A systematic investigation into the role of rare protein-coding biallelic variants in human health and disease

Frederik Heymann Lassen

New College
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy

Michaelmas 2024

Abstract

Exome sequencing has identified rare protein-coding variants associated with many diseases, typically by studying heterozygous and homozygous variation. However, the phenotypic impact of deleterious Compound Heterozygous (CH) variants on trait variation at scale remains largely unexplored. In this thesis, I investigate the role of biallelic variation (CH and homozygotes) in shaping common complex traits. Through *in silico* permutation and association analyses across 311 binary traits, I demonstrate that CH variation can contribute to developing cancer, asthma, dermatitis, and hearing loss. I show that inclusion of CH variation can increase power to discover novel gene-trait associations, such as CH variants in *ATP2C2* associated with increased risk of chronic obstructive pulmonary disease (COPD). Next, I extend my investigation to dominance effects owed to rare deleterious CH and homozygotes variants which is where we expect the largest nonadditive effects. To do this, I impose orthogonality between the additive and nonadditive components of heritability by recoding genotypes. I identify 29 nonadditive associations across proteomic and physiological endpoints, including well-established and novel recessive gene-traits such as *FLG*-Asthma and *RNF123*-HbA1c respectively. I demonstrate that combining additive and nonadditive effects helps explain complex genetic architectures, including the partially recessive inheritance of *NOD2* variants in Crohn's disease, accounting for epidemiologically observed modest risk increase in heterozygotes and high risk in homozygotes. My investigations into biallelic variation culminate in a meta-analysis of rare recessive effects across UK Biobank, Genes and Health, Genomics England, and Biobank Japan. Here, I leverage compound heterozygosity across ancestries, to perform federated recessive gene-based fixed-effects meta-analysis, revealing multiple associations, including biallelic variants in both *DNAI1* and *ODAD1* associated with COPD. Together, this thesis advances our understanding of recessive homozygous and CH effects in shaping trait variation, providing new methodologies to study, and insights into the genetic architecture underlying common complex traits.

A systematic investigation into the role of rare protein-coding biallelic variants in human health and disease



Frederik Heymann Lassen

New College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Michaelmas 2024

Acknowledgements

Personal

First and foremost, I extend my sincere thanks to Professor Cecilia Lindgren and Professor Benedikt Kessler for their supervision and for fostering an inclusive and engaging research environment. I am deeply grateful to Dr. Duncan Palmer for his mentorship and commitment to high-quality science, which were instrumental in this thesis. My thanks also go to Dr. Andreas Damianou for his invaluable assistance during my wetlab work, and to my fellow DPhil colleagues—Samvida, Nik, and Vassi—for their wisdom and discussions.

This academic adventure would not have been possible without the bedrock of support provided by my family. To Hanne, Peter, and Sophie: your unconditional love and encouragement have been my constant source of strength. Your belief in me, regardless of where my aspirations led, has been invaluable. I am forever grateful.

The friendships forged during my DPhil have enriched this experience immeasurably. A special acknowledgment goes to Thomas, whose engaging discussions have inspired my future career path. Finally, I owe an immense debt of gratitude to my partner, Maria. Your unwavering support, patience, and kindness throughout the writing of this thesis have been my anchor in stressful times. Your presence in my life has been a source of joy and motivation, for which I am profoundly thankful. To all those mentioned here, and to the many others who have contributed to this journey, I offer my heartfelt thanks. This thesis stands as a mosaic, each piece carefully placed by the collective support, wisdom, and encouragement you have all provided.

Institutional

I am deeply grateful to New College at the University of Oxford for welcoming me into their vibrant community, enriching my DPhil experience. My sincere thanks also go to the Wellcome Trust. Their generous stipend has been instrumental in broadening my research horizons and making this work possible.

Previously published material

Throughout my doctoral studies, I have been fortunate be embedded in a dynamic research environment that has allowed me to explore a wide range of topics, ranging from medical population genetics to mass spectrometry. The collaborative nature of scientific research has been integral to the production of this thesis, and I have had the privilege of working with exceptional collaborators. Unless explicitly stated otherwise in the relevant sections, all data, scripts, code, and analyses presented in this thesis are my own original work.

Chapter 2 is based on my own first-authored work published in Cell Genomics in 2024 titled 'Exome-wide evidence of Compound Heterozygous effects'. I have received written permission from all co-authors and supervisors, Cecilia Lindgren and Benedikt Kessler, to re-use previously published material. The remaining work is currently being prepared for publication. I have also received confirmation from collaborators that I can use this work in the thesis.

Throughout this thesis, I use the pronouns 'we' and 'I' interchangeably to reflect both the collaborative nature of the research and my individual contributions.

To my parents, Peter and Hanne

Abstract

Exome sequencing has identified rare protein-coding variants associated with many diseases, typically by studying heterozygous and homozygous variation. However, the phenotypic impact of deleterious Compound Heterozygous (CH) variants on trait variation at scale remains largely unexplored. In this thesis, I investigate the role of biallelic variation (CH and homozygotes) in shaping common complex traits. Through *in silico* permutation and association analyses across 311 binary traits, I demonstrate that CH variation can contribute to developing cancer, asthma, dermatitis, and hearing loss. I show that inclusion of CH variation can increase power to discover novel gene-trait associations, such as CH variants in *ATP2C2* associated with increased risk of chronic obstructive pulmonary disease (COPD). Next, I extend my investigation to dominance effects owed to rare deleterious CH and homozygotes variants which is where we expect the largest nonadditive effects. To do this, I impose orthogonality between the additive and nonadditive components of heritability by recoding genotypes. I identify 29 nonadditive associations across proteomic and physiological endpoints, including well-established and novel recessive gene-traits such as *FLG*-Asthma and *RNF123*-HbA1c respectively. I demonstrate that combining additive and nonadditive effects helps explain complex genetic architectures, including the partially recessive inheritance of *NOD2* variants in Crohn's disease, accounting for epidemiologically observed modest risk increase in heterozygotes and high risk in homozygotes. My investigations into biallelic variation culminate in a meta-analysis of rare recessive effects across UK Biobank, Genes and Health, Genomics England, and Biobank Japan. Here, I leverage compound heterozygosity across ancestries, to perform federated recessive gene-based fixed-effects meta-analysis, revealing multiple associations, including biallelic variants in both *DNAI1* and *ODAD1* associated with COPD. Together, this thesis advances our understanding of recessive homozygous and CH effects in shaping trait variation, providing new methodologies to study, and insights into the genetic architecture underlying common complex traits.

Contents

List of Figures **xiii**

List of Tables **xvii**

1 Introduction **1**
1.1 A very brief history of human genetics 1
1.2 Statistical genetics in the 21st century 5
1.2.1 Genome-wide association studies 5
1.2.2 Analysis of rare protein coding variants 7
1.2.3 Biobanks and innovation in phenotyping 9
1.3 The study of biallelic variation 11
1.3.1 Identification and medical relevance of biallelic variation . . 11
1.3.2 Nonadditive effects 14
1.4 Thesis aims 19

2 Exome-wide evidence of compound heterozygous effects **21**
2.1 Introduction 22
2.2 Methods 24
2.2.1 Quality control of genotyping arrays 24
2.2.2 Quality control of 200k exomes 24
2.2.3 Phasing 29
2.2.4 Trio-switch error rates 29
2.2.5 Read-backed phasing 30
2.2.6 Phenotype curation 30
2.2.7 Variant annotation masks 31
2.2.8 Biallelic encoding and recessive models 31
2.2.9 Gene copy dosage encoding and additive models 32
2.2.10 Polygenic risk scores 32
2.2.11 Conditional analysis 34
2.2.12 Permutation of genetic phase 36
2.2.13 Gene-set enrichment of biallelic variation 36
2.2.14 Homozygote and CH down-sampling 37

2.2.15 Power analysis for biallelic association . . . 37
2.2.16 Simulation . . . 38
2.2.17 Age-at-onset analysis . . . 39
2.2.18 Replication analyses . . . 40
2.2.19 Simulation of CH and homozygous variants . . . 40
2.3 Results . . . 42
2.3.1 Accurate phase inference and validation using parent-offspring trios and short-read sequences . . . 42
2.3.2 Identification and examination of CH variation in UK Biobank 45
2.3.3 Systematic evaluation of biallelic effects on common disease 50
2.3.4 Permutation testing establishes the role of genetic phase on disease risk . . . 53
2.3.5 Incorporation of genetic phase boosts power in gene-level rare variant burden analysis . . . 55
2.3.6 CH variants elevate lifetime risk of disease . . . 56
2.3.7 Common complex disorders implicated by CH variation . . . 60
2.4 Discussion . . . 62
2.4.1 Limitations . . . 65

3 Rare variant analysis of dominance 67
3.1 Introduction . . . 68
3.2 Methods . . . 71
3.2.1 QC of the 450k exome release . . . 71
3.2.2 Additional QC of genotyping array data . . . 74
3.2.3 Statistical phasing . . . 74
3.2.4 Variant annotation . . . 74
3.2.5 Encoding nonadditivity at a locus . . . 76
3.2.6 Estimating genetic architectures . . . 78
3.2.7 Association analysis of alternate genotype encodings . . . 79
3.2.8 Simulation of counts of observed biallelic variants . . . 80
3.2.9 Simulation of null trait and analysis . . . 82
3.2.10 Phenotype curation . . . 84
3.2.11 Analysis of proteomic measurements . . . 85
3.2.12 Modelling recessive selection . . . 86
3.3 Results . . . 91
3.3.1 High-quality phased variants enables biallelic analysis . . . 91
3.3.2 Biallelic effects across the plasma proteome . . . 92
3.3.3 Allelic recoding yields nonadditive proteomic effects . . . 97
3.3.4 Analysis of recessive effects across health and disease . . . 107

3.3.5 Simulation of random phenotypes . . . 107
3.3.6 Exome-wide scans of nonadditive effects . . . 111
3.3.7 Deciphering mode of inheritance in binary traits . . . 114
3.3.8 Characterizing the spectrum of recessiveness . . . 119
3.3.9 The burden of nonadditive effects . . . 121
3.3.10 Identifying recessive selection . . . 124
3.4 Discussion . . . 127
3.4.1 Limitations . . . 133

4 Detection of biallelic recessive effects through meta-analysis 135
4.1 Introduction . . . 136
4.2 Methods . . . 137
4.2.1 Study datasets . . . 137
4.2.2 General quality control and phasing guidelines for biobanks . 139
4.2.3 UK Biobank quality control and phasing . . . 143
4.2.4 Genomics England quality control . . . 143
4.2.5 Biallelic variant identification . . . 146
4.2.6 Counting expected and observed biallelic variants . . . 147
4.2.7 Phenotype curation . . . 147
4.2.8 Association analysis . . . 149
4.2.9 Meta-analysis . . . 149
4.2.10 Cauchy Combination of P-values across annotation masks . 150
4.2.11 Power comparison between meta-analysis and UK Biobank . 151
4.2.12 Evaluation of sample overlap across cohorts . . . 152
4.3 Results . . . 153
4.3.1 Patterns of disease across cohorts . . . 153
4.3.2 Characterising biallelic variation across biobanks . . . 161
4.3.3 Federated rare variant meta-analysis . . . 167
4.3.4 Identification of recessive effects through meta-analysis . . . 176
4.3.5 Meta-analysis improves power for select gene-trait associations 183
4.4 Discussion . . . 186
4.4.1 Limitations . . . 188

5 Conclusions and future directions 191
5.1 Conclusions . . . 191
5.2 Discussion and future directions . . . 193
5.2.1 Long-read sequencing and decoding the non-coding genome . 193
5.2.2 Nonadditive effects enable gene therapies and drug development 196
5.2.3 The case for humans as the new model organism . . . 198
5.2.4 Final remarks . . . 200

Appendices

A Supplement to chapter 2 **205**

B Supplement to chapter 3 **237**

C Supplement to chapter 4 **253**

D Supplement to chapter 5 **269**

References **271**

List of Figures

1.1 Determination of biallelic status in a trait with a recessive mode of inheritance . . . 12
1.2 Basic principles of trio-phasing, read-backed phasing, and statistical phasing . . . 13
1.3 Additive model fit across different architectures at two different MAFs 18

2.1 CH variants composed of at least one ultra-rare variant can be robustly identified in large scale biobanks . . . 44
2.2 Agreement between read-backed and statistically phased variants . . 45
2.3 Distribution of variant annotation categories before and after broad consequence categorization and filtering by PP≥0.9 . . . 46
2.4 Gene-set depletion/enrichment modeling for biallelic variation . . . 48
2.5 Count of unique genes affected by homozygous and compound heterozygous variants as a function of sample size . . . 49
2.6 Conditional recessive and additive modeling of gene copy disruption in 311 phenotypes across 176,587 participants . . . 52
2.7 *In-silico* permutation of genetic phase provides evidence for CH-specific effects . . . 54
2.8 Age-at-diagnosis modeling reveals novel recessive effects driven by damaging bi-allelic variants . . . 57
2.8 Trajectories of haplotype disruption in common disease . . . 60

3.1 Assessment of effect size and direction across additive haplotype-level association analysis . . . 95
3.2 Protein abundance patterns for 34 genes with at least five biallelic pLoF variant carriers . . . 98
3.3 Aggregated Manhattan plot for additive and nonadditive effects using rare pLoF pseudo-variants with associated protein abundance . . . 101
3.4 Aggregated Manhattan plot for additive and nonadditive effects using rare pLoF and damaging missense/protein altering pseudo-variants with associated protein abundance . . . 103
3.5 Comparison of inferred genetic architecture and observed protein abundance for *HMCN2* and *MMP10* . . . 104

3.6 Genetic architectures of significant associations with proteomic measurements using pLoF+damaging missense/protein altering mask . 106
3.7 QQ-plots for 600 randomly generated binary and continuous heritable phenotypes tested against pseudo-variants with nonadditive encoding 108
3.8 QQ-plot for simulated null traits stratified by Expected Biallelic Count 110
3.9 Distribution of λ_{GC} in additive and nonadditive analyses before and after EBC filtering 112
3.10 Aggregated Manhattan plot for additive and nonadditive effects using pLoF and damaging missense/protein altering variant mask 113
3.11 Inferred genetic architecture of *MUTYH*-Colorectal cancer and *GJB2* and hearing loss 117
3.12 Estimated genetic architecture of *NOD2*-CD 117
3.13 Inferred genetic architecture of *FLG* and its pleiotropic effects on four traits 118
3.14 Inferred genetic architecture for *RNF123*-HbA1c and *ABHD15*-Sitting height with significant nonadditive contributions 120
3.15 Spectrum of genetic effects across binary traits: Significant nonadditive associations versus randomly sampled additive associations . . 121
3.16 Comparison of variant-based burden analysis and pseudo-variant analysis 123
3.17 Overview of genes that are depleted of compound heterozygous and homozygous variation 125
3.18 QQ-plot for scan of recessive selection in 339,041 individuals across variant annotation masks 126

4.1 Distribution of age in Genomics England for individuals diagnosed with NDD 148
4.2 Flowchart for the BRaVa pilot analysis 154
4.3 Proportion and counts of samples analyzed in the BRaVa pilot . . . 157
4.4 Scatter plot of the pairwise comparison of disease prevalence between biobanks 159
4.5 Assessment and validation of phasing accuracy through trio-switch error rates, read-backed phasing, and simulation of biallelic variants 163
4.6 The compound heterozygous and homozygous pLoF burden across biobanks 165
4.7 Distribution of λ_{GC} and λ_{95} before meta-analysis across investigated variant annotation masks 170
4.8 λ_{GC} and λ_{95} for additive, recessive and nonadditive meta-analyses by variant annotation masks 172
4.9 Manhattan plot for additive, recessive and nonadditive meta-analysis 175

4.10 Overview of significant associations by biobanks, ancestry, and variant annotation . . . 176
4.11 Determining recessive mode of inheritance by comparing additive and recessive P-values . . . 178
4.12 Forest plot of 6 putatively recessive gene-trait associations reported in the meta-analysis . . . 181
4.13 Evaluation of meta-analysis power by comparison of χ^2 across analyses 184
4.14 Projected lower bounds for detectable Odds Ratios with additional BRaVa biobanks . . . 185

A.1 Histogram and scatter-plots of X chromosome F-statistic by collection 206
A.2 Scatter-plots of PCs of UKBB genotype data projected into the PC space defined by 1KGP samples . . . 208
A.3 Distributions of sample metrics following initial variant filtering for 200k ES . . . 209
A.4 Distributions of variant metrics before and after QC for 200k ES . . 210
A.5 Simulation of CH and homozygous events in an outbred population 215
A.6 Variants retained as a function of increasing phasing confidence score 216
A.7 Trio switch error rates by chromosome . . . 217
A.8 Agreement between read-backed phasing and statistical phasing in UKBB . . . 218
A.9 Allele frequencies of variants in the CH state . . . 220
A.10 Allele frequencies of variants in *cis* . . . 221
A.11 Distribution of observed variants across samples by allele frequency 222
A.12 Distribution of unique variants observed by allele frequency . . . 222
A.13 Power analysis to determine the required number of biallelic variants to detect specific ORs . . . 223
A.14 . . . 224
A.14 Simulation study to test our ability to detect biallelic effects in the presence of true effects . . . 225
A.15 Recessive association analysis without accounting for PRS . . . 225
A.16 Effect of excluding singletons prior to calling CH variants for any gene-traits with a $P < 0.001$ from the initial analysis . . . 228
A.17 Number of CH variants after filtering to those that consist of at least one singleton . . . 229
A.18 Association P-values before and after inclusion of PRS as a covariate 230
A.19 Attrition of Bonferroni significant associations after successive conditioning steps and filters . . . 231
A.20 Cox proportional hazards modeling with and without polygenic effects 232
A.21 Kaplan-Meier survival curves for carriers of biallelic variants . . . 233

A.22 Kaplan-Meier survival curves for carriers of CH, homozygous, and heterozygous variants 234
A.23 Co-occurrence of deleterious *ATP2C2* variants by COPD status . . 235

B.1 Overview of sex imputation for 450k exome sequencing data 240
B.2 Principal component analysis with 1000 genomes project for 450k exome sequencing data 241
B.3 Mean genotype quality, allelic depth and call rate for 450k exome sequencing data 242
B.4 Variant metrics before and after sample removals for 450k exome sequencing data 243
B.5 Distribution of θ parameter for genes with MAF > 0.005 245
B.6 Power analysis for gene-level Hardy-Weinberg Equilibrium across different disequilibrium coefficients 245
B.7 Comparison between simulated and empirically observed counts of pLoF carriers 248
B.8 QQ-plot for additive, recessive and nonadditive analysis of Olink data 249

C.1 Comparison of test-statistic inflation across biobanks and annotation categories before meta-analysis using λ_{GC} 257
C.2 Comparison of test-statistic inflation across biobanks and annotation categories before meta-analysis using λ_{95} 258
C.3 Pre-meta-analysis λ_{GC} and λ_{95} for synonymous variants across traits and genetic models 259
C.4 Investigation of sample overlap between biobanks using Z-score correlations 260
C.6 QQ-plots resulting from additive meta-analysis across up to 4 different biobanks 263
C.8 QQ-plots from recessive meta-analysis across up to 4 biobanks . . . 265
C.10 QQ-plots from nonadditive meta-analysis across up to 4 biobanks . 267

List of Tables

2.1 Sample filtering: MAD Intervals 25

3.1 Overview of the additive, recessive and nonadditive encoding leveraged in this chapter 80
3.2 Summary of the effects of variant categories on cognate protein abundance, based on additive haplotype-specific association analyses 93
3.3 λ_{GC} across BC and EBC cutoffs, stratified by trait prevalence . . . 110
3.4 Overview of the 20 significant quantitative and binary gene-trait associations reported 115

4.1 Number of cases ascertained across 32 binary traits in four biobanks 155
4.2 Number of samples with up to 10 quantitative measurements across three biobanks with quantitative measurements 156
4.3 Number of carriers with at least one biallelic genotype in BRaVa stratified by ancestry 166
4.4 Number of observed biallelic genotypes observed across across ancestries 166
4.5 Overview of the 40 associations discovered through meta-analysis of the recessive genetic model 180

A.1 Summary of sample attrition following filtering for 200k UKBB ES release 207
A.2 Summary of variant attrition after imposing filters in 200k UKBB ES release 207
A.3 Comparative SER analysis for various phasing methods 211
A.4 Trio switch errors by MAC bin for 96 parent-offspring trio relationships before and after filtering by PP ≥ 0.90 212
A.5 Trio-SER across genes before and after filtering by PP≥ 0.90 using 96 trio-offspring relationships 213
A.6 Comparison of statistical and physical phasing 214
A.7 Systematic search and conditional analysis of common variants around significant gene-trait associations 219
A.8 Comparison of significant hits from additive haplotype-collapsing association analysis and `SAIGE-GENE+` burden testing 220

A.9 Variant annotation counts before and after filtering to sites with PP≥ 0.90 . . . 221
A.10 Overview of associations before and after conditioning on off-chromosome PRS, nearby common variation, and the burden of rare variation . . 226
A.11 Replication of significant associations from the cross-sectional analysis 227
A.12 Replication analysis of the Bonferroni significant associations from the age-at-onset analysis . . . 227

B.1 Trio-switch error rates across for 450k exome sequencing following phasing . . . 244
B.2 Overview of the UKBB field entry for 114 mostly uncorrelated traits 246
B.3 Number of mono- and biallelic variants identified across annotation masks in 450k exome sequencing dataset in Europeans . . . 247
B.4 Overview of 26 gene-assay associations discovered using recessive or nonadditive models with pLoF annotation mask . . . 250
B.5 68 gene-assay associations discovered using a standard recessive model with pLoF and damaging missense/protein altering mask . . . 251
B.6 Significant quantitative and binary gene-trait associations using pLoF and/or damaging missense/protein altering mask . . . 252

C.1 ICD10 and ICD9 codes used to nominate the binary phenotypes investigated in this analysis . . . 254
C.2 Enrichment of drug targets among genes with at least one biallelic pLoF variant . . . 255
C.3 Enrichment of biallelic pLoF variants in essential and non-essential gene sets . . . 255
C.4 Manually curated HPO and ICD-10 codes for NDD status determination in GEL . . . 256
C.5 Heterogeneity analysis of 40 significant gene-trait associations from recessive meta-analysis . . . 261
C.6 Comparison of significant additive synonymous associations with Genebass for HDLC . . . 268

D.1 FDA-approved cellular and gene therapy products as of October 2024 270

List of abbreviations

1KGP 1000 Genomes. 5, 27, 73, 139, 141, 208

AD Allele Depth. 72

AF Allele Frequency. 7, 74, 76, 91, 145, 222

AFR African. 71, 141, 153, 155, 156

AggV2 Aggregated Variant Call. 143

AI Artificial Intelligence. 8, 201

AMR American/Admixed. 141

AUC area under the curve. 34, 225

BBJ Biobank Japan. 137–139, 142, 147, 149, 151, 153, 155, 158, 160–164, 166, 175, 182, 184, 185, 187

BC Biallelic Count. 82, 83, 108, 110, 149

BF Biallelic Frequency. 83, 109

BMI body mass index. 9, 84

BRaVa Biobank Rare Variant analysis. 10, 136, 137, 153, 161, 165, 185, 186, 188, 194, 255

CD Crohn's Disease. 112, 116, 117, 120, 123, 129, 130, 197, 198

CDF cumulative distribution function. 150

CF Cystic Fibrosis. 197

CH compound heterozygous. 11, 12, 14, 19, 22–24, 32, 36–38, 40–45, 47–51, 53, 54, 56–58, 60–66, 79, 80, 87, 91, 92, 96, 99, 122–124, 127, 130, 131, 161–164, 166, 186, 191–195, 215, 220, 222, 226, 229, 232–234, 247, 248

CI confidence interval. 30, 34, 37, 98, 121, 152, 163, 181, 214, 217, 218, 255, 260

COPD chronic obstructive pulmonary disease. 53, 56, 58, 60, 61, 159, 175, 178, 179, 184, 187, 188, 192, 193, 197, 235

DF degree-of-freedom. 69, 132, 151, 152

DNA Deoxyribonucleic Acid. 5, 7, 11, 85, 92, 138, 201

DP Total Sequencing Depth. 71–73, 140, 242

EAS East Asian. 71, 141, 142, 153, 155, 156, 161, 162

EBC Expected Biallelic Count. 83, 109–112

EHR electronic health record. 9, 50, 138, 198

EMR electronic medical record. 139

ES exome sequencing. 7, 9, 11, 23–26, 28–30, 32–35, 42, 43, 71, 74, 91, 92, 95, 112, 140, 142, 154, 155, 161, 188, 194, 212, 244, 247

EUR European. 27, 71, 141, 142, 153, 155, 156, 161, 162, 175

FDA U.S. Food and Drug Administration. 196, 270

FDR false discovery rate. 125, 126, 171, 175–178, 180, 182, 183, 189, 232, 261, 268

FEV1 Forced expiratory volume in the first second. 114, 116, 118–120, 129–131

GATK genome analysis toolkit. 26

GEL Genomics England. 137, 138, 143, 145, 147–149, 151, 153–155, 157, 158, 160–164, 166, 169, 170, 175, 179, 182–185, 189, 194, 256

GnH Genes and Health. 137, 138, 142, 147, 149, 151, 153, 155, 156, 158, 160–162, 164, 166, 169, 175, 179, 182, 185, 189

GQ genotype quality. 26, 71, 73, 140, 207, 209, 242

GRM genetic relatedness matrix. 6, 32, 82, 107, 149, 189

GWAS genome-wide association study. 5–10, 16, 17, 33, 70, 107, 132, 136, 192, 196

HbA1c Glycated haemoglobin. 114, 118, 120, 131

HC high confidence. 74, 75, 91, 96, 146

HDL high-density lipoprotein. 114, 171, 173, 177, 182, 183, 268

HM3 Haplotype Map 3. 33, 34

HPO Human Phenotype Ontology. 148, 256

HR hazard ratio. 57, 58, 232

HWE Hardy-Weinberg equilibrium. 17, 24, 28, 69, 70, 73, 76–78, 87, 89, 97, 99, 124–126, 132, 140, 144, 145, 192, 207, 210, 238, 239, 245

IBD Identity By Descent. 42

ICD international classification of disease. 31, 84, 107, 111, 147, 148, 254, 256

indel insertions and deletions. 26, 30, 71, 72

INT Inverse Normal Transformation. 81, 86, 107, 149

IQR interquartile range. 170, 172, 257, 258

LC low confidence. 31, 45, 74, 75, 91, 93, 94, 127, 146

LCR low complexity region. 25, 72, 140, 209

LD linkage disequilibrium. 4, 5, 11, 27, 33, 34, 51, 72, 73, 94, 96, 127, 141, 151, 162

LDL low-density lipoprotein. 13, 171

LDSC LD score regression. 34

LOD Limit Of Detection. 85, 86, 92

LoF loss-of-function. 8, 12–15, 22, 49, 64, 74, 129, 136, 146, 164, 167, 186, 196, 198, 199

MAC minor allele count. 32, 35, 42–45, 49, 51, 72, 81, 122, 149, 161, 163, 216, 218, 220, 244

MAF minor allele frequency. 17, 18, 25, 29–33, 35, 42, 44, 45, 49, 51, 62, 65, 69, 72, 74, 81, 89–91, 93, 95, 124, 125, 127, 133, 141, 142, 146, 149, 161, 167, 177, 188, 191, 192, 212, 219, 245, 250, 268

MANE Matched Annotation from the NCBI and EMBL-EBI. 75, 146

Mb mega base pairs. 35, 51, 60

median absolute deviation MAD. 25

NDD Neuro Developmental Disorder. 148, 155, 157, 189, 256

NFE non-Finnish European. 26–30, 32–34, 38, 42, 43, 225

NGS Next Generation Sequencing. 7, 85, 92

NHS National Health Service. 84, 138, 158

NMD Nonsense Mediated Decay. 74, 127

NPX Normalized Protein eXpression. 85, 86, 96, 98, 102, 104

OMIM Online Mendelian Inheritance in Man. 47, 116, 178, 180

OR odds ratio. 37, 50, 152, 153, 181, 185, 223, 255

PC principal component. 25, 27, 32, 33, 39, 80, 86, 141

PCA principal component analysis. 27, 73, 141

PEA Proximity Extension Assay. 85

PL Phred-scaled likelihood. 71

pLoF putative loss of function. xiii, 31, 32, 35, 37–40, 44, 45, 47–49, 51, 53–56, 58, 60, 61, 63, 64, 66, 75, 82, 91–104, 106, 108, 111, 115, 119, 122, 124–129, 133, 146, 151, 162–165, 167–170, 173, 175, 179, 181, 182, 186, 196, 220, 221, 227, 234, 245, 248–250, 252, 255, 263, 265, 267

PP posterior probability. 42–46, 56, 65, 91, 98, 161–163, 193, 211, 213, 216, 217, 220, 221, 244, 247

pQTL protein Quantitative Trait Loci. 93–95, 99, 104

PRS polygenic risk scores. 6, 32–36, 39, 50–52, 56, 57, 63, 201, 226, 230–232

PTV protein truncating variant. 31, 74, 75, 91, 93, 94, 127, 146

QC quality control. 25, 26, 28, 71

QQ Quantile-Quantile. 82, 108, 110, 123, 126, 168, 249

RF random forest. 27, 28, 73, 139, 141, 241

RNA Ribonucleic Acid. 5, 8, 195

ROC-AUC Receiver Operating Charateristic Area Under the Curve. 225

SAIGE Scalable and Accurate Implementation of GEneralized mixed model. 32, 34, 35, 50, 225

SAS Southeast Asian. 71, 141, 142, 153, 155, 156, 161, 162, 175

SD standard deviation. 28, 207

SE standard error. 34

SER switch error rates. 30, 42–44, 74, 91, 161, 163, 211, 212, 218

SNP single nucleotide polymorphism. 5, 25, 27, 33, 34, 75, 140, 141, 143

SNV single nucleotide polymorphism. 30, 144

SPA Saddle Point Approximation. 50, 80, 108, 111, 132, 180

TF Transcription Factor. 195

UK United Kingdom. 9, 138, 158, 160, 200

UK BiLEVE UK Biobank Lung Exome Variant Evaluation. 24, 28

UKB-PPP UK Biobank Pharma Proteomics Project. 10, 85, 92

UKBB UK Biobank. 9, 10, 23–30, 32–34, 38, 39, 42–44, 47, 49–51, 63, 66, 70–74, 84, 85, 89, 91, 92, 107, 111, 113, 124, 127, 129, 137, 139, 140, 142, 143, 147, 149, 151–153, 155, 158, 160–162, 164, 166, 169, 170, 175, 179, 182–185, 191, 192, 198, 206, 209, 246, 248

URV ultra-rare variant. 26, 73, 207

US United States. 9

UTR Untranslated Region. 8, 65, 195

VCF variant call format. 25, 35, 79, 80, 143, 207

VEP Variant Effect Predictor. 31, 46

VQSR Variant Quality Score Recalibration. 140

WGS Whole Genome Sequencing. 9, 74, 91, 138, 143, 154, 161, 162, 194, 195, 200

WHR waist-to-hip ratio. 84

WHRadjBMI WHR adjusted for BMI. 84

1 Introduction

Contents

1.1 A very brief history of human genetics 1
1.2 Statistical genetics in the 21st century 5
1.2.1 Genome-wide association studies 5
1.2.2 Analysis of rare protein coding variants 7
1.2.3 Biobanks and innovation in phenotyping 9
1.3 The study of biallelic variation 11
1.3.1 Identification and medical relevance of biallelic variation 11
1.3.2 Nonadditive effects 14
1.4 Thesis aims 19

1.1 A very brief history of human genetics

Genetics, a term coined in the 19th century[1], is derived from the word, *genesis*, meaning the origin or mode of formation. The discipline of human genetics pertains to the study of the origins of inheritance in humans, a question that has puzzled philosophers and researchers alike for millennia: why do we bear resemblance to our parents? Early speculation about origins of inheritance date back to antiquity, with Aristotle positing that both parents are required to reproduce, because the father provides the blueprint and the mother provides the material to create the offspring[2]. Perhaps more humorously, in the 1700 century, it was thought that miniature men

referred to as 'homunculi' resided in the sperm, and when these entered the womb, they were nurtured to full human size[3]. Early concepts of inheritance began to form with Lamarck in the 1800s, that to the dismay of his contemporary theological naturalists, proposed that organisms could pass on acquired characteristics to their offspring, and not through divine means[4]. While ultimately disproved[†], Lamarck's theory was an important step towards evolutionary thinking. It was not until the mid 1800s that Darwin, during his HMS Beagle expedition to Central and South America, famously observed that finches of the different Galápagos islands had beak shapes adapted to specific food sources on each island[6]. He made the keen observation that traits advantageous to a specific environmental niche were selected for, and therefore persisted in subsequent generations[7]. Upon his return, he published these ideas on 'natural selection' in three papers, including his most famous *On the Origin of Species*[8] between 1858 and 1859, laying the foundation of modern evolutionary theory.

Around the same time as Darwin (1856-1863), Gregor Mendel, a monk from what is now considered the Czech Republic, was conducting experiments on the hybridization of pea plants[9]. Through meticulous crossbreeding experiments, Mendel observed distinct patterns in the way traits were inherited. In hindsight, Mendel was remarkably fortunate in his choice of traits to study. The traits he studied were dichotomous (two possible states), unlinked (not influenced by one another), and monogenic (controlled by a single gene). Nonetheless, this combination allowed him to observe clear, discrete outcomes in his experiments, which would end up contradicting the prevailing 'blending' theory of inheritance. Mendel's results showed that traits could remain distinct across generations, appearing or disappearing in predictable ratios. Based on his results, Mendel postulated a theory, what are now known as Mendel's Laws of Inheritance[10]: 1) The Law of Dominance - that dominant traits mask recessive ones; 2) The Law of Segregation - that each parent passes one randomly selected copy of each trait; and 3) The Law of Independent

[†]While Lamarck was proven wrong, today there is a wealth of literature suggesting and demonstrating the transmission of methylation patterns which can be acquired in the parent as a response to the environment and passed to their offspring, known as epigenetic inheritance[5].

Assortment - that different traits are inherited independently[11]. Traits that are governed by his laws are called Mendelian or monogenic ('single gene') traits, which is a testament to the importance of these ideas. Nonetheless, Mendel's laws initially faded into obscurity after their discovery, only to be rediscovered independently by multiple researchers in the early 1900s[12]. This rediscovery sparked heated debates among scientists with differing views on inheritance[13,14].

The 'biometricians' argued that continuous traits, such as height, couldn't be explained by Mendelian principles[15,16]. They proposed a 'law of ancestral heredity,' suggesting that offspring traits were influenced or 'blended' by multiple generations of ancestors. For height, the mean trait value in the offspring was better predicted with more information on ancestral trait values, which was irreconcilable with Mendelian ideas with single genes of large effects[17,18]. In contrast, the 'Mendelians' defended Mendel's ideas of inheritance, citing his pea experiments where traits exhibited discrete, large effects controlled by single genes. How could it be that both of these theories, while at odds with one another, agreed with empirical observation? R.A. Fisher effectively bridged the gap between Mendelian genetics and biometric observations in 1918 by postulating that multiple genes can contribute to trait variation, each locus obeying Mendel's law of segregation, thus providing an explanation for both discrete and continuous traits[19,20]. As the number of genes influencing a trait increased, their individual effects became smaller, forming the basis of his 'infinitesimal model'. This model explained how traits like height could show a normal distribution in populations, resulting from the combined action of many genetic factors. Fisher's work effectively bridged the gap between discrete Mendelian inheritance and continuous variation, establishing the foundation for quantitative genetics[19,20].

Around the same time as Mendel's work was being rediscovered, Thomas Hunt Morgan and his student Alfred Sturtevant, were studying the inheritance patterns of Drosophila melanogaster, the common fruit fly[21]. Similar to Mendel, they performed crossbreeding experiments of heritable mutations, such as eye color. Morgan observed that the ratio of traits across multiple generations sometimes obeyed

Mendel's laws and sometimes appeared in unexpected ratios. The chromosome theory of inheritance was developing concurrently with their work. This theory proposed that the units of heredity (genes), were located on chromosomes within the cell nucleus[22]. The theory suggested that chromosomes, which were observed to segregate during meiosis, could be the carriers of genetic information. Morgan and Sturtevant realized that the ratios of observed traits were related to the physical distance between the corresponding genes on a chromosome. When traits followed Mendel's laws, they were speculated to be on different chromosomes, allowing for independent assortment[23]. When two traits appear on the same chromosome, the frequency of the coinherited traits in offspring was inversely proportional to their the physical distance between the corresponding genes. This remarkable observation, led to the first 'genetic map' in 1913, laying out the location of six traits in the fruit fly relative to one another[24]. We now know that linkage disequilibrium (LD), the nonrandom inheritance of multiple genetic variants or 'genotypes', collectively referred to as 'haplotypes', are a result of process known as recombination, where maternal and paternal chromosomes exchange segments, creating hybrid chromosomes[25,26].

This understanding profoundly impacted our ability to map genes, and interrogate the inheritance of traits. This methodology became known as 'linkage analysis' and was first used to analyze large family pedigrees and isolate the regions associated with certain traits[27,28]. The method proved particularly useful for identifying linked conditions among traits that easily segregated in families[29,30]. However, linkage mapping saw limited utility for polygenic traits, and those that did not follow a strict dominant or recessive Mendelian dominance patterns[31]. The first 'genome-wide' linkage screening identified 15 loci involved in autism, a number that seemed high at the time[32]. In hindsight, we now know that certain traits, such as height, are extremely polygenic, with 12,111 independently associated loci involved in determining height[33]. It would take technological advancements in sequencing before many of these traits could be appropriately interrogated.

Concurrent with advancements in linkage mapping, crucial discoveries in molecular genetics were reshaping the understanding of inheritance. The identification of Deoxyribonucleic Acid (DNA) as the molecular medium of inheritance[34], followed by the elucidation of DNA transcription to Ribonucleic Acid (RNA) and subsequent translation to protein[34,35], laid the groundwork for modern molecular biology. The genetic code, composed of four nucleotides (A, T, C, G), became readable through the discovery of DNA polymerase[36], enabling the amplification and sequencing of the first human gene[37]. These methodological advances culminated in the completion of the Human Genome Project in 2003, providing the first complete sequence of the human genome, spanning 2.8 billion nucleotides[38,39]. Shortly thereafter, the International HapMap Consortium quantified LD between common variants[26,40,41]. Together, these advances set the stage for a new era genetic studies, with genome-wide association study (GWAS) being the central paradigm.

1.2 Statistical genetics in the 21st century

1.2.1 Genome-wide association studies

In the early 2000s, statistical geneticists shifted from family-based studies to larger cohorts of mostly unrelated individuals. This transition was partially facilitated by the completion of the Human Genome Project, but also the innovation in sequencing technology, which led to an exponential decline in cost. The first sequenced 'draft' human genome came with a price tag of $300 million, while today it costs less than $1000 for a full genome[42]. Through successive iterations of the HapMap project[26,40,41,43], and the establishment of the 1000 Genomes (1KGP)[44,45], we began outlining the LD structure of common variation by mapping chromosomal regions that were shared between individuals. This understanding revealed that a strategically selected subset of index variants could be used to impute the majority of common human genetic variation, exploiting the predictable LD patterns[46–48]. This informed the manufacturing of single nucleotide polymorphism (SNP) genotyping arrays, which were developed in such a way so that the targeted variants were approximately equidistant, exploiting the fact that LD could capture most residual

common variation[49,50]. This marked the dawn of the GWAS era: instead of identifying families in which disease-relevant alleles segregated, it became possible to perform targeted sequencing of individuals with a trait of interest, and test if alleles exhibited higher frequencies in cases compared to controls[51].

The principles behind GWAS are straightforward: when modeling the average effect on a trait in unrelated individuals due to incremental allelic substitution (0, 1, or 2 copies of the minor allele), analysis can be accomplished through simple fixed-effects linear or logistic regression at each locus across the genome. Statistical significance is then determined after multiple test correction, under the assumption of well-calibrated type 1 errors[52]. The first large-scale GWAS, published in 2007[53], marked a significant milestone in the field. This seminal study, involving 2,000 cases and 3,000 controls, identified 24 unique associations across seven major diseases and set the foundations for performing additive and recessive association analysis using case-control design. Subsequently, numerous model developments have emerged, including a shift towards mixed-effects modeling, using a genetic relatedness matrix (GRM) to account for sample relatedness in the model[54,55]. Over time, GWAS has now been applied to associate loci across diverse traits, including type 2 diabetes[56], schizophrenia[57], cancer[58], height[33], sexual orientation[59], infertility[60], and many others[61].

Beyond GWAS

Beyond identifying variant associations, GWAS offers several additional applications. The estimation of effect sizes from GWAS variants enables the construction of polygenic risk scores (PRS). These scores aggregate the effects of numerous common variants to predict individual disease risk or trait values[62,63]. GWAS data also facilitates the assessment of 'genetic correlation' between traits, reflecting shared genetic loci with pleiotropic effects[64–66]. This approach has revealed connections in the genetic architecture across many conditions, broadening our understanding of disease etiology and trait interrelationships. Furthermore, GWAS summary statistics can serve as a foundation for 'Mendelian randomization' studies. These

analyses, rooted in Mendel's laws of inheritance, help determine causal relationships between exposures and outcomes in observational epidemiology[67,68]. This method has become a powerful tool for inferring causality where randomized controlled trials are impractical or unethical[65]. While GWAS significantly advanced our understanding of complex traits, it faced statistical and practical limitations in robustly associating rare variants with diseases[69]. Addressing these challenges required further developments in Next Generation Sequencing (NGS) technologies and analytical methods, paving the way for rare variant association studies and more comprehensive genetic analyses. These following advancements form the foundation of the work conducted in this thesis.

1.2.2 Analysis of rare protein coding variants

Negative selection tends to remove variants with large effects from populations, resulting in an inverse relationship between effect size and Allele Frequency (AF)[70]. Consequently, rare variants often contribute substantial effects to traits, while common variants typically confer low or modest contributions. Nonetheless, rare variants have been documented to contribute to both monogenic and common complex disorders[71]. In the case of schizophrenia, a complex trait, common variant studies have identified 128 common loci associated with risk using genotyping arrays[57]. However, with exome sequencing (ES) technology, we recently discovered that rare protein-coding variants with large effects sizes in 32 genes are sufficient to cause the disorder[72]. This pattern is not unique to schizophrenia; similar findings have emerged for other traits[73]. Similarly, obesity, another complex trait with a well-established polygenic basis, has been found to have monogenic forms caused by rare protein-coding variants with large effect sizes[74].

The identification of rare protein-coding variants that confer substantial risk has been greatly facilitated by (short read) ES technology. ES enables the amplification and sequencing of overlapping (primarily) DNA fragments at a specified coverage depth, enhancing the accuracy of variant detection[75]. This technique primarily targets exons, that contain protein-coding gene regions, which are transcribed into

RNA and retained after intron [†]. The focus on exonic regions stems from the fact that genetic alterations in these areas are more readily interpretable compared to non-coding regions, as exonic changes directly impact the resulting protein structure. Variants causing frameshifts, for instance, are likely to have significant effects due to their deleterious consequences on the encoded protein, potentially manifesting as disease phenotypes that we can test in GWAS[77]. The focus on protein-coding regions also has direct therapeutic implications, because it can be used to guide targeted drug development[78].

Association testing of individual rare variants, especially for less common phenotypes, presents significant statistical challenges. This is because statistical power depends on both variant effect size and frequency, and the multiple testing burden can increase substantially compared to GWAS of common variants[79]. These challenges motivated the development of rare variant collapsing methods. Instead of testing individual variants, these methods aggregate the effects of rare variants within functional units such as genes, noting the presence or absence of any rare variant per individual[80,81]. By combining information across multiple rare variants, these methods can potentially increase statistical power, especially when multiple rare variants in a gene or pathway contribute to a trait or disease risk. In these tests, rare variants are often categorized based on their predicted impact on the resulting protein. Examples include stop-gain or frameshift variants, which subsequently lead to loss-of-function (LoF) in the cognate protein, or missense variants predicted to interfere with protein functionality. These challenges had led to incredible innovations in variant consequence prediction, and recent developments in Artificial Intelligence (AI) have proven highly efficient in predicting the likelihood of missense variants being deleterious[82,83], as well as identifying variants likely to create cryptic splice sites[84]. In rare variant burden testing, functional categories are collapsed separately across genes, and the collective burden of rare variants

[†]Despite a common misconception, exons are not equal to protein-coding regions[76]. In humans, only ∼23% of exonic sequences are protein-coding, while the remainder includes Untranslated Region (UTR)s (∼32%) and non-coding RNAs (∼37%). Whole Exome Sequencing' is therefore a misnomer, as it typically captures less than 25% of the human exome.

is tested, potentially improving statistical power over single rare variant testing. Today, and in this thesis, rare variant burden testing is often performed within a mixed effects framework using software such as `REGENIE`[85] or `SAIGE`[86].

1.2.3 Biobanks and innovation in phenotyping

GWAS and rare variant burden studies have been greatly advanced by the advent of biobanks, which are repositories of biological samples (e.g., blood, urine, saliva) collected from hundreds of thousands of individuals. These samples can be subjected to various analyses, including genotyping, ES[87], Whole Genome Sequencing (WGS)[88], imaging[89], proteomic[90], metabolomics[91], and biochemical profiling. Biobanks often contain comprehensive data on participants' general health, medical history, demographics, and lifestyle. In fact, many participants consent to having linked their electronic health record (EHR), which provides health-related information from primary and secondary care, including prescriptions, operations, and registries of birth, cancer, and death[92]. EHRs typically offer cross-sectional data, but sometimes also longitudinal information, enabling genetic studies on age-of-onset or the factors influencing body mass index (BMI) changes throughout life[93]. The combination of EHR-linkage and sequencing data is particularly powerful for genetic research, facilitating retrospective studies.

The UK Biobank (UKBB), perhaps the most renowned biobank[92], exemplifies the potential of large-scale genetic repositories, and is used throughout this thesis. Comprising approximately 500,000 individuals from across the United Kingdom (UK), it has collected extensive genetic data, including genotyping[92], ES[87], and WGS[88]. Following the creation of UKBB, more than 30 biobanks have been created worldwide, with notable examples in the UK[92,94], United States (US)[95–99], China[100], Mexico[101], Canada[102], Denmark[103,104], Finland[105], Norway[106], Estonia[107], Iceland[108], Qatar[109], Japan[110], Korea[111], the Netherlands[112,113], Uganda[114], Israel[115], Taiwan[116], and Australia[117]. While some biobanks focus on the general population, others target specific communities[118] or those with rare pediatric diseases or cancer[94].

While these biobanks offer new opportunities, it is important to acknowledge their limitations. Studies have demonstrated that UKBB participants are, on average, healthier than non-participants[119], potentially biasing association results and limiting generalizability to broader populations[120,121]. Additionally, sex-differential participation bias observed in UKBB and 23andme[122] leads to disproportionate representation of certain traits or genetic variants between sexes, risking spurious autosomal associations with sex and skewed downstream analyses.

A positive consequence of the rapid increase in sequenced individuals is the emergence of meta-biobank initiatives that analyze results across multiple biobanks through the sharing of GWAS summary statistics. The Global Biobank Meta-analysis Initiative is one such project, investigating the contribution of common variants to disease across 30 unique biobanks[123]. Though useful for the study of common diseases, scale is particularly crucial in rare variant analysis: larger samples provide more opportunities to observe and associate rare variants with trait endpoints. Therefore, the Biobank Rare Variant analysis (BRaVa) consortium aims to examine the rare variant burden contribution to complex diseases across 30 biobanks. Rare variant meta-analysis are the most recent development in large-scale genetic analysis to date, with multiple published in mid 2024[124,125]. These meta-analyses collectively analyze millions of individuals and represent the next major advancement in genetic studies, simply because they are growing in size and becoming accessible for researchers globally. The value of rare variant meta-analysis is considered in chapter 4.

Many biobanks retain biological samples in long-term storage, allowing for the application of new technologies to existing collections without requiring participants to return for additional material collection. A recent example of this foresight's value is the collaboration between the UKBB and the UK Biobank Pharma Proteomics Project (UKB-PPP), which provided proteomic analysis of approximately 3,000 circulating blood proteins for about 50,000 mostly randomly sampled UKBB participants[90]. While this technology has its caveats[126,127], it has enabled researchers to associate genetic variants with the phenotypic expression of the blood plasma

proteome[128]. The plasma proteome, being a product of the environment, lifestyle, and genetics, has given rise to the field of 'proteogenomics'[129,130]. This emerging field aims to interrogate the complex interplay of these factors, identify novel disease biomarkers, conduct Mendelian randomization studies, and drive further biological insights into human disease mechanisms[90,131,132]

1.3 The study of biallelic variation

1.3.1 Identification and medical relevance of biallelic variation

This thesis focuses on rare coding biallelic variants identified through ES. Coding variants in a gene can manifest in several ways. Homozygous variants occur by definition when both gene copies harbor the same variant. Compound heterozygous (CH) variants arise when both copies harbor different variants at distinct locations within the same gene locus. A pair of variants are said to be 'in *cis*' when two variants are located on a single gene copy, whereas variants on opposite gene copies are referred to as '*in trans*'. Homozygous and CH variants are collectively referred to as biallelic variants. While identifying homozygous variants implicated in Mendelian recessive disease is straightforward, detecting CH variants presents a greater challenge. This is because it requires determining whether the constituent heterozygous variants are inherited from both parents or just one, as outlined in Figure 1.1.

The process of resolving the parental origin of a variant is referred to as 'phasing'. This is the separation of an individual's genome by maternally and paternally derived alleles. There are three main approaches to phasing: trio-phasing, which uses Mendelian inheritance principles but is limited by the lack of trio relationships (mother-father-offspring) in many datasets[133,134]; read-backed phasing, which exploits variant overlap across sequencing reads but is constrained by current short-read sequencing limitations[135–137]; and statistical phasing, which uses knowledge of the LD structure, shared segments of DNA between distantly related individuals, and recombination maps to guide phasing[138–141]. Large-scale studies often lack comprehensive trio data, and the most prevalent ES sequencing

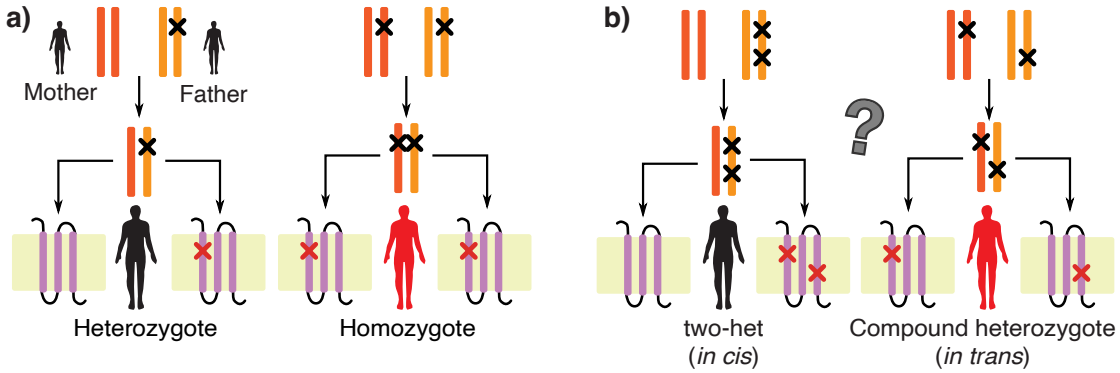


Figure 1.1: Determination of biallelic status in a trait with a recessive mode of inheritance. Given a trait with a recessive mode of inheritance, an individual that inherits a single gene copy is healthy, whereas someone that inherits both gene copies will express the trait. It is straightforward to determine homozygous status in sequencing (a), however, because the parental information is lost during sequencing, it becomes more difficult to determine CH status (b). Unless the haplotype of the offspring is phased into maternally and paternally derived alleles it becomes impossible to distinguish between two variants *in cis* and *in trans*.

technology today produces short reads inadequate for effective read-backed phasing. Consequently, statistical phasing is generally the preferred method. However, the accuracy of statistical phasing depends on several factors, including sample size, genotype accuracy, relatedness, variant density and allele frequency[139]. Accurately phasing rare variants, particularly singletons (variants that occur only once in the population), presents a significant challenge in the community[141]. This difficulty has hindered our ability for large-scale study of the phenotypic consequences of compound heterozygosity[142].

Medical relevance of biallelic coding variants

Individuals carrying biallelic LoF variants, which render both copies of a gene incapable of producing functional protein, have been coined as 'human knockouts'. This terminology mirrors the concept of 'knockout mice', which have long served as model organisms in genetic research[144]. Human knockouts are of great interest to the scientific community, particularly for drug development with multiple compelling examples illustrating their usefulness[145–147]. Perhaps, one of the most famous examples with medical implications is the discovery of healthy individuals with

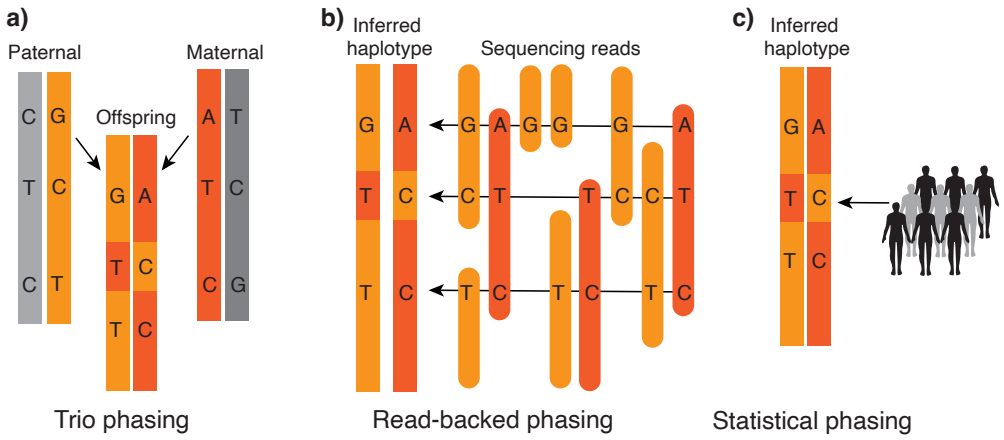


Figure 1.2: Basic principles of trio-phasing, read-backed phasing, and statistical phasing. a) Trio-phasing utilizes Mendelian transmission principles to resolve haplotypes using genetic data from a trio (mother, father, and offspring). The transmitted haplotype is colored, wheras the untransmitted haplotype is shown in grey. b) Read-backed phasing leverages overlapping heterozygous variants in sequencing reads to reconstruct original haplotypes. This is adapted from Wertenbroek *et al.*[143]. c) Statistical phasing employs population-level information, such as haplotype frequencies to infer phase in unrelated individuals.

PCSK9 biallelic LoF variants, who were found to have exceptionally low low-density lipoprotein (LDL) levels. This finding led to the development of therapeutics targeting PCSK9, which are now used as cholesterol-lowering agents[148,149]. Another example occurred with the inadvertent cure of an HIV-infected patient following a bone marrow transplant. Strikingly, it was discovered that the donor was a *CCR5* knockout, lacking an important receptor for viral cellular entry[150]. This discovery has influenced the development of a class of HIV drugs targeting the CCR5 receptor, which are also in use today[151]. These examples illustrate how human knockouts can exhibit protective effects against diseases. As research progresses, more applications continue to emerge[152,153], underscoring the potential of studying human knockouts for medical advancements and drug discovery.

The importance of studying coding biallelic variation, underlying such knockouts, has led to the establishment of specific biobanks built around the premise of excess homozygosity. For instance, the Genes and Health cohort focuses on a community of British-Pakistanis with a high degree of parental relatedness, resulting in an inflation of homozygous variants[118]. This cohort is analyzed in detail in Chapter

4. The FinnGen biobank leverages the unique genetic architecture of the Finnish population, shaped by a historical population bottleneck[105,152]. This demographic event resulted in a founder effect, where disease-predisposing alleles that persisted through the bottleneck became more prevalent. As a consequence, individuals in the current Finnish population are more likely to share these variants, leading to an increased frequency of homozygous genotypes[154]. These biobanks alongside studies in Icelanders[108], Amish[155], play important roles in unraveling the medical relevance of biallelic variation.

Identifying biallelic variants is more difficult in outbred populations, where homozygosity for LoF mutations in disease-relevant genes is generally rare. Instead, it can been speculated whether compound heterozygosity plays a larger role in shaping disease risk. The medical relevance of CH mutations has not been extensively studied in large-scale populations. This gap in research stems from the challenge of inferring whether two heterozygous variants are *in cis* or are *in trans*. As a result, most studies linking CH variation to disease have been limited to small case studies, focusing almost exclusively on Mendelian traits that easily segregate in families[156–161]. This lack of broader exploration has partly motivated this thesis. In Chapters 2 and 4, I explore the diversity of potentially medically relevant CH variants. Lastly, even when biallelic variation is identified in large-scale analysis, and associated with a trait, a framework is needed to determine whether these effects are *recessive* or*additive* with respect to their contribution to the trait.

1.3.2 Nonadditive effects

Theories on the origins of dominance

Mendel's Law of Dominance introduced the terms *dominant* for traits requiring a single allele for phenotypic expression and *recessive* for those requiring two alleles; thus the term 'dominance' encompasses both dominant and recessive modes of inheritance. This phenomenon is widely observed in nature, with the famous case of the peppered moth during the English Industrial Revolution providing a striking example. As industrial soot darkened their habitat, dark-colored moths were

selected for due to superior camouflage against predatory birds, driven by a single heterozygous allele sufficient to fully express the black phenotype[162]. Interestingly, most deleterious alleles in nature tend to be *recessive*, raising a fundamental question in evolutionary genetics: why do mutations that reduce fitness typically exhibit this characteristic? The origins of dominance and the mechanisms underlying this phenomenon have been subjects of ongoing scientific debate for nearly a century[163,164]. Several theories have been proposed to explain this observation, each offering distinct perspectives on the interplay between genetics, metabolism, and evolutionary processes.

Two main theories propose different views on why deleterious mutations tend to be recessive: R.A. Fisher's theory of selection for dominance modifiers, and Wright's metabolic theory based on enzyme kinetics in metabolic pathways. Fisher's theory suggests that natural selection has favored genetic architectures that mask the effects of harmful mutations in heterozygotes[165]. Over evolutionary time, modifier loci have evolved to make wild-type alleles more dominant, thereby protecting organisms from the damaging effects of new mutations. This theory emphasizes the role of Darwinian selection in shaping dominance patterns through the creation of genetic robustness and redundancy. Wright's metabolic control theory, on the other hand, proposes that the intrinsic properties of metabolic pathways naturally buffer against partial loss of function, making most mutations recessive by default[166,167]. This view argues that recessivity is a consequence of the way enzymes share control of metabolic flux, and that substantial changes in fitness only occur when enzyme activity is dramatically reduced, which may require complete inactivation of both gene copies[14,167]. While Wright believed that selection could shape dominance, his later contemporaries took a unique physiological view that dominance was an inevitable consequence of metabolic pathways, and that it did not require an evolutionary explanation[167,168]. Haldane *et al.* proposed a congruent theory, positing that recessivity is a result of Darwinian selection for genes that produce enzymes at higher levels, enabling them to better tolerate acquisition of LoF mutations and deal with unusual fluctuations in the environment[169,170].

These theories have strengths and limitations in explaining observed patterns of dominance[171]. Fisher's theory struggles with the weak selection pressure on dominance modifiers and fails to explain dominance in organisms that selectively express traits in the haploid stage, as is the case for the flagella of Chlamydomonas[168], effectively falsifying Fisher's hypothesis. The metabolic theory, while providing a mechanistic explanation, assumes a linear and positive relationship between metabolic flux and fitness, excluding the possibility of optimal intermediary pathway flux[172], and fails to explain dominance outside of enzymatic reactions and heterozygous advantage, also known as overdominance[171,173]. This debate is still ongoing, and theories that attempt to reconcile these differences continue to emerge[171].

Statistical dominance

While the origins of dominance are debated, statistical methods can effectively model dominance effects. Genetic variance can be partitioned into two main components: the average effect of allelic substitution (known as additive or narrow-sense heritability) and the 'dominance deviation' (referred to as 'nonadditive' in this thesis). The latter represents the additional variance contribution not explained by additive heritability within a locus[14,19]. Epistasis, which involves interactions between segregating loci, has been documented but generally contributes minimally to total genetic variance, is as a result often disregarded in association analyses, and will not be discussed in this thesis[174,175]. However, it is important to emphasize that the definition of nonadditivity used in this work is equivalent to dominance and therefore only considers effects within a single site or gene, thus excluding epistasis.

Despite the Wellcome Trust Case Control Consortium's landmark paper setting the standard for GWAS[53], and advocating for testing both additive and nonadditive effects, large-scale investigations of nonadditive effects remain scarce. This is because additive models perform remarkably well for common variants, even when alleles at a locus influence the trait in a nonadditive manner (Figure 1.3, top row). Additionally, most analyses and theory suggest that nonadditivity contribute very little to overall genetic variance[176,177]. Therefore, for common variants, if the goal

is to discover novel variants, an analysis of the additive effects is often preferred. Perhaps for this reason, it was not until recently that nonadditive effects were systematically interrogated in large-scale cohorts for common complex phenotypes owed to common variation (minor allele frequency (MAF)$>$ 1%)[178,179].

However, the situation becomes more complex when considering rare variants. Statistical power for detecting nonadditivity is lowest for rare variants, precisely where the strongest deviations from additivity are expected to occur[179] (Figure 1.3, bottom row). This is because negative selection keeps rare variants with large effects at low frequencies, and deleterious alleles tend to be recessive[171]. The low frequency of rare variants with large effects limits our ability to observe sufficient homozygous carriers, reducing our statistical power to detect deviations from additivity. Furthermore, assessing the nonadditive contribution of rare variants that often deviate from Hardy-Weinberg equilibrium (HWE)[180–182], the expected genotype frequencies (p^2, $2pq$, and q^2 for AA, Aa, and aa genotypes respectively) in a population under genetic equilibrium, remains inconsistent across the scientific community. GWAS analysis of the recoded genotypes for a canonical Mendelian recessive model ($[0, 1, 2] \rightarrow [0, 0, 1]$) does not yield the nonadditive contribution, because the original and transformed encodings ($[0, 1, 2]$ and $[0, 0, 1]$) are still correlated. Nor does analyzing a joint model with the recessive encoding ($[0, 0, 1]$) together with the additive encoding ($[0, 1, 2]$) and assessing the significance of the effect size estimate of the recessive contribution in this joint model, which remains the 'gold-standard' approach in GWAS.[183]. Collectively, both frameworks risk additive spillover effects. These challenges will be addressed in chapter 3 with the development of a unified framework for reliable analysis of nonadditive effects across the allele frequency spectrum.

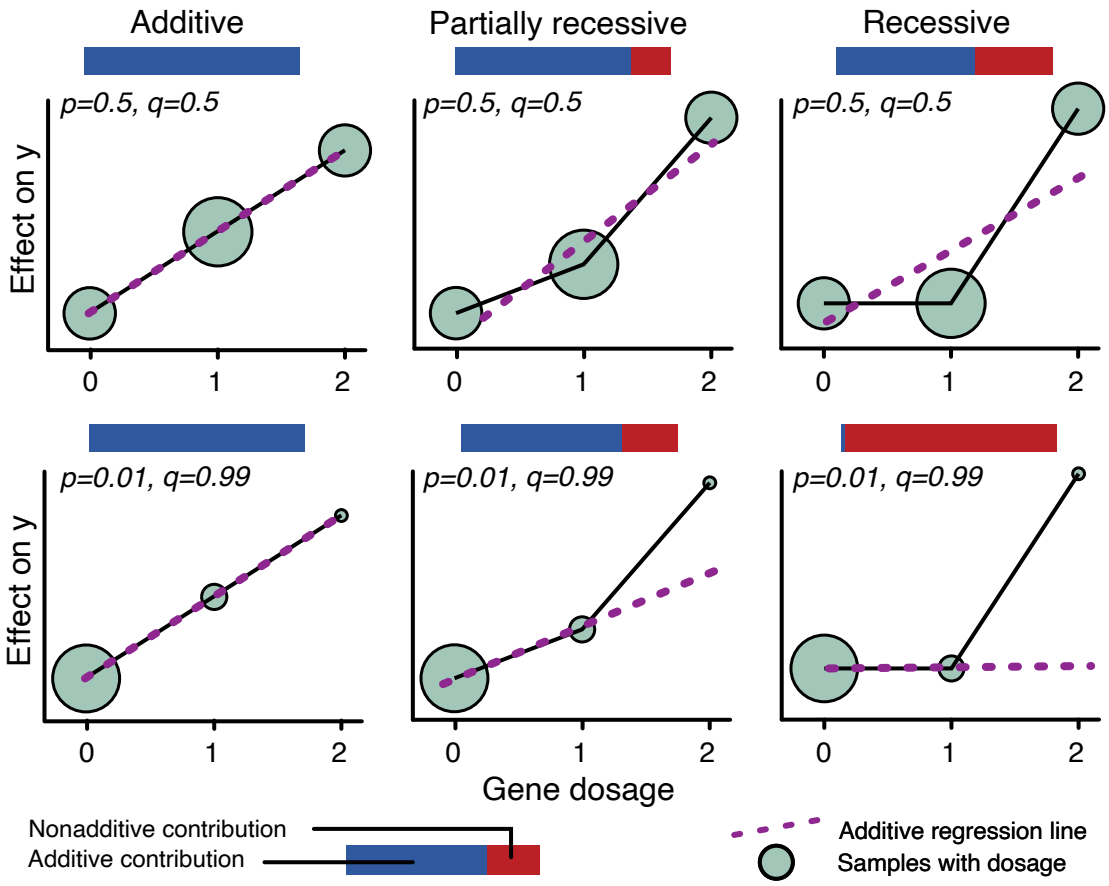


Figure 1.3: Additive model fit across different architectures at two different MAFs. This figure illustrates the relationship between genotype and phenotype for three genetic architectures (additive, partially recessive, and Mendelian recessive) at two different minor allele frequencies (MAF= 0.5 and 0.01). Each column represents a distinct genetic architecture, while rows correspond to different MAFs. The x-axis represents allelic substitutions (0, 1, 2), and the y-axis shows the phenotypic effect. Green circles indicate genotype frequencies, with size proportional to frequency in the population. The dashed purple line represents the fit of a linear additive model. Allele frequencies (p and q) are displayed in the top left corner of each plot. The bar above each plot displays variance explained from an additive model (blue), and the residual variance explained is the nonadditive contribution (red). The genetic architectures, shown from left to right, are: (1) Additive, where the effect on the phenotype doubles with each allelic substitution; (2) Partially recessive, with some effect in heterozygotes but larger in homozygotes; and (3) Mendelian recessive, where the effect is only present in the homozygous alternate state. At MAF= 0.5 (top row), all three architectures has a significant proportion of variance explained by the additive model. However, at MAF= 0.01 (bottom row), the additive model fails to capture the Mendelian recessive architecture effectively with large majority of variance ($\sim$ 98%) explained by the nonadditive model. This figure was inspired by Palmer *et al.*[179] with simulation code available here.

1.4 Thesis aims

In this thesis, I investigate the phenotypic consequences of rare biallelic coding variants in human health and disease. The work is structured in three interconnected chapters: Chapter 2 examines the contribution of CH coding variants to common complex disease risk, expanding the number of identifiable biallelic variants in UK Biobank. Chapter 3 develops a framework for determining the nonadditive, particularly recessive, contribution of rare variants to traits. This method is then applied to two unique datasets to demonstrate its utility and findings. Chapter 4 scales up the analysis through collaborative efforts across UK Biobank, Genomics England, Biobank Japan, and Genes and Health. This larger-scale approach aims to identify novel gene-trait relationships with a recessive mode of inheritance. Through these three chapters, this thesis aims to deepen our understanding of how biallelic coding variation shapes human health and disease, providing insights into the complex landscape of genetic inheritance and its phenotypic consequences.

2

Exome-wide evidence of compound heterozygous effects

Contents

2.1 Introduction 22
2.2 Methods 24
2.2.1 Quality control of genotyping arrays 24
2.2.2 Quality control of 200k exomes 24
2.2.3 Phasing 29
2.2.4 Trio-switch error rates 29
2.2.5 Read-backed phasing 30
2.2.6 Phenotype curation 30
2.2.7 Variant annotation masks 31
2.2.8 Biallelic encoding and recessive models 31
2.2.9 Gene copy dosage encoding and additive models 32
2.2.10 Polygenic risk scores 32
2.2.11 Conditional analysis 34
2.2.12 Permutation of genetic phase 36
2.2.13 Gene-set enrichment of biallelic variation 36
2.2.14 Homozygote and CH down-sampling 37
2.2.15 Power analysis for biallelic association 37
2.2.16 Simulation 38
2.2.17 Age-at-onset analysis 39
2.2.18 Replication analyses 40
2.2.19 Simulation of CH and homozygous variants 40
2.3 Results 42
2.3.1 Accurate phase inference and validation using parent-offspring trios and short-read sequences 42

2.3.2 Identification and examination of CH variation in UK Biobank 45
2.3.3 Systematic evaluation of biallelic effects on common disease 50
2.3.4 Permutation testing establishes the role of genetic phase on disease risk 53
2.3.5 Incorporation of genetic phase boosts power in gene-level rare variant burden analysis 55
2.3.6 CH variants elevate lifetime risk of disease 56
2.3.7 Common complex disorders implicated by CH variation 60
2.4 Discussion 62
2.4.1 Limitations 65

The material from this chapter is adapted from my first-authored paper available in Cell Genomics. I have obtained written permission from my supervisors and all co-authors to re-use this material for my thesis.

2.1 Introduction

Thousands of independent genetic variants have been robustly associated with common, complex human diseases, leading to important advancements in therapeutic development[78]. Naturally occurring variants that disrupt protein-coding sequences are of interest in the context of drug discovery as they modulate potential biological targets with measurable effects on human physiology[184,185]. Thus, individuals who carry LoF variants on both the maternal and paternal copy of a gene, are in principle 'experiments of nature' and their identification can help to determine causality between gene function and phenotype[186–188].

Coding variants in a gene can either be homozygous, where both gene copies harbor the same variant, or CH where both copies harbor different variants, usually at distinct genetic locations within the same gene locus. Alternatively, if two variants are located on a single gene copy, they are said to be 'in *cis*'. Although both copies of a gene are disrupted in two-hit (CH or homozygous) carriers, analyses of the phenotypic impact of coding variation have typically ignored genetic phase information, that is, the separation or 'phasing' of an individual's genome into maternally and paternally derived alleles[189,190]. Large-scale studies of biallelic

damaging variation have generally been restricted to homozygotes in populations with excess homozygosity, such as Icelanders[191], Finns[152,192], and consanguineous populations[193]. In contrast, CH are expected to be more common in outbred populations, but are understudied outside of rare disorders[157,194–197]. While recent efforts have resulted in the characterization of biallelic variation in large-scale population cohorts, they have so far been unable to systematically link CH variants to disease[142,198].

A series of methods exist to infer the genetic phase of two variants. 'Phasing by transmission' employs family member genotyping and Mendelian inheritance principles[140], while 'read-backed phasing' utilizes physical relationships among variants within sequencing reads[199]. In large-scale biobanks, extensively genotyping family members is impractical, and short-read sequencing technologies only allow read-backed phasing for variants in close proximity. Therefore, 'statistical phasing', which models the generative process of newly arising genetic variation subject to recombination and mutation[138–140,200,201], is typically used to phase haplotypes in genetic biobank data. Obtaining high-quality statistically phased genetic data requires large sample sizes, (10^5-10^6 individuals), and tends to require large reference panels[200]. Furthermore, statistical phasing is more error prone for rare variants, which are precisely the collection of variants that we would like to investigate as they are *a priori* more likely to be deleterious variants of large effect under purifying selection. This difficulty in the accurate statistical phasing of rare variation has historically deterred the analysis of CH variants in biobanks. However, recent advancements in statistical phasing[198], achieved by combining common variation across genotyping arrays and exome sequencing to create haplotype 'scaffolds'[201] enables accurate phasing of rare variants. By using this new accurate phase information which extends down into rare allele frequencies, we can identify damaging CH variants to expand the pool of identifiable two-hit carriers and screen for phenotypic consequences.

I describe and apply a systematic analytical approach to test for autosomal biallelic effects, gene-by-gene, across up to 311 traits in the UKBB 200k ES release,

combining both CH and homozygous variation. I iteratively refine candidate associations by adjusting for polygenic background, nearby common variant risk, and rare variant burden within the analyzed gene. Our approach identifies both known and novel biallelic-trait associations that I replicate using a subset of UKBB 450k ES release, distinct from samples in the 200k ES release.

2.2 Methods

2.2.1 Quality control of genotyping arrays

QC performed by UK biobank team

The following description is adapted from Bycroft *et al.*, (2018)[92]. Blood was drawn from 488,377 UKBB participants who were genotyped across two different arrays. 49,950 individuals were involved in the UK Biobank Lung Exome Variant Evaluation (UK BiLEVE) study and were genotyped for 807,411 markers (GRCh37) using the Affymetrix UK BiLEVE Axiom Array[202]. The remainder of 438,427 indiviudals were genotyped for 825,927 markers (GRCh37) using Applied Biosystems UK Biobank Axiom Array. Importantly, the two arrays are related and share an overlap of 95% of markers. A series of ancestry-specific and batch-specific marker-based and sample-based quality controls were then applied. Markers were filtered based on batch effects, plate effects, deviations from HWE, sex effects, array effects, and inconsistencies across control replicates. Samples were filtered based on missing rate, heterozygosity, and inconsistencies between reported and genetic inferred sex. The resulting 488,377 samples and 805,426 variants, spanning two genotyping arrays, passing these aforementioned quality controls metrics were released as filtering masks.

2.2.2 Quality control of 200k exomes

Quality control of the 200k Exome Sequencing data was performed by Nik Baya, student in the Lindgren Lab and Duncan Palmer, senior postdoctoral researcher in the Lindgren lab. I used the resulting filtered dataset.

Sample filters

We evaluated sample-level quality control (QC) metrics on the 200,643 UKBB ES multi-sample project level variant call format (VCF) call-set files[203], Table 2.1. All metrics were calculated for biallelic SNPs, except for metrics involving insertions and deletions. We regressed out the first 21 principal components (PCs)[204], and filtered out sample outliers of the residuals for each metric based on MAD (median absolute deviation) thresholds (Table 2.1). Samples without PC data were subject to more stringent thresholds (Table 2.1).

Metric	**Metric residual (w/ PCs)**	**Raw (w/o PCs)**
call_rate	[24, ∞)	[4, ∞)
n_insertion	[8, 8]	[4, 4]
n_deletion	[8, 8]	[4, 4]
r_insertion_deletion	[8, 8]	[4, 4]
n_het	[12, 12]	[4, 4]
n_hom_var	[12, 12]	[4, 4]
r_het_hom_var	[16, 16]	[4, 4]
n_non_ref	[8, 8]	[4, 4]
n_singleton	(-∞, 16]	(-∞, 4]
n_snp	[8, ∞)	[4, 4]
n_transition	[8, 8]	[4, 4]
n_transversion	[8, 8]	[4, 4]
r_ti_tv	[8, 8]	[4, 4]

Table 2.1: Sample filtering: MAD Intervals. The interval $[a, b]$ represents $\text{median}(X) + \text{MAD}(X)[-a, b]$ for the metric, X. Samples with metrics outside these intervals were removed.

Variant-level filters

Retain variants satisfying all of the following conditions:

1. Not in a low complexity region (LCR)[205].
2. In sequencing target regions ± 50 base pairs.
3. MAF > 0 following genotype QC.
4. Excess heterozygosity (ExcessHet < 54.69) filter: Phred-scaled P-value for exact test of excess heterozygosity[206] in founders as determined by relatedness

estimates and recorded ages of UKBB participants[204]. Variants were retained as recommended in genome analysis toolkit (GATK)[206]

Genotype filters

Multiallelic variants were split into biallelic variants and insertions and deletions (indel) were left-aligned[207]. Genotype calls meeting any of the following criteria were set to missing:

1. Genotype quality (GQ) ≤ 20
2. Total sequencing depth (DP) ≤ 10
3. For heterozygous calls:
   (a) (reference + allele Depth)) + < 0.8
   (b) Indels: alternate allele read depth / total read depth < 0.3
4. Homozygous indel calls: alternate allele read depth / total read depth < 0.7

Additional ES quality control

To perform further QC we use `Hail`, an open-source Python library which focuses on the analysis of large-scale genetic data sets. We used `Hail` to create our own methods, and we take advantage of the functionality that has been rewritten to enable fast and scalable analysis of large exome and genome sequencing projects. Unless otherwise stated, all of the following the data curation and quality control steps were performed in `Hail`[203].

Briefly, we apply a collection of hard-filters on sample metrics. We confirm genotypic sex with reported sex, remove samples with excess ultra-rare variants (URVs), and restrict analysis to samples of genetically ascertained non-Finnish European (NFE) ancestry. Finally we apply a second collection of sample and variant hard filters. As an initial pass to remove low quality and contaminated samples, we filter out samples with call rate < 0.95, mean DP $< 19.5\times$ or mean GQ < 47.8 (Figure A.3).

Sex imputation

To confirm participant sex and calculate PCs, we extracted high quality common variants (allele frequency between 0.01 to 0.99 with high call rate (> 0.98)) and LD prune to pseudo-independent SNPs using `-indep 50 5 2` in PLINK 1.9. When reported sex does not match genotypic sex, it may signal potential sample swaps in the data. Using the F-statistic for each sample using the subset of the non-pseudo autosomal region on chromosome X, we identify and remove samples where reported sex information is not confirmed in the sequence data (Figure A.1). Specifically, we remove samples satisfying at least one of the following criteria:

- Sex is unknown in the phenotype files.
- F-statistic > 0.6 and the sex is female in the phenotype file.
- F-statistic < 0.6 and the sex is male in the phenotype file.
- F-statistic > 0.6 and number of calls on the Y chromosome is < 100.

Defining samples with non-Finnish European ancestry

To ensure adequate case-control for as many traits as possible, we restricted our analysis to a set of genetically ascertained NFE samples. To do this, we perform a number of principal component analysis (PCA) steps to ensure that we have subset down to NFE. We first run PCA on the 1KGP samples (minus the small subset of related individuals within the 1KGP) using subsetting to LD pruned autosomal variants. We then project in the UKBB samples, ensuring that we correctly account for shrinkage bias in the projection[208]. Next, we removed samples outside of the European population European (EUR) using a random forest (RF) classifier: we train a RF on the super-populations labels of 1KGP and predict the super-population for each of the UKBB samples (Figure A.2). We denote strictly defined European subset as those with probability > 0.99 of being European according to the classifier. Another RF classifier is trained following restriction of the 1KGP samples to Europeans to determine NFE, using a classifier probability of

0.95. RF classifiers were trained using the randomForest (4.6) library in R. Samples not assigned to the NFE cluster were removed from downstream analysis.

Final hard filters

For the final variant filtering step, following restriction to the NFE subset, and removal of incorrectly defined sex or unknown sex, and run variant QC. We then filter out variants with call rate < 0.97, variants out of HWE ($P < 1 \times 10^{-6}$), and remove invariant sites following the previous sample based filters. After restricting to these high quality variants, we perform a final set of sample filters to finalize the quality controlled data. We evaluate a collection of sample metrics and remove samples falling outside four standard deviations (SDs) of the sequencing batch mean (Ti/Tv, Het/HomVar, Insertion/Deletion ratios), and remove the collection of samples with over 175 singletons. The resultant curated analysis ready data set consists of 176,935 samples, and 9,169,408 variants (Table A.1-A.2). A summary of sample and variant filters are provided in Table A.1-A.2 and Figure A.4. The high quality ES call-set consisted of 176,935 samples and 9,169,408 variants.

Quality control of genotyping array

The following description is adapted from Bycroft *et al.*, (2018)[92]. Blood was drawn from 488,377 UKBB participants who were genotyped across two different arrays. 49,950 individuals were involved in the UK BiLEVE study and were genotyped for 807,411 markers (GRCh37) using the Affymetrix UK BiLEVE Axiom Array[202]. The remainder of 438,427 indiviudals were genotyped for 825,927 markers (GRCh37) using Applied Biosystems UK Biobank Axiom Array. Importantly, the two arrays are related and share an overlap of 95% of markers. A series of ancestry-specific and batch-specific marker-based and sample-based quality controls were then applied. Markers were filtered based on batch effects, plate effects, deviations from HWE, sex effects, array effects, and inconsistencies across control replicates. Samples were filtered based on missing rate, heterozygosity, and inconsistencies between reported and genetic inferred sex. The resulting 488,377 samples and 805,426

variants, spanning two genotyping arrays, passing these aforementioned quality controls metrics were released as filtering masks.

2.2.3 Phasing

Combining ES data with genotype array data

I combined genotyping array (UK BiLEVE Axiom array and UKBB Axiom array) and exome chip (IDT xGen Exome Research Panel v1.0) variants after general ES quality control using `Hail`[203] and `BCFtools`[209] (1.12). For variants in both data sets, I preferentially retained those on the ES data. For variants on the genotyping array I excluded variants missingness $> 5\%$ after performing a liftover to GRCh38 using `Hail`[203]. To avoid biasing the phasing quality estimates, I excluded parents among trio relationships prior to phasing. I first created a common variant scaffold by phasing variants in the combined (exome sequencing and genotyping array) data with MAF $> 0.1\%$ and otherwise default parameters using `SHAPEIT5_PHASE_COMMON` module. I then phased the remaining rare variants using the common variant scaffold using the `SHAPEIT5_PHASE_RARE` with recommended parameters. To ensure computational tractability, I phased overlapping chunks of 100,000 variants with $\geq 50,000$ variant overlap between consecutive chunks using `Hail`[203]. Following chunk phasing, I then removed the initial and final 22,500 variants from each chunk, so that 5,000 overlapping variants remained between contiguous phased chunks. I then combined the phased chunks, matching haplotype phase using `BCFtools`[209] (1.12) with the `-ligate` option. I then restrict this phased genetic dataset to the set of samples and variants present in the analysis ready NFE subset (Table A.1-A.2).

2.2.4 Trio-switch error rates

I assessed phasing quality by comparing statistically phased genotypes to those implied in 96 trios using Mendelian inheritance logic. Switch errors are determined by traversing the statistically phased and parent-offspring transmitted haplotypes simultaneously and scanning for inconsistencies in phase between pairs of contiguous variants. This method only allows us to consider sites in which the one parent

is heterozygous and the other is homozygous for the reference or alternate allele, and thus do not consider *de novo* variants or Mendelian inconsistencies in the trio data. To assess switch error in a site-specific manner, I modified and recompiled `BCFtools`[209] (1.12) to output errors by genomic position. I then used the modified version to assess switch by variant categories, for example by genetic data modality (genotyping array or ES), or by MAF bins. To evaluate switch errors across different phasing confidence thresholds, I filtered VCF using `Hail`[203] and then repeated the switch error calculation step. I calculated binomial 95% confidence intervals (CIs) for switch error ratess (SERs) using the R-package `HMisc`[210] (4.7).

2.2.5 Read-backed phasing

I performed read-backed phasing with UKBB ES short paired-end read sequences using `.cram` files provided by UKBB. As WhatsHap[211] (1.6) is computationally expensive, I restricted the analysis to pairs of variants on chromosomes 20-22 in 176,586 genetically ascertained NFEs. I phased both single nucleotide polymorphism (SNV) and indel with WhatsHap using the default recommended parameters. WhatsHap outputs lists of phased variants within 'phased sets'. I carried forward reads overlapping no more than two variants, for which phase could be inferred. I combined these phased variants with statistically phased variants from `SHAPEIT5` using `Hail`[203], and determined agreement between estimated phasing in `WhatsHap` and `SHAPEIT5`.

2.2.6 Phenotype curation

I considered a collection of 282 binary quality controlled and publicly available common complex phenotypes for analysis[212]. To complement these, I also considered 28 common complex phenotypes that were obtained through manual curation, resulting in a total of 311 binary phenotypes for analysis. To increase the power for analyses for binary traits, I amalgamated a collection of phenotypes where possible: combining the phenotype curation of Censin et al.[213], with the primary care mappings file provided by UK Biobank `all_lkps_map_v3.xlsx` and my own

manual curation. I aggregated across international classification of disease (ICD)-10, ICD-9, operating codes, nurses interview reports, and self-reported diagnosis by doctor from the main phenotype file, as well as v2 and v3 read codes in the primary care data. As in Censin et al., I made use of the careful definitions of Eastwood *et al.*[214], subsequently applied by Udler *et al.*[215] for diabetes subtype curation. Briefly, the algorithm developed in Eastwood *et al.* bins individuals into putative diabetes status using a collection of phenotypes in the UK Biobank data including self-reported diabetes diagnosis, age of diagnosis, medications, start of insulin within a year of diagnosis. I defined cases as those placed in the probable and possible case categories in the algorithms output. Controls were defined as samples labeled as 'diabetes unlikely' by the algorithm.

2.2.7 Variant annotation masks

I annotated coding variation using Variant Effect Predictor (VEP)[216] (v95) using the worst consequence by gene within 'canonical' transcripts. I classified variants into four categories: protein truncating variants (PTVs), missense variants, synonymous variants, and other variants (Table A.9). I then split PTVs into putative loss of function (pLoF) (HC) and low confidence (LC) loss-of-function variants using `LOFTEE`[217], and labeled missense variants with both `REVEL`[218] score $\geq$ 0.6 and `CADD`[219] score $\geq$ 20 as 'damaging missense' or otherwise as 'other missense'. Finally, I combine the resultant 'damaging missense' category with LC loss-of-function variants, which I denote as 'damaging missense/protein-altering'.

2.2.8 Biallelic encoding and recessive models

Using custom `Hail` scripts, I define and annotate individuals as being 'biallelic' for a gene if they harbor at least one pLoFs or damaging missense variant with MAF $<$ 5% on both inherited copies of the gene. For each sample, I encoded the presence and absence of a damaging biallelic variant for each gene as zero and two, respectively. I encode this information in a `.vcf` file and test for an association between presence of a damaging biallelic variant in a gene and a trait

using Scalable and Accurate Implementation of GEneralized mixed model (SAIGE) (1.1.9)[220], adjusting for sex, age, sex × age, age^2, UKBB centre, genotyping batch and the first 10 PCs. I took relatedness into account using a sparse GRM fitted on NFE. I restrict analysis to (gene, trait) pairs with at least five biallelic variants in the curated ES with non-missing corresponding phenotype data (corresponding to a minimum minor allele count (MAC) ≥ 10), and adjust for multiple testing at Bonferroni significance ($P < 0.05$/gene-trait pairs).

2.2.9 Gene copy dosage encoding and additive models

I define annotate individuals as being 'monoallelic' for a gene if they harbor at least one pLoFs or damaging missense variant with MAF $< 5\%$ on a single copy of the gene. Furthermore, if they harbor at least one pLoF or damaging missense variant on both inherited copies of the gene, I annotate them as 'biallelic'. Using custom `Hail` scripts, I encode wildtypes, monoallelic and biallelic carriers as 0, 1 and 2 respectively, thus representing the number of affected gene copies in an individual. I test for association using SAIGE[220], adjusting for sex, age, sex × age, age^2, UKBB centre, genotyping batch and the first 10 PCs. Again, I took relatedness into account using a sparse GRM fitted on NFE. I restricted to gene-pairs with at least 10 disrupted haplotypes (corresponding to a minimum MAC ≥ 10), and adjust for multiple testing at Bonferroni significance ($P < 0.05$/gene-trait pairs).

2.2.10 Polygenic risk scores

Curation of array-based genetic data

I generated PRSs using imputed genotypes provided by UKBB[204]. In the following, I make the distinction between training and testing data. The first represents the samples that are used for fitting LDPred2[221] weights and parameters while the latter represent the samples with biallelic variant (with homozygous or CH status) information in which I use to assess the predictive accuracy the fitted LDPred based PRS. For the training data, I took the genetically ascertained NFE and filtered to 246,152 unrelated samples (kinship coefficient $< 2^{-4.5}$) that did

not have quality controlled ES data available. NFE samples with high quality ES and imputed genotype data available were used for testing. Where predictive (nominal significant h^2_{snp} and $n_{eff} \geq 5000$), I include PRS as a covariate for downstream biallelic association testing to account for common variant polygenic risk for the trait under investigation.

Genotype variant filtering

I followed best practices from Privé *et al.*[221], and filtered to common Haplotype Map 3 (HM3) SNPs[222]. Additionally, I exclude any variants with genotyping proportion $< 1\%$ and MAF $< 1\%$, resulting in a total of 1,165,296 common autosomal variants for fitting PRS weights. To reduce the likelihood of spurious correlations between low-frequency variants in traits with low case or control count, I restricted to binary phenotypes with at least 1,250 cases and controls. Additionally, I imposed a phenotype specific MAF filter based on the number of cases and controls in a trait, specifically:

$$\text{MAF} > \max\left(0.01, 2 \times \min(n_{\text{cases}}, n_{\text{controls}})\right), \tag{2.1}$$

where n_{cases} and n_{controls} are the numbers of cases and controls with high quality imputed sequence data available, respectively, to guard against non-causal variants that are overrepresented in cases or controls leading to false positive associations.

Common variant association testing

I tested for associations between the 1,165,296 common autosomal HM3 variants and phenotypes using `Hail`[203], running logistic regression (`logistic_regression_rows`) adjusting for sex, age, sex $\times$ age, age^2, UKBB assessment centre, genotyping batch and the first 10 PCs, using a Wald test.

Estimating heritability

I generated LD-scores for HM3 variants in sample, using a random subset 10,000 of 246,152 unrelated genetically ascertained NFEs without haplotype information. Using the GWAS summary statistics and LD-scores, I estimated SNP heritability

h^2_{snp} and standard errors (SEs) using LD score regression (LDSC)[223,224]. I evaluated PRS for phenotypes with nominal significant h^2_{snp} estimates ($P < 0.05$) and restricted to phenotypes with nominally significant ($P < 0.05$) LDSC based SNP heritability estimates and effective sample size $n_{\text{eff}} \geq 5,000$, where:

$$n_{\text{eff}} = \frac{4}{\frac{1}{n_{\text{cases}}} + \frac{1}{n_{\text{controls}}}}. \tag{2.2}$$

Generating PRS using LDPred2

For a given phenotype, I trained a PRS predictor with LDPred2-auto[221], using marginal effect size estimates evaluated on the 246,152 unrelated NFE samples (defined by kinship coefficient $< 2^{-4.5}$) without ES data in the 200k ES UKBB release), h^2_{snp} as estimated by LDSC, and in-sample reference panel to evaluate local LD, as input. I removed any invariant sites and mean-imputed missing genotypes, before training the predictor. Following PRS training, I then predict into the 176,266 samples with ES and high-quality imputed genotype data.

Validation of polygenic risk scores

I assessed the ability of the resulting PRS to discriminate between case status by evaluating area under the curve (AUC) on the held-out unrelated set of samples with both HM3 SNPs and phased exome data. I used the function AUCBoot from the R package bigstatsr[225] (1.5.6) to extract 10,000 bootstrap replicates of individuals and compute the 95% CIs for AUC.

2.2.11 Conditional analysis

Off-chromosome PRS conditional analysis

For each chromosome, C, I evaluated 'off-chromosome' PRS by setting weights on chromosome C to 0. I repeated this for each phenotype with PRS available and fit SAIGE[220] models while controlling for off-chromosome PRS by including it is as a covariate in the null SAIGE model.

Common variant conditional analysis

To assess whether a putative signal in a gene is driven by nearby common variation, I filtered to samples that have both ES and imputed genotypes with MAF $> 1\%$ and imputation INFO score > 0.5. Then, for each gene that passed exome wide significance in the primary analysis ($P < 5 \times 10^{-6}$), I tested for common variant associations in the region (1 mega base pairs (Mb) upstream and downstream of the gene). For each of these regions, I took an iterative approach, testing for common variant associations using SAIGE[220], conditioning on the lead variant and repeating the regression until the conditional P for the newly included variant dropped below 5×10^{-6}, allowing up to 25 'independent' associations in the region. I used the same covariates as in the primary analysis. For every variant that passed exome-wide significance ($P < 5 \times 10^{-6}$), I encoded the genotypes as dosages and embedded them alongside pseudo variants (biallelic variants) in a VCF. I then re-ran the primary analysis twice (with and without controlling for off-chromosome PRS), while conditioning on any nearby common variant signals of association with the phenotype of interest.

Rare and ultra-rare variant conditional analysis

For each significant ($P < 1.68 \times 10^{-7}$) gene-trait associations in the genome-wide analysis after conditioning on PRS and nearby common variants association signals, I considered a further conditioning step. I sought to determine whether the residual signal of association could be explained by additive rare variant effects within the associated gene. To do this, ran further conditioned on rare (MAC$\geq$ 10, MAF$\leq$ 0.05) and ultra-rare (MAC$\leq$ 10) variants annotated as either pLoF or damaging missense within each gene. Because conditioning on ultra-rare variation can lead to convergence issues, I performed a gene-wide collapsing of ultra-rare (MAF$\leq$ 10) variants, thus aggregating them into a single 'super' variant to represent burden of ultra-rare damaging variation in the gene. Following this collapsing, I was able to condition on the ultra rare and rare variant contribution using SAIGE, while also conditioning on PRS and nearby common variant association signals when applicable.

2.2.12 Permutation of genetic phase

To test whether a putative gene-trait association is driven by compound heterozygosity, we designed a permutation-based pipeline that could be systematically applied and scaled across phenotypes and genes. To do this, I label samples that are either CH variants or heterozygous *cis* carriers and then randomly shuffle these labels a series of times. For each permutation, I re-run the association analysis conditioning on covariates as previously discussed (including off chromosome PRS and nearby common variants), and determine the resultant association strength under this label shuffling. Applying this permutation procedure multiple times, I can determine an empirical null for the association strength in the absence of phase information. The result is an empirical distribution of score-statistics and corresponding P-values that reflect the degree of association that would be expected given that the phase is random. I evaluate the one-sided empirical P-value, specifically:

$$P_{\text{empirical}} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(t_i \geq t_{\text{observed}}) \tag{2.3}$$

where n is the number of permutations, t_i is the score-statistic under the i^{th} random label shuffling, and t_{observed} is the observed score-statistic determined using the observed genetic phase. To ensure sampling of score-statistics at a sufficiently large number of configurations of the genetic phase, I analyzed gene-trait pairs with at least ten compound heterozygotes and/or samples with multiple variants on the same haplotype. I permuted up to 100,000 times. To control for multiple testing, I corrected for 5 gene-traits tested (Bonferroni significance threshold $P < 0.05/5 = 0.01$).

2.2.13 Gene-set enrichment of biallelic variation

Analyzed gene-sets

I included the following gene lists in the gene-set enrichment analyses: essential in mice[226], essential gnomAD[227], essential ADaM[228], essential in culture[229], essential CRISPR[230], genes with pLI $>$ 0.9 in gnomAD[227], non-essential in culture[229],

homozygous LoF tolerant[227], and non-essential gnomAD[227] and Curated Cancer Cell Atlas[231].

Poisson regression to assess enrichment of CH variants in gene-sets

I test for depletion and gene-set enrichment using poisson regression. I model the count of biallelic variants across samples as a function of gene-set and mutation frequency using the `glm` function in R.

$$|\text{samples with} \geq 1 \text{ variant of class } x \text{ in gene}| \sim \mathbb{I}(\text{gene-set}) + \text{mutation rate} \quad (2.4)$$

where x is a pair (x_1, x_2): $x_1 \in$ {pLoF, damaging missense, pLoF and/or damaging missense, other missense, synonymous}, $x_2 \in$ {heterozygote, CH, biallelic variants}. For each annotation category I use the transcript-specific mutation rate[232]. 95% confidence intervals are determined using `confint.glm` from the MASS-package (v7.3-58.1).

2.2.14 Homozygote and CH down-sampling

To investigate the number of identifiable CH or homozygous events across varying sample sizes and variant annotations, I performed down-sampling across the total population of 176,587 individuals. To do this, I defined a set of 35 regularly spaced cutoffs between 1,000 and 176,587 samples using increments of 5000. To determine uncertainty in the estimates of the number of unique genes implicated as a homozygote and/or CH, I randomly sampled individuals for each down-sampling 100 times, with replacement. I calculated the 95% CI by taking the 2.5% and 97.5% quantiles for the number of unique genes affected at a given sample size, and repeated across annotations (Figure 2.5).

2.2.15 Power analysis for biallelic association

I perform a power analysis based on biallelic (including both CH and homozygous) variant frequencies in the population. To do this, I adopted code[233] allowing us to determine the effective effect size on the odds ratio (OR) scale across candidate configurations of binary case-control counts by substituting alternate

allele frequencies with biallelic variant frequencies. I calculated effect sizes at 80% power at Bonferroni significance ($P < 1.68 \times 10^{-7}$) for a hypothetical traits with 823 (0.5%), 1766 (1%), 3532 (2%), 5298 (3%), 8829 (5%) cases of 176,587 total samples.

2.2.16 Simulation

Simulation of phenotypes using real genotypes

I performed a series of simulations to test that the pipeline would detect a CH effect in the presence of a true signal. I sampled 100,000 genetically-ascertained NFEs in the UKBB data, and extract chromosome 22 which I then use to simulate phenotypic data with a recessive genetic architecture. To emulate a scenario in which defects in protein coding genes lead to disease, I annotated the filtered UKBB genetic data and determined the collection of samples harboring damaging biallelic variants in each gene (compound heterozygous and homozygous, comprised of variants annotated as pLoF or damaging missense). I then define a n samples $\times$ m genes matrix $\widetilde{\mathbf{B}}$ with entries:

$$\widetilde{\mathbf{B}}_{i,j} = \begin{cases} 1, & \text{if a damaging biallelic variant is present in sample } i \text{ at gene } j \\ 0, & \text{otherwise} \end{cases} \tag{2.5}$$

I then simulated liability under the following model:

$$y_i = \sum_{j=1}^{m} \mathbf{B}_{i,j}\theta_j + \varepsilon_i \tag{2.6}$$

where $\mathbf{B}_{i,j}$ is the $(i,j)^{\text{th}}$ entry of $\mathbf{B}$ after standardizing the columns of $\widetilde{\mathbf{B}}$, $E[\theta_j] = \frac{b}{m}$, $\text{Var}[\theta_j] = \frac{h^2}{m}$, and $\varepsilon_i \sim \mathcal{N}(0, 1-h^2)$. Here, I implicitly assume that presence of at least one homozygous or CH variant of any type within a given gene contributes the same risk to disease, whose average across genes is set by the parameter b. The resultant liability y_i has mean 0 and variance 1. Note that the standardization of $\mathbf{B}$ imposes a frequency dependent relationship between prevalence of biallelic damaging variants in a gene and variance explained. I simulated under the spike-and-slab model:

$$\theta_j \sim \begin{cases} \mathcal{N}(\frac{b}{m\pi_\theta}, \frac{h^2}{m\pi_\theta}), & \text{if } p_j < \pi_\theta \\ 0, & \text{otherwise} \end{cases}$$
$$p_j \sim \text{Bernoulli}(\pi_\theta)$$

in which $\pi_\theta \in [0, 1]$ is the proportion of causal genes with a recessive contribution to the phenotype. Finally, to obtain binary traits I used the liability threshold model assuming a case prevalence of 10%. In the following simulations, I set $\pi_\theta = 0.25$, and considered h^2 values of $h^2 \in \{0, 0.01, 0.02, 0.05, 0.10\}$ and b values of $b \in \{0, 0.5, 1, 2, 10\}$.

2.2.17 Age-at-onset analysis

The following analysis was accomplished in conjunction with S.S. Venkatesh, a fellow PhD student in the Lindgren Lab. S.S. Venkatesh extracted longitudinal health records as described below, and wrote scripts for longitudinal modeling. I adapted and ran these scripts, and analyzed the outputs.

Time-to-event data curation

I curated age-at-diagnosis for 278 binary phenotypes from the UKBB-linked primary care and hospital record data. 251 phenotypes were curated using the mapping tables generated by Kuan *et al.*[212], excluding any codes related to "history of..." events for which accurate age-at-diagnosis could not be extracted. The remaining 27 phenotypes were left-truncated at the age of first record (of any code) in either the primary care or hospital data, and right-censored at the age of the last record.

Cox proportional-hazards modeling

For each gene-trait combination to test, I performed Cox-proportional hazards modeling to estimate differences in lifetime risk of developing the phenotype between heterozygous carriers of pLoF+damaging missense/protein-altering variants in the gene (reference group) and individuals who are biallelic carriers (compound-heterozygous or homozygous), multi-hit *cis*-heterozygous carriers, and wildtypes. All effects were adjusted for sex, the first 10 genetic PCs, birth cohort (in ten-year intervals from 1930-1970), and UKBB assessment center. For phenotypes with a significantly heritable PRS, I additionally adjusted for off-chromosome PRS. I visualized survival probabilities using Kaplan-Meier curves[234]. Finally, for gene-trait combinations where I was powered to detect differences between

compound-heterozygous and multi-hit *cis* heterozygous carriers of variants, i.e. where each group contained at least five cases of the phenotype, I repeated the above analysis with multi-hit *cis* heterozygous carriers as the reference group. Cox proportional-hazards regression was performed using the R package `survival` 3.3.1[235] and Kaplan-Meier plots drawn with the R package `survminer` 0.4.9[236].

2.2.18 Replication analyses

In the initial analysis, I examined 176k individuals from the initial release of 200k exomes. The subsequent release of the full 450k exomes during the drafting of this paper presented an opportunity to replicate the biallelic associations. To do this, I leveraged the gnomAD joint variant called exomes which I combined with genotyping array data for the 450k participants, allowing phasing of rare variants. Before phasing, I performed the same initial quality-control using various hard filters, as in the discovery cohort. My replication cohort consist of the 233,837 individuals not included in the discovery analysis. In the cross-sectional analysis, I performed replication for genes with at least five biallelic (pLoF+damaging missense/protein-altering) variant carriers were present. Similarly, in the longitudinal analysis, I performed replication for gene-trait combinations with at least five biallelic variants (homozygotes or CH) and 100 heterozygotes in the population. In the replication cohort, variant annotation masks were created using VEP (v105).

2.2.19 Simulation of CH and homozygous variants

I conducted simulations to compare the expected and empirically observed number of biallelic variants. I focused on 1,174 genes with at least one homozygous of CH variant in the dataset. Then I generated genotypes for 176,935 individuals. Each genotype simulation assumed that variants were independent and utilized a Bernoulli distribution, where the probability of success was set to the variant's minor allele frequency. I ran the simulations 10 times and derived the expected number of homozygous or CH events with. This average provided us with an estimate for the expected number of biallelic variants, including both homozygotes

and CHs. I then compared these estimates to the actual empirical findings to benchmark the concordance.

2.3 Results

2.3.1 Accurate phase inference and validation using parent-offspring trios and short-read sequences

Gold-standard assessment of phasing accuracy

My analysis of the UKBB 200k release ES data revealed 9,169,408 high-quality variants across 176,935 individuals (Methods 2.2.2). To determine whether variants occurred on the same haplotype (in *cis*) or on opposite haplotypes (in *trans*) for each gene, I employed `SHAPEIT5`[237] to phase both ES and genotype array data jointly (Methods 2.2.3). This choice of phasing software was informed by a comparative analysis of popular phasing tools which included `EAGLE2`[200], `SHAPEIT4`[140] (Table A.3). For rare variants (MAF < 0.001), I assigned a posterior probability (PP) of true haplotype assignment, referred to as the phasing confidence score. I and others[141] observed an inverse relationship between MAF and our confidence in accurate statistical phasing (Table A.4). As rare variants are likely to contribute disproportionately to recessive damaging effects in CH configurations, selecting an appropriate PP threshold involves balancing sensitivity and specificity for identification. Following the phasing process, I focused the analysis on 176,935 individuals of genetically-determined NFE ancestry (Methods 2.2.2).

