NIH Oxford-Cambridge Scholars Program

National Institute on Alcohol Abuse and Alcoholism

Doctor of Philosophy Thesis

Using Mendelian randomization and genetic epidemiology strategies to elucidate novel causal biological mechanisms and pathways connecting risk factors with disease outcomes

Candidate
Daniel B. Rosoff

Supervisors
Dr. Falk W. Lohoff (National Institutes of Health)
Prof. David Ray (University of Oxford)
Dr. Constantinos Christodoulides (University of Oxford)
Prof. George Davey Smith (University of Bristol)

ACKNOWLEDGEMENTS

I would like to thank my supervisory team: Dr. Falk W. Lohoff at the National Institute on Alcohol Abuse and Alcoholism (NIAAA) in the National Institutes of Health (NIH), Dr. David Ray at the Oxford Centre for Diabetes, Endocrinology, and Metabolism (OCDEM), Prof. George Davey Smith at the University of Bristol, and Dr. Costas Christodoulides, also at OCDEM. Your guidance and expertise have been invaluable throughout my PhD, and I am indebted to you all for your exceptional support in helping me develop as a scientist.

I want to thank Prof. Mike Dustin, my Director of Graduate Studies, for his guidance and support and acknowledge the NIH Oxford-Cambridge Scholars Program (Ox Cam) and the OxCam Administrative team for providing the framework and resources that made this work possible.

Lastly, I am deeply grateful to my family for their encouragement and support throughout this journey. Your belief in me has been a constant source of motivation.

AUTHOR'S DECLARATION

I hereby confirm that the research presented in this dissertation was conducted in compliance with the University of Oxford's Regulations and Code of Practice for Research Degree Programs. This work has not been submitted for consideration towards any other academic qualification. Unless explicitly referenced, all work is my own. Contributions made in collaboration with others or with external assistance are appropriately acknowledged. The opinions and interpretations expressed within this dissertation are solely my own.

Daniel B. Rosoff

AUTHORSHIP STATEMENT

Work presented in this thesis has also appeared in peer-reviewed articles published prior to the submission of the thesis. This statement is to confirm the author's contribution to the work as follows:

Chapter 1 includes discussion of published materials where the author of this thesis was either the first author, co-first author or a contributing author. For the first author and co-first author manuscripts DBR confirms lead authorship of the published results. DBR also contributed to the study design, undertook all statistical and computational analyses, drafted the manuscript and generated all figures.

Results and data provided in Chapter 3 also appear in Rosoff et al.[1]

Rosoff DB, Wagner J, Jung J, et al. Multi-Omics Mendelian Randomization Study Investigating the Impact of PCSK9 and HMGCR Inhibition on Type 2 Diabetes Across Five Populations. *Diabetes*. 2024:db240451. doi:10.2337/db24-0451

The author of this thesis confirms lead authorship of the published results and contributions of co-authors are outlined as follows: DBR undertook all statistical and computational analyses, drafted the manuscript and generated all figures. DBR and FWL conceived the design of the study. FWL, DR, CC, and GDS supervised the project and advised on follow-up analyses during the peer-review process.

Results and data provided in Chapter 4 also appear Rosoff et al.[2]

Rosoff DB, Wagner J, Bell AS, Mavromatis LA, Jung J, Lohoff FW. A multi-omics Mendelian randomization study identifies new therapeutic targets for alcohol use disorder and problem drinking. *Nature Human Behaviour*. 2024/11/11 2024;doi:10.1038/s41562-024-02040-1

Permissions to reproduce published figures and materials included in this thesis have been obtained from the respective copyright holders, including BioRender.com (academic licence) and journal publishers. Copies of all license confirmation documents are provided in **Appendix 5** (Publication Licenses and Permissions).

The author of this thesis confirms lead authorship of the published results and contributions of co-authors are outlined as follows: DBR contributed to the study design, undertook all statistical and computational analyses, drafted the manuscript and generated all figures. DBR and FWL conceived the design of the study. FWL supervised the project and advised on follow-up analyses during the peer-review process.

Daniel B. Rosoff

ABSTRACT

Cardiometabolic diseases, including type 2 diabetes (T2D), non-alcoholic fatty liver disease (NAFLD), and coronary artery disease (CAD), impose a significant global health burden due to their shared risk factors and complex etiology. Despite advancements in treatment, therapeutic development has been hindered by an incomplete understanding of the shared mechanisms driving these diseases. This thesis leverages advanced genetic methodologies and multi-omics approaches to elucidate disease mechanisms, validate drug targets, and prioritize new therapeutic strategies for cardiometabolic health. The three aims of this thesis explore distinct aspects of therapeutic development: Aim 1 evaluates the safety of lipid-lowering therapies; Aim 2 investigates the genetic and biological underpinnings of a major cardiovascular risk factor, alcohol consumption; and Aim 3 integrates multivariate analyses to uncover shared genetic mechanisms underlying cardiometabolic multimorbidity. Aim 1 applied drug-target Mendelian Randomization (MR) to investigate the effects of proprotein convertase subtilisin/kexin type 9 (PCSK9) and 3-hydroxy-3-methylglutaryl-CoA reductase (HMGCR) inhibition—key lipid-lowering therapeutic targets—on T2D risk and glycemic traits across multi-ancestry GWAS datasets. Results showed that PCSK9 inhibition had no significant adverse effect on T2D risk, supporting its safety for diverse populations. However, HMGCR inhibition modestly increased T2D risk, emphasizing the need for therapeutic strategies tailored to population-specific genetic contexts. Aim 2 focused on alcohol consumption, a major behavioral risk factor for cardiovascular disease (CVD), and its biological underpinnings. Using MR integrated with cortical proteomic and cell-type transcriptomic data, the study identified novel molecular targets and pathways, such as SAMHD1 and VIPAS39, linking alcohol use behaviors to neural signaling, alcohol metabolism, and cardiometabolic health. The findings offer insights into therapeutic opportunities for addressing alcohol-associated risks in CVD. Aim 3 investigated the shared genetic architecture of NAFLD, T2D, and CAD through multivariate GWAS, uncovering a shared cardiometabolic factor (“*CM-Factor”*) and identifying 523 SNPs across 312 loci. Functional analyses prioritized genes such as COMT, DHX36, and ASPRV1, while drug-target MR validated established targets, including GLP1R and PCSK9, and identified novel candidates like CRY2 and OPRL1. Two-step MR analyses further linked BMI-associated proteins, such as ENO3, to cardiometabolic risk, providing mechanistic insights into the interplay between obesity and disease. This thesis highlights the transformative potential of genetic methodologies to inform therapeutic development for cardiometabolic health. By evaluating the safety of lipid-lowering therapies, uncovering the biological underpinnings of a major cardiovascular risk factor, and identifying shared genetic mechanisms driving cardiometabolic multimorbidity, this work addresses critical gaps in understanding and treatment. The integration of multi-omics data and multivariate frameworks not only validates established therapeutic targets like PCSK9 and GLP1R but also identifies novel opportunities, including CRY2 and OPRL1, for addressing complex disease pathways. Emphasizing the importance of diverse and representative datasets, these findings lay a foundation for precision medicine approaches to improve outcomes for cardiometabolic diseases in diverse patient populations.

Table of Contents

CHAPTER 1: BACKGROUND & RATIONALE.....10

Chapter Overview.....10

1.1. Background and rationale.....10

1.2. Statement of Aims.....12

Aim 1: Investigating the Impact of PCSK9 and HMGCR Inhibition on Type 2 Diabetes Across Diverse Populations.....13

Aim 2: Deciphering the Genetic Basis of Alcohol Consumption and its Role as a Modifiable Cardiometabolic Risk Factor.....13

Aim 3: Multivariate Genome-Wide Analysis of NAFLD, T2D, and CAD to Identify Shared Genetic Architecture.....13

1.3. Aim 1. Investigating the Impact of PCSK9 and HMGCR Inhibition on Type 2 Diabetes Across Five Populations.....14

1.4. Aim 2: Deciphering the Genetic Basis of Alcohol Consumption and its Role as a Modifiable Cardiometabolic Risk Factor.....15

1.5. Aim 3: Multivariate Genome-Wide Analysis of Non-Alcoholic Fatty Liver Disease, Type 2 Diabetes, and Coronary Artery Disease.....17

1.6. Research Foundations and Motivation: Shaping the Focus of My PhD Projects.....18

1.6.1. Why focus on PCSK9 and HMGCR in Aim 1?.....19

1.6.2. Why investigate the genetic underpinnings of alcohol consumption behaviors?. 25

1.6.3. Why explore the genetic links between T2D, CAD, and NAFLD with multivariate GWAS methods?.....32

CHAPTER 2: MATERIALS & METHODS.....36

Chapter Overview.....36

2.1. Mendelian Randomization Background.....36

2.1.1. Mendelian randomization assumptions.....37

2.1.2. Drug-target/cis-instrument MR.....38

2.2. Aim 1 Methods.....40

2.2.1. Data sources.....40

2.2.2. PCSK9, HMGCR, and polygenic LDL-C instruments.....42

2.2.3. PCSK9 protein and PCSK9 expression data.....43

2.2.4. Statistical analysis.....43

2.2.5. Sensitivity analyses: multivariable MR with body mass index.....44

2.2.6. Interpretation of MR results.....44

2.3. Aim 2 Methods.....44

2.3.1. Data sources.....46

2.3.1a. Alcohol consumption behaviors.....46

2.3.1b. Cortical pQTLs.....47

2.3.1c. Cell-type eQTLs in 8 brain cell types.....47

2.3.1d. Structural MRI data for MR analyses.....47

2.3.1e. White matter microstructure (diffusion MRI) for MR analyses.....48

2.3.2. Cis-instrumentation of brain proteins and transcripts.....48

2.3.2a. Cis-instrument construction.....48

2.3.2b. Cis-instrument MR screen....49
2.3.2c. Cis-MR multiverse sensitivity analysis....49
2.3.3. Annotation of the top targets prioritized by the cis-MR screens....50
2.3.3a. Gene ontology and pathway analysis....50
2.3.3b. Prioritized alcohol gene cell-type expression....50
2.3.3c. Gene-drug analyses assessing repurposing opportunities....50
2.3.4. Assessing novelty of the targets identified by the cis-instrument MR screen....51
2.3.4a. Comparison with the GWAS loci of each alcohol-related outcome....51
2.3.4b. H-MAGMA gene-based analyses incorporating cortical chromatin interaction data....51
2.3.4c. FUSION transcriptomic imputation....51
2.3.4d. Comparison with differentially expressed genes in post-mortem brains of AUD patients....52
2.3.4e. Comparison with the methylomic signature of AUD....52
2.3.4f. Comparison with previous TWAS and PWAS for alcohol-related outcomes....52
2.3.4g. Defining novel genes from the cis-instrument MR screens....53
2.3.5. Integrative validation and neurobiological contextualization of alcohol-related genetic findings....53
2.3.5a. Colocalization....53
2.3.5b. Exploring the impact of the alcohol-related genes on brain structure, white-matter tracts, and functional connectivity....53
2.3.5c. Replication and evaluation with EHR alcohol-related problems and diagnoses to prioritize therapeutic targets....54
2.3.5d. Neuropsychiatric contextualization....54

2.4. Aim 3 Methods....54
2.4.1. Data sources....57
2.4.2 Genomic structural equation modeling (GenomicSEM) background and rationale 58
2.4.3. Multivariate GWAS analysis....58
2.4.3a. Effective sample size calculation....59
2.4.3b. SNP-level heterogeneity testing to finalize the CM-Factor....59
2.4.3c. Annotation of CM-Factor loci....60
2.4.3d. Phenotypic characterization of the CM-Factor loci....60
2.4.3e. Sensitivity testing: GWAS-by-subtraction using GenomicSEM to account for obesity....61
2.4.3f. Fine-mapping of CM-Factor loci....64
2.4.3g. Gene prioritization using transcriptomic imputation....64
2.4.3h. Tissue and cell-type enrichment....66
2.4.4. Mendelian randomization applications....67
2.4.4a. MR Study 1: drug-target MR of antidiabetics, lipid-modulating therapeutics, NAFLD/NASH targets, and antihypertensives targets....69
2.4.4b. MR Study 2: Screening the druggable genome for novel targets for the CM-Factor....75
2.4.4c. MR Study 3: Two-step Mendelian randomization to identify circulating proteins mediating the impact of body mass index (BMI) on the CM-Factor....83

CHAPTER 3: AIM 1 RESULTS....90

Chapter Overview....90

3.1. PCSK9 and HMGCR Instrument Strength....90

3.2. The Impact of LDL-C Lowering on T2D by Genetically Mimicked PCSK9 and HMGCR Inhibition....90

3.3. The Impact of LDL-C Lowering on Glycemic Markers by Genetically Mimicked PCSK9 and HMGCR Inhibition....93

3.3. Polygenic LDL-C Results....98
3.4. Aim 1 Discussion....98
3.5. Aim 1 Strengths and Limitations....101
3.6. Aim 1 Future Directions....103
3.6.1 Expanding to more populations and cohorts....103
3.6.2. Investigating T2D complications....104
3.6.3. Two-step MR and mechanistic studies....104
3.6.4. Translating mechanistic insights into clinical practice....104
3.7. Aim 1 Conclusions....105
CHAPTER 4: AIM 2 RESULTS....106
Chapter Overview....106
4.1. Cis-Instrument MR Screens....106
4.2. Biological Characterization of Identified Genes....112
4.2.1. Cis-MR screens identify novel targets for alcohol traits....112
4.2.1a. PWAS/TWAS literature comparison....113
4.2.1b. Assessing directionality in bulk brain tissue....113
4.2.2. Colocalization prioritizes high confidence targets....113
4.2.2a. Distinct genes converge on brain structure and connectivity....115
4.3. Replication and Neuropsychiatric Contextualization....115
4.3.1. Characterizing relationships with neuropsychiatric outcomes....118
4.4. Aim 2 Discussion....125
4.4.1. Drug-gene interaction highlights potential for antidiabetics addressing alcohol behaviors and cardiometabolic disease....126
4.5. Aim 2 Strengths & Limitations....128
4.6. Aim 2 Conclusions....129
CHAPTER 5: AIM 3 RESULTS....130
Chapter Overview....130
5.1. Genetic Correlations and SEM modeling....130
5.2. Multivariate GWAS....130
5.2.1. Novelty and phenotypic analyses of *CM-Factor* loci....134
5.2.2. Sensitivity tests of the *CM-Factor*....134
5.2.3. SNP and gene prioritization with fine-mapping and transcriptomic imputation..135
5.2.4. Pathway, bulk tissue, and cell-type enrichment....138
5.3. Mendelian Randomization Analyses....138
5.3.1. MR Study 1: Drug-target analysis of approved and investigational cardiometabolic therapeutics....139
5.3.1a. Antidiabetics....139
5.3.1b. Lipid modulating targets....140
5.3.1c. NAFLD/NASH and antihypertensives....144
5.3.2. MR Study 2: MR screen of druggable genes prioritizes novel targets for cardiometabolic health....145
5.3.2a. Tissue-level replication....148
5.3.2b. Biomarker mediation....148
5.3.2c. Side-effect profiling for druggable genes....148

5.3.3. MR Study 3: Two-step MR prioritizes ENO3 as a proteomic mediator of the obesity-*CM-Factor* relationship....150
5.3.3a. Replication and triangulation of BMI-associated proteins....153
5.3.3b. Mediation of BMI-associated proteins....153
5.3.3c. BMI-associated protein side-effect profiles....154

5.4. Aim 3 Discussion....155
5.4.1. Drug-target MR assesses cardiometabolic efficacy of approved and investigational therapies and prioritizes targets for therapeutic discovery efforts....157

5.5. Aim 3 Limitations....160

5.6. Aim 1 Conclusion....162

CHAPTER 6: DISCUSSION & FUTURE DIRECTIONS....163

6.1. Leveraging Genetics to Inform Drug Development: Key Contributions from Each Aim....163
6.1.1. Aim 1 examined safety profiles of lipid-lowering drugs to non-European cohorts 163
6.1.2. Aim 2 provided genetic and biological insights into alcohol use behaviors....164
6.1.3. Aim 3 elucidated a shared genetic architecture of cardiometabolic diseases....165

6.2. Scaling Complexity in the PhD: From Drug-Target Validation to Multivariate Insights....166
6.2.1. Implications for the cardiometabolic drug development pipeline....166

6.3. Future Directions....167
6.3.1. Investigating the genetic architecture shared between sleep-related traits, circulating glycemic biomarkers, and type 2 diabetes....167
6.3.2. Addressing pleiotropy to refine causal inference in cardiometabolic disease....169
6.3.2a. What are network-based centrality metrics?....171
6.3.2b. What is PNIMR?....173
6.3.2c. Negative control #1: Vitamin D and CAD....175
6.3.2d. Negative control #2: CRP and CAD....175
6.3.2e. PNIMR correctly shows apparent impact of vitamin D and CRP on CAD is driven by highly connected SNPs....175
6.3.2f. Concluding Remarks for PNIMR....181

6.4. Conclusions....181

REFERENCES....181

APPENDIX 1: PUBLICATIONS....216

Selected first author, or co-first author publications....216
Manuscripts Related to PCSK9 and HMGCR Inhibition....217
Title: Evaluating the Cardiovascular Impact of Genetically Proxied PCSK9 and HMGCR Inhibition in East Asian and European Populations: A Drug-Target Mendelian Randomization Study....217
Title: Assessing the Impact of PCSK9 and HMGCR Inhibition on Liver Function: Drug-Target Mendelian Randomization Analyses in Four Ancestries....221
Title: Mendelian Randomization Study of PCSK9 and HMG-CoA Reductase Inhibition and Cognitive Function....233
Manuscripts Related to Alcohol Consumption Behaviors....243
Title: Association of High-Intensity Binge Drinking With Lipid and Liver Function Enzyme Levels....243
Title: Evaluating the Relationship Between Alcohol Consumption, Tobacco Use, and Cardiovascular Disease: A Multivariable Mendelian Randomization Study....255

Manuscripts Related to Aging and Longevity.....275
Title: Multivariate genome-wide analysis of aging-related traits identifies novel loci and new drug targets for healthy aging.....275
Title: Major Psychiatric Disorders, Substance Use Behaviors, and Longevity.....298
Title: Multi-omic underpinnings of epigenetic aging and human longevity.....311
Manuscripts Related to New Drug-Target Characterization.....326
Title: Bidirectional Mendelian Randomization Highlights Causal Relationships Between Circulating INHBC and Multiple Cardiometabolic Diseases and Traits.....326

Complete Publication List And Poster/Presentation Titles.....337
Publications.....337
Abstracts, Presentations, and Posters.....340

APPENDIX 2 (AP2): SUPPLEMENTARY MATERIALS FOR AIM 1.....344

AP2.1. Aim 1 Supplementary Tables.....344

APPENDIX 3 (AP3): SUPPLEMENTARY MATERIALS FOR AIM 2.....345

AP3.1. Supplementary Aim 1 Discussion.....345
AP3.1.1. Aim 1 Additional Limitations.....345

AP3.2. Aim 2 Supplementary Figures.....346

AP3.3. Aim 2 Supplementary Tables.....365

APPENDIX 4 (AP4): SUPPLEMENTARY MATERIALS FOR AIM 3.....368

AP4.1. Supplementary Methods for Aim 3.....368
AP4.1.1. Genomic Structural Equation Modeling (GenomicSEM).....368
AP4.1.1a. Structured covariance models and factor analysis.....369
AP4.1.1b. Assessing model fit.....372
AP4.1.1c. Confirmatory factor analysis.....372

AP4.2. Supplementary Aim 3 Results.....372
AP4.2.1. Qsnp heterogeneity testing.....372

AP4.3. Aim 3 Supplementary Discussion.....374
AP4.3.1. Extended discussion of the proteins prioritized by the two-step MR study....374
AP4.3.2. Limitations of the transcriptomic imputation gene prioritization analyses.....376

AP4.4. Aim 3 Supplementary Figures.....378

AP4.5. Aim 3 Supplementary Tables.....393

AP5. Biorender Publication Licenses.....395

CHAPTER 1: BACKGROUND & RATIONALE

Chapter Overview

This chapter establishes the scientific foundation and rationale for the research undertaken in this thesis, with a focus on the application of advanced genetic methodologies and multi-omics approaches to address critical gaps in the understanding and treatment of cardiovascular diseases (CVDs) and their comorbid conditions. It begins by contextualizing the global burden of CVDs, highlighting the limitations of existing pharmacological therapies and the challenges associated with the current drug development paradigm. The chapter discusses how emerging genetic tools, such as Mendelian randomization (MR) and multivariate genome-wide association studies (GWAS), offer innovative frameworks to identify causal mechanisms, prioritize therapeutic targets, and address health disparities. These approaches, when combined with multi-omics data, enable deeper exploration of the shared genetic architecture underlying complex phenotypes, including type 2 diabetes (T2D), non-alcoholic fatty liver disease (NAFLD), and alcohol consumption behaviors. By integrating these methods into the context of unmet clinical needs, this chapter lays the groundwork for the subsequent aims and analyses presented in this thesis.

1.1. Background and rationale

Cardiovascular diseases (CVDs) are the leading cause of death worldwide, accounting for approximately 31% of total deaths each year.[3-5] This global burden is driven by several factors, including aging populations, urbanization, and changes in dietary and physical activity patterns. In addition to their rapid rise in developing countries, CVDs have demonstrated a concerning resurgence in high-income nations, highlighting the multifaceted nature of the challenge.[3-5] While public health interventions such as promoting healthy diets, regular physical activity, and smoking cessation remain critical for reducing the overall burden of CVD, pharmacological therapies are indispensable for managing and mitigating disease-related mortality and morbidity.[6]

Despite significant advancements in CVD treatment and prevention strategies, critical gaps persist in addressing many common endpoints, including coronary heart disease (CHD), myocardial infarction (MI), and stroke. Even with established therapies such as statins, angiotensin-converting enzyme (ACE) inhibitors, and antiplatelet agents, a substantial proportion of patients continue to experience disease progression or adverse cardiovascular events. For instance, approximately 10% of hypertensive patients remain treatment-resistant, and between 15% and 30% of individuals with advanced MI present without traditional modifiable risk factors, such as hypertension or hypercholesterolemia.[7,8]

These gaps underscore the need for novel therapeutic approaches and a deeper understanding of the underlying pathophysiology of CVD. The process of developing new drugs for CVD is both resource-intensive and time-consuming, with estimated costs exceeding $1 billion per successful

drug and timelines of approximately 10 years from initial research to market approval.[6,9] Despite these investments, failure rates remain high, with only ~10% of drug candidates successfully progressing from clinical trials to regulatory approval.[10] Many of these failures are due to concerns about safety, lack of efficacy for the primary indication, or adverse side effects.[6,11] Notably, over half of the failures in Phase II and III trials are attributed to insufficient efficacy, while a significant proportion are linked to poor safety profiles.[10] These challenges have contributed to a stagnation in CVD drug development, with the number of cardiovascular drug candidates entering clinical trials declining relative to other therapeutic areas, such as oncology.[12]

One of the critical barriers to successful drug development is a limited understanding of the complex and often heterogeneous pathophysiology underlying CVD. Traditional approaches have primarily focused on conventional risk factors such as lipids, blood pressure, and platelet activity. While these pathways have led to several successful therapies, their scope is limited, and many promising drug candidates fail due to a lack of understanding of causal mechanisms.[13] Additionally, the reliance on biomarkers that may not be causally linked to CVD further reduces the likelihood of success in drug development.

To address these challenges, innovative and complementary methods must be systematically integrated across the drug development pipeline. Advances in human genetics, particularly in the post-genome-wide association study (GWAS) era, have begun to provide powerful tools for elucidating causal pathways and identifying novel therapeutic targets. For example, Mendelian randomization (MR) offers a framework for leveraging genetic variants as proxies to infer causal relationships between exposures (e.g., biomarkers or modifiable risk factors) and disease outcomes.[6,14] MR, which shares conceptual similarities with randomized controlled trials (RCTs), provides an efficient and robust method for prioritizing drug targets, thereby increasing the likelihood of clinical trial success.[15] Similarly, multivariate GWAS approaches, such as Genomic structural equation modeling (GenomicSEM), enable the simultaneous evaluation of shared genetic architecture across related traits and comorbid diseases, facilitating the discovery of novel loci and pathways.[16]

The integration of genetics into CVD drug development is not merely theoretical; its potential is exemplified by the success of lipid-lowering therapies targeting proprotein convertase subtilisin/kexin 9 (PCSK9) and angiopoietin-like protein 3 (ANGPTL3). These therapies, which were initially identified through genetic studies, have demonstrated efficacy in reducing low-density lipoprotein cholesterol (LDL-C) and improving cardiovascular outcomes.[17-19]

Recent advances in omics technologies, such as transcriptomics, proteomics, and metabolomics, further complement genetic approaches by providing comprehensive insights into molecular mechanisms and therapeutic targets. The increasing availability of large-scale disease data from population-based cohorts adds another dimension to this growing toolkit, enabling the systematic incorporation of human genetics into every stage of the CVD drug development pipeline.[20,21]

Nevertheless, key gaps remain in translating genetic insights into clinical applications. For instance, while lipid-lowering therapies have revolutionized CVD prevention, their impact on comorbid conditions, such as type 2 diabetes (T2D), remains incompletely understood, particularly across diverse populations. Similarly, non-conventional risk factors, such as alcohol

consumption, have been implicated in CVD but remain understudied in the context of causal inference and therapeutic intervention. Emerging evidence highlights the need for a broader, more integrative approach to address these challenges, combining cutting-edge genetics methodologies with multi-omics data to uncover novel pathways, prioritize therapeutic targets, and develop innovative interventions for CVD and its comorbid conditions.

1.2. Statement of Aims

The overarching aim of this thesis is to harness advanced genetics methodologies and complementary multi-omics approaches to uncover causal mechanisms underlying CVDs and their comorbid conditions. By addressing critical gaps in current CVD drug development pipelines, this work aims to identify novel therapeutic targets and refine strategies for prevention and treatment. Three specific aims guide the thesis, corresponding to distinct, yet interconnected, areas of research that fall within the drug development pipeline (**Figure 1.1**).

Figure 1.1. Overview of the integration human genetics (with an emphasis on drug-target Mendelian randomization) into the drug development pipeline (adapted from Holmes et al.[6]). *MR analysis serves as a powerful tool throughout the drug development pipeline, offering insights to enhance the likelihood of successful progression at each stage. By leveraging genetic evidence to establish causal links between targets and outcomes, MR can help prioritize promising therapeutic targets, optimize trial design, and identify potential safety concerns early in the process. This approach ultimately aims to increase the efficiency of drug development, reduce late-stage failures, and improve the chances of achieving regulatory approval and marketing authorization. Included in the figures are icons to contextualize the contribution of the thesis research within the broader drug development framework. Created in BioRender. Rosoff, D. (2025) https://BioRender.com/u50rbv5*

Aim 1: Investigating the Impact of PCSK9 and HMGCR Inhibition on Type 2 Diabetes Across Diverse Populations.

This aim focuses on elucidating the relationship between lipid-lowering therapies, particularly PCSK9 and HMGCR inhibition, and type 2 diabetes (T2D) risk across five global populations. By leveraging drug-target MR and multi-omics data, this project investigates long-term safety concerns of these therapies and evaluates their impact on glycemic traits. The findings aim to address gaps in understanding the molecular mechanisms of these targets, especially in non-European populations, and to mitigate health disparities by identifying population-specific therapeutic strategies.

Aim 2: Deciphering the Genetic Basis of Alcohol Consumption and its Role as a Modifiable Cardiometabolic Risk Factor.

Alcohol consumption is a significant yet understudied modifiable risk factor for CVDs. This aim integrates MR and proteomic analyses to explore the genetic basis of problematic alcohol consumption behaviors and their impact on cardiovascular outcomes. By examining cortical proteomic data and single-cell transcriptomics, this project identifies potential therapeutic targets linked to alcohol-related behaviors, including alcohol use disorder (AUD), and evaluates their biological relevance. These insights aim to inform strategies for reducing alcohol consumption as a means of mitigating cardiometabolic risk.

Aim 3: Multivariate Genome-Wide Analysis of NAFLD, T2D, and CAD to Identify Shared Genetic Architecture.

This aim employs GenomicSEM to investigate the shared genetic liability of non-alcoholic fatty liver disease (NAFLD), T2D, and coronary artery disease (CAD). By constructing a multivariate genetic factor and performing extensive annotation, this project aims to uncover novel loci, prioritize therapeutic targets, and identify proteomic mediators linking obesity with cardiometabolic outcomes. These findings are intended to inform future drug discovery and develop targeted interventions to address the intertwined burden of these highly comorbid conditions.

Together, these three aims integrate state-of-the-art genetic and omics approaches to unravel the causal mechanisms of cardiometabolic diseases and their risk factors. The work contributes to advancing precision medicine and improving strategies for CVD prevention and treatment, particularly for underserved populations and complex comorbid conditions.

Below, I expand upon the rationale for each of the projects comprising the 3 Aims in my thesis.

1.3. Aim 1. Investigating the Impact of PCSK9 and HMGCR Inhibition on Type 2 Diabetes Across Five Populations

Type 2 diabetes (T2D) affects more than 410 million people globally[22] with prevalence varying widely across geographical regions and by race/ethnicity.[23-26] Epidemiological data suggests that the risk of developing T2D varies across different ethnic populations,[26] e.g., the T2D prevalence among Hispanic populations (HISP) in the United States is almost double the prevalence of non-Hispanic whites of European (EUR) populations.[25,27] Given clear evidence of differences in T2D risk among populations, there remains a critical need for better understanding of risk factors, molecular mechanisms and treatment approaches across diverse populations. [28-32]

T2D is commonly comorbid with coronary artery disease and other CVDs: CVDs occur in approximately 32% of the T2D cases,[33] and CVDs are the leading cause of morbidity and mortality in patients with T2D.[34] Thus, CVD risk prevention via lipid-lowing therapies, such as statins and proprotein convertase subtilisin/kexin type 9 (PCSK9) inhibitors has become standard of care and is an important aspect of T2D management.[35,36] However, recent meta-analyses of randomized control trials (RCTs) and genetics-based studies have shown that statins are linked with modestly increased risk of T2D.[37-43] In contrast, the relationship between PCSK9 inhibition with monoclonal antibodies (alirocumab and evolocumab)[44] or the recently approved small interfering RNA inhibitor of hepatic PCSK9, inclisiran,[45] and T2D risk has been less clear. While analysis of RCT data failed to find evidence of an adverse impact[46,47] on T2D risk, long-term efficacy and safety data are currently not available and no data exists across populations. Several early drug-target Mendelian randomization (MR)[6,48-51] analyses, which leveraged genetic variants in or near the drug target gene locus to estimate the on-target impact proxying pharmacological modulation of that target,[6,50,51] reported that genetic variants in *PCSK9* are associated with lower low-density lipoprotein cholesterol (LDL-C) and an increase in T2D risk.[39,40,52,53] Conversely, more recent drug-target MRs have failed to replicate these PCSK9-T2D findings.[52,54] Furthermore, these previous studies focused primarily on EUR populations; the long-term impact of genetically proxied statin and PCSK9 inhibition on T2D risk in non-EUR populations remains unknown.

To address some of the possible long-term safety concerns of lipid-lowering therapies and T2D in non-EUR populations, we leveraged summary-level genome-wide association study (GWAS) data from large genomics consortia derived from Eastern (EAS), South Asian (SAS), African (AFR), HISP, and EUR populations and performed drug-target MR analyses to investigate the impact of LDL-C lowering via PCSK9 variants on T2D risk and glycemic traits (glycated hemoglobin [HbA1c]), fasting glucose levels, fasting insulin levels, and 2 hours glucose levels).[55] We compare these MR estimates with analyses proxying LDL-C lowering by the statin drug target HMG-CoA Reductase [HMGCR].[40] Lastly, we used a multi-omics approach to supplement PCSK9 instruments constructed using LDL-C levels (the primary physiological response to pharmacological PCSK9i) by leveraging recently released GWAS data on circulating PCSK9 protein levels[56] and liver *PCSK9* gene expression data[57] to more closely genetically model the mechanisms of action for the protein neutralizing effects of anti-PCSK9 monoclonal

antibodies[58] and the liver-specific *PCSK9* expression-lowering mechanism of inclisiran,[45] respectively. Since recent preclinical work suggested that pancreatic PCSK9 may control β cell LDLR expression and resultant cholesteryl ester metabolism and insulin secretion,[59] we also analyze the glycemic impact of genetically predicted pancreatic PCSK9 expression. Together, the results will inform our understanding of the long-term safety profile of PCSK9 and HMGCR inhibition in diverse populations, and the efficacy of these targets for CVD in the presence of T2D, helping mitigate the ongoing health disparities due, in part, to the underrepresentation of non-European populations in RCTs and genetics studies.[28-32]

1.4. Aim 2: Deciphering the Genetic Basis of Alcohol Consumption and its Role as a Modifiable Cardiometabolic Risk Factor

Alcohol consumption has long been recognized as a significant modifiable risk factor for CVDs. Excessive alcohol consumption, defined as >60 g/day in men and >40 g/day in women, is a well-documented contributor to the global burden of CVD.[60-62] Problematic alcohol drinking and binge patterns significantly increase the risk of adverse CVD outcomes, such as hypertension and stroke, e.g., binge drinking (which the NIAAA defines as consuming 4+ [5+] standard alcoholic drinker per occasion for women [men][63]) is associated with acute rises in blood pressure, heightened risk of myocardial infarction, and increased incidence of hemorrhagic stroke. Moreover, chronic heavy alcohol use is a primary contributor to alcoholic cardiomyopathy, characterized by structural and functional damage to the heart muscle.[60]

Despite the long-standing observational evidence suggesting that light to moderate alcohol intake may confer cardiovascular benefits,[64] MR studies have increasingly challenged this apparent protective impact of light/moderate alcohol consumption in observational studies: MR evidence has pointed to a causal relationship between alcohol consumption and increased cardiovascular risks, contradicting the protective effects reported in epidemiologic studies.[64-66] These findings indicate that even low-to-moderate alcohol intake may elevate the risk of certain conditions, such as hypertension, atrial fibrillation, and ischemic stroke, and raise questions about the validity of the J- or U-shaped associations reported in conventional epidemiological studies,[64] suggesting that the beneficial associations were due to selection bias, e.g., the “sick quitter” hypothesis, or other sources of bias and confounding.[67]

The implications of these findings are profound for public health and prevention strategies. While alcohol abstinence is a well-established goal for mitigating the risk of excessive consumption, understanding whether moderate consumption provides any tangible protective benefit for heart health could influence clinical guidelines, public health messaging, and personal decision-making. Addressing this question requires rigorous investigation beyond observational studies, such as leveraging RCTs and novel epidemiological approaches like MR studies. However, despite these efforts, clear conclusions about the causal relationship between moderate alcohol intake and CVD risk remain elusive. As a result, the development of effective strategies to reduce alcohol consumption to mitigate CVD risk remains a critical area of focus.

In addition to their impact on CVDs, alcohol use disorder (AUD) and problematic alcohol consumption represent their own major public health challenge. They are highly prevalent (e.g., one-third of U.S. adults [~85 million] reporting binge drinking)[68,69] and account for over 5% of global deaths annually,[70] highlighting the need for better prevention and intervention strategies.[71,72] Recently, there has been a shift towards using reductions in drinking quantity and heavy drinking days as study outcomes rather than complete abstinence or a clinical diagnosis of AUD,[73] including in recent clinical trials.[74,75] Currently, there are only three FDA-approved medications for AUD[9], and the condition is often undertreated.[71,72] Alcohol consumption patterns and the risk of transitioning to problematic drinking or AUD vary significantly.[76] Novel computational genetics methods like drug-target/cis-instrument (considered "drug-target MR" when applied to approved therapeutic targets or investigational targets in clinical trials). MR have been used to identify therapeutic targets for clinical trials, increasing success probability.[21,48-50] These methods also enhance understanding of molecular factors in problem drinking and AUD, aiding new therapeutic discoveries.[77-79]

Genome-wide association studies (GWAS) have identified genetic variants associated with problem drinking and AUD.[80-84] However, identifying causal genes via comprehensive multi-omic MR studies remains largely unexplored.[76] Recent MR work using large neuroimaging datasets has shown a causal role of increased cortical gray matter in reducing problematic alcohol consumption.[85] The cortex, linked to addiction through connections with the limbic reward systems and executive functions,[86,87] is crucial for therapeutic development.[86] Most drugs target proteins, and advances in proteomic sequencing have linked the cortical proteome with psychiatric and substance use disorders.[50,88-93] However, cis-instrument MR screens of the cortical proteome for alcohol consumption behaviors like binge drinking have not been performed despite distinct genetic architectures and relationships.[80,94,95] Integrating expression quantitative trait loci (eQTLs) is important for understanding genetic signals from GWAS[35,36], though studies are often cell-specific.[96-98] Existing transcriptomic studies on problematic alcohol consumption and AUD are limited to RNA from bulk brain tissue,[95] complicating gene-level analysis across brain cell types.[97,99-101]

Therefore, in Aim 2, we sought to better understand the genetic basis of alcohol consumption behaviors at the brain proteomic and single-cell transcriptomic levels by applying two-sample cis-instrument MR[48-50] to identify relationships of cortical proteins and gene expression in 8 major brain cell types[102] with AUD and problematic alcohol consumption behaviors. For identified cortical proteins and cell-type genes, we perform biological characterization and compare with other prioritization methods, including transcriptomic imputation[103] and RNA-seq and epigenome-wide association studies (EWAS) on AUD.[104,105] We then perform colocalization,[106] replication using independent alcohol-related GWAS, and a cis-MR screen of magnetic resonance imaging (MRI) cortical and subcortical structures and white matter tracts to further prioritize and characterize targets. Finally, we evaluate behavioral phenotypes to assess relationships of alcohol-related proteins and genes.

1.5. Aim 3: Multivariate Genome-Wide Analysis of Non-Alcoholic Fatty Liver Disease, Type 2 Diabetes, and Coronary Artery Disease

It is estimated that approximately 25% of the global adult population has non-alcoholic fatty liver disease (NAFLD)[107,108]—recently reclassified as metabolic dysfunction-associated steatotic liver disease (MASLD)[109]—making it the most prevalent form of chronic liver disease. The prevalence of NAFLD has increased rapidly over time and is projected to affect more than one-third of U.S. adults and become the leading cause of liver transplants in the U.S. and Europe by 2030.[110,111] NAFLD is commonly associated with other cardiometabolic disorders, including T2D and CAD. For example, a meta-analyses of ~6.85 million participants across 40 countries with NAFLD reported a pooled T2D prevalence of 28.3%,[112] while another meta-analysis of 2,224,144 patients with T2D estimated a global pooled NAFLD prevalence of 65.3%.[113] Moreover, concurrent T2D in NAFLD patients may synergistically exacerbate clinical complications and disease progression, i.e., accelerating the progression of NAFLD to more severe forms of liver disease.[114] Subclinical atherosclerosis is also common among individuals with NAFLD,[115] and patients with fatty livers are at an increased risk for cardiovascular events—even after adjusting for metabolic risk factors such as obesity and insulin resistance—suggesting an important role of NAFLD in CAD.[115,116] Emerging evidence also implicates liver disease, including NAFLD, in the development of cardiovascular dysfunction.[115,116] These epidemiological associations are also supported by Mendelian randomization (MR)[48,49] studies, which provide genetic evidence of causal relationships between NAFLD, T2D, and CAD.[117,118] However, the biological mechanisms linking NAFLD, T2D, and CAD are complex and not yet fully understood.

To advance our understanding of the genetic architecture underlying complex, related traits, multivariate GWAS approaches have been developed to leverage genetic correlations among phenotypes. These methods, such as GenomicSEM,[16] use single-phenotype GWAS summary statistics to model shared genetic liability across multiple traits. By increasing statistical power and enhancing the discovery of novel biological pathways, multivariate GWAS provides an opportunity to identify genetic loci associated with overlapping pathophysiological mechanisms. Recent applications of multivariate GWASs have successfully identified shared genetic architecture across neuropsychiatric disorders,[16] alcohol consumption behaviors,[84] and externalizing behaviors.[119] However, this approach has not yet been applied to investigate the shared genetic liability of NAFLD, T2D, and CAD.

Given the comorbidity and strong genetic correlations among NAFLD, T2D, and CAD, we hypothesized that a multivariate GWAS model could integrate the genetics of these conditions to investigate a shared cardiometabolic disease factor, hereafter referred to as the "*CM-Factor*" (for the remainder of the thesis). This approach aims to facilitate the identification of genomic loci associated with the underlying genetic architecture shared among these cardiometabolic diseases. To this end, we generated a multivariate GWAS of the *CM-Factor* using GenomicSEM[16] and the

largest publicly available univariate GWASs for NAFLD, T2D, and CAD. We extensively annotated the *CM-Factor* GWAS through fine mapping,[120] transcriptomic imputation,[103] partitioned heritability analysis,[121] and cell-type enrichment.[122]

We leveraged the drug-target/cis-instrument MR framework and conducted three separate MR studies aiming to identify and prioritize therapeutics for the genetic risk of these multimorbid cardiometabolic diseases. First, we investigated the genetic relationships between the *CM-Factor* and cardiometabolic targets in several drug classes, including, seven antidiabetic classes (e.g., glucagon-like peptide 1 receptor [GLP1R], and gastric inhibitory polypeptide receptor [GIPR]); 15 lipid-lowering classes (e.g., PCSK9, HMGCR, etc.; seven NAFLD/non-alcoholic steatohepatitis [NASH] targets that are either FDA-approved (thyroid hormone-β receptor [THRB] agonists[123]), in clinical trials, such as fibroblast growth factor 21 (FGF21) analogs,[124,125] or are candidate NAFLD targets; and five classes of FDA-approved antihypertensive drugs. Additionally, we screened ~2,500 genes within the druggable genome to further inform drug discovery efforts.[126]

Finally, the ongoing obesity epidemic presents a critical and escalating global health challenge.[127] While obesity is linked with increased risk for many diseases and adverse health conditions, cardiometabolic diseases are the leading cause of death.[128] Elucidating new biological pathways is essential for developing targeted interventions to reduce the burden of obesity and resultant cardiometabolic disease.[129] Circulating proteins can serve as diagnostic markers and potential therapeutic targets (most approved therapeutics target proteins),[50,126] and recent work has shown that increased body mass index (BMI) and obesity are linked with widespread changes in the plasma proteome.[130,131] Clarifying the link between obesity and cardiometabolic disease and elucidating the circulating proteins that mediate the impact of obesity on cardiometabolic outcomes could highlight new avenues for therapeutic interventions and potential targets to mitigate these risks. Therefore, we used a two-step MR[132] approach leveraging the genetic signature of BMI,[133] > 2,900 circulating proteins from individuals of European ancestry,[134] and our *CM-Factor*, to identify the proteomic consequences of increased BMI and then elucidate the BMI-affected proteins may mediate the impact of BMI on cardiometabolic health.

Together, the *CM-Factor* multivariate GWAS and complementary MR analyses comprising the Aim 3 project aim to characterize the shared genetics across NAFLD, T2D, and CAD, identify new genomic loci and potential causal biomarkers, and prioritize therapeutic targets, and inform future investigations into targeted interventions to reduce the morbidity and mortality associated with these highly prevalent and comorbid conditions.

1.6. Research Foundations and Motivation: Shaping the Focus of My PhD Projects

My research experiences across diverse fields, including neuropsychiatry and addiction, cardiometabolic diseases, and human aging, have directly shaped the focus of the projects included in this thesis. These experiences provided the foundation for developing and applying advanced analytical techniques to complex questions in therapeutic target evaluation. The breadth of my prior work has helped refine my ability to tackle methodological challenges and focus on the translational potential of genetic and genomic research, which is reflected in the studies included in this dissertation.

In the preceding sections, I introduced the broad motivations for my thesis research as it relates to therapeutic target prioritization in cardiometabolic health. Below, I provide more personal background and motivation for the specific topics covered in my PhD as it relates to my broader research experience to date.

I have been fortunate to be successful publishing in several disease areas, e.g., neuropsychiatry and addition, cardiometabolic diseases, and human aging, in my research immediately preceding, and during my PhD studies (31 publications, including 17 first-author publications at the time of writing [December 20th, 2024]) (**Appendix 1).** While many of these papers are not specifically presented in the thesis results chapters, they have played critical roles in my methodological and thematic development for my PhD thesis. In the following sections, I discuss how many of these manuscripts, along with my other research experiences, have informed my interests in drug-target safety/efficacy analyses in non-European ancestry comprising Aim 1, the target prioritization for alcohol use behaviors in Aim 2, and the multivariate GWAS of cardiometabolic multimorbidity comprising Aim 3.

1.6.1. Why focus on PCSK9 and HMGCR in Aim 1?

PCSK9, an important regulator of low-density lipoprotein cholesterol (LDL-C) through its role in low-density lipoprotein receptor (LDLR) (**Figure 1.2**),[135-139] has emerged as an important target for cholesterol-lowering drug development.[140,141] There are currently several US Food and Drug Administration FDA approved pharmacological approaches to PCSK9 inhibition: the first class of PCSK9 inhibitor (PCSK9i) drugs include the monoclonal antibodies alirocumab and evolocumab and the RNA interference PCSK9 drug, inclisiran, are each FDA approved to treat adults with heterozygous familial hypercholesterolemia or clinical atherosclerotic cardiovascular disease, who require additional lowering of LDL-C[142-149] to statins, which are inhibitors of thehydroxy-3-methylglutaryl coenzyme A (HMG-CoA) reductase (HMGCR) (the rate-limiting enzyme in *de novo* cholesterol synthesis) and are the standard of care for hypercholesterolemia.[150,151]

The discovery and approval of PCKS9 inhibition is a hallmark example of a genetically validated drug target.[140] PCSK9 was first identified in 2003 when it was demonstrated that gain-of-function mutations in PCSK9 led to hypercholesterolemia in humans. Later, additional work demonstrated that PCSK9 loss-of-function mutations reduced LDL-C and provided protection against heart disease.[152] Mechanistic studies further revealed that liver cells secrete PCSK9 into the bloodstream, where it binds to the LDL receptor (LDLR), regulating LDL-C metabolism.[153] Among the genetics-based work supporting the candidacy of PCSK9 were early drug-target MR work by Ference et al.[40] using genetic risk scores of variants within the PCSK9 locus to

genetically proxy PCSK9 inhibition investigate the change in risk for cardiovascular disease.[40] Earlier work had shown that HMGCR inhibition could be similarly proxied by assessing the its impact on LDL-C lowering and CAD risk[154] and Ference et al.[40] contextualized the PCSK9 findings with analyses proxying HMGCR inhibition.[40]

Collectively, these findings spurred a surge of interest in PCSK9 as a target for CVD therapy[140] resulting in the successful FDA approval several PCSK9 inhibitors,[142-149] and given the limited safety and efficacy data in non-European participants, motivated my Aim 1 projects focusing on PCSK9 and HMGCR inhibition. In this section, I detail additional background as to why I became interested in PCSK9 and HMGCR inhibition by describing my early research experiences, including both pre-PhD and early-PhD research, in my National Institutes of Health (NIH) lab, the Section on Clinical Genomics and Experimental Therapeutics (CGET; https://www.niaaa.nih.gov/research/division-intramural-clinical-and-biological-research/section-clinical-genomics-and-experimental-therapeutics) in the National Institute on Alcohol Abuse and Alcoholism (NIAAA).

CGET is a translational lab that leverages preclinical, clinical, and in silico approaches focused on genomics and epigenetics to further our understanding of the underlying pathophysiology of alcohol consumption behaviors, alcohol use disorder (AUD), and alcohol-related liver diseases with a focus on therapeutic discovery and repurposing for the treatment of AUD and its systemic sequalae.

We have a long-standing interest in PCSK9 evaluating the central nervous system impact of PCSK9 expression and PCSK9 levels and the safety, efficacy, and underlying biological consequences of PCSK9 inhibitor therapy (by modulation of PCSK9 protein levels with monoclonal antibodies. We were the first to link dysregulation of epigenetic control over PCSK9 expression with AUD in both human and preclinical models.[155] These findings lay the foundation for of translational studies within our lab comprehensively evaluating the biology of PCSK9 and PCSK9 inhibition. As a lab member of CGET since Spring 2017 (three years before starting my PhD in August 2020), I have been involved in many of our PCSK9 projects.

For example, the first manuscripts to which I contributed (pre-PhD) was an analysis looking at the impact of alcohol consumption behaviors and AUD on PCSK9 levels in cerebrospinal fluid.[156] In this paper, we leveraged a subset of the NIAAA in-patient and healthy controls sample (https://clinicalstudies.info.nih.gov/ProtocolDetails.aspx?id=2014-AA-0181) (AUD patients (N=42) versus healthy controls (N=25)) to show that PCSK9 in CSF was significantly higher at both day 5 and day 21 of inpatient care (P-value < 0.0001), with plasma levels positively correlating with CSF levels. These findings suggest that elevated PCSK9 in the brain may play a role in AUD.[156]

More broadly, these findings supported the hypothesized role of PCSK9 role in central nervous system (CNS) functioning. As my lab outlines in a 2022 review (see Bell et al. 2023[157] [I am a contributing author on this review article]), while best known for its role in the regulation of pathways related to the recycling of LDLR (**Figure 1.2**)[135-138] and its importance of cardiometabolic health, PCSK9 was actually first identified as an messenger RNA (mRNA) in

primary cerebellar neurons and originally given the name neural apoptosis-regulated convertase-1 due to its upregulation during apoptosis.[138,158]

There is a growing body of literature linking PCSK9 to important neural functions, such as neural cell apoptosis, neurogenesis, and neuroinflammation, and also disease states (e.g., Alzheimer's Disease, AUD, neural tube defects and stroke),[157] underscoring the importance of further our understanding of PCSK9's role in the CNS and also suggesting potential repurposing opportunities for PCSK9 inhibition in both neurological and psychiatric settings.

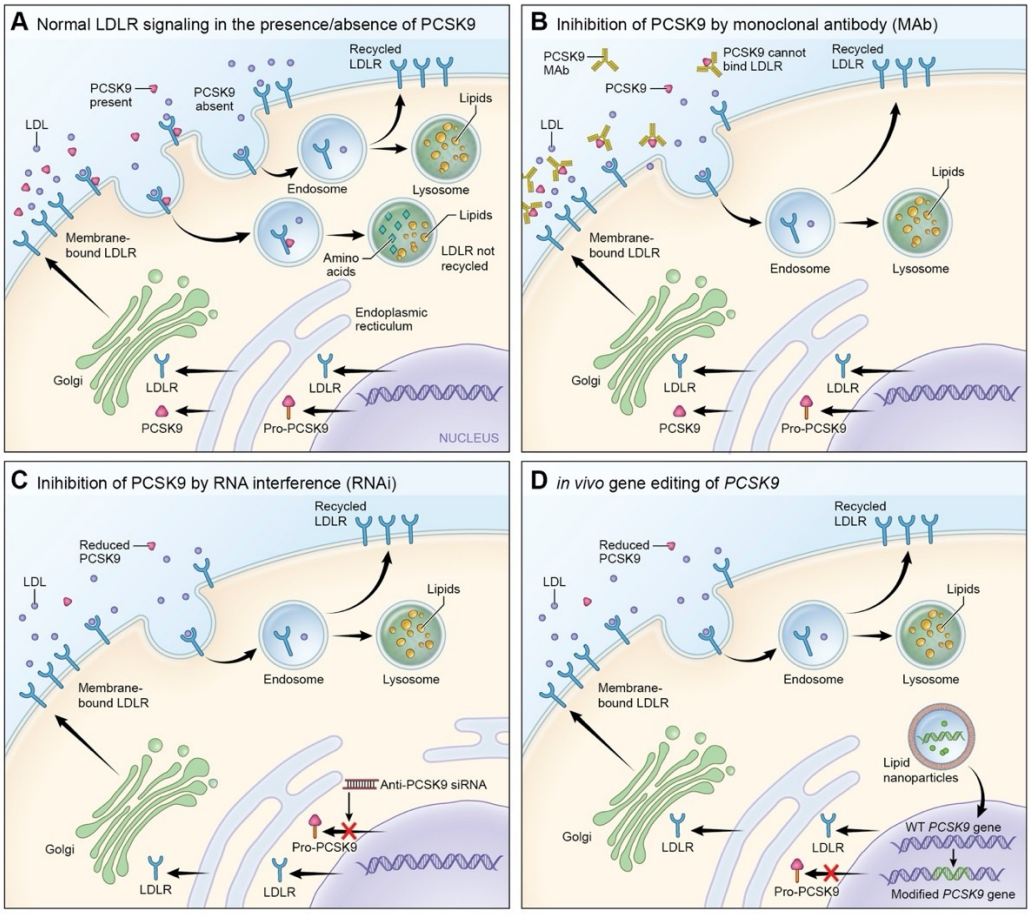


Figure 1.2. PCSK9 mechanism of action. *Figure reprinted from Bell et al. 2023.[157] Cellular mechanisms involved in PCSK9 function and inhibition by various approaches:* ***(A)*** *Standard LDLR signaling with and without PCSK9; PCSK9 directs LDLR to lysosomal degradation.* ***(B)*** *Monoclonal antibodies, like alirocumab, bind to PCSK9, blocking its interaction with LDLR.* ***(C)*** *RNA interference agents, such as inclisiran, inhibit PCSK9 mRNA translation, significantly lowering PCSK9 levels available for LDLR binding.* ***(D)*** *In vivo editing of PCSK9 can induce a PCSK9-/- genotype, effectively reducing active PCSK9 production. Reproduced from Bell AS, Wagner J, Rosoff DB, Lohoff FW. "Proprotein convertase subtilisin/kexin type 9 (PCSK9) in the central nervous system." Neurosci Biobehav Rev. 2023;149:105155. © 2023 Elsevier Ltd. Reproduced with permission.*

Given our preclinical findings, my lab is now conducting a Phase 1 study to assess the safety, tolerability, and biological effects of the PCSK9 inhibitor alirocumab in heavy drinkers who are not actively seeking treatment. The study aims to explore potential therapeutic effects on liver-related outcomes and examine alirocumab's impact on relevant biomarkers (ClinicalTrials ID: NCT04781322; https://clinicaltrials.gov/study/NCT04781322). I was involved in the design of this clinical trial, including informing study design and identifying potential safety concerns. Given that we were focusing the clinical trial on AUD patients, the involvement of PCSK9 in the CNS, and the long-standing concerns in the statin literature regarding neurocognitive impairment associated with statin therapies,[159,160] we were concerned about the potential adverse neurocognitive effects of PCSK9 inhibition on neurocognitive (i.e., cognition, memory, and neurodegeneration) and neuropsychiatric (i.e., related to mood, emotions, etc.) outcomes. Because there were limited RCT data assessing the impact of PCKS9 inhibitors on neurocognitive functioning and no long-term studies given the recency of PCSK9 inhibitor approval, we aimed to use drug-target MR[6,50,161] to provide preliminary assessment of the neurocognitive and neuropsychiatric profiles of genetically modeled PCSK9 inhibition.

Given my existing expertise in MR studies applied to a range of topics, from evaluating the impact of educational attainment and cognition one alcohol consumption behaviors,[162] and suicidal behaviors[163] to assessing the direct effects of alcohol drinking and smoking on cardiovascular health,[164] COVID-19 risk,[165] and the relationships between the genetic liabilities of pain, pain medications, and major depression,[166] I was made the lead analyst for our studies investigating the neurocognitive and neuropsychiatric profiles of PCSK9 inhibition and was responsible for study design, data analysis, and drafting the manuscripts.

These analyses resulted in two publications. In the first manuscript (see Rosoff et al. 2022[167]) (**Figure 1.3**), I focused on the neurocognitive comparison between PCSK9 and HMGCR inhibition using cis-instrumentation methods and modeling the impact of the expected physiological response to PCSK9 inhibition and statin therapies, reduced LDL-C levels,[58] on a range of neurocognitive endpoints ranging from Alzheimer's disease and Tau plaques to cortical gray matter and cognitive tests in adulthood.

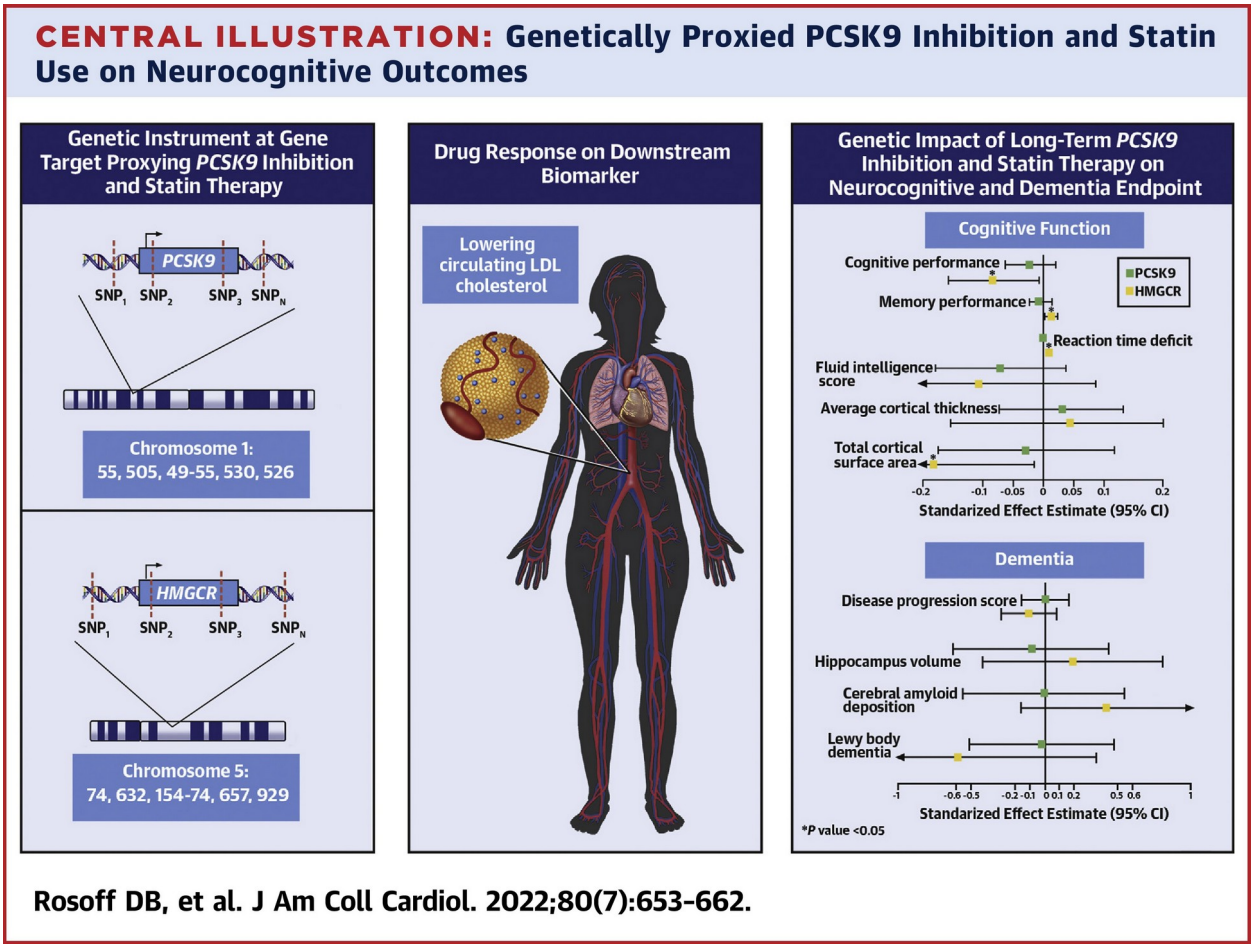


Figure 1.3. Study overview of drug-target MR assessing the neurocognitive impact of PCSK9 and HMGCR inhibition. *The figure is the Central Illustration from my 2022 first-author manuscript in the Journal of the American College of Cardiology (JACC) (Rosoff et al. 2022[167]) proxying of long-term statin therapy (via action of the drug target gene 3-hydroxy-3-methylglutaryl coenzyme A [HMG-CoA] reductase (HMGCR)) and Proprotein convertase subtilisin/kexin type 9 (PCSK9) inhibition. In this study, single nucleotide polymorphisms (SNPs) located within or near the drug target gene were extracted from a genome-wide association study (GWAS) on circulating low-density lipoprotein (LDL) cholesterol. Drug-target MR analysis was conducted for neurocognitive traits related to general cognitive function and dementia. The data presented are standardized MR effect estimates with 95% confidence intervals, representing the impact of a 1 standard deviation (38.7 mg/dL) reduction in LDL cholesterol through the drug target gene. We observed that genetic inhibition of PCSK9 had a neutral effect on cognition, with no significant impact on cognitive performance, memory, or cortical surface area. In contrast, HMGCR inhibition was linked to reduced cognitive performance, slower reaction time, and decreased cortical surface area. Reproduced from Rosoff DB, Bell AS, Jung J, Wagner J, Mavromatis LA, Lohoff FW. "Mendelian Randomization Study of PCSK9 and HMG-CoA Reductase Inhibition and Cognitive Function." J Am Coll Cardiol. 2022;80(7):653–662. © 2022 The Authors. Published by Elsevier Inc. on behalf of the American College of Cardiology Foundation, under the terms of the Creative Commons Attribution License (CC BY 4.0).*

In the second manuscript (I am the second author on this manuscript but performed all of the analyses and wrote more than half of the manuscript), I examined the impact of PCSK9 and HMGCR inhibition on symptoms of mood (i.e., depression, anxiety, irritability, and self-harm) and explored potential differences in the impact between men and women (this was motivated by the well-known sex differences in depression diagnoses, e.g., women are more than twice as likely as men to be diagnosed with depression,[168] and are present with worse depression symptom severity[169-172]). Having gained experience genetically proxying the primary expected physiological of PCSK9 inhibition, I aimed to expand the scope of my assessment of PCSK9 inhibition by incorporating tissue expression and circulating proteomics data into my study design. In addition to constructing PCSK9 instruments using LDL-C, I also leveraged available data sources for circulating protein levels, and disease-relevant tissue gene expression (**Figure 1.4**).

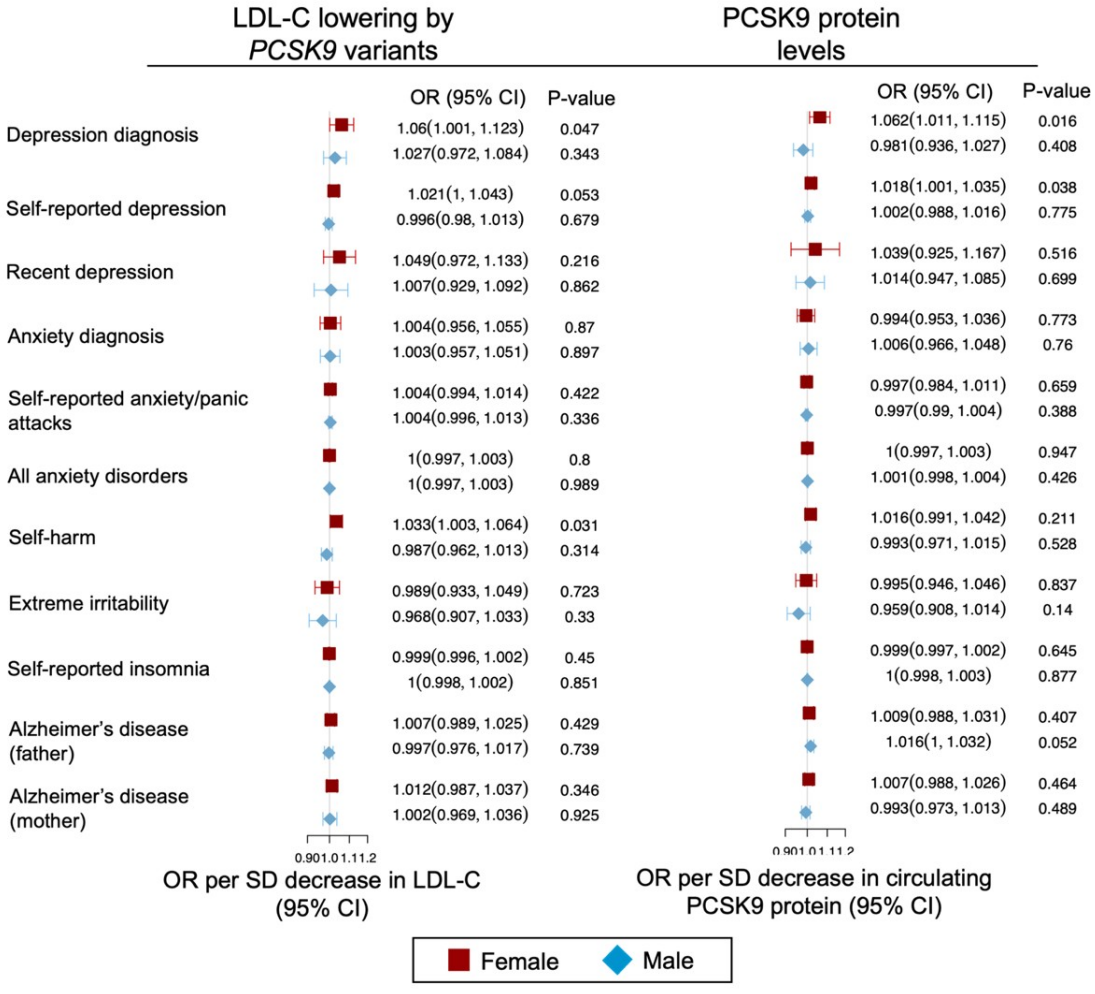


Figure 1.4. Inverse variance weighted (IVW) MR results of genetically proxied PCSK9 in circulating LDL-C and circulating protein levels on neuropsychiatric outcomes for men and

women (from Bell et al.[173] [Figure 2]). *Estimates for the LDL-C lowering impact of PCSK9 inhibition are reported odds ratios (ORs) corresponding to a change in the risk for the neuropsychiatric endpoint for a 1 standard deviation (SD) lowering of genetically proxied circulating low-density lipoprotein cholesterol (LDL-C) levels (corresponding to the primary physiological response of pharmacologic PCSK9 inhibition). For the analyses using circulating PCSK9 protein levels, the ORs correspond for a change in genetically proxied normalized circulating PCSK9 protein levels (i.e., the physiological target of monoclonal anti-PCSK9 inhibitors). Reproduced from Bell AS, Rosoff DB, Mavromatis LA, et al. "Comparing the Relationships of Genetically Proxied PCSK9 Inhibition With Mood Disorders, Cognition, and Dementia Between Men and Women: A Drug-Target Mendelian Randomization Study." J Am Heart Assoc. 2022;11:e026122. © 2022 The Authors. Published by the American Heart Association, Inc., under the terms of the Creative Commons Attribution 4.0 International License (CC BY 4.0).**

Together, these two drug-target MRs did not find evidence of adverse cognitive-related or psychiatric side effects related to inhibition of PCSK9, which did align with preliminary earlier safety assessments based on short-term clinical studies[144,174,175] and early genetic studies.[176] They also provided the experience conducting drug-target MR studies necessary for me to extend to multi-ancestry and multi-omics analyses in Aim 1 as well as the expanded drug-target/cis-instrument MR applications in Aims 2 and 3.

1.6.2. Why investigate the genetic underpinnings of alcohol consumption behaviors?

The NIAAA, which has the following mission statement:

"The mission of the National Institute on Alcohol Abuse and Alcoholism is to generate and disseminate fundamental knowledge about the adverse effects of alcohol on health and well-being, and apply that knowledge to improve diagnosis, prevention, and treatment of alcohol-related problems, including alcohol use disorder, across the lifespan"– (https://www.niaaa.nih.gov/our-work/mission-statement)

My first first-author publication (prior to my PhD) was a manuscript using conventional epidemiological approaches and the NIAAA clinical trial dataset investigate the impact of high-intensity binge drinking (defined by the NIAAA as consuming 2- and 3-times the gender-specific binge drinking thresholds of 5 drinks per occasion for men and 4 drinks per occasion for women)[68] on liver enzymes and cholesterol levels.[177]

The study was motivated by other work at the NIAAA reporting that an estimated 32 million adults in the United States reported high intensity binge drinking at least one time in the previous calendar year.[68] My main findings showed high-intensity binge drinking was associated with a dose-dependent 2- to 8-fold increased odds for clinically high levels of HDL-C, total cholesterol, triglycerides, and all liver function enzymes (gamma-glutamytransferase, aspartate aminotransferase, and alanine aminotransferase) (**Figure 1.5**). In a secondary analysis examining the impact of high-intensity binge drinking patterns accounting for the total amount of alcohol

consumed, I found each additional day of high-intensity binging also increased the odds of clinically high biomarker levels, which is alarming that even one additional day high-intensity binge drinking may increase cardiometabolic risk factor levels given that high-intensity binge drinking is common on the weekends, special events, and holidays.[5]

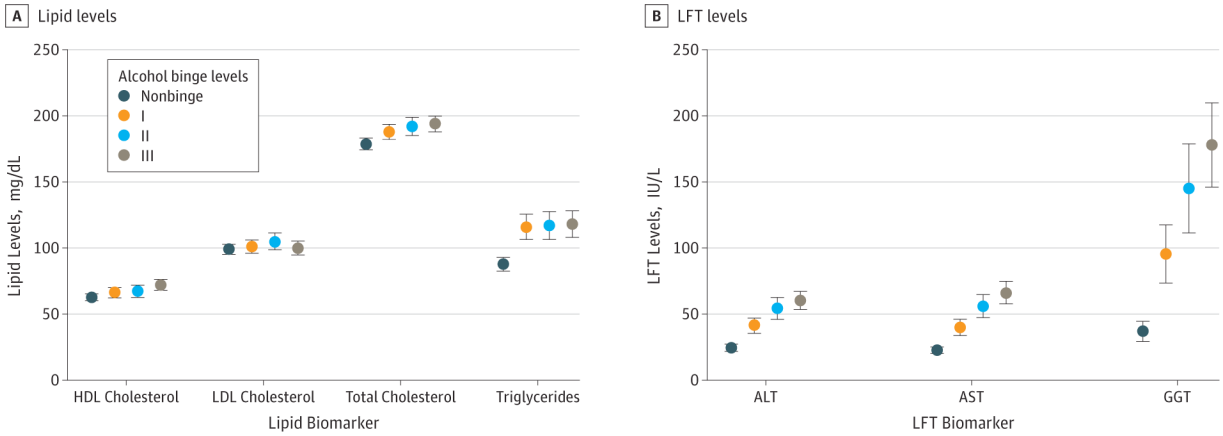


Figure 1.5. Means Levels of Lipid and Liver Function Test (LFT) Biomarkers by Alcohol Binge Levels (Figure 1 in Rosoff et al. 2019[177]). *Means Levels of Lipid and Liver Function Test (LFT) Biomarkers by Alcohol Binge Levels Error bars indicate unadjusted 95% confidence intervals; ALT, alanine aminotransferase; AST, aspartate aminotransferase; GGT, γ-glutamyltransferase; HDL, high-density lipoprotein; and LDL, low-density lipoprotein. Reproduced from Rosoff DB, Charlet K, Jung J, et al. "Association of High-Intensity Binge Drinking With Lipid and Liver Function Enzyme Levels." JAMA Netw Open.* 2019;2(6):e195844. doi:10.1001/jamanetworkopen.2019.5844. © 2019 The Authors. Published by the American Medical Association under the terms of the Creative Commons Attribution License (CC BY 4.0).*

Although alcohol consumption has been linked to alterations in cardiovascular risk factors, such as high-density lipoprotein cholesterol (HDL-C), triglycerides, and liver function enzymes, prior studies have not been able to explore these associations specifically among individuals reporting recent high-intensity binge drinking and my study added to the body of observational literature showing a complex relationship between alcohol consumption and cardiovascular disease.[178]

However, causal inferences difficult in observational studies due to potential confounding and reverse causation.[179] Additionally, in 2019, the Moderate Alcohol and Cardiovascular Health study, a large worldwide multicenter, worldwide, randomized clinical trial designed by the NIAAA to provide causal inference for the hypothesized "J-curve"[180] in the alcohol-CVD relationships by assessing the impact of consuming ~15 grams of alcohol daily (versus a control group abstaining from alcohol consumption), was stopped, due to concerns about study design and future credibility.[181] Early one-sample MR studies investigating genetic variation in the acetylaldehyde dehydrogenase (ALDH) or alcohol dehydrogenase (ADH) genes found that alcohol is associated with increased HDL-C levels,[182-184] while a seminal MR study by Millwood et al.[185] using data from 512,715 Chinese adults in the Kadoori Biobank found that alcohol increases stroke risk but has no effect on myocardial infarction (MI), possibly due to the low MI prevalence in the study population.[185]

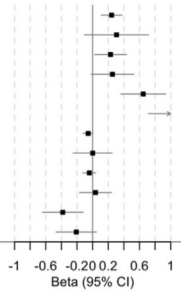
However, there is a complex relationship between alcohol consumption and smoking behaviors with clinical and observational finding them to be highly comorbid behaviors (e.g., it has been estimated that approximately 85% of smokers consume alcohol,[186] and drinkers are 75% more likely to smoke than abstainers[187]) and genetic-based studies finding strongly linked genetic architectures,[188,189] underscoring the importance of assessing the cardiovascular impact of both simultaneously.

Therefore, my lab and I were motivated to use Mendelian randomization approaches to further our understanding of direct impact cardiovascular impact of alcohol consumption and smoking behavior using multivariable MR.[190,191] This results of this study were published in first-author manuscript in PLOS Medicine entitled, "*Evaluating the relationship between alcohol consumption, tobacco use, and cardiovascular disease: A multivariable Mendelian randomization study*" (ref.[192]).

Using polygenic genetic instruments for alcohol consumption behaviors and tobacco smoking analyses, I found that in single-variable MR analyses, genetically predicted alcohol consumption and smoking were linked to an elevated risk of several CVDs, including stroke, myocardial infarction (MI), CHD, peripheral artery disease (PAD), and atrial fibrillation (AF), and systolic blood pressure. When I assessed the direct effects of alcohol consumption and smoking behaviors in multivariable MR, smoking maintained a strong association with these CVDs, while alcohol consumption remained associated with increased systolic blood pressure and increased CHD risk (**Figures 1.6 & 1.7**). The smoking-CVD relationships align with other MR studies showing adverse cardiovascular effects of smoking behaviors,[193,194] and together underscores the importance of identifying targets for heavy drinking reduction and smoking cessation as strategies to alleviate the burden of CVDs.

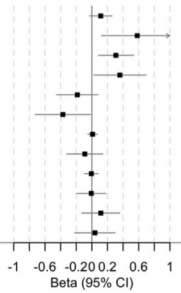
HDL-C (SD mg/dL)

Exposure	Method	Beta (95% CI)	p-value
Alcohol	SVMR IVW	0.243 (0.111,0.375)	3.15e-04
	SVMR Egger	0.306 (-0.104,0.717)	0.15
	SVMR W Median	0.23 (0.025,0.436)	0.28
	SVMR W Mode	0.254 (-0.019,0.527)	0.74
	MVMR IVW	0.648 (0.362,0.933)	8.57e-06
	MVMR Egger	1.07 (0.717,1.423)	2.85e-09
Smoking	SVMR IVW	-0.057 (-0.12,0.005)	0.07
	SVMR Egger	0.001 (-0.249,0.251)	0.99
	SVMR W Median	-0.042 (-0.125,0.041)	0.33
	SVMR W Mode	0.037 (-0.167,0.24)	0.73
	MVMR IVW	-0.382 (-0.642,-0.122)	0.004
	MVMR Egger	-0.209 (-0.467,0.048)	0.11



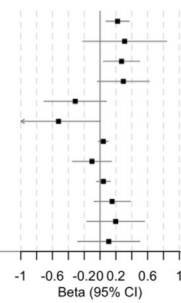
LDL-C (SD mg/dL)

Exposure	Method	Beta (95% CI)	p-value
Alcohol Use	SVMR IVW	0.114 (-0.029,0.258)	0.12
	SVMR Egger	0.582 (0.123,1.042)	0.02
	SVMR W Median	0.308 (0.084,0.533)	0.007
	SVMR W Mode	0.359 (0.024,0.693)	0.04
	MVMR IVW	-0.186 (-0.451,0.078)	0.17
	MVMR Egger	-0.371 (-0.724,-0.018)	0.04
Smoking	SVMR IVW	0.01 (-0.05,0.07)	0.69
	SVMR Egger	-0.09 (-0.32,0.14)	0.44
	SVMR W Median	-0.006 (-0.093,0.08)	0.89
	SVMR W Mode	-0.007 (-0.192,0.177)	0.94
	MVMR IVW	0.115 (-0.128,0.357)	0.35
	MVMR Egger	0.04 (-0.219,0.298)	0.76



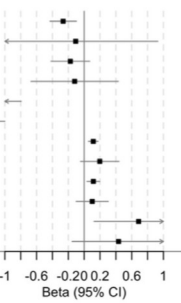
Total-C (SD mg/dL)

Exposure	Method	Beta (95% CI)	p-value
Alcohol	SVMR IVW	0.221 (0.075,0.367)	0.003
	SVMR Egger	0.31 (-0.219,0.84)	0.26
	SVMR W Median	0.271 (0.041,0.501)	0.02
	SVMR W Mode	0.294 (-0.034,0.622)	0.09
	MVMR IVW	-0.313 (-0.706,0.08)	0.12
	MVMR Egger	-0.524 (-1.053,0.006)	0.05
Smoking	SVMR IVW	0.041 (-0.021,0.103)	0.20
	SVMR Egger	-0.102 (-0.348,0.144)	0.42
	SVMR W Median	0.042 (-0.045,0.128)	0.35
	SVMR W Mode	0.152 (-0.076,0.38)	0.20
	MVMR IVW	0.197 (-0.166,0.56)	0.29
	MVMR Egger	0.11 (-0.281,0.5)	0.58



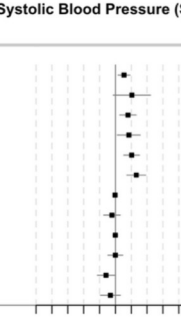
Triglycerides (SD mg/dL)

Exposure	Method	Beta (95% CI)	P-value
Alcohol	SVMR IVW	-0.265 (-0.431,-0.098)	0.002
	SVMR Egger	-0.107 (-1.142,0.929)	0.84
	SVMR W Median	-0.174 (-0.418,0.07)	0.16
	SVMR W Mode	-0.12 (-0.671,0.431)	0.67
	MVMR IVW	-1.411 (-2.028,-0.794)	7.35e-06
	MVMR Egger	-2.036 (-2.842,-1.229)	7.57e-07
Smoking	SVMR IVW	0.115 (0.054,0.176)	2.22e-04
	SVMR Egger	0.197 (-0.046,0.439)	0.12
	SVMR W Median	0.116 (0.036,0.196)	0.004
	SVMR W Mode	0.102 (-0.101,0.304)	0.33
	MVMR IVW	0.687 (0.125,1.249)	0.02
	MVMR Egger	0.434 (-0.152,1.02)	0.15



Systolic Blood Pressure (SD mmHg)

Exposure	Method	Beta (95% CI)	p-value
Alcohol	SVMR IVW	0.108 (0.033,0.182)	0.005
	SVMR Egger	0.208 (-0.028,0.443)	0.09
	SVMR W Median	0.158 (0.054,0.263)	0.003
	SVMR W Mode	0.169 (0.023,0.314)	0.03
	MVMR IVW	0.205 (0.104,0.305)	6.81e-05
	MVMR Egger	0.264 (0.143,0.384)	1.80e-05
Smoking	SVMR IVW	-0.005 (-0.031,0.021)	0.70
	SVMR Egger	-0.044 (-0.148,0.06)	0.41
	SVMR W Median	-0.002 (-0.031,0.027)	0.90
	SVMR W Mode	-0.001 (-0.095,0.092)	0.96
	MVMR IVW	-0.119 (-0.227,-0.01)	0.03
	MVMR Egger	-0.063 (-0.188,0.062)	0.32



Diastolic Blood Pressure (SD mmHg)

Exposure	Method	Beta (95% CI)	p-value
Alcohol	SVMR IVW	0.091 (0.027,0.155)	0.10
	SVMR Egger	0.131 (0.022,0.24)	0.20
	SVMR W Median	0.106 (0.029,0.182)	0.007
	SVMR W Mode	0.11 (0.041,0.179)	0.003
	MVMR IVW	0.154 (0.05,0.258)	0.004
	MVMR Egger	0.198 (0.073,0.323)	0.002
Smoking	SVMR IVW	0.015 (-0.011,0.041)	0.26
	SVMR Egger	-0.06 (-0.169,0.05)	0.29
	SVMR W Median	-0.014 (-0.043,0.015)	0.34
	SVMR W Mode	-0.072 (-0.166,0.022)	0.14
	MVMR IVW	0.015 (-0.097,0.127)	0.80
	MVMR Egger	0.056 (-0.073,0.186)	0.39

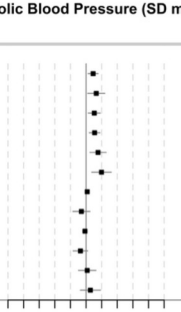


Figure 1.6. Single variable and multivariable MR estimates genetically predicted alcohol consumption and smoking with CVD risk factors (Rosoff et al. 2020[192] ***[Figure 2]).*** *For alcohol, the estimates are reported per 1 standard deviation (SD) increase in log-transformed weekly alcohol consumption (with MVMR adjusting for smoking). For smoking, the estimates are expressed per 1 SD increase in the log odds of being a regular smoker (with MVMR adjusting for alcohol consumption). Abbreviations: CI, confidence interval; CVD, cardiovascular disease; HDL-C, high-density lipoprotein cholesterol; IVW, inverse variance weighted; LDL-C, low-density lipoprotein cholesterol; MVMR, multivariable Mendelian randomization; SD, standard deviation; SNP, single nucleotide polymorphism. Reproduced from Rosoff DB, Davey Smith G, Mehta N, Clarke T-K, Lohoff FW. "Evaluating the relationship between alcohol consumption, tobacco use, and cardiovascular disease: A multivariable Mendelian randomization study." PLoS Med. 2020; 17(12): e1003410. doi:10.1371/journal.pmed.1003410. This article is distributed under the terms of the Creative Commons CC0 Public Domain Dedication.*

Figure 1.7. Single variable and multivariable MR estimates genetically predicted alcohol consumption and smoking with CVDs (Rosoff et al. 2020[192] ***[Figure 3]).*** *The estimates from SVMR, IVW MVMR, and MR-Egger are presented as odds ratios (ORs). For alcohol, the ORs represent the effect per 1 standard deviation (SD) increase in log-transformed weekly alcohol consumption (with MVMR adjusting for smoking). For smoking, the ORs reflect the effect per 1 SD increase in the log odds of being a regular smoker (with MVMR adjusting for alcohol consumption). Abbreviations: CI, confidence interval; CVD, cardiovascular disease; IVW, inverse variance weighted; MVMR, multivariable Mendelian randomization; OR, odds ratio; SD, standard deviation; SNP, single nucleotide polymorphism; SVMR, single-variable Mendelian randomization. Reproduced from Rosoff DB, Davey Smith G, Mehta N, Clarke T-K,*

Lohoff FW. "Evaluating the relationship between alcohol consumption, tobacco use, and cardiovascular disease: A multivariable Mendelian randomization study." PLoS Med. 2020; 17(12): e1003410. doi:10.1371/journal.pmed.1003410.

These findings were supported and extended by more recent work of mine clarifying the direct and total effects of substance use behaviors and major psychiatric illness on healthy aging and epigenetic age acceleration.[195] In Rosoff & Hamandi et al., published in JAMA Psychiatry in June 2024,[195] I used multivariable MR methods to clarify the direct and total effects of substance use behaviors and major psychiatric illness on healthy aging and epigenetic age acceleration. After extensive multivariable MR and sensitivity tests I showed that the direct effects of major psychiatric disorders were most strongly mediated by smoking behaviors (and to a degree alcohol consumption; however, the AUD data was underpowered).[195] As part of this manuscript, I used drug-target MR and brain proteomics to identify 135 cortical proteins significantly associated with smoking behavior, including 27 proteins with evidence of colocalization. Among these, *AKT3*, *LY6H*, and *RIT2* emerged as top candidates for therapeutic development, offering promising targets for smoking cessation interventions. *LY6H* modulates nicotinic acetylcholine receptor activity and showed beneficial neuropsychiatric profiles in phenome-wide MR (Phe-MR), indicating potential dual benefits for smoking cessation and mood regulation.[195]

These analyses motivated a similar, standalone study integrating cortical brain proteomics data (which expand into the single cell transcriptome of canonical brain cell types) with harmful alcohol consumption that is presented as the main study in Aim 2.

1.6.3. Why explore the genetic links between T2D, CAD, and NAFLD with multivariate GWAS methods?

Given the complex interplay between NAFLD, T2D, and CAD, which share overlapping genetic, metabolic, and pathophysiological pathways, it is important to use methods capable of capturing their shared genetic architecture and I gained experience using the recently developed multivariate GWAS method called Genomic Structural Equation Modeling (GenomicSEM)[16] in the context of aging-related traits, as demonstrated in the construction and analysis of the multivariate genetic factor reflecting aging in a first-author publication entitled, "*Multivariate genome-wide analysis of aging-related traits identifies novel loci and new drug targets for healthy aging*," published in 2023 in *Nature Aging*,[196] highlights the utility of this approach in disentangling the genetic underpinnings of interrelated phenotypes (**Figure 1.8).** The ability of GenomicSEM to integrate multiple traits into a single multivariate framework makes it ideally suited for investigating polygenic liability across related traits, enhancing discovery power, and addressing limitations inherent in single-trait GWAS approaches.[16]

Figure 1.8. Multivariate aging GWAS modeled with GenomicSEM (Figure 2 from Rosoff et al. 2023,[196] published in Nature Aging). A *Genetic correlations for structural equation modeling with genomic SEM, displaying pairwise LD Score genetic correlation estimates for the five univariate phenotypes.* ***B*** *Path diagram of the common factor model estimated with genomic SEM, with standardized factor loadings, (standard error in parentheses).* ***C*** *Manhattan plot showing SNP associations ($-\log_{10}$(P-value)) with mvAge, ordered by chromosome. The red dashed line indicates the threshold for conventional genome-wide significance (P-value = 5×10^{-8}). P-values are derived from two-sided Wald tests for each SNP on mvAge. "*" indicates that summary statistics for frailty and PhenoAge (the epigenetic clock variable) were reversed to align with the other longevity-related endpoints. SEM, structural equation modeling. Reproduced from Rosoff DB, Mavromatis LA, Bell AS, Wagner J, Jung J, Marioni RE, Davey*

*Smith G, Horvath S, Lohoff FW. "Multivariate genome-wide analysis of aging-related traits identifies novel loci and new drug targets for healthy aging." Nat Aging 2023; 3: 1020–1035. © 2023 The Authors, licensed under Creative Commons Attribution 4.0 International (CC BY 4.0).**

GenomicSEM was able to model the shared genetic liabilities among diverse phenotypes such as healthspan, lifespan, frailty, and epigenetic aging. These phenotypes, like NAFLD, T2D, and CAD, are highly correlated and reflect interconnected biological processes. GenomicSEM allowed for the identification of novel genetic variants and loci by leveraging genetic correlations among the input GWASs, substantially increasing effective sample size and statistical power. Moreover, this approach enabled insights into the underlying pathways and biological mechanisms driving these traits, providing a robust platform for further bio-annotation and prioritization of potential therapeutic targets. For the analysis of NAFLD, T2D, and CAD, GenomicSEM offers several advantages:

1. **Capturing Shared Genetic Architecture**: Like aging traits, NAFLD, T2D, and CAD are highly interconnected, with overlapping genetic risk factors contributing to their development and progression. The capacity of GenomicSEM to model their shared genetic liability facilitates a comprehensive understanding of the common biological underpinnings.

2. **Enhanced Statistical Power**: By integrating GWAS summary statistics from multiple traits, GenomicSEM boosts statistical power, enabling the discovery of genetic loci and pathways that may not reach significance in univariate analyses, which is especially important for the NAFLD GWAS that have been limited, to date, but small case counts in the available cohorts.[107]

3. **Identification of Trait-Specific and Shared Pathways**: The ability to decompose genetic variance into shared and trait-specific components provides a nuanced understanding of how certain loci and pathways uniquely or jointly contribute to the phenotypes.

4. **Relevance to Multisystem Diseases**: Like aging, NAFLD, T2D, and CAD involve systemic effects and cross-organ interactions, making the systems-level focus of GenomicSEM particularly relevant for exploring their interconnections.

In summary, my diverse research experiences in neuropsychiatry, cardiometabolic health, and human aging have provided a robust foundation for the studies in this thesis. These experiences honed my skills in leveraging genetic and genomic data to tackle complex questions in therapeutic target prioritization and shaped my focus on the translational applications of advanced analytical techniques. This dissertation reflects the integration of these experiences, connecting prior work with the central aims of investigating drug-target safety and efficacy, exploring the genetic underpinnings of alcohol-related behaviors and cardiometabolic health, and applying multivariate methods to disentangle the shared genetic architecture of related diseases. Together, these projects underscore my commitment to advancing the understanding of genetic

and molecular mechanisms underlying human health and disease, with a focus on actionable insights for therapeutic development.

CHAPTER 2: MATERIALS & METHODS

Chapter Overview

Chapter 2 outlines the methodological frameworks and data sources underpinning the analyses presented in this thesis, focusing primarily on MR techniques. It provides a comprehensive overview of MR, its foundational assumptions (relevance, independence, and exclusion restriction), and applications, including conventional MR for identifying causal relationships between biomarkers and disease risk and drug-target MR for therapeutic target validation. The chapter also introduces complementary methods, such as transcriptomic imputation, colocalization, and GenomicSEM, utilized across the specific thesis aims.

Key methodologies include constructing genetic instruments for lipid-lowering targets (e.g., PCSK9, HMGCR), glycemic traits, and cardiometabolic outcomes using population-specific GWAS data. Detailed drug-target MR analyses evaluated the causal impact of lipid-lowering therapies on type 2 diabetes and related glycemic markers. In parallel, cis-instrument MR screens prioritized brain-specific proteins, cell-type-specific gene expression, and their associations with alcohol use behaviors and psychiatric outcomes.

Subsequent sensitivity analyses addressed pleiotropy and robustness, integrating multi-modal data such as neuroimaging, proteomics, and transcriptomics to contextualize findings biologically. Multivariate GWAS approaches, exemplified by the *CM-Factor* integrating shared genetic liability for cardiometabolic traits, further expanded causal inference applications. This multi-disciplinary methodology bridges causal inference, genetic epidemiology, and therapeutic innovation.

2.1. Mendelian Randomization Background

Mendelian Randomization (MR) forms the foundation of the projects presented in this thesis. This section provides an overview of the principles and applications of MR, with a specific focus on its use in drug-target MR for prioritizing therapeutic targets and evaluating their causal effects.[48-50] Other methods, such as transcriptomic imputation[103] and Genomic Structural Equation Modeling (GenomicSEM),[16] are presented in the sections that outline the methods used for the specific Aims of this thesis.

MR is an instrumental variable analysis method used to test causal hypotheses in observational data.[48,197] In MR, genetic variants—typically single nucleotide polymorphisms (SNPs)—serve as instrumental variables for a putative risk factor. The approach is rooted in Mendel's second law

of independent assortment, which ensures the random segregation of alleles during gamete formation, analogous to the random assignment in a randomized controlled trial (RCT). This randomization minimizes confounding and allows causal inference.[48]

In an RCT, participants are randomly assigned to receive either an intervention or a placebo, and any observed differences in outcomes are attributable to the intervention. Similarly, in MR, individuals are effectively "randomized" based on their inherited genetic variants.[48] MR reduces the risk of reverse causation because genetic variants are fixed at conception and are not influenced by disease status, providing robust insights into causality in non-experimental settings.[48]

2.1.1. Mendelian randomization assumptions

MR analyses are subject to the core instrumental variable assumptions. (**Figure 2.1**): (1) relevance, (2) independence, and (3) exclusion restriction.[198] The relevance assumption assumes that genetic variants used as instruments are strongly associated with the exposure. The independence assumption necessitates that these variants are independent of confounders of the exposure-outcome relationship. Finally, the exclusion restriction criterion assumption ensures that the variants affect the outcome solely through the exposure.[198]

Figure 2.1. Mendelian randomization model overview (directed acyclic graph adapted from refs.[48,199]). *Mendelian Randomization (MR) is based on three fundamental assumptions: (1) the genetic variant(s) used as an instrument (the genome-wide genetic variants) must be associated with the exposure of interest (X), a condition known as the relevance assumption (IV1); (2) the instrument must not be influenced by any confounders, whether measured or unmeasured, that also affect the outcome (Y), referred to as the independence assumption (IV2); and (3) the instrument can only affect the outcome (Y) through its effect on the exposure (X), with no alternative causal pathways, an assumption known as the exclusion restriction or the no*

horizontal pleiotropy assumption (IV3). The dotted lines indicate that MR assumes no relationship. Created in BioRender. Rosoff, D. (2025) https://BioRender.com/2syn5uk

2.1.2. Drug-target/cis-instrument MR

As outlined by Holmes et al.,[6] drug-target MR and MR used to assess causal relationships of biomarkers and risk factors serve different purposes, reflecting fundamental differences in their genetic architectures and applications.[6] Conventional MR focuses on complex traits—such as blood pressure, high density lipoprotein cholesterol [HDL-C], or body mass index— that are typically polygenic, meaning their genetic associations arise from many SNPs located throughout the genome. These variants often represent diverse and overlapping biological pathways. By contrast, drug-target MR focuses on specific encoded target (e.g., protein levels, gene expression, or the primary expected physiological response to pharmacological modulation) targeted by therapeutic interventions, which are usually influenced by cis-acting genetic variants near the gene encoding the protein or target of interest, which allows drug-target MR to act as a closer proxy for therapeutic intervention by mimicking the intended mechanism of action of a drug.[6,50] Drug-target/cis-instrument MR has analogies to RCTs modeling therapeutic targets of interest (**Figure 2.2**) and has been both used to both predict efficacy of known drug targets for a range of disease outcomes,[50] also screen for new therapeutic targets across many tissues (see refs.[50,91,200]).

One key distinction lies in the interpretation of results between conventional and drug-target/cis-instrument MR analyses: In conventional MR, the aim is to establish the causal role of a biomarker in disease etiology and evaluate its potential as a modifiable risk factor. For example, MR might evaluate whether lowering LDL cholesterol causally reduces the risk of coronary artery disease. Conversely, drug-target MR evaluates whether a specific drug target (e.g., a protein like PCSK9, HMGCR, or CETP) has the intended therapeutic effect (i.e., lowering LDL-C levels). Even when a biomarker is shown to be causally linked to a disease, the drug targeting that biomarker may have additional effects (target-mediated pleiotropy) that differ from those of the biomarker itself, e.g., CETP inhibitors raise HDL-C but reduce coronary artery disease risk through effects on apolipoprotein B, highlighting how drug effects can diverge from biomarker associations.[50]

Additionally, drug-target MR often involves fewer genetic variants (typically cis-acting variants near the target gene),[50] reducing the risk of horizontal pleiotropy but potentially limiting statistical power. Conventional MR of biomarkers of risk factors, on the other hand, leverages large numbers of variants to capture the full genetic architecture of a trait, which provides higher power but introduces a greater risk of confounding from pleiotropic pathways. Therefore, while both biomarker and drug-target MR provide insights into disease mechanisms, their objectives, methodologies, and interpretations differ significantly. Conventional MR informs public health interventions and identifies modifiable risk factors, whereas drug-target MR directly informs drug development by evaluating the likely success of a therapeutic intervention targeting a specific protein. These differences highlight the complementary roles of the two approaches in advancing precision medicine and therapeutic innovation.[50]

Figure 2.2. Comparison of randomized controlled trial (RCT) and drug-target Mendelian Randomization. *This figure illustrates the conceptual similarities and differences between RCTs and MR in assessing the causal impact of lowering low-density lipoprotein cholesterol (LDL-C) on disease outcomes (here coronary artery disease [CAD])* ***(a)*** *RCT Approach: Participants are randomized to receive either a PCSK9 inhibitor (e.g., alirocumab) or a placebo, ensuring balanced environmental confounders between groups. The intervention reduces LDL-C levels in the treatment group, while the control group maintains normal levels. The outcome (e.g., CAD incidence) is then compared between the groups to evaluate the effect of the treatment.* ***(b)*** *Drug-target MR approach: Genetic variants in the PCSK9 locus associated with lower LDL-C levels (e.g.,* SNP_1*,* SNP_2*) act as instrumental variables. Individuals carrying these variants are compared to those with reference (i.e., the wild-type) alleles. Because genetic variation is randomly inherited and independent of confounding factors, the resulting analysis approximates the effect of life-long lower LDL-C levels on CAD risk. Created in BioRender. Rosoff, D. (2025) https://BioRender.com/l3zvjfn*

2.2. Aim 1 Methods

Figure 2.3 presents an overview of the Aim 1 analyses.

2.2.1. Data sources

For our exposure LDL-C data, we used GWASs of LDL-C levels from the 2021 Global Lipid Genetics Consortium (GLGC) LDL-C population-specific meta-analyses of AFR (N≤ 94,623), EAS (N≤ 82,587), SAS (N≤ 40,472), HISP (N≤ 46,039), and EUR populations (N≤ 1,320,016).[201] For T2D diagnoses in EAS, SAS, and EUR populations, we used the 2022 DIAMANTE population-specific meta-analyses (EAS N=433,540; SAS N=49,492; EUR N=898,132).[202] HISP and AFR data were not available in the DIAMANTE cohorts. For HISP, we used the T2D results (N=10,106) from the Population Architecture using Genomics and Epidemiology (PAGE) study.[203] Because a recent MR evaluated the impact of lipids and lipid-lowering drug-targets on T2D in AFR populations using the results from the MVP cohort,[54] we used another, independent AFR GWAS of T2D (N=4,347).[204]

We used the recent GWASs for HbA1c, fasting glucose levels, fasting insulin levels, and 2-hour glucose levels in the these populations from the MAGIC (Meta-Analyses of Glucose and Insulin-related traits) Consortium[205] to evaluate the glycemic impact of lipid-lowering targets across the 5 populations. 2-hour glucose data was not available for SAS. We also assessed the impact of lipid-lowering targets on insulin-stimulated glucose uptake in EUR cohorts using GWASs of Modified Stumvoll insulin sensitivity index (ISI) and insulin fold change IFC, respectively.[206] See **Table AP2.1** for additional information on the GWAS data used in the Aim 1 analyses.

