Year : 2023

Integrating the genetics of omics and diseases for drug-target and drug-response predictions

Sadler Marie Teresa Cathérine

UNIL | Université de Lausanne

Faculté de biologie
et de médecine

Département de Biologie Computationnelle

Integrating the genetics of omics and diseases for drug-target and drug-response predictions

Thèse de doctorat ès sciences de la vie (PhD)

présentée à la

Faculté de biologie et de médecine
de l'Université de Lausanne

par

Marie Teresa Cathérine SADLER

Master de l'EPFL

Jury

Prof. Petr Broz, Président
Prof. Zoltán Kutalik, Directeur de thèse
Prof. Lili Milani, Experte
Prof. Guillaume Paré, Expert

Lausanne
(2023)

UNIL | Université de Lausanne

Faculté de biologie et de médecine

Département de Biologie Computationnelle

Integrating the genetics of omics and diseases for drug-target and drug-response predictions

Thèse de doctorat ès sciences de la vie (PhD)

présentée à la

Faculté de biologie et de médecine
de l'Université de Lausanne

par

Marie Teresa Cathérine SADLER

Master de l'EPFL

Jury

Prof. Petr Broz, Président
Prof. Zoltán Kutalik, Directeur de thèse
Prof. Lili Milani, Experte
Prof. Guillaume Paré, Expert

Lausanne
(2023)

UNIL | Université de Lausanne

Faculté de biologie
et de médecine

Ecole Doctorale
Doctorat ès sciences de la vie

Imprimatur

Vu le rapport présenté par le jury d'examen, composé de

Président·e	Monsieur	Prof.	Petr	**Broz**
Directeur·trice de thèse	Monsieur	Prof.	Zoltán	**Kutalik**
Expert·e·s	Madame	Prof.	Lili	**Milani**
	Monsieur	Prof.	Guillaume	**Paré**

le Conseil de Faculté autorise l'impression de la thèse de

Marie Teresa Cathérine Sadler

Master in Chemical engineering and biotechnology, EPFL - Ecole Polytechnique Fédérale de Lausanne, Suisse

intitulée

Integrating the genetics of omics and diseases for drug-target and drug-response predictions

Lausanne, le 12 janvier 2024

pour le Doyen
de la Faculté de biologie et de médecine

Petr Brož

Prof. Petr Broz

Integrating the genetics of omics and diseases for drug-target and drug-response predictions

Marie Teresa Cathérine Sadler

Acknowledgements

It has been an incredible journey discovering the world of genomics and going through the learning process comprised of iterative steps of trying, failing and succeeding. At times I felt overwhelmed and stuck, but more often than not it felt very rewarding progressing piece by piece. I am very grateful to all the people who supported me throughout these years and made this a profound scientific and cultural experience.

First and foremost, I would like to thank **Zoltán** who has been an extraordinary mentor and PhD supervisor. I could not have advanced at the same pace without the weekly meetings to discuss both high- and detailed level science. Thank you for teaching me statistical concepts, putting things into perspective when it felt like nothing was working out and keeping me motivated throughout the PhD. I highly appreciated the importance you give to mentoring, work-life balance and the freedom in choosing research projects. Thank you for supporting my research stay abroad, the unparalleled communication and the many collaborative opportunities you enabled.

Next, I would love to thank the whole **SGG** group for these memorable years and great social and scientific exchanges. I could not have been luckier to start my PhD together with **Chiara** who, besides being the best biology advisor I could have wished for, has become a very dear friend with whom to share the ups and downs. Your ambitions and scientific commitments have been truly inspiring. Thank you for all your help in shaping my research work, enabling collaborations and the fun we had during conferences and retreats. **Liza**, while it may have taken a while to get to know the sneaky funny character hiding behind the perfectly organised persona, it was a true pleasure sharing the office with you and having a buddy with an aligned schedule on being a visiting researcher and finishing the thesis. Thank you for all your SGG social engagements, your help whenever needed and introducing me to stargazing. **Adriaan**, I very much enjoyed the reflective conversations we had about science and beyond. Thank you for all your open thoughts and entertaining moments. **Tabea**, your sharing of out of the ordinary experiences has always been a great source of laughter, thank you! **Leona**, while I may not comprehend your athletic performances, I truly admire your energy and dedication to optimise the SGG efficiency. Thank you for being such an active and committed group member! **Samuel**, it has been an immense pleasure collaborating with such a thoughtful and appreciative person. Thank you for all the great exchanges and kind gestures. **Robin**, thank you for all the memorable times during social events and I am very happy you joined the group. **Alexander**, thank you for joining the group and being an outstanding collaborator. Your attention to detail and sharp mind were greatly appreciated. Thank you!

I would also like to thank all the past group members. **Eleonora**, thank you for helping me get started on my PhD. **Ninon**, thank you for answering all my newbie questions. **Sven**, thank you for all the philosophical discussions, the bike rides, and the many times you lit up my day with your constantly good mood. **Marion**, thank you for being a role model on how to keep life and PhD in balance and introducing the group to the most iconic game ever. You and Sven were a true source of inspiration in making me pursue a doctoral degree. **Malú**, it was really great having you in Lausanne, thank you for all the great exchanges and memories. **Kaido**, quiet and humble, but very intelligent, thoughtful and aspiring, I very much enjoyed the honest and profound communication. Your dedication to educating and sharing your research are highly admirable and I have learnt a lot from you, scientifically and personally. Thank you!

A special thanks to the **Helices** and **PharmGKB** team for your warm welcome at Stanford University. Thank you, **Russ**, for being such a great mentor. Your scientific guidance and vibrant energy made my stay at Stanford truly enriching and inspiring. Thank you for facilitating so many networking opportunities and providing advice on career perspectives. **Lu**, you would always surprise with your wittiness. I highly appreciated your curiosity and energy to move things forward, and I have learnt a lot collaborating with you. **Delaney**, it was a great pleasure connecting with a fellow Fulbrighter. Thank you for guiding me around and being such a devoted collaborator!

Many thanks also to the **DBC** people. Besides all the apéros, the pingpong sessions have always been a lot of fun. **Deepak**, **Ioanna**, **Carlos**, **Aisima**, thank you for all your great company in the office. **Diana**, your energy and fearless spirit were always lightening the mood, thank you for all the adventures. Also, many thanks to **Alex Reymond** from the CIG. Thank you for introducing me to the genetics world in the first place and for your advice and support throughout the years.

I am very grateful to my **friends** for all the fun and support outside of work. **Lauriane**, thank you for all the outdoor adventures, crazy or not, we did together. **Philippe** thank you for being such a motivating and fun sports companion. I am also very grateful to the **Rushteam** group for all the trainings and training camps. Many thanks to all the volunteering work and efforts that go into keeping together such an amazing group of people that felt like family over the years. A special thanks to the whole **Master's crew** for all the wonderful get-to-togethers and support. You are the people who make Switzerland feel like home. I would also like to thank my **friends in Luxembourg** - the nostalgic and catch-up reunions feel very precious to me. I am also deeply grateful to **my family** for their unconditional support. I am very thankful for all the moments we share together, your help and advice whenever needed and being the people I can rely on no matter what happens.

Last, but not least, my biggest thanks go to **Raphael** who has been on my side the whole time. I would like to thank you for always listening to me, supporting me whatever career path I chose to take, cheering me up whenever I feel down and being the indispensable cornerstone of my life. You were always understanding when the schedule went overtime and provided balance when I got lost in things that do not matter. Thank you for your support and kindness!

Abstract

Biobanks with genetic and phenotypic information of hundreds of thousands of participants offer new opportunities to study the genetic underpinning of disease aetiology. Beyond genetics and genomics, recent advances in technologies enabled the generation of a variety of *omics* data at an unprecedented scale opening up the possibility of studying the consequences of genetic variation in the mediating molecular space.

Through the lens of statistical genetics, this thesis studies the integration of omics and disease genetics to identify putative novel disease mechanisms and further explores the role genetics can play in drug development pipelines and personalised medicine. In the first part, large-scale omics quantitative trait loci (QTL) data are combined with genome-wide association studies (GWAS) in a causal integrative Mendelian randomisation framework to identify molecular mechanisms mediating the path from genotype to phenotype. Known as 'Nature's clinical trial' whereby individuals who carry specific genetic variants may be at risk or protected against a disease, such integrative genetics approaches also provide a basis for drug target discovery and ultimately drug development. The second part describes the support of human genetics among approved drugs and how diverse molecular data sources and methodologies can contribute towards such genetic support. Not only is genetics gaining increased interest in target identification but also in patient stratification. Inter-individual variability in drug response is known to harbour a genetic component, and understanding the underlying genetics holds promise to move from 'one size fits all' to a more personalised medicine approach. In the third part, large-scale biobanks coupled to electronic health records (EHRs) with longitudinal clinical and prescription data are harnessed to study the pharmacogenetic efficacy of common cardiometabolic medications. Overall, the results demonstrate the value of omics QTL data to study disease pathways, the benefits of consulting both common and rare genetic variants to identify drug targets, and the challenges and promises of EHRs for drug response studies.

In conclusion, genetics serves as an anchor to establish causality and detect molecular disease mechanisms and drug targets, while also enabling patient stratification for more effective and personalised treatment strategies.

Résumé

Les biobanques contenant des informations génétiques et phénotypiques de centaines de milliers de participants offrent de nouvelles opportunités pour étudier l'étiologie génétique des maladies. Au-delà de la génétique et de la génomique, les récents progrès technologiques ont permis la génération d'une variété de données omiques à une échelle sans précédent, ouvrant la possibilité d'étudier les conséquences de la variation génétique dans l'espace moléculaire médiateur.

À travers le prisme de la génétique statistique, cette thèse étudie d'abord l'intégration de la génétique omique et de la génétique des maladies afin d'identifier de nouveaux mécanismes pathologiques putatifs et ensuite le rôle que la génétique peut jouer dans le développement de médicaments et dans la médecine personnalisée. Dans la première partie, des données omiques à grande échelle sur les loci de caractères quantitatifs sont combinées avec des études d'association à l'échelle du génome dans un cadre de randomisation mendélienne afin d'identifier des mécanismes moléculaires médiant le lien entre le génotype au phénotype. Connue sous le nom de 'essai clinique de la nature', où les individus porteurs de variants génétiques spécifiques peuvent être exposés ou protégés contre une maladie, ces approches génétiques intégratives fournissent également une base pour la découverte de cibles médicamenteuses. La deuxième partie décrit le soutien de la génétique humaine parmi les médicaments approuvés et comment diverses méthodologies et sources de données moléculaires peuvent contribuer à un tel soutien génétique. La génétique suscite un intérêt croissant non seulement pour l'identification des cibles, mais aussi pour la stratification des patients. La variabilité interindividuelle de la réponse aux médicaments comporte une composante génétique, et la comprendre promet de passer d'une approche 'taille unique' à une approche de médecine plus personnalisée. Dans la troisième partie, des biobanques couplées à des dossiers de santé électroniques (DSE) contenant des données cliniques et de prescription sont exploitées pour étudier la pharmacogénétique de la réponse aux médicaments cardiométaboliques courants. Dans l'ensemble, les résultats démontrent la valeur des données omiques pour étudier les mécanismes moléculaires, l'importance des variants génétiques communs et rares pour identifier les cibles médicamenteuses, et les défis et promesses des DSE pour les études sur la réponse aux médicaments.

En conclusion, la génétique permet d'établir la causalité pour détecter les mécanismes moléculaires et les cibles médicamenteuses, ainsi que la stratification des patients pour des traitements plus efficaces et personnalisées.

Contents

List of Figures

List of Tables

Acronyms

ADME absorption, distribution, metabolism and excretion.

ADR adverse drug reaction.

AoU All of Us.

CMap Connectivity Map.

CNV copy-number variation.

DNA deoxyribonucleic acid.

DNAm DNA methylation.

EHR electronic health record.

GPCR G-protein-coupled receptor.

GWAS genome-wide association studies.

HDL high-density lipoprotein cholesterol.

HR heart rate.

INR international normalized ratio.

IV instrumental variable.

LD linkage desequilibrium.

LDL low-density lipoprotein cholesterol.

LoF Loss-of-Function.

MACE major adverse cardiovascular events.

MAF minor allele frequency.

MDD major depressive disorder.

ML machine learning.

MP mediation proportion.

MR Mendelian randomisation.

MS mass spectrometry.

MVMR multivariable Mendelian randomisation.

NGS next-generation sequencing.

NMR nuclear magnetic resonance.

PCI percutaneous coronary intervention.

PGx pharmacogenomics.

pheWAS phenome-wide association studies.

PPI protein-protein interaction.

PRS polygenic risk scores.

QTL quantitative trait loci.

RCT randomised controlled trial.

RNA ribonucleic acid.

RNA-seq RNA sequencing.

SBP systolic blood pressure.

scRNA-seq single-cell RNA sequencing.

SNP single-nucleotide polymorphism.

TC total cholesterol.

TSS transcription start site.

TWAS transcriptome-wide association studies.

UKBB UK Biobank.

WES whole-exome sequencing.

WGS whole-genome sequencing.

Chapter 1

Introduction

1.1 Aims of the thesis

Since completion of the *Human Genome Project* in the early 2000s [1, 2], genomics has evolved into a highly multidisciplinary field. From fundamental biology to medicine to biotechnology, genomics provides answers about evolutionary processes and phenotypic variability, and a basis for therapeutic solutions and genetic engineering. Vast amounts of data generated by next-generation sequencing (NGS) and genotyping offer unprecedented opportunities to understand the genetic basis of traits, diseases and other biological phenomena.

High-throughput technologies have given rise not only to large-scale genomics data, but to a variety of *omics* data that cover a wide spectrum of molecules such as transcripts, proteins and metabolites. While genetic associations with biological traits and diseases assessed through so-called genome-wide association studies (GWAS) provide valuable insights into human biology, omics data can inform us about mediating molecular relationships often through so-called omics quantitative trait loci (QTL) data. The molecular space mediates also pharmaceutical effects of drugs and Figure 1.1 shows a simple schematic of how genetics, omics, diseases and drugs can be connected and modelled to gain mechanistic insights into the complex interplay between omics and diseases and understand inter-individual differences in drug response.

Every node in the triangular illustration in Figure 1.1 representing drugs, omics and traits, respectively, exists in a high-dimensional space, and every node and edge represents a sub-discipline itself. As our knowledge in each of these individual fields expands, we will be able to derive integrative models with increased granularity. With this overall goal in mind, this thesis focuses on three major themes: *systems genetics*, *drug target identification* and *pharmacogenomics*. The introduction is structured into three corresponding sections, each explaining fundamental concepts and summarising current knowledge, data and statistical tools. The systems genetics part deals with the mechanistic interaction of omics and traits, the drug target identification part focuses on the role of genetics in detecting effective drug targets and the pharmacogenomics part covers our current understanding of the genetics underlying inter-individual variability in drug response. The following chapters summarise our research and results that have been published in scientific journals or are currently in preparation (Appendix A-C) and that contribute to the three themes: i) Mediation between omics layers and complex traits (Chapter 2), ii) Gene prioritisation approaches to identify drug targets (Chapter 3), and iii) Large-scale biobanks for pharmacogenomic research (Chapter 4). In the final Chapter 5, I will discuss limitations and challenges in complementing gaps in current drug-omics-disease models and how this could be solved in the future.

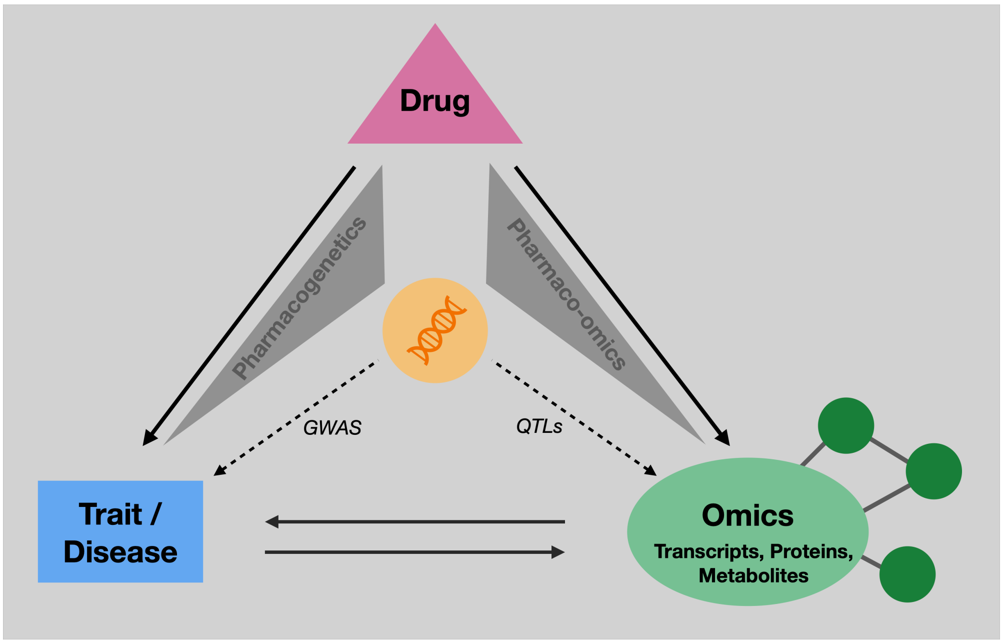


Figure 1.1: Schematic representation of the connections between drugs, traits/diseases, genetics and omics. Genome-wide association studies (GWAS) inform about the genetics of traits while omics quantitative trait loci (QTL) studies inform about the genetics of omics. Pharmacogenetics describes inter-individual drug response due to genetics and the extension to pharmaco-omics expands the genetic dimension to the omics space.

1.2 The path from genotype to phenotype

Each cell in the human body contains an identical copy of deoxyribonucleic acid (DNA). Yet, cells can be very different as a result of how the genetic code is used. Differentiated cell types distinguish themselves from different gene expression patterns with not every gene being expressed in every cell [3]. The central dogma of molecular biology governs the flow of genetic information - from DNA to ribonucleic acid (RNA) to proteins where the latter largely decide the appearance and behaviour of a cell. The central dogma consists of three main steps of which the first one is DNA replication [4]. DNA polymers composed of sequences of any of the four nucleotides A (adenine), T (thymine), G (guanine) and C (cytosine) *replicate* so that each daughter cell receives an identical DNA copy during cell division. DNA is then *transcribed* into RNA by the RNA polymerase, which when coding for a protein is called messenger RNA (mRNA). Lastly, ribosomes *translate* mRNA into peptides or proteins composed of chains of amino acids (Figure 1.2).

The central dogma of molecular biology governs the flow of genetic information.

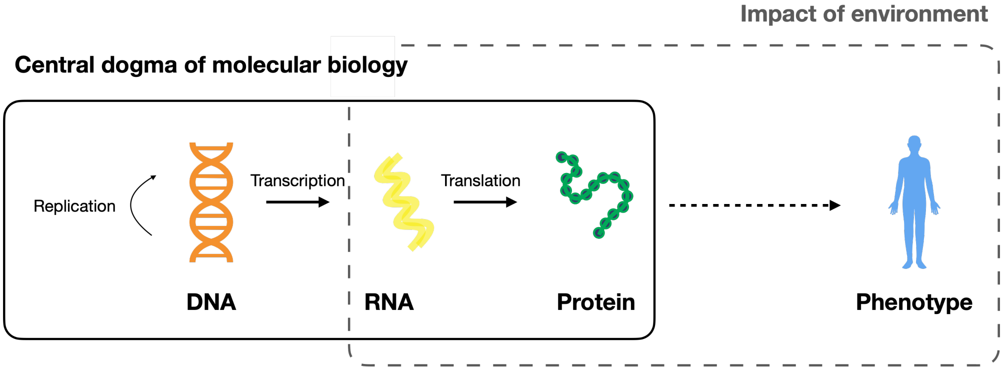


Figure 1.2: The central dogma of molecular biology governs the flow of DNA information from replication to transcription to translation. Environmental factors might influence this process or by themselves shape observed phenotypes.

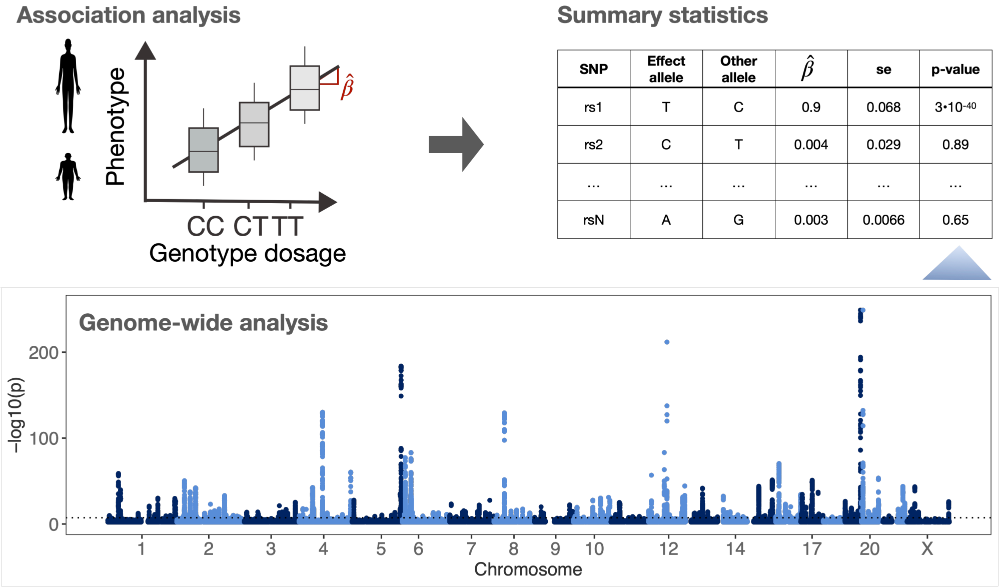
From the central dogma of molecular biology follows that variations in DNA sequences can impact phenotypes through downstream mechanisms affecting gene expression and protein levels. Thus, finding genetic associations with traits and diseases should improve our understanding of the causal gene behind the observed phenotype. Identification of mutations in genes causing monogenic diseases, also called Mendelian diseases, has greatly elucidated our understanding of rare genetic disorders. This success is reflected in the Online Mendelian Inheritance in Man (OMIM) database initiated in the 1960s which covers all known Mendelian disorders and over 16,000 genes (`https://omim.org/`, accessed July 2023). Conversely, common diseases such as cardiovascular, neurodegenerative and auto-immune diseases are typically polygenic, caused by multiple genetic and environmental risk factors.

Genetic associations are more puzzling in polygenic than monogenic disorders.

As we have entered the era of GWAS, hundreds of thousands of genetic variants have been identified to influence disease risk, with common traits involving often over 1,000 independent variants [5]. Yet, the interpretation of GWAS results has proven to be more difficult than expected [6]. In the following, I will expand on the evolution of GWAS and focus on how systems genetics approaches can bridge the path from genotype to the trait of interest.

1.2.1 The GWAS era

GWAS have revolutionised the field of genetics in the past decade (Figure 1.3). Currently, over 500,000 associations between genetic variants and human traits have been reported in the GWAS Catalog (`https://www.ebi.ac.uk/`, accessed May 2023) stemming from $>$ 6,000 publications [7].



SNP	Effect allele	Other allele	$\hat{\beta}$	se	p-value
rs1	T	C	0.9	0.068	$3 \cdot 10^{-40}$
rs2	C	T	0.004	0.029	0.89
...	...	...	...	...	...
rsN	A	G	0.003	0.0066	0.65

Figure 1.3: Illustration of genome-wide association studies (GWAS). Based on individual-level data, a genome-wide scan is performed in which a phenotype (Y) is regressed on the genotype dosage (G) to yield effect sizes ($\hat{\beta}$, corresponding to the least square estimator $(G^TG)^{-1}G^TY$ when assuming a linear model) and p-values that make up summary statistics. Summary statistics no longer contain personally identifiable information and can be shared publicly. A Manhattan plot shown at the bottom displays the association strength (i.e., negative logarithm of the p-values) of the phenotype with the genetic variants along the genomic coordinates.

The steady increase in associations can be attributed to the decreasing cost in genotyping microarray and sequencing technologies giving rise to large-scale population biobanks often coupled to deep phenotyping data [8, 9, 10, 11]. While genotyping microarrays typically cover less than a million genetic variants, imputation techniques that rely on comprehensive reference panels with a diversity of sequenced genomes allow to infer genotypes for millions of variants, thus tremendously increasing genome coverage [12]. As the technologies evolve, so do the types of genetic associations. Large-scale projects such as the UK Biobank (UKBB) have made whole-exome sequencing (WES) and whole-genome sequencing (WGS) data available and the GWAS Catalog harbours now $>$ 5,000 sequencing-based GWAS (seqGWAS) [7]. Sequencing also allows the assessment of rare variants (minor allele frequency (MAF) typically below 0.01) and rare-variant associations tests through the aggregation of multiple variants in a gene or genomic region, so-called gene-based GWAS, are increasingly being reported [13, 14]. Besides rare variant associations, population-scale copy-number variation (CNV) studies have resulted in novel associations and contribute to the diverse genetic landscape [15, 16].

The number and types of genetic associations are steadily increasing.

GWAS unravel the genetic architecture of complex traits and diseases, and as such have a wide applicability from quantifying genomic diversity to disease risk prediction up to causality studies and drug discovery. Polygenic risk scores (PRS) have become popular in recent years to predict an individual's genetic predisposition for a disease which allows for risk stratification and optimised prevention strategies [17]. Other GWAS downstream applications include the estimation of heritability [18], genetic correlations [19], identifying causal relationships between risk factors and health outcomes [20] and interpreting GWAS signals in a biological context [21].

GWAS data can have diverse downstream applications.

Yet, finding the causal single-nucleotide polymorphism (SNP) and biological mechanisms underlying an observed genetic association can be far from trivial [21]. SNPs measured on a microarray may not cause the trait, but may be in linkage desequilibrium (LD) with the causal (unmeasured) SNP. Furthermore, SNPs in a region are often inherited together, forming a so-called haplotype, and multiple close-by SNPs can be equally strongly associated with the trait without being causal. LD patterns can be very complex and several statistical fine-mapping methods have been developed to decipher genetic regions and find genetic variants that are functionally disease-relevant [22]. Once a causal variant has been identified, mapping this variant to a gene can represent an additional challenge. The gene with the closest transcription start site (TSS) could be considered as the causal gene, however, if multiple genes are located in the region, this strategy may fail.

Finding the causal variant and gene in GWAS can be challenging.

The majority of GWAS signals of complex traits fall into non-protein-coding regions of the genome further complicating the task of finding the causal gene in gene-rich regions [23]. These non-coding signals were found to be enriched for regulatory elements such as chromatin accessibility, transcription factor binding, and histone marks associated with transcriptional regulatory activity [24, 23, 25] suggesting that altered levels of molecular traits mediate the path from genotype to phenotype. Omics QTL data provide new opportunities to identify such mediators as genetic associations with molecular traits including DNA methylation (DNAm), transcripts and proteins could furnish the missing link between SNP-trait associations. In the next sections, I will cover recent developments that led to large-scale omics QTL datasets and present statistical methods that allow for QTL and GWAS data integration. Finding the causal gene is at the heart of drug target identification and besides QTL-based approaches, other techniques exist to identify gene-trait associations which will be presented in Section 1.3.

GWAS signals predominantly fall into non-coding regions and are enriched for regulatory elements.

1.2.2 Large-scale omics datasets

The aetiology of complex diseases is very intricate and a combination of different factors including environmental ones are likely to play a role in causing disease [26]. Insights into the underlying biological mechanisms could be gained by assessing diverse omics data, and in recent years, advances in high-throughput and cost-efficient omics technology contributed to the emergence of large-scale omics datasets. A given omics layer covers a whole spectrum of molecules of the same type providing deep insights into biological systems at an unprecedented scale. For instance, the protein space covers around 20,000 protein-coding molecules, of which 92% can be probed with current proteomics technologies and 99% when including transcriptomics (`https://www.proteinatlas.org`, accessed July 2023). The number of omics entities can differ drastically from one level to another nearing a million when considering the number of CpG sites currently measured in the DNAm space [27]. To identify cascading genetic effects, there is a great interest in the genetic basis of molecular data, or omics QTL, that summarise the genetic associations with molecular phenotypes and are key ingredients to most integrative genetics approaches. In recent years, studies analysing the association between genetics and omics entities have expanded both in the number of phenotypes probed in an omics layer and sample size (Figure 1.4).

Omics data have the potential to capture the complexity of biological systems.

In the following, I will briefly explain the specificities of omics technologies used to measure DNAm, transcript, protein and metabolite levels and place the technological progress in the context of large-scale omics datasets (Table 1.1). Genomics constitutes the base layer and compared to the other layers does not undergo changes throughout life. It will not be listed individually as it is integral to QTL associations. Broadly speaking, the phenome could also be considered as an omics layer, though genotype-phenotype associations, or more conventionally GWAS are covered more extensively in Subsection 1.2.1.

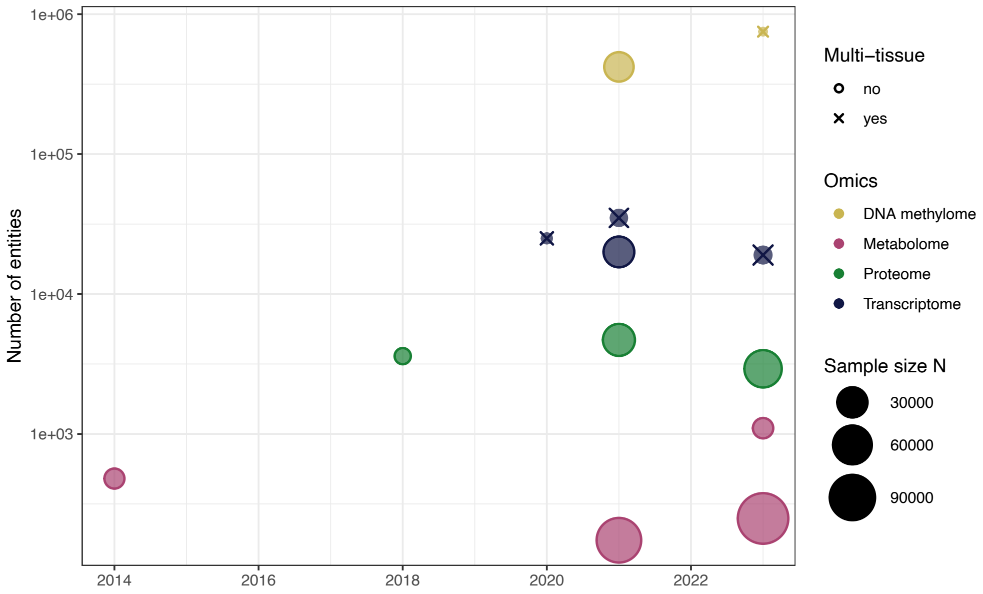


Figure 1.4: Evolution of major large-scale publicly available omics QTL datasets with respect to sample size, number of assessed entities and tissues. Numerical values are in Table 1.1.

DNA methylome: In animals, DNAm occurs mostly on cytosines in CpG sites (DNA segments where a cytosine is followed by a guanine nucleotide along the 5' → 3' direction) of which 60-80% are methylated throughout the genome [41]. Most common techniques to measure DNAm are affinity enrichment or bisulfite conversion based. While whole-genome bisulphite sequencing is considered the gold standard, more cost-effective methods which require less technical expertise such as the Illumina Infinium BeadChips have become popular in recent years. Also bisulfite conversion based, targeted CpG sites are genotyped at single base resolution using probes on a microarray [42, 43]. Studies using the Illumina Infinium HM450 (2011) and EPIC (2016) BeadChip measuring over 450,000 and 850,000 probes, respectively, were included in the Genetics of DNA Methylation Consortium (GoDMC) [28].

Transcriptome: Transcriptomics technologies allow to quantify gene expression by measuring mRNAs, non-coding RNAs and small RNAs. Techniques to quantify the transcriptome are hybridisation- or sequence-based. Hybridisation-based approaches rely on high-density oligo microarrays. Although high-throughput and relatively inexpensive, this technique relies on ex-

Table 1.1: Major large-scale publicly available omics QTL datasets.