I evaluated the quality of the statistical phasing by comparing it to phasing derived from parent-offspring trio data and read-backed phasing using short read sequences. I quantified phasing accuracy before and after applying a PP threshold of ≥ 0.9 at the genotype level. Using 96 parent-offspring trios, I calculated the SER, based on Mendelian transmission, across 2,044,234 distinct variants grouped by MAC (Figure 2.1a, Table A.4). In 93.1% of protein-coding genes, variants were phased without switch errors across the 96 children (Table A.5). Unlike previous statistical phasing tools, `SHAPEIT5` is able to phase singleton (MAC=1) variants. This involves identifying the longest shared Identity By Descent (IBD) segment between the target individual's haplotypes and those in the population. The minor allele is then assigned to the haplotype with the shortest shared IBD segment, reflecting a more

distant common ancestor and thus a higher likelihood of germline variant occurrence. My analysis revealed a singleton SER of 32.4% (95% CI = 30.2 − 45.6%) before PP filtering, aligning with previous estimates from ES and genotyping array data[198]. After applying the PP ≥ 0.9 filter, the SER for singletons (MAC=1) decreased to 12.1% (95% CI = 8.42 − 17.2). For comparison, rare variants with 2 ≤ MAC ≤ 5 showed SERs of 0.27% (0.13–0.53) post-PP-filtering (Figure 2.1a).

Read-backed phasing as an alternative method for validation of statistical phasing

While trio-based SER calculation is considered the gold standard for assessing phasing quality[200], its applicability is constrained by the availability of parent-offspring trios. To complement this approach, I also employed read-backed phasing using `WhatsHap`[211] on UKBB short read sequences from 49,756 NFE individuals, analyzing 453,514 unique variant pairs on autosomes (Figure 2.2a). Read-backed phasing, though limited to variants within the span of one or a few overlapping short reads (typically up to 250 bp), offers an alternative method for validating statistically phased variants. My findings from read-backed phasing corroborated the trio-SER results, showing concordance between statistically and read-backed phased variants as MAC increased (Figure A.8). For singletons, I noted a positive correlation between phasing accuracy and PP, albeit with a decrease in the proportion of retained singletons at higher PP thresholds (Figure 2.2b). When I applied a PP ≥ 0.9 filter at the genotype level, the agreement between read-backed and statistical phasing reached 86.34% (95% CI = 85.68 − 86.98%) for singletons and 99.37% (95% CI = 99.33 − 99.41%) for variants with 2 ≤ MAC ≤ 5 (Table A.6).

My benchmarking analyses indicate that the statistical phasing of the UKBB dataset achieves high accuracy across the spectrum of allele frequencies, from rare to ultra-rare variants, and even singletons. This increases our ability to identify damaging CH variants with confidence. After evaluating the calibration of PP scores and examining their distribution across different MAC bins, I established an empirical threshold of PP ≥ 0.9 at the genotype level. This criterion resulted in the retention of 6,352,396 variants for subsequent characterization and testing, with

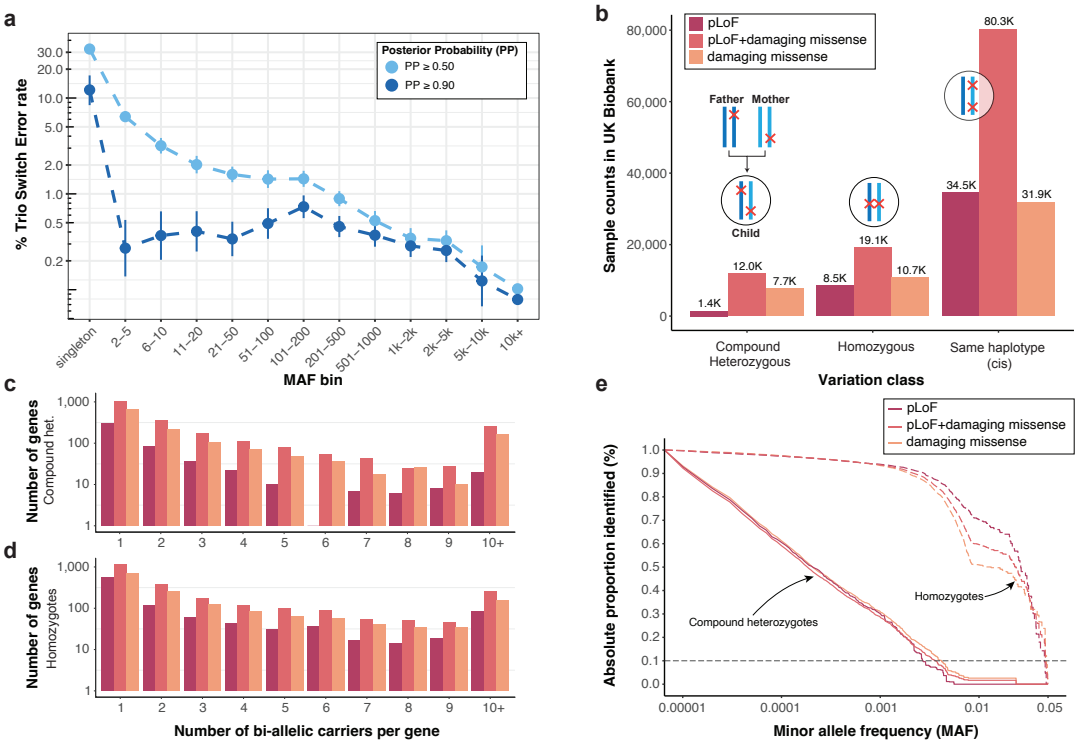


Figure 2.1: CH variants composed of at least one ultra-rare variant can be robustly identified in large scale biobanks. **a)** Trio SER depicted on y-axis as a function of MAC bin (x-axis) for phased variants with MAF $\leq 5\%$, stratified by phasing confidence score PP ≥ 0.5 or PP ≥ 0.9. **b)** Counts of samples harboring different classes of variation with at least two variants in UKBB. Each set of three bars depicts the number of individuals with at least one CH variant, homozygous variant, or multi-hit (*cis*) variant, respectively. Here, I define a CH pLoF+damaging missense variant as any combination of pLoF and/or damaging missense variation on opposite haplotypes. A qualifying carrier for each bar occurs according to the configuration displayed above the bars, and is grouped by variant consequence according to the color legend. **c-d)** Number of CH or homozygous carriers per gene. **e)** 1 - cumulative fraction (y-axis) of homozygous (dashed line) and CH carriers as a function of lowest MAF (x-axis) in bi-allelic variant pairs for which both variants phased at PP ≥ 0.9 (solid line), stratified by variant consequence according to the color key.

singletons comprising 12% of this set (Table A.9). Together, this underscores the robustness of the phasing and provide a solid foundation for downstream analyses of CH genetic variation in the UKBB cohort.

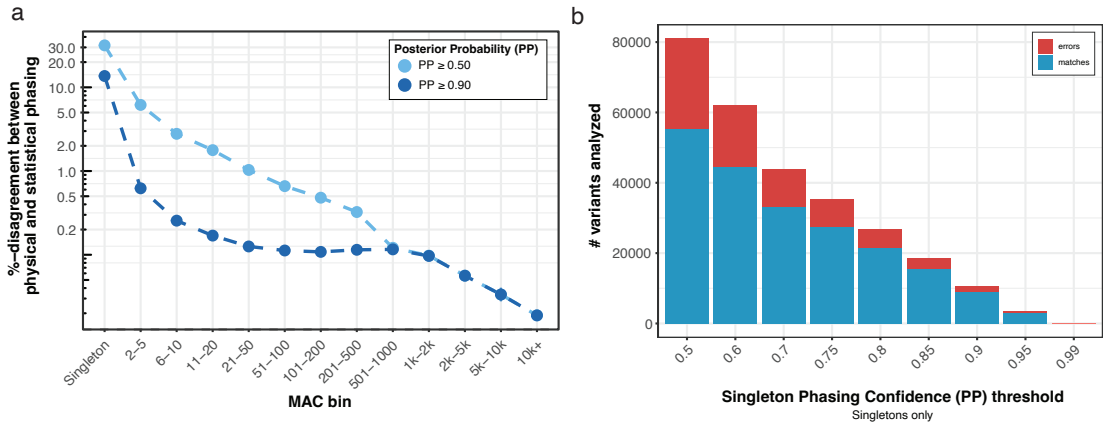


Figure 2.2: Agreement between read-backed and statistically phased variants. a) Agreement between statistically phased and read-backed phased variants across MAC-bins, stratified by phasing confidence scores (PP$\geq$ 0.5 and PP$\geq$ 0.9). b) Agreement between read-backed and statistically phased singleton variants across 49,756 samples. 80,978 reads with at least one singleton (MAC= 1) variant were identified, irrespective of phasing quality. For each PP-bin, agreement (blue) and disagreement (red) between read-backed and statistically phased variants are shown. Higher PP-scores correlate with increased phasing accuracy, while the total number of assessed variants decreases.

2.3.2 Identification and examination of CH variation in UK Biobank

To interrogate the functional role of mono- and biallelic variation in the population, I annotated 6,352,396 variants (of which 762,843 were singletons) with PP $\geq$ 0.9 and MAF $\leq$ 5% across 17,998 autosomal protein-coding genes. I enriched my search for variants with putatively large effect sizes by restricting analyses to two categories of predicted damaging variation: first, I annotated 146,299 (22,380 singletons) high-confidence protein truncating variants, including stop-gain, essential splice and frame-shift variants identified as high-confidence by `LOFTEE`[227], which I refer to as pLoF variants'. Second, I annotated 242,859 (32,199 singletons) missense variants classified as damaging by both `REVEL`[238] score $\geq$ 0.6 and Phred scaled `CADD`[219] score $\geq$ 20, or `LOFTEE` LC protein truncating variants; I refer to these variants collectively as 'damaging missense/protein altering' (Figure 2.3, Table A.9). For each individual, I then determined the set of genes predicted to be affected by pLoFs+damaging missense/protein-altering variants in a CH, homozygous or

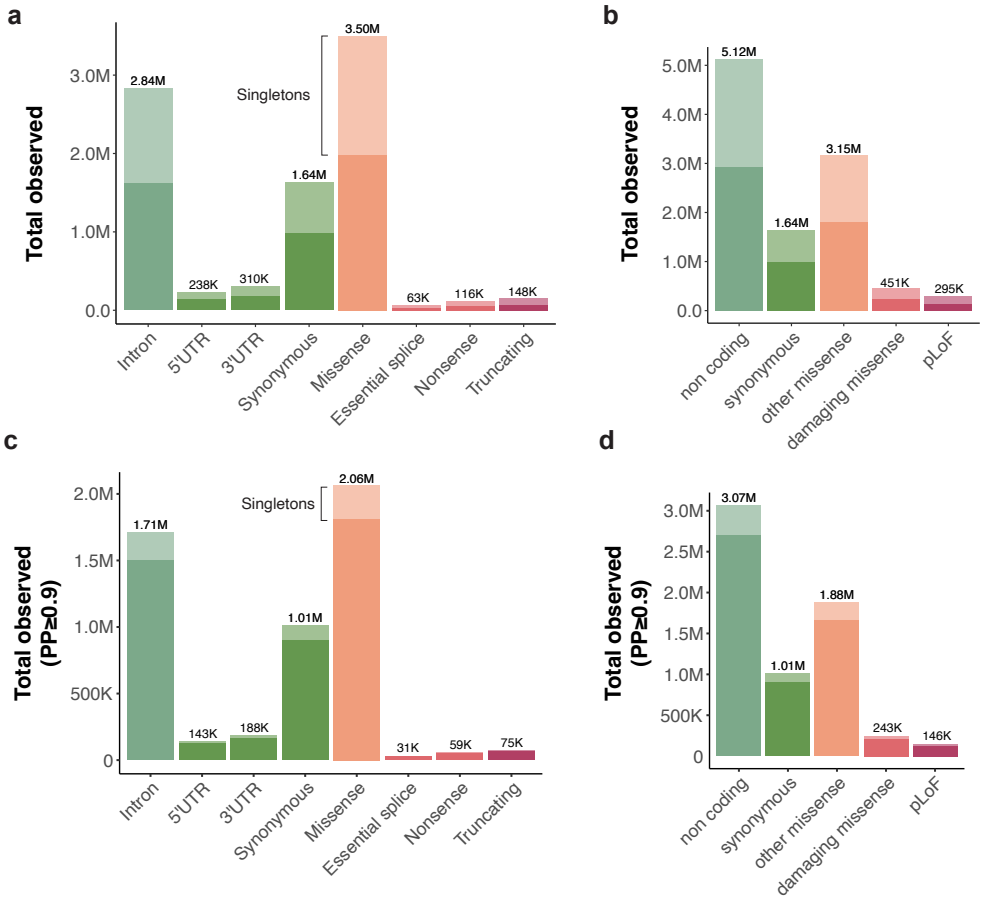


Figure 2.3: Distribution of variant annotation categories before and after broad consequence categorization and filtering by PP≥0.9. Variants are annotated using VEP by the most severe consequence in the canonical transcript. Panels (a) and (b) display the total number of unique variants across a set of consequences before and after broad variant consequence categorization. Panels (c) and (d) show the same distributions after restricting to variants with PP≥0.9. In all panels, green, orange, and red bars indicate low, medium, and high impact respectively, as per the legend. Singleton variants within each class are displayed in a lighter shade. Variant counts for each annotation category are shown above the bars.

in *cis* state on the same haplotype.

As I *a priori* expected that essential genes would be less permissible to biallelic damaging variants when compared to non-essential genes, I investigated tolerance towards predicted biallelic pLoF and pLoF+damaging missense/protein-altering variants across the genome. As some genes carry biallelic variants more often than others (owed to a variety of factors such as gene length and baseline mutation

frequency[232]), I fit counts of the number of individuals carrying biallelic variants per gene using a Poisson regression model accounting for variation in gene length and mutation rate (Methods, Section 2.2.13). Both pLoF and pLoF+damaging missense/protein-altering biallelic variants (homozygous and CH) were significantly depleted in five of the six analyzed essential gene-sets ($P < \frac{0.05}{6} \approx 0.0083$) (Figure 2.4). Conversely, across three non-essential gene-sets, biallelic pLoFs+damaging missense/protein-altering variants were enriched among LoF tolerant genes[227] ($P \leq \frac{0.05}{3} \approx 0.0167$). I found that the degree and direction of effects were consistent across CH, homozygous biallelic, and heterozygous variants (Figure 2.4) and concordant with previous study on CH depletion using the same gene-set definitions[198].

In founder[191] and bottle-necked[192] populations, some alleles are enriched to high frequency by chance, resulting in better powered association studies for the subset of variant alleles that are inherited from the parental population at higher frequency[188]. To explore the diversity of biallelic variation in UKBB, a largely outbred population, I enumerated two-hit carriers across 176,587 individuals. I observed complete biallelic 'knockout' of 1,174 unique genes strictly owed to pLoF variants, identifying 1,431 (0.8%) CH and 8,582 (4.8%) homozygous individuals with biallelic pLoF variants in at least one gene (Figure 2.1b). Across genes, 307 (26.1%) CH and 560 (47.7%) homozygous 'knockouts' were observed only in a single individual (Figure 2.1c-d). My simulations (Methods 2.2.19) closely recapitulated the empirically observed patterns of biallelic pLoF events, with strong correlation between simulated and observed homozygous ($R = 0.996$, $P < 2.2 \times 10^{-16}$) and CH ($R = 0.932$, $P < 2.2 \times 10^{-16}$) pLoF events (Figure A.5). I reasoned that inclusion of damaging missense/protein-altering variants in addition to pLoFs, would expand the number of identifiable damaging biallelic variants compared to assessing the two categories independently. Across 3,288 unique genes, I observed 11,491 (6.5%) CH and 17,863 (10.1%) homozygous carriers of pLoF+damaging missense/protein-altering variants. Of these, 1,112 (0.6%) CH and 436 (0.2%) homozygotes were carriers of biallelic pLoF+damaging missense/protein-altering variants in genes linked to traits with autosomal recessive mode of inheritance in Online Mendelian

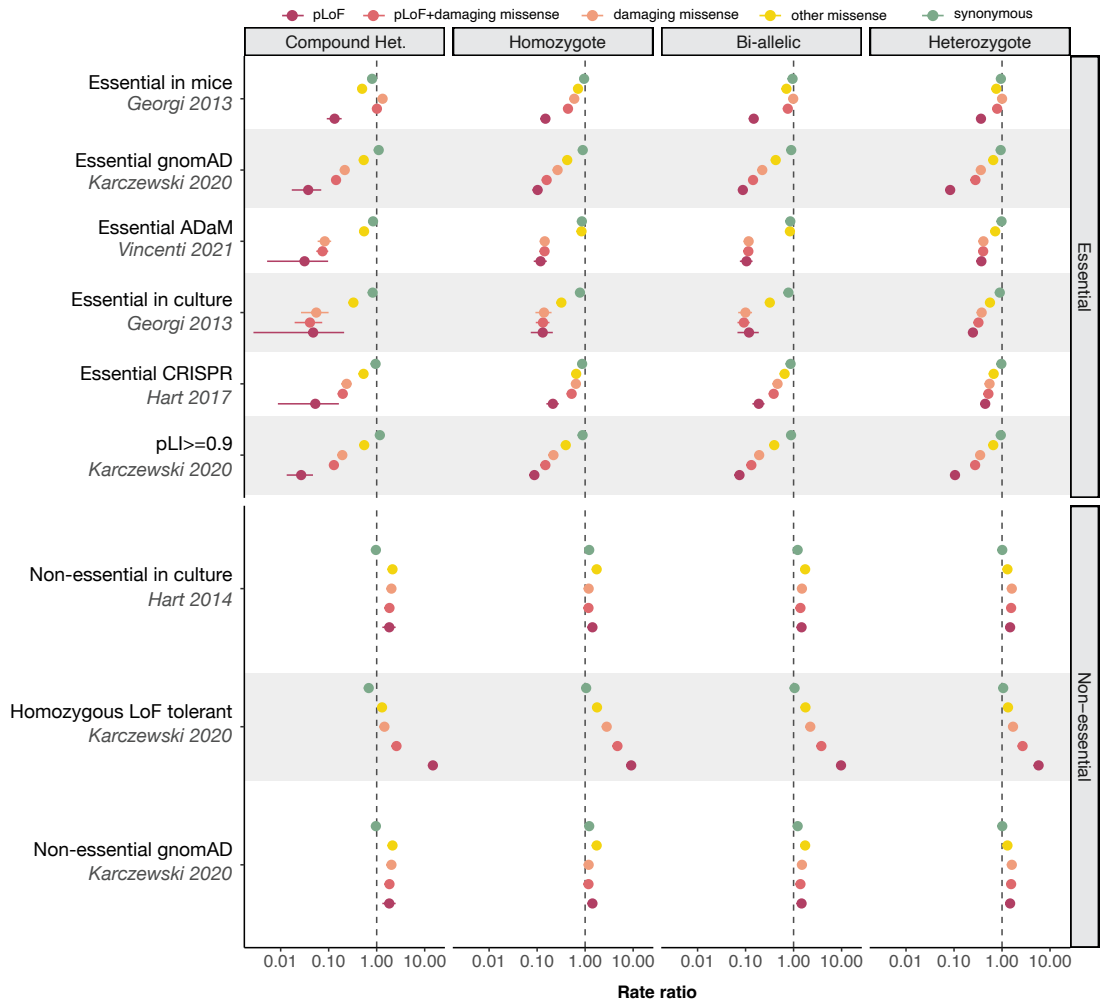


Figure 2.4: Gene-set depletion/enrichment modeling for biallelic variation. Poisson regression model of monoallelic (heterozygous) and biallelic variant (homozygous and compound heterozygous) depletion and enrichment across essential and non-essential gene-sets. Rate ratios are shown for synonymous (green), other missense (yellow), damaging missense (orange), and pLoF (red) variants. The dashed line indicates a rate ratio of 1.

Inheritance in Man (OMIM)[239]. Consistent with previous observations[142], I generally observed a higher prevalence of carriers with variants in *cis* compared to CH, with over a third of individuals (64,555, 36.6%) carrying ≥2 pLoF+damaging missense/protein-altering variants on a single haplotype (Figure 2.1b).

To better understand the evolutionary dynamics giving rise to pathogenic variants in *trans*, I examined the spectrum of allele frequencies of the constituent variants among the confidently called damaging CH variants. CHs variants tend

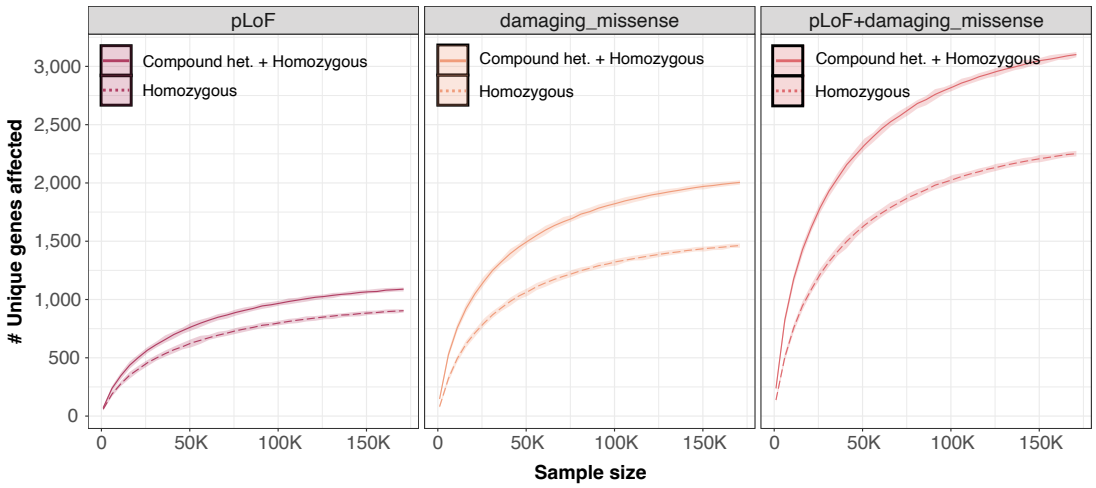


Figure 2.5: Count of unique genes affected by homozygous and compound heterozygous variants as a function of sample size. Starting with the full dataset and progressively down-sampling, I plot the count of unique genes harboring homozygous variants alone, and homozygous or compound heterozygous variants combined, as sample size decreases. Variant classes in each count are denoted according to the key. Each facet represents a specific variant annotation.

to comprise of two variants where one resides on a common haplotype, while the other on a rare haplotype, with a median difference in MAC of 1,181 (Figure A.9-A.10). Approximately 90% of CH-constituent variants have MAF ≤ 0.0038, compared to homozygotes in which 90% are detected at MAF ≥ 0.0022 (Figure 2.1e), suggesting that identifying deleterious biallelic CH variants requires reliable phasing of rare alleles (Figure A.11-A.12).

Multiple studies have assessed the prospects of ascertaining biallelic LoF variation at larger sample sizes in consanguineous, bottle-necked, and outbred populations[188,193]. To investigate empirically how the number of unique genes with biallelic variants scales in an outbred population, I performed down-sampling of UKBB participants. Consistent with previous literature, additional CH and homozygous variants can be inferred by considering both pLoF and damaging missense/protein-altering variation at even larger sample sizes (Figure 2.5).

2.3.3 Systematic evaluation of biallelic effects on common disease

I performed a series of association analyses using SAIGE[240], a generalized mixed model association testing framework that uses a Saddle Point Approximation (SPA) to provide accurate P-values for traits with extreme case-control ratio imbalance. This allowed us to investigate the effects of biallelic variants in 176,587 individuals across 311 phenotypes with varying population prevalence identified from secondary care EHRs (Methods 2.2.6). I restricted to 952 protein-coding genes with at least five individuals carrying biallelic variants in the same gene, which allowed us to detect OR ≥ 10, for traits at approximately 2% population prevalence, with 80% power at exome-wide significance (Bonferroni $P < \frac{0.05}{952 \times 311} \approx 1.68 \times 10^{-7}$) (Figure A.13). Using simulations, I confirmed my ability to detect recessive signals of association with well-calibrated false positive rates across a range of effect sizes (Methods 2.2.16, Figure A.14a-c). I tested a total of 299,854 gene-trait combinations, and identified seven gene-trait associations following stringent Bonferroni correction ($P < 1.68 \times 10^{-7}$) (Table A.10, Figure A.15). I observed that excluding singletons had minimal impact on the resulting gene-trait associations (Figure A.16), likely due to the limited number of CH variants comprised of at least one singleton (Figure A.17). Finally, I performed confirmatory replication analysis in the remainder of UKBB (Methods 2.2.18), and found that 7 of 7 (100%) gene-trait associations with at least five biallelic carriers replicated at $P < 0.05$ (Table A.11).

A recessive gene-trait association may be influenced by a variety of genetic factors unrelated to CH or homozygous status, such as polygenic background or through genetic tagging of a nearby common variant association. To mitigate these factors, I created a pipeline to condition on additional genetic effects within the gene-trait regression model (Methods 2.2.11). First, I trained PRS for 111 significantly heritable traits (h^2_{snp} $P < 0.05$ and $n_{\text{eff}} \geq 5000$) using LDPred2[241] (Methods 2.2.10), a tool that allows PRS derivation based on summary statistics and linkage information. To control for polygenic risk and potentially boost power for association[242], I included the off-chromosome PRS as an additional covariate

(Table A.10). I observed that controlling for PRS had no significant influence on the binary case-control association testing and the resulting P-values were altered by less than a single order of magnitude with the incorporation of PRS (Figure A.18). A possible explanation for this limited impact is that the PRS predictor was trained exclusively on the remainder of 246,152 individuals in UKBB that were not included in our biallelic variant analysis. While this separation was methodologically necessary to prevent overfitting, it may have compromised the predictive power of the PRS in the target population. The resulting predictor may have captured general polygenic effects but lacked sufficient explanatory power within the analysis cohort to meaningfully alter the association statistics. To capture the effects of any causal common variants in LD with the pLoF or damaging missense/protein-altering variants constituting the CH or homozygous variant, I further conditioned on nearby (within 1 Mb of the associated gene) common (MAF $> 1\%$) variant association signals (Table A.7), which abrogated ($P > 0.05$) the signal of a single gene-trait pair.

Lastly, I investigated whether any of the identified, putative recessive, associations could be accounted for by assuming an additive genetic architecture. To do this, I counted the number of gene copies affected by pLoF+damaging missense/protein altering variants in each individual. For each putative recessive association, I re-ran the analysis while simultaneously conditioning on the number of affected haplotypes. I also employed a complementary variant-level approach and repeated the analysis, conditioned on all low-frequency (MAC > 10, MAF $< 5\%$) and ultra-rare (MAC ≤ 10) damaging variants (pLoF+damaging missense/protein-altering), including those that constitute the biallelic variant in question. Among the remaining six Bonferroni significant associations none of the associations were abrogated after conditioning on additive effects (Table A.10).

Together, these analyses refined the list of putative gene-trait associations to six Bonferroni associations after stringent Bonferroni correction and conditioning (Figure 2.6a-2.6b, Table A.10) comprising three unique genes and six traits. Notably, only three of the six associations remained significant (Bonferroni $P < 1.68 \times 10^{-7}$) when restricted to only CH variant-carriers, and just one of six when testing

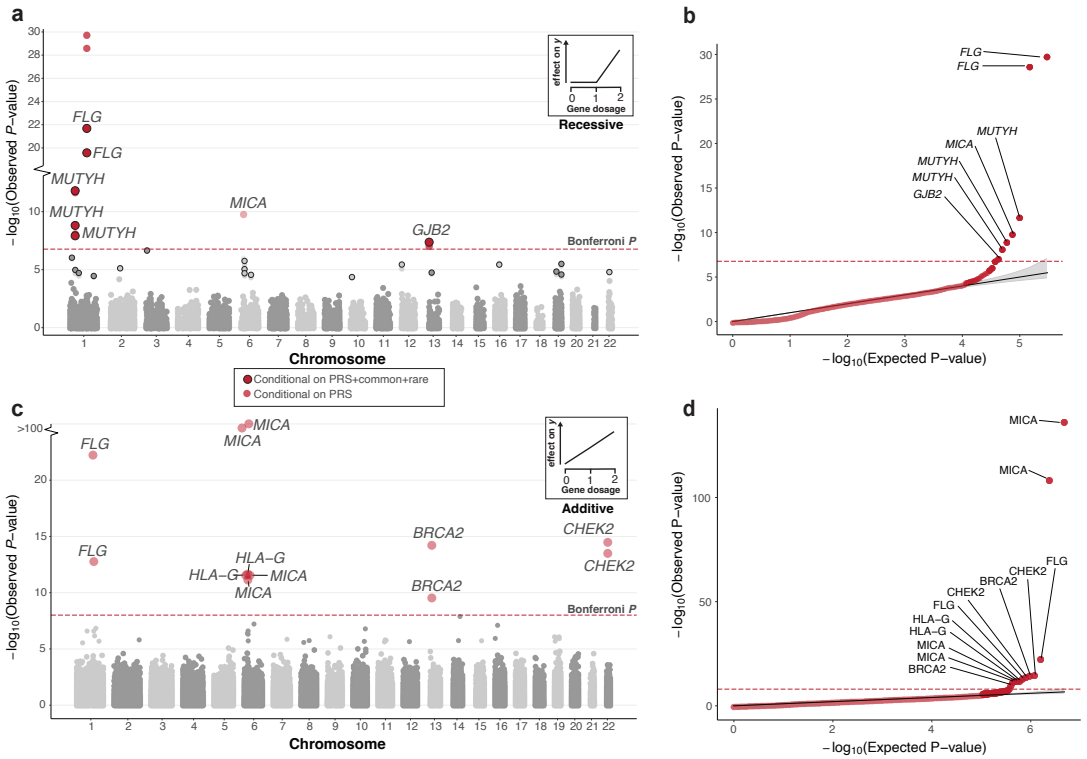


Figure 2.6: Conditional recessive and additive modeling of gene copy disruption in 311 phenotypes across 176,587 participants. a) Recessive Manhattan plot depicting $\log_{10}$-transformed gene-trait association P-values against chromosomal location. Associations are colored red if they are Bonferroni ($P < 1.68 \times 10^{-7}$) significant. Transparent coloring represents the resulting P-value when conditioning only on PRS, whereas solid coloring with black outline represents the P-value derived after conditioning on off-chromosome PRS, nearby (500 kb) common variant association signal, and rare variants within the gene when applicable (Methods 2.2.11). The Bonferroni significance thresholds is also displayed as a red dashed line. A gene may appear multiple times if it is associated with more than one phenotype. A qualifying example of the recessive inheritance pattern is shown in the top right: disruption of both gene copies result in an effect on the phenotype (y). b) QQ-plot for genes with bi-allelic damaging variants after conditioning on off chromosome PRS. The shaded area depicts the 95%CI under the null. Gene-trait associations passing Bonferroni significance are labeled accordingly. c-d) Additive Manhattan plot and corresponding QQ-plot for genes with mono and bi-allelic damaging variants associated with at least one phenotype after conditioning on off chromosome PRS when applicable. No additional conditioning was performed in this analysis. Gene-trait associations are colored red if they are Bonferroni ($P < 9.8 \times 10^{-9}$) significant. The additive inheritance model is depicted in the top right: each affected haplotype result in a incremental effect on the phenotype (y).

homozygous variants alone, underscoring the power of jointly analyzing these variant sets (Figure A.19).

I observed recessive gene-trait relationships across multiple physiological domains (respiratory, neoplasm, skin, ear, and mastoid). All six associations that met the significance threshold after Bonferroni correction and conditioning ($P < 1.68 \times 10^{-7}$) have previously been reported in the literature. For example, individuals with biallelic variants in *MUTYH*, a gene involved in oxidative DNA-damage repair[243], are at significantly increased risk of developing colorectal cancer ($\log_{10}(\text{OR}) = 4.7$ (95% CI $= 3.38 - 6.01$), $P = 2.2 \times 10^{-12}$). I also find that biallelic variants in *FLG* increase risk of both asthma[244] ($\log_{10}(\text{OR}) = 0.33$ $(0.26 - 0.39)$, $P = 2.09 \times 10^{-22}$) and dermatitis[245] ($\log_{10}(\text{OR}) = 0.28$ $(0.22 - 0.33)$, $P = 2.65 \times 10^{-20}$). In addition, I observe that biallelic variants in *GJB2* increase the risk of hearing loss[239] ($\log_{10}(\text{OR}) = 1.66$ $(1.05 - 2.26)$, $P = 9.93 \times 10^{-8}$).

2.3.4 Permutation testing establishes the role of genetic phase on disease risk

To assess the degree to which compound heterozygosity, rather than co-occurring variants on the same haplotype, drives disease risk, I permuted the genetic phase of observed pLoF+damaging missense/protein-altering variants within a gene to generate an empirical distribution of score statistics corresponding to disease-association strength in the absence of phase information (Figure 2.7a-b). To ensure a sufficiently large sampling distribution, I restricted the analysis to six nominally significant ($P < \frac{0.05}{952} \approx 5.25 \times 10^{-5}$) gene-trait combinations with at least ten individuals that are either CH variant-carriers or with two or more pLoF or damaging missense/protein-altering variants on the same haplotype (Methods 2.2.12). I found evidence that incorporation of CH variation significantly (Bonferroni $P = 0.05/6 = 0.0083$) increased the score statistic in three of the six analyzed gene-trait combinations: CH variants in *ATP2C2* are associated with increased risk of chronic obstructive pulmonary disease (COPD) ($P = 0.000238$), and CH variants in *FLG* are associated with increased risk of asthma ($P = 0.002174$), while

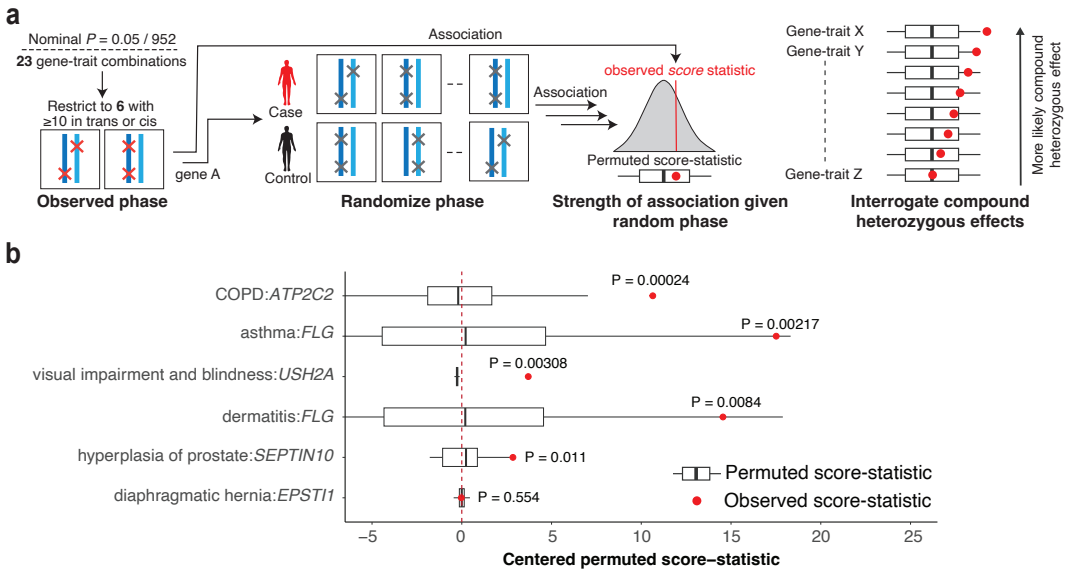


Figure 2.7: *In-silico* permutation of genetic phase provides evidence for CH-specific effects. **a)** Overview of the permutation pipeline. To be sufficiently powered to detect effects, I considered five significant ($P < 0.01$) gene-trait pairs from the genome-wide analysis that have at least ten individuals harboring pLoF or damaging missense/protein-altering variants on the same haplotypes or CH carriers. Then, I shuffled CH *trans* and *cis* labels across samples and re-ran the association analysis, resulting in a null distribution of permuted score-statistics corresponding to the association strength in the absence of phase information. I derive the one-tailed empirical P-value by comparing the observed score-statistics with the empirical null distribution. **b)** The resulting distributions of permuted (white and black box plots) and observed score-statistic (red dot) for each gene-trait and the resulting empirical P-value. P-values shown in bold indicate Bonferroni significance ($P < 0.05/06 = 0.0083$). Box and whisker plots display the quartiles of the empirical null distribution.

CH variants in *USH2A* are associated with increased risk of visual impairment and blindness ($P = 0.00307$) (Figure 2.7b). I identified two additional gene-trait association at nominal significance ($P < 0.05$): CH variants in *FLG* that are associated with increased risk of dermatitis ($P = 0.0092$), and CH variants in *SEPTIN10* which are associated with hyperplasia of prostate ($P = 0.011$). Of these, *FLG*-asthma, *FLG*-dermatitis, and *USH2A*-visual impairment associations have previously been linked to disease in the CH state[246–248]. These observations demonstrate, on a large scale, the effect of compound heterozygosity in driving disease susceptibility, and by extension, how appropriately integrating genetic phase can lead to increased power to discover gene-trait associations.

2.3.5 Incorporation of genetic phase boosts power in gene-level rare variant burden analysis

Complementary to the recessive models described above, rare variant burden testing, which involves the aggregation of rare variants within a gene, has proven to be a robust method to collectively assess the phenotypic impact of rare variation across individuals. Rare variants are aggregated due to their low allele frequency leading to lack of statistical power for detection of single-marker associations. However, these frameworks generally ignore the genetic phase within each individual, and therefore do not differentiate between scenarios in which multiple damaging variants reside on the same (*in cis*) or opposite (*in trans*) haplotypes, despite these two forms having potentially distinct functional and phenotypic effects. I conducted additive genome-wide association analyses by testing for associations between the the number of disrupted gene copies (across 16,363 protein-coding genes with at least ten haplotypes carrying pLoF+damaging missense/protein altering variation) in an individual and case status (across 311 phenotypes) (Methods 2.2.9, Figure 2.6c-2.6d). After adjusting for polygenic contribution, I found 12 significant associations after stringent multiple-testing correction (Bonferroni $P < \frac{0.05}{16{,}363 \times 311} \approx 9.8 \times 10^{-9}$, Table A.10). Among the significant hits are previously reported associations, including association between the number of putatively damaged copies of *BRCA2* ($P = 6.16 \times 10^{-15}$) and *CHEK2* ($P = 3.34 \times 10^{-15}$), and breast cancer. I then compared my approach to an established gene-based method for rare variant burden testing, specifically `SAIGE-GENE+`[86]. I applied `SAIGE-GENE+` to the same variant set used in my study, comparing the P-values derived from this method with those from the haplotype-collapsing approach (Table A.8). This initial comparison revealed 5 of 12 (42%) associations with insufficient evidence to reach Bonferroni significance ($P < 9.8 \times 10^{-9}$) using `SAIGE-GENE+` burden testing. Moreover, among the significant additive haplotype-based gene associations, 11 of 12 (92%) were more strongly associated than in `SAIGE-GENE+` burden testing (one-sided sign test $P = 0.00317$). Recognizing that my initial analysis deviated from the regular usage of `SAIGE-GENE+`, which ignores phasing confidence, I also performed an

analysis without filtering variants by PP. This allowed us to increase the pool of variants and compare the performance of the haplotype collapsing method with the 'standard application' of the burden test. I observed that while `SAIGE-GENE+` now recapitulated all twelve associations at the Bonferroni cutoff ($P < 9.8 \times 10^{-9}$), I found that 12 of 12 (100%) gene-trait associations were more strongly associated using the haplotype-based collapsing method (one-sided sign test $P = 0.00024$).

2.3.6 CH variants elevate lifetime risk of disease

Biallelic effects may be associated with earlier age at onset of disease, which is also often correlated with disease severity. I therefore explored whether CH and homozygous variants had longitudinal effects by evaluating age-at-diagnosis of 278 phenotypes with Cox proportional-hazards models. To identify effects owed to disruption of both gene copies, as opposed to haploinsufficiency, I compared biallelic variant carriers against a reference group comprising carriers of a single heterozygous variant for each gene. I tested 267,400 gene-trait combinations with at least five biallelic variants (homozygotes or CH) and 100 heterozygotes (Figure 2.8a). After adjustment for polygenic risk via off-chromosome PRS, I identified seven gene-trait associations with significantly earlier age-at-diagnosis in biallelic variants compared to heterozygous carriers of pLoF+damaging missense/protein-altering variants (Bonferroni $P < \frac{0.05}{952 \times 278} \approx 1.89 \times 10^{-7}$, Figure 2.8b-c, Figure A.20, Figure A.21-A.22). For six out of the seven Bonferroni significant gene-trait combinations, I found no evidence ($P > 0.05/7 \approx 0.00833$) that carrying a single heterozygous variant altered lifetime disease risk compared to carrying two copies of the reference allele. I also performed confirmatory replication analyses for seven associations with at least five biallelic and 100 heterozygous carriers in the replication cohort (Methods 2.2.18). I found that all seven (100%) Bonferroni significant gene-traits associations replicated at $P < 0.05$, including *ATP2C2*-COPD (Replication $P = 0.013$) which has not previously been reported in the literature (Table A.12).

I further sought to disentangle the effects of homozygous and CH variants on lifetime disease risk from that attributable to multiple damaging rare variant effects

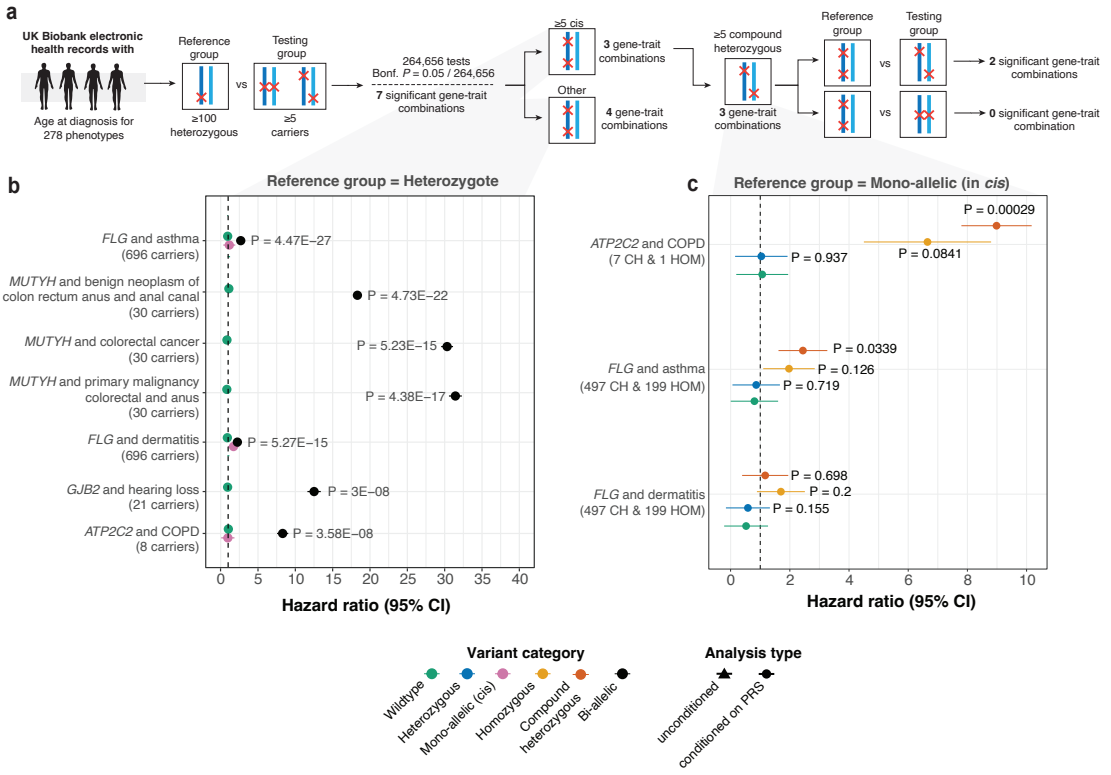


Figure 2.8: Age-at-diagnosis modeling reveals novel recessive effects driven by damaging bi-allelic variants. **a)** Flow diagram of the approach. To investigate whether homozygous and/or CH effects are associated with a difference in lifetime risk of disease development, I perform Cox proportional-hazards modeling for gene-trait combinations in which ≥ 5 samples are two-hit carriers (CH or homozygotes) and ≥ 100 samples that are heterozygotes. Among Bonferroni significant associations ($P < 1.89 \times 10^{-7}$), I filter to gene-trait pairs for which at least five samples carry multiple variants disrupting the same haplotype, and test for an association between CH or homozygous carrier status and lifetime disease risk (corresponding to Hazard ratios (HRs)>1). **b)** HRs when comparing CH and homozygous status versus heterozygous carrier status. Throughout, I display hazard ratios and corresponding P-values after taking the polygenic contribution into account by conditioning on off-chromosome PRSs for heritable traits that pass the quality control cutoffs. P-values following inclusion of polygenic contribution to disease status are provided where PRS are predictive. HRs for gene-traits with two or more individuals with multiple *cis* variants on the same haplotype are displayed in pink. Only associations that pass the stringent Bonferroni significance threshold ($P < 1.89 \times 10^{-7}$) cutoff are illustrated. **c**) HRs when comparing wildtype, heterozygous, CH and homozygous carrier status against individuals that harbor ≥ 2 putatively damaging variants on the same haplotype. 95% CIs are shown in the figure. Number of biallelic carriers is shown for figure (b) and stratified by homozygotes (hom) and compound heterozygotes (CH) in (c). Abbreviations: CC (colorectal cancer), COPD (chronic obstructive pulmonary disease).

on a single haplotype. To do this, I analyzed such effects in the three gene-trait pairs with both (1) at least five CH and/or homozygous variants and (2) at least five individuals harboring ≥2 variants on the same haplotype (Figure 2.8c). Compared to individuals with a single disrupted haplotype, both homozygous and CH carriers of pLoF+damaging missense/protein-altering variants in *ATP2C2* were at increased lifetime risk of developing COPD (homozygote HR = 6.65 (95% CI = 4.5˘8.8); $P = 0.084$, CH HR = 8.98 (7.79˘10.17); $P = 0.00028$). Similarly, both homozygous and CH variants of *FLG* were at increased lifetime risk of asthma (homozygote HR = 1.97 (1.1˘2.84); $P = 0.126$, CH HR = 2.44 (1.61 − 3.26); $P = 0.033$) and dermatitis (homozygote HR = 1.7 (0.88˘2.5); $P = 0.20$, CH HR = 1.16 (0.38˘1.94); $P = 0.7$) (Figure 2.8c). For these gene-trait relationships mono-allelic carriers have no increase in risk of disease compared to wildtypes, whereas biallelic carriers have a significant increase in lifetime risk of disease.

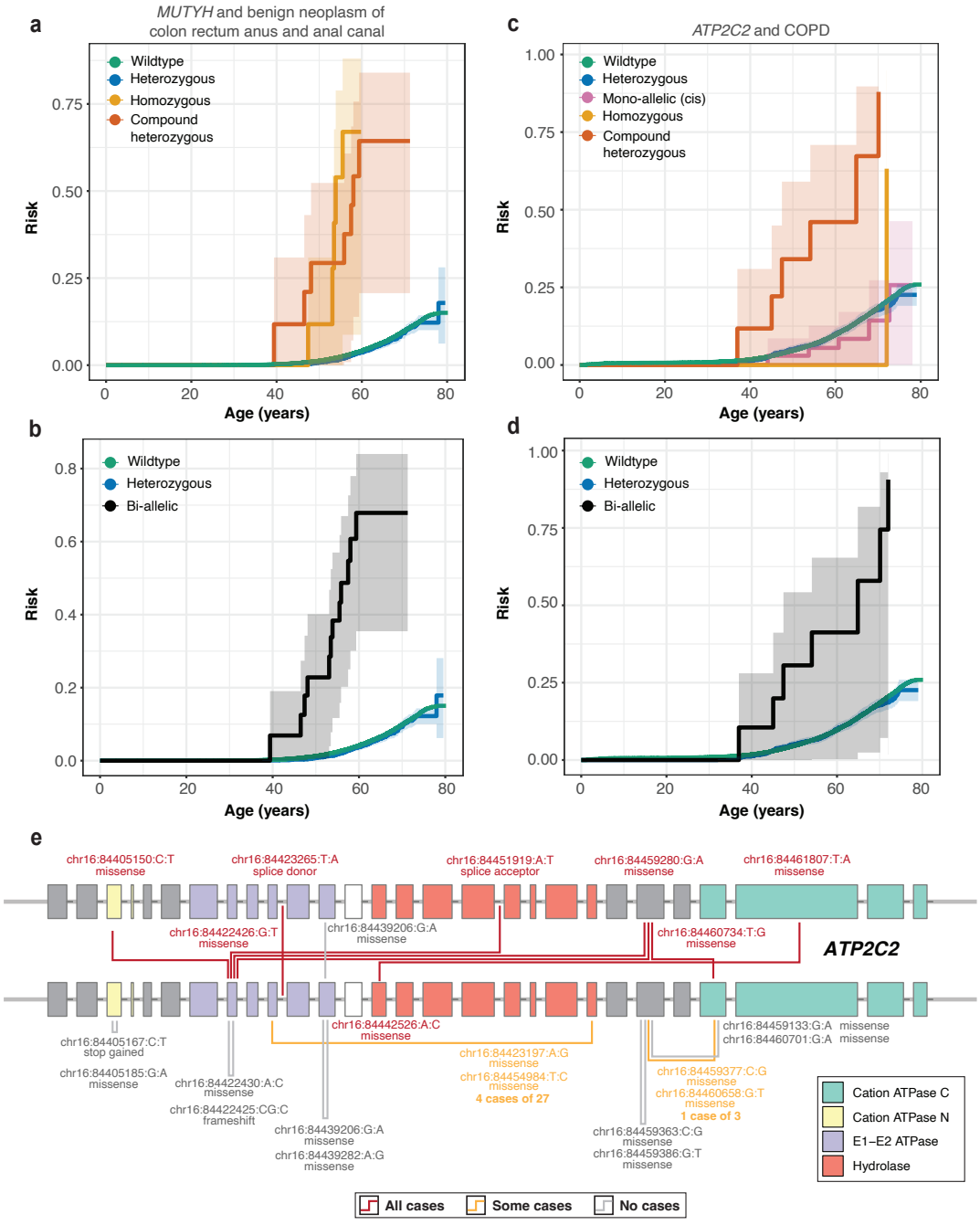

a
MUTYH and benign neoplasm of colon rectum anus and anal canal
Wildtype
Heterozygous
Homozygous
Compound heterozygous
Risk
Age (years)
b
Wildtype
Heterozygous
Bi-allelic
Risk
Age (years)
c
ATP2C2 and COPD
Wildtype
Heterozygous
Mono-allelic (cis)
Homozygous
Compound heterozygous
Risk
Age (years)
d
Wildtype
Heterozygous
Bi-allelic
Risk
Age (years)
e
chr16:84405150:C:T missense
chr16:84423265:T:A splice donor
chr16:84451919:A:T splice acceptor
chr16:84459280:G:A missense
chr16:84461807:T:A missense
chr16:84422426:G:T missense
chr16:84439206:G:A missense
chr16:84460734:T:G missense
ATP2C2
chr16:84442526:A:C missense
chr16:84405167:C:T stop gained
chr16:84405185:G:A missense
chr16:84422430:A:C missense
chr16:84422425:CG:C frameshift
chr16:84439206:G:A missense
chr16:84439282:A:G missense
chr16:84423197:A:G missense
chr16:84454984:T:C missense
4 cases of 27
chr16:84459133:G:A missense
chr16:84460701:G:A missense
chr16:84459377:C:G missense
chr16:84460658:G:T missense
1 case of 3
chr16:84459363:C:G missense
chr16:84459386:G:T missense
Cation ATPase C
Cation ATPase N
E1–E2 ATPase
Hydrolase
All cases
Some cases
No cases

Figure 2.8 *(previous page)*: Trajectories of haplotype disruption in common disease. **a-b**) Kaplan-Meier survival curves for CH (red), homozygous (orange), heterozygous carriers (blue), single disruption of haplotypes (pink) owed to pLoF or damaging missense/protein-altering mutations. Wildtypes and bi-allelic variants (CH or homozygous) are shown with green and black lines respectively. Both CH and homozygous *MUTYH*-variant carriers are at elevated lifetime risk of developing benign neoplasm of the colon compared to heterozygous carriers and wildtypes. **c-d**) Kaplan-Meier survival curves for *ATP2C2* mono and bi-allelic variant carriers. Carriers of CH variants develop COPD more early compared to heterozygotes carriers and wildtypes. Moreover, individuals who harbor a single putatively disrupted haplotype owed to $\geq$2 damaging variants develop COPD at the same frequency as heterozygotes and wildtypes. **e)** Gene plots for *ATP2C2*, displaying protein coding variants for samples that carry $\geq$ 2 pLoF or damaging missense/protein-altering variants stratified by exon or intron. CH variants, multiple variants in *cis*, and homozygous variants are highlighted by lines joining the positions of co-occurring variants in a sample. Lines are colored by number of cases for the shown variant configurations, with gray lines indicating no observed samples are cases; orange lines indicating some some samples are cases; red lines indicate that all observed samples are cases. Variants are labeled by position (GRCh38) and according to inferred consequence (missense, stop gain, splice acceptor/donor). Protein domains are highlighted accordingly[249].

2.3.7 Common complex disorders implicated by CH variation

Six of the seven gene-trait combinations for which I identify Bonferroni significant associations with lifetime disease risk are also significant in my cross-sectional recessive association analysis. The six have previously been reported in the literature, albeit without age-at-onset effects. These include *MUTYH* and colorectal cancer, *GJB2* and hearing loss, and a pleiotropic association of *FLG* with both dermatitis and asthma (Figure 2.8a-b, Figure A.21). I find that *ATP2C2*-COPD is a novel candidate association ($P = 3.58 \times 10^{-8}$) with plausible mechanistic effects.

ATP2C2, a calcium-transporting ATPase linked to surfactant protein D levels via an intronic variant (rs9927461)[250] (a causal risk factor for COPD), is associated with COPD in the gene-trait analyses (HR = 8.3 (95% CI = 7.54˜9.05); $P = 3.56 \times 10^{-8}$). As I did not observe any nearby (1 Mb upstream or downstream) common variants in *ATP2C2* associated with cross-sectional COPD (all $P > 5 \times 10^{-6}$), the association between biallelic variants of *ATP2C2* and COPD is potentially driven by the unique configurations of CH damaging-missense ($n = 7$) and pLoF ($n = 1$) variants that

primarily reside in functional protein domains (Figure 2.8e, Figure A.23). Seven of the eight (87.5%) individuals I identified as biallelic carriers of damaging variation in *ATP2C2* (6 CH and 2 homozygous) were diagnosed with COPD (median age of diagnosis = 54.1 (IQR = 46.2 - 67.5) years) (Figure 2.8c-d). In contrast, only 6.9% of individuals harboring multiple pLoF+damaging missense/protein-altering variants on the same *ATP2C2* haplotype were diagnosed with COPD, with the same median age of diagnosis (60.8 (53.7 - 67.9) years) as both heterozygote carriers (58.0 (48.5 - 64.1) years) and those with wildtype (59.2 (51.3 - 65.1) years).

FLG plays a pivotal role in the differentiation and maintenance of skin barriers[244]. *FLG* variants have been selectively associated with individuals with both asthma and atopic dermatitis, but not with those who have asthma without atopic dermatitis[245]. My findings indicate that individuals carrying a single deleterious *FLG* allele face increased risk of dermatitis (HR = 1.11 (95% CI = $1.05 - 1.19$); $P \approx 7.2 \times 10^{-5}$), but not asthma (HR = 1.06 (95% CI = $1.01 - 0.12$); $P = 0.018$), when compared to wildtypes. In contrast, individuals carrying two variant alleles have an increased risk of developing both dermatitis (HR = 2.23 (95% CI = $2.03 - 2.43$); $P = 5.27 \times 10^{-15}$) and asthma (HR = 2.68 (95% CI = $2.51 - 2.87$); $P = 4.47 \times 10^{-27}$), suggesting a recessive mode of inheritance for *FLG*-related asthma and a semi-dominant inheritance mode for *FLG*-related dermatitis[239]. This implies that the loss of a single *FLG* copy can result in dermatitis, while the loss of both copies can lead to asthma. Together, this may help clarify why *FLG*-related asthma is seldom observed without the presence of *FLG*-related dermatitis.

2.4 Discussion

In this chapter, I systematically interrogate the role of biallelic coding variants in genes conferring risk for common complex diseases. In the cross-sectional and longitudinal recessive analysis I identify a six and seven significant gene-trait associations after stringent Bonferroni correction, respectively. In the cross-sectional analysis only two of the six (33%) associations I identified would have been discovered considering homozygotes alone, whereas in the longitudinal analysis five of the seven (71%) associations would have been discovered without considering compound heterozygosity. Collectively, the associations both replicate established relationships in the literature, and identify a previously unreported gene-trait associations for binary phenotypes across the common disease spectrum.

I show that the 90% of deleterious CH variants occur at MAF $< 0.38\%$. Given that phasing quality is directly correlated with allele frequency, it is essential to filter to the set of variants phased at high confidence to eliminate false positive identifications. Here, I quantified the increase in phasing quality using Mendelian inheritance logic in parent-offspring relationships and compared pairs of statistically phased variants to read-backed phased variants using short read sequences. While read-backed phasing is computationally expensive and restricted to variants in close proximity, I demonstrate that it can be employed to evaluate statistical phasing quality in cohorts that lack trio relationships, with error rates comparable to that of trio switch error rates.

CH disease associations have mainly been explored in rare disorders[157,194–197], but are seldom investigated in the study of common disease. This is due to the low prevalence of variants in the CH state and the genetic architecture of common complex traits, which are typically influenced by environmental factors and numerous loci with low to modest contribution to risk. In this study, I address these challenges and offer multiple lines of evidence to demonstrate the role of CH effects in driving disease risk for common traits. I employed two complementary analyses to detect gene-trait associations: a genome-wide logistic association analysis and a time-to-event model. Through these methods, I identified associations in which variants in

the homozygous or CH state resulted in increased disease risk compared to wildtypes and individuals carrying multiple pathogenic variants on the same haplotype. My findings show that for certain gene-trait pairs, individuals with a single disrupted gene copy have a risk of developing disease that is virtually indistinguishable from that of wildtypes, suggesting non-additive gene dosage effects. Further, by permuting the genetic phase, I find evidence that incorporation of confidently phased CH variants can boost power to detect associations in common disease. Collectively, my results emphasize the importance of considering each individual's specific genetic context when assessing their genetic risk in a clinical setting. Simply identifying the presence of multiple pathogenic variants in a gene, disregarding the phase, may not be sufficient to fully understand an individual's risk profile.

Many common complex traits have polygenic architectures, which should be accounted for when performing gene-trait association testing. The presence of biallelic variants in individuals with such diseases might be coincidental and not causally related to the trait, which may instead be a result of a high polygenic risk. However, across the significant recessive genome-wide associations, I observed that inclusion of PRS as a covariate, affected the resulting association P-value by less than an order of magnitude for the binary traits I analyzed. While I was only able to account for the polygenic contribution to disease development for 111 diseases with significant common variant heritability in the UKBB, due to low case numbers, these observations suggest that incorporation of polygenic background has limited influence on the degree of association when evaluating ultra-rare variation across binary traits.

I found that the majority of biallelic gene-disease associations are driven by variant combinations containing at least one missense variant, which would have been excluded under a stricter high-confidence pLoF criterion. Although the less stringent inclusion threshold enabled us to identify a greater number of biallelic variants, it is likely that some damaging missense or protein-altering variants would incorrectly be predicted as damaging, or may exhibit gain-of-function rather than loss-of-function effects, consequently reducing the signal-to-noise ratio in my

analyses. Even 'knockouts' by *bona fide* pLoF variants may only result in partial gene inactivation, and not necessarily complete gene knockdown. Additionally, pLoF variants may be 'rescued' and not lead to complete or even partial loss-of-function. While I show that including damaging missense/protein-altering variants to define biallelic variants can improve power for certain phenotypic associations, further manual curation and experimental validation will be required to demonstrate that these variants truly result in loss-of-function.

The likelihood of damaging alleles occurring on the same haplotype is influenced by a complex interplay of factors, including population structure and balance between selection, drift, mutation, and recombination. I and others[142] find that damaging CH variants occur less frequently than multiple damaging variants affecting the same haplotype, suggesting that in certain circumstances, natural selection operates on a haplotype level. Once a LoF variant occurs and expands in the population, the affected haplotype has no selection against additional acquisition of damaging mutations. This has implications for association studies investigating CH effects by counting the number of damaging variants in a gene while attributing equal probability to each of affecting each haplotype[190], as such frameworks may overestimate the frequency of CH events.

'Human knockouts' have been extensively discussed in the context of therapeutic development. Examining both biallelic and monoallelic carriers can help assess the safety of therapeutic interventions by analyzing how varying degrees of target modulation affect biological response[185,188]. I showcase several gene-trait relationships where the number of affected haplotypes influences the lifetime risk of disease, potentially representing the manifestation of 'adverse events' which are important endpoint in clinical trials. The absence of adverse events in monoallelic carriers can potentially imply that partial pharmacological inhibition of a target may be a safe and effective approach. However, adverse effects observed in biallelic carriers of damaging variation within the same locus could indicate potential risks associated with complete target inhibition. A natural extension of this work could involve investigating mono and biallelic effects on quantitative outcomes,

such as serum proteins. Changes in biomarkers (or other continuous outcomes) may reflect direct or indirect consequences of gene modulation and could serve as potential pharmacodynamic biomarkers commonly used to assess target engagement in clinical trials.

This work showcases the value of statistical phasing of damaging rare variants, and that association analyses that account for compound heterozygosity can be better-powered for gene-trait discovery. I show that this approach can be employed to discover well-established and novel non-additive and additive gene-trait relationships across a wide range of disease etiologies. From a clinical perspective, I demonstrate the importance of interrogating the genetic phase when dealing with CH variants in traits with recessive mode of inheritance. This is an important step towards uncovering the phenome-wide consequences of biallelic disruption across the human genome.

2.4.1 Limitations

I acknowledge several limitations to this study. Firstly, the accuracy of statistical phasing decreases with the MAF. The study encompasses ultra-rare variation (MAF $< 0.001\%$) which I filtered to confidently phased heterozygous variants (PP $>$ 0.9). This criteria excludes variants that could form CH relationships, potentially leading to an underestimation of compound heterozygous events, especially among those that constitute rare variants and singletons. Secondly, I applied a stringent threshold for statistical significance and replicated all associations. However, caution should be exercised when interpreting findings that are significant below the Bonferroni significance threshold, especially those involving a low frequency of biallelic variants and low case counts. These conditions can result in instability and lower power in the mixed model framework, particularly for rare binary traits within[251]. Thirdly, this work is based on coding mutations within protein-coding regions. Including splicing variants and non-coding variants, such as those in theUTR, could enhance the power of the analyses and potentially lead to further associations.

Lastly, my study was conducted within the UKBB using individuals of European ancestry. It is possible that the number of CH events and their prevalence across the genome may vary across non-European populations. Other population structures could affect the power for association. For example, in populations such as the Finns, where bottlenecks have caused rare variants from founders to rise to appreciable frequencies, there could be increased detection power for those alleles[192]. Similarly, in consanguineous populations, the prevalence of homozygous pLoF genotypes increases due to parental relatedness, could increase power for detection. This is particularly relevant for members of the East London Genes and Health cohort[252], or other cohorts with participants reporting high degree of consanguinity. It will be critical to apply similar approaches in other cohorts encompassing diverse populations to further explore this diversity.

3

Rare variant analysis of dominance

Contents

3.1 Introduction 68
3.2 Methods 71
3.2.1 QC of the 450k exome release 71
3.2.2 Additional QC of genotyping array data 74
3.2.3 Statistical phasing 74
3.2.4 Variant annotation 74
3.2.5 Encoding nonadditivity at a locus 76
3.2.6 Estimating genetic architectures 78
3.2.7 Association analysis of alternate genotype encodings . . 79
3.2.8 Simulation of counts of observed biallelic variants 80
3.2.9 Simulation of null trait and analysis 82
3.2.10 Phenotype curation 84
3.2.11 Analysis of proteomic measurements 85
3.2.12 Modelling recessive selection 86
3.3 Results 91
3.3.1 High-quality phased variants enables biallelic analysis . 91
3.3.2 Biallelic effects across the plasma proteome 92
3.3.3 Allelic recoding yields nonadditive proteomic effects . . 97
3.3.4 Analysis of recessive effects across health and disease . . 107
3.3.5 Simulation of random phenotypes 107
3.3.6 Exome-wide scans of nonadditive effects 111
3.3.7 Deciphering mode of inheritance in binary traits 114
3.3.8 Characterizing the spectrum of recessiveness 119
3.3.9 The burden of nonadditive effects 121
3.3.10 Identifying recessive selection 124
3.4 Discussion 127
3.4.1 Limitations 133

3.1 Introduction

Decomposing trait heritability into additive, dominance and epistatic contributions was first proposed by R.A. Fisher[19]. This framework has since become fundamental in quantitative genetics. The additive contribution is defined as the effects attributed to the average of allelic substitution, while the dominance contribution represents the additional effects not explained by the additive component for pairs of alleles within a single locus or site[179]. Throughout this thesis, the terms 'nonadditive' and 'dominance' are used interchangeably.

As discussed in detail in chapter 1, there is a substantial amount of evidence for nonadditive effects in nature, including examples of the peppered moth during the industrial revolution[162], and the selective expression of the flagella of Chlamydomonas in its haploid stage[168]. Mendelian traits also provide clear instances of nonadditive effects[253]. In completely penetrant Mendelian recessive traits, the variance can be exclusively explained by the nonadditive contribution, as two mutated alleles are required for phenotypic expression. Similarly, for Mendelian dominant traits, a single gene copy is sufficient for phenotypic expression, with no additional effect from a second allele. However, most traits are not Mendelian, and can be described by a combination of additive and nonadditive contributions to genetic variance, resulting in intermediary phenotypes, in which the biallelic effect is either larger or smaller than the heterozygous effect[192,254].

Nonadditive effects are rarely the priority when performing association analysis in large-scale population studies. This omission is primarily due to the effectiveness of the additive model in capturing the majority of trait variation, even when the underlying genetic architecture is truly recessive or dominant[177] (See Figure 1.3 in Chapter 1). In common complex traits, the nonadditive contribution tagged by common variation accounts for as little as 0.5 to 0.8% of the additive

heritability[179,255,256]. For common variants analysis, the additive model has therefore been the default choice for both variant- and gene-based association analyses.

For analysis of rare variants that show recessive effects, additive models may not be sufficiently powered[257], motivating the use of explicit recessive models for genetic discovery. The simplest form of this analysis is to map the additive genotype encoding $[0, 1, 2] \rightarrow [0, 0, 1]$, referred to here as the canonical recessive encoding, and associate it with a trait of interest. However, the additive and recessive genotype encodings are correlated. As a result, a significant fit under the recessive model does not preclude a significant fit under the additive model for the same data, limiting the ability to definitively distinguish between these genetic architectures. To address this limitation, Heyne *et al.* proposed a method comparing P-values from separate additive $[0, 1, 2]$ and canonical recessive $[0, 0, 1]$ analyses applied to the same data[192]. In their framework, an association is classified as 'recessive' if the recessive P-value ($[0, 0, 1]$) is more than two orders of magnitude more significant than the additive P-value ($[0, 1, 2]$). An alternative approach involves testing the 'dominance deviation'. This method models the genotype as $[0, 1, 0]$ or $[0, 0, 1]$ and incorporates it into a joint 2-degree-of-freedom (DF) conditional model with the additive encoding $[0, 1, 2]$. It then assesses the significance of the effect size estimate for the recessive or heterozygous contribution in this joint model[53,258–260]. However, this approach does not fully resolve the issue, as the correlation between the encodings remains, resulting in persistent additive spillover effects[179].

In this chapter, I propose an alternative strategy for isolating the nonadditive contribution with applications in the rare variant space, which has not yet been feasible in the statistical genetics community. This involves an explicit allelic recoding of genotypes that imposes orthogonality between the additive and nonadditive heritability. For a locus, this allows for parallel analysis of the additive or nonadditive contribution to trait variation without spillover effects[176,179]. Recent work has highlighted similar methods[179,255], but their application has been restricted to common (MAF$>1\%$) variants, simply because their encoding is parameterised by the allele frequencies, and therefore rely on HWE, which rare variants tend

to violate. This is because, rare variants are often subject to strong selection pressures, which actively alter their frequencies in the population. Additionally, rare variants can be the result of recent mutation events, that have not sufficient time to reach equilibrium in the population. Furthermore, rare variants often violate HWE due to sampling. The observed genotype counts for rare variants are unlikely to match the expectation under HWE, simply because people come in whole numbers, not fractions[180–182]. The allelic recoding method proposed in this chapter enables exploration of nonadditive contributions across the allele frequency spectrum, including rare variants, which are expected to exhibit the largest nonadditive effects. The reasoning behind this expectation is twofold: firstly, selection tends to maintain variants with large effects at low frequencies in the population, and secondly, deleterious alleles tend to be recessive[171]. I demonstrate that this framework can also be used to infer genetic architectures and can be extended to standard rare variant burden testing, broadening its applicability from GWAS.

Throughout this chapter, I explore the additive and nonadditive contribution of two different datasets. The recently released Olink proteomics[90], and a large range of quantitative traits and disease endpoints from UKBB. My aim is to unpick nonadditivity across distinct data modalities that range from molecular interactions to broader physiological outcomes. This work is enabled by large-scale biobank phasing, and thus builds on the lessons from chapter 2. Together, my analysis demonstrates the spectrum of nonadditive recessive effects attributable to rare genetic variants across a range of traits.

3.2 Methods

Science and data availability are evolving at an ever-increasing pace: At the start of my DPhil, only the 200k ES tranche was available. The release of the 450k ES dataset allowed us to expand our analysis to additional samples. We performed QC on this new data set using the methods already described in chapter 2 with some modifications outlined below.

3.2.1 QC of the 450k exome release

The following QC was carried out by Barney Hill, Duncan Palmer and myself.

Summary of 450k exome sequencing quality control

For the 450k UKBB ES data, we performed a series of initial hard filters on the variant level, on the sample level, and on the genotype level. As previously, we confirmed genetic sex with self-reported sex, and restricted our analysis to individuals that are confidently classified into broad ancestry groups (EUR, East Asian (EAS), Southeast Asian (SAS) and African (AFR)). Following this, we apply a second round of final hard filtering based on the resultant samples and variants. We used `Hail 0.2`[203] and `PLINK 1.9`[261] coupled with bash, C, python, and R scripts, to carry out all QC.

Initial hard genotype-level filters for exome sequencing

Multi-allelic variants were split into biallelic variants using `split_multiallelic` with `hail`, and indel alignments were adjusted to be left-aligned[207]. The following criteria were used to designate genotype calls as missing:

1. GQ $\leq$ 20
2. Total Sequencing Depth (DP) $\leq$ 10
3. Phred-scaled likelihood (PL) of the genotypes $<$ 20
4. For heterozygous calls:

(a) $(AD_{ref} + AD_{alt})/DP < 0.8$

(b) $(AD_{alt}/DP) < 0.2$

5. Homozygous indel calls: $(AD_{alt}/DP) < 0.8$

where AD_{ref} and AD_{alt} refer to the allele depth supporting the reference and the alternate allele respectively.

Initial variant level filers

After genotype-level filtering we performed an initial variant level filter based on the following criteria. Specifically, we discarded any variants that were:

1. not in the official exome capture target regions (See UKBB field 3803) padded with 50bp on either side.

2. located within a LCR[205].

3. invariant, i.e. MAC equal to 0 after genotype quality control.

A total of 18,992,659 variants were left after the initial genotype-level filter and variant level filter.

Imputing sex

We performed sex imputation using the strategy described in Chapter 2, Section 2.2.2. Imputed sex inconsistent with reported sex could indicate sample swap. Briefly, individual's sex was imputed using high-quality, LD-pruned common variants (MAF $> 1\%$) from the X chromosome. We then identified samples with discordant or unknown sex based on their F-statistic. The results of this imputation process are presented in Figure B.1 in the Appendix.

Defining ancestry in samples

We focused our analysis on a set of samples with genetically well-mixed population to ensure sufficient case-control for a wide range of traits. This was accomplished through a series of PCA steps. Initially, we conducted PCA on the 1KGP cohort, excluding related individuals, using LD pruned autosomal variants. Subsequently, we projected the UKBB while accounting for shrinkage bias. To assign ancestry, we employed a RF classifier trained on 1KGP super-population labels. This classifier was used to predict the super-population of each UKBB sample (Figure B.2). We defined a European subset as samples with a classifier-assigned probability exceeding 0.99 of being European. The RF classifiers were implemented using the `randomForest` (4.6) `R` library.

Final Hard Filters

We remove samples with disagreement between reported and imputed sex, remove samples with excess URVs. We then apply a final set of variant-filters, and remove variants that do not satisfy any of the criteria, and the result is illustrated in the Appendix Figure B.3:

- Non-invariant site
- Call rate ≥ 0.97
- HWE $P \geq 10^{-10}$

Then we apply a second collection of sample-level filters. As an initial pass to remove low quality and contaminated samples, we filtered out samples with call rate < 0.95, mean DP $< 19.5\times$ or mean GQ < 47.8. Using `hail`, we then run the *sample_qc* function to obtain and filter by four metrics. This is used to remove sample outliers based on the ratio of 1) transitions to transversions, 2) heterozygous to homozygous variants, 3) insertions to deletions, and 4) number of singletons (Figure B.4). Samples are removed if they are more than four standard deviations away from the mean stratified by ancestry.

3.2.2 Additional QC of genotyping array data

Both variants and samples from the bulk genotyping array underwent quality control by the UKBB team (see Section 2.2.1 in chapter 2). Variants were then lifted over from the GRCh37 to the GRCh38 genome assembly using `liftovervcf` (version 0.0.1), a component of the `SHAPEIT5` toolkit[141], along with the corresponding human genome reference. As an additional quality control step, I compared allele counts between sites in the genotyping array that overlapped with gnomAD (v4.0.1) WGS data[227]. For this comparison, I first restricted the analysis to individuals of predicted European ancestry and recalculated allele frequencies and counts using `PLINK v1.9`. Variants showing significant discrepancies in AF between the genotyping array and gnomAD were removed (two-sided Fisher's exact test, $P < 1 \times 10^{-50}$). This P-value threshold was selected after a comparison of multiple cutoffs and their impact on the resulting concordance between gnomAD and genotyping array AF. This final filtering step resulted in the removal of 4,958 out of 684,941 variants (0.7%) where there was overlap between gnomAD and genotyping array data.

3.2.3 Statistical phasing

I phased the 450k dataset following the same procedure as described in Chapter 2, Section 2.2.3. Briefly, this involved combining ES with genotyping array data. I first jointly phased common variants (MAF$\geq$ 0.001), which were used as a scaffold to phase rare variants. Prior to phasing, 99 probands in trios were removed, allowing for trio-SER assessment, which was conducted using a combination of `bcftools` and my pipelines, as outlined in detail in Chapter 2, Section 2.2.4.

3.2.4 Variant annotation

I annotated variants using `VEP v105`[262]. I employed a series of tools to predict deleteriousness *in silico*. To discriminate between the predicted functional impact of LoF variants, I used `LOFTEE v1.04`[227], which annotates PTV variants as either high confidence (HC) or LC LoF based on their likelihood to escape Nonsense Mediated Decay (NMD) or be annotation artifacts. For putatively damaging missense or

protein-altering variants, I used pre-computed annotations from `DBNSFP v4.3`[263], containing variant predictions from `CADD v1.6`[264] and `REVEL v7`[238]. As indels were not annotated in this version of `DBNSFP`, I manually re-ran `CADD v1.6` to annotate indels in our data. Splicing variants were annotated using `SpliceAI v1.3`[84]. I split all variants by transcript using `bcftools +split-vep`[209] and restricted to protein-coding transcripts defined by Matched Annotation from the NCBI and EMBL-EBI (MANE) Select[265]. If no MANE Select transcript was available for the gene, I restricted to the canonical transcript as defined by `GENCODE V39`[266]. Ordered by predicted impact on the cognate protein, I defined the following annotations based on the hierarchy. Any given variant received one single annotation of the following:

1. **pLoF**: HC variants included stop-gained, essential splice, and frameshift variants that have been filtered by `LOFTEE`.

2. **Damaging missense or protein-altering variants**: Any variant that is not categorised in (1) (HC pLoF) with at least one of the following are true:

   - The variant is an in-frame indel, stop-loss, start loss or missense variant with a `REVEL`-Score ≥ 0.773 or `CADD`-Phred score ≥ 28.1[267].
   - The SNP has a SpliceAI Δ score ≥ 0.5 for the donor loss, donor gain, acceptor loss or acceptor gain category.
   - The variant has been designated as PTV variant designated as LC by `LOFTEE`.

3. **Other missense**: defined as variants that are either in-frame indel, stop-loss, start loss or missense variants that have not been categorised in (2) (damaging missense and/or protein altering).

4. **Synonymous** variants are defined as any synonymous variant with `SpliceAI` Δ score < 0.5.

5. **Non-coding** variants are defined as downstream and upstream gene variants in neither of the above categories.

Finally, we removed annotations for all variants that have a gnomAD[227] Max Broad ancestry AF > 0.05, meaning that any variants with AF above this threshold in a gnomAD super population would be removed.

3.2.5 Encoding nonadditivity at a locus

The models described in this section were developed and tested with Dr Duncan Palmer, a senior post-doctoral researcher in the Lindgren lab. I wrote, adapted and ran scripts for these models, as well as analyzing the output.

Motivation

Nonadditive transformation assuming Hardy Weinberg Equilibrium

I first considered a model incorporating additive and nonadditive (dominance deviation) effects, as described by Palmer *et al.* and elsewhere[176,179,255,256,268]:

$$y = X^A\beta^A + X^D\beta^D + \varepsilon \tag{3.1}$$

where y is the standardized phenotype vector $(n \times 1)$ for n individuals, X^A and X^D are standardized additive and nonadditive genotype encoding matrices $(n \times m)$ for m loci, with corresponding effect size vectors β^A and β^D $(m \times 1)$, respectively. The noise component ε $(n \times 1)$ has mean zero and variance $1 - h^2_A - h^2_D$, where h^2_A and h^2_D are the additive and dominance heritabilities. Assuming HWE, the nonadditive encoding maps genotypes $[0, 1, 2] \rightarrow [-p/(1-p), 1, -(1-p)/p]$, where p is the minor allele frequency[176,179,255,256,268]. The nonadditive encoding is constructed such that it is uncorrelated with the corresponding additive encoding for that variant, which can be verified by testing $X^A \cdot X^D = 0$ (up to floating point precision). To test additive and nonadditive associations at a single locus j, I regress y on the column vectors X^A_j and X^D_j $(n \times 1$ each), interpreting the resulting scalar coefficients as estimates of β^A_j and β^D_j.

Nonadditive transformation without assuming HWE

To extend the approach beyond common variants[179], I consider a general case without assuming HWE. I parameterize the locus using genotype proportions:

$$\begin{bmatrix} r \\ h \\ a \end{bmatrix}, \text{ such that } r + h + a = 1. \tag{3.2}$$

Here, r, h, and a represent the proportions of homozygous reference, heterozygous, and homozygous alternate genotypes, respectively. At the gene level, these are equivalent to the proportions of wildtype, monoallelic, and biallelic carriers. I partition the genetic contributions into intercept, additive, and nonadditive components by defining the following vectors:

$$X^{0\prime} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \; X^{A\prime} = \begin{bmatrix} 0 \\ 1 \\ 2 \end{bmatrix}, \; X^{D\prime} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \tag{3.3}$$

where $X^{A\prime}$ reflects the number of alternate alleles, capturing additive effects, and $X^{D\prime}$ captures effects deviating from additivity. I define the inner product according to the general genotype proportions as:

$$\langle X, Y \rangle = rX_0Y_0 + hX_1Y_1 + aX_2Y_2. \tag{3.4}$$

I then apply the Gram-Schmidt process to orthogonalize these vectors relative to the defined genotype proportions, enabling the partitioning of genetic effects:

$$X^0 = \frac{1}{\left\| X^{0\prime} \right\|} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \tag{3.5}$$

$$X^A = \frac{X^{A\prime} - \langle X^{A\prime}, X^0 \rangle X^0}{\left\| X^{A\prime} - \langle X^{A\prime}, X^0 \rangle X^0 \right\|} = \frac{1}{\sqrt{a + r - (a - r)^2}} \begin{bmatrix} -(h + 2a) \\ 1 - (h + 2a) \\ 2 - (h + 2a) \end{bmatrix} \tag{3.6}$$

$$\tag{3.7}$$

$$X^D = \frac{X^{D\prime} - \langle X^{D\prime}, X^0 \rangle X^0 - \langle X^{D\prime}, X^A \rangle X^A}{\left\| X^{D\prime} - \langle X^{D\prime}, X^0 \rangle X^0 - \langle X^{D\prime}, X^A \rangle X^A \right\|} = \frac{1}{\sqrt{har(a + r - (a - r)^2)}} \begin{bmatrix} -ha \\ 2ar \\ -hr. \end{bmatrix} \tag{3.8}$$

A full derivation can be found in Appendix B. Notably, by substituting p^2 for r, q^2 for a, and $2pq$ for h, I recover the HWE encoding, which is demonstrated in Appendix B, and I confirm that the two components are independent through $\mathbf{X}^A \cdot \mathbf{X}^D = 0$, up to floating point precision.

Applying nonadditive transformation and scaling

The nonadditive encoding of genotypes proposed in equation 3.8 could produce values outside the conventional $[0, 2]$ range, posing challenges for standard genotype representation tools, such as `PLINK`[183]. To normalize the nonadditive genotype encoding matrix $X^D \in \mathbb{R}^{n \times m}$ (n samples, m loci), I apply a linear transformation to map values onto $[0, 2]$ while preserving orthogonality to the additive encoding. For each column j of X^D, let

$$s_j := \min_i X^D_{ij} - \epsilon, \quad b_j := \max_i X^D_{ij} + \epsilon, \tag{3.9}$$

for small epsilon ($0 < \epsilon \ll 1$) where $\epsilon > 0$. I define the transformation:

$$f(x) = 2 \cdot \frac{x - s_j}{b_j - s_j} \tag{3.10}$$

Applied element-wise to the matrix X^D:

$$X^{D''}_{i,j} = f(X^D ij) = 2 \cdot \frac{X^D_{ij} - s_j}{b_j - s_j}. \tag{3.11}$$

This transformation maps the minimum value to $0 + \epsilon$, the maximum to $2 - \epsilon$, and linearly interpolates others, preserving relative relationships while standardizing the range for each separate loci. Given that $X^D_j \cdot X^A_j = 0$ and our transformation is linear, $X^{D''}_j \cdot X^A_j = 0$ holds for each locus j without need for explicit verification.

3.2.6 Estimating genetic architectures

To estimate the genetic architecture of a trait at locus j, two separate models were fitted: an additive model using X^A_j (Equation 3.6) and a nonadditive model using X^D_j (Equation 3.8). These models yield marginal estimates $\hat{\beta}^A_j$ and $\hat{\beta}^D_j$ for the additive and nonadditive effects, respectively. The marginal estimated

genetic effect $\hat{g}_j$ associated to variation at locus j, which combines the additive and nonadditive contributions, is:

$$\hat{g}_j = \hat{\beta}_j^A X_j^A + \hat{\beta}_j^D X_j^D \tag{3.12}$$

The variance of this estimate, denoted as $\mathrm{Var}(\hat{g}_j)$, can be calculated by combining the variances of the additive and nonadditive components, which are derived from the standard errors of the estimators[269]:

$$\mathrm{Var}(\hat{g}_j) = \mathrm{Var}(\hat{\beta}_j^A X_j^A) + \mathrm{Var}(\hat{\beta}_j^D X_j^D) + 2\mathrm{Cov}(\hat{\beta}_j^A X_j^A, \hat{\beta}_j^D X_j^D) \tag{3.13}$$

We have that $\mathrm{Cov}(\hat{\beta}_j^A X_j^A, \hat{\beta}_j^D X_j^D) = 0$ given that the two components are scaled but remain orthogonal. This orthogonality simplifies the variance calculation and allows for the computation of confidence intervals around the estimated genetic effects, providing a measure of precision for the inferences at locus j. The elements of $\hat{g}_j$ represent the estimated summed additive and nonadditive contributions of locus j to the trait for incremental allelic substitution (wildtype, monoallelic, and biallelic, respectively).

3.2.7 Association analysis of alternate genotype encodings

Fast and efficient biallelic variant encoding

I developed a `C++` plugin for `bcftools` to efficiently identify and encode biallelic variants in VCF format (`https://github.com/BRaVa-genetics/biallelic`). This open-source tool rapidly annotates biallelic variations and encodes genotypes into pseudo variants (CH or homozygotes) in VCF format. The tool supports standard additive ($[0, 1, 2]$), recessive ($[0, 0, 1]$), and nonadditive ($[-ha, 2ar, -hr]$) encoding of individual variant, haplotypes or biallelic variants. The tool operates by first processing phased VCF data and a mapping file to generate a list of mono- and biallelic carriers for each gene. It then constructs a modified VCF file suitable for downstream analysis in `SAIGE`[240]. This ensures rapid encoding of genetic data for various research applications, streamlining nonadditive analysis of large genetic datasets.

Table 3.1: Overview of the additive, recessive and nonadditive encoding leveraged in this chapter. AA refers to wildtype, AB refers to monoallelic and BB refers to biallelic. For the (unscaled) nonadditive encoding, the variables r, h and a can be estimated by determining the proportion of individuals with zero, one or two haplotype carriers for a loci.

Genotype	Additive	Recessive	Nonadditive
AA	0	0	- h * a
AB	1	0	2 * a * r
BB	2	1	- h * r

3.2.8 Simulation of counts of observed biallelic variants

To simulate counts of biallelic variants (homozygous and CH, I follow the same framework as outlined in chapter 2, Section 2.2.19.

Single variant and pseudo-variant analysis

I perform both variant-level and gene-level analyses to capture the full spectrum of genetic effects: For variant-level analysis, I estimate for each locus j: r_j, h_j, and a_j as the proportions of homozygous reference, heterozygous, and homozygous alternate genotypes, respectively. In the gene-level analysis, I use phased genotypes and collapse multiple variants by their respective haplotypes. For each gene g, I set r_g, h_g, and a_g as the proportions of individuals with zero, one, and two haplotypes carrying one or more alternate alleles, respectively. For both analyses, I apply the nonadditive transformation, followed by linear re-scaling as per equation 3.10 to restrict 'dosages' to the interval $[0, 2]$. I modeled three encodings using `SAIGE`[240,270]: standard additive $[0, 1, 2]$, recessive $[0, 0, 1]$, and nonadditive (dominance) $[-ha, 2ar, -hr]$ (after transformation and re-scaling) which are outlined in Table 3.1. The resulting nonadditive dosages are encoded in VCF to preserve floating point accuracy and the additive or recessive analysis are encoded in `plink` files (`.bed`, `.bim`, and `.fam`). Single-variant and pseudo-variant (gene-level) association analysis were performed using `SAIGE`, adjusting for the covariates $\text{sex}, \text{age} \times \text{sex}, \text{age}^2, \text{age}^2 \times \text{sex}, \text{PC1} - \text{PC10}$. I discard variants or pseudo-variants with `p.value.NA < 0.05 AND Is.SPA = FALSE`, as these loci have not converged SPA in the model. For quantitative

traits, I applied a deterministic rank-based Inverse Normal Transformation (INT) based on equation (3.14), before analyzing them with `SAIGE`:

$$Y_i^{int} = \Phi^{-1}\left(\frac{r_i - c}{N - 2c + 1}\right), \tag{3.14}$$

r_i denotes the standard rank of the ith observation within the set of N data points, c is a constant offset equal to 0.5, and $\Phi^{-1}(\cdot)$ signifies the inverse function of the standard normal distribution's cumulative distribution function (CDF), which follows $N(0, 1)$, and finally Y_i^{int} is transformed phenotype value for the ith data point.

Set-based analysis

Complementary to modelling pseudo-variant subject to the nonadditive encoding individually, one can employ the nonadditive encoding X^D to perform nonadditive set-based testing using `SAIGE-GENE+`[86], an extension of `SAIGE`. For this thesis, gene-sets simply constitute genes, although one could have considered gene-sets or pathways instead. The set-based analysis aggregates the effect of variants or pseudo-variants across sets S:

$$y_i \sim \sum_{j \in S} w_j X_{i,j}^{D''} \tag{3.15}$$

where y_i is the phenotype for individual i, w_j is the weight for variant j, and $X_{i,j}^{D''}$ is the transformed nonadditive encoding for individual i at variant j. This allows me to conduct various tests, including burden testing, SKAT[271], and SKAT-O[272]. For variants, I define sets based on the canonical transcript (see Section 3.2.4). I specify the per-variant weighting w_j through a beta distribution:

$$w_j = \text{Beta}(\text{MAF}_j, 1, 25). \tag{3.16}$$

This weighting scheme upweights rare variants, as the variance of effect sizes for these variants is expected to be larger[273]. To ensure the stability and validity of my analyses, I remove ultra-rare variants (minor allele count MAC< 10). This step is necessary because `SAIGE` typically collapses these variants into a single pseudo-variant, which is not yet compatible with the nonadditive encoding approach presented here.

3.2.9 Simulation of null trait and analysis

Simulation and analysis of null traits

Because my biallelic analysis deviated from the conventional use of `SAIGE`, specifically by modelling genes instead of single variants, under a nonadditive encoding scheme, I employed a simulation framework to analyze the asymptotic behaviour of the resulting test statistic[251]. I simulated random phenotypes $\mathbf{y}$ that follow the normal distribution with a mean of zero and a covariance structure specified by the sparse GRM derived from `SAIGE` step0 using `sparseMVN v0.2.2`[274]. I let $\mathbf{g} \sim \mathcal{N}(\mathbf{0}, \mathbf{K})$ where $\mathbf{K}$ is GRM. Then the phenotype $\mathbf{y}$ can be simulated as:

$$\mathbf{y} = \sqrt{h^2}\mathbf{g} + \sqrt{1-h^2}\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \tag{3.17}$$

where h^2 represents the heritability of the trait, and $\mathbf{g} \sim \mathcal{N}(0, \sigma_g^2)$ is the 'genetic' contribution that is normally distributed with mean zero and variance σ_g^2. The 'environmental' component $\sqrt{1-h^2}\epsilon_i \sim \mathcal{N}(0, 1-h^2)$ is normally distributed with mean zero and variance $1-h^2$. The resultant phenotype y has mean of 0 and the variance is 1, with covariance structure governed by the relatedness of samples in the dataset. I binarized phenotypes by applying the liability threshold model[275], assuming multiple different population prevalences. Under this framework I simulated a total of 600 random binary and continuous phenotypes for 395,517 individuals across heritabilities (0.01, 0.1, 0.2, 0.5) and prevalences (0.05%, 0.1%, 0.5%, 1%, 10%).

Next, I set out to visualize the test statistic behaviour when associating simulated traits with the nonadditive encoding of genotypes stratified by counts of identified biallelic variants, referred to here as biallelic count or Biallelic Count (BC). To do this, I first performed an analysis using `SAIGE`[240] associating simulated traits with pLoF+damaging missense/protein altering pseudo variants subject to the nonadditive encoding. I generated Quantile-Quantile (QQ)-plots across trait prevalences, filtering to BC intervals: $(0, 5], (5, 10], (10, 25], (25, \infty]$, demonstrated in Figure 3.7 with corresponding λ_{GC}-values found in table 3.3. This investigation revealed higher instability in P-values, particularly for less common biallelic variants

and traits with BC below 5. Consequently, I implemented a BC ≥ 5 filter for subsequent analyses, recognizing that additional filtering criteria, explored in the following section, would be necessary.

Expected biallelic counts in cases

To mitigate the remaining P-values instability, I adopt a strategy initially developed and described by Karczewski *et al.*[227]. Using the frequency of affected haplotypes (Biallelic Frequency (BF)) for each gene, and the number of cases n_{cases} for a given trait, I defined an element-level filter Expected Biallelic Count (EBC) as BF $\times\, n_{\text{cases}}$. For quantitative traits, this number is defined as BF $\times\, n_{\text{defined}}$ where n_{defined} is the number of defined (i.e. non missing) values. Then, for each gene-trait association test, I filter to genes with EBC$\in (0, 0.5], (0.5, 1.0], (1, 10]$ and evaluated λ_{GC} across EBC intervals and population prevalences (0.05%, 0.1%, 0.5%, 1%, 10%) (Figure 3.8, Table 3.3). Quantitative traits were not affected by this filter. However, for binary traits, I observed that λ_{GC} values converged around 1 as the EBC increased. I observed that restricting to gene-traits with EBC≥ 0.5 resulted in well-calibrated test statistics for both rare traits (0.05% trait prevalence, $\lambda_{GC} = 0.937$) and common traits (10% trait prevalence, $\lambda_{GC} = 0.932$), and I therefore restricted all downstream analyses to EBC> 0.5.

Computing the experiment-wise significance threshold

I set out to compute an experiment-wise significance threshold to account for multiple testing. To do this, I leveraged the P-values generated from association analysis between null traits and biallelic genotypes, after restricting to gene-traits with EBC> 0.5, which were well-behaved. I then took the median of the smallest P-values obtained across 100 continuous simulated null traits. This value served as an estimate for the reciprocal of the total number of effective tests and could be used to define a false discovery rate of 1%: experiment-wise $P = 0.01 \times \text{median}(p_{\text{min}})$. The experiment-wise threshold was $P < 3.72 \times 10^{-7}$ for the additive encoding, $P < 1.15 \times 10^{-6}$ for a recessive ($[0, 0, 1]$) encoding and $P < 1.65 \times 10^{-6}$ for the nonadditive encoding. To analyze all gene-traits using the same

threshold, I opted for the most conservative threshold across the three: $P < 3.72 \times 10^{-7}$. In comparison, a conservative Bonferroni threshold would have yielded $P = \frac{0.05}{2214 \text{ genes} \times 827 \text{ traits}} = 2.7 \times 10^{-8}$.

3.2.10 Phenotype curation

Binary trait curation

To generate binary phenotypes, I extracted, amalgamated and translated ICD10 and ICD9 field codes from a variety of sources including primary, secondary care data, death records, and National Health Service (NHS) cancer registry data in UKBB. Each individual ICD9 were mapped to the equivalent ICD10 using mapping files provided by UKBB. A sample was designated as a 'case' if their health record include at least one of the ICD10 codes considered, and was otherwise assigned to the control group. Traits with fewer than 200 cases in the dataset were excluded.

Quantitative traits

I first considered 266 quantitative traits from UKBB, including NMR metabolomics, biomarkers, physical measures, blood biochemistry, anthropometry, body size measures, body composition by impedance, hand grip strength, blood pressure, blood assays, blood count, spirometry, and urine assays (See Appendix B.2). These traits were selected from the UKBB showcase fields based on several criteria: they were classified as integer or continuous value types, had measurements for at least 1,000 participants, did not use percentage as their unit, and had a defined unit of measurement. All traits were obtained during the initial visit of participants to a UKBB assessment center. I also created two derived traits: waist-to-hip ratio (WHR) and WHR adjusted for BMI (WHRadjBMI), defined as the residuals of WHR regressed on BMI. To reduce the computational burden, and to avoid reporting associations for very similar phenotypes, I pruned to a set of 114 relatively uncorrelated continuous phenotypes across the UKBB. To do this, I first determined the pairwise correlation between each phenotype using the R-package `Hmisc v5.1-3`. For all phenotypes with a coefficient of determination of $R^2 > 0.8$, I applied the

`maximal_independent_set` method using `hail v0.2`, resolving ties by selecting traits with fewer missing values.

3.2.11 Analysis of proteomic measurements

Description of the Olink assays

The Olink Explore 3072 Proximity Extension Assay (PEA) platform used complementary oligonucleotides immobilized on antibody probes. When two polyclonal antibodies bind the same protein in close proximity, the oligonucleotides hybridize, enabling DNA amplification of a unique barcode for the antigen, which is then quantified using NGS[276]. The platform comprises 2,941 immunoassays targeting 2,923 unique proteins distributed across multiple panels[90]. These panels consist of different dilution blocks, allowing the detection of proteins at various concentrations in blood[277]. Following extensive quality control by the manufacturer, Olink measurements are reported in Normalized Protein eXpression (NPX), representing relative quantification after $\log_2$ transformation. A negative control is used to calculate the Limit Of Detection (LOD) for each assay on every plate, effectively measuring the background noise present in the samples.

Assessment and Quality Control

Non-fasting blood samples were collected from all UKBB participants at recruitment, with plasma prepared and stored at -80°C. The Olink PEA was performed on 54,291 participants: 46,595 randomly selected samples representative of the general population, 6,376 samples selected by the UKB-PPP, and 1,268 individuals who participated in the COVID-19 repeat-imaging study[90]. Proteins were measured across eight panels, which includes groups of antibodies relevant for the study of different traits (cardiometabolic, cardiometabolic II, inflammation, inflammation II, neurology, neurology II, oncology, and oncology II). The UKBB Olink dataset underwent assessment and quality control by the UKB-PPP consortium team, including outlier removal, likely sample swap identification, removal of data with QC or assay warnings, and normalization. Following the manufacturer's recommendations, I

performed additional quality control: I substituted protein values below the LOD with $\frac{\text{LOD}}{\sqrt{2}}$ and standardized assay protein expression values to have mean zero and variance one. For separate visualization purposes, I also created NPX_{adj} through regression of sex, age, age^2, and $age \times sex$ with NPX by extraction the residuals.

Exome-wide proteomic association analysis

The association analysis was conducted on a subset of 42,217 individuals with European genetic ancestry. I performed gene-level association analysis using `SAIGE`, adjusting for the covariates sex, $age \times sex$, age^2, $age^2 \times sex$, and PC1 − PC10. I transformed raw NPX values using rank-based INT (see equation (3.14)). I carried out analyses at the haplotype level, collapsing variants by annotation as described in Section 3.2.4 across haplotypes in genes. I tested for associations gene-by-gene under three genetic models by recoding genotypes: a standard additive encoding $[0, 1, 2]$; a canonical recessive encoding $[0, 1, 2] \rightarrow [0, 0, 1]$; and a nonadditive encoding. For the nonadditive encoding, I first defined r, h, and a as the proportions of individuals with zero, one, and two haplotypes carrying one or more alternate alleles, respectively. I then mapped genotypes $[0, 1, 2]$ to $[-ha, 2ar, -hr]$ for each pseudo-variant, followed by a linear rescaling to restrict dosages to the interval $[0, 2]$. (3.2.5). I applied a conservative correction for multiple testing that accounted for both the number of Olink assays tested (traits) and the approximately 20,000 protein-coding genes in the human genome. This method differs from approaches used in some previous studies, such as the permutation testing described by Dhindsa *et al.*[128], which was not computationally feasible for this analysis. The significance threshold for initial discovery was set at $P < 8.33 \times 10^{-10}$ ($\frac{0.05}{3000 \text{ traits} \times 20{,}000 \text{ genes}}$).

3.2.12 Modelling recessive selection

The models described in this section were developed and tested together with Dr Duncan Palmer, a senior post-doctoral researcher in the Lindgren lab. I implemented and ran scripts for these models, as well as analyzing the output.

The following work was inspired by Bloemendal's 2015 unpublished work on the Levene Haldene distribution[278]. We let N be the number of individuals, $2N$ the total number of gene copies among them, K the number of copies carrying a deleterious mutation, and M the number of individuals with a biallelic mutation (either CH or homozygous alternate) in the gene. The objective was to determine the probability of observing M or fewer such individuals among N. The total possible configurations of K deleterious mutations among $2N$ gene copies are:

$$\binom{2N}{K}. \tag{3.18}$$

For configurations corresponding to homozygous variants, I computed $\binom{N}{M}$ to choose M affected individuals. These M individuals were assumed to have $2M$ mutated gene copies. The remaining $K - 2M$ mutations are distributed among the $N - M$ individuals, considering each mutation can be maternally or paternally inherited, which gives:

$$\binom{N-M}{K-2M}2^{K-2M}. \tag{3.19}$$

Hence, under the assumption of random mating, the probability that M individuals manifested biallelic variants, given K mutated copies, is expressed as:

$$P(M|K,N) = \frac{\binom{N}{M}\binom{N-M}{K-2M}2^{K-2M}}{\binom{2N}{K}}. \tag{3.20}$$

This haplotype-based probability model is consistent with variant-level models described elsewhere[180]. We refer to equation (3.20) as the gene-level HWE exact test.

Defining a one-sided test to investigate depletion of biallelic variants

Given that we were interested in investigating a depletion of biallelic variants in the gene of interest, I considered a one-sided test. We followed the definitions by Wigginton *et al.*[180] and set out to determine:

$$P(M \leq M'|K,N) \tag{3.21}$$

where M' is the observed number of biallelic variants in the gene that we observe. We can simply sum over the Equation (3.20):

$$P(m \leq M|K,N) = \sum_{m=0}^{M} P(m|K,N) \tag{3.22}$$

$$= \sum_{m=0}^{M} \frac{\binom{N}{m}\binom{N-m}{K-2m}2^{K-2m}}{\binom{2N}{K}} \tag{3.23}$$

As in Wigginton *et al.*[180,279], we could make use of the recurrence relations to increase efficiency, allowing us to proceed without computing multiple factorials.

$$P(N_{AB} = n_{AB} + 2|N, n_A) = P(N_{AB} = n_{AB}|N, n_A)\frac{4n_{AA}n_{BB}}{(n_{AB}+2)(n_{AB}+1)}, \tag{3.24}$$

$$P(N_{AB} = n_{AB} - 2|N, n_A) = P(N_{AB} = n_{AB}|N, n_A)\frac{n_{AB}(n_{AB}-1)}{4(n_{AA}+1)(n_{AB}+1)}. \tag{3.25}$$

Noting that $2n_{AA} + n_{AB} = n_A$, it follows that

$$P(N_{AB} = K - 2M|N, K) = P(N_{AA} = M|N, K) \tag{3.26}$$

and so we may write the recurrence relations in terms of the number of biallelic variants in our notation:

$$P(N_{AA} = M - 1|N, K) = P(N_{AA} = M|N, K)\frac{4M(N-K+M)}{(K-2M+2)(K-2M+1)}, \tag{3.27}$$

$$P(N_{AA} = M + 1|N, K) = P(N_{AA} = M|N, K)\frac{(K-2M)(K-2M-1)}{4(M+1)(N-K+M+1)}. \tag{3.28}$$

Calculating of mid-P values

As noted by Graffelman and Moreno[182], the P-value above has been known to be overly conservative, and they proposed a mid P-value for use with a two-sided test. We also propose the mid P-value for use in our one-sided test for a depletion of biallelic variants in the sample:

$$\text{mid-}P = P(m < M|K,N) + \frac{1}{2}P(m = M|K,M) \tag{3.29}$$

Analysis of recessive selection in UKBB

To analyze recessive selection in the UKBB dataset, I implemented the gene-level HWE exact test in a custom R package called `hardyr` (`https://github.com/frhl/hardyr`). This implementation incorporated equations (3.20) and recurrence relations (3.27), written in logarithmic space and implemented in `C++` to enhance computational efficiency, allowing for the analysis of samples with hundreds of thousands of individuals. The analysis was conducted on a set of 339,041 unrelated individuals of European ancestry from the UKBB. For each gene, I calculated the number of individuals with biallelic variants (M), the total number of mutated copies (K), and the total number of individuals (N). Using these metrics, I evaluated the one-sided HWE mid P-value, assuming heterozygous excess, to test for depletion of biallelic variants. This was accomplished using the `hwe_exact_test` function from the `hardyr` package, enabling potential assessment of recessive selection across the genome.

