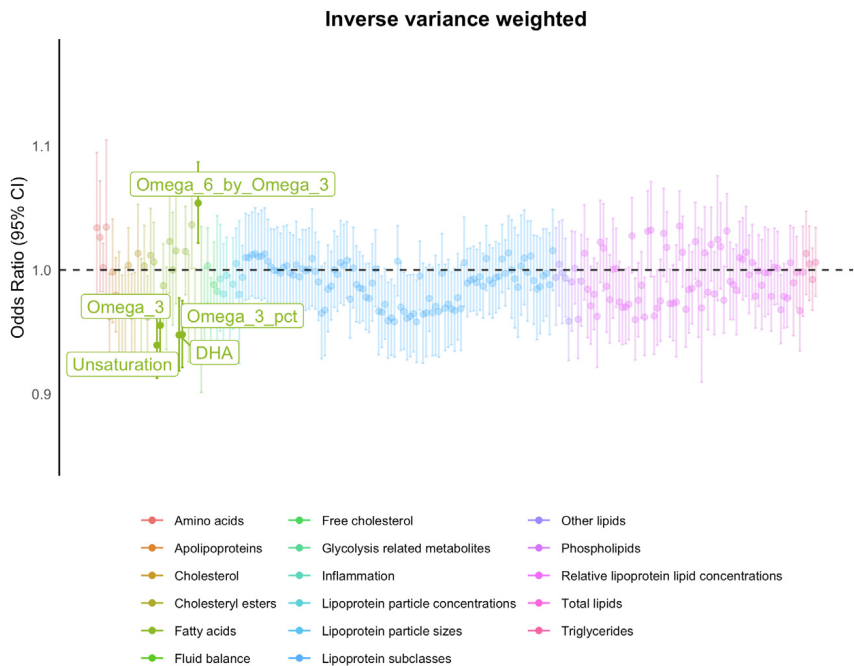


Figure 2. Metabolite to major depressive disorder Mendelian randomization analysis. The odds ratios of metabolite on lifetime major depressive disorder status given by the inverse variance weighted Mendelian randomization method, colored by metabolite group. Significant Mendelian randomization findings (false discovery rate-corrected $p < .05$) are filled in and labeled. Unsaturations indicates the degree of unsaturation, Omega_3_pct indicates omega-3 to total fatty acids (%), Omega_3 indicates omega-3 fatty acids, and Omega_6_by_Omega_3 indicates the omega-6 to omega-3 ratio. DHA, docosahexaenoic acid.

range: 1.01–1.12, $p = 7.58 \times 10^{-7}$ to .011] (Figure S12 in Supplement 1; Table 3). In contrast, mRNA levels of *FEN1* were not significantly causal to MDD status (odds ratio = 0.98, $p = .62$). Subsequent colocalization analysis for the *FADS* cluster eQTLs and MDD found minimal evidence of shared causal variant for both gene expression and MDD. Supporting evidence was strongest for the existence of two independent causal variants for each trait at the same locus (PP.H3 > 0.7) (Table S12 in Supplement 2).

DISCUSSION

This study reports a large-scale ($n = 29,757$) association and causal analysis of blood metabolites with MDD. We found that 191 metabolites were significantly associated with lifetime MDD status, 129 of which withstood further adjustment for common confounders (39). These findings provide greater confidence that the significant metabolite-MDD associations observed in this study are not solely due to the effects of

Table 2. Significant Metabolite to MDD MR Results

Metabolite	Method	N (SNVs)	Beta	SE	p	p_{FDR}	PP.H4
Degree of Unsaturation	MR Egger	25	−0.063	0.021	6.20×10^{-3}	.893	0.026
	WM	25	−0.069	0.016	1.20×10^{-5}	.002	
	IVW	25	−0.063	0.015	1.80×10^{-5}	.004	
Omega-3 Fatty Acids	MR Egger	23	−0.044	0.021	5.10×10^{-2}	.893	0.13
	WM	23	−0.068	0.016	1.30×10^{-5}	.002	
	IVW	23	−0.046	0.013	3.30×10^{-4}	.025	
Docosahexaenoic Acid	MR Egger	24	−0.066	0.025	1.60×10^{-2}	.893	0.89
	WM	24	−0.078	0.019	4.50×10^{-5}	.003	
	IVW	24	−0.054	0.016	6.80×10^{-4}	.039	
Omega-3 Fatty Acids to Total Fatty Acids (%)	MR Egger	23	−0.057	0.021	1.30×10^{-2}	.893	0.16
	WM	23	−0.063	0.018	4.80×10^{-4}	.022	
	IVW	23	−0.054	0.014	2.20×10^{-4}	.025	
Omega-6 Fatty Acids to Omega-3 Fatty Acids Ratio	MR Egger	18	0.062	0.026	3.10×10^{-2}	.893	0.15
	WM	18	0.067	0.019	4.70×10^{-4}	.022	
	IVW	18	0.052	0.016	8.80×10^{-4}	.040	

All 3 MR methods for the 5 metabolites which showed a significant causal relationship with MDD in any MR method. Colocalization analysis was performed for each metabolite for multiple lead SNVs, the top colocalization result for each metabolite is reported in this table under PP.H4. The full table of colocalization results is found in Table S10 in Supplement 3.

IVW, inverse variance weighted; MDD, major depressive disorder; MR, Mendelian randomization; SNVs, single nucleotide variants; WM, weighted median.

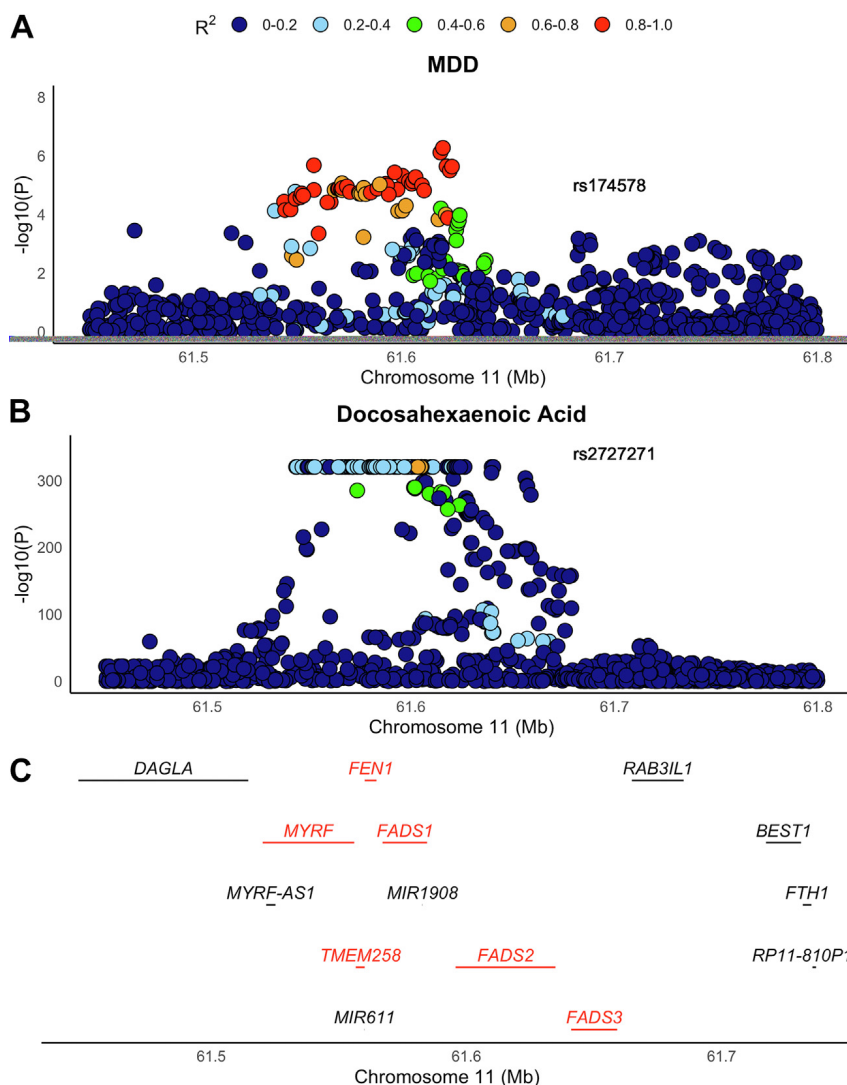


Figure 3. Colocalization of docosahexaenoic acid with MDD. **(A)** The association of single nucleotide variants in the *FADS* cluster in the MDD genome-wide association studies (without UK Biobank). **(B)** The association of single nucleotide variants in the *FADS* cluster in the docosahexaenoic acid genome-wide association studies. **(C)** The genes present in the *FADS* cluster. Genes colored red are those that were present in the GTEx (Genotype-Tissue Expression) data and consequently tested for a causal relationship with MDD in the post hoc analysis. MDD, major depressive disorder.

confounding. Furthermore, we found that 144 metabolites showed a significant genetic correlation with MDD, suggesting a shared genetic architecture between the traits. Bidirectional MR analyses found evidence for potentially causal relationships between 5 metabolite measures relating to LC-PUFAs and MDD. Complementary colocalization analyses between these metabolites and MDD found that an omega-3 fatty acid, DHA colocalized with MDD within the *FADS* region. Subsequent MR analysis of eQTL data of genes in the *FADS* cluster found possible evidence of causality between *MYRF*, *FADS1*, *FADS2*, *FADS3*, and *TMEM258* mRNA abundance and MDD. However, complementary colocalization analysis found evidence that this locus contains separate independent causal variants for mRNA abundance and MDD.

The MR analysis provides evidence that omega-3 LC-PUFAs may have a negative causal effect on lifetime MDD, while the ratio of omega-6 to omega-3 LC-PUFAs may have a positive causal effect on lifetime MDD. Specifically,

Table 3. Expression Quantitative Trait Loci to MDD Wald Ratio MR Results

Gene	SNV	OR	95% CI	MR <i>p</i> Value
<i>FADS1</i>	rs174567	1.08	1.05–1.10	7.58×10^{-7}
<i>FADS2</i>	rs968567	1.01	1.01–1.02	6.33×10^{-5}
<i>FADS3</i>	rs2524288	1.05	1.01–1.10	.011
<i>FEN1</i>	rs10897119	0.98	0.89–1.07	.617
<i>MYRF</i>	rs198462	1.03	1.01–1.05	.001
<i>TMEM258</i>	rs7943728	1.12	1.06–1.19	5.78×10^{-5}

These results indicate a positive causal effect of *FADS1*, *FADS2*, *FADS3*, *MYRF*, and *TMEM258* on lifetime MDD status.

MDD, major depressive disorder; MR, Mendelian randomization; OR, odds ratio; SNV, single nucleotide variant.

colocalization analysis provided further support for a shared causal variant between MDD and DHA. Research on omega-3 and omega-6 levels and MDD has been lengthy and complex

(25,57–62). Although many cross-sectional and prospective studies have observed a reduction in omega-3 fatty acids, and an increase in the omega-6 to omega-3 ratio in those with MDD (30,58,63), it is important to note both the considerable methodological heterogeneity and contradictory evidence (30,61,64). Clinical trial studies testing the efficacy of omega-3 supplements in the prevention and treatment of MDD also show inconsistent findings (65–68). Notably, our findings contradict results from a two-sample MR study in 2019, which found no evidence of a significant causal effect of omega-3 levels on MDD (69). However, this study leveraged summary statistics from smaller GWASs, and therefore may have been underpowered to detect causal effects. The enzymatic oxygenation of omega-3 and omega-6 fatty acids and their derived eicosanoids have opposing anti-inflammatory and proinflammatory effects, respectively (Figure S13 in Supplement 1). Importantly, humans do not show endogenous biosynthesis of active omega-3 and omega-6 fatty acids (eicosapentaenoic acid and arachidonic acid, respectively) and the levels of each are reliant on the intake of their precursors, linoleic acid and alpha-linoleic acid, from the diet (70). Over the last century, the ratio of omega-6 to omega-3 has increased dramatically in many Western diets (20–30:1 compared with 1:1) (71), resulting in an imbalance in proinflammatory and proresolving mediators that may contribute to MDD pathology (72). Furthermore, there is a *FADS* haplotype associated with more efficient LC-PUFA synthesis from the diet (73). While this is evolutionarily advantageous in environments with limited LC-PUFA availability, it may heighten the imbalance between omega-6 and omega-3 levels resulting from modern Western diets (73). Overall, this indicates the possibility that a disproportionate dietary ratio of omega-6 to omega-3 may be contributing to MDD through increased inflammatory processes and that these may be exacerbated by certain genetic haplotypes.

The post hoc MR analysis found evidence of a possible causal association between *MYRF*, *FADS1*, *FADS2*, *FADS3*, and *TMEM258* mRNA levels and MDD. *MYRF* and *TMEM258* each have plausible mechanisms for their association with MDD due to links to abnormal myelination and intestinal stress, respectively (74–76). The *FADS* enzymes however are directly linked to the metabolism of both omega-3 and omega-6 LC-PUFAs (77,78) and have been shown to be central regulators of the ratio of their respective anti-inflammatory and proinflammatory lipid mediators (79). Therefore, this suggests that *FADS* enzymes act as key regulatory molecules in the activation and resolution of inflammatory processes, which may be important in the pathophysiology of MDD. A recent GWAS of bipolar disorder in Japanese populations also found a genetic association with *FADS* intron variants (80), indicating that this association may not be specific to MDD, but rather shared across different psychiatric disorders. Although this aligns with evidence that MDD and bipolar disorder share genetic architecture (81), whether the disorders share the putative causal effect of *FADS* found in this study would require additional analysis dissecting the causal relationship of the *FADS* cluster on bipolar disorder.

Despite these putative causal mechanisms, complementary colocalization analysis between all eQTLs and MDD found

evidence that each trait had distinct, rather than shared, causal variants at the locus. This indicates that the instrumental variable in MR exhibits horizontal pleiotropy on the outcome (a direct violation of the exchangeability assumption for MR), thus weakening the evidence for causality (82,83). Therefore, it cannot be determined with high confidence that the transcript abundance of the genes within the *FADS* cluster has a direct causal relationship with MDD, but rather that it may be commonly seen alongside the presentation of MDD.

This study overcomes various limitations that existed in previous metabolomic analyses of MDD. First, this study analyzed the association and causal relationship between MDD and 249 metabolites in a hypothesis-free manner, and therefore reduces the possibility of confirmation bias, which has been criticized in previous studies with an a priori focus on LC-PUFAs (25). Second, previous studies often had small sample sizes, each with distinct study methodology and often suffering from self-report bias (30). This study utilizes a large sample size from which metabolite levels were biologically measured in the blood using NMR spectroscopy in a standardized, well-documented procedure (35,37). Our findings are also strengthened by the integration of eQTL data from GTEx version 8 (56), which interrogated the role of the gene products from the colocalized *FADS* region.

This study is limited because it tests the association of MDD with plasma metabolite levels, despite MDD primarily being a brain-based disorder. However, blood samples are minimally invasive to collect, which is essential for enabling large-scale high-powered studies and affords them the most promise for future clinical applications. Another limitation to this analysis is the high LD within the *FADS* region (73), meaning that a direct violation of the exclusion-restriction assumption in MR may exist, generating false-positive results (84). However, the complementary use of MR and colocalization analysis in this study acts as a sensitivity analysis (82). Another limitation is the focus of this analysis on European ancestry, especially because the frequencies of the *FADS* haplotypes differ geographically. The divergent *FADS* haplotypes are more commonly seen in non-European populations (85) and are associated with an increased expression of *FADS1* and a more efficient LC-PUFA synthesis (73). Consequently, the mechanistic effects hypothesized in this study may be greater in non-European ancestral groups, which are not currently captured in this analysis. Additionally, although this is a large-scale metabolomic analysis, it is not wholly comprehensive for the human metabolome (86). Indeed, the recent human metabolome database (87) detailed 18,557 metabolites that had been quantified and estimated the presence of thousands more. Finally, as this study focused on MDD prevalence, it may be biased by cases of chronic MDD within the cohort (88). Future work should therefore investigate whether omega-3 fatty acid levels and the omega-6 to omega-3 ratio associate with incident MDD.

In summary, this study interrogated the links between the metabolome and MDD and indicated a protective role for DHA against MDD. Furthermore, this study also suggests a role for an increased ratio of omega-6 to omega-3 fatty acids in the pathophysiology of MDD.

ACKNOWLEDGMENTS AND DISCLOSURES

This work was supported by two Wellcome Trust grants (Investigator Award in Science: 220857/Z/20/Z and Strategic Award 104036/Z/14/Z [to AMM]). Funding from the Biotechnology and Biological Sciences Research Council and Medical Research Council is gratefully acknowledged (Grant Nos. MR/X003434/1 and MR/W014386/1). The Psychiatric Genomics Consortium (including AMM) has received major funding from the U.S. National Institute of Mental Health (Grant No. 5 U01MH109528-03). ED was supported by the UK Research and Innovation (Grant No. EP/S02431X/1), UK Research and Innovation Centre for Doctoral Training in Biomedical AI at the University of Edinburgh, School of Informatics. DAG is supported by funding from the Wellcome Trust 4-year Ph.D. in Translational Neuroscience—training the next generation of basic neuroscientists to embrace clinical research (108890/Z/15/Z). AMM is also supported by a UK Research and Innovation award (Grant No. MR/W014386/1) and by European Union Horizon 2020 funding (Grant No. 847776).

We thank the participants of the UK Biobank. We also thank the UK Biobank team for collecting and preparing the data for analyses. We would like to thank the research participants and investigators who contributed to the previous meta-analysis of GWAS studies of depression, and the research participants and employees of 23andMe Inc., that conducted work enabling this analysis.

Figures S1 and S13 in Supplement 1 were created with licensed BioRender.com.

Data acquisition and analyses were conducted using the UK Biobank Resource under approved project #4844.

Code is available at https://github.com/Elladavyson/Metabolomics_MDD. GWAS summary statistics are available via the GWAS catalog.

DAG is a scientific consultant for Optima Partners. RFH has received consultant fees from Illumina and is a scientific consultant for Optima Partners. RM has received speaker fees from Illumina, is an adviser to the Epigenetic Clock Development Foundation, and a scientific consultant for Optima Partners. All other authors report no biomedical financial interests or potential conflicts of interest.

ARTICLE INFORMATION

From the Division of Psychiatry, University of Edinburgh, Edinburgh, United Kingdom (ED, XS, MA, AMM); and Institute of Genetics and Cancer, University of Edinburgh, Edinburgh, United Kingdom (ED, DAG, EB, RFH, DLM, RM, AMI).

Address correspondence to Andrew M. McIntosh, M.D., at andrew.mcintosh@ed.ac.uk.

Received Jun 30, 2022; revised and accepted Jan 31, 2023.

Supplementary material cited in this article is available online at <https://doi.org/10.1016/j.biopsych.2023.01.027>.

REFERENCES

- GBD 2017 Disease and Injury Incidence and Prevalence Collaborators (2018): Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: A systematic analysis for the Global Burden of Disease Study 2017 [published correction appears in *Lancet* 2019;393:e44]. *Lancet* 392:1789–1858.
- Oladeji BD, Gureje O (2013): The comorbidity between depression and diabetes. *Curr Psychiatry Rep* 15:390.
- Nemeroff CB, Goldschmidt-Clermont PJ (2012): Heartache and heartbreak—The link between depression and cardiovascular disease. *Nat Rev Cardiol* 9:526–539.
- Cizza G, Ravn P, Chrousos GP, Gold PW (2001): Depression: A major, unrecognized risk factor for osteoporosis? *Trends Endocrinol Metab* 12:198–203.
- Glassman AH (2007): Depression and cardiovascular comorbidity. *Dialogues Clin Neurosci* 9:9–17.
- Robinson E (2018): Psychopharmacology: From serendipitous discoveries to rationale design, but what next? *Brain Neurosci Adv* 2: 2398212818812629.
- Souery D, Serretti A, Calati R, Oswald P, Massat I, Konstantinidis A, et al. (2011): Switching antidepressant class does not improve response or remission in treatment-resistant depression. *J Clin Psychopharmacol* 31:512–516.
- Sullivan PF, Neale MC, Kendler KS (2000): Genetic epidemiology of major depression: Review and meta-analysis. *Am J Psychiatry* 157:1552–1562.
- Howard DM, Adams MJ, Clarke TK, Hafferty JD, Gibson J, Shihali M, et al. (2019): Genome-wide meta-analysis of depression identifies 102 independent variants and highlights the importance of the prefrontal brain regions. *Nat Neurosci* 22:343–352.
- Wray NR, Ripke S, Mattheisen M, Trzaskowski M, Byrne EM, Abdellaoui A, et al. (2018): Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nat Genet* 50:668–681.
- Levey DF, Stein MB, Wendt FR, Pathak GA, Zhou H, Aslan M, et al. (2021): Bi-ancestral depression GWAS in the Million Veteran Program and meta-analysis in >1.2 million individuals highlight new therapeutic directions. *Nat Neurosci* 24:954–963.
- McIntosh AM, Sullivan PF, Lewis CM (2019): Uncovering the genetic architecture of major depression. *Neuron* 102:91–103.
- Akiyama M (2021): Multi-omics study for interpretation of genome-wide association study. *J Hum Genet* 66:3–10.
- Barbu MC, Shen X, Walker RM, Howard DM, Evans KL, Whalley HC, et al. (2021): Epigenetic prediction of major depressive disorder. *Mol Psychiatry* 26:5112–5123.
- Shen X, Caramaschi D, Adams MJ, Walker RM, Min JL, Kwong A, et al. (2022): DNA methylome-wide association study of genetic risk for depression implicates antigen processing and immune responses. *Genome Med* 14:36.
- Pain O, Glanville KP, Hagenaars S, Selzam S, Fürtjes A, Coleman JRI, et al. (2021): Imputed gene expression risk scores: A functionally informed component of polygenic risk. *Hum Mol Genet* 30:727–738.
- Fabbri C, Pain O, Hagenaars SP, Lewis CM, Serretti A (2021): Transcriptome-wide association study of treatment-resistant depression and depression subtypes for drug repurposing. *Neuropsychopharmacology* 46:1821–1829.
- van Haeringen M, Milanese Y, Lamers F, Penninx BWJH, Jansen R (2023): Dissection of depression heterogeneity using proteomic clusters. *Psychol Med* 53:2904–2912.
- Wingo TS, Liu Y, Gerasimov ES, Gockley J, Logsdon BA, Duong DM, et al. (2021): Brain proteome-wide association study implicates novel proteins in depression pathogenesis. *Nat Neurosci* 24:810–817.
- Idle JR, Gonzalez FJ (2007): Metabolomics. *Cell Metab* 6:348–351.
- Hillbrand M, Waite BM, Miller DS, Spitz RT, Lingswiler VM (2000): Serum cholesterol concentrations and mood states in violent psychiatric patients: An experience sampling study. *J Behav Med* 23:519–529.
- Holthoff VA, Beuthien-Baumann B, Zündorf G, Triemer A, Lüdecke S, Winiecki P, et al. (2004): Changes in brain metabolism associated with remission in unipolar major depression. *Acta Psychiatr Scand* 110:184–194.
- Humer E, Probst T, Pieh C (2020): Metabolomics in psychiatric disorders: What we learn from animal models. *Metabolites* 10:72.
- Ala-Korpela M, Kangas AJ, Soininen P (2012): Quantitative high-throughput metabolomics: A new era in epidemiology and genetics. *Genome Med* 4:36.
- Mocking RJT, Assies J, Ruhé HG, Schene AH (2018): Focus on fatty acids in the neurometabolic pathophysiology of psychiatric disorders. *J Inher Metab Dis* 41:597–611.
- Kawamura N, Shinoda K, Sato H, Sasaki K, Suzuki M, Yamaki K, et al. (2018): Plasma metabolome analysis of patients with major depressive disorder. *Psychiatry Clin Neurosci* 72:349–361.
- Adams PB, Lawson S, Sanigorski A, Sinclair AJ (1996): Arachidonic acid to eicosapentaenoic acid ratio in blood correlates positively with clinical symptoms of depression. *Lipids* 31(suppl):S157–S161.
- McNamara RK, Liu Y (2011): Reduced expression of fatty acid biosynthesis genes in the prefrontal cortex of patients with major depressive disorder. *J Affect Disord* 129:359–363.
- Kaddurah-Daouk R, Yuan P, Boyle SH, Matson W, Wang Z, Zeng ZB, et al. (2012): Cerebrospinal fluid metabolome in mood disorders—remission state has a unique metabolic profile. *Sci Rep* 2:667.
- Appleton KM, Voyias PD, Sallis HM, Dawson S, Ness AR, Churchill R, Perry R (2021): Omega-3 fatty acids for depression in adults. *Cochrane Database Syst Rev* 11:CD004692.

31. Bot M, Milanese Y, Al-Shehri T, Amin N, Garmaeva S, Onderwater GLJ, *et al.* (2020): Metabolomics profile in depression: A pooled analysis of 230 metabolic markers in 5283 cases with depression and 10,145 controls. *Biol Psychiatry* 87:409–418.
32. Julkunen H, Cichońska A, Slagboom PE, Würtz P; Nightingale Health UK Biobank Initiative (2021): Metabolic biomarker profiling for identification of susceptibility to severe pneumonia and COVID-19 in the general population. *eLife* 10:e63033.
33. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, *et al.* (2015): UK Biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med* 12:e1001779.
34. Elliott P, Peakman TC; UK Biobank (2008): The UK biobank sample handling and storage protocol for the collection, processing and archiving of human blood and urine. *Int J Epidemiol* 37:234–244.
35. Würtz P, Kangas AJ, Soininen P, Lawlor DA, Davey Smith G, Ala-Korpela M (2017): Quantitative serum nuclear magnetic resonance metabolomics in large-scale epidemiology: A primer on -omic technologies. *Am J Epidemiol* 186:1084–1096.
36. Tzoulaki I, Ebbels TMD, Valdes A, Elliott P, Ioannidis JP (2014): Design and analysis of metabolomics studies in epidemiologic research: A primer on -omic technologies. *Am J Epidemiol* 180:129–139.
37. Soininen P, Kangas AJ, Würtz P, Suna T, Ala-Korpela M (2015): Quantitative serum nuclear magnetic resonance metabolomics in cardiovascular epidemiology and genetics. *Circ Cardiovasc Genet* 8:192–206.
38. Davis KAS, Coleman JRI, Adams M, Allen N, Breen G, Cullen B, *et al.* (2020): Mental health in UK Biobank - Development, implementation and results from an online questionnaire completed by 157 366 participants: A reanalysis. *BJPsych Open* 6:e18.
39. Avila C, Holloway AC, Hahn MK, Morrison KM, Restivo M, Anglin R, Taylor VH (2015): An overview of links between obesity and mental health. *Curr Obes Rep* 4:303–310.
40. Han TS, Lean ME (2016): A clinical perspective of obesity, metabolic syndrome and cardiovascular disease. *JRSM Cardiovasc Dis* 5: 2048004016633371.
41. Hebebrand J, Peters T, Schijven D, Hebebrand M, Grasemann C, Winkler TW, *et al.* (2018): The role of genetic variation of human metabolism for BMI, mental traits and mental disorders. *Mol Metab* 12:1–11.
42. Gerstenberg G, Aoshima T, Fukasawa T, Yoshida K, Takahashi H, Higuchi H, *et al.* (2003): Effects of the CYP 2D6 genotype and cigarette smoking on the steady-state plasma concentrations of fluvoxamine and its major metabolite Fluvoxamine acid in Japanese depressed patients. *Ther Drug Monit* 25:463–468.
43. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, *et al.* (2007): PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 81:559–575.
44. Howard DM, Adams MJ, Shirali M, Clarke TK, Marioni RE, Davies G, *et al.* (2018): Genome-wide association study of depression phenotypes in UK Biobank identifies variants in excitatory synaptic pathways [published correction appears in *Nat Commun* 2021;12:2012]. *Nat Commun* 9:1470.
45. Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM (2010): Robust relationship inference in genome-wide association studies. *Bioinformatics* 26(22):2867–2873.
46. Gallois A, Mefford J, Ko A, Vaysse A, Julienne H, Ala-Korpela M, *et al.* (2019): A comprehensive study of metabolite genetics reveals strong pleiotropy and heterogeneity across time and context. *Nat Commun* 10:4788.
47. Yin X, Chan LS, Bose D, Jackson AU, VandeHaar P, Locke AE, *et al.* (2022): Genome-wide association studies of metabolites in Finnish men identify disease-relevant loci. *Nat Commun* 13:1644.
48. Bulik-Sullivan BK, Loh PR, Finucane HK, Ripke S, Yang J, Schizophrenia Working Group of the Psychiatric Genomics Consortium, *et al.* (2015): LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet* 47:291–295.
49. Bulik-Sullivan B, Finucane HK, Anttila V, Gusev A, Day FR, Loh PR, *et al.* (2015): An atlas of genetic correlations across human diseases and traits. *Nat Genet* 47:1236–1241.
50. 1000 Genomes Project Consortium; Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, *et al.* (2012): An integrated map of genetic variation from 1,092 human genomes. *Nature* 491:56–65.
51. Hemani G, Tilling K, Davey Smith GD (2017): Orienting the causal relationship between imprecisely measured traits using GWAS summary data [published correction appears in *PLoS Genet* 2017;13: e1007149]. *PLoS Genet* 13:e1007081.
52. Hemani G, Zheng J, Elsworth B, Wade KH, Haberland V, Baird D, *et al.* (2018): The MR-Base platform supports systematic causal inference across the human phenome. *eLife* 7:e34408.
53. Watanabe K, Taskesen E, van Bochoven A, Posthuma D (2017): Functional mapping and annotation of genetic associations with FUMA. *Nat Commun* 8:1826.
54. Wallace C, Rotival M, Cooper JD, Rice CM, Yang JHM, McNeill M, *et al.* (2012): Statistical colocalization of monocyte gene expression and genetic risk variants for type 1 diabetes. *Hum Mol Genet* 21: 2815–2824.
55. Burgess S, Small DS, Thompson SG (2017): A review of instrumental variable estimators for Mendelian randomization. *Stat Methods Med Res* 26:2333–2355.
56. GTEx Consortium (2015): Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science* 348:648–660.
57. Hibbeln JR (1998): Fish consumption and major depression. *Lancet* 351:1213.
58. Grosso G, Micek A, Marventano S, Castellano S, Mistretta A, Pajak A, Galvano F (2016): Dietary n-3 PUFA, fish consumption and depression: A systematic review and meta-analysis of observational studies. *J Affect Disord* 205:269–281.
59. Liao Y, Xie B, Zhang H, He Q, Guo L, Subramanieapillai M, *et al.* (2019): Efficacy of omega-3 PUFAs in depression: A meta-analysis [published correction appears in *Transl Psychiatry* 2021;11:465]. *Transl Psychiatry* 9:190.
60. Berger ME, Smesny S, Kim SW, Davey CG, Rice S, Samyay Z, *et al.* (2017): Omega-6 to omega-3 polyunsaturated fatty acid ratio and subsequent mood disorders in young people with at-risk mental states: A 7-year longitudinal study. *Transl Psychiatry* 7:e1220.
61. Wani AL, Bhat SA, Ara A (2015): Omega-3 fatty acids and the treatment of depression: A review of scientific evidence. *Integr Med Res* 4:132–141.
62. Grosso G, Galvano F, Marventano S, Malaguamera M, Bucolo C, Drago F, Caraci F (2014): Omega-3 fatty acids and depression: Scientific evidence and biological mechanisms. *Oxid Med Cell Longev* 2014:313570.
63. Li F, Liu X, Zhang D (2016): Fish consumption and risk of depression: A meta-analysis. *J Epidemiol Community Health* 70:299–304.
64. Ellis FR, Sanders TA (1977): Long chain polyunsaturated fatty acids in endogenous depression. *J Neurol Neurosurg Psychiatry* 40:168–169.
65. Thesing CS, Milanese Y, Bot M, Brouwer IA, Owens M, Hegerl U, *et al.* (2020): Supplementation-induced increase in circulating omega-3 serum levels is not associated with a reduction in depressive symptoms: Results from the MoodFOOD depression prevention trial. *Depress Anxiety* 37:1079–1088.
66. Bot M, Brouwer IA, Roca M, Kohls E, Penninx BWJH, Watkins E, *et al.* (2019): Effect of multinutrient supplementation and food-related behavioral activation therapy on prevention of major depressive disorder among overweight or obese adults with subsyndromal depressive symptoms: The MoodFOOD randomized clinical trial. *JAMA* 321:858–868.
67. Hsu MC, Tung CY, Chen HE (2018): Omega-3 polyunsaturated fatty acid supplementation in prevention and treatment of maternal depression: Putative mechanism and recommendation. *J Affect Disord* 238:47–61.
68. Okereke OI, Vyas CM, Mischoulon D, Chang G, Cook NR, Weinberg A, *et al.* (2021): Effect of long-term supplementation with marine omega-3 fatty acids vs placebo on risk of depression or clinically relevant depressive symptoms and on change in mood scores: A randomized clinical trial. *JAMA* 326:2385–2394.

69. Milaneschi Y, Peyrot WJ, Nivard MG, Mbarek H, Boomsma DI, Penninx BWJH (2019): A role for vitamin D and omega-3 fatty acids in major depression? An exploration using genomics. *Transl Psychiatry* 9:219.
70. Saini RK, Keum YS (2018): Omega-3 and omega-6 polyunsaturated fatty acids: Dietary sources, metabolism, and significance - A review. *Life Sci* 203:255–267.
71. Simopoulos AP (1999): Evolutionary aspects of omega-3 fatty acids in the food supply. *Prostaglandins Leukot Essent Fatty Acids* 60:421–429.
72. Simopoulos AP (2008): The importance of the omega-6/omega-3 fatty acid ratio in cardiovascular disease and other chronic diseases. *Exp Biol Med (Maywood)* 233:674–688.
73. Ameer A, Enroth S, Johansson A, Zaboli G, Igl W, Johansson ACV, *et al.* (2012): Genetic adaptation of fatty-acid metabolism: A human-specific haplotype increasing the biosynthesis of long-chain omega-3 and omega-6 fatty acids. *Am J Hum Genet* 90:809–820.
74. Emery B, Agalliu D, Cahoy JD, Watkins TA, Dugas JC, Mulinyawe SB, *et al.* (2009): Myelin gene regulatory factor is a critical transcriptional regulator required for CNS myelination. *Cell* 138:172–185.
75. Poggi G, Boretius S, Möbius W, Moschny N, Baudewig J, Ruhwedel T, *et al.* (2016): Cortical network dysfunction caused by a subtle defect of myelination. *Glia* 64:2025–2040.
76. Graham DB, Lefkovich A, Deelen P, de Klein N, Varma M, Boroughs A, *et al.* (2016): TMEM258 is a component of the oligosaccharyltransferase complex controlling ER stress and intestinal inflammation. *Cell Rep* 17:2955–2965.
77. Marquardt A, Stöhr H, White K, Weber BHF (2000): cDNA cloning, genomic structure, and chromosomal localization of three members of the human fatty acid desaturase family. *Genomics* 66:175–183.
78. Glaser C, Heinrich J, Koletzko B (2010): Role of FADS1 and FADS2 polymorphisms in polyunsaturated fatty acid metabolism. *Metabolism* 59:993–999.
79. Gromovsky AD, Schugar RC, Brown AL, Helsley RN, Burrows AC, Ferguson D, *et al.* (2018): Δ -5 fatty acid desaturase FADS1 impacts metabolic disease by balancing proinflammatory and proresolving lipid mediators. *Arterioscler Thromb Vasc Biol* 38:218–231.
80. Ikeda M, Takahashi A, Kamatani Y, Okahisa Y, Kunugi H, Mori N, *et al.* (2018): A genome-wide association study identifies two novel susceptibility loci and trans population polygenicity associated with bipolar disorder. *Mol Psychiatry* 23:639–647.
81. Coleman JRI, Gaspar HA, Bryois J; Bipolar Disorder Working Group of the Psychiatric Genomics Consortium; Major Depressive Disorder Working Group of the Psychiatric Genomics Consortium, Breen G (2020): The genetics of the mood disorder spectrum: Genome-wide association analyses of more than 185,000 cases and 439,000 controls. *Biol Psychiatry* 88:169–184.
82. Zuber V, Grinberg NF, Gill D, Manipur I, Slob EAW, Patel A, *et al.* (2022): Combining evidence from Mendelian randomization and colocalization: Review and comparison of approaches. *Am J Hum Genet* 109:767–782.
83. Hemani G, Bowden J, Davey Smith G (2018): Evaluating the potential role of pleiotropy in Mendelian randomization studies. *Hum Mol Genet* 27:R195–R208.
84. Slob EAW, Burgess S (2020): A comparison of robust Mendelian randomization methods using summary data. *Genet Epidemiol* 44:313–329.
85. Mathias RA, Sergeant S, Ruczinski I, Torgerson DG, Hugenschmidt CE, Kubala M, *et al.* (2011): The impact of FADS genetic variants on ω 6 polyunsaturated fatty acid metabolism in African Americans. *BMC Genet* 12:50.
86. Pearson H (2007): Meet the human metabolome. *Nature* 446:8.
87. Wishart DS, Feunang YD, Marcu A, Guo AC, Liang K, Vázquez-Fresno R, *et al.* (2018): HMDB 4.0: The human metabolome database for 2018. *Nucleic Acids Res* 46:D608–D617.
88. Bromet EJ (2008): Cross-national comparisons: Problems in interpretation when studies are based on prevalent cases. *Schizophr Bull* 34:256–257.

3.3 Conclusion

Overall, this chapter presents a comprehensive evaluation of the links between 249 circulating metabolites and lifetime MDD in a large-scale sample ($n = 29,757$), expanding our understanding about metabolomic profiles and potential causality in MDD. Our associational analysis found most metabolites to be associated with MDD ($n = 191$), many of which remained associated when adjusting for common confounders ($n = 129$). In addition, many metabolite measures had a significant genetic correlation with MDD, indicating a shared genetic architecture. Although this indicates that there is a distinct metabolomic profile associated with MDD, it does not inform on potential causality between the two.

Bidirectional MR and colocalisation and found evidence of a causal relationship between decreased omega-3 fatty acids and an increased omega-6 to omega-3 ratio in MDD, which was largely dependent on genetic instruments in the *FADS* cluster. This indicates a potential role of omega-6 and omega-3 balance and respective downstream inflammatory functions in MDD. Furthermore, there was evidence of colocalisation between docosahexaenoic acid (DHA) and MDD in a region within the *FADS* locus ($PP.H4 = 0.89$), indicating a high probability of a single shared causal variant. However, the *FADS* locus has extensive linkage disequilibrium (LD) in this region, it is difficult to decipher the shared causal variant and its associated gene. To investigate this further, I extracted whole blood eQTLs for each gene within the colocalised region (*FADS1*, *FADS2*, *FADS3*, *FEN1*, *MYRF* and *TMEM258*), to test for a potential causal relationship between gene expression and MDD. Although MR found evidence of causality, complementary colocalisation analysis indicated the presence of independent causal variants within the region for gene expression and MDD, a violation of MR assumptions.

This indicates that the shared causal variant underlying MDD and DHA are distinct from those influencing the expression of genes in this region.

In **Chapter 4**, I aim to understand more the role of each gene in the colocalised region in the *FADS* locus with both MDD and omega-6/omega-3 metabolites using rare genetic variants.

4 Rare pathogenic *FADS* variants, Major Depressive Disorder and Omega-3 Fatty Acids

In **Chapter 3**, we identified a putative causal association between various omega-3 related metabolite measures and MDD alongside a high probability of a shared causal signal within the *FADS* locus on Chromosome 11. However, the causal variant and related gene underlying this relationship is unclear due to extensive LD within the *FADS* region. In this chapter, we aim to investigate the relationship of the observed colocalised region in the *FADS* locus with both MDD and the omega-3 related measures more in-depth, specifically by analysing rare genetic variants. Rare genetic variants are typically recent mutations which have not yet been removed by selective pressure and therefore do not usually express LD, which can make their associations easier to interpret ²²³. We hypothesize that if we observe a significant burden of rare, potentially pathogenic variants within a particular gene in MDD, it provides further evidence for its functional role in MDD. Furthermore, if there are consistent signals between a burden of rare pathogenic variants in MDD and metabolite measures, it suggests a role of these metabolites in mediating the association of variants in *FADS* on MDD. First, I extracted rare variants ($AF < 0.01$) from each gene (*FADS1*, *FADS2*, *FADS3*, *FEN1*, *MYRF*, *TMEM258*) and annotated them using various in-silico tools and curated databases to prioritise potentially pathogenic or deleterious variants. I then conducted a series of aggregate-variant tests (AVT) using rare, prioritised variants in each gene with lifetime MDD and omega-3 related metabolite measures. Each of these methodologies is in detail in Section 4.1. Then, the Chapter findings are reported in the same way as a manuscript, with an introduction (Section 4.2),

(a shorter) methods (Section 4.3), results (Section 4.4) and a discussion (Section 4.5).

Finally, Section 4.6 will broadly summarise the findings from this Chapter.

Code for this chapter can be found in a GitHub repository:

(https://github.com/Elladavyson/WGS_FADS), supplementary figures and tables for this chapter can be found in Appendix (**Chapter 10**) and the supplementary data files are hosted on Zenodo record (<https://zenodo.org/records/15101441>).

4.1 In-depth Methods

4.1.1 *Variant annotation*

Variants can be annotated with respect to biological, functional and clinical information, and is a critical step in prioritising variants with a potentially pathogenic effect. In these analyses I used a variety of variant annotation tools to prioritise rare variants in *FADS* for subsequent AVT. The information provided by each tool is summarised below.

4.1.1.1 Variant Effect Predictor (VEP)

The variant effect predictor (VEP) toolset³⁸⁸ was developed and is hosted by Ensembl³⁸⁹, a platform which hosts various resources and databases for genomic annotation for various organisms. The VEP returns detailed information regarding genetic variants effect on different transcripts, proteins or regulatory regions. Specifically, VEP utilises the gene annotation data from GENCODE^{390,391} to predict the consequence, defined by sequence ontology terms³⁹², of a genetic variant on all possible transcript isoforms.

Ensembl has also established corresponding “IMPACT” scores (High/Moderate/Low/Modifier) for each consequence term which were used to prioritise variants (Table 4.1). For each gene, I used the VEP annotations corresponding to the canonical transcript, which is defined by Ensembl as the most representative transcript based on its conservation and expression patterns³⁸⁹.

Table 4.1 The Variant Effect Predictor (VEP) consequence terms and associated “IMPACT” term.

For a genetic variant supplied to VEP, its consequence (defined using Sequence Ontology terms) on all overlapping transcripts is assessed using a rule-based approach from the

GENCODE dataset. Each annotated consequence has an assigned impact score. I prioritised variants in **Chapter 4** which had an annotated consequence with a HIGH impact score. UTR = Untranslated region, miRNA = microRNA, TFBS = Transcription factor binding site, TF= Transcription factor.

IMPACT	Consequence (display term)
HIGH	Transcript ablation, Splice acceptor variant, Splice donor variant, stop gained, Frameshift variant, Stop lost, Start lost, Transcript amplification, Feature elongation, Feature truncation
MODERATE	Inframe insertion, Inframe deletion, Missense variant, Protein altering variant
LOW	Splice donor 5th base variant, Splice region variant, Splice donor region variant, Splice polypyrimidine tract variant, Incomplete terminal codon variant, Start retained variant, Stop retained variant, Synonymous variant
MODIFIER	Coding sequence variant, Mature miRNA variant, 5 prime UTR variant, 3 prime UTR variant, Non coding transcript exon variant, Intron variant, NMD transcript variant, Non coding transcript variant, Coding transcript variant, Upstream gene variant, Downstream gene variant, TFBS ablation, TFBS amplification, TF binding site variant, Regulatory region ablation, Regulatory region amplification, Regulatory region variant, Intergenic variant, Sequence variant

4.1.1.2 Loss-of-function transcript effect estimator

Putative loss-of-function (pLoF) are those which are predicted to result in complete loss of functional gene and its downstream products, such as stop-gained, frame-shift or splice-altering variants³⁹³. Identifying pLoF variants is of particular clinical interest due to their well-defined causal role in various mendelian diseases³⁹⁴. However, recent large scale genome sequencing projects have revealed that healthy individuals may carry up to 800 pLoF variants, illustrating that the human genome is relatively robust to pLoF currently identified in the literature³⁹⁵. There are various possibilities for pLoF variants to be found in

healthy individuals including: a severe recessive allele in a heterozygous state, true LoF which have a less deleterious impact, benign LoF variants in redundant genes, or sequencing or annotation artefacts³⁹⁵. True LoF variants with large effects are typically kept to a low frequency due to negative selection, resulting in conserved genomic regions depleted of genomic variation. In contrast, sequencing and annotation error rates are uniformly distributed across the genome. Together, this means pLoF variants are enriched for technical artefacts and many may be false positives (Figure 4.1)^{373,395,396}.

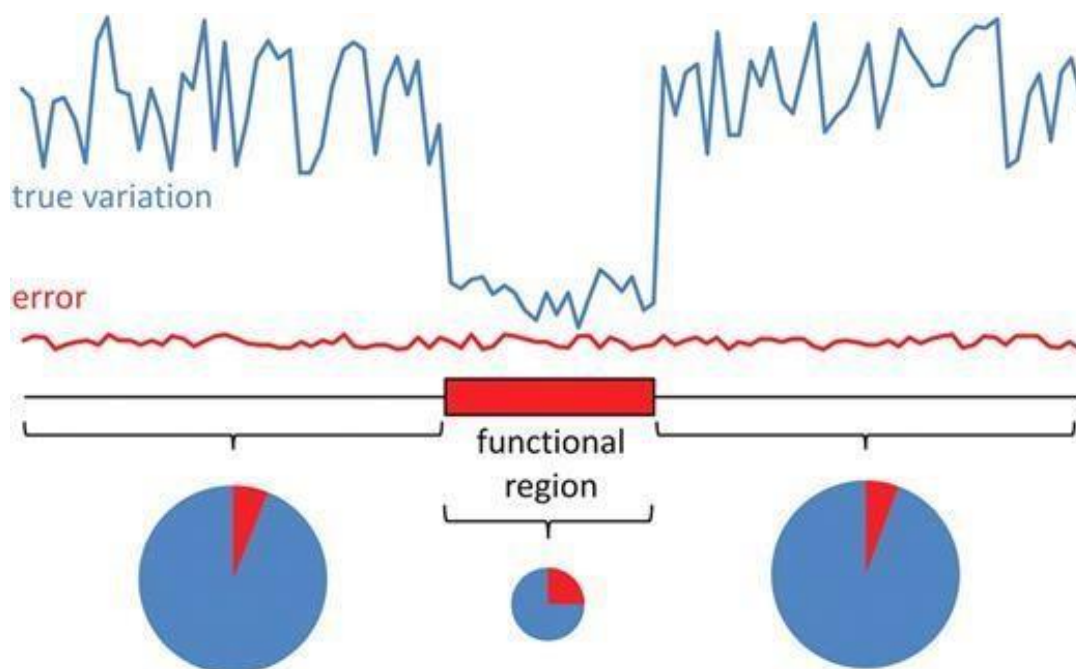


Figure 4.1 Enrichment of error in pLoF variants in functional regions.

A schematic plot showing levels of true sequence variation (blue), with sequencing error (red). Pie charts show the proportion of sequence variation (circle size) due to true sequencing variation (blue) or error variation (from sequencing/annotations). True sequence variation is suppressed in functional regions due to negative selection, but the error rate is approximately uniform across all regions. Therefore, functional regions show

a higher proportion of variation due to error (enrichment) rather than true sequence variation. This Figure was taken from MacArthur & Tyler Smith (2010)³⁹³. Copyright: *Human Molecular Genetics*.

The Loss-Of-Function transcript effect estimator (LOFTEE) is a VEP plugin (<https://github.com/konradjk/loftee>) which can annotate pLoF variants (currently limited to stop-gained, splice-site disrupting or frameshift variants) which are likely to reflect true effects rather than resulting from technical errors. Specifically, LOFTEE uses a set of stringent filters (Table 4.2) to filter out low confidence pLoFs (e.g. terminal truncating or splice variants which are predicted to escape nonsense mediated decay and produce a functional product^{373,397}). In addition, LOFTEE can identify other pLoF variants at non-canonical splice sites (annotated as “Other Splice”).

Table 4.2 Putative loss of function variants filtered out by LOFTEE.

The loss-of-function transcript effect estimator (LOFTEE) identifies high confidence variants by removing those which may be false positives through annotation or technical errors. The filters applied by LOFTEE are shown in this table. UTRs = Untranslated regions, MaxEntScan = Maximum Entropy Scan.

Variant type	“Low Confidence” pLoF variants identified by LOFTEE
Stopgained/frameshift variants	Variants near the end of the transcript (< 50bp), with lower Genomic Evolutionary Profiling (GERP) scores. Full details can be found at <i>Karczewski et al, 2020 Supplement</i> ³⁷³
Splice-site variants	Variants only affecting splicing of UTRs Variants not predicted to affect a donor site Variants with no effect on splicing predicted by MaxEntScan ³⁹⁸ Variants that are “rescued” by nearby, inframe splice sites
All variants	Variants where the purported LoF allele is the ancestral allele (in primates).

Overall LOFTEE is a powerful tool which can be used to identify pLoF variants which have a high confidence of reflecting a true effect. However, LOFTEE is a conservative tool and therefore may result in decreased sensitivity. Also, LOFTEE does not evaluate potential deleterious effects of missense variants or non-coding regulatory variants³⁷³. Therefore, I have also incorporated additional variant annotations from other VEP plugins to evaluate potential pathogenicity of variants within these classes.

4.1.1.3 Rare exon ensemble learner

Missense variants are a type of non-synonymous SNVs in coding regions which result in an amino acid substitution within the translated protein sequence³⁹⁹. Missense variants may change the function of protein products and have downstream implications on a range of biological processes. Therefore, missense mutations can have a large impact on biological function and have an important role in human health. There are many tools

which can be used to predict the functional effect and pathogenicity of missense variants, based on the conservation of nucleotides and amino acids, alongside biochemical properties of the amino acid substitution itself (Table 4.3). These typically involve the incorporation of evolutionary conservation patterns to identify highly conserved DNA or amino-acid positions which are then more likely to have a deleterious functional impact. However, individual tools often present contradictory predictions, in part due to the use of different specific predictive features, parameters and models. Furthermore, many of the commonly used tools are sub-optimal at characterising the pathogenicity of rare variants, with some explicitly excluding variants with an MAF < 0.5% from training data³⁹⁹. The rare exome ensemble learner (REVEL) is an ensemble method combining various individual predictive tools (Table 4.3) using a random-forest algorithm and outputs a REVEL score per variant on a probability scale ranging from 0 (benign) to 1 (pathogenic)³⁹⁹.

Table 4.3 Individual prediction tools which are included in the rare ensemble exome learner (REVEL).

The name, reference and brief description of each of the individual prediction tools which were used in a random-forest ensemble learner, REVEL, to identify potentially pathogenic rare missense variants.

Tool	Full Name	Description
MutPred ⁴⁰⁰	Mutation Prediction	Models the changes to protein sequence (including structural features and functional binding sites) between wild-type and mutant sequences, giving a probability of a loss or gain of function per genetic variant.
FATHMM v2.3 ⁴⁰¹	Functional Analysis through Hidden Markov Models	Predicts the functional effects of protein missense mutations by combining sequence conservation patterns across species in hidden Markov models (HMM) to generate pathogenicity weights, representing the tolerance of the protein/domain to mutations.
VEST 3.0 ⁴⁰²	Variant Effect Scoring Tool	Supervised machine learning (random forest) approach trained on various features of known pathogenic / benign variants in the Human Gene Mutation Database (HGMD 2012v2). Results in a probability score of pathogenicity per variant (0 to 1)
PolyPhen-2 ⁴⁰³	Polymorphism Phenotyping v2	Prediction of an amino acid substitution on the structure and function of a human protein. Uses various sequence, phylogenetic, and structural features to predict probability (naïve bayes classifier) of damaging impact on protein structure/function (from 0 to 1).
PROVEAN ⁴⁰⁴	Protein Variation Effect Analyzer	Predicts whether an amino acid substitution or indel has an impact on biological function of a protein. Uses a alignment-based score to predict damaging effects of variants. More disruptive variants have a lower score. Typical threshold for deleterious is <2.5.
MutationAssessor ⁴⁰⁵	-	A conservation-based approach which uses multiple sequence alignments from diverse species to assess conservation and scores variant based on functional impact. High impact usually from a score threshold > 3.5.

MutationTaster ⁴⁰⁶	-	Predicts disease potential of a genetic variant, using information concerning evolutionarily conservation, splice-site changes, allele frequencies, loss of protein features and changes influencing mRNA expression in a naïve bayes classifier. Trained on known disease mutations from the HGMD and benign mutations in the 1000 Genomes Project (TGP). Classifies variants as disease-causing or polymorphism.
SIFT ⁴⁰⁷	Sorting Intolerant from Tolerant	Uses a sequence-homology based approach to assess the probability of intolerance to amino acid substitutions throughout protein structure. Assumes highly conserved regions are more intolerant to mutations. Using a probability cut-off then classifies variants as Tolerated or Deleterious.
GERP++ ⁴⁰⁸	Genomic Evolutionary Rate Profiling ++	Identifies constrained elements of the genome. Compares orthologous sequences from difference species to identify positions which show fewer substitutions than would be expected under neutral evolution. Positive GERP++ score indicates evidence of constraint.
SiPhy ⁴⁰⁹	Site-specific Phylogenetic Analysis	Identifies constrained genomic regions through analysing patterns of conservation and substitution rates across different species. Calculates a constraint score (strong conservation) and a Base Preference Score to identify substitution preferences.
phyloP ⁴¹⁰	Phylogenetic P-value	Uses multiple sequence alignment from multiple species to identify the evolutionary rate of each nucleotide position. Identifies sites which have a conserved or accelerated conservation rate compared to a neutral evolutionary model (score of 0). Positive scores indicate conservation, and negative scores indicate accelerated rates (positive selection).

phastCons ⁴¹¹	Phylogenetic Hidden Markov Model for Conservation	Uses an HMM on multiple-sequence alignments from 46 species to predict the probability that each nucleotide belongs to a conserved element. Ranges from 0 to 1 (high probability nucleotide is conserved). Considers the conservation of neighbouring nucleotides in its predictions.

REVEL was trained on a putative rare (AF between 0.1% and 1%) neutral and disease missense variants obtained from the Human Gene Mutation Database (HGMD), explicitly excluding variants which were also used to its constituent predictive tools.

REVEL was tested in two independent test sets (SwissVar and ClinVar) and showed improved accuracy of the classification of pathogenic missense variants in comparison with individual tools and other ensemble methods, especially for rarer variants (AF < 3%). Therefore, REVEL represents a useful tool for assessing the potential deleterious effects of rare missense variants. In this thesis, I used the precomputed REVEL scores which are available via the VEP Plugin or can be accessed online (<https://sites.google.com/site/revelgenomics/downloads>).

4.1.1.4 Combined annotation dependent deletion tool

The combined annotation dependent deletion (CADD) tool scores the predicted deleteriousness of SNVs, multi-nucleotide substitutions and indels⁴¹². Importantly, CADD scores are also calculated from the integration of multiple diverse annotation tools and predicts the effects of coding and non-coding variants⁴¹³. CADD scores are based upon the assumption that most genetic variations which have stayed in humans since becoming a distinct species (speciation) are neutral as they have not been eliminated by natural selection⁴¹⁴. Using this logic, CADD scores are calculated by comparing a proxy-neutral set of variants (N = 14.7 million), defined as those present in > 95% of humans but not in the human-ape ancestor, to a proxy-deleterious set of variants⁴¹². The proxy-

set of deleterious variants are a set of *de novo* variants simulated from the proxy-neutral set⁴¹². More than 60 different annotation tools are used to annotate the variants in each set, including conservation scores (PhyloP⁴¹⁰, PhastCons⁴¹¹, GERP++⁴⁰⁸), regulatory information (DNase hypersensitivity) and transcript information (exon-intron boundaries, expression in cell lines and a range of protein-level scores).

Using the myriad of annotation features, a logistic regression model is trained to classify variants in the proxy-neutral or proxy-deleterious set⁴¹². The model is then applied to all ~ 9 billion potential SNVs of the human reference genome GRCh38, to generate a resource of CADD scores for use in the scientific community. The raw CADD score output from the model is an arbitrary unit but can be scaled using a PHRED conversion table, derived by relative ranking of model scores across all other potential SNVs. Specifically, scaled CADD scores range from 1 (low deleteriousness) to 99 (high deleteriousness) based on the rank of each variant relative to all other possible substitutions in the human genome⁴¹³. A general rule of thumb is that scaled CADD scores > 20 and > 30 represent the top 1 and 0.1% most deleterious substitutions in the genome respectively. There have been multiple versions of CADD iteratively released, which include more annotations as technologies progress and datasets accrue. In this thesis, I used precomputed CADD scores from the most recent release (v1.7)⁴¹⁴ which incorporates annotations from several deep-learning methods to improve prediction of a variants' functional impact on protein structure and improve prediction of impact of variants in non-coding regulatory regions.

4.1.2 *Aggregate-variant testing*

There are not usually many carriers of rare pathogenic variants within a dataset. Therefore, the GWAS approach of testing the association of each singular variant with a trait is significantly underpowered for rare variants, unless the sample size or variant effect is

large⁴¹⁵. Aggregate-variant testing (AVT) can boost power in rare-variant association analysis, specifically by evaluating the association of multiple variants in a biologically relevant region (e.g. a gene) together in aggregate (rather than each individually). These analyses used REGENIE⁴¹⁶ software to run several different types of AVT, each of which has distinct advantages and assumptions under specific genetic architectures (Figure 4.2) and are described in detail below. Furthermore, each gene had AVT run on different prespecified sets (or “masks”) of rare pathogenic variants based off different variant annotations. For all methods, the weights for each rare variant j were set by the beta distribution (Equation 4.1), with shape parameters (MAF,1,25) as proposed by Wu et al, (2016)⁴¹⁷. This distribution (Figure 4.3) results in higher weights being assigned to rarer variants.

$$w_j = \text{Beta}(\text{MAF}_j, 1, 25) \quad \text{Equation 4.1}$$

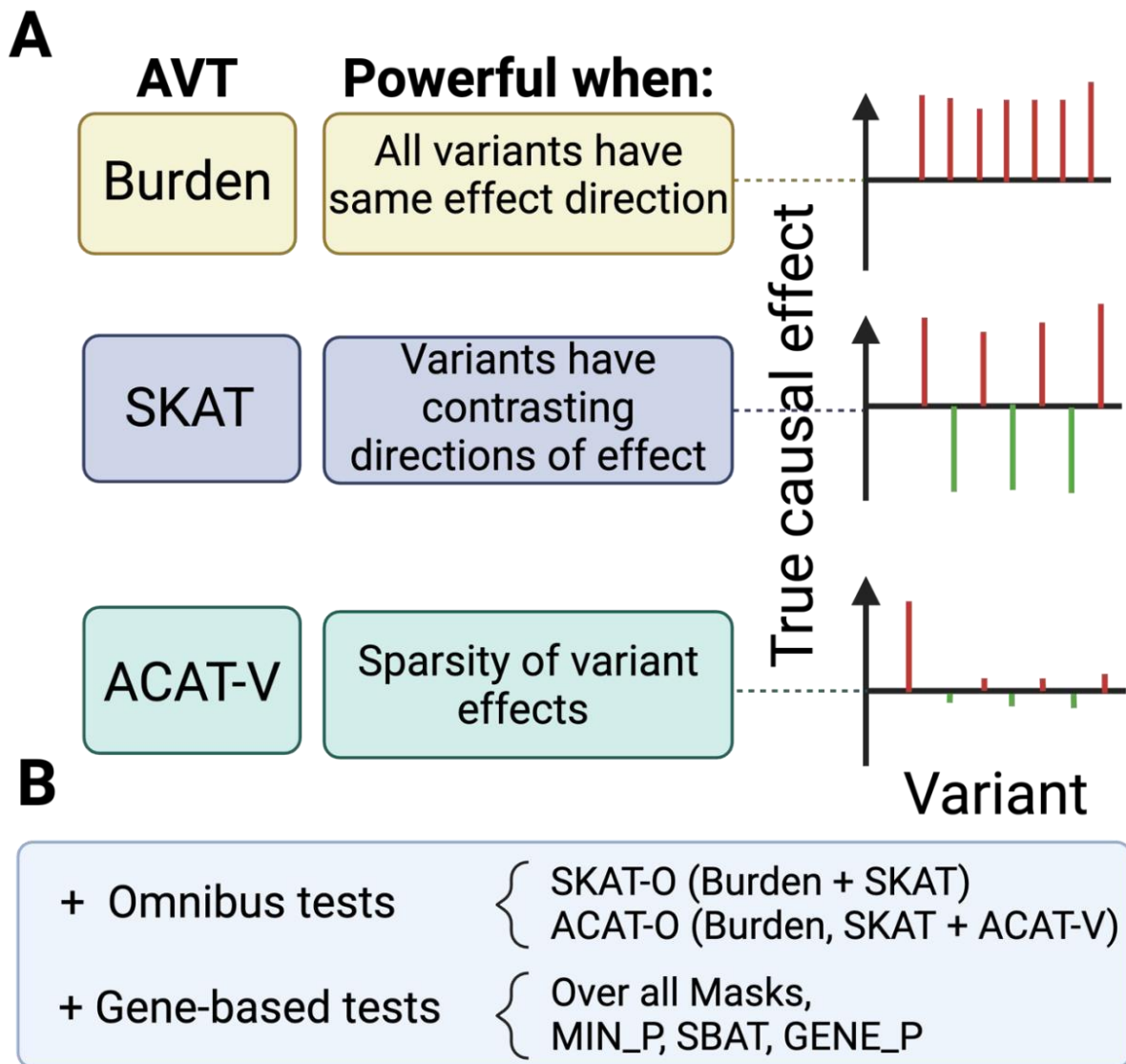


Figure 4.2 Different aggregate variant tests (AVT) employed.

A) The different AVT employed for each mask of rare variants (“AVT”) and the scenario in which this test is most powerful to detect associations between rare variants and a trait (“Powerful when”). The genetic architecture for each scenario is shown in schematics for each test (right), where the x axis reflects individual rare variants included in a mask and the y axis represents the true causal effect of each variant. B) Additional omnibus AVT tests (combinations of the individual tests shown in A), and gene-based tests which combines evidence for each mask per gene. This Figure was made in Biorender.com.

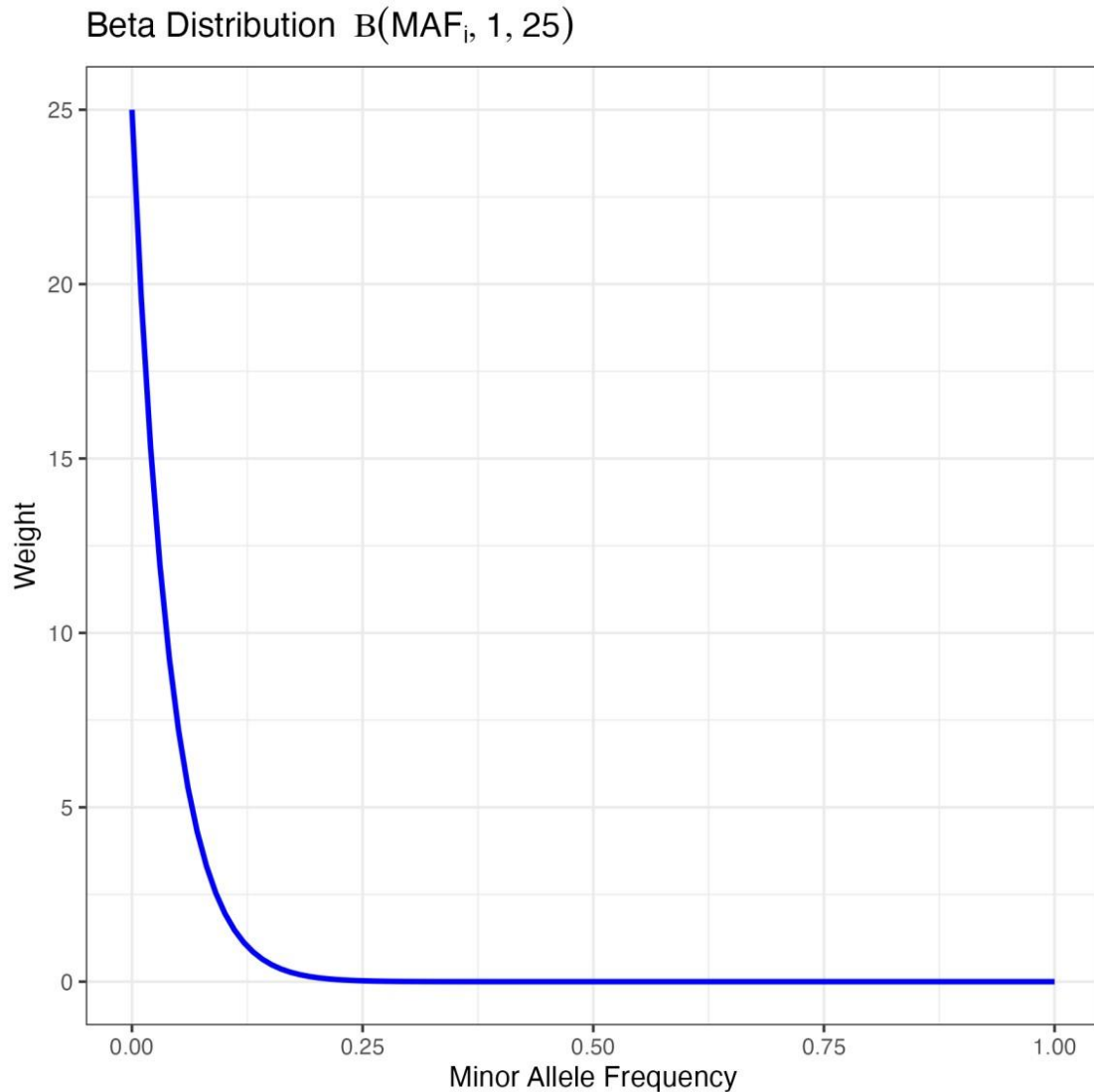


Figure 4.3 Beta distribution with shape (MAF, 1, 25) for assigning weights to rare variants in aggregate variant testing (AVT).

The distribution was used to select weights for calculating the pseudo-genotype in the burden and SKAT tests. This results in alleles with a lower frequency being assigned a higher weight. MAF = Minor allele frequency, SKAT = sequence-kernel association test.

4.1.2.1 Traditional burden test

The traditional burden test⁴¹⁵ calculates an aggregated pseudo-genotype score from selected rare variants and tests its association with a phenotype as in a GWAS.

Specifically, the pseudo-genotype score for the i^{th} individual (C_i) is calculated as a weighted sum of the genotypes (G_{ij}) for m rare variants (Equation 4.2). We used an additive genetic model, whereby the genotypes were 0,1 or 2 depending on the observed count of the minor allele. The burden test tests the null hypothesis $H_0: \beta = 0$ in a simplified regression model, utilizing the pseudo-genotype (Equation 4.3).

$$C_i = \sum_{j=1}^m w_j G_{ij} \quad \text{Equation 4.2}$$

$$h(\mu_i) = \alpha_0 + \alpha' X_i + \beta C_i \quad \text{Equation 4.3}$$

The test statistic under the null hypothesis (Q_{BURDEN}) can be expressed as the square of the weighted sum of score statistics for each variant (S_j) (Equation 4.4) which are derived from the marginal model (Equation 4.5).

$$Q_{BURDEN} = \left(\sum_{j=1}^m w_j S_j \right)^2 \quad \text{Equation 4.4}$$

$$S_j = \sum_{i=1}^n G_{ij} (y_i - \hat{\mu}_i) \quad \text{Equation 4.5}$$

The significance of the burden test can then be obtained by comparison to a chi-square distribution with 1 degree of freedom. The burden test assumes that all variants have the same effect direction and roughly the same magnitude of effect⁴¹⁵. Therefore, the test loses power when variants which are neutral, or which have opposing effects are included in the pseudo genotype score.

4.1.2.2 Sequence kernel association test

The sequence kernel association test (SKAT)⁴¹⁷ evaluates the distribution of the genetic effect estimates from the standard linear model for single variant association tests with the null hypothesis $H_0: \beta = 0$. The SKAT test uses the assumption that the effect estimates for each individual variant (β_j) follow an arbitrary distribution with a mean of 0 and variance of $w_j\tau$, determined by the individual variants' weight (w_j) and a variance component (τ). This assumption makes the null hypothesis equivalent to $H_0: \tau = 0$, which can be tested with a variance-component score test in the mixed model shown in Equation 4.6.

$$h(\mu_i) = \alpha_0 + \alpha'X_i + \beta'G_i \quad \text{Equation 4.6}$$

The test-statistic for the SKAT test (Q_{SKAT}) can be expressed as the weighted sum of squares of single-variant score statistics (S_j) (Equation 4.7).

$$Q_{SKAT} = \sum_{j=1}^m w_j^2 S_j^2 \quad \text{Equation 4.7}$$

Under the null, Q follows a mixture of chi-square distributions (1 degree of freedom) which can be used to obtain a p-value using Davies' exact method⁴¹⁸. If the Davies exact method

fails to converge, a p-value can also be estimated using Kuonen's saddlepoint approximation method⁴¹⁹, which gives a close estimate of the Davies' method⁴¹⁷. The square of the score-statistic makes SKAT more powerful than burden tests when including variants with opposing directions of effects^{415,417}.

4.1.2.3 Sequence kernel association test omnibus test

The omnibus test SKAT-O⁴²⁰ combines the burden test statistic and SKAT test-statistic with different proportions (ρ) (Equation 4.8). In REGENIE, SKAT-O calculates a test-statistic for 8 different values of ρ (0, 0.1², 0.2², 0.3², 0.4², 0.5², 0.5, 1) and selects the value of ρ to maximise the power using numerical integration.

$$Q_{\rho-SKAT-O} = \rho Q_{BURDEN} + (1 - \rho) Q_{SKAT}, \quad 0 \leq \rho \leq 1 \quad \text{Equation 4.8}$$

4.1.2.4 Aggregated Cauchy combination test (ACAT-V)

Both the burden and the SKAT test assume that all variants within the set are causal and lose power under sparse alternatives (i.e. when there is only a small fraction of variants within a mask that exhibit a true effect)⁴²¹. In these instances, the aggregated Cauchy combination (ACAT)-V test⁴²² is more powerful. ACAT is a general approach used to combine p values, which transforms them to Cauchy variables and uses the Cauchy distribution to evaluate the significance. The test statistic is calculated as the weighted sum of transformed p-values (Equation 4.9).

$$T_{ACAT-V} = \sum_{j=1}^m w_{j,ACAT-V}^2 \tan\{(0.5 - p_j)\pi\} \quad \text{Equation 4.9}$$

4.1.2.5 Aggregated Cauchy-combination omnibus test

Finally, we also employed an omnibus test, ACAT-O⁴²² which runs ACAT on the p-values obtained from the Burden test, SKAT, and ACAT-V test (Equation 4.10), which has been shown to be robust under different genetic architectures and sparsity of causal variants within the variant set.

$$T_{ACAT-O} = \frac{1}{3} \tan\{\pi(0.5 - p_{ACAT-V})\} + \frac{1}{3} \tan\{\pi(0.5 - p_{BURDEN})\} + \frac{1}{3} \tan\{\pi(0.5 - p_{SKAT})\} \quad \text{Equation 4.10}$$

4.1.2.6 Joint Aggregated Variant Tests (across all masks)

The AVT tests described so far test the association of a set of rare pathogenic variants (“masks”) in a gene. In practice, these are typically run on various masks per gene defined using different variant annotations and allele frequencies. The REGENIE software also implements various joint-burden tests to provide a gene-level measure of the association across all different masks tested. The most basic of these, referred to as MIN_P, assigns the significance of the burden of that gene as the minimum P-value from the traditional burden test for any mask tested per gene. The sparse burden association test (SBAT)⁴²³ jointly models a set of burden scores using a non-negative least square and was used for the metabolomic traits. This approach is effective in accounting for the correlation between burden scores which can arise when using nested masks, such as the ones implemented in this work. Additionally, the solution in SBAT introduces sparsity which can aid interpretation of which mask is more likely to include causal variants. However, SBAT relies on the assumption that the burden tests for each mask exhibit the same direction of association and has currently only been developed for quantitative traits. Finally, a unified

gene testing strategy, recommended by REGENIE, called GENE_P⁴²³ was used. This test uses ACAT to combine the p-values from individual masks from the SKAT-O, ACAT-V, Burden and SBAT (when applicable) tests to provide an overall assessment of significance for the gene.

4.2 Introduction

The relationship between MDD and circulating long-chain polyunsaturated fatty acids (LC-PUFAs) is not fully understood. Historically, there has been much focus on omega-3 and omega-6 fatty acids in MDD, following observational correlations between dietary omega-3 intake and MDD^{424,425}. However, it has been difficult to establish whether this is a causal relationship and if so, the direction of causality. Indeed, various RCT on the potential benefits of dietary omega-3 supplements on MDD have found conflicting results^{426–431}, which may be due to heterogeneity in design, including different baseline populations, the amount/type of omega-3 supplementation, definitions of MDD and antidepressant usage⁴³⁰. Recent technological advances in methods which enable affordable, high-throughput and accurate measurement of hundreds of circulating metabolites provide an opportunity to examine this relationship with depth and precision^{291,293}. Additionally, in conjunction with genotyping and sequencing technology, the genetic influences on metabolite levels can be elucidated^{432–439}. This also enables the identification of genetic instruments and the use of causal frameworks, such as mendelian randomization (MR) and colocalisation to examine the potential causality between metabolite levels, including LC-PUFA, and MDD or vice versa in large population studies^{436,440–443}.

Using this approach, I have previously identified a potentially causal relationship between decreased omega-3 fatty acids and increased omega-6: omega-3 ratio in MDD, which was largely dependent on genetic instruments in the *FADS* cluster (**Chapter 3**).

Furthermore, we found evidence of colocalisation between an omega-3 fatty acid, DHA, and MDD in a region within the *FADS* locus, indicating a high probability of a single shared causal variant. Since the publication of our findings, various other studies have also established an altered lipid profile and evidence of a role of the *FADS* locus in

MDD^{307,308,310,311,444}. Analysis of metabolites in the Netherlands Study of Depression and Anxiety (NESDA) established the downregulation of long-chain monounsaturated fatty acids and the upregulation of lysophospholipids with MDD at different time points, which were largely independent of BMI and antidepressant medications³¹¹. Furthermore, MR analysis identified a causal relationship with lysophospholipids with MDD, with the top genetic instrument also being in the *FADS* locus. Due to the observed pleiotropy of this locus with many measured metabolites, the authors argue for a broader role of *FADS* on MDD across various metabolomic traits. The *FADS* locus also appears to be underlying the putatively causal association of many metabolites with other neuropsychiatric disorders⁴⁴². This was demonstrated by a recent MR-based study, which investigated the role of circulating metabolites across different neuropsychiatric and neurodegenerative diseases and found 30 metabolite-trait pairs which colocalized in the *FADS* cluster. Specifically, they found consistent evidence of a protective role of arachidonic acid (AA)-containing metabolites whilst linoleic acid (LA)-containing metabolites were associated with increased susceptibility of various neuropsychiatric disorders, including MDD, bipolar disorder (BPD) and SCZ and anxiety disorder (AD).

The pleiotropy of the *FADS* locus can make it difficult to conclude as to whether it is actually AA and LA-containing metabolites which are mediating this effect, or other metabolomic measures which are associated with this locus⁴⁴⁵. However, it should be noted that the associations between AA- and LA-containing metabolites persisted when restricting to metabolites which were not driven by variants in the *FADS* cluster, providing evidence for this association unbiased by the pleiotropy in *FADS*⁴⁴². Another large-scale MR analysis utilizing GWAS summary statistics of > 900 metabolite measures also found significant association between lower levels of AA-containing metabolites alongside

higher levels of LA-containing metabolites in BPD, which were also colocalized at the *FADS* locus⁴⁴³. These associations were also seen for SCZ and MDD, although they are less pronounced and only nominally significant. Interestingly, this study directly investigated whether these associations were being driven by the pleiotropic nature of the *FADS* locus. Specifically, they tested the association of other lipid fractions with the same genetic instrument in the *FADS* cluster with BPD. As they observed no significant causal relationship between these measures and MDD, they hypothesize the *FADS* association with BPD to be mediated by proximal downstream pathways, namely AA synthesis from LA through *FADS1* and *FADS2* enzymes⁴⁴³. Finally, another MR study investigating omega-3 and omega-6 measures also found evidence of a significant protective effect of omega-3 fatty acids on lifetime and recurrent MDD, which also colocalised at *FADS*⁴⁴⁶. This was also seen when using multivariate MR with other genetically correlated lipids, to mitigate potential bias from horizontal pleiotropy⁴⁴⁷. Furthermore, this method demonstrated stronger effect for certain omega-3 fatty acids, namely, eicosapentanoic acid (EPA) over DHA on both MDD outcomes⁴⁴⁶.

Overall, various studies have utilized MR and found a putatively causal relationship between different psychiatric disorders and various omega-3 and omega-6 related metabolites^{307,310,311,439,441,442,446}. Colocalisation analyses have complemented these findings by indicating that there is a single shared causal variant underlying these associations in the *FADS* region, which does not appear to be related to the expression levels of nearby genes^{441,442}. However, it should be noted that whilst this is indicative of causality, accurate inferences from colocalisation relies on the assumption that there is only one causal variant per region³⁸⁶. Furthermore, the single shared causal variant underlying this relationship is difficult to conclusively identify due to the high LD between common genetic variants within

this region⁴⁴⁸. The analysis of rare variants in specific genes can complement findings from GWAS whilst also providing novel insight into disease mechanisms^{223,449}. Rare variants associated with complex traits typically have been found to have large effect estimates^{450–452} and they do not typically exhibit high levels of LD with surrounding variants, making observed associations easier to interpret. As the imputation of rare variants using reference panels is inaccurate^{217,453}, the reliable analysis of rare variation relies on sequencing data. Initial high costs limited analyses to the study of rare variants in candidate genes and/or in small sample sizes^{454,455}. However, like metabolomic technology, there have been advances in next generation sequencing, which have made it possible to apply whole-genome sequencing on largescale cohorts, such as UKB²²⁶.

In this analysis we aimed to investigate the association between the rare potentially pathogenic genetic variants in the *FADS* cluster with both omega-3/omega-6 metabolites and MDD using the whole-genome sequencing data in ~450,000 individuals in UKB²²⁶.

We focused on the six genes located within the colocalised region in

Chapter 3 (*FADS1*, *FADS2*, *FADS3*, *FEN1*, *MYRF* and *TMEM258*), alongside five metabolite measures (omega-3 fatty acids, omega-3 fatty acids to total fatty acids %, DHA, omega-6: omega-3 ratio and the degree of unsaturation) which had significant evidence of a causal relationship on MDD in **Chapter 3**. This enables us to identify potential causal role of the genes in the *FADS* cluster with MDD, and whether this is consistent with changes in omega-3 / omega-6 related measures. This analysis consists of two central parts, first identifying potentially pathogenic or deleterious rare variants in the *FADS* genes, and secondly, testing the burden of these variants on MDD and each metabolomic trait. We prioritized rare variants within each gene as potentially pathogenic or deleterious using various in-silico tools and curated databases. We iteratively placed variants into different groups, referred to as ‘masks’, depending on their variant annotations and allele

frequencies. Then, we comprehensively tested the burden of prioritised rare variants for each gene/mask with MDD and each metabolite measure using traditional statistical tests alongside a range of AVT, each designed for a different genetic architecture. A schematic of methodology for this Chapter is shown in Figure 4.4.

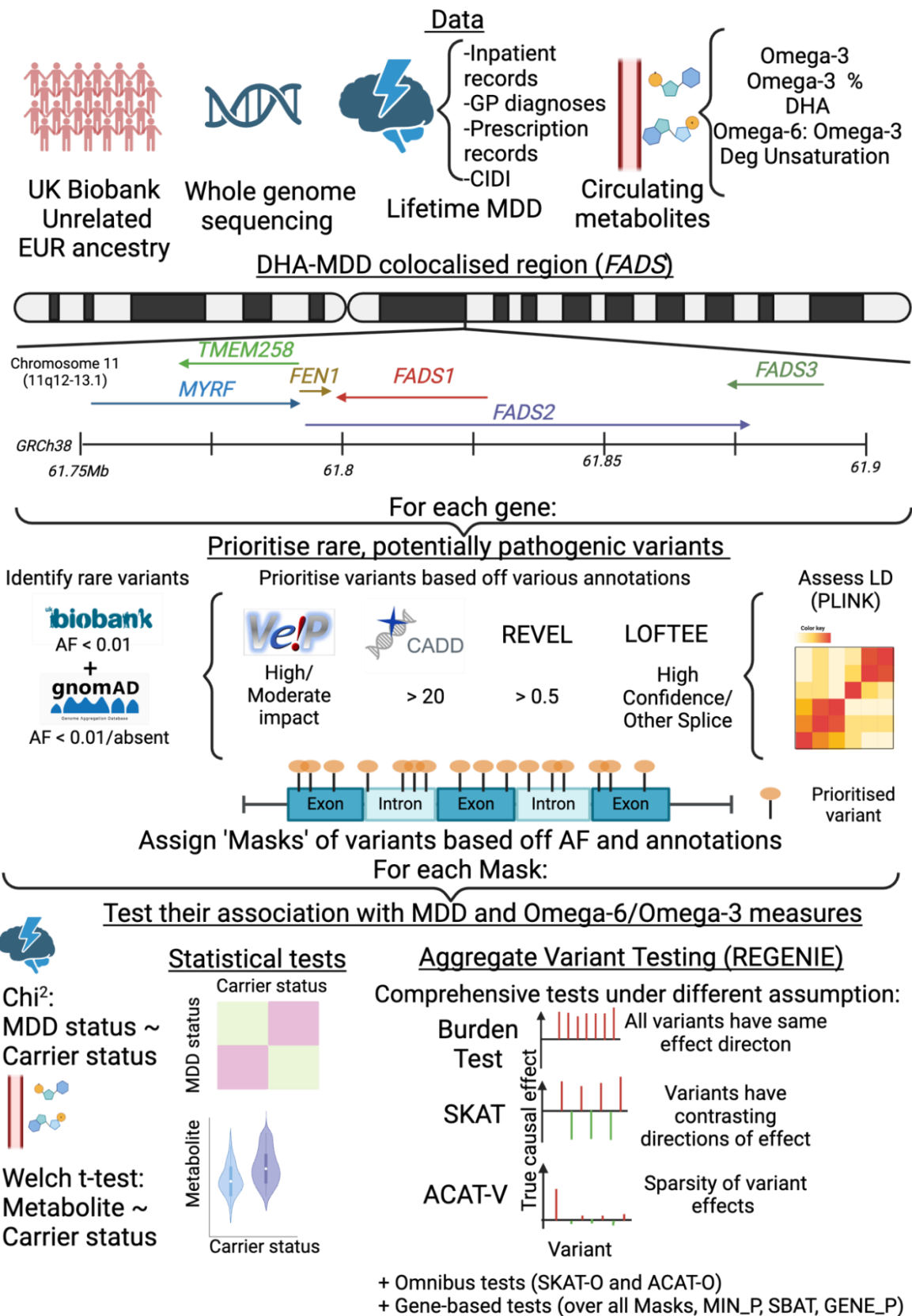


Figure 4.4 Schematic of the analysis presented in Chapter 4.

This chapter focuses on participants in the UK Biobank, specifically utilizing whole-genome sequencing data, an agglomerate measure of lifetime MDD from a range of clinical/questionnaire sources and omega-3 and omega-6 related measures obtained using NMR metabolomics technology. The analysis specifically focused on six genes (*FADS1*, *FADS2*, *FADS3*, *FEN1*, *MYRF* and *TMEM258*) which lie within the colocalised region between docosahexaenoic acid and MDD, observed in **Chapter 3**. For each gene, we first identified rare genetic variants as those with an Allele Frequency (AF) < 0.01 in UKB alongside either being absent or having an AF < 0.01 in gnomAD. Rare variants were then further prioritized as potentially pathogenic or deleterious if they met any of the following criteria: a predicted high or moderate impact from VEP, a CADD score > 20, a REVEL score > 0.5 or a LOFTEE classification of “High Confidence” or “Other splice”. The correlation between prioritized variants in each gene was then assessed using PLINK. Prioritised variants in each gene were assigned to different subgroups, across variant annotation classes and AF bins, referred to as ‘masks’. Each gene/mask combination was tested for an association with MDD and each metabolite trait using both traditional statistical tests alongside a sophisticated aggregate-variant testing (AVT) framework implemented in REGENIE. Statistical tests used include the chi-squared test (for MDD) and Welch’s two-tailed t-test (for each metabolite trait). Various types of AVT were used, including the burden, SKAT and ACAT-V tests, which each have distinct assumptions about the underlying genetic architecture of the prioritized variants. Omnibus tests, such as SKAT-O and ACAT-O were also implemented for each mask. Finally, gene-based tests (across all masks per gene) were used to assess an overall association of rare variants within a gene with each trait.