Figure 2.3. Aim 1 analysis overview. *Presented are details outlining instrument selection, data sources, and analysis plan. Top panel describes populations included in the study and the countries of origin for each dataset (the stars reflect the approximate geographical locations of the datasets included in the publicly available GWAS data). We constructed genetic instruments for Proprotein convertase subtilisin/kexin type 9 (PCSK9) and (3-hydroxy-3-methylglutaryl coenzyme A reductase (HMGCR) extracting variants at the gene target locus (±100 kilobases [kb]) from population-specific summary-level genome-wide association study (GWAS) data of circulating low-density lipoprotein (LDL) cholesterol levels (2021 Global Lipid Genetics*

Consortium (GLGC) meta-analysis GWAS for EAS, SAS, AFR, HISP, and EUR populations). For PCSK9, we also constructed alternate instruments comprised of previously identified functional variants (R46L, E670G). Similarly, we constructed polygenic LDL-C instruments using conventionally genome-wide statistically significant ($P < 5\times10^{-8}$) variants across the genome that were conditionally independent at linkage disequilibrium (LD) $R^2 < 0.001$. We obtained GWAS summary statistics for type 2 diabetes (T2D) and glycemic markers from each population and harmonized the exposure and outcome before performing Mendelian randomization (MR). For the drug-target MR genetically proxying LDL-C lowering via the PCSK9 and HMGCR loci, we used the inverse-variance weighted random-effects (IVW) method accounting for the correlation between the genetic variants, for 2+ SNP instruments, and for single SNP instruments, the Wald ratio, as main methods. We performed colocalization under the single variant and multiple variant models for exposure-outcome pairs that had MR estimates with P-values < 0.05. PCSK9: Proprotein-convertase subtilisin/kexin 9; HMGCR; 3-hydroxy-3-methylglutaryl coenzyme A reductase; T2D: type 2 diabetes; HbA1c: glycated hemoglobin; MR: Mendelian randomization; LDL-C: low-density lipoprotein cholesterol; IVW: inverse variance weighted; SNP: single nucleotide polymorphism. Created in BioRender. Rosoff, D. (2025) https://BioRender.com/hd6tvqb.

2.2.2. PCSK9, HMGCR, and polygenic LDL-C instruments

To construct the PCSK9 and HMGCR instruments we selected genetic variants within 100 kilobases (kb) on either side of gene boundaries that were associated with LDL-C levels (at conventional genome-wide significance P-value $< 5\times10^{-8}$) to proxy the primary physiological response pharmacological inhibition of these targets.[58] We clumped the *PCSK9* and *HMGCR* variants at linkage disequilibrium (LD) $r^2 \leq 0.2$ using a 250 kb window, and the respective population-specific 1000 Genomes Project reference panels (1000G).[207] That is, for each population-specific analysis, we applied the corresponding reference panels from the 1000G: the EAS panel for East Asian analyses, the SAS panel for South Asian analyses, the AFR panel for African analyses, the HISP panel for Hispanic analyses, and the EUR panel for European analyses. For PCSK9, we also created instruments comprised of only functional variants (the gain-of-function R46L [rs505151][47] and the loss-of-function E670G [rs11591147][208]). Both E670G and R46L were available for analysis in SAS, AFR, HISP, and EUR populations, whereas only R46L was available in the EAS population because E670G was not present in the EAS LDL-C GWAS data. Detailed information for each drug-target instrument is shown in **Tables AP2.2** and **AP2.3**.

Change in LDL-C levels is the primary biomarker measured to assess the physiological response to PCSK9 inhibition and statin therapy,[58,209] and dyslipidemia has also been associated with T2D risk.[205] Therefore, we investigated the relationships of circulating LDL-C and both T2D and glycemic traits in the five populations using polygenic LDL-C instruments. For the polygenic LDL-C instruments, we identified variants associated in respective population-specific 2021 GLGC GWASs of LDL-C levels at conventional genome-wide significance (P-value $< 5\times10^{-8}$) located throughout the genome. We clumped the variants at LD $r^2 \leq 0.001$ (within a 10,000 kb window) using the appropriate 1000G reference panel.[207]

2.2.3. PCSK9 protein and PCSK9 expression data

In a separate set of drug-target analyses to further explore the potential impact of PCKS9 inhibition on T2D and glycemic traits, we performed a multi-omics drug-target MR analysis genetically modeling PCSK9 inhibition leveraging both expression quantitative trait loci (eQTL) from Genotype-Tissue Expression version 8 (GTEx v8)[57] and protein quantitative loci (pQTL) from deCODE (N=35,559).[56] We constructed two tissue-specific PCSK9 expression instruments: one using PCSK9 eQTLs in liver tissue (N=178) and another using PCSK9 eQTLs in pancreatic tissue (N=243) from postmortem data (EUR).[57] Cis-PCSK9 variants were extracted and clumped per the same criteria used constructing the PCSK9 instruments derived from LDL-C data (i.e., variants within ±100kb clumped at LD $r^2 < 0.2$ using a 250kb window and the European 1000G reference panel[207]). PCSK9 protein levels were measured in normalized protein units.[56] eQTL data is measured in transcripts per million (TPM).[210,211] pQTL and eQTL data for circulating PCSK9 and tissue-specific *PCSK9* expression were not available for non-EUR cohorts. Therefore, these analyses are limited to the T2D and glycemic markers in the EUR cohorts.

2.2.4. Statistical analysis

To adhere to the relevance assumption, we assessed the strength of the population-specific instruments by calculating F-statistics and R^2 (variance in trait explained by the variant) for each variant comprising the population-specific drug target instruments,[48,212] retaining instruments of sufficient strength (by convention, F-statistic exceeding 10). Complementary MR methods along with alternate instruments facilitate assessment of our adherence to exchangeability and exclusion restriction assumptions.

For drug target instruments with a single variant, we used the Wald ratio method.[213] For instruments with 2 or more variants, we performed MR using the inverse variance weighted (IVW) MR, MR Egger, and Maximum Likelihood methods[213] accounting for the LD between variants by incorporating correlation matrices generated using the 1000 Genomes Project data to both assess evidence for relationship of PCSK9 and HMGCR with the outcomes, and also evaluate potential violations of the MR assumptions[214]—consistency of estimates across these MR methods suggests an unbiased estimate.[48,49,215] For the polygenic LDL-C analyses using instruments clumped at LD $r^2 \leq 0.001$ (and also the PCSK9 instruments comprised of the R46L and E670G functional variants), we performed MR IVW as the main method for instruments with 2+ SNPs and the Wald method for the single-SNP instruments. Again, we included several complementary MR methods (i.e., MR Egger, weighted median, and weighted mode MR) as sensitivity analyses to assess the robustness of the MR IVW results—important for strengthening causal inference[49] because these complementary MR methods help evaluate the sensitivity of the results to different patterns of violations of the MR assumptions (e.g., horizontal pleiotropy).[48] We also used the MR Egger intercept test[216] and Cochran Q heterogeneity test[217] to assess heterogeneity and the MR Steiger test to evaluate the hypothesized causal direction between circulating LDL-C levels and outcomes.[49] We used the MR LASSO method,[218] when applicable, to remove variants identified as outliers in the analyses using polygenic LDL-C instruments.

2.2.5. Sensitivity analyses: multivariable MR with body mass index

The MAGIC GWAS glycemic traits data were each adjusted for body mass index (BMI),[55] which may introduce bias into the SNP associations of the GWAS results.[219] This adjustment for a heritable covariate may, in turn, introduce biases into effect estimates of MR analyses.[220] Multivariable MR (MVMR), which builds upon conventional single-variable MR by estimating the direct impact of multiple exposures on an outcome, taking into account the influence of each exposure in relation to the others,[191] has been shown to attenuate the bias from heritable covariates by incorporating the heritable covariate used to adjust the GWAS within the MVMR analysis.[221] Therefore, to assess the robustness of the glycemic trait analyses, we performed MVMR incorporating the genetics of BMI into the PCSK9, HMGCR, and polygenic LDL-C models. GWAS BMI data was available for all populations (EAS N=256,450,[222] SAS N=8,646,[222] AFR N=6,545,[223] HISP N=56,161,[224] and EUR N=694,649[133]) and MVMR instruments for PCSK9, HMGCR, and polygenic LDL-C levels were constructed using the same instrumentation strategies as outlined for the drug-target and single variable MR methods described above. We were unable to construct an MVMR instrument for PCSK9 in the SAS population to confirm the finding with HbA1c due to the PCSK9 SAS instrument only having a single SNP and therefore were unable to test these analyses.

2.2.6. Interpretation of MR results

We report the MR 95% confidence interval estimates as odds ratios (ORs) for T2D risk and regression effect estimates for the continuous glycemic traits. We aligned the direction of the estimates with the physiological impact of PCSK9 inhibitors and statins by transforming the reported MR estimates to a standard deviation lowering in LDL-C. We advise against interpreting study findings based solely on a P-value threshold;[225] however, to account for multiple testing within each population, we used a Bonferroni-corrected P-value threshold of 0.005 (0.05/5 main outcomes and 2 drug targets per population) as a heuristic to define ‘strong evidence’ of evidence for a genetics-based relationship. We considered findings with P-values ≥ 0.005 and P-value < 0.05 as ‘weak evidence’ for a genetics-based relationship in the main findings. We used a P-value threshold of 0.05 for the MVMR sensitivity analyses. To assess consistency and robustness, we examined whether the estimates agreed in both direction and magnitude, indicated by overlapping confidence intervals, across complementary MR methods.

2.3. Aim 2 Methods

Figure 2.4 provides the Aim 2 study overview.

Figure 2.4. Aim 2 Study overview. *The first panel presents data sources and sample sizes of the genome-wide association studies (GWASs) used for problematic alcohol use (PAU) and the three alcohol consumption behaviors. The second panel outlines Stage 1, comprising the initial cis-instrument Mendelian randomization (MR) with cortical protein data (N=722 from three cortical regions: dorsolateral prefrontal cortex, orbitofrontal cortex, and parahippocampal region, and single cell eQTL data from 8 brain cell types (N=192 and the cell types are derived from bulk tissue in the prefrontal cortex, temporal cortex, and white matter tracts), sensitivity cis-instrument MR analyses, and colocalization screens. Stage 1 also used several gene mapping methods, transcriptomic imputation using the FUSION method, whole blood epigenome-wide data that identified the epigenomic signatures of AUD (compared to healthy controls), and bulk RNAseq data from postmortem brain tissue AUD patients to characterize the proteins and genes identified in the cis-instrument MR screens and determine if the cis-instrument screens identified novel biological underpinnings of problematic alcohol consumption compared to other gene prioritization methods. Stage 2 is outlined in the third and fourth panels. In Stage 2 we annotated targets with biological pathway and ontology analyses (third panel), analyzed neuroimaging traits with cis-instrument MR characterizing their neurophysiological impact. We also performed replication analyses using independent GWAS data from electronic health records related to the psychiatric and physical consequences of problematic alcohol consumption and then expanded the analyses to other neuropsychiatric outcomes to both contextualize the proteins and genes, assessing their potential neuropsychiatric side effect profiles with 65 curated to neuropsychiatric outcomes ranging from clinical diagnoses to self-reported mood and behaviors and further prioritize potential the targets for therapeutic development. AUD: alcohol use disorder; GWAS: genome-wide association study; AST: astrocytes; EXC: excitatory neurons; INH: inhibitory neurons; MIC: microglia; OLI: oligodendrocytes; OPC: oligodendrocyte precursor cells; PER: pericytes; END: endothelial cells; IVW: inverse variance weighted; LD: linkage disequilibrium; SNP: single nucleotide polymorphism. Created in BioRender. Rosoff, D. (2025) https://BioRender.com/1srvk2j*

2.3.1. Data sources

We obtained summary-level GWAS data for the four alcohol use behaviors from publicly available GWASs in populations of predominantly European ancestry. Similarly, we obtained proteomic and transcriptomic data from their respective sources (**Table AP.1**). These datasets have existing ethical permissions from their respective institutional review boards and include participant informed consent with rigorous quality control. In sections **2.3.1.1a-f**, we provide phenotypic details for all GWAS data included in the Aim 2 study.

2.3.1a. Alcohol consumption behaviors. We included 4 GWAS endpoints encompassing aspects of alcohol consumption and AUD: alcohol intake frequency (AIF), drinks per week (DPW), binge drinking (consuming six or more units of alcohol per occasion), and problematic alcohol use (PAU). AIF data were derived from UK Biobank participants of European ancestry (N=462,346).[102] The DPW summary statistics were derived from a recent meta-analysis GWAS of 29 cohorts (N=537,349) of predominantly European ancestry.[226] The phenotype is measured in log transformed DPW to prevent outliers from biasing the analyses.[226] Binge drinking (N=143,658) was assessed using Question 3 from the AUDIT questionnaire ("How often do you have more than 6 drinks on one occasion?").[223] PAU data was derived from

a meta-analysis combining a GWAS of AUD in the MVP cohort of European Americans (N=267,391, diagnosed per ICD-9/10, abuse and dependence), a GWAS of DSM-IV alcohol dependence (AD) in multiple cohorts of European ancestry (N=46,568), and a GWAS of Alcohol Use Disorders Identification Test-Problems (AUDIT-P scores, a measure of problematic from a UK Biobank cohort) (N=121,604).[82]

2.3.1b. Cortical pQTLs. The brain cortex has been shown to be important in the underlying biology of addiction for its role in cue-elicited drug response and initiation of drug abuse,[86] including alcohol use behaviors,[85,227] More generally, the neuroproteome has been used extensively to identify novel neuropsychiatric disorder mechanisms and is considered important for drug target development for these disorders.[91,228-230] Therefore, for the proteome-wide MR analysis, we used pQTL data derived from three cortical regions (dorsolateral prefrontal cortex, orbitofrontal cortex, and parahippocampal region) of 722 study participants of European descent from a recent meta-analysis of brain cortex data (7,376 proteins).[228] Sample composition, proteomic profiling, and genotyping of the cortical data has been previously described.[228]

2.3.1c. Cell-type eQTLs in 8 brain cell types. We used the recently released single-cell eQTL data from 8 brain major cell types derived from post-mortem samples of 192 individuals of European ancestry for which single cell RNA-seq and genotype data were available.[102] Bryois et al. used tissue from prefrontal cortex, temporal cortex and white matter tracts and analyzed gene expression from 7,208 protein coding genes and 10,846 non-coding genes in excitatory and inhibitory neurons (EXC and INH), astrocytes (AST), microglia (MIC), oligodendrocytes (OLI), oligodendrocyte precursor cells (OPCs), endothelial cells (END), and pericytes (PER), and found 6,108 unique genes (accounting for genes that were expressed in more than one cell type) with variants within 1 Megabase (Mb) windows around the transcription start sites that were associated with gene expression.[102]

2.3.1d. Structural MRI data for MR analyses. We aimed to further contextualize the genes associated with the genetic liabilities for PAU and alcohol consumption behaviors in the cortical proteins cell-type genes cis-instrument Mendelian randomization (MR) analyses by assessing their impact the structure and connectivity of the brain. Therefore, we investigated the relationships of the alcohol-related genes that surpassed Bonferroni correction for multiple comparisons and also demonstrated evidence of a shared causal variant with follow up colocalization[106] analyses (i.e., PP.H4 >0.7)[106] in structural MRI data in both cortical and subcortical areas. These analyses were used to both assess the neurophysiological role of the cell-type transcriptomic findings and investigate whether the non-overlapping genes identified with the PAU and the alcohol consumption behaviors GWASs demonstrated shared or distinct relationships with brain gray matter, white matter connectivity, and resting state networks.

For these analyses, we obtained GWAS data of cortical thickness and surface area from 34 regions in addition to global cortical thickness (GCT) and global cortical surface area (GCSA) from a recent Enhancing NeuroImaging Genetics through Meta-Analysis (ENIGMA) consortium GWAS of T1-weighted MRI images from 1.5-3 Tesla scans (N=33,709).[231] The MR images were processed using FreeSurfer[232] and the ENIGMA processing pipeline and the cortical regions were defined using the Desikan-Killiany atlas.[232] GCT and GCSA were the averaged results of 34 cortical regions. Grasby et al. defined GCT as the average distance between white matter and pial surfaces across both cortical hemispheres and GCSA measured at the grey-white matter

boundary.[231] For subcortical gray matter volumes, we obtained GWAS summary statistics (N≤ 33,536 participants of European ancestry for 8 subcortical structures – hippocampus, amygdala, caudate, putamen, pallidum, brainstem, thalamus, and nucleus accumbens.[233] The subcortical MR images were also processed with the ENIGMA pipeline as described for the Grasby et al. GWAS on cortical structures.[234,235] Left and right volumes were first calculated and then averaged for the mean subcortical volumes (in mm^3).[233] We extracted cis-variants comprising the instruments of the alcohol consumption related genes identified with our initial analyses. We harmonized the gene and protein exposures and outcome data and performed MR analyses as described in the main manuscript.

2.3.1e. White matter microstructure (diffusion MRI) for MR analyses.

We obtained summary GWAS data for 110 diffusion MRI (dMRI) endpoints from a recent study investigating the genetic architecture of white matter microstructural differences in 35,101 European participants.[236] For a detailed discussion of the methods used to derive the dMRI endpoints, see Zhao et al. (2020).[236] Briefly, we prepared GWAS data for the 5 tract-averaged endpoints of common diffusion tensor imaging (DTI)-based parameters – fractional anisotropy (FA), mean diffusivity (MD), axial diffusivity (AD), radial diffusivity (RD), and mode of anisotropy (MO)) – from 21 white-matter tracts and the whole brain;[236] and 105 tract-specific outcomes (21 tracts in each of the 5 DTI parameters = 105 endpoints).[236] We extracted cis-variants comprising the instruments of the alcohol consumption related genes identified with our initial analyses. We harmonized the gene and protein exposures and outcome data and performed MR analyses as described in the main manuscript

2.3.2. Cis-instrumentation of brain proteins and transcripts

Sections **2.3.2a-c** outline the use of the cis-instrumentation MR framework and instrument construction to investigate the causal relationships between brain-specific proteins, cell-type gene expression, and alcohol consumption behaviors. The analysis covered thousands of brain proteins and genes across multiple neural and non-neural cell types, harmonizing these with alcohol consumption behaviors. Causal estimates were derived using MR methods tailored to the available SNPs per instrument, with stringent thresholds for statistical significance. Sensitivity analyses further assessed the robustness of findings across alternative parameter settings. The results provide a comprehensive view of brain-specific molecular pathways linked to alcohol consumption, setting the stage for follow-up investigations in target discovery and validation.

2.3.2a. Cis-instrument construction. Based upon the cis-instrument MR framework outlined by Schmidt et al.,[50] and used extensively for target discovery and validation,[196,200,237-241] we extracted SNPs within or near the genomic coordinates (±100 kb) and given that the cis-regions comprise only small proportions of the genome, we used a relaxed P-value threshold of 5×10^{-4} for SNP selection.[50] This threshold also ensured that all SNPs included in the primary instruments had F-statistics >10 (the conventional cutoff for determining the variant is strong and will be unlikely to be subject to weak instrument bias).[242] F-statistics were used to assess the MR relevance assumption by evaluating instrument strength.[48] We filtered variants using the 1000 Genomes Project Phase 3 European reference panel[243] and clumped the

variants at linkage disequilibrium (LD) $R^2 < 0.2$ (250 kb window). This LD r^2 threshold was selected to maximize power and decrease variability while also remaining below the thresholds (LD $r^2 > \sim 0.4$) that previous simulations and applied studies have shown to result in unstable estimates.[50,244] We also constructed correlation matrices to account for SNP-SNP LD.[237] This instrumentation process resulted in 3,562 proteins that were taken forward to harmonization with each alcohol consumption behavior (**Table AP4.2**). We were able to instrument 4,604 genes in excitatory (EXC) neurons, 2,493 in inhibitory (INH) neurons, 2,406 in astrocytes (AST), 1,470 in microglia (MIC), 1,199 in endothelial cells (END), 3,174 in oligodendrocytes (OLI), 1,905 in oligodendrocyte progenitor cells (OPCs), and 617 in pericytes (PER) (**Tables AP4.3-AP4.10**).

2.3.2b. Cis-instrument MR screen. First, to assess the relevance assumption,[48] we calculated F-statistics to evaluate instrument strength for each QTL instrument. We only included SNPs with F-statistics exceeding the conventional cutoff of 10, suggesting minimal evidence for bias from weak instruments.[242] We harmonized the cell-type eQTL instruments with each alcohol consumption behavior. For QTL instruments with 2+ SNPs, we incorporated LD correlation matrices calculated with the 1000 Genomes Project European reference panel[243] between the instrument variants in the MR IVW estimator.[50,245] If the QTL instrument had 3+ SNPs, we were able to use the MR Maximum Likelihood,[213] and MR Egger method (also incorporating the LD correlation matrix into the estimators) to assess pleiotropy.[246] We also calculated the Cochran's Q heterogeneity test.[217] For QTL instruments with only 1 SNP, we used the Wald ratio method[49] to obtain the effect estimate. MR estimates correspond to an increase in the respective alcohol consumption behavior (e.g., a positive MR estimate for PAU is interpreted as evidence that increased expression of that gene is related to increased PAU risk). P-values for the MR estimates are derived from two-sided t-tests. We used Bonferroni correction for multiple comparisons for each alcohol consumption behavior analyzed as a heuristic to allow for follow-up analyses on a plausible number of findings. Some of the alcohol consumption behavior outcome GWASs did not contain all instrument SNPs (or their proxies). Therefore, these genes were not included in the analyses for that outcome. We used thresholds adjusting for the number of proteins or cell-type/genes assessed per behavior: 3,562 cortical proteins and 17,877 cell-type gene MR analyses used two-sided P-value thresholds of 1.40×10^{-5} and 2.80×10^{-6}, respectively. The four primary alcohol-related traits are continuous measures, and therefore, we report the MR estimates in the manuscript using the regression coefficients (β), 95% confidence intervals, and P-values.

2.3.2c. Cis-MR multiverse sensitivity analysis. To evaluate the robustness of results from the initial cis-instrument MR screens, cortical proteins and cell-type genes with primary MR estimates (IVW or Wald ratio) that surpassed the Bonferroni-adjusted P-value thresholds were subjected to a comprehensive "multiverse"[200,247] sensitivity analysis evaluating the robustness of the relationships across a range of instrument selection criteria and MR methods. We constructed additional cis-instruments for each cortical protein and cell-type gene using 8 combinations of LD r^2 clumping thresholds (0.001, 0.1), P-value thresholds (5×10^{-4}, 5×10^{-5}, 5×10^{-6}, and 5×10^{-8}), and the LD clumping window of 10,000 kb, including the instrumentation parameters used in conventional polygenic instrument construction in MR studies of complex traits (i.e., LD r^2 =0.001, LD window=10,000 kb, and P-value threshold=5×10^{-8}).[48] For instruments with a single variant, we used the Wald ratio.[50,245] For instruments constructed using the LD r^2 clumping threshold of 0.001 and 0.1, we used conventional two-sample MR methods (i.e., the standard IVW estimator).[48] We took forward all cortical proteins

and cell-type genes that were directionally consistent between the primary cis-instrument MR screen and the sensitivity analyses for novelty assessment, bio-annotation, colocalization, and replication.

2.3.3. Annotation of the top targets prioritized by the cis-MR screens

Sections **2.3.3a-c** outline the integrative bioinformatic approaches employed to annotate the high-confidence targets identified in our study of alcohol consumption behaviors and problems associated with alcohol use. Three key analyses were conducted to provide functional insights into the prioritized targets:

2.3.3a. Gene ontology and pathway analysis. We used EnrichR[248] to perform gene ontology and pathway enrichment analyses with SynGo,[249] Gene Ontology (GO),[250] DisGeNET,[251] KEGG,[252] and the GWAS Catalog.[253] We analyzed the proteins/genes linked with each of the alcohol consumption behaviors separately and compared their results to assess whether the proteins and genes converged to shared ontologies and pathways.

2.3.3b. Prioritized alcohol gene cell-type expression. We aimed to further assess the cell-type signature of the high-confidence genes linked with PAU and alcohol consumption behaviors. Single-cell RNA sequencing data (GEO accession code: GSE207334) was comprised of ~600,000 single-nucleus transcriptomes from the dorsolateral prefrontal cortex[254] analyzed Using Seurat 5.0.1:[255] data were first processed, then counts were normalized and scaled using the *NormalizeData* and *ScaleData* functions, and finally, differential expression analysis was performed. We used the Ma et al. ontology and nomenclature[254] and the Seurat adjusted P-value threshold (two-sided from Wilcoxon rank sum tests) to define whether the cis-instrument MR findings were differentially expressed in the cell-types.

2.3.3c. Gene-drug analyses assessing repurposing opportunities. We conducted several gene-drug investigations as a final bio-annotation of the cis-instrument MR results. First, we selected compounds that were currently FDA-approved, in clinical trials (active or withdrawn), or experimental use compounds from the Broad Institute Repurposing Hub.[256] We also evaluated the most recent interaction data from DGIdb[257] ("interactions.tsv", "genes.tsv", "drugs.tsv" from April 2024) to determine if the cis-instrument MR genes were targets for therapeutics. DGIdb interaction scores are measures combining supporting publications with the specificity of drug and gene involvement: higher scores indicate stronger endorsement of a drug-gene interaction,[257] and we retained only drug-gene pairs with interaction scores > 0.5 for the analysis.[257]

We also used the Connectivity Map (CMap) Drug database[256] to identify potential drug targets for the cortical proteins and cell-type genes associated PAU and alcohol consumption behaviors. For the cortical proteins and cell-type genes, we combined the AIF, DPW, and binge genes surpassing correction for multiple comparisons into an alcohol consumption signature and the genes associated with PAU into a signature for problems related to alcohol consumption. We collapsed all unique cell-type genes into a single signature (i.e., included genes from all of the cell-types). The CMap database was accessed via the CLUE.IO platform (URL in Code

Availability section), which compared the input gene lists against the CMap touchstone dataset of CMap comprised reference gene signatures across nine cell lines treated with ~3000 well-annotated small molecule drugs. CMap connectivity scores range from ±100 and indicate the similarity between the CMap reference expression signature and the list of input genes representing the disease signature.[256] A negative connectivity core indicates that trait-associated gene expression profile will be normalized by the identified molecule, suggesting possible a repurposing opportunity for the disease of interest (e.g., a connectivity scores of -90 indicates that the drug's signature reverses the expression of the input disease gene sets more than 90% of the other drug sets evaluated).[256] We also performed correlation analysis of the PAU and alcohol consumption behavior drug repositioning signatures, i.e., the PAU cortical signature versus the AIF/DPW/binge drinking signature to evaluate whether there were overall similarities in the drug repositioning scores.

2.3.4. Assessing novelty of the targets identified by the cis-instrument MR screen

We aimed to contextualize and investigate the novel genes identified by cis-instrument MR by comparing them to other gene-based prioritization methods for GWAS studies and complementary approaches that have transcriptomic and epigenomic signatures of AUD. In sections **3.1.4a-g** below, we outline the gene-based methods used to assess the novelty of our cis-instrument MR results.

2.3.4a. Comparison with the GWAS loci of each alcohol-related outcome. Our initial test of novelty was to compare the identified cis-instrument MR genes associated with PAU, AIF, binge drinking, and DPW with the respective genomic loci (represented by the lead SNPs) of the respective alcohol use behavior GWASs. For this initial comparison, we constructed genomic windows (±500 kb) around the lead, independent SNPs for the respective alcohol use behavior GWASs (P-values $< 5\times10^{-8}$, LD $r^2 < 0.1$) and compared the genomic windows with the genomic coordinates of the identified genes. For each alcohol use behavior, we considered the gene to be not captured by the original GWAS signature if it was located beyond any of these genomic windows.

2.3.4b. H-MAGMA gene-based analyses incorporating cortical chromatin interaction data. To assess the novelty of the proteins and genes identified by the cis-instrument MR and colocalization screen, we used an extension of MAGMA, termed "H-MAGMA"[258] (https://github.com/thewonlab/H-MAGMA), which incorporates Hi-C data, to supplement the MAGMA gene mapping comparison of the alcohol-related genes identified by the cis-instrument MR analyses. We analyzed three Hi-C datasets from adult brain,[259] midbrain dopaminergic neurons,[260] and cortical neurons.[261]

2.3.4c. FUSION transcriptomic imputation. We next performed transcriptomic imputation using the FUSION method and following FUSION protocol default settings on autosomal chromosomes[262] to further investigate genes related to the genetic liability of problem alcohol consumption behaviors prioritized by the proteomic and cell-type cis-instrument MR and colocalization analyses. These analyses aimed to investigate the bulk-tissue transcriptomic

relationships of the genes, complement the cis-instrument MR methods, and explore the underlying transcriptomic associations across cortical and subcortical structures given the strong brain-level enrichment we observed in the S-LDSC partitioned heritability analyses outlined above.

For the transcriptomic imputation, we obtained FUSION pre-computed expression quantitative trait loci (eQTL) features from GTEx Version 8.[57] To assess potential bulk-tissue level relationships across brain, we used tissue specific weights for 12 brain regions: amygdala, anterior cingulate cortex (Broadman's Area 24), caudal basal ganglia, cerebellar hemisphere, cerebellum, cortex, frontal cortex (Broadman's Area 9), hippocampus, hypothalamus, nucleus accumbens, putamen, and substantia nigra. As with the cis-instrument MR analyses, the 1000 Genomes Project Phase 3 European subpopulation was used for estimation of linkage disequilibrium (LD).[262] The FUSION pipeline is comprised of three steps. First, it identifies gene expression features that are cis-heritable (i.e., variants associated with gene expression within or near the genomic locus). Second, linear predictors for each cis-heritable gene are constructed – that is, a SNP-based prediction weight of the gene feature. Third, FUSION uses penalized several linear regression and Bayesian sparse linear mixed models (e.g., GBLUP, LASSO, Elastic Net, BLSMM) and computes an out-of-sample R^2 statistics to identify the best model via a cross-validation of each gene-GWAS model to calculate test-statistics incorporating these SNP-based prediction weights and summary-level GWAS Z-scores.[262]

In addition to evaluating whether the cis-instrument MR genes were captured by the FUSION method, we also assessed whether the direction of the MR and FUSION estimates aligned (i.e., were directionally consistent), to evaluate whether the genes demonstrated consistent associations with the respective alcohol use behavior across brain tissues.

2.3.4d. Comparison with differentially expressed genes in post-mortem brains of AUD patients. We also compared the prioritized proteins and genes identified by the cis-instrument MR and colocalization screens with differentially expressed genes (DEG) from 8 brain regions (amygdala, caudate nucleus, cerebellum, hippocampus, nucleus accumbens, prefrontal cortex, putamen, and ventral tegmental area) in a sample of 12 patients with AUD (and 12 European controls).[104] We obtained and processed DEGs from the 8 brain regions using the GREIN package.[263] We extracted the genes prioritized by the cis-instrument MR screen and considered them to be identified by the DEG analysis (i.e., not novel) if they had P-values < 0.05.

2.3.4e. Comparison with the methylomic signature of AUD. We obtained epigenomic signatures (i.e., 2,504 CpG sites significantly associated with AUD) from a recent epigenome-wide association study of AUD using whole blood data from 8,161 participants.[105] As with the GWAS lead loci comparison described above, we constructed genomic windows (±500 kb) around the genomic positions of the 2,504 CpG sites and assessed the genomic windows with the genomic coordinates of the identified genes from the cis-instrument MR screen. We considered the gene to be not captured by the original GWAS signature if it was located beyond any of the CpG genomic windows.

2.3.4f. Comparison with previous TWAS and PWAS for alcohol-related outcomes. For the final comparison, we obtained lists of genes that have been associated with alcohol-related outcomes from the existing proteome-wide association study (PWAS) and transcriptome-wide association study (TWAS) literature (see **Table AP2.1** for a list of the included studies). For inclusion, we considered analyses that were hypothesis free (i.e., not assessing specific pathways or targets of interest), and had used either cis-instrument MR, or a related method capable of integrating GWAS data with proteomic or transcriptomic QTL data, such as FUSION, or S-MultiXcan, S-PrediXcan,[264] etc.). We compared our cis-instrument findings with the targets considered significant by the original study-defined significant thresholds (i.e., we used the existing P-value thresholds designated by each of the studies). If our finding was considered significant by one or more of these previous PWAS/TWAS studies for alcohol-related outcomes, then it was classified in our analyses as "not novel" for alcohol-related outcomes.

2.3.4g. Defining novel genes from the cis-instrument MR screens. We defined novel genes for the respective alcohol use behavior from the cis-instrument MR screen to be those that were not previously captured by any of the above methods or datasets. We considered the gene to be previously mapped if the MAGMA, H-MAGMA, or FUSION transcriptomic imputation P-value surpassed stringent Bonferroni-corrected thresholds used in the initial cis-MR screens (1.40×10^{-5} for cortical proteins and 2.8×10^{-6} for cell-type genes). We defined cis-instrument genes to be novel if they were not captured in any of the methods or datasets at the specified thresholds outlined above.

2.3.5. Integrative validation and neurobiological contextualization of alcohol-related genetic findings

Sections **2.3.5.5a-d** outline methods related to validating and contextualizing genetic findings linked to alcohol consumption behaviors identified through cis-instrument MR. Colocalization analyses were used to confirm shared causal variants, while multi-modal MRI data explored the neurobiological impact of these genes on brain structure and connectivity. Replication efforts utilized independent GWAS data from the FinnGen cohort, incorporating EHR-based diagnoses to prioritize therapeutic targets. Lastly, neuropsychiatric contextualization linked these genes to a broad range of neuropsychiatric and behavioral outcomes, offering insights into their clinical and biological relationships.

2.3.5a. Colocalization. For each gene with primary cis-MR estimates surpassing correction for multiple comparisons, we used colocalization to further evaluate whether targets identified in the cis-instrument MR stage were likely causal (or instead confounded by LD patterns).[265] Colocalization was also used to evaluate the MR exclusion restriction assumption for the prioritized targets. We performed colocalization analysis testing for evidence of a single causal variant within the locus using the *coloc* R package (version 5.2.3) package and default priors.[266] We calculated the posterior probability for each of the genes identified and included all variants ±100 kb of the gene start and end positions; we considered a posterior probability > 0.7

used previously[267,268] as suggestive that cortical proteins and cell-type genes and the respective alcohol consumption behavior share one or more causal variants in the gene region.

2.3.5b. Exploring the impact of the alcohol-related genes on brain structure, white-matter tracts, and functional connectivity. Given the low overlap between the colocalized cortical proteins and cell-type genes, we sought to evaluate whether these non-overlapping proteins/genes demonstrated evidence of shared or distinct neurobiological relationships. Thus, we explored their neurobiological impact using multi-modal MRI data, including global and regional cortical structures (34 regions), subcortical structures,[231,233] and 110 diffusion MRI (dMRI) endpoints (5 tract-averaged endpoints and 21 tracts).[236]

2.3.5c. Replication and evaluation with EHR alcohol-related problems and diagnoses to prioritize therapeutic targets. We next aimed to replicate the top colocalized genes using independent alcohol-related GWAS data. We used the FinnGen cohort (N=377,277)[269] to investigate whether the top cortical proteins and cell-type genes demonstrated genetic relationships with EHR-based outcomes coded as the consequences of alcohol use, including acute International Classification of Disease, Tenth Revision (ICD-10) diagnosed AUD (15,715 cases), alcohol intoxication (8,957 cases), alcoholic liver disease (2,761 cases), alcohol-induced chronic pancreatitis (1,794 cases), alcoholic polyneuropathy (249 cases), and a combined alcohol-related diseases and deaths endpoint (22,186 cases). The AUD Swedish definition includes electronic health record codes related to a range of physical and psychiatric consequences of alcohol use behavior (https://risteys.finregistry.fi/endpoints/AUD_SWEDISH). The ICD-10 AUD definition included electronic health record hospital discharges and causes of death related to mental and behavioral disorders due to use of alcohol. Given the potential biases in self-reported data,[270] in addition to providing an opportunity for replication using independent alcohol-related GWAS data, these EHR-based data also serve both as an important sensitivity test for the self-reported data and a validation of these findings for potentially reducing the clinical burden of problematic alcohol consumption. Cis-instrument MR analyses were performed using the same MR methods as the initial cis-instrument MR screen. The FinnGen data are clinical diagnoses, and therefore, the estimates are reported as odds ratios, 95% CIs, and P-values.

2.3.5d. Neuropsychiatric contextualization. Given the common comorbidity and genetic relationships of alcohol consumption behaviors, neuropsychiatric disorders,[271] and other behavioral outcomes,[188,272] we performed MR analyses on 65 curated neuropsychiatric, neurologic, and behavioral outcomes (the list of included outcomes are presented in **Table AP4.1**) for the cortical protein levels and cell-type gene expression that were also linked with the EHR-based consequences of alcohol consumption. We extracted cis-acting variants for the cortical proteins and cell-type gene expression from each outcome GWAS and performed cis-instrument MR using the same methods as the initial MR screen. We used a Bonferroni-corrected P-value threshold of 7.69×10^{-4} (0.05/65) and compared the directionality of the observed MR estimates with the corresponding MR estimates for the PAU and the other alcohol-related outcomes in the initial cis-MR screen.

2.4. Aim 3 Methods

In the following sections, we present the study overview and detailed methods for the project comprising the 3rd thesis aim. This section is structured to outline the multivariate GWAS framework for the cardiometabolic factor (termed "*CM-Factor*"), followed by the annotation of genetic loci and culminating with three distinct applications of MR designed to assess causal relationships and therapeutic insights. A study overview is provided in **Figure 2.5**.

The project began by constructing a multivariate GWAS model to capture the shared genetic architecture underlying three interrelated cardiometabolic conditions: NAFLD, T2D and coronary artery disease (CAD). We leveraged GenomicSEM to integrate these traits and identify SNP associations representing a broad genetic liability factor, referred to as the *CM-Factor.* The methodological pipeline included rigorous quality control, cross-trait genetic covariance estimation, and SNP-level heterogeneity testing to assess the shared genetic underpinnings of these conditions.

After construction of the *CM-Factor*, we performed comprehensive genomic annotation of identified loci using fine-mapping, transcriptomic imputation, and bulk, and cell-type enrichment. These steps elucidated the biological mechanisms underlying the *CM-Factor* loci by pinpointing candidate causal variants, prioritizing genes, and characterizing functional relevance across tissues and cell types. Finally, we applied MR methods in three distinct studies:

1. **Drug-target MR:** Evaluating the therapeutic potential of approved (by the United States Food & Drug Administration [FDA] and European Medicines Agency [EMA]) and investigational and drug targets for cardiometabolic diseases in 4 major drug classes: anti-diabetics, lipid modulating drugs, NAFLD/NASH therapies, and antihypertensives.

2. **Cis-instrument MR screen:** Identifying novel therapeutic targets within the druggable genome.

3. **Two-step MR:** Investigating proteomic mediators linking body mass index (BMI) to the *CM-Factor*.

Figure 2.5. Aim 3 Project Overview. *An overview of the univariate GWAS study data sources, construction of the multivariate CM-Factor, bio-annotation, and Mendelian randomization (MR) analytical flow and methodology. In the first stage, the broad cardiometabolic liability underlying the univariate input GWASs of non-alcoholic fatty liver disease (NAFLD), type 2 diabetes (T2D), and coronary artery disease (CAD) (i.e., the cardiometabolic factor, or "CM-Factor ") was modeled using Genomic Structural Equation Modeling (GenomicSEM) methods. The CM-Factor model was fit to individual SNPs, generating a multivariate CM-Factor GWAS. In the second stage of the study, we conducted extensive bio-annotation CM-Factor, including fine-mapping to characterize causal variants, performing a transcriptome-wide association study to prioritize causal genes, and evaluating bulk tissue and cell-type enrichment of the CM-Factor signature. Created in BioRender. Rosoff, D. (2025) https://BioRender.com/s33725i*

2.4.1. Data sources

Each GWAS included in the study has existing ethical permissions from its respective institutional review boards and includes participant informed consent with rigorous quality control. For our *CM-Factor* multivariate GWAS, we used a recent meta-analysis NAFLD GWAS derived from 542,997 participants of European ancestry (8,464 cases and 536,533 controls). NAFLD diagnoses were based upon electronic health records (see **Table AP5.1** for complete ICD-10 codes used to define NAFLD cases) and exclusion criteria included alcohol-related liver disease, alcohol use disorder, liver transplantation, hepatitis, LCAT deficiency, inborn errors of metabolism.[107] This meta-analysis had a sample prevalence of 1%,[107] which is substantially lower than the estimated global prevalence of NAFLD of ~25%,[107,108] suggesting a potential underdiagnosis of NAFLD among the controls resulting in control mis-classification. To address these concerns and enhance the diagnostic precision of NAFLD in genetic studies, incorporating liver fat percentage measurements obtained through MRI offered a promising approach: MRI, being a highly sensitive and non-invasive tool for quantifying liver fat, provides a more accurate and direct assessment of liver steatosis compared to ICD codes,[273] which can often miss mild to moderate cases.[273] We therefore augmented the NAFLD GWAS data by combining the NAFLD data with the results of a GWAS on liver fat percentage measured via resonance imaging (MRI) (N=33,235).[273] We meta-analyzed the NAFLD and liver fat percentage GWASs using the multi-trait association of genome-wide association studies (MTAG) method,[274] which is an approach to conduct bi-variate GWAS analyses. We performed MTAG using default settings: first we munged the GWAS data and filtered SNPs (retaining variants with minor allele frequency > 0.1% and sample sizes of $2/3 \times 90^{th}$ percentile of the total sample size); second, we merged the clean NAFLD and liver fat percentage data, keeping only SNPs common to both GWASs; third, we estimated the residual covariance matrix of two datasets using LD Score regression[275] and estimated the genetic covariance matrix; and finally, we performed the MTAG analysis with the European 1000 Genomes Project reference population.[207] For T2D, we used the 2024 European DIAMANTE meta-analysis of 36 cohorts (242,282 T2D cases and 1,569,734 controls)[276]; and for CAD, we used the CAD meta-analysis of 181,522 CAD cases and 984,168 controls from the UK Biobank, CARDIoGRAMplusC4D, and 9 other studies.[277] See **Table AP2.1** for additional information for each of the GWASs used to generate the *CM-Factor* models, including ICD-10 and Office of Population Censuses and Surveys Classification of Interventions and Procedures, version 4 (OPCS-4) codes used to define CAD cases.

2.4.2 Genomic structural equation modeling (GenomicSEM) background and rationale

See **Appendix 4 (Section 4.1)** for an extended presentation of the GenomicSEM methods. Briefly, we used GenomicSEM[16] to perform the multivariate GWAS analysis of NAFLD (meta-analyzed with liver fat percentage), T2D, and CAD to investigate the genetics representing a broad genetic liability underlying these cardiometabolic diseases. Genomic-SEM is a recently developed multivariate GWAS method that facilitates assessment of multiple potential multivariate models of the underlying architecture of the traits of interest (here NAFLD, T2D, and CAD), prior to performing the GWAS analysis.[16] Importantly, given of the inclusion of the UK Biobank participants in each of the univariate GWASs, GenomicSEM is not biased by sample overlap imbalanced sample size.[16] GenomicSEM also performs SNP-level heterogeneity testing to investigate whether the multivariate SNP-association is consistent with the hypothesis that it influences the univariate GWAS traits through the broad cross-trait liability underlying them (see section **2.4.3b.** below).[16]

GenomicSEM analysis is performed in two stages. In Stage 1, we estimated the empirical genetic covariance matrix (and corresponding sampling covariance matrix) for the univariate GWASs. We prepared the NAFLD, T2D, and CAD GWAS summary statistics for stage 1 using a multivariate extension of cross-trait linkage disequilibrium score regression (LDSC)[16,275] to generate the empirical genetic covariance matrix between the three traits, as input for the SEM common factor model.[16] In Stage 2, we specified an SEM maximizing fit between hypothesized covariance matrix and the empirical covariance matrix calculated in Stage 1.[16] Because we aimed to identify a genetic signature underlying the three cardiometabolic diseases, we evaluated a one-factor model estimating the genetic associations between the underlying *CM-Factor* and three CM diseases. We tested the model fit using the standardized root mean square residual (SRMR), model χ^2, Akaike Information Criterion (AIC), and Comparative Fit Index (CFI).[278] After we identified the *CM-Factor* SEM, we proceeded to applying the model to perform the multivariate GWAS identifying SNP associations of the shared covariance across the three cardiometabolic diseases.

2.4.3. Multivariate GWAS analysis

First, we performed quality control (QC) using recommended defaults in LD score regression, including removing SNPs with an MAF < 0.01 (LDSC inflates standard errors of estimates) and information scores < 0.9, and then filtering SNPs to HapMap3, with LD scores estimated from 1000 Genomes Phase 3 reference panel.[275] We use all variants from the three univariate cardiometabolic disease GWASs passing recommended default QC filters for the multivariate GWAS analysis, filtering to the 1000 Genomes Phase 3 European ancestry reference panel. Following the GenomicSEM guidelines, we removed variants with minor allele frequency (MAF) < 0.01 (these variants are prone to error due to fewer samples within the genotype cluster,[16] and also LDSC standard errors of these variants' effects tend to be high[16]); variants with effect values estimated to be exactly equal zero (this avoids compromising matrix inversion required to extend the SEM model to the GWAS data); SNPs not found in the reference panel; and SNPs with mismatched alleles. After merging across the NAFLD, T2D, and CAD summary-level data using listwise SNP deletion, performing QC, and merging with the reference file, there

were 5,547,254 SNPs left to perform the multivariate GWAS analysis. After QC, individual SNP associations are incorporated into the genetic and associated sample covariance matrices and the *CM-Factor* model generated in stages 1 and 2 is applied to the SNPs to obtain their multivariate associations of the shared covariance comprising the *CM-Factor* GWAS data.

After constructing the multivariate GWAS for the *CM-Factor*, we proceeded to calculate the effective sample size and perform several other sensitivity tests, including SNP-level heterogeneity testing, and "GWAS-by-subtraction" (GBS)[279] to test the robustness of the CM-Factor multivariate associations to the genetics of obesity. We outline these analyses below in the following sections.

2.4.3a. Effective sample size calculation. Genomic SEM is not biased by sample overlap,[16] which is important given that the NAFLD, T2D, and CAD included data from overlapping samples (i.e., the UK Bank). We performed a per-SNP calculation for the effective sample size of the *CM-Factor* using the method outlined by the developers.[243] Specifically, given the *CM-Factor* GWAS effect estimate for SNP *j*,

$$\beta_j = \frac{Z_j}{\sqrt{n_j \times 2 \times MAF_j(1-MAF_j)}}$$

In this equation, Z_j is the *CM-Factor* GWAS association statistic for SNP *j*. n_j is the effective sample size for SNP *j* of interest, and MAF_j is the minor allele frequency of SNP *j*; the SNP *j* variance (σ_j^2) is $2 \times MAF_j(1-MAF_j)$¿,∧¿ rearranging,

$$n_j = \frac{(Z_j/\beta_j)^2}{\sigma_j^2}.$$

The, effective sample size, N_{eff}, is considered to be approximately equal to the mean n_j for *m* SNPs meeting the MAF thresholds (restricted in our analyses, as recommended, to MAF between 10% and 40% because the effective calculations are inflated):[16]

$$N_{eff} \approx \frac{1}{m} \sum_{MAF=a}^{b} n_j.$$

2.4.3b. SNP-level heterogeneity testing to finalize the CM-Factor. Genome-wide SNP-level heterogeneity testing (Q_{SNP}) was performed to investigate whether the *CM-Factor* SNP associations were appropriately modeled by a broad underlying genetic factor shared across cardiometabolic diseases or if certain loci exhibited discordant or disease-specific effects that deviated from the shared multivariate genetic architecture. A key feature of the GenomicSEM approach is its ability to assess whether multivariate GWAS SNP associations are best explained by a shared causal pathway underlying the *CM-Factor* or by independent trait-specific pathways. This is achieved using SNP-level Q_{SNP} heterogeneity test statistics, which evaluate the extent to which individual SNP effects are mediated by the common factor or deviate from it.

The Q_{SNP} heterogeneity test produces a χ^2-distributed statistic, where the null hypothesis posits that the SNP effect is fully mediated by the *CM-Factor*. Highly statistically significant Q_{SNP} P-values (e.g., Q_{SNP} P-values $< 5\times10^{-8}$) indicate that the SNP effect is better explained by independent, trait-specific pathways rather than the shared genetic liability modeled by the *CM-Factor*. The test also examines the consistency of the GWAS associations in terms of both the direction (regression coefficients) and the magnitude of the effect estimates across the traits included in the multivariate model. Loci showing significant heterogeneity therefore provide critical insight into regions of the genome where the underlying biology may diverge from the shared genetic liability across the three cardiometabolic diseases.

In our construction of the *CM-Factor*, Q_{SNP} heterogeneity testing was used not only to refine the *CM-Factor* GWAS results but also to exclude loci with significant heterogeneity (Q_{SNP} P-values $< 5\times10^{-8}$) prior to subsequent analyses. This step ensured that the remaining SNP associations reflected the shared genetic risk across NAFLD, T2D, and CAD, thereby enhancing the biological interpretability of downstream bio-annotation and Mendelian Randomization (MR) applications. By applying this filtering criterion, we finalized the *CM-Factor* GWAS for the comprehensive analyses discussed in subsequent sections, ensuring a focus on loci representing the shared genetic architecture of the three cardiometabolic diseases.

2.4.3c. Annotation of CM-Factor loci. We identified genomic loci and lead SNPs (P-values $< 5\times10^{-8}$ and LD $r^2 < 0.1$) associated with the *CM-Factor* using FUMA v1.5.2[280] with default settings. We defined a genomic locus by considering lead SNPs within a 250 kb range and all SNPs in LD ($R^2 > 0.6$) with at least one independent SNP. Independent significant SNPs that were in LD with the same lead SNPs and had LD blocks within 250 kb of each other were consolidated into a single locus. To assess whether the lead SNPs were previously captured in the single-phenotype GWASs of NAFLD, T2D, and CAD, we compared the *CM-Factor* lead SNPs and loci with lead SNPs and loci in underlying univariate GWASs and considered loci novel if they were >1 Mb from previously identified lead SNPs of the input NAFLD, T2D, or CAD GWAS data. Other FUMA-based loci annotation used included ANNOVAR categories, i.e., the functional consequence of SNPs on genes; Combined Annotation Dependent Depletion (CADD) scores (scores >12.37 are the suggested threshold to classify a SNP as deleterious[280]); RegulomeDB scores, i.e., biological evidence that the SNP is a regulatory element (scores with 1a representing the strongest evidence[280]), and MAGMA[281] (Multi-marker Analysis of GenoMic Annotation) with data from GTEx (version 8) to perform gene-based analyses (18,649 protein coding genes within 10 kilobases of the lead *CM-Factor* variants) to inform the gene-set, bulk tissue, and cell-type enrichment analyses (described in the following sections).

2.4.3d. Phenotypic characterization of the CM-Factor loci. We investigated the *CM-Factor* lead SNPs evidenced pleiotropic associations by performing queries of published GWAS significant associations (P-value $< 5\times10^{-8}$) in the GWAS catalog (query date: October 20th, 2024, using FUMA version v1.5.2).[282] To fully account for linkage disequilibrium and remove redundant associations among the independent significant SNPs, we followed this procedure: If the top lead SNP showed any clinical associations, it represented the current locus. If no clinical associations were found for the top lead SNP, we examined the

independent significant SNPs that were highly correlated with it, starting with the most significant SNPs, until we identified known associations.

2.4.3e. Sensitivity testing: GWAS-by-subtraction using GenomicSEM to account for obesity. Given the role of obesity in cardiometabolic diseases,[283] we estimated the SNP associations of the *CM-Factor* after adjusting for BMI. We next aimed to assess the robustness of the *CM-Factor* loci after accounting aspects of obesity. GenomicSEM has been further extended to provide a statistical framework capable of disentangling the genetic architecture of correlated, observed behavior or disease endpoints using GWAS summary statistics in a method termed "GWAS-by-subtraction" (GBS)[279] (**Figure 2.6**). In effect, GBS 'subtracts' the genetic component of one variable from another. For example, in developing the GBS method, Demange et al. used GBS to disentangle the genetic influence of cognition with educational attainment resulting in new, previously unmeasured GWAS data corresponding to the "cognitive component of EA" and the "noncognitive component of EA".[279] Further details regarding GBS can be found in Demange et al.,[279] the corresponding GitHub depository (https://github.com/PerlineDemange/non-cognitive), and the GBS tutorial https://rpubs.com/MichelNivard/565885. As a sensitivity analysis, we leveraged GenomicSEM to perform GBS to account for the genetic architecture of body mass index (BMI) among the lead variants associated with the *CM-Factor*. The aim of this sensitivity analysis was to estimate the SNP associations of the *CM-Factor* after adjusting for body mass index (BMI) given the given the role of obesity in cardiometabolic diseases.[283] We used the BMI summary statistics from the BMI meta-analysis of the GIANT Consortium and UK Biobank by Pulit et al.[133] We also performed GBS using the waist-to-hip ratio adjusted for BMI (WHRadjBMI , also from Pulit et al.[133]). Except for a difference in model specification (here the aim is to construct an SEM reflecting the genetics underlying the trait of interest [i.e., the *CM-Factor*] that is independent of the additional, mediating trait [i.e., BMI or WHRadjBMI])[279], GBS analysis proceeds in a similar fashion as with the GenomicSEM method to estimate the broad genetic liability underlying complex traits: first, a genetic covariance matrix is estimated between the input GWAS data (here the *CM-Factor* and BMI summary statistics); second, a model is fit to estimate the genetic component of the trait of interest that is independent of the second trait.[279] Finally, we fit the GBS model to the lead variants in the *CM-Factor* to assess the extent to which their associations with the broad genetic liability underlying NAFLD, T2D, and CAD are mediated via BMI (or WHRadjBMI). We used two P-value thresholds: a stringent conventional genome wide statistical significance [5×10^{-8}] and an additional P-value threshold correcting for the number of lead variants in the *CM-Factor* to evaluate the robustness of the SNP-*CM-Factor* associations to accounting for BMI or WHRadjBMI.

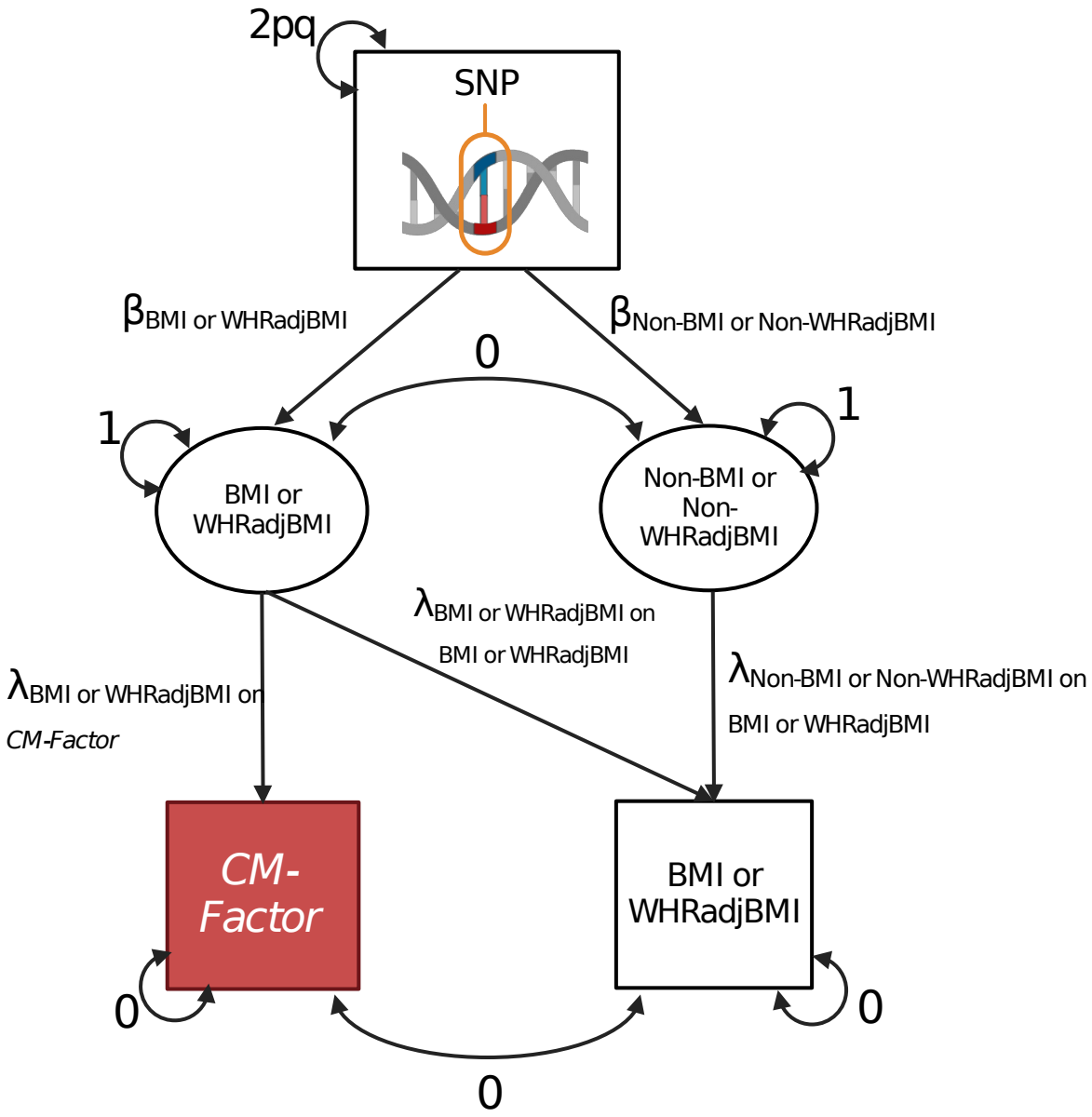


Figure 2.6. GWAS-by-subtraction model to assess the components of the CM-Factor influenced by body mass and adiposity. *The Cholesky model of the GWAS-by-subtraction analysis (adapted from Demange et al.[279]) was fitted in GenomicSEM, incorporating path estimates for a single SNP as an example. Two separate GWAS-by-subtraction analyses were performed: one using body mass index (BMI) and another using waist-to-hip ratio adjusted for BMI (WHRadjBMI). The observed variables, SNP, CM-Factor and BMI/or WHRadjBMI, are derived from GWAS summary statistics. The genetic covariance between CM-Factor and BMI/or WHRadjBMI is estimated using their GWAS summary statistics. BMI/or WHRadjBMI and Non-BMI/or Non-WHRadjBMI are latent (unobserved) variables, with the covariances between CM-Factor and BMI/or WHRadjBMI and between BMI/or WHRadjBMI and Non-BMI/or Non-WHRadjBMI fixed to 0. The SNP variance is fixed to 2pq (p = reference allele frequency, q = alternative allele frequency, based on 1000 Genomes Project phase 3). The residual variances of*

the CM-Factor and BMI/or WHRadjBMI are fixed to 0, attributing all variance to the latent factors, which have variances fixed at 1. Created in BioRender. Rosoff, D. (2025) https://BioRender.com/iozw6hy

2.4.3f. Fine-mapping of CM-Factor loci.

SuSiE (Sum of Single Effects Linear Regression) was used for fine-mapping (*susieR* package, version 0.12.35).[284,285] We fine-mapped all regions with variants with effect P-values < 5×10^{-8}, with regions defined taking a 1.5 Mb window around each lead variant, allowing 10 independent signals per locus. SuSiE reports a 95% credible set for each independent signal. We fit the SuSiE regression using the LD matrix from the 1000 Genomes Phase 3 European reference panel,[243] and report the 95% credible set as an independent signal within the locus when it includes the GWAS lead independent SNP (P-value < 5×10^{-8}), and also secondary sets that include SNP(s) at genome wide significance and log Bayes factor (BF) >2 (so as to exclude multiple credible sets for a single strong signal). We used the *haploR* R package (version 4.0.7)[286] to map each of the fine-mapped SNPs to the datasets available in HaploReg[286] to provide functional annotation for the candidate causal SNPs.

2.4.3g. Gene prioritization using transcriptomic imputation.

We conducted a transcriptome-wide association study (TWAS) to prioritize gene-level associations of the *CM-Factor* genetic signature using the TWAS FUSION method[262] with the FUSION protocol default settings on autosomal chromosomes.[262] For our TWAS, we downloaded from the FUSION website (http://gusevlab.org/projects/fusion/) pre-computed expression quantitative trait loci (eQTL) features in cardiometabolic disease relevant tissues from the GTEx (Genotype-Tissue Expression Project) Version 8[57,287] and METSIM (Metabolic Syndrome in Men)[288] studies derived from donors of European ancestry: GTEx V8 subcutaneous adipose (N=479); GTEx V8 adipose visceral omentum (N=393); METSIM adipose (N=563); aortic artery (N=329); coronary artery (N=175), tibial artery (N=476); heart tissue, including the atrial appendage (N=316) and left ventricle (N=327); liver (N=178); skeletal muscle (N=588); pancreas (N=243); and whole blood (N=558) (all GTEx V8).[57,287]

We used the FUSION method with the same settings as our transcriptomic imputation analyses in Aim 2 (albeit with different transcriptomic weights and panels). Here, we briefly present the FUSION methods for convenience: per the FUSION protocol, we used the 1000 Genomes Project Phase 3 European subpopulation for LD estimation,[262] and excluded genes located within the major histocompatibility complex (MHC) (chromosome 6) given the complex LD structure of that region). FUSION analyses proceed in three stages. First, identify gene expression features that are cis-heritable (variants associated with gene expression within or near the genomic locus). Second, construct a linear predictor for each cis-heritable gene providing SNP-based prediction weight of the gene feature in that tissue. Third, calculate the TWAS test-statistics incorporating these SNP-based prediction weights (the TWAS weights) and summary-level GWAS Z-scores.[262] FUSION performs the regression with several methods: penalized several linear regression and Bayesian sparse linear mixed models (e.g., GBLUP, LASSO, Elastic Net, BLSMM) and then computes an out-of-sample R^2 statistics for each mode, which facilitates identification of the best model via a cross-validation of each gene-GWAS model. We used a Bonferroni-corrected-statistical thresholds determined by the number of genes tested across tissues, i.e., P-value < 6.15×10^{-7} (0.05/81,246 tissue-gene features available for testing in *CM-Factor*), as a heuristic to guide the follow-up analysis on a plausible number of TWAS findings.

We took each of the genes associated with *CM-Factor* surpassing Bonferroni correction for multiple comparisons forward for colocalization to assess evidence of a shared causal variant between the gene expression and the *CM-Factor* in the locus of the gene under analysis.

Colocalization uses a Bayesian approach to estimate posterior probabilities (PP) for SNP associations between two outcome (here *CM-Factor* and the TWAS gene expression) at distinct loci are driven by a shared causal SNP,[106] which facilitates determination of whether the observed associations are driven by horizontal pleiotropy, i.e., a single SNP impacting both gene expression and *CM-Factor* (posterior probability PP.H4) or linkage disequilibrium (LD), i.e., 2 separate SNPs in LD impacting gene expression and GWAS signal separately (posterior probability PP.H3).[106] We then conducted colocalization analyses between each *CM-Factor*-associated gene that surpassed Bonferroni correction for multiple comparisons in the TWAS (TWAS P-value < 6.15×10^{-7} [0.05/81,246 total tests]) using default priors the *coloc* R package (version 5.1.0).[106] We categorized results using a PP.H4 threshold of > 0.6 to define evidence for a shared causal variant between eQTL and the *CM-Factor* signal at that genomic locus.[106] Finally, for features with TWAS P-values surpassing correction for multiple testing and demonstrating evidence of colocalization, we performed conditional testing to identify which genes are conditionally among multiple associated features in a locus. We used the FUSION *FUSION.post_process.R* provided as part of the FUSION pipeline script to perform these analyses we prioritized "high-confidence" *CM-Factor* genes identified by FUSION if the genes also demonstrated evidence of colocalization and were conditionally significant. For high-confidence genes with associations across more than one tissue weight (i.e., if the gene was associated with *CM-Factor* in two or more tissues), we filtered results for the gene/tissue combination based upon variance explained by the gene/tissue pair.

Novelty assessment for the high-confidence TWAS genes. We investigated whether the TWAS results captured additional biology beyond the genetic loci of the input univariate T2D, NAFLD (augmented with liver fat percentage), and CAD GWASs. To assess and define novelty for the high-confidence TWAS genes, we compared the genomic coordinates of the 243 genes with the lead variants for the genomic loci of the T2D, NAFLD, and CAD GWASs. We considered the high-confidence gene to be novel if it was located beyond 500 kb from any of the lead variants comprising the T2D, NAFLD, CAD, or *CM-Factor* GWASs. We also performed a GWAS Catalog look-up using FUMA (performed on October 20th, 2024) for associations of these genes with CAD, T2D, NAFLD, and liver fat. If the gene was beyond 500 kb of the input univariate GWAS loci, the *CM-Factor* loci, and not in the GWAS Catalog, we defined the druggable gene as a novel finding.

Gene-drug-disease network enrichment. We next undertook several gene-drug interaction analyses to leverage the high-confidence *CM-Factor* gene signature to identify potential gene-drug associations. First, we investigated the enrichment of the high-confidence *CM-Factor* genes identified with the TWAS, colocalization, and conditional testing within targeted gene sets for different drug categories found in the DrugBank database.[289] To identify potentially repositionable drugs, we constructed a gene-drug-disease network. The GREP package[290] was used to perform Fisher's exact tests, assessing whether the high-confidence *CM-Factor* genes were significantly enriched in the gene sets targeted by drugs for specific diseases or conditions. All tests were subjected to Bonferroni correction to account for multiple comparisons. This comprehensive approach enabled us to systematically prioritize drugs that may be repurposed based on their genetic targets. We used a P-value threshold of 0.05 as evidence for enrichment in the drug category.

Gene-drug signature matching investigation assess potential therapeutic repurposing opportunities. Finally, we queried the high-confidence genes list against the drug/compound signatures in the Connectivity Map (CMap) Drug database.[256] We accessed CMap via the CLUE.IO platform (https://clue.io/), compared the input gene lists against the CMap touchstone dataset, which comprises reference gene signatures from nine cell lines treated with ~3,000 well-annotated small molecule drugs. We used the transcriptomic signature matching approach[291] where the high-confidence *CM-Factor* gene signature (i.e., their Z scores) are queried against each of the transcriptomic signatures of the ~3,000 small molecules in the CMap database to identify and prioritize small molecule compounds that have transcriptomic signatures that may offset the *CM-Factor* transcriptomic signature, reflecting a connection indicating potential relevance for the underlying disease processes.[291] To refine the *CM-Factor* signature, we accounted for any duplicate genes that were identified in more than one bulk tissue panel by removing the Z scores of the gene/tissue that had a lower TWAS model R^2 (assessed from the TWAS summary statistics). CMap connectivity scores range from ±100, indicating the similarity between the CMap reference expression signature and the list of input genes representing the disease signature.[256] A negative connectivity score indicates that the trait-associated gene expression profile will be normalized by the identified molecule, suggesting a possible repurposing opportunity for the disease of interest (e.g., a connectivity score of -90 suggests that the drug's signature reverses the expression of the input disease gene sets more than 90% of the other drug sets evaluated).[256] We therefore prioritized small molecule compounds with scores between -90 and -100.

2.4.3h. Tissue and cell-type enrichment. We sought to investigate the tissue-level enrichment of the *CM-Factor*. We leveraged stratified LD Scored regression (S-LDSC)[292] and SNP-based heritability estimation, which enables the partitioned heritability of SNPs according to genomic properties, to facilitate identification of tissues contributing to the polygenic signature the *CM-Factor*.[292] We used the S-LDSC pipeline to evaluate the *CM-Factor* in S-LDSC using 489 datasets from the ENCODE project[293] and Roadmap Epigenomics[294] assessing the tissue specific annotations of active chromatin and enhancers (marked by H3K9ac, H3K27ac, DNase hypersensitivity sites, and H3K4me1). We also tested the *CM-Factor* enrichment of genes expressed in 53 GTEx V7 tissues[57] and the 152 tissues derived from several RNA-sequencing studies.[295,296] We used a Bonferroni-corrected threshold to adjust the P-value (two-sided) to account for multiple comparisons for each analysis.

To complement the partitioned heritability S-LDSC analyses across bulk tissues, we next aimed to identify etiological cell types associated with the *CM-Factor*. We integrated single-cell RNA-sequencing (scRNA-seq) data using CELLECT (CELL-type Expression-specific integration for Complex Traits),[122] using scRNA-seq data from the Tabula Muris study.[297] The Tabula Muris study includes transcriptomic data from 100,000 cells and 20 organs and tissues throughout the mouse. We cleaned and prepared the Tabula Muris scRNA-seq data for CELLECT analysis using CELLEX.[122] CELLEX calculates expression specificity likelihood (ESμ) scores for each gene following normalization and pre-processing.[122] Using CELLECT's default settings, we performed the cell-type enrichment with MAGMA and S-LDSC methods. In CELLECT, MAGMA measures the extent to which genetic associations with a phenotype increase as a function of gene expression specificity for a given cell type, while S-LDSC quantifies the effects of each cell type ESμ with the heritability of the *CM-Factor* GWAS.[122] We categorized our cell

types following the nomenclature used in the original Tabula Muris study.[297] We compared the MAGMA and S-LDSC results and looked for concordance across the two methods, using a Bonferroni-corrected threshold to adjust the P-value (two-sided) to account for multiple comparisons.

2.4.4. Mendelian randomization applications

In addition to conducting a screen of druggable genes with whole blood gene expression data to supplement the TWAS results and gene prioritization of the *CM-Factor*, we leveraged the MR framework[48,49] for several applications related to therapeutic efficacy of approved and investigational cardiometabolic targets, and identification of novel proteomic mediators linking obesity and the *CM-Factor* **(Figure 2.7)**. In the sections below, we detail the specific methods for the three types of MR analyses performed in the study: (1) drug-target MR approved and prospective therapeutic targets for NAFLD, T2D, and CAD using cis-instrumentation in the primary physiological response to the therapy of the primary biomarker of interest; (2) a screen of "druggable" genes in the circulating transcriptome; and (3) a two-step MR to assess the proteomic mediators of the effect of body mass index (BMI) on the *CM-Factor*.

Figure 2.7. Overview of Aim 3 Mendelian randomization (MR) studies. *After bio-annotation, we proceeded to perform the studies leveraging the MR framework to inform drug prioritization and discovery. First, we curated lists of approved and investigational cardiometabolic drug targets across several classes of therapies (antidiabetics, lipid-modulating drugs, NAFLD/non-alcoholic steatohepatitis [NASH] targets, and antihypertensives) and performed drug-target MR using variants located within or near the respective genomic loci encoding the drug target to evaluate the relationships of the expected physiological responses to each therapeutic class (e.g., lowered LDL-C levels in response to PCSK9 inhibition). We performed colocalization to assess evidence of a shared causal variant for drug targets surpassing correction for multiple comparisons. In the second MR study, we screened transcriptomic levels of ~2,500 genes that have been characterized as amenable for modulation by therapeutic compounds to prioritize candidate targets for the CM-Factor. Finally, in the third study, we conducted a two-step MR analysis to characterize the proteomic mediators of BMI-CM-Factor relationships. In the first stage, we identified proteins associated with increased BMI in the first release of the UK Biobank Pharma Proteomics Project (UKB-PPP) circulating proteome data. We then took forward prioritized BMI-associated proteins and performed cis-instrument MR of these proteins on the CM-Factor. Replication using independent proteomic data, colocalization, and mediation*

analyses were also conducted as part of the two-step MR analyses. Created in BioRender. Rosoff, D. (2025) https://BioRender.com/pb26h2z

2.4.4a. MR Study 1: drug-target MR of antidiabetics, lipid-modulating therapeutics, NAFLD/NASH targets, and antihypertensives targets.

Because genetics-based studies are useful to identify repurposing opportunities and identify adverse side effect profiles,[6,51] and genetics-based support of therapeutic targets in clinical trials improves the probability of clinical success for candidate compounds entering clinical trials,[21] we performed drug-target MR[50] of approved, clinical-stage, or preclinical targets in the drug development pipelines for NAFLD, T2D, and CAD. See **Figure 2.8** for an overview of the drug-target MR screen, including complete lists of the targets analyzed. For instrumentation, first, we obtained summary level data for glycated hemoglobin (HbA1c), body mass index (BMI) circulating lipid levels (low-density lipoprotein cholesterol (LDL-C), high density lipoprotein cholesterol (HDL-C), triglycerides (TG), lipoprotein A (Lp(a))), liver fat percentage calculated by magnetic resonance imaging (MRI data), and systolic blood pressure (SBP) from participants of European ancestry[201,223,298,299] (N range: 33,824 to 1,320,016) to construct our drug-target genetic instruments proxying pharmacological modulation of the expected physiological responses for each class of therapeutic targets. Below we outline instrument selection for the respective targets.

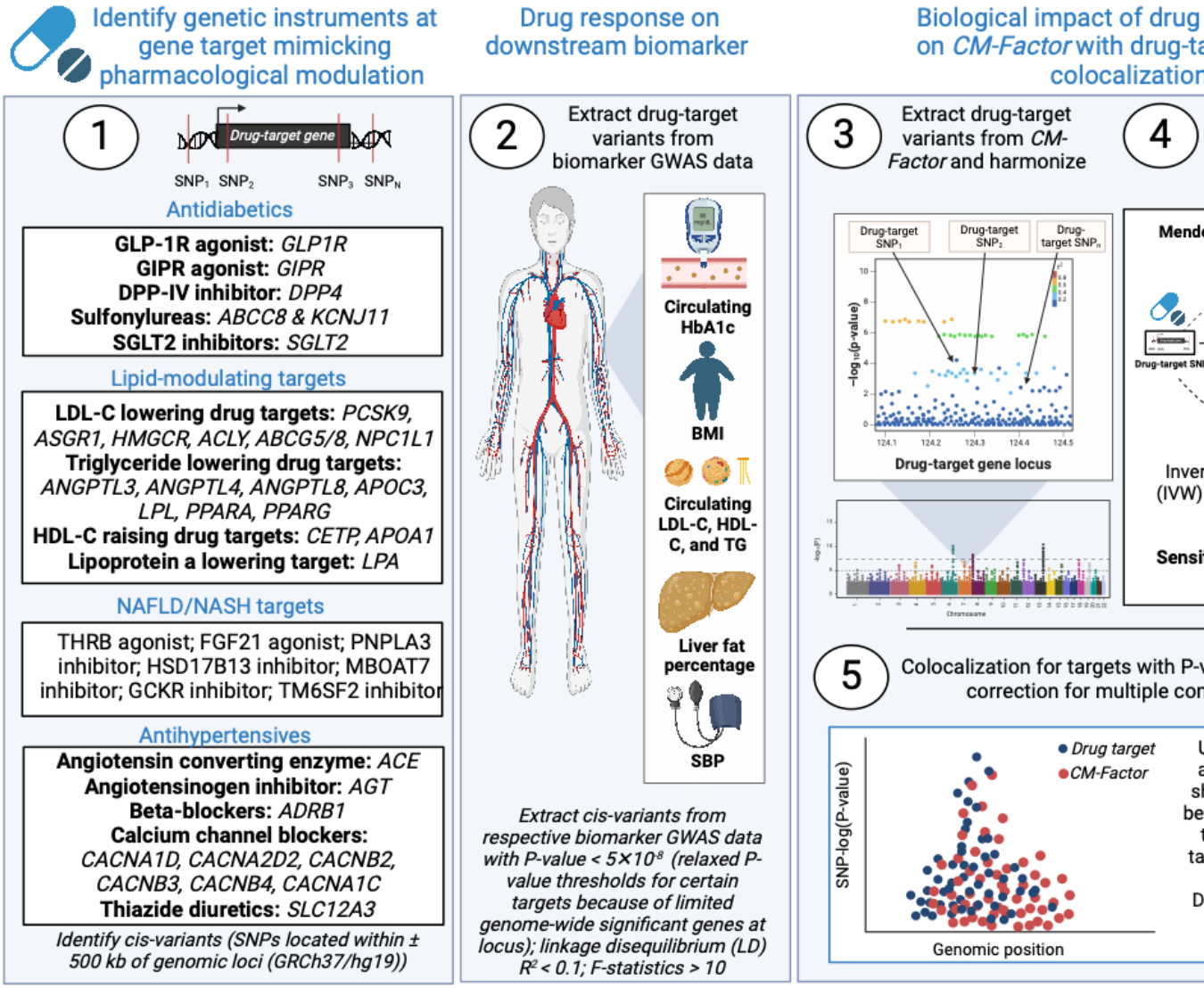


Figure 2.8. Drug-target Mendelian randomization analysis overview of antidiabetics, lipid-modulating targets, NAFLD/NASH targets, and antihypertensives for the first MR study of Aim 3. *The figure presents a flow diagram and details of the drug-target MR analyses of the drug-targets on the CM-Factor. In Step 1, cis-instrumentation was performed using genome-wide association study (GWAS) of biomarkers that are the primary indications of pharmacological modulation of these targets. For the antidiabetics, we used circulating levels of HbA1c (N=344,182). Body mass index (BMI) was also used for instrumentation of GLP1R and GIPR. For lipid-modulating targets, we used several lipid subfractions (N ~ 1.3 million) including LDL-C, triglycerides, and HDL-C. Because non-alcoholic fatty liver disease (NAFLD) targets lower liver fat and/or alanine aminotransferase levels, we used GWASs for expected physiological response (e.g., reduced liver fat [N=32,858], or alanine aminotransferase levels [for HSD17B13, N=344,136]) to proxy the NAFLD targets. Lastly, for the antihypertensives, we used GWAS data of SBP (N ~ 1.1 million). Independent variants (LD $r^2 < 0.1$) at P-values < 5×10^{-8}) were extracted, and cis-instruments constructed for each target, which exposure variants were then extracted from the CM-Factor GWAS, harmonized, and then analyzed using*

multiple MR methods (steps 3 and 4). In step 5, we used colocalization to assess evidence of a shared causal variant between the drug targets that had MR estimate P-values surpassing correction for multiple comparisons. MR, Mendelian randomization; LD, linkage disequilibrium. LDL-C, low-density lipoprotein cholesterol; HDL-C, high-density lipoprotein cholesterol; TG, triglycerides; SBP, systolic blood pressure. Created in BioRender. Rosoff, D. (2025) https://BioRender.com/djcerbt

Instrumenting anti-diabetic drug targets. We identified 7 classes of antidiabetic medications with known drug targets guide our genetic instrumentation:[300] glucagon-like peptide-1 receptor (GLP1R) analogs, gastric inhibitor polypeptide receptor (GIPR) analogs, insulin analogs, sulfonylureas, dipeptidyl peptidase 4 (DPP-IV) inhibitors, and sodium-glucose cotransporter-2 (SGLT2) inhibitors. We identified the targets for sulfonylureas (ABCC8 and KCNJ11), insulin analogs (INSR), and thiazolidinediones (TZD) (PPARG) with the DrugBank and ChEMBL databases.[301,302]

We found that there were suitable SNPs located within or near the genomic coordinates of the target boundaries (i.e., 500 kb on either side of start and end positions) within the glycated hemoglobin (HbA1c) GWAS data derived from participants in the UK Biobank.[223] We instrumented antidiabetics with the HbA1c biomarker because lowered glycated Hba1c is considered an expected physiological response to the antidiabetic effects of these targets and has been used in previous drug-target MR studies of antidiabetic targets.[196,303]

We used three data GWAS data sources to genetically instrument GLP1R agonism. First, we proxied the impact of GLP1R agonism using independent (linkage disequilibrium $R^2 < 0.1$) genome-wide significant (P-value < 5×10^{-8}) SNPs located ±500kb of the GLP1R gene position (chromosome 6:39,016,574–39,055,519 on GRCh37/hg19) associated with glycated hemoglobin (Hba1c) levels (mmol/mol) among participants in the UK Biobank (N= 344,182),[223] because one of the primary effects of GLPR1 agonists is to improve glucose homeostasis via actions on pancreatic cells.[304] Also, because GLP1R and GIPR also act in the central nervous system to reduce appetite and body weight,[223] we constructed other GLP1R and GIPR instruments using body mass index (BMI) data from a meta-analysis of GIANT and the UK Biobank (N=806,834) by Pulit et al.[133] As there were no genome-wide significant associations for BMI within the *GLP1R* locus, we selected independent SNPs using a relaxed P-value threshold of 5×10^{-6}. For our genetic instruments proxying GLP1R analogs, GIPR analogs, sulfonylureas, SGLT2 inhibitors, and thiazolidinediones were extracted variants at conventional genome-wide statistical significance (P-value < 5×10^{-8}) that were within/near the cis-acting loci of the target gene boundaries. For DPP-IV inhibitors and INSR analogs, we used a relaxed P-value threshold (5×10^{-4}) because there were no variants at conventional genome-wide statistical significance in either the *INSR* locus or the *DPP4* locus.

Instrumenting lipid-modulating drug targets. We analyzed both the targets of approved lipid-modulating therapies and also investigational lipid-modulating targets identified from a literature look-up (see ref.[305]). We identified 18 potential targets for to evaluate for potential instrumentation. As with our instrumentation of antidiabetics using HbA1c and BMI GWAS data as the exposure, we genetically mimicked pharmacological modulation of the lipid drug targets by extracting genetic variants associated with their respective biomarker considered the primary physiological response to pharmacological modulation of that target, i.e., low-density lipoprotein

cholesterol (LDL-C) lowering therapies, triglyceride lowering therapies, high-density lipoprotein cholesterol (HDL-C) raising therapies, and lipoprotein a (LP(a)) lowering therapies. For the targets that modulate LDL-C, triglycerides, or HDL-C, we used the 2021 Global Lipids Genomics Consortium (GLGC) meta-analysis GWAS data from European participants (≤ 1,320,016).[201] For every target we evaluated whether there were suitable genetic variants located within the cis-acting loci of the target gene boundaries (±500 kb of genomic boundaries) for their respective lipid. *DGAT1* and *DGAT2* genes, which are important in triglyceride metabolism,[306] had no variants within these genomic loci with suitable association statistics with the triglyceride GWAS data of triglycerides levels, and therefore, *DGAT1* and *DGAT2* were excluded from further analysis, leaving 15 lipid-modulating targets for our analyses. For 14 of the remaining 16 lipid-modulating drug targets, we extracted variants at conventional genome-wide statistical significance (P-value < 5×10^{-8}) that were within the cis-acting loci of the target gene boundaries (±500 kb of gene boundaries). We used a relaxed P-value threshold (P-value < 5×10^{-4}) for instrumenting *ACLY* in LDL-C data due to the lack of SNPs associated with LDL-C at conventional genome-wide statistical significance in the *ACLY* locus.

Instrumenting NAFLD/NASH targets. There is only a single currently approved therapeutic for non-alcoholic steatohepatitis (NASH), Resmiterom, which is an analog for thyroid hormone receptor (THR)-β.[307] Several other NAFLD therapeutics are in Phase 1 or 2 clinical trials. For example, results from Phase 2b trials support new fibroblast growth factor 21 (FGF21) analogs, Pegozafermin and Efruximermin, as therapeutic interventions for NAFLD,[124,125] and other Phase 1/2 clinical trial results demonstrate that inhibiting 17β-Hydroxysteroid dehydrogenase type 13 (HSD17B13), lowers alanine aminotransferase levels, a marker of hepatic dysfunction.[308] Similarly, patatin-like phospholipase domain-containing 3 (PNPLA3) is another target under consideration in preclinical and clinical trials.[309-311] In addition to these targets, genetics-based studies and other lines of evidence have highlighted several other potential targets for NAFLD treatment, including Transmembrane 6 superfamily member 2 (TM6SF2), Membrane Bound O-Acyltransferase Domain Containing 7 (MBOAT7), and Glucokinase Regulator (GCKR),[107] which together underscores the need to investigate the potential long-term efficacy of these NAFLD targets for cardiometabolic disease.

Reducing liver fat content or percentage is one of the expected physiological responses to pharmacological modulation for many of the NAFLD targets. For example, PNPLA3 is associated with formation of hepatic droplets and liver fat content, and reductions in liver fat levels are being used as an efficacy marker in ongoing PNPLA3 inhibitor trials.[309-312] Therefore, for FGF21, PNPLA3, TM6SF2, MBOAT6, and GCKRT, our primary drug-target instruments were constructed using variants within or near their respective genomic loci (±500 kb) associated at conventional genome wide significance with magnetic resonance imaging (MRI)-derived quantification of liver fat percentage (N=32,858, UK Biobank participants of European ancestry)[313] to instrument PNPLA3, TM6SF2, and GCKR. Due to a lack of conventional genome-wide significant SNPs in the loci of FGF21, THRB, and MBOAT7, we used less stringent P-value thresholds to construct the instruments, i.e., a P-value=0.01 for FGF21, 5×10^{-6} for THRB, and 5×10^{-5} for MBOAT7.

HSD17B13 is not associated with triglyceride formation in the liver,[314] and there were no sufficiently strong variants (i.e., F-statistics >10) within or near the *HSD17B13* locus associated

with the GWAS data for liver fat percentage. Therefore, we did not instrument HSD17B13 using MRI-derived liver fat data. Instead, we used an additional marker of liver function and NAFLD – alanine aminotransferase (ALT) – commonly used to evaluate NAFLD risk in clinical settings.[314] NAFLD is associated with chronically elevated ALT levels, and while ALT elevation does not cause NAFLD or cirrhosis, the underlying biological mechanisms responsible for increased ALT levels are responsible for NAFLD and cirrhosis.[314] Also, the current clinical trials investigating the novel inhibitors of HSD17B13 and PNPLA3 include ALT as one of the primary liver-related biomarkers assessing the physiological response to the pharmacological inhibition of these targets.[309-311] Therefore, to genetically-proxy inhibition of HSD17B13 and PNPLA3, we extracted cis-instruments for HSD17B13 from the Neale Lab GWAS on circulating alanine aminotransferase levels (ALT) in UK Biobank participants of European ancestry, N=361,194).[223] As with the lipid-lowering drug-targets, we extracted variants within ±500 kb of the gene loci associated with ALT levels at P-value < 5×10^{-8}.

Antihypertensive drug targets. As with the selection processes for antidiabetic, lipid-modulating, and NAFLD, targets, we began by identifying classes of antihypertensive drugs, referencing DrugBank and ChEMBL databases[301,302] to pinpoint genes associated with the targets of 5 antihypertensive classes: angiotensin converting enzyme (ACE) inhibitors, angiotensinogen (AGT) inhibitors, beta blockers, calcium channel blockers (CCBs), and thiazide diuretics (TDs).[315] We used the 2024 GWAS meta-analysis of SBP performed in participants of European ancestry (N=1,028,980)[299] for the exposure instruments. As above, we considered SNPs located within 500 kb of the gene boundaries as cis-instrument proxies to potentially reduce systolic blood pressure (SBP) through these targets. For the CCBs, rather than analyzing each gene target independently, we merged several gene targets (i.e., CACNA1D, CACNA2D2, CACNB2, CACNB3) into a unified CCB instrument, following prior methodologies.[316] Additionally, we adopted a less stringent P-value threshold (< 5×10^{-6}) due to the absence of SNPs with P-values < 5×10^{-8} near the *CACNB4* and *SLC12A3* loci, where no genome-wide significant variants related to SBP were found.

Drug-target MR statistical methods and colocalization. First, we extracted SNPs at the LD r^2 < 0.1 threshold (10,000 kb) using the 1000 Genomes Project EUR reference population,[207] and calculated F-statistics to evaluate instrument strength and assess the first MR assumption. We used the cutoff of F-statistic >10 as suggesting the resulting estimates would be subject to minimal bias from weak instruments.[242] After harmonization with the *CM-Factor*, we performed MR IVW (random-effects analysis performed when there were more than three variants) as the main method and the Wald ratio[245] as the main method for instruments comprised of a single SNP. Heterogeneity test (Cochran's Q) and MR Steiger directionality tests investigated evidence of pleiotropic SNPs and reverse causality.[212,317] For instruments with 3+ SNPs, complementary methods (MR-Egger, weighted median, weighted mode) were also used. If the Cochran's Q P-value indicated heterogeneity (Cochran's Q P-value < 0.05), we used the MR LASSO method[318], which applies a lasso-type penalization to refine the genetic instruments by identifying outlier SNPs, to provide IVW estimates corrected for heterogeneity.

For the NAFLD/NASH targets instrumented using liver fat percentage GWAS data (THRB, FGF21, PNPLA3, TM6SF2, MBOAT7, and GCKR), which data were integrated into the main *CM-Factor* under investigation, we performed these drug-target MR analyses using a second

CM-Factor that did not contain the liver fat percentage GWAS data as part of its construction supplementing the GWAS of NAFLD diagnosis (i.e., it was constructed using only the NAFLD, CAD, and T2D datasets and not the NAFLD + liver fat percentage meta-analysis). The drug-target MR analysis of HSD17B13 inhibition proxied using lowered ALT levels as the exposure used the primary *CM-Factor* outcome.

To facilitate interpretation of these classes of drugs, we oriented the MR estimates to reflect the direction of the expected physiological responses to the pharmacological modulation of the drug targets. For the antidiabetics, we oriented estimates to correspond to the HbA1c-lowering and BMI-lowering for GLP1R and GIPR. For the LDL-C and TG lowering drug-targets, we oriented MR estimates to correspond to lower circulating LDL-C levels (for LPL, this is achieved by enhancing LPL activity). We oriented CETP inhibition estimates to correspond to an increase in circulating HDL-C levels. Similarly, for FGF21, PNPLA3, and the other NAFLD targets, we oriented MR estimates to correspond to lowering of liver fat percentage, while for HSD17B13, estimates were oriented to lowering of circulating liver fat percentage (or ALT levels for HSD17B13). Finally, for antihypertensives, we oriented MR estimates to align with lowering SBP.

For drug targets with MR estimates on the *CM-Factor* surpassing nominal P-value threshold (P-value < 0.05), we performed colocalization analyses to assess whether the biomarker GWAS in which the drug-target gene was instrumented and the *CM-Factor* show evidence of a shared causal variant at the gene locus. For colocalization, we used the coloc R package (version 5.0.0),[106] and included all variants within 500 kb of the drug target genomic boundaries. We calculated the posterior probabilities using the default priors: ($p1 = p2 = 10^{-4}$ and $p12 = 10^{-5}$) in *coloc* and considered a posterior probability hypothesis H4 (PP.H4) > 0.6 as evidence that both traits share a single causal variant within the genomic locus of the target gene. Low posterior probabilities for the third (H3, both trait 1 and trait 2 are associated, but with separate SNPs) and fourth hypotheses and a corresponding high posterior probability for the first hypothesis (H1, only trait 1 has a genetic association in the locus) suggest that the colocalization analysis is underpowered, potentially because the outcome dataset does not have sufficiently strong genetic signals in the locus.[265] Therefore, for colocalization results meeting these criteria (low PP.H3 and PP.H4 and a high PP.H1), we included an conditional PP.H4 by calculating PP.H4/(PP.H4+ PP.H3) that has been used previously in MR studies using eQTL and pQTL exposure sources.[265,319,320]

2.4.4b. MR Study 2: Screening the druggable genome for novel targets for the CM-Factor. In the second MR study (overview in **Figure 2.9**), we again used drug-target MR using cis-expression quantitative trait loci (eQTL) data from whole blood to identify potential therapeutic targets for the *CM-Factor*, focusing on genes previously identified as druggable. We started with a comprehensive list of 4,676 druggable genes, which includes approved drug targets and candidates in clinical development (Tier 1), genes with known bioactivity for small molecules (Tier 2), and genes linked to druggable families such as G-protein coupled receptors and kinases (Tiers 3A-3B).[126] After harmonizing these eQTL data with the *CM-Factor*, we performed drug-target MR analyses on 2,546 genes, using stringent statistical thresholds to ensure robustness. We further conducted colocalization analyses to determine shared causal variants between these genes and the *CM-Factor*. Genes surpassing these

thresholds were advanced for further investigation, including assessing their novel contributions to cardiometabolic health and evaluating their therapeutic potential through phenome-wide MR analyses across 366 diseases and biomarkers.

Figure 2.9. Aim 3 MR Study 2 overview drug-target MR screening the druggable genome for novel targets of the CM-Factor. *This figure describes the second MR study included in the downstream exploration of the CM-Factor, which used drug-target Mendelian randomization (MR) to identify potential therapeutic targets for the CM-Factor using cis-expression quantitative trait loci (eQTL) data from whole blood. The study focused on a list of 4,676 druggable genes, categorized into tiers based on their status in drug development, including approved drug targets and candidates in clinical trials (Tier 1), genes with known bioactivity for small molecules (Tier 2), and genes linked to druggable families such as G-protein coupled receptors and kinases (Tiers 3A-3B) (defined by Finan et al.[126]). After harmonizing eQTL data with the CM-Factor, MR analyses were performed on 2,546 genes. The study flow incorporated colocalization and mediation analyses for prioritized genes and evaluated their relationships with established cardiometabolic biomarkers such as lipids, glucose, blood pressure, and BMI. Prioritized genes were subjected to additional analyses, including novelty assessment, replication evaluation of their effects on cardiometabolic health and biomarker relationships through phenome-wide MR across 366 diseases and biomarkers. This integrative approach aimed to uncover novel therapeutic targets and assess the potential clinical relevance of these genes for cardiometabolic health. Created in BioRender. Rosoff, D. (2025) https://BioRender.com/seh9gpx*

MR analysis and colocalization. We performed the drug-target MR screen using the same MR methods as described in the preceding sections for the Study 1 MR analyses assessing the impact of approved and investigational therapeutic targets on the *CM-Factor*. We took all genes surpassing P-value=1.96×10^{-5} (0.05/2,546 druggable genes tested) forward for colocalization analyses using the *coloc* R package.[106] The colocalization settings were the same as those described in MR Study 1 and druggable genes with MR estimate P-values < 1.96×10^{-5}, and colocalization PP.H4 or conditional PP.H4 values > 0.6 were further prioritized for the additional analyses outlined in the sections below. As with the TWAS results, we evaluated whether the screen of the druggable genome captured biology beyond the genetic loci of the input univariate GWASs of T2D, NAFLD (augmented with liver fat percentage by meta-analysis using the MTAG methods), and CAD. For each of the druggable genes with evidence of colocalization, we compared the genomic coordinates of the genes with the loci of T2D, NAFLD, and CAD and defined the gene to be novel for the cardiometabolic diseases if it was located > 500 kb upstream or downstream of the lead variants defining the T2D, NAFLD, or CAD genomic loci as determined by the FUMA package[280] using the same settings that defined the *CM-Factor* loci. We also performed a GWAS Catalog look-up using FUMA (performed on Sept 20th, 2024) for associations of these genes with CAD, T2D, NAFLD, and liver fat. If the gene was beyond 500 kb of the input univariate GWAS loci, and the *CM-Factor* loci, and was not in the GWAS Catalog, we defined the druggable gene as a novel finding.

Integrating biomarker data relevant for cardiometabolic health. For many of the prioritized drug-target genes, their downstream effects on cardiometabolic biomarkers and cardiometabolic health remain uncertain. Therefore, to further stratify the drug-targets from the drug-target MR and colocalization screen, we sought to link the drug targets to cardiometabolic biomarker data to enhance the potential biological validation of genes by determining how the gene expression of the prioritized targets influences cardiometabolic health biomarkers such as lipids, glucose, blood pressure, and BMI. In addition, because the causal role of biomarkers in disease elucidated

through MR may identify intermediate exposures (and correspondingly underlying causal genes) that are both causal and amenable to therapeutic modification,[21,161,282] this approach may also inform our understanding of the biological pathways, providing valuable insights into disease mechanisms in the causal pathways from gene expression to biomarker to cardiometabolic risk (i.e., linking the drug-target gene expression with established cardiometabolic biomarkers that are amenable to pharmacological modulation would increase the confidence in the interpretation that the drug-target genes have relevant cardiometabolic biology).[161]

We first calculated 95% credible sets for the colocalized drug-target genes to identify the candidate causal SNPs in the colocalized loci. We took genes with variants with posterior inclusion probabilities (PIPs) of > 0.5 forward to a phenotypic screen using *LDtrait* (part of the *LDlinkR* package).[321] *LDtrait* performs look-pups of SNPs (or variants in LD with those SNPs) in the EBI-EMBL GWAS Catalog.[282] We queried *LDtrait* for GWAS Catalog associations surpassing conventional genome-wide significance (P-value < 5×10^{-8}) (query performed on October 1st, 2024) with established cardiometabolic biomarkers used in clinical practice and clinical trials to assess the physiological responses to cardiometabolic therapeutics. Specifically, we looked for associations with BMI, SBP, glycemic markers (HbA1c, fasting glucose, fasting insulin, etc.), major lipid subfractions (LDL-C, HDL-C, and triglycerides), and liver fat content, liver enzyme tests (ALT, aspartate aminotransferase [AST], gamma-glutamyltransferase [GGT]).

To supplement this GWAS Catalog-based look-up of the causal variants, we also performed SNP-based analyses applying an additive genetic model using data from the Oxford Biobank (OBB) (N≤ 8,000) (see below for additional details on the OBB).[322] For each causal SNP, we fit a linear regression model with anthropometric variables, major circulating biomarkers, and adiposity depots measured by Dual X-ray Absorptiometry (DXA) as the outcome(s) and the SNP genotype (coded as 0, 1, or 2 for the number of minor alleles) as the predictor. We ran several models adjusting for potential confounders, including age, sex, and body mass index (BMI), in the regression model and looked for associations with relevant biomarkers aligning with the GWAS Catalog query. We looked for consistent evidence of relationships between the causal variants and the biomarkers from the two analyses (P-value < 0.05) and took forward the drug-target genes with evidence of associations between their causal variants and the cardiometabolic biomarkers for further MR analyses, including mediation to determine the proportion of the gene's effect on the *CM-Factor* is mediated by the respective biomarker(s) and also phenome-wide MR analyses to explore potential side-effects across clinical endpoints.

Additional information on participants and data management in Oxford Biobank (OBB).

For a detailed overview of the OBB, refer to Karpe et al.[322] Established in 1999 and approved by the Oxfordshire Clinical Research Ethics Committee (reference 08/H0606/107+5), the OBB collects comprehensive health data, physical metrics, and biological samples from about 8,000 participants. These participants, ranging in age from 30 to 50 years, male and female, were primarily in good health and recruited from Oxfordshire via letters distributed by the Thames Valley Primary Care Agency. Individuals with severe health conditions such as heart disease, diabetes, cancer, autoimmune or psychiatric disorders, or addiction issues are generally excluded from the OBB. Consent is secured at the outset during a detailed screening at the Oxford Centre for Endocrinology, Diabetes and Metabolism, Churchill Hospital, Oxford, UK. This screening includes baseline health evaluations like body measurements and blood pressure checks.