Power analysis

To establish a lower bound for MAF at which departures from HWE could be detected, I conducted a power analysis based on varying degrees of biallelic depletion. I followed a framework initially described by Rohlf and Weir[280]. While this framework has been implemented in the `HardyWeinberg` (1.4.6) `R`-package, its application is limited to populations with less than 1000 samples[281], motivating an efficient implementation (See 3.2.12). Central to HWE power testing is the $\theta = P_{Aa}/\sqrt{P_{AA}Paa}$ parameter, which depends on the genotype probabilities in the population, and describes the degree of disequilibrium. Under HWE, $\theta = 4$, with $\theta > 4$ indicating heterozygous excess and $\theta < 4$ indicating homozygous or compound heterozygous excess. I evaluated the degree of disequilibrium ($\theta \in \{8, 12, 16, 32\}$) I was powered to detect given 339,041 samples, across variants with MAF ranging from 0.001% to 1% using HWE exact test (Figure B.6). The selection of these θ-coefficients was guided by empirical inspection of the data (Figure B.5). Based on the

results from the power analysis, I restricted the analysis to sites with MAF ≥ 0.005, as these yielded 80% power to detect nominal significance ($P = 0.05$) at $\theta = 16$.

3.3 Results

3.3.1 High-quality phased variants enables biallelic analysis

Characterization of biallelic variants in the UKBB 450k ES release

I leveraged 20,737,653 high quality ES variants across canonical transcripts for 450k ES release after stringent quality control (Methods 3.2.1). This included multiple hard filters for variants, samples and individual genotypes, sample filtering steps within ancestry-specific batches, and requiring concordance between genetic and reported sex. For genotyping arrays, beyond the quality control performed by UKBB team[92], I excluded 4,958 of 805,426 sites that showed discrepancy between UKBB and gnomAD[227] WGS AFs to guard against genotyping errors (Two-sided fisher's exact test for AF difference $P < 1 \times 10^{-50}$, methods 3.2.2). I jointly phased combined quality-controlled exome and genotyping array data using `SHAPEIT5`[237] and quantified phasing quality using gold-standard SER across 100 trio-families (Table B.1). I verified that exome-wide SER were comparable to that of my previous investigation in chapter 2 and to what has been reported elsewhere[198], and restricted to confidently phased variants (PP$\geq$0.9).

I restricted to protein coding variants with MAF$<$ 0.05 and set out to identify mono and biallelic (homozygous and CH) genotypes using my `bcftools` plugin (Method 3.2.7). I annotated two different classes of putatively damaging variation. Firstly, I created an annotation mask for `LOFTEE` HC pLoF variants. Secondly, I created a mask for 'damaging missense/protein altering' variants that included putatively deleterious missense variants following recommendation by Pejaver *et al.*[267](`REVEL`$\geq$ 0.773, `CADD`$\geq$ 28.1), splice site variants (SpliceAI Δ score ≥ 0.5) and `LOFTEE` LC PTVs. Across 399,943 genetically ascertained Europeans (Methods 3.2.1), I observed 47,642 homozygous, 10,757 CH and 119,566 with pLoF variants *in cis*. My simulations corroborated the empirically observed frequencies of biallelic pLoF variants in a population of unrelated individuals of European ancestry, where I identified a strong positive correlation between simulated and observed counts for both homozygous ($R = 0.994$, $P < 2.2 \times 10^{-16}$) and compound heterozygous (CH)

($R = 0.903$, $P < 2.2 \times 10^{-16}$) pLoF events (Figure B.7, Methods 3.2.8). Combining pLoF and damaging missense/protein altering variant masks can potentially boost power if the additional variants contribute to the trait in a manner consistent with the pLoF variants: When combining the two variant annotation masks, I observed a total 131,224 homozygous carriers, 48,259 CH carriers and 221,453 carriers with variants *in cis*.

3.3.2 Biallelic effects across the plasma proteome

Proteogenomic study population description

I set out to explore the phenotypic consequences of biallelic inactivation on the protein level. To do this, I took advantage of the recent release of Olink proteomics, which provides a comprehensive quantitative assessment of 2,942 proteins in the blood across 54,219 individuals in the UKBB[90]. The Olink assays combine antibody pairs with attached oligonucleotides that, when bound to a target protein, hybridize to form a unique DNA barcode, which is then amplified and quantified via NGS to measure protein levels[276]. Consistent with observations from a recent study[282], I found that a considerable proportion of assays had protein levels below the LOD: On average, 7,648 (18.1%) individuals had values below the limit of detection across the 2,914 assays investigated. I first restricted my analysis to the subset of 46,596 randomized participants, excluding those pre-selected by UKB-PPP-consortium members or for COVID-19 repeat-imaging visits. Finally, I restricted to 42,217 individuals of European genetic ancestry with phased ES data (Methods 3.2.3). The 42,217 participants in the dataset were aged 40-70 years at baseline (mean 57.1 ± 8.1 years). The subset had a slight over representation of female participants (22,673; 53.7%), similar to the full UKBB dataset (273,474; 54.3%)[119].

Comparative analysis of variant annotation mask and the effect on protein abundance

I first characterized the impact of various genetic alterations gene-by-gene on their cognate protein level through additive association between the number of affected haplotypes and protein abundance. This allowed me to estimate the proportion of

Variant category	Total assays	|β| > 0		β > 0		β < 0	
		N	%	N	%	N	%
pLoF	1103	494	44.80%	9	1.80%	485	98.20%
LC	91	32	35.20%	1	3.10%	31	96.90%
Damaging missense and/or protein altering	2230	488	21.90%	23	4.70%	465	95.30%
synonymous	2729	214	7.80%	74	34.60%	140	65.40%

Table 3.2: Summary of the effects of variant categories on cognate protein abundance, based on additive haplotype-specific association analyses. Analyses between genes with at least 10 biallelic carriers and its cognate protein. The analysis was restricted to 42,217 Europeans with Olink proteomic data. The number of significant associations between a gene and the abundance of the cognate protein (N) is displayed alongside the percentage of total assays that have are significant. I stratify by direction of effect (β) and functional category. Note that the 'Total Assays' is different across variant categories as the number of biallelic carriers differ for each gene-assay pair, and as a result, does not always pass the haplotype count criterion (i.e. at least 10 affected haplotypes required).

variants in each functional category (pLoF and damaging missense/protein altering) that may have an effect on protein abundance, and by extension potentially be implicated in health and disease.

I interrogated the impact on protein abundance across annotation categories: pLoF, damaging missense and/or protein altering and synonymous variants with MAF$<5\%$. I also included LC PTV variants annotated with `LOFTEE` to better understand the ability of `LOFTEE` in classifying PTV variants. This category consisted of 59% (end-truncating) frameshift variants. Using phased genotypes, I collapsed variants with the same variant annotation mask across each haplotype, allowing me to perform additive gene-level analyses to test the association between number of affected gene copies with the corresponding protein abundance. I evaluated 2,741 unique gene-assay pairs across functional categories (1,103 pLoF, 2,230 damaging missense and/or protein altering, 91 LC PTVs, and 2,729 synonymous) with at least 10 affected haplotype in the population, regressing protein abundance on haplotype carrier status using `SAIGE`[240] (Method 3.2.11). To test the direct effects of variants that lie within the protein's cognate gene (referred to as *cis* protein Quantitative Trait Locis (pQTLs)), I restricted to 6,153 gene-trait pairs.

For this analysis, I observed well-calibrated type 1 errors across all annotation masks with combined $\lambda_{GC} = 0.99 \pm 0.5$ across all assays and annotation categories

prior to restricting to *cis* pQTLs). My results demonstrated that pLoF carriers are significantly more likely to exhibit changes in protein abundance (Bonferroni $P < 8.33 \times 10^{-10}$) compared to the other tested annotation categories. Of the 1,103 tested pLoF associations, 494 (44.8%) showed significant changes, with 485 of these (98.2%) associated to reduced protein levels (Table 3.2). In contrast, only 214 of 2,729 (7.8%) synonymous associations were deemed significant, with 140/214 resulting in reduced abundance. LC PTV variants identified by `LOFTEE` also showed a substantial impact: 32/91 (35.1%) associations were considered significant ($P < 8.33 \times 10^{-10}$), with 31/32 (96.8%) resulting in reduced expression. Damaging missense and/or protein-altering variants, which combined the categories `LOFTEE` LC PTV, variants with `SpliceAI` $\Delta > 0.50$, `REVEL` score > 0.773, and `CADD` > 28.1), displayed the lowest impact on protein abundance across all non-synonymous variant categories tested with only 488 of 2,230 (21.9%) significant associations after multiple testing correction ($P < 8.33 \times 10^{-10}$). Excluding synonymous variants, this category had the largest proportion of putative gain-of-function variants, with 23 of 488 (4.7%) associated with increased abundance in the cognate protein. This is consistent with observations elsewhere[128], suggesting several possible mechanisms. These variants may represent genuine gain-of-function mutations that enhance protein production or stability, or they could influence regulatory elements that increase gene expression. Alternatively, these variants might be in LD with nearby common variants that upregulate gene expression. To distinguish between these possibilities and elucidate the precise mechanisms underlying the increased protein abundance, conditional analysis controlling for common variant effects would be necessary.

I observed a hierarchy in effect sizes linked to the functional impact of variants on protein structure, as illustrated in Figure 3.1. pLoF variants generally exhibited the largest effect size (average $\beta_{\text{pLoF}} = -1.68 \pm 0.74$), followed by damaging missense/protein-altering variants (average $\beta_{\text{damaging}} = -1.29 \pm 0.80$), which were predicted to have a lower impact on the resulting protein. Synonymous variants demonstrated the smallest effect size (average $\beta_{\text{synonymous}} = -0.18 \pm 0.50$). The difference between effect sizes for pLoF and synonymous variants was statistically

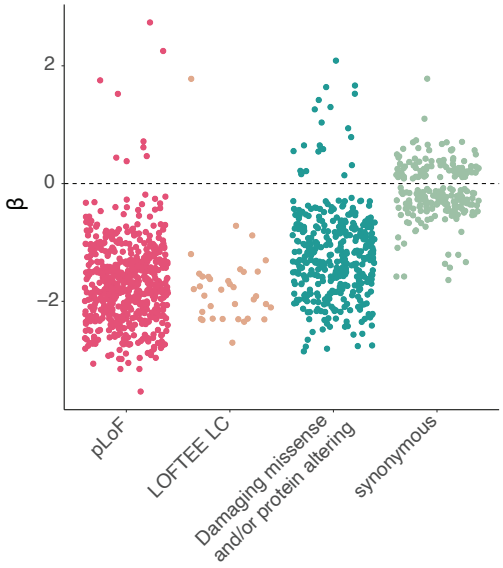


Figure 3.1: Assessment of effect size and direction across additive haplotype-level association analysis. Demonstrates the direction and effect β for significant ($P < 8.33 \times 10^{-10}$) *cis* pQTLs between protein coding genes and and cognate protein assuming an additive genetic architecture.

significant (two-sided t-test $P = 2.2 \times 10^{-16}$), further supporting the relationship between effect sizes and the functional impact of variants exerted on the protein structure.

Characterizing effects of haplotype disruption on cognate protein levels

I next sought to investigate whether I could observe differences in cognate protein abundance among wildtypes (i.e. individuals without any mutated gene copies according to the annotation masks), monoallelic carriers, and biallelic carriers. Specifically, I set out to inspect the effect on protein abundances following acquisition of multiple pLoF variants, with the goal to visualize potential linear or nonlinear relationships.

I restricted to variants with MAF<0.05 and identified all biallelic pLoF carriers in the dataset. I then aggregated the protein abundance following incremental variant acquisition in both gene copies stratified by functional annotation. In my subset of 42,217 individuals with proteomics and ES data available, I identified

8,721 (20.7%) participants with (1,692 CH and 7,029 homozygous) HC pLoF biallelic variants across 786 autosomal genes. I restricted to 34 genes in which I observed at least five individuals that are biallelic pLoF carriers. This resulted in 1,496 (341 CH and 1,155 homozygous) individuals being putative complete gene knockouts. As a negative control, I also considered 2,096 (939 CH and 1,157 homozygotes) carriers of synonymous variants in 22 of the 34 selected genes with $\geq$ 5 biallelic carriers in the population.

Given that each haplotype encodes one copy of a cognate protein, I reasoned that I should be able to observe a dosage-dependent effect on protein abundance proportional to the number of disrupted gene copies by pLoF variants (Figure 3.2). I first investigated how the acquisition of a single pLoF variant impacted NPX after regressing out sex and age, as these have considerable influence on abundance[90,283]. For 26/34 (76.4%) assays I observed a significant reduction in abundance for pLoF heterozygous carriers (two-sided Welch t-test, Bonferroni $P < 0.05/34$). When testing assays with heterozygous synonymous variant carriers, I found 11 (32.3%) assays with a significant change in abundance. For biallelic pLoF carriers I identified 22 of 34 (64%) assays that had a significant (Two-sided Welch t-test, Bonferroni $P < 0.05/34$) reduction in protein abundance (Figure 3.2). In contrast, for the 22 genes in which I had at least 5 biallelic synonymous carriers, 3/22 (13.6%) assays displayed a significant ($P < 0.05/34$) change in protein abundance, potentially driven by LD with nearby deleterious variants.

Empirical observations from Figure 3.2 revealed that many genes exhibited a nonlinear decrease in cognate protein abundance following incremental pLoF acquisition. For instance, biallelic inactivation of GIMPAP8 resulted in a more than twofold reduction in protein levels (adjusted NPX= -1.14) compared to monoallelic inactivation (adjusted NPX= -0.35). Interestingly, one gene (*HMCN2*) deviated from the typical pattern, failing to show a consistent decrease (or increase) in protein abundance with additional pLoF variants in each gene copy. For *HMCN2*, only monoallelic pLoF carriers exhibited a significant reduction in protein abundance ($P = 1.5 \times 10^{-35}$). In contrast, heterozygous synonymous ($P = 0.0051$), homozygous

synonymous ($P = 0.996$), and homozygous pLoF ($P = 0.138$) carriers showed no significant effect after correcting for multiple testing. Nonetheless, these observations necessitated a more sophisticated analytical framework to assess nonadditive effects while accounting for additional covariates.

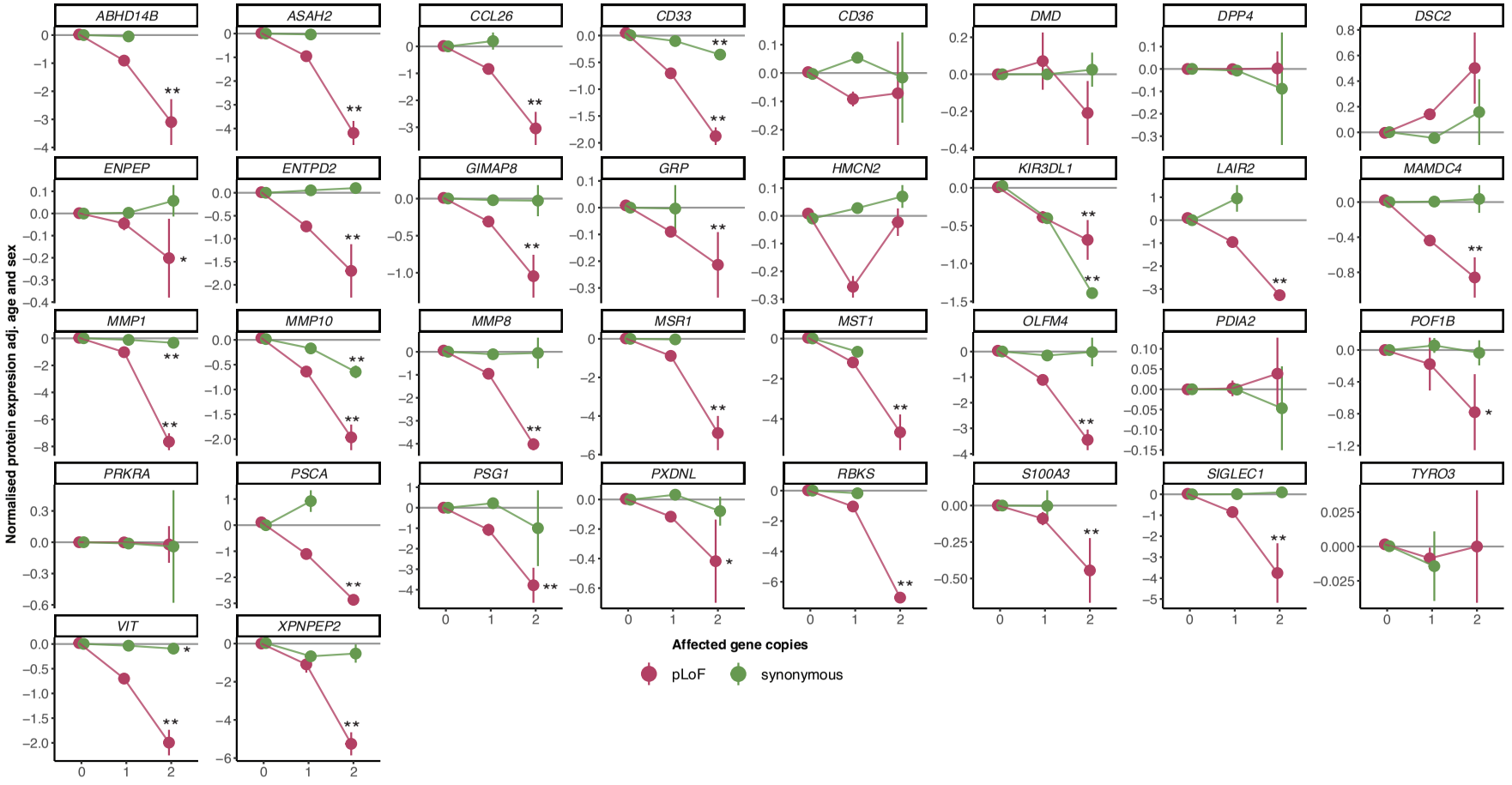
3.3.3 Allelic recoding yields nonadditive proteomic effects

Imposing orthogonality between additive and nonadditive components of heritability

Heritability quantifies the proportion of trait variance attributable to genetics[19]. It comprises additive, nonadditive, and epistatic components. Additive heritability represents trait variation explained by additive genetic effects, while nonadditive heritability captures additional variance from within-locus allelic interactions for a locus. My observations in Section 3.3.2 suggested potential nonadditive effects between biallelic variants and protein abundances. A commonly utilised approach for testing nonadditive effects (or the 'dominance deviation') incorporates $[0, 1, 0]$ or $[0, 0, 1]$ encoding into a joint model with the additive encoding $[0, 1, 2]$[258–260]. However, this approach does not yield the true nonadditive contribution due to correlation between the $[0, 1, 2]$ and $[0, 1, 0]$ genotype encodings, leading to additive spillover[179]. While a joint model of $[0, 1, 2]$ and $[0, 0, 1]$ shows less correlation due to fewer shared homozygous variants, the issue of additive spillover persists. To accurately test nonadditive effects, a genotype recoding that is completely uncorrelated with the additive encoding would be required.

Here, I propose a framework for recoding the allelic representation at a locus that imposes orthogonality between the additive and nonadditive components of heritability. This allows for explicit testing of the nonadditive and additive contributions to trait variance without spillover from additive effects[176,179]. This method improves upon previous approaches that imposed orthogonality under the assumption of HWE by parameterizing nonadditivity using allele frequencies[176,179,255,256,268] (Section 3.2.5). As highlighted in the introduction, the HWE assumption is often violated for rare variants due to strong selection pressures and the lower likelihood of

Figure 3.2: Protein abundance patterns for 34 genes with at least five biallelic pLoF variant carriers. The y-axis depicts protein abundance (NPX), adjusted for sex and age. The x-axis represents the number of affected gene copies, determined using phased genotypes (PP$\geq$ 0.90) for pLoF (red) and synonymous (green) carriers. Point ranges indicate 95% CI for the mean. On haplotypes with both synonymous and pLoF variants, pLoF takes precedence. Significant differences in mean NPX between wildtype and biallelic carriers are denoted by '**' (Bonferroni-corrected $P < 0.05/34$) and '*' (nominal $P < 0.05$). Note: For the X-chromosome gene *XPNPEP2*, only female carriers' NPX values are plotted.



achieving exact genotype proportions expected under HWE[284,285]. Instead, the proposed transformation is based on the *proportions* of wildtype, heterozygous, and homozygous alleles at a given locus. As a consequence, it avoids the HWE assumption, enabling testing of nonadditive effects for rare variants.

I applied the nonadditive recoding (Methods 3.2.5) to the genotypes of pseudo-variants (homozygotes and CH) in my dataset. For each pseudo-variant, I defined r, h, and a as the proportions of individuals with zero, one, and two haplotypes carrying one or more alternate alleles, respectively. I then mapped the genotypes $[0, 1, 2]$ to $[-ha, 2ar, -hr]$ for each pseudo-variant, and performed a linear rescaling to restrict dosages to $[0, 2]$ (Methods 3.2.5).

Testing nonadditive effects between biallelic variants and protein abundances

I set out to perform additive and nonadditive association analysis between protein abundances and pseudo-variants using `SAIGE` (Methods 3.2.8). This analysis expanded to gene-traits beyond the cognate protein, hence considering both *cis* and *trans* pQTLs. I first tested exactly 200 genes with at least five biallelic pLoF carriers across 2,910 unique assays, examining the additive, nonadditive, and the canonical recessive ($[0, 0, 1]$) genetic contributions. I performed a total of 33,577,856 tests throughout. To account for multiple testing, I applied a conservative Bonferroni correction, assuming 20,000 protein-coding genes and approximately 3,000 traits, yielding a significance threshold of $P < 8.33 \times 10^{-10}$ ($\frac{0.05}{3{,}000 \times 20{,}000}$). My tests were well controlled (Figure B.8), and the resulting P-values had well-calibrated type-1 errors for both additive ($\lambda_{GC} = 1.03$) and nonadditive ($\lambda_{GC} = 0.96$) tests (Figure B.8). I identified 168 additive and 4 nonadditive significant ($P < 8.33 \times 10^{-10}$) associations across 72 unique genes and 141 unique traits (Figure 3.3, Table B.4). The four significant nonadditive associations included were *GIMAP8*-GIMAP7, *HMCN2*-HMCN2, *MMP10*-MMP10, and *VIT*-VIT. These were also significant under an additive model (Table B.4). Interestingly, *HMCN2*-HMCN2 was only detected by the additive ($P = 4.01 \times 10^{-21}$) and nonadditive ($P = 4.63 \times 10^{-25}$) models but not the canonical recessive $[0, 0, 1]$ model (Recessive $P = 0.28$). Three of

these associations (HMCN2, MMP10, and VIT) demonstrated a direct relationship between the gene and its encoded protein, suggesting potential nonadditive *cis*-regulatory mechanisms, while the *GIMAP8*-GIMAP7 association indicates a possible *trans* regulatory effect. Although these associations were previously reported in the flagship Olink rare variant paper[128], that study only described the additive contributions, which I indeed replicate for 4 of 4 of these gene-assays. However, the nonadditive effects have not been reported in the literature.

Next, I included damaging missense/protein altering variants alongside pLoF variants, increasing the number of analyzable genes from 200 to 720. I again confirmed appropriately calibrated type-1 errors (Figure B.8), and identified 562 additive and 9 nonadditive significant associations across 186 genes and 436 assays (Figure 3.4, Table B.5). Of the 9 significant nonadditive associations, 3 of the 4 gene-trait associations from the pLoF-only analysis remained significant, with *VIT*-VIT (nonadditive $P = 3.5 \times 10^{-8}$) falling short of my stringent threshold ($P < 8.33 \times 10^{-10}$) when including damaging missense/protein altering variants. Additional gene-assay associations included *FLG* with GPR15L, PYDC1 and KLK8, alongside *CWH43*-SBSN and *DXO*-DXO most of which reflected nonadditive effects beyond the cognate gene. While these associations had previously been reported in additive analyses, their nonadditive contributions had not been examined. Moreover, both the additive and nonadditive association between *FLG* and GPR15L had not been reported in existing literature or the Olink flagship paper[128].

Nonadditive analysis reveals 8 of 9 gene-traits that would have been incorrectly classified as additive using alternate methods

Heyne *et al.* proposed identifying recessive effects by comparing P-values from additive $[0, 1, 2]$ and canonical recessive $[0, 0, 1]$ association analyses for the same locus[192]. They classified an association as 'recessive' if its recessive P-value was two orders of magnitude more significant than its additive P-value. I evaluated how many associations this approach might miss compared to analyzing P-values derived from nonadditive encoding. I performed recessive $[0, 0, 1]$ association analysis, identifying 68 significant associations ($P < 8.33 \times 10^{-10}$) (Tables B.4-B.5). Comparing additive

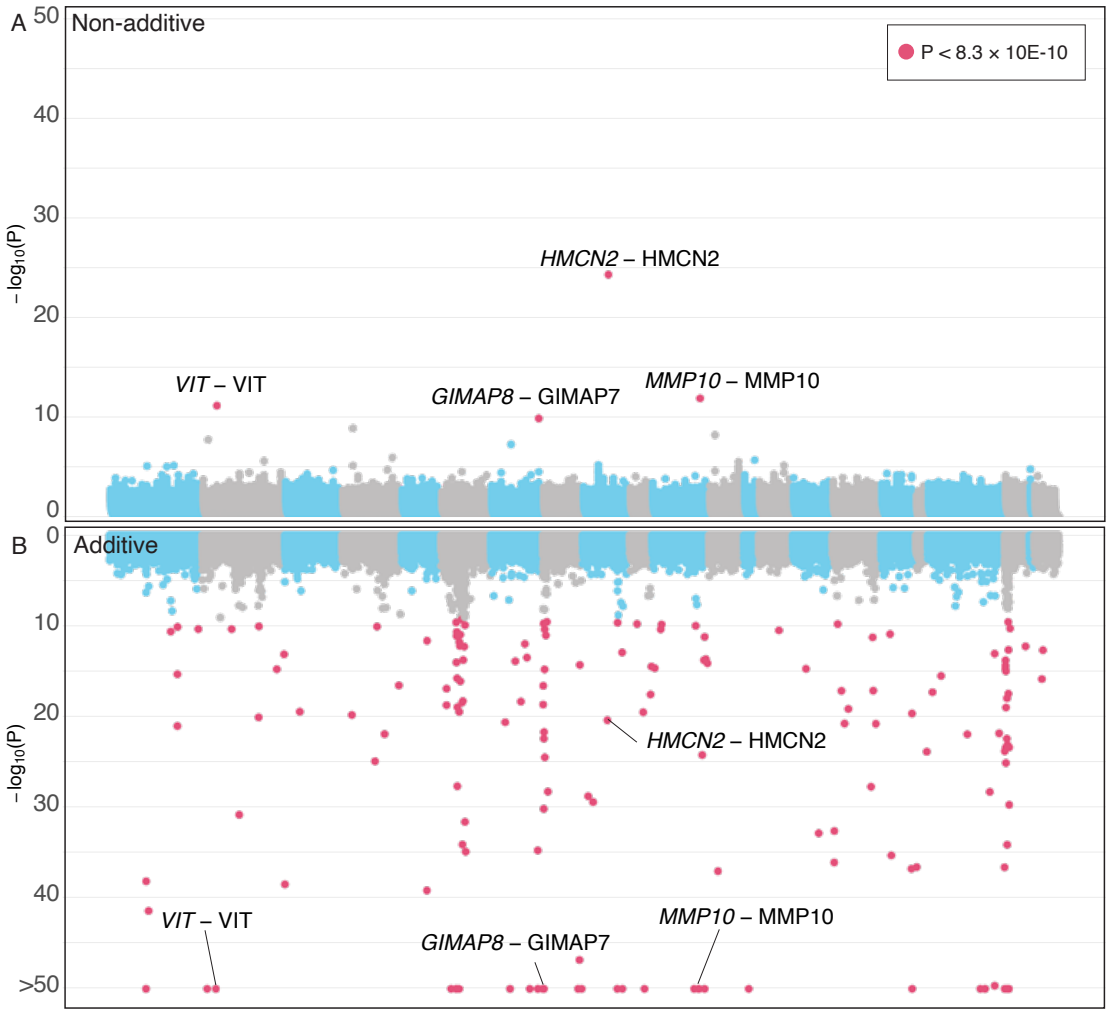


Figure 3.3: Aggregated Manhattan plot for additive and nonadditive effects using rare pLoF pseudo-variants with associated protein abundance. Aggregate Manhattan plot for additive and nonadditive effects across 2,910 assays and 200 genes with at least five biallelic pLoF carriers. The nonadditive (A) and the corresponding additive (B) P-values are shown. Odd and even chromosomes are colored blue and grey respectively. Red indicates significant ($P < 8.33 \times 10^{-10}$) after stringent Bonferroni correction for multiple testing. The Y-axis has been truncated at $P < 10^{-50}$ to aid presentation.

and recessive P-values, only 2 of 68 (2.9%) were more significant in the recessive analysis: *APOBEC1*-TMPRSS15 (Additive $P = 0.0015$, recessive $P = 2.38 \times 10^{-11}$, nonadditive $P = 1.47 \times 10^{-10}$) and *ENTPD2*-GFAP (Additive $P = 0.00065$, Recessive $P = 2.27 \times 10^{-10}$, nonadditive $P = 4.16 \times 10^{-9}$). Notably, *ENTPD2*-GFAP only passed the P-value threshold in the recessive model. While this framework classified two associations as potentially recessive, the nonadditive model revealed a more complex picture. It identified nine total associations with significant nonadditive contributions. Moreover, among these nine, only one (*CWH43*-SBSN) would have been classified as 'recessive' when comparing additive and recessive P-values, suggesting that the majority would have been missed.

Estimating genetic architecture of quantitative proteomic traits

Analysis of pLoF and damaging missense/protein altering variants revealed nonadditive associations between gene copy perturbations and protein abundance across nine gene-trait pairs. To estimate the genetic architectures underlying these traits due to deleterious variation at a locus, I evaluated the combined additive and nonadditive contributions to the trait at incremental allelic substitution (Methods 3.2.6). By summing the additive and nonadditive components, we can examine how different allelic configurations relate to protein abundance.

Two compelling examples illustrated this framework: *MMP10*-MMP10 and *HMCN2*-HMCN2. For each, I compared the estimated genetic architecture to the empirically observed protein abundances. If the inferred genetic architecture was accurate, I expected it to closely approximate the observed NPX values. The *MMP10*-MMP10 association exhibited significant negative additive ($P < 1 \times 10^{-320}$) and nonadditive ($P = 3.5 \times 10^{-12}$) contributions to MMP10 abundance. The estimated genetic architecture revealed that biallelic carriers of pLoF and damaging missense/protein altering variants saw more than twice the reduction in cognate MMP10 compared to monoallelic carriers (Figure 3.5a). This estimation closely aligned with empirical observations (Figure 3.5b), validating the framework for this gene-trait pair. In contrast, variants in *HMCN2* demonstrated a positive additive

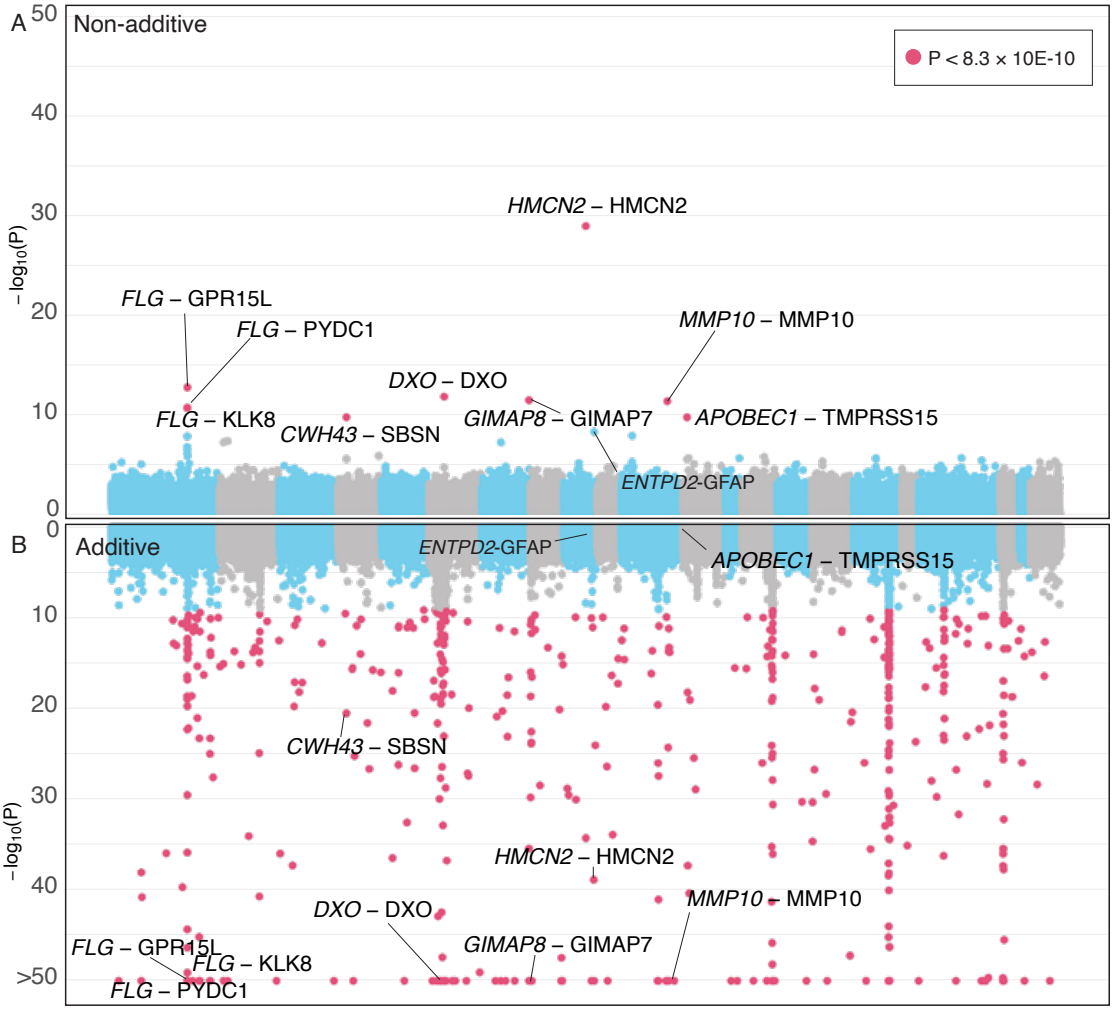


Figure 3.4: Aggregated Manhattan plot for additive and nonadditive effects using rare pLoF and damaging missense/protein altering pseudo-variants with associated protein abundance. Overview of the derived P-values from the nonadditive (a) and additive (b) analyses, exclusively for gene-traits with at least five biallelic pLoFs and damaging missense/protein altering variants in a gene. Odd and even chromosomes are colored blue and grey respectively. Red indicates significant associations ($P < 8.33 \times 10^{-10}$) after conservative Bonferroni correction for multiple testing. Note that an additional gene-assay association has been labelled here, namely *ENTPD2*-GFAP, as this was only significant in the recessive model. The y-axis has been truncated at $P < 10^{-50}$ to aid presentation.

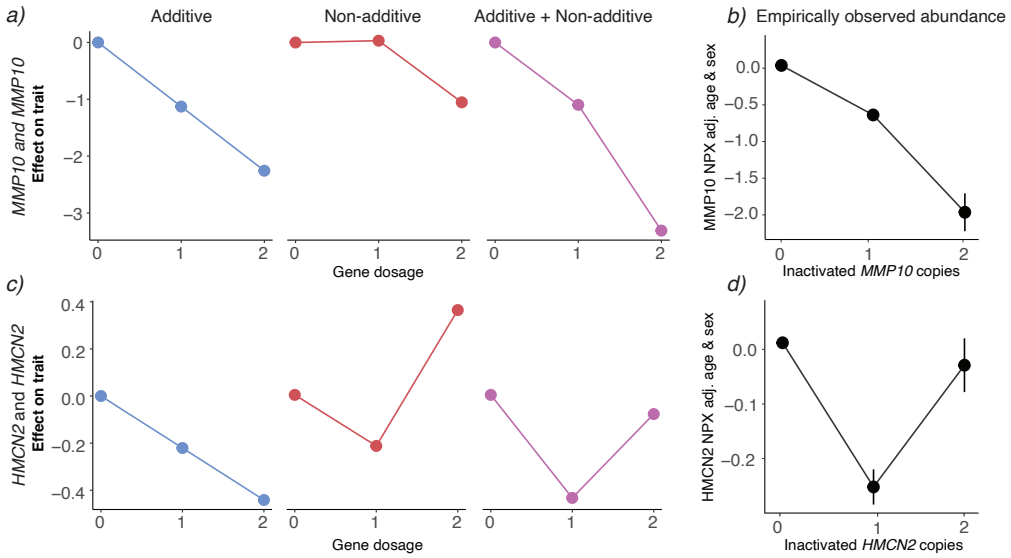


Figure 3.5: Comparison of inferred genetic architecture and observed protein abundance for *HMCN2* and *MMP10*. Panels (a) and (c) show marginal effects of gene dosage on trait expression using additive, nonadditive, and combined models for *HMCN2* and *MMP10*, respectively. Panels (b) and (d) display empirically observed abundance of HMCN2 and MMP10 proteins (NPX, adjusted for age and sex) relative to inactivated gene copies. The x-axis in all panels represents the number of inactivated gene copies (0, 1, or 2), corresponding to pLoF and damaging missense/protein altering variants. The y-axis shows the effect on trait or protein abundance. This figure demonstrates the concordance between estimated genetic models and empirically observed protein abundances.

($P = 4.01 \times 10^{-21}$) and a negative nonadditive ($P = 4.64 \times 10^{-25}$) contribution to HMCN2 abundance, resulting in a putative overdominant relationship. Therefore, individuals with a single affected gene copy exhibited significant reduction in cognate protein, whereas those with both gene copies affected demonstrated only a subtle reduction in cognate protein abundance compared to wildtypes (Figure 3.5c). Again, this pattern recapitulated the empirically observed protein abundance across increasing allelic dosages (Figure 3.5d).

I also estimated diverse genetic architectures for the remaining gene-trait associations (Figure 3.6), with a general trend that *cis* pQTLs resulted in a decrease in protein abundance, whereas *trans* pQTLs led to an increase in abundance for their associated protein. While I observed a significant additive association between *GIMAP8* and its cognate protein (additive $P = 5.07 \times 10^{-198}$), the nonadditive contribution was not significant (nonadditive $P = 0.0057$) and was

therefore not picked up in the initial analysis. Instead, my results indicated a nonadditive association between *GIMAP8* and a reduction in GIMAP7 (nonadditive P=2.79×10^{-12}). I also observed significant pleiotropic nonadditive associations between variants in *FLG* and increases in abundances for GPR15L (nonadditive $P = 1.47 \times 10^{-13}$), KLK8 (nonadditive $P = 1.7 \times 10^{-11}$), and PYDC1 (nonadditive $P = 1.56 \times 10^{-11}$). Additionally, I identified a significant nonadditive association (additive $P = 0.0015$, nonadditive $P = 1.46 \times 10^{-10}$) between biallelic variants in *APOBEC1* and TMPRSS15 abundance, suggesting that complete biallelic inactivation of *APOBEC1* may lead to a reduction in TMPRSS15 levels.

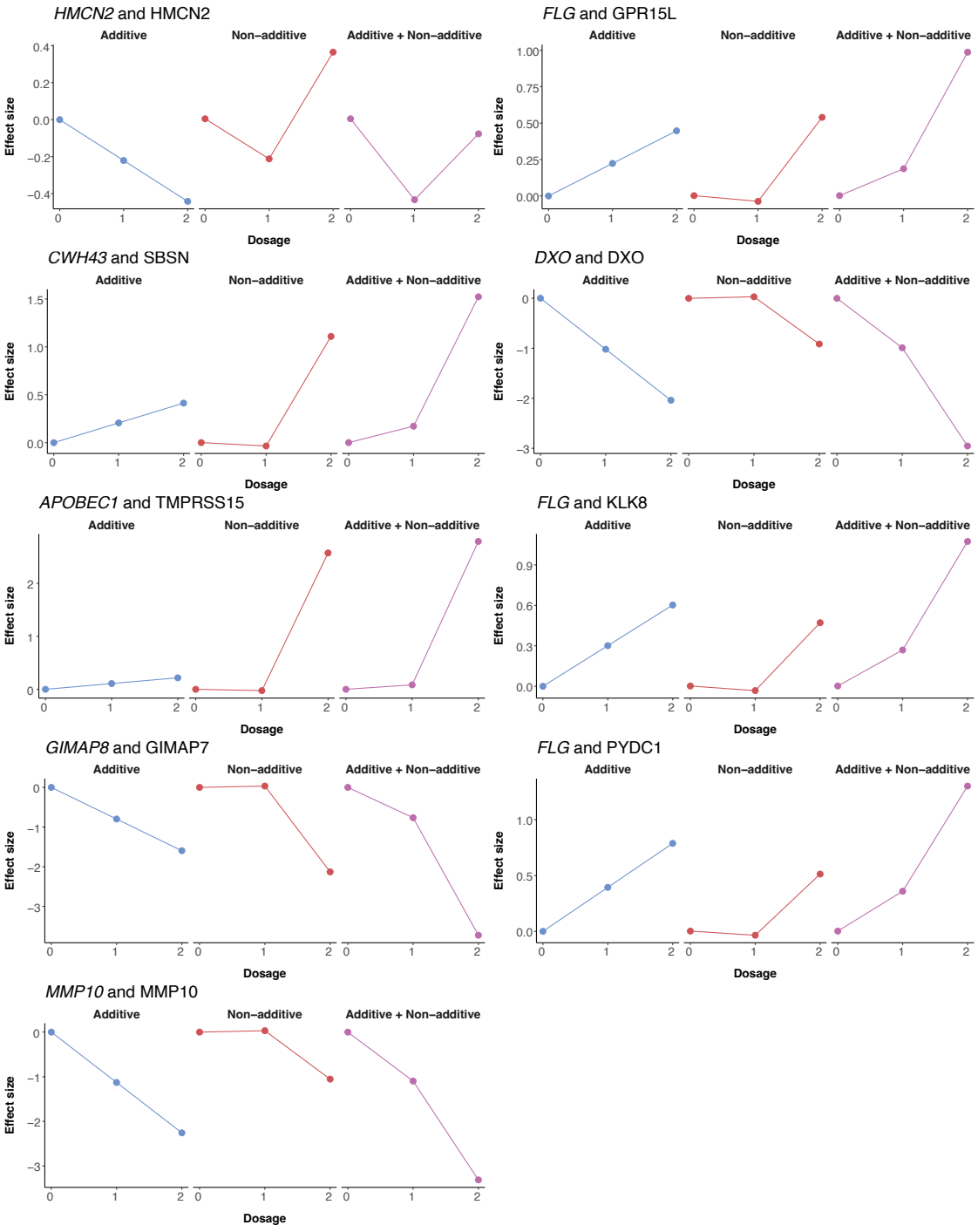


Figure 3.6: Genetic architectures of significant associations with proteomic measurements using pLoF+damaging missense/protein altering mask. The x-axis represents the number of inactivated gene copies (0, 1, or 2), corresponding to pLoF and damaging missense/protein altering variants. The y-axis shows the effect on protein abundance. Each panel depicts the additive model (blue), nonadditive model (red), and combined additive and nonadditive effects (purple). The purple line provides a generalized estimation of the genetic architecture for each locus, incorporating both additive and nonadditive effects.

3.3.4 Analysis of recessive effects across health and disease

Data description

The previous proteomic analysis was limited to 42,217 individuals, not utilizing the full UKBB. Moreover, while it explored molecular associations, it did not examine broader physiological outcomes that could have medical relevance. To address these limitations, I conducted an exhaustive search for nonadditive effects across continuous biomarkers and primary/secondary care disease endpoints in 395,517 individuals of European genetic ancestry. Binary traits were chosen by aggregating two digit ICD-codes restricted to at least 200 cases (corresponding to a disease prevalence of approximately 0.05% in the dataset), resulting in 798 traits to be tested. I also analyzed continuous traits across 9 broad categories, which I pruned to a set of 114 less-correlated traits (Methods Section 3.2.10, Appendix B.2). For quantitative traits, I analyzed rank-based INT transformed values assuming an additive, canonical recessive ($[0, 0, 1]$) and nonadditive genetic architectures. In this section, I tested a combined 912 quantitative and binary traits.

3.3.5 Simulation of random phenotypes

Testing type-1 errors for binary traits with biallelic nonadditive genotype encoding

While quantitative traits typically behave predictably in GWAS, the impact of nonadditive analysis on rare binary traits and their resulting test statistics is less clear. To address this uncertainty, I implemented a simulation-based framework to analyze both simulated quantitative and binary traits. This framework was first designed to examine the asymptotic behavior of associations between binary traits and the nonadditive encoding, following a framework described by Karczewski *et al.*[251]. Specifically, I simulated null phenotypes for each of the 395,517 UKBB participants. These simulated phenotypes followed a normal distribution with zero mean and a covariance structure specified by a sparse GRM. I used the liability threshold model to convert 500 of 600 into binary traits assuming population prevalences of (0.05%, 0.1%, 0.5%, 1%, 10%) using the liability threshold model[275].

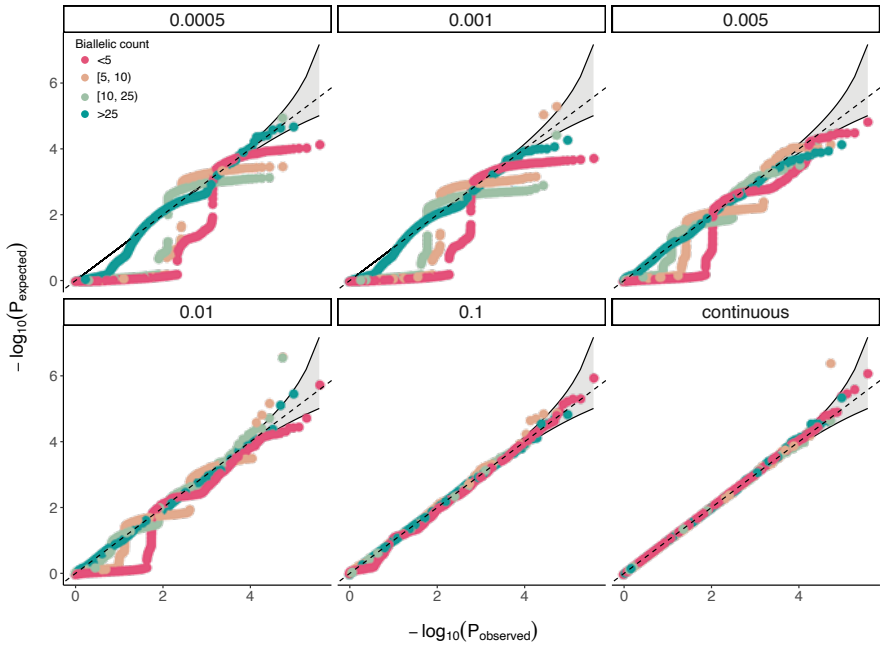


Figure 3.7: QQ-plots for 600 randomly generated binary and continuous heritable phenotypes tested against pseudo-variants with nonadditive encoding. Traits were generated with heritabilities ranging from 0.01 to 0.5. Each facet represents a population prevalence, from rare (0.0005) to common traits (0.1), including continuous traits. Associations are stratified based on biallelic count in cases, binned as $[0, 5], (5, 10], (10, 25], (25, \infty)$. For each bin and population prevalence stratum, observed P-values are plotted against expected P-values.

With these binary and quantitative phenotypes, I then proceeded to perform pLoF+damaging missense/protein altering nonadditive biallelic association testing with `SAIGE` and examined the behaviour of the resulting P-values under various phenotype prevalences and heritabilities. I inspected the behaviour of these tests using QQ-plots of the P-values (Figure 3.7), splitting the results from the results from the into BC intervals of $[0, 5], (5, 10], (10, 25], (25, \infty)$. For binary traits, `SAIGE` uses SPA to mitigate potential unbalanced case-control ratios which can results in inflation of type-1 errors[240,286]. 213 of 501,264 (0.042%) associations did not achieve SPA convergence and were therefore excluded from downstream analysis. I note that the affected markers were only observed as singletons, suggesting that these are already too rare to reliably test. When examining random phenotypes associated with biallelic variants subject to the nonadditive encoding I observed considerable P-value deflation for binary traits, particularly for rare biallelic variants

and for rare binary traits, but no inflation for quantitative traits (Figure 3.7, top left). This had been observed elsewhere[86,240,251]: the statistical power is directly correlated with the variant frequency and trait prevalence. The low power for these rare biallelic variants and rare outcomes underscored the value of imposing a threshold on the minimum number of biallelic carriers, and restricting to binary traits with sufficient sample sizes. For this reason, I imposed a cutoff of 5 biallelic variants for the nonadditive analysis and 10 haplotypes in the additive analysis for binary and continuous endpoints.

Imposing a dynamic filtering threshold that resolves remaining *P*-value instabilities

While I imposed a cutoff based on absolute biallelic count, I still observed some instability in *P*-values, particularly for rarer traits (Figure 3.7). To address this, I imposed an additional filter to account for case-control imbalance in rare traits, by introducing EBC, defined as BF $\times$ n_{cases}, where n_{cases} represent the number of cases and BF is the biallelic frequency (Methods 3.2.9). This filter allowed retaining variants that were sufficiently rare if the investigated trait had a substantial number of cases, or conversely, traits that were sufficiently rare but had a high number of biallelic carriers. As a result, the threshold adapted to both the scarcity of the biallelic variants and the prevalence of the trait in the study population, effectively balancing the trade-off between the two. I evaluated λ_{GC} across EBC intervals of $(0, 0.5], (0.5, 1](1, 10](10, \infty)$ for each random phenotype, characterizing inflation and deflation patterns by each stratum (Figure 3.8, Table 3.3). I observed that the lowest interval (EBC$<$ 0.5) explained the majority of the *P*-value instability. This suggested that markers with EBC$<$ 0.5 could be excluded when performing association analysis of biallelic variants under a nonadditive encoding scheme, resulting in more well-behaved *P*-values, albeit at the cost of removing genes with a low number of biallelic variants in rare traits.

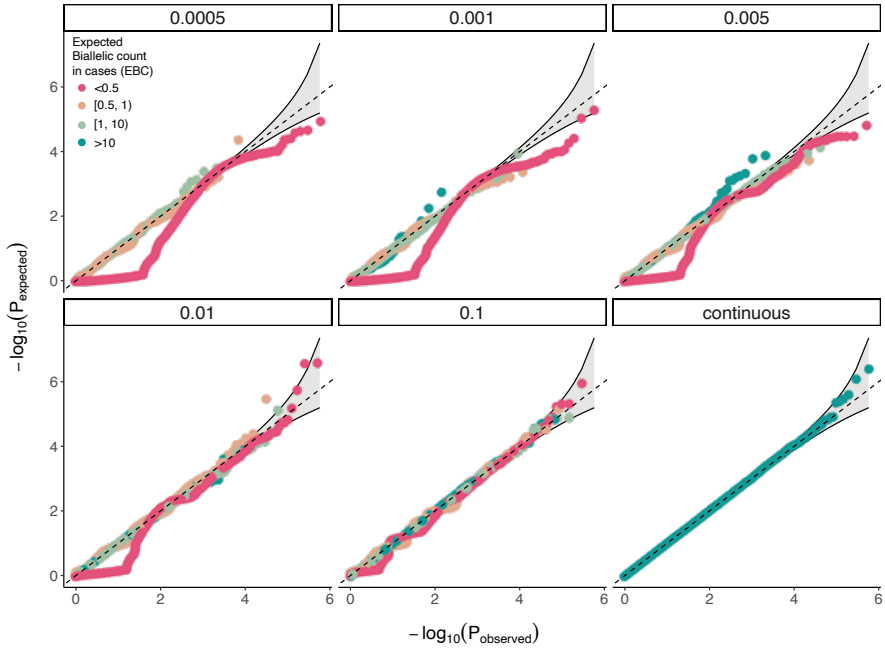


Figure 3.8: QQ-plot for simulated null traits stratified by Expected Biallelic Count. QQ-plots for randomly generated heritable phenotypes tested against pseudo-variants with nonadditive encoding, stratified by EBC. Facets represent population prevalences from rare (0.0005) to common (0.1) traits, including continuous traits. Simulated traits have heritabilities ranging from 0.01 to 0.5.

		λ_{gc} by trait prevalence					
	BC or EBC Interval	0.0005	0.001	0.005	0.01	0.1	continuous
Biallelic count (BC)	<5	0.003	0.005	0.017	0.034	0.367	1.011
	[5, 10)	0.010	0.016	0.082	0.176	1.075	0.980
	[10, 25)	0.017	0.033	0.187	0.388	0.975	0.994
	>25	0.108	0.219	0.717	0.964	1.012	1.008
	[5, ∞)	0.026	0.052	0.222	0.362	0.952	0.994
Expected Biallelic Count (EBC)	<0.5	0.005	0.008	0.030	0.052	0.308	NA
	[0.5, 1)	0.894	0.863	0.784	0.741	0.684	NA
	[1, 10)	1.068	1.052	1.030	1.026	0.982	NA
	>10	NA	0.890	0.946	1.029	1.015	1.006
	[0.5, ∞)	0.938	0.958	0.964	0.961	0.933	1.006

Table 3.3: λ_{GC} across BC and EBC cutoffs, stratified by trait prevalence. λ_{GC} values for simulated phenotypes, grouped by population prevalence (including continuous traits), and stratified by Expected Biallelic Count (EBC) or Biallelic Count (BC) filters. NA values indicate bins with no evaluable P-values for a given interval or prevalence stratum.

3.3.6 Exome-wide scans of nonadditive effects

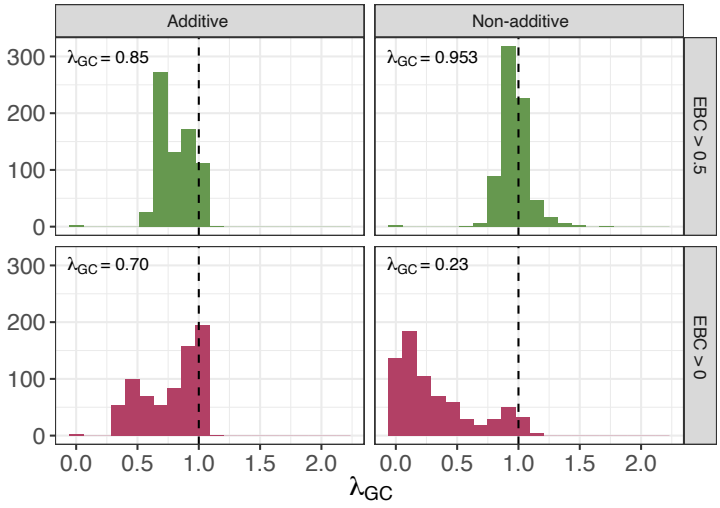
Identification of nonadditive effects across broad physiological endpoints

Next, I set out to test nonadditive effects gene-by-gene across a diverse collection of phenotypes in UKBB. I analyzed a total of 827 (716 binary and 111 quantitative) traits derived from primary care records, hospital records and visits at UKBB assessment centres. I curated binary traits from two digit ICD-codes with more than > 200 cases. For quantitative traits, I started with a broader set across eight categories and pruned them by removing highly correlated traits ($R^2 > 0.8$) to create a more independent set of measurements (Methods 3.2.10-3.2.10).

I analyzed 2,214 unique genes with at least 10 pLoF and damaging missense/protein altering variant carriers in the population. I performed a total of 1,830,978 nonadditive association tests, and compared the resulting P-values with that obtained assuming an additive genetic architecture for same genes (Figure 3.9). I accounted for multiple testing using the simulations of 600 random phenotypes setting a significance threshold (Experiment-wise $P < 3.72 \times 10^{-7}$) based on the median of the most extreme P-value observed across well-behaved random phenotypes (Methods 3.2.9). Guided by my simulations, and to ensure well-behaved type-1 errors, I restricted gene-trait pairs to those with EBC≥ 0.5, leading to an increase in λ_{GC} from 0.211 to 0.953 across the nonadditive, and 0.703 to 0.836 for the additive binary gene-trait analysis (Figure 3.9). In total, this filter removed 605,753 (37.8%) nonadditive gene-trait associations that would otherwise have been analysed. Finally, I removed exactly 300 nonadditive gene-trait associations in which SPA did not converge.

I identified 20 significant nonadditive associations (Experiment-wise $P < 3.72 \times 10^{-7}$) across 19 traits and 12 unique genes. These associations were evenly split between binary traits (10) and quantitative markers (10) (Figure 3.10, Table B.6). Notably, applying a more stringent Bonferroni-corrected significance threshold ($P = \frac{0.05}{2214 \text{ genes} \times 827 \text{ traits}} = 2.7 \times 10^{-8}$) would have resulted in 16 significant gene-trait associations (Table B.6). Among the binary traits, I identified and replicated several well-established Mendelian traits. These include *GJB2* and hearing loss (nonadditive

Figure 3.9: Distribution of λ_{GC} in additive and nonadditive analyses before and after EBC filtering. λ_{GC} distributions for 716 binary traits analyzed using additive and nonadditive genetic models, before and after filtering by Expected Biallelic Count (EBC). The top left of each facet shows the overall λ_{GC} derived from P-values across all gene-trait pairs.



$P = 1.02 \times 10^{-21}$), *MUTYH* and colorectal cancer (nonadditive $P = 1.13 \times 10^{-16}$), *FLG* and asthma (nonadditive $P = 1.01 \times 10^{-63}$), and *FLG* and atopic dermatitis (nonadditive $P = 5.07 \times 10^{-26}$). Additionally, I identified gene-trait associations that were not found in the initial 200k ES analysis (Chapter 2) but have been reported elsewhere. These include a significant additive and nonadditive ($P = 6.44 \times 10^{-11}$) association between damaging biallelic variants in *NOD2* and the incidence of Crohn's Disease (CD), a well-established relationship[287,288]. I also replicated an association between biallelic variants in *CHADL* and the incidence of osteoarthritis of the hip (nonadditive $P = 9.02 \times 10^{-12}$)[289,290]. Furthermore, I recapitulated a significant association (Additive $P = 6.82 \times 10^{-6}$, nonadditive $P = 3.67 \times 10^{-13}$) between biallelic variants in *FLG* and vasomotor and allergic rhinitis[291]. Finally, I observed a well-established relationship between biallelic variants in *CLYBL* and deficiency of vitamin B (nonadditive $P = 3.69 \times 10^{-10}$), which has been documented specifically in human carriers of deleterious biallelic variation[292].

Among the 10 significant quantitative gene-trait associations, I identified both well-established and potentially novel relationships. A notable example is the association between biallelic variants in *HAL* and increased histidine levels (nonadditive

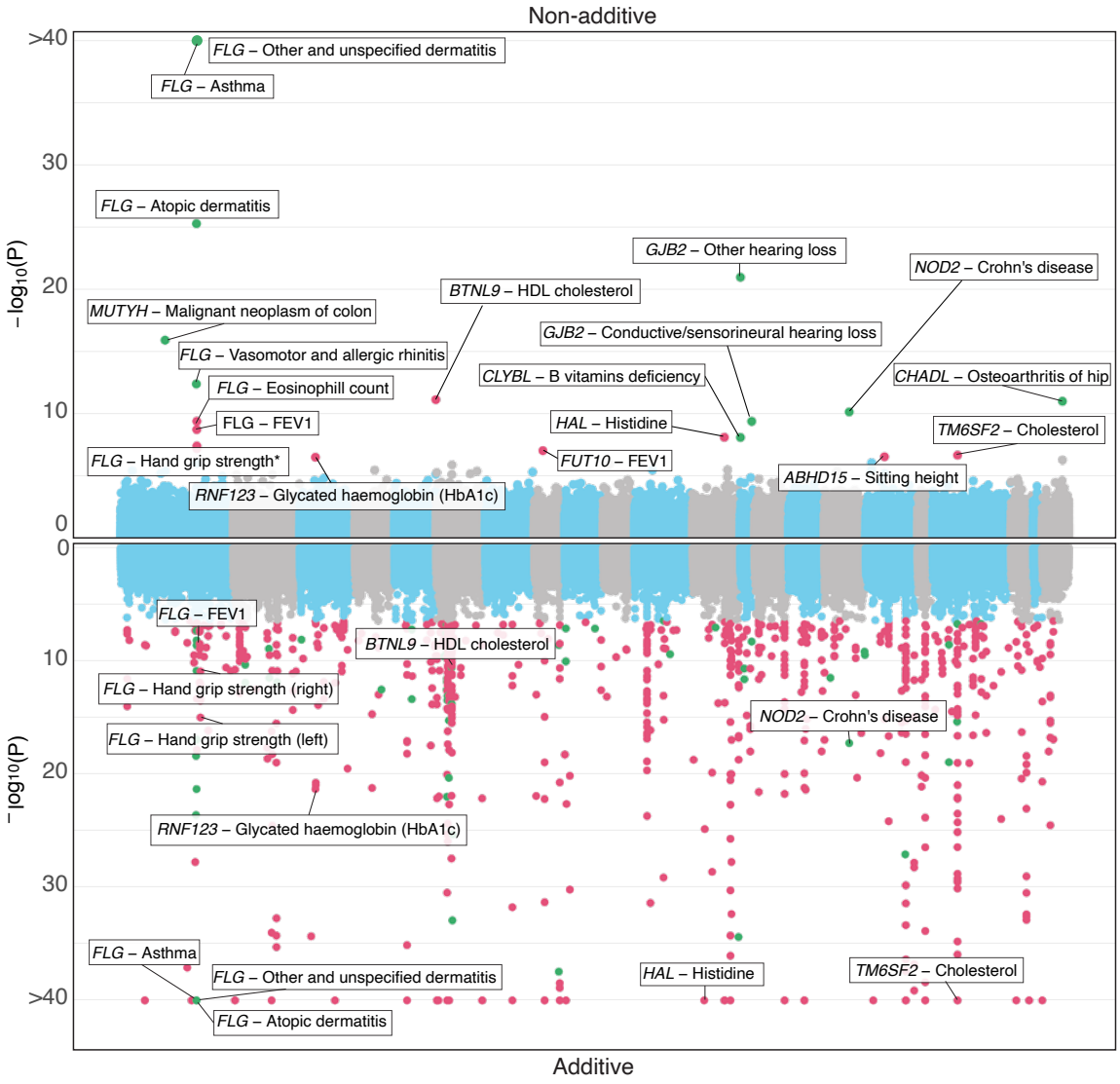


Figure 3.10: Aggregated Manhattan plot for additive and nonadditive effects using pLoF and damaging missense/protein altering variant mask. Nonadditive (upper) and additive (lower) contributions to quantitative (red) and binary (green) traits in UKBB. Y-axis shows $-\log_{10}(P)$, truncated at $P < 10^{-40}$ for clarity. Odd and even chromosomes are represented by blue and grey points, respectively. Significant nonadditive gene-trait associations are highlighted in the upper panel. In the lower panel, these associations are highlighted only if they are also significant in the additive model (Experiment-wise $P < 3.72 \times 10^{-7}$ for both models).

$P = 7.01 \times 10^{-9}$). This aligns with known biology, as *HAL* encodes histidase, an enzyme primarily responsible for histidine breakdown[293]. Biallelic variants in *FLG* showed multiple associations: reduced Forced expiratory volume in the first second (FEV1) (nonadditive $P = 1.68 \times 10^{-9}$), increased eosinophil count (nonadditive $P = 3.64 \times 10^{-10}$), and reduced hand grip strength (nonadditive $P = 5.67 \times 10^{-8}$). I confirmed an association between *TM6SF2* and cholesterol levels, previously reported using an additive genetic model[294], with significant effects under both additive ($P = 1.13 \times 10^{-51}$) and nonadditive ($P = 1.89 \times 10^{-7}$) models. Similarly, *BTNL9* highlighted a nonadditive association with high-density lipoprotein (HDL) cholesterol levels (nonadditive $P = 6.69 \times 10^{-12}$), supporting recent findings linking variants in this gene to lipid metabolism in Polynesian populations[295]. My investigation also revealed potentially novel gene-trait associations. Biallelic variants in *RNF123* are associated with Glycated haemoglobin (HbA1c) levels (nonadditive $P = 2.77 \times 10^{-7}$), suggesting a potential role in glucose metabolism. *ABHD15* is associated with sitting height (nonadditive $P = 2.53 \times 10^{-7}$). *FUT10* shows a nonadditive association with FEV1 (nonadditive $P = 8.20 \times 10^{-8}$), similar to the aforementioned *FLG*-FEV1 associations, suggesting multiple genetic factors may contribute to lung function in a nonadditive manner.

Of the 20 identified associations, only 9 (45%) gene-trait pairs would have been deemed significant (Experiment-wise $P < 3.72 \times 10^{-7}$) using an additive model (Table B.6). Finally, if the mode of inheritance had been determined by comparing additive and recessive P-values, as proposed by Heyne *et al.*, 8 of the 20 putatively recessive associations would have been overlooked (Table B.6). These findings underscore the importance of considering nonadditive contributions, not only in increasing the number of identifiable gene-trait associations but also in revealing potentially important recessive effects that might otherwise go undetected.

3.3.7 Deciphering mode of inheritance in binary traits

Modes of inheritances for traits are often categorized as recessive, dominant or additive based on statistical model fit, often by comparing P-values[192,296]. However,

Table 3.4: Overview of the 20 significant quantitative and binary gene-trait associations reported. Overview of significant (Experiment-wise $P < 3.72 \times 10^{-7}$) quantitative and binary gene-trait associations in the analysis. The total number of affected haplotypes is shown, alongside the number of total biallelic carriers of pLoF and damaging missense/protein altering variants. The effect sizes, standard errors and P-values are shown for the additive $[0, 1, 2]$, recessive $[0, 0, 1]$ and nonadditive encoding.

trait	Gene	Haplotypes	Biallelic	Additive analysis			Recessive analysis			Non-additive analysis		
				BETA	SE	P	BETA	SE	P	BETA	SE	P
Other and unspecified dermatitis	*FLG*	50338	1604	0.385	0.014	7.25E-164	0.755	0.030	3.62E-141	-0.6312	0.0324	1.66E-84
Asthma	*FLG*	50338	1604	0.217	0.013	3.74E-67	0.585	0.028	2.35E-94	-0.5035	0.0299	1.01E-63
Atopic dermatitis	*FLG*	50338	1604	0.408	0.025	1.10E-58	0.687	0.050	3.75E-43	-0.5876	0.0557	5.07E-26
Other hearing loss	*GJB2*	7550	30	0.216	0.053	4.43E-05	2.040	0.210	2.64E-22	-2.0284	0.2118	1.02E-21
Malignant neoplasm of colon	*MUTYH*	7697	45	0.278	0.081	5.89E-04	1.842	0.220	6.02E-17	-1.8526	0.2235	1.13E-16
Vasomotor and allergic rhinitis	*FLG*	50338	1604	0.067	0.015	6.82E-06	0.290	0.036	4.34E-16	-0.2664	0.0367	3.67E-13
HDL cholesterol	*BTNL9*	9097	80	-0.060	0.010	2.81E-10	-0.399	0.051	2.73E-15	0.3468	0.0505	6.69E-12
Osteoarthritis of hip	*CHADL*	21622	314	0.083	0.028	3.33E-03	0.614	0.086	1.26E-12	-0.6042	0.0886	9.02E-12
Crohn's disease	*NOD2*	15365	145	0.681	0.079	5.53E-18	1.379	0.194	1.17E-12	-1.3425	0.2055	6.44E-11
Eosinophill count	*FLG*	48833	1565	0.001	0.005	8.67E-01	0.074	0.012	2.88E-09	-0.0778	0.0124	3.64E-10
Deficiency of other B group vitamins	*CLYBL*	21523	314	0.277	0.061	5.79E-06	0.952	0.144	3.78E-11	-0.9424	0.1504	3.69E-10
Forced expiratory volume, Best measure	*FLG*	37852	1215	-0.018	0.004	3.82E-06	-0.075	0.010	4.38E-13	0.0624	0.0104	1.68E-09
Histidine	*HAL*	4162	22	0.574	0.015	3.12E-302	1.130	0.106	1.04E-26	-0.6121	0.1057	7.01E-09
Conductive/sensorineural hearing loss	*GJB2*	7550	30	0.160	0.081	4.93E-02	1.581	0.271	5.62E-09	-1.586	0.2741	7.22E-09
Hand grip strength (right)	*FLG*	50140	1598	0.024	0.003	4.65E-14	-0.022	0.009	1.02E-02	0.047	0.0085	3.49E-08
Hand grip strength (left)	*FLG*	50117	1597	0.028	0.003	1.43E-18	-0.018	0.009	3.98E-02	0.0461	0.0085	5.67E-08
Forced expiratory volume, Best measure	*FUT10*	2488	10	-0.042	0.015	3.64E-03	-0.660	0.115	1.01E-08	0.6173	0.1151	8.20E-08
Cholesterol	*TM6SF2*	11090	71	-0.139	0.009	1.13E-51	-0.425	0.057	8.53E-14	0.297	0.057	1.89E-07
Sitting height	*ABHD15*	10391	57	0.033	0.007	8.12E-06	0.286	0.050	8.72E-09	-0.2563	0.0497	2.53E-07
Glycated haemoglobin (HbA1c)	*RNF123*	16923	199	0.072	0.007	4.81E-22	-0.100	0.034	3.15E-03	0.1739	0.0338	2.77E-07

these classifications often fail to capture the full complexity of genetic architecture, which can involve multifactorial modes of inheritance, which includes partial recessivity. Following the approach outlined in section 3.3.3, I estimated the genetic architectures for each gene-trait association by evaluating the orthogonal additive and nonadditive contributions to the trait across wildtype, monoallelic, and biallelic states (Methods 3.2.6).

The mode of inheritance between variants in *MUTYH* and malignant neoplasm of colon has primarily been reported 'recessive' in the literature[239,297]. Indeed, my analysis suggests that a primarily nonadditive recessive contribution to colon cancer risk, and I infer that biallelic carriers are disproportionately at risk compared to wildtype and heterozygotes (Figure 3.11a). For *GJB2* and hearing loss, I observed a significant nonadditive contribution, consistent with its well-established recessive mode of inheritance[239,298]. The impact on hearing loss remains minimal for wildtype and heterozygous genotypes but increases disproportionately in biallelic carriers, supporting a recessive inheritance model for *GJB2*-related hearing loss (Figure 3.11b).

I also observed disease-gene pairs with more complex genetic architectures. For example, *NOD2* variants have been reported to exhibit a complex relationship with CD risk: Heterozygous carriers of risk alleles have 2- to 4-fold increased risk, while biallelic carriers face a 15- to 40-fold higher risk[299]. This observation suggested a genetic mechanism involving both additive and nonadditive components, rather than a simple recessive model[300], explaining why the OMIM database[239] annotates *NOD2* as having 'multifactorial' inheritance with respect to CD. Indeed, my hybrid model of *NOD2* confirmed this complexity, illustrating a significant additive ($P = 5.53 \times 10^{-18}$) and nonadditive ($P = 6.43 \times 10^{-11}$) contribution to the trait (Figure 3.12). Additionally, analysis of *FLG* revealed diverse inheritance patterns across related allergic and respiratory traits. *FLG* demonstrated predominantly nonadditive effects on asthma and FEV1 (Figure 3.13A-B), with significant differences particularly evident in individuals with variants in the biallelic state. The impacts on asthma risk and FEV1 were negatively correlated. This was consistent with prior studies

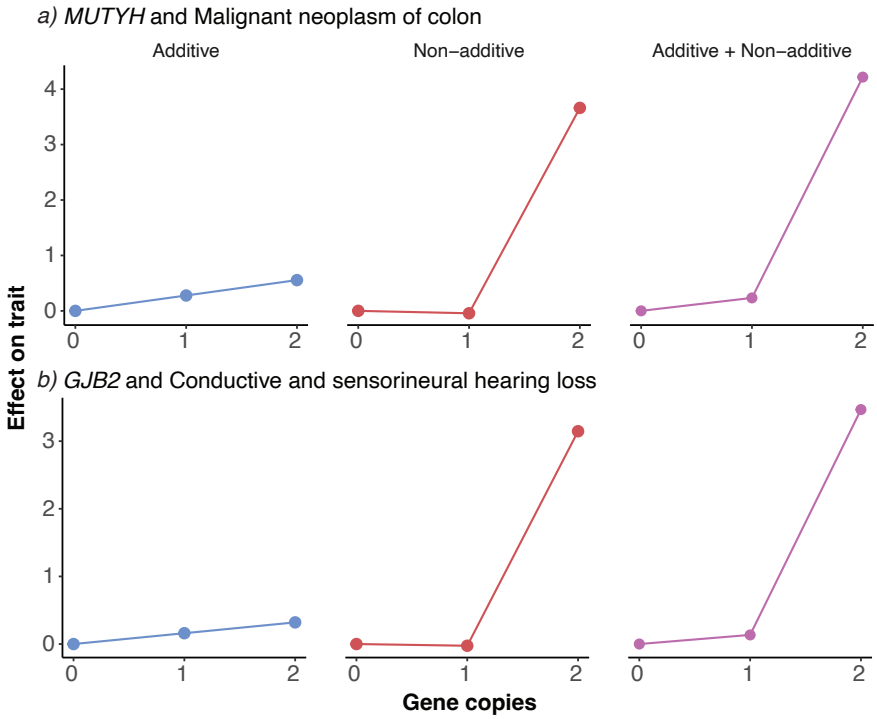


Figure 3.11: Inferred genetic architecture of *MUTYH*-Colorectal cancer and *GJB2* and hearing loss. The genetic architecture of well-known recessive gene-trait associations for genes estimated using the combined additive and nonadditive marginal effects. *MUTYH*-Colorectal cancer and *GJB2* and hearing loss are shown. The x-axis indicates number of gene copies affected, and y-axis indicates effect on the trait for a given number of affected gene copies.

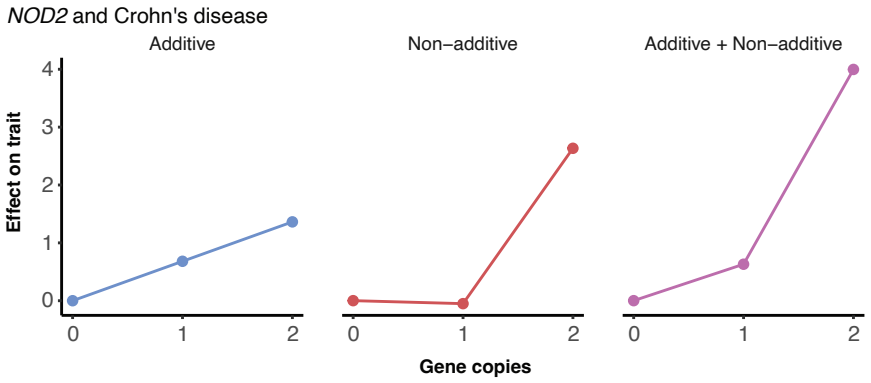


Figure 3.12: Estimated genetic architecture of *NOD2*-CD. The estimated genetic architecture of *NOD2*-CD. The plot depicts both significant additive ($p = 5.53 \times 10^{-18}$) and nonadditive ($p = 6.43 \times 10^{-11}$) components, supporting a complex inheritance model beyond simple recessivity.

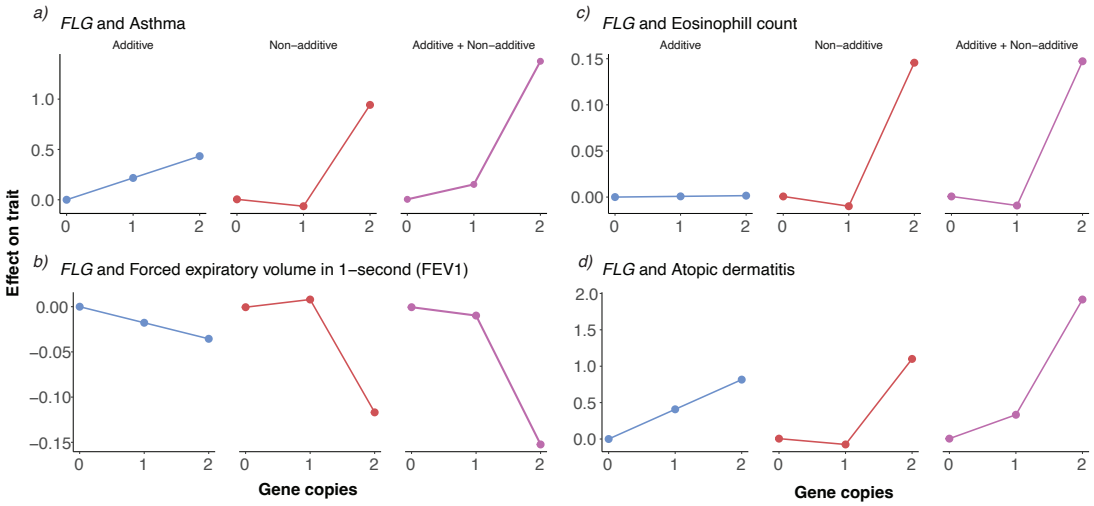


Figure 3.13: Inferred genetic architecture of *FLG* and its pleiotropic effects on four traits. The genetic architecture of various conditions influenced by mono and biallelic variants in *FLG* as determined by the nonadditive and additive contribution to the traits. a) The genetic architecture with Asthma; b) The genetic architecture with FEV1; c) the genetic architecture with Eosinophil count; and d) the genetic architecture of *FLG*-related dermatitis.

which have established clear cause-and-effect relationship between asthma severity and reduced FEV1[301]. Moreover, the eosinophil count, which is a critical biomarker of allergic inflammation, also showed a strong additive ($P = 3.74 \times 10^{-67}$) and nonadditive ($P = 1.01 \times 10^{-63}$) pattern when associated with *FLG* (Figure 3.13C). This finding is particularly relevant for eosinophilic asthma, an endotype of asthma characterized by elevated eosinophil levels[302], adding to growing line of evidence that *FLG* plays a role in modulating immune responses in allergic and respiratory conditions[303,304]. Finally, my analysis of the genetic architecture of *FLG* and dermatitis (Figure 3.13D) suggested an additive ($P = 1.09 \times 10^{-58}$) and nonadditive ($P = 5.06 \times 10^{-26}$) contribution to disease. I also observed an interesting relationship between *RNF123*, an E3-ubiquitin ligase highly expressed in red blood cells[305], and HbA1c. Here, I noted a positive marginal additive ($P = 4.8 \times 10^{-22}$) contribution and a negative nonadditive ($P = 2.77 \times 10^{-7}$) contribution, leading to putative overdominance, where heterozygous carriers had a slight elevation in HbA1c, while biallelic carriers showed an overall reduction (Figure 3.14a). This relationship was primarily driven by damaging missense/protein altering variants, as I identified only

one pLoF *RNF123* biallelic carrier in the tested population. Additionally, I observed a putatively completely recessive (additive $P = 0.004$, nonadditive $P = 8.19 \times 10^{-8}$) relationship between biallelic variants in *FUT10* and FEV1: only individuals in the biallelic state appeared to experience reduction in FEV1 (Figure 3.14b). This relationship remains unreported, and the literature on the gene is limited. Lastly, biallelic variants in *ABHD15*, a gene known to regulate adipocyte lipolysis[306], were associated (additive $P = 8.12 \times 10^{-6}$, nonadditive $P = 2.53 \times 10^{-7}$) with an increase in sitting height (Figure 3.14c), which also remain unreported. Notably, standing height was not included in this analysis, and I was therefore unable to check whether biallelic variants in *ABHD15* were also associated with standing height.

3.3.8 Characterizing the spectrum of recessiveness

In a recent paper[192], Heyne *et al.* identified several variants with disease effects in both heterozygous and homozygous states in known Mendelian disease genes, concluding that the usual labels of recessive and dominant are insufficient to describe their modes of inheritance. They highlighted the need for a nomenclature that integrates Mendelian and complex genetic effects on diseases. In response to this need, I propose a method that contrasts the marginal biallelic effects from additive and nonadditive models in traits, offering a potential framework to evaluate multifactorial modes of inheritance. This approach has three key features that address the limitations identified by Heyne *et al.*. First, it captures a larger spectrum of genetic effects, from purely additive to purely recessive, including intermediate states such as partial recessivity, while also allowing for overdominance. Second, it avoids the potential misclassification that can occur when simply comparing P-values from different models, as demonstrated in Section 3.3.3 and 3.3.7. Third, it provides a visual representation of the relationship between additive and nonadditive effects, allowing for intuitive interpretation of potentially complex inheritance patterns.

To put this approach into practice, I compared 10 significant nonadditive gene-trait reported in previous sections, with 10 randomly sampled significant additive gene-trait associations (Figure 3.15). Purely additive traits clustered along the $y = 0$

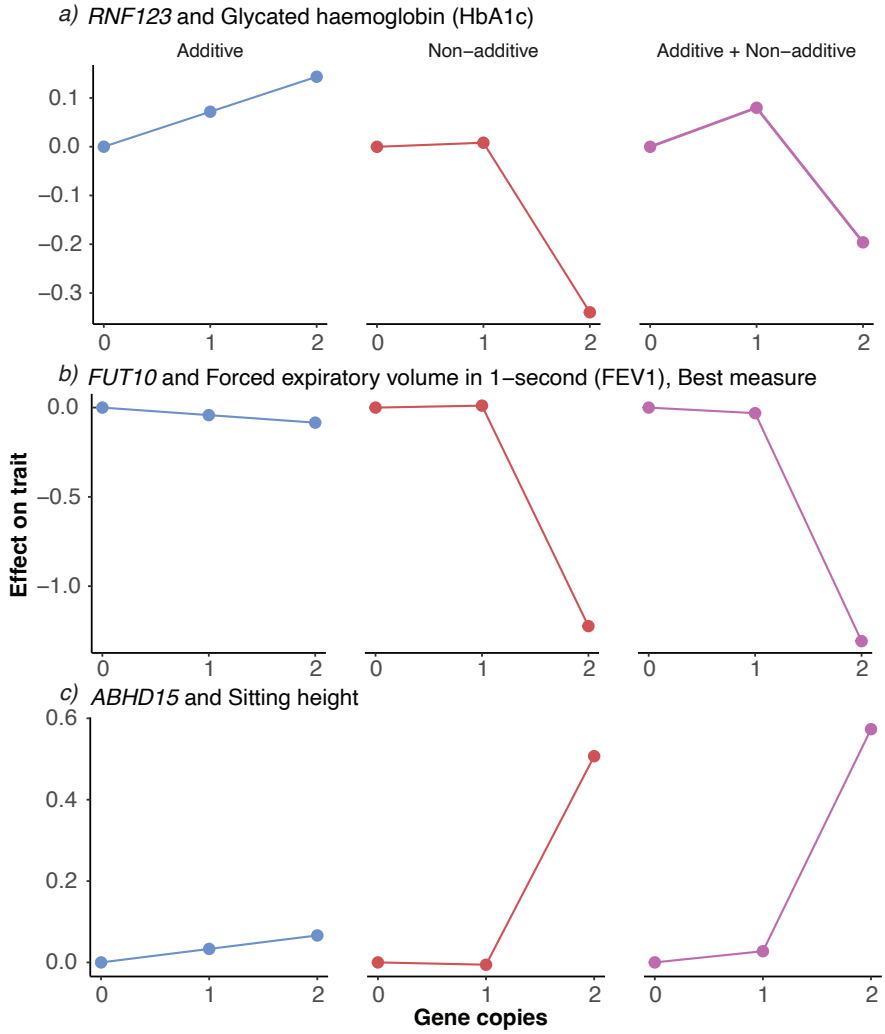


Figure 3.14: Inferred genetic architecture for *RNF123*-HbA1c and *ABHD15*-Sitting height with significant nonadditive contributions. The genetic architecture of two potentially novel gene-trait associations for genes: *RNF123*-HbA1c (first row), *FUT10*-FEV1 (2nd row) and *ABHD15*-Sitting height (third row). The x-axis indicates number of gene copies affected, and y-axis indicates effect on the trait for a given number of affected gene copies.

line, including previously established gene-traits such as *RPL3L*-Atrial fibrillation[307] and *C4B*-reduced Asthma risk[308]. In contrast, well-established Mendelian traits, such as *GJB2*-Hearing loss and *MUTYH*-colorectal cancer, clustered closer to the 'recessive region' (high x, low y values). Other traits, including *FLG*-Asthma, *FLG*-Dermatitis, and *NOD2*-CD, exhibited substantial additive contributions alongside their nonadditive effects, suggesting more complex inheritance patterns, such as partial recessivity. By considering gene-traits in this framework, one may achieve a

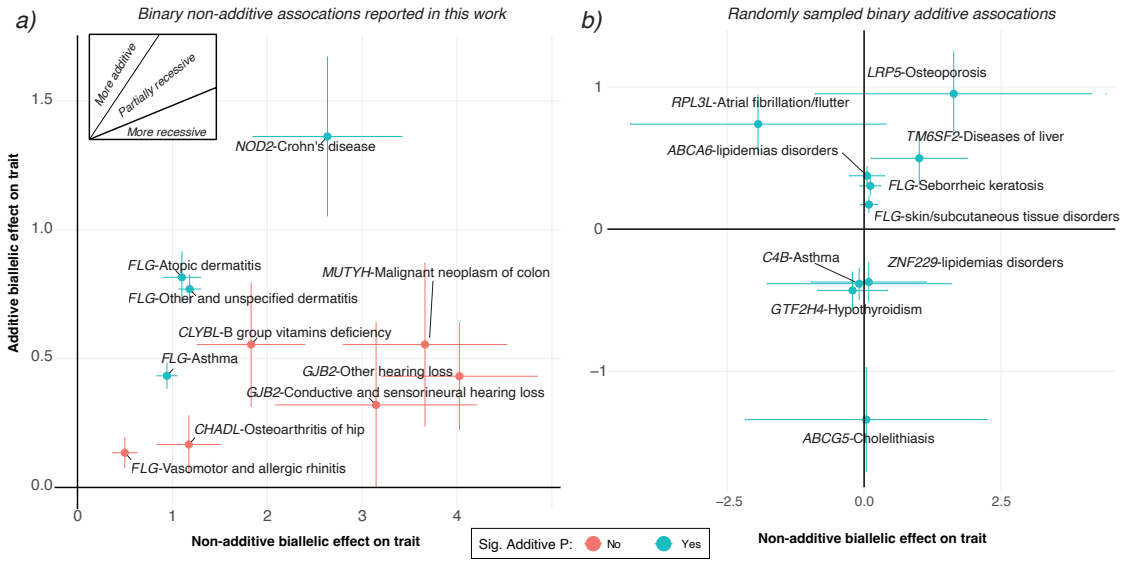


Figure 3.15: Spectrum of genetic effects across binary traits: Significant nonadditive associations versus randomly sampled additive associations. Comparison of additive (y-axis) and nonadditive (x-axis) effects for selected and randomly sampled gene-traits. (a) Comparison of additive and nonadditive effects for 10 binary gene-trait pairs with significant nonadditive contributions. The top-left sketch illustrates how to interpret the effects: Gene-traits closer to the x-axis suggest a recessive mode of inheritance; those closer to $y = x$ indicate partial recessivity; and those near $x = 0$ imply an additive mode of inheritance. (b) Comparison of additive and nonadditive effects in 10 randomly selected gene-traits with significant additive but nonsignificant nonadditive contributions. 95% CI intervals are shown for both additive (vertical) and nonadditive (horizontal) effects. Note the different axis limits within and between plots.

better representation of the underlying genetic architecture, potentially revealing a continuum of effects rather than a simple binary classification.

3.3.9 The burden of nonadditive effects

In most large-scale rare variant analyses, phasing haplotypes and modeling gene-traits as individual pseudo-variants is not feasible. Instead, additive gene-based rare variant burden models are employed to enhance statistical power, using tools such as `SAIGE-GENE+`[86] or `REGENIE`[85,86]. In theory, a similar approach could be applied to study nonadditive effects, where the nonadditive contributions of multiple rare variants are aggregated across genes. Furthermore, such a method would facilitate nonadditive analyses within familiar frameworks, allowing them to be conducted alongside traditional additive analyses.

To demonstrate the application of such a framework, I applied the nonadditive encoding to individual variants, rather than gene-based pseudo-variants, and modeled them using `SAIGE-GENE+` burden testing (Methods 3.2.8). This framework may sacrifice the information encoded in the phase and is potentially less powerful than the pseudo-variant approach. That is, unless the model assumed by this framework is true. To evaluate this trade-off, I focused on the 19 unique (20 total associations) traits that had previously demonstrated at least one significant non-additive association in my pseudo-variant analysis. I assessed the combined burden of pLoF and damaging missense/protein-altering variant categories, mirroring my previous analysis. `SAIGE-GENE+` deals with ultra-rare variants (MAC$\leq$ 10), by combining them into pseudo-variants. At the time of writing this thesis, I was not yet able to adapt this framework to support the nonadditive encoding, and therefore excluded such ultra-rare variants. I also restricted the analysis to the same putatively well-powered gene-trait pairs examined in the pseudo-variant approach, resulting in 2745 total genes assessed.

The analysis exhibited slightly deflated type-1 errors ($\lambda_{GC} = 0.891$) across both binary and quantitative gene-trait associations (Figure 3.16a). Only 10 of the 20 (50%) gene-trait associations that were significant in the pseudo-variant analysis reached significance in the nonadditive burden analysis (Experiment-wise $P < 3.72 \times 10^{-7}$). To elucidate the factors driving this discrepancy, I conducted a direct comparison of P-values between the two analyses (Figure 3.16b). I found that 18 of 20 (90%) associations resulted in lower P-values in the pseudo-variant analysis. This observation suggested two potential explanations: 1) The phased information contributed significantly to detecting these associations; and 2) Ultra-rare variants (MAC$\leq$ 10), which were included in the pseudo-variant analysis but excluded from the variant-based analysis, could be driving some of these gene-trait associations. The validity of these hypotheses was supported by the relationship between deviation from the $x = y$ line and the proportion of compound heterozygous (CH) identifications illustrated in the figure. For instance, the *HAL*-Histidine association, significant only in the pseudo-variant analysis, showed a high

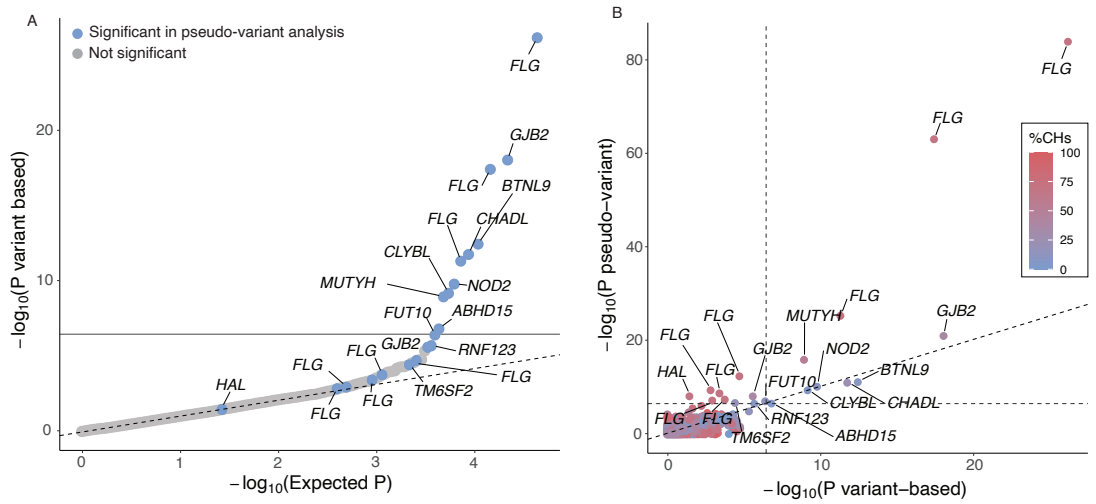


Figure 3.16: Comparison of variant-based burden analysis and pseudo-variant analysis. a) QQ-plot of observed versus expected P-values in the variant-based burden analysis of nonadditive effects. The horizontal line indicates the significance threshold (Experiment-wise $P < 3.72 \times 10^{-7}$). Gene-trait associations significant in the original pseudo-variant analysis are highlighted in blue and labelled. b) Scatter plot comparing P-values from the pseudo-variant analysis (y-axis) and the variant-based burden analysis (x-axis). Dashed lines represent the significance threshold ($P < 3.72 \times 10^{-7}$). Points are colored by the proportion of biallelic variants identified as CH. Significant gene-trait associations from the pseudo-variant analysis are labelled. The line $x = y$ is shown in both plots for reference.

proportion of CH carriers (66%, with 24 CH and 12 homozygotes). In contrast, the *NOD2*-CD association, which was significant in both analyses, had a much lower proportion of CH carriers (10%, with 15 CHs and 130 homozygotes). This demonstrates that the nonadditive encoding can be successfully applied within a standard gene-based burden framework to systematically search for nonadditive effects. This nonadditive burden approach complements existing analyses that lack phased data. The observed differences in significance and the relationship with CH proportions underscore the potential importance of phased information and rare variants in detecting nonadditive gene-trait associations.

3.3.10 Identifying recessive selection

Motivation

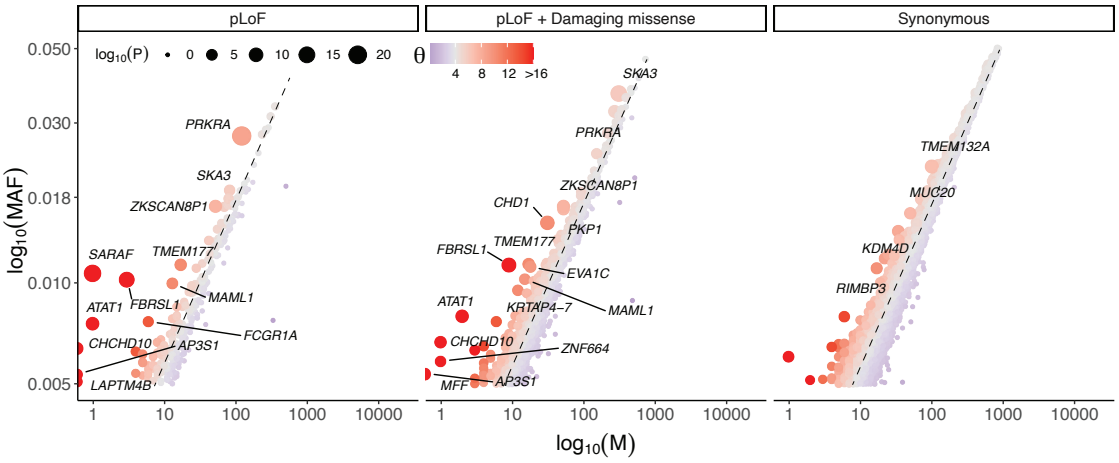
Nonadditive effects can only be assessed if a sufficient number of biallelic carriers are present. Finally, in this section I now turn my attention to a complementary analysis: leveraging the *absence* of biallelic variants to detect genes potentially subject to purifying recessive selection. Here, I examine genes that exhibit appreciable allele frequencies yet paradoxically have few or no biallelic (CH or homozygous) variants in the population, which would be expected under HWE. This discrepancy serves as a potential indicator of purifying selection acting on deleterious biallelic variants, offering insights into the evolutionary forces shaping genetic diversity.

Biallelic depletion in UKBB

I set out to investigate the depletion of biallelic (homozygous and CH) variants in the UKBB dataset, which could indicate selection against recessive alleles. I confined my analysis to 339,041 unrelated individuals of European ancestry. For each gene, I quantified the number of individuals with biallelic variants (M) and the total number of mutated copies (K). I then conducted gene-by-gene tests using the HWE exact test (Methods 3.2.12). I report mid P-values[182] that represent the evidence for biallelic variant depletion compared to expectations under HWE. I applied this method to three variant annotation masks: high-impact variants (pLoF), moderate-impact variants (pLoF and damaging missense/protein altering), and low-impact variants (synonymous). To facilitate comparison with a recent study on recessive selection by Oddsson *et al.* involving 1.52 million individuals[309], I restricted the analysis to variants with MAF$<$ 0.02.

Central to the biallelic depletion analysis is the θ parameter, which quantifies the degree of HWE disequilibrium. Under HWE, $\theta = 4$, with $\theta > 4$ indicating heterozygous excess and $\theta < 4$ indicating homozygous or CH excess. I conducted power analysis (Figure B.6) to determine an appropriate MAF threshold given 339,041 samples, following an investigation into empirically observable θ parameters (Figure B.5). This analysis revealed that MAF > 0.005 provided approximately

Figure 3.17: Overview of genes that are depleted of compound heterozygous and homozygous variation. Comparison of M (number of biallelic carriers) versus MAF, with points colored by θ, which quantifies the degree of disequilibrium. Results are shown for three variant categories: pLoF, pLoF and damaging missense/protein altering, and synonymous variants. The dashed line at $\theta = 4$ indicates equilibrium under HWE. Redder hues signify heterozygous excess ($\theta > 4$), while bluer hues indicate homozygous excess ($\theta < 4$). For visualization purposes, θ values are truncated at 16. Point size reflects the one-sided gene-level HWE mid P-value, testing for heterozygous excess. Genes with FDR$<$ 0.1 are labeled accordingly.



80% power to detect heterozygous excess at $\alpha = 0.05$ when $\theta = 16$ (Methods 3.2.12), leading to the imposition of this cutoff. I proceeded to calculate gene-level HWE mid-P values for 6927 unique genes across pLoF (282 genes), pLoF and damaging missense/protein altering (1124 genes) and synonymous variants (6502 genes), imposing a false discovery rate (FDR)$<$ 0.1 cutoff for each annotation category to account for multiple testing.

I observed λ_{GC} of 0.93, 0.75 and 0.68 for pLoF, pLoF+damaging missense/protein altering, and synonymous biallelic variants, respectively. I searched for gene-level recessive selection and found 19 significant deviations (12 pLoF, 15 pLoF+damaging missense/protein altering and 4 synonymous) across 13 unique genes at FDR$<$ 0.1 (Figure 3.18-3.17). Had I imposed a conservative Bonferroni cutoff ($P = \frac{0.05}{6927} \approx 7.21 \times 10^{-}6$), I would have identified 7 genes with a significant depletion of biallelic pLoF variants. In contrast, Oddsson *et al.* explicitly report they would have identified 10 genes with significant deficit of homozygosity ($P < \frac{0.05}{1258} = 4 \times 10^{-5}$) assuming a Poisson distribution[309]. Instead, they report 19 genes with a

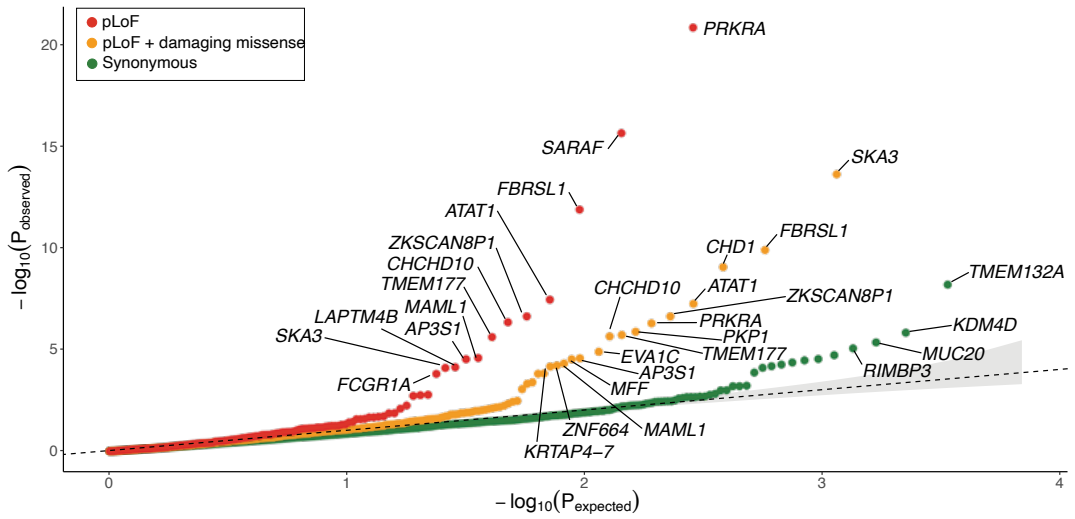


Figure 3.18: QQ-plot for scan of recessive selection in 339,041 individuals across variant annotation masks. QQ-plot for gene-level HWE mid P-value stratified by variant annotation mask: pLoF (green), pLoF and damaging missense/protein altering (yellow), and synonymous (green). Genes that passes the FDR<0.1 cutoff are labeled on the plot. Deviations from the null are potentially caused by recessive selection.

'significant' deficit of homozygosity defined as less than 10% homozygotes expected under HWE. Of these, I was able to replicate 2 at nominal significance, namely *CCDC59* ($P = 0.028$) and *PNKP* ($P = 0.023$). Among the 12 genes depleted of biallelic pLoF variants in my analysis, two had previously been annotated as recessive[227]: *PRKA* (pLoF $P = 2.02 \times 10^{-21}$), which had been associated with early-onset generalized dystonia under a recessive mode of inheritance[310,311], and *PKP1* (pLoF and damaging missense/protein altering $P = 1.73 \times 10^{-6}$) which had been associated with ectodermal dysplasia/skin fragility syndrome under a recessive mode of inheritance[239]. Of the 282 genes tested for depletion of biallelic variants, 11 were considered essential for cell growth, of which 2 were among genes significantly (FDR<0.1) depleted for biallelic pLoF variants. Comparing genes significantly depleted of biallelic variation to those that were not, I did not observe a significant enrichment among 'essential' genes for this gene set (Two-sided Fisher's exact test $P = 0.074$). Collectively, this analysis demonstrates the utility of large-scale genomic data in identifying genes potentially subject to nonadditive effects acting on recessive alleles, contributing to our understanding of recessive selection.

3.4 Discussion

In this chapter, I examined rare additive and nonadditive effects (MAF $< 5\%$) using two distinct datasets within UKBB: 42,217 individuals with ~ 3000 proteomic measurements each, and 399,943 individuals across a wide range of disease endpoints and quantitative traits, including anthropometrics, biomarkers, and spirometry. These datasets enable me to investigate nonadditive effects across multiple biological dimensions, from molecular interactions to broader physiological outcomes. By making use of genetic phase information, I increase the number of identifiable biallelic pLoF carriers from 47,642 homozygous carriers to 58,399 homozygous and CH carriers (a 22% increase).

I first investigated additive effects in the proteomic dataset. This examination revealed significant impacts of functional variant annotations on protein abundance, with varying effects across different variant types. pLoF variants demonstrated the most substantial influence, resulting in significant changes in abundance for 44.8% of 1,103 tested associations ($P < 8.33 \times 10^{-10}$). A smaller proportion of LC PTV variants (35.2% of 91 associations), primarily end-truncating frameshift variants in the penultimate exon, also affected protein abundance. This aligns with the understanding that PTV variants in the final exon often escape NMD, resulting in truncated but potentially functional proteins detectable by the Olink assay. In contrast, pLoF variants are more frequently subject to NMD, possibly explaining their proportionally larger impact on protein abundance. Damaging missense variants demonstrated less consistent effects, decreasing abundance in only 21.9% of 2,230 associations. This does not necessarily imply a lack of impact on protein functionality, as variants in catalytic or allosteric sites may exert deleterious effects through other mechanisms while maintaining normal abundance levels. 7.8% of 2,729 synonymous variant associations induced significant changes in protein abundance, which could result from LD tagging of nearby deleterious variation, or by creating non-optimal codons that reduce transnational efficiency or mRNA stability[312].