Omics	Cohort/ Consortium	N	# Entitities	Tissue	Year (published)	Reference
DNA methylome	GoDMC	27,750	∼420,000	whole blood	2021	[28]
	GTEx	424	∼750,000	9 tissues	2023	[29]
Transcriptome	GTEx	838	∼25,000	49 cell types and tissues	2020	[30]
	eQTLGen	31,684	∼20,000	whole blood	2021	[31]
	eQTL Catalogue	5,714	∼35,000	69 cell types and tissues	2021	[32]
	Metabrain	6,523	∼19,000	7 brain tissues	2023	[33]
Proteome	INTERVAL	3,301	∼3,600	blood plasma	2018	[34]
	deCODE	35,559	∼4,700	blood plasma	2021	[35]
	UK Biobank	54,306	∼2,923	blood plasma	2023	[36]
Metabolome	KORA/TwinsUK	7,824	∼480	whole blood	2014	[37]
	26 cohorts	86,507	174	whole blood	2021	[38]
	UK Biobank	118,461	249	whole blood	2023	[39]
	CLSA	8,299	∼1,100	whole blood	2023	[40]

isting genome sequences, and suffers from a limited detection range due to technical issues such as cross-hybridisation and saturation of signals [44]. On the other hand, RNA sequencing (RNA-seq) allows for *de novo* assembly and thus the discovery of new transcripts. Furthermore, background signals are not an issue and there is no upper limit for quantification [44]. In the eQTLGen Consortium, studies utilising both microarray (N = 25,263, 79.7%) and RNA-seq (N = 6,422, 20.3%) methods were meta-analysed [31]. RNA-seq methods have constantly evolved over the past decade, and new opportunities emerge with single-cell RNA sequencing (scRNA-seq), short-read, long-read and direct RNA-seq as well as spatial transcriptomics [45].

Proteome: Affinity proteomics is widely used for probing the plasma proteome. Paired, nucleotide-labeled antibody probes (Olink) and single-strand DNA aptamer reagents (SomaScan) are the main assays used in large-scale human studies. In the Olink assay, the antibody pair which is labelled with unique complimentary oligonucleotides (probes) is required to bind to the target protein. Upon binding, the probes hybridise due to the close proximity and are quantified using NGS. In the SomaScan assay, aptamers bind to the target proteins and corresponding DNA aptamer concentrations that correlate with protein concentrations are quantified on a DNA microarray. While both platforms allow for high-throughput profiling, they suffer from lower specificities than liquid chromatography–mass spectrometry (LC-MS)–based methods

[46]. The Olink Explore 3072 platform has been used to measure ∼3,000 proteins in the UKBB [36]. The SomaScan assay v4 which targets ∼5,000 unique human proteins was used in the deCODE study [35].

Metabolome: Metabolites are predominantly measured through either nuclear magnetic resonance (NMR) or mass spectrometry (MS). NMR spectrometry is highly reproducible, allows for structure elucidation and is suitable for a great variety of (not pure) samples. However, its low sensitivity and signal overlap can lead to ambiguous read-outs. MS, targeted or untargeted, gas and/or liquid chromatography based, has a great selectivity and sensitivity, and allows to analyse a wide spectrum of metabolites, although untargeted approaches can make metabolite identification challenging [47]. In the UKBB, Nightingale Health's metabolic biomarker platform (NMR) was used to quantify 249 metabolic measures of which most relate to lipoprotein metabolism [39]. The Metabolon platform based on MS has been used to measure up to ∼1,300 metabolites (untargeted) in several large-scale genomic studies covering a broader spectrum of metabolites (amino acids, carbohydrates, cofactors and vitamins, energy, lipids, nucleotides, peptides and xenobiotics) [37, 48, 40].

1.2.3 Integrating the genetics of omics and diseases

Genetics provides a basis to disentangle omics-trait correlations.

Interpreting GWAS signals can be difficult as they often fall into non-coding regions and can harbour a multitude of variants in LD that mask the causal one (Subsection 1.2.1). Furthermore, if the region contains multiple genes it can be hard to ascertain the gene underlying the disease association. The wealth of omics QTL data available in the public domain provides new opportunities to find diverse mediators of SNP-trait associations in a high-throughput, data-driven fashion. While omics themselves, transcript or protein levels, can be correlated to a trait of interest to identify associated genes, correlations can suffer from confounding and reverse causality. As genetic variants remain unaltered throughout life, SNP-trait associations should reflect a cause and not a consequence of a disease (a SNP-trait association is not caused by the trait itself). Based on QTL data and Mendelian randomisation (MR) (Box 1, Figure 1.7), omics-trait correlations can be disentangled into cause, consequence and mere confounding. Through the application of MR, we have demonstrated that observed gene expression-trait correlations are more likely to arise due to reverse causality (i.e., trait-induced) and others concluded the same for DNAm [49, 28].

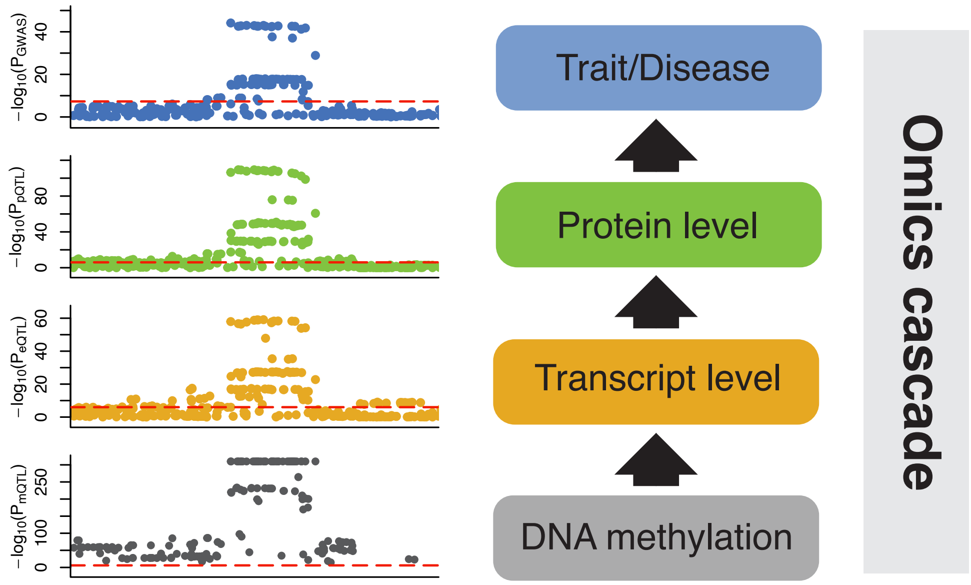


Figure 1.5: Illustration of how the integration of genetic association data across multiple omics levels can result in the identification of molecular mechanisms underlying GWAS signals. The x-axis represents the chromosomal position for a window typically smaller than 1Mb and the y-axis represents the association strength with molecular and disease traits.

As illustrated in Figure 1.5, molecular disease mechanisms can be detected by following genetic downstream effects from one omics layer to the next. In Box 1, I summarise widely-used omics-disease integration methods

Box 1. Statistical methods for omics-trait associations

Colocalisation (coloc)

Colocalisation assesses whether observed association signals at the same locus in two different studies (e.g. omics QTL and GWAS region) are consistent with a shared causal variant. A colocalisation test calculates the posterior probability for each hypothesis, of which there are five when assuming a single causal variant (e.g. coloc method [50]): $\mathbb{H}_0$, no association with either trait; $\mathbb{H}_1$, association with trait 1, not with trait 2; $\mathbb{H}_2$, association with trait 2, not with trait 1; $\mathbb{H}_3$, association with both traits but at separate causal variants; $\mathbb{H}_4$, association with both traits at a shared causal variant [50]. A large posterior probability for hypothesis 4 supports colocalisation. This method is Bayesian as it sums up all possible configurations supporting a given hypothesis. Extension to multiple causal variants are implemented in eCAVIAR [51] and coloc-Sum of Single Effects (SuSiE) [52].

Mendelian randomisation (MR)

MR techniques calculate the causal effect of an exposure (i.e., omics trait) on an outcome trait (i.e., GWAS trait) by instrumenting the exposure through genetic variants. Analogously to a randomised controlled trial (RCT), where the effect of an intervention is assessed through randomisation (i.e., all potential confounders are distributed evenly among two or more groups), genetic variants inherited randomly at birth and strongly associated with the risk factor of interest allow for an unbiased causal effect estimate [20]. MR relies on strong assumptions which are outlined in Figure 1.7 and which may not hold in practice. Thus, sensitivity analyses and the use of multiple MR methods less sensitive to the violation of one or several assumptions are usually recommended to assess the robustness of results [53]. The most commonly used MR method is the inverse-variance weighted (IVW) method. Other methods suited for omics MR include multivariable Mendelian randomisation (MVMR) accounting for nearby genes in the setting of gene expression MR [54] and principal component analysis (PCA)-MR [55].

SMR-HEIDI

The summary data–based Mendelian randomisation (SMR)-heterogeneity in dependent instruments (HEIDI) method estimates a causal effect of an exposure on an outcome based on the Wald ratio estimate (i.e., SNP-outcome association divided by the SNP-exposure association) of the most significant QTL. The HEIDI test statistic assesses homogeneity in Wald ratio estimates of correlated genetic variants in the region to corroborate vertical pleiotropy [56]. SMR-HEIDI falls into the category of proportional non-Bayesian colocalisation methods [57].

Transcriptome-wide association studies (TWAS)

TWAS, e.g. PrediXcan, correlate the PRS of an omics level with an outcome trait [58, 59]. TWAS can be considered as an *ad hoc* method for causal inference and is fundamentally equivalent to MR without taking into account corresponding assumptions (Figure 1.7) and other theoretical concepts such as weak instrument bias [60].

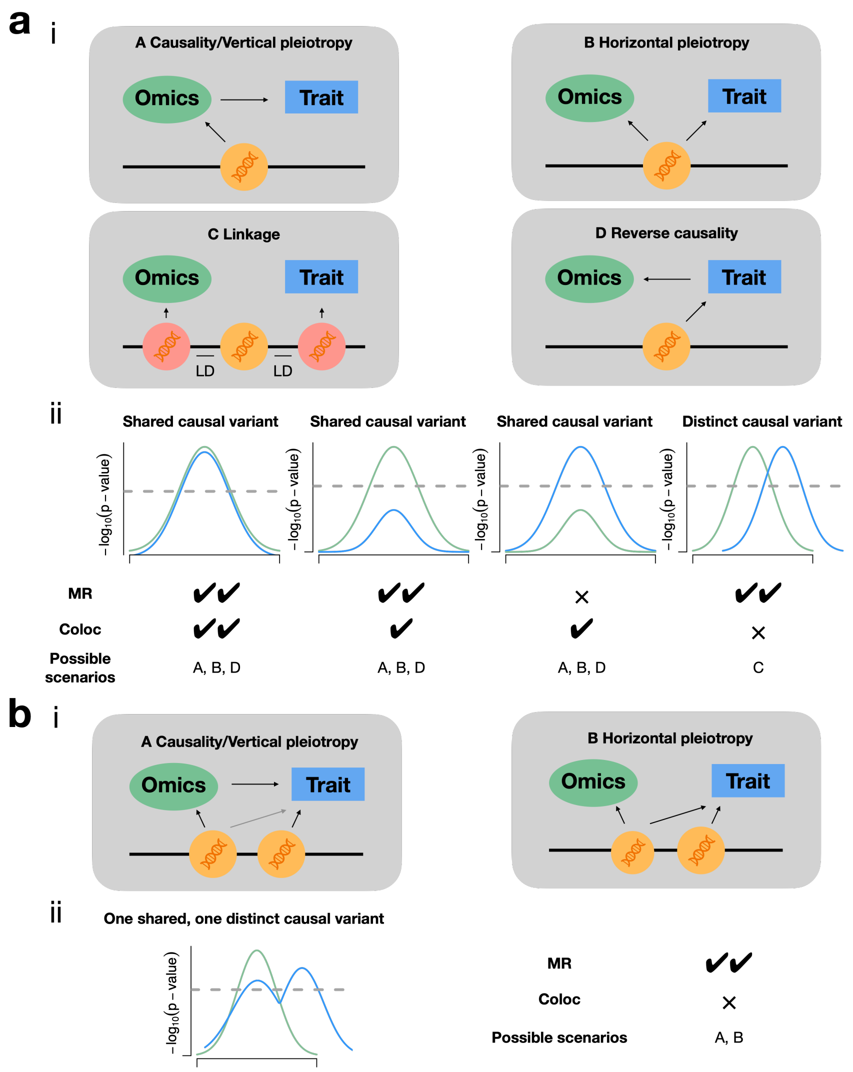


Figure 1.6: **a** Possible pleiotropic scenarios for detecting significant Mendelian randomisation (MR) and colocalisation (coloc) effects with a single genetic variant. i) **A** Causal path from the omics exposure to the outcome trait; **B** Genetic variant affects unrelated exposure and outcome traits independently; **C** Genetic variant is in LD with distinct causal variants that affect the exposure and outcome independently; **D** Effect on the exposure trait is mediated by the outcome trait. ii) Illustrations of observed omics (green) and trait (blue) signals that could result in significant MR or coloc effects depending on whether the underlying causal genetic variants reach significance (horizontal dashed line) and assuming that coloc correctly identifies the underlying causal variants. A single compared to two check marks refers to potential power issues in detection. **b** Possible pleiotropic scenarios for observing significant MR and coloc effects with two genetic variants that are not in LD. i) **A** Causal path from the omics exposure to the outcome trait; **B** Genetic variants affect unrelated exposure and outcome traits independently. ii) Although there is colocalisation at the shared variant, there is none at the distinct variant which results in a significant MR, but no significant coloc effect. Like MR, coloc is unable to distinguish between scenario **A** and **B** (see also **a**), but unlike MR, coloc will not detect pleiotropy in the presence of an additional variant only impacting the exposure or outcome trait. Regional association plots represent the association strength of variants with each omics/trait plotted against chromosomal position. Illustration adapted from [57].

that make use of GWAS and omics QTL summary data to unravel such mechanisms. These methods are suited for the integration of genetic associations across datasets in *cis*, which typically implies a region of 0.5-1Mb, although some of the methods can also be extended to *trans*/genome-wide settings.

While TWAS are comparable to MR studies, there are conceptual differences between MR and colocalisation tests. Colocalisation techniques can be viewed as an extension of fine-mapping applied to multiple traits, and identify shared QTL and GWAS signals. However, shared QTL and GWAS signals can also result from reverse causality, i.e., the trait is causing differential molecular trait levels or horizontal pleiotropy, i.e., the shared variant influences two unrelated traits. In Figure 1.6, several scenarios are illustrated under which MR or coloc or both detect a significant effect between an omics and complex trait. While both methods are ill-suited to distinguish vertical (scenario **A**) from horizontal (scenario **B**) pleiotropy, especially in the case of a single genetic variant, MR is better equipped to distinguish between forward (scenario **A**) and reverse causality (scenario **D**). MR makes a distinction between the exposure and the outcome, and by instrumenting the exposure with (usually) multiple SNPs and making sure that IVs are stronger associated with the exposure than the outcome (Steiger filter [61]), it is more likely to identify forward causal relationships [62]. However, omics exposures can often only be instrumented by a single genetic variant and if this variant is in high LD with distinct causal variants of the exposure and outcome trait a significant MR effect can be identified even if both traits are unrelated (scenario **C**). Colocalisation methods are more likely to attribute such scenarios to $\mathbb{H}_3$ (i.e., association with both traits but at separate causal variants). Zuber *et al.*, compared both methods in a comprehensive review and supported the use of colocalisation as a sensitivity analysis since MR results may suffer from higher false positive rates when there is a limited number of instrumental variables (IVs) in a genetic region of interest [57]. With highly-powered QTL data and multiple IVs, MR estimates become more robust whereas highly-powered data, notably outcome GWAS, may be of disadvantage to colocalisation methods: variants only impacting outcome and not exposure traits do not support the colocalisation hypothesis $\mathbb{H}_4$ even though a genetic region may affect the assessed outcome through multiple pathways (e.g. a coding variant not altering expression levels in addition to a regulatory variant; Figure 1.6b).

MR and colocalisation tests are conceptually different, and provide complementary information.

Methods mentioned so far identify a single molecular trait that mediates genotype-to-phenotype relationships. Identifying causal chains through multiple omics layers is challenging as QTL effects weaken in each consecutive

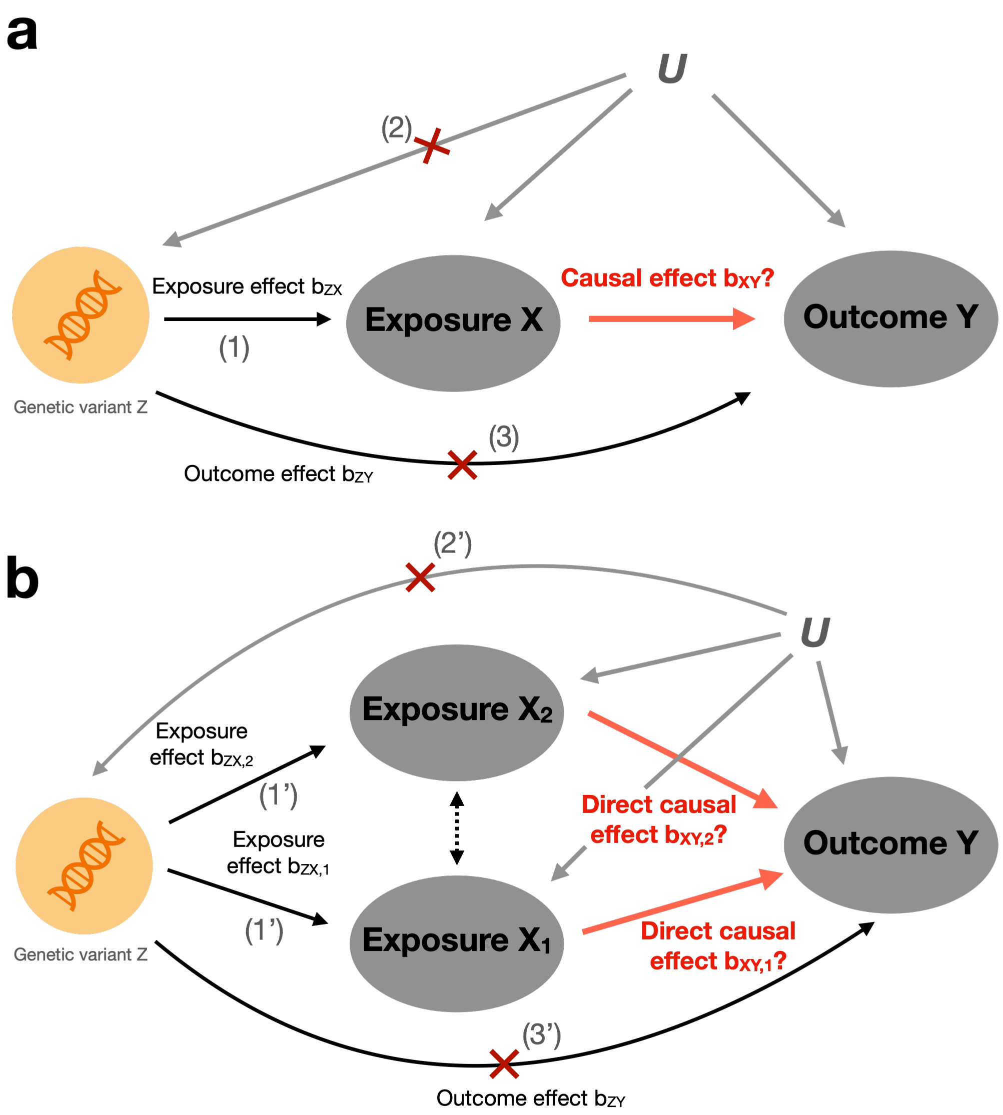


Figure 1.7: Graphs and assumptions of **a** univariable and **b** multivariable Mendelian randomisation (MVMR) to estimate causal effect estimates b_{XY} in the presence of unobserved confounders U of the exposure (X)-outcome (Y) relationships. For simplicity, two exposures are presented, but there can be many more. In both graphs, the genetic variant Z represents a group of instrumental variables (IVs) associated with (at least one of) the exposure(s). In **b** no assumptions are made about the X$_1$ and X$_2$ relationship, however, if X$_2$ were the mediator, we would expect a unidirectional link from X$_1$ to X$_2$. Compared to the causal effect or total causal effect estimated in **a**, the direct causal effect in **b** represents the causal effect conditional on the other exposures included in the model which allows to disentangle individual contributions and mediating relationships.

Under the univariable MR assumptions, Z used as IV must be 1) strongly associated with X ($b_{ZX} \neq 0$), 2) independent of any confounder of the X − Y relationship, 3) conditionally independent of Y given X.

In the MVMR setting, the assumptions remain largely unchanged. Z used as IV must 1') strongly predict X$_i$ conditional on all other included exposures, 2') be independent of all confounders of any of the X$_i$ − Y relationships and 3') be conditionally independent of Y given all included exposures X$_i$ [63].

layer [64]. However, increased omics sample sizes allow to overcome statistical power issues and approaches have emerged to identify multiple molecular mediators. In the simplest case, pairwise associations are combined to infer mechanisms of the scheme: omics trait 1 → omics trait 2 → outcome trait. In the context of MR, this is also known as two-step MR [65]. Colocalisation methods accommodating multiple omics layers have also been developed using a Bayesian statistical framework [66]. However, colocalisation methods do not provide effect sizes and extensions to multiple mediators generally rely on a single causal variant underlying multiple associations [66, 64, 67].

Fewer methods exist that extend to multiple molecular mediators.

MVMR approaches can mitigate some of these issues. They can integrate multiple exposures and/or mediators, and estimate the conditional contribution of each on an outcome. In a mediation analysis, the total causal effect of an exposure on an outcome is dissected into a direct and indirect effect through the mediators. The direct effect is estimated by the MVMR model and the indirect effect can be derived by multiplying the exposure-to-mediator and the mediator-to-outcome effects. Instrumenting the exposure and mediators allows for robust causal inference even in the presence of confounders [65, 63]. In Figure 1.7, MR and MVMR with their respective assumptions are visualised.

MVMR approaches can accommodate multiple exposures and/or mediators.

MR and MVMR methods rely on strong assumptions among which the validity of IVs. IVs should be strongly and directly (not via any confounder or the outcome) associated with the exposure, which in practice is achieved by selecting variants with strong genetic associations. Yet, biased genetic associations can arise due to demographic factors such as population stratification and assortative mating [68, 69] as well as parental effects causing indirect genetic effects [70]. Statistical methods like family-based GWAS designs exist to account for demographic and indirect effects and these have shown that indirect genetic effects can indeed bias MR estimates, although their impact on molecular phenotypes was found to be low [70].

Biased genetic associations can violate MR assumptions.

In Chapter 2, I describe an MVMR framework developed during this thesis work that integrates two molecular traits and a disease outcome with application to DNAm, transcript levels and 50 complex traits. In addition to detecting putative molecular mechanisms, we quantified the role of transcript levels in mediating DNAm-to-complex trait effects.

1.3 Genomics in drug discovery

Studying the path from genotype to phenotype can greatly elucidate our understanding of disease mechanisms and help to pinpoint genes responsible for disease risk. While there is a fundamental aspect to uncovering disease mechanisms, this field has a direct application in drug discovery. Reversing the effect of a "defect" gene could be achieved by a pharmaceutical, and finding the right target or causal gene underlying the disease is key in this process.

In the following, I will briefly summarise the history of drug discovery and then highlight the different roles genetics and statistical genetics can take on in the drug development process and in particular in the early stages of drug target discovery. I will end this section with a technical part on gene prioritisation methods that have emerged over the past years and that complement those presented in Subsection 1.2.3.

1.3.1 Brief history of drug discovery

Drug discovery has been around for much longer than technologies able to probe the genome and some successful drugs still used in the clinics today have emerged over 200 years ago.

Early drug discovery relied on nature's rich sources and serendipitous discoveries.

Morphine isolated from opium in the early 1800s can be considered as the first modern, pharmaceutical medicine as it was the first time an active pharmacological compound has been isolated from natural sources in a pure state [71]. A series of serendipitous discoveries, or accidental discoveries, in the 19th and 20th centuries followed and led to the clinical use of chloral hydrate as a hypnotic (1869, first synthetic drug), the use of lithium as mood disorder treatment although initially tested for the treatment of gout (1896) and the famous discovery of penicillin by Alexander Fleming in 1928 [72]. In the early 20th century, Paul Ehrlich, considered the founder of immunology, introduced the receptor theory according to which receptors, initially called side chains, associated with either cells or more generally located in the bloodstream are able to bind to distinct toxins. His "magic bullet" concept states that drugs selectively target disease-causing agents while sparing healthy tissues [73].

In the mid-20th century, the drug receptor theory evolved into the receptor-occupancy theory which was further refined to account for the intrinsic activity of drugs (i.e., their ability to induce an effect after binding), spare receptors (i.e., maximal drug response can be obtained with less than all receptors occupied) and binding affinities. The distinction of the α and β receptors eventually led to the introduction of propranolol, the first clinically useful β-receptor blocker [74]. In the late 20th century, advances in biochemistry and molecular biology such as X-ray crystallography and NMR greatly enhanced the understanding of the structure and function of proteins and molecules [75]. Recombinant DNA technology played a major role in studying the pharmacology and function of G-protein-coupled receptors (GPCRs), the largest family of membrane-bound receptors that represents targets of a third of FDA-approved drugs [76, 77]. These cell surface receptors can detect chemical signals in a highly selective way and transmit the signal to generate intracellular responses. Subsequently, GPCR or other high-throughput assays were screened for drug interactions where the development of combinatorial libraries containing millions of potential drugs marked another major milestone in the early 1990s [75]. Over time, drug discovery has transitioned from forward pharmacology, or phenotypic screening, where drugs with often unknown molecular mechanisms of action have been selected based on their therapeutic impact, to reverse pharmacology, or target-based screening where proteins identified to play a role in disease are tested for interaction with small molecules or biologicals such as monoclonal antibodies [78].

Technological advances in biology and chemistry revolutionised drug discovery.

The Human Genome project provided the great promise of revolutionising the field and tremendously helping in identifying disease-causing genes, and ultimately new therapeutic targets [75]. Whether biobanks of genome sequences have entirely lived up to this promise is debatable, however, it is undeniable that genomics has played and is likely to play an important part in drug target discovery and beyond.

1.3.2 Genetics for effective drug development

If a genetic variation exists within a population that leaves one group at risk and the other group protected against a disease, then the gene associated with this variation could serve as a potential drug target. This hypothesis was tested in a systematic way for approved drugs and their targets in 2015, and at the time it was found that drug targets are two-fold enriched for genetic support from GWAS which represented 8.2% of the investigated target-indication pairs [79]. Genetic support increased from phase I to the approval stage suggesting a substantial benefit from genetics in terms of efficacy. A replication study on updated data confirmed these findings in 2019 [80]. As genetics becomes integral to drug development pipelines, the statistics brought forward in these studies are bound to change over time. Of the 50 FDA-approved drugs in 2021,

Drugs with genetic support are more likely to be approved.

33 (66%) target a gene that either has direct or indirect (through physical interaction) genetic support for its indication or closely related phenotype [81]. This study was based on genetic evidence from the Open Targets Platform which besides GWAS data integrates several other genetic resources [82].

Genetics provides an anchor to study causality and identify effective drug targets.

High-throughput omics techniques can cover the full protein-coding transcriptome and a significant fraction of the proteome space. Thus, one could argue that finding disease-causing genes could be achieved by bypassing genomics altogether and correlating gene products with disease status. However, correlation does not imply causation, and correlations can also arise because of reverse causation or the presence of a confounder (see Subsection 1.2.3). As a drug should target the underlying cause of a disease to be therapeutically effective, genetics can be helpful in providing this evidence.

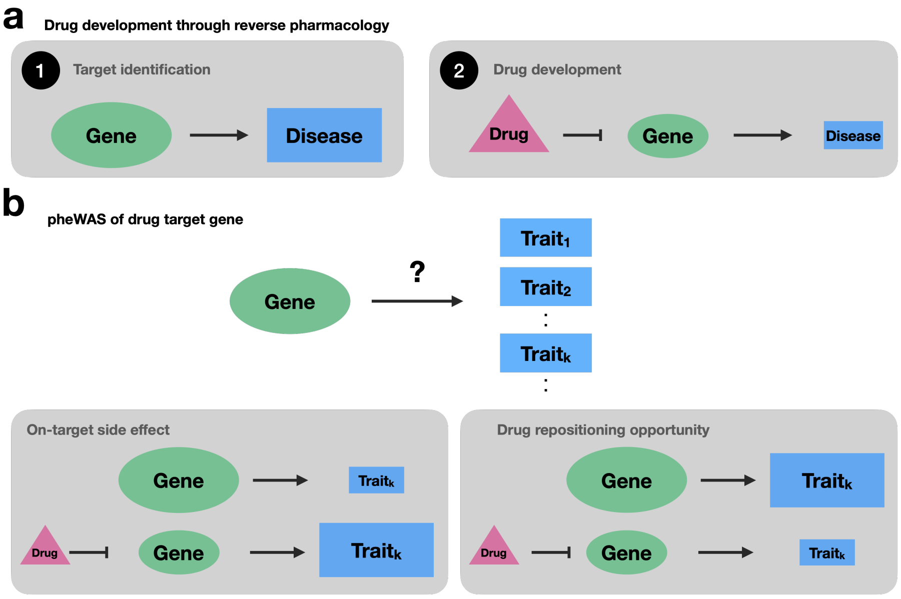


Figure 1.8: The role of gene-disease links in candidate target identification. **a** In reverse pharmacology, first a target is identified that associates with disease risk. In a second step, a drug is developed that interacts with the target through e.g. inhibition, resulting in disease treatment. **b** A phenome-wide association study (pheWAS) of the drug target gene can identify potential side effects as well as drug repositioning opportunities which manifest themselves upon target perturbation.

In all the illustrations, a bigger green gene circle represents a fully functional gene at normal abundance, whereas smaller circles represent gene products with reduced function or lower abundance. Likewise, the size of the blue trait rectangles varies to represent either increased or decreased disease risk.

Identifying robust gene-trait relationships has applications beyond the discovery of efficient drug target candidates. While this could be described as the first step in the process (i.e., screening disease genetics to find candidate genes), the second step would be to expand the analysis to the whole phenome. In so-called phenome-wide association studies (pheWAS), candidate genes can be assessed for associations with other traits, fulfilling two purposes: 1) early-on identification of potential side effects warranting caution for pursuing the target [83], 2) identification of additional conditions associated with the target that could enlarge the drug use spectrum. In the case of drug development, the existence of other conditions would make this target more attractive as future drugs could be marketed for multiple indications. If a drug targeting this gene already exists, this strategy could be employed for drug repositioning (i.e., finding a purpose for a drug other than the one it was originally indicated for) [84]. Importantly, directionality has to be taken into account to properly distinguish between side effects and conditions that could benefit from target perturbation (i.e., inhibiting protein A could make trait X a side effect and trait Y an additional indication, while activating protein A would reverse the roles of X and Y).

pheWAS can be predictive of side effects and multiple indications.

Along the drug development process, genetics can also play a role in identifying who is most at risk of side effects and who would benefit the most from a given medication. Pharmacogenetics deals with the study of inter-individual variability in drug response due to genetics and in Section 1.4, I will cover this topic and present opportunities offered by large-scale biobanks to study drug-gene interactions.

1.3.3 Gene prioritisation methods

While the benefits of leveraging genetics for drug target discovery are obvious, prioritising genes based on genetic evidence is not as trivial. As sample sizes increase, thousands of significant GWAS hits can be identified [85]. Not only does the sheer number of associations call for prioritisation methods, also mapping these associations to genes requires post-processing. In Box 2, I summarise major gene prioritisation methods and in Box 3, miscellaneous methods that were derived based on multiple scores and/or rely on external datasets such as networks.

Gene prioritisation scores resulting from methods listed in Boxes 2 and 3 can vary significantly and may even be completely independent due to the diverse sources used in their derivation. In recent years, more and more scores have emerged that combine different methods and data sources [86, 87, 88, 89]. Although I classified the algorithms into broad categories, there can be ambiguity. For instance, PoPs [88] also integrates network features in the form

Recent Gene prioritisation methods integrate a variety of features.

of a binary variable which indicates whether a gene is its first-degree neighbour or not, and thus constitutes also a network approach. Likewise, the PI [89] also makes use of diverse functional annotations making it a combined approach.