Participants complete a thorough questionnaire with the help of research nurses to detail their lifestyle choices and family health history, especially relating to cardiovascular diseases and T2D. After fasting, blood samples are taken for various analyses, and additional examinations such as metabolic and genetic profiling, as well as body composition analysis using DXA scans, are conducted. Participants also consent to be available for future research based on specific genetic traits or physical characteristics.[322]

Druggable gene mediation analysis. We conducted mediation analysis to calculate the proportion of the effect of druggable genes on the *CM-Factor* mediated by the respective biomarkers with MR (**Figure 2.10**). For each of the druggable genes identified by the drug-target MR and colocalization steps that had associations with clinically-relevant biomarkers, we used the product of coefficient method[132] to estimate the mediation of the druggable gene by the biomarker.

These mediation analyses entailed performing both drug-target MR and conventional two-sample MR analyses. For the biomarker GWAS data, we identified the largest available GWAS summary statistics (from studies comprised of participants of European ancestry) for each of these biomarkers. For BMI, HbA1c, circulating lipid subfractions, ALT, and SBP, we used the same data sources as those described in the preceding drug-target MR analyses of approved and investigational cardiometabolic therapeutics using the exposures modeling the primary expected physiological responses to pharmacological modulation. For the GGT and AST, we used the Neale Lab release of the UK Biobank biochemistry data (N=344,136), and for fasting insulin and fasting glucose data, we used the recent GWAS data from the MAGIC (Meta-Analyses of Glucose and Insulin-related traits) Consortium[205] (N≤ 200,622).

First, we estimated the effect of each druggable gene on the biomarker(s) prioritized by the GWAS Catalog and Oxford Biobank assessment by performing drug-target MR of the druggable gene onto the biomarker, then multiplied this by the effect of biomarker on the *CM-Factor* from single-variable MR models using standard two-sample MR methods: each biomarker exposure was instrumented using conventional single-variable MR thresholds for SNP selection (i.e., independent SNPs associated with their respective biomarker at P-value < 5×10^{-8} and LD r^2 < 0.001). We used the inverse-variance weighted estimator (IVW) as our primary MR method, and also included the MR-Egger, weighted median, weighted mode, and simple mode estimators, which rely on different assumptions than IVW[323-325] because comparing the estimates across the methods facilitates the cross-validation of results under different sets of assumptions, enhancing confidence in the findings if results are consistent across methods.[48] Because heterogeneity in the MR estimates indicated violations of the instrumental variable assumptions,[317] we performed the Cochran Q heterogeneity test.[317] As with the drug-target MR analyses of approved or investigational targets, where Cochran's Q P-value indicated presence of heterogeneity in the IVW estimate (Q P-value < 0.05), we used the MR LASSO[218] method to identify outlier SNPs and provide MR estimates with heterogenous SNPs removed. We also used the Steiger directionality test[212] to evaluate whether there was evidence for reverse directionality between the biomarkers and *CM-Factor*.

After calculating the effect of each biomarker on the *CM-Factor*, the proportion of the total effect of the druggable gene on *CM-Factor* mediated by respective biomarker(s) was estimated

by dividing the biomarker-mediated effect ($\beta_{\text{biomarker -to-}CM\text{-}Factor}$) by the total effect ($\beta_{\text{druggable gene-to-}CM\text{-}Factor}$).[326,327] We used the product of coefficients method when estimating the effect of the biomarker mediator on the outcome ($\beta_{\text{biomarker -to-}CM\text{-}Factor}$) to avoid weak instrument bias, which has been used in previous mediation MR analyses that include exposures constructed with cis-instrument and conventional polygenic methods in the MR model.[132,326] Because there are cis-instrumental variables for each of the druggable genes and tens-to-hundreds of instrumental variables for the biomarkers, the association between the druggable gene and the *CM-Factor* would be substantially weakened in conventional multivariable MR models due to the large number of instrumental variables for the biomarkers.[326]

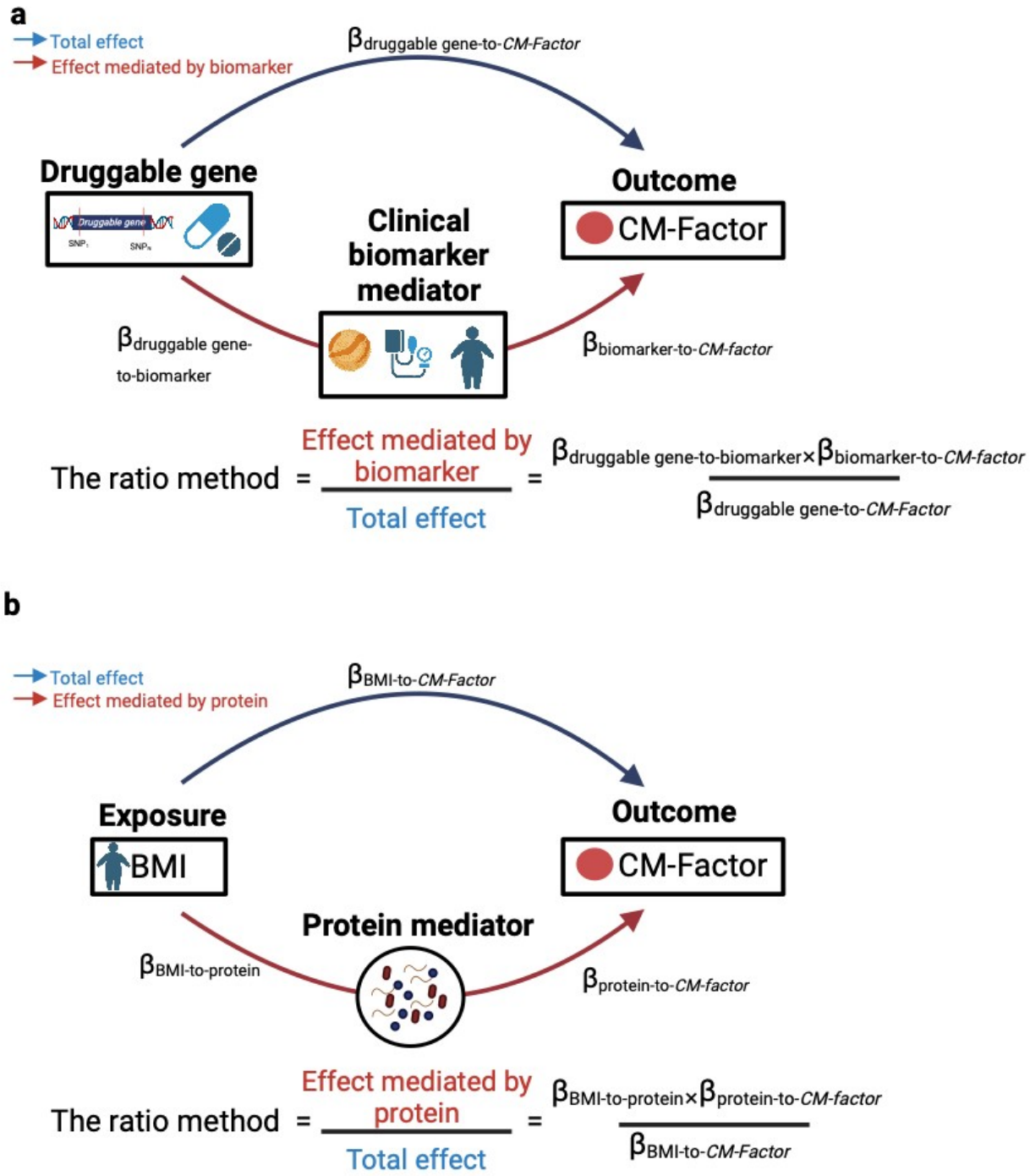


Figure 2.10. Overview of mediation analyses in (a) Aim 3 MR Study 1: Drug target MR screen and (b) Aim 3 MR Study 2: Two-step MR assessing the effect of body mass index (BMI) on the CM-Factor by the proteome. (a) *the dark blue arrow indicates the total impact of the druggable gene on the CM-Factor. The red arrow shows the influence of the druggable gene on CM-Factor that is mediated through the clinical biomarker (e.g., BMI, circulating lipids, blood pressure, or glycemic traits). To determine the extent of mediation by biomarker, we use the product of coefficients method, which involves multiplying* $\beta_{druggable\ gene\text{-}to\text{-}biomarker}$ *and* $\beta_{biomarker\text{-}to\text{-}CM\text{-}Factor}$*, then*

dividing the result by $\beta_{druggable\ gene\text{-}to\text{-}\ CM\text{-}Factor}$. Here, $\beta_{druggable\ gene\text{-}to\text{-}biomarker}$ represents the impact of the druggable on its biomarker identified via GWAS Catalog look-up and SNP-association tests in the Oxford Biobank; $\beta_{biomarker\text{-}to\text{-}\ CM\text{-}Factor}$ represents the influence of the respective protein on CM-Factor risk; and $\beta_{druggable\ gene\text{-}to\text{-}\ CM\text{-}Factor}$ denotes the total impact of druggable gene on the CM-Factor. We assessed the mediated proportion for the effect of the druggable genes on the CM-Factor for each of the biomarkers linked with the druggable genes in the GWAS Catalog look-up and SNP-association tests in the Oxford Biobank. ***(b)*** *As in* ***(a)****, the dark blue arrow indicates the total impact of BMI on the CM-Factor. The red arrow shows the influence of BMI on CM-Factor that is mediated through protein. To determine the extent of mediation by protein, we use the product of coefficients method, which involves multiplying $\beta_{BMI\text{-}to\text{-}protein}$ and $\beta_{protein\text{-}to\text{-}\ CM\text{-}Factor}$, then dividing the result by $\beta_{BMI\text{-}to\text{-}\ CM\text{-}Factor}$. Here, $\beta_{BMI\text{-}to\text{-}protein}$ represents the impact of BMI on one of the proteins prioritized by the two-step Mendelian randomization (MR) analyses; $\beta_{protein\text{-}to\text{-}CM\text{-}Factor}$ represents the influence of the respective protein on CM-Factor risk; and $\beta_{BMI\text{-}to\text{-}CM\text{-}Factor}$ denotes the total impact of BMI on the CM-Factor. We assessed the mediated proportion for the effect of BMI on the CM-Factor for each of the proteins prioritized by steps 1 and 2 in the two-step MR analyses. Created in BioRender. Rosoff, D. (2025) https://BioRender.com/efhdhzj*

Replication of the top druggable genes in disease-relevant tissues. We selected the whole blood eQTLGen data as the primary resource for our MR screening of the druggable genome due to its large sample size (N ~31,000), significantly surpassing the available bulk tissue eQTL datasets from potentially relevant tissues, such as the largest muscle eQTL dataset (e.g., ~500 samples from GTEx V8).[287] This larger sample size enhances the statistical power for target identification and prioritization. Furthermore, prior studies have demonstrated a strong correlation in gene expression and eQTLs across various tissues,[287,328,329] supporting the utility of whole blood data in target discovery efforts. In fact, the eQTLGen whole blood dataset has already been effectively utilized as the primary exposure in drug-target MR studies for outcomes such as Parkinson's disease, instead of relying on brain eQTLs for similar reasons.[240]

However, tissue-specific gene expression and regulatory mechanisms can vary widely, meaning that drug targets identified in one tissue may not have the same effects in another.[330,331] Including tissues that are directly related to the disease or condition of interest may facilitate the identification of genes and pathways that play a critical role in disease development and progression. This approach not only improves the accuracy of target prioritization but also increases the likelihood that the identified targets will have clinical relevance and therapeutic potential in treating the disease.[330,331]

Therefore, we aimed to incorporate analyses using disease-relevant tissue into the MR Study 2 study design. We extracted eQTLs and created cis-instruments proxying the bulk tissue gene expression for each of the 41 targets prioritized by the drug-target MR and colocalization of the 2,567 druggable genes in the eQTLGen whole blood data. We used the following bulk tissue eQTL data from GTEx V8 (European ancestry):[57,287] subcutaneous adipose (N=479); adipose visceral omentum (N=393); METSIM (METabolic Syndrome In Men) adipose (N=563); aortic artery (N=329); coronary artery (N=175); tibial artery (N=476); heart tissue, including the atrial appendage (N=316) and left ventricle (N=327); liver (N=178); skeletal muscle (N=588); pancreas (N=243); and whole blood (N=558). The same methods were used to create the bulk tissue instruments and MR was performed as in the initial screen (and in each of the

drug-target/cis-instrument MR analyses in MR Studies 1-3). Across all tissues we were able to successfully instrument 36 of the 41 targets with the instrumentation thresholds used for the main analyses. No tissue had eQTLs for all 36 genes (the GTEx whole blood dataset had the most eQTLs for replication [30] and the GTEx liver tissue had the fewest [7 eQTLs]). Overall, we performed 186 tests for replication. We used several P-value thresholds: the stringent threshold of 1.96×10^{-5} used in the main screen of the druggable genome as very strong evidence for replication; a threshold of 2.27×10^{-4} (0.05/186 total tests) for strong evidence of replication, and a nominal P-value threshold of 0.05 as suggestive evidence of replication. We looked for directional consistency among the MR estimates from the eQTLGen results and took results with MR estimate P-values < 2.27×10^{-4} forward for colocalization.

Phenome-wide MR to assess potential side-effect profiles druggable genes. Due to the role adverse side effects play in the failure of therapeutics during drug development,[332] we aimed to enhance our understanding of the therapeutic potential of the 41 druggable genes. Therefore, we performed phenome-wide MR studies involving 366 diseases and biomarkers (**Table AP5.1**). Inclusion criteria comprised studies conducted in cohorts of European ancestry, sample sizes of at least 1,000 participants, with a minimum of 100 cases for binary variables, and availability of summary statistics (betas, standard errors, effect alleles) for 100,000 SNPs. The cis-instrument MR analysis, outlined in the preceding sections, was used. A Bonferroni-corrected P-value threshold of 1.37×10^{-4} (0.05/366 outcomes) was used to determine associations with biomarkers or diseases. We evaluated the therapeutic implications of gene expression changes by comparing the directions of cis-instrument MR estimates against the direction indicated by their effects on the CM-Factor. Specifically, a positive MR estimate for a gene on the CM-Factor suggests that reducing the gene's expression would be beneficial, while a negative estimate implies that increasing its expression may be favorable. Traits with MR estimates aligning with this therapeutic direction were classified as "beneficial effects," whereas those in the opposite direction were classified as "adverse effects." This approach provides a nuanced understanding of the role of druggable genes across phenotypes, identifying traits where gene expression changes may support or oppose therapeutic goals.

2.4.4c. MR Study 3: Two-step Mendelian randomization to identify circulating proteins mediating the impact of body mass index (BMI) on the CM-Factor.

Background. The ongoing obesity epidemic presents a critical and escalating global health challenge.[127] The prevalence has increased substantially in the last several decades, now affecting >1 billion adults globally.[333] This rise is especially pronounced in Western countries, where statistics such as a jump from 30.5% to 42.4% in the United States from 2000 to 2018 highlight the scale of the crisis. Europe and increasingly, lower-income nations, are witnessing similar trends.[127,333] While obesity is linked with increased risk for many diseases and adverse health conditions, cardiometabolic diseases are the leading causes of death.[128]

While there are many mechanisms through which obesity increases cardiometabolic disease risk, including obesity-related effects on metabolism, the endocrine system, immune functioning, and hemodynamic dysregulation,[129] there is also a growing body of literature suggesting that obesity impacts cardiometabolic disease risk through indirect mechanisms,[129] and elucidating new

biological pathways is essential for developing targeted interventions that can reduce the burden of obesity and resultant cardiometabolic disease.[129]

Circulating proteins can serve as diagnostic markers and potential therapeutic targets (most of the approved therapeutics target proteins),[50,126] targets and recent work has shown that increased body mass index (BMI) and obesity are linked with widespread changes in the plasma proteome,[130,131] suggesting that clarifying link between obesity and cardiometabolic disease and elucidating circulating proteins that mediate the impact of obesity on cardiometabolic outcomes, such as the *CM-Factor*, could highlight new avenues for therapeutic interventions, offering potential targets to mitigate these risks. Therefore, we used a two-step MR[132] approach (**Figure 2.11**) leveraging the genetic signature from the largest available BMI GWAS,[334] data on > 2,900 circulating proteins from individuals of European ancestry,[134] and our *CM-Factor*, to first identify the proteomic consequences of increased BMI and then elucidate the BMI-affected proteins may mediate the impact of BMI on cardiometabolic health.

Figure 2.11. Overview of the two-step MR leveraging the circulating proteome to characterize the impact of BMI on the CM-Factor. *These analyses used a two-step MR approach to investigate the influence of BMI on the CM-Factor through alterations in circulating proteins. Initially, a genetic signature was derived from a large BMI meta-analysis and data on ~2,900 proteins from the UKB Pharma Proteomics Project (Olink) to identify proteins affected by BMI. Conventional MR methods identified independent SNPs associated with BMI. In the second step, BMI-associated proteins were analyzed using cis-instrumentation to ensure direct influence on protein levels, successfully instrumenting 455 proteins. These were harmonized with the CM-Factor, and mediation analysis was conducted to quantify the proteins' mediation effects. Colocalization analyses validated the causality of MR results, with further validation using*

observational data from the INTERVAL study and replication with deCODE proteomics data by Ferkingstad et al. measured using the SomaScan version 4 assay (SomaLogic),[56] and obesity diagnoses in the FinnGen cohort (release 11). This robust approach identified proteins mediating the relationship between BMI and the CM-Factor, highlighting potential therapeutic targets. Created in BioRender. Rosoff, D. (2025) https://BioRender.com/qy7zvq9

Step 1: identify proteins with levels affected by BMI. In Step 1, we evaluated the impact of genetically increased BMI on ~2,900 circulating proteins (N≤ 34,557) measured on the Olink Explore proteomics assay comprising the first release of the UKB Pharma Project (UKB-PPP).[134] For the BMI exposure, we used the meta-analysis of the GIANT and UKB performed by Pulit et al. (N=694,649).[133] We used conventional two-sample MR instrumentation methods to genetically model increased BMI: we extracted independent SNPs (LD $r^2 < 0.001$) associated with BMI at P-value < 5×10^{-8} located throughout the genome and then extracted these variants from the downloaded and prepared UKB-PPP data. Because the *CM-Factor* GWAS only included autosomes, we excluded the 86 proteins in the UKB-PPP data that were located on the X chromosome because we would not be able to perform cis-instrument MR of these proteins onto the *CM-Factor* in Step 2. Therefore, our Step 1 analyses included 2,835 circulating proteins. We next harmonized the BMI exposure and each of the UKB-PPP protein datasets and performed Steiger filtering[212] to identify and remove outlier SNPs, ensuring appropriate directionality of the BMI instrument. We tested the instrumental variable relevance assumption by calculating the variance explained by the instruments and F-statistics for each SNP comprising the polygenic BMI instrument.

We used the inverse-variance weighted estimator (IVW) as our primary MR method, and also included the MR-Egger, weighted median, weighted mode, and simple mode estimators, which rely on different assumptions than IVW[323-325] because comparing estimates across the methods cross-validates results under different sets of assumptions, enhancing confidence in the findings if results are consistent across methods.[48] Because heterogeneity in the MR estimates indicated violations of the instrumental variable assumptions,[317] we performed the Cochran Q heterogeneity test.[317] As with the drug-target MR analyses of approved or investigational targets, if Cochran's Q P-value indicated presence of heterogeneity in the IVW estimate (Cochran Q P-value < 0.05), we used the MR LASSO[218] method to identify outlier SNPs and provide MR estimates with heterogenous SNPs removed. We also used the Steiger directionality test[212] to evaluate whether there was evidence for reverse directionality between the BMI exposure and protein outcomes. To account for multiple comparisons, we used a Bonferroni-corrected P-value threshold of 1.76×10^{-5} (0.05/2,835 proteins) as a heuristic to guide the screening for Step 2.

Step 2: Screen BMI-associated proteins for effects on the CM-Factor. In Step 2, we focused our instrumentation of the BMI-associated proteins on variants located within or near each protein's corresponding genomic regions (i.e., cis-pQTLs) because this proximity improvise the likelihood that the genetic variant is more likely to influence the target proteins exposure directly rather than other genes, reducing the risk of confounding and improves the accuracy of the causal inference regarding the target.[6,50] Cis-instrumentation for screening potential drug-targets also improves biological plausibility, making interpretation of the causal pathway more straightforward, and more likely provides a direct link to the gene encoding the drug target, providing more relevant insights into the potential efficacy and safety of modulating that target.

This relevance is particularly important for identifying and validating novel drug targets.[6,50] We were able to successfully cis-instrument 455 of the 506 circulating proteins with associated with BMI from Step 1 using pQTLs (LD $r^2 < 0.1$) associated with the protein at P-value $\leq 5\times10^{-8}$ within a ±500 kb window of the protein coding genomic locus coordinates. We harmonized these datasets with the *CM-Factor* and removed any SNP that indicated incorrect directionality by Steiger filtering. We performed cis-instrument MR of these 455 proteins with the *CM-Factor* as the outcome using the IVW estimator as the primary methods for cis-instruments with 2+ SNPs and the Wald ratio for cis-instruments comprised of a single variant. We used the Steiger directionality test[212] to assess reverse causation of the MR estimates. For instrumental variables with 2+ SNPs, we conducted Cochran Q heterogeneity tests[317] and for instrument with 3+ SNPs, we also performed MR with the MR-Egger, weighted median, weighted mode, and simple mode estimators.[323-325] We used a Bonferroni-corrected P-value threshold of 1.10×10^{-4} (0.05/455 [number of proteins tested in the Step 2 MR]) as a heuristic to guide follow-up and validation analyses and looked for proteins with estimates surpassing correction for multiple testing in both Step 1 and Step 2 that also demonstrated directions consistent with mediation of the adverse impact of increased BMI on the *CM-Factor*, which was expected given the literature associating obesity with increased risk for cardiometabolic disease, and the direction of the single-variable MR results of BMI on the *CM-Factor* performed as part of the two-step MR analyses.

Mediation analysis. As with the MR Study 2, we undertook mediation analysis to calculate the proportion of the effect of BMI on the *CM-Factor* mediated by the respective circulating proteins using network MR (**Figure 2.10**). For each of the proteins surpassing correction for multiple comparisons and demonstrating directionally consistent MR estimates between each step, we used the product of coefficient method[132] to estimate the circulating protein-mediated effect (the effect of BMI on *CM-Factor* that was accounted for by circulating proteins identified in steps 1 and 2) and the same instrumental variables as described above.

First, we estimated the effect of BMI on each of the circulating protein prioritized by Steps 1 and 2, then multiplied this by the effect of circulating protein on the *CM-Factor*. Subsequently, the proportion of the total effect of BMI on *CM-Factor* mediated by circulating protein was estimated by dividing the circulating protein-mediated effect ($\beta_{\text{circulating protein-to-}CM\text{-}Factor}$) by the total effect ($\beta_{\text{BMI-to-}CM\text{-}Factor}$).[326,327] We used the product of coefficients method without adjusting for BMI when estimating the effect of the mediator on the outcome ($\beta_{\text{circulating protein-to-}CM\text{-}Factor}$) to avoid weak instrument bias, which has been used in previous two-step MR analyses.[132,326] We did so because the exposure adjustment requires multivariable MR using circulating protein and BMI as exposures; however, there are only cis-instrumental variables for each circulating protein (i.e., cis-pQTLs) and hundreds of instrumental variables for BMI. Thus, when using multivariable MR, which includes instrumental variables from both the polygenic BMI and cis-instrument proteomic exposures in the model, the association between plasma levels of circulating protein and the genetic variants would be substantially weakened due to the large number of instrumental variables for BMI decreasing the strength of the association between plasma levels of circulating protein and the genetic variants).[326]

Replication. We also aimed to replicate the prioritized targets from Steps 1 and 2 using independent proteomic data from deCODE (N=35,559; Icelandic ancestry) measured using the SomaScan version 4 assay (SomaLogic).[56] deCODE had full GWAS data for 6 of the 8

prioritized proteins (ENO3, MSR1, FSTL3, SHBG, PTPRR, and CD34). We performed replication for both Steps 1 and 2 (BMI onto the 6 proteins using the single variable MR methods described above and cis-instrument MR of the proteins onto the *CM-Factor* using cis-instrument MR methods). For each of the cis-instrument findings that replicated at P-value < 0.05, we also performed colocalization as with the primary analyses. We also performed an additional replication of the prioritized proteins using another exposure data source, i.e., we used obesity diagnoses in the latest release of the FinnGen cohort (Release 11) (27,711 obesity cases/425,881 controls)[269] and assessed the impact of diagnosed obesity on the 8 proteins measured in the UKB-PPP data. These two-sample MR analyses were performed with the same methods as those used in the primary Step 1 MR analyses with BMI.

Validation with observational data. We used conventional observational data from the INTERVAL study[335] to investigate whether the top proteins demonstrated evidence of being associated with increased BMI. See the original publication[335] for additional details regarding the INTERVAL study. Briefly, between 2012 and 2014 the INTERVAL study recruited about 50,000 English participants of primarily European ancestry without self-reported major diseases. Previous work by Goudswaard et al. used the SomaScan assay to measure 3,622 plasma protein levels in a subset of the INTERVAL participants (N=2,737) to investigate associations between BMI and plasma protein levels, adjusting for age, sex, and other confounders, with the regression estimates representing a normalized standard deviation unit difference in each protein level per one standard deviation (4.8 kg/m²) increase in BMI.[130] 5 of the 8 top proteins from the two-step MR were found in the INTERVAL data (i.e., MSR1, FSTL3, SHBG, PTPRR, and CD34). We looked for consistency in the direction of the Step 1 estimates and the estimates of the observational regression models of BMI on each of the proteins and used a nominal P-value threshold (P-value=0.05) to define whether the protein level was impacted by BMI in the INTERVAL data.

Colocalization of top proteins with CM-Factor. We next conducted colocalization analyses to evaluate whether the MR results from Step 2 were likely to be causal and not biased by LD, which may cause confounding.[265] We used the *coloc* package[266] and included all variants ±500 kb of the gene's genomic boundaries in the analysis. We calculated the posterior probabilities using the default priors (i.e., $p1 = p2 = 10^{-4}$ and $p12 = 10^{-5}$) in *coloc* and considered a posterior probability hypothesis 4 (PP.H4) >0.6 as evidence that both traits (here a pQTL and the *CM-Factor*) share a single causal variant within the genomic locus of the target gene. Low posterior probabilities for the third (H3, both trait 1 and trait 2 are associated, but with separate SNPs) and fourth hypotheses and a corresponding high posterior probability for the first hypothesis (H1, only trait 1 has a genetic association in the locus) suggest that the colocalization analysis is underpowered, potentially because the outcome dataset does not have sufficiently strong genetic signals in the locus.[265] Therefore, for colocalization results meeting these criteria (low PP.H3 and PP.H4 and a high PP.PP. H4 by calculating PP.H4/(PP.H3+PP.H4) that has been used previously in MR studies using eQTL and pQTL exposure sources.[265,319,320]

Phenome-wide MR to assess potential side-effect profiles of BMI-associated proteins with directionally consistent effects on the CM-Factor. As with the MR Study 2 prioritizing druggable genes for involvement with the *CM-Factor*, we sought to improve our understanding of the therapeutic potential of the 4 proteomic mediators with directionally consistent MR

estimates aligning with the adverse BMI-to-*CM-Factor* relationship (i.e., GGT1, PTPRR, ENO3, and SHGB). Therefore, we performed phenome-wide MR studies for the 4 proteomic mediators on 366 diseases and biomarkers as described in the methods for MR Study 2, using a Bonferroni-corrected P-value threshold of 1.37×10^{-4} (0.05/366 outcomes), and comparing the directions of the cis-instrument MR estimates with the indicated direction of the cis-instrument MR estimate that would be therapeutically indicated by its effects on the *CM-Factor*.

CHAPTER 3: AIM 1 RESULTS

Chapter Overview

This chapter explores the application of drug-target Mendelian Randomization (MR) to investigate the causal effects of genetically proxied LDL-C lowering through PCSK9 and HMGCR inhibition on type 2 diabetes (T2D) risk and glycemic markers across diverse global populations. By leveraging large-scale genetic data from East Asian (EAS), South Asian (SAS), African (AFR), Hispanic (HISP), and European (EUR) cohorts, the chapter provides a comprehensive evaluation of the therapeutic safety profiles of these lipid-lowering strategies, highlighting population-specific differences and their implications for clinical practice.

3.1. PCSK9 and HMGCR Instrument Strength

F-statistics for genetic variants comprising the *PCSK9* and *HMGCR* drug-target instruments in each population were strong (**Table AP2.2**): the average F-statistics for LDL-C lowering via *PCSK9* genetic variants were 112.45, 40.13, 109.3, 64.15, and 279 for EAS, SAS, AFR, HISP, and EUR, respectively, while the average F-statistics for LDL-C lowering via genetic *HMGCR* were 60.1, 72.2, 55.3, 83.1, and 241.3 for EAS, SAS, AFR, HISP, and EUR, respectively. The alternate PCSK9 instruments comprised of functional variants (**Table AP2.3**) and the QTL instruments (**Table AP2.4**) were similarly strong, suggesting that the drug-target instruments are unlikely to be subject to weak instrument bias[212] (**Table AP2.5**). Further, for PCSK9 and HMGCR, the instruments generally explained comparable variance in circulating LDL-C levels: on average, PCSK9 instruments accounted for 1.5% of the variance, with the EAS PCSK9 instrument explaining the most variance (3.9%) and the SAS instrument explaining the least (0.5%). HMGCR instruments explained on average 0.36% (AFR explained the least with 0.12% and EAS explained the most with 0.62% of the variance) of the LDL-C levels. Instruments for each polygenic LDL-C instrument were also strong (**Table AP2.2**).

3.2. The Impact of LDL-C Lowering on T2D by Genetically Mimicked PCSK9 and HMGCR Inhibition

Estimates of genetically proxied PCSK9 inhibition with the risk for T2D in the SAS, EAS, HISP, and EUR populations each included the null (**Figure 3.1, Table AP2.6**). These results broadly aligned with the estimates derived using functional variants (analyses of both the gain-of-function R46L [rs505151][47] and loss-of-function E670G [rs11591147][208] variants in the SAS, HISP, and EUR populations and the gain-of-function R46L in the EAS population) within the *PCSK9* locus (**Table AP2.7**). In AFR, there was weak evidence that LDL-C lowering via PCSK9 variants increased T2D risk (OR per 1-SD lower LDL-C levels=1.53, CI=[1.058, 2.22], P-

value=0.024); however, estimates were not consistent across instruments: the PCSK9 AFR estimate derived from functional variants included the null. By contrast, 2 of the 5 population-specific estimates (EAS and EUR) demonstrated strong evidence that genetically proxied HMGCR inhibition increased T2D risk, which also aligned in direction with the SAS estimate that showed weak evidence of a genetics-based relationship (OR=1.698, CI=[1.051,2.743], P-value=0.031). HMGCR estimates were generally consistent across MR methods and the MR Egger intercept estimates did not demonstrate evidence for pleiotropy.

Regarding colocalization of the T2D results with P-values < 0.05, we observed evidence of a shared causal variant between LDL-C and T2D in the EAS, EUR, and SAS populations (**Table AP2.8**). The PCSK9-T2D finding in AFR did not colocalize (PP.H4=0.033). However, we observed evidence of a single shared causal variant between LDL-C and T2D risk in the *HMGCR* locus in both the EAS and SAS populations (EAS PP.H4=0.682 and SAS PP.H4=0.813) while the initial single-variant colocalization of the *HMGCR* locus in EUR suggested multiple causal variants (PP.H3=0.98). Follow-up SuSiE colocalization confirmed evidence of shared multiple causal variants between LDL-C and T2D in the *HMGCR* locus among EUR (SuSiE PP.H4=0.76) (**Table AP2.8**).

Figure 3.1. Mendelian randomization results of the impact of PCSK9 and HMGCR inhibition on T2D risk. *Results report the inverse variance weighted (IVW) estimates from MR analyses that incorporated correlation between SNPs. Because there were only 2 SNPs in the AFR and HISP HMGCR instruments, the MR Egger method was not performed (requires 3+ SNPs). Results for the type 2 diabetes (T2D) are reported as odds ratio (OR) change (with 95% confidence interval [CI]) in T2D risk per standard deviation decrease in the low-density lipoprotein cholesterol (LDL-C) levels via variants within the PCSK9 and HMGCR genomic loci. IVW MR estimates surpassing correction for multiple comparisons (P-value < 0.005 (0.05/10 tests performed per population) are indicated with an asterisk ("*"). "# SNPs" is the number of genetic variants used in the drug-target MR analysis. OR: odds ratio; CI: confidence interval; PCSK9: Proprotein-convertase subtilisin/kexin 9; HMGCR; 3-hydroxy-3-methylglutaryl coenzyme A reductase. Reproduced from Rosoff DB, Wagner J, Jung J, Pacher P, Christodoulides C, Davey Smith G, Ray DW, Lohoff FW. "Multiomic Mendelian Randomization Study Investigating the Impact of PCSK9 and HMGCR Inhibition on Type 2 Diabetes Across Five Populations." Diabetes 2025; 74(1): 120–130. © 2025 The Authors. Licensed under Creative Commons Attribution 4.0 International (CC BY 4.0).*

3.3. The Impact of LDL-C Lowering on Glycemic Markers by Genetically Mimicked PCSK9 and HMGCR Inhibition

IVW and Wald ratio results from the primary drug-target MR analyses assessing the impact of LDL-C lowering via either the *PCSK9* or *HMGCR* loci on glycemic markers are presented in **Figure 3.2**. While none of the PCSK9 estimates surpassed the multiple comparisons threshold used to define strong evidence, we did observe several PCSK9 estimates with weak evidence for an impact on glycemic markers, including reduced HbA1c in SAS, and increased fasting insulin in EAS (**Table AP2.6**), the latter robust and directionally consistent in MVMR correcting for potential bias in the adjustment for the heritable covariates BMI in the MAGIC glycemic trait GWASs[219] (**Tables AP2.9-AP2.13**) (MVMR not available for SAS). The PCSK9-HbA1c finding in the SAS population was supported with evidence of colocalization (PP.H4=0.835), and there was some evidence, albeit not surpassing the study threshold of PP.H4 >0.6, for colocalization between LDL-C levels and HbA1c in the EAS population (PP.H4=0.537) (**Table AP2.8**). As with the T2D analyses, estimates on glycemic markers for PCSK9 instruments comprised of the R46L and E670G functional variants spanned the null (**Table AP2.7**). We found strong evidence that genetically proxied HMGCR inhibition increased HbA1c in HISP (β=0.167, CI=[0.059,0.275], P-value=0.002), which was supported by evidence of a single shared causal variant (PP.H4=0.813) (**Table AP2.8**) and robust in MVMR accounting for BMI (**Table AP2.12**), and also aligned with weak evidence for an increase in HbA1c in EUR (β=0.040, CI=[0.001,0.0979], P-value=0.043) (however the EUR finding was not robust in MVMR [**Table AP2.13**]). By contrast, there was no evidence of colocalization in the HMGCR locus for any of the other glycemic traits with drug-target MR estimate P-values < 0.05. Finally, for each population, the PCSK9 and HMGCR drug-target estimates were generally consistent across the complementary drug-target MR methods, strengthening causal inference (**Table AP2.4**).

Figure 3.2. Mendelelian randomization results of PCSK9 and HMGCR inhibition on glycemic markers. *Presented are MR results of the impact of PCSK9* ***(A)*** *and HMGCR* ***(B)*** *inhibition on glycemic traits results are reported with the inverse variance weighted (IVW) or Wald ratio estimates from MR analyses. 2-hour glucose levels were not available for SAS. Results for the glycemic markers are reported as the regression coefficient (β) (with 95% confidence interval [CI]) in the respective glycemic marker per standard deviation decrease in the low-density lipoprotein cholesterol (LDL-C) levels via variants within the PCSK9 and HMGCR genomic loci. MR estimates surpassing correction for multiple comparisons (P-value < 0.005 (0.05/10 tests performed per population) are indicated with an asterisk ("*"). "# SNPs" is the number of genetic variants used in the drug-target MR analysis. OR: odds ratio; CI: confidence interval; PCSK9: Proprotein-convertase subtilisin/kexin 9; HMGCR; 3-hydroxy-3-methylglutaryl coenzyme A reductase: HbA1C: glycated hemoglobin; AFR: African; EAS: East Asian; SAS: South Asian; HISP: Hispanic; EUR: European. Reproduced from Rosoff DB, Wagner J, Jung J, Pacher P, Christodoulides C, Davey Smith G, Ray DW, Lohoff FW. "Multiomic Mendelian Randomization Study Investigating the Impact of PCSK9 and HMGCR Inhibition on Type 2 Diabetes Across Five Populations." Diabetes 2025; 74(1): 120–130. © 2025 The Authors. Licensed under Creative Commons Attribution 4.0 International (CC BY 4.0).*

Regarding measures of insulin resistance after a glucose challenge in EUR, there was weak evidence that a 1-SD reduction in LDL-C via *PCSK9* variants was associated with increased IFC (β=0.14, CI=[0.03,034], P-value=0.009) but not ISI (β=-0.07, CI=[-0.16,0.012], P-value=0.092), while *HMGCR* variants were weakly linked with reduced ISI (β=-0.16, CI=[-0.28,-0.03], P-value=0.013) (**Table AP2.14**). These MR estimates aligned with the results using PCSK9 instruments comprised of functional variants for the populations in which these functional variants were available (that is, both the gain-of-function R46L and loss-of-function E670G were analyzed in SAS, AFR, HISP, and EUR, while only R46L was analyzed in EAS because E670G was not present in the EAS LDL-C GWAS data) (**Tables AP2.3, AP2.7**).

Given the mechanisms of action of the approved PCSK9 inhibitors (i.e., anti-PCSK9 monoclonal antibodies lowering circulating PCSK9 protein levels[56] and inclisiran inhibiting hepatic PCSK9 expression),[57] we investigated the impact of genetically lowered circulating PCSK9 protein, hepatic *PCSK9* expression, and pancreatic *PCSK9* expression on T2D and glycemic markers. These results generally aligned with results from analyses evaluating LDL-C lowering via variants within the *PCSK9* locus (**Figures 3.3, 3.4**, **Table AP2.15**). We did observe weak evidence that genetically lowered circulating PCSK9 protein levels were linked with reduced HbA1c levels and 2-hour glucose; however, we failed to find corresponding relationships with the other glycemic markers, and the PCSK9-HbA1c finding was not replicated in analyses using instruments proxying hepatic (or pancreatic) *PCSK9* expression. In the IFC and ISI analyses, the PCSK9 pQTL instruments aligned with the results of the LDL-C and functional instruments discussed above: there was weak evidence that PCSK9 inhibition increased IFC (β=0.04, CI=[0.009,0.078], P-value=0.013) (**Table AP2.15**).

Figure 3.3. Mendelelian randomization results of additional PCSK9 instruments (functional variants and QTLs) on T2D risk. *Presented are MR results (either IVW or Wald ratio depending on number of cis-variants in the instrument) of the impact of the additional PCSK9 instruments (functional variants lowering LDL-C, tissue-specific PCSK9 expression, and circulating PCSK9 protein levels) on T2D risk. Results for T2D are reported as the odds ratio (OR) (with 95% confidence interval [CI]) in the respective glycemic marker per standard deviation decrease in either LDL-C, circulating PCSK9 protein levels, or transcripts per million in pancreas and liver PCSK9 expression. "# SNPs" is the number of genetic variants used in the drug-target MR analysis. PCSK9: Proprotein-convertase subtilisin/kexin 9; QTL: quantitative trait loci; HbA1c: glycated hemoglobin; AFR: African; EAS: East Asian; SAS: South Asian; HISP: Hispanic; EUR: European. Reproduced from Rosoff DB, Wagner J, Jung J, Pacher P, Christodoulides C, Davey Smith G, Ray DW, Lohoff FW. "Multiomic Mendelian Randomization Study Investigating the Impact of PCSK9 and HMGCR Inhibition on Type 2 Diabetes Across Five Populations." Diabetes 2025; 74(1): 120–130. © 2025 The Authors. Licensed under Creative Commons Attribution 4.0 International (CC BY 4.0).*

Figure 3.4. Mendelelian randomization (MR) results of additional PCSK9 instruments (functional variants and QTLs) on glycemic traits. *Presented are MR results (either IVW or Wald ratio depending on number of cis-variants in the instrument) of the impact of the additional PCSK9 instruments (functional variants lowering LDL-C, tissue-specific PCSK9 expression, and*

circulating PCSK9 protein levels) for glycemic traits. Glycemic traits results are reported the β's (with 95% CI) per standard deviation decrease in either LDL-C, circulating PCSK9 protein levels, or transcripts per million in pancreas and liver PCSK9 expression. Note that 2-hour glucose levels were not available for SAS. "# SNPs" is the number of genetic variants used in the drug-target MR analysis. PCSK9: Proprotein-convertase subtilisin/kexin 9; QTL: quantitative trait loci; HbA1c: glycated hemoglobin; AFR: African; EAS: East Asian; SAS: South Asian; HISP: Hispanic; EUR: European. Reproduced from Rosoff DB, Wagner J, Jung J, Pacher P, Christodoulides C, Davey Smith G, Ray DW, Lohoff FW. "Multiomic Mendelian Randomization Study Investigating the Impact of PCSK9 and HMGCR Inhibition on Type 2 Diabetes Across Five Populations." Diabetes 2025; 74(1): 120–130. © 2025 The Authors. Licensed under Creative Commons Attribution 4.0 International (CC BY 4.0).

3.3. Polygenic LDL-C Results.

Full results of the polygenic lipid instrument are presented in **Table AP2.10**. Prior to interpretation, we aligned each polygenic LDL-C estimate to correspond to the expected physiological response to pharmacological lipid-lowering therapy, i.e., a change in T2D risk or glycemic marker level per unit SD lowering in LDL-C. In line with the results observed using HMGCR variants, we observed strong evidence that lower LDL-C levels increased risk for T2D in SAS (OR=1.34, 95% CI=[1.166,1.529], P-value=2.81×10^{-5}) and in EUR (OR=1.056, 95% 95% CI=[1.013,1.101], P-value=0.001). Polygenic LDL-C estimates in the other three populations included the null.

3.4. Aim 1 Discussion

We used drug-target MR to compare the relationships of genetic LDL-C lowering via *PCSK9* and *HMGCR* variant therapies with T2D and glycemic markers using data from five populations. We found a neutral safety profile for PCSK9 inhibition on T2D in SAS, EAS, HISP, and EUR populations, adding to the growing body of genetics-based literature finding generally safe side effect profiles of long-term PCSK9 lowering.[52,157,167,336] Sensitivity analyses using functional PCSK9 variant R46L, PCSK9 protein levels, and both hepatic and pancreatic *PCSK9* gene expression similarly yielded null results. Our assessment of the genetic PCSK9-T2D relationships across cohorts representing 5 populations using complementary MR methods, sensitivity analyses based upon functional PCSK9 variants, and analyses using multi-omic PCSK9 instruments, along with several other lines of evidence failing to find an adverse increase in T2D risk by PCSK9 inhibition,[52,54] further strengthens our inference of the neutral side-effect profile and should be reassuring for any concerns regarding T2D diabetes risk from pharmacological PCSK9 inhibition. Importantly, while the global T2D prevalence is high, it varies widely across geographical regions and by race/ethnicity,[23-26] and epidemiological data suggest that certain populations may have higher or lower risk of developing T2D.[26] Nevertheless, despite the need for more diversity in all clinical trials and genetics-based studies,[28-32] apart from a recent study investigating the relationships of lipids, lipid-lowering targets, and T2D risk among African Americans,[54] the existing PCSK9-T2D literature is based

primarily upon analysis of individuals in European populations, highlighting the need for population-specific work to inform our understanding of PCSK9 inhibition and the risk for T2D.

The main analyses found weak evidence (i.e., P-value < 0.05 but not surpassing correction for multiple testing) that PCSK9 inhibition was associated with increased T2D risk in AFR; however, the result was not robust to our sensitivity analyses using functional PCSK9 variants as instruments. In addition, we did not find a relationship of HMGCR and T2D risk in the AFR data. Our AFR T2D results were based upon cohorts from continental Africa, and neither the PCSK9 nor the HMGCR finding align with the recent MR study by Soremekun et al. assessing the impact of LDL-C lowering by PCSK9 and HMGCR inhibition on T2D risk among African Americans in the Million Veterans Program (MVP).[54] Soremekun et al. found that HMGCR inhibition increased T2D risk (OR per SD decrease in LDL-C=1.68, CI=[1.03,2.72], P-value=0.04) but PCKS9 inhibition did not.[54] These discrepancies may also reflect the impact of genetic admixture, the mixing of different ancestral populations,[337] which can influence genetic associations by introducing variability in allele frequencies and genetic backgrounds across populations.[337] In the context of genetic studies, admixture can lead to confounding effects, where associations identified in one population may not hold true in another due to differences in genetic architecture.[337] This variability can obscure or inflate the true effects of genetic variants on traits such as T2D risk. For example, in populations with a high degree of admixture, such as African American populations,[338,339] the presence of alleles from different ancestral backgrounds may alter the expression and impact of genes targeted by therapies like HMGCR inhibitors or PCSK9 inhibitors. As a result, the observed genetic associations in a more genetically homogenous population may differ when studied in an admixed population, leading to discrepancies in the findings. More broadly, Africans possess significantly more genetic and linguistic diversity, with over 3,000 indigenous languages, largely shaped by geography. However, over 90% of these ethnolinguistic groups lack genetic data.[28] Focusing on African diaspora populations and broadly categorizing them as African ancestry overlooks Africa's genetic diversity, perpetuating imbalances and health disparities,[28] underscoring the critical importance of future investigation and replication with these populations when the data becomes available.

For HMGCR inhibition, our results replicate and extend reports of increased T2D risk from RCTs evaluating statin use and MR analyses using variants in the *HMGCR* region as proxies for long-term HMGCR inhibition,[37-42] by also finding adverse relationships between HMGCR and T2D risk in EAS (having strong genetics evidence with drug-target MR estimates surpassing correction for multiple comparisons and also demonstrating evidence of colocalization) and SAS (weak evidence with a less precise drug-target MR estimate but evidence of colocalization). They also extend recent population genetics work in diverse populations that applied a clustering-based method to GWAS of T2D along with other cardiometabolic diseases and glycemic markers to develop genetic signatures underlying subtypes of T2D using GWAS data from EUR and non-EUR cohorts.[324] One of the clusters was enriched for variants involved in increased LDL-C levels, including SNPs in the *HMGCR* locus.[324] Further, the directions of the SNP associations were consistent with LDL-C lowering increasing T2D risk, which is in line with the directionality of our HMGCR findings, suggesting that our findings may be driven by a T2D subtype, which warrants future investigation. As we did not find corresponding evidence for adverse effects of HMGCR inhibition on glycemic traits in either EAS or SAS population (in

fact, HMGCR inhibition reduced fasting glucose in EAS, which result was robust in MVMR to correct bias for the BMI adjustment in the fasting glucose GWAS data, suggesting some potential glycemic benefits), it is possible that the adverse impact on T2D may be via potential pathways, such as weight gain associated with statin use, that have been previously reported.[340]

As the heterogenous HMGCR findings suggested adverse relationships in 3 of the 5 study populations, suggesting potential population specificity and biological mechanisms, it is possible that the observed differences in the estimates reflect differing allele frequencies of the HMGCR variants across the populations; however, the R^2 values for HMGCR instruments were generally comparable across populations, and the variants used for population-specific instruments were largely distinct, likely capturing the genetic architecture of the *HMGCR* locus specific to each population. Further, two-sample MR studies may be biased by population differences, which includes differences in allele frequencies, between the exposure and outcome data.[199] While we matched populations between the exposure and outcome pairs (i.e., EAS exposures with EAS outcomes), we cannot eliminate the possibility that there is remaining population stratification (i.e., phenotypic and genetic differences) present between the GLGC LDL-C, DIAMANTE T2D, and MAGIC glycemic trait cohorts that may influence the HMGCR-T2D and other findings in our study. Therefore, future studies are necessary with additional data sources to replicate and confirm the suggested population-level differences in lifelong HMGCR inhibition.

We underscore what previous studies have discussed regarding the comparative risks for T2D and cardiovascular benefits of statin therapy: that this modest adverse increase in T2D does not outweigh the substantial cardiovascular benefit of statin therapy.[39] While the drug-target MR study design does not enable investigation of the underlying mechanism, it has been suggested that the increase in T2D risk by statins may be explained, in part, by both impaired pancreatic insulin secretion because of a blockade by statins of L-type Ca^{2+} channels in pancreatic β cells and reduced insulin sensitivity in adipose tissue (mediated by downregulation of the glucose transporter type 4).[341] We also highlight that the genetic risk corresponding to long-term HMGCR inhibition—and the associated T2D risk—may not correspond to shorter periods of statin therapy. Notably, in exploratory analyses assessing the impact of HMGCR and PCSK9 inhibition on fasting insulin and insulin response to oral glucose tests, HMGCR inhibition was not linked with either fasting insulin or the IFC or Stumvoll ISI, both markers of postprandial insulin resistance,[206] suggesting that the mechanism linking HMGCR and T2D is not via increased insulin resistance. Conversely, these analyses suggested a potential beneficial role of PCSK9 inhibition on reduced insulin resistance (i.e., lower PCSK9 was linked with increased IFC). The analyses were consistent across several PCSK9 instruments—altered LDL-C levels via PCSK9 variants and lowered PCSK9 protein levels—and may motivate future follow-up study to determine whether PCSK9 inhibition does indeed stimulate insulin secretion, potentially mitigating the risk for T2D.

Finally, our polygenic LDL-C findings are generally in line with previous observational and genetics-based work finding that lower levels of circulating LDL-C are linked with higher T2D risk.[342] Interestingly, it has also been shown that among individuals with CVD, higher LDL-C levels are associated with reduced T2D, and families with hypercholesterolemia have demonstrated reduced T2D risk,[343,344] supporting a growing body of literature suggesting biological relationships between lipids and T2D, and providing important insight into

mechanisms underlying diabetogenesis.[343,344] One potential explanation for the observed discrepancy in how different loci affecting LDL-C influence T2D and glycemic control is that standard measurements of LDL-C do not distinguish between its various subtypes. LDL-C is a heterogeneous particle with different subtypes, each of which might have distinct biological effects on metabolic processes, including insulin sensitivity and glucose regulation.[345] Additionally, LDL-C undergoes changes in its composition and structure, such as altered electrophoretic mobility, increased triglyceride and ceramide content, prolonged retention of modified LDL-C in the blood, enhanced uptake by macrophages, and the formation of foam cells; the inability to differentiate these subtypes, or more fine-grained aspects of LDL-C dynamics, in routine clinical measurements could obscure the specific pathways through which certain loci, such as *PCSK9* and *HMGCR*, influence T2D risk.[346,347] By providing a detailed profile of lipid subtypes and their interactions with other metabolites, metabolomics can help identify the specific LDL- C subtypes that are most strongly associated with T2D risk.[346] For instance, metabolomics data could reveal whether certain LDL-C subtypes are more prevalent in individuals with T2D or whether they interact differently with glucose and insulin metabolism compared to other subtypes. Metabolomics can capture a broader spectrum of lipid-related metabolites, offering insights into the downstream effects of LDL-C subtypes on metabolic pathways involved in glycemic control.[348] This improved resolution could help to clarify why certain genetic variants associated with LDL-C have differing impacts on T2D.[347] For future research, exploring the complexities of metabolomics data could be invaluable in disentangling these relationships, leading to a more refined understanding of the metabolic consequences of LDL-C variability and potentially guiding more targeted therapeutic strategies.

3.5. Aim 1 Strengths and Limitations

Aim 1 has several important strengths. The primary strength of Aim 1 is the use of non-EUR data to perform parallel population-specific MR analyses in the most comprehensive genetics-based investigation of the risk for T2D associated with lipid-lowering drug-targets to date. A recent meta-analysis of 20,692 RCTs in the United States listed in ClinicalTrials.gov from the years 2000-2020 found that while there have been positive trends in RCT enrollment, there are still substantial racial/ethnic differences in minority recruitment, reporting, and representation.[349] Among the 20,692 RCTs (generating data from more than 4.76 million participants), Turner et al. found that European populations comprised almost 80% of all participants while only 10% of participants were AFR, 6% Hispanics/Latino, and 1% Asian (EAS and SAS combined).[349] This study and others like it may be an important step to help address the current imbalances in representation for minority race/ethnicity participants in RCTs,[349] and resulting health disparities[28,350] (at least with regards to lipid-lowering therapies, including statins, which are the most prescribed for noncommunicable diseases drug worldwide[351,352]). Another strength is our use of both proteomic and transcriptomic data to construct additional PCSK9 instruments in therapeutically and T2D-relevant tissues. For example, in mice, tissue-specific investigation of PCSK9 inhibition found that PCSK9 expression in the pancreas may be related to impaired β cell dysfunction while liver-derived PCSK9 was not related β cell dysfunction or insulin secretion.[59] However, our PCSK9 instruments derived from pancreatic tissue (in addition to the liver tissue PCSK9 instruments) failed to find evidence of pancreatic or hepatic PCSK9 expression on T2D or glycemic markers using human data, which should reassure clinicians and patients regarding the PCSK9-T2D relationship. Other strengths include leveraging complementary MR methods

incorporating genetic matrices of the underlying LD structure between drug target instrument variants to improve instrument precision, which is crucial for causal inference in MR.[48] Employing complementary MR methods, heterogeneity tests, and alternate instruments—and observing consistent MR estimates across them—further strengthens causal inference.[48,49,215]

There are important limitations for the analyses in Aim 1. First, the population specific GWAS for glycemic markers we analyzed in our study included adjustment for BMI in their models. It has been suggested that adjustment for covariates in GWAS models may bias the resultant GWAS.[219] Therefore, future multi-population MR studies should re-evaluate these results when sufficiently large unadjusted glycemic GWAS data becomes available. The Aim 1 project uses data derived from five populations. Participants included in the GWASs in these five populations hailed from a combined 34 countries and four continents, which, while an important step to improve our understanding of the relationships of lipid-lowering therapies and T2D risk in these populations, nevertheless is still not applicable to other populations not represented (e.g., although a large number of participants in this study are from the United States, interpretation of these results should not be generalized to Indigenous Americans, a population with an estimated 13.6% T2D prevalence[353]), or even geographically distinct members of one of the populations included (e.g., the HISP results should not be generalized to Hispanic/Latino populations located in South America). Further, given the nature of the GWAS data, we were not able to evaluate potential country-to-country heterogeneity within the respective populations, or even ethnolinguistic divisions within each population. Further still, we emphasize that causal inference requires triangulating study designs,[354] and while improving racial/ethnic diversity in genetics-based studies represents important advancements in our understanding of health and disease, long-term RCTs across diverse populations are required to further our understanding of the lipid-lowering therapeutic-T2D relationships. While we observed several heterogenous population-based estimates in the T2D and glycemic data suggesting potential population-specific relationships, we cannot eliminate the hypotheses that these differences may instead reflect differences in access to healthcare and glycemic control across the populations and the genetics data derived from them. For example, we observed heterogeneity between the PCSK9 and HMGCR impact on HbA1c in HISP with HMGCR inhibition (but not PCSK9 inhibition) linked with increased HbA1c levels. While the biological mechanisms are unclear, the HISP cohort may reflect challenges in healthcare access and T2D management reported among HISP populations, resulting in comparably poor glycemic control and worse clinical presentation of HISP T2D patients.[355] In the US, one analysis of ~66,000 electronic health records found that HISP patients who preferred speaking Spanish more poorly controlled HbA1c compared to HISP patients who preferred speaking English,[356] underscoring the need for programs designed to engage HISP and address organizational-level barriers to glycemic control for HISP patients[356] and the resulting biobanks. Other study limitations are inherent to the drug-target MR framework, e.g., the inability to evaluate potential off-target effects and pathways of PCSK9 and HMGCR inhibition, other than their intended lipid-lowering mechanisms.[357] Therefore, future non-EUR MRs will be necessary when the data becomes available in these populations. Also, while using *cis*-instruments in MR is less affected by possible violations of MR assumption than conventional polygenic MR,[6,50] it is not possible to completely rule out bias due to confounding or pleiotropy; however, colocalization and robust estimates across the multiple instrument sets improves confidence in the findings.

3.6. Aim 1 Future Directions

3.6.1 Expanding to more populations and cohorts

To advance our understanding of the effects of HMGCR inhibition on T2D risk, future genetic studies must prioritize inclusivity by expanding analyses to underrepresented populations. Current genetic research remains predominantly focused on European populations, which limits the generalizability of findings to other ancestries. This gap not only perpetuates health disparities but also hinders the identification of population-specific genetic effects and mechanisms. For example, individuals of African ancestry exhibit greater genetic diversity, including unique allele frequencies and loci associated with metabolic traits, yet remain understudied in lipid-related and T2D research.[30,358] Expanding datasets to include other cohorts of African, South Asian, Hispanic/Latino, and Indigenous populations is essential to uncovering genetic variants that may drive differential responses to HMGCR inhibition. Moreover, population-specific studies can illuminate the role of environmental and lifestyle factors that may interact with genetic predispositions to influence T2D risk.

Importantly, such expansions require concerted global efforts to establish and integrate data from diverse biobanks. Collaborative initiatives like the H3Africa Consortium and All of Us Research Program are critical starting points. These efforts could facilitate the inclusion of ancestries historically excluded from genetic research and provide insights into T2D prevalence and complications across different groups. Addressing these gaps is not only ethically imperative but also scientifically enriching, as it could lead to the discovery of novel mechanisms and therapeutic targets specific to these populations.[359] Ultimately, the findings from more inclusive studies will empower clinicians to tailor lipid-lowering therapies, such as statins or PCSK9 inhibitors, based on ancestry-specific risk profiles.

More generally, there exists a need to improve race/ancestry representation in genetics-based studies across all clinical disciplines.[28,30,360-364] As outlined by Fatumo et al.,[28] the imperative for increased genetic diversity in genomic studies is underscored by the prevailing imbalance, where the majority of data comes from individuals of European ancestry, leaving other populations underrepresented.[28] This European-centric bias not only raises ethical concerns but also results in missed scientific opportunities and health disparities.[28] For example, inadequate representation impedes the identification of population-specific variants and, in the application of MR and other genetics-based studies, potential ancestry-specific differences in the causal roles of important risk factors and biomarkers in disease risk. It also limits the accuracy of polygenic risk scores for diverse populations, and overlooks clinically important variants discovered exclusively in underrepresented groups.[28] Addressing the inequalities in genomic studies requires a concerted global effort to implement a roadmap for increased diversity.[28] These initiatives should leverage existing research infrastructure, capacity, expertise, and leadership within local institutions. Further, overcoming historical injustices, building trust, and considering ethical, legal, and social implications in study design are essential for engaging diverse populations in genomic research.[28,30,360-364] Ultimately, fostering genetic diversity is not only an ethical imperative but also crucial for advancing scientific understanding, reducing health disparities, and ensuring the applicability of genetic insights across a broad spectrum of populations.

3.6.2. Investigating T2D complications

Beyond assessing the primary relationship between HMGCR inhibition and T2D risk, future studies should evaluate its effects on T2D complications, including, diabetic nephropathy, retinopathy, and neuropathy. These downstream outcomes represent significant contributors to morbidity and mortality among individuals with T2D.[358] For example, while statins and other lipid-lowering therapies provide well-documented cardiovascular benefits, their potential to exacerbate or mitigate T2D complications remains underexplored. Understanding how HMGCR inhibition influences these secondary outcomes is critical for comprehensive risk-benefit assessments.

To achieve this, longitudinal studies linking genetic proxies of HMGCR inhibition to biomarkers and clinical endpoints of T2D complications are necessary. This approach could uncover whether increased T2D risk is offset by protective effects on microvascular or macrovascular outcomes. Additionally, stratified analyses based on T2D subtypes or glycemic control could reveal heterogeneity in these effects. For example, some evidence suggests that statin-induced increases in T2D risk may arise from mechanisms such as weight gain or impaired insulin secretion.[365] Future studies should explore whether these mechanisms differentially impact patients already at high risk of complications, such as those with preexisting kidney disease or poor glycemic control.

3.6.3. Two-step MR and mechanistic studies

To elucidate the biological mechanisms linking HMGCR inhibition to T2D, future research should leverage two-step MR using proteomic data. This approach, which this thesis uses in Aim 3 to elucidate proteomic mediators between BMI and cardiometabolic risk and has also successfully applied to identify protein mediators in other disease contexts (e.g., between BMI and COVID-19 risk[326] and also proteomic mediators of the atherosclerotic benefits of Interleukin 6 inhibition,[239] involves identifying proteins influenced by HMGCR variants (Step 1) and testing their causal relationship with T2D outcomes (Step 2). By integrating large-scale proteomic datasets such as those from deCODE and GTEx, two-step MR can help identify circulating proteins or tissue-specific markers that mediate HMGCR's effects on T2D risk. For example, proteins involved in insulin signaling, glucose metabolism, or inflammation could be highlighted as key intermediaries.

3.6.4. Translating mechanistic insights into clinical practice

While identifying mechanisms is crucial, translating these findings into clinical practice will require robust validation in diverse cohorts and real-world settings. Current evidence suggests that the cardiovascular benefits of HMGCR inhibitors often outweigh their modest impact on T2D risk.[366] However, stratified analyses are needed to identify patient subgroups that may be at heightened risk for adverse metabolic effects. For instance, individuals with obesity, prediabetes, or specific genetic predispositions may require closer monitoring or alternative therapies.

Ultimately, integrating mechanistic insights with clinical data could lead to the development of targeted interventions. Proteins identified as mediators in two-step MR studies could serve as biomarkers for T2D risk stratification or therapeutic targets. For example, interventions designed to modulate these proteins—either through pharmacological agents or lifestyle changes —could mitigate the metabolic side effects of HMGCR inhibitors. Multivariable MR analyses could further clarify how modifiable factors such as BMI, diet, and physical activity interact with these pathways, providing actionable strategies for reducing T2D risk while preserving cardiovascular benefits.

3.7. Aim 1 Conclusions

In conclusion, the Aim 1 project did not find an adverse impact of genetically proxied lowering of LDL-C levels by PCSK9 variants on T2D or markers of glycemia in EAS, SAS, HISP, or EUR cohorts. We do find an adverse relationship between LDL-C lowering via PCSK9 inhibition and T2D in AFR; however, because the increased risk of PCSK9 inhibition on T2D in AFR was not robust in sensitivity analyses and does not align with recent results using data from African Americans in the US-based MVP cohort, we emphasize the need for additional investigation into this potential relationship. For HMGCR, we replicate the previously observed slight increase in T2D risk and emphasize that this increase likely does not offset the substantial cardiovascular benefit of statin therapy. While replication with data from more countries, additional populations, and additional age groups is necessary, these findings should help inform clinicians and patients considering lipid-lowering therapies who may be concerned about the possibility for increased T2D risk.

CHAPTER 4: AIM 2 RESULTS

Chapter Overview

This chapter explores the genetic and biological underpinnings of problematic alcohol use (PAU) and alcohol consumption behaviors, utilizing cis-instrument Mendelian Randomization (MR) and complementary analyses across cortical proteomic and cell-type transcriptomic data. Screening identified 217 cortical proteins and 255 cell-type genes, including 36 novel proteins and 37 novel genes, with robust instrument validity and minimal weak instrument bias. Key findings include strong associations of MAPT and MTOR with multiple alcohol-related outcomes, emphasizing their roles in synaptic function and neurogenesis. Colocalization analyses highlighted high-confidence targets such as SAMHD1, VIPAS39, and NRBP1, with replication in the FinnGen cohort confirming these associations. Gene-set enrichment analyses linked findings to pathways in alcohol metabolism, neural signaling, and blood-brain barrier integrity. Druggable targets like SLC4A8 and CAB39L suggest therapeutic potential, while drug repurposing identified prazosin and memantine as candidates. Neuropsychiatric profiling revealed shared pathways with disorders like schizophrenia and Alzheimer's, underscoring the broader clinical relevance. These findings advance understanding of alcohol-related behaviors and provide a foundation for targeted therapeutic strategies.

4.1. Cis-Instrument MR Screens

Variants for the cortical proteins had an average F-statistic of 42.0 (range: 12.12-825.70) (**Table AP3.2**) and variants comprising the single cell transcriptomic instruments explained on average ~9.0% of the variance in the instrumented gene expression in that cell type, and the average F-statistic for individual variants was 20.21 (range: 12.12-238.60) (**Tables AP3.3-AP3.10**). The strong F-statistics for the cortical and cell-type instruments suggest minimal evidence that the MR estimates are subject to weak instrument bias.[242] In the proteome, 2269/3562 cortical proteins were instrumented by 2+ SNPs; and in the cell-type transcriptome, as follows: AST, 782/2406 genes; EXC 1936/4606 genes INH, 763/2493 genes; OPC, 610/1905 genes; PER, 97/617 genes; END, 245/1199 genes; MIC, 491/1479 genes; OLI, 1266/3174 genes.

In the cortical proteome, we found 293 total proteins surpassing correction for multiple comparisons (219 individual proteins) whose cis-regulated protein levels were associated with the alcohol-related outcomes surpassing correction for multiple comparisons (**Figure 4.1a; Figures AP3.1-AP3.4, Tables AP3.11-AP3.14**). 48 proteins associated with PAU, 139 with AIF, 79 with DPW, and 27 with binge drinking. 57 proteins were pleiotropic (i.e., linked with more one alcohol outcome) (**Figure AP3.5**), and estimates for these proteins were generally directionally consistent across the outcomes, e.g., MAPT was associated with increased AIF (β =0.180, CI=[0.127,0.233], P-value=2.92×10^{-11}), increased DPW (β=0.149, CI=[0.116,0.183], P-value=1.94×10^{-17}) and increased binge drinking (β=0.278, CI=[0.193,0.363], P-value=1.49×10^{-10}); MTOR was associated with increased AIF (β=0.064, CI=[0.0544,0.083], P-value=8.11×10^{-11}) and DPW (β=0.031, CI=[0.018,0.043], P-value=9.52×10^{-7}) (**Table AP3.15**). The MTOR

finding aligns with preclinical work suggesting brain-specific mTOR pathway inhibition reduces alcohol intake in mice.[367] Results for cortical proteins surpassing correction for multiple comparisons generally aligned across complementary MR methods used to test the MR assumptions, and the MR Egger intercept did not demonstrate evidence for pleiotropy.

We identified 486 cell-type-gene combinations, representing 255 unique genes (unique used here to indicate the gene count has been adjusted for any duplicates in the results across the different alcohol consumption behaviors and cell-types) across eight cell types—230 unique to cell types and 25 overlapping with cortical protein findings (**Figure 4.1a-b, Figures AP3.6-AP3.9, Tables AP3.16-AP3.19**). MR estimates for 68 genes surpassed correction for multiple comparisons in multiple cell types (**Table AP3.20**). For instance, MR estimates for RBM6 in five cell types were consistent with increased AIF, while FAM118A expression in all cell types was linked with reduced AIF. These findings highlight both behavior-specific and overlapping gene associations across alcohol-related behaviors (**Figure AP3.10**). MR results aligned across complementary methods, with no evidence of pleiotropy.

We used a "multiverse"[200,247] sensitivity analysis assessing the stability of the cis-MR estimates over a range of instrument selection parameters[200,247] for each of the proteins and cell-type genes for which the primary cis-MR analyses suggested underlying causal effects on the alcohol-related outcomes. We found the cis-MR estimates directionally concordant for 217 of the 219 cortical proteins and all the cell-type genes and took these targets forward to the downstream analyses (**Tables AP3.21-AP3.28**). AIF was linked with most proteins and genes and binge drinking with the fewest (**Figure 4.1b**). The proteomic and cell-type transcriptomic results were largely distinct, with only 25 genes overlapping across all alcohol-related phenotypes (**Table 4.1**, **Figure 4.1c**), highlighting differences in the genetic underpinnings of gene expression and protein levels in the brain.[368] Combining all cortical proteins and genes in the cell-type transcriptome showed that only 13 genes (2.9% of top proteins/genes) were shared across the 4 alcohol-related outcomes (**Figure 4.1d**), underscoring that PAU and alcohol consumption behaviors have distinct proteomic and transcriptomic underpinnings, which aligns with previous GWAS analyses.[80,94]

Figure 4.1. Manhattan plots for cis-MR screens of cortical proteomic and single-cell transcriptomic architecture in PAU and alcohol consumption behaviors. (a) *shows the cortical proteomic and transcriptomic MR results for each behavior, with highlighted values surpassing the adjusted P-value threshold. All cis-instrument MR estimate P-values are based upon used two-sided Wald tests. P-values were not adjusted for multiple comparisons. In the cell-type analyses, some genes were associated with multiple behaviors, but the flow chart focuses on unique genes (adjusted for duplicates).* ***(b)*** *summarizes the cortical proteomic and cell-type*

transcriptomic findings for PAU and alcohol behaviors, including proteins and genes specific to each behavior or associated with multiple outcomes. ***(c)*** *presents a Venn diagram of the 217 unique cortical proteins and 255 cell-type genes across the four alcohol-related outcomes identified in the cis-MR screen and sensitivity analyses.* ***(d)*** *shows another Venn diagram outlining the overlap of unique cortical proteins and cell-type genes for each outcome.* ***(e)*** *displays the percentage of proteins/genes identified by the cis-MR screen that were captured by other methods, with percentages above the bar plots indicating how many cortical proteins and cell-type genes were identified by each method. Proteins and genes not captured by other methods were considered novel for alcohol use behaviors. PAU: problematic alcohol use; AIF: alcohol intake frequency; DPW: drinks per week; AUD: alcohol use disorder; EWAS: epigenome-wide association study; MAGMA: Multivariate Analysis of Genomic Annotation; H-MAGMA: Hi-C Multivariate Analysis of Genomic Annotation. Reproduced from Rosoff DB, Wagner J, Bell AS, Mavromatis LA, Jung J & Lohoff FW. "A multi-omics Mendelian randomization study identifies new therapeutic targets for alcohol use disorder and problem drinking." Nature Human Behaviour 2025; 9: 188–207.*

Table 4.1. Cis-instrument MR estimates for 25 overlapping targets surpassing correction for multiple comparisons in the cortical proteome and cell-type transcriptome

			Cortical protein estimates		Cell-type transcriptome estimates	
Gene	**Gene position**	**Druggability Tier**	**Alcohol-related outcomes**	**Proteome β [95% CI]**	**Alcohol-related outcomes**	**Cell-type transcriptome β [95% CI]**
ACTR1B	2: 98272431-98280570	--	DPW	-0.024, [-0.034, -0.014]	AIF (OLI), DPW (OLI, EXC)	0.051, [0.034, 0.068]; 0.032, [0.022, 0.043]; 0.026, [0.016, 0.037]
ARFGAP3	22: 43192508-43254112	--	AIF	-0.067, [-0.096, -0.038]	AIF (END, OLI, EXC)	-0.023, [-0.032, -0.013]; 0.048, [0.031, 0.066]; 0.055, [0.034, 0.075]
C18orf8	18: 21083473-21111746	--	AIF	0.079, [0.06, 0.097]	AIF (END, INH, EXC)	0.074, [0.058, 0.089]; 0.039, [0.023, 0.055]; 0.113, [0.088, 0.137]
CAB39L	13: 49882786-50018262	--	PAU, Binge, AIF, DPW	0.012, [0.008, 0.017]; 0.018, [0.011, 0.025]; 0.022, [0.016, 0.027]; 0.013, [0.009, 0.016]	AIF (AST)	0.038, [0.027, 0.049]
CCDC25	8: 27590835-27630170	--	AIF, DPW	-0.028, [-0.035, -0.021]; -0.016, [-0.022, - 0.01]	AIF (INH)	-0.039, [-0.055, -0.023]
DDHD2	8: 38082736-38133076	--	DPW	-0.066, [-0.093, -0.039]	DPW (INH, EXC)	-0.043, [-0.06, -0.025]; -0.0343, [-0.048, -0.02]
EML6	2: 54950636-55199157	--	DPW	-0.036, [-0.048, -0.023]	DPW (EXC)	-0.032, [-0.045, -0.02]
GMPPB	3: 49754277-49761384	--	PAU	0.014, [0.009,0.019]	AIF (INH)	0.075, [0.053, 0.096]
HIBADH	7: 27565061-27702614	--	DPW	-0.013, [-0.017, -0.008]	DPW (OLI)	-0.009, [-0.012, -0.005]
LARS	5: 145492601-145562223	Tier 2	AIF	-0.045, [-0.062, -0.027]	AIF (OPC)	0.036, [0.022,0.05]
MAPT	17: 43971748-44105700	Tier 1	AIF, Binge, DPW	0.18, [0.127, 0.233]; 0.278, [0.193, 0.363]; 0.149, [0.115, 0.183]	AIF (AST), Binge (AST), DPW (AST)	-0.046, [-0.057, -0.036]; -0.059, [-0.083, -0.034]; -0.031, [-0.038, -0.024]
NPC1	18: 21086148-21166862	Tier 1	Binge	0.026, [0.015, 0.038]	PAU (EXC)	0.037, [0.023, 0.051]
NT5C1A	1: 40124793-40137710	--	DPW	0.021, [0.016, 0.025]	DPW (EXC)	0.021, [0.013, 0.03]
PACSIN2	22: 43231418-43411151	--	Binge	0.026, [0.015, 0.038]	PAU (EXC)	0.037, [0.023, 0.051]
RAB3C	5: 57878048-58155213	--	DPW	0.021, [0.016, 0.025]	DPW (EXC)	0.021, [0.013, 0.03]
RABGAP1L	1: 174128548-174964445	--	PAU	-0.059, [-0.082, -0.036]	PAU (EXC), AIF (EXC), Binge (EXC), DPW (EXC)	-0.046, [-0.055, -0.038]; -0.108, [-0.116, -0.1]; -0.108, [-0.121, -0.095]; -0.067, [-0.072, -0.062]
SH2B1	16: 28857921-28885526	--	AIF, DPW	0.233, [0.188,0.278]; 0.097, [0.069,0.126]	AIF (INH, EXC), DPW (INH, EXC)	0.114, [0.093, 0.135]; 0.125, [0.102, 0.148]; 0.045, [0.031, 0.058]; 0.0491, [0.034, 0.064]
SHMT1	17: 18231187-18266856	--	AIF, DPW	0.021, [0.015,0.027]; 0.011, [0.008,0.015]	AIF (AST)	0.046, [0.027,0.065]
SIPA1L1	14: 71787166-72207946	--	DPW	0.031, [0.02,0.042]	AIF (OLI), Binge (OLI), DPW (OLI)	-0.008, [-0.011, -0.005]; -0.005, [-0.006, -0.003]; -0.017, [-0.021, -0.013]
SLC5A6	2: 27422455-27435826	Tier 1	PAU, AIF, DPW	-0.022, [-0.031, -0.012]; -0.052, [-0.061, -0.042]; -0.026, [-0.032, -0.02]	AIF (EXC)	-0.085, [-0.115, -0.055]
SNTB2	16: 69221032-69342955	-	AIF, DPW	-0.077, [-0.111, -0.044]; -0.138, [-0.19, -0.086]	DPW (EXC)	0.045, [0.026, 0.063]
STAT6	12: 57489191-57525922	Tier 2	AIF	0.044, [0.029,0.058]	AIF (EXC)	0.073, [0.048, 0.097]
SULT1A2	16: 28603264-28608430	Tier 1	AIF, DPW	0.049, [0.04, 0.058]; 0.019, [0.013, 0.025]	AIF (END, INH, EXC, OPC); DPW (END, AST)	-0.072, [-0.083, -0.061]; -0.035, [-0.046, -0.024]; -0.038, [-0.049, -0.027]; -0.053, [-0.072, -0.034]; -0.027, [-0.034, -0.02]; -0.0358, [-0.047, -0.025]
TMEM245	9: 111777432-111882225	--	DPW	0.008, [0.005, 0.012]	AIF (OLI)	0.055, [0.04,0.07]

UGDH	4: 39500375-39529931	--	AIF, DPW	-0.08, [-0.097, -0.062]; -0.037, [-0.048, -0.026]	AIF (OPC), DPW (OPC)	0.052, [0.04, 0.065]; 0.045, [0.037, 0.053]

Notes: *The 25 targets included in the table had cis-instrument MR estimates surpassing Bonferroni-corrected P-value thresholds of 1.41×10^{-5} and 2.80×10^{-6} for the cortical proteome and cell-type transcriptome screens, respectively. All cis-instrument MR estimate P-values are based upon used two-sided Wald tests. P-values were not adjusted for multiple comparisons. Boldfaced indicates target that has directionally inconsistent estimates between cortical proteome and cell-type transcriptome. The columns labeled "Alcohol-related outcomes" are the alcohol-related GWASs with which the target was associated in the cis-instrument MR screen (either PAU, AIF, Binge, or DPW). The corresponding columns with the Mendelian randomization (MR) estimates (β's and corresponding 95% confidence intervals [CIs]) for that alcohol-related outcome. For cells with more than one alcohol-related outcome and related MR estimates, the MR estimates are presented in the same order as corresponding alcohol-related outcome. Druggability tiers are defined by Finan et al.[126] Tier 1 genes include targets of approved small molecules and drugs in clinical trials and Tier 2 genes consist of genes with documented bioactivity for drug-like small molecules. PAU: problematic alcohol use; AIF: alcohol intake frequency; DPW: drinks per week; AST: astrocytes; EXC: excitatory neurons; INH: inhibitory neurons; OLI: oligodendrocytes; OPC: oligodendrocyte precursor cells; PER: pericytes; END: endothelial cells.*

4.2. Biological Characterization of Identified Genes

Gene-set enrichment (GSEA) implicated a range of biological processes and pathways, including alcohol catabolism, neurogenesis/neurodegeneration, and other neural processes related to synaptic functioning (e.g., exocytic insertion of neurotransmitter receptor to the postsynaptic membrane (overlapping gene: SNAP47), neuronal dense cores vesicles (gene: STXBP5L), and regulation of postsynaptic neurotransmitter receptors) (**Tables AP3.29**, **AP3.30**). As expected, GWAS Catalog look-up found overlap with alcohol-related phenotypes and other substance use behaviors, including cannabis use (genes: *CADM2* and *SMG6*) and tobacco smoking (genes: *CADM2*, *SMG6*, and *MAPT*), consistent with the shared genetic architecture of substance use behaviors.[369] Additional biological characterization using human brain single-cell datasets from Li et al.[254] demonstrated widespread differential expression among 28 cortical cell-types, with generally increased expression in excitatory neurons and inhibitory neurons across several cortical layers (i.e., L2, L3, L5, and L6) and reduced expression in other cell types (**Figure AP3.11, Tables AP3.31, AP3.32**).

We conducted several drug-gene analyses to prioritize repurposing opportunities among approved therapeutics and candidate compound (**Tables AP3.33-AP3.36**). Because DRD2 was among the top genes in EXC for PAU (β=-0.047, CI=[-0.066,-0.028], P-value=1.36×10^{-6}), antipsychotics and Anti-Parkinson's drugs were represented in the cross-reference of the cMAP[256] and DGIdb databases[257] (**Table AP3.34**); however, the direction of effect for PAU indicated by our cis-MR screen—an increased DRD2 expression in EXC—suggested the dopaminergic receptor agonists comprising the Anti-Parkinson's drugs would be beneficial while the dopaminergic receptor antagonists would not, aligning with previous work.[370,371] Other drugs with directionally relevant effects on the targets were the antidiabetics glyburide (DGIdb interaction score=0.53 with ABCC5), voglibose (DGIdb interaction score=5.36 with MGAM), and acarabose (DGIdb interaction score=3.57 with MGAM) (**Table AP3.34**). 57 of the cortical proteins and 42 of the cell-type genes were considered druggable by small molecules (**Table AP3.35**), and signature matching found potential therapeutic candidates with high negative connectivity scores, including prazosin and memantine, for reversing the proteomic and transcriptomic signatures of PAU and alcohol consumption (**Table AP3.36**). Weakly positive correlations between the connectivity scores for PAU and alcohol consumption suggested the bulk proteomic and cell-type transcriptomic signatures may reflect distinct repurposing opportunities (**Figures AP3.12, AP3.13**).

4.2.1. Cis-MR screens identify novel targets for alcohol traits

We next investigated whether the proteins and genes identified with the cis-instrument MR screens represented novel alcohol-related signals, or are captured by the alcohol outcome GWAS signatures, tagged by other gene-prioritization methods ("Multi-marker Analysis of GenoMic Annotation" [MAGMA],[281] H-MAGMA,[258] and FUSION transcriptomic imputation[262]), or have been implicated by complementary multi-omics approaches elucidating the biological

underpinnings of AUD or previous studies integrating GWAS data for alcohol-related outcomes transcriptomic and proteomic data to identify alcohol-related genes.[95,228,230,372,373] 36 of the 217 cortical proteins and 37 of the 255 cell-type genes were classified as novel (**Figure 4.1e**, **Tables AP3.37-AP3.39**), suggesting that the cis-MR screens captured additional biological underpinnings of alcohol use behaviors not previously captured by the complementary approaches. Novel proteins and genes were located throughout the genome. Of the 4 alcohol-related outcomes, AIF had the most novel genes; however, there were novel targets for each alcohol-related GWAS. Several of the proteins and genes were associated with multiple alcohol-related outcomes, and while not novel across all alcohol-related outcomes, were, in fact, novel for one or more of the alcohol-related GWASs. For example, CAB39L protein abundance, which surpassed correction for multiple comparisons for the 4 alcohol-related outcomes, were not novel for AIF or DPW, but were novel for PAU and binge drinking. Many of the top proteins and genes (155 of 217 cortical proteins and 173 of 255 cell-type genes) were located within the loci of the epigenetic signatures of AUD identified by Lohoff et al.,[105] indicating a strong overlap between the transcriptomic, proteomic, and epigenetic signatures of AUDs. We found less overlap among differentially expressed genes (DEGs) in postmortem brain tissue from AUD patients, suggesting cis-MR and postmortem DEG may be capturing distinct aspects of AUDs.

4.2.1a. PWAS/TWAS literature comparison. In addition, using previous proteome wide association study (PWAS) and transcriptome wide association study (TWAS) literature integrating omics data with GWASs of alcohol-related outcomes for gene prioritization to classify the novelty of our cis-instrument MR findings, we also contextualize our results with existing PWAS/TWAS literature for other neuropsychiatric outcomes. 6 novel proteins and 9 novel genes have been linked with other neuropsychiatric outcomes (depression, neuroticism, post-traumatic stress disorder, bipolar disorder, and schizophrenia) (**Table AP3.40**), underscoring many of the targets, including VIPAS39, as novel findings.

4.2.1b. Assessing directionality in bulk brain tissue. Among targets with FUSION[103] transcriptomic imputation estimates surpassing correction for multiple comparisons, imputation estimates were generally consistent across 12 brain tissues and directionally consistent across the bulk brain tissue and the proteomic or cell-type MR results, suggesting generally conserved associations throughout the brain (**Figures AP3.14-AP3.17**, **Tables AP3.41, AP3.42**). We found fewer bulk transcriptomic relationships for the cortical proteins than among the cell-type genes, supporting previous work showing differences in the genetic underpinnings of gene expression and circulating proteins in the brain.[368]

4.2.2. Colocalization prioritizes high confidence targets

Among cortical proteins, across the 4 outcomes, there was evidence of colocalization (PP.H4 >0.7) for 21.5% of the cortical proteins (47 of 217 unique proteins). 35 AIF proteins colocalized, including the novel protein VIPAS39, 17 DPW proteins, 1 for binge drinking, and 5 for PAU (**Figure 4.2a**, **Table AP3.43**). MAPT colocalized with AIF, DPW, and binge drinking, and 7 proteins colocalized for both AIF and DPW (**Figure 4.2a**)**.** CADM2 colocalized with PAU and had a PP.H4 almost surpassing the designated threshold for DPW (PP.H4=0.65). Only 2 proteins

colocalized for both PAU and drinking behaviors (SLC4A8 colocalized with PAU and DPW; SLC5A6, with PAU, DPW, and AIF).

In the cell-type transcriptome, there was evidence of colocalization for 22.7% or 58 of the 255 unique genes (**Table AP3.44**). As with the cortical proteins and likely because of the comparably large sample size and large number of independent genomic loci surpassing conventional genome-wide statistical significance (P-value $< 5\times10^{-8}$), AIF had the most cell-type genes (47 total genes), including the novel protein coding gene, PACRG, and the RNA gene ENSG00000259420, surpassing correction for multiple comparisons and showing evidence of colocalization. DPW had 17, binge drinking had 6, and PAU had 7. Among genes that surpassed correction for multiple comparisons in more than one cell type, most colocalized in multiple cell types (e.g., RBM6 for AIF in astrocytes, excitatory neurons, oligodendrocytes, OPCs, and microglia; DDHD2 for DPW in excitatory and inhibitory neurons; and MTCH2 for PAU in oligodendrocytes and OPCs, etc.). Conversely, several other genes, e.g., TNRC6A, only colocalized with AIF in OPCs and not inhibitory neurons, suggesting cell-type specificity for these genes.

Figure 4.2. Colocalization results and number of overlapping neuroimaging traits between the cortical proteins and cell-type genes. *Panel* ***(a)*** *shows scatter plots of the colocalization results for all proteins/genes surpassing correction for multiple comparisons. The red dashed lines indicate the PP.H4 cutoff of 0.7 used to define evidence of shared causal variant between the protein or gene and respective alcohol-related outcome. The 7 labeled results are those with*

*evidence of colocalization in both the proteomic and transcriptomic quantitative trait loci (QTL) data sources. Panel **(b)** presents the counts of the results of the neuroimaging cis-instrument MR analysis across 2 imaging modalities (grey matter structure and white-matter microstructure [diffusor tensor imaging – DTI]) for the colocalized proteins/genes that did not demonstrate evidence of colocalization in both QTL sources. To be included in the analyses, the cis-MR estimate for the protein/gene needed to pass the same P-value threshold adjusted for multiple comparisons as in the original cis-MR screens. AIF: alcohol intake frequency. AST: astrocytes; EXC: excitatory neurons; INH: inhibitory neurons; OLI: oligodendrocytes; OPC: oligodendrocyte precursor cells; PER: pericytes; END: endothelial cells. Reproduced from Rosoff DB, Wagner J, Bell AS, Mavromatis LA, Jung J & Lohoff FW. "A multi-omics Mendelian randomization study identifies new therapeutic targets for alcohol use disorder and problem drinking." Nature Human Behaviour 2025; 9: 188–207. © 2025 The Authors, licensed under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).*

22.4% (13 of the 58) genes colocalized with more than one alcohol-related behavior. Overlapping colocalization was generally conserved across cell-types. For example, FUT2 in excitatory neurons colocalized with all 4 outcomes (and further with PAU, AIF, and DPW in astrocytes); SYT14 in excitatory neurons colocalized with PAU, AIF, and binge drinking; NTN5 in astrocytes colocalized with both PAU and DPW; and PLEKHM1 expression in oligodendrocytes colocalized with both AIF and binge drinking. Further, in line with the overlap findings from the initial cis-MR screen, there was minimal overlap across the cortical proteins and cell-type genes (**Table AP3.45**): only 7 genes colocalized at both the proteomic and cell-type levels with their respective alcohol-related outcome (DDHD2, CAB39L, LARS, SNBTB2, SHMT1, STAT6, and ARFGAP3).

4.2.2a. Distinct genes converge on brain structure and connectivity.

Given the minimal overlap of the proteomic and cell-type transcriptomic mediators of the genetic predisposition for PAU and alcohol consumption behaviors, we next sought to evaluate whether there was evidence as to whether the non-overlapping features demonstrated convergent relationships with magnetic resonance imaging (MRI) data derived from recent GWASs of gray matter structures[231,234,235] and white matter connectivity[85] (**Tables AP3.46-AP3.53**). Exploratory MR analyses of the cortical proteins and cell-type genes on brain MRI data indicated that there were convergent gray matter structures and white matter tracts on which the non-overlapping genes impacted (**Figure 4.2b**), suggesting shared neurophysiological pathways of proteomic and transcriptomic mediators of the genetic predisposition for PAU and alcohol consumption behaviors despite their minimal overlap of individual genes. For instance, cortical proteins and cell-type genes each demonstrated relationships with overall cortical thickness and cortical surface area, both recently shown in MR analyses to impact the predisposition to alcohol consumption behaviors.[85]

4.3. Replication and Neuropsychiatric Contextualization

We next aimed to replicate and prioritize the colocalized cortical proteins and cell-type genes using independent electronic health record-based (EHR) data for assessing the psychiatric and physical consequences of problematic alcohol consumption among FinnGen[269] cohort participants (**Figure AP3.18**). 35.5% of cortical proteins (16 of 45) (**Table 4.2**) and 18.9% of cell-type genes (11 of 58) (**Figure 4.3a-b, Table 4.3**) replicated with at least one FinnGen alcohol-related outcome, including the novel protein VIPAS39 (**Tables AP3.54-AP3.57**). We saw strong replication with AUD diagnosis among FinnGen participants, providing additional evidence that these prioritized, high-confidence cortical proteins and cell-type genes are involved in mediating the genetic predisposition of AUD and PAUs. As expected, cis-MR estimates for the FinnGen consequences of alcohol consumption were generally directionally consistent with the cis-MR observed MR estimates of the alcohol-related outcomes from the initial screen (i.e., if the therapeutically indicated direction for cortical protein or cell-type gene with reducing PAU or alcohol consumption behaviors was inhibition of the target, then it was also inhibition for the FinnGen replication) (**Tables AP3.55, AP3.57**).

For instance, RHOA showed consistent directions of association across 3 different FinnGen outcomes, with effect sizes (odds ratios [ORs]) ranging from 0.58–0.63 [0.458,0.772] and associated P-values $< 9.34\times10^{-6}$, indicating a robust association with alcohol-related outcomes. Similarly, the protein SAMHD1 exhibited activation with ORs of 0.82–0.88 [0.738,0.932] across 2 FinnGen outcomes (**Figure 4.3a**). Conversely, the cortical proteins GBA2, CAB39L, VIPAS39, and ULK3 had heterogenous estimates across their respective outcomes, while among cell-type genes, only *AGRN* had heterogenous estimates (**Figure 4.3b**).

Figure 4.3. Results of colocalized, high-confidence proteins and genes on the physical and psychiatric consequences of alcohol consumption in the FinnGen cohort. *Forest plots of the primary Mendelian randomization (MR) estimates. The center of the error bars are the odds ratios (ORs) (IVW or Wald ratio depending on the number of instrument variants and the error bars are the lower and upper 95% confidence intervals for results surpassing correction for multiple comparisons. All cis-instrument MR estimate P-values are based upon used two-sided Wald tests. P-values were not adjusted for multiple comparisons.* ***(a)*** *presents the cortical proteins, and* ***(b)*** *presents the cell-type gene results. Results are aligned with increased protein levels or gene expression, i.e., a positive OR indicates that increased protein levels/cell-type expression was associated with increased risk of the hypothetical outcome. Data used for the exposures were the cortical proteins (N=722) and single cell gene expression (N=192) and the outcomes were derived from the 9th release of the FinnGen cohort (N=377,277). The AUD Swedish definition includes electronic health record codings related to a range of physical and psychiatric consequences of alcohol use behavior while the ICD-10 AUD definition included electronic health record hospital discharges and causes of death related to mental and behavioral disorders due to use of alcohol. The "*" indicates the genes with discordant MR estimate directions between the alcohol-related outcome in the initial cis-instrument MR screen and the replication outcome. All other results were directionally consistent, i.e., the direction of the cis-MR estimates aligned for the target in both the primary and replication analyses, between the initial and replication analyses for all outcomes that surpassed correction for multiple comparisons. AUD: alcohol use disorder, ICD: International Classification of Disease; AST: astrocytes; EXC: excitatory neurons; INH: inhibitory neurons; MIC: microglia; OLI: oligodendrocytes; OPC: oligodendrocyte precursor cells; PER: pericytes; END; ICD-10; International Classification of Disease, Tenth Revision. Reproduced from Rosoff DB, Wagner J, Bell AS, Mavromatis LA, Jung J & Lohoff FW. "A multi-omics Mendelian randomization study identifies new therapeutic targets for alcohol use disorder and problem drinking." Nature Human Behaviour 2025; 9: 188–207. © 2025 The Authors, licensed under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).*

4.3.1. Characterizing relationships with neuropsychiatric outcomes

Next, we phenotypically characterized the colocalized proteins and genes, assessing their impact on 65 psychiatric, neurologic, and behavior outcomes. As expected, given strong clinical and genetic relationships of alcohol use behaviors and neuropsychiatric comorbidities,[80,94] we observed strong associations with the 65 outcomes (**Figure 4.4**, **Tables 4.2, 4.3, AP3.58, AP3.59**). For example, cortical MAPT proteins levels were linked with 31 outcomes, aligning with its noted role in neurodegenerative diseases and neuropsychiatric disorders.[374] In addition to the highly pleiotropic neuropsychiatric impact of MAPT, the directions of the cis-MR estimates for MAPT on these outcomes were directionally opposite to that of the direction for PAU and alcohol consumption behaviors, i.e., while the cis-MR suggests that increased MAPT levels would increase PAU, it would also reduce the risk for neuropsychiatric outcomes, including feelings of depressed mood (β=-0.050, CI=[-0.073,-0.028], P-value=1.46×10^{-5}), and mood swings (β=-0.086, CI=[-0.104,-0.067], P-value=3.79×10^{-20}), suggesting that the therapeutically relevant direction for PAU (lowered MAPT levels) would increase these outcomes, providing an

example of a target where the complexities of the broader neuropsychiatric profile may make it a challenging target for pharmacological modulation. By contrast, for C1QTNF4 protein levels linked with 21 neuropsychiatric outcomes, the direction of the relationships indicated by cis-MR generally aligned with the direction for PAU, i.e., increased C1QTNF4 would increase DPW, PAU, and neuropsychiatric outcomes like self-reported feeling tense (OR=1.043, CI=[1.025, 1.062], P-value=2.74×10^{-6}), feeling miserable (OR=1.097, CI=[1.063,1.132], P-value=1.22×10^{-8}), Alzheimer's disease (OR=1.80, CI=[1.48,2.20], P-value=4.98×10^{-9}), and tobacco smoking (β =0.030, CI=[0.014,0.045], P-value=1.95×10^{-4}), suggesting that lowered C1QTN4 would be the therapeutically relevant direction for PAU and these other neuropsychiatric outcomes (**Table AP3.58**).

Other cortical proteins with generally aligned and favorable neuropsychiatric profiles included the novel protein VIPAS39, which had an impact on 5 neuropsychiatric outcomes that all aligned with the hypothetical therapeutic direction (increased VIPAS39 levels) on AUD risk (FinnGen AUD OR=0.91, CI=[0.862, 0.957], P-value=9.2×10^{-5}), and also SAMHD1 (cis-MR analyses suggested that higher levels were associated with reduced AIF and AUD risk), which was linked with increased cognitive performance (β=0.051, CI=[1.027, 1.082], P-value=8.43×10^{-5}) and reduced risk of attention deficit/hyperactivity disorder (OR=0.81, CI=[0.714, 0.908], P-values=3.97×10^{-4}) (**Tables 4.2, AP3.58**), suggesting that potential pleiotropic relationships from targeting these cortical proteins would be generally favorable. Finally, the two solute carrier transporters, SLC4A8 and SLC5A6, also demonstrated directionally consistent neuropsychiatric profiles, with the indicated clinical direction of pharmacological modulation that would be required to have efficacy for AUD, aligning with recent work highlighting the potential for SLCs for AUD.[375]

We observed similar patterns of pleiotropic associations among many of the cell-type genes, frequently extending previous transcriptomic-neuropsychiatric associations to single cell resolution, e.g., NRBP1 expression in INH was associated with lower Parkinson's disease (PD) risk (OR=0.75, CI=[0.649,0.859], P-value=4.62×10^{-5}), supporting recent bulk transcriptomic imputation work[376] (**Figure 4.4, Tables 4.3, AP3.59**). For NRBP1, the PD finding was the only neuropsychiatric relationship that surpassed correction for multiple comparisons. Further, it was directionally consistent with the cis-MR estimates for PAU, DPW, and binge drinking, providing genetics support that therapeutically increasing NRBP1 expression in INH may be beneficial for reducing alcohol consumption with a favorable neuropsychiatric side-effect profile. CAB39L expression in AST, INO80E expression in EXC, and NUP160 expression in MIC had similarly favorable neuropsychiatric profiles (INO80E expression in EXC reduced schizophrenia risk, and NUP160 reduced self-reported experiences of mood swings, feeling guilty, feeling miserable, and increased sleep duration) when aligning their cis-MR estimates to the direction required for reducing DPW (lower expression). By contrast, ARL17B expression in AST, INH, and MIC demonstrated highly pleiotropic relationships, including many estimates suggesting that the required direction for therapeutic targeting to reduce PAU and alcohol consumption behaviors (increased ARL17B expression) would have an adverse neuropsychiatric impact, which would diminish its potential suitability as a therapeutic target for PAU.

