CONTRIBUTIONS OF RARE GENETIC VARIANTS TO BIOMARKERS AND DISEASES

CONTRIBUTIONS OF RARE GENETIC VARIANTS TO BIOMARKERS AND DISEASES

By NAZIA PATHAN, B.Sc. (Hons.), MSc.

A Thesis Submitted to the School of Graduate Studies in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy

McMaster University DOCTOR OF PHILOSOPHY (2024) Hamilton, Ontario, Canada. (Medical Sciences Program- Department of Health Sciences)

TITLE: Contributions of rare genetic variants to biomarkers and diseases

AUTHOR: Nazia Pathan, BSc (University of Toronto), MSc (University of Toronto)

SUPERVISOR: Dr. Guillaume Paré, MD, MSc

NUMBER OF PAGES: 145

LAY ABSTRACT

This thesis explores the role of genetic mutations (found in less than 1% of the population) in vascular dementia (VaD) and 49 other traits using data from over 167,000 individuals in the UK. Two tools, RARity and RARity-β, were developed to measure how much the mutations cause traits, like height, circulating blood proteins, and diseases to vary among individuals.

Review of existing research on VaD emphasizes the complexity of the disease, and highlights the need for large-scale, collaborative research efforts. RARity and RARity-β showed that mutations have significant influence on measurable traits and diseases, with height being the most affected. These tools enabled discovery of genes linked to traits and assessment of current predictive tools to decide which mutations are detrimental to health.

Overall, the findings suggest that rare genetic mutations play a crucial role in human health and emphasizes the need for better predictive tools to identify detrimental mutations.

ABSTRACT

Most human genetic variants are rare (minor allele frequency, MAF <1%). This thesis investigates the significance of rare coding variants (RV), first with a literature review of vascular dementia (VaD), and subsequently in 31 continuous and 18 binary traits, utilizing whole exome sequence from the UK Biobank (N=167,348 and N=173,688, respectively). This was enabled with the development of the rare variant heritability (RARity) estimator and RARity-β.

Genetic determinants of VaD are explored through genome-wide association studies, polygenic risk scores, heritability estimates, and family studies. Complexity and heterogeneity of the disease are highlighted, emphasizing the need for large-scale collaborations and integromics approaches to enhance discoveries.

RARity estimates RV heritability ($\hat{h}^2_{RV}$) without assuming a specific genetic architecture. It revealed a significant loss of heritability (79%) due to gene-level RV aggregation. For 27 traits, $\hat{h}^2_{RV}$ exceeded 5%, with height showing the highest at 21.9%. VaD risk factors such as ApoA-I, BMI, blood pressure, LDL-cholesterol, and triglycerides had $\hat{h}^2_{RV}$ of 4.6% to 9.9%. RARity showed RVs as the source of "missing heritability", identified 11 new gene-phenotype associations using gene-level heritability estimates, and showed that current pathogenicity predictors do not adequately enrich for RVs contributing to trait variance, indicating a need for better predictive algorithms.

RARity-β estimates overall ($\hat{h}^2_{RV\text{-}liab}$) and gene-level heritability of binary traits on a liability scale. Significant $\hat{h}^2_{RV\text{-}liab}$ was found for hypothyroidism, asthma, hypercholesterolemia, and essential hypertension, identifying 77 genes with significant contributions to $\hat{h}^2_{RV\text{-}liab}$, including 70 new gene-trait relationships. The *PEPB1* gene's role in atrial fibrillation and the *TSHR* gene's link to hypothyroidism and sciatica are discussed. Results suggest that genes contributing significantly to $\hat{h}^2_{RV\text{-}liab}$ have functional consequences. Overall, this thesis provides novel methodologies and insights into the understanding of complex traits and diseases.

ACKNOWLEDGEMENTS

Bringing a new life into the world, shifting back and forth between virtual and in-person learning for both children and me, on top of the demands of a PhD, during a global pandemic was an experience like no other. This also makes me incredibly thankful and accomplished for making this small contribution to the world. I am thankful to my family for their unwavering support and understanding throughout this journey. Their constant encouragement and love have been my source of strength and motivation. I am forever indebted to my parents for instilling the value of resilience and knowledge, and for their countless sacrifices that have shaped me into the individual I am today.

I extend my heartfelt thanks to the members of my thesis committee, Dr. John Waye and Dr. Peter Kavsak for their insightful feedback, and dedication to excellence. Special acknowledgment goes to Dr. Kavsak, who not only served as my advisor during the comprehensive examination but also provided insights into biomarkers, particularly troponin. Additionally, Drs. John Waye, Marie Pigeyre, Michael Chong, Wei Deng and Jason Roberts have been incredibly supportive, and I deeply value their knowledge in molecular and epidemiological genetics, and well as their role in creating a conducive environment for academic inquiry and scholarly pursuits.

I highly appreciate all my colleagues and peers for their camaraderie, intellectual exchange, and moral support. Their friendship and encouragement have made the journey of pursuing a PhD more enjoyable and rewarding. Last, but most importantly, I extend my profound gratitude to my supervisor, Dr. Guillaume Paré for his guidance, encouragement, and support throughout the course of my PhD. His expertise, wisdom, and mentorship have been invaluable in shaping my research and academic growth, enabled me to pursue my academic goals and contribute to the advancement of knowledge in my field. Thank you!

Table of Contents

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

AD	Alzheimer's Disease
ADM	Adrenomedullin
APOB	Apolipoprotein-B
APOE	Apolipoprotein-E
APP	Amyloid Precursor Protein
BG	Basal Ganglia
BHR	Burden Heritability Regression
BMI	Body Mass Index
CAA	Cerebral Amyloid Angiopathy
CAD	Coronary Artery Disease
CADASIL	Cerebral Autosomal Dominant Arteriopathies with Aubcortical Infarcts and Leukoencephalopathy
CADD	Combined Annotation-Dependent Depletion
CARASAL	Cathepsin-A-related Arteriopathy with Strokes and Leukoencephalopathy
CARASIL	Cerebral Autosomal Recessive Arteriopathy with Subcortical Infarcts and Leukoencephalopathy
CBI	Covert (or silent) Brain Infarcts
CFH	Complement Factor H
CGRP	Calcitonin Gene-Related Peptide
CHD	Coronary Heart Disease
CI	Confidence Interval
CNS	Central Nervous System
CSF	Cerebrospinal Fluid
CSVD	Cerebral Small Vessel Diseases
CV	Common Variant
DNA	Deoxyribonucleic Acid
DPYD	Dihydropyrimidine Dehydrogenase
EMMAX	Efficient Mixed-Model Association eXpedited
FDR	False Discovery Rate
GCK	Glucokinase
GO	Gene Ontology
GPU	Graphics Processing Unit
GREML	Genetic Relatedness Estimation through Maximum Likelihood
GRS	Genetic Risk Score

GWAS	Genome-Wide Association Study
GWS	Genome-Wide Significant
HDL	High Density Lipoprotein
HR	Hazard Ratio
HUGO	Human Genome Organization
ICH	Intracerebral Hemorrhage
IS	Ischemic Stroke
LAMNA	Laminin Subunit Alpha
LD	Linkage Disequilibrium
LDL	Low Density Lipoprotein
LDLR	Low Density Lipoprotein Receptor
LDSC	LD Score Regression
LMM	Linear Mixed Model
LOEUF	Loss-of-function Observed/Expected Upper bound Fraction
LPA	Lipoprotein(A)
LPL	Lipoprotein Lipase
LRLD	Long-range LD
LSH	Liability Scale heritability
MAC	Minor allele count
MAF	Minor allele frequency
MAGMA	Multi-marker Analysis of GenoMic Annotation
MCAP	Mendelian Clinically Applicable Pathogenicity
MR	Mendelian Randomization
MRI	Magnetic Resonance Imaging
MZ	Monozygotic
NGS	Next Generation Sequencing
OLS	Ordinary Least Square
OR	Odds Ratio
PGS	Polygenic Score
PINT	P53 Induced Transcript
PLINK	A toolset for whole-genome association and population-based linkage analysis
POLMM	Proportional Odds Logistic Mixed Model
PPARA	Peroxisome Proliferator Activated Receptor Alpha
PRS	Polygenic Risk Score
PVS	Perivascular Spaces
QC	Quality Control

REVEL	Rare Exome Variant Ensemble Learner
RV	Rare variants
SAIGE	Scalable and Accurate Implementation of GEneralized mixed model
SCS	Set Count and Sizes
SD	Standard Deviation
SKAT	Sequence Kernel Association Test
SNP	Single Nucleotide Polymorphism
STRING	Database of protein-protein interactions
SVD	Small Vessel Disease
TFAM	Transcription Factor A, Mitochondrial
TSH	Thyroid Stimulating Hormone
TSHB	Thyroid Stimulating Hormone, beta
TSHR	Thyroid Stimulating Hormone Receptor
UCSC	University of Santa Cruz
UK	United Kingdom
UKB	UK Biobank
US	United States
VCAN	Versican
VCI	Vascular Cognitive Impairment
WES	Whole Exome Sequence
WGS	Whole Genome Sequence
WMH	White Matter Hyperintensities

DECLARATION OF ACADEMIC ACHIEVEMENT

FORMAT AND ORGANIZATION OF THESIS

This thesis is prepared in the "sandwich" format as outlined in the School of Graduate Studies' Guide for the Preparation of Theses. It includes a general introduction, an overview of hypotheses and objectives, three independent studies prepared in journal article format, and an overall discussion. The candidate is the first author on all manuscripts. At the time of thesis preparation, Chapter 3 and 4 were published, and 5 is in preparation for submission in peer-review journal.

CONTRIBUTION TO PAPERS WITH MULTIPLE AUTHORSHIP

CHAPTER 3 (STUDY1):

Pathan, N., Kharod, M. K., Nawab, S., Di Scipio, M., Paré, G., & Chong, M. (2024). Genetic Determinants of Vascular Dementia. The Canadian Journal of Cardiology, 40 (8), P1412-1423.

NP, MC and GP contributed to the conception and design of the study. **NP** was the principal writer of the manuscript. MC and GP facilitated project administration and supervision. All authors contributed to the drafting and revision of the final article. All authors approved the final submitted version of the manuscript.

CHAPTER 4 (STUDY2):

Pathan, N., Deng, W. Q., Di Scipio, M., Khan, M., Mao, S., Morton, R. W., Lali, R., Pigeyre, M., Chong, M. R., & Paré, G. (2024). A method to estimate the contribution of rare coding variants to complex trait heritability. Nature Communications, 15 (1), Article number: 1245.

N.P. conceived and implemented the project, performed computational analyses, analyzed results, and wrote the manuscript. G.P. conceived and supervised the overall project. W.D. and M.K. provided expert guidance and feedback on statistical analyses and writing. M.D.S. and S.M. performed the CV related computational analyses. M.P. and R.L. curated the list of medications for diabetes and lipids. M.R.C. performed primary QC of genotype data and R.W.M. created Fig. 1. All authors reviewed and edited the manuscript.

CHAPTER 5 (STUDY 3):

In preparation for submission, Pathan, N.; Chaudhari, A.; Feiner, J.; Parrot, N.; Paré, G. Rare variant heritability estimator to identify genes implicated in disease status in the UK Biobank.

NP and GP contributed to the conception and design of the study, and methodological development. **NP** conducted bioinformatic and/or statistical analyses. AC and NP assisted with secondary analyses. JF assisted with data acquisition. GP facilitated project administration and supervision. **NP** was the principal writer of the manuscript. All authors contributed to the drafting and revision of the final article.

CHAPTER 1:
INTRODUCTION

CHAPTER 1: INTRODUCTION

This chapter sets the stage for the analytical methodologies discussed in subsequent chapters. It introduces the fundamental principles of quantitative genetics, genotyping methodologies, association study frameworks, estimations of trait heritability, development of genetic risk scores, techniques for causal inference, and network and pathway analyses. By the end of this chapter, readers will have a solid foundation in these essential concepts, facilitating a deeper understanding of the material presented in the later chapters.

1.1 COMPLEX TRAITS GENETICS

Diseases caused by single genes, referred to as monogenic traits, such as cystic fibrosis, sickle-cell disease, and Huntington's disease, adhere to Mendelian mode of inheritance, constituting a relatively simple genetic etiology[1]. This category of human traits often serves as the primary focus for genetic testing and therapeutic interventions as they are easier to detect and more conventionally understood. In contrast, most human traits are complex in nature, being multifactorial and influenced by a myriad of genetic variants, and environmental factors. Such genetic variants are dispersed throughout the genome, across hundreds of genes and intergenic regions[2]. Examples of complex traits include height, circulating blood biomarker levels, BMI, diabetes, and cardiovascular diseases. A comprehensive understanding of the factors influencing these traits are crucial for the development of effective prevention strategies, diagnostic tools, and targeted treatments tailored to individual genetic profiles, ultimately advancing personalized medicine, and improving healthcare outcomes.

1.2 COMMON AND RARE VARIANTS IN POPULATION GENETICS

Over the past decade, technological advances with microarray technology, next generation sequencing, and large-scale biomarker measurements have enabled the establishment of large biobanks, such as Trans-Omics for Precision Medicine (TOPMed), UK Biobank, FinGenn, and All of US, with a range of genotype and phenotype data at the population level. This in turn has enabled advancements in state-of-the art genetic methodologies and novel discoveries. Traditionally, the genetic variations have been categorized broadly into two classes: common variants (CVs) and rare variants (RVs). In this thesis, variants with minor allele frequency (MAF) <1% are defined as RVs, and MAF>=1% as CVs. In the biological context, variations in a trait within a population may be a product of both common and rare variants, as well as the environmental factors and interactions between environmental factors and genetic variants[3]. There is increasing evidence that pathogenicity may be caused by either a single rare variant with large, monogenic effect, multiple rare variants with modest, oligogenic effects or accumulation of very low effect common variants[4-6]. The contributions from each of these factors to trait variation may vary widely, which necessitates application of appropriate and robust methodology for their study.

1.3 HIGH-THROUGHPUT GENOMIC TECHNOLOGIES

Microarray and next generation sequencing (NGS) are two main categories of genomic profiling technologies used in large population studies. Microarrays utilize a grid of fixed DNA probes to hybridize with target DNA, enabling high-throughput analysis of thousands of genes simultaneously[7]. This technique is relatively fast and cost-effective for large-scale studies but is limited to known sequences, with lower sensitivity and specificity compared to NGS. In contrast, NGS involves massively parallel sequencing, providing detailed nucleotide sequences of DNA. This technology boasts high sensitivity and specificity, capable of detecting rare, low-abundance sequences and discovering novel variants[8]. Although NGS requires more extensive data analysis, higher costs, and significant

computational resources, it offers comprehensive genomic insights. Therefore, NGS is generally the preferred method for thorough and accurate RV detection in scientific and clinical settings, especially for high-throughput sequencing needs.

1.4 STATISTICAL GENETICS

Statistical genetics is a field of study aimed at the development and application of analytical methods to derive inferences from genetic data. It involves the use of statistical models and algorithms to explain the genetic basis of phenotypic variations and thus identify the genetic architecture of a trait. The genetic architecture of a trait can be described in terms of genetic features, such as the magnitude and the effect of the genetic variations, minor allele frequencies, linkage disequilibrium, and evolutionary constraints on the variations in a trait within a population[9]. The genetic architecture of a trait not only informs on the sources of variation between individuals but has broader impacts on diagnosis through molecular and genetic testing, development of personalized treatments, and disease predictions. As such, statistical genetics forms the backbone of modern population genetic research. The following sections provide brief summaries of common statistical approaches that are amenable to quantifying the risk and statistical significance of putatively disease-causing variants. Applications of these methods are also described in chapter 3 in the context of understanding the genetics of vascular dementia, a highly complex and heterogeneous disease.

1.5 BRIEF OVERVIEW OF CURRENT GENETIC METHODS APPLIED TO COMMON VARIANTS

The microarray technology required to detect CVs was developed first and was more widely used as compared to the next generation sequencers (NGS). Consequently, the advancements in genetic methodologies are far more progressive for common variants compared to rare variants. Some of the widely used methods for studying the complex relationship between common variants and phenotypes are Genome wide association

studies (GWAS), SNP-heritability, polygenic scores (PGS), and Mendelian randomization (MR), each of which are described below.

1.5.1 GENOME-WIDE ASSOCIATION STUDY (GWAS)

GWAS are conducted by leveraging the abundance of high-density genetic markers throughout the genome and the principle of linkage disequilibrium (LD). The genetic markers are typically single nucleotide polymorphic (SNPs) variants, representing genetic loci that are in LD, either in the exonic or intronic regions of the genome[10]. Linkage disequilibrium occurs due to recombination between homologous chromosomes over many generations, such that the SNPs are highly correlated and inherited together as a unit. GWAS results in identification of genomic risk loci with statistical tests between the genetic markers and phenotypes, reporting blocks of correlated SNPs that all show a statistically significant association with the trait of interest. Ultimately, the genetic markers are prioritized based on association testing with a phenotype of interest. Since the test incorporates all SNPs in the genome, *vs* pre-specified candidate variants/ genes, it is characterized as the first hypothesis-free method of genetic studies.

The statistical methods for GWAS have evolved over time. The first generation GWAS relied on linear or logistic regression models using PLINK and SNPTEST; the second generation models such as GCTA-MLMA, GEMMA, EMMAX were dependent on mixed models to improve power, and account for population structure and relatedness; while the third generation was necessitated with large-scale biobanks (N>25,000) and includes linear mixed models using BOLT-LMM and Fast-GWA-LMM, as well as REGENIE, SAIGE and POLMM which utilize linear/logistic mixed models allowing for correction of case-control imbalance[11]. GWAS typically analyze each SNP-phenotype pair, typically corrected for covariates such as age, sex, population structure, and for multiple-testing using Bonferroni corrected *p-value* to avoid false positive results, i.e. for testing 1 million common variants in the human genome, significant SNPs have p-value $< 5\times10^{-8}$.

Since 2005, when the first GWAS identified a common risk allele in the *CFH* gene with large effect size for age-related macular degeneration[12], many additional genomic risk loci have been associated with diseases and traits. Some examples include the identification of the 9p21 loci associated with myocardial infarction, CAD, carotid atherosclerosis, stroke, aneurysms, congestive heart failure and CV mortality[13-15], *FTO* for obesity and *PTPN22* for autoimmune diseases[10]. As of May 2024, the GWAS catalog reports 10378 entries on the SNPs associated with cardiovascular disease (CVD) and 5498 entries on the SNPs associated with type-2 diabetes (T2D) mellitus, even after applying a stringent Bonferroni correction[16].

Despite the identification of numerous risk alleles, there are several limitations to GWAS. First, the coverage of variants in a GWAS is limited by the number of predefined sites on the microarray. It has been shown that most traits are affected by thousands of SNPs, each individually conferring low risk and often demonstrating pleiotropic effect on multiple traits[17]. Additionally, these risk loci are not necessarily located in the coding region and a strongly associated variant may not necessarily be causal; instead it may be in LD with a functionally relevant allele that is located in close proximity at this locus[18]. Together, these factors make it difficult to derive unambiguous biological meaning of the results.

In the post-GWAS era, several methods have been developed to apply, interpret, and identify the impact of the hundreds or thousands of SNPs associated with each trait by GWAS. These methods include SNP based heritability, polygenic risk scores (PRS), Mendelian randomization (MR) studies, and complex network and pathway analyses, each of these methods are described below [19].

1.5.2 HERITABILITY

The common proverb that human traits are the result of nature and nurture can be restated using statistical genetic terms: phenotypes are the result of genetic and environmental factors. The total phenotypic variance (Vp) in given population can be expressed as the sum of the variance due to genetic variance Vg and the environmental variance Ve:

$V\text{p} = V\text{g} + V\text{e}$

The genotypic variance Vg, can be further explained as being the sum of the additive (a), dominant (d), and epistatic (i) variance components:

$V\text{g} = V\text{a} + V\text{d} + V\text{i}$

However, research has shown that the effects of non-additive genetic effects (Vd + Vi) are minute, and thus broad sense heritability (H^2) quantifies the proportion of phenotypic variance explained by both genetic and environmental variance in a population[3].

$$H^2 = \frac{V_g}{V_p}$$

Conversely, narrow-sense heritability (h^2) can be defined as estimating only the effect of additive genetic variation on the phenotypic variations in a population. In the simplest form h^2 is defined as:

$$h^2 = \frac{V_a}{V_p}$$

Heritability ranges from 0, where genetic variation does not explain any of the phenotypic variations, to 1, meaning that genetics explains all the variations in the phenotype. In other

words, it informs how well a trait could be predicted from the genetics in a population. It is important to note that heritability is a characteristic of a population and not individuals, and the measurements are not constant over time. Different populations are likely to have different heritability estimates for the same trait due to differences in their genetics and environmental exposures, although some traits may present a stable heritability across species and populations. Homogenous populations are more likely to display larger heritability effect sizes as the phenotypic variation (*Vp*) is more precise.

Traditionally, heritability was measured in related individuals using family and twin studies, whereby researchers examine the resemblance between relatives (e.g., siblings, parents, offspring) with varying degrees of genetic relatedness. In this model, heritability was estimated by comparing the phenotypic similarity among relatives, such as the correlation of offspring and parental phenotypes, the correlation of full or half siblings, and the difference in the correlation of monozygotic (MZ) and dizygotic (DZ) twin pairs. For instance, if a trait shows greater similarity between biological relatives than unrelated individuals, or between MZ than DZ, it suggests a genetic component influencing that trait [20].

Genome-wide significant heritability emerged with the need to assess the total contributions of the SNPs identified in GWAS and led to the development of Genome-wide significant heritability ($\hat{h}^2_{GWS}$) [21]. With this method, it was observed that the estimated variance explained by genome-wide significant (GWS) SNPs discovered in GWAS was only a fraction of the estimated heritability from family or twin studies[22], an issue known as the problem of "missing heritability". For example, the $\hat{h}^2_{GWS}$ of human height was around 15%, compared to the $\hat{h}^2$ of 80% from family or twin studies[23-25]; similarly by 2016, over 700 variants identified for cardiovascular disease accounted for no more than 10% of the heritable risk[18]. This sparked concerns about the usefulness and cost-effectiveness of GWAS. Potential causes of missing heritability include exclusion of SNPs with small effects that do not meet the Bonferroni threshold, rare variants that are not captured by

GWAS, gene-by-gene interactions, gene-by-environment interactions, epigenetics, and perhaps, family studies with very few sample sizes had overestimated the heritability of these complex phenotypes.

Consequently, with the hypothesis that even the non-significant SNPs may contribute to the overall trait heritability, SNP-heritability ($\hat{h}^2_{SNP}$) was developed. SNP-heritability is the estimation of the variance explained by all SNPs used in a genome-wide association study (GWAS) in unrelated individuals, using either individual or summary level GWAS data. This has allowed estimation of the overall contributions of common variants to trait heritability, provided that all SNPs are represented on the genomic array. Through this effort, it was discovered that the SNP heritability is significantly larger than GWS heritability, such as in height where the estimates are 10-20% and 40-50%, respectively[25]. It was concluded that for complex traits there is a large number of common variants with effect sizes that are too small to pass the strict Bonferroni threshold, consistent with a model of polygenic inheritance. A conclusion that prompted establishment of large biobanks to capture variants of smaller effect sizes, and the development of PRS models.

Over the years, heritability for common variants have gone through several iterations in methodology, each with their own strengths and weaknesses, some of which are discussed in chapter 3.

1.5.3 POLYGENIC SCORE

GWAS and $\hat{h}^2_{SNP}$ offer compelling evidence supporting the polygenic architecture of complex traits, giving hope to precision medicine in the form of establishing Genetic risk scores (GRS), PRS for diseases, or PGS for other traits. The objective of these risk scores is to enable prediction of traits based on a person's genetic makeup and is the only approach that provides an estimate of genetic liability to a trait at the individual level. GRS is the cumulative risk estimates derived from the aggregation of risk contributions of GWS SNPs

towards a trait, thus, it is statistically defined as the sum of alleles weighted by their estimated effect sizes from GWAS summary statistics[26]. This is statistically represented as:

$$r_i = \sum_{j=1}^{k} \beta_j x_{ij}$$

Where r_i represents the risk; i, represents the individual number; j is the SNP number ranging from 1 to k; β_j is the weight for each SNP derived from the associated GWAS, and x_{ij} corresponds to the allele number for the j^{th} SNP of the i^{th} individual. GRS can be extended to PRS / PGS when non-GWS SNPs are included to reflect the polygenic nature of traits[27].

PRS may be particularly useful for complex diseases such as coronary artery disease, atrial fibrillation, type 2 diabetes, inflammatory bowel disease, and breast cancer, as they are shown to have risk factors similar to monogenic disorders[28]. Individuals with higher PRS was shown to increase the risk of early onset coronary artery disease by two-fold in 23% of the participants (N=7 of 30)[29]. In another study, patients with higher PRS for schizophrenia tended to have less improvement with antipsychotic drug treatment[30].

Over the years PRS have gone through many iterations in methodology to achieve the optimal level of predictability. This has involved varying the selection criteria for SNPs, weighting scheme, LD parameter selection, validation algorithms, and more. A comparison of these models, applied to psychiatric disorders is summarized by Ni, et al. (2021)[31]. Currently PRSs require further validation for clinical use, but are widely utilized in biomedical research, to examine shared etiology of phenotypes, assess clinical utility, and to compare experimental outcomes[26].

1.5.4 MENDELIAN RANDOMIZATION

Associations, or correlations between two variables, does not axiomatically denote causality. This is because associations are susceptible to unknown confounding factors influencing the exposure or outcome as well as reverse causality. To identify causality between the exposure, biomarkers and disease, the MR strategic framework may be employed. MR relies on genetic instruments exhibiting robust association with an exposure such that these instruments can act as proxies for the exposure. This approach is enabled by the fundamental principle that genotypes are not generally susceptible to reverse causation and confounding, due to their fixed nature and Mendel's First and Second Laws of Inheritance. In essence, the random assortment of alleles from parents to offspring effectively segregates participants into groups based on the presence or absence of specific genetic variants[32]. This process mirrors the randomization principle employed in randomized controlled trials (RCTs).

In order for a genetic variant to qualify as a valid instrument for causal inference in MR study, it must satisfy three core assumptions: 1) the genetic variant must be robustly associated with the exposure, as ensured by using genome-wide significance threshold for association in an independent sample ($P < 5 \times 10^{-8}$), 2) the genetic variant should not be associated with confounders of exposure-outcome relationships, and 3) the genetic-instrument should be associated with the outcome only through the exposure, i.e. no pleiotropic effects. However, even when these assumptions are fulfilled, limitations in MR include low power of studies, biases due to population stratifications, and weak instrumental biases. There are various tools to assess these assumptions and address the limitations[32]. The most basic MR method uses a single SNP instrument and is based on the following statistical model, implementing ratio coefficient method or Wald's ratio[33]:

$$Ratio\ estimate\ of\ causal\ effect = \frac{\hat{\beta}_{Y|Z}}{\hat{\beta}_{Z|X}}$$

Where $\hat{\beta}_{Y|Z}$ and $\hat{\beta}_{Z|X}$ are the slope estimates from the regressions of the outcome and exposure, respectively, on the SNP instrument, each of which can be obtained from GWAS summary data. Note that other methods are required for more complex issues, such when using multiple SNP instruments[34,35].

When appropriately implemented, MR is a powerful strategy to find causal inferences, and has been successfully utilized to validate drug targets. For example, MR has suggested causal association between LDL-C and the risk of CHD, mediated by SNPs in the *NPC1L1*, *HMGCR*, and *PCSK9* genes[36,37]. Furthermore, NPC1L1, HMGCR, and PCSK9 proteins are targets of the drugs ezetimibe, statins, and PCSK9-inhibitors, respectively, to reduce LDL-C levels, as confirmed with RCTs. Consequently, MR studies, especially when using previously published summary-level genetic association data, provide a rapid and affordable approach to evaluating causal questions, especially when RCTs cannot be implemented due to logistical or ethical reasons[38].

1.5.5 NETWORK AND PATHWAY ANALYSES

Network and pathway analyses help in understanding the biological implications of the observations in 'omics data and their interdependencies. Biological pathways are interactions among molecules in a cell that leads to a certain product or a change in a cell. Some curated databases of pathways include (Kyoto Encyclopedia of Genes and Genomes (KEGG), Reactome, WikiPathways, gene ontology (GO), transcription factor database (TRANSFAC) and miRTarBase[39]. Results from GWAS and other multi-omics data, such as exome-wide associations, mRNA/miRNA expression, protein expression, and DNA methylation. typically lead to a list of genes, which are sometimes prioritized according to a ranking system. The enrichment of these genes in the various pathways using tools such as g:profiler, GSEA, Enrichr and MAGMA provide mechanistic insights into the biological relevance of these genes[39,40]. When a gene or its' products are involved in multiple pathways in a context-specific manner, the result is a network. Networks are particularly

powerful as they allow visual mappings of many interactions across different data-types. The intuitive display for biological data in networks is established with the biological elements represented in nodes, which are linked to each other by directional or non-directional edges. This system allows automation of analyses, enables ease of interpretation, identification of hidden patterns and creation of emergent models to explain experimental observations. Currently, application programs are built to demonstrate networks from single-data types, such as GeneMania for gene-gene interactions and STRING for protein-protein interactions[39]. Additionally, platforms like Cytoscape have been built that integrate multiple application programs to visualize complex networks and integrate these with any type of attribute data targeting any specific case use[41].

1.6 STRATEGIES TO DETECT RARE VARIANTS

Compared to common variants, rare variant studies are a relatively new field. For a given effect size, as the allele frequency decreases the power to detect genes or variants of interest also decreases. Several platforms and methods have been used in the rare-variant associations with biomarkers and diseases, this includes extreme-phenotype sampling, family-based studies, GWAS chip and imputation, targeted region sequence, whole exome sequencing, and whole genome sequencing.

The extreme-phenotype sampling method is done by selecting participants with extreme traits, defined by a specific threshold. Extreme-phenotype sampling potentially enriches the samples with the presence of causal rare variants and thus requires a smaller sample size, for example, the *ABCA1*, *APOA1*, and *LCAT* genes have been related to low HDL-C, found by studying families with extreme high-density lipoprotein (HDL) phenotypes[42]. However, with this method, the outcome may not be generalized in the underlying population, it is sensitive to the outliers, it introduces a sampling bias, assumptions are made about the normal phenotypic characteristics of the specific condition under study, and also there is a reduced power to detect loci with smaller effects[43,44].

Family-based studies in which several family members share the same phenotype may provide more statistical power than regular population-based case-controls studies[45,46]. This type of study has been very successful in identifying large effect, highly penetrant, and monogenic Mendelian disorders[47]. In addition, a major benefit of family-based studies is the common genetic background of the studied subjects, which means that there are no issues related to population stratification. This was the method used to detect the role of PCSK9 mutation on autosomal dominant hypercholesterolemia[48] and most other mutations reported in Online Mendelian Inheritance in Man database (OMIM, https://www.omim.org/). There are several challenges and limitations to family based studies, including establishing and analyzing the complex and incomplete pedigree sequence, low sample size, a challenge of replication in large cohorts, inadequate capture the full spectrum of RVs present in the population, and considering the late-onset nature of many diseases it is also difficult to obtain genetic information from parents[49]. Put together, this makes it difficult to study complex traits with family-based studies. Nonetheless, as of May 2024, the OMIM reports over 7,528 of Mendelian diseases linked to rare variants, though their validity in the context of population genetics and replicability has been questioned. Consequently, efforts to examine these gene-disease associations in large biobanks and to curate them via efforts like the ClinGen Gene-Disease Clinical Validity Curation processes are underway[50].

Currently, exome-wide genotyping arrays are available that test thousands of exonic variants at a modest cost. Exome chips have been successful in identifying rare coding variants associated with numerous diseases, including insulin traits, liver disease and lipid levels[47]. GWAS based RV associations have also proved to be successful when followed by targeted region sequencing. Some examples include novel associations between rare variants in *APP* and Alzheimer's disease and between rare variants in *PDX1* and T2D[51,52]. Some of the most exciting discoveries arising from rare variant association with biomarkers in diseases were made from targeted gene sequencing experiments which identified rare coding variants with strong effects on phenotypic variation. These included genetic

variation in *ABCA1* and *PCSK9* associated with HDL and LDL-C levels, respectively[42,53]. The biggest drawback of using targeted region based platforms is that they are not as exhaustive as whole exome sequencing and will miss a large amount of very rare genetic variation[47].

Regardless of the platform used to study rare variants, the overarching challenge has been to move beyond handfuls of candidate gene studies to unbiased gene discovery.

1.7 OVERCOMING THE CURRENT LIMITATIONS OF RVS WITH WHOLE EXOME SEQUENCING

NGS has been used to mitigate the limitations of the above-mentioned methods in a large number of unrelated populations. NGS, with the ability to read every sequence with whole genome sequencing (WGS) or whole exome sequencing (WES) provides a much better opportunity to detect rare variant association[54]. Recently, sequencing studies have identified hundreds of genes containing rare coding variants, and these variants can have much larger effect sizes than CVs[55-58]. The exomes constitute 1%- 2% of the genome and house a majority of the rare protein-altering mutations, which are responsible for disrupting the function of the coded proteins and are therefore thought to be responsible for the majority of the deleterious phenotypes. This makes WES more cost-effective than WGS and better at studying rare variants. WES in human medicine also benefits from the availability of many large databases of single-nucleotide-variants (SNVs), known pathogenic variants, and control genomes[54]. These databases provide a wealth of information such as genotype quality, allele frequency, putative variant consequences, and pre-calculated algorithms to estimate likely pathogenicity of each genetic variant.

1.8 CHALLENGES SPECIFIC TO RVS

Studying RVs requires specialized statistical methods that can effectively handle the challenges posed by low variant frequencies, including low statistical power, biases due to uncorrected population stratification, cryptic relatedness, and the computational burden of testing a large number of variants simultaneously. Conventional methods designed for common variants may lack power or result in bias when applied to RVs. Despite these challenges, large-scale biobanks provide extensive genetic and phenotypic data, enabling researchers to conduct sophisticated statistical methods to identify and characterize RVs associated with complex traits.

1.9 METHODS TO STUDY RVS

1.9.1 VARIANT PRIORITIZATION

In the human genome, the number of rare variants is far greater than common variants, most of which are likely non-functional. Effective quality control steps reduce the number of variants, for example, by removing intronic, singletons, and common variants. This reduces the burden of multiple testing and single-to-noise ratio, thereby reducing false positive results or type-1 error. Prioritization of variants may also increase biological interpretability of variants. Annotation tools such as ANNOVAR[59], Variant Effect Predictor (VEP), and (Database for Nonsynonymous SNPs' Functional Predictions (dbNSFP)[60], utilize a wide array of publicly available databases to enable prioritization of variants. A fair number of pathogenicity scores has also been developed to prioritize variants, based on evolutionary constraints, predicted protein alterations, disease severity, experimentally measured regulatory effects and complex trait associations, or a combination of various factors. These methods usually leverage supervised or unsupervised machine learning techniques to enable assessment and prioritize variants. A few examples of pathogenicity scores include Combined Annotation Dependent Depletion (CADD)[61], Mendelian Clinically Applicable Pathogenicity (M-CAP) score[62], Rare Exome Variant

Ensemble Learner (REVEL)[63], Sorting Intolerant From Tolerant (SIFT)[64], Polymorphism Phenotyping (PolyPhen)[65], and Variant Annotation, Analysis & Search Tool (VAAST)[66].

Once RVs are selected, statistical methods are applied to identify variant-trait relationships. Commonly used statistical methods for studying RVs are centered around association testing. The most widely used methods are gene-burden tests and variations of SKAT, as described below.

1.9.2 GENE BURDEN

This is one of the earliest and simplest methods used in RV association testing. Under the autosomal dominant additive model, each variant per individual is scored as 0 for exhibiting two copies of the reference allele, 1 for heterozygous status with one reference and 1 alternative allele, and 2 for homozygous with both alternative alleles. The gene burden score is simply the sum of scores for all qualifying RVs within the gene and serves as the unit of association with a phenotype of interest. This effectively aggregates the variants to a single gene-score, thus reducing the burden of multiple testing and potentially increasing the power of association using a regression model. Most burden tests assume that the variants are either deleterious or protective and act in one direction. The weighted burden test is a variation of the gene-burden tests, where variants are weighted based on rarity and pathogenicity under additional assumptions about effect sizes[67].

1.9.3 SEQUENCE KERNEL ASSOCIATION TEST (SKAT) AND ITS VARIATIONS

SKAT aggregates information across multiple variants using a kernel matrix, allowing for more powerful detection of associations, especially when variants may have differing directions and magnitudes of effects. However, SKAT might be less powerful than burden tests in scenarios where most variants are causal and have effects in the same direction.

SKAT-O combines burden and SKAT methodologies to address SKATs limitations but might be slightly less powerful than burden or variance-component tests if their underlying assumptions are held. While SKAT and SKAT-O focus on association testing, gene-burden testing variant aggregation are broadly applicable in any situation that necessitates treating a gene as a unified entity[67,68].

1.9.4 OTHER STATISTICAL METHODS

Other statistical methods such as heritability, PRS and MR for rare variants are not as advanced as they are for CVs. In fact, currently there are no known published PRS of MR methods for RVs, and RV heritability has only recently been investigated in a few publications which will be explained in detail in relevant chapters 4 and 5.

1.10 HUMAN VARIABLES

1.10.1 TYPES OF VARIABLES

Identification and description of human variables used in a study is a necessary component in genetic research. Statistical analyses rely on the type of variables that are involved in the study. These variables can be categorized as either quantitative or qualitative.

Quantitative variables, which can be measured numerically, are further divided into continuous variables or discrete variables. Continuous variables can take any value within a range, such as anthropomorphic traits and blood biomarker measurements. Discrete variables represent specific values (e.g., number of hospital visits, cigarettes smoked per day, etc.). Qualitative variables, on the other hand, are non-numerical. These include nominal variables, which do not have a natural order (e.g., blood group, sex and disease status), and ordinal variables, which have an inherent order but undefined intervals between categories (e.g., disease severity stages, severity of pain and risk levels). Numerical values

assigned to different qualitative variables are useful solely for identification purposes, such as 1 for male and 2 for female. When a qualitative variable has only two categories, such as alive or dead, male or female, diabetic or non-diabetic, it is called a binary or dichotomous variable.

The nature of variables—whether categorical, continuous, or a mix—guides the selection of appropriate statistical tests and models. For example, categorical variables are often analyzed using frequency tables, chi-square tests, or logistic regression, while continuous variables can be analyzed with techniques like t-tests, ANOVA, correlation or linear regression.

1.10.2 BIOMARKERS, ANTHROPOMORPHIC TRAITS AND, DISEASES

Although understanding the genetic basis of disease status is important for direct diagnosis and prognosis of diseases, combining human genetics with biomarkers could help bridge the gap between the human genome and diseases[69], accelerate pre-clinical diagnosis, prognosis, improve disease subclass, identify and validate therapeutic targets[70,71], help predict long-term consequences of pharmacological intervention[72], enhance patient stratification[73], and enable repurposing of existing drugs[74]. Anthropomorphic traits (such BMI, waist-to-hip-ratio and blood pressure), blood proteins, lipids, glucose and hormones are some of the most important biomarkers of the state of human health, as they dictate the onset, severity and progression of diseases. A prime example of the benefits of gene-biomarker studies is the discovery of the relationship between proprotein convertase subtilisin/ kexin type-9 (*PCSK9*) gene and low-density lipoprotein cholesterol (LDL-C) [75]. LDL-C is a well-established biomarker for cardiovascular health, where higher levels of LDL are associated with an increased risk of coronary heart disease (CHD)[76]. A gain-of-function missense variant of *PCSK9* was associated with an autosomal dominant hypercholesterolemia[48]. Conversely, a loss-of-function variant in the *PCSK9* gene present in 2-4% of some ethnic populations, was shown to result in significantly lower levels of LDL cholesterol, and thus reduced the risk of CHD[77,78], these relationships were later

proven through functional studies[79,80]. This discovery of PCSK9's role in cholesterol metabolism led to the development of PCSK9 inhibitors, a new class of cholesterol-lowering drugs to effectively reduce LDL cholesterol levels sustainably over the long term and have shown promise in reducing the risk of CHD and incidence of heart attacks and other cardiovascular events[81,82]. Furthermore, genetic screening for *PCSK9* mutations can help stratify patients who would benefit most from PCSK9 inhibitors versus other lipid-lowering therapies[83].

In this thesis, I studied the contributions of genetic variants on dichotomous traits in the form of disease status (Chapters 3 and 5), as well as continuous variables, including anthropomorphic traits and blood biomarkers (Chapter 4).

1.11 REFERENCES

1 Jackson, M., Marks, L., May, G. H. W. & Wilson, J. B. The genetic basis of disease. *Essays Biochem* **62**, 643-723 (2018). https://doi.org/10.1042/EBC20170053

2 Lappalainen, T., Li, Y. I., Ramachandran, S. & Gusev, A. Genetic and molecular architecture of complex traits. *Cell* **187**, 1059-1075 (2024). https://doi.org/10.1016/j.cell.2024.01.023

3 Zhu, H. & Zhou, X. Statistical methods for SNP heritability estimation and partition: A review. *Comput Struct Biotechnol J* **18**, 1557-1568 (2020). https://doi.org/10.1016/j.csbj.2020.06.011

4 Wang, T., Zhao, P. A. & Eichler, E. E. Rare variants and the oligogenic architecture of autism. *Trends Genet* **38**, 895-903 (2022). https://doi.org/10.1016/j.tig.2022.03.009

5 Tadros, R. *et al.* Large scale genome-wide association analyses identify novel genetic loci and mechanisms in hypertrophic cardiomyopathy. *medRxiv* (2023). https://doi.org/10.1101/2023.01.28.23285147

6 Urpa, L. *et al.* Evidence for the additivity of rare and common variant burden throughout the spectrum of intellectual disability. *Eur J Hum Genet* **32**, 576-583 (2024). https://doi.org/10.1038/s41431-024-01581-3

7 Bumgarner, R. Overview of DNA microarrays: types, applications, and their future. *Curr Protoc Mol Biol* **Chapter 22**, Unit 22.21. (2013). https://doi.org/10.1002/0471142727.mb2201s101

8 Satam, H. *et al.* Next-Generation Sequencing Technology: Current Trends and Advancements. *Biology (Basel)* **12** (2023). https://doi.org/10.3390/biology12070997

9 Timpson, N. J., Greenwood, C. M. T., Soranzo, N., Lawson, D. J. & Richards, J. B. Genetic architecture: the shape of the genetic contribution to human traits and disease. *Nat Rev Genet* **19**, 110-124 (2018). https://doi.org/10.1038/nrg.2017.101

10 Uffelmann, E. *et al.* Genome-wide association studies. *Nature Reviews Methods Primers*, 1:59 (2021).

11 Mbatchou, J. *et al.* Computationally efficient whole-genome regression for quantitative and binary traits. *Nat Genet* **53**, 1097-1103 (2021). https://doi.org/10.1038/s41588-021-00870-7

12 Klein, R. J. *et al.* Complement factor H polymorphism in age-related macular degeneration. *Science* **308**, 385-389 (2005). https://doi.org/10.1126/science.1109557

13 Holdt, L. M. & Teupser, D. Recent studies of the human chromosome 9p21 locus, which is associated with atherosclerosis in human populations. *Arterioscler Thromb Vasc Biol* **32**, 196-206 (2012). https://doi.org/10.1161/ATVBAHA.111.232678

14 Anand, S. S. *et al.* A Digital Health Intervention to Lower Cardiovascular Risk: A Randomized Clinical Trial. *JAMA Cardiol* **1**, 601-606 (2016). https://doi.org/10.1001/jamacardio.2016.1035

15 Paquette, M. *et al.* The 9p21.3 locus and cardiovascular risk in familial hypercholesterolemia. *J Clin Lipidol* **11**, 406-412 (2017). https://doi.org/10.1016/j.jacl.2017.01.012

16 Sollis, E. *et al.* The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. *Nucleic Acids Res* **51**, D977-D985 (2023). https://doi.org/10.1093/nar/gkac1010

17 Watanabe, K. *et al.* A global overview of pleiotropy and genetic architecture in complex traits. *Nat Genet* **51**, 1339-1348 (2019). https://doi.org/10.1038/s41588-019-0481-0

18 Kessler, T., Vilne, B. & Schunkert, H. The impact of genome-wide association studies on the pathophysiology and therapy of cardiovascular disease. *EMBO Mol Med* **8**, 688-701 (2016). https://doi.org/10.15252/emmm.201506174

19 Adam, Y., Samtal, C., Brandenburg, J. T., Falola, O. & Adebiyi, E. Performing post-genome-wide association study analysis: overview, challenges and recommendations. *F1000Res* **10**, 1002 (2021). https://doi.org/10.12688/f1000research.53962.1

20 Mayhew, A. J. & Meyre, D. Assessing the Heritability of Complex Traits in Humans: Methodological Challenges and Opportunities. *Curr Genomics* **18**, 332-340 (2017). https://doi.org/10.2174/1389202918666170307161450

21 Yang, J., Zeng, J., Goddard, M. E., Wray, N. R. & Visscher, P. M. Concepts, estimation and interpretation of SNP-based heritability. *Nat Genet* **49**, 1304-1310 (2017). https://doi.org/10.1038/ng.3941

22 Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747-753 (2009). https://doi.org/10.1038/nature08494

23 Macgregor, S., Cornes, B. K., Martin, N. G. & Visscher, P. M. Bias, precision and heritability of self-reported and clinically measured height in Australian twins. *Hum Genet* **120**, 571-580 (2006). https://doi.org/10.1007/s00439-006-0240-z

24 Silventoinen, K. *et al.* Heritability of adult body height: a comparative study of twin cohorts in eight countries. *Twin Res* **6**, 399-408 (2003). https://doi.org/10.1375/136905203770326402

25 Wood, A. R. *et al.* Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat Genet* **46**, 1173-1186 (2014). https://doi.org/10.1038/ng.3097

26 Choi, S. W., Mak, T. S. & O'Reilly, P. F. Tutorial: a guide to performing polygenic risk score analyses. *Nat Protoc* **15**, 2759-2772 (2020). https://doi.org/10.1038/s41596-020-0353-1

27 Igo, R. P., Kinzy, T. G. & Cooke Bailey, J. N. Genetic Risk Scores. *Curr Protoc Hum Genet* **104**, e95 (2019). https://doi.org/10.1002/cphg.95

28 Khera, A. V. *et al.* Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat Genet* **50**, 1219-1224 (2018). https://doi.org/10.1038/s41588-018-0183-z

29 Thériault, S. *et al.* Polygenic Contribution in Individuals With Early-Onset Coronary Artery Disease. *Circ Genom Precis Med* **11**, e001849 (2018). https://doi.org/10.1161/CIRCGEN.117.001849

30 Zhang, J. P. *et al.* Schizophrenia Polygenic Risk Score as a Predictor of Antipsychotic Efficacy in First-Episode Psychosis. *Am J Psychiatry* **176**, 21-28 (2019). https://doi.org/10.1176/appi.ajp.2018.17121363

31 Ni, G. *et al.* A Comparison of Ten Polygenic Score Methods for Psychiatric Disorders Applied Across Multiple Cohorts. *Biol Psychiatry* **90**, 611-620 (2021). https://doi.org/10.1016/j.biopsych.2021.04.018

32 Zheng, J. *et al.* Recent Developments in Mendelian Randomization Studies. *Curr Epidemiol Rep* **4**, 330-345 (2017). https://doi.org/10.1007/s40471-017-0128-6

33 Burgess, S., Small, D. S. & Thompson, S. G. A review of instrumental variable estimators for Mendelian randomization. *Stat Methods Med Res* **26**, 2333-2355 (2017). https://doi.org/10.1177/0962280215597579

34 Boehm, F. J. & Zhou, X. Statistical methods for Mendelian randomization in genome-wide association studies: A review. *Comput Struct Biotechnol J* **20**, 2338-2351 (2022). https://doi.org/10.1016/j.csbj.2022.05.015

35 Sanderson, E. *et al.* Mendelian randomization. *Nat Rev Methods Primers* **2** (2022). https://doi.org/10.1038/s43586-021-00092-5

36 Ference, B. A., Majeed, F., Penumetcha, R., Flack, J. M. & Brook, R. 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: a 2 × 2 factorial Mendelian randomization study. *J Am Coll Cardiol* **65**, 1552-1561 (2015). https://doi.org/10.1016/j.jacc.2015.02.020

37 Navarese, E. P. *et al.* Effects of Proprotein Convertase Subtilisin/Kexin Type 9 Antibodies in Adults With Hypercholesterolemia: A Systematic Review and Meta-analysis. *Ann Intern Med* **163**, 40-51 (2015). https://doi.org/10.7326/M14-2957

38 Würtz, P. *et al.* Metabolomic Profiling of Statin Use and Genetic Inhibition of HMG-CoA Reductase. *J Am Coll Cardiol* **67**, 1200-1210 (2016). https://doi.org/10.1016/j.jacc.2015.12.060

39 Reimand, J. *et al.* Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap. *Nat Protoc* **14**, 482-517 (2019). https://doi.org/10.1038/s41596-018-0103-9

40 de Leeuw, C. A., Mooij, J. M., Heskes, T. & Posthuma, D. MAGMA: generalized gene-set analysis of GWAS data. *PLoS Comput Biol* **11**, e1004219 (2015). https://doi.org/10.1371/journal.pcbi.1004219

41 Barrio-Hernandez, I. & Beltrao, P. Network analysis of genome-wide association studies for drug target prioritisation. *Curr Opin Chem Biol* **71**, 102206 (2022). https://doi.org/10.1016/j.cbpa.2022.102206

42 Peloso, G. M. *et al.* Phenotypic extremes in rare variant study designs. *Eur J Hum Genet* **24**, 924-930 (2016). https://doi.org/10.1038/ejhg.2015.197

43 Lee, S., Abecasis, R., Gonçalo, Boehnke, M. & Lin, X. Rare-Variant Association Analysis: Study Designs and Statistical Tests. *The American Journal of Human Genetics* **95**, 5-23 (2014). https://doi.org/10.1016/j.ajhg.2014.06.009

44 Chen, W., Coombes, B. J. & Larson, N. B. Recent advances and challenges of rare variant association analysis in the biobank sequencing era. *Front Genet* **13**, 1014947 (2022). https://doi.org/10.3389/fgene.2022.1014947

45 Li, M., Boehnke, M. & Abecasis, G. R. Efficient study designs for test of genetic association using sibship data and unrelated cases and controls. *Am J Hum Genet* **78**, 778-792 (2006). https://doi.org/10.1086/503711

46 Kazma, R. & Bailey, J. N. Population-based and family-based designs to analyze rare variants in complex diseases. *Genet Epidemiol* **35 Suppl 1**, S41-47 (2011). https://doi.org/10.1002/gepi.20648

47 Auer, P. L. & Lettre, G. Rare variant association studies: considerations, challenges and opportunities. *Genome Med* **7**, 16 (2015). https://doi.org/10.1186/s13073-015-0138-2

48 Abifadel, M. *et al.* Mutations in PCSK9 cause autosomal dominant hypercholesterolemia. *Nat Genet* **34**, 154-156 (2003). https://doi.org/10.1038/ng1161

49 Fernández, M. V. *et al.* Evaluation of Gene-Based Family-Based Methods to Detect Novel Genes Associated With Familial Late Onset Alzheimer Disease. *Front Neurosci* **12**, 209 (2018). https://doi.org/10.3389/fnins.2018.00209

50 Rehm, H. L. *et al.* ClinGen--the Clinical Genome Resource. *N Engl J Med* **372**, 2235-2242 (2015). https://doi.org/10.1056/NEJMsr1406261

51 Steinthorsdottir, V. *et al.* Identification of low-frequency and rare sequence variants associated with elevated or reduced risk of type 2 diabetes. *Nat Genet* **46**, 294-298 (2014). https://doi.org/10.1038/ng.2882

52 Jonsson, T. *et al.* A mutation in APP protects against Alzheimer's disease and age-related cognitive decline. *Nature* **488**, 96-99 (2012). https://doi.org/10.1038/nature11283

53 Seidah, N. G. & Prat, A. The biology and therapeutic targeting of the proprotein convertases. *Nat Rev Drug Discov* **11**, 367-383 (2012). https://doi.org/10.1038/nrd3699

54 Warr, A. *et al.* Exome Sequencing: Current and Future Perspectives. *G3 (Bethesda)* **5**, 1543-1550 (2015). https://doi.org/10.1534/g3.115.018564

55 Sun, B. B. *et al.* Genetic associations of protein-coding variants in human disease. *Nature* **603**, 95-102 (2022). https://doi.org/10.1038/s41586-022-04394-w

56 Wang, Q. *et al.* Rare variant contribution to human disease in 281,104 UK Biobank exomes. *Nature* **597**, 527-532 (2021). https://doi.org/10.1038/s41586-021-03855-y

57 Backman, J. D. *et al.* Exome sequencing and analysis of 454,787 UK Biobank participants. *Nature* **599**, 628-634 (2021). https://doi.org/10.1038/s41586-021-04103-z

58 Karczewski, K. J. *et al.* Systematic single-variant and gene-based association testing of thousands of phenotypes in 394,841 UK Biobank exomes. *Cell Genom* **2**, 100168 (2022). https://doi.org/10.1016/j.xgen.2022.100168

59 Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* **38**, e164 (2010). https://doi.org/10.1093/nar/gkq603

60 Liu, X., Wu, C., Li, C. & Boerwinkle, E. dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs. *Hum Mutat* **37**, 235-241 (2016). https://doi.org/10.1002/humu.22932

61 Teotónio, H., Chelo, I. M., Bradić, M., Rose, M. R. & Long, A. D. Experimental evolution reveals natural selection on standing genetic variation. *Nat Genet* **41**, 251-257 (2009). https://doi.org/10.1038/ng.289

62 Jagadeesh, K. A. *et al.* M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. *Nat Genet* **48**, 1581-1586 (2016). https://doi.org/10.1038/ng.3703

63 Ioannidis, N. M. *et al.* REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. *Am J Hum Genet* **99**, 877-885 (2016). https://doi.org/10.1016/j.ajhg.2016.08.016

64 Ng, P. C. & Henikoff, S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res* **31**, 3812-3814 (2003). https://doi.org/10.1093/nar/gkg509

65 Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nat Methods* **7**, 248-249 (2010). https://doi.org/10.1038/nmeth0410-248

66 Hu, H. *et al.* VAAST 2.0: improved variant classification and disease-gene identification using a conservation-controlled amino acid substitution matrix. *Genet Epidemiol* **37**, 622-634 (2013). https://doi.org/10.1002/gepi.21743

67 Curtis, D. A weighted burden test using logistic regression for integrated analysis of sequence variants, copy number variants and polygenic risk score. *Eur J Hum Genet* **27**, 114-124 (2019). https://doi.org/10.1038/s41431-018-0272-6

68 Wu, M. C. *et al.* Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet* **89**, 82-93 (2011). https://doi.org/10.1016/j.ajhg.2011.05.029

69 Suhre, K., McCarthy, M. I. & Schwenk, J. M. Genetics meets proteomics: perspectives for large population-based studies. *Nat Rev Genet* **22**, 19-37 (2021). https://doi.org/10.1038/s41576-020-0268-2

70 Schmidt, A. F. *et al.* Genetic drug target validation using Mendelian randomisation. *Nat Commun* **11**, 3255 (2020). https://doi.org/10.1038/s41467-020-16969-0

71 Wang, R. C. & Wang, Z. Precision Medicine: Disease Subtyping and Tailored Treatment. *Cancers (Basel)* **15** (2023). https://doi.org/10.3390/cancers15153837

72 Nguyen, P. A., Born, D. A., Deaton, A. M., Nioi, P. & Ward, L. D. Phenotypes associated with genes encoding drug targets are predictive of clinical trial side effects. *Nat Commun* **10**, 1579 (2019). https://doi.org/10.1038/s41467-019-09407-3

73 Christiansen, M. K. *et al.* Polygenic Risk Score-Enhanced Risk Stratification of Coronary Artery Disease in Patients With Stable Chest Pain. *Circ Genom Precis Med* **14**, e003298 (2021). https://doi.org/10.1161/CIRCGEN.120.003298

74 Nabirotchkin, S. *et al.* Next-generation drug repurposing using human genetics and network biology. *Curr Opin Pharmacol* **51**, 78-92 (2020). https://doi.org/10.1016/j.coph.2019.12.004

75 Lagace, T. A. PCSK9 and LDLR degradation: regulatory mechanisms in circulation and in cells. *Curr Opin Lipidol* **25**, 387-393 (2014). https://doi.org/10.1097/MOL.0000000000000114

76 Abifadel, M. *et al.* Mutations and polymorphisms in the proprotein convertase subtilisin kexin 9 (PCSK9) gene in cholesterol metabolism and disease. *Hum Mutat* **30**, 520-529 (2009). https://doi.org/10.1002/humu.20882

77 Cohen, J. *et al.* Low LDL cholesterol in individuals of African descent resulting from frequent nonsense mutations in PCSK9. *Nat Genet* **37**, 161-165 (2005). https://doi.org/10.1038/ng1509

78 Cohen, J. C., Boerwinkle, E., Mosley, T. H. & Hobbs, H. H. Sequence variations in PCSK9, low LDL, and protection against coronary heart disease. *N Engl J Med* **354**, 1264-1272 (2006). https://doi.org/10.1056/NEJMoa054013

79 Lagace, T. A. *et al.* Secreted PCSK9 decreases the number of LDL receptors in hepatocytes and in livers of parabiotic mice. *J Clin Invest* **116**, 2995-3005 (2006). https://doi.org/10.1172/JCI29383

80 Zhang, D. W. *et al.* Binding of proprotein convertase subtilisin/kexin type 9 to epidermal growth factor-like repeat A of low density lipoprotein receptor decreases receptor recycling and increases degradation. *J Biol Chem* **282**, 18602-18612 (2007). https://doi.org/10.1074/jbc.M702027200

81 Kwon, H. J., Lagace, T. A., McNutt, M. C., Horton, J. D. & Deisenhofer, J. Molecular basis for LDL receptor recognition by PCSK9. *Proc Natl Acad Sci U S A* **105**, 1820-1825 (2008). https://doi.org/10.1073/pnas.0712064105

82 Chan, J. C. *et al.* A proprotein convertase subtilisin/kexin type 9 neutralizing antibody reduces serum cholesterol in mice and nonhuman primates. *Proc Natl Acad Sci U S A* **106**, 9820-9825 (2009). https://doi.org/10.1073/pnas.0903849106

83 Reay, W. R. & Cairns, M. J. Advancing the use of genome-wide association studies for drug repurposing. *Nat Rev Genet* **22**, 658-671 (2021). https://doi.org/10.1038/s41576-021-00387-z

CHAPTER 2:
GENERAL HYPOTHESIS, OBJECTIVE, RATIONALE, AND APPROACH

CHAPTER 2: HYPOTHESIS, OBJECTIVE, RATIONALE, & APPROACH

2.1 GENERAL HYPOTHESIS

I hypothesize that rare coding variants contribute significantly to complex trait heritability and can provide a hypothesis-free method to identifying gene-trait relationships.