Following this, I used an explicit nonadditive allelic recoding to perform association analysis gene-by-gene in the proteomic dataset. Firstly, I found that putatively

damaging biallelic variants in *GIMAP8* were associated with an additive decrease of GIMAP8 and simultaneously a nonadditive decrease of GIMAP7, suggesting a complex regulatory network within the GIMAP protein family which is still under active investigation[313]. Such cross-regulation could suggest that the presence of GIMAP8 is important for maintaining normal levels of GIMAP7. In this framework, the nonadditive decrease in GIMAP7 levels I observed when both *GIMAP8* alleles are perturbed could indicate a threshold effect, where a critical amount of GIMAP8 is required to maintain normal GIMAP7 protein levels.

Secondly, my proteomic analysis revealed compelling examples of pleiotropic nonadditive effects, particularly in the context of skin physiology and related disorders. For instance, putatively deleterious (pLoF and damaging missense/protein altering) biallelic variants in *FLG* were associated with a nonlinear increase in the abundance of GPR15L, KLK8, and PYDC1. These proteins are known to play critical roles in regulating inflammation, skin barrier function, and cellular differentiation[314–316]. While additive pleiotropic effects related to *FLG* (including for the three aforementioned proteins) have been previously demonstrated[128], the nonadditive relationship between *FLG* variants and these proteins suggests the existence of a potential threshold effect involving complete biallelic *FLG* inactivation. This observation aligns with previous research indicating that while mono-allelic *FLG* variant carriers are predisposed to dermatitis, biallelic carriers exhibit a disproportionately increased risk for asthma[245,317]. These nonadditive associations may provide insight into the molecular mechanisms underlying the gene-trait relationship between *FLG* and asthma. Functional studies support the role of KLK8 and GPR15L in inflammation: KLK8 actively converts the chemokine CCL14 into an active state which promotes inflammation[318], while GPR15L, another chemokine, is required for immune homeostasis[316,319]. One can hypothesize that GPR15L, KLK8, and PYDC1, which show a nonlinear response to deleterious *FLG* variants, could play a role in triggering asthma onset. Further investigation into these proteins may open new avenues for understanding asthma pathogenesis and potentially developing targeted therapies.

Thirdly, I identified a putative overdominant relationship between pLoF variants in *HMCN2* gene and its corresponding protein HMCN2. Little is known about the function of this protein, but it is believed to be involved in extracellular matrix architecture, organisation and guiding morphology of specialized cells[320]. Single-gene overdominant (or underdominant) effects attributable to pLoF variants are rarely documented in humans[321,322], and have never been demonstrated for rare putatively LoF variants in large-scale studies[323]. The validity of this observation requires further investigation. Replication studies are necessary to confirm whether this relationship is a genuine biological phenomenon or a potential batch effect or artifact of the analysis. If confirmed, the functional implications and potential evolutionary significance of reduced HMCN2 expression would need to be elucidated. The literature is limited and studies on *Hmcn2* mice have not revealed an overt phenotype associated with knockouts[320], which may or may not be indicative of its role in humans.

Following the investigation of additive and nonadditive effects in the proteomic dataset, the analysis was extended to encompass a broader spectrum of phenotypes. This expansion facilitated a more comprehensive examination of the influence of rare genetic variants across a wide range of physiological and pathological outcomes available for most UKBB participants. To this end, a phenome-wide analysis was conducted utilizing a substantially larger cohort of 399,412 individuals of European ancestry. Of the 20 significant (Experiment-wise $P < 3.72 \times 10^{-7}$) gene-trait associations discovered (10 binary and 10 quantitative), at least 14 have been previously documented in the literature. These associations span a range of conditions including colorectal cancer, hearing loss, asthma, dermatitis, FEV1, and CD, all of which demonstrated significant nonadditive contributions. Intriguingly, for conditions exhibiting both significant additive and nonadditive contributions, I observed more nuanced inheritance patterns than a purely recessive model would have suggested.

For example, the relationship between *NOD2* risk alleles and CD has been the subject of extensive research, with numerous publications advocating for a

recessive mode of inheritance[287,324–326]. However, a completely recessive model fails to fully account for an important clinical observation: heterozygous carriers of *NOD2* risk alleles exhibit a 2-4 fold elevated risk of CD compared to individuals with wild-type alleles[327–329]. This increased risk in heterozygotes suggests that a purely recessive model may be an oversimplification of the underlying genetic mechanism. A partially recessive model, incorporating both additive and strong nonadditive components, appears to more accurately capture the complex risk profile associated with *NOD2* variants. The hybrid model proposed here can account for the modest risk increase in heterozygous carriers while also explaining the more pronounced risk in homozygous or CH individuals.

Similarly, my analyses suggested both additive and nonadditive components contributing to trait variance for biallelic carriers of damaging *FLG* variants in asthma and dermatitis, revealing a complex relationship to multiple related traits. For dermatitis, my findings align with literature reports of a nonlinear relationship: studies have shown that individuals with one affected *FLG* copy experience dermatitis at 26.2% incidence, compared to 22.9% in wild-type individuals, while those with two affected copies show a dramatically higher incidence of 88.9%[330]. In contrast, *FLG*-related asthma incidence appears to have a threshold effect, with limited epidemiological evidence of increased incidence or even mild symptoms in heterozygous carriers but significant incidence in biallelic carriers. This complexity is further illustrated by my finding where eosinophil count—a critical marker of allergic inflammation—showed a strong nonadditive relationship with *FLG* haplotype inactivation (Figure 3.13C). Consistent with previous speculation[303], this finding could indicate that complete *FLG* inactivation triggers a specific endotype of asthma, called eosinophilic asthma, which is distinct from allergic asthma, and is driven by eosinophil differentiation, maturation, and migration[302,331]. My analysis also revealed an association between *FLG* variants and hand grip strength, which may be correlated with or confounded by FEV1, a key measure of lung function. Recent research by Hesselberg *et al.* has shown significant associations between hand grip strength and effort-dependent lung function measures, particularly FEV1,

in both asthmatic and non-asthmatic adolescents[332]. Overall, my results underscore the wide range of *FLG*-related phenotypes, and how further investigation is needed to unravel its complexities.

I also report three potentially novel associations, all characterized by strong nonadditive contributions to trait variance, which may explain why they have not been previously identified. First, I observed a putative overdominant relationship between biallelic variants in *RNF123*, which is highly expressed in blood[305], and HbA1c, driven primarily by damaging missense/protein altering variants. Heterozygous carriers exhibit slightly elevated protein levels compared to wildtypes, while biallelic carriers show significantly reduced levels. The increased expression in heterozygotes might represent a compensatory response, suggesting a potential feedback mechanism to maintain adequate levels of this protein. Secondly, I found a potentially novel association between putatively damaging biallelic variants in *ABHD15* and sitting height (nonadditive $P = 2.53 \times 10^{-7}$). Although *ABHD15* is well-studied in adipogenesis, its mechanism relating to sitting height remains unclear. Sitting height, which reflects body proportions rather than overall stature, is often used to examine skeletal development and growth disorders[333,334]. It is possible that *ABHD15* may be involved in vertebral growth or trunk adiposity, thus affecting sitting height. However, further analysis of height and adiposity phenotypes will be required to confirm this. Thirdly, I find that biallelic variants in *FUT10* significantly reduce FEV1. This finding corroborates a previous study that associated common variants near *FUT10* with subclinical interstitial lung disease[335], supporting the functional impact of the gene on lung function. While this suggests a potential role for *FUT10* in respiratory function, additional studies are necessary to validate and understand the underlying mechanisms.

I demonstrated that 10 out of 20 reported findings in 399,412 Europeans using a pseudo-variant approach (combining CH and homozygotes), could be replicated within a standard gene-based burden framework where single variants are jointly modeled across genes. While it is feasible to conduct nonadditive burden analysis, I show that this approach may result in reduced statistical power, partly due to the

loss of information in phased genotypes. This suggests that modeling approaches incorporating phased information may more accurately reflect biological reality for the interrogated gene-traits. Nevertheless, the nonadditive burden test offers a significant advantage: it can be readily implemented within a standard gene-level framework without requiring phased data. This is achieved through simple recoding of alleles, which can then be utilized in set-based analysis tools such as `SAIGE`[86] or `REGENIE`[85]. Consequently, the nonadditive burden test becomes a potentially valuable option for researchers, as it can easily be evaluated in parallel to an additive gene-based burden model.

Throughout this chapter, I have fitted additive and nonadditive models in parallel. These are both 1-DF models, capturing the orthogonal additive and nonadditive components of heritability respectively. Either of these models will perform better depending on the additive or nonadditive nature of the tested genetic architecture. However, in the presence of partial recessivity or more complex nonadditive genetic architectures, a more flexible 2-DF model that incorporates both the additive and nonadditive components could theoretically improve power to detect gene-trait associations. Because the two variance components are orthogonal, multicollinearity is not a concern when combining them in a single model. This joint model allows for the simultaneous estimation of both additive and nonadditive effects, theoretically increasing power to detect associations, and could initially be implemented through likelihood ratio testing. While 2-DF models offer greater flexibility, it is important to note that this comes at the cost of an additional DF, which may reduce power if the true genetic architecture is simply canonical additive or recessive. Nevertheless, future work should explore the application of such 2-DF models in large-scale GWAS, employing techniques like SPA for accurate P-value calculation, especially in the context of rare variants or traits with imbalance between cases and controls.

Finally, I show that nonadditive effects can also be detected through absence of biallelic variants. I implement a method based on the Levene-Haldene distribution to detect recessive selection by identifying a deficit of biallelic variants compared to expectations under HWE. This approach leverages compound heterozygosity, which

should, in principle, demonstrate increased statistical power over models that only consider homozygosity. However, to validate this assertion, explicit modeling of homozygosity deficit and comparison with the current model would be necessary. If this hypothesis holds true, future analyses of recessive selection should consider the benefits of exome-wide haplotype estimation. Further research is required to quantify the precise advantages of incorporating compound heterozygosity in this context and to determine whether it indeed enhances statistical power as theorized.

3.4.1 Limitations

This work has explored the additive and nonadditive contributions within a single locus to trait variance. There are several limitations to this work. This analysis has focused on rare (MAF$< 5\%$) genetic variation at a locus and does not consider nonadditive effects in common variation, as this has been examined elsewhere[179]. Future analysis should consider bottlenecked populations such as Finnish[105] and consanguineous populations[193], as these populations may have biallelic variants with large nonadditive effects at higher frequencies. Furthermore, this analysis has been conducted in Europeans, and future analyses should work towards analyzing nonadditive effects in ancestrally diverse cohorts. Cohorts with a different ancestral background may exhibit allele frequency differences, allowing us to recapitulate gene-trait associations not observable in European populations.

One assumption for the pseudo-variant analysis demonstrated here is that variants within a gene act in the same direction and with the same mode of inheritance. However, it is possible that two or more variants within the same gene operate differently: some additively and others nonadditively with respect to the trait. Combining such variants as pseudo-variants may obscure their individual effects and potentially mask important biological insights. This assumption therefore warrants careful scrutiny, particularly when combining multiple variant annotation masks, such as pLoF and damaging missense/protein-altering variants. As demonstrated earlier in this chapter, the latter is more likely to result in gain-of-function effects, introducing potential heterogeneity in the direction of effect. Similar caution is

necessary for the nonadditive rare variant burden analysis, where an assumption is that variants act uniformly in direction and magnitude of effect. This assumption may not always hold true across all variants within a gene for the same reasons as discussed here. It will be important to replicate these effects in external datasets, whether employing gene-based burden analysis or the pseudo-variant framework. Ideally, analysis at the level of individual variants would provide the most granular and potentially informative results; however, this would be at the cost reduced power compared to gene-based analysis.

4

Detection of biallelic recessive effects through meta-analysis

Contents

4.1 Introduction 136
4.2 Methods 137
4.2.1 Study datasets 137
4.2.2 General quality control and phasing guidelines for biobanks 139
4.2.3 UK Biobank quality control and phasing 143
4.2.4 Genomics England quality control 143
4.2.5 Biallelic variant identification 146
4.2.6 Counting expected and observed biallelic variants 147
4.2.7 Phenotype curation 147
4.2.8 Association analysis 149
4.2.9 Meta-analysis 149
4.2.10 Cauchy Combination of P-values across annotation masks 150
4.2.11 Power comparison between meta-analysis and UK Biobank 151
4.2.12 Evaluation of sample overlap across cohorts 152
4.3 Results 153
4.3.1 Patterns of disease across cohorts 153
4.3.2 Characterising biallelic variation across biobanks 161
4.3.3 Federated rare variant meta-analysis 167
4.3.4 Identification of recessive effects through meta-analysis . 176
4.3.5 Meta-analysis improves power for select gene-trait associations 183
4.4 Discussion 186
4.4.1 Limitations 188

4.1 Introduction

The scale of genetic studies is crucial for associating rare variants with complex traits and diseases. Larger sample sizes are required to find rare variants, and to achieve the statistical power needed to robustly associate them with a trait. This is particularly true for detecting recessive effects, which require mutated alleles in both gene copies, and are consequently more elusive than additive effects involving only a single mutated gene copy. Larger datasets combining genetic information with deep phenotyping provide the statistical power needed to identify both additive and recessive genetic associations. This recognition has led to the creation of many biobanks worldwide, collectively profiling millions of individuals. Ideally, analyzing these biobanks together could provide valuable insights. However, such initiatives face challenges related to privacy, security, and formatting issues. Consequently, with two recent exception[124,125], most rare variant analyses have been restricted to single biobanks, limiting the potential for broader discoveries.

The BRaVa consortium represents a 'meta biobank' initiative, aiming to integrate genetic and phenotypic data from 5.7 million individuals across diverse ancestries. This increased sample size is expected to enhance the power to detect both additive and recessive effects (Figure 4.14). By phasing each biobank and incorporating compound heterozygosity, biobanks can further improve detection of rare recessive effects by increasing the number of available biallelic deleterious variants for association analysis. Beyond GWAS, there is value in identifying and characterising even a single individual with biallelic deleterious variants, as this can aid drug discovery by anticipating safety concerns[145,146,188,336]. A prime example is the case of a 'healthy' individual with biallelic LoF in the *HAO1* gene, which supported the safety of HAO1 inhibition as a chronic therapy for primary hyperoxaluria type 1 and provided crucial *in vivo* mechanistic understanding[337], and as of 2024 at least one HAOI-inhibitors is now at late-stage clinical trial[338,339].

In this chapter, I leverage the lessons learned about compound heterozygosity in chapter 2, and analyzing nonadditive effects in chapter 3, to combine and interrogate data from 4 biobanks in the BRaVa consortium: UKBB, Genomics England (GEL), Genes and Health (GnH), and Biobank Japan (BBJ). Together with All of Us, BioMe and Penn Medicine Biobank, these biobanks form the spearhead of the BRaVa recessive subgroup, with the ultimate goal to collectively analyze the role of rare biallelic recessive effects in human health and disease. To this end, we conduct biobank-wide phasing to explore compound heterozygosity across diverse ancestries (European, South Asian, Central Asian, East Asian, African, and Admixed American), and perform additive, recessive, and nonadditive federated gene-based recessive testing to investigate the phenotypic consequences of biallelic gene inactivation. The work presented in this chapter represents the pilot recessive analysis of 42 traits, demonstrating the utility of phasing multiple large-scale cohorts to uncover phenome-wide recessive associations and highlights the power gained through global collaboration.

4.2 Methods

This thesis chapter is part of the BRaVa recessive consortium, a global collaborative initiative led by Dr. Duncan Palmer. To date, the study incorporates data from UKBB, GEL, GnH, and BBJ. Analyses in BioMe, All of Us and Penn Medicine biobank are ongoing. I performed all analyses of UKBB and GEL. Key contributors from other institutions include Georgios Kalantzis, David Van-Heel, Hilary Martin from GnH; and Kyuto Sonehara, Shinichi Namba, Yukinori Okada from BBJ, who kindly used tools developed in my thesis, and shared resultant summary statistics. All downstream analysis, including meta-analysis, was performed by myself.

4.2.1 Study datasets

Genes and Health

The GnH cohort is a community-based study focusing on British Bangladeshi and British Pakistani communities in east London[252,340]. With over 39,148 volunteers

enrolled and a target of 100,000, GnH combines exome sequencing and genotyping array data with participants' EHR from both primary and secondary care[252]. This comprehensive dataset enables GnH to actively address the underrepresentation of South Asian populations in genetics research[252,341,342]. The study examines genetic influences on disease, including effects of autozygosity resulting from higher degrees of parental relatedness[343,344]. GnH incorporates recall-by-genotype[337,345], allowing for follow-up and further investigation of individuals with genotypes that are extremely rare in outbred populations of millions of individuals[337,346].

Genomics England

Genomics England[347,348], launched by the UK government in 2013 as part of the 100,000 Genomes Project[349], is a large-scale biobank initiative within the UK NHS. It focuses on collecting WGS and clinical data from NHS patients with rare diseases, cancers, and infections[350,351]. It also recruits relatives for family-based study of rare disease[347], and the dataset therefore has many families: 12,723 individuals in duos, 45,617 in trios, and the remaining being unrelated individuals. While being named the '100,000' genomes project, the project spans 88,492 individuals, containing 72,869 germline genomes from rare disease program, and 15,623 germline and somatic genomes from cancer program as of December 2023. The GEL cohort has details on clinical phenotyping, and health linkage with NHS EHRs allowing for retrospective genetic studies, which are leveraged in this chapter. As of 2023, the project has expanded beyond its initial 100,000 genomes target, continuing to collect and analyze genomic and health data to improve rare disease diagnosis and personalized medicine within the NHS.

Biobank Japan

BioBank Japan[352–355] is a large-scale, multi-institutional hospital-based registry launched in 2003 to facilitate personalized medicine through genomic research[356]. BBJ has collected DNA, serum samples, and clinical information from approximately 270,000 participants diagnosed with one or more of 47 common diseases. Initial recruitment involved 66 hospitals across 12 cooperating medical institutions in

Japan, only enrolling Japanese participants based on set of selected diagnoses that has been deemed to be of interest[356]. The project collects annual clinical information and serum samples, which are linked to electronic medical record (EMR). As of 2023, BBJ has genotyped approximately 179,000 participants[355] and has exome-sequencing data for 11,462 patients, which are the focus of this chapter. BBJ aims to identify disease-related genes, drug response factors, and support the implementation of personalized medicine[356].

UK Biobank

UKBB data was incorporated into this analysis. The UKBB has been comprehensively described in previous chapters of this thesis (Chapters 2–3). The quality-controlled sample set from Chapter 3 was utilized, along with individuals of African, Southeast Asian, and East Asian ancestry, as detailed in Section 3.2.1. In summary, to determine ancestry, I employed a RF classifier trained on 1KGP super-population labels. This classifier was applied to predict the super-population of each UKBB sample (Figure B.2). An individual's ancestry was defined as the super-population for which the RF classifier assigned a probability greater than 0.99.

4.2.2 General quality control and phasing guidelines for biobanks

Each biobank performed quality control separately. However, we provided general guidelines for variant, sample-level quality control, defining ancestry, and phasing, which is described here.

Variant calling

Biobanks were recommended to process exome or whole-genome sequencing reads using `GATK`[357] v4.2.6.1, though alternative tools were allowed. The suggested workflow included aligning reads to the GRCh38DH reference genome using `BWA-MEM`[358] v0.7.15, and marking duplicate reads with `Picard`[359] v2.4.1. Base quality score recalibration could be performed using `GATK BaseRecalibrator`. Following these steps, variant calling was recommended to be conducted using `GATK HaplotypeCaller`

(v4.1.4.1 or 4.1.7.1) to generate gVCF files. Joint genotyping could then be performed using `GenomicsDBImport` and `GenotypeGVCFs`. It was advised to apply Variant Quality Score Recalibration (VQSR) separately for SNPs and indels using `VariantRecalibrator` and `ApplyVQSR` from `GATK`, though biobanks were free to use alternative methods if deemed appropriate.

Recommended sample and variant quality control

We recommend performing quality control using `Hail`[203] in multiple steps. For ES data, we suggest filtering sites where at least 85% of samples have a mean coverage of 20x or higher. Individual genotypes should be filtered based on allelic depth (DP≥10) and genotype quality (GQ≥20). Variants should be excluded if they fall within LCR, fail VQSR, or lie outside of padded target intervals (using 50bp padding). Samples should be excluded based on any of the following criteria: call rate (<93%), mean depth (DP<30), mean genotype quality (GQ<55), `FREEMIX` contamination estimates (>2%), and proportion of chimeric reads (>1.5%). These thresholds were determined by examining the distributions of the corresponding metrics in the UKBB ES 450k dataset. However, we strongly recommend that each biobank conduct its own investigation to determine appropriate thresholds based on their avilable data and quality standards.

After these initial filters, we recommend performing a final pass through the data. Remaining sites with low call rates (<97%) or significant deviation from HWE ($P < 1^{-10}$) should be excluded. We advise removing additional sample outliers (based on 4 standard deviations) based on metrics like number of SNPs, transition/transversion ratio, and heterozygous/homozygous ratio, using median absolute deviation within each ancestry group and sequencing platform. Ideally, this step should be carried out in batches based on ancestries and/or recruitment centres where applicable.

Sex imputation

We recommended that each biobank perform sex imputation to confirm that the reported sex matched the genetically inferred sex, as discrepancies could signal potential sample swaps in the data. We recommended extracting high-quality

common variants (MAF>0.01, call rate >0.98) and performing LD pruning to obtain pseudo-independent SNPs using `PLINK` 1.9 (`-indep 50 5 2`). We recommended calculating an F-statistic for each sample using the non-pseudo autosomal region on chromosome X. We advised removing samples where:

- sex is unknown in the phenotype files,
- F-statistic > 0.6 for reported females
- F-statistic < 0.6 for reported males
- F-statistic > 0.6 with < 100 calls on the Y chromosome

Importantly, the same set of high-quality, LD-pruned SNPs could also be used to calculate PCs for downstream analysis allowing the ascertainment of genetic ancestry.

Defining ancestry

We recommended a multi-step process for population stratification and ancestry definitions. First, we advised performing PCA on the 1KGP[360] samples, using LD-pruned autosomal variants. We then suggested projecting the biobank samples onto this PCA space. While acknowledging that ancestry is not truly categorical[124,361], we advocated for implementing a RF classifier trained on the super-population labels of 1KGP (EUR, AFR, EAS, SAS, American/Admixed (AMR)). We proposed using this classifier to predict the super-population for each biobank sample. We recommended selecting samples with a probability > 0.99 of being classified as a particular super-population according to the RF model. This conservative approach excludes admixed samples, which otherwise can contribute to inflation of the test statistic, when not adequately controlled for[362,363]. This cautious approach allow for robust identification and selection of genetically homogeneous cohorts within each ancestry group, and allow for for more well calibrated test statistics.

Phasing GnH and BBJ

The following Snakemake pipeline was based on my work from chapter 2 and implemented by G. Kalantzis (analyst from GnH) and shared with collaborators.
Creating pipelines and tools that can be easily run by collaborators is crucial for global research efforts. To this end, and to facilitate rapid phasing of exome sequencing and genotyping array data in biobanks, we developed a `SHAPEIT5`[141] wrapper using `snakemake`. This wrapper was designed for easy sharing and integration into various biobank workflows. The resulting pipeline was successfully implemented to phase genotypes in GnH and BBJ. In this pipeline, phasing is conducted on single-chromosome chunks after combining ES and genotyping array data, with genotyping array variants prioritized in cases of overlap. The phasing process employs a two-stage approach: first, common variants (MAF > 0.001) are phased, which then serve as a scaffold for phasing the remaining rare variants. This method enables efficient and accurate phasing of both common and rare genetic variants across large-scale biobank datasets.

Validating of phasing accuracy

Accuracy of phasing was assessed using read-backed phasing and trio-switch error rates. For GnH, we calculated trio-switch error rates using 100 trios (300 samples) of SAS ancestry, while for UKBB, we used 99 trios (297 samples) of EUR ancestry. In the case of BBJ, where trios were not available, we conducted read-backed phasing on all autosomes using `whatshap`[211] across 1000 randomly sampled individuals of EAS ancestry. We then used custom `R` scripts to determine the agreement between read-backed phased and statistically inferred haplotypes. This process involved identifying pairs of variants in close proximity (150-250bp) on short-read sequences from `.bam` or `.cram` files using `WhatsHap`[211] with default parameters. Two variants on the same short-sequencing read must originate from the same haplotype (*in cis*), while variants on different reads are likely to originate from opposite haplotypes (*in trans*). Given this, phasing accuracy was estimated as the proportion of read-backed configurations of variant pairs that agreed with the configuration of statistically

inferred variants. Phasing accuracy was not explicitly validated for GEL for three reasons: 1) trios had already been used to perform transmission phasing in GEL (See section 4.2.4) and could therefore not be used for validation; 2) read-backed phasing of GEL WGS was extremely computationally demanding; and 3) estimates of imputation accuracy, which serves as a proxy for phasing accuracy, had previously been published for this exact dataset[364] (See Section 4.2.4).

4.2.3 UK Biobank quality control and phasing

The quality control pipeline for UKBB has already been extensively described in chapter 3, Section 3.2.1. This includes definition of ancestry (section 3.2.1), sex imputation (Section 3.2.1), and multiple stages of variant-level and sample-level quality control (Section 3.2.1), and phasing (Section 3.2.3).

4.2.4 Genomics England quality control

Variant and sample-level QC

The following workflow is adapted from the GEL documentation available online. To avoid computationally cumbersome recalling of variant calls in WGS from GEL, we leveraged the quality-controlled Aggregated Variant Call (AggV2) dataset, a multi-sample VCF comprising 78,195 germline genomes and 722 million SNPs and small indels ($\leq$ 50bp) upon release. This dataset was constructed by merging single-sample gVCF files, with sample and variant-level quality control described below. All quality control procedures described in this section were carried out by the GEL team. Samples were sequenced on Illumina HiSeq X instruments using 150bp paired-end reads, one lane per sample. Data processing employed the Illumina `North Star` v4 Whole Genome Sequencing Workflow (NSV4, v2.6.53.23), utilizing `iSAAC Aligner` (v03.16.02.19) and `Starling Small Variant Caller` (v2.4.7). Reads were aligned to the Homo sapiens NCBI GRCh38 assembly with decoys. Samples were individually quality controlled and required to meet the following criteria:

- Sample contamination (`freemix`) < 0.03

- Ratio of SNV heterozygous to homozygous calls < 3
- Total number of SNVs between 3.2M-4.7M
- Array concordance $> 90\%$
- Median fragment size > 250bp
- Excess of chimeric reads $< 5\%$
- Percentage of mapped reads $> 60\%$
- Percentage of AT dropout $< 10\%$

Additionally, samples were required to have at least 95% of the genome covered at 15X or above with well-mapped reads (mapping quality > 10) after discarding duplicates. Variants were split from multi-allelic into biallelic and indels were left-aligned using `vt` v0.57721. Variants were filtered based on the following criteria:

- Missingness $\leq 5\%$
- Median depth ≥ 10
- Median genotype quality ≥ 15
- $\geq 25\%$ of heterozygous calls showing no significant allele imbalance
- $\geq 50\%$ of sites with complete genotype data
- HWE mid P-value $\geq 1^{-5}$ in unrelated samples of inferred European ancestry

Finally, sex-specific QC was performed for the X chromosome. In summary, 722,342,407 (74.8%) autosomal variants passed all QC filters.

Phasing

We utilized phased genotypes provided by Shi *et al.*[364], who performed additional variant- and sample-level quality control prior to phasing, and the following description is adapted from their work. This quality control process evaluated genotype quality, depth, missingness, allelic balance, Mendel errors, HWE equilibrium, and gnomAD[227] allele frequency concordance. Notably, Shi *et al.* removed singletons in unrelated individuals, as these could not be phased statistically using `SHAPEIT4`. Overall, these stringent filters reduced the number of variants from 722M to 342M. The phasing strategy, as described by Shi *et al.*[364], involved a multi-stage strategy that incorporated the high relatedness owed to many families within GEL:

1. The `makeScaffold` tool (`https://github.com/odelaneau/makeScaffold`) was employed to determine the phase of duo and trio samples using direct transmission phasing, resolving most sites in related samples. The remaining unphased genotypes in related samples then were inferred using `SHAPEIT4` (v4.2.2), utilizing the phased genotypes from the transmission phasing step as a scaffold.

2. For unrelated samples, common variants (AF ≥ 0.01) were phased chromosome-wise using `SHAPEIT4`, employing the phased genotypes from steps 1 as a reference panel. Then, the remaining rare variants in unrelated samples were phased using `SHAPEIT4`, leveraging the phased common variants as a scaffold, and the phased related samples as a reference panel.

Phasing of related individuals (step 1) was conducted at the chromosome level, while step 2 were carried out in regions of approximately 300,000 sites, with 30,000 sites on each side as a buffer. The resulting phased regional segments were merged and concatenated using `bcftools ligate`. Shi *et al.*[364] assessed phasing accuracy by evaluating the performance of GEL as a reference panel through genotype imputation in 589 trios from the 1000 Genomes Project, using `SHAPEIT4` (v4.2.2). Phasing accuracy was then evaluated by comparing imputed haplotypes to those inferred from Mendelian inheritance patterns in each trio.

4.2.5 Biallelic variant identification

Variant masks

Variants were annotated using `VEP v105`[262], using the same definitions as in chapter 3. In summary, we employed `LOFTEE v1.04`[227] to classify PTVs as HC or LC LoF variants. Damaging missense and protein-altering variants were assessed using `CADD v1.6`[264] and `REVEL v7`[238] scores from `DBNSFP v4.3`[263], with `CADD v1.6` separately applied to indels. `SpliceAI v1.3` was used for splicing variant annotation. We restricted our analysis to MANE Select transcripts when available, otherwise defaulting to `GENCODE V39`[266] canonical transcripts. Variants were hierarchically classified as: (1) pLoF: HC PTVs filtered by `LOFTEE`; (2) damaging missense/protein-altering: including LC LoF, variants with `REVEL`-Score ≥ 0.773 or `CADD`-Phred score ≥ 28.1, and splicing variants with SpliceAI Δ score ≥ 0.5; (3) other missense; (4) synonymous (SpliceAI Δ score < 0.5); (5) non-coding variants; and (6) a nonsynonymous variants, which included all pLoF, damaging missense/protein altering and other missense variants. Synonymous variants were included as a negative control in the analysis. All biobanks restricted to variants with MAF≤ 0.05.

Biallelic variant encoding

Biallelic variants were identified as described in chapter 3. Briefly, I developed a custom `C++` toolset (`https://github.com/BRaVa-genetics/biallelic`), that could be shared across biobank platforms with minimal dependencies, to identify and annotate individuals with biallelic variants. Using phased data, we collapsed rare variants (MAF $< 5\%$) across individual haplotypes or gene copies. An individual was classified a monoallelic carrier if only a single gene copy is affected, and biallelic if both gene copies are affected. Using this terminology, we considered the following models for analysis:

1. Additive encoding ($[0, 1, 2]$)
2. Recessive encoding ($[0, 0, 1]$)
3. Nonadditive encoding ($[-ha, 2ar, -hr]$, see below)

The additive encoding considers the number of gene copies affected, whereas the recessive recoding takes into account the number of biallelic carriers. We encoded nonadditivity at each locus following the same approach as described in detail in section 3.2.5-3.2.8. In summary, for each pseudo-variant, we defined r, h, and a as the proportions of individuals with zero, one, and two haplotypes carrying one or more alternate alleles, respectively. We then mapped the genotypes $[0, 1, 2]$ to $[-ha, 2ar, -hr]$ for each pseudo-variant, and performed a rescaling to restrict dosages to $[0, 2]$. This nonadditive recoding of genotypes was only carried out in UKBB, GEL and GnH, because BBJ had yet to share summary statistics generated using this nonadditive encoding.

4.2.6 Counting expected and observed biallelic variants

This method has already been described in section Chapter 2, section 2.2.19.

4.2.7 Phenotype curation

Overview of phenotyping efforts

To motivate participation, phenotypes were democratically selected by representatives from each biobank through nomination of ICD10 (and ICD9) case and control inclusion/exclusion criteria (Table C.1). This resulted in a list of 32 binary and 11 quantitative phenotypes which can be found in Tables 4.1 and 4.2, which included sex-specific traits. For the five sex-specific endpoints (Benign and in situ cervical and uterine neoplasms, Cervical cancer, Breast cancer, Excess, frequent and irregular menstrual bleeding, Female Infertility, Maternal hemorrhage) only females were included in the study sample. Only GnH and UKBB had access to and were able to share quantitative phenotypes for this analysis. One trait were only analyzed in a single biobank (Alcohol consumption) and was therefore excluded from the meta-analysis.

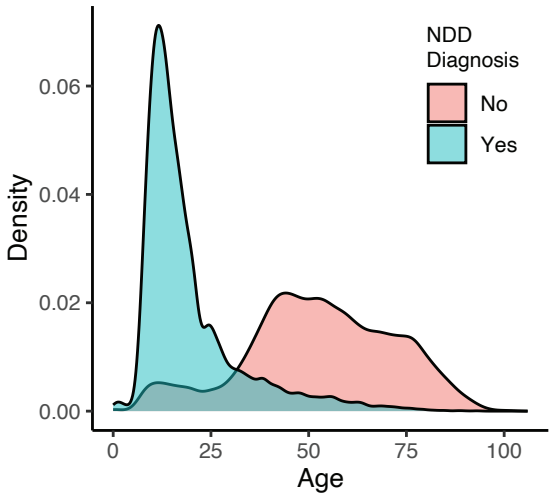


Figure 4.1: Distribution of age in Genomics England for individuals diagnosed with NDD. Distribution of age in 72,841 samples in the GEL cohort stratified by NDD diagnosis status (red) versus no diagnosis (blue). NDD is defined as having either global developmental delay, autism, and/or intellectual disability (See Table C.4).

Additional control exclusions with Neuro Developmental Disorder (NDD)

We excluded probands in GEL diagnosed with either global developmental delay (n=12,191), intellectual disability (n=9,149), or autism (n=4,609), identified through manual curation of ICD and Human Phenotype Ontology (HPO) codes (see Table C.4). Throughout this thesis, we refer to these terms collectively as NDD. The exclusion was necessary due to two key factors: firstly, NDD patients in our cohort were predominantly very young (median age 15 years - see Figure 4.1), making them more likely to be classified as controls for many common adult-onset diseases investigated here. Secondly, NDD patients have been reported to be enriched for rare biallelic germline variants[365]. The combination of these factors could potentially lead to an artificial enrichment of rare biallelic variants in our control group, risking spurious associations where the absence of these variants might appear to increase disease risk. To mitigate this bias, we excluded a total of 13,976 NDD probands, reducing our sample size from 72,841 to 58,865, while retaining 81% of the original GEL samples.

4.2.8 Association analysis

UKBB, GnH, and GEL performed biallelic association testing using `SAIGE`[240] and BBJ performed association analysis using `REGENIE`[85]. Both of these tools scaled to biobank-level data and accounted for sample relatedness and case-control imbalance. Each analysis was carried out seperately for each super ancestry group. For analysis with `SAIGE`, we recommended constructing a sparse GRM using 2000 randomly selected markers and a relatedness cutoff of 0.05. The suggested covariates for the null models were age, age-squared, sex, age-by-sex interaction, age^2-by-sex interaction, and first ten principal components, as well as biobank specific covariates, such as genotyping batches and recruitment centres. Quantitative traits were subject to INT. For binary traits, Firth correction was applied for variants with P <0.01. Minimum BC count was set to 5 (corresponding to MAC of 10), and no minimum MAF filter was imposed.

4.2.9 Meta-analysis

Fixed-effects meta-analysis

Summary statistics were shared across biobanks by collaborators. We then conducted a fixed-effects meta-analysis using Stouffer's method[366]. This approach weights P-values from each study by their effective sample size (n_{eff}). Specifically, the method converts P-values to signed Z-scores, which are then combined in a weighted sum according to equation 4.1:

$$Z = \frac{\sum_{i=1}^{k} \sqrt{n_{\text{eff}}} \cdot (-\Phi^{-1}(p_i)) \cdot \text{sign}(\beta_i)}{\sqrt{\sum_{i=1}^{k} n_{\text{eff}}}} \tag{4.1}$$

where k is the number of studies, n_{eff} is the effective sample size of study i, p_i is the P-value from study i, Φ^{-1} is the inverse standard normal cumulative distribution function, and β_i is the direction of effect in study i. The effective sample size n_{eff} is defined using the number of cases N_{cases} and controls N_{controls} in each cohort, calculated by:

$$n_{\text{eff},i} = \frac{4}{\frac{1}{N_{\text{cases},i}} + \frac{1}{N_{\text{controls},i}}} \tag{4.2}$$

The meta-analysis P-value was then computed as $p = 1 - \Phi(Z)$ for one-tailed tests or $p = 2(1 - \Phi(|Z|))$ for two-tailed tests, where Φ is the standard normal cumulative distribution function (CDF). This approach is flexible, as we do not require the effect sizes and corresponding standard errors.

Heterogeneity testing

Heterogeneity among studies was assessed using Cochran's Q test[367]. The meta-analytic effect size was calculated as a weighted average:

$$\bar{\beta} = \frac{\sum_{i=1}^{k} w_i \beta_i}{\sum_{i=1}^{k} w_i} \tag{4.3}$$

where β_i is the effect size of study i, and w_i is the weight assigned to study i. Weights were calculated as the square root of the effective sample (n_{eff}) size ($w_i = \sqrt{n_{eff}}$). Then, Cochran's Q statistic was computed as:

$$Q = \sum_{i=1}^{k} w_i (\beta_i - \bar{\beta})^2 \tag{4.4}$$

Under the null hypothesis of homogeneity, Q follows a chi-square distribution with $k - 1$ degrees of freedom, where k is the number of studies. The P-value for the heterogeneity test was calculated as:

$$P_{het} = P(\chi^2_{k-1} > Q) \tag{4.5}$$

This test quantifies the extent of between-study variability in effect sizes beyond what would be expected by chance alone.

4.2.10 Cauchy Combination of *P*-values across annotation masks

The following description is adapted from Liu *et al.*[368,369]. To compute a single P-value per gene-trait pair, we employ Cauchy combinations, combining P-values

from up to four annotation masks (pLoF, damaging missense/protein altering, pLoF+damaging missense/protein altering, and nonsynonymous variants). The P-values derived from these annotation masks are expected to have moderate to strong correlation, partly because of LD. Let P_i be individual P-values, for $i = 1, 2, \ldots, d$. The cauchy combination test statistic is defined as:

$$T_{\text{Cauchy}} = \sum_{i=1}^{d} \omega_i \tan\left((0.5 - P_i)\,\pi\right), \tag{4.6}$$

where the weights $\omega_i \geq 0$ and $\sum_{i=1}^{d} \omega_i = 1$. The transformation $\tan\left((0.5 - P_i)\pi\right)$ follows the standard Cauchy distribution if P_i follows the null distribution, i.e., is uniformly distributed between 0 and 1. The combined P-value can therefore be evaluated as:

$$P_{\text{Cauchy}} = F_C(T_{\text{Cauchy}}) \tag{4.7}$$

where $F_C(\cdot)$ is the cumulative distribution function of the standard Cauchy distribution. This provides an efficient method for combining potentially correlated P-values from multiple annotation masks into a single, interpretable P-value for each gene-trait pair, while maintaining type 1 errors[369].

4.2.11 Power comparison between meta-analysis and UK Biobank

To assess the gain in statistical power achieved through meta-analysis compared to using the UKBB dataset alone, we conducted a comparative analysis of chi-squared (χ^2) statistics derived from both approaches. Our meta-analysis incorporated data from multiple biobanks: the UKBB, GnH, BBJ, and GEL. For each gene-trait association pair, we converted P-values from both the UKBB analysis (P_{UKBB}) and the meta-analysis (P_{META}) to χ^2 statistics (1-DF) using the inverse cumulative distribution function (F^{-1}) of the χ^2 distribution:

$$\chi^2_{\text{UKBB}} = F^{-1}(1 - P_{\text{UKBB}}) \tag{4.8}$$

$$\chi^2_{\text{META}} = F^{-1}(1 - P_{\text{META}}) \tag{4.9}$$

We then calculated the ratio of these χ^2 statistics ($\chi^2_{\text{META}}/\chi^2_{\text{UKBB}}$), which approximates the increase in relative effective sample sizes under the assumption of consistent effect sizes across studies. This ratio serves as an estimate of power gain, with a ratio greater than 1 indicating increased statistical power in the meta-analysis compared to the UKBB alone. This approach provides a quantitative measure of the enhancement in statistical power achieved through our multi-biobank meta-analysis strategy, allowing us to assess the value added by incorporating additional datasets beyond the UKBB.

4.2.12 Evaluation of sample overlap across cohorts

To evaluate potential sample overlap and/or cryptic relatedness between individual biobanks, I compared Z-scores between summary statistics of synonymous variants across biobanks and ancestries using the additive genetic model. This approach is adapted from the software `METAL`[370], and leverages the principle that if overlap exists between biobanks, Z-scores are likely to be correlated under the null hypothesis. Synonymous variants were chosen for this analysis, as nonsynonymous variants are more susceptible to correlation due to true shared underlying genetic effects between genes and traits. The analysis was conducted in `R` using the `boot` library, employing ordinary bootstrap of Pearson correlations with 1000 replicates to determine 95% CIs. *P*-values were calculated using `stats::cor.test`.

Quantifying power by effective sample size

To quantify a lower limit for detectable ORs, I used an approximation at a pre-specified significance threshold expressed as:

$$\beta = \sqrt{\frac{\chi^2}{n_{\text{eff}}}} \tag{4.10}$$

where β is the effect size, n_{eff} is the effective sample size, and for binary traits, OR $= e^{\beta}$. Assuming a genome-wide significance threshold of $P = 1 \times 10^{-7}$, we determine the corresponding χ^2 statistic using the inverse cumulative distribution function of the χ^2 distribution (1-DF). With χ^2 fixed, I calculate β for the cumulative n_{eff} as we

sequentially include each biobank. This approach estimates the minimum detectable effect size (OR) for a given n_{eff}, demonstrating that as n_{eff} increases, we are powered to detect gradually smaller effect sizes (ORs closer to 1) at the same significance level.

4.3 Results

4.3.1 Patterns of disease across cohorts

Selection of 42 traits for downstream analysis

In this analysis, we examined 519,594 exome and whole genome sequenced samples across four biobanks participating in the BRaVa consortium: UK Biobank (UKBB), Genes and Health (GnH), Biobank Japan (BBJ), and Genomics England (GEL). This large-scale study represents the first combined analysis of rare variant contributions to health and disease across these diverse cohorts. Each biobank brings unique characteristics to the consortium: GnH is a community-based study of British Bangladeshi and British Pakistani people in east London[252]; BBJ is a hospital-based national biobank primarily recruiting Japanese patients with 47 diseases[371]; and GEL focuses on rare disease and cancer patients and their relatives. Following separate quality control procedures in each biobank (Methods 4.2.2), our analysis included 410,030 samples from UKBB, 58,865 from GEL, 39,152 from GnH, and 11,462 from BBJ.

We categorized ancestry into four super populations (EUR, EAS, SAS, AFR) based on genetic similarity to reference populations from the 1000 Genomes Project[372] (Methods 4.2.2). The combined BRaVa cohort was predominantly of European genetic ancestry (85.9% European, n=446,402), largely due to UKBB's composition (96.4% European, n=395,495) (Figure 4.3, Table 4.1). However, the other biobanks contributed significant genetic diversity: 51,327 individuals of South Asian ancestry, 8,317 of African ancestry, and 13,548 of East Asian ancestry. Notably, GnH focused exclusively on individuals of SAS ancestry, while BBJ analyzed only individuals with EAS ancestry.

The BRaVa consortium democratically defined 32 binary health endpoints and 10 quantitative traits for our pilot analysis (Figure 4.2, Methods 4.2.7). While the

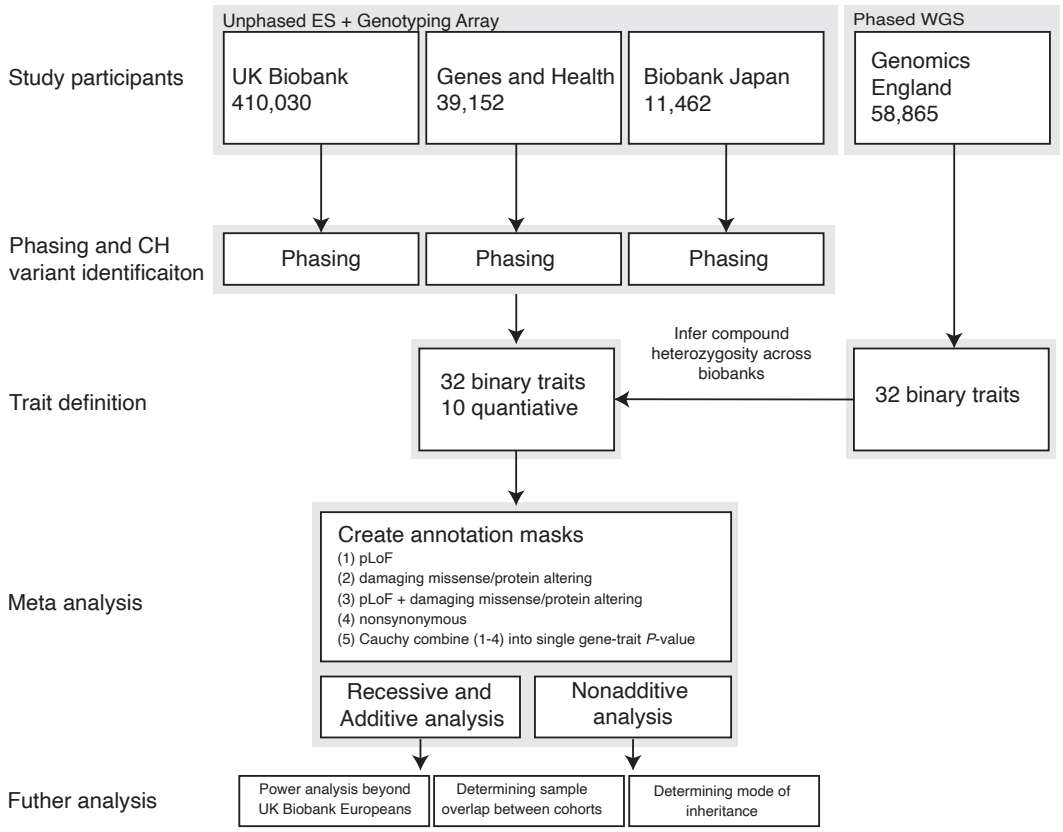


Figure 4.2: Flowchart for the BRaVa pilot analysis. UK biobank, Genes and Health, Genomics England and Biobank Japan was included in our study, totalling 519,594 samples. Only Genomics England had phased WGS, and the remainder biobanks employed joint phasing of ES and genotyping array data. 42 traits were ascertained across biobanks, with GEL only providing binary endpoints. We performed additive, recessive and nonadditive analysis across up to four annotation masks, annd the resulting P-values were combined using the Cauchy distribution for each gene-trait pair.

binary endpoints were available across all biobanks (Table 4.1), the quantitative measurements, including biochemical assays and anthropometric measurements, were available for up to three of the four biobanks, excluding GEL from this analysis (Table 4.2). The selected binary endpoints encompassed a broad spectrum of health conditions, including cardiovascular diseases, respiratory disorders, various cancers, and age-related conditions. These ranged from common conditions such as hypertension (prevalence of 31.5% with 172,945 cases across all cohorts) and type 2 diabetes (11.0%, 60,422 cases), to more rarely reported and sex-specific conditions like cervical cancer (0.2%, 1,346 cases) and female infertility (0.5%, 2,507 cases).

Table 4.1: Number of cases ascertained across 32 binary traits in four biobanks. This study analyzed 32 binary endpoints across 519,594 quality-controlled samples from four biobanks: UKBB 450K ES release, Genomics England (GEL), Genes and Health (GnH), and Biobank Japan (BBJ). Case counts for 32 binary traits, selected for pilot analysis, were stratified by ancestry (AFR, EAS, SAS and EUR). GEL data excludes 14,245 NDD probands and does not represent the full biobank. Phenotypes with fewer than five cases were redacted.

Trait	ID	Cases	Controls	Prev.	UK Biobank (450K)					Genomics England					Genes & Health		Biobank Japan	
					410,113	6,600	1,649	395,495	6,369	58,865	1,717	437	50,907	5,804	39,154	39,154	11,462	11,462
Total samples:	519,594				All	AFR	EAS	EUR	SAS	All	AFR	EAS	EUR	SAS	All	SAS	All	EAS
Age-related macular degeneration	AMD	11,517	536,808	0.021	8,969	112	39	8,654	164	928	17	7	842	62	303	303	0	0
Asthma	Asth	59,828	488,497	0.109	39,935	687	104	38,336	808	8,594	212	44	7,632	706	5,788	5,788	91	91
Atrial Fibrillation	AFib	40,550	507,775	0.074	34,175	237	46	33,595	297	3,636	59	11	3,452	114	569	569	100	100
in situ cervical and uterine neoplasm	CervUterNeo	17,920	530,405	0.033	13,626	270	64	13,094	198	1,941	65	19	1,857	NA	1,250	1,250	0	0
in situ intestinal neoplasms	BenIntNeo	52,937	495,388	0.097	43,328	520	141	42,117	550	5,636	135	28	5,240	233	1,813	1,813	0	0
Breast cancer	BreastCanc	17,131	531,194	0.031	15,200	155	51	14,838	156	361	NA	NA	361	NA	224	224	258	258
Cervical cancer	CervCanc	4,894	543,431	0.009	744	16	3	719	6	3,922	NA	43	3,879	NA	78	78	0	0
Chronic obstructive pulmonary dis.	COPD	31,075	517,250	0.057	23,890	188	20	23,421	261	3,630	52	11	3,441	126	986	986	33	33
Chronic Renal Failure	CRF	31,471	516,854	0.057	20,082	423	31	19,228	400	3,837	196	36	3,338	267	1,963	1,963	950	950
Colon and rectum cancer	olonRectCanc	28,233	520,092	0.051	15,033	243	59	14,529	202	4,125	107	28	3,839	151	266	266	240	240
Coronary artery disease	CAD	28,030	520,295	0.051	21,727	189	24	20,916	598	2,107	33	7	1,911	156	1,722	1,722	1,564	1,564
Excess, frequent,irregular menstrua	EFRMB	23,484	524,841	0.043	15,078	473	70	14,268	267	23	NA	23	NA	NA	5,874	5,874	0	0
Female infertility	FemInf	2,137	546,188	0.004	1,217	37	13	1,139	28	0	NA	NA	NA	NA	687	687	0	0
Gout	Gout	12,715	535,610	0.023	9,459	148	23	9,120	168	1,219	62	15	1,064	78	831	831	128	128
Heart Failure	HF	22,765	525,560	0.042	16,344	228	19	15,767	330	2,349	85	11	2,123	130	1,012	1,012	284	284
Hypertension	HTN	172,945	375,380	0.315	131,959	2,713	355	126,216	2,675	14,596	611	115	12,833	1,037	9,988	9,988	2,733	2,733
Inflammatory bowel disease	IBD	33,385	514,940	0.061	25,240	277	52	24,467	444	4,290	85	11	3,853	341	1,384	1,384	0	0
Inguinal, femoral, and abdom. hern	IFHern	79,453	468,872	0.145	67,576	712	102	65,851	911	7,421	187	26	6,795	413	2,436	2,436	0	0
Interstitial lung dis. + pulmonary sar	ILDSarc	6,662	541,663	0.012	4,428	106	8	4,212	102	806	23	5	730	48	275	275	18	18
Maternal hemorrhage	MatHem	7,582	540,743	0.014	3,115	103	22	2,927	63	46	NA	46	NA	NA	4,023	4,023	0	0
Non-rheumatic valvular heart dis.	NonRheuValv	18,415	529,910	0.034	13,829	155	26	13,451	197	1,967	58	11	1,809	89	684	684	0	0
Pancreatitis	Pancreat	8,087	540,238	0.015	5,724	81	12	5,545	86	1,010	25	7	915	63	397	397	38	38
Peptic ulcer disease	PeptUlcer	15,968	532,357	0.029	12,167	192	54	11,666	255	1,298	55	15	1,124	104	1,102	1,102	591	591
Peripheral artery disease	PAD	12,043	536,282	0.022	7,509	77	10	7,300	122	908	27	NA	836	45	340	340	178	178
Psoriasis	Psori	8,192	540,133	0.015	5,067	22	10	4,940	95	866	11	NA	800	55	1,259	1,259	204	204
Rheumatic heart disease	RheumHeaDis	11,605	536,720	0.021	8,871	139	24	8,559	149	1,448	47	13	1,314	74	584	584	0	0
Rheumatoid arthritis	RheumArth	2,567	545,758	0.005	1,278	15	1	1,237	25	207	NA	NA	187	20	497	497	200	200
Stroke	Stroke	12,595	535,730	0.023	9,352	154	17	9,027	154	952	25	6	866	55	406	406	442	442
Type 2 diabetes	T2Diab	60,422	487,903	0.110	35,339	1,251	130	32,168	1,790	4,581	215	37	3,685	644	10,144	10,144	2,523	2,523
Urolithiasis	Urolith	14,584	533,741	0.027	9,583	113	36	9,252	182	1,752	35	23	1,551	143	1,650	1,650	61	61
Varicose Veins	VaricVeins	20,118	528,207	0.037	15,339	108	25	15,031	175	1,391	12	5	1,307	67	1,450	1,450	0	0
Venous Thromboembolism (VTE)	VTE	14,870	533,455	0.027	12,131	203	14	11,812	102	2,338	91	8	2,136	103	401	401	0	0

Table 4.2: Number of samples with up to 10 quantitative measurements across three biobanks with quantitative measurements. Three biobanks contributed with up to 10 biomarker measurements to the final analysis: UK Biobank, Genes and Health, and Biobank Japan. This table illustrates the number of defined quantitative values from each biobank, stratified by ancestries: AFR, EAS, SAS and EUR. Note, that 'Alcohol consumption' was only provided by GnH and therefore excluded from meta-analysis.

Trait description	ID	Total	UK Biobank						Genes and Health		Biobank Japan	
			All	AFR	AMR	EAS	EUR	SAS	All	SAS	All	EAS
Alanine transaminase	ALT	425,394	391,104	6,188	0	1,569	377,312	6,035	30,922	30922	3,368	3,368
Alcohol consumption (drinks per week)	AlcCons	19,312	0	0	0	0	0	0	19,312	19,312	0	0
Aspartate aminotransferase (AST)	AST	405,587	389,795	6,153	0	1,558	376,061	6,023	12,412	12412	3,380	3,380
BMI	BMI	442,602	408,573	6,496	0	1,639	394,200	6,238	30,408	30408	3,621	3,621
C-reactive protein (CRP)	CRP	416,020	390,429	6,178	0	1,563	376,663	6,025	23,616	23616	1,975	1,975
HDLC	HDLC	388,100	358,132	5,739	0	1,423	345,476	5,494	27,482	27482	2,486	2,486
Height	Height	444,703	409,024	6,507	0	1,640	394,632	6,245	32,043	32043	3,636	3,636
LDLC	LDLC	419,182	390,503	6,175	0	1,568	376,730	6,030	26,118	26118	2,561	2,561
Total cholesterol	TChol	260,844	229,743	3,319	0	860	222,306	3,258	27,750	27750	3,351	3,351
Triglycerides	TG	420,152	390,935	6,184	0	1,569	377,145	6,037	26,084	26084	3,133	3,133

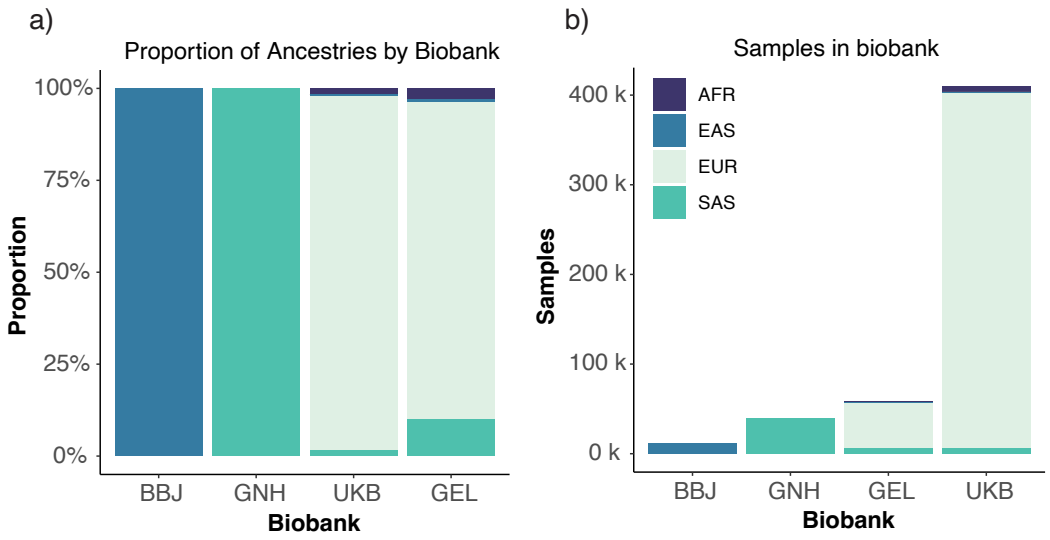


Figure 4.3: Proportion and counts of samples analyzed in the BRaVa pilot. Overview of samples analyzed in this chapter stratified by biobank and ancestry assignment. a) Proportion of ancestries per biobank. b) Number of quality controlled individuals per ancestry per biobank, which are included for downstream analyses.

Challenges and considerations in biobank case ascertainment and analysis of disease prevalence

Our study faced challenges related to selection bias in biobank-wide association studies[373]. Specifically, I was concerned with a form of bias that shares characteristics with collider bias. In our scenario, the GEL cohort is enriched for pediatric rare disease patients with NDD, and these patients have been reported to be enriched for rare germline biallelic variation[365]. Given that these are young NDD patients, with a median age of 15 (Figure 4.1), they would likely be classified as controls for many of the adult-onset common diseases under study, potentially leading to an enrichment of rare biallelic variants in the control group. This could lead to a spurious association suggesting that the absence of these rare biallelic variants is linked to an increased risk of the investigated trait. To mitigate this potential bias, I excluded 13,976 probands diagnosed with NDD from our analysis (Methods 4.2.7). This refinement reduced our sample size for association analysis from 72,841 to 58,865, retaining 81% of the original number of samples in GEL.

Despite potential differences in recruitment strategies, I observed a generally high concordance in the proportion of cases across the four biobanks. This consistency

was evident in the high pairwise correlations between biobank case prevalences for traits defined in multiple biobanks (Figure 4.4). The strongest correlation was found between UKBB and GEL (Spearman $r = 0.82$), likely reflecting their shared use of NHS for health linkage and similar ancestry compositions reflected by the population in the United Kingdom. In support of this, I observed relatively high correlation between GnH with both GEL (Spearman $r = 0.6$) and UKBB (Spearman $r = 0.58$). The weakest correlation was observed between GEL and BBJ (Spearman $r = 0.38$), potentially indicating differences in both case ascertainment strategies and ancestry composition. All three UK-based cohorts had low correlation with BBJ (maximum Spearman correlation $r = 0.44$), which is potentially a reflection of the fact that the UK cohorts incorporate both primary and secondary care data, while BBJ relies on hospital records[374]. Nonetheless, the general high concordance suggests a robust overall approach to case identification across the consortium.

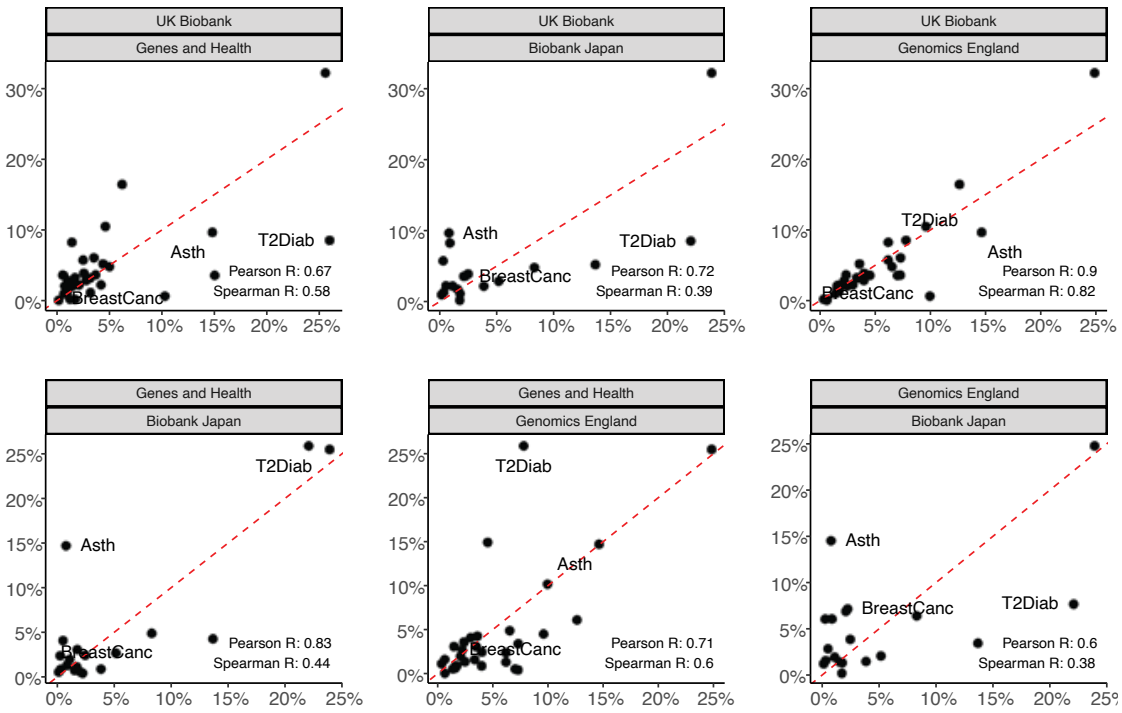


Figure 4.4: Scatter plot of the pairwise comparison of disease prevalence between biobanks. The x-axis represents the prevalence of a specific trait in one biobank (upper facet), while the y-axis represents the prevalence of the same trait in another biobank (lower facet). Each point corresponds to a specific trait aggregated across all ancestry groups. The diagonal dashed line indicates where the prevalences would be equal across biobanks. Pearson and Spearman correlation coefficients are shown in each panel, quantifying the strength and direction of the linear and rank-based relationships, respectively. Up to six selected traits are highlighted with labels: Chronic obstructive pulmonary disease (COPD), in situ intestinal neoplasms (BenIntNeo), Coronary Artery Disease (CAD), Hypertension (HTN), Asthma (Asth), and Type 2 Diabetes (T2Diab). The red dashed line indicates $x = y$.

Next, I investigated specific instances where disease incidence differed markedly across biobanks. These differences reflected both established epidemiological trends and the distinct characteristics of each biobank. A striking example is the distribution of Type 2 Diabetes (T2Diab) cases. The global adult T2Diab prevalence is projected to reach 7.7% by 2030[375], significantly lower than the rates observed in both biobanks. The prevalence in GnH (25.9%) is more than three times the projected global rate, while the UKBB prevalence, although lower (8.6%), still exceeds the global projection. This disparity aligns with the documented higher prevalence of T2Diab in South Asian populations[376,377] and aging individuals[378], as well as in diverse urban settings[379,380]. These findings highlight the diversity across biobanks in terms of ancestral composition and disease prevalence, presenting unique opportunities for trait-specific studies.

Further examples underscore how biobank-specific factors can significantly influence observed patterns of disease occurrence. Asthma prevalence demonstrates this variation clearly: while the global prevalence was around 3.3% in 2021[381], we observed markedly different rates across biobanks. BBJ, a hospital-based registry focused on specific diseases, showed only 0.8% (91 cases) diagnosed with asthma, far below Japan's population prevalence of 3.5%[381]. In contrast, UKBB, GEL, and GnH exhibited much higher rates at 9.7%, 14.5%, and 14.7%, respectively. The UKBB figure aligns closely with the UK population average of 10.0%[381], while GEL and GnH show clear enrichment.

Similarly, cancer prevalence patterns reflect biobank-specific recruitment strategies. In GEL, 7.3% (4,276 cases) were diagnosed with breast cancer, nearly double the 3.7% observed in UKBB. This elevation aligns with GEL's focus on cancer patients as part of its rare disease and cancer sequencing initiative[347,348]. These examples highlight how the design and focus of each biobank can lead to significant variations in disease prevalence, underscoring the importance of understanding these factors when interpreting and comparing data across different biobank resources.

4.3.2 Characterising biallelic variation across biobanks

Phasing biobanks in the BRaVa consortium

To facilitate the identification of CH variants, each biobank undertook efforts to statistically phase rare (MAF$<$ 0.05) variants in ES dataset. Recognizing that phasing accuracy can be substantially enhanced by combining ES and array data, UKBB, GnH, and BBJ jointly phased ES and genotyping arrays using `SHAPIT5`[141]. After phasing, each biobank restricted the analysis to confidently called genotypes, defined as those with a posterior probability (PP$\geq$0.90) on the genotype entry level. This step ensured a high level of confidence in CH variant calls across all participating biobanks. In the case of GEL, I leveraged quality-controlled and phased WGS data from a prior study[364] (Methods 4.2.4), as the computational demands of phasing WGS data independently were prohibitive, requiring 6,500 CPU days for autosomal phasing[364]. Moreover, the available GEL data had been phased using a rigorous multi-step approach ensuring high-quality haplotypes. Specifically, 34,617 (44.27%) individuals in trios and 12,273 (16.27%) in duos were phased using Mendelian transmission, allowing almost perfect haplotype estimation. This provided a high-confidence scaffold for subsequent statistical phasing of genotypes (MAF$>$ 1%) in the remaining 30,855 (39.5%) unrelated individuals using `SHAPEIT4`.

To support federated association analysis of biallelic variants, including both homozygous and CH variation, we conducted biobank-specific investigations into phasing accuracy across diverse ancestries (EUR, SAS, EAS) and biobanks (UKBB, GnH, and BBJ). For UKBB and GnH, we evaluated accuracy using 99 and 100 trios respectively, assessing trio-SER. As expected, singletons (MAC=1) showed the highest switch error estimates: 0.0412% (95% CI = 0.0281 − 0.0601) in GnH and 0.0563% (0.0385–0.0789) in UKBB. In contrast, common variants (MAC$\geq$ 10000) exhibited significantly lower SER: 0.0019% (0.0019–0.0019) and 0.0032% (0.0032–0.0032) respectively (Figure 4.5a-b). For BBJ, lacking trios, we employed read-backed phasing of short-read sequencing data, quantifying the agreement in phase between multiple heterozygous variants on the same sequencing read with that of statistically phased haplotypes (Methods 4.2.2). Despite having

fewer samples than UKBB (N=11,462 versus N=410,030), BBJ demonstrated high phasing accuracy across the allele frequency spectrum, particularly after filtering to PP$\geq$ 0.9 (Figure 4.5c). However, singletons remained poorly phased 0.207% (0.186–0.230), even after this filtering step. Phasing accuracy in GEL was not explicitly verified in this thesis, as all 46,890 individuals in trios or duos had been used to inform statistical phasing, a drawback of the aforementioned multi-stage phasing strategy. While we considered read-backed phasing for GEL, it was deemed too computationally intensive for WGS data.

To further validate our phasing and ensure accurate biallelic variant identification, we compared empirically observed counts of CH and homozygous pLoF events with expected counts using simulation (Figure 4.5d). While estimation of expected homozygous variants is trivial using allele frequencies, the combinatorial complexity of multiple variants within a gene, makes analytical estimation of CH variants more challenging, motivating a simple simulation-based approach. Each biobank simulated biallelic events gene-by-gene based for individual pLoF and damaging missense/protein altering variants (Methods 4.2.6). These analyses were carried out in the ancestry groups with most samples in each biobank, and did not account for LD between variants, and thus represent an approximation of the expected number of biallelic carriers. Nonetheless, our analyses revealed high concordance between observed and simulated biallelic events across all biobanks. For homozygous variants, correlations ranged from $R^2 = 0.84$ in GnH (SAS) to $R^2 = 0.99$ in UKBB (EUR), with the lower correlation in GnH due to higher parental relatedness not accounted for in our simulation. I observed high concordance for simulated and observed CH identifications across diverse biobanks and ancestries: GEL (EUR $R^2 = 0.98$), BBJ (EAS $R^2 = 0.73$), and GnH (SAS $R^2 = 0.75$). Reassuringly, in GEL, in which we did not have phasing accuracy, we observed high concordance between simulated and observed CH variants, indirectly suggesting that the phasing was of reasonable quality.

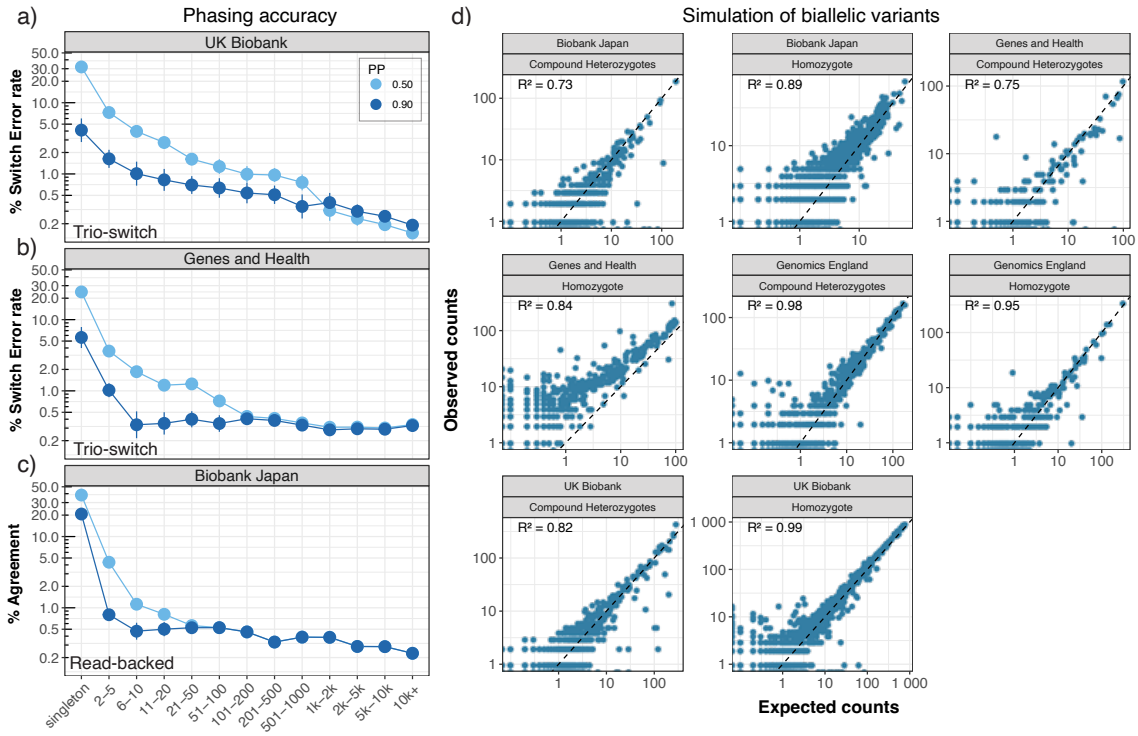


Figure 4.5: Assessment and validation of phasing accuracy through trio-switch error rates, read-backed phasing, and simulation of biallelic variants. a-c) Phasing accuracy assessment for three biobanks: The x-axis represents the MAC bin for the variants, while the y-axis shows phasing accuracy, measured by switch errors (a-b) or agreement between read-backed and statistically inferred haplotypes (c). Phasing accuracy is shown before (PP$\geq$ 0.5) and after (PP$\geq$ 0.9) filtering to confidently phased variants. CIs indicate 95% binomial confidence intervals. d) Comparison between simulated (x-axis) and empirically observed (y-axis) CH and homozygous variants for individuals of the majority ancestry in each biobank. Each point represents a gene. All biobanks used pLoF variants, except for BBJ, which used pLoF plus damaging missense/protein-altering variants due to insufficient CH pLoF variants. Note that while we did not explicitly verify SER in GEL, we observed high concordance between empirically observed and simulated biallelic variants.

Together, these results highlight the robustness of our phasing process and CH estimates across populations and biobanks, validating the reliability of our approach and supporting the use of these data in subsequent analyses.

Inclusion of compound heterozygosity increases biallelic pLoF carriers by 17%

Our study identified pLoF variants in 17,185 unique human protein-coding genes across four biobanks, with 5,226 unique genes observed in the biallelic pLoF

state. 2,077 (39%) distinct genes were observed as biallelic variant carriers in only a single individual, suggesting that most biallelic pLoF events are very rare. Without biobanks-wide phasing, studies are limited to homozygous pLoF variant identification. Under such a scenario, we would have identified 4,986 unique genes harbouring at least one homozygous pLoF variants. However, by considering compound heterozygosity, we uncovered CH pLoF variants in 1,440 genes, revealing an additional 240 genes that would have been missed through strict consideration of homozygosity.

The contribution of biallelic variants varied across biobanks, with many identified in only a single biobank (Figure 4.6). Of the 5,226 genes observed with at least one biallelic pLoF variant, the vast majority (n=3,056, 58.4%) were only observed in the homozygous or CH state in one biobank. Only 132 genes were observed in the pLoF biallelic state at least once across all four biobanks. GnH yielded the highest number of unique genes with biallelic pLoF variants through homozygosity (2,437 genes), slightly surpassing UKBB (2,446 genes) despite its smaller cohort size, likely due to higher parental relatedness in the GnH cohort[193]. Conversely, UKBB contributed the highest number of unique genes with pLoF CH variants (981 genes), followed by GEL (789 genes), and GnH (345 genes). We observed no CH pLoF variants in BBJ, probably due to its relatively small size (11,462 individuals)

To date, five studies have systematically identified and reported rare predicted biallelic LoF carriers, collectively encompassing 4,794 unique autosomal genes[191,193,227,309,340]. Our study identified 2,796 unique genes with biallelic pLoF variants not observed in any of these previous studies, thus significantly expanding the number of genes in humans with putatively observed biallelic gene inactivation.

Integrating CH variants can be particularly important for mixed-effect association analysis of rare biallelic variants, as an appreciable number of carriers are required for statistical power. Across all cohorts, including CH variants increased the number of biallelic pLoF carriers by 17% from 93,135 to 108,991 (Table 4.3). I observed increases of 19%, 14.3%, and 20.5% in the number of identifiable carriers for GEL, GnH, and UKBB, respectively. By incorporating CH variants, we expanded the

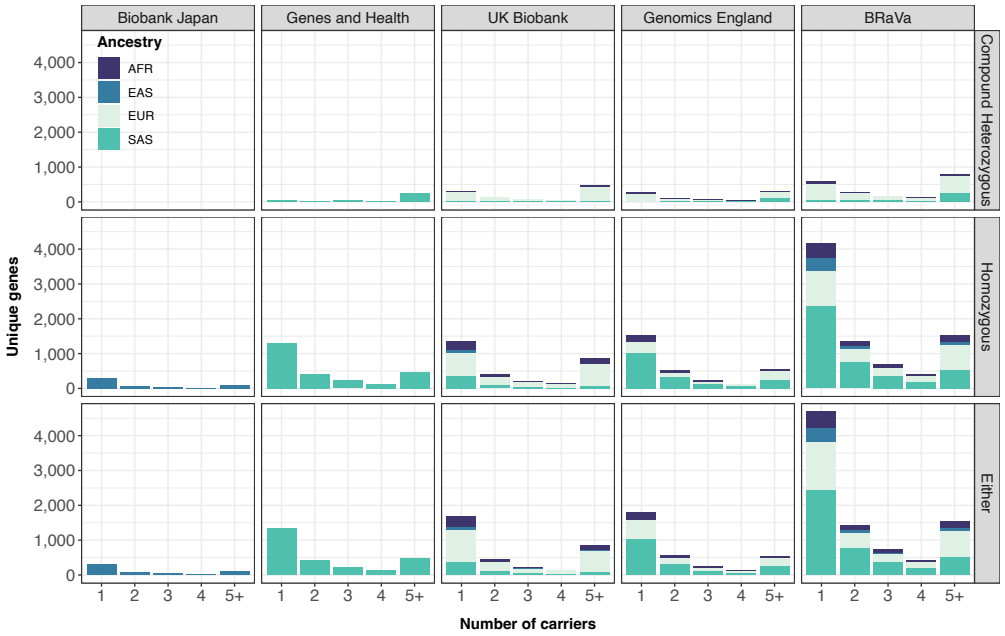


Figure 4.6: The compound heterozygous and homozygous pLoF burden across biobanks. Illustrates the number of unique genes with biallelic pLoF variants (y-axis) identified at a specific carrier count (x-axis). It can be observed, that most biallelic variants are only observed in a single individual. Counts are stratified by compound heterozygosity (top row), homozygosity (middle row) or either of the two (last row) across biobanks and the combined BRaVa pilot investigation.

number of genes with at least five carriers from 1,137 to 1,244, allowing us to test an additional 9.4% unique genes (Figure 4.6).

Genes affected with biallelic pLoF variants are less likely to be drug targets

The concept of 'healthy human knockouts'—individuals without clear disease manifestations despite having biallelic gene inactivation—has been proposed as a potential source of *in vivo* evidence for the safety of therapeutic target inhibition[337]. To investigate human knockouts in the context of drug development, I compiled a list of approved, investigational, and withdrawn gene targets from DrugBank[382], categorizing them into small molecule drugs and biologics, and investigated enrichment among genes with at least one biallelic pLoF carrier across biobanks (Table C.2). Contrary to the initial hypothesis that knockouts might be enriched among drug targets due to potentially favorable safety profiles, I observed a significant *depletion* of knockouts among approved (Two-sided fisher's exact test: OR=0.76 (95%

Table 4.3: Number of carriers with at least one biallelic genotype in BRaVa stratified by ancestry. Number of unique individuals that have at least one homozygous, CH or variants in *in cis* across the biobanks UK Biobank (UKBB), Genes and Health (GnH), Genomics England (GEL) and Biobank Japan (BBJ) stratified by ancestry. The first column indicates the variant type and annotation, with 'damag' is short for 'damaging missense/protein altering' functional category.

variant:annotation	UKBB				GNH	GEL				BBJ
	AFR	EUR	EAS	SAS	SAS	AFR	EUR	EAS	SAS	EAS
chet:pLoF	1,061	27,410	72	491	5,151	891	11,276	102	1,722	1,560
chet:damag	1,936	48,259	113	856	8,702	1,214	16,989	157	2,550	2,257
chet:pLoFIdamag	1,931	46,325	185	1,085	10,534	566	8,029	96	2,842	1,548
hom:pLoF	3,341	84,899	312	2,054	18,645	1,503	26,564	242	5,153	5,034
hom:damag	4,065	113,115	492	2,213	14,664	500	7,215	41	652	3,662
cis:pLoF	2,445	117,392	431	2,125	13,338	1,334	24,176	149	2,928	5,112
cis:damag	5,493	221,453	855	4,069	26,017	1,665	33,803	241	4,119	7,711
cis:pLoFIdamag	1,061	27,410	72	491	5,151	891	11,276	102	1,722	1,560

Table 4.4: Number of observed biallelic genotypes observed across across ancestries. Total number of observed biallelic genotypes across biobanks: homozygous, CH or variants *in cis*. The first column indicates the variant type and annotation, with 'damag' indicating damaging missense/protein altering variants. The biobanks included are cross the following biobanks: UK Biobank (UKBB), Genes and Health (GnH), Genomics England (GEL), and Biobank Japan (BBJ)

variant:annotation	UKBB				GNH	GEL				BBJ
	AFR	EUR	EAS	SAS	SAS	AFR	EUR	EAS	SAS	EAS
chet:pLoF	449	9,588	14	152	1,479	227	2,415	19	332	208
chet:damag	516	28,466	72	516	5,563	1,225	12,831	111	2,004	1,673
chet:pLoFIdamag	2,311	51,627	116	927	9,937	1,926	20,314	182	3,184	2,493
hom:pLoF	2,377	49,874	205	1,247	14,793	694	8,978	112	4,837	1,744
hom:damag	2,793	96,790	386	2,793	38,316	3,398	39,854	390	18,547	7,307
cis:pLoF	6,239	134,116	571	2,659	18,366	564	7,613	43	675	4,394
cis:damag	2,552	137,353	485	2,552	16,286	2,252	32,732	177	3,914	6,565
cis:pLoFIdamag	11,778	322,976	1,170	6,392	42,394	3,632	50,668	334	6,261	12,584

CI $= 0.65 - 0.90$), $P = 0.0011$), investigational (OR=0.78 (95% CI $= 0.66 - 0.92$), $P = 0.004$), but not withdrawn (OR=0.77 (95% CI $= 0.54 - 1.08$), $P = 0.13$) drug targets. This depletion was primarily driven by small molecule drugs (OR=0.78 (95% CI $= 0.66 - 0.92$), $P = 0.0028$), while biologics showed no significant depletion (OR=0.86 (95% CI $= 0.63 - 1.18$), $P = 0.40$). These findings suggest that the relationship between gene knockouts and drug targetability may be more complex than initially hypothesized. It could also indicate that the absence of a gene product might not always translate directly into a desirable drug effect. Finally, genes that can be knocked out without severe consequences might be under less evolutionary pressure, as evidenced by enrichment across homozygous LoF tolerant genes[227] (OR=9.17 (95% CI $= 8.13 - 10.35$), $P < 10^{-320}$), and therefore less likely to be central to disease processes that we aim to target with small molecules (Table C.3). These findings collectively suggest that the presence of knockouts for a gene is not a straightforward predictor of that gene's potential as a drug target. It is possible

4.3.3 Federated rare variant meta-analysis

Assessment of association quality across biobanks

We performed gene-based association analysis between pseudo-variants and up to 42 endpoints (32 binary and 10 quantitative traits), analyzing female participants exclusively for female-specific traits. To create pseudo-variants, each biobank amalgamated rare variants (MAF< 0.05) across one or a combination of five masks of variation, including pLoF, damaging missense/protein altering, pLoF-damaging missense/protein altering, nonsynonymous and synonymous variants. Three types of genetic models were run in parallel within each biobank: an additive model, a recessive model and a nonadditive model. These analyses were carried out on the pseudo-variant level, i.e. we collapsed variants across haplotypes in genes based on the variant annotation masks. In this context, an additive model associated the incremental count of affected haplotypes ($[0, 1, 2]$), and a recessive model associated the count of both haplotypes affected ($[0, 0, 1]$) with the trait outcome. The nonadditive model imposed orthogonality between additive and nonadditive

components of heritability by recoding genotypes at a single locus or site, and has been described in more detail in the following sections.

One way to ascertain the quality of summary statistics is to interrogate the distribution of resulting P-values[52]. Under the null hypothesis, P-values are drawn from the uniform distribution, and deviations from uniformity can potentially indicate uncorrected confounders, unresolved population stratification, or polygenicity[383–385]. For ultra-rare variants, or traits with very few cases, the possible combinations of allele counts across cases and controls are limited, leading to a discrete set of achievable P-values. The resulting distribution is therefore rarely uniform under the null, an issue which is even more pronounced for P-values closer to 1, leading to characteristic dips or 'steps' at the tail-end of the resulting QQ-plot. One implication of this is that the genomic inflation factor (λ_{GC}), calculated by dividing the median observed χ^2 statistic with the expected χ^2 statistic under the null hypothesis, may be insufficient to capture these nuances for rare variants, simply because the median observed χ^2 may suffer from this discreteness. As a result the λ_{GC} can look artificially deflated ($\lambda_{GC} << 1$), even in the presence of actual inflation. One way to address this, is by calculating the genomic inflation factor at a higher percentile, such as the 95th (λ_{95}), where we expect the test statistic to be more likely to be drawn from a uniform distribution under the null[124]. Selecting a higher percentile is a trade-off. Assessing λ_{95} works best for traits with few causal variants, whereas for polygenic traits, we expect to see inflation at higher percentiles due to the presence of many true, small effects. This inflation may reflect genuine departure from the null hypothesis rather than an artifact of the analysis method.

Given this, I set out to assess the quality of summary statistics by calculating both λ_{GC} and λ_{95} for the combinations of 42 traits, 4 biobanks, 4 variant annotations across available ancestries. As expected, I observed deflation in λ_{GC} for recessive tests compared to additive tests, particularly when assessing masks for pLoF with damaging missense/protein altering (all $\lambda_{GC} > 0.11$), and even synonymous variants ($\lambda_{GC} > 0.19$) (Figure 4.7a). As explained previously, this deflation can likely be attributed to the lower allele count in recessive tests, which model only

biallelic variants, whereas additive tests model individual haplotypes, the count of which is more common than biallelic variants. Importantly, when assessing λ_{95}, I observed a substantial improvement in the calibration of the test statistic, with pLoF+damaging missense/protein altering variants showing all $\lambda_{95} > 0.65$ and synonymous variants showing all $\lambda_{95} > 0.81$ (Figure 4.7b). This dramatic improvement with λ_{95} demonstrates its effectiveness in addressing the discreteness issue inherent in rare variant analyses. Moreover, when assessing only binary traits using the recessive model with pLoF+damaging missense/protein altering variant mask, no single traits showed evidence of inflation of the test statistic across biobanks (All $\lambda_{95} < 1.14$). However, when including quantitative traits, I observed the highest inflation across all traits and biobanks for 'Height' in GnH ($\lambda_{GC} = 1.20$, $\lambda_{95} = 1.23$), which may be explained by its well-documented polygenic architecture[19,383,386]. Of note, height was not analyzed in UKBB. The patterns discussed in this section were consistent across the remaining annotation masks (Appendix Figure C.1-C.2).

Finally, given that 3 of 4 biobanks included in this analysis are located in the United Kingdom (UKBB, GEL and GnH), there was increased possibility of shared samples and/or cryptic relatedness between the cohorts, which can lead to biased meta-analysis estimates[124]. To test for this, I compared Z-scores of synonymous variants derived under the additive model, and performed a pairwise bootstrapped Pearson correlation analysis (Methods 4.2.12). For synonymous variants, which are generally expected to follow the null, if sample overlap and/or cryptic relatedness is present, the Z-scores will be correlated[124]. Across biobanks, I observed no significant ($P > 0.05$) correlation in Z-scores from synonymous variants, suggesting limited or no evidence of sample overlap between the cohorts (Appendix C.4).

40 putatively recessive gene-trait associations are identified in meta-analysis of 5 biobanks

Given the well-calibrated summary statistics provided by each biobank, with little or no sample overlap, I performed a fixed-effects meta-analysis across 42 endpoints (32 binary and 10 quantitative traits) and pseudo-variants using four different variant annotation masks: pLoF, damaging missense/protein altering, pLoF combined

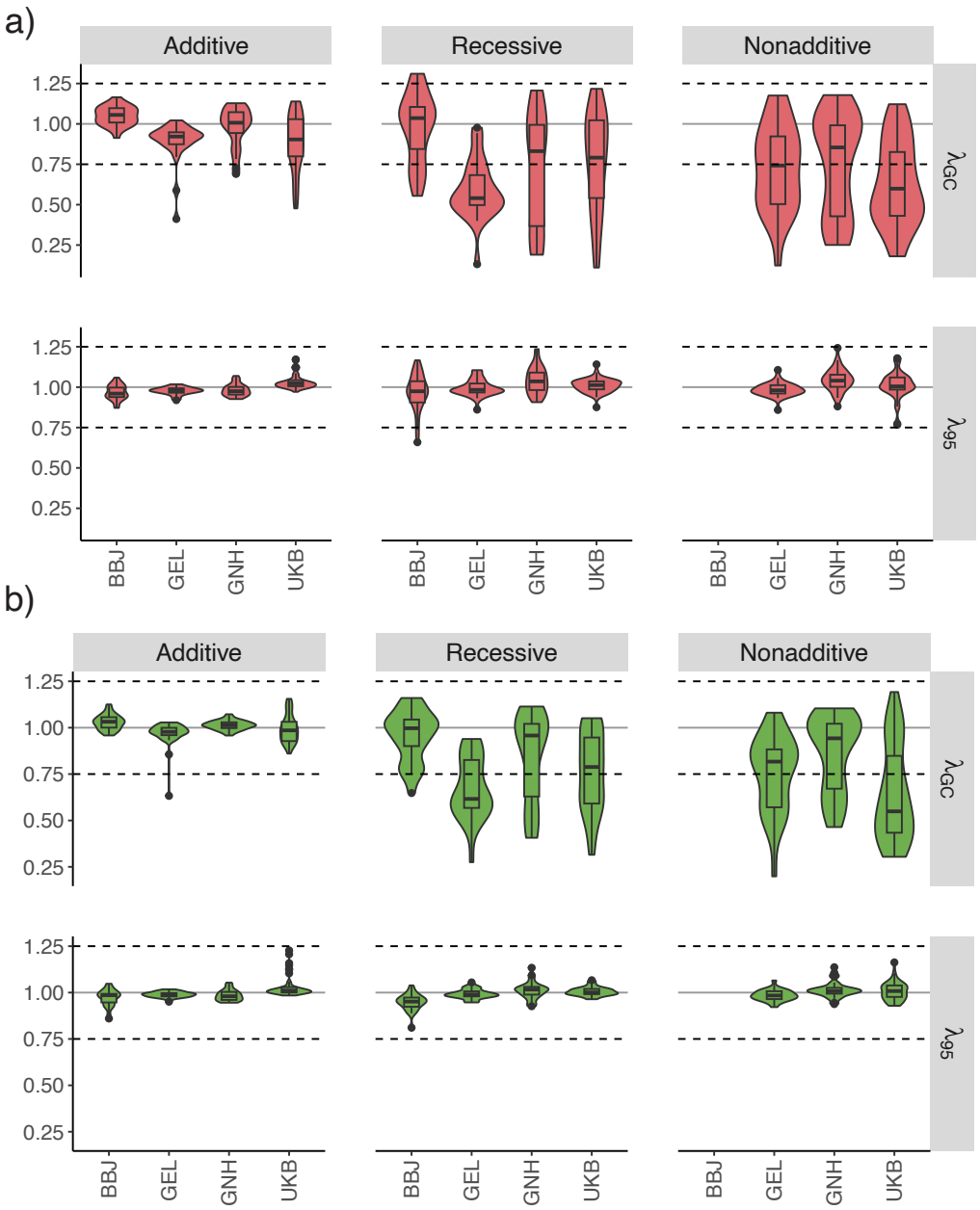


Figure 4.7: Distribution of λ_{GC} and λ_{95} before meta-analysis across investigated variant annotation masks. Violin plots with overlaid boxplot showing the distribution of λ_{GC} and λ_{95} (lambda estimated at the 95th percentile) for additive, recessive and nonadditive analyses across biobanks. For biobanks with multiple ancestries (UKBB and GEL), the genomic inflation factors are evaluated across all ancestries. a) showcasing inflation for pLoF+damaging missense/protein altering variants, and b) depicting inflation for synonymous variants. Dashed lines indicate the $\lambda = 1.25$ and $\lambda = 0.75$ cutoffs on the y-axis. The violin plots depict the probability density, while the boxplot illustrates the median, quartiles (Q1 and Q3) as hinges, and whiskers extending to $Q1 - 1.5\times$interquartile range (IQR) and $Q3 + 1.5\times$IQR. Data points beyond the whiskers (outliers) are plotted individually.

with damaging missense/protein altering, and nonsynonymous variants. I also included synonymous variants as a negative control. I then combined P-values for gene-trait pairs across annotation masks (excluding synonymous variants) using the Cauchy distribution (Methods 4.2.10). This yielded a single, P-value for each gene-trait association.

I first considered the meta-analysis under the recessive ($[0, 0, 1]$) genetic model. The meta-analyzed traits demonstrated reasonable calibration of the test statistics, and I consistently observed no inflation of a single trait for Cauchy combined P-values (all $0.78 < \lambda_{95} < 1.16$, Figure 4.8). Similar to the analysis of each individual biobank, the most inflated trait was Height ($\lambda_{95} = 1.16$), followed by HDL cholesterol ($\lambda_{95} = 1.06$). The calibration was consistent across annotation categories, including the synonymous variant mask, with $0.77 < \lambda_{95} < 1.23$, still driven by height (damaging missense/protein altering mask $\lambda_{95} = 1.23$). Manual inspection of QQ-plots confirmed a well calibrated recessive meta-analysis (Figure X, Appendix Figures C.6-C.10).

The analysis encompassed 42 traits across up to 16,279 genes, resulting in 615,368 total tests after Cauchy combination. To account for multiple testing, I applied a Benjamini–Hochberg correction with FDR< 0.01. The least significant P-value meeting this threshold was 5.87×10^{-7}. For comparison, a conservative Bonferroni correction would have imposed a more stringent P-value cutoff of $\frac{0.05}{16{,}279 \text{ genes} \times 42 \text{ traits}} \approx 5.85 \times 10^{-8}$. Using the Cauchy-combined P-values under a recessive model, I identified 40 significant associations (FDR< 0.01) across 37 unique genes and 18 traits (Figure 4.9, Table 4.5). No gene-trait associations in the sex-specific analysis passed our significance threshold. The analysis of the synonymous variant mask yielded two significant gene-trait associations at FDR< 0.01. These were C-reactive protein and *TOMM40* (Recessive $P = 1.74 \times 10^{-13}$), and LDL Cholesterol and *TOMM40* (Recessive $P = 2.09 \times 10^{-16}$). I hypothesize that these associations are a result from genetic tagging, which would be in agreement with observations elsewhere. For example, recent work by Koyama *et al.* found that associations of lipid traits across a million individuals resulted in a large number of

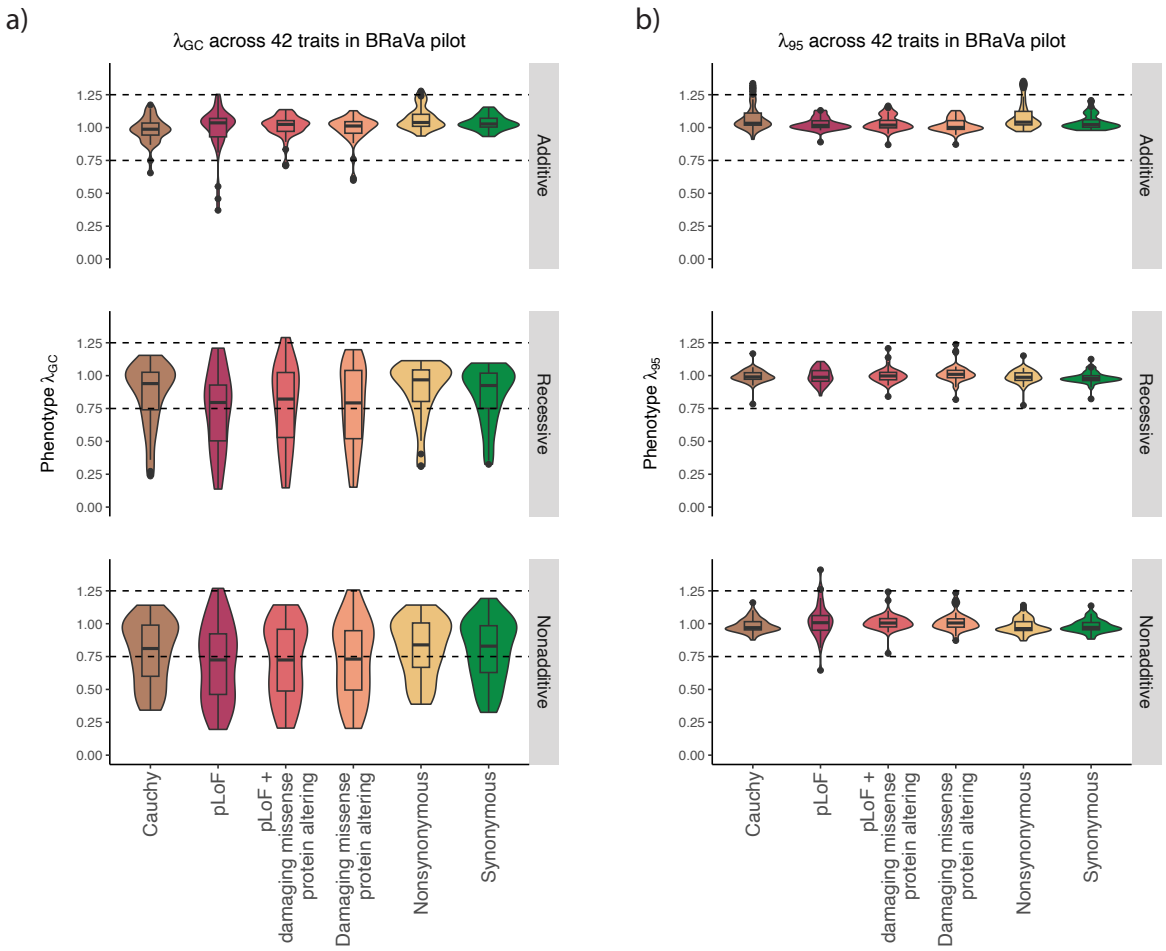


Figure 4.8: λ_{GC} and λ_{95} for additive, recessive and nonadditive meta-analyses by variant annotation masks. Violin plot with overlaid boxplot showing the distribution of inflation across additive, recessive and nonadditive analyses on the y-axis for (a) λ_{GC} (genomic control lambda) and (b) λ_{95} (lambda estimated at the 95th percentile instead of 50th percentile). The x-axis represents different variant annotation masks, including combinations of p-values using the Cauchy distribution. Dashed lines indicate the $\lambda = 1.25$ and $\lambda = 0.75$ cutoffs on the y-axis. The violin plot's outer shape depicts the probability density, while the boxplot visualizes the median, quartiles (Q1 and Q3 as hinges), and whiskers extending to $Q1 - 1.5\times\text{IQR}$ and $Q3 + 1.5\times\text{IQR}$. Data points beyond the whiskers (outliers) are plotted individually.

synonymous findings for single-variant analysis[125]. This is also in agreement with rare variant gene-based analysis by Karczewski *et al.*[251] , which reports many lipid traits (Triglycerides, HDL cholsterol and C-reactive protein) associated with burden of synonymous variants. In fact, *TOMM40* has been associated with C-reactive protein (Genebass $P = 1.7 \times 10^{-18}$) under an additive model. Lastly, I observed that 8 of 40 (20%) gene-traits using the nonsynonymous mask showed heterogeneity in effect size (Heterogeneity P $\leq 0.05/40$, Appendix Table C.5), whereas for the pLoF variant mask 0 of 8 showed heterogeneity in effect size. This could suggest either lack of pLoF variants across cohorts or more consistent direction of effects from pLoF variants compared to nonsynonymous variant.

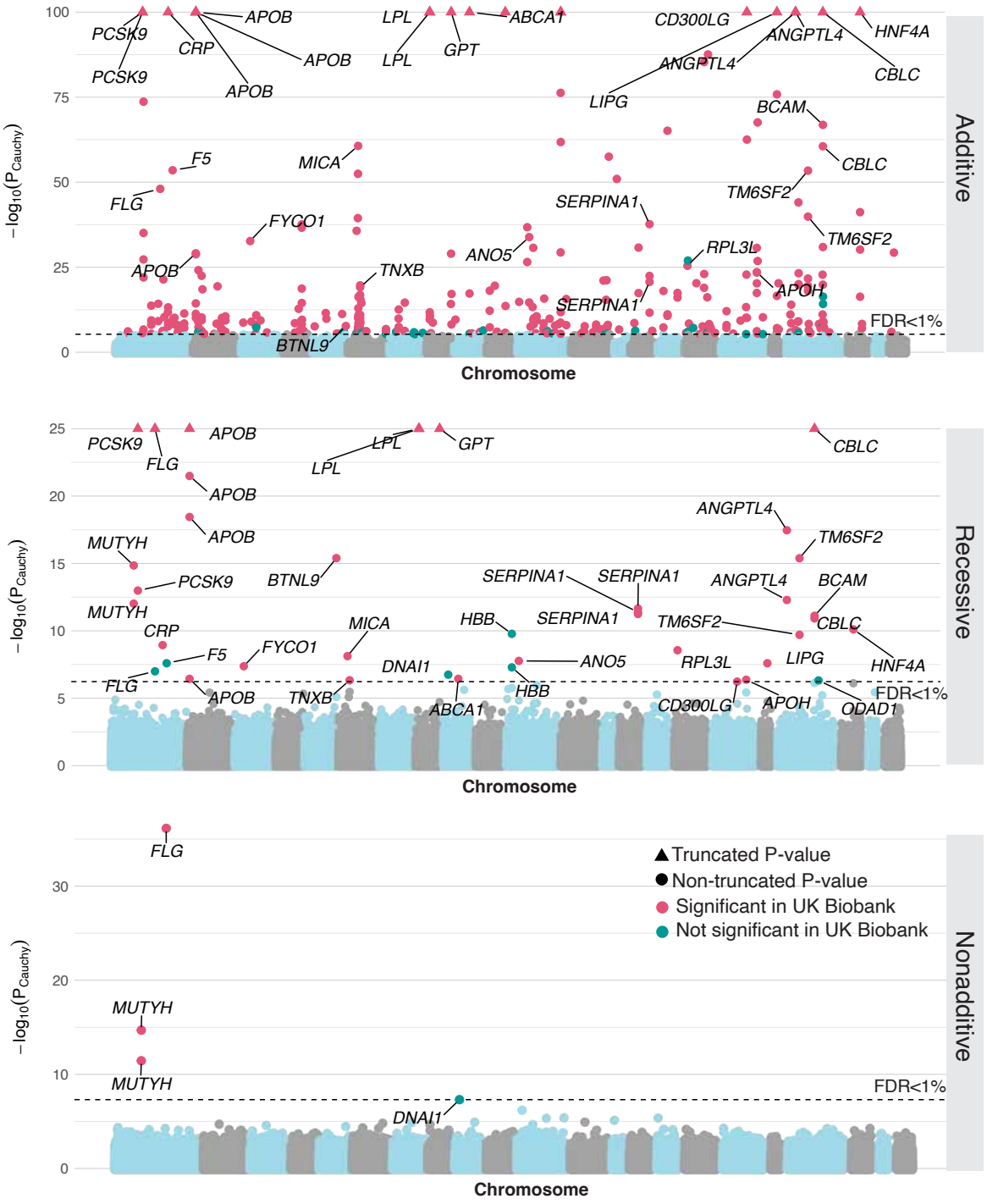

Additive
Recessive
Nonadditive
$-\log_{10}(P_{Cauchy})$
Chromosome
FDR<1%
▲Truncated P-value
●Non-truncated P-value
Significant in UK Biobank
Not significant in UK Biobank
PCSK9
APOB
CRP
LPL
ABCA1
GPT
CD300LG
ANGPTL4
HNF4A
CBLC
LIPG
BCAM
F5
MICA
FLG
FYCO1
SERPINA1
TM6SF2
ANO5
RPL3L
TNXB
APOH
BTNL9
MUTYH
HBB
DNAI1
ODAD1

Figure 4.9 ***(previous page)*****: Manhattan plot for additive, recessive and nonadditive meta-analysis**. Manhattan plot for additive, recessive and nonadditive meta-analysis across four biobanks and up to four super ancestry groups. The meta-analysis was performed first across four different variant annotation masks, followed by combination of P-values using the Cauchy distribution, which are plotted here on the Y-axis. Each point represents a gene-trait association. Points are colored red if the gene-trait associations have been deemed significant (FDR<0.01) in an analysis of Europeans in UKBB alone, and are otherwise colored green. To aid visualisation, P-values are truncated in the additive and recessive facets. Such truncated P-values are denoted by a triangle, and non-truncated P-values are circles. For the nonadditive analysis BBJ was not included in the meta-analysis, and across the included biobanks. Additionally, for the nonadditive analysis only Cauchy-combined P-values from pLoF and damaging missense/protein altering masks were considered.