Figure 4.4. Selected results of neuropsychiatric contextualization of cortical proteins and cell-type genes. *Each panel presents the Z scores (β/se) of the primary Mendelian randomization (MR) estimate (IVW or Wald ratio, depending on the number of instrument variants) on 65 neuropsychiatric disorders and traits for the cortical proteins (left) and cell-type genes (right) that demonstrated evidence of replication in the FinnGen cohort. Estimates surpassing correction for multiple comparisons (P-value=7.69×10^{-4} [0.05/65 outcomes]) are presented as Z scores and all cis-instrument MR estimate P-values are based upon used two-sided Wald tests. P-values were not adjusted for multiple comparisons. AST: astrocytes; END: endothelial; EXC: excitatory; MIC: microglia; INH: inhibitory; OLI: oligodendrocytes. Reproduced from Rosoff DB, Wagner J, Bell AS, Mavromatis LA, Jung J & Lohoff FW. "A multi-omics Mendelian randomization study identifies new therapeutic targets for alcohol use disorder and problem drinking." Nature Human Behaviour 2025; 9: 188–207. © 2025 The Authors, licensed under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).*

Table 4.2. High-confidence cortical proteins demonstrating evidence of colocalization and replication

Cortical protein	Alcohol-related outcome	β [95% CI]	P-value	Colocalization (H4)	FinnGen replication (effect direction)	# hits in Phe-MR	General background information on target
SLC4A8	PAU	0.1, [0.058, 0.142]	2.71×10^{-6}	0.97	Alcohol related diseases/death (+); AUD Swedish definition (+)	7	A membrane protein that transports sodium and bicarbonate ions to regulate pH in neurons.
SLC4A8	DPW	0.055, [0.033, 0.077]	9.81×10^{-7}	0.97	--	--	--
RHOA	PAU	-0.114, [-0.162, -0.066]	3.09×10^{-6}	0.96	Alcohol related diseases/death (-); AUD Swedish definition (-); ICD-10 AUD (-)	22	Rho family GTPase involved in signaling pathways that regulate actin cytoskeleton reorganization and are linked to tumor proliferation and metastasis; multiple splice variants exist
GBA2	DPW	-0.012, [-0.015, -0.01]	2.20×10^{-21}	0.79	Alcohol-related pancreatitis (+); Acute intoxication (-)	8	Microsomal beta-glucosidase that breaks down bile acid 3-O-glucosides. Primarily localized to the endoplasmic reticulum and plays a role in carbohydrate transport and metabolism.
SAMHD1	AIF	0.11, [0.083, 0.138]	7.08×10^{-15}	0.98	AUD Swedish definition (-); ICD-10 AUD (-)	3	Involved in innate immune response regulation. Upregulated during viral infections and potentially mediating proinflammatory response.
CAB39L	DPW	0.013, [0.009, 0.016]	2.13×10^{-13}	0.77	AUD Swedish definition (-); Alcohol-related gastritis (-), Alcohol-related liver disease (+)	4	Enables protein serine/threonine kinase activator activity, and involved in intracellular signal transduction, including mTOR complex.
EFNB3	DPW	0.04, [0.03, 0.051]	1.11×10^{-13}	0.70	Alcohol-related polyneuropathy (+)	10	Plays critical role in brain development and function, particularly in the forebrain, and interacts with EPH receptors, which are key in developmental signaling in the nervous system
EFNB3	AIF	0.077, [0.061, 0.094]	2.01×10^{-19}	0.71	--	--	--
SLC5A6	PAU	-0.022, [-0.031, -0.012]	1.13×10^{-5}	0.91	Alcohol related diseases/death (-); AUD Swedish definition (-); ICD-10 AUD (-)	8	Enables biotin and pantothenate transmembrane transport, involved in anion transport and crossing the blood-brain barrier, located in the plasma membrane
SLC5A6	DPW	-0.026, [-0.032, -0.02]	1.44×10^{-17}	0.98	--	--	--
SLC5A6	AIF	-0.052, [-0.061, -0.042]	2.57×10^{-27}	0.85	--	--	--
ULK3	AIF	-0.014, [-0.018, -0.01]	3.26×10^{-11}	0.81	AUD Swedish definition (+)	13	Facilitates protein serine/threonine kinase activity. Involved in fibroblast activation, protein autophosphorylation, and regulating the smoothened signaling pathway, located in the cytoplasm.
HDGF	DPW	0.023, [0.018, 0.029]	4.04×10^{-16}	0.94	Alcohol related diseases/death (+); AUD Swedish definition (+); Alcohol related gastritis (+)	6	Hepatoma-derived growth factor protein that promotes cellular proliferation and differentiation.
CCDC25	DPW	-0.016, [-0.022, -0.01]	1.08×10^{-7}	0.81	AUD Swedish definition (-); ICD-10 AUD (-)	10	Enables DNA binding activity/enhances cell motility. Integral component of the plasma membrane. Located within the endomembrane system.
DOC2A	DPW	-0.032, [-0.044, -0.021]	8.02×10^{-8}	0.99	Alcohol related diseases/death (-); AUD Swedish definition (-); ICD-10 AUD (-)	9	Involved in Ca^{2+}-dependent neurotransmitter release.
DOC2A	AIF	-0.075, [-0.094, -0.057]	7.19×10^{-16}	0.99	--	--	--
VIPAS39	AIF	0.03, [0.018, 0.041]	4.38×10^{-7}	0.89	AUD Swedish definition (-); ICD-10 AUD (-)	6	Facilitates endosome-to-lysosome and intracellular protein transport, contributes to collagen metabolism and peptidyl-lysine hydroxylation. Maintains apical-basolateral polarity.
DPYSL4	AIF	-0.039, [-0.048, -0.03]	6.47×10^{-17}	1.00	Acute intoxication (-)	11	Enables filamin binding. Involved in nervous system development. Located in the cytosol.
C1QTNF4	PAU	0.111, [0.075, 0.147]	1.01×10^{-9}	0.85	Alcohol related diseases/death (+); AUD Swedish definition (+)	21	Enables cytokine activity. Involved in cytokine production enhancement and signal transduction. Role in regulating food intake and energy balance.
SHMT1	AIF	0.021, [0.015, 0.027]	3.46×10^{-12}	0.90	Alcohol-related liver disease (+)	23	Serine hydroxymethyltransferase, catalyzing key reactions for one-carbon unit synthesis in methionine, thymidylate, and purine production,
MAPT	DPW	0.149, [0.115, 0.183]	1.94×10^{-17}	0.86	Alcohol related diseases/death (+); AUD Swedish definition (+); Alcohol related liver disease (+)	31	Promotes microtubule assembly and stability, acting as a linker protein between axonal microtubules and neural plasma membrane components to establish and maintain neuronal polarity.
MAPT	AIF	0.18, [0.127, 0.233]	2.92×10^{-11}	0.86	--	--	--
MAPT	Binge	0.278, [0.193, 0.363]	1.49×10^{-10}	0.86	--	--	--

Notes: *The 16 targets included in the table had cis-instrument MR estimates P-values [CIs]) surpassing Bonferroni-corrected thresholds of 1.41×10^{-5} for the cortical proteome, evidence of a shared causal variant with the respective alcohol-related outcome with colocalization (PP.H4 >0.7), and replication in the FinnGen cohort. All cis-instrument MR estimate P-values are based upon used two-sided Wald tests. P-values were not adjusted for multiple comparisons. Reported are the β's and corresponding 95% confidence intervals. The (+) and (-) signs indicate the direction of the cis-instrument MR estimate of the protein on the FinnGen outcome. The AUD Swedish definition includes electronic health record codings related to a range of physical and psychiatric consequences of alcohol use behavior while the ICD-10 AUD definition included electronic health record hospital discharges and causes of death related to mental and behavioral disorders due to use of alcohol. The # of hits in the phenome-wide MR (Phe-MR) are the number of traits surpassing Bonferroni correction (P-value=7.69×10^{-4} [0.05/65 outcomes]) that were associated with the cortical protein in cis-instrument MR analyses evaluating the neuropsychiatric landscape of the protein across 65 traits. The general information column is information about the protein from the GeneCard Human Gene Database.*[377] *PAU: problematic alcohol use; AIF: alcohol intake frequency; DPW: drinks per week; AUD: alcohol use disorder; ICD-10; International Classification of Disease, Tenth Revision.*

Table 4.3. High-confidence genes with single-cell expression demonstrating evidence of colocalization and replication

Gene	Cell type	Alcohol-related outcome	β [95% CI]	P-value	Colocalization (H4)	FinnGen replication (effect direction)	# hits in Phe-MR	General background information on target
AGRN	OLI	AIF	-0.027, [-0.036, -0.017]	2.42×10^{-8}	0.81	Alcohol-related cardiomyopathy (+)	6	Plays diverse roles in neural processes, including neuromuscular junction dynamics, and stimulating dendritic filopodia formation in brain neurons, reflecting their distinct contributions to neural development and function.
ARL17B	INH	DPW	-0.013, [-0.017, -0.01]	3.43×10^{-15}	0.90	AUD, ICD-coded (-); AUD, Swedish definition (-)	33	Involved in intracellular protein transport and vesicle-mediated transport.
ARL17B	MIC	DPW	-0.014, [-0.017, -0.01]	3.43E-15	0.90	AUD, ICD-coded (-); AUD, Swedish definition (-)	33	--
C2orf82	AST	AIF	0.017, [0.011, 0.024]	3.87×10^{-7}	0.90	AUD, Swedish definition (-)	14	Encodes proteoglycan transmembrane protein highly in white matter, spinal cord, and medulla oblongata and in astrocytes and excitatory neurons.
CAB39L	AST	DPW	0.018, [0.011, 0.025]	4.31×10^{-7}	0.73	Alcohol-induced polyneuropathy (+)	4	Has serine/threonine kinase activator activity. Involved intracellular signal transduction. Role in modulating mTORc.
INO80E	EXC	DPW	0.037, [0.024, 0.05]	1.62×10^{-8}	0.87	Alcohol-related diseases/death (+); AUD, Swedish definition (+); AUD, ICD-coded (+)	14	Involved in DNA recombination; DNA repair; and chromatin remodeling.
LRRC37A	MIC	DPW	-0.045, [-0.056, -0.034]	2.29×10^{-15}	0.91	AUD, Swedish definition (-); AUD, ICD-coded (-)	32	Integral membrane component. Involved in immune and inflammatory responses, cellular migration, and synapse formation.
LRRC37A	OLI	DPW	-0.035, [-0.043, -0.026]	9.65×10^{-16}	0.92	AUD, Swedish definition (-); AUD, ICD-coded (-); Acute alcohol intoxication (+); Alcohol pancreatitis (+)	33	--
LRRC37A	AST	Binge	-0.052, [-0.068, -0.037]	8.23×10^{-11}	0.87	AUD, Swedish definition (-); AUD, ICD-coded (-)	33	--
NPIPB9	AST	DPW	0.0372, [0.026, 0.048]	7.83×10^{-11}	0.75	Alcohol-related pancreatitis (+)	15	Unknown function.
NRBP1	INH	PAU	-0.078, [-0.099, -0.057]	3.34×10^{-13}	0.93	Alcohol-related diseases/death (-); AUD, Swedish definition (-); AUD, ICD-coded (-)	4	Role in endoplasmic reticulum-to-Golgi vesicle transport. Controls synaptic growth through mTOR signaling.
NUP160	MIC	AIF	0.029, [0.022, 0.036]	1.60×10^{-14}	0.90	Alcohol-related diseases/death (+); AUD, Swedish definition (+); AUD, ICD-coded (+)	21	Component of the nuclear pore complex.
SH2B1	EXC	DPW	0.0491, [0.034, 0.064]	7.83×10^{-11}	0.84	Alcohol-related pancreatitis (+)	15	Mediates kinase activation, including nerve growth factor (NGF), brain-derived neurotrophic factor (BDNF), glial cell line-derived neurotrophic factor (GDNF). Role in cytokine and growth factor receptor signaling. Important for brain growth and behaviors, such as energy balance, and body weight.
SH2B1	INH	DPW	0.045, [0.031, 0.058]	7.83×10^{-11}	0.91	Alcohol-related pancreatitis (+)	15	--
SYT14	EXC	Binge	-0.071, [-0.083, -0.059]	6.75×10^{-33}	0.72	Alcohol-related diseases/death (-); AUD, Swedish definition (-); AUD, ICD-coded (-); alcohol-related liver disease (-)	42	Mediates membrane trafficking in synaptic transmission. Dysregulation is linked with impaired neurodevelopment.

Notes: *The 12 targets included in the table had cis-instrument MR estimates P-values [CIs]) surpassing Bonferroni-corrected thresholds of 2.80×10^{-6} in the screen of gene expression in the 8 brain cell types, evidence of a shared causal variant with the respective alcohol-related outcome with colocalization (PP.H4 >0.7), and replication in the FinnGen cohort. All cis-instrument MR estimate P-values are based upon used two-sided Wald tests and were not adjusted for multiple comparisons. Reported are the β's and corresponding 95% confidence intervals. The (+) and (-) signs indicate the direction of the cis-instrument MR estimate of the protein on the FinnGen outcome. The AUD Swedish definition includes electronic health record codings related to a range of physical and psychiatric consequences of alcohol use behavior while the ICD-10 AUD definition included electronic health record hospital discharges and causes of death related to mental and behavioral disorders due to use of alcohol. The # of hits in the phenome-wide*

MR (Phe-MR) are the number of traits surpassing Bonferroni correction (P-value=7.69×10^{-4} [0.05/65 outcomes]) that were associated with the cortical protein in cis-instrument MR analyses evaluating the neuropsychiatric landscape of the protein across 65 traits. The general information column is information about the protein from the GeneCard Human Gene Database.[377] AST: astrocytes; EXC: excitatory neurons; INH: inhibitory neurons; OLI: oligodendrocytes; OPC: oligodendrocyte precursor cells; PER: pericytes; END: endothelial cells; PAU: problematic alcohol use; AIF: alcohol intake frequency; DPW: drinks per week; AUD: alcohol use disorder; ICD-10; International Classification of Disease, Tenth Revision.

4.4. Aim 2 Discussion

We utilized the cis-instrument MR statistical framework[48,50] to analyze how genetic components of cortical proteins and cell-type gene expression levels influence AUD risk and predispositions toward problematic alcohol consumption behaviors. Our study identified 217 cortical proteins and 255 cell-type genes, including 36 novel proteins and 37 genes not previously captured by GWASs of alcohol consumption behavior, other conventional gene-mapping methods, post-mortem brain tissue DEG from AUD patients, or whole blood epigenetic signatures of AUD. This multi-omics approach is critical for uncovering the biological mechanisms behind neuropsychiatric outcomes identified in GWASs,[93,378-380] enhancing our understanding of AUD and alcohol consumption behaviors by revealing tissue and cell-type specific relationships. Our comparison between cortical proteome and cell-type transcriptome results showed minimal overlap, consistent with previous findings that protein levels and gene expression are underpinned by different genetic factors.[228] This suggests that each data type provides unique biological insights important for variant-to-function studies.[381] Moreover, this comprehensive analysis not only aids in understanding the genetic predisposition to AUD but also highlights potential mediators that could link AUD with common clinical comorbidities, such as neurodegeneration and other psychiatric disorders noted in genes like MAPT and INO80E, that will inform future investigation into the biological mechanisms and therapeutic development for these comorbidities. Given the diverse nature of these endpoints, the targets identified from our analyses for a particular alcohol-related outcome may have different implications than those identified for another outcome. For instance, targets linked with binge drinking and weekly DPW might offer insights into acute intervention and harm reduction strategies,[382] whereas those associated with PAU could guide long-term treatment approaches. Thus, while all endpoints are clinically relevant, their importance varies based on the specific aspect of alcohol-related behavior they address and their potential utility in guiding therapeutic interventions. This nuanced approach in our study aims to provide a more comprehensive understanding of the genetic underpinnings of alcohol-related outcomes, facilitating the development of targeted and effective treatment strategies.

Regarding our comparison of the top targets for each alcohol-related outcome, we found minimal overlap of the identified proteins and genes or biological pathways across PAU and alcohol consumption behaviors, aligning with previous work finding distinct phenotypic signatures of different alcohol consumption behaviors,[94] and suggesting that these differences extend to both the proteomic and transcriptomic levels in the brain. However, our cis-MR of brain gray matter structures, white matter tracts, and resting state functional connectivity did suggest some convergence of function of the alcohol-related proteins and genes on brain physiology. Nonetheless, the minimal proteomic and transcriptomic overlap may have important implications for development of therapeutics aimed at reducing problematic alcohol consumption, suggesting that strategies aimed at targets underlying AUD may be different than targeting genes related to binge drinking. Future studies will be important to further clarify the roles of the apparent differences in the biological underpinnings of genetically predisposed AUD and alcohol consumption. We found generally consistent directional relationships among the genes that had MR estimates surpassing correction for multiple comparisons in more than one cell type (e.g., *RBM6* in 5 cell types, and *FAM118A*), suggesting conserved impact of expression across different cell types, aligning with recent QTL-based cell-type analyses in other psychiatric

outcomes.[102] We extend these observations by showing similarly conserved directional relationships between single cell and bulk transcriptomic tissues (and across brain regions), which may have implications for therapeutic development, as these results suggest that pharmacological approaches (i.e., inhibition or activation) for the intended proteins or genes within these tissues may have directionally consistent relationships regardless of their tissue type.

4.4.1. Drug-gene interaction highlights potential for antidiabetics addressing alcohol behaviors and cardiometabolic disease

Cross-referencing the cMAP[256] and DGIdb databases,[257] we identified potential drug repurposing candidates for alcohol use behaviors. Increased DRD2 expression in excitatory neurons correlates with beneficial outcomes in reducing alcohol use, supported by evidence from drugs like cabergoline and sumanirole used in Parkinson's disease treatment. Cabergoline has been shown to reduce alcohol consumption and relapse risk in rats by modulating neurotrophic pathways.[383] Conversely, sumanirole demonstrated initial decreases in alcohol intake followed by increases, suggesting limited utility for persistent alcohol use reduction.[384] Additionally, some antidiabetics, previously linked to neuropsychiatric benefits, also showed potential. For example, glyburide, voglibose, and acarbose have effects on depression and cognitive impairment in animal models.[385-387] These findings, together with recent links between antidiabetics like GLP-1R analogs and reduced alcohol consumption,[388] suggest a significant connection between cardiometabolic metabolic processes and substance use disorders along the gut-brain axis that may be important for addressing behaviors linking these comorbidities.[388] Moreover, our study highlights several targets amenable to small molecule intervention, although druggability definitions are evolving with biotechnological advances, including biologics like monoclonal antibodies and RNA therapeutics.[126,389] Our signature matching approach[256] supported candidates such as prazosin and memantine for alcohol-related behaviors, based on their negative connectivity with alcohol consumption signatures and existing literature on their effects in reducing alcohol cravings and intake in clinical and preclinical settings.[390-392] Our findings advocate for further validation and investigation of these drugs for repurposing, reinforced by genetics-based evidence and consistent results across different studies.[393]

After assessment for novelty, colocalization, and replication, we prioritized 18 high-confidence proteins and 12 cell-type genes showing strong associations with PAU and alcohol consumption behaviors. These proteins and genes include potential therapeutic targets for AUD, such as the neurotransmitter-regulating solute carrier transporters SLC4A8 and SLC5A6, which may influence dopamine and serotonin balance.[375] Additionally, CAB39L, involved in the mTOR pathway,[367] supports the hypothesis from rodent studies that mTOR inhibition could reduce alcohol consumption. This is corroborated by human genetics linking CAB39L and MTOR with alcohol use, suggesting a targeted mTOR inhibition strategy in the brain could minimize side effects typically associated with systemic mTOR inhibition.[367] VIPAS39, novel in our stringent gene prioritization, is implicated in regulating endothelial cell polarity and maintaining the blood-brain barrier (BBB).[394,395] Disruption in BBB integrity is a feature of various

neurodegenerative and psychiatric illnesses,[396] making VIPAS39 a promising candidate for further investigation into its role in AUD.

Among the 12 high-confidence cell-type genes, NRBP1 in INH, NUP160 in MIC, and INO80E in EXC demonstrated protective effects for AUD and reduced drinking behaviors. NRBP1, a synaptic stability regulator found in Drosophila, controls synaptic growth through mTOR signaling via cap-dependent translation (eukaryotic initiation factor 4E),[397] linking it to alcohol behavior effects. NRBP1 dysfunction is linked to neurodegeneration and synaptic problems,[95,398,399] with NRBP1 knockouts in Drosophila showing increased neurodegeneration,[397] aligning with our cis-MR estimates suggesting increased NRBP1 in INH would be the therapeutically relevant direction, potentially by mitigating AUD-related synaptic damage. NUP160, part of the nuclear pore complexes (NPCs) on the nuclear membrane,[400] is critical for nucleo-cytoplasmic transport and is implicated in neurodegenerative diseases due to its role in forming tau-positive neurofibrillary tangles near the nuclear envelope.[401] Postmortem studies in chronic AUD patients have found these tangles in the basal forebrain,[402] underscoring NPCs' role in alcohol-induced neurodegeneration. Our neuropsychiatric cis-MR screen linked INO80E with schizophrenia, building on previous work that identified genetic overlap between AUD and schizophrenia.[403] Approximately 30% of individuals with schizophrenia also have an AUD diagnosis,[404] increasing the risk for clinical complications and reducing medication adherence.[405] INO80E is involved in DNA repair and chromatin remodeling[406,407] and is a potential target to address accelerated aging observed in AUD populations[408-410] that may be a driver of cardiometabolic diseases.

The integration of Aim 2 alcohol-related findings into the broader cardiometabolic framework of this PhD and my first-author research undertaken during my PhD[195,196,411] (**Appendix 1**) investigating the genetic underpinnings of healthy aging provides a compelling opportunity to identify novel therapeutic targets for CVDs by modulating alcohol consumption. Recent MR studies indicate that mTOR pathway inhibition can modulate metabolic traits such as body mass index (BMI) and basal metabolic rate (BMR), alongside protective effects on lifespan, positioning the mTOR signaling pathway as a pivotal intersection linking alcohol use behaviors, systemic metabolic regulation, and longevity.[320] The identification of CAB39L, connected to mTOR signaling, underscores the dual benefits of targeting this pathway to reduce alcohol consumption and mitigate cardiometabolic risk factors, including elevated BMI and BMR. Excessive alcohol consumption disrupts metabolic pathways, increases oxidative stress, and promotes systemic inflammation, all which drive CVD progression, making CAB39L a promising therapeutic target for both alcohol use reduction and cardiometabolic health improvement. These shared molecular pathways, such as mTOR signaling, provide a framework for cross-cutting targets that influence both alcohol use and cardiometabolic traits, creating synergistic opportunities for cardiovascular risk reduction. The reduction in BMI and BMR observed with mTOR inhibition complements the metabolic benefits of decreased alcohol consumption, reinforcing the potential of alcohol-related mechanisms to yield far-reaching benefits that extend beyond CVD prevention to impact systemic health and lifespan. Leveraging tools such as MR and multi-omics approaches, this work identifies novel therapeutic targets and lays the groundwork for precision medicine strategies addressing the intertwined burdens of alcohol use and cardiometabolic diseases, emphasizing the importance of translating these findings into clinical applications that balance efficacy and safety.

4.5. Aim 2 Strengths & Limitations

The Aim 2 project has several strengths. First, it comprehensively sourced different levels of "-omics" for direct comparison across various biological resolutions. Second, complementary statistical gene prioritization methods were used to evaluate the robustness of findings and characterize novel targets. Third, it is the first to resolve proteomic and transcriptomic underpinnings across alcohol consumption behaviors and contextualize these findings with neuropsychiatric disorders. For example, while previous studies linked MAPT to alcohol consumption and neuropsychiatric disorders,[95] we found widespread pleiotropy for proteins and genes associated with neuropsychiatric disorders and behavioral outcomes, explaining the genetic correlations and comorbidities of problematic alcohol consumption and neuropsychiatric outcomes.[188,271,272] Additionally, since poor safety profiles cause about 24% of drug failures,[412] our analyses on neuropsychiatric outcomes help screen for adverse neuropsychiatric relationships, prioritizing therapeutic targets.[88] For instance, ARL17B expression in AST, INH, and MIC, while potentially involved in alcohol consumption, showed highly pleiotropic relationships with neuropsychiatric outcomes, suggesting adverse consequences for reducing alcohol consumption behaviors. Further work is needed to clarify the role of ARL17B, but this profiling suggests its reduced potential as a therapeutic target.

There are study limitations (see **Appendix 3 Supplementary Discussion** for an extended presentation). First, although cis-instrument MR is less prone to horizontal pleiotropy than polygenic MR,[6,50] potential biases due to confounding or pleiotropy remain. Sensitivity analyses showed consistent MR estimates, but interpreting these results requires considering MR assumptions and constraints, such as 'on-target' gene effects and the absence of gene-environment interactions.[50] Second, our instrument construction followed the cis-instrument/drug-target MR framework outlined by Schmidt et al.,[50] using cis-QTLs with P-values $< 5\times10^{-4}$, which may reduce precision[50] but improve model performance.[50,200,237,239,413,414] Although the instruments had F-statistics >10, weak-instrument bias remains a concern but would attenuate results toward the null in two-sample MR analyses.[242,415] Third, single-cell eQTL data has technical challenges, including amplification bias[416] and data sparsity,[417] affecting data analyses. Conventional scRNA-seq methods lose spatial information,[418] though newer methods promise improvements but introduce classification challenges.[419] Additionally, the QTL and GWAS data from common variants prevented assessment against rare variants, important for understanding AUD pathophysiology.[420] Fourth, colocalization assumed a single causal variant, potentially missing signals due to data constraints. Advanced methods like SuSiE *coloc* require high SNP density but risk inflated false positives without sufficient SNPs.[421] Our data's SNP range (40 to ~600 per locus) was insufficient for these methods.[421] Fifth, cortical proteomic data from three regions necessitate future analyses with more brain regions. Gene expression instruments varied across cell types, possibly missing causal genes. Sixth, self-reported alcohol consumption data may be biased,[185] but our EHR-based replication using the FinnGen cohort helps address this potential bias source. Moreover, alcohol-related variables reflect recent behaviors, not long-term patterns.[422] Finally, analyses were based on white, European ancestry participants, limiting generalizability. The AIF, DPW, and binge drinking phenotypes were primarily from educated, healthier UK Biobank participants,[423] not representing global drinking behavior diversity.[424]

4.6. Aim 2 Conclusions

We used cis-instrument MR and colocalization to study the genetic liabilities in alcohol consumption behaviors, identifying novel targets for PAU, AIF, DPW, and binge drinking. Differences in the proteomic and transcriptomic architecture across alcohol behaviors, tissues, and omics levels underscore the need for comprehensive evaluation. Replication highlighted genes involved in mTOR signaling (CAB39L and NRBP1), supporting the role of mTOR inhibition in alcohol behaviors.[425] Other targets like SAMHD1, VIPAS39, NUP160, and INO80E will further inform future efforts to reduce AUD and potentially its cardiometabolic sequalae.

CHAPTER 5: AIM 3 RESULTS

Chapter Overview

This chapter presents Aim 3's comprehensive investigation into the shared genetic architecture of cardiometabolic diseases, focusing on non-alcoholic fatty liver disease (NAFLD), type 2 diabetes (T2D), and coronary artery disease (CAD). The analysis utilized advanced methods, including Linkage Disequilibrium Score Regression and Genomic structural equation modeling (GenomicSEM), to model shared genetic liabilities across these diseases, resulting in the identification of the cardiometabolic factor ("*CM-Factor"*). This multivariate approach uncovered 523 significant SNPs across 312 loci, with substantial novelty in NAFLD and CAD. SNP-level heterogeneity tests revealed loci uniquely influencing individual traits, while bioinformatics analyses identified key genes like *COMT, DHX36,* and *ASPRV1* as critical players in cardiometabolic health. Drug-target Mendelian randomization (MR) analyses validated therapeutic targets, such as GLP1R and PCSK9, and prioritized novel druggable genes, including *CRY2* and *OPRL1,* for cardiometabolic health interventions. Two-step MR further linked proteins like ENO3 to the impact of obesity on cardiometabolic disease. These findings underscore the utility of multivariate frameworks for elucidating genetic risk factors and guiding therapeutic strategies for cardiometabolic multimorbidity.

5.1. Genetic Correlations and SEM modeling

LDSC indicated that the univariate GWASs representing the genetic liabilities of NAFLD, T2D, and CAD were positively correlated (**Figure 5.1a, Tables AP4.2, AP4.3**). Therefore, we proceeded to perform genomic structural equation modeling (SEM) for the multivariate GWAS. The common factor model fit of the implied genetic covariance matrix between NAFLD, T2D and CAD with the empirical covariance matrix was excellent according to the GenomicSEM model criteria[16] (CFI=1 and SRMR=2.09×10^{-9}) (**Figure 5.1b, Table AP4.4**), supporting the hypothesized shared genetic *CM-Factor*.

5.2. Multivariate GWAS

We next fit the SEM model of the general *CM-Factor* to individual variants and conducted a multivariate GWAS using GenomicSEM estimating SNP-level associations for 5,547,254 SNPs remaining after GenomicSEM quality control (QC). LDSC estimated *CM-Factor* total observed scale heritability (h^2)=0.103 (standard error [SE]=0.0033), mean χ^2=2.44 and λ_{GC}=1.98; the LD score intercept=1.028 (SE=0.0033),[16,275] suggesting GWAS inflation was due predominantly to polygenic heritability rather than population stratification bias (**Figures AP4.1, AP4.2, Table AP4.4**).[16,275] The effective sample size was calculated to be 716,916 and the genetic signature was comprised of 523 independent SNPs (P-value < 5×10^{-8}; LD r^2 < 0.1) in 312 genomic loci (**Figure 2c**, **Table AP4.5**).[18] As expected, most of the lead SNPs were intronic (**Figure AP4.3**);

however, there were 77 lead SNPs considered deleterious, as indicated by Combined Annotation Dependent Depletion (CADD) scores >10. 24 of these SNPs were exonic, including rs10305420 in the *GLP1R* locus and rs12140153 in the *PATJ/INADL* locus (**Figures AP4.4-AP4.8, Table AP4.6).**

Figure 5.1. ***The CM-Factor GWAS modeled with GenomicSEM. (a)*** *Genetic correlations for the three input, univariate GWASs, i.e., non-alcoholic fatty liver disease (NAFLD), type 2 diabetes (T2D), and coronary artery disease (CAD), calculated using structural equation modeling with GenomicSEM, displaying pairwise LD Score genetic correlation estimates above each bar plot (r_g and corresponding standard error [SE]).* ***(b)*** *Path diagram of the common factor model GWAS assessing the genetics shared across NAFLD, T2D, and CAD estimated with Genomic SEM—termed the cardiometabolic (CM) factor or "CM-Factor"—with standardized*

factor loadings (SEs in parentheses). ***(c)*** *Top panel is the Manhattan plot showing SNP associations (−log_{10}(P-values)) with the CM-Factor GWAS, ordered by chromosome; and the bottom panel is the genome-wide Q_{SNP} heterogeneity plot with −log_{10} Q_{SNP} P-values on the y-axis. The red dashed line indicates the threshold for conventional genome-wide significance (P-value=5×10^{-8}). The presented P-values are two-sided and have not been adjusted for multiple testing.*

If the *CM-Factor* loci reflected a shared genetic liability to cardiometabolic disease, it would be expected that heterogeneity in the SNP-level associations would be primarily in regions of the genome not associated with the *CM-Factor*. Therefore, GenomicSEM was also used to perform genome-wide SNP-level heterogeneity tests (**Figure 5.1c**) to assess whether a SNP had associations with the three-input univariate cardiometabolic GWASs operating only through the *CM-Factor*. There were 151 heterogenous loci (258 independent lead Q_{SNPs}). Only ~4% of 19,983 total genome-wide significant *CM-Factor* SNPs were also highly heterogenous (i.e., Q_{SNP} P-values < 5×10^{-8}) (**Figure 5.2**), suggesting that the *CM-Factor* reflects a shared cardiometabolic architecture. Further, there was strong representation of canonical lipid-related loci and genes among the heterogenous Q_{SNP} loci capturing multiple aspects of lipid and cholesterol metabolism (**Tables AP4.7-AP4.8**). For example, in addition to the exonic variant rs429358 in the *APOE* locus, with the strongest Q_{SNP} signal (two-sided Q_{SNP} P-value=4.15×10^{-68}), other Q_{SNP} loci were involved in lipid and cholesterol metabolism,[426] i.e., rs12740374 near *CELSR2-PSRC1-MYBPHL-SORT1*, rs55997232 in the *LDLR* locus, rs11591147 in the *PCSK9* locus, highlighting pathways in lipid metabolism that might operate independently of the broad *CM-Factor* to impact the cardiometabolic health. PNPLA3 variants were also among the heterogenous loci, likely reflecting the role of PNPLA3 in hepatic lipid sequestration and reported inverse relationship between NAFLD and CAD.[427] See **Appendix 4 (Section 4.2)** for an extended presentation of these results.

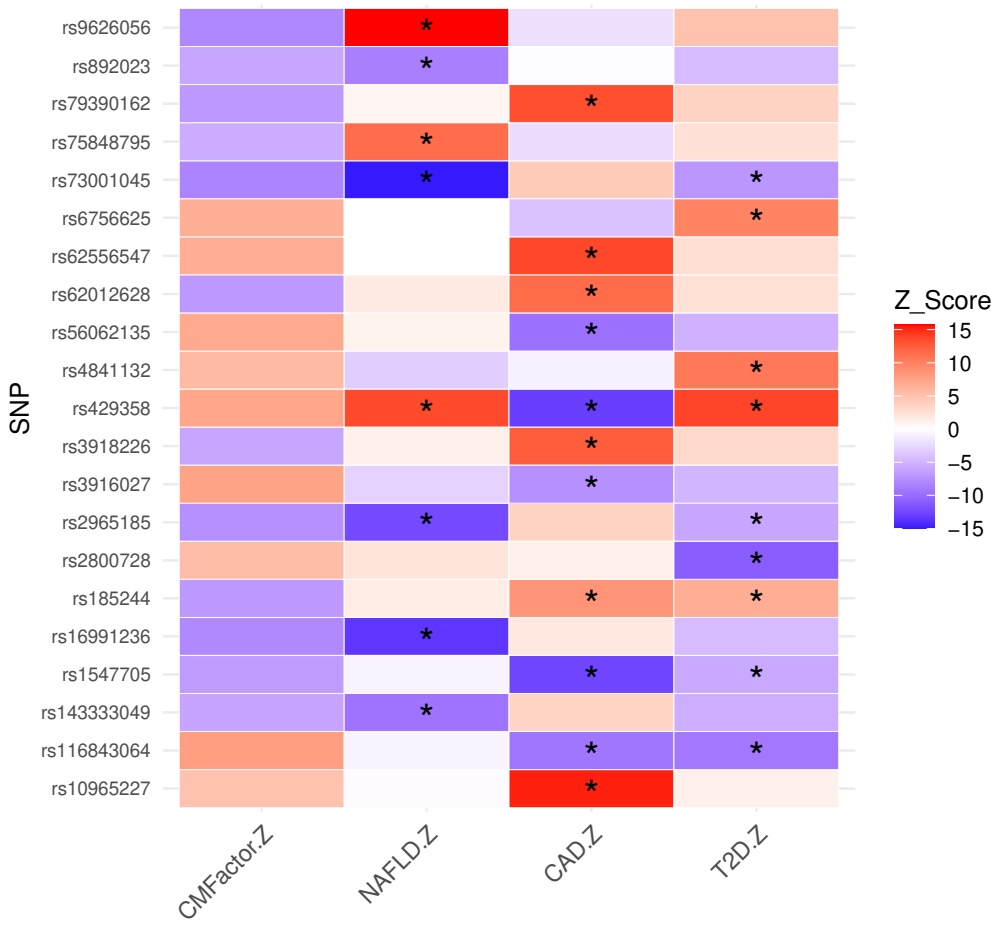


Figure 5.2. Comparison of the effect estimates of independent SNPs with Q_{SNP} P-values < 5× 10^{-8} that were also genome-wide significant in the CM-Factor. *Heatmap plots the Z scores for the multivariate CM-Factor GWAS summary statistics (Beta/SE) and the univariate NAFLD, T2D, and CAD GWAS summary statistics used to generate the CM-Factor. The SNPs are the lead SNPs Q_{SNP} in the genome-wide Q_{SNP} analysis that were also genome-wide significant for the CM-Factor. Asterisks indicate that the SNP surpassed conventional genome-wide significance in the univariate GWAS (P-value < 5×10^{-8}).*

Next, we evaluated the concordance in estimate directions (i.e., the signs) for the 523 *CM-Factor* SNPs in the input GWASs (**Table AP4.9**). This analysis of concordance among cardiometabolic factor (*CM-Factor*) lead SNPs reveals distinct patterns in SNP associations across three traits: T2D, CAD, and NAFLD. In total, 75.7% (396 SNPs) show concordance across all three traits, with various configurations of significance, including cases where all three traits present a P-value < 0.05 (N=94). Additionally, 20.85% (109 SNPs) exhibit concordance across two traits, with the most common being T2D and CAD (N=145) showing P-values < 0.05. Conversely, discordance between two traits with significant P-values is rare, occurring in only 7 variants (1.3%). In these instances, the *CM-Factor* coefficient tracked the direction of the T2D coefficient, and the other 2 traits were either non-significant or only nominally significant. The

assessment of sign concordance in the input GWASs further supports the interpretation that the *CM-Factor* reflects a shared genetic liability of cardiometabolic disease.

5.2.1. Novelty and phenotypic analyses of *CM-Factor* loci

Comparing the *CM-Factor* lead SNPs with the genetic signatures of the input NAFLD, T2D, and CAD, we found only 9 of the lead variants had been identified by the genetic signature of NAFLD while only 108 were genome-wide significant in the Aragam et al. GWAS of CAD[277] (**Table AP4.9**). Many of the novel loci for NAFLD and CAD had some level of statistical association (i.e., P-values < 0.05) in the univariate NAFLD (152 of 514 novel) or CAD (185 of 415) GWASs (**Table AP4.5**), suggesting they would be captured in larger univariate GWASs of these diseases. The Suzuki et al. T2D[276] signature was strongly represented in the *CM-Factor*, likely reflecting its large sample size and substantial contribution to the genetic liability of cardiometabolic health; however, 6 *CM-Factor* loci were still not captured in the univariate T2D GWAS. Further, 3 loci (with lead variants rs77424687, rs139562826, and rs1737897) (**Figures AP4.6-4.8**) were not captured by any of the input GWASs. Further, as the multivariate *CM-Factor* genetic signature prioritized some, but not all, of the ~660 T2D loci (1,457 lead variants) identified by Suzuki et al., it can be inferred that the *CM-Factor* has prioritized specific T2D loci for their broader relationships with cardioembolic health, underscoring the ability of GenomicSEM to inform loci discovery and prioritization in multi-trait applications.

We aimed to investigate phenotypic associations of the *CM-Factor* loci by performing a phenome-wide association query in the EMBL-EBI GWAS Catalog using the built-in GWAS Catalog query in the FUMA ("functional mapping and annotation of genetic associations") package (v1.5.2[280]) and its default settings to define independent genetic signals and account for linkage disequilibrium. As expected, there was widespread representation of the *CM-Factor* loci in the GWAS Catalog (**Table AP4.10**). 48 *CM-Factor* loci were associated with >100 traits in the catalog (e.g., the genomic locus with lead variants rs7841189 and rs11204087 was associated with 1,328 traits), while others, such as the locus with lead variant rs302864 on chromosome 17 was only associated with T2D (**Table AP4.11**). Similarly, when evaluating the pleiotropy by trait, top traits were primarily related to biomarkers and risk factors related to cardiometabolic health. For example, BMI was associated with 109 loci, high-density lipoprotein (HDL-C) and triglycerides with 88 and 85 loci, respectively, and glycated hemoglobin (HbA1c) with 75 loci (**Table AP4.12**), indicating a strong representation of major cardiometabolic biomarkers and risk factors among the *CM-Factor* loci.

5.2.2. Sensitivity tests of the *CM-Factor*

Given the strong enrichment of the *CM-Factor* for traits related to adiposity observed in the GWAS Catalog query, we performed sensitivity tests of the *CM-Factor* genetic signature by incorporating genetics of BMI into the multivariate modeling with GWAS-by-subtraction (GBS) models[279] implemented in GenomicSEM (**Figure 2.6, Tables AP4.13-AP4.15**). 259 of the lead variants remained genome-wide significant after removing the genetic contribution of BMI and another 148 surpassed the more relaxed P-value threshold of 9.56×10^{-5} (0.05/523 lead variants

tested). The estimates were similar in additional GBS models conditioning on the genetic effects of waist-to-hip-ratio adjusted for BMI (265 lead variants surpassed conventional genome-wide significance and additional 218 variants at the relaxed P-value threshold). In addition, for both models, the effect directions of the SNP associations were consistent after accounting for BMI or WHRadjBMI, suggesting the loci effects on cardiometabolic health are robust when accounting for body mass (**Tables AP4.14-AP4.15**).

5.2.3. SNP and gene prioritization with fine-mapping and transcriptomic imputation

We conducted fine-mapping for the *CM-Factor* loci with SuSieR.[284] 82 of the loci fine-mapped with generally good 95% credible set resolution: 45 credible sets resolved to single variants and the maximum credible set size was 9 (**Table AP4.16**). There were 81 genome-wide significant SNPs that also had posterior inclusion probabilities (PIPs) ≥0.8 (**Table AP4.17**) that further clarified the causal roles of the *CM-Factor* loci: e.g., the exonic variants identified as lead SNPs in the *INADL/PATJ* and *GLP1R* loci each demonstrated perfect posterior inclusion probabilities with the *CM-Factor* (PIPs=1). The *INADL/PATJ* SNP rs12140153 also demonstrated both epigenomic associations (e.g., multiple enhancer histone marks such as H3K4me1 and H3K27ac), and is enriched with motifs for several transcription factors, suggesting a regulatory role in *INADL/PATJ* expression (**Table AP4.18**). Similarly, the *GLP1R* SNP rs10305420 features extensive enhancer histone marks like H3K4me1, as well as promoter histone marks like H3K4me3 and H3K9ac, and linked with multiple transcription factor binding sites, indicating its potential impact on gene expression regulation across various biological contexts, and suggesting that rs10305420 plays a significant regulatory role in *GLP1R* expression. There were several other loci with causal variants in genes encoding drug targets, e.g., rs2074311 in *ABCC8/KCNJ11*, the target for sulfonylureas; rs17036160 in *PPARG*, the target for thiazolidinediones (TZDs)[301,302]; and rs59325138 near *APOC1*, which is an investigational lipid-modulating target.[305] Further, many of the loci demonstrated extensive mapping to epigenomic features and corresponding gene expression in tissue relevant for cardiometabolic health (**Table AP4.17**), highlighting their potential roles in gene regulation and chromatin modifications, and altogether contributing to the understanding of complex biological processes underlying broad cardiometabolic disease risk.

Next, we performed a transcriptome-wide association study (TWAS) with the FUSION method[103] to identify gene-level associations with the *CM-Factor* (**Figure 5.3, Tables AP4.19-AP4.21**). Across the cardiometabolic tissues used in the analyses, we identified 789 genes surpassing the Bonferroni correction for multiple comparisons (P-value < 6.25×10^{-7} [0.05/81,246 total tests]). We took these genes forward for further testing, including colocalization[106] and conditional testing as part of the FUSION TWAS pipeline.[103] 486 of these genes colocalized under the single causal variant assumption at PP.H4 ≥0.6, supporting their evidence of causal role in the *CM-Factor*. We then performed conditional testing on the TWAS genes because TWAS often identify many associated genes in the locus/tissue, making it difficult to resolve the conditionally independent gene driving the association.[103] Conditional testing prioritized 243 conditionally independent signals that we considered “high-confidence” gene-level associations, including 112 that would not have been captured by using more conventional gene-prioritization methods (i.e., MAGMA) (**Table AP4.20**). TWAS may identify novel biology not captured by

the underlying genetic architecture of the traits of interest,[103] and among the high-confidence genes, 11 were novel/not captured by genomic loci of the input T2D, NAFLD, or CAD GWASs, the *CM-Factor*, or in a GWAS Catalog look-up (**Table AP4.21**), including 5 protein coding genes, *ASPRV1*, *COMT*, *DHX36*, *CCDC188*, and *PIP4K2A* (**Table 5.1**).

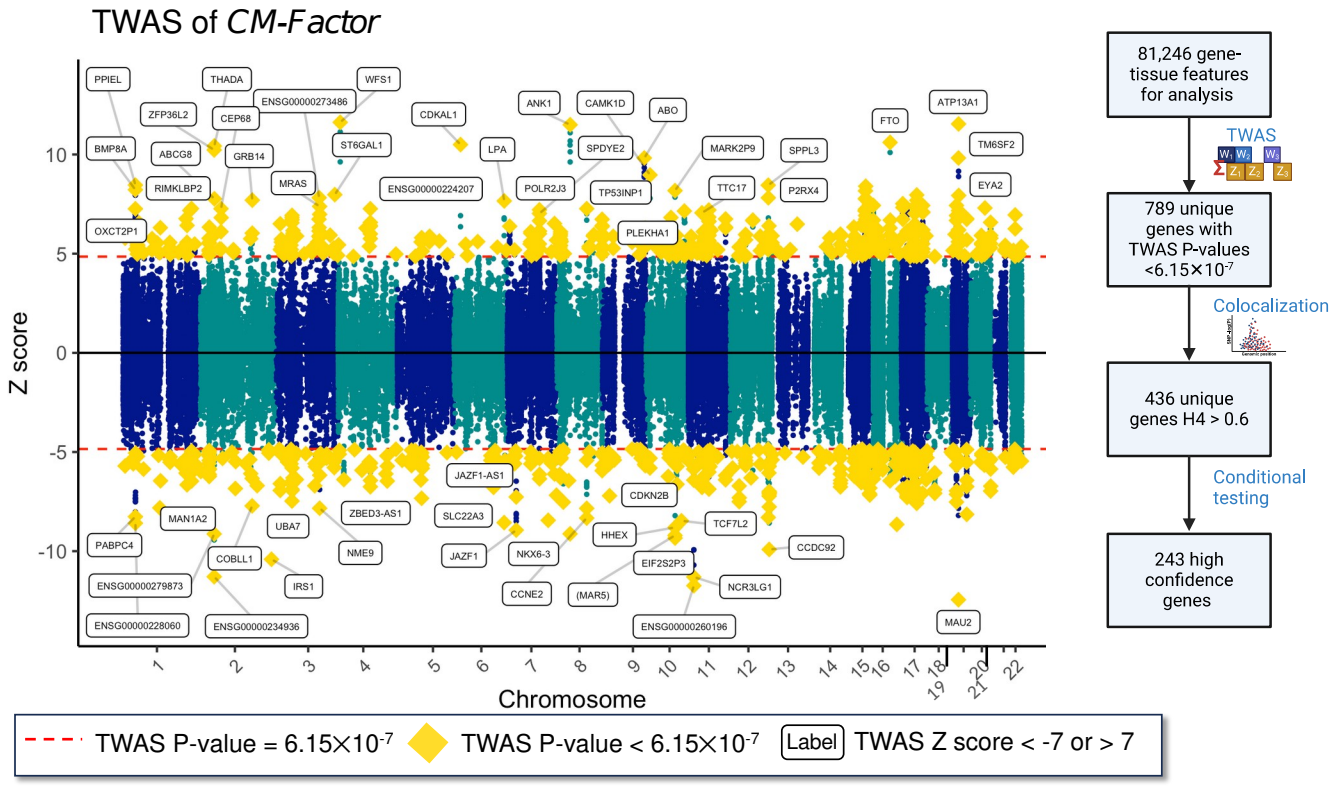


Figure 5.3. Transcriptomic imputation results for the CM-Factor. *Manhattan plot presents the FUSION transcriptome wide association study (TWAS) results using expression quantitative trait loci (eQTL) data from cardiometabolic-relevant tissues or the CM-Factor. The red dotted lines indicate the TWAS study Bonferroni correction for multiple comparisons (P-value < 6.15×10^{-7} [0.05/81,246 total tests in the TWAS analysis]). Flow chart outlines the results of the individual steps in the transcriptomic imputation, colocalization, and conditional testing that prioritized high confidence CM-Factor genes. The labeled genes are those with TWAS Z scores above 7 or less than -7. PP.H4 is the posterior probability of colocalization between the gene expression and CM-Factor (i.e., there is a single shared genetic variant explains the associations for both traits within the specified region) and we considered PP.H4 value > 0.6 as evidence of colocalization, suggesting that the observed associations for both traits may be due to a common causal SNP. FUSION uses a linear regression model to test the association between predicted gene expression (derived from reference transcriptome data) and a phenotype of interest. The Z scores indicate the strength of the association (not adjusted for multiple comparisons), and their estimates include two-sided P-value testing for associations in both directions (positive and negative).*

Another advantage of TWAS versus other gene prioritization approaches is that it provides a regression estimate with a direction linking the genetically-predicted gene expression with the outcome, facilitating inference towards how expression of the gene in the tissue influences the *CM-Factor*.[103] We found 128 high-confidence genes with negative TWAS estimates (Z-scores), indicating that increased expression of these genes reduces *CM-Factor* risk (e.g., the results indicated increased insulin receptor substrate 1 [IRS1] expression in adipose tissue would reduce *CM-Factor* risk, aligning with its canonical role in improving insulin sensitivity and lipid metabolism[428]). Also, comparing the estimates across multiple bulk tissue expression panels showed strong consistency in the direction of the gene expression associations with the *CM-Factor*. For example, Ankyrin 1 (*ANK1*), which functions to connect integral membrane proteins to the underlying spectrin-actin cytoskeleton in muscle and erythroid cells,[429] had strong positive TWAS associations in adipose, heart (both left ventricle and atrial appendage), and muscle tissue, suggesting conserved mechanisms of action across multiple tissue types for cardiometabolic health. Drug-gene enrichment analyses confirmed involvement of the high-confidence genes in ischemic heart diseases and lipid modifying agents (**Tables AP4.22-AP4.23**), and transcriptomic signature matching with ~3,000 drug-gene signatures in the CMap touchstone dataset[429] identified strong negative connectivity scores with HO-013, a PPAR receptor agonist, 4-hydroxyretinoic-acid, a retinoid receptor binder, and TWS-119, a glycogen synthase kinase inhibitor, suggesting these compounds have signatures that would offset the transcriptomic signature of the *CM-Factor* (**Table AP4.24**), and underscoring the potential utility of TWAS in uncovering novel gene associations and guiding therapeutic strategies aimed at mitigating cardiometabolic risks.

Table 5.1. Novel protein coding genes identified with transcriptomic imputation, colocalization, and conditional testing

				Transcriptomic imputation test statistics			*Colocalization*	*Conditional testing*	
Gene	**Name**	**Tissue source**	**Genomic position**	**Gene heritability**	**Z score**	**P-value**	**H4**	**Conditional Z score**	**Conditional P-value**
ASPRV1	Aspartic Peptidase Retroviral Like 1	Whole blood	chr2:70,187,226-70,189,397	10.4%	-5.3	1.02×10^{-7}	0.94	-5.3	1.00×10^{-7}
COMT	Catechol-O-Methyltransferase	Adipose (METSIM study)	chr22:19,929,130-19,957,498	8.3%	-5.3	1.13×10^{-7}	0.73	-3.1	0.0023
DHX36	DEAH-Box Helicase 36	Tibial artery	chr3:153,990,335-154,042,286	4.6%	-5.3	1.16×10^{-7}	0.83	-5.3	1.20×10^{-7}
CCDC188	Coiled-Coil Domain Containing 188	Tibial artery	chr22:20,135,950-20,138,399	16.2%	-5.47	4.45×10^{-8}	0.79	-3.3	0.00084
PIP4K2A	Phosphatidylinositol-5-Phosphate 4-Kinase Type 2 Alpha	Tibial artery	chr10:22,823,778-23,003,484	13.5%	5.48	4.21×10^{-8}	0.65	5.5	4.20×10^{-8}

Notes: *This table highlights five novel protein-coding genes identified with transcriptomic imputation by performing a transcriptome wide association study followed by colocalization analyses and conditional testing. Each gene was prioritized as a "high-confidence" association for the CM-Factor based on stringent statistical criteria: Bonferroni correction for TWAS results (P-value < 6.15×10^{-7}; P-values based upon two-sided tests, not adjusted for multiple*

comparisons); evidence of a shared causal variant (colocalization posterior probability of the fourth hypothesis that a single shared causal variant exists for both the gene expression and CM-Factor [PP.H4] > 0.6); and conditional significance testing to resolve independence across multiple signals within a given locus. These genes were not found from prior univariate GWAS loci for NAFLD, T2D, CAD, or the CM-Factor, underscoring their novelty in the genetic architecture of cardiometabolic multimorbidity. Gene heritability refers to the proportion of variability in gene expression levels that can be attributed to genetic factors, specifically the effects of cis-genetic variants (variants located within or near the gene's locus) and quantifies how much of the gene expression is heritable and thus influenced by genetic variants.

5.2.4. Pathway, bulk tissue, and cell-type enrichment

We proceeded to perform bio-annotation of the *CM-Factor*. First, we investigated its enrichment in pathways and biological processes, bulk tissue, and cell types. Biological processes and pathways enrichment analyses found 14 gene-sets with P-values < 2.94×10^{-6} (0.05/17,009 gene-sets tested) (**Table AP4.25**). The identified biological processes and pathways, such as the regulation of RNA metabolic processes, cell differentiation, Notch signaling, macromolecule biosynthesis, and transcription by RNA polymerase II, are critical to maintaining cellular and physiological homeostasis in cardiometabolic tissues,[430,431] suggesting that the genetic signature of the *CM-Factor* is capturing fundamental biological processes with strong relevance to cardiometabolic health. Partitioned heritability using annotations from the ENCODE project[293] and Roadmap Epigenomics[294] identified 12 tissue types surpassing Bonferroni correction for multiple comparisons. Top hits were related to pancreatic islets and brain tissues (**Figure AP4.9-AP4.10, Table AP4.26**). There was also evidence of enrichment in the liver and adipose nuclei at P-value < 0.05. S-LDSC results from the gene expression data broadly aligned with the chromatin-based tissue enrichment but were less precise (12 tissues in mainly pancreatic, brain, and liver tissue had enrichment P-values < 0.05, but none surpassed stringent Bonferroni correction) (**Table AP4.27**). The single-cell analyses across 115 cell types demonstrated similar patterns of enrichment (**Figures AP4.11-AP.4.12, Tables AP4.28-AP4.29)**. For example, partitioned heritability of the *CM-Factor* at single cell resolution found enrichment predominantly in pancreatic cell types (5 pancreatic cell types surpassed stringent multiple testing correction), including pancreatic α cells (P-value=4.67×10^{-6}), PP cells (P-value=8.08×10^{-6}), and β cells (P-value=1.75×10^{-5}).

5.3. Mendelian Randomization Analyses

As outlined in Chapter 2 (**Aim 3 Methods section 2.4.4**), we use the MR framework in several contexts, including (1) assessing the cardiometabolic efficacy of approved therapeutic targets and those presently in clinical trials with drug-target MR;[432] (2) screening the druggable genome[126] to inform cardiometabolic drug discovery; and (3) implementing a two-step MR approach leveraging the circulating proteome to elucidate proteomic mediators that connect body mass index (BMI) with cardiometabolic health outcomes.

5.3.1. MR Study 1: Drug-target analysis of approved and investigational cardiometabolic therapeutics

We curated 34 approved and investigational cardiometabolic drug targets to assess their efficacy for the broad *CM-Factor* genetic liability with drug-target MR, which may inform their efficacy for treating cardiometabolic multimorbidity (**Figure 27**).[50] Results for each drug-target MR analysis are oriented to correspond to the physiological responses of the biomarker to pharmacological modulation in each drug target: the antidiabetic target estimates to the change in risk of the *CM-Factor* associated with lowering of HbA1c or reduced BMI; the lipid-modulating targets estimates to a lowering of LDL-C, triglycerides, or increase in HDL-C (depending on the lipid subfraction used for the exposure); the NAFLD/NASH target to reduced liver fat percentage or alanine aminotransferase (ALT) levels (for HSD17B13, see **Chapter 2 Aim 3 Methods [section 2.4.3]**); and lowered systolic blood pressure (SBP) for the antihypertensives. Overall, drug-target instruments were strong: apart from FGF21, which used a relaxed P-value threshold for instrumentation, all SNPs used as instruments had F-statistics >10) (**Tables AP4.30-AP4.33**) suggesting there is minimal bias due to weak instruments in these analyses.[242]

5.3.1a. Antidiabetics. First, 6 of the 7 antidiabetics tests demonstrated evidence for protective effects against the *CM-Factor* (**Figure 5.4, Table AP4.34**). The MR estimate for lowered HbA1c via INSR (proxying insulin analogs) was directionally consistent with a protective relationship (β=-0.10, [-0.39, 0.19]); however, the confidence interval included the null. The estimates for GIPR and GLP1R, which were modeled proxying both their effects on glycemic markers (with lowered HbA1c) and their central-acting effects on adiposity and appetite (with reduced BMI),[223,304] were consistent between exposures, suggesting beneficial effects on cardiometabolic health through both mechanisms-of-action. Colocalization analyses indicated that HbA1c and the *CM-Factor* shared single causal variants in the *ABCC8/KCNJ11* locus (PP.H4=0.999), *GIPR* (PP.H3=0.997), and *DPP4* (PP.H4=0.65) loci (**Table AP4.35**). Conversely, BMI and the *CM-Factor* demonstrated evidence of colocalization with multiple variants in the *GIPR* (PP.H3=1.0) locus. There was also some evidence of colocalization with multiple variants between the *CM-Factor* and both HbA1c and BMI in the *GLP1R* locus (PP.H3's of 0.58 and 0.53, respectively).

Figure 5.4. Drug-target MR results for antidiabetics targets on the CM-Factor. *Results are reported as effect estimates (β) and corresponding 95% confidence intervals (CIs) from the inverse variance weighted (IVW) methods and aligned to a lowering of HbA1c or reduced BMI. The centers of the error bars are the β's and the error bars are the lower and upper 95% confidence intervals. All cis-instrument MR estimate P-values are based upon used two-sided Wald tests. P-values were not adjusted for multiple comparisons.*

5.3.1b. Lipid modulating targets. Estimates for the 15 lipid-modulating targets generally indicated beneficial effects related to expected directions of the therapeutics being modeled (**Figure 5.5, Table AP4.35**). Apart from HMGCR (the target for statins) and NPC1L1 (the target for ezetimibe), each of the targets for approved therapeutics (PCSK9, ACLY, PPARA, APOC3, and ANGTPL3) indicated protective effects on the *CM-Factor* (the ANGPTL3 estimate was protective but less precise than the study cutoff of P-value < 0.05). The estimate indicated that LDL-C lowering via HMGCR inhibition increased the risk for the *CM-Factor*. As statins and the HMGCR locus have been previously linked with increased T2D risk,[38,433] this signal is likely driven by the T2D component of the *CM-Factor*. Estimates for several of the investigational targets also supported their efficacy for cardiometabolic health. For example, LDL-C lowering via ABCG5/ABCG8, which encodes the sterol transporter G5G8 that is body's main defense against the accumulation of neutral sterols,[434] was strongly protective against *CM-Factor* risk, underscoring the potential of sterol transport as a therapeutically-relevant pathway for cardiometabolic drug development. Similarly, lowered triglycerides levels by both ANPGTL4 inhibition and increased LPL (ANGPTL4 is an inhibitor of LPL, which blocks triglyceride clearance from the plasma, resulting in increased triglycerides levels[435]) provides human genetics evidence supporting its potentials benefits from liver-specific inactivation.[436] Genetically-predicted triglyceride levels by the *ANGPTL4*, *LPL*, and *PPARA* loci each

demonstrated strong evidence of colocalization with the *CM-Factor* (PP.H4's >0.9) (**Table AP4.34**). Colocalization analyses also suggested HDL-C levels and the *CM-Factor* share a causal variant in the *CETP* locus while APOC3 and APOA1 had H4's of 0.57 and 0.55, respectively. Interestingly, several of the LDL-C-lowering targets (HMGCR, ACLY, and ABCG5/8) and the *LPA* locus had H3 posterior probabilities close to 1.0, suggesting that there is more than one causal variant linking the lipid subfractions in these loci, which may be due to there being more than one independent SNP in these loci (**Table AP4.31**).

Figure 5.5. Drug-target MR results for lipid-modulating and NAFLD/NASH targets on the CM-Factor. *Results are reported as effect estimates (β) and corresponding 95% confidence intervals from the inverse variance weighted (IVW) methods. The centers of the error bars are the β's and the error bars are the lower and upper 95% confidence intervals. All cis-instrument MR estimate P-values are based upon used two-sided Wald tests. P-values were not adjusted for multiple comparisons. Lipid-modulating target MR estimates are aligned to the corresponding expected physiological change in the lipid subfraction for that drug class: lowered LDL-C, increased HDL-C, lowered TG, and lowered LP(a) levels, respectively. MR estimates for the 7*

NAFLD/NASH targets instrumented using MRI-derived liver fat percentage data are aligned to reduced liver fat percentage, and HSD17B13 inhibition is proxied using GWAS data for in alanine aminotransferase (ALT) and aligned with reduced ALT levels.

5.3.1c. NAFLD/NASH and antihypertensives. We next investigated the cardiometabolic efficacy of NAFLD/NASH targets (**Figure 5.5, Table AP4.37**). Reduced liver fat percentage through PNPLA3, TM6SF2, and MBOAT7 reduced *CM-Factor* risk while GCKR increased *CM-Factor* risk. THRB, the target for the recently approved THRB partial agonist resmetirom[307] had estimates suggesting beneficial effects related to reduced liver fat; however, the confidence intervals included the null. PNPLA3 and TM6SF2 also had strong evidence of colocalization (PP.H4's=0.994 and 0.942, respectively), supporting the drug-target MR evidence by reducing the likelihood that the MR estimates are confounded by LD patterns.[265] Finally, among the antihypertensives, ACE inhibition and SLC12A3 (the thiazide diuretic target) demonstrated the strongest evidence for a genetics-based impact on *CM-Factor* risk (**Figure 5.6, Table AP4.38**); however, neither colocalized with a single shared causal variant (rather, colocalization indicated multiple causal variants between SBP and the *CM-Factor* in the *ACE* locus [PP.H3=0.99]) (**Table AP4.34**). Several of the individual calcium channels targeted by calcium channel blockers (CCBs) (CANA1C, CACNA1D, CACNA2D2) were linked with reduced *CM-Factor* risk. The overall CCB instrument comprised of all calcium channel targets also had a beneficial effect estimate; however, it was less precise. Like antidiabetics and lipid-modulating targets, NAFLD/NASH candidates and antihypertensives showed little evidence of heterogeneity and were generally consistent across the complementary MR methods, further strengthening causal inference from these results.

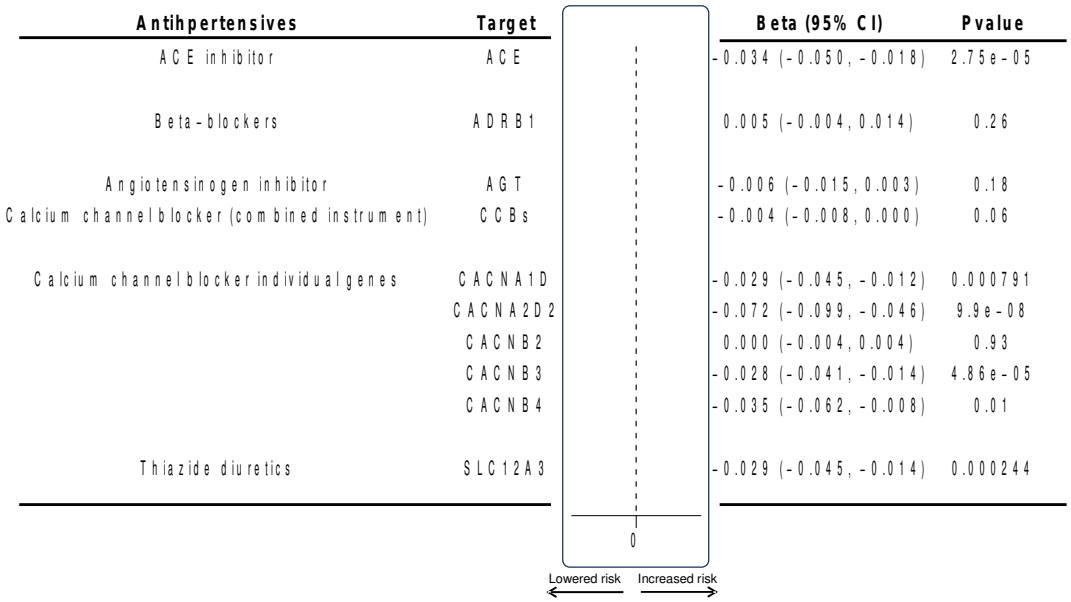


Figure 5.6. Drug-target MR results for antihypertensive targets on the CM-Factor. *Results are reported as effect estimates (beta) and corresponding 95% confidence intervals (CIs) from the inverse variance weighted methods and aligned to a lowering of systolic blood pressure. All cis-instrument MR estimate P-values are based upon used two-sided Wald tests. P-values were not adjusted for multiple comparisons. The CCB combined instrument represents a multi-gene exposure proxying SBP lowering via the 5 individual calcium channel genes, which were also instrumented separately in SBP levels and shown below the combined CCB estimate.*

5.3.2. MR Study 2: MR screen of druggable genes prioritizes novel targets for cardiometabolic health

An overview of these analyses is presented in **Figure 2.8**. SNPs comprising the cis-instruments for each potential drug target were strong with an average F-statistic of 358.16 (range:29.72–37,527.36), suggesting minimal weak instrument bias.[242] Results from the drug-target MR scan of the druggable genes[126] found 91 genes surpassing correction for multiple comparisons (**Figure 5.7a-b**, **Table AP4.39**). Colocalization testing of these 91 genes for evidence of a shared causal variant with the *CM-Factor* highlighted 41 druggable genes with PP.H4 >0.60, including *ACE*, *HMGCR*, and *LPL* expression, which aligns with the MR Study 1 findings assessing the impact of lowered SBP via *ACE*, LDL-C via *HMGCR*, and triglycerides via *LPL*, and additional targets for cardiometabolic health, including 13 novel genes that were not captured by the input univariate GWASs, including *AOAH*, *LAMC1*, and *CDK5R1* (**Tables AP4.40-AP4.42**).

Figure 5.7. Results of the MR screen of the druggable genome (MR Study 2). (a) *outlines MR Study 2 analyses stages and number of targets taken forward from each step.* ***(b)*** *is a Manhattan plot of the cis-instrument MR screen of the gene expression (whole blood) with the gene coordinates on the x-axis and* $-log_{10}$ *P-values for the MR estimates on the y-axis. All cis-instrument MR estimate P-values are based upon used two-sided Wald tests. P-values were not adjusted for multiple comparisons. Labeled genes are the targets surpassing correction for multiple comparisons in the cis-instrument MR screen that also had evidence of colocalization (PP.H4 >0.6) and had relationships with established cardiometabolic biomarkers. The red dashed line is the cis-instrument MR screen Bonferroni-corrected P-value threshold of* 1.96×10^{-5} *(0.05/2,546 druggable genes tested).* ***(c)*** *is a mediation plot of the MR estimates (β's and 95% confidence intervals [CIs]) relating CRY2 gene expression, fasting glucose, and the CM-Factor.* ***(d)*** *is the phenome-wide MR screen of CRY2 expression on 366 biomarkers and diseases. The labeled traits are those that surpassed Bonferroni correction for multiple comparisons (P-value=*1.37×10^{-4}*; 0.05/366), which is indicated by the red dashed line.*

5.3.2a. Tissue-level replication. We aimed to incorporate replication analyses using gene expression instrumented in disease-relevant tissues. 36 of the 41 druggable genes were available for instrumentation (**Table AP4.43**) and we were able to perform a total of 186 replication tests across all tissues. 30 of the 36 genes assessed replicated at P-value < 0.05, while 21 had estimates that surpassed the more stringent threshold adjusting for the 186 replication tests (P-value < 2.69×10^{-4}) (**Table AP4.44**). For genes available in multiple tissues, we did observe some tissue-specificity, e.g., *CAMK1D* was analyzed in 9 tissues, but only replicated in the tibial artery, heart left ventricle, and whole blood (**Table AP4.44**); *AOAH* (one of the druggable genes considered novel for the *CM-Factor*) was analyzed in 11 tissues, but only replicated in subcutaneous adipose and tibial artery tissues; *CRY2* replicated in whole blood but not aortic tissue. The direction of the estimates for the druggable genes that successfully replicated generally aligned across each tissue and the eQTLGen[437] whole blood gene expression data used for the main screen, e.g., increased *AOAH* expression associated with reduced *CM-Factor* risk in the initial screen and in the subcutaneous adipose and tibial artery tissues, and increased CAMK1D expression associated with increased *CM-Factor* risk in the initial screen and in whole blood, tibial artery, and heart tissues (both atrial appendage and left ventricle) (**Table AP4.45**). 9 of the 21 targets also demonstrated evidence of colocalization at PP.H4 >0.70, further strengthening causal inference (**Table AP4.46**).

5.3.2b. Biomarker mediation. 28 of the 41 druggable genes were associated with established cardiometabolic biomarkers listed in the GWAS Catalog and Oxford Biobank,[322] some with only a few biomarkers, e.g., *CRY2* with fasting glucose, *LAMC1* with blood pressure and major lipid subfractions, and *CDK5R1* with systolic blood pressure; others, conversely, with an array of biomarkers, e.g., *GPBAR1* and *OPRL1* were associated with a range of metabolic syndrome traits, including blood pressure, lipid subfractions, glycemic traits, and measures of adiposity (**Tables AP4.47**-**AP4.48**).

Many of these associations were confirmed with drug-target MR analyses of the druggable gene onto the biomarkers (**Table AP4.49**). For example, *CRY2* whole blood gene expression was associated with fasting glucose levels (β=0.051, [0.038, 0.064], P-value=5.99×10^{-15}); *LAMC1*, with levels of LDL-C, HDL-C, apolipoprotein B, apolipoprotein A1, and LP(a); *CDK5R1* with SBP; and *OPRL1*, with metabolic syndrome traits (**Table AP4.49**). These druggable-gene biomarker associations motivated further analyses to assess the extent to which the total impact of the druggable gene on the *CM-Factor* was mediated by these established biomarkers (**Table AP4.50**). Fasting glucose levels mediated 81.1% of the total impact of *CRY2* on the *CM-Factor*; fasting insulin mediated 93.1% of *GPBAR1*'s impact (**Figure 5.7c**); while the impact of *OPRL1* expression was mediated by several biomarkers, including SBP (19.3%), circulating ALT levels (13.7%), and fasting glucose (12.5%).

5.3.2c. Side-effect profiling for druggable genes. Phenome-wide MR (PheMR) analyses for the 41 genes characterized their potential side-effect profiles and further clarified their therapeutic potential. Broadly, we observed relationships surpassing correction for multiple comparisons for each of the 41 genes, suggesting pleiotropic effects (**Figure 5.8**, **Figures AP4.13**-**AP4.14, Tables AP4.51**-**AP4.53**). For several of the genes, the direction of the MR estimate aligned with the therapeutically relevant direction for the *CM-Factor* indicated by the initial drug-target MR screen. For example, increased *OPRL1* expression was related to

increased *CM-Factor* risk and other cardiometabolic diseases, including angina and myocardial infarction. It was also associated with reduced cognitive performance and increased laryngeal cancer risk. *CGREF1* expression had a similarly neutral phenotypic profile. In contrast, other genes generally had favorable phenotypic profiles, with exceptions suggesting caution for therapeutic applications. For instance, lowered *CRY2* expression was linked to increased gallstone risk (**Figure 5.8d**); lowered *CDK5R1* expression was associated with increased schizophrenia risk; lowered *GPBAR1* expression correlated with increased cholelithiasis risk; and increased *LAMC1* expression was associated with heightened colorectal cancer risk. However, given the challenges in translating the magnitude of MR estimates—which reflect lifelong relationships of the gene—from short-term therapeutic modulation,[50] it remains unknown whether these side effects would outweigh the cardiometabolic benefits.

Figure 5.8. Bar plot of the phenome-wide Mendelian randomization (MR) results for the 41 genes prioritized by the MR Study 2 screen of the druggable genome. *Plotted are counts of the total number of traits that were impacted by gene expression of druggable gene (X-axis) that surpassed the phenome-wide MR P-value (two-sided tests) adjustment for multiple comparisons. The "adverse effect" count indicated by the blue columns are the traits with MR estimate direction corresponding to the opposite of the direction that was therapeutically indicated by the MR estimate on the CM-Factor. Similarly, "the beneficial" effect indicated by the orange columns are the traits with MR estimate direction that aligned with of the direction that was therapeutically indicated by the MR estimate on the CM-Factor.*

5.3.3. MR Study 3: Two-step MR prioritizes ENO3 as a proteomic mediator of the obesity-*CM-Factor* relationship

We next performed a two-step MR study[132] (**Figure 2.10**) to identify proteomic mediators of the impact of obesity (modeled with BMI) on the *CM-Factor*. First, we confirmed that genetically predicted BMI increased the risk for the CM-Factor (**Figure 5.9**). We next analyzed the impact of BMI on the circulating proteome, finding 506 proteins whose levels are influenced by BMI (as indicated by their MR estimates surpassing correction for multiple comparisons (P-values < 1.76×10^{-5} [0.05/2,835 proteins tested in Step 1]) (**Figures 5.10a-5.10b**). We did not find heterogeneity for any of the BMI-influenced proteins (all Cochran's Q P-values > 0.05), and all analyses passed the Steiger directionality test (**Table AP4.54**). Broadly, MR estimates from the complementary MR methods (weighted median, weighted mode, and MR-Egger) were directionally consistent with the main inverse variance weighted estimates, including for the proteins prioritized by Step 2.

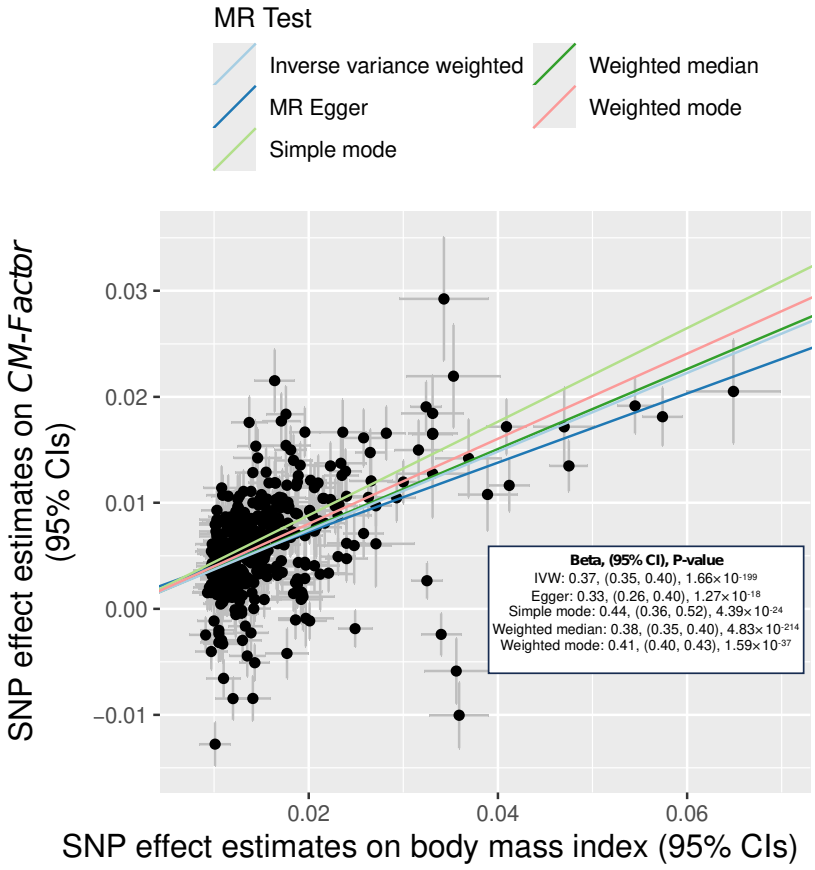


Figure 5.9. SNP-SNP plot of the single-variable MR estimate of BMI on CM-Factor. *Single variable MR was performed using conventional MR methods (the same methods used to assess the impact of BMI on the proteome in Step 1. All cis-instrument MR estimate P-values are based upon used two-sided Wald tests. P-values were not adjusted for multiple comparisons. Each*

point represents the GWAS association statistic for BMI (x-axis) and the CM-Factor (y-axis) with the corresponding 95% confidence intervals (CIs). The box in the main body of the plot presents the results for each MR method plotted.

In Step 2, we used drug-target MR to assess the causal impact of the BMI-influenced proteins on the *CM-Factor*. We focused Step 2 analyses on proteins with cis-pQTLs to increase the likelihood that the genetic variants are relevant to the protein being studied and reduce the chance of bias due to horizontal pleitropy.[50] Using the same cis-instrumentation selection procedure as outlined in the above sections for approved and investigational therapeutics, we were able construct instrumental variables for 455 of the 506 proteins identified in Step 1. The instruments for the proteins were strong (all F-statistics >10), reducing the risk of weak instrument bias.[242] MR estimates for 8 proteins (labeled proteins in **Figure 5.9b**) surpassed correction for multiple comparisons, and Steiger directionality tests, used to assess potential bias from reverse causation, supported a causal role of these proteins on the *CM-Factor* (**Table AP4.53**). 4 of these proteins—enolase 3 (ENO3), sex hormone binding globulin (SHBG), protein tyrosine phosphatase receptor type R (PTPRR), and gamma-glutamyltransferase 1 (GGT1)—had estimates for steps 1 and 2 that were directionally concordant with direction of the single variable MR, demonstrating that BMI increases the risk of the *CM-Factor* (β=0.137, CI=[0.35, 0.40], P-value=1.66×10^{-199}) (**Figure 5.10c**, **Figure AP4.27**, **Table AP4.56**).

Figure 5.10. Results of the two-step MR characterizing the proteomic pathways between BMI and the CM-Factor. (a) *outlines MR Study 3 (two-step MR) analyses stages and number of proteins taken forward from each step.* ***(b)*** *is a volcano plot of the Step 1 proteome-wide MR identifying proteins whose levels are impacted by BMI with the Z score for the MR estimates on the x-axis and the* $-log_{10}$ *P-values for the MR estimates on the y-axis. Yellow points are the proteins surpassing correction for multiple comparisons, and labeled proteins are those that were prioritized further by Step 2 and the other downstream analyses.* ***(c)*** *are forest plots for the MR estimates (β's and 95% confidence intervals [CIs]) of the eight proteins prioritized by Step 1 and Step 2. The centers of the error bars are the β's and the error bars are the lower and upper 95% CIs. All cis-instrument MR estimate P-values are based upon used two-sided Wald tests. P-values were not adjusted for multiple comparisons. Step 1 estimates are oriented to the impact on the circulating protein levels from increased BMI, and Step 2 estimates are oriented to the impact of increased proteins levels on the CM-Factor. '*' indicates the four proteins with Step 1 and Step 2 estimates that aligned with the adverse impact of increased BMI on the CM-Factor.* ***(d)*** *depicts the locus association plot for the colocalization results for the ENO3 locus between ENO3 protein levels measured in the UK Biobank and the CM-Factor. The x-axis represents the genomic coordinates for each ENO3 locus SNP included in the colocalization results, and the y-axes represents the* $-log_{10}$ *P-values for the SNP GWAS associations for ENO3 levels (lower panel) and the CM-Factor (upper panel).*

5.3.3a. Replication and triangulation of BMI-associated proteins.

We next aimed to replicate the two-step MR findings for the 8 proteins using the independent deCODE proteomics dataset from Ferkingstad et al.[438] NUCB3 and GGT1 were not present in the deCODE data. In Step 1 replication, all estimates were directionally consistent with the primary data and had P-values < 0.05 (**Table AP4.57**). The estimate for ENO3 surpassed the same stringent Bonferroni correction from the primary analyses. Estimates from additional replication using obesity diagnoses in the FinnGen cohort (27,711 cases/425,881 controls)[269] as the exposure were similarly consistent—only NUCB3 failed to replicate with the FinnGen obesity exposure. Step 2 replication was similarly consistent, with 4 of the 6 proteins (ENO3, FSTL3, MSR1, and SHBG) demonstrating evidence for an impact on the *CM-Factor* (**Table AP4.58**). Step 2 replication estimates for ENO3 (β=0.024, CI=[0.017, 0.031], P-value=6.57×10^{-11}) and SHBG (β=-0.048, CI=[-0.061, -0.034], P-value=3.63×10^{-12}) would have also surpassed same stringent Bonferroni correction from Step 2. For the 5 proteins available in the observational INTERVAL study data, 4 were impacted by BMI in the same direction as the Step 1 results at P-value < 0.05. The association between BMI and CD34 levels was directionally consistent between our Step 1 findings and the observational data (increase BMI reduced CD34 levels); however, the observational association confidence interval spanned the null (P-value=0.36) (**Table AP4.59**).

5.3.3b. Mediation of BMI-associated proteins.

We assessed the extent to which these proteins were involved in the BMI-*CM-Factor* relationship by performing mediation analyses using the product of coefficients method (**Figure 2.9**), which showed that the 4 proteins with MR estimates consistent with an increased risk of the *CM-Factor* from BMI mediated between 1.0% (GGT1) and 9.8% (PTPRR) of the BMI effect (**Table AP4.56**). SHBG and ENO3 mediated 1.3% and 2.6%, respectively. The colocalization analyses indicated that ENO3 and SHBG each had high posterior probabilities of a shared causal variant with the *CM-Factor*

(PP.H4's of 0.75, and 0.86, respectively) (**Table AP4.60**). These proteins also colocalized using the deCODE data used as independent replication (ENO3 PP.H4=0.9 and SHBG PP.H4=0.96). GGT1 and PTPRR each demonstrated strong evidence of colocalization but with multiple variants (PP.H3=0.73 and 0.99, respectively). CD34 colocalized (PP.H4=0.96) in the primary UKB-PPP data, but not in the deCODE data (**Figures AP4.15-AP4.16**).

5.3.3c. BMI-associated protein side-effect profiles. Finally, we performed Phe-MRs for the 4 proteomic mediators with MR estimates consistent with the adverse impact of obesity on the *CM-Factor:* ENO3, PTPRR, SHBG, and GGT1, to characterize potential side effects. For ENO3, PTPRR, and GGT1, Phe-MR results suggested generally favorable side-effect profiles, with most estimates indicating beneficial effects from pharmacological modulation of the proteins in in the direction indicated by the two-step MR to be therapeutically relevant (**Figure 5.11**, **Figures AP4.17-AP4.19**, **Tables AP4.62-AP4.64**). The profiles for ENO3 and PTPRR were particularly neutral with only a single trait surpassing correction for multiple comparisons for ENO3 – platelet width (**Figure 5.11, Table AP4.61**). By contrast, SHBG demonstrated effects on 22 traits, including several potential adverse effects on blood pressure, myasthenia gravis, and arrhythmia (**Figure AP4.17**, **Table AP4.63**), reflecting the complex role SHBG plays in many physiological processes.[439] As expected, the top findings for GGT1 proteins levels were related to liver function enzymes (GGT levels and alanine aminotransferase levels), and the other traits surpassing correction for multiple comparisons included calcium levels, appendicitis, and acute pancreatitis (in the adverse direction) (**Table AP4.64**).

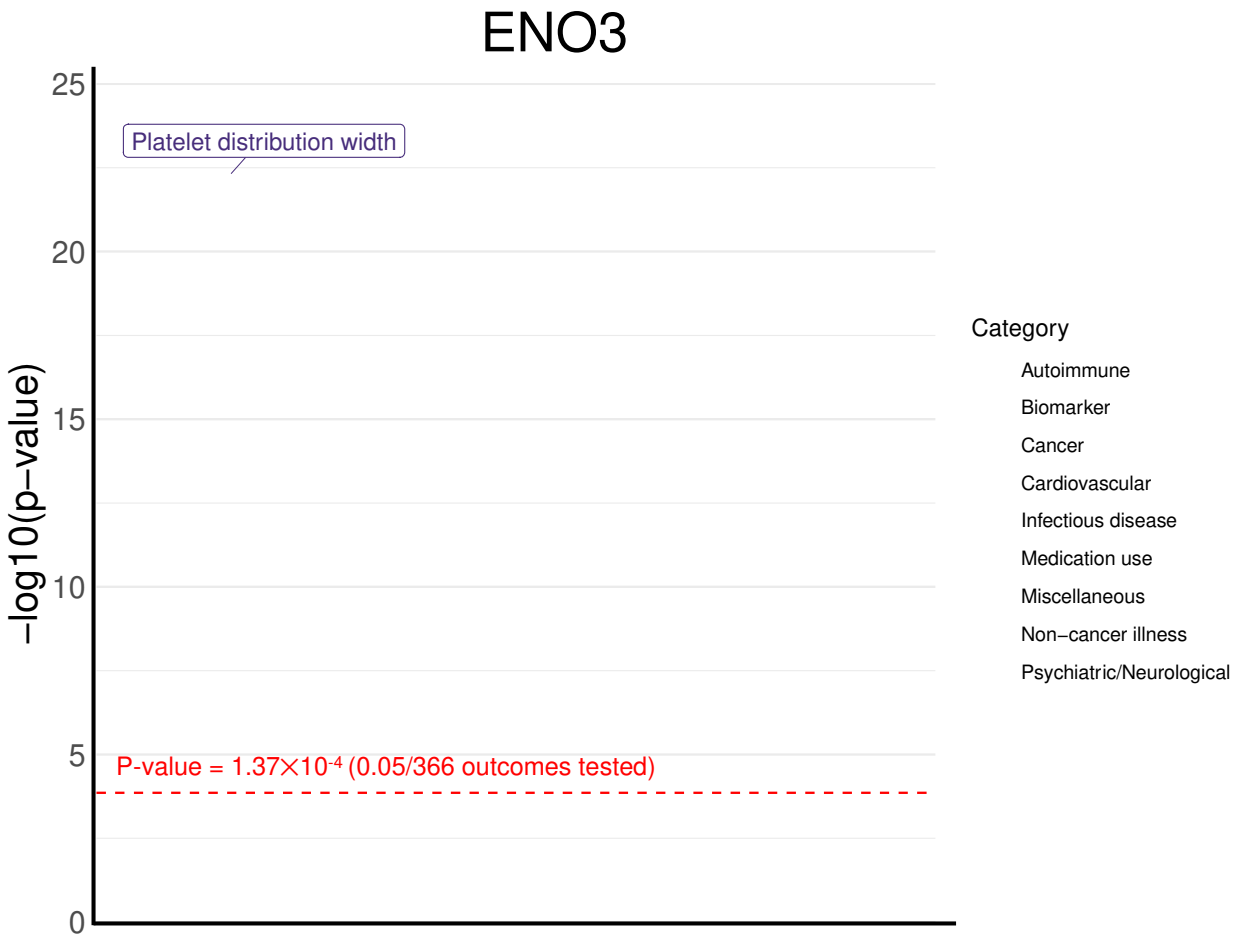


Figure 5.11. Phe-MR results of circulating ENO3 protein levels (UKB-PPP). *Results are the −log10 P-values for the main drug-target MR estimates (either inverse variance weighted or Wald ratios) for the ENO3 protein on 366 outcomes curated for the Phe-MR analysis. The outcomes are grouped by clinical category. All cis-instrument MR estimate P-values are based upon used two-sided Wald tests. P-values were not adjusted for multiple comparisons. The red dashed lined is the Bonferroni-corrected P-value threshold (1.37×10^{-4} [0.05/366 outcomes]), and the labeled outcomes are those that surpassed the Bonferroni correction for multiple comparisons.*

5.4. Aim 3 Discussion

Aim 3 used recently developed multivariate GWAS methods that leverage genetic correlations among correlated univariate GWASs to investigate the genetic architecture of NAFLD, T2D, and CAD. Multivariate methods have been shown to increase statistical power and boost loci discovery[16] and our *CM-Factor* GWAS identified 523 lead variants in 312 genomic loci shared across NAFLD, T2D, and CAD. Given that the average cardiometabolic patient is aged 65+ and presents with 2 or more cardiometabolic diseases,[14,440] these results underscore the importance of research beginning to elucidate the biological mechanism linking these multimorbid cardiometabolic diseases.[115,441,442]

CM-Factor SNPs showed strong concordance in effect directions across NAFLD, T2D, and CAD, supporting the idea that the *CM-Factor* represents a shared genetic risk for cardiometabolic diseases. Discordances generally occurred in SNPs with weaker associations. Comparison of lead SNPs identified by the *CM-Factor* with the genetic signatures from NAFLD, T2D, and CAD revealed that the *CM-Factor* identified novel loci for each of the diseases, including substantial novelty for NAFLD (514 novel loci) and CAD (415 novel loci). Many of these novel loci for NAFLD and CAD showed some statistical association in the univariate GWASs, indicating their potential to be detected in larger studies. Given the relatively small sample sizes of current NAFLD GWASs compared to those for CAD and T2D GWASs, this gain in statistical power with the GenomicSEM method represents an important opportunity for loci prioritization for NAFLD. Additionally, while the T2D genetics was strongly represented in the shared genetic architecture, approximately 50% of the T2D loci identified by Suzuki et al.[276] were not prioritized, suggesting the *CM-Factor* effectively captures key T2D loci while also clarifying specific variants with broader relevance to cardiometabolic health, highlighting its utility in multi-trait genetic analyses.

The SNP-level heterogeneity testing provided further insight into the shared cardiometabolic architecture by identifying 151 heterogeneous loci that likely influence each of the individual cardiometabolic diseases through distinct pathways, or had directionally inconsistent effect directions (i.e., increasing risk for one cardiometabolic disease while decreasing the risk for the others), and not the broad, directionally concordant shared architecture linking them. Several of the heterogenous loci are canonical cardiometabolic loci, and this analysis has contextualized and clarified their relationships within the larger cardiometabolic disease landscape with a statistically stringent and hypothesis-free approach. For example, *PNPLA3*, a canonical NALFD locus,[443] was genome-wide significant in the initial multivariate GWAS of the *CM-Factor* (*CM-Factor* P-value= 5.74×10^{-15}). However, the SNP-level heterogeneity testing flagged it as a highly heterogenous locus. As expected, each variant within the locus was strongly associated with the meta-analyzed NAFLD + liver fat outcome used to construct the *CM-Factor* (P-values from 5.84×10^{-17} to 6.84×10^{-53}). PNPLA3 catalyzes the hydrolysis of hepatic triglycerides, and impaired *PNPLA3* function is major genetic risk factor for all stages of liver disease. For example, homozygous carriers of the I148M variant in *PNPLA3* show elevated liver fat content due to accumulated lipid content, as well as increased risk for fibrosis and hepatocellular carcinoma.[443] However, because impaired *PNPLA3* function sequesters lipids in the liver, increasing NAFLD risk, it may also be associated with lower circulating lipid and cholesterol levels, reducing CAD risk, and variants in the *PNPLA3* locus have been linked to reduced CAD risk.[427] Consistent with this, several of the *PNPLA3* variants identified as heterogenous for the *CM-Factor* have been linked with increased liver fat content, but lower circulating lipids (e.g., rs9626056 is linked with lower total cholesterol, LDL-C, apolipoprotein B, etc.).[201] Our heterogeneity testing also flagged several well-known loci in CAD, e.g., *LDLR* and *CELSR2-PSRC1-MYBPHL-SORT1*, suggesting these components of lipid metabolism and signaling may impact CAD rather than directly impact T2D or NAFLD, providing additional insight into how these canonical loci influence CAD within the broader context of cardiometabolic diseases.

Our bioinformatic characterization investigated the *CM-Factor* genetics at the variant, gene, tissue, cell, and pathway levels. Comparison of these results with the corresponding underlying

univariate NAFLD, T2D, and CAD GWASs facilitated identification of biological dimensions unique to the *CM-Factor* as well as those distinct to the individual cardiometabolic diseases. Among the 243 high-confidence genes prioritized by the TWAS were established cardiometabolic genes (e.g., *IRS1, LPL*) and novel genes not captured by the genetic signatures of NAFLD, T2D, or CAD. One such novel gene was *COMT* (Catechol-O-Methyltransferase), which plays a critical role in the metabolism of catecholamines, mediating the cardiometabolic impact of the sympathetic nervous system.[444] Early genetics-based studies linked the *COMT* Val158Met polymorphism, the primary cause of variation in *COMT* activity, with increased SBP, waist-to-hip ratio, and acute coronary events.[444] Preclinical studies further support the link between *COMT* and blood pressure: compared with Wistar-Kyoto rats, spontaneously hypertensive rats exhibit lower hepatic *COMT* expression and impaired methylation of catecholamines.[445]

The other novel targets have less literature support, but nevertheless may reflect novel biological targets in cardiometabolic health. For example, *DHX36*, a DEAH-box helicase involved in nucleic acid binding and processing,[446] could influence cardiometabolic risk through its role in RNA metabolism and gene expression regulation. Research suggests that *DHX36* may impact pathways related to inflammation, including the innate immune response,[447] which has been implicated in cardiometabolic disease.[448] By modulating the stability and translation of mRNAs encoding proteins critical to immune processes, *DHX36* could indirectly affect the development of conditions such as atherosclerosis, insulin resistance, and obesity. Furthermore, alterations in *DHX36* activity or expression levels could disrupt normal cellular functions, leading to an imbalance in metabolic homeostasis, thereby increasing the risk of developing cardiometabolic diseases. *ASPRV1* (Aspartic Peptidase Retroviral Like 1), another novel *CM-Factor* gene, is implicated in the immune response through its role in modulating cytokine production and immune cell activation.[449] These processes are crucial in the pathogenesis of atherosclerosis and insulin resistance,[450] suggesting new inflammatory mechanisms underlying cardiometabolic health. Future research will be needed to further elucidate the mechanisms linking *COMT*, *DHX36*, *ASPRV1*, and other novel genes to cardiometabolic health.

5.4.1. Drug-target MR assesses cardiometabolic efficacy of approved and investigational therapies and prioritizes targets for therapeutic discovery efforts

Drug-target MR is an important component of the drug discovery paradigm in cardiometabolic disease.[6] We included several applications of the drug-target MR framework motivated by the current stagnation in the development of new therapeutics for cardiometabolic disease relative to other specialities,[12] a stagnation attributed to several factors, including a still limited understanding of the underlying pathophysiology resulting from inappropriate disease models, poor patient characterization,[14,451] and drug candidates targeting non-causal biomarkers.[13]

We first conducted drug-target MR assessing the general cardiometabolic efficacy of 34 approved and investigational cardiometabolic targets and found protective roles of many of the targets through their primary expected physiological responses to pharmacological modulation,

including recently approved therapeutics such as genetically mimicked *GLP1R* and *GIPR* agonists, *PCSK9* inhibitors, and fibrates (*PPAR-alpha* agonists) along with investigational targets like *ANGPTL4*, *PNPLA3*, and *TM6SF2*. These findings will help improve understanding of their broader efficacy in these multimorbid patient populations. For some of these targets, these results may also help inform ongoing concerns regarding side effects. For example, genetically proxied *PCSK9* inhibition has been previously linked with increased T2D risk.[39,40,52,53] However, here we have shown that by combining the genetics of CAD with T2D and NAFLD, we still find a protective impact of *PCSK9* inhibition, suggesting that long-term *PCSK9* inhibition will be beneficial in patients with CAD and one or more cardiometabolic comorbidities.

By contrast, we did not find protective relationships of *HMGCR* inhibition with the *CM-Factor*, which is likely due to the adverse effects of *HMGCR* inhibition on T2D, possibly through mechanisms such as increased body weight or insulin resistance.[452] We emphasize that this potential increase in the risk of T2D does not offset the lipid-lowering benefits of statin therapy. For *GLP1R* and *GIPR*, we found evidence supporting protective roles of their genetic agonism, likely reflecting their beneficial effects on pancreatic cells and glucose homeostasis[304] as well as central nervous system mechanisms that reduce appetite and body weight.[223] Although HbA1c and BMI did not colocalize with the *CM-Factor* in the *GLP1R* locus, fine mapping of the *CM-Factor* provided strong evidence for causal variants in the *GLP1R* locus. Similar strong evidence was also observed in loci associated with sulfonylureas, TZDs, and APOC1 inhibitors, further strengthening causal inference for the drug-target effects.

In the Aim 3 MR Study 2, we extended the drug-target MR framework to inform target discovery across the druggable genome,[126] aiming to guide future research into therapies for reducing the risk of cardiometabolic diseases. Leveraging this approach, we prioritized 41 targets from over 2,500 included in the initial screen, including 13 that were novel genes, including *AOAH*, *LAMC1*, and *CDK5R1*, for these cardiometabolic diseases. *AOAH* (acyloxyacyl hydrolase) plays a critical role in modulating inflammatory pathways by detoxifying lipopolysaccharide (LPS) and hydrolyzing acyl chains, thereby reducing systemic inflammation,[453] suggesting a potential to mitigate chronic inflammation, a major driver of cardiovascular pathology, including atherosclerosis and myocardial remodeling.[454] Further, *AOAH* has broader enzymatic activity, including phospholipase functions,[453] suggesting it may regulate lipid metabolism and oxidative stress, key factors implicated in cardiometabolic disease. By attenuating inflammation and promoting lipid homeostasis, *AOAH* may provide a promising pathway for the development of interventions addressing cardiometabolic risks.

Many of the prioritized druggable genes are implicated in pathways not yet targeted by existing cardiometabolic therapeutics, despite existing lines of evidence for their involvement in cardiometabolic diseases, reflecting potential new therapeutic strategies. For instance, *CRY2* is a gene involved in circadian clock regulation, which is important for physiological homeostasis,[455] and in animal models, *CRY2* modulation impacts metabolic homeostasis, including glucose metabolism and liver fat accumulation,[456,457] while in cell-based assays, small molecular modulators of *CRY2* inhibited glucagon-induced gluconeogenesis in primary hepatocytes,[458] suggesting that targeting these circadian rhythm-related regulators is a potential therapeutic approach for cardiometabolic disease.

Our screen of relevant biomarkers provided further clarification of potential pathways for *CRY2*. We found that the impact of *CRY2* expression on the *CM-Factor* is likely mediated by fasting glucose, but not by insulin or blood pressure (two other biomarkers previously linked to *CRY2* variants). Overall, the *CRY2* results provide human genetics support for the literature demonstrating that circadian rhythms influence plasma glucose concentration (potentially by autonomous circadian rhythms in relevant tissues like pancreatic islet cells[459]). Given that misalignment of endogenous cellular biological clocks, i.e., circadian disruption, may result in dysregulation of glucose homeostasis and T2D,[460] these findings together suggest that the core circadian clock may represent an important therapeutic domain for cardiometabolic drug developments.

Shifting focus from circadian regulation to the opioid system, *OPRL1* (nociceptin/orphanin FQ receptor [NOP receptor]) emerged as another promising target. *OPRL1* plays a crucial role in the endogenous opioid system, which, while predominantly studied for its role in pain, depression, and substance abuse, has also been implicated in cardiometabolic health.[461] This includes its involvement in the structure and function of cardiometabolic tissues (e.g., cryoprotection of the myocardium[462]) and its role in modulating obesity through feeding-related behaviors.[463] Together, these functions make the opioid system an attractive and novel avenue for therapeutic modulation.[462]

Our results indicated that antagonism of *OPRL1* would be the relevant direction for therapeutic benefit for the *CM-Factor*. Several small-molecule *OPRL1* antagonists have already been developed for other diseases, such as LY2940094 by Eli Lilly, which completed phase 2 clinical trials for depression and alcohol use disorder.[463] Given the clinical overlap and shared biological underpinnings between cardiometabolic health and depression[464]—such as dysregulation of the hypothalamic-pituitary-adrenal (HPA) axis,[465] which is partially modulated by *OPRL1*[463]—these *OPRL1* antagonists may represent repurposing opportunities capable of providing therapeutic efficacy for psychiatric and cardiometabolic comorbidities. As several pharmacological agents targeting *OPRL1* (but not LY2940094) have been hindered by side effects,[463] it is important that our Phe-MR analysis found genetic inhibition of *OPRL1* to have a safe side effect profile, suggesting that safety issues may relate to the specific pharmacological agents rather than pharmacological modulation of *OPRL1* itself. These findings highlight *OPRL1* as a promising avenue for future research and drug development, offering new strategies to address the complex mechanisms underlying these conditions.