4.3 Methods

All analyses involving individual-level data and WGS were performed on the UKB Research Analysis Platform (<https://ukbiobank.dnanexus.com>).

4.3.1 *Population cohort*

This chapter used data from the UK Biobank (UKB), which is described in detail in Chapter 2. Briefly, the UKB is a large prospective study (N = 502, 492) of participants recruited from around the United Kingdom between the ages of 40 and 69³⁴⁴. Baseline data collection occurred between 2006 and 2010 and included extensive phenotyping for a range of health outcomes alongside genotyping. The UK Biobank data was collected following informed consent obtained for all participants. The North-West Research Ethics Committee reviewed and approved UKB's scientific protocol and operational procedures (11/NW/0382) The analysis and data acquisition for this chapter were conducted under application #4844.

4.3.2 *Whole genome sequencing data*

Whole genome sequencing (WGS) data was generated for 490,549 individuals in UK Biobank through a public-private partnership between UK Biobank, UK Research and Innovation, Wellcome and four industry partners²²⁶. Sequencing was performed in various batches at two centres, deCODE facility in Reykjavik, Iceland (n = 247,522) and Wellcome Sanger Institute in Cambridge UK (n = 243,027, 49,934 in the Vanguard phase and 193,093 in the main phase) using Illumina NovaSeq 6000 instruments, with paired-end sequencing using S4 flowcells (v1.0 chemistry). The WGS data used in this chapter was processed using DRAGEN at AstraZeneca as previously described.

Specifically, sequence alignment and variant calling were performed using the Illumina DRAGEN Bio-IT Platform Germline Pipeline v3.7.8, run within the Amazon Web Services cloud platform. The sequence reads were aligned to GRCh38 graph genome reference and single nucleotide variants (SNVs) and short insertions and deletions (indels) were called at a sample level. Several variant-level quality control (QC) steps were applied to define high-quality variants, including depth of coverage ($> 10X$), various genotype and mapping quality scores, DRAGEN variant status, the proportion of alternate allele reads for heterozygous calls, and the proportion of samples failing these QC criteria. Detailed information regarding the pipeline can be found elsewhere (Illumina Dragen 3.7.8 User Guide)²²⁶. The DRAGEN Iterative gVCF Genotyper was used to aggregate samples per batch of 1000 and perform genotyping across 490,541 individuals, including all autosomes, sex chromosomes, mitochondria and 3341 ALT contigs of human genome build GRCh38. These analyses were conducted on Illumina Analytics (ICA) Platform.

4.3.3 Assessment of Major Depressive Disorder

We implemented a broader approach in the assessment of MDD compared to **Chapter 3**, which focused on a lifetime diagnosis of MDD based on the Composite International Diagnostic Interview (CIDI) questionnaire only. Here we use the MDD phenotype used for UKB in the most recent PGC MDD GWAS from Adams et al, (2025)^{183,226}, which encompasses additional data from primary care records such as GP diagnoses, prescription data and hospital records. Participants were defined as cases if they matched two or more of the following criteria: self-reported MDD, meeting CIDI case criteria (described in **Chapter 2**), self-declared antidepressant treatment, one to two GP diagnoses of MDD or three or more episodes of GP antidepressant prescribing.

Participants were also defined as cases if they matched one of the following criteria: three or more GP diagnoses of MDD or inpatient diagnoses of MDD. Cases were excluded if they had a self-reported or electronic health record of schizophrenia, psychosis, mania, bipolar disorder, personality disorder, schizotypal disorder or delusional disorder.

4.3.4 *Variant filtering and extraction*

Variants were extracted from the aggregated variant call files (pVCF) for six genes in the *FADS* cluster: *FADS1*, *FADS2*, *FADS3*, *FEN1*, *MYRF*, and *TMEM258* (Supplementary Table 1). For each gene, we selected the aggregated variant call files (pVCF, each 20kb) which overlapped its genomic region and filtered any files that extended beyond the gene boundaries, including the 3' and 5' UTR, from Ensembl⁴⁵⁶. This created a gene-level pVCF, which contained only variants within the gene boundaries for all participants (Supplementary Table 2, Supplementary Figure 1). All gene pVCFs were left-aligned and multi-allelic sites were split using bcftools (-norm and -m flags) using the the DRAGEN GRCh38 reference genome file (GRCh38_full_analysis_set_plus_decoy_hla.fa). Variants that had an exact conditional HWE P-value < 1×10^{-100} were discarded²²⁶. Variant-level missingness was calculated using bcftools⁴⁵⁷ (+fill-tags, F_MISSING) and those with a missingness > 0.1 were discarded. Additionally, per-sample missingness was calculated using vcftools⁴⁵⁸ (-missing -indv flag), and those with > 0.1 missingness were discarded. Samples were further filtered to unrelated individuals of European ancestry as in **Chapter 3**, and as previously reported¹⁷⁷. The sample was also then filtered to include only those with a non-missing MDD phenotype. Following sample-QC, variants with an allele-count of 0 in the QC'd sample were also removed.

4.3.5 Variant Annotation

Variants were annotated using VEP³⁸⁸ (release 113, GRCh38), to establish their consequence and associated impact on function. Additionally, various VEP plugins were used to further annotate variants with different metrics of deleteriousness. LOFTEE⁴⁵⁹, a tool which distinguishes high confidence putative loss-of-function (pLoF) variants from annotation and technical artefacts, was used to annotate stop-gained, splice-site disrupting and frameshift variants. Further annotations for deleteriousness include REVEL³⁹⁹, which uses an ensemble method of individual tools to predict the pathogenicity of missense variants, and the CADD (version 1.7)⁴¹⁴ score, which scores the predicted deleteriousness of coding and non-coding variants based on numerous genomic features, such as cross-species conservation, biochemical action and predicted impact on protein structure⁴¹⁴. Finally, variants were annotated with the observed allele frequencies from the gnomAD genomes (version 4)⁴⁵⁹.

4.3.6 Variant Prioritisation

First, rare-variant sets per gene were created by filtering based on AF in the UKB sample and in gnomAD. Specifically, variants were retained if they had an AF < 1% in the UKB sample and if they were either absent or had an AF < 1% in gnomAD. These rare-variant sets were then further prioritised based on meeting any of the following criteria: a high/moderate VEP impact score, a LOFTEE score of High Confidence (HC) or Other Splice (OS), a CADD score > 20 and a REVEL score > 0.5. For the prioritised variants in each gene, we assessed the linkage disequilibrium between the UKB samples using

PLINK⁴⁶⁰(--r2 flag). Prioritised variants were then placed into five different masks depending on variant annotations and allele frequencies (Table 4.4).

Table 4.4 The criteria for each mask created per gene.

Note that Mask 5 incorporates all prioritisation criteria. LoF = LoFTEE Loss-Of-Function, VEP = Variant Effect Predictor, REVEL = rare exome variant ensemble learner, CADD = combined annotation dependent deletion.

Allele Frequency Criteria	Mask	Annotations Criteria	Allele Frequency	Mask Name
UKB: AF < 0.01 gnomAD: AF < 0.01/ absent	Mask 1	LoF == High Confidence/Other Splice	< 0.01	Mask1.0.01
			Singletons	Mask1.singletons
	Mask 2	LoF == High Confidence/Other Splice VEP HIGH impact	< 0.01	Mask2.0.01
			Singletons	Mask2.singletons
	Mask 3	LoF == High Confidence/Other Splice VEP HIGH impact REVEL score > 0.5	< 0.01	Mask3.0.01
			Singletons	Mask3.singletons
	Mask 4	LoF == High Confidence/Other Splice VEP HIGH impact REVEL score > 0.5 CADD score > 20	< 0.01	Mask4.0.01
			Singletons	Mask4.singletons
	Mask 5	LoF == High Confidence/Other Splice VEP high impact REVEL score > 0.5 CADD score > 20 VEP moderate impact	< 0.01	Mask5.0.01
			Singletons	Mask5.singletons

4.3.7 *Metabolic biomarkers*

We assessed the metabolite levels of the five metabolites identified with a causal association with MDD in **Chapter 3**: docosahexaenoic acid (DHA, Field ID: 23450), omega-3 fatty acids (omega-3, Field ID: 23444), omega-3 to total fatty acids % (omega3 %, Field ID: 23451), omega-6: omega-3 ratio (omega-6: omega-3, Field ID: 23459) and the degree of unsaturation (DegUnsat, Field ID: 23443). We utilized both Phase 1 and Phase 2 of the baseline metabolomic data in UKB (N ~ 275,000). This provides an additional 157,000 participants from those studied in **Chapter 3**. The metabolites were quantified in non-fasting baseline EDTA samples collected at baseline assessment between April 2007 and December 2010⁴⁶¹ and measured between June 2019 and April 2020 (Phase 1) and April 2020 and June 2022 (Phase 2), using eight spectrometers at Nightingale Health, Finland. The measurement was performed using the NMR spectroscopy platform, developed by Nightingale Health Ltd, Finland, with detailed methodology described elsewhere²⁸⁶. Prior to analysis, the levels of each metabolite were transformed to a standard normal distribution using rank-based inverse normalization, as in **Chapter 3**. We restricted our sample to unrelated individuals of European ancestry (N = 207,643) and to those present in the QC'd whole-genome sequencing data (N = 206,483).

4.3.8 *Sub-cohorts*

As there are variations in data availability between the metabolomic data and the MDD phenotype, we analysed the association of rare, prioritised variants within each gene with these traits using different subsets of UKB, referred to here as cohorts. The MDD cohort (used to test the association with MDD) was comprised of all unrelated participants of European ancestry with WGS and a non-missing MDD phenotype (n = 362,202). The

metabolite cohort (used to test the association with each metabolite) comprised of all unrelated participants of European ancestry with WGS and baseline metabolomic data (n = 206,483).

4.3.9 Statistical tests

For each gene and mask, we implemented Pearson's Chi-squared test with Yates continuity correction to test the association of carrying prioritized variants with MDD in the MDD cohort, and a two-tailed Welch t-test to test the association of carrying prioritized variants with the different metabolic traits in the metabolite-cohort. For the latter, we applied a BH FDR correction to account for multiple testing⁴⁶².

4.3.10 Aggregate variant testing

Aggregate variant testing (AVT) was performed using the REGENIE software⁴¹⁶. REGENIE follows a two-step framework, whereby the first step computes trait predictions from a whole-genome regression model using genotype array data and step 2 computes variant-phenotype associations conditioned on the prediction values calculated in step 1. This approach can account for the effects of population stratification and sample relatedness on variant-phenotype associations. Both steps of REGENIE were run separately for the MDD and metabolite cohorts. The first step utilised the genotype data in UKB (field ID: 22418, N = 487,921). Variant-level quality control steps included removing variants which had a MAF < 0.01, MAC < 20, HWE $p < 1e^{-15}$ or with > 10% missingness across all samples. Sample-level quality control included removing samples which were not present in the phenotype file (metabolite or MDD), alongside removing those with more than 10% missingness across all variants.

For MDD, we also implemented a fast Firth logistic regression in REGENIE ('—firth' flag) to account for the imbalance in the case-control ratio^{463,464}. Covariates for each analysis included age (Field ID: 21003), sex (Field ID: 31), genotype array (Field ID: 22000), assessment centre (Field ID: 54) and the top 10 genetic principal components (Field IDs: 22009). Analysis on the metabolite traits also covaried for spectrometer (Field ID: 23650).

REGENIE Step 2 conducted AVT for the prioritised variants in each mask per gene, using the same covariates as Step 1. Several different types of AVT were used, including the traditional burden test, the SKAT, and the ACAT-V test. Additionally, omnibus tests which combine features from these tests, such as SKAT-O and ACAT-O were also applied. The variety of tests were employed as each has distinct advantages and assumptions (described in Section 4.1.2) and therefore can inform about potential genetic architectures underlying rare-variant associations. Furthermore, joint-burden tests (MIN_P, SBAT and GENE_P) were implemented to get a gene-level measure of association across all masks.

4.4 Results

4.4.1 Identification of rare pathogenic variants in *FADS*

4.4.1.1 Variant filtering and extraction

There was a total of 82,237 variants identified across all six genes (*FADS1* $n_{\text{variants}} = 13,978$, *FADS2* = 34,621, *FADS3* = 7,925, *FEN1* = 1,644, *MYRF* = 14,556 and *TMEM258* = 9,513), which was reduced to 76,341 (*FADS1* $n_{\text{variants}} = 11,387$, *FADS2* = 31,529, *FADS3* = 7,787, *FEN1* = 1,205, *MYRF* = 14,502 and *TMEM258* = 9,492) following variant quality control (Supplementary Figure 2) for 490,541 participants.

Restricting to unrelated participants of European ancestry reduced the sample size to 371,437. Further sample-level quality control reduced the unrelated sample size by 6.2% in all genes but *FADS1* (*FADS2* $n_{\text{participants}} = 369,130$ and *FADS3*, *FEN1*, *MYRF*, *TMEM258*: $n_{\text{participants}} = 369,135$) (Supplementary Figure 3). Variants within *FADS1* displayed a much higher rate of missingness per-person (Supplementary Figure 4), and therefore sample quality control filters reduced the sample size by 38% ($n_{\text{participants}} = 227,592$). Filtering for participants with a non-missing MDD status further reduced the sample size by 1.9% across all genes (*FADS1* $n_{\text{participants}} = 223,316$, *FADS2* = 362,201, *FADS3*, *FEN1*, *MYRF*, *TMEM258* = 362,206). Variants which then had an allele count of 0 in the QC'd sample were also removed per gene before variant annotation, resulting in a 43% reduction in variants in *FADS1* ($n_{\text{variants}} = 6,490$), and ~26% decrease in variants for *FADS2* ($n_{\text{variants}} = 23,323$), *FADS3* ($n_{\text{variants}} = 5,574$), *FEN1* ($n_{\text{variants}} = 1,205$), *MYRF* ($n_{\text{variants}} = 10,631$) and *TMEM258* ($n_{\text{variants}} = 6,925$) (Supplementary Figure 5). Across all genes,

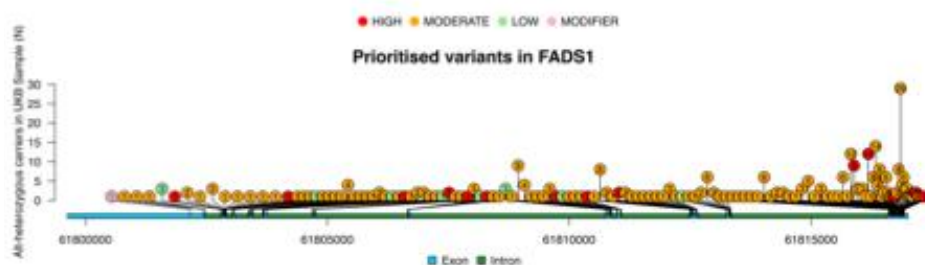
most of these variants were rare ($AF < 0.01$) within the UKB sample (Supplementary Figure 6) and there was a large proportion of singleton variants (Supplementary Figure 7).

4.4.1.2 Variant prioritisation

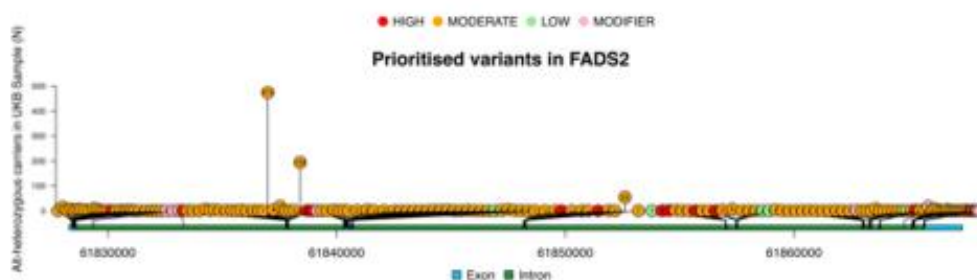
For each gene, the output from VEP was filtered to those annotated to the canonical transcript of the gene of interest (Supplementary Table 1). For *FADS3*, *FEN1* and *MYRF*, every variant had an annotation to the canonical transcript of the corresponding gene, however only 73, 65 and 35% of the variants for *FADS1*, *FADS2* and *TMEM258* had annotations for the corresponding canonical transcript respectively. There were 1,408 variants which met at least one prioritization criteria per gene (*FADS1* $n_{\text{variants}} = 142$, *FADS2* = 177, *FADS3* = 233, *FEN1* = 176, *MYRF* = 623, *TMEM258* = 77, Figure 4.5). The proportion of rare variants per gene which were prioritised varied from 1% (*FADS2*) to 15% (*FEN1*). Across all genes, the prioritised variants were predominantly missense singletons (Figure 4.6). Most of the prioritised variants were prioritised based off the high/moderate impact predictions from VEP (Supplementary Figure 8-Supplementary Figure 13). In-depth information regarding the annotations of prioritised variants can be found in Supplementary Data 1-6. There was no correlation between any prioritised variants in *FADS2* ($\max R^2 = 7.6 \times 10^{-7}$), *TMEM258* ($\max R^2 = 5.6 \times 10^{-6}$), or *FADS3* ($\max R^2 = 1.86 \times 10^{-8}$) (Supplementary Figure 14 - Supplementary Figure 16).

There was one, four and 11 pairs of prioritised variants in *FEN1*, *FADS1* and *MYRF* respectively which showed complete correlation ($R^2 = 1$) and two pairs of variants in *MYRF* which exhibited low correlations ($R^2 = 0.12-0.25$) (Supplementary Table 3, Supplementary Figure 17). All other variant pairs in these genes exhibited no correlations (*FEN1*: $\max R^2 = 1.64 \times 10^{-6}$, *FADS1*: $\max R^2 = 4.65 \times 10^{-9}$, *MYRF*: $\max R^2 = 3.75 \times 10^{-3}$).

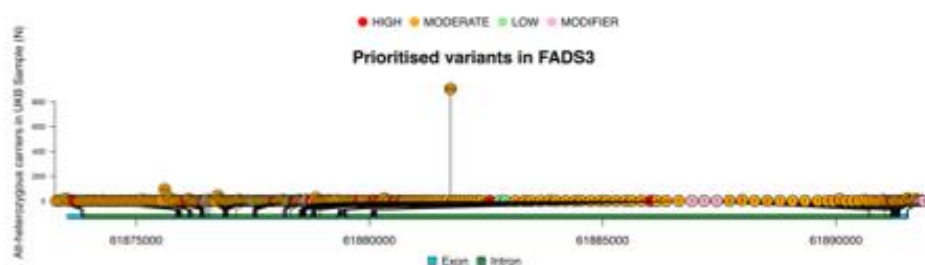
A)



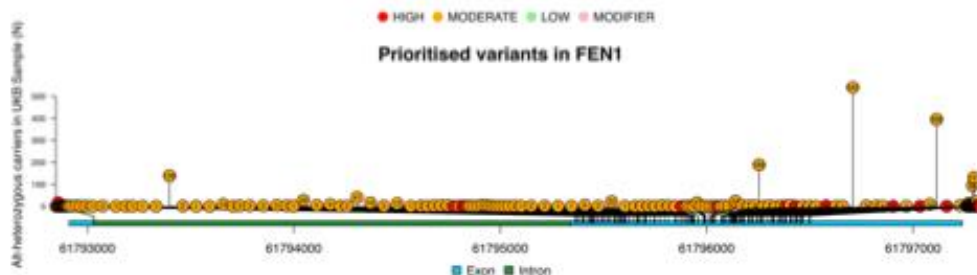
B)



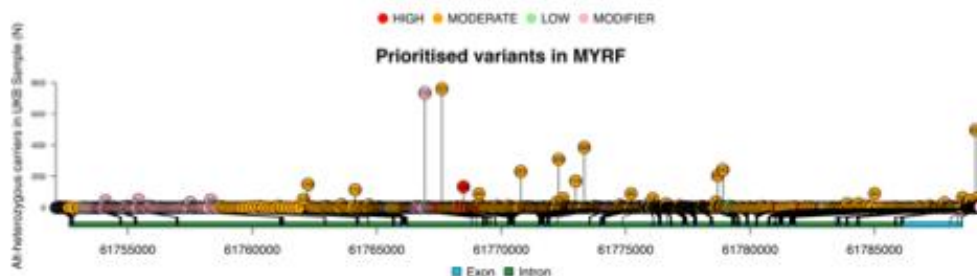
C)



D)



E)



F)

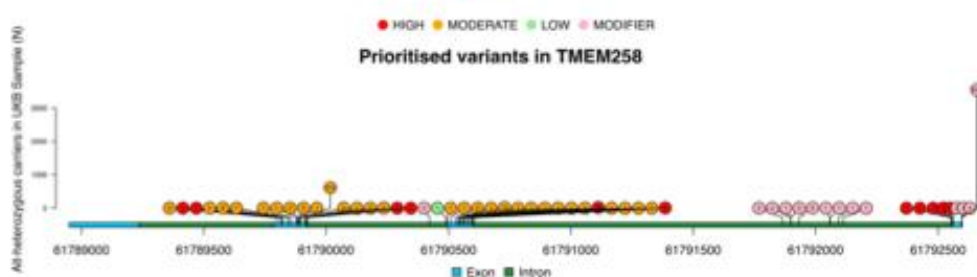
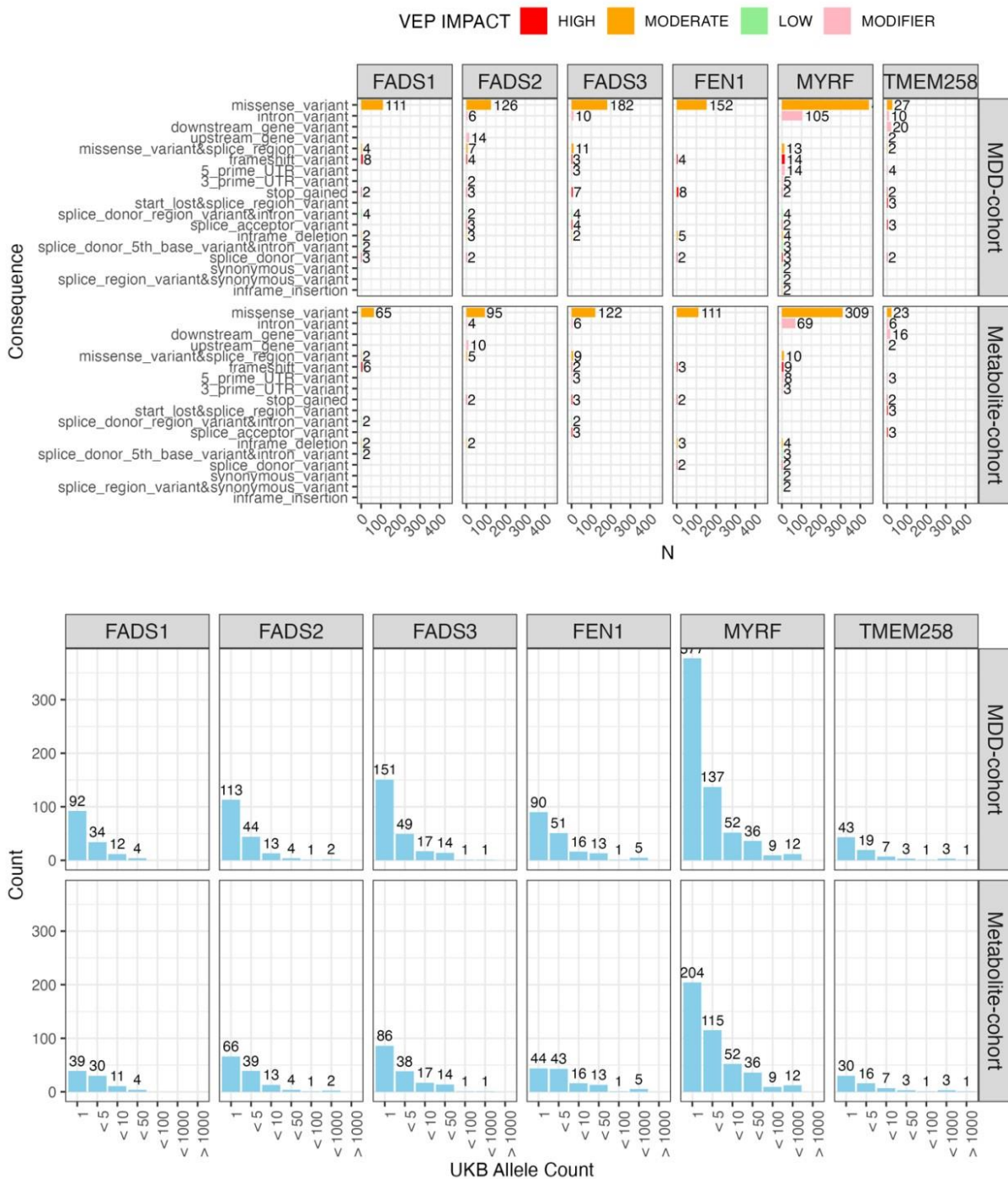


Figure 4.5 Prioritised variants in each gene in the FADS cluster.

Lollipop plots show the prioritised variants per gene, coloured by their predicted impact from VEP. The number of alternate heterozygous carriers of each variant in the MDD-cohort is represented by the y-axis and labelled per variant. VEP = Ensembl Variant Effect Predictor.



A) The predicted consequence of prioritised variants from VEP. B) The allele counts of the prioritised variants in each cohort.

4.4.1.3 Carriers per cohort

All the prioritised variants were present in the MDD cohort, carried in a total of 15,751 participants. Most prioritised variants across all genes were present as alternate heterozygous (0/1 or 1/0), however there were a few alternate homozygous carriers (1/1) of prioritised variants in *FADS3* ($n_{\text{participants}} = 2$), *MYRF* ($n_{\text{participants}} = 9$) and *TMEM258* ($n_{\text{participants}} = 19$) (Figure 4.7). Additionally, there were 83 participants who carried > 1 prioritised variant per gene in an alternate heterozygous state (*FADS1* = 18, *FADS3* = 2, *FEN1* = 5, *MYRF* = 51 and *TMEM258* = 7) (Figure 4.7). Due to the lower sample size, only 961 of the prioritised variants were present in the metabolite cohort, carried by a total of 8,693 participants (*FADS1* $n_{\text{variants}} = 84$, *FADS2* = 125, *FADS3* = 157, *FEN1* = 122, *MYRF* = 428, *TMEM258* = 61) (Figure 4.7). Like the MDD cohort, most participants carried a single prioritised variant as alternate heterozygous although there were a few alternate homozygous carriers of prioritised variants in *TMEM258* ($n_{\text{participants}} = 13$) and *MYRF* ($n_{\text{participants}} = 5$). There were also 46 participants who carried > 1 prioritised variant per gene in an alternate heterozygous state (*FADS1* $n_{\text{participants}} = 11$, *FADS3* = 2, *FEN1* = 4, *MYRF* = 24, *TMEM258* = 5) (Figure 4.7).

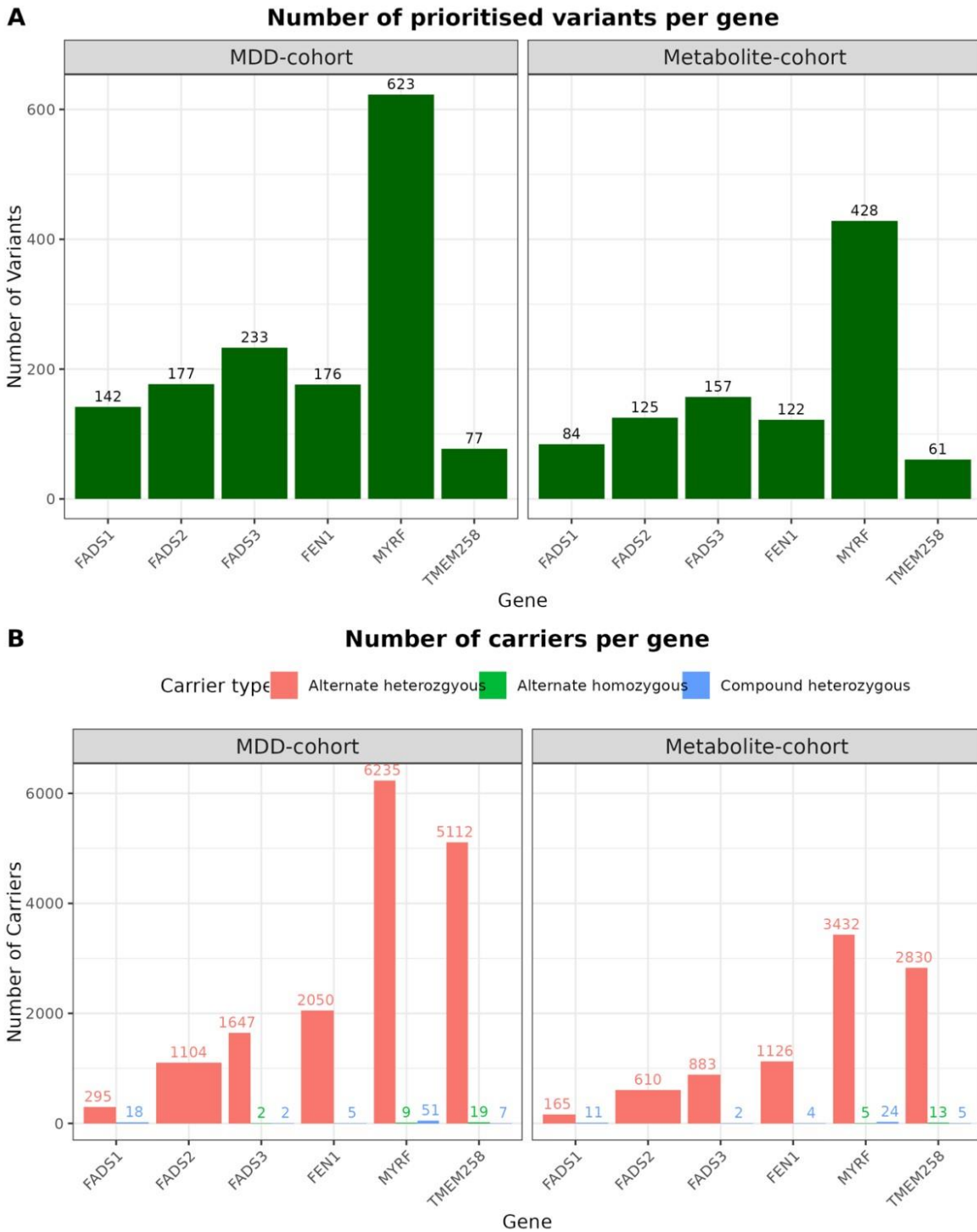


Figure 4.7 Summary of the prioritised variants and participant carriers in each cohort.

- A) The number of prioritised variants per gene carried in any participant of the cohort.
 - B) The number of carriers of prioritised variants per gene, coloured by their zygosity.
- Here, compound heterozygous is used to represent participants who carry > 1 prioritised variant.

There were 796 and 411 participants who carried prioritised variants in more than one *FADS* gene in the MDD and metabolite cohort respectively. *TMEM258* and *MYRF* were the most frequently co-occurring genes with prioritised variants in these participants (Supplementary Figure 20).

4.4.1.4 Variant masks

The number of variants within each gene-mask in the MDD cohort and Metabolite cohort and gene are shown in Figure 4.8. In general, Mask 1 had very low number of variants (*FADS1* $n_{\text{variants}} = 12$, *FADS3* = 5, *FEN1* = 0, *MYRF* = 5, *TMEM258* = 5), and carriers across all genes. As expected, as the mask criteria became less stringent, there were more variants and carriers available for burden testing. The number of carriers and carrier types per gene-mask in each cohort are shown in Supplementary Figure 21. Lolliplots visualising the variants included in each mask per each gene are shown in Supplementary Figure 22 - Supplementary Figure 33.

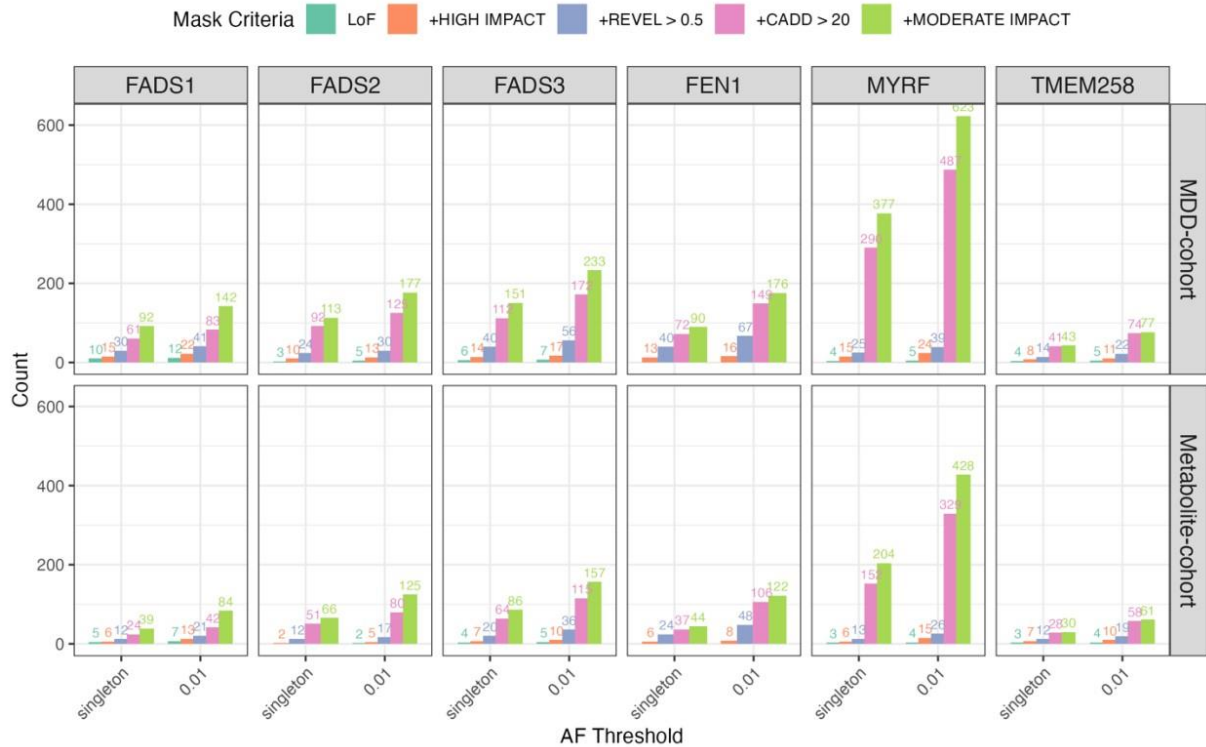


Figure 4.8 Variant masks per gene for variant testing.

The counts of variants present within each mask (coloured) for each gene (columns) in each cohort (rows).

4.4.2 Testing the association of rare-prioritised variants with MDD and metabolites

4.4.2.1 Demographics of MDD and metabolite cohorts

The MDD-cohort and metabolite-cohort showed similar demographic profiles (Table 4.5). Both cohorts have the same mean age (56), have a slightly higher proportion of females (~ 54%), are predominantly of white ethnicity and have an average BMI of 27.4. Both cohorts also show similar proportions of current smokers (~ 10%), University/College attendees (~11%), and MDD cases (~ 15%). These demographic characteristics remain

largely consistent between carriers and non-carriers of prioritised variants in each gene within both the MDD and metabolite cohort (Supplementary Table 4).

Table 4.5 The demographics of the MDD cohort and metabolite-cohort.

Continuous traits (Age, BMI and SES) are displayed as the mean values within the cohorts whilst categorical traits are displayed as the counts and % of the sample.

Smoking status shows the ratio of current, previous and never smokers, socioeconomic status was measured using the Townsend Deprivation Index (TDI), described in Section 2.1.6. Ethnicity shows the ratio of White, Mixed and Other populations, and MDD shows the ratio of MDD cases and controls within the sample.

Cohort	MDD cohort	Metabolite cohort
Total	362202	206483
Age	56.8 (8.01)	56.7 (8)
BMI	27.4 (4.75)	27.4 (4.76)
SES	-1.48 (2.99)	-1.51 (2.98)
% Female	195746 (54%)	110693 (53.6%)
% Attended University/College	40891 (11.3%)	22363 (10.8%)
Smoking status (Current:Previous:Never)	37066 (10.2%): 128469 (35.5%): 195207 (53.9%)	21337 (10.3%): 73083 (35.4%): 111268 (53.9%)
Ethnicity (White: Mixed: Other)	359963 (99.4%): 463 (0.128%): 755 (0.208%)	205237 (99.4%): 270 (0.131%): 418 (0.202%)
MDD status (Cases: Controls)	54793 (15.1%): 307409 (84.9%)	31167 (15.1%): 171499 (83.1%)

4.4.2.2 Statistical tests

4.4.2.2.1 MDD-cohort

The counts of MDD cases and MDD controls per gene-mask combination in the MDD cohort are shown in Figure 4.9. The Chi-squared test found no evidence of an association between carrying any prioritised variant in any gene with MDD (P range: 0.117 to 1, Supplementary Table 5).

4.4.2.2.2 Metabolite-cohort

The Welch t-test found that carrying a prioritised variant in *TMEM258* Mask 4 and Mask 5 (both $AF < 0.01$) was significantly associated with all metabolite traits (P_{BH} range: 4.54×10^{-21} to 3.44×10^{-10}). Specifically, carriers of these variants showed higher levels of DHA (Cohens $D = 0.13$), omega-3 (Cohens $D = 0.15-0.18$), omega3 % (Cohens $D = 0.19$), and the DegUnsat (Cohens $D = 0.14$), and a lower ratio of omega-6: omega-3 (Cohens $D = -0.17$) (Supplementary Figure 34 - Supplementary Figure 35). There was also a weaker significant association between carrying a prioritised variant in *FEN1* Mask 5 ($AF < 0.01$) with omega-3 % ($t(1140) = -3.16$, $P_{BH} = 0.038$, $D = 0.097$) and the omega-6: omega-3 ($t(1140) = 3.49$, $P_{BH} = 0.013$, $D = 0.1$) (Supplementary Figure 36). Finally, there was another weak association between carrying a prioritised variant in *FADS3* Mask 3 ($AF < 0.01$) with the DegUnsat ($t(19) = -3.61$, $P_{BH} = 0.041$, $D = 0.84$) (Supplementary Figure 37). The significant t-tests are shown in Supplementary Table 6, and all the t-test results per gene-mask with each metabolite can be found in Supplementary Data 7.

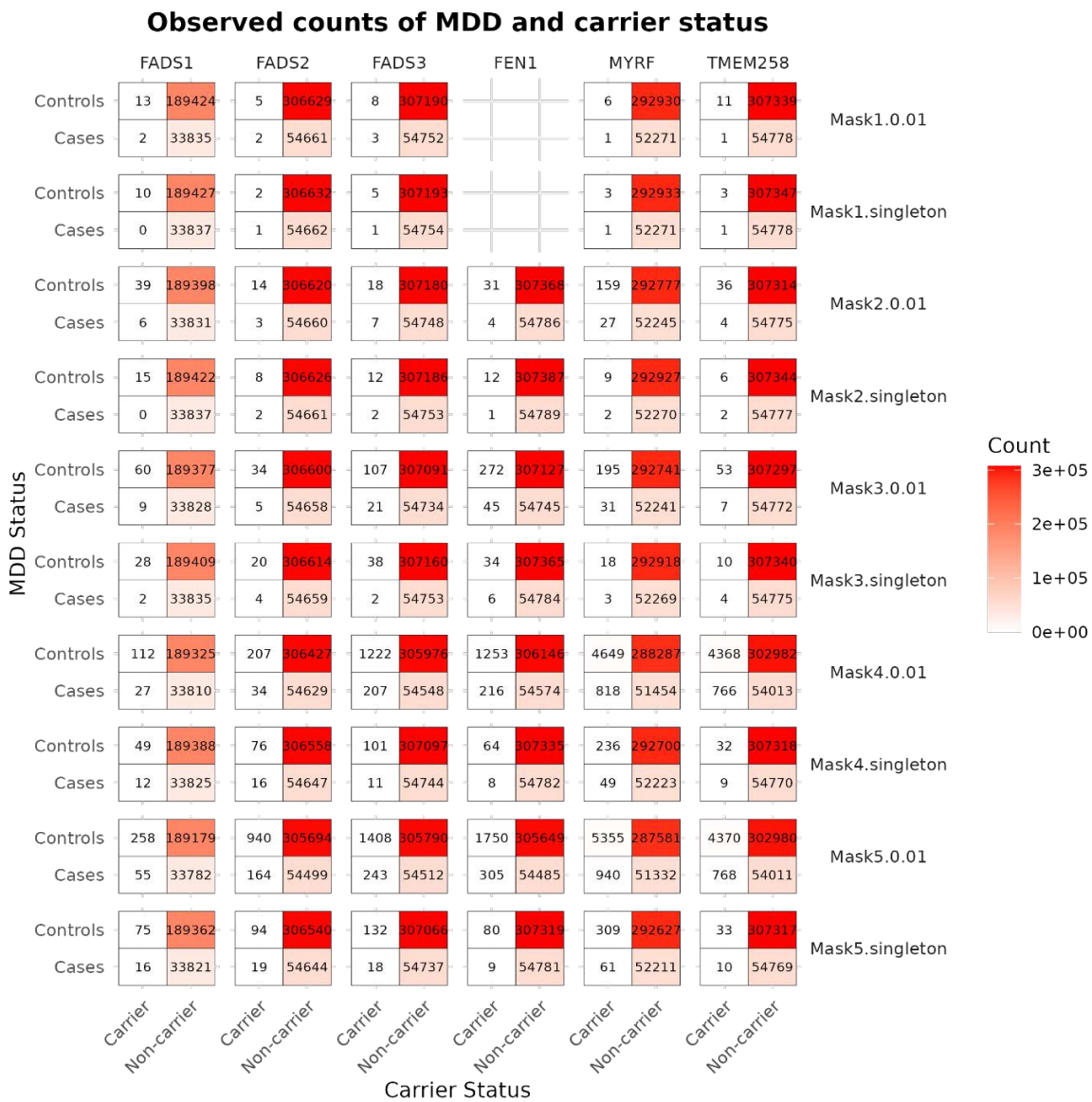


Figure 4.9 Counts of MDD status in carriers/non-carriers of prioritised variants in the FADS cluster.

The observed counts of MDD cases (bottom) and controls (top) in carriers (left) and non-carriers (right) of prioritised variants in each mask (rows) across all genes (columns). The shading of red in each box represents the count of the participants in each category.

4.4.2.3 Aggregate-Variant-Testing

The significance of different AVT tests per gene-mask combination is shown for MDD and metabolite traits in Figure 4.10. The effect size estimate for aggregated pseudo-genotype used in the traditional burden test is shown per gene-mask combination for each trait in Supplementary Figure 38. All results per gene-mask combination for each trait can be found in Supplementary Data 8.

4.4.2.3.1 MDD-cohort

Using the ACAT-V test, there was a nominally significant association between carrying all prioritised variants in *FADS1* (Mask 5, $AF < 0.01$) and MDD status ($P_{ACAT-V} = 0.024$), which was not seen using the other AVT tests ($P_{ACAT-O} = 0.12$, $P_{BURDEN} = 0.13$, $P_{SKATO} = 0.15$, $P_{SKAT} = 0.47$). The traditional burden test identified associations which approached nominal significance for Mask 3 (singletons) in *FADS3* ($\beta_{BURDEN} = -0.8$, $P_{BURDEN} = 0.083$) and Mask 4 variants ($AF < 0.01$) in *FADS1* with MDD ($\beta_{BURDEN} = 0.41$, $P_{BURDEN} = 0.094$). All other masks in all FADS genes showed null associations with MDD across all AVT tests (P range: 0.15-1.0). Although joint burden testing across all masks per gene showed no associations with MDD following correction for multiple testing, the strongest signal was for *FADS1* ($P_{MINP} = 0.094$, $P_{GENE\_P} = 0.67$) and *FADS3* ($P_{MINP} = 0.083$, $P_{GENE\_P} = 0.46$). All other genes showed no evidence of an association of carrying prioritised variants with MDD (P_{MINP} range = 0.15-0.38, P_{GENE_P} range = 0.86-0.99) (Figure 4.10).

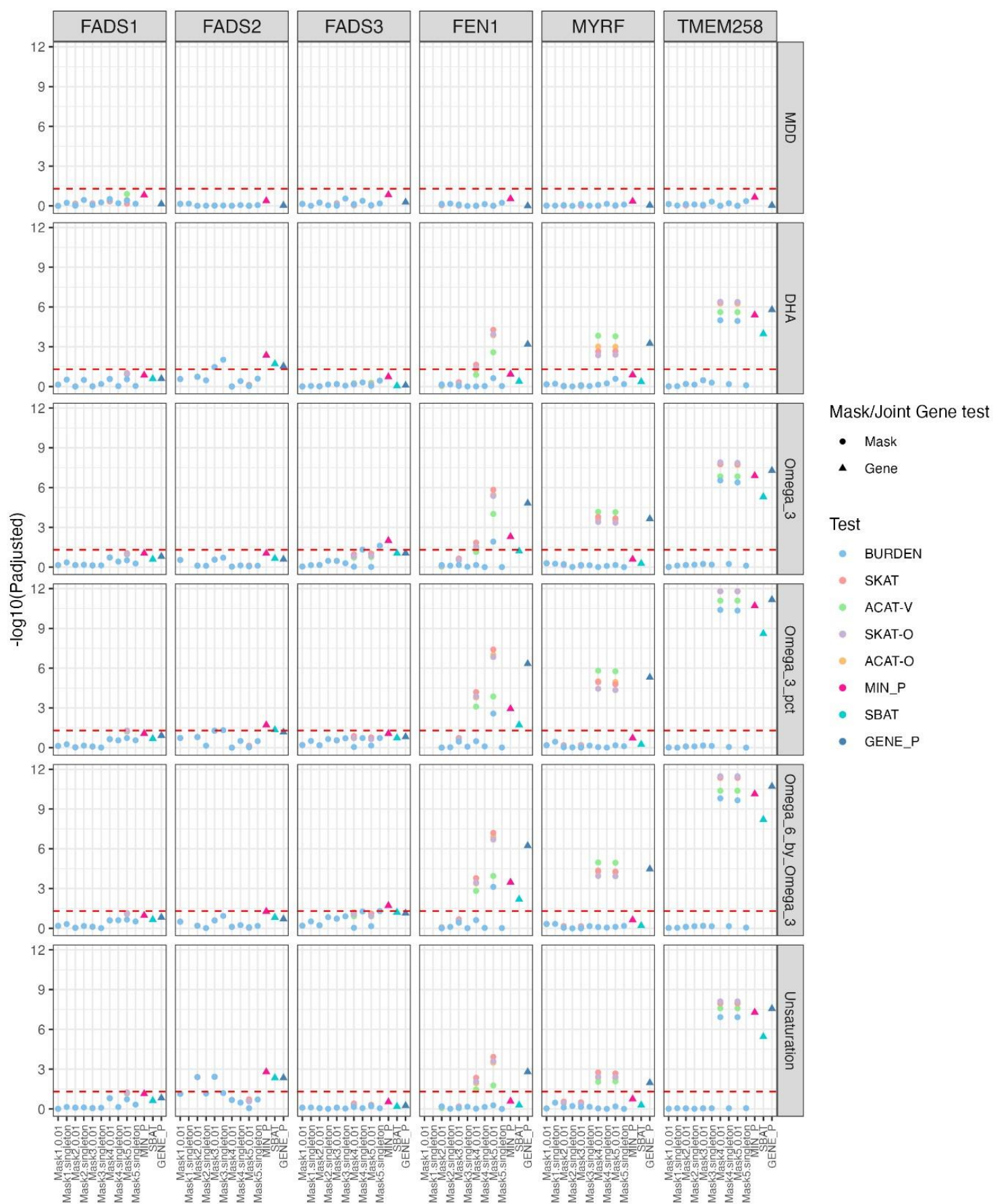


Figure 4.10 Aggregate-variant-testing of prioritised variants within each gene with MDD and omega-3 related measures.

The $-\log_{10}$ -transformed p value of each test is shown on the y-axis for each mask and for the joint tests across all masks per gene (x-axis). The red dashed line indicates the threshold for nominal significance ($P < 0.05$). The colour of the point denotes which test and the shape denotes whether the test is applied per individual mask or a joint test across all masks (per gene). Tests applied per mask include: BURDEN = Traditional burden test, SKAT = Sequence kernel association test, ACAT=Aggregated Cauchy combination test, SKAT-O= SKAT-Burden omnibus test, ACAT-O = ACAT omnibus test which combines SKAT, ACAT-V and Burden test p values using ACAT. Joint burden test across all masks per gene include: MIN_P = Minimum P value, SBAT= Sparse burden association tests and GENE_P = REGENIE joint gene test. Note that SBAT is currently only developed for quantitative traits and therefore is not applied to MDD.

4.4.2.3.2 Metabolite cohort

Whilst all the FADS genes showed a significant association with each metabolomic trait in at least one mask, *TMEM258*, *MYRF* and *FEN1* consistently demonstrated much stronger associations with the metabolomic traits across all masks. The prioritised variants in Mask 4 and Mask 5 ($AF < 0.01$) consistently demonstrated the strongest signal across all genes, apart from *FADS2*, where Mask 3 ($AF < 0.01$ and singletons) was the strongest. *TMEM258* Mask 4 and Mask 5 variants ($AF < 0.01$) demonstrated the most robust association with all the metabolomic traits, strongest in omega-3 % (P_{adj} range: 1.38×10^{-12} to 4.32×10^{-11}). Both *MYRF* and *FEN1* Mask 4 and Mask 5 variants ($AF < 0.01$) also showed significant associations with each metabolite in all tests but the burden test

(*MYRF* P_{adj} range: 1.46×10^{-6} to 8.65×10^{-3} and *FEN1* P_{adj} range: 3.8×10^{-8} to 0.049). *FADS1* Mask 4 and Mask 5 (AF < 0.01) displayed much weaker associations with each metabolite (P_{adj} range: 0.048 to 0.25). *FADS3* Mask 4 and Mask 5 variants (singletons) also had borderline significant associations with omega-3 (P_{adj} range: 0.023 to 0.046) and omega-6: omega-3 (P_{adj} range: 0.048 to 0.0515). Finally, *FADS2* variants in Mask 3 (AF < 0.01 and singletons) showed a significant association with DHA (P_{adj} range: 9.1×10^{-3} to 0.033) whilst variants in Mask 2 and Mask 3 (AF < 0.01) had significant associations with the DegUnsat (P_{adj} range: 3.67×10^{-3} to 3.86×10^{-3}).

The results from joint burden testing methods across all masks per gene is shown in Figure 4.10 and in

Supplementary Table 7. *TMEM258* demonstrated robust associations with each metabolite across all methods (P_{MINP} range = 1.97×10^{-11} to 4.01×10^{-6} , P_{SBAT} range = 2.49×10^{-9} to 1.1×10^{-4} , $P_{\text{GENE\_P}}$ range = 6.84×10^{-12} to 1.61×10^{-6}). *MYRF* and *FEN1* showed evidence of an association with all metabolites using the GENE_P measure only (*MYRF*: P_{MINP} range = 0.13 to 0.26, P_{SBAT} range = 0.44 to 0.64, $P_{\text{GENE\_P}}$ range = 4.95×10^{-6} to 0.011 and *FEN1*: P_{MINP} range = 3.38×10^{-4} to 0.26, P_{SBAT} range = 6.52×10^{-3} to 0.52, $P_{\text{GENE\_P}}$ range = 4.62×10^{-7} to 1.61×10^{-3}). Additionally, *FADS3* showed a borderline significant association with both omega-6: omega-3 (P_{MINP} = 0.019, P_{SBAT} = 0.062, $P_{\text{GENE\_P}}$ = 0.072) and omega-3 (P_{MINP} = 0.01, P_{SBAT} = 0.092, $P_{\text{GENE\_P}}$ = 0.089) only. Additionally, *FADS2* showed significant associations with DHA (P_{MINP} = 4.4×10^{-3} , P_{SBAT} = 0.019, $P_{\text{GENE\_P}}$ = 0.029) and the DegUnsat (P_{MINP} = 1.6×10^{-3} , P_{SBAT} = 4.55×10^{-3} , $P_{\text{GENE\_P}}$ = 4.54×10^{-3}) across all methods and also an association with omega-3 % and omega-6: omega-3 ratio using the MIN_P and SBAT measures only (P_{MINP} = 0.019, P_{SBAT} = 0.045, $P_{\text{GENE\_P}}$ = 0.067).

Finally, *FADS1* showed no significant associations with any of the metabolite measures (P_{MINP} range = 0.072 to 0.14, P_{SBAT} range=0.21 to 0.26, $P_{\text{GENE\_P}}$ range: 0.12 to 0.26).

4.4.2.4 Leave-one-variant out analysis (LOVO)

Although not significant, we noticed a small consistent signal between *FADS1* variants in Mask 5 ($AF < 0.01$) with MDD and the metabolite traits. To investigate further, we ran the LOVO analysis on this mask for MDD and the metabolite traits, selecting the best performing test per trait (ACAT-V for MDD and SKAT for all metabolomic traits). The LOVO analysis creates a series of burden masks for a set of variants, whereby one variant is left out of the mask at a time, which are then tested for an association with a trait (MDD and all metabolite traits). If there is a large decrease in statistical significance when a singular variant is left out of the mask, this indicates this variant is a main driver of the association. Therefore, this approach can help to evaluate whether an association is driven by multiple variants or a single variant.

4.4.2.4.1 LOVO analyses on *FADS1* Mask 5 ($AF < 0.01$) and MDD

Full results of the LOVO analysis for each trait are shown in Supplementary Data 9 and Figure 4.11. For MDD, the LOVO analysis identified a missense variant in *FADS1* exon 1 (chr11:61816814:G:C) which contributed the most to the burden association of prioritised variants in *FADS1* with MDD (Mask 5 $AF < 0.01$, $P_{\text{ACAT-V}}=0.025$). Excluding this variant from the burden mask led to a large drop in statistical significance ($P_{\text{ACATV}}=0.636$). Notably, removing any other variants in the mask made a minimal difference to the association results ($P_{\text{ACAT-V}}$ range= 0.021 to 0.025), indicating that this variant may be the sole driver of the association. There were 14 alternate heterozygous carriers of this

variant in the UKB MDD cohort, 7 of which had MDD (50%) (Table 4.6). A two-tailed Fisher's exact test found a statistically significant relationship between chr11:61816814:G:C carrier status and MDD status (OR = 5.54 [1.68-18.71], $p = 0.0023$). Eight of these carriers were also present in the metabolite cohort (Supplementary Figure 39). Welch t-tests were applied to assess the relationship between chr11:61816814:G:C carrier status and metabolite levels, which found no evidence of an association (P range: 0.58 to 0.73) (Supplementary Table 8).

Table 4.6 The demographics of the carriers and non-carriers of variant chr11_61816814_G_C in the MDD cohort.

Continuous traits (Age, BMI and TDI) are displayed as the mean values within the cohorts whilst categorical traits are displayed as the counts and % of the sample. Smoking status shows the ratio of current, previous and never smokers, Ethnicity shows the ratio of White, Mixed and Other populations, and MDD shows the ratio of MDD cases and controls within the sample.

chr11_61816814_G_C	Non-carrier	Carrier
Total	223260	14
Age	56.8 (8)	55.1 (9.61)
BMI	27.4 (4.75)	26.3 (3.71)
TDI	-1.49 (2.98)	0.035 (3.34)
% Female	120523 (54%)	9 (64.3%)
% Attended University/ College	25130 (11.3%)	3 (21.4%)
Smoking status	22760 (10.2%):79117 (35.4%):120488 (54%)	1 (7.14%):7 (50%):6 (42.9%)
Ethnicity	221859 (99.4%):291 (0.13%):498 (0.223%)	14 (100%):0 (0%):0 (0%)
MDD	33830 (15.2%):189430 (84.8%)	7 (50%):7 (50%)

4.4.2.4.2 LOVO analysis on FADS1 Mask 5 (AF < 0.01) and metabolomic traits

The LOVO analysis also identified another missense variant in exon 5 of *FADS1* (chr11:16810815:C:A) which was a main driver in the burden associations with prioritised variants in *FADS1* (Mask5, AF < 0.01) with DHA (P_{SKAT} = 0.017), omega-3 (P_{SKAT} = 0.016) and omega-6: omega-3 (P_{SKAT} = 0.011). Excluding this variant led to a small drop in significance for each metabolite; DHA (P_{SKAT} = 0.055), omega-3 (P_{SKAT} = 0.054) and omega-6: omega-3 (P_{SKAT} = 0.043). Notably, excluding other variants did also reduce the P-value for DHA (P_{SKAT} range: 0.011 to 0.036), omega-3 (P_{SKAT} range: 6.7×10^{-3} to 0.033) and omega-6: omega-3 (P_{SKAT} range: 0.058 to 0.025), indicating that burden association was driven by various variants in the burden mask. There were 8 carriers of this variant in the metabolite cohort (Supplementary Table 9). Overall, carriers of this variant had lower levels of DHA, omega-3 and omega-3 %, whilst also exhibiting higher omega-6: omega-3 (Supplementary Figure 40). However, using Welch's t-test, there was no evidence of a significant association between chr11:16810815:C:A carrier status and the metabolite levels (P range: 0.11 to 0.28, Supplementary Table 10). Notably, a single outlier in the carrier group for each metabolite except the DegUnsat, was identified using the interquartile range (IQR) method (Supplementary Table 11). When excluding this outlying sample, there was a significant negative association between chr11:16810815:C:A carrier status and DHA ($P = 7.61 \times 10^{-3}$), omega-3 ($P = 7.62 \times 10^{-3}$), omega-3 % ($P = 3.45 \times 10^{-3}$) and a significant positive association with omega-6: omega-3 ($P = 1.44 \times 10^{-3}$) (Supplementary Figure 41, Supplementary Table 12).

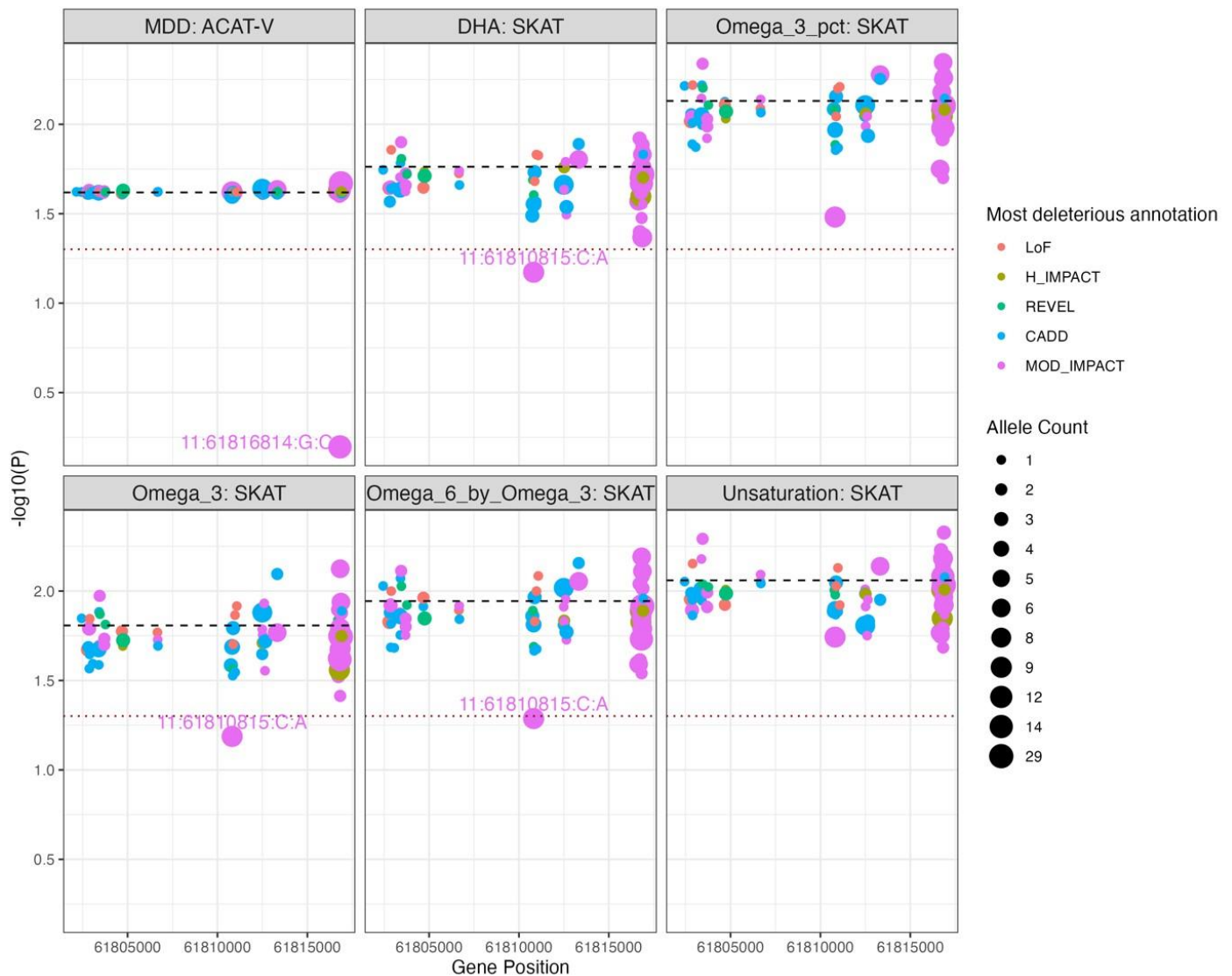


Figure 4.11 Leave one variant out analysis for prioritised variants in *FADS1* (Mask 5 AF < 0.01) with MDD and all metabolomic traits.

The LOVO analysis was conducted on the most significant burden association test per trait (labelled). The LOVO p-values are plotted against the variant positions. Each variant is coloured based on the most deleterious annotation (LoF = Loss of Function from LOFTEE, H_IMPACT = High Impact from VEP, REVEL = REVEL score > 0.5, CADD = Scaled CADD score > 20 and MOD_IMPACT = Moderate impact from VEP) assigned by

VEP. The size of each variable is based on the allele count in the UK Biobank MDD cohort. LOFTEE=loss-of-function transcript effect estimator, CADD=combined annotation dependent deletion, REVEL=rare exome variant ensemble learner, VEP=Variant Effect Predictor.

4.5 Discussion

This chapter uses whole genome sequencing data to investigate the role of rare potentially pathogenic variants in the colocalised *FADS* region identified in **Chapter 3**, on both MDD and omega-3 related measures. Following quality control, we identified 52,148 rare variants across all genes, with 1,428 prioritised as potentially deleterious based on various annotation criteria. Most of these variants were missense, with low correlations, and were prioritised based on the VEP impact score. All 1,428 and 961 prioritised variants were carried in 15,751 and 8,693 participants in the MDD cohort and metabolite cohorts respectively. In both cohorts, most of the carriers carried one prioritised variant in a single gene as alternate heterozygous, although there were a few instances of alternate homozygous carriers and carriers of multiple prioritised variants within and across genes. Traditional statistical tests found no significant association between carrying a prioritised variants and MDD, but significant associations were observed with metabolite traits for *TMEM258*, *FEN1* and *FADS3*. Aggregate variant testing (AVT) found a nominally significant association between prioritised variants in *FADS1* with MDD only. There was consistent, strong evidence for an association of prioritised variants in *TMEM258*, *MYRF* and *FEN1* with all the metabolite traits, whilst *FADS1* showed a small non-significant association. A LOVO analysis into prioritised variants in *FADS1* with MDD and metabolite traits indicated that these associations were driven by two distinct missense variants in Exon 1 (chr11:61816814:G:C) and Exon 5 (chr11:16810815:C:A) respectively. Finally, chr11:61816814:G:C carrier status showed a significant association with MDD, with null associations with any metabolite traits.

4.5.1 Minimal association of rare variation in *FADS1/2/3* genes with circulating LC-PUFAs

The association of common genetic variants at the *FADS* locus on LC-PUFAs is well established⁴⁶⁵⁻⁴⁶⁸. However, the common variants within this region exhibit high LD ($r^2 > 0.8$), making it a challenge to identify potentially causal genes which may underlie observed associations⁴⁶⁹. As *FADS1* and *FADS2* encode FADS1 and FADS2 enzymes which are directly involved in the LC-PUFA synthesis pathway (Figure 6.1), they represent a biologically plausible mechanism for genetic variation to affect LC-PUFA levels and are therefore usually hypothesized to be underlying associations of LCPUFAs at the *FADS* cluster. Our analysis showed that changes in the omega-3 related measures were not significantly associated with rare variation in *FADS1*, *FADS2* and *FADS3*. This may reflect different contributions of common and rare variation within these *FADS* genes on LC-PUFA levels.

A recent study demonstrated that humans have two common haplotypes in the *FADS* region, defined by 28 SNPs over a 38.9kb region, which are associated with dramatic differences in LC-PUFA synthesis⁴⁷⁰. Importantly, this study also demonstrated that there were no independent effects of rare variants over common variants in *FADS* on LC-PUFA levels⁴⁷⁰. Also, another study using WGS in multi-ethnic populations found associations between common *FADS* variants with LC-PUFAs but no evidence of increased burden of rare variants in any gene within this region on LC-PUFAs⁴⁷¹. Although this is further evidence that rare variants in the *FADS* genes are not associated with changes in LC-PUFAs, it does not support our finding of the association of LC-PUFAs with rare variation in nearby genes (*TMEM258*, *MYRF* and *FEN1*). However, this study is considerably less

powered (N = 11,840) and used MS (rather than NMR) to measure metabolites, making direct comparison difficult. A recent WES analysis which used NMR metabolomic data in UKB and the Estonian Biobank (N = 207,836) also found no association of LC-PUFA levels with rare variants in the *FADS1/2/3* genes but did observe a significant association of rare variants in *MYRF* and other genes in the *FADS* locus including *BEST1*, *DAGLA* and *RAB3IL1*⁴⁷². Notably, all the metabolite measures included in our analyses also had significant association of rare variants in various other genes across the genome, including *LPA* (Chr 6), *RRN3* (Chr 16), *TM6SF2* and *TOMM40* (Chr 19), demonstrating their polygenicity.

Our results indicate that rare variants in nearby genes such as *TMEM258*, *MYRF* and *FEN1* may play a role in circulating LC-PUFAs. A recent GWAS of omega-3 and omega-6 fatty acids in Hispanic and African American populations indicated that there were distinct genetic signals regulating LC-PUFAs within the *FADS* locus, with many fine-mapped associations outside of the *FADS1/2/3* genes, including *TMEM258* and *MYRF*⁴⁷³. They also found a significant association between *TMEM258* expression in whole blood with various LC-PUFAs alongside evidence of chromatin physical contact between *TMEM258* and *FADS1*. Taken together, the authors argue for the role of numerous genes in the *FADS* locus in regulating LC-PUFAs outside of the traditional *FADS* genes⁴⁷³. Overall, our analyses align with evidence that, unlike common variants, rare variants in *FADS1/2/3* genes do not primarily influence circulating LC-PUFAs, and that other genes within the locus may contribute to their regulation.

4.5.2 Distinct signals in *FADS* for MDD and metabolite traits

We observed a nominally significant association between rare variants in *FADS1* and MDD and metabolite traits. However, compared to other genes tested (*TMEM258*, *MYRF* and *FEN1*), the effect on the omega-3 related metabolites was negligible. Importantly, these genes which had strong associations with all the circulating metabolite measures had null associations with MDD. Together this indicates that there may be disparate signals in rare variants in this region associated with MDD (*FADS1*) and circulating LC-PUFAs (*TMEM258*, *MYRF* and *FEN1*) respectively. This could also indicate evidence against the causal role of circulating LC-PUFAs in MDD, as if there was a causal relationship, the strong associations observed with circulating LC-PUFAs (*TMEM258*, *MYRF* and *FEN1*) may be expected to be reflected in MDD.

One possibility is that the rare variants in *FADS1* associated with MDD have tissue-specific metabolic effects in brain tissue which would not be captured by circulating metabolite levels. Notably, the most recent GWAS of MDD in ~4.8 million participants identified *FADS1* as a high confidence gene alongside colocalisation of this association with *FADS1* expression in the dorsolateral prefrontal cortex¹⁸³, indicating that its role in MDD may be tissue-specific. Also, common variants in the *FADS* locus have been shown to have a tissue-specific impact on expression of *FADS1* and *FADS2*, which is highest in the brain and the adrenal gland⁴⁷⁴. Additionally, another study found evidence of significant interaction between common genetic variants in *FADS1*, *FADS2*, *FADS3*, *FEN1*, *MYRF* and *TMEM258* and dietary fatty acid supplementation on the levels of LC-PUFAs, which was also associated with tissue-specific expression of *FADS1*, *TMEM258* and *FEN1*⁴⁷⁵. Specifically, they observed a significant interaction between fatty acid

supplementation and *FADS1* expression in cerebellar tissue, whereas the significant interaction with *TMEM258* and *FEN1* was found in non-brain tissue⁴⁷⁵. This indicates that genetic variants in this region may influence the effects of fatty acid metabolism in different tissues in the body.

Another possibility is that the *FADS1* association with MDD may be specific to metabolic effects at different life periods depending on the *FADS1* expression over the life course. It is known that LC-PUFAs accumulate in the brain in the prenatal period, particularly in the third trimester, which coincides with an intensive period of neurite outgrowth and synaptic wiring⁴⁷⁶. As well, in-vitro studies of zebrafish have shown spatial and temporal specific expression of *FADS1* in the head area during embryogenesis⁴⁷⁷. Furthermore, LC-PUFAs and their downstream metabolites have been linked to several neurodevelopmental processes such as neurogenesis⁴⁷⁸, synaptogenesis⁴⁷⁹ and microglial inflammatory response⁴⁸⁰, which are all biological factors proposed to be important in the etiology of MDD⁴⁸¹⁻⁴⁸⁵. The role of rare variation in *FADS1* on these factors requires further analyses using brain transcriptome data at different neurodevelopmental stages.

4.5.3 *Driver variant for FADS1 association with MDD*

Although our results did indicate a nominal significant effect of rare variation in *FADS1* with MDD, this association did not survive correction for multiple testing. However, it is important to note that the analysis of rare variation in *FADS1* was substantially less powered than other genes due to the high proportion of missingness (N ~ 100,000 fewer participants). A recent meta-analysis performed burden-testing of rare coding variation across hundreds of diseases using exome sequencing data from UKB, the AoU research

program and the Mass General Brigham Biobank (MGB)⁴⁸⁶. This analysis did not find any significant associations of rare variation with MDD, or any significant burden of *FADS1* on any other disease tested⁴⁸⁷. However, this analysis applied the traditional burden test only, which assumes each variant within a tested set has an equal effect on the outcome⁴¹⁵. Importantly, the nominally significant association of rare variation in *FADS1* with MDD in these analyses was detected using the ACAT-V test, which is especially powerful compared to the burden test when there is sparsity in the true effects in the tested set of variants⁴²². Moreover, the LOVO analysis indicated that there was sparsity in the effects of tested variants, as the association was primarily being driven by a singular missense variant on Exon 1 (chr11:61816814:G:C, which is annotated to identifier *rs555414942*), initially prioritized due to a moderate impact predicted by VEP only (CADD score= 5.2, REVEL=0.017).

The Fishers' Exact test identified a significant association between carrying a single copy of the alternate allele at *rs555414942* and MDD, with exactly 50% of carriers within the sample also having MDD. This variant has been observed as alternate heterozygous only in participants in other large-scale sequencing datasets such as gnomAD, the Regeneron Genetic Center Million Exomes (RGC-ME) dataset⁴⁸⁸ and AoU⁴⁸⁹. Notably, these datasets show that the frequency of this variant is relatively similar across different ancestries (AF range: 1.18×10^{-4} to 8.69×10^{-5})

(<https://rgc/research.regeneron.com/me/rsid/rs555414942> and

<https://databrowser.researchallofus.org/snvsindels/11-61816814-G-C>). This genetic variation results in a C to G substitution at nucleotide position 116 (c.116C>G), in turn resulting in a Proline at amino acid position 39 being replaced by Arginine (p.Pro39Arg).

However, the clinical significance of this alteration is unclear with a singular entry on ClinVar⁴⁹⁰ denoting it as of Uncertain Significance (Accession: VCV002361985.3). Interestingly, this genomic position is multi-allelic, and another variation, chr11:61816814_G_T, which results in a C to A substitution (c.116C>A) and Proline replaced with Glutamine (p.Pro39Gln) was also prioritized due to VEP impact score (CADD score: 4.3, REVEL = 0.026). However, the LOVO analysis found that this variant was not a driver of the association of rare variation in *FADS1* and MDD. Therefore, this indicates that if this variant does have a functional impact on MDD, it is specific to the Proline to Arginine substitution at position 39.

4.5.4 Strengths

This study presents an in-depth examination of the role of rare variation in six genes at the *FADS* cluster on both MDD and circulating omega-3 related metabolomic traits. Although the availability of metabolomic and WGS data in large-scale cohorts is rapidly increasing, the UKB is currently unique in having joint availability of both these data with MDD phenotypic information at such a large scale, presenting a novel opportunity for a well-powered investigation between these factors. Furthermore, this analysis was informed by the findings of a colocalised signal in the *FADS* region between MDD and DHA in Chapter 1⁴⁴¹. This a-priori focus allows for a more in-depth examination of the effects of rare variants in this region on both MDD and omega-3 circulating metabolites which may be missed by broad phenome-wide and genome-wide approaches. Additionally, this study makes use of various cutting-edge variant annotation tools to prioritise potentially pathogenic rare variation within each gene^{491,492}. Finally, the study is strengthened by the incorporation of various burden tests and the LOVO analysis, which

helps to elucidate the genetic architecture underlying rare variant associations and ultimately aids their interpretation.

4.5.5 *Limitations*

This study also has various limitations. In keeping consistent with **Chapter 3**, this analysis was restricted to unrelated participants of European ancestry within UKB, which removed a substantial portion of individuals and rare variants to power these analyses. Also, the 2-step framework of REGENIE can account for different ancestries and relatedness, making this approach rather conservative⁴¹⁶. Therefore, future analyses on an unrestricted sample would gain power and boost discovery, whilst also enabling the effect of rare variation across different ancestries to be assessed. However, as the UKB predominantly consists of participants of European ancestry (~94.4%), it would still be limited in its power to do so. This links to the next limitation, which is that our analysis is conducted only on UKB, which is biased towards white, wealthy, highly educated individuals⁴⁹³. It is paramount to replicate our findings in other more representative cohorts to assess the robustness and generalisability of our findings across different sociodemographic populations. Although large cohorts combining WGS, metabolomic data and MDD phenotypic information are rare, there are several large cohorts with WGS and either MDD data⁴⁸⁹ or metabolomic⁴⁹⁴ across different populations. These different datasets could be used to assess the robustness of our findings with respect to the association of *rs555942141* with MDD and the association of rare variants in *TMEM258*, *MYRF* and *FEN1* and with circulating metabolites separately. Also, our analysis did not investigate the association of common and rare variation in these genes together and therefore is unable to assess the interplay of each. Further analyses could model both common and rare variation simultaneously to

understand more about how these factors may influence each other in their associations with MDD and circulating metabolomic levels. Furthermore, although we used various variant annotation tools, most of the prioritised variants were within coding regions, which reflects the increased knowledge base for annotating coding variants⁴⁹⁵. The incorporation of annotation tools specifically for non-coding⁴⁹⁵ and regulatory variants⁴⁹⁶ may capitalise more on the high resolution afforded by WGS. Finally, the metabolite measures available in UK Biobank are from the plasma samples from whole blood, which is not thought to be a functionally relevant tissue involved in the pathology of MDD, a brain-based disorder. Based on our results of differential signals for metabolites in the blood and MDD, it is vital to assess the role of rare variation in *FADS1* in MDD alongside tissue-specific multi-omics data⁴⁹⁷.

4.5.6 Summary

Overall, this study investigated the association between rare, potentially pathogenic variants in the *FADS* locus on both MDD and omega-3 related measures. Our findings support the presence of disparate signals of rare genetic variation in the investigated *FADS* region on MDD (*FADS1*) and circulating omega-3 related levels (*TMEM258*, *MYRF* and *FEN1*) respectively. Furthermore, our LOVO analysis revealed that the nominally significant association between *FADS1* and MDD was primarily being driven by a singular missense variant, *rs555942141*. Following validation of these findings in an external cohort, the integration of tissue specific multi-omics data to assess the functional impact of *rs555942141* and its potential role in MDD would be of interest.

4.6 Conclusion

This chapter directly builds upon observed evidence of a shared causal variant between an omega-3 fatty acid (DHA) and MDD located in the *FADS* region, identified in **Chapter 3**. Specifically, this analysis aimed to identify potential causal genes within this colocalised region (*FADS1*, *FADS2*, *FADS3*, *FEN1*, *MYRF*, *TMEM258*), by evaluating the associations of rare deleterious variants with both MDD and a range of circulating omega-3 related metabolite measures. Using various variant annotations, I identified 1,428 rare ($AF < 0.01$) potentially pathogenic/deleterious variants (“prioritised”), each carried in at least one participant in UKB ($N_{\text{total carriers}} = 15,751$). Aggregate variant testing (AVT) found strong associations between prioritised variants in *TMEM258*, *MYRF* and *FEN1* with a range of circulating omega-3 related measures, whilst no effect was seen in MDD. In contrast, there was a nominally significant signal for prioritised variants in *FADS1* with MDD, which appeared to be dependent on a single variant in Exon 1 (*rs555942141*). Interestingly, 50% of the carriers of *rs555942141* had lifetime MDD, revealing a statistically significant association. These findings indicate that regulation of omega-3 related measures in the *FADS* cluster by rare genetic variation is primarily influenced by genes outside of the core *FADS1/2/3* genes. Additionally, the disparate signals for MDD (*FADS1*) and circulating omega-3 related measures (*TMEM258*, *MYRF* and *FEN1*) indicates that a potentially causal role of these metabolites in MDD may be specific to certain tissues of developmental time points. Furthermore, our findings pinpoint a specific missense variant in *FADS1* which may have a potentially causal influence in MDD. Important further study includes the validation of our findings in an external sample, alongside assessing the interplay between common and rare variants at this locus in MDD.

In both **Chapter 3** and **Chapter 4** I have utilised metabolomics data to understand more about the underlying biological processes in MDD. Another way of investigating this is to evaluate the biological impact of antidepressant treatments. This approach can gain insight into potential therapeutic mechanisms and novel treatment strategies but may also help to identify factors that may contribute to more severe side effects in certain individuals. In addition, as more multi-omics studies of MDD are conducted, it is important to establish potential omics associations with antidepressant themselves, to disentangle MDD-relevant or antidepressant induced effects. In **Chapter 5**, I focus on learning more about the potential impact of antidepressant exposure on biological systems, specifically through the association with DNAm.

5 DNA methylation and antidepressant exposure

5.1 Introduction

Various neurobiological theories of MDD converge on the idea that altered gene expression may lead to sustained maladaptive changes in neural function. Additionally, the delayed therapeutic action of antidepressants may be due to slow adaptive changes to counteract this. Environmental factors, such as medication use, can influence gene expression through epigenetic mechanisms, the most studied of which is DNA methylation (DNAm). Also, there is experimental evidence that commonly prescribed antidepressants interact with DNAm enzymes in vitro and in vivo in animal models. However, large epidemiological studies of this relationship have been limited in size and scope due to the expense of measuring both DNAm at scale and measuring antidepressant exposure reliably. In this chapter, I aim to comprehensively assess associations between antidepressant exposure and DNAm in Generation Scotland (GS, N = 16,531), presenting a large epidemiological investigation of antidepressant exposure and DNAm.

I first ran a methylome-wide association study (MWAS) of self-reported (N = 16,531) and prescription-derived (N = 7,951) antidepressant exposure to assess the association of each CpG site individually. I then investigated potential confounding by MDD indication (subset samples to lifetime MDD cases) and any sex-specific effects (sex-stratified analyses) in the MWAS. Following this, I conducted regional analysis and a range of downstream tests to assess the enrichment of associated CpGs and mapped genes in specific tissue-expression profiles and gene ontologies. Also, using prescription-derived measures, I assessed the relationship between time in treatment and DNAm at

significantly associated CpGs. Finally, I calculated a methylation profile score (MPS) to represent DNAm associations with self-reported antidepressant exposure and meta-analysed its association with measured antidepressant use in eight external cohorts. The core methods involved in each of these analyses are presented in Section 5.2.

This study has been published in *Nature Communications* in 2025 and is included in full in Section 5.3. My role in this project was conducting all local analyses, co-ordinating the meta analysis with independent cohorts and drafting the manuscript. Code for this work is available at the GitHub repository: https://github.com/Elladavyson/Antidepressant_MWAS. The summary statistics for each MWAS and downstream analysis data are available in a Zenodo record: <https://zenodo.org/records/14216136>. Supplementary figures, tables and data can be found online accompanying the publication (<https://www.nature.com/articles/s41467024-55356-x>).

5.2 Methods

5.2.1 *Methylome-wide association study*

Methylome-wide association studies (MWAS) are analogous to GWAS, except the association tested is with the methylation M-value at each singular CpG rather than the genotype at each SNV. However, in contrast to genetic data, which is fixed and identical in every cell, DNAm is a dynamic process with distinct signatures in different cell types which can confound MWAS and lead to spurious associations^{498,499}. For DNAm measured from peripheral blood samples (as in GS), the DNAm is measured from a heterogeneous collection of different cell types, each with distinct DNAm patterns^{499,500}. Therefore, the cell-type composition (CTC) of a sample is important as it often correlates with cell-type specific DNAm markers which can be associated with and therefore confound various traits and diseases³⁶⁵. Different statistical algorithms can be used to predict the CTC of blood samples from DNAm, using a predefined reference DNAm profile of each cell type; CTC can then be included in a model as a fixed covariate⁵⁰⁰. However, this approach does not correct for the residual CTC variation not captured by reference panels or any additional unobserved confounders which can lead to correlations between distal CpGs³⁶⁵. Alternatively, reference-free methods which usually involve fitting surrogate variables or principal components as covariates can be more effective at accounting for uncharacterised confounders³⁶⁵. The Mixed-linear-model Omics Association (MOA) model, from the OmicS-data-based Complex trait Analysis (OSCA) software, is a reference-free method which aims to account for unobserved confounding and correlations between CpG sites to provide unbiased estimates of CpG-trait associations³⁶⁵.

5.2.1.1 Mixed-linear-model Omics based Analysis (MOA) model

In summary, the MOA model fits a correlation matrix of all distal CpG probes, called an omics-relatedness matrix (ORM), in the model as a random effect, successfully accounting for unobserved confounding effects and correlations between distal CpG probes. The MOA model is shown in Equation 5.1, where y is a vector of phenotype values, w_i is a vector of standardised DNAm values at probe i , b_i is the estimated fixed effect of probe i , C is a matrix of known covariate values, β is a matrix of fixed effects for each covariate, W is a matrix of standardised DNAm measures of probe i and all distal probes (> 50bp), u is the joint effect of all probes and ϵ is a vector of residual errors.

$$y = w_i b_i + C\beta + Wu + \epsilon \quad \text{Equation 5.1}$$

Various simulations have shown the MOA model to be powerful at controlling the false positive rate³⁶⁵. The MOA model assumes that the random effects of all distal probes on the trait follow a single distribution. If there are regions of CpGs much more strongly associated with a phenotype than others, extensions of the MOA model can be applied which fit the random effects with multiple distributions³⁶⁵. In this analysis the phenotype was GRM-corrected antidepressant exposure (generated using best-linear unbiased prediction described below) and covariates included age, sex, AHR probe M-values to proxy for smoking status and predicted monocyte and lymphocyte cell proportions. The null hypothesis ($b_i=0$) can then be tested by a two-sided t-test. The threshold for significance of a CpG-trait association was set at 9.42×10^{-8} , as was recently recommended for case-control MWAS analyses³³¹.

5.2.1.2 Best Linear Unbiased Prediction (BLUP)

As GS is a family-based cohort, there is a high degree of relatedness between individuals which could bias associations. To account for this, I regressed the antidepressant exposure phenotypes on a GRM using the Best Linear Unbiased Prediction (BLUP) tool from the GCTA software⁵⁰¹. The BLUP method was developed to estimate random effects in MLMs and has widely been used in agricultural science to quantify the genetic merit (“breeding value”) of an organism for a certain phenotype⁵⁰². In GCTA, BLUP models all SNVs in a given GRM together and can provide an unbiased estimate of the total genetic effect on a phenotype (here, antidepressant exposure)⁵⁰¹. The residuals from this model represent phenotypic variance not explained by genetic effects modelled in the GRM. Accordingly, I used the residuals as the phenotype for the MWAS, as these represent variation in antidepressant exposure independent of modelled genetic effects.