To assess the impact of meta-analysis versus individual biobank analysis, I investigated how many associations would have been discovered if each biobank were to be analyzed separately. I recalculated the FDR for each gene-trait association within individual biobanks across each ancestry, accounting for the different number of genes tested in each biobank. Following the methodology used in the meta-analysis, I combined P-values across annotation masks using the Cauchy distribution. As expected, UKBB, the biobank with the largest sample size, yielded the highest number of significant associations (n=37) after Cauchy-combination of P-values under a recessive model. A putative recessive association between *FLG* and COPD (Recessive $P = 1.79 \times 10^{-7}$), was the only one that did not reach significance in any individual biobank, but emerged as significant through recessive meta-analysis (Table 4.5). This suggested that many of the observed associations in our meta-analysis would have been significant in an analysis of UKBB alone (Figure 4.9). The other biobanks reported fewer significant associations: GnH (SAS) would have identified 4 significant associations, GEL would have found 3 associations (1 in SAS and 2 in EUR populations), and BBJ would have detected a single significant association across the traits investigated (Figure 4.10).

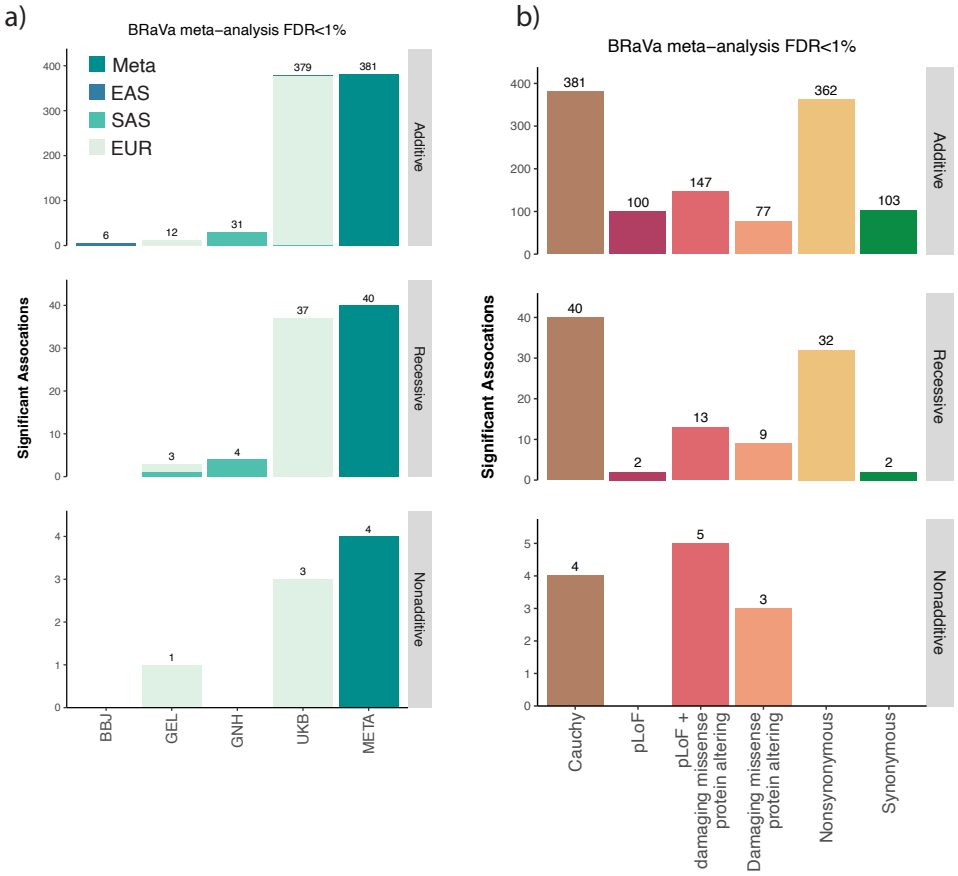


Figure 4.10: Overview of significant associations by biobanks, ancestry, and variant annotation. (a) Number of significant (FDR<0.01) hits discovered in each biobank by ancestry for additive,recessive and nonadditive analyses. (b) Number of significant hits in meta-analysis by individual rare variant masks, and combination of P-values across masks (except synonymous) using the Cauchy distribution.

4.3.4 Identification of recessive effects through meta-analysis

Identifying recessive associations through comparison of genetic models

In global meta-analyses, streamlining processes, and minimizing collaborator effort are crucial for encouraging participation. This often limits the complexity of analyses that can be performed. Given these constraints, I sought to identify recessive genetic effects using only additive and recessive P-values, as more sophisticated analyses may not always be feasible to carry out. One method described by Heyne *et al.*[192] for determining whether a gene-trait is more likely to follow a recessive mode of inheritance involves comparing the fit between an additive ($[0, 1, 2]$) and a recessive ($[0, 0, 1]$) model for the same gene-trait pairs. They propose a heuristic, that a

gene-trait is often considered to be under a recessive mode of inheritance if the recessive association P-value is more than two orders of magnitude smaller than its additive counterpart for the same gene.

Given that running a recessive and additive analysis is straightforward for collaborators, as well as comparing the resulting meta-analyzed P-values under each model, I set out to investigate this approach in our cohorts. To facilitate the analysis, I leveraged the additive analysis, i.e. collapsing of variants by individual haplotypes, and associating the number of gene copies (or haplotypes) with the 42 trait in each biobank. This was followed by meta-analysis across annotation categories for the same 16,285 genes investigated in the recessive analysis. The resulting additive P-values were also combined using the Cauchy distribution, and then compared to the corresponding Cauchy-combined recessive P-values derived in section 4.3.3. Similar to the recessive analysis, the additive analysis was also well calibrated (median Cauchy-combined $\lambda_{GC} = 1.02$ across traits) as illustrated by QQ-plots of Cauchy-combined P-values and across individual variant annotation masks (Figure C.6). In this additive analysis, I found 691 significant gene-trait (470 genes, 39 traits) associations at FDR<0.01 (Figure 4.9). I did observe 103 synonymous gene-trait associations (Figure 4.10), potentially due to tagging resulting from a lenient MAF$<5\%$ cutoff imposed across biobanks. Synonymous associations between cholesterol-related traits were observed elsewhere: for HDL cholesterol which exhibited the highest (n=32) number of synonymous additive associations, 16 of the same associations were reported in Genebass[251] with synonymous Burden $P < 0.05$ despite only assessing variants at MAF$<1\%$ (Appendix C.6), for example for *CD300LG* (Genebass synonymous Burden $P = 3.2 \times 10^{-22}$) or *SIDT2* (Genebass synonymous Burden $P = 3.2 \times 10^{-22}$). Moreover, 43 of 103 gene-traits were also associated at $P < 0.05$ when considering nonsynonymous variation in our own analysis. Collectively, this suggested that many of the observed synonymous associations were likely due to tagging, which was expected under an additive model, especially at higher MAF cutoffs.

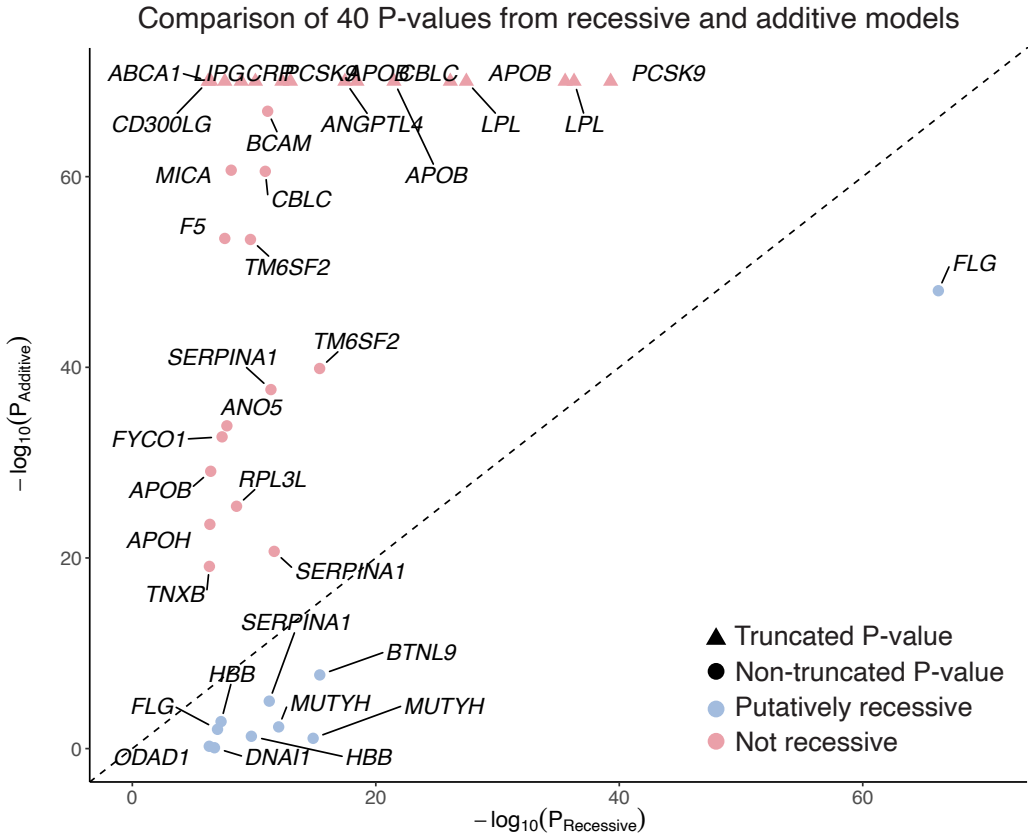


Figure 4.11: Determining recessive mode of inheritance by comparing additive and recessive P-values. Scatter plot of P-values from recessive (x-axis) vs. additive (y-axis) genetic models. Associations with recessive P-values two or more orders of magnitude smaller than additive P-values are considered putatively recessive (blue). Others are non-recessive (red). Dashed line indicates $x = y$.

I compared Cauchy-combined additive and recessive P-values from the meta-analysis. Of the 40 original gene-traits discovered, 10 (7 genes, 6 traits) exhibited more significant recessive P-values at more than two orders of magnitude (Figure 4.11). Five of these putatively recessive significant (FDR<0.01) gene-trait associations have previously been reported in OMIM[239], or had plausibly related entries, under a recessive mode of inheritance (Table 4.5). For example, of the 7 binary recessive traits identified, 5 had entries, including COPD with *ODAD1* (Recessive $P = 4.8 \times 10^{-7}$) and *DNAI1* (Recessive $P = 1.79 \times 10^{-7}$) both reported to be associated with autosomal recessive forms of primary ciliary dyskinesia, and COPD-*SERPINA1* (Recessive $P = 5.65 \times 10^{-12}$) reported to be associated with emphysema due to alpha trypsin deficiency. Of the 10 recessive gene-trait associations reported

in the meta-analysis, only a single was not significant in any single biobank, namely COPD-*FLG* (Table 4.5). I also observed recessive associations in a single biobank: the association between COPD with *ODAD1* and *DNAI1* was only observed in GEL. Additionally, we observed a recessive association between biallelic variants in *HBB* and Total cholesterol (Recessive $P = 5.16 \times 10^{-8}$) exclusively in GnH.

To better understand the contributions of individual biobanks, ancestries, and variant annotations to our findings, I created forest plots for six of the seven gene-trait associations identified, excluding benign neoplasm of intestines due to its phenotypic correlation with colorectal cancer (Figure 4.12). These visualizations revealed that, with the exception of *DNAI1* and COPD, 5 of 6 gene-trait associations demonstrated at least nominally significant associations ($P < 0.05$) with complementary directions of effect across multiple biobanks and/or ancestries. The inclusion of nonsynonymous and pLoF+damaging missense/protein altering variants was also complementary, as illustrated by *ODAD1*-COPD: In UKBB only nonsynonymous variants achieved significance, owed to a lack of biallelic pLoF+damaging missese/protein altering variants. In contrast, in GEL only pLoF+damaging missense/protein aletering annotation mask achieved nominal significance (Figure (Figure 4.12). Interestingly, this was only observed in individuals of Southeast Asian ancestry across both UKBB and GEL, suggesting that *ODAD1*-related COPD is caused by alleles with higher frequencies in individuals of Southeast Asian ancestry.

Investigation of nonadditive effects across three biobanks

While comparing P-values derived from additive and recessive models can potentially lead to identification of recessive effects, more robust approaches are needed to reliably identify gene-traits with a recessive mode of inheritance. As discussed in chapter 3, one approach involves jointly modeling the allelic encoding $[0, 1, 0]$ with the additive encoding $[0, 1, 2]$. This has been used to provide 'recessive' or dominance deviation P-values[259], but such a framework is susceptible to additive spillover effects[179]. Instead, we can leverage the method developed in chapter 3, by imposing orthogonality between additive and nonadditive components of

Table 4.5: Overview of the 40 associations discovered through meta-analysis of the recessive genetic model. P-values are shown for the recessive, additive, and nonadditive model. The column 'mode' indicates whether the P-value derived from recessive model is more than two orders of magnitude more significant than the P-value from the additive model for each gene-trait pair. Genes with plausibly related OMIM entries are highlighted alongside their mode of inheritance, either autosomal recessive (AR) or autosomal dominant (AD). The final three columns indicate whether the gene-trait associations would have been discovered through analysis of a single biobank alone at a FDR<0.01 cutoff. For the nonadditive P-values, some entries are designated as 'n/a'. These did not converge SPA in `SAIGE`, and the P-values have therefore been omitted.

TRAIT	SYMBOL	OMIM	Putative mode of inheritance	Meta-analaysis Cauchy *P*			Single biobanks with FDR < 1%		
				Recessive	Additive	Nonadditive	Additive	Recessive	Nonadditive
Afib	*RPL3L*	619371 (AR)	Additive	2.76E-09	3.79E-26	2.02E-01	UKB	UKB	
ALT	*GPT*		Additive	1.17E-71	7.50E-305	8.46E-02	GNH, UKB	UKB	
ALT	*SERPINA1*		Additive	4.12E-12	2.18E-38	1.67E-01	UKB	UKB	
AST	*APOH*		Additive	4.27E-07	2.99E-24	6.20E-01	UKB	UKB	
AST	*FYCO1*		Additive	4.25E-08	1.98E-33	6.81E-01	UKB	UKB	
AST	*ANO5*		Additive	1.71E-08	1.37E-34	2.19E-01	UKB	UKB	
AST	*SERPINA1*		Additive	2.22E-12	2.05E-21	2.63E-02	UKB	UKB	
Asth	*FLG*		Recessive	6.13E-67	9.26E-49	7.20E-37	GEL, UKB	GEL, UKB	UKB
Asth	*TNXB*		Additive	4.67E-07	7.69E-20	8.66E-01	UKB	UKB	
BenIntNeo	*MUTYH*	608456 (AR)	Recessive	9.55E-13	5.28E-03	3.58E-12	n/a	UKB	UKB
COPD	*ODAD1*	615067 (AR)	Recessive	4.80E-07	5.70E-01	4.23E-02	-	GEL	
COPD	*DNAI1*	244400 (AR)	Recessive	1.79E-07	8.11E-01	4.87E-08		GEL	GEL
COPD	*FLG*		Recessive	1.01E-07	9.50E-03	3.46E-04			
COPD	*SERPINA1*	613490 (AR)	Recessive	5.65E-12	1.05E-05	3.57E-01		UKB	
CRP	*CRP*		Additive	1.14E-09	2.25E-304	5.67E-02	GNH, UKB	UKB	
ColonRectCanc	*MUTYH*	608456 (AR)	Recessive	1.41E-15	8.32E-02	2.03E-15		UKB	UKB
HDLC	*APOB*		Additive	3.65E-07	8.10E-30	2.89E-01	UKB	UKB	
HDLC	*HNF4A*		Additive	7.75E-11	6.89E-116	3.70E-01	GNH, UKB	UKB	
HDLC	*LIPG*		Additive	2.55E-08	1.30E-276	1.33E-01	UKB	UKB	
HDLC	*CD300LG*		Additive	5.88E-07	1.23E-111	7.28E-01	GNH, UKB	UKB	
HDLC	*ABCA1*	604091 (AD)	Additive	3.64E-07	4.38E-173	2.98E-01	GNH, UKB	UKB	
HDLC	*BTNL9*		Recessive	4.02E-16	1.87E-08	4.46E-02	UKB	UKB	
HDLC	*ANGPTL4*		Additive	5.15E-13	4.14E-122	2.09E-01	UKB	UKB	
HDLC	*LPL*		Additive	5.17E-37	2.25E-304	7.29E-02	GNH, UKB	UKB	
HDLC	*HBB*		Recessive	1.67E-10	5.10E-02	6.55E-07	GNH	GNH	
LDLC	*APOB*		Additive	2.79E-36	2.25E-304	8.04E-01	GNH, UKB	UKB	
LDLC	*CBLC*		Additive	7.55E-27	3.29E-143	1.32E-01	UKB	UKB	
LDLC	*PCSK9*		Additive	5.10E-40	2.25E-304	2.09E-02	GNH, UKB	UKB	
LDLC	*BCAM*		Additive	7.65E-12	1.38E-67	7.52E-01	UKB	UKB	
LDLC	*TM6SF2*		Additive	4.11E-16	1.33E-40	n/a	UKB	UKB	
Psori	*MICA*		Additive	7.56E-09	2.11E-61	1.22E-03	GEL, UKB	UKB	
TChol	*APOB*	144010 (AD)	Additive	3.56E-19	1.46E-155	3.89E-01	GNH, UKB	UKB	
TChol	*CBLC*		Additive	1.20E-11	2.77E-61	2.10E-02	UKB	UKB	
TChol	*PCSK9*	603776 (AD)	Additive	1.03E-13	3.04E-174	6.68E-01	GNH, UKB	UKB	
TChol	*HBB*		Recessive	5.16E-08	1.46E-03	1.28E-04	GNH	GNH	
TG	*APOB*	144010 (AD)	Additive	3.23E-22	7.31E-173	n/a	UKB	UKB	
TG	*ANGPTL4*	615881 (AD)	Additive	3.45E-18	7.48E-169	n/a	UKB	UKB	
TG	*LPL*		Additive	3.64E-28	1.57E-291	n/a	UKB	UKB	
TG	*TM6SF2*		Additive	1.97E-10	3.89E-54	n/a	UKB	UKB	
VTE	*F5*	188055 (AD)	Additive	2.54E-08	3.04E-54	6.58E-01	UKB		

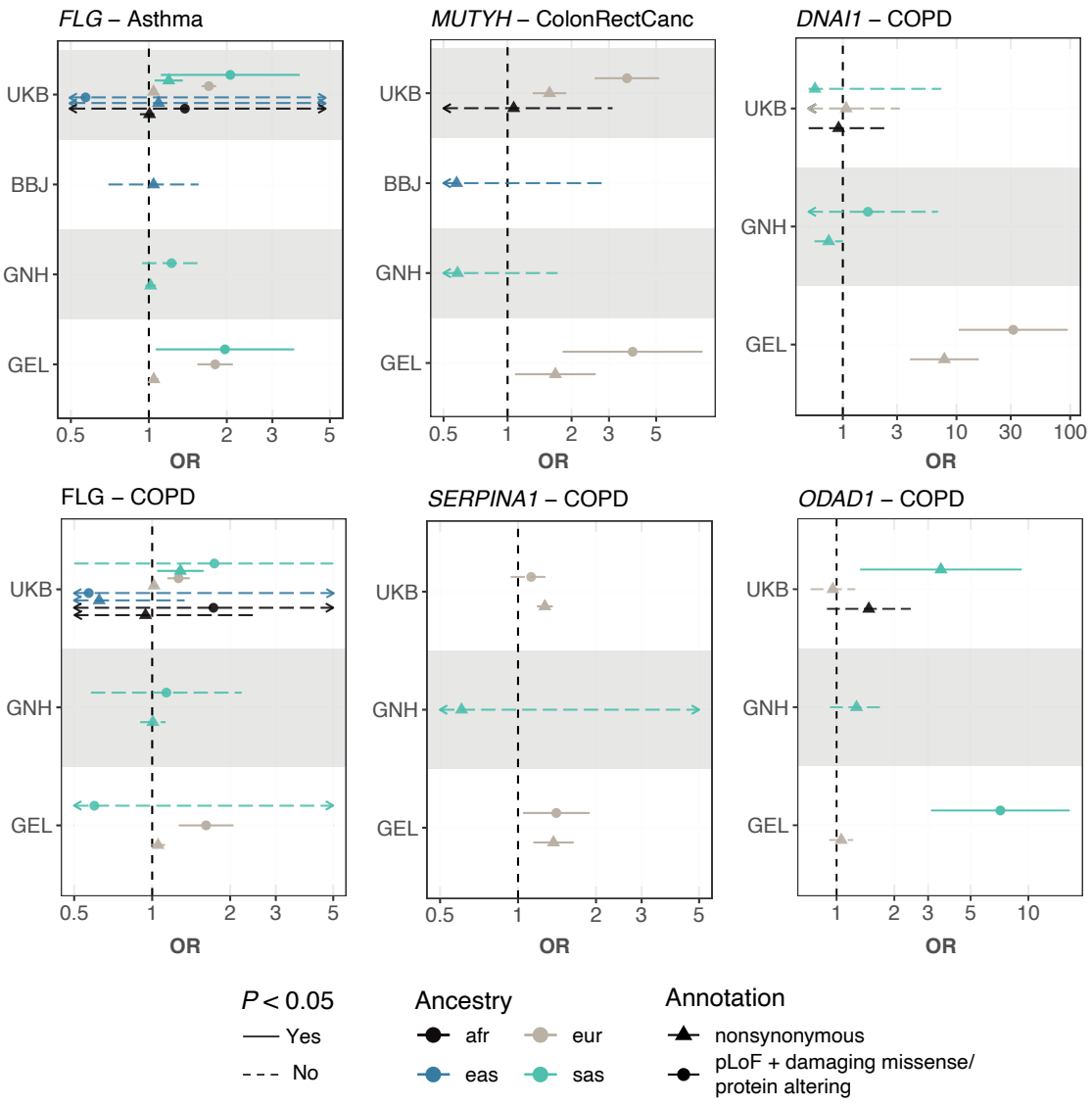


Figure 4.12: Forest plot of 6 putatively recessive gene-trait associations reported in the meta-analysis. Forest plot of ORs and 95% CIs across six of seven recessive gene-trait associations reported in the meta-analysis. The plot includes: 1) pLoF and damaging missense/protein altering variants (circles); and 2) nonsynonymous variants (triangles). The y-axis indicates the biobank, with each association stratified by ancestry and variant annotation. Associations with $P < 0.05$ are demarcated with a solid line, otherwise with a dashed line. To aid visualization, large CIs have been truncated, and arrows indicate whether the CI extends beyond the plot borders. Benign intestinal neoplasm (BenIntNeo) is not shown, in favour of depicting colorectal cancer (ColonRectCanc) associated with biallelic variants in *MUTYH*.

heritability through recoding of genotypes at each single locus. We applied this method in individual biobank by defining r, h, and a as the proportions of individuals with zero (wildtype), one (monoallelic), and two haplotypes carrying one or more alternate alleles (biallelic), respectively. Genotypes $[0, 1, 2]$ were then mapped to $[-ha, 2ar, -hr]$ in each biobank, which were associated with the trait of interest using `SAIGE` or `REGENIE` (Methods 4.2.8). P-values were then shared, and I performed fixed-effects meta-analysis using Stouffer's method. UKBB, GnH, and GEL tested for nonadditive effects gene-by-gene using this framework and shared summary statistics. At the time of writing, BBJ had not yet shared their summary statistics, and was therefore not included in the following analyses. Additionally, only pLoF and damaging missense/protein altering, and synonymous variant masks were shared and therefore tested.

I tested up to 8,047 genes across three different masks: pLoF, damaging missense/protein altering and a combination of the two, which were combined through Cauchy combinations of P-values. Similar to the recessive analysis, the nonadditive meta-analysis demonstrated decent calibration of the test statistic for Cauchy combined P-values across individual traits with $0.87 < \lambda_{95} < 1.16$ (Figure 4.8), with Height ($\lambda_{95} = 1.16$) and Triglycerides ($\lambda_{95} = 1.10$) being the two most traits with highest lambdas. From this analysis, I identified 4 significant (FDR< 0.01) gene-trait associations across 3 genes and 4 traits (Figure 4.9, Table 4.5). I did not find any nonadditive synonymous gene-trait associations; however, I did observe a single gene-trait association *HBB*-HDL cholesterol, which only passed FDR< 0.01 in the pLoF+damaging missense mask, but not in the final Cauchy-combined P-values (Figure 4.10b). Lastly, if I had only tested the 10 associations reported in section 4.3.4, disregarding the lack of BBJ data and nonsynonymous variants, 7 could potentially have been considered to have a significant nonadditive contribution (nonadditive $P < \frac{0.05}{10} = 0.005$).

4.3.5 Meta-analysis improves power for select gene-trait associations

Comparison of power between meta-analysis and analysis of indiviudal biobanks

UKBB contributed the highest number of significant associations to our analysis (Figure 4.10a). To investigate whether our meta-analysis provided an overall boost in power, I compared χ^2 statistics derived from P-values in recessive ($[0, 0, 1]$) analysis of UKBB participants of European ancestry to the χ^2 from the meta-analysis (Figure 4.13a). Of the original 40 significant (FDR$<$ 0.01) gene-trait associations, I was able to compare 38 gene-trait pairs with at least five biallelic carriers in Europeans of UKBB. Among these 38 gene-trait pairs, 20 demonstrated higher χ^2 statistics in the meta-analysis compared to UKBB. However, 8 showed no change (to one decimal place precision), while 10 had higher χ^2 in UKBB compared to the meta-analysis. This could indicate the presence of the same biallelic variants in other biobanks, but with no cases for the given trait. The highest absolute increase in χ^2 statistics was observed for the *FLG*-asthma association ($\chi^2_{\text{UKBB}} = 263.4$, $\chi^2_{\text{meta}} = 298.8$). Conversely, the largest change favoring UKBB was observed for the *LIPG*-HDL cholesterol association ($\chi^2_{\text{UKBB}} = 40.5$, $\chi^2_{\text{meta}} = 31.0$). When repeating this analysis, comparing meta-analyzed χ^2 to that of the largest ancestry group of other biobanks (Figure 4.13c-d), only *DNAI1* was more significant in Europeans of GEL compared to the meta-analysis. These results suggest that while our meta-analysis may generally have improved statistical power, the extent of improvement varied across gene-trait pairs, with some associations showing stronger signals in the UKBB alone. This does not imply that the associations identified in UKBB are false positives. Rather, it could indicate that the sample sizes in other cohorts may be insufficient to detect effects with the same magnitude as in UKBB.

Power beyond current pilot analysis

The current analysis leverages four of the seven biobanks planned to be included in the recessive meta-analysis. The remaining biobanks includes quality-controlled

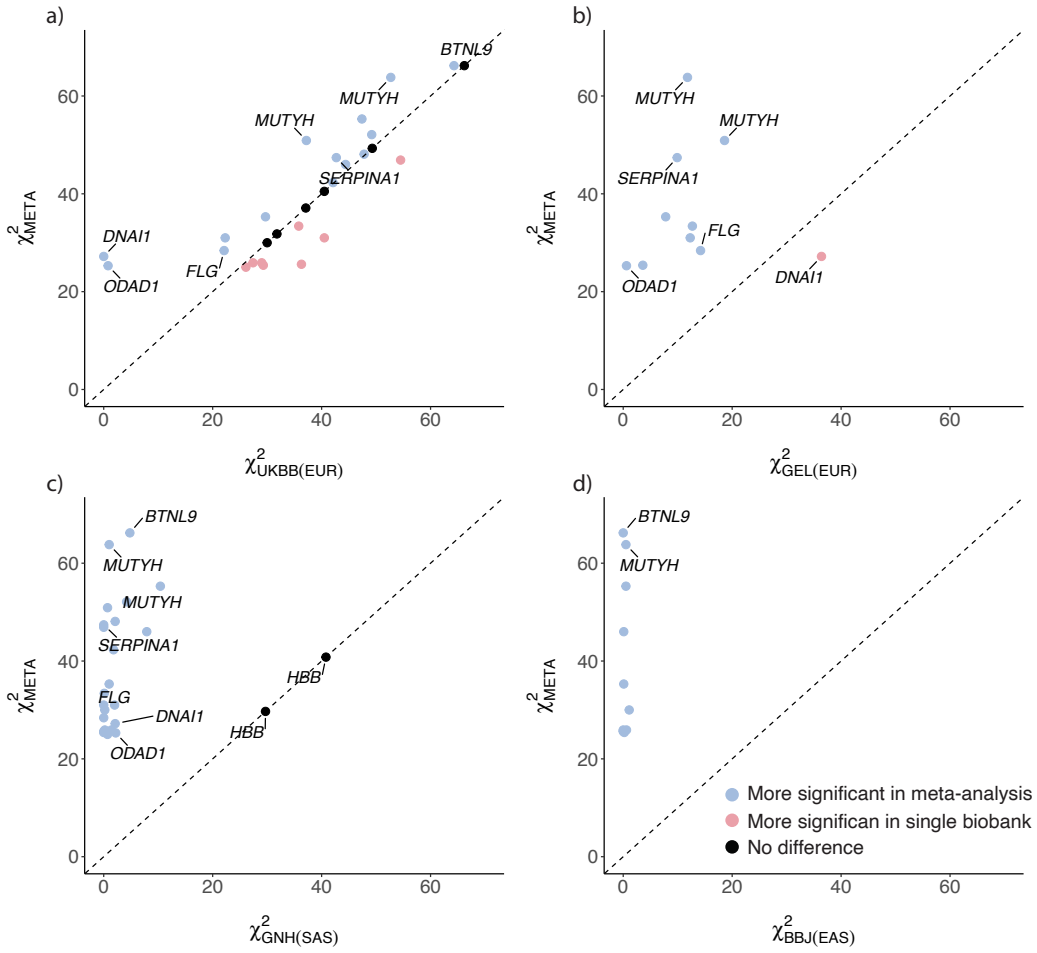


Figure 4.13: Evaluation of meta-analysis power by comparison of χ^2 across analyses. Comparison of χ^2 derived from recessive P-values from the meta analysis, with χ^2 of the a) Europeans in UKBB, b) Europeans in GEL, c) Southeast Asians in GEL and d) East Asians in BBJ. A dashed line indicate $x = y$, and where power is exactly equal between the meta-analysis and corresponding biobank. Each dot is a gene-trait associations, and up to 40 are shown in the plot. Blue points indicate gene-traits that are better powered in the meta analysis, compared to the biobank in question. Red points indicate gene-trait associations that are better powered in a single biobank alone. Black points indicate gene-traits that are equally powered in meta-analysis and corresponding biobank (down to 1-decimal precision). Note that *ODAD1*-COPD association is driven by individuals of Southeast Asian ancestry in GEL, which is not depicted in this plot.

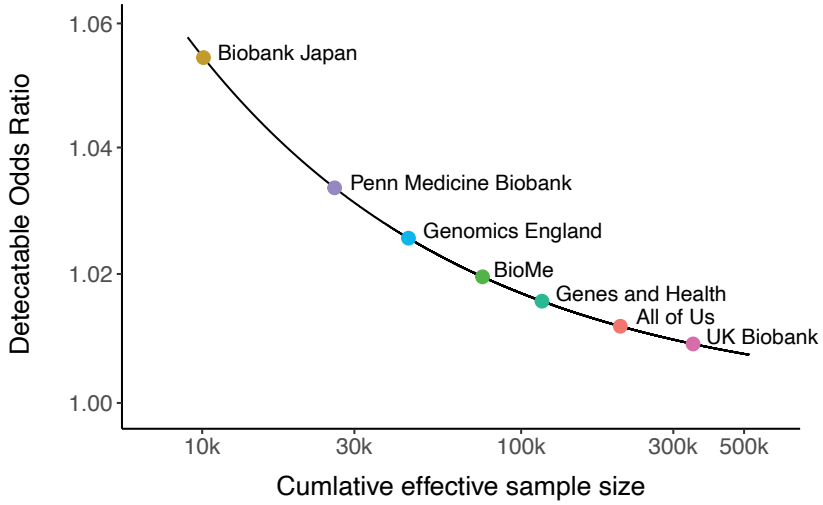


Figure 4.14: Projected lower bounds for detectable Odds Ratios with additional BRaVa biobanks. Projection of hypothetical detectable ORs for Type 2 Diabetes, as a function of increased effective sample (n_{eff}) size across biobanks in the BRaVa recessive consortium. As we increase the number of samples, we are better powered to detect lower odds ratios. Only GnH, GEL, BBJ and UKBB are included in this chapter, however All Of Us, BioMe, and Penn Medicine biobank are currently preparing summary statistics. Cases are ascertained using counts from each biobank except for in All of Us and Penn Medicine Biobank which have been based on publicly available data.

samples with phased genetic data from BioMe (28,731 individuals), Penn Medicine Biobank (43,731 indiviudals) and All Of Us (242,902 indiviudals). Together with the cohorts analyzed in this chapter, this would allow us to analyze 834,875 samples, instead of the current 519,594 samples, which corresponds to a 60% increase in sample size. This will give us increased power to analyze these 42 traits and allow us to detect even more subtle OR. However, given that power scales with the square root of effective sample size, we will begin to see diminishing returns with respect to the lowest possible OR detectable (Figure 4.14). Nonetheless, the true value of this expansion lies not just in increased numbers, but in the diversity of traits and ancestries these biobanks offer. The unique phenotypic and ancestral composition of each biobank offers an opportunity to explore the phenotypic consequences across different backgrounds and allele frequencies, potentially revealing recessive effects that are population-specific or more pronounced in certain contexts.

4.4 Discussion

In this chapter, I have presented pilot work for the BRaVa recessive subgroup, investigating up to 42 phenotypes across four biobanks, as well as providing a map of both homozygous and compound heterozygous variants across multiple biobanks, enabling a glimpse into the diversity of biallelic variation across ancestries.

Together, our pilot BRaVa analysis demonstrated the utility of phasing biobanks to infer CH variation at scale, which we validated using trio-switch error rates, read-backed phasing and simulation. I also highlighted that CH variants provided a modest 10-20% increase in the number of identifiable biallelic pLoF carriers, and that inclusion of compound heterozygosity, only allowed consideration of a small proportion of unique genes that would otherwise not have been discovered through homozygosity alone. Consistent with previous studies[188,227], we observed that biallelic gene inactivation tend to occur in genes that were tolerant to LoF mutations, making them potentially less suited for drug development. In agreement with this, I also found that current small molecule and approved drug targets from DrugBank[382] are depleted among genes with at least one biallelic pLoF variant, indicating that genes that are frequently 'knocked out' across populations, may have limited influence on human physiology and health, and are therefore not necessarily compelling drug targets.

I performed two separate analyses to identify putative recessive gene-trait associations through meta-analysis. Following a framework by Heyne *et al.*, I compared P-values derived from recessive and additive models, and deemed associations recessive if they were two orders of magnitude below the corresponding additive association P-value. One problem with this analysis is that P-values depend on power, and this approach therefore assumes equally powered additive and recessive analysis for each locus, which is likely not the case. This problem is exaggerated in the context of meta-analysis, as power may also deviate within and between multiple studies due to differences in case ascertainment and variant allele frequencies. While such model fit comparison can indeed identify recessive associations of large effects, the framework is likely not optimally designed for

identifying more subtle recessive effects. For this reason, I also carried out a meta-analysis of nonadditive effects. This analysis involved recoding genotypes to impose orthogonality between the additive and nonadditive components of heritability, allowing for independent assessment of recessive effects that were not confounded by additive spillover. Despite analyzing only three biobanks and a subset of annotation masks, I observed that of the 10 original associations deemed 'recessive' by comparison of additive and recessive models, 4 of them remained significant in the nonadditive model. While this could be due to the exclusion of BBJ in this analysis, its small sample size (n=11,462) makes this unlikely. Instead, the difference is probably due to the lack of testing of the nonsynonymous variant mask. Nonetheless, associations deemed recessive through comparison of P-values, or through nonadditive testing, should still be interpreted with great care, and require replication. Future work should be directed towards interrogating the genetic architecture of meta-analyzed gene-trait associations using the standard additive and nonadditive recoding of genotypes, which has never been accomplished before.

I identified putatively recessive associations between biallelic variants in *ODAD1* and *DNAI1* associated with COPD. Both of these genes have previously been implicated in primary ciliary dyskinesia under a recessive mode of inheritance[239]. Ciliary dysfunction is a hallmark of COPD, suggesting that the association we observe is likely to be genuine. Through meta-analysis, we identified a single significant gene-trait association between *FLG* and COPD, which was not detected as significant in any individual biobank study. This association seems plausible given the established role of *FLG* in respiratory disorders, such as asthma, which has been reported extensively in this thesis. I also observed multiple associations with lipid and cholesterol-related traits involving biallelic variants in *HBB* and *BTNL9*, both of which have recently been reported in a large-scale rare-variant meta-analysis of lipid traits, although without providing a potential explanation for the latter[125]. I speculate that the *HBB*-total cholesterol association is confounded by potential β-thalassemia treatment, which has been documented to influence lipid

concentrations[387]. This hypothesis could be further investigated by conditioning on β-thalassemia case status.

While many associations were observed across multiple biobanks, some gene-trait associations are unique to specific populations. For instance, deleterious alleles in *ODAD1* and *DNAI1* associated with COPD were only detected in individuals of Southeast Asian ancestry. This can be attributed to several factors. Genetic drift can lead to differences in allele frequencies between populations, potentially increasing the frequency of certain alleles in specific ancestries. Moreover, the statistical power to detect associations is proportional to the square of the MAF. Consequently, even a modest increase in MAF can increase our ability to detect associations, especially for recessive traits. Such observations underscore the importance of diverse cohorts and suggest that ancestry-specific analyses may reveal additional recessive effects.

In conclusion, this pilot analysis of 519,594 samples for BRaVa recessive subgroups highlights the opportunities and challenges in identifying rare recessive effects. From this analysis, it is evident that the number of associations identified through analysis of 42 traits across 4 biobanks is still low, and that larger sample sizes, and more diverse cohorts are required to meaningfully and robustly associate novel genes and traits. This is particularly true when assessing 2nd order effects such as nonadditivity. At the time of writing this discussion, two additional biobanks (Penn Medicine Biobank and BioMe) have phased ES data and are in the process of running association analysis. Similarly, I am working on phasing All of Us Biobank. By the end of the year, our collaborative efforts will yield a diverse cohort of approximately 800,000 individuals. Hopefully, with the expanded sample size, encompassing multiple ancestries, we will be able to detect and analyze recessive effects of rare variants that would be unobservable in any single biobank alone.

4.4.1 Limitations

There are several limitations to my federated analysis of biallelic variants across biobanks. Firstly, in the current fixed-effects meta-analysis, I combine P-values across biobanks, weighting each one of them by the corresponding effective sample

size (N_{eff}). An important assumption in this calculation is that N_{eff} is derived assuming all individuals are unrelated. For biobanks such as GnH with high parental relatedness, this assumption does not hold, and as a consequence, the biobank has a larger weight than it should have. Future, analysis should therefore consider more accurate estimation of sample size in such a way that accounts for the relatedness structure in the population, prior to meta-analysis, for example through assessment of the GRM. Secondly, for GEL we exclude 13,976 rare disease probands diagnosed with NDD, to reduce the likelihood of artifactual association signals arising from biallelic variants enriched in this subset of control subjects. However, as demonstrated in GEL there is also an enrichment of cancer patients, which could have similar effect on the analysis. Future work should include sensitivity analyses of both NDD and cancer patients to investigate how their inclusion or exclusion relates to inflation in the test statistic and if that leads to spurious associations. Finally, I impose a liberal cutoff of FDR$<$ 0.01 in my analysis. Because of this and the aforementioned reasons, any particular gene-trait association identified under this threshold should be subject to replication and validation in an external cohort. Following the strategy undertaken chapter 3, and in other studies[251], a potentially better way to choose a threshold is to meta-analyze null phenotypes across biobanks and use the most significant P-values as an estimator for the inverse of the number of tests.

5 Conclusions and future directions

Contents

5.1 Conclusions 191
5.2 Discussion and future directions 193
5.2.1 Long-read sequencing and decoding the non-coding genome 193
5.2.2 Nonadditive effects enable gene therapies and drug development 196
5.2.3 The case for humans as the new model organism 198
5.2.4 Final remarks 200

5.1 Conclusions

In this thesis, I have demonstrated the utility of phasing large-scale genetic datasets linked with electronic health records to infer the phenotypic consequences of biallelic inactivation. Building upon many decades of genetics research, I provide new tools and ideas to advance our understanding of additive and nonadditive biallelic effects.

In Chapter 2, I demonstrated the rationale for performing rare variant (MAF$\sim$ 0.001%) phasing in UKBB, enabling the identification of deleterious CH variants. Together with homozygous variants, CH variants expand the number of identifiable biallelic variants, which increases power in association analysis. This increase in power can be demonstrated through *in silico* permutation of the genetic phase,

to establish that CH variants, rather than co-occurring variants on the same haplotype, drive disease risk across many of the common complex traits investigated. Building on this, I conducted cross-sectional recessive association analysis and complementary age-at-onset models for up to 311 common binary traits available in UKBB. My analyses demonstrate that biallelic variants, and specifically CH variation, confer susceptibility to common complex phenotypes such as colorectal cancer, asthma, dermatitis, and hearing loss. I demonstrated that my findings are broadly reproducible by successfully replicating all reported gene-trait associations, including a novel link between biallelic variation in *ATP2C2* and COPD, in the remainder of UKBB. Together, my analyses showcase the value of interrogating CH variation in an outbred population, while demonstrating the role of biallelic variants in shaping common complex disease.

Genetic trait variance can be decomposed into independent additive and nonadditive contributions[19]. In chapter 3, I proposed a method to assess the nonadditive contribution of rare variation by imposing orthogonality between the additive and nonadditive components of heritability. Unlike previous investigations of nonadditivity[179,255,256], this method does not rely on the assumption of HWE, which enables its application to rare variation (MAF$<1\%$). I demonstrated the utility of this framework through a GWAS investigation of nonadditive effects in the recently released Olink proteomics data, followed by an analysis of hundreds of binary and quantitative traits in UKBB. I also demonstrated that the genetic architecture can be estimated by fitting additive and nonadditive models to a locus. Finally, I generalised this framework to the standard rare-variant gene-based burden framework, testing the collective variant burden of nonadditive effects using standard mixed-effects models. Lastly, I presented work that suggests that, with sufficient data, the absence of biallelic (CH and homozygotes) variation can be used to infer the presence of selection against recessive alleles. Together, this chapter proposes methods with particular relevance for assessing the presence or absence of recessive effects owed to rare protein coding variation.

My investigations into biallelic effects culminate in chapter 4 through a multi-biobank collaboration and investigation into rare recessive effects. We infer compound heterozygosity across four unique biobanks, and show a modest (10-17%) increase in number of significant gene-level recessive associations exome-wide. Through meta-analysis of additive, recessive, and nonadditive effects, I showcased the utility of meta-biobank initiatives for identification of recessive effects, and conclude the need for even larger samples sizes. Some gene-trait associations are found in specific biobanks and are sometimes only powered for detection in specific ancestries, such as deleterious alleles in *ODAD1*-COPD in those of Southeast Asian ancestry. Meta-analysis may not boost power for all gene-trait associations, particularly those with alleles private to specific ancestry group. Finally, this pilot presents an initial framework for the first gene-based meta-analysis of rare variant recessive effects, and demonstrates the diversity of biallelic variants across populations and ancestries.

Through my three chapters, this thesis contributes to the field of medical population genetics, shedding light on the mechanisms of inheritance. Throughout the next few sections, I will collectively discuss the broader implications of these findings, and propose possible directions for future research.

5.2 Discussion and future directions

5.2.1 Long-read sequencing and decoding the non-coding genome

This thesis contributes to a growing body of scientific research aimed at unraveling the phenotypic impact of biallelic variation[192,259,260]. Recent advancements in phasing methods have enabled this work, in particular the integration of exome sequencing and genotyping array data from the same individuals to increase the density of variants. While current models can now phase singletons[141], albeit with lower accuracy, I did not observe a significant contribution to gene-trait associations from these variants, possibly due to imposing the PP-criteria on phasing accuracy, which removed many singletons from the association analysis. Nonetheless, future research should assess the contribution of CH variants with at least one singleton,

as these are expected to potentially have large effects on traits. Future efforts should work towards investigating the contribution of such ultra-rare variants in shaping the deleteriousness of CH variation.

Such analyses will be enabled by innovation in sequencing technology. CH variants can in principle be resolved with near-perfect accuracy through long-read sequencing followed by read-backed phasing. Sequencing costs across all technologies are rapidly declining, including high quality long-read WGS, which now costs ∼US$1300 per sample (as of 2021, PacBio[388]). The increasing adoption of long-read sequencing in rare disease cohorts, such as GEL[94], and its prototyping in large-scale general population cohorts such as All of Us[99], is a testament towards this potential. A recent study[389] highlighted three key advantages of long-read sequencing over short-read sequencing: (1) superior detection of structural variants; (2) higher accuracy in challenging but medically relevant genomic regions; and (3) improvement in overall variant calling. Given these advancements, I hypothesise that long-read sequencing will eventually replace short-read sequencing, beginning with rare disease cohorts[390,391]. This technology will overcome current limitations in CH variant identification due to limitations in statistical phasing accuracy.

In this thesis, I demonstrated that some common complex diseases are driven by biallelic recessive contributions to the trait. Similar observations have been made in Finnish populations[192], and with initiatives such as the BRaVa consortium, we will start to map the biallelic contribution of compound heterozygosity and homozygosity to common complex diseases. These ongoing investigations have been restricted to coding variants, which is a result of both the low cost of ES and the fact that functional consequences of protein-coding changes are generally easier to predict and interpret compared to non-coding variants. This focus is further supported by the increasing number of tools that can predict deleteriousness *in silico,* including `SpliceAI`, `AlphaMissense` and `REVEL`. However, it seems unlikely that compound heterozygous effects are constrained to coding variants. Using Mendelian Transmission phasing it was recently shown that certain developmental disorders were caused by coding variants *in trans* with non-coding variants, specifically in

the 5'UTR regions across several genes. It makes sense to begin the investigation of non-coding CH effects in the these regions: 5'UTR regions are easier to interpret than intergenic regions simply because they are part of the mature RNA transcript, and variants may impact the stability of mRNA molecule[392], or on the protein sequence itself through extension of the open reading frame, or by creating 'small open reading frames' upstream of the main transcript[393].

If compound heterozygous effects exist for UTRs variants, then it is logical to extend this concept to intergenic and intronic regions which constitute the majority of our genome[88]. Variants in intergenic regions on one haplotype that disrupt enhancer-promoter activity, combined with coding variants on the other haplotype, are almost certainly a naturally occurring phenomenon, which should warrant its investigation. Such an effort will not only require significant expansion of current WGS sequencing efforts to unlock the intergenic regions, but also functional studies aimed towards unraveling the underpinnings of transcriptional regulation. Genes are uniquely expressed and spatiotemporally controlled in specific cell types as a result of a complex interplay between Transcription Factor (TF) and *cis*-regulatory elements, such as enhancers[394,395]. Mapping these complex interactions in a tissue-specific manner will be crucial in uncovering how intergenic CH variation can contribute to the development of disease. Multiple efforts are being carried out *in vitro*: for example, recent modifications of the CRIPSR technology allowed assays that systemically test $\geq 3,500$ enhancer-gene connections across 30 genes[396]. Even more recently, the regulatory network of the Drosophila brain was investigated and dissected, revealing more than 95,000 regulatory regions across multiple cell types[397]. This is reminiscent of the days of linkage mapping, as discussed in chapter 1, in which the location of just a few disease genes were determined. The next frontier of genetic analysis will require mapping of these regulatory features of the genome across human tissues and cell types, which is a task that will pose significant challenges.

5.2.2 Nonadditive effects enable gene therapies and drug development

The identification of nonadditive genetic effects, particularly biallelic recessive effects of rare variants, has significant implications for drug development and gene therapies. This thesis provides a unified framework for detecting recessive effects and discovering novel Mendelian gene-trait associations, which are critical to therapeutic success. Drug targets with Mendelian inheritance evidence are 3.7 times more likely to gain U.S. Food and Drug Administration (FDA) approval compared to those without, outstripping the 2.6-fold increase from GWAS-implicated genes[78,398].

In my analysis, I observe that genes with biallelic pLoF variants are depleted among drug targets, likely because these variants predominantly occur in LoF-tolerant genes—where complete loss is observed in healthy individuals—suggesting reduced biological essentiality and relevance as targets[227]. However, this depletion may vary when stratifying by drug type: antagonists, which dominate approved drugs and mimic protective pLoF effects (PCSK9 inhibitors), may underrepresent these genes, while agonists, which restore function in deleterious pLoF contexts (CFTR modulators), could show less depletion[188]. Future work should stratify drug targets by LoF-tolerance status and mode of action (agonist versus antagonist) to clarify these trends.

The impact of Mendelian recessive inheritance is particularly evident in the field of gene therapy[399,400]. As of late 2024, the FDA has approved 38 cellular and gene therapies, of which 12 specifically target genetic disorders (Table D.1). Strikingly, all 12 of these treatments (100%) target disorders with recessive modes of inheritance. Examples include therapies for Spinal Muscular Atrophy[401], inherited retinal dystrophy[402], β-Thalassemia[403], and Haemophilia B[404]. The effectiveness of these therapies stems from their mechanism: replacing a dysfunctional or adding a functional copy of the gene of interest in a specific cellular environment. Due to the recessive nature of the targeted traits, expression of a single functional gene copy achieves a threshold effect, effectively curing the patient, or significantly reducing symptoms. This principle extends beyond gene therapies to some conventional

small-molecule drugs, such as Ivacaftor currently used in treating Cystic Fibrosis (CF), by enhancing the function of the CFTR protein[405,406]. Recessive effects, therefore, represent the 'low-hanging fruit' in drug development, particularly in gene therapy. The nonadditive framework developed in this thesis offers a systematic approach to identifying nonadditive effects attributed to rare deleterious variants, which is where the largest effects tend to exist. Gene-based studies of nonadditive effects could play a crucial role in designing targeted gene therapies. Such therapies will aim to elevate gene function above the threshold required for homeostasis, potentially leading to the amelioration of disease.

Recent advancements in gene therapy have shown promising results in treating recessive genetic conditions in model organisms, many of which align with the nonadditive effects identified in this thesis. For instance, studies have demonstrated that hearing loss associated with biallelic variants in *GJB2* can be rescued in mice through gene therapy[407]. Similarly, recent investigations have revealed that a subtype of CD can be potentially cured in mice by replacing *NOD2* through gene therapy[408]. Both of these gene-trait associations were implicated in my analysis, exhibiting significant nonadditive effects. In addition to these findings, I uncovered other nonadditive effects, such as those in COPD linked with rare variants in *ATP2C2*, *DNAI1*, and *ODAD1*. While no current gene therapy studies exist for these specific genes for the treatment of COPD, they represent excellent candidates for future research. These discoveries open new avenues for potential long-term treatments for COPD, particularly if biallelic variations are identified early in life. Early identification through sequencing could pave the way for timely interventions through gene therapy or small molecule treatments, once such therapies are developed, tested and approved.

The scope of gene therapy research and development is rapidly expanding to include many more recessive disorders, reflecting a broader shift towards precision medicine enabled by large-scale genetic studies[409–413]. Through systematic understanding of nonadditive genetic effects, as provided by the framework developed

in this thesis, we can potentially broaden treatment strategies for a wide range of genetic conditions that have historically been challenging to address.

Finally, a critical consideration when interpreting nonadditive effects is the potential influence of phenotype misclassification. In population-based studies, such as those in UKBB, phenotypes are often derived from EHRs or self-reported diagnoses, which may not distinguish between rare Mendelian subtypes, and common complex forms of the same disease. When a Mendelian subtype of a specific disease is present, with variants at appreciable frequencies, it may be possible to observe inflated contributions of nonadditivity. For instance, in conditions like CD or hearing loss, the strong nonadditive signals observed for genes like *NOD2* and *GJB2* might partially reflect the inclusion of Mendelian cases within broader phenotype definitions. This heterogeneity could mask the true genetic architecture. Therefore, future studies would benefit from deeper clinical characterization, perhaps through recall-by-genotype approaches, and analytical methods capable of distinguishing between rare variants causing Mendelian forms and common variants contributing to complex forms of the same disease category.

5.2.3 The case for humans as the new model organism

In recent years, advancements in biomedical research have led some to question the necessity of traditional model organisms. Nobel Laureate Sydney Brenner famously suggested that humans have become the new model organism, rendering other models obsolete[414]. This quote should be interpreted in the context of the rapid innovations in biobanking, sequencing, proteomics, metabolomics, and imaging, combined with extensive EHR linkage, which has allowed for deep phenotyping at an unprecedented scale. As mentioned in Chapter 2 of this thesis and in other papers[184], 'human knockouts' represent naturally occurring *in vivo* experiments that can potentially be used to assess the phenotypic consequences of pharmacological inhibition. While not exclusively a result of LoF variants, in this thesis I demonstrate many associations that are the result of putative biallelic inactivation, highlighting its prospect in identifying gene-trait relationship.

However, human knockouts may offer several additional advantages over traditional animal models, such as mice and primates[415]. While mice used in research are often inbred strains designed to ensure phenotypic homogeneity and manage external noise[416], human knockouts exist within the complex genetic background of diverse populations. This diversity can reveal unexpected phenotypes or gene-environment interactions that might not be apparent in homogeneous animal models[414]. These observations can provide valuable insights into the full spectrum of effects resulting from gene inactivation in humans, potentially informing drug development and predicting possible outcomes of pharmacological inhibition. Secondly, human knockouts provide insights directly applicable to human physiology and pathology, avoiding the need to translate and extrapolate from animal models which may be prone to errors[417]. Such 'errors in translation' could partially explain the higher attrition rates when transitioning from pre-clinical to clinical trials[418]. Thirdly, studying naturally occurring human knockouts circumvents the ethical concerns associated with creating transgenic animals or exposing them to potentially dangerous molecules[419].

Despite such a potential of human knockouts, there are several limitations to its direct application for drug development, and more specifically for validating safety profiles. Most LoF variants are extremely rare, making it unlikely to observe them in the biallelic state, even across billions of individuals in outbred populations[188]. Some variants are embryonically lethal in the biallelic state, precluding their observation in living humans due to recessive selection[309]. As demonstrated in this work and by others[416], enrichment of variants tends to occur in genes tolerant to homozygous LoF, which may not be of medical interest[227,420]. Furthermore, suspected variants leading to putative knockouts would need to be rigorously validated *in vitro*, to ensure that they are *bona fide* LoF variants. Comprehensive phenotyping of affected individuals is necessary to identify unexpected effects, and recall-by-genotype systems will need to be put into place to ensure that the individual does not exhibit 'unexpected' traits, validating that they are a healthy knockouts. At the time of writing this thesis, there is only one example of a human biallelic carrier being recalled and

used as evidence for safety in the context of drug development[337]. However, this is in contrast to numerous papers from the past years that suggested the potential of utilising human knockouts in drug development[145,146,421–424].

One reason for this discrepancy could be that pharmacological inhibition differs from complete loss of gene expression, and human knockouts may not fully replicate drug-induced effects. Animal models will likely remain important for assessing gradual inhibition, identifying severe off-target effects, and informing first human dose through pharmacometric modeling, balancing efficacy and potential side effects[425]. While the study of leveraging naturally occurring human knockouts is still in its infancy, it shows promise for validating therapeutic hypotheses through association analysis, and we have some evidence that it might assist in instilling confidence in safety profiles[337]. This approach could be particularly valuable given the current challenges in the pharmaceutical industry, exemplified by Eroom's Law[426], which observes that the number of new drugs approved per billion US$ spent has halved approximately every nine years. By providing *in vivo* evidence of safety in humans, knockout studies could potentially reduce these costs.

5.2.4 Final remarks

In the two decades since the first complete human genome was sequenced[39], biomedical research has advanced remarkably. These developments have enabled us not only to decode but also to manipulate the medium of inheritance. However, the rapid pace of technological progress raises critical questions about our capacity to address the associated medical, ethical, and societal implications that comes with implementation in a health care setting.

Recent initiatives, such as the UK's plan to screen 100,000 newborns for 200 treatable monogenic disorders using WGS, exemplify this trend[427]. While promising for early intervention, concerns have been raised about high false-positive rates, uncertain predictive value of genetic variants, potential for unnecessary long-term monitoring, and whether it truly improves outcomes compared to symptomatic diagnosis[428]. As genetic screening becomes more accessible, its clinical application

for common complex diseases is imminent, likely through PRS. This technology has the potential to slow disease progression by improving population screening and fostering risk-reducing behaviors[429]. However, disparities in predictive accuracy between European and non-European populations raise important concerns about exacerbating health inequalities[341]. Concurrently, advances in AI promise to revolutionize diagnosis[430], for example with Google's `Med-PaLM`, a large language model that can answer complex medical questions[431]. Yet, such models face challenges in promoting fair and equitable medical advice that do not perpetuate existing biases[432,433]. Implementing these technologies equitably in healthcare remains a critical challenge, demanding thoughtful consideration.

Lastly, our ability to manipulate the genetic code has advanced remarkably, moving beyond merely reading DNA to editing it with unprecedented precision. The CRISPR-Cas system, introduced in 2012[434], has undergone significant evolution, culminating in techniques like CRISPR 'prime editing'[435]. This breakthrough enables high fidelity genome editing, potentially treating conditions once deemed incurable, from Duchenne Muscular Dystrophy to coronary artery disease[436]. The impact of gene editing extends far beyond medicine. In agriculture, it promises to revolutionize crop development, increasing resilience to harsh conditions and improving nutritional value, which will be crucial for securing global food supply in the face of climate change[437–439]. However, this technology is not without risks and ethical concerns. The ability to fundamentally alter genomes in plants, insects, or animals raises the prospect of unintended consequences in both human applications and environmental interventions[440]. Moreover, recent developments in insect-based delivery of genetically modified viruses for crop editing have sparked debates about biosecurity and potential weaponization[441]. While gene editing holds immense promise for treating genetic disorders, its ease of use presents significant regulatory challenges and biosecurity risks.

Despite the scientific and ethical dilemmas, there is reason for cautious optimism. Collaborative efforts between researchers, ethicists, and policymakers are shaping

frameworks for responsible innovation in genomics, balancing healthcare implementation with societal values. As we navigate the complexities of genetic research, our efforts today lay the groundwork for a future where precision medicine and sustainable agriculture are realities, securing a healthier and more prosperous world for all.

Appendices

A

Supplement to chapter 2

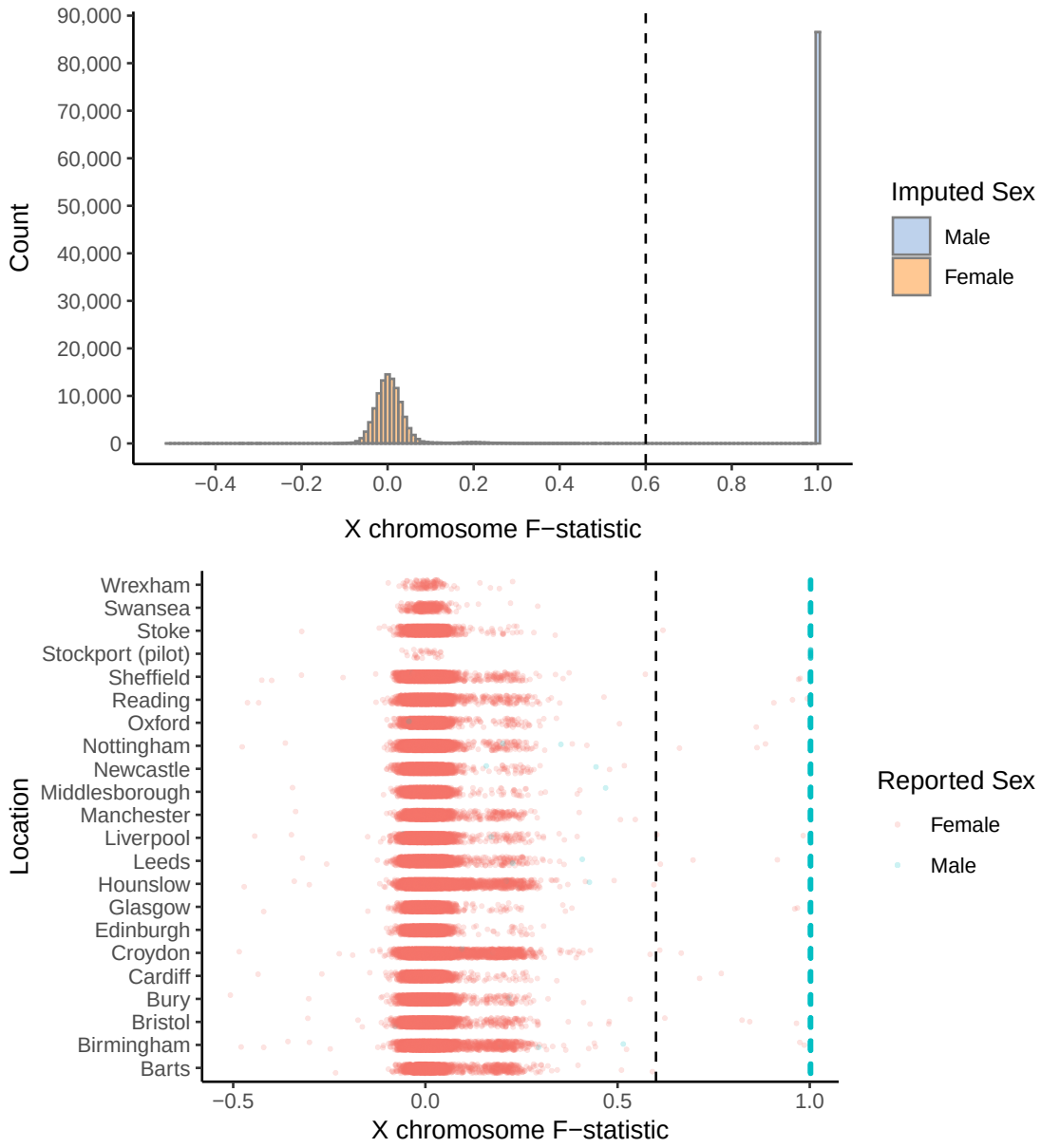


Figure A.1: Histogram and scatter-plots of X chromosome F-statistic by collection. Samples lying to the left and right of the dashed line were called as female and male respectively, according to the imputed sex colorings in the upper histogram. Reported sex, split by UKBB recruitment center are shown in the lower jittered scatter-plots: red if the sample is reported as female, and blue if the sample is reported as male.

Filter	**Samples**	**Batch 1**	**Batch 2**	**%**
Initial samples in raw UKBB VCF	200,643	NA	NA	100.0
Initial samples in filtered VCF	199,795	49,759	150,036	99.6
Sample call rate <0.95	7,400	4,780	2,620	3.7
Mean DP <19.5	3,253	511	2,742	1.6
Mean GQ <47.8	2,123	834	1,289	1.1
Samples with sex swap	85	24	61	0.0
Samples with excess URVs	76	6	70	0.0
PCA based filters	13,537	3,390	10,147	6.7
Within batch Ti/Tv ratio outside 4 SDs	13	3	10	0.0
Within batch Het/HomVar ratio outside 4 SDs	251	46	205	0.1
Within batch Insertion/Deletion ratio outside 4 SDs	9	4	5	0.0
n singletons >175	19	2	17	0.0
Samples after all sample filters	**176,935**	**41,371**	**135,564**	**88.2**

Table A.1: Summary of sample attrition following filtering for 200k UKBB ES release. Moving down through the rows of the table, we move through QC filtering steps.

Filter	**Variants**	**%**
Initial variants in raw UKBB VCF	15,922,704	100.0
Variants removed in initial filters	2,883,660	18.1
Invariant sites after sample filters	2,744,044	17.2
Overall variant call rate <0.97	1,122,987	7.1
Variants failing HWE filter	5,237	0.0
Variants remaining after all filters	**9,169,408**	**57.6**

Table A.2: Summary of variant attrition after imposing filters in 200k UKBB ES release. Moving down through the rows of the table, we move through QC filtering steps

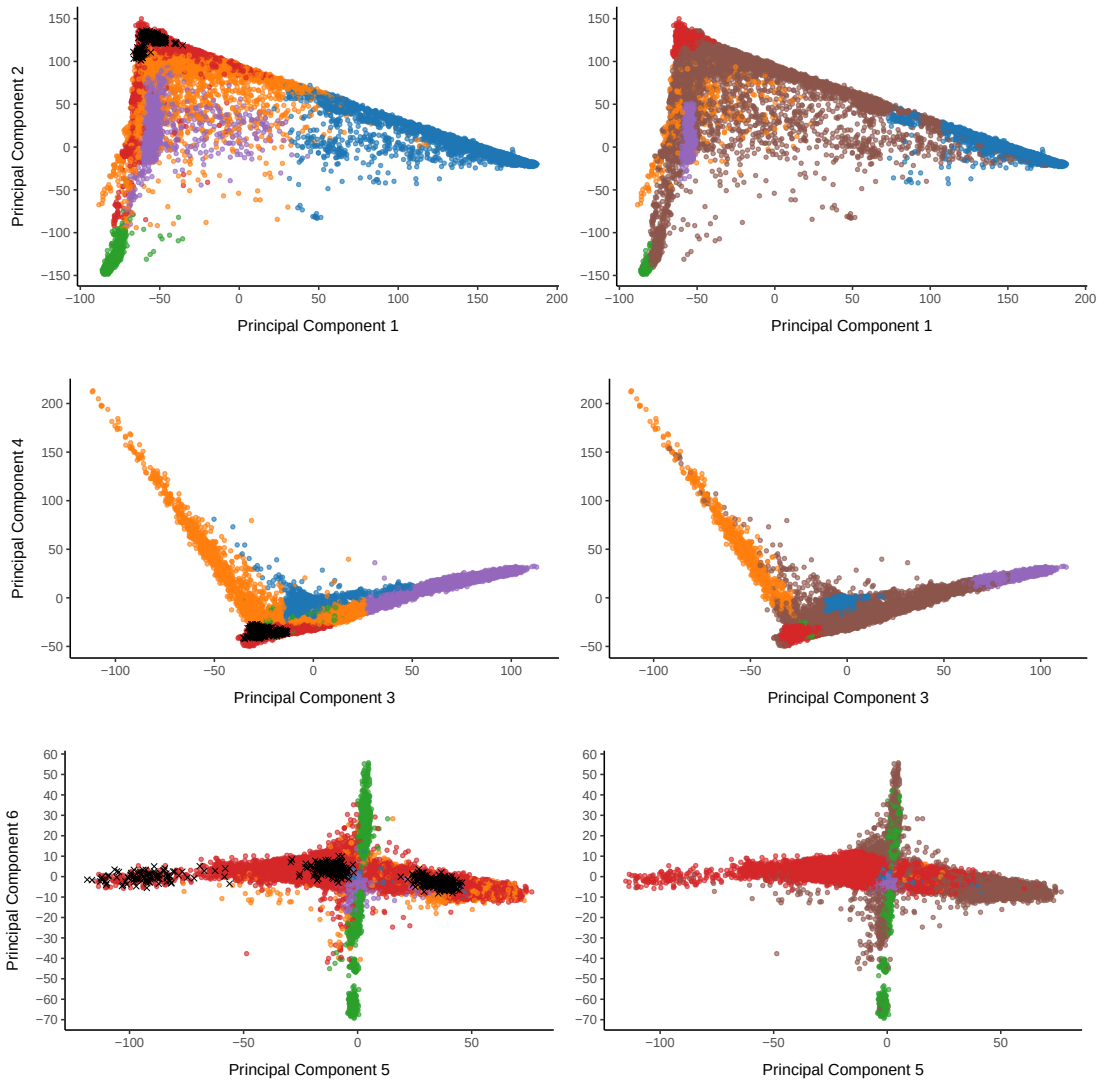


Figure A.2: Scatter-plots of PCs of UKBB genotype data projected into the PC space defined by 1KGP samples. Points are colored according to sample collection, with 1KGP samples colored in blue. 1KGP super-populations labels were used to train a random forest classifier.

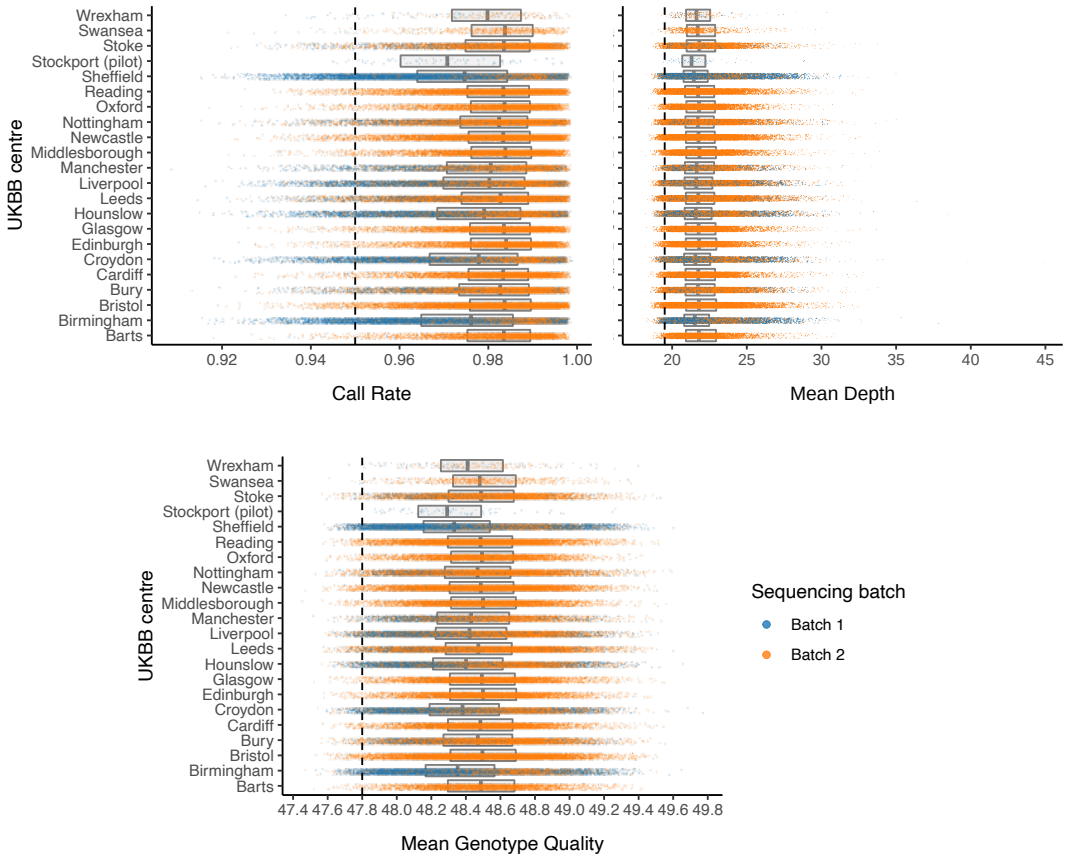


Figure A.3: Distributions of sample metrics following initial variant filtering for 200k ES. Distributions of sample metrics following initial restriction to variants, lying outside LCRs and inside the padded (50 bp) target intervals, and prior to the initial hard sample filters (call rate > 0.95, mean depth > 19.5, mean GQ > 47.8). In each plot, jittered scatters display the distribution for each UKBB recruitment center, colored according to sequencing batch. Box-plots behind the scatter display the median and interquartile range for each sequencing batch. Hard-filtering thresholds are denoted by the dashed vertical line.

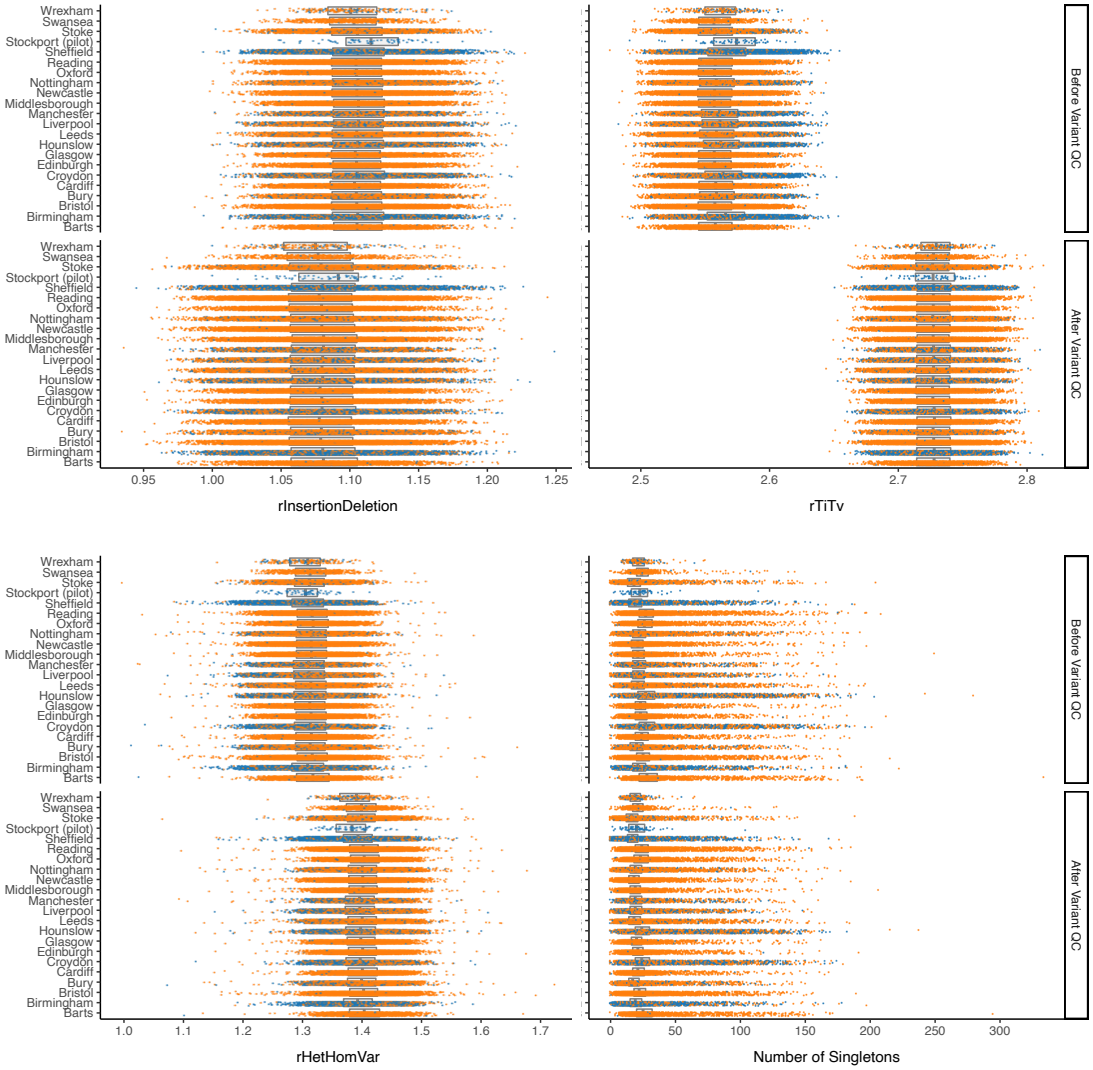


Figure A.4: Distributions of variant metrics before and after QC for 200k ES. Distributions of variant metrics before and after the removal of invariant sites, variants with call rate < 0.97, and variants out of HWE ($P < 1 \times 10^{-6}$). In each plot, jittered scatters display the distribution for each sequencing batch colored by sequencing batch. Box-plots behind the scatter display the median and interquartile range for each sequencing batch. Points shown are following variants hard-filters and prior to removal of variants with metrics outside four standard deviations of the sequencing batch mean.

Table A.3: Comparative SER analysis for various phasing methods. Assessment of SER point estimates across `SHAPEIT4`, `SHAPEIT5` (pre- and post-PP filtering), and `Eagle2` phasing algorithms. `SHAPEIT5` analysis encompassed all autosomes, while `SHAPEIT4` and `Eagle2` evaluations were restricted to chromosomes 20-22. The data presents 95% binomial confidence intervals for each method's SER estimates.

		SHAPEIT5 (Using chromosomes 1-22)										SHAPEIT4 (Using chromosomes 20-22)					Eagle2 (Using chromosomes 20-22)				
		No phasing confidence cutoff					Phasing confidence (PP) >= 0.90														
MAC Bin	Type	Errors	Tested	Estimate	95%LowerCI	95%UpperCI	Errors	Tested	Estimate	95%LowerCI	95%UpperCI	Errors	Tested	Estimate	95%LowerCI	95%UpperCI	Errors	Tested	Estimate	95%LowerCI	95%UpperCI
singleton	wes	569	1754	0.324401368	0.3028968753	0.346673342	26	214	0.1214953271	0.0842755044	0.1720644005	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
2-5	wes	248	3880	0.063917526	0.0566439682	0.072053731	8	2949	0.0027127840	0.0013752544	0.0053441916	24	262	0.091603053	6.232983e-02	0.132679097	37	261	0.141762452	0.104629735	0.189287461
6-10	wes	106	3334	0.031793641	0.0263560616	0.038308921	11	2995	0.0036727880	0.0020520907	0.0065650566	9	195	0.046153846	2.446889e-02	0.085374693	16	193	0.082901554	0.051671409	0.130411467
11-20	wes	85	4209	0.020194821	0.0163621901	0.024902467	16	3937	0.0040640081	0.0025031468	0.0065917278	8	266	0.030075188	1.531679e-02	0.058213275	12	266	0.045112782	0.025991911	0.077185187
21-50	wes	108	6768	0.015957447	0.0132348443	0.019229214	22	6495	0.0033872209	0.0022380032	0.0051235334	11	425	0.025882353	1.451259e-02	0.045746184	13	424	0.030660377	0.018004106	0.051744765
51-100	wes	84	5890	0.014261460	0.0115345461	0.017621559	28	5704	0.0049088359	0.0033985308	0.0070855479	0	435	0.000000000	0.000000e+00	0.008753637	6	434	0.013824885	0.006351061	0.029829749
101-200	wes	102	7106	0.014354067	0.0118392238	0.017393701	51	6949	0.0073391855	0.0055867222	0.0096360410	3	426	0.007042254	2.397833e-03	0.020497722	6	424	0.014150943	0.006501214	0.030525256
201-500	wes	117	13123	0.008915644	0.0074448966	0.010673815	59	12934	0.0045616205	0.0035383049	0.0058791438	8	814	0.009828010	4.988248e-03	0.019272516	8	814	0.009828010	0.004988248	0.019272516
501-1000	wes	71	13506	0.005256923	0.0041701182	0.006625083	50	13499	0.0037039781	0.0028108657	0.0048794756	1	904	0.001106195	5.674037e-05	0.006239151	3	907	0.003307607	0.001125509	0.009679290
1001-2000	wes	65	18857	0.003446996	0.0027055923	0.004390669	54	18854	0.0028641137	0.0021959536	0.0037348131	8	1155	0.006926407	3.513827e-03	0.013607980	8	1154	0.006932409	0.003516875	0.013619719
2001-5000	wes	62	19095	0.003246923	0.0025338592	0.004159817	49	19095	0.0025661168	0.0019417214	0.0033906156	4	1161	0.003445306	1.340603e-03	0.008825123	4	1157	0.003457217	0.001345241	0.008855514
5001-10000	wes	14	8100	0.001728395	0.0010298808	0.002899300	10	8100	0.0012345679	0.0006707488	0.0022712460	0	500	0.000000000	0.000000e+00	0.007624340	0	500	0.000000000	0.000000000	0.007624340
10001-Inf	wes	1475	1443202	0.001022033	0.0009712134	0.001075509	1141	1443202	0.0007906031	0.0007460571	0.0008378066	119	100982	0.001178428	9.849549e-04	0.001409851	132	100986	0.001307112	0.001102441	0.001549721
singleton	array	155	423	0.366430260	0.321906922	0.413357782	12	56	0.214285714	0.127059042	0.338194636	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
2-5	array	36	428	0.084112150	0.061373572	0.114249813	7	309	0.022653722	0.011015868	0.046014485	3	26	0.115384615	0.0400324462	0.289759032	5	26	0.192307692	0.085070611	0.378762571
6-10	array	11	252	0.043650794	0.024546025	0.076459727	10	213	0.046948357	0.025698400	0.084250402	0	13	0.000000000	0.0000000000	0.228095372	0	13	0.000000000	0.000000000	0.228095372
11-20	array	17	283	0.060070671	0.037839940	0.094084709	9	253	0.035573123	0.018826009	0.066212672	1	13	0.076923077	0.0039456380	0.333139509	1	13	0.076923077	0.003945638	0.333139509
21-50	array	21	500	0.042000000	0.027632310	0.063351586	9	470	0.019148936	0.010106380	0.035988065	2	30	0.066666667	0.0184770238	0.213234584	1	30	0.033333333	0.001709776	0.166703910
51-100	array	9	349	0.025787966	0.013625288	0.048276334	7	333	0.021021021	0.010219086	0.042747846	0	11	0.000000000	0.0000000000	0.258832967	0	11	0.000000000	0.000000000	0.258832967
101-200	array	8	287	0.027874564	0.014190770	0.054030105	6	281	0.021352313	0.009821825	0.045793179	0	12	0.000000000	0.0000000000	0.242494007	0	12	0.000000000	0.000000000	0.242494007
201-500	array	13	486	0.026748971	0.015697700	0.045222947	11	474	0.023206751	0.013006747	0.041072819	1	28	0.035714286	0.0018319034	0.177121977	2	28	0.071428571	0.019812064	0.226453626
501-1000	array	12	835	0.014371257	0.008239771	0.024950599	12	835	0.014371257	0.008239771	0.024950599	1	44	0.022727273	0.0011657567	0.118077097	1	44	0.022727273	0.001165757	0.118077097
1001-2000	array	22	2438	0.009023790	0.005966753	0.013625616	22	2438	0.009023790	0.005966753	0.013625616	1	127	0.007874016	0.0004038842	0.043254008	2	126	0.015873016	0.004363802	0.056028764
2001-5000	array	208	69778	0.002980882	0.002602835	0.003413651	197	69778	0.002823239	0.002455984	0.003245233	9	3694	0.002436383	0.0012823391	0.004624204	18	3696	0.004870130	0.003082865	0.007685558
5001-10000	array	934	371929	0.002511232	0.002355441	0.002677299	889	371929	0.002390241	0.002238363	0.002552399	55	20025	0.002746567	0.0021108907	0.003572986	78	20028	0.003894548	0.003121934	0.004857435
10001-Inf	array	21822	8907371	0.002449881	0.002417630	0.002482561	21159	8907371	0.002375448	0.002343693	0.002407633	1413	524663	0.002693157	0.0025565178	0.002837079	2189	524661	0.004172218	0.004001396	0.004350301

Table A.4: Trio switch errors by MAC bin for 96 parent-offspring trio relationships before and after filtering by PP ≥ 0.90. Overview of the trio-SER stratified by MAF-bin for ES data. 95% binomial confidence intervals are shown.

		All variants				
Dataset	MAC	Switch errors	Evaluated sites	Switch Error Rate	Switch Error Rate (95% lower CI)	Switch Error Rate (95% upper CI)
wes	singleton	569	1754	0.324401368	0.3028968753	0.346673342
wes	2-5	248	3880	0.063917526	0.0566439682	0.072053731
wes	6-10	106	3334	0.031793641	0.0263560616	0.038308921
wes	11-20	85	4209	0.020194821	0.0163621901	0.024902467
wes	21-50	108	6768	0.015957447	0.0132348443	0.019229214
wes	51-100	84	5890	0.014261460	0.0115345461	0.017621559
wes	101-200	102	7106	0.014354067	0.0118392238	0.017393701
wes	201-500	117	13123	0.008915644	0.0074448966	0.010673815
wes	501-1000	71	13506	0.005256923	0.0041701182	0.006625083
wes	1k-2k	65	18857	0.003446996	0.0027055923	0.004390669
wes	2k-5k	62	19095	0.003246923	0.0025338592	0.004159817
wes	5k-10k	14	8100	0.001728395	0.0010298808	0.002899300
wes	10k+	1475	1443202	0.001022033	0.0009712134	0.001075509
		Restricted to varaints with Phasing Confidence (PP) >= 0.90				
Dataset	MAC	Switch errors	Evaluated sites	Switch Error Rate	Switch Error Rate (95% lower CI)	Switch Error Rate (95% upper CI)
wes	singleton	26	214	0.1214953271	0.0842755044	0.1720644005
wes	2-5	8	2949	0.0027127840	0.0013752544	0.0053441916
wes	6-10	11	2995	0.0036727880	0.0020520907	0.0065650566
wes	11-20	16	3937	0.0040640081	0.0025031468	0.0065917278
wes	21-50	22	6495	0.0033872209	0.0022380032	0.0051235334
wes	51-100	28	5704	0.0049088359	0.0033985308	0.0070855479
wes	101-200	51	6949	0.0073391855	0.0055867222	0.0096360410
wes	201-500	59	12934	0.0045616205	0.0035383049	0.0058791438
wes	501-1000	50	13499	0.0037039781	0.0028108657	0.0048794756
wes	1k-2k	54	18854	0.0028641137	0.0021959536	0.0037348131
wes	2k-5k	49	19095	0.0025661168	0.0019417214	0.0033906156
wes	5k-10k	10	8100	0.0012345679	0.0006707488	0.0022712460
wes	10k+	1141	1443202	0.0007906031	0.0007460571	0.000837807

Table A.5: Trio-SER across genes before and after filtering by PP≥ 0.90 using 96 trio-offspring relationships. Overview of switch errors per protein coding gene stratified by PP.

		All variants					Restricted to varaints with Phasing Confidence (PP) >= 0.90					
chromosome	Total genes (n)	Genes with zero errors (n)	Genes with ≥1 switch errors (n)	Genes with zero errors (%)	mean sites per gene (n)	mean switch errors per gene (n)	Total genes (n)	Genes with zero errors (n)	Genes with ≥1 switch errors (n)	Genes with zero errors (%)	mean sites per gene (n)	mean switch errors per gene (n)
chr1	2150	1880	270	0.8744186	81.19023	0.1632558	2142	2012	130	0.9393091	81.26937	0.07329599
chr2	1367	1199	168	0.8771031	89.44770	0.1726408	1362	1291	71	0.9478708	89.52203	0.06240822
chr3	1219	1081	138	0.8867925	82.28056	0.1460213	1217	1146	71	0.9416598	82.21775	0.06573541
chr4	782	690	92	0.8823529	90.60614	0.1508951	779	724	55	0.9293967	90.74069	0.08600770
chr5	965	844	121	0.8746114	106.52021	0.1875648	962	904	58	0.9397089	106.61019	0.06964657
chr6	1123	1002	121	0.8922529	92.32769	0.1308994	1116	1050	66	0.9408602	92.74014	0.06810036
chr7	981	852	129	0.8685015	87.32212	0.1875637	980	909	71	0.9275510	87.19694	0.09081633
chr8	727	639	88	0.8789546	88.95461	0.1636864	721	673	48	0.9334258	89.48405	0.07905687
chr9	825	712	113	0.8630303	93.09091	0.2012121	824	764	60	0.9271845	92.92597	0.09344660
chr10	829	715	114	0.8624849	91.44270	0.1917973	828	760	68	0.9178744	91.32488	0.10386473
chr11	1422	1249	173	0.8783404	89.45640	0.1687764	1416	1320	96	0.9322034	89.64689	0.09322034
chr12	1178	1018	160	0.8641766	78.74024	0.1833616	1168	1097	71	0.9392123	79.20377	0.08047945
chr13	342	305	37	0.8918129	88.33626	0.1286550	342	322	20	0.9415205	88.14035	0.05847953
chr14	680	587	93	0.8632353	87.12059	0.1808824	679	627	52	0.9234168	87.00884	0.08689249
chr15	679	587	92	0.8645066	87.28277	0.2179676	677	629	48	0.9290990	87.29690	0.10044313
chr16	972	826	146	0.8497942	95.82922	0.2047325	968	899	69	0.9287190	95.95764	0.10020661
chr17	1339	1150	189	0.8588499	88.27483	0.1904406	1336	1234	102	0.9236527	88.24027	0.09131737
chr18	315	263	52	0.8349206	93.92063	0.2285714	315	282	33	0.8952381	93.72063	0.13650794
chr19	1627	1394	233	0.8567916	100.34173	0.1948371	1626	1489	137	0.9157442	100.17651	0.10516605
chr20	594	520	74	0.8754209	87.32997	0.1666667	591	543	48	0.9187817	87.57530	0.09983080
chr21	231	186	45	0.8051948	115.51082	0.2900433	229	205	24	0.8951965	116.21834	0.13100437
chr22	493	420	73	0.8519270	97.88844	0.1845842	491	447	44	0.9103870	98.02444	0.09979633
Total	**20840**	**18119**	**2721**	**0.8694338**	**90.07135**	**0.1780230**	**20769**	**19327**	**1442**	**0.9305696**	**90.15499**	**0.08594540**

Table A.6: Comparison of statistical and physical phasing. Analysis of haplotype predictions from `SHAPEIT5` phasing versus read-backed phasing from `WhatsHap`. The comparison is constrained to variant pairs within individual reads, encompassing 49,756 individuals across all autosomes. 'Agreement pairs' denotes the count of concordant phase assignments between `SHAPEIT5` statistically phased variants and `WhatsHap` read-backed phased variants. 95% binomial CIs are shown.

MAC group	Phasing confidence (PP)	Agreement pairs	Total pairs	Agreement (%)	Agreement (lower CI95%)	Agreement (upper CI95%)
Singleton	0.5	55294	80978	68.28274346	67.96135324	68.60239916
2-5	0.5	179110	190877	93.83529708	93.7265151	93.94231469
6-10	0.5	147028	151237	97.21695088	97.13284473	97.29865844
11-20	0.5	188443	191852	98.22310948	98.16302103	98.28126682
21-50	0.5	287101	290091	98.96928895	98.93188143	99.00539955
51-100	0.5	236817	238391	99.33974017	99.30642528	99.37146496
101-200	0.5	253158	254380	99.51961632	99.4919891	99.54574795
201-500	0.5	433555	434964	99.67606515	99.65873407	99.69251879
501-1000	0.5	523025	523659	99.87892884	99.86913739	99.8879885
1k-2k	0.5	819559	820360	99.90235994	99.89536378	99.90888875
2k-5k	0.5	928519	929042	99.94370545	99.93867127	99.94832662
5k-10k	0.5	337111	337224	99.96649112	99.95971854	99.97212533
10k+	0.5	86033459	86049711	99.98111324	99.98082066	99.98140136
Singleton	0.6	44754	62119	72.04558992	71.69132435	72.39712904
2-5	0.6	164461	172057	95.58518398	95.48709679	95.68123568
6-10	0.6	141759	144304	98.23636212	98.16715434	98.30300182
11-20	0.6	184124	186009	98.98660817	98.94007039	99.03112266
21-50	0.6	282693	284263	99.44769456	99.41977409	99.47427861
51-100	0.6	234271	235030	99.6770625	99.65329904	99.6992021
101-200	0.6	251102	251714	99.75686692	99.73685351	99.77536165
201-500	0.6	431439	432291	99.80291054	99.78923967	99.8156963
501-1000	0.6	522970	523588	99.88196827	99.87229485	99.89090974
1k-2k	0.6	819535	820332	99.90284421	99.89586469	99.90935637
2k-5k	0.6	928518	929041	99.94370539	99.9386712	99.94832656
5k-10k	0.6	337111	337224	99.96649112	99.95971854	99.97212533
10k+	0.6	86033453	86049705	99.98111324	99.98082066	99.98140136
Singleton	0.7	33374	43939	75.95530167	75.55345635	76.35260899
2-5	0.7	158320	162977	97.14254159	97.06053659	97.22232428
6-10	0.7	139499	141080	98.87935923	98.823084	98.93297267
11-20	0.7	182258	183445	99.35293957	99.31520108	99.38861114
21-50	0.7	281043	282046	99.64438425	99.62172912	99.66568709
51-100	0.7	233315	233836	99.77719427	99.75724875	99.79550434
101-200	0.7	250325	250774	99.82095433	99.80362744	99.83675487
201-500	0.7	430700	431384	99.84144057	99.82911531	99.85287817
501-1000	0.7	522946	523560	99.88272595	99.8730821	99.89163781
1k-2k	0.7	819528	820323	99.90308695	99.8961158	99.90959073
2k-5k	0.7	928518	929040	99.94381297	99.93878338	99.94842954
5k-10k	0.7	337111	337224	99.96649112	99.95971854	99.97212533
10k+	0.7	86033449	86049700	99.9811144	99.98082183	99.98140251
Singleton	0.8	21455	26686	80.39796148	79.91730169	80.86987093
2-5	0.8	152813	155235	98.43978484	98.37692561	98.50024675
6-10	0.8	137445	138304	99.37890444	99.33610542	99.4189605
11-20	0.8	180688	181304	99.66023916	99.63238158	99.68599238
21-50	0.8	279590	280194	99.78443507	99.76656637	99.8009387
51-100	0.8	232605	232970	99.84332747	99.82642441	99.85858681
101-200	0.8	249753	250105	99.85925911	99.84378114	99.8732055
201-500	0.8	430155	430731	99.86627385	99.85490665	99.87675159
501-1000	0.8	522923	523535	99.88310237	99.87347314	99.89199957
1k-2k	0.8	819522	820316	99.90320803	99.89624105	99.90970764
2k-5k	0.8	928517	929039	99.94381291	99.93878331	99.94842948
5k-10k	0.8	337111	337224	99.96649112	99.95971854	99.97212533
10k+	0.8	86033448	86049699	99.9811144	99.98082183	99.98140251
Singleton	0.9	9176	10627	86.34609956	85.68013318	86.98579858
2-5	0.9	147445	148368	99.3778982	99.33659124	99.41664829
6-10	0.9	135450	135797	99.74447153	99.71617651	99.76995226
11-20	0.9	179165	179469	99.83061141	99.81049005	99.84859962
21-50	0.9	278208	278558	99.87435292	99.86049218	99.88683809
51-100	0.9	231913	232174	99.88758431	99.87310359	99.90041422
101-200	0.9	249183	249453	99.89176318	99.87806867	99.90392109
201-500	0.9	429661	430154	99.88538988	99.87482352	99.89506526
501-1000	0.9	522908	523515	99.88405299	99.8744612	99.8929127
1k-2k	0.9	819517	820311	99.90320744	99.89624041	99.90970709
2k-5k	0.9	928517	929039	99.94381291	99.93878331	99.94842948
5k-10k	0.9	337111	337224	99.96649112	99.95971854	99.97212533
10k+	0.9	86033446	86049693	99.98111905	99.98082651	99.98140712

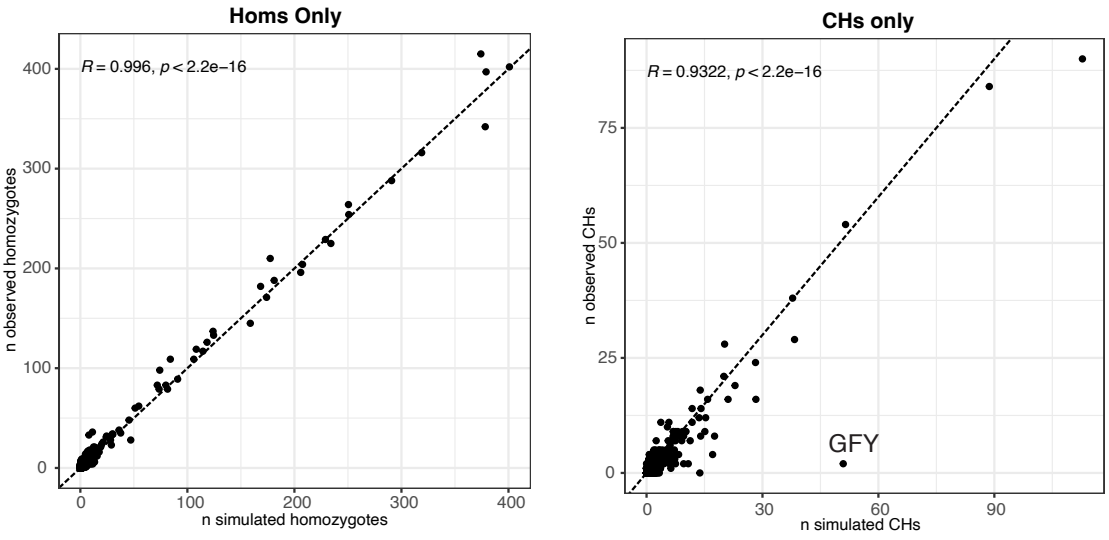


Figure A.5: Simulation of CH and homozygous events in an outbred population. I generated genotypes for 1174 genes using allele frequencies derived from observed pLoF variants. For each gene, I simulated genotypes for 176,935 individuals. I averaged the number of biallelic variants across 10 simulations. This served as an estimate for the expected count (x-axis) of biallelic variants, against which I compared the empirically observations (y-axis). The first panel (left) is the comparison between observed and simulated homozygous variants. The second panel (right) is the comparison between observed and simulated CH variants.

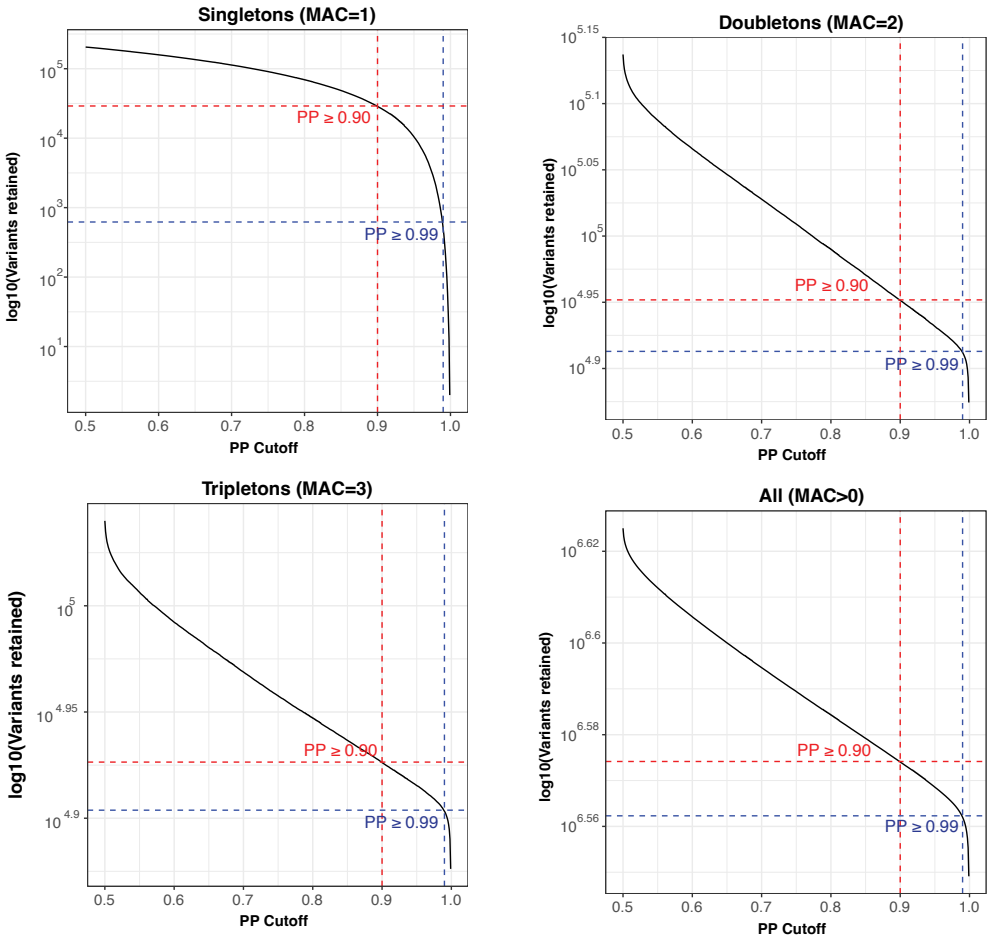


Figure A.6: Variants retained as a function of increasing phasing confidence score. Each subplot displays the number of variants retained on the $\log_{10}$ scale as the PP is increased, split by singletons (MAC= 1), doubletons (MAC= 2), tripletons (MAC= 3) and all (MAC> 0) variants. Dotted red and blue lines highlight the number of variants retained after imposing PP cutoffs of 0.9 and 0.99, respectively.

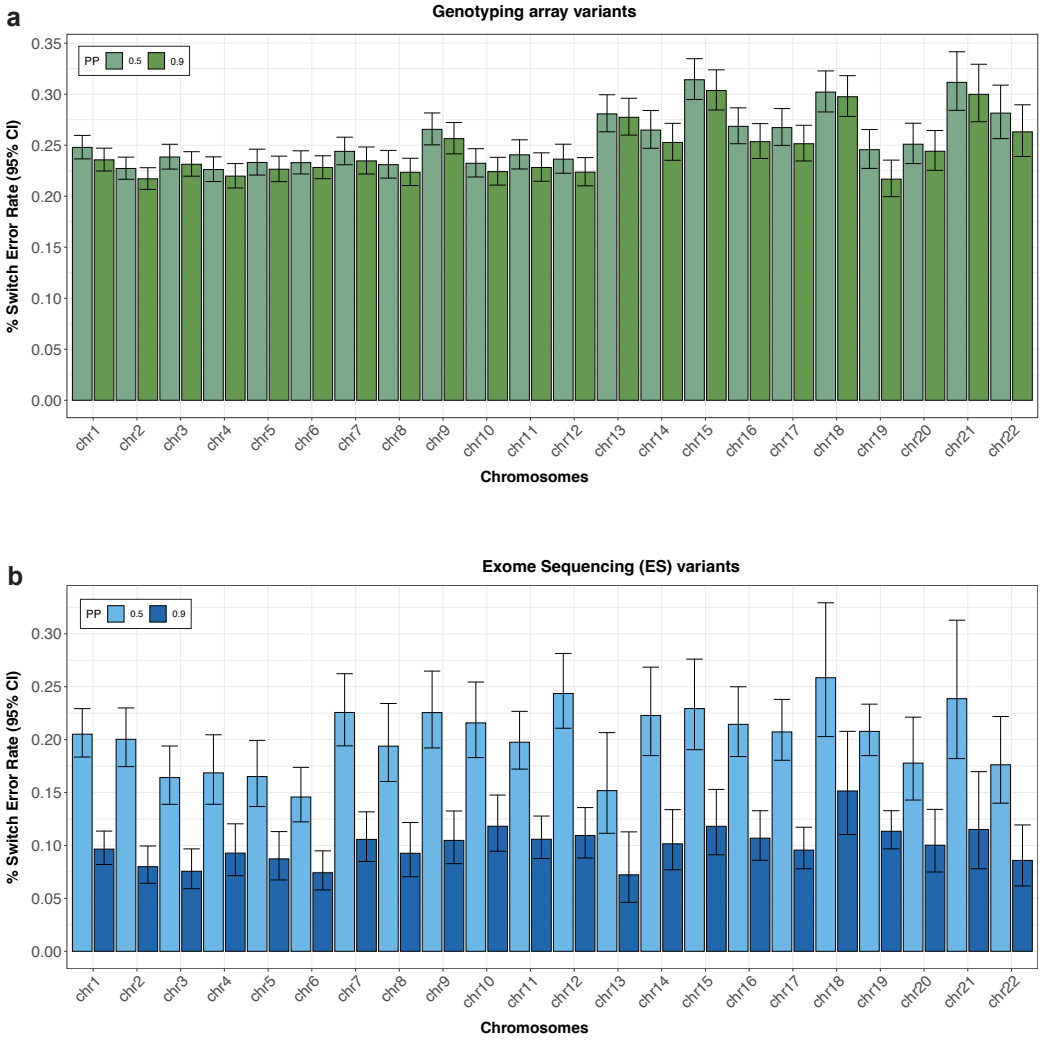


Figure A.7: Trio switch error rates by chromosome. Parent-offspring trios are used to determine switch error rates for variants that originate from the genotyping array (a) and exome sequencing data (b). I stratify by phasing confidence (PP) according to the color legends. Mean switch error rates are plotted, with whiskers enclosing the 95% binomial CI.