Box 2. Major gene prioritisation methods

GWAS - Fine-mapping & (non -) functional annotations

GWAS fine-mapping approaches allow to pinpoint causal SNPs in a region through methods such as stepwise conditional analysis (e.g. GCTA-COJO [90]) or Bayesian fine-mapping methods that estimate the posterior probability of a vast space of possible causal configurations given the data at hand and choose the one with the highest probability (e.g. FINEMAP [91], SuSiE [92]). The resulting credible set of SNPs can be mapped to genes through various annotations. Non-functional annotations based on distance include the closest TSS or gene body and functional annotations exon [93], promoter and enhancer locations [94, 95, 96]. QTL annotations also work with this strategy with colocalisation approaches (Box 1) being a special case that apply fine-mapping to both GWAS and omics trait. The proposed combined SNP-to-gene strategy (cS2G) weighs and integrates multiple annotations to optimise heritability coverage [86].

GWAS - Gene scoring

GWAS gene scoring methods calculate a test statistic for each gene by aggregating GWAS summary statistics of the SNPs falling into the respective gene region. There are various tools performing gene scoring based on summary statistics and an LD-reference panel such as MAGMA [97] and Pascal [98] which compute a gene-based test statistic based on the sum of squared SNP z-scores (T_{sum}) following a weighted χ_1^2 distribution. Although both methods rely on the same theoretical principle, they differ in how the decomposition of the local genetic correlation matrix is dealt with in the computational implementation. Another method, LDAK-GBAT computes a test statistic based on a restricted maximum likelihood model with the null distribution being approximated via permutation [99].

QTL-GWAS integration

Methods that prioritise genes based on QTL-GWAS integration include colocalisation, MR, SMR-HEIDI and TWAS which are described in Box 1 (Subsection 1.2.3).

Rare variant gene tests

Rare variants obtained from WES or WGS data can be collapsed into gene burden masks which can then be tested for association with the phenotype of interest. Masks are determined by the MAF cut-off and types of included variants (Loss-of-Function (LoF), missense variants). Simple burden tests combine rare variants linearly, assuming all variants are either trait-increasing or trait-lowering. Variance-component tests such as the sequence kernel association test (SKAT) [100] and SKATO (combination of SKAT and burden tests) [101] can account for variants with effects in opposite directions.

Box 3. Miscellaneous gene prioritisation methods

Combination approaches

Several methods can be combined to yield gene scores. For instance, the Polygenic Priority Score (PoPS) combines gene scores (computed by MAGMA [97]) and gene features coming from gene expression data (scRNA-seq), biological pathways and predicted protein-protein interaction (PPI) networks. This method proceeds by estimating the weights of each gene feature through regression of gene scores on the gene features. PoPS is then the linear combination of all features passing a significance selection threshold [88].

Machine learning approaches

Machine learning (ML) approaches combine multiple gene features from different sources with their weights determined through ML models that train on a gold standard dataset. For instance, Open Target uses an XGBoost gradient-boosting classifier to train a 'locus to gene' (L2G) score based on a manually curated set of 445 gold-standard-positive genes at GWAS loci for which there was strong prior knowledge about the causal gene and 9,171 gold-standard-negative genes [87].

Network approaches

Network approaches leverage molecular interactions to propagate initial gene scores either continuous or binary (in the latter case, genes with a non-zero initial value are referred to as 'seed genes') to neighbouring genes. In the simplest case, the closest genes (i.e., first-degree neighbour) can be included in the set of prioritised seed genes, whereas more complex algorithms take into account the full graph topology to diffuse scores through algorithms such as random walks. Examples include the priority index (PI) where seed genes defined from GWAS and functional annotations were propagated on a PPI network through a random walk with restart algorithm [89].

Consulting more than one gene prioritisation approach increases confidence in determining disease genes, and allows one to capitalise on their respective advantages and disadvantages. QTL-GWAS approaches which have been introduced more extensively in Subsection 1.2.3 have the advantage of providing mechanistic insights as to whether altered levels of the gene product increase or decrease disease risk. This information is not readily available from other gene prioritisation approaches. Methods that aggregate multiple scores or sources were shown to be superior to individual methods in validation sets composed of drug targets [89, 87], putative causal genes defined through fine-mapping [88] and exon/promoter information [86]. Yet, they may lack interpretability as to the statistical significance and may not convey the driving feature which could hide potential sources of bias.

Complex gene prioritisation methods can be more accurate, but less transparent.

In Chapter 3, I will present a benchmarking study where we compared various gene prioritisation approaches (GWAS, QTL and Exome-based gene scores alone and in combination with network approaches) and evaluated their ability to identify approved drug targets across thirty clinical traits.

1.4 Pharmacogenomics

While the preceding section dealt with natural genetic variation to identify drug targets, genetic variation can also serve as predictor of drug response. Pharmacogenomics (PGx), often used interchangeably with pharmacogenetics, is the field that studies inter-individual variability in drug response (efficacy and/or safety) due to genetic factors [102]. In the following, I will first explain the notions of pharmacokinetics and pharmacodynamics and then focus on the current state of PGx and its challenges in demonstrating evidence for drug-gene interactions. Finally, I will introduce opportunities offered by large-scale biobanks and present recent studies that have leveraged these immense data resources to study PGx. Some of the elements in this section are taken from a review article that I co-authored with Chiara Auwerx entitled “From pharmacogenetics to pharmaco-omics: Milestones and future directions” which was published in *Human Genetics and Genomics Advances* in 2022 [103].

1.4.1 Pharmacokinetics and pharmacodynamics

PGx variants mostly reside in genes involved in pharmacokinetics and pharmacodynamics as well as in regions related to immune response [104]. Pharmacokinetics can be broadly defined as “what the body does to the drug” and pharmacodynamics as “what the drug does to the body”. Pharmacokinetics evolves around absorption, distribution, metabolism and excretion (ADME) which influences the drug’s concentration and activity in the body over time. On the other hand, pharmacodynamics is concerned with the variability in drug response not related to drug concentration, but to the mechanisms of action of drugs and how they interact with specific receptors or molecular targets (Figure 1.9).

Single variants in key pharmacokinetic genes can have a substantial impact on active drug concentration in two scenarios (Figure 1.9a) [105]. In the first scenario, a prodrug needs to be bioactivated to its active drug metabolite. Variation in the responsible enzymes can lead to toxicity or lower the pharmacologic effect. Examples include the prodrug codeine that is biotransformed to morphine by the CYP2D6 enzyme [106] as well as the antiplatelet drug clopidogrel bioactivated by CYP2C19 where LoF carriers have an increased risk of major adverse cardiovascular events (MACE) and bleeding [107]. In the second scenario, variation in enzymes responsible for the elimination of active drugs with a narrow therapeutic window (i.e., small margin between therapeutic and toxic doses) can cause large PGx effects if a single pathway is

Most pharmacogenetic variants have a pharmacokinetic basis.

responsible for inactivation or drug efflux. For instance, LoF genetic variants in CYP2C9 which is responsible for the metabolic clearance of the anticoagulant drug warfarin can increase the incidence of severe bleeding [108]. Note that most of the clinically actionable pharmacogenetic variants described to date have a pharmacokinetic basis [105].

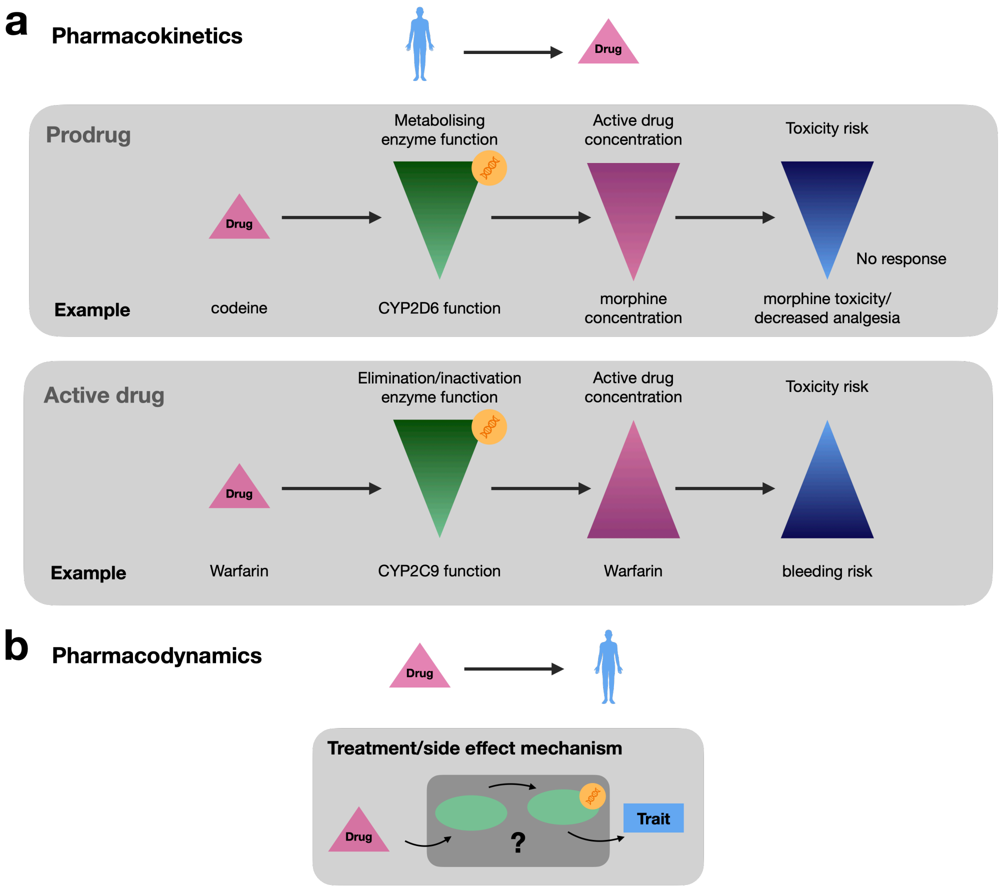


Figure 1.9: Genetic variants that influence pharmacokinetics and pharmacodynamics can impact drug response. **a** Pharmacokinetics studies how an organism affects the drug thereby modulating its concentration. Genetic variants in genes responsible for bioactivation ("prodrug scenario") or elimination/inactivation ("active drug scenario") can lead to toxicity due to high active concentrations or no response in the absence of the active metabolite (adapted from [105]). **b** Pharmacodynamics studies how the drug affects an organism thereby causing molecular and physiological effects. Pharmacodynamics is traditionally studied through drug-response curves and binding affinity experiments. Understanding the molecular basis through which drugs elicit a response can explain treatment and side effect mechanisms as well as inter-individual drug responses due to pharmacodynamic variants.

Pharmacodynamic variants usually reside in genes implicated in disease mechanisms and can imply the drug target gene itself (Figure 1.9b) [105]. For instance, variants in *VKORC1*, which is inhibited by warfarin, and which is involved in clotting factor activation, determine the required maintenance dose, and can also cause warfarin resistance [109]. *VKORC1* is one of the key pharmacogenes and the variant rs9923231C>T which is estimated to account for 15-30% of the variability is highly ancestry-specific (T allele frequency of 39%, 5% and 88% in European, African and South East Asian populations, respectively; `https://www.pharmgkb.org/`, accessed August 2023). Other pharmacodynamic PGx mechanisms include variants in *RYR1* and *CACNA1S* genes that increase the risk of malignant hyperthermia upon exposure to potent volatile anaesthetics [110].

Pharmacodynamic variants usually reside in genes implicated in disease mechanisms.

The field of pharmacogenetics is adopting a standard set of definitions to describe pharmacogenetic phenotypes [111]. Historically, the star allele nomenclature is being used to designate pharmacogenetic alleles, although efforts are being undertaken to harmonise the notation with alternative naming conventions such as 'rsIDs' [112]. In the star allele nomenclature *1 describes a fully functional/wild-type haplotype whereas any other number refers to either a decreased or increased activity based on the presence of a single or a combination of alternative alleles. From the diplotype, a pharmacogenetic phenotype can be determined which in the case of drug-metabolising pharmacokinetic enzymes will be "normal metaboliser" (corresponding to a *1/*1 diplotype or the combination of a normal and decreased function allele), "intermediate metaboliser", "poor metaboliser", "rapid metaboliser" and "ultrarapid metaboliser". For transporters such as SLCO1B1, which facilitates the hepatic uptake of statins, possible phenotypes are "increased", "normal", "decreased" and "poor" function, and for pharmacodynamic genes, phenotype definitions are "positive" if a high-risk allele is detected and "negative" otherwise [111].

Allele functional status and phenotypes are used to describe pharmacogenetic variation.

1.4.2 Current state in pharmacogenomics

As of today, the Pharmacogenomics Knowledgebase (PharmGKB) reports 201 clinical guideline annotations published by the Clinical Pharmacogenetics Implementation Consortium (CPIC) [113], the Dutch Pharmacogenetics Working Group (DPWG) [114], and other professional societies including the Canadian Pharmacogenomics Network for Drug Safety (CPNDS, `https://cpnds.ubc.ca`) and the French National Network of Pharmacogenetics (RNPGx) [115] (`pharmgkb.org`, accessed August 2023). These guidelines may have differing levels of evidence and not all are attributed to a recommendation. Also, the methodology used in scoring the level of evidence may differ between individual working groups [116]. The gold standard in establishing causality between genetic variants and clinical outcomes, and demonstrating the superiority of genotype-guided treatment over the standard of care remain indisputably

Various working groups provide clinical guideline annotations about drug-gene interactions.

RCTs. A number of large-scale RCTs have been conducted of which major trials are summarised in Table 1.2.

Table 1.2: Major randomised control trials conducted to test the benefit of pharmacogenetics testing to guide treatment.

Name	Year (completion)	N	Drug	Gene	Endpoints	Aim	Study result	Reference
PREDICT-1	2006	1,956	abacavir	HLA-B	Hypersensitivity reaction	Assessing the effectiveness of prospective HLA-B*5701 screening to prevent the hypersensitivity reaction to abacavir in HIV patients	HLA-B*5701 screening reduced the risk of hypersensitivity reaction to abacavir	[117]
COAG	2013	1,015	warfarin	CYP2C9, VKORC1	Percentage of time the INR was in the therapeutic range	Assessing whether dosing algorithm that included both clinical variables and genotype data was superior to one that included clinical variables only	Genotype-guided dosing of warfarin did not improve anticoagulation control.	[118]
EU-PACT	2013	455	warfarin	CYP2C9, VKORC1	Percentage of time the INR was in the therapeutic range	Assessing whether genotype-guided warfarin dosing was superior to standard dosing	Genotype-guided dosing was superior than standard dosing during the initiation of warfarin therapy	[119]
GUIDED	2017	1,167	antidepressants	CYP1A2, CYP2C9, CYP2C19, CYP3A4, CYP2B6, CYP2D6, HTR2A, SLC6A4	Symptom improvement, response and remission of depressive symptoms	Assessing whether PGx testing affects antidepressant medication selection and whether such testing leads to better clinical outcomes in patients with MDD	PGx testing did not significantly improve mean symptoms but did significantly improve response and remission rates	[120]
POPular Genetics	2019	2,488	clopidogrel	CYP2C19	MACE after PCI	Assessing the clinical utility of a genotype-guided selection of oral P2Y12 inhibitors after PCI with respect to adverse clinical events	Genotype–guided strategy resulted in lower incidence of thrombotic events and bleeding	[121]
TAILOR-PCI	2019	5,302	clopidogrel	CYP2C19	MACE after PCI	Assessing the clinical utility of a genotype-guided selection of oral P2Y12 inhibitors after PCI with respect to adverse clinical events	No statistically significant reduction in MACE following genotype-guided prescription, but lower risk of bleeding	[122]
U-PGx-PREPARE	2020	6,944	42 drugs	12-gene panel CYP2B6, CYP2C9, CYP2C19, CYP2D6, CYP3A5, DPYD, F5, HLA-B, SLCO1B1, TPMT, UGT1A1, VKORC1	Adverse reactions	Assessing the clinical utility of a pre-emptive genotyping strategy in a real-world situation	PGx-guided prescribing resulted in a 30% reduction of clinically relevant ADRs	[123]
PRIME Care	2021	1,944	antidepressants	CYP1A, CYP2B6, CYP2C19, CYP2C9, CYP3A4, CYP2D6, UGT1A4, UGT2B1, SLC6A4, HTR2A, HLA-A, HLA-B	Proportion of prescriptions with a predicted drug-gene interaction and remission of depressive symptoms	Assessing whether PGx testing affects antidepressant medication selection and whether such testing leads to better clinical outcomes in patients with MDD	PGx testing reduced prescription of medications with predicted drug-gene interactions compared with usual care. Remission rates were not significantly higher.	[124]

While some of the trials showed a clear advantage of a pharmacogenetically adapted treatment strategy, other trials yielded conflicting outcomes. For instance, genotype-guided warfarin dosing resulted in a significant improvement in minimising the risk of bleeding as assessed by the international normalized ratio (INR) in the EU-PACT, but not in the COAG trial (both completed in 2013) [119, 118]. A later trial in 2016, demonstrated the advantage of genotype-guided warfarin strategy using adverse drug reactions (ADRs) as endpoints (major bleeding, INR>4, venous thromboembolism, and death) rather than INR in 1,650 individuals [125]. Patients carrying CYP2C19*2 or *3 LoF variants were found to have an increased risk of ischemic events when treated with antiplatelet medication clopidogrel, however, whereas a first trial could demonstrate the benefits of genotype-guided selection of oral P2Y12 inhibitors after primary percutaneous coronary intervention (PCI) [121], a second trial did not [122]. Conflicting results were also obtained in the GUIDED [120] and PRIME care [124] trials on patients with major depressive disorder (MDD), where the selection of antidepressant medication in the pharmacogenomic-guided group was adapted to have a lower potential of drug-gene interactions. The GUIDED trial found an improvement in symptoms and remission rates whereas the PRIME care trial which was longer in duration did not. PREPARE, a large-scale trial (N = 6,944) across seven European countries which tested the clinical utility of a pre-emptive genotyping strategy through a pharmacogenetic passport across 12 genes, found a 30% reduction in clinically relevant ADRs (in absolute numbers 152 (21%) of 725 in the study group with an actionable test result and 231 (28%) of 833 in the control group experienced an ADR) [123].

Several trials testing the same or a similar PGx interaction reported conflicting results.

Overall, clinical trials in PGx are a challenging undertaking due to several reasons. First, pharmacogenetic variants often have low allele frequencies which requires large sample sizes to test for statistical significance. However, due to high costs and recruitment difficulties, this may not be feasible in practice. Moreover, not all of the genetic variation determining the PGx phenotype (e.g. low metaboliser phenotype) may be known and tested. The PGx phenotype may be more polygenic than initially assumed, thus requiring a more comprehensive assessment of genetic variation to guide treatment. Related, evidence of PGx interactions often stems from studies in European ancestries and an analysis of the UKBB showed that non-European populations carry a higher frequency of predicted deleterious variants not captured by current PGx allele definitions [126]. It was suspected that the failure of the warfarin COAG trial was partly due to the large proportion of African-Americans (27%) who

Major challenges in testing PGx variants in RCTs include their low frequencies, cost and transferability across ancestries.

have a lower frequency of the tested *CYP2C9* variants, but may carry other variants not assessed in the study that influence dosage [127]. Finally, not only PGx variants, but many other factors can influence drug response among which polypharmacy. In fact, the presence of a second drug that inhibits a key enzyme can mimic the effect of a LoF variant. Although studies have been conducted to investigate such drug–drug interactions, or drug-induced phenoconversion, their scope remains limited and the real-world impact of drug-induced phenoconversion remains largely unknown [128, 129]. In addition to comedication, other factors such as sex, age, diet and comorbidities can influence drug metabolism and cause environment-drug-gene interactions [128].

1.4.3 Large-scale biobanks for pharmacogenomic research

While RCTs remain the gold standard in providing clinical evidence, many of the pharmacogenetic clinical guidelines have not been tested in prospective clinical trials and updates occur regularly as new evidence emerges [116]. Large-scale biobanks can play a pivotal role in complementing PGx research and provide new lines of evidence for drug-gene interactions. Despite being of a retrospective nature, deep phenotypic longitudinal data from electronic health records (EHRs) that encompass medical diagnoses, drug prescriptions and laboratory results open up new possibilities to study PGx when coupled with genetic data.

PGx research in biobanks could address issues related to cost, comorbidities, polypharmacy, restricted clinical endpoints and unknown PGx variants.

Biobanks provide the opportunity to verify reported drug-gene interactions in much larger sample sizes and potentially discover novel relationships in a more cost-effective manner [130]. Furthermore, increased sample sizes allow for increased levels of stratification, by considering for instance concomitant medication and co-occurring conditions that could induce PGx interactions. Biobank-based analyses also allow for much longer follow-up times which makes it possible to assess ’hard’ instead of surrogate endpoints [105]. For example, two of the three warfarin trials mentioned previously used the INR test being in the therapeutic range as endpoint (i.e., blood clotting tendency), but harder endpoints such as major bleeding or even death could be of greater value when deciding on clinical guidelines [105]. Importantly, genome-wide genetic data allows for agnostic methods such as GWAS to screen for new PGx variants that may have been missed in earlier candidate gene studies. Large sequencing projects in biobanks, including long-read WGS in the All of Us (AoU) biobank, open up new opportunities to assess not only common, but also rare variants and their association with variable drug response. In Box 4, major biobanks suitable for PGx research are listed.

Box 4. Large-scale biobanks with EHRs

Several large-scale biobank projects exist that are suited for PGx analyses. The **Estonian Biobank** (N $\approx$ 200,000) contains genotype and sequencing (for a subset of individuals) data and contains EHR data for all its participants with detailed information about drug purchase and disease incidences collected since 2000 [9, 131]. The **UK Biobank** (N $\approx$ 500,000) made available genotype and sequencing data (WES and WGS) for all of its participants (as of Q4 2023) and links the data to the primary care records for ~230,000 participants dating back to 1990 (`https://www.ukbiobank.ac.uk`) [8]. A number of biobanks are still recruiting and/or genotyping/sequencing samples. **FinnGen** has nearly completed genotyping of $>$ 500,000 participants who are all linked to their national health registry data including the Finnish drug purchase registry which contains all prescription drug purchases starting from 1995 [11]. The **Million Veteran Program** in the US is recruiting up to a million participants who have their medical data recorded in the Veteran Affairs EHR [10, 132]. Similarly, the **All of Us** research program is recruiting up to a million participants who have their EHRs linked through participating healthcare provider organisations [133]. Genotyping has been completed for ~350,000 and short-read and long-read WGS for ~250,000 and ~1,000 participants, respectively (as of August 2023, `https://www.researchallofus.org`). **BioVU**, Vanderbilt University Medical Center's biobank, is constantly growing through their "opt-out" model that started in 2007 and links individuals to de-identified EHRs [134]. The biobank collects about 500 DNA samples per week totalling over 300,000 biological samples in 2023 (`https://victr.vumc.org`).

Various PGx studies have been undertaken in biobanks that can be classified into three types: 1) Analysis of medication use and relationship with underlying disease, 2) characterisation of PGx variation in single or across genetic ancestries, 3) association of (PGx) variants/PGx phenotypes with drug-related phenotypes (dosage, ADRs). Studies that fall into the first category include GWAS of self-reported medication use in the UKBB and comparison with the underlying disease phenotypes through PRS stratification, comparison of genetic architecture and MR analysis [135]. On a larger scale, GWAS on longitudinal patterns of medication use for cardiometabolic conditions extracted from

EHRs were conducted in FinnGen and meta-analysed with results from the Estonian and UK biobanks [136]. As in the previous study, strong positive genetic correlations were observed for the total number of purchases and underlying disease as well as positive correlations between disease PRS and medication use. Interestingly, the analysis of changing or discontinuing medication also showed a strong relationship with the underlying risk factors. Discontinuation of lipid-lowering medication was associated with variants in the *PCSK9*, *LDLR* and *APOE* loci, but in the opposite direction than low-density lipoprotein cholesterol (LDL)-associated variants. This trend was corroborated by a negative correlation between LDL PRS and the proportion of individuals stopping statin use quickly. A third study conducted in the UKBB and three Scottish cohorts constructed a "dose-decrease" in addition to a drug discontinuation phenotype to replicate known and identify novel drug-gene interactions [137]. While all these studies analysed drug prescriptions which by themselves do not constitute a formal drug response phenotype, the use of a certain medication class rather than another as well as drug discontinuation and changes can potentially proxy the presence of ADRs or sub-optimal drug responses based on genetic variation.

Atypical medication patterns can reveal PGx associations.

The second type of analysis which deals with PGx variation in the general population is motivated by the increased availability of sequencing data in biobanks and a better representation of diverse genetic ancestries. PGx star alleles and their associated phenotypes were analysed across 14 clinically significant genes for all participants in the UK Biobank using genotype and the available WES data at that time (N = 50,000). This analysis revealed notable distinctions among African, East Asian, European, and South Asian populations, such as differences for genes like *VKORC1*, which is the target of warfarin, and *CYP3A5*, which plays a crucial role in the metabolism of tacrolimus — an immunosuppressive agent widely used in kidney transplantation [126]. Analysis of WGS data of 2,240 Estonian Biobank participants identified novel LoF and missense variants in 64 very important pharmacogenes [131]. This study additionally investigated associations between genetic variants and ADRs extracted from EHRs making it also fall within the realm of the final category of PGx studies. Besides replicating known drug-gene-ADR relationships, the authors identified a novel association between *CTNNA3* and myopathy among individuals taking nonsteroidal anti-inflammatory oxicams [131].

Large-scale sequencing data enable the full characterisation of PGx variation.

The third type of study that links phenotypic variation to drug-gene interactions is most likely to identify actionable PGx variants. Whereas the previous study analysed drug-gene-ADR relationships on the variant level, a study in the UKBB correlated PGx metaboliser phenotypes derived from PGx haplotype information with maintenance dose and ADRs extracted from EHRs to identify known and potentially novel PGx interactions[138]. On a larger scale, a study combining data from the Estonian Biobank, UKBB and BioVu, conducted a GWAS on self-reported penicillin allergy — extracted from EHRs — and uncovered an association with the HLA-B*55:01 allele and a missense variant in the *PTPN22* gene, which is associated with other autoimmune diseases [139].

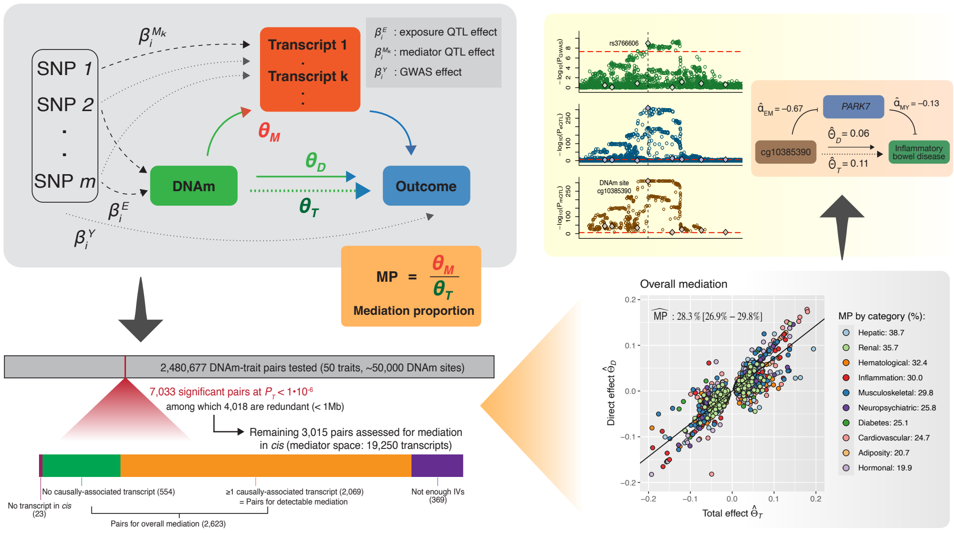
Analysis of phenotypic and medication data in EHRs can identify actionable PGx variants.

While biobank-scale studies have linked genetic variation to ADRs, the integration of longitudinal phenotypic and medication data to screen for genetic predictors of drug efficacy remains underexplored. In Chapter 4, I present a study where we analysed cardiometabolic drug response PGx using EHRs from the UKBB and AoU program. From the records, we extracted baseline and post-treatment measures of LDL, HbA1c, systolic blood pressure (SBP) and heart rate (HR) for statin, metformin and antihypertensive users and conducted GWAS analyses on the biomarker difference. By comparing drug response to baseline genetics as well as the genetics of longitudinal biomarker change in medication-naive individuals, we disentangled disease and medication-specific genetic components.

Chapter 2

Mediation between omics layers and complex traits

In this Chapter, I will give an overview of the omics integration framework that we published in the article “Quantifying the role of transcript levels in mediating DNA methylation effects on complex traits and diseases” in *Nature Communications* (see Appendix A) [140]. In this work, we proposed an MVMR framework to quantify the mediation of DNAm-to-complex trait effects through transcripts. This framework can be generalized to other omics layers and in addition to quantifying mediation proportions, it allows to identify causal pathways that could explain GWAS signals (Figure 2.1).



$\beta_i^{M_k}$
SNP 1
SNP 2
SNP m
Transcript 1
Transcript k
β^E : exposure QTL effect
$\beta_i^{M_k}$: mediator QTL effect
β_i^Y : GWAS effect
θ_M
DNAm
θ_D
θ_T
Outcome
β_i^E
β_i^Y
MP = $\frac{\theta_M}{\theta_T}$
Mediation proportion
2,480,677 DNAm-trait pairs tested (50 traits, ~50,000 DNAm sites)
7,033 significant pairs at $P_T < 1\cdot10^{-6}$
among which 4,018 are redundant (< 1Mb)
Remaining 3,015 pairs assessed for mediation in *cis* (mediator space: 19,250 transcripts)
No causally-associated transcript (554)
≥1 causally-associated transcript (2,069) = Pairs for detectable mediation
Not enough IVs (369)
No transcript in *cis* (23)
Pairs for overall mediation (2,623)
rs3766606
$-\log_{10}(P_{GWAS})$
$-\log_{10}(P_{eQTL})$
$-\log_{10}(P_{mQTL})$
DNAm site cg10385390
$\hat{\alpha}_{EM} = -0.67$
PARK7
$\hat{\alpha}_{MY} = -0.13$
$\hat{\Theta}_D = 0.06$
$\hat{\Theta}_T = 0.11$
cg10385390
Inflammatory bowel disease
Overall mediation
$\widehat{MP}$: 28.3 % [26.9% – 29.8%]
Direct effect $\hat{\Theta}_D$
Total effect $\hat{\Theta}_T$
MP by category (%):
Hepatic: 38.7
Renal: 35.7
Hematological: 32.4
Inflammation: 30.0
Musculoskeletal: 29.8
Neuropsychiatric: 25.8
Diabetes: 25.1
Cardiovascular: 24.7
Adiposity: 20.7
Hormonal: 19.9

Figure 2.1: Summary of the omics mediation study applied to the mediation of DNAm-to-trait effects through transcript levels.

Mediation framework

We developed a three-sample MVMR (3S-MVMR) framework that takes as input an omics exposure (here DNAm site), omics mediators (here transcripts in *cis*, ± 500 kB of the DNAm site) and an output disease or complex trait. Leveraging genetic effect sizes on each of these entities and instrumenting the exposure and mediators by following the MR methodology, we can calculate causal effect estimates from the exposure on the outcome through the mediators. First, we calculated the total effect θ_T of the exposure on the outcome (complex trait) in a univariable MR analysis based on exposure-associated SNPs only. Then, the total effect was dissected into a direct effect θ_D and indirect θ_M effect in an MVMR analysis based on all valid instruments. A mediation proportion (MP) was derived as the proportion of θ_M over θ_T, or through the regression of θ_D on θ_T causal effects over multiple exposure-outcome pairs to increase statistical power. In the latter case, the MP was defined as 1-$\hat{\gamma}$ where $\hat{\gamma}$ is the regression slope. We applied this framework to 50 complex traits and diseases by using the largest publicly available GWAS (N $>$ 320,000). As omics QTLs, we used mQTL data from the GoDMC consortium (N = 32,851) [28], which contains $>$ 170,000 whole blood DNAm sites with at least one significant *cis*-mQTL (P $<$ 1e-6) and *cis*-eQTL data from the eQTLGen consortium (N = 31,684) [31] which includes *cis*-eQTLs for 19,250 transcripts.