2.2 GENERAL OBJECTIVE

The overall objective of this PhD thesis is to estimate the collective contributions of genetic rare variants to trait heritability and to prioritize genes based on gene-level rare variant heritability estimates.

2.3 RATIONALE AND APPROACH

Complex traits range from quantitative measurements of biomarkers to heterogeneous diseases that are challenging to diagnose. While clinical measurements of biomarkers provide a practical framework for diagnosis and treatment, genetics and biology reveal that diseases are often interconnected at a molecular level, challenging the rigid boundaries of clinical classifications. Furthermore, genetic studies that offer insights into biological mechanisms can potentially lead to the discovery of new diagnostic biomarkers that are more accessible and therapeutic targets that are more personalized.

In Chapter 3, using vascular dementia (VaD) as an example of a highly complex trait, genetic advancements in understanding the disease are discussed. The study aims to bridge the gap between genetic research and clinical practice by providing a comprehensive overview of the current knowledge on genetic markers, risk factors, and molecular pathways involved in VaD. This includes the genetics of cerebrovascular risk factors and associated diseases such as stroke, small vessel disease (SVD), and cerebral amyloid angiopathy (CAA). Consequently, this chapter offers a thorough examination of contemporary genetic methodologies employed in deciphering disease mechanisms and

diagnosis. Additionally, the review identifies challenges, opportunities, and future directions for research aimed at earlier diagnosis, targeted interventions, and therapeutic advancements in VaD.

Like nearly all complex traits, the most compelling evidence for genetic determinants of VaD comes from studies on common variants or rare variants with monogenic effects, often discovered through family studies. Limited data are available from twin and other epidemiological studies to estimate the heritability of VaD directly. Heritability results from contributions of common variants identified through GWAS indicate significant heritability of the pathological processes and risk factors underlying VaD[1,2]. However, the contributions of rare variants to the heritability estimates of VaD, its subtypes or risk factors, remain largely unknown, this is because of a lack in methodological advancements.

The vast majority of variants identified are rare, and yet their functional consequences are not known. For example, observations of the UKB WGS (N=150,119) reveal that out of the 710,913,648 variants, only 4.4% are common variants. In contrast, 43.9% are singletons (carried by a single sequenced individual) and 51.7% are rare variants (non-singleton, MAF<0.1%)[3]. Similar conclusion was also made by Taliun, et al. 2021 using the WGS from TopMed program [4].

Most known pathogenic variants are rare, and generally have large magnitudes of effects on traits. For example, RVs in *RB1*, *BRCA1/BRCA2*, *LMNA* and *NOTCH3* are well known predictors of retinoblastoma[5], breast and ovarian cancer[6], familial lipodystrophy[7,8], and Cerebral Autosomal Dominant Arteriopathies with Subcortical Infarcts and Leukoencephalopathy (CADASIL)[9], respectively.

Several algorithms have been developed to predict the pathogenicity of rare variants. While many of these algorithms perform well for Mendelian traits, as they were primarily built

and/or benchmarked using these traits, their effectiveness in the context of complex traits, which may be oligogenic in nature, still needs to be assessed.

Furthermore, pharmacogenomics has shown several common variants playing a vital role in drug efficacy and safety, some examples include *CYP2C9* involved in warfarin metabolism, *SLCO1B1* impacting statin uptake, and *DPYD* influencing the metabolism of fluoropyrimidine drugs, used in cancer treatment. There are a large number of putatively pathogenic RVs within these pharmacogenes, likely accounting for a substantial part of the unexplained inter-individual differences in drug metabolism phenotypes, however, the consequences on diseases and biomarkers remain largely unexplored[10].

As previously discussed, common variants do not account for all the heritability observed in family-based studies. This "missing heritability" may be explained by RVs[11], leading to the common-disease/rare-variant (CD-RV) hypothesis[12]. As per CD-RV hypothesis, common diseases may result from many RVs on multiple genes, independent of the CVs. This has been a motivation to examine the contributions of rare variants to common complex traits. Two recent methods to estimate rare variants are genomic residual maximum likelihood analysis (GREML) and burden heritability regression (BHR). RVs are shown to account for unexplained heritability in height and BMI[13] using the GREML method, however the estimates are based on both coding and non-coding variants using WGS, restricts to RVs with 0.01%< MAF, and also makes assumptions about MAF distributions of causal variants, which can result in biased estimates if the assumptions are violated[14]. BHR is based on aggregated allele scores, with estimates much smaller than the GREML method. A method to estimate the contributions of truly RVs in a fast and accurate manner, without prior assumptions, has been lacking, and this has been the primary focus of chapters 4 and 5.

Chapter 4 describes the development of the rare variant heritability estimator (RARity), a novel statistical approach to enable fast and accurate estimation of the contributions of rare

coding variants to complex, continuous trait heritability. Here we examine the effect of aggregation of variants into gene-burden on estimated trait heritability. RARity was established as an unbiased estimation of heritability for 31 complex traits, including 26 blood biomarkers and 5 anthropomorphic traits using WES from British and non-British Caucasian population in the UKB ($n = 167{,}348$). Additionally, gene-level heritability estimates are explored to discover gene-trait relationships and investigate the efficiency of pathogenicity algorithms. Finally, I investigated how WES data can help address the missing heritability question.

In chapter 5, RARity-β, an adaptation of RARity for binary trait heritability, is utilized to examine the contributions of RVs on disease status on a liability scale. With this I identified many canonical genes associated with diseases and delved into the common genetic pathways to identify secondary causes of diseases via genetics.

In summary, this thesis investigates the potential benefits of using rare variant heritability in human genetics.

2.4 REFERENCES

1 Hoggart, C. *et al.* Identification of novel locus associated with coronary artery aneurysms and validation of loci for susceptibility to Kawasaki disease. *Eur J Hum Genet* **29**, 1734-1744 (2021). https://doi.org/10.1038/s41431-021-00838-5

2 Markus, H. S. & Schmidt, R. Genetics of Vascular Cognitive Impairment. *Stroke* **50**, 765-772 (2019). https://doi.org/10.1161/STROKEAHA.118.020379

3 Halldorsson, B. V. *et al.* The sequences of 150,119 genomes in the UK Biobank. *Nature* **607**, 732-740 (2022). https://doi.org/10.1038/s41586-022-04965-x

4 Taliun, D. *et al.* Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **590**, 290-299 (2021). https://doi.org/10.1038/s41586-021-03205-y

5 Yun, J., Li, Y., Xu, C. T. & Pan, B. R. Epidemiology and Rb1 gene of retinoblastoma. *Int J Ophthalmol* **4**, 103-109 (2011). https://doi.org/10.3980/j.issn.2222-3959.2011.01.24

6 Mavaddat, N. *et al.* Cancer risks for BRCA1 and BRCA2 mutation carriers: results from prospective analysis of EMBRACE. *J Natl Cancer Inst* **105**, 812-822 (2013). https://doi.org/10.1093/jnci/djt095

7 Chen, W., Coombes, B. J. & Larson, N. B. Recent advances and challenges of rare variant association analysis in the biobank sequencing era. *Front Genet* **13**, 1014947 (2022). https://doi.org/10.3389/fgene.2022.1014947

8 Vigouroux, C. *et al.* Lipodystrophic syndromes due to LMNA mutations: recent developments on biomolecular aspects, pathophysiological hypotheses and therapeutic perspectives. *Nucleus* **9**, 235-248 (2018). https://doi.org/10.1080/19491034.2018.1456217

9 Grami, N. *et al.* Global Assessment of Mendelian Stroke Genetic Prevalence in 101 635 Individuals From 7 Ethnic Groups. *Stroke* **51**, 1290-1293 (2020). https://doi.org/10.1161/STROKEAHA.119.028840

10 Ingelman-Sundberg, M., Mkrtchian, S., Zhou, Y. & Lauschke, V. M. Integrating rare genetic variants into pharmacogenetic drug response predictions. *Hum Genomics* **12**, 26 (2018). https://doi.org/10.1186/s40246-018-0157-3

11 Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747-753 (2009). https://doi.org/10.1038/nature08494

12 Schork, N. J., Murray, S. S., Frazer, K. A. & Topol, E. J. Common vs. rare allele hypotheses for complex diseases. *Curr Opin Genet Dev* **19**, 212-219 (2009). https://doi.org/10.1016/j.gde.2009.04.010

13 Wainschtein, P. *et al.* Assessing the contribution of rare variants to complex trait heritability from whole-genome sequence data. *Nat Genet* **54**, 263-273 (2022). https://doi.org/10.1038/s41588-021-00997-7

14 Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet* **88**, 76-82 (2011). https://doi.org/10.1016/j.ajhg.2010.11.011

CHAPTER 3:
GENETIC DETERMINANTS OF VASCULAR DEMENTIA

Published in the *Canadian Journal of Cardiology.* (August 2024*)*

Canadian Journal of Cardiology 40 (2024) 1412–1423

Review

Genetic Determinants of Vascular Dementia

Nazia Pathan, MSc,[a,b] Muskaan Kaur Kharod, BSc,[a] Sajjha Nawab, BSc,[a] Matteo Di Scipio, BSc,[a,c] Guillaume Paré, MD, MSc, FRCPc,[a,b,c,d,e,‡] and Michael Chong, MSc, PhD[a,b,d,‡]

[a] *Population Health Research Institute, David Braley Cardiac, Vascular and Stroke Research Institute, Hamilton Health Sciences and McMaster University, Hamilton, Ontario, Canada*

[b] *Department of Pathology and Molecular Medicine, McMaster University, Michael G. DeGroote School of Medicine, Hamilton, Ontario, Canada*

[c] *Department of Medicine, Faculty of Health Sciences, McMaster University, Hamilton, Ontario, Canada*

[d] *Thrombosis and Atherosclerosis Research Institute, David Braley Cardiac, Vascular and Stroke Research Institute, Hamilton, Ontario, Canada*

[e] *Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, Ontario, Canada*

ABSTRACT

Vascular dementia (VaD) is a prevalent form of cognitive impairment with underlying vascular etiology. In this review, we examine recent genetic advancements in our understanding of VaD, encompassing a range of methodologies including genome-wide association studies, polygenic risk scores, heritability estimates, and family studies for monogenic disorders revealing the complex and heterogeneous nature of the disease. We report well known genetic associations and highlight potential pathways and mechanisms implicated in VaD and its pathological risk factors, including stroke, cerebral small vessel disease, and cerebral amyloid angiopathy. Moreover, we discuss important modifiable risk factors such as hypertension, diabetes, and dyslipidemia, emphasizing the importance of a multifactorial approach in prevention, treatment, and understanding the genetic basis of VaD. Last, we outline several areas of scientific advancements to improve clinical care, highlighting that large-scale collaborative efforts, together with an integromics approach can enhance the robustness of genetic discoveries. Indeed, understanding the genetics of VaD and its pathophysiological risk factors hold the potential to redefine VaD on the basis of molecular mechanisms and to generate novel diagnostic, prognostic, and therapeutic tools.

RÉSUMÉ

La démence vasculaire (DVa) est une forme prévalente de déficience cognitive avec une étiologie vasculaire sous-jacente. Dans cette revue de littérature, nous examinons les avancées génétiques récentes dans notre compréhension de la Dva, englobant une gamme de méthodologies comprenant des études d'association pangénomiques, des scores de risque polygénique, des estimations de l'héritabilité et des études familiales pour les troubles monogéniques, révélant la nature complexe et hétérogène de la maladie. Nous rapportons des associations génétiques bien connues et mettons en évidence des voies et mécanismes potentiels impliqués dans la DVa et ses facteurs de risque pathologiques, notamment l'accident vasculaire cérébral, la maladie des petits vaisseaux cérébraux et l'angiopathie amyloïde cérébrale. En outre, nous discutons d'importants facteurs de risque modifiables tels que l'hypertension, le diabète et la dyslipidémie, en soulignant l'importance d'une approche multifactorielle dans la prévention, le traitement et la compréhension de la base génétique de la DVa. Enfin, nous décrivons plusieurs domaines d'avancées scientifiques pour améliorer les soins cliniques, en soulignant que des efforts de collaboration à grande échelle, associés à une approche intégromique, peuvent améliorer la robustesse des découvertes génétiques. En effet, la compréhension de la génétique de la DVa et de ses facteurs de risque physiopathologiques pourrait permettre de redéfinir la DVa sur la base de mécanismes moléculaires et de créer de nouveaux outils diagnostiques, pronostiques et thérapeutiques.

Vascular dementia (VaD) is a severe form of vascular cognitive impairment (VCI), defined as having clinically significant deficits in at least 1 cognitive domain, along with pronounced disruption in activities of daily living.[1,2] After Alzheimer disease (AD), VaD is the second most common major neurocognitive disorder. In North America and Europe, VaD accounts for approximately 15%-20% of all clinically diagnosed neurocognitive diseases, and the incidence might be higher in Asia and developing countries.[3,4]

There are 4 main subtypes of VaD: (1) poststroke dementia, in which irreversible cognitive decline occurs within 6

Received for publication February 5, 2024. Accepted March 29, 2024.

‡These authors are co-senior authors.

Corresponding author: Guillaume Paré, Clinical Research Laboratory & Biobank — Genetic & Molecular Epidemiology Laboratory (CRLB-GMEL), 237 Barton Street East, Hamilton, Ontario L8L2X2, Canada. Tel.: +1-905-527-4322 ×40356.

E-mail: pareg@mcmaster.ca

Corresponding author: Michael Chong, CRLB-GMEL, 237 Barton Street East, Hamilton, Ontario L8L2X2, Canada. Tel.: +1-905-527-4322 ×41091.

E-mail: michael.chong@crlb-gmel.ca

See page 1420 for disclosure information.

https://doi.org/10.1016/j.cjca.2024.03.025

months of a stroke; (2) subcortical ischemic VaD driven by stenosis and occlusion of small vessels; (3) multi-infarct (cortical) dementia characterized by multiple "silent" strokes and white matter hyperintensities (WMH); and (4) mixed dementia, representing a combination of vascular and neurodegenerative disease.[1,2] Magnetic resonance imaging (MRI) is a "gold-standard" requirement for a definitive clinical diagnosis of VaD. However, a diagnosis of "probable" or "possible" VaD is assigned in the presence or absence of computed tomography evidence, respectively.[2,5]

In statistical genetics, heritability refers to the proportion of phenotypic variation in a trait that can be attributed to genetic factors within a population. Limited data are available from twin and other epidemiological studies to directly estimate the heritability of VaD. Nevertheless, a compelling line of evidence from epidemiological and genetic studies indicate a significant heritability of the pathological processes and risk factors underlying VaD. Genome-wide association studies (GWAS) suggest heritability estimates range from 11.9% for cerebral amyloid angiopathy (CAA) to 40% for stroke,[6,7] whereas twin and family studies propose heritability estimates for WMH lesion volume between 50% and 80%.[6,7] However, the pathological processes and risk factors related to VaD are highly heterogeneous, and pose significant challenges in diagnosis and management.

As clinicians encounter an increasing prevalence of VaD, because of an aging population, a nuanced understanding of the genetic underpinnings of this disorder become essential in more effective and personalized patient care. In this review, we aim to bridge the gap between genetic research and clinical practice, by providing a comprehensive overview of the current state of knowledge regarding the genetic markers, risk factors, and molecular pathways that contribute to its onset and pathogenesis. This review incorporates the genetics of the cerebrovascular risk factors and diseases associated with VaD, including stroke, small vessel disease (SVD), and CAA[8] (Fig. 1). This inclusive approach aims to establish an up-to-date overview of the genetic determinants of VaD, ensuring the inclusion of well known genes associated with VaD. In addition, we identify the challenges, opportunities, and future directions for research endeavours toward earlier diagnosis, targeted interventions, and therapeutic advancements.

Heritable Risk Factors for VaD

The high degree of heterogeneity in the clinical presentation can be attributed, in part, to individual baseline cognitive capacities and their lifetime exposure to known vascular risk factors. Major vascular risk factors linked to VaD include: blood pressure, glucose levels, lipid profiles, atrial fibrillation status, diet, body mass index, physical activity levels, and smoking status.[4,9] All of these risk factors have varying levels of genetic heritability, whereby variation in the risk factors are influenced by variations in population genetics. These risk factors range from moderate (5%) to highly (66%) heritable.[10,11] Blood pressure, type 2 diabetes, low-density lipoprotein (LDL) cholesterol, triglycerides, atrial fibrillation, diet, body mass index, exercise, and smoking are epidemiologically associated with greater VaD risk and have significant estimated heritabilites at 15%, 66%, 8.3%, 21.8%, 14.4%, 4.8%, 24.9%, 7%, and 15.1%, respectively.[10,11]

Genetic Epidemiology Associations With VaD

The search for common variants underlying these strong heritability signals has been facilitated by GWAS. Powered by microarray technology, GWAS systematically and agnostically investigate millions of polymorphisms for association with disease. A recent study conducted the largest known GWAS to date on VaD, which meta-analyzed 293,544 individuals from 12 diverse cohorts (with 2935 cases of VaD) and discovered 5 novel loci, in addition to the *Apolipoprotein E* (*APOE*) locus. One locus (*ASTN2*) was previously associated with hippocampal volume, verbal memory, and cerebrospinal fluid amyloid levels; the other nearby genes were associated with hypertension, diabetes, and dyslipidemias.[12]

The unique combination of disease-associated variants from GWAS enables a quantitative measurement of an individual's genetic predisposition to diseases, in the form of polygenic risk scores (PRS). In a study involving 198,965 initially dementia-free participants aged 60 years or older, genetically determined hypertension was associated with a 19% increased risk of dementia over a 15-year follow-up period (hazard ratio [HR], 1.18; 95% confidence interval [CI], 1.11-1.27). Individuals with higher PRS for all-cause dementia risk exhibited greater risk difference in dementia incidence in hypertensive vs normotensive individuals.[13] Furthermore, PRS of type 2 diabetes, fasting glucose, fasting insulin, and hemoglobin A1c are linked to a higher risk of all-cause dementia, particularly in individuals with VaD, suggesting that a subset of individuals with type 2 diabetes, because of their genetic makeup, are more susceptible to developing diabetes-associated dementia.[14] Another study leveraging PRS and social risk factors showed that each additional social risk score is associated with a 24% increased risk of VaD, with lower incidence rates observed among participants with high genetic risk and low social risk compared to those with high social risk.[15] To date, emerging PRS are continually improving the risk prediction performance for all-cause dementia and their subtypes, in addition to the risk conferred by *APOE* genotypes alone.[16]

Another valuable genetic epidemiological method to assess causal relationships of risk factors and disease is Mendelian randomization (MR). MR uses genetic variants as instrumental variables to infer causation, by taking advantage of the random distribution of exposure-influencing alleles. An MR study conducted from the US Veterans Affairs Million Veteran program (N = 334,672) showed evidence of causality of diabetes with VaD in non-Hispanic white (odds ratio [OR], 1.11; 95% CI, 1.07-1.15) and in non-Hispanic black participants (OR, 1.11; 95% CI, 1.04-1.19).[17] This same causal association for type 2 diabetes and VaD was also found in 213,370 Danish individuals with MR (OR, 1.53; 95% CI, 1.48-1.59).[18] For dyslipidemias, some MR evidence shows that lowering LDL-cholesterol levels by statin use might be effective at reducing VaD risk.[19] Continual improvements in polygenic quantification of VaD risk will further our diagnostic yield and interventional capacities in this form of dementia. Although the genetics of heritable vascular risk factors contribute to VaD, a deeper understanding of the genetic determinants of pathological risk factors, namely, stroke, SVD, and CAA, is crucial for comprehensively assessing VaD pathogenesis.

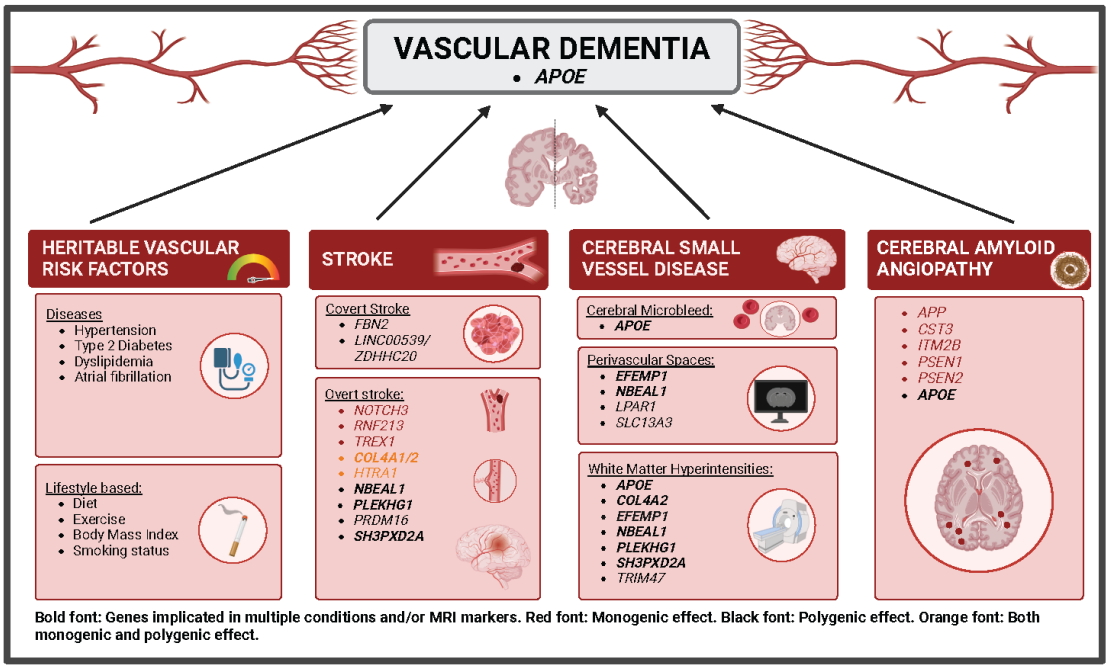


Figure 1. Etiological factors and key genetic markers associated with vascular dementia. Only genes that are most frequently implicated in the diseases are shown. MRI, magnetic resonance imaging. Created with BioRender.com.

Stroke and VaD

Stroke, an acute neurological deficit lasting more than 24 hours, caused by an ischemic blockage or hemorrhagic rupture, often precipitates VaD. Poststroke dementia, defined as a diagnosis of dementia within 6 months of stroke, is one of the main subtypes of VaD.[1] Even beyond this 6-month window, ischemic stroke (IS) dramatically increases the risk of being diagnosed with dementia. Indeed, after a stroke, cognition only partially recovers in the short-term because there is an unfortunate "turning point" at approximately 12 months when global cognition and many of its constituents (attention and processing speed, memory, language, perceptual motor function) begin a course of steady decline.[20] In the Atherosclerosis Risk in Communities study (N = 15,792), dementia incidence was > 4 times higher in stroke patients, with dementia diagnoses occurring an average of 7 years after the initial stroke event.[21] Furthermore, a dose-response relationship of stroke severity (as well as stroke recurrence) with subsequent risk of dementia was observed. Even after adjustment for vascular risk factors, HRs for incident dementia in individuals with 1 minor to mild stroke, 1 moderate to severe stroke, 2 or more minor to mild strokes, and 2 or more moderate to severe strokes were 1.76, 3.47, 3.48, and 6.68, respectively.[21] Ultimately, stroke is a major cause and risk factor for VaD, and therefore, understanding the genetic determinants of stroke is inextricable to understanding the genetic etiology of VaD.

Rare monogenic causes of VaD—not so rare anymore?

The genetic basis for VaD is evident from early family studies focused on rare hereditary forms of early-onset stroke. The most extensively studied monogenic stroke disorder, initially termed, "chronic familial vascular encephalopathy" in 1977,[22] is now recognized as cerebral autosomal dominant arteriopathies with subcortical infarcts and leukoencephalopathy (CADASIL). CADASIL is characterized by recurrent strokes (ischemic and hemorrhagic), migraine with aura, mood disorders, depression, epilepsy, and dementia. CADASIL patients exhibit SVD markers on MRI scans including lacunar infarcts, cerebral microbleeds (CMBs), and WMH, typically affecting the external capsule and anterior temporal lobe. The condition results from cysteine-altering mutations in the *NOTCH3* gene. These mutations disrupt normal formation of disulfide bridges in the *notch3* protein, destabilizing the protein, and leading to the accumulation of protein aggregates in cerebral vessel walls. This accumulation damages smooth muscle cells, pericytes, and endothelial cells, compromising vascular integrity and causing progressive thickening of cerebral vessel walls, ultimately leading to cerebrovascular insufficiency. CADASIL is among one of several monogenic conditions resulting in increased risk of stroke and VaD (Table 1).

Although monogenic causes of stroke were discovered by studying rare families with a preponderance of stroke, large-scale sequencing initiatives have revealed a higher prevalence

Table 1. Monogenic causes of stroke and vascular dementia[24,26,33,112]

Gene	Disease	Inheritance	Pathogenic mutation carrier prevalence*	Stroke risk (OR)	Dementia risk (OR)
NOTCH3	CADASIL	AD	0.09%-1.2%	2.65 (IS) 2.42 (ICH)	2.26 (All-cause) 5.42 (Vascular)
HTRA1	CARASIL	AR	N/A	N/A	N/A
	CADASIL2	AD	0.03%-0.08%	2.01 (IS)	2.17 (All-cause) 2.49 (Vascular)
COL4A1, *COL4A2*, *COLGALT1*	Collagen IV-related brain small vessel disease	AD	0.07%-0.26%	3.56 (ICH)	N/A
		AD	< 0.01%	N/A	N/A
		AR	< 0.01%	N/A	N/A
CTSA	CARASAL	AD	< 0.01%	N/A	N/A
TREX1	Retinal vasculopathy with cerebral leukodystrophy	AD	< 0.01%	N/A	N/A
GLA	Fabry disease	X-linked	0-0.02%	N/A	N/A
CECR1	ADA2 deficiency	AR	0.05%-0.18%	N/A	N/A
ABCC6	Pseudoxanthoma elasticum	AR	0.25%-0.76%	4.9 (IS)	N/A
COL3A1	Vascular Ehlers-Danlos syndrome (type IV)	AD	0-0.02%	N/A	N/A
KRIT1	Familial cerebral cavernous malformations	AD	0-0.03%	N/A	N/A
RNF213	Moyamoya syndrome	AD	0-0.78%	3.58 (LAA) 1.91 (IS)	N/A

AD, autosomal dominant; AR, autosomal recessive; CADASIL, cerebral autosomal dominant arteriopathies with subcortical infarcts and leukoencephalopathy; CARASAL, cathepsin A-related arteriopathy with strokes and leukoencephalopathy; CARASIL, cerebral autosomal recessive arteriopathies with subcortical infarcts and leukoencephalopathy; ICH, intracerebral hemorrhage; IS, ischemic stroke; LAA, large artery atherosclerosis; N/A, not applicable; OR, odds ratio.

*Prevalence range in the Genome Aggregation Database superpopulations (non-Finnish European, Latin American, East Asian, South Asian, or African).

of pathogenic mutations in the general population than traditionally diagnosed. For instance, although clinically diagnosed CADASIL is extremely rare (2 cases per 100,000 individuals), *NOTCH3* pathogenic mutation carriers are > 100 times more prevalent in the general population, ranging from approximately 2 carriers per 1000 individuals of European descent to as high as approximately 10 carriers per 1000 individuals of East Asian or South Asian descent.[23-25] Despite increased mutation prevalence, carriers are still at substantial risk of stroke (OR, approximately 2-4) and VaD (OR, approximately 5).[26]

Similarly, more prominent roles for other Mendelian stroke genes in sporadic stroke susceptibility have recently emerged.[27-33] Classically, biallelic loss of function mutations in *HTRA1* cause a severe recessive form of SVD called "severe recessive form of small vessel disease," with heterozygote carriers initially considered asymptomatic.[34] However, it is now known that heterozygote carriers experience a milder, later-onset form of SVD,[28,35] expanding the proportion of individuals with symptomatic SVD attributable to *HTRA1* mutations. Indeed, *HTRA1* carriers are relatively common (1 carrier per 1000) and have a twofold increased risk of stroke and VaD.[26] Another example is the *RNF213* p.R4810K mutation causing Moyamoya syndrome. Even in the absence of Moyamoya vasculopathy, this variant still confers substantial risk for large artery atherosclerosis (OR, 3.58), strokes of undetermined etiology, and strokes of other determined etiology.[33,36] Remarkably, this mutation is found in 20%-30% of Japanese patients with intracranial artery stenosis/occlusion and approximately 2% of all East Asian individuals.[37]

The observation of a higher-than-expected mutation frequency prompted us to systematically reassess the genetic prevalence of rare pathogenic mutations across all known Mendelian stroke genes.[24] In a public sequencing repository of 101,635 ethnically diverse individuals, we estimated the total carrier prevalence of Mendelian stroke mutations to be approximately 1%-3% depending on genetic ancestry.[24] This challenges the perception of monogenic causes of stroke as extremely rare disorders, highlighting a prevalent subgroup at extreme risk for VaD. The disparity between genetic and clinical prevalence suggests reason for optimism, because carrying a "pathogenic" mutation, rather than deterministic, is a strong risk factor for SVD and dementia. Contemporary research focuses on identifying disease modifiers including genetic and nongenetic factors. For example, approximately 5% of the variance in WMH severity among CADASIL patients could be explained by an independent common genetic predisposition for WMH.[38] Finally, conventional risk factors, such as hypertension and smoking, remain vital, exacerbating disease risk and progression among *NOTCH3* and *HTRA1* carriers, reinforcing the importance of aggressive management and treatment of vascular risk factors for pathogenic mutation carriers.[26]

Common genetic variants associated with stroke culminate in novel genomic risk predictors

In contrast to monogenic disorders, the polygenic model of disease inheritance stipulates that many common genetic variants (on the order of hundreds of thousands) with individually weak effects (OR, < 1.20) combine to explain the most of the genetic predisposition to complex diseases. For stroke, studies estimate that the totality of common polymorphisms, with common being defined as having a minor allele frequency > 1%, explain 16%-44% of the overall variance in stroke liability (ie, heritability), depending on the subtype. Specifically, heritability for intracerebral hemorrhage (ICH), large artery atherosclerosis, cardioembolic stroke, cryptogenic stroke, and small vessel stroke has been estimated at 44%, 40%, 33%, 24%, and 16%, respectively.[39-41] Altogether, this indicates that a substantial proportion of stroke risk is explained by common genetic risk factors.

GWAS with large sample sizes are necessary to uncover the specific susceptibility loci mediating these strong heritability signals, demanding a coordinated effort of researchers worldwide via the International Stroke Genetics Consortium. The

first collaborative International Stroke Genetics Consortium GWAS published in 2012 included 3548 stroke cases and 5972 controls of European ancestry.[42] This seminal study identified 4 susceptibility loci for large artery atherosclerosis (*HDAC9*, *CDKN2A/B*) and cardioembolic stroke (*PITX2*, *ZFHX3*), all of which have since been replicated.

Fast forward a decade, and recent stroke GWAS are several orders of magnitude larger and more ethnically diverse. Among the largest, the GIGASTROKE initiative conducted a cross-ancestry GWAS meta-analysis of 110,182 stroke cases and > 1.5 million control participants.[43] Cases included individuals of European (66.7%), East Asian (24.8%), African American (3.7%), South Asian (3.3%), and Hispanic (1.4%) ancestry. GIGASTROKE identified 89 stroke susceptibility loci, which were enriched for genes highly expressed in brain vascular endothelial cells and astrocytes, as well as those involved in the formation of blood clots and new blood vessels. Of the 89 loci, 69 were associated with all stroke, 45 with IS, 15 with large artery stroke, 13 with cardioembolic stroke, and 10 with small vessel stroke. Small vessel stroke loci included variants within or nearby genes associated with: (1) monogenic causes of stroke and VaD (*COL4A2*, *HTRA1*, *FOXF2*); (2) MRI markers of SVD (*NBEAL1*, *PMF1*, *PTCH1*, *SH3PXD2A*, *ZCCH14*); and (3) blood pressure-related traits (*COL4A2*, *NBEAL1*, *PMF1*, *PRDM16*, *SH3PXD2A*).[44,45]

Numerous studies have shown the potential utility of PRS for stroke risk stratification with a few common themes emerging. First, PRS are strongly predictive of IS incidence (HR, approximately 1.30 per standard deviation increase), with effects comparable with established clinical risk factors, such as hypertension and smoking.[43,46,47] Second, risk captured by PRS is complementary to that of clinical factors, and incorporating PRS with them jointly in a statistical model maximizes predictiveness. Third, PRS performance varies substantially according to ethnicity, with PRS in understudied populations performing worse (HR for African descent, 1.11 per 1 standard deviation increase) than in better represented groups (HR for European descent, 1.26; HR for East Asian descent, 1.33).[43] Fourth, PRS identifies a significant proportion of the population at extreme risk of stroke, equivalent to that of carrying rare pathogenic mutations. The most recent GIGASTROKE PRS identified 2.5% of European, 2% of East Asian, and 0.1% of African individuals to have double the risk of stroke compared with those with an average PRS in the middle decile.[43] PRS might also help to fulfil one of the promises of precision medicine, specifically, to predict medication-associated adverse effects. For example, Mayerhofer et al. reported that a PRS predicted ICH risk within anticoagulant users (HR, 1.24).[48] Also, this PRS improved risk stratification when used in addition to a clinical risk score for bleeding (HR, 1.33). Similarly, lower LDL-cholesterol has been epidemiologically and genetically associated with greater risk of ICH, and a PRS predicting LDL response during statin treatment was associated with higher ICH risk (HR, 1.16) even after adjusting for statin dose.[49] Ultimately, as stroke GWAS continue to grow in sample size and ethnic diversity, PRS will become increasingly predictive and robust, perhaps warranting more routine consideration alongside other clinical risk assessments.

Covert stroke

Approximately 10%-30% of elderly individuals aged older than 40 years experience covert (or silent) brain infarcts (CBIs).[50] CBIs are defined as cerebral ischemic events measuring ≥ 3 mm in diameter on MRI scans and lack clinical symptoms associated with overt stroke.[50,51] Notably, CBIs occur 5 times more frequently than symptomatic strokes, and have an estimated heritability of 29%.[51,52] Their pathophysiology varies, with atherosclerosis, atrial fibrillation, and cerebral SVD (CSVD) being prominent causes, often leading to embolic or thrombotic arterial occlusion.[53] CBI is a risk factor for cognitive decline, dementia, and stroke; those with CBIs have an estimated fivefold risk of incident dementia and fourfold risk of incident stroke compared with those without.[54-56]

Genetic research on covert stroke is still in its early stages,[52,57] with the largest transethnic meta-analysis including 20,949 individuals from 18 population-based cohorts.[52] Of the MRI-detected brain infarcts analyzed, approximately 90% of them were covert, and hence will be referred to as CBI herein. Analysis yielded 2 genome-wide significant (GWS) loci, with signals near the *FBN2* and *LINC00539/ZDHHC20* gene regions. Although validation efforts in a smaller sample (N = 6862) failed to replicate these findings, these gene regions remain biologically plausible.[52] The single nucleotide polymorphism (SNP) rs39938 (C > T) in the *FBN2* gene, has been associated with an elevated risk of CBI (OR, 1.21; 95% CI, 1.13-1.30).[52] Intriguingly, independent SNPs in *FBN2* have also been reported to be associated with systolic blood pressure (rs6595838) and overall stroke risk (rs55670004).[52] *FBN2* encodes for fibrillin 2, a glycoprotein essential in microfibril formation and the maintenance of vascular integrity within the extracellular matrix.[58] Rare mutations in *FBN2* represent a monogenic cause of a connective tissue disorder, congenital contractural arachnodactyly, and age-related macular degeneration, a condition characterized by damage to small blood vessels that supply the macula.[59,60] The second CBI locus (rs12583648; G > C; OR, 1.21; 95% CI, 1.13-1.29) was an intergenic SNP located downstream of *LINC00539* and *ZDHHC20*.[52] A modestly correlated SNP ($r^2 = 0.10$) within the same region has also been associated with WMH at suggestive significance (rs155076).[61] Notably, in GWAS of the adverse metabolic effects of the antihypertensive medication, hydrochlorothiazide, *LINC00539/ZDHHC20* further emerged as a suggestive locus regulating plasma triglyceride levels.[62] One notable SNP with biological plausibility emerged at suggestive significance: rs9371194 (G > T) in *PLEKHG1* (OR, 1.19; 95% CI, 1.11-1.28).[52] A moderately correlated SNP ($r^2 = 0.44$) within the same region, rs275350 (G > C), has been associated with WMH,[63] suggesting this as an important susceptibility locus for CSVD. The *PLEKHG1* gene encodes for a rho guanine nucleotide exchange factor, a protein instrumental in the cell reorientation within the vascular endothelium.[63] Studies have shown strong epidemiological associations of various vascular risk factors with the occurrence of CBI, especially hypertension (OR, 1.62; 95% CI, 1.48-1.78).[52] Indeed, PRS for systolic blood pressure is positively associated with risk of CBI (OR, 1.03 per mm Hg increase).[52]

CAA

CAA is a progressive, age-related condition, affecting the blood vessels in the brain. It is characterized by the accumulation of abnormal amyloid proteins, leading to the formation of amyloid plaques on the walls of the small and medium blood vessels in cerebral cortex and leptomeninges.[64] The most common clinical presentations include spontaneous lobar ICH, transient focal neurological episodes, convexity subarachnoid hemorrhage, CMBs, cognitive impairment/dementia, neurological symptoms (including headaches, confusion, and difficulty with speech or language), and seizures.[64,65] The current standardized assessment and reporting of CAA is outlined in the Boston criteria (version 2.0).[64] In this method, definitive diagnosis is CAA is only possible with histopathological analysis from a brain autopsy, whereas the second tier identifies probable CAA with supporting pathology from evacuated hematoma or biopsy samples. The third tier and fourth tiers identify probable and possible CAA, respectively, using noninvasive MRI markers in patients (age 50 years or older).[64]

CAA mostly occurs in sporadic form, in mid to later life, resulting from a culmination of risk factors, instead of a single causal contributor. Two distinct types of sporadic CAA have been identified. CAA-type 1 is characterized by the presence of amyloid deposits in a wide range of blood vessels within the brain, including cortical capillaries, leptomeningeal and cortical arteries, arterioles, veins, and venules. Conversely, CAA-type 2 is primarily associated with amyloid deposits in leptomeningeal and cortical vessels, sparing cortical capillaries.[66] Early-onset CAA, however, might affect patients younger than the age of 50 years,[67] typically caused by rare monogenic mutations, with severe clinical manifestations.[68] Although the prevalence of early-onset CAA is difficult to estimate, the overall prevalence of CAA increases with age, affecting 2.3% for patients between the ages of 65 and 74 years, 8.0% for those age 75-84 years, and 12.1% in patients older than the age of 85 years.[68] CAA is highly prevalent in individuals with AD, with postmortem examinations of subjects affected by AD showing signs of CAA in 90% of cases.[65,69]

As of 2022, 45 human proteins have been reported to form amyloid fibrils, however only 6 have been described to affect the central nervous system[70]; of these, amyloid beta (Aβ) protein plays a major role in the etiology of CAA, but a few non-Aβ CAA also exist. The amyloid precursor protein (APP) is a transmembrane glycoprotein, housing the Aβ peptide. Under normal circumstances, APP assists with synaptic formation and repair. Aβ peptide is cleaved in several ways by proteolytic processes via α-, β-, or γ-secretase. The enzyme α-secretase cleaves APP within the Aβ domain, generating a soluble extracellular domain and an intercellular domain that is further cleaved by γ-secretase to release a nontoxic p3-peptide.[71] Alternatively, cleavage by β- or γ-secretase produces intact $A\beta_{40}$ and $A\beta_{42}$ peptides, respectively, which are subsequently transported across the blood brain barrier, facilitated by *lrp-1*, *rage*, *bcrp*, and *p-gp* proteins, and degraded by enzymes, such as matrix metalloproteins and APOE.[72] The precise mechanisms by which alleles influence the development and progression of CAA are still under investigation. It is postulated that the causal alleles result in increased production, impaired transport, and degradation of amyloid peptides, causing the accumulation of amyloid fibrils on the small- and medium-calibre leptomeningeal and cortical artery walls.[65] The amyloid fibrils form plaques, which weaken the cerebral blood vessels and disrupt normal blood flow.

The *APOE ε4* allele is the only genetic risk factor robustly associated with sporadic CAA. A comprehensive meta-analysis of 24 studies (N = 3520) of *APOE* genotypes and pathologically diagnosed sporadic CAA, showed a significant association of *APOE ε4* with CAA. Generally, the *APOE ε4* allele is strongly associated with CAA type 1 (where Aβ deposition is found in cortical capillaries), cognitive phenotypes, and AD pathology,[67,73] however, irrespective of the dementia status, the strength of the association of CAA and *APOE ε4* increases in a dose-dependent manner, as the OR of heterozygous carriers increases from 2.09 (95% CI, 1.69-2.58) to 6.6 (95% CI, 4.47-9.75) for homozygous individuals.[73] Interestingly, an intronic variant of *LINC-PINT* plays a protective role against CAA in the absence of *APOE ε4* allele.[74] The relationship between *APOE ε2* and CAA is less convincing. A few studies have identified *APOE ε2* as a risk factor of ICH in CAA patients,[68,75] and is observed to affect bleeding-prone leptomeningeal vessels,[67,73] conversely, on the basis of the meta-analysis of 11 studies (N = 1640), the overall association of CAA with *APOE ε2* is nonsignificant and with effect in the opposite direction (OR, 0.73; 95% CI, 0.53-1.00).[73]

There are several subtypes of hereditary, monogenic CAA, named after the regions where they were first diagnosed. The most common form of monogenic CAA is caused by the mutation in the gene encoding the APP,[76,77] which is linked to the Dutch, Italian, Piedmont, Iowa, Flemish, and Arctic familial CAA, with the Dutch type being the most prevalent hereditary and well studied cause of CAA.[67] Mutations of APP that primarily cause AD can also show significant pathological evidence of CAA, including the London, Indiana, and the Swedish mutations.[67] Other monogenic CAA genes like *CST3* and *ITM2B*, are inherited in an autosomal dominant manner, and induce amyloidosis via non-Aβ pathways; *CST3* causes Icelandic CAA, and *ITM2B* mutations result in the British and Danish types.[67] Rarer forms of monogenic CAA are linked to *TGF-β1*, *TOMM40* and *CR1*, *PSEN1*, and *PSEN2* genes. A number of genes with limited evidence for association with CAA include *LRP1*, *ACT*, *ACE*, *BCHE*, *DXS1047* locus, *APOE* promoter, *A2M*, *PON1*, *NEP*, *OLR1*, *LRP*, *CYP46*, *CH25H*1*, *VEGF*, interleukin genes (*IL-1A*, *IL-1B*, *IL-33*), *GSTO1-1*, *CTSD*, *AβPP*, *AβPp* promoter, *ABCA7*, *CLU*, and *CR1*.[67,73]

Although there are currently no effective therapeutics available to cure or halt the progression of CAA, early probable diagnosis empowers the implementation of management strategies. These strategies aim to mitigate the risk of first-time or recurrent lobar ICH, which, in turn involves managing hypertension, comorbidities including cardiovascular risk factors, and advocating the avoidance of lifelong systemic anticoagulant use.[78] Ongoing research focused on therapeutic approaches are aimed at the reduction or clearance of amyloid deposition, with antiamyloid antibodies, β-secretase inhibitors, and γ-secretase.[65] In recent years, the phase III trials of lecanemab and donanemab resulted in promising prospects for AD and might be studied further in the context of CAA.[79,80]

Covert Cerebral SVD—Silent But Deadly

Cerebral SVD (CSVD) is a highly prevalent and chronic vascular condition that affects small brain vessels, such as the arterioles, capillaries, and venules.[45,81] Causes of CSVD are heterogeneous and include arteriosclerosis, CAA, venous collagenosis, genetic small-vessel angiopathy, and inflammation and immune-mediated SVD.[81] CSVD causes 25% of strokes, while contributing up to 40%-45% of all dementia cases, making it imperative to identify markers of CSVD.[82] CSVD is classified into overt or covert subtypes. Although overt CSVD requires a presence of clinical symptoms such as stroke and dementia for diagnosis, covert CSVD is detected with MRI in the absence of observable neurological symptoms; MRI-detected markers, such as WMH, CMBs, and perivascular spaces (PVS) are hallmarks for CSVD. These markers exhibit a high degree of heritability, with the highest being WMH heritability reported at 71% in a male twin study,[83] and a more recent heritability estimate at 55%.[84] Dilated PVS burden has an estimated heritability of 59%.[85] Overall, research indicates a high degree of genetic overlap in the different manifestations of CSVD, with 34% of the GWS loci being associated with at least 2 CSVD phenotypes.[45] Furthermore, covert CSVD is a risk factor for future ischemic and hemorrhagic stroke, VCI, and dementia.

WMH

WMH are regions in the brain that appear particularly bright on fluid-attenuated inversion recovery T2-weighted sequence MRI scans, and are an indication of pathological changes in the white matter resulting from demyelination and axonal loss of nerve fibres.[86,87] Past studies have identified hypertension, aging, and smoking status as risk factors for WMH development.[61,86,87] In addition, MR analyses reveal that genetically determined WMH burden is putatively causally associated with stroke, AD, and hypertension.[88] WMH are associated with a twofold increased risk of dementia and threefold increased risk of stroke,[89] and consequently, are major contributors of VaD. Family- and population-based studies show that WMH volume is associated with AD, non-AD related dementias, and mild cognitive impairment.[90] Also, an increased burden of WMH is a hallmark of monogenic forms of VaD including the severe recessive form of small vessel disease and CADASIL.[90]

GWAS on WMH have revealed numerous loci with shared pathways involving the vascular, glial, and other neuronal cells, several of which are also linked to cell structural components and cell apoptosis.[61,88,90,91] Some examples of GWS loci for WMH include genes that encode for matrisome proteins, such as *COL4A2*, *EFEMP1*, *NID2*, and *VCAN* implicated in the largest multiancestry GWAS meta-analysis of WMH (N = 50,970).[88] Among these loci, *EFEMP1* supports tumour cell motility and invasion, and was upregulated in malignant gliomas; *COL4A2/1* mutations affect basement membrane collagen, and are a known cause of monogenic SVD; *VCAN* assists with cell adhesion, oligodendrocyte maturation, and remyelination; *PLEKHG1* is associated with WMH and IS[63,92]; and the chr1q22 region (containing *PMF1-BGLAP*) is also a susceptibility locus for ICH.[88]

More recently, Persyn et al. identified 19 independent loci related to WMH (N = 42,310), accounting for 1.79% variance in the UK Biobank.[91] SNPs in proximity to *APOE*, *COL4A2*, *TRIM47*, *SH3PXD2A*, *EFEMP1*, and *NBEAL1* were associated with WMH recapitulating evidence from previous GWAS.[61,88,91,93] *NBEAL1* and *COL4A2* are noteworthy, because they are associated with WMH in stroke and stroke-free populations, and linked to cholesterol metabolism and increased coronary artery disease risk, highlighting common genetic susceptibility.[93] A transcriptome-wide association study performed to identify tissue-specific gene expression associations with traits, revealed that genetically elevated arterial expression levels of *ADAMTSL4* and *SLC25A44* were associated with greater WMH burden, whereas decreased arterial expression of *CALCRL* was associated with increased WMH burden.[91] Interestingly, *CALCRL* is involved in a vasodilation pathway relating to the calcitonin gene related peptide and adrenomedullin receptors, with ameliorating effects on cardiovascular diseases, and a potential therapeutic target for cerebral ischemia.[91]

CMBs

CMBs are chronic hemorrhages that occur in the small vessels of the brain, typically measuring a few millimeters to centimeters in diameter. These microbleeds cause accumulation of a paramagnetic material called hemosiderin, a byproduct of blood degradation located within macrophages. Hemosiderin, and thus CMBs, are detected with T2*-gradient recalled echo imaging.[94] The prevalence of CMBs is positively correlated with age, affecting 6.5% of patients aged 35-40 years, and 35.7% of patients aged 80 years and older.[95] Furthermore, individuals with CMBs have a twofold greater risk of developing dementia.[96] Occurrence of multiple CMBs increase the odds of VaD (OR, 3.10; 95% CI, 1.11-8.62), showing a profile of VCI by decreasing executive function and processing speeds.[97]

CMBs can be subcategorized by their location, as strictly lobar, deep, or a mixture of both. Lobar CMBs are distinct to cortical-subcortical areas in the brain lobes and cerebellum and are generally thought to be a result of CAA.[94] CMBs in deep white matter regions of the basal ganglia (BG), thalami, brainstem, or cerebellum are a characteristic result of hypertensive angiopathy. Patients with mixed CMBs and deep white matter CMBs have a threefold greater risk of developing dementia, compared with those with lobar CMBs, who have an approximately twofold increased risk of dementia.[98]

GWAS have facilitated identification of common genetic variants associated with CMBs, providing insight on the genetic basis of CMBs, which have received less attention compared with other markers of CSVD, like WMH. Knol et al. conducted a GWAS meta-analysis involving 11 population-based cohorts and 3 stroke cohorts (N = 25,862).[99] They identified multiple loci associated with CMBs, with 1 locus reaching genome-wide significance: *APOE* (rs769449). Additionally other loci previously associated with WMH, deep ICH, and lacunar stroke were suggestively associated with CMBs.