5.2.2 *Differentially methylated regions*

Commonly used commercial DNAm arrays only cover a fraction of CpG sites in the genome and have therefore been designed to strategically cover specific regions, such as gene enhancers, promoters, transcription factor binding sites and regions previously linked to disease mechanisms^{503,504}. There is a complex correlational structure between DNAm at CpG sites captured by commonly used Illumina arrays, and often single CpG trait associations are clustered together in differentially methylated regions (DMR). Searching for associations of DNAm at the level of genomic regions rather than individual CpGs can improve the specificity (reduced false positive rate) of MWAS⁵⁰⁵. A variety of algorithms have been proposed to identify DMRs, such as “*bumphunter*”⁵⁰⁵, “*DMRcate*”⁵⁰⁴

and “*dmrff*”⁵⁰⁴, most of which combine MWAS summary statistics (in different ways) from nearby CpGs to give an overall statistic for a given region. I used the “*dmrff*” approach as this was shown to be more effective at controlling false positives and more computationally efficient than aforementioned methods⁵⁰⁴. For a given region, *dmrff* runs an IVW meta-analysis of MWAS beta and standard error estimates, adjusting for estimate uncertainty and the correlational structure between probes (using a correlation matrix). The test-statistic for *dmrff* is B/S, defined using Equation 5.2 - Equation 5.4), that follows a standard normal distribution under the null hypothesis. Here, β and σ are vectors of MWAS effect estimates and standard errors for each CpG respectively, ρ is a correlation matrix of all CpG sites and $\mathbf{1}$ represents a vector of 1's.

$$\Omega = \sigma\sigma^T \times \rho \quad \text{Equation 5.2}$$

$$B = (\mathbf{1}^T \Omega^{-1} \mathbf{1})^{-1} \mathbf{1}^T \Omega^{-1} \beta \quad \text{Equation 5.3}$$

$$S = \sqrt{((\mathbf{1}^T \Omega^{-1} \mathbf{1})^{-1})} \quad \text{Equation 5.4}$$

Candidate genomic regions to test are defined as those with sets of CpGs with at most 500bp between each site, with nominally significant MWAS p-values (< 0.05) and effect estimates all in the same direction. The test statistic is calculated for each region, and sub-regions which are greedily selected based on the most extreme p-value within the region which is non-overlapping with previous sub-regions. Multiple-testing adjustment in

dmrff is performed using a Bonferroni correction. In this analysis, I classed DMRs achieving Bonferroni-corrected p-value < 0.05 to be statistically significant.

5.2.3 *Enrichment (Hypergeometric tests)*

Following the MWAS analysis, I extracted the top 100 CpGs and their annotated genes to test for potential enrichment in specific tissues and/or specific gene ontologies to learn more about the functional impact of top methylation signals. All the enrichment analyses were based on a hypergeometric test, a statistical method which can test the probability of an observed overlap between two gene-lists, given a background set of genes⁵⁰⁶. Various gene-sets have been created to represent tissue-specific signatures or genes linked to specific biological processes. The hypergeometric test can be used to assess whether there is an overrepresentation of genes linked to a particular gene-set within a list of priority genes (here the genes linked to the top 100 CpGs), which can inform on potentially impacted biological processes and tissues. Equation 5.5 and Equation 5.6 shows how the probability of the overlap is calculated, where m and n are the number of genes in the background gene set in/not in a specific gene-ontology respectively, k is the number of genes in the priority list and x is the number of genes within the priority list which overlap with the gene-set of interest.

$$p(x) = \frac{\binom{m}{x} \binom{n}{k-x}}{\binom{m+n}{k}} \quad \text{Equation 5.5}$$

$$\binom{m}{n} = \frac{m!}{x!(m-x)!}$$

Equation 5.6

Specifically, I used the “GENE2FUNC” tool in the functional mapping and annotation (FUMA)⁵⁰⁷ web-tool to test the enrichment of priority genes from the antidepressant MWAS with differentially expressed gene sets (DEGs). DEGs represent genes which are significantly up or down-regulated in a specific tissue ($P \leq 0.05$ and log fold change ≥ 0.58), calculated for 54 specific tissues in the GTEx dataset³⁷⁰. Furthermore, I also tested the enrichment of priority genes with expert-curated synaptic gene ontologies hosted on SynGo (<https://www.syngoportal.org/>)⁵⁰⁸. Finally, I tested the enrichment of priority genes with a host of gene-ontologies related to different biological processes identified in the “*msigdb*” database using the “*gsameth*” function in the “*missMethyl*” R package⁵⁰³. This uses a Wallenius central hypergeometric test and explicitly accounts for the number of CpG sites per gene on the EPIC arrays, a factor which can bias gene-set analyses⁵⁰³.

5.2.4 Methylation profile score (MPS) analysis

5.2.4.1 Calculating the score

Methylation scores have been calculated for a range of traits, including biological age (“epigenetic clocks”)⁵⁰⁹, smoking status (“EpiSmokEr”)⁵¹⁰, and circulating protein levels (“EpiScores”)⁵¹¹. A common way to calculate methylation scores is through the application of penalised regression models, which can add regularisation penalties and distinguish non important and important features for the prediction of a trait of interest. A least absolute shrinkage and selection operator (LASSO) penalised regression model was

used to identify CpGs and their corresponding weights to constitute a MPS of self-reported antidepressant exposure. A LASSO model simultaneously fits all CpGs to the model and applies a L1 regularisation penalty (λ) so that coefficients of CpGs which do not contribute to the model are shrunk towards, or to zero. The penalty term is a model hyperparameter, which can be tuned using cross-validation. I performed LASSO regression on the self-reported antidepressant exposure phenotype. First, I regressed the phenotype on GRM (using BLUP) and all other covariates which were used in the MOA MWAS model. The LASSO model was performed on all CpGs that were in both the 450K and EPIC Illumina arrays (N = 365,912) and ten-fold cross-validation was applied. All CpGs with a non-zero weight (N = 212) were then selected to constitute the MPS of antidepressant exposure, and their corresponding regression coefficients were selected as weights. This data was then supplied to external cohorts for the calculation of the MPS in each individual as the weighted sum of the standardised DNAm M-values at the 212 selected probes.

5.2.4.2 Meta-analysis

The MPS was calculated for each individual from eight external cohorts: Finn Twin Cohort (FTC)^{512,513}, Study of Health in Pomerania (SHIP-Trend)⁵¹⁴, Lothian Birth Cohort 1936 (LBC1936)⁵¹⁵, FOR2107⁵¹⁶, Netherlands Twin Register (NTR)⁵¹⁷, Avon Longitudinal Study of Parents and Children (ALSPAC)⁵¹⁸, Munich Antidepressant Response Study-Unipolar Depression Study (MARS-UniDep)⁵¹⁹ and the Environmental Risk (E-Risk) Longitudinal Twin Study⁵²⁰, alongside a prospective sample of GS GS: Stratifying Depression and Resilience Longitudinally (STRADL)³⁶⁹. Each cohort then tested the

association of the derived MPS with measured antidepressant exposure using either a binomial generalised linear model (unrelated cohorts), binomial generalised mixed linear model (family-based cohorts), or binomial generalised estimation equation model (twin-cohorts). All models included the binary antidepressant exposure as the outcome, and age, sex, white-blood cell proportions or counts, AHRR probe M-values and technical factors (batch/set) as covariates. The effect estimates and standard errors from each associational model were then meta-analysed to assess the overall association between the MPS and antidepressant exposure using the '*meta*' R package. Specifically, I used a random-effects meta-analysis which assumes that true effect estimates from each individual study vary due to heterogeneous populations and that given an infinite number of studies, would follow a standard normal distribution (Figure 5.1)⁵²¹.

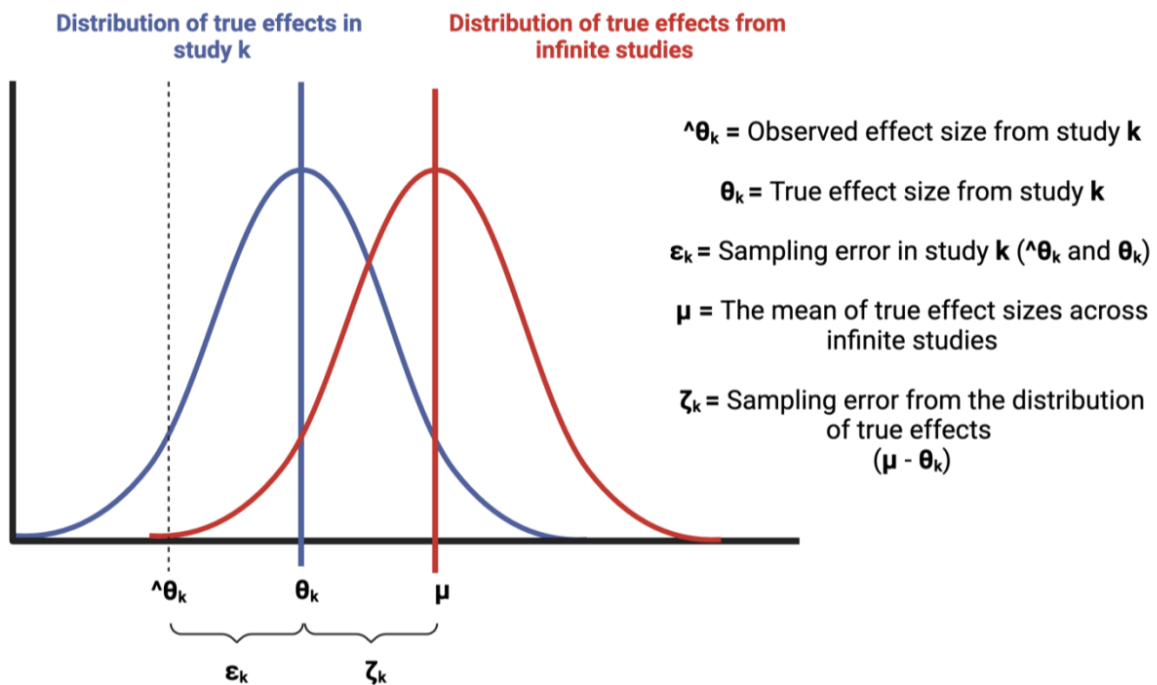


Figure 5.1 Random effects meta-analysis.

The random effects meta-analysis assumes that the true effect size in each study may vary and follow a standard normal distribution. Therefore, this meta-analysis estimates the pooled true effect (μ) considering both the sampling error in each individual study (ε_k) and variation between-studies (ζ_k). This figure was made using Biorender.com.

This assumes that the observed effect estimate from each study k (θ_k) is the true effect size (μ) plus variance from sampling error within each study itself (ε_k) plus variance between the true effect size (across infinite studies) and the true effect size in study k (ζ_k) (Equation 5.7).

$$\hat{\theta}_k = \mu + \zeta_k + \varepsilon_k \quad \text{Equation 5.7}$$

The weight given to each study k (w_k) is a function of the variance of its estimated effect (σ_k^2) and the variance of true effect estimates between studies (τ^2) (Equation 5.8).

$$w_k = \frac{1}{\sigma_k^2 + \tau^2} \quad \text{Equation 5.8}$$

The variance of true effect estimates across studies (τ^2) can be estimated using a Dersimonian-Laird estimator⁵²², which uses the Cochran Q statistic to measure the heterogeneity between studies (Equation 5.9)⁵²². The parameter θ is the weighted mean of effect estimates under a fixed-effects meta-analysis.

$$Q = \sum_{k=1}^k \frac{1}{\sigma_k^2} (\hat{\theta}_k - \hat{\theta}_{fixed-effects})^2 \quad \text{Equation 5.9}$$

Following the estimation of τ^2 and the calculation of individual weights (w_k), the pooled effect estimate in the random effects meta-analysis can be calculated (Equation 5.10)⁵²¹.

$$\hat{\theta}_{random-effects} = \frac{\sum_{k=1}^k \hat{\theta}_k w_k}{\sum_{k=1}^k w_k} \quad \text{Equation 5.10}$$

5.3 Insights from a methylome-wide association study of antidepressant exposure

Insights from a methylome-wide association study of antidepressant exposure

Received: 21 May 2024

Accepted: 9 December 2024

Published online: 24 February 2025



E. Davyson^{1,2}, X. Shen¹, F. Huider^{3,4,5}, M. J. Adams¹, K. Borges¹, D. L. McCartney^{1,2}, L. F. Barker⁶, J. van Dongen^{3,4,5,7}, D. I. Boomsma^{3,4,7}, A. Weihs^{8,9}, H. J. Grabe^{8,9}, L. Kühn⁸, A. Teumer^{8,10}, H. Völzke^{10,11}, T. Zhu^{12,13}, J. Kaprio¹², M. Ollikainen^{12,13}, F. S. David^{14,15}, S. Meinert^{16,17}, F. Stein^{15,18}, A. J. Forstner^{14,19,20}, U. Dannlowski¹⁶, T. Kircher^{15,18}, A. Tapuc^{21,22}, D. Czamara²², E. B. Binder²², T. Brückl²², A. S. F. Kwong^{1,23}, P. Yousefi^{23,24,25}, C. C. Y. Wong²⁶, L. Arseneault²⁶, H. L. Fisher^{26,27}, J. Mill²⁸, S. R. Cox²⁹, P. Redmond²⁹, T. C. Russ^{1,30,31}, E. J. C. G. van den Oord³², K. A. Aberg³², B. W. J. H. Penninx³³, R. E. Marioni², N. R. Wray^{6,34} & A. M. McIntosh¹✉

This study tests the association of whole-blood DNA methylation and antidepressant exposure in 16,531 individuals from Generation Scotland (GS), using self-report and prescription-derived measures. We identify 8 associations and a high concordance of results between self-report and prescription-derived measures. Sex-stratified analyses observe nominally significant increased effect estimates in females for four CpGs. There is observed enrichment for genes expressed in the Amygdala and annotated to synaptic vesicle membrane ontology. Two CpGs (cg15071067; *DGUOK-AS1* and cg26277237; *KANK1*) show correlation between DNA methylation with the time in treatment. There is a significant overlap in the top 1% of CpGs with another independent methylome-wide association study of antidepressant exposure. Finally, a methylation profile score trained on this sample shows a significant association with antidepressant exposure in a meta-analysis of eight independent external datasets. In this large investigation of antidepressant exposure and DNA methylation, we demonstrate robust associations which warrant further investigation to inform on the design of more effective and tolerated treatments for depression.

Major Depressive Disorder (MDD) is predicted to become the leading cause of disability worldwide by 2030¹, partly due to the limitations of current treatments². Although antidepressants are commonly prescribed effective treatments³, they prove to be ineffective in a high proportion of cases, with an estimated 40% of those presenting with MDD developing treatment-resistant depression^{4,5}. Furthermore, many treatments are commonly accompanied by side effects, including weight changes, dizziness, fatigue and sexual dysfunction². There is

a need for more effective and better-tolerated antidepressant treatments and to target existing treatments to those most likely to respond. Advances are hampered by poor mechanistic understanding of both MDD itself^{6,7} and how currently prescribed antidepressants lead to therapeutic effects⁸.

The mechanism of currently prescribed antidepressants is incompletely understood. Initial theories surmised that their therapeutic effects were primarily due to the inhibition of monoamine reuptake in

the synapse, leading to an increase in monoamine concentrations in the brain⁹. However, antidepressant treatment has a delayed onset for symptomatic improvement, which does not reflect the immediate effect on monoamine levels¹⁰. This casts doubt on the simple role of monoamines as a causal factor in MDD^{7,8,11}, although other experimental paradigms continue to suggest their importance¹². Another prominent theory suggests that antidepressants exert their therapeutic effects by increasing brain-derived neurotrophic factor (BDNF), leading to synaptic remodelling¹³ and enhanced neuronal plasticity^{8,14}. A recent study found evidence of antidepressants binding directly the BDNF receptor (neurotrophic tyrosine kinase receptor 2; TRKB), proposing this as a potential mechanism of action independent of changes in monoamine concentrations¹⁵. Additionally, sex differences in MDD risk^{16,17}, antidepressant efficacy and side-effects are well-documented¹⁸, which may reflect sex-specific differences in neuronal circuitry¹⁹. However, precise mechanisms of these differences are unclear.

DNAm, the addition of a methyl group at a cytosine-phosphate-guanine (CpG) site, regulates gene expression and impacts cellular function^{20,21}. The evidence for the effect of antidepressants on DNA methylation (DNAm) is growing^{22,23}. In vitro studies found that the antidepressant paroxetine interacted with DNA methyltransferase (DNMT), a key enzyme involved in DNAm²⁴. Furthermore, studies of chronically stressed rodent models have found that stress-induced DNAm and behavioural changes are reversed through both antidepressant treatment²⁵ and DNMT inhibitors²⁶. In 2022, Barbu et al.²⁷ performed a methylome-wide association study (MWAS) of self-reported antidepressant exposure in a subset of participants in Generation Scotland (GS, $N=6428$) and the Netherlands Twin Register (NTR, $N=2449$), and identified altered DNAm near to genes involved in the innate immune response in those exposed to antidepressants²⁷. As self-report measures may be unreliable due to volunteer recall bias, a poor understanding of the medication nosology, and non-disclosure^{28–30}, Barbu et al.²⁷ also performed an MWAS of antidepressant exposure based on recorded antidepressant prescriptions in the last 12 months. However, this assumes continuous treatment, potentially overestimating exposure due to general low adherence to antidepressant medication³¹. The calculation of active treatment periods from consecutive prescribing events could provide a potentially more reliable identification of antidepressant exposure³². Here, adherence to the medication is assumed given regular prescription dispensations at an expected frequency (given their amount and dosage), rather than a singular prescribing event.

In our study, we build upon previous analyses by Barbu et al.²⁷ by analysing a larger sample of GS ($N=16,531$), and by estimating active treatment periods from prescribing records to identify those exposed to antidepressants at DNAm measurement. First, an MWAS was performed on both the self-report and prescription-derived measures of antidepressant exposure. Second, the MWAS analyses were restricted to MDD cases only to assess potential confounding by MDD, and sex-stratified analyses were conducted to investigate any sex-specific effects. Third, functional follow-up analysis of differentially methylated CpG sites was performed. Fourth, we investigated the enrichment of top CpGs in GS and an independent MWAS conducted in the Netherlands Study of Depression and Anxiety (NESDA). Fifth, the relationship between time in treatment and DNAm at significant CpG sites was investigated. Finally, a methylation profile score (MPS) for self-report antidepressant exposure was trained in GS and tested for an association with antidepressant exposure in eight independent external datasets: Finn Twin Cohort (FTC), Study of Health in Pomerania (SHIP-Trend), Lothian Birth Cohort 1936 (LBC1936), FOR2107, NTR, Avon Longitudinal Study of Parents and Children (ALSPAC), Munich Antidepressant Response Study-Unipolar Depression Study (MARS-UniDep) and the Environmental Risk (E-Risk) Longitudinal Twin Study, alongside a prospective sample of GS: Stratifying Depression and Resilience Longitudinally (STRADL) (Fig. 1).

Results

The demographics of both antidepressant exposure phenotypes in the GS are shown in Table 1. To assess demographic differences between the exposed and unexposed group, we used Welch's independent samples t-tests for continuous variables (age, BMI) and chi-squared tests for categorical variables (sex, smoking status and lifetime MDD status).

In both self-report and prescription-derived measures, the antidepressant exposed group were significantly older ($t_{\text{self-report}}(1970) = 10.6$, $p_{\text{self-report}} = 1.01 \times 10^{-25}$, $t_{\text{prescription-derived}}(1180) = 7.35$, $p_{\text{prescription-derived}} = 3.63 \times 10^{-13}$) and had significantly higher BMI measurements ($t_{\text{self-report}}(1700) = 11.4$, $p_{\text{self-report}} = 5.45 \times 10^{-29}$, $t_{\text{prescription-derived}}(980) = 9.82$, $p_{\text{prescription-derived}} = 8.69 \times 10^{-22}$). Additionally, both phenotypes showed that the antidepressant-exposed group had significantly different proportions of current, former and never smokers ($\chi^2_{\text{self-report}}(2) = 160$, $p_{\text{self-report}} = 1.69 \times 10^{-35}$, $\chi^2_{\text{prescription-derived}}(2) = 114$, $p_{\text{prescription-derived}} = 1.49 \times 10^{-25}$), a significantly higher proportion of females ($\chi^2_{\text{self-report}}(1) = 214$, $p_{\text{self-report}} = 1.53 \times 10^{-48}$, $\chi^2_{\text{prescription-derived}}(1) = 129$, $p_{\text{prescription-derived}} = 7.62 \times 10^{-30}$) and significantly higher proportion of those with lifetime-MDD ($\chi^2_{\text{self-report}}(1) = 2170$, $p_{\text{self-report}} < 1 \times 10^{-320}$, $\chi^2_{\text{prescription-derived}}(1) = 1450$, $p_{\text{prescription-derived}} = 2.72 \times 10^{-318}$).

Methylome-wide association studies

The self-report MWAS (Fig. 2A, Table 2) and prescription-derived MWAS (Fig. 2B, Supplementary Table 1) identified seven and four hypermethylated CpGs respectively, in those exposed to antidepressants (Supplementary Figs. 1, 2). The effect estimates from all CpGs in both analyses were significantly correlated ($R=0.54$, $p < 2.2 \times 10^{-16}$) (Supplementary Fig. 3).

In the MDD-subgroup self-report MWAS, only cg08527546 exhibited a significant association with antidepressant exposure ($\beta = 0.050$, $p = 3.57 \times 10^{-8}$); no CpGs were significantly associated in the MDD-subgroup prescription-derived MWAS (Supplementary Table 2, Supplementary Fig. 4). For both phenotypes, there was a significant correlation between CpG effect estimates in the full and MDD-subgroup analyses ($R_{\text{self-report}} = 0.57$, $p_{\text{self-report}} < 2.2 \times 10^{-16}$, $R_{\text{prescription}} = 0.43$, $p_{\text{prescription}} < 2.2 \times 10^{-16}$) (Supplementary Fig. 5). Notably, restricting the analyses to MDD cases resulted in an average 2.5-fold and 2.1-fold increase in the self-report and prescription-derived effect sizes of the significant CpGs respectively (Fig. 2C).

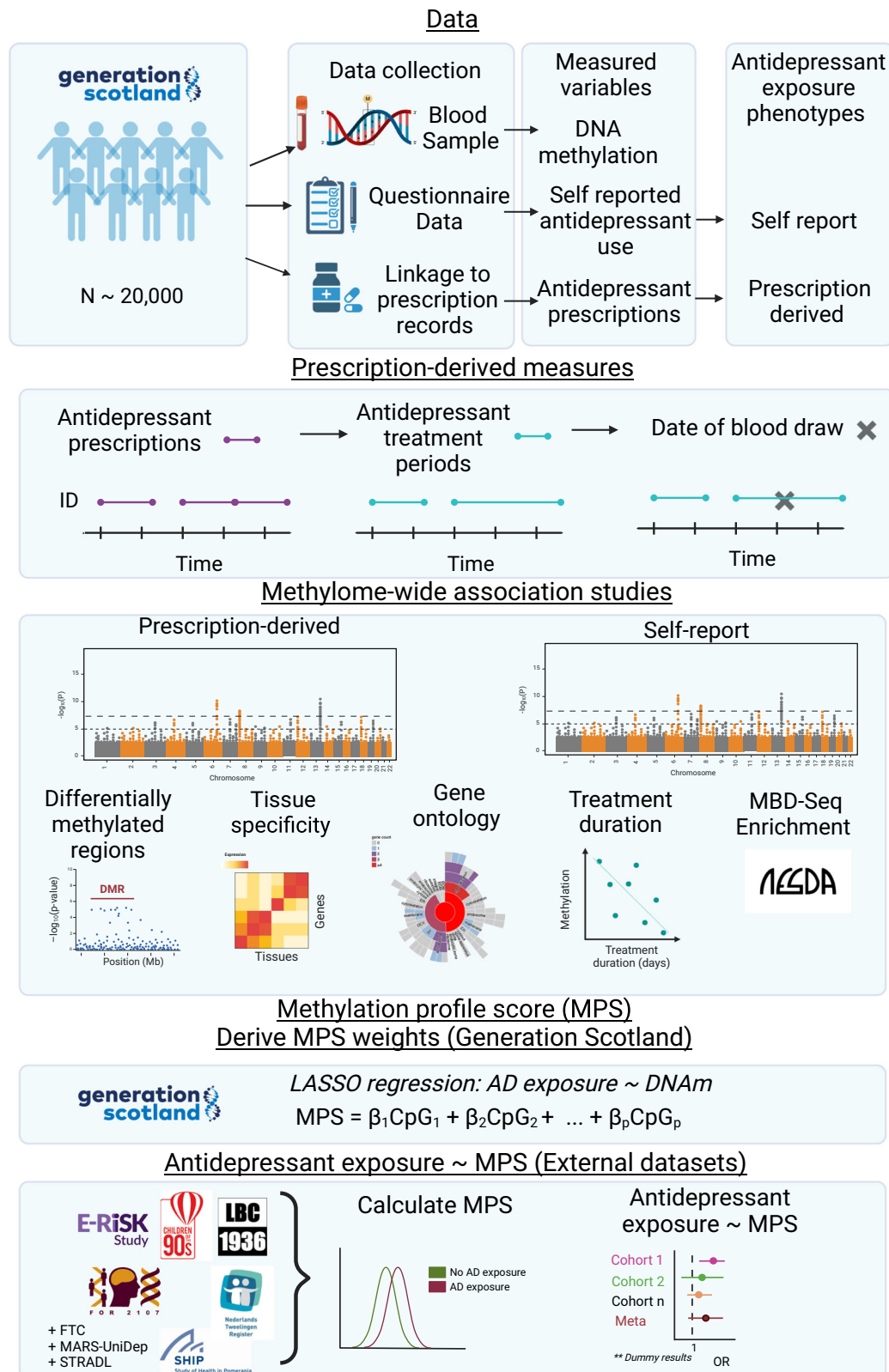
In the sex-stratified analyses, two CpGs were significantly associated with self-report antidepressant exposure in females, cg26277237 ($\beta = 0.030$, $p = 1.77 \times 10^{-10}$) and cg02183564 ($\beta = 0.023$, $p = 4.29 \times 10^{-9}$) (Supplementary Table 3) and there were no significant associations with antidepressant exposure in males (Supplementary Fig. 6). The effect sizes of CpGs in the male-only and female-only analyses were significantly correlated ($R=0.039$, $p < 2.2 \times 10^{-16}$) (Supplementary Fig. 7). Of the eight CpG sites significantly associated with antidepressant exposure in the overall analysis, all demonstrated a larger effect size in females (Supplementary Figs. 7–8). Four of these sites showed nominally significant sex differences (cg26277237; $p = 0.0081$, cg02183564; $p = 0.0132$, cg08907118; $p = 0.0335$, cg03222540; $p = 0.0305$) (Supplementary Data 1).

Differentially methylated regions

There were 719,506 candidate regions considered in the analysis. The self-report MWAS had one significant DMR ($\beta = 0.096$, $p_{\text{adj}} = 4.98 \times 10^{-3}$) (Supplementary Data 2), consisting of two CpGs (cg01964004 and cg15071067), which maps to deoxyguanosine kinase antisense RNA1 (*DGUOK-AS1*) (Supplementary Fig. 9). The prescription-derived MWAS identified no significant DMRs (Supplementary Data 3).

Functional Annotation

The most significant 100 CpGs from the self-report and prescription-derived MWAS were annotated to 77 and 83 genes respectively



(Supplementary Data 4-7). There was a significant overlap between both the CpG-sets ($N_{\text{overlap}} = 17$, $p = 1.95 \times 10^{-48}$, Supplementary Fig. 10) and the gene-sets ($N_{\text{overlap}} = 16$, $p = 1.3 \times 10^{-25}$, Supplementary Fig. 11). The self-report gene-set was significantly enriched for the genes expressed in the amygdala ($p_{\text{FDR}} = 0.043$,

Supplementary Data 8, Supplementary Fig. 12), whilst the prescription-derived gene-set had no significant enrichment (Supplementary Data 9, Supplementary Fig. 13). There was significant enrichment of the self-report gene-set in GO:0008021; “synaptic vesicle membrane” ($p_{\text{FDR}} = 0.030$, Supplementary Figs. 14-16, Supplementary

Fig. 1 | Study design of investigating antidepressant exposure and DNA methylation. Data: Participants in Generation Scotland (GS) provided blood samples from which DNA methylation was measured. Their antidepressant exposure status was measured using both self-report questionnaires and prescription-derived measures. Prescription-derived measures: Repeated regular prescriptions over time (X axis) for antidepressants (purple bars) are merged to form active antidepressant treatment periods (blue bars). Individuals in an active treatment period at the time of blood sample (black cross) are classed as antidepressant exposed. Methylome-wide association studies: A methylome-wide association study (MWAS), subsequent

regional analysis and functional annotation was performed for both measures of antidepressant exposure. Additionally, an enrichment analysis was done using MBD-Sequencing data in the Netherlands Depression and Anxiety (NESDA) cohort for the self-report antidepressant exposure. Methylation Profile Score: Weights for a methylation profile score (MPS) of self-reported antidepressant exposure was calculated in GS using a least absolute shrinkage and selection operator (LASSO) model. Eight independent datasets then tested the association of this MPS with self-reported antidepressant exposure. Created in BioRender. Davyson, E. (2024) <https://BioRender.com/s43g100>.

Tables 4-5, Supplementary Data 10-11). There was no significant enrichment for any GO:Biological Processes pathways tested for either gene set (Supplementary Figs. 17-18, Supplementary Data 12-13).

Enrichment analysis: Netherlands Study of Depression and Anxiety

We tested whether top findings from our self-reported MWAS on all participants were also more likely to be among the top findings in an independent MWAS of antidepressant exposure in the Netherlands Depression and Anxiety (NESDA) cohort³³. Notably, all participants in this cohort had a recent (~6 months) MDD DSM-IV diagnosis, obtained using the Composite International Diagnostic Interview, including single-episode and recurrent MDD (Supplementary Methods). DNAm was assayed on whole-blood samples from 812 MDD patients, 398 of which reported using one or more antidepressants, using methyl-CG binding domain sequencing (MBD-Seq)³⁴⁻³⁶. Data quality control and the MWAS was performed using RaMWAS³⁷ (Supplementary Methods). Enrichment tests were performed to assess the top findings from GS and NESDA MWAS. The enrichment analyses were performed using the ‘shiftR’ package³⁸ that performs circular permutations to account for having correlated methylation levels between CpG sites. For both MWAS, CpGs were filtered to those which showed a concordant direction of effect and three thresholds (0.1, 0.5 and 1%) were used to define top (by *p*-value) findings. We corrected for this “multiple testing” by using the same thresholds in the permutations and selecting the most significant result from each permutation to generate the “empirical” null distribution. Results suggested a small (OR: 1.39) but significant (*P* < 0.042) enrichment between the top 1% of results from both MWAS’ (Supplementary Data 14).

Methylation differences by time in treatment

To investigate the relationship of DNAm at significantly associated CpGs (*n* = 8) with the length of antidepressant exposure, a two-tailed Spearman correlation test was performed between the DNAm levels and time in current treatment for those within a treatment period (*N* = 863). Two probes, cg15071067 (*DGUOK-AS1*) and cg26277237 (*KANK1*), showed a significant correlation between methylation and time in treatment (cg15071067: *p* = 0.085, *p* = 0.012, cg26277237: *p* = 0.087, *p* = 0.011) (Supplementary Figs. 19-21, Supplementary Table 6), with the same direction of effect found in the MWAS.

Methylation profile score

There were 212 CpGs identified by the LASSO regression model (Supplementary Figs. 22-24, Supplementary Data 15), used to calculate the MPS in external cohorts (Fig. 3A, Supplementary Figs. 25-34). All cohorts bar one (NTR) showed a positive relationship between antidepressant MPS and antidepressant exposure ($\beta_{\text{FTC}} = 0.156$, $\beta_{\text{SHIP-Trend}} = 0.134$, $\beta_{\text{STRADL}} = 0.149$, $\beta_{\text{LBC1936}} = 0.228$, $\beta_{\text{FOR2107}} = 0.349$, $\beta_{\text{MARS}} = 0.263$, $\beta_{\text{ALSPAC}} = 0.170$, $\beta_{\text{ERISK}} = 0.342$, $\beta_{\text{NTR}} = -0.031$) (Supplementary Fig. 35, Supplementary Table 7). Nagelkerke’s pseudo *R*² estimates ranged from 1.11×10^{-3} (NTR) to 0.03 (LBC1936) (Fig. 3B). The random-effects meta-analysis (Fig. 3C) found a significant association between antidepressant exposure and the MPS (pooled β [95%CI]: 0.196 [0.105, 0.288],

Table 1 | Demographics and structured clinical interview of the DSM (SCID) diagnoses of antidepressant exposed and unexposed individuals using the prescription-derived and self-reported antidepressant exposure phenotypes in Generation Scotland

	Prescription-derived phenotype		Self-report phenotype	
	Unexposed	Exposed	Unexposed	Exposed
<i>N</i>	7090	861	15028	1508
Age: M (SD)	47.6 (15.2)	51.1 (12.7)	46.6 (14.9)	50.2 (12.4)
BMI: M (SD)	26.4 (4.78)	28.5 (6.11)	26.4 (5.01)	28.2 (6.05)
Sex				
Female: <i>N</i> (%)	3747 (53)	631 (73)	8556 (57)	1154 (77)
Male: <i>N</i> (%)	3343 (47)	230 (27)	6467(43)	354 (23)
Smoking behaviours				
Current: <i>N</i> (%)	963 (14)	214 (25)	2448 (16)	402 (27)
Former: <i>N</i> (%)	1942 (27)	273 (32)	4183 (28)	481 (32)
Never: <i>N</i> (%)	3913 (55)	328 (38)	8114 (54)	576 (38)
Pack years: M (SD)	6.3 (13.1)	12 (17.1)	6.55 (13.3)	11.6 (16.9)
Structured Clinical Interview for DSM IV (SCID)				
No Major Disorder: <i>N</i> (%)	6514 (92)	373 (43)	12704 (85)	556 (37)
Bipolar Disorder	9 (0.1)	9 (1)	29 (0.2)	23 (1.5)
Single Episode MDD <i>N</i> (%)	281 (4)	137 (16)	880 (6)	260 (17)
Recurrent MDD: <i>N</i> (%)	131 (2)	243 (28)	622 (4)	506 (34)
MDD cases (Recurrent & Single Episode): <i>N</i> (%)	412 (6)	380 (44)	1502 (10)	766 (51)

M Mean, *SD* Standard Deviation, *MDD* Major Depressive Disorder.

p < 1×10^{-4}), with low heterogeneity between studies (*I*² [95%CI] = 0% [0, 64.8%]) (Supplementary Table 8).

Discussion

This study presents a large investigation of antidepressant exposure and the methylome³⁹. There was evidence of hypermethylation at eight individual CpGs and a singular region on Chromosome 2 (BP: 74196550-74196572) in those exposed to antidepressants using both self-report and prescription-derived measures³². Sex-stratified analyses indicated larger effect estimates in females compared to males. Functional annotation found that genes annotated to the top 100 differentially methylated CpGs in the self-report MWAS were significantly enriched in both a synaptic vesicle membrane gene-ontology and for genes expressed in the amygdala. Preliminary analysis indicated a significant correlation between the time in treatment and methylation at two probes (cg15071067; *DGUOK-AS1* and cg26277237; *KANK1*). Furthermore, there was significant enrichment in the top 1% of findings from an independent MWAS using MBD-Seq profiling. Finally, an MPS trained on GS data, demonstrated a robust association with antidepressant exposure in a

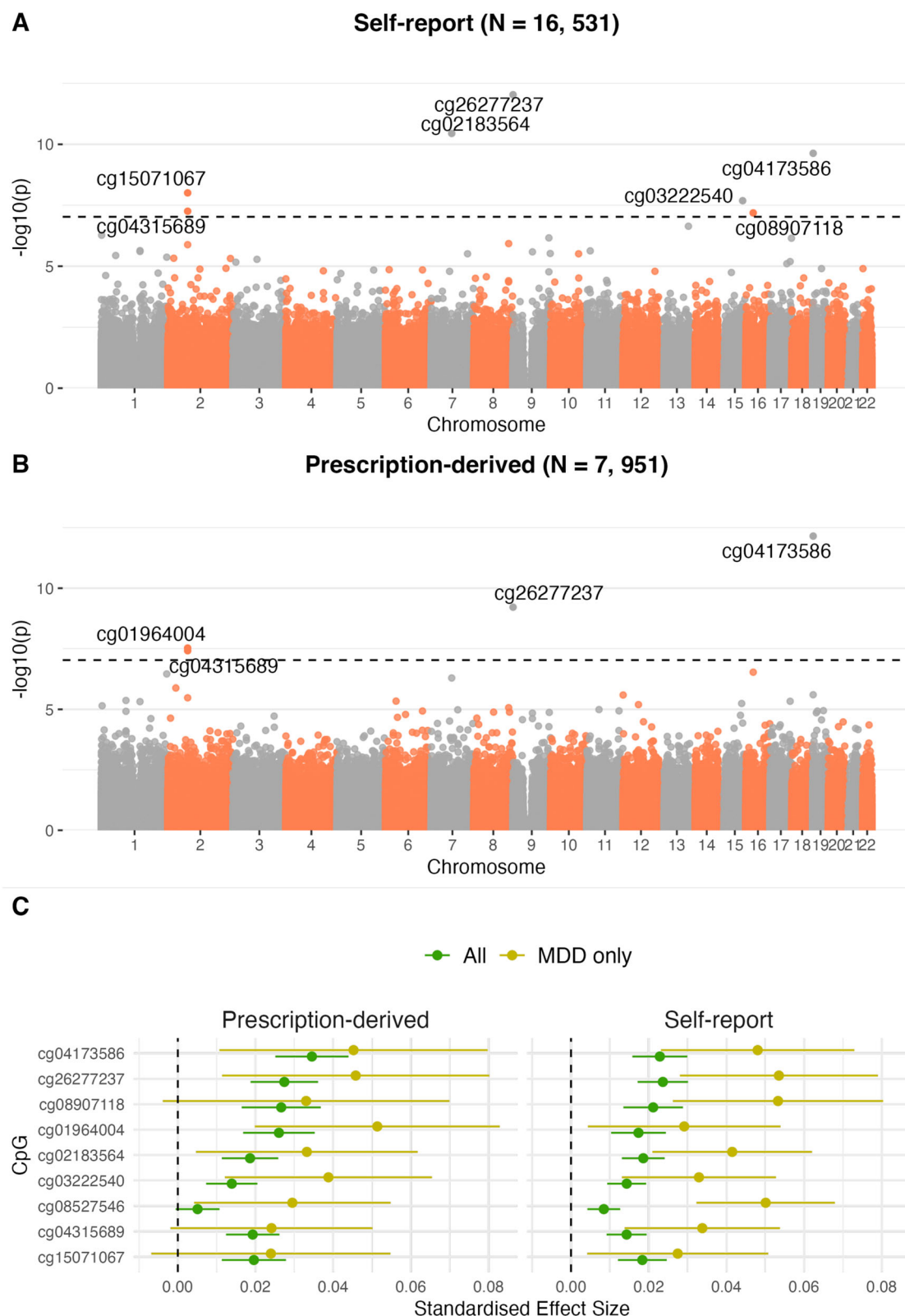


Fig. 2 | Methyome-wide association studies (MWAS) of self-reported and prescription-derived antidepressant exposure. Manhattan plots of the MWAS of self-report (**A**) and prescription-derived (**B**) antidepressant exposure using a Mixed-linear-model Omics-based Analysis (MOA) model. Significance was assessed using the p -value threshold 9.42×10^{-8} , as recommended for case-control MWAS analyses⁶⁷. **C** The MOA MWAS standardised effect sizes and 95% confidence-

intervals (effect estimate $\pm 1.96 \times \text{Standard Error}$) for associated CpGs ($p < 9.42 \times 10^{-8}$) for the full sample ($N_{\text{self-report}} = 16,531$, $N_{\text{prescription-derived}} = 7951$) and MDD-subgroup sample ($N_{\text{self-report}} = 2268$, $N_{\text{prescription-derived}} = 792$). Data are pre-sented as the MWAS effect estimates $\pm$ the 95% confidence intervals. Effect sizes represent a per-1 standard-deviation increase in CpG methylation M-values.

Table 2 | Eight CpGs associated with self-reported (SR) and/or prescription-derived (PD) antidepressant use

Probe	Significant in SR / PD MWAS	Chr	BP	Gene	β_{SR}	SE _{SR}	P _{SR}	β_{PD}	SE _{PD}	P _{PD}	Other traits associated with the CpG (EWAS catalog) ⁷⁹
cg26277237	SR + PD	9	631910	KANK1	0.024	0.0033	9.3×10^{-13}	0.027	0.0044	6.1×10^{-10}	C-reactive protein levels ⁴⁰ (basic model), Chronic kidney disease (basic model) ⁸¹ , Type 2 diabetes (basic model) ⁸² , Estimated glomerular filtration rate ⁸²
cg04173586	SR + PD	19	2167496	DOT1L	0.023	0.0036	2.35×10^{-10}	0.034	0.0048	7×10^{-13}	Alcohol consumption ⁸³ , COVID-19 severity (basic model) ⁸⁴ , Sex ⁸⁴ , Age ⁸⁵ , Alzheimer's disease Braak stage ⁸⁶ , Estimated glomerular filtration rate ⁸² , Schizophrenia ⁸⁷
cg08907118	SR only	16	27482516	GTF3C1	0.021	0.0039	6.46×10^{-8}	0.027	0.0052	3.0×10^{-7}	C-reactive protein levels (basic model) ⁸⁰ , Chronic pain (basic model) ⁸¹ , Type 2 diabetes (basic model) ⁸¹ , Type 2 diabetes (basic model) ⁸¹ , Sex ⁸⁴ , Age ⁸⁵
cg02183564	SR only	7	76874892	CCDC146	0.019	0.0028	3.59×10^{-11}	0.019	0.0037	5.17×10^{-7}	Birthweight ⁸⁸ , C-reactive protein levels (basic model) ⁸⁰ , Gestational age ⁸⁸ , Chronic pain (basic model) ⁸¹ , Type 2 diabetes (basic model) ⁸¹ , Age ⁸⁵
cg15071067	SR-only	2	74196550	DGUOK-AS1	0.018	0.0032	9.84×10^{-9}	0.020	0.0042	3.3×10^{-6}	C-reactive protein levels (basic model) ⁸⁰ , Incident chronic pain (basic model) ⁸¹ , Incident type 2 diabetes (basic & fully adjusted model) ⁸¹ , Chronic pain (self-report) ⁸¹ , Type 2 diabetes (basic model) ⁸¹
cg01964004	PD only	2	74196572	DGUOK-AS1	0.017	0.0036	1.3×10^{-6}	0.026	0.0047	3.0×10^{-8}	C-reactive protein levels (basic model) ⁸⁰ , Incident type 2 diabetes (basic & fully adjusted model) ⁸¹ , Chronic pain (basic model) ⁸¹ , Ischemic heart disease (basic model) ⁸¹ , Type 2 diabetes (basic model) ⁸¹
cg03222540	SR only	15	90575080	ZNF710	0.014	0.002	2.05×10^{-8}	0.014	0.0034	3.7×10^{-7}	C-reactive protein levels (basic model) ⁸⁰ , Chronic kidney disease (basic model) ⁸¹ , Type 2 diabetes (basic & fully adjusted model) ⁸¹
cg04315689	SR + PD	2	74198896	DGUOK-AS1	0.014	0.0026	5.56×10^{-8}	0.019	0.0035	3.85×10^{-8}	C-reactive protein levels (basic model) ⁸⁰ , Incident type 2 diabetes (basic model) ⁸¹ , Chronic kidney disease (basic & fully adjusted model) ⁸¹ , Chronic pain (basic model) ⁸¹ , Type 2 diabetes (basic & fully adjusted model) ⁸¹

The effect estimates (β), standard-errors (SE) and P-values (P) for the self-report (_{SR}) and prescription-derived (_{PD}) MWAS, calculated using a Mixed-linear-model Omics-based Analysis (MOA) model. The EWAS catalog was searched using the 'ewascatlog' R package⁷⁹ for other studies ($n > 1000$) which report a significant CpG-trait association, accessed on 17/03/2024.

meta-analysis of eight external cohorts and a prospective wave of data from GS, indicating the generalisability of our findings.

The CpG with the highest significance and the largest effect size, cg26277237, mapped to KN motif and ankyrin repeat domains 1 (KANK1), and was previously reported by Barbu et al.²⁷ on a smaller sample of GS. KANK1 facilitates the formation of the actin cytoskeleton and has an active role in neurite outgrowth and neurodevelopment⁴⁰. A meta-analysis of copy-number variant association studies found a significant duplication in KANK1 in those with five different neurodevelopmental disorders, including MDD⁴¹. The DMR analysis indicated antidepressant exposure is also significantly associated with hypermethylation near *DGUOK-AS1*, a long non-coding RNA (lncRNA). *DGUOK-AS1* has an inhibitory role on the expression of a nearby gene *DGUOK*⁴², which encodes a mitochondrial enzyme involved in the production of mitochondrial DNA⁴², and has previously been implicated as a risk gene in schizophrenia⁴³ and Alzheimer disease⁴⁴. A recent review reported evidence that antidepressants do influence mitochondrial function, although the effects are heterogeneous between different types of antidepressants, independent of their current classification^{45,46}. Functional validation of the associations of antidepressant exposure with DNAm at KANK1 and *DGUOK-AS1* would strengthen our findings. The integration of additional multi-omic data alongside in-vitro experiments could further assess the impact of these associations on biologically relevant processes, such as neuroplasticity.

Our results show broadly consistent findings between self-report and prescription-derived measures. However, self-report measures showed stronger signal in downstream functional annotation analyses. Due to the strong correlation between effect estimates and several overlapping significant signals, this is likely due to the increased sample size and power in the self-report cohort (+8580 individuals). The top CpGs in the self-report MWAS were significantly enriched for genes expressed in the amygdala, an important component of emotional brain circuits, specifically in regulating fear and stress responses⁴⁷. Genomic studies of MDD have shown enrichment in neural synaptic pathways⁴⁸ and brain regions in the meso-limbic system, including the prefrontal cortex, nucleus accumbens, hippocampus and the amygdala⁴⁹. Several meta-analyses have found evidence of amygdala hyperactivity in those with MDD^{50–52}, while other studies have demonstrated the amygdala response to stress can be dampened through neuroplastic processes following antidepressant treatment⁵³ or cognitive behavioural therapy⁵⁴. Furthermore, a recent systematic review reported sex-specific differences in amygdala activity and grey matter volume in MDD¹⁶. Notably, our sex-stratified analyses found nominally significant sex differences in the DNAm-antidepressant exposure associations at four significant CpGs, with larger effects observed in females. Although there is no clear consensus regarding sex differences in antidepressant efficacy, reports have found that women generally respond better to selective serotonin reuptake inhibitors (SSRIs) than men¹⁹. The nominally significant sex-differences in DNAm associations with antidepressant exposure and the significant enrichment of genes in the amygdala observed in this study highlight the potential role of the amygdala in mediating sex-specific responses to antidepressants. Future studies could investigate sex-specific DNAm-profiles of antidepressant exposure and their functional impact on the amygdala using functional imaging data.

There are several strengths of this study. The comparison of self-report and prescription-derived measures is valuable to the research community. Self-reported measures are often cheaper and easier to obtain in large-scale cohort studies³¹. Equally, the methods used in this study to derive of medication exposure using prescription records could enable passive data collection, enabling more generalisable analyses on whole populations outside of biases which influence participation in biobank cohorts⁵⁵. Furthermore, the MDD-subgroup analysis indicates that the hypermethylation associated with antidepressant use is not

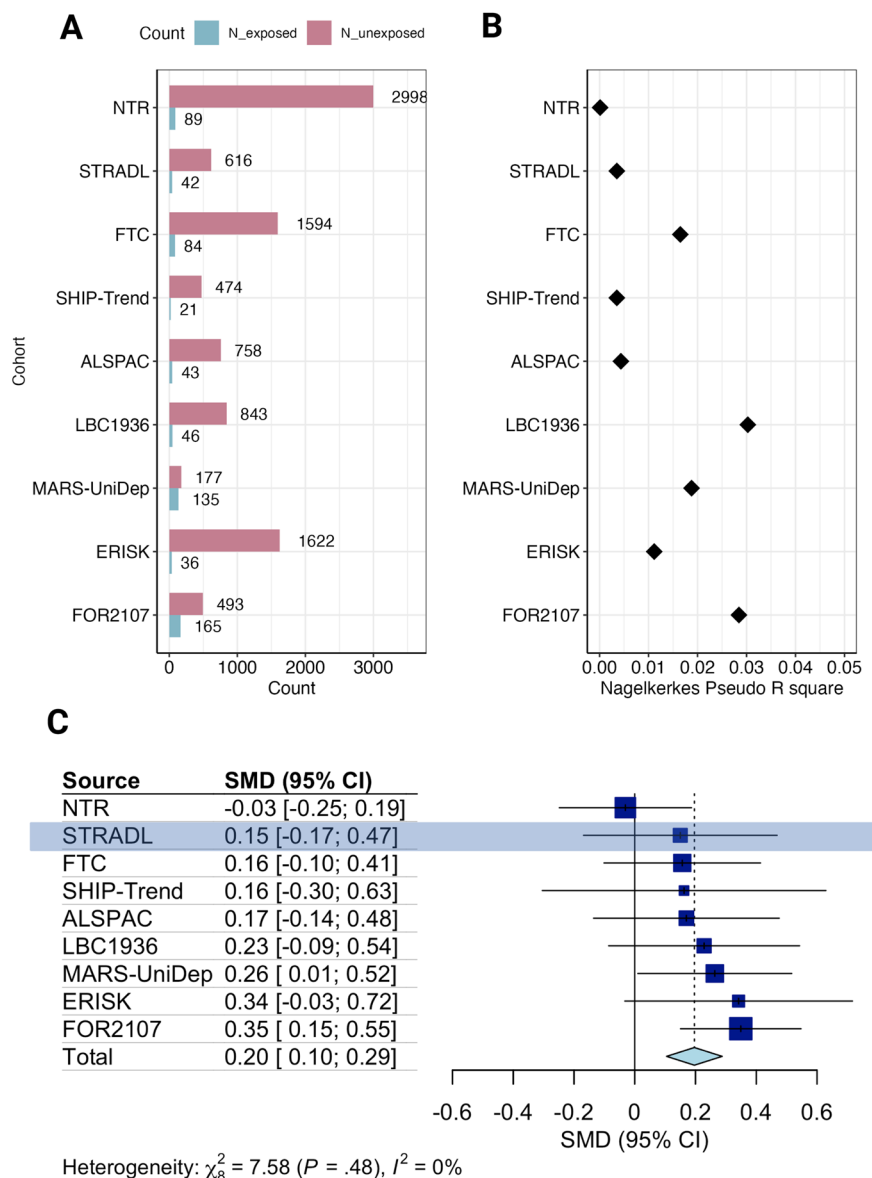


Fig. 3 | An antidepressant exposure methylation profile score (MPS) and antidepressant exposure in external cohorts. A The number of participants exposed and unexposed to antidepressants in each cohort ($N_{NTR} = 3087$, $N_{STRADL} = 658$, $N_{FTC} = 1678$, $N_{SHIP-TREND} = 495$, $N_{ALSPAC} = 801$, $N_{LBC1936} = 889$, $N_{MARS-UniDep} = 312$, $N_{ERISK} = 1658$, $N_{FOR2107} = 658$). **B** Nagelkerke's pseudo R^2 , representing an estimate of how much variance in the antidepressant exposure outcome that is explained by the MPS in each cohort. **C** The effect estimates and 95% confidence intervals (effect estimate $\pm 1.96 \times$ Standard Error) of MPS - antidepressant exposure association in

each cohort, using either a generalised linear model (FOR2107 and ALSPAC), a generalised linear mixed model (SHIP-Trend, LBC1936, MARS-UniDep, STRADL and E-Risk) or a generalised estimation equation model (FTC and NTR). The square size = study weight in the random-effects meta-analysis. The pooled effect estimate interval, calculated from a random-effects meta-analysis, is represented by the blue diamond ($N_{pooled} = 10,236$). Created in BioRender. Davyson, E. (2024) <https://BioRender.com/q70l792>.

primarily driven by MDD indication. Additionally, the performance of the GS-trained MPS in discriminating antidepressant exposure across eight external datasets, alongside the significant enrichment of top findings with an independent MWAS, demonstrates that this may be a generalisable biomarker indicative of antidepressant exposure.

This study has various limitations. Firstly, both measures of antidepressant exposure do not discriminate between antidepressant drugs, classes, or dosages. However, we anticipate the opportunity to investigate more medication-specific effects on the methylome using prescription-linkage data as biobanks increase in size. Secondly, all the cohorts used primarily consist of European ancestry. It is paramount that this analysis is conducted in non-European ancestral groups to further verify our findings and disentangle any ancestry-specific effects^{56–58}. Thirdly, our epidemiological

analyses do not adjust for various other potential confounders, such as comorbidities and concomitant treatments. However, adjusting for all possible confounds such as these may bias findings due to collinearity and collider bias. Fourthly, by design, this epidemiological study cannot directly address causality between antidepressant exposure and DNAm. The integration of DNAm analysis into randomised controlled trials of antidepressants is important to establish the exact nature of the association and to inform potential new targets for antidepressant therapy. Finally, this study does not examine recently approved rapid-acting antidepressants, such as ketamine. Future research could build on this by exploring the associations between DNAm and rapid-acting antidepressants, while comparing these effects with those of slower-acting antidepressants discussed in this study.

This study indicates that antidepressant exposure is associated with hypermethylation at *DGUK-AS1* and *KANK1*, which have roles in mitochondrial metabolism and neurite outgrowth respectively. Future research should include more cohorts of non-European ancestry, alongside the incorporation of DNAm in randomised trials of antidepressants to establish causality. If replicated, targeting of these genes could inform the design of more effective and better tolerated treatments for depression.

Methods

Ethics declaration

All components of Generation Scotland: The Scottish Family Health Study (GS) received ethical approval from the NHS Tayside Committee on Medical Research Ethics (REC Reference Number: 05/S1401/89). Generation Scotland has also been granted Research Tissue Bank status by the East of Scotland Research Ethics Service (REC Reference Number: 20-ES-0021), providing generic ethical approval for a wide range of uses within medical research. All participants included in the current study provided informed consent for the use of their data for biomedical research. The research presented in this study complies with these ethical regulations. The study protocol for the Netherlands Study of Anxiety and Depression (NESDA) was approved centrally by the Ethical Review Board of the VU University Medical Centre and subsequently by local review boards of each participating centre (METC number 2003-183). After full verbal and written information about the study, written informed consent was obtained from all participants at the start of baseline assessment. Ethics approvals have been granted for multiple studies concerning the Finnish Twin Cohort (FTC) twins by the ethics committees of Helsinki University Central Hospital (113/E3/2001, 249/E5/2001, 346/E0/05, 270/13/03/01/2008, and 154/13/03/00/2011) with the last one on the transfer of biological samples to the THL Biobank in 2018 (HU51179912017). Participants in the FTC were given information on the study procedures and of freedom to participate or to decline at any point in both oral and written form. Informed consent was obtained upon the contact with the study subjects before new questionnaire information was collected, and when clinical investigations were undertaken with sampling of biological material. The Study of Health in Pomerania (SHIP) received ethical approval from the ethics Committee of the University Medicine Greifswald, Germany (BB 39/08). All participants provided written informed consent before any assessment and/or sampling took place. FOR2107 received ethical approval from the ethics committees of the Medical Faculties, University of Marburg (AZ: 07/14) and University of Münster (2014-422-b-S). Participants provided written consent before any assessment and/or sampling took place. The Netherlands Twin Register was approved by the Central Ethics Committee on Research Involving Human Subjects of the VU university Medical Centre, Amsterdam, an Institutional Review Board certified by the U.S Office of Human Research Protections (IRB number IRB00002991 under Federal-wide Assurance-FWA00017598; IRB/institute codes, NTR 03-180). Informed consent was obtained from all participants. The Munich Antidepressant Response (MARS) study was approved by the local Ethics Committee of the Ludwig Maximilians University, Munich, Germany (318/00; 244/01) and carried out in accordance with the latest version of the Declaration of Helsinki. All participants provided written consent after the study protocol and potential risks were explained. The UniPolar Depression (UniDep) study was approved by the Bavarian State Medical Association (BLAEK) (01217) and written informed consent was obtained from all subjects. The Lothian Birth Cohort 1936 (LBC1936) received ethical approval from the Multicentre Research Ethics Committee for Scotland (baseline, MREC/01/0/56), the Lothian Research Ethics Committee (age 70, LREC/2003/2/29), and the Scotland A Research Ethics Committee (ages 73, 76, 79, 07/MRE00/58). All participants provided written

informed consent. Ethical approval for the Avon Longitudinal Study of Parents and Children (ALSPAC) was obtained from the ALSPAC Ethics and Law Committee and the Local Research Ethics Committees under proposal B3818. Consent for biological samples has been collected in accordance with the Human Tissue Act (2004). Each study phase of the Environmental risk (E-risk) Longitudinal Twin Study received ethical approval from the Joint South London and Maudsley and Institute of Psychiatry Research Ethics Committee (NRES 1997/122). Parents gave informed consent, and twins gave assent between 5 and 12 years and then informed consent at age 18. All components of the Stratifying Resilience and Depression Longitudinally (STRADL) received formal national ethical approval from the NHS Tayside committee on research ethics (reference 14/SS/0039).

Generation Scotland

Generation Scotland: The Scottish Family Health Study (GS) is a family-based cohort study ($N = 24,000$), investigating the genetic and environmental factors influencing health within Scotland (Supplementary Methods)^{59,60}. Data collection including biological sampling occurred between February 2006 and March 2011.

Methylation data

DNAm was profiled from baseline blood samples using the Illumina MethylationEPIC array for 19,390 individuals. Methylation typing was performed in 4 distinct sets between 2017 and 2021. Sets 1, 2 and 4 included related individuals within and between each set while all individuals in set 3 were unrelated to each other and to individuals in set 1 (genetic relationship matrix (GRM) threshold <0.05). Following quality control (QC) (Supplementary Methods), the sample sizes were Set 1: 5087, Set 2: 459, Set 3: 4450 and Set 4: 8873 individuals. The sets were combined and dasen normalisation was performed across all individuals⁶¹. In total, 752,741 probes and 18,869 individuals passed QC. Beta-values were transformed to M-values using the 'beta2M()' function in the 'lumi' R package⁶², and standardised using the 'std-probe' flag in the OSCA software⁶³.

Antidepressant exposure phenotypes

Prescription-derived antidepressant exposure was measured using antidepressant prescriptions from Public Health Scotland (Supplementary Methods), according to the British National Formulary (BNF) paragraph code '040303', which largely corresponds to Anatomical Therapeutic Chemical (ATC) subclass 'N06A' (Supplementary Table 9) ($N_{\text{prescriptions}} = 174,454$, $N_{\text{people}} = 7544$) (Supplementary Table 10, Supplementary Figs. 36-43). After removing ambiguous prescriptions ($N_{\text{prescriptions}} = 5484$, $N_{\text{people}} = 171$) (Supplementary Methods, Supplementary Table 11), active treatment periods were defined by consecutive dispensing of antidepressant medications (Supplementary Methods, Supplementary Fig. 44-46). Those who had their blood-draw appointment ≥ 7 days after treatment initiation or < 7 days after the end of a treatment period were defined as exposed ($N_{\text{exposed}} = 861$) (Supplementary Fig. 47). Those with no antidepressant prescriptions on record were defined as unexposed ($N_{\text{unexposed}} = 7090$).

Self-reported antidepressant exposure was derived from questionnaires sent 1–2 weeks before venepuncture (Supplementary Methods, Supplementary Table 12, Supplementary Fig. 48). Those who did/did not self-report antidepressant use were defined as exposed and unexposed respectively ($N_{\text{exposed}} = 1508$, $N_{\text{unexposed}} = 15,023$). Out of 6473 individuals with both self-report and prescription-derived phenotypes, 6355 exhibited concordant classification of antidepressant exposure (Supplementary Fig. 49). The MDD-only phenotypes were derived by stratifying the samples to those with a lifetime MDD diagnosis, ascertained by the Structured Clinical Interview of the Diagnostic and Statistical Manual, version IV (SCID)⁶⁴ (prescription-derived: $N_{\text{exposed}} = 380$, $N_{\text{unexposed}} = 412$, self-report: $N_{\text{exposed}} = 766$, $N_{\text{unexposed}} = 1502$) (Supplementary Methods, Supplementary Fig. 50).

Methylome-wide association study

The MWAS were performed using a Mixed-linear-model Omics-based Analysis (MOA) in the OSCA software⁶³. To account for relatedness within GS, each phenotype was regressed on a GRM⁶⁵ using the Best Linear Unbiased Prediction (BLUP) tool in GCTA software⁶⁶ ('remi-pred-rand' flag). The residuals were entered into a MOA model, which included a methylation omics-relatedness matrix as a random effect and age, sex, AHRR probe (cg05575921) M-values to proxy for smoking status, and predicted monocyte and lymphocyte cell proportions as fixed effects. Statistical significance was assessed using the p -value threshold 9.42×10^{-8} , as recommended for case-control MWAS analyses⁶⁷. Effect sizes represent a per-1 standard-deviation increase in CpG methylation M-values. For any significant CpGs, we searched the EWAS catalog to assess their associations with other traits in the literature. For sex-stratified analyses, we divided our participants by self-reported sex (male/female) and used the same MOA model without sex as a covariate.

Differentially methylated regions (DMRs) were identified using the 'dmrff' R package⁶⁸, which performs an inverse-variance-weighted meta-analysis of MWAS beta and standard-error estimates per region, adjusting for estimate uncertainty and the correlational structure between probes. Candidate DMRs are identified as sets (>2) of CpGs ≤ 500 bp apart with nominal significance ($P < 0.05$) and consistent effect direction. DMRs achieving Bonferroni-corrected p -value < 0.05 were considered statistically significant.

Functional annotation

Gene-sets for both MWAS were collated by annotating the top 100 CpGs (by p -value) by the Infinium MethylationEPIC BeadChip database⁶⁹. Hypergeometric tests, using 'phyper()', were used to assess the overlap of CpGs and gene-sets from both analyses. The background set consisted of CpGs and genes on or annotated to the EPIC array (Supplementary Data 16). The 'GENE2FUNC' analysis in functional mapping and annotation web-tool (FUMA) was used to assess enrichment of both gene-sets across 54 specific tissues in the Genotype-Tissue Expression (GTEx)^{70,71} Project. Both gene-sets were tested for enrichment in synapse-related GO terms using the SynGo web tool⁷² and for enrichment in GO: Biological Processes gene-sets (20 <ngenes <500) in the msigdb database⁷³, using the 'gsameth()' function from 'missMethyl' R package⁷⁴ (Supplementary Data 17).

Methylation profile score

A least absolute shrinkage and selection operator (LASSO) penalised regression model was applied using 'cv.biglasso()' from the 'glmnet' R package⁷⁵ on the GS sample to generate weights for a MPS of antidepressant exposure. First, the self-report phenotype was regressed against the GRM (using BLUP) to account for relatedness, and then on age, sex, AHRR probe M-values as a proxy for smoking status, batch, and white blood cell (monocyte and lymphocyte) proportions. The extracted residuals were used as the dependent variable and the standardised CpG sites on both EPIC and 450 K Illumina arrays ($N = 365,912$) were included as independent variables. Ten-fold cross-validation was applied, and the mixing parameter was set to 1. The non-zero weighted CpGs identified in the LASSO model were used to calculate a weighted-sum MPS in external datasets (FTC, SHIP-Trend, FOR2107, NTR, LBC1936, ALSPAC, MARS-UniDep and E-Risk, $N_{\text{total}} = 9578$, $N_{\text{exposed}} = 619$) and a prospective sample of GS (STRADL, $N_{\text{total}} = 658$, $N_{\text{exposed}} = 42$) (Supplementary Methods). The association between antidepressant exposure and the MPS was assessed using: generalised linear mixed models, generalised linear models and generalised estimation equations, depending on the cohort's population structure (i.e., twin studies vs unrelated participants) and DNAm pre-processing (Supplementary Methods). In-depth information about each cohort's sample, DNAm pre-processing and associational model can be found in the supplementary information (Supplementary

Methods, Supplementary Data 18–20). All models have antidepressant exposure as the outcome and include age at blood sampling, sex, batch (where applicable), white blood cell proportions/counts and M-values at the AHRR probe as covariates. A random-effects meta-analysis using a DerSimonian-Laird estimator was performed to assess the overall association between the MPS and antidepressant exposure, using the 'meta' R package⁷⁶.

Statistics and reproducibility

No statistical method was used to predetermine the sample size in either the self-report or prescription-derived MWAS analysis. Sample size was determined by the availability of eligible individuals (with DNA methylation and phenotypic information available) in each cohort. For all the MWAS, samples that failed quality control checks (Supplementary Methods) were excluded from the analysis. For the MDD-subgroup analyses, individuals who fulfilled criteria for bipolar disorder in the SCID assessment ($N_{\text{prescription-derived}} = 18$, $N_{\text{self-report}} = 52$) were discarded. For external cohorts which utilised inpatient and outpatient samples (FOR2107, MARS and UniDep) there were further exclusionary criteria. In FOR2107, individuals with a bipolar, schizophrenia or schizoaffective disorder diagnosis were excluded. In MARS, individuals were excluded if they displayed hypomanic symptoms, a diagnosis of alcohol dependence, a history of illicit drug use or having depressive symptoms secondary to another medical or neurological condition. In UniDep, individuals were excluded if they had manic or hypomanic episodes, mood incongruent psychotic symptoms, a lifetime diagnosis of intravenous drug abuse and depressive symptoms which are secondary to a substance abuse disorder or due to medical illness or medication. The study design did not involve randomization as this is an observational study utilizing data from population-based cohorts. The investigators were not blinded to the exposure or outcome variables as this was an observational study with no experimental allocation of treatments.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The data generated in this study (MWAS summary statistics, downstream functional results and weights in the MPS) have been deposited in a Zenodo repository (14203229 [<https://doi.org/10.5281/zenodo.14203229>])⁷⁷. Additionally, the source data for main figures and all Supplementary Figs. which are not sharing individual-level data are provided in this repository. The raw data from all cohorts (Generation Scotland, Stratifying Anxiety and Depression Longitudinally, Netherlands Study of Anxiety and Depression, Finnish Twin Cohort, Study of Health in Pomerania, FOR2107, Netherlands Twin Register, Munich Antidepressant Response Study, UniPolar Depression Study, Avon Longitudinal Study of Parents and Children, E-risk Longitudinal Twin Study and Lothian Birth Cohorts) are not publicly available due to them containing information that could compromise participant consent and confidentiality. Access information for each cohort can be found below. Generation Scotland is run as a resource for the research community. Requests to use the Resource are made from: Academic collaborators: employees who are party to the Generation Scotland Collaboration Agreement, or researchers or employees of an academic institution or the NHS. Commercial organisations: specific arrangements have been defined to allow commercial organisations to access Generation Scotland resources. Data can be obtained from the data owners. Instructions for accessing Generation Scotland data can be found here: <https://genscot.ed.ac.uk/for-researchers/access>; the GS Access Request Form can be downloaded from this site. Completed request forms must be sent to access@generationscotland.org to be approved by the Generation Scotland Access Committee. Upon

submission, applications are reviewed within 6–8 weeks. For any further correspondence and material requests please contact genscot@ed.ac.uk. The Netherlands Study of Anxiety and Depression (NESDA) is run as a resource for the research community and is open to data-use requests from bona fide international researchers. Instructions for gaining access to NESDA can be found here: https://www.nesda.nl/nesda/wp-content/uploads/2024/09/NESDA_policy_data_access.pdf. Applications involve submitting a research proposal, including specific research questions, methodology and proposed statistical analysis, to the NESDA management committee (nesda@amsterdamumc.nl). Data access forms can be downloaded here: www.nesda.nl. The review process by the management committee for data access may take up to 6 weeks. E-Risk is run as a resource for the research community and is free to access by researchers from all over the world who are based at universities or research institutions. E-Risk operates a managed access process to protect the privacy of the participants. Instructions for accessing The Environmental-Risk Longitudinal Twin Study (E-Risk) data can be found here: <https://eriskstudy.com/data-access>. In brief following the reading of a data-sharing protocol, a concept paper will need to be submitted outlining the proposed analyses and required variables via this link: <https://redcap.link/ERiskConceptPaperForm> for consideration by the E-Risk Steering Committee. A decision will usually be made within 1 month. Once approved, the concept paper will be made public and the data sent securely to those requiring access within 1–2 months after a formal data use agreement has been signed. For any questions regarding the data available and the access process, please email eriskstudy@kcl.ac.uk. The data of the Study of Health in Pomerania study cannot be made publicly available due to the informed consent of the study participants, but it can be accessed through a data application form available at <https://transfer.ship-med.uni-greifswald.de/> for researchers who meet the criteria for access to confidential data. In detail, access to the data requires a proposal submitted to the University Medicine Greifswald represented by the steering committee of the Research Network Community Medicine (FVCM), which meets once a month. A successful data application provides data usage permission for three years at maximum and needs to be extended afterwards if needed. In accordance with the Finnish Biobank Act, the data used in the analysis from the Finnish Twin Cohort (FTC) is deposited in the Biobank of the Finnish Institute for Health and Welfare (<https://thl.fi/en/research-and-development/thl-biobank/for-researchers>). It is available to researchers from academia and companies after written application and following the relevant Finnish legislation. To ensure the protection of privacy and compliance with national data protection legislation, a data use/transfer agreement is needed, the content and specific clauses of which will depend on the nature of the requested data. Allow 2–4 weeks for initial response from the Biobank. Avon Longitudinal Study of Parents and Children (ALSPAC) is run as a resource for the research community. Instructions for accessing ALSPAC data can be found here: <https://www.bristol.ac.uk/alspac/researchers/access/>. A research proposal must be submitted via the research proposal system for consideration by the ALSPAC Executive Committee. For any questions regarding accessing data or samples please email alspac-data@bristol.ac.uk (data) or bbl-info@bristol.ac.uk (samples). Approval may take up to two weeks. Lothian Birth Cohort 1936 (LBC1936) is run as a resource for the research community which actively collaborates with research experts in the UK and internationally. Instructions for how to access LBC1936 data can be found here: <https://lothian-birth-cohorts.ed.ac.uk/data-access-collaboration>. In brief, identify the variables you will require for the analysis from the LBC1936 data dictionary (which can be downloaded from the link). Following this, prepare a data request form including the provisional title of the study, the principle researcher and institution and the brief rationale for the study, research method and main variables. Email the LBC request form to Professor Simon Cox (simon.cox@ed.ac.uk), the director of Lothian

Birth Cohorts for approval from the study investigator team. Approval may take up to 4 weeks. Data will be shared on the basis of a Data or Material Transfer Agreement between provider and recipient institutions. The Netherlands Twin Register (NTR) is run as a resource for the research community and is open to data-use requests from bona fide international researchers. Information on accessing NTR data can be found here: <https://ntr-data-request.psy.vu.nl/>. To submit a data sharing request, complete a data sharing request form (<https://ntr-data-request.psy.vu.nl/DSR-forms.html>) and send it to NTR Data management team (ntr.datamanagement.fgb@vu.nl), which will check the request for feasibility and completeness and pass on the request to the data access committee (DAC) for approval. The review process by the DAC may take up to 4 weeks. The raw data collected in the FOR2107 study is not openly accessible to protect participant consent and confidentiality. Nevertheless, the FOR2107 study serves as a valuable resource for the global research community. It is accessible, in principle, to all scientific researchers affiliated with non-commercial research organizations worldwide. Researchers seeking access to the study's data must submit a formal research proposal. This proposal should outline the specific research questions, methodology, and planned statistical analyses. Applications are reviewed by the principal investigators of FOR2107, Professors Tilo Kircher (tilo.kircher@staff.uni-marburg.de) and Udo Dannlowski (udo.dannlowski@uni-muenster.de), within 6–8 weeks of submission. The raw data collected in the Munich Antidepressant Response Study-Unipolar Depression Study (MARS-UniDep) is not publicly available but can be shared with researchers upon request. Access to the data can be requested in a data transfer agreement, please contact Darina Czamara (darina@psych.mpg.de). Only research questions related to psychiatric disorders can be addressed directly. The expected time frame for response to access requests is 2 weeks. Data access will be granted until the end of the requested project, that is upon publication of the related manuscript.

Code availability

All code and analysis scripts are available on GitHub (https://github.com/Elladavyson/Antidepressant_MWAS) and as a linked Zenodo object (14185886 [<https://doi.org/10.5281/zenodo.14185886>])⁷⁸. Furthermore, a capsule of the code processing the MWAS summary statistics, downstream analyses and the meta-analysis results is provided at Code Ocean (<https://doi.org/10.24433/CO.6842127.v2>).

References

1. GBD 2017 Dis Injury Incidence Pr, James, S. L. G. et al. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. *The Lancet*. 2018;392:1789–1858. [https://doi.org/10.1016/S0140-6736\(18\)32279-7](https://doi.org/10.1016/S0140-6736(18)32279-7).
2. Braund, T. A. et al. Antidepressant side effects and their impact on treatment outcome in people with major depressive disorder: an iSPOT-D report. *Transl. Psychiatry* **11**, 1–8 (2021).
3. Cipriani, A. et al. Comparative efficacy and acceptability of 21 antidepressant drugs for the acute treatment of adults with major depressive disorder: a systematic review and network meta-analysis. *Lancet* **391**, 1357–1366 (2018).
4. Souery, D. et al. Switching antidepressant class does not improve response or remission in treatment-resistant depression. *J. Clin. Psychopharmacol.* **31**, 512–516 (2011).
5. McIntyre, R. S. et al. Treatment-resistant depression: definition, prevalence, detection, management, and investigational interventions. *World Psychiatry* **22**, 394–412 (2023).
6. McIntosh, A. M., Sullivan, P. F. & Lewis, C. M. Uncovering the Genetic Architecture of Major Depression. *Neuron* **102**, 91–103 (2019).

7. Robinson, E. Psychopharmacology: From serendipitous discoveries to rationale design, but what next. *Brain Neurosci. Adv.* **2**, 2398212818812629 (2018).
8. Harmer, C. J., Duman, R. S. & Cowen, P. J. How do antidepressants work? New perspectives for refining future treatment approaches. *Lancet Psychiatry* **4**, 409–418 (2017).
9. Hirschfeld R. M. A. History and Evolution of the Monoamine Hypothesis of Depression. *J. Clin. Psychiatry* **61**, 4–6 (2000).
10. Berton, O. & Nestler, E. J. New approaches to antidepressant drug discovery: beyond monoamines. *Nat. Rev. Neurosci.* **7**, 137–151 (2006).
11. Ślifirski, G., Król, M. & Turło, J. 5-HT Receptors and the Development of New Antidepressants. *Int. J. Mol. Sci.* **22**, 9015 (2021).
12. Jauhar, S. et al. A leaky umbrella has little value: evidence clearly indicates the serotonin system is implicated in depression. *Mol. Psychiatry Published online June* **16**, 1–4 (2023).
13. Bessa, J. M. et al. The mood-improving actions of antidepressants do not depend on neurogenesis but are associated with neuronal remodeling. *Mol. Psychiatry* **14**, 764–773 (2009).
14. Duman, R. S. & Aghajanian, G. K. Synaptic Dysfunction in Depression: Potential Therapeutic Targets. *Science* **338**, 68–72 (2012).
15. Casarotto, P. C. et al. Antidepressant drugs act by directly binding to TRKB neurotrophin receptors. *Cell* **184**, 1299–1313.e19 (2021).
16. Mohammadi, S. et al. Brain-based Sex Differences in Depression: A Systematic Review of Neuroimaging Studies. *Brain Imaging Behav. Published online April* **14**, 1–29 (2023).
17. Rubinow, D. R. & Schmidt, P. J. Sex differences and the neurobiology of affective disorders. *Neuropsychopharmacol* **44**, 111–128 (2019).
18. Bigos, K. L., Pollock, B. G., Stankevich, B. A. & Bies, R. R. Sex differences in the pharmacokinetics and pharmacodynamics of antidepressants: An updated review. *Gen. Med.* **6**, 522–543 (2009).
19. LeGates, T. A., Kvarta, M. D. & Thompson, S. M. Sex differences in antidepressant efficacy. *Neuropsychopharmacol* **44**, 140–154 (2019).
20. Nabais, M. F. et al. An overview of DNA methylation-derived trait score methods and applications. *Genome Biol.* **24**, 28 (2023).
21. Bird, A. Perceptions of epigenetics. *Nature* **447**, 396–398 (2007).
22. Webb, L. M., Phillips, K. E., Ho, M. C., Veldic, M. & Blacker, C. J. The Relationship between DNA Methylation and Antidepressant Medications: A Systematic Review. *Int. J. Mol. Sci.* **21**, 826 (2020).
23. Menke, A. & Binder, E. B. Epigenetic alterations in depression and antidepressant treatment. *Dialogues Clin. Neurosci.* **16**, 395–404 (2014).
24. Gassen, N. C. et al. Chaperoning epigenetics: FKBP51 decreases the activity of DNMT1 and mediates epigenetic effects of the antidepressant paroxetine. *Sci. Signal.* **8**, ra119 (2015).
25. Le François, B. et al. Chronic mild stress and antidepressant treatment alter 5-HT1A receptor expression by modifying DNA methylation of a conserved Sp4 site. *Neurobiol. Dis.* **82**, 332–341 (2015).
26. Sales, A. J. & Joca, S. R. L. Effects of DNA methylation inhibitors and conventional antidepressants on mice behaviour and brain DNA methylation levels. *Acta Neuropsychiatrica* **28**, 11–22 (2016).
27. Barbu, M. C. et al. Methylome-wide association study of antidepressant use in Generation Scotland and the Netherlands Twin Register implicates the innate immune system. *Mol. Psychiatry* **27**, 1647–1657 (2022).
28. Klungel, O. H. et al. Influence of question structure on the recall of self-reported drug use. *J. Clin. Epidemiol.* **53**, 273–277 (2000).
29. Nielsen, M. W., Søndergaard, B., Kjølner, M. & Hansen, E. H. Agreement between self-reported data on medicine use and prescription records vary according to method of analysis and therapeutic group. *J. Clin. Epidemiol.* **61**, 919–924 (2008).
30. Stirratt, M. J. et al. Self-report measures of medication adherence behavior: recommendations on optimal use. *Transl. Behav. Med.* **5**, 470–482 (2015).
31. McClintock, H. F., BeKampis, A. N., Hartmann, E. & Bogner, H. R. Adherence to Antidepressants in Underserved Communities: A Comparison of Electronic Monitoring and Self-report Measures. *Community Ment. Health J.* **56**, 727–734 (2020).
32. Hafferty, J. D. et al. Self-reported medication use validated through record linkage to national prescribing data. *J. Clin. Epidemiol.* **94**, 132–142 (2018).
33. Penninx BWJH, Beekman A. T. F. et al. The Netherlands Study of Depression and Anxiety (NESDA): rationale, objectives and methods. *Int. J. Methods Psychiatr. Res.* **17**, 121–140 (2008).
34. Aberg, K. A. et al. A MBD-seq protocol for large-scale methylome-wide studies with (very) low amounts of DNA. *Epigenetics* **12**, 743–750 (2017).
35. Aberg, K. A., Chan, R. F. & van den Oord, E. J. C. G. MBD-seq - realities of a misunderstood method for high-quality methylome-wide association studies. *Epigenetics* **15**, 431–438 (2019).
36. Chan, R. F. et al. Enrichment methods provide a feasible approach to comprehensive and adequately powered investigations of the brain methylome. *Nucleic Acids Res.* **45**, e97 (2017).
37. Shabalin, A. A. et al. RaMWAS: fast methylome-wide association study pipeline for enrichment platforms. *Bioinformatics* **34**, 2283–2285 (2018).
38. Cabrera, C. P. et al. Uncovering Networks from Genome-Wide Association Studies via Circular Genomic Permutation. *G3 Genes|Genomes|Genet.* **2**, 1067–1075 (2012).
39. Monster, T. B. M., Janssen, W. M. T., de Jong, P. E. & de Jong-van den Berg, L. T. W. Group for the PS. Pharmacy data in epidemiological studies: an easy to obtain and reliable tool. *Pharmacoepidemiology Drug Saf.* **11**, 379–384 (2002).
40. Chandra Roy, B., Kakinuma, N. & Kiyama, R. Kank attenuates actin remodeling by preventing interaction between IRSp53 and Rac1. *J. Cell Biol.* **184**, 253–267 (2009).
41. Glessner, J. T. et al. Copy number variation meta-analysis reveals a novel duplication at 9p24 associated with multiple neurodevelopmental disorders. *Genome Med.* **9**, 106 (2017).
42. Fazel-Najafabadi, M. et al. Discovery and Functional Characterization of Two Regulatory Variants Underlying Lupus Susceptibility at 2p13.1. *Genes (Basel)* **13**, 1016 (2022).
43. Lm, H. et al. Gene expression imputation across multiple brain regions provides insights into schizophrenia risk. *Nat. Genet.* **51**, 659–674 (2019).
44. Nazarian, A., Yashin, A. I. & Kulminski, A. M. Methylation-wide association analysis reveals AIM2, DGUOK, GNAI3, and ST14 genes as potential contributors to the Alzheimer's disease pathogenesis. *Published online May* **15**, 322503 (2018).
45. Emmerzaal, T. L. et al. Effect of neuropsychiatric medications on mitochondrial function: For better or for worse. *Neurosci. Biobehav. Rev.* **127**, 555–571 (2021).
46. Holper, L., Ben-Shachar, D. & Mann, J. J. Psychotropic and neurological medication effects on mitochondrial complex I and IV in rodent models. *Eur. Neuropsychopharmacol.* **29**, 986–1002 (2019).
47. Ressler, K. J. Amygdala Activity, Fear, and Anxiety: Modulation by Stress. *Biol. Psychiatry* **67**, 1117–1119 (2010).
48. Howard, D. M. et al. Genome-wide association study of depression phenotypes in UK Biobank identifies variants in excitatory synaptic pathways. *Nat. Commun.* **9**, 1470 (2018).
49. Levey, D. F. et al. Bi-ancestral depression GWAS in the Million Veteran Program and meta-analysis in >1.2 million individuals highlight new therapeutic directions. *Nat. Neurosci.* **24**, 954–963 (2021).
50. McTeague, L. M. et al. Identification of Common Neural Circuit Disruptions in Emotional Processing Across Psychiatric Disorders. *Am. J. Psychiat.* <https://doi.org/10.1176/appi.ajp.2019.18111271> (2020).