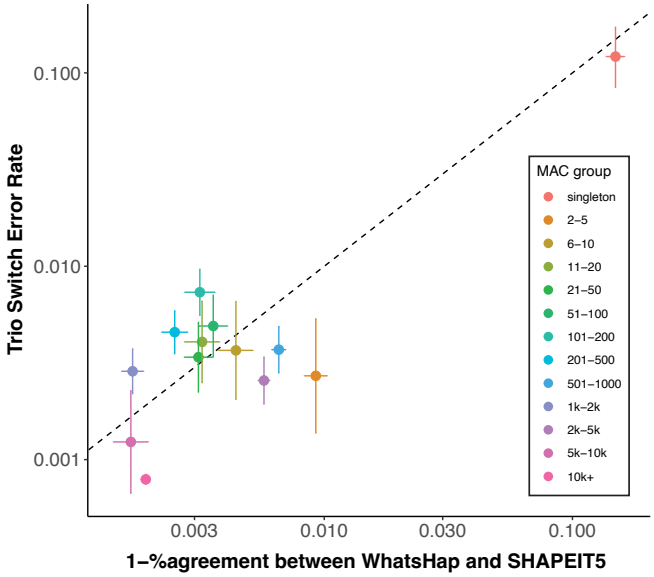


Figure A.8: Agreement between read-backed phasing and statistical phasing in UKBB. Illustrates the disagreement between `WhatsHap` (read-backed phasing) and `SHAPEIT5` (statistical phasing) on the x-axis against SER phase estimates implied by trio-based phasing on the y-axis. Variants are binned and colored based on their MAC. Horizontal and vertical lines enclose 95% binomial CIs around mean estimates. The dotted indicates $y = x$.

Table A.7: Systematic search and conditional analysis of common variants around significant gene-trait associations. This table presents the results of our systematic search for common variants (MAF $> 1\%$) located within 1 MB upstream and downstream of any significant gene-trait associations ($P < 5.25 \times 10^{-5}$). Upon identifying a significant common variant ($P < 5 \times 10^{-6}$), a conditional analysis was performed using that variant. If other significant variants remained, they were included in subsequent iterations of the analysis until either the signal was exhausted, or 25 iterations were completed. The table enumerates all the resulting gene-traits and variants upon which common variant conditional analysis took place.

Trait	Iteration	Conditional *P*-value	Marker	List of markers
NAFLD_combined	1	2.17E-06	chr1:16201723:G:T	chr1:16201723:G:T
NAFLD_combined	2	2.49E-09	chr1:16006647:GAATA:C	chr1:16006647:GAATA:G,chr1:16201723:G:T
NAFLD_combined	3	2.22E-07	chr1:16053036:C:CT	chr1:16006647:GAATA:G,chr1:16053036:C:CT,chr1:16201723:G:T
NAFLD_combined	4	2.74E-55	chr1:15498700:T:C	chr1:15498700:T:C,chr1:16006647:GAATA:G,chr1:16053036:C:CT,chr1:16201723:G:T
NAFLD_combined	5	3.62E-06	chr1:16185896:C:T	chr1:15498700:T:C,chr1:16006647:GAATA:G,chr1:16053036:C:CT,chr1:16185896:C:T,chr1:16201723:G:T
NAFLD_combined	6	3.86E-06	chr1:16187251:T:C	chr1:15498700:T:C,chr1:16006647:GAATA:G,chr1:16053036:C:CT,chr1:16185896:C:T,chr1:16187251:T:C,chr1:16201723:G:T
PSOR_combined	1	6.54E-200	chr6:31351254:G:T	chr6:31351254:G:T
PSOR_combined	2	2.95E-21	chr6:31249331:CA:C	chr6:31249331:CA:C,chr6:31351254:G:T
PSOR_combined	3	7.07E-25	chr6:31271630:G:T	chr6:31249331:CA:C,chr6:31271630:G:T,chr6:31351254:G:T
PSOR_combined	4	2.27E-19	chr6:31447530:A:G	chr6:31249331:CA:C,chr6:31271630:G:T,chr6:31351254:G:T,chr6:31447530:A:G
PSOR_combined	5	9.37E-12	chr6:31175734:G:A	chr6:31175734:G:A,chr6:31249331:CA:C,chr6:31271630:G:T,chr6:31351254:G:T,chr6:31447530:A:G
PSOR_combined	6	2.76E-10	chr6:31493266:G:A	chr6:31175734:G:A,chr6:31249331:CA:C,chr6:31271630:G:T,chr6:31351254:G:T,chr6:31447530:A:G,chr6:31493266:G:A
PSOR_combined	7	8.05E-11	chr6:31274414:A:G	chr6:31175734:G:A,chr6:31249331:CA:C,chr6:31271630:G:T,chr6:31274414:A:G,chr6:31351254:G:T,chr6:31447530:A:G,chr6:31493266:G:A
PSOR_combined	8	3.14E-08	chr6:31133092:G:A	chr6:31133092:G:A,chr6:31175734:G:A,chr6:31249331:CA:C,chr6:31271630:G:T,chr6:31274414:A:G,chr6:31351254:G:T,chr6:31447530:A:G,chr6:31493266:G:A
PSOR_combined	9	4.10E-06	chr6:31392478:A:C	chr6:31133092:G:A,chr6:31175734:G:A,chr6:31249331:CA:C,chr6:31271630:G:T,chr6:31274414:A:G,chr6:31351254:G:T,chr6:31392478:A:C,chr6:31447530:A:G,chr6:31493266:G:A
PSOR_combined	10	3.81E-06	chr6:31113632:T:C	chr6:31113632:T:C,chr6:31133092:G:A,chr6:31175734:G:A,chr6:31249331:CA:C,chr6:31271630:G:T,chr6:31274414:A:G,chr6:31351254:G:T,chr6:31392478:A:C,chr6:31447530:A:G,chr6:31493266:G:A
PSOR_combined	11	3.31E-08	chr6:30991224:G:A	chr6:30991224:G:A,chr6:31113632:T:C,chr6:31133092:G:A,chr6:31175734:G:A,chr6:31249331:CA:C,chr6:31271630:G:T,chr6:31274414:A:G,chr6:31351254:G:T,chr6:31392478:A:C,chr6:31447530:A:G,chr6:31493266:G:A
PSOR_combined	12	1.65E-06	chr6:31968280:G:T	chr6:30991224:G:A,chr6:31113632:T:C,chr6:31133092:G:A,chr6:31175734:G:A,chr6:31249331:CA:C,chr6:31271630:G:T,chr6:31274414:A:G,chr6:31351254:G:T,chr6:31392478:A:C,chr6:31447530:A:G,chr6:31493266:G:A,chr6:31968280:G:T
PSOR_combined	13	1.10E-47	chr6:31472229:C:T	chr6:30991224:G:A,chr6:31113632:T:C,chr6:31133092:G:A,chr6:31175734:G:A,chr6:31249331:CA:C,chr6:31271630:G:T,chr6:31274414:A:G,chr6:31351254:G:T,chr6:31392478:A:C,chr6:31447530:A:G,chr6:31472229:C:T,chr6:31493266:G:A,chr6:31968280:G:T
spiro_asthma	1	2.88E-06	chr1:152714582:C:T	chr1:152714582:C:T
spiro_asthma	2	1.58E-08	chr1:152335767:T:C	chr1:152335767:T:C,chr1:152714582:C:T
spiro_asthma	3	3.63E-09	chr1:152307634:C:T	chr1:152307634:C:T,chr1:152335767:T:C,chr1:152714582:C:T
spiro_asthma	4	4.54E-32	chr1:152167184:G:A	chr1:152167184:G:A,chr1:152307634:C:T,chr1:152335767:T:C,chr1:152714582:C:T
spiro_asthma	5	3.85E-11	chr1:152911635:G:A	chr1:152167184:G:A,chr1:152307634:C:T,chr1:152335767:T:C,chr1:152714582:C:T,chr1:152911635:G:A
spiro_dermatitis	1	3.27E-08	chr1:152298743:A:T	chr1:152298743:A:T
spiro_dermatitis	2	6.17E-178	chr1:151948993:G:A	chr1:151948993:G:A,chr1:152298743:A:T
spiro_dermatitis	3	5.90E-07	chr1:151769943:GC:G	chr1:151769943:GC:G,chr1:151948993:G:A,chr1:152298743:A:T
spiro_dermatitis	4	1.34E-09	chr1:151883939:G:A	chr1:151769943:GC:G,chr1:151883939:G:A,chr1:151948993:G:A,chr1:152298743:A:T
spiro_dermatitis	5	4.99E-07	1:152079770:TCTATCCC	chr1:151769943:GC:G,chr1:151883939:G:A,chr1:151948993:G:A,chr1:152079770:TCTATCCCA:T,chr1:152298743:A:T
spiro_psoriasis	1	9.68E-129	chr6:31346230:AC:A	chr6:31346230:AC:A
spiro_psoriasis	2	2.98E-37	chr6:31149411:C:T	chr6:31149411:C:T,chr6:31346230:AC:A
spiro_psoriasis	3	5.89E-32	chr6:31364750:G:A	chr6:31149411:C:T,chr6:31346230:AC:A,chr6:31364750:G:A
spiro_psoriasis	4	9.19E-28	chr6:31274414:A:G	chr6:31149411:C:T,chr6:31274414:A:G,chr6:31346230:AC:A,chr6:31364750:G:A
spiro_psoriasis	5	1.05E-12	chr6:31331276:A:AG	chr6:31149411:C:T,chr6:31274414:A:G,chr6:31331276:A:AG,chr6:31346230:AC:A,chr6:31364750:G:A
spiro_psoriasis	6	1.22E-13	chr6:31493276:T:C	chr6:31149411:C:T,chr6:31274414:A:G,chr6:31331276:A:AG,chr6:31346230:AC:A,chr6:31364750:G:A,chr6:31493276:T:C
spiro_psoriasis	7	4.68E-06	chr6:32005239:A:G	chr6:31149411:C:T,chr6:31274414:A:G,chr6:31331276:A:AG,chr6:31346230:AC:A,chr6:31364750:G:A,chr6:31493276:T:C,chr6:32005239:A:G
spiro_psoriasis	8	1.81E-06	chr6:32046798:A:G	chr6:31149411:C:T,chr6:31274414:A:G,chr6:31331276:A:AG,chr6:31346230:AC:A,chr6:31364750:G:A,chr6:31493276:T:C,chr6:32005239:A:G,chr6:32046798:A:G
spiro_psoriasis	9	4.44E-06	chr6:31393157:C:T	chr6:31149411:C:T,chr6:31274414:A:G,chr6:31331276:A:AG,chr6:31346230:AC:A,chr6:31364750:G:A,chr6:31393157:C:T,chr6:31493276:T:C,chr6:32005239:A:G,chr6:32046798:A:G
spiro_psoriasis	10	1.49E-11	chr6:31343256:A:G	chr6:31149411:C:T,chr6:31274414:A:G,chr6:31331276:A:AG,chr6:31343256:A:G,chr6:31346230:AC:A,chr6:31364750:G:A,chr6:31393157:C:T,chr6:31493276:T:C,chr6:32005239:A:G,chr6:32046798:A:G
spiro_psoriasis	11	1.49E-06	chr6:31271630:G:T	chr6:31149411:C:T,chr6:31271630:G:T,chr6:31274414:A:G,chr6:31331276:A:AG,chr6:31343256:A:G,chr6:31346230:AC:A,chr6:31364750:G:A,chr6:31393157:C:T,chr6:31493276:T:C,chr6:32005239:A:G,chr6:32046798:A:G
spiro_psoriasis	12	1.62E-06	chr6:31268870:C:T	chr6:31149411:C:T,chr6:31268870:C:T,chr6:31271630:G:T,chr6:31274414:A:G,chr6:31331276:A:AG,chr6:31343256:A:G,chr6:31346230:AC:A,chr6:31364750:G:A,chr6:31393157:C:T,chr6:31493276:T:C,chr6:32005239:A:G,chr6:32046798:A:G
spiro_psoriasis	13	1.06E-07	chr6:31357042:C:T	chr6:31149411:C:T,chr6:31268870:C:T,chr6:31271630:G:T,chr6:31274414:A:G,chr6:31331276:A:AG,chr6:31343256:A:G,chr6:31346230:AC:A,chr6:31357042:C:T,chr6:31364750:G:A,chr6:31393157:C:T,chr6:31493276:T:C,chr6:32005239:A:G,chr6:32046798:A:G
spiro_psoriasis	14	1.58E-08	chr6:31105287:A:G	chr6:31105287:A:G,chr6:31149411:C:T,chr6:31268870:C:T,chr6:31271630:G:T,chr6:31274414:A:G,chr6:31331276:A:AG,chr6:31343256:A:G,chr6:31346230:AC:A,chr6:31357042:C:T,chr6:31364750:G:A,chr6:31393157:C:T,chr6:31493276:T:C,chr6:32005239:A:G,chr6:32046798:A:G
spiro_splenomegaly	1	4.88E-07	chr19:57963032:T:TA	chr19:57963032:T:TA
spiro_splenomegaly	2	3.54E-10	chr19:57822158:C:CT	chr19:57822158:C:CT,chr19:57963032:T:TA
spiro_splenomegaly	3	1.36E-48	chr19:58198807:CT:C	chr19:57822158:C:CT,chr19:57963032:T:TA,chr19:58198807:CT:C

Table A.8: Comparison of significant hits from additive haplotype-collapsing association analysis and `SAIGE-GENE+` burden testing. I first compare well phased variants (PP > 0.90). Recognizing that this analysis deviated from the regular usage of `SAIGE-GENE+`, which ignores phasing confidence, I also performed an analysis without filtering variants by PP (PP > 0.50). The column 'P1<P2' indicates whether the haplotype-based collapsing approach produces lower P-values than `SAIGE-GENE+` burden testing. This table contains all Bonferroni significant ($P < 9.8 \times 10^{-9}$) hits reported in the chapter.

Trait	Gene	P-values at PP>0.90			P-values at PP>=0.50		
		Haplotype (P1)	Burden (P2)	P1 < P2	Haplotype (P1)	Burden (P2)	P1 < P2
PSOR_combined	*MICA*	8.93E-137	9.75E-01	Yes	1.04E-136	4.38E-98	Yes
PSOR_combined	*HLA-G*	2.68E-12	2.53E-12	No	3.62E-12	6.41E-12	Yes
spiro_psoriasis	*MICA*	7.58E-109	4.13E-01	Yes	8.45E-109	1.01E-81	Yes
spiro_psoriasis	*HLA-G*	2.06E-12	2.94E-12	Yes	3.83E-13	3.87E-13	Yes
spiro_dermatitis	*FLG*	5.89E-23	1.76E-11	Yes	4.71E-24	2.35E-21	Yes
BC_combined	*BRCA2*	6.16E-15	4.55E-11	Yes	2.28E-19	3.31E-19	Yes
BC_combined	*CHEK2*	3.34E-15	4.52E-15	Yes	2.44E-15	3.43E-15	Yes
spiro_primary_malignancy_breast	*BRCA2*	2.93E-10	6.40E-08	Yes	2.66E-13	3.51E-13	Yes
spiro_primary_malignancy_breast	*CHEK2*	3.19E-14	4.30E-14	Yes	1.22E-13	1.78E-13	Yes
spiro_asthma	*FLG*	1.71E-13	4.53E-12	Yes	8.12E-14	4.22E-13	Yes
spiro_psoriatic_arthropathy	*MICA*	2.80E-12	6.08E-02	Yes	2.83E-12	1.42E-10	Yes
spiro_hypo_or_hyperthyroidism	*MICA*	6.83E-12	9.09E-01	Yes	6.54E-12	2.09E-09	Yes

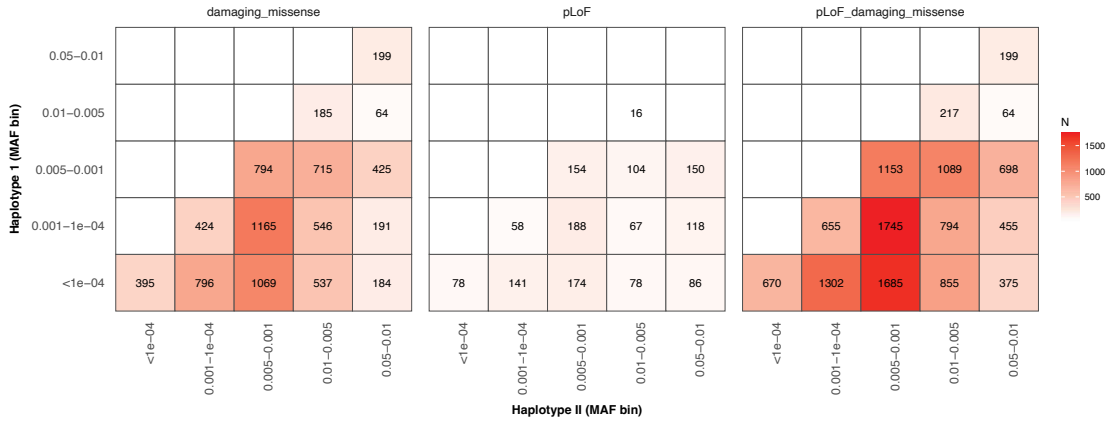


Figure A.9: Allele frequencies of variants in the CH state. Heatmap of allele counts for variants in CH state stratified by predicted variant consequence (damaging missense, pLoF or pLoF+damaging missense). I plot the MAC for variants residing on the most common haplotype (y-axis) versus the rarest haplotype (x-axis).

Table A.9: Variant annotation counts before and after filtering to sites with PP$\geq$ 0.90. Distribution of most severe consequences for variants in canonical transcripts across 22 autosomes, before and after filtering to PP$\geq$ 0.90. The table shows counts for quality-controlled variants and samples, categorized by their predicted functional impact.

Category	Predicted consequence	VEP annotation	PP >= 0.5			PP >= 0.90		
			n (MAC>0)	n (MAC=1)	% (MAC=1)	n (MAC>0)	n (MAC=1)	% (MAC=1)
PTV	LC	splice_acceptor_variant	942	504	53.5	489	74	15.1
PTV	LC	splice_donor_variant	633	306	48.3	350	54	15.4
PTV	LC	stop_gained	12788	5693	44.5	7398	864	11.7
PTV	LC	frameshift_variant	18533	8841	47.7	10465	1301	12.4
PTV	Total LC		32896	15344	46.6	18702	2293	12.3
PTV	pLoF	splice_acceptor_variant	26578	15267	57.4	12833	2201	17.2
PTV	pLoF	splice_donor_variant	34631	18810	54.3	17391	2726	15.7
PTV	pLoF	stop_gained	103690	53166	51.3	52026	7339	14.1
PTV	pLoF	frameshift_variant	129817	72272	55.7	64049	10114	15.8
PTV	Total pLoF		294716	159515	54.1	146299	22380	15.3
Total PTV			327612	174859	53.4	165001	24673	15
Missense	damaging_missense	start_lost	1075	568	52.8	551	90	16.3
Missense	damaging_missense	missense_variant	417314	202488	48.5	223606	29816	13.3
Missense	Total damaging_missense		418389	203056	48.5	224157	29906	13.3
Missense	other_missense	stop_lost	3375	1795	53.2	1802	301	16.7
Missense	other_missense	start_lost	9215	4418	47.9	5038	663	13.2
Missense	other_missense	inframe_insertion	17389	6828	39.3	10957	1095	10
Missense	other_missense	inframe_deletion	41645	17347	41.7	25128	2784	11.1
Missense	other_missense	missense_variant	3081420	1317019	42.7	1836789	218884	11.9
Missense	Total other_missense		3153044	1347407	42.7	1879714	223727	11.9
Total Missense			3571433	1550463	43.4	2103871	253633	12.1
Synonymous	synonymous	stop_retained_variant	1718	833	48.5	968	141	14.6
Synonymous	synonymous	synonymous_variant	1635045	653611	40	1012937	114709	11.3
Synonymous	Total synonymous		1636763	654444	40	1013905	114850	11.3
Total Synonymous			1636763	654444	40	1013905	114850	11.3
Non_coding	non_coding	5_prime_UTR_variant	237528	102095	43	143280	18069	12.6
Non_coding	non_coding	3_prime_UTR_variant	310372	131734	42.4	188128	22954	12.2
Non_coding	non_coding	intron_variant	2836382	1218740	43	1709955	212588	12.4
Non_coding	non_coding	upstream_gene_variant	692245	299032	43.2	408302	47137	11.5
Non_coding	non_coding	downstream_gene_variant	1040319	439582	42.3	619954	68939	11.1
Total Non_coding			5116846	2191183	42.8	3069619	369687	12
Total			10652654	4570949	42.9	6352396	762843	12.01

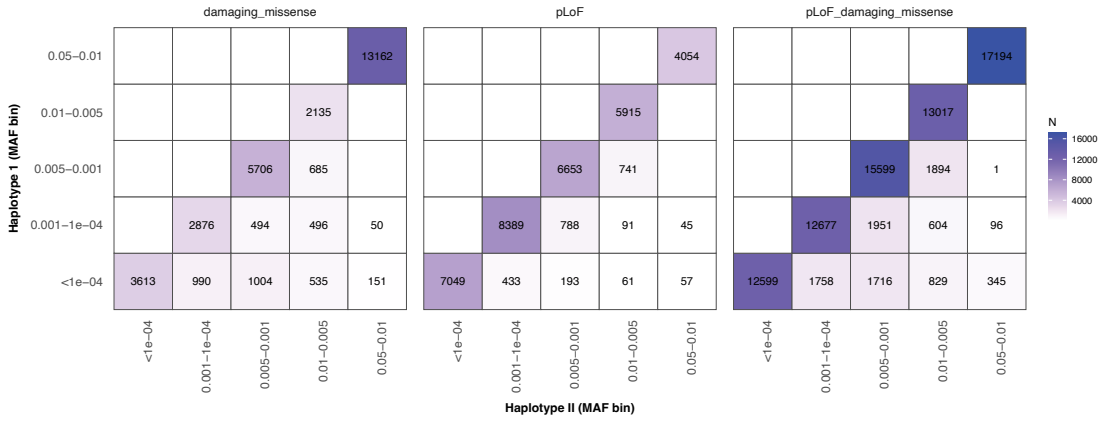


Figure A.10: Allele frequencies of variants in *cis*. Heatmap of allele counts for co-occurring variants on the same haplotype stratified by predicted variant consequence (damaging missense, pLoF or pLoF+damaging missense). The most common variant on the haplotype versus the least common are plotted.

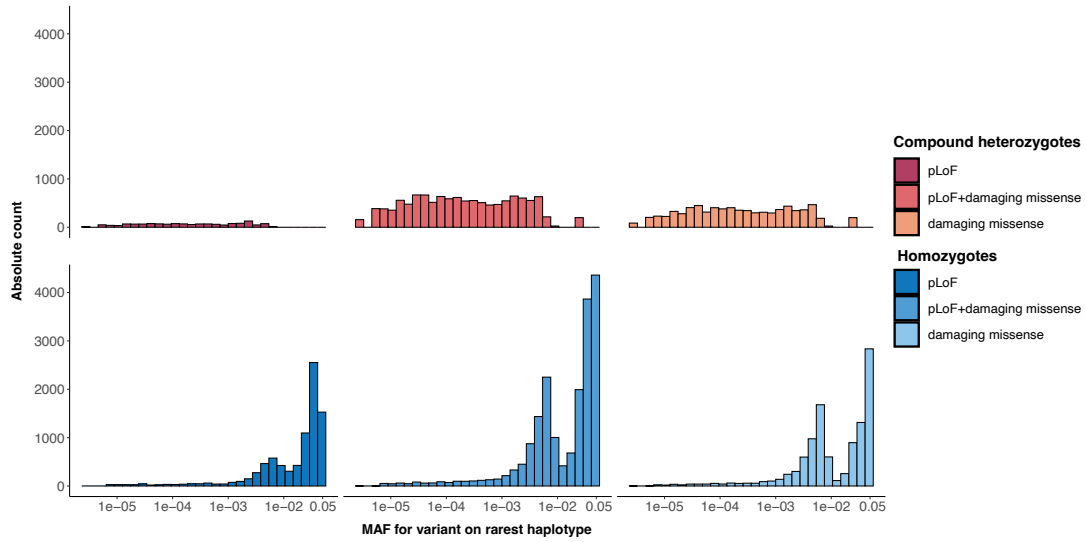


Figure A.11: Distribution of observed variants across samples by allele frequency. Histogram of unique biallelic variant (CH and homozygotes) prevalence across the allele frequency spectrum. For a qualifying CH variant, the AF corresponding to the alternate allele on the rarest haplotype are plotted.

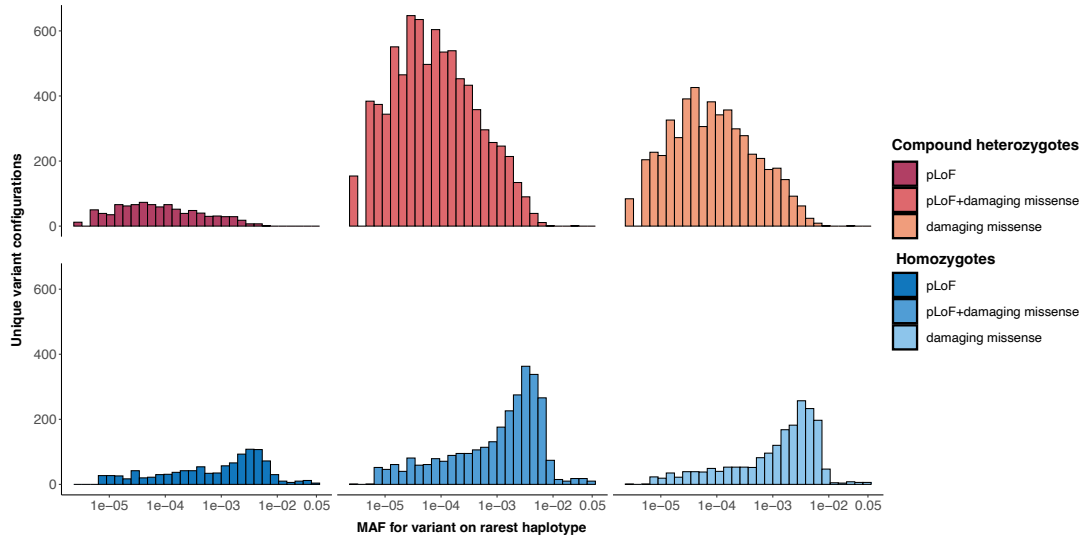


Figure A.12: Distribution of unique variants observed by allele frequency. Histogram of biallelic variant (CH and homozygotes) count for all gene-samples pairs in the analysis. For a qualifying CH variant, the AF corresponding to the alternate allele on the rarest haplotype are plotted.

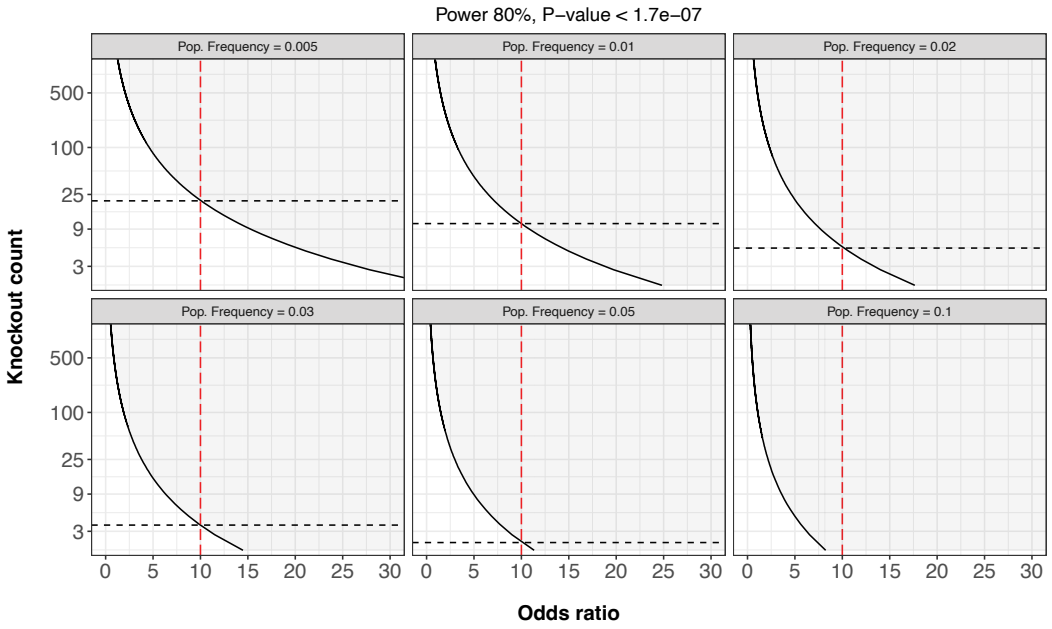


Figure A.13: Power analysis to determine the required number of biallelic variants to detect specific ORs. Power analysis to determine the required number of biallelic variants to detect specific ORs at 80% power at Bonferroni significance ($P < 1.7 \times 10^{-7}$). I repeat the analysis while varying trait population prevalence assuming 823 (0.5%), 1766 (1%), 3532 (2%), 5298 (3%), 8829 (5%) cases out of 176,587 total individuals. The dashed red lines in the plot demonstrate the required number of biallelic variants to detect an OR ≥ 10.

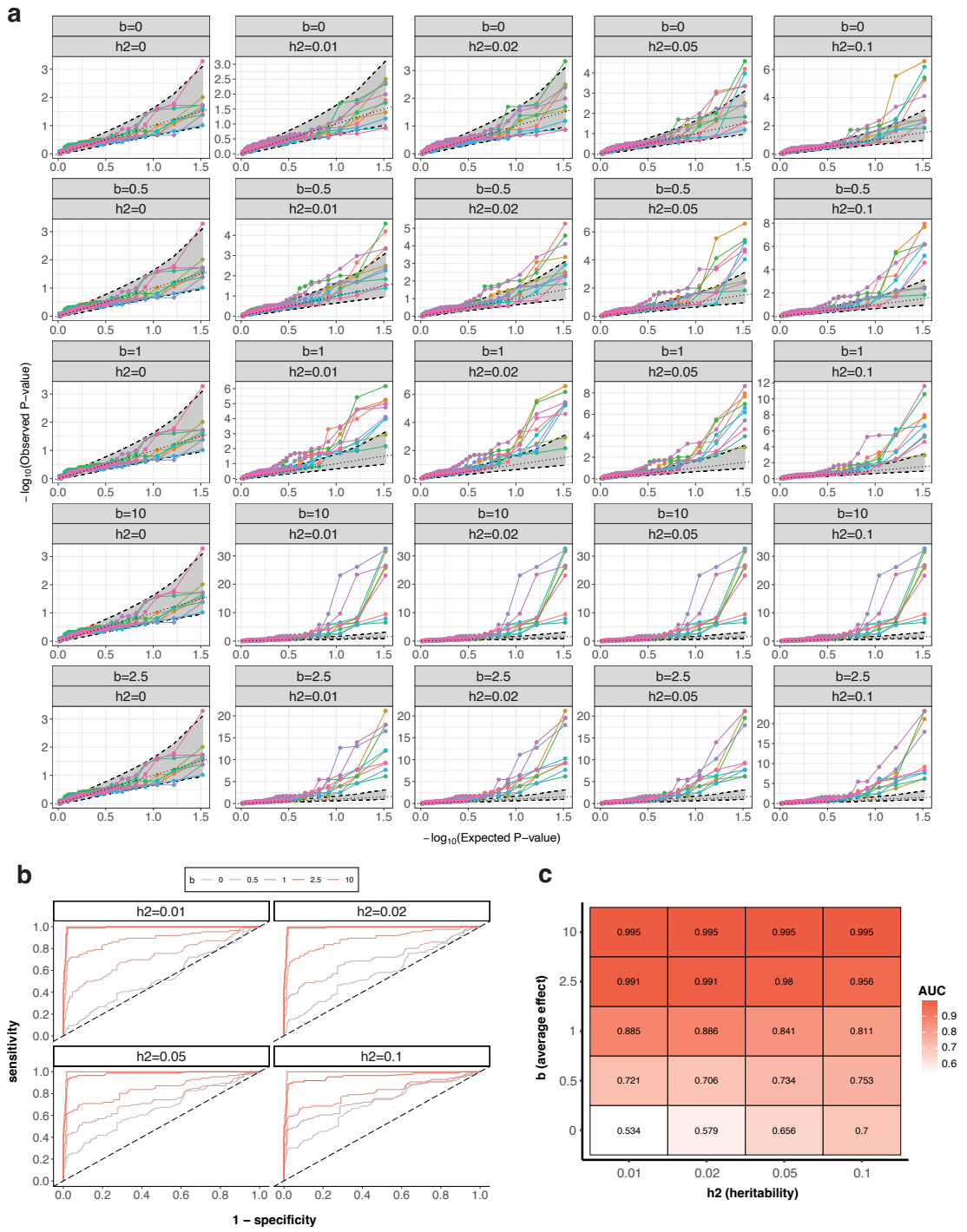

a
b=0 h2=0
b=0 h2=0.01
b=0 h2=0.02
b=0 h2=0.05
b=0 h2=0.1
b=0.5 h2=0
b=0.5 h2=0.01
b=0.5 h2=0.02
b=0.5 h2=0.05
b=0.5 h2=0.1
b=1 h2=0
b=1 h2=0.01
b=1 h2=0.02
b=1 h2=0.05
b=1 h2=0.1
b=10 h2=0
b=10 h2=0.01
b=10 h2=0.02
b=10 h2=0.05
b=10 h2=0.1
b=2.5 h2=0
b=2.5 h2=0.01
b=2.5 h2=0.02
b=2.5 h2=0.05
b=2.5 h2=0.1
−log10(Observed P−value)
−log10(Expected P−value)
b
b 0 0.5 1 2.5 10
h2=0.01
h2=0.02
h2=0.05
h2=0.1
sensitivity
1 − specificity
c
0.995 0.995 0.995 0.995
0.991 0.991 0.98 0.956
0.885 0.886 0.841 0.811
0.721 0.706 0.734 0.753
0.534 0.579 0.656 0.7
b (average effect)
h2 (heritability)
AUC
0.9
0.8
0.7
0.6

Figure A.14

Figure A.14 *(previous page)*: Simulation study to test our ability to detect biallelic effects in the presence of true effects. I simulate phenotypic data applied to 100,000 genetically-ascertained NFE on chromosome 22 (Methods) under the liability-threshold model assuming a spike and slab genetic architecture. I assume a 10% disease prevalence and 25% causal genes, and consider varying levels of phenotypic variance explained by these effects $\in \{0, 0.01, 0.02, 0.05, 0.10\}$. I then apply SAIGE to the simulated phenotypes, testing for an association between presence of a biallelic variant in each gene and case status. a) Each panel indicates a set of simulations assuming varying levels of heritability and average effect as labeled in the subtitles. In each panel, I plot the true effect size in the simulation for a given gene on x-axis against the corresponding $-\log_{10}(P)$ value of association. Areas of circles correspond to the number of samples harboring biallelic damaging variants in the 100,000 samples according to the legend. b) To assess the sensitivity and specificity of our approach, I created Receiver Operating Charateristic Area Under the Curve (ROC-AUC) curves for each combination of increasing phenotypic variance explained (facet) and increasing average affect (red lines). c) For each ROC-AUC curve from b, I calculate the AUC. White indicates low AUC and red indicates higher AUC.

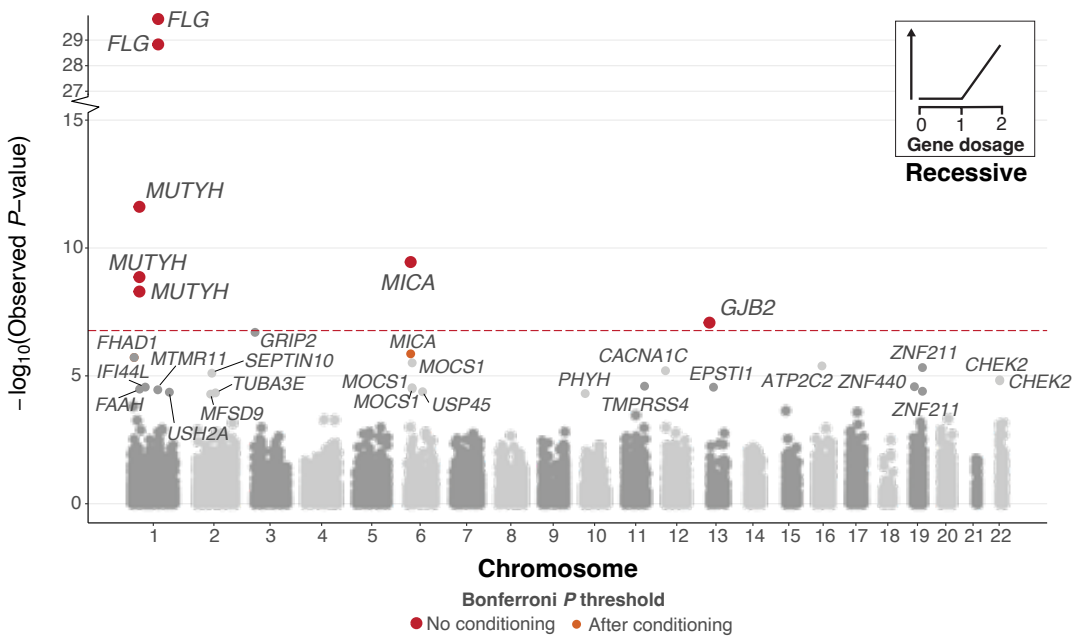


Figure A.15: Recessive association analysis without accounting for PRS. Manhattan plot of recessive gene-trait associations. -$\log_{10}(P)$ values plotted against chromosomal location. Red points indicate Bonferroni significant associations ($P < 1.68 \times 10^{-7}$). Gene symbols shown for associations with $P < 3.05 \times 10^{-6}$ (nominal significance). No additional conditioning was applied in this analysis.

Phenotype	Gene	No conditioning			Conditional on PRS			Conditional on PRS, common variants and rare variants			Conditional on nearby common variants		
		β	SE	*P*-value	β	SE	*P*-value	β	SE	*P*-value	β	SE	*P*-value
Asthma	FLG	8.39E-01	7.47E-02	1.48E-29	8.36E-01	7.48E-02	2.62E-29	7.54E-01	7.80E-02	2.09E-22	0.820429	7.45E-02	1.56E-28
Dermatitis (atopc/conta	FLG	7.60E-01	6.65E-02	1.50E-30	7.58E-01	6.64E-02	1.95E-30	6.34E-01	6.92E-02	2.65E-20	0.719503	6.65E-02	1.49E-27
Colorectal Cancer	MUTYH	1.08E+01	1.56E+00	2.45E-12	1.07E+01	1.55E+00	2.29E-12	1.08E+01	1.55E+00	1.55E-12	NA	NA	NA
Primary Malignancy_co	MUTYH	13.0624	2.20E+00	1.37E-09	1.30E+01	2.19E+00	1.36E-09	1.28E+01	2.16E+00	1.54E-09	NA	NA	NA
Benign neoplasm of co	MUTYH	3.42754	5.98E-01	5.05E-09	3.24E+00	5.75E-01	8.45E-09	3.17E+00	5.68E-01	1.20E-08	NA	NA	NA
Hearing loss	GJB2	3.82E+00	7.31E-01	8.31E-08	3.72E+00	7.16E-01	9.93E-08	3.81E+00	7.12E-01	4.25E-08	NA	NA	NA
Pulmonary collapse (e:	GRIP2	9.37E+00	1.85E+00	1.94E-07	9.37E+00	1.85E+00	1.94E-07	9.12E+00	1.81E+00	2.27E-07	NA	NA	NA
NAFLD	FHAD1	2.97E+01	6.42E+00	1.91E-06	2.96E+01	6.39E+00	1.88E-06	2.97E+01	6.24E+00	9.27E-07	29.4976	6.20E+00	9.68E-07
Lower Respiratory Trac	MOCS1	6.23E+00	1.38E+00	3.10E-06	5.93E+00	1.32E+00	3.69E-06	6.05E+00	1.30E+00	1.76E-06	NA	NA	NA
Splenomegaly	ZNF211	1.15E+02	2.60E+01	4.74E-06	1.15E+02	2.60E+01	4.74E-06	1.14E+02	2.53E+01	3.16E-06	114.776	2.51E+01	2.45E-06
Fatty Liver	CACNA1C	2.24E+01	5.13E+00	6.28E-06	1.98E+01	4.60E+00	8.64E-06	2.03E+01	4.52E+00	3.61E-06	NA	NA	NA
COPD	ATP2C2	2.19E+00	5.05E-01	6.94E-06	2.13E+00	4.95E-01	8.63E-06	2.30E+00	5.05E-01	2.65E-06	NA	NA	NA
Hyperplasia of prostate	SEPTIN10	4.44E+00	1.03E+00	7.87E-06	4.56E+00	1.05E+00	6.80E-06	4.41E+00	1.02E+00	7.64E-06	NA	NA	NA
Other or unspecified in	MOCS1	3.94E+00	9.84E-01	3.20E-05	3.91E+00	9.79E-01	3.21E-05	4.14E+00	9.66E-01	9.01E-06	NA	NA	NA
Dementia	FAAH	4.36E+01	1.09E+01	3.29E-05	5.24E+01	1.28E+01	2.15E-05	5.28E+01	1.24E+01	1.06E-05	NA	NA	NA
Anorectal prolapse	ZNF440	5.27E+01	1.30E+01	2.65E-05	5.27E+01	1.30E+01	2.65E-05	5.26E+01	1.26E+01	1.46E-05	NA	NA	NA
Leukaemia	CHEK2	6.37E+01	1.53E+01	1.56E-05	6.37E+01	1.53E+01	1.56E-05	6.04E+01	1.45E+01	1.63E-05	NA	NA	NA
Diaphragmatic hernia	EPSTI1	9.90E-01	2.46E-01	2.78E-05	1.01E+00	2.47E-01	2.39E-05	9.99E-01	2.42E-01	1.79E-05	NA	NA	NA
depression	IFI44L	5.57351	1.38E+00	2.75E-05	5.35E+00	1.34E+00	3.38E-05	5.33E+00	1.29E+00	1.93E-05	NA	NA	NA
Pulmonary embolism	MOCS1	1.38E+01	3.44E+00	2.92E-05	1.34E+01	3.35E+00	3.12E-05	1.33E+01	3.25E+00	2.04E-05	NA	NA	NA
Leukaemia	ZNF211	4.31E+01	1.09E+01	4.00E-05	4.31E+01	1.09E+01	4.00E-05	4.28E+01	1.06E+01	2.61E-05	NA	NA	NA
Diverticular disease of	TUBA3E	3.87E+00	9.90E-01	4.62E-05	3.42E+00	9.05E-01	7.84E-05	3.58E+00	8.85E-01	2.64E-05	NA	NA	NA
Actinic keratosis	USP45	4.00E+00	1.02E+00	4.17E-05	3.88E+00	9.96E-01	4.87E-05	3.93E+00	9.74E-01	2.79E-05	NA	NA	NA
Visual impairment and	USH2A	1.33E+01	3.39E+00	4.35E-05	1.33E+01	3.39E+00	4.35E-05	1.33E+01	3.34E+00	3.50E-05	NA	NA	NA
Benign neoplasm of st	PHYH	1.14E+01	2.94E+00	4.92E-05	1.14E+01	2.94E+00	4.92E-05	1.10E+01	2.80E+00	4.36E-05	NA	NA	NA
Hearing loss	MFSD9	2.92E+00	7.52E-01	5.18E-05	2.83E+00	7.37E-01	5.98E-05	2.76E+00	7.07E-01	4.86E-05	NA	NA	NA
Type 2 Diabetes	TMPRSS4	1.29E+01	3.19E+00	2.53E-05	9.45E+00	2.46E+00	6.24E-05	9.17E+00	2.36E+00	5.06E-05	NA	NA	NA
Primary Malignancy_O	CHEK2	5.93E+01	1.42E+01	1.43E-05	5.93E+01	1.42E+01	1.43E-05	5.13E+01	1.36E+01	8.30E-05	NA	NA	NA
Type 2 Diabetes	MTMR11	1.92E+00	4.83E-01	3.51E-05	1.69E+00	4.47E-01	7.67E-05	1.60E+00	4.30E-01	9.61E-05	NA	NA	NA
Psoriasis	MICA	1.07E+00	2.29E-01	1.38E-06	1.10E+00	2.31E-01	1.04E-06	-2.57E-01	1.96E-01	1.90E-01	-2.60E-01	1.95E-01	1.82E-01
psoriasis	MICA	1.70E+00	2.76E-01	3.55E-10	1.77E+00	2.82E-01	1.75E-10	3.60E-02	2.18E-01	8.69E-01	9.97E-02	2.15E-01	6.43E-01

Phenotype	Gene	Additive association analysis conditional on PRS			Conditional on additive encoding of haplotypes and PRS			Only CH and conditional on PRS			Only homozygotes and conditional on PRS		
		β	SE	*P*-value	β	SE	*P*-value	β	SE	*P*-value	β	SE	*P*-value
Asthma	FLG	1.77E-01	2.44E-02	1.71E-13	7.62E-01	7.79E-02	7.42E-23	9.06E-01	8.93E-02	1.70E-24	6.59E-01	1.36E-01	6.51E-07
Dermatitis (atopc/conta	FLG	2.20E-01	2.25E-02	5.89E-23	6.27E-01	6.89E-02	4.32E-20	7.13E-01	7.81E-02	3.40E-20	8.64E-01	1.26E-01	3.02E-12
Colorectal Cancer	MUTYH	2.51E-01	1.34E-01	3.08E-02	1.07E+01	1.54E+00	1.70E-12	1.21E+01	2.03E+00	1.19E-09	8.45E+00	2.33E+00	1.44E-04
Primary Malignancy_co	MUTYH	5.43E-01	1.63E-01	4.30E-04	1.28E+01	2.15E+00	1.43E-09	1.22E+01	2.74E+00	4.17E-06	1.42E+01	3.53E+00	2.78E-05
Benign neoplasm of co	MUTYH	6.99E-02	6.85E-02	3.08E-01	3.27E+00	5.70E-01	5.05E-09	3.09E+00	7.29E-01	1.14E-05	3.48E+00	9.22E-01	8.07E-05
Hearing loss	GJB2	1.06E-01	6.38E-02	9.57E-02	3.68E+00	7.01E-01	7.78E-08	2.35E+00	7.19E-01	5.36E-04	9.84E+00	2.17E+00	2.81E-06
Pulmonary collapse (e:	GRIP2	6.17E-01	2.01E-01	1.08E-03	9.13E+00	1.81E+00	2.22E-07	NA	NA	NA	NA	NA	NA
NAFLD	FHAD1	1.46E-01	2.12E-01	4.91E-01	2.97E+01	6.23E+00	9.50E-07	NA	NA	NA	NA	NA	NA
Lower Respiratory Trac	MOCS1	5.72E-02	1.16E-01	6.23E-01	6.00E+00	1.30E+00	1.82E-06	NA	NA	NA	NA	NA	NA
Splenomegaly	ZNF211	1.51E+00	6.92E-01	1.48E-02	1.15E+02	2.52E+01	2.58E-06	NA	NA	NA	NA	NA	NA
Fatty Liver	CACNA1C	-9.03E-02	2.34E-01	6.99E-01	2.02E+01	4.50E+00	3.66E-06	NA	NA	NA	NA	NA	NA
COPD	ATP2C2	-3.99E-02	5.94E-02	5.02E-01	2.22E+00	4.86E-01	2.58E-06	2.24E+00	5.37E-01	1.54E-05	NA	NA	NA
Hyperplasia of prostate	SEPTIN10	3.18E-01	1.23E-01	4.81E-03	4.35E+00	1.01E+00	9.03E-06	NA	NA	NA	NA	NA	NA
Other or unspecified in	MOCS1	-4.66E-02	9.37E-02	6.19E-01	4.04E+00	9.55E-01	1.18E-05	NA	NA	NA	NA	NA	NA
Dementia	FAAH	9.22E-02	3.39E-01	7.85E-01	5.28E+01	1.24E+01	1.04E-05	NA	NA	NA	NA	NA	NA
Anorectal prolapse	ZNF440	6.31E-01	3.70E-01	8.82E-02	5.26E+01	1.26E+01	1.42E-05	NA	NA	NA	NA	NA	NA
Leukaemia	CHEK2	2.62E+00	5.77E-01	2.69E-06	6.19E+01	1.46E+01	1.15E-05	NA	NA	NA	NA	NA	NA
Diaphragmatic hernia	EPSTI1	5.05E-02	4.27E-02	2.37E-01	9.97E-01	2.42E-01	1.85E-05	NA	NA	NA	NA	NA	NA
depression	IFI44L	7.99E-02	8.39E-02	3.41E-01	5.33E+00	1.29E+00	1.95E-05	NA	NA	NA	NA	NA	NA
Pulmonary embolism	MOCS1	3.86E-01	2.20E-01	7.92E-02	1.33E+01	3.23E+00	1.93E-05	NA	NA	NA	NA	NA	NA
Leukaemia	ZNF211	4.61E-01	3.84E-01	2.30E-01	4.32E+01	1.05E+01	2.08E-05	NA	NA	NA	NA	NA	NA
Diverticular disease of	TUBA3E	1.45E-03	8.20E-02	9.86E-01	3.48E+00	8.75E-01	3.55E-05	4.76E+00	1.47E+00	6.22E-04	2.32E+00	1.10E+00	1.75E-02
Actinic keratosis	USP45	1.63E-02	1.01E-01	8.73E-01	3.95E+00	9.68E-01	2.22E-05	5.74E+00	1.46E+00	4.15E-05	1.48E+00	1.01E+00	1.41E-01
Visual impairment and	USH2A	8.64E-01	2.93E-01	1.59E-03	1.28E+01	3.24E+00	3.73E-05	NA	NA	NA	NA	NA	NA
Benign neoplasm of st	PHYH	1.46E-01	1.69E-01	3.89E-01	1.15E+01	2.83E+00	2.62E-05	NA	NA	NA	NA	NA	NA
Hearing loss	MFSD9	9.96E-02	7.21E-02	1.67E-01	2.79E+00	7.08E-01	4.15E-05	NA	NA	NA	NA	NA	NA
Type 2 Diabetes	TMPRSS4	3.38E-01	1.42E-01	8.57E-03	9.21E+00	2.35E+00	4.44E-05	NA	NA	NA	NA	NA	NA
Primary Malignancy_O	CHEK2	1.05E+00	5.39E-01	2.61E-02	5.92E+01	1.37E+01	8.04E-06	NA	NA	NA	NA	NA	NA
Type 2 Diabetes	MTMR11	1.26E-01	6.61E-02	5.60E-02	1.62E+00	4.29E-01	7.95E-05	1.99E+00	5.49E-01	1.44E-04	1.04E+00	6.32E-01	9.94E-02
Psoriasis	MICA	1.45E+00	6.56E-02	#######	-3.21E-01	1.96E-01	1.01E-01	9.93E-01	6.15E-01	1.06E-01	1.11E+00	2.43E-01	2.58E-06
psoriasis	MICA	1.84E+00	7.40E-02	#######	7.77E-03	2.18E-01	9.72E-01	3.72E-01	6.66E-01	5.76E-01	1.92E+00	3.02E-01	9.40E-11

Table A.10: Overview of associations before and after conditioning on off-chromosome PRS, nearby common variation, and the burden of rare variation. The analysis was also carried out for CHs and homozygotes independently while conditioning on PRS when applicable. In addition, I aggregated rare variants by haplotype and modeled the number of affected haplotypes in each individual. Subsequently, I performed two analyses: (1) an additive association analysis using the haplotype encoding, and (2) a recessive association analysis, conditioned on the additive encoding of haplotypes. Gene-trait associations with $P < 3.05 \times 10^{-6}$ are displayed in this table.

phenotype	hgnc symbol	Haplotypes affetected	BETA	SE	*P*-value	*P*<0.05
CC_combined	*MUTYH*	56	1.62083	0.282558	2.42E-09	Yes
spiro_benign_neoplasm_of_colon_rectum_anus_and_anal	*MUTYH*	56	0.572266	0.324138	0.019369875	Yes
spiro_primary_malignancy_colorectal_and_anus	*MUTYH*	56	1.30639	0.423845	0.000513621	Yes
spiro_asthma	*FLG*	1862	0.500134	0.0441844	2.63E-30	Yes
spiro_dermatitis	*FLG*	1862	0.484834	0.0420288	2.18E-31	Yes
spiro_hearing_loss	*GJB2*	356	0.833672	0.10661	1.32E-15	Yes
PSOR_combined	*MICA*	922	0.736622	0.10604	9.35E-13	Yes
spiro_psoriasis	*MICA*	922	0.689308	0.105103	1.36E-11	Yes

Table A.11: Replication of significant associations from the cross-sectional analysis. Only hits that achieved Bonferroni significance ($P < 1.68 \times 10^{-7}$) in the discovery analysis were carried forward for replication. 233,837 samples are analyzed in the replication cohort. I restricted the replication analysis to gene-traits with at least five biallelic pLoF/damaging missense or protein altering variant carriers.

phenotype	hgnc symbol	ensemble gene id	Hazard Ratio	Std. Error	one sided *P*-value
asthma	*FLG*	ENSG00000143631	2.31310517	0.084623724	1.8899829305939e-23
hearing_loss	*GJB2*	ENSG00000165474	5.38645491	0.189846006	3.66536860917759e-19
dermatitis	*FLG*	ENSG00000143631	1.9165276	0.091461524	5.7002742239104e-13
primary_malignancy_colorectal_and_anus	*MUTYH*	ENSG00000132781	16.8503307	0.597385407	1.13434578237927e-06
CC	*MUTYH*	ENSG00000132781	12.6516329	0.592350084	9.1671056963636e-06
benign_neoplasm_of_colon_rectum_anus_and_anal	*MUTYH*	ENSG00000132781	3.62935946	0.504531871	0.005310078
COPD	*ATP2C2*	ENSG00000064270	2.03271939	0.318948894	0.013070831

Table A.12: Replication analysis of the Bonferroni significant associations from the age-at-onset analysis. Bonferroni significant ($P < 1.89 \times 10^{-7}$) associations were carried forward for replication. Up to 233,837 samples are analyzed in the replication cohort. I restricted the replication analysis to gene-traits with at least five biallelic pLoF/damaging missense or protein altering variant carriers.

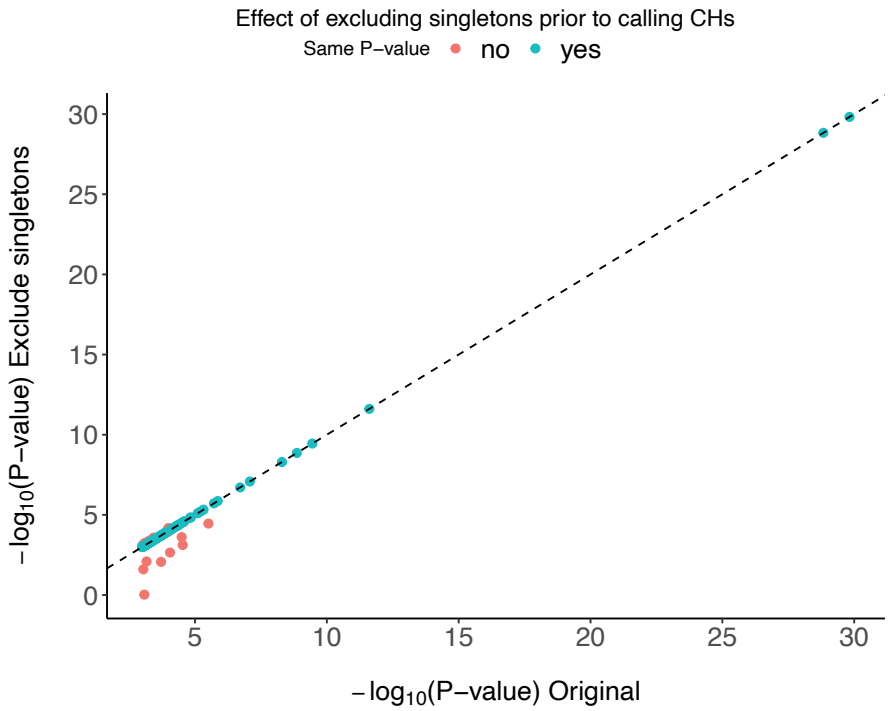


Figure A.16: Effect of excluding singletons prior to calling CH variants for any gene-traits with a $P < 0.001$ from the initial analysis. The P-value from the original cross-sectional analysis is shown on the x-axis, while the P-value for the same associated gene-traits without singletons is depicted on the y-axis. The dots are colored by whether the significance P-values change after excluding singletons.

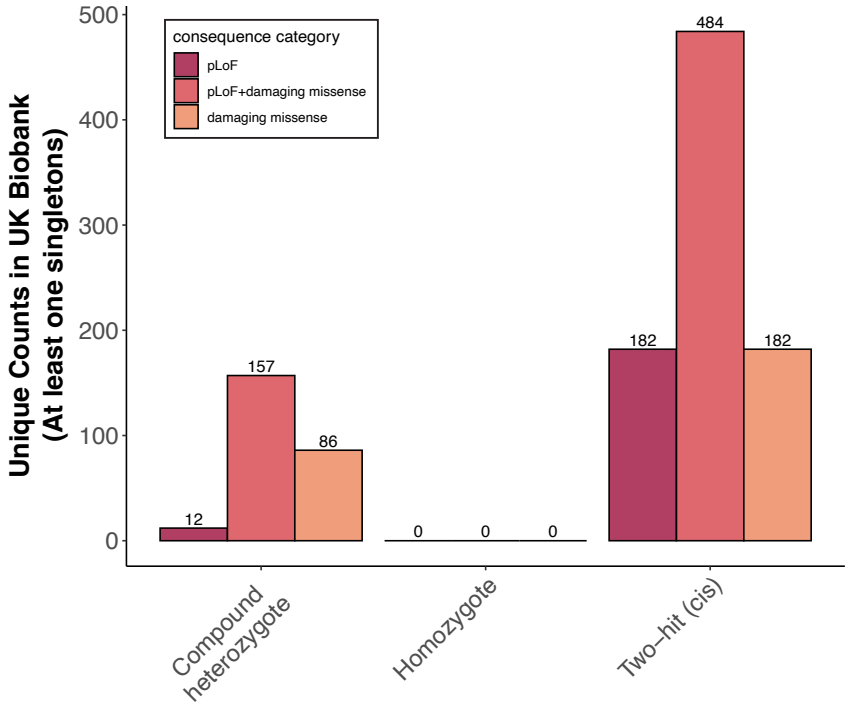


Figure A.17: Number of CH variants after filtering to those that consist of at least one singleton. I filter to CH variants that are comprised of at least one singleton across compound heterozygotes, homozygotes and carriers with variants *in cis*. The actual number of empirically observed carriers for each category is displayed on top of the each bar.

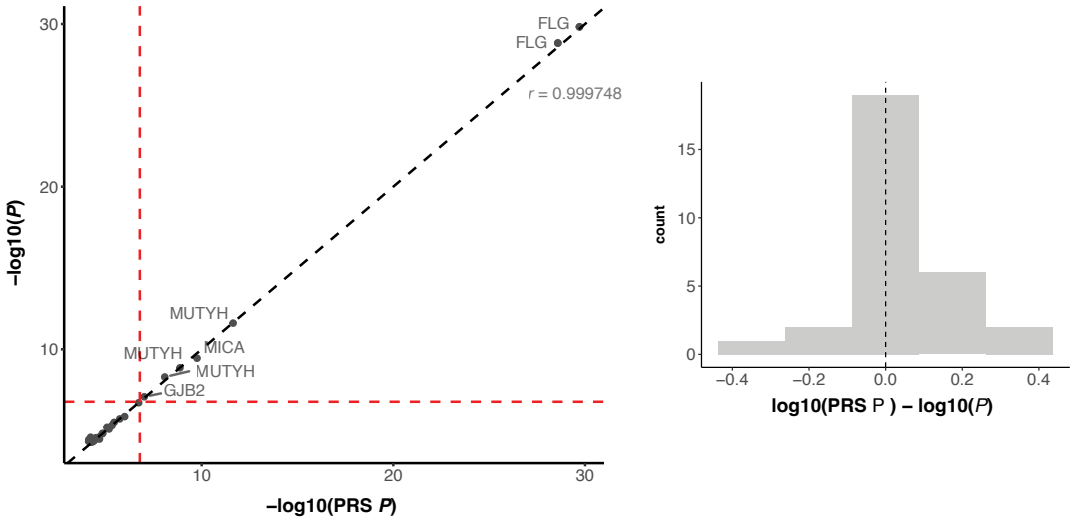


Figure A.18: Association P-values before and after inclusion of PRS as a covariate. The scatter plot depicts the association P-values both before and after PRS was included as a covariate. The y-axis represents the P-value prior to PRS adjustment, while the x-axis demonstrates the P-value afterPRS adjustment. On the right, the difference in log-transformed P-values before and after PRS adjustment is displayed. The plot highlights gene-trait associations that were considered Bonferroni significant in the recessive analysis.

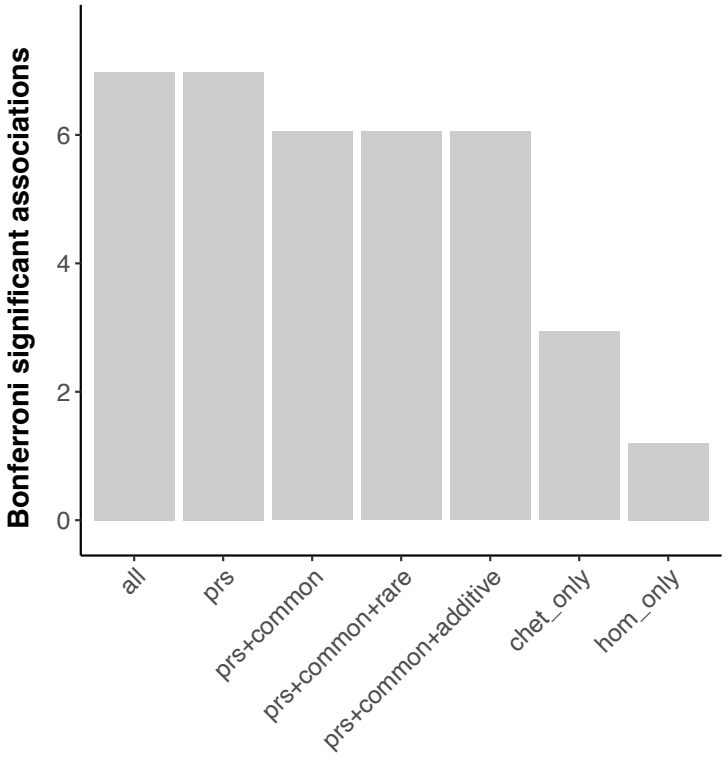


Figure A.19: Attrition of Bonferroni significant associations after successive conditioning steps and filters. Bar chart showing remaining Bonferroni significant associations through a series of steps. Starting with all initial significant associations, it shows the impact of conditioning on off-chromosome PRS, then accounting for nearby common variants (eliminating two gene-trait pairs). Subsequent bars show the effects of conditioning on rare variants in the gene and applying an additive model of affected haplotypes, neither of which reduces associations. The final bars separate associations remaining after filtering to compound heterozygous or homozygous variants, respectively. Each bar represents the number of associations remaining after its corresponding step or filter.

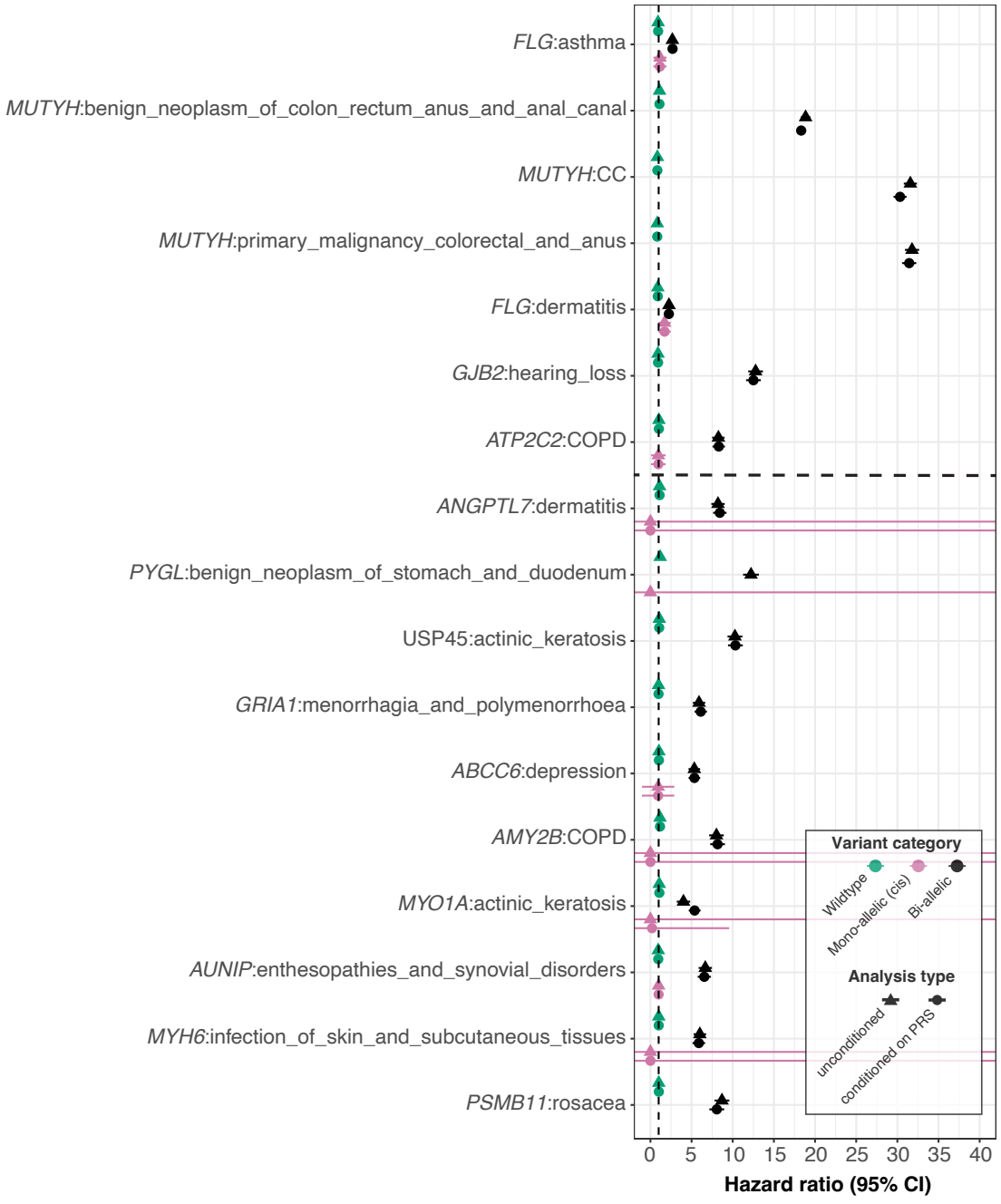


Figure A.20: Cox proportional hazards modeling with and without polygenic effects. HRs comparing CH and homozygous status versus heterozygous carrier status. Hazard ratios shown with (circles) and without (triangles) accounting for polygenic contribution by conditioning on off-chromosome PRSs for heritable traits passing quality control. Pink symbols indicate gene-traits with one or more individuals having multiple *cis* variants on the same haplotype. Dashed lines demarcate Bonferroni significance ($P < 1.89 \times 10^{-7}$, top) and FDRs < 0.1 (bottom) cutoffs. Abbreviations: CC (colorectal cancer), COPD (chronic obstructive pulmonary disease).

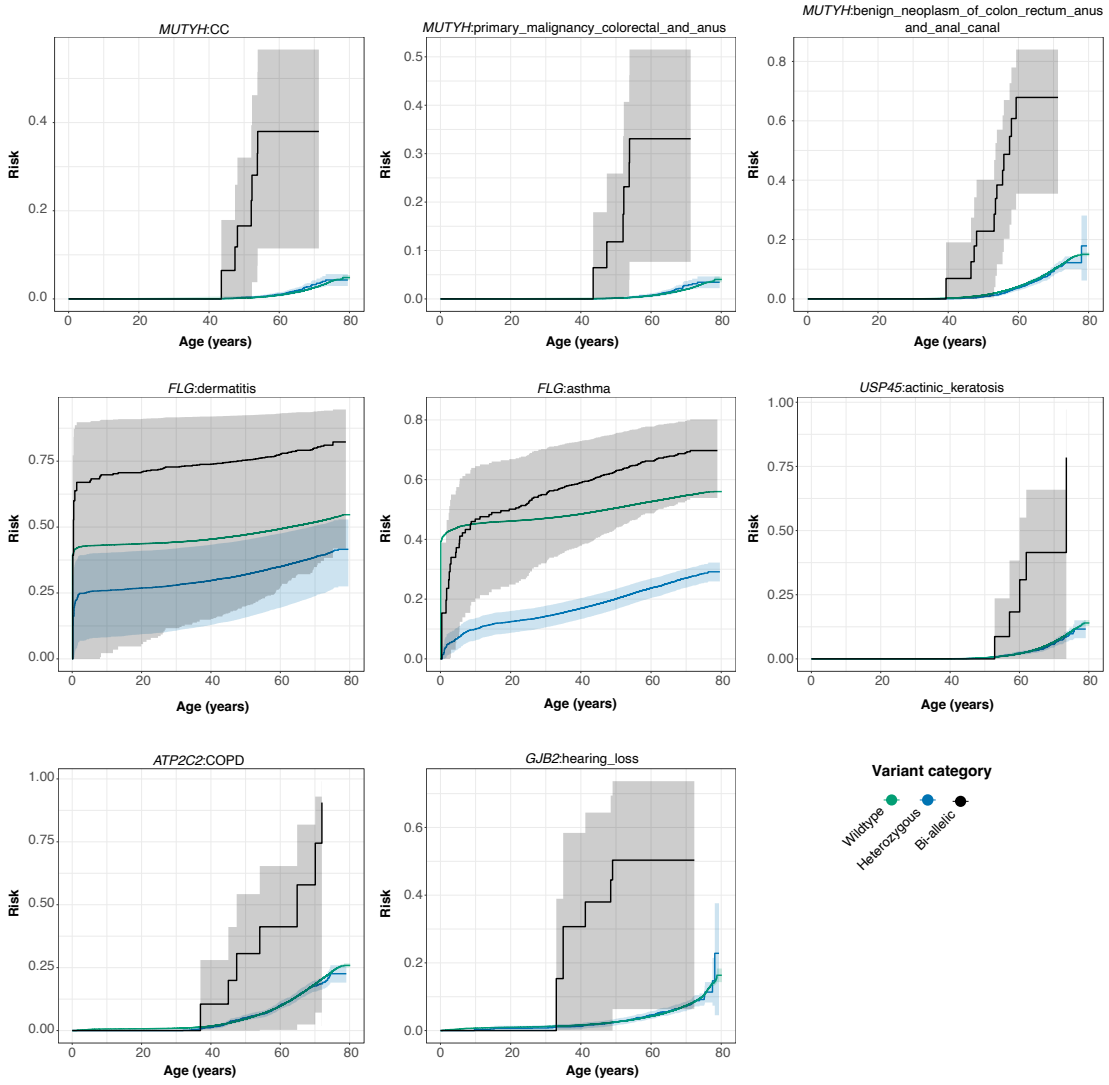


Figure A.21: Kaplan-Meier survival curves for carriers of biallelic variants. Trajectories for wildtypes (green) and biallelic (CH or homozygous) carriers (black) of damaging missense/protein-altering mutations. For traits with >50% left-censored cases, confidence interval estimates are unreliable and omitted. Thus, wildtype confidence intervals for *FLG*-Asthma and *FLG*-Dermatitis are not displayed.

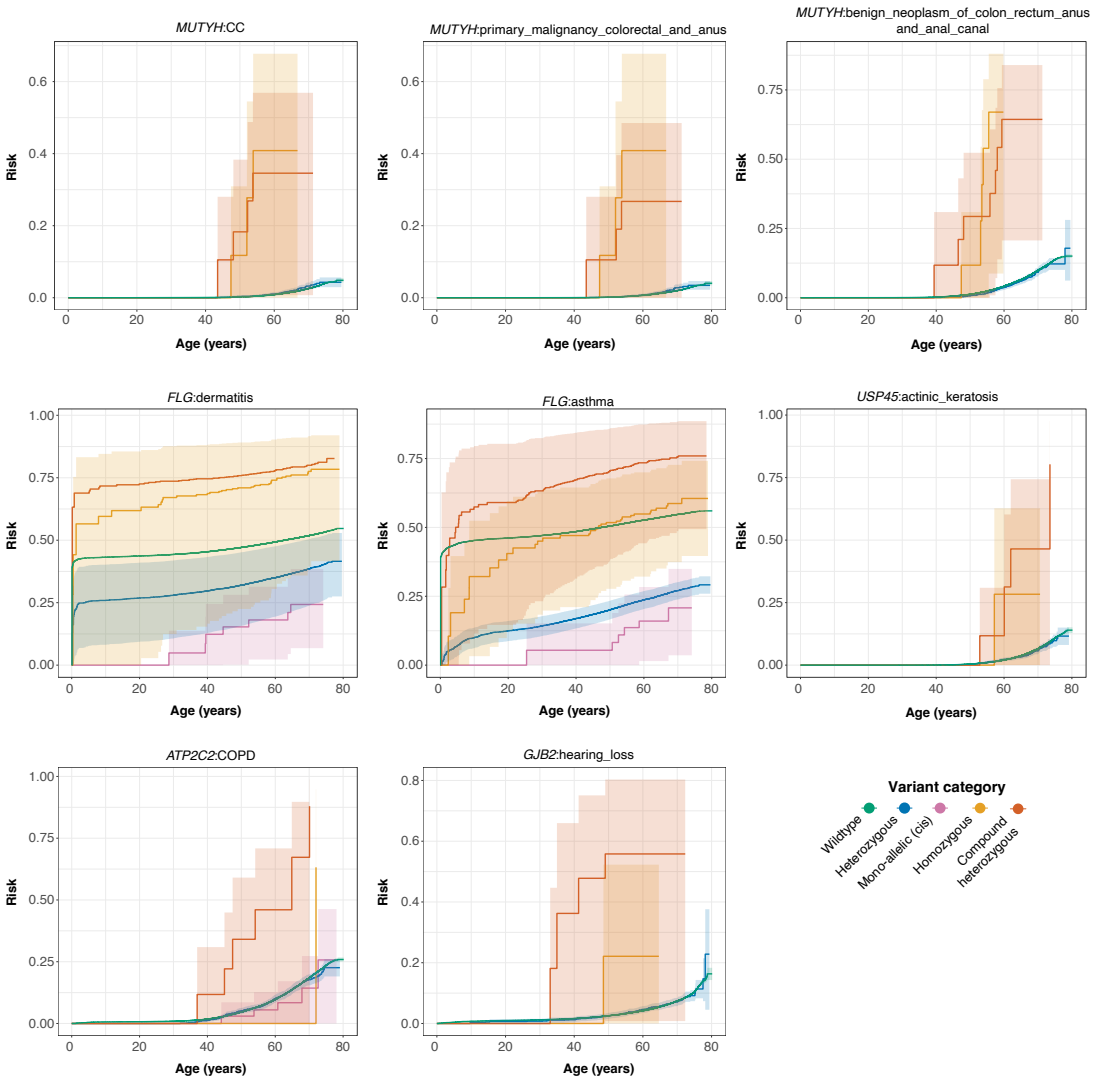


Figure A.22: Kaplan-Meier survival curves for carriers of CH, homozygous, and heterozygous variants. CH (red), homozygous (orange), heterozygous (blue), and single haplotype disruption (pink) carriers of pLoF or damaging missense/protein-altering mutations are shown. Wildtypes are in green. For traits with >50% left-censored cases, confidence interval estimates are unreliable and omitted. Thus, wildtype confidence intervals for *FLG*-Asthma and both wildtype and CH confidence intervals for *FLG*-Dermatitis are not displayed.

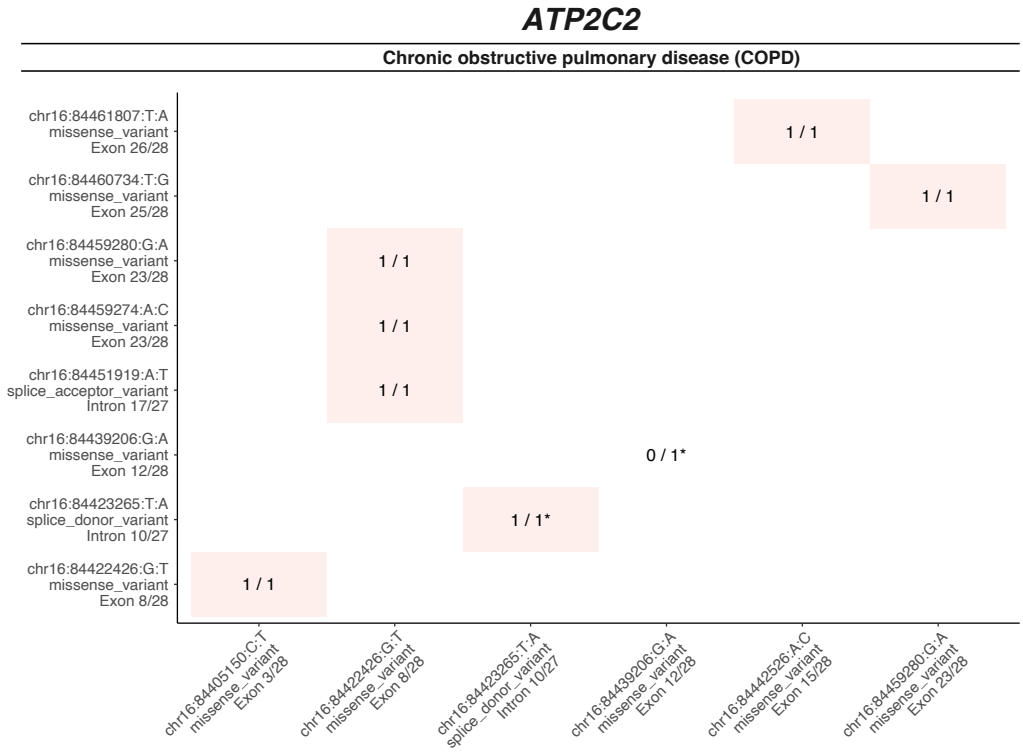


Figure A.23: Co-occurrence of deleterious *ATP2C2* variants by COPD status. Biallelic variant occurrence in *ATP2C2* for COPD cases. Each tile shows the number of COPD cases out of total biallelic carriers. Variants affecting both gene copies are displayed with their consequences and involved exon or intron. Stars (*) in labels indicate homozygosity.

B

Supplement to chapter 3

Full derivation of non-additive genotype transformation

The following derivation was completed by Dr Duncan Palmer. We consider the general case in which the only restriction is that the sum of proportions is 1:

$$\begin{bmatrix} r \\ h \\ a \end{bmatrix}, \text{ such that } r + h + a = 1. \tag{B.1}$$

We define:

$$X^{0\prime} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \; X^{A\prime} = \begin{bmatrix} 0 \\ 1 \\ 2 \end{bmatrix}, \; X^{D\prime} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}. \tag{B.2}$$

We define an inner product according to the general genotype proportions:

$$\langle X, Y \rangle = rX_0Y_0 + hX_1Y_1 + aX_2Y_2. \tag{B.3}$$

Under this encoding then, determine the additive component (equivalent to centering and scaling the additive encoding assuming the genotype proportions) using Gram-Schmidt:

$$X^0 = \frac{1}{\left\| X^{0\prime} \right\|} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \tag{B.4}$$

$$X^A = \frac{X^{A\prime} - \langle X^{A\prime}, X^0\rangle X^0}{\left\|X^{A\prime} - \langle X^{A\prime}, X^0\rangle X^0\right\|} = \frac{1}{\sqrt{-h^2 - 4ha + h - 4a^2 + 4a}} \left(\begin{bmatrix} 0 \\ 1 \\ 2 \end{bmatrix} - (h + 2a) \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \right) \tag{B.5}$$

$$= \frac{1}{\sqrt{4ar + h(a + r)}} \begin{bmatrix} -(h + 2a) \\ 1 - (h + 2a) \\ 2 - (h + 2a) \end{bmatrix} \tag{B.6}$$

$$= \frac{1}{\sqrt{a + r - (a - r)^2}} \begin{bmatrix} -(h + 2a) \\ 1 - (h + 2a) \\ 2 - (h + 2a) \end{bmatrix}. \tag{B.7}$$

Finally, determine the dominance encoding without assuming HWE by completing the basis:

$$X^D = \frac{X^{D\prime} - \langle X^{D\prime}, X^0\rangle X^0 - \langle X^{D\prime}, X^A\rangle X^A}{\left\|X^{D\prime} - \langle X^{D\prime}, X^0\rangle X^0 - \langle X^{D\prime}, X^A\rangle X^A\right\|} \tag{B.8}$$

$$= \frac{\frac{1}{(h^2+4ha-h+4a^2-4a)} \begin{bmatrix} 2ha \\ 4a(h + a - 1) \\ -2h(h + a - 1) \end{bmatrix}}{\left\|X^{D\prime} - \langle X^{D\prime}, X^0\rangle X^0 - \langle X^{D\prime}, X^A\rangle X^A\right\|} \tag{B.9}$$

$$= \frac{\frac{1}{(h^2+4ha-h+4a^2-4a)} \begin{bmatrix} 2ha \\ -4ar \\ 2hr \end{bmatrix}}{2\sqrt{\frac{har}{-(h^2+4ha-h+4a^2-4a)}}} \tag{B.10}$$

$$= \frac{1}{\sqrt{-har(h^2 + 4ha - h + 4a^2 - 4a)}} \begin{bmatrix} -ha \\ 2ar \\ -hr \end{bmatrix} \tag{B.11}$$

$$= \frac{1}{\sqrt{har(4ar + h(a + r))}} \begin{bmatrix} -ha \\ 2ar \\ -hr \end{bmatrix} \tag{B.12}$$

$$= \frac{1}{\sqrt{har(a + r - (a - r)^2)}} \begin{bmatrix} -ha \\ 2ar \\ -hr \end{bmatrix}. \tag{B.13}$$

Confirmation of HWE encoding recovery

To confirm that substitution recovers the HWE encoding, let's substitute the HWE proportions into the general case equations and see if we get the expected results.

Under HWE:

$$\begin{aligned} r &= p^2 \text{ (homozygous reference)} \\ h &= 2pq = 2p(1-p) \text{ (heterozygous)} \\ a &= q^2 = (1-p)^2 \text{ (homozygous alternate)}, \end{aligned}$$

where p is the frequency of the reference allele, and $q = 1 - p$ is the frequency of the alternate allele. We have X^D as defined previously:

$$X^D = \frac{1}{\sqrt{har(a + r - (a - r)^2)}} \begin{bmatrix} -ha \\ 2ar \\ -hr \end{bmatrix}. \tag{B.14}$$

I substitute the HWE proportions:

$$\begin{aligned} X^D &= \frac{1}{\sqrt{(2pq)(q^2)(p^2)(2pq)}} \begin{bmatrix} -2pq \cdot q^2 \\ 2p^2 \cdot q^2 \\ -2pq \cdot p^2 \end{bmatrix} \\ &= \frac{1}{2p^2q^2} \begin{bmatrix} -2pq^3 \\ 2p^2q^2 \\ -2p^3q \end{bmatrix} \\ &= \begin{bmatrix} -q^2/p \\ 1 \\ -p^2/q \end{bmatrix} \end{aligned}$$

I simplify this further by substituting $q = 1 - p$:

$$\begin{aligned} X^D &= \begin{bmatrix} -(1-p)^2/p \\ 1 \\ -p^2/(1-p) \end{bmatrix} \\ &= \begin{bmatrix} -p/(1-p) \\ 1 \\ -(1-p)/p \end{bmatrix}. \end{aligned}$$

This matches the expected HWE encoding for the dominance component: $[-p/(1-p), 1, -(1-p)/p]$ reported elsewhere and we are done. Thus, we can confirm that substituting the HWE proportions into the general case equations does indeed recover the expected HWE encoding for the dominance component.

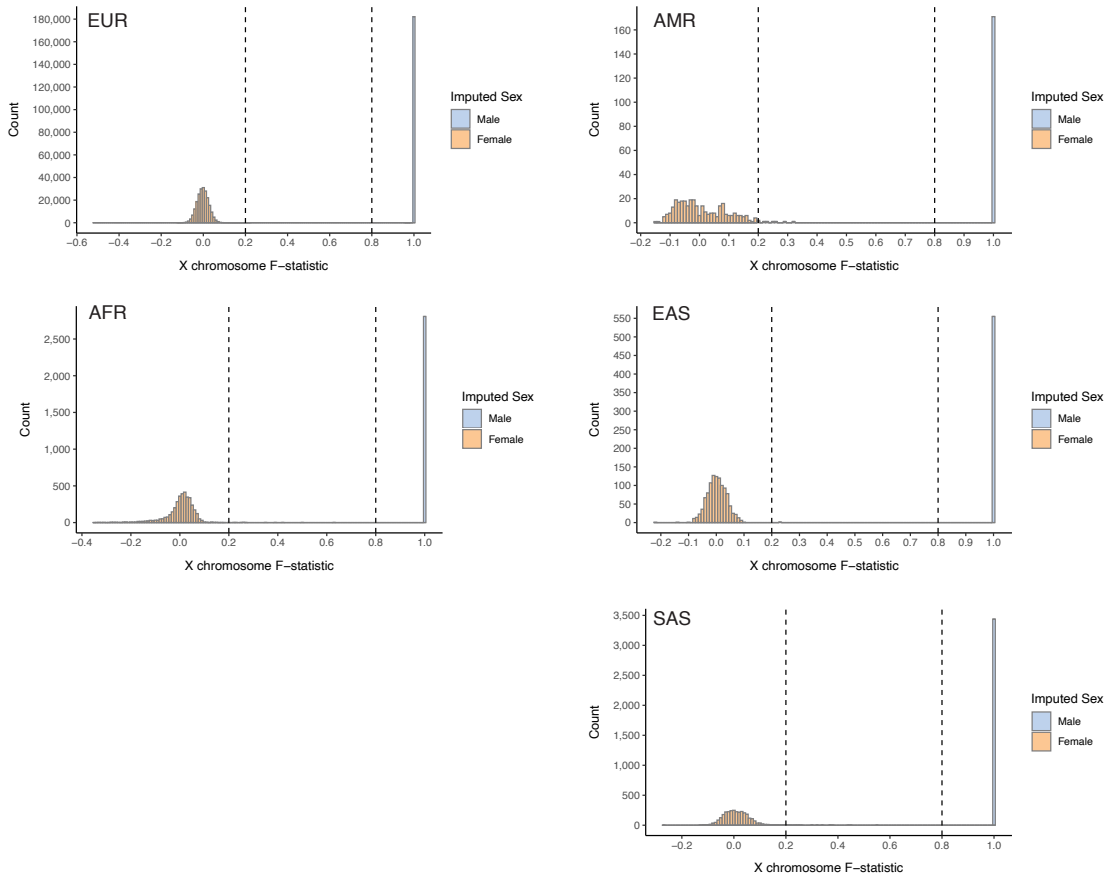


Figure B.1: Overview of sex imputation for 450k exome sequencing data. Sex imputation in UK Biobank participants using X chromosome variants. The F-statistic, derived from a pruned set of X chromosome variants, is used to impute sex across different ancestries. The dashed line (F-statistic = 0.6) represents the threshold for sex classification: values above indicate females, below indicate males.

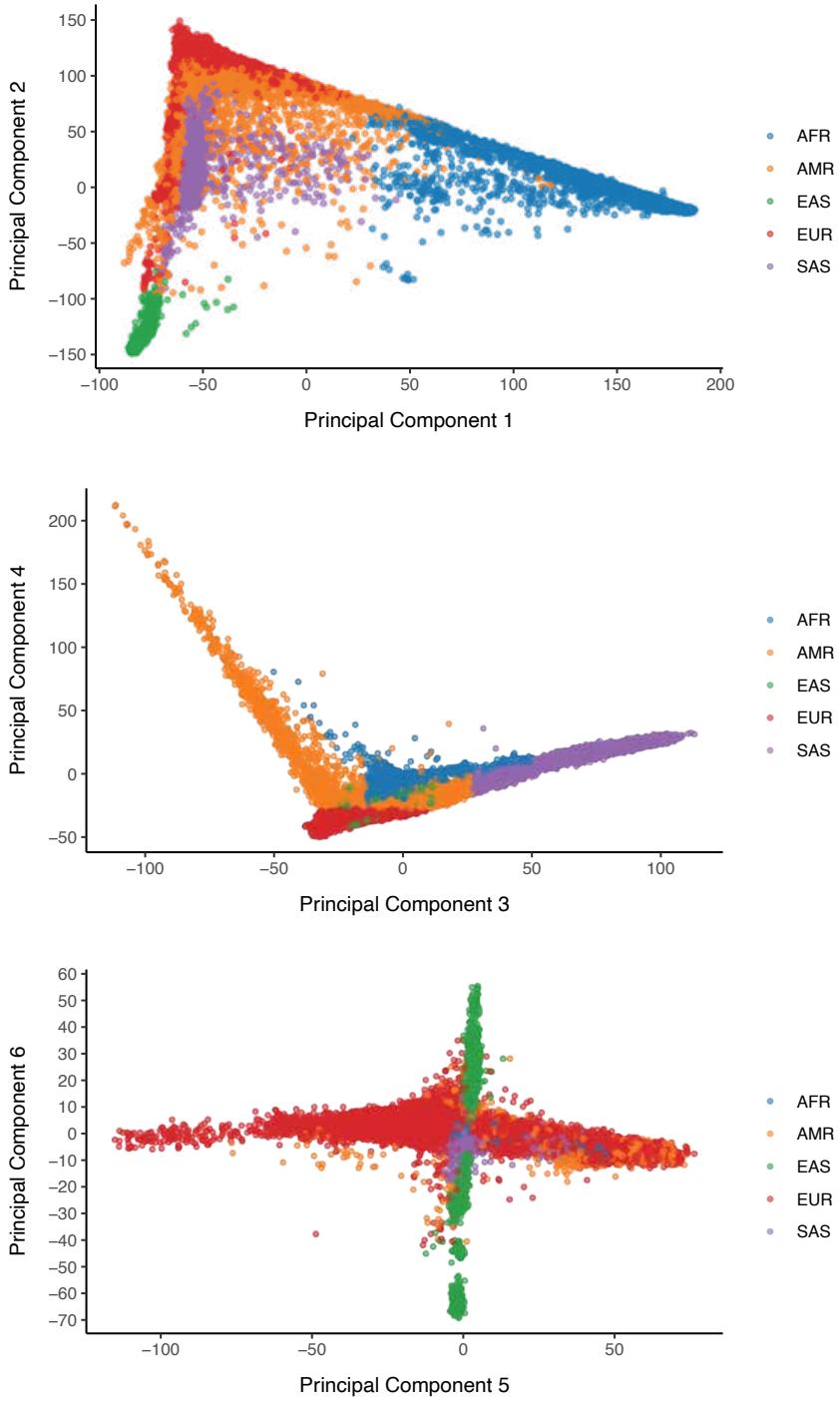


Figure B.2: Principal component analysis with 1000 genomes project for 450k exome sequencing data. Principal component analysis of genetic variation using 1000 Genomes data. The first six principal components are shown, derived from a pruned set of variants. Colors represent predicted ancestries assigned by a random forest (RF) classifier.

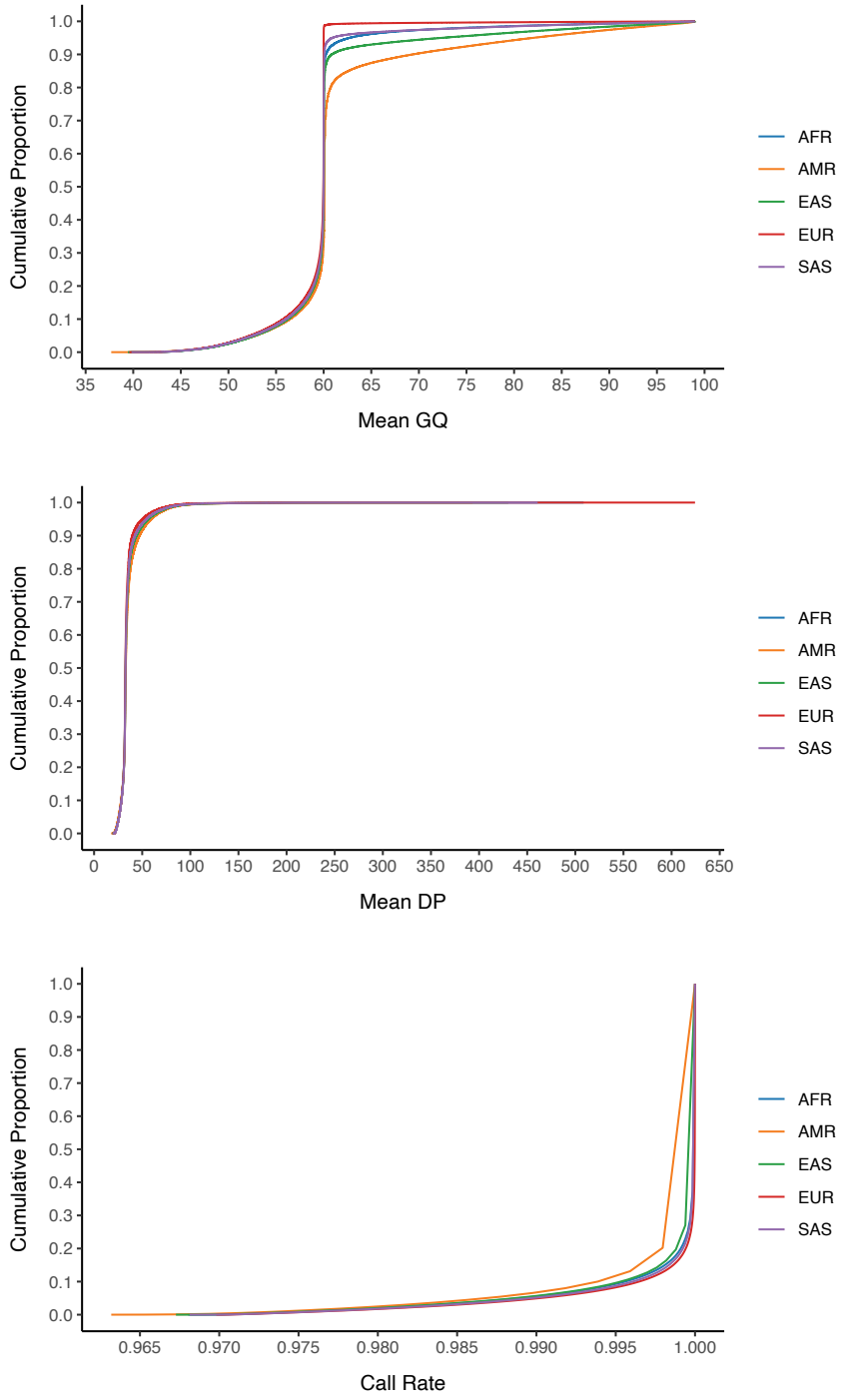


Figure B.3: Mean genotype quality, allelic depth and call rate for 450k exome sequencing data. Cumulative distributions of key sequencing quality metrics across major ancestry groups. From top to bottom, the panels show: mean genotype quality (GQ), mean sequencing depth (DP), and call rate. Each colored line represents a distinct major ancestry group.

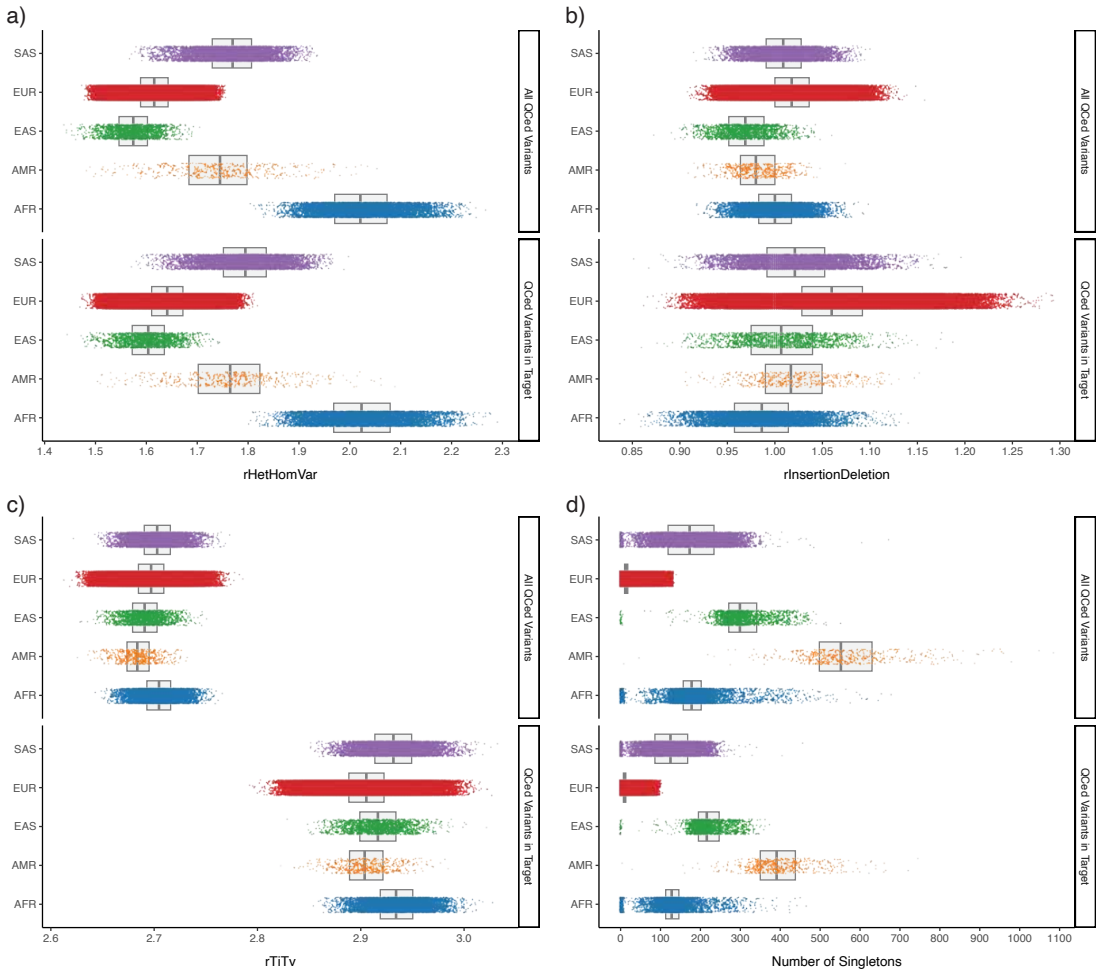


Figure B.4: Variant metrics before and after sample removals for 450k exome sequencing data. Distribution of key genomic metrics before and after sample removal. Each panel illustrates: a) ratio of heterozygous to homozygous variants ($r_{\text{HetHomVar}}$); b) ratio of insertions to deletions ($r_{\text{InsertionDeletion}}$); c) ratio of transitions to transversions (r_{TiTv}); and d) number of singletons. Upper and lower parts of each panel represent distributions before and after sample removal, respectively.

Table B.1: Trio-switch error rates across for 450k exome sequencing following phasing. Trio-switch error rates across final reconstituted chromosomes for variants included in the 450k ES release across MAC bins. 'Switches' is the number of errors observed, and 'Tested' is the total number of sites assessed. We include the switch error rates before (PP$\geq$ 0.50) and after (PP$\geq$ 0.90) restricting to variants that are phased with high-confidence. All confidence intervals are 95% binomial confidence intervals.

MAC Bin	PP>0.50			PP>0.90		
	Switches	Tested	SER estimate	Switches	Tested	SER estimate
singleton	400	1259	0.3177 (0.2926 - 0.344)	25	607	0.0412 (0.0281 - 0.0601)
2-5	205	2814	0.0729 (0.0638 - 0.083)	44	2699	0.0163 (0.0122 - 0.0218)
6-10	95	2399	0.0396 (0.0325 - 0.0482)	25	2475	0.0101 (0.0069 - 0.0149)
11-20	94	3405	0.0276 (0.0226 - 0.0337)	30	3634	0.0083 (0.0058 - 0.0118)
21-50	95	5889	0.0161 (0.0132 - 0.0197)	46	6579	0.007 (0.0052 - 0.0093)
51-100	66	5199	0.0127 (0.01 - 0.0161)	38	5981	0.0064 (0.0046 - 0.0087)
101-200	55	5549	0.0099 (0.0076 - 0.0129)	34	6294	0.0054 (0.0039 - 0.0075)
201-500	75	7764	0.0097 (0.0077 - 0.0121)	43	8426	0.0051 (0.0038 - 0.0069)
501-1000	57	7493	0.0076 (0.0059 - 0.0098)	25	7155	0.0035 (0.0024 - 0.0052)
1001-2000	34	11069	0.0031 (0.0022 - 0.0043)	39	9844	0.004 (0.0029 - 0.0054)
2001-5000	69	29080	0.0024 (0.0019 - 0.003)	78	26105	0.003 (0.0024 - 0.0037)
5001-10000	170	87657	0.0019 (0.0017 - 0.0023)	203	79855	0.0025 (0.0022 - 0.0029)
10001-Inf	16706	11278675	0.0015 (0.0015 - 0.0015)	17550	9191016	0.0019 (0.0019 - 0.0019)

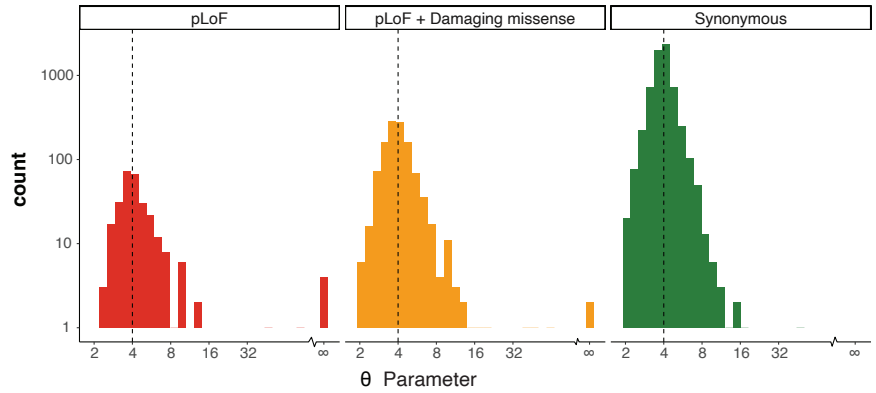


Figure B.5: Distribution of θ parameter for genes with MAF $>$ 0.005. θ signifies the degree of disequilibrium. $\theta = 4$ indicates Hardy-Weinberg Equilibrium (HWE), marked by a dashed line. $\theta > 4$ shows heterozygous excess, while $\theta < 4$ indicates homozygous excess compared to HWE expectations. Some genes in the pLoF and damaging missense/protein altering category have $\theta = \infty$, indicating no observed homozygous carriers. The x-axis has been truncated at $\theta < 2$ to aid visualization.