Key results

When evaluating 2,623 DNAm-trait pairs with significant total effects ($P_T <$ 1e-6) among the tested 50 complex traits, we observe that at least 28.3% (95% CI: [26.9%–29.8%]) of DNAm-to-trait effects are mediated through transcripts in the *cis*-region. When restricting the analysis to pairs with at least 1 causally-associated transcript (2,069 pairs), a condition fulfilled when a significant MR effect was detected from the DNAm site to the transcript, the $\widehat{MP}$ increased to 37.8% (95% CI: [36.0%-39.5%]. MPs were highest for hepatic and renal biomarkers, and lowest for adiposity-related and hormonal traits. MVMR sensitivity analyses corroborated that these $\widehat{MP}$ estimates were robust and not influenced by outlier or single strong, potentially invalid, IVs (sensitivity tests: conditional F-statistic, heterogeneity Q-statistic, excluding the strongest instrumental variable). We further conducted simulation studies that indicated that these $\widehat{MP}$s were likely lower bounds. Low mediator sample sizes as well as weak exposure- and mediator-associated instruments were shown to result in underestimated $\widehat{MP}$s.

In line with studies that showed a high fraction of positively correlated DNAm-transcript pairs (i.e., presence of DNAm favouring gene expression), we found that DNAm increased transcript levels for ~22,000 of ~47,000 significant DNAm- transcript pairs (46.6%). DNAm sites situated in the gene body were particularly enriched for increasing transcription. Yet, $\widehat{MP}$s were higher when DNAm was decreasing transcript levels.

Besides quantifying mediation through transcript levels, our analysis also brought forward many putative regulatory mechanisms. For instance, we found that methylation of the promoter probe cg10385390 (chr1:8’022’505) increases the risk for inflammatory bowel disease by reducing *PARK7* expression and methylation of cg09070378 (chr1:161’183’762) decreases asthma risk by reducing *FCER1G* expression a gene listed in the KEGG pathway for asthma.

A main limitation of this study was that omics data came from whole blood which is not necessarily the most relevant tissue. Furthermore, we only concentrate on the strongest DNAm site in a region, ignoring nearby methylation sites whose effects may be mediated by transcripts to a different degree. Related to this, we only conducted mediation analyses on DNAm-trait pairs with large, detectable causal effects for which MPs may be different than for pairs with weaker causal links.

Contribution of the author

This study was conceived and designed by Zoltán Kutalik, Eleonora Porcu and myself. Based on omics MR analyses and scripts previously developed by Eleonora Porcu and Kaido Lepik, I added univariable and multivariable MR functionalities into the SMR software (`https://cnsgenomics.com/software/smr` [64] written in C++) which is available at `https://github.com/masadler/smrivw` and allows for fast and parallel computations. I carried out simulations, statistical analyses on real data as well as sensitivity analyses with the help of Eleonora Porcu and Zoltán Kutalik. Interpretation of results and manuscript writing was done by Zoltán Kutalik, Eleonora Porcu and myself. Chiara Auwerx contributed to the interpretation and writing of the biological mechanisms.

Related work

This framework was applied in several collaborations. I conducted mediation analyses through transcript levels with DNAm as exposure in a study investigating the genetics of cancer (“Novel discoveries and enhanced genomic prediction from modelling genetic risk of cancer age-at-onset”, currently under review; preprint available at [141]) as well as longevity (“Causal Epigenetic Age Uncouples Damage and Adaptation”, currently under review; preprint available at [142]). Furthermore, a variation of this framework with transcripts as exposures and metabolites as mediators was published in the article “Exploiting the mediating role of the metabolome to unravel transcript-to-phenotype associations” in *eLife* [143].

Chapter 3

Gene prioritization approaches to identify drug targets

In this Chapter, I will summarise our findings on benchmarking gene prioritisation methods in identifying known drug targets which was published in the article “Multi-layered genetic approaches to identify approved drug targets” in *Cell Genomics* (see Appendix B) [144]. While previous studies have reported that drugs with a genetically informed target have a 2-fold enrichment for being approved [79, 80], these studies solely relied on GWAS data. With the advent of more diverse data sources such as large-scale QTL and WES data, there are new opportunities to establish genetic support. In this study, we quantified enrichment of disease genes determined by various methods with drug targets across 30 clinical traits (Figure 3.1).

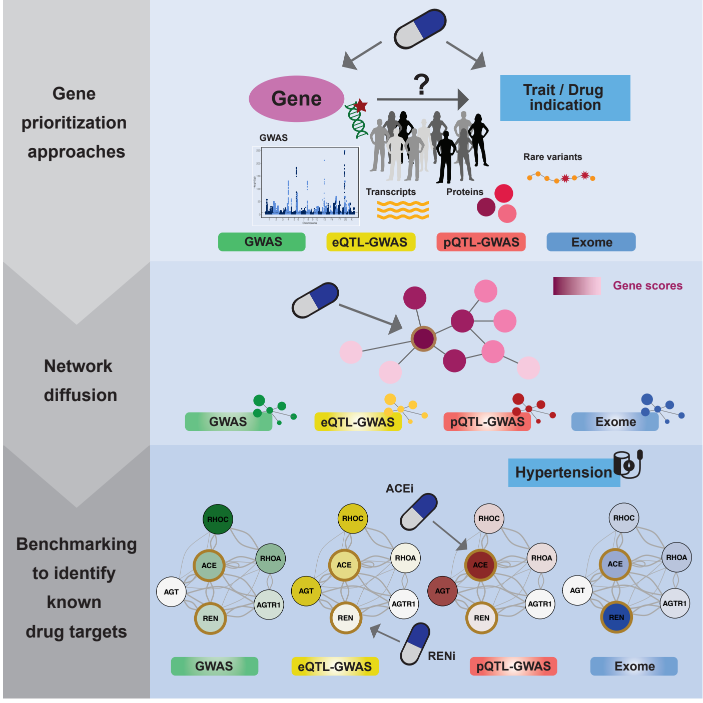


Figure 3.1: Summary of the drug target identification study. Graphical abstract from [144].

Gene prioritisation methods and study design

We considered four main gene prioritisation methods: 1) gene scores computed from GWAS summary statistics through Pascal (*GWAS*, see Subsection 1.3.3), 2) gene scores computed by combining tissue-wide eQTL and GWAS data through MR (*eQTL-GWAS*, see Subsection 1.2.3), 3) MR combining plasma protein QTL with GWAS (*pQTL-GWAS*) and 4) WES burden tests computed in the UKBB (*Exome*, see Subsection 1.3.3). Scores from these four methods served as seed genes for diffusion in three different networks: 1) the STRING PPI network, 2) an RNA-sequencing co-expression network (CoXRNAseq), and (3) an RNA-seq co-expression and proteomics network (FAVA). Network diffusion was based on the Markov random walk algorithm that relies on a restart parameter r determining diffusion strength.

All four methods and their combination with the networks (12x) were tested for enrichment with approved drug targets. Drug targets were defined from several databases that provide drug-indication and drug-target links (ChEMBL, DrugBank, Ruiz *et al.*, DGIdb and STITCH).

Key results

First, we assessed the agreement between pairs of methods and found that genes prioritised by QTL-GWAS and GWAS had a high agreement, whereas agreement with the Exome method was generally low.

Enrichment (OR) for drug targets was 2.17, 2.04, 1.81, and 1.31 for the GWAS, eQTL-GWAS, Exome, and pQTL-GWAS methods, respectively. These main enrichment results were derived based on all testable genes, a number that can differ between methods, and consortia data when available. For instance, the number of testable for the pQTL-GWAS method was ~1,870 so adjustment were made when comparing the pQTL-GWAS to the GWAS method which could test ~19,000 genes. Similarly, the Exome method was restricted to the UKBB, whereas GWAS data is available for disease-specific consortia that provide meta-analyses with much larger case counts. Adjusting for all these differences, we found that GWAS outperformed e/pQTL-GWAS, but not the Exome approach.

Network diffusion relies on a restart parameter r which determines whether the signal is propagated to nearby (high r) or more distant genes (low r). At high r values, initial gene scores are dominant whereas at low r values the network structure becomes more important. At the extreme, r = 0, the gene score is solely determined by the node degree, i.e., the number of neighbours a gene has in the network. We found that diffusion on the STRING PPI network significantly boosted performance, to the extent where the network degree was the best predictor (OR = 8.7). Same results were obtained with the area under the receiver operating characteristic curve (AUC) metric which increased from 54.3% at no diffusion to 77.6% at r = 0 for the GWAS method. Conversely, improvements were modest, although statistically significant, when diffusing on the co-expression networks. For the GWAS method, AUC increased to 54.9% and 55.9% at r = 0.6 in the CoXRNAseq and FAVA networks, respectively.

This massive improvement in the STRING network is due to its biased network structure. Drug targets are more extensively studied than non-drug targets, and hence more interactions are reported for these proteins. Indeed, an analysis of the network connectivity revealed that drug target genes were much more likely to be hub genes (mean log-degree = 13.0 *vs* 12.3, P_{diff} = 6.6e-284).

To conclude, this analysis demonstrated the usefulness of integrating various data sources such as QTLs to gain mechanistic insights, sequencing data to assess rare variants, GWAS when molecular QTL signals are absent, and network propagation to harness gene-gene interactions.

Contribution of the author

This study was conceived and designed by Zoltán Kutalik and myself. With the help of Chiara Auwerx, I extracted drug-indication and drug-target data. Network data, GWAS, QTL and WES summary statistics were retrieved from various data sources which I used to compute gene prioritisation scores and calculate enrichments. Patrick Deelen provided guidance and Zoltán Kutalik supervised all statistical analyses. Zoltán Kutalik and I drafted the manuscript and all the authors contributed by providing advice on interpretation of results and feedback on the final manuscript.

Related work

A current collaboration with the co-author Patrick Deelen is developing a new gene prioritisation method that integrates GWAS scores with regulatory networks. I am contributing by testing its performance in identifying drug target genes and comparing it to GWAS scores alone. A preprint “Linking common and rare disease genetics through gene regulatory networks” is available at [145].

Chapter 4

Drug response pharmacogenomics using EHRs from biobanks

In this Chapter, I will summarise our results on cardiometabolic drug response pharmacogenetics using EHRs from biobanks (manuscript in preparation, see Appendix C). Given the richness of longitudinal data on medication prescriptions and clinical measures found in the UKBB and AoU research program, we assessed the analysis potential of these resources to find genetic predictors of drug efficacy and compared drug response genetics to disease and disease progression genetics in medication-naive individuals (Figure 4.1).

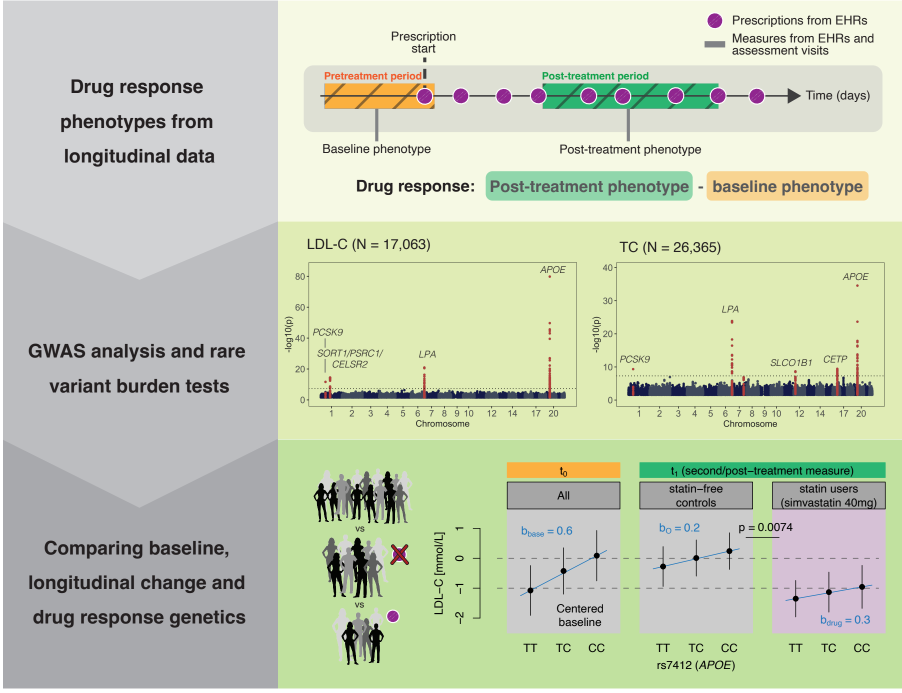


Figure 4.1: Summary of the drug response pharmacogenomics study. Based on longitudinal data from EHRs ten cardiometabolic drug response phenotypes were defined and tested for genetic associations.

Defining drug response phenotypes from biobank data

We constructed drug response cohorts by extracting prescriptions and clinical measures from EHRs focusing on ten cardiometabolic medication-phenotype pairs: statin-lipids (LDL, high-density lipoprotein cholesterol (HDL), total cholesterol (TC)), metformin-HbA1c, antihypertensive-SBP (SBP; by antihypertensive class (ACEi, CCB, thiazide diuretics) and all classes combined), beta blocker-SBP and beta blocker-HR. Participants were part of a PGx cohort if a phenotype measurement was available before (baseline) and after (post-treatment) drug initiation. Several filtering steps were applied to ensure consistent drug adherence throughout the study period (i.e., no treatment changes and regular prescriptions) and to make sure that the first prescription corresponded to treatment initiation.

Key results

In the discovery GWAS conducted in the UKBB, we identified 14 independent signals to influence drug response, all of which were from the lipid response to statin GWAS (N = 17,063-26,365). Of these 14 signals, 7 replicated in the AoU at the Bonferroni-corrected replication threshold of 0.05/14 = 0.00357 and 10 at a nominal p-value of 0.05 (all directionally concordant). Among replicated signals, *PCSK9* was identified as a novel genetic determinant of LDL cholesterol response to statins. No genetic variant passed genome-wide significance level in the HbA1c response to metformin, SBP response to antihypertensives and HR response to beta blocker GWAS, likely because of lower sample sizes (N = 780-6,199).

We further extracted genome-wide significant signals (p-value $<$ 5e-8) from the literature to assess whether GWAS derived from EHRs are coherent with those identified in RCT and observational studies of dedicated PGx cohorts. *APOE*, *LPA*, and *SORT1* reported in earlier studies to influence LDL response to statin passed genome-wide significance in the UKBB analyses, whereas *SLCO1B1* only passed that threshold in the TC response analyses for which sample sizes were larger (26,365 vs 17,063). We could not find evidence for a fifth LDL-related locus *ABCG2*, but replicated a genome-wide significant signal at *CETP* which was found to influence HDL response. None of the loci reported to influence HbA1c response to metformin could be replicated (p-value $>$ 0.05) which could be due to lower statistical power, or be a true biological absence aligned with other GWAS studies that failed to replicate them. Overall, concordance with cohort-derived drug efficacy loci was very high.

Furthermore, we tried to address an open question as to whether rare variants play a bigger role in drug response phenotypes compared to common ones. To this end, we conducted rare variant burden tests based on WES in the UKBB and WGS data in the AoU. Only two genes, *PCSK9* for LDL and *ABCA1* for HDL response to statins survived multiple-testing correction suggesting that rare variants only have a modest impact on drug efficacy. In the

AoU, both genes replicated.

Finally, we compared drug response genetics to baseline and longitudinal change genetics in medication-naive individuals. Longitudinal change analyses were conducted in individuals part of the primary care data that did not have any drug prescription indicated for the investigated disease/surrogate end point and who had two available measures equally spaced as baseline and post-treatment measures. We found strong similarities between drug response and longitudinal change genetics with 7 out of 14 signals being general prognostic (disease-specific) and not drug-specific genetic markers. Furthermore, we demonstrated that for same baseline levels, individuals with a higher PRS tended to have reduced treatment efficacy.

Contribution of the author

This study was conceived and designed by Zoltán Kutalik and myself. I performed statistical analyses in the UK Biobank and Alexander Apostolov conducted replication analyses in the All of Us research program under Russ Altman's and my supervision. Diogo Ribeiro provided guidance on analyzing rare variants from sequencing data and Zoltán Kutalik supervised all statistical analyses. The manuscript in its current state was drafted by Zoltán Kutalik and myself.

Related work

This study started during my stay as a visiting researcher at Stanford University in Russ Altman's research group. Prior conducting efficacy PGx analyses, I worked on calling PGx haplotypes and metaboliser phenotypes in the AoU biobank. This work was later taken up by Alexander Apostolov (co-author of this study) who systematically assessed differences in PGx haplotypes across diverse populations for 12 important pharmacogenes in the AoU. During my stay at Stanford, I also engaged in a collaboration within the research group where I characterized DNA methylation profiles for 10 *CYP* genes. This work was published in a review article “Promises and challenges in pharmacoepigenetics” in *Cambridge Prisms: Precision Medicine* [146].

Chapter 5

Discussion

In this thesis, our contributions to the integrative modelling of drugs, omics, and diseases covered three aspects: i) development of a framework for detecting causal molecular chains, ii) comparison of the overlap between drug targets and disease genes prioritised by various methods, iii) identification of genetic predictors of drug response by leveraging EHRs. Figure 5.1 illustrates how these themes connect to constitute the pharmacokinetics and pharmacodynamics of a drug and how improved modelling of drug and disease mechanisms on the molecular level can improve our understanding of treatment mechanisms and emergence of side effects. Further on, genetic variations within each of the genes involved in these processes can potentially lead to deviations from expected drug responses, influencing treatment efficacy, or predisposing individuals to ADRs.

In the following sections, I will summarise limitations encountered during this research, discuss future directions to enhance proposed models, and conclude with some final remarks.

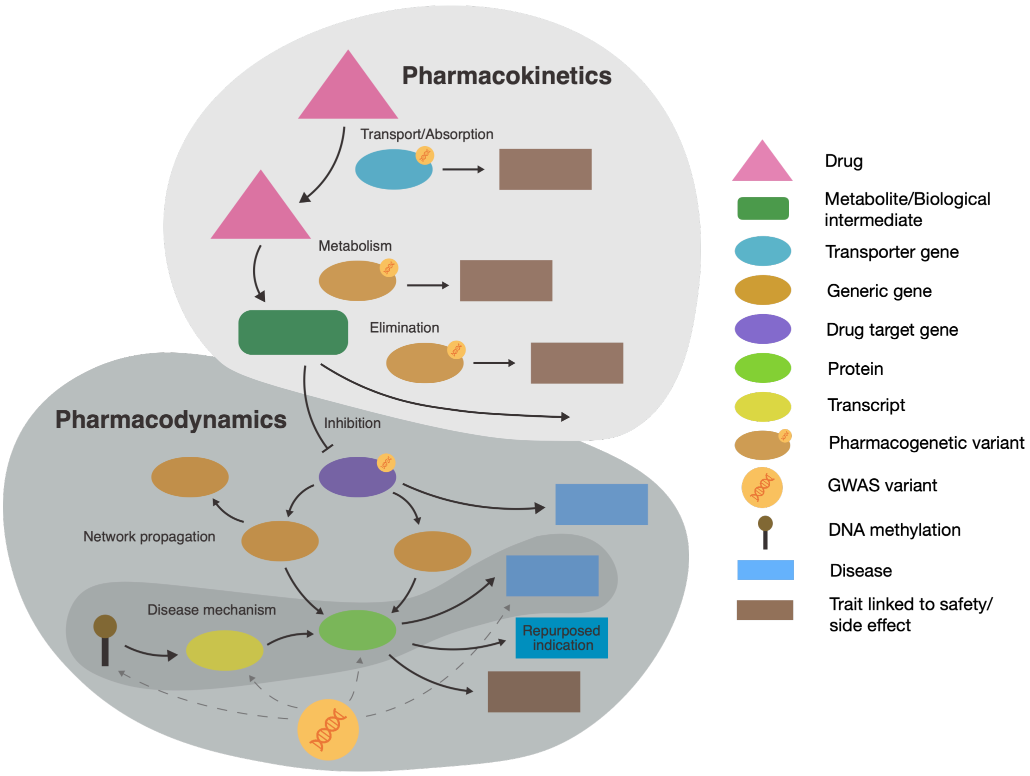


Figure 5.1: Schematic depicting a generalised view of the genetics underlying the pharmacokinetics and pharmacodynamics of a drug. Genetic variation in genes linked to absorption, diffusion (not shown here), metabolism of the drug to its active ingredient and ultimate elimination can modulate the concentration of the drug within the body and and compromise treatment efficacy and/or lead to side effects (Chapter 4).
Once the active drug metabolite has reached the tissue of action, the most common mechanism includes inhibition of its primary target which either has a direct treatment effect or indirect effect on the trait through molecular interactions (Chapter 3). A pharmaceutical can treat multiple diseases, but also cause side effects, either on-target side effects (not shown here) or indirect side effects through molecular interactions.
Identifying disease genes that could serve as drug targets can be done through genetically informed methods that allow for causal inference. Mendelian randomisation methods that instrument molecular and disease traits with genetic variants enable the identification of disease mechanisms across multiple omics layers (Chapter 2).

Unless explicitly specified, a gene refers to both transcript and protein gene products.

5.1 Data as the major limitation

The major limitation we encountered during this research was the lack of data on several levels. In the following, I will outline data limitations that currently hinder the development of comprehensive integrative drug-omics-disease models. The discussion will be based on the research results presented in Chapters 2-4 across the topics of systems genetics, drug target identification and pharmacogenomics.

5.1.1 QTL data

In Chapter 2, I presented an MVMR model to integrate two omics layers with a complex trait to find molecular mechanisms underlying diseases. Through an overall quantification, we estimated that about 28% of DNAm-to-trait effects are mediated through transcripts in the *cis*-region. This proportion leaves many of the DNAm-trait effects unexplained by regulatory gene expression effects, in line with an earlier study that estimated that only 11% of disease heritability is mediated by differential gene expression levels [147]. A recent study assessing mQTLs across tissues corroborated a lack of expression mediation by observing that mQTL-GWAS colocalisations were often missing a corresponding eQTL-GWAS link (i.e., 749 of 1,505 DNAm-trait pairs had evidence of a transcript-trait colocalisation) [29]. These consistent findings across studies suggest that while gene expression may indeed mediate genetic effects to a lesser extent than expected because of other mediators such as protein-coding changes, splicing, chromatin state or expression levels in *trans*, it is also likely that current assays do not capture the full picture of gene expression. In addition to technical limitations, especially for microarray-based technologies (see Subsection 1.2.2) and limited statistical power due to low sample sizes (see simulations in Chapter 2), steady-state expression levels measured in bulk assayed tissues may miss the causal cell type and state as well as other environmental contexts [148]. Indeed, expression changes induced by stimuli such as activation, hypoxic stress and drug treatment could capture relevant environmental factors [148]. In the future, scRNA-seq across tissues might be able to capture causal cell types and could help in finding context-dependent eQTLs, which would ultimately increase the proportion of disease heritability and/or DNAm-trait effects mediated by expression. Practically speaking, availability of post-mortem tissues from the causal organ is likely to remain limited, and blood cell type proxies will probably continue to be the major contributor to such studies. While the space of measuring expression levels across cells and time points grows exponentially, mQTL data could help in prioritising experiments. Compared to eQTL, mQTL effects are more likely to be shared across tissues and be stable across time [29]. Thus, observing an mQTL-trait effect with a missing eQTL-trait link could call for further transcriptomics experiments.

Steady-state expression levels measured in bulk may miss the causal cell type and state.

When comparing gene prioritisation methods, our results showed that QTL-GWAS approaches were not as performant in identifying approved drug target genes as the GWAS approach. Compared to GWAS alone, QTL-GWAS methods provide mechanistic insights which can be very valuable to understand disease aetiology and guide treatment strategies in terms of activating or inhibiting gene products. Again, context-specific QTL-data could fill the gap and explain mechanisms behind GWAS-prioritised disease genes currently lacking QTL evidence.

Unlike eQTL data, pQTL data are only about to take off. While sample sizes are now reaching those of large-scale eQTL datasets, there are still technical limitations in measuring the entire protein-coding space at scale (see Subsection 1.2.2, Figure 1.4). Besides technical issues linked to low specificities in affinity-based methods [46], proteins are downstream of transcripts and thus are expected to have attenuated genetic effects [64]. Polygenicity proxied by the number of independent SNPs was found to be higher for proteins than transcripts (Chapter 3) which points towards a stronger negative selection on the protein than transcript level and leads to a genetic architecture closer

pQTL data are not yet capturing the same breadth as eQTL data.

to that of complex traits and common diseases where critical regions are deprived from strong common-variant associations [149]. We attempted to conduct mediation analyses through protein levels, however, the largest pQTL dataset available at the time was too low in terms of sample size and number of available entities [34].

There was no significant difference in performance between eQTL-GWAS and pQTL-GWAS integration methods in identifying approved drug targets, provided we restricted the comparison to the testable protein space. However, far fewer proteins than transcripts could be tested, and for the same sample size and same set of genes (~4,700 genes from the SomaScan platform), more genes have associated eQTLs than pQTLs. Collectively, this suggests that larger sample sizes are required to detect pQTLs when compared to eQTLs and omics-wide applications will need to wait for the technology to cover a larger protein space.

5.1.2 Drug data

A major challenge in establishing and validating drug-omics-disease models is obtaining a reliable and comprehensive data source for the drug-omics dimension.

Drug mechanism of action data is scarce and not always publicly available.

In Chapter 4, we used drug-protein target data as gold standards to benchmark statistical methods in their ability to identify disease genes. However, there are many cases where drug-protein interactions are not clearly defined and where disagreements or changes in the accepted drug target occur over time [150]. Furthermore, drugs can bind to additional proteins that may not be efficacy targets. These interacting proteins are generally less well-documented, but may be of great importance when studying off-target side effects [150]. An additional layer of complexity is added when considering the drug-indication dimension. Not all drugs are disease-modifying with some being symptom-managing drugs. Such a classification is not readily available, but could be of immense value when studying treatment mechanisms, as disease genetics may be of less importance to symptom-managing drugs [151]. Unfortunately, drug-indication and drug-protein interaction data often sit behind paywalls, examples being Citeline Pharmaprojects (`https://www.citeline.com`) and DrugBank (`https://www.drugbank.com/`) which includes a free and commercial version that additionally provides side-effect data and neatly formatted tables (as opposed to a single XML file that requires post-parsing steps). Fortunately, publicly available databases such as ChEMBL (https://www.ebi.ac.uk/chembl/) and Open Targets (https://www.opentargets.org) are constantly expanding available information on drugs, their targets and indications. It is important to note that current drug targets are largely biased towards GPCR proteins [76], which may influence the validation of putative gene-trait links with drug target data. In the future, drug targets may become more diverse as new therapeutic agents enter the market such as double-stranded RNA-mediated interference (RNAi) and antisense oligonucleotides (ASOs) that can suppress gene expression [152].

While the majority of drugs exert their therapeutic effect by inhibiting proteins, pharmacological effects through gene expression modulation and direct RNA binding have also been reported [153, 154]. However, systematic data of drug regulatory effects in humans are lacking. The Connectivity Map (CMap) is a library that contains the expression profile of 1,000 genes upon perturbation by ~20,000 small molecules across multiple, mostly cancer, cell lines [155]. However, the quality of this data resource has been called into question, as the reproducibility of transcriptomic signatures both between CMap and within CMap versions was found to be very low [156]. In our research analyses, we could not find an enrichment of disease genes for drug-perturbed genes prioritised by CMap (data not shown) which could reflect a true absence of regulatory pathways underlying treatment mechanisms or be the result of low signal-to-noise ratio within the database. On the other hand, regulatory pharmacological effects could rather be responsible for off-target effects [154]. Ideally, omics-level data before and after drug initiation in humans would be available to relate differential transcript and/or protein levels to disease mechanisms. Such longitudinal omics data could stem from clinical trials, but also from biobanks where large sample sizes makes it possible to stratify participants according to their medication regimen.

Drug perturbation data on expression levels in humans could elucidate pharmacological mechanisms.

5.1.3 Networks

Gene co-regulation and physical interactions between proteins can result in concerted gene effects that can be mathematically described by networks. In Chapter 3, we demonstrate the benefits of leveraging molecular networks to identify drug targets as cascading pharmacological effects may mediate treatment mechanisms. Throughout the residence time of a drug in a system, molecular interactions govern ADME and therapeutic processes (Figure 5.1). Drug → gene → trait or drug → gene → side effects may be direct or mediated through network interactions.

Comprehensive, unbiased networks could identify key molecular interactions in drug and disease mechanisms.

When modelling disease network mechanisms in Chapter 3, we consulted the STRING PPI network and two co-expression networks. Bias in the STRING network caused by drug targets being well-studied and having more reported interactions artificially inflated performance. Conversely, improvement in identifying drug targets was limited when using co-expression networks stemming from high-throughput experiments. Although derived from an unbiased data source, such networks may be more noisy due to technical issues. Furthermore, co-expression networks do not necessarily capture physical protein interactions and they may also be tissue- or cell type-specific. Given a strong incentive to integrate networks, there is a need for high-quality, unbiased networks that connect molecular traits within or even across different omics layers.

5.1.4 Rare variants

Sample sizes of current sequencing data do not yet match those of genotyping microarrays.

With the advent of large-scale sequencing data such as WES and WGS in the UKBB and AoU program [13, 133, 157], rare-variant analyses start to gain traction. In Chapter 3 and 4, we compared the effect of common and rare variants on drug-target and drug-response predictions, respectively. While rare-variant burden tests on WES were equally good as gene scores from GWAS data in predicting drug targets, this was only the case when restricting GWAS to UKBB data to match case/control count with the WES data. Using consortia data, GWAS achieved a higher performance. Compared to disease genetics, the impact of rare variants in pharmacogenetics is even less studied. In Chapter 4, we conducted rare-variant burden tests on drug-response phenotypes with only two genes surviving multiple testing. In both studies, a larger sample size would be needed to assess the full impact of rare variation. We found that Exome-prioritised and GWAS-prioritised genes usually point towards different drug targets, highlighting the value in consulting both data sources. Indeed, WES burden heritability was found to be strongly concentrated in constrained genes, in line with the paradigm that ’flattening’ due to negative selective deprives important genes from harbouring common variants which therefore may be missed by GWAS [149]. Although, WES data proves promising to identify complementary disease genes/drug targets to those identified by GWAS, a recent analysis of stopped clinical trials shows that trials are more likely to stop for safety reasons if the drug target gene is highly constrained [158]. Future large-scale WES and WGS data will likely unlock the full potential of rare variants in drug target discovery and demonstrate their importance in PGx.

5.1.5 Electronic health records

EHRs coupled to biobanks greatly enlarge the phenotypic space through longitudinal medication prescriptions/purchases, physical and biochemical measures and diagnosis information. However, EHRs have often been set up for billing and not scientific research purposes which can impact reported phenotype codes. In Chapter 4, we analysed EHRs from the UKBB and AoU to extract drug response phenotypes and encountered several challenges related to data quality and quantity:

- Missing data: EHRs are only available for 45% of participants from the UKBB, although the data exists for remaining participants and was briefly made available for Covid-related research (plans are underway to make the data again available for the entire cohort). In the AoU, only records from participating EHR sites are included which means that not every clinical and medication record figures in the EHRs.

- Data harmonisation: In the UKBB, EHRs come from four different data providers (England (Vision), Scotland, England (TPP) and Wales) that use different clinical and prescription codes (British National Formulary (BNF), National Health Service (NHS) dictionary of medicines and devices (DM+D), Read V2 and Clinical Terms Version 3 (CTV3)). This heterogeneity in reporting greatly complicates data harmonisation, and can even be an error-prone task for researchers unfamiliar with the structure. Furthermore, not all systems provide the same level of detail which can result in inaccurate or incomplete phenotype definitions. Since 2018, SNOMED CT, a structured clinical vocabulary, was introduced by the NHS which may be available in future primary care data releases (records in the current release end in 2016-2017). A harmonised system has already been adopted in the AoU, where all participant data are transformed into Observational Medical Outcomes Partnership (OMOP) standard vocabulary (e.g. SNOMED for conditions and physical measurements and RxNORM for drugs) which enormously facilitates phenotype and medication data retrieval.