51. Janiri, D. et al. Shared Neural Phenotypes for Mood and Anxiety Disorders: A Meta-analysis of 226 Task-Related Functional Imaging Studies. *JAMA Psychiatry* **77**, 172 (2020).
52. Li, X. & Wang, J. Abnormal neural activities in adults and youths with major depressive disorder during emotional processing: a meta-analysis. *Brain Imaging Behav.* **15**, 1134–1154 (2021).
53. Goldstein-Piekarski, A. N. et al. Human amygdala engagement moderated by early life stress exposure is a biobehavioral target for predicting recovery on antidepressants. *Proc. Natl Acad. Sci.* **113**, 11955–11960 (2016).
54. Månsson, K. N. T. et al. Neuroplasticity in response to cognitive behavior therapy for social anxiety disorder. *Transl. Psychiatry* **6**, e727–e727 (2016).
55. Schoeler, T. et al. Participation bias in the UK Biobank distorts genetic associations and downstream analyses. *Nat. Hum. Behav.* **7**, 1216–1227 (2023).
56. Genetics for all. *Nat Genet.* **51**, 579–579 <https://doi.org/10.1038/s41588-019-0394-y> (2019).
57. Elliott, H. R. et al. Characterisation of ethnic differences in DNA methylation between UK-resident South Asians and Europeans. *Clin. Epigenetics* **14**, 130 (2022).
58. Li, B. et al. Incorporating local ancestry improves identification of ancestry-associated methylation signatures and meQTLs in African Americans. *Commun. Biol.* **5**, 1–12 (2022).
59. Smith, B. H. et al. Cohort Profile: Generation Scotland: Scottish Family Health Study (GS:SFHS). The study, its participants and their potential for genetic research on health and illness. *Int. J. Epidemiol.* **42**, 689–700 (2013).
60. Smith, B. H. et al. Generation Scotland: the Scottish Family Health Study; a new resource for researching genes and heritability. *BMC Med Genet.* **7**, 74 (2006).
61. Pidsley, R. et al. LC. A data-driven approach to preprocessing Illumina 450 K methylation array data. *BMC Genomics* **14**, 293 (2013).
62. Du, P., Kibbe, W. A. & Lin, S. M. lumi: a pipeline for processing Illumina microarray. *Bioinformatics* **24**, 1547–1548 (2008).
63. Zhang, F. et al. OSCA: a tool for omic-data-based complex trait analysis. *Genome Biol.* **20**, 107 (2019).
64. First, M. B. et al. The Structured Clinical Interview for DSM-III-R Personality Disorders (SCID-II). Part II: Multi-Site Test-Retest Reliability Study. *J. Personal. Disord.* **9**, 92–104 (1995).
65. Hall, L. S. et al. Genome-wide meta-analyses of stratified depression in Generation Scotland and UK Biobank. *Transl. Psychiatry* **8**, 1–12 (2018).
66. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: A Tool for Genome-wide Complex Trait Analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).
67. Campagna, M. P. et al. Epigenome-wide association studies: current knowledge, strategies and recommendations. *Clin. Epigenetics* **13**, 214 (2021).
68. Suderman, M. et al. dmrff: identifying differentially methylated regions efficiently with power and control. <https://doi.org/10.1101/508556> (2018).
69. Hansen K. D. IlluminaHumanMethylationEPICanno.ilm10b2.hg19: Annotation for Illumina's EPIC methylation arrays. R package version 0.6.0. (2016). https://bitbucket.com/kasperdanielhansen/Illumina_EPIC.
70. Watanabe, K., Taskesen, E., van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat. Commun.* **8**, 1826 (2017).
71. THE GTEx CONSORTIUM, Ardlie, K. G. et al. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science* **348**, 648–660 (2015).
72. Koopmans, F. et al. SynGO: An Evidence-Based, Expert-Curated Knowledge Base for the Synapse. *Neuron* **103**, 217–234.e4 (2019).
73. Dolgalev I. msigdb: MSigDB Gene Sets for Multiple Organisms in a Tidy Data Format. (2022). <https://CRAN.R-project.org/package=msigdb>.
74. Phipson, B., Maksimovic, J. & Oshlack, A. missMethyl: an R package for analyzing data from Illumina's HumanMethylation450 platform. *Bioinformatics* **32**, 286–288 (2016).
75. Friedman, J. H., Hastie, T. & Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J. Stat. Softw.* **33**, 1–22 (2010).
76. Balduzzi, S., Rücker, G. & Schwarzer, G. How to perform a meta-analysis with R: a practical tutorial. *Evid. Based Ment. Health* **22**, 153–160 (2019).
77. Davyson, E. Insights from a Methylome-Wide Association Study of Antidepressant Exposure [Data set]. In *Nat. Commun.* <https://doi.org/10.5281/zenodo.14203230> (2024).
78. Davyson E. Insights from a Methylome-Wide Association Study of Antidepressant Exposure: Code. <https://doi.org/10.5281/zenodo.14185885> (2024).
79. Battram, T. et al. The EWAS Catalog: a database of epigenome-wide association studies. *Wellcome Open Res.* **7**, 41 (2022).
80. Hillary, R. F. et al. Blood-based epigenome-wide analyses of chronic low-grade inflammation across diverse population cohorts. <https://doi.org/10.1101/2023.11.02.23298000>.
81. Hillary, R. F. et al. Blood-based epigenome-wide analyses on the prevalence and incidence of nineteen common disease states. <https://doi.org/10.1101/2023.01.10.23284387>.
82. Breeze, C. E. et al. Epigenome-wide association study of kidney function identifies trans-ethnic and ethnic-specific loci. *Genome Med.* **13**, 74 (2021).
83. Dugué, P. A. et al. Alcohol consumption is associated with wide-spread changes in blood DNA methylation: Analysis of cross-sectional and longitudinal data. *Addict. Biol.* **26**, e12855 (2021).
84. Singmann, P. et al. Characterization of whole-genome autosomal differences of DNA methylation between men and women. *Epigenetics Chromatin* **8**, 43 (2015).
85. Mulder, R. H. et al. Epigenome-wide change and variation in DNA methylation in childhood: trajectories from birth to late adolescence. *Hum. Mol. Genet.* **30**, 119–134 (2021).
86. Zhang, L. et al. Epigenome-wide meta-analysis of DNA methylation differences in prefrontal cortex implicates the immune processes in Alzheimer's disease. *Nat. Commun.* **11**, 6114 (2020).
87. Hannon, E. et al. DNA methylation meta-analysis reveals cellular alterations in psychosis and markers of treatment-resistant schizophrenia. *Elife* **10**, e58430 (2021).
88. Spiers, H. et al. Methylomic trajectories across human fetal brain development. *Genome Res.* **25**, 338–352 (2015).

Acknowledgements

This work has made use of the resources provided by the Edinburgh Compute and Data Facility (ECDF, 2023) (<http://www.ecdf.ed.ac.uk/>). E.D. was supported by the United Kingdom Research and Innovation (grant EP/S02431X/1), UKRI Centre for Doctoral Training in Biomedical AI at the University of Edinburgh, School of Informatics. For the purpose of open access, the author has applied a creative commons attribution (CC BY) licence to any author accepted manuscript version arising. A.M.M. is supported by Wellcome (226770/Z/22/Z, 220857/Z/20/Z), UK Research and Innovation awards (Grant No. MR/W014386/1, MR/Z000548/1, MR/Z503563/1) and by European Union Horizon 2020 funding (Grant No. 847776). X.S. is supported by Wellcome (220857/Z/20/Z and 226770/Z/22/Z) and a UK Research and Innovation award (Grant no. MR/Z50354X/1). M.J.A. is supported by Wellcome (220857/Z/20/Z). F.H. is supported by the Netherlands Organization for Scientific Research (NWO 480-15-001/674) and Biobanking and Biomolecular Resources Research Infrastructure (BBMRI-NL: 184.021.007; 184.033.111) J.V.D. is supported by the Biobanking and Biomolecular Resources Research Infrastructure

(BBMRI-NL: 184.021.007). D.I.B. is supported by the Royal Dutch Academy for Arts and Science (KNAW) Academy Professor Award (PAH/6635); PI the Netherlands Organization for Scientific Research (NWO 480-15-001/674) and CoPI: Biobanking and Biomolecular Resources Research Infrastructure (BBMRI-NL: 184.021.007; 184.033.111). J.K. is supported by the Academy of Finland (grants 100499, 205585, 118555, 141054, 264146, 308248, 265240, 263278) and the Sigrid Juselius Foundation (grant #1057). M.O. is supported by the Academy of Finland (grants 328685, 307339, 297908 and 251316) and the Sigrid Juselius Foundation (grant #1105). A.S.F.K. is supported by the Wellcome Trust Early Career Award (227063/Z/23/Z). P.Y. is supported by National Institute for Health and Care Research (NIHR) Bristol Biomedical Research Centre and the Medical Research Council Integrative Epidemiology Unit at the University of Bristol (MC_UU_00032/4, MC_UU_00032/6). The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care. C.C.Y.W. is supported by the National Institute for Health and Care Research (NIHR) Maudsley Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London [NIHR203318]. H.L.F. is supported by the Economic and Social Research Council (ESRC) Centre for Society and Mental Health at King's College London [ES/S012567/1]. The views expressed are those of the authors and not necessarily those of the NIHR, the Department of Health and Social Care, the ESRC, or King's College London. J.M. has received support from the National Institute for Health and Care Research (NIHR) Exeter Biomedical Research Centre (BRC, NIHR203320). The views expressed are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care. S.R.C. is supported by the Biotechnology and Biological Sciences Research Council and the Economic and Social Research Council [BB/W008793/1, Age UK [Disconnected Mind project], the Medical Research Council [MR/R024065/1], the Milton Damerel Trust [Lothian Birth Cohorts], Sir Henry Dale Fellowship jointly funded by the Wellcome Trust and the Royal Society (221890/Z/20/Z). T.C.R. is supported by the Biotechnology and Biological Sciences Research Council and the Economic and Social Research Council [BB/W008793/1] and Age UK [Disconnected Mind project]. E.J.C.G.v.d.O and K.A.A. are supported by the National Institute of Mental Health (NIMH: R01MH124981). N.R.W. is supported by the Australian National Health and Medical Research Council (NHMRC: 1173790, 1113400). Generation Scotland (GS) & Stratifying Resilience and Depression Longitudinally (STRADL): We thank the participants in Generation Scotland for making this research possible. We also acknowledge the team at Generation Scotland for collecting and preparing the data for analyses. This work was supported by 3 Wellcome Trust grants (220857/Z/20/Z, 226770/Z/22/Z and 104036/Z/14/Z [AMM]). Funding from the Biotechnology and Biological Sciences Research Council and Medical Research Council (Grant Nos. MR/X003434/1 and MR/W014386/1) and the European Union (Grant agreement 847776) is also gratefully acknowledged. The Psychiatric Genomics Consortium (including AMM) has received major funding from the U.S. National Institute of Mental Health (Grant No. 5 U01MH109528-03). Netherlands Study of Anxiety and Depression (NESDA): The NESDA study is supported by the Geestkracht program of the Netherlands Organization for Health Research and Development (Zon-Mw, grant number 10-000-1002) and the participating institutions (VU University Medical Center, Leiden University Medical Center, University Medical Center Groningen). The current methylation project was supported by grant R01MH099110 from the National Institute of Mental Health. The sponsors had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication. Finnish Twin Cohort (FTC): Phenotype and genotype data collection in FT12 and FT16 studies of the Finnish twin cohort has been supported by the Wellcome Trust Sanger Institute, the Broad Institute, ENGAGE – European Network for Genetic and Genomic Epidemiology, FP7-HEALTH-

F4-2007, grant agreement number 201413, National Institute of Alcohol Abuse and Alcoholism (grants AA-12502, AA-00145, and AA-09203 to R J Rose; AA15416 and K02AA018755 to D M Dick; R01AA015416 to J Salvatore) and the Academy of Finland (grants 100499, 205585, 118555, 141054, 264146, 308248, 265240, 263278 to J.K., 328685, 307339, 297908 and 251316 to M.O., and Centre of Excellence in Complex Disease Genetics grants 312073, 336823, and 352792 to J.K.), NIH/NHLBI (grant HL104125 to X Wang), and the Sigrid Juselius Foundation to M.O and J.K. Study of Health in Pomerania (SHIP-Trend): SHIP is part of the Community Medicine Research net of the University of Greifswald which is funded by the Federal Ministry of Education and Research (01ZZ9603, 01ZZ0103, and 01ZZ0403), the Ministry of Cultural Affairs and the Social Ministry of the Federal State of Mecklenburg-West Pomerania. DNA methylation data have been supported by the DZHK (grant 81*3400104). The University of Greifswald is a member of the Caché Campus program of the InterSystems GmbH. FOR2107: We are deeply indebted to all study participants and staff. A list of acknowledgments can be found here: www.for2107.de/acknowledgements. The German multicenter consortium “Neurobiology of Affective Disorders. A translational perspective on brain structure and function” is funded by the German Research Foundation (Research Unit FOR2107). Principal investigators are Tilo Kircher (speaker FOR2107, DFG grant numbers KI588/14-1, KI588/14-2, KI588/20-1, KI588/22-1, KI 588/15-1, KI 588/17-1), Udo Dannlowski (co-speaker FOR2107; DA1151/5-1, DA1151/5-2, DA1151/6-1), Axel Krug (KR3822/5-1, KR3822/7-2), Igor Nenadic (NE2254/1-2, NE2254/2-1, NE2254/3-1, NE2254/4-1), Carsten Konrad (KO4291/3-1), Marcella Riettschel (RI 908/11-1, RI 908/11-2), Markus Nöthen (NO 246/10-1, NO 246/10-2), Stephanie Witt (WI 3439/3-1, WI 3439/3-2). This work was funded in part by the consortium grant Trajectories of Affective Disorders from the German Research Foundation (DFG) SFB/TRR 393 (project grant no 521379614). Biosamples and corresponding data were sampled, processed and stored in the Marburg Biobank CBBMR. Netherlands Twin Register (NTR): We warmly thank all twin families of the Netherlands Twin Register who make this research possible. This work was supported by the Royal Dutch Academy for Arts and Science (KNAW) Academy Professor Award (PAH/6635) to D.I.B.; the Netherlands Organization for Scientific Research (NWO 480-15-001/674) and Biobanking and Biomolecular Resources Research Infrastructure (BBMRI-NL: 184.021.007; 184.033.111). Munich Antidepressant Response Study / UniPolar Depression Study (MARS-UniDep): We would like to thank all contributors to the research project including physicians, psychologists, study nurses, researchers and research assistants, and of course patients of the hospital of the Max Planck Institute of Psychiatry in Munich and psychiatric hospitals in Augsburg and Ingolstadt. The MARS cohort was sponsored by the Max Planck Society. The UniDep cohort was funded by the Bavarian Ministry of Commerce and by the Federal Ministry of Education and Research in the framework of the National Genome Research Network, Foerderkennzeichen 01GS0481 and the Bavarian Ministry of Commerce. DNA methylation analysis of a subset of both cohorts was financed by ERA-NET NEURON. Avon Longitudinal Study of Parents and Children (ALSPAC): We thank all the families who took part in this study, the midwives for their help in recruiting them, and the whole ALSPAC team, which includes interviewers, computer and laboratory technicians, clerical workers, research scientists, volunteers, managers, receptionists and nurses. The UK Medical Research Council and Wellcome (Grant Ref: 217065/Z/19/Z) and the University of Bristol provide core support for ALSPAC. A comprehensive list of grants funding is available on the ALSPAC website (<http://www.bristol.ac.uk/alspac/external/documents/grant-acknowledgements.pdf>). This publication is the work of the authors and E.D., P.Y. & A.S.F.K. will serve as guarantors for the ALSPAC contents of this paper. E-risk Longitudinal Twin Study (E-risk): We are grateful to the E-Risk study mothers and fathers, the twins, and the twins' teachers for their participation. Our thanks to Professors Terrie Moffitt and Avshalom Caspi, the founders of the E-Risk study, and to the E-Risk team for their dedication, hard work,

and insights. The E-Risk Study is funded by grants from the UK Medical Research Council [G1002190; MR/X010791/1]. Additional support was provided by the US National Institute of Child Health and Human Development [HD077482] and the Jacobs Foundation. Generation of DNA methylation data in E-Risk was supported by a Senior Investigator Award from the American Asthma Foundation (AAF) to J.M. High-performance computing was supported by an MRC Clinical Infrastructure award (MR/M008924/1) and we acknowledge the contribution and use of the CREATE high-performance computing cluster at King's College London (King's College London (2022). King's Computational Research, Engineering and Technology Environment (CREATE), <https://doi.org/10.18742/rnvf-m076>.) Lothian Birth Cohort 1936 (LBC1936): The LBC1936 is supported by the Biotechnology and Biological Sciences Research Council, and the Economic and Social Research Council [BB/W008793/1], Age UK (Disconnected Mind project), the Medical Research Council [G0701120, G1001245, MR/M013111/1, MR/R024065/1], the Milton Damerel Trust, and the University of Edinburgh. Methylation typing of was supported by Centre for Cognitive Ageing and Cognitive Epidemiology (Pilot Fund award), Age UK, The Wellcome Trust Institutional Strategic Support Fund, The University of Edinburgh, and The University of Queensland.

Author contributions

E.D., A.M.M., R.E.M., and N.R.W. were responsible for the conception and the design of the study. E.D. conducted the primary data analyses, the MPS-meta-analysis and was primary author of the manuscript. X.S., M.J.A., D.L.M., and L.F.B., advised on methodology and data analysis. K.B. contributed to the derivation of prescription-derived antidepressant phenotypes. F.H., J.V.D., and D.B., ran the out-of-sample analysis in the Netherlands Twin Register. A.W., H.J.G., L.K., A. Teumer., and H.V., ran the out-of-sample analysis in Study of Health in Pomerania. T.Z., J.K., and M.O., ran the out-of-sample analysis in the Finnish Twin Cohort. F.S.D., S.M., F.S., A.J.F., U.D., and T.K., ran the out-of-sample analysis in FOR2107. A. Tapuc., D.C., E.B.B., and T.B. ran the out-of-sample analysis in the Munich Antidepressant Response Study / UniPolar Depression Study. A.S.F.K., and P.Y., ran the out-of-sample analysis in the Avon Longitudinal Study of Parents and Children. C.C.Y.W., ran the out-of-sample analysis in the E-Risk Longitudinal Twin Study. L.A., H.L.F., and J.M., secured funding, oversaw data collection and the derivation of variables for the E-Risk Longitudinal Twin Study. S.R.C., P.R., and T.C.R., provided guidance and expertise on running the out-of-sample analysis in the Lothian Birth Cohorts. E.J.C.G.v.d.O., K.A.A., and B.W.J.H.P., ran the enrichment analysis in NESDA. A.M.M., R.E.M., N.R.W., X.S., and M.J.A. co-supervised the project. All authors read and approved the final manuscript.

Competing interests

R.E.M. is an advisor to the Epigenetic Clock Development Foundation and Optima Partners. D.L.M. was employed by Optima Partners Ltd in a

part-time capacity. H.J.G. has received travel grants and speaker honoraria from Fresenius Medical Care, Neuraxpharm, Servier, Indorsia and Janssen Cilag, as well as research funding from Fresenius Medical Care. H.J.G. had personal contracts approved by the university administration for speaker honoraria and one IIT with Fresenius Medical Care. T.K. received unrestricted educational grants from Servier, Janssen, Recordati, Aristot, Otsuka, neuraxpharm. B.W.J.H.P. has received research funding (not related to the current paper) from Boehringer Ingelheim and Jansen Research. All other authors report no biomedical financial interests or potential conflicts of interest.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41467-024-55356-x>.

Correspondence and requests for materials should be addressed to A. M. McIntosh.

Peer review information *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2025

¹Division of Psychiatry, Centre for Clinical Brain Sciences, University of Edinburgh, Edinburgh, UK. ²Institute of Genetics and Cancer, University of Edinburgh, Edinburgh, UK. ³Complex Trait Genetics, Center of Neurogenomics and Cognitive Research, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands. ⁴Amsterdam Public Health Research Institute, Amsterdam, The Netherlands. ⁵Department of Biological Psychiatry, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands. ⁶Institute for Molecular Bioscience, University of Queensland, Brisbane, QLD 4072, Australia. ⁷Amsterdam Reproduction & Development, Research Institute, Amsterdam, The Netherlands. ⁸Department of Psychiatry and Psychotherapy, University Medicine Greifswald, 17475 Greifswald, Germany. ⁹German Center for Neurodegenerative Diseases (DZNE), Site Rostock/Greifswald, 17489 Greifswald, Germany. ¹⁰German Centre for Cardiovascular Research (DZHK), Partner Site Greifswald, 17489 Greifswald, Germany. ¹¹Department SHIP/Clinical-Epidemiological Research, Institute for Community Medicine, University Medicine Greifswald, 17475 Greifswald, Germany. ¹²Institute for Molecular Medicine Finland FIMM, HiLIFE, University of Helsinki, Helsinki, Finland. ¹³Minerva Foundation Institute for Medical Research, Helsinki, Finland. ¹⁴Institute of Human Genetics, University of Bonn, School of Medicine & University Hospital Bonn, Bonn, Germany. ¹⁵Department of Psychiatry and Psychotherapy, University of Marburg, Marburg, Germany. ¹⁶Institute for Translational Psychiatry, University of Münster, Münster, Germany. ¹⁷Institute for Translational Neuroscience, University of Münster, Münster, Germany. ¹⁸Center for Mind, Brain and Behavior, University of Marburg, Marburg, Germany. ¹⁹Institute of Neuroscience and Medicine (INM-1), Research Center Jülich, Jülich, Germany. ²⁰Center for Human Genetics, University of Marburg, Marburg, Germany. ²¹Max Planck School of Cognition, Leipzig, Germany. ²²Max-Planck-Institute of Psychiatry, Department Genes and Environment, Munich, Germany. ²³Medical Research Council Integrative Epidemiology Unit at the University of

Bristol, University of Bristol, Bristol, UK. ²⁴Population Health Science, Bristol Medical School, University of Bristol, Bristol, UK. ²⁵NIHR Bristol Biomedical Research Centre, University Hospitals Bristol and Weston NHS Foundation Trust and University of Bristol, Bristol, UK. ²⁶Social, Genetic & Developmental Psychiatry Centre, Institute of Psychiatry, Psychology & Neuroscience, King's College London, London, UK. ²⁷ESRC Centre for Society and Mental Health, King's College London, London, UK. ²⁸Department of Clinical & Biomedical Sciences, University of Exeter Medical School, University of Exeter, Exeter, UK. ²⁹Lothian Birth Cohorts, Department of Psychology, University of Edinburgh, Edinburgh, UK. ³⁰Alzheimer Scotland Dementia Research Centre, University of Edinburgh, Edinburgh, UK. ³¹Neuroprogressive and Dementia Network, NHS Research Scotland, Scotland, UK. ³²Center for Biomarker Research and Precision Medicine (BPM), Virginia Commonwealth University, Virginia, USA. ³³Department of Psychiatry, Amsterdam UMC, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands. ³⁴Department of Psychiatry, University of Oxford, Oxford, UK. ✉ e-mail: andrew.mcintosh@ed.ac.uk

5.4 Conclusion

This analysis demonstrates robust associations between antidepressant exposure and DNAm. Specifically, the MWAS found evidence of hypermethylation at eight CpGs across the genome in those exposed to antidepressants. These results demonstrated a concordance of findings when using self-reported or prescription-derived measures of antidepressant exposure, which has important implications for future epidemiological research on antidepressant use. Specifically, the methods for deriving prescription-derived exposure at a certain time point presents a useful framework for future studies on antidepressant exposure and DNAm. The most significant CpG with the largest effect mapped to *KANK1*, which has an active role in neurite outgrowth. Additionally, regional analysis found evidence of a significantly hypermethylated region near the long non-coding RNA (lncRNA), *DGUOK-AS1*, which inhibits a mitochondrial enzyme. Functional annotation also found an enrichment of top signals for genes in a synaptic vesicle ontology and for genes expressed in the amygdala. Furthermore, I found nominally significant evidence of a larger effect (more hypermethylation) in females compared to males, which could be linked to the reported differential antidepressant response between the two sexes. Future experimental analysis could further assess the functional impact of these hypermethylated sites, whilst the integration of imaging data could provide insight into their potential effect on amygdala function.

Importantly, this analysis found that DNAm associations appear specific to antidepressant use rather than MDD itself, as an MWAS on MDD cases alone found similar effect sizes for significant CpGs. In addition, the MPS was significantly associated with antidepressant exposure in external independent datasets,

demonstrating that these are robust and generalisable findings. Given these findings, it is possible that antidepressant exposure may be a confounding factor in DNAm studies of neuropsychological traits, or any trait where there may be a mismatch in antidepressant exposure between cases or controls. Therefore, adjustment for antidepressant exposure or sensitivity analyses in antidepressant naïve individuals may be necessary to delineate trait specific DNAm effects in the future for these cases.

An important limitation of these findings is that all cohorts include individuals of predominantly European ancestry, and potential ancestry-specific effects remain to be investigated. Finally, whether these observed associations are due to causality require additional causal inference methods (e.g. MR), or ultimately the integration of DNAm into randomised controlled trials of antidepressant exposure.

6 Discussion

In this thesis, I have illustrated how the analysis of multi-omics data in large population studies can provide biological insights to MDD and antidepressant exposure. For MDD, I systematically examined the associational and potential causal impact of 249 circulating metabolites (**Chapter 3**). This analysis found evidence of a causal relationship between an omega-3 fatty acid and MDD, colocalised in the *FADS* region. In **Chapter 4**, I attempted to identify potential causal genes underlying this relationship by analysing the burden of rare pathogenic variants in each gene, and their effects on MDD and omega-3 related measures. Finally, in **Chapter 5**, I comprehensively assessed the association of antidepressant exposure with differential DNAm. In this chapter, I discuss the findings from the analysis in **Chapters 3-5**. I begin with an evaluation of the metabolomic links to MDD (**Chapters 3 and 4**), followed by a discussion of antidepressant exposure associations with DNAm (**Chapter 5**). Then, I systematically evaluate limitations to this thesis, and potential future directions which can mitigate these and build upon these findings.

6.1 The role of metabolites in MDD

6.1.1 *Summary of findings*

My analysis in **Chapter 3** showed that MDD was associated with many different circulating metabolites, indicating a distinct metabolomic profile. However, only five omega-3 related measures had a potentially causal impact on MDD from MR.

Specifically, it was found that omega-3 fatty acids and their concentrations had a negative (protective) effect whilst the ratio of omega-6: omega-3 had a positive effect on MDD. Notably, only DHA had complementary causal evidence from colocalisation, with a high probability of a shared causal variant located in the *FADS* cluster. The *FADS* cluster is evolutionarily conserved, and therefore common variants exhibit high levels of LD, making the identification of the causal gene in this region challenging. In **Chapter 4**, I attempted to identify potential causal genes underlying the relationship between DHA and MDD by using rare genetic variants identified from WGS. This analysis found distinct signals in the *FADS* cluster for MDD and circulating omega-3 related measures, indicating they may not share a causal pathway. Specifically, there were strong associations with rare genetic variants in *TMEM258*, *MYRF* and *FEN1* on DHA concentrations, but no observed association with MDD. There was a nominal association of rare genetic variants in *FADS1* with MDD, which appeared dependent on a singular variant in Exon 1.

6.1.2 *Omega-3 and omega-6 Fatty Acids in MDD*

Omega-3 and omega-6 fatty acids are both long-chain ($n > 18$) polyunsaturated acids, with a carbon-carbon double bond in the 3rd and 6th position of the methyl group respectively. They can be present in circulation in their free form or esterified into other lipid molecules (e.g. phospholipids). Both omega-3 and omega-6 acids have been implicated in a range of different physiological functions, including energy production, tumour suppression, metabolic flux, membrane fluidity, cell signalling and mediating inflammatory responses⁵²³. Importantly, omega-3 and omega-6 fatty acids cannot be endogenously synthesised in the human body, and their production depends on the dietary intake of precursor molecules, alpha-linoleic acid (ALA) and linoleic acid (LA) respectively (Figure 6.1)^{523,524}. LA is found in high quantities in vegetable and seed oils and can be metabolised into arachidonic acid (AA), via a series of intermediate desaturation (making a carbon-carbon double bond) and elongation (extending the carbon chain) steps⁵²⁵. ALA, commonly found in flaxseeds and nuts, can be metabolised into eicosapentaenoic acid (EPA) in the same metabolic process, which can then be converted into docosapentaenoic acid (DPA) and docosahexaenoic acid (DHA). Notably, downstream omega-3 (EPA, DPA, DHA) and omega-6 (AA) fatty acids can also be obtained directly from the diet, from fatty fish and fish oil supplements (EPA, DPA, DHA) and from poultry and eggs (AA)⁵²³. Generally, it is thought that free omega-6 and omega-3 fatty acids have opposing effects on inflammatory processes. Omega-6 fatty acids (LA and AA) produce various pro-inflammatory small molecules called eicosanoids whereas omega-3 fatty acids (EPA and DHA) produce many anti-inflammatory small molecules such as resolvins and maresins⁵²⁵. Crucially, the

metabolism of both omega-6 and omega-3 precursors (LA and ALA) relies upon the FADS1/2 enzymes⁴⁶⁵. Therefore, the ratio of dietary omega-6: omega-3 is important as they directly compete for the FADS enzymes and may result in an imbalance in downstream pro/anti-inflammatory mediators⁵²⁶. Since the 1950s, there has been a dramatic increase in the typical dietary ratio of omega-6: omega-3 fatty acids (~ 10-20:1 compared to 1:1), through a mirrored increase in polyunsaturated oils in industrial food production⁵²⁷. This increase was partly due to recommendations that the replacement of saturated fats (from animal products) by polyunsaturated oils may prevent atherosclerosis and risk of cardiovascular events⁵²⁸. However, a large omega-6: omega-3 ratio and subsequent imbalance of downstream mediators may lead to dysregulated inflammatory responses and has been implicated in a range of chronic diseases, including asthma, rheumatoid arthritis and various cancers⁵²⁶.

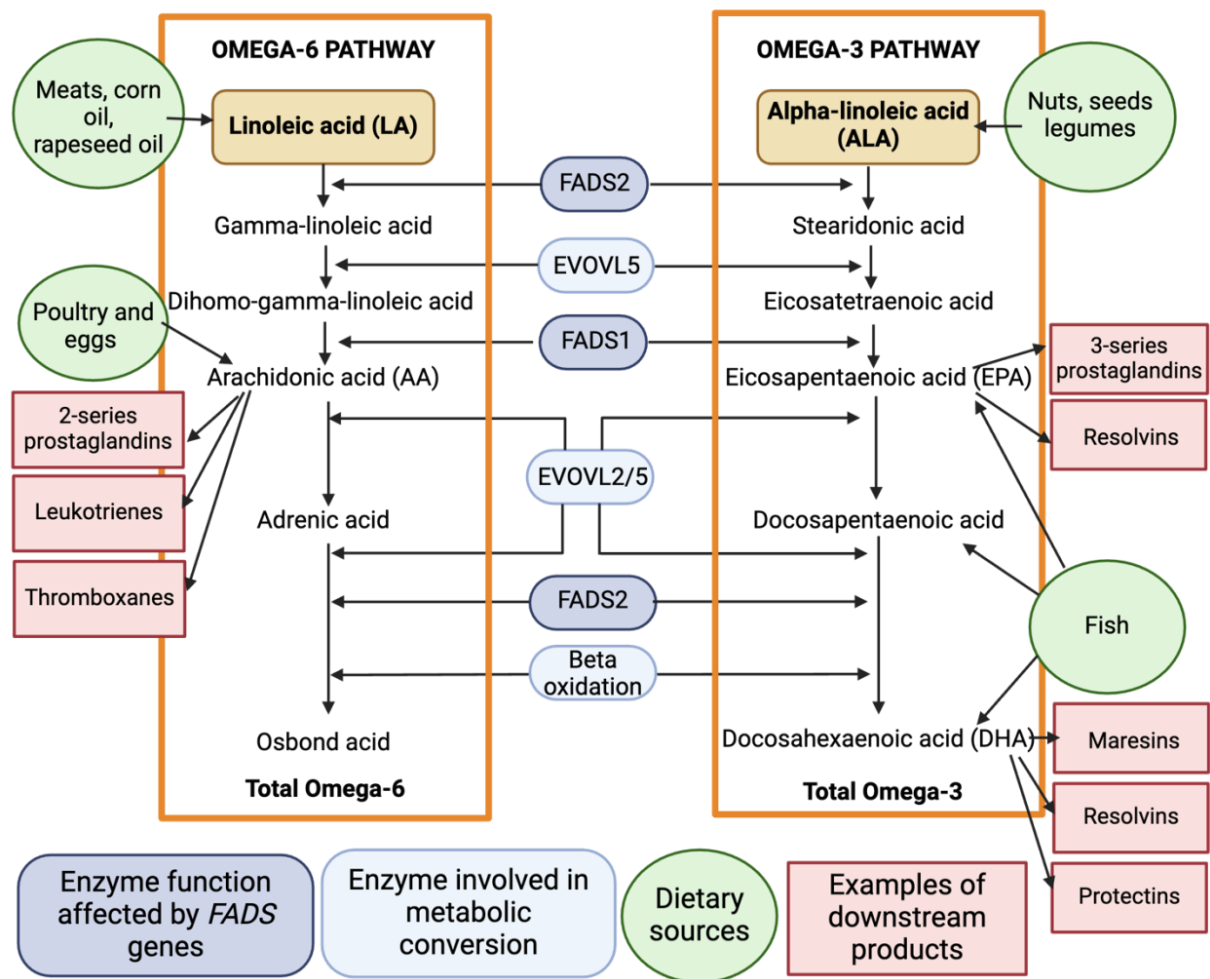


Figure 6.1 Metabolic pathways of omega-3 and omega-6 fatty acids from dietary precursors.

This figure was made using Biorender and adapted from Carnegie et al (2024)⁴⁴⁶.

Copyright: *Translational Psychiatry*.

In **Chapter 3**, we observed a potential causal effect of higher omega-6: omega-3 ratio on MDD, whilst omega-3 measures were causally linked to a lowered odds-ratio of MDD. This could indicate that increasing omega-3 levels would be protective against MDD, primarily through its anti-inflammatory actions. However, many randomised clinical trials of omega-3 supplementation have had mixed findings, attributed in part due to heterogenous methodological design and small sample sizes⁵²⁹. A recent meta-analysis indicated a beneficial effect of EPA and EPA+DHA supplementation in MDD at certain doses but also found evidence of publication bias which may influence the results⁵²⁹. In addition, a recent large-scale RCT (VITAL-Dep, n ~ 18,000) found long-term supplementation of omega-3 (EPA or DHA) did not reduce the risk of MDD but was in fact associated with a small (but significant) increase in clinical depressive symptoms⁴²⁶. Therefore, the exact nature of the causal relationship between omega-3 and omega-6 fatty acid metabolism and MDD appears more complex than an imbalance between omega-3 and omega-6 fatty acids which can be reversed by dietary supplementation.

Omega-3 fatty acids and omega-6 fatty acids are often grouped together with generalised roles in inflammation^{523,526}. Yet specific omega-3 and omega-6 fatty acids have different biological functions which can also depend on their cellular context (free or incorporated into different lipid molecules)^{523,524}. However, the NMR measures used in this thesis reflect total omega-3 and omega-6 concentrations so cannot investigate specific roles of individual fatty acids or more complex structures containing them. Mass-spectrometry based methods can identify more metabolites and may help to disentangle this relationship²⁸⁵. A recent large MR study utilised mass spectrometry metabolomics data to assess the causal nature of ~ 1000 metabolites in a range of

psychiatric disorders⁴⁴². Notably, nearly all the lipids colocalised with psychiatric disorders contained omega-6 fatty acids LA or AA, however LA-containing metabolites increased risk and AA-containing metabolites were protective. Another independent MR study found genetically predicted AA and another intermediate, gamma-linoleic acid (GLA) were protective against SCZ⁴³⁹ and BPD⁴⁴³. These data indicate that the dichomatisation of omega-3 and omega-6 fatty acids as “good” and “bad” for risk of psychiatric traits based on broad inflammatory effects is likely oversimplifying a complex intricate system. Also, untargeted MR studies have implicated other complex lipid species in MDD, such as sphingomyelins and lysophospholipids^{311,442}.

It is important to acknowledge that the primary MR analyses in this thesis and the majority of metabolite-MDD MR estimates in the literature were using genetic instruments in the *FADS* cluster in Chromosome 11, a highly pleiotropic and LD-dense region. The *FADS* locus and the impact of genetic instruments from *FADS* in MR interpretation will therefore be discussed.

6.1.3 *Fatty Acid Desaturase (FADS) cluster*

The *FADS* cluster includes three fatty acid desaturase genes (*FADS1*, *FADS2* and *FADS3*) which encode desaturase enzymes (*FADS1*, *FADS2* and *FADS3*)⁴⁶⁵. Both *FADS1* and *FADS2* have well-defined roles as the rate-limiting enzymes in omega-3 and omega-6 metabolism pathways (Figure 6.1), whereas the exact function of *FADS3* is unclear, but has been implicated in DHA accumulation in the brain during development⁴⁶⁸. The *FADS* cluster is a highly evolutionarily conserved genetic region

and therefore displays high LD⁴⁷⁰. Two distinctive haplotypes in the *FADS* region have been identified, which each have varying frequencies across different ancestral backgrounds^{470,530}. The derived haplotype (“D”) is associated with more efficient synthesis of fatty acids, which is evolutionarily advantageous in environments with limited access to their dietary precursors⁴⁷⁰. However, coupled with the increased abundance of omega-6 in modern western diets, the D haplotype may heighten the imbalance between omega-6: omega-3 derived inflammatory markers.

Various recent analyses have shown the widespread effect of genetic variants in *FADS* on hundreds of different metabolomic traits, indicating the high pleiotropy of the *FADS* region^{351,531,532}. One particular analysis found that common SNPs in *FADS* were significantly associated with 52 complex lipid species and that the direction of effect appeared to be dependent on the activity of *FADS1*⁵³⁰. Specifically, they observed that individual minor allele variants were associated with higher concentrations and lower concentrations of lipid species upstream and downstream of *FADS1* enzymatic action respectively. Therefore, they hypothesise *FADS1* to be a central genetic control point for a range of circulating molecular lipid species including phospholipids, lysophospholipids, free fatty acids and endocannabinoids which all have important physiological roles⁵³⁰. Furthermore, phenome-wide association studies (PheWAS) of a SNP with high LD in the *FADS* (*rs174546*) region was associated with a range of cardiovascular, immunological and psychiatric traits³¹¹. This broad pleiotropy of *FADS* means genetic instruments in this region could be affecting a trait (such as MDD) through a range of different mechanisms (Figure 6.2) and therefore violates the exclusion-restriction assumption and may bias causal estimates from MR³⁸³. Therefore,

many have argued for caution when interpreting MR causal estimates which rely genetic instruments within the *FADS* cluster^{311,445}. Several studies have attempted to mitigate the potential bias from horizontal pleiotropy in *FADS* using multi-variate MR (MVMR) methods for genetically correlated metabolites^{446,447}, as well as utilising genetic instruments outside of the *FADS* region⁴⁴³. Notably, Carnegie et al, (2024)⁴⁴⁶ observed protective causal effect of omega-3 fatty acids (EPA and DHA) which was maintained when including various genetically correlated lipids in MVMR analysis⁴⁴⁶. Also, the MVMR findings indicated a stronger effect of EPA than DHA on MDD, which may be a factor in mixed results from EPA and DHA supplementation trials⁴⁴⁶. Overall, there is robust evidence from various studies of the causal impact of dysregulated *FADS* metabolism in MDD. However, as *FADS* metabolism tightly controls many downstream mediators and related lipid species, each with diverse biological functions, the identification of exact pathological mechanism for MDD is challenging (Figure 6.2).

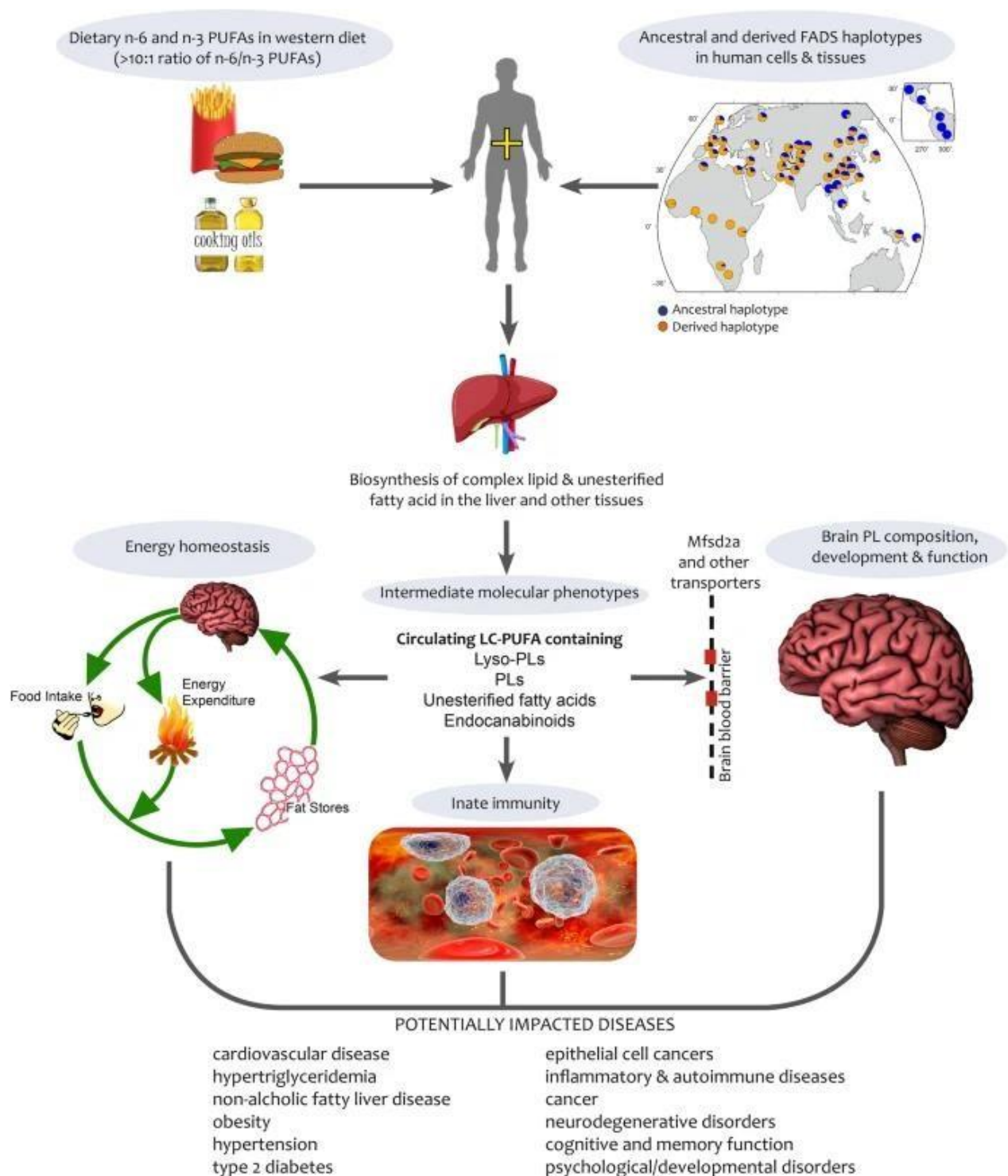


Figure 6.2 Impact of dietary omega-3 and omega-6 levels and *FADS* haplotypes on lipid species and physiological processes.

A schematic of how levels of long-chain polyunsaturated fatty acids (LC-PUFAs) can interact with global variation in *FADS* to impact levels of various LC-PUFA containing metabolites, including lysophospholipids (lyso-PL), phospholipids (PL), endocannabinoids and unesterified fatty acids. Changes in circulating and tissue levels of these lipids can impact energy homeostasis, immune responses, and brain PL composition. These mechanisms influence a range of physiological processes and potentially impact a wide variety of human diseases. This figure is taken from Reynolds et al (2020)⁵³⁰. Copyright: *Nature Scientific Reports*.

6.1.4 *Potential mechanisms*

In **Chapter 4**, various rare, prioritised variants in the *FADS* cluster were strongly associated with changes in circulating omega-3 levels alongside the omega-6: omega-3 ratio. However, these same variants were not associated with MDD. Whilst consistent associations are not proof of causality, this lack of shared signal indicates that circulating omega-3 and omega-6 levels captured by NMR spectroscopy may not be the causal effectors in MDD pathophysiology. One possibility is tissue specific effects of altered *FADS* metabolism on omega-3 and omega-6 levels. The brain has a unique fatty acid composition, with brain phospholipids being specifically enriched for DHA and AA compared to other omega-3 fatty acids (DPA and EPA)⁵³³. High levels of DHA and AA in neuronal cell membranes can affect membrane fluidity which in turn influences a range of synaptic processes⁵³³. In particular, DHA was found to accumulate in cholesterol-rich lipid rafts which impacts downstream cellular signalling and synaptic activity^{533,534}. A recent study reported the role of cholesterol-rich lipid rafts in mediating the response of

antidepressants at the TrkB receptor, through prolonging signalling and facilitating BDNF colocalisation to synapses⁵⁸. Therefore, changes in brain DHA levels may contribute to the pathophysiology of MDD through altered neuronal membrane composition and downstream effects on synaptic plasticity. Various rodent models have shown that depletion of DHA is related to altered brain transmission⁵³⁵, reduced *BDNF* expression⁵³⁶, deficits in working memory⁵³⁷ and abnormal behaviour in the forced swim test⁵³⁵. Additionally, DHA accumulates early in development and has effects on various processes related to plasticity, such as neurogenesis, synaptic pruning and neurite outgrowth⁵³³.

As the brain has very low de novo synthesis of DHA and primarily relies on hepatic metabolism and transport across the blood-brain barrier (BBB), blood DHA levels may be expected to reflect the composition of DHA in the brain⁵³⁸. However, it has been shown that the transport of DHA across the BBB is by a transporter, *Mfsd2a*, which specifically transports fatty acids in the form of lysophosphatidylcholines (LPC), rather than free unesterified fatty acids⁵³⁸. Therefore, circulating omega-3 and omega-6 levels measured in bulk do not necessarily capture the content of these fatty acids in the brain. In **Chapter 4**, I identified a nominally significant association between rare pathogenic variants in *FADS1* and MDD, which appeared to be driven by a singular missense variant in Exon 1. This variant converts a Proline into an Arginine at position 39 of *FADS1* enzyme, however little is known about its functional impact on *FADS1* activity and downstream metabolites. Experimental functional assays could help to establish any functional effects on *FADS1* enzymatic activity, whilst genetically modified in-vivo

models could investigate the effect of this variant on neuronal membrane properties which may be relevant to MDD.

Overall, my thesis and various other studies have demonstrated the potential causality between dysregulated fatty acid metabolism in MDD, with colocalisation in the *FADS* region. Furthermore, my thesis indicates a potential driver variant in Exon 1 of *FADS1* which may be related to MDD. However, there is growing evidence of *FADS* being a central gatekeeper for a host of diverse lipid species which have a range of biological functions, making the identification of effector mechanisms relevant to MDD difficult. Potential mechanisms include chronic inflammatory response and altered synaptic plasticity, both processes which are implicated in MDD. However, given the complexity of the fatty acid metabolism pathway, measurement of bulk fatty acid levels cannot inform on precise mechanisms. More comprehensive measurement of fatty acids in their various forms through untargeted metabolomics and the incorporation of brain tissue could inform more on the potential impact from *FADS* relevant to MDD.

6.2 The association of antidepressants with DNAm

6.2.1 *Summary of findings*

In **Chapter 5**, I conducted the largest study to date of the association between DNA methylation and antidepressant exposure in GS. There were 8 CpGs which were significantly hypermethylated in those exposed to antidepressants, which had nominally significant larger effects in females and showed no evidence of confounding by indication. A differentially methylated region was also identified, mapping to *DGUOKAS1*, a lncRNA. There was also correlation between methylation levels at cg15071067 (*DGUOK-AS1*) and cg26277237 (*KANK1*) and time taking antidepressants prior to blood draw. Additionally, top associated CpGs and annotated genes had significant enrichment for genes expressed in the amygdala and for those in a “synaptic vesicle membrane” ontology. Finally, a MPS based off observed DNAm associations in GS was significantly associated with antidepressant exposure in multiple independent cohorts, indicating generalisability of these findings.

6.2.2 *Biological implications of associations*

The majority of previous studies on antidepressant treatment and DNAm have been conducted in animal models and have had a focus on candidate genes or regions hypothesised to be involved in antidepressant mechanisms, such as *BDNF*⁵³⁹ or *SLC6A4*⁵⁴⁰. However, methylome-wide approaches, which can give a more comprehensive evaluation of the effects of antidepressants in humans have not identified differential methylation at these genes³⁴³. A previous MWAS in GS was

conducted in self-report antidepressant exposure in a subset of participants (N = 6,428)³⁴³. This analysis also identified hypermethylation in those who self-report antidepressant exposure at *cg26277237*, alongside various DNAm associations with roles in the innate immune system³⁴³. The analysis in this thesis (N = 16,531) found hypermethylation at several CpGs annotated to genes with roles in transcriptional regulation (*DOT1L*, *GTF3C1* and *ZNP710*), neurite outgrowth (*KANK1*) and mitochondrial function (*DGUOK-AS1*). Some of these genes have also previously been linked to other neuropsychiatric and neurodegenerative disorders⁵⁴¹⁻⁵⁴³. For example, *DOT1L* is involved in catalysing the methylation of histone H3 and is involved in susceptibility to chronic social defeat in animal models following early life stress, a prominent risk factor in MDD⁵⁴⁴. However, as a broad post-translational modification, *DOT1L* also has a role in a range of biological functions, including orchestrating immune response⁵⁴⁵ and regulating the cell cycle, and is involved in several cancers⁵⁴⁶.

Overall, there was strongest evidence of hypermethylation at *DGUOK-AS1*, a lncRNA, from individual CpGs (*cg15071067* and *cg01964004*), regional analysis, and correlation between DNAm and time in treatment. Long non-coding RNAs have historically been overlooked but are increasingly being shown to have diverse regulatory roles including chromatin remodelling, histone modifications and post-transcriptional regulation⁵⁴⁷. Specifically, *DGUOK-AS1* has many different downstream targets, including several microRNAs (miRNA) involved in immune function^{548,549}. Most notably, *DGUOK-AS1* affects miR-204-5p, which in turn primarily targets the cytokine interleukin-11 (IL-11). Studies have shown that through this axis, *DGUOK-AS1* increases IL-11 secretion, leading to increased angiogenesis, macrophage migration and the progression of

various cancers^{549,550}. Moreover, the expression levels of *DGUOK-AS1* and *DGUOK* are negatively correlated, indicating a regulatory role of *DGUOK-AS1* on the expression of *DGUOK*, common to lncRNA target gene pairs^{547,550,551}. *DGUOK* encodes a mitochondrial enzyme involved in the mitochondrial purine metabolism pathway, specifically as a rate limiting step in the production of mitochondrial DNA (mtDNA) required for adenosine triphosphate (ATP) production^{550,552}. A recent animal study investigated the effects of fluoxetine treatment on gene expression across various brain regions, using bulk and single cell RNA sequencing (RNA-seq) alongside chromatin immunoprecipitation sequencing (ChIP-seq)⁵⁵³. They observed regional and cell specific shifts in gene expression and chromatin states, which may not be captured using the bulk tissue data presented in this thesis. However, alongside cell-specific effects, they observed increased energy metabolism via mitochondrial changes in the majority of brain regions following treatment, indicating a more global effect of antidepressant treatment across the brain on mitochondrial bioenergetics⁵⁵³.

6.2.3 *The potential role of mitochondrial dysfunction in MDD*

Mitochondrial dysfunction has been hypothesised to play a role in MDD⁵⁵⁴⁻⁵⁵⁶, due to the high comorbidity of MDD with mitochondrial disorders and the impact of mitochondria on biological processes relevant to MDD⁵⁵⁴. The mitochondria are known as the “powerhouse” of the cell, through the production of ATP which is required for all cellular processes⁵⁵⁶. The brain is the most energy intensive organ in the body, estimated to expend up to 20% of the oxygen in the body, whilst representing 2% of the body weight⁵⁵⁷. Furthermore, the brain is more reliant on aerobic ATP production by

mitochondria as it has limited glycogen energy stores⁵⁵⁴⁻⁵⁵⁶. Together, this makes the brain very vulnerable to any impairment in mitochondrial energy production, and deficits may disproportionately affect energy-intensive processes such as synaptic plasticity⁵⁵⁶. Mitochondria also play a critical role in regulating cell death (“apoptosis”) and intracellular Ca^{2+} ions, which are both involved in processes related to synaptic plasticity. Controlled apoptosis is required for synaptic pruning, whereby unnecessary synaptic connections are removed to optimise neural networks, which underlies learning and memory and is especially important during development⁵⁵⁸. Furthermore, intracellular Ca^{2+} can regulate downstream cellular signalling following neurotransmission, most notably for the NMDA receptor, which is required for long-term potentiation (LTP) of synaptic connections, and which is targeted by Ketamine for rapid-acting antidepressant effects⁶⁴. As well, mitochondria release reactive oxygen species (ROS) free radicals as a by-product of ATP, which when dysregulated can lead to intracellular damage, increased apoptosis and rise in inflammation⁵⁵⁴. Notably, mitochondrial function is also linked to lipid metabolism, as β -oxidation of long-chain fatty acids is estimated to account for ~ 20% of the brains’ ATP production⁵⁵⁹. Specifically, L-carnitines, lipid molecules which regulate fatty acid oxidation in mitochondria have been found to be dysregulated in MDD in two independent cohorts³¹⁰. Therefore, mitochondrial function in the brain has roles in synaptic plasticity and neuroinflammation, two processes which are hypothesised to be core to MDD, and has links to lipid metabolism, which is increasingly being shown to be dysregulated in MDD.

Several studies have found that neuropsychiatric medications, including antidepressants have an influence on mitochondrial bioenergetics and function⁵⁶⁰. However, the nature

and direction of effect is unclear, with heterogeneous results between different antidepressant drugs, even those in the same class. For example, two SSRIs, escitalopram and sertraline consistently show negative and positive impact on mitochondrial function respectively^{561,562}. However, these studies largely use electron transport chain (ETC) activity to measure mitochondrial function, but there are various other mechanisms which can impact mitochondrial function with no direct action on ETC activity, e.g. mitochondrial bioenergetics⁵⁶⁰. Additionally, as there is no current clinically validated and non-invasive test of mitochondrial function in humans, the current evidence is limited to rodent models, which may not be representative of antidepressant effects in humans⁵⁶⁰.

In **Chapter 5**, hypermethylation at *DGUOK-AS1* was significantly associated with antidepressant exposure, which may be linked to increased *DGUOK* expression and mitochondrial function. However, whether these associations are due to a true causal relationship requires investigation of DNAm measures in RCTs of antidepressants, or the use of causal inference methods such as MR. Furthermore, although *DGUOK-AS1* is linked to immune regulation and mitochondrial function, the exact impact of hypermethylation at *DGUOK-AS1* on biological function is not well characterised and requires further investigation. Finally, antidepressant exposure in **Chapter 5** did not distinguish between drug types, dosages or medication history and therefore the findings represent broad DNAm associations with commonly prescribed antidepressants. However, it is likely that the integration of EHR and DNAm in large population cohorts will enable the study any medication-specific effects in the future.

6.2.4 *Insights into potential side effects*

Commonly prescribed antidepressants are accompanied by a range of side effects, including nausea, dizziness, insomnia, tremor and sexual dysfunction⁹¹. The side effect profile can limit their use in certain individuals and contribute to treatment resistance in MDD¹²¹. Furthermore, antidepressants are often prescribed continuously and there is growing concern about impact of long-term exposure to antidepressants⁵⁶³⁻⁵⁶⁶. One recent large study in UKB observed an association between long-term antidepressant use and heightened risk of coronary heart disease, cardiovascular disease, mortality and all-cause mortality at 10 years follow-up⁵⁶⁶. As well, these findings differed between antidepressants, with atypical and SNRI antidepressants (e.g. mirtazapine, venlafaxine and duloxetine) being associated with much higher risk of these outcomes than SSRIs. There was also evidence that SSRIs were associated with reduced risk of Type 2 Diabetes (T2D)⁵⁶⁶. Notably, many of the CpGs significantly associated with antidepressant use in **Chapter 5** have been reported previously to be significantly associated with incident and/or prevalent T2D²⁶² with the same direction of effect. Various other large epidemiological studies have indicated that antidepressant use leads to an increased risk of T2D onset in a time and dose dependent manner^{564,567,568}. Whilst this study confirms overlapping DNAm changes associated with T2D and antidepressant exposure, it is unable to establish the cause and nature of the overlap. Additionally, the significant association with T2D was also observed in a separate study of the same cohort (GS) and therefore do not represent wholly independent findings²⁶². Future prospective and longitudinal research into the link between antidepressant use, DNAm and T2D is required.

There have also been increased reports of post-SSRI sexual dysfunction (PSSD), when sexual dysfunction side effects of SSRIs persist following discontinuation of treatment^{563,569,570}. However, the prevalence of PSSD is unknown with varied estimates from 2 to 50%⁵⁷¹ and there is a lack of knowledge regarding potential mechanisms underlying PSSD⁵⁶⁹. Further analysis on whether the DNAm associations are sustained after cessation of antidepressant treatment may provide insight into long-term biological impact of treatment, which may be relevant to PSSD. One CpG significantly associated with antidepressant exposure was annotated to *CCDC146*, which has recently been linked to spermatogenesis and male infertility⁵⁷². Specifically, a cohort of 167 infertile male patients with multiple morphological defects in their sperm were found to have various biallelic mutations in *CCDC146*, and *CCDC146* knock out mice models were also infertile, with similar morphological phenotypes⁵⁷². A recent meta-analysis found consistent evidence of SSRIs having an adverse effect on sperm quality, although they conclude there is inconsistent evidence on general effects of SSRIs on fertility across the sexes⁵⁷³. The relationship between antidepressant exposure and *CCDC146* DNAm observed in **Chapter 5** therefore warrants further research to ascertain potential effects on sperm quality and male infertility.

6.2.5 *Treatment response*

The research in this thesis focuses on antidepressant exposure rather than antidepressant response. Therefore, the DNAm associations with antidepressant exposure observed in this thesis may not necessarily be associated with their therapeutic effects. However, investigating antidepressant exposure increases the

power of the discovery, but also enables a comprehensive overview of DNAm associations with antidepressant use, which can then inform and focus future studies on antidepressant response. So far, DNAm studies on antidepressant response have had heterogeneous findings, however they also exhibit either an a-priori focus on genes hypothesised to be involved in MDD and/or small samples, limiting the power of these analyses to comprehensively evaluate potential DNAm mediators of antidepressant response^{323,574-576}. This thesis presents empirical evidence of DNAm associations with antidepressant exposure which were generalisable across other cohorts. Therefore, these DNAm sites could be used as an a-priori focus instead of hypothesised genes. Overall, a comprehensive understanding of a medication's effects on DNAm is useful to then go on and disentangle associations which may underlie therapeutic effects.

In summary, **Chapter 5** identified robust DNAm associations with antidepressant exposure across different cohorts of European ancestry, indicating that antidepressant use may confound DNAm studies of psychiatric traits if not matched between groups or adjusted for. Additionally, these findings indicate that antidepressant response is associated with mitochondrial function, a factor hypothesised to be relevant to MDD. Furthermore, our findings can also inform on potential mechanisms of side effects of antidepressants. Finally, this work may provide a useful knowledge base for extended analyses on potential medication/dosage specific effects alongside work investigating DNAm and treatment response in the future.

6.3 Limitations & Future Directions

6.3.1 *Integrative analysis of multi-omics*

Although this thesis examines multiple omics data types (genomics, metabolomics, epigenomics and eQTL data), they have been segregated in their analysis and typically involve testing one omics type at a time. Whilst this analysis can give valuable insight into markers of disease and relevant biological pathways, without added context of other biological layers they may often reflect associated reactive effects rather than true causative factors^{254,278,577}. The integration of different omics data together can a more holistic view of potential disease-causing pathways which may transcend a single omics layer alone²⁵⁴. For instance, understanding how DNAm associations with antidepressant exposure identified in **Chapter 5** relate to transcriptomics, proteomics and metabolomics would inform more about downstream biological impact of treatment. However, the analysis of multiple omics simultaneously presents various statistical and methodological challenges. The incorporation of more omics data vastly increases the number of variables to test (the dimensionality) which can increase the multiple testing burden and also lead to overfitting of models to “noise” rather than meaningful signal^{578,579}. Various dimensionality reduction techniques can be used, such as PCA, however biological interpretation of these derived variables is not straightforward and limits the potential clinical applications⁵⁷⁹. Furthermore, this analysis requires multi-omics data to be measured and available in the same sample, which should be sufficiently large to account for the high dimensionality of multi-omics data^{254,579}.

Although the cost of omics technologies is reducing, this is still a significant barrier to conducting this type of integrative multi omics research at a large scale.

Another way to assess shared biological effects across multiple omics layers without relying on a single sample is through assessing the causal relationship between molQTLs and a trait using MR and colocalisation techniques¹⁹². This approach was implemented in **Chapter 3**, which tested the potential causal relationship between gene transcript levels in *FADS* (using eQTLs from GTEx) with MDD (using GWAS summary statistics). As this approach is from summary-level data, which can be derived from various samples, it is therefore more tractable to large analyses regarding multiple different omics layers. However, performing this analysis for various different molQTLs is cumbersome and computationally intensive, especially given the rapid increase in molQTLs which are being reported, such as variants which affect chromatin accessibility (caQTLs), histone modifications (hisQTLs) and alternative splicing (sQTLs)⁵⁸⁰. Recent methods now allow for the colocalisation of multiple molQTLs simultaneously, which may identify causal links across different omic layers⁵⁸⁰⁻⁵⁸³. However, many of these are limited by the number of molQTLs they can assess computationally⁵⁸¹⁻⁵⁸³ or are subject to assumptions of a single causal variant underlying all molQTLs in a locus⁵⁸². A recent methodology, Omics PIEiotRopic Association (OPERA), presents progress in both these areas⁵⁸⁰. OPERA can integrate up to 10 molQTLs to identify plausible molecular mechanisms which may underlie a GWAS locus. Specifically, OPERA uses a Bayesian generalisation approach on MR estimates of molQTLs on a trait to calculate the posterior probability of a pleiotropic association of multiple exposures (molQTLs) with said trait. This approach was shown to outperform other methods, whilst also being

much more efficient computationally⁵⁸⁰. In the future, OPERA could be implemented to understand more about the shared signals between omics layers relevant to both MDD and antidepressant exposure.

This thesis implemented association-wide approaches whereby each individual molecule (e.g. metabolite and CpGs) were tested individually. This reductionist approach may limit discovery in two ways. Firstly, metabolite measures investigated in **Chapter 3** consist of related metabolites which are highly correlated and do not represent independent variables. Therefore, multiple testing adjustment in this case is conservative and may increase type II errors^{254,579}. Perhaps more importantly, metabolites are increasingly being recognised as signalling molecules operating in dense networks²⁸⁵. Each individual metabolite may have a small effect on a trait, but modelled together may represent network perturbation which is significantly associated with the disease and may be more informative regarding relevant biological mechanisms. A more systems-level approach could involve the derivation of metabolite networks and systematic comparison of network characteristics between MDD cases and controls. For example, weighted correlation network analysis could identify highly correlated clusters of metabolites, which can then each be tested for an association with MDD^{584,585}. This may be especially useful in untargeted metabolomics whereby many metabolites of unknown function have been profiled. In the case of well characterised metabolic pathways, the known biochemistry can also be incorporated to aid interpretation of associations. The analyses in this thesis and other studies have indicated a dysregulation of *FADS* activity, however the pleiotropy makes it difficult to decipher exact mechanisms^{311,441}. The incorporation of known biochemical reactions

between related metabolites in the *FADS* pathway can be used to decipher pleiotropic mechanisms which may be relevant to MDD. Smith et al, (2022)⁴³⁸ illustrated how overall genetic correlations (r_g) between related metabolite pairs in a network can reflect shared upstream/downstream pathways, whereas specific variants which had discordant effects (to overall r_g) represent specific biological mechanisms between the two. Following this framework, comprehensive analysis of genetic correlations between metabolite pairs involved in fatty acid metabolism pathways could map associated variants to specific steps in the metabolic pathway. This would further help to interpret variants implicated in *FADS* metabolism which may be relevant to MDD.

Overall, my analyses consider single omics layers and individual molecules in isolation. Integrative multi-omics approaches may provide deeper insight into potential causal biological cascades relevant to MDD. Additionally, network-based analyses- whether derived from correlations within measured omics data or from established biochemical pathways- could offer a more holistic perspective of perturbed biological systems and aid in interpretation of observed associations.

6.3.2 *Longitudinal/prospective changes*

This thesis is also limited by its cross-sectional design, as all analyses examine associations between multi-omics, MDD and antidepressant exposure at a single time point. With the exception of genomics, multi-omics data is dynamic and can be influenced by a myriad of transient factors such as the season, time of day, dietary intake, infection and environmental stimuli at the point of measurement²⁵⁴. Therefore, measurement of multi-omics from a single timepoint provides only a snapshot of

biological variance, which can make it difficult to ascertain disease- or medication-relevant effects from background noise. Consistent findings across multiple time points would strengthen confidence that observed associations reflect trait-related processes. This is especially relevant to metabolomics analyses of MDD in **Chapter 3 & 4** as variability tends to increase further down the central dogma, due to accumulating biological regulation²⁷⁹. Studies have suggested that DNAm markers can serve as proxies for circulating protein levels (“EpiScores”) and can offer a more stable alternative to proteomic data, which is often susceptible to transient fluctuations⁵⁸⁶. Therefore, future studies could incorporate DNAm markers predictive of metabolomic levels to enhance robustness of associations and reduce the potential influence of short-term biological variance which is unrelated to MDD.

Additionally, data from a single time point are limited to assessing associations with already established conditions (the prevalence) rather than associations with the future development of a condition (the incidence). As such, the associational analysis of metabolites with lifetime MDD (**Chapter 3**) cannot disentangle whether observed associations are a cause or an effect of MDD. Although causal inference methods such as MR can be implemented to investigate the bidirectional relationship, as discussed these are subject to core assumptions which are often impossible to prove³⁷⁹. Future investigation of prospective multi-omics changes and the incidence of MDD (from longitudinal MDD and multi-omics follow-up) can explicitly identify associations which may be linked to the development of MDD. Also, as complex diseases such as MDD are usually developed over time, longitudinal data could also be used to assess biological trajectories leading up to a MDD diagnosis. However, MDD typically develops in

adolescence or early adulthood¹, well below the average age of participants in UKB at recruitment (~56 years). Therefore, even longitudinal data in this cohort may not be the most informative for investigating the development of MDD in the first instance.

Longitudinal multi-omics analysis in birth cohorts^{518,587} are more appropriate for modelling the development of first-episode MDD and can also be compared to older cohorts such as UKB to establish common or divergent MDD processes throughout the life course.

Linking multi-omics data to EHR can be valuable for assessing the longitudinal changes following baseline assessment. For example, in **Chapter 5**, I examined the correlation between DNAm at specific CpGs with antidepressant treatment duration, which was inferred from EHR data, specifically prescription patterns over time. However, without repeated multi-omics data, this analysis is limited to correlation between individuals and cannot distinguish general between-person variation in DNAm to within-person changes in DNAm which may be more relevant to treatment duration. Additionally, since the potential long-term effects of DNAm are yet to be established, it is important assess longitudinal changes of antidepressant exposure in individuals who were previously antidepressant naïve. Comprehensive evaluation of prior lifetime antidepressant exposure was not possible in my analysis as EHR linkage in GS does not precede 2009. Regardless, GS also predominantly consists of older adults (~ 47 years), and given the high prescribing rate of antidepressants, a significant portion are likely to have had prior exposure at one point in their lives, limiting the availability of truly antidepressant-naïve individuals. Again, longitudinal multi-omics analysis in birth cohorts

could be instrumental here in assessing DNAm effects of first-time antidepressant exposure.

In summary, this thesis uses cross-sectional multi-omics data which provides only a static snapshot of biological variation. Therefore, these analyses are unable to capture temporal changes relevant to development of MDD or long-term effects of antidepressant exposure on DNAm. Future integration of repeated multi-omic measurements, EHR linkage, and comparison with developmental cohorts is likely to provide a powerful tool in mapping biological trajectories underlying MDD onset, progression, and the impact antidepressants throughout the life course.

6.3.3 *More representative cohorts*

This thesis is limited as all analysis is conducted in cohorts of predominantly European ancestry, and therefore findings may not be generalisable to other ancestries. This reflects a historical Eurocentric bias in the field of genomics, where the majority of studies are conducted in individuals of European ancestry (~86% in 2021)^{178,588}. Whilst genomics studies in other populations are increasing, especially in those of Asian ancestries, various populations often from lower-middle income countries or indigenous or minority communities remain consistently underrepresented¹⁷⁸. Discoveries solely driven by European-dominated cohorts may lead to inequitable healthcare and widen clinical disparities between populations. For example, treatments derived based on genomics research on individuals of European ancestry may not be effective or be associated with specific side effects in other ancestries. Moreover, the incorporation of diverse ancestries, each with distinct properties and LD patterns, can lead to increased

power in genomics research and aid in both functional annotation and the identification of causal variants¹⁷⁸. As multi-omics is an emerging field expanding on genomics, it has typically been incorporated into established population cohorts and therefore largely suffers from the same Eurocentric bias. In addition, the majority of reference expression datasets (such as GTEx), which are increasingly incorporated in multi-omics studies also predominantly consist of those of European ancestry.

There is a concerted global effort to increase representation in genomics and multi-omics research, and many datasets are beginning to emerge which will enable the multi-omics analysis in multiple ancestries at scale⁴⁹⁴. For example, the Trans-Omics for Precision Medicine (TOPMed) programme consists of > 80 participating studies and currently ~ 155,000 participants, of which ~31% and ~15% are individuals of African ancestry and Asian ancestry⁴⁹⁴. This programme has collected WGS data for all participants, and is currently in the process of profiling a range of omics data, including transcriptomics (N~ 40,000), DNAm (N ~ 57,000), metabolomics (N ~ 37,000) and proteomics (N ~ 4,000)⁵⁸⁹. This will serve as a rich resource for the analysis of multi-omics in diverse ancestral groups and gives the opportunity to assess multiple omics in the same samples which can aid interpretation. In particular, this thesis finds a link between variants in *FADS* and MDD, which is linked to dysregulation of various fatty acid metabolism. The two *FADS* haplotypes, A(ncestral) and D(erived), have varied frequencies in different populations around the world (Figure 6.2)⁴⁷⁰. In European populations, the A haplotype is present in between 25-50% of individuals, whereas this haplotype has a very low frequency in African populations (~ 1%) and is very high in Native American populations (~97%). The D haplotype is associated with more

increased expression of *FADS1* and more efficient fatty acid synthesis from dietary precursors. Therefore, the effect *FADS1* and *FADS* dysregulation in MDD may be stronger in populations with a high frequency of the “D” haplotype⁴⁷⁰. Furthermore, the differences in LD and effect estimates across populations may be informative in fine-mapping variants and identifying potential mechanisms¹⁷⁸. Future analysis of the role of *FADS* in MDD across different populations is imperative to understand more about the mechanistic impact and perhaps to identify individuals who may be more susceptible to MDD following *FADS* dysregulation.

Finally, it is recognised that main cohorts used in this analysis (UKB and GS) suffer from “healthy volunteer bias”, whereby recruited participants are not characteristic of the background population, but are often biased towards white, middle-class, highly educated individuals who are typically more health-conscious^{345,590}. This is often attributed to increased barriers to participation facing underrepresented and typically more socially disadvantaged groups, such as location, literacy skills, and a mistrust of research⁵⁹¹. Therefore, it is a concern that research findings from these biobanks may not be truly generalisable and may act to perpetuate existing social inequalities. Various new cohorts are beginning to emerge which are explicitly designed in recruitment to minimise bias, such as AoU in the US²²⁷ and “Our Future Health” (OFH) in the UK⁵⁹². Since the latter launched in 2022, it has accrued nearly 2 million consenting volunteers and over 1.1 million donated blood samples from a nationally representative sample. Additionally, OFH includes linkage to health-related records including primary and secondary care, which would enable analysis of antidepressant exposure analogous to **Chapter 5** using prescription records. Projected to collect up to 5 million volunteers in

total, and with several planned future omics' assays, OFH is set to become an unprecedented resource for assessing multi-omics links to disease (including MDD) in representative samples.

Overall, the findings from this thesis are from population cohorts of European ancestry which are also not representative of the general population. Future analysis of *FADS* dysregulation in MDD and DNAm associations with antidepressant exposure in more inclusive cohorts are paramount for establishing the generalisability of these results across different populations and to gain power in discovery.

6.3.4 *Tissue specific analyses*

Another limitation to this thesis is the use of bulk whole-blood multi-omics data, which may not fully capture biologically relevant associations with MDD and antidepressant exposure. Aside from genomics, multi-omics signatures vary significantly across tissues and cell types due to inherent biological differentiation²⁵⁴. Consequently, associations detected in whole-blood may not reflect the same biological variation in other tissues. Given that MDD is primarily a brain-based disorder, multi-omics profiling of brain tissue may provide deeper insights into disease mechanisms and the effects of antidepressants. However, as direct access to living brain tissue is not feasible, most brain-specific multi-omics data are obtained from post-mortem brain banks, which have far smaller sample sizes than population cohorts⁵⁹³. Nonetheless, post-mortem brain banks offer the most direct opportunity to study multi-omics changes in brain tissue and could be used in the future to complement the associations in blood presented here. A notable example includes the Religious Orders Study/Memory and Aging Project

(ROSMAP)⁵⁹⁴, which includes longitudinal blood-based multi-omics and clinical data before death, as well as post-mortem brain tissue multi-omics data.

In **Chapters 3 & 4**, findings suggested a causal role of FADS and dysregulated fatty acid metabolism in MDD, alongside a potential rare MDD-driver variant in *FADS1*. The vast majority of fatty acid metabolism by *FADS* occurs in the liver and is transported in the blood across the blood-brain-barrier (BBB) to the brain⁵²³. Therefore, theoretically fatty acid levels in the blood may be a useful biomarker for fatty acid changes in the brain. However, the transport of fatty acids across the BBB is complex and often involves intermediate molecules which may be better proxies for brain metabolite levels than bulk plasma measurements. For instance, while DHA is enriched in the brain and is implicated in a range of processes relevant to MDD such as synaptic plasticity and neuroinflammation, it is transported across the BBB in the form of LPCs⁵³⁸. Therefore, circulating DHA-containing LPCs might be a more accurate indicator of brain DHA levels measuring all DHA-containing metabolites collectively. Increased resolution of metabolomic profiling, either using mass-spectrometry or 2-D NMR²⁹¹, would be beneficial for assessing the transport of DHA across the BBB and for estimating relevant effects in the brain from blood tissue. However, this relies on mechanistic knowledge of the transport across the BBB, which is still an active area of research for various fatty acids⁵⁹⁵. In addition, a small portion of fatty acid metabolism does occur in the brain *de novo*, meaning that some changes in brain lipid composition would not be captured through peripheral blood measurements⁵⁹⁵.

To date, there is no systematic reference dataset focused on the comparison of metabolomic profiles across different tissues, although various studies have compared

cross-tissue associations with specific diseases. Research on Alzheimer's Disease in ROSMAP found that blood and brain metabolomic profiles showed minimal overlap, suggesting that tissue-specific metabolic regulation in AD⁵⁹⁶. However, whether this holds true for MDD-associated metabolites- particularly fatty acids downstream of *FADS* activity- remains to be investigated. Another recent study performed GWAS on metabolomics data from the cerebrospinal fluid (CSF, N = 2,602) and post-mortem brain tissue (N = 1,016), to identify tissue-specific metabolite molQTLs (metQTLs)⁵⁹⁷. Their findings suggested strong tissue-specific genetic regulation of metabolites, with most novel associations in the CSF (~ 97%) and brain (~ 72%) belonging to metabolites which had previously been analysed using peripheral samples. Future integration of brain-specific metQTLs into causal inference methods (e.g. MR) could elucidate brain specific metabolic pathways which may be causal in MDD. A comprehensive reference dataset of metabolomic profiles across different tissues, analogous to GTEx²⁶⁷, would be instrumental in robustly identifying metQTLs and potential tissue-specific metabolomic signatures relevant to disease.

The same tissue-specific limitations also apply to the DNAm changes associated with antidepressant exposure presented in **Chapter 5**. Recent animal models showed that the antidepressant fluoxetine exerts cell- and region-specific effects on gene expression across the brain⁵⁵³. Therefore, blood based DNAm associations may not be capturing all relevant epigenetic modifications resulting from antidepressant exposure. Compared to metabolomics, more work has been done to systematically assess the correspondence between blood and brain DNAm. Several studies have directly compared blood with post-mortem brain DNAm profiles (N=12-80) and have found low overlap, with only 1.4-

9.7% of array CpGs having comparable DNAm⁵⁹⁸⁻⁶⁰⁰. Studies using brain samples from live donors undergoing neurosurgery (N=19-27) find higher correlations overall between blood and brain samples ($r \sim 0.86$), indicating that postmortem degradation may limit comparability^{601,602}. However, within-person variability was still high (~ 20% of CpGs were significantly correlated), and the levels of correlation between blood and brain were highly variable between specific genes and CpGs. Whilst informative, these studies are limited by small sample sizes and potential confounding due to the conditions that required the neurosurgery itself (e.g. intractable epilepsy). Nonetheless, future analyses could cross-reference the antidepressant-associated CpGs identified in **Chapter 5** with existing blood-brain DNAm studies to assess their relevance for brain-specific effects.

While this thesis provides valuable insights, its reliance on multi-omics from blood tissue may not capture biological changes relevant to MDD and antidepressant exposure in other tissues, particularly the brain. Future research could incorporate tissue-specific molQTLs to interrogate any causal links specific to certain tissues. Despite its limitations, blood-based multi-omics remains the most practical approach for clinical translation as it is minimally invasive, scalable and cost-effective. However, to fully harness this potential, a deeper understanding of blood and brain multi-omics and the relationship between them is essential. Comprehensive reference datasets comparing blood and brain multi-omics profiles would be an invaluable for identifying peripheral biomarkers relevant to MDD and antidepressant exposure effects in the brain. The GTEx project²⁶⁷ has begun this work for transcriptomic profiles in different tissues and

provides a useful blueprint for extending this analysis to other emerging multi-omics (e.g. metabolomics) in the future.

6.3.5 *Trans-diagnostic approaches*

Another limitation to this thesis is the binary categorization of lifetime MDD, which does not distinguish between symptomatic profiles, chronicity, age-of-onset or current symptom severity, all of which may have specific multi-omics associations. Recurrent MDD and early-onset MDD are associated with heightened treatment resistance and significant lower quality of life⁶⁰³. It would therefore be beneficial to understand biological differences which may underpin these different subtypes of MDD to inform potential preventative measures, treatment stratification and the development of novel therapeutics. Emerging evidence indicates that early and adult-onset of MDD may have distinct metabolomic signatures, with differential regulation of phosphatidylcholines, carnitines, biogenic amines and amino acids⁶⁰⁴. Future analysis could incorporate prior information regarding the number of MDD episodes or the age at first diagnosis to delineate whether the observed findings regarding *FADS* dysregulation are more/less prominent in these groups. Furthermore, lifetime MDD in this thesis does not differentiate between those suffering from MDD symptoms at the time of measurement and those who are now in remission from a previous episode. Given the transient nature of multi-omics, a comparison of symptom severity at the time of measurement may give a more accurate indicator of multi-omics changes proximal to MDD symptoms.

Therefore, future work could incorporate continuous symptom scales such as the Patient Health Questionnaire-9 (PHQ-9) to investigate the links between associations in this thesis and MDD severity. However, the use of sum-scores such as the PHQ-9

inherently assumes that all possible MDD symptoms are interchangeable and does not account for possible symptom-specific effects⁶⁰⁵.

Notably, the potential causal role of *FADS* dysregulation in lifetime MDD identified in **Chapter 3** has also been reported for other psychiatric disorders, such as BPD and SCZ^{307,606}. Therefore, *FADS* dysregulation may be a shared biological mechanism which may underline shared symptoms between these disorders. Future analysis of the association of fatty acids and *FADS* regulation on symptoms in a trans-diagnostic approach may be informative in understanding the impact of *FADS* regulation on different aspects of psychopathology. The heterogeneity in DSM defined MDD is also hypothesised to be a major contributor to the high levels of treatment resistance⁸⁰, as antidepressants may only have beneficial effects on certain biological subtypes of the disorder. It would therefore be interesting to assess whether DNAm associations with antidepressant exposure in **Chapter 5** are linked to changes in specific symptoms of MDD. Furthermore, future work could also assess whether DNAm associations with antidepressant exposure are heightened in those with distinct symptomatic profiles.

In summary, as this thesis evaluates the broad measure of lifetime MDD, it cannot disentangle any multi-omics effects which may be distinctive to symptom severity, chronicity or age of onset. Future research could integrate symptom-level information and episode history to assess whether associations identified here are generalizable across different subtypes of MDD. Furthermore, trans-diagnostic approaches can be implemented to evaluate whether associations identified here are linked to certain symptomatic profiles across psychiatric disorders.

6.3.6 *Technological advances*

Finally, this thesis is also limited by current technological platforms and assays available to measure the multi-omics data, specifically genomics, epigenomics and metabolomics. Each of these fields are at different stages of development and have different considerations.

Genomics is the oldest omics field and as such has seen the most technological advancement. Indeed, the reduced cost and speed of NGS approaches have enabled the analysis of whole-genomes at scale in population cohorts such as UKB^{222,226}. This data was utilised in **Chapter 4** to complement findings from GWAS (**Chapter 3**). However, these technologies also have limitations²²⁰. Specifically, the WGS data was generated using short-read sequencing, which sequences DNA by assembling short segments (typically < 300 bp) to the reference genome. However, short reads cannot capture large genomic structural variation (> 300bp), such as copy number variants, inversions, segmental duplications, tandem repeats or regions with a high guaninecytosine content⁶⁰⁷. Structural variants typically have larger effect, and these regions typically exhibit higher mutation rates⁶⁰⁷. Therefore, variable genomic regions of potentially large-effect are understudied using short-read technologies²²². However, long-read sequencing (LRS) techniques such as those from Pacific Biosciences (PacBio) or Oxford Nanopore Technologies (ONT), can generate continuous DNA sequences spanning kilobases or even several megabases (Mb) and consequently detect large structural variants⁶⁰⁷. The impact of structural variation in *FADS* or isoform-specific effects on downstream metabolites and MDD may give further insight into pathological mechanisms. However, LRS is currently limited in its application in large-

scale population cohorts due to high cost of these technologies compared to short-read NGS⁶⁰⁷. Perhaps the most immediate benefit of LRS to population genomics research is the derivation of the Human Pangenome⁶⁰⁸. This refers to a reference genome generated constructed from 47 phased diploid assemblies of a genetically diverse cohort. Compared to the commonly use GRCh38 reference, it adds 119 million base pairs (~ 90 million from structural variations) and 1,115 gene duplications⁶⁰⁸. Reference genomes are the backbone of genomic analysis, as they are essential for imputation from genotyping arrays in GWAS and variant calling in NGS^{217,608}. Increased diversity and resolution from LRS in the pangenome will boost discovery, power and generalisability of genomic analyses. Implementation into GRCh38-based workflows have shown the pangenome to reduce small-variant discovery errors and increase the number of structural variants detected using short-read NGS data⁶⁰⁸. Therefore, future incorporation of the pangenome in genomic analyses of MDD will boost discovery.

Epigenomics studies modifiable non-nucleotide changes to DNA which can affect gene expression, the most studied of which is DNAm²⁷⁷. The findings from **Chapter 5** are based on the measurement of DNAm using Illumina arrays which have been specifically designed to capture DNAm CpGs predominantly at gene promoter regions. Nonetheless, DNAm arrays only capture a small proportion of potential DNAm sites (~ 2%). Each iteration of Illumina arrays has improved on coverage of regulatory regions; however the EPIC array used in this analysis (~850K CpG sites) only covers ~27% of proximal regulatory elements and ~7% of distal regulatory elements³³¹. Therefore, these data may not capture potentially relevant DNAm changes associated with antidepressant exposure, especially those in regulatory regions. Whole-genome bisulphite sequencing

(WGBS) can detect a higher density of CpGs and can also measure DNAm at non-CpG sites, providing increased breadth of DNAm measurement. However, even with decreasing costs, WGBS remains prohibitively expensive for applications in large-scale population studies³³¹. As well, both these methods require pretreatment of DNA with sodium bisulphite, which can result in DNA fragmentation which can impact downstream analysis⁶⁰⁹. However, both PacBio and ONT sequencers have been shown to be able to detect DNAm in long-reads of untreated DNA^{331,609}. Although these are currently techniques are less accurate than DNAm measurements from micro-arrays, they present exciting opportunities for the simultaneous detection of DNAm and DNA sequencing in long stretches of untreated DNA in the future. Future research using sequencing technologies such as WGBS and LRS can assess the association of antidepressant exposure more broadly across the methylome. Additionally, DNAm represents only one epigenetic mechanism, and other factors such as chromatin accessibility, histone acetylation and long-range chromosomal interactions can also influence gene expression. ChIP-seq allows for the detection of various histone modifications and chromatin states²⁷⁷ and could be implemented to assess the association of antidepressant exposure on the epigenomic landscape beyond DNAm alone.

Among the various multi-omics fields, metabolomics is one of the most recent to emerge, and it remains an area of rapid technological development²⁷⁹. Unlike genomics or epigenomics, which have more established methodologies and extensive reference datasets, metabolomics is still evolving in terms of standardisation, sensitivity and coverage. Different mainstream commercial platforms that measure metabolite levels have variable coverage, use different analytical pipelines and metabolomic profiling

techniques. For instance, Metabolon utilises untargeted metabolomic analysis using both LC and GC coupled with MS and can identify thousands of metabolites with high precision⁶¹⁰. In contrast, targeted methodologies utilised by Biocrates and Nightingale screen up to ~250 and ~650 preselected metabolites respectively⁶¹¹. Specifically, Biocrates measures metabolites using LC-MS methods, whereas Nightingale utilises NMR spectroscopy. Various studies have shown that generally, there is high correlation between levels overlapping metabolites on each platform⁶¹¹. However, as each platform is variable in coverage, direct comparison is not possible for most measured analytes, limiting the ability to assess replication and perform meaningful meta analyses³⁰⁵. Furthermore, changes to analysis pipelines in specific platforms could also affect inter-platform comparability and the comparability with previously assayed measurements. A recent technical report found that a re-calibration in Nightingale platform update significant affected the accuracy of previously curated metabolite scores predicting age (“MetaboAge”), which had to be re-trained using the updated measurements⁶¹². As technologies progress and become more established, improved standardisation across platforms and pipeline versions is required to enable more direct comparison of associated metabolites with MDD. To this end, a meta-analysis of the role of metabolites in MDD across multiple cohorts (using different profiling platforms) would be beneficial to assess the generalisability of *FADS* dysregulation in MDD.

Furthermore, the full extent of the human metabolome is yet to be realised⁶¹³. The Human Metabolome Database (HMDB) aims to provide a resource detailing all identified and quantified metabolites in the human body. Since its launch in 2007, the number of annotated metabolites on the HMDB has increased from 2,180 to

217,920⁶¹³. Additionally, the HMDB has over 1 million detected metabolites which are unannotated from GC-MS and of unknown function⁶¹³. Therefore, there may be many metabolite species relevant to MDD which are not captured in the 249 metabolites assayed on the Nightingale platform. As the HMDB and our knowledge of different metabolite species expands, the insights gained from metabolomic studies will continue to grow. Future research may benefit from the increasing knowledge of metabolite species in the body and can build on these findings to further elucidate their potential roles in MDD.