In MR Study 3, we integrated two-step MR—leveraging polygenic MR, drug-target MR, and the plasma proteome—colocalization, and sensitivity analyses to screen proteomic mediators, aiming to further our mechanistic understanding of the relationship between obesity and cardiometabolic diseases. We prioritized four proteins—*ENO3*, *GGT1*, *SHBG*, and *PTPRR*—that showed estimates at each step consistent with an adverse impact of increased BMI on the *CM-Factor*. Given that BMI is a highly polygenic trait with over 530 associated loci and influences approximately 490 circulating proteins in the UKB-PPP data, it is notable that these four proteins could account for a significant portion of the adverse impact of BMI and other obesity-related traits on the *CM-Factor*. *ENO3* (enolase 3) encodes the homodimer of the subunit β-enolase, responsible for the penultimate step in glycolysis (the conversion of 2-phosphoglycerate to phosphoenolpyruvate).[466] *ENO3* plays a role in striated muscle development and regeneration[466]

and has also been implicated in NAFLD, potentially through dysregulated cholesterol ester accumulation, accelerating NASH progression by negatively regulating hepatic ferroptosis.[467] While we could not assess observational evidence of BMI and *ENO3* due to its absence in the INTERVAL study, previous research has shown that *ENO3* levels are elevated in insulin resistance and obesity.[468] Our Phe-MR analysis indicated a neutral side effect profile for genetically modulating *ENO3* protein levels, suggesting that it may represent a key target for coordinating responses to obesity and metabolic dysregulation that is amenable to pharmacological intervention. This finding warrants further investigation and validation. We have focused our discussion on *ENO3* because it colocalized, replicated, and demonstrated a neutral on-target safety profile. However, the other prioritized proteins also have existing evidence supporting their roles in metabolism (e.g., *PTPRR* in insulin signaling and secretion[468]). See **Appendix 4** (**Section 4.3.1**) for an extended discussion of these proteins.

5.5. Aim 3 Limitations

We acknowledge several limitations of our study. First, like standard GWAS methods, GenomicSEM assumes an additive model for common genetic variants,[16] which may not capture the effects of rare variants with large impacts on cardiometabolic health.[469-473] Additionally, the etiology of NAFLD, T2D, and CAD reflects complex interactions between environmental and genetic factors,[474-476] including the induction of epigenetic changes,[477] which cannot be fully modeled with the *CM-Factor*. Consequently, future studies incorporating interaction analyses and exploring the shared genetics of NAFLD, T2D, and CAD will be essential to elucidate these etiological relationships.

Second, another limitation in our study is the potential confounding in the input GWASs due to the high epidemiological overlap between traits, particularly T2D and NAFLD. Given that approximately 65% of patients with T2D also have NAFLD, the univariate GWAS results may partially reflect shared comorbidity rather than fully independent genetic influences for each trait. This overlap could contribute to the inflated genetic correlations between traits, observed in our LDSC analyses. Although the GenomicSEM multivariate framework is designed to manage overlapping genetic architectures, this confounding factor should still be considered when interpreting the findings.[16] Furthermore, while our multivariate model identifies distinct shared genetic components beyond those captured by each individual GWAS, unmeasured trait overlap may limit our ability to fully disentangle the unique genetic contributions of T2D, NAFLD, and CAD. Future research employing additional stratification or covariate adjustment in GWAS analyses could help address this limitation and further refine the identification of unique genetic influences.

Third, in our *CM-Factor* multivariate GWAS, we used a recent meta-analysis NAFLD GWAS with a sample prevalence of ~1%,[107] which is substantially lower than the estimated NAFLD global prevalence of 25%.[107,108] While the UK Biobank has been shown to represent a healthier population compared to the general population,[423] this discrepancy raises concerns about the accuracy of control classification and suggests a potential underdiagnosis of NAFLD among the controls. Additionally, the effective sample and case count for the NAFLD GWAS was substantially lower than those for T2D or CAD, potentially affecting the power and precision of our *CM-Factor* and increasing the risk of type 1 errors.[478] Although GenomicSEM has been

shown to be robust to both sample overlap and sample size imbalance,[84] we emphasize the need to reproduce these analyses as larger and more representative NAFLD data becomes available. To mitigate potential control mis-classification and sample imbalance issues, we incorporated data related to liver fat content based upon MRI, which, because it is a highly sensitive and non-invasive method for quantifying liver fat, provides a more precise and direct assessment of liver steatosis than ICD codes.[479,480] However, there remains the potential for residual misclassification and underrepresentation of NAFLD genetics, particularly in individuals with mild steatosis undetected by current thresholds, which could impact the power and precision of our GWAS findings, highlighting the importance of larger, more accurate datasets in future research.

Fourth, while the prevalence and burden of NAFLD are growing worldwide,[110,111] there still exists substantial heterogeneity in the prevalence of NAFLD—as well as T2D and CAD—among populations of different ancestries, likely due to variations in diet, lifestyle, comorbid conditions, and genetic predispositions.[481-485] For instance, in the United States, studies indicate a higher prevalence of NAFLD in Hispanic populations, linked to genetic factors such as the *PNPLA3* variant, and a lower prevalence in Black populations, potentially due to differences in fat distribution and other metabolic factors.[486] Similar variability is observed globally. For example, Uyghur individuals in China have a higher prevalence of NAFLD than other ethnic groups, and studies in multiethnic settings in Europe and Asia have highlighted differences in NAFLD susceptibility related to both genetic and lifestyle factors.[486] We were unable to perform replication analyses in non-European populations due to the lack of availability of NAFLD diagnoses or liver fat in non-European ancestry cohorts. As a result, we caution before extension of these findings to non-European ancestry populations, and we emphasize the need for future work to investigate the shared genetics of NAFLD, T2D, and CAD across diverse populations. **Appendix 4** elaborates on the challenges of limited non-European participant data in genetics-based studies.

The variant and gene prioritization methods used in this study also have important limitations. For instance, transcriptomic imputation may mis-prioritize genes due to eQTL sharing, tissue-specific expression variability, and linkage disequilibrium confounding.[487] Predictive models often assume linear relationships and may overlook complex regulatory mechanisms, potentially leading to inaccuracies. Moreover, environmental influences, temporal changes in gene expression, and the limited resolution for pinpointing causal variants further complicate the interpretation of findings[487] (**Appendix 4; Section 4.3.2**). Similarly, there are limitations when interpreting the results of drug-target MR analyses. Specifically, drug-target MR using downstream biomarker data cannot fully mimic potential heterogeneity in drug responses associated with tissue-specificity or mechanisms of action for specific therapeutics. For example, inclisiran, a recently approved small interfering RNA inhibitor of hepatic *PCSK9* expression,[8] may exert liver-specific effects that are not captured by our drug-target MR estimates, which rely on *PCSK9* variants derived from circulating LDL-C levels. This limitation highlights the need for caution when extrapolating findings to therapeutic interventions with highly specific mechanisms of action.

More broadly, causal inference requires triangulation of study designs,[354,488] as well as replication and clinical validation. Future studies involving patients with cardiometabolic comorbidities will be essential to test the precision and robustness of our drug-target MR findings. Nonetheless, our

findings support and extend previous genetics-based studies that have validated both approved and investigational therapeutic targets for a range of cardiometabolic outcomes.

5.6. Aim 1 Conclusion

The Aim 3 project integrated genetic data for NAFLD, T2D, and CAD into a multivariate GWAS framework to model the broad genetic architecture shared across these multimorbid diseases, facilitating the discovery and prioritization of genomic loci. Heterogeneity testing provided further insights into whether the identified loci modeled shared or distinct cardiometabolic pathways. Gene prioritization implicated both established and novel genes, such as *COMT*, *DHX36*, and *ASPRV1*, as critical for cardiometabolic health. Drug-target MR analyses, genetically proxying the physiological responses of approved and investigational cardiometabolic drugs, identified protective relationships for *GIPR*, *GLP1R*, *PCSK9*, *ANGPTL4*, and other targets with the *CM-Factor*, suggesting their potential therapeutic benefits for cardiometabolic multimorbidity. Hypothesis-generating screening of approximately 2,500 druggable genes prioritized 41 targets for cardiometabolic health, including 28 with strong links to major cardiometabolic biomarkers. These genes implicated fundamental biological mechanisms, such as *CRY2* (a core circadian clock regulator) and *OPRL1* (a key player in endogenous opioid signaling), offering valuable insights to guide future drug discovery efforts. Finally, two-step MR analyses prioritized *ENO3*, a protein important for striated muscle functioning, along with three other proteins, as potential mediators linking BMI to the *CM-Factor*. These findings advance our understanding of how obesity impacts the *CM-Factor* and its role in cardiometabolic disease.

Overall, Aim 3 demonstrates the utility of integrating GWAS data in multi-trait frameworks to elucidate the genetic signatures that influence the development of cardiometabolic diseases. These insights underscore the importance for future research and therapeutic development targeting cardiometabolic multimorbidity.

CHAPTER 6: DISCUSSION & FUTURE DIRECTIONS

The overarching motivation of this PhD was to leverage cutting-edge genetic methodologies and multi-omics approaches to inform the drug development pipeline for CVDs and their comorbidities. Cardiometabolic diseases, including T2D, NAFLD and CAD, represent a leading global health burden due to their high prevalence, shared risk factors, and multifactorial etiology. These conditions are underpinned by a complex interplay of genetic, environmental, and lifestyle factors. Despite advancements in treatment, there is a growing need for innovative therapeutic strategies targeting the shared pathophysiological mechanisms underlying these diseases. This thesis addressed this challenge by using advanced genetic tools to elucidate disease mechanisms and prioritize therapeutic targets at different stages of the drug development pipeline. Each aim tackled a distinct aspect of this challenge, scaling in complexity and culminating in the integration of multi-trait genetic analyses to understand cardiometabolic multimorbidity.

6.1. Leveraging Genetics to Inform Drug Development: Key Contributions from Each Aim

6.1.1. Aim 1 examined safety profiles of lipid-lowering drugs to non-European cohorts

The first aim focused on using drug-target MR to evaluate the causal effects of LDL-C lowering via PCSK9 and HMGCR inhibition on T2D risk and glycemic markers. This aim applied a relatively straightforward but robust genetic methodology to address the long-standing question of whether the benefits of LDL-C lowering come at the expense of adverse glycemic effects. The findings revealed no adverse effects of PCSK9 inhibition on T2D risk in most populations, supporting its role as a safe therapeutic target for LDL-C reduction. However, some evidence suggested that HMGCR inhibition might increase T2D risk, underscoring the need for population-specific analyses and careful consideration of comorbidities in clinical practice.

A critical limitation highlighted by Aim 1 is the lack of diversity in RCTs and genetic studies, which restricts the generalizability of findings to populations that reflect the true clinical diversity of individuals likely to receive these therapies.[28,489] Ensuring diversity in genetic datasets is crucial not only for improving the precision of efficacy estimates but also for addressing safety concerns, as genetic responses to drug targets may differ across ancestries. Expanding representation in genetic studies is therefore imperative for informing safe and effective drug development strategies for all populations.

Further, Aim 1 exemplifies how genetic tools can refine our understanding of existing drug mechanisms, providing evidence to support their efficacy across populations that more completely reflect the diversity of patients that use lipid-modulations therapies. These results are particularly relevant for cardiometabolic multimorbidity, where therapeutic interventions often need to balance efficacy for one condition against potential adverse effects on another. By validating the safety of PCSK9 inhibitors, this work reinforces their utility in managing cardiovascular risk across diverse populations and sets the stage for incorporating similar approaches into early drug development pipelines.

Section 1.6 in **Chapter 1** outlined the motivations for my focus on PCSK9 and HMGCR and future work inspired by Aim 1 will extend these analyses to additional approved and investigational lipid-lowering therapies in non-European populations with new non-European data sources. One of the key challenges I encountered in my multi-ancestry analyses during Aim 1 was the substantial disparity in sample sizes across ancestries for many of the traits studied, underscoring the critical need for large, diverse datasets to enable robust trans-ancestry genetic analyses.[28] The recent release of ~2,000 GWAS datasets from the Million Veteran Program (MVP)[490] provides an unparalleled opportunity to extend my work on understanding the genetic architecture of complex traits and diseases, especially in underrepresented populations. With over 635,000 participants and nearly 30% of the cohort genetically similar to African, Admixed American, and East Asian populations, the MVP allows for unprecedented exploration of non-European ancestry genetic variation. Using MVP data, I plan to again leverage drug-target MR approaches to ensure that causal inference and therapeutic predictions are robust across ancestries.

6.1.2. Aim 2 provided genetic and biological insights into alcohol use behaviors

The second aim expanded the scope by integrating cis-instrument MR with cortical proteomic and transcriptomic data to investigate the genetic and biological underpinnings of alcohol use behaviors. This aim not only identified 217 cortical proteins and 255 cell-type genes associated with problematic alcohol use but also prioritized novel therapeutic targets, such as SAMHD1 and VIPAS39, which are implicated in synaptic function and neurogenesis. Importantly, the findings highlighted pathways relevant to alcohol metabolism, neural signaling, and cardiometabolic health, further linking alcohol use behaviors to broader health outcomes.

This aim demonstrated the power of integrating multi-omics data to identify actionable therapeutic targets and potential drug repurposing opportunities. For instance, the identification of high-confidence targets such as SLC4A8 and CAB39L, along with validated drug repurposing candidates like prazosin and memantine, showcases how genetic analyses can accelerate target discovery and therapeutic innovation. Moreover, these findings bridge the gap between neuropsychiatric and cardiometabolic diseases, emphasizing the need for holistic approaches in addressing comorbid conditions.

6.1.3. Aim 3 elucidated a shared genetic architecture of cardiometabolic diseases

The third and most complex aim employed multivariate GWAS methods to model the shared genetic architecture of NAFLD, T2D, and CAD. This analysis revealed the cardiometabolic factor ("*CM-Factor*"), identifying 523 SNPs across 312 loci and implicating key genes such as COMT, DHX36, and ASPRV1. Drug-target MR validated the efficacy of several approved and investigational cardiometabolic therapeutics, including GLP1R, GIPR, PCSK9, and ANGPTL4, while also highlighting novel targets like CRY2 and OPRL1.

This aim represents a significant leap in complexity, integrating multi-trait genetic data to uncover the shared biology underlying cardiometabolic multimorbidity. The identification of heterogenous loci and novel genes not only enhances our understanding of disease mechanisms but also informs target prioritization for drug development. For example, the finding that PCSK9 inhibition has protective effects despite its association with T2D risk underscores the importance of context-specific genetic analyses in evaluating therapeutic potential. Similarly, the prioritization of CRY2 and OPRL1 as novel targets illustrates the potential of leveraging genetics to uncover previously unrecognized therapeutic opportunities.

I aim to continue to learn new methods to address multimorbidity. For example, an exciting future direction for understanding cardiometabolic multimorbidity involves leveraging advanced topic modeling techniques, such as treeLFA, introduced by Zhang et al. (2023). treeLFA (tree-structured Latent Factor Allocation) is a Bayesian topic modeling framework designed to identify latent disease clusters in binary diagnostic data. It extends traditional topic modeling approaches like Latent Dirichlet Allocation (LDA) by incorporating an informative prior derived from medical ontologies, such as the ICD-10 hierarchy. This prior ensures that diseases closely related on the tree structure (e.g., within the same chapter or subcategory of ICD-10) are more likely to cluster together, enhancing both the stability and biological relevance of the inferred multimorbidity topics. By treating individuals as “documents” and diseases as “words,” treeLFA captures patterns of disease co-occurrence, such as those representing cardiometabolic syndromes, while also allowing for the identification of unique clusters like the “healthy topic,” which relates to distinct lifestyle behaviors and protective genetic factors. Applying treeLFA to cardiometabolic multimorbidity could provide a deeper understanding of how metabolic and cardiovascular conditions cluster and interact, while also shedding light on disease progression and the shared pathways driving multimorbidity.

Moreover, integrating treeLFA with genetic analyses, as demonstrated through their topic-GWAS approach, offers a powerful method for identifying genetic loci associated with multimorbid traits. Zhang et al. (2023) demonstrated that using topic weights as quantitative traits in GWAS uncovered 128 loci associated with multimorbidity clusters, including loci that were not significant in single-disease GWASs. For example, topic-GWAS improved power for detecting associations by aggregating the shared genetic contributions of diseases within the same cluster, providing a more holistic view of genetic architecture. For cardiometabolic multimorbidity, this approach could reveal novel pleiotropic loci shared across conditions like type 2 diabetes, obesity, and cardiovascular disease. Additionally, polygenic risk scores (PRS) constructed from topic-GWAS results demonstrated enhanced predictive power for certain

disease outcomes compared to standard PRS, suggesting potential applications in personalized risk prediction and prevention. By combining the interpretability of disease clusters with the strength of genetic association studies, treeLFA represents a promising avenue for uncovering the systemic dysregulation underlying cardiometabolic multimorbidity and guiding targeted therapeutic strategies.

As highlighted above, the recently-released MVP GWAS data[490] has the potential to inform non-European ancestry analyses, and MVP's breadth of phenotypic and genomic data offers the potential to uncover ancestry-specific risk factors for diseases that disproportionately affect non-European populations. The combination of treeLFA and GenomicSEM further enhances the ability to explore multimorbidity and the shared genetic architecture of co-occurring diseases. I am excited to begin integrating MVP data into my research as it will significantly enhance the breadth and depth of my analyses, allowing for tailored insights into the genetic and environmental contributors to health disparities. By improving drug target validation, uncovering ancestry-specific risk factors, and employing cutting-edge methodologies to study multimorbidity, this work will bridge critical gaps in understanding and addressing the genetic underpinnings of disease in diverse populations

6.2. Scaling Complexity in the PhD: From Drug-Target Validation to Multivariate Insights

This PhD followed a deliberate trajectory, starting with relatively straightforward genetic analyses in Aim 1 and culminating in the more complex, integrative approaches used in Aim 3. Aim 1's drug-target MR analyses provided critical insights into the safety and efficacy of existing therapies, laying the groundwork for more ambitious investigations. Aim 2 built on this foundation by integrating multi-omics data to uncover novel targets for alcohol use behaviors, demonstrating the utility of combining genetic and proteomic data to inform drug discovery. Aim 3 tackled the most challenging questions, using multivariate frameworks to model shared genetic liabilities across multiple diseases and identify targets with broad relevance to cardiometabolic health.

6.2.1. Implications for the cardiometabolic drug development pipeline

One of the important contributions of this thesis is the validation of 34 approved and investigational cardiometabolic drug targets through genetic analyses. These findings have direct implications for the drug development pipeline, supporting the broader efficacy of established therapies and identifying novel opportunities for therapeutic innovation. For example:

- **Broad Cardiometabolic Efficacy:** The validation of targets like GLP1R, PCSK9, and ANGPTL4 highlights their potential to address cardiometabolic multimorbidity,

providing robust evidence for their inclusion in treatment regimens targeting diverse patient populations.

- **Target Discovery:** The prioritization of novel genes such as CRY2 and OPRL1 underscores the potential of genetics to uncover new therapeutic avenues, particularly in areas like circadian regulation and opioid signaling, which remain underexplored in cardiometabolic health.

- **Pathway Elucidation:** The integration of multi-omics data in Aim 2 and the identification of shared genetic mechanisms in Aim 3 offer valuable insights into the biological pathways driving cardiometabolic diseases, informing the development of targeted interventions.

6.3. Future Directions

The findings of this thesis highlight several opportunities for future research. In addition to the future directions outlined in this discussion, I have also begun work on several projects that are thematically and methodologically linked to the aims presented in this thesis. These projects build upon the foundational approaches and insights developed throughout the PhD and aim to address unresolved challenges in the field. Below, I present some of the preliminary results from one of these ongoing projects.

6.3.1. Investigating the genetic architecture shared between sleep-related traits, circulating glycemic biomarkers, and type 2 diabetes

Aim 2 of this PhD focused on one such behavioral factor—alcohol use—and demonstrated how integrating genetic and multi-omic approaches could uncover effector genes and pathways linking alcohol behaviors to metabolic and cardiometabolic health. Building on this framework, I have begun work focusing on sleep traits, a key behavioral risk factor strongly associated with T2D. Sleep disturbances, including insomnia, reduced sleep duration, and altered chronotype, are linked to insulin resistance and glycemic dysregulation. However, the genetic mechanisms underlying these associations remain poorly understood.

To address this, the study employed a two-step analytical approach to uncover shared genetic mechanisms. First, pairwise Bayesian colocalization analyses were used to systematically screen 890 genomic loci to identify regions where sleep traits, glycemic markers, and T2D share causal genetic variants (**Figure 6.1**). Next, loci showing evidence of colocalization were subjected to multi-trait colocalization analyses incorporating expression quantitative trait loci (eQTL) data from key metabolic tissues (e.g., liver, pancreas, adipose) and brain regions. This integration enabled the prioritization of effector genes by linking genetic signals to their regulatory effects on gene expression.

Figure 6.1. Overview of the study design investigating the genetic architecture linking sleep-related traits, glycemic markers, and T2D. *The study screened 890 genomic loci using colocalization to identify 25 trait-locus pairs with shared causal variants between sleep traits (insomnia, sleep duration, chronotype) and glycemic traits (T2D, HbA1c, fasting glucose, fasting insulin). Global genetic correlation and Mendelian randomization analyses provided evidence of genetic relationships. Bio-annotation analyses included local genetic correlation, cis-instrument MR, pathway and cell-type enrichment, tissue-of-action mapping, and multi-trait colocalization integrating eQTL data. Further validation involved cardiometabolic phenotyping, circadian gene expression analyses, and functional testing via CRISPR knockout of PATJ in zebrafish, prioritizing candidate causal genes and pathways for therapeutic insight.*

Preliminary findings identified 25 trait-locus pairs with evidence of shared causal variants. The *PATJ* locus emerged as a key finding, with a missense variant (rs12140153) implicated in both insomnia and T2D (**Figure 6.2**). Functional validation using CRISPR knockout experiments in zebrafish confirmed *PATJ*'s role in regulating both sleep behaviors and glycemic traits. Additional loci, including *FTO*, *CACNA1D*, and *RBM6*, were implicated in pathways related to circadian regulation and energy metabolism.

This integrative approach demonstrates how combining large-scale genomic screens with functional annotations, such as eQTL data, can uncover the shared genetic architecture of complex traits. By focusing on sleep as a modifiable risk factor, this study extends the framework of Aim 2 and highlights the role of genetic regulation in the interplay between behavioral traits and T2D, underscoring the potential for targeting sleep-related pathways to mitigate T2D risk and inform therapeutic development.

Figure 6.2. Colocalization results for sleep traits, glycemic markers, and T2D.
(a) *Pairwise Bayesian colocalization results for genomic regions showing evidence of shared causal variants between sleep traits (insomnia, sleep duration, and chronotype), glycemic markers (HbA1c, fasting glucose, and fasting insulin), and T2D. Each point represents a colocalized locus, plotted by the posterior probability for hypothesis H4 (PP-H4, indicating shared causal variants) on the y-axis and the size of the 95% credible set on the x-axis. The size of the points reflects the number of variants in the 95% credible set. Colors denote trait pairs, as indicated in the legend.* ***(b)*** *Regional colocalization analysis for the PATJ locus on chromosome 1 (62,208,149–62,629,592), which demonstrated evidence of shared causal variants for T2D (top panel) and chronotype (bottom panel). The locus is represented as a regional association plot showing GWAS -log10(P) values for T2D (top) and chronotype (bottom), with the lead variant, rs12140153, highlighted in pink. This missense variant in PATJ was identified as the most likely causal variant based on colocalization and credible set analysis. These findings highlight PATJ as a shared genetic locus linking sleep traits and T2D, with implications for understanding their genetic and biological relationships.*

6.3.2. Addressing pleiotropy to refine causal inference in cardiometabolic disease

Identifying causal risk factors and biomarkers is critical to advancing precision medicine for cardiometabolic diseases,[197] and MR is powerful tool for causal inference by leveraging genetic variants as instrumental variables to minimize biases from confounding and reverse causation.

MR faces a significant challenge: pleiotropy, particularly horizontal pleiotropy, where genetic variants influence outcomes through pathways independent of the exposure of interest. This issue

poses a serious threat to the validity of MR findings by violating instrumental variable assumptions, potentially leading to biased estimates or erroneous conclusions. In the example provided by Larsson et al.in their review of MR for cardiometabolic diseases recently published in the *European Heart Journal*,[197] suppose SNPs strongly associated with smoking are included in an MR analysis aimed at evaluating the causal effect of coffee consumption on CAD. If these smoking-associated SNPs are correlated with coffee consumption (here, the exposure of interest) but also independently influence CAD through mechanisms unrelated to coffee, this introduces horizontal pleiotropy.[197]

Pleiotropy often arises in polygenic MR studies, where hundreds or thousands of genetic variants are included in the analysis.[491,492] The growing size of genome-wide association studies and the identification of distal variants with smaller effect sizes exacerbate this challenge,[493] as these variants often have smaller effect sizes and are more distally mediated through intermediate phenotypes making them more likely to act through confounders.[491,492] This increased polygenicity introduces several challenges, including distally mediated effects where small effects on the exposure are more likely to influence other traits, making them prone to horizontal pleiotropy; and heritable confounding where variants may simultaneously influence confounders of the exposure and outcome, leading to biased MR estimates in the direction of observational associations. Recent studies have highlighted the potential for heritable confounding in MR, where genetic variants associated with a confounder of the exposure and outcome reintroduce bias, aligning MR estimates with observational associations.[493] Conventional MR methods, including inverse variance weighting and MR Egger, are susceptible to these biases, particularly when horizontal pleiotropy is widespread,[493] necessitating new approaches for selecting and grouping instruments to minimize pleiotropic effects, because misattributing causal relationships due to horizontal pleiotropy could hinder our understanding of cardiometabolic disease mechanisms and impair efforts to identify effective therapeutic targets.[493]

I have been working on developing a method—Phenotypic Network-Informed Mendelian Randomization (PNIMR)—that aims to refine genetic instrument selection and improve causal inference in MR studies. This approach leverages the integration of Phenome-wide Association Studies (PheWAS) data and network-based centrality metrics[494] to systematically evaluate the phenotypic connectivity of genetic variants. By contextualizing genetic variants within phenotypic networks, PNIMR may help prioritize SNPs that are more likely to act directly through the exposure of interest while minimizing bias from pleiotropic effects. I hypothesize that this network-informed MR framework improves causal inference by reducing bias from heritable confounding and pleiotropy, and then demonstrate how this approach clarifies instrumentation of polygenic exposures and informs the exposure-outcome relationships using several positive and negative controls.

This work is ongoing, but aims to add to the growing body of MR methods that may be used to inform instrument selection (e.g., MR-CORGE[495]). Below we a brief background of network centrality metrics, PNIMR, and the preliminary analyses results of negative control analyses where PNIMR provides insight into two cardiometabolic risk factor exposures, C-reactive protein (CRP) and vitamin D, where polygenic MR studies have suggested genetic associations likely driven by the inclusion of variants linked to other biomarkers, while randomized controlled trials and cis-instrumentation of the causal gene loci have demonstrated null effects.[496-

[500] Using these exposures, we show that PNIMR accurately identifies null effects when using instruments from low-centrality groupings, whereas instruments constructed with more highly phenotypically connected variants replicate the effects observed in polygenic MR analyses.

6.3.2a. What are network-based centrality metrics? Centrality metrics are important tools in network science, offering insights into the roles and influence of nodes within a network (**Figure 6.3**).[494] These metrics prioritize network components to highlight key or significant elements and have been extensively applied across diverse fields, including social sciences, biology, and systems biology, to understand how elements interact within complex systems. In biological systems, centrality metrics have been pivotal in exploring protein-protein interaction networks, gene regulatory networks, and disease-gene associations.[501] For instance, nodes with high centrality values often correspond to genes or proteins playing essential roles in cellular processes or disease pathways.[501] Identifying such nodes can guide experimental prioritization and therapeutic target discovery. Similarly, in ecological networks, centrality metrics help pinpoint keystone species crucial for ecosystem stability, highlighting the versatility of these measures across domains.[494] By quantifying the importance of nodes, centrality metrics facilitate inference into structural hierarchies and critical points of influence within networks, making them invaluable for analyzing intricate relationships.

Figure 6.3. Overview of network centrality metrics. *This figure illustrates the application of network centrality metrics to a network composed of nodes and edges, where nodes represent entities (e.g., variables or elements of interest) and edges represent significant relationships or associations between them. Centrality metrics provide insights into the roles and influence of nodes within the network: Degree Centrality reflects the number of direct connections a node has, identifying highly connected nodes; Betweenness Centrality captures nodes that act as bridges between disparate parts of the network, highlighting their role in connecting distinct clusters; Closeness Centrality measures how efficiently a node is connected to all other nodes, indicating its integration within the network; Eigenvector Centrality emphasizes nodes connected to other highly connected nodes, identifying influential nodes within key clusters; and PageRank centrality ranks nodes based on the importance of their neighbors and the strength of*

their connections. These metrics collectively characterize the structure and dynamics of the network, enabling the identification of key nodes and their roles within the broader system. Figure adapted from ref. [494] *and made with BioRender).*

6.3.2b. What is PNIMR? PNIMR applies network-informed principles to support MR analyses by leveraging phenotypic connectivity metrics to inform instrument selection. It operates on the premise that SNPs with lower connectivity (i.e., lower centrality scores) are more likely to act directly through the exposure of interest, while SNPs with higher connectivity (i.e., higher centrality scores) may influence multiple traits, increasing the likelihood of pleiotropy. Using centrality metrics derived from PheWAS data, PNIMR helps identify and prioritize SNPs with lower connectivity, aiming to reduce bias from horizontal pleiotropy. The PNIMR code is compatible with existing MR analysis pipelines in the TwoSampleMR R package and utilizes publicly available genetic resources, facilitating its integration into broader research workflows. In the following sections, the PNIMR procedure is outlined.

Step 1: Conduct PheWAS for polygenic MR instrument SNPs. To construct genetic instruments for MR analyses, we followed a systematic approach leveraging publicly available datasets in the MRC Integrative Epidemiology Unit (IEU) Open GWAS Project.[49] Initially, SNPs strongly associated with the exposure of interest were extracted using TwoSampleMR R package (version 6.8).[49] We used conventional single-variable MR instrumentation thresholds (i.e., conventional genome-wide significance (P-value $< 5\times10^{-8}$) from the Open GWAS database. Instrument SNPs were required to meet stringent inclusion criteria, such as independence based on linkage disequilibrium clumping (LD $r^2 < 0.001$ within the conventional 10 Mb window) and robust statistical support in the exposure GWAS.

For each instrument, a PheWAS was then conducted using the *ieugwasr* R package function *phewas()* the Open GWAS Project to explore their associations with a broad range of phenotypes across multiple datasets. PheWAS provided insights into the phenotypic landscape of each SNP, allowing us to identify their pleiotropic effects and potential horizontal pleiotropy. A P-value threshold of $< 5\times10^{-5}$ was applied to the PheWAS results to ensure that only robust SNP-trait associations were retained for quality control.

Step 2: Composite score and centrality metrics calculation. The filtered SNP-trait associations from PheWAS were used to construct a bipartite network representing the relationship between SNPs and phenotypes. We hypothesized that the centrality metrics would inform direct pathway identification, i.e., SNPs with low centrality scores are less likely to influence unrelated traits and pathways, making them stronger candidates for causal inference and provide biological insights, e.g., high-centrality SNPs, which connect multiple traits, can be identified as pleiotropic hubs, shedding light on shared biological mechanisms between traits. To elucidate the phenotypic connectivity of SNPs, we calculated a composite score integrating several graph-theoretic centrality metrics. These metrics were derived from a weighted network constructed using SNP-trait associations identified through PheWAS data, where nodes represent traits or SNPs and edges are each SNP-trait association in the filtered PheWAS. We computed four core centrality metrics to characterize SNPs' roles within the network (Degree, Betweenness, Closeness, Eigenvector, and PageRank).

We then calculated a composite score for each SNP included in the PheWAS. The composite score integrates multiple centrality metrics to characterize the phenotypic connectivity of SNPs in a weighted network. This approach leverages the power of graph theory to identify SNPs with varying levels of phenotypic connectivity and assesses their influence within a broader phenotypic landscape. Each centrality metric was normalized to a 0–1 scale to allow comparability:

$$Normalized\ Metric = \frac{Metric\ Value}{Maximum\ Metric\ Value}$$

The composite score for each SNP was then calculated as the weighted average of these normalized metrics:

$$Composite\ Score = (0.25 \times Degree\ Norm + 0.25 \times Betweenness\ Norm + 0.25 \times Closeness\ Norm + 0.25 \times Eigenvector$$

This weighting equally emphasizes direct connectivity, bridging capacity, network efficiency, and global influence, providing a holistic measure of phenotypic connectivity.

Step 3. Instrument construction for stratified network-based, and cumulative, groupings.
SNPs were stratified into discrete instrument groups based on deciles of the composite score. For example, SNPs within the lowest decile (composite score ≤ 0.10) were considered core SNPs, likely to act directly through the exposure. For group-specific analyses, MR estimates were derived individually for each decile, facilitating sensitivity analyses to assess the robustness of causal estimates across varying levels of network centrality. Successive decile thresholds grouped SNPs with broader phenotypic connections and created instruments for each decile. Cumulative groupings were also constructed by aggregating SNPs from successive deciles, enabling sensitivity analyses to evaluate how including more pleiotropic SNPs affects instrument composition and causal inference. The composite score and centrality-based stratification framework were applied in the subsequent MR analyses. Each discrete decile and cumulative grouping were used as an exposure set to estimate causal effects. Group-specific analyses allowed for sensitivity testing by focusing on core SNPs with direct biological roles, while cumulative analyses examined how the inclusion of peripheral variants influenced causal estimates. This approach provides a robust means to differentiate SNPs with direct biological connections to the exposure from those contributing to horizontal pleiotropy.

After constructing the PNIMR-informed instrument groupings, we used inverse-variance weighted MR as the primary method and applied additional techniques to assess the robustness of findings and address potential violations of MR assumptions. These included MR-Egger, weighted median, penalized weighted median, and weighted mode methods to detect and adjust for pleiotropy or invalid instruments. Heterogeneity in MR estimates was evaluated using the MR-Egger intercept and Cochran's Q tests, with significant heterogeneity (Cochran's Q P-value < 0.05) addressed through MR-LASSO. MR-LASSO applies penalization to outlier SNPs, refining instrument selection and generating corrected inverse-variance weighted estimates for improved causal inference.[318]

6.3.2c. Negative control #1: Vitamin D and CAD.

While observational studies have consistently linked lower vitamin D levels to a reduced risk of CAD, RCTs of vitamin D supplementation have not demonstrated significant cardiovascular benefits in the general population.[499] Polygenic MR studies, using variants distributed throughout the genome, have often recapitulated these observational associations, likely due to the inclusion of pleiotropic variants associated with other biomarkers or pathways.[502] In contrast, MR studies employing more stringent, biologically informed instrumentation approaches—such as focusing on variants within the genetic loci of genes directly involved in vitamin D metabolism—have aligned with the null findings of RCTs, reinforcing the absence of a causal relationship.[502,503]

We hypothesize that the PNIMR method will demonstrate that vitamin D instruments with low centrality—representing SNPs more likely to act directly through vitamin D metabolism—will replicate the null findings observed in RCTs and biologically informed MR analyses with careful instrumentation of loci involved in vitamin D metabolism, while vitamin D instruments composed of increasingly highly connected SNPs (variants with pleiotropic effects) will recapitulate the observational associations and results of polygenic MR studies, which are likely influenced by genetic correlations with other traits or biomarkers. To further validate this hypothesis, we aim to show that removing these highly connected, pleiotropic variants from single-variable, genome-wide vitamin D instruments will eliminate the apparent causal effect, aligning the MR estimates with the null results from RCTs.

6.3.2d. Negative control #2: CRP and CAD.

CRP is a marker of systemic inflammation commonly linked to CAD risk in observational studies.[504] However, cis-instrument MR[496-498] analyses examining variants within the *CRP* locus with cis-in have shown no causal relationship with CAD.[496-498] By contrast, MR studies using polygenic CRP instruments have revealed widespread genetic associations with conditions such as stroke, Alzheimer's disease, and CAD,[505-508] indicating that these associations may be driven by pleiotropic variants distributed across the genome. Collectively, this evidence suggests that CRP functions as a biomarker of systemic inflammation rather than a causal factor. We hypothesize that variants with lower network centrality scores, which are more likely to serve as direct instruments for CRP, will show null associations with CAD. Conversely, CRP instruments enriched with SNPs exhibiting higher network centrality scores, indicative of pleiotropy, are expected to drive adverse associations with CAD.

6.3.2e. PNIMR correctly shows apparent impact of vitamin D and CRP on CAD is driven by highly connected SNPs.

For CRP, instruments constructed from SNPs in the lowest centrality deciles (i.e., less phenotypically connected SNPs) produced null associations with CAD across all MR methods, including IVW MR-Egger, weighted median, and penalized weighted median (**Figure 6.4a**). By contrast, instruments including SNPs from higher centrality deciles replicated the positive associations observed in previous polygenic MR analyses, which likely reflect pleiotropic effects. For example, the second CRP instrument grouping was the lowest group to demonstrate and adverse relationship with CAD and these SNPs had substantial pleiotropic relationships with other cardiometabolic risk factors such as traits related to adiposity and cholesterol levels (**Figure 6.4b**). Among the SNPs in CRP group 1 instrument set was a variant in the *CRP* locus, which given the previous MR work showing that the cis-CRP variants have null relationships with CAD,[496-498] suggests that

the CRP group 1 instrument set reflects a genetic signal likely capturing CRP levels and supports its validity as an instrumentation approach.

The results with vitamin D and CAD showed a similar pattern (**Figure 6.5a**): Instruments constructed from SNPs in the lowest centrality deciles showed no evidence of a causal relationship with CAD, consistent with findings from randomized controlled trials. Instruments incorporating SNPs from higher centrality deciles, however, suggested significant associations, which are likely driven by pleiotropic loci linked to other traits or biomarkers. For example, in **Figure 6.5b**, we highlight the network connectivity for the SNPs in the 7th decile vitamin D instrument, as this instrument grouping represents the first vitamin D instrument grouping to demonstrate a genetic impact on CAD aligning with the previously reported but confounded protective effect. This network visualization shows extensive pleiotropy, with highly connected SNPs linked to multiple traits, emphasizing the role of pleiotropic loci in driving these associations. The attenuation of these associations when using low-centrality instruments underscores the ability of PNIMR to refine instrument selection and minimize bias.

Figure 6.4. PNIMR-Informed instrument groupings for CRP and their genetic estimates on CAD. (a) *shows the forest plot of odds ratios (95% CI) for CAD per 1-SD increment in genetically predicted CRP levels, stratified by PNIMR instrument groupings based on centrality deciles. Results are presented for individual (Instrument Groups 1–10) and cumulative (Cumulative Instrument Groups 1–10) groupings, with methods including inverse variance weighted (IVW), MR-Egger, weighted median, weighted mode, and MR-LASSO. Low-centrality instruments show null effects, while higher-centrality instruments suggest significant associations, likely reflecting pleiotropy.* ***(b)*** *illustrates the connectivity network for Instrument Group 2, highlighting SNP-trait associations. Nodes represent SNPs (orange) and traits (blue), with edges denoting significant SNP-trait associations. Node size corresponds to composite*

centrality scores, with larger nodes indicating higher phenotypic connectivity. This network demonstrates limited pleiotropy for Instrument Group 2, supporting its robustness as an instrument for CRP in MR analyses. Together, these panels showcase PNIMR's ability to identify low-pleiotropy instruments for more reliable causal inference.

Figure 6.5. Application of the PNIMR framework to evaluate vitamin D's causal relationship with CAD. (a) *The forest plot displays odds ratios (95% CI) for CAD per 1-SD increase in genetically predicted vitamin D levels across instrument groupings, stratified by centrality deciles (1–10). Results for cumulative groupings are also shown, highlighting the effects of including more highly connected SNPs. Methods include inverse variance weighted (IVW), MR-Egger, weighted median, weighted mode, and MR-LASSO. Instruments composed of lower-centrality SNPs (e.g., decile 1) yield null associations, aligning with findings from randomized controlled trials, while higher-centrality groupings replicate previously reported observational associations, likely influenced by pleiotropic effects.* ***(b)*** *The network visualization represents the pleiotropic connections within the 7th decile instrument grouping (the first vitamin D*

instrument grouping to demonstrate a genetics impact on CAD), with nodes indicating either traits (e.g., phenotypes) or SNPs, and edges representing SNP-trait associations identified through PheWAS data. Node size reflects composite centrality scores, with higher scores indicating greater phenotypic connectivity.

6.3.2f. Concluding Remarks for PNIMR. These preliminary results demonstrate the potential of PNIMR as a significant advancement in refining MR instrumentation. By integrating phenotypic connectivity metrics derived from PheWAS data, PNIMR offers a systematic approach to stratify genetic instruments based on their pleiotropic potential, addressing a key limitation in MR analyses. The findings for CRP and vitamin D underscore the ability of PNIMR to differentiate robust, low-centrality instruments from highly connected, pleiotropic SNPs that may bias causal estimates. This framework not only aligns with null results from randomized controlled trials and cis-instrument studies but also reveals the pleiotropic pathways driving spurious associations in polygenic MR.

While these results are preliminary, they provide compelling evidence that PNIMR can improve the precision and validity of causal inference in cardiometabolic disease research. The ability to prioritize causal risk factors and correctly attribute MR results has broad implications for therapeutic target identification and translational research, particularly in addressing complex diseases like CAD. Further validation and expansion of this approach across additional traits and ancestries could establish PNIMR as a valuable tool for advancing precision medicine and bridging gaps in our understanding of cardiometabolic health.

6.4. Conclusions

This thesis demonstrates the transformative potential of leveraging genetic data to inform the drug development pipeline for cardiometabolic diseases. By validating established targets, uncovering novel therapeutic opportunities, and elucidating shared genetic mechanisms, this work provides a comprehensive framework for addressing the complex challenges of cardiometabolic multimorbidity. Each aim contributes a unique piece to this puzzle, highlighting the power of genetic methodologies to advance our understanding of disease and inform therapeutic innovation. Moreover, the emphasis on diversity and representation underscores the importance of ensuring that genetic discoveries translate into equitable clinical benefits, paving the way for safer and more effective therapies that reflect the diversity of real-world patient populations.

REFERENCES

1. Rosoff DB, Wagner J, Jung J, et al. Multi-omics mendelian randomization study investigating the impact of pcsk9 and hmgcr inhibition on type 2 diabetes across five populations. *Diabetes*. 2024:db240451. doi:10.2337/db24-0451
2. Rosoff DB, Wagner J, Bell AS, Mavromatis LA, Jung J, Lohoff FW. A multi-omics mendelian randomization study identifies new therapeutic targets for alcohol use disorder and problem drinking. *Nature Human Behaviour*. 2024/11/11 2024;doi:10.1038/s41562-024-02040-1
3. Timmis A, Townsend N, Gale CP, et al. European society of cardiology: Cardiovascular disease statistics 2019. *Eur Heart J*. Jan 1 2020;41(1):12-85. doi:10.1093/eurheartj/ehz859
4. Virani SS, Alonso A, Benjamin EJ, et al. Heart disease and stroke statistics-2020 update: A report from the american heart association. *Circulation*. Mar 3 2020;141(9):e139-e596. doi:10.1161/cir.0000000000000757

5. Roth GA, Mensah GA, Johnson CO, et al. Global burden of cardiovascular diseases and risk factors, 1990-2019: Update from the gbd 2019 study. *J Am Coll Cardiol*. Dec 22 2020;76(25):2982-3021. doi:10.1016/j.jacc.2020.11.010
6. Holmes MV, Richardson TG, Ference BA, Davies NM, Davey Smith G. Integrating genomics with biomarkers and therapeutic targets to invigorate cardiovascular drug development. *Nature Reviews Cardiology*. 2021/06/01 2021;18(6):435-453. doi:10.1038/s41569-020-00493-1
7. Vernon ST, Coffey S, Bhindi R, et al. Increasing proportion of st elevation myocardial infarction patients with coronary atherosclerosis poorly explained by standard modifiable risk factors. *Eur J Prev Cardiol*. Nov 2017;24(17):1824-1830. doi:10.1177/2047487317720287
8. Vernon ST, Coffey S, D'Souza M, et al. St-segment-elevation myocardial infarction (stemi) patients without standard modifiable cardiovascular risk factors-how common are they, and what are their outcomes? *J Am Heart Assoc*. Nov 5 2019;8(21):e013296. doi:10.1161/jaha.119.013296
9. Administation USFD. The drug development process. Accessed November 25, 2021. https://www.fda.gov/patients/drug-development-process/step-3-clinical-research
10. Harrison RK. Phase ii and phase iii failures: 2013-2015. *Nat Rev Drug Discov*. Dec 2016;15(12):817-818. doi:10.1038/nrd.2016.184
11. Wouters OJ, McKee M, Luyten J. Estimated research and development investment needed to bring a new medicine to market, 2009-2018. *JAMA*. 2020;323(9):844-853. doi:10.1001/jama.2020.1166
12. Pammolli F, Magazzini L, Riccaboni M. The productivity crisis in pharmaceutical r&d. *Nature Reviews Drug Discovery*. 2011/06/01 2011;10(6):428-438. doi:10.1038/nrd3405
13. Fordyce CB, Roe MT, Ahmad T, et al. Cardiovascular drug development: Is it dead or just hibernating? *J Am Coll Cardiol*. Apr 21 2015;65(15):1567-82. doi:10.1016/j.jacc.2015.03.016
14. Figtree GA, Broadfoot K, Casadei B, et al. A call to action for new global approaches to cardiovascular disease drug solutions. *European Heart Journal*. 2021;42(15):1464-1475. doi:10.1093/eurheartj/ehab068
15. Davey Smith G, Ebrahim S. What can mendelian randomisation tell us about modifiable behavioural and environmental exposures? *Bmj*. May 7 2005;330(7499):1076-9. doi:10.1136/bmj.330.7499.1076
16. Grotzinger AD, Rhemtulla M, de Vlaming R, et al. Genomic structural equation modelling provides insights into the multivariate genetic architecture of complex traits. *Nature Human Behaviour*. 2019/05/01 2019;3(5):513-525. doi:10.1038/s41562-019-0566-x
17. Fda approves add-on therapy for patients with genetic form of severely high cholesterol. 02/11/2021, Accessed May 5, 2021. https://www.fda.gov/drugs/drug-safety-and-availability/fda-approves-add-therapy-patients-genetic-form-severely-high-cholesterol
18. Elguindy A, Yacoub MH. The discovery of pcsk9 inhibitors: A tale of creativity and multifaceted translational research. *Glob Cardiol Sci Pract*. 2013;2013(4):343-347. doi:10.5339/gcsp.2013.39
19. Graham MJ, Lee RG, Brandt TA, et al. Cardiovascular and metabolic effects of angptl3 antisense oligonucleotides. *New England Journal of Medicine*. 2017/07/20 2017;377(3):222-232. doi:10.1056/NEJMoa1701329

20. Szustakowski JD, Balasubramanian S, Kvikstad E, et al. Advancing human genetics research and drug discovery through exome sequencing of the uk biobank. *Nature Genetics*. 2021/07/01 2021;53(7):942-948. doi:10.1038/s41588-021-00885-0
21. Nelson MR, Tipney H, Painter JL, et al. The support of human genetic evidence for approved drug indications. *Nature Genetics*. 2015/08/01 2015;47(8):856-860. doi:10.1038/ng.3314
22. Federation ID. Idf diabetes atlas, 7th ed. Accessed August 4, 2017. http://www.diabetesatlas.org/
23. Golden SH, Yajnik C, Phatak S, Hanson RL, Knowler WC. Racial/ethnic differences in the burden of type 2 diabetes over the life course: A focus on the USA and india. *Diabetologia*. Oct 2019;62(10):1751-1760. doi:10.1007/s00125-019-4968-0
24. Khan MAB, Hashim MJ, King JK, Govender RD, Mustafa H, Al Kaabi J. Epidemiology of type 2 diabetes - global burden of disease and forecasted trends. *J Epidemiol Glob Health*. Mar 2020;10(1):107-111. doi:10.2991/jegh.k.191028.001
25. Spanakis EK, Golden SH. Race/ethnic difference in diabetes and diabetic complications. *Curr Diab Rep*. Dec 2013;13(6):814-23. doi:10.1007/s11892-013-0421-9
26. Barroso I. The importance of increasing population diversity in genetic studies of type 2 diabetes and related glycaemic traits. *Diabetologia*. 2021/12/01 2021;64(12):2653-2664. doi:10.1007/s00125-021-05575-4
27. Cusi K, Ocampo GL. Unmet needs in hispanic/latino patients with type 2 diabetes mellitus. *Am J Med*. Oct 2011;124(10 Suppl):S2-9. doi:10.1016/j.amjmed.2011.07.017
28. Fatumo S, Chikowore T, Choudhury A, Ayub M, Martin AR, Kuchenbaecker K. A roadmap to increase diversity in genomic studies. *Nature Medicine*. 2022/02/01 2022;28(2):243-250. doi:10.1038/s41591-021-01672-4
29. Clark LT, Watkins L, Piña IL, et al. Increasing diversity in clinical trials: Overcoming critical barriers. *Current Problems in Cardiology*. 2019/05/01/ 2019;44(5):148-172. doi:https://doi.org/10.1016/j.cpcardiol.2018.11.002
30. Sirugo G, Williams SM, Tishkoff SA. The missing diversity in human genetic studies. *Cell*. 2019;177(1):26-31. doi:10.1016/j.cell.2019.02.048
31. Clarke SL, Assimes TL, Tcheandjieu C. The propagation of racial disparities in cardiovascular genomics research. *Circulation: Genomic and Precision Medicine*. 2021/10/01 2021;14(5):e003178. doi:10.1161/CIRCGEN.121.003178
32. Tanne JH. Us must urgently correct ethnic and racial disparities in clinical trials, says report. *BMJ*. 2022;377:o1292. doi:10.1136/bmj.o1292
33. Einarson TR, Acs A, Ludwig C, Panton UH. Prevalence of cardiovascular disease in type 2 diabetes: A systematic literature review of scientific evidence from across the world in 2007–2017. *Cardiovascular Diabetology*. 2018/06/08 2018;17(1):83. doi:10.1186/s12933-018-0728-6
34. Chatterjee S, Khunti K, Davies MJ. Type 2 diabetes. *Lancet*. Jun 03 2017;389(10085):2239-2251. doi:10.1016/S0140-6736(17)30058-2
35. Ma C-X, Ma X-N, Guan C-H, Li Y-D, Mauricio D, Fu S-B. Cardiovascular disease in type 2 diabetes mellitus: Progress toward personalized management. *Cardiovascular Diabetology*. 2022/05/14 2022;21(1):74. doi:10.1186/s12933-022-01516-6
36. Elnaem MH, Mohamed MHN, Huri HZ, Azarisman SM, Elkalmi RM. Statin therapy prescribing for patients with type 2 diabetes mellitus: A review of current evidence and challenges. *J Pharm Bioallied Sci*. Apr-Jun 2017;9(2):80-87. doi:10.4103/jpbs.JPBS_30_17

37. Sattar N, Preiss D, Murray HM, et al. Statins and risk of incident diabetes: A collaborative meta-analysis of randomised statin trials. *Lancet*. Feb 27 2010;375(9716):735-42. doi:10.1016/s0140-6736(09)61965-6
38. Preiss D, Seshasai SR, Welsh P, et al. Risk of incident diabetes with intensive-dose compared with moderate-dose statin therapy: A meta-analysis. *Jama*. Jun 22 2011;305(24):2556-64. doi:10.1001/jama.2011.860
39. Schmidt AF, Swerdlow DI, Holmes MV, et al. Pcsk9 genetic variants and risk of type 2 diabetes: A mendelian randomisation study. *Lancet Diabetes Endocrinol*. Feb 2017;5(2):97-105. doi:10.1016/s2213-8587(16)30396-5
40. Ference BA, Robinson JG, Brook RD, et al. Variation in pcsk9 and hmgcr and risk of cardiovascular disease and diabetes. *N Engl J Med*. Dec 1 2016;375(22):2144-2153. doi:10.1056/NEJMoa1604304
41. Cho Y, Choe E, Lee YH, et al. Risk of diabetes in patients treated with hmg-coa reductase inhibitors. *Metabolism*. Apr 2015;64(4):482-8. doi:10.1016/j.metabol.2014.09.008
42. Liu G, Shi M, Mosley JD, et al. A mendelian randomization approach using 3-hmg-coenzyme-a reductase gene variation to evaluate the association of statin-induced low-density lipoprotein cholesterol lowering with noncardiovascular disease phenotypes. *JAMA Network Open*. 2021;4(6):e2112820-e2112820. doi:10.1001/jamanetworkopen.2021.12820
43. Swerdlow DI, Preiss D, Kuchenbaecker KB, et al. Hmg-coenzyme a reductase inhibition, type 2 diabetes, and bodyweight: Evidence from genetic analysis and randomised trials. *The Lancet*. 2015;385(9965):351-361. doi:10.1016/S0140-6736(14)61183-1
44. Sabatine MS, Giugliano RP, Keech AC, et al. Evolocumab and clinical outcomes in patients with cardiovascular disease. *New England Journal of Medicine*. 2017;376(18):1713-1722. doi:10.1056/NEJMoa1615664
45. Ray KK, Wright RS, Kallend D, et al. Two phase 3 trials of inclisiran in patients with elevated ldl cholesterol. *N Engl J Med*. Apr 16 2020;382(16):1507-1519. doi:10.1056/NEJMoa1912387
46. Sabatine MS, Leiter LA, Wiviott SD, et al. Cardiovascular safety and efficacy of the pcsk9 inhibitor evolocumab in patients with and without diabetes and the effect of evolocumab on glycaemia and risk of new-onset diabetes: A prespecified analysis of the fourier randomised controlled trial. *Lancet Diabetes Endocrinol*. Dec 2017;5(12):941-950. doi:10.1016/s2213-8587(17)30313-3
47. Chen Q, Wu G, Li C, Qin X, Liu R, Zhang M. Safety of proprotein convertase subtilisin/kexin type 9 monoclonal antibodies in regard to diabetes mellitus: A systematic review and meta-analysis of randomized controlled trials. *Am J Cardiovasc Drugs*. Aug 2020;20(4):343-353. doi:10.1007/s40256-019-00386-w
48. Sanderson E, Glymour MM, Holmes MV, et al. Mendelian randomization. *Nature Reviews Methods Primers*. 2022/02/10 2022;2(1):6. doi:10.1038/s43586-021-00092-5
49. Hemani G, Zheng J, Elsworth B, et al. The mr-base platform supports systematic causal inference across the human phenome. *Elife*. May 30 2018;7doi:10.7554/eLife.34408
50. Schmidt AF, Finan C, Gordillo-Marañón M, et al. Genetic drug target validation using mendelian randomisation. *Nature communications*. 2020;11(1):3255-3255. doi:10.1038/s41467-020-16969-0
51. Walker VM, Davey Smith G, Davies NM, Martin RM. Mendelian randomization: A novel approach for the prediction of adverse drug events and drug repurposing opportunities. *Int J Epidemiol*. Dec 1 2017;46(6):2078-2089. doi:10.1093/ije/dyx207

52. Rao AS, Lindholm D, Rivas MA, Knowles JW, Montgomery SB, Ingelsson E. Large-scale phenome-wide association study of pcsk9 variants demonstrates protection against ischemic stroke. *Circ Genom Precis Med*. Jul 2018;11(7):e002162. doi:10.1161/circgen.118.002162
53. Paneni F, Costantino S. Pcsk9 in diabetes: Sweet, bitter or sour? *European Heart Journal*. 2019;40(4):369-371. doi:10.1093/eurheartj/ehy432
54. Soremekun O, Karhunen V, He Y, et al. Lipid traits and type 2 diabetes risk in african ancestry individuals: A mendelian randomization study. *EBioMedicine*. Apr 2022;78:103953. doi:10.1016/j.ebiom.2022.103953
55. Chen J, Spracklen CN, Marenne G, et al. The trans-ancestral genomic architecture of glycemic traits. *Nat Genet*. Jun 2021;53(6):840-860. doi:10.1038/s41588-021-00852-9
56. Ferkingstad E, Sulem P, Atlason BA, et al. Large-scale integration of the plasma proteome with genetics and disease. *Nat Genet*. Dec 2021;53(12):1712-1721. doi:10.1038/s41588-021-00978-w
57. Consortium GT. The genotype-tissue expression (gtex) project. *Nature genetics*. 2013;45(6):580-585. doi:10.1038/ng.2653
58. Sabatine MS. Pcsk9 inhibitors: Clinical evidence and implementation. *Nature Reviews Cardiology*. 2019/03/01 2019;16(3):155-165. doi:10.1038/s41569-018-0107-8
59. Da Dalt L, Ruscica M, Bonacina F, et al. Pcsk9 deficiency reduces insulin secretion and promotes glucose intolerance: The role of the low-density lipoprotein receptor. *Eur Heart J*. Jan 21 2019;40(4):357-368. doi:10.1093/eurheartj/ehy357
60. Hoek AG, van Oort S, Mukamal KJ, Beulens JWJ. Alcohol consumption and cardiovascular disease risk: Placing new data in context. *Curr Atheroscler Rep*. Jan 2022;24(1):51-59. doi:10.1007/s11883-022-00992-1
61. Fernández-Solà J. Cardiovascular risks and benefits of moderate and heavy alcohol consumption. *Nat Rev Cardiol*. Oct 2015;12(10):576-87. doi:10.1038/nrcardio.2015.91
62. Di Castelnuovo A, Costanzo S, Bagnardi V, Donati MB, Iacoviello L, de Gaetano G. Alcohol dosing and total mortality in men and women: An updated meta-analysis of 34 prospective studies. *Arch Intern Med*. Dec 11-25 2006;166(22):2437-45. doi:10.1001/archinte.166.22.2437
63. NIAAA. What is binge drinking? https://www.niaaa.nih.gov/publications/brochures-and-fact-sheets/binge-drinking#:~:text=What%20Is%20Binge%20Drinking%3F,alcohol%20per%20deciliter%E2%80%93or%20higher.
64. Biddinger KJ, Emdin CA, Haas ME, et al. Association of habitual alcohol intake with risk of cardiovascular disease. *JAMA Network Open*. 2022;5(3):e223849-e223849. doi:10.1001/jamanetworkopen.2022.3849
65. Millwood IY, Walters RG, Mei XW, et al. Conventional and genetic evidence on alcohol and vascular disease aetiology: A prospective study of 500 000 men and women in china. *The Lancet*. 2019;393(10183):1831-1842. doi:10.1016/S0140-6736(18)31772-0
66. Holmes MV, Dale CE, Zuccolo L, et al. Association between alcohol and cardiovascular disease: Mendelian randomisation analysis based on individual participant data. *Bmj*. Jul 10 2014;349:g4164. doi:10.1136/bmj.g4164
67. Mukamal KJ, Rimm EB. Alcohol's effects on the risk for coronary heart disease. *Alcohol Res Health*. 2001;25(4):255-61.

68. Hingson RW, Zha W, White AM. Drinking beyond the binge threshold: Predictors, consequences, and changes in the u.S. *American Journal of Preventive Medicine*. 2017/06/01/ 2017;52(6):717-727. doi:https://doi.org/10.1016/j.amepre.2017.02.014
69. Azagba S, Shan L, Latham K, Manzione L. Trends in binge and heavy drinking among adults in the united states, 2011-2017. *Subst Use Misuse*. 2020;55(6):990-997. doi:10.1080/10826084.2020.1717538
70. World Health O. Global status report on alcohol and health 2018. Accessed February 10, 2022. https://www.who.int/publications/i/item/9789241565639
71. Charlet K, Heinz A. Harm reduction-a systematic review on effects of alcohol reduction on physical and mental symptoms. *Addict Biol*. Sep 2017;22(5):1119-1159. doi:10.1111/adb.12414
72. Saitz R, Larson MJ, Labelle C, Richardson J, Samet JH. The case for chronic disease management for addiction. *J Addict Med*. Jun 2008;2(2):55-65. doi:10.1097/ADM.0b013e318166af74
73. Lohoff FW. Targeting unmet clinical needs in the treatment of alcohol use disorder. *Front Psychiatry*. 2022;13:767506. doi:10.3389/fpsyt.2022.767506
74. Bogenschutz MP, Ross S, Bhatt S, et al. Percentage of heavy drinking days following psilocybin-assisted psychotherapy vs placebo in the treatment of adult patients with alcohol use disorder: A randomized clinical trial. *JAMA Psychiatry*. 2022;79(10):953-962. doi:10.1001/jamapsychiatry.2022.2096
75. Witkiewitz K, Kranzler HR, Hallgren KA, et al. Drinking risk level reductions associated with improvements in physical health and quality of life among individuals with alcohol use disorder. *Alcohol Clin Exp Res*. Dec 2018;42(12):2453-2465. doi:10.1111/acer.13897
76. Witkiewitz K, Litten RZ, Leggio L. Advances in the science and treatment of alcohol use disorder. *Science Advances*. 5(9):eaax4043. doi:10.1126/sciadv.aax4043
77. Tawa EA, Hall SD, Lohoff FW. Overview of the genetics of alcohol use disorder. *Alcohol Alcohol*. Sep 2016;51(5):507-14. doi:10.1093/alcalc/agw046
78. Sanchez-Roige S, Palmer AA. Emerging phenotyping strategies will advance our understanding of psychiatric genetics. *Nature Neuroscience*. 2020/04/01 2020;23(4):475-480. doi:10.1038/s41593-020-0609-7
79. Gupta I, Dandavate R, Gupta P, Agrawal V, Kapoor M. Recent advances in genetic studies of alcohol use disorders. *Curr Genet Med Rep*. Jun 2020;8(2):27-34. doi:10.1007/s40142-020-00185-9
80. Kranzler HR, Zhou H, Kember RL, et al. Genome-wide association study of alcohol consumption and use disorder in 274,424 individuals from multiple populations. *Nature Communications*. 2019/04/02 2019;10(1):1499. doi:10.1038/s41467-019-09480-8
81. Liu M, Jiang Y, Wedow R, et al. Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use. *Nat Genet*. Feb 2019;51(2):237-244. doi:10.1038/s41588-018-0307-5
82. Zhou H, Sealock JM, Sanchez-Roige S, et al. Genome-wide meta-analysis of problematic alcohol use in 435,563 individuals yields insights into biology and relationships with other traits. *Nature Neuroscience*. 2020/07/01 2020;23(7):809-818. doi:10.1038/s41593-020-0643-5
83. Walters RK, Polimanti R, Johnson EC, et al. Transancestral gwas of alcohol dependence reveals common genetic underpinnings with psychiatric disorders. *Nat Neurosci*. Dec 2018;21(12):1656-1669. doi:10.1038/s41593-018-0275-1

84. Mallard TT, Savage JE, Johnson EC, et al. Item-level genome-wide association study of the alcohol use disorders identification test in three population-based cohorts. *American Journal of Psychiatry*. 2021:appi.ajp.2020.20091390. doi:10.1176/appi.ajp.2020.20091390
85. Mavromatis LA, Rosoff DB, Cupertino RB, Garavan H, Mackey S, Lohoff FW. Association between brain structure and alcohol use behaviors in adults: A mendelian randomization and multiomics study. *JAMA Psychiatry*. 2022;79(9):869-878. doi:10.1001/jamapsychiatry.2022.2196
86. Goldstein RZ, Volkow ND. Dysfunction of the prefrontal cortex in addiction: Neuroimaging findings and clinical implications. *Nat Rev Neurosci*. Oct 20 2011;12(11):652-69. doi:10.1038/nrn3119
87. Lin Z, Nie C, Zhang Y, Chen Y, Yang T. Evidence accumulation for value computation in the prefrontal cortex during decision making. *Proceedings of the National Academy of Sciences*. 2020/12/01 2020;117(48):30728-30737. doi:10.1073/pnas.2019077117
88. Holmes MV, Richardson TG, Ference BA, Davies NM, Davey Smith G. Integrating genomics with biomarkers and therapeutic targets to invigorate cardiovascular drug development. *Nat Rev Cardiol*. Jun 2021;18(6):435-453. doi:10.1038/s41569-020-00493-1
89. Johnson ECB, Dammer EB, Duong DM, et al. Large-scale proteomic analysis of alzheimer's disease brain and cerebrospinal fluid reveals early changes in energy metabolism associated with microglia and astrocyte activation. *Nature Medicine*. 2020/05/01 2020;26(5):769-780. doi:10.1038/s41591-020-0815-6
90. Wingo AP, Dammer EB, Breen MS, et al. Large-scale proteomic analysis of human brain identifies proteins associated with cognitive trajectory in advanced age. *Nature Communications*. 2019/04/08 2019;10(1):1619. doi:10.1038/s41467-019-09613-z
91. Wingo TS, Gerasimov ES, Liu Y, et al. Integrating human brain proteomes with genome-wide association data implicates novel proteins in post-traumatic stress disorder. *Mol Psychiatry*. Jul 2022;27(7):3075-3084. doi:10.1038/s41380-022-01544-4
92. Pathak GA, Singh K, Wendt FR, et al. Genetically regulated multi-omics study for symptom clusters of posttraumatic stress disorder highlights pleiotropy with hematologic and cardio-metabolic traits. *Molecular Psychiatry*. 2022/03/01 2022;27(3):1394-1404. doi:10.1038/s41380-022-01488-9
93. Liu J, Li X, Luo XJ. Proteome-wide association study provides insights into the genetic component of protein abundance in psychiatric disorders. *Biol Psychiatry*. Dec 1 2021;90(11):781-789. doi:10.1016/j.biopsych.2021.06.022
94. Marees AT, Smit DJA, Ong JS, et al. Potential influence of socioeconomic status on genetic correlations between alcohol consumption measures and mental health. *Psychol Med*. Feb 2020;50(3):484-498. doi:10.1017/s0033291719000357
95. Kapoor M, Chao MJ, Johnson EC, et al. Multi-omics integration analysis identifies novel genes for alcoholism with potential overlap with neurodegenerative diseases. *Nature Communications*. 2021/08/20 2021;12(1):5071. doi:10.1038/s41467-021-25392-y
96. Neavin D, Nguyen Q, Daniszewski MS, et al. Single cell eqtl analysis identifies cell type-specific genetic control of gene expression in fibroblasts and reprogrammed induced pluripotent stem cells. *Genome Biology*. 2021/03/05 2021;22(1):76. doi:10.1186/s13059-021-02293-3
97. Li M, Santpere G, Imamura Kawasawa Y, et al. Integrative functional genomic analysis of human brain development and neuropsychiatric risks. *Science*. Dec 14 2018;362(6420)doi:10.1126/science.aat7615

98. Trevino AE, Müller F, Andersen J, et al. Chromatin and gene-regulatory dynamics of the developing human cerebral cortex at single-cell resolution. *Cell*. Sep 16 2021;184(19):5053-5069.e23. doi:10.1016/j.cell.2021.07.039
99. Schwartzentruber J, Cooper S, Liu JZ, et al. Genome-wide meta-analysis, fine-mapping and integrative prioritization implicate new alzheimer's disease risk genes. *Nature Genetics*. 2021/03/01 2021;53(3):392-402. doi:10.1038/s41588-020-00776-w
100. Walker RL, Ramaswami G, Hartl C, et al. Genetic control of expression and splicing in developing human brain informs disease mechanisms. *Cell*. Oct 17 2019;179(3):750-771.e22. doi:10.1016/j.cell.2019.09.021
101. Erickson EK, Grantham EK, Warden AS, Harris RA. Neuroimmune signaling in alcohol use disorder. *Pharmacol Biochem Behav*. Feb 2019;177:34-60. doi:10.1016/j.pbb.2018.12.007
102. Bryois J, Calini D, Macnair W, et al. Cell-type-specific cis-eqtls in eight human brain cell types identify novel risk genes for psychiatric and neurological disorders. *Nat Neurosci*. Aug 2022;25(8):1104-1112. doi:10.1038/s41593-022-01128-z
103. Gusev A, Ko A, Shi H, et al. Integrative approaches for large-scale transcriptome-wide association studies. *Nat Genet*. Mar 2016;48(3):245-52. doi:10.1038/ng.3506
104. Lim Y, Beane-Ebel JE, Tanaka Y, et al. Exploration of alcohol use disorder-associated brain mirna–mrna regulatory networks. *Translational Psychiatry*. 2021/10/02 2021;11(1):504. doi:10.1038/s41398-021-01635-w
105. Lohoff FW, Clarke T-K, Kaminsky ZA, et al. Epigenome-wide association study of alcohol consumption in n = 8161 individuals and relevance to alcohol use disorder pathophysiology: Identification of the cystine/glutamate transporter slc7a11 as a top target. *Molecular Psychiatry*. 2022/03/01 2022;27(3):1754-1764. doi:10.1038/s41380-021-01378-6
106. Giambartolomei C, Vukcevic D, Schadt EE, Franke L, Hingorani AD, Wallace C, Plagnol V. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet*. May 2014;10(5):e1004383. doi:10.1371/journal.pgen.1004383
107. Ghodsian N, Abner E, Emdin CA, et al. Electronic health record-based genome-wide meta-analysis provides insights on the genetic architecture of non-alcoholic fatty liver disease. *Cell Rep Med*. Nov 16 2021;2(11):100437. doi:10.1016/j.xcrm.2021.100437
108. Sumida Y, Yoneda M. Current and future pharmacological therapies for nafld/nash. *J Gastroenterol*. Mar 2018;53(3):362-376. doi:10.1007/s00535-017-1415-1
109. Hsu CL, Loomba R. From nafld to masld: Implications of the new nomenclature for preclinical and clinical research. *Nature Metabolism*. 2024/04/01 2024;6(4):600-602. doi:10.1038/s42255-024-00985-1
110. Estes C, Razavi H, Loomba R, Younossi Z, Sanyal AJ. Modeling the epidemic of nonalcoholic fatty liver disease demonstrates an exponential increase in burden of disease. *Hepatology*. Jan 2018;67(1):123-133. doi:10.1002/hep.29466
111. Pais R, Barritt ASt, Calmus Y, et al. Nafld and liver transplantation: Current burden and expected challenges. *J Hepatol*. Dec 2016;65(6):1245-1257. doi:10.1016/j.jhep.2016.07.033
112. Cao L, An Y, Liu H, et al. Global epidemiology of type 2 diabetes in patients with nafld or mafld: A systematic review and meta-analysis. *BMC Med*. Mar 6 2024;22(1):101. doi:10.1186/s12916-024-03315-0
113. Younossi ZM, Golabi P, Price JK, Owrangi S, Gundu-Rao N, Satchi R, Paik JM. The global epidemiology of nonalcoholic fatty liver disease and nonalcoholic steatohepatitis among patients with type 2 diabetes. *Clinical Gastroenterology and Hepatology*. 2024/10/01/ 2024;22(10):1999-2010.e8. doi:https://doi.org/10.1016/j.cgh.2024.03.006

114. Xia MF, Bian H, Gao X. Nafld and diabetes: Two sides of the same coin? Rationale for gene-based personalized nafld treatment. *Front Pharmacol*. 2019;10:877. doi:10.3389/fphar.2019.00877
115. Deprince A, Haas JT, Staels B. Dysregulated lipid metabolism links nafld to cardiovascular disease. *Mol Metab*. Dec 2020;42:101092. doi:10.1016/j.molmet.2020.101092
116. Matyas C, Haskó G, Liaudet L, Trojnar E, Pacher P. Interplay of cardiovascular mediators, oxidative stress and inflammation in liver disease and its complications. *Nat Rev Cardiol*. Feb 2021;18(2):117-135. doi:10.1038/s41569-020-0433-5
117. Martin S, Sorokin EP, Thomas EL, Sattar N, Cule M, Bell JD, Yaghootkar H. Estimating the effect of liver and pancreas volume and fat content on risk of diabetes: A mendelian randomization study. *Diabetes Care*. Feb 1 2022;45(2):460-468. doi:10.2337/dc21-1262
118. Ren Z, Simons PIHG, Wesselius A, Stehouwer CDA, Brouwers MCGJ. Relationship between nafld and coronary artery disease: A mendelian randomization study. https://doi.org/10.1002/hep.32534. *Hepatology*. 2022/04/20 2022;n/a(n/a)doi:https://doi.org/10.1002/hep.32534
119. Karlsson Linnér R, Mallard TT, Barr PB, et al. Multivariate analysis of 1.5 million people identifies genetic associations with traits related to self-regulation and addiction. *Nat Neurosci*. Oct 2021;24(10):1367-1376. doi:10.1038/s41593-021-00908-3
120. Wallace C. Eliciting priors and relaxing the single causal variant assumption in colocalisation analyses. *PLOS Genetics*. 2020;16(4):e1008720. doi:10.1371/journal.pgen.1008720
121. Finucane HK, Bulik-Sullivan B, Gusev A, et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nature Genetics*. 2015/11/01 2015;47(11):1228-1235. doi:10.1038/ng.3404
122. Timshel PN, Thompson JJ, Pers TH. Genetic mapping of etiologic brain cell types for obesity. *Elife*. Sep 21 2020;9doi:10.7554/eLife.55851
123. Harrison SA, Bedossa P, Guy CD, et al. A phase 3, randomized, controlled trial of resmetirom in nash with liver fibrosis. *N Engl J Med*. Feb 8 2024;390(6):497-509. doi:10.1056/NEJMoa2309000
124. Harrison SA, Ruane PJ, Freilich BL, et al. Efruxifermin in non-alcoholic steatohepatitis: A randomized, double-blind, placebo-controlled, phase 2a trial. *Nature Medicine*. 2021/07/01 2021;27(7):1262-1271. doi:10.1038/s41591-021-01425-3
125. Loomba R, Sanyal AJ, Kowdley KV, et al. Randomized, controlled trial of the fgf21 analogue pegozafermin in nash. *New England Journal of Medicine*. 2023;389(11):998-1008. doi:10.1056/NEJMoa2304286
126. Finan C, Gaulton A, Kruger FA, et al. The druggable genome and support for target identification and validation in drug development. *Sci Transl Med*. Mar 29 2017;9(383)doi:10.1126/scitranslmed.aag1166
127. Valenzuela PL, Carrera-Bastos P, Castillo-García A, Lieberman DE, Santos-Lozano A, Lucia A. Obesity and the risk of cardiometabolic diseases. *Nature Reviews Cardiology*. 2023/07/01 2023;20(7):475-494. doi:10.1038/s41569-023-00847-5
128. Afshin A, Forouzanfar MH, Reitsma MB, et al. Health effects of overweight and obesity in 195 countries over 25 years. *N Engl J Med*. Jul 6 2017;377(1):13-27. doi:10.1056/NEJMoa1614362
129. Lopez-Jimenez F, Almahmeed W, Bays H, et al. Obesity and cardiovascular disease: Mechanistic insights and management strategies. A joint position paper by the world heart

federation and world obesity federation. *European Journal of Preventive Cardiology*. 2022;29(17):2218-2237. doi:10.1093/eurjpc/zwac187
130. Goudswaard LJ, Bell JA, Hughes DA, et al. Effects of adiposity on the human plasma proteome: Observational and mendelian randomisation estimates. *International Journal of Obesity*. 2021/10/01 2021;45(10):2221-2229. doi:10.1038/s41366-021-00896-1
131. Zaghlool SB, Sharma S, Molnar M, et al. Revealing the role of the human blood plasma proteome in obesity using genetic drivers. *Nature Communications*. 2021/02/24 2021;12(1):1279. doi:10.1038/s41467-021-21542-4
132. Relton CL, Davey Smith G. Two-step epigenetic mendelian randomization: A strategy for establishing the causal role of epigenetic processes in pathways to disease. *Int J Epidemiol*. Feb 2012;41(1):161-76. doi:10.1093/ije/dyr233
133. Pulit SL, Stoneman C, Morris AP, et al. Meta-analysis of genome-wide association studies for body fat distribution in 694 649 individuals of european ancestry. *Hum Mol Genet*. Jan 1 2019;28(1):166-174. doi:10.1093/hmg/ddy327
134. Sun BB, Chiou J, Traylor M, et al. Plasma proteomic associations with genetics and health in the uk biobank. *Nature*. 2023/10/01 2023;622(7982):329-338. doi:10.1038/s41586-023-06592-6
135. Surdo PL, Bottomley MJ, Calzetta A, et al. Mechanistic implications for ldl receptor degradation from the pcsk9/ldlr structure at neutral ph. *EMBO reports*. 2011;12(12):1300-1305. doi:https://doi.org/10.1038/embor.2011.205
136. Poirier S, Mayer G, Benjannet S, et al. The proprotein convertase pcsk9 induces the degradation of low density lipoprotein receptor (ldlr) and its closest family members vldlr and apoer2*. *Journal of Biological Chemistry*. 2008/01/25/ 2008;283(4):2363-2372. doi:https://doi.org/10.1074/jbc.M708098200
137. Benjannet S, Rhainds D, Essalmani R, et al. Narc-1/pcsk9 and its natural mutants: Zymogen cleavage and effects on the low density lipoprotein (ldl) receptor and ldl cholesterol *. *Journal of Biological Chemistry*. 2004;279(47):48865-48875. doi:10.1074/jbc.M409699200
138. Seidah NG, Benjannet S, Wickham L, et al. The secretory proprotein convertase neural apoptosis-regulated convertase 1 (narc-1): Liver regeneration and neuronal differentiation. *Proceedings of the National Academy of Sciences*. 2003;100(3):928-933. doi:10.1073/pnas.0335507100
139. Sabatine MS. Pcsk9 inhibitors: Clinical evidence and implementation. *Nat Rev Cardiol*. 2019/03/01 2019;16(3):155-165. doi:10.1038/s41569-018-0107-8
140. Mullard A. Nine paths to pcsk9 inhibition. *Nature Reviews Drug Discovery*. 2017/05/01 2017;16(5):299-301. doi:10.1038/nrd.2017.83
141. Shapiro MD, Tavori H, Fazio S. Pcsk9. *Circ Res*. 2018;122(10):1420-1438. doi:doi:10.1161/CIRCRESAHA.118.311227
142. Raal F, Scott R, Somaratne R, Bridges I, Li G, Wasserman SM, Stein EA. Low-density lipoprotein cholesterol-lowering effects of amg 145, a monoclonal antibody to proprotein convertase subtilisin/kexin type 9 serine protease in patients with heterozygous familial hypercholesterolemia: The reduction of ldl-c with pcsk9 inhibition in heterozygous familial hypercholesterolemia disorder (rutherford) randomized trial. *Circulation*. Nov 13 2012;126(20):2408-17. doi:10.1161/circulationaha.112.144055
143. Robinson JG, Farnier M, Krempf M, et al. Efficacy and safety of alirocumab in reducing lipids and cardiovascular events. *N Engl J Med*. Apr 16 2015;372(16):1489-99. doi:10.1056/NEJMoa1501031

144. Sabatine MS, Giugliano RP, Keech AC, et al. Evolocumab and clinical outcomes in patients with cardiovascular disease. *N Engl J Med*. May 4 2017;376(18):1713-1722. doi:10.1056/NEJMoa1615664
145. Toth PP, Descamps O, Genest J, et al. Pooled safety analysis of evolocumab in over 6000 patients from double-blind and open-label extension studies. *Circulation*. May 9 2017;135(19):1819-1831. doi:10.1161/circulationaha.116.025233
146. Administratio UFD. Fda approves add-on therapy to lower cholesterol among certain high-risk adults. Accessed March 27, 2022. https://www.fda.gov/drugs/news-events-human-drugs/fda-approves-add-therapy-lower-cholesterol-among-certain-high-risk-adults
147. Leiter LA, Teoh H, Kallend D, et al. Inclisiran lowers ldl-c and pcsk9 irrespective of diabetes status: The orion-1 randomized clinical trial. *Diabetes Care*. 2019;42(1):173-176. doi:10.2337/dc18-1491
148. Raal FJ, Kallend D, Ray KK, et al. Inclisiran for the treatment of heterozygous familial hypercholesterolemia. *N Engl J Med*. 2020;382(16):1520-1530. doi:10.1056/NEJMoa1913805
149. Ray KK, Wright RS, Kallend D, et al. Two phase 3 trials of inclisiran in patients with elevated ldl cholesterol. *New England Journal of Medicine*. 2020;382(16):1507-1519. doi:10.1056/NEJMoa1912387
150. Ray S, Jindal AK, Sengupta S, Sinha S. Statins: Can we advocate them for primary prevention of heart disease? *Med J Armed Forces India*. 2014/07/01/ 2014;70(3):270-273. doi:https://doi.org/10.1016/j.mjafi.2013.05.008
151. Reiner Ž. Statins in the primary prevention of cardiovascular disease. *Nat Rev Cardiol*. 2013/08/01 2013;10(8):453-464. doi:10.1038/nrcardio.2013.80
152. Guella I, Asselta R, Ardissino D, et al. Effects of pcsk9 genetic variants on plasma ldl cholesterol levels and risk of premature myocardial infarction in the italian population. *Journal of Lipid Research*. 2010/11/01/ 2010;51(11):3342-3349. doi:https://doi.org/10.1194/jlr.M010009
153. Lagace TA. Pcsk9 and ldlr degradation: Regulatory mechanisms in circulation and in cells. *Curr Opin Lipidol*. Oct 2014;25(5):387-93. doi:10.1097/mol.0000000000000114
154. Ference Brian A, Majeed F, Penumetcha R, Flack John M, Brook Robert D. Effect of naturally random allocation to lower low-density lipoprotein cholesterol on the risk of coronary heart disease mediated by polymorphisms in npc1l1, hmgcr, or both. *Journal of the American College of Cardiology*. 2015/04/21 2015;65(15):1552-1561. doi:10.1016/j.jacc.2015.02.020
155. Lohoff FW, Sorcher JL, Rosen AD, et al. Methylomic profiling and replication implicates deregulation of pcsk9 in alcohol use disorder. *Mol Psychiatry*. Sep 2018;23(9):1900-1910. doi:10.1038/mp.2017.168
156. Lee JS, Rosoff D, Luo A, et al. Pcsk9 is increased in cerebrospinal fluid of individuals with alcohol use disorder. *Alcohol Clin Exp Res*. Jun 2019;43(6):1163-1169. doi:10.1111/acer.14039
157. Bell AS, Wagner J, Rosoff DB, Lohoff FW. Proprotein convertase subtilisin/kexin type 9 (pcsk9) in the central nervous system. *Neurosci Biobehav Rev*. Jun 2023;149:105155. doi:10.1016/j.neubiorev.2023.105155
158. Chiang LW, Grenier JM, Ettwiller L, et al. An orchestrated gene expression component of neuronal programmed cell death revealed by cdna array analysis. *Proceedings of the National Academy of Sciences*. 2001;98(5):2814-2819. doi:10.1073/pnas.051630598
159. Samaras K, Makkar SR, Crawford JD, et al. Effects of statins on memory, cognition, and brain volume in the elderly. *J Am Coll Cardiol*. Nov 26 2019;74(21):2554-2568. doi:10.1016/j.jacc.2019.09.041

160. Bradley CK, Wang TY, Li S, et al. Patient-reported reasons for declining or discontinuing statin therapy: Insights from the palm registry. *J Am Heart Assoc*. 2019/04/02 2019;8(7):e011765. doi:10.1161/JAHA.118.011765
161. Gill D, Georgakis MK, Walker VM, et al. Mendelian randomization for studying the effects of perturbing drug targets. *Wellcome Open Res*. 2021;6:16. doi:10.12688/wellcomeopenres.16544.2
162. Rosoff DB, Clarke T-K, Adams MJ, McIntosh AM, Davey Smith G, Jung J, Lohoff FW. Educational attainment impacts drinking behaviors and risk for alcohol dependence: Results from a two-sample mendelian randomization study with ~780,000 participants. *Molecular Psychiatry*. 2021/04/01 2021;26(4):1119-1132. doi:10.1038/s41380-019-0535-9
163. Rosoff DB, Kaminsky ZA, McIntosh AM, Davey Smith G, Lohoff FW. Educational attainment reduces the risk of suicide attempt among individuals with and without psychiatric disorders independent of cognition: A bidirectional and multivariable mendelian randomization study with more than 815,000 participants. *Transl Psychiatry*. Nov 9 2020;10(1):388. doi:10.1038/s41398-020-01047-2
164. Rosoff DB, Bell AS, Mavromatis LA, et al. Evaluating the cardiovascular impact of genetically proxied pcsk9 and hmgcr inhibition in east asian and european populations: A drug-target mendelian randomization study. *Circulation: Genomic and Precision Medicine*. 2024/02/01 2024;17(1):e004224. doi:10.1161/CIRCGEN.122.004224
165. Rosoff DB, Yoo J, Lohoff FW. Smoking is significantly associated with increased risk of covid-19 and other respiratory infections. *Commun Biol*. Oct 28 2021;4(1):1230. doi:10.1038/s42003-021-02685-y
166. Rosoff DB, Smith GD, Lohoff FW. Prescription opioid use and risk for major depressive disorder and anxiety and stress-related disorders: A multivariable mendelian randomization analysis. *JAMA Psychiatry*. Feb 1 2021;78(2):151-160. doi:10.1001/jamapsychiatry.2020.3554
167. Rosoff Daniel B, Bell Andrew S, Jung J, Wagner J, Mavromatis Lucas A, Lohoff Falk W. Mendelian randomization study of pcsk9 and hmg-coa reductase inhibition and cognitive function. *Journal of the American College of Cardiology*. 2022/08/16 2022;80(7):653-662. doi:10.1016/j.jacc.2022.05.041
168. Riecher-Rössler A. Sex and gender differences in mental disorders. *The Lancet Psychiatry*. 2017;4(1):8-9. doi:10.1016/S2215-0366(16)30348-0
169. Bogren M, Brådvik L, Holmstrand C, Nöbbelin L, Mattisson C. Gender differences in subtypes of depression by first incidence and age of onset: A follow-up of the lundby population. *European Archives of Psychiatry and Clinical Neuroscience*. 2018/03/01 2018;268(2):179-189. doi:10.1007/s00406-017-0778-x
170. Eid RS, Gobinath AR, Galea LAM. Sex differences in depression: Insights from clinical and preclinical studies. *Progress in Neurobiology*. 2019/05/01/ 2019;176:86-102. doi:https://doi.org/10.1016/j.pneurobio.2019.01.006
171. Marcus SM, Kerber KB, Rush AJ, et al. Sex differences in depression symptoms in treatment-seeking adults: Confirmatory analyses from the sequenced treatment alternatives to relieve depression study. *Comprehensive Psychiatry*. 2008/05/01/ 2008;49(3):238-246. doi:https://doi.org/10.1016/j.comppsych.2007.06.012
172. Helton SG, Lohoff FW. Serotonin pathway polymorphisms and the treatment of major depressive disorder and anxiety disorders. *Pharmacogenomics*. 2015;16(5):541-53. doi:10.2217/pgs.15.15

173. Bell AS, Rosoff DB, Mavromatis LA, Jung J, Wagner J, Lohoff FW. Comparing the relationships of genetically proxied pcsk9 inhibition with mood disorders, cognition, and dementia between men and women: A drug-target mendelian randomization study. *Journal of the American Heart Association*. 2022/11/01 2022;11(21):e026122. doi:10.1161/JAHA.122.026122
174. Mannarino MR, Sahebkar A, Bianconi V, Serban MC, Banach M, Pirro M. Pcsk9 and neurocognitive function: Should it be still an issue after fourier and ebbinghaus results? *J Clin Lipidol*. Sep-Oct 2018;12(5):1123-1132. doi:10.1016/j.jacl.2018.05.012
175. Giugliano RP, Mach F, Zavitz K, et al. Cognitive function in a randomized trial of evolocumab. *N Engl J Med*. 2017;377(7):633-643. doi:10.1056/NEJMoa1701131
176. Lyall DM, Ward J, Banach M, et al. Pcsk9 genetic variants and cognitive abilities: A large-scale mendelian randomization study. journal article. *Arch Med Sci*. 2021;17(1):241-244. doi:10.5114/aoms/127226
177. Rosoff DB, Charlet K, Jung J, et al. Association of high-intensity binge drinking with lipid and liver function enzyme levels. *JAMA Network Open*. 2019;2(6):e195844-e195844. doi:10.1001/jamanetworkopen.2019.5844
178. Piano MR. Alcohol's effects on the cardiovascular system. *Alcohol Res*. 2017;38(2):219-241.
179. Lawlor DA, Harbord RM, Sterne JAC, Timpson N, Davey Smith G. Mendelian randomization: Using genes as instruments for making causal inferences in epidemiology. *Statistics in Medicine*. 2008/04/15 2008;27(8):1133-1163. doi:10.1002/sim.3034
180. Rimm EB, Williams P, Fosher K, Criqui M, Stampfer MJ. Moderate alcohol intake and lower risk of coronary heart disease: Meta-analysis of effects on lipids and haemostatic factors. *Bmj*. Dec 11 1999;319(7224):1523-8. doi:10.1136/bmj.319.7224.1523
181. NIH. Nih to end funding for moderate alcohol and cardiovascular health trial. June 15, 2018, Accessed June 21, 2019. https://www.nih.gov/news-events/news-releases/nih-end-funding-moderate-alcohol-cardiovascular-health-trial
182. Lawlor DA, Nordestgaard BG, Benn M, Zuccolo L, Tybjaerg-Hansen A, Davey Smith G. Exploring causal associations between alcohol and coronary heart disease risk factors: Findings from a mendelian randomization study in the copenhagen general population study. *European heart journal*. Aug 2013;34(32):2519-28. doi:10.1093/eurheartj/eht081
183. Holmes MV, Dale CE, Zuccolo L, et al. Association between alcohol and cardiovascular disease: Mendelian randomisation analysis based on individual participant data. *BMJ : British Medical Journal*. 2014;349:g4164. doi:10.1136/bmj.g4164
184. Vu KN, Ballantyne CM, Hoogeveen RC, Nambi V, Volcik KA, Boerwinkle E, Morrison AC. Causal role of alcohol consumption in an improved lipid profile: The atherosclerosis risk in communities (aric) study. *PLoS One*. 2016;11(2):e0148765-e0148765. doi:10.1371/journal.pone.0148765
185. Millwood IY, Walters RG, Mei XW, et al. Conventional and genetic evidence on alcohol and vascular disease aetiology: A prospective study of 500 000 men and women in china. *The Lancet*. 2019/05/04/ 2019;393(10183):1831-1842. doi:https://doi.org/10.1016/S0140-6736(18)31772-0
186. Patten CA, Martin JE, Owen N. Can psychiatric and chemical dependency treatment units be smoke free? *Journal of Substance Abuse Treatment*. 1996;13(2):107-118. doi:10.1016/0740-5472(96)00040-2

187. Touchette JC, Lee AM. Assessing alcohol and nicotine co-consumption in mice. *Oncotarget*. 2017;8(4):5684-5685. doi:10.18632/oncotarget.14603
188. Marees AT, Smit DJA, Ong JS, et al. Potential influence of socioeconomic status on genetic correlations between alcohol consumption measures and mental health. *Psychol Med*. Mar 15 2019:1-15. doi:10.1017/s0033291719000357
189. Karlsson Linnér R, Biroli P, Kong E, et al. Genome-wide association analyses of risk tolerance and risky behaviors in over 1 million individuals identify hundreds of loci and shared genetic influences. *Nature Genetics*. 2019/02/01 2019;51(2):245-257. doi:10.1038/s41588-018-0309-3
190. Sanderson E, Davey Smith G, Windmeijer F, Bowden J. An examination of multivariable mendelian randomization in the single-sample and two-sample summary data settings. *International Journal of Epidemiology*. 2018;48(3):713-727. doi:10.1093/ije/dyy262
191. Sanderson E. Multivariable mendelian randomization and mediation. *Cold Spring Harb Perspect Med*. Feb 1 2021;11(2)doi:10.1101/cshperspect.a038984
192. Rosoff DB, Davey Smith G, Mehta N, Clarke T-K, Lohoff FW. Evaluating the relationship between alcohol consumption, tobacco use, and cardiovascular disease: A multivariable mendelian randomization study. *PLOS Medicine*. 2020;17(12):e1003410. doi:10.1371/journal.pmed.1003410
193. Larsson SC, Burgess S, Michaëlsson K. Smoking and stroke: A mendelian randomization study. *Annals of Neurology*. 2019;86(3):468-471. doi:https://doi.org/10.1002/ana.25534
194. Larsson SC, Mason AM, Bäck M, Klarin D, Damrauer SM, Michaëlsson K, Burgess S. Genetic predisposition to smoking in relation to 14 cardiovascular diseases. *Eur Heart J*. Sep 14 2020;41(35):3304-3310. doi:10.1093/eurheartj/ehaa193
195. Rosoff DB, Hamandi AM, Bell AS, et al. Major psychiatric disorders, substance use behaviors, and longevity. *JAMA Psychiatry*. Sep 1 2024;81(9):889-901. doi:10.1001/jamapsychiatry.2024.1429
196. Rosoff DB, Mavromatis LA, Bell AS, et al. Multivariate genome-wide analysis of aging-related traits identifies novel loci and new drug targets for healthy aging. *Nature Aging*. 2023/08/01 2023;3(8):1020-1035. doi:10.1038/s43587-023-00455-5
197. Larsson SC, Butterworth AS, Burgess S. Mendelian randomization for cardiovascular diseases: Principles and applications. *European Heart Journal*. 2023;44(47):4913-4924. doi:10.1093/eurheartj/ehad736
198. Davies NM, Holmes MV, Smith GD. Reading mendelian randomisation studies: A guide, glossary, and checklist for clinicians. *Bmj*. 2018;362
199. Lawlor DA, Harbord RM, Sterne JA, Timpson N, Davey Smith G. Mendelian randomization: Using genes as instruments for making causal inferences in epidemiology. *Stat Med*. Apr 15 2008;27(8):1133-63. doi:10.1002/sim.3034
200. Henry A, Gordillo-Marañón M, Finan C, et al. Therapeutic targets for heart failure identified using proteomics and mendelian randomization. *Circulation*. Apr 19 2022;145(16):1205-1217. doi:10.1161/circulationaha.121.056663
201. Graham SE, Clarke SL, Wu KH, et al. The power of genetic diversity in genome-wide association studies of lipids. *Nature*. Dec 2021;600(7890):675-679. doi:10.1038/s41586-021-04064-3
202. Mahajan A, Spracklen CN, Zhang W, et al. Multi-ancestry genetic study of type 2 diabetes highlights the power of diverse populations for discovery and translation. *Nat Genet*. May 2022;54(5):560-572. doi:10.1038/s41588-022-01058-3

203. Wojcik GL, Graff M, Nishimura KK, et al. Genetic analyses of diverse populations improves discovery for complex traits. *Nature*. Jun 2019;570(7762):514-518. doi:10.1038/s41586-019-1310-4
204. Chen J, Sun M, Adeyemo A, et al. Genome-wide association study of type 2 diabetes in africa. *Diabetologia*. Jul 2019;62(7):1204-1211. doi:10.1007/s00125-019-4880-7
205. Mooradian AD. Dyslipidemia in type 2 diabetes mellitus. *Nat Clin Pract Endocrinol Metab*. Mar 2009;5(3):150-9. doi:10.1038/ncpendmet1066
206. Williamson A, Norris DM, Yin X, et al. Genome-wide association study and functional characterization identifies candidate genes for insulin-stimulated glucose uptake. *Nat Genet*. Jun 2023;55(6):973-983. doi:10.1038/s41588-023-01408-9
207. Auton A, Abecasis GR, Altshuler DM, et al. A global reference for human genetic variation. *Nature*. 2015/10/01 2015;526(7571):68-74. doi:10.1038/nature15393
208. Qiu C, Zeng P, Li X, et al. What is the impact of pcsk9 rs505151 and rs11591147 polymorphisms on serum lipids level and cardiovascular risk: A meta-analysis. *Lipids Health Dis*. Jun 12 2017;16(1):111. doi:10.1186/s12944-017-0506-6
209. Mihaylova B, Emberson J, Blackwell L, et al. The effects of lowering ldl cholesterol with statin therapy in people at low risk of vascular disease: Meta-analysis of individual data from 27 randomised trials. *Lancet*. Aug 11 2012;380(9841):581-90. doi:10.1016/s0140-6736(12)60367-5
210. Consortium G. Human genomics. The genotype-tissue expression (gtex) pilot analysis: Multitissue gene regulation in humans. *Science*. May 8 2015;348(6235):648-60. doi:10.1126/science.1262110
211. de Klein N, Tsai EA, Vochteloo M, et al. Brain expression quantitative trait locus and network analysis reveals downstream effects and putative drivers for brain-related diseases. *bioRxiv*. 2021:2021.03.01.433439. doi:10.1101/2021.03.01.433439
212. Hemani G, Tilling K, Davey Smith G. Orienting the causal relationship between imprecisely measured traits using gwas summary data. *PLOS Genetics*. 2017;13(11):e1007081. doi:10.1371/journal.pgen.1007081
213. Smith GD, Ebrahim S. 'Mendelian randomization': Can genetic epidemiology contribute to understanding environmental determinants of disease? *Int J Epidemiol*. Feb 2003;32(1):1-22. doi:10.1093/ije/dyg070
214. Hemani G, Zheng J, Elsworth B, et al. The mr-base platform supports systematic causal inference across the human phenome. *eLife*. 2018/05/30 2018;7:e34408. doi:10.7554/eLife.34408
215. Davey Smith G, Hemani G. Mendelian randomization: Genetic anchors for causal inference in epidemiological studies. *Hum Mol Genet*. Sep 15 2014;23(R1):R89-98. doi:10.1093/hmg/ddu328
216. Bowden J, Del Greco MF, Minelli C, Smith GD, Sheehan N, Thompson J. A framework for the investigation of pleiotropy in two-sample summary data mendelian randomization. *Statistics in Medicine*. May 20 2017;36(11):1783-1802. doi:10.1002/sim.7221
217. Bowden J, Del Greco MF, Minelli C, et al. Improving the accuracy of two-sample summary-data mendelian randomization: Moving beyond the nome assumption. *Int J Epidemiol*. Jun 1 2019;48(3):728-742. doi:10.1093/ije/dyy258
218. Rees JMB, Wood AM, Dudbridge F, Burgess S. Robust methods in mendelian randomization via penalization of heterogeneous causal estimates. *PLoS One*. 2019;14(9):e0222362. doi:10.1371/journal.pone.0222362