Although the investigation of genetic mechanisms underlying CMBs remains crucial, equally imperative is the exploration of risk factors for CMBs, because of their

significant contribution to the overall risk of dementia. Liu et al. reported that genetically determined risks of IS (OR, 1.47; 95% CI, 1.04-2.07) and small-vessel occlusion stroke (OR, 1.62; 95% CI, 1.07-2.47) were associated with greater risk of CMBs.[100] With 25%-30% of stroke patients going on to develop VCI or VaD, thus, it is essential to further understand the independent and consequential contribution of CMBs to VCI and VaD outcomes.[101]

Perivascular space burden

PVS are normal cerebrospinal fluid-filled spaces around the blood vessels as they penetrate the brain's parenchyma.[85,102] The perivascular space burden refers to the extent of pathological changes observed in these spaces, detected using MRI. PVS play a prominent role in the brain drainage system, waste clearance, blood flow regulation, and interstitial fluid exchange.[85,103] White matter PVS (WM-PVS) burden, hippocampus (HIP) PVS (HIP-PVS) burden, and BG PVS (BG-PVS) burden have an estimated heritability of 11%, 8%, and 5%, respectively.[104] Large PVS (diameter > 3 mm) have been associated with a greater risk of CMBs (adjusted risk ratio, 1.43; 95% CI, 1.18-1.72), subcortical infarcts (adjusted risk ratio, 2.54; 95% CI, 1.76-3.68), a larger WMH 5-year progression, and quadrupling the risk of VaD in at a 5-year follow-up.[105] PVS burden was reported to increase with age (OR, 1.29; 95% CI, 1.28-1.30) in a neuroimaging study of approximately 40,000 individuals.[106] Compared with WMH and CMBs, PVS detection is more sensitive in early stages of CSVD.[106] Several studies have examined the relationship of PVS burden with cardiovascular risk factors. Hypertension, particularly diastolic blood pressure, was associated with more PVS in all regions (OR, 1.04-1.05).[107] Lacunae and WMH volume were associated with PVS in different areas (ORs, 1.13 and 1.10, respectively), particularly in the BG.

The largest GWAS meta-analysis of PVS burden in 18 European cohort studies (N = 40,095)[104] identified 22 GWS variants associated with PVS burden, and 2 risk loci for WM-PVS. BG-PVS and HIP-PVS had a moderate genetic correlation (genetic correlation [r_g], 0.63; standard error [SE], 0.14]), and WM-PVS had a correlation with BG-PVS (r_g, 0.24; SE, 0.12) and HIP-PVS (r_g, 0.27; SE, 0.09). Of the 24 PVS burden loci, 5 were associated with WMH, and another 5 were associated with blood pressure factors. Individuals with higher PRS for systolic and diastolic blood pressures showed a greater burden of PVS, specifically in the BG and HIP regions. Other markers of CSVD also exhibited moderate to high correlation with BG- and HIP-PVS. The study highlighted GWS PVS loci including *FOXF2*, *LPAR1*, *EFEMP1*, *KCNK2*, and *NBEAL1-ICA1L* as risk factors for WMH and small vessel stroke, honing in on possible biological pathways implicated with PVS burden. Additionally, the *SLC13A3* gene was identified as GWS for WM-PVS. *SLC13A3* encodes for a Na^+/dicarboxylate cotransporter that is expressed in the kidneys, astrocytes, and chord plexus. The study identified that solute carriers might contribute to interstitial fluid accumulation near PVS. In addition, WM-PVS were associated with lower lysophosphatidic acid receptor 1 (an important regulator of glial function) expression in vascular tissue.

Current Challenges, Opportunities, and Future Directions

Despite being the second most prevalent cause of dementia, research on the genetics of VaD is limited, because of the syndromic nature of VaD and the heterogeneity in disease pathology. The relationship of clinical subtypes with genetic or biological factors in VaD highlights the complexity of categorizing diseases solely on the basis of clinical manifestations. Although clinical subtypes provide a practical framework for diagnosis and treatment, genetics and biology show that diseases are often interconnected at a molecular level, challenging the rigid boundaries of clinical classifications. A case in point, the association of *APOE ε4* with AD and CAA, as well as the association of *NBEAL1* with AD and WMH. Although there are many genes linked to cerebrovascular diseases associated with VaD, the distinction between those that progress to dementia and those that do not, combined with the utility of multiomics data identifying specific pathways, might hold the key to addressing the heterogeneity and provide a more precise diagnosis.

Existing global consortia related to VaD are few and lack a genetic focus.[108] Furthermore, many of the reported genetic associations for VaD diagnoses rely on small- to midsized single studies without rigourous replication initiatives. Large biobanks (eg, UK Biobank and All of Us) provide a unique opportunity to understand the genetic determinants of many health outcomes; however, the low prevalence of VaD, and further lower representation of multiethnic groups in these consortia, represents a major obstacle. The largest GWAS meta-analysis including 9000 cases only identified the already-known *APOE* locus in GWAS.[109] This is in stark contrast to 75 loci identified for AD.[110] One way to overcome the challenges posed by the reduced power is not only to initiate accessible, focused consortia, with health records and biomarker readings, but also to use statistical tools to integrate the various consortia, multiethnic studies, and genetically related traits.

As more VaD-related genes are identified, it is vital to acknowledge that the evidence spans decades, marked by various technological advancements and study designs. These factors raise concerns about gene validity for clinical diagnosis, however, using a standardized framework like those created by the National Human Genome Research Institute (ClinGen) to ascertain gene-disease validity can aid gene prioritization in clinical contexts, improving genetic data usability and dissemination.[111]

Studying the full spectrum of genetic contributions to VaD, including common and rare variants, alongside epigenomics, transcriptomics, and proteomics, holds promise in revolutionizing our understanding of VaD. Identifying specific genetic factors associated with an increased risk provides an opportunity for early risk assessment. The current gold standard for diagnosing VaD, involving MRI-based identification of cerebrovascular pathology, faces pragmatic challenges in settings that lack adequate facilities, expertise, or resources. Genetic studies that provide insight into the biological mechanisms, through applications of integromics, can potentially lead to the discovery of diagnostic biomarkers that are more accessible and therapeutic targets that are more personalized.

Ethics Statement

The research reported has adhered to the relevant ethical guidelines.

Patient Consent

The authors confirm that patient consent is not applicable to this article, because this is a review article consisting of previously published data.

Funding Sources

None.

Disclosures

Dr Guillaume Paré received research support from Bayer and honoraria from Amgen, Bayer, Illumina, Novartis, and Sanofi. Dr Michael Chong received honoraria from Bayer. The remaining authors have no conflicts of interest to disclose.

References

1. Iadecola C, Duering M, Hachinski V, et al. Vascular cognitive impairment and dementia: JACC Scientific Expert Panel. J Am Coll Cardiol 2019;73:3326.
2. Hachinski V, Iadecola C, Petersen RC, et al. National Institute of Neurological Disorders and Stroke-Canadian Stroke Network vascular cognitive impairment harmonization standards. Stroke 2006;37:2220-41.
3. Legdeur N, van der Lee SJ, de Wilde M, et al. The association of vascular disorders with incident dementia in different age groups. Alzheimers Res Ther 2019;11:47.
4. Sanders AE, Schoo C, Kalish VB. Vascular Dementia. Treasure Island, FL: StatPearls Publishing, 2023.
5. Skrobot OA, Black SE, Chen C, et al. Progress toward standardized diagnosis of vascular cognitive impairment: guidelines from the Vascular Impairment of Cognition Classification Consensus Study. Alzheimers Dement 2018;14:280-92.
6. Hoggart C, Shimizu C, Galassini R, et al. Identification of novel locus associated with coronary artery aneurysms and validation of loci for susceptibility to Kawasaki disease. Eur J Hum Genet 2021;29:1734-44.
7. Markus HS, Schmidt R. Genetics of vascular cognitive impairment. Stroke 2019;50:765-72.
8. Gorelick PB, Scuteri A, Black SE, et al. Vascular contributions to cognitive impairment and dementia: a statement for healthcare professionals from the American Heart Association/American Stroke Association. Stroke 2011;42:2672-713.
9. Filler J, Georgakis MK, Dichgans M. Risk factors for cognitive impairment and dementia after stroke: a systematic review and meta-analysis. Lancet Healthy Longev 2024;5:e31-44.
10. Bulik-Sullivan BK, Loh PR, Finucane HK, et al. LD score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat Genet 2015;47:291-5.
11. Di Scipio M, Khan M, Mao S, et al. A versatile, fast and unbiased method for estimation of gene-by-environment interaction effects on biobank-scale datasets. Nat Commun 2023;14:5196.
12. Fongang B, Weinstein G, Guðjónsson A, et al. A meta-analysis of genome-wide association studies identifies new genetic loci associated with all-cause and vascular dementia. Alzheimers Dement 2021;17(suppl 3):e056081.
13. Littlejohns TJ, Collister JA, Liu X, Clifton L, Tapela NM, Hunter DJ. Hypertension, a dementia polygenic risk score, APOE genotype, and incident dementia. Alzheimers Dement 2023;19:467-76.
14. Dybjer E, Kumar A, Nägga K, et al. Polygenic risk of type 2 diabetes is associated with incident vascular dementia: a prospective cohort study. Brain Commun 2023;5:fcad054.
15. Li Y, Zeng Z, Zhuang Z, et al. Polysocial and polygenic risk scores and all-cause dementia, Alzheimer's disease, and vascular dementia. J Gerontol A Biol Sci Med Sci 2024;79:glad262.
16. Yu C, Ryan J, Orchard SG, et al. Validation of newly derived polygenic risk scores for dementia in a prospective study of older individuals. Alzheimers Dement 2023;19:5333-42.
17. Litkowski EM, Logue MW, Zhang R, et al. Mendelian randomization study of diabetes and dementia in the Million Veteran Program. Alzheimers Dement 2023;19:4367-76.
18. Thomassen JQ, Tolstrup JS, Benn M, Frikke-Schmidt R. Type-2 diabetes and risk of dementia: observational and Mendelian randomisation studies in 1 million individuals. Epidemiol Psychiatr Sci 2020;29:e118.
19. Zhang X, Geng T, Li N, et al. Associations of lipids and lipid-lowering drugs with risk of vascular dementia: a Mendelian randomization study. Nutrients 2022;15:69.
20. Lo JW, Crawford JD, Desmond DW, et al. Long-term cognitive decline after stroke: an individual participant data meta-analysis. Stroke 2022;53:1318-27.
21. Koton S, Pike JR, Johansen M, et al. Association of ischemic stroke incidence, severity, and recurrence with dementia in the Atherosclerosis Risk in Communities Cohort Study. JAMA Neurol 2022;79:271-80.
22. Stevens DL, Hewlett RH, Brownell B. Chronic familial vascular encephalopathy. Lancet 1977;1:1364-5.
23. Chong M, O'Donnell M, Thijs V, et al. Mendelian genes and risk of intracerebral hemorrhage and small-vessel ischemic stroke in sporadic cases. Stroke 2017;48:2263-5.
24. Grami N, Chong M, Lali R, et al. Global assessment of Mendelian stroke genetic prevalence in 101 635 individuals from 7 ethnic groups. Stroke 2020;51:1290-3.
25. Lee YC, Chung CP, Chang MH, Wang SJ, Liao YC. Cysteine-altering variant is an important risk factor for stroke in the Taiwanese population. Neurology 2020;94:e87-96.
26. Cho BPH, Harshfield EL, Al-Thani M, Tozer DJ, Bell S, Markus HS. Association of vascular risk factors and genetic factors with penetrance of variants causing monogenic stroke. JAMA Neurol 2022;79:1303-11.
27. Lee YC, Chung CP, Chao NC, et al. Characterization of heterozygous mutations in Taiwanese patients with cerebral small vessel disease. Stroke 2018;49:1593-601.
28. Verdura E, Hervé D, Scharrer E, et al. Heterozygous HTRA1 mutations are associated with autosomal dominant cerebral small vessel disease. Brain 2015;138:2347-58.
29. Nozaki H, Kato T, Nihonmatsu M, et al. Distinct molecular mechanisms of HTRA1 mutants in manifesting heterozygotes with CARASIL. Neurology 2016;86:1964-74.

30. Chung J, Hamilton G, Kim M, et al. Rare missense functional variants at *COL4A1* and *COL4A2* and in sporadic intracerebral hemorrhage. Neurology 2021;97:e236-47.

31. Vahedi K, Kubis N, Boukobza M, et al. COL4A1 mutation in a patient with sporadic, recurrent intracerebral hemorrhage. Stroke 2007;38: 1461-4.

32. Meuwissen MEC, Halley DJJ, Smit LS, et al. The expanding phenotype of COL4A1 and COL4A2 mutations: clinical data on 13 newly identified families and a review of the literature. Genet Med 2015;17: 843-53.

33. Okazaki S, Morimoto T, Kamatani Y, et al. Moyamoya disease susceptibility variant RNF213 p.R4810K increases the risk of ischemic stroke attributable to large-artery atherosclerosis. Circulation 2019;139: 295-8.

34. Fukutake T. CARASIL: Identification of the clinical concept [in Japanese]. Rinsho Shinkeigaku 2010;50:849-51.

35. Coste T, Hervé D, Neau JP, et al. Heterozygous HTRA1 nonsense or frameshift mutations are pathogenic. Brain 2021;144:2616-24.

36. Park HK, Lee KJ, Park JM, et al. Prevalence of mutations in Mendelian stroke genes in early onset stroke patients. Ann Neurol 2023;93:768-82.

37. Miyawaki S, Imai H, Shimizu M, et al. Genetic variant RNF213 c. 14576G > A in various phenotypes of intracranial major artery stenosis/ occlusion. Stroke 2013;44:2894-7.

38. Opherk C, Gonik M, Duering M, et al. Genome-wide genotyping demonstrates a polygenic risk score associated with white matter hyperintensity volume in CADASIL. Stroke 2014;45:968-72.

39. Devan WJ, Falcone GJ, Anderson CD, et al. Heritability estimates identify a substantial genetic contribution to risk and outcome of intracerebral hemorrhage. Stroke 2013;44:1578.

40. Bevan S, Traylor M, Adib-Samii P, et al. Genetic heritability of ischemic stroke and the contribution of previously reported candidate gene and genomewide associations. Stroke 2012;43:3161-7.

41. Pulit SL, Weng LC, McArdle PF, et al. Atrial fibrillation genetic risk differentiates cardioembolic stroke from other stroke subtypes. Neurol Genet 2018;4:e293.

42. International Stroke Genetics Consortium (ISGC), Wellcome Trust Case Control Consortium 2 (WTCCC2), Bellenguez C, et al. Genome-wide association study identifies a variant in HDAC9 associated with large vessel ischemic stroke. Nat Genet 2012;44:328-33.

43. Mishra A, Malik R, Hachiya T, et al. Stroke genetics informs drug discovery and risk prediction across ancestries. Nature 2022;611: 115-23.

44. Crawford KM, Gallego-Fabrega C, Kourkoulis C, et al. Cerebrovascular Disease Knowledge Portal: an open-access data resource to accelerate genomic discoveries in stroke. Stroke 2018;49:470-5.

45. Bordes C, Sargurupremraj M, Mishra A, Debette S. Genetics of common cerebral small vessel disease. Nat Rev Neurol 2022;18:84-101.

46. Abraham G, Malik R, Yonova-Doing E, et al. Genomic risk score offers predictive performance comparable to clinical risk factors for ischaemic stroke. Nat Commun 2019;10:5819.

47. Lu X, Niu X, Shen C, et al. Development and validation of a polygenic risk score for stroke in the Chinese population. Neurology 2021;97: e619.

48. Mayerhofer E, Parodi L, Prapiadou S, et al. Genetic risk score improves risk stratification for anticoagulation-related intracerebral hemorrhage. Stroke 2023;54:791-9.

49. Mayerhofer E, Malik R, Parodi L, et al. Genetically predicted on-statin LDL response is associated with higher intracerebral haemorrhage risk. Brain 2022;145:2677-86.

50. Meinel TR, Kaesmacher J, Roten L, Fischer U. Covert brain infarction: towards precision medicine in research, diagnosis, and therapy for a silent pandemic. Stroke 2020;51:2597-606.

51. Vermeer SE, Longstreth WT Jr, Koudstaal PJ. Silent brain infarcts: a systematic review. Lancet Neurol 2007;6:611-9.

52. Chauhan G, Adams HHH, Satizabal CL, et al. Genetic and lifestyle risk factors for MRI-defined brain infarcts in a population-based setting. Neurology 2019;92:e486-503.

53. Durrani R, Hill MD, Smith EE. Preventing covert brain infarct-related cognitive impairment and dementia. Can J Neurol Sci 2020;47:456-63.

54. NeuroVISION Investigators: Perioperative covert stroke in patients undergoing non-cardiac surgery (NeuroVISION): a prospective cohort study. Lancet 2019;394:1022-9.

55. Kent DM, Leung LY, Zhou Y, et al. Association of incidentally discovered covert cerebrovascular disease identified using natural language processing and future dementia. J Am Heart Assoc 2023;12: e027672.

56. Vermeer SE, Hollander M, van Dijk EJ, et al. Silent brain infarcts and white matter lesions increase stroke risk in the general population: the Rotterdam Scan Study. Stroke 2003;34:1126-9.

57. Debette S, Bis JC, Fornage M, et al. Genome-wide association studies of MRI-defined brain infarcts: meta-analysis from the CHARGE Consortium. Stroke 2010;41:210-7.

58. Mead TJ, Martin DR, Wang LW, et al. Proteolysis of fibrillin-2 microfibrils is essential for normal skeletal development. Elife 2022;11: e71142.

59. Li AL, He JQ, Zeng L, et al. Case report: identification of novel fibrillin-2 variants impacting disulfide bond and causing congenital contractural arachnodactyly. Front Genet 2023;14:1035887.

60. Ratnapriya R, Zhan X, Fariss RN, et al. Rare and common variants in extracellular matrix gene fibrillin 2 (FBN2) are associated with macular degeneration. Hum Mol Genet 2014;23:5827-37.

61. Verhaaren BFJ, Debette S, Bis JC, et al. Multiethnic genome-wide association study of cerebral white matter hyperintensities on MRI. Circ Cardiovasc Genet 2015;8:398-409.

62. Del-Aguila JL, Beitelshees AL, Cooper-Dehoff RM, et al. Genome-wide association analyses suggest NELL1 influences adverse metabolic response to HCTZ in African Americans. Pharmacogenomics J 2014;14:35-40.

63. Traylor M, Tozer DJ, Croall ID, et al. Genetic variation in PLEKHG1 is associated with white matter hyperintensities (n = 11,226) [erratum in: 2019;93:608]. Neurology 2019;92:e749-57.

64. Charidimou A, Boulouis G, Frosch MP, et al. The Boston criteria version 2.0 for cerebral amyloid angiopathy: a multicentre, retrospective, MRI-neuropathology diagnostic accuracy study. Lancet Neurol 2022;21:714-25.

65. Cozza M, Amadori L, Boccardi V. Exploring cerebral amyloid angiopathy: insights into pathogenesis, diagnosis, and treatment. J Neurol Sci 2023;454:120866.

66. Thal DR, Ghebremedhin E, Rüb U, Yamaguchi H, Del Tredici K, Braak H. Two types of sporadic cerebral amyloid angiopathy. J Neuropathol Exp Neurol 2002;61:282-93.

67. Banerjee G, Collinge J, Fox NC, et al. Clinical considerations in early-onset cerebral amyloid angiopathy. Brain 2023;146:3991-4014.

68. Biffi A, Greenberg SM. Cerebral amyloid angiopathy: a systematic review. J Clin Neurol 2011;7:1-9.

69. Jäkel L, De Kort AM, Klijn CJM, Schreuder FHBM, Verbeek MM. Prevalence of cerebral amyloid angiopathy: a systematic review and meta-analysis. Alzheimers Dement 2022;18:10-28.

70. Buxbaum JN, Dispenzieri A, Eisenberg DS, et al. Amyloid nomenclature 2022: update, novel proteins, and recommendations by the International Society of Amyloidosis (ISA) Nomenclature Committee. Amyloid 2022;29:213-9.

71. MacLeod R, Hillert EK, Cameron RT, Baillie GS. The role and therapeutic targeting of α-, β- and γ-secretase in Alzheimer's disease. Future Sci OA 2015;1:FSO11.

72. Gatti L, Tinelli F, Scelzo E, et al. Understanding the pathophysiology of cerebral amyloid angiopathy. Int J Mol Sci 2020;21:3435.

73. Rannikmäe K, Samarasekera N, Martínez-Gonzâlez NA, Al-Shahi Salman R, Sudlow CLM. Genetics of cerebral amyloid angiopathy: systematic review and meta-analysis. J Neurol Neurosurg Psychiatry 2013;84:901-8.

74. Reddy JS, Allen M, Ho CCG, et al. Genome-wide analysis identifies a novel LINC-PINT splice variant associated with vascular amyloid pathology in Alzheimer's disease. Acta Neuropathol Commun 2021;9:93.

75. Goldberg TE, Huey ED, Devanand DP. Associations of APOE e2 genotype with cerebrovascular pathology: A postmortem study of 1275 brains. J Neurol Neurosurg Psychiatry 2021;92:7-11.

76. Revesz T, Ghiso J, Lashley T, et al. Cerebral amyloid angiopathies: a pathologic, biochemical, and genetic view. J Neuropathol Exp Neurol 2003;62:885-98.

77. Revesz T, Holton JL, Lashley T, et al. Genetics and molecular pathogenesis of sporadic and hereditary cerebral amyloid angiopathies. Acta Neuropathol 2009;118:115-30.

78. Kozberg MG, Perosa V, Gurol ME, van Veluw SJ. A practical approach to the management of cerebral amyloid angiopathy. Int J Stroke 2021;16:356-69.

79. Sims JR, Zimmer JA, Evans CD, et al. Donanemab in early symptomatic Alzheimer disease: the TRAILBLAZER-ALZ 2 randomized clinical trial. JAMA 2023;330:512-27.

80. Volloch V, Rits-Volloch S. Effect of lecanemab in early Alzheimer's disease: mechanistic interpretation in the amyloid cascade hypothesis 2. 0 perspective. J Alzheimers Dis 2023;93:1277-84.

81. Chojdak-Łukasiewicz J, Dziadkowiak E, Zimny A, Paradowski B. Cerebral small vessel disease: a review. Adv Clin Exp Med 2021;30: 349-56.

82. Cannistraro RJ, Badi M, Eidelman BH, Dickson DW, Middlebrooks EH, Meschia JF. CNS small vessel disease: a clinical review. Neurology 2019;92:1146-56.

83. Carmelli D, DeCarli C, Swan GE, et al. Evidence for genetic variance in white matter hyperintensity volume in normal elderly male twins. Stroke 1998;29:1177-81.

84. Atwood LD, Wolf PA, Heard-Costa NL, et al. Genetic variation in white matter hyperintensity volume in the Framingham study. Stroke 2004;35:1609-13.

85. Duperron MG, Tzourio C, Sargurupremraj M, et al. Burden of dilated perivascular spaces, an emerging marker of cerebral small vessel disease, is highly heritable. Stroke 2018;49:282-7.

86. Prins ND, Scheltens P. White matter hyperintensities, cognitive impairment and dementia: an update. Nat Rev Neurol 2015;11:157-65.

87. Wong EC, Chang CH. Vascular cognitive impairment and Dementia. Continuum 2022;28:750-80.

88. Sargurupremraj M, Suzuki H, Jian X, et al. Cerebral small vessel disease genomics and its implications across the lifespan. Nat Commun 2020;11:6285.

89. Wardlaw JM, Valdés Hernández MC, Muñoz-Maniega S. What are white matter hyperintensities made of?: Relevance to vascular cognitive impairment [erratum in: 2016;5:e002006]. J Am Heart Assoc 2015;4: 001140.

90. Ho WM, Wu YY, Chen YC. Genetic variants behind cardiovascular diseases and dementia. Genes (Basel) 2020;11:1514.

91. Persyn E, Hanscombe KB, Howson JMM, Lewis CM, Traylor M, Markus HS. Genome-wide association study of MRI markers of cerebral small vessel disease in 42,310 participants. Nat Commun 2020;11:1-12.

92. Abiko H, Fujiwara S, Ohashi K, et al. Rho guanine nucleotide exchange factors involved in cyclic-stretch-induced reorientation of vascular endothelial cells. J Cell Sci 2015;128:1683-95.

93. Traylor M, Zhang CR, Adib-Samii P, et al. Genome-wide meta-analysis of cerebral white matter hyperintensities in patients with stroke. Neurology 2016;86:146.

94. Martinez-Ramirez S, Greenberg SM, Viswanathan A. Cerebral microbleeds: overview and implications in cognitive impairment. Alzheimers Res Ther 2014;6:1-7.

95. Poels MMF, Vernooij MW, Arfan Ikram M, et al. Prevalence and risk factors of cerebral microbleeds. Stroke 2010;41(10 suppl):S103-6.

96. Hussein AS, Shawqi M, Bahbah EI, et al. Do cerebral microbleeds increase the risk of dementia? A systematic review and meta-analysis. IBRO Neurosci Rep 2023;14:86-94.

97. Qiu C, Cotch MF, Sigurdsson S, et al. Cerebral microbleeds, retinopathy, and dementia: the AGES-Reykjavik Study. Neurology 2010;75: 2221-8.

98. Romero JR, Beiser A, Himali JJ, Shoamanesh A, DeCarli C, Seshadri S. Cerebral microbleeds and risk of incident dementia: the Framingham Heart Study. Neurobiol Aging 2017;54:94-9.

99. Knol MJ, Lu D, Traylor M, et al. Association of common genetic variants with brain microbleeds: a genome-wide association study. Neurology 2020;95:e3331-43.

100. Liu R, Shi X, Feng J, et al. Ischemic stroke and cerebral microbleeds: a two-sample bidirectional Mendelian randomization study. Neurol Ther 2023;12:1299-308.

101. Kalaria RN, Akinyemi R, Ihara M. Stroke injury, cognitive impairment and vascular dementia. Biochim Biophys Acta 2016;1862:915-25.

102. Gouveia-Freitas K, Bastos-Leite AJ. Perivascular spaces and brain waste clearance systems: relevance for neurodegenerative and cerebrovascular pathology. Neuroradiology 2021;63:1581-97.

103. Yu L, Hu X, Li H, Zhao Y. Perivascular spaces, glymphatic system and MR. Front Neurol 2022;13:844938.

104. Duperron MG, Knol MJ, Le Grand Q, et al. Genomics of perivascular space burden unravels early mechanisms of cerebral small vessel disease. Nat Med 2023;29:950-62.

105. Ding J, Sigurðsson S, Jónsson PV, et al. Large perivascular spaces visible on magnetic resonance imaging, cerebral small vessel disease

progression, and risk of dementia: the Age, Gene/Environment Susceptibility—Reykjavik Study. JAMA Neurol 2017;74:1105-12.

106. Evans TE, Knol MJ, Schwingenschuh P, et al. Determinants of perivascular spaces in the general population: a pooled cohort analysis of individual participant data. Neurology 2023;100:e107-22.

107. Deramecourt V, Slade JY, Oakley AE, et al. Staging and natural history of cerebrovascular pathology in dementia. Neurology 2012;78:1043-50.

108. Gladman JT, Corriveau RA, Debette S, et al. Vascular contributions to cognitive impairment and dementia: research consortia that focus on etiology and treatable targets to lessen the burden of dementia worldwide. Alzheimers Dement 2019;5:789-96.

109. Fongang B, Sargurupremraj M, Jian X, et al. Genetic insights of all-cause and vascular dementia through genome-wide association studies. Alzheimers Dement 2022;18:e067165.

110. Andrews SJ, Renton AE, Fulton-Howard B, Podlesny-Drabiniok A, Marcora E, Goate AM. The complex genetic architecture of Alzheimer's disease: novel insights and future directions. eBioMedicine 2023;90:104511.

111. Rehm HL, Berg JS, Brooks LD, et al. ClinGen–the clinical genome resource. N Engl J Med 2015;372:2235-42.

112. De Vilder EYG, Cardoen S, Hosen MJ, et al. Pathogenic variants in the ABCC6 gene are associated with an increased risk for ischemic stroke. Brain Pathol 2018;28:822-31.

CHAPTER 4:
A METHOD TO ESTIMATE THE CONTRIBUTION OF RARE CODING VARIANTS TO COMPLEX TRAIT HERITABILITY.

Published in *Nature Communications.* (February 2024)

nature communications

Article

https://doi.org/10.1038/s41467-024-45407-8

A method to estimate the contribution of rare coding variants to complex trait heritability

Received: 28 October 2022

Accepted: 22 January 2024

Published online: 09 February 2024

Check for updates

Nazia Pathan[1,2], Wei Q. Deng[3,4], Matteo Di Scipio[1,5], Mohammad Khan[1,5], Shihong Mao[1], Robert W. Morton[1,2], Ricky Lali[1,6], Marie Pigeyre[1,5], Michael R. Chong[1,2,7] & Guillaume Paré[1,2,6,7]

It has been postulated that rare coding variants (RVs; MAF < 0.01) contribute to the "missing" heritability of complex traits. We developed a framework, the Rare variant heritability (RARity) estimator, to assess RV heritability (h^2_{RV}) without assuming a particular genetic architecture. We applied RARity to 31 complex traits in the UK Biobank (n = 167,348) and showed that gene-level RV aggregation suffers from 79% (95% CI: 68-93%) loss of h^2_{RV}. Using unaggregated variants, 27 traits had h^2_{RV} > 5%, with height having the highest h^2_{RV} at 21.9% (95% CI: 19.0-24.8%). The total heritability, including common and rare variants, recovered pedigree-based estimates for 11 traits. RARity can estimate gene-level h^2_{RV}, enabling the assessment of gene-level characteristics and revealing 11, previously unreported, gene-phenotype relationships. Finally, we demonstrated that in silico pathogenicity prediction (variant-level) and gene-level annotations do not generally enrich for RVs that over-contribute to complex trait variance, and thus, innovative methods are needed to predict RV functionality.

Rare protein coding variants, herein defined as those having minor allele frequency (MAF) < 1% and referred to as RVs, represent an important and understudied component of non-Mendelian complex trait genetics[1]. Despite efforts to functionally characterize RVs, the biological impact of roughly 400 rare, putatively disruptive mutations carried by each individual remains largely unknown[2]. Classification of RVs is challenging, and current algorithms do not always correctly predict their pathogenic characteristics[3,4]. Indeed, existing tools to classify RVs have typically been trained on conditions of Mendelian inheritance[5–7], while most human phenotypes are complex and non-Mendelian in nature. This gives rise to an unmet need to assess existing classifications in the context of complex traits.

Genome-wide association studies (GWAS) have been fruitful for characterizing common variants with regards to complex traits; however, a similar approach lacks the statistical power to study rare variants, unless sample sizes or effect sizes are very large[7]. Consequently, to improve statistical power, RV association testing often relies on gene-level variant aggregation methods to perform gene-burden tests, or variance component tests such as Sequence Kernel Association Test (SKAT)[8] and its variations (e.g., SKAT-O or "Optimal SKAT"). A limitation of gene-burden tests is the assumption that all RVs influencing a trait are homogeneous in terms of direction and magnitude of effects[7]. SKAT, on the other hand, aggregates the associations between variants and the phenotype through a kernel matrix[8], is a powerful tool for

[1]Population Health Research Institute, David Braley Cardiac, Vascular and Stroke Research Institute, Hamilton Health Sciences and McMaster University, Hamilton, Canada. [2]Department of Pathology and Molecular Medicine, McMaster University, Michael G. DeGroote School of Medicine, Hamilton, Canada. [3]Peter Boris Centre for Addictions Research, St. Joseph's Healthcare Hamilton, Hamilton, Canada. [4]Department of Psychiatry and Behavioural Neurosciences, McMaster University, Hamilton, Canada. [5]Department of Medicine, Faculty of Health Sciences, McMaster University, Hamilton, Canada. [6]Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, Canada. [7]Thrombosis and Atherosclerosis Research Institute, David Braley Cardiac, Vascular and Stroke Research Institute, Hamilton, Canada. e-mail: pareg@mcmaster.ca

Article https://doi.org/10.1038/s41467-024-45407-8

association testing in the presence of variants acting in opposing directions, but can be less powerful than burden tests when most variants are causal, and effects are in the same direction. SKAT-O combines both burden and SKAT to overcome the limitations of SKAT but can be slightly less powerful than burden or variance-component tests if their assumptions are held. While SKAT and SKAT-O are solely designed for association testing, variant aggregation in gene-burden testing may be used in any scenario requiring a gene to be treated as a single unit. However, the impact of RV aggregation on phenotypic variance explained has never been empirically evaluated.

RVs are postulated to contribute significantly to the "missing heritability" of complex traits, i.e. rare coding variants, when combined with common variants (CVs) may help recover the difference between current SNP-based heritability and heritability estimates from the traditional pedigree-based studies[9,10], yet this hypothesis has only been assessed for a handful of traits[11–17]. The RV contribution to narrow-sense heritability, h^2_{RV}, defined as the proportion of phenotypic variance attributable to their additive genetic effects, has been estimated in several recent studies[11–17]. However, these studies are limited by either the use of genotyping array data, small sample size, or models that make specific assumptions about the underlying genetic architecture. RV heritability estimates utilizing genotype data[12,13,15,16] are limited by the selection of rare genetic variation captured on the array, and algorithms to impute missing variants resulted in low accuracy for truly rare alleles[16,18]. As next-generation sequencing, particularly, whole exome sequencing (WES), are becoming a common place to accurately detect RVs[11,14,17], methods to evaluate RV heritability that do not rely on assumptions about the genomic architecture are needed.

We propose an approach to estimate RV narrow-sense heritability (h^2_{RV}), the Rare-variant heritability (RARity) estimator. A common strategy to estimate heritability is by assuming a random genetic effect model, whereby each (unknown) causal variant contributes to the total phenotypic variance according to a statistical distribution that might depend on its MAF, linkage disequilibrium (LD) with nearby variants, and functional properties, among others. These models aim to evaluate the overall contribution from a large number of variants, without identifying variant-specific effects[19]. RARity, in contrast, is based on multiple linear regression to estimate heritability, where the genetic effects are estimated conditional on the observed genotypes, and thus does not make any assumptions about the distribution of individual genetic effects nor the joint distribution of genotypes. The RARity framework simultaneously evaluates a large number of smaller regions of the chromosomes, referred to as blocks. This step is necessary to avoid estimation of high-dimensional linear models as the large number of rare variants would otherwise make matrix calculations intractable even with modern computational capabilities. An important and critical feature of RARity is the pruning of variants to minimize inflation in heritability due to LD spillage between blocks (see Methods).

The availability of large WES data from the UK Biobank (UKB) provides a unique opportunity to study the overall and gene-level contribution of RVs to complex trait heritability. We hypothesize that the study of exome-wide and gene-level RV heritability will provide insights into the functional characteristics of RVs. Specifically, by estimating the contribution of rare coding variants to narrow-sense heritability in unrelated individuals, our objectives were to (1) characterize the variant-level characteristics (allele frequency, disruptiveness according to a variety of algorithms, and clinical pathogenicity) that best predict phenotypic variations, (2) characterize the gene-level properties (membership to gene sets and biological pathways, evolutionary constraint, gene length) that best predict biological effects and, (3) identify genes associated with the traits.

Using WES data from UK Biobank (n = 167,348), we report h^2_{RV} for 31 complex continuous traits, including 26 biomarkers and 5 anthropometric traits, and we demonstrate the utility of RARity estimator to understand whether existing in silico pathogenicity prediction (variant-level) and gene-level annotations could be enriched for RVs that disproportionately contribute to the complex trait. This study has major implications for our understanding of the genetic architecture of complex traits in the context of RV functions, which we expect would ultimately facilitate the discovery of new disease pathogenesis.

Results

Overview and testing of the RARity method

The RARity method entails parallel computing of the adjusted R^2 based on an ordinary least square (OLS) multiple linear regression as an unbiased estimator of block-wise heritability for each consecutive genetic block. Adjusted R^2 estimates are then summed over all blocks as the overall heritability estimate. Overview of RARity is shown in Fig. 1 and technical details provided in Methods and Supplementary Fig. 1. The current sample size provided at least 80% power to detect 4% h^2_{RV}, at an empirical type-I error rate of 0.05 (Supplementary Fig. 2). Extensive simulations were performed using real genotype data to identify an approach for estimating heritability that is robust under realistic scenarios. Through this endeavour, we discovered that h^2_{RV} could be affected by the presence of long-range LD (LRLD), occurring at a much greater distance than what is observed for CVs[20] (Supplementary Fig. 3). LRLD complications were controlled by using a stringent LD threshold over a large, empirically derived window size ($r^2 > 0.1$, window size = 50 Mb, step size = 500 bases). This stringent pruning method was enforced in all subsequent analyses involving RVs to ensure a well-calibrated estimate of h^2_{RV}. The simulation studies on the effects of varying MAFs, heritability, proportion of causal genes or variants indicated that RARity is largely unbiased but tends to underestimate when the number of causal variants or genes is low (<1%; Supplementary Fig. 3).

Comparison of gene-burden, gene-wise, and exome-wide heritability

To estimate the amount of information lost when aggregating rare variants within each gene, we compared h^2_{RV} estimates using the gene-burden approach to exome-wide estimations for 31 complex traits using RARity. We created blocks of genotype data in the following manners: (1) gene-burden blocks, derived by summing the number of rare alleles within each gene for an individual, which produced a single block containing all gene-burden scores as predictors; (2) gene-wise blocks, consisted of un-aggregated RVs partitioned by gene, such that each block contained all the variants within a single gene; (3) exome-wide blocks were created by partitioning RVs in each chromosome into blocks of ~5000 adjacent RVs. RV heritability estimates were then derived from each type of block construct using RARity. Design and applications of RARity, using these constructs is provided in Fig. 1, and a detailed computational pipeline is illustrated in Supplementary Fig. 1.

Our results show that the overall estimated RV heritability from gene-burden blocks ($h^2_{RV\text{-}burden}$) is on average 79.3% (95% CI: 76.5%–82.0%) less than heritability based on either gene-wise ($h^2_{RV\text{-}gene\text{-}tot}$), or exome-wide blocks (h^2_{RV}; Fig. 2, Supplementary Table 3). Because gene-level estimates ($h^2_{RV\text{-}gene}$) are useful for secondary analyses, we further tested whether $h^2_{RV\text{-}gene\text{-}tot}$ could in fact substitute for h^2_{RV} from exome-wide blocks. A potential caveat of using gene-wise heritability to determine total heritability is the possibility of LD between variants in two or more genes inflating the total heritability, as would be the case when genes overlap. We observed consistent results between $h^2_{RV\text{-}gene\text{-}tot}$ and h^2_{RV} (Fig. 2b and Supplementary Table 3), with $h^2_{RV\text{-}gene\text{-}tot}$ being 1.5% (95% CI: 0.0%–4.0%) higher than h^2_{RV} on average, indicating only a slight inflation that could be due to chance. However, the advantage of gene-level blocks is the smaller number of variants per block, which makes it computationally less intensive and 3× faster to compute than h^2_{RV}.

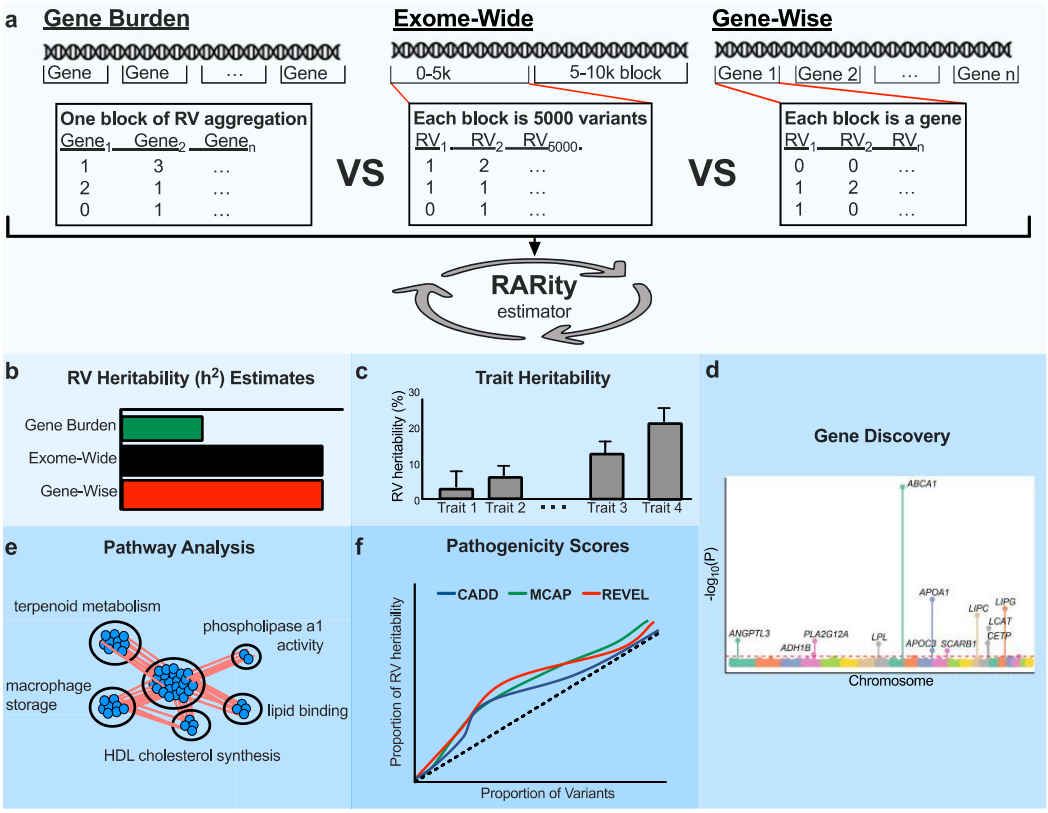


Fig. 1 | A schematic diagram illustrating the design and applications of Rare variant heritability (RARity) estimator. Following the initial quality control steps (Methods), the protein-altering and LoF RVs in each chromosome were LD pruned and partitioned into consecutive regions, referred to as genotype blocks. **a** Diagrammatic representation of the three genotype block constructs implemented by RARity. Applications of RARity included: **b** Comparison of the block constructs to empirically test the loss of information when aggregating rare variants. **c** Rare variant heritability estimates of complex traits using exome-wide blocks. **d** Gene-based RV heritability to discover gene-phenotype associations. **e** Network and pathway analyses of the genes with significant RV heritability to provide insights into their biological relevance. **f** Enrichment of RV heritability by pathogenicity scores. Three pathogenicity scores, CADD, M-CAP and REVEL were applied to retain progressively more deleterious variants and tested on all 31 traits for enrichment in RV heritability. This figure was generated using GraphPad Prism version 6.04 for Windows[66].

RV contribution to heritability estimates for 31 complex traits

Here we utilized two genetic datasets from the UK Biobank: RVs were derived from the WES data (UKB data field: 23155) and CVs were extracted from the imputed genotype data (data field: 22418; v3 release). The quality control steps for each dataset are provided in Methods. Applying RARity using the exome-blocks (MAF < 0.01), we estimated RV heritability for 31 complex traits, with 29 traits showing meaningful contribution at a nominal significant threshold of 0.05, 27 traits showing $h^2_{RV} > 5\%$ and height having the highest h^2_{RV} at 21.9% (95% CI: 19.0%–24.8%) (Table 1; Fig. 3, and Supplementary Table 3). The lower overall heritability estimate for glucose, 1.8% (95% CI: −1.0%–4.7%), observed in our study was most likely due to sample collection in non-fasting states, as opposed to the fasting glucose level used for heritability estimates in the pedigree studies[21]. Sex stratification, performed on all 31 traits, showed some heterogeneity in h^2_{RV} between the sexes (Supplementary Fig. 4), but the apparent differences were statistically nonsignificant (p-value > 0.05).

Since RARity poses no upper restriction on the MAFs, we additionally assessed the contribution CV (h^2_{CV}) and the combined CV and RV (h^2_{tot}) to these 31 traits by concatenating common and rare variants (Methods). We observed that h^2_{CV} estimated using RARity was consistent with BOLT[22], but higher than LDSC[23] (Supplementary Fig. 5). Although common variants contributed more to overall heritability, as compared to RVs (Table 1), estimated h^2_{RV} generally increased proportionally to h^2_{CV} (Supplementary Fig. 6), except for height, alkaline phosphatase and Lp(a). Lp(a) particularly stands out with a much higher h^2_{CV} in relation to h^2_{RV}. The low concordance between CV and RV heritability in Lp(a) may be due to the unique genetic architecture of this trait, with most genetic variance attributed to the *LPA* locus itself and the highly polymorphic kringle IV type 2 copy number variation having an outsized impact on concentration[24].

The estimated h^2_{tot} for 11 of the 31 traits were consistent with previously reported heritability from pedigree or twin-based studies (Supplementary Data 1). The difference between h^2_{RV} and h^2_{RV}-adjusted-for-h^2_{CV} were heterogenous (Table 1), indicating that the degree of tagging or LD between the rare and common variants can vary by trait. For example, h^2_{tot} (from CV and RV) was 87.8% (95% CI: 84.5–91.1%) for height and 22.5 (95% CI: 19.3–25.7%) for albumin, with height showing a greater degree of LD between RVs and CVs, as compared to albumin, determined by a 5.3% reduction in h^2_{RV} for height, but only 0.6% reduction in h^2_{RV} for albumin following adjustment for h^2_{CV} (Table 1).

 https://doi.org/10.1038/s41467-024-45407-8

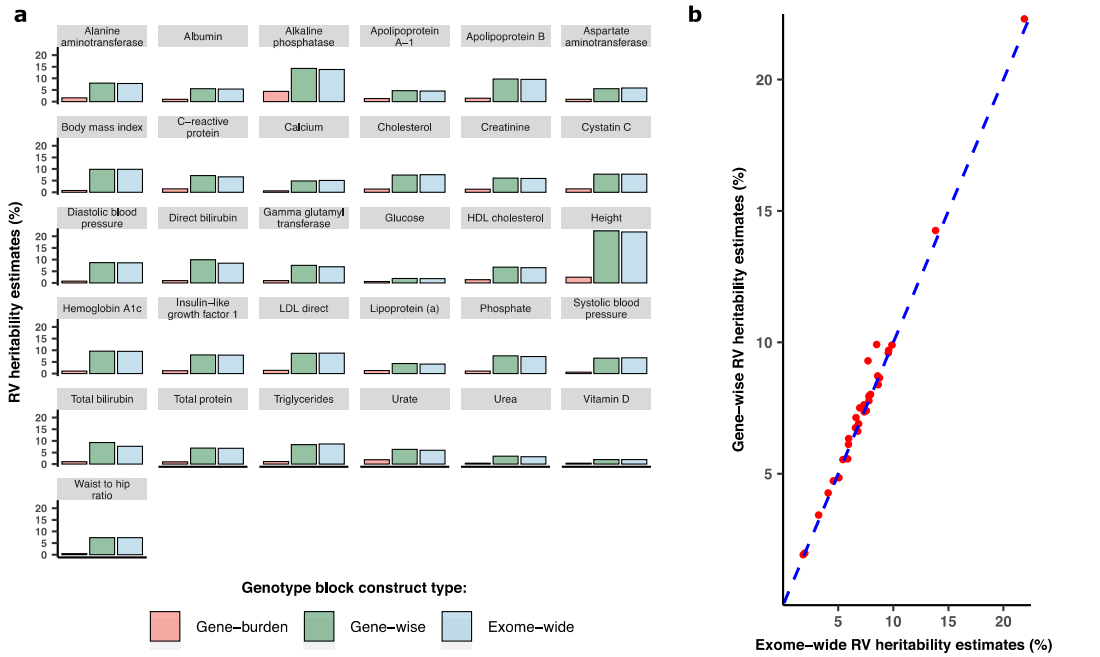


Fig. 2 | Effect of aggregating variants on RV heritability estimates. a Comparison of RV heritability estimates derived from aggregation of variants in gene-burden blocks, exome-wide blocks (blocks of 5000 unaggregated variants) and gene-wise blocks with un-aggregated variants. The *y*-axis corresponds to the rare coding variant contribution to percentage of heritability estimates for each trait. **b** Correlation between heritability estimates from gene-wise blocks represented on the *y*-axis versus exome-wide blocks, represented on the *x*-axis.

RV heritability to characterize pathogenicity scores

Various in silico tools have been developed to predict the effects of rare coding variants on risk of Mendelian disorders. Whether such pathogenicity tools are useful for prioritizing variants contributing to complex trait variance in the general population remains uncertain. As such, we tested the association between commonly used pathogenicity scores, namely Combined Annotation-Dependent Depletion (CADD), Mendelian Clinically Applicable Pathogenicity (MCAP), and Rare Exome Variant Ensemble Learner (REVEL), and the fraction of trait RV heritability explained, hypothesizing that variants predicted to be more deleterious account for more trait variance explained and vice versa.

For most traits, the fraction of heritability explained by an increasing proportion of RVs (added from the highest to the lowest level of predicted pathogenicity), was largely uniform and independent of pathogenicity score (sub-plot Fig. 4, and Supplementary Fig. 7), indicating little enrichment of trait-associated RVs by pathogenicity score. Across traits (Supplementary Fig. 7), the average RV heritability explained by the top 25% most deleterious variants was slightly higher than expected at 36% (CADD 34%; MCAP 36.7%; REVEL 37.2%). However, the top 50% of most deleterious variants accounted for 42.9% of heritability explained, which is lower than the fraction of heritability explained by the bottom 50% set of variants (54%). The magnitude of enrichment by pathogenicity score also varied by trait, for example, the top 25% most deleterious variants by M-CAP score explained as little as 32.3% RV heritability for height but as much as 48.4% RV heritability for ApoA-I. Further, the allele frequency of variants seemed to have very little impact on the magnitude of enrichment, as we observed similar results across the three MAF categories (<0.001, <0.005, <0.01; Fig. 4, Supplementary Fig. 7). The specified MAF cut-offs were used instead of MAF-bins as the $h^2_{\text{RV-gene-tot}}$ in these bins ($0.01 > \text{MAF} \geq 0.005$, $0.005 > \text{MAF} \geq 0.001$, and $0.01 > \text{MAF} \geq 0.001$) were <5% (Supplementary Table 4) and consequently would have produced unreliable results when further stratified by pathogenicity scores.

Characterizing genes based on heritability estimates

To investigate gene-level characteristics of RVs, we used RARity to determine $h^2_{\text{RV-gene}}$ for all genes with qualifying variants and derived corresponding *p*-values for each gene using an *F*-test. After Bonferroni correction (p-value $< 2.75 \times 10^{-6}$), 152 of the 18,214 genes had significant $h^2_{\text{RV-gene}}$ for one or more traits (herein referred to as significant $h^2_{\text{RV-gene}}$), representing 218 distinct gene-biomarker relationships. A list of these genes with the corresponding $h^2_{\text{RV-gene}}$ and *p*-values are presented in Supplementary Data 2. We identified many genes that recapitulated previously reported associations and discovered 11 previously unidentified gene-traits relationships. Some examples of well-established relationships with significant $h^2_{\text{RV-gene}}$ include *PCSK9*, which is known to regulate Apolipoprotein B (ApoB) and Low density cholesterol (LDL); *MC4R* gene affecting body mass index (BMI), and the association of *LPL* with Apolipoprotein A-I (ApoA-I) levels. Some of the previously unidentified relationships include *PPARA* with ApoB, *TFAM* with alkaline phosphatase, *TMEM43* with hemoglobin A1c (HbA1c), and *NR1I2* with total bilirubin. A list of the significant heritability genes, highlighting the gene-biomarker relationships is provided in Table 2. In addition, Manhattan plots (Fig. 5) are included for 6 randomly selected traits: Alkaline phosphatase, HbA1c, Low density lipoprotein direct (LDL), Insulin like growth-factor 1 (IGF-1), and ApoA-I.