Overall, this thesis is constrained by current genomics, epigenomics and metabolomics technologies, each at different stages of development. Advances in NGS have improved genomic analysis, but limitations in detecting structural variation persist, with the LRS and the pangenome promising future improvements. Epigenomics studies in population cohorts is largely dominated by DNAm studies using microarrays which only cover a subset of possible DNAm sites. Reduced cost of sequencing technologies and LRS can increase the breadth of DNAm measurement. Furthermore, increased incorporation of other epigenetic markers in population cohorts such as histone modifications can increase discovery of associated epigenetic regulation with antidepressant exposure. Metabolomics faces challenges in the standardisation and platform comparability, hindering cross-study replication. Increased overlap between platforms or a comprehensive linking strategy between them is required to increase the potential of metabolomic studies. Furthermore, increased characterisation and understanding of the human metabolome may aid in the identification of metabolites linked to MDD which

were hitherto unknown. Ongoing technological progress will strengthen multi-omics research into MDD and antidepressant effects.

7 Responsible Research and Innovation

This research aims to understand more about biological systems in MDD pathology and antidepressant exposure. The potential to then use this knowledge to identify novel therapeutic targets and aid clinical patient stratification is the main motivation for this work. This outcome would be a positive change for many who do not respond to treatment or suffer from intolerant side effects from current antidepressant medications. However, there are several ethical considerations regarding the clinical translation of multi-omics findings to the treatment of MDD.

One, which has been discussed in depth (Section 6.3.3), is the disproportionate amount of research performed on biobanks developed in western countries with a Eurocentric focus¹⁷⁸, as well as these datasets displaying “healthy volunteer bias”. Therefore, clinical translation based off these findings alone could be biased towards healthier, wealthier individuals of European ancestry, and widen healthcare disparities. This is an issue pertinent to the genomics and multi-omics field as a whole and needs to be addressed to ensure equitable benefit from multi-omics discoveries. In this research, I have capitalised on recent large scale multi-omics data, which are in cohorts of predominantly European ancestry. It is paramount to consider this in the interpretation of these findings. This is especially relevant for findings in **Chapter 3 & 4** which focus in on a *FADS*, a complex genomic region which varies significantly between different ancestries⁴⁷⁰. As multi-omics data becomes more prolific, and datasets become more diverse, future analysis investigating the MDD and *FADS* in non-European ancestries will complement these studies and allow for a more holistic interpretation.

Another ethical consideration for this thesis is the use of prescription-linkage data to determine antidepressant exposure. Data linkage can be hugely challenging to perform, especially when combining data from multiple different healthcare providers⁶¹⁴. In countries which do not have a national health service or appropriate infrastructure, this could also increase disparities seen in health research. Although linkage of multi-omics to health records gives rich resource of epidemiological information, the safety, management and appropriate use of the data is a major concern. A meta-analysis of studies focused on public opinion of the sharing of health data for research found that although the public were aware of the many benefits, there remained significant anxiety around potential security breaches and abuse of their sensitive data⁶¹⁵. Health data is particularly sensitive and breaches could result in identity theft, fraudulent activity, increased stigma and potential discrimination⁶¹⁶. Furthermore, breaches could reduce medical safety, if EHRs are missing or altered in systems used by healthcare providers⁶¹⁶. The continued use of encryption, de-identification and storage of data in specialised data havens is critical to avoid data breaches and maintain patient trust⁶¹⁶.

The motivation behind this research is based off the growing prevalence of MDD, resistance to treatments and unknown therapeutic mechanisms of antidepressants, with the goal of improving quality of life for those who suffer from MDD. However, this research did not directly involve members of the public or those with lived experience of MDD and/or antidepressant use to shape these research aims. Patient and Public Involvement and Engagement (PPIE) in research has been growing over the last decade, with PPIE groups being co-collaborators and taking an active role in shaping research questions and analyses⁶¹⁷. Future involvement of PPIE in MDD and

antidepressant research will ensure that analyses are aligned with patient priorities, and contribute to ultimately providing effective and equitable outcomes.

8 Conclusion

Major Depressive Disorder (MDD) is common, frequently disabling and remains a global health priority. Antidepressant treatments were first identified based on serendipitous observations in the 1950s. Many treatment episodes are sub-optimal and can be accompanied by intolerant side effects. Lack of mechanistic knowledge regarding the biological basis of MDD hinders the development of novel therapeutic agents, and the potential targeting of treatments to those most likely to respond. This thesis expands knowledge on metabolomic links to MDD and the association of antidepressant exposure with DNAm.

Multi-omics research provides an unprecedented opportunity to investigate the biological basis of complex diseases such as MDD, alongside the evaluation of environmental factors such as antidepressant medications. This thesis expands knowledge on metabolomic links to MDD and the association of antidepressant exposure with DNAm. Specifically, this work capitalises on recent large-scale metabolomic data and WGS in the UKB to elucidate metabolomic signatures in MDD. Specifically, heightened levels of circulating omega-3 fatty acids and the omega-6: omega-3 ratio were found to have a potentially causal protective and risk-inducing effect respectively. However, these findings are hinged upon genetic variants in the *FADS* cluster, a conserved region which is being increasingly shown to be highly pleiotropic. Therefore, this thesis finds evidence of a causal role of dysregulated *FADS* metabolism in MDD but cannot empirically identify the metabolite specie(s) which may be specifically affecting MDD. Rare-variant analysis found evidence of disparate signals in

FADS for MDD and circulating omega-3 related metabolites, indicating that these circulating measures may not be primary causal mechanism of *FADS* dysregulation on MDD. Additionally, the rare-variant analysis identified a rare missense variant in *FADS1* which was significantly associated with MDD. Overall, these findings contribute to a growing body of evidence of *FADS* dysregulation in MDD and have identified a potential causal variant for MDD in *FADS1*. If replicated, understanding the downstream impact of this variant on *FADS* enzymatic activity and downstream metabolite pathways may pinpoint processes relevant to MDD.

Additionally, this thesis finds evidence that antidepressant exposure is robustly associated with hypermethylation in 8 different CpG sites, which show links to neurite outgrowth (*KANK1*), immune regulation (*DOT1L*, *DGUOK-AS1*), mitochondrial function (*DGUOK-AS1*) and male infertility (*CCDC146*). This research demonstrates how EHRs can be incorporated to gain reliable measures of antidepressant exposure, which may aid future research. Furthermore, this provides a useful knowledge base to focus future analyses on antidepressant response and for delineating DNAm associations which may be confounded by antidepressant use in future MWAS of psychiatric traits.

There are several limitations to this work, most of which pair with exciting future research directions. Notably, the use of more representative cohorts with longitudinal data can give more generalisable estimates of multi-omics associations beyond a singular time point. In particular, the use of developmental cohorts may enable detection of the development of first-episode MDD alongside evaluating the effects of antidepressants in antidepressant-naïve individuals. The analysis of tissue-specific multi-omics data or molQTLs in the future can also help to establish whether

associations with circulating markers are representative of those in the brain, which may be more relevant to MDD and antidepressant exposure. Additionally, trans-diagnostic approaches may enable detection of multi-omics signatures at a symptom level, aiding in potential future stratification. Ongoing technological advances will also provide further insight into both MDD and antidepressant exposure in the future. For example, LRS in population cohorts may enable analysis of structural variation with MDD, and more broadly the release of the pangenome will boost discovery for all genomic analyses. Emerging integrative multi-omics analyses will also be transformative in the future as many cohorts collect more layers of omics data.

In summary, this thesis leverages multi-omics data to assess the metabolomic signature of MDD, whilst also providing novel insight into the epigenetic impact of antidepressant exposure. The findings highlight the potential causal role of FADS metabolism in MDD and reveal a potentially causal rare missense variant in *FADS1*. Also, this work shows robust associations of antidepressant exposure on DNAm which indicate potential mechanisms underlying both therapeutic and side effects. Future directions include the incorporation of more representative and longitudinal cohorts, tissue-specific analyses and the use of emerging technologies to assess additional omics information. Ultimately, this research contributes to a growing body of knowledge regarding the biological impact of MDD and antidepressant treatment which will support potential precision medicine approaches for MDD in the future.

9 References

1. Solmi M, Radua J, Olivola M, et al. Age at onset of mental disorders worldwide: large-scale meta-analysis of 192 epidemiological studies. *Mol Psychiatry*. 2022;27(1):281-295. doi:10.1038/s41380-021-01161-7
2. Grigoriadis S, Erlick Robinson G. Gender Issues in Depression. *Annals of Clinical Psychiatry*. 2007;19(4):247-255. doi:10.3109/10401230701653294
3. ter Meulen WG, Draisma S, van Hemert AM, et al. Depressive and anxiety disorders in concert—A synthesis of findings on comorbidity in the NESDA study. *Journal of Affective Disorders*. 2021;284:85-97. doi:10.1016/j.jad.2021.02.004
4. Grant BF, Stinson FS, Dawson DA, et al. Prevalence and Co-occurrence of Substance Use Disorders and Independent Mood and Anxiety Disorders: Results From the National Epidemiologic Survey on Alcohol and Related Conditions. *Archives of General Psychiatry*. 2004;61(8):807-816. doi:10.1001/archpsyc.61.8.807
5. Wimberley T, Horsdal HT, Brikell I, et al. Temporally ordered associations between type 2 diabetes and brain disorders – a Danish register-based cohort study. *BMC Psychiatry*. 2022;22(1):573. doi:10.1186/s12888-022-04163-z
6. Possidente C, Fanelli G, Serretti A, Fabbri C. Clinical insights into the cross-link between mood disorders and type 2 diabetes: A review of longitudinal studies and Mendelian randomisation analyses. *Neuroscience & Biobehavioral Reviews*. 2023;152:105298. doi:10.1016/j.neubiorev.2023.105298
7. Mayer MA, Gutierrez-Sacristan A, Leis A, et al. Using Electronic Health Records to Assess Depression and Cancer Comorbidities. In: *Informatics for Health: Connected Citizen-Led Wellness and Population Health*. IOS Press; 2017:236-240. doi:10.3233/978-1-61499-753-5-236
8. Huang JW, Cao CA, Zheng WH, et al. The mechanism of cancer-depression comorbidity. *Neuroscience*. 2024;556:25-30. doi:10.1016/j.neuroscience.2024.07.040
9. Marrie RA, Bernstein CN. Psychiatric comorbidity in immune-mediated inflammatory diseases. *World Psychiatry*. 2021;20(2):298-299. doi:10.1002/wps.20873
10. Nemeroff CB, Goldschmidt-Clermont PJ. Heartache and heartbreak—the link between depression and cardiovascular disease. *Nat Rev Cardiol*. 2012;9(9):526-539. doi:10.1038/nrcardio.2012.91

11. Clayborne ZM, Varin M, Colman I. Systematic Review and Meta-Analysis: Adolescent Depression and Long-Term Psychosocial Outcomes. *Journal of the American Academy of Child & Adolescent Psychiatry*. 2019;58(1):72-79. doi:10.1016/j.jaac.2018.07.896
12. Campbell D, Green MJ, Davies N, et al. Effects of depression on employment and social outcomes: a Mendelian randomisation study. *J Epidemiol Community Health*. 2022;76(6):563-571. doi:10.1136/jech-2021-218074
13. Cai W, Mueller C, Li YJ, Shen WD, Stewart R. Post stroke depression and risk of stroke recurrence and mortality: A systematic review and meta-analysis. *Ageing Research Reviews*. 2019;50:102-109. doi:10.1016/j.arr.2019.01.013
14. Lépine JP, Briley M. The increasing burden of depression. *Neuropsychiatric Disease and Treatment*. 2011;7(sup1):3-7. doi:10.2147/NDT.S19617
15. Moffitt TE, Caspi A, Taylor A, et al. How common are common mental disorders? Evidence that lifetime prevalence rates are doubled by prospective versus retrospective ascertainment. *Psychological Medicine*. 2010;40(6):899-909. doi:10.1017/S0033291709991036
16. Ferrari AJ, Santomauro DF, Aali A, et al. Global incidence, prevalence, years lived with disability (YLDs), disability-adjusted life-years (DALYs), and healthy life expectancy (HALE) for 371 diseases and injuries in 204 countries and territories and 811 subnational locations, 1990–2021: a systematic analysis for the Global Burden of Disease Study 2021. *The Lancet*. 2024;403(10440):2133-2161. doi:10.1016/S0140-6736(24)00757-8
17. Vos T, Barber RM, Bell B, et al. Global, regional, and national incidence, prevalence, and years lived with disability for 301 acute and chronic diseases and injuries in 188 countries, 1990–2013: a systematic analysis for the Global Burden of Disease Study 2013. *The Lancet*. 2015;386(9995):743-800. doi:10.1016/S01406736(15)60692-4
18. James SL, Abate D, Abate KH, et al. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. *The Lancet*. 2018;392(10159):1789-1858. doi:10.1016/S0140-6736(18)32279-7
19. *World Mental Health Report: Transforming Mental Health for All*. World Health Organization; 2022.
20. Greenberg P, Chitnis A, Louie D, et al. The Economic Burden of Adults with Major Depressive Disorder in the United States (2019). *Adv Ther*. 2023;40(10):4460-4479. doi:10.1007/s12325-023-02622-x

21. Herrman H, Kieling C, McGorry P, Horton R, Sargent J, Patel V. Reducing the global burden of depression: a Lancet–World Psychiatric Association Commission. *The Lancet*. 2019;393(10189):e42–e43. doi:10.1016/S0140-6736(18)32408-5
22. Cui L, Li S, Wang S, et al. Major depressive disorder: hypothesis, mechanism, prevention and treatment. *Sig Transduct Target Ther*. 2024;9(1):1–32. doi:10.1038/s41392-024-01738-y
23. Schildkraut JJ. The catecholamine hypothesis of affective disorders: a review of supporting evidence. *AJP*. 1965;122(5):509–522. doi:10.1176/ajp.122.5.509
24. Stahl SM. Blue Genes and the Monoamine Hypothesis of Depression: (Brainstorms). *J Clin Psychiatry*. 2000;61(2):77–78. doi:10.4088/JCP.v61n0201
25. Robinson E. Psychopharmacology: From serendipitous discoveries to rationale design, but what next? *Brain and Neuroscience Advances*. 2018;2:2398212818812629. doi:10.1177/2398212818812629
26. Hirschfeld RMA. History and Evolution of the Monoamine Hypothesis of Depression. *J Clin Psychiatry*. 2000;(61):Suppl 6, 4–6.
27. Berton O, Nestler EJ. New approaches to antidepressant drug discovery: beyond monoamines. *Nat Rev Neurosci*. 2006;7(2):137–151. doi:10.1038/nrn1846
28. Perez-Caballero L, Torres-Sanchez S, Romero-López-Alberca C, González-Saiz F, Mico JA, Berrocoso E. Monoaminergic system and depression. *Cell Tissue Res*. 2019;377(1):107–113. doi:10.1007/s00441-018-2978-8
29. Moncrieff J, Cooper RE, Stockmann T, Amendola S, Hengartner MP, Horowitz MA. The serotonin theory of depression: a systematic umbrella review of the evidence. *Mol Psychiatry*. 2023;28(8):3243–3256. doi:10.1038/s41380-022-01661-0
30. Jauhar S, Arnone D, Baldwin DS, et al. A leaky umbrella has little value: evidence clearly indicates the serotonin system is implicated in depression. *Mol Psychiatry*. 2023;28(8):3149–3152. doi:10.1038/s41380-023-02095-y
31. Bartova L, Lanzenberger R, Rujescu D, Kasper S. Reply to: “The serotonin theory of depression: a systematic umbrella review of the evidence” published by Moncrieff J, Cooper RE, Stockmann T, Amendola S, Hengartner MP, Horowitz MA in *Molecular Psychiatry* (2022 Jul 20. doi: 10.1038/s41380-022-01661-0). *Mol Psychiatry*. 2023;28(8):3153–3154. doi:10.1038/s41380-023-02093-0
32. Almulla AF, Maes M. Although serotonin is not a major player in depression, its precursor is. *Mol Psychiatry*. 2023;28(8):3155–3156. doi:10.1038/s41380-02302092-1

33. Shyn SI, Hamilton SP. The genetics of major depression: Moving beyond the monoamine hypothesis. *Psychiatr Clin North Am.* 2010;33(1):125-140. doi:10.1016/j.psc.2009.10.004
34. Bullmore E. Inflamed depression. *The Lancet.* 2018;392(10154):1189-1190. doi:10.1016/S0140-6736(18)32356-0
35. Troubat R, Barone P, Leman S, et al. Neuroinflammation and depression: A review. *European Journal of Neuroscience.* 2021;53(1):151-171. doi:10.1111/ejn.14720
36. Dantzer R. Cytokine-induced sickness behaviour: a neuroimmune response to activation of innate immunity. *European Journal of Pharmacology.* 2004;500(1):399-411. doi:10.1016/j.ejphar.2004.07.040
37. Raison CL, Capuron L, Miller AH. Cytokines sing the blues: inflammation and the pathogenesis of depression. *Trends Immunol.* 2006;27(1):24-31. doi:10.1016/j.it.2005.11.006
38. Patten SB, Beck CA, Williams JVA, Barbu C, Metz LM. Major depression in multiple sclerosis. *Neurology.* 2003;61(11):1524-1527. doi:10.1212/01.WNL.0000095964.34294.B4
39. Kurina LM, Goldacre MJ, Yeates D, Gill LE. Depression and anxiety in people with inflammatory bowel disease. *Journal of Epidemiology & Community Health.* 2001;55(10):716-720. doi:10.1136/jech.55.10.716
40. Haapakoski R, Mathieu J, Ebmeier KP, Alenius H, Kivimäki M. Cumulative metaanalysis of interleukins 6 and 1 β , tumour necrosis factor α and C-reactive protein in patients with major depressive disorder. *Brain, Behavior, and Immunity.* 2015;49:206-215. doi:10.1016/j.bbi.2015.06.001
41. O'Brien SM, Scott LV, Dinan TG. Antidepressant therapy and C-reactive protein levels. *The British Journal of Psychiatry.* 2006;188(5):449-452. doi:10.1192/bjp.bp.105.011015
42. O'Brien SM, Scully P, Fitzgerald P, Scott LV, Dinan TG. Plasma cytokine profiles in depressed patients who fail to respond to selective serotonin reuptake inhibitor therapy. *Journal of Psychiatric Research.* 2007;41(3):326-331. doi:10.1016/j.jpsychires.2006.05.013
43. Khandaker GM, Pearson RM, Zammit S, Lewis G, Jones PB. Association of Serum Interleukin 6 and C-Reactive Protein in Childhood With Depression and Psychosis in Young Adult Life: A Population-Based Longitudinal Study. *JAMA Psychiatry.* 2014;71(10):1121-1128. doi:10.1001/jamapsychiatry.2014.1332

44. Gimeno D, Kivimäki M, Brunner EJ, et al. Associations of C-reactive protein and interleukin-6 with cognitive symptoms of depression: 12-year follow-up of the Whitehall II study. *Psychological Medicine*. 2009;39(3):413-423. doi:10.1017/S0033291708003723
45. MIYAOKA H, OTSUBO T, KAMIJIMA K, ISHII M, ONUKI M, MITAMURA K. Depression From Interferon Therapy in Patients With Hepatitis C. *AJP*. 1999;156(7):1120-1120. doi:10.1176/ajp.156.7.1120
46. Capuron L, Gummnick JF, Musselman DL, et al. Neurobehavioral Effects of Interferon- α in Cancer Patients: Phenomenology and Paroxetine Responsiveness of Symptom Dimensions. *Neuropsychopharmacology*. 2002;26(5):643-652. doi:10.1016/S0893-133X(01)00407-9
47. Udina M, Castellví, P, Moreno-España J, et al. Interferon-Induced Depression in Chronic Hepatitis C: A Systematic Review and Meta-Analysis. *The Journal of Clinical Psychiatry*. 2012;73(8):9260.
48. Kappelmann N, Lewis G, Dantzer R, Jones PB, Khandaker GM. Antidepressant activity of anti-cytokine treatment: a systematic review and meta-analysis of clinical trials of chronic inflammatory conditions. *Mol Psychiatry*. 2018;23(2):335-343. doi:10.1038/mp.2016.167
49. Raison CL, Rutherford RE, Woolwine BJ, et al. A Randomized Controlled Trial of the Tumor Necrosis Factor Antagonist Infliximab for Treatment-Resistant Depression: The Role of Baseline Inflammatory Biomarkers. *JAMA Psychiatry*. 2013;70(1):31-41. doi:10.1001/2013.jamapsychiatry.4
50. Akhondzadeh S, Jafari S, Raisi F, et al. Clinical trial of adjunctive celecoxib treatment in patients with major depression: a double blind and placebo controlled trial. *Depression and Anxiety*. 2009;26(7):607-611. doi:10.1002/da.20589
51. Abbasi SH, Hosseini F, Modabbernia A, Ashrafi M, Akhondzadeh S. Effect of celecoxib add-on treatment on symptoms and serum IL-6 concentrations in patients with major depressive disorder: Randomized double-blind placebocontrolled study. *Journal of Affective Disorders*. 2012;141(2):308-314. doi:10.1016/j.jad.2012.03.033
52. Berk M, Dean OM, Cotton SM, et al. The efficacy of adjunctive N-acetylcysteine in major depressive disorder: a double-blind, randomized, placebo-controlled trial. *The Journal of clinical psychiatry*. 2014;75(6):5936.
53. Husain MI, Strawbridge R, Stokes PR, Young AH. Anti-inflammatory treatments for mood disorders: Systematic review and meta-analysis. <https://doi.org/101177/0269881117725711>. Published online August 31, 2017. doi:10.1177/0269881117725711

54. Duman RS, Monteggia LM. A Neurotrophic Model for Stress-Related Mood Disorders. *Biological Psychiatry*. 2006;59(12):1116-1127. doi:10.1016/j.biopsych.2006.02.013
55. Drevets WC, Price JL, Furey ML. Brain structural and functional abnormalities in mood disorders: implications for neurocircuitry models of depression. *Brain Struct Funct*. 2008;213(1):93-118. doi:10.1007/s00429-008-0189-x
56. Miguel-Hidalgo JJ, Rajkowska G. Morphological Brain Changes in Depression. *Mol Diag Ther*. 2002;16(6):361-372. doi:10.2165/00023210-200216060-00001
57. Skaper SD. Neurotrophic Factors: An Overview. In: Skaper SD, ed. *Neurotrophic Factors: Methods and Protocols*. Springer; 2018:1-17. doi:10.1007/978-1-49397571-6_1
58. Casarotto P, Umemori J, Castrén E. BDNF receptor TrkB as the mediator of the antidepressant drug action. *Front Mol Neurosci*. 2022;15:1032224. doi:10.3389/fnmol.2022.1032224
59. Castrén E, Rantamäki T. The role of BDNF and its receptors in depression and antidepressant drug action: Reactivation of developmental plasticity. *Developmental Neurobiology*. 2010;70(5):289-297. doi:10.1002/dneu.20758
60. Lopez JP, Mamdani F, Labonte B, et al. Epigenetic regulation of BDNF expression according to antidepressant response. *Mol Psychiatry*. 2013;18(4):398-399. doi:10.1038/mp.2012.38
61. Adachi M, Barrot M, Autry AE, Theobald D, Monteggia LM. Selective Loss of BrainDerived Neurotrophic Factor in the Dentate Gyrus Attenuates Antidepressant Efficacy. *Biological Psychiatry*. 2008;63(7):642-649. doi:10.1016/j.biopsych.2007.09.019
62. Gatt JM, Nemeroff CB, Dobson-Stone C, et al. Interactions between BDNF Val66Met polymorphism and early life stress predict brain and arousal pathways to syndromal depression and anxiety. *Mol Psychiatry*. 2009;14(7):681-695. doi:10.1038/mp.2008.143
63. Kowiański P, Lietzau G, Czuba E, Waśkow M, Steliga A, Moryś J. BDNF: A Key Factor with Multipotent Impact on Brain Signaling and Synaptic Plasticity. *Cell Mol Neurobiol*. 2018;38(3):579-593. doi:10.1007/s10571-017-0510-4
64. Duman RS, Li N. A neurotrophic hypothesis of depression: role of synaptogenesis in the actions of NMDA receptor antagonists. *Philosophical Transactions of the Royal Society B: Biological Sciences*. 2012;367(1601):2475-2484. doi:10.1098/rstb.2011.0357

65. Rolls ET, Cheng W, Gilson M, et al. Effective Connectivity in Depression. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*. 2018;3(2):187-197. doi:10.1016/j.bpsc.2017.10.004
66. Miljevic A, Bailey NW, Murphy OW, Perera MPN, Fitzgerald PB. Alterations in EEG functional connectivity in individuals with depression: A systematic review. *Journal of Affective Disorders*. 2023;328:287-302. doi:10.1016/j.jad.2023.01.126
67. Kessels HW, Malinow R. Synaptic AMPA Receptor Plasticity and Behavior. *Neuron*. 2009;61(3):340-350. doi:10.1016/j.neuron.2009.01.015
68. Citri A, Malenka RC. Synaptic Plasticity: Multiple Forms, Functions, and Mechanisms. *Neuropsychopharmacol*. 2008;33(1):18-41. doi:10.1038/sj.npp.1301559
69. Akhondzadeh S. Hippocampal synaptic plasticity and cognition. *Journal of Clinical Pharmacy and Therapeutics*. 1999;24(4):241-248. doi:10.1046/j.13652710.1999.00231.x
70. Noworyta K, Cieslik A, Rygula R. Neuromolecular Underpinnings of Negative Cognitive Bias in Depression. *Cells*. 2021;10(11):3157. doi:10.3390/cells10113157
71. Li N, Liu RJ, Dwyer JM, et al. Glutamate *N*-methyl-D-aspartate Receptor Antagonists Rapidly Reverse Behavioral and Synaptic Deficits Caused by Chronic Stress Exposure. *Biological Psychiatry*. 2011;69(8):754-761. doi:10.1016/j.biopsych.2010.12.015
72. Berman RM, Cappiello A, Anand A, et al. Antidepressant effects of ketamine in depressed patients. *Biological Psychiatry*. 2000;47(4):351-354. doi:10.1016/S00063223(99)00230-9
73. Li N, Lee B, Liu RJ, et al. mTOR-Dependent Synapse Formation Underlies the Rapid Antidepressant Effects of NMDA Antagonists. *Science*. 2010;329(5994):959-964. doi:10.1126/science.1190287
74. Hoeffer CA, Klann E. mTOR signaling: At the crossroads of plasticity, memory and disease. *Trends in Neurosciences*. 2010;33(2):67-75. doi:10.1016/j.tins.2009.11.003
75. Goldberg D. The heterogeneity of “major depression.” *World Psychiatry*. 2011;10(3):226-228.
76. Association AP. Diagnostic and Statistical Manual of Mental Disorders. *Text Revision*. Published online 2000. Accessed January 31, 2025. <https://cir.nii.ac.jp/crid/1573950399819987840>

77. Østergaard SD, Jensen SOW, Bech P. The heterogeneity of the depressive syndrome: When numbers get serious. *Acta Psychiatrica Scandinavica*. 2011;124(6):495-496. doi:10.1111/j.1600-0447.2011.01744.x
78. Łojko D, Rybakowski JK. Atypical depression: current perspectives. *Neuropsychiatric Disease and Treatment*. 2017;13:2447-2456. doi:10.2147/NDT.S147317
79. Milaneschi Y, Lamers F, Peyrot WJ, et al. Polygenic dissection of major depression clinical heterogeneity. *Mol Psychiatry*. 2016;21(4):516-522. doi:10.1038/mp.2015.86
80. Byrne EM, Carrillo-Roa T, Penninx BWJH, et al. Applying polygenic risk scores to postpartum depression. *Arch Womens Ment Health*. 2014;17(6):519-528. doi:10.1007/s00737-014-0428-5
81. Insel T, Cuthbert B, Garvey M, et al. Research Domain Criteria (RDoC): Toward a New Classification Framework for Research on Mental Disorders. *AJP*. 2010;167(7):748-751. doi:10.1176/appi.ajp.2010.09091379
82. Lynall ME, McIntosh AM. The Heterogeneity of Depression. *AJP*. 2023;180(10):703-704. doi:10.1176/appi.ajp.20230574
83. Mandal S, Moudgil M, Mandal SK. Rational drug design. *European Journal of Pharmacology*. 2009;625(1):90-100. doi:10.1016/j.ejphar.2009.06.065
84. Freis ED. Mental Depression in Hypertensive Patients Treated for Long Periods with Large Doses of Reserpine. *New England Journal of Medicine*. 1954;251(25):1006-1008. doi:10.1056/NEJM195412162512504
85. Muller JC. DEPRESSION AND ANXIETY OCCURRING DURING RAUWOLFIA THERAPY. *JAMA*. 1955;159(9):836. doi:10.1001/jama.1955.02960260006002
86. CRANE GE. The psychiatric side-effects of iproniazid. *AJP*. 1956;112(7):494-501. doi:10.1176/ajp.112.7.494
87. Feighner JP. Mechanism of Action of Antidepressant Medications. *Journal of Clinical Psychiatry*. 1999;60(4):4-13.
88. Gardner DM, Shulman KI, Walker SE, Taylor SA. The making of a user friendly MAOI diet. *J Clin Psychiatry*. 1996;57(3):99-104.
89. Brown WA, Rosdolsky M. The Clinical Discovery of Imipramine. *AJP*. 2015;172(5):426-429. doi:10.1176/appi.ajp.2015.14101336
90. KUHN R. The treatment of depressive states with g 22355 (imipramine hydrochloride). *AJP*. 1958;115(5):459-464. doi:10.1176/ajp.115.5.459

91. Santarsieri D, Schwartz TL. Antidepressant efficacy and side-effect burden: a quick guide for clinicians. *Drugs Context*. 2015;4:212290. doi:10.7573/dic.212290
92. Nutt DJ. Death by tricyclic: the real antidepressant scandal? *J Psychopharmacol*. 2005;19(2):123-124. doi:10.1177/0269881105051987
93. Wong DT, Perry KW, Bymaster FP. The Discovery of Fluoxetine Hydrochloride (Prozac). *Nat Rev Drug Discov*. 2005;4(9):764-774. doi:10.1038/nrd1821
94. Stokes PE. Ten years of fluoxetine. *Depression & Anxiety (1091-4269)*. 1998;8:1-4. doi:10.1002/(SICI)1520-6394(1998)8:1+<1::AID-DA1>3.0.CO;2-P
95. Jakubovski E, Johnson JA, Nasir M, Müller-Vahl K, Bloch MH. Systematic review and meta-analysis: Dose-response curve of SSRIs and SNRIs in anxiety disorders. *Depression and Anxiety*. 2019;36(3):198-212. doi:10.1002/da.22854
96. Bakker A, Van Balkom AJLM, Spinhoven P. SSRIs vs. TCAs in the treatment of panic disorder: a meta-analysis. *Acta Psychiatrica Scandinavica*. 2002;106(3):163-167. doi:10.1034/j.1600-0447.2002.02255.x
97. Bloch MH, McGuire J, Landeros-Weisenberger A, Leckman JF, Pittenger C. Metaanalysis of the dose-response relationship of SSRI in obsessive-compulsive disorder. *Mol Psychiatry*. 2010;15(8):850-855. doi:10.1038/mp.2009.50
98. Hilbert A, Hoek HW, Schmidt R. Evidence-based clinical guidelines for eating disorders: international comparison. *Current Opinion in Psychiatry*. 2017;30(6):423. doi:10.1097/YCO.0000000000000360
99. Patetsos E, Horjales-Araujo E. Treating Chronic Pain with SSRIs: What Do We Know? *Pain Research and Management*. 2016;2016(1):2020915. doi:10.1155/2016/2020915
100. Healy D. Lines of Evidence on the Risks of Suicide with Selective Serotonin Reuptake Inhibitors. *Psychotherapy and Psychosomatics*. 2003;72(2):71-79. doi:10.1159/000068691
101. Hammad TA, Laughren T, Racoosin J. Suicidality in Pediatric Patients Treated With Antidepressant Drugs. *Archives of General Psychiatry*. 2006;63(3):332-339. doi:10.1001/archpsyc.63.3.332
102. Murphy SE, Capitão LP, Giles SLC, Cowen PJ, Stringaris A, Harmer CJ. The knowns and unknowns of SSRI treatment in young people with depression and anxiety: efficacy, predictors, and mechanisms of action. *The Lancet Psychiatry*. 2021;8(9):824-835. doi:10.1016/S2215-0366(21)00154-1

103. Cleare A, Pariante CM, Young AH, et al. Evidence-based guidelines for treating depressive disorders with antidepressants: A revision of the 2008 British Association for Psychopharmacology guidelines. *J Psychopharmacol*. 2015;29(5):459-525. doi:10.1177/0269881115581093
104. Gabriel FC, Melo DO de, Fráguas R, Leite-Santos NC, Silva RAM da, Ribeiro E. Pharmacological treatment of depression: A systematic review comparing clinical practice guideline recommendations. *PLOS ONE*. 2020;15(4):e0231700. doi:10.1371/journal.pone.0231700
105. National Institute for Health and Care Excellence. Depression in adults: treatment and management. Accessed February 21, 2025. <https://www.nice.org.uk/guidance/ng222>
106. Bogowicz P, Curtis HJ, Walker AJ, Cowen P, Geddes J, Goldacre B. Trends and variation in antidepressant prescribing in English primary care: a retrospective longitudinal study. *BJGP Open*. 2021;5(4):BJGPO.2021.0020. doi:10.3399/BJGPO.2021.0020
107. Pirraglia PA, Stafford RS, Singer DE. Trends in Prescribing of Selective Serotonin Reuptake Inhibitors and Other Newer Antidepressant Agents in Adult Primary Care. *Prim Care Companion J Clin Psychiatry*. 2003;5(4):153-157.
108. Lambert O, Bourin M. SNRIs: mechanism of action and clinical features. *Expert Review of Neurotherapeutics*. 2002;2(6):849-858. doi:10.1586/14737175.2.6.849
109. Stahl SM, Grady MM, Moret C, Briley M. SNRIs: The Pharmacology, Clinical Efficacy, and Tolerability in Comparison with Other Classes of Antidepressants. *CNS Spectrums*. 2005;10(9):732-747. doi:10.1017/S1092852900019726
110. Nutt D. Mirtazapine: pharmacology in relation to adverse effects. *Acta Psychiatrica Scandinavica*. 1997;96(s391):31-37. doi:10.1111/j.1600-0447.1997.tb05956.x
111. Quitkin FM, Taylor BP, Kremer C. Does mirtazapine have a more rapid onset than SSRIs? *J Clin Psychiatry*. 2001;62(5):358-361. doi:10.4088/jcp.v62n0509
112. Thase ME, Entsuah AR, Rudolph RL. Remission rates during treatment with venlafaxine or selective serotonin reuptake inhibitors. *The British Journal of Psychiatry*. 2001;178(3):234-241. doi:10.1192/bjp.178.3.234
113. Cipriani A, Furukawa TA, Salanti G, et al. Comparative efficacy and acceptability of 21 antidepressant drugs for the acute treatment of adults with major depressive disorder: a systematic review and network meta-analysis. *Lancet*. 2018;391(10128):1357-1366. doi:10.1016/S0140-6736(17)32802-7
114. Sackeim HA. The definition and meaning of treatment-resistant depression. *J Clin Psychiatry*. 2001;62 Suppl 16:10-17.

115. Trivedi MH, Rush AJ, Wisniewski SR, et al. Evaluation of outcomes with citalopram for depression using measurement-based care in STAR*D: implications for clinical practice. *Am J Psychiatry*. 2006;163(1):28-40. doi:10.1176/appi.ajp.163.1.28
116. Rush AJ, Trivedi MH, Wisniewski SR, et al. Acute and Longer-Term Outcomes in Depressed Outpatients Requiring One or Several Treatment Steps: A STAR*D Report. *AJP*. 2006;163(11):1905-1917. doi:10.1176/ajp.2006.163.11.1905
117. Souery D, Serretti A, Calati R, et al. Switching antidepressant class does not improve response or remission in treatment-resistant depression. *J Clin Psychopharmacol*. 2011;31(4):512-516. doi:10.1097/JCP.0b013e3182228619
118. Judd LL, Akiskal HS, Maser JD, et al. Major depressive disorder: A prospective study of residual subthreshold depressive symptoms as predictor of rapid relapse. *Journal of Affective Disorders*. 1998;50(2):97-108. doi:10.1016/S01650327(98)00138-4
119. Israel JA. The Impact of Residual Symptoms in Major Depression. *Pharmaceuticals (Basel)*. 2010;3(8):2426-2440. doi:10.3390/ph3082426
120. Fabbri C. Treatment-resistant depression: role of genetic factors in the perspective of clinical stratification and treatment personalisation. *Mol Psychiatry*. Published online January 18, 2025:1-9. doi:10.1038/s41380-025-02899-0
121. Howes OD, Thase ME, Pillinger T. Treatment resistance in psychiatry: state of the art and new directions. *Mol Psychiatry*. 2022;27(1):58-72. doi:10.1038/s41380-02101200-3
122. Israel JA. The Impact of Residual Symptoms in Major Depression. *Pharmaceuticals (Basel)*. 2010;3(8):2426-2440. doi:10.3390/ph3082426
123. Mrazek DA, Hornberger JC, Altar CA, Degtiar I. A Review of the Clinical, Economic, and Societal Burden of Treatment-Resistant Depression: 1996–2013. *PS*. 2014;65(8):977-987. doi:10.1176/appi.ps.201300059
124. Crown WH, Finkelstein S, Berndt ER, et al. The impact of treatment-resistant depression on health care utilization and costs. *Journal of Clinical Psychiatry*. 2002;63(11):963-971.
125. Gaynes BN, Warden D, Trivedi MH, Wisniewski SR, Fava M, Rush AJ. What Did STAR*D Teach Us? Results From a Large-Scale, Practical, Clinical Trial for Patients With Depression. *PS*. 2009;60(11):1439-1445. doi:10.1176/ps.2009.60.11.1439
126. Simon GE, Perlis RH. Personalized Medicine for Depression: Can We Match Patients With Treatments? *AJP*. 2010;167(12):1445-1455. doi:10.1176/appi.ajp.2010.09111680

127. Vekhova KA, Namiot ED, Jonsson J, Schiöth HB. Ketamine and Esketamine in Clinical Trials: FDA-Approved and Emerging Indications, Trial Trends With Putative Mechanistic Explanations. *Clin Pharmacol Ther.* 2024;117(2):374-386. doi:10.1002/cpt.3478
128. aan het Rot M, Collins KA, Murrough JW, et al. Safety and Efficacy of Repeated-Dose Intravenous Ketamine for Treatment-Resistant Depression. *Biological Psychiatry.* 2010;67(2):139-145. doi:10.1016/j.biopsych.2009.08.038
129. Xu Y, Hackett M, Carter G, et al. Effects of Low-Dose and Very Low-Dose Ketamine among Patients with Major Depression: a Systematic Review and MetaAnalysis. *International Journal of Neuropsychopharmacology.* 2016;19(4):pyv124. doi:10.1093/ijnp/pyv124
130. Price RB, Nock MK, Charney DS, Mathew SJ. Effects of Intravenous Ketamine on Explicit and Implicit Measures of Suicidality in Treatment-Resistant Depression. *Biological Psychiatry.* 2009;66(5):522-526. doi:10.1016/j.biopsych.2009.04.029
131. Swainson J, Klassen LJ, Brennan S, et al. Non-parenteral Ketamine for Depression: A Practical Discussion on Addiction Potential and Recommendations for Judicious Prescribing. *CNS Drugs.* 2022;36(3):239-251. doi:10.1007/s40263022-00897-2
132. Smith-Apeldoorn SY, Veraart JK, Spijker J, Kamphuis J, Schoevers RA. Maintenance ketamine treatment for depression: a systematic review of efficacy, safety, and tolerability. *The Lancet Psychiatry.* 2022;9(11):907-921. doi:10.1016/S2215-0366(22)00317-0
133. Zhu W, Ding Z, Zhang Y, Shi J, Hashimoto K, Lu L. Risks Associated with Misuse of Ketamine as a Rapid-Acting Antidepressant. *Neurosci Bull.* 2016;32(6):557-564. doi:10.1007/s12264-016-0081-2
134. Minikel EV, Painter JL, Dong CC, Nelson MR. Refining the impact of genetic evidence on clinical success. *Nature.* 2024;629(8012):624-629. doi:10.1038/s41586-024-07316-0
135. Nelson MR, Tipney H, Painter JL, et al. The support of human genetic evidence for approved drug indications. *Nat Genet.* 2015;47(8):856-860. doi:10.1038/ng.3314
136. Weissman MM, Wickramaratne P, Nomura Y, et al. Families at High and Low Risk for Depression: A 3-Generation Study. *Archives of General Psychiatry.* 2005;62(1):29-36. doi:10.1001/archpsyc.62.1.29
137. Sullivan PF, Neale MC, Kendler KS. Genetic Epidemiology of Major Depression: Review and Meta-Analysis. *AJP.* 2000;157(10):1552-1562. doi:10.1176/appi.ajp.157.10.1552

138. Zhu M, Zhao S. Candidate Gene Identification Approach: Progress and Challenges. *Int J Biol Sci*. 2007;3(7):420-427.
139. Kwon JM, Goate AM. The Candidate Gene Approach. *Alcohol Res Health*. 2000;24(3):164-168.
140. Heils A, Teufel A, Petri S, et al. Allelic Variation of Human Serotonin Transporter Gene Expression. *Journal of Neurochemistry*. 1996;66(6):2621-2624. doi:10.1046/j.1471-4159.1996.66062621.x
141. Zhu M, Zhao S. Candidate Gene Identification Approach: Progress and Challenges. *Int J Biol Sci*. 2007;3(7):420-427.
142. Border R, Johnson EC, Evans LM, et al. No Support for Historical Candidate Gene or Candidate Gene-by-Interaction Hypotheses for Major Depression Across Multiple Large Samples. *AJP*. 2019;176(5):376-387. doi:10.1176/appi.ajp.2018.18070881
143. Caspi A, Sugden K, Moffitt TE, et al. Influence of Life Stress on Depression: Moderation by a Polymorphism in the 5-HTT Gene. *Science*. 2003;301(5631):386-389. doi:10.1126/science.1083968
144. Culverhouse RC, Saccone NL, Horton AC, et al. Collaborative meta-analysis finds no evidence of a strong interaction between stress and 5-HTTLPR genotype contributing to the development of depression. *Mol Psychiatry*. 2018;23(1):133-142. doi:10.1038/mp.2017.44
145. Fergusson DM, Horwood LJ, Miller AL, Kennedy MA. Life stress, 5-HTTLPR and mental disorder: findings from a 30-year longitudinal study. *The British Journal of Psychiatry*. 2011;198(2):129-135. doi:10.1192/bjp.bp.110.085993
146. Keller MC. Gene × Environment Interaction Studies Have Not Properly Controlled for Potential Confounders: The Problem and the (Simple) Solution. *Biological Psychiatry*. 2014;75(1):18-24. doi:10.1016/j.biopsych.2013.09.006
147. Bosker FJ, Hartman CA, Nolte IM, et al. Poor replication of candidate genes for major depressive disorder using genome-wide association data. *Mol Psychiatry*. 2011;16(5):516-532. doi:10.1038/mp.2010.38
148. Lander ES, Linton LM, Birren B, et al. Initial sequencing and analysis of the human genome. *Nature*. 2001;409(6822):860-921. doi:10.1038/35057062
149. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature*. 2004;431(7011):931-945. doi:10.1038/nature03001
150. Visscher PM, Brown MA, McCarthy MI, Yang J. Five Years of GWAS Discovery.

- The American Journal of Human Genetics*. 2012;90(1):7-24.
doi:10.1016/j.ajhg.2011.11.029
151. Collins FS, Guyer MS, Chakravarti A. Variations on a Theme: Cataloging Human DNA Sequence Variation. *Science*. 1997;278(5343):1580-1581.
doi:10.1126/science.278.5343.1580
152. Hill WG, Robertson A. Linkage disequilibrium in finite populations. *Theoret Appl Genetics*. 1968;38(6):226-231. doi:10.1007/BF01245622
153. The International HapMap Consortium. The International HapMap Project. *Nature*. Published online December 18, 2003. doi:10.1038/nature02168
154. Altshuler D, Donnelly P, The International HapMap Consortium. A haplotype map of the human genome. *Nature*. 2005;437(7063):1299-1320. doi:10.1038/nature04226
155. Burton PR, Clayton DG, Cardon LR, et al. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*. 2007;447(7145):661-678. doi:10.1038/nature05911
156. Yang J, Manolio TA, Pasquale LR, et al. Genome partitioning of genetic variation for complex traits using common SNPs. *Nat Genet*. 2011;43(6):519-525.
doi:10.1038/ng.823
157. Yang J, Benyamin B, McEvoy BP, et al. Common SNPs explain a large proportion of the heritability for human height. *Nat Genet*. 2010;42(7):565-569.
doi:10.1038/ng.608
158. Wray NR, Lin T, Austin J, et al. From Basic Science to Clinical Application of Polygenic Risk Scores: A Primer. *JAMA Psychiatry*. 2021;78(1):101-109.
doi:10.1001/jamapsychiatry.2020.3049
159. Cerezo M, Sollis E, Ji Y, et al. The NHGRI-EBI GWAS Catalog: standards for reusability, sustainability and diversity. *Nucleic Acids Research*. 2025;53(D1):D998-D1005. doi:10.1093/nar/gkae1070
160. Rietschel M, Mattheisen M, Frank J, et al. Genome-wide association-, replication-, and neuroimaging study implicates HOMER1 in the etiology of major depression. *Biol Psychiatry*. 2010;68(6):578-585. doi:10.1016/j.biopsych.2010.05.038
161. Sullivan PF, de Geus EJC, Willemsen G, et al. Genome-wide association for major depressive disorder: a possible role for the presynaptic protein piccolo. *Mol Psychiatry*. 2009;14(4):359-375. doi:10.1038/mp.2008.125
162. Lewis CM, Ng MY, Butler AW, et al. Genome-Wide Association Study of Major Recurrent Depression in the U.K. Population. *AJP*. 2010;167(8):949-957.
doi:10.1176/appi.ajp.2010.09091380

163. Shi J, Potash JB, Knowles JA, et al. Genome-wide association study of recurrent early-onset major depressive disorder. *Mol Psychiatry*. 2011;16(2):193-201. doi:10.1038/mp.2009.124
164. Muglia P, Tozzi F, Galwey NW, et al. Genome-wide association study of recurrent major depressive disorder in two European case-control cohorts. *Mol Psychiatry*. 2010;15(6):589-601. doi:10.1038/mp.2008.131
165. Wray NR, Pergadia ML, Blackwood DHR, et al. Genome-wide association study of major depressive disorder: new results, meta-analysis, and lessons learned. *Mol Psychiatry*. 2012;17(1):36-48. doi:10.1038/mp.2010.109
166. Ripke S, Wray N, Lewis C, et al. A mega-analysis of genome-wide association studies for major depressive disorder. *Mol Psychiatry*. 2013;18(4):10.1038/mp.2012.21. doi:10.1038/mp.2012.21
167. Lee SH, Ripke S, Neale BM, et al. Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. *Nat Genet*. 2013;45(9):984-994. doi:10.1038/ng.2711
168. Hek K, Demirkan A, Lahti J, et al. A Genome-Wide Association Study of Depressive Symptoms. *Biological Psychiatry*. 2013;73(7):667-678. doi:10.1016/j.biopsych.2012.09.033
169. Purcell SM, Wray NR, Stone JL, et al. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*. 2009;460(7256):748-752. doi:10.1038/nature08185
170. Mullins N, Lewis CM. Genetics of Depression: Progress at Last. *Curr Psychiatry Rep*. 2017;19(8):43. doi:10.1007/s11920-017-0803-9
171. CONVERGE consortium. Sparse whole genome sequencing identifies two loci for major depressive disorder. *Nature*. 2015;523(7562):588-591. doi:10.1038/nature14659
172. Okbay A, Baselmans BML, De Neve JE, et al. Genetic variants associated with subjective well-being, depressive symptoms, and neuroticism identified through genome-wide analyses. *Nat Genet*. 2016;48(6):624-633. doi:10.1038/ng.3552
173. Hyde CL, Nagle MW, Tian C, et al. Identification of 15 genetic loci associated with risk of major depression in individuals of European descent. *Nat Genet*. 2016;48(9):1031-1036. doi:10.1038/ng.3623
174. Cai N, Revez JA, Adams MJ, et al. Minimal phenotyping yields genome-wide association signals of low specificity for major depression. *Nat Genet*. 2020;52(4):437-447. doi:10.1038/s41588-020-0594-5

175. Wray NR, Ripke S, Mattheisen M, et al. Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nat Genet.* 2018;50(5):668-681. doi:10.1038/s41588-018-0090-3
176. Howard DM, Adams MJ, Shirali M, et al. Genome-wide association study of depression phenotypes in UK Biobank identifies variants in excitatory synaptic pathways. *Nat Commun.* 2018;9(1):1470. doi:10.1038/s41467-018-03819-3
177. Howard DM, Adams MJ, Clarke TK, et al. Genome-wide meta-analysis of depression identifies 102 independent variants and highlights the importance of the prefrontal brain regions. *Nat Neurosci.* 2019;22(3):343-352. doi:10.1038/s41593018-0326-7
178. Fatumo S, Chikowore T, Choudhury A, Ayub M, Martin AR, Kuchenbaecker K. A roadmap to increase diversity in genomic studies. *Nat Med.* 2022;28(2):243-250. doi:10.1038/s41591-021-01672-4
179. Schaid DJ, Chen W, Larson NB. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat Rev Genet.* 2018;19(8):491-504. doi:10.1038/s41576-018-0016-z
180. Levey DF, Stein MB, Wendt FR, et al. Bi-ancestral depression GWAS in the Million Veteran Program and meta-analysis in >1.2 million individuals highlight new therapeutic directions. *Nat Neurosci.* 2021;24(7):954-963. doi:10.1038/s41593-02100860-2
181. Meng X, Navoly G, Giannakopoulou O, et al. Multi-ancestry genome-wide association study of major depression aids locus discovery, fine mapping, gene prioritization and causal inference. *Nat Genet.* 2024;56(2):222-233. doi:10.1038/s41588-023-01596-4
182. Huang QQ, Sallah N, Dunca D, et al. Transferability of genetic loci and polygenic scores for cardiometabolic traits in British Pakistani and Bangladeshi individuals. *Nat Commun.* 2022;13(1):4664. doi:10.1038/s41467-022-32095-5
183. Adams MJ, Streit F, Meng X, et al. Trans-ancestry genome-wide study of depression identifies 697 associations implicating cell types and pharmacotherapies. *Cell.* Published online January 14, 2025. doi:10.1016/j.cell.2024.12.002
184. Hindorff LA, Sethupathy P, Junkins HA, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences.* 2009;106(23):9362-9367. doi:10.1073/pnas.0903103106
185. Bunt M van de, Cortes A, Consortium I, Brown MA, Morris AP, McCarthy MI. Evaluating the Performance of Fine-Mapping Strategies at Common Variant GWAS

- Loci. *PLOS Genetics*. 2015;11(9):e1005535. doi:10.1371/journal.pgen.1005535
186. Zaykin DV, Zhivotovsky LA. Ranks of Genuine Associations in Whole-Genome Scans. *Genetics*. 2005;171(2):813-823. doi:10.1534/genetics.105.044206
187. Spain SL, Barrett JC. Strategies for fine-mapping complex traits. *Human Molecular Genetics*. 2015;24(R1):R111-R119. doi:10.1093/hmg/ddv260
188. Cano-Gamez E, Trynka G. From GWAS to Function: Using Functional Genomics to Identify the Mechanisms Underlying Complex Diseases. *Front Genet*. 2020;11. doi:10.3389/fgene.2020.00424
189. Maurano MT, Humbert R, Rynes E, et al. Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. *Science*. 2012;337(6099):1190-1195. doi:10.1126/science.1222794
190. Farh KKH, Marson A, Zhu J, et al. Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature*. 2015;518(7539):337-343. doi:10.1038/nature13835
191. Musunuru K, Strong A, Frank-Kamenetsky M, et al. From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. *Nature*. 2010;466(7307):714-719. doi:10.1038/nature09266
192. Zhu Z, Zhang F, Hu H, et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat Genet*. 2016;48(5):481-487. doi:10.1038/ng.3538
193. Sanyal A, Lajoie BR, Jain G, Dekker J. The long-range interaction landscape of gene promoters. *Nature*. 2012;489(7414):109-113. doi:10.1038/nature11279
194. Smemo S, Tena JJ, Kim KH, et al. Obesity-associated variants within FTO form long-range functional connections with IRX3. *Nature*. 2014;507(7492):371-375. doi:10.1038/nature13138
195. Claussnitzer M, Dankel SN, Kim KH, et al. FTO Obesity Variant Circuitry and Adipocyte Browning in Humans. *New England Journal of Medicine*. 2015;373(10):895-907. doi:10.1056/NEJMoa1502214
196. Risch N, Merikangas K. The Future of Genetic Studies of Complex Human Diseases. *Science*. 1996;273(5281):1516-1517. doi:10.1126/science.273.5281.1516
197. Maher B. Personal genomes: The case of the missing heritability. *Nature*. 2008;456(7218):18-21. doi:10.1038/456018a

198. Macgregor S, Cornes BK, Martin NG, Visscher PM. Bias, precision and heritability of self-reported and clinically measured height in Australian twins. *Hum Genet.* 2006;120(4):571-580. doi:10.1007/s00439-006-0240-z
199. Silventoinen K, Sammalisto S, Perola M, et al. Heritability of Adult Body Height: A Comparative Study of Twin Cohorts in Eight Countries. *Twin Research and Human Genetics.* 2003;6(5):399-408. doi:10.1375/twin.6.5.399
200. Visscher PM. Sizing up human height variation. *Nat Genet.* 2008;40(5):489-490. doi:10.1038/ng0508-489
201. Eichler EE, Flint J, Gibson G, et al. Missing heritability and strategies for finding the underlying causes of complex disease. *Nat Rev Genet.* 2010;11(6):446-450. doi:10.1038/nrg2809
202. Matthews LJ, Turkheimer E. Three Legs of the Missing Heritability Problem. *Stud Hist Philos Sci.* 2022;93:183-191. doi:10.1016/j.shpsa.2022.04.004
203. Manolio TA, Collins FS, Cox NJ, et al. Finding the missing heritability of complex diseases. *Nature.* 2009;461(7265):747-753. doi:10.1038/nature08494
204. Kaprio J. Twins and the mystery of missing heritability: the contribution of gene-environment interactions. *Journal of Internal Medicine.* 2012;272(5):440-448. doi:10.1111/j.1365-2796.2012.02587.x
205. Zuk O, Hechter E, Sunyaev SR, Lander ES. The mystery of missing heritability: Genetic interactions create phantom heritability. *Proceedings of the National Academy of Sciences.* 2012;109(4):1193-1198. doi:10.1073/pnas.1119675109
206. Young AI. Solving the missing heritability problem. *PLOS Genetics.* 2019;15(6):e1008222. doi:10.1371/journal.pgen.1008222
207. Visscher PM, Medland SE, Ferreira MAR, et al. Assumption-Free Estimation of Heritability from Genome-Wide Identity-by-Descent Sharing between Full Siblings. *PLOS Genetics.* 2006;2(3):e41. doi:10.1371/journal.pgen.0020041
208. Pritchard JK. Are Rare Variants Responsible for Susceptibility to Complex Diseases? *The American Journal of Human Genetics.* 2001;69(1):124-137. doi:10.1086/321272
209. Manolio TA. Genomewide Association Studies and Assessment of the Risk of Disease. *New England Journal of Medicine.* 2010;363(2):166-176. doi:10.1056/NEJMra0905980
210. Pe'er I, Yelensky R, Altshuler D, Daly MJ. Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. *Genetic Epidemiology.* 2008;32(4):381-385. doi:10.1002/gepi.20303

211. Yang J, Zeng J, Goddard ME, Wray NR, Visscher PM. Concepts, estimation and interpretation of SNP-based heritability. *Nat Genet.* 2017;49(9):1304-1310. doi:10.1038/ng.3941
212. Yang J, Benyamin B, McEvoy BP, et al. Common SNPs explain a large proportion of the heritability for human height. *Nat Genet.* 2010;42(7):565-569. doi:10.1038/ng.608
213. Yengo L, Vedantam S, Marouli E, et al. A saturated map of common genetic variants associated with human height. *Nature.* 2022;610(7933):704-712. doi:10.1038/s41586-022-05275-y
214. Visscher PM, Goddard ME, Derks EM, Wray NR. Evidence-based psychiatric genetics, AKA the false dichotomy between common and rare variant hypotheses. *Mol Psychiatry.* 2012;17(5):474-485. doi:10.1038/mp.2011.65
215. Momozawa Y, Mizukami K. Unique roles of rare variants in the genetics of complex diseases in humans. *J Hum Genet.* 2021;66(1):11-23. doi:10.1038/s10038-02000845-2
216. McClellan J, King MC. Genetic Heterogeneity in Human Disease. *Cell.* 2010;141(2):210-217. doi:10.1016/j.cell.2010.03.032
217. Das S, Abecasis GR, Browning BL. Genotype Imputation from Large Reference Panels. *Annual Review of Genomics and Human Genetics.* 2018;19(Volume 19, 2018):73-96. doi:10.1146/annurev-genom-083117-021602
218. Cortes A, Brown MA. Promise and pitfalls of the Immunochip. *Arthritis Res Ther.* 2011;13(1):101. doi:10.1186/ar3204
219. Voight BF, Kang HM, Ding J, et al. The Metabochip, a Custom Genotyping Array for Genetic Studies of Metabolic, Cardiovascular, and Anthropometric Traits. *PLOS Genetics.* 2012;8(8):e1002793. doi:10.1371/journal.pgen.1002793
220. Alekseyev YO, Fazeli R, Yang S, et al. A Next-Generation Sequencing Primer—How Does It Work and What Can It Do? *Acad Pathol.* 2018;5:2374289518766521. doi:10.1177/2374289518766521
221. Shendure J, Ji H. Next-generation DNA sequencing. *Nat Biotechnol.* 2008;26(10):1135-1145. doi:10.1038/nbt1486
222. Hu T, Chitnis N, Monos D, Dinh A. Next-generation sequencing technologies: An overview. *Human Immunology.* 2021;82(11):801-811. doi:10.1016/j.humimm.2021.02.012
223. Sazonovs A, Barrett JC. Rare-Variant Studies to Complement Genome-Wide Association Studies. *Annual Review of Genomics and Human Genetics.*

- 2018;19(Volume 19, 2018):97-112. doi:10.1146/annurev-genom-083117-021641
224. Cirulli ET, Goldstein DB. Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nat Rev Genet.* 2010;11(6):415-425. doi:10.1038/nrg2779
225. Backman JD, Li AH, Marcketta A, et al. Exome sequencing and analysis of 454,787 UK Biobank participants. *Nature.* 2021;599(7886):628-634. doi:10.1038/s41586021-04103-z
226. Li S, Carss KJ, Halldorsson BV, Cortes A, Consortium UBWGS. Whole-genome sequencing of half-a-million UK Biobank participants. Published online December 8, 2023:2023.12.06.23299426. doi:10.1101/2023.12.06.23299426
227. The All of Us Research Program Investigators. The “All of Us” Research Program. *New England Journal of Medicine.* 2019;381(7):668-676. doi:10.1056/NEJMSr1809937
228. Bick AG, Metcalf GA, Mayo KR, et al. Genomic data in the All of Us Research Program. *Nature.* 2024;627(8003):340-346. doi:10.1038/s41586-023-06957-x
229. Sazonovs A, Barrett JC. Rare-Variant Studies to Complement Genome-Wide Association Studies. *Annu Rev Genom Hum Genet.* 2018;19(1):97-112. doi:10.1146/annurev-genom-083117-021641
230. Gorlov I, Gorlova O, Frazier M, Spitz M, Amos C. Evolutionary evidence of the effect of rare variants on disease etiology. *Clinical Genetics.* 2011;79(3):199-206. doi:10.1111/j.1399-0004.2010.01535.x
231. Povysil G, Petrovski S, Hostyk J, Aggarwal V, Allen AS, Goldstein DB. Rare-variant collapsing analyses for complex traits: guidelines and applications. *Nat Rev Genet.* 2019;20(12):747-759. doi:10.1038/s41576-019-0177-4
232. Auer PL, Lettre G. Rare variant association studies: considerations, challenges and opportunities. *Genome Med.* 2015;7(1):16. doi:10.1186/s13073-015-0138-2
233. McCaw ZR, O’Dushlaine C, Sominen H, et al. An allelic-series rare-variant association test for candidate-gene discovery. *Am J Hum Genet.* 2023;110(8):1330-1342. doi:10.1016/j.ajhg.2023.07.001
234. Reich DE, Lander ES. On the allelic spectrum of human disease. *Trends in Genetics.* 2001;17(9):502-510. doi:10.1016/S0168-9525(01)02410-6
235. Plenge RM, Scolnick EM, Altshuler D. Validating therapeutic targets through human genetics. *Nat Rev Drug Discov.* 2013;12(8):581-594. doi:10.1038/nrd4051

236. Kotowski IK, Pertsemlidis A, Luke A, et al. A Spectrum of PCSK9 Alleles Contributes to Plasma Levels of Low-Density Lipoprotein Cholesterol. *Am J Hum Genet.* 2006;78(3):410-422.
237. Cohen JC, Boerwinkle E, Mosley TH, Hobbs HH. Sequence Variations in PCSK9, Low LDL, and Protection against Coronary Heart Disease. *New England Journal of Medicine.* 2006;354(12):1264-1272. doi:10.1056/NEJMoa054013
238. Sabatine MS, Giugliano RP, Keech AC, et al. Evolocumab and Clinical Outcomes in Patients with Cardiovascular Disease. *New England Journal of Medicine.* 2017;376(18):1713-1722. doi:10.1056/NEJMoa1615664
239. Wang Q, Dhindsa RS, Carss K, et al. Rare variant contribution to human disease in 281,104 UK Biobank exomes. *Nature.* 2021;597(7877):527-532. doi:10.1038/s41586-021-03855-y
240. Bomba L, Walter K, Soranzo N. The impact of rare and low-frequency genetic variants in common disease. *Genome Biol.* 2017;18(1):77. doi:10.1186/s13059017-1212-4
241. Chen CY, Tian R, Ge T, et al. The impact of rare protein coding genetic variation on adult cognitive function. *Nat Genet.* 2023;55(6):927-938. doi:10.1038/s41588023-01398-8
242. Jurgens SJ, Wang X, Choi SH, et al. Rare coding variant analysis for human diseases across biobanks and ancestries. *Nat Genet.* 2024;56(9):1811-1820. doi:10.1038/s41588-024-01894-5
243. Jurgens SJ, Choi SH, Morrill VN, et al. Analysis of rare genetic variation underlying cardiometabolic diseases and traits among 200,000 individuals in the UK Biobank. *Nat Genet.* 2022;54(3):240-250. doi:10.1038/s41588-021-01011-w
244. Szustakowski JD, Balasubramanian S, Kvikstad E, et al. Advancing human genetics research and drug discovery through exome sequencing of the UK Biobank. *Nat Genet.* 2021;53(7):942-948. doi:10.1038/s41588-021-00885-0
245. Deng YT, Wu BS, Yang L, et al. Large-scale whole-exome sequencing of neuropsychiatric diseases and traits in 350,770 adults. *Nat Hum Behav.* 2024;8(6):1194-1208. doi:10.1038/s41562-024-01861-4
246. Andreassen OA, Hindley GFL, Frei O, Smeland OB. New insights from the last decade of research in psychiatric genetics: discoveries, challenges and clinical implications. *World Psychiatry.* 2023;22(1):4-24. doi:10.1002/wps.21034
247. Ganna A, Satterstrom FK, Zekavat SM, et al. Quantifying the Impact of Rare and Ultra-rare Coding Variation across the Phenotypic Spectrum. *The American Journal of Human Genetics.* 2018;102(6):1204-1211. doi:10.1016/j.ajhg.2018.05.002

248. Hamming RW. Error detecting and error correcting codes. *The Bell System Technical Journal*. 1950;29(2):147-160. doi:10.1002/j.1538-7305.1950.tb00463.x
249. Pinheiro HP, de Souza Pinheiro A, Sen PK. Comparison of genomic sequences using the Hamming distance. *Journal of Statistical Planning and Inference*. 2005;130(1):325-339. doi:10.1016/j.jspi.2003.03.002
250. Yu C, Arcos-Burgos M, Baune BT, et al. Low-frequency and rare variants may contribute to elucidate the genetics of major depressive disorder. *Transl Psychiatry*. 2018;8(1):1-8. doi:10.1038/s41398-018-0117-7
251. Tian R, Ge T, Kweon H, et al. Whole-exome sequencing in UK Biobank reveals rare genetic architecture for depression. *Nat Commun*. 2024;15(1):1755. doi:10.1038/s41467-024-45774-2
252. Cheng S, Cheng B, Liu L, et al. Exome-wide screening identifies novel rare risk variants for major depression disorder. *Mol Psychiatry*. 2022;27(7):3069-3074. doi:10.1038/s41380-022-01536-4
253. Li ZY, Fei CJ, Yin RY, et al. Whole exome sequencing identified six novel genes for depressive symptoms. *Mol Psychiatry*. Published online October 29, 2024:1-12. doi:10.1038/s41380-024-02804-1
254. Hasin Y, Seldin M, Lusis A. Multi-omics approaches to disease. *Genome Biology*. 2017;18(1):83. doi:10.1186/s13059-017-1215-1
255. Wu H, Eckhardt CM, Baccarelli AA. Molecular mechanisms of environmental exposures and human disease. *Nat Rev Genet*. 2023;24(5):332-344. doi:10.1038/s41576-022-00569-3
256. Crick F. Central dogma of molecular biology. *Nature*. 1970;227(5258):561-563. doi:10.1038/227561a0
257. The ENCODE Project Consortium. An Integrated Encyclopedia of DNA Elements in the Human Genome. *Nature*. 2012;489(7414):57-74. doi:10.1038/nature11247
258. Gibney ER, Nolan CM. Epigenetics and gene expression. *Heredity*. 2010;105(1):4-13. doi:10.1038/hdy.2010.54
259. Nichols RG, Davenport ER. The relationship between the gut microbiome and host gene expression: a review. *Hum Genet*. 2021;140(5):747-760. doi:10.1007/s00439020-02237-0
260. Smoleńska Ż, Zdrojewski Z. Metabolomics and its potential in diagnosis, prognosis and treatment of rheumatic diseases. *Reumatologia*. 2015;53(3):152-156. doi:10.5114/reum.2015.53137

261. Lempp M, Farke N, Kuntz M, Freibert SA, Lill R, Link H. Systematic identification of metabolites controlling gene expression in *E. coli*. *Nat Commun*. 2019;10(1):4463. doi:10.1038/s41467-019-12474-1
262. Hillary RF, McCartney DL, Bernabeu E, et al. Blood-based epigenome-wide analyses on the prevalence and incidence of nineteen common disease states. Published online January 11, 2023:2023.01.10.23284387. doi:10.1101/2023.01.10.23284387
263. Aguet F, Alasoo K, Li YI, et al. Molecular quantitative trait loci. *Nat Rev Methods Primers*. 2023;3(1):1-22. doi:10.1038/s43586-022-00188-6
264. Brandt M, Lappalainen T. SnapShot: Discovering Genetic Regulatory Variants by QTL Analysis. *Cell*. 2017;171(4):980-980.e1. doi:10.1016/j.cell.2017.10.031
265. Nica AC, Dermitzakis ET. Expression quantitative trait loci: present and future. *Philos Trans R Soc Lond B Biol Sci*. 2013;368(1620):20120362. doi:10.1098/rstb.2012.0362
266. Vösa U, Claringbould A, Westra HJ, et al. Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. *Nat Genet*. 2021;53(9):1300-1310. doi:10.1038/s41588021-00913-z
267. Lonsdale J, Thomas J, Salvatore M, et al. The Genotype-Tissue Expression (GTEx) project. *Nat Genet*. 2013;45(6):580-585. doi:10.1038/ng.2653
268. Stranger BE, Nica AC, Forrest MS, et al. Population genomics of human gene expression. *Nat Genet*. 2007;39(10):1217-1224. doi:10.1038/ng2142
269. Emilsson V, Thorleifsson G, Zhang B, et al. Genetics of gene expression and its effect on disease. *Nature*. 2008;452(7186):423-428. doi:10.1038/nature06758
270. The GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science*. 2020;369(6509):1318-1330. doi:10.1126/science.aaz1776
271. Carithers LJ, Ardlie K, Barcus M, et al. A Novel Approach to High-Quality Postmortem Tissue Procurement: The GTEx Project. *Biopreservation and Biobanking*. 2015;13(5):311-319. doi:10.1089/bio.2015.0032
272. Yang H, Liu D, Zhao C, et al. Mendelian randomization integrating GWAS and eQTL data revealed genes pleiotropically associated with major depressive disorder. *Transl Psychiatry*. 2021;11:225. doi:10.1038/s41398-021-01348-0
273. Barrett LW, Fletcher S, Wilton SD. Regulation of eukaryotic gene expression by the untranslated gene regions and other non-coding elements. *Cell Mol Life Sci*.

- 2012;69(21):3613-3634. doi:10.1007/s00018-012-0990-9
274. Moreno-Moya JM, Vilella F, Simón C. MicroRNA: key gene expression regulators. *Fertility and Sterility*. 2014;101(6):1516-1523. doi:10.1016/j.fertnstert.2013.10.042
275. Garcia-Blanco MA, Baraniak AP, Lasda EL. Alternative splicing in disease and therapy. *Nat Biotechnol*. 2004;22(5):535-546. doi:10.1038/nbt964
276. Karve TM, Cheema AK. Small Changes Huge Impact: The Role of Protein Posttranslational Modifications in Cellular Homeostasis and Disease. *Journal of Amino Acids*. 2011;2011(1):207691. doi:10.4061/2011/207691
277. Wang KC, Chang HY. Epigenomics—Technologies and Applications. *Circ Res*. 2018;122(9):1191-1199. doi:10.1161/CIRCRESAHA.118.310998
278. Yang X. Multi-tissue Multi-omics Systems Biology to Dissect Complex Diseases. *Trends Mol Med*. 2020;26(8):718-728. doi:10.1016/j.molmed.2020.04.006
279. Patti GJ, Yanes O, Siuzdak G. Metabolomics: the apogee of the omics trilogy. *Nat Rev Mol Cell Biol*. 2012;13(4):263-269. doi:10.1038/nrm3314
280. Pu J, Yu Y, Liu Y, et al. ProMENDA: an updated resource for proteomic and metabolomic characterization in depression. *Transl Psychiatry*. 2024;14(1):1-7. doi:10.1038/s41398-024-02948-2
281. Rimti FH, Shahbaz R, Bhatt K, Xiang A. A review of new insights into existing major depressive disorder biomarkers. *Heliyon*. 2023;9(8). doi:10.1016/j.heliyon.2023.e18909
282. Tzoulaki I, Ebbels TMD, Valdes A, Elliott P, Ioannidis JPA. Design and Analysis of Metabolomics Studies in Epidemiologic Research: A Primer on -Omic Technologies. *American Journal of Epidemiology*. 2014;180(2):129-139. doi:10.1093/aje/kwu143
283. Baker M. Metabolomics: from small molecules to big ideas. *Nat Methods*. 2011;8(2):117-121. doi:10.1038/nmeth0211-117
284. Shah S, McRae AF, Marioni RE, et al. Genetic and environmental exposures constrain epigenetic drift over the human life course. *Genome Res*. 2014;24(11):1725-1733. doi:10.1101/gr.176933.114
285. Baker SA, Rutter J. Metabolites as signalling molecules. *Nat Rev Mol Cell Biol*. 2023;24(5):355-374. doi:10.1038/s41580-022-00572-w
286. Würtz P, Kangas AJ, Soininen P, Lawlor DA, Davey Smith G, Ala-Korpela M. Quantitative Serum Nuclear Magnetic Resonance Metabolomics in Large-Scale Epidemiology: A Primer on -Omic Technologies. *American Journal of Epidemiology*.

- 2017;186(9):1084-1096. doi:10.1093/aje/kwx016
287. Wishart DS, Feunang YD, Marcu A, et al. HMDB 4.0: the human metabolome database for 2018. *Nucleic Acids Res.* 2018;46(Database issue):D608-D617. doi:10.1093/nar/gkx1089
 288. Pearson H. Meet the human metabolome. *Nature.* 2007;446(7131):8-8. doi:10.1038/446008a
 289. Lindon JC, Nicholson JK. Analytical technologies for metabonomics and metabolomics, and multi-omic information recovery. *TrAC Trends in Analytical Chemistry.* 2008;27(3):194-204. doi:10.1016/j.trac.2007.08.009
 290. Shah SH, Kraus WE, Newgard CB. Metabolomic Profiling for the Identification of Novel Biomarkers and Mechanisms Related to Common Cardiovascular Diseases. *Circulation.* 2012;126(9):1110-1120. doi:10.1161/CIRCULATIONAHA.111.060368
 291. Gowda GAN, Raftery D. NMR Based Metabolomics. *Adv Exp Med Biol.* 2021;1280:19-37. doi:10.1007/978-3-030-51652-9_2
 292. Emwas AH, Roy R, McKay RT, et al. NMR Spectroscopy for Metabolomics Research. *Metabolites.* 2019;9(7):123. doi:10.3390/metabo9070123
 293. Ala-Korpela M, Kangas AJ, Soininen P. Quantitative high-throughput metabolomics: a new era in epidemiology and genetics. *Genome Medicine.* 2012;4(4):36. doi:10.1186/gm335
 294. van den Akker EB, Trompet S, Barkey Wolf JJH, et al. Metabolic Age Based on the BBMRI-NL 1H-NMR Metabolomics Repository as Biomarker of Age-related Disease. *Circulation: Genomic and Precision Medicine.* 2020;13(5):541-547. doi:10.1161/CIRCGEN.119.002610
 295. Nightingale Health Biobank Collaborative Group, Barrett JC, Esko T, et al. Metabolomic and genomic prediction of common diseases in 477,706 participants in three national biobanks. Published online June 12, 2023:2023.06.09.23291213. doi:10.1101/2023.06.09.23291213
 296. Yu B, Zanetti KA, Temprosa M, et al. The Consortium of Metabolomics Studies (COMETS): Metabolomics in 47 Prospective Cohort Studies. *American Journal of Epidemiology.* 2019;188(6):991-1012. doi:10.1093/aje/kwz028
 297. Sabatine MS, Liu E, Morrow DA, et al. Metabolomic Identification of Novel Biomarkers of Myocardial Ischemia. *Circulation.* 2005;112(25):3868-3875. doi:10.1161/CIRCULATIONAHA.105.569137
 298. Wang TJ, Larson MG, Vasan RS, et al. Metabolite profiles and the risk of developing diabetes. *Nat Med.* 2011;17(4):448-453. doi:10.1038/nm.2307

299. Grauslys A, Phelan MM, Broughton C, et al. NMR-based metabolic profiling provides diagnostic and prognostic information in critically ill children with suspected infection. *Sci Rep*. 2020;10(1):20198. doi:10.1038/s41598-020-77319-0
300. Mallol R, Rombauts A, Abelenda-Alonso G, Gudiol C, Balsalobre M, Carratalà J. Metabolomic profile of severe COVID-19 and a signature predictive of progression towards severe disease status: a prospective cohort study (METCOVID). *Sci Rep*. 2025;15(1):4963. doi:10.1038/s41598-025-87288-x
301. Buergel T, Steinfeldt J, Ruyoga G, et al. Metabolomic profiles predict individual multidisease outcomes. *Nat Med*. Published online September 22, 2022:1-12. doi:10.1038/s41591-022-01980-3
302. Luo H, Jiang ZL, Ren Y. Therapy Management of Metabolic Disorder Comorbidity With Depression. *Front Psychol*. 2021;12:683320. doi:10.3389/fpsyg.2021.683320
303. Jokela M, Hamer M, Singh-Manoux A, Batty GD, Kivimäki M. Association of metabolically healthy obesity with depressive symptoms: pooled analysis of eight studies. *Mol Psychiatry*. 2014;19(8):910-914. doi:10.1038/mp.2013.162
304. Pu J, Yu Y, Liu Y, et al. MENDA: a comprehensive curated resource of metabolic characterization in depression. *Briefings in Bioinformatics*. 2020;21(4):1455-1464. doi:10.1093/bib/bbz055
305. Schulte EC, Schulze TG. How do lipids fit into the picture of severe mental health disorders? A missing pathophysiologic link or just a consequence of behavior and treatment. *European Neuropsychopharmacology*. 2024;83:27-29. doi:10.1016/j.euroneuro.2024.02.007
306. Setoyama D, Kato TA, Hashimoto R, et al. Plasma Metabolites Predict Severity of Depression and Suicidal Ideation in Psychiatric Patients-A Multicenter Pilot Analysis. *PLOS ONE*. 2016;11(12):e0165267. doi:10.1371/journal.pone.0165267
307. Tkachev A, Stekolshchikova E, Vanyushkina A, et al. Lipid Alteration Signature in the Blood Plasma of Individuals With Schizophrenia, Depression, and Bipolar Disorder. *JAMA Psychiatry*. 2023;80(3):250-259. doi:10.1001/jamapsychiatry.2022.4350
308. Wang J, Hu X, Li Y, et al. Impaired lipid homeostasis and elevated lipid oxidation of erythrocyte membrane in adolescent depression. *Redox Biology*. 2025;80:103491. doi:10.1016/j.redox.2025.103491
309. MacDonald K, Krishnan A, Cervenka E, Hu G, Guadagno E, Trakadis Y. Biomarkers for major depressive and bipolar disorders using metabolomics: A systematic review. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*. 2019;180(2):122-137. doi:10.1002/ajmg.b.32680

310. Zacharias HU, Hertel J, Johar H, et al. A metabolome-wide association study in the general population reveals decreased levels of serum laurycarnitine in people with depression. *Mol Psychiatry*. Published online June 4, 2021:1-12. doi:10.1038/s41380-021-01176-0
311. Jansen R, Milaneschi Y, Schraner D, et al. The metabolome-wide signature of major depressive disorder. *Mol Psychiatry*. 2024;29(12):3722-3733. doi:10.1038/s41380-024-02613-6
312. Kawamura N, Shinoda K, Sato H, et al. Plasma metabolome analysis of patients with major depressive disorder. *Psychiatry and Clinical Neurosciences*. 2018;72(5):349-361. doi:10.1111/pcn.12638
313. Huang T, Balasubramanian R, Yao Y, et al. Associations of depression status with plasma levels of candidate lipid and amino acid metabolites: a meta-analysis of individual data from three independent samples of US postmenopausal women. *Mol Psychiatry*. 2021;26(7):3315-3327. doi:10.1038/s41380-020-00870-9
314. Gu X, Ke S, Wang Q, et al. Energy metabolism in major depressive disorder: Recent advances from omics technologies and imaging. *Biomedicine & Pharmacotherapy*. 2021;141:111869. doi:10.1016/j.biopha.2021.111869
315. Pu J, Liu Y, Wu H, et al. Characterizing metabolomic and proteomic changes in depression: a systematic analysis. *Mol Psychiatry*. Published online February 15, 2025:1-9. doi:10.1038/s41380-025-02919-z
316. Bot M, Milaneschi Y, Al-Shehri T, et al. Metabolomics Profile in Depression: A Pooled Analysis of 230 Metabolic Markers in 5283 Cases With Depression and 10,145 Controls. *Biol Psychiatry*. 2020;87(5):409-418. doi:10.1016/j.biopsych.2019.08.016
317. de Kluiver H, Jansen R, Penninx BWJH, Giltay EJ, Schoevers RA, Milaneschi Y. Metabolomics signatures of depression: the role of symptom profiles. *Transl Psychiatry*. 2023;13(1):1-10. doi:10.1038/s41398-023-02484-5
318. Sato S, Yu Z, Sakai M, et al. Decreased β -hydroxybutyrate and ketogenic amino acid levels in depressed human adults. *European Journal of Neuroscience*. 2023;57(6):1018-1032. doi:10.1111/ejn.15931
319. Zhang S, Wang Z, Wang Y, et al. A metabolomic profile of biological aging in 250,341 individuals from the UK Biobank. *Nat Commun*. 2024;15(1):8081. doi:10.1038/s41467-024-52310-9
320. Liu L, Li Y, Tollefsbol TO. Gene-Environment Interactions and Epigenetic Basis of Human Diseases. *Current Issues in Molecular Biology*. 2008;10(1-2):25-36. doi:10.21775/cimb.010.025

321. Petronis A. Epigenetics as a unifying principle in the aetiology of complex traits and diseases. *Nature*. 2010;465(7299):721-727. doi:10.1038/nature09230
322. Lisoway AJ, Zai CC, Tiwari AK, Kennedy JL. DNA methylation and clinical response to antidepressant medication in major depressive disorder: A review and recommendations. *Neuroscience Letters*. 2018;669:14-23. doi:10.1016/j.neulet.2016.12.071
323. Webb LM, Phillips KE, Ho MC, Veldic M, Blacker CJ. The Relationship between DNA Methylation and Antidepressant Medications: A Systematic Review. *International Journal of Molecular Sciences*. 2020;21(3):826. doi:10.3390/ijms21030826
324. Bollepalli S, Korhonen T, Kaprio J, Anders S, Ollikainen M. Epismoker: A Robust Classifier to Determine Smoking Status from DNA Methylation Data. *Epigenomics*. 2019;11(13):1469-1486. doi:10.2217/epi-2019-0206
325. Matías-García PR, Ward-Caviness CK, Raffield LM, et al. DNAm-based signatures of accelerated aging and mortality in blood are associated with low renal function. *Clinical Epigenetics*. 2021;13(1):121. doi:10.1186/s13148-021-01082-w
326. Sun D, Zhang T, Su S, et al. Body Mass Index Drives Changes in DNA Methylation. *Circulation Research*. 2019;125(9):824-833. doi:10.1161/CIRCRESAHA.119.315397
327. Liu C, Marioni RE, Hedman ÅK, et al. A DNA methylation biomarker of alcohol consumption. *Mol Psychiatry*. 2018;23(2):422-433. doi:10.1038/mp.2016.192
328. Tsankova N, Renthal W, Kumar A, Nestler EJ. Epigenetic regulation in psychiatric disorders. *Nat Rev Neurosci*. 2007;8(5):355-367. doi:10.1038/nrn2132
329. Klengel T, Pape J, Binder EB, Mehta D. The role of DNA methylation in stress-related psychiatric disorders. *Neuropharmacology*. 2014;80:115-132. doi:10.1016/j.neuropharm.2014.01.013
330. Moran S, Arribas C, Esteller M. Validation of a DNA Methylation Microarray for 850,000 CpG Sites of the Human Genome Enriched in Enhancer Sequences. *Epigenomics*. 2016;8(3):389-399. doi:10.2217/epi.15.114
331. Campagna MP, Xavier A, Lechner-Scott J, et al. Epigenome-wide association studies: current knowledge, strategies and recommendations. *Clinical Epigenetics*. 2021;13(1):214. doi:10.1186/s13148-021-01200-8
332. Du P, Zhang X, Huang CC, et al. Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics*. 2010;11(1):587. doi:10.1186/1471-2105-11-587

333. Aberg KA, Chan RF, Shabalín AA, et al. A MBD-seq protocol for large-scale methylome-wide studies with (very) low amounts of DNA. *Epigenetics*. 2017;12(9):743-750. doi:10.1080/15592294.2017.1335849
334. Aberg KA, McClay JL, Nerella S, et al. MBD-seq as a cost-effective approach for methylome-wide association studies: demonstration in 1500 case-control samples. *Epigenomics*. 2012;4(6):605-621. doi:10.2217/epi.12.59
335. Shen X, Barbu M, Caramaschi D, et al. A methylome-wide association study of major depression with out-of-sample case-control classification and trans-ancestry comparison. Published online October 28, 2023:2023.10.27.23297630. doi:10.1101/2023.10.27.23297630
336. Shen X, Caramaschi D, Adams MJ, et al. *DNA Methylome-Wide Association Study of Genetic Risk for Depression Implicates Antigen Processing and Immune Responses.*; 2021:2021.06.30.21259731. doi:10.1101/2021.06.30.21259731
337. Starnawska A, Tan Q, Soerensen M, et al. Epigenome-wide association study of depression symptomatology in elderly monozygotic twins. *Transl Psychiatry*. 2019;9(1):1-14. doi:10.1038/s41398-019-0548-9
338. Hüls A, Robins C, Conneely KN, et al. Association between DNA methylation levels in brain tissue and late-life depression in community-based participants. *Transl Psychiatry*. 2020;10(1):1-10. doi:10.1038/s41398-020-00948-6
339. Story Jovanova O, Nedeljkovic I, Spieler D, et al. DNA Methylation Signatures of Depressive Symptoms in Middle-aged and Elderly Persons: Meta-analysis of Multiethnic Epigenome-wide Studies. *JAMA Psychiatry*. 2018;75(9):949-959. doi:10.1001/jamapsychiatry.2018.1725
340. Zimmermann N, Zschocke J, Perisic T, Yu S, Holsboer F, Rein T. Antidepressants inhibit DNA methyltransferase 1 through reducing G9a levels. *Biochemical Journal*. 2012;448(1):93-102. doi:10.1042/BJ20120674
341. Le François B, Soo J, Millar AM, et al. Chronic mild stress and antidepressant treatment alter 5-HT1A receptor expression by modifying DNA methylation of a conserved Sp4 site. *Neurobiology of Disease*. 2015;82:332-341. doi:10.1016/j.nbd.2015.07.002
342. Sales AJ, Biojone C, Terceti MS, Guimarães FS, Gomes MV, Joca SR. Antidepressant-like effect induced by systemic and intra-hippocampal administration of DNA methylation inhibitors. *Br J Pharmacol*. 2011;164(6):1711-1721. doi:10.1111/j.1476-5381.2011.01489.x
343. Barbu MC, Huider F, Campbell A, et al. Methylome-wide association study of antidepressant use in Generation Scotland and the Netherlands Twin Register