- Incomplete records: Both in the UKBB and AoU, entries are often incomplete. For instance, a cholesterol record may miss the measured value as well as its unit (e.g. mmol/L or mg/dL). While the unit can often be inferred from the value, missing values are hard/impossible to impute. Incomplete medication records can also limit analysis potential. A medication entry by itself is useful information, but even more so when the prescription dosage is available, especially when considering drug response phenotypes. Through natural language processing, we extracted medication dose from the name, and if the quantity was available (number of pills and packages prescribed), dosage can be derived by integrating prescription frequency. While dose and quantity information was available for $>$ 95% of the assessed prescriptions in the UKBB, this was only the case for ~70% of the prescriptions in the AoU. Related, reasons for stopping or changing a medication are often not reported, but would be of great value to study side effects and drug efficacy.

Missing and incomplete data resulted in a massive drop in sample size when deriving drug response phenotypes. As an example, of the ~65,000 participants with a statin prescription in the UKBB primary care data, 63% could not be considered for the LDL-response analysis because of missing baseline and/or post-treatment measures. Numbers were similar in the AoU biobank. While overall massive sample sizes can compensate for these data losses, a focus on complete longitudinal phenotyping in future biobanks could immensely increase their value and analysis potential.

5.1.6 Cohort diversity

Population genetics data is largely biased towards European ancestry with approximately 80% of the participants in the GWAS Catalog being of European descent, despite this group constituting only about 16% of the global population (statistics from 2019) [159]. This lack of diversity has important downstream implications on the research and clinical applications of genetics. Allele frequency can largely differ between populations as do LD patterns. Increased natural variation could be of great benefit to identify disease genes as biologically important genes may be missed if the frequency of associated variants is too low in the studied population. Furthermore, differing LD patterns could improve our fine-mapping abilities under the assumption that non-causal variants have differing effect sizes across populations which can make it easier to identify the causal SNP. Deriving PRS based on causal and not merely correlated SNPs is a much needed step to improve their performance and transferability across ancestries, and hence their wide-spread clinical use [159]. As research on the clinical applicability of PGx advances, it becomes crucial to conduct large-scale PGx studies within diverse populations and assess the full spectrum of PGx variants. A study conducted in the UKBB on 14 pharmacogenes revealed that non-European populations carry a higher frequency of variants predicted to be functionally deleterious than individuals of European descent of which many are not captured by current PGx allele definitions [126]. Thus, both the introduction of PRS and PGx-guided prescribing in

Cohort diversity is needed to foster genetic discoveries and ensure equitable benefits from PRS and PGx passports in clinical use.

the clinics could disproportionally benefit individuals of European descent and exacerbate health disparities [159]. The AoU program is actively addressing the issue by encouraging historically understudied populations to participate and is becoming one of the most diverse biobank. As the data emerge, so will most likely the methods to conduct multi-ancestry genetic studies.

5.2 Future work

While currently available data do not contain the molecular interactions necessary to model the complete scope of drug effects, there remain opportunities for improving existing models and exploring unknowns with the data at hand.

Improved gene prioritisation scores In Chapter 3, we compared gene prioritisation methods in their ability to identify drug targets. One of the conclusion was that the GWAS and Exome methods performed equally well while also prioritising different drug targets. A logical next step would be to combine scores across methods to get an optimal consensus gene score that would result in the best performance. Furthermore, in the introductory Subsection 1.2.3, I presented different QTL-GWAS approaches that allow the computation of gene scores with MR being associated to higher false positive rates than colocalisation methods. In our benchmarking study we only considered MR-IVW as the QTL-GWAS method and it is possible that other QTL-GWAS integration methods yield higher performances.

Drug repositioning Once confident gene-trait relationships have been established, candidate drug target genes can be identified and further assessed for potential side effects and multiple indications in a pheWAS (Figure 1.8). In Chapter 3, we created a library of disease-related genes and compared them to existing drug targets. These relationships could be effectively utilised for drug repositioning by identifying diseases with which drug targets are associated beyond their approved indications. Ideally, this analysis is conducted phenome-wide to include a maximum of conditions and also side effects. To evaluate drug candidates and their repurposed therapeutic effects *in silico*, one could additionally consult EHRs and verify whether individuals taking these drugs have these repurposed indications (or proxies thereof) measured and whether they associate with improved outcomes compared to matched controls [160, 161].

Regulatory drug effects Not all drugs have their molecular mechanism well understood, which is for instance the case for metformin [162]. Large-scale perturbation studies like CMap have attempted to address this issue, however, concerns about their quality have been raised (see Subsection 5.1.2). Nonetheless, drugs can exert regulatory effects, for instance as a consequence of protein or RNA binding, i.e., a protein/transcript whose function is inhibited by a small molecule can interrupt downstream molecular pathways and cause regulatory changes [154]. Indeed, a recent study showed that drugs pervasively interact with the human transcriptome and suggested that RNA off-targets may contribute to toxicity [154]. While we conducted preliminary analyses on drug regulatory effects using CMap, more systematic analyses are needed to assess the potential of this resource. For instance, one could calculate the enrichment of drug-induced CMap gene signatures with network-diffused drug target signals to test the hypothesis that drug-target binding influences abundance of interacting genes which, if affirmative, could increase confidence in reported gene signatures. Ultimately, one could test whether drug regulatory effects are more likely to induce side effects rather than contribute to treatment mechanisms.

PGx of drug targets In Chapter 4, we screened the genome for predictors of drug efficacy for selected drug-indication pairs. While statistical power was certainly limiting the ability of finding all relevant genes, our study as well as similar studies in the field, did not identify drug targets themselves as top hits. *APOE* was the top signal in the LDL-response to statin GWAS whereas the statin target *HMGCR* did not reach genome-wide significance (p-value $>$ 1e-6). An analysis of GPCR drug targets found a wide spectrum of natural variation within functional regions such as drug- and effector-binding sites [77]. However, it is less clear to what extent drug target variation influences drug response and whether it is sensible to enrich (early-phase) RCTs with patients carrying variants within the target region under the assumption that they are predisposed to respond better [152]. By extracting drug response data from EHRs similar to the study design in Chapter 4, patients could be stratified by their target genetic variation, both on regulatory (most likely common) and coding (most likely rare) variants, to assess the effect on drug response and gain new insights into the PGx of drug targets.

Polypharmacy A major issue in assessing the genetic basis of inter-individual drug efficacy and safety is polypharmacy. A pooled analysis of 106 studies across the globe (59 from Europe) identified polypharmacy, defined as the concurrent prescription of five or more drugs, to have a prevalence of 45% in individuals aged $\geq$ 65 years with polypharmacy increasing with age [163]. A drug can have a similar effect as a LoF variant (i.e., drug-induced phenoconversion, see Subsection 1.4.2), and accounting for multiple drugs could reveal drug-drug-gene interactions. Cardiometabolic drug response studies such as the ones conducted in Chapter 4 could be further refined to account for concomitant medication (e.g. concomitant antidiabetic and antilipemic prescriptions) to identify effects of polypharmacy on the genetics of drug response.

5.3 Conclusion

It has been an extraordinary journey uncovering the genetics of omics and diseases for drug-target and drug-response predictions. While the vast amount of data within biobanks and beyond holds immense potential for drug development and personalised medicine, analysing and modelling the data made me slowly, but surely realise that this field is profoundly complex, and that we are still only scratching the surface of molecular mechanisms underlying disease and therapeutic effects. While data scarcity has certainly been the limit in many perspectives, the multidisciplinary nature of this field means that genetics may only be able to solve a small part of the problem.

The increase in success rate for drugs with a genetically informed target is undeniable, but in terms of personalised medicine, genetics may play a modest role in patient stratification compared to other clinical factors. Studying the role of omics not only as a mediator of genetics, but also as a biomarker of disease status and surrogate of environmental components such as lifestyle, diet and concomitant medication could refine our understanding of treatment efficacy and safety, and ultimately lead to better informed treatment strategies.

Bibliography

[1] Eric S Lander, Lauren M Linton, Bruce Birren, Chad Nusbaum, Michael C Zody, Jennifer Baldwin, Keri Devon, Ken Dewar, Michael Doyle, William FitzHugh, et al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, 2001.

[2] J Craig Venter, Mark D Adams, Eugene W Myers, Peter W Li, Richard J Mural, Granger G Sutton, Hamilton O Smith, Mark Yandell, Cheryl A Evans, Robert A Holt, et al. The sequence of the human genome. *Science*, 291(5507):1304–1351, 2001.

[3] Bruce Alberts, Dennis Bray, Karen Hopkin, Alexander D Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. *Essential cell biology*. Garland Science, 2015.

[4] Francis Crick. Central dogma of molecular biology. *Nature*, 227(5258):561–563, 1970.

[5] Omer Weissbrod, Farhad Hormozdiari, Christian Benner, Ran Cui, Jacob Ulirsch, Steven Gazal, Armin P Schoech, Bryce Van De Geijn, Yakir Reshef, Carla Márquez-Luna, et al. Functionally informed fine-mapping and polygenic localization of complex trait heritability. *Nature genetics*, 52(12):1355–1363, 2020.

[6] Peter M Visscher, Matthew A Brown, Mark I McCarthy, and Jian Yang. Five years of GWAS discovery. *The American Journal of Human Genetics*, 90(1):7–24, 2012.

[7] Elliot Sollis, Abayomi Mosaku, Ala Abid, Annalisa Buniello, Maria Cerezo, Laurent Gil, Tudor Groza, Osman Güneş, Peggy Hall, James Hayhurst, et al. The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. *Nucleic Acids Research*, 51(D1):D977–D985, 2023.

[8] Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O'Connell, et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203–209, 2018.

[9] Liis Leitsalu, Toomas Haller, Tõnu Esko, Mari-Liis Tammesoo, Helene Alavere, Harold Snieder, Markus Perola, Pauline C Ng, Reedik Mägi, Lili Milani, et al. Cohort profile: Estonian Biobank of the Estonian Genome Center, University of Tartu. *International Journal of Epidemiology*, 44(4):1137–1147, 2015.

[10] John Michael Gaziano, John Concato, Mary Brophy, Louis Fiore, Saiju Pyarajan, James Breeling, Stacey Whitbourne, Jennifer Deen, Colleen Shannon, Donald Humphries, et al. Million Veteran Program: A mega-biobank to study genetic influences on health and disease. *Journal of Clinical Epidemiology*, 70:214–223, 2016.

[11] Mitja I Kurki, Juha Karjalainen, Priit Palta, Timo P Sipilä, Kati Kristiansson, Kati M Donner, Mary P Reeve, Hannele Laivuori, Mervi Aavikko, Mari A Kaunisto, et al. FinnGen provides genetic insights from a well-phenotyped isolated population. *Nature*, 613(7944):508–518, 2023.

[12] Olivier Delaneau, Jonathan Marchini, and Jean-François Zagury. A linear complexity phasing method for thousands of genomes. *Nature methods*, 9(2):179–181, 2012.

[13] Joshua D Backman, Alexander H Li, Anthony Marcketta, Dylan Sun, Joelle Mbatchou, Michael D Kessler, Christian Benner, Daren Liu, Adam E Locke, Suganthi Balasubramanian, et al. Exome sequencing and analysis of 454,787 UK Biobank participants. *Nature*, 599(7886):628–634, 2021.

[14] Konrad J Karczewski, Matthew Solomonson, Katherine R Chao, Julia K Goodrich, Grace Tiao, Wenhan Lu, Bridget M Riley-Gillis, Ellen A Tsai, Hye In Kim, Xiuwen Zheng, et al. Systematic single-variant and gene-based association testing of thousands of phenotypes in 394,841 UK Biobank exomes. *Cell Genomics*, 2(9), 2022.

[15] Tomas Fitzgerald and Ewan Birney. CNest: a novel copy number association discovery method uncovers 862 new associations from 200,629 whole-exome sequence datasets in the UK Biobank. *Cell Genomics*, 2(8), 2022.

[16] Chiara Auwerx, Maarja Lepamets, Marie C Sadler, Marion Patxot, Miloš Stojanov, David Baud, Reedik Mägi, Tõnu Esko, Andres Metspalu, Lili Milani, et al. The individual and global impact of copy-number variants on complex human traits. *The American Journal of Human Genetics*, 109(4):647–668, 2022.

[17] Ali Torkamani, Nathan E Wineinger, and Eric J Topol. The personal and clinical utility of polygenic risk scores. *Nature Reviews Genetics*, 19(9):581–590, 2018.

[18] Brendan K Bulik-Sullivan, Po-Ru Loh, Hilary K Finucane, Stephan Ripke, Jian Yang, Schizophrenia Working Group of the Psychiatric Genomics Consortium, Nick Patterson, Mark J Daly, Alkes L Price, and Benjamin M Neale. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature genetics*, 47(3):291–295, 2015.

[19] Brendan Bulik-Sullivan, Hilary K Finucane, Verneri Anttila, Alexander Gusev, Felix R Day, Po-Ru Loh, ReproGen Consortium, Psychiatric Genomics Consortium, Genetic Consortium for Anorexia Nervosa of the Wellcome Trust Case Control Consortium 3, Laramie Duncan, et al. An atlas of genetic correlations across human diseases and traits. *Nature genetics*, 47(11):1236–1241, 2015.

[20] George Davey Smith and Shah Ebrahim. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *International Journal ofEpidemiology*, 32(1):1–22, 2003.

[21] Michael D Gallagher and Alice S Chen-Plotkin. The post-GWAS era: from association to function. *The American Journal of Human Genetics*, 102(5):717–730, 2018.

[22] Daniel J Schaid, Wenan Chen, and Nicholas B Larson. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nature Reviews Genetics*, 19(8):491–504, 2018.

[23] Matthew T Maurano, Richard Humbert, Eric Rynes, Robert E Thurman, Eric Haugen, Hao Wang, Alex P Reynolds, Richard Sandstrom, Hongzhu Qu, Jennifer Brody, et al. Systematic localization of common disease-associated variation in regulatory DNA. *Science*, 337(6099):1190–1195, 2012.

[24] Dan L Nicolae, Eric Gamazon, Wei Zhang, Shiwei Duan, M Eileen Dolan, and Nancy J Cox. Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS genetics*, 6(4):e1000888, 2010.

[25] Marc A Schaub, Alan P Boyle, Anshul Kundaje, Serafim Batzoglou, and Michael Snyder. Linking disease associations with regulatory information in the human genome. *Genome research*, 22(9):1748–1759, 2012.

[26] Yehudit Hasin, Marcus Seldin, and Aldons Lusis. Multi-omics approaches to disease. *Genome biology*, 18(1):1–15, 2017.

[27] Inc. Illumina. Infinium MethylationEPIC v2.0 Kit. `https://emea.illumina.com/products/by-type/microarray-kits/infinium-methylation-epic.html`, 2023. Accessed on August, 2023.

[28] Josine L Min, Gibran Hemani, Eilis Hannon, Koen F Dekkers, Juan Castillo-Fernandez, René Luijk, Elena Carnero-Montoro, Daniel J Lawson, Kimberley Burrows, Matthew Suderman, et al. Genomic and phenotypic insights from an atlas of genetic effects on DNA methylation. *Nature genetics*, 53(9):1311–1321, 2021.

[29] Meritxell Oliva, Kathryn Demanelis, Yihao Lu, Meytal Chernoff, Farzana Jasmine, Habibul Ahsan, Muhammad G Kibriya, Lin S Chen, and Brandon L Pierce. DNA methylation QTL mapping across diverse human tissues provides molecular links between genetic variation and complex traits. *Nature genetics*, 55(1):112–122, 2023.

[30] GTEx Consortium. The GTEx consortium atlas of genetic regulatory effects across human tissues. *Science*, 369(6509):1318–1330, 2020.

[31] Urmo Võsa, Annique Claringbould, Harm-Jan Westra, Marc Jan Bonder, Patrick Deelen, Biao Zeng, Holger Kirsten, Ashis Saha, Roman Kreuzhuber, Harm Brugge, et al. Large-scale cis-and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. *Nature genetics*, pages 1–11, 2021.

[32] Nurlan Kerimov, James D Hayhurst, Kateryna Peikova, Jonathan R Manning, Peter Walter, Liis Kolberg, Marija Samoviča, Manoj Pandian Sakthivel, Ivan Kuzmin, Stephen J Trevanion, et al. A compendium of uniformly processed human gene expression and splicing quantitative trait loci. *Nature genetics*, 53(9):1290–1299, 2021.

[33] Niek de Klein, Ellen A Tsai, Martijn Vochteloo, Denis Baird, Yunfeng Huang, Chia-Yen Chen, Sipko van Dam, Roy Oelen, Patrick Deelen, Olivier B Bakker, et al. Brain expression quantitative trait locus and network analyses reveal downstream effects and putative drivers for brain-related diseases. *Nature genetics*, 55(3):377–388, 2023.

[34] Benjamin B Sun, Joseph C Maranville, James E Peters, David Stacey, James R Staley, James Blackshaw, Stephen Burgess, Tao Jiang, Ellie Paige, Praveen Surendran, et al. Genomic atlas of the human plasma proteome. *Nature*, 558(7708):73–79, 2018.

[35] Egil Ferkingstad, Patrick Sulem, Bjarni A Atlason, Gardar Sveinbjornsson, Magnus I Magnusson, Edda L Styrmisdottir, Kristbjorg Gunnarsdottir, Agnar Helgason, Asmundur Oddsson, Bjarni V Halldorsson, et al. Large-scale integration of the plasma proteome with genetics and disease. *Nature Genetics*, 53(12):1712–1721, 2021.

[36] Benjamin B Sun, Joshua Chiou, Matthew Traylor, Christian Benner, Yi-Hsiang Hsu, Tom G Richardson, Praveen Surendran, Anubha Mahajan, Chloe Robins, Steven G Vasquez-Grinnell, et al. Plasma proteomic associations with genetics and health in the UK Biobank. *Nature*, pages 1–10, 2023.

[37] So-Youn Shin, Eric B Fauman, Ann-Kristin Petersen, Jan Krumsiek, Rita Santos, Jie Huang, Matthias Arnold, Idil Erte, Vincenzo Forgetta, Tsun-Po Yang, et al. An atlas of genetic influences on human blood metabolites. *Nature genetics*, 46(6):543–550, 2014.

[38] Luca A Lotta, Maik Pietzner, Isobel D Stewart, Laura BL Wittemans, Chen Li, Roberto Bonelli, Johannes Raffler, Emma K Biggs, Clare Oliver-Williams, Victoria PW Auyeung, et al. A cross-platform approach identifies genetic regulators of human metabolism and health. *Nature Genetics*, 53(1):54–64, 2021.

[39] Heli Julkunen, Anna Cichońska, Mika Tiainen, Harri Koskela, Kristian Nybo, Valtteri Mäkelä, Jussi Nokso-Koivisto, Kati Kristiansson, Markus Perola, Veikko Salomaa, et al. Atlas of plasma nmr biomarkers for health and disease in 118,461 individuals from the UK Biobank. *Nature Communications*, 14(1):604, 2023.

[40] Yiheng Chen, Tianyuan Lu, Ulrika Pettersson-Kymmer, Isobel D Stewart, Guillaume Butler-Laporte, Tomoko Nakanishi, Agustin Cerani, Kevin YH Liang, Satoshi Yoshiji, Julian Daniel Sunday Willett, et al. Genomic atlas of the plasma metabolome prioritizes metabolites implicated in human diseases. *Nature Genetics*, 55(1):44–53, 2023.

[41] Matthew D Schultz, Yupeng He, John W Whitaker, Manoj Hariharan, Eran A Mukamel, Danny Leung, Nisha Rajagopal, Joseph R Nery, Mark A Urich, Huaming Chen, et al. Human body epigenome maps

reveal noncanonical DNA methylation variation. *Nature*, 523(7559):212–216, 2015.

[42] Ruth Pidsley, Elena Zotenko, Timothy J Peters, Mitchell G Lawrence, Gail P Risbridger, Peter Molloy, Susan Van Djik, Beverly Muhlhausler, Clare Stirzaker, and Susan J Clark. Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. *Genome biology*, 17(1):1–17, 2016.

[43] Benjamin D Singer. A practical guide to the measurement and analysis of DNA methylation. *American journal of respiratory cell and molecular biology*, 61(4):417–428, 2019.

[44] Zhong Wang, Mark Gerstein, and Michael Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews genetics*, 10(1):57–63, 2009.

[45] Rory Stark, Marta Grzelak, and James Hadfield. RNA sequencing: the teenage years. *Nature Reviews Genetics*, 20(11):631–656, 2019.

[46] Daniel H Katz, Jeremy M Robbins, Shuliang Deng, Usman A Tahir, Alexander G Bick, Akhil Pampana, Zhi Yu, Debby Ngo, Mark D Benson, Zsu-Zsu Chen, et al. Proteomic profiling platforms head to head: leveraging genetics and clinical traits to compare aptamer-and antibody-based methods. *Science Advances*, 8(33):eabm5164, 2022.

[47] M Fraga-Corral, M Carpena, P Garcia-Oliveira, AG Pereira, MA Prieto, and J Simal-Gandara. Analytical metabolomics and applications in health, environmental and food science. *Critical Reviews in Analytical Chemistry*, 52(4):712–734, 2022.

[48] Elena V Feofanova, Michael R Brown, Taryn Alkis, Astrid M Manuel, Xihao Li, Usman A Tahir, Zilin Li, Kevin M Mendez, Rachel S Kelly, Qibin Qi, et al. Whole-genome sequencing analysis of human metabolome in multi-ethnic populations. *Nature Communications*, 14(1):3111, 2023.

[49] Eleonora Porcu, Marie C. Sadler, Kaido Lepik, Chiara Auwerx, Andrew R. Wood, Antoine Weihs, Maroun S. Bou Sleiman, Diogo M. Ribeiro, Stefania Bandinelli, Toshiko Tanaka, Matthias Nauck, Uwe Völker, Olivier Delaneau, Andres Metspalu, Alexander Teumer, Timothy Frayling, Federico A. Santoni, Alexandre Reymond, and Zoltán Kutalik. Differentially expressed genes reflect disease-induced rather than disease-causing changes in the transcriptome. *Nature Communications*, 12(1):5647, September 2021.

[50] Claudia Giambartolomei, Damjan Vukcevic, Eric E Schadt, Lude Franke, Aroon D Hingorani, Chris Wallace, and Vincent Plagnol. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet*, 10(5):e1004383, 2014.

[51] Farhad Hormozdiari, Martijn Van De Bunt, Ayellet V Segre, Xiao Li, Jong Wha J Joo, Michael Bilow, Jae Hoon Sul, Sriram Sankararaman, Bogdan Pasaniuc, and Eleazar Eskin. Colocalization of GWAS and eQTL signals detects target genes. *The American Journal of Human Genetics*, 99(6):1245–1260, 2016.

[52] Chris Wallace. A more accurate method for colocalisation analysis allowing for multiple causal variants. *PLoS genetics*, 17(9):e1009440, 2021.

[53] Veronika W Skrivankova, Rebecca C Richmond, Benjamin AR Woolf, James Yarmolinsky, Neil M Davies, Sonja A Swanson, Tyler J VanderWeele, Julian PT Higgins, Nicholas J Timpson, Niki Dimou, et al. Strengthening the reporting of observational studies in epidemiology using Mendelian randomization: the STROBE-MR statement. *Jama*, 326(16):1614–1621, 2021.

[54] Eleonora Porcu, Sina Rüeger, Kaido Lepik, Federico A Santoni, Alexandre Reymond, and Zoltán Kutalik. Mendelian randomization integrating GWAS and eQTL data reveals genetic determinants of complex and clinical traits. *Nature communications*, 10(1):1–12, 2019.

[55] Stephen Burgess, Verena Zuber, Elsa Valdes-Marquez, Benjamin B Sun, and Jemma C Hopewell. Mendelian randomization with fine-mapped genetic data: choosing from large numbers of correlated instrumental variables. *Genetic epidemiology*, 41(8):714–725, 2017.

[56] Zhihong Zhu, Futao Zhang, Han Hu, Andrew Bakshi, Matthew R Robinson, Joseph E Powell, Grant W Montgomery, Michael E Goddard, Naomi R Wray, Peter M Visscher, et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nature genetics*, 48(5):481–487, 2016.

[57] Verena Zuber, Nastasiya F Grinberg, Dipender Gill, Ichcha Manipur, Eric AW Slob, Ashish Patel, Chris Wallace, and Stephen Burgess. Combining evidence from Mendelian randomization and colocalization: Review and comparison of approaches. *The American Journal of Human Genetics*, 2022.

[58] Alexander Gusev, Arthur Ko, Huwenbo Shi, Gaurav Bhatia, Wonil Chung, Brenda WJH Penninx, Rick Jansen, Eco JC De Geus, Dorret I Boomsma, Fred A Wright, et al. Integrative approaches for large-scale transcriptome-wide association studies. *Nature genetics*, 48(3):245–252, 2016.

[59] Alvaro N Barbeira, Scott P Dickinson, Rodrigo Bonazzola, Jiamao Zheng, Heather E Wheeler, Jason M Torres, Eric S Torstenson, Kaanan P Shah, Tzintzuni Garcia, Todd L Edwards, et al. Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. *Nature communications*, 9(1):1–20, 2018.

[60] Kaido Lepik. *Inferring causality between transcriptome and complex traits*. Phd thesis, Institute of Computer Science, Faculty of Science and Technology, University of Tartu, Estonia., Tartu, Estonia, May 2021. Available at `http://dspace.ut.ee/bitstream/handle/10062/71645/lepik_kaido.pdf`.

[61] Gibran Hemani, Kate Tilling, and George Davey Smith. Orienting the causal relationship between imprecisely measured traits using GWAS summary data. *PLoS genetics*, 13(11):e1007081, 2017.

[62] Stephen Burgess, Dylan S Small, and Simon G Thompson. A review of instrumental variable estimators for Mendelian randomization. *Statistical methods in medical research*, 26(5):2333–2355, 2017.

[63] Eleanor Sanderson. Multivariable Mendelian randomization and mediation. *Cold Spring Harbor perspectives in medicine*, 11(2):a038984, 2021.

[64] Yang Wu, Jian Zeng, Futao Zhang, Zhihong Zhu, Ting Qi, Zhili Zheng, Luke R Lloyd-Jones, Riccardo E Marioni, Nicholas G Martin, Grant W Montgomery, et al. Integrative analysis of omics summary data reveals putative mechanisms underlying complex traits. *Nature communications*, 9(1):1–14, 2018.

[65] Alice R Carter, Eleanor Sanderson, Gemma Hammerton, Rebecca C Richmond, George Davey Smith, Jon Heron, Amy E Taylor, Neil M Davies, and Laura D Howe. Mendelian randomisation for mediation analysis: current methods and challenges for implementation. *European journal of epidemiology*, 36(5):465–478, 2021.

[66] Claudia Giambartolomei, Jimmy Zhenli Liu, Wen Zhang, Mads Hauberg, Huwenbo Shi, James Boocock, Joe Pickrell, Andrew E Jaffe, CommonMind Consortium, Bogdan Pasaniuc, et al. A Bayesian framework for multiple trait colocalization from summary association statistics. *Bioinformatics*, 34(15):2538–2545, 2018.

[67] Eilis Hannon, Tyler J Gorrie-Stone, Melissa C Smart, Joe Burrage, Amanda Hughes, Yanchun Bao, Meena Kumari, Leonard C Schalkwyk, and Jonathan Mill. Leveraging DNA-methylation quantitative-trait loci to characterize the relationship between methylomic variation, gene expression, and complex traits. *The American Journal of Human Genetics*, 103(5):654–665, 2018.

[68] John Novembre, Toby Johnson, Katarzyna Bryc, Zoltán Kutalik, Adam R Boyko, Adam Auton, Amit Indap, Karen S King, Sven Bergmann, Matthew R Nelson, et al. Genes mirror geography within europe. *Nature*, 456(7218):98–101, 2008.

[69] Matthew R Robinson, Aaron Kleinman, Mariaelisa Graff, Anna AE Vinkhuyzen, David Couper, Michael B Miller, Wouter J Peyrot, Abdel Abdellaoui, Brendan P Zietsch, Ilja M Nolte, et al. Genetic evidence of assortative mating in humans. *Nature Human Behaviour*, 1(1):0016, 2017.

[70] Laurence J Howe, Michel G Nivard, Tim T Morris, Ailin F Hansen, Humaira Rasheed, Yoonsu Cho, Geetha Chittoor, Rafael Ahlskog, Penelope A Lind, Teemu Palviainen, et al. Within-sibship genome-wide association analyses decrease bias in estimates of direct genetic effects. *Nature genetics*, 54(5):581–592, 2022.

[71] Alan Wayne Jones. Early drug discovery and the rise of pharmaceutical chemistry. *Drug testing and analysis*, 3(6):337–344, 2011.

[72] Thomas A Ban. The role of serendipity in drug discovery. *Dialogues in clinical neuroscience*, 8(3):335–344, 2006.

[73] Klaus Strebhardt and Axel Ullrich. Paul Ehrlich's magic bullet concept: 100 years of progress. *Nature Reviews Cancer*, 8(6):473–480, 2008.

[74] Andreas-Holger Maehle, Cay-Rüdiger Prüll, and Robert F Halliwell. The emergence of the drug receptor theory. *Nature Reviews Drug Discovery*, 1(8):637–641, 2002.

[75] Leland J Gershell and Joshua H Atkins. A brief history of novel drug discovery technologies. *Nature Reviews Drug Discovery*, 2(4):321–327, 2003.

[76] Stephen J Hill. G-protein-coupled receptors: past, present and future. *British journal of pharmacology*, 147(S1):S27–S37, 2006.

[77] Alexander S Hauser, Sreenivas Chavali, Ikuo Masuho, Leonie J Jahn, Kirill A Martemyanov, David E Gloriam, and M Madan Babu. Pharmacogenomics of GPCR drug targets. *Cell*, 172(1-2):41–54, 2018.

[78] David C Swinney and Jason Anthony. How were new medicines discovered? *Nature reviews Drug discovery*, 10(7):507–519, 2011.

[79] Matthew R Nelson, Hannah Tipney, Jeffery L Painter, Judong Shen, Paola Nicoletti, Yufeng Shen, Aris Floratos, Pak Chung Sham, Mulin Jun Li, Junwen Wang, et al. The support of human genetic evidence for approved drug indications. *Nature genetics*, 47(8):856–860, 2015.

[80] Emily A King, J Wade Davis, and Jacob F Degner. Are drug targets with genetic support twice as likely to be approved? Revised estimates of the impact of genetic support for drug mechanisms on the probability of drug approval. *PLoS genetics*, 15(12):e1008489, 2019.

[81] David Ochoa, Mohd Karim, Maya Ghoussaini, David G Hulcoop, Ellen M McDonagh, and Ian Dunham. Human genetics evidence supports two-thirds of the 2021 FDA-approved drugs. *Nat Rev Drug Discov*, 21(8):551, 2022.

[82] David Ochoa, Andrew Hercules, Miguel Carmona, Daniel Suveges, Asier Gonzalez-Uriarte, Cinzia Malangone, Alfredo Miranda, Luca Fumis, Denise Carvalho-Silva, Michaela Spitzer, et al. Open Targets Platform: supporting systematic drug–target identification and prioritisation. *Nucleic acids research*, 49(D1):D1302–D1310, 2021.

[83] Phuong A Nguyen, David A Born, Aimee M Deaton, Paul Nioi, and Lucas D Ward. Phenotypes associated with genes encoding drug targets are predictive of clinical trial side effects. *Nature communications*, 10(1):1579, 2019.

[84] Jitao David Zhang, Lisa Sach-Peltason, Christian Kramer, Ken Wang, and Martin Ebeling. Multiscale modelling of drug mechanism and safety. *Drug Discovery Today*, 25(3):519–534, 2020.

[85] Loïc Yengo, Sailaja Vedantam, Eirini Marouli, Julia Sidorenko, Eric Bartell, Saori Sakaue, Marielisa Graff, Anders U Eliasen, Yunxuan Jiang, Sridharan Raghavan, et al. A saturated map of common genetic variants associated with human height. *Nature*, 610(7933):704–712, 2022.

[86] Steven Gazal, Omer Weissbrod, Farhad Hormozdiari, Kushal K Dey, Joseph Nasser, Karthik A Jagadeesh, Daniel J Weiner, Huwenbo Shi, Charles P Fulco, Luke J O'Connor, et al. Combining SNP-to-gene linking strategies to identify disease genes and assess disease omnigenicity. *Nature Genetics*, 55:827–836, 2022.