219. Aschard H, Vilhjálmsson BJ, Joshi AD, Price AL, Kraft P. Adjusting for heritable covariates can bias effect estimates in genome-wide association studies. *Am J Hum Genet*. Feb 5 2015;96(2):329-39. doi:10.1016/j.ajhg.2014.12.021
220. Hartwig FP, Tilling K, Davey Smith G, Lawlor DA, Borges MC. Bias in two-sample mendelian randomization when using heritable covariable-adjusted summary associations. *International Journal of Epidemiology*. 2021;50(5):1639-1650. doi:10.1093/ije/dyaa266
221. Gilbody J, Borges MC, Smith GD, Sanderson E. Multivariable mr can mitigate bias in two-sample mr using covariable-adjusted summary associations. *medRxiv*. 2022:2022.07.19.22277803. doi:10.1101/2022.07.19.22277803
222. Chen CY, Chen TT, Feng YA, et al. Analysis across taiwan biobank, biobank japan, and uk biobank identifies hundreds of novel loci for 36 quantitative traits. *Cell Genom*. Dec 13 2023;3(12):100436. doi:10.1016/j.xgen.2023.100436
223. Neale-Lab. Uk biobank gwas. Accessed June 2019. http://www.nealelab.is/uk-biobank/
224. Fernández-Rhodes L, Graff M, Buchanan VL, et al. Ancestral diversity improves discovery and fine-mapping of genetic loci for anthropometric traits-the hispanic/latino anthropometry consortium. *HGG Adv*. Apr 14 2022;3(2):100099. doi:10.1016/j.xhgg.2022.100099
225. Amrhein V, Greenland S, McShane B. Scientists rise up against statistical significance. *Nature*. Mar 2019;567(7748):305-307. doi:10.1038/d41586-019-00857-9
226. Liu M, Jiang Y, Wedow R, et al. Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use. *Nature Genetics*. 2019/02/01 2019;51(2):237-244. doi:10.1038/s41588-018-0307-5
227. Hatoum AS, Johnson EC, Agrawal A, Bogdan R. Brain structure and problematic alcohol use: A test of plausible causation using latent causal variable analysis. *Brain Imaging Behav*. Dec 2021;15(6):2741-2745. doi:10.1007/s11682-021-00482-z
228. Wingo TS, Liu Y, Gerasimov ES, et al. Shared mechanisms across the major psychiatric and neurodegenerative diseases. *Nat Commun*. Jul 26 2022;13(1):4314. doi:10.1038/s41467-022-31873-5
229. Cartas-Cejudo P, Cortés A, Lachén-Montes M, et al. Mapping the human brain proteome: Opportunities, challenges, and clinical potential. *Expert Review of Proteomics*. 2024/03/03 2024;21(1-3):55-63. doi:10.1080/14789450.2024.2313073
230. Toikumo S, Xu H, Gelernter J, Kember RL, Kranzler HR. Integrating human brain proteomic data with genome-wide association study findings identifies novel brain proteins in substance use traits. *Neuropsychopharmacology*. Aug 8 2022;doi:10.1038/s41386-022-01406-1
231. Grasby KL, Jahanshad N, Painter JN, et al. The genetic architecture of the human cerebral cortex. *Science*. 2020;367(6484):eaay6690. doi:10.1126/science.aay6690
232. Chen C-H, Fiecas M, Gutiérrez ED, et al. Genetic topography of brain morphology. *Proceedings of the National Academy of Sciences*. 2013;110(42):17089-17094. doi:doi:10.1073/pnas.1308091110
233. Satizabal CL, Adams HHH, Hibar DP, et al. Genetic architecture of subcortical brain structures in 38,851 individuals. *Nature Genetics*. 2019/11/01 2019;51(11):1624-1636. doi:10.1038/s41588-019-0511-y
234. Grasby KL, Jahanshad N, Painter JN, et al. The genetic architecture of the human cerebral cortex. *Science*. Mar 20 2020;367(6484)doi:10.1126/science.aay6690

235. Hibar DP, Adams HHH, Jahanshad N, et al. Novel genetic loci associated with hippocampal volume. *Nature Communications*. 2017/01/18 2017;8(1):13624. doi:10.1038/ncomms13624
236. Zhao B, Li T, Yang Y, et al. Common genetic variation influencing human white matter microstructure. *Science*. 2021;372(6548):eabf3736. doi:doi:10.1126/science.abf3736
237. Gordillo-Marañón M, Zwierzyna M, Charoen P, et al. Validation of lipid-related therapeutic targets for coronary heart disease prevention using human genetics. *Nature Communications*. 2021/10/21 2021;12(1):6120. doi:10.1038/s41467-021-25731-z
238. Schmidt AF, Bourfiss M, Alasiri A, et al. Druggable proteins influencing cardiac structure and function: Implications for heart failure therapies and cancer cardiotoxicity. *Science Advances*. 2023;9(17):eadd4984. doi:doi:10.1126/sciadv.add4984
239. Prapiadou S, Živković L, Thorand B, et al. Proteogenomic data integration reveals cxcl10 as a potentially downstream causal mediator for il-6 signaling on atherosclerosis. *Circulation*. 2024;149(9):669-683. doi:doi:10.1161/CIRCULATIONAHA.123.064974
240. Storm CS, Kia DA, Almramhi MM, et al. Finding genetically-supported drug targets for parkinson's disease using mendelian randomization of the druggable genome. *Nature Communications*. 2021/12/20 2021;12(1):7342. doi:10.1038/s41467-021-26280-1
241. Bouras E, Karhunen V, Gill D, et al. Circulating inflammatory cytokines and risk of five cancers: A mendelian randomization analysis. *BMC Med*. Jan 11 2022;20(1):3. doi:10.1186/s12916-021-02193-0
242. Burgess S, Thompson SG. Avoiding bias from weak instruments in mendelian randomization studies. *Int J Epidemiol*. Jun 2011;40(3):755-64. doi:10.1093/ije/dyr036
243. Auton A, Brooks LD, Durbin RM, et al. A global reference for human genetic variation. *Nature*. Oct 1 2015;526(7571):68-74. doi:10.1038/nature15393
244. Burgess S, Zuber V, Valdes-Marquez E, Sun BB, Hopewell JC. Mendelian randomization with fine-mapped genetic data: Choosing from large numbers of correlated instrumental variables. *Genet Epidemiol*. Dec 2017;41(8):714-725. doi:10.1002/gepi.22077
245. Burgess S, Small DS, Thompson SG. A review of instrumental variable estimators for mendelian randomization. *Stat Methods Med Res*. Oct 2017;26(5):2333-2355. doi:10.1177/0962280215597579
246. Yavorska OO, Burgess S. Mendelianrandomization: An r package for performing mendelian randomization analyses using summarized data. *International journal of epidemiology*. 2017;46(6):1734-1739. doi:10.1093/ije/dyx034
247. Steegen S, Tuerlinckx F, Gelman A, Vanpaemel W. Increasing transparency through a multiverse analysis. *Perspect Psychol Sci*. Sep 2016;11(5):702-712. doi:10.1177/1745691616658637
248. Kuleshov MV, Jones MR, Rouillard AD, et al. Enrichr: A comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res*. Jul 8 2016;44(W1):W90-7. doi:10.1093/nar/gkw377
249. Koopmans F, van Nierop P, Andres-Alonso M, et al. Syngo: An evidence-based, expert-curated knowledge base for the synapse. *Neuron*. 2019;103(2):217-234.e4. doi:10.1016/j.neuron.2019.05.002
250. Ashburner M, Ball CA, Blake JA, et al. Gene ontology: Tool for the unification of biology. *Nature Genetics*. 2000/05/01 2000;25(1):25-29. doi:10.1038/75556

251. Piñero J, Ramírez-Anguita JM, Saüch-Pitarch J, Ronzano F, Centeno E, Sanz F, Furlong LI. The disgenet knowledge platform for disease genomics: 2019 update. *Nucleic Acids Research*. 2020;48(D1):D845-D855. doi:10.1093/nar/gkz1021
252. Kanehisa M, Furumichi M, Sato Y, Kawashima M, Ishiguro-Watanabe M. Kegg for taxonomy-based analysis of pathways and genomes. *Nucleic Acids Res*. Jan 6 2023;51(D1):D587-d592. doi:10.1093/nar/gkac963
253. Sollis E, Mosaku A, Abid A, et al. The nhgri-ebi gwas catalog: Knowledgebase and deposition resource. *Nucleic Acids Research*. 2023;51(D1):D977-D985. doi:10.1093/nar/gkac1010
254. Ma S, Skarica M, Li Q, et al. Molecular and cellular evolution of the primate dorsolateral prefrontal cortex. *Science*. Sep 30 2022;377(6614):eabo7257. doi:10.1126/science.abo7257
255. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology*. 2018/05/01 2018;36(5):411-420. doi:10.1038/nbt.4096
256. Lamb J, Crawford ED, Peck D, et al. The connectivity map: Using gene-expression signatures to connect small molecules, genes, and disease. *Science*. Sep 29 2006;313(5795):1929-35. doi:10.1126/science.1132939
257. Griffith M, Griffith OL, Coffman AC, et al. Dgidb: Mining the druggable genome. *Nat Methods*. Dec 2013;10(12):1209-10. doi:10.1038/nmeth.2689
258. Sey NYA, Hu B, Mah W, et al. A computational tool (h-magma) for improved prediction of brain-disorder risk genes by incorporating brain chromatin interaction profiles. *Nature Neuroscience*. 2020/04/01 2020;23(4):583-593. doi:10.1038/s41593-020-0603-0
259. Wang D, Liu S, Warrell J, et al. Comprehensive functional genomic resource and integrative model for the human brain. *Science*. 2018;362(6420):eaat8464. doi:10.1126/science.aat8464
260. Sey NYA, Hu B, Iskhakova M, et al. Chromatin architecture in addiction circuitry elucidates biological mechanisms underlying cigarette smoking and alcohol use traits. *bioRxiv*. 2021:2021.03.18.436046. doi:10.1101/2021.03.18.436046
261. Hu B, Won H, Mah W, et al. Neuronal and glial 3d chromatin architecture informs the cellular etiology of brain disorders. *Nature Communications*. 2021/06/25 2021;12(1):3968. doi:10.1038/s41467-021-24243-0
262. Feng H, Mancuso N, Gusev A, Majumdar A, Major M, Pasaniuc B, Kraft P. Leveraging expression from multiple tissues using sparse canonical correlation analysis and aggregate tests improve the power of transcriptome-wide association studies. *bioRxiv*. 2020:2020.07.03.186247. doi:10.1101/2020.07.03.186247
263. Mahi NA, Najafabadi MF, Pilarczyk M, Kouril M, Medvedovic M. Grein: An interactive web platform for re-analyzing geo rna-seq data. *Scientific Reports*. 2019/05/20 2019;9(1):7580. doi:10.1038/s41598-019-43935-8
264. Barbeira AN, Pividori M, Zheng J, Wheeler HE, Nicolae DL, Im HK. Integrating predicted transcriptome from multiple tissues improves association detection. *PLOS Genetics*. 2019;15(1):e1007889. doi:10.1371/journal.pgen.1007889
265. Zuber V, Grinberg NF, Gill D, et al. Combining evidence from mendelian randomization and colocalization: Review and comparison of approaches. *Am J Hum Genet*. May 5 2022;109(5):767-782. doi:10.1016/j.ajhg.2022.04.001

266. Tissink E, Werme J, de Lange SC, et al. The genetic architectures of functional and structural connectivity properties within cerebral resting-state networks. *eneuro*. 2023:ENEURO.0242-22.2023. doi:10.1523/ENEURO.0242-22.2023
267. Foley CN, Staley JR, Breen PG, Sun BB, Kirk PDW, Burgess S, Howson JMM. A fast and efficient colocalization algorithm for identifying shared genetic risk factors across multiple traits. *Nature Communications*. 2021/02/03 2021;12(1):764. doi:10.1038/s41467-020-20885-8
268. Thom CS, Voight BF. Genetic colocalization atlas points to common regulatory sites and genes for hematopoietic traits and hematopoietic contributions to disease phenotypes. *BMC Med Genomics*. 2020/06/29 2020;13(1):89. doi:10.1186/s12920-020-00742-9
269. Kurki MI, Karjalainen J, Palta P, et al. Finngen provides genetic insights from a well-phenotyped isolated population. *Nature*. 2023/01/01 2023;613(7944):508-518. doi:10.1038/s41586-022-05473-8
270. Rosenman R, Tennekoon V, Hill LG. Measuring bias in self-reported data. *Int J Behav Healthc Res*. Oct 2011;2(4):320-332. doi:10.1504/ijbhr.2011.043414
271. Kessler RC, Avenevoli S, Costello EJ, et al. Prevalence, persistence, and sociodemographic correlates of dsm-iv disorders in the national comorbidity survey replication adolescent supplement. *Arch Gen Psychiatry*. Apr 2012;69(4):372-80. doi:10.1001/archgenpsychiatry.2011.160
272. Rosoff DB, Yoo J, Lohoff FW. Smoking is significantly associated with increased risk of covid-19 and other respiratory infections. *Communications Biology*. 2021/10/28 2021;4(1):1230. doi:10.1038/s42003-021-02685-y
273. van der Meer D, Gurholt TP, Sønderby IE, et al. The link between liver fat and cardiometabolic diseases is highlighted by genome-wide association study of mri-derived measures of body composition. *Communications Biology*. 2022/11/19 2022;5(1):1271. doi:10.1038/s42003-022-04237-4
274. Turley P, Walters RK, Maghzian O, et al. Multi-trait analysis of genome-wide association summary statistics using mtag. *Nature Genetics*. 2018/02/01 2018;50(2):229-237. doi:10.1038/s41588-017-0009-4
275. Bulik-Sullivan BK, Loh P-R, Finucane HK, et al. Ld score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics*. 2015/03/01 2015;47(3):291-295. doi:10.1038/ng.3211
276. Suzuki K, Hatzikotoulas K, Southam L, et al. Genetic drivers of heterogeneity in type 2 diabetes pathophysiology. *Nature*. 2024/03/01 2024;627(8003):347-357. doi:10.1038/s41586-024-07019-6
277. Aragam KG, Jiang T, Goel A, et al. Discovery and systematic characterization of risk variants and genes for coronary artery disease in over a million participants. *Nature Genetics*. 2022/12/01 2022;54(12):1803-1815. doi:10.1038/s41588-022-01233-6
278. Savalei V, Bentler PM. A two-stage approach to missing data: Theory and application to auxiliary variables. *Structural Equation Modeling: A Multidisciplinary Journal*. 2009/07/14 2009;16(3):477-497. doi:10.1080/10705510903008238
279. Demange PA, Malanchini M, Mallard TT, et al. Investigating the genetic architecture of noncognitive skills using gwas-by-subtraction. *Nature Genetics*. 2021/01/01 2021;53(1):35-44. doi:10.1038/s41588-020-00754-2
280. Watanabe K, Taskesen E, van Bochoven A, Posthuma D. Functional mapping and annotation of genetic associations with fuma. *Nature Communications*. 2017/11/28 2017;8(1):1826. doi:10.1038/s41467-017-01261-5

281. de Leeuw CA, Mooij JM, Heskes T, Posthuma D. Magma: Generalized gene-set analysis of gwas data. *PLOS Computational Biology*. 2015;11(4):e1004219. doi:10.1371/journal.pcbi.1004219
282. Buniello A, MacArthur JAL, Cerezo M, et al. The nhgri-ebi gwas catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res*. Jan 8 2019;47(D1):D1005-d1012. doi:10.1093/nar/gky1120
283. Metabolic mediators of the effects of body-mass index, overweight, and obesity on coronary heart disease and stroke: A pooled analysis of 97 prospective cohorts with 1·8 million participants. *The Lancet*. 2014;383(9921):970-983. doi:10.1016/S0140-6736(13)61836-X
284. Wang G, Sarkar A, Carbonetto P, Stephens M. A simple new approach to variable selection in regression, with application to genetic fine mapping. https://doi.org/10.1111/rssb.12388. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2020/12/01 2020;82(5):1273-1300. doi:https://doi.org/10.1111/rssb.12388
285. Zou Y, Carbonetto P, Wang G, Stephens M. Fine-mapping from summary data with the "sum of single effects" model. *PLOS Genetics*. 2022;18(7):e1010299. doi:10.1371/journal.pgen.1010299
286. Ward LD, Kellis M. Haploreg v4: Systematic mining of putative causal variants, cell types, regulators and target genes for human complex traits and disease. *Nucleic Acids Res*. Jan 4 2016;44(D1):D877-81. doi:10.1093/nar/gkv1340
287. Aguet F, Anand S, Ardlie KG, et al. The gtex consortium atlas of genetic regulatory effects across human tissues. *Science*. 2020/09/11 2020;369(6509):1318-1330. doi:10.1126/science.aaz1776
288. Laakso M, Kuusisto J, Stančáková A, et al. The metabolic syndrome in men study: A resource for studies of metabolic and cardiovascular diseases. *J Lipid Res*. Mar 2017;58(3):481-493. doi:10.1194/jlr.O072629
289. Knox C, Wilson M, Klinger CM, et al. Drugbank 6.0: The drugbank knowledgebase for 2024. *Nucleic Acids Res*. Jan 5 2024;52(D1):D1265-d1275. doi:10.1093/nar/gkad976
290. Sakaue S, Okada Y. Grep: Genome for repositioning drugs. *Bioinformatics*. 2019;35(19):3821-3823. doi:10.1093/bioinformatics/btz166
291. Iorio F, Rittman T, Ge H, Menden M, Saez-Rodriguez J. Transcriptional data: A new gateway to drug repositioning? *Drug Discov Today*. Apr 2013;18(7-8):350-7. doi:10.1016/j.drudis.2012.07.014
292. Finucane HK, Bulik-Sullivan B, Gusev A, et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat Genet*. Nov 2015;47(11):1228-35. doi:10.1038/ng.3404
293. Dunham I, Kundaje A, Aldred SF, et al. An integrated encyclopedia of DNA elements in the human genome. *Nature*. 2012/09/01 2012;489(7414):57-74. doi:10.1038/nature11247
294. Kundaje A, Meuleman W, Ernst J, et al. Integrative analysis of 111 reference human epigenomes. *Nature*. 2015/02/01 2015;518(7539):317-330. doi:10.1038/nature14248
295. Pers TH, Karjalainen JM, Chan Y, et al. Biological interpretation of genome-wide association studies using predicted gene functions. *Nature Communications*. 2015/01/19 2015;6(1):5890. doi:10.1038/ncomms6890
296. Fehrmann RSN, Karjalainen JM, Krajewska M, et al. Gene expression analysis identifies global gene dosage sensitivity in cancer. *Nature Genetics*. 2015/02/01 2015;47(2):115-125. doi:10.1038/ng.3173

297. Schaum N, Karkanias J, Neff NF, et al. Single-cell transcriptomics of 20 mouse organs creates a tabula muris. *Nature*. 2018/10/01 2018;562(7727):367-372. doi:10.1038/s41586-018-0590-4
298. Richardson TG, Sanderson E, Palmer TM, Ala-Korpela M, Ference BA, Davey Smith G, Holmes MV. Evaluating the relationship between circulating lipoprotein lipids and apolipoproteins with risk of coronary heart disease: A multivariable mendelian randomisation analysis. *PLOS Medicine*. 2020;17(3):e1003062. doi:10.1371/journal.pmed.1003062
299. Keaton JM, Kamali Z, Xie T, et al. Genome-wide analysis in over 1 million individuals of european ancestry yields improved polygenic risk scores for blood pressure traits. *Nature Genetics*. 2024/05/01 2024;56(5):778-791. doi:10.1038/s41588-024-01714-w
300. Chaudhury A, Duvoor C, Reddy Dendi VS, et al. Clinical review of antidiabetic drugs: Implications for type 2 diabetes mellitus management. *Front Endocrinol (Lausanne)*. 2017;8:6. doi:10.3389/fendo.2017.00006
301. Mendez D, Gaulton A, Bento AP, et al. Chembl: Towards direct deposition of bioassay data. *Nucleic Acids Research*. 2019;47(D1):D930-D940. doi:10.1093/nar/gky1075
302. Wishart DS, Knox C, Guo AC, et al. Drugbank: A comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res*. Jan 1 2006;34(Database issue):D668-72. doi:10.1093/nar/gkj067
303. Tang B, Wang Y, Jiang X, Thambisetty M, Ferrucci L, Johnell K, Hägg S. Genetic variation in targets of antidiabetic drugs and alzheimer disease risk: A mendelian randomization study. *Neurology*. Aug 16 2022;99(7):e650-e659. doi:10.1212/wnl.0000000000200771
304. Ussher JR, Drucker DJ. Glucagon-like peptide 1 receptor agonists: Cardiovascular benefits and mechanisms of action. *Nature Reviews Cardiology*. 2023/07/01 2023;20(7):463-474. doi:10.1038/s41569-023-00849-3
305. Ference BA, Kastelein JJP, Ray KK, et al. Association of triglyceride-lowering lpl variants and ldl-c–lowering ldlr variants with risk of coronary heart disease. *JAMA*. 2019;321(4):364-373. doi:10.1001/jama.2018.20045
306. Bhatt-Wessel B, Jordan TW, Miller JH, Peng L. Role of dgat enzymes in triacylglycerol metabolism. *Arch Biochem Biophys*. Oct 1 2018;655:1-11. doi:10.1016/j.abb.2018.08.001
307. Harrison SA, Taub R, Neff GW, et al. Resmetirom for nonalcoholic fatty liver disease: A randomized, double-blind, placebo-controlled phase 3 trial. *Nat Med*. Nov 2023;29(11):2919-2928. doi:10.1038/s41591-023-02603-1
308. Mak LY, Gane E, Schwabe C, et al. A phase i/ii study of aro-hsd, an rna interference therapeutic, for the treatment of non-alcoholic steatohepatitis. *J Hepatol*. Apr 2023;78(4):684-692. doi:10.1016/j.jhep.2022.11.025
309. Pharmaceuticals A. A study of aln-hsd in healthy adult subjects and adult patients with nonalcoholic steatohepatitis (nash). Accessed 10/15, 2022. https://clinicaltrials.gov/ct2/show/NCT04565717
310. Pharmaceuticals R. A study to evaluate the efficacy and safety of aln-hsd in adult participants with non-alcoholic steatohepatitis (nash) with fibrosis with genetic risk factors (nashgen-2). Accessed 10/20, 2022. https://clinicaltrials.gov/ct2/show/NCT05519475
311. AstraZeneca. A study to assess safety, tolerability, pk and pd of azd2693 in non-alcoholic steatohepatitis patients
. Accessed September 20, 2022. https://www.clinicaltrials.gov/ct2/show/NCT04483947

312. Romeo S, Sanyal A, Valenti L. Leveraging human genetics to identify potential new treatments for fatty liver disease. *Cell Metabolism*. 2020/01/07/ 2020;31(1):35-45. doi:https://doi.org/10.1016/j.cmet.2019.12.002
313. Liu Y, Basty N, Whitcher B, et al. Genetic architecture of 11 organ traits derived from abdominal mri using deep learning. *Elife*. Jun 15 2021;10doi:10.7554/eLife.65554
314. Romeo S, Sanyal A, Valenti L. Leveraging human genetics to identify potential new treatments for fatty liver disease. *Cell Metab*. Jan 7 2020;31(1):35-45. doi:10.1016/j.cmet.2019.12.002
315. Suchard MA, Schuemie MJ, Krumholz HM, et al. Comprehensive comparative effectiveness and safety of first-line antihypertensive drug classes: A systematic, multinational, large-scale analysis. *The Lancet*. 2019;394(10211):1816-1826. doi:10.1016/S0140-6736(19)32317-7
316. Gill D, Georgakis MK, Koskeridis F, et al. Use of genetic variants related to antihypertensive drugs to inform on efficacy and side effects. *Circulation*. 2019;140(4):270-279. doi:doi:10.1161/CIRCULATIONAHA.118.038814
317. Bowden J, Del Greco M F, Minelli C, et al. Improving the accuracy of two-sample summary-data mendelian randomization: Moving beyond the nome assumption. *International journal of epidemiology*. 2019;48(3):728-742.
318. Rees JMB, Wood AM, Burgess S. Extending the mr-egger method for multivariable mendelian randomization to correct for both measured and unmeasured pleiotropy. *Stat Med*. Dec 20 2017;36(29):4705-4718. doi:10.1002/sim.7492
319. Chen BY, Bone WP, Lorenz K, Levin M, Ritchie MD, Voight BF. Colocquial: A qtl-gwas colocalization pipeline. *Bioinformatics*. Sep 15 2022;38(18):4409-4411. doi:10.1093/bioinformatics/btac512
320. Maria KS, Tom RG. Evaluating the life-extending potential and safety profile of rapamycin: A mendelian randomization study of the mtor pathway. *medRxiv*. 2023:2023.10.02.23296427. doi:10.1101/2023.10.02.23296427
321. Myers TA, Chanock SJ, Machiela MJ. Ldlinkr: An r package for rapidly calculating linkage disequilibrium statistics in diverse populations. *Front Genet*. 2020;11:157. doi:10.3389/fgene.2020.00157
322. Karpe F, Vasan SK, Humphreys SM, Miller J, Cheeseman J, Dennis AL, Neville MJ. Cohort profile: The oxford biobank. *International Journal of Epidemiology*. 2017;47(1):21-21g. doi:10.1093/ije/dyx132
323. Bowden J, Davey Smith G, Burgess S. Mendelian randomization with invalid instruments: Effect estimation and bias detection through egger regression. *Int J Epidemiol*. Apr 2015;44(2):512-25. doi:10.1093/ije/dyv080
324. Bowden J, Davey Smith G, Haycock PC, Burgess S. Consistent estimation in mendelian randomization with some invalid instruments using a weighted median estimator. *Genet Epidemiol*. May 2016;40(4):304-14. doi:10.1002/gepi.21965
325. Hartwig FP, Davey Smith G, Bowden J. Robust inference in summary data mendelian randomization via the zero modal pleiotropy assumption. *International Journal of Epidemiology*. 2017;46(6):1985-1998. doi:10.1093/ije/dyx102
326. Yoshiji S, Butler-Laporte G, Lu T, et al. Proteome-wide mendelian randomization implicates nephronectin as an actionable mediator of the effect of obesity on covid-19 severity. *Nature Metabolism*. 2023/02/01 2023;5(2):248-264. doi:10.1038/s42255-023-00742-w

327. Burgess S, Daniel RM, Butterworth AS, Thompson SG. Network mendelian randomization: Using genetic variants as instrumental variables to investigate mediation in causal pathways. *Int J Epidemiol*. Apr 2015;44(2):484-95. doi:10.1093/ije/dyu176
328. Kosti I, Jain N, Aran D, Butte AJ, Sirota M. Cross-tissue analysis of gene and protein expression in normal and cancer tissues. *Scientific Reports*. 2016/05/04 2016;6(1):24799. doi:10.1038/srep24799
329. Feng H, Mancuso N, Gusev A, Majumdar A, Major M, Pasaniuc B, Kraft P. Leveraging expression from multiple tissues using sparse canonical correlation analysis and aggregate tests improves the power of transcriptome-wide association studies. *PLOS Genetics*. 2021;17(4):e1008973. doi:10.1371/journal.pgen.1008973
330. Loewa A, Feng JJ, Hedtrich S. Human disease models in drug development. *Nature Reviews Bioengineering*. 2023/08/01 2023;1(8):545-559. doi:10.1038/s44222-023-00063-3
331. Duffy Á, Verbanck M, Dobbyn A, et al. Tissue-specific genetic features inform prediction of drug side effects in clinical trials. *Science Advances*. 2020;6(37):eabb6242. doi:doi:10.1126/sciadv.abb6242
332. Sun D, Gao W, Hu H, Zhou S. Why 90% of clinical drug development fails and how to improve it? *Acta Pharm Sin B*. Jul 2022;12(7):3049-3062. doi:10.1016/j.apsb.2022.02.002
333. Phelps NH, Singleton RK, Zhou B, et al. Worldwide trends in underweight and obesity from 1990 to 2022: A pooled analysis of 3663 population-representative studies with 222 million children, adolescents, and adults. *The Lancet*. 2024;403(10431):1027-1050. doi:10.1016/S0140-6736(23)02750-2
334. Yengo L, Sidorenko J, Kemper KE, et al. Meta-analysis of genome-wide association studies for height and body mass index in ~700000 individuals of european ancestry. *Hum Mol Genet*. Oct 15 2018;27(20):3641-3649. doi:10.1093/hmg/ddy271
335. Di Angelantonio E, Thompson SG, Kaptoge S, et al. Efficiency and safety of varying the frequency of whole blood donation (interval): A randomised trial of 45 000 donors. *The Lancet*. 2017;390(10110):2360-2371. doi:10.1016/S0140-6736(17)31928-1
336. Rosoff DB, Bell AS, Wagner J, et al. Assessing the impact of pcsk9 and hmgcr inhibition on liver function: Drug-target mendelian randomization analyses in four ancestries. *Cellular and Molecular Gastroenterology and Hepatology*. 2023;doi:10.1016/j.jcmgh.2023.09.001
337. Korunes KL, Goldberg A. Human genetic admixture. *PLOS Genetics*. 2021;17(3):e1009374. doi:10.1371/journal.pgen.1009374
338. Zakharia F, Basu A, Absher D, et al. Characterizing the admixed african ancestry of african americans. *Genome Biol*. 2009;10(12):R141. doi:10.1186/gb-2009-10-12-r141
339. Kachuri L, Mak ACY, Hu D, et al. Gene expression in african americans, puerto ricans and mexican americans reveals ancestry-specific patterns of genetic architecture. *Nature Genetics*. 2023/06/01 2023;55(6):952-963. doi:10.1038/s41588-023-01377-z
340. Redberg RF. Statins and weight gain. *JAMA Internal Medicine*. 2014;174(7):1046-1046. doi:10.1001/jamainternmed.2014.1994
341. Yu Q, Chen Y, Xu CB. Statins and new-onset diabetes mellitus: Ldl receptor may provide a key link. *Front Pharmacol*. 2017;8:372. doi:10.3389/fphar.2017.00372
342. Klimentidis YC, Arora A, Newell M, Zhou J, Ordovas JM, Renquist BJ, Wood AC. Phenotypic and genetic characterization of lower ldl cholesterol and increased type 2 diabetes risk in the uk biobank. *Diabetes*. Oct 2020;69(10):2194-2205. doi:10.2337/db19-1134
343. Sacks FM, Tonkin AM, Craven T, et al. Coronary heart disease in patients with low ldl-cholesterol: Benefit of pravastatin in diabetics and enhanced role for hdl-cholesterol and

triglycerides as risk factors. *Circulation*. Mar 26 2002;105(12):1424-8. doi:10.1161/01.cir.0000012918.84068.43
344. Besseling J, Kastelein JJ, Defesche JC, Hutten BA, Hovingh GK. Association between familial hypercholesterolemia and prevalence of type 2 diabetes mellitus. *Jama*. Mar 10 2015;313(10):1029-36. doi:10.1001/jama.2015.1206
345. Wolska A, Remaley AT. Measuring ldl-cholesterol: What is the best way to do it? *Curr Opin Cardiol*. Jul 2020;35(4):405-411. doi:10.1097/hco.0000000000000740
346. Bonilha I, Hajduch E, Luchiari B, Nadruz W, Le Goff W, Sposito AC. The reciprocal relationship between ldl metabolism and type 2 diabetes mellitus. *Metabolites*. Nov 28 2021;11(12)doi:10.3390/metabo11120807
347. Morze J, Wittenbecher C, Schwingshackl L, Danielewicz A, Rynkiewicz A, Hu FB, Guasch-Ferré M. Metabolomics and type 2 diabetes risk: An updated systematic review and meta-analysis of prospective cohort studies. *Diabetes Care*. Apr 1 2022;45(4):1013-1024. doi:10.2337/dc21-1705
348. Djekic D, Nicoll R, Novo M, Henein M. Metabolomics in atherosclerosis. *IJC Metabolic & Endocrine*. 2015/09/01/ 2015;8:26-30. doi:https://doi.org/10.1016/j.ijcme.2014.11.004
349. Turner BE, Steinberg JR, Weeks BT, Rodriguez F, Cullen MR. Race/ethnicity reporting and representation in us clinical trials: A cohort study. *The Lancet Regional Health – Americas*. 2022;11doi:10.1016/j.lana.2022.100252
350. Patin E, Lopez M, Grollemund R, et al. Dispersals and genetic adaptation of bantu-speaking populations in africa and north america. *Science*. 2017/05/05 2017;356(6337):543-546. doi:10.1126/science.aal1988
351. Lippi G, Mattiuzzi C, Cervellin G. Statins popularity: A global picture. *Br J Clin Pharmacol*. Jul 2019;85(7):1614-1615. doi:10.1111/bcp.13944
352. Lin S-y, Baumann K, Zhou C, Zhou W, Cuellar AE, Xue H. Trends in use and expenditures for brand-name statins after introduction of generic statins in the us, 2002-2018. *JAMA Network Open*. 2021;4(11):e2135371-e2135371. doi:10.1001/jamanetworkopen.2021.35371
353. Pleis JR, Lethbridge-Cejku M. Summary health statistics for u.S. Adults: National health interview survey, 2005. *Vital Health Stat 10*. Dec 2006;(232):1-153.
354. Hammerton G, Munafò MR. Causal inference with observational data: The need for triangulation of evidence. *Psychol Med*. Mar 2021;51(4):563-578. doi:10.1017/s0033291720005127
355. Fortmann AL, Gallo LC, Philis-Tsimikas A. Glycemic control among latinos with type 2 diabetes: The role of social-environmental support resources. *Health Psychol*. May 2011;30(3):251-8. doi:10.1037/a0022850
356. Aceves B, Ezekiel-Herrera D, Marino M, Datta R, Lucas J, Giebultowicz S, Heintzman J. Disparities in hba1c testing between aging us latino and non-latino white primary care patients. *Preventive Medicine Reports*. 2022/04/01/ 2022;26:101739. doi:https://doi.org/10.1016/j.pmedr.2022.101739
357. Williams DM, Finan C, Schmidt AF, Burgess S, Hingorani AD. Lipid lowering and alzheimer disease risk: A mendelian randomization study. *Annals of Neurology*. 2020;87(1):30-39. doi:https://doi.org/10.1002/ana.25642
358. Harding JL, Pavkov ME, Magliano DJ, Shaw JE, Gregg EW. Global trends in diabetes complications: A review of current evidence. *Diabetologia*. Jan 2019;62(1):3-16. doi:10.1007/s00125-018-4711-2

359. Zheng Y, Ley SH, Hu FB. Global aetiology and epidemiology of type 2 diabetes mellitus and its complications. *Nature Reviews Endocrinology*. 2018/02/01 2018;14(2):88-98. doi:10.1038/nrendo.2017.151
360. Michos ED, Reddy TK, Gulati M, et al. Improving the enrollment of women and racially/ethnically diverse populations in cardiovascular clinical trials: An aspc practice statement. *Am J Prev Cardiol*. Dec 2021;8:100250. doi:10.1016/j.ajpc.2021.100250
361. Loree JM, Anand S, Dasari A, et al. Disparity of race reporting and representation in clinical trials leading to cancer drug approvals from 2008 to 2018. *JAMA Oncology*. 2019;5(10):e191870-e191870. doi:10.1001/jamaoncol.2019.1870
362. Zheng J, Xu M, Walker V, et al. Evaluating the efficacy and mechanism of metformin targets on reducing alzheimer's disease risk in the general population: A mendelian randomization study. *medRxiv*. 2022:2022.04.09.22273625. doi:10.1101/2022.04.09.22273625
363. Bentley AR, Callier S, Rotimi CN. Diversity and inclusion in genomic research: Why the uneven progress? *J Community Genet*. Oct 2017;8(4):255-266. doi:10.1007/s12687-017-0316-6
364. Borrell LN, Elhawary JR, Fuentes-Afflick E, et al. Race and genetic ancestry in medicine — a time for reckoning with racism. *New England Journal of Medicine*. 2021;384(5):474-480. doi:10.1056/NEJMms2029562
365. Ference BA, Ginsberg HN, Graham I, et al. Low-density lipoproteins cause atherosclerotic cardiovascular disease. 1. Evidence from genetic, epidemiologic, and clinical studies. A consensus statement from the european atherosclerosis society consensus panel. *Eur Heart J*. Aug 21 2017;38(32):2459-2472. doi:10.1093/eurheartj/ehx144
366. Landmesser U, Chapman MJ, Farnier M, et al. European society of cardiology/european atherosclerosis society task force consensus statement on proprotein convertase subtilisin/kexin type 9 inhibitors: Practical guidance for use in patients at very high cardiovascular risk. *Eur Heart J*. Aug 1 2017;38(29):2245-2255. doi:10.1093/eurheartj/ehw480
367. Ehinger Y, Zhang Z, Phamluong K, Soneja D, Shokat KM, Ron D. Brain-specific inhibition of mtorc1 eliminates side effects resulting from mtorc1 blockade in the periphery and reduces alcohol intake in mice. *Nature Communications*. 2021/07/27 2021;12(1):4407. doi:10.1038/s41467-021-24567-x
368. Yang C, Farias FHG, Ibanez L, et al. Genomic atlas of the proteome from brain, csf and plasma prioritizes proteins implicated in neurological disorders. *Nat Neurosci*. Sep 2021;24(9):1302-1312. doi:10.1038/s41593-021-00886-6
369. Hatoum AS, Colbert SMC, Johnson EC, et al. Multivariate genome-wide association meta-analysis of over 1 million subjects identifies loci underlying multiple substance use disorders. *Nature Mental Health*. 2023/03/01 2023;1(3):210-223. doi:10.1038/s44220-023-00034-y
370. Noble EP. Alcoholism and the dopaminergic system: A review. *Addict Biol*. 1996;1(4):333-48. doi:10.1080/1355621961000124956
371. Kishi T, Sevy S, Chekuri R, Correll CU. Antipsychotics for primary alcohol dependence: A systematic review and meta-analysis of placebo-controlled trials. *J Clin Psychiatry*. Jul 2013;74(7):e642-54. doi:10.4088/JCP.12r08178
372. Spencer BH, Ami SI, Qingyue Y, Chelsie EB-B, Rohan HCP. Genome- and transcriptome-wide splicing associations with problematic alcohol use and alcohol use disorder. *bioRxiv*. 2021:2021.03.31.437932. doi:10.1101/2021.03.31.437932

373. Marees AT, Gamazon ER, Gerring Z, et al. Post-gwas analysis of six substance use traits improves the identification and functional interpretation of genetic risk loci. *Drug Alcohol Depend*. Jan 1 2020;206:107703. doi:10.1016/j.drugalcdep.2019.107703
374. Desikan RS, Schork AJ, Wang Y, et al. Genetic overlap between alzheimer's disease and parkinson's disease at the mapt locus. *Mol Psychiatry*. Dec 2015;20(12):1588-95. doi:10.1038/mp.2015.6
375. McColl ER, Piquette-Miller M. Slc neurotransmitter transporters as therapeutic targets for alcohol use disorder: A narrative review. *Alcohol Clin Exp Res*. Oct 2020;44(10):1965-1976. doi:10.1111/acer.14445
376. Li J, Amoh BK, McCormick E, et al. Integration of transcriptome-wide association study with neuronal dysfunction assays provides functional genomics evidence for parkinson's disease genes. *Human Molecular Genetics*. 2023;32(4):685-695. doi:10.1093/hmg/ddac230
377. Stelzer G, Rosen N, Plaschkes I, et al. The genecards suite: From gene data mining to disease genome sequence analyses. *Curr Protoc Bioinformatics*. Jun 20 2016;54:1.30.1-1.30.33. doi:10.1002/cpbi.5
378. Chatzinakos C, Georgiadis F, Daskalakis NP. Gwas meets transcriptomics: From genetic letters to transcriptomic words of neuropsychiatric risk. *Neuropsychopharmacology*. 2021/01/01 2021;46(1):255-256. doi:10.1038/s41386-020-00835-0
379. Wingo TS, Liu Y, Gerasimov ES, et al. Brain proteome-wide association study implicates novel proteins in depression pathogenesis. *Nat Neurosci*. Jun 2021;24(6):810-817. doi:10.1038/s41593-021-00832-6
380. Hall LS, Medway CW, Pain O, et al. A transcriptome-wide association study implicates specific pre- and post-synaptic abnormalities in schizophrenia. *Human Molecular Genetics*. 2020;29(1):159-167. doi:10.1093/hmg/ddz253
381. Lichou F, Trynka G. Functional studies of gwas variants are gaining momentum. *Nature Communications*. 2020/12/08 2020;11(1):6283. doi:10.1038/s41467-020-20188-y
382. Charlet K, Heinz A. Harm reduction—a systematic review on effects of alcohol reduction on physical and mental symptoms. https://doi.org/10.1111/adb.12414. *Addiction Biology*. 2017/09/01 2017;22(5):1119-1159. doi:https://doi.org/10.1111/adb.12414
383. Carnicella S, Ahmadiantehrani S, He DY, Nielsen CK, Bartlett SE, Janak PH, Ron D. Cabergoline decreases alcohol drinking and seeking behaviors via glial cell line-derived neurotrophic factor. *Biol Psychiatry*. Jul 15 2009;66(2):146-53. doi:10.1016/j.biopsych.2008.12.022
384. Spoelder M, Baars AM, Rotte MD, Vanderschuren LJ, Lesscher HM. Dopamine receptor agonists modulate voluntary alcohol intake independently of individual levels of alcohol intake in rats. *Psychopharmacology (Berl)*. Jul 2016;233(14):2715-25. doi:10.1007/s00213-016-4330-x
385. Su WJ, Peng W, Gong H, et al. Antidiabetic drug glyburide modulates depressive-like behavior comorbid with insulin resistance. *J Neuroinflammation*. Oct 30 2017;14(1):210. doi:10.1186/s12974-017-0985-4
386. Sonsalla MM, Babygirija R, Johnson M, et al. Acarbose ameliorates western diet-induced metabolic and cognitive impairments in the 3xtg mouse model of alzheimer's disease. *Alzheimer's & Dementia*. 2023;19(S13):e078561. doi:https://doi.org/10.1002/alz.078561
387. Rajkumar M, Kannan S, Thangaraj R. Voglibose attenuates cognitive impairment, aβ aggregation, oxidative stress, and neuroinflammation in streptozotocin-induced alzheimer's disease rat model. *Inflammopharmacology*. Oct 2023;31(5):2751-2771. doi:10.1007/s10787-023-01313-x

388. Jerlhag E. Alcohol-mediated behaviours and the gut-brain axis; with focus on glucagon-like peptide-1. *Brain Research*. 2020/01/15/ 2020;1727:146562. doi:https://doi.org/10.1016/j.brainres.2019.146562
389. Paunovska K, Loughrey D, Dahlman JE. Drug delivery systems for rna therapeutics. *Nature Reviews Genetics*. 2022/05/01 2022;23(5):265-280. doi:10.1038/s41576-021-00439-4
390. Milivojevic V, Angarita GA, Hermes G, Sinha R, Fox HC. Effects of prazosin on provoked alcohol craving and autonomic and neuroendocrine response to stress in alcohol use disorder. *Alcohol Clin Exp Res*. Jul 2020;44(7):1488-1496. doi:10.1111/acer.14378
391. Rogawski MA, Wenk GL. The neuropharmacological basis for the use of memantine in the treatment of alzheimer's disease. *CNS Drug Rev*. Fall 2003;9(3):275-308. doi:10.1111/j.1527-3458.2003.tb00254.x
392. Montemitro C, Angebrandt A, Wang T-Y, Pettorruso M, Abulseoud OA. Mechanistic insights into the efficacy of memantine in treating certain drug addictions. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*. 2021/12/20/ 2021;111:110409. doi:https://doi.org/10.1016/j.pnpbp.2021.110409
393. Lawlor DA, Tilling K, Davey Smith G. Triangulation in aetiological epidemiology. *International Journal of Epidemiology*. 2016;45(6):1866-1886. doi:10.1093/ije/dyw314
394. Gissen P. 1265vps33b, vipas39, and the arthrogryposis, renal dysfunction, and cholestasis syndrome. *Epstein's Inborn Errors of Development: The Molecular Basis of Clinical Disorders of Morphogenesis*. Oxford University Press; 2016. Accessed 1/21/2024. https://doi.org/10.1093/med/9780199934522.003.0193
395. Worzfeld T, Schwaninger M. Apicobasal polarity of brain endothelial cells. *J Cereb Blood Flow Metab*. Feb 2016;36(2):340-62. doi:10.1177/0271678x15608644
396. Morris G, Fernandes BS, Puri BK, Walker AJ, Carvalho AF, Berk M. Leaky brain in neurological and psychiatric disorders: Drivers and consequences. *Australian & New Zealand Journal of Psychiatry*. 2018/10/01 2018;52(10):924-948. doi:10.1177/0004867418796955
397. Mushtaq Z, Aavula K, Lasser DA, Kieweg ID, Lion LM, Kins S, Pielage J. Madm/nrbp1 mediates synaptic maintenance and neurodegeneration-induced presynaptic homeostatic potentiation. *Cell Reports*. 2022/11/29/ 2022;41(9):111710. doi:https://doi.org/10.1016/j.celrep.2022.111710
398. Nalberczak-Skóra M, Beroun A, Skonieczna E, et al. Impaired synaptic transmission in dorsal dentate gyrus increases impulsive alcohol seeking. *Neuropsychopharmacology*. 2023/02/01 2023;48(3):436-447. doi:10.1038/s41386-022-01464-5
399. Heymann D, Stern Y, Cosentino S, Tatarina-Nulman O, Dorrejo JN, Gu Y. The association between alcohol use and the progression of alzheimer's disease. *Curr Alzheimer Res*. 2016;13(12):1356-1362. doi:10.2174/1567205013666160603005035
400. Belgareh N, Rabut G, Baï SW, et al. An evolutionarily conserved npc subcomplex, which redistributes in part to kinetochores in mammalian cells. *J Cell Biol*. Sep 17 2001;154(6):1147-60. doi:10.1083/jcb.200101081
401. Coyne AN, Rothstein JD. Nuclear pore complexes - a doorway to neural injury in neurodegeneration. *Nat Rev Neurol*. Jun 2022;18(6):348-362. doi:10.1038/s41582-022-00653-6
402. Cullen KM, Halliday GM. Neurofibrillary tangles in chronic alcoholics. *Neuropathol Appl Neurobiol*. Aug 1995;21(4):312-8. doi:10.1111/j.1365-2990.1995.tb01065.x
403. Johnson EC, Kapoor M, Hatoum AS, et al. Investigation of convergent and divergent genetic influences underlying schizophrenia and alcohol use disorder. *Psychol Med*. Mar 2023;53(4):1196-1204. doi:10.1017/s003329172100266x

404. Castillo-Carniglia A, Keyes KM, Hasin DS, Cerdá M. Psychiatric comorbidities in alcohol use disorder. *Lancet Psychiatry*. Dec 2019;6(12):1068-1080. doi:10.1016/s2215-0366(19)30222-6
405. Drake RE, Wallach MA. Substance abuse among the chronic mentally ill. *Hosp Community Psychiatry*. Oct 1989;40(10):1041-6. doi:10.1176/ps.40.10.1041
406. Khurana S, Oberdoerffer P. Replication stress: A lifetime of epigenetic change. *Genes*. 2015;6(3):858-877. doi:10.3390/genes6030858
407. López-Otín C, Blasco MA, Partridge L, Serrano M, Kroemer G. The hallmarks of aging. *Cell*. Jun 6 2013;153(6):1194-217. doi:10.1016/j.cell.2013.05.039
408. Jung J, McCartney DL, Wagner J, et al. Additive effects of stress and alcohol exposure on accelerated epigenetic aging in alcohol use disorder. *Biological Psychiatry*. 2022/07/16/ 2022;doi:https://doi.org/10.1016/j.biopsych.2022.06.036
409. Jung J, McCartney DL, Wagner J, et al. Alcohol use disorder is associated with DNA methylation-based shortening of telomere length and regulated by tespa1: Implications for aging. *Molecular Psychiatry*. 2022/06/15 2022;doi:10.1038/s41380-022-01624-5
410. Luo A, Jung J, Longley M, et al. Epigenetic aging is accelerated in alcohol use disorder and regulated by genetic variation in apol2. *Neuropsychopharmacology*. Jan 2020;45(2):327-336. doi:10.1038/s41386-019-0500-y
411. Mavromatis LA, Rosoff DB, Bell AS, Jung J, Wagner J, Lohoff FW. Multi-omic underpinnings of epigenetic aging and human longevity. *Nature Communications*. 2023/04/19 2023;14(1):2236. doi:10.1038/s41467-023-37729-w
412. Harrison RK. Phase ii and phase iii failures: 2013–2015. *Nature Reviews Drug Discovery*. 2016/12/01 2016;15(12):817-818. doi:10.1038/nrd.2016.184
413. Chong M, Sjaarda J, Pigeyre M, et al. Novel drug targets for ischemic stroke identified through mendelian randomization analysis of the blood proteome. *Circulation*. Sep 9 2019;140(10):819-830. doi:10.1161/circulationaha.119.040180
414. Dudbridge F. Power and predictive accuracy of polygenic risk scores. *PLOS Genetics*. 2013;9(3):e1003348. doi:10.1371/journal.pgen.1003348
415. Burgess S, Davey Smith G, Davies NM, et al. Guidelines for performing mendelian randomization investigations: Update for summer 2023. *Wellcome Open Res*. 2019;4:186. doi:10.12688/wellcomeopenres.15555.3
416. Lähnemann D, Köster J, Szczurek E, et al. Eleven grand challenges in single-cell data science. *Genome Biology*. 2020/02/07 2020;21(1):31. doi:10.1186/s13059-020-1926-6
417. Hicks SC, Townes FW, Teng M, Irizarry RA. Missing data and technical variability in single-cell rna-sequencing experiments. *Biostatistics*. Oct 1 2018;19(4):562-578. doi:10.1093/biostatistics/kxx053
418. Williams CG, Lee HJ, Asatsuma T, Vento-Tormo R, Haque A. An introduction to spatial transcriptomics for biomedical research. *Genome Medicine*. 2022/06/27 2022;14(1):68. doi:10.1186/s13073-022-01075-1
419. Tanay A, Regev A. Scaling single-cell genomics from phenomenology to mechanism. *Nature*. 2017/01/01 2017;541(7637):331-338. doi:10.1038/nature21350
420. Egervari G, Siciliano CA, Whiteley EL, Ron D. Alcohol and the brain: From genes to circuits. *Trends in Neurosciences*. 2021/12/01/ 2021;44(12):1004-1015. doi:https://doi.org/10.1016/j.tins.2021.09.006
421. Wallace C. A more accurate method for colocalisation analysis allowing for multiple causal variants. *PLOS Genetics*. 2021;17(9):e1009440. doi:10.1371/journal.pgen.1009440

422. Britton A, Ben-Shlomo Y, Benzeval M, Kuh D, Bell S. Life course trajectories of alcohol consumption in the united kingdom using longitudinal data from nine cohort studies. *BMC Medicine*. 2015/03/06 2015;13(1):47. doi:10.1186/s12916-015-0273-z
423. Fry A, Littlejohns TJ, Sudlow C, et al. Comparison of sociodemographic and health-related characteristics of uk biobank participants with those of the general population. *American Journal of Epidemiology*. Nov 1 2017;186(9):1026-1034. doi:10.1093/aje/kwx246
424. Sudhinaraset M, Wigglesworth C, Takeuchi DT. Social and cultural contexts of alcohol use: Influences in a social-ecological framework. *Alcohol Res*. 2016;38(1):35-45.
425. Ehinger Y, Zhang Z, Phamluong K, Soneja D, Shokat KM, Ron D. Brain-specific inhibition of mtorc1 eliminates side effects resulting from mtorc1 blockade in the periphery and reduces alcohol intake in mice. *Nat Commun*. Jul 27 2021;12(1):4407. doi:10.1038/s41467-021-24567-x
426. Zhang H. Lysosomal acid lipase and lipid metabolism: New mechanisms, new questions, and new therapies. *Curr Opin Lipidol*. Jun 2018;29(3):218-223. doi:10.1097/mol.0000000000000507
427. Brouwers MCGJ, Simons N, Stehouwer CDA, Isaacs A. Non-alcoholic fatty liver disease and cardiovascular disease: Assessing the evidence for causality. *Diabetologia*. 2020/02/01 2020;63(2):253-260. doi:10.1007/s00125-019-05024-3
428. Copps KD, White MF. Regulation of insulin sensitivity by serine/threonine phosphorylation of insulin receptor substrate proteins irs1 and irs2. *Diabetologia*. Oct 2012;55(10):2565-2582. doi:10.1007/s00125-012-2644-8
429. Scott LJ, Erdos MR, Huyghe JR, et al. The genetic regulatory signature of type 2 diabetes in human skeletal muscle. *Nature Communications*. 2016/06/29 2016;7(1):11764. doi:10.1038/ncomms11764
430. Gao C, Wang Y. Mrna metabolism in cardiac development and disease: Life after transcription. *Physiol Rev*. Apr 1 2020;100(2):673-694. doi:10.1152/physrev.00007.2019
431. Rusanescu G, Weissleder R, Aikawa E. Notch signaling in cardiovascular disease and calcification. *Curr Cardiol Rev*. Aug 2008;4(3):148-56. doi:10.2174/157340308785160552
432. Davies NM, Holmes MV, Davey Smith G. Reading mendelian randomisation studies: A guide, glossary, and checklist for clinicians. *BMJ*. 2018;362:k601. doi:10.1136/bmj.k601
433. Smith K, Deutsch AJ, McGrail C, et al. Multi-ancestry polygenic mechanisms of type 2 diabetes. *Nat Med*. Apr 2024;30(4):1065-1074. doi:10.1038/s41591-024-02865-3
434. Sabeva NS, Liu J, Graf GA. The abcg5 abcg8 sterol transporter and phytosterols: Implications for cardiometabolic disease. *Curr Opin Endocrinol Diabetes Obes*. Apr 2009;16(2):172-7. doi:10.1097/med.0b013e3283292312
435. Kersten S. Role and mechanism of the action of angiopoietin-like protein angptl4 in plasma lipid metabolism. *Journal of Lipid Research*. 2021/01/01/ 2021;62:100150. doi:https://doi.org/10.1016/j.jlr.2021.100150
436. Deng M, Kutrolli E, Sadewasser A, et al. Angptl4 silencing via antisense oligonucleotides reduces plasma triglycerides and glucose in mice without causing lymphadenopathy. *J Lipid Res*. Jul 2022;63(7):100237. doi:10.1016/j.jlr.2022.100237
437. Võsa U, Claringbould A, Westra HJ, et al. Large-scale cis- and trans-eqtl analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. *Nat Genet*. Sep 2021;53(9):1300-1310. doi:10.1038/s41588-021-00913-z
438. Burgess S, Thompson SG. Interpreting findings from mendelian randomization using the mr-egger method. *Eur J Epidemiol*. May 2017;32(5):377-389. doi:10.1007/s10654-017-0255-x

439. Arathimos R, Millard LAC, Bell JA, Relton CL, Suderman M. Impact of sex hormone-binding globulin on the human phenome. *Hum Mol Genet*. Jul 21 2020;29(11):1824-1832. doi:10.1093/hmg/ddz269
440. Rahimi K, Lam CSP, Steinhubl S. Cardiovascular disease and multimorbidity: A call for interdisciplinary research and personalized cardiovascular care. *PLOS Medicine*. 2018;15(3):e1002545. doi:10.1371/journal.pmed.1002545
441. Targher G, Bertolini L, Padovani R, et al. Prevalence of nonalcoholic fatty liver disease and its association with cardiovascular disease among type 2 diabetic patients. *Diabetes Care*. May 2007;30(5):1212-8. doi:10.2337/dc06-2247
442. Muzurović E, Peng CC-H, Belanger MJ, Sanoudou D, Mikhailidis DP, Mantzoros CS. Nonalcoholic fatty liver disease and cardiovascular disease: A review of shared cardiometabolic risk factors. *Hypertension*. 2022/07/01 2022;79(7):1319-1326. doi:10.1161/HYPERTENSIONAHA.122.17982
443. Pingitore P, Romeo S. The role of pnpla3 in health and disease. *Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids*. 2019/06/01/ 2019;1864(6):900-906. doi:https://doi.org/10.1016/j.bbalip.2018.06.018
444. Bastos P, Gomes T, Ribeiro L. Catechol-o-methyltransferase (comt): An update on its role in cancer, neurological and cardiovascular diseases. *Rev Physiol Biochem Pharmacol*. 2017;173:1-39. doi:10.1007/112_2017_2
445. Tsunoda M, Tenhunen J, Tilgmann C, Arai H, Imai K. Reduced membrane-bound catechol-o-methyltransferase in the liver of spontaneously hypertensive rats. *Hypertens Res*. Nov 2003;26(11):923-7. doi:10.1291/hypres.26.923
446. Sauer M, Juranek SA, Marks J, et al. Dhx36 prevents the accumulation of translationally inactive mrnas with g4-structures in untranslated regions. *Nature Communications*. 2019/06/03 2019;10(1):2421. doi:10.1038/s41467-019-10432-5
447. Yoo J-S, Takahasi K, Ng CS, et al. Dhx36 enhances rig-i signaling by facilitating pkr-mediated antiviral stress granule formation. *PLOS Pathogens*. 2014;10(3):e1004012. doi:10.1371/journal.ppat.1004012
448. Jaén RI, Val-Blasco A, Prieto P, et al. Innate immune receptors, key actors in cardiovascular diseases. *JACC Basic Transl Sci*. Jul 2020;5(7):735-749. doi:10.1016/j.jacbts.2020.03.015
449. Barruet E, Morales BM, Cain CJ, et al. Nf-κb/mapk activation underlies acvr1-mediated inflammation in human heterotopic ossification. *JCI Insight*. Nov 15 2018;3(22)doi:10.1172/jci.insight.122958
450. Razani B, Chakravarthy MV, Semenkovich CF. Insulin resistance and atherosclerosis. *Endocrinol Metab Clin North Am*. Sep 2008;37(3):603-21, viii. doi:10.1016/j.ecl.2008.05.001
451. Pott A, Rottbauer W, Just S. Streamlining drug discovery assays for cardiovascular disease using zebrafish. *Expert Opin Drug Discov*. Jan 2020;15(1):27-37. doi:10.1080/17460441.2020.1671351
452. Swerdlow DI, Preiss D, Kuchenbaecker KB, et al. Hmg-coenzyme a reductase inhibition, type 2 diabetes, and bodyweight: Evidence from genetic analysis and randomised trials. *Lancet*. Jan 24 2015;385(9965):351-61. doi:10.1016/s0140-6736(14)61183-1
453. Zou B, Goodwin M, Saleem D, et al. A highly conserved host lipase deacylates oxidized phospholipids and ameliorates acute lung injury in mice. *eLife*. 2021/11/16 2021;10:e70938. doi:10.7554/eLife.70938

454. Alfaddagh A, Martin SS, Leucker TM, et al. Inflammation and cardiovascular disease: From mechanisms to therapeutics. *American Journal of Preventive Cardiology*. 2020/12/01/ 2020;4:100130. doi:https://doi.org/10.1016/j.ajpc.2020.100130
455. Takahashi JS. Transcriptional architecture of the mammalian circadian clock. *Nature Reviews Genetics*. 2017/03/01 2017;18(3):164-179. doi:10.1038/nrg.2016.150
456. Miller S, Kesherwani M, Chan P, et al. Cry2 isoform selectivity of a circadian clock modulator with antiglioblastoma efficacy. *Proceedings of the National Academy of Sciences*. 2022/10/04 2022;119(40):e2203936119. doi:10.1073/pnas.2203936119
457. Mirzaei K, Xu M, Qi Q, de Jonge L, Bray GA, Sacks F, Qi L. Variants in glucose- and circadian rhythm-related genes affect the response of energy expenditure to weight-loss diets: The pounds lost trial. *Am J Clin Nutr*. Feb 2014;99(2):392-9. doi:10.3945/ajcn.113.072066
458. Hirota T, Lee JW, St John PC, et al. Identification of small molecule activators of cryptochrome. *Science*. Aug 31 2012;337(6098):1094-7. doi:10.1126/science.1223710
459. Kalsbeek A, la Fleur S, Fliers E. Circadian control of glucose metabolism. *Mol Metab*. Jul 2014;3(4):372-83. doi:10.1016/j.molmet.2014.03.002
460. Mason IC, Qian J, Adler GK, Scheer F. Impact of circadian disruption on glucose metabolism: Implications for type 2 diabetes. *Diabetologia*. Mar 2020;63(3):462-472. doi:10.1007/s00125-019-05059-6
461. Chow SL, Sasson C, Benjamin IJ, et al. Opioid use and its relationship to cardiovascular disease and brain health: A presidential advisory from the american heart association. *Circulation*. 2021;144(13):e218-e232. doi:doi:10.1161/CIR.0000000000001007
462. Rawal H, Patel BM. Opioids in cardiovascular disease: Therapeutic options. *J Cardiovasc Pharmacol Ther*. Jul 2018;23(4):279-291. doi:10.1177/1074248418757009
463. Ubaldi M, Cannella N, Borruto AM, et al. Role of nociceptin/orphanin fq-nop receptor system in the regulation of stress-related disorders. *Int J Mol Sci*. Nov 30 2021;22(23)doi:10.3390/ijms222312956
464. Torgersen K, Rahman Z, Bahrami S, et al. Shared genetic loci between depression and cardiometabolic traits. *PLoS Genet*. May 2022;18(5):e1010161. doi:10.1371/journal.pgen.1010161
465. Jokinen J, Nordström P. Hpa axis hyperactivity and cardiovascular mortality in mood disorder inpatients. *J Affect Disord*. Jul 2009;116(1-2):88-92. doi:10.1016/j.jad.2008.10.025
466. Wu J, Zhou D, Deng C, Wu X, Long L, Xiong Y. Characterization of porcine eno3: Genomic and cdna structure, polymorphism and expression. *Genet Sel Evol*. Sep-Oct 2008;40(5):563-79. doi:10.1186/1297-9686-40-5-563
467. Lu D, Xia Q, Yang Z, et al. Eno3 promoted the progression of nash by negatively regulating ferroptosis via elevation of gpx4 expression and lipid accumulation. *Ann Transl Med*. Apr 2021;9(8):661. doi:10.21037/atm-21-471
468. Giebelstein J, Poschmann G, Højlund K, et al. The proteomic signature of insulin-resistant human skeletal muscle reveals increased glycolytic and decreased mitochondrial enzymes. *Diabetologia*. 2012/04/01 2012;55(4):1114-1127. doi:10.1007/s00125-012-2456-x
469. Wain LV. Rare variants and cardiovascular disease. *Briefings in Functional Genomics*. 2014;13(5):384-391. doi:10.1093/bfgp/elu010
470. Flannick J, Mercader JM, Fuchsberger C, et al. Exome sequencing of 20,791 cases of type 2 diabetes and 24,440 controls. *Nature*. 2019/06/01 2019;570(7759):71-76. doi:10.1038/s41586-019-1231-2

471. Flannick J. The contribution of low-frequency and rare coding variation to susceptibility to type 2 diabetes. *Curr Diab Rep*. Apr 8 2019;19(5):25. doi:10.1007/s11892-019-1142-5
472. Baselli GA, Jamialahmadi O, Pelusi S, et al. Rare atg7 genetic variants predispose patients to severe fatty liver disease. *Journal of Hepatology*. 2022;77(3):596-606. doi:10.1016/j.jhep.2022.03.031
473. Santoro L, Marchelli D, Cherubini A, et al. Increased burden of inherited irf3 rare genetic variants in europeans with severe nafld. *Digestive and Liver Disease*. 2022;54:S17. doi:10.1016/j.dld.2022.01.032
474. Cornelis MC, Hu FB. Gene-environment interactions in the development of type 2 diabetes: Recent progress and continuing challenges. *Annu Rev Nutr*. Aug 21 2012;32:245-59. doi:10.1146/annurev-nutr-071811-150648
475. Lanktree MB, Hegele RA. Gene-gene and gene-environment interactions: New insights into the prevention, detection and management of coronary artery disease. *Genome Medicine*. 2009/02/26 2009;1(2):28. doi:10.1186/gm28
476. Jonas W, Schürmann A. Genetic and epigenetic factors determining nafld risk. *Molecular Metabolism*. 2021/08/01/ 2021;50:101111. doi:https://doi.org/10.1016/j.molmet.2020.101111
477. Wattacheril JJ, Raj S, Knowles DA, Greally JM. Using epigenomics to understand cellular responses to environmental influences in diseases. *PLOS Genetics*. 2023;19(1):e1010567. doi:10.1371/journal.pgen.1010567
478. Zhou W, Nielsen JB, Fritsche LG, et al. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat Genet*. Sep 2018;50(9):1335-1341. doi:10.1038/s41588-018-0184-y
479. Long L, Zhou X. Defining severe nafld based on icd codes in large cohorts: Balancing feasibility and limitations. *Journal of Hepatology*. 2023;79(6):e232-e233. doi:10.1016/j.jhep.2023.04.034
480. Haas ME, Pirruccello JP, Friedman SN, et al. Machine learning enables new insights into genetic contributions to liver fat accumulation. *Cell Genom*. Dec 8 2021;1(3)doi:10.1016/j.xgen.2021.100066
481. Rich NE, Oji S, Mufti AR, et al. Racial and ethnic disparities in nonalcoholic fatty liver disease prevalence, severity, and outcomes in the united states: A systematic review and meta-analysis. *Clin Gastroenterol Hepatol*. Feb 2018;16(2):198-210.e2. doi:10.1016/j.cgh.2017.09.041
482. Nagar SD, Nápoles AM, Jordan IK, Mariño-Ramírez L. Socioeconomic deprivation and genetic ancestry interact to modify type 2 diabetes ethnic disparities in the united kingdom. *eClinicalMedicine*. 2021;37doi:10.1016/j.eclinm.2021.100960
483. Cheng YJ, Kanaya AM, Araneta MRG, et al. Prevalence of diabetes by race and ethnicity in the united states, 2011-2016. *JAMA*. 2019;322(24):2389-2398. doi:10.1001/jama.2019.19365
484. Chaturvedi N. Ethnic differences in cardiovascular disease. *Heart*. Jun 2003;89(6):681-6. doi:10.1136/heart.89.6.681
485. Kurian AK, Cardarelli KM. Racial and ethnic differences in cardiovascular disease risk factors: A systematic review. *Ethn Dis*. Winter 2007;17(1):143-52.
486. Riazi K, Swain MG, Congly SE, Kaplan GG, Shaheen AA. Race and ethnicity in non-alcoholic fatty liver disease (nafld): A narrative review. *Nutrients*. Oct 28 2022;14(21)doi:10.3390/nu14214556

487. Wainberg M, Sinnott-Armstrong N, Mancuso N, et al. Opportunities and challenges for transcriptome-wide association studies. *Nature Genetics*. 2019/04/01 2019;51(4):592-599. doi:10.1038/s41588-019-0385-z
488. Lawlor DA, Tilling K, Davey Smith G. Triangulation in aetiological epidemiology. *Int J Epidemiol*. Dec 1 2016;45(6):1866-1886. doi:10.1093/ije/dyw314
489. Schwartz AL, Alsan M, Morris AA, Halpern SD. Why diverse clinical trial participation matters. *New England Journal of Medicine*. 2023;388(14):1252-1254. doi:doi:10.1056/NEJMp2215609
490. Verma A, Huffman JE, Rodriguez A, et al. Diversity and scale: Genetic architecture of 2068 traits in the va million veteran program. *Science*. 385(6706):eadj1182. doi:10.1126/science.adj1182
491. Boyle EA, Li YI, Pritchard JK. An expanded view of complex traits: From polygenic to omnigenic. *Cell*. Jun 15 2017;169(7):1177-1186. doi:10.1016/j.cell.2017.05.038
492. Abdellaoui A, Yengo L, Verweij KJH, Visscher PM. 15 years of gwas discovery: Realizing the promise. *Am J Hum Genet*. Feb 2 2023;110(2):179-194. doi:10.1016/j.ajhg.2022.12.011
493. Sanderson E, Rosoff D, Palmer T, Tilling K, Smith GD, Hemani G. Bias from heritable confounding in mendelian randomization studies. *medRxiv*. 2024:2024.09.05.24312293. doi:10.1101/2024.09.05.24312293
494. Koschützki D, Schreiber F. Centrality analysis methods for biological networks and their application to gene regulatory networks. *Gene Regul Syst Bio*. May 15 2008;2:193-201. doi:10.4137/grsb.s702
495. Zhang W, Su C-Y, Yoshiji S, Lu T. Mr corge: Sensitivity analysis of mendelian randomization based on the core gene hypothesis for polygenic exposures. *Bioinformatics*. 2024;40(11):btae666. doi:10.1093/bioinformatics/btae666
496. Said S, Pazoki R, Karhunen V, et al. Genetic analysis of over half a million people characterises c-reactive protein loci. *Nature Communications*. 2022/04/22 2022;13(1):2198. doi:10.1038/s41467-022-29650-5
497. Collaboration CRPCHDG. Association between c reactive protein and coronary heart disease: Mendelian randomisation analysis based on individual participant data. *BMJ*. 2011;342:d548. doi:10.1136/bmj.d548
498. Zacho J, Tybjærg-Hansen A, Jensen JS, Grande P, Sillesen H, Nordestgaard BG. Genetically elevated c-reactive protein and ischemic vascular disease. *New England Journal of Medicine*. 2008;359(18):1897-1908. doi:10.1056/NEJMoa0707402
499. Barbarawi M, Kheiri B, Zayed Y, et al. Vitamin d supplementation and cardiovascular disease risks in more than 83 000 individuals in 21 randomized clinical trials: A meta-analysis. *JAMA Cardiology*. 2019;4(8):765-776. doi:10.1001/jamacardio.2019.1870
500. Bahrami LS, Ranjbar G, Norouzy A, Arabi SM. Vitamin d supplementation effects on the clinical outcomes of patients with coronary artery disease: A systematic review and meta-analysis. *Scientific Reports*. 2020/07/31 2020;10(1):12923. doi:10.1038/s41598-020-69762-w
501. Ashtiani M, Salehzadeh-Yazdi A, Razaghi-Moghadam Z, Hennig H, Wolkenhauer O, Mirzaie M, Jafari M. A systematic survey of centrality measures for protein-protein interaction networks. *BMC Systems Biology*. 2018/07/31 2018;12(1):80. doi:10.1186/s12918-018-0598-2
502. Burgess S, Gill D. Genetic evidence for vitamin d and cardiovascular disease: Choice of variants is critical. *Eur Heart J*. May 7 2022;43(18):1740-1742. doi:10.1093/eurheartj/ehab870

503. Manousaki D, Mokry LE, Ross S, Goltzman D, Richards JB. Mendelian randomization studies do not support a role for vitamin d in coronary artery disease. *Circ Cardiovasc Genet*. Aug 2016;9(4):349-56. doi:10.1161/circgenetics.116.001396
504. Liu C, Li C. C-reactive protein and cardiovascular diseases: A synthesis of studies based on different designs. *European Journal of Preventive Cardiology*. 2023;30(15):1593-1596. doi:10.1093/eurjpc/zwad116
505. Qing X, Jiang J, Yuan C, Wang K. Mendelian randomization analysis identifies a genetic casual association between circulating c-reactive protein and intracerebral hemorrhage. *Journal of Stroke and Cerebrovascular Diseases*. 2024;33(2)doi:10.1016/j.jstrokecerebrovasdis.2023.107554
506. Habibi D, Daneshpour MS, Asgarian S, Kohansal K, Hadaegh F, Mansourian M, Akbarzadeh M. Effect of c-reactive protein on the risk of heart failure: A mendelian randomization study. *BMC Cardiovasc Disord*. Mar 7 2023;23(1):112. doi:10.1186/s12872-023-03149-3
507. Kuppa A, Tripathi H, Al-Darraji A, Tarhuni WM, Abdel-Latif A. C-reactive protein levels and risk of cardiovascular diseases: A two-sample bidirectional mendelian randomization study. *Int J Mol Sci*. May 23 2023;24(11)doi:10.3390/ijms24119129
508. Prins BP, Abbasi A, Wong A, et al. Investigating the causal relationship of c-reactive protein with 32 complex somatic and psychiatric outcomes: A large-scale cross-consortium mendelian randomization study. *PLOS Medicine*. 2016;13(6):e1001976. doi:10.1371/journal.pmed.1001976
509. Darrous L, Hemani G, Davey Smith G, Kutalik Z. Phewas-based clustering of mendelian randomisation instruments reveals distinct mechanism-specific causal effects between obesity and educational attainment. *Nature Communications*. 2024/02/15 2024;15(1):1420. doi:10.1038/s41467-024-45655-8
510. Hu Lt, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*. 1999/01/01 1999;6(1):1-55. doi:10.1080/10705519909540118
511. Lumsden AL, Mulugeta A, Zhou A, Hyppönen E. Apolipoprotein e (apoe) genotype-associated disease risks: A phenome-wide, registry-based, case-control study utilising the uk biobank. *eBioMedicine*. 2020;59doi:10.1016/j.ebiom.2020.102954
512. Cooper LA, Page ST, Amory JK, Anawalt BD, Matsumoto AM. The association of obesity with sex hormone-binding globulin is stronger than the association with ageing--implications for the interpretation of total testosterone measurements. *Clin Endocrinol (Oxf)*. Dec 2015;83(6):828-33. doi:10.1111/cen.12768
513. Basualto-Alarcón C, Llanos P, García-Rivas G, Troncoso MF, Lagos D, Barrientos G, Estrada M. Classic and novel sex hormone binding globulin effects on the cardiovascular system in men. *Int J Endocrinol*. 2021;2021:5527973. doi:10.1155/2021/5527973
514. Yang J, Zhou J, Liu H, et al. Blood lipid levels mediating the effects of sex hormone-binding globulin on coronary heart disease: Mendelian randomization and mediation analysis. *Scientific Reports*. 2024/05/25 2024;14(1):11993. doi:10.1038/s41598-024-62695-8
515. Saez-Lopez C, Villena JA, Simó R, Selva DM. Sex hormone-binding globulin overexpression protects against high-fat diet-induced obesity in transgenic male mice. *J Nutr Biochem*. Nov 2020;85:108480. doi:10.1016/j.jnutbio.2020.108480

516. Sevillano J, Sánchez-Alonso MG, Pizarro-Delgado J, Ramos-Álvarez MDP. Role of receptor protein tyrosine phosphatases (rptps) in insulin signaling and secretion. *Int J Mol Sci*. May 28 2021;22(11)doi:10.3390/ijms22115812
517. Wondmkun YT. Obesity, insulin resistance, and type 2 diabetes: Associations and therapeutic implications. *Diabetes Metab Syndr Obes*. 2020;13:3611-3616. doi:10.2147/dmso.S275898
518. Sharma C, Kim Y, Ahn D, Chung SJ. Protein tyrosine phosphatases (ptps) in diabetes: Causes and therapeutic opportunities. *Arch Pharm Res*. Mar 2021;44(3):310-321. doi:10.1007/s12272-021-01315-9
519. Targher G. Elevated serum gamma-glutamyltransferase activity is associated with increased risk of mortality, incident type 2 diabetes, cardiovascular events, chronic kidney disease and cancer - a narrative review. *Clin Chem Lab Med*. Feb 2010;48(2):147-57. doi:10.1515/cclm.2010.031
520. Alissa EM. Relationship between serum gamma-glutamyltransferase activity and cardiometabolic risk factors in metabolic syndrome. *J Family Med Prim Care*. Mar-Apr 2018;7(2):430-434. doi:10.4103/jfmpc.jfmpc_194_17
521. Paolicchi A, Emdin M, Ghliozeni E, Ciancia E, Passino C, Popoff G, Pompella A. Images in cardiovascular medicine. Human atherosclerotic plaques contain gamma-glutamyl transpeptidase enzyme activity. *Circulation*. Mar 23 2004;109(11):1440. doi:10.1161/01.Cir.0000120558.41356.E6
522. Emdin M, Passino C, Michelassi C, et al. Prognostic value of serum gamma-glutamyl transferase activity after myocardial infarction. *Eur Heart J*. Oct 2001;22(19):1802-7. doi:10.1053/euhj.2001.2807
523. Dominici S, Valentini M, Maellaro E, et al. Redox modulation of cell surface protein thiols in u937 lymphoma cells: The role of gamma-glutamyl transpeptidase-dependent h2o2 production and s-thiolation. *Free Radic Biol Med*. Sep 1999;27(5-6):623-35. doi:10.1016/s0891-5849(99)00111-2
524. Harris EH. Elevated liver function tests in type 2 diabetes. *Clinical Diabetes*. 2005;23(3):115-119. doi:10.2337/diaclin.23.3.115
525. Takemura K, Board PG, Koga F. A systematic review of serum γ-glutamyltransferase as a prognostic biomarker in patients with genitourinary cancer. *Antioxidants*. 2021;10(4):549.
526. Brancaccio M, Russo M, Masullo M, Palumbo A, Russo GL, Castellano I. Sulfur-containing histidine compounds inhibit γ-glutamyl transpeptidase activity in human cancer cells. *Journal of Biological Chemistry*. 2019;294(40):14603-14614. doi:10.1074/jbc.RA119.009304
527. Shimamura Y, Takeuchi I, Terada H, Makino K. Therapeutic effect of ggstop, selective gamma-glutamyl transpeptidase inhibitor, on a mouse model of 5-fluorouracil-induced oral mucositis. *Anticancer Research*. 2019;39(1):201. doi:10.21873/anticanres.13098

APPENDIX 1: PUBLICATIONS

Provided in this appendix are full PDFs of selected first author or co-first author publications completed during my PhD studies (Summer 2020—December 2024) and a complete curriculum vitae formatted list of first-author or contributing author publications and presentations.

Selected first author, or co-first author publications

In this section, I provide completed published manuscript files (PDF versions) of selected first author or co-first author publications. I have organized the publications in the following themes:

1. PCSK9 and HMGCR drug-target MR studies
2. Alcohol consumption
3. Aging and longevity
4. Drug-target characterization

Manuscripts Related to PCSK9 and HMGCR Inhibition

Title: Evaluating the Cardiovascular Impact of Genetically Proxied PCSK9 and HMGCR Inhibition in East Asian and European Populations: A Drug-Target Mendelian Randomization Study

Title: Assessing the Impact of PCSK9 and HMGCR Inhibition on Liver Function: Drug-Target Mendelian Randomization Analyses in Four Ancestries

Title: Mendelian Randomization Study of

PCSK9 and HMG-CoA Reductase Inhibition and Cognitive Function

Manuscripts Related to Alcohol Consumption Behaviors

Title: Association of High-Intensity Binge Drinking With Lipid and Liver Function Enzyme Levels

Title: Evaluating the Relationship Between Alcohol Consumption, Tobacco Use, and Cardiovascular Disease: A Multivariable

Mendelian Randomization Study

Manuscripts Related to Aging and Longevity

Title: Multivariate genome-wide analysis of aging-related traits identifies novel loci and new drug targets for healthy aging

Title: Major Psychiatric Disorders, Substance Use Behaviors, and Longevity

Title: Multi-omic underpinnings of epigenetic aging and human longevity

Manuscripts Related to New Drug-Target Characterization

Title: Bidirectional Mendelian Randomization Highlights Causal Relationships Between Circulating INHBC and Multiple Cardiometabolic Diseases and Traits

Complete Publication List And Poster/Presentation Titles

This section contains the references to my complete list of publications (both first author and contributing author) and conference/meeting (poster and presentations).

Publications

1. **Rosoff DB**, Wagner J, Bell AS, Mavromatis LA, Jung J, Lohoff FW. A multi-omics Mendelian randomization study identifies new therapeutic targets for alcohol use disorder and problem drinking. *Nature Human Behaviour*. 2024/11/11 2024;doi:10.1038/s41562-024-02040-1

2. **Rosoff DB**, Wagner J, Jung J, et al. Multi-Omics Mendelian Randomization Study Investigating the Impact of PCSK9 and HMGCR Inhibition on Type 2 Diabetes Across Five Populations. *Diabetes*. 2024:db240451. doi:10.2337/db24-0451

3. Loh NY*, **Rosoff DB*,** Richmond R, et al. Bidirectional Mendelian randomization highlights causal relationships between circulating INHBC and multiple cardiometabolic diseases and traits. *Diabetes*. 2024:db240168. doi:10.2337/db24-0168 **(*Shared first-authorship)**

4. **Rosoff DB*,** Hamandi AM*, Bell AS, et al. Major Psychiatric Disorders, Substance Use Behaviors, and Longevity. *JAMA Psychiatry*. Sep 1 2024;81(9):889-901. doi:10.1001/jamapsychiatry.2024.1429 **(*Shared first-authorship)**

5. **Rosoff DB,** Bell AS, Mavromatis LA, et al. Evaluating the Cardiovascular Impact of Genetically Proxied PCSK9 and HMGCR Inhibition in East Asian and European Populations: A Drug-Target Mendelian Randomization Study. *Circulation: Genomic and Precision Medicine*. 2024/02/01 2024;17(1):e004224. doi:10.1161/CIRCGEN.122.004224

6. **Rosoff DB,** Mavromatis LA, Bell AS, et al. Multivariate genome-wide analysis of aging-related traits identifies novel loci and new drug targets for healthy aging. *Nature Aging*. 2023/08/01 2023;3(8):1020-1035. doi:10.1038/s43587-023-00455-5

7. Mavromatis LA,* **Rosoff DB,*** Bell AS, Jung J, Wagner J, Lohoff FW. Multi-omic underpinnings of epigenetic aging and human longevity. *Nature Communications*. 2023/04/19 2023;14(1):2236. doi:10.1038/s41467-023-37729-w (*shared first-authorship)

8. **Rosoff DB,* Bell AS,*** Wagner J, et al. Assessing the Impact of PCSK9 and HMGCR Inhibition on Liver Function: Drug-Target Mendelian Randomization Analyses in Four Ancestries. *Cellular and Molecular Gastroenterology and Hepatology*. doi:10.1016/j.jcmgh.2023.09.001 (*shared first-authorship)

9. Wagner J, Park LM, Mukhopadhyay P, … **Rosoff DB**, et al. PCSK9 inhibition attenuates alcohol-associated neuronal oxidative stress and cellular injury. *Brain Behav Immun*. Jul 2024;119:494-506. doi:10.1016/j.bbi.2024.04.022

10. Cuozzo F, Viloria K, Shilleh AH, … **Rosoff DB**, et al. LDHB contributes to the regulation of lactate levels and basal insulin secretion in human pancreatic β cells. *Cell Reports*. 2024;43(4)doi:10.1016/j.celrep.2024.114047

11. Loh NY, **Rosoff D**, Noordam R, Christodoulides C. Investigating the impact of metabolic syndrome traits on telomere length: a Mendelian randomization study. *Obesity*. 2023;31(8):2189-2198. doi:https://doi.org/10.1002/oby.23810

12. Bell, A. S., **D. B. Rosoff**, L. A. Mavromatis, J. Jung, J. Wagner and F. W. Lohoff (2022). "Comparing the Relationships of Genetically Proxied PCSK9 Inhibition With Mood Disorders, Cognition, and Dementia Between Men and Women: A Drug-Target Mendelian Randomization Study." Journal of the American Heart Association **11**(21): e026122.

13. Mavromatis, L. A., **D. B. Rosoff**, R. B. Cupertino, H. Garavan, S. Mackey and F. W. Lohoff (2022). "Association Between Brain Structure and Alcohol Use Behaviors in Adults: A Mendelian Randomization and Multiomics Study." JAMA Psychiatry **79**(9): 869-878.

14. **Rosoff Daniel, B**., S. Bell Andrew, J. Jung, J. Wagner, A. Mavromatis Lucas and W. Lohoff Falk (2022). "Mendelian Randomization Study of PCSK9 and HMG-CoA Reductase Inhibition and Cognitive Function." Journal of the American College of Cardiology **80**(7): 653-662.

15. Jung, J., D. L. McCartney, J. Wagner, **D. B. Rosoff**, M. Schwandt, H. Sun, C. E. Wiers, L. M. de Carvalho, N. D. Volkow, R. M. Walker, A. Campbell, D. J. Porteous, A. M. McIntosh, R. E. Marioni, S. Horvath, K. L. Evans and F. W. Lohoff (2022). "Alcohol use disorder is associated with DNA methylation-based shortening of telomere length and regulated by TESPA1: implications for aging." Molecular Psychiatry.

16. Jung, J., D. L. McCartney, J. Wagner, J. Yoo, A. S. Bell, L. A. Mavromatis, **D. B. Rosoff**, C. A. Hodgkinson, H. Sun, M. Schwandt, N. Diazgranados, A. K. Smith, V. Michopoulos, A. Powers, J. Stevens, B. Bradley, N. Fani, R. M. Walker, A. Campbell, D. J. Porteous, A. M. McIntosh, S. Horvath, R. E. Marioni, K. Evans, D. Goldman and F. W. Lohoff "Additive effects of stress and alcohol exposure on accelerated epigenetic aging in Alcohol Use Disorder." Biological Psychiatry.

17. Lohoff, F. W., T.-K. Clarke, Z. A. Kaminsky, R. M. Walker, M. L. Bermingham, J. Jung, S. W. Morris, **D. Rosoff**, A. Campbell, M. Barbu, K. Charlet, M. Adams, J. Lee, D. M. Howard, E. M. O'Connell, H. Whalley, D. J. Porteous, A. M. McIntosh and K. L. Evans (2022). "Epigenome-wide association study of alcohol consumption in N = 8161 individuals and relevance to alcohol use disorder pathophysiology: identification of the cystine/glutamate transporter SLC7A11 as a top target." Molecular Psychiatry **27**(3): 1754-1764.

18. **Rosoff, D. B.,** J. Yoo and F. W. Lohoff (2021). "Smoking is significantly associated with increased risk of COVID-19 and other respiratory infections." Communications Biology **4**(1): 1230.

19. **Rosoff DB**, Davey Smith G, Mehta N, Clarke T-K, Lohoff FW. Evaluating the relationship between alcohol consumption, tobacco use, and cardiovascular disease: A multivariable Mendelian randomization study. *PLOS Medicine.* 2020;17(12):e1003410.

20. **Rosoff DB**, Smith GD, Lohoff FW. Prescription Opioid Use and Risk for Major Depressive Disorder and Anxiety and Stress-Related Disorders: A Multivariable Mendelian Randomization Analysis. *JAMA Psychiatry.* 2020.

21. **Rosoff DB**, Kaminsky ZA, McIntosh AM, Davey Smith G, Lohoff FW. Educational attainment reduces the risk of suicide attempt among individuals with and without psychiatric disorders independent of cognition: a bidirectional and multivariable Mendelian randomization study with more than 815,000 participants. *Translational Psychiatry.* 2020;10(1):388.

22. **Rosoff DB**, Clarke T-K, Adams MJ, et al. Educational attainment impacts drinking behaviors and risk for alcohol dependence: results from a two-sample Mendelian randomization study with ~780,000 participants. *Molecular Psychiatry.* 2021;26(4):1119-1132.

23. **Rosoff DB**, Charlet K, Jung J, et al. Lipid profile dysregulation predicts alcohol withdrawal symptom severity in individuals with alcohol use disorder. *Alcohol.* 2020;86:93-101.

24. Jung J*, **Rosoff DB***, Muench C, et al. Adverse Childhood Experiences are Associated with High-Intensity Binge Drinking Behavior in Adulthood and Mediated by Psychiatric Disorders. *Alcohol Alcohol.* 2020;55(2):204-214 (*shared first-authorship).

25. Lohoff FW, Roy A, Jung J, et al. (**Rosoff DB 4th author)** Epigenome-wide association study and multi-tissue replication of individuals with alcohol use disorder: evidence for abnormal glucocorticoid signaling pathway gene regulation. *Molecular Psychiatry.* 2020.

26. **Rosoff DB**, Charlet K, Jung J, et al. Association of High-Intensity Binge Drinking With Lipid and Liver Function Enzyme Levels. *JAMA Network Open.* 2019;2(6):e195844-e195844.

27. Luo, A., J. Jung, M. Longley, **D. B. Rosoff**, K. Charlet, C. Muench, J. Lee, C. A. Hodgkinson, D. Goldman, S. Horvath, Z. A. Kaminsky and F. W. Lohoff (2019). "Epigenetic aging is accelerated in alcohol use disorder and regulated by genetic variation in APOL2." Neuropsychopharmacology. https://www.nature.com/articles/s41386-019-0500-y

28. Lee, J. S., **D. Rosoff**, A. Luo, M. Longley, M. Phillips, K. Charlet, C. Muench, J. Jung and F. W. Lohoff (2019). "PCSK9 is Increased in Cerebrospinal Fluid of Individuals With Alcohol Use Disorder." Alcohol Clin Exp Res **43**(6): 1163-1169. https://onlinelibrary.wiley.com/doi/abs/10.1111/acer.14039

29. Lee, J. S., J. L. Sorcher, A. D. Rosen, R. Damadzic, H. Sun, M. Schwandt, M. Heilig, J. Kelly, K. L. Mauro, A. Luo, **D. Rosoff**, C. Muench, J. Jung, Z. A. Kaminsky and F. W. Lohoff (2018). "Genetic Association and Expression Analyses of the Phosphatidylinositol-4-Phosphate 5-Kinase (PIP5K1C) Gene in Alcohol Use Disorder-Relevance for Pain Signaling and Alcohol Use." Alcoholism, clinical and experimental research **42**(6): 1034-1043.https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6134400/

30. Muench, C., A. Luo, K. Charlet, J. Lee, **D. B. Rosoff**, H. Sun, S. J. Fede, J. Jung, R. Momenan and F. W. Lohoff (2019). "Lack of Association Between Serotonin Transporter Gene (SLC6A4) Promoter Methylation and Amygdala Response During Negative Emotion Processing in

Individuals With Alcohol Dependence." Alcohol and Alcoholism **54**(3): 209-215.https://academic.oup.com/alcalc/article/54/3/209/5476013

31. Mauro, K. L., S. G. Helton, **D. B. Rosoff**, A. Luo, M. Schwandt, J. Jung, J. Lee, C. Muench and F. W. Lohoff (2018). "Association Analysis Between Genetic Variation in GATA Binding Protein 4 (GATA4) and Alcohol Use Disorder." Alcohol and Alcoholism **53**(4): 361-367. https://academic.oup.com/alcalc/article-lookup/doi/10.1093/alcalc/agx120

Abstracts, Presentations, and Posters

1. **Rosoff, D.B.** Cholesterol lowering and Type 2 DM risk: a Mendelian randomization study (presentation), Regional Lipidologists Meeting, Pembroke College, University of Oxford, United Kingdom, November 21st, 2024

2. **Rosoff, D. B**., Ray, D., Lohoff, F. W., & Davey Smith, G. (2024). Evaluating the impact of C-reactive protein levels on cardiovascular disease: a comparison of Mendelian randomization estimates across different methods and instrument inclusion. Paper presented at the *Mendelian Randomization Conference*, University of Bristol, United Kingdom, June 19, 2024.

3. Reitz, J.,* **Rosoff, D. B.*,** Perlstein, T., Jung, J., Wagner, J., & Lohoff, F. W. (2024). Genetically proxied fibroblast growth factor 21 (FGF21) and cannabis use disorder: a Mendelian randomization analysis. Paper presented at the NIAAA Fellows Day (short talk) and NIH Post-baccalaureate Poster Day, Bethesda, MD, May 2024. (*shared first authorship)

4. Perlstein, T.,* **Rosoff, D. B.,*** Reitz, J., Jung, J., Wagner, J., & Lohoff, F. W. (2024). Assessing the neurocognitive profile of fibroblast growth factor 21: a drug-target Mendelian randomization study. Paper presented at the NIAAA Fellows Day (short talk) and NIH Post-baccalaureate Poster Day, Bethesda, MD, May 2024. (*shared first authorship)

5. **Rosoff, D.,** Wagner, J., Bell, A., Mavromatis, L., Jung, J., & Lohoff, F. (2024). Multi-Omics Analyses of Cortical Proteome and Single Cell-Type Transcriptome Identifies Novel Drug Targets for Problematic Drinking Behavior and Alcohol Use Disorder. Paper presented at the *NIDA Genetics and Epigenetics Cross-Cutting Research Team Meeting*, Bethesda, Maryland, May 23, 2024.

6. **Rosoff, D.,** Wagner, J., Bell, A., Mavromatis, L., Jung, J., & Lohoff, F. (2024). Multi-Omics Analyses of Cortical Proteome and Single Cell-Type Transcriptome Identifies Novel Drug Targets for Problematic Drinking Behavior and Alcohol Use Disorder. Paper presented at the *Biological Psychiatry Annual Meeting*, Austin, Texas, May 10, 2024. doi:10.1016/j.biopsych.2024.02.055

7. **Rosoff, D. B**., Ray, D., Lohoff, F. W., & Davey Smith, G. (2024). Evaluating the impact of C-reactive protein levels on cardiovascular disease: a comparison of Mendelian randomization estimates across different methods and instrument inclusion.

Paper presented at the *Mendelian Randomization Methods Meeting*, University of Bristol, United Kingdom, January 26, 2024

8. **Rosoff, D. B**., Wagner, J., Park, L., Hamandi, A., Perlstein, T., Jung, J., Pacher, P., Christodoulides, C., Davey Smith, G., Ray, D., & Lohoff, F. W. (2024). Multi-ancestry Mendelian randomization study investigating relationships of PCSK9 and HMGCR inhibition with type 2 diabetes. Paper presented at the 9th Meeting of the Study Group on Genetics of Diabetes (SGGD), University of Exeter, United Kingdom, April 17, 2024.

9. **Rosoff, D. B.,** Mavromatis, L. A., Bell, A. S., Wagner, J., Jung, J., Marioni, R. E., Davey Smith, G., Horvath, S., & Lohoff, F. W. (2023). Multivariate genome-wide analysis of aging-related traits identifies novel loci and new drug targets for healthy aging. Paper presented at the *ACNP 2023, December 5, Tampa, Florida*

10. **Rosoff, D. B.,** Mavromatis, L. A., Park, L., Bell, A. S., Hamandi, A., Mukhopadhyay, P., Pacher, P., Jung, J., Wagner, J., & Lohoff, F. W. (2023). Association between plasma PCSK9 levels and stroke risk: A multivariable Mendelian randomization study. Paper presented at Society for Neuroscience, November 113, 2023, Washington DC, USA.

11. **Rosoff, D. B.,** Mavromatis, L. A., Park, L., Bell, A. S., Hamandi, A., Mukhopadhyay, P., Pacher, P., Jung, J., Wagner, J., & Lohoff, F. W. (2023). Association between plasma PCSK9 levels and stroke risk: A multivariable Mendelian randomization study. Paper presented at WCPG, October 12, 2023, Montreal, Canada.

12. **D.B. Rosoff,** F. Lohoff, D. Ray. "Evaluating the long-term hepatic efficacy and cardiometabolic safety profiles of novel non-alcoholic fatty liver disease therapeutics: a drug-target Mendelian randomization study." 6/21/2023, EASL Congress, Vienna, Austria.

13. L. Daniels, M. Neville, L. Hodson**, D.B. Rosoff**, M. Ruby, D. Ray. "Circadian Regulation of Hepatic Energy Metabolism: ROR regulation of Glucose Signaling Pathways." 6/2023, Novo Nordisk Fellow Symposium, Denmark.

14. L. Daniels, M. Neville, L. Hodson, **D.B. Rosoff,** M. Ruby, D. Ray. "Circadian Regulation of Hepatic Energy Metabolism: ROR regulation of Glucose Signaling Pathways." 3/6/2023, Radcliffe Department of Medicine Symposium, University of Oxford, UK.

15. **D.B. Rosoff,** F. Lohoff, D. Ray. “Evaluaing the long-term hepatic efficacy and cardiometabolic safety profiles of novel non-alcoholic fatty liver disease therapeutics: a drug-target Mendelian randomization study”. 3/6/2023. Radcliffe Department of Medicine Symposium, University of Oxford, UK

16. **D.B. Rosoff.** F. Lohoff, D. Ray. "Proteomic underpinnings of insomnia: a cis-instrument Mendelian randomization analysis." 4/17/2023, UK Clock Club, Surrey, UK

17. **D. B. Rosoff**, J. Wagner, A. Bell, L. Mavromatis, J. Jung, F. Lohoff. "Novel Therapeutic Drug Target Discoveries for Alcohol Use Disorder and Problem Drinking: A Multi-Level Proteome-Wide Mendelian Randomization Study." 4/28/2023, Society of Biological Psychiatry, San Diego, CA.

18. **Rosoff, D.,** J. Wagner, A. Bell, L. Mavromatis, J. Jung and F. Lohoff. "Novel Therapeutic Drug Target Discoveries for Alcohol Use Disorder and Problem Drinking: A Multi-Level Proteome-Wide Mendelian Randomization Study." 12/5/2022, Pheonix, AZ

19. **Rosoff, D.,** J. Yoo, A. Bell, L. Mavromatis, J. Jung, J. Wagner and F. Lohoff (2022). "P16. A Multivariable Mendelian Randomization Study Disentangling the Relationships Between Neuropsychiatric Disorders, Substance Use Behaviors, and Longevity." Society of Biological Psychiatry

20. **D. B. Rosoff,** A. Bell, L. Mavromatis, J. Jung, J. Wagner, F. Lohoff. Transcriptome-Wide Association Study Examining Alcohol Consumption, Tobacco Smoking, and Cannabis Use Yields Insight into the Etiology of Substance Use, and Identifies Drug Repositioning Opportunities." 9/15/2022, ISPG, World Congress of Psychiatric Genetics, Florence, Italy

21. **D. B. Rosoff,** J. Yoo, A. Bell, L. Mavromatis, J. Jung, J. Wagner, F. Lohoff. "A Multivariable Mendelian Randomization Study Disentangling the Relationships Between Neuropsychiatric Disorders, Substance Use Behaviors, and Longevity." September 15, 2022, ISPG, World Congress of Psychiatric Genetics, Florence, Italy

22. **Rosoff, D.B**, J. Yoo, A. Bell, L. Mavromatis, J. Jung, J. Wagner and F. Lohoff. "A Multivariable Mendelian Randomization Study Disentangling the Relationships Between Neuropsychiatric Disorders, Substance Use Behaviors, and Longevity." 4/29/2022, Society of Biological Psychiatry, virtual

23. J. Yoo, **D.B. Rosoff**, F. Lohoff. "A Genetically-Informed Evaluation of the Relationships Between Tobacco Smoking, Cannabis Use, Alcohol Consumption, and Respiratory Infections, Including COVID-19." 10/13/2021. ISPG, World Congress of Psychiatric Genetics, virtual

24. **D.B. Rosoff,** G.D. Smith, F.W. Lohoff. Prescription Opioid Use and Risk for Major Depressive Disorder and Anxiety and Stress-Related Disorders: A Multivariable

Mendelian Randomization Analysis. 4/30/2021, Society of Biological Psychiatry Conference, virtual.