- implicates the innate immune system. *Mol Psychiatry*. 2022;27(3):1647-1657. doi:10.1038/s41380-021-01412-7
344. Sudlow C, Gallacher J, Allen N, et al. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLOS Medicine*. 2015;12(3):e1001779. doi:10.1371/journal.pmed.1001779
 345. Smith BH, Campbell H, Blackwood D, et al. Generation Scotland: the Scottish Family Health Study; a new resource for researching genes and heritability. *BMC Med Genet*. 2006;7:74. doi:10.1186/1471-2350-7-74
 346. Bycroft C, Freeman C, Petkova D, et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature*. 2018;562(7726):203-209. doi:10.1038/s41586-018-0579-z
 347. McCarthy S, Das S, Kretzschmar W, et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet*. 2016;48(10):1279-1283. doi:10.1038/ng.3643
 348. Huang J, Howie B, McCarthy S, et al. Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. *Nat Commun*. 2015;6(1):8111. doi:10.1038/ncomms9111
 349. Illumina. DRAGEN Bio-IT Platform 3.7 User Guide. https://support.illumina.com/content/dam/illumina-support/documents/documentation/software_documentation/dragen-bio-it/IlluminaDRAGEN-Bio-IT-Platform-User-Guide-1000000141465-00.pdf
 350. Nightingale Health, UK Biobank. Nightingale Health Metabolic Biomarkers. Accessed March 6, 2025. https://biobank.ndph.ox.ac.uk/showcase/ukb/docs/NMR_companion_phase2.pdf
 351. Julkunen H, Cichońska A, Tiainen M, et al. Atlas of plasma NMR biomarkers for health and disease in 118,461 individuals from the UK Biobank. *Nat Commun*. 2023;14(1):604. doi:10.1038/s41467-023-36231-7
 352. Kessler RC, Üstün TB. The World Mental Health (WMH) Survey Initiative version of the World Health Organization (WHO) Composite International Diagnostic Interview (CIDI). *Int J Methods Psychiatr Res*. 2006;13(2):93-121. doi:10.1002/mpr.168
 353. Kessler RC, Andrews G, Mroczek D, Ustun B, Wittchen HU. The World Health Organization Composite International Diagnostic Interview short-form (CIDI-SF). *International Journal of Methods in Psychiatric Research*. 1998;7(4):171-185. doi:10.1002/mpr.47
 354. Gigantesco A, Morosini P. Development, reliability and factor analysis of a selfadministered questionnaire which originates from the World Health

- Organization's Composite International Diagnostic Interview – Short Form (CIDI-SF) for assessing mental disorders. *Clin Pract Epidemiol Ment Health*. 2008;4:8. doi:10.1186/17450179-4-8
355. Davis KAS, Coleman JRI, Adams M, et al. Mental health in UK Biobank – development, implementation and results from an online questionnaire completed by 157 366 participants: a reanalysis. *BJPsych Open*. 2020;6(2):e18. doi:10.1192/bjo.2019.100
 356. Townsend P. Deprivation. *Journal of Social Policy*. 1987;16(2):125-146. doi:10.1017/S0047279400020341
 357. Min JL, Hemani G, Davey Smith G, Relton C, Suderman M. Meffil: efficient normalization and analysis of very large DNA methylation datasets. *Bioinformatics*. 2018;34(23):3983-3989. doi:10.1093/bioinformatics/bty476
 358. Pidsley R, Y Wong CC, Volta M, Lunnon K, Mill J, Schalkwyk LC. A data-driven approach to preprocessing Illumina 450K methylation array data. *BMC Genomics*. 2013;14(1):293. doi:10.1186/1471-2164-14-293
 359. Fortin JP, Fertig E, Hansen K. shinyMethyl: interactive quality control of Illumina 450k DNA methylation arrays in R. Published online September 19, 2014. doi:10.12688/f1000research.4680.2
 360. McCartney DL, Walker RM, Morris SW, McIntosh AM, Porteous DJ, Evans KL. Identification of polymorphic and off-target probe binding sites on the Illumina Infinium MethylationEPIC BeadChip. *Genom Data*. 2016;9:22-24. doi:10.1016/j.gdata.2016.05.012
 361. Zhou W, Laird PW, Shen H. Comprehensive characterization, annotation and innovative use of Infinium DNA methylation BeadChip probes. *Nucleic Acids Res*. 2017;45(4):e22. doi:10.1093/nar/gkw967
 362. Aryee MJ, Jaffe AE, Corrada-Bravo H, et al. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics*. 2014;30(10):1363-1369. doi:10.1093/bioinformatics/btu049
 363. Reinius LE, Acevedo N, Joerink M, et al. Differential DNA Methylation in Purified Human Blood Cells: Implications for Cell Lineage and Studies on Disease Susceptibility. *PLoS One*. 2012;7(7):e41361. doi:10.1371/journal.pone.0041361
 364. Du P, Kibbe WA, Lin SM. lumi: a pipeline for processing Illumina microarray. *Bioinformatics*. 2008;24(13):1547-1548. doi:10.1093/bioinformatics/btn224
 365. Zhang F, Chen W, Zhu Z, et al. OSCA: a tool for omic-data-based complex trait analysis. *Genome Biology*. 2019;20(1):107. doi:10.1186/s13059-019-1718-z

366. Kerr SM, Campbell A, Marten J, et al. Electronic health record and genome-wide genetic data in Generation Scotland participants. *Wellcome Open Res.* 2017;2:85. doi:10.12688/wellcomeopenres.12600.1
367. Alvarez-Madrazo S, McTaggart S, Nangle C, Nicholson E, Bennie M. Data Resource Profile: The Scottish National Prescribing Information System (PIS). *International Journal of Epidemiology.* 2016;45(3):714-715f. doi:10.1093/ije/dyw060
368. Spitzer RL, Williams JBW, Gibbon M, First MB. The Structured Clinical Interview for DSM-III-R (SCID): I: History, Rationale, and Description. *Archives of General Psychiatry.* 1992;49(8):624-629. doi:10.1001/archpsyc.1992.01820080032005
369. Navrady LB, Wolters MK, MacIntyre DJ, et al. Cohort Profile: Stratifying Resilience and Depression Longitudinally (STRADL): a questionnaire follow-up of Generation Scotland: Scottish Family Health Study (GS:SFHS). *Int J Epidemiol.* 2018;47(1):13-14g. doi:10.1093/ije/dyx115
370. The GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science.* 2020;369(6509):1318-1330. doi:10.1126/science.aaz1776
371. The GTEx Consortium, Ardlie KG, Deluca DS, et al. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science.* 2015;348(6235):648-660. doi:10.1126/science.1262110
372. Aguet F, Brown AA, Castel SE, et al. Genetic effects on gene expression across human tissues. *Nature.* 2017;550(7675):204-213. doi:10.1038/nature24277
373. Karczewski KJ, Francioli LC, Tiao G, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature.* 2020;581(7809):434-443. doi:10.1038/s41586-020-2308-7
374. Gudmundsson S, Singer-Berk M, Watts NA, et al. Variant interpretation using population databases: Lessons from gnomAD. *Human Mutation.* 2022;43(8):10121030. doi:10.1002/humu.24309
375. Uffelmann E, Huang QQ, Munung NS, et al. Genome-wide association studies. *Nat Rev Methods Primers.* 2021;1(1):1-21. doi:10.1038/s43586-021-00056-9
376. Bulik-Sullivan BK, Loh PR, Finucane HK, et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet.* 2015;47(3):291-295. doi:10.1038/ng.3211
377. Bulik-Sullivan B, Finucane HK, Anttila V, et al. An atlas of genetic correlations across human diseases and traits. *Nat Genet.* 2015;47(11):1236-1241. doi:10.1038/ng.3406

378. Richmond RC, Davey Smith G. Mendelian Randomization: Concepts and Scope. *Cold Spring Harb Perspect Med*. 2022;12(1):a040501. doi:10.1101/cshperspect.a040501
379. Sanderson E, Glymour MM, Holmes MV, et al. Mendelian randomization. *Nat Rev Methods Primers*. 2022;2(1):1-21. doi:10.1038/s43586-021-00092-5
380. Davey Smith G, Ebrahim S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease?*. *International Journal of Epidemiology*. 2003;32(1):1-22. doi:10.1093/ije/dyg070
381. Davey Smith G, Holmes MV, Davies NM, Ebrahim S. Mendel's laws, Mendelian randomization and causal inference in observational data: substantive and nomenclatural issues. *Eur J Epidemiol*. 2020;35(2):99-111. doi:10.1007/s10654020-00622-7
382. Bowden J, Davey Smith G, Burgess S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *International Journal of Epidemiology*. 2015;44(2):512-525. doi:10.1093/ije/dyv080
383. Carter AR, Sanderson E, Hammerton G, et al. Mendelian randomisation for mediation analysis: current methods and challenges for implementation. *Eur J Epidemiol*. 2021;36(5):465-478. doi:10.1007/s10654-021-00757-1
384. Burgess S, Butterworth A, Thompson SG. Mendelian Randomization Analysis With Multiple Genetic Variants Using Summarized Data. *Genetic Epidemiology*. 2013;37(7):658-665. doi:10.1002/gepi.21758
385. Bowden J, Davey Smith G, Haycock PC, Burgess S. Consistent Estimation in Mendelian Randomization with Some Invalid Instruments Using a Weighted Median Estimator. *Genetic Epidemiology*. 2016;40(4):304-314. doi:10.1002/gepi.21965
386. Zuber V, Grinberg NF, Gill D, et al. Combining evidence from Mendelian randomization and colocalization: Review and comparison of approaches. *The American Journal of Human Genetics*. Published online April 2022:S0002929722001495. doi:10.1016/j.ajhg.2022.04.001
387. Giambartolomei C, Vukcevic D, Schadt EE, et al. Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. *PLOS Genetics*. 2014;10(5):e1004383. doi:10.1371/journal.pgen.1004383
388. McLaren W, Gil L, Hunt SE, et al. The Ensembl Variant Effect Predictor. *Genome Biology*. 2016;17(1):122. doi:10.1186/s13059-016-0974-4
389. Harrison PW, Amode MR, Austine-Orimoloye O, et al. Ensembl 2024. *Nucleic Acids Research*. 2024;52(D1):D891-D899. doi:10.1093/nar/gkad1049

390. Harrow J, Denoeud F, Frankish A, et al. GENCODE: producing a reference annotation for ENCODE. *Genome Biol.* 2006;7(Suppl 1):S4. doi:10.1186/gb-20067-s1-s4
391. Mudge JM, Carbonell-Sala S, Diekhans M, et al. GENCODE 2025: reference gene annotation for human and mouse. *Nucleic Acids Research.* 2025;53(D1):D966D975. doi:10.1093/nar/gkae1078
392. Eilbeck K, Lewis SE, Mungall CJ, et al. The Sequence Ontology: a tool for the unification of genome annotations. *Genome Biology.* 2005;6(5):R44. doi:10.1186/gb-2005-6-5-r44
393. MacArthur DG, Tyler-Smith C. Loss-of-function variants in the genomes of healthy humans. *Human Molecular Genetics.* 2010;19(R2):R125-R130. doi:10.1093/hmg/ddq365
394. Chong JX, Buckingham KJ, Jhangiani SN, et al. The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities. *The American Journal of Human Genetics.* 2015;97(2):199-215. doi:10.1016/j.ajhg.2015.06.009
395. MacArthur DG, Balasubramanian S, Frankish A, et al. A Systematic Survey of Loss-of-Function Variants in Human Protein-Coding Genes. *Science.* 2012;335(6070):823-828. doi:10.1126/science.1215040
396. Lek M, Karczewski KJ, Minikel EV, et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature.* 2016;536(7616):285-291. doi:10.1038/nature19057
397. Chang YF, Imam JS, Wilkinson MF. The Nonsense-Mediated Decay RNA Surveillance Pathway. *Annual Review of Biochemistry.* 2007;76(Volume 76, 2007):51-74. doi:10.1146/annurev.biochem.76.050106.093909
398. Yeo G, Burge CB. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J Comput Biol.* 2004;11(2-3):377-394. doi:10.1089/1066527041410418
399. Ioannidis NM, Rothstein JH, Pejaver V, et al. REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. *Am J Hum Genet.* 2016;99(4):877-885. doi:10.1016/j.ajhg.2016.08.016
400. Li B, Krishnan VG, Mort ME, et al. Automated inference of molecular mechanisms of disease from amino acid substitutions. *Bioinformatics.* 2009;25(21):2744-2750. doi:10.1093/bioinformatics/btp528
401. Shihab HA, Gough J, Cooper DN, et al. Predicting the Functional, Molecular, and Phenotypic Consequences of Amino Acid Substitutions using Hidden Markov Models. *Human Mutation.* 2013;34(1):57-65. doi:10.1002/humu.22225

402. Carter H, Douville C, Stenson PD, Cooper DN, Karchin R. Identifying Mendelian disease genes with the Variant Effect Scoring Tool. *BMC Genomics*. 2013;14(Suppl 3):S3. doi:10.1186/1471-2164-14-S3-S3
403. Adzhubei I, Jordan DM, Sunyaev SR. Predicting Functional Effect of Human Missense Mutations Using PolyPhen-2. *Curr Protoc Hum Genet*. 2013;0 7:Unit7.20. doi:10.1002/0471142905.hg0720s76
404. Choi Y, Sims GE, Murphy S, Miller JR, Chan AP. Predicting the Functional Effect of Amino Acid Substitutions and Indels. *PLOS ONE*. 2012;7(10):e46688. doi:10.1371/journal.pone.0046688
405. Reva B, Antipin Y, Sander C. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res*. 2011;39(17):e118. doi:10.1093/nar/gkr407
406. Schwarz JM, Rödelberger C, Schuelke M, Seelow D. MutationTaster evaluates disease-causing potential of sequence alterations. *Nat Methods*. 2010;7(8):575576. doi:10.1038/nmeth0810-575
407. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc*. 2009;4(7):10731081. doi:10.1038/nprot.2009.86
408. Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. Identifying a High Fraction of the Human Genome to be under Selective Constraint Using GERP++. *PLOS Computational Biology*. 2010;6(12):e1001025. doi:10.1371/journal.pcbi.1001025
409. Garber M, Guttman M, Clamp M, Zody MC, Friedman N, Xie X. Identifying novel constrained elements by exploiting biased substitution patterns. *Bioinformatics*. 2009;25(12):i54-i62. doi:10.1093/bioinformatics/btp190
410. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res*. 2010;20(1):110-121. doi:10.1101/gr.097857.109
411. Siepel A, Bejerano G, Pedersen JS, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res*. 2005;15(8):1034-1050. doi:10.1101/gr.3715005
412. Kircher M, Witten DM, Jain P, O’Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet*. 2014;46(3):310-315. doi:10.1038/ng.2892

413. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Research*. 2019;47(D1):D886-D894. doi:10.1093/nar/gky1016
414. Schubach M, Maass T, Nazaretyan L, Röner S, Kircher M. CADD v1.7: using protein language models, regulatory CNNs and other nucleotide-level scores to improve genome-wide variant predictions. *Nucleic Acids Research*. 2024;52(D1):D1143-D1154. doi:10.1093/nar/gkad989
415. Lee S, Abecasis GR, Boehnke M, Lin X. Rare-Variant Association Analysis: Study Designs and Statistical Tests. *American Journal of Human Genetics*. 2014;95(1):5. doi:10.1016/j.ajhg.2014.06.009
416. Mbatchou J, Barnard L, Backman J, et al. Computationally efficient whole-genome regression for quantitative and binary traits. *Nat Genet*. 2021;53(7):1097-1103. doi:10.1038/s41588-021-00870-7
417. Wu B, Guan W, Pankow JS. On efficient and accurate calculation of significance p-values for sequence kernel association testing of variant set. *Annals of human genetics*. 2016;80(2):123. doi:10.1111/ahg.12144
418. Davies RB. The Distribution of a Linear Combination of χ^2 Random Variables. *Journal of the Royal Statistical Society Series C: Applied Statistics*. 1980;29(3):323-333. doi:10.2307/2346911
419. Kuonen D. Miscellanea. Saddlepoint approximations for distributions of quadratic forms in normal variables. *Biometrika*. 1999;86(4):929-935. doi:10.1093/biomet/86.4.929
420. Lee S, Wu MC, Lin X. Optimal tests for rare variant effects in sequencing association studies. *Biostatistics*. 2012;13(4):762-775. doi:10.1093/biostatistics/kxs014
421. Bick AG, Metcalf GA, Mayo KR, et al. Genomic data in the All of Us Research Program. *Nature*. Published online February 19, 2024:1-7. doi:10.1038/s41586023-06957-x
422. Liu Y, Chen S, Li Z, Morrison AC, Boerwinkle E, Lin X. ACAT: A Fast and Powerful p Value Combination Method for Rare-Variant Analysis in Sequencing Studies. *The American Journal of Human Genetics*. 2019;104(3):410-421. doi:10.1016/j.ajhg.2019.01.002
423. Ziyatdinov A, Mbatchou J, Marcketta A, et al. Joint testing of rare variant burden scores using non-negative least squares. *The American Journal of Human Genetics*. 2024;111(10):2139-2149. doi:10.1016/j.ajhg.2024.08.021
424. Hibbeln JR. Fish consumption and major depression. *The Lancet*.

1998;351(9110):1213. doi:10.1016/S0140-6736(05)79168-6

425. Grosso G, Micek A, Marventano S, et al. Dietary n-3 PUFA, fish consumption and depression: A systematic review and meta-analysis of observational studies. *Journal of Affective Disorders*. 2016;205:269-281. doi:10.1016/j.jad.2016.08.011
426. Okereke OI, Vyas CM, Mischoulon D, et al. Effect of Long-term Supplementation With Marine Omega-3 Fatty Acids vs Placebo on Risk of Depression or Clinically Relevant Depressive Symptoms and on Change in Mood Scores: A Randomized Clinical Trial. *JAMA*. 2021;326(23):2385-2394. doi:10.1001/jama.2021.21187
427. Lc R, Em B, Db H, H S, Ar P. A pilot randomized controlled trial testing supplements of omega-3 fatty acids, probiotics, combination or placebo on symptoms of depression, anxiety and stress. *Journal of Affective Disorders Reports*. 2021;5:100141. doi:10.1016/j.jadr.2021.100141
428. Bot M, Brouwer IA, Roca M, et al. Effect of Multinutrient Supplementation and Food-Related Behavioral Activation Therapy on Prevention of Major Depressive Disorder Among Overweight or Obese Adults With Subsyndromal Depressive Symptoms: The MoodFOOD Randomized Clinical Trial. *JAMA*. 2019;321(9):858868. doi:10.1001/jama.2019.0556
429. Liao Y, Xie B, Zhang H, et al. Efficacy of omega-3 PUFAs in depression: A metaanalysis. *Transl Psychiatry*. 2019;9(1):1-9. doi:10.1038/s41398-019-0515-5
430. Mocking RJT, Harmsen I, Assies J, Koeter MWJ, Ruhé HG, Schene AH. Metaanalysis and meta-regression of omega-3 polyunsaturated fatty acid supplementation for major depressive disorder. *Transl Psychiatry*. 2016;6(3):e756e756. doi:10.1038/tp.2016.29
431. Deane KHO, Jimoh OF, Biswas P, et al. Omega-3 and polyunsaturated fat for prevention of depression and anxiety symptoms: systematic review and metaanalysis of randomised trials. *The British Journal of Psychiatry*. 2021;218(3):135142. doi:10.1192/bjp.2019.234
432. Gallois A, Mefford J, Ko A, et al. A comprehensive study of metabolite genetics reveals strong pleiotropy and heterogeneity across time and context. *Nat Commun*. 2019;10(1):4788. doi:10.1038/s41467-019-12703-7
433. Hagenbeek FA, Pool R, van Dongen J, et al. Heritability estimates for 361 blood metabolites across 40 genome-wide association studies. *Nat Commun*. 2020;11(1):39. doi:10.1038/s41467-019-13770-6
434. Kettunen J, Demirkan A, Würtz P, et al. Genome-wide study for circulating metabolites identifies 62 loci and reveals novel systemic effects of LPA. *Nat Commun*. 2016;7(1):11122. doi:10.1038/ncomms11122

435. Shin SY, Fauman EB, Petersen AK, et al. An atlas of genetic influences on human blood metabolites. *Nat Genet.* 2014;46(6):543-550. doi:10.1038/ng.2982
436. Yin X, Chan LS, Bose D, et al. Genome-wide association studies of metabolites in Finnish men identify disease-relevant loci. *Nat Commun.* 2022;13(1):1644. doi:10.1038/s41467-022-29143-5
437. Tambets R, Kronberg J, Abner E, et al. Genome-wide association study for circulating metabolites in 619,372 individuals. Published online October 16, 2024:2024.10.15.24315557. doi:10.1101/2024.10.15.24315557
438. Smith CJ, Sinnott-Armstrong N, Cichońska A, et al. Integrative analysis of metabolite GWAS illuminates the molecular basis of pleiotropy and genetic correlation. Published online April 4, 2022:2022.04.02.486791. doi:10.1101/2022.04.02.486791
439. Gao Y, Hu X, Wang D, et al. Association between Arachidonic Acid and the Risk of Schizophrenia: A Cross-National Study and Mendelian Randomization Analysis. *Nutrients.* 2023;15(5):1195. doi:10.3390/nu15051195
440. Hemani G, Zheng J, Elsworth B, et al. The MR-Base platform supports systematic causal inference across the human phenome. Loos R, ed. *eLife.* 2018;7:e34408. doi:10.7554/eLife.34408
441. Davyson E, Shen X, Gadd DA, et al. Metabolomic Investigation of Major Depressive Disorder Identifies a Potentially Causal Association With Polyunsaturated Fatty Acids. *Biological Psychiatry.* 2023;0(0). doi:10.1016/j.biopsych.2023.01.027
442. Gilchrist L, Mutz J, Hysi P, et al. Evaluating metabolome-wide causal effects on risk for psychiatric and neurodegenerative disorders. Published online November 7, 2024:2024.11.06.24316481. doi:10.1101/2024.11.06.24316481
443. Stacey D, Benyamin B, Lee SH, Hyppönen E. A Metabolome-Wide Mendelian Randomization Study Identifies Dysregulated Arachidonic Acid Synthesis as a Potential Causal Risk Factor for Bipolar Disorder. *Biological Psychiatry.* 2024;96(6):455-462. doi:10.1016/j.biopsych.2024.02.1005
444. DeLong LN, Fleetwood K, Prigge R, Galdi P, Guthrie B, Fleuriot JD. Investigating associations between physical multimorbidity clusters and subsequent depression: cluster and survival analysis of UK Biobank data. Published online July 7, 2024:2024.07.05.24310004. doi:10.1101/2024.07.05.24310004
445. Borges MC, Haycock PC, Zheng J, et al. Role of circulating polyunsaturated fatty acids on cardiovascular diseases risk: analysis using Mendelian randomization and fatty acid genetic association data from over 114,000 UK Biobank participants. *BMC Medicine.* 2022;20(1):210. doi:10.1186/s12916-022-02399-w

446. Carnegie R, Borges MC, Jones HJ, et al. Omega-3 fatty acids and major depression: a Mendelian randomization study. *Transl Psychiatry*. 2024;14(1):1-10. doi:10.1038/s41398-024-02932-w
447. Burgess S, Freitag DF, Khan H, Gorman DN, Thompson SG. Using Multivariable Mendelian Randomization to Disentangle the Causal Effects of Lipid Fractions. *PLOS ONE*. 2014;9(10):e108891. doi:10.1371/journal.pone.0108891
448. Ye K, Gao F, Wang D, Bar-Yosef O, Keinan A. Dietary adaptation of FADS genes in Europe varied across time and geography. *Nat Ecol Evol*. 2017;1(7):1-11. doi:10.1038/s41559-017-0167
449. Spence JP, Mostafavi H, Ota M, et al. Specificity, length, and luck: How genes are prioritized by rare and common variant association studies. Published online December 16, 2024:2024.12.12.628073. doi:10.1101/2024.12.12.628073
450. Goswami C, Chattopadhyay A, Chuang EY. Rare variants: data types and analysis strategies. *Ann Transl Med*. 2021;9(12):961. doi:10.21037/atm-21-1635
451. Wainschein P, Jain D, Zheng Z, et al. Assessing the contribution of rare variants to complex trait heritability from whole-genome sequence data. *Nat Genet*. 2022;54(3):263-273. doi:10.1038/s41588-021-00997-7
452. Wang Q, Dhindsa RS, Carss K, et al. Rare variant contribution to human disease in 281,104 UK Biobank exomes. *Nature*. 2021;597(7877):527-532. doi:10.1038/s41586-021-03855-y
453. Momozawa Y, Mizukami K. Unique roles of rare variants in the genetics of complex diseases in humans. *J Hum Genet*. 2021;66(1):11-23. doi:10.1038/s10038-02000845-2
454. Cohen JC, Kiss RS, Pertsemlidis A, Marcel YL, McPherson R, Hobbs HH. Multiple Rare Alleles Contribute to Low Plasma Levels of HDL Cholesterol. *Science*. 2004;305(5685):869-872. doi:10.1126/science.1099870
455. Cirulli ET, Goldstein DB. Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nat Rev Genet*. 2010;11(6):415-425. doi:10.1038/nrg2779
456. Dyer SC, Austine-Orimoloye O, Azov AG, et al. Ensembl 2025. *Nucleic Acids Research*. 2025;53(D1):D948-D957. doi:10.1093/nar/gkae1071
457. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*. 2011;27(21):2987-2993. doi:10.1093/bioinformatics/btr509
458. Danecek P, Auton A, Abecasis G, et al. The variant call format and VCFtools. *Bioinformatics*. 2011;27(15):2156-2158. doi:10.1093/bioinformatics/btr330

459. Karczewski KJ, Francioli LC, Tiao G, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*. 2020;581(7809):434-443. doi:10.1038/s41586-020-2308-7
460. Purcell S, Neale B, Todd-Brown K, et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics*. 2007;81(3):559-575. doi:10.1086/519795
461. Elliott P, Peakman TC, on behalf of UK Biobank. The UK Biobank sample handling and storage protocol for the collection, processing and archiving of human blood and urine. *International Journal of Epidemiology*. 2008;37(2):234-244. doi:10.1093/ije/dym276
462. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B (Methodological)*. 1995;57(1):289-300.
463. Heinze G, Schemper M. A solution to the problem of separation in logistic regression. *Statistics in Medicine*. 2002;21(16):2409-2419. doi:10.1002/sim.1047
464. FIRTH D. Bias reduction of maximum likelihood estimates. *Biometrika*. 1993;80(1):27-38. doi:10.1093/biomet/80.1.27
465. Lattka E, Illig T, Koletzko B, Heinrich J. Genetic variants of the FADS1 FADS2 gene cluster as related to essential fatty acid metabolism. *Current Opinion in Lipidology*. 2010;21(1):64-69. doi:10.1097/MOL.0b013e3283327ca8
466. Glaser C, Heinrich J, Koletzko B. Role of FADS1 and FADS2 polymorphisms in polyunsaturated fatty acid metabolism. *Metabolism*. 2010;59(7):993-999. doi:10.1016/j.metabol.2009.10.022
467. Mathias RA, Pani V, Chilton FH. Genetic Variants in the FADS Gene: Implications for Dietary Recommendations for Fatty Acid Intake. *Curr Nutr Rep*. 2014;3(2):139148. doi:10.1007/s13668-014-0079-1
468. Zhang JY, Qin X, Liang A, et al. Fads3 modulates docosahexaenoic acid in liver and brain. *Prostaglandins, Leukotrienes and Essential Fatty Acids*. 2017;123:25-32. doi:10.1016/j.plefa.2017.07.001
469. Ye K, Gao F, Wang D, Bar-Yosef O, Keinan A. Dietary adaptation of FADS genes in Europe varied across time and geography. *Nat Ecol Evol*. 2017;1(7):1-11. doi:10.1038/s41559-017-0167
470. Ameur A, Enroth S, Johansson Å, et al. Genetic Adaptation of Fatty-Acid Metabolism: A Human-Specific Haplotype Increasing the Biosynthesis of LongChain Omega-3 and Omega-6 Fatty Acids. *The American Journal of Human Genetics*. 2012;90(5):809-820. doi:10.1016/j.ajhg.2012.03.014

471. Feofanova EV, Brown MR, Alkis T, et al. Whole-Genome Sequencing Analysis of Human Metabolome in Multi-Ethnic Populations. *Nat Commun.* 2023;14(1):3111. doi:10.1038/s41467-023-38800-2
472. Meer D van der, Rahman Z, Ottas A, et al. Pleiotropic and sex-specific genetic architecture of circulating metabolic markers. Published online August 1, 2024:2024.07.30.24311254. doi:10.1101/2024.07.30.24311254
473. Yang C, Veenstra J, Bartz TM, et al. Genome-wide association studies and finemapping identify genomic loci for n-3 and n-6 polyunsaturated fatty acids in Hispanic American and African American cohorts. *Commun Biol.* 2023;6(1):1-15. doi:10.1038/s42003-023-05219-w
474. Reynolds LM, Howard TD, Ruczinski I, et al. Tissue-specific impact of FADS cluster variants on FADS1 and FADS2 gene expression. *PLoS One.* 2018;13(3):e0194610. doi:10.1371/journal.pone.0194610
475. Ihejirika SA, Chiang AH, Singh A, Stephen E, Chen H, Ye K. A multi-level gene-diet interaction analysis of fish oil supplementation and 14 circulating polyunsaturated fatty acids-related traits identifies the FADS and GRP12 loci. Published online December 22, 2024:2024.12.12.24318956. doi:10.1101/2024.12.12.24318956
476. Martinat M, Rossitto M, Di Miceli M, Layé S. Perinatal Dietary Polyunsaturated Fatty Acids in Brain Development, Role in Neurodevelopmental Disorders. *Nutrients.* 2021;13(4):1185. doi:10.3390/nu13041185
477. Monroig Ó, Rotllant J, Sánchez E, Cerdá-Reverter JM, Tocher DR. Expression of long-chain polyunsaturated fatty acid (LC-PUFA) biosynthesis genes during zebrafish *Danio rerio* early embryogenesis. *Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids.* 2009;1791(11):1093-1101. doi:10.1016/j.bbalip.2009.07.002
478. Coti Bertrand P, O'Kusky JR, Innis SM. Maternal Dietary (n-3) Fatty Acid Deficiency Alters Neurogenesis in the Embryonic Rat Brain. *The Journal of Nutrition.* 2006;136(6):1570-1575. doi:10.1093/jn/136.6.1570
479. Martin RE, Bazan NG. Changing Fatty Acid Content of Growth Cone Lipids Prior to Synaptogenesis. *Journal of Neurochemistry.* 1992;59(1):318-325. doi:10.1111/j.1471-4159.1992.tb08906.x
480. Madore C, Leyrolle Q, Morel L, et al. Essential omega-3 fatty acids tune microglial phagocytosis of synaptic elements in the mouse developing brain. *Nat Commun.* 2020;11(1):6133. doi:10.1038/s41467-020-19861-z
481. Cao P, Chen C, Liu A, et al. Early-life inflammation promotes depressive symptoms in adolescence via microglial engulfment of dendritic spines. *Neuron.*

2021;109(16):2573-2589.e9. doi:10.1016/j.neuron.2021.06.012

482. Wang H, He Y, Sun Z, et al. Microglia in depression: an overview of microglia in the pathogenesis and treatment of depression. *J Neuroinflammation*. 2022;19(1):132. doi:10.1186/s12974-022-02492-0
483. Rial D, Lemos C, Pinheiro H, et al. Depression as a Glial-Based Synaptic Dysfunction. *Frontiers in Cellular Neuroscience*. 2016;9. Accessed August 22, 2023. <https://www.frontiersin.org/articles/10.3389/fncel.2015.00521>
484. Duman RS, Aghajanian GK, Sanacora G, Krystal JH. Synaptic plasticity and depression: New insights from stress and rapid-acting antidepressants. *Nat Med*. 2016;22(3):238-249. doi:10.1038/nm.4050
485. Gao YN, Zhang YQ, Wang H, Deng YL, Li NM. A New Player in Depression: MiRNAs as Modulators of Altered Synaptic Plasticity. *International Journal of Molecular Sciences*. 2022;23(9):4555. doi:10.3390/ijms23094555
486. Karlson EW, Boutin NT, Hoffnagle AG, Allen NL. Building the Partners HealthCare Biobank at Partners Personalized Medicine: Informed Consent, Return of Research Results, Recruitment Lessons and Operational Considerations. *Journal of Personalized Medicine*. 2016;6(1):2. doi:10.3390/jpm6010002
487. Jurgens SJ, Wang X, Choi SH, et al. Rare coding variant analysis for human diseases across biobanks and ancestries. *Nat Genet*. Published online August 29, 2024:1-10. doi:10.1038/s41588-024-01894-5
488. Sun KY, Bai X, Chen S, et al. A deep catalog of protein-coding variation in 985,830 individuals. Published online November 2, 2023:2023.05.09.539329. doi:10.1101/2023.05.09.539329
489. The All of Us Research Program Investigators. The “All of Us” Research Program. *New England Journal of Medicine*. 2019;381(7):668-676. doi:10.1056/NEJMSr1809937
490. Landrum MJ, Lee JM, Benson M, et al. ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Research*. 2016;44(D1):D862-D868. doi:10.1093/nar/gkv1222
491. Yourshaw M, Taylor SP, Rao AR, Martín MG, Nelson SF. Rich annotation of DNA sequencing variants by leveraging the Ensembl Variant Effect Predictor with plugins. *Briefings in Bioinformatics*. 2015;16(2):255-264. doi:10.1093/bib/bbu008
492. Li X, Li Z, Zhou H, et al. Dynamic incorporation of multiple in silico functional annotations empowers rare variant association analysis of large whole-genome sequencing studies at scale. *Nat Genet*. 2020;52(9):969-983. doi:10.1038/s41588020-0676-4

493. Schoeler T, Speed D, Porcu E, Pirastu N, Pingault JB, Kutalik Z. Participation bias in the UK Biobank distorts genetic associations and downstream analyses. *Nat Hum Behav.* 2023;7(7):1216-1227. doi:10.1038/s41562-023-01579-9
494. Taliun D, Harris DN, Kessler MD, et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature.* 2021;590(7845):290-299. doi:10.1038/s41586-021-03205-y
495. Gelfman S, Wang Q, McSweeney KM, et al. Annotating pathogenic non-coding variants in genic regions. *Nat Commun.* 2017;8(1):236. doi:10.1038/s41467-01700141-2
496. Lu H, Ma L, Quan C, et al. RegVar: Tissue-specific Prioritization of Non-coding Regulatory Variants. *Genomics Proteomics Bioinformatics.* 2023;21(2):385-395. doi:10.1016/j.gpb.2021.08.011
497. Shen EH, Overly CC, Jones AR. The Allen Human Brain Atlas: Comprehensive gene expression mapping of the human brain. *Trends in Neurosciences.* 2012;35(12):711-714. doi:10.1016/j.tins.2012.09.005
498. Teschendorff AE, Breeze CE, Zheng SC, Beck S. A comparison of referencebased algorithms for correcting cell-type heterogeneity in Epigenome-Wide Association Studies. *BMC Bioinformatics.* 2017;18(1):105. doi:10.1186/s12859017-1511-5
499. Jaffe AE, Irizarry RA. Accounting for cellular heterogeneity is critical in epigenomewide association studies. *Genome Biol.* 2014;15(2):R31. doi:10.1186/gb-2014-152-r31
500. Houseman EA, Accomando WP, Koestler DC, et al. DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics.* 2012;13(1):86. doi:10.1186/1471-2105-13-86
501. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: A Tool for Genome-wide Complex Trait Analysis. *Am J Hum Genet.* 2011;88(1):76-82. doi:10.1016/j.ajhg.2010.11.011
502. Henderson CR. Best Linear Unbiased Estimation and Prediction under a Selection Model. *Biometrics.* 1975;31(2):423-447. doi:10.2307/2529430
503. Phipson B, Maksimovic J, Oshlack A. missMethyl: an R package for analyzing data from Illumina's HumanMethylation450 platform. *Bioinformatics.* 2016;32(2):286-288. doi:10.1093/bioinformatics/btv560
504. Suderman M, Staley JR, French R, Arathimos R, Simpkin A, Tilling K. dmrff: identifying differentially methylated regions efficiently with power and control. Published online December 31, 2018:508556. doi:10.1101/508556

505. Jaffe AE, Murakami P, Lee H, et al. Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. *International Journal of Epidemiology*. 2012;41(1):200-209. doi:10.1093/ije/dyr238
506. Falcon S, Gentleman R. Hypergeometric Testing Used for Gene Set Enrichment Analysis. In: Hahne F, Huber W, Gentleman R, Falcon S, eds. *Bioconductor Case Studies*. Springer; 2008:207-220. doi:10.1007/978-0-387-77240-0_14
507. Watanabe K, Taskesen E, van Bochoven A, Posthuma D. Functional mapping and annotation of genetic associations with FUMA. *Nat Commun*. 2017;8(1):1826. doi:10.1038/s41467-017-01261-5
508. Koopmans F, van Nierop P, Andres-Alonso M, et al. SynGO: An Evidence-Based, Expert-Curated Knowledge Base for the Synapse. *Neuron*. 2019;103(2):217234.e4. doi:10.1016/j.neuron.2019.05.002
509. Horvath S. DNA methylation age of human tissues and cell types. *Genome Biol*. 2013;14(10):R115. doi:10.1186/gb-2013-14-10-r115
510. Bollepalli S, Korhonen T, Kaprio J, Anders S, Ollikainen M. Epismoker: A Robust Classifier to Determine Smoking Status from DNA Methylation Data. *Epigenomics*. 2019;11(13):1469-1486. doi:10.2217/epi-2019-0206
511. Gadd DA, Hillary RF, McCartney DL, et al. Epigenetic scores for the circulating proteome as tools for disease prediction. *eLife*. 2022;11:e71802. doi:10.7554/eLife.71802
512. Rose RJ, Salvatore JE, Aaltonen S, et al. FinnTwin12 Cohort: An Updated Review. *Twin Res Hum Genet*. 2019;22(5):302-311. doi:10.1017/thg.2019.83
513. Kaidesoja M, Aaltonen S, Bogl LH, et al. FinnTwin16: A Longitudinal Study from Age 16 of a Population-Based Finnish Twin Cohort. *Twin Research and Human Genetics*. 2019;22(6):530-539. doi:10.1017/thg.2019.106
514. Jürgens C, Völzke H, Tost F. Study of Health in Pomerania (SHIP-Trend). *Ophthalmologe*. 2014;111(5):443-447. doi:10.1007/s00347-013-2924-9
515. Deary IJ, Gow AJ, Pattie A, Starr JM. Cohort Profile: The Lothian Birth Cohorts of 1921 and 1936. *International Journal of Epidemiology*. 2012;41(6):1576-1584. doi:10.1093/ije/dyr197
516. Kircher T, Wöhr M, Nenadic I, et al. Neurobiology of the major psychoses: a translational perspective on brain structure and function—the FOR2107 consortium. *Eur Arch Psychiatry Clin Neurosci*. 2019;269(8):949-962. doi:10.1007/s00406-0180943-x

517. Ligthart L, Beijsterveldt CEM van, Kevenaar ST, et al. The Netherlands Twin Register: Longitudinal Research Based on Twin and Twin-Family Designs. *Twin Research and Human Genetics*. 2019;22(6):623-636. doi:10.1017/thg.2019.93
518. Golding, Pembrey, Jones, Team TAS. ALSPAC–The Avon Longitudinal Study of Parents and Children. *Paediatric and Perinatal Epidemiology*. 2001;15(1):74-87. doi:10.1046/j.1365-3016.2001.00325.x
519. Hennings JM, Owashi T, Binder EB, et al. Clinical characteristics and treatment outcome in a representative sample of depressed inpatients – Findings from the Munich Antidepressant Response Signature (MARS) project. *Journal of Psychiatric Research*. 2009;43(3):215-229. doi:10.1016/j.jpsychires.2008.05.002
520. Moffitt TE, Team the ERS. Teen-aged mothers in contemporary Britain. *Journal of Child Psychology and Psychiatry*. 2002;43(6):727-742. doi:10.1111/14697610.00082
521. Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods*. 2010;1(2):97-111. doi:10.1002/jrsm.12
522. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Controlled Clinical Trials*. 1986;7(3):177-188. doi:10.1016/0197-2456(86)90046-2
523. Calder PC. Functional Roles of Fatty Acids and Their Effects on Human Health. *Journal of Parenteral and Enteral Nutrition*. 2015;39(1S):18S-32S. doi:10.1177/0148607115595980
524. Saini RK, Keum YS. Omega-3 and omega-6 polyunsaturated fatty acids: Dietary sources, metabolism, and significance — A review. *Life Sciences*. 2018;203:255267. doi:10.1016/j.lfs.2018.04.049
525. Schmitz G, Ecker J. The opposing effects of *n*-3 and *n*-6 fatty acids. *Progress in Lipid Research*. 2008;47(2):147-155. doi:10.1016/j.plipres.2007.12.004
526. Simopoulos AP. The importance of the ratio of omega-6/omega-3 essential fatty acids. *Biomed Pharmacother*. 2002;56(8):365-379. doi:10.1016/s07533322(02)00253-6
527. Blasbalg TL, Hibbeln JR, Ramsden CE, Majchrzak SF, Rawlings RR. Changes in consumption of omega-3 and omega-6 fatty acids in the United States during the 20th century. *Am J Clin Nutr*. 2011;93(5):950-962. doi:10.3945/ajcn.110.006643
528. Central Committee for Medical and Community Program of the American Heart Association. Dietary Fat and Its Relation to Heart Attacks and Strokes. *JAMA*. 1961;175(5):389-391. doi:10.1001/jama.1961.63040050001011

529. Kelaiditis CF, Gibson EL, Dyll SC. Effects of long-chain omega-3 polyunsaturated fatty acids on reducing anxiety and/or depression in adults; A systematic review and meta-analysis of randomised controlled trials. *Prostaglandins, Leukotrienes and Essential Fatty Acids*. 2023;192:102572. doi:10.1016/j.plefa.2023.102572
530. Reynolds LM, Dutta R, Seeds MC, et al. FADS genetic and metabolomic analyses identify the $\Delta 5$ desaturase (FADS1) step as a critical control point in the formation of biologically important lipids. *Sci Rep*. 2020;10(1):15873. doi:10.1038/s41598020-71948-1
531. Chen Y, Lu T, Pettersson-Kymmer U, et al. Genomic atlas of the plasma metabolome prioritizes metabolites implicated in human diseases. *Nat Genet*. 2023;55(1):44-53. doi:10.1038/s41588-022-01270-1
532. Lotta LA, Pietzner M, Stewart ID, et al. A cross-platform approach identifies genetic regulators of human metabolism and health. *Nat Genet*. 2021;53(1):54-64. doi:10.1038/s41588-020-00751-5
533. Dyll SC. Long-chain omega-3 fatty acids and the brain: a review of the independent and shared effects of EPA, DPA and DHA. *Front Aging Neurosci*. 2015;7:52. doi:10.3389/fnagi.2015.00052
534. Williams JA, Batten SE, Harris M, et al. Docosahexaenoic and Eicosapentaenoic Acids Segregate Differently between Raft and Nonraft Domains. *Biophysical Journal*. 2012;103(2):228-237. doi:10.1016/j.bpj.2012.06.016
535. Able JA, Liu Y, Jandacek R, Rider T, Tso P, McNamara RK. Omega-3 fatty acid deficient male rats exhibit abnormal behavioral activation in the forced swim test following chronic fluoxetine treatment: Association with altered 5-HT1A and alpha2A adrenergic receptor expression. *Journal of Psychiatric Research*. 2014;50:42-50. doi:10.1016/j.jpsychires.2013.11.008
536. Cardoso HD, Santos Junior EF dos, de Santana DF, et al. Omega-3 deficiency and neurodegeneration in the substantia nigra: Involvement of increased nitric oxide production and reduced BDNF expression. *Biochimica et Biophysica Acta (BBA) - General Subjects*. 2014;1840(6):1902-1912. doi:10.1016/j.bbagen.2013.12.023
537. Lim SY, Hoshiba J, Salem Jr N. An extraordinary degree of structural specificity is required in neural phospholipids for optimal brain function: n-6 docosapentaenoic acid substitution for docosahexaenoic acid leads to a loss in spatial task performance. *Journal of Neurochemistry*. 2005;95(3):848-857. doi:10.1111/j.14714159.2005.03427.x
538. Nguyen LN, Ma D, Shui G, et al. Mfsd2a is a transporter for the essential omega-3 fatty acid docosahexaenoic acid. *Nature*. 2014;509(7501):503-506. doi:10.1038/nature13241

539. Pathak H, Borchert A, Garaali S, Burkert A, Frieling H. BDNF exon IV promoter methylation and antidepressant action: a complex interplay. *Clin Epigenetics*. 2022;14(1):187. doi:10.1186/s13148-022-01415-3
540. Moon YK, Kim H, Kim S, Lim SW, Kim DK. Influence of antidepressant treatment on SLC6A4 methylation in Korean patients with major depression. *Am J Med Genet B Neuropsychiatr Genet*. 2023;192(1-2):28-37. doi:10.1002/ajmg.b.32921
541. Glessner JT, Li J, Wang D, et al. Copy number variation meta-analysis reveals a novel duplication at 9p24 associated with multiple neurodevelopmental disorders. *Genome Medicine*. 2017;9(1):106. doi:10.1186/s13073-017-0494-1
542. Huckins LM, Dobbyn A, Ruderfer DM, et al. Gene expression imputation across multiple brain regions provides insights into schizophrenia risk. *Nat Genet*. 2019;51(4):659-674. doi:10.1038/s41588-019-0364-4
543. Nazarian A, Yashin AI, Kulminski AM. Methylation-wide association analysis reveals AIM2, DGUOK, GNAI3, and ST14 genes as potential contributors to the Alzheimer's disease pathogenesis. Published online May 15, 2018:322503. doi:10.1101/322503
544. Kronman H, Torres-Berrio A, Sidoli S, et al. Long-term behavioral and cell-typespecific molecular effects of early life stress are mediated by H3K79me2 dynamics in medium spiny neurons. *Nat Neurosci*. 2021;24(5):667-676. doi:10.1038/s41593021-00814-8
545. Kealy L, Runting J, Thiele D, Scheer S. An emerging maestro of immune regulation: how DOT1L orchestrates the harmonies of the immune system. *Front Immunol*. 2024;15. doi:10.3389/fimmu.2024.1385319
546. Park G, Gong Z, Chen J, Kim JE. Characterization of the DOT1L Network: Implications of Diverse Roles for DOT1L. *Protein J*. 2010;29(3):213-223. doi:10.1007/s10930-010-9242-8
547. Statello L, Guo CJ, Chen LL, Huarte M. Gene regulation by long non-coding RNAs and its biological functions. *Nat Rev Mol Cell Biol*. 2021;22(2):96-118. doi:10.1038/s41580-020-00315-9
548. Feng Y, Wu F, Wu Y, Guo Z, Ji X. LncRNA DGUOK-AS1 facilitates non-small cell lung cancer growth and metastasis through increasing TRPM7 stability via m6A modification. *Translational Oncology*. 2023;32:101661. doi:10.1016/j.tranon.2023.101661
549. Liang Y, Ye F, Wang Y, et al. DGUOK-AS1 acts as a tumor promoter through regulating miR-204-5p/IL-11 axis in breast cancer. *Molecular Therapy - Nucleic Acids*. 2021;26:1079-1091. doi:10.1016/j.omtn.2021.10.018

550. Fazel-Najafabadi M, Rallabandi HR, Singh MK, et al. Discovery and Functional Characterization of Two Regulatory Variants Underlying Lupus Susceptibility at 2p13.1. *Genes (Basel)*. 2022;13(6):1016. doi:10.3390/genes13061016
551. Fukunaga T, Iwakiri J, Ono Y, Hamada M. LncRRlsearch: A Web Server for lncRNA-RNA Interaction Prediction Integrated With Tissue-Specific Expression and Subcellular Localization Data. *Frontiers in Genetics*. 2019;10. Accessed August 21, 2023. <https://www.frontiersin.org/articles/10.3389/fgene.2019.00462>
552. Sang L, He YJ, Kang J, et al. Mitochondrial Deoxyguanosine Kinase Regulates NAD⁺ Biogenesis Independent of Mitochondria Complex I Activity. *Front Oncol*. 2020;10:570656. doi:10.3389/fonc.2020.570656
553. Rayan NA, Kumar V, Aow J, et al. Integrative multi-omics landscape of fluoxetine action across 27 brain regions reveals global increase in energy metabolism and region-specific chromatin remodelling. *Mol Psychiatry*. 2022;27(11):4510-4525. doi:10.1038/s41380-022-01725-1
554. Allen J, Romay-Tallon R, Brymer KJ, Caruncho HJ, Kalynchuk LE. Mitochondria and Mood: Mitochondrial Dysfunction as a Key Player in the Manifestation of Depression. *Frontiers in Neuroscience*. 2018;12. Accessed August 22, 2023. <https://www.frontiersin.org/articles/10.3389/fnins.2018.00386>
555. Pei L, Wallace DC. Mitochondrial Etiology of Neuropsychiatric Disorders. *Biological Psychiatry*. 2018;83(9):722-730. doi:10.1016/j.biopsych.2017.11.018
556. Klinedinst NJ, Regenold WT. A mitochondrial bioenergetic basis of depression. *J Bioenerg Biomembr*. 2015;47(1):155-171. doi:10.1007/s10863-014-9584-6
557. Kety SS. Circulation and metabolism of the human brain in health and disease. *The American Journal of Medicine*. 1950;8(2):205-217. doi:10.1016/00029343(50)90363-9
558. Roth KA, D'Sa C. Apoptosis and brain development. *Mental Retardation and Developmental Disabilities Research Reviews*. 2001;7(4):261-266. doi:10.1002/mrdd.1036
559. Panov A, Orynbayeva Z, Vavilin V, Lyakhovich V. Fatty Acids in Energy Metabolism of the Central Nervous System. *BioMed Research International*. 2014;2014(1):472459. doi:10.1155/2014/472459
560. Emmerzaal TL, Nijkamp G, Veldic M, et al. Effect of neuropsychiatric medications on mitochondrial function: For better or for worse. *Neuroscience & Biobehavioral Reviews*. 2021;127:555-571. doi:10.1016/j.neubiorev.2021.05.001
561. Gonçalves CL, Rezin GT, Ferreira GK, et al. Differential effects of escitalopram administration on metabolic parameters of cortical and subcortical brain regions of

- Wistar rats. *Acta Neuropsychiatrica*. 2012;24(3):147-154.
doi:10.1111/j.16015215.2011.00592.x
562. Gaur V, Kumar A. Behavioral, biochemical and cellular correlates in the protective effect of sertraline against transient global ischemia induced behavioral despair: Possible involvement of nitric oxide-cyclic guanosine monophosphate study pathway. *Brain Research Bulletin*. 2010;82(1):57-64.
doi:10.1016/j.brainresbull.2010.01.010
563. Cox D. 'It feels like we've been lobotomised': the possible sexual consequences of SSRIs. *The Observer*.
<https://www.theguardian.com/society/2024/mar/02/ssriantidepressants-sexual-dysfunction-side-effects-consequences-libido>. March 2, 2024. Accessed March 19, 2025.
564. Andersohn F, Schade R, Suissa S, Garbe E. Long-Term Use of Antidepressants for Depressive Disorders and the Risk of Diabetes Mellitus. *AJP*. 2009;166(5):591-598.
doi:10.1176/appi.ajp.2008.08071065
565. Cartwright C, Gibson J, Kerry, Read J, John, Cowan J, Ondria, and Dehar T. Long-term antidepressant use: patient perspectives of benefits and adverse effects. *Patient Preference and Adherence*. 2016;10:1401-1407. doi:10.2147/PPA.S110632
566. Bansal N, Hudda M, Payne RA, Smith DJ, Kessler D, Wiles N. Antidepressant use and risk of adverse outcomes: population-based cohort study. *BJPsych Open*. 2022;8(5):e164. doi:10.1192/bjo.2022.563
567. Miidera H, Enomoto M, Kitamura S, Tachimori H, Mishima K. Association Between the Use of Antidepressants and the Risk of Type 2 Diabetes: A Large, Population-Based Cohort Study in Japan. *Diabetes Care*. 2020;43(4):885-893.
doi:10.2337/dc19-1175
568. Frisard C, Gu X, Whitcomb B, et al. Marginal structural models for the estimation of the risk of Diabetes Mellitus in the presence of elevated depressive symptoms and antidepressant medication use in the Women's Health Initiative observational and clinical trial cohorts. *BMC Endocr Disord*. 2015;15:56. doi:10.1186/s12902-0150049-7
569. Reisman Y. Post-SSRI sexual dysfunction. Published online February 27, 2020.
doi:10.1136/bmj.m754
570. Bala A, Nguyen HMT, Hellstrom WJG. Post-SSRI Sexual Dysfunction: A Literature Review. *Sexual Medicine Reviews*. 2018;6(1):29-34.
doi:10.1016/j.sxmr.2017.07.002

571. Healy D, Mangin D. Post-SSRI sexual dysfunction: barriers to quantifying incidence and prevalence. *Epidemiology and Psychiatric Sciences*. 2024;33:e40. doi:10.1017/S2045796024000441
572. Muroňová J, Kherraf ZE, Giordani E, et al. Lack of CCDC146, a ubiquitous centriole and microtubule-associated protein, leads to non-syndromic male infertility in human and mouse. *eLife*. 2024;12:RP86845. doi:10.7554/eLife.86845
573. Sylvester C, Menke M, Gopalan P. Selective Serotonin Reuptake Inhibitors and Fertility: Considerations for Couples Trying to Conceive. *Harv Rev Psychiatry*. 2019;27(2):108-118. doi:10.1097/HRP.0000000000000204
574. Engelmann J, Zillich L, Frank J, et al. Epigenetic signatures in antidepressant treatment response: a methylome-wide association study in the EMC trial. *Transl Psychiatry*. 2022;12(1):1-9. doi:10.1038/s41398-022-02032-7
575. Schiele MA, Crespo Salvador O, Lipovsek J, et al. Epigenome-Wide DNA Methylation in Unipolar Depression: Predictive Biomarker of Antidepressant Treatment Response? *International Journal of Neuropsychopharmacology*. 2024;27(11):pyae045. doi:10.1093/ijnp/pyae045
576. Powell TR, Smith RG, Hackinger S, et al. DNA methylation in interleukin-11 predicts clinical response to antidepressants in GENDEP. *Transl Psychiatry*. 2013;3(9):e300-e300. doi:10.1038/tp.2013.73
577. Karczewski KJ, Snyder MP. Integrative omics for health and disease. *Nat Rev Genet*. 2018;19(5):299-310. doi:10.1038/nrg.2018.4
578. Johnstone IM, Titterton DM. Statistical challenges of high-dimensional data. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*. 2009;367(1906):4237-4253. doi:10.1098/rsta.2009.0159
579. Perng W, Aslibekyan S. Find the Needle in the Haystack, Then Find It Again: Replication and Validation in the 'Omics Era. *Metabolites*. 2020;10(7):286. doi:10.3390/metabo10070286
580. Wu Y, Qi T, Wray NR, Visscher PM, Zeng J, Yang J. Joint analysis of GWAS and multi-omics QTL summary statistics reveals a large fraction of GWAS signals shared with molecular phenotypes. *Cell Genom*. 2023;3(8):100344. doi:10.1016/j.xgen.2023.100344
581. Giambartolomei C, Zhenli Liu J, Zhang W, et al. A Bayesian framework for multiple trait colocalization from summary association statistics. *Bioinformatics*. 2018;34(15):2538-2545. doi:10.1093/bioinformatics/bty147
582. Foley CN, Staley JR, Breen PG, et al. A fast and efficient colocalization algorithm for identifying shared genetic risk factors across multiple traits. *Nat Commun*.

2021;12(1):764. doi:10.1038/s41467-020-20885-8

583. Gleason KJ, Yang F, Pierce BL, He X, Chen LS. Primo: integration of multiple GWAS and omics QTL summary statistics for elucidation of molecular mechanisms of trait-associated SNPs and detection of pleiotropy in complex traits. *Genome Biol.* 2020;21(1):236. doi:10.1186/s13059-020-02125-w
584. Jahagirdar S, Saccenti E. On the Use of Correlation and MI as a Measure of Metabolite—Metabolite Association for Network Differential Connectivity Analysis. *Metabolites.* 2020;10(4):171. doi:10.3390/metabo10040171
585. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics.* 2008;9(1):559. doi:10.1186/1471-2105-9-559
586. Gadd DA, Hillary RF, McCartney DL, et al. Epigenetic scores for the circulating proteome as tools for disease prediction. Lo YD, Ferrucci L, eds. *eLife.* 2022;11:e71802. doi:10.7554/eLife.71802
587. van Dijk MT, Murphy E, Posner JE, Talati A, Weissman MM. Association of Multigenerational Family History of Depression With Lifetime Depressive and Other Psychiatric Disorders in Children: Results from the Adolescent Brain Cognitive Development (ABCD) Study. *JAMA Psychiatry.* 2021;78(7):778-787. doi:10.1001/jamapsychiatry.2021.0350
588. Popejoy AB, Fullerton SM. Genomics is failing on diversity. *Nature.* 2016;538(7624):161-164. doi:10.1038/538161a
589. TOPMed OMICS Progress Report. Accessed March 21, 2025. <https://nhlbi.sph.umich.edu/omics/index.php#>
590. Fry A, Littlejohns TJ, Sudlow C, et al. Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. *American Journal of Epidemiology.* 2017;186(9):1026-1034. doi:10.1093/aje/kwx246
591. Pictor M, Teare HJA, Kaye J. Equitable Participation in Biobanks: The Risks and Benefits of a “Dynamic Consent” Approach. *Front Public Health.* 2018;6:253. doi:10.3389/fpubh.2018.00253
592. Cook MB, Sanderson SC, Deanfield JE, et al. Our Future Health: a unique global resource for discovery and translational research. *Nat Med.* 2025;31(3):728-730. doi:10.1038/s41591-024-03438-0
593. Danner B, Gonzalez AD, Corbett WC, et al. Brain banking in the United States and Europe: Importance, challenges, and future trends. *Journal of Neuropathology & Experimental Neurology.* 2024;83(4):219-229. doi:10.1093/jnen/nlae014

594. Bennett DA, Buchman AS, Boyle PA, Barnes LL, Wilson RS, Schneider JA. Religious Orders Study and Rush Memory and Aging Project. *J Alzheimers Dis.* 2018;64(Suppl 1):S161-S189. doi:10.3233/JAD-179939
595. Pifferi F, Laurent B, Plourde M. Lipid Transport and Metabolism at the Blood-Brain Interface: Implications in Health and Disease. *Front Physiol.* 2021;12. doi:10.3389/fphys.2021.645646
596. Huo Z, Yu L, Yang J, Zhu Y, Bennett DA, Zhao J. Brain and blood metabolome for Alzheimer's dementia: Findings from a targeted metabolomics analysis. *Neurobiol Aging.* 2020;86:123-133. doi:10.1016/j.neurobiolaging.2019.10.014
597. Wang C, Yang C, Western D, et al. Genetic architecture of cerebrospinal fluid and brain metabolite levels and the genetic colocalization of metabolites with human traits. *Nat Genet.* 2024;56(12):2685-2695. doi:10.1038/s41588-024-01973-7
598. Hannon E, Lunnon ,Katie, Schalkwyk ,Leonard, and Mill J. Interindividual methylomic variation across blood, cortex, and cerebellum: implications for epigenetic studies of neurological and neuropsychiatric phenotypes. *Epigenetics.* 2015;10(11):1024-1032. doi:10.1080/15592294.2015.1100786
599. Walton E, Hass J, Liu J, et al. Correspondence of DNA Methylation Between Blood and Brain Tissue and Its Application to Schizophrenia Research. *Schizophrenia Bulletin.* 2016;42(2):406-414. doi:10.1093/schbul/sbv074
600. Edgar RD, Jones MJ, Meaney MJ, Turecki G, Kobor MS. BECon: a tool for interpreting DNA methylation findings from blood in the context of brain. *Transl Psychiatry.* 2017;7(8):e1187-e1187. doi:10.1038/tp.2017.171
601. Braun PR, Han S, Hing B, et al. Genome-wide DNA methylation comparison between live human brain and peripheral tissues within individuals. *Transl Psychiatry.* 2019;9(1):1-10. doi:10.1038/s41398-019-0376-y
602. Nishitani S, Isozaki M, Yao A, et al. Cross-tissue correlations of genome-wide DNA methylation in Japanese live human brain and blood, saliva, and buccal epithelial tissues. *Transl Psychiatry.* 2023;13(1):1-10. doi:10.1038/s41398-023-02370-0
603. Souery D, Oswald P, Massat I, et al. Clinical factors associated with treatment resistance in major depressive disorder: results from a European multicenter study. *J Clin Psychiatry.* 2007;68(7):1062-1070. doi:10.4088/jcp.v68n0713
604. Grant CW, Barreto EF, Kumar R, et al. Multi-Omics Characterization of Early- and Adult-Onset Major Depressive Disorder. *Journal of Personalized Medicine.* 2022;12(3):412. doi:10.3390/jpm12030412

605. Fried EI, Nesse RM. Depression sum-scores don't add up: why analyzing specific depression symptoms is essential. *BMC Med*. 2015;13(1):72. doi:10.1186/s12916015-0325-4
606. Yamamoto H, Lee-Okada HC, Ikeda M, et al. GWAS-identified bipolar disorder risk allele in the FADS1/2 gene region links mood episodes and unsaturated fatty acid metabolism in mutant mice. *Mol Psychiatry*. 2023;28(7):2848-2856. doi:10.1038/s41380-023-01988-2
607. Logsdon GA, Vollger MR, Eichler EE. Long-read human genome sequencing and its applications. *Nat Rev Genet*. 2020;21(10):597-614. doi:10.1038/s41576-0200236-x
608. Liao WW, Asri M, Ebler J, et al. A draft human pangenome reference. *Nature*. 2023;617(7960):312-324. doi:10.1038/s41586-023-05896-x
609. Simpson JT, Workman RE, Zuzarte PC, David M, Dursi LJ, Timp W. Detecting DNA cytosine methylation using nanopore sequencing. *Nat Methods*. 2017;14(4):407-410. doi:10.1038/nmeth.4184
610. Ryals J, Lawton ,Kay, Stevens ,Daniel, and Milburn M. Metabolon, Inc. *Pharmacogenomics*. 2007;8(7):863-866. doi:10.2217/14622416.8.7.863
611. Yet I, Menni C, Shin SY, et al. Genetic Influences on Metabolite Levels: A Comparison across Metabolomic Platforms. *PLOS ONE*. 2016;11(4):e0153672. doi:10.1371/journal.pone.0153672
612. Bizzarri D, Reinders MJT, Beekman M, Slagboom PE, van den Akker EB. Technical Report: A Comprehensive Comparison between Different Quantification Versions of Nightingale Health's 1H-NMR Metabolomics Platform. *Metabolites*. 2023;13(12):1181. doi:10.3390/metabo13121181
613. Wishart DS, Guo A, Oler E, et al. HMDB 5.0: the Human Metabolome Database for 2022. *Nucleic Acids Res*. 2021;50(D1):D622-D631. doi:10.1093/nar/gkab1062
614. Harron K. Data linkage in medical research. *BMJ Medicine*. 2022;1(1). doi:10.1136/bmjmed-2021-000087
615. Kalkman S, Van Delden J, Banerjee A, Tyl B, Mostert M, Van Thiel G. Patients' and public views and attitudes towards the sharing of health data for research: a narrative review of the empirical evidence. *J Med Ethics*. 2022;48(1):3-13. doi:10.1136/medethics-2019-105651
616. Ozair FF, Jamshed N, Sharma A, Aggarwal P. Ethical issues in electronic health records: A general overview. *Perspect Clin Res*. 2015;6(2):73-76. doi:10.4103/2229-3485.153997

617. De Simoni A, Jackson T, Inglis Humphrey W, et al. Patient and public involvement in research: the need for budgeting PPI staff costs in funding applications. *Research Involvement and Engagement*. 2023;9(1):16. doi:10.1186/s40900-023-00424-

10 Appendix

This Appendix includes the supplementary figures and tables accompanying the analysis in **4**.

10.1 Supplementary Figures

Supplementary Figure 1: FADS cluster genes investigated in Chapter 4.....	337
Supplementary Figure 2: Variant quality control filters per-gene.	338
Supplementary Figure 3: Sample quality control filters per-gene.	339
Supplementary Figure 4: Sample-level missingness per gene.....	340
Supplementary Figure 5: Variants allele count (AC) per gene following sample quality control	341
Supplementary Figure 6: Allele frequency of variants in each gene following quality control. Histograms showing the number of variants within different bins of allele frequency per gene.	342
Supplementary Figure 7: Allele count of variants in each gene following quality control Histograms showing the number of variants within different bins of allele count per gene.	343
Supplementary Figure 8: Prioritised variants in FADS1 in the MDD and Metabolite cohort	345
Supplementary Figure 9: Prioritised variants in FADS2 in the MDD and Metabolite cohort	346
Supplementary Figure 10: Prioritised variants in FADS3 in the MDD and Metabolite cohort.	347
Supplementary Figure 11: Prioritised variants in FEN1 in the MDD and Metabolite cohort.....	348
Supplementary Figure 12: Prioritised variants in MYRF in the MDD and Metabolite cohort.	349
Supplementary Figure 13: Prioritised variants in TMEM258 in the MDD and Metabolite cohort.....	350
Supplementary Figure 14: Linkage Disequilibrium matrix of all prioritised variants in FADS2.	352
Supplementary Figure 15: Linkage Disequilibrium matrix of all prioritised variants in TMEM258.	353
Supplementary Figure 16: Linkage Disequilibrium matrix of all prioritised variants in FADS3.	354
Supplementary Figure 17: Linkage Disequilibrium matrix of all prioritised variants in FEN1.....	355
Supplementary Figure 18:Linkage Disequilibrium matrix of all prioritised variants in FADS1).	356
Supplementary Figure 19: Linkage disequilibrium matrix of all prioritized variants in MYRF.....	357

Supplementary Figure 20: UpSet plots showing the distribution of prioritised variants per carried by participants in the MDD-cohort (A) and Metabolite-cohort (B).....	358
Supplementary Figure 21: Carriers of prioritised variants in each mask.	359
Supplementary Figure 22: FADS1 prioritised variants with AF < 0.01 in each mask.	361
Supplementary Figure 23: FADS1 prioritised singleton variants in each mask.	362
Supplementary Figure 24: FADS2 prioritised variants with AF < 0.01 in each mask.	363
Supplementary Figure 25: FADS2 prioritised singleton variants in each mask.	364
Supplementary Figure 26: FADS3 prioritised variants with AF < 0.01 in each mask.	365
Supplementary Figure 27: FADS3 prioritised singleton variants in each mask.	366
Supplementary Figure 28: FEN1 prioritised variants with AF < 0.01 in each mask.	367
Supplementary Figure 29: FEN1 prioritised singleton variants in each mask.	368
Supplementary Figure 30: MYRF prioritised variants with AF < 0.01 in each mask.	369
Supplementary Figure 31: MYRF prioritised singleton variants in each mask.	370
Supplementary Figure 32: TMEM258 prioritised variants with AF < 0.01 in each mask.	371
Supplementary Figure 33: TMEM258 prioritised singleton variants in each mask.....	372
Supplementary Figure 34: Violin plots of each metabolite trait in carriers and non-carriers of prioritised variants in TMEM258 assigned to Mask 4.	373
Supplementary Figure 35: Violin plots of each metabolite trait in carriers and non-carriers of prioritised variants in TMEM258 assigned to Mask 5 (AF < 0.01).	374
Supplementary Figure 36: Violin plots of each metabolite trait in carriers and non-carriers of prioritised variants in FEN1 assigned to Mask 5 (AF < 0.01).	375
Supplementary Figure 37: Violin plots of each metabolite trait in carriers and non-carriers of prioritised variants in FADS3 assigned to Mask 3 (AF < 0.01).	376

Supplementary Figure 38: The effect size of the pseudo-genotype on each trait for each mask in the traditional burden test..... 377

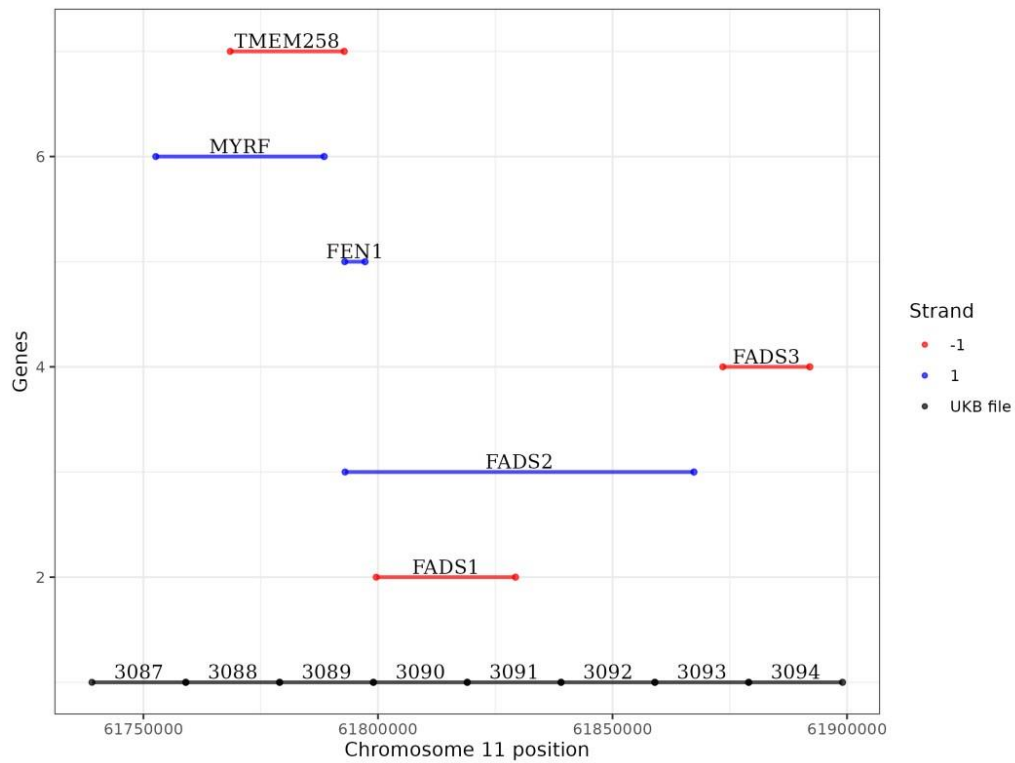
Supplementary Figure 39: Violin plots of each metabolite trait in carriers and non-carriers of chr11:61816814:G:C variant. 378

Supplementary Figure 40: Violin plots of each metabolite trait in carriers and non-carriers of chr11:61810815:C:A variant. 380

Supplementary Figure 41: Violin plots of each metabolite trait in carriers and non-carriers of chr11:61810815:C:A variant (Excluding the outlying sample)..... 381

10.1.1 Methods

10.1.1.1 Variant filtering and extraction

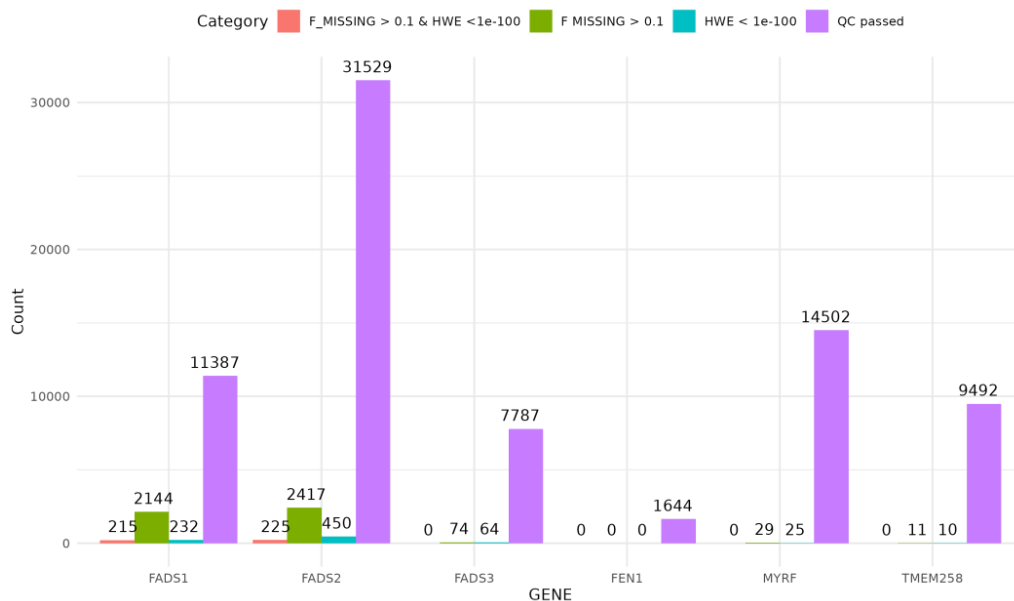


Supplementary Figure 1: FADS cluster genes investigated in Chapter 4.

A dumb-bell plot visualising the genes position within the FADS cluster (Chr11). The pVCF files containing the variants chunks which overlay these genes are shown on the bottom file. Blue = Forward Strand (1) and Red = Reverse Strand (-1).

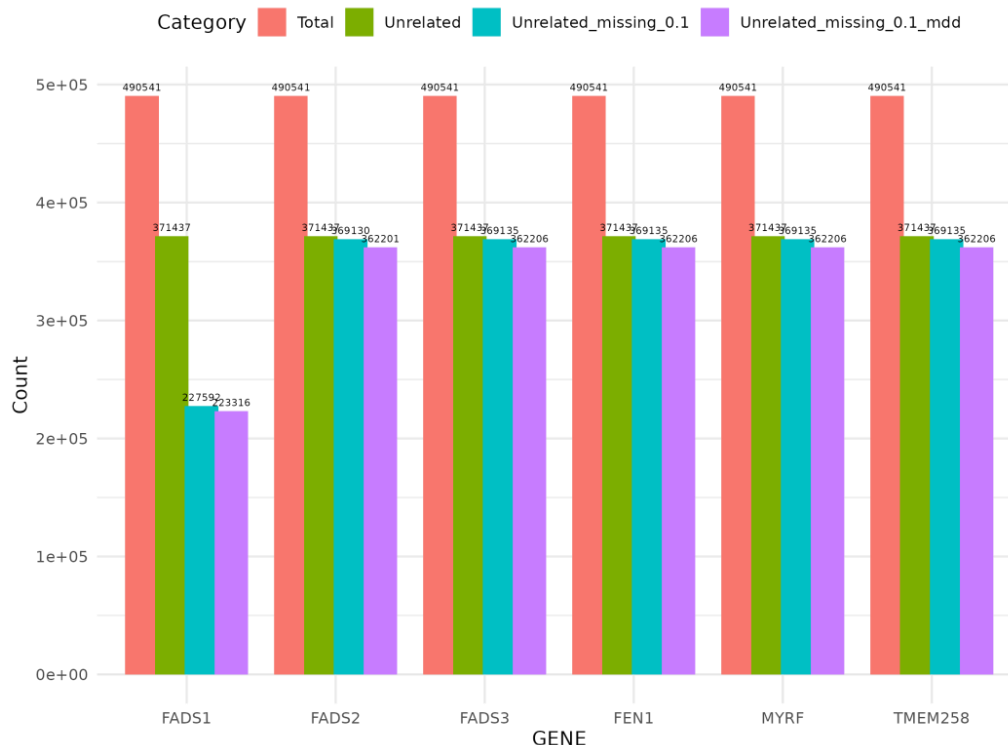
10.1.2 Results

10.1.2.1 Variant filtering and extraction



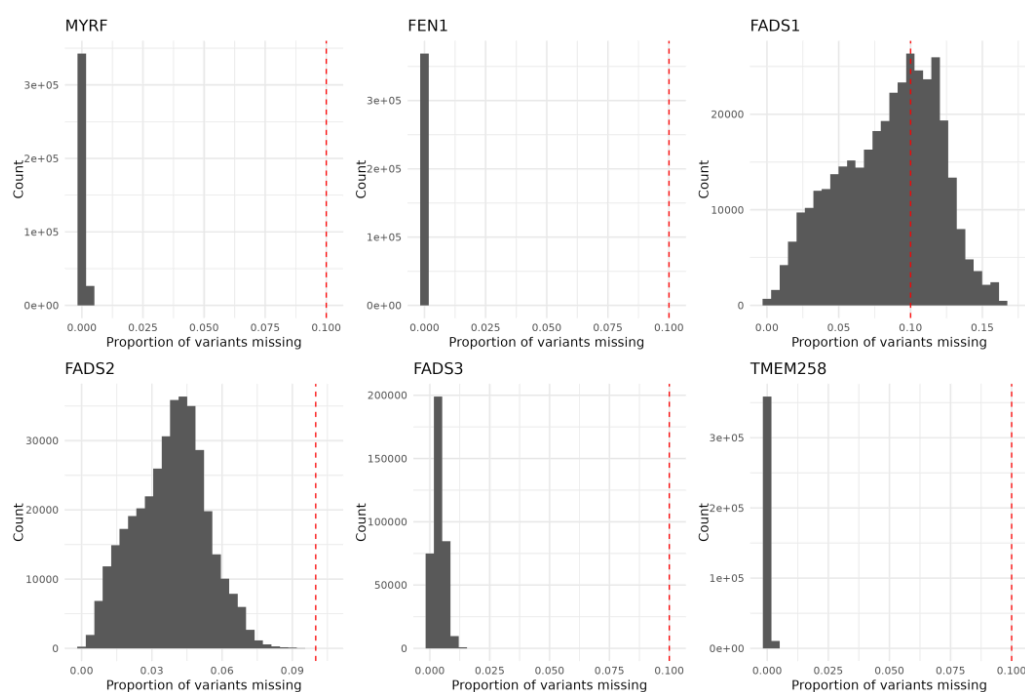
Supplementary Figure 2: Variant quality control filters per-gene.

A bar chart demonstrating the number of variants which fit different quality control (QC) criteria. The green bar (F_MISSING > 0.1) indicates the number of variants which had a missingness > 0.1 across all samples. The blue bar (HWE < $1e^{-100}$) indicates the number of variants which do not fit the HardyWeinburg Equilibrium with a P value < $1e^{-100}$. The red bar indicates the number of variants which have a missingness > 0.1 across all samples and have a HWE p-value < $1e^{-100}$. All these variants are removed in the QC step, leaving the purple bar, which indicates the number of variants which passed QC.



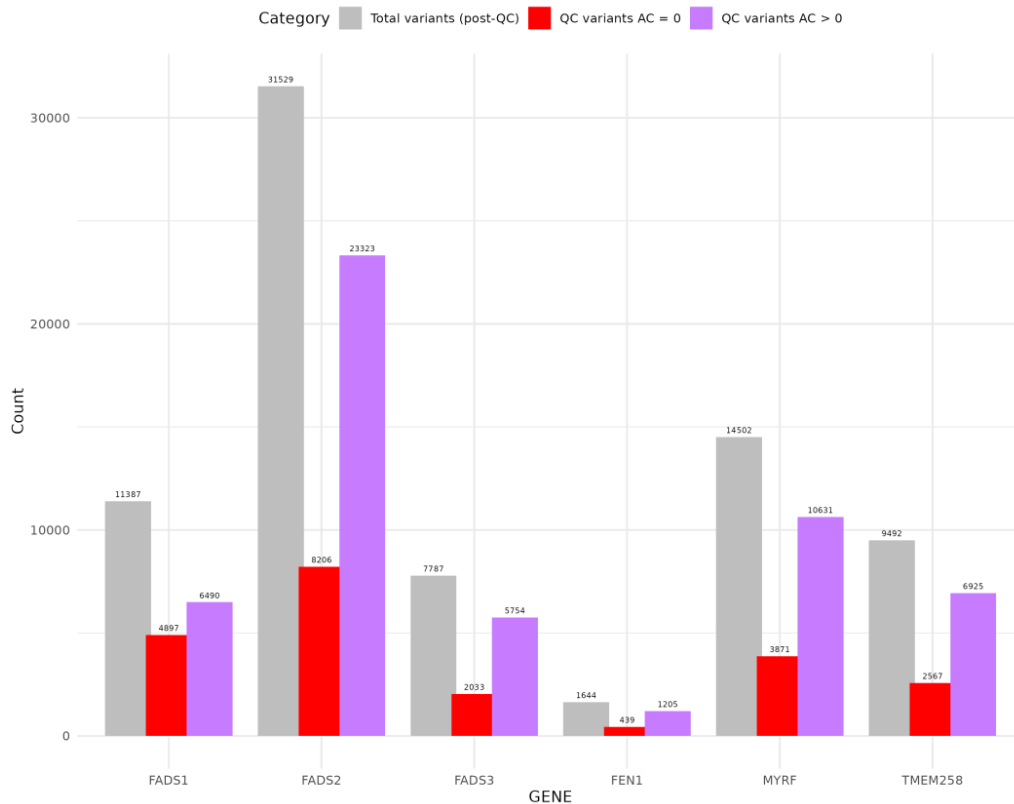
Supplementary Figure 3: Sample quality control filters per-gene.

A bar chart demonstrating the number of participants retained /excluded in the sample following different quality control (QC) criteria. The red bar (Total) indicates the total number of participants with whole genome sequencing variants within the gene, the green bar (Unrelated) indicates the number of unrelated participants, the blue bar indicates the number of unrelated participants with an overall missingness rate < 0.1 within the gene, and the purple bar indicates the number of unrelated participants with overall missingness < 0.1 and with a non-missing Major Depressive Disorder status.



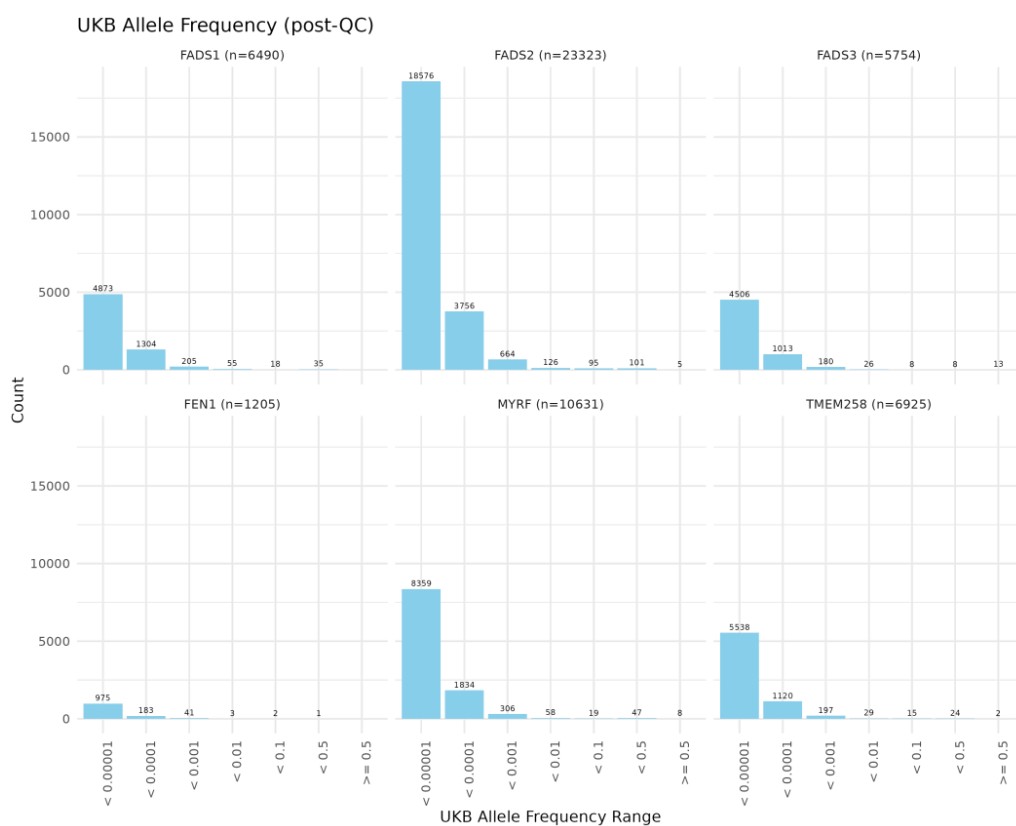
Supplementary Figure 4: Sample-level missingness per gene

A histogram per gene denoting the proportion of variants missing per participant. The red line indicates the quality control (QC) criteria of missingness of 0.1 for which participants will be removed. Note the higher proportion of samples with missingness > 0.1 for FADS1.