We further investigated the role of the 152 genes contributing significantly to trait heritability (p-value $< 2.75 \times 10^{-6}$) in diseases and

Table 1 | Heritability estimates derived from rare coding and common variants

Traits	h^2_{RV} (%) (95% CI)	h^2_{CV} (%) (95% CI)	h^2_{tot} (%) (95% CI)	h^2_{RV} adjusted for h^2_{CV} (%) (95% CI)	Difference between h^2_{RV} and h^2_{RV} adjusted for h^2_{CV} (%)
Albumin	5.4 (2.6,8.3)	17.7 (15.5,20.0)	22.5 (19.3,25.7)	4.8 (1.7,7.8)	0.6
Alkaline phosphatase	13.8 (11.0,16.7)	36.1 (33.9,38.4)	46.8 (43.6,50.1)	10.7 (7.6,13.8)	3.1
Alanine aminotransferase	7.8 (4.9,10.6)	17.1 (14.8,19.3)	23.0 (19.7,26.2)	5.9 (2.8,9.0)	1.9
Apolipoprotein A	4.6 (1.7,7.4)	27.1 (24.8,29.4)	29.9 (26.6,33.1)	2.8 (−0.3,5.8)	1.8
Apolipoprotein B	9.6 (6.7,12.5)	38.9 (36.6,41.2)	44.0 (40.8,47.3)	5.1 (2.0,8.2)	4.5
Aspartate aminotransferase	5.9 (3.0,8.7)	19.9 (17.6,22.1)	24.0 (20.8,27.3)	4.2 (1.1,7.2)	1.7
Body mass index	9.9 (7.0,12.8)	31.8 (29.6,34.1)	39.5 (36.3,42.8)	7.7 (4.6,10.7)	2.2
Calcium	5.1 (2.2,7.9)	16.1 (13.9,18.4)	20.7 (17.5,24.0)	4.6 (1.6,7.7)	0.4
Cholesterol	7.6 (4.7,10.4)	28.9 (26.6,31.2)	32.2 (29.0,35.4)	3.3 (0.2,6.3)	4.3
Creatinine	5.9 (3.1,8.8)	28.6 (26.3,30.9)	32.6 (29.3,35.8)	3.9 (0.9,7.0)	2.0
C-reactive protein	6.6 (3.8,9.5)	27.3 (25.0,29.6)	33.7 (30.5,37.0)	6.4 (3.4,9.5)	0.2
Cystatin C	7.8 (4.9,10.7)	33.5 (31.3,35.8)	38.9 (35.7,42.2)	5.4 (2.3,8.5)	2.4
Diastolic Blood Pressure	8.6 (5.7,11.4)	26.1 (23.9,28.4)	34.6 (31.4,37.8)	8.5 (5.4,11.5)	0.1
Direct bilirubin	8.5 (5.6,11.4)	35.0 (32.7,37.3)	39.4 (36.1,42.6)	4.4 (1.3,7.4)	4.1
Gamma glutamyl transferase	7.0 (4.1,9.8)	29.4 (27.1,31.7)	35.0 (31.7,38.2)	5.6 (2.5,8.6)	1.4
Glucose	1.8 (−1.0,4.7)	9.8 (7.6,12.1)	12.5 (9.3,15.7)	2.7 (−0.4,5.7)	-0.8
Height	21.9 (19.0,24.8)	71.3 (69.0,73.6)	87.8 (84.5,91.1)	16.5 (13.5,19.6)	5.4
Hemoglobin A1c	9.5 (6.7,12.4)	31.4 (29.1,33.6)	38.3 (35.1,41.6)	7.0 (3.9,10.0)	2.6
HDL cholesterol	6.6 (3.7,9.4)	31.5 (29.2,33.7)	36.2 (32.9,39.4)	4.7 (1.6,7.8)	1.9
IGF-1	7.9 (5.1,10.8)	28.1 (25.9,30.4)	33.8 (30.5,37.0)	5.6 (2.6,8.7)	2.3
LDL direct	8.7 (5.9,11.6)	32.4 (30.2,34.7)	36.7 (33.5,39.9)	4.2 (1.2,7.3)	4.5
Lipoprotein (a)	4.1 (1.2,6.9)	61.1 (58.9,63.4)	61.6 (58.4,64.8)	0.5 (−2.6,3.5)	3.6
Phosphate	7.3 (4.5,10.2)	17.1 (14.8,19.4)	23.3 (20.1,26.6)	6.2 (3.2,9.3)	1.1
Systolic Blood Pressure	6.8 (3.9,9.6)	25.4 (23.1,27.7)	33.4 (30.2,36.7)	8.0 (5.0,11.1)	−1.3
Total bilirubin	7.7 (4.8,10.6)	40.3 (38.0,42.6)	44.3 (41.1,47.6)	4.1 (1.0,7.1)	3.6
Total protein	6.9 (4.0,9.7)	23.0 (20.7,25.2)	29.4 (26.1,32.6)	6.4 (3.3,9.5)	0.4
Triglycerides	8.6 (5.8,11.5)	26.3 (24.0,28.5)	31.9 (28.6,35.1)	5.6 (2.5,8.7)	3.0
Urea	3.2 (0.4,6.1)	17.1 (14.8,19.4)	20.0 (16.8,23.2)	2.9 (−0.2,6.0)	0.3
Urate	6.0 (3.1,8.8)	28.0 (25.7,30.3)	32.5 (29.3,35.8)	4.5 (1.5,7.6)	1.4
Vitamin D	2.0 (-0.9,4.8)	13.3 (11.0,15.5)	15.0 (11.7,18.2)	1.7 (−1.4,4.7)	0.3
Waist to hip ratio	7.4 (4.5,10.2)	21.0 (18.7,23.2)	28.2 (25.0,31.5)	7.2 (4.2,10.3)	0.1

CI confidence interval, h^2_{CV} heritability estimates due to common variants, h^2_{RV} heritability estimates due to rare coding variants, h^2_{tot} heritability estimates due to combined rare coding and common variants.

biological pathways. DisGenet[25] revealed 115 of the target genes influencing 2137 disease pathways (Supplementary Data 3). A heatmap of the diseases associated with the significant genes for ApoB is presented in Fig. 6a. Interestingly, searching through the Drug Gene Interaction database (DGIdb)[26] revealed that 93 of the 152 of the target genes (i.e., 61%) belong in the "druggable genome" category, 16 of which are "clinically actionable", including *APOB, FGFR3* and *LDLR* (Supplementary Data 4). The target genes also appear to be significantly overrepresented in many biologically relevant pathways (multiple-testing correction via the g:Profiler g:SCS algorithm, adjusted p-value < 0.05). For example, the genes contributing significantly towards h^2_{RV} of ApoB results in enrichment of 156 pathways, all of which are highly interconnected in an elaborate network and includes well-known pathways such as the LDL receptor binding pathways and pathways related to atherosclerosis (Fig. 6b, Supplementary Data 5).

We explored whether $h^2_{RV\text{-}gene}$ is associated with gene-length and evolutionary constraint, where gene length was derived from the RefSeq transcripts with the greatest length and gene-level evolutionary constraint was determined using gnomAD pLoF Metrics[27]. For most traits, neither gene-length (Supplementary Fig. 8, Supplementary Table 5) nor evolutionary constraint (Supplementary Fig. 9, Supplementary Table 6) was significantly associated with $h^2_{RV\text{-}gene}$. Height was the only trait where $h^2_{RV\text{-}gene}$ was significantly associated with gene-length (0.24% variance explained, p-value = 3.3×10^{-11}) and evolutionary constraint (0.27% variance explained, p-value = 1.5×10^{-11}). Evolutionary constraint was suggestively associated with the $h^2_{RV\text{-}gene}$ of BMI (p-value = 3.7×10^{-4}) and waist-to-hip ratio (p-value = 9.4×10^{-3}), but no other significant association was observed. The highly conserved gene cluster regions, with short repeats, such as the hox, histone, protocadherin and hemoglobin gene clusters did not contribute significantly to h^2_{RV} (Supplementary Table 7). The overall heritability estimates, $h^2_{RV\text{-}gene\text{-}tot}$, were consistent between genes that are transcribed from either the positive or the negative strand (Supplementary Fig. 10), which is not surprising, considering that there are nearly equal number of genes on either strand.

Discussion

We established a method, the RARity estimator, to accurately estimate the contribution of RVs to the heritability of complex quantitative traits, and to characterize the gene-level and variant level characteristics of RVs, as demonstrated with results from 31 continuous traits form the UK Biobank. RARity is a versatile method, as it does not make any prior assumptions about the genetic architecture of the selected variants, making it applicable to both common and rare variants. Calibration of RV contributions was empirically confirmed using

Article https://doi.org/10.1038/s41467-024-45407-8

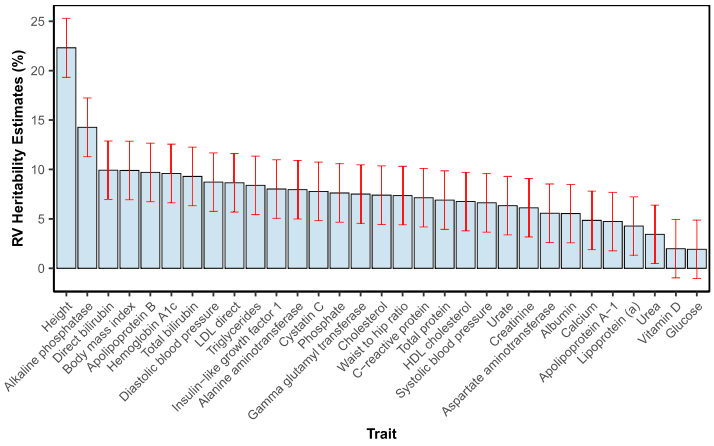


Fig. 3 | Estimated phenotypic variance explained by RVs across 31 complex traits. Bar chart illustrating the exome-wide RV heritability estimates (h^2_{RV}) +/− 95% confidence interval. Contribution of rare coding variants on 31 complex traits were based on $n = 167{,}348$ Caucasian individuals from the UK Biobank, as estimated using RARity via the exome-wide block construction. Traits were standardized for age, sex, and the first 20 genetic principal components. 95% confidence interval of trait h^2_{RV} is denoted with red, vertical error bars.

extensive simulation studies and proved to produce robust estimates across a wide range of analytical scenarios.

The performance of RARity for CV contributions was comparable with BOLT. The estimated h^2_{CV} by LDSC on the other hand was lower than RARity for each trait. While LDSC is a powerful tool that allows estimation h^2_{CV} from GWAS summary-statistics, with adequate control for population stratification, it is prone to underestimation as discussed in several studies[28–31]. Comparison of the h^2_{CV} for height and BMI, with other SNP-heritability methods in the literature, such as LDSC[23], LDAK[32], GCTA[33], GRE[19] and SumHer[34], shows that our estimates are on the higher end (but not the highest) of what has been historically reported in comparison studies[19,30]. The apparent higher h^2_{CV} for height and BMI using RARity may be a characteristic of the subset of the population selected in this study. All current methods make assumptions about the genetic architecture, such as the effect size, variance, LD/MAF ratios, etc., since the true genetic architecture is unknown, it remains unclear which estimates in the literature are reliable. More importantly, none of these methods are suitable for estimation of rare variant heritability.

It was observed that for most traits, RVs account for a noticeable proportion of trait heritability independent of CV contribution. One of the technical challenges in combining CV and RV contribution is the presence of LD and its interplay with allele frequencies. The results suggested that the amount of LD between CVs and RVs is heterogeneous, and consequently, their independent contribution to heritability is trait dependent. Indeed, the "missing heritability" can be almost perfectly recapitulated by incorporating RVs for 11 of the 31 traits, including height and BMI (Supplementary Data 1), reaffirming the conclusion from a recent study that used whole genome sequencing (WGS) data[14]. Meanwhile, the apparent lack of recovery of the pedigree-based heritability for other traits may be due to the inherent characteristics of the pedigree-based studies. Generally, pedigree-based studies report higher heritability as compared to SNP-based heritability from population studies[21]. This gap is most likely due to limited sample size in the early genome-wide association studies, variability in sample collection, population characteristics, the exclusion of sex chromosomes[35], rare intronic, structural and non-coding regulatory variants, in addition to the non-additive effects which are captured by pedigree-based analysis but not necessarily captured by RARity, nor by the most commonly used CV heritability models.

Several methods have been used to increase the power of detecting gene-trait associations, the most popular methods being gene burden testing and Sequence Kernel Association Test (SKAT)[8]. Since SKAT aggregates the associations between variants and the phenotype through a kernel matrix[8] it is solely designed to test the strength of association via p-values without providing an effect size, and thus rendering it incomparable to RARity. The variant aggregation method used in gene-burden testing do not have such a limitation, and consequently, we were able to examine the effect of aggregating RVs on trait variance, which on average resulted in a 79.3% loss on the estimated h^2_{RV}. These results suggest that burden tests may have limited ability to predict traits. On the other hand, the use of unaggregated variants in RARity, not only captures more genetic variance, but also offers the practical advantage of characterizing genes based on $h^2_{RV\text{-}gene}$. Furthermore, with fewer variants per block, $h^2_{RV\text{-}gene}$ can be computed more efficiently to obtain the total heritability and still produce consistent results with those derived using larger blocks of exome data (Fig. 2b).

Next, we leveraged heritability estimates to assess the performance of RV pathogenicity algorithms in the context of complex traits. We see very modest, if any, enrichment of h^2_{RV} when variants are filtered according to pathogenicity scores. This is likely because the pathogenicity scores for RVs currently used (MCAP, CADD, REVEL, etc.) are largely based on Mendelian diseases, have a modest impact on complex trait h^2_{RV}, and are limited in distinguishing the variants with any biological effects from those that do not. This affirms our conclusion that there is a need for alternative methods to study the functional consequences of RVs.

One of the useful features of RARity is that it can be tailored to identify genes significantly enriched for heritability and help with gene discovery and characterization. We show this by identifying 152 genes significantly enriched for heritability. To the best of our knowledge, 11 of these associations have not been previously described through exome wide association studies or GWAS (Table 2, Supplementary Data 2). We further demonstrate that the 152 genes with significant $h^2_{RV\text{-}gene}$ across 31 traits are enriched in various biological and disease pathways. For example, 11 significant genes contributing to heritability of ApoB are involved in pathways ranging from abnormal arterial stenosis to chylomicron, LDL and lipoprotein clearance, as well as hyperlipidemia and coronary artery diseases (Fig. 6, Supplementary

Article https://doi.org/10.1038/s41467-024-45407-8

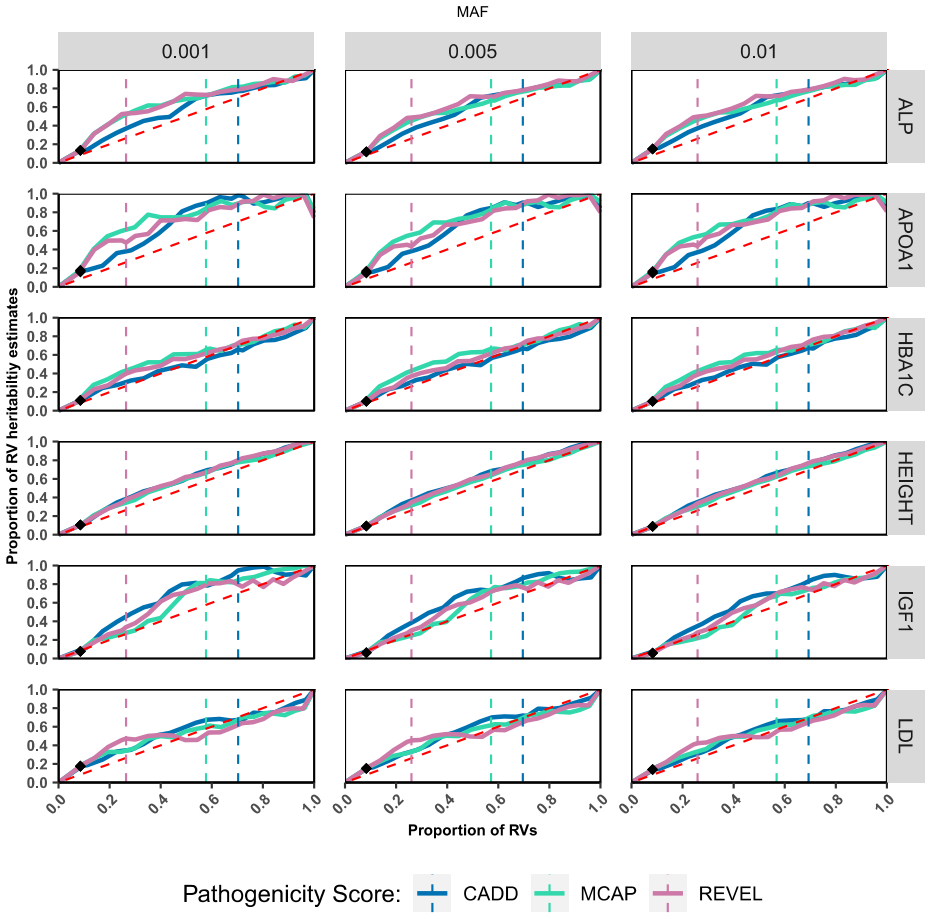


Fig. 4 | Impact of pathogenicity scores on variance explained by RVs for 6 representative traits. Plots illustrating the proportion of RV heritability explained (*y*-axis) as a function of incorporating increasingly "deleterious" genetic variants (*x*-axis). Proportion of heritability estimates is the fraction of estimates in relation to all protein altering and LoF variants within the MAF categories. The vertical, dashed lines represent the binary thresholds recommended to define pathogenicity for CADD (blue), M-CAP (green) and REVEL (magenta). The black diamond marks the point of last inclusion of LoF variants, which were prioritized before missense mutations. The diagonal dashed red lines represent the scenario wherein RVs uniformly contribute to h^2_{RV}, irrespective of pathogenicity score. ALP alkaline phosphatase, HBA1C hemoglobin A1c, LDL low density lipoprotein direct, IGF1 insulin line growth-factor 1, APOA1 apolipoprotein A-I.

Data 4 and 5). Interestingly, previous association studies did not identify the *PPARA* gene as genome-wide significant, even though *PPARA* significantly contributes to the h^2_{RV} of ApoB and that both *PPARA* and *APOB* are involved in several lipid-related pathways (Supplementary Data 5), fatty liver disease, and dyslipidemias (Fig. 6), and identified as "druggable genome" in the DGIdb database (Supplementary Data 3). Notably, *PPARA* is also the target of the lipid-lowering drug class known as fibrates[36], making it a strong candidate for further pharmacogenomic studies. Indeed, most of the significant genes play important roles in disease etiology, as observed through DisGenet analyses. Further examples include the contribution of *TFAM* on alkaline phosphatase heritability and its role in hepatocerebral mitochondrial DNA depletion syndrome, and the contribution of *TMEM43* on HbA1c heritability and macular degeneration. These results imply that genes involved in diseases are also likely to contribute significantly to biomarker heritability. However, whether these associations can be used as markers of pathogenic mutations or represent causal mediations through biomarker concentrations will require further investigations.

As an application of the gene-wise blocks, we observed that for most traits, longer genes do not explain higher trait variance, this is likely because the relative paucity of larger genes obscuring any underlying relationship[16]. Our observations also show a modest increase in $h^2_{RV\text{-}gene}$ with higher RV evolutionary constraint (indicated by lower LOEUF scores[27]) for only a few traits, and from this, we can conclude that RVs contributing to the variance in complex traits are likely well-tolerated and are not selected against in general population.

This study has several limitations. First, exclusion of non-coding, singleton, and doubleton variants, as well as LD pruning of potentially functional variants, may lead to an underestimate of the h^2_{RV}. Although, numerous studies have shown that rare coding variants have major functional impacts in direct way[37], it is quite possible that some of the observed h^2_{RV} is due to the coding RVs being in LD with the non-coding variants that are not represented in WES. On the other hand,

Table 2 | Genes with significant heritability estimates for each trait

Trait	Genes with significant RV heritability estimates	Not reported in exome-wide associations in UKB**	Not reported in GWAS Catalogue
Albumin	*ALB; FCGRT; TBC1D2B*		
Alkaline phosphatase	*ALDH5A1; ALPL; ASGR1; GBGT1; GPLD1; HSPG2; NBPF3; TDP2; TFAM; ZNF800*	***TFAM***	***TFAM***
Alanine aminotransferase	*GPT; MFSD3*		*MFSD3*
Apolipoprotein A-I	*ABCA1; ADH1B; ANGPTL3; APOA1; APOC3; CETP; LCAT; LIPC; LIPG; LPL; PLA2G12A; SCARB1*	*CETP*	
Apolipoprotein B	*APOB; APOE; BCAM; CEACAM20; CLASRP; LDLR; NECTIN2; NKPD1; PCSK9; PPARA; ZNF229*	***PPARA***	***PPARA***
Aspartate aminotransferase	*ANO5; GOT1*		
BMI	*MC4R*		
Calcium	*ALB; CASR; FCGRT*		
Cholesterol	*ABCA1; ABCG5; ANGPTL3; APOB; FGB; JAK2; KHDRBS2; LDLR; LIPG; NECTIN2; NKPD1; PCSK9; ZNF229*	***KHDRBS2****; ZNF229*	***KHDRBS2***
Creatinine	*LRP2; SLC22A2; SLC22A7*		
C-reactive protein	*ABCA1; APCS; CRP; JAK1; TMED8*	***TMED8***	***TMED8***
Cystatin C	*CGNL1; CST3; SH2B3*		
Direct bilirubin	*ATG16L1; DGKD; DNAJB3; HJURP; MROH2A; NR1I2; SAG; SLCO1B1; SLCO1B3; SLCO1B3-SLCO1B7; TRPM8; UGT1A1; UGT1A10; UGT1A3; UGT1A4; UGT1A5; UGT1A6; UGT1A7; UGT1A8; UGT1A9; USP40*	*NR1I2; SLCO1B3-SLCO1B7*	
Gamma glutamyl transferase	*A1CF; GGT1; LRRC75B; RORC; SYNJ2*		
Glucose	*G6PC2; GCK*		
Hemoglobin A1c	*ADGRE5; APEH; CTU2; G6PC2; GCK; JAK2; PFKL; PFKM; PIEZO1; RHAG; SPTB; TMC8; TMEM43*	***ADGRE5; TMEM43***	***ADGRE5****; APEH; JAK2;****TMEM43***
HDL cholesterol	*ABCA1; APOA1; APOA5; APOC3; CETP; LCAT; LIPC; LIPG; LPL; NR1H3; PLA2G12A; SCARB1*		
Height	*CRISPLD2; DDR2; FGF2; FGFR3; GH1; GHRH; GRAMD2A; HAPLN3; IHH; NPR2; NPR3; SCMH1; STC2; ZFAT*	*GH1; GHRH*	
IGF-1	*GH1; IGFALS; IGFBP3; PARPBP; ZNF12*		
Low-density lipoprotein	*ABCG5; ANGPTL3; APOB; APOE; LDLR; NECTIN2; NKPD1; PCSK9; ZNF229*		
Lipoprotein(a)	*ACAT2; AGPAT4; ARID1B; EZR; FNDC1; IGF2R; LPA; MAP3K4; MAS1; MRPL18; PLG; PNLDC1; SLC22A1; SLC22A2; SLC22A3; SOD2; SYNJ2; SYTL3; TMEM181; TULP4; WTAP*	*ACAT2; ARID1B;****EZR****; SYNJ2; TULP4*	***EZR****; TMEM181*
Phosphate	*ALPL; CDR2; ENPP1; HLA-DPA1; TTK*	***CDR2; HLA-DPA1; TTK***	***CDR2; HLA-DPA1; TTK***
Total bilirubin	*ATG16L1; DGKD; DNAJB3; HJURP; MROH2A; NR1I2; SAG; SLCO1B1; SLCO1B3; SLCO1B3-SLCO1B7; TRPM8; UGT1A1; UGT1A10; UGT1A3; UGT1A4; UGT1A5; UGT1A6; UGT1A7; UGT1A8; UGT1A9; USP40*	***NR1I2;*** *SLCO1B3-SLCO1B7*	***NR1I2***
Total protein	*FCGR2B; FCGRT; SNX8; TNFRSF13B*		
Triglycerides	*ANGPTL3; APOA5; APOB; APOC3; LPL; PLA2G12A; SIK3; ZPR1*		
Urate	*LGALS13; PDZK1; SLC22A11; SLC22A12; SLC2A9; WDR1*		
Urea	*RBM47; CYP2R1; HAL; PDE3B*		
Vitamin D	*CYP2R1; HAL; PDE3B*		

Systolic blood pressure, diastolic blood pressure and waist-to-hip ratio did not have significant $h^2_{RV\text{-}gene}$. Bolded font indicates gene-trait relationships unidentified in previous studies. ** Wang, Q. et al.[64] and Backman, J.D. et al. [65].

very little is known about the rare non-coding regions, which are difficult to define and even more difficult to assess functionality[38], consequently there is no direct comparison of the contribution of rare-coding vs. non-coding variants in the literature. This is indeed a question of high importance and may be answered with WGS. We anticipate that RARity can be applied to WGS with further calibrations, to study effects of other variants not discussed here. One of the challenges of WGS data analysis is that there is no natural biological unit available in the intergenic regions. With RARity, there is no need for defining the biological units, as the estimates can be based on blocks that are agnostic of genetic borders.

Currently, the RARity model is fitted for to continuous traits. For dichotomous trait heritability, the model will require slight adjustments to transform the observed heritability to a liability scale heritability to account for the case-control imbalance in population study[39]. We did not find any significant difference in the RV heritability between males and females across the 31 quantitative traits studied (Supplementary Fig. 4). In contrast, SNP-based studies showed slight differences in heritability of selected traits between the sexes[40–42]. Since the analysis using RVs was limited by the reduced power, further research of larger sample sizes is required to better understand the sex effect on RV heritability. We have only explored a fraction of the pathogenicity scores and further research will benefit from using RARity as a tool to evaluate the algorithms to distinguish functional vs. non-functional variants. In addition, heritability describes the phenotypic variance at the population level, and thus cannot be generalized to a different population, and nor does it inform the prediction of phenotypic variations of an individual. Therefore, methods to estimate

Article https://doi.org/10.1038/s41467-024-45407-8

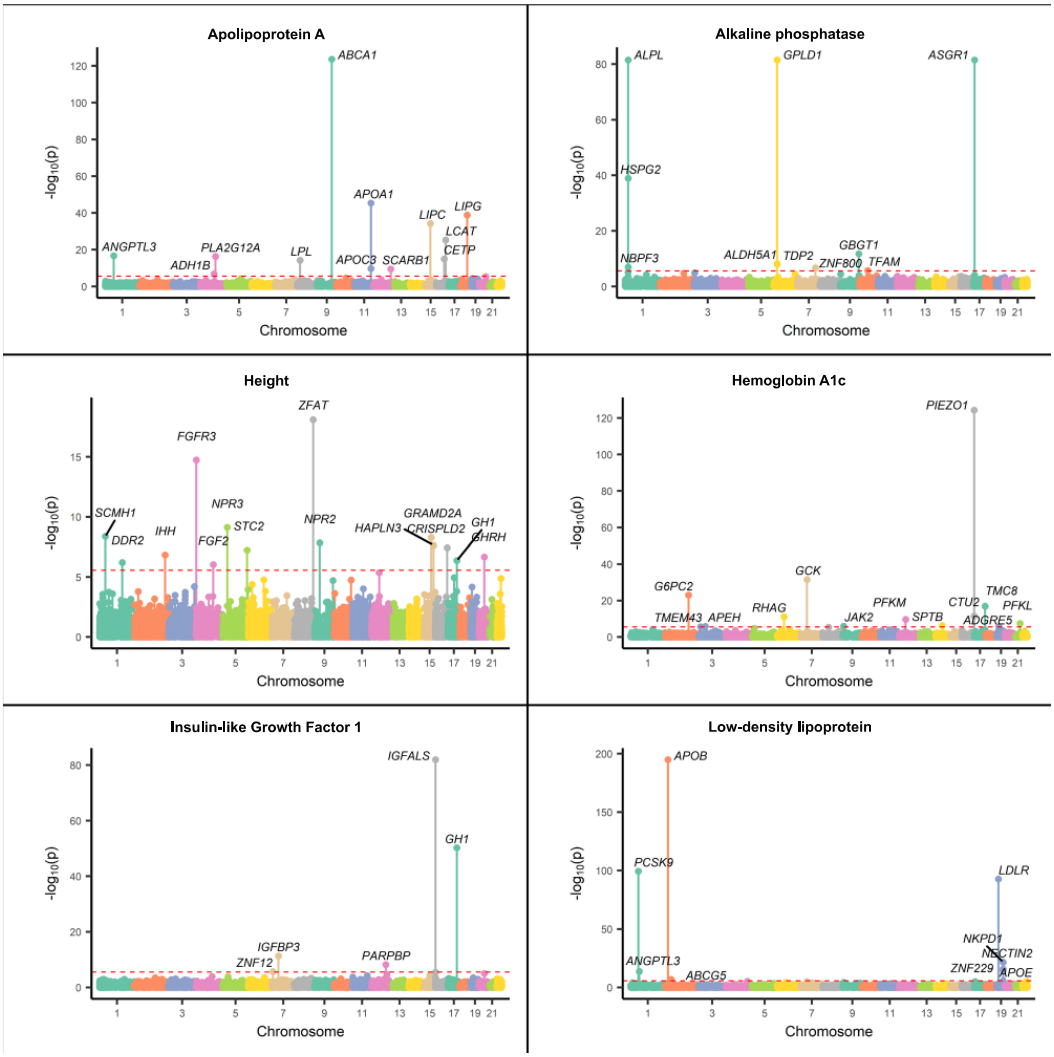


Fig. 5 | Manhattan plot illustrating genes with significant heritability for selected traits. RARity was used to determine $h^2_{RV\text{-}gene}$ for all genes with qualifying RVs. Each dot represents a single gene, with genes ordered on the *x*-axis according to their genomic position. The *y*-axis represents the significance of $h^2_{RV\text{-}gene}$ measured as $-\log_{10}$ transformed *p*-values, where the *p*-values were derived using *F*-test. Red, horizontal dashed lines mark the Bonferroni's *p*-value significance threshold corrected for 18,214 genes (p-value < 2.75×10^{-6}). Genes with significant $h^2_{RV\text{-}gene}$ are labelled.

RV heritability, such as RARity, are not intended to replace RV association methods, but rather to complement.

Together, these results confirmed that (1) RVs can account for a significant portion of the complex trait heritability, (2) gene-level RV aggregation (gene burden) leads to a substantial loss of information, (3) identification of genes significantly enriched for $h^2_{RV\text{-}gene}$ can help with gene discovery, and (4) innovative methods are needed to predict variant-level functionality. In conclusion, the high trait variance explained by RVs makes it imperative to continue to invest in the study of RVs and understand their impact on health and diseases. As such, future studies extending the methodology to analysis of dichotomous traits, particularly disease status, are in a pressing need.

Methods

Study population

The UK Biobank (UKB) study is a prospective cohort comprising of approximately 500,000 participants (ages 40–69 years) with extensive genotypic and phenotypic data from consenting individuals[43]. All UKB data included in our analyses were accessed as part of our approved application #15255. Here we utilized two main genetic datasets from the UK Biobank. First, our primary source of RVs was the WES data with 17,975,236 variants on 200,643 participants (UKB data field: 23155). Second, CVs were extracted from imputed genotype data on 488,264 individuals (data field: 22418; v3 release). The acquisition and primary quality control (QC) of both genetic data are described elsewhere[44]. Briefly, out of the 200,643 samples with WES data,

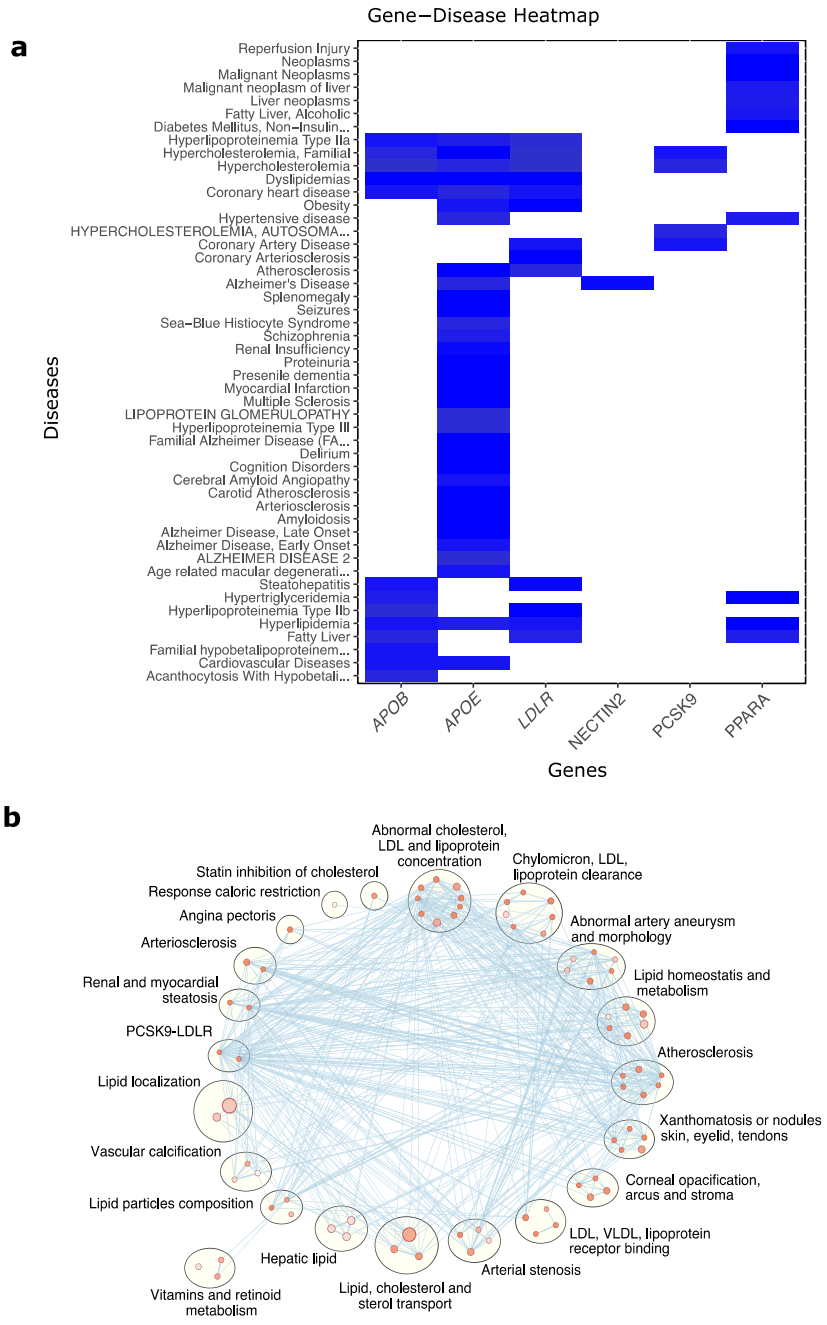


Fig. 6 | Network and pathway analyses of genes contributing significantly to ApoB. a Gene-disease heatmap, created with DisGeNET, for the genes contributing significantly to the RV heritability of ApoB. The intensity of colour is proportional to the strength of evidence for gene-disease-association, with darker colour representing a greater level of evidence. **b** Network cluster representing biological mechanisms influenced by genes contributing significantly to ApoB heritability estimates. The nodes represent gene-sets, the edges represent known interaction between biological pathways. Node size reflects the number of genes in the gene-set. More intense colour indicates higher enrichment value.

individuals were excluded based on: consent withdrawal ($n = 11$), call rates less than 99% ($n = 2$), discordance between genetic and reported sex ($n = 18$), a departure from putative ancestorial clusters based on the first two genetic principal components ($n = 3$), assigned cluster membership to a continental population with less than 5000 samples ($n = 12,765$, of which, South Asian = 3395; African = 3168; Other = 6202), and 3rd degree or closer relatedness ($n = 14,156$). In the remaining 173,688 individuals, an additional 6340 were removed following QC of biomarker data (as described below). We focused on the 167,348 unrelated Caucasian participants to estimate narrow-sense heritability contributed by CVs, RVs, or the combined CV and RVs.

Biomarker and anthropometric data

The blood biomarkers in UKB represent clinical diagnostic measures and established risk factors for diseases. For example, HbA1c is used in the diagnosis of diabetes and lipids are needed for risk stratification of cardiovascular diseases. Besides the standard QC steps implemented by the UKB study team[45], we applied additional steps to curate the final list of 26 biomarkers and 5 anthropometric traits (height, body mass index [BMI], waist-to-hip ratio, and systolic and diastolic blood pressure measured automatically) (Supplementary Table 1, Supplementary Fig. 1). Briefly, sex-specific biomarkers, such as sex hormone binding globulin (SHBG) and testosterone, as well as those biomarkers with >80% missingness (e.g. oestradiol and rheumatoid factors) were excluded from the analyses. Next, we winsorized values that were either above or below the detectable range, using the reportability fields for each biomarker. Out of the 173,688 participants with WES data, we removed 598 individuals with missing values for all 26 biomarkers. Since many of the examined biomarkers can be altered by specific medications including blood glucose, HbA1c, lipids, and blood pressure, we applied corrections for medication status (Supplementary Table 2). As a result, 5,742 individuals on glucose-lowering drugs were removed, meanwhile, individuals on statins had their baseline low-density lipoprotein cholesterol (LDL-C) and ApoB values adjusted by dividing by 0.7, their total cholesterol adjusted by dividing by 0.8, and their ApoA-I and high-density lipoprotein (HDL) cholesterol adjusted by dividing by 1.06 and 1.05, respectively[46–49]. Blood pressure-lowering medications were adjusted by adding 10 mmHg and 15 mmHg to their diastolic blood pressure (DBP) and systolic blood pressure (SBP), respectively[50,51]. Further, missing biomarker or trait values were imputed by their mean values. All traits were then quantile transformed to resemble a standard normal distribution and further adjusted for age, sex, and the first 20 genetic principal components (PCs) to account for any effects of sub-population structure within UKB[52], and finally standardized to have mean 0 and variance 1. Biomarker treatment for sex-stratified analysis was performed in an identical manner.

Genotype data quality control

Rare coding variants. Genetic variants were called from WES data following the Functional Equivalent pipeline[53]. All monomorphic variants ($m = 83,700$), variants with missing genotypes in more than 10% samples ($m = 369,215$), and those deviating significantly from Hardy-Weinberg Equilibrium (p-value $< 5 \times 10^{-6}$; $m = 35,317$) were removed. Remaining variants were annotated with predicted pathogenicity scores, and amino-acid changes using ANNOVAR *geneanno* pipeline with the refGene database[54]. Variants were annotated with MAF based on the UKB samples, as well as the five major ancestries identified in the Genome aggregation database (gnomAD 2.11): Latino, non-Finnish European, African/African American, East Asian, South Asian[27]. Qualifying RVs were defined as variants that were nonsynonymous single nucleotide variants, frameshift deletions or insertions, in-frame deletions or insertions, stop-gain, stop-loss and start-loss variants, with a minor allele count (MAC, the number of minor alleles at each locus in the population being studied) >2 and MAF below the cut-off (<1%, <0.5% or <0.1%) in all gnomAD subpopulations, and locally in the UKB samples. Within these variants, the stop gain/loss variants and frameshift variants were defined as the LoF variants, and the rest are referred to as protein-altering variants. RVs were also subset into MAF bins (MAF = 0.01–0.005, 0.005–0.001 or 0.01–0.001) to examine the contribution of different MAF categories to h^2_{RV}.

To reduce the influence of long-range LD between variants that would otherwise inflate the overall heritability estimate, as we show in subsequent simulations ("Calibrating RARity"), highly correlated RVs were removed using PLINK1.9[55], by LD pruning with a Pearson's r^2 threshold > 0.1 within a window of 50 Mb that was shifted by 500 bases at the end of each step. Next, individuals on glucose-lowering medications were removed and a MAC filter was applied once more to retain RVs with MAC > 2, leading to a final analytical dataset including 167,348 individuals and 1,592,257 variants with MAF < 1% in 18,213 genes. All other analyses were based on this analytical dataset. Individual level genotypes were extracted with PLINK1.9, assuming an additive model for all variants, and thus allotting a score of 2 for rare allele homozygous variants, 1 for heterozygous variants, and 0 otherwise. Multiallelic variants were treated as bi-allelic by considering the presence or absence of (any) rare variant. While RVs with missing genotype in >10% samples were removed during the early QC steps, mean imputation was employed to fill in the missing genotypes in the remaining samples. Finally, genotypes were standardized to have mean 0 and variance 1. The genotype and phenotype data processing steps are illustrated schematically in Supplementary Fig. 1.

Common variants. CVs originated from the third release of the UK Biobank genotype data in 2017. The imputed genotypes are based on the Genome Reference Consortium Human Build 37 (GRCH37), and further filtered to retain CVs with imputation quality score greater than 0.7, those with no significant deviation from Hardy-Weinberg equilibrium (p-value > 1×10^{-10}). We further filtered genotype data by LD pruning with $r^2 > 0.9$ and a rolling window of 1 Mb, that were shifted in steps of 500 bases, leaving 1,030,594 CVs in 159,058 participants, for whom we also had WES data. These pruning parameters were selected based on simulations using a very similar approach, using 325,989 participants in the UKB by Di Scipio and Khan et al.[31]. For compatibility with RVs, the CVs were lifted to the GRCH38 assembly using UCSC LiftOver[56]. Similar to RVs, individual level genotypes were extracted with PLINK1.9, assuming an additive model, mean imputation was also employed to fill in the missing genotypes (in <10% samples), followed by standardized to have mean 0 and variance 1.

Combined common and rare coding variants. We concatenated the derived RVs and CVs, and then LD pruned once again with the same parameters as CVs ($r^2 > 0.9$, window size = 1 Mb, step size = 500 bases). We already implemented a more stringent LD pruning schema for RVs ($r^2 > 0.1$, window size = 50 Mb, step size = 500 bases), conversely, an overly stringent LD pruning schema applied to CVs would remove most of the variants, and thus, for the combined CV + RV we implemented the more relaxed LD pruning similar to that for CV ($r^2 > 0.9$, window size = 1 Mb, step size = 500 bases) and blocks of 20,000 variants were constructed to enable distribution of both CVs and RVs within each block.

Statistical model to estimate RV heritability using RARity

We developed a method, RARity, to compute heritability estimates based on aggregating linear regression models over large genetic regions including up to thousands of variants. The method is based on a multivariate linear regression model:

$$\boldsymbol{Y} = \boldsymbol{G}\boldsymbol{\beta}_{\boldsymbol{G}} + \boldsymbol{\epsilon}, \quad (1)$$

The RARity method entails computing the multiple linear regression solutions using an ordinary least square (OLS) for each of the non-overlapping genetic blocks $(1,\ldots,k,\ldots,K)$ in parallel under the condition that n is much larger than the number of genetic variants (p_k) in the k th block, while ensuring the between block correlation, due to linkage disequilibrium (LD) spillage between blocks, is minimized. In other words, RARity approximates the linear solution to $\boldsymbol{\beta_G}$ by setting the observed $(\boldsymbol{G}'\boldsymbol{G}) \in R^{m\times m}$ to a block diagonal matrix, where $m = \sum p_k \gg n$. Given the observed quantitative trait y, the OLS estimate of the genetic effects vector for the k th block $\boldsymbol{G_k}$ is:

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{G_k}} = (\boldsymbol{G}_{\boldsymbol{k}}'\boldsymbol{G_k})^{-1}\boldsymbol{G}_{\boldsymbol{k}}'\boldsymbol{y}, \tag{2}$$

and the fitted value, denoted by $\hat{y}$, can be computed as:

$$\hat{\boldsymbol{y}} = \boldsymbol{G_k}\hat{\boldsymbol{\beta}}_{\boldsymbol{G_k}}. \tag{3}$$

The estimated heritability associated with the G matrix is simply the amount of variance explained by the fitted value:

$$R^2{}_{G_k} = \hat{\boldsymbol{y}}'\hat{\boldsymbol{y}}/\boldsymbol{y}'\boldsymbol{y}. \tag{4}$$

Since $R^2{}_{G_k}$ increases as the number of predictors increase, the adjusted R^2, denoted by $\bar{R}^2$, is used as our estimate for the proportion of variance explained. Total RV heritability of the trait is estimated by the sum of $\bar{R}_k{}^2$ over all K blocks:

$$\hat{h}^2 = \sum_{k=1}^{K}\bar{R}^2_{G_k} = \sum_{k=1}^{K}\left[1-\left(1-R^2{}_{G_k}\right)(n-1)/(n-p_k-1)\right]. \tag{5}$$

The 95% confidence interval (CI) of $R^2{}_{G_k}$ can be approximated for each block using the asymptotic properties described by Algina[57] using the Wald's method, where the variance of $R^2{}_{G_k}$ is given by:

$$\widehat{Var}\left(R^2{}_{G_k}\right) = \frac{4R^2{}_{G_k}\left(1-R^2{}_{G_k}\right)^2(n-k-1)^2}{(n^2-1)(n+3)} \tag{6}$$

The 95% CI for the adjusted R^2 of a single block can then be derived accordingly:

$$\bar{R}^2{}_{G_k} \pm 1.96\sqrt{\widehat{Var}\left(\bar{R}^2{}_{G_k}\right)}, \tag{7}$$

where

$$\widehat{Var}\left(\bar{R}^2{}_{G_k}\right) = \left(\frac{n-1}{n-p_k-1}\right)^2\widehat{Var}(R^2{}_{G_k}). \tag{8}$$

To estimate the 95% CI for $\hat{h}^2$, we approximated the asymptotic variance by the sum of the individual sampling variance, $\widehat{Var}(\bar{R}^2{}_{G_k})$ for each block, assuming each block is roughly uncorrelated of others after controlling for LD spillage:

$$\widehat{Var}\left(\hat{h}^2\right) = \sum_{k=1}^{K}\left(\frac{n-1}{n-p_k-1}\right)^2\widehat{Var}\left(R^2{}_{G_i}\right), \tag{9}$$

which then translates to a Wald's 95% CI for the $\hat{h}^2$:

$$\hat{h}^2 \pm 1.96\sqrt{\widehat{Var}\left(\hat{h}^2\right)}. \tag{10}$$

Since the model is conditional on the observed genotype matrix, RARity requires no parametrization nor assumptions regarding the genetic architecture of traits analyzed (such as polygenicity of effects or relationships between MAF/LD and effect size). The only assumption of block-wise independence was addressed by a stringent LD pruning to avoid long-range LD. The main computational burden for biobank scale datasets is the inversion of the $\boldsymbol{G}'\boldsymbol{G}$ matrix, which is of size $m \times m$, but the matrix calculation becomes quite manageable for blocks of $\boldsymbol{G}$, where each block contains p_k = 5000–10,000 variants and n ~ 200,000 in a standard high-performance computing environment. For example, for a typical analysis of 5000 RVs and 167,348 samples, the heritability estimation took approximately 7 minutes for all 31 traits, running on a single core and 7.3 Gb memory. All statistical analyses were performed using the statistical programming language R (version 3.6.0)[58].

Statistical power

Statistical power of $h^2{}_{RV}$ was estimated empirically from the variance of 10,000 simulated $h^2{}_{RV}$ $(\widehat{Var}(\hat{h}^2))$, under 230 conditions of sample sizes and true set $h^2{}_{RV}$. Non-central F-distributions were used to simulate the observed genetic effects at each genotype block, and exome-wide $h^2{}_{RV}$ was derived as described above. The true set $h^2{}_{RV}$ ranged from 2 to 25%, with increments of 5%. Sample size was varied from 25,000 to 250,000 individuals by increments of 10,000. For each condition, the statistical power was calculated as the proportion of observed p-values less than 0.05 out of the 10,000 simulations.

Calibrating RARity

Rarity can be used with both common and rare variants. The effect of LD on over-estimation of heritability is well-known for common variants[20] and can be minimized by choosing a suitable set of pruning parameters. In our preliminary analyses, we also observed over-estimation of $h^2{}_{RV}$, which occurs when rare variants in different blocks are in LD, thus, it was necessary to perform pruning and calibrate the pruning parameters to empirically minimize bias.

Here we conducted simulation studies using observed genotypes to identify the most suitable LD pruning parameters, including the Pearson's r^2 coefficient threshold, the window size, and the step size, to reduce bias in estimating heritability. To this end, RVs on each chromosome, with MAF < 0.01, were LD pruned under 7 different scenarios that varied in window sizes (ws) and LD r^2 threshold in the following combinations (ws = 1 Mb with r^2 > 0.9, 0.5 or 0.1; ws = 20 Mb with r^2 > 0.1; ws = 50 with r^2 > 0.9, 0.5 or 0.1), with a fixed step size of 500 bases. Pruned variants in each scenario were then partitioned into exome-wide blocks, with 5000 variants per block. We assumed that a random subset of 20% of the RVs in each block had an independent effect associated with the simulated trait of interest and their contribution to true set $h^2{}_{RV}$ was 0.05 for the whole exome. The unobserved genetic effects were simulated from a standard normal distribution while the errors were sampled independently from a normal distribution with mean 0 and variance 0.95.

The simulated phenotype ($\mathbf{Y_{sim}}$) was then computed as:

$$\boldsymbol{Y}_{\mathbf{sim}} = \boldsymbol{G}\boldsymbol{\beta}_{\boldsymbol{G}} + \boldsymbol{\epsilon} \tag{11}$$

We generated 20 phenotypes ($\mathbf{Y_{sim}}$) under each of the 7 scenarios to assess the overall impact of LD pruning parameters on RARity estimates and proceeded to estimating exome-wide $h^2{}_{RV}$ as described in the section "Statistical model to estimate RV heritability using RARity". In addition, the calibration was repeated with the assumption that the true $h^2{}_{RV}$ originated from 10% of RVs within 5% of the genes, instead of homogeneous distribution of $h^2{}_{RV}$ across all blocks.

To test whether the more relaxed LD pruning for CVs (r^2 > 0.9, ws = 1 Mb, step size = 500 bases) would impact the estimation of $h^2{}_{CV}$, we benchmarked RARity against alternative methods designed for CVs, namely BOLT[22] and LDSC[23].

Testing the effects of genetic architecture

To assess whether RARity is sensitive to MAF thresholds, values of true h^2_{RV}, or varying fractions of causal variants or genes, we further tested RARity with simulations utilizing real genotype data, pruned with the default LD pruning threshold ($r^2 > 0.1$ within a window size of 50 Mb). For each simulation, exome-wide h^2_{RV} was estimated for $\mathbf{Y_{sim}}$. A total of 20 simulations were carried out for each scenario as follows. The effect of varying MAF was tested by assuming that 10% of variants within 5% of genes are causal in each of the MAF categories (MAF < 0.001, <0.005 and <0.01). The effect of varying h^2_{RV} was tested for MAF < 0.01 with the true h^2_{RV} set to either 2, 5, 10, 20 or 25%. Next, keeping a consistent MAF < 0.01 and true h^2_{RV} at 10%, the effect of varying fractions of causal genes (0.01, 0.05, 0.1 and 0.2) was tested with 10% of RVs with true effects. Finally, keeping a consistent MAF < 0.01, and true set h^2_{RV} at 10%, we tested the effect of varying fractions of causal variants (0.001, 0.05, 0.1 and 0.2) in 5% of randomly selected genes.

Application to UKB data

We estimated RV, CV (h^2_{CV}) and combined CV and RV heritability (h^2_{tot}) of 31 complex traits, including 26 biomarkers and 5 anthropometric traits in 167,348 UKB samples. For RV heritability, we additionally examined the influence of block construction on the final heritability estimate. The qualifying RVs were arranged in three different ways to create blocks of genotype data that were used to estimate heritability: (1) gene-burden blocks were derived by summing the number of rare alleles within each gene for an individual, which produced a single block containing all gene-burden scores as predictors. Heritability estimates from this type of block is denoted with $h^2_{RV\text{-}burden}$; (2) gene-wise blocks, consisted of un-aggregated RVs partitioned by gene, such that each block contained all the variants within a single gene. In other words, there were as many blocks as the number of genes. Gene heritability from this type of block is referred to as $h^2_{RV\text{-}gene}$, and the total trait heritability based on all genes is denoted as $h^2_{RV\text{-}gene\text{-}tot}$; (3) exome-wide blocks, were created by partitioning RVs in each chromosome into blocks of ~5000 adjacent RVs. This type of construct results in blocks that are gene-agnostic, i.e., independent of gene borders with varying number of genes per block, for example, the number of genes ranged from 17–170 per exome-wide block when protein altering and LoF variants (<0.01 MAF) were selected. The total heritability estimates from exome-wide blocks are denoted with h^2_{RV}. In each case, LD spillage was minimized through pruning prior to creating the blocks (as described above, in the "Genotype data quality control" section).

The impact of block size on h^2_{RV} was tested using the first 4 blocks of ~5000 protein altering and LoF RVs (MAF < 0.01) from chromosome 22, where 2 consecutive blocks were combined to create blocks of ~10,000 RVs and all four blocks were combined to form a single block of ~20,000 RVs. Since the size of blocks did not impact the accuracy of h^2_{RV} (Supplementary Fig. 11), the choice of 5000 RVs per block in an exome-wide block construct was motivated by computational efficiency.