[87] Edward Mountjoy, Ellen M Schmidt, Miguel Carmona, Jeremy Schwartzentruber, Gareth Peat, Alfredo Miranda, Luca Fumis, James Hayhurst, Annalisa Buniello, Mohd Anisul Karim, et al. An open approach to systematically prioritize causal variants and genes at all published human GWAS trait-associated loci. *Nature Genetics*, 53(11):1527–1533, 2021.

[88] Elle M Weeks, Jacob C Ulirsch, Nathan Y Cheng, Brian L Trippe, Rebecca S Fine, Jenkai Miao, Tejal A Patwardhan, Masahiro Kanai, Joseph Nasser, Charles P Fulco, et al. Leveraging polygenic enrichments of gene features to predict genes underlying complex traits and diseases. *Nature Genetics*, pages 1–10, 2023.

[89] Hai Fang, Hans De Wolf, Bogdan Knezevic, Katie L Burnham, Julie Osgood, Anna Sanniti, Alicia Lledó Lara, Silva Kasela, Stephane De Cesco, Jörg K Wegner, et al. A genetics-led approach defines the drug target landscape of 30 immune-related traits. *Nature genetics*, 51(7):1082–1091, 2019.

[90] Jian Yang, Teresa Ferreira, Andrew P Morris, Sarah E Medland, Genetic Investigation of ANthropometric Traits (GIANT) Consortium, DIAbetes Genetics Replication, Meta analysis (DIAGRAM) Consortium, Pamela AF Madden, Andrew C Heath, Nicholas G Martin, Grant W Montgomery, et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nature genetics*, 44(4):369–375, 2012.

[91] Christian Benner, Chris CA Spencer, Aki S Havulinna, Veikko Salomaa, Samuli Ripatti, and Matti Pirinen. FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics*, 32(10):1493–1501, 2016.

[92] Gao Wang, Abhishek Sarkar, Peter Carbonetto, and Matthew Stephens. A simple new approach to variable selection in regression, with application to genetic fine mapping. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(5):1273–1300, 2020.

[93] Adam Frankish, Sílvia Carbonell-Sala, Mark Diekhans, Irwin Jungreis, Jane E Loveland, Jonathan M Mudge, Cristina Sisu, James C Wright, Carme Arnan, If Barnes, et al. GENCODE: reference annotation for the human and mouse genomes in 2023. *Nucleic acids research*, 51(D1):D942–D949, 2023.

[94] Fergal J Martin, M Ridwan Amode, Alisha Aneja, Olanrewaju Austine-Orimoloye, Andrey G Azov, If Barnes, Arne Becker, Ruth Bennett, Andrew Berry, Jyothish Bhai, et al. Ensembl 2023. *Nucleic acids research*, 51(D1):D933–D941, 2023.

[95] Roadmap Epigenomics Consortium Integrative analysis coordination Kundaje Anshul 1 2 3 Meuleman Wouter 1 2 Ernst Jason 1 2 4 Bilenky Misha 5, Scientific program management Chadwick Lisa H. 53, and Principal investigators Bernstein Bradley E. 2 26 42 Costello Joseph F. 14 Ecker Joseph R. 9 Hirst Martin 5 18 Meissner Alexander 2 6 Milosavljevic Aleksandar 7 Ren Bing 8 13 Stamatoyannopoulos John A. 10 Wang Ting 21 Kellis Manolis 1 2. Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317–330, 2015.

[96] Carles A Boix, Benjamin T James, Yongjin P Park, Wouter Meuleman, and Manolis Kellis. Regulatory genomic circuitry of human disease loci by integrative epigenomics. *Nature*, 590(7845):300–307, 2021.

[97] Christiaan A de Leeuw, Joris M Mooij, Tom Heskes, and Danielle Posthuma. MAGMA: generalized gene-set analysis of GWAS data. *PLoS computational biology*, 11(4):e1004219, 2015.

[98] David Lamparter, Daniel Marbach, Rico Rueedi, Zoltán Kutalik, and Sven Bergmann. Fast and rigorous computation of gene and pathway scores from SNP-based summary statistics. *PLoS computational biology*, 12(1):e1004714, 2016.

[99] Takiy-Eddine Berrandou, David Balding, and Doug Speed. LDAK-GBAT: fast and powerful gene-based association testing using summary statistics. *The American Journal of Human Genetics*, 110(1):23–29, 2023.

[100] Michael C Wu, Seunggeun Lee, Tianxi Cai, Yun Li, Michael Boehnke, and Xihong Lin. Rare-variant association testing for sequencing data

with the sequence kernel association test. *The American Journal of Human Genetics*, 89(1):82–93, 2011.

[101] Seunggeun Lee, Mary J Emond, Michael J Bamshad, Kathleen C Barnes, Mark J Rieder, Deborah A Nickerson, ESP Lung Project Team, David C Christiani, Mark M Wurfel, Xihong Lin, et al. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *The American Journal of Human Genetics*, 91(2):224–237, 2012.

[102] Munir Pirmohamed. Pharmacogenomics: Current status and future perspectives. *Nature Reviews Genetics*, pages 1–13, 2023.

[103] Chiara Auwerx, Marie C Sadler, Alexandre Reymond, and Zoltán Kutalik. From pharmacogenetics to pharmaco-omics: Milestones and future directions. *Human Genetics and Genomics Advances*, 2022.

[104] Volker M Lauschke, Yitian Zhou, and Magnus Ingelman-Sundberg. Novel genetic and epigenetic factors of importance for inter-individual differences in drug disposition, response and toxicity. *Pharmacology & Therapeutics*, 197:122–152, 2019.

[105] Dan M Roden, Howard L McLeod, Mary V Relling, Marc S Williams, George A Mensah, Josh F Peterson, and Sara L Van Driest. Parmacogenomics. *Lancet*, 394:521–532, 2019.

[106] Pierre Dayer, Jules Desmeules, Thierry Leemann, and Rita Striberni. Bioactivation of the narcotic drug codeine in human liver is mediated by the polymorphic monooxygenase catalyzing debrisoquine 4-hydroxylation (cytochrome P-450 dbl/bufI). *Biochemical and biophysical research communications*, 152(1):411–416, 1988.

[107] Jessica L Mega, Tabassome Simon, Jean-Philippe Collet, Jeffrey L Anderson, Elliott M Antman, Kevin Bliden, Christopher P Cannon, Nicolas Danchin, Betti Giusti, Paul Gurbel, et al. Reduced-function CYP2C19 genotype and risk of adverse clinical outcomes among patients treated with clopidogrel predominantly for PCI: a meta-analysis. *Jama*, 304(16):1821–1830, 2010.

[108] Derek Van Booven, Sharon Marsh, Howard McLeod, Michelle Whirl Carrillo, Katrin Sangkuhl, Teri E Klein, and Russ B Altman. Cytochrome P450 2C9-CYP2C9. *Pharmacogenetics and genomics*, 20(4):277, 2010.

[109] Simone Rost, Andreas Fregin, Vytautas Ivaskevicius, Ernst Conzelmann, Konstanze Hörtnagel, Hans-Joachim Pelz, Knut Lappegard, Erhard Seifried, Inge Scharrer, Edward GD Tuddenham, et al. Mutations in VKORC1 cause warfarin resistance and multiple coagulation factor deficiency type 2. *Nature*, 427(6974):537–541, 2004.

[110] Stephen G Gonsalves, Robert T Dirksen, Katrin Sangkuhl, Rebecca Pulk, Maria Alvarellos, Teresa Vo, Keiko Hikino, Dan Roden, Teri E Klein, S Mark Poler, et al. Clinical Pharmacogenetics Implementation Consortium (CPIC) guideline for the use of potent volatile anesthetic agents and succinylcholine in the context of RYR 1 or CACNA 1S genotypes. *Clinical Pharmacology & Therapeutics*, 105(6):1338–1344, 2019.

[111] Kelly E Caudle, Henry M Dunnenberger, Robert R Freimuth, Josh F Peterson, Jonathan D Burlison, Michelle Whirl-Carrillo, Stuart A Scott, Heidi L Rehm, Marc S Williams, Teri E Klein, et al. Standardizing terms for clinical pharmacogenetic test results: consensus terms from the Clinical Pharmacogenetics Implementation Consortium (CPIC). *Genetics in Medicine*, 19(2):215–223, 2017.

[112] Andrea Gaedigk, Magnus Ingelman-Sundberg, Neil A Miller, J Steven Leeder, Michelle Whirl-Carrillo, Teri E Klein, and PharmVar Steering Committee. The Pharmacogene Variation (PharmVar) Consortium: incorporation of the human cytochrome P450 (CYP) allele nomenclature database. *Clinical Pharmacology & Therapeutics*, 103(3):399–401, 2018.

[113] MV Relling and TE Klein. CPIC: clinical pharmacogenetics implementation consortium of the pharmacogenomics research network. *Clinical Pharmacology & Therapeutics*, 89(3):464–467, 2011.

[114] JJ Swen, I Wilting, AL De Goede, L Grandia, H Mulder, DJ Touw, A De Boer, JMH Conemans, TCG Egberts, OH Klungel, et al. Pharmacogenetics: from bench to byte. *Clinical Pharmacology & Therapeutics*, 83(5):781–787, 2008.

[115] Nicolas Picard, Jean-Christophe Boyer, Marie-Christine Etienne-Grimaldi, Chantal Barin-Le Guellec, Fabienne Thomas, Marie-Anne Loriot, and French National Network of Pharmacogenetics. Pharmacogenetics-based personalized therapy: levels of evidence and recommendations from the French Network of Pharmacogenetics (RNPGx). *Therapies*, 72(2):185–192, 2017.

[116] Heshu Abdullah-Koolmees, Antonius M Van Keulen, Marga Nijenhuis, and Vera HM Deneer. Pharmacogenetics guidelines: overview and comparison of the DPWG, CPIC, CPNDS, and RNPGx guidelines. *Frontiers in pharmacology*, 11:595219, 2021.

[117] Simon Mallal, Elizabeth Phillips, Giampiero Carosi, Jean-Michel Molina, Cassy Workman, Janez Tomažič, Eva Jägel-Guedes, Sorin Rugina, Oleg Kozyrev, Juan Flores Cid, et al. HLA-B* 5701 screening for hypersensitivity to abacavir. *The New England Journal of Medicine*, 358(6):568–579, 2008.

[118] Stephen E Kimmel, Benjamin French, Scott E Kasner, Julie A Johnson, Jeffrey L Anderson, Brian F Gage, Yves D Rosenberg, Charles S Eby, Rosemary A Madigan, Robert B McBane, et al. A pharmacogenetic versus a clinical algorithm for warfarin dosing. *New England Journal of Medicine*, 369(24):2283–2293, 2013.

[119] Munir Pirmohamed, Girvan Burnside, Niclas Eriksson, Andrea L Jorgensen, Cheng Hock Toh, Toby Nicholson, Patrick Kesteven, Christina Christersson, Bengt Wahlström, Christina Stafberg, et al. A randomized trial of genotype-guided dosing of warfarin. *New England Journal of Medicine*, 369(24):2294–2303, 2013.

[120] John F Greden, Sagar V Parikh, Anthony J Rothschild, Michael E Thase, Boadie W Dunlop, Charles DeBattista, Charles R Conway, Brent P Forester, Francis M Mondimore, Richard C Shelton, et al. Impact of pharmacogenomics on clinical outcomes in major depressive disorder in the GUIDED trial: a large, patient-and rater-blinded, randomized, controlled study. *Journal of psychiatric research*, 111:59–67, 2019.

[121] Daniel MF Claassens, Gerrit JA Vos, Thomas O Bergmeijer, Renicus S Hermanides, Arnoud WJ Van't Hof, Pim Van Der Harst, Emanuele Barbato, Carmine Morisco, Richard M Tjon Joe Gin, Folkert W Asselbergs, et al. A genotype-guided strategy for oral P2Y12 inhibitors in primary PCI. *New England Journal of Medicine*, 381(17):1621–1631, 2019.

[122] Naveen L. Pereira, Michael E. Farkouh, Derek So, Ryan Lennon, Nancy Geller, Verghese Mathew, Malcolm Bell, Jang-Ho Bae, Myung Ho Jeong, Ivan Chavez, Paul Gordon, J. Dawn Abbott, Charles Cagin, Linnea Baudhuin, Yi-Ping Fu, Shaun G. Goodman, Ahmed Hasan, Erin Iturriaga, Amir Lerman, Mandeep Sidhu, Jean-Francois Tanguay, Liewei Wang, Richard Weinshilboum, Robert Welsh, Yves Rosenberg, Kent

Bailey, and Charanjit Rihal. Effect of genotype-guided oral P2Y12 inhibitor selection vs conventional clopidogrel therapy on ischemic outcomes after percutaneous coronary intervention: The TAILOR-PCI randomized clinical trial. *JAMA*, 324(8):761–771, 08 2020.

[123] Jesse J Swen, Cathelijne H van der Wouden, Lisanne EN Manson, Heshu Abdullah-Koolmees, Kathrin Blagec, Tanja Blagus, Stefan Böhringer, Anne Cambon-Thomsen, Erika Cecchin, Ka-Chun Cheung, et al. A 12-gene pharmacogenetic panel to prevent adverse drug reactions: an open-label, multicentre, controlled, cluster-randomised crossover implementation study. *The Lancet*, 401(10374):347–356, 2023.

[124] David W Oslin, Kevin G Lynch, Mei-Chiung Shih, Erin P Ingram, Laura O Wray, Sara R Chapman, Henry R Kranzler, Joel Gelernter, Jeffrey M Pyne, Annjanette Stone, et al. Effect of pharmacogenomic testing for drug-gene interactions on medication selection and remission of symptoms in major depressive disorder: the PRIME Care randomized clinical trial. *Jama*, 328(2):151–161, 2022.

[125] Brian F Gage, Anne R Bass, Hannah Lin, Scott C Woller, Scott M Stevens, Noor Al-Hammadi, Juan Li, Tomás Rodríguez, J Philip Miller, Gwendolyn A McMillin, et al. Effect of genotype-guided warfarin dosing on clinical events and anticoagulation control among patients undergoing hip or knee arthroplasty: the GIFT randomized clinical trial. *Jama*, 318(12):1115–1124, 2017.

[126] Gregory M McInnes, Adam Lavertu, Katrin Sangkuhl, Teri E Klein, Michelle Whirl-Carrillo, and Russ B Altman. Pharmacogenetics at scale: An analysis of the UK Biobank. *Clinical Pharmacology & Therapeutics*, 109(6):1528–1537, 2021.

[127] Nita A Limdi, Todd M Brown, Qi Yan, Jonathan L Thigpen, Aditi Shendre, Nianjun Liu, Charles E Hill, Donna K Arnett, and T Mark Beasley. Race influences warfarin dose changes associated with genetic factors. *Blood, The Journal of the American Society of Hematology*, 126(4):539–545, 2015.

[128] Rashmi R Shah and Robert L Smith. Addressing phenoconversion: the Achilles' heel of personalized medicine. *British journal of clinical pharmacology*, 79(2):222–240, 2015.

[129] Sylvia D Klomp, Martijn L Manson, Henk-Jan Guchelaar, and Jesse J Swen. Phenoconversion of cytochrome P450 metabolism: a systematic review. *Journal of Clinical Medicine*, 9(9):2890, 2020.

[130] Erica Bowton, Julie R Field, Sunny Wang, Jonathan S Schildcrout, Sara L Van Driest, Jessica T Delaney, James Cowan, Peter Weeke, Jonathan D Mosley, Quinn S Wells, et al. Biobanks and electronic medical records: enabling cost-effective research. *Science translational medicine*, 6(234):234cm3–234cm3, 2014.

[131] Tõnis Tasa, Kristi Krebs, Mart Kals, Reedik Mägi, Volker M Lauschke, Toomas Haller, Tarmo Puurand, Maido Remm, Tõnu Esko, Andres Metspalu, et al. Genetic variation in the Estonian population: pharmacogenomics study of adverse drug effects using electronic health records. *European Journal of Human Genetics*, 27(3):442–454, 2019.

[132] Haley Hunter-Zinck, Yunling Shi, Man Li, Bryan R Gorman, Sun-Gou Ji, Ning Sun, Teresa Webster, Andrew Liem, Paul Hsieh, Poornima Devineni, et al. Genotyping array design and data quality control in the Million Veteran Program. *The American Journal of Human Genetics*, 106(4):535–548, 2020.

[133] The All of Us Research Program Investigators. The “All of Us” research program. *The New England Journal of Medicine*, 381(7):668–676, 2019.

[134] Dan M Roden, Jill M Pulley, Melissa A Basford, Gordon R Bernard, Ellen W Clayton, Jeffrey R Balser, and Dan R Masys. Development of a large-scale de-identified dna biobank to enable personalized medicine. *Clinical Pharmacology & Therapeutics*, 84(3):362–369, 2008.

[135] Yeda Wu, Enda M Byrne, Zhili Zheng, Kathryn E Kemper, Loic Yengo, Andrew J Mallett, Jian Yang, Peter M Visscher, and Naomi R Wray. Genome-wide association study of medication-use and associated disease in the UK Biobank. *Nature communications*, 10(1):1891, 2019.

[136] Tuomo Kiiskinen, Pyry Helkkula, Kristi Krebs, Juha Karjalainen, Elmo Saarentaus, Nina Mars, Arto Lehisto, Wei Zhou, Mattia Cordioli, Sakari Jukarainen, et al. Genetic predictors of lifelong medication-use patterns in cardiometabolic diseases. *Nature Medicine*, 29(1):209–218, 2023.

[137] Mustafa Adnan Malki, Adem Y Dawed, Caroline Hayward, Alex Doney, and Ewan R Pearson. Utilizing large electronic medical record data sets to identify novel drug–gene interactions for commonly used drugs. *Clinical Pharmacology & Therapeutics*, 110(3):816–825, 2021.

[138] Gregory McInnes and Russ B Altman. Drug response pharmacogenetics for 200,000 UK Biobank participants. *Pacific Symposium on Biocomputing*, 26:184–195, 2021.

[139] Kristi Krebs, Jonas Bovijn, Neil Zheng, Maarja Lepamets, Jenny C Censin, Tuuli Jürgenson, Dage Särg, Erik Abner, Triin Laisk, Yang Luo, et al. Genome-wide study identifies association between HLA-B* 55:01 and self-reported penicillin allergy. *The American Journal of Human Genetics*, 107(4):612–621, 2020.

[140] Marie C Sadler, Chiara Auwerx, Kaido Lepik, Eleonora Porcu, and Zoltán Kutalik. Quantifying the role of transcript levels in mediating DNA methylation effects on complex traits and diseases. *Nature Communications*, 13(1):1–14, 2022.

[141] Ekaterina S Maksimova, Sven E Ojavee, Kristi Läll, Marie C Sadler, Reedik Mägi, Zoltan Kutalik, and Matthew R Robinson. Novel discoveries and enhanced genomic prediction from modelling genetic risk of cancer age-at-onset. *medRxiv*, pages 2022–03, 2022.

[142] Kejun Ying, Hanna Liu, Andrei E Tarkhov, Ake T Lu, Steve Horvath, Zoltan Kutalik, Xia Shen, and Vadim N Gladyshev. Causal epigenetic age uncouples damage and adaptation. *bioRxiv*, pages 2022–10, 2022.

[143] Chiara Auwerx, Marie C Sadler, Tristan Woh, Alexandre Reymond, Zoltán Kutalik, and Eleonora Porcu. Exploiting the mediating role of the metabolome to unravel transcript-to-phenotype associations. *Elife*, 12:e81097, 2023.

[144] Marie C Sadler, Chiara Auwerx, Patrick Deelen, and Zoltan Kutalik. Multi-layered genetic approaches to identify approved drug targets. *Cell Genomics*, 3(7):100341, 2023.

[145] Olivier B Bakker, Annique Claringbould, Harm-Jan Westra, Henry Wiersma, Floranne Boulogne, Urmo Võsa, Sophie Mulcahy Symmons, Iris H Jonkers, Lude Franke, and Patrick Deelen. Linking common and rare disease genetics through gene regulatory networks. *medRxiv*, pages 2021–10, 2021.

[146] Delaney A Smith, Marie C Sadler, and Russ B. Altman. Promises and challenges in pharmacoepigenetics. *Cambridge Prisms: Precision Medicine*, 1:e18, 2023.

[147] Douglas W Yao, Luke J O'Connor, Alkes L Price, and Alexander Gusev. Quantifying genetic effects on disease mediated by assayed gene expression levels. *Nature Genetics*, 52(6):626–633, 2020.

[148] Joyce B Kang, Alessandro Raveane, Aparna Nathan, Nicole Soranzo, and Soumya Raychaudhuri. Methods and insights from single-cell expression quantitative trait loci. *Annual Review of Genomics and Human Genetics*, 24, 2023.

[149] Luke J O'Connor, Armin P Schoech, Farhad Hormozdiari, Steven Gazal, Nick Patterson, and Alkes L Price. Extreme polygenicity of complex traits is explained by negative selection. *The American Journal of Human Genetics*, 105(3):456–476, 2019.

[150] Rita Santos, Oleg Ursu, Anna Gaulton, A Patrícia Bento, Ramesh S Donadi, Cristian G Bologa, Anneli Karlsson, Bissan Al-Lazikani, Anne Hersey, Tudor I Oprea, et al. A comprehensive map of molecular drug targets. *Nature reviews Drug discovery*, 16(1):19–34, 2017.

[151] Eric Vallabh Minikel, Jeffrey L Painter, Coco Chengliang Dong, and Matthew R Nelson. Refining the impact of genetic evidence on clinical success. *medRxiv*, pages 2023–06, 2023.

[152] Katerina Trajanoska, Claude Bhérer, Daniel Taliun, Sirui Zhou, J Brent Richards, and Vincent Mooser. From target discovery to clinical drug development with human genetics. *Nature*, 620(7975):737–745, 2023.

[153] Adriana Heguy, Alexander A Stewart, John D Haley, David E Smith, and J Gordon Foulkes. Gene expression as a target for new drug discovery. *Gene expression*, 4(6):337, 1995.

[154] Linglan Fang, Willem A Velema, Yujeong Lee, Xiao Lu, Michael G Mohsen, Anna M Kietrys, and Eric T Kool. Pervasive transcriptome interactions of protein-targeted drugs. *Nature Chemistry*, pages 2022–07, 2023.

[155] Aravind Subramanian, Rajiv Narayan, Steven M Corsello, David D Peck, Ted E Natoli, Xiaodong Lu, Joshua Gould, John F Davis, Andrew A Tubelli, Jacob K Asiedu, et al. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell*, 171(6):1437–1452, 2017.

[156] Nathaniel Lim and Paul Pavlidis. Evaluation of Connectivity Map shows limited reproducibility in drug repositioning. *Scientific Reports*, 11(17624), 2021.

[157] Bjarni V Halldorsson, Hannes P Eggertsson, Kristjan HS Moore, Hannes Hauswedell, Ogmundur Eiriksson, Magnus O Ulfarsson, Gunnar Palsson, Marteinn T Hardarson, Asmundur Oddsson, Brynjar O Jensson, et al. The sequences of 150,119 genomes in the UK Biobank. *Nature*, 607(7920):732–740, 2022.

[158] Olesya Razuvayevskaya, Irene Lopez, Ian Dunham, and David Ochoa. Why clinical trials stop: The role of genetics. *medRxiv*, pages 2023–02, 2023.

[159] Alicia R Martin, Masahiro Kanai, Yoichiro Kamatani, Yukinori Okada, Benjamin M Neale, and Mark J Daly. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nature genetics*, 51(4):584–591, 2019.

[160] Patrick Wu, QiPing Feng, Vern Eric Kerchberger, Scott D Nelson, Qingxia Chen, Bingshan Li, Todd L Edwards, Nancy J Cox, Elizabeth J Phillips, C Michael Stein, et al. Integrating gene expression and clinical data to identify drug repurposing candidates for hyperlipidemia and hypertension. *Nature Communications*, 13(1):46, 2022.

[161] Alice Taubes, Phil Nova, Kelly A Zalocusky, Idit Kosti, Mesude Bicak, Misha Y Zilberter, Yanxia Hao, Seo Yeon Yoon, Tomiko Oskotsky, Silvia Pineda, et al. Experimental and real-world evidence supporting the computational repurposing of bumetanide for APOE4-related Alzheimer's disease. *Nature Aging*, 1(10):932–947, 2021.

[162] Traci E LaMoia and Gerald I Shulman. Cellular and molecular mechanisms of metformin action. *Endocrine reviews*, 42(1):77–96, 2021.

[163] Mahin Delara, Lauren Murray, Behnaz Jafari, Anees Bahji, Zahra Goodarzi, Julia Kirkham, Mohammad Chowdhury, and Dallas P Seitz. Prevalence and factors associated with polypharmacy: a systematic review and meta-analysis. *BMC geriatrics*, 22(1):601, 2022.

Appendix A

Quantifying the role of transcript levels in mediating DNA methylation effects on complex traits and diseases

This article (Sadler *et al.*, 2022, *Nature Communications*) is presented in Chapter 2.

nature communications

Article

https://doi.org/10.1038/s41467-022-35196-3

Quantifying the role of transcript levels in mediating DNA methylation effects on complex traits and diseases

Received: 3 February 2022

Accepted: 22 November 2022

Published online: 07 December 2022

Check for updates

Marie C. Sadler[1,2,3], Chiara Auwerx[1,2,3,4], Kaido Lepik[1,2,3], Eleonora Porcu[1,2,3,4,5] & Zoltán Kutalik[1,2,3,5]

High-dimensional omics datasets provide valuable resources to determine the causal role of molecular traits in mediating the path from genotype to phenotype. Making use of molecular quantitative trait loci (QTL) and genome-wide association study (GWAS) summary statistics, we propose a multivariable Mendelian randomization (MVMR) framework to quantify the proportion of the impact of the DNA methylome (DNAm) on complex traits that is propagated through the assayed transcriptome. Evaluating 50 complex traits, we find that on average at least 28.3% (95% CI: [26.9%–29.8%]) of DNAm-to-trait effects are mediated through (typically multiple) transcripts in the *cis*-region. Several regulatory mechanisms are hypothesized, including methylation of the promoter probe cg10385390 (chr1:8′022′505) increasing the risk for inflammatory bowel disease by reducing *PARK7* expression. The proposed integrative framework can be extended to other omics layers to identify causal molecular chains, providing a powerful tool to map and interpret GWAS signals.

In the past decade, genome-wide association studies (GWASs) have identified thousands of genetic variants associated with complex traits[1], however, linking these variants to molecular pathways still remains challenging[2]. GWAS signals of common diseases predominantly fall into the non-coding genome[3] and both their enrichment in regulatory elements (e.g., quantitative trait loci (QTL)[3,4]), as well as advances in omics technology[5], have motivated the establishment of large-scale consortia providing publicly available QTL datasets for molecular phenotypes such as DNA methylation (DNAm)[6], transcript[7,8], protein[9–11] and metabolite[12,13] levels.

Integrative statistical methods combining GWAS and omics QTL summary data include colocalization tests[14,15], summary versions of transcriptome-wide association studies (TWAS)[16,17] and Mendelian randomization (MR) studies[18,19]. Colocalization methods identify shared QTL and GWAS signals, and while this might indicate causality between the molecular and GWAS trait, signal overlap can also arise due to reverse causality (i.e., causal effect of the GWAS trait on the molecular trait[20]) or horizontal pleiotropy (i.e., the identified shared genetic variant drives the molecular and trait perturbation independently). In comparison, MR studies, which are conceptually similar to TWAS, use multiple genetic variants as instrumental variables (IVs) and are less prone to reverse causality and artefacts arising from LD patterns[21] - although horizontal pleiotropy can never be ruled out entirely. In addition, MR analyses allow the quantification - direction and magnitude - of the causal effect of the omic on the outcome trait.

With the advent of QTL datasets with increased sample sizes[6,8], opportunities to integrate GWAS data with multiple molecular traits are no longer hampered by low statistical power. Previous efforts integrating multiple QTL omics data either adopted colocalization strategies[22,23] or combined pairwise MR associations (two-step MR)[24,25] testing only a single molecular mediator. Multivariable MR (MVMR) approaches have been proposed to identify multiple mediators of

[1]University Center for Primary Care and Public Health, Lausanne, Switzerland. [2]Swiss Institute of Bioinformatics, Lausanne, Switzerland. [3]Department of Computational Biology, University of Lausanne, Lausanne, Switzerland. [4]Center for Integrative Genomics, University of Lausanne, Lausanne, Switzerland. [5]These authors jointly supervised this work: Eleonora Porcu, Zoltán Kutalik. e-mail: marie.sadler@unil.ch; zoltan.kutalik@unil.ch

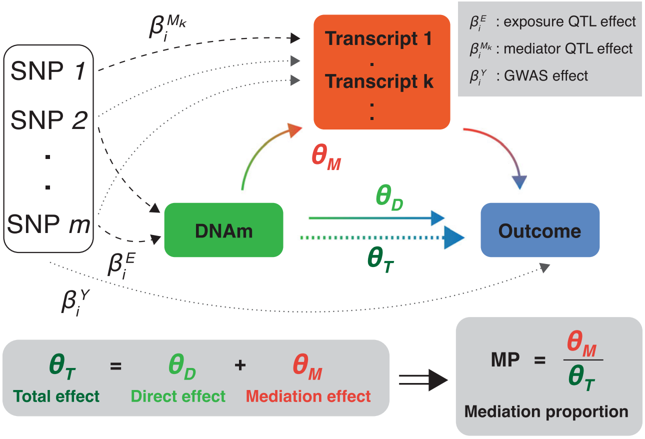


Fig. 1 | Overview of the three-sample multivariable Mendelian randomization (3S-MVMR) design to quantify mediation of complex traits through DNA methylation (DNAm) and transcripts. Genetic instruments (SNPs) are selected to be directly and significantly associated (dashed arrows) with either the exposure (DNAm, green) or any mediator k (transcript in *cis*, red). The total effect θ_T (green-blue dotted arrow) of the exposure on the outcome (complex trait, blue) is estimated in a univariable MR analysis based on exposure-associated SNPs only. The direct effect θ_D (green-blue arrow) is estimated in an MVMR analysis on all valid instruments. The mediation effect θ_M (green-red and red-blue arrows) results from the difference between θ_T and θ_D, and allows to calculate the mediation proportion (MP). The genetic effect sizes β on the exposure, mediator and outcome come from m/eQTL and GWAS summary statistics, respectively. Transcripts were required to be causally associated to the DNAm-exposure to be included as mediators.

exposure-outcome relationships[26,27]. These approaches enable the dissection of the total causal effect of an exposure on an outcome into a direct and indirect effect measured via mediators. Similar to classical MR approaches, the use of genetic instruments allows for robust causal inference and MVMR has proven to be an unbiased approach for mediation analyses, even in the presence of confounders[26,27]. Hence, in addition to identifying causal effects through multiple layers, MVMR allows the quantification of mediation effects.

Here, we propose a three-sample MVMR (3S-MVMR) framework to quantify the role of *cis*-transcripts in mediating DNAm → complex trait causal relationships (Fig. 1). To do so we integrated methylation and transcript QTLs (mQTLs and eQTLs, respectively) with GWAS summary data of 50 clinically relevant traits to estimate global mediation proportions (MPs), i.e., the proportion of transcript-mediated causal effect relative to the total effect of DNAm on complex traits. In contrast with previous multi-omics integration methods, each 3S-MVMR regression analysis makes use of at least 5 near-independent instrumental variables (IVs) allowing for more robust causal inference and post-hoc sensitivity analyses. We performed simulation studies to assess biases of the 3S-MVMR estimates for MP under various parameter settings. In addition to quantifying the regulatory connectivity between DNAm and transcript levels, we investigated underlying factors driving high MPs, and hypothesized several mechanistic pathways between DNAm, gene expression and complex traits.

Results

Overview of the methods

We performed univariable and multivariable MR to estimate total ($\hat{\theta}_T$) and direct ($\hat{\theta}_D$) causal effects, respectively, with MP (mediation proportion) estimates being calculated as the ratio of the indirect effect (i.e., mediated through the molecular mediators) to the total effect of the exposure on the outcome trait[28] (Fig. 1; Eqs. (1) and (3)). If weak genetic instruments can introduce a bias towards the null in a univariable MR setting[29], this bias can be in any direction for MVMR studies[30]. Both sample size and choice of instruments and mediators can introduce a bias in any direction[30], leading to under- or overestimations of the MP. To quantify these biases and assess the sensitivity of estimated $\widehat{\text{MP}}$s, we conducted simulation studies mimicking settings that emerge from real data applications (Methods; Supplementary Fig. 4).