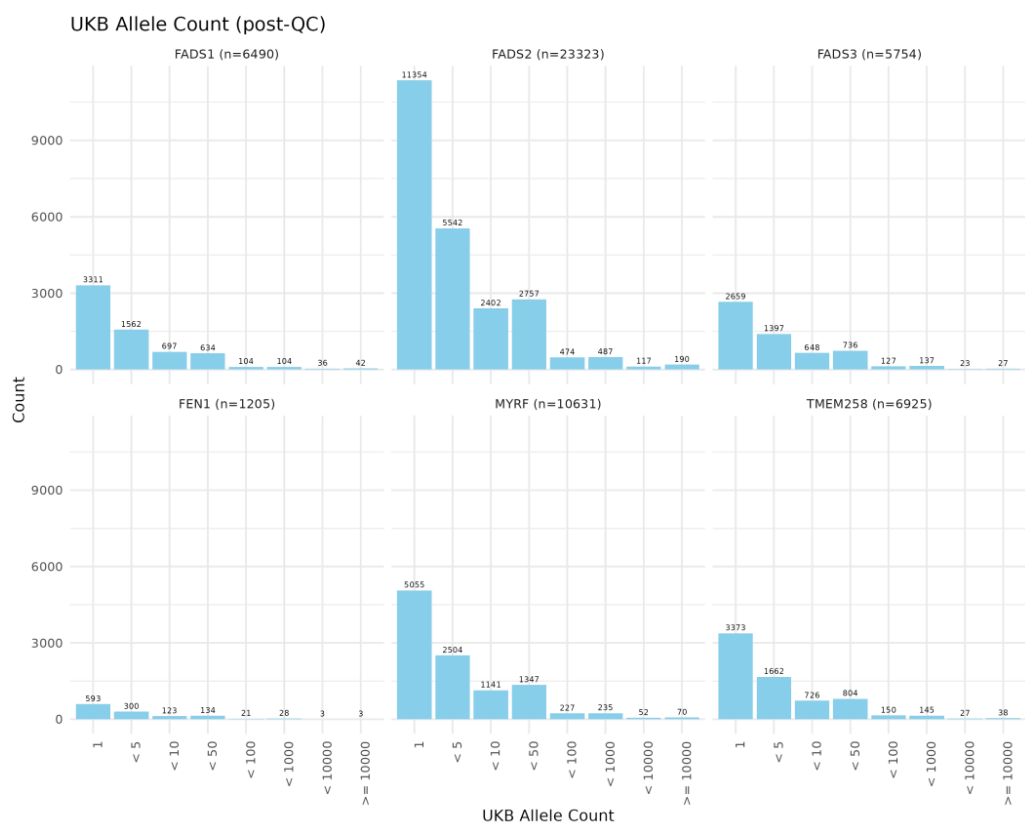
Supplementary Figure 5: Variants allele count (AC) per gene following sample quality control

A bar chart per gene indicating the number of variants removed following sample quality control due to a 0 AC. The gray bar indicates the total number of variants following variant level quality control, the red bars indicates the number of variants with AC of 0 following sample-level quality control which were excluded from the study, and the purple bar indicates the number of variants with an AC > 0 following sample quality control and were therefore retained in the study.



Supplementary Figure 6: Allele frequency of variants in each gene following quality control.

Histograms showing the number of variants within different bins of allele frequency per gene.



Supplementary Figure 7: Allele count of variants in each gene following quality control

Histograms showing the number of variants within different bins of allele count per gene.

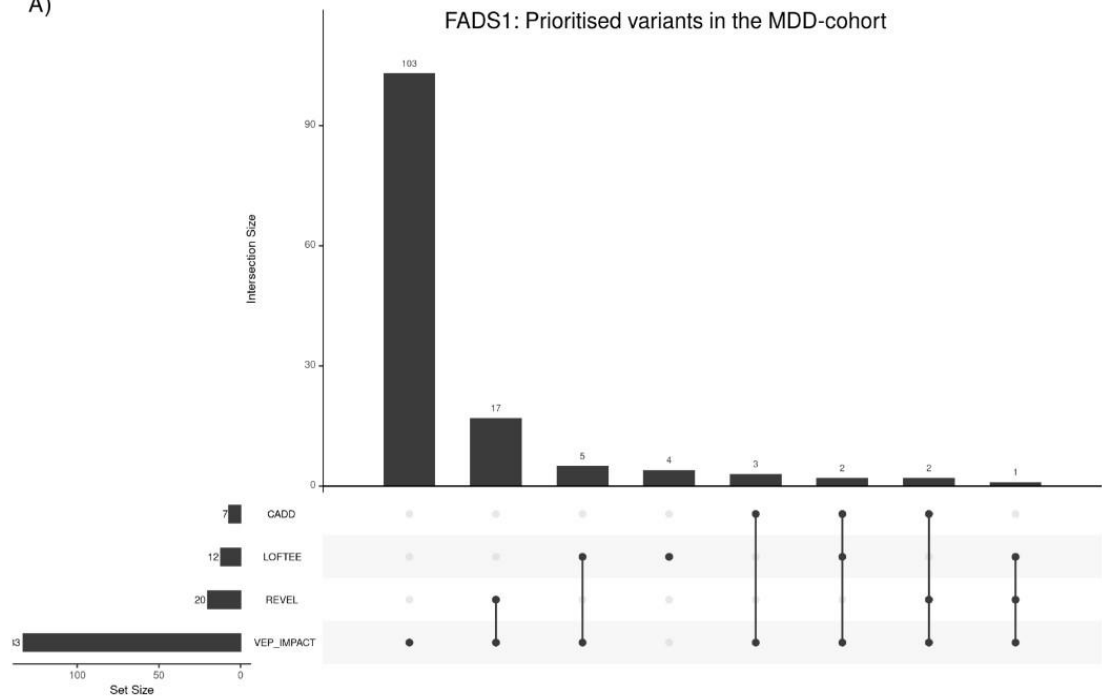
10.1.2.2 Variant prioritisation

10.1.2.2.1 Prioritisation criteria

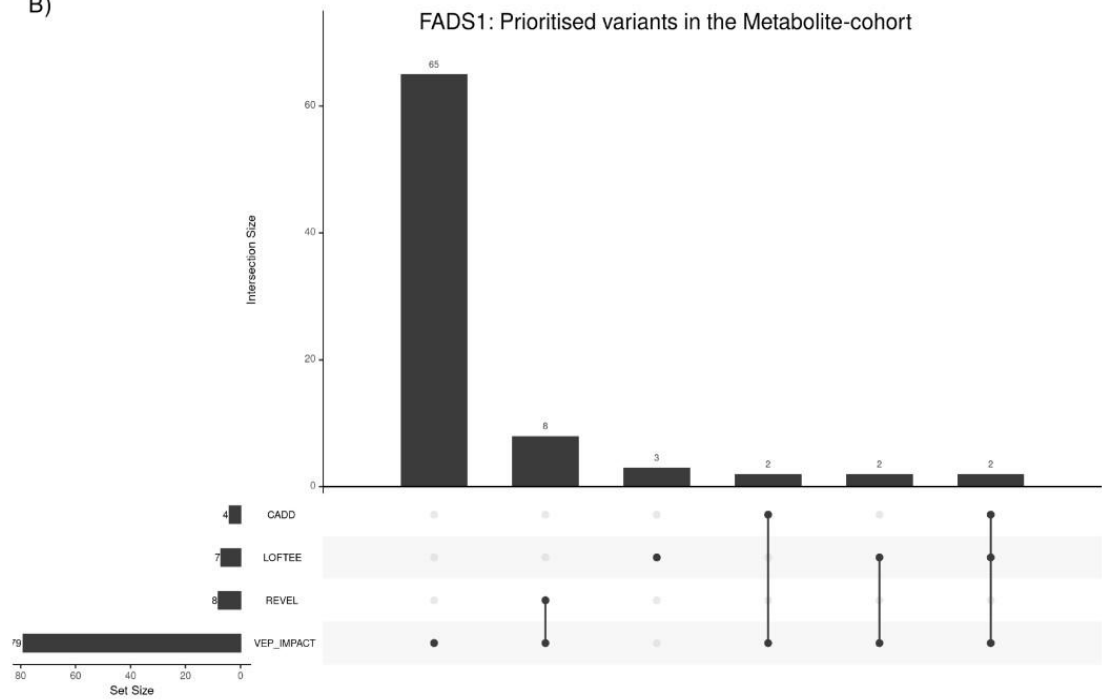
Supplementary plots 8-13 shows UpSet plots of the prioritisation criteria by prioritised variants carried in the MDD cohort (A) and the metabolite cohort (B) for each gene.

FADS1: Prioritised variants

A)



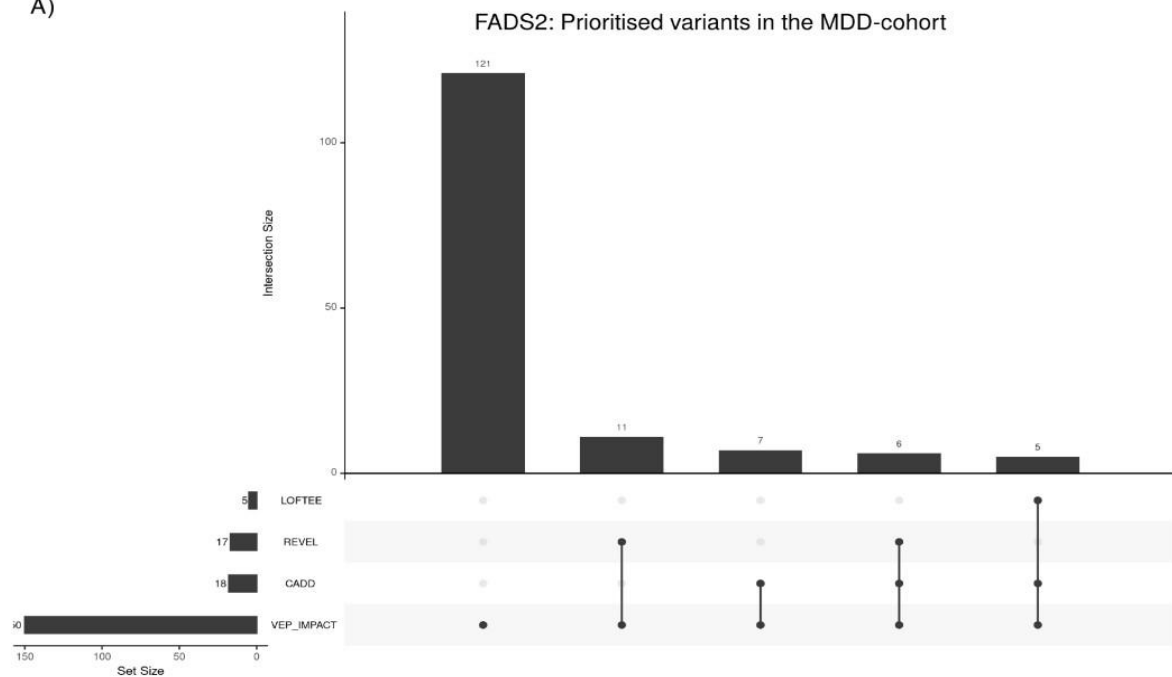
B)



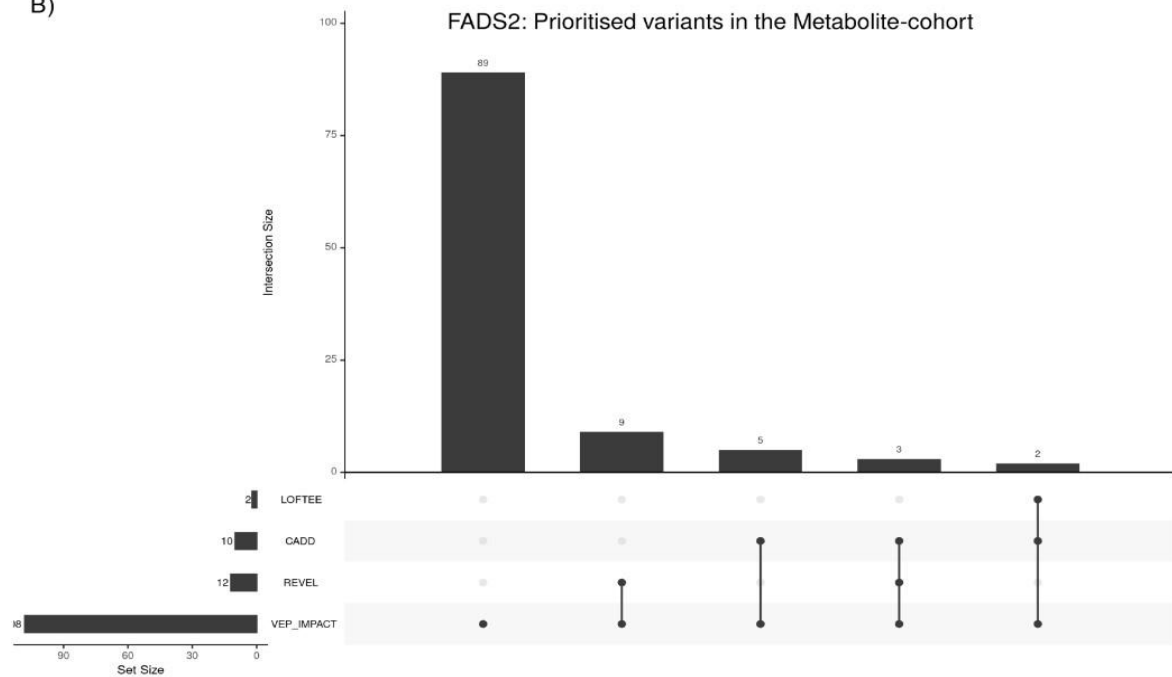
Supplementary Figure 8: Prioritised variants in FADS1 in the MDD and Metabolite cohort

FADS2: Prioritised variants

A)

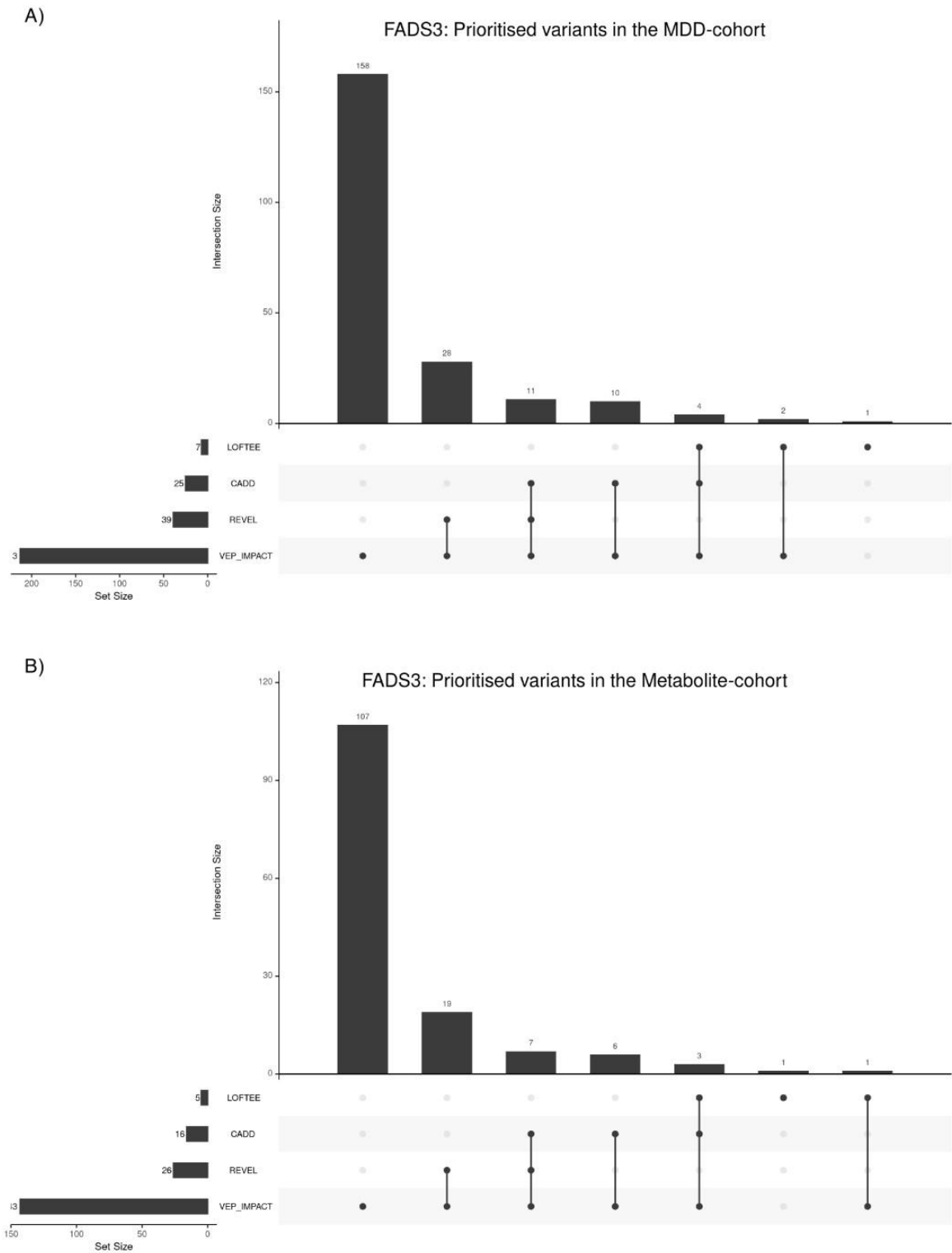


B)



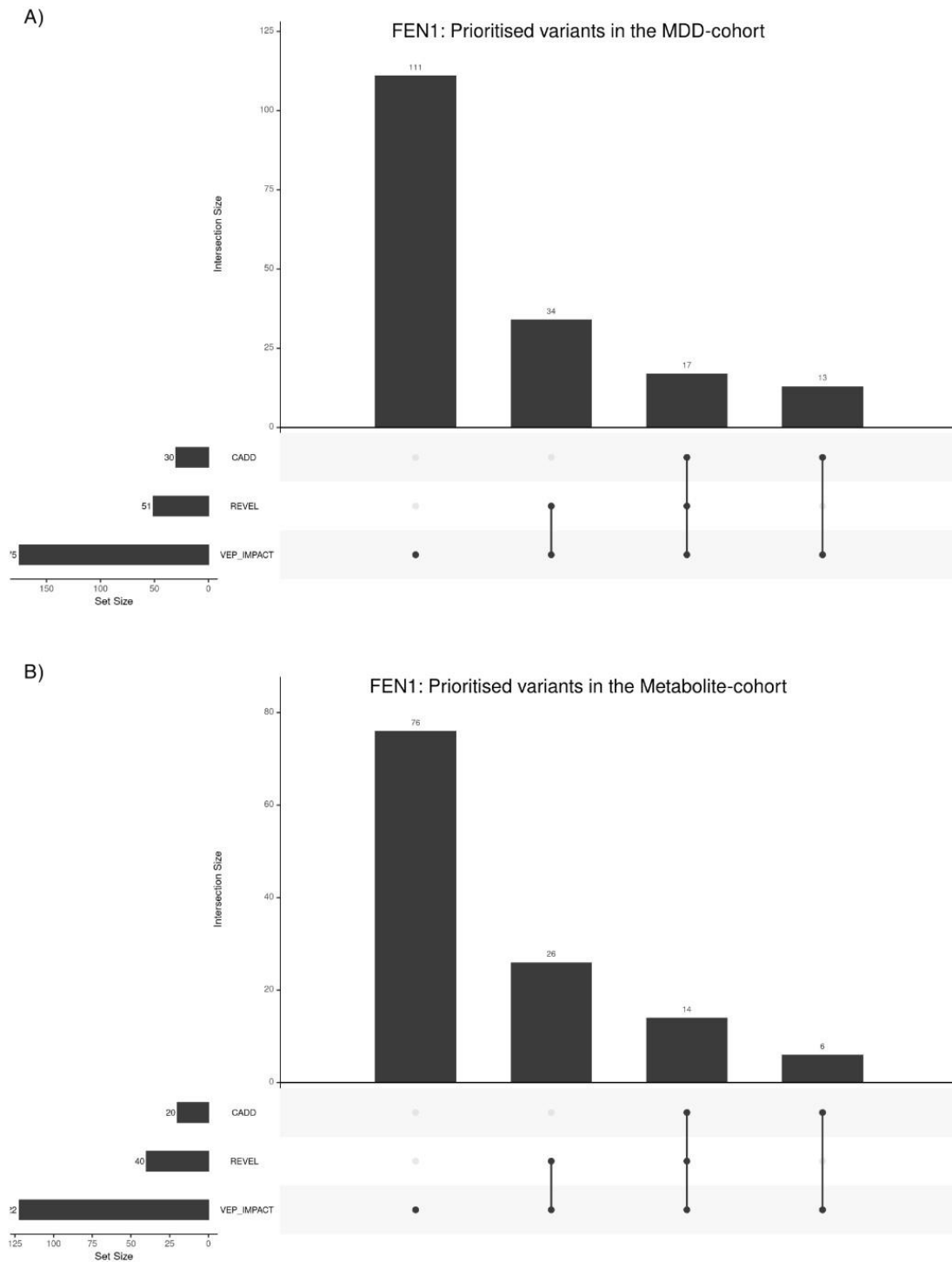
Supplementary Figure 9: Prioritised variants in FADS2 in the MDD and Metabolite cohort

FADS3: Prioritised variants



Supplementary Figure 10: Prioritised variants in FADS3 in the MDD and Metabolite cohort.

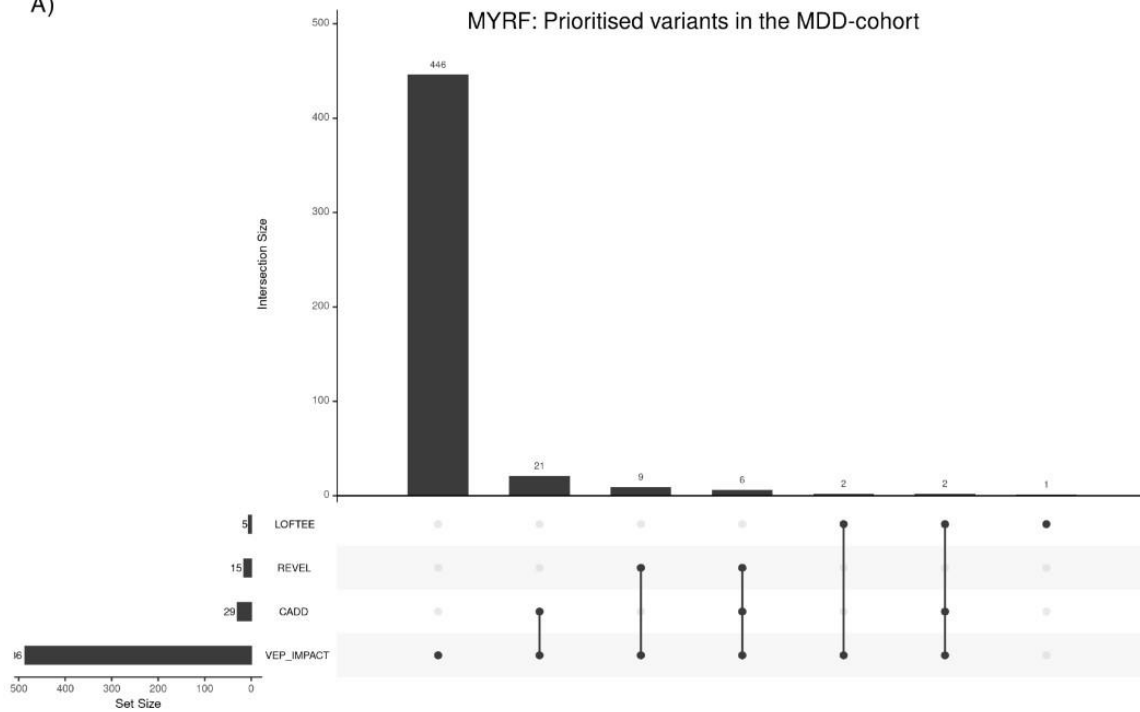
FEN1: Prioritised variants



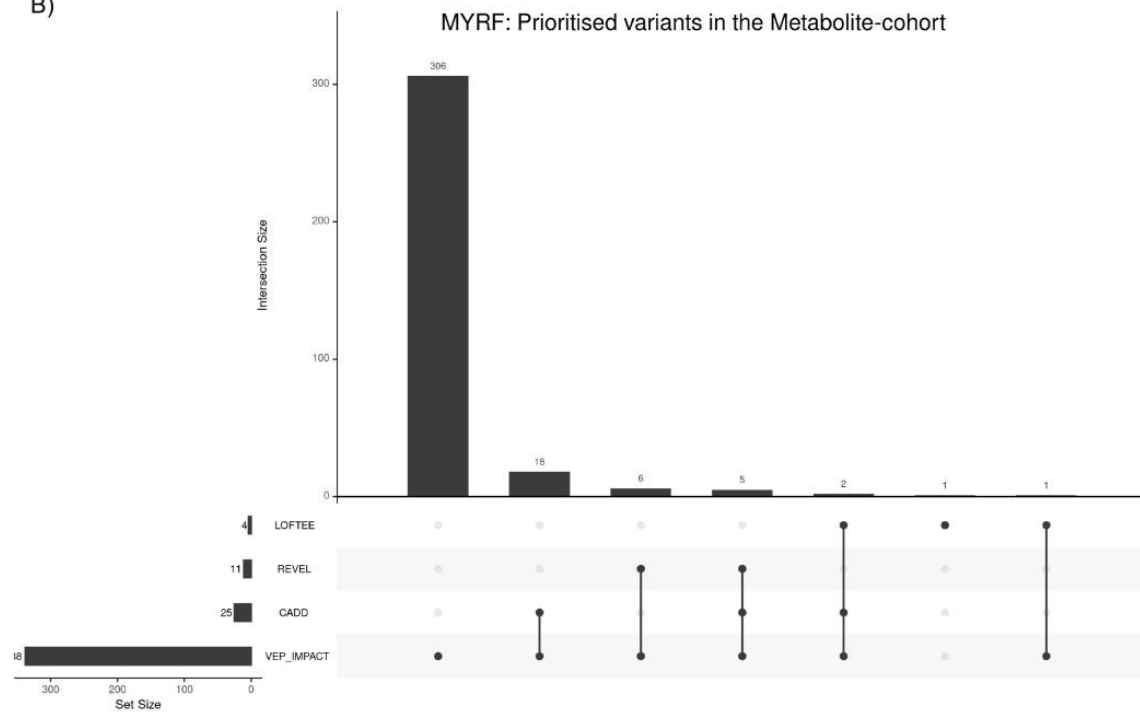
Supplementary Figure 11: Prioritised variants in FEN1 in the MDD and Metabolite cohort.

MYRF: Prioritised variants

A)



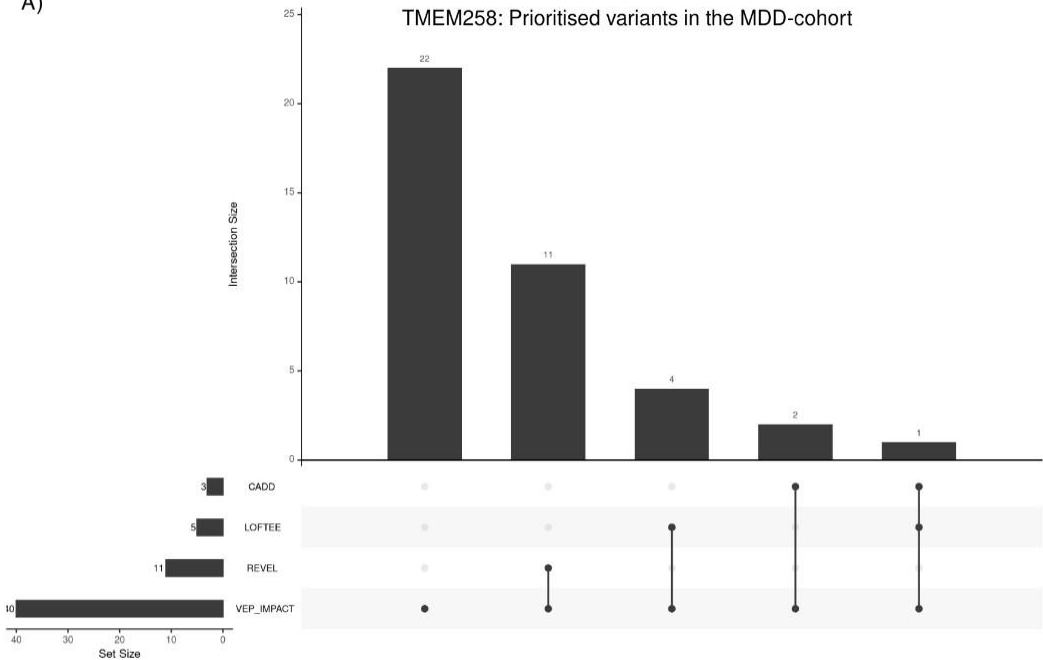
B)



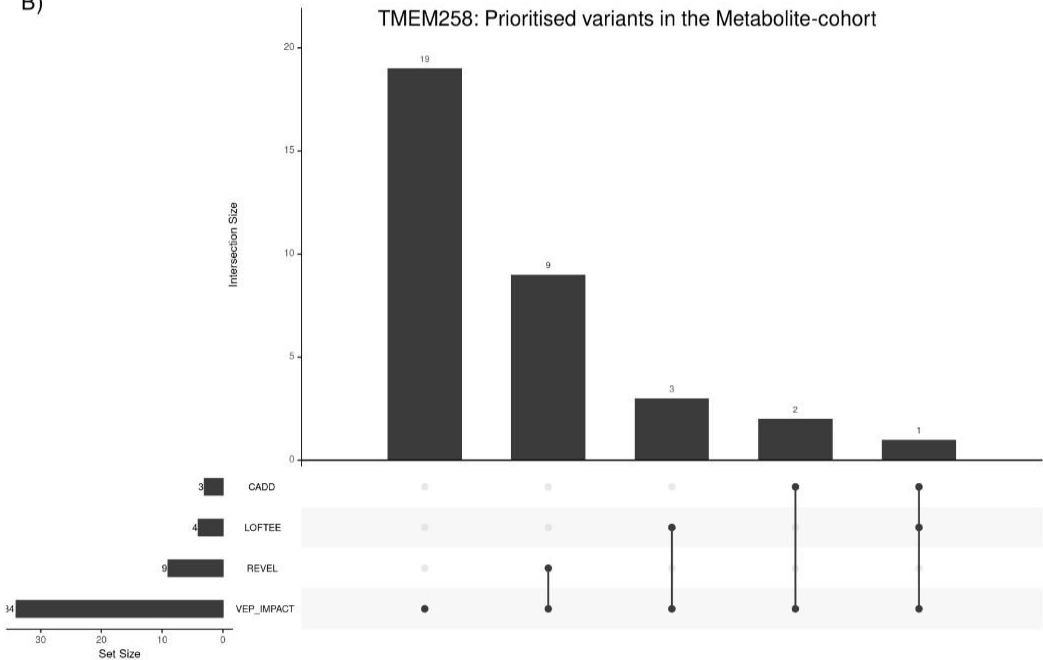
Supplementary Figure 12: Prioritised variants in MYRF in the MDD and Metabolite cohort.

TMEM258: Prioritised variants

A)



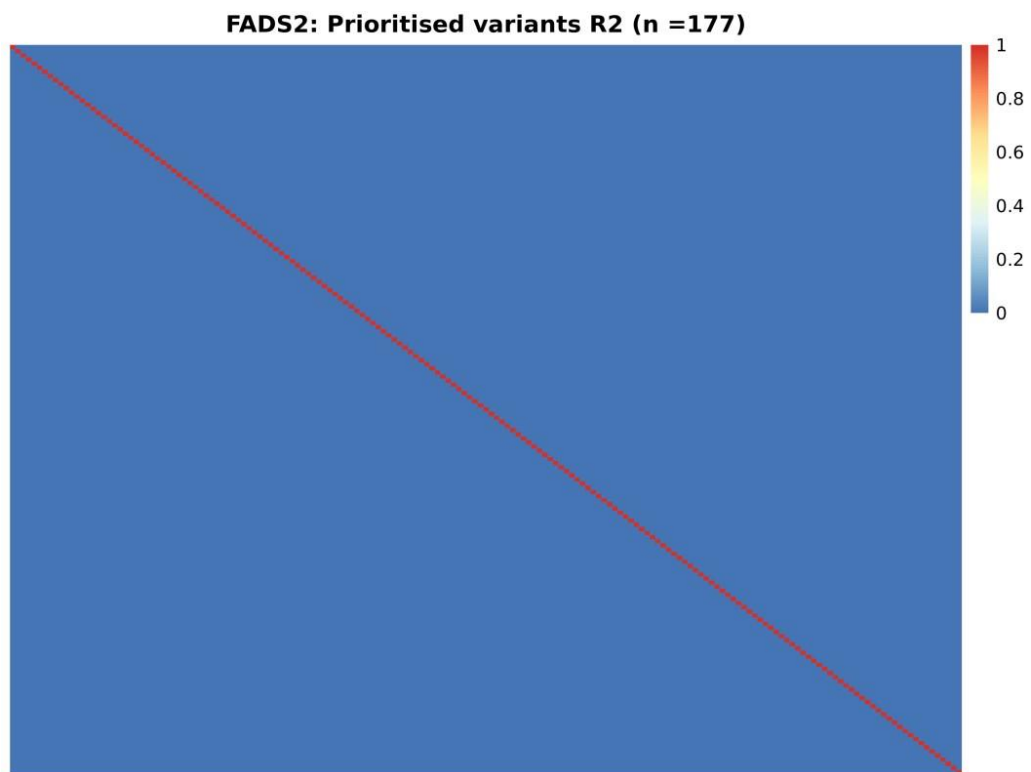
B)



Supplementary Figure 13: Prioritised variants in TMEM258 in the MDD and Metabolite cohort.

10.1.2.2.2 Linkage disequilibrium matrices

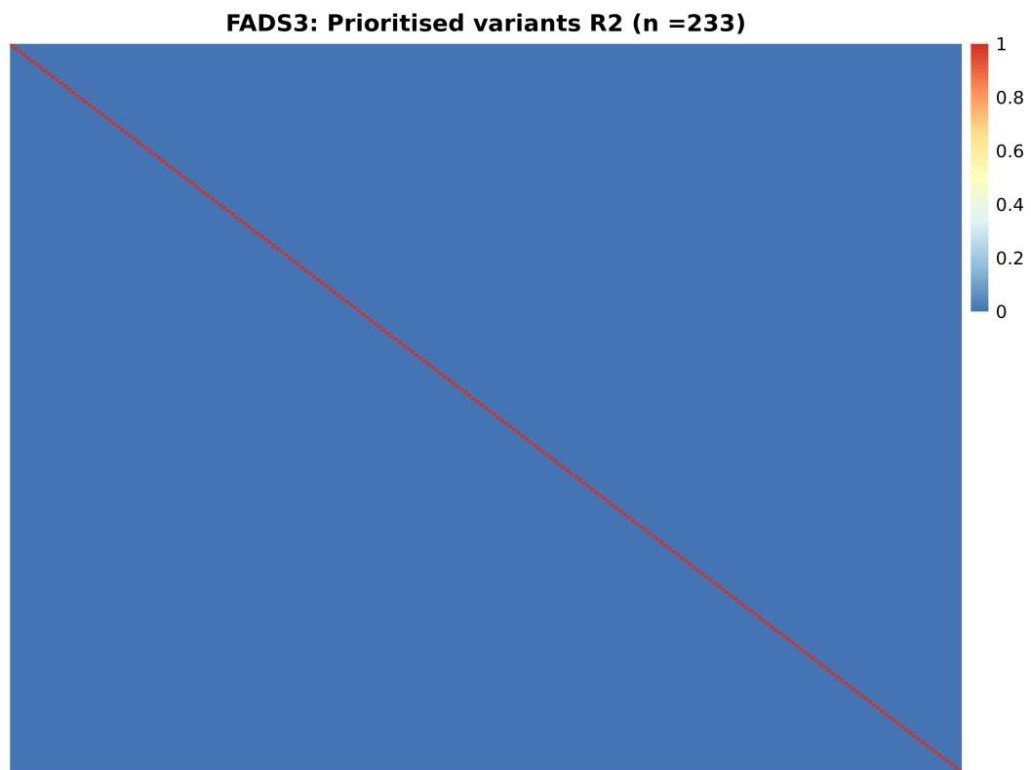
Supplementary Figures 14-19 shows the correlation (R^2) of all prioritized variant pairs in each gene, calculated using plink (-r2 flag). For genes with variant pairs with a correlation > 0.01 there are two plots: 1) shows all prioritized variants and B) shows a zoomed in correlation matrix of variant pairs with > 0.01 correlation.



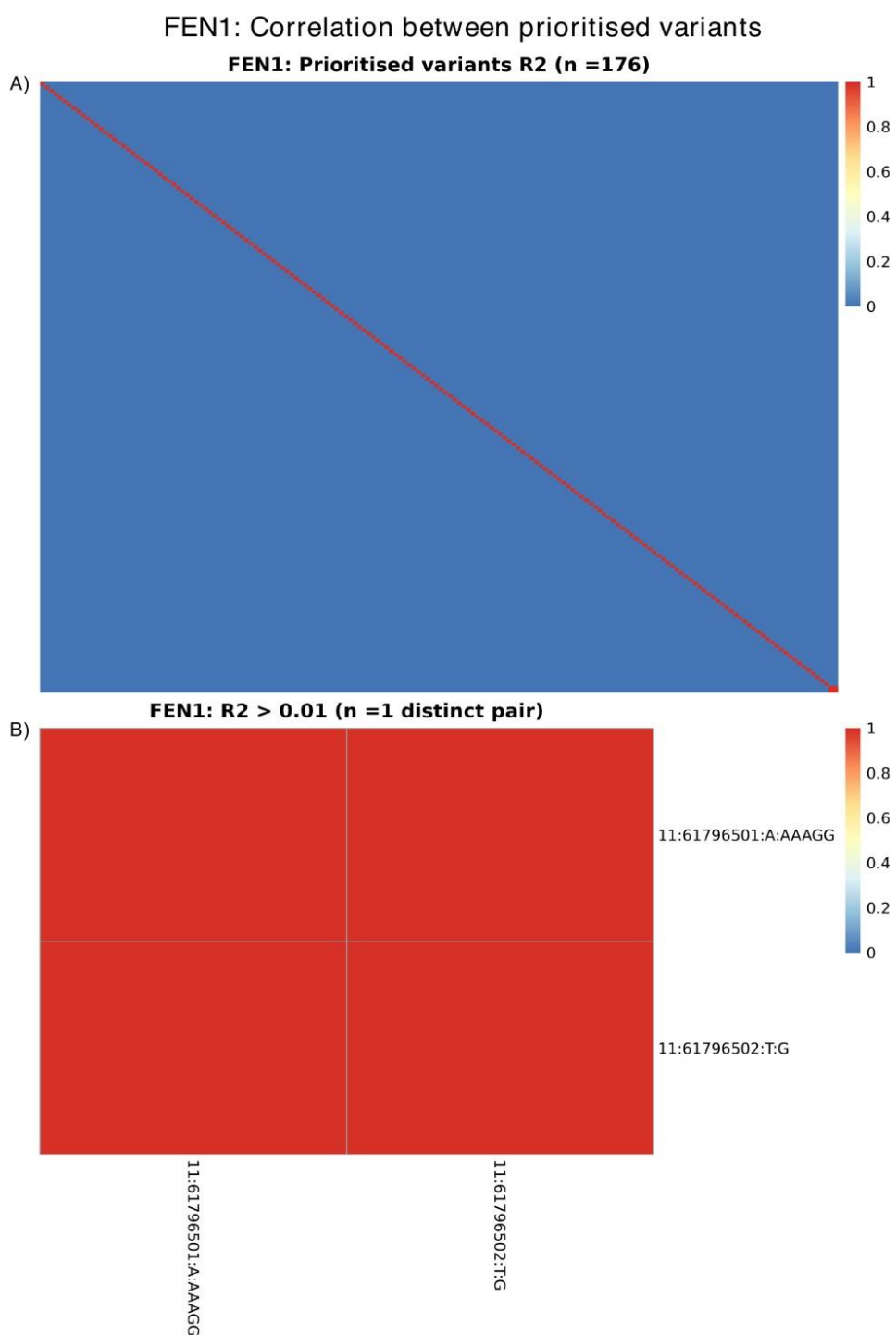
Supplementary Figure 14: Linkage Disequilibrium matrix of all prioritised variants in FADS2.



Supplementary Figure 15: Linkage Disequilibrium matrix of all prioritised variants in TMEM258.

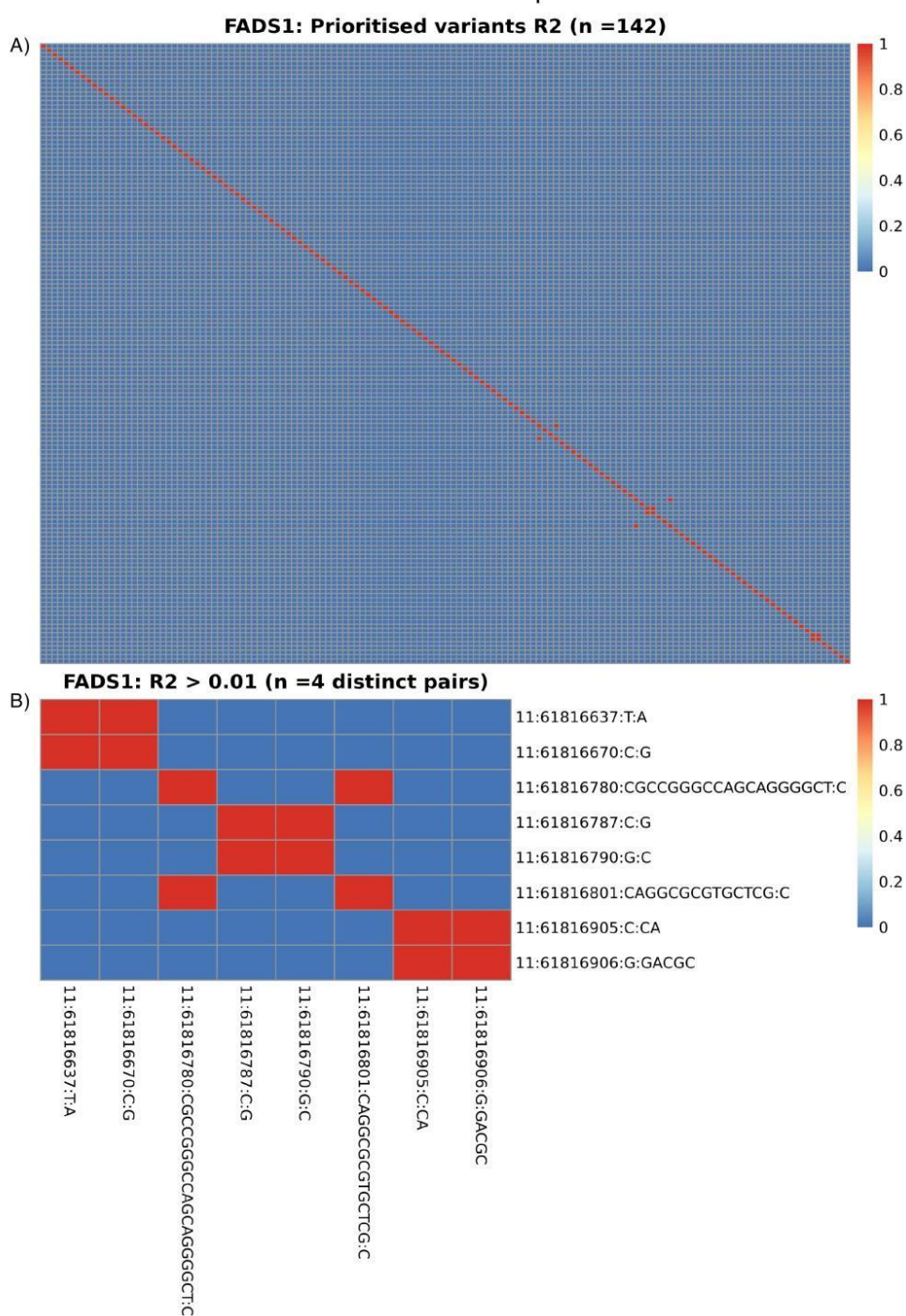


Supplementary Figure 16: Linkage Disequilibrium matrix of all prioritised variants in FADS3.

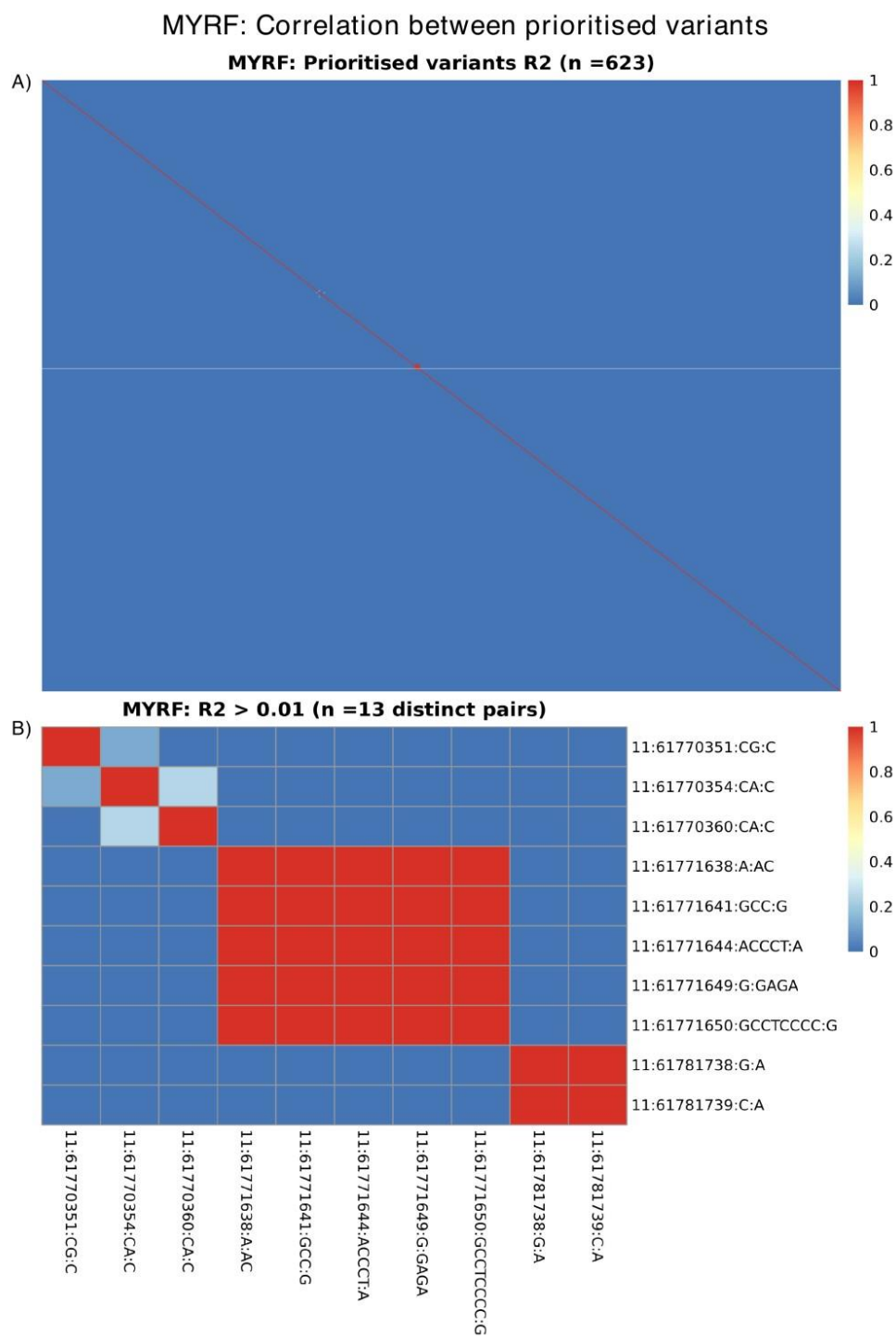


Supplementary Figure 17: Linkage Disequilibrium matrix of all prioritised variants in FEN1.

FADS1: Correlation between prioritised variants



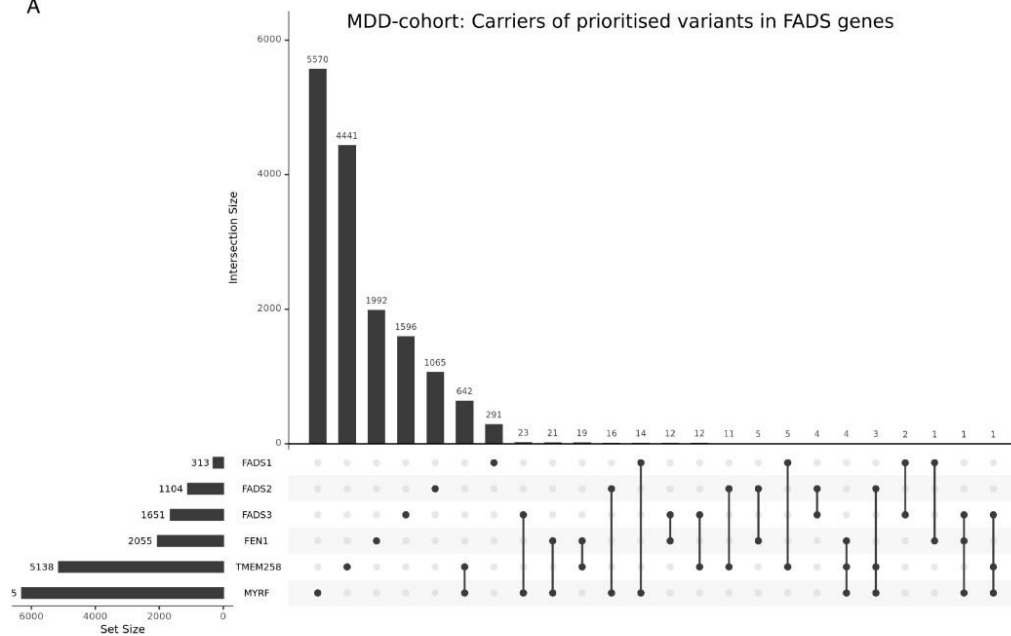
Supplementary Figure 18:Linkage Disequilibrium matrix of all prioritised variants in FADS1).



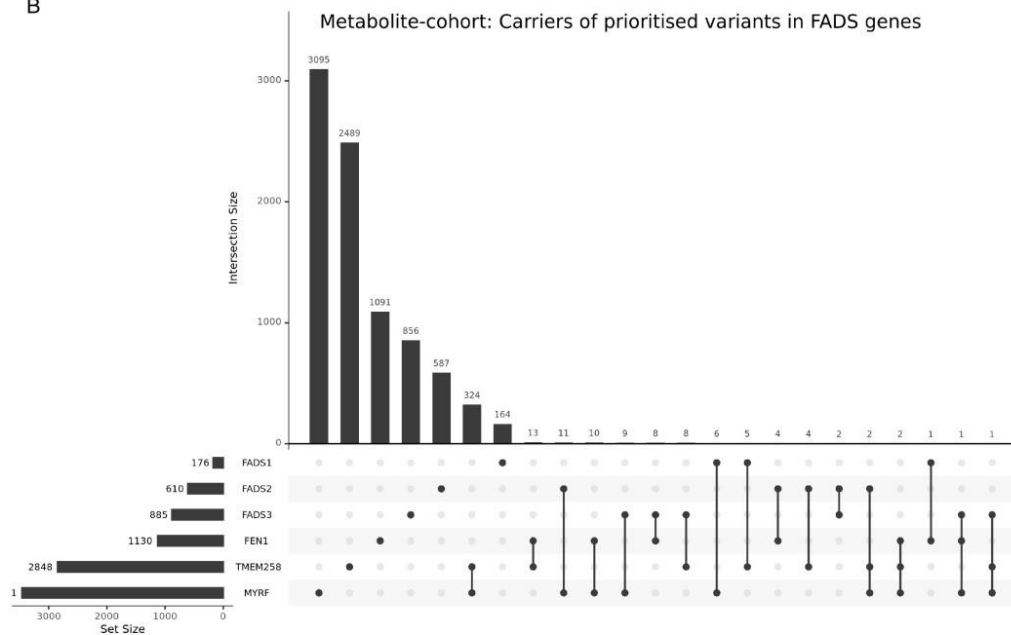
Supplementary Figure 19: Linkage disequilibrium matrix of all prioritized variants in MYRF.

UpSet plots of carriers of prioritised variants

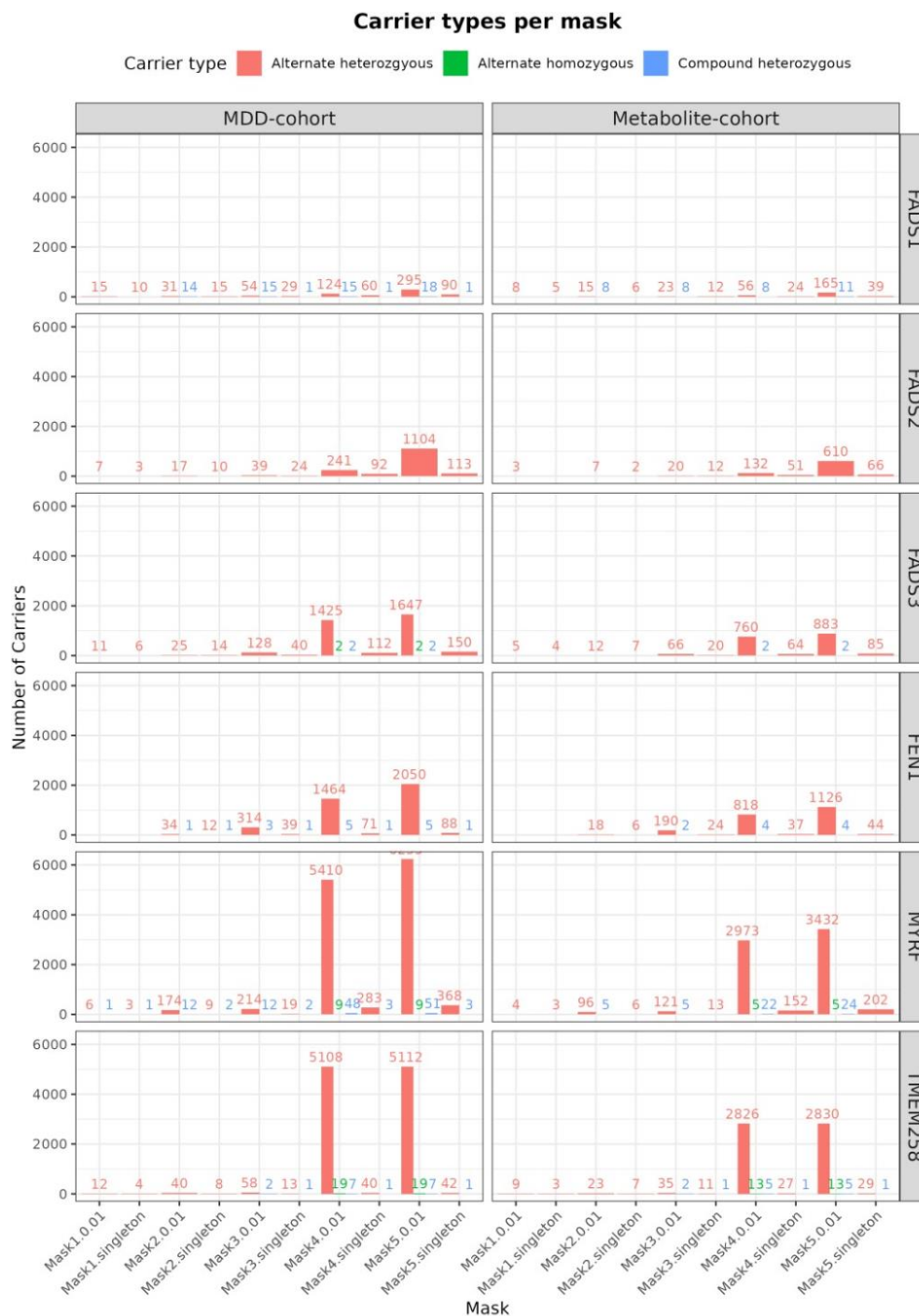
A



B



Supplementary Figure 20: UpSet plots showing the distribution of prioritised variants per carried by participants in the MDD-cohort (A) and Metabolite-cohort (B).



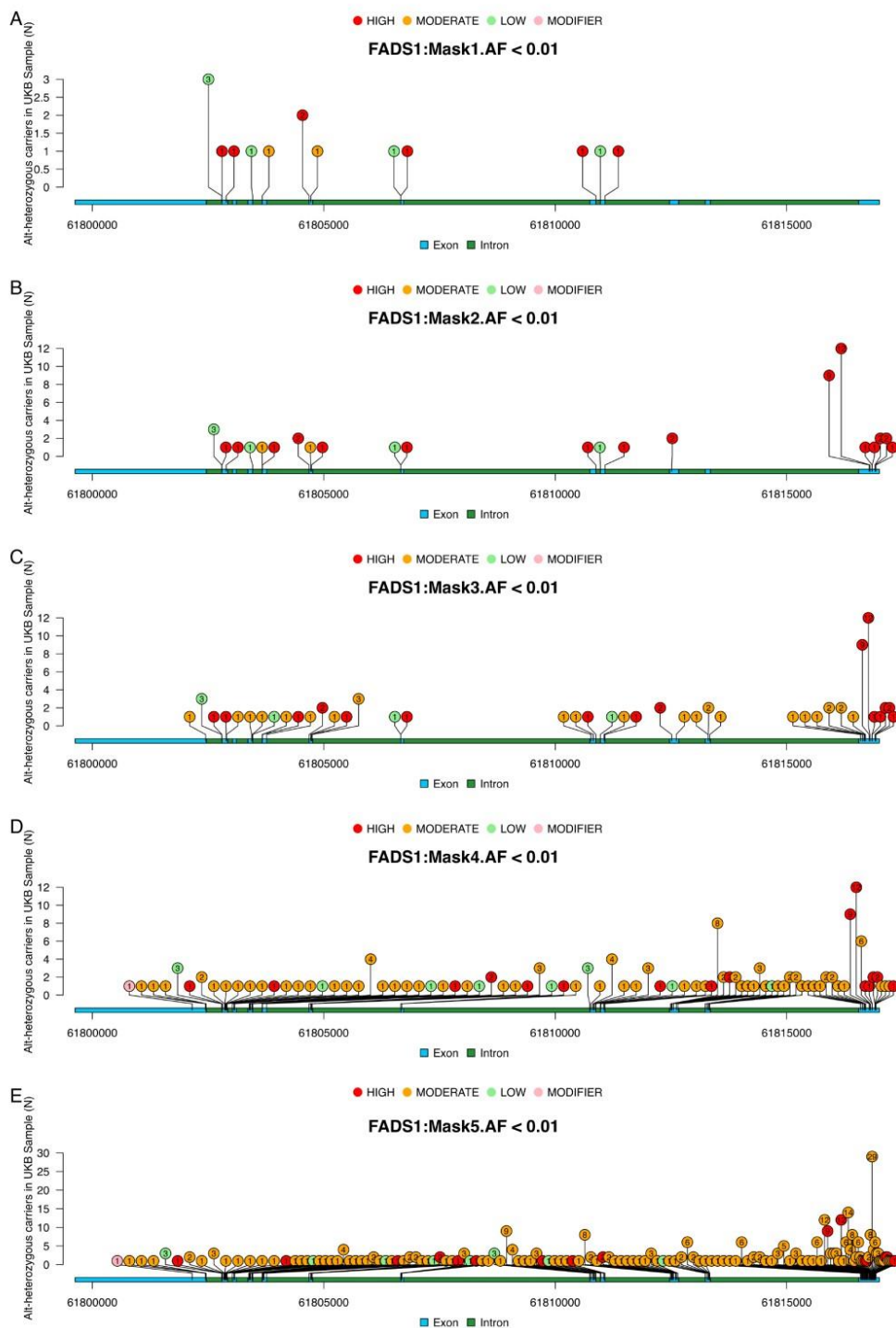
Supplementary Figure 21: Carriers of prioritised variants in each mask.

The number of carriers (yaxis) of prioritised variants within each mask (x-axis) in the MDD-cohort and the Metabolite-cohort (columns) for each gene (rows).

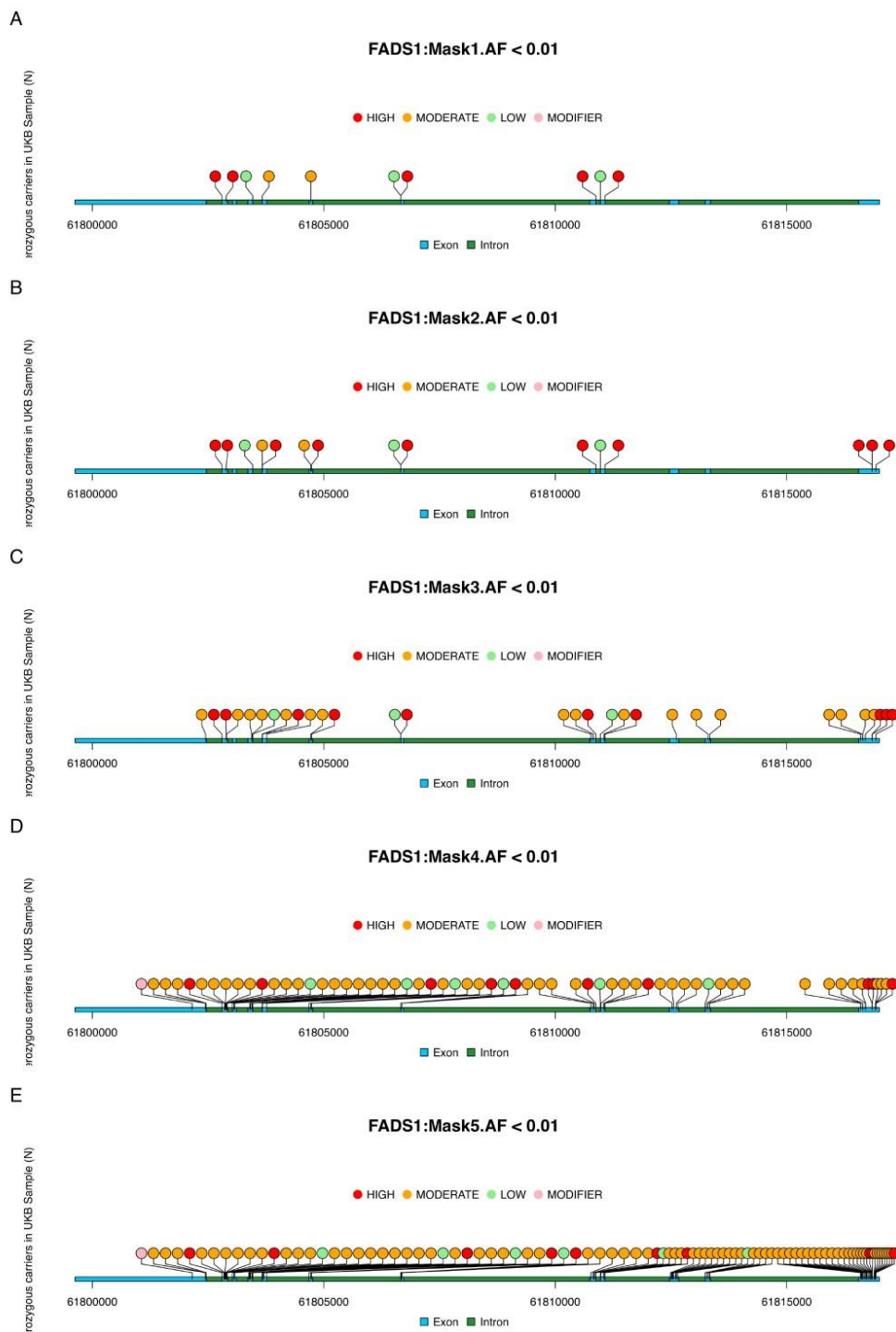
10.1.2.3 Variant masks

10.1.2.3.1 Mask-specific lolliplots

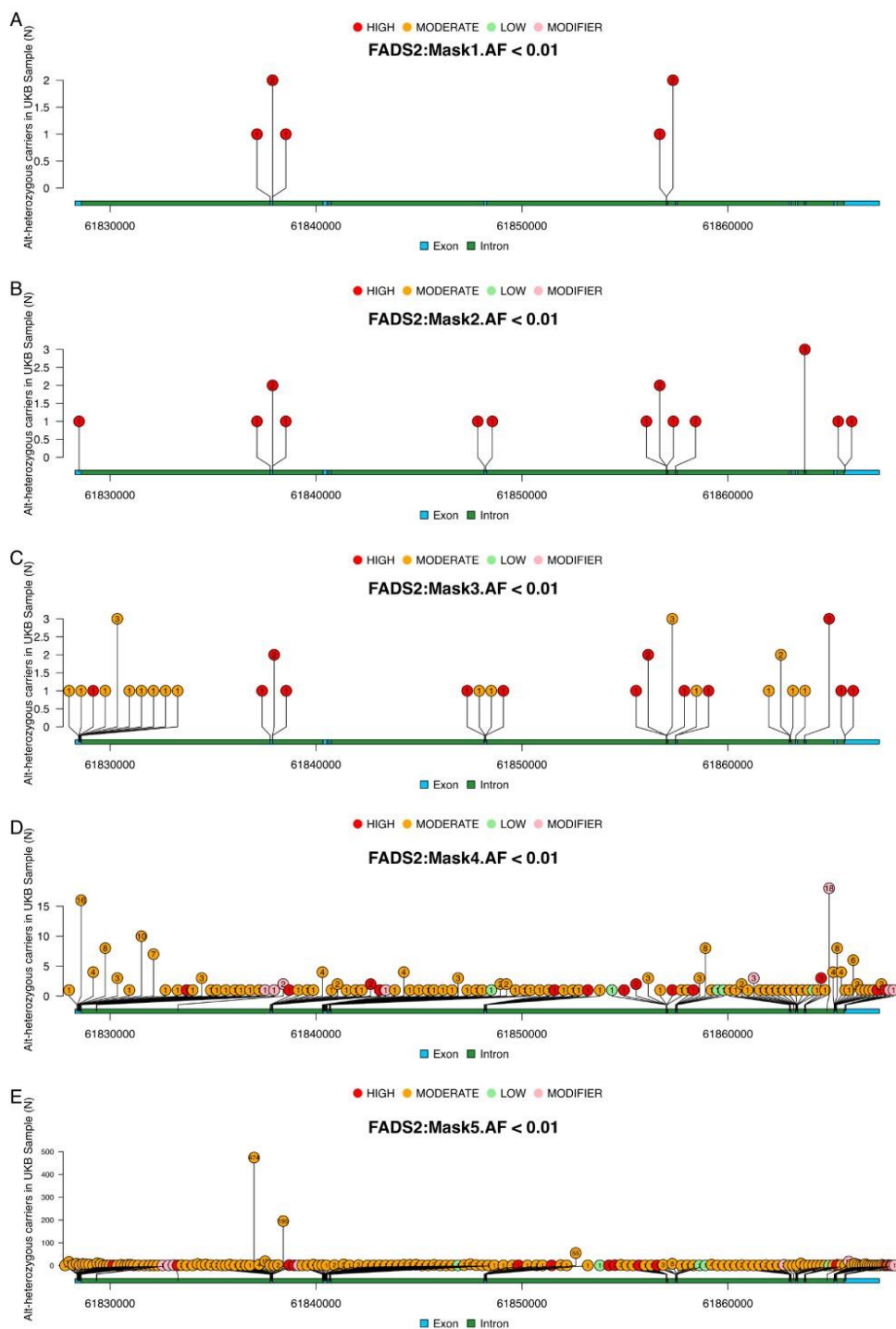
Supplementary Figures 22-33 are a series of subplots showing the location of either rare ($AF < 0.01$) or singleton prioritised variants in each gene. The subplots in each graph denote the specific prioritised variants within each Mask, specifically: Mask 1 (A), Mask 2 (B), Mask 3 (C), Mask 4 (D) and Mask 5 (E). Each variant is coloured by the predicted impact from VEP. The number of alternate heterozygous carriers for each variant is both represented by the height of the bar and is displayed in each lollipop. AF = Allele Frequency, VEP = Variant Effect Predictor.



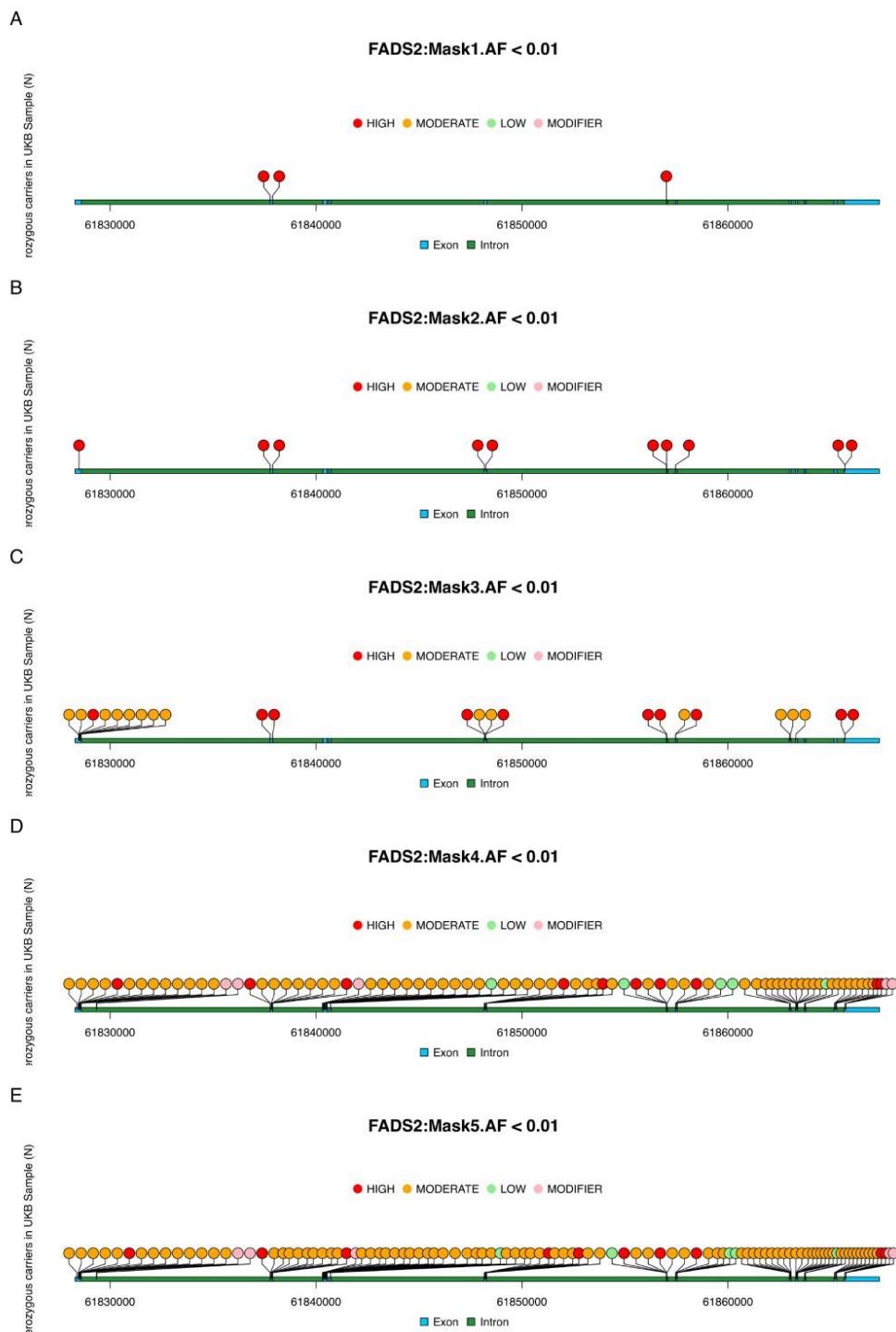
Supplementary Figure 22: FADS1 prioritised variants with AF < 0.01 in each mask.



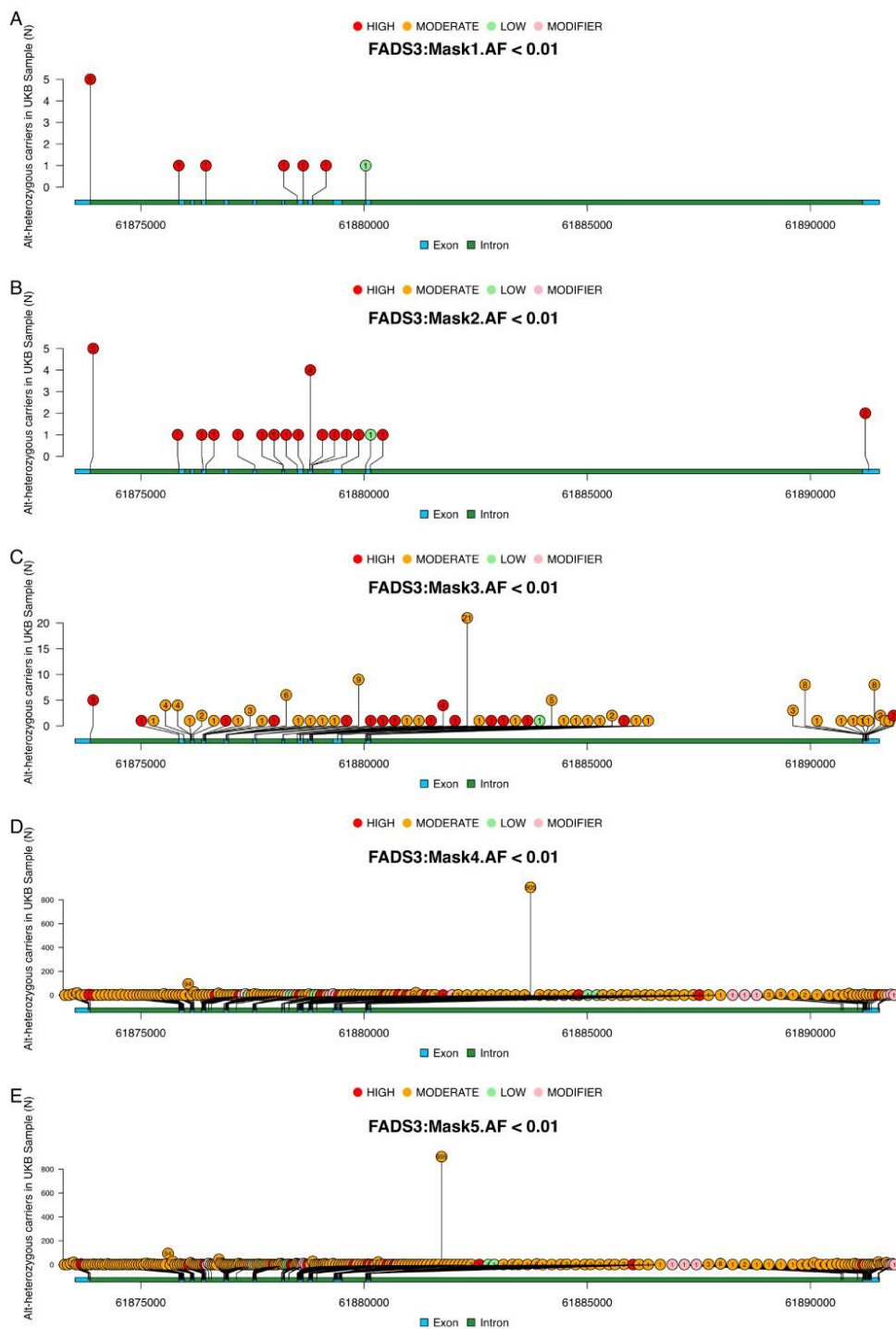
Supplementary Figure 23: FADS1 prioritised singleton variants in each mask.



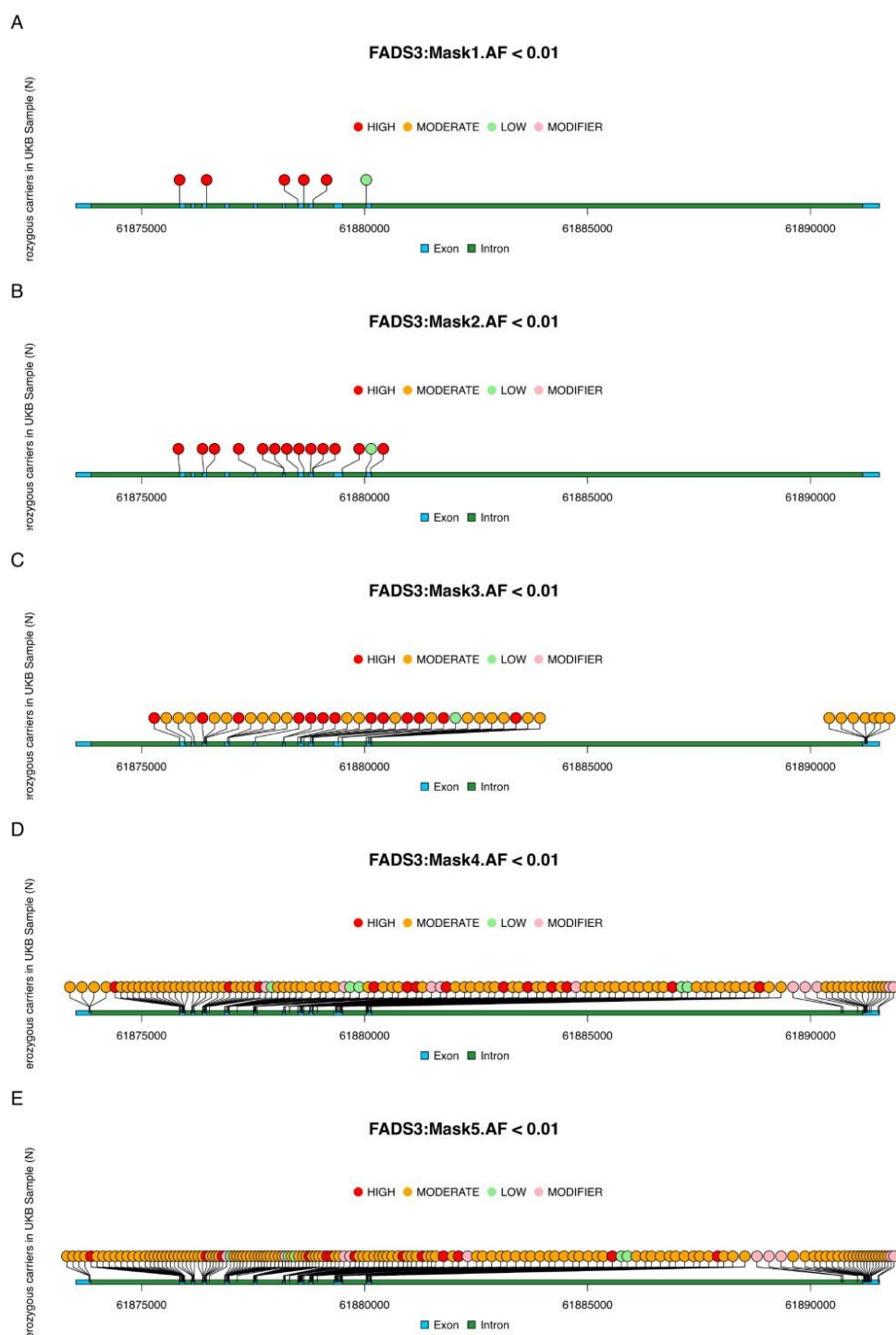
Supplementary Figure 24: FADS2 prioritised variants with AF < 0.01 in each mask.



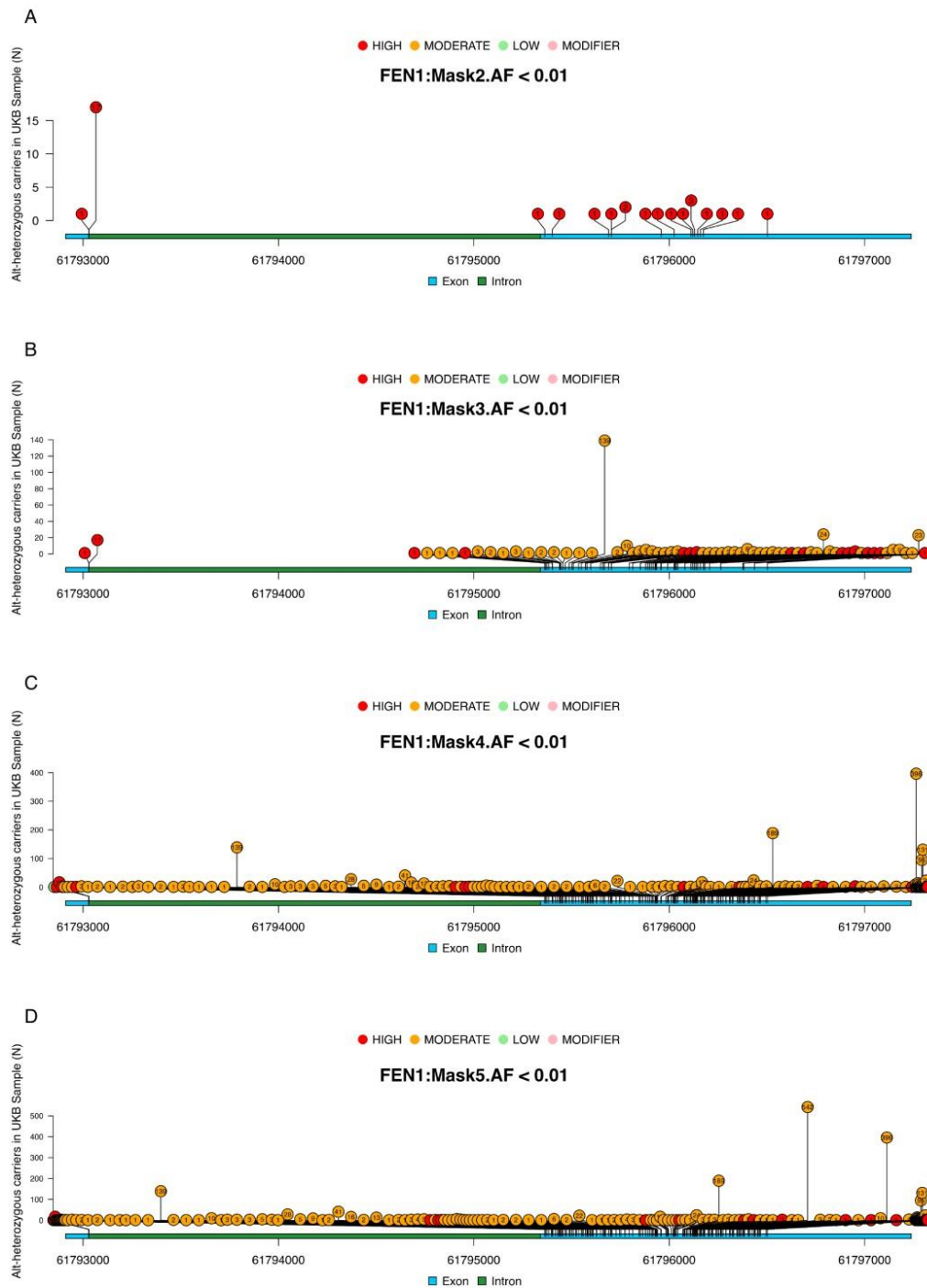
Supplementary Figure 25: FADS2 prioritised singleton variants in each mask.