Enrichment analysis (pathogenicity and gene set)

Identification of significant genes. For each trait, we assessed 18,214 genes for their contribution to the total heritability and prioritized those with significant contribution for functional enrichment. A statistically significant contribution was determined by an *F*-test for regression models, against the null hypothesis that the gene-level heritability was zero, at a Bonferroni corrected gene-wide significance threshold of $\alpha = 0.05/18{,}214 = 2.75 \times 10^{-6}$.

Pathogenicity scores. We evaluated three well-known pathogenicity scores to classify deleterious RVs[59], namely, Combined Annotation Dependent Depletion (CADD) scores (v.1.6)[6], Mendelian Clinically Applicable Pathogenicity (M-CAP 1.3) scores[60], and rare exome variant ensemble learner (REVEL)[5]. We created subsets of RVs based on the default thresholds indicative of deleteriousness (CADD > 20, M-CAP > 0.025, and REVEL > 0.5[5,6,60]), as well as decreasing proportion of variants, by increasing the pathogenicity scores every 5th percentile. This was repeated for all three MAF cut-offs (<1%, <0.5% or <0.1%), and thus creating 63 subsets of RVs (21 pathogenicity cut-offs × 3 MAF thresholds) for each class of pathogenicity score. To derive an independent set of deleterious variants, we applied LD-clumping (Pearson's $r^2 > 0.1$, window size = 50 Mb) within each subset and retained the more pathogenic, independent variants with the highest score. We then constructed gene-wise blocks and derived $h^2_{RV\text{-}gene\text{-}tot}$, to efficiently estimate the total heritability for each subset of RVs. The proportion of $h^2_{RV\text{-}gene\text{-}tot}$ was measured in relation to all protein altering and LoF variants within each MAF categories. The proportion of $h^2_{RV\text{-}gene\text{-}tot}$ explained as a function of incorporating increasingly "deleterious" genetic variants was used to measure the performance of pathogenicity scores to identify sets of functional RVs.

Network and pathway analyses. The g:ProfileR[61] web tool for functional profiling, g:GOST, was used to test the enrichment of the genes with significant heritability ($h^2_{RV\text{-}gene}$ p-value < 2.75×10^{-6}), against gene-sets in common databases. The significant heritability genes for each trait were treated as separate gene lists for independent query, and statistical tests were conducted within a domain scope of only annotated genes, considering the GO biological process, GO molecular function, GO cellular component, KEGG, Reactome, TRANSFAC, miRNA, CORUM, HP, HPA, and WikiPathways data sources, and removing electronic GO annotations. This analysis resulted in a list of statistically significant enriched terms for each gene list, adjusted for multiple testing using g:SCS (set counts and sizes) p-value < 0.05, integral to the g:Profiler server. An Enrichment Map was then created with the results from g:Profiler, using Cytoscape version 3.10.1[62], with an FDR q-value cut-off value set to 0.001, and medium connectivity, producing networks where the nodes represent enriched pathways, and the edges represent all pairwise connections. These threshold values were chosen to ensure that we captured highly enriched pathways while being able to observe structure in the network. Next, we clustered the nodes to distinguish the major theme in the enriched pathways for each trait. DisGeNET[25] was used to retrieve and explore the gene-disease associations for the significant list of genes, resulting in a disease-heat-map for each trait. In addition, we ran a query of the target genes in the Drug-Gene Interaction database (DGIdb)[26] to identify potentially druggable and clinically actionable genes.

Heritability estimates in relation to sex and gene-level annotations

Sex-stratified analyses were performed using gene-wise blocks including RVs with MAF < 0.01, including 92,963 females and 74,385 males. Differences in the heritability estimates between the genetic sexes was determined with *t*-test. In order to investigate the influence of gene length and evolutionary constraint on RV heritability, we tested the association between these independent predictors and heritability contribution from each gene towards RV trait h^2 ($h^2_{RV\text{-}gene}$). The predictors included (1) gene-level evolutionary constraint, which was determined using gnomAD pLoF Metrics (v2.1.1)[27]; and (2) gene length, which was based on the curated RefSeq transcripts obtained via the UCSC genome browser. For each gene, the transcript with the longest transcript length was selected to represent the length of the gene. Since the distribution of the gene lengths are skewed, with some genes being much longer than others, the lengths of all genes were log transformed. Heritability estimates for both analyses were calculated using the same set of 1,592,257 RVs with MAF < 1%.

To examine the enrichment of h^2_{RV} in highly conserved regions, we estimated the $h^2_{RV\text{-}gene\text{-}tot}$ of height originating from the conserved

gene clusters, such as the hox, histone, protocadherin, and the hemoglobin gene clusters. The list of genes in these clusters was obtained from the HUGO Gene Nomenclature Committee (HGCN) database (www.genenames.org) and cross-listed with the genes carrying RVs in the UKB. The impact of gene orientation on h^2_{RV} was examined by comparing the $h^2_{RV\text{-}gene\text{-}tot}$ contribution of the genes transcribed from the positive vs. negative strand, where the strand of each gene was annotated using the RefGene database in the UCSC genome browser[56], with roughly equal number of genes present on either strand.

Reporting summary

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

Data availability

Individual genetic and phenotypic data were obtained from the UK Biobank (http://www.ukbiobank.ac.uk/), under application #15255. The UK Biobank study received approval from the National Health Service National Research Ethics Service North West. Access to the UK Biobank individual-level data is not publicly available and must be obtained via an application (https://www.ukbiobank.ac.uk/register-apply/). UCSC genome browser (https://genome.ucsc.edu/) was utilized to access LiftOver for the conversion of Genome Reference Consortium Human Build 37 (GRCH37) to Genome Reference Consortium Human Build 38 built, and to obtain additional gene-level annotations such as gene-length, strand orientation and the gnomAD pLoF Metrics. Databases for gene-disease associations (DisGeNet, https://www.disgenet.org/), Drug Gene Interaction database (DGIdb, version 4.2.0, https://www.dgidb.org/) and HUGO Gene Nomenclature Committee (HGCN, https://www.genenames.org/) were utilized to inform on the importance of the target genes. Variant level annotations and pathogenicity scores, such as Mendelian Clinically Applicable Pathogenicity (M-CAP) Score, and Rare exome variant ensemble learner (REVEL) were obtained using ANNOVAR or downloaded directly from the web-based platform, such as the Combined Annotation Dependent Depletion (CADD) scores. Source data for all main and supplemental figures are provided with this paper.

Code availability

All custom code and relevant documentation to run RARity is available in a public GitHub repository (https://zenodo.org/records/10426710)[63].

References

1. Auer, P. L. & Lettre, G. Rare variant association studies: considerations, challenges and opportunities. *Genome Med.* **7**, 16 (2015).
2. Xue, Y. et al. Deleterious- and disease-allele prevalence in healthy individuals: insights from current predictions, mutation databases, and population-scale resequencing. *Am. J. Hum. Genet.* **91**, 1022–1032 (2012).
3. Gunning, A. C. et al. Assessing performance of pathogenicity predictors using clinically relevant variant datasets. *J. Med. Genet.* **58**, 547–555 (2021).
4. Wang, D., Li, J., Wang, Y. & Wang, E. A comparison on predicting functional impact of genomic variants. *NAR Genom. Bioinform.* **4**, lqab122 (2022).
5. Ioannidis, N. M. et al. REVEL: an ensemble method for predicting the pathogenicity of rare missense variants. *Am. J. Hum. Genet.* **99**, 877–885 (2016).
6. Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* **47**, D886–D894 (2019).
7. Lee, S., Abecasis, R., Gonçalo, Boehnke, M. & Lin, X. Rare-variant association analysis: study designs and statistical tests. *Am. J. Hum. Genet.* **95**, 5–23 (2014).
8. Wu, M. C. et al. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.* **89**, 82–93 (2011).
9. Zuk, O. et al. Searching for missing heritability: Designing rare variant association studies. *Proc. Natl Acad. Sci. USA* **111**, E455–E464 (2014).
10. Young, A. I. Solving the missing heritability problem. *PLoS Genet.* **15**, e1008222 (2019).
11. Hernandez, R. D. et al. Ultrarare variants drive substantial cis heritability of human gene expression. *Nat. Genet.* **51**, 1349–1355 (2019).
12. Mancuso, N. et al. The contribution of rare variation to prostate cancer heritability. *Nat. Genet.* **48**, 30–35 (2016).
13. Marouli, E. et al. Rare and low-frequency coding variants alter human adult height. *Nature* **542**, 186–190 (2017).
14. Wainschtein, P. et al. Assessing the contribution of rare variants to complex trait heritability from whole-genome sequence data. *Nat. Genet.* **54**, 263–273 (2022).
15. Yang, J. et al. Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nat. Genet.* **47**, 1114–1120 (2015).
16. Burch, K. S. et al. Partitioning gene-level contributions to complex-trait heritability by allele frequency identifies disease-relevant genes. *Am. J. Hum. Genet.* **109**, 692–709 (2022).
17. Jang, S. K. et al. Rare genetic variants explain missing heritability in smoking. *Nat. Hum. Behav.* **6**, 1577–1586 (2022).
18. Zhang, Z., Xiao, X., Zhou, W., Zhu, D. & Amos, C. I. False positive findings during genome-wide association studies with imputation: influence of allele frequency and imputation accuracy. *Hum. Mol. Genet.* **31**, 146–155 (2021).
19. Hou, K. et al. Accurate estimation of SNP-heritability from biobank-scale data irrespective of genetic architecture. *Nat. Genet.* **51**, 1244–1251 (2019).
20. Park, L. Population-specific long-range linkage disequilibrium in the human genome and its influence on identifying common disease variants. *Sci. Rep.* **9**, 11380 (2019).
21. Freeman, M. S., Mansfield, M. W., Barrett, J. H. & Grant, P. J. Heritability of features of the insulin resistance syndrome in a community-based study of healthy families. *Diabet. Med.* **19**, 994–999 (2002).
22. Loh, P. R. et al. Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.* **47**, 284–290 (2015).
23. Bulik-Sullivan, B. K. et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
24. Coassin, S. & Kronenberg, F. Lipoprotein(a) beyond the kringle IV repeat polymorphism: the complexity of genetic variation in the LPA gene. *Atherosclerosis* **349**, 17–35 (2022).
25. Piñero, J. et al. The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Res.* **48**, D845–D855 (2020).
26. Freshour, S. L. et al. Integration of the Drug-Gene Interaction Database (DGIdb 4.0) with open crowdsource efforts. *Nucleic Acids Res.* **49**, D1144–D1151 (2021).
27. Karczewski, K. J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
28. Evans, L. M. et al. Comparison of methods that use whole genome data to estimate the heritability and genetic architecture of complex traits. *Nat. Genet.* **50**, 737–745 (2018).
29. Ni, G., Moser, G., Wray, N. R., Lee, S. H. & Consortium, S.W.G.o.t.P.G. Estimation of genetic correlation via linkage disequilibrium score

regression and genomic restricted maximum likelihood. *Am. J. Hum. Genet.* **102**, 1185–1194 (2018).
30. Srivastava, A. K., Williams, S. M. & Zhang, G. Heritability estimation approaches utilizing genome-wide data. *Curr. Protoc.* **3**, e734 (2023).
31. Di Scipio, M. et al. A versatile, fast and unbiased method for estimation of gene-by-environment interaction effects on biobank-scale datasets. *Nat. Commun.* **14**, 5196 (2023).
32. Speed, D. et al. Reevaluation of SNP heritability in complex human traits. *Nat. Genet.* **49**, 986–992 (2017).
33. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).
34. Speed, D. & Balding, D. J. SumHer better estimates the SNP heritability of complex traits from summary statistics. *Nat. Genet.* **51**, 277–284 (2019).
35. Wise, A. L., Gyi, L. & Manolio, T. A. eXclusion: toward integrating the X chromosome in genome-wide association analyses. *Am. J. Hum. Genet.* **92**, 643–647 (2013).
36. Hogue, J. C. et al. Differential effect of fenofibrate and atorvastatin on in vivo kinetics of apolipoproteins B-100 and B-48 in subjects with type 2 diabetes mellitus with marked hypertriglyceridemia. *Metabolism* **57**, 246–254 (2008).
37. Chen, W., Coombes, B. J. & Larson, N. B. Recent advances and challenges of rare variant association analysis in the biobank sequencing era. *Front. Genet.* **13**, 1014947 (2022).
38. Bocher, O. & Génin, E. Rare variant association testing in the non-coding genome. *Hum. Genet.* **139**, 1345–1362 (2020).
39. Lee, S. H., Wray, N. R., Goddard, M. E. & Visscher, P. M. Estimating missing heritability for disease from genome-wide association studies. *Am. J. Hum. Genet.* **88**, 294–305 (2011).
40. Bernabeu, E. et al. Sex differences in genetic architecture in the UK Biobank. *Nat. Genet.* **53**, 1283–1289 (2021).
41. Rawlik, K., Canela-Xandri, O. & Tenesa, A. Evidence for sex-specific genetic architectures across a spectrum of human complex traits. *Genome Biol.* **17**, 166 (2016).
42. Ge, T., Chen, C. Y., Neale, B. M., Sabuncu, M. R. & Smoller, J. W. Correction: Phenome-wide heritability analysis of the UK Biobank. *PLoS Genet.* **14**, e1007228 (2018).
43. Bycroft, C. et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
44. Chong, M. et al. GWAS and ExWAS of blood mitochondrial DNA copy number identifies 71 loci and highlights a potential causal role in dementia. *Elife* **11**, e70382 (2022).
45. Biobank, U. UK Biobank Serum Biochemistry (2019).
46. Khera, A. V. et al. Diagnostic yield and clinical utility of sequencing familial hypercholesterolemia genes in patients with severe hypercholesterolemia. *J. Am. Coll. Cardiol.* **67**, 2578–2589 (2016).
47. Klarin, D. et al. Genetics of blood lipids among ~300,000 multi-ethnic participants of the Million Veteran Program. *Nat. Genet.* **50**, 1514–1523 (2018).
48. Barter, P. J., Brandrup-Wognsen, G., Palmer, M. K. & Nicholls, S. J. Effect of statins on HDL-C: a complex process unrelated to changes in LDL-C: analysis of the VOYAGER Database. *J. Lipid Res.* **51**, 1546–1553 (2010).
49. Jones, P. H. et al. Effects of rosuvastatin versus atorvastatin, simvastatin, and pravastatin on non-high-density lipoprotein cholesterol, apolipoproteins, and lipid ratios in patients with hypercholesterolemia: additional results from the STELLAR trial. *Clin. Ther.* **26**, 1388–1399 (2004).
50. Tobin, M. D., Sheehan, N. A., Scurrah, K. J. & Burton, P. R. Adjusting for treatment effects in studies of quantitative traits: antihypertensive therapy and systolic blood pressure. *Stat. Med.* **24**, 2911–2935 (2005).
51. Paz, M. A. et al. Treatment efficacy of anti-hypertensive drugs in monotherapy or combination: ATOM systematic review and meta-analysis of randomized clinical trials according to PRISMA statement. *Medicine* **95**, e4071 (2016).
52. Galinsky, K. J., Loh, P. R., Mallick, S., Patterson, N. J. & Price, A. L. Population structure of UK Biobank and ancient eurasians reveals adaptation at genes influencing blood pressure. *Am. J. Hum. Genet.* **99**, 1130–1139 (2016).
53. Regier, A. A. et al. Functional equivalence of genome sequencing analysis pipelines enables harmonized variant calling across human genetics projects. *Nat. Commun.* **9**, 4038 (2018).
54. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).
55. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
56. Kuhn, R. M., Haussler, D. & Kent, W. J. The UCSC genome browser and associated tools. *Brief. Bioinform.* **14**, 144–161 (2013).
57. Algina, J. A comparison of methods for constructing confidence intervals for the squared multiple correlation coefficient. *Multivar. Behav. Res.* **34**, 493–504 (1999).
58. Core-Team. R Foundation for Statistical Computing (Vienna, Austria, 2022).
59. Li, J. et al. Performance evaluation of pathogenicity-computation methods for missense variants. *Nucleic Acids Res.* **46**, 7793–7804 (2018).
60. Jagadeesh, K. A. et al. M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. *Nat. Genet.* **48**, 1581–1586 (2016).
61. Raudvere, U. et al. g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res.* **47**, W191–W198 (2019).
62. Shannon, P. et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).
63. Pathan, N. A method to estimate the contribution of rare coding variants to complex trait heritability. *GitHub Repository*. https://doi.org/10.5281/zenodo.10426709. (2023).
64. Wang, Q. et al. Rare variant contribution to human disease in 281,104 UK Biobank exomes. *Nature* **597**, 527–532 (2021).
65. Backman, J. D. et al. Exome sequencing and analysis of 454,787 UK Biobank participants. *Nature* **599**, 628–634 (2021).
66. GraphPad Software. https://www.graphpad.com/. (2022).

Acknowledgements

The authors are thankful or all the funders, participants, and investigators of the UK Biobank. All UK Biobank data included in our analyses were accessed as part of our approved application #15255. We would also like to thank Dr. Andrew McArthur and his laboratory at McMaster University for providing GPU support.

Author contributions

N.P. conceived and implemented the project, performed computational analyses, analyzed results, and wrote the manuscript. G.P. conceived and supervised the overall project. W.D. and M.K. provided expert guidance and feedback on statistical analyses and writing. M.D.S. and S.M. performed the CV related computational analyses. M.P. and R.L. curated the list of medications for diabetes and lipids. M.R.C. performed primary QC of genotype data and R.W.M. created Fig. 1. All authors reviewed and edited the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-024-45407-8.

Correspondence and requests for materials should be addressed to Guillaume Paré.

Peer review information *Nature Communications* thanks Seunggeun Lee, Sara Rashkin and Gang Wu for their contribution to the peer review of this work. A peer review file is available.

CHAPTER 5:
RARE VARIANT HERITABILITY ESTIMATOR TO IDENTIFY GENES IMPLICATED IN DISEASE STATUS IN THE UK BIOBANK

In preparation

Rare variant heritability estimator to identify genes implicated in disease status in the UK Biobank

Nazia Pathan,[1,2] Alia Chaudhary[1], James Feiner[1,3], Nicolas Parrot[1,2], Guillaume Paré[1,2,4,5*]

[1]Population Health Research Institute, David Braley Cardiac, Vascular and Stroke Research Institute, Hamilton Health Sciences and McMaster University, Hamilton, Canada

[2]Department of Pathology and Molecular Medicine, McMaster University, Michael G. DeGroote School of Medicine, Hamilton, Canada

[3]Department of Biochemistry and Biomedical Sciences, McMaster University

[4]Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, Canada.

[5]Thrombosis and Atherosclerosis Research Institute, David Braley Cardiac, Vascular and Stroke Research Institute, Hamilton, Canada

*Correspondence: pareg@mcmaster.ca

ABSTRACT

The heritability of disease status due to rare coding variants (RVs, MAF<1%) remains unclear due to limited power. We introduce the rare variant heritability estimator-β (RARity-β), an adaptation of RARity for continuous traits, to estimate the overall ($\hat{h}^2_{RV\text{-}liab}$) and gene-level heritability ($\hat{h}^2_{RV\text{-}liab\text{-}gene}$) of binary traits on a liability scale. RARity-β was applied to 18 binary traits (cases >10,000) and RVs across 18,217 genes from whole exome sequence data, in the UK Biobank Caucasian participants (N=173,688). Significant

heritability was observed in hypothyroidism, asthma, hypercholesterolemia, and essential hypertension, with respective $\hat{h}^2_{RV\text{-}liab}$ of 12.5% (95% CI: 1.7-23.4%), 12.5% (95% CI: 4.0-20.9%), 7.9% (95% CI: 1.1-14.7%), and 6.4% (95% CI: 1.4-11.3%). Gene-level analysis identified 77 genes with significant $\hat{h}^2_{RV\text{-}liab\text{-}gene}$ for one or more traits, with 70 previously unidentified gene-trait relationships, including the contributions of an autophagy gene, *PEPB1,* to atrial fibrillation and the TSHR gene's contribution to hypothyroidism and sciatica. The *TSHR* gene and several mutations in this gene increases levels of thyroid-stimulating hormone-beta (TSHB), even in the absence of clinically diagnosed thyroid disorder. Our findings confirm the functional role of the *TSHR* gene in regulating TSHB and suggest that hypothyroidism is an under-recognized cause of sciatica.

INTRODUCTION

The emergence of large biobanks, such as the UK Biobank (UKB) and AllofUS have dramatically increased the number of rare variant (minor allele frequency, MAF<1%) associations that have been detected, especially at the gene-level, illustrating significant contributions of the rare variants to complex traits. Numerous methods of gene-phenotype associations, such as Collapsing Analysis, SKAT, Burden test, BOLT-LMM, SAIGE-GENE+, SKAT-O, STAAR-SKAT, aSPU and ACAT-V have helped in the identification of significant RV-phenotype associations, especially at the gene-level[1], many of these results are now readily available through catalogs such as RAVAR[1], Genebass[2] and AstraZeneca PheWAS portal[3]. The significance of rare variant (RV) associations is underscored by the statistical observation that genes harboring RVs are linked to an average

of 3.75 distinct traits. Concurrently, most traits demonstrate interactions with numerous genes, with a median of seven associated genes per trait[1]. Historically, the RV contributions to diseases have been linked to Mendelian forms of diseases, however most diseases are complex in nature. Even though RVs play an important role in human diseases and are postulated to explain missing heritability in many complex traits[4], rare variant contributions to the heritability estimates of disease status remain unknown, largely because of the lack of power with fewer prevalent cases. Current methods to estimate RV heritability for binary traits use gene-burden heritability regression[5,6] or relies on prior assumptions related to distribution of MAF or effect sizes[7], however, aggregating rare variants can cause substantial loss of variance[8], while assessing heritability based on assumptions such as beta and distribution of MAF, may lead to biases when the assumptions are violated. Thus, there is a need to develop a robust method to assess the overall and gene-level contributions of RVs to complex diseases.

In our previous study, we developed a Rare variant heritability estimator(RARity) which has proven to be successful in delineating the extent of heritability attributable to rare coding variants, and unraveling the genetic contributions to diverse continuous traits, including the discovery of several susceptibility genes related to lipids and anthropomorphic straits[8]. Here we extend the method to estimate the narrow sense, liability scale heritability (LHS) of binary traits. For case-control phenotypes, the heritability estimates on the observed binary scale are dependent on the case prevalence, hence it is preferred to transform the observed scale heritability into liability scale heritability (LSH).

LSH standardizes the genetic contribution across different diseases and traits, allowing for better comparability; providing a biologically interpretable measure of genetic risk, and mitigates ascertainment bias[9,10]. In this study, we developed RARity-β, an adaptation of RARity to estimate overall and gene-level heritability of binary traits. We hypothesize that gene-level heritability estimation of binary traits will enable prioritization of genes underlying disease mechanisms. With this objective, we applied RARity-β to estimate liability scaled heritability in 18 binary traits in the UKB, each with more than 10,000 cases.

METHODS

Statistical model to estimate RV heritability using RARity-β

Like the original RARity, RARity-β computes heritability estimates based on aggregating linear regression models over large genetic regions, including thousands of variants, using an ordinary least square (OLS) for each of the non-overlapping genetic blocks (1,...,*k*,...,*K*) in parallel under the condition that *n* is much larger than the number of genetic variants (p_k) in the k^{th} block, while ensuring the between block correlation, due to linkage disequilibrium (LD) spillage between blocks, is minimized. Here, we use each gene as a block to estimate RV heritability, and apply an empirically derived, LD pruning threshold size ($r^2 > 0.1$, window size = 50Mb, step size= 500 bases) to ensure robust results. The statistical details for computing the RV heritability ($\acute{h}^2$=adjusted R^2) of the trait and the 95% confidence interval (CI) for RV heritability are explained in detail in Pathan, *et. al*, 2024[8]. For quantitative traits the scale of measurement is the same as the scale on which heritability is expressed. In the context of binary traits, such as disease status, the phenotypes (case-

control status) are measured on the 0–1 scale, but heritability is most interpretable on a scale of liability[5]. LSH ensures that the estimates are not affected by ascertainment biases or disease prevalence. RARity-β primarily differs from RARity in application of a liability scale to the observed heritability and confidence limits, as described by Lee, et. al and Ojavee, et.al[5,10] and shown in equations 1 and 2, respectively.

$$\hat{h}^2_{liab} = \hat{h}^2 \frac{K(1-K)}{z^2}\frac{K(1-K)}{P(1-P)} \qquad (1)$$

$$95\%\, CL_{liab} = CL\frac{K(1-K)}{z^2}\frac{K(1-K)}{P(1-P)} \qquad (2)$$

Generally, *P* is the prevalence of cases in sample population and *K* is prevalence of cases in full population, and z^2 is is the squared probability density function of the standard normal distribution evaluated at the K^{th} quantile of the inverse cumulative density function of the standard normal distribution, i.e.:

$$z^2 = \varphi(\phi^{-1}(K))^2 \qquad (3)$$

With a large sample size, as we use here, we assume that *P*=*K*.

Identification of significant genes

For each trait, we assessed 18,217 genes for their contribution to the total heritability ($\hat{h}^2_{RV\text{-}liab}$) and prioritized those gene-heritability ($\hat{h}^2_{RV\text{-}liab\text{-}gene}$) with significant contribution for functional enrichment. A statistically significant contribution was determined by an *F*-test for regression models, against the null hypothesis that the gene-level heritability was zero, at a suggestive exome-wide significance threshold of $\alpha = 0.05/18{,}217$ genes

$= 2.74\times10^{-6}$, and a strict Bonferroni corrected exome-wide significance threshold of $\alpha = 0.05/(18{,}217 \text{ genes}\times 18 \text{ traits}) = 1.52\times10^{-7}$. Known disease associations of the significant genes were investigated with the DisGenet database.

Statistical power

Statistical power of $\hat{h}^2_{RV\text{-}liab}$ was estimated empirically from the variance of 10,000 simulated $\hat{h}^2_{RV\text{-}liab}$, under 272 conditions of number of cases and true set $\hat{h}^2_{RV\text{-}liab}$, keeping a constant sample size of 173,651 participants. Since each gene has a varying number of variants, power calculation was simplified with the assumption of having 317 blocks with 5,000 variants per-block. Non-central *F*-distributions were used to simulate the observed genetic effects at each genotype block, where the non-central parameter was calculated as the product of the number of blocks and sample size. Total $\hat{h}^2_{RV\text{-}liab}$ was derived as described above. The true set $\hat{h}^2_{RV\text{-}liab}$ ranged from 2 to 10%, with increments of 1%. Number of cases varied from 2,000 to 10,000 individuals by increments of 500. For each condition, the statistical power was calculated as the proportion of observed *p*-values less than 0.05 out of the 10,000 simulations.

Testing for bias

Permutation tests were done to assess Type I errors by randomly generating null distributions of traits, under the assumption that significant $\hat{h}^2_{RV\text{-}liab}$ will be observed if the null hypothesis is true. We generated 20 phenotypes (Y_{sim}) via random permutation of cases, under 4 scenarios of varying number of cases to assess the overall impact of type 1

error on $\hat{h}^2_{RV\text{-}liab}$. The number of cases varied from 5,000 to 20,000 individuals by increments of 5000, thus leading to 80 simulations. By comparing the observed $\hat{h}^2_{RV\text{-}liab}$ with the mean $\hat{h}^2_{RV\text{-}liab}$ for each scenario, it was determined whether the observed $\hat{h}^2_{RV\text{-}liab}$ are statistically significant or if they could have occurred by chance. P-value for the permutation tests for each scenario was calculated as the proportion of test-statistics >=5%. All statistical analyses were performed using the statistical programming language R (version 3.6.0)[11].

Application in the UK Biobank Data

UKB Sample

The UK Biobank (UKB) is a prospective population-based cohort encompassing approximately 500,000 individuals aged 40–69 years, serves as a rich source of genotypic and phenotypic data, all collected with participant consent under our approved application #15255. Leveraging Whole Exome Sequencing (WES) data from 200,643 participants, containing 17,975,236 variants (as annotated in the UKB data field: 23155), we focused our investigation on rare variants. Quality control procedures, described extensively elsewhere, were applied to ensure data integrity[8]. Among the samples with WES data, exclusions were made based on withdrawal of consent (n = 11), call rates below 99% (n = 2), discordance in reported sex (n = 18), deviation from ancestral clusters (n = 3), assignment to continental populations with fewer than 5000 samples (n = 12,765, subdivided as follows: South Asian = 3395; African = 3168; Other = 6202), and relatedness up to the 3rd degree (n = 14,156). Our analysis then focused on 173,688 unrelated British

and non-British Caucasian participants to estimate the narrow-sense liability scaled heritability ($\hat{h}^2_{RV\text{-}liab}$) contributed by rare variants within the UK Biobank cohort.

Genetic Data

Genetic variants were called from WES data following the Functional Equivalent pipeline[12] and treated identically as described previously, with the exception of adjustment for medications. MAC filter was applied once more to retain RVs with MAC > 2, leading to a final analytical dataset including 173,688 participants and variants with MAF < 1% in 18,217 genes. Individual level genotypes were extracted with PLINK1.9, assuming an additive model for all variants, and thus allotting a score of 2 for rare allele homozygous variants, 1 for heterozygous variants, and 0 otherwise. Finally, genotypes were standardized to have mean 0 and variance 1.

Phenotype Data

Disease outcomes

Health data was gathered from self-reported medical histories and physical assessments upon enrollment, with ongoing updates obtained through connected electronic health records. Disease entities were defined by the PheWAS Codes (PheCode) derived from the occurrence of related International Classification of Disease (10th Revision) and Related Health Problem codes, in electronic health records from hospital inpatient admissions (field identifier 41270), cancer registry (field identifier 40006), underlying cause of death registry (field identifier 40001), and contributory cause of death registry (field identifier 40002). Sex-specific outcomes and outcomes with fewer than 10,000 cases were excluded because

of the lack of statistical power. Also, given the inherent redundancy between PheCodes, we selected either the child PheCodes for specific diagnosis or the parent PheCodes, whereby the child PheCodes were collapsed to parent PheCodes, forming the representative trait, while maximizing the sample size[12,13]. This left us with 18 PheCodes for analyses with sufficiently powered prevalent cases, where cases were coded as "1" and the controls as "0". Each disease outcome was residualized with age, sex, 20PCs and then standardized to a mean zero and standard deviation 1.

Putative role of the TSHR gene

TSHR was one of the lead genes, contributing significantly towards $\hat{h}^2_{RV\text{-}liab}$ of hypothyroidism. We investigated the putative role of the *TSHR* gene in individuals without any form of thyroid disorders with a linear regression between the Thyrotropin subunit beta (TSHB) plasma protein as the outcome variable and both *TSHR* gene-burden, and individual *TSHR* variants, as explanatory variables, while correcting for age, sex and 20PCs. TSHB, which is the beta subunit of the TSH diagnostic biomarker for thyroid disfunction, was obtained from the UKB plasma proteomic profiles of 54,219 UK Biobank participants. TSHB was measured with Olink proteomics assay, of which 18,714 also had WES data to allow this analysis. A phenome-wide study on the heritability contributions of the *TSHR* was carried out, exploring 360 different phenotypes, each comprising over 2000 cases from the UK Biobank.

RESULTS

We previously developed the RARity model to estimate the heritability of complex continuous traits explained by protein coding rare variants in a WES of unrelated individuals[8]. To quantify the heritability of complex binary traits that can be explained RVs with RARity-β, we first needed to select traits with sufficient power and investigate whether this approach can provide an unbiased estimate of $\hat{h}^2_{RV\text{-}liab}$ using WES data. With access to limited traits with sufficient cases, we contended to using diagnoses of 18 common diseases, each with at least 10,000 cases, which ensures 50-80% power to observe 10% to 14% $\hat{h}^2_{RV\text{-}liab}$, respectively (Figure1). The baseline characteristics of the disease outcomes selected for this study are presented in Table1.

Simulations were performed based on a WES data set, comprising 173,651 unrelated British and non-British Caucasian participants after QC (Methods). The simulation results showed that when the traits were permuted randomly, estimate of $\hat{h}^2_{RV\text{-}liab}$ using real genotype data was unbiased (Figure 2). By unbiased we mean that the mean estimate of $\hat{h}^2_{RV\text{-}liab}$ from 20 permutations for each case size was always zero in the absence of meaningful phenotypes, however, fewer cases (<10,000) showed a high degree of error, and this confirms our choice of disease outcomes with at least 10,000 cases.

Heritability of 18 common diseases

We analyzed UK Biobank exome sequences in up to 173,651 British and non-British Caucasian for 18 common diseases and. Four of these traits had significant RV contributions to $\hat{h}^2_{RV\text{-}liab}$, include hypothyroidism, asthma, hypercholesterolemia, and essential hypertension, with respective $\hat{h}^2_{RV\text{-}liab}$ of 12.5% (95% CI: 1.7-23.4%), 12.5% (95% CI: 4.0-20.9%), 7.9% (95% CI: 1.1-14.7%), 6.4%(95%CI: 1.4-11.3%) (Figure 3, Table 2).

To investigate gene-level contributions of RVs, we used RARity-β to determine $\hat{h}^2_{RV\text{-}liab\text{-}gene}$ for all genes with qualifying variants and derived corresponding *p*-values for each gene. 242 of the 18,217 genes had significant $\hat{h}^2_{RV\text{-}liab\text{-}gene}$ (*p*-value $< 2.75 \times 10^{-6}$), for one or more traits (herein referred to as significant $\hat{h}^2_{RV\text{-}liab\text{-}gene}$), representing 244 distinct gene-disease relationships. 77 significant genes remained after applying a more stringent Bonferroni correction (*p*-value < (0.05/ (18217genes x 18 traits) = 1.52×10^{-7}). A list of these 77 genes with the corresponding $\hat{h}^2_{RV\text{-}liab\text{-}gene}$ and *p*-values are presented in Appendix-B, Supplementary Table 1, as well as in Manhattan plots (Figure 4). We identified several genes that replicated previously reported associations while 70 gene-traits relationships are previously unidentified (Table 3). Some examples of well-established gene-disease relationships captured by significant $\hat{h}^2_{RV\text{-}liab\text{-}gene}$ include association between *GCK* (*glucokinase*) and Type-2-diabetes, *TSHR* (*thyroid-stimulating hormone receptor)* and hypothyroidism, and *LDLR* (*low-density lipoprotein receptor*) and hypercholesterolemia. Previously unidentified gene-trait relationships include *PEBP1* (*phosphatidylethanolamine*

binding protein 1) with atrial fibrillation and flutter, and *IDH2 (Isocitrate Dehydrogenase (NADP (+)) 2)* with renal failure as well as pneumonia.

Putative role of TSHR in hypothyroidism

To illustrate the importance of RV heritability in disease etiology, we investigated the potential involvement of the *TSHR* gene in hypothyroidism, given its significant contribution to the $\hat{h}^2_{RV\text{-}liab\text{-}gene}$ of this condition. Consistent with previous findings, the *TSHR* gene is notably associated with multiple thyroid-related diseases (Figure 5). Linear regression of 219 RVs in the *TSHR* gene with hypothyroidism revealed that 4 variants are significantly associated with hypothyroidism (Table 4). However, when investigating which variants contribute to the overall $\hat{h}^2_{RV\text{-}liab}$ estimates of hypothyroidism, we found that at least 17 variants contribute to the exome-wide heritability estimates (Figure 6). We further leveraged the UK Biobank proteomics datasets to explore the influence of the *TSHR* gene on TSHB levels in individuals with and without any form of thyroid disorder. The results show strong, positive association between the *TSHR* gene burden and TSHB, and between 11 individual variants within the gene and TSHB, despite the absence of thyroid disorder (Table 5). Conversely, 4 variants and the *TSHR* gene burden were associated with TSHB when individuals with thyroid disorders were not excluded (Appendix-B, Supplementary Table 2). We also found a strong association between sciatica and hypothyroidism (*p-value* = 1.11×10^{-15}). Hypothyroidism is diagnosed in 270 of the 2309 participants with sciatica, prompting further investigation into the contributions of the *TSHR* gene to the heritability of various phenotypes, particularly those with more than 2000

cases. Our results revealed that the *TSHR* gene positively contributes to sciatica and limb pain, among other traits (Table 6). The putative functional and epidemiological consequences of the TSHR mutations are summarized in Appendix-B, Supplementary figure 1.

DISCUSSION

In this study, we aimed to estimate the heritability of complex binary traits explained by RVs using WES data. Building upon our previous work where the RARity model was developed to estimate the heritability of continuous traits, we sought to assess its utility in quantifying the heritability of binary traits in unrelated individuals. The robustness of the method was evaluated through simulations. We observed genome-wide significant estimates for 4 traits and gene-level estimates revealed important gene-disease associations for 244 genes.

The significance of our approach is underscored by its ability to capture significant overall heritability estimates for 4 traits, as well as well-established gene-disease relationships, as evidenced by significant $\hat{h}^2_{RV\text{-}liab\text{-}gene}$ estimates. Notably, our analysis identified several compelling examples of such associations. One prominent example is the association between *GCK* and Type 2 diabetes. *GCK* plays a crucial role in glucose metabolism, and variants in this gene have been consistently implicated in the pathogenesis of Type 2 diabetes[14,15]. Similarly, *LDLR* contributes significantly to hypercholesterolemia, and this is

a well-established gene-disease relationship with profound implications for cardiovascular health. Variants in *LDLR* disrupt lipid metabolism, leading to elevated levels of LDL cholesterol and increased cardiovascular risk[16].

The *PEBP1* gene contributes significantly towards atrial fibrillation and flutter, but has not been previously reported, however, there is some evidence of its role in cardiac hypertrophy[17]. Furthermore, the *PEBP1* is expressed at medium levels in heart muscles[18], supporting its relevance to cardiac function and diseases. Under normal circumstance, the PEBP1 protein interacts with phosphatidylethanolamine in cellular membranes and regulates the initiation of autophagy through interactions with several other autophagy related proteins[19]. Autophagy, a vital cellular process that degrades unfolded and misfolded proteins, plays a protective role in cardiomyocytes against AF. This process is highly conserved across evolution and is considered a potential therapeutic target and biomarker for various cardiovascular diseases, including AF. Recent studies suggest that regulating autophagy could help reduce the incidence of AF[20-22]. Together, the results implicate the *PEBP1* gene as a potential therapeutic target for AF, and merit further studies.

The highly significant association between *TSHR* and hypothyroidism was particularly noteworthy, prompting a focused examination of the relationship between *TSHR* variants and the manifestation of hypothyroidism. *TSHR* has been implicated in several Mendelian forms of thyroid dysfunction. We identified 4 nonsynonymous variants in the *TSHR* gene that are significantly associated hypothyroidism (Table 4), of which, the variants *TSHR:NM_000369:exon10:c.G1637A:p.W546X* and

TSHR:NM_000369:exon1:c.G122C:p.C41S have been previously implicated in Mendelian forms of congenital hypothyroidism[23-25]. In our study, we provide further evidence reaffirming the association between these variants and hypothyroidism at the population level. However, it is important to recognize that the overall heritability estimate of this gene is not explained entirely by only the significant variants. In fact, we show at least 17 variants within this gene that contribute to the overall heritability estimates (Figure 6), underscoring the polyallelic nature of the rare variants.

Under normal circumstances, the TSH produced by the pituitary gland binds to the TSHR in the thyroid gland, releasing the T4 and T3 hormones. Once sufficient levels of the thyroid hormones are released, further production of TSH by the pituitary gland is prevented with a negative feedback loop. Mutations in the *TSHR* gene impair the receptor functionality, resulting in overproduction of TSH. Detection of a strong, positive association between the 11 *TSHR* variants, as well as the *TSHB* gene burden and the TSHB levels in a subset of the population without any form of thyroid disorder, confirms decrease in function of the TSH receptor, in the absence of diagnosed thyroid conditions. Interestingly, we also found strong association between sciatica and hypothyroidism, this warranted investigation of the contributions of the *TSHR* gene on the heritability of other phenotypes, including all available phenotypes with more than 2000 cases. We found significant contributions of the *TSHR* gene in several other traits, including sciatica and limb pain. The observed interrelationship between the *TSHR* gene, sciatica and hypothyroidism presents a possible

clinical implication for individuals afflicted with sciatica. Our findings suggest that hypothyroidism is putatively an under-recognized cause of sciatica.

Important limitations need to be acknowledged. Specifically, the power calculations revealed that the sample size posed a constraint, thereby limiting our ability to detect heritability estimates below 14%, when traits have a minimum of 10,000 cases. This limitation emphasizes the need for caution when interpreting results, particularly in scenarios where the true heritability falls below this threshold, and the observed heritability on liability scales are negative. While negative heritability has been observed and dismissed as statistical artifacts in other studies[26], here, the estimates deviate significantly and justify further investigations into the extraneous factors driving this anomaly. Some of the factors driving negative heritability include over-adjustment for LD, underpowered analysis due to low allele count, population structure, and case-control imbalance associated with large data banks on generally healthy population, etc. Efforts to address these constraints, such as increasing sample sizes to ~1 million WES by pooling several large biobanks, or refining analytical strategies, may enhance the sensitivity of our method to detect lower levels of heritability and further strengthen its utility in genetic studies of complex traits.

Overall, our findings support the feasibility and validity of using the RARity-β model in conjunction with WES data to estimate the heritability of complex binary traits explained by RVs. By leveraging large-scale genotype data and carefully selecting traits with sufficient statistical power, we have demonstrated the utility of our approach in advancing our understanding of the genetic architecture of common diseases. Future research can

build upon these findings by applying similar methods to additional trait categories and expanding the scope of analysis to diverse populations.

FIGURES AND TABLES

Table1: Baseline characteristics of Disease outcomes in the UK Biobank WES cohort of 173651 participants.

	Females (N=95412)	**Males (N=78239)**	**Overall (N=173651)**
AGE			
Mean (SD)	56.5 (7.96)	56.9 (8.11)	56.7 (8.03)
Median [Min, Max]	58.0 [40.0, 70.0]	58.0 [38.0, 72.0]	58.0 [38.0, 72.0]
Asthma	10373 (10.9%)	6744 (8.6%)	17117 (9.9%)
Atrial fibrillation and flutter	5227 (5.5%)	8805 (11.3%)	14032 (8.1%)
Cataract	14057 (14.7%)	9646 (12.3%)	23703 (13.6%)
Cholelithiasis and cholecystitis	7101 (7.4%)	3596 (4.6%)	10697 (6.2%)
Coronary atherosclerosis	3922 (4.1%)	9868 (12.6%)	13790 (7.9%)
Diabetes mellitus	6454 (6.8%)	8949 (11.4%)	15403 (8.9%)
Essential hypertension	25912 (27.2%)	28453 (36.4%)	54365 (31.3%)
Gastrointestinal hemorrhage	7316 (7.7%)	6731 (8.6%)	14047 (8.1%)
GERD	11905 (12.5%)	9205 (11.8%)	21110 (12.2%)
Hemorrhoids	5548 (5.8%)	5172 (6.6%)	10720 (6.2%)
Hypercholesterolemia	10970 (11.5%)	14574 (18.6%)	25544 (14.7%)
Hypothyroidism	8939 (9.4%)	2307 (2.9%)	11246 (6.5%)
Obesity	7288 (7.6%)	6210 (7.9%)	13498 (7.8%)
Osteoarthrosis	13670 (14.3%)	8739 (11.2%)	22409 (12.9%)
Pneumonia	5158 (5.4%)	6820 (8.7%)	11978 (6.9%)
Renal failure	6327 (6.6%)	7668 (9.8%)	13995 (8.1%)
Skin cancer	8874 (9.3%)	9021 (11.5%)	17895 (10.3%)
Urinary tract infection	5769 (6.0%)	4741 (6.1%)	10510 (6.1%)

Table 2: Scaled RV heritability estimates of disease status

Trait	**Scaled heritability estimates (%)**	**95% LCL**	**95% UCL**
Hypothyroidism	12.50	1.66	23.35
Asthma	12.46	4.03	20.90
Obesity	9.21	-0.49	18.91
Type 2 diabetes	8.33	-1.12	17.79
Hypercholesterolemia	7.94	1.15	14.72
Hemorrhoids	7.50	-3.67	18.67
Essential hypertension	6.37	1.46	11.28
Gastrointestinal hemorrhage	2.82	-6.64	12.28
Cholelithiasis and cholecystitis	2.48	-8.70	13.65
Coronary atherosclerosis	-0.15	-9.71	9.41
Cataract	-5.86	-12.89	1.17
GERD	-6.40	-13.88	1.08
Renal failure	-9.06	-18.52	0.40
Pneumonia	-11.29	-21.69	-0.89
Urinary tract infection	-17.50	-28.77	-6.23
Atrial fibrillation and flutter	-17.70	-27.13	-8.26
Osteoarthrosis	-20.93	-31.31	-10.54
Skin cancer	-21.14	-29.32	-12.96

Table 3: Genes with significant heritability for each trait

Trait	**Genes with significant RV heritability estimates**	**Non-significant in exome-wide associations in UKB****	**Non-significant in GWAS Catalogue**
Asthma	***EIF4A1***	*EIF4A1*	*EIF4A1*
Atrial Fibrillation and Flutter	***MAD2L1, PEBP1, YPEL2***	*MAD2L1, PEBP1, YPEL2*	*MAD2L1, PEBP1, YPEL2*
Cataract	***SAMD5***	*SAMD5*	*SAMD5*
Cholelithiasis and cholecystitis	***FRS2, XCL2, INPP5K, SLC38A3, LETM1, EDC3, KTN1, JAK3,*** *UGT1A5,* ***NVL, C1QB***	*NA*	*NA*
Coronary atherosclerosis	***LOC101927572, NKX3-1***	*LOC101927572, NKX3-1*	*LOC101927572, NKX3-1*
Gastrointestinal Hemorrhage	***NAIFNA7, TMEM151B***	*NAIFNA7, TMEM151B*	*NAIFNA7, TMEM151B*
Hemorrhoids	***ME1, ZNF619****,* ***RUNDC3B****,* ***MB21D2****,* ***TP53TG3D****,* ***KCNV1****,* ***KIF5C, PALD1, TTL, DHX16, ZC3H18, CDC42, TVP23C-CDRT4, ZNF292***	*NA*	*ME1, ZNF619, RUNDC3B, MB21D2, TP53TG3D, KCNV1, KIF5C, PALD1, TTL, DHX16, ZC3H18, CDC42, TVP23C-CDRT4, ZNF292*
Hypercholesterolemia	*LDLR*		
Hypothyroidism	*TSHR,* ***COL5A2, ZNF583, ZNF490, TATDN1, TMEM256-PLSCR3, REEP5***	*COL5A2, ZNF583, ZNF490, TATDN1, TMEM256-PLSCR3, REEP5*	*COL5A2, ZNF583, ZNF490, TATDN1, TMEM256-PLSCR3, REEP5*
Obesity	***PPP1R21, SKA1, SCARB1***	*PPP1R21, SKA1, SCARB1*	*PPP1R21, SKA1, SCARB1*
Pneumonia	***SRSF2, OR10A7, CPB2, PANK3, BST1, RBAK-RBAKDN, ERLIN2, IDH2***	*SRSF2, OR10A7, CPB2, PANK3, BST1, RBAK-RBAKDN, ERLIN2, IDH2*	*SRSF2, OR10A7, CPB2, PANK3, BST1, RBAK-RBAKDN, ERLIN2, IDH2*
Renal Failure	***SRSF5****, SRSF2,* ***LIX1L, ABCC9, IDH2***	*SRSF5, LIX1L, ABCC9, IDH2*	*SRSF5, SRSF2, LIX1L, ABCC9, IDH2*

Type 2 Diabetes	*GCK*, ***AGXT2, HRH1, FOLH1B, SCO1, MYRF, PDK1,*** *DNMT3A*	*AGXT2, HRH1, FOLH1B, SCO1, MYRF, PDK1, DNMT3A*	*AGXT2, HRH1, FOLH1B, SCO1, PDK1, MYRF*
Urinary tract infection	***CD38, PDCL, VPS41, STMN2, STXBP3, ZNF587B, C12orf43, CBX6, ANO8, AADACL4, TEX22***	*CD38, PDCL, VPS41, STMN2, STXBP3, ZNF587B, C12orf43, CBX6, ANO8, AADACL4, TEX22*	*CD38, PDCL, VPS41, STMN2, STXBP3, ZNF587B, C12orf43, ANO8, AADACL4, TEX22*

NA=trait not listed. Bold font represents genes that are not identified in either GWAS or ExWAS as per **[23] **[26].

Table 4: Variant level association (linear regression). Four *TSHR* variants are significantly associated with hypothyroidism in this study population.

HGVS annotation	**Estimate**	**Standard Error**	**P-value**	**Mutation type**	**Known Disease associations**
TSHR:NM_000369:exon10:c.G1637A:p.W546X	0.0145	0.0024	1.37E-09	stopgain	Hyperthyroidism nonautoimmune/congenital/ nongoitrous ClinVar-Conflicting with 5 submitter showing pathogenic and 1 benign
TSHR:NM_000369:exon10:c.C1532T:p.T511M	0.0102	0.002398	2.05E-05	nonsynonymous SNV	Not applicable
TSHR:NM_000369:exon10:c.C1600T:p.R534C	0.00947	0.002398	7.81E-05	nonsynonymous SNV	Not applicable
TSHR:NM_000369:exon1:c.G122C:p.C41S	0.00941	0.002406	9.25E-05	nonsynonymous SNV	Epilepsy, Developmental delay, autistic features, Hypothyroidism congenital/nongoitrous - Pathogenic in ClinVar

Table 5: Association between the blood plasma biomarker TSHB and *TSHR* in thyroid disease-free population (N=17,107)

GENE / variant ID	CMAC	CMAF	Estimate	Standard error	P-value
TSHR gene burden	450	0.013153	0.113829	0.018124	3.45E-10
14:81143695:G:A_A	14	0.000409	0.875351	0.203134	1.65E-05
14:80955786:G:C_C	292	0.008535	0.168581	0.044804	1.69E-04
14:81143715:G:A_A	4	0.000117	1.358058	0.379324	3.44E-04
14:81143227:G:T_T	2	5.85E-05	1.390554	0.536496	9.55E-03
14:81143283:G:T_T	1	2.92E-05	1.878784	0.759074	1.33E-02
14:81143460:C:T_T	1	2.92E-05	1.809892	0.758792	1.71E-02
14:81143250:T:A_A	1	2.92E-05	1.795922	0.758673	1.79E-02
14:80955841:C:T_T	3	8.77E-05	1.026459	0.43816	1.92E-02
14:81092547:C:G_G	10	0.000292	0.551467	0.239981	2.16E-02
14:81144036:A:G_G	2	5.85E-05	-1.21162	0.536594	2.40E-02

CMAC= cumulative minor allele count, CMAF= cumulative allele frequency, based on CMAC.

Table 6: *TSHR* contributes significantly to other phenotypes.

Traits	**Scaled Heritability**	**P-value**
Hypothyroidism NOS	0.0042	3.74E-18
Hypothyroidism	0.0039	2.35E-17
Sciatica	0.0077	7.27E-10
Right bundle branch block	0.008	1.09E-09
Bronchiectasis	0.0073	3.83E-09
Umbilical hernia	0.0065	5.99E-09
Duodenal ulcer	0.0078	7.60E-09
Pain in limb	0.004	3.08E-08
Acute renal failure	0.0024	3.58E-07

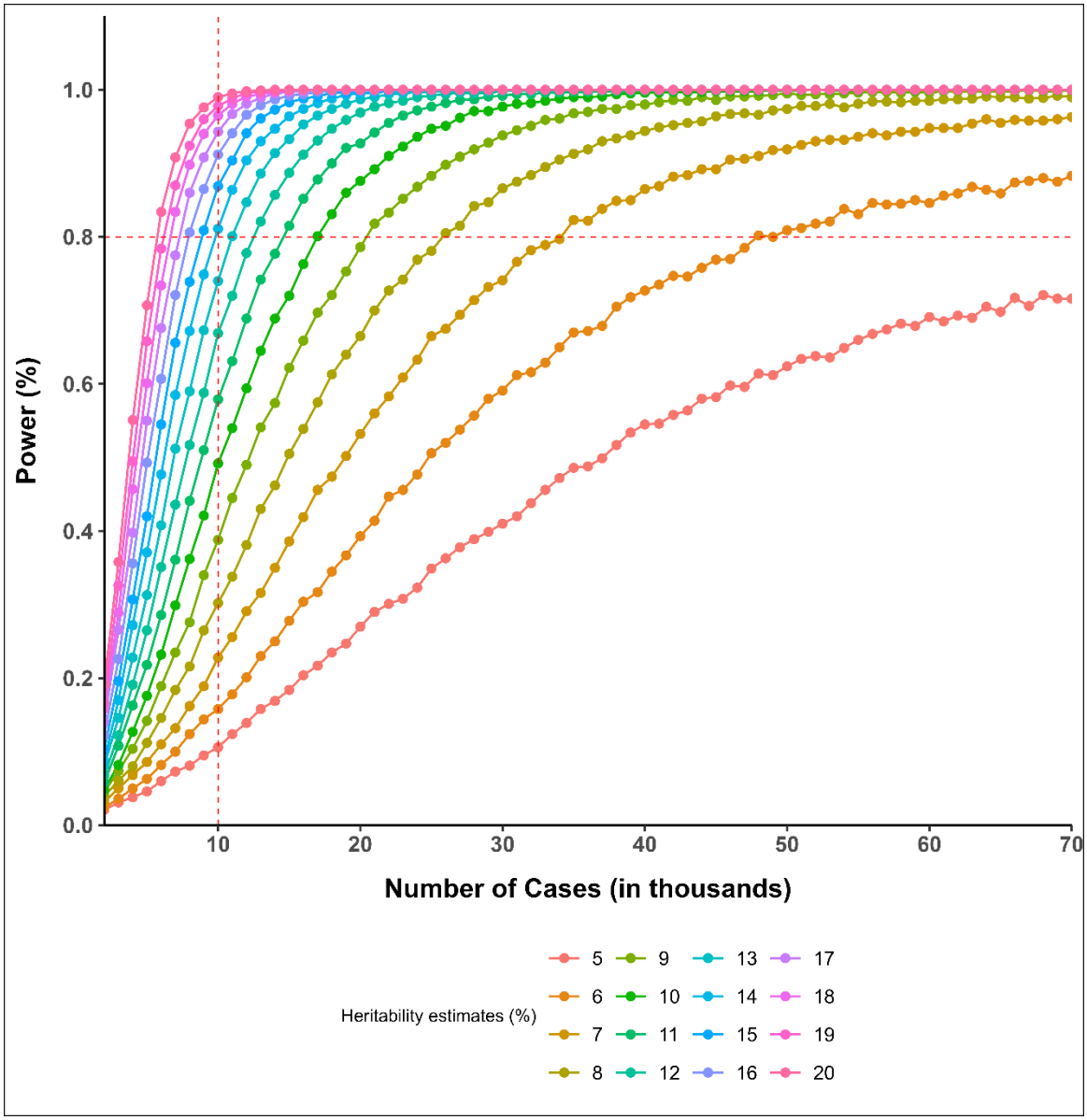


Figure 1: Power of Rare variant heritability estimates on a liability scale.
Based on 173,651British and non-British participants, with the true set $\hat{h}^2_{RV\text{-}liab}$ ranged from 2 to 10%, with increments of 1%. Number of cases varied from 2,000 to 10,000 individuals by increments of 500. The red-dashed horizontal line marks 80% power, while the red-dashed vertical line marks 10,000 cases, as the threshold for traits selection in this study.