25. **D.B. Rosoff,** G. Davey Smith, N. Mehta, T.-K. Clarke, F.W. Lohoff. Evaluating the relationship between alcohol consumption, tobacco use, and cardiovascular disease: A multivariable Mendelian randomization study. 6/21/2021. Royal Society on Alcoholism Conference, virtual.

26. **D.B. Rosoff,** G.D. Smith, F.W. Lohoff. Prescription Opioid Use and Risk for Major Depressive Disorder and Anxiety and Stress-Related Disorders: A Multivariable Mendelian Randomization Analysis. 3/9/2021. National Institute on Drug Abuse (NIDA) Genomics Consortium Conference, virtual.

27. F. Lohoff, A. Roy, J. Jung, M. Longley, **D. B. Rosoff**, A. Luo, E. P'Connell, J. Sorcher, H. Sun, M. Schwandt, C. Hodgkinson, D. Goldman, R. Momenan, A. McIntosh, M. Adams, R. Walker, K. Evans, D. Porteous, A. Smith, J. Lee, C. Muench, K. Charlet, T. Clarke and Z. Kaminsky. "Epigenome-Wide Association Study and Multi-Tissue Replication of Individuals With Alcohol Use Disorder: Evidence for Abnormal Glucocorticoid Signaling Pathway Gene Regulation." 2020 Society of Biological Psychiatry, virtual

28. Z. Kaminsky, M. Longley, **D.B. Rosoff**, A. Luo, H. Sun, C. Hodgkinson, D. Goldman, M. Schwandt, I. Lucki, C. Browne, J. Lee, K. Charlet, C. Muench and F. Lohoff (2020). "Epigenome-Wide Association Study of Individuals With Alcohol Use Disorder and Suicidal Behavior: Identification and Replication of TIMP2 as Novel Target." 2020 Society of Biological Psychiatry, virtual

29. **D.B. Rosoff,** T.-K. Clarke, M.J. Adams, A.M. McIntosh, G.D. Smith, J. Jung, F.W. Lohoff. "Educational attainment impacts drinking behaviors and risk for alcohol dependence: results from a two-sample Mendelian randomization study with ~780,000 participants". 10/30/2019, World Congress of Psychiatric Genetics Conference, Anaheim, CA.

30. **D.B. Rosoff,** T.-K. Clarke, M.J. Adams, A.M. McIntosh, G.D. Smith, J. Jung, F.W. Lohoff. "Educational attainment impacts drinking behaviors and risk for alcohol dependence: results from a two-sample Mendelian randomization study with ~780,000 participants". 7/14/2019, MRC-IEU Mendelian Randomization Conference, Bristol, UK.

31. C. Muench, K. Charlet, J. S. Lee, J. Jung, **D. B. Rosoff,** M. Longley, N. L. Balderston, C. R. Cortes, C. Grillon, R. Momenan, F. W. Lohoff. "Functional Neuronal Alterations During Fear Conditioning and Extinction Recall in Alcohol-Dependent and Healthy

Individuals With and Without Early Life Stress." 5/17, 2019, Society of Biological Psychiatry, Chicago, IL.

32. **D.B. Rosoff,** J. Jung, C. Muench, A. Luo, M. Longley, J.-S. Lee, K. Charlet, F.W. Lohoff. "Adverse childhood experiences predict high-intensity binge drinking behavior in adulthood." 5/3/2018, NIH Postbaccalaureate Poster Day, Bethesda, MD.

33. Luo, J. Lee, **D.B. Rosoff**, K. Mauro, B.A. Blank, L.F. Vendruscolo, G.F. Koob, F.W. Lohoff. "Regulation of PCSK9 in Rodent Models Using Liquid Alcohol Diet." 5/3/2018, NIH Postbaccalaureate Poster Day, Bethesda, MD.

34. K. Mauro, C. Muench, J. Lee, M. Schwandt, A. Rosen, J. Sorcher, **D.B. Rosoff**, F.W. Lohoff. "Association analysis between genetic variation in the FKBP5 gene and cortisol levels in alcohol dependence." 11/14/2017, Society For Neuroscience Conference, Washington, DC

APPENDIX 2 (AP2): SUPPLEMENTARY MATERIALS FOR AIM 1

AP2.1. Aim 1 Supplementary Tables

Below are titles to each for the Aim 1 Supplementary Tables that are formatted as Excel files uploaded separately.

Table AP3.1. Description and sources of phenotypes GWAS summary statistics

Table AP3.2. Multi-ancestry PCSK9 and HMGCR instruments constructed in ancestry-specific GWAS of LDL-C levels

Table AP3.3. Gain-of-function, loss-of-function SNPs used as instruments in sensitivity analyses

Table AP3.4. eQTL and pQTL PCSK9 instruments constructed using GWASs of circulating PCSK9 protein levels, pancreatic *PCSK9* gene expression, and liver *PCSK9* expression

Table AP3.5. Average R2 and F-statistic information for instruments used in the study

Table AP3.6. Drug-target Mendelian randomization on type 2 diabetes and glycemic traits

Table AP3.7. Single variable Mendelian randomization of gain-of-function PCSK9 instruments on type 2 diabetes and glycemic traits

Table AP3.8. ABF and SuSiE colocalization for ancestry specific PCSK9 and HMGCR drug target MR estimates surpassing P-value threshold 0.05

Table AP3.9. Eastern ancestry multivariable MR of PCSK9 and HMGCR cis-instruments in LDL-C and BMI

Table AP3.10. South Asian ancestry multivariable MR of PCSK9 and HMGCR cis-instruments in LDL-C and BMI

Table AP3.11. African ancestry multivariable MR of PCSK9 and HMGCR cis-instruments in LDL-C and BMI

Table AP3.12. Hispanic ancestry multivariable MR of PCSK9 and HMGCR cis-instruments in LDL-C and BMI

Table AP3.13. European ancestry multivariable MR of PCSK9 and HMGCR cis-instruments in LDL-C and BMI

Table AP3.14. Impact of Mendelian randomization of LDL-C lowering via PCSK9 and HMGCR variants on Stumvoll insulin sensitivity index (ISI) and insulin fold change (IFC) (European ancestry only)

Table AP3.15. Correlated Mendelian randomization of quantitative trait loci (QTL) PCSK9 instruments on type 2 diabetes and glycemic traits in EUR participants

APPENDIX 3 (AP3): SUPPLEMENTARY MATERIALS FOR AIM 2

AP3.1. Supplementary Aim 1 Discussion

In this supplementary discussion section, we expand upon the Aim 2 limitations outlined in the **Chapter 4 Discussion**. First, while cis-instrument MR is less prone to horizontal pleiotropy than polygenic MR,[6,50] and sensitivity analyses showed consistent MR estimates across methods used

to evaluate these assumptions, there remains potential biases due to confounding or pleiotropy. Further, while results from complementary MR methods and sensitivity analyses (i.e., colocalization) suggested that the MR assumptions are plausible for these analyses, we underscore the importance of interpreting the MR results through the lens of the underlying assumptions and additional constraints of the cis-instrument MR model, including restricted interpretation of the identified relationships to the 'on-target' effects of the genes, the assumption that there is no gene-environment or gene-gene interactions, and that the relationships are linear.[50] Second, we based our instrument construction upon the cis-instrument/drug-target MR framework outlined by Schmidt et al.,[50] which included using cis-QTLs associated with their respective protein or gene at lower P-values than the conventional genome-wide significance (P-value < 5×10^{-8}), which may increase the chance of including variants in the instrumental variables that do not impact the exposure;[50] however, this method may instead diminish power and reduce precision,[50] and many drug-target MR and other genetics-based studies have been shown including these variants improve model performance.[50,200,237,239,413,414] While each of the SNPs had F-statistics >10, the conventional threshold to indicate whether there is evidence for weak-instrument bias,[239] one concern with including SNPs associated with their exposure at a relaxed P-value threshold is weak-instrument bias, but in two-sample MR analyses like those performed in this study any possible weak-instrument bias would attenuate results towards the null.[242,415] Further, while the findings from the eQTL of the 8 canonical brain cell types provide new insights into cellular diversity and function related to the genetic liability of PAU and other alcohol use behaviors, the single-cell data eQTL data sourced for this study, and more broadly, single-cell transcriptomic data in general has several technical and methodological challenges[416] that are important to consider.

AP3.1.1. Aim 1 Additional Limitations

There are additional limitations related to the single-cell transcriptomic data, which includes potential amplification bias due to the necessity of amplifying the minute RNA quantities found in individual cells, which can skew interpretations of cellular states and biological variability.[416] Additionally, data sparsity resulting from dropout events, where transcripts present in cells are not detected, can lead to an overrepresentation of zero counts, affecting data analyses.[417] Conventional scRNA-seq methods (like those used to generate eQTLs for the eight brain cell types[102] sourced for this study) also result in the loss of spatial information within bulk tissue.[418] Although newer spatial transcriptomics methods[418] promise to retain this spatial context and will further inform cell-type eQTL analysis, they too introduce challenges in how to effectively utilize the additional spatial information for cell type identification and classification.[419] Moreover, the current classification of identified cells relies on unsupervised clustering due to the incompleteness of cell type reference atlases. This reliance can lead to misinterpretations in the biological meaning of clusters identified through these algorithms, potentially complicating the understanding of cellular identities and functions.[416] Also, because the QTL and GWAS data were comprised from common variants, we were unable to assess, or compare, these relationships against data from rare variants. Because rare variants might have a more direct functional impact on protein function and gene regulation, which is important to inform our understanding the pathophysiology of AUD,[420] we underscore the need to perform these comparisons when the data becomes available. Further, colocalization assuming a single causal variant is likely overly conservative and may have missed potential signals. While this

assumption might lead to overlooking certain genetic signals, we chose this method due to the constraints of our data. The recent advancements in colocalization methods, such as SuSiE coloc, require a high density of SNPs.[421] However, these methods also have a known risk of inflated false positive rates without sufficient SNPs in the region of interest.[421] Here, the data derived from the cortical pQTLs provided a SNP range of 40 to ~600 per locus, which was not sufficient for the application of these newer methods.[120] Additionally, considering that the majority of proteome-wide or transcriptome-wide studies have used the standard colocalization method with a single causal variant assumption, our approach aligns with the established practices in the field, potentially enhancing the comparability of our findings with the existing body of literature. Also, the cortical proteomic data were derived from three cortical regions, so future proteomics-based analyses are necessary when proteomic data from other brain regions become available. Relatedly, available gene expression instruments in the cell-type analyses varied, with many more genes available in, e.g., EXC and INH cell types than in PER or END, and our resulting screen may have missed causal genes in these cell types. Third, apart from PAU GWAS data, which included AUD diagnoses, the other alcohol consumption data were derived from self-report; as with any study in the substance use field using self-reported data, these variables may be either under- or over-reported,[185] which could impact the findings; however, our replication of the targets in EHR-based data, may help address this potential bias. Further, these alcohol-related variables only account for recent drinking behaviors (i.e., within the past 12 months), and because alcohol consumption may vary over the life course,[422] our results may not be generalizable to longer-term alcohol consumption patterns. Finally, analyses were based upon participants of white, European ancestry. More specifically, the AIF, DPW, and binge drinking phenotypes were derived primarily from participants of European ancestry in the UK Biobank, reportedly more educated, with healthier lifestyles, and fewer health problems than the general UK population.[423] Therefore, we underscore the limited generalizability of study findings to other populations, especially as it has been shown that drinking behaviors vary across race/ethnicity around the world.[424]

AP3.2. Aim 2 Supplementary Figures

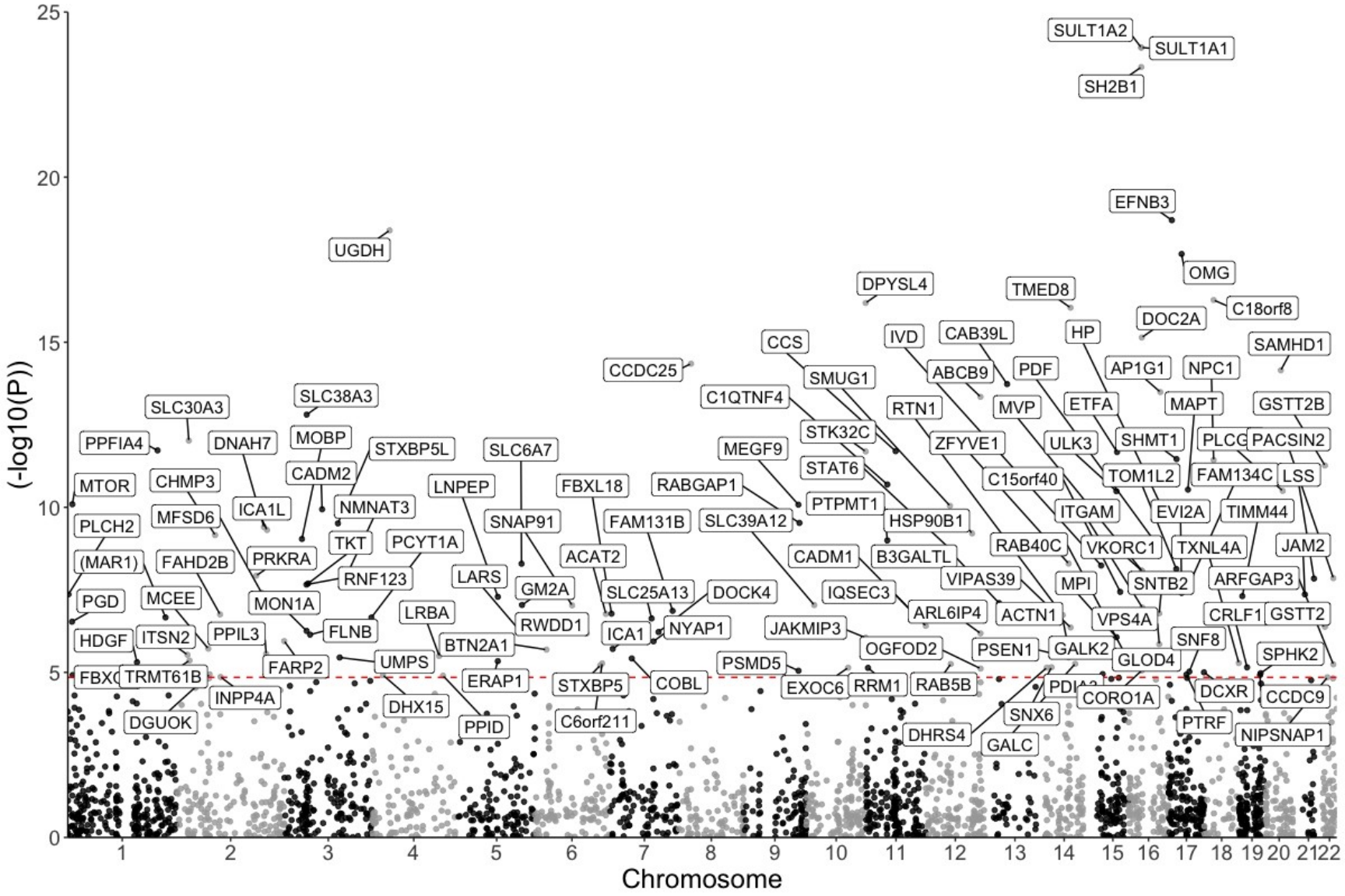


Figure AP3.1: Manhattan plot of the cortical proteomic cis-instrument Mendelian randomization results on alcohol intake frequency. *Labeled genes surpassed correction for multiple comparisons (P-value < 1.41×10^{-5}.)*

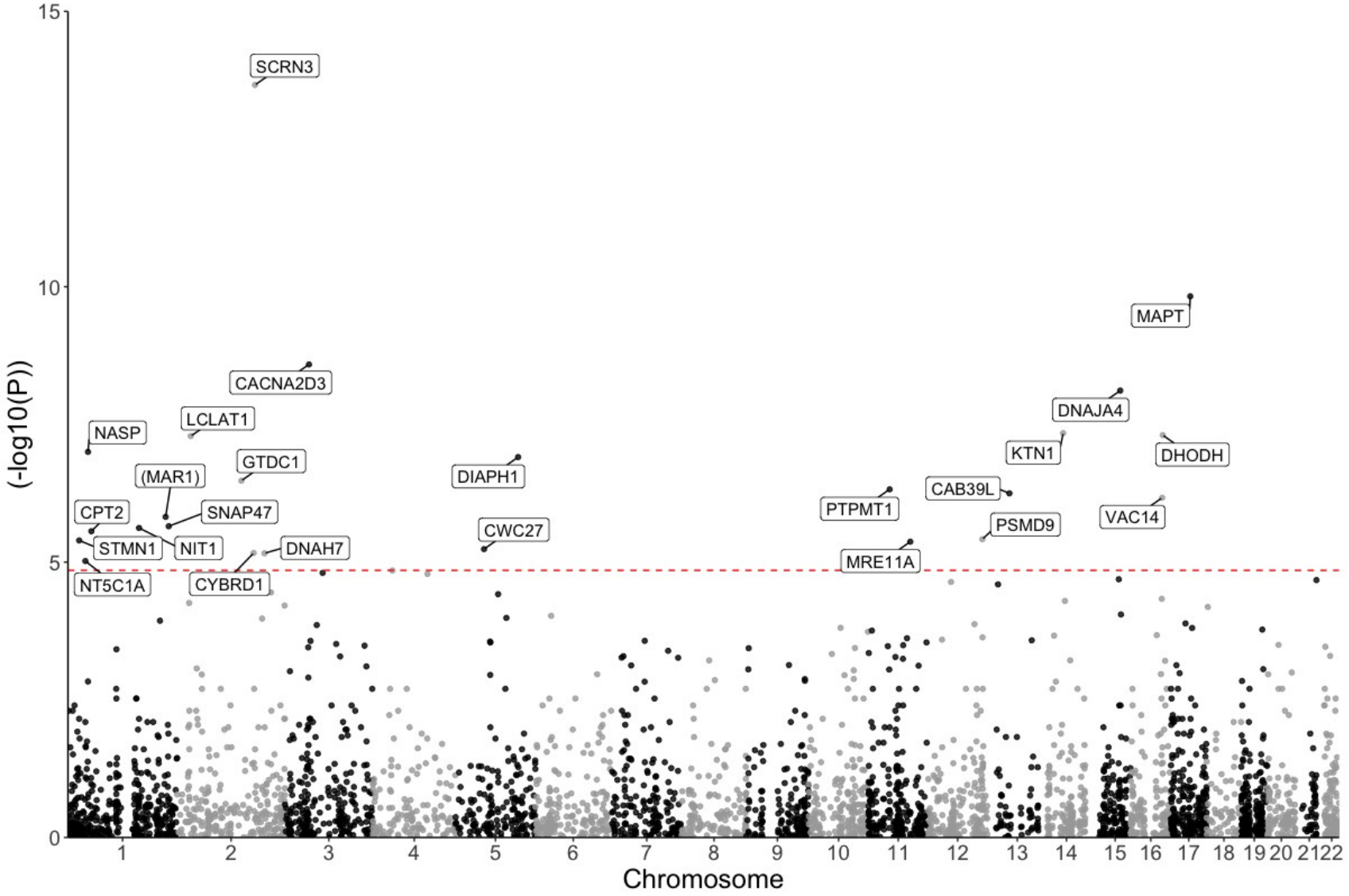


Figure AP3.2: Manhattan plot of the cortical proteomic cis-instrument Mendelian randomization results on binge drinking. *Labeled genes surpassed correction for multiple comparisons (P-value < 1.41×10^{-5}).*

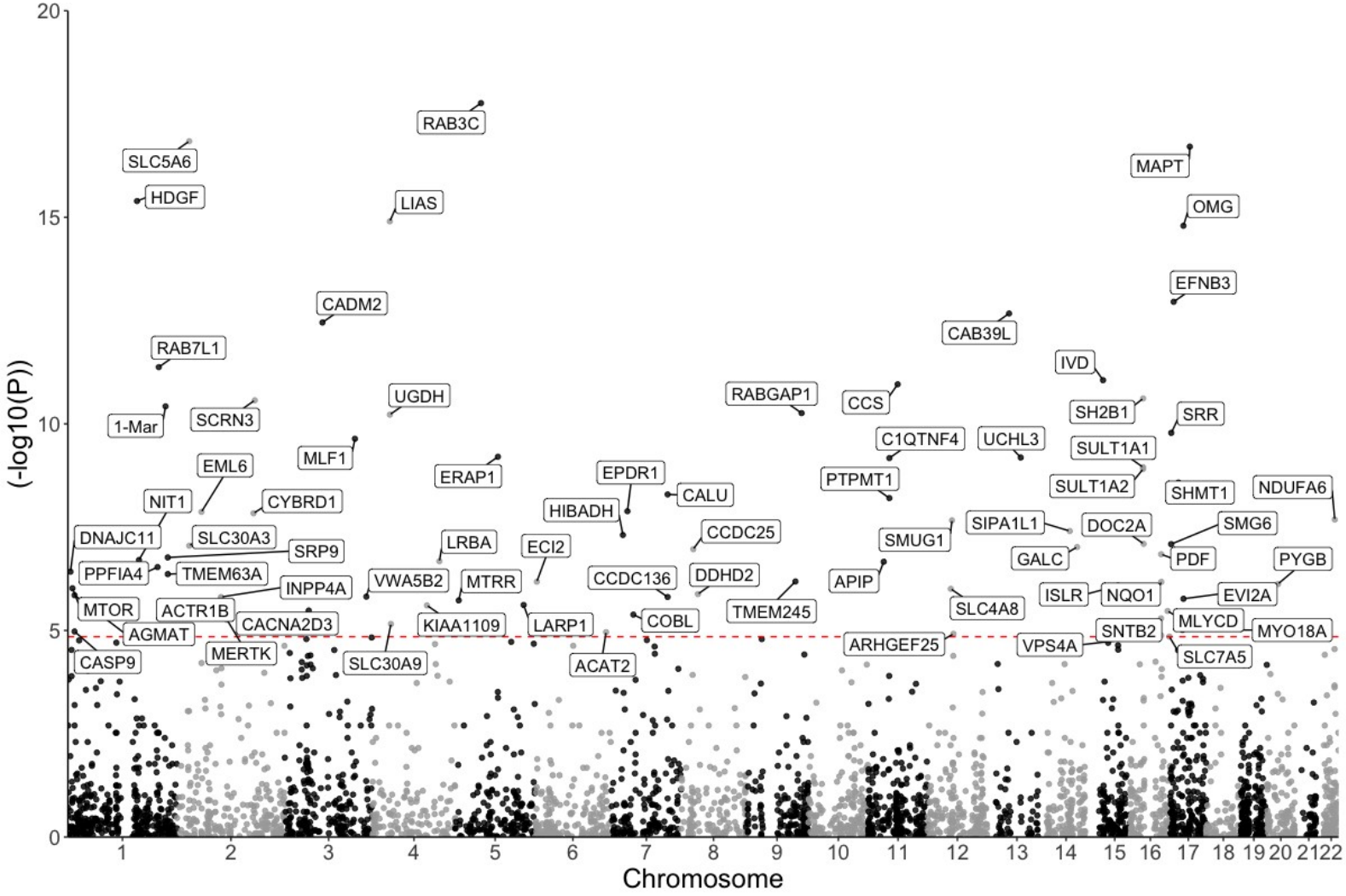


Figure AP3.3: Manhattan plot of the cortical proteomic cis-instrument Mendelian randomization results on self-reported alcoholic drinks per week. *Labeled genes surpassed correction for multiple comparisons (P-value < 1.41×10^{-5}).*

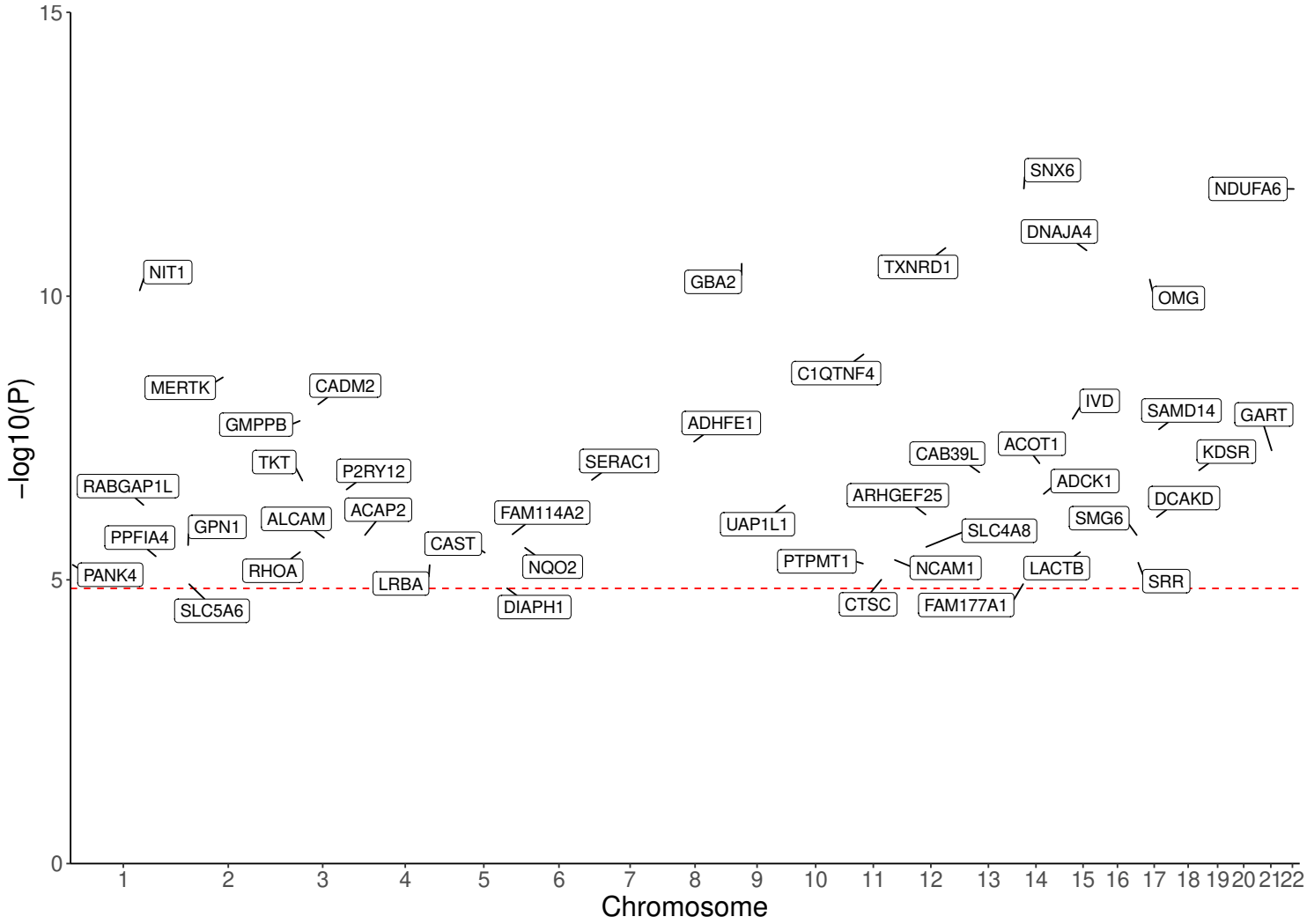


Figure AP3.4: Manhattan plot of the cortical proteomic cis-instrument Mendelian randomization results on problematic alcohol use. *Labeled genes surpassed correction for multiple comparisons (P-value < 1.41×10^{-5}).*

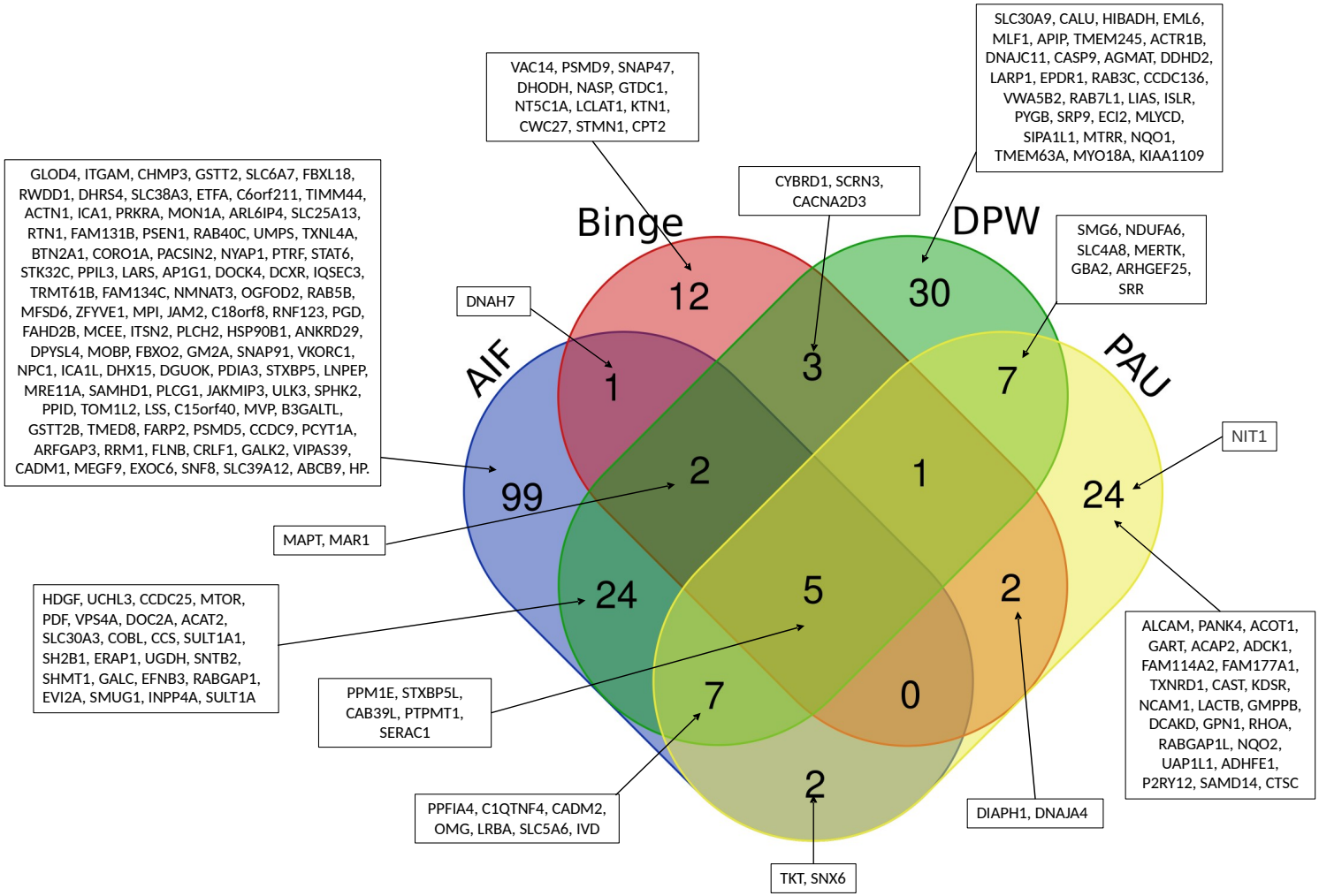


Figure AP3.5. Venn diagram outlining the overlap of cortical proteins surpassing correction for multiple comparisons for each alcohol consumption behavior. *These counts include all the unique proteins across the cortical (i.e., those that both colocalized and did not colocalize at PP.H4 >0.70, see the* ***Methods*** *section in the main manuscript for additional details).*
Abbreviations: *AIF, alcohol intake frequency; DPW: drinks per week; PAU: problematic alcohol use.*

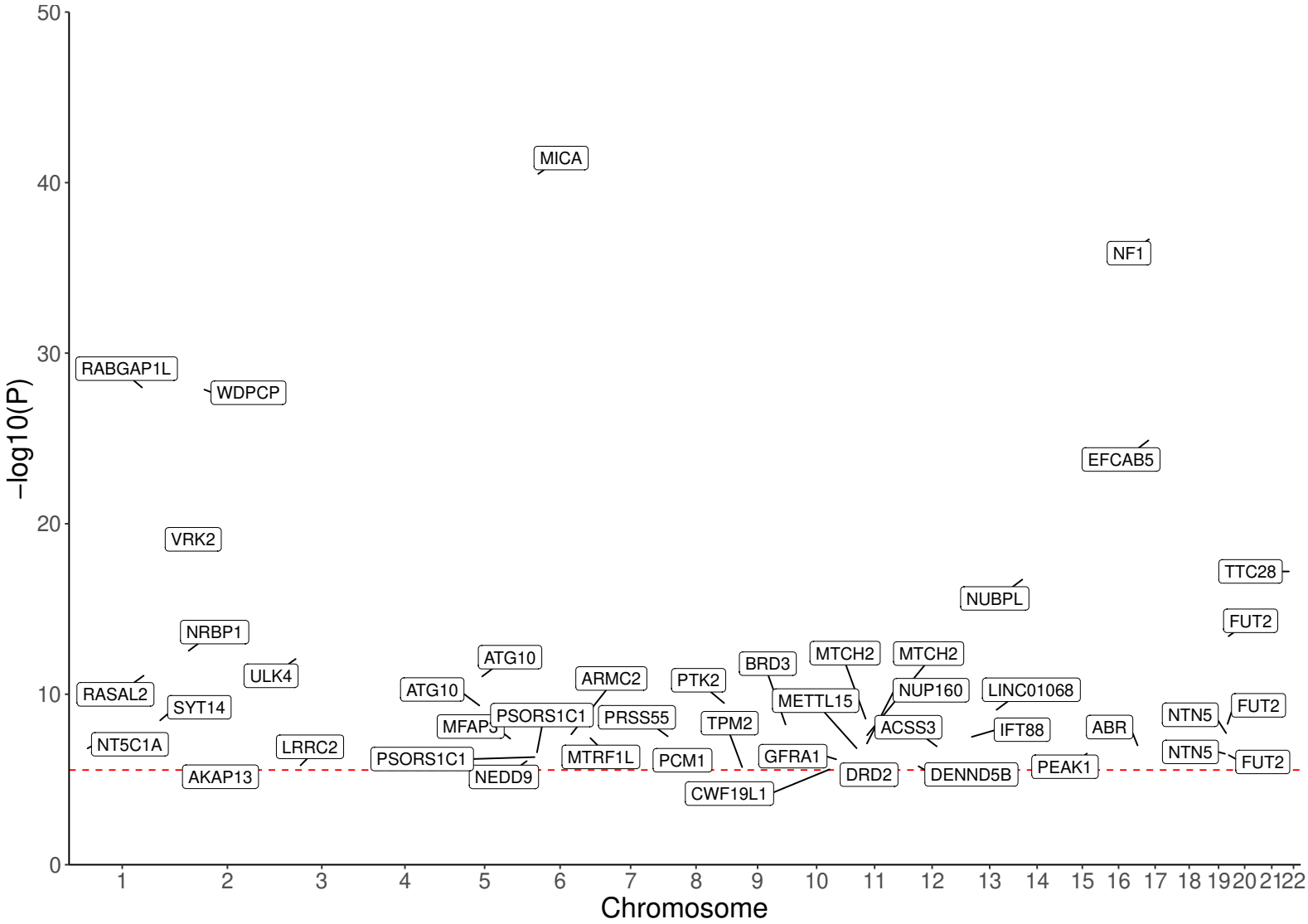


Figure AP3.6: Manhattan plot of the cell-type cis-instrument Mendelian randomization results on problematic alcohol use (PAU). *Labeled genes surpassed correction for multiple comparisons.*

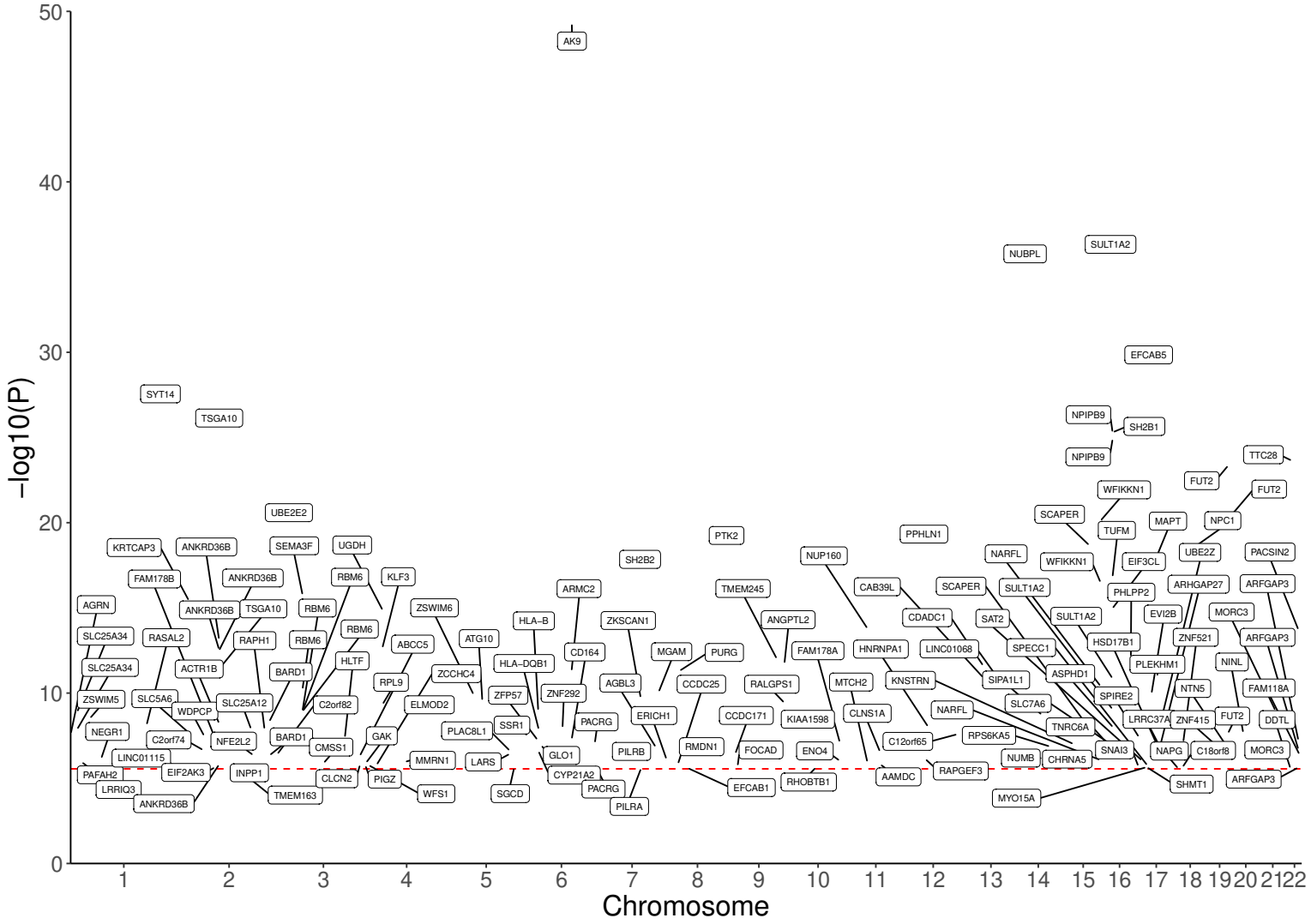


Figure AP3.7: Manhattan plot of the cell-type cis-instrument Mendelian randomization results on the alcohol intake frequency (AIF). *Labeled genes surpassed correction for multiple comparisons.*

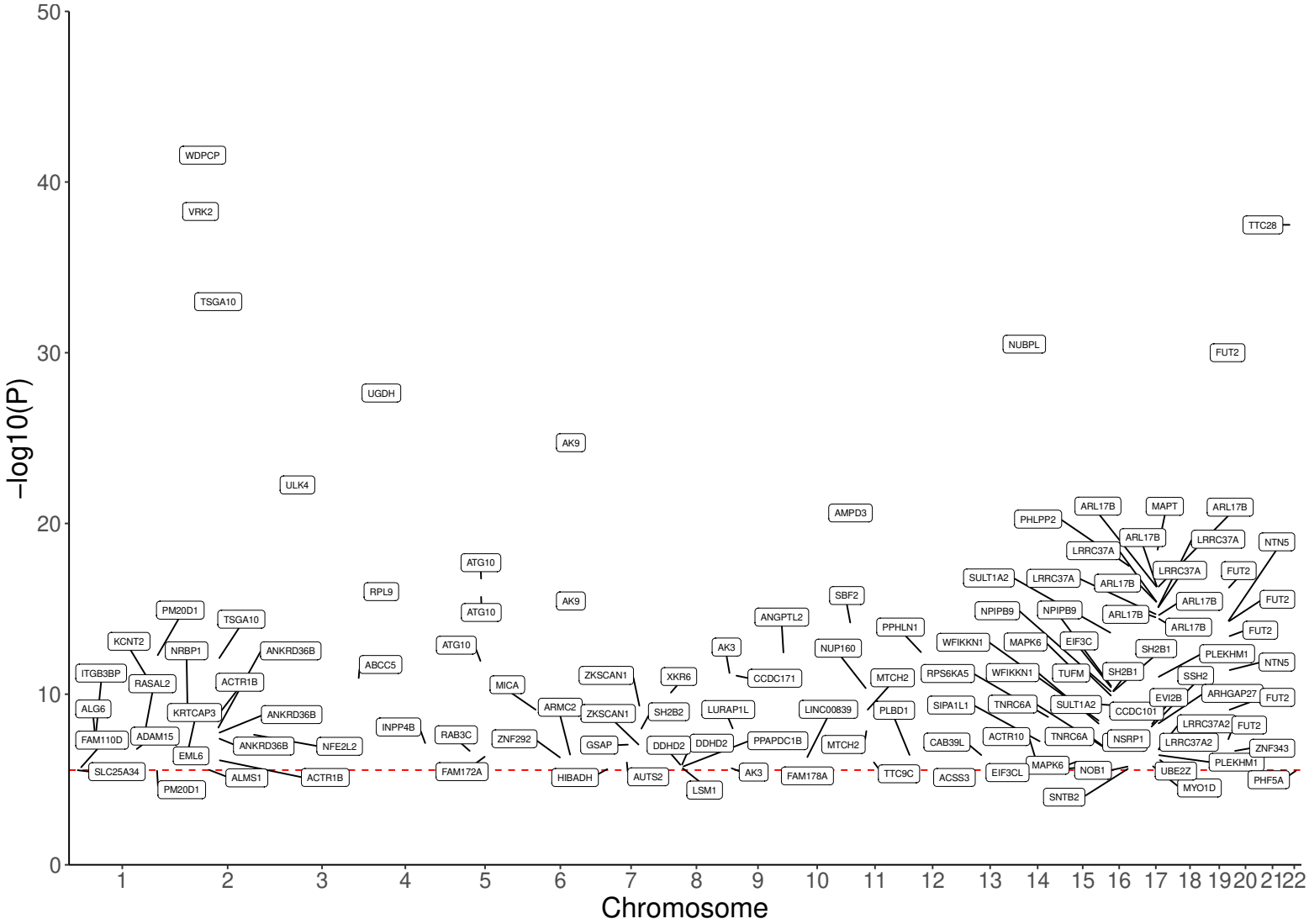


Figure AP3.8: Manhattan plot of the cell-type cis-instrument Mendelian randomization results on the weekly alcoholic drinks per week (DPW). *Labeled genes surpassed correction for multiple comparisons.*

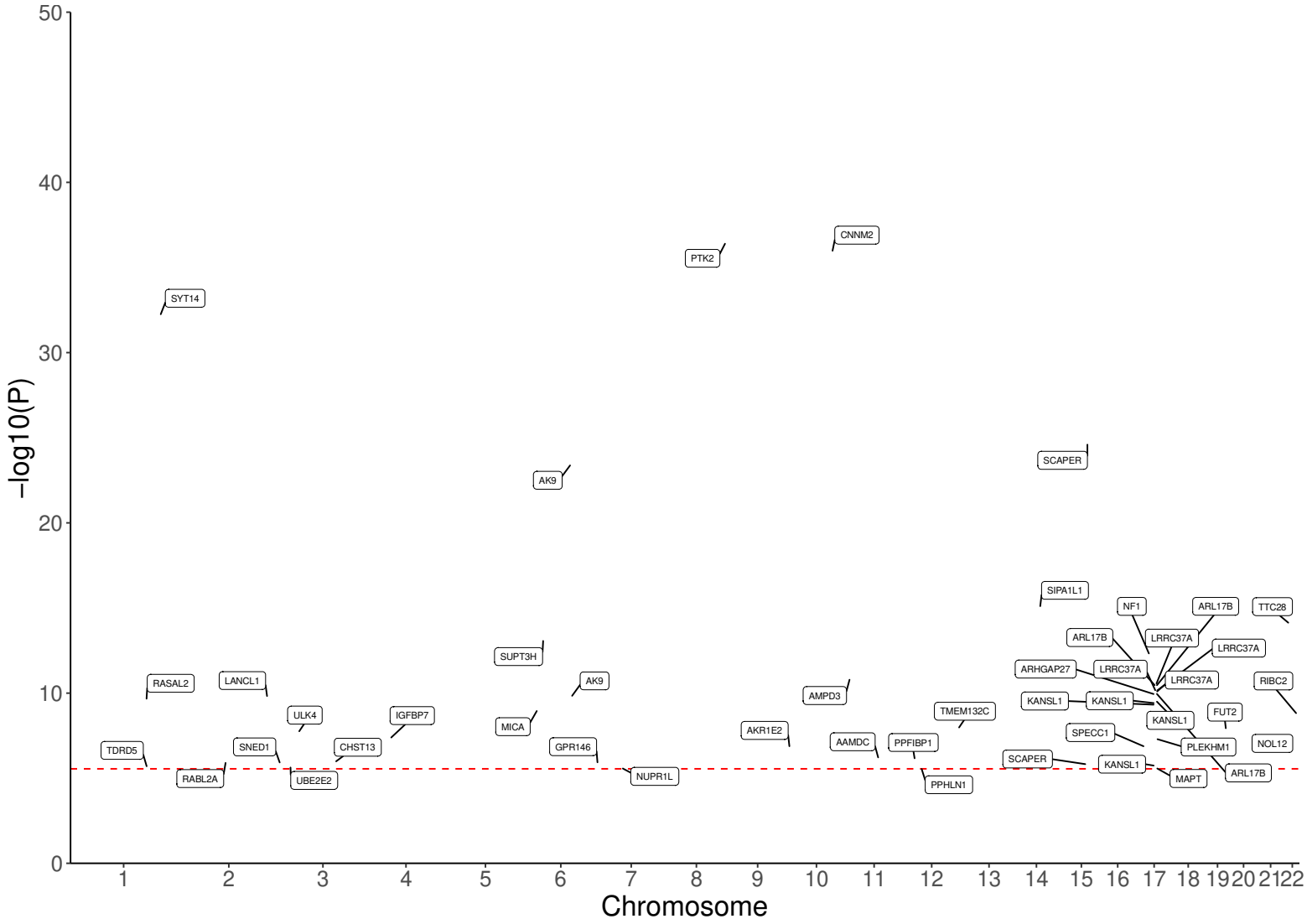


Figure AP3.9: Manhattan plot of the cell-type cis-instrument Mendelian randomization results on the binge drinking. *Labeled genes surpassed correction for multiple comparisons.*

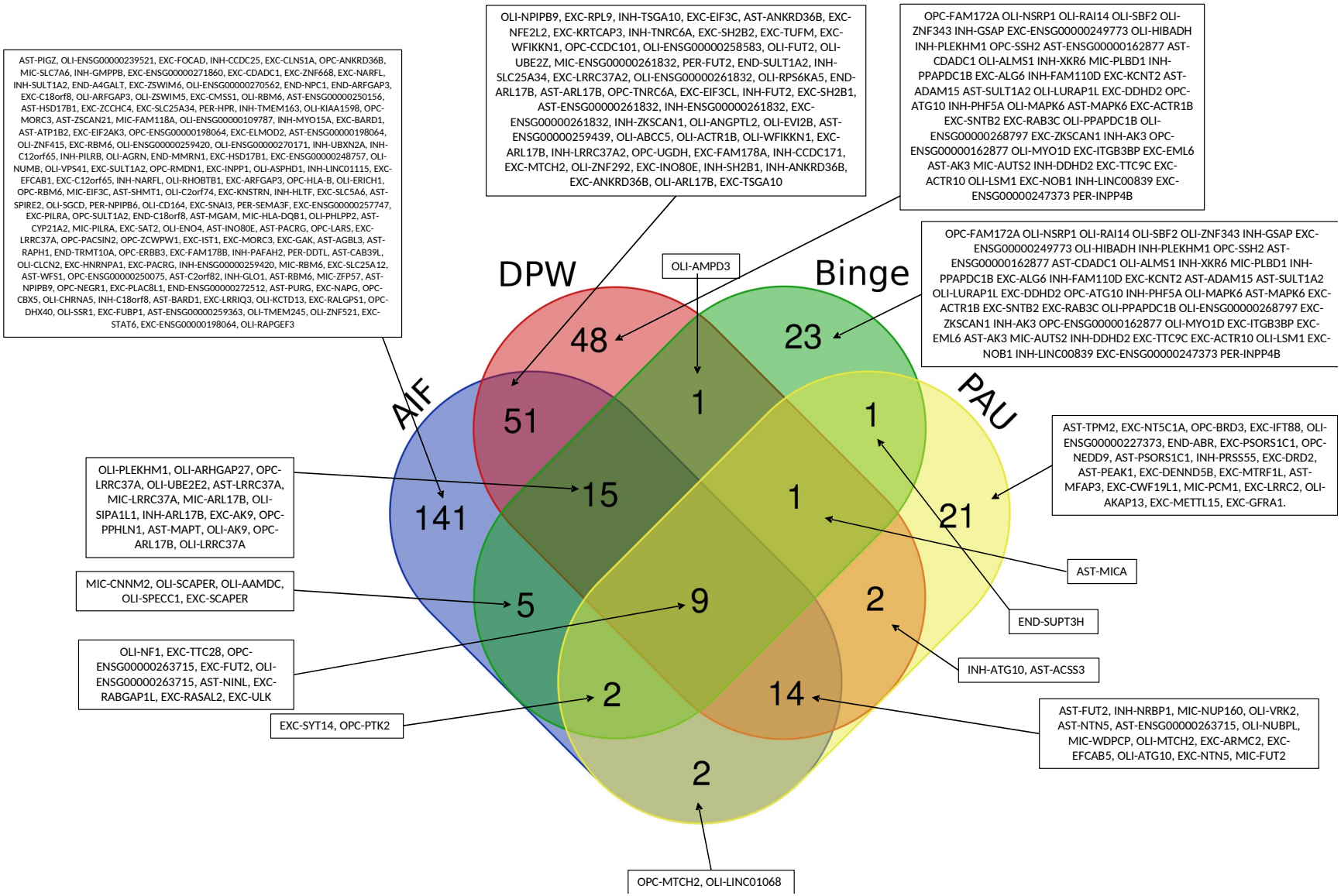

AST-PIGZ, OLI-ENSG00000239521, EXC-FOCAD, INH-CCDC25, EXC-CLNS1A, OPC-ANKRD36B, MIC-SLC7A6, INH-GMPPB, EXC-ENSG00000271860, EXC-CDADC1, EXC-ZNF668, EXC-NARFL, INH-SULT1A2, END-A4GALT, EXC-ZSWIM6, OLI-ENSG00000270562, END-NPC1, END-ARFGAP3, EXC-C18orf8, OLI-ARFGAP3, OLI-ZSWIM5, EXC-CMSS1, OLI-RBM6, AST-ENSG00000250156, AST-HSD17B1, EXC-ZCCHC4, EXC-SLC25A34, PER-HPR, INH-TMEM163, OLI-KIAA1598, OPC-MORC3, AST-ZSCAN21, MIC-FAM118A, OLI-ENSG00000109787, INH-MYO15A, EXC-BARD1, AST-ATP1B2, EXC-EIF2AK3, OPC-ENSG00000198064, EXC-ELMOD2, AST-ENSG00000198064, OLI-ZNF415, EXC-RBM6, OLI-ENSG00000259420, OLI-ENSG00000270171, INH-UBXN2A, INH-C12orf65, INH-PILRB, OLI-AGRN, END-MMRN1, EXC-HSD17B1, EXC-ENSG00000248757, OLI-NUMB, OLI-VPS41, EXC-SULT1A2, OPC-RMDN1, EXC-INPP1, OLI-ASPHD1, INH-LINC01115, EXC-EFCAB1, EXC-C12orf65, INH-NARFL, OLI-RHOBTB1, EXC-ARFGAP3, OPC-HLA-B, OLI-ERICH1, OPC-RBM6, MIC-EIF3C, AST-SHMT1, OLI-C2orf74, EXC-KNSTRN, INH-HLTF, EXC-SLC5A6, AST-SPIRE2, OLI-SGCD, PER-NPIPB6, OLI-CD164, EXC-SNAI3, PER-SEMA3F, EXC-ENSG00000257747, EXC-PILRA, OPC-SULT1A2, END-C18orf8, AST-MGAM, MIC-HLA-DQB1, OLI-PHLPP2, AST-CYP21A2, MIC-PILRA, EXC-SAT2, OLI-ENO4, AST-INO80E, AST-PACRG, OPC-LARS, EXC-LRRC37A, OPC-PACSIN2, OPC-ZCWPW1, EXC-IST1, EXC-MORC3, EXC-GAK, AST-AGBL3, AST-RAPH1, END-TRMT10A, OPC-ERBB3, EXC-FAM178B, INH-PAFAH2, PER-DDTL, AST-CAB39L, OLI-CLCN2, EXC-HNRNPA1, EXC-PACRG, INH-ENSG00000259420, MIC-RBM6, EXC-SLC25A12, AST-WFS1, OPC-ENSG00000250075, AST-C2orf82, INH-GLO1, AST-RBM6, MIC-ZFP57, AST-NPIPB9, OPC-NEGR1, EXC-PLAC8L1, END-ENSG00000272512, AST-PURG, EXC-NAPG, OPC-CBX5, OLI-CHRNA5, INH-C18orf8, AST-BARD1, EXC-LRRIQ3, OLI-KCTD13, EXC-RALGPS1, OPC-DHX40, OLI-SSR1, EXC-FUBP1, AST-ENSG00000259363, OLI-TMEM245, OLI-ZNF521, EXC-STAT6, EXC-ENSG00000198064, OLI-RAPGEF3
OLI-NPIPB9, EXC-RPL9, INH-TSGA10, EXC-EIF3C, AST-ANKRD36B, EXC-NFE2L2, EXC-KRTCAP3, INH-TNRC6A, EXC-SH2B2, EXC-TUFM, EXC-WFIKKN1, OPC-CCDC101, OLI-ENSG00000258583, OLI-FUT2, OLI-UBE2Z, MIC-ENSG00000261832, PER-FUT2, END-SULT1A2, INH-SLC25A34, EXC-LRRC37A2, OLI-ENSG00000261832, OLI-RPS6KA5, END-ARL17B, AST-ARL17B, OPC-TNRC6A, EXC-EIF3CL, INH-FUT2, EXC-SH2B1, AST-ENSG00000261832, INH-ENSG00000261832, EXC-ENSG00000261832, INH-ZKSCAN1, OLI-ANGPTL2, OLI-EVI2B, AST-ENSG00000259439, OLI-ABCC5, OLI-ACTR1B, OLI-WFIKKN1, EXC-ARL17B, INH-LRRC37A2, OPC-UGDH, EXC-FAM178A, INH-CCDC171, EXC-MTCH2, OLI-ZNF292, EXC-INO80E, INH-SH2B1, INH-ANKRD36B, EXC-ANKRD36B, OLI-ARL17B, EXC-TSGA10
OPC-FAM172A OLI-NSRP1 OLI-RAI14 OLI-SBF2 OLI-ZNF343 INH-GSAP EXC-ENSG00000249773 OLI-HIBADH INH-PLEKHM1 OPC-SSH2 AST-ENSG00000162877 AST-CDADC1 OLI-ALMS1 INH-XKR6 MIC-PLBD1 INH-PPAPDC1B EXC-ALG6 INH-FAM110D EXC-KCNT2 AST-ADAM15 AST-SULT1A2 OLI-LURAP1L EXC-DDHD2 OPC-ATG10 INH-PHF5A OLI-MAPK6 AST-MAPK6 EXC-ACTR1B EXC-SNTB2 EXC-RAB3C OLI-PPAPDC1B OLI-ENSG00000268797 EXC-ZKSCAN1 INH-AK3 OPC-ENSG00000162877 OLI-MYO1D EXC-ITGB3BP EXC-EML6 AST-AK3 MIC-AUTS2 INH-DDHD2 EXC-TTC9C EXC-ACTR10 OLI-LSM1 EXC-NOB1 INH-LINC00839 EXC-ENSG00000247373 PER-INPP4B
OPC-FAM172A OLI-NSRP1 OLI-RAI14 OLI-SBF2 OLI-ZNF343 INH-GSAP EXC-ENSG00000249773 OLI-HIBADH INH-PLEKHM1 OPC-SSH2 AST-ENSG00000162877 AST-CDADC1 OLI-ALMS1 INH-XKR6 MIC-PLBD1 INH-PPAPDC1B EXC-ALG6 INH-FAM110D EXC-KCNT2 AST-ADAM15 AST-SULT1A2 OLI-LURAP1L EXC-DDHD2 OPC-ATG10 INH-PHF5A OLI-MAPK6 AST-MAPK6 EXC-ACTR1B EXC-SNTB2 EXC-RAB3C OLI-PPAPDC1B OLI-ENSG00000268797 EXC-ZKSCAN1 INH-AK3 OPC-ENSG00000162877 OLI-MYO1D EXC-ITGB3BP EXC-EML6 AST-AK3 MIC-AUTS2 INH-DDHD2 EXC-TTC9C EXC-ACTR10 OLI-LSM1 EXC-NOB1 INH-LINC00839 EXC-ENSG00000247373 PER-INPP4B
OLI-AMPD3
DPW
Binge
AIF
PAU
48
23
51
1
1
1
141
15
1
21
5
9
2
2
14
2
AST-TPM2, EXC-NT5C1A, OPC-BRD3, EXC-IFT88, OLI-ENSG00000227373, END-ABR, EXC-PSORS1C1, OPC-NEDD9, AST-PSORS1C1, INH-PRSS55, EXC-DRD2, AST-PEAK1, EXC-DENND5B, EXC-MTRF1L, AST-MFAP3, EXC-CWF19L1, MIC-PCM1, EXC-LRRC2, OLI-AKAP13, EXC-METTL15, EXC-GFRA1.
OLI-PLEKHM1, OLI-ARHGAP27, OPC-LRRC37A, OLI-UBE2E2, AST-LRRC37A, MIC-LRRC37A, MIC-ARL17B, OLI-SIPA1L1, INH-ARL17B, EXC-AK9, OPC-PPHLN1, AST-MAPT, OLI-AK9, OPC-ARL17B, OLI-LRRC37A
AST-MICA
MIC-CNNM2, OLI-SCAPER, OLI-AAMDC, OLI-SPECC1, EXC-SCAPER
END-SUPT3H
OLI-NF1, EXC-TTC28, OPC-ENSG00000263715, EXC-FUT2, OLI-ENSG00000263715, AST-NINL, EXC-RABGAP1L, EXC-RASAL2, EXC-ULK
INH-ATG10, AST-ACSS3
EXC-SYT14, OPC-PTK2
AST-FUT2, INH-NRBP1, MIC-NUP160, OLI-VRK2, AST-NTN5, AST-ENSG00000263715, OLI-NUBPL, MIC-WDPCP, OLI-MTCH2, EXC-ARMC2, EXC-EFCAB5, OLI-ATG10, EXC-NTN5, MIC-FUT2
OPC-MTCH2, OLI-LINC01068

Figure AP3.10. Venn diagram outlining the overlap of cell-type genes surpassing correction for multiple comparisons for each alcohol consumption behavior. *These counts include all the unique proteins across the 8 cell-types (i.e., those that both colocalized and did not colocalize at PP.H4 >0.70, see the* ***Methods*** *section in the main manuscript for additional details).*
Abbreviations: *AIF, alcohol intake frequency; DPW: drinks per week; PAU: problematic alcohol use; EXC: excitatory; INH: inhibitory; AST: astrocytes; MIC: microglia; PERI: pericytes; ENDO: endothelial cells; OLIGO: oligodendrocytes; OPCs: oligodendrocyte-precursor cells.*

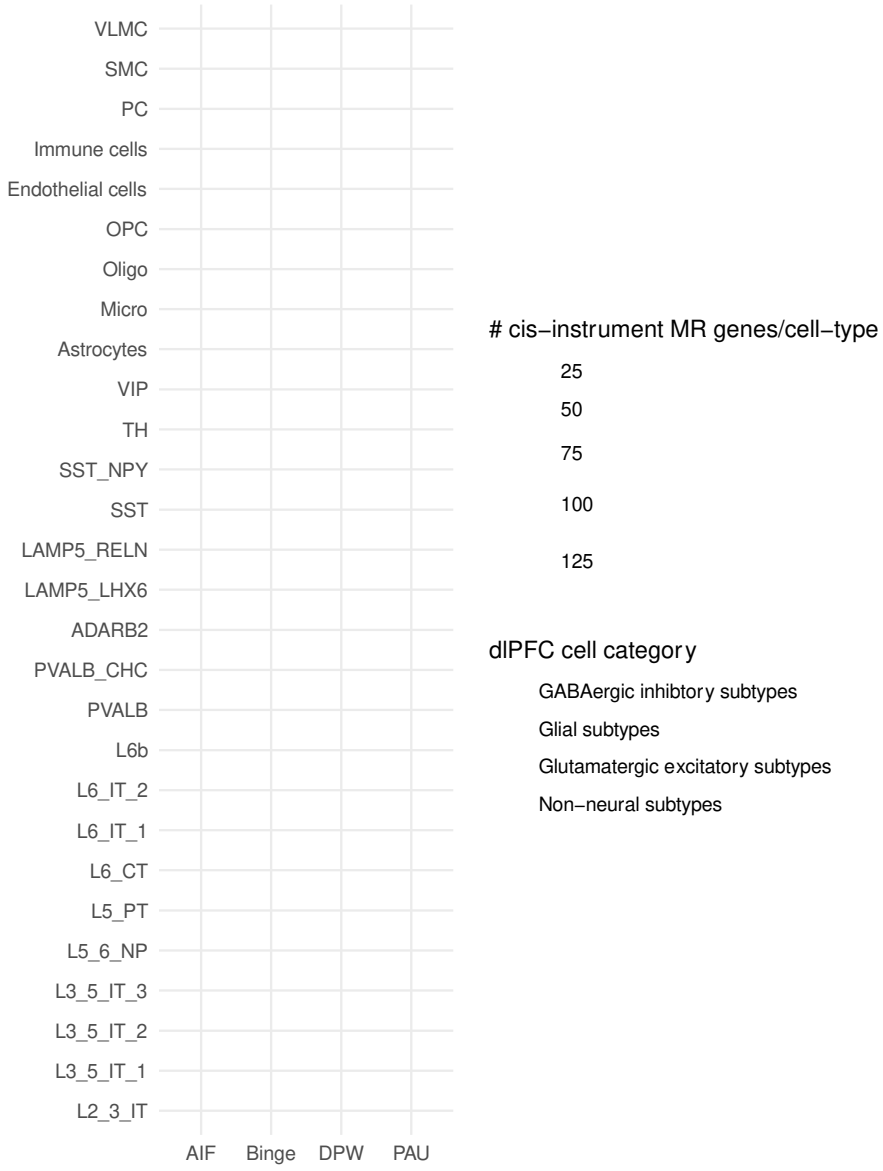


Figure AP3.11: scRNA-seq expression profiles of the cortical proteins surpassing correction for multiple comparisons. *The bubble plot depicts the expression profiles in 28 human cortical cell types of the cortical proteins associated with PAU and the 3 alcohol consumption behaviors. Only cortical proteins identified by the cis-instrument MR screen surpassing correction for multiple comparisons were analyzed (see **Methods**) and the bubble sizes depict the number of genes per cell type that surpassed correction for multiple comparisons for differential expression analysis of the human brain cell types. Cell-types are organized by their broad cortical cell-type ontology and nomenclature defined by the original study[254] (GEO accession code: GSE207334).*

Abbreviations: *scRNA: single-cell RNA; dlPFC: dorsolateral prefrontal cortex; PAU: problematic alcohol use; AIF: alcohol intake frequency; DPW: drinks per week; MR: Mendelian randomization; Micro: microglia; OPC: oligodendrocyte precursor cells; Oligo: oligodendrocytes.*

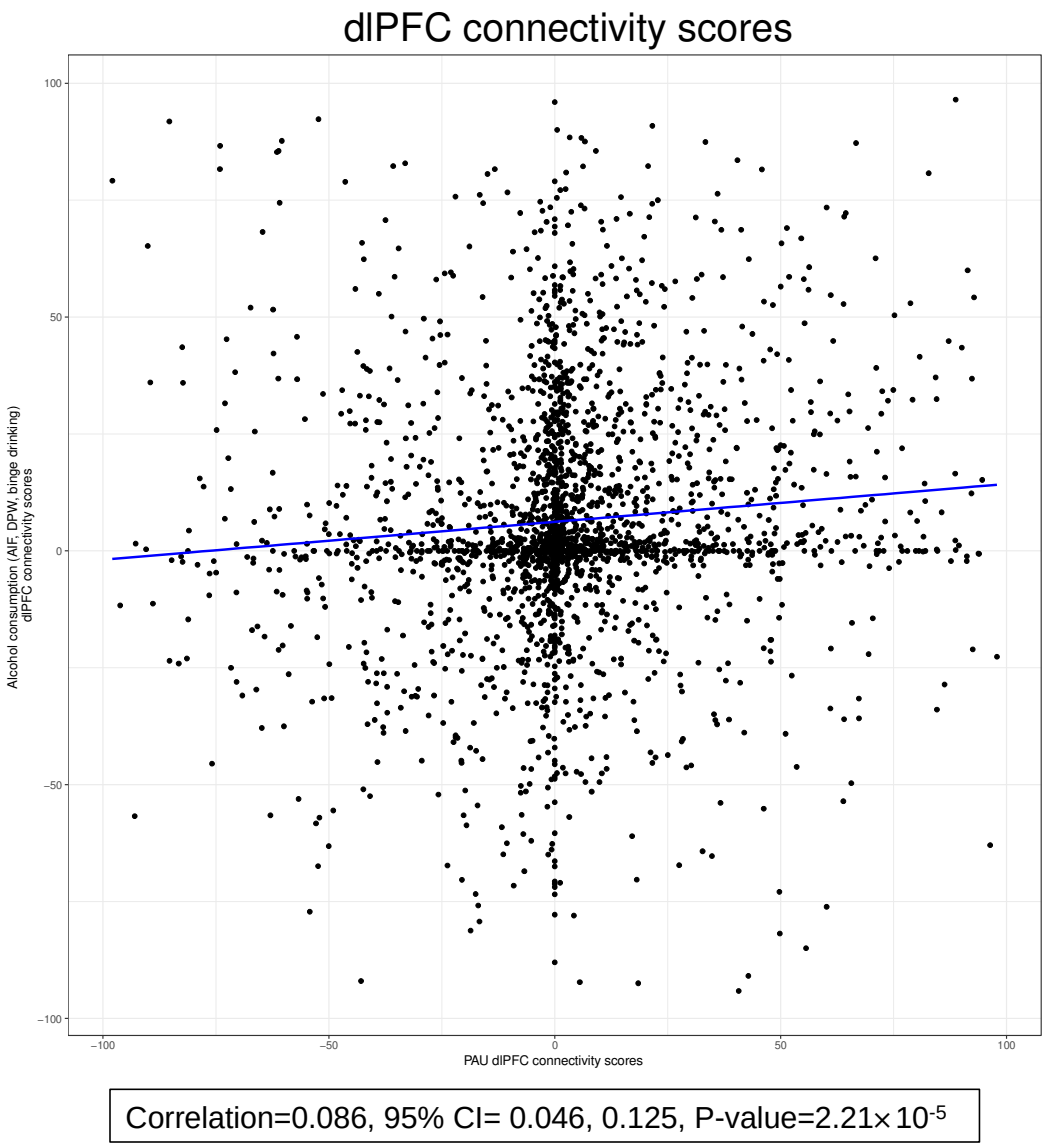


Figure AP3.12. Cortical proteome CMAP connectivity score correlation comparison. *Correlation (Pearson's) correlation of CMAP drug-set signature results (see Supplementary Methods) for the PAU cortical proteomic signature (all proteins surpassing correction for multiple comparisons) versus a combined alcohol consumption behavior cortical proteome signature (non-overlapping proteins for AIF, DPW, and binge drinking).*

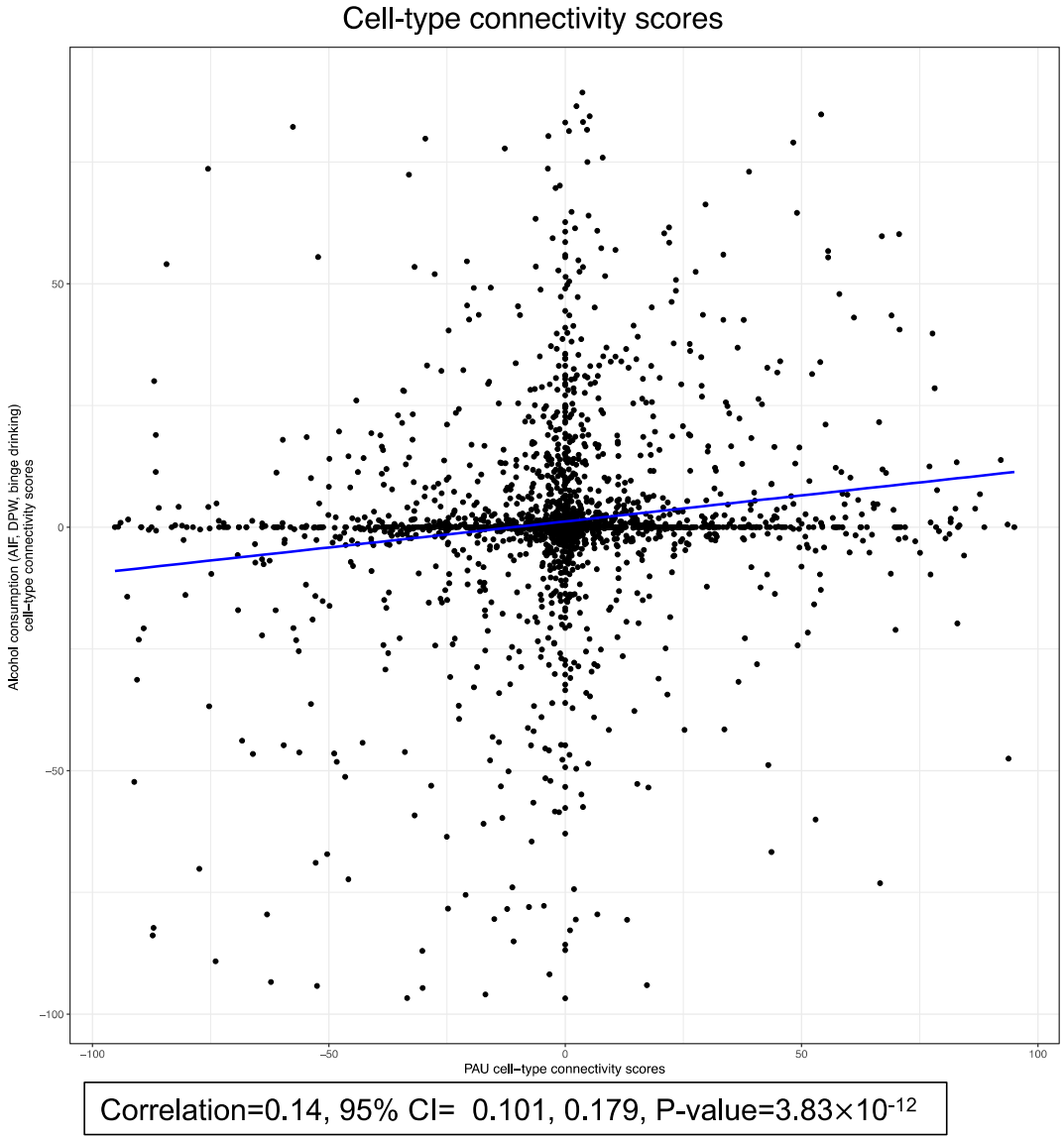


Figure AP3.13. Cell-type transcriptome CMAP connectivity score correlation comparison. *Correlation (Pearson's) correlation of CMAP drug-set signature results (see Supplementary Methods) for the PAU cell-type transcriptomic signature (all cell-type genes surpassing correction for multiple comparisons) versus a combined alcohol consumption behavior cell-type transcriptomic signature (non-overlapping proteins for AIF, DPW, and binge drinking).*

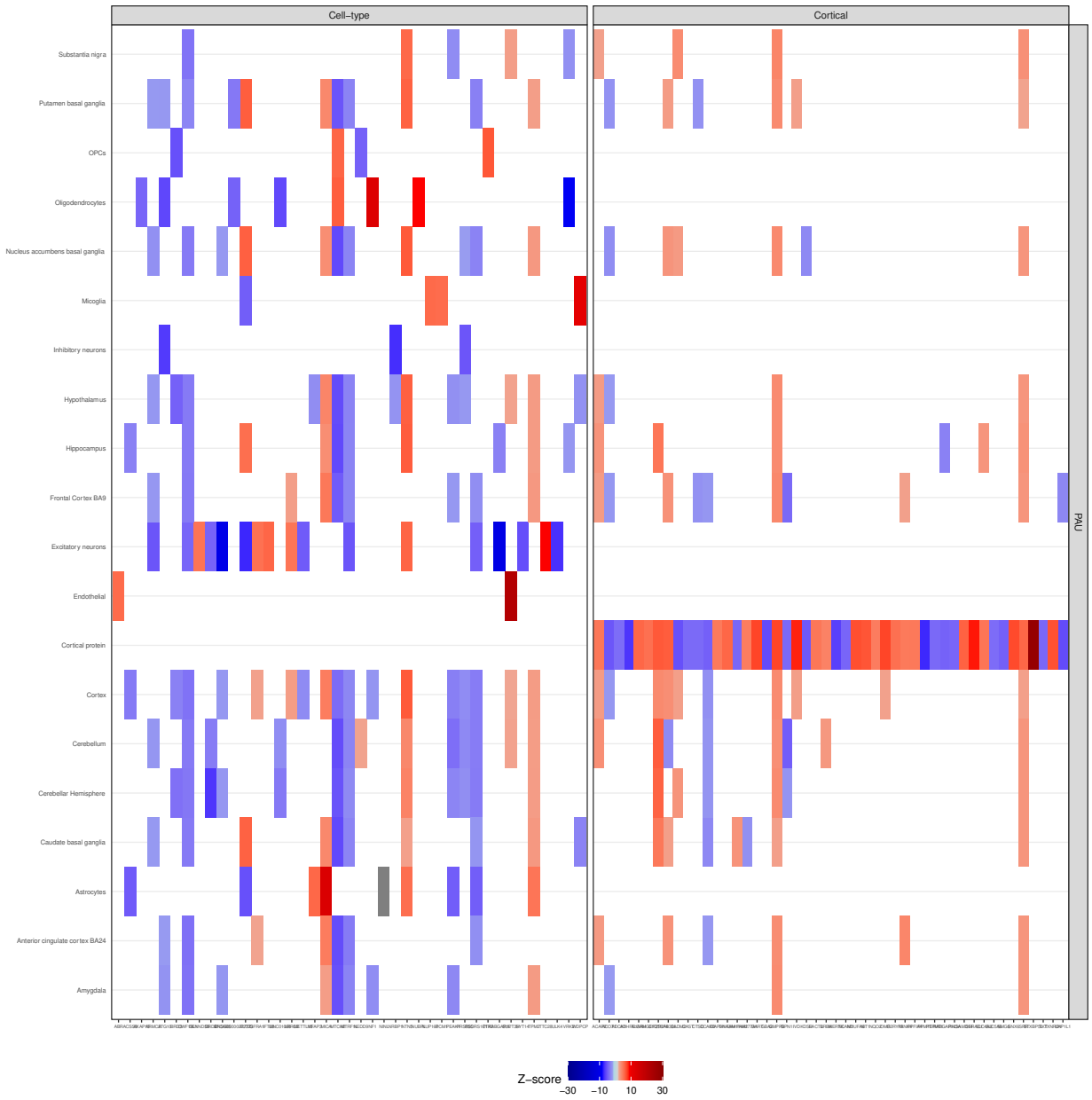


Figure AP3.14. Heatmap of Z-scores of cortical proteins and cell-type genes associated with problematic alcohol use (PAU). *Z scores from the cis-instrument MR analyses (beta/se) and the FUSION transcriptomic imputation analyses for FUSION analyses with the P-value < 0.05 from bulk RNAseq data in12 brain regions. The x-axis plots each of the cortical proteins and cell-type genes associated with PAU in the initial cis-MR screen and the y-axis are the regions.*

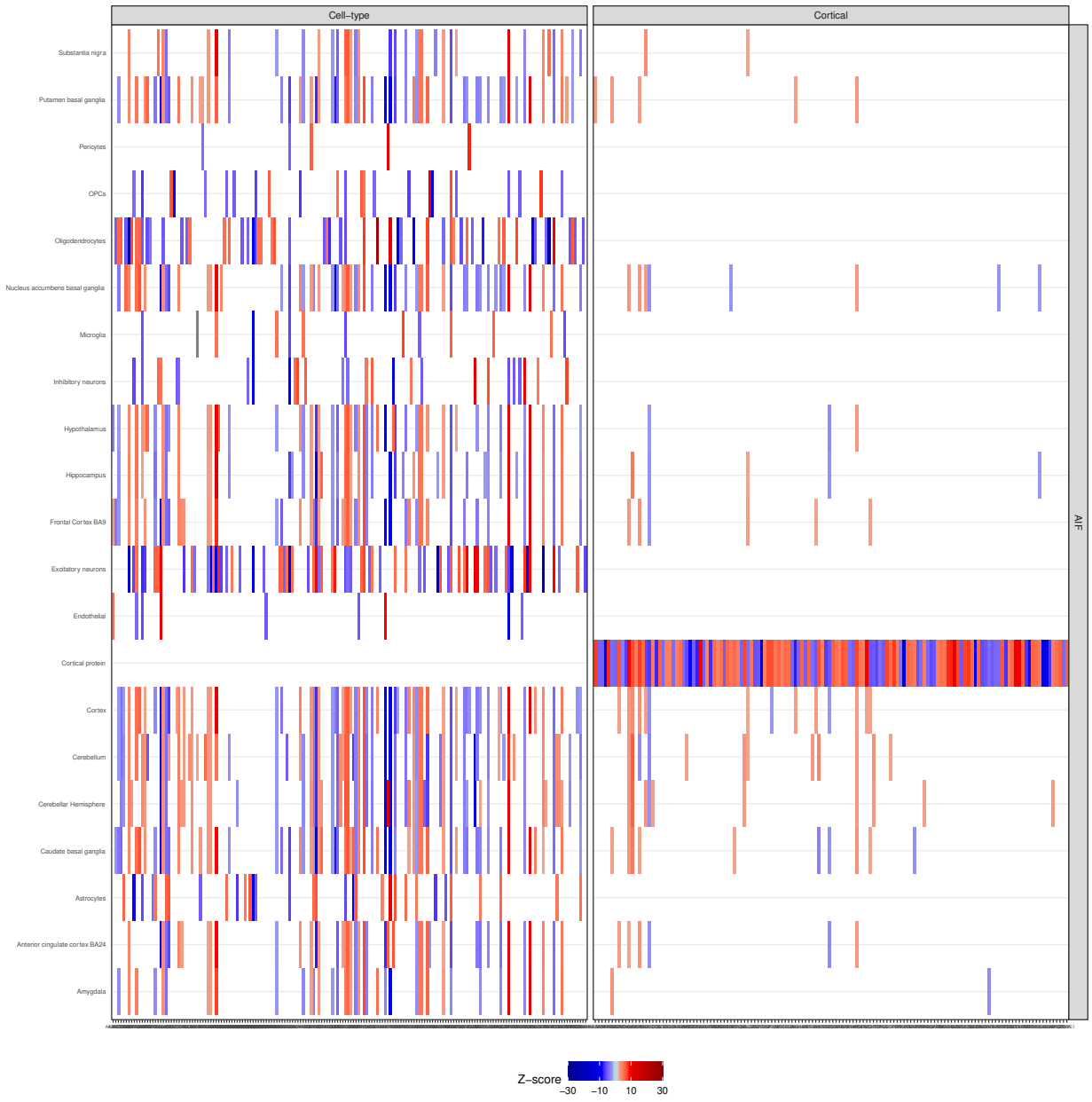


Figure AP3.15. Heatmap of Z-scores of cortical proteins and cell-type genes associated with alcohol intake frequency (AIF). *Z scores from the cis-instrument MR analyses (beta/se) and the FUSION transcriptomic imputation analyses for FUSION analyses with the P-value < 0.05 from bulk RNAseq data in12 brain regions. The x-axis plots each of the cortical proteins and cell-type genes associated with AIF in the initial cis-MR screen and the y-axis are the regions.*

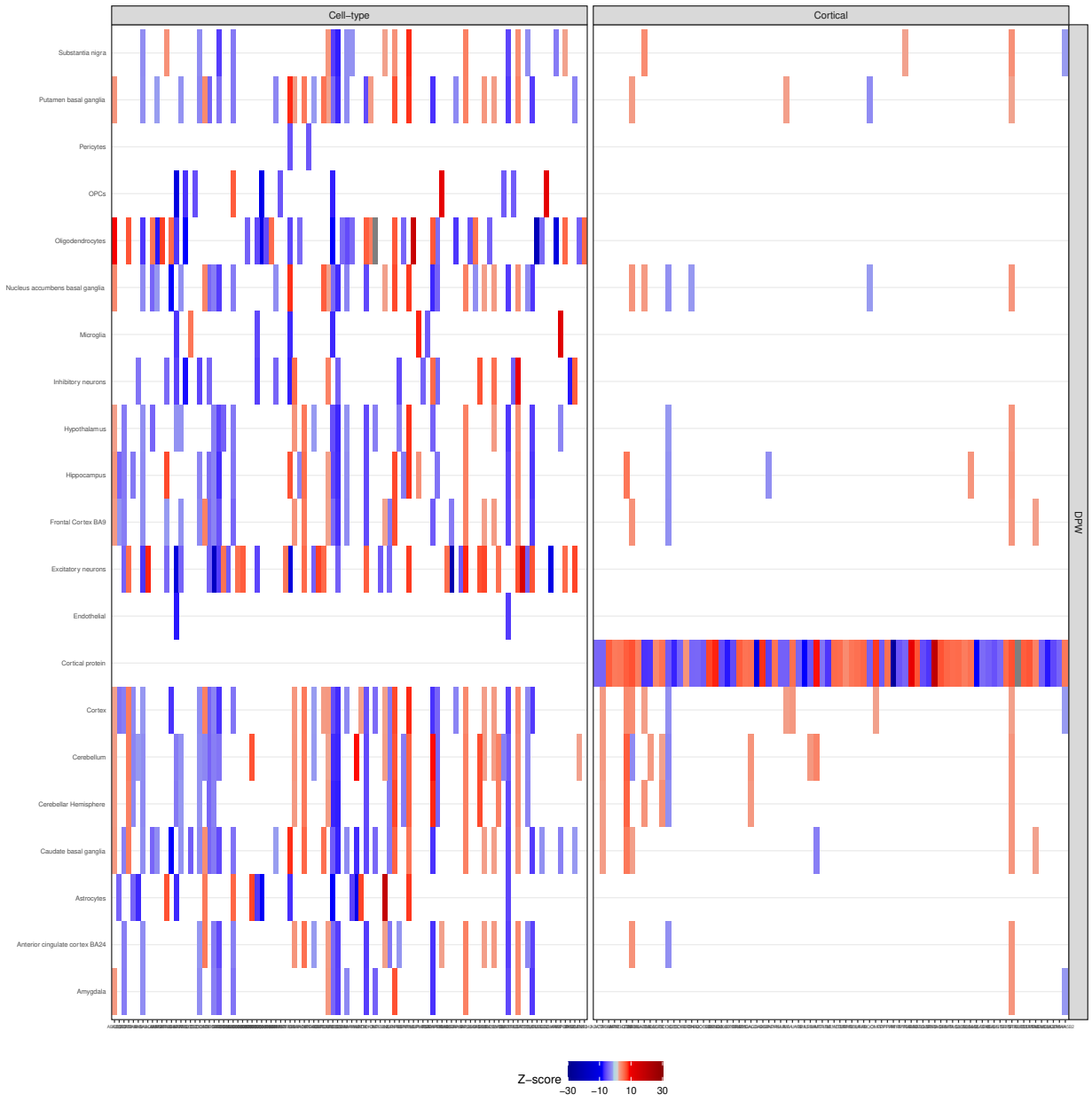


Figure AP3.16. Heatmap of Z-scores of cortical proteins and cell-type genes associated with alcoholic drinks per week (DPW). *Z scores from the cis-instrument MR analyses (beta/se) and the FUSION transcriptomic imputation analyses for FUSION analyses with the P-value < 0.05 from bulk RNAseq data in12 brain regions. The x-axis plots each of the cortical proteins and cell-type genes associated with DPW in the initial cis-MR screen and the y-axis are the regions.*

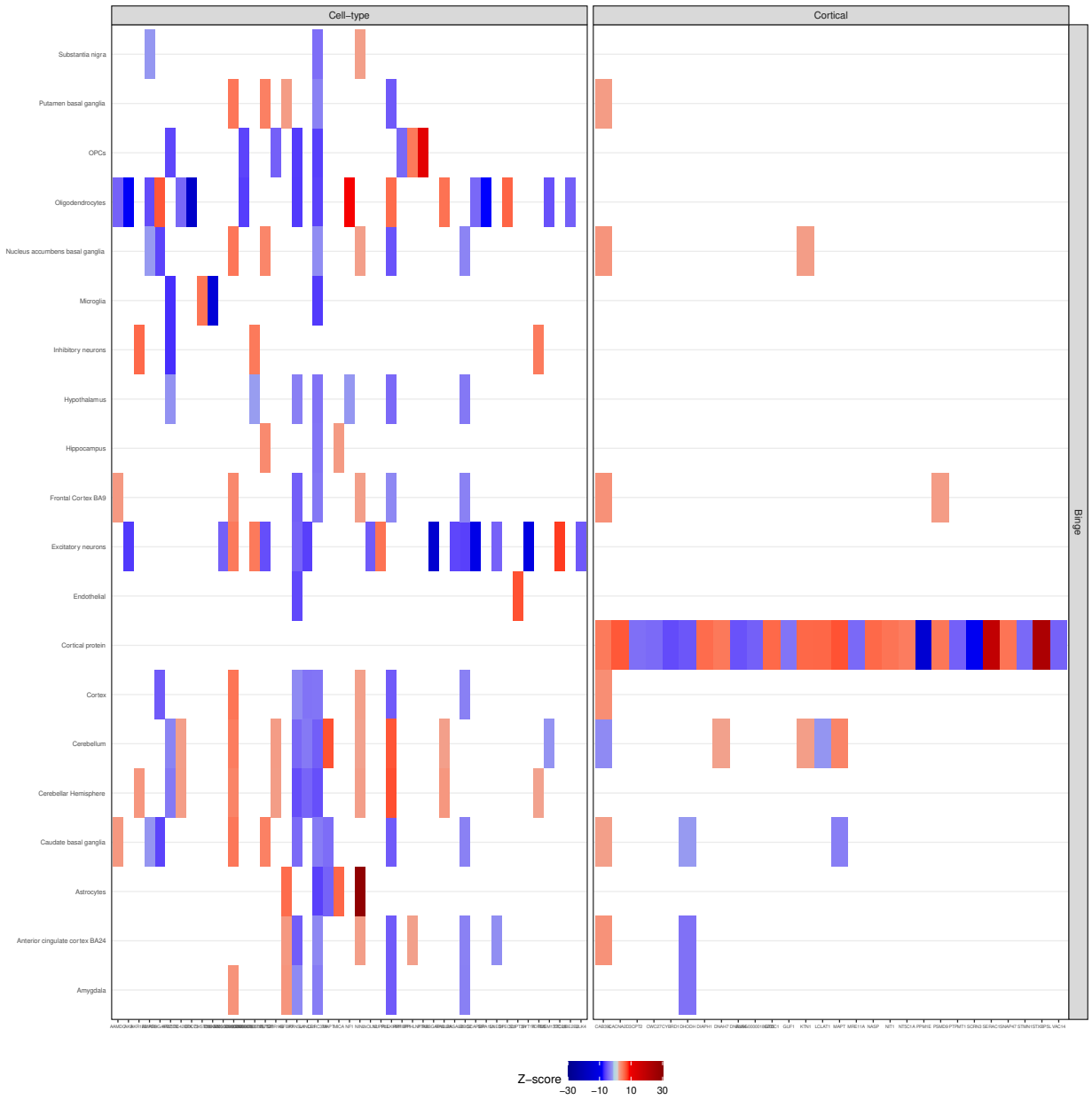


Figure AP3.17. Heatmap of Z-scores of cortical proteins and cell-type genes associated with problematic binge drinking. *Z scores from the cis-instrument MR analyses (beta/se) and the FUSION transcriptomic imputation analyses for FUSION analyses with the P-value < 0.05 from bulk RNAseq data in12 brain regions. The x-axis plots each of the cortical proteins and cell-type genes associated with binge drinking in the initial cis-MR screen and the y-axis are the regions.*

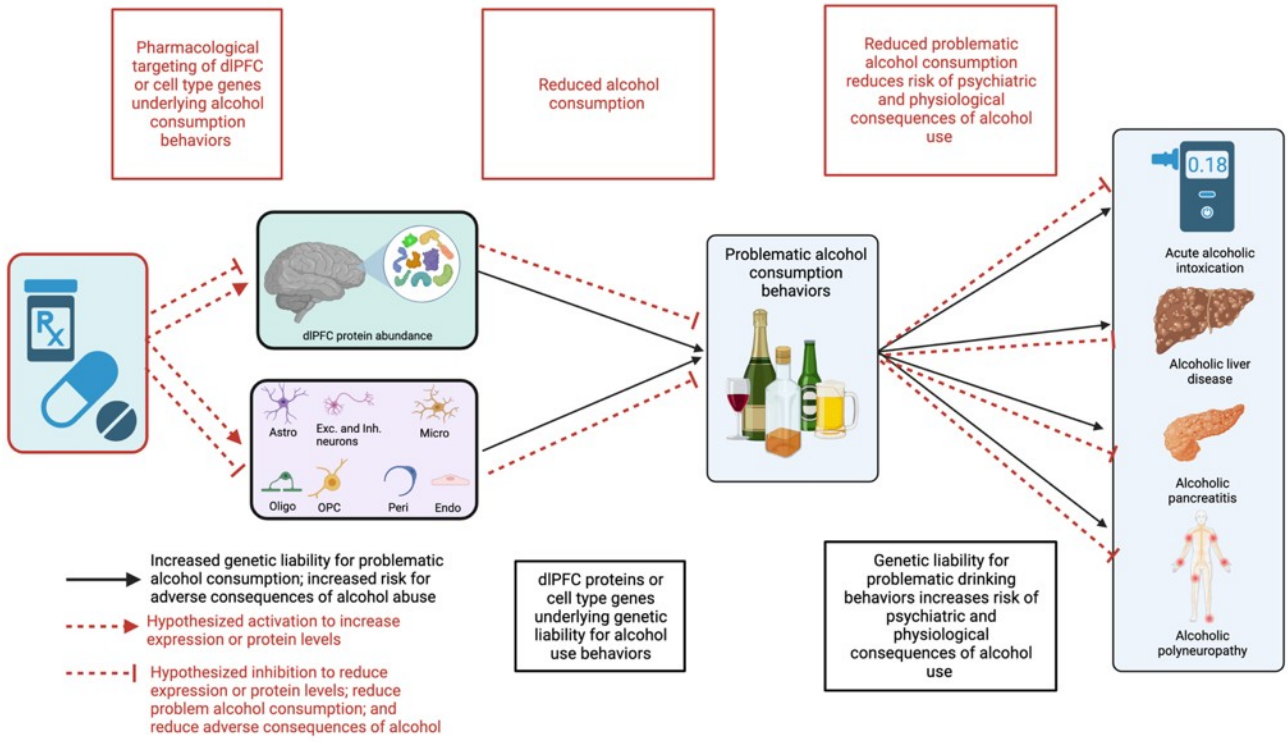


Figure AP3.18. Overview of rationale for analyzing cortical proteins and brain cell type genes on the psychiatric and physical consequences of problematic alcohol consumption. *Flow diagram depicting potential pathway through which cortical proteins and brain cell type genes related to the genetic liabilities for increased problematic alcohol consumption behaviors may impact the risk for the psychiatric and physical consequences of problematic alcohol consumption. Therapeutically targeting these cortical proteins and cell-type genes may result in less problematic alcohol consumption behaviors, which may, in turn, reduce the risk for the psychiatric and physical consequences. Therefore, in addition to providing an independent dataset for replication, the FinnGen electronic-health records-based information regarding psychiatric and physical consequences provided further potential prioritization of candidate dlPC protein and cell-type targets for future study.*

AP3.3. Aim 2 Supplementary Tables

Below are titles to each for the Aim 2 Supplementary Tables that are formatted as Excel files uploaded separately.

Table AP4.1. Data sources

Table AP4.2. ROSMAP cortical cis-pQTL instruments

Table AP4.3. Astrocyte cell-type cis-eQTL instruments

Table AP4.4. Pericyte cell-type cis-eQTL instruments

Table AP4.5. Endocytes cell-type cis-eQTL instruments

Table AP4.6. Excitatory cell-type cis-eQTL instruments

Table AP4.7. Microglia cell-type cis-eQTL instruments

Table AP4.8. Oligodendrocytes cell-type cis-eQTL instruments

Table AP4.9. Inhibitory cell-type cis-eQTL instruments

Table AP4.10. Oligodendrocyte progenitor cell-types cis-eQTL instruments

Table AP4.11. Results for drug-target MR screen of cortical proteome on Alcohol Intake Frequency (AIF)

Table AP4.12. Results for drug-target MR screen of cortical proteome on outcome alcoholic drinks consumed per week (DPW)

Table AP4.13. Results for drug-target MR screen of cortical proteome on outcome frequency of consuming ≥ 6 alcoholic drinks on occasion (Binge)

Table AP4.14. Results for drug-target MR screen of cortical proteome on Problematic Alcohol Use (PAU)

Table AP4.15. Cortical proteome top hits on alcohol consumption behaviors

Table AP4.16. Results for drug-target MR screen in 8 brain cell-types on alcohol intake frequency (AIF)

Table AP4.17. Results for drug-target MR screen in 8 brain cell-types on alcoholic Drinks Per Week (DPW)

Table AP4.18. Results for drug-target MR screen in 8 brain cell-types on frequency of consuming ≥ 6 alcoholic drinks (Binge)

Table AP4.19. Results for drug-target MR screen in 8 brain cell-types on problematic alcohol use (PAU)

Table AP4.20. Cell-type transcriptomic genes surpassing correction for multiple comparisons in alcohol consumption behaviors

Table AP4.21. Sensitivity MR analysis for cortical proteome on alcohol intake frequency (AIF)

Table AP4.22. Sensitivity MR analysis of cortical proteome on drinks per week (DPW)

Table AP4.23. Sensitivity MR analyses of cortical proteome on frequency of consuming ≥ 6 alcoholic drinks on occasion (Binge)

Table AP4.24. Sensitivity MR analysis of cortical proteome on problematic alcohol use (PAU)

Table AP4.25. Sensitivity MR analyses of cell-type transcriptome on alcohol intake frequency (AIF)

Table AP4.26. Sensitivity MR analysis of cell-type transcriptome on drinks per week (DPW)

Table AP4.27. Sensitivity MR analysis of cell-type transcriptome on frequency of consuming ≥ 6 alcoholic drinks on occasion (Binge)

Table AP4.28. Sensitivity MR analysis of cell-type transcriptome on problematic alcohol use (PAU)

Table AP4.29. Gene-set enrichment results of cortical proteins for each alcohol-related phenotype (PAU, AIF, DPW, and Binge drinking)

Table AP4.30. Gene-set enrichment results of brain cell-type genes for each alcohol-related phenotype (PAU, AIF, DPW, and Binge drinking)

Table AP4.31. Human brain single-cell differential expression results for cortical proteins associated with alcohol behaviors in the cortical proteomic cis-instrument MR screen

Table AP4.32. Human brain single-cell differential expression results for genes associated with alcohol behaviors in the cell type transcriptome cis-instrument MR screen

Table AP4.33. DGIdb interaction scores > 0.5 for cortical proteins linked with PAU and alcohol consumption behaviors

Table AP4.34. DGIdb interaction scores > 0.5 for cell-type genes linked with PAU and alcohol consumption behaviors

Table AP4.35. Druggability results for cortical proteins and cell-type genes

Table AP4.36. CMAP drug repositioning analyses of cis-MR signatures for PAU and other alcohol consumption behaviors

Table AP4.37. Gene mapping and novelty of top genes in cortical proteome compared to GWAS lead SNPs

Table AP4.38. Gene mapping and novelty of top genes in cell-type transcriptome drug-target

Mendelian randomization analysis compared to GWAS lead SNPs

Table AP4.39. Novel cortical proteins and cell-type genes from the drug-target Mendelian randomization screens

Table AP4.40. Comparison of cortical proteins from current analysis of alcohol-related outcomes with previous PWAS/TWAS studies for non-alcohol neuropsychiatric outcomes

Table AP4.41. FUSION transcriptomic imputation results in 12 brain tissues for genes related to the alcohol behaviors identified in the drug-target Mendelian randomization analysis with the cortical proteome

Table AP4.42. FUSION transcriptomic imputation results in 12 brain tissues for genes related to the alcohol behaviors identified in the drug-target Mendelian randomization analysis in the cell-type transcriptome

Table AP4.43. Colocalization of top genes identified in drug-target Mendelian randomization of the cortical proteome

Table AP4.44. Colocalization of cell-type genes associated with alcohol behaviors

Table AP4.45. Comparison of the alcohol-related cortical proteins and unique cell-type genes identified in the drug-target Mendelian randomization analyses also demonstrating evidence of colocalization

Table AP4.46. Correlated MR results for cortical proteome colocalized top hits on ENIGMA3 subcortical gray matter volumes

Table AP4.47. Correlated MR results of cell-type gene colocalized top hits on ENIGMA3 subcortical gray matter volumes

Table AP4.48. Correlated MR results for cortical proteome colocalized top hits on ENIGMA3 mean full brain surface area and thickness

Table AP4.49. Correlated MR results for cell-type gene colocalized top hits on ENIGMA3 surface area and average thickness of the whole cortex

Table AP4.50. Correlated MR results for cortical proteome colocalized top hits on mean surface area and thickness of 34 regions of the cortex

Table AP4.51. Correlated MR results cell-type gene colocalized top hits on ENIGMA3 mean surface area and thickness of 34 regions of the cortex

Table AP4.52. Correlated MR results for cortical proteome colocalized top hits on UKB diffuser

tensor imaging measurements

Table AP4.53. Correlated MR results for cell-type gene colocalized top hits con UKB diffuser tensor imaging measurements

Table AP4.54. Correlated MR results for cortical proteome colocalized top hits on FinnGen Freeze 8 and 10 alcohol-related outcomes

Table AP4.55. Inverse variance weighted (IVW) or Wald ratio estimates for the cortical proteome colocalized top hits surpassing correction for multiple comparisons in the FinnGen alcohol-related outcomes

Table AP4.56. Correlated MR results for cell-type gene colocalized top hits on FinnGen Freeze 8 and 9 alcohol-related outcomes

Table AP4.57. Inverse variance weighted (IVW) or Wald ratio estimates for the cell-type genes surpassing correction for multiple comparisons in the FinnGen alcohol-related outcomes

Table AP4.58. Correlated MR results for cortical proteomic colocalized top hits on behavioral and psychiatric/neurological outcomes

Table AP4.59. Correlated Mendelian randomization of cell-type gene colocalized top hits on behavioral and psychiatric/neurological outcomes

APPENDIX 4 (AP4): SUPPLEMENTARY MATERIALS FOR AIM 3

AP4.1. Supplementary Methods for Aim 3

AP4.1.1. Genomic Structural Equation Modeling (GenomicSEM)

Genomic structural equation modeling (GenomicSEM) is an innovative genetics-based approach that utilizes summary statistics from genome-wide association studies of related, single traits to explore their shared genetic structures.[16] Essentially, it applies structural equation modeling (SEM) – a method that investigates variance and covariance among related variables – in a two-

stage process specific to GWAS data.[16] This technique has proven resilient against variations in sample sizes and overlaps of the input GWAS datasets, making it highly adaptable and effective in enhancing statistical power by leveraging increased effective sample sizes.[16]

We used GenomicSEM to first construct a unified genetic framework for cardiometabolic disease risk by concurrently analyzing single trait GWAS of NAFLD, T2D, and CAD. Subsequently, we generated a GWAS that pinpointed individual SNP associations with a generalized latent factor underlying these observed traits that we termed the "cardiometabolic factor" or "*CM-Factor*". In addition to providing a robust approach to modeling the multivariate architecture linking observed traits, GenomicSEM also has integrated methods capable of further refining the resultant multivariate GWAS signature to remove SNPs that are unlikely to impact the observed data through the latent modeled multivariate GWAS factors, which would indicate that these SNPs and loci are likely not to link NAFLD, T2D, or CAD, through the *CM-Factor,* which facilitates improved identification of the appropriate loci for annotation and follow-up analyses.