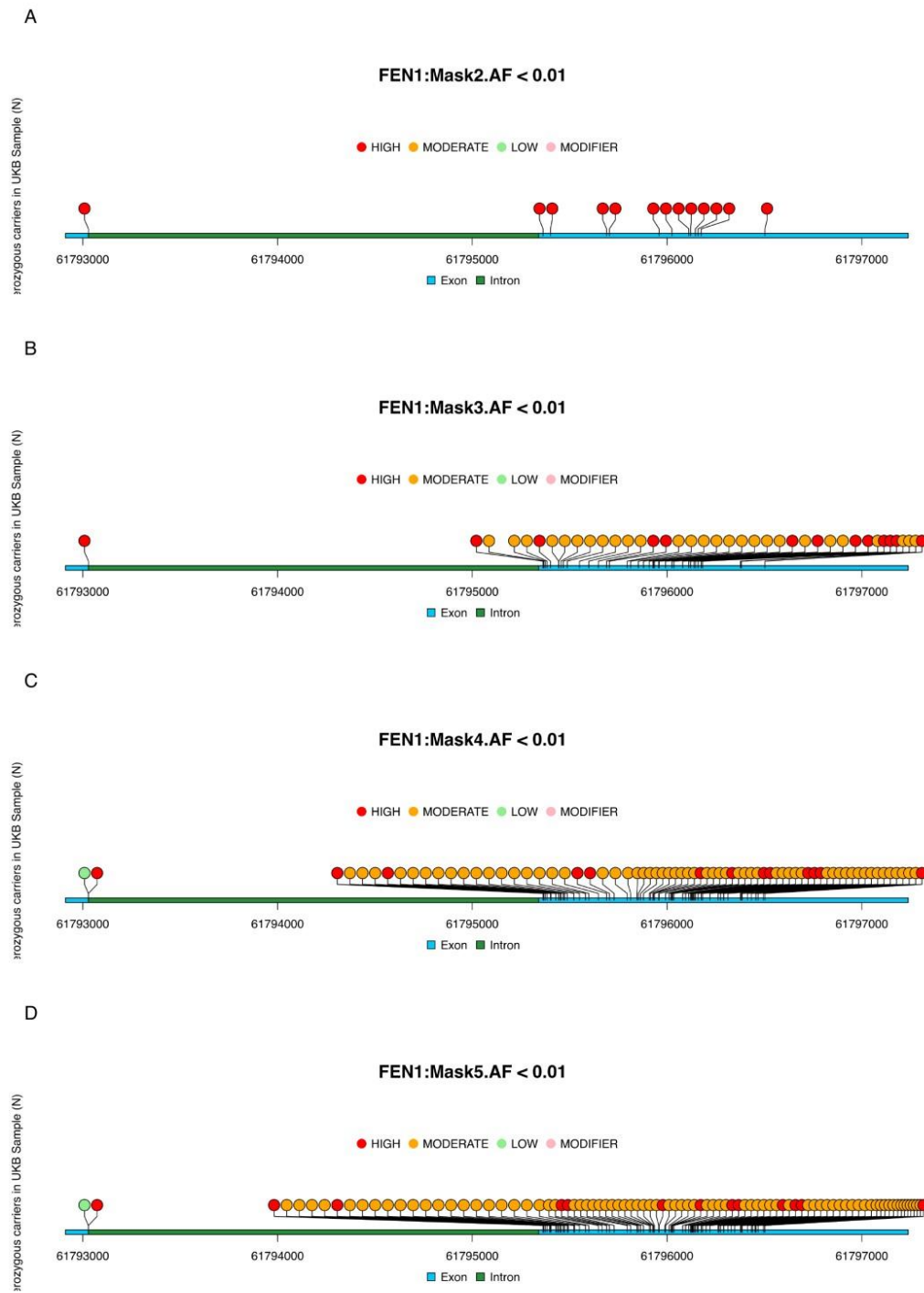
Supplementary Figure 26: FADS3 prioritised variants with AF < 0.01 in each mask.



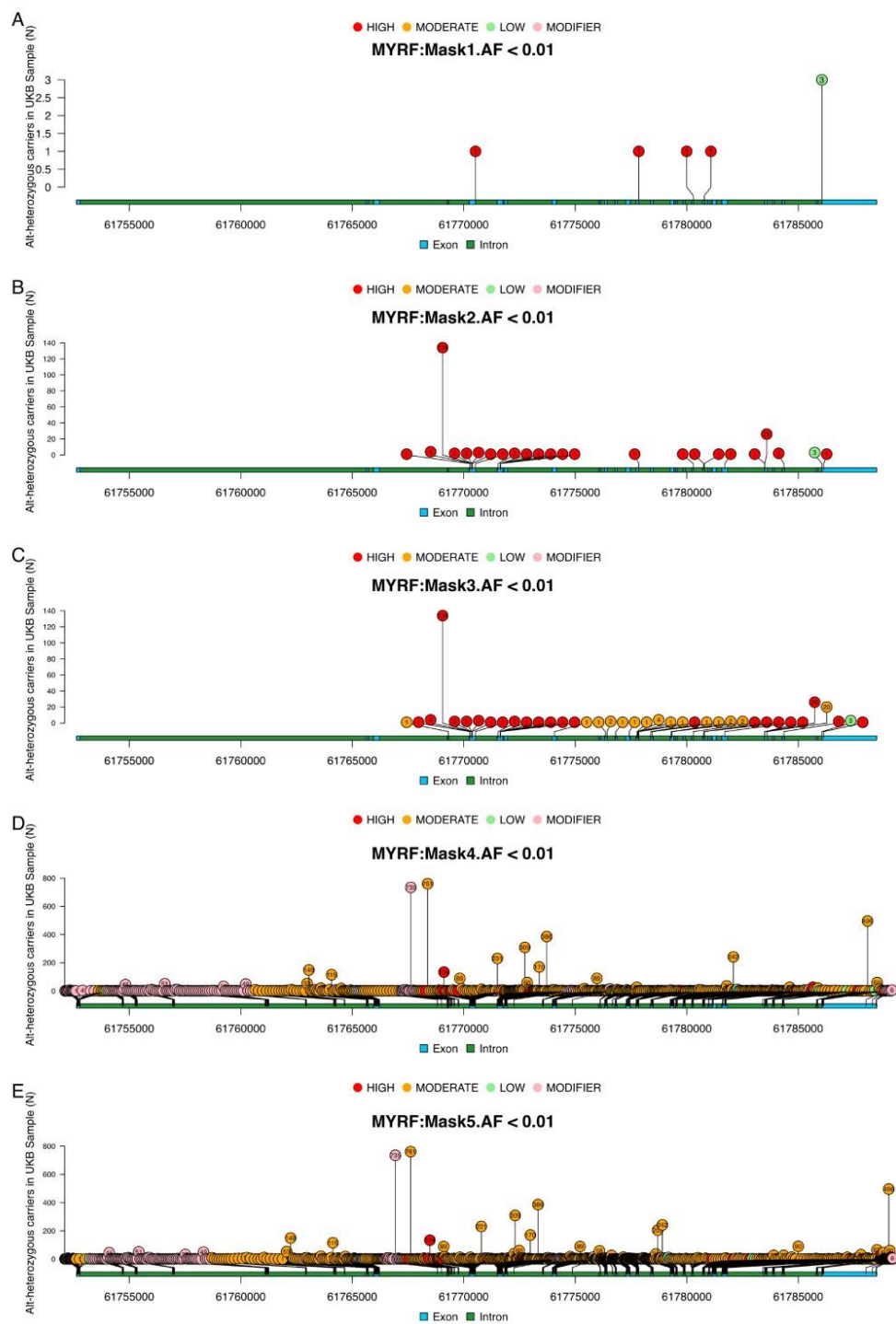
Supplementary Figure 27: FADS3 prioritised singleton variants in each mask.



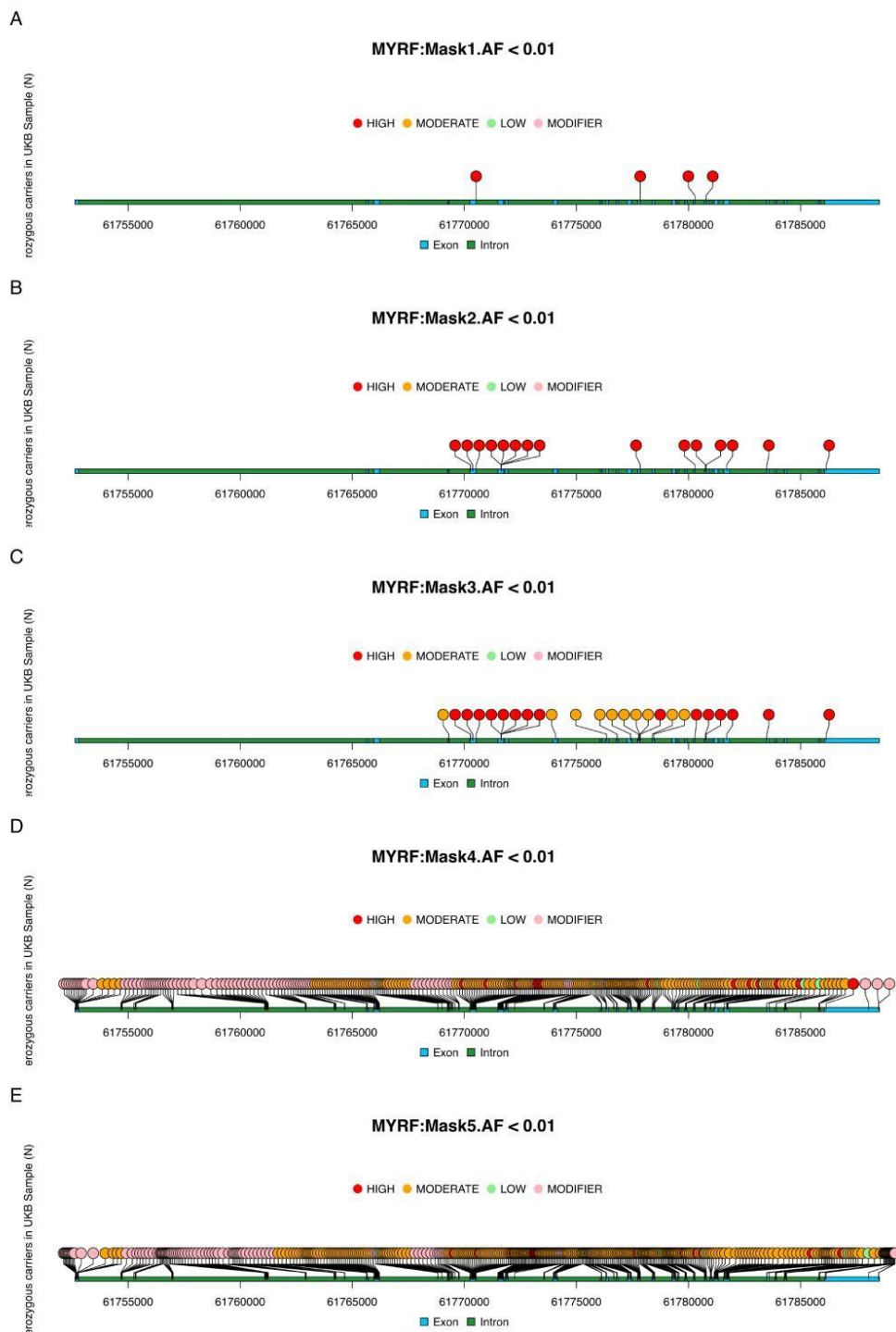
Supplementary Figure 28: FEN1 prioritised variants with AF < 0.01 in each mask.



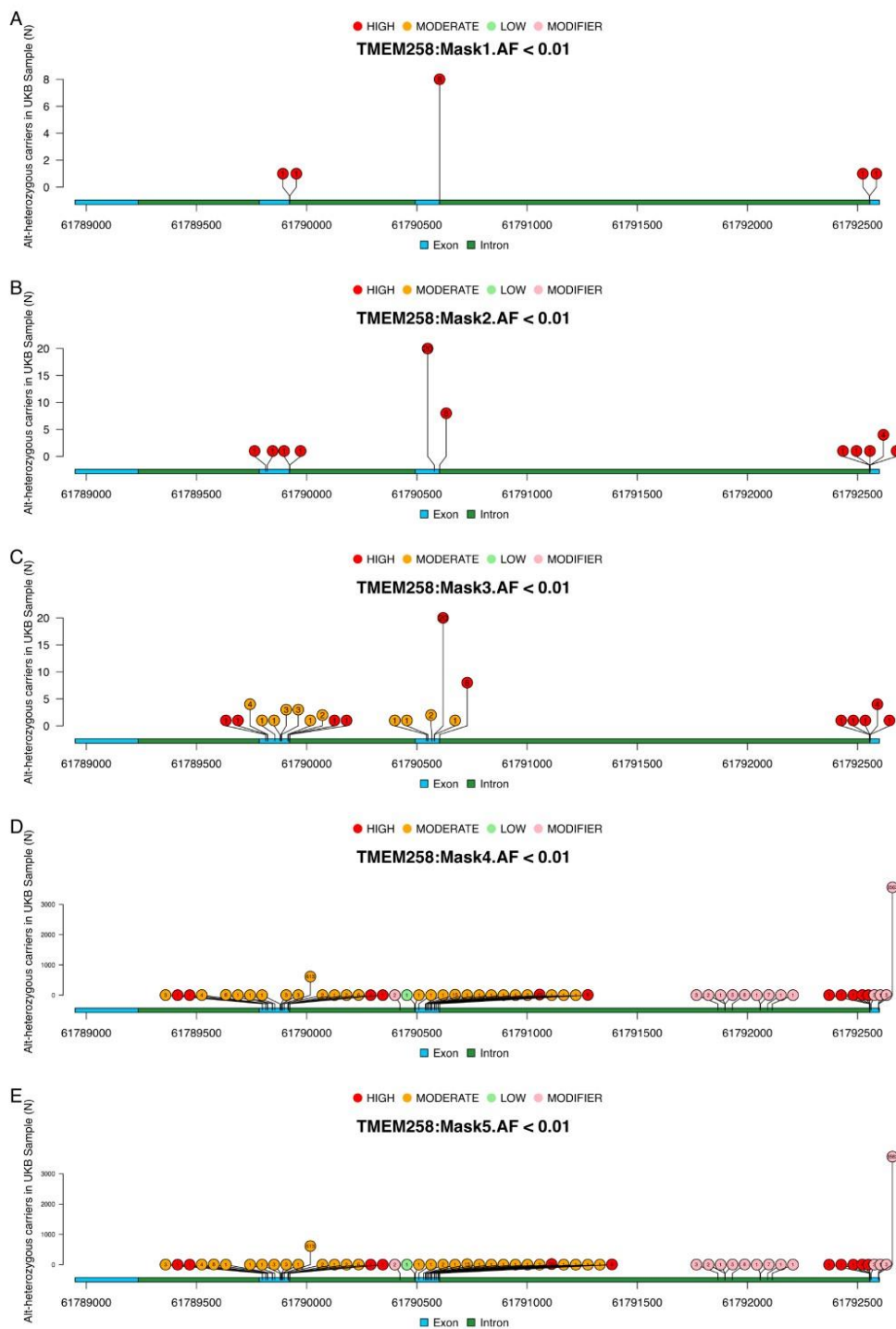
Supplementary Figure 29: FEN1 prioritised singleton variants in each mask.



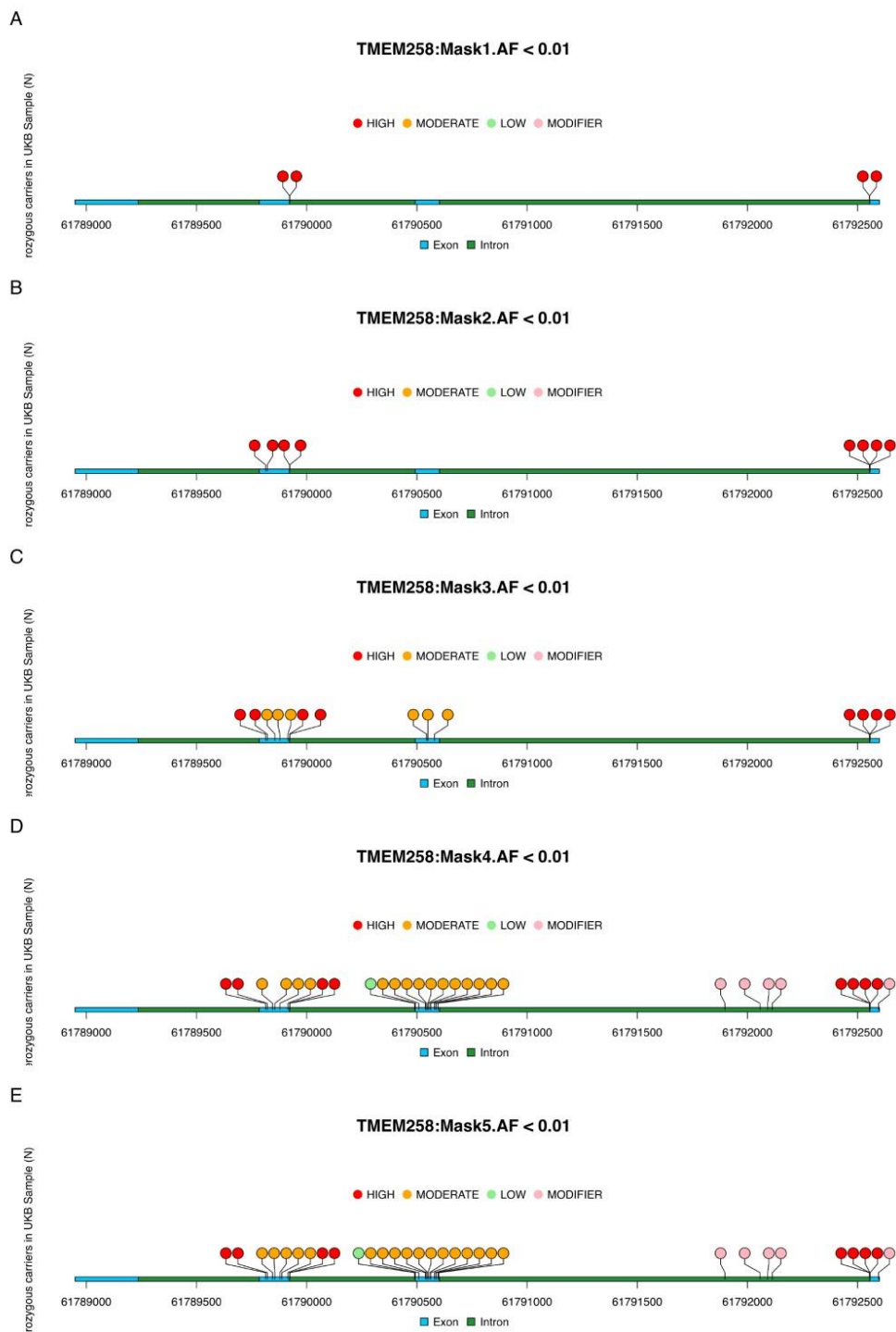
Supplementary Figure 30: MYRF prioritised variants with AF < 0.01 in each mask.



Supplementary Figure 31: MYRF prioritised singleton variants in each mask.

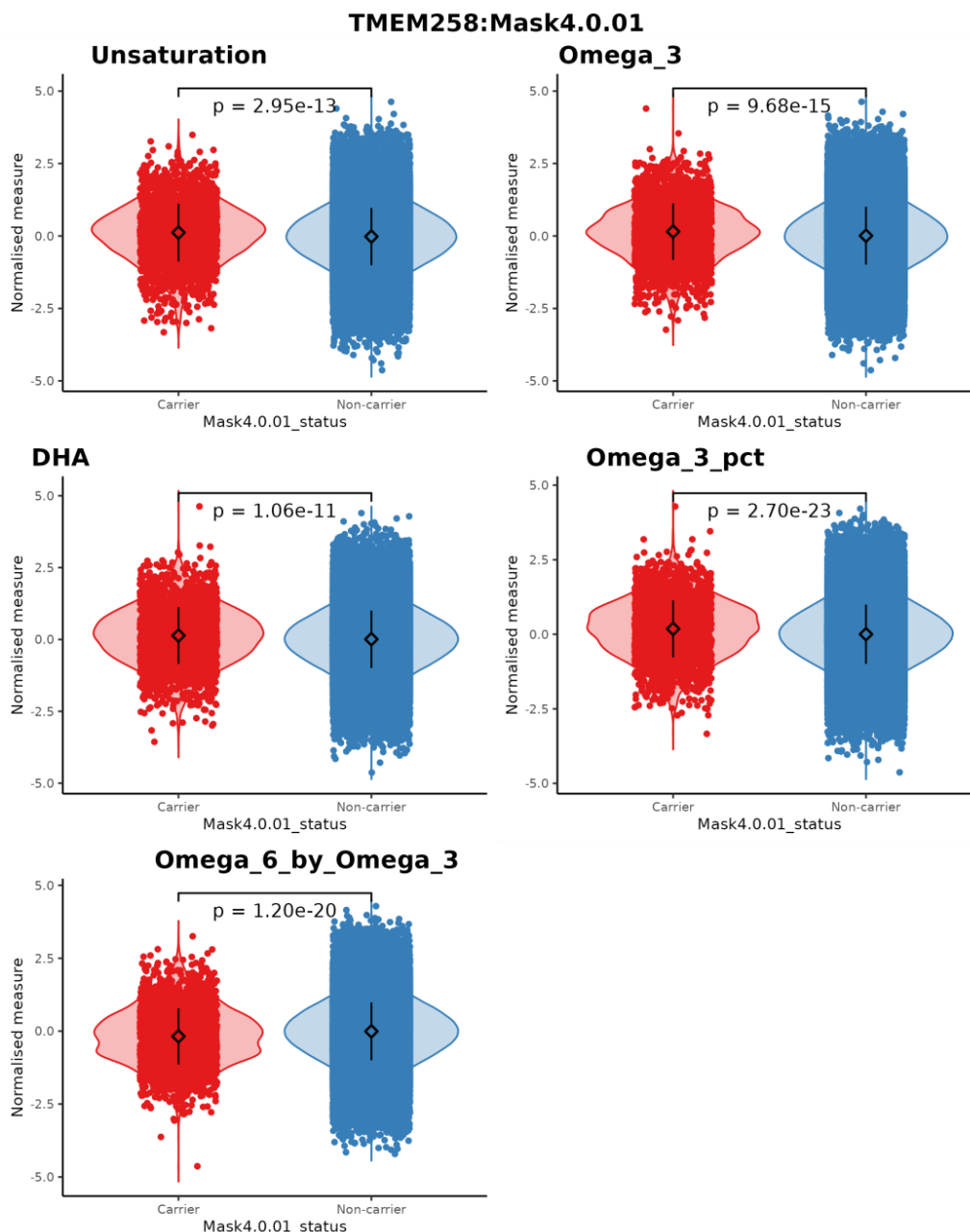


Supplementary Figure 32: TMEM258 prioritised variants with AF < 0.01 in each mask.



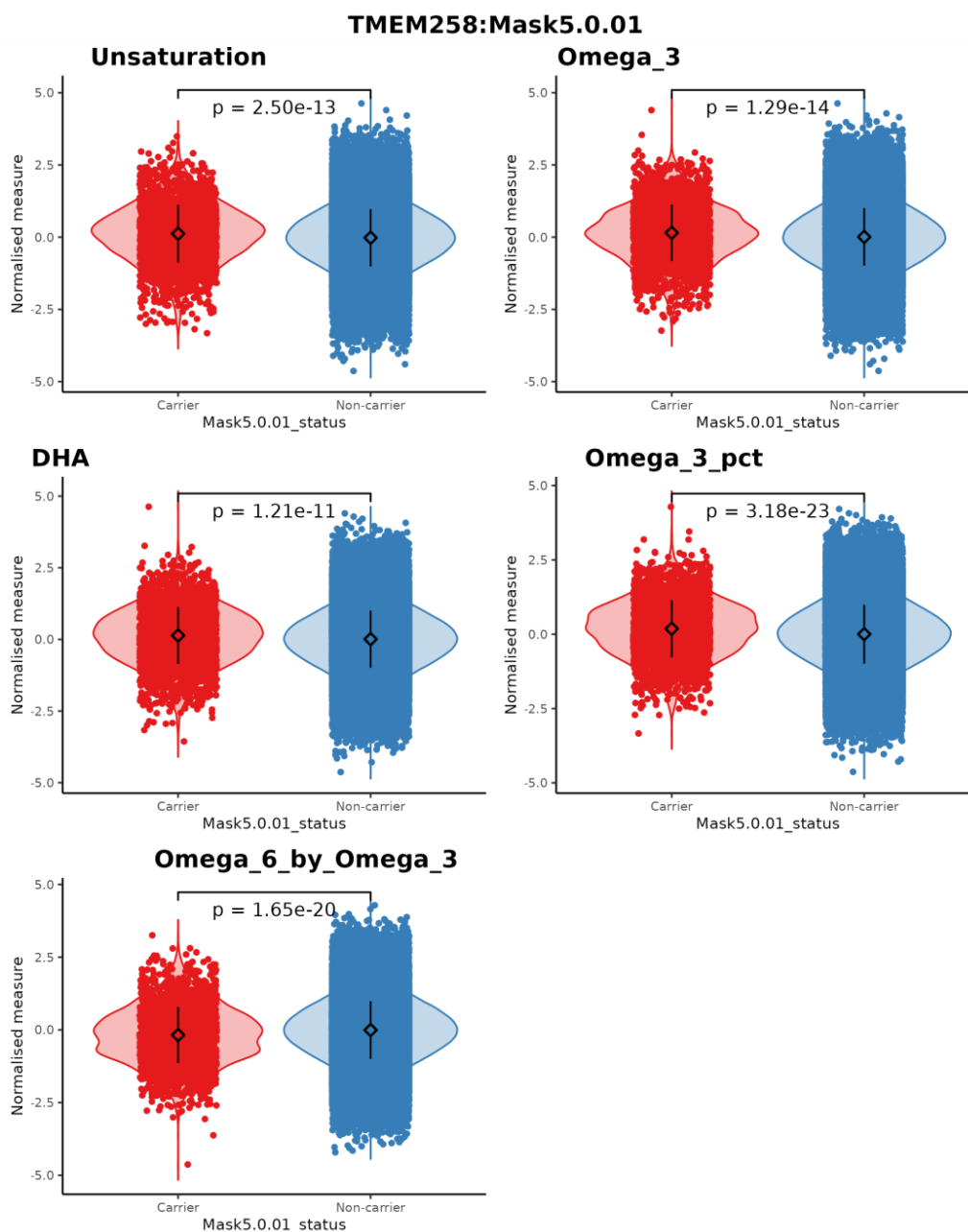
Supplementary Figure 33: TMEM258 prioritised singleton variants in each mask.

10.1.2.4 Statistical tests



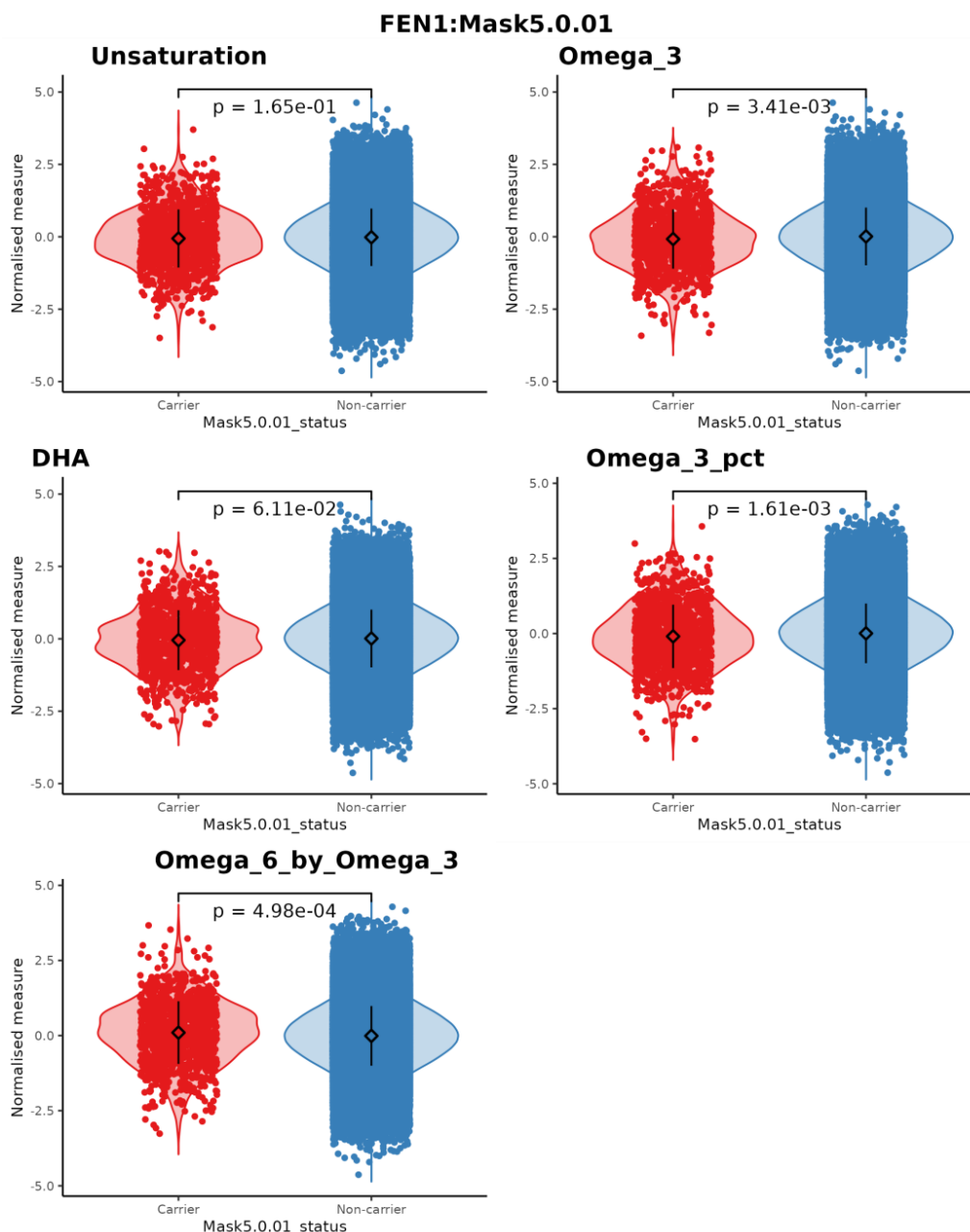
Supplementary Figure 34: Violin plots of each metabolite trait in carriers and non-carriers of prioritised variants in TMEM258 assigned to Mask 4.

The distribution of each metabolite trait in the carriers and non-carriers of TMEM258- Mask 4 (AF < 0.01) prioritised variants. The statistical significance of the difference in distributions was tested by Welch's two-tailed t-test, with the p value shown on each graph.



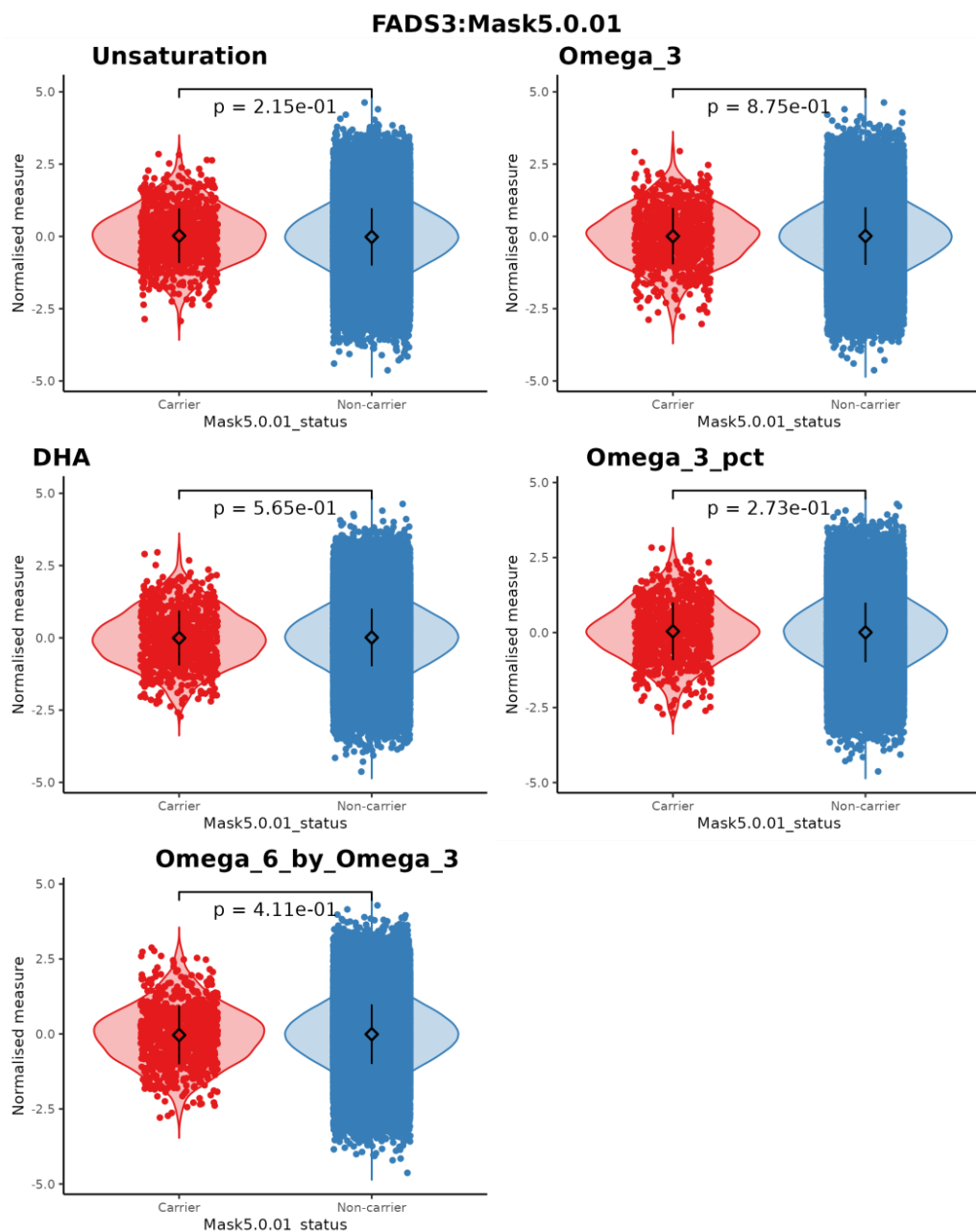
Supplementary Figure 35: Violin plots of each metabolite trait in carriers and non-carriers of prioritised variants in TMEM258 assigned to Mask 5 (AF < 0.01).

The distribution of each metabolite trait in the carriers and non-carriers of TMEM258 Mask 5 (AF < 0.01) prioritised variants. The statistical significance of the difference in distributions was tested by Welch's two-tailed t-test, with the p value shown on each graph.



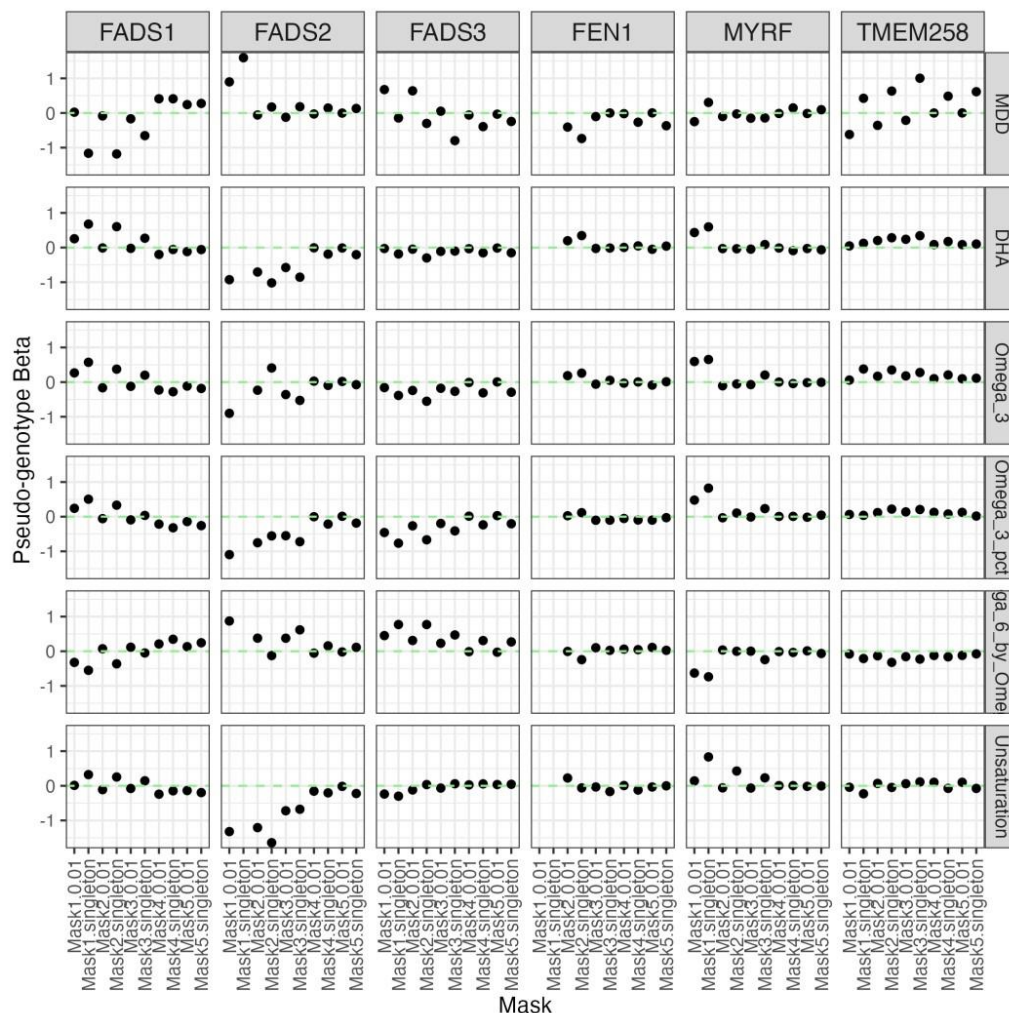
Supplementary Figure 36: Violin plots of each metabolite trait in carriers and non-carriers of prioritised variants in FEN1 assigned to Mask 5 (AF < 0.01).

The distribution of each metabolite trait in the carriers and non-carriers of FEN1 Mask 5 (AF < 0.01) prioritised variants. The statistical significance of the difference in distributions was tested by Welch's two-tailed t-test, with the p value shown on each graph.



Supplementary Figure 37: Violin plots of each metabolite trait in carriers and non-carriers of prioritised variants in FADS3 assigned to Mask 3 ($AF < 0.01$).

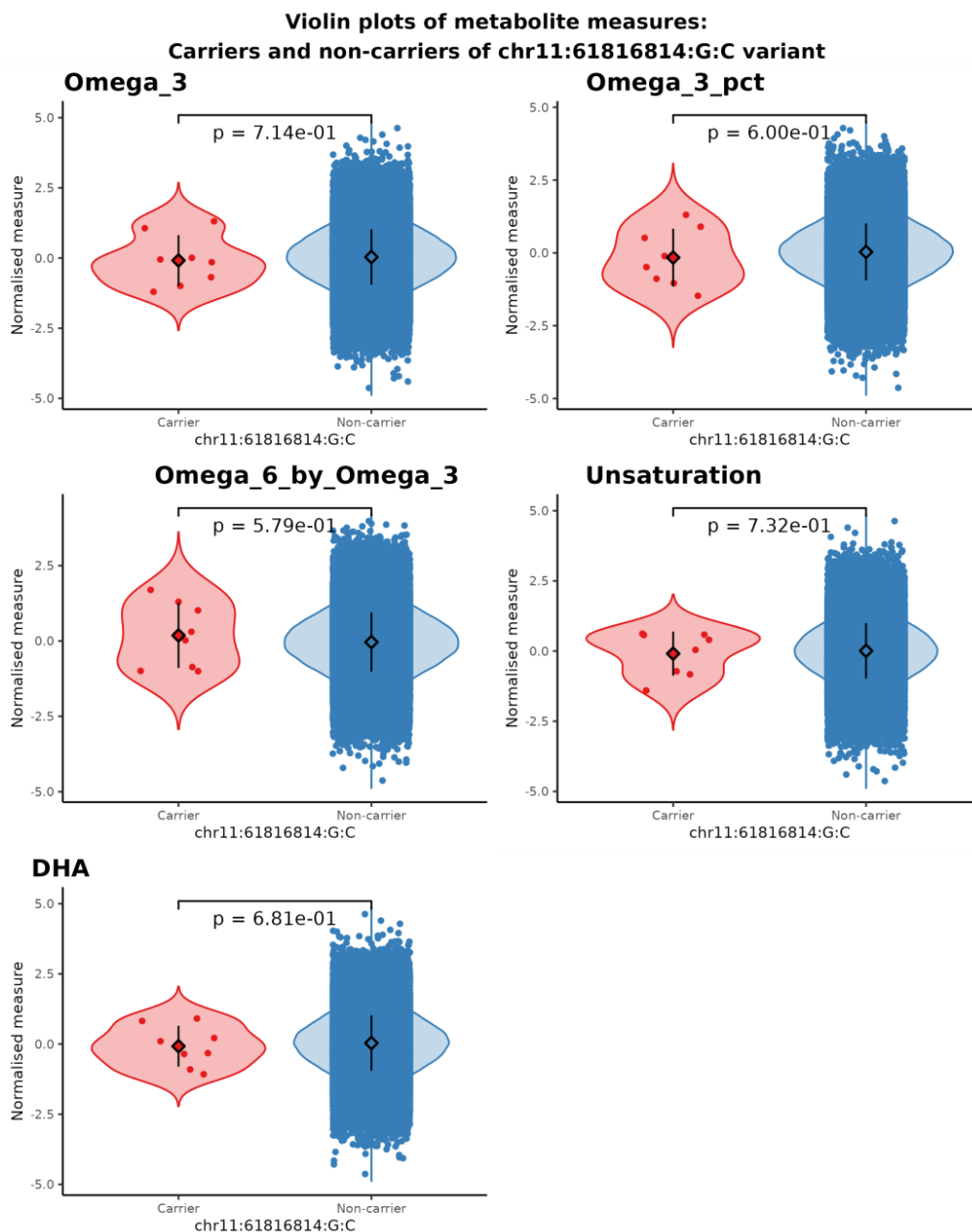
The distribution of each metabolite trait in the carriers and non-carriers of FADS3 Mask 5 ($AF < 0.01$) prioritised variants. The statistical significance of the difference in distributions was tested by Welch's two-tailed t-test, with the p value shown on each graph.



Supplementary Figure 38: The effect size of the pseudo-genotype on each trait for each mask in the traditional burden test.

The effect size of the pseudo genotype (y-axis) comprised from each mask (x-axis), split by the gene (columns) and trait (rows). The green line denotes an effect size of 0.

10.1.2.5 Leave one variant out (LOVO) analysis

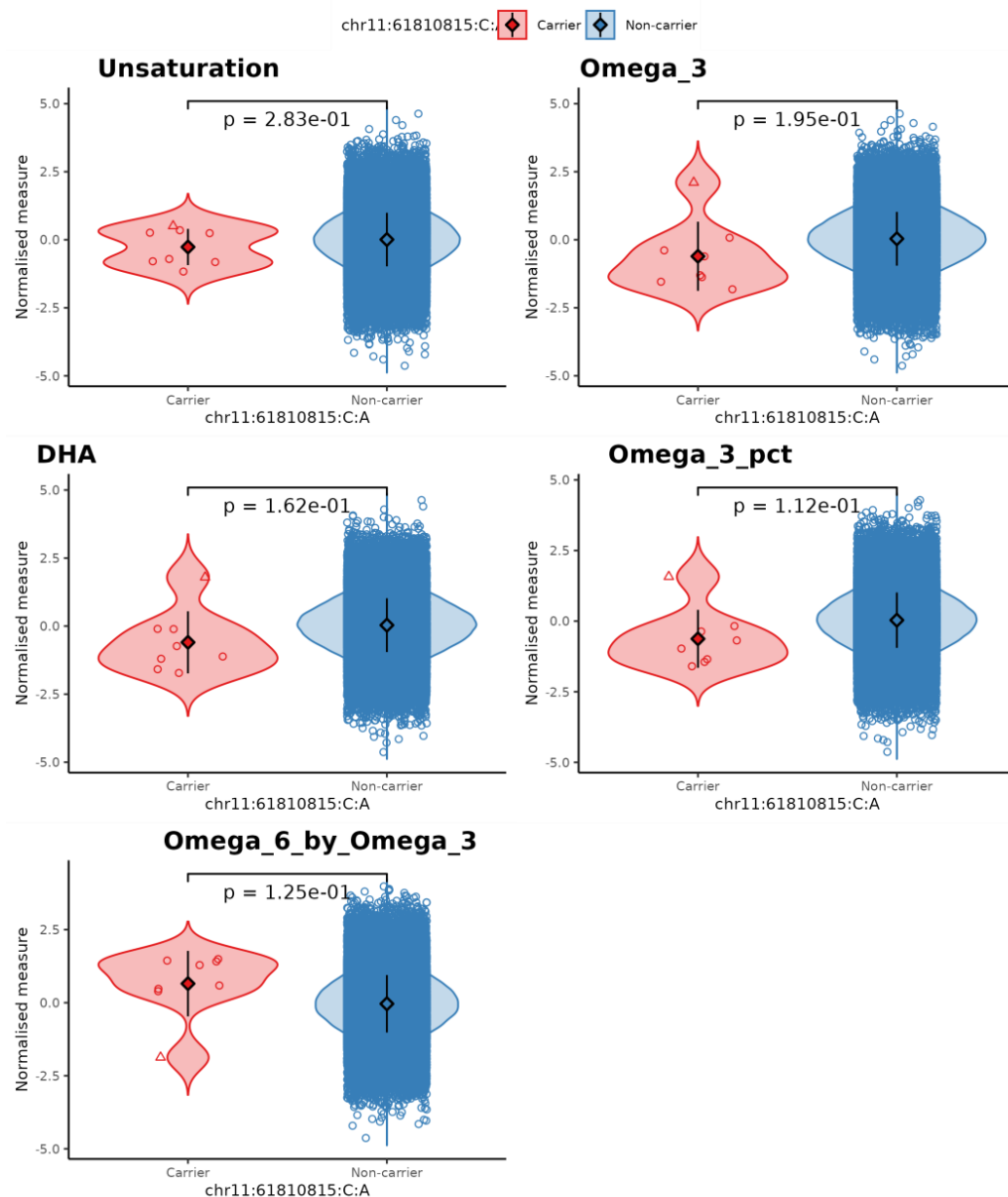


Supplementary Figure 39: Violin plots of each metabolite trait in carriers and non-carriers of chr11:61816814:G:C variant.

The distribution of each metabolite trait in the carriers and non-carriers of chr11:61816814:G:C. This variant was the main driver of the nominal significant burden between prioritised variants in

FADS1 and MDD. The statistical significance of the difference in distributions was tested by Welch's two-tailed t-test, with the p value shown on each graph.

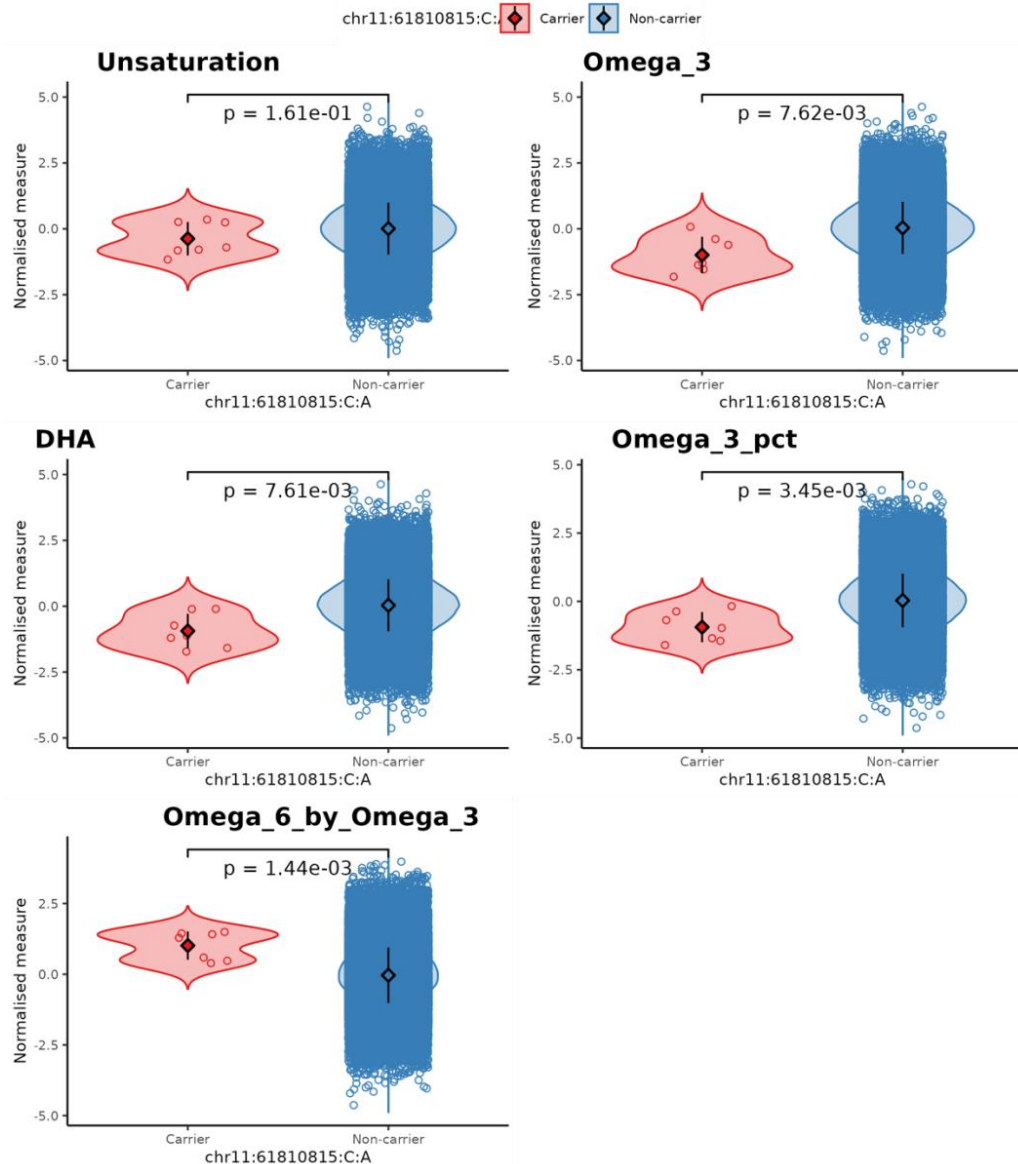
Violin plots of metabolite measures: Carriers and non-carriers of chr11:61810815:C:A variant



Supplementary Figure 40: Violin plots of each metabolite trait in carriers and non-carriers of chr11:61810815:C:A variant.

The distribution of each metabolite trait in the carriers and non-carriers of chr11:61810815:C:A. This variant was the main driver of the nominal significant burden between prioritised variants in FADS1 and each metabolite trait. The statistical significance of the difference in distributions was tested by Welch's two-tailed t-test, with the p value shown on each graph. An outlying sample was identified in all metabolites (except the Degree of Unsaturation), which is shown with a triangle on each graph.

**Violin plots of metabolite measures:
Carriers and non-carriers of chr11:61810815:C:A variant
(After removing outlying sample based on IQR)**



Supplementary Figure 41: Violin plots of each metabolite trait in carriers and non-carriers of chr11:61810815:C:A variant (Excluding the outlying sample).

The distribution of each metabolite trait in the carriers and non-carriers of chr11:61810815:C:A. This variant was the main driver of the nominal significant burden between prioritised variants in FADS1 and each metabolite trait. The statistical significance of the difference in distributions was tested by Welch's two-tailed t-test, with the p value shown on each graph. An outlying sample was

identified in all metabolites (except the Degree of Unsaturation), which has been removed in this graph.

10.2 Supplementary Tables

Supplementary Table 1: Genes tested in the FADS cluster.	384
Supplementary Table 2: UKB pVCF files used to generate the gene-level pVCF.	384
Supplementary Table 3: The variant pairs in FADS1, FEN1 and MYRF which exhibited a correlation (R^2) > 0.01.....	385
Supplementary Table 4: Demographic variables for the carriers and non-carriers of prioritised variants in the FADS genes (Gene) in both the MDD and metabolite cohorts (Cohort).....	386
Supplementary Table 5: The association between carrying rare prioritised variants in FADS genes and MDD.....	389
Supplementary Table 6: Significant associations between carrying rare prioritised variants in FADS genes and metabolite traits.	391
Supplementary Table 7: Joint burden tests across all masks per gene with MDD and each of the metabolomic traits.	392
Supplementary Table 8: Welchs t-test results for the association between chr11:61861814:G:C carrier status and metabolite levels.	395
Supplementary Table 9: The demographics of the carriers and non-carriers of the variant chr11_61810815_C_A in the Metabolite cohort.....	396
Supplementary Table 10: Welchs t-test results for the association between chr11:61810815:C:A carrier status and metabolite levels.	397
Supplementary Table 11: The distribution of metabolite traits in the carriers of the chr11:61810815:C:A variant.	398
Supplementary Table 12: Welchs t-test results for the association between chr11:61810815:C:A carrier status and metabolite levels after excluding the outlying sample.	399

10.2.1 Methods

Supplementary Table 1: Genes tested in the FADS cluster.

The start position, end position, strand and canonical transcript used for each gene investigated in **Chapter 4**.

Gene	chr	Start	End	Strand	ensembl_gene_id	Transcript	transcript_is_canonical
FADS1	chr11	61,799,627	61,829,318	-1	ENSG00000149485	ENST00000350997	1
FADS2	chr11	61,792,980	61,867,354	1	ENSG00000134824	ENST00000278840	1
FADS3	chr11	61,873,519	61,892,051	-1	ENSG00000221968	ENST00000278829	1
FEN1	chr11	61,792,911	61,797,238	1	ENSG00000168496	ENST00000305885	1
MYRF	chr11	61,752,636	61,788,518	1	ENSG00000124920	ENST00000278836	1
TMEM258	chr11	61,768,501	61,792,802	-1	ENSG00000134825	ENST00000537328	1

Supplementary Table 2: UKB pVCF files used to generate the gene-level pVCF.

Gene	UKB_files
FADS1	ukb24310_c11_b3090_v1.vcf.gz:ukb24310_c11_b3091_v1.vcf.gz
FADS2	ukb24310_c11_b3089_v1.vcf.gz:ukb24310_c11_b3090_v1.vcf.gz:ukb24310_c11_b3091_v1.vcf.gz:ukb24310_c11_b3092_v1.vcf.gz:ukb24310_c11_b3093_v1.vcf.gz
FADS3	ukb24310_c11_b3093_v1.vcf.gz:ukb24310_c11_b3094_v1.vcf.gz
FEN1	ukb24310_c11_b3089_v1.vcf.gz
MYRF	ukb24310_c11_b3087_v1.vcf.gz:ukb24310_c11_b3088_v1.vcf.gz:ukb24310_c11_b3089_v1.vcf.gz
TMEM258	ukb24310_c11_b3088_v1.vcf.gz:ukb24310_c11_b3089_v1.vcf.gz

10.2.2 Results

Supplementary Table 3: The variant pairs in FADS1, FEN1 and MYRF which exhibited a correlation (R^2) > 0.01.

Correlation was computed using the PLINK --r2 flag.

Gene	Variant 1	Variant 2	R2
FADS1	11:61816637:T:A	11:61816670:C:G	1.000000
	11:61816780:CGCCGGGCCAGCAGGGGCT:C	11:61816801:CAGGCGCGTGCTCG:C	1.000000
	11:61816787:C:G	11:61816790:G:C	1.000000
	11:61816905:C:CA	11:61816906:G:GACGC	1.000000
FEN1	11:61796501:A:AAAGG	11:61796502:T:G	1.000000
MYRF	11:61770351:CG:C	11:61770354:CA:C	0.124996
	11:61770354:CA:C	11:61770360:CA:C	0.249997
	11:61771638:A:AC	11:61771641:GCC:G	1.000000
	11:61771638:A:AC	11:61771644:ACCCT:A	1.000000
	11:61771638:A:AC	11:61771649:G:GAGA	1.000000
	11:61771638:A:AC	11:61771650:GCCTCCCC:G	1.000000
	11:61771641:GCC:G	11:61771644:ACCCT:A	1.000000
	11:61771641:GCC:G	11:61771649:G:GAGA	1.000000
	11:61771641:GCC:G	11:61771650:GCCTCCCC:G	1.000000
	11:61771644:ACCCT:A	11:61771649:G:GAGA	1.000000
	11:61771644:ACCCT:A	11:61771650:GCCTCCCC:G	1.000000
	11:61771649:G:GAGA	11:61771650:GCCTCCCC:G	1.000000
	11:61781738:G:A	11:61781739:C:A	1.000000

Supplementary Table 4: Demographic variables for the carriers and non-carriers of prioritised variants in the FADS genes (Gene) in both the MDD and metabolite cohorts (Cohort).

Continuous traits (Age, BMI and SES) are displayed as the mean values within the cohorts whilst categorical traits are displayed as the counts and % of the sample. Smoking status shows the ratio of current, previous and never smokers, Ethnicity shows the ratio of White, Mixed and Other populations, and MDD shows the ratio of MDD cases and controls within the sample.

Gene	Cohort	Carrier status	Total	Age	BMI	SES	% Female	% University/College	Smoking status	Ethnicity	MDD
FADS1	MDD cohort	Carrier	313	57.7 (7.75)	27.3 (4.45)	-1.67 (2.92)	159 (50.8%)	37 (11.8%)	29 (9.27%):116 (37.1%):166 (53%)	311 (99.4%):0 (0%):0 (0%)	55 (17.6%):258 (82.4%)
		Non-carrier	222,961	56.8 (8)	27.4 (4.75)	-1.49 (2.98)	120373 (54%)	25096 (11.3%)	22732 (10.2%):79008 (35.4%):120328 (54%)	221562 (99.4%):291 (0.131%):498 (0.223%)	33782 (15.2%):189179 (84.8%)
	Metabolite cohort	Carrier	176	58.1 (7.5)	27.5 (4.6)	-1.59 (2.95)	87 (49.4%)	16 (9.09%)	17 (9.66%):62 (35.2%):96 (54.5%)	175 (99.4%):0 (0%):0 (0%)	35 (19.9%):141 (80.1%)
		Non-carrier	125,016	56.8 (7.99)	27.4 (4.74)	-1.53 (2.96)	67189 (53.7%)	13443 (10.8%)	12654 (10.1%):44272 (35.4%):67619 (54.1%)	124275 (99.4%):157 (0.126%):271 (0.217%)	19362 (15.5%):105654 (84.5%)
FADS2	MDD cohort	Carrier	1,104	56.6 (7.98)	27.3 (4.63)	-1.62 (2.94)	563 (51%)	140 (12.7%)	88 (7.97%):446 (40.4%):564 (51.1%)	1095 (99.2%):2 (0.181%):5 (0.453%)	164 (14.9%):940 (85.1%)
		Non-carrier	360,193	56.8 (8.01)	27.4 (4.75)	-1.48 (2.99)	194719 (54.1%)	40636 (11.3%)	36875 (10.2%):127692 (35.5%):194175 (53.9%)	357972 (99.4%):459 (0.127%):746 (0.207%)	54499 (15.1%):305694 (84.9%)
	Metabolite cohort	Carrier	610	56.6 (8.01)	27.4 (4.79)	-1.54 (2.97)	309 (50.7%)	67 (11%)	44 (7.21%):258 (42.3%):305 (50%)	604 (99%):2 (0.328%):3 (0.492%)	98 (16.1%):512 (83.9%)
		Non-carrier	201,553	56.8 (7.99)	27.4 (4.75)	-1.52 (2.97)	108296 (53.7%)	21783 (10.8%)	20581 (10.2%):71459 (35.5%):108740 (54%)	200369 (99.4%):251 (0.125%):395 (0.196%)	31003 (15.4%):170550 (84.6%)
FADS3	MDD cohort	Carrier	1,651	56.9 (8.14)	27.3 (4.8)	-1.48 (2.93)	892 (54%)	183 (11.1%)	159 (9.63%):619 (37.5%):868 (52.6%)	1641 (99.4%):6 (0.363%):2 (0.121%)	243 (14.7%):1408 (85.3%)
		Non-carrier	360,302	56.8 (8.01)	27.4 (4.75)	-1.48 (2.99)	194720 (54%)	40664 (11.3%)	36882 (10.2%):127766 (35.5%):194201 (53.9%)	358073 (99.4%):457 (0.127%):753 (0.209%)	54512 (15.1%):305790 (84.9%)

Gene	Cohort	Carrier status	Total	Age	BMI	SES	% Female	% University/ College	Smoking status	Ethnicity	MDD
	Metabolite cohort	Carrier	885	57 (8.19)	27.5 (5)	-1.53 (2.93)	480 (54.2%)	90 (10.2%)	76 (8.59%):349 (39.4%):456 (51.5%)	882 (99.7%):2 (0.226%):0 (0%)	135 (15.3%):750 (84.7%)
		Non-carrier	201,644	56.8 (7.99)	27.4 (4.75)	-1.52 (2.97)	108306 (53.7%)	21798 (10.8%)	20596 (10.2%):71483 (35.5%):108794 (54%)	200455 (99.4%):251 (0.124%):399 (0.198%)	31013 (15.4%):170631 (84.6%)
FEN1	MDD cohort	Carrier	2,055	57.1 (7.91)	27.4 (4.82)	-1.63 (2.92)	1100 (53.5%)	266 (12.9%)	187 (9.1%):705 (34.3%):1147 (55.8%)	2048 (99.7%):0 (0%):1 (0.0487%)	305 (14.8%):1750 (85.2%)
		Non-carrier	360,134	56.8 (8.01)	27.4 (4.75)	-1.48 (2.99)	194641 (54%)	40625 (11.3%)	36877 (10.2%):127761 (35.5%):194053 (53.9%)	357902 (99.4%):463 (0.129%):754 (0.209%)	54485 (15.1%):305649 (84.9%)
	Metabolite cohort	Carrier	1,130	57.1 (7.96)	27.6 (4.96)	-1.59 (2.94)	599 (53%)	152 (13.5%)	97 (8.58%):398 (35.2%):629 (55.7%)	1128 (99.8%):0 (0%):0 (0%)	171 (15.1%):959 (84.9%)
		Non-carrier	201,533	56.8 (7.99)	27.4 (4.75)	-1.52 (2.97)	108258 (53.7%)	21755 (10.8%)	20587 (10.2%):71488 (35.5%):108689 (53.9%)	200343 (99.4%):253 (0.126%):399 (0.198%)	30995 (15.4%):170538 (84.6%)
MYRF	MDD cohort	Carrier	6,295	56.8 (7.9)	27.3 (4.73)	-1.41 (3)	3468 (55.1%)	791 (12.6%)	630 (10%):2173 (34.5%):3468 (55.1%)	6240 (99.1%):8 (0.127%):30 (0.477%)	0 (0%):940 (14.9%)
		Non-carrier	338,913	56.8 (8.01)	27.4 (4.76)	-1.48 (2.99)	183062 (54%)	38140 (11.3%)	34656 (10.2%):120298 (35.5%):182603 (53.9%)	336845 (99.4%):431 (0.127%):686 (0.202%)	0 (0%):51332 (15.1%)
	Metabolite cohort	Carrier	3,461	56.8 (7.82)	27.5 (4.82)	-1.46 (2.98)	1916 (55.4%)	377 (10.9%)	339 (9.79%):1191 (34.4%):1919 (55.4%)	3436 (99.3%):4 (0.116%):16 (0.462%)	0 (0%):529 (15.3%)
		Non-carrier	189,801	56.8 (7.99)	27.4 (4.75)	-1.52 (2.97)	101918 (53.7%)	20514 (10.8%)	19339 (10.2%):67346 (35.5%):102390 (53.9%)	188696 (99.4%):235 (0.124%):358 (0.189%)	0 (0%):29223 (15.4%)
TMEM258	MDD cohort	Carrier	5,138	57.1 (7.78)	27.2 (4.59)	-1.49 (2.96)	2782 (54.1%)	686 (13.4%)	505 (9.83%):1838 (35.8%):2776 (54%)	5083 (98.9%):2 (0.0389%):33 (0.642%)	768 (14.9%):4370 (85.1%)
		Non-carrier	356,991	56.8 (8.01)	27.4 (4.76)	-1.48 (2.99)	192932 (54%)	40198 (11.3%)	36551 (10.2%):126608 (35.5%):192391 (53.9%)	354807 (99.4%):461 (0.129%):722 (0.202%)	54011 (15.1%):302980 (84.9%)
	Metabolite cohort	Carrier	2,848	57.2 (7.67)	27.2 (4.45)	-1.52 (2.94)	1510 (53%)	356 (12.5%)	277 (9.73%):1033 (36.3%):1529 (53.7%)	2813 (98.8%):0 (0%):21 (0.737%)	413 (14.5%):2435 (85.5%)
		Non-carrier	199,779	56.8 (8)	27.4 (4.75)	-1.52 (2.97)	107330 (53.7%)	21548 (10.8%)	20406 (10.2%):70838	198622 (99.4%):253	30745 (15.4%):169034 (84.6%)

Gene	Cohort	Carrier status	Total	Age	BMI	SES	% Female	% University/ College	Smoking status	Ethnicity	MDD
									(35.5%):10776 8 (53.9%)	(0.127%):378 (0.189%)	

Supplementary Table 5: The association between carrying rare prioritised variants in FADS genes and MDD.

Chi-squared test statistics from testing the association between carrying the prioritised variants within a certain mask and MDD status.

Gene	Mask	Statistic	Df	P
FADS3	Mask3.singleton	2.456	1	0.117
FADS3	Mask2.0.01	2.302	1	0.129
FADS3	Mask4.singleton	2.061	1	0.151
FADS1	Mask4.0.01	1.654	1	0.198
FADS1	Mask2.singleton	1.630	1	0.202
TMEM258	Mask5.singleton	1.625	1	0.202
FEN1	Mask5.singleton	1.375	1	0.241
FADS1	Mask5.0.01	1.242	1	0.265
FADS1	Mask3.singleton	1.086	1	0.297
TMEM258	Mask3.singleton	1.063	1	0.303
TMEM258	Mask4.singleton	1.003	1	0.317
FADS3	Mask5.singleton	0.913	1	0.339
FADS1	Mask1.singleton	0.802	1	0.370
MYRF	Mask4.singleton	0.735	1	0.391
FADS1	Mask4.singleton	0.649	1	0.421
FEN1	Mask4.singleton	0.619	1	0.431
FADS3	Mask1.0.01	0.495	1	0.482
TMEM258	Mask2.0.01	0.468	1	0.494
FADS3	Mask4.0.01	0.412	1	0.521
MYRF	Mask5.singleton	0.364	1	0.546
TMEM258	Mask3.0.01	0.323	1	0.570
MYRF	Mask3.0.01	0.255	1	0.613
FADS1	Mask5.singleton	0.250	1	0.617
FADS2	Mask1.0.01	0.216	1	0.642
FADS2	Mask4.singleton	0.212	1	0.646
MYRF	Mask5.0.01	0.203	1	0.652
FADS3	Mask5.0.01	0.186	1	0.667
FEN1	Mask4.0.01	0.174	1	0.676
TMEM258	Mask4.0.01	0.158	1	0.691
FEN1	Mask3.0.01	0.148	1	0.700
FEN1	Mask2.0.01	0.141	1	0.708
FADS2	Mask5.singleton	0.136	1	0.712
FEN1	Mask2.singleton	0.130	1	0.718
MYRF	Mask4.0.01	0.126	1	0.723
FADS2	Mask4.0.01	0.125	1	0.724
TMEM258	Mask5.0.01	0.117	1	0.732

FEN1	Mask5.0.01	0.110	1	0.740
FADS1	Mask3.0.01	0.103	1	0.748
TMEM258	Mask2.singleton	0.082	1	0.775
FADS3	Mask3.0.01	0.079	1	0.779
TMEM258	Mask1.0.01	0.065	1	0.800
FADS2	Mask5.0.01	0.045	1	0.831
FADS2	Mask3.0.01	0.032	1	0.858
MYRF	Mask2.0.01	0.018	1	0.892
FADS1	Mask2.0.01	0.018	1	0.894
FADS2	Mask1.singleton	5.52e-03	1	0.941
TMEM258	Mask1.singleton	3.87e-22	1	1.000
Gene	Mask	Statistic	Df	P
FADS3	Mask1.singleton	2.45e-22	1	1.000
MYRF	Mask2.singleton	2.36e-22	1	1.000
FEN1	Mask3.singleton	3.54e-23	1	1.000
FADS3	Mask2.singleton	2.53e-23	1	1.000
FADS1	Mask1.0.01	5.11e-24	1	1.000
MYRF	Mask1.0.01	2.86e-24	1	1.000
FADS2	Mask3.singleton	1.89e-24	1	1.000
FADS2	Mask2.singleton	7.86e-25	1	1.000
MYRF	Mask3.singleton	2.13e-27	1	1.000
FADS2	Mask2.0.01	9.54e-28	1	1.000
MYRF	Mask1.singleton	3.64e-28	1	1.000

Supplementary Table 6: Significant associations between carrying rare prioritised variants in FADS genes and metabolite traits.

Welchs t-test statistics for the masks with a significant association with a metabolite trait. Full results can be found in the Supplementary Data. P_BH= refers to p-values corrected for multiple comparisons using the Benjamini-Hochberg procedure.

Gene	Metabolite	Mask	Mean Difference	T statistic [95% CI]	P	P_BH	Cohens D [95% CI]
FADS2	Degree of Unsaturation	Mask3.0.01	-0.872	-3.615 [-1.377,-0.367]	1.85e-03	0.040	-0.842 [-1.362,-0.305]
FEN1	Omega-6 Fatty Acids to Omega-3 Fatty Acids ratio	Mask5.0.01	0.109	3.492 [0.048,0.17]	4.98e-04	0.013	0.107 [0.047,0.167]
	Omega-3 Fatty Acids to Total Fatty Acids percentage	Mask5.0.01	-0.099	-3.161 [-0.161,-0.038]	1.61e-03	0.038	-0.097 [-0.157,-0.037]
TMEM258	Degree of Unsaturation	Mask5.0.01	0.138	7.353 [0.101,0.175]	2.50e-13	1.02e-11	0.139 [0.102,0.176]
		Mask4.0.01	0.138	7.331 [0.101,0.175]	2.95e-13	1.05e-11	0.139 [0.102,0.176]
	Omega-6 Fatty Acids to Omega-3 Fatty Acids ratio	Mask4.0.01	-0.171	-9.387 [-0.207,-0.135]	1.20e-20	1.14e-18	-0.175 [-0.211,-0.138]
		Mask5.0.01	-0.170	-9.353 [-0.206,-0.134]	1.65e-20	1.17e-18	-0.174 [-0.211,-0.137]
	Docosahexaenoic Acid	Mask4.0.01	0.127	6.826 [0.091,0.164]	1.06e-11	3.34e-10	0.128 [0.091,0.165]
		Mask5.0.01	0.127	6.807 [0.09,0.163]	1.21e-11	3.43e-10	0.128 [0.091,0.165]
	Omega-3 Fatty Acids to Total Fatty Acids percentage	Mask4.0.01	0.182	10.028 [0.146,0.217]	2.70e-23	4.54e-21	0.186 [0.149,0.223]
		Mask5.0.01	0.181	10.012 [0.146,0.216]	3.18e-23	4.54e-21	0.186 [0.149,0.222]
	Omega-3 Fatty Acids	Mask4.0.01	0.143	7.784 [0.107,0.179]	9.68e-15	5.52e-13	0.145 [0.109,0.182]
		Mask5.0.01	0.142	7.747 [0.106,0.178]	1.29e-14	6.13e-13	0.144 [0.108,0.181]

Supplementary Table 7: Joint burden tests across all masks per gene with MDD and each of the metabolomic traits.

MIN_P=Minimum P value, SBAT=Sparse Burden Association Test, GENE_P=unified strategy combining multiple tests proposed by REGENIE.

GENE	TRAIT	TEST	EXTRA	CHISQ	LOG10P	P
FADS1	DHA	GENE_P	DF=4;STRONGEST_MASK=Mask5	1.667	0.706	0.197
		MIN_P	DF=10	3.087	1.103	0.079
		SBAT	DF=10	1.604	0.688	0.205
	MDD	GENE_P	DF=3;STRONGEST_MASK=Mask5	0.182	0.174	0.669
		MIN_P	DF=10	2.809	1.028	0.094
	Omega_3	GENE_P	DF=4;STRONGEST_MASK=Mask5	2.717	1.003	0.099
		MIN_P	DF=10	4.025	1.348	0.045
		SBAT	DF=10	1.633	0.696	0.201
	Omega_3_pct	GENE_P	DF=4;STRONGEST_MASK=Mask5	3.394	1.184	0.065
		MIN_P	DF=10	4.249	1.406	0.039
		SBAT	DF=10	2.124	0.839	0.145
	Omega_6_by_Omega_3	GENE_P	DF=4;STRONGEST_MASK=Mask5	2.944	1.065	0.086
		MIN_P	DF=10	3.668	1.256	0.055
		SBAT	DF=10	2.004	0.804	0.157
	Unsaturation	GENE_P	DF=4;STRONGEST_MASK=Mask5	2.839	1.036	0.092
		MIN_P	DF=10	4.563	1.486	0.033
		SBAT	DF=10	1.830	0.754	0.176
FADS2	DHA	GENE_P	DF=4;STRONGEST_MASK=Mask3	6.435	1.951	0.011
		MIN_P	DF=9	10.478	2.918	1.21e-03
		SBAT	DF=9	7.285	2.158	6.95e-03
	MDD	GENE_P	DF=3;STRONGEST_MASK=Mask1	0.019	0.050	0.892
		MIN_P	DF=10	0.922	0.472	0.337
	Omega_3	GENE_P	DF=4;STRONGEST_MASK=Mask3	1.688	0.712	0.194
		MIN_P	DF=9	3.952	1.330	0.047
		SBAT	DF=9	1.976	0.796	0.160
	Omega_3_pct	GENE_P	DF=4;STRONGEST_MASK=Mask3	4.765	1.537	0.029
		MIN_P	DF=9	7.320	2.166	6.82e-03
		SBAT	DF=9	5.636	1.755	0.018
	Omega_6_by_Omega_3	GENE_P	DF=4;STRONGEST_MASK=Mask3	2.234	0.870	0.135
		MIN_P	DF=9	5.297	1.670	0.021
		SBAT	DF=9	2.868	1.044	0.090
	Unsaturation	GENE_P	DF=4;STRONGEST_MASK=Mask3	10.355	2.889	1.29e-03
		MIN_P	DF=9	12.412	3.370	4.27e-04
		SBAT	DF=9	10.290	2.874	1.34e-03
FADS3	DHA	GENE_P	DF=4;STRONGEST_MASK=Mask5	0.084	0.113	0.771
		MIN_P	DF=10	2.351	0.902	0.125
		SBAT	DF=10	0.034	0.069	0.854
	MDD	GENE_P	DF=3;STRONGEST_MASK=Mask3	0.546	0.337	0.460
		MIN_P	DF=10	3.006	1.081	0.083
	Omega_3	GENE_P	DF=4;STRONGEST_MASK=Mask5	4.109	1.370	0.043
		MIN_P	DF=10	8.666	2.489	3.24e-03
		SBAT	DF=10	3.973	1.335	0.046

GENE	TRAIT	TEST	EXTRA	CHISQ	LOG10P	P
	Omega_3_pct	GENE_P	DF=4;STRONGEST_MASK=Mask5	2.846	1.038	0.092
		MIN_P	DF=10	4.200	1.393	0.040
		SBAT	DF=10	2.445	0.928	0.118
	Omega_6_by_Omega_3	GENE_P	DF=4;STRONGEST_MASK=Mask5	4.593	1.493	0.032
		MIN_P	DF=10	7.218	2.142	7.22e-03
		SBAT	DF=10	4.971	1.589	0.026
	Unsaturation	GENE_P	DF=4;STRONGEST_MASK=Mask4	0.420	0.287	0.517
		MIN_P	DF=10	1.376	0.618	0.241
		SBAT	DF=10	0.251	0.210	0.616
FEN1	DHA	GENE_P	DF=4;STRONGEST_MASK=Mask5	14.333	3.815	1.53e-04
		MIN_P	DF=8	3.494	1.211	0.062
		SBAT	DF=8	0.903	0.466	0.342
	MDD	GENE_P	DF=3;STRONGEST_MASK=Mask5	1.07e-04	3.60e-03	0.992
		MIN_P	DF=8	1.425	0.633	0.233
	Omega_3	GENE_P	DF=4;STRONGEST_MASK=Mask5	22.046	5.575	2.66e-06
		MIN_P	DF=8	10.022	2.811	1.55e-03
		SBAT	DF=8	4.943	1.582	0.026
	Omega_3_pct	GENE_P	DF=4;STRONGEST_MASK=Mask5	29.724	7.303	4.98e-08
		MIN_P	DF=8	13.174	3.547	2.84e-04
		SBAT	DF=8	7.235	2.146	7.15e-03
	Omega_6_by_Omega_3	GENE_P	DF=4;STRONGEST_MASK=Mask5	29.088	7.160	6.92e-08
		MIN_P	DF=8	15.730	4.136	7.31e-05
		SBAT	DF=8	9.509	2.689	2.04e-03
	Unsaturation	GENE_P	DF=4;STRONGEST_MASK=Mask5	12.460	3.381	4.16e-04
		MIN_P	DF=8	1.576	0.679	0.209
		SBAT	DF=8	0.552	0.340	0.458
MYRF	DHA	GENE_P	DF=4;STRONGEST_MASK=Mask4	14.683	3.896	1.27e-04
		MIN_P	DF=10	3.226	1.140	0.072
		SBAT	DF=10	0.803	0.432	0.370
	MDD	GENE_P	DF=3;STRONGEST_MASK=Mask4	0.030	0.064	0.863
		MIN_P	DF=10	0.760	0.417	0.383
	Omega_3	GENE_P	DF=4;STRONGEST_MASK=Mask4	16.592	4.334	4.63e-05
		MIN_P	DF=10	1.671	0.707	0.196
		SBAT	DF=10	0.502	0.320	0.478
	Omega_3_pct	GENE_P	DF=4;STRONGEST_MASK=Mask4	24.300	6.084	8.24e-07
		MIN_P	DF=10	2.376	0.909	0.123
		SBAT	DF=10	0.436	0.293	0.509
	Omega_6_by_Omega_3	GENE_P	DF=4;STRONGEST_MASK=Mask4	20.407	5.203	6.26e-06
		MIN_P	DF=10	1.890	0.771	0.169
		SBAT	DF=10	0.290	0.229	0.590
	Unsaturation	GENE_P	DF=4;STRONGEST_MASK=Mask4	8.493	2.448	3.57e-03
		MIN_P	DF=10	2.491	0.941	0.114
		SBAT	DF=10	0.545	0.337	0.460
TMEM258	DHA	GENE_P	DF=4;STRONGEST_MASK=Mask4	26.979	6.687	2.06e-07
		MIN_P	DF=10	24.947	6.230	5.89e-07
		SBAT	DF=10	18.129	4.685	2.06e-05
	MDD	GENE_P	DF=3;STRONGEST_MASK=Mask5	0.017	0.048	0.895
		MIN_P	DF=10	2.024	0.810	0.155
	Omega_3	GENE_P	DF=4;STRONGEST_MASK=Mask4	34.355	8.338	4.59e-09
		MIN_P	DF=10	32.403	7.902	1.25e-08

GENE	TRAIT	TEST	EXTRA	CHISQ	LOG10P	P
	Omega_3_pct	SBAT	DF=10	24.303	6.085	8.23e-07
		GENE_P	DF=4;STRONGEST_MASK=Mask4	56.154	13.174	6.70e-14
		MIN_P	DF=10	51.911	12.236	5.81e-13
		SBAT	DF=10	41.434	9.914	1.22e-10
	Omega_6_by_Omega_3	GENE_P	DF=4;STRONGEST_MASK=Mask4	52.698	12.410	3.89e-13
		MIN_P	DF=10	48.783	11.544	2.86e-12
		SBAT	DF=10	39.248	9.428	3.73e-10
	Unsaturation	GENE_P	DF=4;STRONGEST_MASK=Mask5	36.096	8.726	1.88e-09
		MIN_P	DF=10	34.410	8.350	4.46e-09
		SBAT	DF=10	25.303	6.310	4.90e-07

Supplementary Table 8: Welchs t-test results for the association between chr11:61861814:G:C carrier status and metabolite levels.

Metabolite	Mean Difference	T statistic [95% CI]	P	Cohens D [95% CI]
Omega-3 Fatty Acids	-0.122	-0.382 [-0.877,0.633]	0.714	-0.129 [-0.789,0.540]
Omega-3 Fatty Acids to Total Fatty Acids percentage	-0.192	-0.549 [-1.020,0.635]	0.600	-0.195 [-0.893,0.515]
Omega-6 Fatty Acids to Omega-3 Fatty Acids ratio	0.222	0.581 [-0.681,1.124]	0.579	0.215 [-0.526,0.941]
Degree of Unsaturation	-0.099	-0.356 [-0.756,0.558]	0.732	-0.111 [-0.722,0.507]
Docosahexaenoic Acid	-0.111	-0.429 [-0.720,0.499]	0.681	-0.127 [-0.709,0.463]

Supplementary Table 9: The demographics of the carriers and non-carriers of the variant chr11_61810815_C_A in the Metabolite cohort.

Continuous traits (Age, BMI and SES) are displayed as the mean values within the cohorts whilst categorical traits are displayed as the counts and % of the sample. Smoking status shows the ratio of current, previous and never smokers, Ethnicity shows the ratio of White, Mixed and Other populations, and MDD shows the ratio of MDD cases and controls within the sample.

chr11_61810815_C_A.carrier	No	Yes
Total	125184	8
Age	56.8 (7.99)	57.4 (8.19)
BMI	27.4 (4.74)	27.8 (5.76)
SES	-1.53 (2.96)	-2.85 (1.98)
% Female	67272 (53.7%)	4 (50%)
% Attended University/College	13459 (10.8%)	0 (0%)
Smoking status	12671 (10.1%):44332 (35.4%):67709 (54.1%)	0 (0%):2 (25%):6 (75%)
Ethnicity	124442 (99.4%):157 (0.125%):271 (0.216%)	8 (100%):0 (0%):0 (0%)
MDD	0 (0%):19395 (15.5%)	0 (0%):2 (25%)
Degree of Unsaturation	1.36 (0.0797)	1.34 (0.0512)
Docosahexaenoic Acid	0.239 (0.0838)	0.197 (0.0933)
Omega-3 Fatty Acids	0.539 (0.223)	0.435 (0.296)
Omega-3 Fatty Acids to Total Fatty Acids %	4.42 (1.53)	3.56 (1.5)
Omega-6 Fatty Acids to Omega-3 Fatty Acids Ratio	9.65 (4.02)	12.3 (3.9)

Supplementary Table 10: Welchs t-test results for the association between chr11:61810815:C:A carrier status and metabolite levels.

Metabolite	Mean Difference	T statistic [95% CI]	P	Cohens D [95% CI]
Omega-3 Fatty Acids	-0.644	-1.434 [-1.705,0.418]	0.195	-0.566 [-1.375,0.278]
Omega-3 Fatty Acids to Total Fatty Acids percentage	-0.660	-1.820 [-1.518,0.197]	0.112	-0.658 [-1.425,0.146]
Omega-6 Fatty Acids to Omega-3 Fatty Acids ratio	0.690	1.742 [-0.247,1.627]	0.125	0.655 [-0.174,1.446]
Degree of Unsaturation	-0.273	-1.163 [-0.830,0.283]	0.283	-0.326 [-0.889,0.259]
Docosahexaenoic Acid	-0.631	-1.564 [-1.584,0.323]	0.162	-0.591 [-1.374,0.227]

Supplementary Table 11: The distribution of metabolite traits in the carriers of the chr11:61810815:C:A variant.

The outlying sample was identified using the IQR method, which classes outliers as data points which lie below $Q1 - 1.5(IQR)$ or above $Q3 + 1.5(IQR)$. There was one outlying sample which was identified in all metabolites apart from the Degree of Unsaturation. $Q1$ = Lower Quartile, $Q3$ = Upper Quartile, IQR = Interquartile Range.

Metabolite	Q1	Q3	IQR (Q3-Q1)	Lower Bound ($Q1 - 1.5 \cdot IQR$)	Upper Bound ($Q3 + 1.5 \cdot IQR$)	Number of outliers
Degree of Unsaturation	-0.798	0.282	1.080	-2.418	1.902	None
Omega-3 Fatty Acids	-1.415	0.274	1.140	-3.125	1.437	1
Docosahexaenoic Acid	-1.297	0.108	1.189	-3.080	1.676	1
Omega-3 Fatty Acids to Total Fatty Acids %	-1.373	0.318	1.054	-2.954	1.263	1
Omega-6 to Omega-3 Fatty Acids Ratio	0.454	1.420	0.966	-0.995	2.868	1

Supplementary Table 12: Welchs t-test results for the association between chr11:61810815:C:A carrier status and metabolite levels after excluding the outlying sample.

Metabolite	Mean Difference	T statistic [95% CI]	P	Cohens D [95% CI]
Degree of Unsaturation	-0.384	-1.599 [-0.971, 0.204]	0.161	-0.463 [-1.071, 0.175]
Omega-3 Fatty Acids	-1.031	-3.941 [-1.671,-0.391]	7.62e-03	-1.209 [-2.086,-0.290]
Docosahexaenoic Acid	-0.972	-3.941 [-1.576,-0.369]	7.61e-03	-1.162 [-2.005,-0.279]
Omega-3 Fatty Acids to Total Fatty Acids percentage	-0.975	-4.664 [-1.486,-0.463]	3.45e-03	-1.226 [-2.059,-0.360]
Omega-6 Fatty Acids to Omega-3 Fatty Acids ratio	1.051	5.552 [0.588, 1.514]	1.44e-03	1.348 [0.453, 2.217]