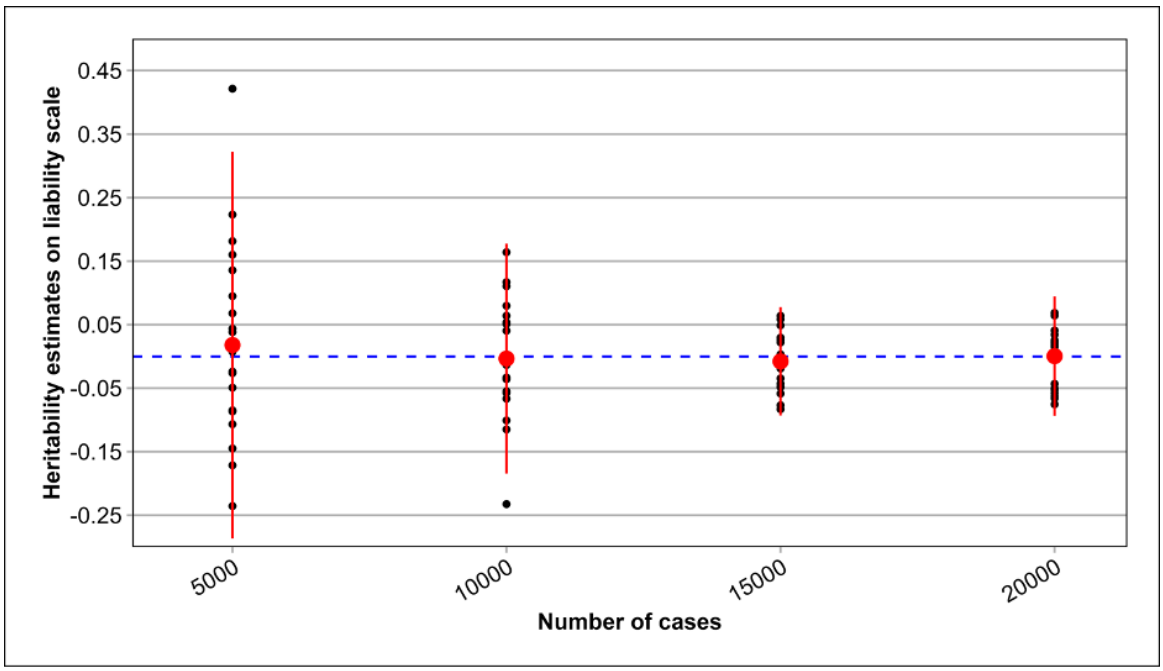


Figure 2: Simulation with permutation.
Here we simulated phenotypes using permutations using various numbers of cases, and always observed $\hat{h}^2_{RV\text{-}liab}$ of mean zero, in the absence of meaningful phenotypes. The central red dot represents the mean $\hat{h}^2_{RV\text{-}liab}$ of 20 simulations via phenotype permutations with the specified number of cases. The red vertical lines represent the standard deviation of the estimates.

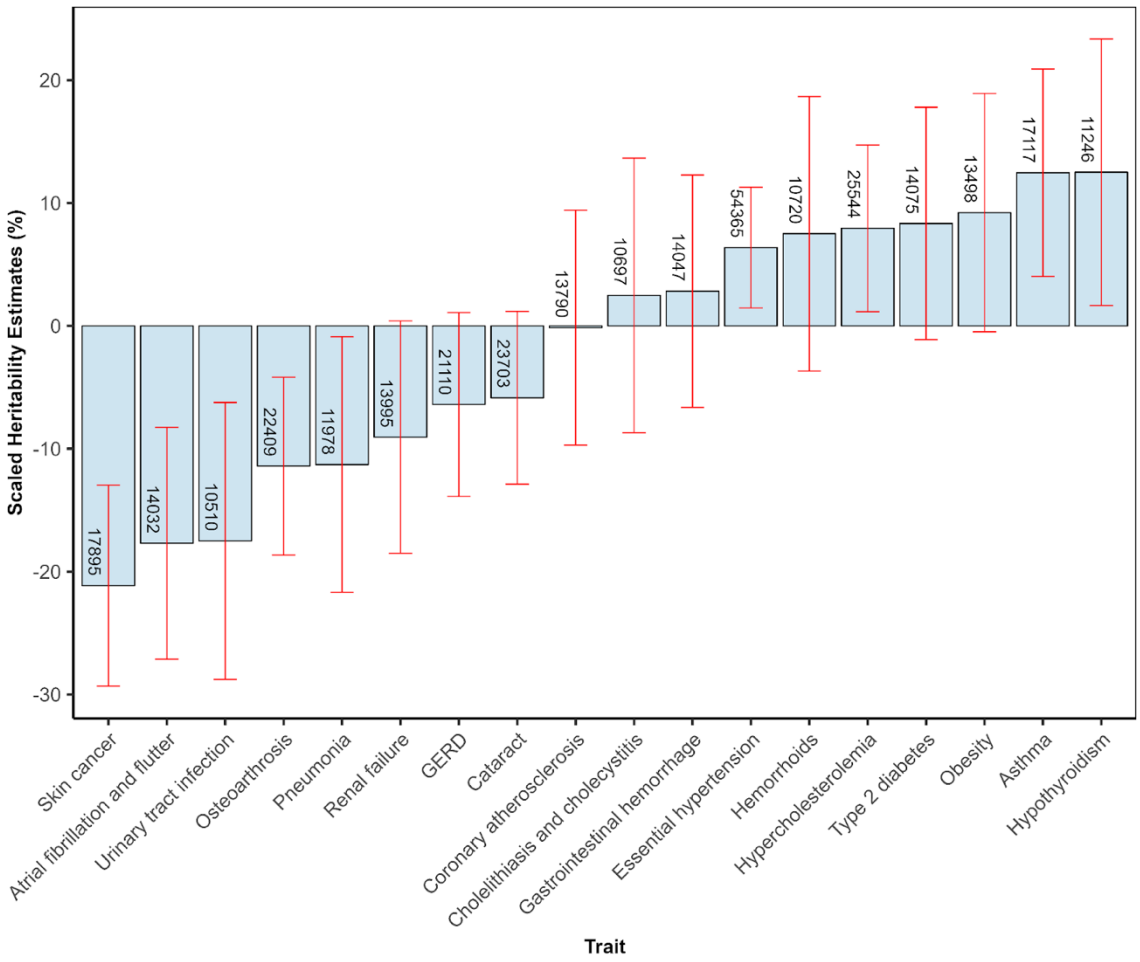
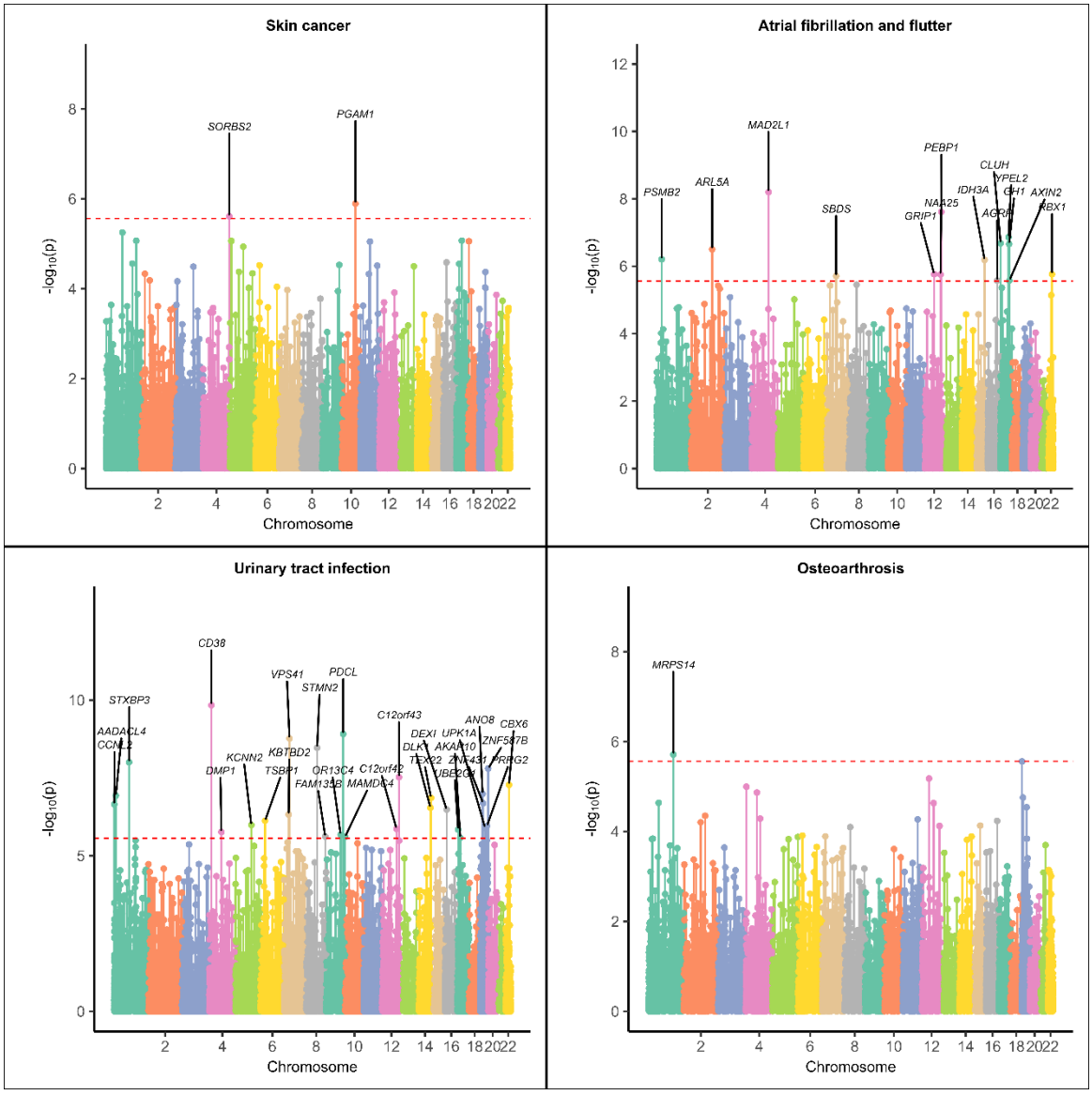
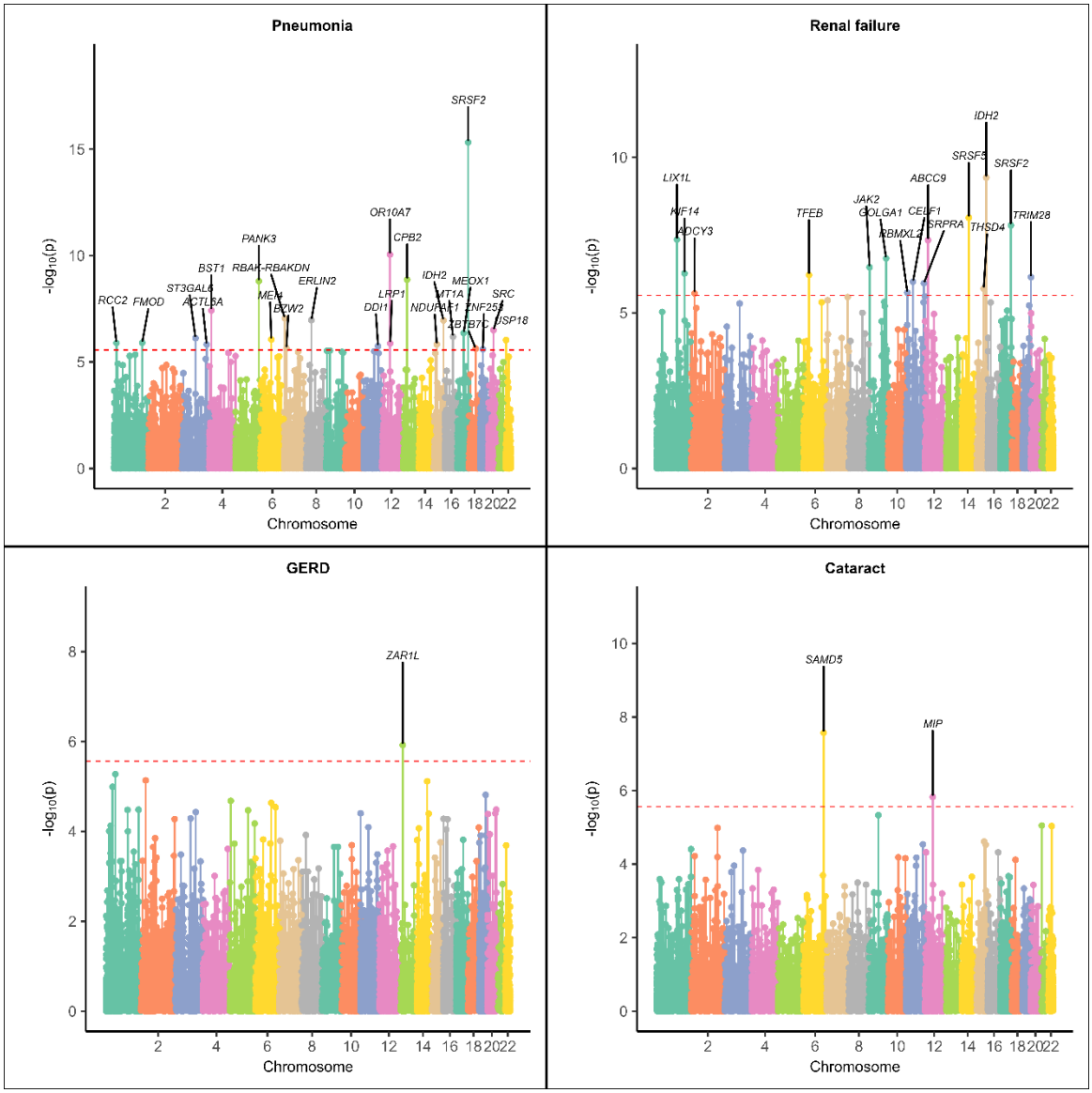
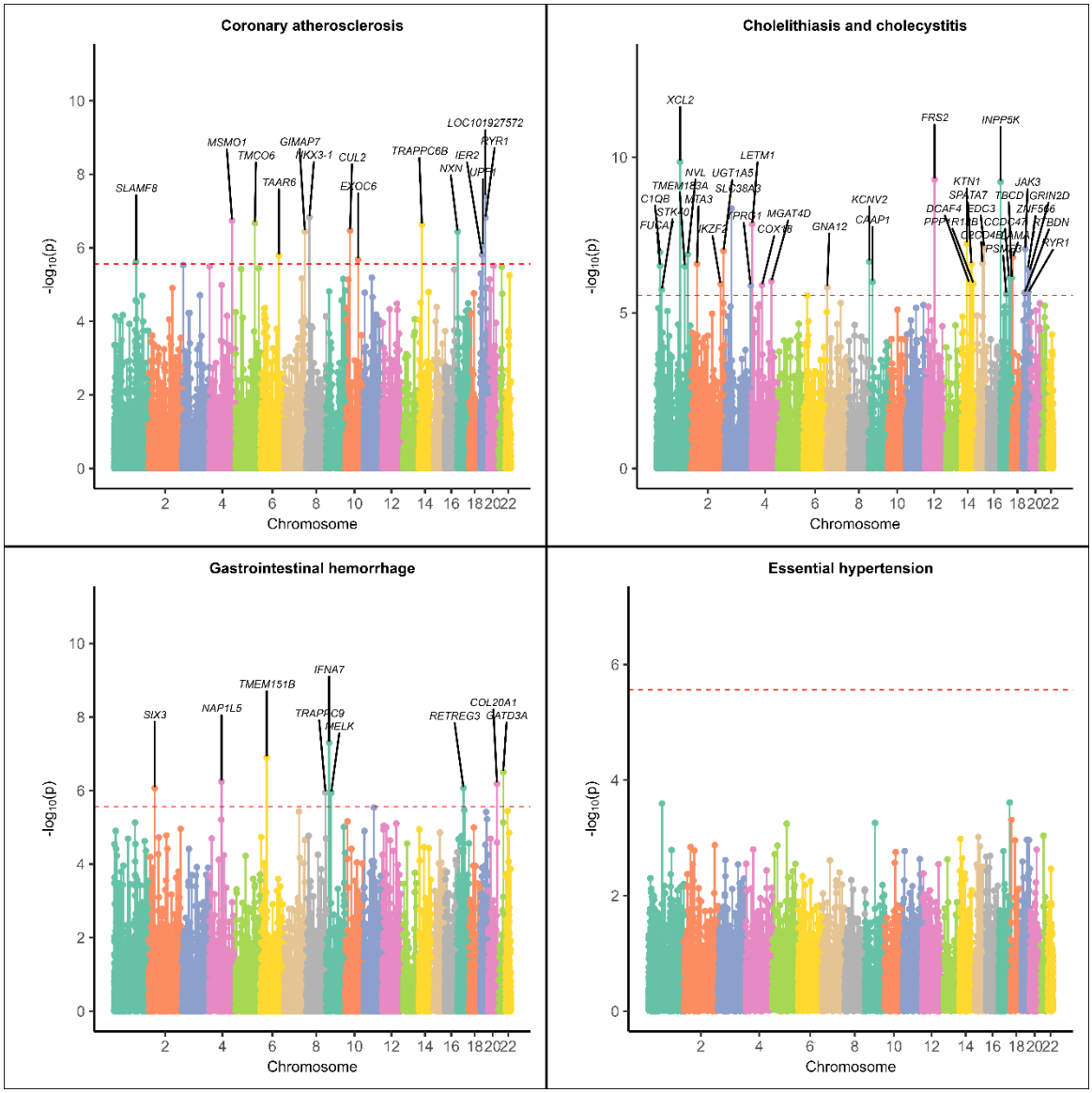


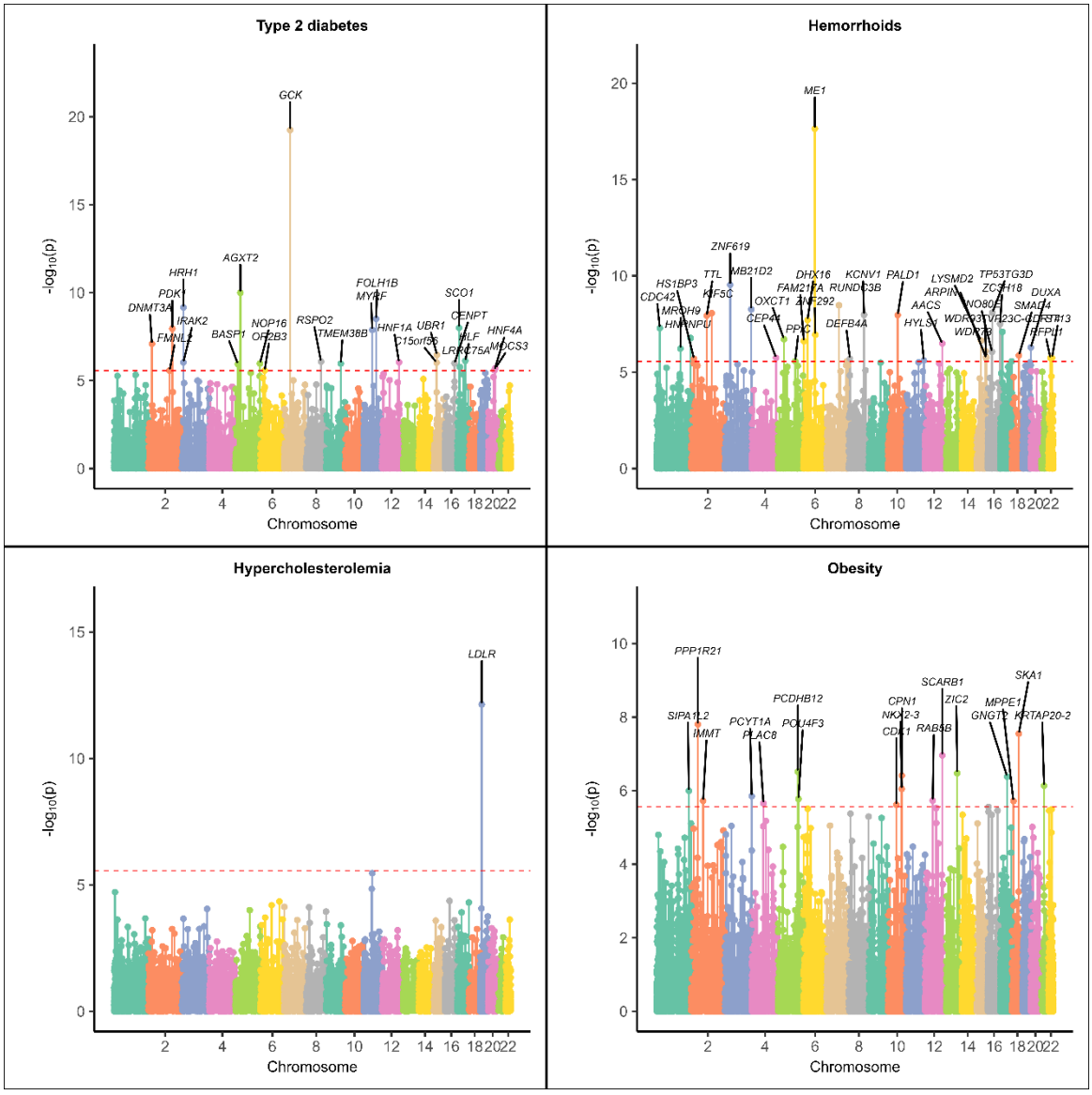
Figure3: Bar chart illustrating the exome-wide RV heritability estimates ($\hat{h}^2_{RV\text{-}liab}$) +/− 95% confidence interval.

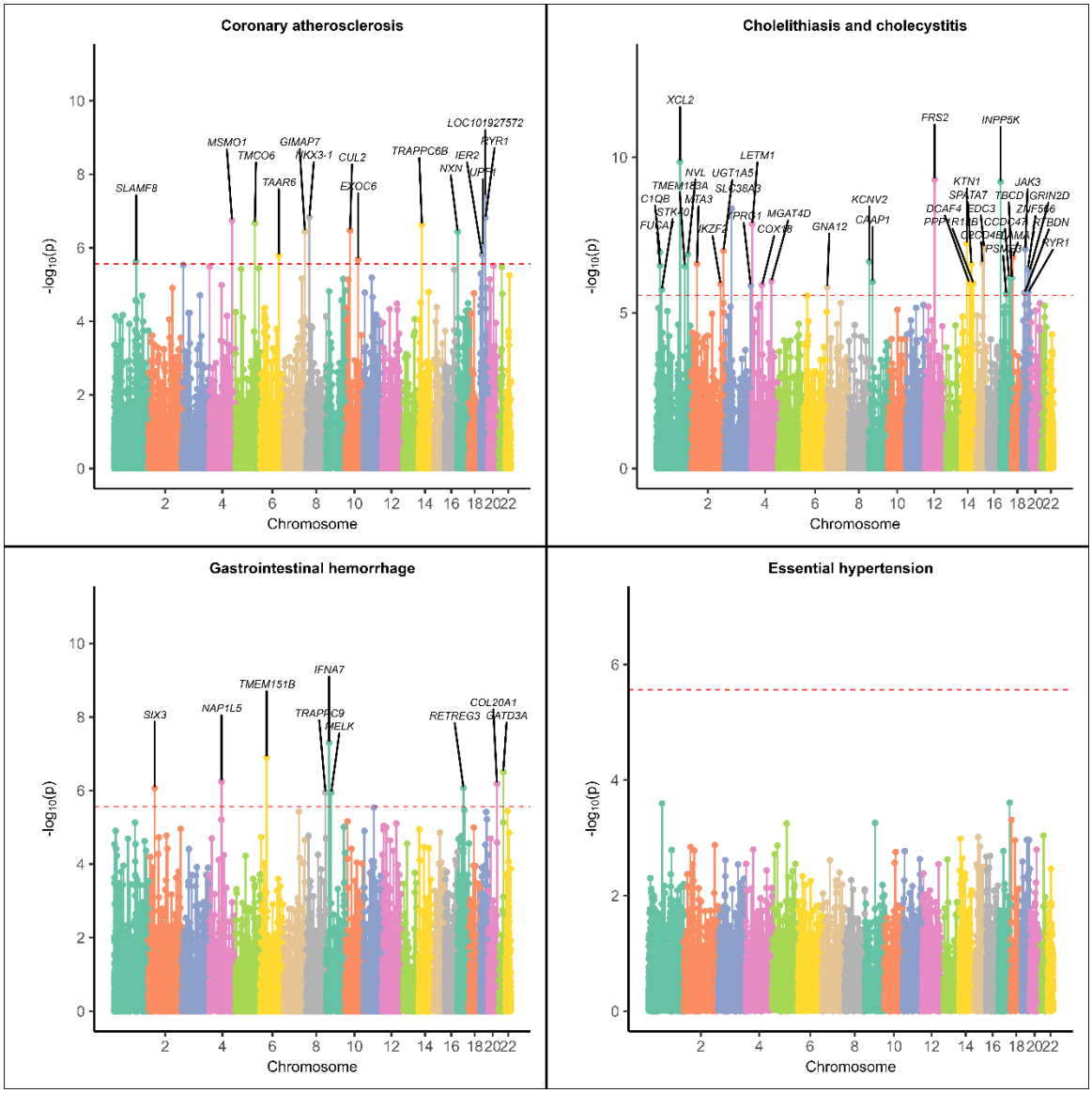
Contribution of 1,606,565 qualifying rare coding variants on 31 complex traits were based on $n = 173{,}651$ Caucasian individuals from the UK Biobank, as estimated using RARity-β. Traits were standardized for age, sex, and the first 20 genetic principal components. 95% confidence interval of trait $\hat{h}^2_{RV\text{-}liab}$ is denoted with red, vertical error bars. The number of cases for each trait is indicated next to the error bar.


Skin cancer
SORBS2
PGAM1
Atrial fibrillation and flutter
MAD2L1
PEBP1
PSMB2
ARL5A
SBDS
CLUH
YPEL2
IDH3A
AXIN2
RBX1
Urinary tract infection
CD38
VPS41
PDCL
STMN2
STXBP3
C12orf43
ANO8
CBX6
AADACL4
DMP1
KCNN2
KBTBD2
TSBP1
OR13C4
C12orf42
FAM133B
MAMDC4
ZNF587B
Osteoarthrosis
MRPS14
-log10(p)
Chromosome


Pneumonia
SRSF2
OR10A7
CPB2
PANK3
BST1
RBAK-RBAKDN
ERLIN2
IDH2
MEOX1
ST3GAL6
ACTL8A
MEI4
LRP1
SRC
RCC2
FMOD
BZW2
DDI1
MT1A
USP18
-log10(p)
Chromosome
Renal failure
IDH2
SRSF5
SRSF2
LIX1L
ABCC9
JAK2
KIF14
TFEB
GOLGA1
CELF1
TRIM28
ADCY3
SRPRA
THSD4
RBMXL2
GERD
ZAR1L
Cataract
SAMD5
MIP


Coronary atherosclerosis
SLAMF8
MSMO1
TMCO6
TAAR6
GIMAP7
IKX3-1
CUL2
EXOC6
TRAPPC6B
IER2
NXN
UPF1
RYR1
LOC101927572
-log10(p)
Chromosome
Cholelithiasis and cholecystitis
XCL2
C1QB
STK40
FUCA1
TMEM163A
NVL
MTA3
IKZF2
UGT1A5
SLC38A3
LETM1
PRG1
MGAT4D
COX18
GNA12
KCNV2
CAAP1
FRS2
DCAF4
PPP4R1B
INPP5K
KTN1
SPATA7
EDC3
CCDC47
QRICH1
PSMA6
TBCD
JAK3
GRIN2D
ZNF506
RTBDN
RYR1
-log10(p)
Chromosome
Gastrointestinal hemorrhage
SIX3
NAP1L5
TMEM151B
IFNA7
TRAPPC9
MELK
COL20A1
RETREG3
GATD3A
-log10(p)
Chromosome
Essential hypertension
-log10(p)
Chromosome


Type 2 diabetes
GCK
AGXT2
HRH1
PDK1
DNMT3A
IRAK2
FMNL2
BASP1
NOP16
OR2B3
RSPO2
TMEM38B
FOLH1B
MYRF
HNF1A
UBR1
C15orf56
SCO1
CENPT
WLF
LRRC75A
HNF4A
MOCS3
-log10(p)
Chromosome
Hemorrhoids
ME1
ZNF619
HS1BP3
CDC42
MRPH9
HMGNP1
TTL
KIF5C
MB21D2
OXCT1
CEP44
FAM21A
ANKF292
PPIC
DHX16
DEFB4A
KCNV1
RUNDC3B
PALD1
HYLS1
LYSMD2
ARPIN
AACS
TP53TG3D
ZC3H18
WDR73
DUXA
SMAD4
RFPL1
-log10(p)
Chromosome
Hypercholesterolemia
LDLR
-log10(p)
Chromosome
Obesity
PPP1R21
SIPA1L2
IMMT
PCYT1A
PLAC8
PCDHB12
POU4F3
CPN1
NKX2-3
CDH1
SCARB1
RAB3B
ZIC2
MPPE1
GNGT2
SKA1
KRTAP20-2
-log10(p)
Chromosome


Coronary atherosclerosis
SLAMF8
MSMO1
TMCO6
TAAR6
GIMAP7
IKX3-1
CUL2
EXOC6
TRAPPC6B
IER2
NXN
UPF1
RYR1
LOC101927572
Cholelithiasis and cholecystitis
XCL2
FUCA1
STK40
C1QB
TMEM183A
NVL
MTA3
UGT1A5
SLC38A3
IKZF2
LETM1
PRG1
COX18
MGAT4D
GNA12
KCNV2
CAAP1
FRS2
DCAF4
PPP4R1B
KTN1
SPATA7
EDC3
CCDC47
QRICH1
PSMA6
INPP5K
TBCD
JAK3
GRIN2D
ZNF506
RTBDN
RYR1
Gastrointestinal hemorrhage
SIX3
NAP1L5
TMEM151B
IFNA7
TRAPPC9
MELK
COL20A1
RETREG3
GATD3A
Essential hypertension
-log10(p)
Chromosome

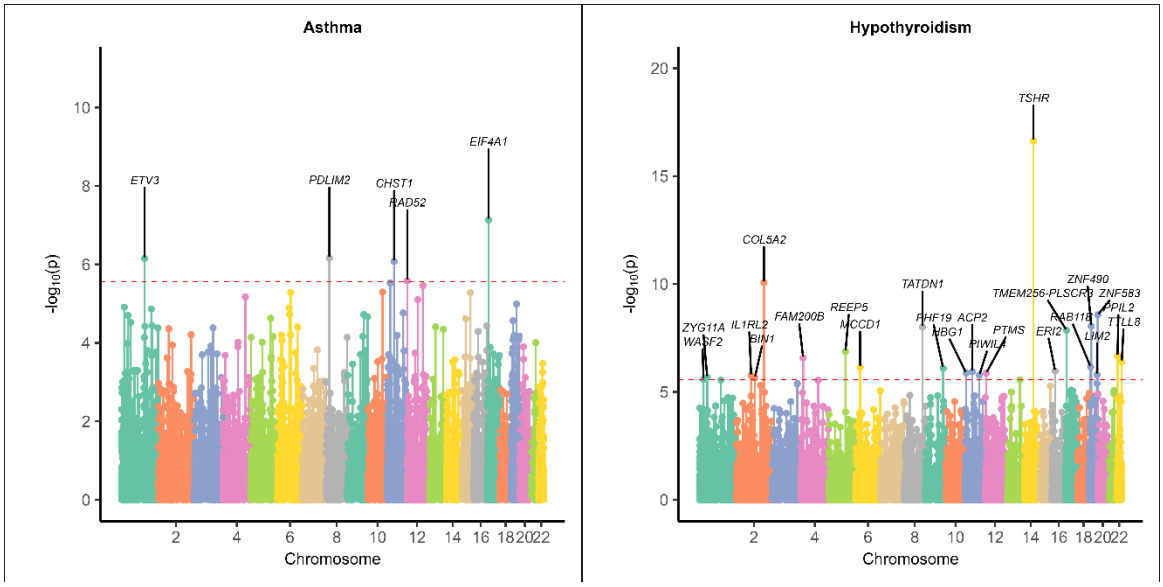


Figure 4: RARity-β to determine $\hat{h}^2_{RV\text{-}liab\text{-}gene}$ for all genes with qualifying RVs.
Each dot represents a single gene, with genes ordered on the x-axis according to their genomic position. The y-axis represents the significance of $\hat{h}^2_{RV\text{-}liab\text{-}gene}$ measured as -$\log_{10}$ transformed p-values, where the p-values were derived using F-test. Red, horizontal dashed lines mark the Bonferroni's p-value significance threshold corrected for 18,169 genes (p-value $< 2.75 \times 10^{-6}$). Genes with significant $\hat{h}^2_{RV\text{-}liab\text{-}gene}$ are labeled.

Figure 5: Gene-disease heatmap, for the genes contributing significantly to the RV heritability of hypothyroidism. The intensity of color is proportional to the strength of evidence for gene-disease-association, with darker color representing a greater level of evidence.

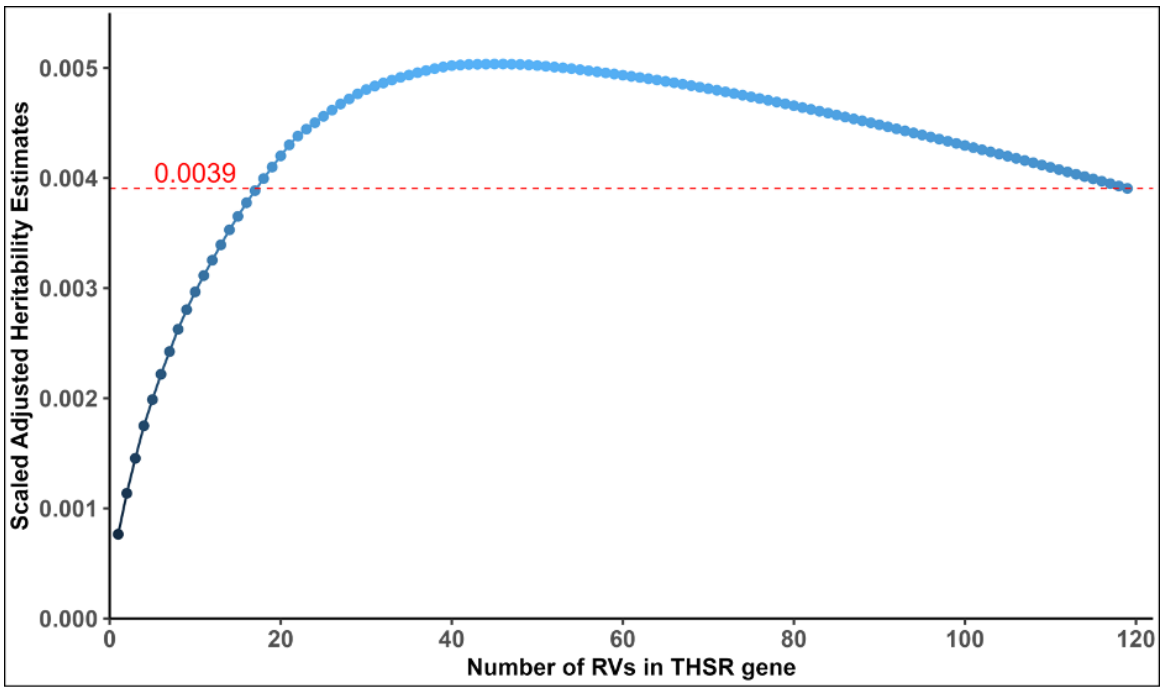


Figure 6: Rare variants in the *TSHR* gene contributing to the entire heritability estimate conferred by the gene on hypothyroidism. Effect of increasing the number of variants in order of most to least significant heritability contributions on total heritability. At least 17 variants contribute to the overall RV heritability inferred by the *TSHR* gene.

REFERENCES

1 Cao, C. *et al.* RAVAR: a curated repository for rare variant-trait associations. *Nucleic Acids Res* **52**, D990-D997 (2024). https://doi.org/10.1093/nar/gkad876

2 Karczewski, K. J. *et al.* Systematic single-variant and gene-based association testing of thousands of phenotypes in 394,841 UK Biobank exomes. *Cell Genom* **2**, 100168 (2022). https://doi.org/10.1016/j.xgen.2022.100168

3 Wang, Q. *et al.* Rare variant contribution to human disease in 281,104 UK Biobank exomes. *Nature* **597**, 527-532 (2021). https://doi.org/10.1038/s41586-021-03855-y

4 Zuk, O. *et al.* Searching for missing heritability: Designing rare variant association studies. *Proceedings of the National Academy of Sciences* **111**, E455-E464 (2014). https://doi.org/10.1073/pnas.1322563111

5 Lee, D. S. M. *et al.* Common- and rare-variant genetic architecture of heart failure across the allele frequency spectrum. *medRxiv* (2023). https://doi.org/10.1101/2023.07.16.23292724

6 Weiner, D. J. *et al.* Polygenic architecture of rare coding variation across 394,783 exomes. *Nature* **614**, 492-499 (2023). https://doi.org/10.1038/s41586-022-05684-z

7 Burch, K. S. *et al.* Partitioning gene-level contributions to complex-trait heritability by allele frequency identifies disease-relevant genes. *Am J Hum Genet* **109**, 692-709 (2022). https://doi.org/10.1016/j.ajhg.2022.02.012

8 Pathan, N. *et al.* A method to estimate the contribution of rare coding variants to complex trait heritability. *Nat Commun* **15**, 1245 (2024). https://doi.org/10.1038/s41467-024-45407-8

9 Lee, S. H., Wray, N. R., Goddard, M. E. & Visscher, P. M. Estimating missing heritability for disease from genome-wide association studies. *Am J Hum Genet* **88**, 294-305 (2011). https://doi.org/10.1016/j.ajhg.2011.02.002

10 Ojavee, S. E., Kutalik, Z. & Robinson, M. R. Liability-scale heritability estimation for biobank studies of low-prevalence disease. *Am J Hum Genet* **109**, 2009-2017 (2022). https://doi.org/10.1016/j.ajhg.2022.09.011

11 Team, R. C. *R: A language and environment for statistical computing*, <https://www.R-Project.org> (2019).

12 Regier, A. A. *et al.* Functional equivalence of genome sequencing analysis pipelines enables harmonized variant calling across human genetics projects. *Nat Commun* **9**, 4038 (2018). https://doi.org/10.1038/s41467-018-06159-4

13 Zhou, W. *et al.* Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat Genet* **50**, 1335-1341 (2018). https://doi.org/10.1038/s41588-018-0184-y

14 Yu, R., Zhang, H. & Xiao, X. Partial GCK gene deletion mutations causing maturity-onset diabetes of the young. *Acta Diabetol* **61**, 107-115 (2024). https://doi.org/10.1007/s00592-023-02173-1

15 Bonnefond, A. *et al.* Monogenic diabetes. *Nat Rev Dis Primers* **9**, 12 (2023). https://doi.org/10.1038/s41572-023-00421-w

16 Goldstein, J. L. & Brown, M. S. The LDL receptor. *Arterioscler Thromb Vasc Biol* **29**, 431-438 (2009). https://doi.org/10.1161/ATVBAHA.108.179564

17 Hu, X., Wu, P., Liu, B., Lang, Y. & Li, T. RNA-binding protein CELF1 promotes cardiac hypertrophy via interaction with PEBP1 in cardiomyocytes. *Cell Tissue Res* **387**, 111-121 (2022). https://doi.org/10.1007/s00441-021-03541-5

18 Uhlén, M. *et al.* A human protein atlas for normal and cancer tissues based on antibody proteomics. *Mol Cell Proteomics* **4**, 1920-1932 (2005). https://doi.org/10.1074/mcp.M500279-MCP200

19 Post, J. A., Bijvelt, J. J. & Verkleij, A. J. Phosphatidylethanolamine and sarcolemmal damage during ischemia or metabolic inhibition of heart myocytes. *Am J Physiol* **268**, H773-780 (1995). https://doi.org/10.1152/ajpheart.1995.268.2.H773

20 van Bavel, J. J. A., Vos, M. A. & van der Heyden, M. A. G. Cardiac Arrhythmias and Antiarrhythmic Drugs: An Autophagic Perspective. *Front Physiol* **9**, 127 (2018). https://doi.org/10.3389/fphys.2018.00127

21 Zhou, J., Dong, Y., Cai, X., Yang, H. & Guo, T. Identification and Validation of Autophagy-Related Genes as Potential Biomarkers and Therapeutic Targets in Atrial Fibrillation. *Int J Gen Med* **14**, 7783-7796 (2021). https://doi.org/10.2147/IJGM.S337855

22 Liu, A., Jia, K., Liang, H. & Jin, Q. Comprehensive analysis of autophagy-related genes and patterns of immune cell infiltration in valvular atrial fibrillation. *BMC Cardiovasc Disord* **21**, 132 (2021). https://doi.org/10.1186/s12872-021-01939-1

23 Alberti, L. *et al.* Germline mutations of TSH receptor gene as cause of nonautoimmune subclinical hypothyroidism. *J Clin Endocrinol Metab* **87**, 2549-2555 (2002). https://doi.org/10.1210/jcem.87.6.8536

24 de Roux, N. *et al.* A neomutation of the thyroid-stimulating hormone receptor in a severe neonatal hyperthyroidism. *J Clin Endocrinol Metab* **81**, 2023-2026 (1996). https://doi.org/10.1210/jcem.81.6.8964822

25 Jordan, N. *et al.* The W546X mutation of the thyrotropin receptor gene: potential major contributor to thyroid dysfunction in a Caucasian population. *J Clin Endocrinol Metab* **88**, 1002-1005 (2003). https://doi.org/10.1210/jc.2002-021301

26 Sollis, E. *et al.* The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. *Nucleic Acids Res* **51**, D977-D985 (2023). https://doi.org/10.1093/nar/gkac1010

CHAPTER 6:
DISCUSSION

CHAPTER 6: DISCUSSION

6.1 GENERAL OVERVIEW

This thesis focuses on the contributions of rare genetic variants to complex traits and diseases. The syndromic nature of diseases, diverse pathology, polygenicity and the lack of appropriate genetic tools to study complex traits complicate genetic research. In study 1 (chapter 3), the genetic determinants of vascular dementia were examined, allowing a comprehensive understanding of how genetic methodologies are currently applied to complex traits. Study 2 (Chapter 4) describes the developed of a rare variants heritability estimator (RARity) to examine the contributions of rare variants to complex continuous traits, including 5 anthropomorphic and 26 blood biomarkers in the UKB. In study 3 (Chapter 5) RARity method was extended to develop RARity-β, to study binary trait heritability on a liability scale and its implications in 18 diseases in the UKB. In the subsequent sections, I will i) summarize the main findings from each chapter, ii) outline the significance, and iii) describe the limitations, and future areas of investigation.

6.2 STUDY 1 (CHAPTER 3)

6.2.1 SUMMARY

Study 1, "Genetic determinants of vascular dementia", is a review article exploring recent genetic research on VaD, utilizing methods such as genome-wide association studies (GWAS), polygenic risk scores (PRS), heritability estimates, and family studies for Mendelian forms of the disease. It highlights key genetic associations and potential pathways involved in VaD, focusing on pathological risk factors like stroke, cerebral small vessel diseases, and cerebral amyloid angiopathy. The review also emphasizes significant modifiable risk factors, such as hypertension, diabetes, and dyslipidemia, advocating for a multifactorial approach in VaD prevention and treatment. Furthermore, it identifies areas for scientific advancement to enhance clinical care, suggesting that large-scale collaborations and an integromics approach can strengthen genetic discoveries.

Understanding the genetics of VaD and its risk factors could revolutionize its definition based on molecular mechanisms, leading to new diagnostic, prognostic, and therapeutic tools.

6.2.2 SIGNIFICANCE

While clinical manifestations provide a practical framework for diagnosis of a highly heterogeneous complex diseases such as VaD, genetics offer insights into the molecular mechanisms and shows that diseases are often interconnected at a molecular level, challenging the rigid boundaries of clinical classifications. For example, the *APOE ε4* common variant is implicated in CAA, and AD; and the *NBEAL1* is associated with stroke, WMH, and AD[1]. Furthermore, the current gold standard for diagnosing VaD, which involves MRI-based identification of cerebrovascular pathology, faces practical challenges in settings lacking adequate facilities, expertise, or resources. Genetic studies that provide insights into the biological mechanisms, through understanding both common and rare variant contributions, as well as other omics, can enable earlier predictions and potentially lead to the discovery of diagnostic biomarkers that are more accessible and therapeutic targets that are more personalized.

It has been shown that CV contributions to pathological risk factors of VaD ranges from 11.9% to 40% for CAA and stroke, respectively while twin and family studies propose heritability estimates for WMH lesion volume between 50% and 80%[2,3]. Blood pressure, type 2 diabetes, LDL-cholesterol, triglycerides, atrial fibrillation, diet, BMI, exercise, and smoking are epidemiologically associated with greater VaD risk and have significant estimated CV heritabilities at 15%, 66%, 8.3%, 21.8%, 14.4%, 4.8%, 24.9%, 7%, and 15.1%, respectively[4,5].From this, it is apparent that a comprehensive understanding of the genetic determinants requires examining the genetic factors affecting the disease directly, as well as the pathological and modifiable risk factors. Advancements in genomic methodologies related to common variants, has been a great starting point for understanding

the genetic determinants of VaD, consequently, numerous genetic associations and pathways have been implicated in VaD and its risk factor.

6.2.3 LIMITATIONS AND FUTURE OPPORTUNITIES

Despite being the second most common cause of dementia, research on the genetics of vascular dementia (VaD) is limited. Major challenges to studying and implementing the genetics of VaD include a lack of large-scale, multi-ethnic consortia that examines both genetic and non-genetic aspects of VaD. Many genetic associations for VaD come from small studies with limited replication. The lower prevalence of VaD and underrepresentation of multi-ethnic groups in research further complicate efforts. Additionally, the studies on the genetics of VaD spans decades, marked by various technological advancements and study designs, raising concerns about gene validity. Importantly, beyond pedigree and association studies, not much is known about the oligogenic or polygenic contributions of rare variants to the risk factors of VaD.

With technological advances, it is now possible to navigate the above-mentioned challenges. Accessible, focused consortia that integrate health records and biomarker data, and utilize advanced statistical tools to combine data from various sources and multi-ethnic studies, will overcome the limitation of sample size and accelerate research in this area. Furthermore, studies spanning decades of evidence need to be evaluated with a standardized framework, such as those created by the Clinical Genome Resource (ClinGen), to ascertain gene-disease validity and aid gene prioritization in clinical contexts, improving genetic data usability and dissemination. Finally, there is a need to develop and implement appropriate methods to study rare variant contributions, such as heritability, MR, and PRS. This has been one of the motivating factors for study 2 and 3 in this thesis, which describes the developments and application of methods to estimate rare variant contributions to complex traits. Integration of genetics with other multi-omics data would provide a holistic approach to truly understanding VaD.

6.3 STUDY 2 (CHAPTER 4)

6.3.1 SUMMARY

Study 2 (chapter 4) describes a novel method to estimate rare variant heritability from whole exome sequence data, “RARity”, and its application to the UK Biobank Caucasian population. This method is versatile and does not require prior assumptions about the genetic architecture of selected variants, and thus has been applied to both common and rare variants.

Several discoveries were made with RARity for continuous traits, enhancing our understanding of the genetic architecture of RVs, identifying traits with significant RV contributions, prioritizing genes based on RV heritability and also assess the capacity of current methods to capture trait variance, such as gene-burden testing and pathogenicity scores. When applied to 31 complex, continuous traits in the UKB, I discovered that RVs act in long range LD, at a much greater distance compared to common variants, and thus a stringent LD pruning was required to avoid overestimation of h^2_{RV}. Secondly, I discovered that gene burden testing, which aggregates RVs, leads to a substantial loss of information compared to the use of unaggregated variants in RARity. Furthermore, RARity offers practical advantages in characterizing genes based on RV heritability. For most traits RV account for a significant portion of heritability (>5% in 27 of the 31 traits), so much so that, incorporation of RVs almost perfectly recapitulated the "missing heritability" for 11 traits, reaffirming conclusions from recent studies using whole genome sequencing (WGS) data. With RARity we were able to assess and identify the limited ability of pathogenicity scores to distinguish variants with biological effects from those without. This emphasizes the need for alternative methods to study RV functional consequences.

6.3.2 SIGNIFICANCE

RARity proved to be a fast, robust and versatile tool for rare variant studies. It allows estimation of RV heritability at the genome-wide level, and gene level. Thus, it was demonstrated that a substantial contribution of RVs to trait heritability, while enabling assessment of variant level and gene-level characteristics of RVs. More importantly, the identification of genes significantly enriched for heritability offered a hypothesis free method of gene discovery. Finally, utility of RARity was demonstrated with evaluation of pathogenicity scores, discussing a key flaw in current algorithms to distinguish truly pathogenic variants from those without any consequences. Ultimately, the study emphasizes the need to incorporate RV analysis into genetic studies of complex traits.

6.3.3 LIMITATIONS AND FUTURE OPPORTUNITIES

This study has several limitations. The exclusion of non-coding, singleton, and doubleton variants, along with LD pruning of potentially functional variants, may lead to an underestimation of $\hat{h}^2_{RV}$. The contribution of rare non-coding versus coding variants is not well understood and is challenging to assess, but calibrating RARity for WGS is a future opportunity to provide insights. It's also important to recognize that heritability describes phenotypic variance at the population level and cannot predict individual variations. Therefore, RARity is intended to complement, rather than replace, RV association methods.

While no significant differences in RV heritability were found between males and females, larger sample sizes are needed for more accurate analysis. This study only explored some pathogenicity scores, indicating that further research using RARity could help evaluate functional versus non-functional variants. Additionally, the RARity model in this study suited for continuous traits, would need adjustments for dichotomous traits, which is the focus of study 3 (Chapter 5).

6.4 STUDY 3 (CHAPTER 5)

6.4.1 SUMMARY

Rare variant heritability estimator-β (RARity-β), an adaptation of RARity for continuous traits, was developed to estimate estimates the overall ($\hat{h}^2_{RV\text{-}liab}$) and gene-level heritability ($\hat{h}^2_{RV\text{-}liab\text{-}gene}$) of binary traits on a liability scale. This method was applied to 18 binary traits using WES data from 173,688 Caucasian participants in the UK Biobank. Significant heritability was found for hypothyroidism, asthma, hypercholesterolemia, and essential hypertension, with respective $\hat{h}^2_{RV\text{-}liab}$ of 12.5%, 12.5%, 7.9%, and 6.4%. Gene-level analysis revealed 77 genes with significant $\hat{h}^2_{RV\text{-}liab\text{-}gene}$ for one or more traits, identifying 70 new gene-trait relationships. The *TSHR* gene, notably contributing to hypothyroidism heritability, was linked to increased thyroid-stimulating hormone levels and associated with sciatica, suggesting hypothyroidism as an under-recognized cause of sciatica.