We then applied our framework in a genome-wide screen to estimate $\hat{\theta}_T$ of DNAm sites on 50 outcomes and contrasted them to the effects not mediated by transcripts in *cis* ($\hat{\theta}_D$). Genetic effect sizes on the DNAm and transcript levels came from the largest publicly available mQTL and eQTL datasets, respectively, derived from whole blood[6,8]. MP estimates were then computed only for DNAm-trait pairs with significant Bonferroni-corrected $\hat{\theta}_T$ effects, grouped by trait, trait category and all pairs combined. We present MP results for DNAm-trait pairs with at least one mediator significantly associated to the exposure ("detectable mediation"), but also for pairs, including the ones without a significant causal effect on any potential transcript ("overall mediation"). The overall MP quantifies more accurately the role of *cis*-transcripts in mediating DNAm effects, as the restriction to only DNAm-trait pairs with a mediator could introduce a selection bias towards higher MPs. Additionally, we performed various sensitivity analyses on these MR results to assess the robustness of the MP estimates: assessing weak instruments (through conditional F-statistics), heterogeneity tests (through heterogeneity Q-statistics and leaving the strongest instrument out) and estimating bias due to by-chance signal overlap (through simulations).

Simulation results

We performed simulation studies to assess the bias in estimated MPs ($\widehat{\text{MP}}$) by exploring a wide range of realistic parameter settings which cover at least the interquartile range as observed in real data (Supplementary Figs. 4-5; Supplementary Tables 1-2; Methods). Using default settings (i.e., median values for each parameter such as 2 true mediators N_{med} and a true MP of 35%), the bias in $\widehat{\text{MP}}$ is minimal with the mean $\widehat{\text{MP}}$ equalling 33.5% (95% CI: [32.0%–35.0%]; Supplementary Fig. 6; Supplementary Table 2). A determining factor in accurately estimating MPs was the available sample size to derive the mediator QTL effects. Low sample sizes resulted in significant underestimations of the MP, with mediator sample size of 3000 compared to 30,000 resulting in a 17% relative decrease (6% in absolute values) of the estimated $\widehat{\text{MP}}$ (Fig. 2a). The reason for this significant underestimation was not only weak instrument bias, but also the omission of relevant mediators with on average only 1.17 ($N_{\text{med,sig}}$) out of the 2 (N_{med}) relevant mediators detected at a sample size of 3000 (Fig. 2a). We further tested the robustness of the $\widehat{\text{MP}}$ with respect to the number of included mediators by varying the mediator selection threshold P_{EM}. Among a set of 20 potential mediators, those not passing the P_{EM} as determined by univariable MR effects of the exposure on each of these mediators were excluded from the MVMR model (Methods). Using a too lenient or too stringent P_{EM} threshold resulted in downward biased $\widehat{\text{MP}}$s (Fig. 2b), as the former leads to the inclusion of too many non-mediators in the model (giving rise to weak instrument bias), while the latter case fails to include relevant mediators in the model. The used mQTL and eQTL datasets provide SNP effect sizes in *cis* of the assessed DNAm probe and transcript levels, respectively, and were primarily restricted to significant mQTLs for the former. Thus, in the MVMR analysis SNP-exposure effects for mediator instruments are often non-significant (hence unreported) and set to zero to reduce regression dilution bias (i.e., weak instrument bias). Our simulation studies, which mimicked this scenario by setting non-significant effects to zero (Methods), confirmed that this did not introduce any bias.

Furthermore, we investigated weak instrument bias of both exposure- and mediator-associated IVs. When mediator-associated IVs were weak (i.e., low direct mediator heritabilities ($h^2_{M,\text{direct}}$; Methods), a high variability and significant underestimation of the $\widehat{\text{MP}}$ was observed (Fig. 2c). In case of low mediator heritability, the conditional

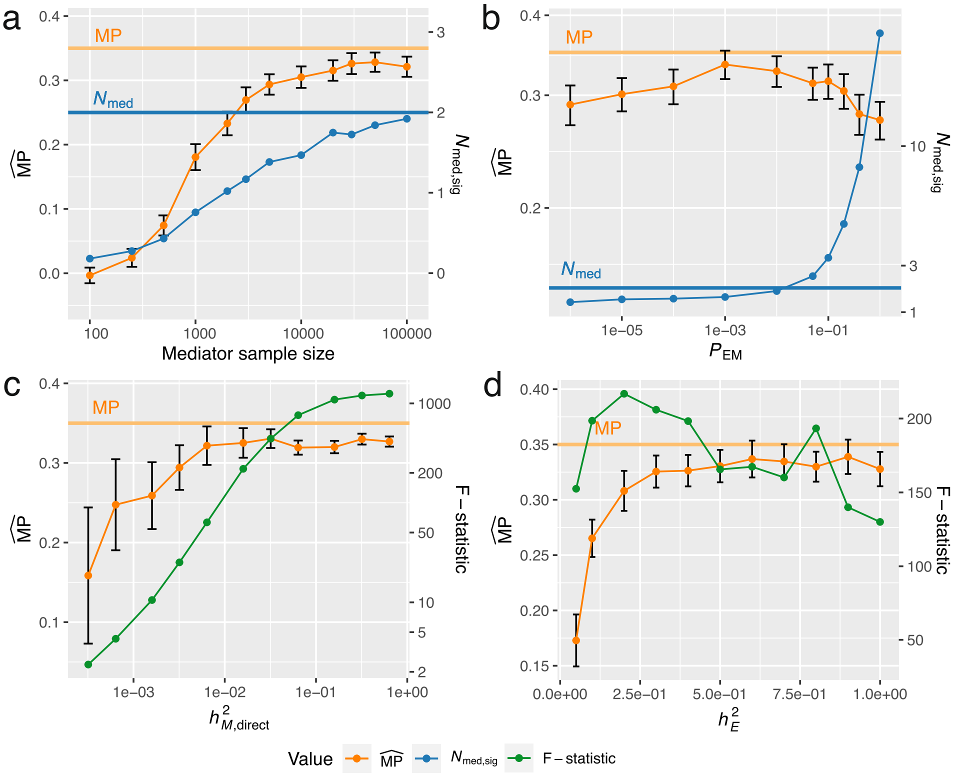


Fig. 2 | Simulation results to assess the bias in estimated mediation proportions ($\widehat{MP}$s) in real data settings. a Influence of the mediator sample size on the estimated $\widehat{MP}$ (orange) and number of selected mediators ($N_{med,sig}$, blue). **b** Influence of the mediator selection threshold P_{EM}. **c**, **d** Sensitivity of $\widehat{MP}$s in settings of weak instruments, simulated by low mediator ($h^2_{M,direct}$) and exposure (h^2_E) *cis*-heritabilities, respectively. Conditional F-statistics (green) of the exposure allow to test for weak instrument bias (critical values are defined at a threshold of F-statistic < 10). For a given parameter setting, 500 exposure-outcome pairs were simulated on which an $\widehat{MP}$ (orange points) and 95% CI (black error bars) were estimated. The true MP of the model was 0.35 (horizontal orange lines), and the true number of relevant mediators N_{med} was 2 (horizontal blue lines) among a set of 12 potential mediators (20 in **b**).

F-statistics of the exposure was also below the critical threshold of < 10 (Methods) indicating weak instruments. Similarly, for low exposure heritability (h^2_E), underestimated $\widehat{MP}$s were obtained, even in case of high conditional F-statistics (>120; Fig. 2d). Additional simulation studies with more polygenic exposures and increased number of relevant mediators N_{med} for different exposure and mediator heritabilities corroborated the findings of underestimated $\widehat{MP}$s in case of weak instruments (Supplementary Fig. 7).

Application to 50 complex traits

We first estimated the causal effects of DNAm probes on 50 complex traits, ranging from biomarkers indicative for diseases, such as low-density lipoprotein (LDL) and glucose levels, to diseases such as asthma and schizophrenia (Supplementary Data 1). DNAm-trait pairs with a significant total causal MR effect (P_T < 1e-6) were then further assessed to examine what fraction of the DNAm → trait causal effect is mediated by transcripts in *cis* (Fig. 3a; Supplementary Fig. 1). Mediation analyses could be conducted for 2069 pairs, for which at least 1 transcript was causally associated to the DNAm exposure (detectable mediation). First, we regressed $\hat{\theta}_D$ against $\hat{\theta}_T$ within each trait influenced by at least 10 DNAm probes while accounting for regression dilution bias[31] (Eq. (6)). $\widehat{MP}$s estimated for each of these 41 traits ranged from 18.0 to 78.0% (mean: 36.9%, 95% CI: [13.5%–60.3%]) with the trait with the highest $\widehat{MP}$ being grip strength and the one with the lowest testosterone level (Fig. 3b). Regressing $\hat{\theta}_D$ against $\hat{\theta}_T$ for all pairs combined yielded an $\widehat{MP}$ of 37.8% (95% CI: [36.0%–39.5%]) (Fig. 3c). Grouping the traits into 10 physiological categories (Supplementary Data 1) showed that the $\widehat{MP}$ was highest for hepatic biomarkers (mean: 46.6%, 95%CI: [41.5%–51.7%]), followed by renal biomarkers (mean: 43.5%, 95%CI: [37.5%–49.5%]). In contrast, adiposity-related and hormonal traits exhibited the lowest $\widehat{MP}$ (Fig. 3b; Supplementary Fig. 8).

In addition to the 2069 DNAm-trait pairs with detectable mediation, there were 554 pairs testable for mediation, but with no detectable causally implicated transcript (Fig. 3a). Setting $\hat{\theta}_D$ to $\hat{\theta}_T$ for these pairs and regressing $\hat{\theta}_D$ against $\hat{\theta}_T$ for all 2623 DNAm-trait pairs combined reduced the $\widehat{MP}$ to 28.3% (95% CI: [26.9%–29.8%]) (Fig. 3d). We refer to this $\widehat{MP}$ as the overall $\widehat{MP}$, as it is a more objective measure of the importance of the transcriptome in mediating DNAm-to-phenotype effects. While more reflective of mediated DNAm effects, it may also be overly conservative since the set of testable transcript mediators (N = 19,250[8]) is a magnitude lower than that of the whole transcriptome[32].

The average number of mediator transcripts, potentially correlated, was 3.3 per methylation-trait pair with detectable mediation, indicating that the impact of methylation is not mediated by a single transcript. To further explore this observation, we assessed the extent to which DNAm → trait effects were mediated by the single most significantly DNAm-associated transcript ("top" transcript; Methods), as opposed to all transcripts in *cis*. This resulted in an $\widehat{MP}_{top}$ of 26.0% (range: [13.0%–46.8%]) averaged across the 41 traits, and an $\widehat{MP}_{top}$ of 26.6% (95% CI: [25.1%–28.1%]) when aggregating the 2069 DNAm-trait pairs (Supplementary Fig. 9). This significant drop in the $\widehat{MP}$ (P_{diff} < 5e-21) corroborates our initial hypothesis that DNAm sites regulate the expression of multiple transcripts in the *cis* region.

MVMR sensitivity analyses

We conducted MVMR sensitivity analyses to assess potential sources of bias of the MP estimates such as weak instruments and pleiotropy.

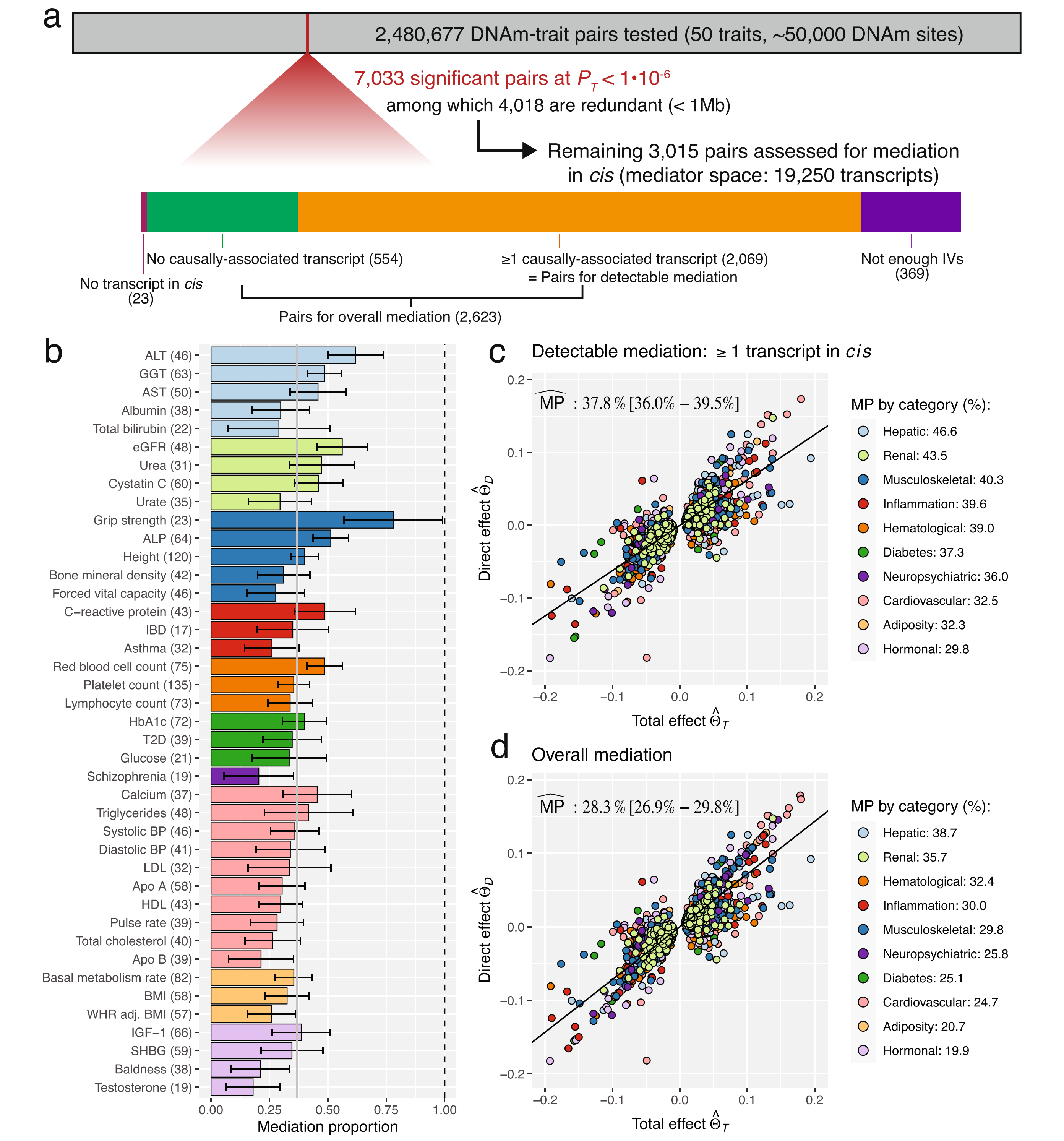


Fig. 3 | Mediation proportions ($\widehat{\text{MP}}$s) for transcripts in *cis* mediating DNAm-to-trait effects. a Flowchart describing the selection of DNAm-trait pairs retained for mediation analyses. Among the total of 2,480,677 pairs tested, 2069 pairs (orange) with a significant total causal effect (P_T) and at least 1 causally-associated transcript were assessed for mediation (**b**, **c**). 554 pairs without any transcript causally linked to the DNAm site (green) were included in the calculation of the overall mediation. Pairs without any transcript in the *cis* region (pink) were omitted in mediation analyses as were pairs without sufficient instrumental variables (IVs, purple). **b**$\widehat{\text{MP}}$s by trait where error bars denote the 95% CI, and the grey vertical bar shows the mean $\widehat{\text{MP}}$ across the traits ($\widehat{\text{MP}}$s per trait were derived by regressing $\hat{\theta}_D$ against $\hat{\theta}_T$). Only traits with ≥10 DNAm-trait pairs with detectable mediation are displayed (41 traits with number of evaluated pairs indicated in parentheses), colour-coded by their physiological category as defined in the legends of **c** and **d**. **c** Detectable mediation: All DNAm-trait pairs assessed in the mediation analyses (2069) with traits being grouped into 10 physiological categories. The global $\widehat{\text{MP}}$ in percentage with 95% CI is shown in the plotting area and individual category $\widehat{\text{MP}}$s in the legend. **d** Overall mediation: Same analysis as in **c**, but including all 2623 DNAm-trait pairs with at least 1 transcript present in the *cis* region. For these additional pairs, the direct effect was set to the total effect.

To test whether the MVMR estimates suffer from weak instrument bias, we calculated conditional F-statistics[33]. These statistics reflect whether genetic variants sufficiently explain the variance in the exposure given the presence of mediators. As demonstrated by Sanderson et al., direct effect estimates ($\hat{\theta}_D$) of exposure-trait pairs for which the F-statistic is ≤10 might be biased[33]. Among the 2069 DNAm-trait pairs, 1061 had an F-statistic > 10 with an $\widehat{\mathrm{MP}}$ of 35.5% (95% CI: [33.6%–37.5%]) which was not significantly lower than the one for all pairs combined (P_{diff}=0.09). Pairs with F-statistics ≤10 (N=1008) had significantly more mediators (4.32 vs 2.35, two-sided t-test: P=2.13e-64), but not a significantly higher $\widehat{\mathrm{MP}}$ (mean: 40.9%, 95% CI: [37.8%–44.0%]; P_{diff}= 0.08; Supplementary Figs. 10-11).

Pleiotropic IVs violate MR assumptions and heterogeneity tests, such as the Cochran's Q-statistic, can be used to detect them, assuming that most IVs are valid[34]. We calculated Q-statistics for the IV sets in both the univariable and multivariable MR analyses. Out of the 2069 DNAm-trait pairs, 1757 showed no signs of heterogeneity in the univariable MR analyses ($P_{\mathrm{HET}} > 0.01$) and 1405 in neither the univariable nor multivariable analyses. The $\widehat{\mathrm{MP}}$ of these 1405 pairs was not significantly different from the overall one (mean: 38.3%, 95% CI: [36.1%–40.6%]; P_{diff}=0.7; Supplementary Fig. 12).

Next, we assessed the influence of the p-value threshold P_{EM} to select mediators based on the exposure-to-mediator causal effect (default P_{EM}=0.01 for which N=2069 DNAm-trait pairs with at least 1 mediator were found). With a more lenient threshold (P_{EM}=0.05), more DNAm-trait pairs with mediators emerged (N=2189). Conversely, with a more stringent threshold (P_{EM}=0.001), less pairs were detected (N =1881). No differences in MPs between the three settings were found in these detectable mediation analyses ($P_{\mathrm{diff}} > 0.05$; Supplementary Fig. 13), but when calculating the overall MP (i.e., inclusion of all DNAm-trait pairs with potential transcript mediators in the cis-region) on a common set of DNAm-trait pairs (N=2543, $\widehat{\mathrm{MP}}_{\mathrm{overall,P01}}$=27.6% (95% CI: [26.1%–29.2%])), a significantly higher MP for the more lenient threshold ($\widehat{\mathrm{MP}}_{\mathrm{overall,P05}}$=32.0% (95% CI: [30.4%–33.6%]); P_{diff}=1.1e-4), and significantly lower MP for the more stringent threshold were observed ($\widehat{\mathrm{MP}}_{\mathrm{overall,P001}}$=24.6% (95% CI: [23.2%–26.1%]); P_{diff}=4.8e-3; Supplementary Fig. 14).

Finally, we conducted sensitivity analyses to determine whether significant MR associations were due to horizontal pleiotropy. Regulatory pathways between DNAm exposure probes and transcript mediators were assessed in cis. As such, SNPs in LD with significant QTLs for both quantities could give rise to an association merely because of horizontal pleiotropy (i.e., due to random overlap between cis-QTLs in close vicinity), an issue further exacerbated by the fact that molecular omics entities generally have fewer associated IVs than complex traits. To assess whether mediation results are only based on a single strong genetic instrument, we repeated the mediation analysis excluding the top IV (i.e., exposure-associated IV with the lowest p-value) from both the total effect θ_T and direct effect θ_D calculations (Methods). The results show that while MR effect estimates remain concordant in magnitude and effect direction, the estimates are noisier due to the much weaker instruments (significantly lower F-statistics; two-sided t-test: P=5.37e-11; Supplementary Fig. 15). MP estimates were also higher when the top IV was excluded ($\widehat{\mathrm{MP}}$ =47.3% (95% CI: [38.4%–56.2%]); P_{diff}=0.023; Supplementary Fig. 15), however, this was no longer the case when controlling for conditional F-statistics > 10 ($\widehat{\mathrm{MP}}$ =40.9% (95% CI: [29.3%–52.4%]); P_{diff}=0.48). Additionally, we performed simulation analyses to assess the possibility of significant DNAm-transcript associations caused by cis-mQTL and -eQTL signals being in LD (Methods). The analysis shows that randomly picked eQTL-SNPs in the region result in slightly inflated, but much weaker MR associations than using the original eQTL data (Supplementary Figs. 16-17). The results indicate that by-chance LD between cis-mQTLs and-eQTLs can yield false positive findings, but those signals are substantially weaker than the ones observed in real data. In other words, mQTL and eQTL IVs are in much higher LD than expected by chance.

Overall, these sensitivity analyses showed that the estimated MPs remain robust when removing DNAm-trait pairs that potentially violate MVMR assumptions, while also suggesting that the set P_{EM} threshold of 0.01 may lead to underestimated MP estimates. Finally, we found strong evidence that molecular associations mediating DNAm-trait effects are predominantly due to vertical pleiotropy, even when only a limited number of IVs were available.

Determining factors of mediation proportions

We explored underlying factors driving high MPs through transcript levels (Fig. 4a). $\widehat{\mathrm{MP}}_{\mathrm{top}}$ decreased with increased distances between the DNAm site and the gene transcription start site (TSS) of the top transcript ($\rho = -0.076$, $P = 5.2\mathrm{e}{-4}$; Fig. 4b). This distance was also negatively correlated to the DNAm-to-transcript MR squared effect size, α^2_{EM}, ($\rho = -0.13$, $P = 3.1\mathrm{e}{-19}$; Fig. 4c), which in turn was a good predictor for high MPs ($\rho = 0.39$, P = 2.5e-75; Fig. 4d). The mediation proportion was the highest for DNAm sites residing in the first exon, followed by those in the 5'UTR, within 200 bp of the TSS and lowest for those within 1500 bp of the TSS and in the gene body (Supplementary Fig. 18).

DNAm inhibiting the binding of transcription factors (TFs) thereby repressing gene expression is often alluded to as the classical mechanism of action for DNAm[35]. From the 1,066,307 unique DNAm-to-transcript causal effects assessed, 47,445 were significant at $P < 4.7\mathrm{e}{-8}$. Although negative effects had a larger magnitude than positive ones (two-sided t-test: $P = 0.0082$) only 53.4% of DNAm → transcript causal effects were negative. Stratifying DNAm sites with respect to their location on the assessed transcript, we found that DNAm sites situated in the first exon and nearby the TSS were enriched for negative effects (P = 2.7e-3, 1.2e-5 and 3.8e-4 for 1st exon, TSS ± 1500 bp and TSS ± 200 bp, respectively), whereas those in the gene body were enriched for positive ones (P = 2.2e-10; Supplementary Table 3). These observations are in line with previous studies that only showed a slight trend for negative methylation-gene expression correlations[36–39]. We further tested whether the MR DNAm-to-transcript causal effects correlated with reported methylation-transcript correlations[37] and found a strong agreement (ρ = 0.39, P = 2.6e-18, 471 DNAm-transcript pairs).

Consistent with higher MPs when mediating through multiple transcripts, we found a strong correlation between the number of mediators and the MP (ρ = 0.39, P = 4.4e-75; Fig. 4e). Many of these mediators were correlated amongst each other, which in theory should be accounted for by the MVMR model. To ensure that this was the case, we repeated the mediation analysis with uncorrelated mediators ($R_{\mathrm{med}} < 0.3$; Methods). The mean number of selected mediators dropped by more than half, from 3.3 to 1.2 (Supplementary Fig. 19), and the $\widehat{\mathrm{MP}}$ across all DNAm-trait pairs decreased ($\widehat{\mathrm{MP}}_{\mathrm{uncorrelated}}$ = 30.5% (95% CI: [28.8%–32.1%])), while remaining significantly higher than $\widehat{\mathrm{MP}}_{\mathrm{top}}$ (P_{diff} = 6.6e-4). Decreasing the R_{med} threshold to 0.2 and 0.1 did not significantly decrease $\widehat{\mathrm{MP}}_{\mathrm{uncorrelated}}$ ($P_{\mathrm{diff}} > 0.05$), which stabilized at 29.2% (95% CI: [27.5%–30.8%]) for $R_{\mathrm{med}} < 0.1$ (Supplementary Fig. 19).

Furthermore, we investigated whether $\widehat{\mathrm{MP}}$s are dependent on the DNAm → transcript causal effect directions following the logic of a recent DNAm-transcript correlation study[39]. To this end, we stratified DNAm-trait pairs by the α_{EM} sign and number of mediators (Table 1). If there was only a single mediator, $\widehat{\mathrm{MP}}$s were significantly higher if the DNAm was decreasing expression (P_{diff} = 3.49e-8). This is consistent with the observation that negative effects α_{EM} were larger than positive ones and the positive correlation between α_{EM} magnitudes and high $\widehat{\mathrm{MP}}$s (Fig. 4d). When there were multiple mediators, most DNAm sites had negative effects on some transcripts and positive effects on others. These bivalent DNAm probes exhibited the highest $\widehat{\mathrm{MP}}$s

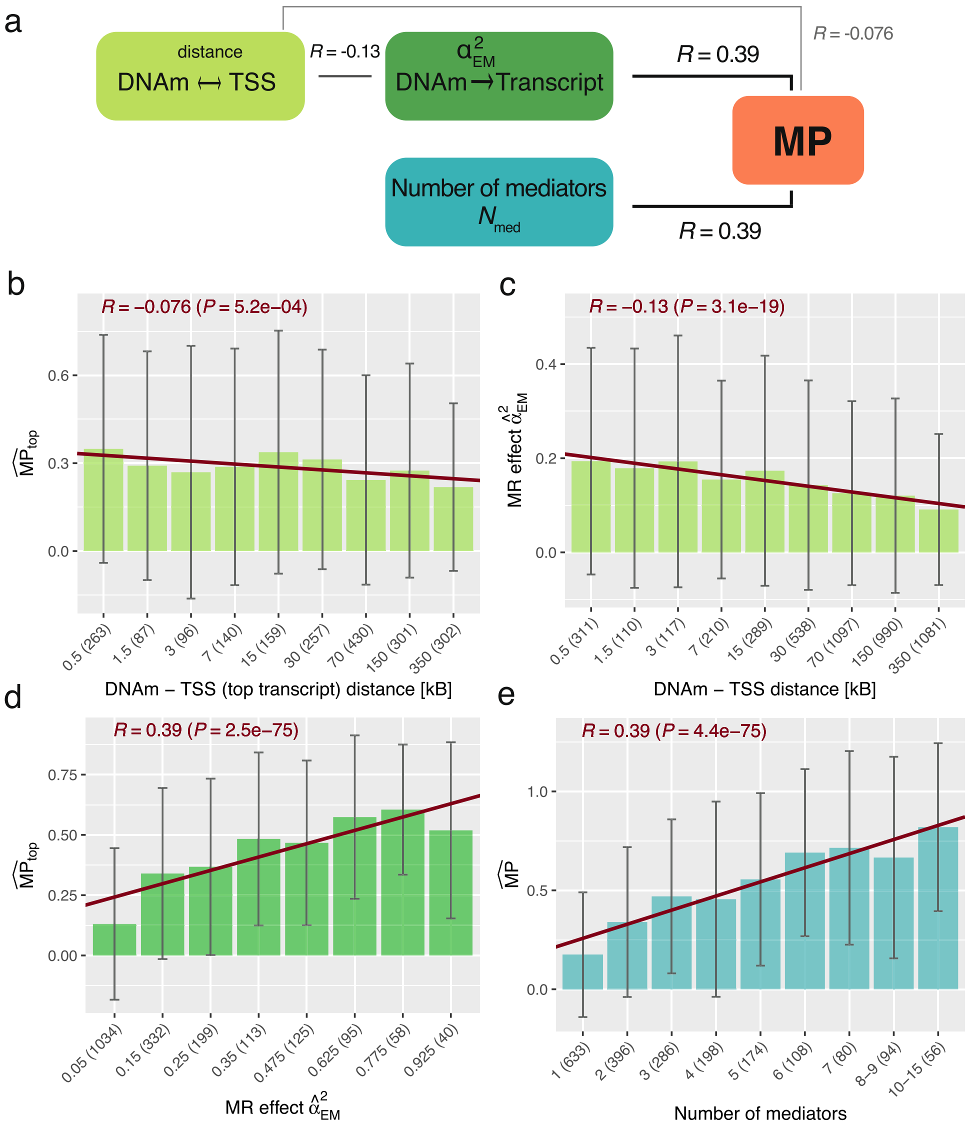


Fig. 4 | Exposure-to-mediator regulatory strength and number of mediators explaining mediation proportions (MPs). **a** Summary of the correlations (R) between MP (red) and DNA methylation (DNAm)-to-transcript causal MR effects ($\hat{\alpha}^2_{\mathrm{EM}}$, dark green), distance between the DNAm site and transcription start site (TSS, light green) and number of mediators (N_{med}, blue). **b** Average MP through top transcript ($\widehat{\mathrm{MP}}_{\mathrm{top}}$) of DNAm-trait pairs stratified according to the distance between the DNAm site and the TSS of the top transcript. All DNAm-trait pairs with at least one mediator were included (2069 pairs). **c** Average MR causal effects ($\hat{\alpha}^2_{\mathrm{EM}}$) of DNAm-transcript pairs stratified according to the distance between the DNAm site and the TSS. Unique DNAm-transcript mediator pairs across all DNAm-trait pairs were included (4743 pairs). **d** Average $\widehat{\mathrm{MP}}_{\mathrm{top}}$ of DNAm-trait pairs stratified according to DNAm-to-top transcript MR causal effect size $\hat{\alpha}^2_{\mathrm{EM}}$. All DNAm-trait pairs with at least one mediator were included (2069 pairs). **e** Average $\widehat{\mathrm{MP}}$ of DNAm-trait pairs stratified according to the number of mediators. All DNAm-trait pairs with at least one mediator were included (2069 pairs). The reported p-values (P) for the corresponding Pearson correlations (R) arise from a two-sided t-test and were calculated between the two respective quantities on DNAm-trait/DNAm-transcript pairs prior stratification. Bin height represents mean within each bin and the error bars the corresponding standard deviations (number of evaluated pairs within each bin is indicated in parentheses). The red slope represents the regression fit between the bin's positions and heights, and serves merely for visualization purposes.

($\widehat{\mathrm{MP}}$ = 53.9% (95% CI: [51.2%–56.5%])) - a consequence of being causally associated to more mediators than average (5.01 vs 3.31), with N_{med} being a strong predictor for high $\widehat{\mathrm{MP}}$s (Fig. 4e). Combining DNAm-trait pairs with single and multiple mediators, but with consistent negative or positive α_{EM} values, the observation of higher $\widehat{\mathrm{MP}}$s when DNAm was decreasing transcript levels persisted (P_{diff} = 0.020).

Putative regulatory mechanisms of action

In addition to providing insights into global patterns governing the mediation between different intermediate phenotypic layers and functional traits, our analyses generated plausible hypotheses regarding specific biological pathways. We chose to follow-up putative regulatory mechanisms of DNAm-to-complex traits through transcript levels which showed both strong total effects ($|\hat{\theta}_T|$>0.02) and substantial mediation proportion ($\widehat{\mathrm{MP}}$>0.2; complete list in Supplementary Data 2).