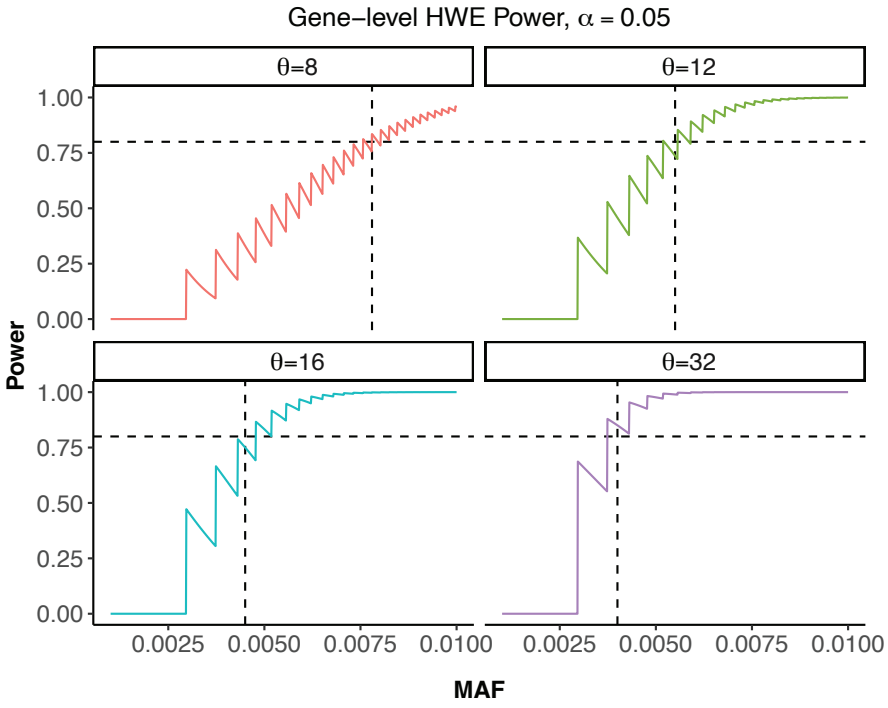


Figure B.6: Power analysis for gene-level Hardy-Weinberg Equilibrium across different disequilibrium coefficients. $\theta = 4$ indicates HWE, while increasing θ values represent greater homozygous depletion relative to HWE expectations. Horizontal dashed lines indicate 80% power. Vertical dashed lines correspond to Minor Allele Frequencies (MAFs) of 0.0078, 0.0055, 0.0045, and 0.004, representing the approximate MAFs required to achieve 80% power at $\alpha = 0.05$ for each θ value.

Table B.2: Overview of the UKBB field entry for 114 mostly uncorrelated traits. Overview of 114 mostly uncorrelated ($R^2 < 0.8$) quantitative traits and their corresponding category and UKBB entry field codes. Description on how 'other' traits were defined, can be found in methods.

Category	Category Field	Entry Field and Name of trait
nmr_metabolomics	220	p23428.conc_of_vldl_particles , p23432.average_diameter_for_ldl_particles , p23433.average_diameter_for_hdl_particles , p23407.total_triglycerides , p23420.free_cholesterol_in_vldl , p23473.citrate , p23474.3_hydroxybutyrate , p23475.acetate , p23476.acetoacetate , p23477.acetone , p23478.creatinine_from_ , p23479.albumin_from_ , p23480.glycoprotein_acetyls , p23441.apolipoprotein_b_to_apolipoprotein_a1_ratio , p23443.degree_of_unsaturation , p23459.omega_6_fatty_acids_to_omega_3_fatty_acids_ratio , p23460.alanine , p23461.glutamine , p23462.glycine , p23463.histidine , p23466.leucine , p23467.valine , p23468.phenylalanine , p23469.tyrosine , p23470.glucose_from_ , p23471.lactate , p23472.pyruvate , p23516.conc_of_very_small_vldl_particles , p23517.total_lipids_in_very_small_vldl , p23519.cholesterol_in_very_small_vldl , p23521.free_cholesterol_in_very_small_vldl , p23522.triglycerides_in_very_small_vldl , p23492.cholesteryl_esters_in_very_large_vldl , p23450.docosahexaenoic_acid , p23458.polyunsaturated_fatty_acids_to_monounsaturated_fatty_acids_ratio, p23499.cholesteryl_esters_in_large_vldl , p23503.total_lipids_in_medium_vldl , p23509.conc_of_small_vldl_particles , p23510.total_lipids_in_small_vldl , p23553.phospholipids_in_very_large_hdl , p23554.cholesterol_in_very_large_hdl , p23555.cholesteryl_esters_in_very_large_hdl , p23564.triglycerides_in_large_hdl , p23529.triglycerides_in_idl , p23536.triglycerides_in_large_ldl , p23551.conc_of_very_large_hdl_particles , p23552.total_lipids_in_very_large_hdl , p23571.triglycerides_in_medium_hdl , p23572.conc_of_small_hdl_particles , p23573.total_lipids_in_small_hdl , p23576.cholesteryl_esters_in_small_hdl , p23577.free_cholesterol_in_small_hdl
blood_biochemistry	17518	p30830.shbg , p30840.total_bilirubin , p30850.testosterone , p30680.calcium , p30600.albumin_from_ , p30610.alkaline_phosphatase , p30620.alanine_aminotransferase , p30860.total_protein , p30650.aspartate_aminotransferase, p30810.phosphate , p30690.cholesterol , p30700.creatinine_from_ , p30710.c_reactive_protein , p30870.triglycerides , p30880.urate , p30890.vitamin_d , p30750.glycated_haemoglobin_hba1c, p30760.hdl_cholesterol_from_ , p30770.igf_1 , p30670.urea , p30790.lipoprotein_a , p30720.cystatin_c , p30730.gamma_glutamyltransferase , p30740.glucose_from_
body_composition_by_impedance	100009	p23107.impedance_of_leg_right, p23110.impedance_of_arm_left , p23128.trunk_fat_mass , p23112.leg_fat_mass_right , p23102.whole_body_water_mass
body_size_measures	100010	p21001.body_mass_index_bmi_from_, p48.waist_circumference , p49.hip_circumference , p20015.sitting_height , p21002.weight_from_ , p51.seated_height
hand_grip_strength	100019	p47.hand_grip_strength_right, p46.hand_grip_strength_left
spirometry	100020	p20150.forced_expiratory_volume_in_1_second_fev1_best_measure, p20153.forced_expiratory_volume_in_1_second_fev1_predicted
blood_count	100081	p30150.eosinophill_count , p30120.lymphocyte_count , p30130.monocyte_count , p30140.neutrophill_count , p30260.mean_reticulocyte_volume , p30160.basophill_count , p30250.reticulocyte_count , p30060.mean_corpuscular_haemoglobin_conc , p30270.mean_sphered_cell_volume , p30280.immature_reticulocyte_fraction , p30300.high_light_scatter_reticulocyte_count, p30000.white_blood_cell_leukocyte_count , p30080.platelet_count , p30100.mean_platelet_thrombocyte_volume , p30040.mean_corpuscular_volume , p30050.mean_corpuscular_haemoglobin , p30010.red_blood_cell_erythrocyte_count , p30020.haemoglobin_conc
urine_assays	100083	p30510.creatinine , p30520.potassium, p30530.sodium
Other	Other	BMI_adj_WHR, WHR

Table B.3: Number of mono- and biallelic variants identified across annotation masks in 450k exome sequencing dataset in Europeans. Overview of the number of variants that are present in the 450k ES release for individuals of European genetic ancestry before and after filtering by PP. Note, that CHs indicate Compound heterozygous occurrences. Additionally, if two heterozygous variants have been counted as a CH or *cis*, they will not be double counted in the 'heterozygous column'.

Annotation	PP≥0.50				PP≥0.90			
	CHs	Homs.	Cis	Hets.	CHs	Homs.	Cis	Hets.
pLoF	17,718	66,882	184,724	8,536,164	16,058	65,276	174,040	7,961,600
damaging missense	35,810	114,397	174,790	18,738,594	32,746	110,498	165,836	17,775,276
pLoF + damaging_missense	79,469	181,138	440,905	27,015,997	72,590	175,631	417,504	25,488,652
other missense	4,349,683	2,541,701	8,122,758	216,964,644	4,177,798	2,476,592	7,942,149	212,148,726
synonymous	2,018,800	1,680,667	5,281,930	160,548,614	1,954,424	1,638,110	5,217,337	157,994,456
nonsynonymous	5,174,892	2,715,174	9,583,760	240,002,423	4,944,143	2,644,720	9,323,478	233,899,738

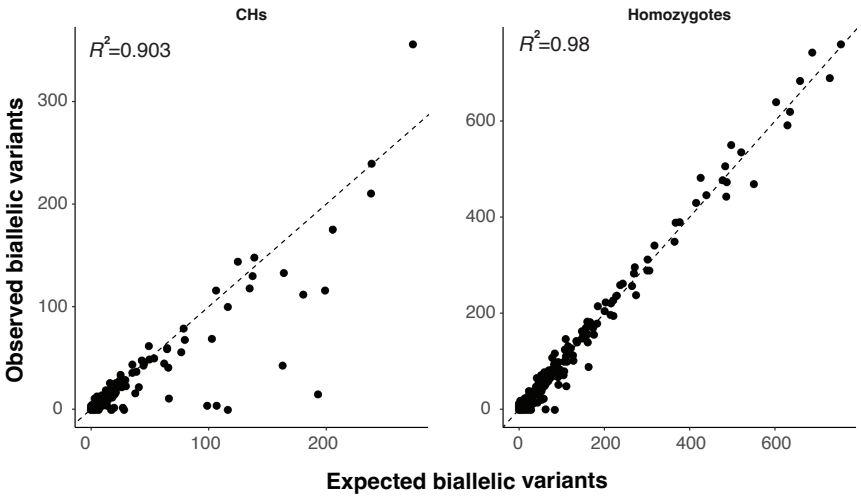


Figure B.7: Comparison between simulated and empirically observed counts of pLoF carriers. Observed versus expected CH (left) and homozygous (right) pLoF events in 339,041 unrelated Europeans from UKBB. Each dot represents a gene. X-axis shows expected events; Y-axis shows observed events. Coefficient of determination (R^2) is displayed in the top left corner of each plot.

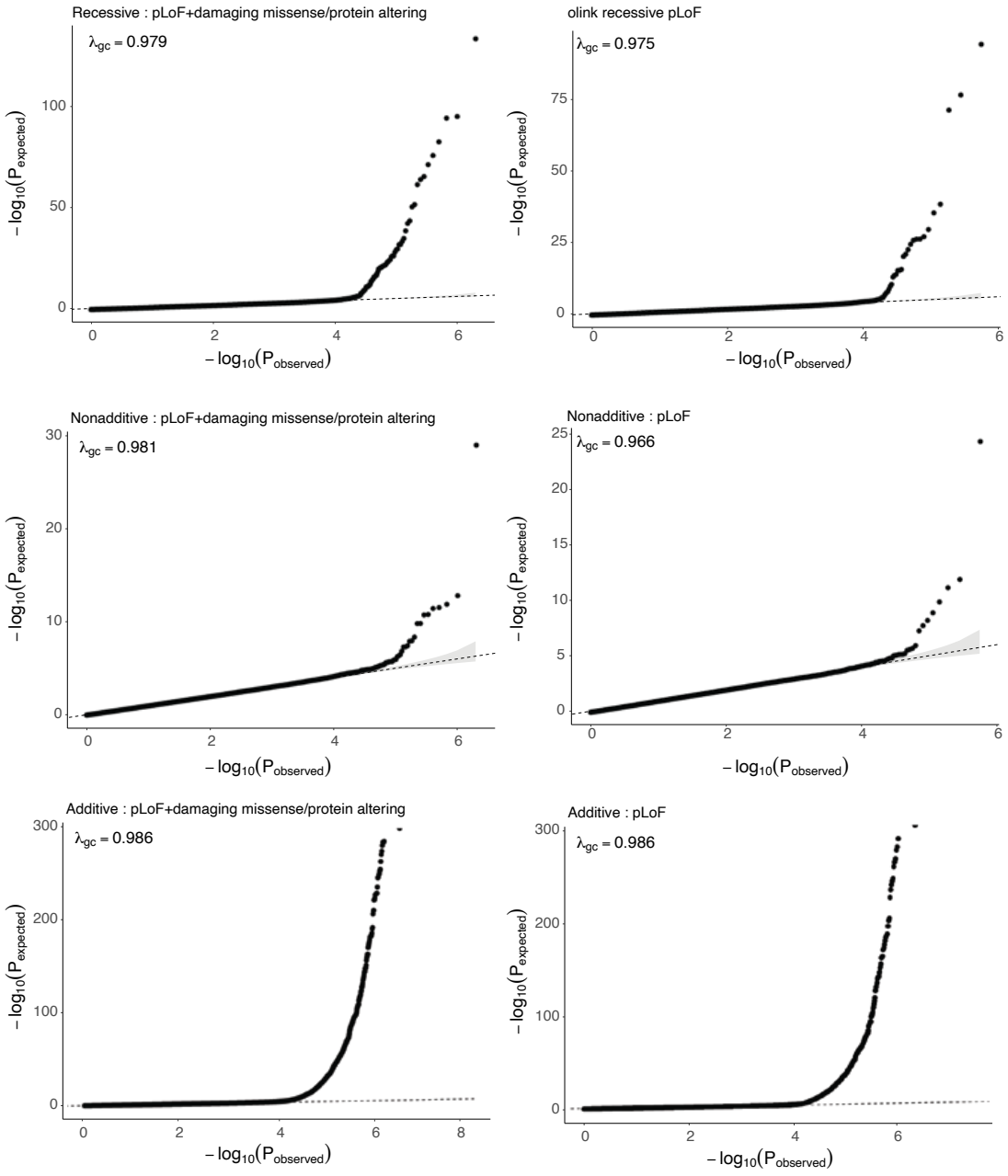


Figure B.8: QQ-plot for additive, recessive and nonadditive analysis of Olink data. QQ-plots for Olink data using recessive, additive, and nonadditive genotype encodings. Analysis covers pLoF+damaging missense/protein altering (first column) and pLoF (second column) annotation masks. The corresponding λ_{GC} is shown in the top left corner for each panel. X-axis shows expected P-values (assuming uniform distribution under null hypothesis); Y-axis shows observed P-values.

Table B.4: Overview of 26 gene-assay associations discovered using recessive or nonadditive models with pLoF annotation mask. This table shows all recessive or nonadditive associations that passed the significance threshold ($P < 8.33 \times 10^{-10}$). 'Affected gene' shows genes with at least five biallelic pLoF carriers, 'Olink assay' indicates measured proteins, 'Samples' gives total non-missing values, 'Haplotypes' shows mutated haplotypes (equivalent to MAF), and 'Biallelics' denotes biallelic carrier count. Discovery used the recessive $[0, 0, 1]$ model, with additional analyses using nonadditive recoding and additive encoding for significant associations. Rows are ordered by nonadditive P-values.

Affected gene	Olink assay	Samples	Haplotypes	Biallelics	P-value		
					Additive	Recessive	Nonadditive
HMCN2	HMCN2	34665	1884	420	4.01E-21	2.83E-01	4.63E-25
MMP10	MMP10	41210	2272	41	0.00E+00	3.80E-72	1.16E-12
VIT	VIT	35540	951	10	0.00E+00	2.50E-36	6.33E-12
GIMAP8	GIMAP7	34731	1103	9	5.42E-160	3.25E-27	1.21E-10
CWH43	SBSN	35459	2076	29	1.54E-20	2.64E-16	1.15E-09
RBKS	RBKS	40334	739	7	0.00E+00	1.45E-30	1.63E-08
CCL26	CCL26	41229	759	6	1.90E-175	2.89E-16	4.80E-08
FCN3	MASP1	40927	1376	17	0.00E+00	2.18E-25	7.17E-06
MMP8	MMP8	41588	2224	28	0.00E+00	2.06E-77	6.83E-05
ASAH2	ASAH2	40988	698	6	1.94E-288	6.55E-15	2.29E-04
LAIR2	LAIR2	40988	3608	87	0.00E+00	4.54E-95	1.76E-03
ENTPD2	ENTPD2	40307	657	5	0.00E+00	1.97E-11	1.77E-03
MAMDC4	MAMDC4	35357	1629	28	5.15E-272	3.70E-21	2.00E-03
MSR1	LGALS3BP	35579	815	6	0.00E+00	9.83E-11	3.54E-03
SIGLEC1	SIGLEC1	41486	1088	7	0.00E+00	1.60E-23	1.25E-02
OLFM4	OLFM4	33435	976	9	1.75E-168	1.30E-14	2.18E-02
CD33	CD33	41427	2661	44	0.00E+00	2.38E-39	2.44E-02
GIMAP8	GIMAP8	34228	1088	7	8.34E-193	5.60E-14	4.37E-02
LAIR2	LILRA2	41308	3648	89	3.78E-210	3.50E-27	1.19E-01
MMP1	MMP1	39830	1479	8	0.00E+00	8.17E-27	1.34E-01
MSR1	C1QTNF5	34731	789	6	0.00E+00	1.22E-16	1.74E-01
MSR1	MSR1	41660	942	7	0.00E+00	4.90E-28	3.45E-01
PSCA	PSCA	35369	3120	67	1.55E-207	6.27E-22	8.19E-01

Table B.5: 68 gene-assay associations discovered using a standard recessive model with pLoF and damaging missense/protein altering mask. Summary of 68 gene-assay associations identified using a standard recessive model. Rows are ordered by nonadditive *P*-values. Most gene-assay combinations are significant under an additive model. Those significant under a nonadditive model are often also significant assuming additive genetic architecture.

Affected gene	Olink assay	Samples	Haplotypes	Biallelics	P-value		
					Additive	Recessive	Nonadditive
HMCN2	HMCN2	34665	2360	429	5.30E-35	1.96E-01	1.00E-29
FLG	GPR15L	35369	4560	148	1.49E-53	2.91E-34	1.47E-13
DXO	DXO	35450	2385	56	0.00E+00	7.87E-65	1.24E-12
GIMAP8	GIMAP7	34731	1117	10	2.03E-161	5.05E-29	2.79E-12
MMP10	MMP10	41210	2278	41	0.00E+00	3.80E-72	3.52E-12
FLG	PYDC1	34591	4471	145	1.35E-149	1.82E-52	1.56E-11
FLG	KLK8	39943	5133	165	1.79E-104	3.43E-43	1.75E-11
CWH43	SBSN	35459	2142	31	3.06E-21	1.20E-17	1.44E-10
APOBEC1	TMPRSS15	41118	833	6	1.54E-03	2.38E-11	1.47E-10
ENTPD2	GFAP	40243	3325	75	6.59E-03	2.78E-10	4.16E-09
FLG	KLK7	33831	4392	145	4.78E-58	3.58E-27	1.17E-08
VIT	VIT	35540	1213	15	0.00E+00	1.58E-39	3.59E-08
RBKS	RBKS	40334	763	7	0.00E+00	1.45E-30	4.67E-08
CCL26	CCL26	41229	759	6	1.90E-175	2.89E-16	4.80E-08
FLG	LY6D	41095	5288	171	4.18E-107	8.69E-36	1.36E-07
FLG	IL36G	34352	4449	147	7.64E-50	8.83E-23	3.28E-07
FLG	SLURP1	35262	4562	147	3.79E-55	4.54E-24	4.24E-07
FLG	PSAPL1	35369	4560	148	2.49E-267	2.57E-62	5.48E-07
FLG	WFDC12	41564	5345	175	0.00E+00	3.46E-134	1.12E-06
KLK14	KLK14	41639	1612	20	3.97E-241	5.52E-25	3.03E-06
FCN3	MASP1	40927	1386	17	0.00E+00	2.18E-25	7.58E-06
ASAH2	ASAH2	40988	740	8	7.95E-286	5.51E-17	1.63E-05
FLG	CEACAM19	35369	4560	148	4.57E-45	1.35E-18	2.24E-05
FLG	CDHR1	40347	5185	169	2.93E-30	4.92E-15	2.89E-05
ABCA6	DCBLD2	41427	1980	21	8.11E-30	1.38E-10	2.93E-05
SIGLEC1	CD63	40918	1501	10	2.02E-50	6.49E-13	3.39E-05
FLG	DSG3	41505	5332	175	1.28E-57	1.05E-20	3.47E-05
MMP8	MMP8	41588	2244	29	0.00E+00	1.19E-76	7.95E-05
NAIP	IL18	41146	572	25	3.41E-37	1.82E-15	1.39E-04
DXO	C2	41400	2781	61	7.67E-199	1.90E-21	9.43E-04
NUMA1	FOLR3	41297	2245	28	0.00E+00	2.25E-24	9.49E-04
DXO	CFB	34974	2342	55	7.65E-229	3.14E-22	1.15E-03
SERPINA1	CTRL	35450	3515	91	5.16E-42	5.55E-12	1.38E-03
ADIPOQ	ADIPOQ	35579	972	11	9.62E-165	3.91E-12	1.43E-03
FLG	LGALS7_LGALS7B	41165	5290	174	6.95E-94	2.48E-23	1.60E-03
CTSH	CTSH	40412	1743	15	0.00E+00	7.49E-11	2.15E-03
MAMDC4	MAMDC4	35357	1685	29	1.12E-276	2.94E-21	2.47E-03
FGL1	FGL1	35586	1802	26	3.25E-88	4.06E-12	2.94E-03
LAIR2	LAIR2	40988	3632	87	0.00E+00	4.54E-95	3.09E-03
DXO	TNXB	41308	2786	62	5.10E-194	2.06E-19	3.68E-03
GIMAP8	GIMAP8	34228	1102	8	5.07E-198	9.69E-16	5.71E-03
FLG	FGFBP1	41427	5329	172	1.37E-36	6.60E-12	7.05E-03
ENTPD2	ENTPD2	40307	3329	75	0.00E+00	2.15E-83	9.04E-03
FLG	SBSN	35459	4578	149	4.48E-47	1.88E-13	9.53E-03
ABCA6	GAL	41486	1991	21	5.95E-256	1.72E-21	1.09E-02
SUMF2	SUMF2	41660	1905	25	0.00E+00	2.77E-66	1.51E-02
OLFM4	OLFM4	33435	979	9	1.85E-170	1.30E-14	2.26E-02
FLG	ENDOU	32811	4236	139	3.56E-53	6.60E-13	3.67E-02
KIRREL2	KIRREL2	40347	3043	25	1.52E-247	1.27E-17	3.85E-02
CRISP2	CRISP2	41639	1628	14	0.00E+00	2.46E-51	4.66E-02
AZU1	AZU1	40108	2168	26	1.38E-274	1.64E-22	5.58E-02
NHLRC3	NHLRC3	33883	1435	13	3.87E-271	4.22E-17	5.87E-02
LAIR2	LILRA2	41308	3672	89	2.49E-207	3.50E-27	1.23E-01
CFHR5	CFHR5	35111	720	5	0.00E+00	6.92E-17	1.39E-01
CD33	CD33	41427	3345	72	0.00E+00	9.08E-33	1.52E-01
VWA7	LILRB4	40545	7467	356	1.31E-33	3.90E-10	1.57E-01
CAPN2	DTX3	41118	1319	17	7.02E-189	3.32E-10	1.75E-01
SON	IFNGR2	40988	1555	13	0.00E+00	1.26E-12	1.96E-01
MMP1	MMP1	39830	1513	8	0.00E+00	8.17E-27	2.05E-01
MSR1	LGALS3BP	35579	971	10	0.00E+00	1.39E-14	2.89E-01
MSR1	MSR1	41660	1120	10	0.00E+00	2.46E-22	3.23E-01
FLG	CDSN	41620	5355	176	0.00E+00	1.80E-44	4.22E-01
STK19	LRPAP1	41660	1294	14	3.85E-270	7.11E-10	4.27E-01
ABCA6	MEGF10	41277	1984	21	4.89E-298	3.37E-12	5.30E-01
SIGLEC1	SIGLEC1	41486	1537	11	0.00E+00	9.94E-33	5.74E-01
SERPINA1	SERPINA1	34904	3448	87	2.28E-112	4.41E-12	6.32E-01
KLK15	KLK15	35294	1805	23	0.00E+00	3.59E-30	7.46E-01
MSR1	C1QTNF5	34731	943	10	0.00E+00	8.87E-21	8.31E-01
PSCA	PSCA	35369	3136	67	7.83E-209	6.27E-22	9.24E-01

Table B.6: Significant quantitative and binary gene-trait associations using pLoF and/or damaging missense/protein altering mask. Summary of significant (Experiment-wise $P < 3.72 \times 10^{-7}$) gene-trait associations assuming recessive ($[0, 0, 1]$) or nonadditive genetic architecture. For each association, the table shows the total number of affected haplotypes, the number of biallelic carriers of pLoF and damaging missense/protein altering variants, and effect sizes, standard errors, and P-values for additive, recessive, and nonadditive encodings.

Trait	Gene	Haplotypes	Biallelic	Additive analysis			Recessive analysis			Nonadditive analysis		
				BETA	SE	P	BETA	SE	P	BETA	SE	P
Other and unspecified dermatitis	*FLG*	50338	1604	0.385155	0.014118	7.25E-164	0.754701	0.029836	3.62E-141	-0.631153	0.032402	1.66E-84
Asthma	*FLG*	50338	1604	0.216775	0.012521	3.74E-67	0.584867	0.028381	2.35E-94	-0.50351	0.029878	1.01E-63
Atopic dermatitis	*FLG*	50338	1604	0.407928	0.025255	1.10E-58	0.687052	0.049887	3.75E-43	-0.587552	0.055691	5.07E-26
Other hearing loss	*GJB2*	7550	30	0.215901	0.052867	4.43E-05	2.03982	0.209999	2.64E-22	-2.02839	0.211839	1.02E-21
Malignant neoplasm of colon	*MUTYH*	7697	45	0.277552	0.080767	5.89E-04	1.84221	0.220233	6.02E-17	-1.85262	0.223468	1.13E-16
Vasomotor and allergic rhinitis	*FLG*	50338	1604	0.067406	0.014982	6.82E-06	0.289705	0.035639	4.34E-16	-0.266431	0.036663	3.67E-13
HDL cholesterol	*BTNL9*	9097	80	-0.0601	0.009526	2.81E-10	-0.399302	0.050527	2.73E-15	0.346816	0.050526	6.69E-12
Osteoarthritis of hip	*CHADL*	21622	314	0.083381	0.028408	3.33E-03	0.61355	0.086436	1.26E-12	-0.604206	0.088576	9.02E-12
Crohn's disease	*NOD2*	15365	145	0.681345	0.078842	5.53E-18	1.37944	0.194058	1.17E-12	-1.34251	0.205492	6.44E-11
Eosinophill count	*FLG*	48833	1565	0.00077	0.004605	8.67E-01	0.07386	0.012438	2.88E-09	-0.077812	0.012413	3.64E-10
Deficiency of other B vitamins	*CLYBL*	21523	314	0.277427	0.061187	5.79E-06	0.951754	0.143932	3.78E-11	-0.942355	0.150381	3.69E-10
FEV1	*FLG*	37852	1215	-0.017753	0.003842	3.82E-06	-0.075171	0.010378	4.38E-13	0.062428	0.010359	1.68E-09
Histidine	*HAL*	4162	22	0.573514	0.015434	3.12E-302	1.12998	0.105622	1.04E-26	-0.612065	0.105696	7.01E-09
Conductive+sensorineural hearing loss	*GJB2*	7550	30	0.160243	0.08149	4.93E-02	1.58119	0.271331	5.62E-09	-1.58599	0.274118	7.22E-09
Hand grip strength (right)	*FLG*	50140	1598	0.023755	0.00315	4.65E-14	-0.021943	0.008539	1.02E-02	0.046984	0.008519	3.49E-08
Hand grip strength (left)	*FLG*	50117	1597	0.027632	0.003142	1.43E-18	-0.01751	0.008518	3.98E-02	0.046136	0.008498	5.67E-08
FEV1	*FUT10*	2488	10	-0.04217	0.014504	3.64E-03	-0.66021	0.115248	1.01E-08	0.617314	0.115113	8.20E-08
Cholesterol	*TM6SF2*	11090	71	-0.139431	0.009219	1.13E-51	-0.425255	0.056991	8.53E-14	0.296965	0.057006	1.89E-07
Sitting height	*ABHD15*	10391	57	0.033199	0.00744	8.12E-06	0.2861	0.049723	8.72E-09	-0.256294	0.049714	2.53E-07
HbA1c	*RNF123*	16923	199	0.07172	0.00743	4.81E-22	-0.099985	0.033859	3.15E-03	0.173893	0.033844	2.77E-07
Osteoarthritis of knee	*CHADL*	21622	314	0.032267	0.023599	1.72E-01	0.403529	0.07843	2.67E-07	-0.400679	0.079521	4.69E-07
Other epidermal thickening	*FLG*	50338	1604	0.431399	0.044693	4.80E-22	0.632524	0.090253	2.41E-12	-0.503225	0.102238	8.56E-07
Seated height	*ABHD15*	10391	57	0.034719	0.007459	3.25E-06	0.258355	0.04986	2.20E-07	-0.226737	0.049852	5.41E-06
Creatinine	*LRP2*	6007	21	-0.108407	0.010437	2.86E-25	-0.494188	0.087662	1.73E-08	0.395035	0.087679	6.62E-06
Citrate	*PCK2*	12617	173	0.085365	0.00879	2.68E-22	0.23661	0.03696	1.54E-10	-0.159632	0.036963	1.57E-05
Aspartate aminotransferase	*FYCO1*	10359	71	0.115899	0.009382	4.65E-35	0.317892	0.056182	1.53E-08	-0.207924	0.056182	2.15E-04
Triglycerides	*TM6SF2*	11088	70	-0.144548	0.009225	2.49E-55	-0.317146	0.057383	3.26E-08	0.18074	0.057389	1.64E-03
Seated height	*MYH7B*	44787	1275	-0.03078	0.003672	5.23E-17	-0.054933	0.010546	1.90E-07	0.027467	0.010524	9.06E-03
Platelet count	*SH2B3*	2326	6	0.573548	0.019976	2.75E-181	1.04989	0.195792	8.22E-08	-0.498186	0.195796	1.09E-02
IGF-1	*PDE11A*	38426	987	0.063178	0.005042	5.15E-36	0.080388	0.015311	1.52E-07	-0.021886	0.015315	1.53E-01
Albumin	*SERPINA1*	33534	788	0.09889	0.005465	3.51E-73	0.111845	0.017359	1.17E-10	-0.019567	0.017383	2.60E-01
Mean corpuscular haemoglobin conc.	*ANK1*	29583	546	0.11153	0.005795	1.53E-82	0.129038	0.020885	6.48E-10	-0.021345	0.020895	3.07E-01
Lymphocyte count	*VWA7*	69975	3297	0.058515	0.003944	8.27E-50	0.059784	0.008682	5.73E-12	-0.00717	0.008658	4.08E-01
Mean reticulocyte volume	*ANK1*	29057	536	-0.196864	0.005898	2.60E-244	-0.179891	0.021239	2.46E-17	-0.011827	0.021263	5.78E-01
Mean sphered cell volume	*ANK1*	29057	536	-0.227421	0.005806	0.00E+00	-0.229839	0.020904	4.05E-28	0.009376	0.020933	6.54E-01
Albumin	*SERPINA1*	21348	497	0.144021	0.006932	7.08E-96	0.145039	0.02214	5.72E-11	-0.008098	0.022143	7.15E-01
HbA1c	*ANK1*	29119	545	-0.14848	0.005718	1.16E-148	-0.138433	0.020444	1.27E-11	-0.003895	0.02048	8.49E-01
Mean platelet (thrombocyte) volume	*IQGAP2*	16237	164	0.228209	0.00796	9.47E-181	0.230529	0.038962	3.28E-09	-0.006047	0.03906	8.77E-01
Albumin	*ABCA6*	16868	181	-0.194845	0.007613	1.75E-144	-0.193117	0.03615	9.18E-08	0.005027	0.036263	8.90E-01
Reticulocyte count	*ANK1*	29057	536	0.145901	0.005904	7.58E-135	0.142186	0.02126	2.26E-11	-0.000472	0.021285	9.82E-01
Impedance of arm (left)	*FLG*	49447	1576	0.03599	0.003256	2.14E-28	0.092436	0.00882	1.06E-25	NA	NA	NA
Vitamin D	*FLG*	45964	1464	0.141244	0.004733	1.03E-195	0.142089	0.012833	1.72E-28	NA	NA	NA

C

Supplement to chapter 4

Table C.1: ICD10 and ICD9 codes used to nominate the binary phenotypes investigated in this analysis. The 32 nominated phenotypes by ICD9 and ICD10 case and control inclusions and exclusions. Either ICD10, ICD9 or or a combination of the codes were utilized in each biobank to define cases and controls.

Phenotype ID	Sex	ICD10_control_exclude	ICD10_case_include	ICD9_control_exclude	ICD9_case_include
AMD	Both		H35.3		362.5
Asth	Both	J40,J41,J41.0,J41.1,J41.8,J42,J43,J43.0,J43.1,J43.2,J43.8,J43.9,J44,J44.0,J44.1,J44.8,J44.9,J47,J44.1,J44,J44.0,J47.1,J47.9,J47.0,J47,J43.2,J43.0,J43.1,J43.8,J43.9,J43,J40,J41.8,J42,J41.1,J41.0,J41,J98.01,J44.9	J45,J45.0,J45.1,J45.8,J45.9,J46,J45.31,J45.901,J45.42,J45.41,J45.51,J45.990,J45.902,J45.21,J45.52,J45.32,J45.22,J45.909,J45.998,J45.40,J45.20,J45,J45.991,J45.90,J45.30,J45.50	490,491,491.0,491.1,491.2,491.20,491.21,491.22,491.8,491.9,492,492.0,492.8,494,494.0,494.1,496,519.11	493,493.0,493.00,493.01,493.02,493.1,493.10,493.11,493.12,493.2,493.20,493.21,493.22,493.8,493.81,493.82,493.9,493.90,493.91,493.92
AFib	Both		I48,I48.0,I48.1,I48.2,I48.4,I48.9,I49		427.3
BenCervUterNeo	Female		D07.0,D39.0,D06,D06.0,D06.1,D06.7,D06.9,D26.0,D26.1,D26.7,D26.9,N84.0,N84.1,N87.0,N87.1,N87.2,N87.9		622.7,219,219.0,219.1,219.8,219.9,233.1,233.2,622.1,622.2
BenIntNeo	Both		D13.9,E34.0,K63.5,D01,D01.0,D01.1,D01.2,D01.3,D01.4,D01.5,D01.7,D01.9,D12,D12.0,D12.1,D12.2,D12.3,D12.4,D12.5,D12.6,D12.7,D12.8,D12.9,D13.3,D37.2,D37.3,D37.4,D37.5,K51.4,K62.0,K62.1,K62.2,K62.3		235.2,236.0,569.0,211.2,211.3,211.4,230.3,230.4,230.5,230.6,230.7,610,610.0,610.1,610.2,610.3,610.4,610.8,610.9
BreastCanc	Female		C50		174.0, 175, 174.1, 174.2, 174.3, 174.4, 174.5, 174.6, 174.8, 174.9
CervCanc	Female		Z85.41,C53,C53.0,C53.1,C53.8,C53.9,D06,D06.0,D06.1,D06.7,D06.9		V10.4.1,V72.3.2,180,180.0,180.1,180.8,180.9
COPD	Both	J40,J45,J45.0,J45.1,J45.8,J45.9,J46,J45.31,J45.901,J45.42,J45.41,J45.51,J45.990,J45.902,J45.21,J45.52,J45.32,J45.22,J40,J45.909,J45.998,J45.40,J45.20,J45,J45.991,J45.90,J45.30,J45.50,J98.01	J41,J41.0,J41.1,J41.8,J42,J43,J43.0,J43.1,J43.2,J43.8,J43.9,J44,J44.0,J44.1,J44.8,J44.9,J47,J44.1,J44,J44.0,J47.1,J47.9,J47.0,J47,J43.2,J43.0,J43.1,J43.8,J43.9,J43,J41.8,J42,J41.1,J41.0,J41,J44.9	490,493,493.0,493.00,493.01,493.02,493.1,493.10,493.11,493.12,493.2,493.20,493.21,493.22,493.8,493.81,493.82,493.9,493.90,493.91,493.92,519.11	491,491.0,491.1,491.2,491.20,491.21,491.22,491.8,491.9,492,492.0,492.8,494,494.0,494.1,496
CRF	Both		N18		585.9
ColonRectCanc	Both		C20,Z86.010,C18,C18.0,C18.1,C18.2,C18.3,C18.4,C18.5,C18.6,C18.7,C18.8,C18.9,C19,C21,C21.0,C21.1,C21.2,C21.8,Z12.1		V76.4.1,153,153.0,153.1,153.2,153.3,153.4,153.5,153.6,153.7,153.8,153.9,154,154.0,154.1,154.2,154.3,154.8
CAD	Both		I21.0, I21.1, I21.2, I21.3, I21.4, I21.9, I22.0, I22.1, I22.2, I22.8, I22.9, I23.0, I23.1, I23.2, I23.3, I23.4, I23.5, I23.6, I23.7, I23.8, I24.1, I25.2		410, 4109, 411, 4119, 412, 4129
EFRMB	Female		N92		626.2
FemInf	Female		N97, N97.0, N97.1, N97.2, N97.3, N97.4, N97.8, N97.9		6280, 6283, 6284, 6288, 6289
Gout	Both	M11.0,M11.1,M11.2,M11.8,M11.9,M67.9,M11.8,M11.049,M11.831,M11.052,M11.07,M11.86,M11.81,M11.842,M11.849,M11.01,M11.85,M11.83,M11.852,M11.019,M11.819,M11.08,M11.05,M11.061,M11.011,M11.821,M11.069,M11.839,M11.862,M11.042,M11.812,M11.87,M11.9,M11.829,M11.861,M11.879,M11.031,M11.82,M11.021,M11.051,M11.03,M11.022,M11.039,M11.04,M11.02,M11.029,M11.832,M11.041,M11.822,M11.89,M11.85,M11.811,M11.032,M11.871,M11.872,M11.06,M11.851,M11.00,M11.071,M11.079,M11.80,M11.88,M11.841,M11.89,M11.012,M11.072,M11.059,M11.84,M11.869,M11.062,M11.139,M11.17,M11.232,M11.149,M11.272,M11.29,M11.141,M11.1,M11.171,M11.251,M11.18,M11.212,M11.28,M11.142,M11.269,M11.112,M11.21,M11.131,M11.13,M11.129,M11.26,M11.231,M11.11,M11.179,M11.2,M11.152,M11.22,M11.279,M11.24,M11.259,M11.229,M11.161,M11.159,M11.211,M11.261,M11.23,M11.169,M11.132,M11.221,M11.122,M11.19,M11.242,M11.271,M11.12,M11.239,M11.10,M11.27,M11.219,M11.14,M11.241,M11.252,M11.16,M11.15,M11.20,M11.121,M11.249,M11.119,M11.162,M11.222,M11.25,M11.172,M11.262,M11.151,M11.111	M10,M10.00,M10.10,M10.20,M10.30,M10.40,M10.462,M10.449,M10.441,M10.40,M10.421,M10,M10.42,M10.451,M1A.9,M10.411,M10.431,M10.459,M10.452,M10.422,M10.469,M10.429,M10.48,M10.471,M10.479,M10.412,M10.45,M10.47,M10.439,M10.42,M10.419,M10.4,M10.432,M10.44,M10.49,M1A.9XX0,M10.46,M10.9,M10.41,M1A.39,M10.0,M1A.4291,M1A.4290,M1A.38X0,M10.139,M1A.3121,M10.049,M10.38,M1A.2611,M1A.042,M1A.029,M1A.2210,M10.041,M10.219,M1A.4511,M1A.451,M1A.212,M1A.3,M1A.221,M1A.379,M10.10,M1A.469,M10.311,M10.312,M1A.0490,M1A.4721,M1A.2XX0,M1A.342,M1A.3791,M1A.412,M10.26,M10.37,M1A.0121,M1A.4790,M1A.28X0,M1A.3120,M1A.0221,M1A.039,M1A.30,M10.25,M1A.3591,M10.28,M1A.3621,M10.39,M1A.041,M1A.28,M10.052,M10.169,M1A.3310,M1A.352,M1A.271,M1A.0001,M1A.022,M1A.29X1,M1A.0220,M10.332,M1A.490,M10.261,M1A.2391,M1A.450,M1A.339,M1A.459,M1A.4610,M10.06,M1A.0791,M1A.3410,M1A.011,M1A.249,M10.051,M10.162,M1A.079,M1A.361,M10.02,M1A.3711,M1A.3390,M10.361,M1A.461,M10.119,M1A.341,M1A.2410,M1A.0610,M1A.429,M1A.4111,M1A.469,M1A.38X1,M10.352,M1A.3620,M10.079,M1A.31,M1A.2720,M10.08,M1A.2120,M1A.219,M1A.2190,M1A.09,M1A.4791,M1A.0311,M1A.41,M10.141,M1A.319,M1A.3520,M10.33,M1A.20,M1A.42,M1A.2621,M1A.37,M10.269,M1A.0391,M1A.411,M1A.4211,M1A.2,M1A.4390,M1A.23,M1A.0410,M10.319,M10.372,M1A.27,M1A.2XX1,M1A.4290,M1A.2690,M10.09,M10.35,M1A.331,M1A.2191,M1A.071,M10.112,M1A.2221,M1A.49,M10.242,M10.122,M1A.3710,M1A.0691,M1A.4421,M1A.069,M1A.21,M10.00,M1A.0290,M1A.4491,M1A.259,M1A.0711,M10.179,M1A.2790,M1A.43,M1A.30X1,M1A.269,M10.129,M1A.231,M1A.4121,M1A.22,M1A.0611,M1A.04,M1A.341,M1A.3691,M10.232,M1A.0420,M1A.290,M1A.0411,M1A.33,M1A.2390,M1A.439,M1A.419,M10.2,M10.342,M1A.03,M10.019,M1A.3420,M1A.47,M1A.02,M10.021,M1A.4411,M10.36,M1A.2311,M10.331,M1A.46,M1A.4311,M10.359,M1A.08	712,712.1,712.10,712.11,712.12,712.13,712.14,712.15,712.16,712.17,712.18,712.19,712.2,712.20,712.21,712.22,712.23,712.24,712.25,712.26,712.27,712.28,712.29,712.3,712.30,712.31,712.32,712.33,712.34,712.35,712.36,712.37,712.38,712.39,712.8,712.81,712.82,712.83,712.84,712.85,712.86,712.87,712.88,712.89,712.9,712.90,712.91,712.92,	274,274.0,274.00,274.01,274.02,274.03,274.1,274.10,274.11,274.19,274.8,274.81,274.82,274.89,274.9
HF	Both	I50.0,I97.0,R09.8,R94.3,T86.3,Z94.1,Z94.3,R09.89,R00,I50.9,I09.81,I97.79,I97.1!0,I97.191,I97.71,I97.130,I97.111,I97.131,I97.791,I97.89,I97.190,I97.121,I97.88,I97.120,I97.790,I97.711,I97.710,I97.7,R94.30,R94.39,R94.3,I50.3,I50.30,I50.31,I50.33,I50.32,I25.812,I97.0,I25.769,I25.760,Z95.818,T86.23,T86.39,Z94.3,I25.761,I25.811,T86.2,I25.750,T86.3,T86.32,T86.20,T86.30,T86.22,T86.298,Z94.1,T86.31,I25.768,I25.758,Z95.812,Z95.811,T86.21,T86.29,I25.751,T86.33,T86.290,I25.759,I50.20,I50.40,I50.41,I50.23,I50.42,I50.21,I50.22,I50.4,I50.43,I50.2	I50,I50.1,I50.9,I50.82,I50.81,I50.8,I50.84,I50.810,I50.812,I50.83,I50.89,I50.814,I50,I50.811,I50.1,I50.813	398.91,414.06,414.07,428,428.0,428.00,428.2,428.20,428.21,428.22,428.23,428.3,428.30,428.31,428.32,428.33,428.4,428.40,428.41,428.42,428.43,429.4,429.81,429.82,785,785.9,794.3,794.30,794.39,99 6.83,997.1,V15.1,V42.1,V43.2,V43.21,V43.22	428,1,428.9
HTN	Both		I10		401.0, 401.1, 401.9,I10
IBD	Both		K50,K50.0,K50.1,K50.8,K50.9,K51,K51.0,K51.1,K51.2,K51.3,K51.5,K51.8,K51.9,K52,K52.8,K52.9		564.1,569.5,555,555.0,555.1,555.2,555.9,556,556.9,558,558.9
IFHern	Both		K40,K40.0,K40.1,K40.2,K40.3,K40.4,K40.9,K41,K41.0,K41.1,K41.2,K41.3,K41.4,K41.9,K42,K42.0,K42.1,K42.9,K44,K44.0,K44.1,K44.9,K45,K45.0,K45.1,K45.8,K46,K46.0,K46.1,K46.9		550,550.0,550.1,550.9,551,551.0,551.1,551.3,551.8,551.9,552,552.0,552.1,552.3,552.8,552.9,553,553.0,553.1,553.3,553.8,553.9
ILDSarc	Both		D86.9,D86,D86.0,D86.1,D86.2,I84,I84.0,I84.1,I84.8,I84.9		515,135,135.9,515.9,516,516.0,516.1,516.2,516.3,516.8,516.9
MatHem	Female		O62.2,O20,O20.0,O20.8,O20.9,O43.2,O44,O44.0,O44.1,O45,O45.0,O45.8,O45.9,O46,O46.0,O46.8,O46.9,O67,O67.0,O67.8,O67.9,O72,O72.0,O72.1,O72.2,O72.3		640,640.0,640.8,640.9,641,641.0,641.1,641.2,641.3,641.8,641.9,661.2,666,666.0,666.1,666.2,666.3
NonRheuValv	Both		I34,I34.0,I34.1,I34.2,I34.8,I34.9,I35,I35.0,I35.1,I35.2,I35.8,I35.9,I36,I36.0,I36.1,I36.2,I36.8,I36.9,I37,I37.0,I37.1,I37.2,I37.8,I37.9		424,0,424.1,424.2,424.3
Pancreat	Both		K85,K85.0,K85.1,K85.2,K85.3,K85.8,K85.9,K86,K86.0,K86.1,K86.2,K86.3,K86.8,K86.9		577,577.0,577.1,577.2,577.8,577.9
PeptUlcer	Both		K25,K25.0,K25.1,K25.2,K25.3,K25.4,K25.5,K25.6,K25.7,K25.9,K26,K26.0,K26.1,K26.2,K26.3,K26.4,K26.5,K26.6,K26.7,K26.9,K27,K27.0,K27.1,K27.2,K27.3,K27.4,K27.5,K27.6,K27.7,K27.9,K28,K28.0,K28.1,K28.2,K28.3,K28.4,K28.5,K28.6,K28.7,K28.9		531,531.0,531.1,531.2,531.3,531.4,531.5,531.6,531.7,531.9,532,532.0,532.1,532.2,532.3,532.4,532.5,532.6,532.7,532.9,533,533.0,533.1,533.2,533.3,533.4,533.5,533.6,533.7,533.9,534,534.0,534.1,534.2,534.3,534.4,534.5,534.6,534.7,534.9
PAD	Both		I70.0**,I70.2**,I70.3**,I70.4**,I70.5**,I70.6**,I70.7**,I70.9**,I73.9**		440,0,440.2,440.21,440.22,440.23,440.24,440.29,440.3,440.31,440.32,440.4,440.8,443.9
Psori	Both		L40,L40.0,L40.1,L40.2,L40.3,L40.4,L40.5,L40.8,L40.9,L41,L41.0,L41.1,L41.2,L41.3,L41.4,L41.5,L41.8,L41.9		696,696.0,696.1,696.2,696.3,696.4,696.5,696.8
RheumHeaDis	Both		I02.0,I01,I01.0,I01.1,I01.2,I01.8,I01.9,I05,I05.0,I05.1,I05.2,I05.8,I05.9,I06,I06.0,I06.1,I06.2,I06.8,I06.9,I07,I07.0,I07.1,I07.2,I07.8,I07.9,I08,I08.0,I08.1,I08.2,I08.3,I08.8,I08.9,I09,I09.0,I09.1,I09.2,I09.8,I09.9		392.0,391,391.0,391.1,391.2,391.8,391.9,393,393.9,394,394.0,394.1,394.2,394.9,395,395.0,395.1,395.2,395.9,396,396.9,397,397.0,397.1,397.9,398,398.0,398.9
RheumArth	Both		M05,M05.0,M05.00,M05.01,M05.02,M05.03,M05.04,M05.05,M05.06,M05.07,M05.08,M05.09,M05.1,M05.10,M05.11,M05.12,M05.13,M05.14,M05.15,M05.16,M05.17,M05.18,M05.19,M05.2,M05.20,M05.21,M05.22,M05.23,M05.24,M05.25,M05.26,M05.27,M05.28,M05.29,M05.3,M05.30,M05.31,M05.32,M05.33,M05.34,M05.35,M05.36,M05.37,M05.38,M05.39,M05.8,M05.80,M05.81,M05.82,M05.83,M05.84,M05.85,M05.86,M05.87,M05.88,M05.89,M05.9,M08,M08.0,M08.00,M08.01,M08.02,M08.03,M08.04,M08.05,M08.06,M08.07,M08.08,M08.09,M08.1,M08.10,M08.11,M08.12,M08.13,M08.14,M08.15,M08.16,M08.17,M08.18,M08.19,M08.2,M08.20,M08.21,M08.22,M08.23,M08.24,M08.25,M08.26,M08.27,M08.28,M08.29,M08.3,M08.30,M08.31,M08.32,M08.33,M08.34,M08.35,M08.36,M08.37,M08.38,M08.39,M08.4,M08.40,M08.41,M08.42,M08.43,M08.44,M08.45,M08.46,M08.47,M08.48,M08.49,M08.8,M08.80,M08.81,M08.82,M08.83,M08.84,M08.85,M08.86,M08.87,M08.88,M08.89,M08.9,M08.90,M08.91,M08.92,M08.93,M08.94,M08.95,M08.96,M08.97,M08.98,M08.99,M09,M09.0,M09.00,M09.01,M09.02,M09.03,M09.04,M09.05,M09.06,M09.07,M09.08,M09.09,M09.1,M09.10,M09.11,M09.12,M09.13,M09.14,M09.15,M09.16,M09.17,M09.18,M09.19,M09.2,M09.20,M09.21,M09.22,M09.23,M09.24,M09.25,M09.26,M09.27,M09.28,M09.29,M09.8		714,714.0,714.00,714.01,714.02,714.03,714.04,714.05,714.06,714.07,714.08,714.09,714.1,714.10,714.11,714.12,714.13,714.14,714.15,714.16,714.17,714.18,714.19,714.2,714.20,714.21,714.22,714.23,714.24,714.25,714.26,714.27,714.28,714.29,714.3,714.30,714.31,714.32,714.33,714.34,714.35,714.36,714.37,714.38,714.39,714.4,714.40,714.41,714.42,714.43,714.44,714.45,714.46,714.47,714.48,714.49,714.8,714.80,714.81,714.82,714.83,714.84,714.85,714.86,714.87,714.88,714.89,714.9
Stroke	Both	I63,I63.0,I63.1,I63.2,I63.3,I63.4,I63.5,I63.6,I63.8,I64,I63.349,I63.431,I63.533,I63.423,I63.439,I63.023,I63.421,I63.313,I63.311,I63.522,I63.34,I63.422,I63.513,I63.529,I63.413,I63.342,I63.43,I63.412,I63.54,I63.341,I63.542,I63.441,I63.319,I63.539,I63.5,I63.52,I63.32,I63.39,I63.31,I63.541,I63.523,I63.411,I63.49,I63.322,I63.443,I63.4,I63.33,I63.532,I63.44,I63.519,I63.339,I63.432,I63.449,I63.549,I63.433,I63.30,I63.42,I63.321,I63.0,I63.442,I63.323,I63.531,I63.511,I63.51,I63.312,I63.331,I63.512,I63.42,I63.50,I63.332,I63.33,I63.521,I63.343,I63.41,I63.419,I63.9,I63.5	346.6,346.60,346.61,346.62,346.63,430,430.0,431,431.0,432,432.0,432.1,432.9,433,433.0,433.00,433.01,433.1,433.10,433.11,433.2,433.20,433.21,433.3,433.30,433.31,433.8,433.80,433.81,433.9,433.90,433.91,434,434.0,434.00,434.1,434.10,434.9,434.90,435,435.0,435.1,435.2,435.3,435.8,435.9,436,437,437.0,437.1,437.3,437.4,437.5,437.6,437.7,437.8,437.9,438,438.0,438.1,438.10,438.11,438.12,438.13,438.14,438.19,438.2,438.20,438.21,438.22,438.3,438.30,438.31,438.32,438.4,438.40,438.41,438.42,438.43,438.49,438.5,438.50,438.51,438.52,438.53,438.6,438.7,438.8,438.81,438.82,438.83,438.84,438.85,438.89,438.9	434.01,434.11,434.91
T2Diab	Both	E10.**,E12.**, E13.**	E11.**, O24.1**	249.**,250.**	250.00,250.02,250.20,250.22,250.30,250.32,250.40,250.42,250.50,250.52,250.60,250.62,250.70,250.72, 250.80,250.82,250.90,250.92
Urolith	Both		N20,N20.0,N20.1,N20.2,N20.9,N21,N21.0,N21.1,N21.8,N21.9,N22,N22.0,N22.8,N23		592,592.0,592.1,592.9,594,594.0,594.1,594.2,594.8,594.9
VaricVeins	Both		I83,I83.*		454,454.0,454.1,454.2,454.8,454.9
VTE	Both	I81, I82.0, I80.0, I80.3, I80.8, I80.9, D68	I80.1, I80.2, I82.2 , I26.0, I26.9	452,453.0,451.0,451.89,451.9,286	451.11, 453.40, 453.2, 453.77, 453.87, 415.1

Table C.2: Enrichment of drug targets among genes with at least one biallelic pLoF variant. Analysis of genes with at least one biallelic pLoF carrier in the BRaVa cohort, enriched or depleted among genes that are considered drug target. Odds ratios with 95% confidence intervals (CIs) from two-sided Fisher's exact tests are shown. Drug target categories include approved, investigational, withdrawn, illicit, nutraceutical, experimental, small molecule, and biotech (biological) drugs.

Odds Ratio	Lower CI	Upper CI	P	Drug Target category
0.769339265	0.652680885	0.903972233	0.001187771	approved
0.773468821	0.545111159	1.080703789	0.133293593	withdrawn
0.812158106	0.43809273	1.436698314	0.503417185	illicit
0.783249728	0.657541918	0.929611306	0.004555073	investigational
0.782832695	0.662616097	0.921837134	0.002849984	small_molecule
0.869555517	0.630692998	1.184833858	0.408761573	biotech
0.618825077	0.345947424	1.053846412	0.076268541	nutraceutical
0.963246697	0.695642609	1.319028071	0.876330649	experimental

Table C.3: Enrichment of biallelic pLoF variants in essential and non-essential gene sets. Analysis of genes harboring at least one biallelic pLoF variant in the BRaVa cohort, potentially enriched or depleted across essential and non-essential gene sets. Odds ratios (OR) with 95% confidence intervals (CIs) from two-sided Fisher's exact tests are shown.

Odds Ratio	Lower CI	Upper CI	P	Geneset
0.227534	0.179856	0.284683	5.60E-55	essential_adam_vincenti2021
0.232856	0.171778	0.309982	2.04E-33	essential_crispr_hart2017
0.192385	0.159114	0.230916	4.71E-105	essential_gnomad_karczewski2020
0.231942	0.147072	0.35107	9.30E-17	essential_in_culture_hart2014
0.384753	0.340231	0.434149	1.61E-64	essential_in_mice_georgi2013
9.174716	8.134115	10.35982	0	homozygous_lof_tolerant_karczewski2020
1.784197	1.51185	2.103721	5.66E-12	non_essential_gnomad_karczewski2020
1.784197	1.51185	2.103721	5.66E-12	non_essential_in_culture_hart2014

Table C.4: Manually curated HPO and ICD-10 codes for NDD status determination in GEL. Summary of proband cases with neurodevelopmental disorders (NDD), including global developmental delay, intellectual disability, autism, or any combination thereof. Case inclusion criteria based on ICD-10 and HPO codes in the GEL biobank.

Description	Samples (Probands)	ICD10 case include	ICD10 control exclude	HPO case include	HPO control exclude
global developmental delay	12,191	R62,F80,F81,F82,F83,F88,F89	NA	HP:0001263,HP:0012736,HP:0011344,HP:0011343, HP:0011342	NA
intellectual disability	9,149	F70,F71,F72,F73,F78,F79	NA	HP:0001249,HP:0002187,HP:0010864,HP:0002342, HP:0001256,HP:0006887,HP:0006889	NA
autism	4,609	F84	NA	HP:0000717,HP:0000729,HP:0000753	NA
Neurodevelopmental disorder (NDD)	13,976	R62,F80,F81,F82,F83,F88,F89, F70,F71,F72,F73,F78,F79,F84	NA	HP:0001263,HP:0012736,HP:0011344,HP:0011343, HP:0011342,HP:0001249,HP:0002187,HP:0010864, HP:0002342,HP:0001256,HP:0006887,HP:0006889, HP:0000717,HP:0000729,HP:0000753	NA

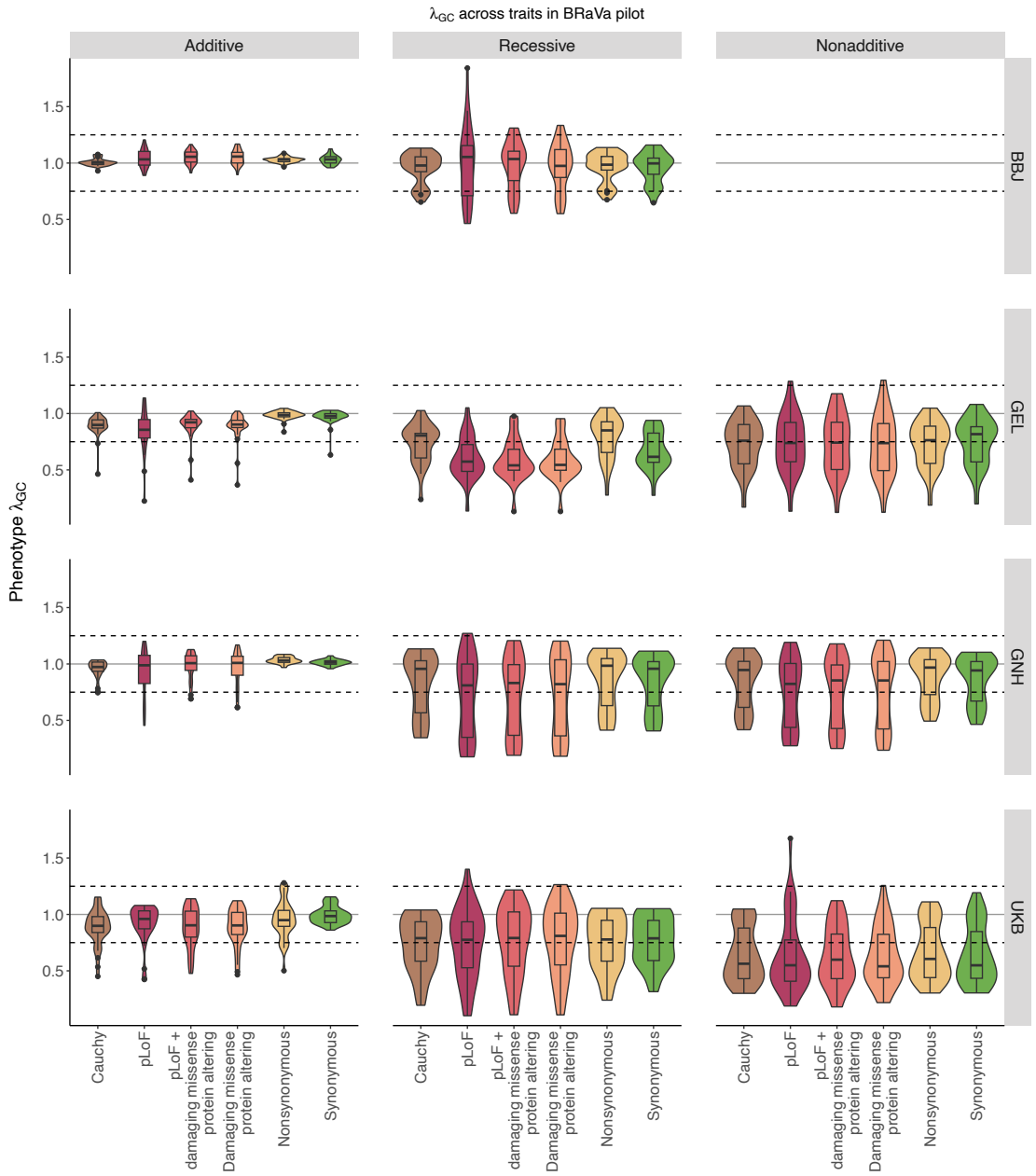


Figure C.1: Comparison of test-statistic inflation across biobanks and annotation categories before meta-analysis using λ_{GC}. Violin plot with overlaid boxplot showing the distribution of inflation (λ_{GC}) on the y-axis. X-axis represents different variant annotation masks, including Cauchy-combined P-values. Dashed lines indicate $\lambda = 1.25$ and $\lambda = 0.75$ cutoffs. Violin shape depicts probability density; boxplot shows median, quartiles (Q1, Q3 as hinges), and whiskers extending to $Q1 - 1.5\times\text{IQR}$ and $Q3 + 1.5\times\text{IQR}$. Outliers beyond whiskers are plotted individually.

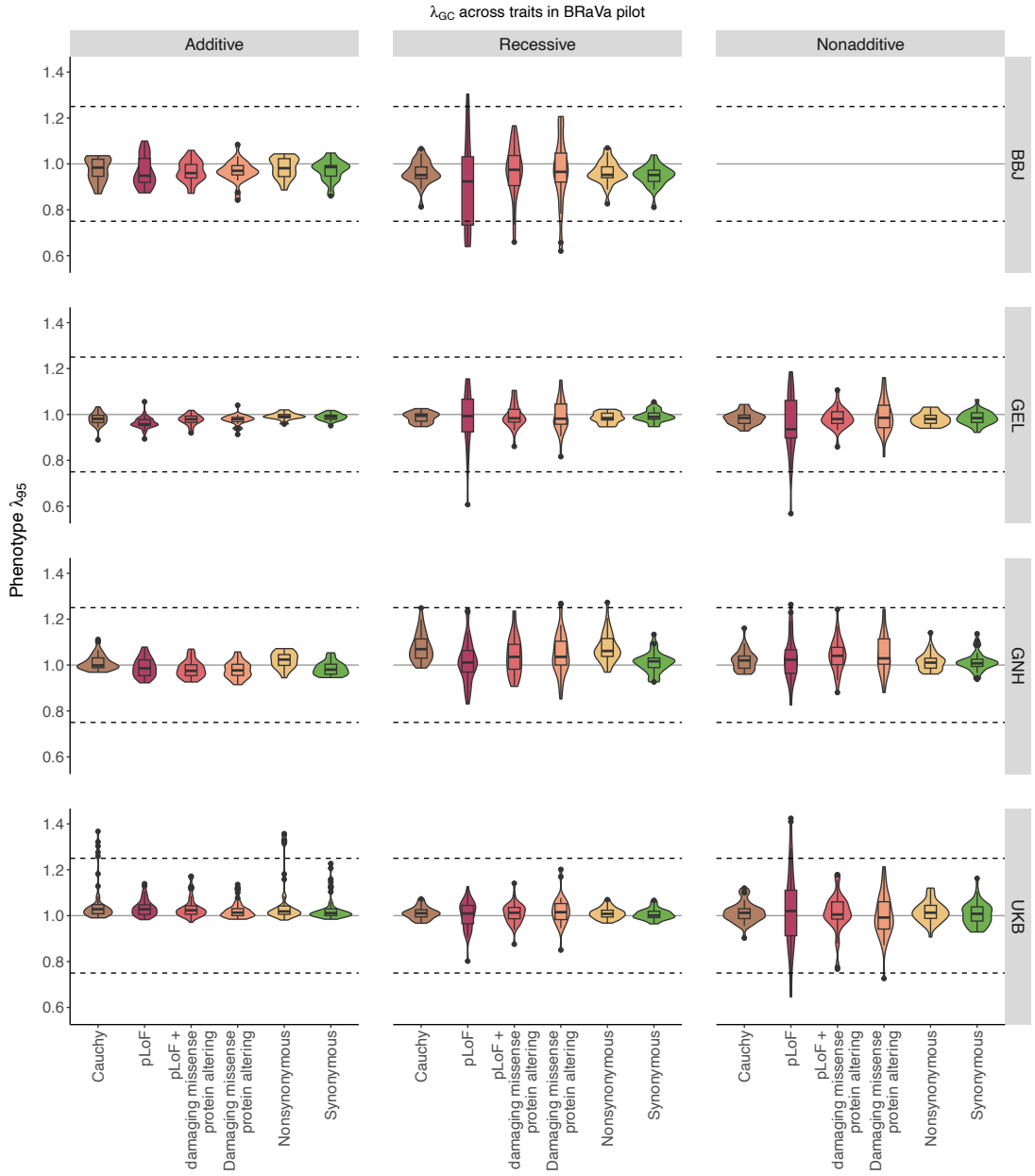


Figure C.2: Comparison of test-statistic inflation across biobanks and annotation categories before meta-analysis using λ_{95}. Violin plot with overlaid boxplot showing the distribution of inflation (λ_{95}, lambda at 95th percentile) on the y-axis. X-axis represents different variant annotation masks, including Cauchy-combined p-values. Dashed lines indicate $\lambda = 1.25$ and $\lambda = 0.75$ cutoffs. Violin shape depicts probability density; boxplot shows median, quartiles (Q1, Q3 as hinges), and whiskers extending to $Q1 - 1.5\times\text{IQR}$ and $Q3 + 1.5\times\text{IQR}$. Outliers beyond whiskers are plotted individually.

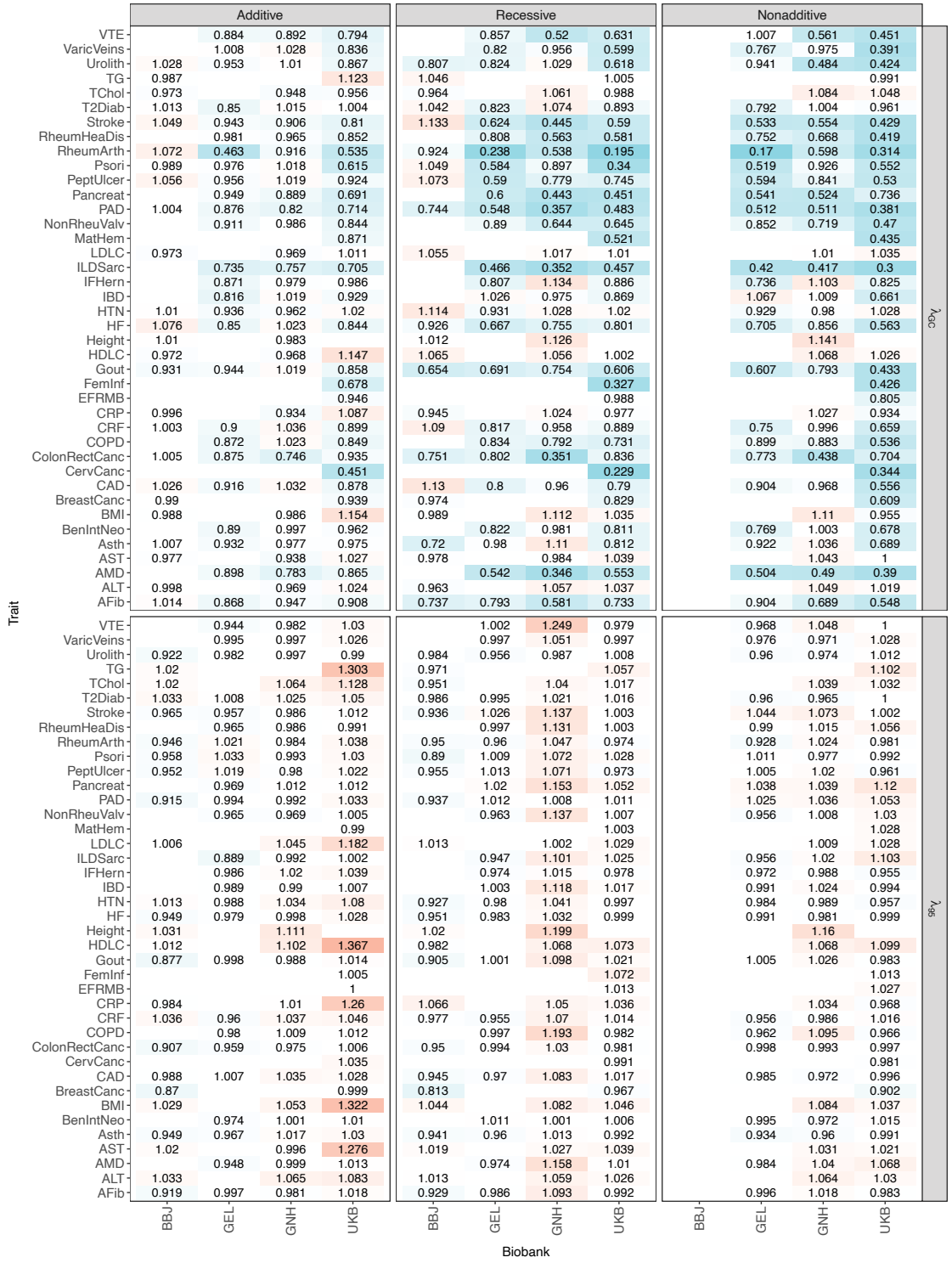


Figure C.3: Pre-meta-analysis λ_{GC} and λ_{95} for synonymous variants across traits and genetic models. λ_{GC} and λ_{95} (lambda at 95th percentile) values for synonymous variants across traits, shown for additive, recessive, and nonadditive genetic models. Color scale: red indicates higher lambda (inflation), blue indicates lower lambda (deflation).

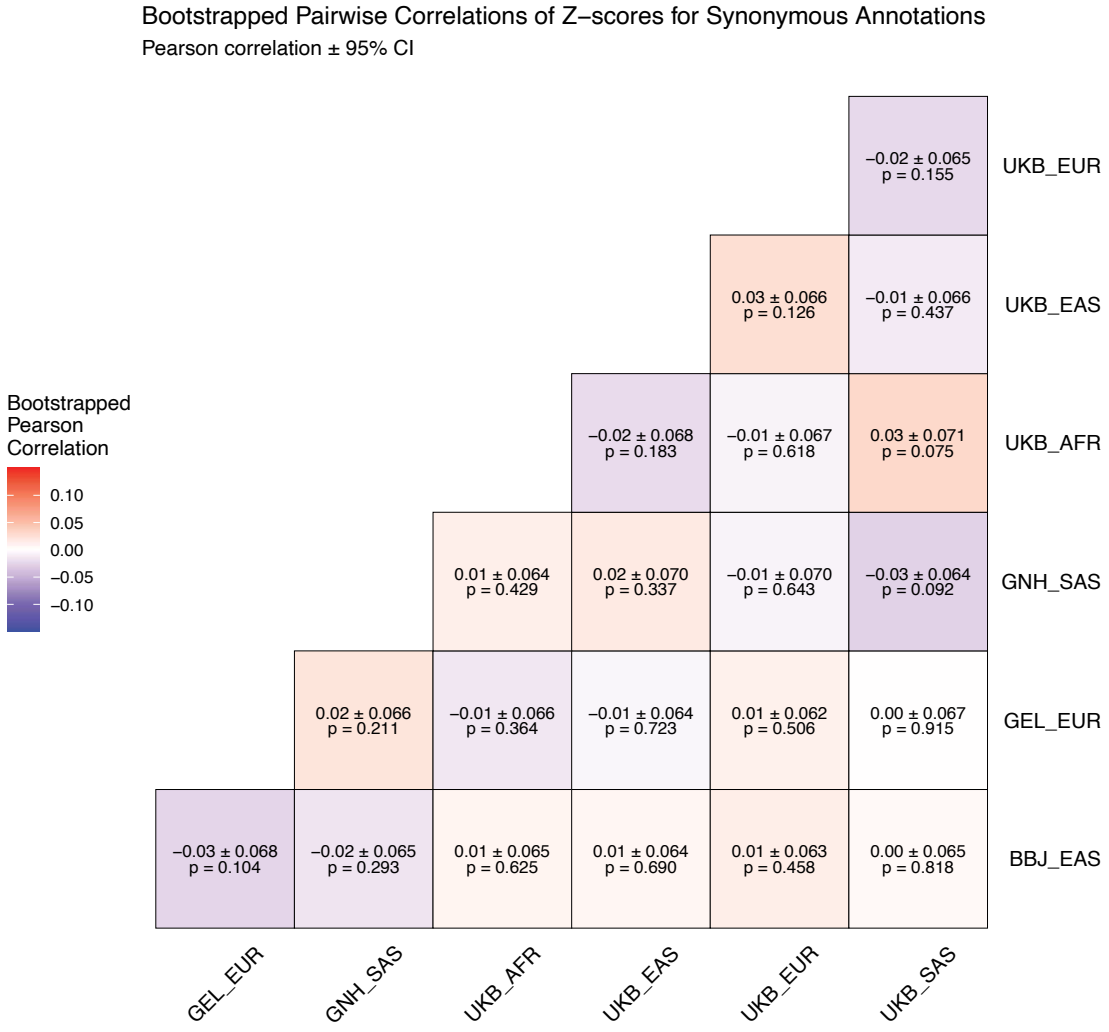
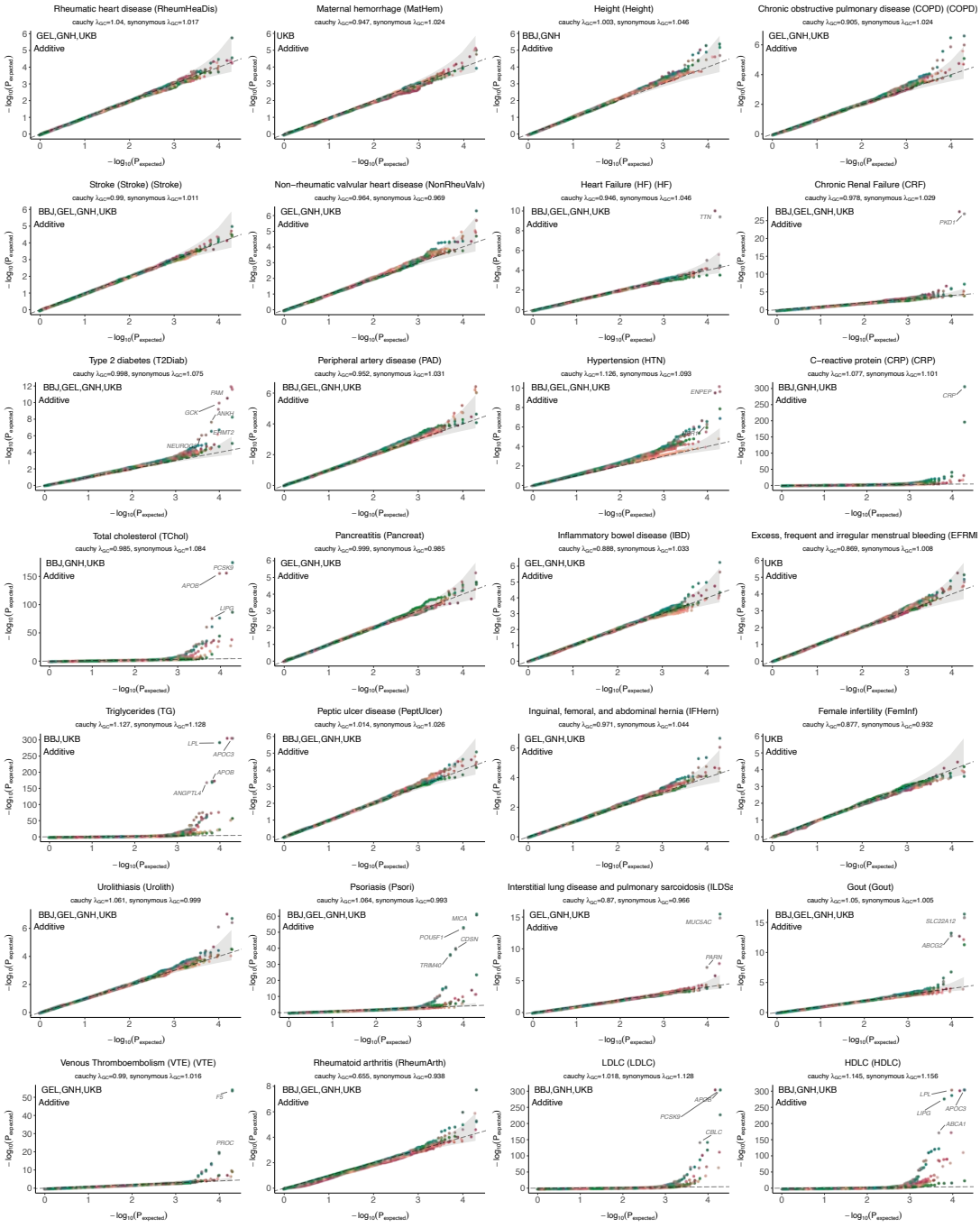


Figure C.4: Investigation of sample overlap between biobanks using Z-score correlations. Pearson correlations of Z-scores ($\frac{\text{BETA}}{\text{SE}}$) from synonymous variants across biobanks and ancestries. Each tile displays: correlation coefficient, 95% CI (from 1000 ordinary bootstrap replicates), and P-value. Color intensity indicates correlation strength (red: higher, blue: lower).

Table C.5: Heterogeneity analysis of 40 significant gene-trait associations from recessive meta-analysis. Cochran's Q heterogeneity *P*-values for 40 significant (FDR< 0.01) gene-trait associations identified using the recessive genetic model. All annotation masks are included, each with its corresponding heterogeneity *P*-value. 'NA' indicates no gene-trait associations for a given annotation mask.

Trait	Gene	*P* Heterogenity by annotation				
		pLoF	pLoF + damaging missense	damaging_missense	nonsynonymous	synonymous
AFib	RPL3L	1.00E+00	1.00E+00	NA	1.86E-02	4.37E-07
ALT	GPT	NA	NA	NA	2.52E-01	1.00E+00
ALT	SERPINA1	NA	1.00E+00	1.00E+00	9.55E-02	2.32E-01
AST	ANO5	NA	NA	NA	1.00E+00	1.00E+00
AST	APOH	1.00E+00	1.02E-01	1.00E+00	2.24E-01	3.35E-02
AST	FYCO1	NA	1.00E+00	1.00E+00	8.39E-01	9.16E-01
AST	SERPINA1	NA	1.00E+00	1.00E+00	1.00E+00	NA
Asth	FLG	NA	2.13E-07	2.39E-07	9.79E-01	7.85E-01
Asth	TNXB	NA	8.81E-10	1.00E+00	5.93E-01	8.95E-02
BenIntNeo	MUTYH	NA	2.62E-06	1.05E-06	1.92E-05	1.60E-26
COPD	DNAI1	NA	3.42E-77	1.00E+00	7.36E-87	2.05E-01
COPD	FLG	NA	2.60E-05	2.61E-05	5.32E-01	5.94E-01
COPD	ODAD1	1.00E+00	1.00E+00	NA	1.28E-10	2.37E-08
COPD	SERPINA1	NA	4.35E-02	3.45E-02	1.46E-07	7.79E-07
CRP	CRP	NA	NA	NA	0.00E+00	1.74E-01
ColonRectCanc	MUTYH	NA	5.77E-01	5.48E-01	4.20E-12	9.91E-01
HDLC	ABCA1	NA	NA	NA	2.62E-01	8.92E-01
HDLC	ANGPTL4	NA	NA	NA	7.32E-01	1.00E+00
HDLC	APOB	NA	NA	NA	9.16E-01	8.63E-01
HDLC	BTNL9	2.81E-01	8.81E-01	NA	5.20E-01	7.28E-01
HDLC	CD300LG	NA	NA	NA	8.74E-01	NA
HDLC	HBB	NA	1.00E+00	0.00E+00	1.00E+00	1.00E+00
HDLC	HNF4A	NA	NA	NA	3.82E-01	9.26E-01
HDLC	LIPG	NA	1.00E+00	1.00E+00	2.45E-03	NA
HDLC	LPL	NA	NA	NA	7.86E-01	5.10E-03
LDLC	APOB	NA	NA	NA	8.68E-01	9.86E-01
LDLC	BCAM	NA	NA	NA	4.58E-01	7.02E-01
LDLC	CBLC	1.00E+00	8.18E-02	1.16E-06	1.36E-01	0.00E+00
LDLC	PCSK9	NA	NA	NA	1.08E-01	5.07E-01
LDLC	TM6SF2	NA	1.00E+00	1.00E+00	1.00E+00	1.00E+00
Psori	MICA	1.00E+00	1.00E+00	NA	1.50E-04	4.64E-04
TChol	APOB	NA	NA	NA	9.68E-01	9.36E-01
TChol	CBLC	1.00E+00	1.01E-02	1.72E-11	4.52E-01	1.00E+00
TChol	HBB	NA	1.00E+00	0.00E+00	0.00E+00	NA
TChol	PCSK9	NA	NA	NA	2.80E-01	9.86E-01
TG	ANGPTL4	NA	NA	NA	1.00E+00	NA
TG	APOB	NA	NA	NA	8.59E-01	4.10E-01
TG	LPL	NA	NA	NA	1.00E+00	1.00E+00
TG	TM6SF2	NA	1.00E+00	1.00E+00	1.00E+00	1.00E+00
VTE	F5	NA	NA	NA	3.59E-02	9.86E-06



Continued on next page

↓

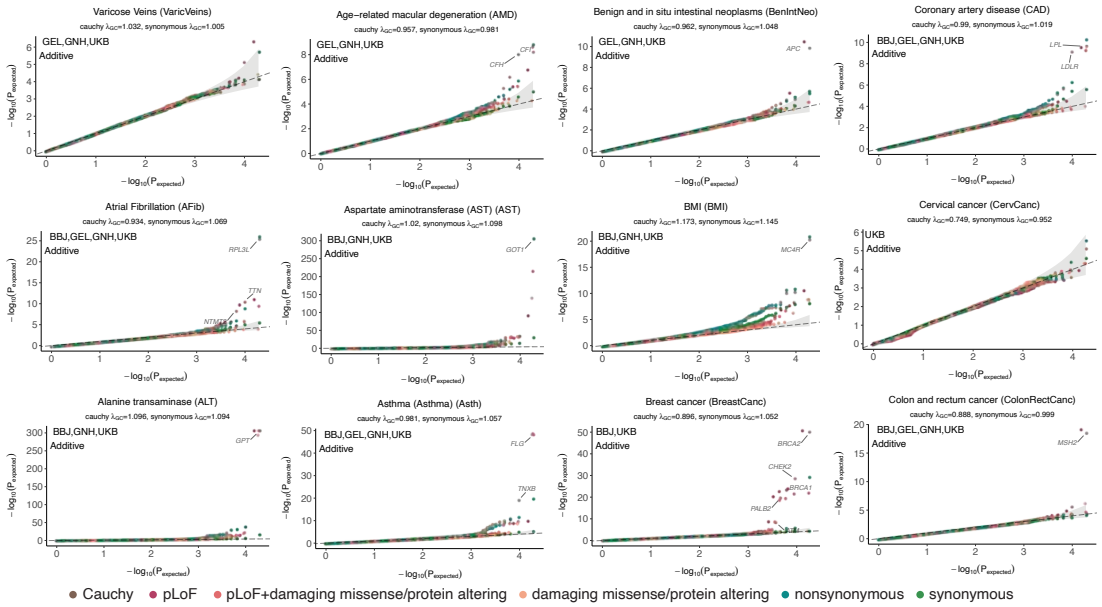
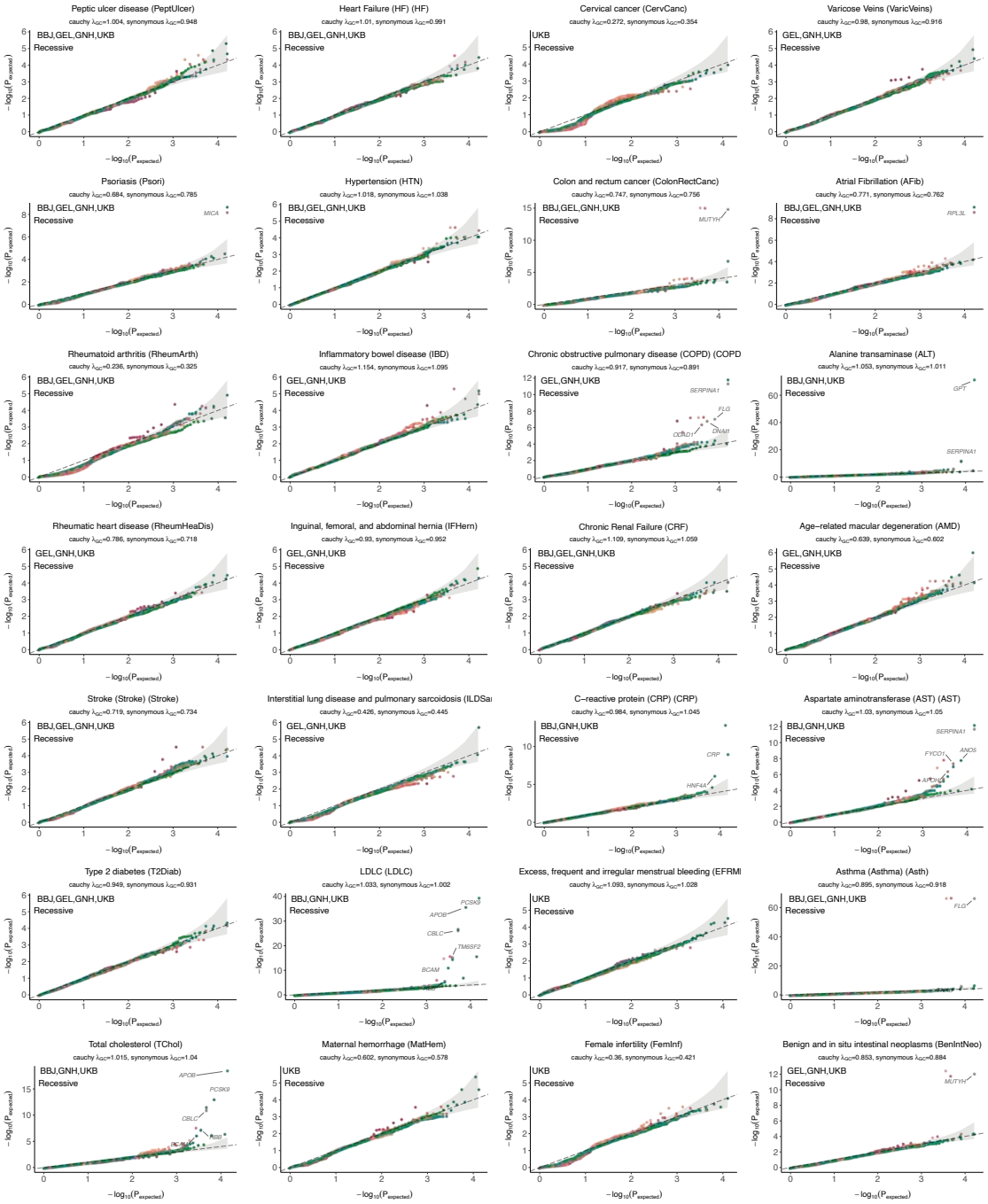


Figure C.6: QQ-plots resulting from additive meta-analysis across up to 4 different biobanks. The biobanks included in the analysis are shown in the subtitle. Five different annotation categories are annotated: pLoF (dark red), pLoF+damaging missense/protein altering (light red), damaging missense/protein altering (orange) nonsynonymous (darkgreen), and synonymous (lightgreen) variants. x-axis indicate the expected P-value whereas y-axis indicate the observed.



Continued on next page

↓

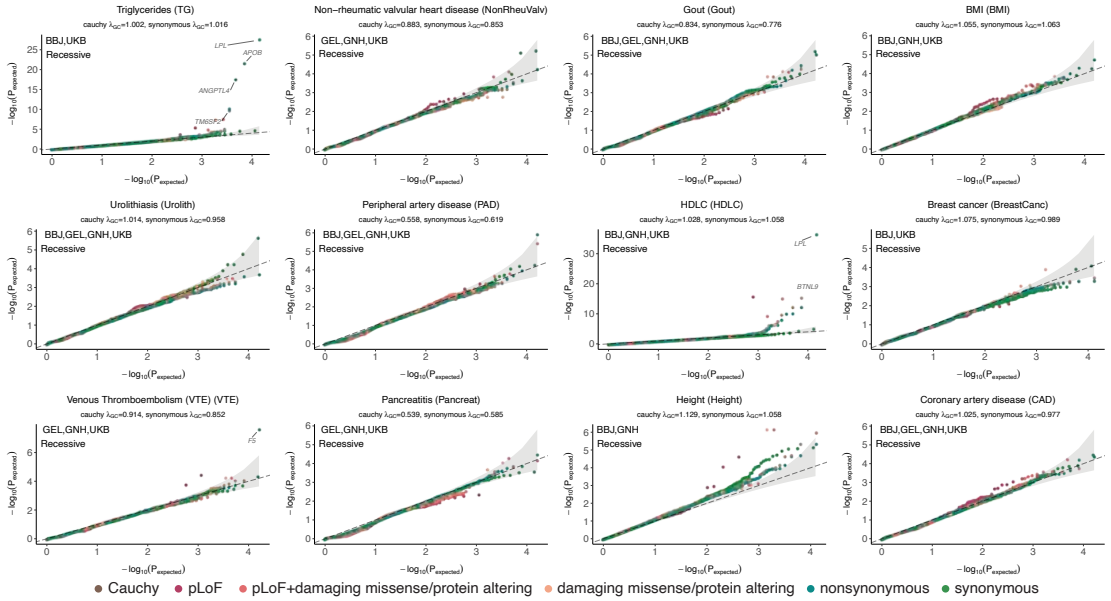
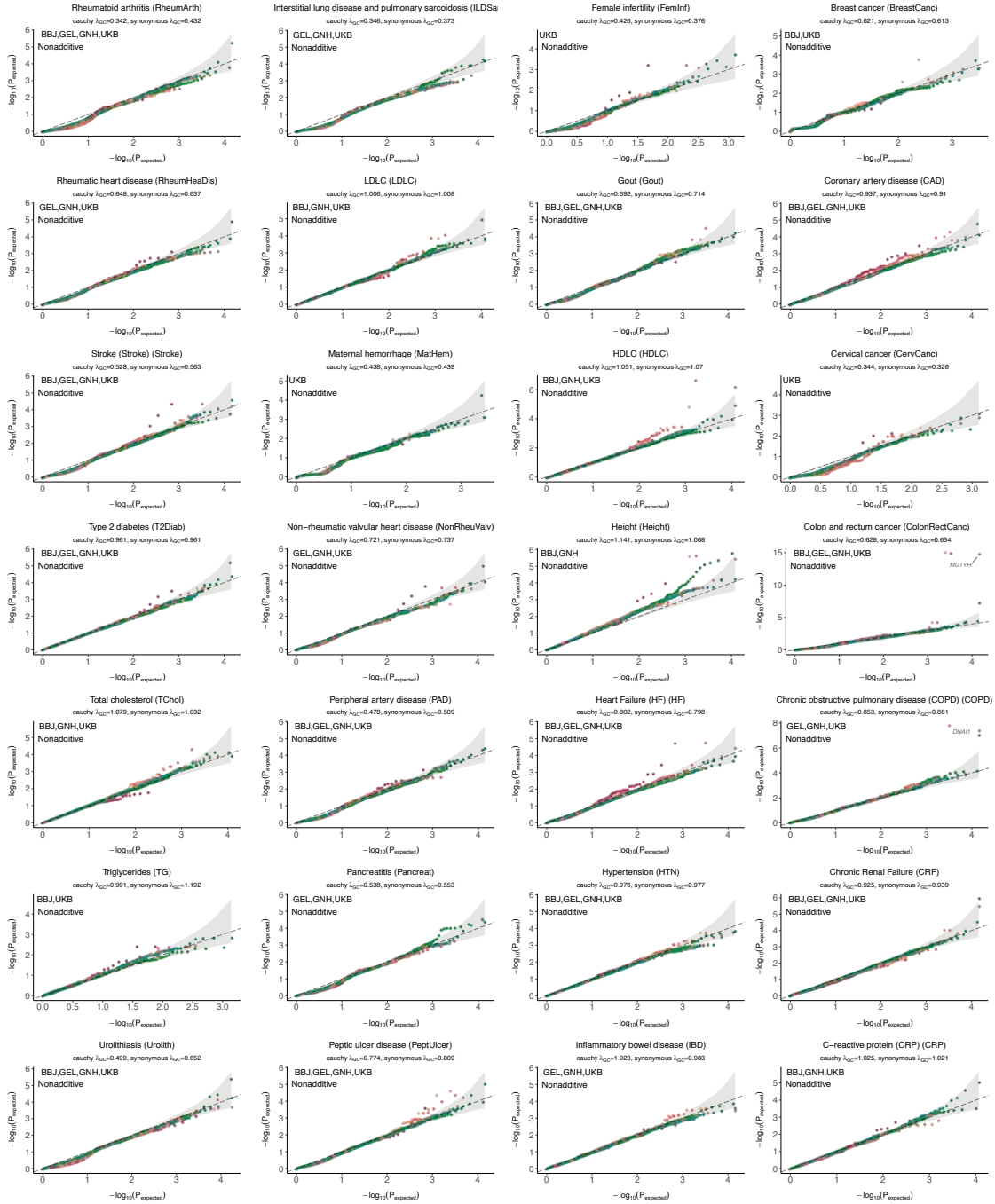


Figure C.8: QQ-plots from recessive meta-analysis across up to 4 biobanks. Biobanks included are specified in the subtitle. Five annotation categories are shown: pLoF (dark red), pLoF+damaging missense/protein altering (light red), damaging missense/protein altering (orange), nonsynonymous (dark green), and synonymous (light green) variants. x-axis shows expected P-values; y-axis shows observed P-values.



Continued on next page

↓

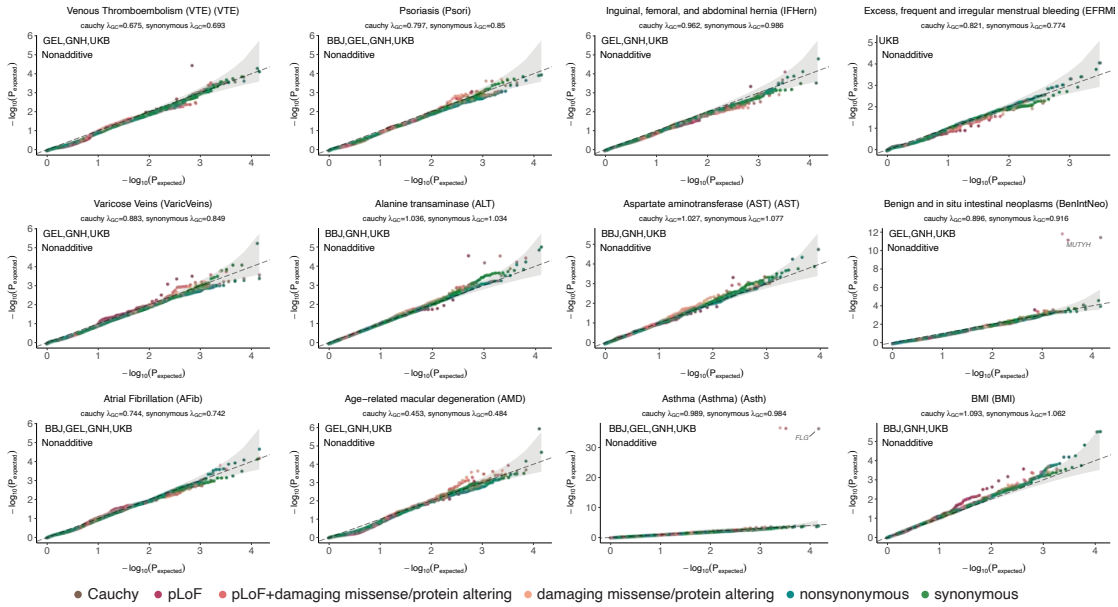


Figure C.10: QQ-plots from nonadditive meta-analysis across up to 4 biobanks. Biobanks included are specified in the subtitle. Five annotation categories are shown: pLoF (dark red), pLoF+damaging missense/protein altering (light red), damaging missense/protein altering (orange), nonsynonymous (dark green), and synonymous (light green) variants. x-axis shows expected P-values; y-axis shows observed P-values.

Table C.6: Comparison of significant additive synonymous associations with Genebass for HDLC. Summary of 32 significant synonymous associations (FDR< 0.01) with HDL cholesterol from our meta-analysis, compared to synonymous associations for the same genes in Genebass[251]. Note: Genebass imposes a MAF cutoff of 1%.

SYMBOL	TRAIT	ANNOTATION	META P	Genebass reported P-values			*P*<0.05
				SKAT	Burden	SKATO	
AGRP	HDLC	synonymous	6.47E-06	1.49E-05	0.000353386	1.60E-05	YES
APOA4	HDLC	synonymous	6.57E-26	7.30E-05	0.102758978	0.000169245	YES
B4GALNT4	HDLC	synonymous	2.28E-07	0.729569092	0.128123256	0.225239552	NO
BAP1	HDLC	synonymous	1.39E-05	0.000124591	0.000314178	0.000133916	YES
CD300LG	HDLC	synonymous	2.17E-18	1.98E-26	3.20E-22	2.07E-26	YES
CDK12	HDLC	synonymous	5.59E-06	0.213455351	0.041628223	0.07228885	YES
CEACAM19	HDLC	synonymous	2.11E-09	0.005299156	0.119852104	0.007622831	YES
CETP	HDLC	synonymous	2.63E-19	1.59E-16	0.004594274	1.11E-15	YES
CTCF	HDLC	synonymous	1.14E-11	0.21031531	0.026530833	0.047087236	YES
ELMO2	HDLC	synonymous	3.38E-06	8.80E-08	3.64E-05	6.86E-09	YES
GLB1L	HDLC	synonymous	3.98E-06	0.493206844	0.938170228	0.702952438	NO
GNAI2	HDLC	synonymous	2.79E-08	0.265352541	0.341141692	0.434970583	NO
GOLM2	HDLC	synonymous	1.91E-18	0.393596637	0.590483061	0.609628345	NO
HERPUD1	HDLC	synonymous	2.30E-05	0.721136497	0.888769535	0.908431478	NO
LIMS1	HDLC	synonymous	7.50E-06	0.000664852	0.008217232	0.000910171	YES
MAD2L1BP	HDLC	synonymous	2.06E-05	0.306048468	0.43035874	0.465179997	NO
MYPOP	HDLC	synonymous	1.25E-09	0.345192786	0.098399091	0.162481789	NO
MYRF	HDLC	synonymous	1.17E-05	0.000556471	0.007061972	0.00068429	YES
NBPF10	HDLC	synonymous	3.17E-06	0.822777247	0.691020797	0.8844062	NO
NFKB1	HDLC	synonymous	4.25E-06	0.341953201	0.028010039	0.048792117	YES
OR13C8	HDLC	synonymous	1.19E-11	0.330507562	0.417047279	0.457483422	NO
RIPOR1	HDLC	synonymous	1.56E-05	3.41E-11	0.70300627	2.17E-10	YES
SEMA3G	HDLC	synonymous	3.37E-06	2.14E-05	0.127256536	3.30E-05	YES
SIDT2	HDLC	synonymous	7.64E-06	1.87E-30	8.56E-22	6.49E-30	YES
SLC4A1	HDLC	synonymous	9.44E-13	0.344034098	0.00102435	0.00188356	YES
SULF2	HDLC	synonymous	4.62E-07	0.977907113	0.480989377	0.700125393	NO
TLR9	HDLC	synonymous	6.05E-06	3.92E-05	0.000136804	3.25E-05	YES
TMEM262	HDLC	synonymous	7.60E-09	0.954241247	0.822645518	1	NO
TNKS1BP1	HDLC	synonymous	7.27E-11	0.384205635	0.828992147	0.585858691	NO
TOMM40	HDLC	synonymous	7.80E-13	0.181556969	0.313702772	0.286491778	NO
TRIM64C	HDLC	synonymous	3.38E-06	0.165757285	0.850321006	0.290112282	NO
ZDHHC1	HDLC	synonymous	1.54E-05	0.221024145	0.163016266	0.269448136	NO

D

Supplement to chapter 5

Table D.1: FDA-approved cellular and gene therapy products as of October 2024. Overview of 38 cellular and gene therapies approved by the FDA. The table includes target gene (if applicable), treated condition, mode of inheritance, and producing company for each therapy.