6.4.2 SIGNIFICANCE

The significance of RARity-β is highlighted by its ability to capture significant overall heritability estimates for four of the 18 diseases, along with well-established gene-disease relationships. Notably, the analyses identified several compelling examples of such associations. For instance, the gene *GCK* is significantly associated with Type 2 diabetes, a finding consistent with its known role in glucose metabolism. Similarly, the *LDLR* gene is linked to hypercholesterolemia, disrupting lipid metabolism and increasing cardiovascular risk. The autophagy related *PEBP1* gene shows a significant association with atrial fibrillation and flutter, and could be a potential therapeutic target, pending further investigation.

A particularly noteworthy finding is the highly significant association between *TSHR* and hypothyroidism. Hypothyroidism, a prevalent endocrine disorder characterized by insufficient thyroid hormone production, has diverse clinical manifestations. Four nonsynonymous variants in the TSHR gene were significantly associated with

hypothyroidism. Among these, variants previously implicated in congenital hypothyroidism were reaffirmed at the population level. This study shows the polyallelic nature of rare variants, revealing that multiple variants within the TSHR gene contribute to overall heritability estimates. A number of variants and the *TSHR* gene-burden are positively associated with the TSHB levels in a population devoid of hypothyroidism, confirming the functional consequence of the RVs in *TSHR*. Furthermore, the *TSHR* gene is also found to contribute significantly towards sciatica, indicative of hypothyroidism being an unrecognized cause of sciatica.

6.4.3 LIMITATIONS AND FUTURE OPPORTUNITIES

This study has several limitations in addition to those described in Study 2. The sample size posed a constraint, limiting our ability to detect heritability estimates below 14% for traits with a minimum of 10,000 cases. This emphasizes the need for caution when interpreting results, especially when the true heritability is below this threshold and observed heritability on liability scales is negative. Increasing sample sizes to approximately 1 million WES by pooling data from several large biobanks or refining analytical strategies may enhance the sensitivity of our method to detect lower levels of heritability. Expanding the scope of analysis to diverse populations and additional trait categories will further advance our understanding of the genetic architecture of common diseases.

Overall, our findings support the feasibility and validity of using the RARity-β model with WES data to estimate the heritability of complex binary traits explained by RVs. By leveraging large-scale genotype data and carefully selecting traits with sufficient statistical power, we have demonstrated the utility of our approach in advancing our understanding of the genetic architecture of common diseases.

6.5 FINAL REFLECTIONS AND CONCLUSION

This thesis began by examining statistical genetic methods to study complex traits. Overall, through three studies, significant advancements have been made in understanding the

genetic contributions to complex diseases, by developing novel methodologies for estimating heritability from rare variants and applying these methods to both continuous and binary traits.

The first study reviewed the genetic determinants of vascular dementia, providing insights into genetic associations, genetic risk scores, causality through Mendelian randomization, pathways, and the impact of modifiable risk factors. This study underscores the potential of genetics to redefine VaD based on molecular mechanisms, leading to improved diagnostic, prognostic, and therapeutic strategies. The second study introduced the RARity method, a novel tool for estimating rare variant heritability from whole exome sequence data. Applied to continuous traits in the UK Biobank, RARity demonstrated significant contributions of rare variants to heritability, offering a hypothesis-free approach for gene discovery and highlighting the limitations of current pathogenicity scores. By the end of this study, I as able to measure the contributions of rare coding variants to several risk factors of VaD, including ApoA-I, BMI, diastolic/systolic blood pressure, LDL-cholesterol, and triglycerides, with estimated rare variant heritability at 4.6%, 9.9%, 8.6-6.8%, 8.7%, and 8.6%, respectively. The third study extended the RARity method to binary traits through the development of RARity-β. This adaptation enabled the estimation of heritability on a liability scale for 18 diseases and identifying numerous gene-trait associations, including new gene discoveries for common diseases like hypothyroidism and atrial fibrillation. Identification of significant gene-trait associations, such as the role of *TSHR* in hypothyroidism and sciatica, and *PEBP1* in atrial fibrillation, underscores the potential for these genetic insights to inform clinical practice and therapeutic development.

The collective findings from these studies emphasize the important role of rare variants in the genetic architecture of complex traits and diseases. They also highlight the necessity of integrating rare variant analysis into genetic research, leveraging large-scale biobank data, and applying advanced statistical tools to enhance our understanding of complex traits. The development of RARity and RARity-β provides robust tools for rare variant analysis,

facilitating gene discovery and enhancing the understanding of genetic contributions to both continuous and binary traits.

Despite these advances, the studies face several limitations. The limited sample size and underrepresentation of multi-ethnic groups in some analyses highlight the need for larger, more diverse cohorts to improve the robustness and generalizability of findings. Singletons and doubletons were excluded in the studies because they are difficult to analyze with confidence due to their low frequency, leading to high variability and noise in the results, as they often exhibit considerable heterogeneity in their effects. The exclusion of sex-chromosomes, non-coding, singleton, and doubleton variants in some analyses may underestimate heritability, necessitating future research to incorporate these variants and explore their functional consequences.

Singletons and doubletons are typically excluded from genetic studies due to their low frequency and the challenges they present in analysis. These variants offer limited statistical power and are prone to high variability in results, leading to significant uncertainty in their association with traits or diseases. This rarity may increase the risk of false positive results. Although aggregation of singletons has been associated with several phenotypes[6,7], it is difficult to distinguish the truly functional singleton variants from the non-functional ones, especially in the absence of replication in independent cohorts and functional validation such as experimentation with cell or animal models. As a result, any heritability model that includes these types of variants will require special considerations and additional simulations to create realistic scenarios and avoid biases. The availability of larger, more diverse multi-omics databases, combined with advancements in methods for high-throughput functional assessments of mutations—such as massively parallel Perturb-seq for functional assessment of large numbers of mutations simultaneously[8,9]—may improve our understanding of singletons in the future.

Impact of genetic sex on heritability estimates, based on Study 2, were inconclusive due to the lack of power to observe any significant difference, this once more emphasizes the need to reassess the question of genetic sex on RV contributions to heritability with higher sample size. Furthermore, the sex chromosomes excluded from this study, may also contribute observed phenotypic variations between males and females. Unlike autosomes, the difference in dosage of X and Y chromosomes in males vs females can complicate the initial QC steps that evaluate HWE, as well as the statistical analysis, making it difficult to interpret genetic associations consistently across sexes[10]. Furthermore, the X-inactivation in females, and smaller Y chromosomes, additional population stratification issues due to differences in allele frequencies between males and females adds layers of complexities. For these reasons, genetic studies focus primarily on autosomes, where the inheritance patterns are simpler and more consistent across the population. However, to observe the sex-differences in heritability estimates, there is a need RARity, to include sex chromosomes in genetic analyses. This may be achieved via sex-stratified analyses, exclusion of HWE for sex chromosomes and further simulations.

Nonetheless, versatility of the method means that RARity can be adapted for common variants, as shown in study 2, and may be extended to CNVs and WGS, so long as the assumptions of the OLS are not violated, and appropriate calibrations are performed. However, it is important to note that, since most of the missing heritability estimates are captured by the rare variants included in this study, we anticipate a low contribution, if any, due to the excluded variants. Consequently, studying the contributions of the excluded variants to trait heritability will likely require a much larger sample size, and computational resources.

An important goal should be to produce a maximally enriched heritability, such that the top ranked genes for each trait explains most of the heritability. Future research should also aim to integrate genetics with other omics data to provide a comprehensive understanding of disease mechanisms and extend the study to a diverse population. Further investigations

are necessary to better understand the underlying biological mechanisms of the target genes identified with of RARity and RARity-β to further interrogate their validity as potential diagnostic or therapeutic targets. I also established that there is a need for a pathogenicity score that enrich for functional rare variants in complex traits. This may be achieved by a combination of variant-level and gene-level annotations, where the variant selection criteria are determined by a machine learning algorithm optimized for rare variant heritability, and thus identifying RVs that are functionally important to a given trait, ultimately improving our understanding of genetic predisposition to complex trait susceptibility.

To conclude, in this study, I have shown that RV heritability estimates with RARity and RARity-β provide valuable insights into the genetic basis of complex traits and diseases, offering several important advantages in genetic research. By focusing on genes that contribute substantially to heritability, researchers can streamline their efforts toward those genes most likely to play a critical role in the trait or disease of interest, making research more efficient. Additionally, incorporating high-priority genes into predictive models based on their heritability contributions can improve the accuracy of genetic risk predictions and variant-level pathogenicity scores, which are crucial for personalized medicine. This approach also optimizes resource use, directing experimental and computational efforts towards the most promising candidates, thereby facilitating more efficient functional studies that could lead to new therapeutic targets. Ultimately, gene prioritization based on RV heritability estimates will significantly enhance the efficiency, accuracy, and impact of genetic research. Overall, this thesis contributes significantly to the field of genetic research, offering novel methodologies, and insights that pave the way for future discoveries and applications in understanding complex traits and diseases.

6.6 REFERENCES

1 Pathan, N. *et al.* Genetic Determinants of Vascular Dementia. *Can J Cardiol* (2024). https://doi.org/10.1016/j.cjca.2024.03.025

2 Markus, H. S. & Schmidt, R. Genetics of Vascular Cognitive Impairment. *Stroke* **50**, 765-772 (2019). https://doi.org/10.1161/STROKEAHA.118.020379

3 Hoggart, C. *et al.* Identification of novel locus associated with coronary artery aneurysms and validation of loci for susceptibility to Kawasaki disease. *Eur J Hum Genet* **29**, 1734-1744 (2021). https://doi.org/10.1038/s41431-021-00838-5

4 Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet* **47**, 291-295 (2015). https://doi.org/10.1038/ng.3211

5 Di Scipio, M. *et al.* A versatile, fast and unbiased method for estimation of gene-by-environment interaction effects on biobank-scale datasets. *Nat Commun* **14**, 5196 (2023). https://doi.org/10.1038/s41467-023-40913-7

6 Bobbili, D. R., Banda, P., Krüger, R. & May, P. Excess of singleton loss-of-function variants in Parkinson's disease contributes to genetic risk. *J Med Genet* **57**, 617-623 (2020). https://doi.org/10.1136/jmedgenet-2019-106316

7 Genovese, G. *et al.* Increased burden of ultra-rare protein-altering variants among 4,877 individuals with schizophrenia. *Nat Neurosci* **19**, 1433-1441 (2016). https://doi.org/10.1038/nn.4402

8 Dixit, A. *et al.* Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. *Cell* **167**, 1853-1866.e1817 (2016). https://doi.org/10.1016/j.cell.2016.11.038

9 Ursu, O. *et al.* Massively parallel phenotyping of coding variants in cancer with Perturb-seq. *Nat Biotechnol* **40**, 896-905 (2022). https://doi.org/10.1038/s41587-021-01160-7

10 Sun, L., Wang, Z., Lu, T., Manolio, T. A. & Paterson, A. D. eXclusionarY: 10 years later, where are the sex chromosomes in GWASs? *Am J Hum Genet* **110**, 903-912 (2023). https://doi.org/10.1016/j.ajhg.2023.04.009

APPENDIX A:
SUPPLEMENTARY INFORMATION FOR STUDY 2 (CHAPTER 4)

Supplementary Information

A Method to Estimate the Contribution of Rare Coding Variants to Complex Trait Heritability

Nazia Pathan,[1,2] Wei Q. Deng,[3,4] Matteo Di Scipio,[1,5] Mohammad Khan,[1,5] Shihong Mao,[1] Robert W. Morton,[1,2] Ricky Lali,[1,6] Marie Pigeyre,[1,5] Michael R. Chong,[1,2,7] Guillaume Paré[1,2,6,7*]

*Corresponding author

[1]Population Health Research Institute, David Braley Cardiac, Vascular and Stroke Research Institute, Hamilton Health Sciences and McMaster University, Hamilton, Canada.
[2]Department of Pathology and Molecular Medicine, McMaster University, Michael G. DeGroote School of Medicine, Hamilton, Canada.
[3]Peter Boris Centre for Addictions Research, St. Joseph's Healthcare Hamilton, Hamilton, Canada.
[4]Department of Psychiatry and Behavioural Neurosciences, McMaster University, Hamilton, Canada.
[5]Department of Medicine, Faculty of Health Sciences, McMaster University, Hamilton, Canada.
[6]Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, Canada.
[7]Thrombosis and Atherosclerosis Research Institute, David Braley Cardiac, Vascular and Stroke Research Institute, Hamilton, Canada.

Supplemental Figures: 11

Supplemental Tables: 7

Corresponding Author:

Guillaume Paré MD, MSc, FRCPC

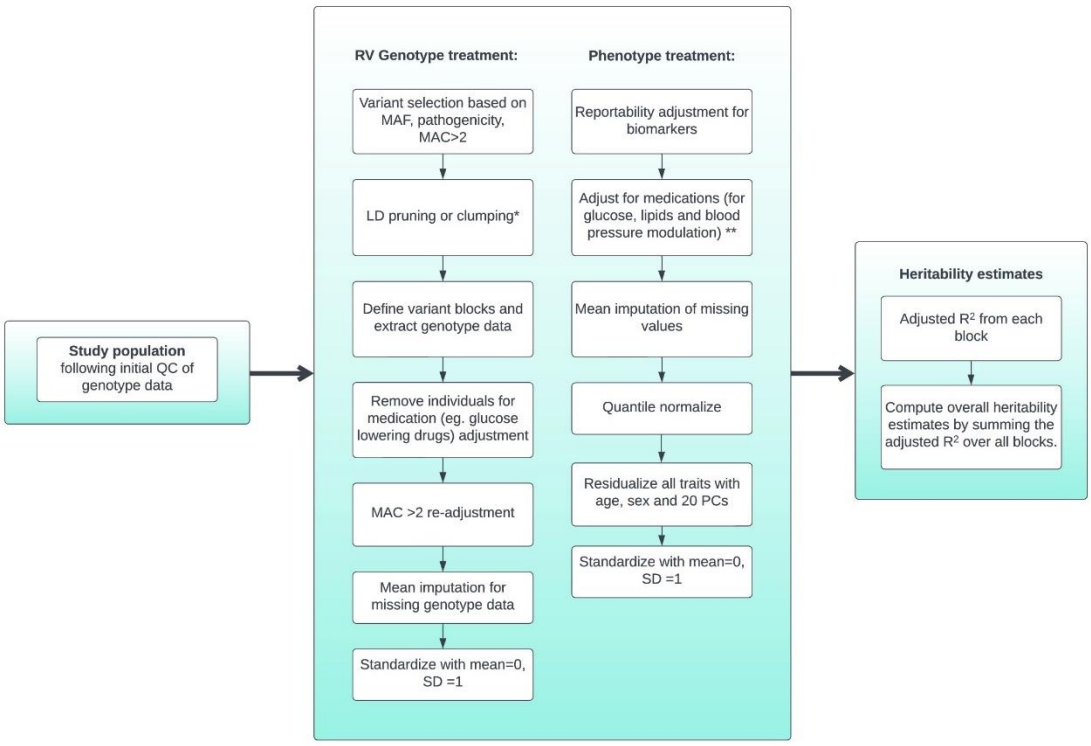
McMaster University Population Health Research Institute

David Braley Cardiac, Vascular, and Stroke Research Institute

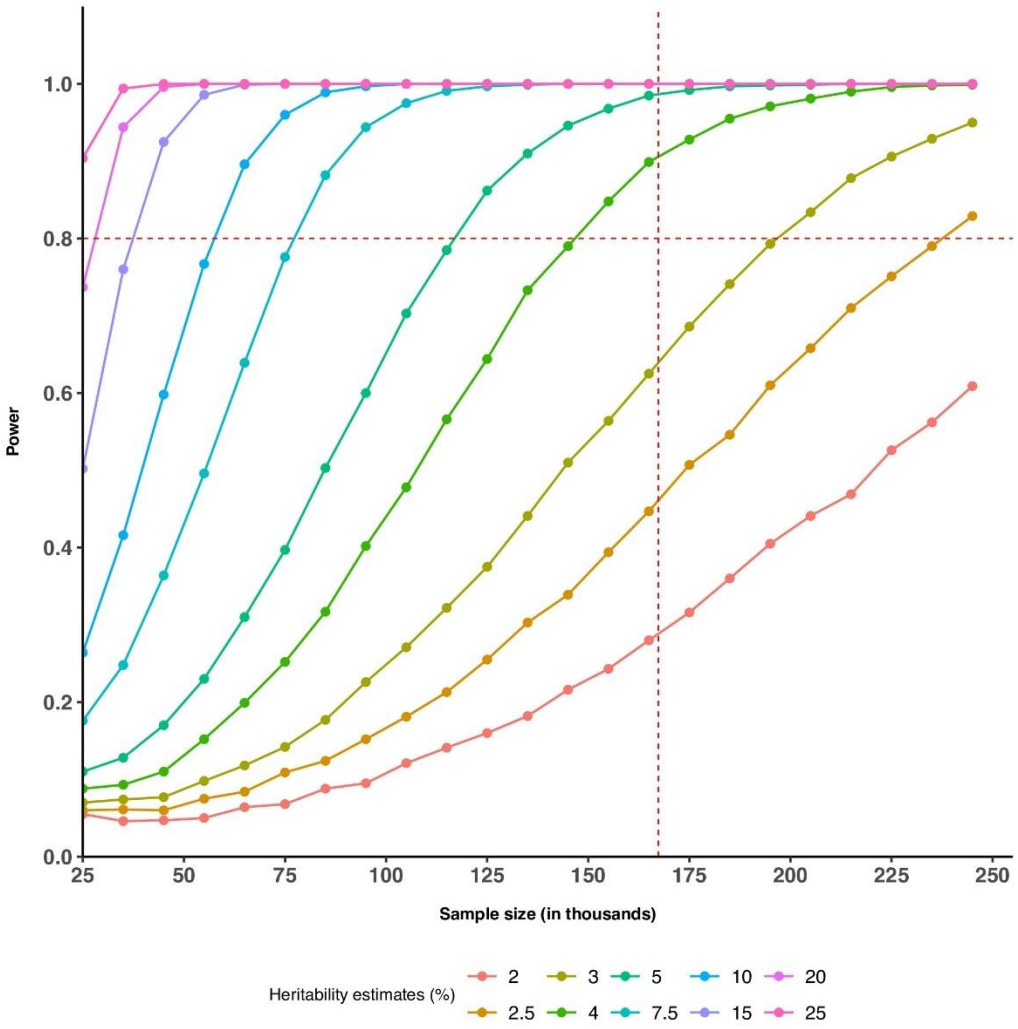
237 Barton Street East – C4 126

E-mail: pareg@mcmaster.ca

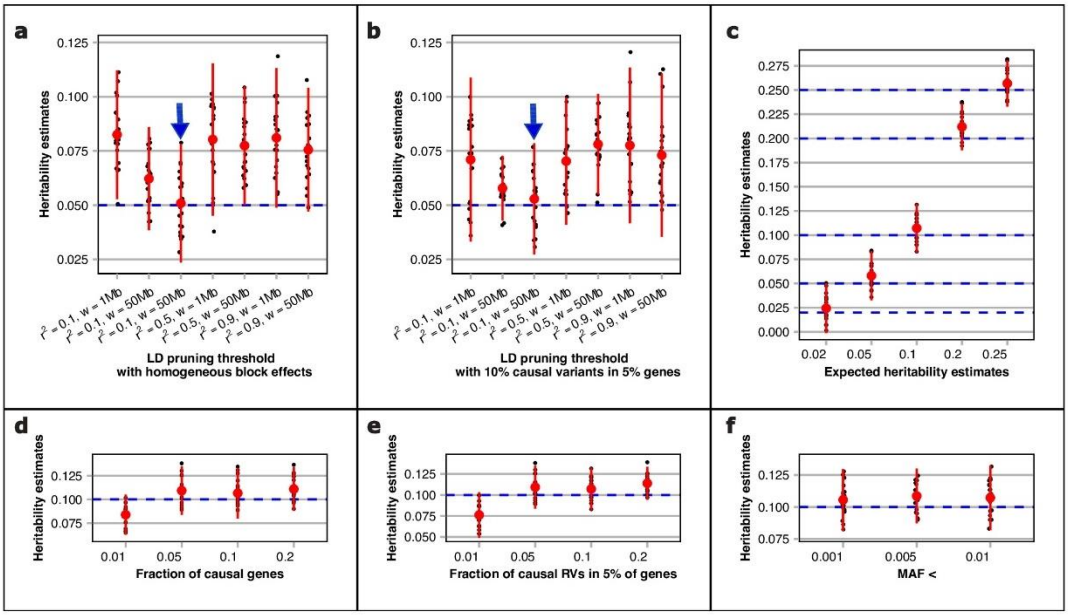
Contents



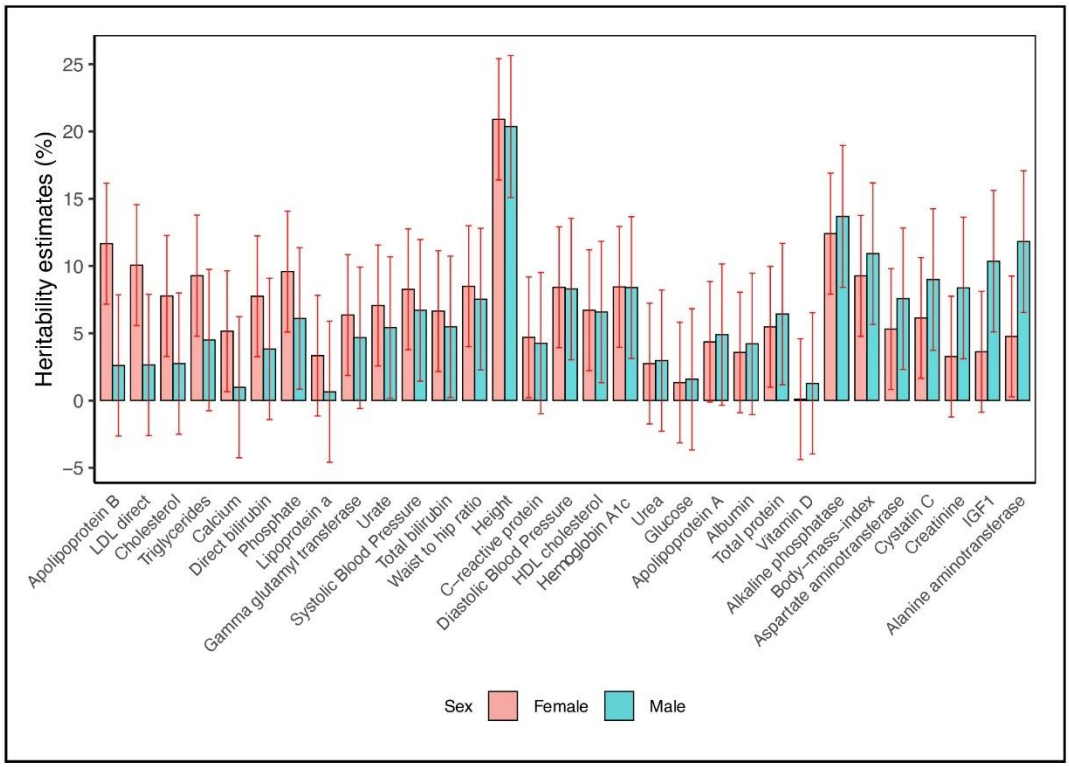
Supplementary Fig.1: Summary of the Rare variant heritability (RARity) estimator pipeline. The RARity pipeline constitutes pre-treatment of the genotype and the phenotype data, followed by application of the statistical model to each block, and finally estimation of the total heritability from all blocks. *The model may be modified to prioritize variants by implementing LD clumping instead of pruning. In addition, the pruning parameters are dependent on the selection of common *vs* rare variants for the analysis. **Adjustment of medications may involve implementing correction factors or removal of individuals using the medications. MAF = minor allele frequency, MAC = minor allele count, LD = linkage disequilibrium, SD = standard deviation, PCs = principal components.



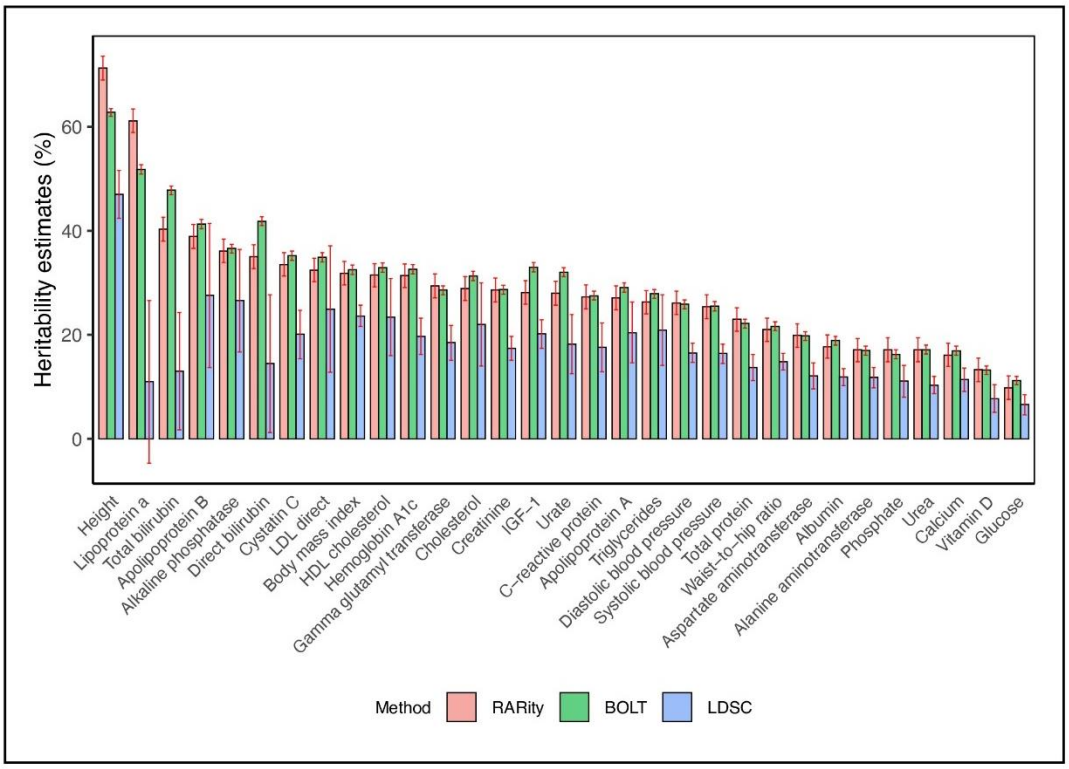
Supplementary Fig.2: Statistical power of RARity. The statistical power of rare coding variant heritability with RARity is displayed as a function of sample size for different levels of RV heritability estimates. Statistical power was estimated empirically from the variance of 10,000 simulated h^2_{RV}, under 230 conditions of sample sizes and true set h^2_{RV}, with each condition being represented by a dot. The red, horizontal dashed line corresponds to an empirical 80% power with alpha-level of 0.05. The red, vertical dashed line marks the sample size used in the current study.



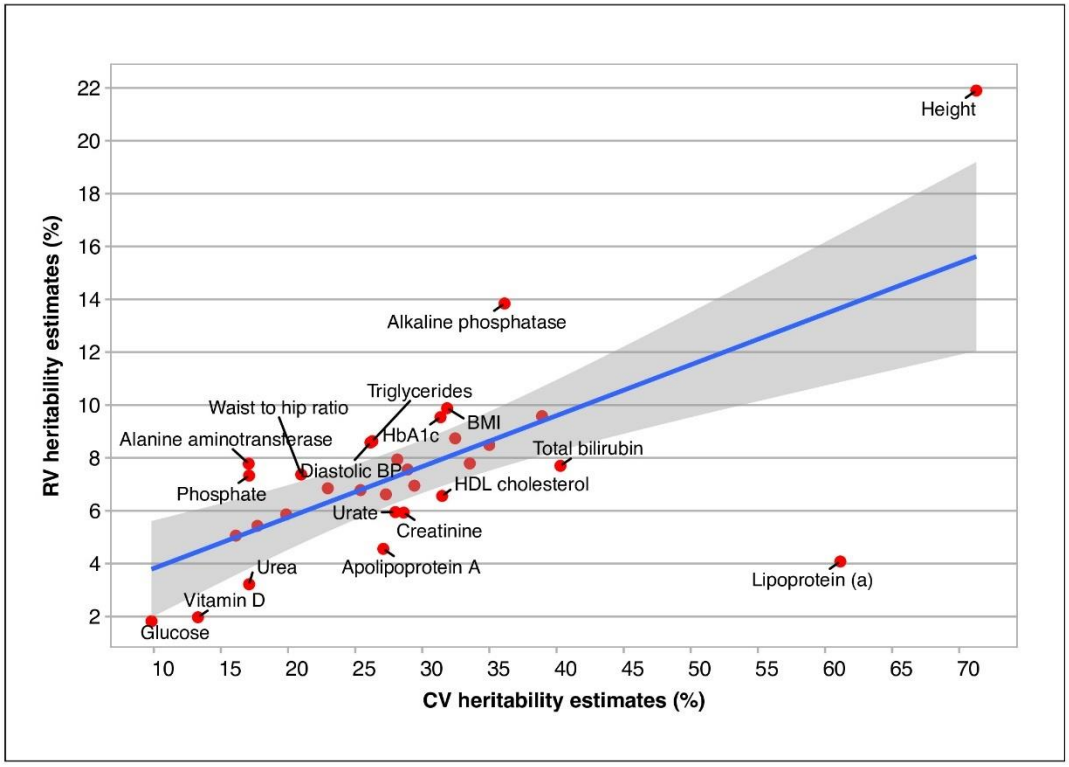
Supplementary Fig.3: Simulation of rare variant heritability estimates under varying LD pruning, fraction of causal genes and variants, and MAF conditions. Calibration of RARity for linkage disequilibrium, where r^2 = 'coefficient of correlation' and w = 'window size' thresholds used for LD pruning, (a) assuming homogenous distribution of heritability across all exome-wide blocks, and (b) assuming 10% of causal variants in 5% genes. In both cases, unbiased estimation of h^2_{RV} was observed when RVs were pruned with $r^2 > 0.1$ within a window size of 50Mb (blue arrows) and is the default LD pruning threshold used for all analyses. Figures (c-f) examined the sensitivity of RARity to varying (c) true h^2_{RV} values; (d) fraction of causal genes with 10% RVs with effects; (e) fraction of causal RVs in 5% of the genes and (f) MAF thresholds assuming 10% causal RVs within 5% of genes has an effect. Each dot in the figure represents a single simulation, with 20 exome-wide simulations performed for each scenario. The red vertical lines represent the 95% CIs, and the blue dashed, horizontal lines represent the true h^2_{RV}. Except for (f), RVs with MAF < 0.01 were selected for all analyses.



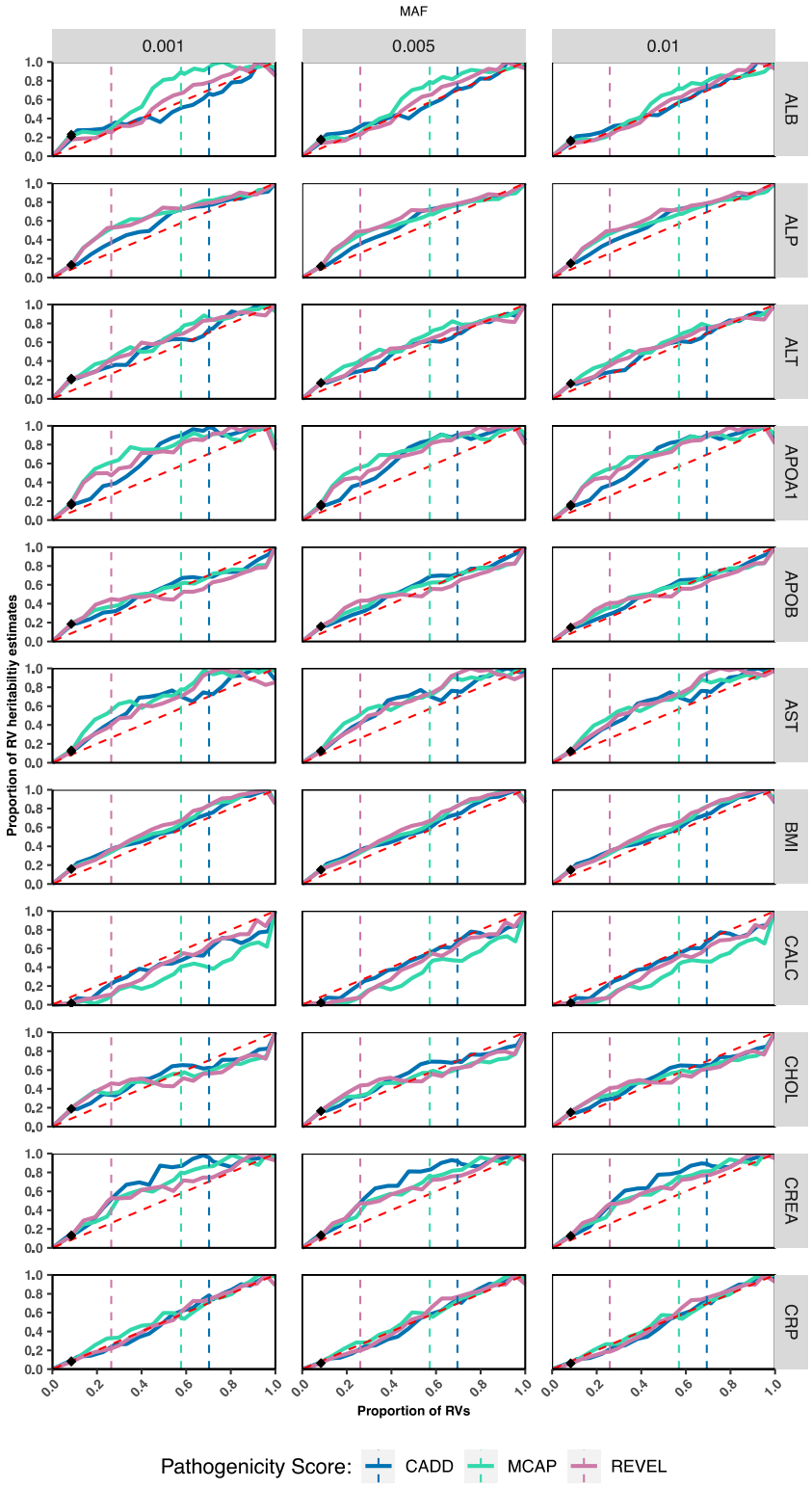
Supplementary Fig.4: Comparison of RV exome-wide heritability estimates between the sexes for 31 continuous traits. Estimation of heritability based on RVs with MAF <0.01 in 92,963 females (light red) and 74,385 males (blue). Red error bar represents 95% confidence intervals of the estimated heritability. Differences in heritability estimates between the sexes appear heterogenous but remains statistically non-significant (p-value > 0.05).



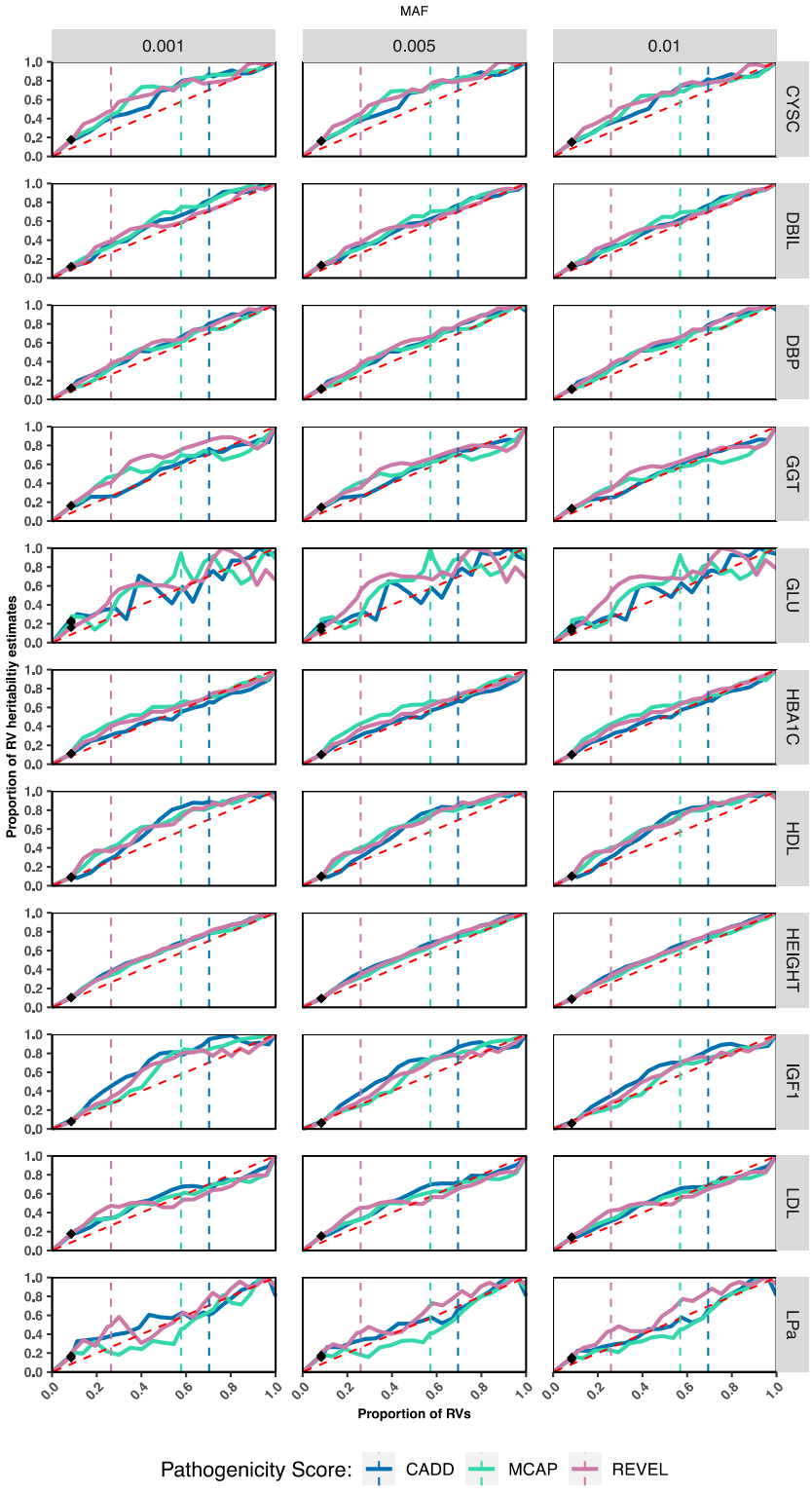
Supplementary Fig.5: Comparison of methods to estimate common variant (MAF>0.01) heritability estimates. The methods under comparison are RARity (light red), BOLT (green), and LDSC (blue) in 31 complex traits, that were adjusted for age, sex and the first 20 PCs. See online Methods for a description of each method. The red error bar represents 95% confidence intervals of the heritability estimates.

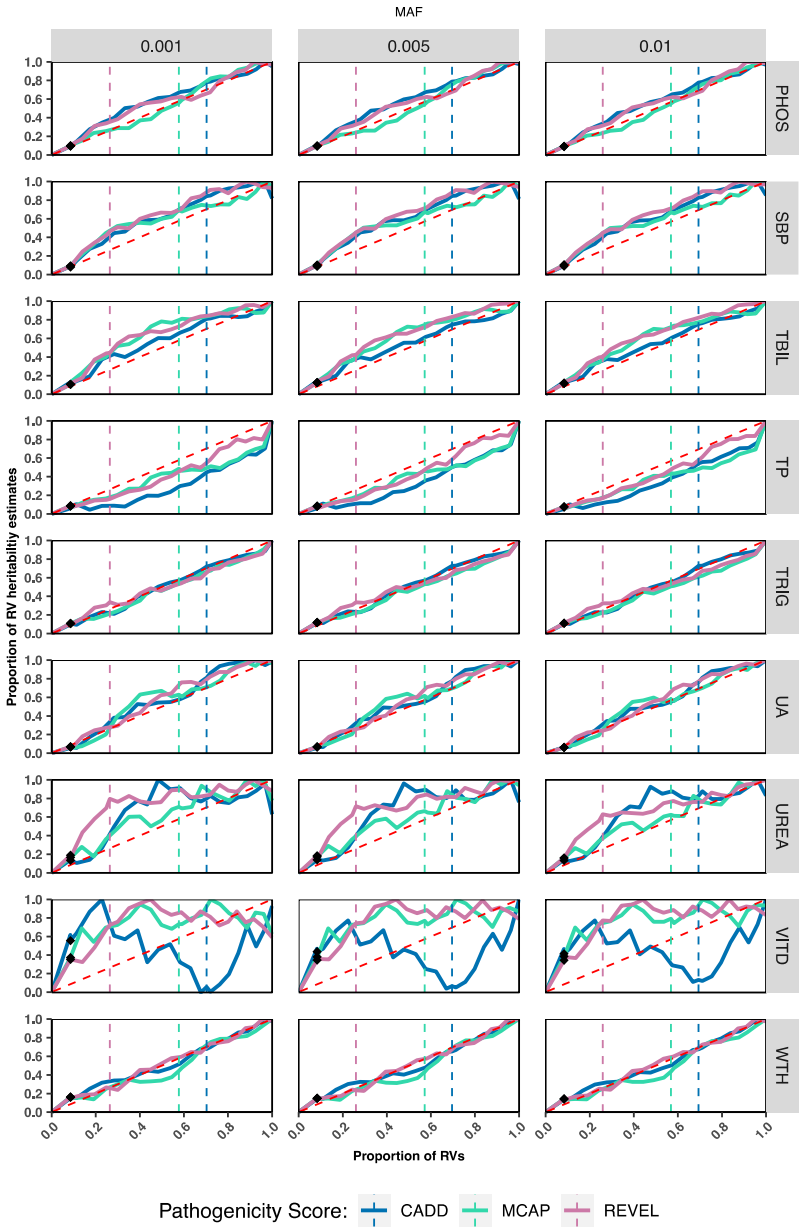


Supplementary Fig.6: Correlation between the contribution of RV and CV to complex traits heritability. Both CV and RV contributions to heritability was estimated with RARity, however the LD pruning threshold was varied (Methods) depending on the variant type. Exome-wide blocks with 5,000 RVs/ block were used to estimate h^2_{RV}, while genome-wide blocks with 20,000 CVs/block were used to estimate h^2_{CV}. Uncertainty in the relationship is expressed as 95% CI, denoted with the grey, shaded area.

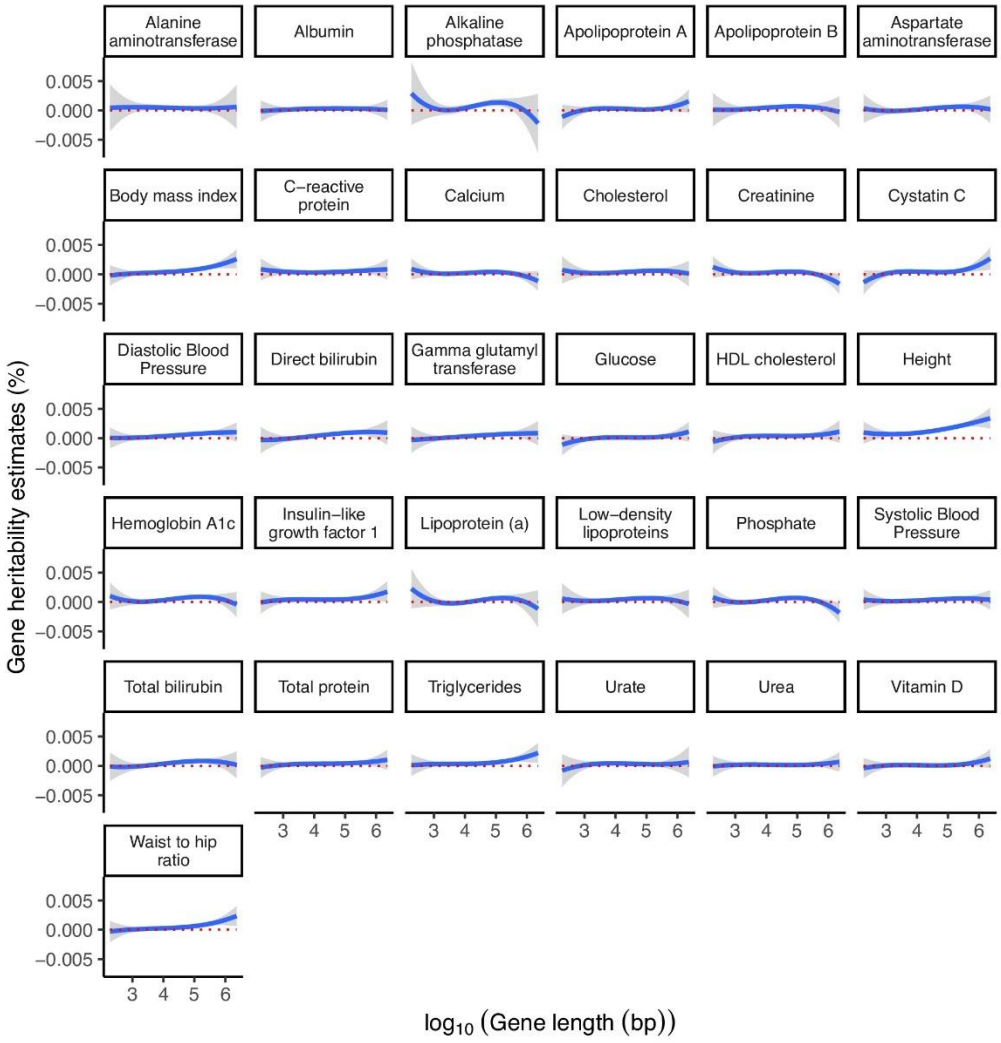


Supplementary

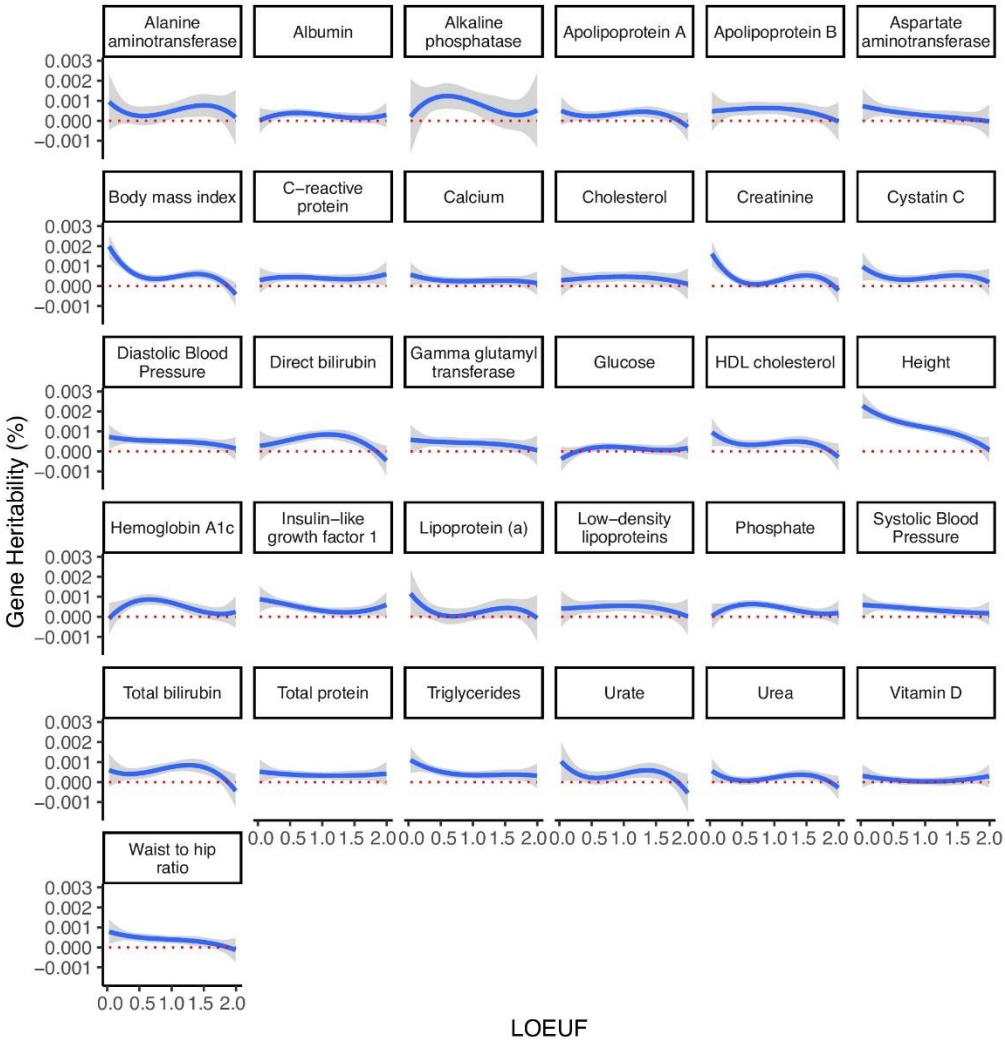

MAF
0.001
0.005
0.01
CYSC
DBIL
DBP
GGT
GLU
HBA1C
HDL
HEIGHT
IGF1
LDL
LPa
Proportion of RV heritability estimates
Proportion of RVs
Pathogenicity Score:
CADD
MCAP
REVEL



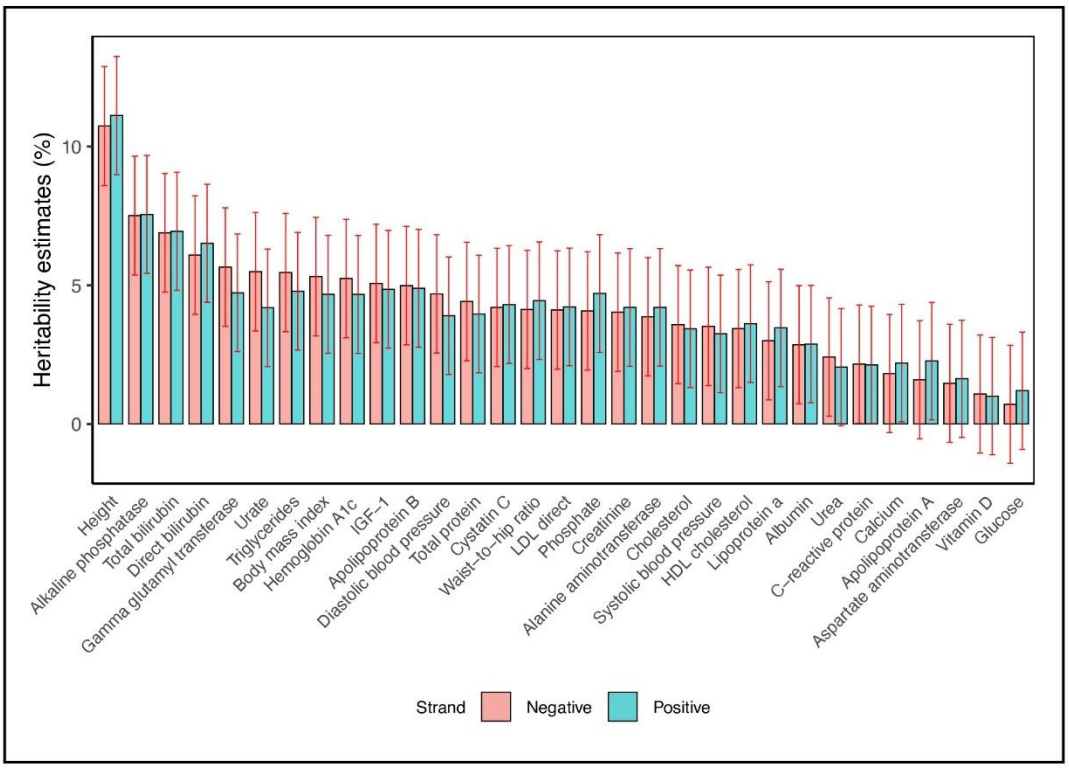
Supplementary Fig.7: Impact of pathogenicity scores on variance explained by RVs for 31 complex traits. Illustration of the proportion of RV heritability explained (*y* axis) as a function of incorporating increasingly "deleterious" genetic variants (*x* axis). Proportion of heritability estimates is the fraction of estimates in relation to all protein altering and LoF variants within the MAF categories. The vertical, dashed lines represent the binary thresholds recommended to define pathogenicity for CADD (dark blue), M-CAP (green) and REVEL (magenta). The black diamond marks the point of last inclusion of LoF variants, which were prioritized before missense mutations. The diagonal dashed red lines represent the scenario wherein RVs uniformly contribute to h^2_{RV}, irrespective of pathogenicity score. Description of the abbreviated traits are available in Supplementary Table 1.