GenomicSEM analysis is performed in two stages. In Stage 1, we estimated the empirical genetic covariance matrix (and corresponding sampling covariance matrix) for the univariate GWASs. We prepared the NAFLD, T2D, and CAD GWAS summary statistics for stage 1 using a multivariate extension of cross-trait linkage disequilibrium score regression (LDSC)[16,275] to generate the empirical genetic covariance matrix between the three traits, as input for the SEM common factor model.[16] In Stage 2, we specified an SEM maximizing fit between hypothesized covariance matrix and the empirical covariance matrix calculated in Stage 1.[16] Because we aimed to identify a genetic signature underlying the three cardiometabolic diseases, we evaluated a one-factor model estimating the genetic associations between the underlying *CM-Factor* and the three cardiometabolic diseases. We tested the model fit using the standardized root mean square residual (SRMR), model χ^2, Akaike Information Criterion (AIC), and Comparative Fit Index (CFI).[278] After we identified the *CM-Factor* SEM, we proceeded to applying the model to perform the multivariate GWAS identifying SNP associations of the shared covariance across these cardiometabolic diseases.

In the following sections, we elaborate on the key principles and specifics of SEM within the GenomicSEM context, adhering to and adopting the mathematical notation and methods presented in the foundational GenomicSEM paper by Grotzinger et al. 2019.[16]

AP4.1.1a. Structured covariance models and factor analysis.

GenomicSEM provides the flexibility to select the most appropriate model fit for SEM.[509] SEM operates through two distinct equations: the measurement model and the structural model.The measurement model, utilized in factor analysis (FA), delineates the variance and covariance among observed variables (indicators) in relation to latent factors. Essentially, FA models the shared variance across these observed variables.[16] In this context, the measurement model is expressed as $y=\Lambda\eta+ E$. Here, k observed phenotypes (indicators) are represented as a linear combination of m latent variables. In this formula, y is a vector of k observed variables, E is a vector of residuals for these observed variables, η represents the latent variables in an $m \times 1$ vector, and Λ is a matrix of regression coefficients that link latent factors to the observed variables.[16] In our application of GenomicSEM, we modeled the input univariate GWASs for

NAFLD, T2D, and CAD as a function of a singular latent variable modeling the *CM-Factor* demonstrating the utility of GenomicSEM in analyzing complex genomic data.[16]

In SEM, structural models are used to define the relationships between latent factors through directed regression coefficients, as represented by the equation $\eta=B\eta+\zeta$.[16] Within these models, B is an m × m matrix that includes the regression coefficients detailing how these latent variables interact with each other. Meanwhile, ζ is an m × 1 vector that represents the residuals of the latent factors. The covariance matrix of the observed variables implied by the model is denoted as $\sum(\theta)= \Lambda(I-B)^{(-1)} \Psi[(I-B)^{(-1)}]' \Lambda' + \Theta$, where I stands for a k × k identity matrix. Comprehensive SEM models utilize a series of linear equations to approximate the empirical matrix, connecting the observed variables to the latent variables and delineating the relationships among the latent variables themselves.[16]

GenomicSEM utilizes the frameworks of measurement and structural models to analyze the genetic covariances among selected GWAS phenotypes.[16] The first stage of this process involves estimating the empirical genetic covariance matrix using the LDSC package, which is adapted to consider potential sample overlaps. In the next stage, GenomicSEM specifies a system of multivariate regression and covariance that connects the observed phenotypes with one or more latent factors. This approach allows for the modeling of multiple latent factors. The goal is then to identify the parameters (θ) that optimally reduce the discrepancy between the covariance matrix implied by the model (∑(θ)) and the corresponding empirical covariance matrix, denoted as "S."[16]

GenomicSEM constructs a multivariable version of cross-trait linkage disequilibrium score regression (LDSC)[275] to estimate the genetic covariance matrix for each of the included GWAS traits (here NAFLD, T2D, and CAD):

$$S_{\mathrm{LDSC}} = \begin{bmatrix} h_1^2 & & & \\ \sigma_{g1,g2} & h_2^2 & & \\ \vdots & & \ddots & \\ \sigma_{g1,gk} & \sigma_{g2,gk} & \cdots & h_k^2 \end{bmatrix}$$

k is number of observed phenotypes (*k* = 3 for our analyses) and the diagonal elements are heritabilities for the common SNPs for each SNP included in the munged summary statistics used for the LDSC input implemented in GenomicSEM.[16] The other elements are the genetic covariances between NAFLD, T2D, and CAD.

To acquire unbiased estimates and their standard errors, the non-redundant components of the matrix "S" are employed to construct an asymptotic sampling covariance matrix named "VSLDSC." This matrix is composed of regression estimates derived from LDSC. Notably, VSLDSC is a symmetric matrix of order k*, where its diagonal elements represent the sampling variances and its off-diagonal elements indicate the sampling covariances.[16]The V_{SLDSC} matrix is written as:

$$
V_{S_{\text{LDSC}}} = \begin{bmatrix}
\text{s. e.}(h_1^2)^2 & & & & & \\
\text{cov}(h_1^2, \sigma_{g1,g2}) & \text{s. e.}(\sigma_{g1,g2})^2 & & & & \\
\vdots & \vdots & \ddots & & & \\
\text{cov}(h_1^2, \sigma_{g1,gk}) & \text{cov}(\sigma_{g1,g2}, \sigma_{g1,gk}) & & \text{s. e.}(\sigma_{g1,gk})^2 & & \\
\vdots & \vdots & & \vdots & \ddots & \\
\text{cov}(h_1^2, h_j^2) & \text{cov}(\sigma_{g1,g2}, h_j^2) & & \text{cov}(\sigma_{g1,gk}, h_j^2) & & \text{s. e.}(h_j^2)^2 \\
\vdots & \vdots & & \vdots & & \\
\text{cov}(h_1^2, \sigma_{gj,gk}) & \text{cov}(\sigma_{g1,g2}, \sigma_{gj,gk}) & & \text{cov}(\sigma_{g1,gk}, \sigma_{gj,gk}) & & \text{cov}(h_j^2, \sigma_{gj,gk}) \\
\text{cov}(h_1^2, h_k^2) & \text{cov}(\sigma_{g1,g2}, h_k^2) & & \text{cov}(\sigma_{g1,gk}, h_k^2) & & \text{cov}(h_j^2, h_k^2)
\end{bmatrix}
$$

Diagonal elements in V_{SLDSC} are estimated using a jackknife resampling procedure from the LDSC package[275] extended in the GenomicSEM package.[16] Next the effects of the individual SNPs are incorporated: first, the initial input genetic covariance matrix is extended to model covariances between individuals SNPs and each observed phenotype by incorporating a vector of SNP-phenotype covariances, denoted "S_{SNP}" to S_{LDSC}:

The diagonal elements of the VSLDSC matrix are estimated through a jackknife resampling procedure, an approach extended from the LDSC package[275] and incorporated into the GenomicSEM R package. This process is followed by the inclusion of individual SNP effects. Initially, the genetic covariance matrix that serves as the primary input is expanded to model the covariances between individual SNPs and each observed phenotype. This expansion is achieved by integrating a vector of SNP-phenotype covariances, referred to as "S_{SNP}" to S_{LDSC}", which is given by the matrix below:

$$
S_{\text{Full}} = \begin{bmatrix}
\sigma^2_{\text{SNP}} & & & & & \\
\sigma_{\text{SNP},g1} & h_1^2 & & & & \\
\sigma_{\text{SNP},g2} & \sigma_{g1,g2} & h_2^2 & & & \\
\sigma_{\text{SNP},g3} & \sigma_{g1,g3} & \sigma_{g2,g3} & h_3^2 & & \\
\vdots & \vdots & & & \ddots & \\
\sigma_{\text{SNP},gk} & \sigma_{g1,gk} & \sigma_{g2,gk} & \sigma_{g3,gk} & \cdots & h_k^2
\end{bmatrix}
$$

"V_{SFULL}" (i.e., the sampling covariance matrix) associated with S_{FULL} is given by the following:

$$
V_{S_{\text{Full}}} = \begin{bmatrix} V_{S_{\text{SNP}}} & \\ 0 & V_{S_{\text{LDSC}}} \end{bmatrix}
$$

The V_{SFULL} matrix in GenomicSEM analysis is composed of two distinct blocks. The first block, V_{SLDSC}, encompasses the SNP heritabilities, which include the sampling variances and covariances of the latent genetic variances, as well as the genetic covariances determined by the multivariable LDSC analysis. The second block of V_{SFULL} contains the "V_{SSNP}" matrix, which is a representation of the sampling covariance matrix between SNP effects and phenotypes.[16]

For the purposes of calculating SNP variance, we rely on the 1000 Genomes Phase 3 European reference panel,[243] treating this variance as fixed. Consequently, both the SNP sample variance and sampling covariance are set to zero. We then construct the sampling covariances for the SNP genotype covariances using the intercepts from cross-trait multivariate LDSC analysis.[16]

The final section of the V_{SFULL} matrix corresponds to the sampling covariance between the SNP genotype covariances and the genetic variances and covariances. These values are fixed at zero, operating under the assumption that the SNP genotype covariance is independent of the test statistics from other linkage disequilibrium (LD) blocks, except for the specific block where the SNPs are located. Lastly, we derive the sampling variance of the heritabilities and genetic correlations from the sampling variability present in the test statistics across all LD blocks. Correspondingly, the sampling covariances between individual SNPs and these elements are also presumed to be zero.

In Stage 2, parameters of the user-specified SEM model are estimated with either weighted least squares (WLS) or maximum likelihood (ML) estimators using the S_{LDSC} matrix (from stage 1). WLS and ML estimators weigh the matrix information differently, but both estimators minimize the fit error between the model-implied and empirical genetic covariances.[16] We use the WLS estimator as recommended and used in recent GenomicSEM GWAS studies.[84,119] WLS optimizes the fit function by using the diagonal elements in the V_{SLDSC} matrix and adjusting standard errors of the estimates with the off-diagonal elements, indices for the correlations among the sampling errors of the summary statistics.[16] These features make GenomicSEM unbiased and robust up to 100% sample overlap and also unbalanced GWAS sample sizes.[16]

AP4.1.1b. Assessing model fit. To assess the fit of our model, we employed standard indices commonly used in SEM modeling – and have been incorporated into the GenomicSEM framework –including the model χ^2 (chi-square) statistics, the Akaike Information Criterion (AIC), the Comparative Fit Index (CFI), and the Standardized Root Mean Square Residual (SRMR). Except for the χ^2 statistic, each of these indices is interpreted in the traditional SEM context.[16] In GWAS analysis, large sample sizes can potentially inflate the sensitivity of the χ^2 test, making it more likely to yield statistically significant results.[16] Recognizing this, we adhered to the recommendations provided by the developers of GenomicSEM,[510] using the χ^2 statistic more as a comparative metric for model fit, rather than as a strict indicator of statistical significance. Consistent with established guidelines, we considered CFI values greater than 0.90 and SRMR values less than 0.08 as indicative of a good model fit. This approach aligns with standard practices and thresholds in SEM analysis, ensuring that our model evaluation is both rigorous and relevant to the context of our study.[510]

AP4.1.1c. Confirmatory factor analysis. GenomicSEM then uses confirmatory factor analysis (CFA) to estimate the strength of the relationships between the *CM-Factor* and the input/observed GWASs.[16] CFA also facilitates investigation into model fit,[16] and it provided

evidence confirming a good fit for the *CM-Factor* for NAFLD, T2D, and CAD, which confirms that there is a strong genetic architecture linking these traits that is suitable for multivariate GWAS analysis.

AP4.2. Supplementary Aim 3 Results

AP4.2.1. Qsnp heterogeneity testing

In addition to performing the multivariate GWAS of the *CM-Factor*, GenomicSEM was also used to perform SNP-level tests of heterogeneity (Q_{SNP}) to investigate whether each SNP had consistent, pleiotropic effects on the seven input phenotypes that effectively only operate via the shared genetic liability *CM-Factor*. The *CM-Factor* represents a broad genetic liability of T2D, NAFLD, and CAD. To evaluate this potential heterogeneity in SNP effects, we estimated genome-wide Q_{SNP} statistics for each SNP in the multivariate GWAS, which are χ2-distributed test statistics. The null hypothesis of the Q_{SNP} test is that SNP effects on the constituent phenotypes are completely mediated via a common pathway through the *CM-Factor*, so a significant Q_{SNP} test indicates that a given SNP's effects are better explained by trait-specific pathways independent of the *CM-Factor*. In other words, in the absence of heterogeneity, it is expected that a given SNP's effects on the input phenotypes should scale proportionally to the unstandardized factor loadings.

As described by Grotzinger et al.,[16] larger values for Q_{SNP} reflect a violation of the null hypothesis that a SNP is mediated through the latent factor. Genome-wide results for 6,550,214 Q_{SNP} tests for which the method converged are presented in the Manhattan plot in **Figure 5.1** in Chapter 5 (section **2.4.3**) and a Q-Q plot (**Figure AP4.1** and **AP4.2**). We calculated the mean χ^2 and genomic inflation factor (λ_{GC}) for the Q_{SNP} results to be 1.54 and 1.32, respectively, using the ~6 million SNPs for the genome-wide heterogeneity testing. These results indicate that the Q_{SNP} analysis had sufficient power to detect substantial heterogeneity across the genome, but reassuringly, this did not affect our main findings, aligning with the expectation of modeling a latent common factor with SEM, where the *CM-Factor* should primarily capture shared variance rather than the unique features of the model indicators. Additionally, we estimated an LD Score regression intercept of 1.0023 (SE = 0.0212), indicating that the observed inflation in the Q_{SNP} test statistics is not due to bias from population stratification.[16,275]

We applied the clumping algorithm described in the Aim 3 methods sections in Chapter 2 sections to define independent genomic loci to the Q_{SNP} results and found 258 near-independent genome-wide significant Q_{SNPs} (two-sided test P-value < 5×10^{-8}) in 151 loci. Only 21 of these SNPs were also genome-wide significant for the *CM-Factor*, indicating minimal overlap between the *CM-Factor* loci and the heterogenous loci. In fact, only 805 of the 19,983 total SNPs (~4%) surpassing genome-wide significance for the *CM-Factor* also had Q_{SNP} P-values < 5×10^{-8}, suggesting that the *CM-Factor* signature captures a shared dimension of the cardiometabolic genetic liability across T2D, NAFLD, and CAD.

Notably, there was strong representation of lipid-related SNPs and genes among the Q_{SNP} loci. For example, the strongest and most salient example of a trait-specific association is SNP

rs429358 (two-sided Q_{SNP} P-value=4.15×10^{-68}; nearest gene: APOE).[511] Other significant SNPs include rs7454157 (two-sided Q_{SNP} P-value=1.95×10^{-59}; nearest gene: PHACTR1), rs12740374 (two-sided Q_{SNP} P-value=1.07×10^{-57}; nearest gene: CELSR2, part of the *CELSR2-PSRC1-MYBPHL-SORT1* locus that has implicated in cardiovascular disease development[511]), and rs148812085 (two-sided Q_{SNP} P-value=2.86×10^{-49}; nearest gene: WDR12) (**Table AP2.7**). Additionally, the SNP rs1051338 near the LIPA gene, rs55997232 in the LDLR gene, are critical for lipid and cholesterol metabolism,[426] rs11591147 in the PCSK9 gene, known for its influence on LDL cholesterol levels,[46] and the nearest gene of rs2294917, PNPLA3, is strongly implicated in liver fat content and lipid levels,[309-312] highlight specific pathways in lipid metabolism and cardiovascular health (or NAFLD for the PNPLA3 locus) that might operate independently of the broad *CM-Factor* genetic liability to impact CAD or NAFLD.

AP4.3. Aim 3 Supplementary Discussion

AP4.3.1. Extended discussion of the proteins prioritized by the two-step MR study

In Chapter 5, we focus our discussion of the 4 proteomic mediators prioritized by the two-step MR study on ENO3. Here, we present an extended discussion of the other 3 proteomic mediators of the obesity-*CM-Factor* relationship.

In line with previous work, we found that increased BMI reduced SHBG (sex hormone-binding globulin),[512] and, in turn, that lowered SHBG levels increase the risk for cardiometabolic disease.[513,514] The SHBG findings was robust in replication analyses with the deCODE and FinnGen data and SHBG protein levels demonstrated evidence for a shared casual variant with the *CM-Factor* in the SHBG locus, together providing compelling genetics-based evidence linking SHBG in the BMI-to-*CM-Factor* pathway. SHBG is a hepatic-derived protein responsible for the systemic transport and bioavailability regulation of androgens and estrogens.[512] Plasma SHBG levels are strongly impacted by nutritional state, metabolism, and hormonal factors,[513] and while the mechanisms linking SHBG with cardiometabolic disease remain largely unknown, SHBG levels have demonstrated bi-directional relationships with circulating lipids in directions consistent lowered SHBG being adverse for cardiometabolic health (i.e., reduced SHBG and increased LDL-C, reduced HDL-C, and increased levels of atherogenic lipid subfractions).[514] There are also enzymatic links between SHBG and lipids: hepatic lipase, involved in the breakdown of HDL-C, is stimulated by androgens and high levels of SHBG reduce free androgen levels, leading to decreased hepatic lipase activity and increased HDL-C levels. Additionally, studies suggest that the interaction between AMP-activated protein kinase and peroxisome proliferator-activated receptor can regulate hepatic nuclear factor-4α (HNF-4α) expression, which in turn upregulates SHBG expression. HNF-4α also influences the transcription of various genes related to lipid metabolism,[513,514] potentially explaining the correlation between circulating SHBG levels and lipid metabolism and supporting previous work suggesting that low SHBG levels are early indicators of cardiovascular risk in obesity and metabolic syndrome and may be a useful clinical biomarker for cardiometabolic disease risk.[513]

It has been shown that overexpression of SHBG prevents weight gain and fat accumulation caused by a high-fat diet. Furthermore, it eliminates the rise in insulin, leptin, and resistin levels, providing protection against high-fat diet-induced obesity,[515] supporting our two-step MR findings emphasizing its mediating role between obesity and the *CM-Factor*. Unsurprisingly, given its widespread physiological role in human health, our Phe-MR of SHBG found widespread effects of SHBG protein levels in many clinical domains, e.g., SHBG impacted aspects of neurocognition, self-reported mood, and myasthenia gravis, as well as a range of cardiometabolic traits like blood pressure, peripheral artery disease, heart rate (both arrhythmia diagnosis and electrocardiogram-measured resting heart rate). The effects on heart rate are notable given that SHBG expression in cardiomyocytes is linked with dilated cardiomyopathy, potentially to control testosterone levels in the myocardium and activate androgen signaling pathways.[515] Further, experimentally, suppressing SHBG causes cardiac disorders by mimicking low testosterone conditions, while physiological SHBG and testosterone levels have cardioprotective effects, indicating that dysfunctional SHBG production could explain adverse cardiac metabolic effects of androgen deficiency.[515]

PTPRR, a member of the R7 subfamily of RPTPs, plays an important role in modulating insulin signaling pathways.[516] It contains a single intracellular protein tyrosine phosphatases (PTP) domain and is known to interact with components of the mitogen-activated protein kinase (MAPK) pathway.[516] All isoforms of PTPRR have a kinase interacting motif, which enables the dephosphorylation of MAPKs like ERK1/2/5 and p38, leading to their inactivation and preventing their translocation to the nucleus.[516] The involvement of PTPRR in insulin signaling suggests that it acts as a negative regulator, modulating the activity of MAPKs that are crucial for various cellular processes, including glucose metabolism.[516] By dephosphorylating these kinases, PTPRR can impact insulin signaling pathways, potentially influencing insulin sensitivity and glucose homeostasis. The regulatory function of PTPRR on MAPKs indicates its potential role in the pathophysiology of T2D. Insulin resistance, often linked with obesity[517] and a hallmark of T2D and broad cardiometabolic disease,[516] involves impaired insulin signaling, often due to disruptions in the balance of phosphorylation-dephosphorylation processes regulated by kinases and phosphatases. Given PTPRR's ability to modulate key signaling pathways, its dysregulation could contribute to the development of insulin resistance and cardiometabolic disease, potentially because of its dysregulation due to increased BMI.

PTPs are considered a potential pathway for the development of new antidiabetic drugs; however, preliminary efforts in finding a PTP inhibitor suitable for clinical use have failed (only small molecule and antisense PTP1B inhibitors have been assessed in early phase clinical trials, see Sharma et al. for a detailed overview[518]) and have demonstrated some efficacy in T2D. For example, PTP1B antisense therapies have reached phase 2 clinical trials (ISIS-113715 and IONIS-PTP-1B$_{Rx}$) and demonstrated the ability to sensitize insulin, normalize blood glucose levels without causing hypoglycemia, and reduce LDL-C levels and weight in T2D patients.[518] PTPRR has not been targeted, to date, however, as with ENO3, our Phe-MR of PTPRR protein levels revealed a neutral side effect profile across 366 biomarkers and disease outcomes, supporting follow up investigation and validation as a potential therapeutic target.

The final proteomic mediator highlighted by the two-step MR was the liver biomarker GGT1 (gamma-glutamyltransferase, commonly termed “GGT”). GGT, while generally considered an indicator of hepatobiliary dysfunction, has also been recently found to play important role in key pathophysiological processes, such as oxidative stress and lipid peroxidation, which are crucial for the development of insulin resistance and metabolic syndrome, and predict T2D[519,520] underscoring the potential of GGT as a biomarker not only for liver-related conditions but also for broader cardiometabolic health and disease.[520] It has been therefore suggested that elevated GGT levels may be useful in implementing preventive measures and monitoring strategies to mitigate the risk of developing T2D or multimorbid cardiometabolic diseases.

While GGT has been primarily considered an indicator of liver diseases in clinical practice, circulating GGT levels have been shown to be involved in cardiometabolic disease mechanisms, suggesting targeting GGT or its pathways may facilitate the development of lead to new preventive and treatment strategies.[520] There are now compelling lines of evidence implicating GGT in atherosclerotic plaque formation and rupture. For example, it has shown that catalytically active GGT is present within atherosclerotic coronary plaques from autopsies and surgical endoarterectomies,[521] leading to the hypothesis that serum GGT may be partially adsorbed onto LDL lipoproteins, which can carry GGT activity into plaques, corresponding with serum GGT levels.[522] Further, GGT-mediated reactions catalyze the oxidation of LDL lipoproteins, contributing to oxidative events that influence plaque evolution and rupture,[523] and GGT also is believed to play a central role in forming the fibrous cap, inducing apoptosis of lesion cellular elements, causing plaque erosion and rupture, and enhancing platelet aggregation and thrombosis.[522] By contrast, the biological mechanisms linking GGT with T2D are less well studied, but it has been suggested that it may be involved in the pathogenesis of T2D through inflammation and oxidative stress or be a marker of insulin resistance given its link with BMI, NAFLD, and general cardiometabolic functioning (observed in the literature and confirmed with our Phe-MR).[524] While primarily studied in oncology,[525] GGT inhibition is considered a viable therapeutic target, having demonstrated efficacy for various cancer types (elevated GGT expression on tumor cells increases cell proliferation and chemotherapy resistance[526]) and chemotherapy side effects in preclinical models (e.g., GGT inhibition promoted collagen production in oral mucosa, reducing the severity and improving the recovery time for 5-Fluorouracil-induced oral mucositis in mice[527]). Current GGT inhibitor have proven to be too toxic for human use; however, ongoing efforts are underway to identify safe GGT inhibitors for applications to a range of tumor types,[526] suggesting there may be future potential cardiometabolic repurposing opportunities once classes of GGT inhibitors have been more developed.

AP4.3.2. Limitations of the transcriptomic imputation gene prioritization analyses

Transcriptomic imputation,[103] which form the methodological basis of performing transcriptome-wide association studies (TWASs) has advanced our understanding of the genetic basis of complex traits and diseases by integrating gene expression data with GWASs. However, it has several limitations that are necessary for interpretating the 243 high-confidence genes for the *CM-Factor* prioritized by the transcriptomic imputation, colocalization, and conditional testing.

One of the primary limitations of TWAS is the potential for spurious prioritization of genes due to the sharing of eQTLs. eQTLs are genomic loci that explain variation in gene expression levels, and they can affect multiple genes within a locus. This means that a single eQTL can influence the expression of several genes, leading TWAS to prioritize multiple genes at the same locus. Some of these genes may not be causally related to the trait of interest but are instead identified due to their shared regulation by the same eQTL.[487] Second, TWAS accuracy is highly dependent on the tissue specificity of the expression data used. Expression levels and eQTL strengths can vary significantly across different tissues. Using expression data from non-trait-related tissues can lead to substantial inaccuracies in gene prioritization. For instance, genes that are not expressed or have weak eQTLs in the tissue used for TWAS analysis may be missed even if they are causally related to the trait in another, more relevant tissue. This tissue bias can result in both false positives and false negatives, complicating the interpretation of TWAS results.[487]

Linkage disequilibrium (LD) can also confound TWAS results. LD refers to the non-random association of alleles at different loci. In GWAS, LD can obscure the identification of causal variants because nearby variants are often inherited together. Similarly, in TWAS, LD can lead to the co-regulation of genes, where multiple genes appear to be associated with the trait due to their physical proximity and shared regulatory elements rather than due to a direct causal relationship. This can result in the identification of multiple significant genes within a single locus, complicating the pinpointing of the true causal gene.[487] Third, TWAS relies on predictive models of gene expression that are trained using reference transcriptome datasets. These models often assume a linear relationship between genotype and expression, which may not capture the full complexity of gene regulation. Non-linear interactions, epigenetic modifications, and post-transcriptional processes can all influence gene expression but are typically not accounted for in these models. As a result, the imputed gene expression levels may not accurately reflect the true biological variability, leading to potential misidentification of causal genes.[487] Fourth, Co-regulation of genes, where multiple genes are regulated by the same eQTL or regulatory elements, can lead to false-positive associations in TWAS. Genes with correlated expression due to shared regulatory mechanisms may be identified as associated with the trait even if only one of them is truly causal. This problem is exacerbated when the expression prediction models include variants that are in LD with the causal variant, leading to correlated predicted expression levels that do not reflect actual causality.[487] Finally, TWAS typically does not account for environmental influences or temporal changes in gene expression. Gene expression can be highly dynamic, responding to environmental factors, physiological states, and temporal changes such as circadian rhythms. The static nature of the reference expression data used in TWAS means that these dynamic factors are not considered, potentially leading to incomplete or inaccurate associations between gene expression and the trait.[487]

AP4.4. Aim 3 Supplementary Figures

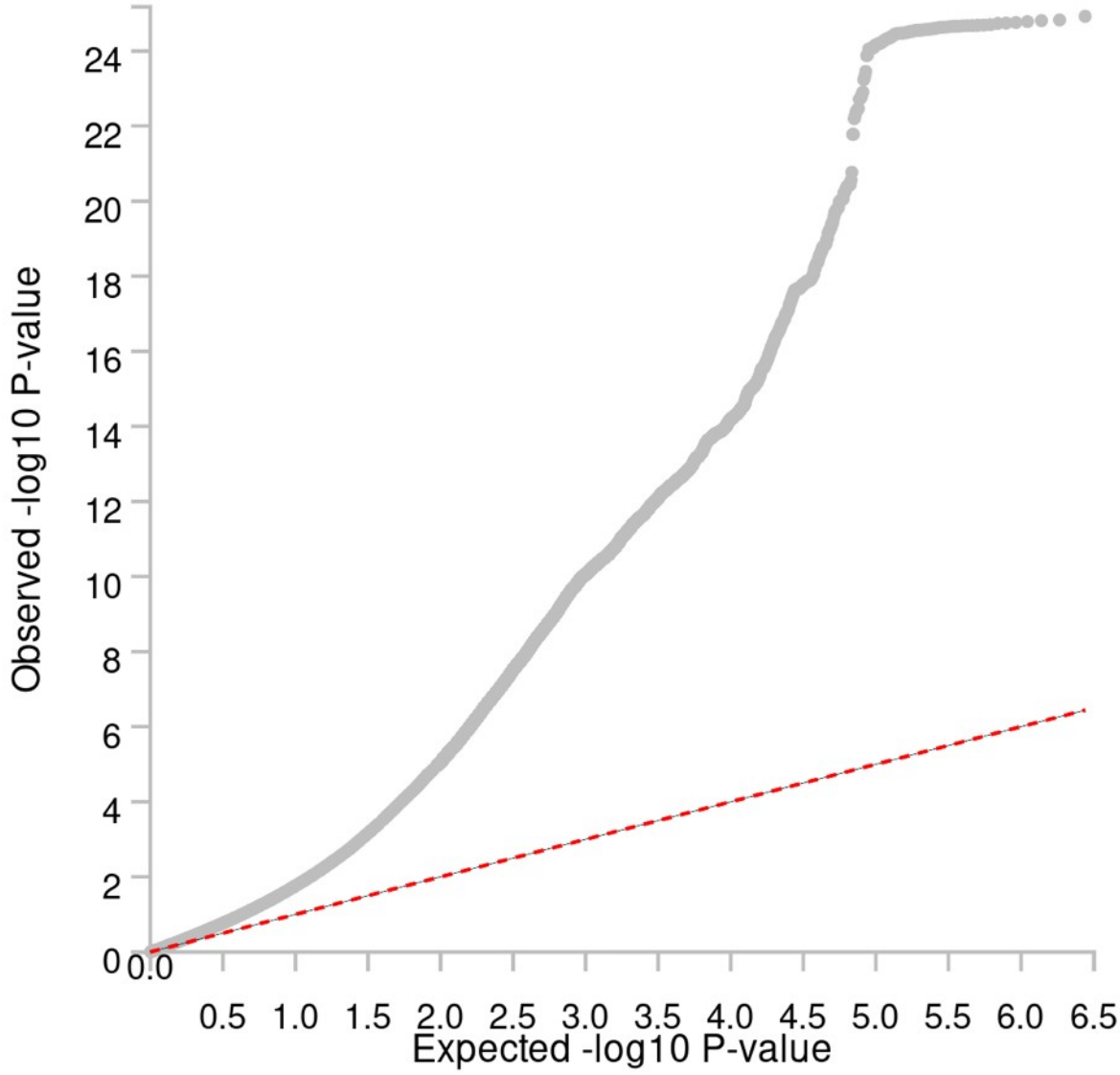


Figure AP4.1. The CM-Factor GWAS Q-Q plot for the P-values. *The plot displays observed versus expected −log10(P-values) for the heterogeneity test across genetic variants. Points falling along the diagonal line indicate concordance with the null hypothesis of no heterogeneity, while deviations suggest variants with heterogeneity in effects across studies or subgroups. The distribution provides insights into the presence of heterogeneity, where substantial deviations above the line indicate potential heterogeneity signals.*

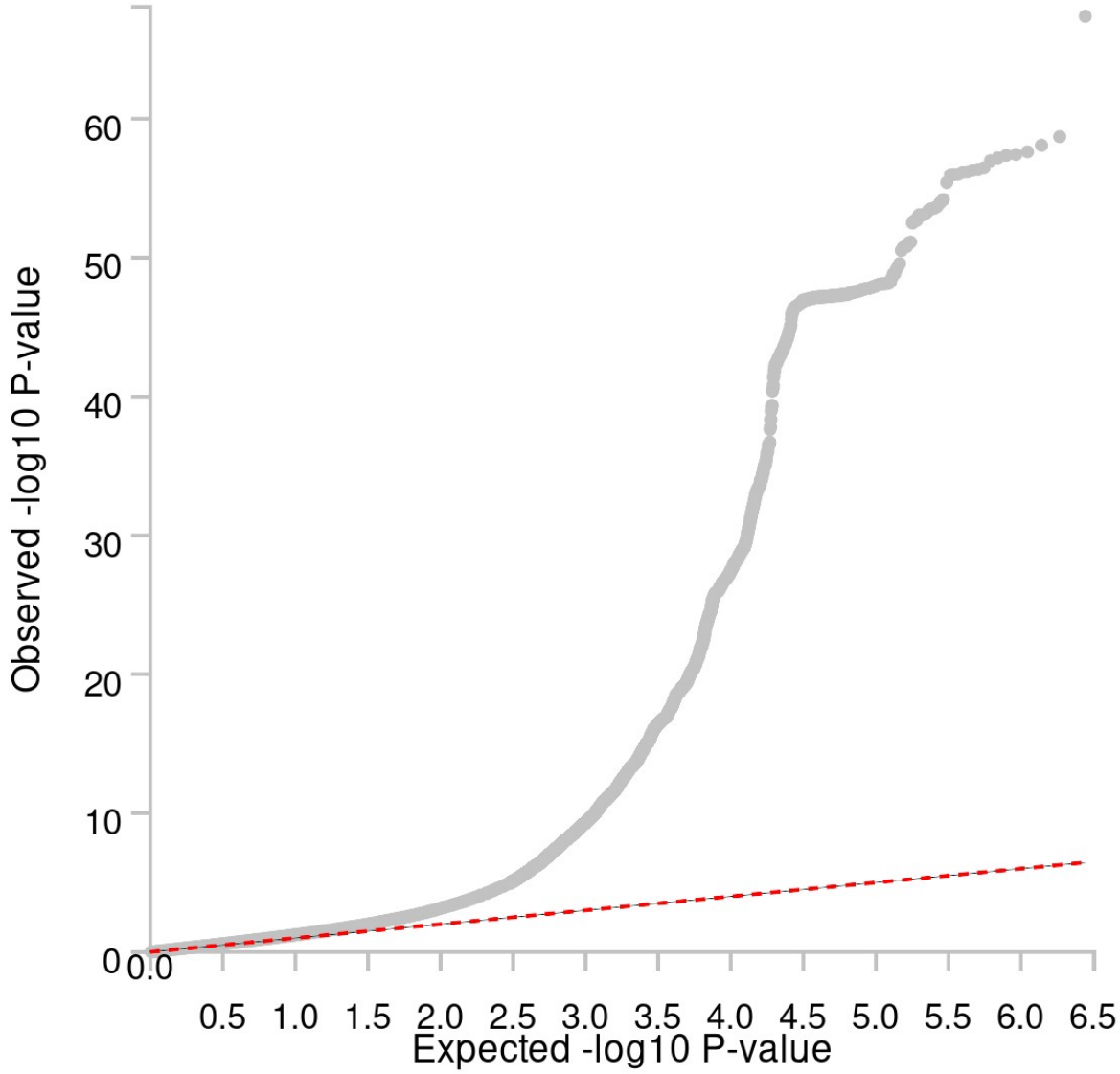


Figure AP4.2. The CM-Factor GWAS Q-Q plot for the Q_{SNP} P-values (heterogeneity tests). *The plot displays observed versus expected −log10(Q SNP P-values) for the heterogeneity test across genetic variants. Points falling along the diagonal line indicate concordance with the null hypothesis of no heterogeneity, while deviations suggest variants with heterogeneity in effects across studies or subgroups. The distribution provides insights into the presence of heterogeneity, where substantial deviations above the line indicate potential heterogeneity signals.*

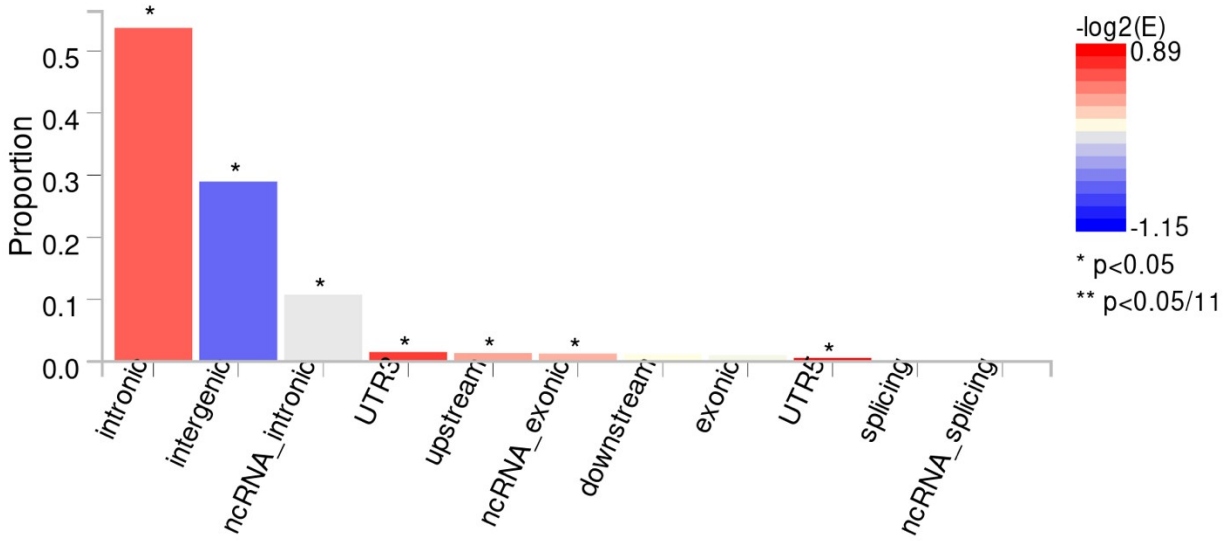


Figure AP4.3. Overview of the functional consequences of the lead loci SNPs of the CM-Factor. *Plotted are the proportion of the 523 independent lead SNPs comprising the CM-Factor genetic signature. The x-axis lists the SNP categories. Bars are colored by log_2-transformed enrichment value (E is the proportion of genome-wide significant SNPs in a category divided by the proportion of all analyzed SNPs in the same functional category). Asterisks indicate enrichment or depletion in comparison (red is enrichment, blue depletion) with all analyzed SNPs based on Fisher's exact test (two-sided). ncRNA_intronic: non-coding RNA are intronic regions encode for non-coding RNAs that have functional roles without being translated into proteins. ncRNA_splicing: SNP has position involved in the regulation and facilitation of RNA splicing UTR: untranslated region (UTR5 is upstream of the gene start position and UTR3 indicates the SNP position is downstream of the gene coding region).*

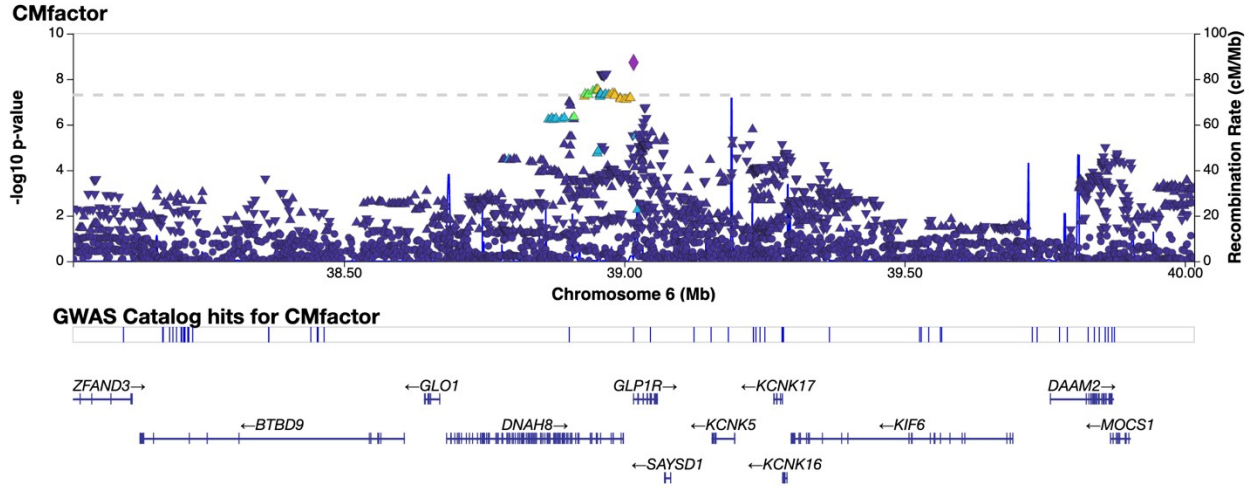


Figure AP4.4. LocusZoom plot of the locus with the exonic lead variant, rs77424687 (GLP1R locus (chromosome 6:39016636)*. The x-axis is the genomic coordinates of the locus, and the y-axis is the −log10 P-value of the multivariate GWAS SNP association analyses. The presented P-values are two-sided and have not been adjusted for multiple testing. The variants are colored according to the SNP-SNP linkage disequilibrium (LD) (shown in the legend on the figure) and the genes located within the window are shown on the bottom track (labeled using their HGNC symbols).*

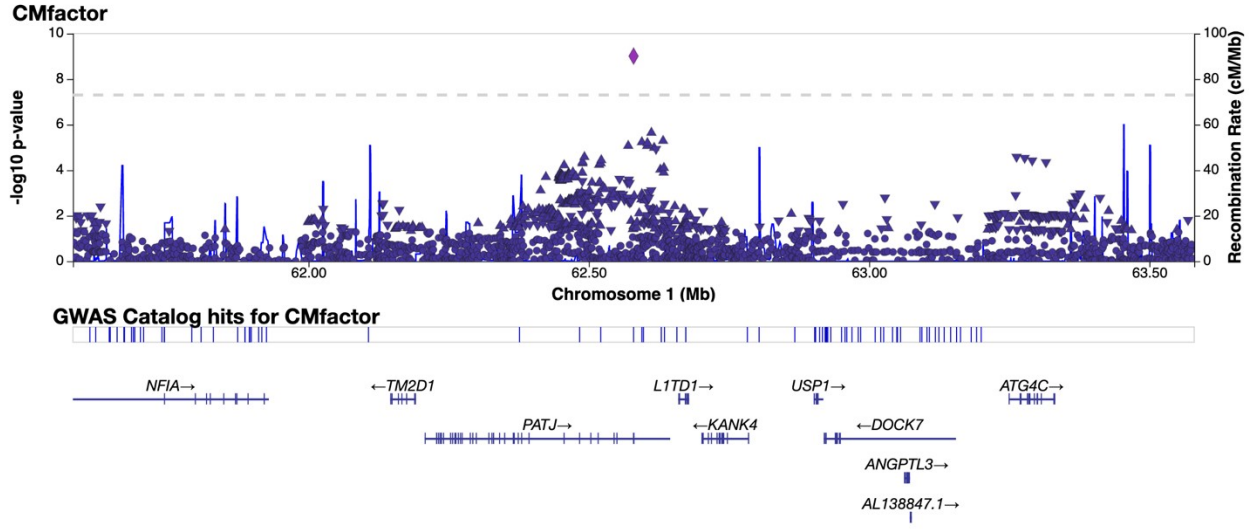


Figure AP4.5. LocusZoom plot of the locus with the exonic lead variant rs12140153 (INADL/ PATJ) locus (chromosome 1:62579891). *The x-axis is the genomic coordinates of the locus, and the y-axis is the −log10 P-value of the multivariate GWAS SNP association analyses. The presented P-values are two-sided and have not been adjusted for multiple testing. The variants are colored according to the SNP-SNP linkage disequilibrium (LD) (shown in the legend on the figure) and the genes located within the window are shown on the bottom track (labeled using their HGNC symbols).*

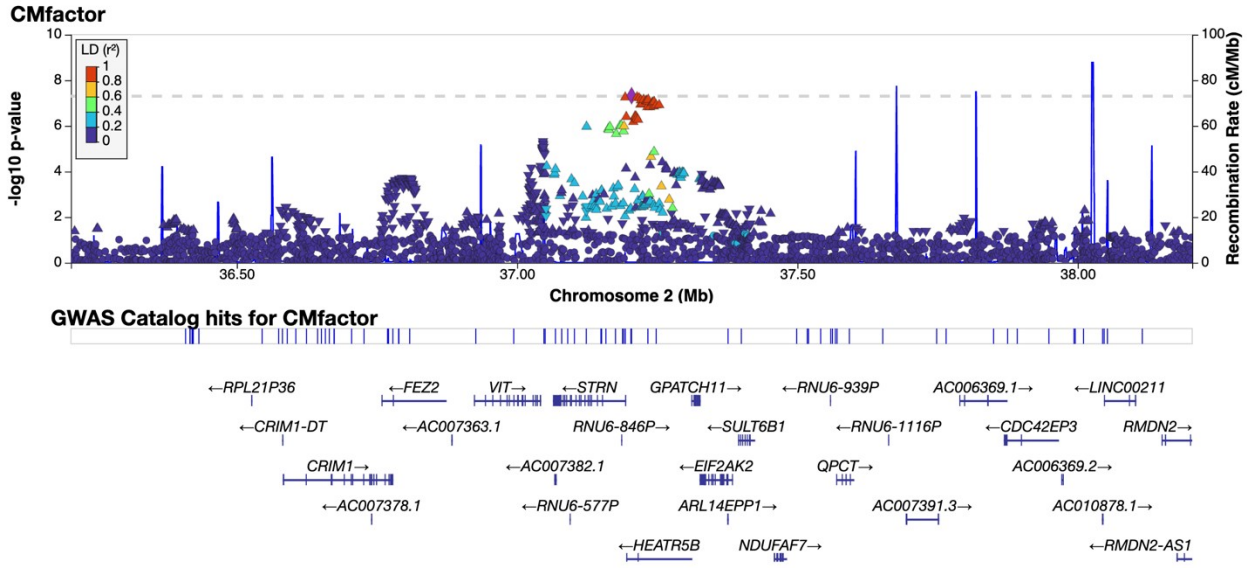


Figure AP4.6. LocusZoom plot of the locus with the lead variant, rs77424687 (chromosome 2:37204168). *The x-axis is the genomic coordinates of the locus, and the y-axis is the −log10 P-value of the multivariate GWAS SNP association analyses. The presented P-values are two-sided and have not been adjusted for multiple testing. The variants are colored according to the SNP-SNP linkage disequilibrium (LD) (shown in the legend on the figure) and the genes located within the window are shown on the bottom track (labeled using their HGNC symbols).*

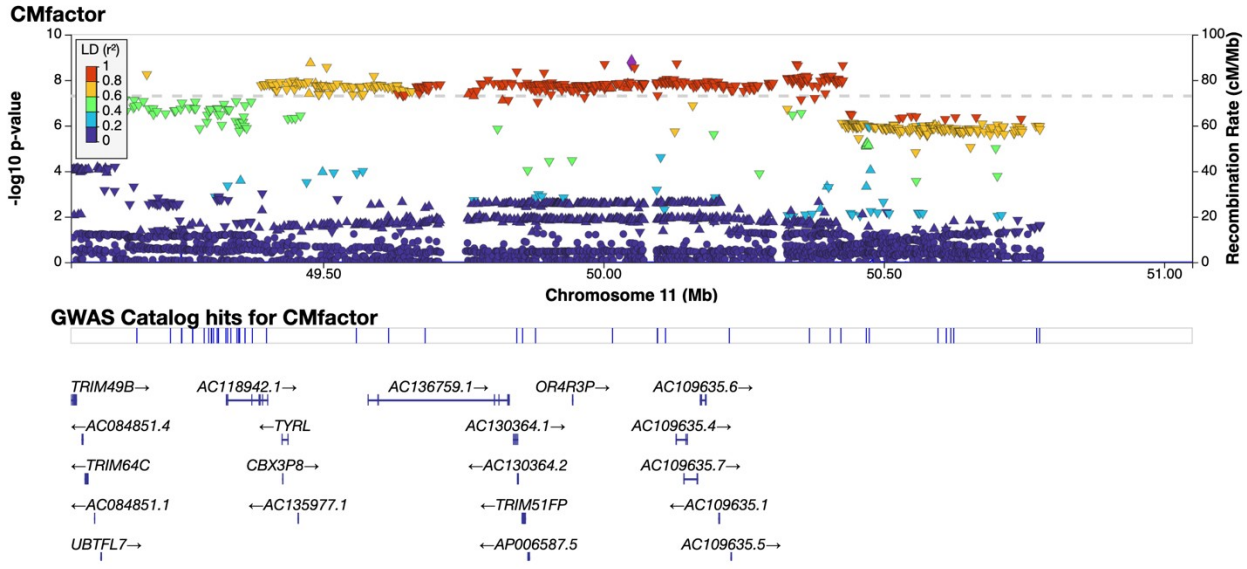


Figure AP4.7. LocusZoom plot of the locus with the lead variant, rs139562826 (chromosome 11:50050097). *The x-axis is the genomic coordinates of the locus, and the y-axis is the −log10 P-value of the multivariate GWAS SNP association analyses. The presented P-values are two-sided and have not been adjusted for multiple testing. The variants are colored according to the SNP-SNP linkage disequilibrium (LD) (shown in the legend on the figure) and the genes located within the window are shown on the bottom track (labeled using their HGNC symbols).*

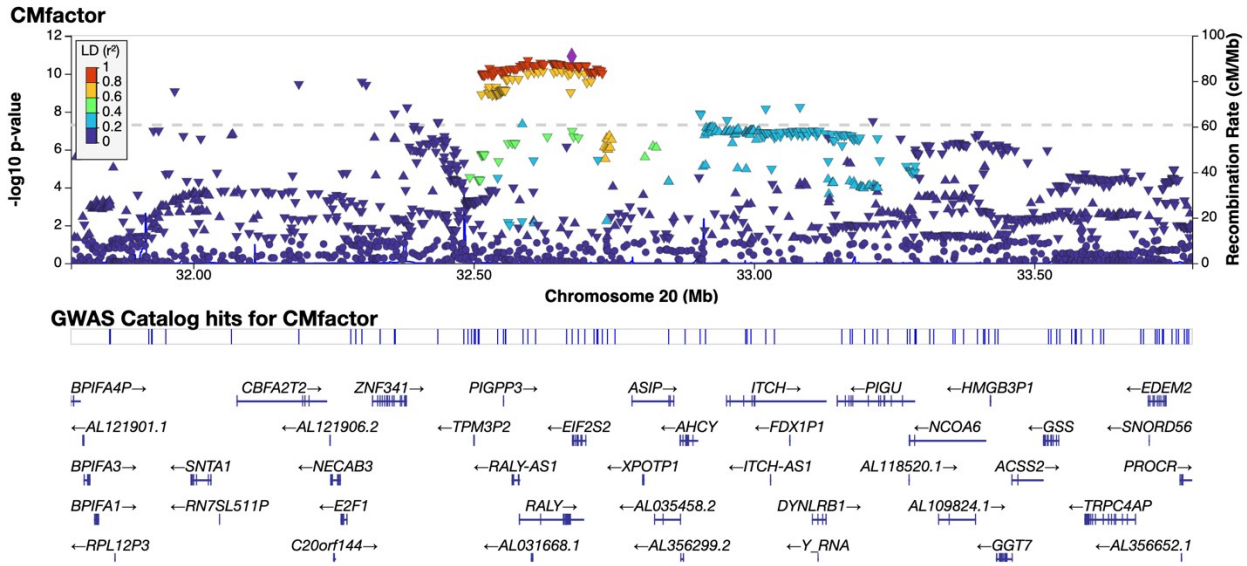


Figure AP4.8. LocusZoom plot of the locus with the lead variant, rs1737897 (chromosome 20:31028723). *The x-axis is the genomic coordinates of the locus, and the y-axis is the −log10 P-value of the multivariate GWAS SNP association analyses. The presented P-values are two-sided and have not been adjusted for multiple testing. The variants are colored according to the SNP-SNP linkage disequilibrium (LD) (shown in the legend on the figure) and the genes located within the window are shown on the bottom track (labeled using their HGNC symbols).*

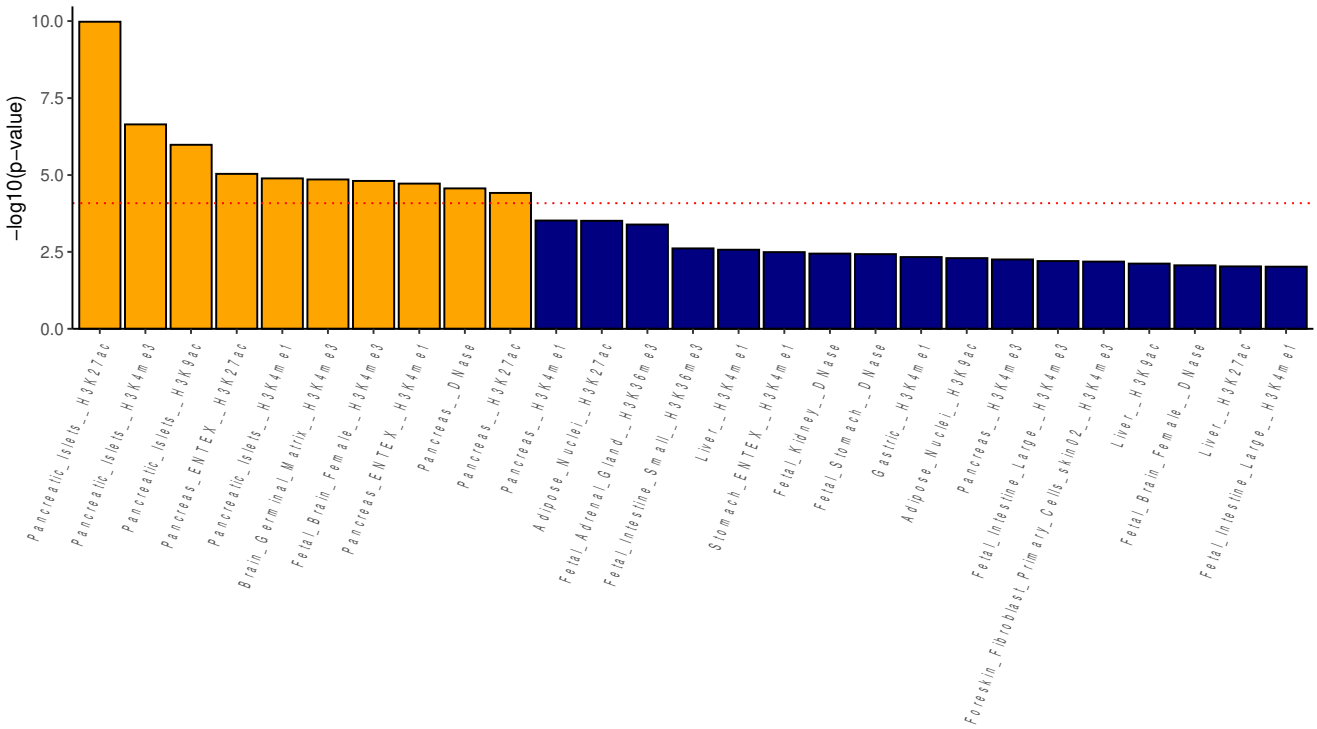


Figure AP4.9. Bulk tissue enrichment of the CM-Factor using S-LDSC. *The bar plots depict the $-log_{10}$ P-values on the y-axis and tissue/marker type on the x-axis. Each bar represents results from S-LDSC analyses using chromatin-based datasets related to the pancreas, liver, cardiovascular, and adipose tissues. Plotted are relevant tissues with enrichment P-values < 0.01. The presented P-values are one-sided and have not been adjusted for multiple testing. Tissue-specific regulatory elements are indicated by histone 3 acetylation (H3K[X]ac) or DNase hypersensitivity (for open chromatin) and H3K4me1 (for enhancers). The Bonferroni-adjusted P-value threshold is 8.18×10^{-5} (0.05/489 tissues) (red dotted line). Enrichment surpassing the Bonferroni correction are indicated in orange.*

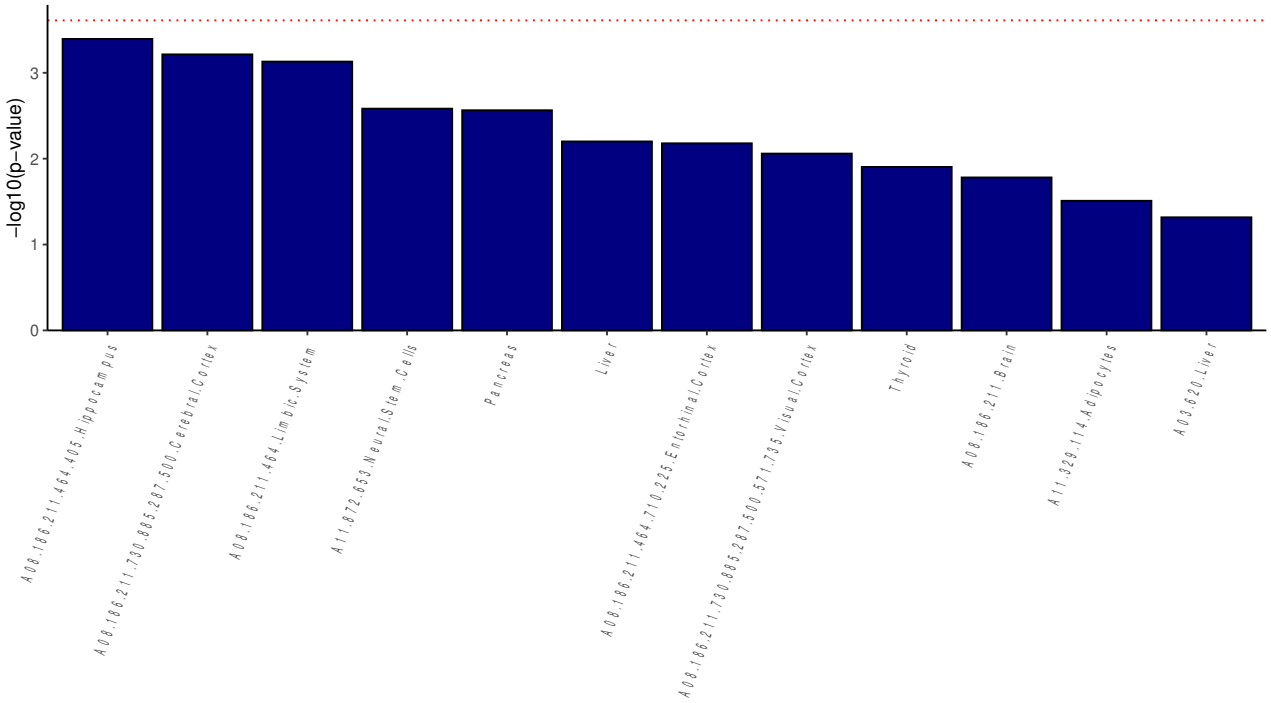


Figure AP4.10. Bulk tissue enrichment of the CM-Factor using S-LDSC. *Bars represent the −log10 of the unadjusted P-values of enrichment for the CM-Factor. Plotted are all tissues with enrichment P-values < 0.05. The presented P-values are one-sided and have not been adjusted for multiple testing. The red dotted line is the correction for multiple comparisons (0.05/205 tissue datasets).*

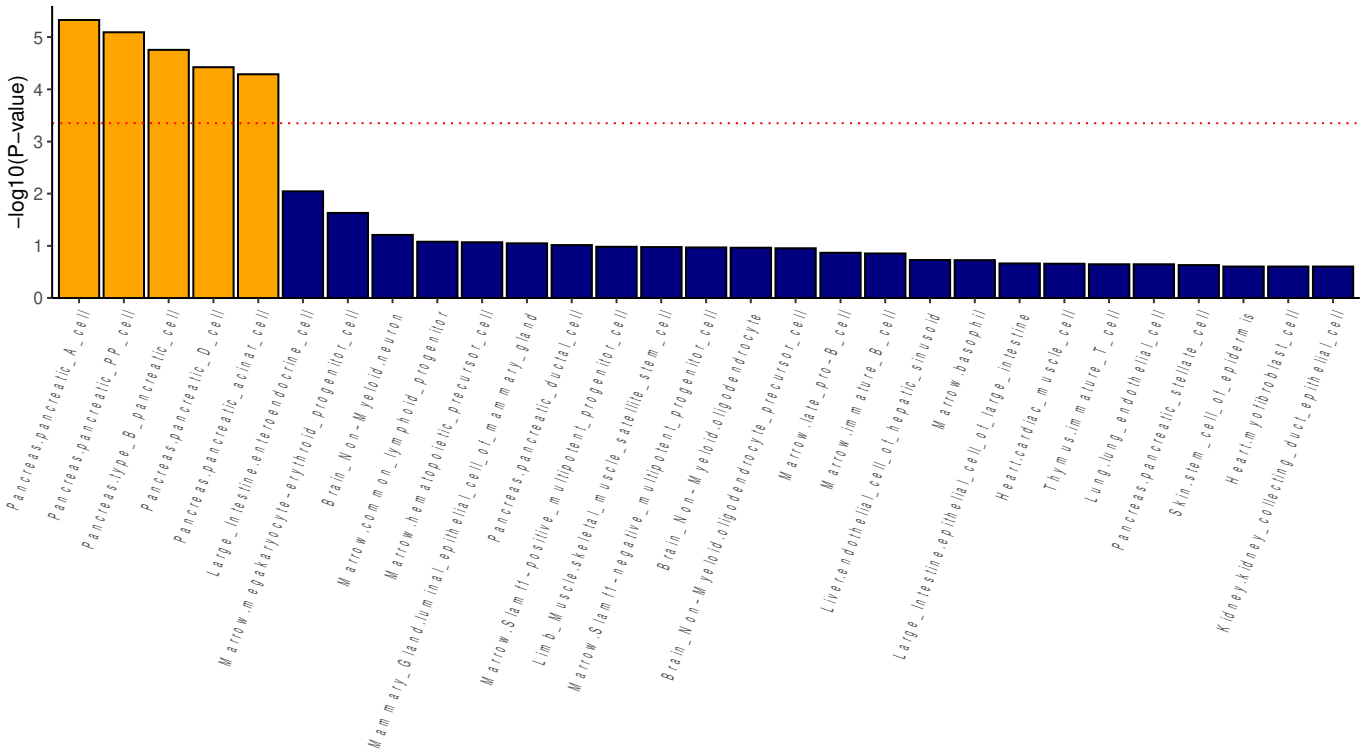


Figure AP4.11. Cell-type enrichment of the CM-Factor using the CELLECT pipeline and S-LDSC. *Bars represent the −log10 of the unadjusted P-values of enrichment for the CM-Factor in the 115 cell types comprising the Tabula Muris adult scRNA-seq dataset. Plotted are all tissues with enrichment P-values < 0.25. The presented P-values are one-sided and have not been adjusted for multiple testing. The red dotted line is the correction for multiple comparisons (0.05/115 cell types).*

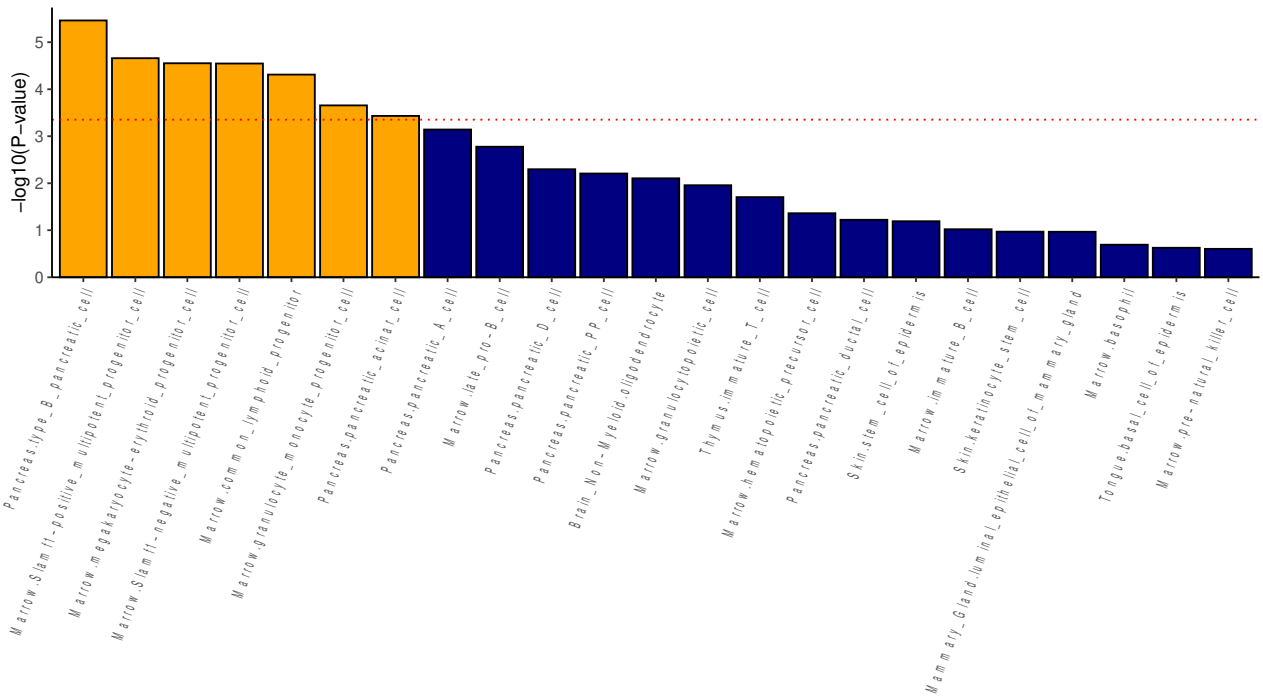


Figure AP4.12. Cell-type enrichment of the CM-Factor using the CELLECT pipeline and MAGMA gene prioritization. *Bars represent the −log10 of the unadjusted P-values of enrichment for the CM-Factor in the 115 cell types comprising the Tabula Muris adult scRNA-seq dataset. The presented P-values are one-sided and have not been adjusted for multiple testing. In contrast to the enrichment based upon the partitioned heritability scores of the CM-Factor, these enrichment results used the MAGMA gene prioritization method. Plotted are all tissues with enrichment P-values < 0.25. The red dotted line is the correction for multiple comparisons (0.05/115 cell types).*

Figure AP4.13. Bubble plots and bar plots for MR Study 2 phenome-wide Mendelian randomization (MR) outcomes with beneficial relationships aligned with the therapeutically indicated direction for the CM-Factor. (a) *presents the beneficial relationships for the 41 druggable genes grouped by the clinical category. Only traits that had MR estimate P-values (from two-sided tests) surpassing Phe-MR analysis multiple correction threshold (1.37×10^{-4} [0.05/366 outcomes]), were included in the.* ***(b)*** *provides a circular bar plot with counts of beneficial relationships for each gene.*

Figure AP4.14. Bubble plots and bar plots for MR Study 2 phenome-wide Mendelian randomization (MR) outcomes with adverse relationships aligned with the therapeutically indicated direction for the CM-Factor. (a) *presents the beneficial relationships for the 41 druggable genes grouped by the clinical category. Only traits that had MR estimate P-values (from two-sided tests) surpassing Phe-MR analysis multiple correction threshold (1.37×10^{-4} [0.05/366 outcomes]), were included in the.* ***(b)*** *provides a circular bar plot with counts of adverse relationships for each gene.*

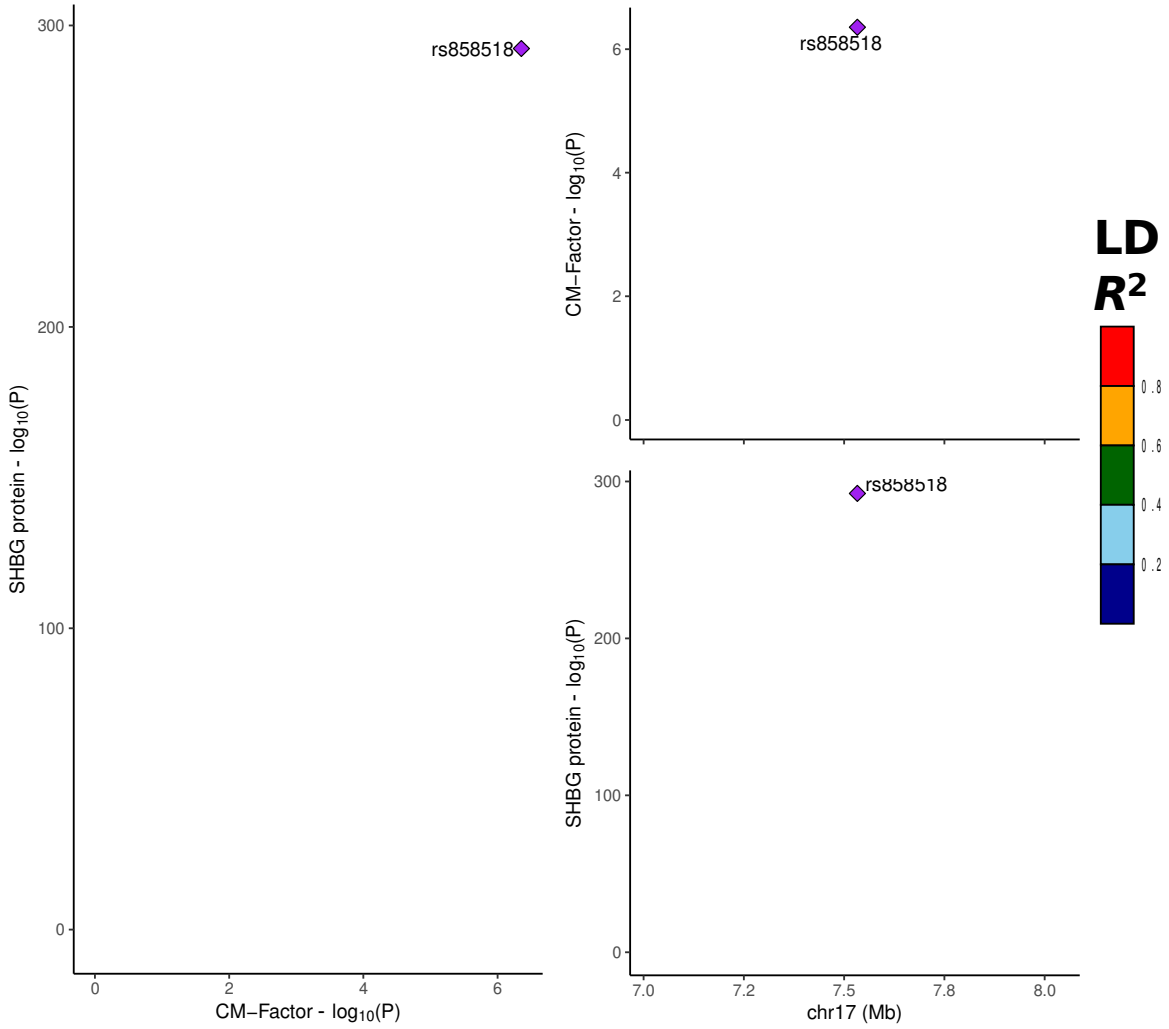


Figure AP4.15. Regional plot of the colocalization between SHBG protein levels and the CM-Factor. *Sex hormone binding globulin (SHBG) was one of the 8 proteins to emerge from the two-step MR analyses investigating the pathway linking body mass index and the CM-Factor. SHBG demonstrated strong evidence of colocalization in both the primary and replication datasets (PP.H4 >0.8 in both proteomic dataset). Plotted are the locus-based results of the primary proteomic SHBG data (from first release of the UKB Pharma Proteomics Project [PPP] data). On the left, the x-axis is the $-\log_{10}$ (P-values) of the SNP associations with the CM-factor and the y-axis is the $-\log_{10}$ (P-values) of the SNP associations with the SHBG protein levels. On the right panels, the x-axes are the genomic locations of the SHBG SNPs and y-axes are the $-\log_{10}$ (P-values) of the SNPs with the SHBG protein (bottom) and CM-Factor (top panel), respectively. Colocalization was performed under the single causal variant assumption using 500 kb windows around the genomic coordinates of the SHBG locus. SNPs are colored based upon their SNP-SNP linkage disequilibrium (LD) R^2.*

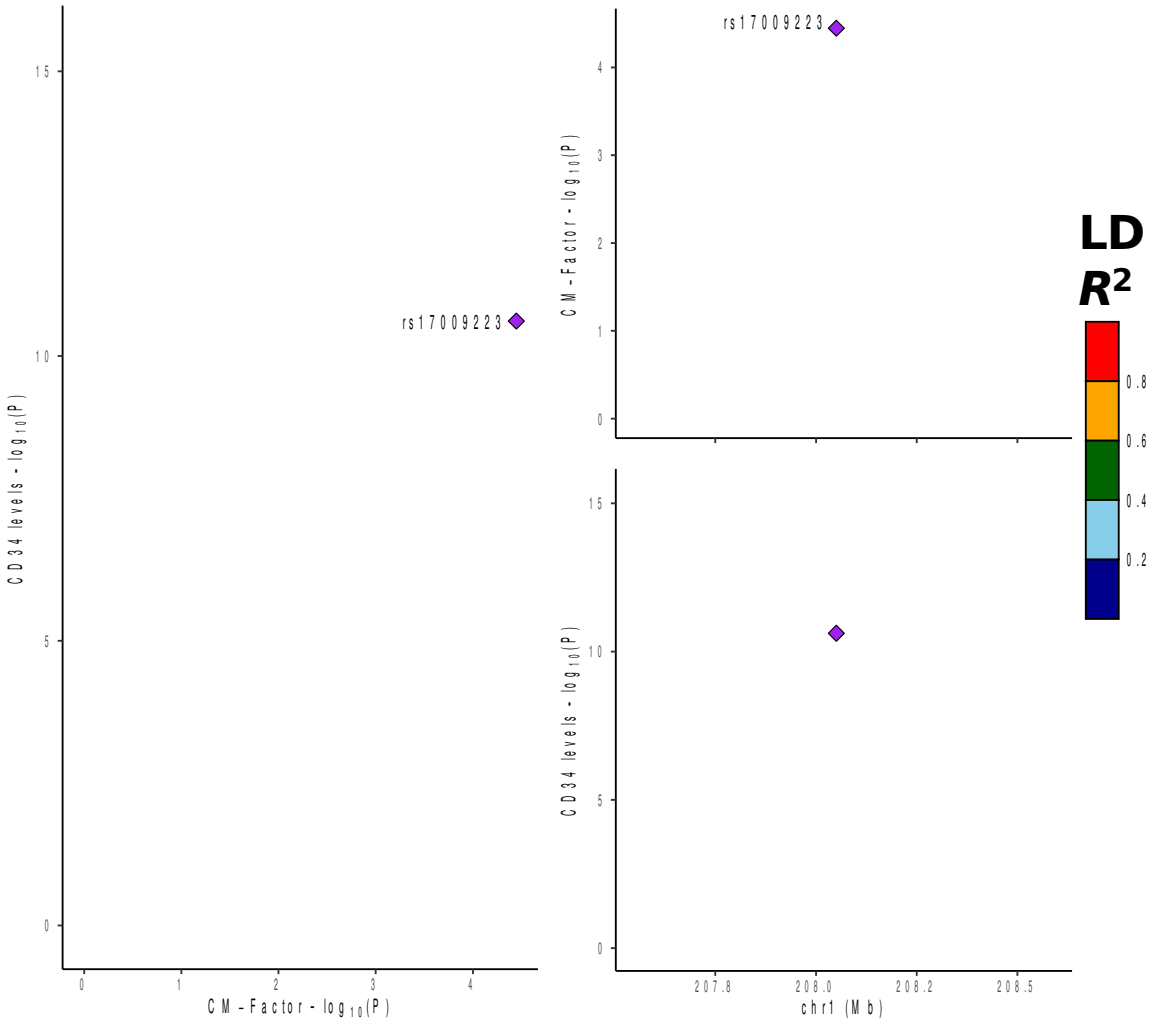


Figure AP4.16. Regional plot of the colocalization between CD34 protein levels and the CM-Factor. *The CD34 antigen (CD34) was one of the 8 proteins to emerge from the two-step MR analyses investigating the pathway linking body mass index and the CM-Factor. CD34 demonstrated strong evidence of colocalization only in the primary (PP.H4 >0.8) in proteomic dataset). Plotted are the locus-based results of the primary proteomic CD34 data (from first release of the UKB Pharma Proteomics Project [PPP] data). On the left, the x-axis is the $-\log_{10}$ (P-values) of the SNP associations with the CM-factor and the y-axis is the $-\log_{10}$ (P-values) of the SNP associations with the CD34 protein levels. On the right panels, the x-axes are the genomic locations of the CD34 SNPs and y-axes are the $-\log_{10}$ (P-values) of the SNPs with the CD34 protein (bottom) and CM-Factor (top panel), respectively. Colocalization was performed under the single causal variant assumption using 500 kb windows around the genomic coordinates of the CD34 locus. SNPs are colored based upon their SNP-SNP linkage disequilibrium (LD) R^2.*

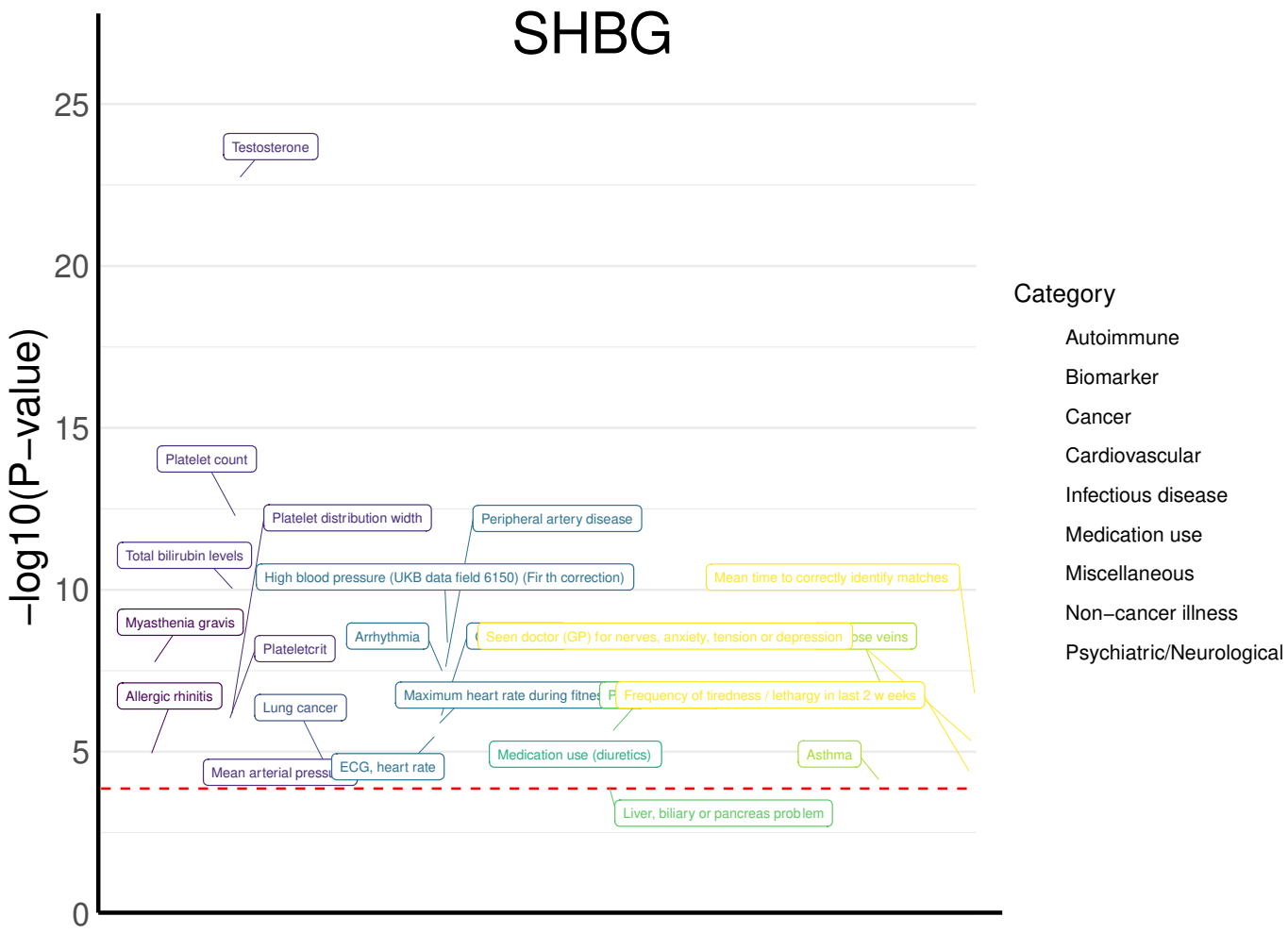


Figure AP4.17. Phe-MR results of circulating SHBG protein levels (UKB-PPP). *Results are the −log10 P-values (y-axis) for the main drug-target MR estimates (either inverse variance weighted or Wald ratios) for the SHBG protein on 366 outcomes curated for the Phe-MR analysis. The outcomes are grouped by clinical category. All cis-instrument MR estimate P-values are based upon used two-sided Wald tests. P-values were not adjusted for multiple comparisons. The red dashed lined is the Bonferroni-corrected P-value threshold (1.37×10^{-4} [0.05/366 outcomes]), and the labeled outcomes are those that surpassed the Bonferroni correction for multiple comparisons.*

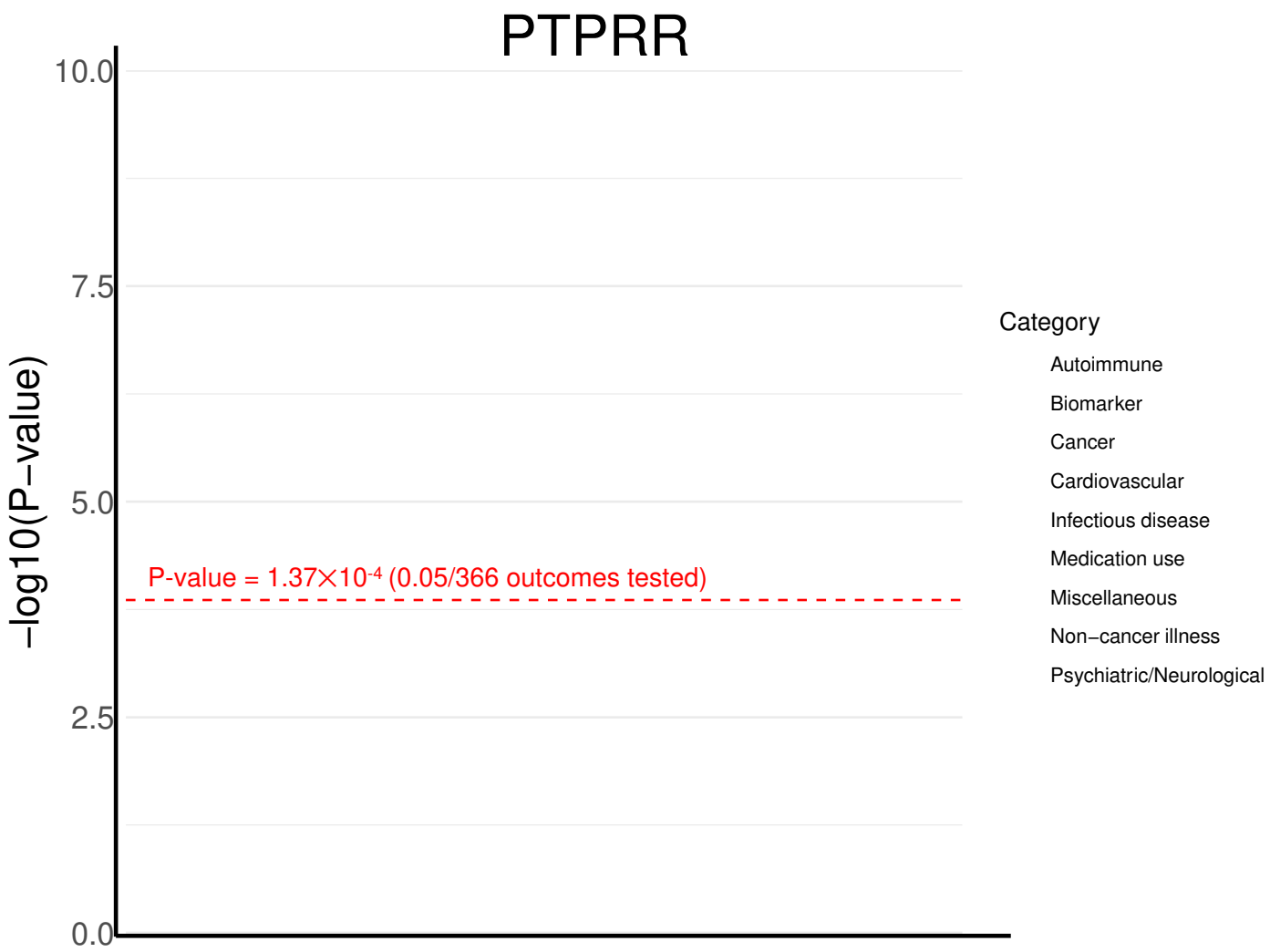


Figure AP4.18. Phe-MR results of circulating PTPRR protein levels (UKB-PPP). *Results are the −log10 P-values (y-axis) for the main drug-target MR estimates (either inverse variance weighted or Wald ratios) for the PTPRR protein on 366 outcomes curated for the Phe-MR analysis. The outcomes are grouped by clinical category. All cis-instrument MR estimate P-values are based upon used two-sided Wald tests. P-values were not adjusted for multiple comparisons. The red dashed lined is the Bonferroni-corrected P-value threshold (1.37×10^{-4} [0.05/366 outcomes]), and the labeled outcomes are those that surpassed the Bonferroni correction for multiple comparisons.*

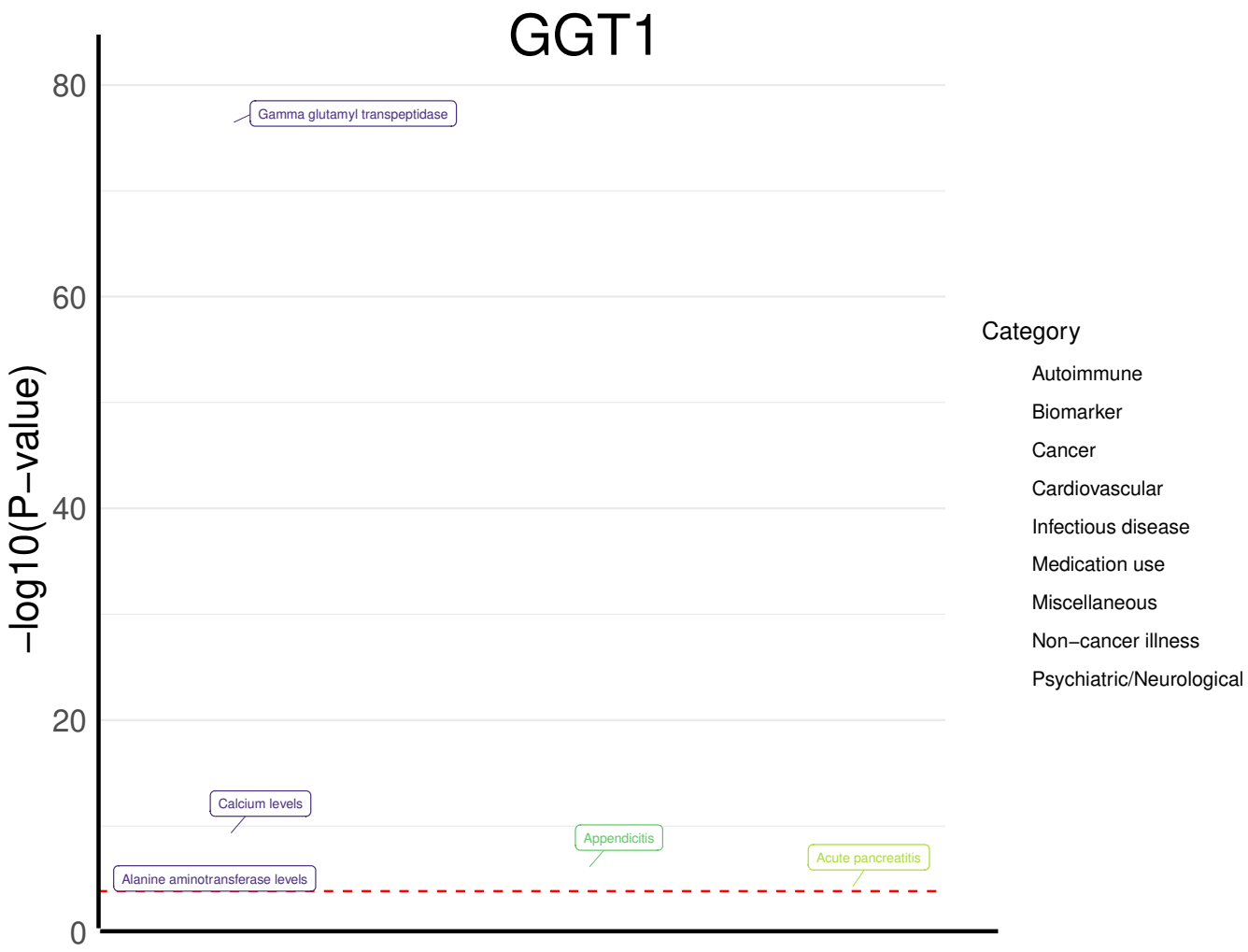


Figure AP4.19. Phe-MR results of circulating GGT1 protein levels (UKB-PPP). *Results are the −log10 P-values (y-axis) for the main drug-target MR estimates (either inverse variance weighted or Wald ratios) for the GTT1 protein on 366 outcomes curated for the Phe-MR analysis. The outcomes are grouped by clinical category. All cis-instrument MR estimate P-values are based upon used two-sided Wald tests. P-values were not adjusted for multiple comparisons. The red dashed lined is the Bonferroni-corrected P-value threshold (1.37×10^{-4} [0.05/366 outcomes]), and the labeled outcomes are those that surpassed the Bonferroni correction for multiple comparisons.*

AP4.5. Aim 3 Supplementary Tables

Below are titles to each for the Aim 3 Supplementary Tables that are formatted as Excel files uploaded separately.

Table AP5.1. Summary of NAFLD, T2D, CAD, and LPDFF phenotypes included in genomic SEM

Table AP5.2. Multivariate LD Score regression estimates of disease phenotypes included in GSEM analysis

Table AP5.3. LD Score regression genetic correlation estimates

Table AP5.4. Common factor model of NAFLD, T2D, and CAD using the GenomicSEM method

Table AP5.5. Genomic risk loci for lead SNPs

Table AP5.6. Annotation of the independent lead CM-Factor SNPs

Table AP5.7. Heterogenous genomic loci and the heterogenous lead SNPs for the CM-Factor

Table AP5.8. Qsnp alignment with univariate GWASs

Table AP5.9. CM-factor lead SNPs in proximity to lead SNPs in univariate GWASs

Table AP5.10. Look-up of CM-Factor lead SNPs in the GWAS Catalog performed October 20, 2024

Table AP5.11. CM-Factor genomic loci GWAS Catalog query summary of the number of traits/CM-Factor locus

Table AP5.12. GWAS Catalog query summary of CM-Factor loci and their genome-wide significant (P-value < 5e-8) pleiotropic associations in the GWAS Catalog

Table AP5.13. Multivariate LD Score regression and genetic correlation estimates for GBS analysis

Table AP5.14. BMI and non-BMI contributions to CM-factor (lead SNPs)

Table AP5.15. Waist-hip ratio and non-WHR contributions to CM-factor (lead SNPs)

Table AP5.16. Fine-mapping results for the CM-Factor loci

Table AP5.17. Credible set information for the fine-mapping results for CM-Factor

Table AP5.18. HaploReg annotation of CM-Factor fine-mapped SNPs

Table AP5.19. Full summary statistics of transcriptome-wide association study of *CM-Factor*

Table AP5.20. High-confidence TWAS genes associated with the CM-Factor (TWAS significant, evidence of shared causal variant, and conditionally significant) and comparison with MAGMA gene-based results

Table AP5.21. Novelty assessment for the high-confidence TWAS genes

Table AP5.22. GREP (Genome for REPositioning drugs) drug-gene enrichment analysis of the high-confidence TWAS genes using ICD classifications

Table AP5.23. GREP (Genome for REPositioning drugs) drug-gene enrichment analysis of the high-confidence TWAS genes using ATC classifications

Table AP5.24. Gene-drug transcriptomic signature matching results comparing the high confidence gene signature against the CMap Touchstone data

Table AP5.25. Gene-set enrichment of CM-Factor GWAS using the MAGMA gene-based results and the FUMA GSEA (Gene-Set Enrichment Analysis) method)

Table AP5.26. Partitioned heritability (S-LDSC) using chromatin ENCODE and Roadmap datasets

Table AP5.27. Partitioned heritability (S-LDEC) using gene expression from 205 tissue datasets

Table AP5.28. Cell-type enrichment using CELLECT and partitioned heritability gene prioritization method and the Tabula Muris dataset

Table AP5.29. Cell-type enrichment using CELLECT and MAGMA gene prioritization method and the Tabula Muris dataset

Table AP5.30. Antidiabetic drug target instruments

Table AP5.31. Lipid-modulating drug target instruments

Table AP5.32. NAFLD/NASH drug target instruments

Table AP5.33. Antihypertensive cis-instruments for drug-target MR analyses

Table AP5.34. Drug-target MR results of antidiabetic targets on the CM-Factor

Table AP5.35. Colocalization results of the drug targets with drug-target Mendelian randomization estimate P-values < 0.05

Table AP5.36. Results of Mendelian randomization analysis of lipid modulating drug targets on CM-factor

Table AP5.37. Drug-target MR results of NAFLD targets on the CM-Factor

Table AP5.38. Drug-target MR results of antihypertensive targets on the CM-Factor

Table AP5.39. Full drug-target MR results of druggable genes on the CM-Factor

Table AP5.40. Colocalization results of 91 druggable genes surpassing correction for multiple comparisons

Table AP5.41. Novelty assessment for 41 druggable genes identified by the screen of the druggable genome (drug-target MR + colocalization)

Table AP5.42. Credible set results of the druggable genes in the circulating eQTLGen data demonstrating evidence of colocalization with the CM-Factor

Table AP5.43. Tissue-specific top druggable gene instruments

Table AP5.44. Summary of tissue-level replication of 41 druggable genes prioritized by the initial screen using whole blood gene expression

Table AP5.45. Cis-instrument Mendelian randomization of top tissue-specific druggable genes on CM-factor

Table AP5.46. Druggable gene top hits colocalization

Table AP5.47. LDLink results prioritizing established cardiometabolic biomarkers in the EBI GWAS Catalog of druggable gene credible set SNPs (query performed on October 1st, 2024)

Table AP5.48. Druggable gene credible set SNP PIP > 0.5 biomarker/risk factor annotation in Oxford Biobank (all results with P-value < 0.05)

Table AP5.49. SVMR gene targets on outcome biomarkers

Table AP5.50. Two-step mediation for druggable genes with established cardiometabolic biomarkers

Table AP5.51. PheWAS for druggable genes prioritized by drug target screen on the CM-Factor

Table AP5.52. Step 1 MR of BMI onto circulating protein levels

Table AP5.53. Step 2 cis-MR of BMI-associated proteins onto CM-Factor

Table AP5.54. Mediation analyses of the BMI-associated proteins identified by the two-step MR

Table AP5.55. Two-sample MR Step 1 replication

Table AP5.56. Two-sample MR Step 2 replication (effects of BMI-associated proteins from the primary analyses on the CM-Factor)

Table AP5.57. Observational analyses of proteins associated with increased body mass index (BMI) from the INTERVAL cohort

Table AP5.58. Colocalization results (single causal variant assumption) of the BMI-associated proteins on the CM-Factor

Table AP5.59. Phenome-wide MR results of ENO3 protein levels

Table AP5.60. Phenome-wide MR results of SHBG protein levels

Table AP5.61. Phenome-wide MR result PTPRR protein levels

Table AP5.62. Phenome-wide MR result GGT1 protein levels

Appendix 5. Biorender Publication Licenses