Gene Therapy	Gene Symbol	Condition Treated	Mode of Inheritance	Company
ABECMA	*BCMA*	Multiple myeloma	NA	Celgene Corporation
ADSTILADRIN	Interferon alfa-2b	Bladder cancer	NA	Ferring Pharmaceuticals A/S
ALLOCORD	N/A	Hematopoietic disorders	NA	SSM Cardinal Glennon Children's Medical Center
AMTAGVI	N/A	Melanoma	NA	Iovance Biotherapeutics, Inc.
BEQVEZ	*F9*	Hemophilia B	X-linked recessive	Pfizer, Inc.
BREYANZI	*CD19*	Large B-cell lymphoma	NA	Juno Therapeutics
CARVYKTI	*BCMA*	Multiple myeloma	NA	Janssen Biotech, Inc.
CASGEVY	*HBB*	Sickle cell disease and beta-thalassemia	Autosomal recessive	Vertex Pharmaceuticals
CLEVECORD	N/A	Hematopoietic disorders	NA	Cleveland Cord Blood Center
Ducord	N/A	Hematopoietic disorders	NA	Duke University School of Medicine
ELEVIDYS	*DMD*	Duchenne muscular dystrophy	X-linked recessive	Sarepta Therapeutics, Inc.
GINTUIT	N/A	Mucogingival conditions	NA	Organogenesis Incorporated
HEMACORD	N/A	Hematopoietic disorders	NA	New York Blood Center
HEMGENIX	*F9*	Hemophilia B	X-linked recessive	CSL Behring LLC
HPC, Cord Blood	N/A	Hematopoietic disorders	NA	Clinimmune Labs, University of Colorado
HPC, Cord Blood	N/A	Hematopoietic disorders	NA	MD Anderson Cord Blood Bank
HPC, Cord Blood	N/A	Hematopoietic disorders	NA	LifeSouth Community Blood Centers, Inc.
HPC, Cord Blood	N/A	Hematopoietic disorders	NA	Bloodworks
IMLYGIC	N/A	Melanoma	NA	BioVex
KYMRIAH	*CD19*	B-cell acute lymphoblastic leukemia and large B-cell lymphoma	NA	Novartis Pharmaceuticals
LANTIDRA	N/A	Type 1 diabetes	NA	CellTrans Inc.
LAVIV	N/A	Nasolabial fold wrinkles	NA	Fibrocell Technologies
LENMELDY	*ARSA*	Metachromatic leukodystrophy	Autosomal recessive	Orchard Therapeutics
LUXTURNA	*RPE65*	Inherited retinal dystrophy	Autosomal recessive	Spark Therapeutics
LYFGENIA	*HBB*	Sickle cell disease	Autosomal recessive	bluebird bio, Inc.
MACI	N/A	Cartilage defects	NA	Vericel Corp.
OMISIRGE	N/A	Hematopoietic stem cell transplantation	NA	Gamida Cell Ltd.
PROVENGE	N/A	Prostate cancer	NA	Dendreon Corp.
RETHYMIC	N/A	Congenital athymia	NA	Enzyvant Therapeutics GmbH
ROCTAVIAN	*F8*	Hemophilia A	X-linked recessive	BioMarin Pharmaceutical
SKYSONA	*ABCD1*	Cerebral adrenoleukodystrophy	X-linked recessive	bluebird bio, Inc.
STRATAGRAFT	N/A	Deep partial-thickness burns	NA	Stratatech Corporation
TECARTUS	*CD19*	Mantle cell lymphoma and B-cell acute lymphoblastic leukemia	NA	Kite Pharma, Inc.
TECELRA	*MAGE-A4*	Synovial sarcoma	NA	Adaptimmune LLC
VYJUVEK	*COL7A1*	Dystrophic epidermolysis bullosa	Autosomal recessive	Krystal Biotech, Inc.
YESCARTA	*CD19*	Large B-cell lymphoma and follicular lymphoma	NA	Kite Pharma, Inc.
ZYNTEGLO	*HBB*	Beta-thalassemia	Autosomal recessive	bluebird bio, Inc.
ZOLGENSMA	*SMN1*	Spinal muscular atrophy	Autosomal recessive	Novartis Gene Therapies

References

1. Bateson, W. & Gregory, R. P. On the inheritance of heterostylism in Primula. *Proceedings of the Royal Society of London. Series B, Containing Papers of a Biological Character* **76.** Publisher: Royal Society, 581–586. doi:10.1098/rspb.1905.0049 (1905).
2. Harper, P. S. & Harper, P. S. *A Short History of Medical Genetics* (Oxford University Press, Oxford, New York, 2008).
3. Mukherjee, S. *The Gene: An Intimate History* en (2016).
4. Lamarck, J.-B.-P.-A. d. M. d. *Philosophie zoologique; ou, Exposition des considérations relatives à l'histoire naturelle des animaux. Nouv. éd., rev. et précédée d'une introd. biographique par Charles Martins* Nouv. éd. doi:10.5962/bhl.title.26297 (F. Savy, Paris, 1873).
5. Skvortsova, K., Iovino, N. & Bogdanović, O. Functions and mechanisms of epigenetic inheritance in animals. en. *Nature Reviews Molecular Cell Biology* **19.** Publisher: Nature Publishing Group, 774–790. doi:10.1038/s41580-018-0074-2 (2018).
6. Grant, P. R. & Grant, B. R. Evolution of Character Displacement in Darwin's Finches. *Science* **313.** Publisher: American Association for the Advancement of Science, 224–226. doi:10.1126/science.1128374 (2006).
7. Grant, P. R. Natural Selection and Darwin's Finches. *Scientific American* **265.** Publisher: Scientific American, a division of Nature America, Inc., 82–87 (1991).
8. Darwin, C. in *British Politics and the Environment in the Long Nineteenth Century* Num Pages: 9 (Routledge, 2023).
9. Ellis, T. H. N., Hofer, J. M. I., Swain, M. T. & van Dijk, P. J. Mendel's pea crosses: varieties, traits and statistics. *Hereditas* **156,** 33. doi:10.1186/s41065-019-0111-y (2019).
10. Mendel, G. & Mendel, G. *Versuche über Pflanzen-Hybriden* Pages: 1-464. doi:10.5962/bhl.title.61004 (Im Verlage des Vereines, Brünn, 1866).
11. Weldon, W. F. R. Mendel's Laws of Alternative Inheritance in Peas. *Biometrika* **1.** Publisher: [Oxford University Press, Biometrika Trust], 228–254. doi:10.2307/2331488 (1902).
12. Simunek, M. V., Mielewczik, M., Levit, G. S. & Hossfeld, U. Armin von Tschermak-Seysenegg (1870–1952): Physiologist and Co-'Rediscoverer' of Mendel's laws. en. *Theory in Biosciences* **136,** 59–67. doi:10.1007/s12064-016-0236-8 (2017).

13. Visscher, P. M. Commentary: Height and Mendel's theory: the long and the short of it. *International Journal of Epidemiology* **42,** 944–945. doi:10.1093/ije/dyt069 (2013).

14. Visscher, P. M. & Goddard, M. E. From R.A. Fisher's 1918 Paper to GWAS a Century Later. *Genetics* **211,** 1125–1130. doi:10.1534/genetics.118.301594 (2019).

15. Galton, F. Regression Towards Mediocrity in Hereditary Stature. *The Journal of the Anthropological Institute of Great Britain and Ireland* **15.** Publisher: [Royal Anthropological Institute of Great Britain and Ireland, Wiley], 246–263. doi:10.2307/2841583 (1886).

16. Pearson, K. LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* **2.** Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/14786440109462720, 559–572. doi:10.1080/14786440109462720 (1901).

17. Yule, G. U. Mendel's Laws and Their Probable Relations to Intra-Racial Heredity (Continued). *The New Phytologist* **1.** Publisher: [Wiley, New Phytologist Trust], 222–238 (1902).

18. Provine, W. B. *The Origins of Theoretical Population Genetics: With a New Afterword* en. Google-Books-ID: 8Mj0ptluGeQC (University of Chicago Press, 2001).

19. Fisher, R. A. The Correlation between Relatives on the Supposition of Mendelian Inheritance. en. *Earth and Environmental Science Transactions of The Royal Society of Edinburgh* **52,** 399–433. doi:10.1017/S0080456800012163 (1919).

20. Visscher, P. M., Brown, M. A., McCarthy, M. I. & Yang, J. Five Years of GWAS Discovery. *American Journal of Human Genetics* **90,** 7–24. doi:10.1016/j.ajhg.2011.11.029 (2012).

21. Shine, I. & Wrobel, S. *Thomas Hunt Morgan: Pioneer of Genetics* en. Google-Books-ID: 1rUeBgAAQBAJ (University Press of Kentucky, 2014).

22. Hegreness, M. & Meselson, M. What Did Sutton See?: Thirty Years of Confusion Over the Chromosomal Basis of Mendelism. *Genetics* **176,** 1939–1944. doi:10.1534/genetics.104.79723 (2007).

23. Morgan, T. H. Sex Limited Inheritance in Drosophila. *Science* **32.** Publisher: American Association for the Advancement of Science, 120–122. doi:10.1126/science.32.812.120 (1910).

24. Sturtevant, A. H. The linear arrangement of six sex-linked factors in Drosophila, as shown by their mode of association. en. *Journal of Experimental Zoology* **14.** _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jez.1400140104, 43–59. doi:10.1002/jez.1400140104 (1913).

25. Pääbo, S. The mosaic that is our genome. en. *Nature* **421.** Publisher: Nature Publishing Group, 409–412. doi:10.1038/nature01400 (2003).

26. Gibbs, R. A. *et al.* The International HapMap Project. en. *Nature* **426.** Publisher: Nature Publishing Group, 789–796. doi:10.1038/nature02168 (2003).

27. Fisher, R. A. & Russell, E. J. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* **222.** Publisher: Royal Society, 309–368. doi:10.1098/rsta.1922.0009 (1922).

28. Smith, C. a. B. Some simple methods for linkage analysis. en. *Annals of Human Genetics* **53.** _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1469-1809.1989.tb01795.x, 277–283. doi:10.1111/j.1469-1809.1989.tb01795.x (1989).

29. Feder, J. N. *et al.* A novel MHC class I–like gene is mutated in patients with hereditary haemochromatosis. en. *Nature Genetics* **13.** Publisher: Nature Publishing Group, 399–408. doi:10.1038/ng0896-399 (1996).

30. Botstein, D. & Risch, N. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. en. *Nature Genetics* **33,** 228–237. doi:10.1038/ng1090 (2003).

31. Lander, E. S. & Botstein, D. Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* **121,** 185–199. doi:10.1093/genetics/121.1.185 (1989).

32. Risch, N. *et al.* A Genomic Screen of Autism: Evidence for a Multilocus Etiology. English. *The American Journal of Human Genetics* **65.** Publisher: Elsevier, 493–507. doi:10.1086/302497 (1999).

33. Yengo, L. *et al.* A saturated map of common genetic variants associated with human height. en. *Nature* **610.** Publisher: Nature Publishing Group, 704–712. doi:10.1038/s41586-022-05275-y (2022).

34. Watson, J. D. & Crick, F. H. C. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. en. *Nature* **171.** Publisher: Nature Publishing Group, 737–738. doi:10.1038/171737a0 (1953).

35. Brenner, S., Jacob, F. & Meselson, M. An Unstable Intermediate Carrying Information from Genes to Ribosomes for Protein Synthesis. en. *Nature* **190.** Publisher: Nature Publishing Group, 576–581. doi:10.1038/190576a0 (1961).

36. Sanger, F. & Coulson, A. R. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. eng. *Journal of Molecular Biology* **94,** 441–448. doi:10.1016/0022-2836(75)90213-2 (1975).

37. Anderson, S. *et al.* Sequence and organization of the human mitochondrial genome. en. *Nature* **290.** Publisher: Nature Publishing Group, 457–465. doi:10.1038/290457a0 (1981).

38. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. en. *Nature* **409.** Publisher: Nature Publishing Group, 860–921. doi:10.1038/35057062 (2001).

39. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. en. *Nature* **431.** Publisher: Nature Publishing Group, 931–945. doi:10.1038/nature03001 (2004).

40. Frazer, K. A. *et al.* A second generation human haplotype map of over 3.1 million SNPs. en. *Nature* **449.** Publisher: Nature Publishing Group, 851–861. doi:10.1038/nature06258 (2007).

41. Altshuler, D. M. *et al.* Integrating common and rare genetic variation in diverse human populations. en. *Nature* **467.** Publisher: Nature Publishing Group, 52–58. doi:10.1038/nature09298 (2010).

42. (2024). The Cost of Sequencing a Human Genome. URL: https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost.

43. Durbin, R. M. *et al.* A map of human genome variation from population-scale sequencing. en. *Nature* **467.** Publisher: Nature Publishing Group, 1061–1073. doi:10.1038/nature09534 (2010).

44. Altshuler, D., Donnelly, P. & The International HapMap Consortium. A haplotype map of the human genome. en. *Nature* **437.** Publisher: Nature Publishing Group, 1299–1320. doi:10.1038/nature04226 (2005).

45. Buchanan, C. C., Torstenson, E. S., Bush, W. S. & Ritchie, M. D. A comparison of cataloged variation between International HapMap Consortium and 1000 Genomes Project data. *Journal of the American Medical Informatics Association* **19,** 289–294. doi:10.1136/amiajnl-2011-000652 (2012).

46. Neale, B. M. Introduction to Linkage Disequilibrium, the HapMap, and Imputation. en. *Cold Spring Harbor Protocols* **2010.** Publisher: Cold Spring Harbor Laboratory Press, pdb.top74. doi:10.1101/pdb.top74 (2010).

47. Pei, Y.-F., Zhang, L., Li, J. & Deng, H.-W. Analyses and Comparison of Imputation-Based Association Methods. en. *PLOS ONE* **5.** Publisher: Public Library of Science, e10827. doi:10.1371/journal.pone.0010827 (2010).

48. Gao, X. *et al.* Genotype Imputation for Latinos Using the HapMap and 1000 Genomes Project Reference Panels. English. *Frontiers in Genetics* **3.** Publisher: Frontiers. doi:10.3389/fgene.2012.00117 (2012).

49. Spencer, C. C. A., Su, Z., Donnelly, P. & Marchini, J. Designing Genome-Wide Association Studies: Sample Size, Power, Imputation, and the Choice of Genotyping Chip. en. *PLOS Genetics* **5.** Publisher: Public Library of Science, e1000477. doi:10.1371/journal.pgen.1000477 (2009).

50. Nothnagel, M., Ellinghaus, D., Schreiber, S., Krawczak, M. & Franke, A. A comprehensive evaluation of SNP genotype imputation. en. *Human Genetics* **125,** 163–171. doi:10.1007/s00439-008-0606-5 (2009).

51. Waksmunski, A. R., Main, L. R. & Haines, J. L. in *Genetics and Genomics of Eye Disease* (ed Gao, X. R.) 7–23 (Academic Press, 2020). doi:10.1016/B978-0-12-816222-4.00002-2.

52. Devlin, B. & Roeder, K. Genomic Control for Association Studies. *Biometrics* **55,** 997–1004. doi:10.1111/j.0006-341X.1999.00997.x (1999).

53. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. **447.** Publisher: Nature Publishing Group, 661–678. doi:10.1038/nature05911 (2007).

54. Yu, J. *et al.* A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. en. *Nature Genetics* **38.** Publisher: Nature Publishing Group, 203–208. doi:10.1038/ng1702 (2006).

55. Kang, H. M. *et al.* Efficient Control of Population Structure in Model Organism Association Mapping. *Genetics* **178,** 1709–1723. doi:10.1534/genetics.107.080101 (2008).

56. Xue, A. *et al.* Genome-wide association analyses identify 143 risk variants and putative regulatory mechanisms for type 2 diabetes. en. *Nature Communications* **9.** Publisher: Nature Publishing Group, 2941. doi:10.1038/s41467-018-04951-w (2018).

57. Ripke, S. *et al.* Biological insights from 108 schizophrenia-associated genetic loci. en. *Nature* **511.** Publisher: Nature Publishing Group, 421–427. doi:10.1038/nature13595 (2014).

58. Sud, A., Kinnersley, B. & Houlston, R. S. Genome-wide association studies of cancer: current insights and future perspectives. en. *Nature Reviews Cancer* **17.** Publisher: Nature Publishing Group, 692–704. doi:10.1038/nrc.2017.82 (2017).

59. Ganna, A. *et al.* Large-scale GWAS reveals insights into the genetic architecture of same-sex sexual behavior. *Science* **365.** Publisher: American Association for the Advancement of Science, eaat7693. doi:10.1126/science.aat7693 (2019).

60. Venkatesh, S. S. *et al.,* (2024). Genome-wide analyses identify 21 infertility loci and over 400 reproductive hormone loci across the allele frequency spectrum. URL: https://www.medrxiv.org/content/10.1101/2024.03.19.24304530v1.

61. Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Research* **42,** D1001–D1006. doi:10.1093/nar/gkt1229 (2014).

62. Aragam, K. G. *et al.* Discovery and systematic characterization of risk variants and genes for coronary artery disease in over a million participants. en. *Nature Genetics* **54.** Publisher: Nature Publishing Group, 1803–1815. doi:10.1038/s41588-022-01233-6 (2022).

63. Patel, A. P. *et al.* A multi-ancestry polygenic risk score improves risk prediction for coronary artery disease. en. *Nature Medicine* **29.** Publisher: Nature Publishing Group, 1793–1803. doi:10.1038/s41591-023-02429-x (2023).

64. Lee, S. H., Yang, J., Goddard, M. E., Visscher, P. M. & Wray, N. R. Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. eng. *Bioinformatics (Oxford, England)* **28,** 2540–2542. doi:10.1093/bioinformatics/bts474 (2012).

65. Visscher, P. M. *et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation. English. *The American Journal of Human Genetics* **101.** Publisher: Elsevier, 5–22. doi:10.1016/j.ajhg.2017.06.005 (2017).

66. Van Rheenen, W., Peyrot, W. J., Schork, A. J., Lee, S. H. & Wray, N. R. Genetic correlations of polygenic disease traits: from theory to practice. en. *Nature Reviews Genetics* **20.** Publisher: Nature Publishing Group, 567–581. doi:10.1038/s41576-019-0137-z (2019).

67. Pingault, J.-B. *et al.* Using genetic data to strengthen causal inference in observational research. en. *Nature Reviews Genetics* **19.** Publisher: Nature Publishing Group, 566–580. doi:10.1038/s41576-018-0020-3 (2018).

68. Davey Smith, G. & Hemani, G. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Human Molecular Genetics* **23,** R89–R98. doi:10.1093/hmg/ddu328 (2014).

69. Fernández-Marmiesse, A., Gouveia, S. & Couce, M. L. NGS Technologies as a Turning Point in Rare Disease Research, Diagnosis and Treatment. *Current Medicinal Chemistry* **25,** 404–432. doi:10.2174/0929867324666170718101946 (2018).

70. Zuk, O. *et al.* Searching for missing heritability: Designing rare variant association studies. *Proceedings of the National Academy of Sciences* **111.** Publisher: Proceedings of the National Academy of Sciences, E455–E464. doi:10.1073/pnas.1322563111 (2014).

71. Kosmicki, J. A., Churchhouse, C. L., Rivas, M. A. & Neale, B. M. Discovery of rare variants for complex phenotypes. en. *Human Genetics* **135,** 625–634. doi:10.1007/s00439-016-1679-1 (2016).

72. Singh, T. *et al.* Rare coding variants in ten genes confer substantial risk for schizophrenia. en. *Nature* **604.** Publisher: Nature Publishing Group, 509–516. doi:10.1038/s41586-022-04556-w (2022).

73. Gibson, G. Rare and common variants: twenty arguments. en. *Nature Reviews Genetics* **13.** Publisher: Nature Publishing Group, 135–145. doi:10.1038/nrg3118 (2012).

74. Hinney, A., Vogel, C. I. G. & Hebebrand, J. From monogenic to polygenic obesity: recent advances. en. *European Child & Adolescent Psychiatry* **19,** 297–310. doi:10.1007/s00787-010-0096-6 (2010).

75. Warr, A. *et al.* Exome Sequencing: Current and Future Perspectives. *G3 Genes|Genomes|Genetics* **5,** 1543–1550. doi:10.1534/g3.115.018564 (2015).

76. Aspden, J. L., Wallace, E. W. J. & Whiffin, N. Not all exons are protein coding: Addressing a common misconception. English. *Cell Genomics* **3.** Publisher: Elsevier. doi:10.1016/j.xgen.2023.100296 (2023).

77. Singleton, A. B. Exome sequencing: a transformative technology. English. *The Lancet Neurology* **10.** Publisher: Elsevier, 942–946. doi:10.1016/S1474-4422(11)70196-X (2011).

78. Nelson, M. R. *et al.* The support of human genetic evidence for approved drug indications. en. *Nature Genetics* **47,** 856–860. doi:10.1038/ng.3314 (2015).

79. Bansal, V., Libiger, O., Torkamani, A. & Schork, N. J. Statistical analysis strategies for association studies involving rare variants. en. *Nature Reviews Genetics* **11.** Publisher: Nature Publishing Group, 773–785. doi:10.1038/nrg2867 (2010).

80. Povysil, G. *et al.* Rare-variant collapsing analyses for complex traits: guidelines and applications. en. *Nature Reviews Genetics* **20,** 747–759. doi:10.1038/s41576-019-0177-4 (2019).

81. Cirulli, E. T. *et al.* Genome-wide rare variant analysis for thousands of phenotypes in over 70,000 exomes from two cohorts. en. *Nature Communications* **11.** Publisher: Nature Publishing Group, 542. doi:10.1038/s41467-020-14288-y (2020).

82. Cheng, J. *et al.* Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science* **381.** Publisher: American Association for the Advancement of Science, eadg7492. doi:10.1126/science.adg7492 (2023).

83. Parry, D. A. *et al.,* (2024). PrimateAI-3D outperforms AlphaMissense in real-world cohorts. URL: https://www.medrxiv.org/content/10.1101/2024.01.12.24301193v1.

84. Jaganathan, K. *et al.* Predicting Splicing from Primary Sequence with Deep Learning. English. *Cell* **176.** Publisher: Elsevier, 535–548.e24. doi:10.1016/j.cell.2018.12.015 (2019).

85. Mbatchou, J. *et al.* Computationally efficient whole-genome regression for quantitative and binary traits. en. *Nature Genetics* **53.** Publisher: Nature Publishing Group, 1097–1103. doi:10.1038/s41588-021-00870-7 (2021).

86. Zhou, W. *et al.* SAIGE-GENE+ improves the efficiency and accuracy of set-based rare variant association tests. en. *Nature Genetics* **54.** Publisher: Nature Publishing Group, 1466–1469. doi:10.1038/s41588-022-01178-w (2022).

87. Backman, J. D. *et al.* Exome sequencing and analysis of 454,787 UK Biobank participants. en. *Nature* **599.** Publisher: Nature Publishing Group, 628–634. doi:10.1038/s41586-021-04103-z (2021).

88. Halldorsson, B. V. *et al.* The sequences of 150,119 genomes in the UK Biobank. en. *Nature* **607.** Publisher: Nature Publishing Group, 732–740. doi:10.1038/s41586-022-04965-x (2022).

89. Elliott, L. T. *et al.* Genome-wide association studies of brain imaging phenotypes in UK Biobank. en. *Nature* **562.** Publisher: Nature Publishing Group, 210–216. doi:10.1038/s41586-018-0571-7 (2018).

90. Sun, B. B. *et al.* Plasma proteomic associations with genetics and health in the UK Biobank. en. *Nature* **622.** Publisher: Nature Publishing Group, 329–338. doi:10.1038/s41586-023-06592-6 (2023).

91. Julkunen, H. *et al.* Atlas of plasma NMR biomarkers for health and disease in 118,461 individuals from the UK Biobank. en. *Nature Communications* **14.** Publisher: Nature Publishing Group, 604. doi:10.1038/s41467-023-36231-7 (2023).

92. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. en. *Nature* **562.** Publisher: Nature Publishing Group, 203–209. doi:10.1038/s41586-018-0579-z (2018).

93. Venkatesh, S. S. *et al.* Characterising the genetic architecture of changes in adiposity during adulthood using electronic health records. en. *Nature Communications* **15.** Publisher: Nature Publishing Group, 5801. doi:10.1038/s41467-024-49998-0 (2024).

94. 100,000 Genomes Project Pilot Investigators *et al.* 100,000 Genomes Pilot on Rare-Disease Diagnosis in Health Care - Preliminary Report. en. *N. Engl. J. Med.* **385,** 1868–1880 (2021).

95. Gaziano, J. M. *et al.* Million Veteran Program: A mega-biobank to study genetic influences on health and disease. English. *Journal of Clinical Epidemiology* **70.** Publisher: Elsevier, 214–223. doi:10.1016/j.jclinepi.2015.09.016 (2016).

96. Johnson, R. *et al.* Leveraging genomic diversity for discovery in an electronic health record linked biobank: the UCLA ATLAS Community Health Initiative. *Genome Medicine* **14,** 104. doi:10.1186/s13073-022-01106-x (2022).

97. Boutin, N. T. *et al.* The Evolution of a Large Biobank at Mass General Brigham. en. *Journal of Personalized Medicine* **12.** Number: 8 Publisher: Multidisciplinary Digital Publishing Institute, 1323. doi:10.3390/jpm12081323 (2022).

98. Zawistowski, M. *et al.* The Michigan Genomics Initiative: A biobank linking genotypes and electronic clinical records in Michigan Medicine patients. English. *Cell Genomics* **3.** Publisher: Elsevier. doi:10.1016/j.xgen.2023.100257 (2023).

99. Bick, A. G. *et al.* Genomic data in the All of Us Research Program. en. *Nature* **627.** Publisher: Nature Publishing Group, 340–346. doi:10.1038/s41586-023-06957-x (2024).

100. Chen, Z. *et al.* China Kadoorie Biobank of 0.5 million people: survey methods, baseline characteristics and long-term follow-up. *International Journal of Epidemiology* **40,** 1652–1666. doi:10.1093/ije/dyr120 (2011).

101. Sohail, M. *et al.* Mexican Biobank advances population and medical genomics of diverse ancestries. en. *Nature* **622.** Publisher: Nature Publishing Group, 775–783. doi:10.1038/s41586-023-06560-0 (2023).

102. Kirsh, V. A. *et al.* Cohort Profile: The Ontario Health Study (OHS). *International Journal of Epidemiology* **52,** e137–e151. doi:10.1093/ije/dyac156 (2023).

103. Hansen, T. F. *et al.* DBDS Genomic Cohort, a prospective and comprehensive resource for integrative and temporal analysis of genetic, environmental and lifestyle factors affecting health of blood donors. en. *BMJ Open* **9.** Publisher: British Medical Journal Publishing Group Section: Genetics and genomics, e028401. doi:10.1136/bmjopen-2018-028401 (2019).

104. Laursen, I. H. *et al.* Cohort profile: Copenhagen Hospital Biobank - Cardiovascular Disease Cohort (CHB-CVDC): Construction of a large-scale genetic cohort to facilitate a better understanding of heart diseases. en. *BMJ Open* **11.** Publisher: British Medical Journal Publishing Group Section: Genetics and genomics, e049709. doi:10.1136/bmjopen-2021-049709 (2021).

105. Kurki, M. I. *et al.* FinnGen provides genetic insights from a well-phenotyped isolated population. en. *Nature* **613.** Publisher: Nature Publishing Group, 508–518. doi:10.1038/s41586-022-05473-8 (2023).

106. Åsvold, B. O. *et al.* Cohort Profile Update: The HUNT Study, Norway. *International Journal of Epidemiology* **52,** e80–e91. doi:10.1093/ije/dyac095 (2023).

107. Leitsalu, L. *et al.* Cohort Profile: Estonian Biobank of the Estonian Genome Center, University of Tartu. *International Journal of Epidemiology* **44,** 1137–1147. doi:10.1093/ije/dyt268 (2015).

108. Hakonarson, H., Gulcher, J. R. & Stefansson, K. deCODE genetics, Inc. *Pharmacogenomics* **4.** Publisher: Taylor & Francis _eprint: https://doi.org/10.1517/phgs.4.2.209.22627, 209–215. doi:10.1517/phgs.4.2.209.22627 (2003).

109. Al Thani, A. *et al.* Qatar Biobank Cohort Study: Study Design and First Results. *American Journal of Epidemiology* **188,** 1420–1433. doi:10.1093/aje/kwz084 (2019).

110. Nagai, A. *et al.* Overview of the BioBank Japan Project: Study design and profile. *Journal of Epidemiology. SI: Overview of BBJ cohort* **27,** S2–S8. doi:10.1016/j.je.2016.12.005 (2017).

111. Cho, S. Y. *et al.* Opening of the National Biobank of Korea as the Infrastructure of Future Biomedical Science in Korea. *Osong Public Health and Research Perspectives* **3,** 177–184. doi:10.1016/j.phrp.2012.07.004 (2012).

112. Sijtsma, A. *et al.* Cohort Profile Update: Lifelines, a three-generation cohort study and biobank. *International Journal of Epidemiology* **51,** e295–e302. doi:10.1093/ije/dyab257 (2022).

113. Scholtens, S. *et al.* Cohort Profile: LifeLines, a three-generation cohort study and biobank. *International Journal of Epidemiology* **44,** 1172–1180. doi:10.1093/ije/dyu229 (2015).

114. Nsubuga, G. *et al.* Biobanking in East and Central Africa: A case of the Integrated Biorepository of H3Africa Uganda. *Open Research Africa* **5,** 33. doi:10.12688/openresafrica.13495.1 (2022).

115. Siegal, G. Genomic Databases and Biobanks in Israel. en. *Journal of Law, Medicine & Ethics* **43,** 766–775. doi:10.1111/jlme.12318 (2015).

116. Fan, C.-T., Lin, J.-C. & Lee, C.-H. Taiwan Biobank: A Project Aiming to Aid Taiwan's Transition into a Biomedical Island. *Pharmacogenomics* **9.** Publisher: Taylor & Francis _eprint: https://doi.org/10.2217/14622416.9.2.235, 235–246. doi:10.2217/14622416.9.2.235 (2008).

117. Olsen, C. M. *et al.* Cohort profile: The QSkin Sun and Health Study. *International Journal of Epidemiology* **41,** 929–929i. doi:10.1093/ije/dys107 (2012).

118. Hawkes, N. Genetic project calls for 100 000 volunteers from Pakistani and Bangladeshi communities in east London. en. *BMJ* **350.** Publisher: British Medical Journal Publishing Group Section: News, h1415. doi:10.1136/bmj.h1415 (2015).

119. Fry, A. *et al.* Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. *American Journal of Epidemiology* **186,** 1026–1034. doi:10.1093/aje/kwx246 (2017).

120. Schoeler, T. *et al.* Participation bias in the UK Biobank distorts genetic associations and downstream analyses. en. *Nature Human Behaviour* **7.** Publisher: Nature Publishing Group, 1216–1227. doi:10.1038/s41562-023-01579-9 (2023).

121. Van Alten, S., Domingue, B. W., Faul, J., Galama, T. & Marees, A. T. Reweighting UK Biobank corrects for pervasive selection bias due to volunteering. *International Journal of Epidemiology* **53,** dyae054. doi:10.1093/ije/dyae054 (2024).

122. Pirastu, N. *et al.* Genetic analyses identify widespread sex-differential participation bias. en. *Nature Genetics* **53.** Publisher: Nature Publishing Group, 663–671. doi:10.1038/s41588-021-00846-7 (2021).

123. Zhou, W. *et al.* Global Biobank Meta-analysis Initiative: Powering genetic discovery across human disease. English. *Cell Genomics* **2.** Publisher: Elsevier. doi:10.1016/j.xgen.2022.100192 (2022).

124. Jurgens, S. J. *et al.* Rare coding variant analysis for human diseases across biobanks and ancestries. en. *Nature Genetics.* Publisher: Nature Publishing Group, 1–10. doi:10.1038/s41588-024-01894-5 (2024).

125. Koyama, S. *et al.,* (2024). Exome wide association study for blood lipids in 1,158,017 individuals from diverse populations. URL: http://medrxiv.org/lookup/doi/10.1101/2024.09.17.24313718.

126. Pietzner, M. *et al.* Synergistic insights into human health from aptamer- and antibody-based proteomic profiling. en. *Nature Communications* **12.** Publisher: Nature Publishing Group, 6822. doi:10.1038/s41467-021-27164-0 (2021).

127. Rooney, M. R. *et al.,* (2024). Plasma proteomic comparisons change as coverage expands for SomaLogic and Olink. URL: https://www.medrxiv.org/content/10.1101/2024.07.11.24310161v1.

128. Dhindsa, R. S. *et al.* Rare variant associations with plasma protein levels in the UK Biobank. en. *Nature* **622.** Publisher: Nature Publishing Group, 339–347. doi:10.1038/s41586-023-06547-x (2023).

129. Nesvizhskii, A. I. Proteogenomics: concepts, applications and computational strategies. en. *Nature Methods* **11.** Publisher: Nature Publishing Group, 1114–1125. doi:10.1038/nmeth.3144 (2014).

130. Ferkingstad, E. *et al.* Large-scale integration of the plasma proteome with genetics and disease. en. *Nature Genetics* **53.** Publisher: Nature Publishing Group, 1712–1721. doi:10.1038/s41588-021-00978-w (2021).

131. You, J. *et al.* Plasma proteomic profiles predict individual future health risk. en. *Nature Communications* **14.** Publisher: Nature Publishing Group, 7817. doi:10.1038/s41467-023-43575-7 (2023).

132. Gadd, D. A. *et al.* Blood protein assessment of leading incident diseases and mortality in the UK Biobank. en. *Nature Aging* **4.** Publisher: Nature Publishing Group, 939–948. doi:10.1038/s43587-024-00655-7 (2024).

133. Marchini, J. *et al.* A Comparison of Phasing Algorithms for Trios and Unrelated Individuals. English. *The American Journal of Human Genetics* **78.** Publisher: Elsevier, 437–450. doi:10.1086/500808 (2006).

134. Miller, D. B. & Piccolo, S. R. trioPhaser: using Mendelian inheritance logic to improve genomic phasing of trios. en. *BMC Bioinformatics* **22,** 559. doi:10.1186/s12859-021-04470-4 (2021).

135. Kojima, K. *et al. HapMonster: A Statistically Unified Approach for Variant Calling and Haplotyping Based on Phase-Informative Reads* en. in *Algorithms for Computational Biology* (eds Dediu, A.-H., Martín-Vide, C. & Truthe, B.) (Springer International Publishing, Cham, 2014), 107–118. doi:10.1007/978-3-319-07953-0_9.

136. Castel, S. E., Mohammadi, P., Chung, W. K., Shen, Y. & Lappalainen, T. Rare variant phasing and haplotypic expression from RNA sequencing with phASER. en. *Nature Communications* **7.** Publisher: Nature Publishing Group, 12817. doi:10.1038/ncomms12817 (2016).

137. Martin, M. *et al.,* (2016). WhatsHap: fast and accurate read-based phasing. URL: https://www.biorxiv.org/content/10.1101/085050v2.

138. Li, N. & Stephens, M. Modeling Linkage Disequilibrium and Identifying Recombination Hotspots Using Single-Nucleotide Polymorphism Data. *Genetics* **165,** 2213–2233. doi:10.1093/genetics/165.4.2213 (2003).

139. Browning, S. R. & Browning, B. L. Haplotype phasing: existing methods and new developments. en. *Nature Reviews Genetics* **12.** Publisher: Nature Publishing Group, 703–714. doi:10.1038/nrg3054 (2011).

140. Delaneau, O., Zagury, J.-F., Robinson, M. R., Marchini, J. L. & Dermitzakis, E. T. Accurate, scalable and integrative haplotype estimation. en. *Nature Communications* **10.** Number: 1 Publisher: Nature Publishing Group, 5436. doi:10.1038/s41467-019-13225-y (2019).

141. Hofmeister, R. J., Ribeiro, D. M., Rubinacci, S. & Delaneau, O. Accurate rare variant phasing of whole-genome and whole-exome sequencing data in the UK Biobank. en. *Nature Genetics* **55.** Publisher: Nature Publishing Group, 1243–1249. doi:10.1038/s41588-023-01415-w (2023).

142. Guo, M. H. *et al.,* (2023). Inferring compound heterozygosity from large-scale exome sequencing data. URL: https://www.biorxiv.org/content/10.1101/2023.03.19.533370v1.

143. Wertenbroek, R., Hofmeister, R. J., Xenarios, I., Thoma, Y. & Delaneau, O. Improving population scale statistical phasing with whole-genome sequencing data. en. *PLOS Genetics* **20.** Publisher: Public Library of Science, e1011092. doi:10.1371/journal.pgen.1011092 (2024).

144. Capecchi, M. R. Gene targeting in mice: functional analysis of the mammalian genome for the twenty-first century. en. *Nature Reviews Genetics* **6.** Publisher: Nature Publishing Group, 507–512. doi:10.1038/nrg1619 (2005).

145. Alkuraya, F. S. Human knockout research: new horizons and opportunities. *Trends in Genetics* **31,** 108–115. doi:10.1016/j.tig.2014.11.003 (2015).

146. Perdigoto, C. Dawn of the Human Knockout Project. en. *Nature Reviews Genetics* **18.** Publisher: Nature Publishing Group, 328–329. doi:10.1038/nrg.2017.35 (2017).

147. Alsalem, A. B., Halees, A. S., Anazi, S., Alshamekh, S. & Alkuraya, F. S. Autozygome Sequencing Expands the Horizon of Human Knockout Research and Provides Novel Insights into Human Phenotypic Variation. en. *PLOS Genetics* **9.** Publisher: Public Library of Science, e1004030. doi:10.1371/journal.pgen.1004030 (2013).

148. Robinson, J. G. *et al.* Effect of Evolocumab or Ezetimibe Added to Moderate- or High-Intensity Statin Therapy on LDL-C Lowering in Patients With Hypercholesterolemia: The LAPLACE-2 Randomized Clinical Trial. *JAMA* **311,** 1870–1883. doi:10.1001/jama.2014.4030 (2014).

149. Blom, D. J. *et al.* A 52-Week Placebo-Controlled Trial of Evolocumab in Hyperlipidemia. *New England Journal of Medicine* **370.** Publisher: Massachusetts Medical Society _eprint: https://www.nejm.org/doi/pdf/10.1056/NEJMoa1316222, 1809–1819. doi:10.1056/NEJMoa1316222 (2014).

150. Hütter, G. *et al.* Long-Term Control of HIV by CCR5 Delta32/Delta32 Stem-Cell Transplantation. *New England Journal of Medicine* **360.** Publisher: Massachusetts Medical Society _eprint: https://www.nejm.org/doi/pdf/10.1056/NEJMoa0802905, 692–698. doi:10.1056/NEJMoa0802905 (2009).

151. Mohamed, H. *et al.* Targeting CCR5 as a Component of an HIV-1 Therapeutic Strategy. English. *Frontiers in Immunology* **12.** Publisher: Frontiers. doi:10.3389/fimmu.2021.816515 (2022).

152. Lim, E. T. *et al.* Distribution and Medical Impact of Loss-of-Function Variants in the Finnish Founder Population. en. *PLoS Genetics* **10** (ed Cutler, D.) e1004494. doi:10.1371/journal.pgen.1004494 (2014).

153. Jørgensen, A. B., Frikke-Schmidt, R., Nordestgaard, B. G. & Tybjærg-Hansen, A. Loss-of-Function Mutations in APOC3 and Risk of Ischemic Vascular Disease. *New England Journal of Medicine* **371.** Publisher: Massachusetts Medical Society _eprint: https://www.nejm.org/doi/pdf/10.1056/NEJMoa1308027, 32–41. doi:10.1056/NEJMoa1308027 (2014).

154. Pedersen, C.-E. T. *et al.* The Effect of an Extreme and Prolonged Population Bottleneck on Patterns of Deleterious Variation: Insights from the Greenlandic Inuit. *Genetics* **205,** 787–801. doi:10.1534/genetics.116.193821 (2017).

155. Strauss, K. A. & Puffenberger, E. G. Genetics, Medicine, and the Plain People. en. *Annual Review of Genomics and Human Genetics* **10.** Publisher: Annual Reviews, 513–536. doi:10.1146/annurev-genom-082908-150040 (2009).

156. Tsukahara, A. *et al.* Compound heterozygosity for two novel mutations in a severe factor XI deficiency. en. *American Journal of Hematology* **73.** _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ajh.10378, 279–284. doi:10.1002/ajh.10378 (2003).

157. Hague, S. *et al.* Early-onset Parkinson's disease caused by a compound heterozygous DJ-1 mutation. en. *Annals of Neurology* **54.** _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ana.10663, 271–274. doi:10.1002/ana.10663 (2003).

158. Welch, K. O., Marin, R. S., Pandya, A. & Arnos, K. S. Compound heterozygosity for dominant and recessive GJB2 mutations: Effect on phenotype and review of the literature. en. *American Journal of Medical Genetics Part A* **143A.** _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ajmg.a.31701, 1567–1573. doi:10.1002/ajmg.a.31701 (2007).

159. Bonneau, D. *et al.* Early-onset Behr syndrome due to compound heterozygous mutations in OPA1. en. *Brain* **137,** e301–e301. doi:10.1093/brain/awu184 (2014).

160. Piscosquito, G. *et al.* Mutational mechanisms in -related neuropathy: compound heterozygosity for recessive and semidominant mutations. en. *Journal of the Peripheral Nervous System* **20.** _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/jns.12145, 380–386. doi:10.1111/jns.12145 (2015).

161. Brodsky, N. N. *et al.* Novel compound heterozygous variants in NHLRC2 in a patient with FINCA syndrome. en. *Journal of Human Genetics* **65.** Publisher: Nature Publishing Group, 911–915. doi:10.1038/s10038-020-0776-0 (2020).

162. Hof, A. E. v. *et al.* The industrial melanism mutation in British peppered moths is a transposable element. en. *Nature* **534.** Publisher: Nature Publishing Group, 102–105. doi:10.1038/nature17951 (2016).

163. Mayo, O. & Bürger, R. The Evolution of Dominance: A Theory Whose Time Has Passed? en. *Biological Reviews* **72.** _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1469-185X.1997.tb00011.x, 97–110. doi:10.1111/j.1469-185X.1997.tb00011.x (1997).

164. De Visser, J. A. G. M. *et al.* Perspective: Evolution and Detection of Genetic Robustness. en. *Evolution* **57.** _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.0014-3820.2003.tb00377.x, 1959–1972. doi:10.1111/j.0014-3820.2003.tb00377.x (2003).

165. Fisher, R. A. The Possible Modification of the Response of the Wild Type to Recurrent Mutations. *The American Naturalist* **62.** Publisher: The University of Chicago Press, 115–126. doi:10.1086/280193 (1928).

166. Wright, S. Fisher's Theory of Dominance. en. *The American Naturalist.* Publisher: Science Press. doi:10.1086/280260 (1929).

167. Kacser, H. & Burns, J. A. THE MOLECULAR BASIS OF DOMINANCE. *Genetics* **97,** 639–666. doi:10.1093/genetics/97.3-4.639 (1981).

168. Orr, H. A. A test of Fisher's theory of dominance. *Proceedings of the National Academy of Sciences of the United States of America* **88,** 11413–11415 (1991).

169. Haldane, J. B. S. A Note on Fisher's Theory of the Origin of Dominance, and on a Correlation between Dominance and Linkage. *The American Naturalist* **64.** Publisher: The University of Chicago Press, 87–90. doi:10.1086/280299 (1930).

170. Haldane, J. B. S. The theory of the evolution of dominance. en. *Journal of Genetics* **37,** 365–374. doi:10.1007/BF02982734 (1939).

171. Manna, F., Martin, G. & Lenormand, T. Fitness Landscapes: An Alternative Theory for the Dominance of Mutation. *Genetics* **189,** 923–937. doi:10.1534/genetics.111.132944 (2011).

172. Hurst, L. D. & Randerson, J. P. Dosage, Deletions and Dominance: Simple Models of the Evolution of Gene Expression. *Journal of Theoretical Biology* **205,** 641–647. doi:10.1006/jtbi.2000.2095 (2000).

173. Gemmell, N. J. & Slate, J. Heterozygote Advantage for Fecundity. en. *PLOS ONE* **1.** Publisher: Public Library of Science, e125. doi:10.1371/journal.pone.0000125 (2006).

174. Mäki-Tanila, A. & Hill, W. G. Influence of Gene Interaction on Complex Trait Variation with Multilocus Models. *Genetics* **198,** 355–367. doi:10.1534/genetics.114.165282 (2014).

175. Hill, W. & Mäki-Tanila, A. Expected influence of linkage disequilibrium on genetic variance caused by dominance and epistasis on quantitative traits. en. *Journal of Animal Breeding and Genetics* **132.** _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/jbg.12140, 176–186. doi:10.1111/jbg.12140 (2015).

176. Zhu, Z. *et al.* Dominance Genetic Variation Contributes Little to the Missing Heritability for Human Complex Traits. English. *The American Journal of Human Genetics* **96.** Publisher: Elsevier, 377–385. doi:10.1016/j.ajhg.2015.01.001 (2015).

177. Hill, W. G., Goddard, M. E. & Visscher, P. M. Data and Theory Point to Mainly Additive Genetic Variance for Complex Traits. en. *PLOS Genetics* **4.** Publisher: Public Library of Science, e1000008. doi:10.1371/journal.pgen.1000008 (2008).

178. Yengo, L., Wray, N. R. & Visscher, P. M. Extreme inbreeding in a European ancestry sample from the contemporary UK population. en. *Nat. Commun.* **10,** 3719 (2019).

179. Palmer, D. S. *et al.* Analysis of genetic dominance in the UK Biobank. *Science* **379.** Publisher: American Association for the Advancement of Science, 1341–1348. doi:10.1126/science.abn8455 (2023).

180. Wigginton, J. E., Cutler, D. J. & Abecasis, G. R. A Note on Exact Tests of Hardy-Weinberg Equilibrium. *The American Journal of Human Genetics* **76,** 887–893. doi:10.1086/429864 (2005).

181. Li, M. & Li, C. Assessing departure from Hardy-Weinberg equilibrium in the presence of disease association. en. *Genetic Epidemiology* **32.** _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/gepi.20335, 589–599. doi:10.1002/gepi.20335 (2008).

182. Graffelman, J. & Moreno, V. The mid p-value in exact tests for Hardy-Weinberg equilibrium. en. *Statistical Applications in Genetics and Molecular Biology* **12.** Publisher: De Gruyter, 433–448. doi:10.1515/sagmb-2012-0039 (2013).

183. Purcell, S. *et al.* PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. English. *The American Journal of Human Genetics* **81.** Publisher: Elsevier, 559–575. doi:10.1086/519795 (2007).

184. Plenge, R. M., Scolnick, E. M. & Altshuler, D. Validating therapeutic targets through human genetics. en. *Nature Reviews Drug Discovery* **12.** Number: 8 Publisher: Nature Publishing Group, 581–594. doi:10.1038/nrd4051 (2013).

185. Whiffin, N. *et al.* The effect of LRRK2 loss-of-function variants in humans. en. *Nature Medicine* **26,** 869–877. doi:10.1038/s41591-020-0893-5 (2020).

186. Tobert, J. A. Lovastatin and beyond: the history of the HMG-CoA reductase inhibitors. en. *Nature Reviews Drug Discovery* **2.** Number: 7 Publisher: Nature Publishing Group, 517–526. doi:10.1038/nrd1112 (2003).

187. Do, R. Q., Vogel, R. A. & Schwartz, G. G. PCSK9 Inhibitors: potential in cardiovascular therapeutics. eng. *Current Cardiology Reports* **15,** 345. doi:10.1007/s11886-012-0345-z (2013).

188. Minikel, E. V. *et al.* Evaluating drug targets through human loss-of-function genetic variation. en. *Nature* **581,** 459–464. doi:10.1038/s41586-020-2267-z (2020).

189. Van Hout, C. V. *et al.* Exome sequencing and characterization of 49,960 individuals in the UK Biobank. en. *Nature* **586,** 749–756. doi:10.1038/s41586-020-2853-0 (2020).

190. DeBoever, C. *et al.* Medical relevance of protein-truncating variants across 337,205 individuals in the UK Biobank study. eng. *Nature Communications* **9,** 1612. doi:10.1038/s41467-018-03910-9 (2018).

191. Sulem, P. *et al.* Identification of a large set of rare complete human knockouts. en. *Nature Genetics* **47.** Number: 5 Publisher: Nature Publishing Group, 448–452. doi:10.1038/ng.3243 (2015).

192. Heyne, H. O. *et al.* Mono- and biallelic variant effects on disease at biobank scale. eng. *Nature* **613,** 519–525. doi:10.1038/s41586-022-05420-7 (2023).

193. Saleheen, D. *et al.* Human knockouts and phenotypic analysis in a cohort with a high rate of consanguinity. en. *Nature* **544,** 235–239. doi:10.1038/nature22034 (2017).

194. De Rosa, M. *et al.* Evidence for a recessive inheritance of Turcot's syndrome caused by compound heterozygous mutations within the PMS2 gene. en. *Oncogene* **19,** 1719–1723. doi:10.1038/sj.onc.1203447 (2000).

195. Robinson, J. P. *et al.* Evidence for an Association between Compound Heterozygosity for Germ Line Mutations in the Hemochromatosis (*HFE*) Gene and Increased Risk of Colorectal Cancer. en. *Cancer Epidemiology, Biomarkers & Prevention* **14,** 1460–1463. doi:10.1158/1055-9965.EPI-04-0672 (2005).

196. Maffei, L. *et al.* A novel compound heterozygous mutation of the aromatase gene in an adult man: reinforced evidence on the relationship between congenital oestrogen deficiency, adiposity and the metabolic syndrome. en. *Clinical Endocrinology* **67,** 218–224. doi:10.1111/j.1365-2265.2007.02864.x (2007).

197. Wang, X.-H. *et al.* Identification of Novel Compound Heterozygous MYO15A Mutations in Two Chinese Families with Autosomal Recessive Nonsyndromic Hearing Loss. eng. *Neural Plasticity* **2021,** 9957712. doi:10.1155/2021/9957712 (2021).

198. Hofmeister, R. J., Ribeiro, D. M., Rubinacci, S. & Delaneau, O. *Accurate rare variant phasing of whole-genome and whole-exome sequencing data in the UK Biobank* en. 2022.

199. Maestri, S. *et al.* A Long-Read Sequencing Approach for Direct Haplotype Phasing in Clinical Settings. en. *International Journal of Molecular Sciences* **21.** Number: 23 Publisher: Multidisciplinary Digital Publishing Institute, 9177. doi:10.3390/ijms21239177 (2020).

200. Loh, P.-R. *et al.* Reference-based phasing using the Haplotype Reference Consortium panel. en. *Nature Genetics* **48.** Number: 11 Publisher: Nature Publishing Group, 1443–1448. doi:10.1038/ng.3679 (2016).

201. Barton, A. R., Sherman, M. A., Mukamel, R. E. & Loh, P.-R. Whole-exome imputation within UK Biobank powers rare coding variant association and fine-mapping analyses. en. *Nature Genetics* **53.** Number: 8 Publisher: Nature Publishing Group, 1260–1269. doi:10.1038/s41588-021-00892-1 (2021).

202. Wain, L. V. *et al.* Novel insights into the genetics of smoking behaviour, lung function, and chronic obstructive pulmonary disease (UK BiLEVE): a genetic association study in UK Biobank. *The Lancet Respiratory Medicine* **3,** 769–781. doi:10.1016/S2213-2600(15)00283-0 (2015).

203. Hail Team, (2022). Hail.

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

205. Li, H. Toward better understanding of artifacts in variant calling from high-coverage samples. en. *Bioinformatics* **30,** 2843–2851 (2014).

206. Van der Auwera, G. A. & O'Connor, B. D. *Genomics in the Cloud: Using Docker, GATK, and WDL in Terra* en ("O'Reilly Media, Inc.", 2020).

207. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. en. *Bioinformatics* **27,** 2987–2993 (2011).

208. Zhang, D., Dey, R. & Lee, S. Fast and robust ancestry prediction using principal component analysis. en. *Bioinformatics* **36,** 3439–3446 (2020).

209. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. eng. *GigaScience* **10,** giab008. doi:10.1093/gigascience/giab008 (2021).

210. Harrell Jr, F. E. & Harrell Jr, M. F. E. Package 'hmisc'. *CRAN2018* **2019,** 235–236 (2019).

211. Martin, M. *et al. WhatsHap: fast and accurate read-based phasing* en. 2016.

212. Kuan, V. *et al.* A chronological map of 308 physical and mental health conditions from 4 million individuals in the English National Health Service. English. *The Lancet Digital Health* **1.** Publisher: Elsevier, e63–e77. doi:10.1016/S2589-7500(19)30012-3 (2019).

213. Censin, J. C. *et al.* Causal relationships between obesity and the leading causes of death in women and men. en. *PLoS Genet.* **15,** e1008405 (2019).

214. Eastwood, S. V. *et al.* Algorithms for the Capture and Adjudication of Prevalent and Incident Diabetes in UK Biobank. en. *PLoS One* **11,** e0162388 (2016).

215. Udler, M. S., McCarthy, M. I., Florez, J. C. & Mahajan, A. Genetic Risk Scores for Diabetes Diagnosis and Precision Medicine. en. *Endocr. Rev.* **40,** 1500–1520 (2019).

216. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. en. *Genome Biol.* **17,** 122 (2016).

217. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. en. *Nature* **581,** 434–443 (2020).

218. Ioannidis, N. M. *et al.* REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. en. *Am. J. Hum. Genet.* **99,** 877–885 (2016).

219. Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. CADD: predicting the deleteriousness of variants throughout the human genome. en. *Nucleic Acids Res.* **47,** D886–D894 (2019).

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

221. Privé, F. *et al.* High-resolution portability of 245 polygenic scores when derived and applied in the same cohort. *medRxiv* (2021).

222. International HapMap 3 Consortium *et al.* Integrating common and rare genetic variation in diverse human populations. en. *Nature* **467,** 52–58 (2010).

223. Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. en. *Nat. Genet.* **47,** 291–295 (2015).

224. Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. en. *Nat. Genet.* **47,** 1228–1235 (2015).

225. Privé, F., Aschard, H., Ziyatdinov, A. & Blum, M. G. B. Efficient analysis of large-scale genome-wide data with two R packages: bigstatsr and bigsnpr. en. *Bioinformatics* **34,** 2781–2787 (2018).

226. Georgi, B., Voight, B. F. & Bućan, M. From mouse to human: evolutionary genomics analysis of human orthologs of essential genes. eng. *PLoS genetics* **9,** e1003484. doi:10.1371/journal.pgen.1003484 (2013).

227. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. en. *Nature* **581,** 434–443. doi:10.1038/s41586-020-2308-7 (2020).

228. Vinceti, A. *et al.* CoRe: a robustly benchmarked R package for identifying core-fitness genes in genome-wide pooled CRISPR-Cas9 screens. en. *BMC Genomics* **22,** 828 (2021).

229. Hart, T. *et al.* High-Resolution CRISPR Screens Reveal Fitness Genes and Genotype-Specific Cancer Liabilities. en. *Cell* **163,** 1515–1526. doi:10.1016/j.cell.2015.11.015 (2015).

230. Hart, T. *et al.* Evaluation and Design of Genome-Wide CRISPR/SpCas9 Knockout Screens. *G3 Genes|Genomes|Genetics* **7,** 2719–2727. doi:10.1534/g3.117.041277 (2017).

231. Gavish, A. *et al.* Hallmarks of transcriptional intratumour heterogeneity across a thousand tumours. en. *Nature* **618.** Number: 7965 Publisher: Nature Publishing Group, 598–606. doi:10.1038/s41586-023-06130-4 (2023).

232. Samocha, K. E., Kosmicki, J. A., Karczewski, K. J., *et al.* Regional missense constraint improves variant deleteriousness prediction. *BioRxiv* (2017).

233. Pirinen, M., (2023). GWAS 3: Statistical power. URL: https://www.mv.helsinki.fi/home/mjxpirin/GWAS_course/material/GWAS3.html.

234. Kaplan, E. L. & Meier, P. Nonparametric Estimation from Incomplete Observations. *J. Am. Stat. Assoc.* **53,** 457 (1958).

235. Therneau, T. M., until 2009), T. L. (S.->. p. a. R. m., Elizabeth, A. & Cynthia, C., (2024). survival: Survival Analysis. URL: https://cran.r-project.org/web/packages/survival/index.html.

236. Kassambara, A., Kosinski, M., Biecek, P. & Fabian, S., (2021). survminer: Drawing Survival Curves using 'ggplot2'. URL: https://cran.r-project.org/web/packages/survminer/index.html.

237. Delaneau, O., Zagury, J.-F., Robinson, M. R., Marchini, J. L. & Dermitzakis, E. T. Accurate, scalable and integrative haplotype estimation. en. *Nat. Commun.* **10,** 5436 (2019).

238. Ioannidis, N. M. *et al.* REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. *The American Journal of Human Genetics* **99,** 877–885. doi:10.1016/j.ajhg.2016.08.016 (2016).

239. Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A. & McKusick, V. A. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research* **33,** D514–D517. doi:10.1093/nar/gki033 (2005).

240. Zhou, W. *et al.* Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. en. *Nature Genetics* **50,** 1335–1341. doi:10.1038/s41588-018-0184-y (2018).

241. Privé, F., Arbel, J. & Vilhjálmsson, B. J. LDpred2: better, faster, stronger. en. *Bioinformatics* **36** (ed Schwartz, R.) 5424–5431. doi:10.1093/bioinformatics/btaa1029 (2021).

242. Jurgens, S. J. *et al.* Adjusting for common variant polygenic scores improves yield in rare variant association analyses. eng. *Nature Genetics* **55,** 544–548. doi:10.1038/s41588-023-01342-w (2023).

243. Kavec, M. J. *et al.* Oxidative Damage in Sporadic Colorectal Cancer: Molecular Mapping of Base Excision Repair Glycosylases MUTYH and hOGG1 in Colorectal Cancer Patients. *International Journal of Molecular Sciences* **23,** 5704. doi:10.3390/ijms23105704 (2022).

244. Rice, N. E. *et al.* Filaggrin gene mutations are associated with asthma and eczema in later life. en. *The Journal of allergy and clinical immunology* **122.** Publisher: NIH Public Access, 834. doi:10.1016/j.jaci.2008.07.027 (2008).

245. Palmer, C. N. A. *et al.* Common loss-of-function variants of the epidermal barrier protein filaggrin are a major predisposing factor for atopic dermatitis. en. *Nature Genetics* **38.** Number: 4 Publisher: Nature Publishing Group, 441–446. doi:10.1038/ng1767 (2006).

246. Carlsen, B. C. *et al.* Filaggrin compound heterozygous patients carry mutations in trans position. eng. *Experimental Dermatology* **22,** 572–575. doi:10.1111/exd.12199 (2013).

247. Riethmuller, C. *et al.* Filaggrin breakdown products determine corneocyte conformation in patients with atopic dermatitis. en. *Journal of Allergy and Clinical Immunology* **136,** 1573–1580.e2. doi:10.1016/j.jaci.2015.04.042 (2015).

248. Liu, X. *et al.* Novel USH2A compound heterozygous mutations cause RP/USH2 in a Chinese family. eng. *Molecular Vision* **16,** 454–461. doi:10.1167/3.9.454 (2010).

249. Finn, R. D. *et al.* Pfam: the protein families database. eng. *Nucleic Acids Research* **42,** D222–230. doi:10.1093/nar/gkt1223 (2014).

250. Obeidat, M. *et al.* Surfactant protein D is a causal risk factor for COPD: results of Mendelian randomisation. en. *European Respiratory Journal* **50.** Publisher: European Respiratory Society Section: Original articles. doi:10.1183/13993003.00657-2017 (2017).

251. Karczewski, K. J. *et al.* Systematic single-variant and gene-based association testing of thousands of phenotypes in 394,841 UK Biobank exomes. English. *Cell Genomics* **2.** Publisher: Elsevier. doi:10.1016/j.xgen.2022.100168 (2022).

252. Finer, S. *et al.* Cohort Profile: East London Genes & Health (ELGH), a community-based population genomics and health study in British Bangladeshi and British Pakistani people. *International Journal of Epidemiology* **49,** 20–21i. doi:10.1093/ije/dyz174 (2020).

253. Antonarakis, S. E. & Beckmann, J. S. Mendelian disorders deserve more attention. en. *Nature Reviews Genetics* **7.** Publisher: Nature Publishing Group, 277–282. doi:10.1038/nrg1826 (2006).

254. Barton, A. R., Hujoel, M. L. A., Mukamel, R. E., Sherman, M. A. & Loh, P.-R. A spectrum of recessiveness among Mendelian disease variants in UK Biobank. eng. *American Journal of Human Genetics* **109,** 1298–1307. doi:10.1016/j.ajhg.2022.05.008 (2022).

255. Pazokitoroudi, A., Chiu, A. M., Burch, K. S., Pasaniuc, B. & Sankararaman, S. Quantifying the contribution of dominance deviation effects to complex trait variation in biobank-scale data. English. *The American Journal of Human Genetics* **108.** Publisher: Elsevier, 799–808. doi:10.1016/j.ajhg.2021.03.018 (2021).

256. Hivert, V. *et al.* Estimation of non-additive genetic variance in human complex traits from a large sample of unrelated individuals. English. *The American Journal of Human Genetics* **108.** Publisher: Elsevier, 786–798. doi:10.1016/j.ajhg.2021.02.014 (2021).

257. Lettre, G., Lange, C. & Hirschhorn, J. N. Genetic model testing and statistical power in population-based association studies of quantitative traits. en. *Genetic Epidemiology* **31.** _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/gepi.20217, 358–362. doi:10.1002/gepi.20217 (2007).

258. Guindo-Martínez, M. *et al.* The impact of non-additive genetic associations on age-related complex diseases. en. *Nature Communications* **12.** Publisher: Nature Publishing Group, 2436. doi:10.1038/s41467-021-21952-4 (2021).

259. Heng, T. H. *et al.,* (2024). Widespread recessive effects on common diseases in a cohort of 44,000 British Pakistanis and Bangladeshis with high autozygosity. URL: https://www.medrxiv.org/content/10.1101/2024.04.03.24305256v1.

260. Lassen, F. H. *et al.* Exome-wide evidence of compound heterozygous effects across common phenotypes in the UK Biobank. English. *Cell Genomics* **4.** Publisher: Elsevier. doi:10.1016/j.xgen.2024.100602 (2024).

261. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. en. *Gigascience* **4,** 7 (2015).

262. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biology* **17,** 122. doi:10.1186/s13059-016-0974-4 (2016).

263. Liu, X., Li, C., Mou, C., Dong, Y. & Tu, Y. dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs. *Genome Medicine* **12,** 103. doi:10.1186/s13073-020-00803-9 (2020).

264. Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. CADD: predicting the deleteriousness of variants throughout the human genome. en. *Nucleic Acids Research* **47,** D886–D894. doi:10.1093/nar/gky1016 (2019).

265. Morales, J. *et al.* A joint NCBI and EMBL-EBI transcript set for clinical genomics and research. en. *Nature* **604.** Publisher: Nature Publishing Group, 310–315. doi:10.1038/s41586-022-04558-8 (2022).

266. Frankish, A. *et al.* GENCODE: reference annotation for the human and mouse genomes in 2023. *Nucleic Acids Research* **51,** D942–D949. doi:10.1093/nar/gkac1071 (2023).

267. Pejaver, V. *et al.* Calibration of computational tools for missense variant pathogenicity classification and ClinGen recommendations for PP3/BP4 criteria. *The American Journal of Human Genetics* **109,** 2163–2177. doi:10.1016/j.ajhg.2022.10.013 (2022).

268. Vitezica, Z. G., Varona, L. & Legarra, A. On the Additive and Dominant Variance and Covariance of Individuals Within the Genomic Selection Scope. *Genetics* **195,** 1223–1230. doi:10.1534/genetics.113.155176 (2013).

269. Sokal, R. R. & Rohlf, F. J. *Biometry: The Principles and Practice of Statistics in Biological Research* en (W. H. Freeman, 1981).

270. Zhou, W. *et al.* Scalable generalized linear mixed model for region-based association tests in large biobanks and cohorts. en. *Nature Genetics* **52,** 634–639. doi:10.1038/s41588-020-0621-6 (2020).

271. Wu, M. C. *et al.* Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test. English. *The American Journal of Human Genetics* **89.** Publisher: Elsevier, 82–93. doi:10.1016/j.ajhg.2011.05.029 (2011).

272. Lee, S., Wu, M. C. & Lin, X. Optimal tests for rare variant effects in sequencing association studies. *Biostatistics* **13,** 762–775. doi:10.1093/biostatistics/kxs014 (2012).

273. Moutsianas, L. *et al.* The Power of Gene-Based Rare Variant Methods to Detect Disease-Associated Variation and Test Hypotheses About Complex Disease. en. *PLOS Genetics* **11.** Publisher: Public Library of Science, e1005165. doi:10.1371/journal.pgen.1005165 (2015).

274. Braun, M., (2021). sparseMVN: Multivariate Normal Functions for Sparse Covariance and Precision Matrices. URL: https://cran.r-project.org/web/packages/sparseMVN/index.html.

275. Falconer, D. S. The inheritance of liability to certain diseases, estimated from the incidence among relatives. en. *Annals of Human Genetics* **29.** _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1469-1809.1965.tb00500.x, 51–76. doi:10.1111/j.1469-1809.1965.tb00500.x (1965).

276. Wik, L. *et al.* Proximity Extension Assay in Combination with Next-Generation Sequencing for High-throughput Proteome-wide Analysis. English. *Molecular & Cellular Proteomics* **20.** Publisher: Elsevier. doi:10.1016/j.mcpro.2021.100168 (2021).

277. Sun, B. B. *et al.,* (2022). Genetic regulation of the human plasma proteome in 54,306 UK Biobank participants. URL: https://www.biorxiv.org/content/10.1101/2022.06.17.496443v1.

278. Bloemendal, A. The Levene-Haldane Distribution. en (2015).

279. Guo, S. W. & Thompson, E. A. Performing the Exact Test of Hardy-Weinberg Proportion for Multiple Alleles. *Biometrics* **48.** Publisher: [Wiley, International Biometric Society], 361–372. doi:10.2307/2532296 (1992).

280. Rohlfs, R. V. & Weir, B. S. Distributions of Hardy–Weinberg Equilibrium Test Statistics. *Genetics* **180,** 1609–1616. doi:10.1534/genetics.108.088005 (2008).

281. Graffelman, J. & Weir, B. S. Testing for Hardy–Weinberg equilibrium at biallelic genetic markers on the X chromosome. en. *Heredity* **116.** Publisher: Nature Publishing Group, 558–568. doi:10.1038/hdy.2016.20 (2016).

282. Eldjarn, G. H. *et al.* Large-scale plasma proteomics comparisons through genetics and disease associations. en. *Nature* **622.** Publisher: Nature Publishing Group, 348–358. doi:10.1038/s41586-023-06563-x (2023).

283. Zhong, W. *et al.* Next generation plasma proteome profiling to monitor health and disease. en. *Nature Communications* **12.** Publisher: Nature Publishing Group, 2493. doi:10.1038/s41467-021-22767-z (2021).

284. Graffelman, J. & Weir, B. S. The transitivity of the Hardy–Weinberg law. *Forensic Science International: Genetics* **58,** 102680. doi:10.1016/j.fsigen.2022.102680 (2022).

285. Chen, B., Cole, J. W. & Grond-Ginsbach, C. Departure from Hardy Weinberg Equilibrium and Genotyping Error. English. *Frontiers in Genetics* **8.** Publisher: Frontiers. doi:10.3389/fgene.2017.00167 (2017).

286. Kuonen, D. Saddlepoint Approximations for Distributions of Quadratic Forms in Normal Variables. *Biometrika* **86.** Publisher: [Oxford University Press, Biometrika Trust], 929–935 (1999).

287. Horowitz, J. E. *et al.* Mutation spectrum of NOD2 reveals recessive inheritance as a main driver of Early Onset Crohn's Disease. en. *Scientific Reports* **11.** Publisher: Nature Publishing Group, 5595. doi:10.1038/s41598-021-84938-8 (2021).

288. Ashton, J. J., Seaby, E. G., Beattie, R. M. & Ennis, S. NOD2 in Crohn's Disease—Unfinished Business. *Journal of Crohn's & Colitis* **17,** 450–458. doi:10.1093/ecco-jcc/jjac124 (2022).

289. Tillgren, V., Ho, J. C. S., Önnerfjord, P. & Kalamajski, S. The Novel Small Leucine-rich Protein Chondroadherin-like (CHADL) Is Expressed in Cartilage and Modulates Chondrocyte Differentiation *. English. *Journal of Biological Chemistry* **290.** Publisher: Elsevier, 918–925. doi:10.1074/jbc.M114.593541 (2015).

290. Styrkarsdottir, U. *et al.* Whole-genome sequencing identifies rare genotypes in COMP and CHADL associated with high risk of hip osteoarthritis. en. *Nature Genetics* **49.** Publisher: Nature Publishing Group, 801–805. doi:10.1038/ng.3816 (2017).

291. Osawa, R., Akiyama, M. & Shimizu, H. Filaggrin Gene Defects and the Risk of Developing Allergic Disorders. en. *Allergology International* **60,** 1–9. doi:10.2332/allergolint.10-RAI-0270 (2011).

292. Shen, H. *et al.* The Human Knockout Gene CLYBL Connects Itaconate to Vitamin B12. English. *Cell* **171.** Publisher: Elsevier, 771–782.e11. doi:10.1016/j.cell.2017.09.051 (2017).

293. Yu, B. *et al.* Association of Rare Loss-Of-Function Alleles in HAL, Serum Histidine Levels and Incident Coronary Heart Disease. *Circulation. Cardiovascular genetics* **8,** 351–355. doi:10.1161/CIRCGENETICS.114.000697 (2015).

294. Holmen, O. L. *et al.* Systematic evaluation of coding variation identifies a candidate causal variant in TM6SF2 influencing total cholesterol and myocardial infarction risk. en. *Nature Genetics* **46.** Publisher: Nature Publishing Group, 345–351. doi:10.1038/ng.2926 (2014).

295. Carlson, J. C. *et al.* A stop-gain variant in BTNL9 is associated with atherogenic lipid profiles. English. *Human Genetics and Genomics Advances* **4.** Publisher: Elsevier. doi:10.1016/j.xhgg.2022.100155 (2023).

296. Scherer, N., Sekula, P., Pfaffelhuber, P. & Schlosser, P. pgainsim: an R-package to assess the mode of inheritance for quantitative trait loci in GWAS. *Bioinformatics* **37,** 3061–3063. doi:10.1093/bioinformatics/btab150 (2021).

297. Georgeson, P. *et al.* Identifying colorectal cancer caused by biallelic MUTYH pathogenic variants using tumor mutational signatures. en. *Nature Communications* **13.** Publisher: Nature Publishing Group, 3254. doi:10.1038/s41467-022-30916-1 (2022).

298. Beach, R. *et al.* GJB2 Mutations Linked to Hearing Loss Exhibit Differential Trafficking and Functional Defects as Revealed in Cochlear-Relevant Cells. English. *Frontiers in Cell and Developmental Biology* **8.** Publisher: Frontiers. doi:10.3389/fcell.2020.00215 (2020).

299. Economou, M., Trikalinos, T. A., Loizou, K. T., Tsianos, E. V. & Ioannidis, J. P. A. Differential Effects of NOD2 Variants on Crohn's Disease Risk and Phenotype in Diverse Populations: A Metaanalysis. en-US. *Official journal of the American College of Gastroenterology | ACG* **99,** 2393 (2004).

300. Eckmann, L. & Karin, M. NOD2 and Crohn's Disease: Loss or Gain of Function? English. *Immunity* **22.** Publisher: Elsevier, 661–667. doi:10.1016/j.immuni.2005.06.004 (2005).

301. Tuomisto, L. E., Ilmarinen, P., Lehtimäki, L., Tommola, M. & Kankaanranta, H. Immediate bronchodilator response in FEV1 as a diagnostic criterion for adult asthma. en. *European Respiratory Journal* **53.** Publisher: European Respiratory Society Section: Reviews. doi:10.1183/13993003.00904-2018 (2019).

302. McBrien, C. N. & Menzies-Gow, A. The Biology of Eosinophils and Their Role in Asthma. English. *Frontiers in Medicine* **4.** Publisher: Frontiers. doi:10.3389/fmed.2017.00093 (2017).

303. Gao, W. *et al.* The Pathogenesis of Eosinophilic Asthma: A Positive Feedback Mechanism That Promotes Th2 Immune Response via Filaggrin Deficiency. English. *Frontiers in Immunology* **12.** Publisher: Frontiers. doi:10.3389/fimmu.2021.672312 (2021).

304. Gonzalez, J. R. *et al.* FLG Deficiency in Mice Alters the Early-Life CD4+ T-Cell Response to Skin Commensal Bacteria. *Journal of Investigative Dermatology* **143,** 790–800.e12. doi:10.1016/j.jid.2022.10.019 (2023).

305. Chaturvedi, P., Khanna, R. & Parnaik, V. K. Ubiquitin ligase RNF123 mediates degradation of heterochromatin protein 1α and β in lamin A/C knock-down cells. eng. *PloS One* **7,** e47558. doi:10.1371/journal.pone.0047558 (2012).

306. Stöckli, J. *et al.* ABHD15 regulates adipose tissue lipolysis and hepatic lipid accumulation. *Molecular Metabolism* **25,** 83–94. doi:10.1016/j.molmet.2019.05.002 (2019).

307. Thorolfsdottir, R. B. *et al.* Coding variants in RPL3L and MYZAP increase risk of atrial fibrillation. en. *Communications Biology* **1.** Publisher: Nature Publishing Group, 1–9. doi:10.1038/s42003-018-0068-9 (2018).

308. Fattah, M. A., El Baz, M., Sherif, A. & Adel, A. Complement components (C3, C4) as inflammatory markers in asthma. en. *The Indian Journal of Pediatrics* **77,** 771–773. doi:10.1007/s12098-010-0117-y (2010).

309. Oddsson, A. *et al.* Deficit of homozygosity among 1.52 million individuals and genetic causes of recessive lethality. en. *Nature Communications* **14.** Publisher: Nature Publishing Group, 3453. doi:10.1038/s41467-023-38951-2 (2023).

310. Camargos, S. *et al.* DYT16, a novel young-onset dystonia-parkinsonism disorder: identification of a segregating mutation in the stress-response protein PRKRA. English. *The Lancet Neurology* **7.** Publisher: Elsevier, 207–215. doi:10.1016/S1474-4422(08)70022-X (2008).

311. Zech, M. *et al.* DYT16 revisited: Exome sequencing identifies PRKRA mutations in a European dystonia family. en. *Movement Disorders* **29.** _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/mds.25981, 1504–1510. doi:10.1002/mds.25981 (2014).

312. Dhindsa, R. S., Copeland, B. R., Mustoe, A. M. & Goldstein, D. B. Natural Selection Shapes Codon Usage in the Human Genome. English. *The American Journal of Human Genetics* **107.** Publisher: Elsevier, 83–95. doi:10.1016/j.ajhg.2020.05.011 (2020).

313. Limoges, M.-A., Cloutier, M., Nandi, M., Ilangumaran, S. & Ramanathan, S. The GIMAP Family Proteins: An Incomplete Puzzle. English. *Frontiers in Immunology* **12.** Publisher: Frontiers. doi:10.3389/fimmu.2021.679739 (2021).

314. Kishibe, M. *et al.* Kallikrein 8 Is Involved in Skin Desquamation in Cooperation with Other Kallikreins *. English. *Journal of Biological Chemistry* **282.** Publisher: Elsevier, 5834–5841. doi:10.1074/jbc.M607998200 (2007).

315. Lachner, J., Mlitz, V., Tschachler, E. & Eckhart, L. Epidermal cornification is preceded by the expression of a keratinocyte-specific set of pyroptosis-related genes. en. *Scientific Reports* **7.** Publisher: Nature Publishing Group, 17446. doi:10.1038/s41598-017-17782-4 (2017).

316. Tseng, P.-Y. & Hoon, M. A. GPR15L is an epithelial inflammation-derived pruritogen. *Science Advances* **8.** Publisher: American Association for the Advancement of Science, eabm7342. doi:10.1126/sciadv.abm7342 (2022).

317. Smieszek, S. P. *et al.* Correlation of age-of-onset of Atopic Dermatitis with Filaggrin loss-of-function variant status. en. *Scientific Reports* **10.** Publisher: Nature Publishing Group, 2721. doi:10.1038/s41598-020-59627-7 (2020).

318. Grünberg, M. *et al.* Kallikrein-related peptidases are activators of the CC chemokine CCL14. *European Journal of Immunology* **48,** 1592–1594. doi:10.1002/eji.201747452 (2018).

319. Suply, T. *et al.* A natural ligand for the orphan receptor GPR15 modulates lymphocyte recruitment to epithelia. *Science Signaling* **10.** Publisher: American Association for the Advancement of Science, eaal0180. doi:10.1126/scisignal.aal0180 (2017).

320. Welcker, D. *et al.* Hemicentin-1 is an essential extracellular matrix component of the dermal–epidermal and myotendinous junctions. en. *Scientific Reports* **11.** Publisher: Nature Publishing Group, 17926. doi:10.1038/s41598-021-96824-4 (2021).

321. Hedrick, P. W. What is the evidence for heterozygote advantage selection? *Trends in Ecology & Evolution* **27,** 698–704. doi:10.1016/j.tree.2012.08.012 (2012).

322. Morissette, J. *et al.* Homozygotes carrying an autosomal dominant TIGR mutation do not manifest glaucoma. en. *Nature Genetics* **19.** Publisher: Nature Publishing Group, 319–321. doi:10.1038/1203 (1998).

323. Zschocke, J., Byers, P. H. & Wilkie, A. O. M. Mendelian inheritance revisited: dominance and recessiveness in medical genetics. en. *Nature Reviews Genetics* **24.** Publisher: Nature Publishing Group, 442–463. doi:10.1038/s41576-023-00574-0 (2023).

324. Forabosco, P. *et al.* Combined segregation and linkage analysis of inflammatory bowel disease in the IBD1 region using severity to characterise Crohn's disease and ulcerative colitis. en. *European Journal of Human Genetics* **8.** Publisher: Nature Publishing Group, 846–852. doi:10.1038/sj.ejhg.5200542 (2000).

325. Cuthbert, A. P. *et al.* The contribution of NOD2 gene mutations to the risk and site of disease in inflammatory bowel disease. en. *Gastroenterology* **122,** 867–874. doi:10.1053/gast.2002.32415 (2002).

326. Ashton, J. J. *et al.* Prediction of Crohn's Disease Stricturing Phenotype Using a NOD2-derived Genomic Biomarker. *Inflammatory Bowel Diseases* **29,** 511–521. doi:10.1093/ibd/izac205 (2023).

327. Hugot, J.-P. *et al.* Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. en. *Nature* **411.** Publisher: Nature Publishing Group, 599–603. doi:10.1038/35079107 (2001).

328. Andreoletti, G. *et al.* Exome Analysis of Rare and Common Variants within the NOD Signaling Pathway. en. *Scientific Reports* **7.** Publisher: Nature Publishing Group, 46454. doi:10.1038/srep46454 (2017).

329. Kayali, S. *et al.* NOD2 and Crohn's Disease Clinical Practice: From Epidemiology to Diagnosis and Therapy, Rewired. *Inflammatory Bowel Diseases.* doi:10.1093/ibd/izae075 (2024).

330. Brown, S. J. *et al.* Filaggrin null mutations and childhood atopic eczema: A population-based case-control study. *Journal of Allergy and Clinical Immunology* **121,** 940–946.e3. doi:10.1016/j.jaci.2008.01.013 (2008).

331. Possa, S. S., Leick, E. A., Prado, C. M., Martins, M. A. & Tibério, I. F. Eosinophilic Inflammation in Allergic Asthma. English. *Frontiers in Pharmacology* **4.** Publisher: Frontiers. doi:10.3389/fphar.2013.00046 (2013).

332. Hesselberg, L. M. *et al.* Handgrip strength associates with effort-dependent lung function measures among adolescents with and without asthma. en. *Scientific Reports* **13.** Publisher: Nature Publishing Group, 13044. doi:10.1038/s41598-023-40320-4 (2023).

333. Chan, Y. *et al.* Genome-wide Analysis of Body Proportion Classifies Height-Associated Variants by Mechanism of Action and Implicates Genes Important for Skeletal Development. English. *The American Journal of Human Genetics* **96.** Publisher: Elsevier, 695–708. doi:10.1016/j.ajhg.2015.02.018 (2015).

334. Fredriks, A. M. *et al.* Nationwide age references for sitting height, leg length, and sitting height/height ratio, and their diagnostic value for disproportionate growth disorders. en. *Archives of Disease in Childhood* **90.** Publisher: BMJ Publishing Group Ltd Section: Community child health, public health, and epidemiology, 807–812. doi:10.1136/adc.2004.050799 (2005).

335. Manichaikul, A. *et al.* Genome-wide association study of subclinical interstitial lung disease in MESA. *Respiratory Research* **18,** 97. doi:10.1186/s12931-017-0581-2 (2017).

336. Nguyen, P. A., Born, D. A., Deaton, A. M., Nioi, P. & Ward, L. D. Phenotypes associated with genes encoding drug targets are predictive of clinical trial side effects. en. *Nature Communications* **10.** Publisher: Nature Publishing Group, 1579. doi:10.1038/s41467-019-09407-3 (2019).

337. McGregor, T. L. *et al.* Characterising a healthy adult with a rare HAO1 knockout to support a therapeutic strategy for primary hyperoxaluria. *eLife* **9** (eds Goate, A., Weigel, D., Ryten, M. & Blackburn, N.) Publisher: eLife Sciences Publications, Ltd, e54363. doi:10.7554/eLife.54363 (2020).

338. Maag, H. *et al.* en (2020).

339. Mackinnon, S. R., Zarganes-Tzitzikas, T., Adams, C. J., Brennan, P. E. & Yue, W. W. Luminescence-based complementation assay to assess target engagement and cell permeability of glycolate oxidase (HAO1) inhibitors. *Biochimie.* doi:10.1016/j.biochi.2024.08.011 (2024).

340. Narasimhan, V. M. *et al.* Health and population effects of rare gene knockouts in adult humans with related parents. *Science* **352.** Publisher: American Association for the Advancement of Science, 474–477. doi:10.1126/science.aac8624 (2016).

341. Martin, A. R. *et al.* Clinical use of current polygenic risk scores may exacerbate health disparities. en. *Nature Genetics* **51,** 584–591. doi:10.1038/s41588-019-0379-x (2019).

342. Wall, J. D. *et al.* South Asian medical cohorts reveal strong founder effects and high rates of homozygosity. en. *Nature Communications* **14.** Publisher: Nature Publishing Group, 3377. doi:10.1038/s41467-023-38766-1 (2023).

343. Narasimhan, V. M. *et al.* Estimating the human mutation rate from autozygous segments reveals population differences in human mutational processes. en. *Nature Communications* **8.** Publisher: Nature Publishing Group, 303. doi:10.1038/s41467-017-00323-y (2017).

344. Malawsky, D. S. *et al.* Influence of autozygosity on common disease risk across the phenotypic spectrum. *Cell* **186,** 4514–4527.e14. doi:10.1016/j.cell.2023.08.028 (2023).

345. Corbin, L. J. *et al.* Formalising recall by genotype as an efficient approach to detailed phenotyping and causal inference. en. *Nature Communications* **9.** Publisher: Nature Publishing Group, 711. doi:10.1038/s41467-018-03109-y (2018).

346. Jacobs, B. *et al.,* (2023). Genetic architecture of routinely acquired blood tests in a UK cohort of South Asian ancestry reveals ancestry-specific causal variants. URL: https://www.researchsquare.com/article/rs-3438851/v1.

347. Null null, n. 100,000 Genomes Pilot on Rare-Disease Diagnosis in Health Care — Preliminary Report. *New England Journal of Medicine* **385.** Publisher: Massachusetts Medical Society _eprint: https://www.nejm.org/doi/pdf/10.1056/NEJMoa2035790, 1868–1880. doi:10.1056/NEJMoa2035790 (2021).

348. Sosinsky, A. *et al.* Insights for precision oncology from the integration of genomic and clinical data of 13,880 tumors from the 100,000 Genomes Cancer Programme. en. *Nature Medicine* **30.** Publisher: Nature Publishing Group, 279–289. doi:10.1038/s41591-023-02682-0 (2024).

349. (). U.K. Unveils Plan to Sequence Whole Genomes of 100,000 Patients. URL: https://www.science.org/content/article/uk-unveils-plan-sequence-whole-genomes-100000-patients.

350. Turnbull, C. Introducing whole-genome sequencing into routine cancer care: the Genomics England 100 000 Genomes Project. English. *Annals of Oncology* **29.** Publisher: Elsevier, 784–787. doi:10.1093/annonc/mdy054 (2018).

351. Turnbull, C. *et al.* The 100 000 Genomes Project: bringing whole genome sequencing to the NHS. *BMJ: British Medical Journal* **361.** Publisher: BMJ (2018).

352. Akiyama, M. *et al.* Genome-wide association study identifies 112 new loci for body mass index in the Japanese population. en. *Nature Genetics* **49.** Publisher: Nature Publishing Group, 1458–1467. doi:10.1038/ng.3951 (2017).

353. Kanai, M. *et al.* Genetic analysis of quantitative traits in the Japanese population links cell types to complex human diseases. en. *Nature Genetics* **50.** Publisher: Nature Publishing Group, 390–400. doi:10.1038/s41588-018-0047-6 (2018).

354. Akiyama, M. *et al.* Characterizing rare and low-frequency height-associated variants in the Japanese population. en. *Nature Communications* **10.** Publisher: Nature Publishing Group, 4393. doi:10.1038/s41467-019-12276-5 (2019).

355. Sakaue, S. *et al.* A cross-population atlas of genetic associations for 220 human phenotypes. en. *Nature Genetics* **53.** Publisher: Nature Publishing Group, 1415–1424. doi:10.1038/s41588-021-00931-x (2021).

356. Hirata, M. *et al.* Overview of BioBank Japan follow-up data in 32 diseases. *Journal of Epidemiology. SI: Overview of BBJ cohort* **27,** S22–S28. doi:10.1016/j.je.2016.12.006 (2017).

357. McKenna, A. *et al.* The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* **20,** 1297–1303. doi:10.1101/gr.107524.110 (2010).

358. Li, H., (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. URL: http://arxiv.org/abs/1303.3997.

359. (2019). Picard toolkit.

360. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. en. *Nature* **526,** 68–74 (2015).

361. Ding, Y. *et al.* Polygenic scoring accuracy varies across the genetic ancestry continuum. en. *Nature* **618.** Publisher: Nature Publishing Group, 774–781. doi:10.1038/s41586-023-06079-4 (2023).

362. Derks, E. M., Zwinderman, A. H. & Gamazon, E. R. The Relation Between Inflation in Type-I and Type-II Error Rate and Population Divergence in Genome-Wide Association Analysis of Multi-Ethnic Populations. en. *Behavior Genetics* **47,** 360–368. doi:10.1007/s10519-017-9837-3 (2017).

363. Atkinson, E. G. *et al.* Tractor uses local ancestry to enable the inclusion of admixed individuals in GWAS and to boost power. en. *Nature Genetics* **53.** Publisher: Nature Publishing Group, 195–204. doi:10.1038/s41588-020-00766-y (2021).

364. Shi, S. *et al.* A Genomics England haplotype reference panel and imputation of UK Biobank. en. *Nature Genetics.* Publisher: Nature Publishing Group, 1–4. doi:10.1038/s41588-024-01868-7 (2024).

365. Lim, E. T. *et al.* Rare Complete Knockouts in Humans: Population Distribution and Significant Role in Autism Spectrum Disorders. English. *Neuron* **77.** Publisher: Elsevier, 235–242. doi:10.1016/j.neuron.2012.12.029 (2013).

366. Stouffer, S. A., Suchman, E. A., Devinney, L. C., Star, S. A. & Williams Jr., R. M. *The American soldier: Adjustment during army life. (Studies in social psychology in World War II), Vol. 1* Pages: xii, 599 (Princeton Univ. Press, Oxford, England, 1949).

367. Higgins, J. P. T. & Thompson, S. G. Quantifying heterogeneity in a meta-analysis. en. *Statistics in Medicine* **21.** _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.1186, 1539–1558. doi:10.1002/sim.1186 (2002).

368. Liu, Y. *et al.* ACAT: A Fast and Powerful p Value Combination Method for Rare-Variant Analysis in Sequencing Studies. English. *The American Journal of Human Genetics* **104.** Publisher: Elsevier, 410–421. doi:10.1016/j.ajhg.2019.01.002 (2019).

369. Liu, Y. & Xie, J. Cauchy Combination Test: A Powerful Test With Analytic p-Value Calculation Under Arbitrary Dependency Structures. *Journal of the American Statistical Association* **115.** Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/01621459.2018.1554485, 393–402. doi:10.1080/01621459.2018.1554485 (2020).

370. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26,** 2190–2191. doi:10.1093/bioinformatics/btq340 (2010).

371. Miyazawa, K. *et al.* Cross-ancestry genome-wide analysis of atrial fibrillation unveils disease biology and enables cardioembolic risk prediction. en. *Nature Genetics* **55.** Publisher: Nature Publishing Group, 187–197. doi:10.1038/s41588-022-01284-9 (2023).

372. Auton, A. *et al.* A global reference for human genetic variation. en. *Nature* **526.** Publisher: Nature Publishing Group, 68–74. doi:10.1038/nature15393 (2015).

373. Mezei, G. & Kheifets, L. Selection bias and its implications for case–control studies: a case study of magnetic field exposure and childhood leukaemia. en. *International Journal of Epidemiology* **35,** 397–406. doi:10.1093/ije/dyi245 (2006).

374. Chen, C.-Y. *et al.* Analysis across Taiwan Biobank, Biobank Japan, and UK Biobank identifies hundreds of novel loci for 36 quantitative traits. English. *Cell Genomics* **3.** Publisher: Elsevier. doi:10.1016/j.xgen.2023.100436 (2023).

375. Chen, L., Magliano, D. J. & Zimmet, P. Z. The worldwide epidemiology of type 2 diabetes mellitus—present and future perspectives. en. *Nature Reviews Endocrinology* **8.** Publisher: Nature Publishing Group, 228–236. doi:10.1038/nrendo.2011.183 (2012).

376. Carter, P., Gray, L. J., Morris, D. H., Davies, M. J. & Khunti, K. South Asian individuals at high risk of type 2 diabetes have lower plasma vitamin C levels than white Europeans. *Journal of Nutritional Science* **2,** e21. doi:10.1017/jns.2013.15 (2013).

377. Gujral, U. P. & Kanaya, A. M. Epidemiology of diabetes among South Asians in the United States: lessons from the MASALA study. *Annals of the New York Academy of Sciences* **1495,** 24–39. doi:10.1111/nyas.14530 (2021).

378. Bradley, D. & Hsueh, W. Type 2 Diabetes in the Elderly: Challenges in a Unique Patient Population. *Journal of geriatric medicine and gerontology* **2,** 14. doi:10.23937/2469-5858/1510014 (2016).

379. Dendup, T., Feng, X., Clingan, S. & Astell-Burt, T. Environmental Risk Factors for Developing Type 2 Diabetes Mellitus: A Systematic Review. *International Journal of Environmental Research and Public Health* **15,** 78. doi:10.3390/ijerph15010078 (2018).

380. Gaskin, D. J. *et al.* Disparities in diabetes: the nexus of race, poverty, and place. eng. *American Journal of Public Health* **104,** 2147–2155. doi:10.2105/AJPH.2013.301420 (2014).

381. (). Prevalence of asthma. URL: https://ourworldindata.org/grapher/asthma-prevalence.

382. Knox, C. *et al.* DrugBank 6.0: the DrugBank Knowledgebase for 2024. *Nucleic Acids Research* **52,** D1265–D1275. doi:10.1093/nar/gkad976 (2024).

383. Yang, J. *et al.* Genomic inflation factors under polygenic inheritance. en. *European Journal of Human Genetics* **19.** Publisher: Nature Publishing Group, 807–812. doi:10.1038/ejhg.2011.39 (2011).

384. Zeng, P. *et al.* Statistical analysis for genome-wide association study. *Journal of Biomedical Research* **29,** 285–297. doi:10.7555/JBR.29.20140007 (2015).

385. Wang, J., Yu, J., Lipka, A. E. & Zhang, Z. en. in *Genome-Wide Association Studies* (eds Torkamaneh, D. & Belzile, F.) 63–80 (Springer US, New York, NY, 2022). doi:10.1007/978-1-0716-2237-7_5.

386. Sohail, M. *et al.* Polygenic adaptation on height is overestimated due to uncorrected stratification in genome-wide association studies. *eLife* **8** (eds Nordborg, M., McCarthy, M. I., Nordborg, M., Barton, N. H. & Hermisson, J.) Publisher: eLife Sciences Publications, Ltd, e39702. doi:10.7554/eLife.39702 (2019).

387. Haghpanah, S., Davani, M., Samadi, B., Ashrafi, A. & Karimi, M. Serum lipid profiles in patients with beta-thalassemia major and intermedia in southern Iran. *Journal of Research in Medical Sciences : The Official Journal of Isfahan University of Medical Sciences* **15,** 150–154 (2010).

388. De Coster, W., Weissensteiner, M. H. & Sedlazeck, F. J. Towards population-scale long-read sequencing. en. *Nature Reviews Genetics* **22.** Publisher: Nature Publishing Group, 572–587. doi:10.1038/s41576-021-00367-3 (2021).

389. Mahmoud, M. *et al.* Utility of long-read sequencing for All of Us. en. *Nature Communications* **15.** Publisher: Nature Publishing Group, 837. doi:10.1038/s41467-024-44804-3 (2024).

390. Mantere, T., Kersten, S. & Hoischen, A. Long-Read Sequencing Emerging in Medical Genetics. English. *Frontiers in Genetics* **10.** Publisher: Frontiers. doi:10.3389/fgene.2019.00426 (2019).

391. Mitsuhashi, S. & Matsumoto, N. Long-read sequencing for rare human genetic diseases. en. *Journal of Human Genetics* **65.** Publisher: Nature Publishing Group, 11–19. doi:10.1038/s10038-019-0671-8 (2020).

392. Jia, L. *et al.* Decoding mRNA translatability and stability from the 5 prime UTR. en. *Nature Structural & Molecular Biology* **27.** Publisher: Nature Publishing Group, 814–821. doi:10.1038/s41594-020-0465-x (2020).

393. Wieder, N. *et al.* Differences in 5'untranslated regions highlight the importance of translational regulation of dosage sensitive genes. en. *Genome Biology* **25,** 111. doi:10.1186/s13059-024-03248-0 (2024).

394. Long, H. K., Prescott, S. L. & Wysocka, J. Ever-Changing Landscapes: Transcriptional Enhancers in Development and Evolution. English. *Cell* **167.** Publisher: Elsevier, 1170–1187. doi:10.1016/j.cell.2016.09.018 (2016).

395. Hamamoto, K. & Fukaya, T. Molecular architecture of enhancer–promoter interaction. *Current Opinion in Cell Biology* **74,** 62–70. doi:10.1016/j.ceb.2022.01.003 (2022).

396. Fulco, C. P. *et al.* Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations. en. *Nature Genetics* **51.** Publisher: Nature Publishing Group, 1664–1669. doi:10.1038/s41588-019-0538-0 (2019).

397. Janssens, J. *et al.* Decoding gene regulation in the fly brain. en. *Nature* **601.** Publisher: Nature Publishing Group, 630–636. doi:10.1038/s41586-021-04262-z (2022).

398. Minikel, E. V., Painter, J. L., Dong, C. C. & Nelson, M. R. Refining the impact of genetic evidence on clinical success. en. *Nature* **629.** Publisher: Nature Publishing Group, 624–629. doi:10.1038/s41586-024-07316-0 (2024).

399. Collins, M. & Thrasher, A. Gene therapy: progress and predictions. *Proceedings of the Royal Society B: Biological Sciences* **282.** Publisher: Royal Society, 20143003. doi:10.1098/rspb.2014.3003 (2015).

400. Dunbar, C. E. *et al.* Gene therapy comes of age. *Science* **359.** Publisher: American Association for the Advancement of Science, eaan4672. doi:10.1126/science.aan4672 (2018).

401. Mendell, J. R. *et al.* Single-Dose Gene-Replacement Therapy for Spinal Muscular Atrophy. *New England Journal of Medicine* **377.** Publisher: Massachusetts Medical Society _eprint: https://www.nejm.org/doi/pdf/10.1056/NEJMoa1706198, 1713–1722. doi:10.1056/NEJMoa1706198 (2017).

402. Russell, S. *et al.* Efficacy and safety of voretigene neparvovec (AAV2-hRPE65v2) in patients with RPE65-mediated inherited retinal dystrophy: a randomised, controlled, open-label, phase 3 trial. English. *The Lancet* **390.** Publisher: Elsevier, 849–860. doi:10.1016/S0140-6736(17)31868-8 (2017).

403. Thompson, A. A. *et al.* Gene Therapy in Patients with Transfusion-Dependent β-Thalassemia. *New England Journal of Medicine* **378.** Publisher: Massachusetts Medical Society _eprint: https://www.nejm.org/doi/pdf/10.1056/NEJMoa1705342, 1479–1493. doi:10.1056/NEJMoa1705342 (2018).

404. Pipe, S. W. *et al.* Gene Therapy with Etranacogene Dezaparvovec for Hemophilia B. *New England Journal of Medicine* **388.** Publisher: Massachusetts Medical Society _eprint: https://www.nejm.org/doi/pdf/10.1056/NEJMoa2211644, 706–718. doi:10.1056/NEJMoa2211644 (2023).

405. Ramsey, B. W. *et al.* A CFTR Potentiator in Patients with Cystic Fibrosis and the G551D Mutation. *New England Journal of Medicine* **365.** Publisher: Massachusetts Medical Society _eprint: https://www.nejm.org/doi/pdf/10.1056/NEJMoa1105185, 1663–1672. doi:10.1056/NEJMoa1105185 (2011).

406. Fanen, P., Wohlhuter-Haddad, A. & Hinzpeter, A. Genetics of cystic fibrosis: CFTR mutation classifications toward genotype-based CF therapies. *The International Journal of Biochemistry & Cell Biology. Cystic Fibrosis: From o-mics to cell biology, physiology, and therapeutic advances* **52,** 94–102. doi:10.1016/j.biocel.2014.02.023 (2014).

407. Iizuka, T. *et al.* Perinatal Gjb2 gene transfer rescues hearing in a mouse model of hereditary deafness. *Human Molecular Genetics* **24,** 3651–3661. doi:10.1093/hmg/ddv109 (2015).

408. Enjalbert, F. *et al.* HEMATOPOIETIC STEM CELL GENE THERAPY AS A NOVEL THERAPEUTIC APPROACH FOR SEVERE CROHN'S DISEASE ASSOCIATED WITH NOD2-DEFICIENCY. *Inflammatory Bowel Diseases* **30,** S42. doi:10.1093/ibd/izae020.087 (2024).

409. Zimran, A. *et al.* Phase 1/2 and extension study of velaglucerase alfa replacement therapy in adults with type 1 Gaucher disease: 48-month experience. eng. *Blood* **115,** 4651–4656. doi:10.1182/blood-2010-02-268649 (2010).

410. MacLaren, R. E. *et al.* Retinal gene therapy in patients with choroideremia: initial findings from a phase 1/2 clinical trial. English. *The Lancet* **383.** Publisher: Elsevier, 1129–1137. doi:10.1016/S0140-6736(13)62117-0 (2014).

411. Akil, O. *et al.* Dual AAV-mediated gene therapy restores hearing in a DFNB9 mouse model. *Proceedings of the National Academy of Sciences* **116.** Publisher: Proceedings of the National Academy of Sciences, 4496–4501. doi:10.1073/pnas.1817537116 (2019).

412. Richards, D. Y. *et al.* AAV-Mediated CRISPR/Cas9 Gene Editing in Murine Phenylketonuria. English. *Molecular Therapy Methods & Clinical Development* **17.** Publisher: Elsevier, 234–245. doi:10.1016/j.omtm.2019.12.004 (2020).

413. Kanter, J. *et al.* Biologic and Clinical Efficacy of LentiGlobin for Sickle Cell Disease. *New England Journal of Medicine* **386.** Publisher: Massachusetts Medical Society _eprint: https://www.nejm.org/doi/pdf/10.1056/NEJMoa2117175, 617–628. doi:10.1056/NEJMoa2117175 (2022).

414. FitzGerald, G. *et al.* The future of humans as model organisms. *Science* **361.** Publisher: American Association for the Advancement of Science, 552–553. doi:10.1126/science.aau7779 (2018).

415. Phillips, K. A. *et al.* Why primate models matter. en. *American Journal of Primatology* **76.** _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ajp.22281, 801–827. doi:10.1002/ajp.22281 (2014).

416. Sigmund, C. D. Viewpoint: Are Studies in Genetically Altered Mice Out of Control? *Arteriosclerosis, Thrombosis, and Vascular Biology* **20.** Publisher: American Heart Association, 1425–1429. doi:10.1161/01.ATV.20.6.1425 (2000).

417. Perlman, R. L. Mouse models of human disease: An evolutionary perspective. *Evolution, Medicine, and Public Health* **2016,** 170–176. doi:10.1093/emph/eow014 (2016).

418. Paul, S. M. *et al.* How to improve R&D productivity: the pharmaceutical industry's grand challenge. en. *Nature Reviews Drug Discovery* **9.** Publisher: Nature Publishing Group, 203–214. doi:10.1038/nrd3078 (2010).

419. Ormandy, E. H., Dale, J. & Griffin, G. Genetic engineering of animals: Ethical issues, including welfare concerns. *The Canadian Veterinary Journal* **52,** 544–550 (2011).

420. Rausell, A. *et al.* Common homozygosity for predicted loss-of-function variants reveals both redundant and advantageous effects of dispensable human genes. en. *Proceedings of the National Academy of Sciences* **117,** 13626–13636. doi:10.1073/pnas.1917993117 (2020).

421. Narasimhan, V. M., Xue, Y. & Tyler-Smith, C. Human Knockout Carriers: Dead, Diseased, Healthy, or Improved? English. *Trends in Molecular Medicine* **22.** Publisher: Elsevier, 341–351. doi:10.1016/j.molmed.2016.02.006 (2016).

422. Maddirevula, S. *et al.* GWAS signals revisited using human knockouts. *Genetics in Medicine* **20,** 64–68. doi:10.1038/gim.2017.78 (2018).

423. Spedicati, B. *et al.* Natural human knockouts and Mendelian disorders: deep phenotyping in Italian isolates. en. *European Journal of Human Genetics* **29.** Publisher: Nature Publishing Group, 1272–1281. doi:10.1038/s41431-021-00850-9 (2021).

424. Carss, K. J. *et al.* Using human genetics to improve safety assessment of therapeutics. en. *Nature Reviews Drug Discovery* **22.** Publisher: Nature Publishing Group, 145–162. doi:10.1038/s41573-022-00561-w (2023).

425. Standing, J. F. Understanding and applying pharmacometric modelling and simulation in clinical practice and research. en. *British Journal of Clinical Pharmacology* **83.** _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/bcp.13119, 247–254. doi:10.1111/bcp.13119 (2017).

426. Scannell, J. W., Blanckley, A., Boldon, H. & Warrington, B. Diagnosing the decline in pharmaceutical R&D efficiency. en. *Nature Reviews Drug Discovery* **11.** Publisher: Nature Publishing Group, 191–200. doi:10.1038/nrd3681 (2012).

427. Spiekerkoetter, U. *et al.* Genomic newborn screening: Are we entering a new era of screening? en. *Journal of Inherited Metabolic Disease* **46.** _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jimd.12650, 778–795. doi:10.1002/jimd.12650 (2023).

428. Turnbull, C. *et al.* Population screening requires robust evidence—genomics is no exception. *The Lancet* **403,** 583–586. doi:10.1016/S0140-6736(23)02295-X (2024).

429. Adeyemo, A. *et al.* Responsible use of polygenic risk scores in the clinic: potential benefits, risks and gaps. en. *Nature Medicine* **27.** Publisher: Nature Publishing Group, 1876–1884. doi:10.1038/s41591-021-01549-6 (2021).

430. Tu, T. *et al.,* (2024). Towards Conversational Diagnostic AI. URL: http://arxiv.org/abs/2401.05654.

431. Singhal, K. *et al.,* (2023). Towards Expert-Level Medical Question Answering with Large Language Models. URL: https://arxiv.org/abs/2305.09617v1.

432. Cottingham, M. D. & Fisher, J. A. Gendered Logics of Biomedical Research: Women in U.S. Phase I Clinical Trials. *Social problems* **69,** 492–509. doi:10.1093/socpro/spaa035 (2022).

433. Nazer, L. H. *et al.* Bias in artificial intelligence algorithms and recommendations for mitigation. en. *PLOS Digital Health* **2.** Publisher: Public Library of Science, e0000278. doi:10.1371/journal.pdig.0000278 (2023).

434. Jinek, M. *et al.* A Programmable Dual-RNA–Guided DNA Endonuclease in Adaptive Bacterial Immunity. *Science* **337.** Publisher: American Association for the Advancement of Science, 816–821. doi:10.1126/science.1225829 (2012).

435. Anzalone, A. V. *et al.* Search-and-replace genome editing without double-strand breaks or donor DNA. en. *Nature* **576.** Publisher: Nature Publishing Group, 149–157. doi:10.1038/s41586-019-1711-4 (2019).

436. Chen, P. J. & Liu, D. R. Prime editing for precise and highly versatile genome manipulation. en. *Nature Reviews Genetics* **24.** Publisher: Nature Publishing Group, 161–177. doi:10.1038/s41576-022-00541-1 (2023).

437. Wang, J. Y. & Doudna, J. A. CRISPR technology: A decade of genome editing is only the beginning. *Science* **379.** Publisher: American Association for the Advancement of Science, eadd8643. doi:10.1126/science.add8643 (2023).

438. Dilawari, R. *et al.* en. in *Harsh Environment and Plant Resilience: Molecular and Functional Aspects* (ed Husen, A.) 45–72 (Springer International Publishing, Cham, 2021). doi:10.1007/978-3-030-65912-7_3.

439. Kumar, M. *et al.* Application of CRISPR/Cas9-mediated gene editing for abiotic stress management in crop plants. English. *Frontiers in Plant Science* **14.** Publisher: Frontiers. doi:10.3389/fpls.2023.1157678 (2023).

440. Ben Ouagrham-Gormley, S. & Fye-Marnien, S. R. en. in *Defense Against Biological Attacks: Volume I* (eds Singh, S. K. & Kuhn, J. H.) 233–251 (Springer International Publishing, Cham, 2019). doi:10.1007/978-3-030-03053-7_12.

441. Reeves, R. G., Voeneky, S., Caetano-Anollés, D., Beck, F. & Boëte, C. Agricultural research, or a new bioweapon system? *Science* **362.** Publisher: American Association for the Advancement of Science, 35–37. doi:10.1126/science.aat7664 (2018).