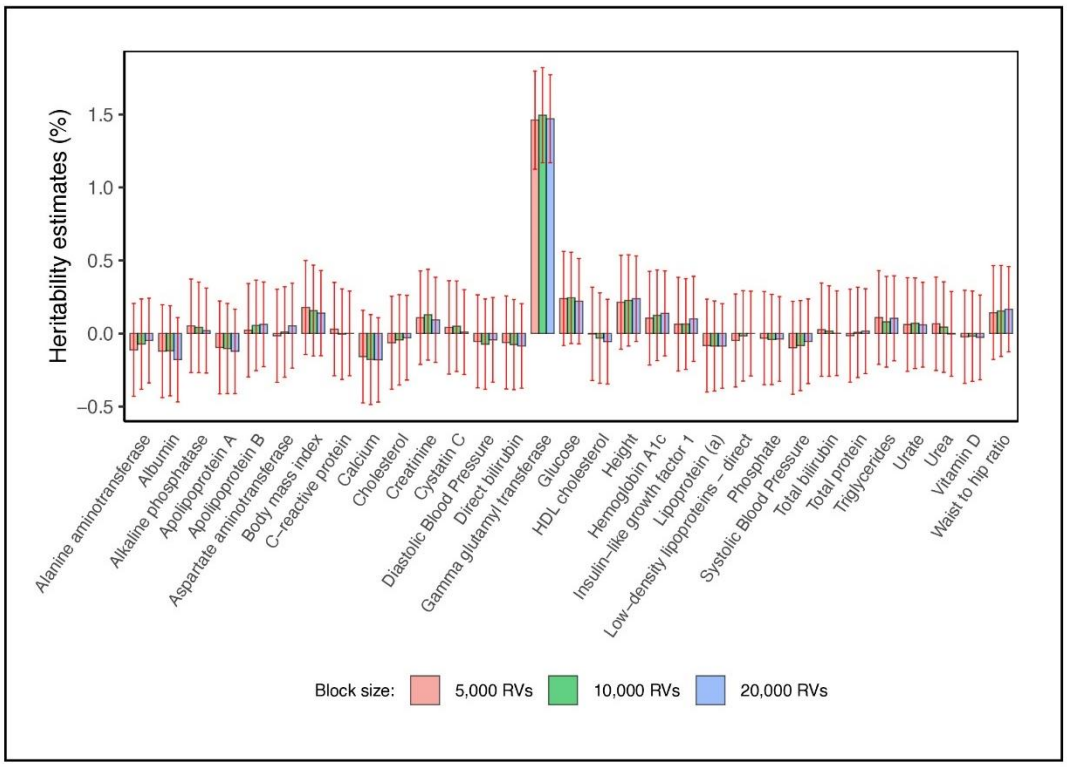
Supplementary Fig.8: Spline plots for the associations of $\log_{10}$ (Gene length (bp)) with $h^2_{\text{gene-RV}}$. Transcripts with the largest length were used as a measure of gene-length. A base spline graph with three degrees of freedom has been used to show the relationship between gene length and $h^2_{\text{gene-RV}}$. Red-dotted lines represent the expected $h^2_{\text{gene-RV}} = 0$, under the null-hypothesis. Uncertainty in the relationship is expressed by a 95% CI (grey band).



Supplementary Fig.9: Spline plots for the associations of evolutionary constraint (LOEUF) with $h^2_{\text{gene-RV}}$. Relationship between the gene evolutionary constraints and $h^2_{\text{gene-RV}}$ is illustrated using a base spline with three degrees of freedom. Low LOEUF scores indicate strong selection against predicted loss-of-function (pLoF) variation in a given gene, while high LOEUF scores suggest a relatively higher tolerance to inactivation. Red-dotted lines represent the expected $h^2_{\text{gene-RV}} = 0$, under the null-hypothesis. Uncertainty in the relationship is expressed by a 95% CI (grey band).



Supplementary Fig.10: Comparison of RV heritability estimates between the genes encoded in the positive vs negative strands. Estimation of trait heritability ($h^2_{\text{RV-gene-tot}}$) +/- 95% CI, was based on RVs with MAF < 0.01, belonging to genes on either the positive strand (blue) or the negative strands (light red). Gene-wise block construct was utilized for the estimation of each $h^2_{\text{RV-gene}}$, followed by the estimation of the $h^2_{\text{RV-gene-tot}}$ as described in methods. Bright-red error bars denote 95% CI of the $h^2_{\text{RV-gene-tot}}$.



Supplementary Fig.11: Impact of block size on RV heritability estimates. Estimation of heritability (h^2_{RV}) +/- 95% CI, based on protein altering and LoF RVs with MAF < 0.01, using 20,000 variants from chromosome 22, partitioned into blocks of either 5,000 (light red), 10,000 (green) or 20,000 (blue) consecutive RVs. Bright-red error bars denote 95% confidence intervals of the h^2_{RV}.

Supplementary Table 1: Abbreviations used to describe the phenotypes. The list of phenotypes in this study includes 26 biomarkers and 5 anthropomorphic traits in 167,348 unrelated Caucasian participants from the UKB.

Index	TRAITS	ABBREVIATIONS
1	Albumin	ALB
2	Alkaline phosphatase	ALP
3	Alanine aminotransferase	ALT
4	Apolipoprotein A	APOA1
5	Apolipoprotein B	APOB
6	Aspartate aminotransferase	AST
7	Calcium	CALC
8	Cholesterol	CHOL
9	Creatinine	CREA
10	C-reactive protein	CRP
11	Cystatin C	CYSC
12	Direct bilirubin	DBIL
13	Gamma glutamyl transferase	GGT
14	Glucose	GLU
15	Hemoglobin A1c	HBA1C
16	HDL cholesterol	HDL
17	Insulin-like growth factor 1	IGF1
18	Low-density lipoproteins - direct	LDL
19	Lipoprotein (a)	LPa
20	Phosphate	PHOS
21	Total bilirubin	TBIL
22	Total protein	TP
23	Triglycerides	TRIG
24	Urea	UREA
25	Urate	UA
26	Vitamin D	VITD
27	Height	HIGHT
28	Waist to hip ratio	WTH
29	Systolic Blood Pressure	SBP
30	Diastolic Blood Pressure	DBP
31	Body mass index	BMI

Supplementary Table 2: Medications used for adjusting biomarkers values. Field ids and the medication code used for glucose lowering, cholesterol lowering and hypertension medications. Adjustment for these medications is further described in methods.

Glucose lowering medication		**Cholesterol lowering medications**		**Hypertension medications**	
Field 20003 code	**Description**	**Field 20003 code**	**Description**	**Field 6177 code**	**Description**
1140868902	acarbose	1141146234	atorvastatin	2	Blood pressure medication
1140857584	acetohexamide	1141192414	crestor 10mg tablet		
1141171652	actos 15mg tablet	1141192736	ezetimibe	**Field 6153 code**	**Description**
1140868866	bromocriptine	1140888594	fluvastatin	2	Blood pressure medication
1140874706	chlorpropamide	1141146138	lipitor 10mg tablet		
1140874746	diamicron 80mg tablet	1140888648	pravastatin		
1140866568	disopyramide	1141192410	rosuvastatin		
1141157186	disopyramide product	1140861958	simvastatin		
1140857518	eudemine 50mg tablet	1140881748	zocor 10mg tablet		
1140874718	glibenclamide	1141200040	zocor heart-pro 10mg tablet		
1140857494	glibornuride	1140864592	lescol 20mg capsule		
1140874744	gliclazide				
1141152590	glimepiride	**Field 6177 code**	**Description**		
1140874646	glipizide	1	Cholesterol lowering medications		
1141157284	glipizide product	**Field 6153 code**	**Description**		
1140874658	gliquidone	1	Cholesterol lowering medications		
1140874686	glucophage 500mg tablet				
1140857500	glymidine				
1140874754	guar gum				
1140884600	metformin				
1140869112	mifepristone				
1141157302	mifepristone product				
1140874652	minodiab 2.5mg tablet=glipizide				
1141173882	nateglinide				
1140884338	pentamidine				
1141171646	pioglitazone				
1140874420	quinine				
1141168660	repaglinide				
1141177600	rosiglitazone				
1141189090	rosiglitazone 1mg / metformin 500mg tablet				
1141173786	starlix 60mg tablet=nateglinide				
1141182110	sulfadiazine				
1140874664	tolazamide				
1140874674	tolbutamide				
1141153254	troglitazone				
Field 6177 code	**Description**				
3	Insulin				
Field 6153 code	**Description**				
3	Insulin				

Supplementary Table 3: Comparison of RV heritability estimates derived using gene-burden, gene-wise and exome-wide blocks. $h^2_{RV\text{-}burden}$ = RV heritability estimates derived using gene-burden block construct, $h^2_{RV\text{-}gene\text{-}tot}$ = RV heritability estimates derived using gene block construct, h^2_{RV} = RV heritability estimates derived using exome-wide block construct with ~5000 consecutive variants in each block. LCL= lower confidence level, UCL=upper confidence level. Comparison shows percentage difference in trait heritability between the methods.

	Gene Burden			Gene-wise			Exome-wide			Comparison of RV heritability		
Trait	$h^2_{RV\text{-}burden}$	LCL	UCL	$h^2_{RV\text{-}gene\text{-}tot}$	LCL	UCL	h^2_{RV}	LCL	UCL	Exome-wise vs Gene burden	Gene-wise vs Gene burden	Gene-wise vs Exome-wise
ALB	0.0099	0.0070	0.0128	0.0554	0.0258	0.0849	0.0543	0.0257	0.0829	82%	82%	2%
ALP	0.0438	0.0406	0.0469	0.1426	0.1129	0.1723	0.1384	0.1097	0.1671	68%	69%	3%
ALT	0.0161	0.0132	0.0191	0.0795	0.0499	0.1091	0.0778	0.0492	0.1064	79%	80%	2%
APOA1	0.0129	0.0100	0.0158	0.0473	0.0177	0.0768	0.0456	0.0170	0.0742	72%	73%	4%
APOB	0.0151	0.0122	0.0180	0.0970	0.0674	0.1266	0.0958	0.0672	0.1245	84%	84%	1%
AST	0.0101	0.0072	0.0130	0.0557	0.0261	0.0853	0.0586	0.0301	0.0872	83%	82%	-5%
CALC	0.0065	0.0036	0.0093	0.0485	0.0189	0.0781	0.0506	0.0221	0.0792	87%	87%	-4%
CHOL	0.0137	0.0108	0.0166	0.0740	0.0444	0.1036	0.0756	0.0470	0.1042	82%	81%	-2%
CREA	0.0130	0.0101	0.0159	0.0612	0.0317	0.0908	0.0593	0.0308	0.0879	78%	79%	3%
CRP	0.0142	0.0113	0.0171	0.0714	0.0418	0.1009	0.0662	0.0376	0.0948	79%	80%	7%
CYSC	0.0144	0.0115	0.0173	0.0778	0.0482	0.1074	0.0779	0.0493	0.1065	81%	81%	0%
DBIL	0.0100	0.0071	0.0129	0.0992	0.0695	0.1288	0.0849	0.0563	0.1135	88%	90%	14%
GGT	0.0100	0.0071	0.0129	0.0751	0.0455	0.1047	0.0695	0.0409	0.0981	86%	87%	7%
GLU	0.0056	0.0027	0.0084	0.0192	-0.0103	0.0487	0.0182	-0.0103	0.0467	69%	71%	5%
HBA1C	0.0104	0.0075	0.0133	0.0960	0.0663	0.1256	0.0954	0.0667	0.1240	89%	89%	1%
HDL	0.0140	0.0111	0.0169	0.0675	0.0379	0.0971	0.0656	0.0370	0.0942	79%	79%	3%
IGF1	0.0116	0.0087	0.0145	0.0802	0.0506	0.1098	0.0794	0.0508	0.1081	85%	86%	1%
LDL	0.0138	0.0109	0.0167	0.0865	0.0569	0.1161	0.0874	0.0588	0.1160	84%	84%	-1%
LPa	0.0128	0.0099	0.0157	0.0427	0.0132	0.0723	0.0408	0.0123	0.0694	69%	70%	4%
PHOS	0.0107	0.0078	0.0136	0.0762	0.0466	0.1058	0.0733	0.0446	0.1019	85%	86%	4%
TBIL	0.0103	0.0074	0.0131	0.0930	0.0633	0.1226	0.0770	0.0484	0.1056	87%	89%	17%
TP	0.0093	0.0064	0.0122	0.0690	0.0394	0.0986	0.0685	0.0399	0.0971	86%	86%	1%
TRIG	0.0108	0.0079	0.0137	0.0839	0.0543	0.1135	0.0863	0.0577	0.1149	88%	87%	-3%
UREA	0.0044	0.0015	0.0072	0.0343	0.0048	0.0639	0.0322	0.0037	0.0608	87%	87%	6%
UA	0.0191	0.0162	0.0221	0.0634	0.0338	0.0930	0.0595	0.0309	0.0881	68%	70%	6%
VITD	0.0044	0.0016	0.0073	0.0198	-0.0097	0.0493	0.0197	-0.0088	0.0482	77%	78%	0%
SBP	0.0060	0.0031	0.0089	0.0662	0.0366	0.0958	0.0678	0.0392	0.0964	91%	91%	-2%
DBP	0.0073	0.0044	0.0101	0.0872	0.0576	0.1168	0.0858	0.0572	0.1144	92%	92%	2%
BMI	0.0076	0.0047	0.0104	0.0990	0.0693	0.1286	0.0988	0.0702	0.1275	92%	92%	0%
WTH	0.0050	0.0022	0.0079	0.0736	0.0440	0.1032	0.0737	0.0451	0.1023	93%	93%	0%
HEIGHT	0.0248	0.0218	0.0278	0.2231	0.1932	0.2529	0.2190	0.1902	0.2478	89%	89%	2%

Supplementary Table 4: RV heritability by MAF bins calculated using gene-wise blocks. N variants= number of variants present in each MAF bin. LCL= lower confidence level, UCL=upper confidence level.

MAF	0.01 to 0.005				0.005 to 0.001				0.01 to 0.001				< 0.001			
Trait	N variants	$h^2_{RV\text{-}gene\text{-}tot}$	LCL	UCL	N variants	$h^2_{RV\text{-}gene\text{-}tot}$	LCL	UCL	N variants	$h^2_{RV\text{-}gene\text{-}tot}$	LCL	UCL	N variants	$h^2_{RV\text{-}gene\text{-}tot}$	LCL	UCL
ALB	6509	0.0063	0.0043	0.0084	23714	0.0093	0.0056	0.0131	29552	0.0146	0.0104	0.0188	1514538	0.0355	0.0066	0.0643
ALP	6509	0.0096	0.0075	0.0117	23714	0.0247	0.0208	0.0286	29552	0.0332	0.0288	0.0376	1514538	0.1110	0.0820	0.1400
ALT	6509	0.0040	0.0021	0.0060	23714	0.0136	0.0099	0.0174	29552	0.0177	0.0135	0.0220	1514538	0.0539	0.0250	0.0827
APOA1	6509	0.0049	0.0029	0.0069	23714	0.0143	0.0105	0.0181	29552	0.0181	0.0138	0.0223	1514538	0.0327	0.0038	0.0615
APOB	6509	0.0139	0.0117	0.0161	23714	0.0180	0.0142	0.0218	29552	0.0299	0.0256	0.0343	1514538	0.0710	0.0421	0.0999
AST	6509	0.0060	0.0040	0.0080	23714	0.0093	0.0055	0.0130	29552	0.0146	0.0104	0.0188	1514538	0.0398	0.0109	0.0686
CALC	6509	0.0040	0.0020	0.0060	23714	0.0077	0.0040	0.0114	29552	0.0120	0.0079	0.0162	1514538	0.0367	0.0079	0.0656
CHOL	6509	0.0101	0.0080	0.0122	23714	0.0146	0.0108	0.0184	29552	0.0232	0.0189	0.0275	1514538	0.0541	0.0253	0.0830
CREA	6509	0.0067	0.0047	0.0088	23714	0.0136	0.0098	0.0173	29552	0.0195	0.0152	0.0237	1514538	0.0450	0.0162	0.0739
CRP	6509	0.0062	0.0041	0.0082	23714	0.0134	0.0096	0.0171	29552	0.0188	0.0145	0.0230	1514538	0.0542	0.0253	0.0831
CYSC	6509	0.0064	0.0043	0.0084	23714	0.0128	0.0091	0.0166	29552	0.0177	0.0135	0.0219	1514538	0.0620	0.0331	0.0909
DBIL	6509	0.0084	0.0063	0.0105	23714	0.0094	0.0056	0.0131	29552	0.0171	0.0129	0.0213	1514538	0.0766	0.0476	0.1055
GGT	6509	0.0093	0.0072	0.0114	23714	0.0124	0.0086	0.0161	29552	0.0207	0.0165	0.0250	1514538	0.0491	0.0203	0.0780
GLU	6509	0.0032	0.0013	0.0052	23714	0.0046	0.0010	0.0083	29552	0.0079	0.0038	0.0120	1514538	0.0138	-0.0151	0.0426
HBA1C	6509	0.0099	0.0078	0.0121	23714	0.0164	0.0126	0.0202	29552	0.0258	0.0215	0.0301	1514538	0.0772	0.0483	0.1061
HDL	6509	0.0048	0.0028	0.0068	23714	0.0146	0.0108	0.0184	29552	0.0185	0.0143	0.0227	1514538	0.0515	0.0226	0.0804
IGF1	6509	0.0071	0.0050	0.0091	23714	0.0161	0.0123	0.0199	29552	0.0222	0.0179	0.0264	1514538	0.0542	0.0253	0.0831
LDL	6509	0.0113	0.0091	0.0134	23714	0.0151	0.0113	0.0189	29552	0.0245	0.0202	0.0288	1514538	0.0644	0.0355	0.0933
LPa	6509	0.0066	0.0045	0.0086	23714	0.0177	0.0138	0.0215	29552	0.0230	0.0187	0.0272	1514538	0.0235	-0.0053	0.0524
PHOS	6509	0.0078	0.0057	0.0099	23714	0.0113	0.0076	0.0151	29552	0.0180	0.0138	0.0223	1514538	0.0618	0.0329	0.0907
TBIL	6509	0.0095	0.0074	0.0116	23714	0.0112	0.0074	0.0149	29552	0.0202	0.0159	0.0244	1514538	0.0686	0.0397	0.0975
TP	6509	0.0091	0.0070	0.0112	23714	0.0099	0.0061	0.0136	29552	0.0179	0.0137	0.0221	1514538	0.0503	0.0214	0.0791
TRIG	6509	0.0073	0.0052	0.0094	23714	0.0116	0.0078	0.0153	29552	0.0182	0.0140	0.0224	1514538	0.0629	0.0340	0.0918
UREA	6509	0.0045	0.0025	0.0065	23714	0.0042	0.0005	0.0078	29552	0.0086	0.0045	0.0128	1514538	0.0197	-0.0091	0.0486
UA	6509	0.0059	0.0039	0.0079	23714	0.0099	0.0061	0.0136	29552	0.0151	0.0110	0.0193	1514538	0.0490	0.0202	0.0779
VITD	6509	0.0018	-0.0002	0.0037	23714	0.0066	0.0029	0.0103	29552	0.0083	0.0042	0.0125	1514538	0.0120	-0.0168	0.0409
SBP	6509	0.0053	0.0033	0.0073	23714	0.0089	0.0052	0.0126	29552	0.0135	0.0093	0.0177	1514538	0.0549	0.0260	0.0838
DBP	6509	0.0060	0.0040	0.0080	23714	0.0082	0.0045	0.0119	29552	0.0136	0.0094	0.0178	1514538	0.0783	0.0494	0.1072
BMI	6509	0.0054	0.0034	0.0074	23714	0.0089	0.0052	0.0127	29552	0.0141	0.0099	0.0183	1514538	0.0869	0.0580	0.1158
WTH	6509	0.0051	0.0030	0.0071	23714	0.0073	0.0036	0.0110	29552	0.0120	0.0079	0.0162	1514538	0.0647	0.0358	0.0936
HEIGHT	6509	0.0171	0.0148	0.0193	23714	0.0301	0.0262	0.0341	29552	0.0460	0.0415	0.0506	1514538	0.1726	0.1435	0.2016

Supplementary Table 5: Rare coding variant heritability as a function of $\log_{10}$(Gene-length). *P-values* were calculated using a multivariable linear regression model, with $\log_{10}$ (gene-length) as the independent predictor and RV gene-heritability estimates as the outcome variable, adjusted for sex and the first 20 principal components of ancestry.

Trait	Estimate	Standard error	R^2	*P-value*
HEIGHT	7.68×10^{-2}	1.16×10^{-2}	2.41×10^{-3}	3.34×10^{-11}
BMI	5.40×10^{-2}	1.16×10^{-2}	1.19×10^{-3}	3.17×10^{-6}
DBIL	4.72×10^{-2}	1.16×10^{-2}	9.11×10^{-4}	4.62×10^{-5}
WTH	4.55×10^{-2}	1.16×10^{-2}	8.47×10^{-4}	8.58×10^{-5}
DBP	4.20×10^{-2}	1.16×10^{-2}	7.21×10^{-4}	2.91×10^{-4}
TBIL	3.65×10^{-2}	1.16×10^{-2}	5.45×10^{-4}	1.63×10^{-3}
HBA1C	3.58×10^{-2}	1.16×10^{-2}	5.24×10^{-4}	2.00×10^{-3}
GGT	3.19×10^{-2}	1.16×10^{-2}	4.17×10^{-4}	5.87×10^{-3}
TRIG	3.10×10^{-2}	1.16×10^{-2}	3.92×10^{-4}	7.54×10^{-3}
PHOS	3.09×10^{-2}	1.16×10^{-2}	3.92×10^{-4}	7.56×10^{-3}
AST	2.97×10^{-2}	1.16×10^{-2}	3.61×10^{-4}	1.04×10^{-2}
SBP	2.65×10^{-2}	1.16×10^{-2}	2.87×10^{-4}	2.23×10^{-2}
Lpa	2.38×10^{-2}	1.16×10^{-2}	2.31×10^{-4}	4.01×10^{-2}
ALP	2.14×10^{-2}	1.16×10^{-2}	1.88×10^{-4}	6.46×10^{-2}
CHOL	1.82×10^{-2}	1.16×10^{-2}	1.36×10^{-4}	1.16×10^{-1}
LDL	1.68×10^{-2}	1.16×10^{-2}	1.15×10^{-4}	1.48×10^{-1}
GLU	1.45×10^{-2}	1.16×10^{-2}	8.65×10^{-5}	2.10×10^{-1}
TP	1.44×10^{-2}	1.16×10^{-2}	8.44×10^{-5}	2.15×10^{-1}
APOB	1.42×10^{-2}	1.16×10^{-2}	8.29×10^{-5}	2.19×10^{-1}
CYSC	1.33×10^{-2}	1.16×10^{-2}	7.24×10^{-5}	2.51×10^{-1}
IGF1	1.19×10^{-2}	1.16×10^{-2}	5.84×10^{-5}	3.03×10^{-1}
CRP	1.08×10^{-2}	1.16×10^{-2}	4.75×10^{-5}	3.52×10^{-1}
HDL	9.66×10^{-3}	1.16×10^{-2}	3.82×10^{-5}	4.04×10^{-1}
CALC	8.28×10^{-3}	1.16×10^{-2}	2.80×10^{-5}	4.75×10^{-1}
ALB	7.65×10^{-3}	1.16×10^{-2}	2.39×10^{-5}	5.09×10^{-1}
VITD	6.48×10^{-3}	1.16×10^{-2}	1.72×10^{-5}	5.76×10^{-1}
TEST	-6.41×10^{-3}	1.16×10^{-2}	1.68×10^{-5}	5.80×10^{-1}
ALT	-4.93×10^{-3}	1.16×10^{-2}	9.94×10^{-6}	6.70×10^{-1}
UA	-1.55×10^{-3}	1.16×10^{-2}	9.86×10^{-7}	8.93×10^{-1}
CREA	1.54×10^{-3}	1.16×10^{-2}	9.73×10^{-7}	8.94×10^{-1}
APOA1	9.02×10^{-4}	1.16×10^{-2}	3.33×10^{-7}	9.38×10^{-1}
UREA	-3.36×10^{-4}	1.16×10^{-2}	4.62×10^{-8}	9.77×10^{-1}

Supplementary Table 6: Rare coding variant heritability as a function of evolutionary constraint. *P-values* were calculated using a multivariable linear regression model, with gene evolutionary constraint (LOEUF score) as the independent predictor and RV gene-heritability estimates as the outcome variable, adjusted for sex and the first 20 principal components of ancestry.

Trait	Estimate	Standard error	R^2	*P-value*
HEIGHT	-8.77×10^{-6}	1.30×10^{-6}	2.68×10^{-3}	1.71×10^{-11}
BMI	-4.29×10^{-6}	1.20×10^{-6}	7.52×10^{-4}	3.66×10^{-4}
WTH	-3.24×10^{-6}	1.25×10^{-6}	4.00×10^{-4}	9.37×10^{-3}
IGF1	-2.71×10^{-6}	1.30×10^{-6}	2.56×10^{-4}	3.77×10^{-2}
AST	-3.48×10^{-6}	1.73×10^{-6}	2.40×10^{-4}	4.40×10^{-2}
TRIG	-2.46×10^{-6}	1.28×10^{-6}	2.20×10^{-4}	5.41×10^{-2}
SBP	-2.29×10^{-6}	1.22×10^{-6}	2.10×10^{-4}	5.96×10^{-2}
HBA1C	-2.83×10^{-6}	1.59×10^{-6}	1.86×10^{-4}	7.60×10^{-2}
DBP	-2.01×10^{-6}	1.21×10^{-6}	1.64×10^{-4}	9.60×10^{-2}
PHOS	-1.98×10^{-6}	1.26×10^{-6}	1.46×10^{-4}	1.16×10^{-1}
CREA	-1.61×10^{-6}	1.26×10^{-6}	9.75×10^{-5}	2.00×10^{-1}
HDL	-1.78×10^{-6}	1.40×10^{-6}	9.54×10^{-5}	2.04×10^{-1}
GGT	-1.92×10^{-6}	1.52×10^{-6}	9.47×10^{-5}	2.06×10^{-1}
ALP	-4.53×10^{-6}	3.80×10^{-6}	8.42×10^{-5}	2.33×10^{-1}
APOB	-1.87×10^{-6}	1.99×10^{-6}	5.28×10^{-5}	3.45×10^{-1}
CALC	-9.30×10^{-7}	1.23×10^{-6}	3.39×10^{-5}	4.49×10^{-1}
DBIL	-1.13×10^{-6}	1.54×10^{-6}	3.17×10^{-5}	4.64×10^{-1}
ALB	-9.06×10^{-7}	1.25×10^{-6}	3.10×10^{-5}	4.69×10^{-1}
LDL	-1.09×10^{-6}	1.82×10^{-6}	2.11×10^{-5}	5.51×10^{-1}
ALT	1.67×10^{-6}	2.83×10^{-6}	2.05×10^{-5}	5.56×10^{-1}
UA	-1.10×10^{-6}	1.95×10^{-6}	1.88×10^{-5}	5.73×10^{-1}
GLU	5.13×10^{-7}	1.22×10^{-6}	1.05×10^{-5}	6.74×10^{-1}
TP	-5.10×10^{-7}	1.23×10^{-6}	1.01×10^{-5}	6.79×10^{-1}
APOA1	-5.57×10^{-7}	1.44×10^{-6}	8.88×10^{-6}	6.99×10^{-1}
CYSC	-5.08×10^{-7}	1.41×10^{-6}	7.66×10^{-6}	7.19×10^{-1}
CHOL	-5.09×10^{-7}	1.61×10^{-6}	5.94×10^{-6}	7.51×10^{-1}
VITD	-1.81×10^{-7}	1.18×10^{-6}	1.39×10^{-6}	8.78×10^{-1}
LPa	-3.53×10^{-7}	2.39×10^{-6}	1.30×10^{-6}	8.82×10^{-1}
TBIL	-1.60×10^{-7}	1.70×10^{-6}	5.24×10^{-7}	9.25×10^{-1}
UREA	1.07×10^{-7}	1.22×10^{-6}	4.58×10^{-7}	9.30×10^{-1}
CRP	-2.04×10^{-8}	1.29×10^{-6}	1.49×10^{-8}	9.87×10^{-1}

Supplementary Table 7: Heritability of height originating from RVs in selected gene clusters. LCL= lower confidence level, UCL=upper confidence level.

Cluster	Number of genes	Number of variants	RV heritability	LCL	UCL
Hemoglobin	10	159	0.000160	-0.000159	0.000479
Histone	77	2099	0.000415	-0.000675	0.00150
HOX	226	14735	0.00118	-0.001679	0.00404
Olfactory	366	18366	0.00315	-0.0000697	0.00637
Protocadherin	59	7034	-0.00124	-0.00317	0.000690

LCL= lower confidence limit, UCL=upper confidence limit

Supplementary Data 1: Heritability estimates from twin and pedigree studies

In this table, selected reports of pedigree/twin-based trait heritability are presented for comparison with RARity based heritability estimates from combined rare and common variants (h^2_{tot}). Rows highlighted in green indicate the recovery of pedigree/twin-based heritability estimates through RARity, as these estimates lie within the range historically reported in the referenced literature.

Trait	Reference	Cohort	Sample size	Heritability (%)	RARity h^2_{tot} (%) (95% CI)
Alanine aminotransferase	van Beek, J.H. *et al.* The genetic architecture of liver enzyme levels: GGT, ALT and AST. *Behav Genet* **43**, 329-39 (2013).	NTR biobankstudy	Twins, their siblings, parents and spouses (N = 8,371; age range 18–90).	22 (Females) - 40 (Males)	23.0 (19.7,26.2)
Albumin	Loomba, R. *et al.* Heritability of Hepatic Fibrosis and Steatosis Based on a Prospective Twin Study. *Gastroenterology* **149**, 1784-93 (2015).	Southern California	60 pairs of twins	45	
Albumin	Whitfield, J.B. & Martin, N.G. The effects of inheritance on constituents of plasma: a twin study on some biochemical variables. *Ann Clin Biochem* **21 (Pt 3)**, 176-83 (1984).	The Australian NH & MRC Twin Registry for a study of alcohol metabolism Twin Registry for a study of alcohol metabolism and susceptibility to intoxication	206 pairs of twins	73	22.5 (19.3,25.7)
Alkaline phosphatase	Whitfield, J.B. & Martin, N.G. Determinants of variation in plasma alkaline phosphatase activity: a twin study. *Am J Hum Genet* **35**, 978-86 (1983).	Australian NH&MRC Twin Registry	204 pairs of twins	71-90	
	Loomba, R. *et al.* Heritability of Hepatic Fibrosis and Steatosis Based on a Prospective Twin Study. *Gastroenterology* **149**, 1784-93 (2015).	Southern California	60 pairs of twins	50	46.8 (43.6,50.1)

Full table available at https://doi.org/10.1038/s41467-024-45407-8

Supplementary Data 2: Genes contributing significantly to the rare coding variant heritability estimates of complex traits.

RARity was used to determine the heritability estimates for all genes with qualifying RVs <MAF 0.01. This table shows the significant genes passing the Bonferroni's *p-value* significance threshold corrected for 18,214 genes ($h^2_{\text{RV-gene}}$ *p*-value < 2.75×10^{-6}), where the $h^2_{\text{RV-gene}}$ *p*-values were derived using *F*-test. 152 genes had significant $h^2_{RV\text{-}gene}$ for one or more traits, representing 218 distinct gene-biomarker relationships.

Phenotype	Gene	$h^2_{RV\text{-}gene}$	P-value
ALB	*ALB*	1.96E-03	7.94E-47
ALB	*FCGRT*	2.60E-03	7.68E-75
ALB	*TBC1D2B*	6.20E-04	2.05E-07
ALP	*ALDH5A1*	5.65E-04	7.94E-09
ALP	*ALPL*	2.83E-02	0
ALP	*ASGR1*	2.95E-03	3.44E-82
ALP	*GBGT1*	7.41E-04	1.97E-12
ALP	*GPLD1*	1.23E-02	0
ALP	*HSPG2*	3.94E-03	1.26E-39
ALP	*NBPF3*	5.57E-04	1.12E-07
ALP	*TDP2*	4.71E-04	9.73E-09
ALP	*TFAM*	3.17E-04	1.86E-06
ALP	*ZNF800*	4.34E-04	1.88E-07
ALT	*GPT*	2.22E-02	0
ALT	*MFSD3*	4.12E-04	2.43E-06
APOA1	*ABCA1*	5.72E-03	2.19E-124
APOA1	*ADH1B*	4.68E-04	1.65E-07
APOA1	*ANGPTL3*	9.01E-04	1.96E-17
APOA1	*APOA1*	1.69E-03	4.06E-46
APOA1	*APOC3*	4.15E-04	1.98E-10
APOA1	*CETP*	8.91E-04	1.03E-15
APOA1	*LCAT*	1.21E-03	5.51E-26
APOA1	*LIPC*	1.81E-03	6.06E-35
APOA1	*LIPG*	1.76E-03	1.47E-39
APOA1	*LPL*	8.04E-04	6.18E-15
APOA1	*PLA2G12A*	7.46E-04	5.81E-17
APOA1	*SCARB1*	6.86E-04	3.52E-10
APOB	*APOB*	1.22E-02	3.70E-250
APOB	*APOE*	6.34E-04	1.22E-11
APOB	*BCAM*	7.75E-04	2.21E-10
APOB	*CEACAM20*	4.99E-04	4.41E-08
APOB	*CLASRP*	6.63E-04	1.19E-08
APOB	*LDLR*	3.75E-03	1.03E-88
APOB	*NECTIN2*	1.48E-03	1.15E-28

Full table available at https://doi.org/10.1038/s41467-024-45407-8

Supplementary Data 3: Association between the genes contributing significantly to RV heritability and diseases based on DisGeNET platform.

The genes with significant contributions to trait heritability, as identified through RARity, were examined for disease associations using the disgenet2r package from the DisGeNET platform. Among the 152 genes with significant $h^2_{\text{RV-gene}}$, 115 genes are linked to multiple diseases, as presented in this table. The data is specifically filtered to include only literature-supported connections (GDA Score >= 0.3).

disease_type	gene_pli	el	disease_class_name	ei	e	gene_symbol	source	geneid	protein_class	year_final	gene_dpi	diseaseid	uniprotid	disease_semantic_type	disease_name	disease_class	score	year_initial	gene_dsi
disease	1.45E-13		Congenital, Hereditary, and	0.991	Transporter	*ABCA1*	CURATED	19	DTO_05007405	2019	0.846	C0039292	O95477	Disease or Syndrome	Tangier Disease	C16;C18;C10	1	1999	0.467
disease	1.45E-13	limited	Congenital, Hereditary, and	0.972	Transporter	*ABCA1*	CURATED	19	DTO_05007405	2017	0.846	C1704429	O95477	Disease or Syndrome	Hypoalphalipopr	C16;C18	0.8	1999	0.467
disease	1.45E-13	limited	Nutritional and Metabolic D	1	Transporter	*ABCA1*	CURATED	19	DTO_05007405	2017	0.846	C0020443	O95477	Disease or Syndrome	Hypercholestero	C18	0.55	2002	0.467
disease	1.45E-13		Digestive System Diseases; I	1	Transporter	*ABCA1*	CURATED	19	DTO_05007405	2018	0.846	C0009402	O95477	Neoplastic Process	Colorectal Carcir	C06;C04	0.52	2012	0.467
disease	1.45E-13	limited	Congenital, Hereditary, and	1	Transporter	*ABCA1*	CURATED	19	DTO_05007405	2019	0.846	C0745103	O95477	Disease or Syndrome	Hyperlipoproteir	C16;C18	0.52	2005	0.467
phenotype	1.45E-13		Congenital, Hereditary, and	1	Transporter	*ABCA1*	CURATED	19	DTO_05007405	2012	0.846	C0342898	O95477	Pathologic Function	Apolipoprotein /	C16;C18	0.3	2004	0.467
phenotype	1.45E-13		Cardiovascular Diseases	1	Transporter	*ABCA1*	CURATED	19	DTO_05007405	2011	0.846	C1563937	O95477	Pathologic Function	Atherogenesis	C14	0.3	2011	0.467
disease	1.45E-13		Cardiovascular Diseases	0.992	Transporter	*ABCA1*	CURATED	19	DTO_05007405	2020	0.846	C0004153	O95477	Disease or Syndrome	Atherosclerosis	C14	0.4	2000	0.467
group	1.45E-13		Digestive System Diseases; I	1	Transporter	*ABCA1*	CURATED	19	DTO_05007405	2012	0.846	C0009404	O95477	Neoplastic Process	Colorectal Neopl	C06;C04	0.3	2012	0.467
disease	1.45E-13		Congenital, Hereditary, and	1	Transporter	*ABCA1*	CURATED	19	DTO_05007405	2017	0.846	C2931838	O95477	Disease or Syndrome	Familial HDL defi	C16;C18	0.4	1999	0.467
disease	1.45E-13		Congenital, Hereditary, and	1	Transporter	*ABCA1*	CURATED	19	DTO_05007405	2001	0.846	C3711531	O95477	Disease or Syndrome	Hdl Deficiency, T	C16;C18	0.3	1999	0.467
disease	1.45E-13		Congenital, Hereditary, and	1	Transporter	*ABCA1*	CURATED	19	DTO_05007405	2019	0.846	C0020445	O95477	Disease or Syndrome	Hypercholestero	C16;C18	0.33	2003	0.467
disease	1.45E-13		Congenital, Hereditary, and	1	Transporter	*ABCA1*	CURATED	19	DTO_05007405	2005	0.846	C1704417	O95477	Disease or Syndrome	Hyperlipoproteir	C16;C18	0.3	2005	0.467
disease	1.45E-13		Congenital, Hereditary, and	1	Transporter	*ABCA1*	CURATED	19	DTO_05007405	2020	0.846	C0473527	O95477	Disease or Syndrome	Hypoalphalipopr	C16;C18	0.4	2002	0.467
disease	1.45E-13		Female Urogenital Diseases a	1	Transporter	*ABCA1*	CURATED	19	DTO_05007405	2010	0.846	C0022661	O95477	Disease or Syndrome	Kidney Failure, C	C13;C12	0.3	2010	0.467
group	1.45E-13		Nervous System Diseases	1	Transporter	*ABCA1*	CURATED	19	DTO_05007405	2011	0.846	C4721453	O95477	Disease or Syndrome	Peripheral Nervo	C10	0.3	2011	0.467

Full table available at https://doi.org/10.1038/s41467-024-45407-8

Supplementary Data 4: List of druggable and clinically actionable genes retrieved from Drug gene interaction database (DGIdb)

RARIty was utilized to identify the genes contributing significantly towards trait heritability, and subsequently the DGIdb (v4.2.0) database was employed to categorize the genes as "clinically actionable" or part of the "druggable genome." The table presents the genes that fall into either the clinically actionable or druggable categories.

search_term	match_term	match_type	category	source databses
APOB	APOB	Definite	CLINICALLY ACTIONABLE	Tempus
ARID1B	ARID1B	Definite	CLINICALLY ACTIONABLE	MskImpact|Tempus
CASR	CASR	Definite	CLINICALLY ACTIONABLE	Tempus
DDR2	DDR2	Definite	CLINICALLY ACTIONABLE	MskImpact|FoundationOneGenes|CarisMolecularIntelligence|Oncomine|Tempus
EZR	EZR	Definite	CLINICALLY ACTIONABLE	FoundationOneGenes|CarisMolecularIntelligence
FCGR2B	FCGR2B	Definite	CLINICALLY ACTIONABLE	FoundationOneGenes
FGF2	FGF2	Definite	CLINICALLY ACTIONABLE	Tempus
FGFR3	FGFR3	Definite	CLINICALLY ACTIONABLE	MskImpact|FoundationOneGenes|CarisMolecularIntelligence|Oncomine|Tempus|CIViC
HLA-DPA1	HLA-DPA1	Definite	CLINICALLY ACTIONABLE	Tempus
JAK1	JAK1	Definite	CLINICALLY ACTIONABLE	MskImpact|FoundationOneGenes|CarisMolecularIntelligence|Oncomine|Tempus
JAK2	JAK2	Definite	CLINICALLY ACTIONABLE	MskImpact|FoundationOneGenes|CarisMolecularIntelligence|Oncomine|Tempus
LDLR	LDLR	Definite	CLINICALLY ACTIONABLE	Tempus
SH2B3	SH2B3	Definite	CLINICALLY ACTIONABLE	CarisMolecularIntelligence|Tempus
SOD2	SOD2	Definite	CLINICALLY ACTIONABLE	Tempus
UGT1A1	UGT1A1	Definite	CLINICALLY ACTIONABLE	Tempus|CIViC
UGT1A9	UGT1A9	Definite	CLINICALLY ACTIONABLE	Tempus
ABCA1	ABCA1	Definite	DRUGGABLE GENOME	RussLampel|HingoraniCasas
ABCG5	ABCG5	Definite	DRUGGABLE GENOME	RussLampel
ADGRE5	ADGRE5	Definite	DRUGGABLE GENOME	HopkinsGroom|RussLampel|HingoraniCasas
ADH1B	ADH1B	Definite	DRUGGABLE GENOME	RussLampel|HingoraniCasas|HopkinsGroom
ALB	ALB	Definite	DRUGGABLE GENOME	HingoraniCasas
ALDH5A1	ALDH5A1	Definite	DRUGGABLE GENOME	HopkinsGroom|HingoraniCasas|RussLampel
ALPL	ALPL	Definite	DRUGGABLE GENOME	HingoraniCasas
ANGPTL3	ANGPTL3	Definite	DRUGGABLE GENOME	RussLampel|HingoraniCasas|HopkinsGroom

Full table available at https://doi.org/10.1038/s41467-024-45407-8

Supplementary Data 5: Functional annotation of the genes contributing significantly to trait heritability, performed using g:Profiler.

The g:ProfileR web tool for functional profiling, g:GOST, was used to test the enrichment of the genes with significant heritability ($h^2_{RV\text{-}gene}$ p-value < 2.75×10^{-6}), against gene-sets in common databases. The significant heritability genes for each phenotype were treated as separate gene lists for independent query, and statistical tests were conducted within a domain scope of only annotated genes (Methods). This table shows a list of statistically significant enriched terms for each gene list, corresponding to the phenotypes, adjusted for multiple testing using g:SCS (set counts and sizes) p-value < 0.05.

Phenotype	Term ID	Term name	Adjusted P-value	FDR	Intersection genes
Alanine aminotransferase	KEGG:01210	2-Oxocarboxylic acid metabolism	0.036491318	0.0364913	*GPT*
Alanine aminotransferase	KEGG:00220	Arginine biosynthesis	0.042253105	0.0422531	*GPT*
Alanine aminotransferase	REAC:R-HSA-8964540	Alanine metabolism	0.012490351	0.0124904	*GPT*
Alkaline phosphatase	GO:0042578	phosphoric ester hydrolase activity	0.033053819	0.0330538	*GPLD1,ALPL,TDP2*
Alkaline phosphatase	GO:0004621	glycosylphosphatidylinositol phospholipase D activity	0.049921872	0.0499219	*GPLD1*
Alkaline phosphatase	GO:0004777	succinate-semialdehyde dehydrogenase (NAD+) activity	0.049921872	0.0499219	*ALDH5A1*
Alkaline phosphatase	GO:0009013	succinate-semialdehyde dehydrogenase [NAD(P)+] activity	0.049921872	0.0499219	*ALDH5A1*
Alkaline phosphatase	GO:0036317	tyrosyl-RNA phosphodiesterase activity	0.049921872	0.0499219	*TDP2*
Alkaline phosphatase	GO:0050187	phosphoamidase activity	0.049921872	0.0499219	*ALPL*
Alkaline phosphatase	GO:0070260	5'-tyrosyl-DNA phosphodiesterase activity	0.049921872	0.0499219	*TDP2*
Apolipoprotein A-I	GO:0034185	apolipoprotein binding	3.96E-11	3.96E-11	*ABCA1,LIPC,LCAT,LPL,SCARB1*
Apolipoprotein A-I	GO:0034186	apolipoprotein A-I binding	5.50E-07	5.50E-07	*ABCA1,LCAT,SCARB1*
Apolipoprotein A-I	GO:0052689	carboxylic ester hydrolase activity	2.32683E-06	2.327E-06	*LIPG,LIPC,LCAT,PLA2G12A,LPL*
Apolipoprotein A-I	GO:0016298	lipase activity	2.63821E-06	2.638E-06	*LIPG,LIPC,LCAT,PLA2G12A,LPL*
Apolipoprotein A-I	GO:0015485	cholesterol binding	7.61813E-06	7.618E-06	*ABCA1,APOA1,CETP,APOC3*
Apolipoprotein A-I	GO:0004465	lipoprotein lipase activity	9.06086E-06	9.061E-06	*LIPG,LIPC,LPL*
Apolipoprotein A-I	GO:0032934	sterol binding	1.56396E-05	1.564E-05	*ABCA1,APOA1,CETP,APOC3*
Apolipoprotein A-I	GO:0008970	phospholipase A1 activity	3.06921E-05	3.069E-05	*LIPG,LIPC,LPL*
Apolipoprotein A-I	GO:0043178	alcohol binding	3.57829E-05	3.578E-05	*ABCA1,APOA1,CETP,APOC3*
Apolipoprotein A-I	GO:0005496	steroid binding	5.62843E-05	5.628E-05	*ABCA1,APOA1,CETP,APOC3*

Full table available at https://doi.org/10.1038/s41467-024-45407-8

APPENDIX B:
SUPPLEMENTARY INFORMATION FOR STUDY 3 (CHAPTER 5)

Supplementary Table 1: Significant genes contributing to trait heritability

Trait	**Chr**	**Gene**	$\hat{h}^2_{RV\text{-}liab\text{-}gene}$	***P-value***
Asthma	17	*EIF4A1*	8.750E-04	7.361E-08
Atrial fibrillation and flutter	4	*MAD2L1*	1.283E-03	6.316E-09
Atrial fibrillation and flutter	12	*PEBP1*	1.197E-03	2.442E-08
Atrial fibrillation and flutter	17	*YPEL2*	6.060E-04	1.356E-07
Cataract	6	*SAMD5*	7.920E-04	2.689E-08
Cholelithiasis and cholecystitis	1	*XCL2*	1.450E-03	1.403E-10
Cholelithiasis and cholecystitis	12	*FRS2*	2.108E-03	5.223E-10
Cholelithiasis and cholecystitis	17	*INPP5K*	2.492E-03	6.139E-10
Cholelithiasis and cholecystitis	3	*SLC38A3*	1.645E-03	4.403E-09
Cholelithiasis and cholecystitis	4	*LETM1*	2.403E-03	1.409E-08
Cholelithiasis and cholecystitis	14	*KTN1*	2.730E-03	6.006E-08
Cholelithiasis and cholecystitis	15	*EDC3*	1.553E-03	7.797E-08
Cholelithiasis and cholecystitis	19	*JAK3*	2.245E-03	8.899E-08
Cholelithiasis and cholecystitis	2	*UGT1A5*	2.141E-03	1.016E-07
Cholelithiasis and cholecystitis	1	*C1QB*	1.440E-03	1.123E-07
Cholelithiasis and cholecystitis	1	*NVL*	2.070E-03	1.316E-07
Coronary atherosclerosis	19	*LOC101927572*	9.190E-04	3.949E-08
Coronary atherosclerosis	8	*NKX3-1*	1.280E-03	1.470E-07
Gastrointestinal hemorrhage	9	*IFNA7*	1.331E-03	5.032E-08
Gastrointestinal hemorrhage	6	*TMEM151B*	1.550E-03	1.266E-07
Hemorrhoids	6	*ME1*	3.799E-03	2.288E-18
Hemorrhoids	3	*ZNF619*	2.376E-03	2.911E-10
Hemorrhoids	7	*RUNDC3B*	1.886E-03	3.360E-09

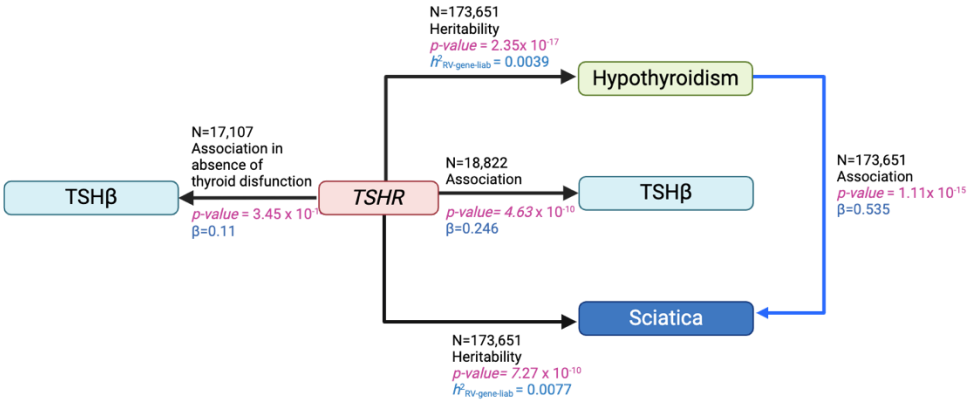
Hemorrhoids	3	*MB21D2*	1.933E-03	5.469E-09
Hemorrhoids	16	*TP53TG3D*	1.226E-03	8.067E-09
Hemorrhoids	2	*KIF5C*	2.057E-03	8.719E-09
Hemorrhoids	8	*KCNV1*	1.330E-03	1.047E-08
Hemorrhoids	10	*PALD1*	2.688E-03	1.083E-08
Hemorrhoids	2	*TTL*	1.427E-03	1.163E-08
Hemorrhoids	6	*DHX16*	2.353E-03	1.970E-08
Hemorrhoids	16	*ZC3H18*	2.811E-03	3.103E-08
Hemorrhoids	1	*CDC42*	9.330E-04	5.106E-08
Hemorrhoids	17	*TVP23C-CDRT4*	1.320E-03	8.116E-08
Hemorrhoids	6	*ZNF292*	3.322E-03	1.138E-07
Hypercholesterolemia	19	*LDLR*	2.058E-03	7.270E-13
Hypothyroidism	14	*TSHR*	3.906E-03	2.355E-17
Hypothyroidism	2	*COL5A2*	3.311E-03	8.581E-11
Hypothyroidism	19	*ZNF583*	1.809E-03	2.706E-09
Hypothyroidism	19	*ZNF490*	1.784E-03	9.011E-09
Hypothyroidism	8	*TATDN1*	1.887E-03	9.790E-09
Hypothyroidism	17	*TMEM256-PLSCR3*	6.810E-04	1.387E-08
Hypothyroidism	5	*REEP5*	1.249E-03	1.355E-07
Obesity	2	*PPP1R21*	2.213E-03	1.578E-08
Obesity	18	*SKA1*	1.257E-03	2.817E-08
Obesity	12	*SCARB1*	1.822E-03	1.096E-07
Pneumonia	17	*SRSF2*	1.969E-03	4.769E-16
Pneumonia	12	*OR10A7*	1.839E-03	9.027E-11
Pneumonia	13	*CPB2*	1.786E-03	1.350E-09
Pneumonia	5	*PANK3*	1.462E-03	1.590E-09
Pneumonia	4	*BST1*	1.658E-03	3.781E-08
Pneumonia	7	*RBAK-RBAKDN*	1.055E-03	8.854E-08
Pneumonia	8	*ERLIN2*	1.094E-03	1.085E-07
Pneumonia	15	*IDH2*	1.637E-03	1.104E-07
Renal failure	15	*IDH2*	1.848E-03	4.490E-10
Renal failure	14	*SRSF5*	1.319E-03	8.760E-09
Renal failure	17	*SRSF2*	1.011E-03	1.533E-08
Renal failure	1	*LIX1L*	1.192E-03	4.428E-08
Renal failure	12	*ABCC9*	1.912E-03	4.594E-08
Type 2 diabetes	7	*GCK*	2.641E-03	5.676E-20

Type 2 diabetes	5	*AGXT2*	2.107E-03	1.008E-10
Type 2 diabetes	3	*HRH1*	1.803E-03	7.054E-10
Type 2 diabetes	11	*FOLH1B*	1.662E-03	3.107E-09
Type 2 diabetes	17	*SCO1*	1.559E-03	9.983E-09
Type 2 diabetes	2	*PDK1*	1.724E-03	1.119E-08
Type 2 diabetes	11	*MYRF*	2.050E-03	1.318E-08
Type 2 diabetes	2	*DNMT3A*	1.885E-03	8.198E-08
Urinary tract infection	4	*CD38*	2.173E-03	1.452E-10
Urinary tract infection	9	*PDCL*	1.567E-03	1.227E-09
Urinary tract infection	7	*VPS41*	2.462E-03	1.699E-09
Urinary tract infection	8	*STMN2*	1.195E-03	3.381E-09
Urinary tract infection	1	*STXBP3*	1.913E-03	9.840E-09
Urinary tract infection	19	*ZNF587B*	1.953E-03	1.574E-08
Urinary tract infection	12	*C12orf43*	1.701E-03	2.990E-08
Urinary tract infection	22	*CBX6*	1.624E-03	5.274E-08
Urinary tract infection	19	*ANO8*	2.663E-03	1.028E-07
Urinary tract infection	1	*AADACL4*	1.951E-03	1.167E-07
Urinary tract infection	14	*TEX22*	1.302E-03	1.380E-07

Supplementary Table 2- Association between *TSHR* and TSHB

Biomarker	Variant/ Gene	CMAC	CMAF	Estimate	Standard error	P-Value
TSHB	14:80955786:G:C_C	319	0.008474	0.187	0.049	1.58 E-04
TSHB	14:81139699:T:A_A	1	0.000027	-3.353	0.860	9.70E-05
TSHB	14:81143695:G:A_A	15	0.000398	0.873	0.230	1.5 E-04
TSHB	*TSHR* gene burden	496	0.013176	0.246	0.039	4.63E-10

N=18822



Supplementary Figure 1: Putative Functional and epidemiological consequences of TSHR mutation. The results are obtained with a linear regression model, correcting for age, sex and 20 PCs. N Sciatica cases=2,309. N hypothyroidism cases=11,246.