Involvement of the anti-oxidant and anti-inflammatory protein PARK7 in inflammatory bowel disease (IBD) has recently been brought to light[40–43]. While the exact role of the protein in the disease remains debated, reduced intestinal expression of *PARK7* was observed in patients and mouse models for IBD[43]. Moreover, *Park7* knockout mice were shown to have increased levels of pro-colitis bacterial species in their microbiome[42,44] and experience aggravated symptoms of experimentally-induced colitis[43]. In line with these observations, DNAm of the *PARK7* promoter probe cg10385390 (chr1:8'022'505) decreased *PARK7* transcript expression ($\hat{\alpha}_{\mathrm{EM}}$ = −0.675, P = 2.7e-4;

Table 1 | Exposure-to-mediator effect direction and number of mediators explaining MPs

	Negative	Positive	Bivalent	Total
Mono	$N = 370$ (17.9%)	$N = 276$ (13.3%)	0, by definition	$N = 646$ (31.2%)
	$\widehat{MP}$ =20.8%	$\widehat{MP}$ =7.97%		$\widehat{MP}$ =14.7%
	(95% CI: [17.3%–24.2%])	(95% CI: [5.04%–10.9%])		(95% CI: [12.3%–17.1%])
Multi	$N = 239$ (11.6%)	$N = 207$ (10.0%)	$N = 977$ (47.2%)	$N = 1423$ (68.8%)
	$\widehat{MP}$ =42.8%	$\widehat{MP}$ =42.8%	$\widehat{MP}$ =53.9%	$\widehat{MP}$ =50.3%
	(95% CI: [38.3%–47.2%],	(95% CI: [38.3%–47.3%],	(95% CI: [51.2%–56.5%],	(95% CI: [48.2%–52.4%],
	mean (N_{med})=3.01)	mean (N_{med})=2.84)	mean (N_{med})=5.01)	mean (N_{med})=4.35)
Total	N=609 (29.4%)	N =483 (23.3%)	N=977 (47.2%)	N=2069 (100%)
	$\widehat{MP}$ =27.7%	$\widehat{MP}$ =22.8%	$\widehat{MP}$ =53.9%	$\widehat{MP}$ =37.8%
	(95% CI: [24.8%–30.5%],	(95% CI: [19.8%–25.8%],	(95% CI: [51.2%–56.5%],	(95% CI: [36.0%–39.5%],
	mean (N_{med})=1.79)	mean (N_{med})=1.79)	mean (N_{med})=5.01)	mean (N_{med})=3.31)

DNAm-trait pairs were stratified by the number of mediators (N_{med}; "mono" if $N_{med} = 1$ and "multi" if $N_{med} > 1$) and by the exposure-to-mediator α_{EM} causal effect sign ("Negative" and "Positive" for DNAm decreasing and increasing transcript levels, respectively, and "Bivalent" if a given DNAm site is affecting transcript levels in both directions). For each stratum, the number of DNAm-trait pairs (N), the estimated mediation proportion ($\widehat{MP}$) and mean N_{med} is shown.

Fig. 5a). High transcript levels decrease IBD risk ($\hat{\alpha}_{MY} = -0.131$, $P = 1.7\text{e-}7$) resulting in an overall increased IBD risk upon DNAm ($\hat{\theta}_T = 0.114$, $P = 8.2\text{e-}9$).

Despite often being associated with decreased expression[35], our data provides examples of methylation boosting expression. For instance, DNAm of cg13428477 (chr3:122′748′086) increased *PDIA5* expression ($\hat{\alpha}_{EM} = 0.333$, $P = 7.3\text{e-}11$), whose levels subsequently increased platelet count ($\hat{\alpha}_{MY} = 0.062$, $P = 0.018$), so that DNAm resulted in significantly increased platelet count ($\hat{\theta}_T = 0.056$, $P = 1.3\text{e-}43$) (Fig. 5b). Association between the *PDIA5* locus and platelet count was reported through GWAS[45]. Platelets are small cell fragments produced by megakaryocytes, which themselves are derived from hematopoietic stem cells. Accordingly, *PDIA5* has a binding site for the hematopoietic stem and progenitor cell TF MEIS1[46] and is overexpressed in megakaryocytes as compared to other blood cell types[47]. Further studies showed that *pdia5* protein knockdown in zebrafish resulted in strongly decreased platelet count[48], matching our findings and confirming the role of *PDIA5* in thrombopoiesis.

In another example, we observed that DNAm of cg09070378 (chr1:161′183′762) decreased asthma risk ($\hat{\theta}_T = -0.031$, $P = 8.1\text{e-}11$) by reducing *FCER1G* expression ($\hat{\alpha}_{EM} = -1.0$, $P = 3.5\text{e-}18$), a gene listed in the KEGG pathway for asthma (hsa05310) and whose expression associated with an increased risk for asthma ($\hat{\alpha}_{MY} = 0.019$, $P = 3\text{e-}12$) (Supplementary Fig. 21). The *FCER1G* promoter was found to be hypomethylated in patients with atopic dermatitis, with DNAm levels correlating negatively with the gene's expression[49], suggesting a broad role of *FCER1G* in allergic disorders. Our data also supports and provides a mechanistical explanation for the recent finding that reduced *IFNAR2* expression causally decreases the odds of severe coronavirus disease 2019 (COVID-19)[50,51], which was later supported by the increased susceptibility for severe COVID-19 in individuals with rare loss-of-function mutations in *IFNAR2*[52]. Indeed, we found that DNAm of the *IFNAR2* promoter probe cg13208562 (chr21:34′603′264) decreased the gene's expression ($\hat{\alpha}_{EM} = -0.446$, $P = 2.4\text{e-}19$) (Supplementary Fig. 22). As *IFNAR2* expression protects against hospitalization following COVID-19 infection ($\hat{\alpha}_{MY} = -0.090$, $P = 4.2\text{e-}6$), DNAm of the locus increased the risk of severe infection ($\hat{\theta}_T = 0.064$, $P = 8.5\text{e-}13$).

Discussion

We presented a framework to quantify mediation of complex trait-impacting effects through an omics layer and demonstrated its application to assayed blood-derived DNAm (exposure) and transcript levels (as mediator). Evidence for mediation of DNAm-to-trait effects through transcripts in *cis* was found to be at least 28.3% for the 2623 DNAm-trait pairs with significant total causal effects that could be assessed. While many robust methods are available for univariable MR, it is not the case for MVMR[26,27]. Still, we could confirm the robustness of our MVMR estimates through various sensitivity analyses (conditional F-statistic, heterogeneity Q-statistic, excluding the strongest IV) that could not pinpoint any factor drastically biasing our MP estimates. Importantly, simulation studies indicated that MP estimates were likely to be lower bounds. Low sample size was shown to lead to MP underestimations, as do weak instruments, both for exposure- and mediator-associated IVs.

Additionally, we quantified the causal connectivity and directionality between DNAm and transcript levels and its impact on MPs. We found that 46.6% of significant DNAm-to-transcript effects were of positive sign (i.e., DNAm increasing transcription), particularly so when the DNAm site was situated in the gene body ($P_{Enrichment}$=2.15e-10). Interestingly, MPs were higher when DNAm was downregulating rather than upregulating transcripts. Previous genome-wide methylation and gene expression association studies reported high fractions of positive correlations (30−41%)[36,37,39] and further investigations indicated that our estimated methylation-to-transcript causal effects agree strongly with the respective correlations reported by Grundberg et al. (P=2.6e-18). While poorly understood[38], several mechanisms have been proposed to explain the phenomenon of DNAm induced transcription: preferential binding of some transcription factors to methylated DNA[53,54], prevention of repressor binding indirectly leading to increased expression through looping DNA[24,55], or DNAm in the gene body promoting elongation efficiency and preventing spurious initiation of transcription[56]. Furthermore, MP estimates indicated that DNAm sites typically regulate multiple transcripts in *cis* and that mediation through transcripts decreased the further away the TSS of the mediator transcript was from the DNAm site. Collectively, these results describe a more diverse picture of the transcription machinery, going beyond the classical views that DNAm solely reduces gene expression in the TSS region.

Statistical methods to integrate GWAS with omics data have seen a surge in recent years. Namely, colocalization methods based on a single genetic signal or corroborated by a secondary one, as well as methods supported by the SMR HEIDI statistic have been previously used in the study of DNAm-to-complex trait effects[6,14,24]. In the most recent publication of the GoDMC consortium, the former strategy was applied to systematically evaluate DNAm and GWAS co-localizing signals and compare them to MR[6]. This revealed a relatively poor overlap between colocalization and MR results, as both approaches have their weaknesses in detecting causal relationships. The major weakness of colocalization analysis is that it cannot detect directionality and does not estimate causal effect size. Colocalization of local association signals of two traits may be due to causal effects in either direction, common local confounder effect (e.g., shared regulatory mechanism) or causal markers in very high LD. Lack of colocalization can happen even if there is a true causal relationship, but there are additional associations impacting only the outcome trait. On the other hand, the major weakness of MR is that it may falsely detect a causal relationship when the causal variants for each trait are in reasonably high LD. The comparison of these two approaches is out of the scope of this work, but to explore the above-mentioned weakness in our study, we performed simulations tailored to detect by-chance overlaps in the association signals for methylation and gene expression (see pleiotropy sensitivity analyses for details). These analyses indicated that indeed elevated false positive rates are expected for MR, but the resulting MR

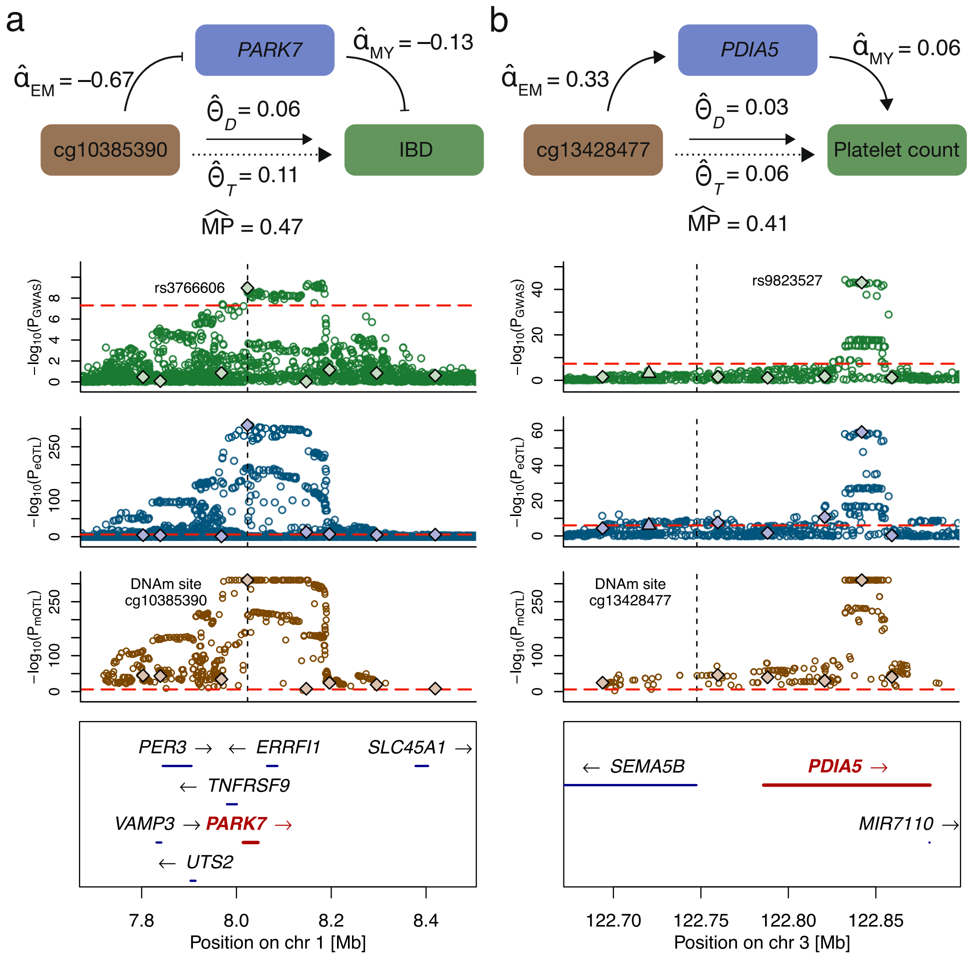


Fig. 5 | Plausible DNAm-transcript-trait regulatory mechanisms. a Mechanism involving DNAm probe cg10385390, *PARK7* and inflammatory bowel disease (IBD). **b** Mechanism involving DNAm probe cg13428477, *PDIA5* and platelet count. The top row displays a schematic of the mechanism with estimated univariable (total effect $\hat{\theta}_T$, DNAm-to-transcript effect $\hat{\alpha}_{EM}$ and transcript-to-outcome effect $\hat{\alpha}_{MY}$) and multivariable (direct effect $\hat{\theta}_D$) MR effects (displayed mediation proportions ($\widehat{MPs}$) are derived from $\hat{\theta}_D$ and $\hat{\theta}_T$ estimates). The three following rows show the regional SNP associations (-log$_{10}$(p-values)) with the trait (GWAS, green), transcript (eQTL, blue) and DNAm (mQTL, brown) probe, respectively. Solid diamonds represent DNAm-associated instruments used in the univariable (for $\hat{\theta}_T$ calculation) and multivariable (for $\hat{\theta}_D$ calculation) MR analyses. Upwards pointing triangles are transcript-associated SNPs that were additionally included in the MVMR instrument set. Red dashed lines indicate the significance thresholds of the respective SNP associations and the vertical black dashed line represents the DNAm probe position. Bottom row illustrates the positions and strand direction of the genes in the locus.

p-values under the null are much less significant than the ones observed for real methylation-transcript data.

Mapping genetic variants identified by GWASs to biological processes is notoriously difficult[2]. In particular, a challenge in identifying causal chains through omics layers is the attenuation in the genetic association strengths when moving up along layers. In a linear model, the genetic effect on the phenotype is assumed to be the product of causal effects between the preceding layers and it was previously shown that the variance explained by the top associated QTL of the first layer weakens with each successive omics layer[24]. In line with this observation, the examples depicted in Fig. 5 visualize the decrease in the genetic associations from the DNAm to the complex trait level. While in the future our 3S-MVMR framework could be applied to further mediating layers (e.g. proteins or metabolites), current QTL datasets for these omics layers lack the dimensionality - both in terms of sample size and number of assessed entities. Once larger datasets become available, these could be used to support mechanistic findings resulting from transcript data.

While our method highlights candidate pathways and provides MP estimates, several limitations are to be considered. First, our MP estimates are based on a selection of 2623 DNAm-trait pairs with significant total effects (P_T < 1e-6), which inherently focuses on DNAm-trait pairs with larger (and hence detectable) effects. In theory, MPs could depend on the magnitude of the total causal effect, thus the reported MP may differ for weaker total effects. A special case of these weaker total effects is when direct and indirect effects differ in sign, leading to a weak total effect with an MP potentially outside the [0,1] range. Furthermore, selected DNAm sites were those with the strongest DNAm-trait signal in their region (up to 1Mb). Thus, we omit secondary methylation signals, which may be mediated by transcripts to a different degree. Second, as for all MR-IVW approaches, included IVs might be pleiotropic, i.e., violating MR assumptions and potentially biasing effect estimates. Although, filtering out DNAm-trait pairs with signs of heterogeneous IV sets did not change MP estimates, the presence of invalid IVs cannot be entirely excluded and could therefore compromise causal effect estimates[57,58]. In particular, since selected IVs are in *cis* of the investigated molecular trait, they might be based on a single (pleiotropic) haplotype signal. Third, we select mediators based on their association to the exposure without taking into account their mediator potential, i.e., whether or not the mediator is additionally causally linked to the trait. Phrased differently, selected mediators are simply candidates and such selection serves as a first filter to remove non-mediators. In line with our simulations, it has been shown that an extremely large number of such "false" mediators (88 out of 92) can cause MVMR regression models to fail[30], indicating that our framework is less suitable for large numbers of molecular mediators unless the selection threshold P_{EM} is made more stringent. Finally, while molecular mechanisms ought to be tissue- or even cell type-specific, QTL data used in this study were derived from whole blood. However, not correcting for blood cell types when analyzing gene expression data can introduce important artefacts[59]. It is also known that different tissues express different isoforms[60], with many splicing and expression QTLs shown to differ across tissues[61]. Accordingly, MPs for blood biomarkers were generally higher than those for diseases, for which blood might not be the most relevant tissue. Differences between biomarker and disease MPs might also be due to the fact that indirect pathways, through unmeasured mediators, play a greater role for the latter trait category. Once tissue-stratified multi-omics datasets of larger sample size become available, more accurate, and potentially higher MPs will be obtained in trait-relevant tissues.

To conclude, by adapting existing MVMR mediation techniques to molecular exposures and mediators, we quantified the causal connectivity between DNAm and transcript levels, and their importance in shaping complex traits. Overall, we found solid evidence that almost a third of DNAm-to-complex trait effects are mediated by transcripts in *cis*. Our integrative omics framework can be extended to other omics-GWAS combinations and provide a powerful tool for mapping GWAS signals to biological pathways and prioritizing functional follow-up experiments.

Methods

Univariable and multivariable Mendelian randomization

Univariable Mendelian randomization (MR) was applied to estimate the total causal effect (θ_T) and multivariable MR (MVMR) to estimate the direct causal effect (θ_D) of an exposure E on an outcome Y. The mediation proportion (MP) was defined as $1-\theta_D/\theta_T$. Under the MR assumptions, genetic variants G used as IVs must be i) associated with E, ii) independent of any confounder of the $E-Y$ relationship, iii) conditionally independent of Y given E. We analysed exposures with at least five LD-pruned ($r^2 < 0.05$) IVs associated ($P < 1\text{e-}6$) with the molecular exposure and located in *cis* (< 1 Mb). To estimate θ_T we used the inverse-variance weighted (IVW) MR method, while accounting for (mildly) correlated instruments[19,62] as follows:

$$\hat{\theta}_T = (\boldsymbol{\beta}_E'\mathbf{C}^{-1}\boldsymbol{\beta}_E)^{-1}\boldsymbol{\beta}_E'\mathbf{C}^{-1}\boldsymbol{\beta}_Y \tag{1}$$

where $\boldsymbol{\beta}_E$ and $\boldsymbol{\beta}_Y$ are vectors of genetic effect sizes obtained from summary statistics for E and Y, respectively. $\mathbf{C}$ is the linkage disequilibrium (LD) matrix with pairwise correlations between IVs estimated from the UK10K reference panel[63]. Sensitivity analyses confirmed that accounting for the LD-matrix safeguards against MR estimates being influenced by the pruning threshold r^2 (Supplementary Figs. 2-3). Since in the following MVMR model more IVs than mediators are required, we chose a more lenient pruning threshold ($r^2 < 0.05$), including IVs in mild LD (Supplementary Fig. 2). Prior to the causal effect calculations, IVs were Steiger-filtered to avoid that the IV's effect on Y is significantly larger than it is on E[64] and were thus required to pass a threshold $t_{\mathrm{rev}} < \frac{|\beta_{E_i}|-|\beta_{Y_i}|}{\sqrt{\mathrm{var}(\beta_{E_i})+\mathrm{var}(\beta_{Y_i})}}$ with t_{rev} set at −2, equivalent to a one sided test p-value threshold of 0.023[34]. IVs not passing this threshold are prone to violating the third MR assumption of horizontal pleiotropy since they are more directly linked to the outcome. As a result, MR estimates including such IVs would potentially mix up forward and reverse causal effects. The standard error (SE) of θ_T can be approximated by the Delta method[65]:

$$\mathrm{SE}(\hat{\theta}_T) = \sqrt{(\boldsymbol{\beta}_E'\mathbf{C}^{-1}\boldsymbol{\beta}_E)^{-1}\boldsymbol{\beta}_E'\mathbf{C}^{-1/2}\boldsymbol{\Sigma}\mathbf{C}^{-1/2}\boldsymbol{\beta}_E(\boldsymbol{\beta}_E'\mathbf{C}^{-1}\boldsymbol{\beta}_E)^{-1}} \tag{2}$$

where $\boldsymbol{\Sigma}$ is a diagonal matrix with each diagonal element i equalling the maximum of the regression variance s^2 and $\mathrm{var}(\beta_{Y_i})$[34].

Through the inclusion of mediators M_k and their associated *cis* genetic variants ($r^2 < 0.05, P < 1\text{e-}6$), θ_D can be estimated analogously to θ_T using a multivariable regression model[28] as the first element of $\boldsymbol{\theta_D}$:

$$\hat{\boldsymbol{\theta}}_{\boldsymbol{D}} = (\boldsymbol{B}'\mathbf{C}^{-1}\boldsymbol{B})^{-1}\boldsymbol{B}'\mathbf{C}^{-1}\boldsymbol{\beta}_Y \tag{3}$$

where $\boldsymbol{B}$ is a matrix with $k+1$ columns containing the effect sizes of the IVs on the exposure in the first column and on each mediator in the subsequent columns. The remaining elements of $\boldsymbol{\theta_D}$ represent the direct effects of the mediators on the outcome and were referred to as $\alpha_{\mathrm{MY},k}$. In the estimation of MPs, we were not interested in $\alpha_{\mathrm{MY},k}$ values per se, but we took these effect sizes into account for inferring molecular mechanisms. If the number of mediator-associated instruments was sufficient (≥3) to conduct a univariable MR from the mediator to the outcome, we estimated $\alpha_{\mathrm{MY},k}$ from this analysis instead. In fact, the (marginal) contribution of an individual mediator can be better disentangled in univariable analyses, when mediators are highly correlated.

As our MVMR model assumes a chain of causal effects from the exposure to the mediator and then to the outcome, we conducted several Steiger filtering steps to reduce biases due to reverse causation. Although it has been proposed that DNAm could be a consequence of gene expression in the same locus[66], our model investigates the commonly assumed concept of DNAm regulating gene expression. In addition to meeting the Steiger criterion described above, exposure-associated IVs were required to pass that same threshold t_{rev} of no larger mediator than exposure effects for each of the mediators M_k. Similarly, to mitigate reverse causal effects from the outcome on the mediators, mediator-associated instruments with larger Y than M effects were removed if not passing the t_{rev} threshold. The SE of $\hat{\theta}_D$ was derived analogously to the univariable form (Eq. (2)) as shown in[19].

MVMR sensitivity analyses

Conditional F-statistic. Conditional F-statistics of the exposure were calculated following the approach of Sanderson et al.[33]. This method involves the regression of the exposure on the mediators based on the IV effect sizes on each of these quantities. The residuals of this regression are then used to derive the conditional F-statistic. The original method additionally includes the phenotypic correlation matrix between the exposure and mediators, which we omitted by default due to the lack of these data and thus used the identity matrix instead. However, as a sensitivity analysis, we calculated conditional F-statistics incorporating the phenotypic correlations between transcript mediators. Transcript correlations were calculated on RNAseq data from the Cohorte Lausannoise (CoLaus) based on 555 samples[67]. Transcript correlations could be estimated for 19,517 transcript of which 15,021 overlapped with the eQTL dataset (Methods: Omics and trait summary statistics)[8]. We then calculated conditional F-statistics that included mediator correlations for all DNAm-trait pairs with at least 2 mediators and for which at least half of them had available correlation data. Conditional F-statistics > 10 allow to reject the null hypothesis that the IVs are too weak to reliably estimate the multivariable effect of the exposure in the presence of the mediators.

Heterogeneity Q-statistic. Heterogeneity Q-statistics were computed as implemented in the TwoSampleMR package (v0.5.6, IVW-method)[34]. This test statistic quantifies the deviation of MR effect estimates of each individual IV from the IVW-estimate based on all IVs[68]. The null hypothesis of homogeneity within the IV set follows a chi-squared distribution with $m-1$ degrees of freedom for the univariable MR, and $m-k$ degrees of freedom for the MVMR, where m is the number of IVs and k the number of mediators.

Mediator selection threshold P_{EM}. For transcripts to be included as mediators in the MVMR regression model they had to be i) in *cis* of the DNAm exposure probe (±500kb) and ii) causally associated to the DNAm probe. This latter condition was verified by univariable MR analyses (Eq. (1)) of the DNAm exposure probe on each mediator transcript k in the region estimating the effect sizes $\alpha_{EM,k}$ and p-values $P_{EM,k}$. Transcripts satisfying $P_{EM,k} < P_{EM}$ were included as mediators with the default threshold equalling 0.01. To assess the sensitivity of this threshold, we also tested milder and more stringent thresholds (P_{EM}=0.05 and 1e-3).

Pleiotropy sensitivity analyses. To quantify whether significant MR estimates between the exposure and mediators were observed due to horizontal pleiotropy, we conducted two sensitivity analyses. First, we repeated the mediation analysis excluding the top IV (i.e., exposure-associated IV with the lowest p-value) from both the total effect θ_T and direct effect θ_D calculations. This analysis allowed to assess whether mediation results are solely driven by a single strong IV. Second, we performed simulation analyses to quantify the possibility that causal links between DNAm probes and transcripts are driven by increased horizontal pleiotropy stemming from potential LD between methylation and transcript instruments due to their close genomic distance.

In the following, we outline step-by-step the workflow of the horizontal pleiotropy simulation study for which a schematic representation is shown in Supplementary Fig. 16. First, we considered DNAm-transcript pairs with a significant MR effect at $P_{EM} < 1e\text{-}6$. For each of these selected DNAm-transcript pairs, we first fixed the SNP-DNAm and SNP-transcript effects as observed in the data. Then, using near-independent significant *cis*-eQTLs ($r^2 < 0.05$, $P < 1e\text{-}6$) with observed marginal (univariable) effect sizes $\boldsymbol{\beta_M}$ (a vector of size m_M) and the corresponding pair-wise local LD matrix $\mathbf{C}_M$, we calculated the multivariable SNP effects on the transcript, $\boldsymbol{\beta}_{\text{multi}}$, as:

$$\boldsymbol{\beta}_{\text{multi}} = \mathbf{C}_M^{-1}\boldsymbol{\beta}_M \tag{4}$$

Using the original data, we performed DNAm-transcript MR on m_E exposure (i.e., DNAm-associated) IVs, yielding the causal effect α_{EM} with corresponding p-value, P_{EM}. We then performed simulation analyses as follows to obtain MR effects for a hypothetical transcript with identical multivariable eQTL effect size distribution as the real transcript. To achieve this, for each simulation j, we randomly selected m_M leniently pruned ($r^2 < 0.5$) SNPs and assigned $\boldsymbol{\beta}_{\text{multi}}$ as their multivariable eQTL effects. Hence the marginal SNP-transcript effects for the m_E exposure-associated SNPs can be calculated as follows:

$$\boldsymbol{\beta}_{\text{marginal},j} = \mathbf{C}_{E,M_j}\boldsymbol{\beta}_{\text{multi}} \tag{5}$$

where $\mathbf{C}_{E,M_j}$ is the LD-matrix between the m_E exposure-associated SNPs and the m_M randomly chosen SNPs (with multivariable SNP-transcript effect $\boldsymbol{\beta}_{\text{multi}}$). This way we assign marginal SNP-transcript effect sizes for the m_E exposure-associated instruments, while keeping the multivariable eQTL effect size distribution identical to the one observed for the real transcript (but they are assigned to other SNPs). Univariable DNAm-transcript MR analyses could then be conducted (Eq. (1)) for each hypothetical transcript j, by using $\boldsymbol{\beta}_{\text{marginal},j}$ as the outcome effect size vector. Thus, we generated MR estimates ($\alpha_{EM,j}$ and $P_{EM,j}$) for 100,000 (N_{sim}) hypothetical transcripts for 100 randomly selected DNAm-transcript pairs throughout the genome. The simulation p-value was then derived as $P_{\text{sim}} = \#(P_{EM,j} < P_{EM})/N_{\text{sim}}$.

DNAm-to-trait mediation analysis

A diagram of the workflow with each of the following steps is shown in Supplementary Fig. 1. First, univariable MRs were conducted to estimate the total causal effect $\hat{\theta}_T$ of the DNAm sites on each trait. We assessed the impact of ~50,000 DNAm probes with ≥ 5 near-independent ($r^2 < 0.05$) mQTLs after harmonization of the datasets. DNAm probes significantly associated to the outcome ($P_T < 0.05/50000$=1e-6) were clumped based on the p-value of the total causal effect $\hat{\theta}_T$, P_T (distance-pruning at 1 Mb), to be independent of each other.

Second, MVMR analyses were performed to estimate the direct effect $\hat{\theta}_D$. Selected transcripts (see "Mediator selection threshold P_{EM}") were included as mediators as well as their associated SNPs as additional instruments. Steiger filtering on mediator-associated IVs was applied using the same t_{rev} threshold as for exposure-associated IVs. Remaining IVs were then clumped based on a rank score determined as follows: 1) for each mediator, IVs were ranked according to their association p-value to the mediator and assigned an integer score, 2) for each IV, a final score was calculated as the sum of its individual mediator scores. Following the establishment of the $\boldsymbol{B}$ effect size matrix, $\hat{\theta}_D$ was calculated, as well as $\hat{\theta}_{D,\text{top}}$ which was estimated from a MVMR model that includes the transcript with the lowest $P_{EM,k}$ as sole mediator. If no transcript causally associated with the DNAm probe, mediation is not detectable, and hence $\hat{\theta}_D$ was set to $\hat{\theta}_T$ for that probe (inclusion of such probes in MP calculation was termed "overall mediation proportion"). As the Steiger filter removed exposure-associated instruments with larger mediator than exposure effects (see "Univariable and multivariable Mendelian randomization"), the number of initial exposure-associated instruments ($m_E \geq 5$) could decrease. Therefore, to avoid scenarios of reverse causality where the mediator exerts an effect on the outcome through the exposure, we required ≥ 3 exposure-associated IVs.

We additionally conducted mediation analyses on independent mediators. To this end, selected mediators (those that passed P_{EM}) were clumped at various correlation thresholds R_{med} (default $R_{\text{med}} < 0.3$, with 0.2 and 0.1 being tested as well). Correlations among mediators were calculated based on QTL effect sizes of independent exposure and mediator IVs and priority was given to the mediator with the lowest $P_{EM,k}$.

Estimating and comparing mediation proportions

Mediation proportions (MPs) were estimated on sets of DNAm-trait pairs with significant total causal effects $\hat{\theta}_T$, either grouped by trait (if there were at least 10 such pairs within a given trait), trait category (e.g. hepatic traits, inflammatory traits/diseases) or combining all pairs together. MPs were then calculated by regressing $\hat{\theta}_D$ on $\hat{\theta}_T$ (without intercept) to estimate for the unmediated proportion, $\hat{\gamma}$, which after correcting for regression dilution bias[31] (Eq. (6)):

$$\hat{\gamma}_{\text{cor}} = \frac{\hat{\gamma}}{\sqrt{1 - \frac{\sum \text{SE}^2(\hat{\theta}_T)}{\sum \hat{\theta}_T^2}}} \tag{6}$$

yielded $\widehat{\text{MP}} = 1 - \hat{\gamma}_{\text{cor}}$ for a defined set of DNAm-trait pairs, together with a standard error. For individual DNAm-trait pairs, we report the $\widehat{\text{MP}}$ as $1 - \hat{\theta}_D/\hat{\theta}_T$, without providing its variance estimate since this would require individual-level data[26]. Note that $\widehat{\text{MP}}$ is an estimator of the true underlying MP and values outside the expected [0-1] range can be observed, especially if $\hat{\theta}_D$ and $\hat{\theta}_T$ estimates are of opposite sign. Such situations are expected to be rare in our analysis, as the total effect would be expected to be small and hence non-detectable.

In our approach, indirect effects θ_M are estimated by subtracting direct effects from total effects, which is also referred to as the difference in coefficients method[26]. Alternatively, the indirect effect can be estimated by the product of coefficients method[26], where univariable MR estimates from the exposure on the mediator are multiplied with the direct effects of the mediator on the outcome (Eq. (3)) and summed across mediators. Direct effects of the exposure on the outcome can then be obtained by the difference between the total and indirect effect. As demonstrated earlier[26], the two approaches yield highly concordant results (Supplementary Fig. 20).

To test the statistical significance between $\widehat{\text{MP}}$s estimated on two different sets of exposure-trait pairs (e.g. $\widehat{\text{MP}}$ of a given physiological category vs all categories combined) or on the same exposure-trait pairs, but with different parameter settings (e.g. changing P_{EM}), we made use of $\hat{\gamma}$ and its corresponding standard error se($\hat{\gamma}$) obtained from regressing $\hat{\theta}_D$ on $\hat{\theta}_T$ (both of which being corrected for regression dilution bias (Eq. (6))) to yield $\hat{\gamma}_{\text{cor}}$ and se($\hat{\gamma}$). We then performed a two-sided z-test based on the following test statistic:

$$\frac{\hat{\gamma}_{\text{cor}}^{(1)} - \hat{\gamma}_{cor}^{(2)}}{\sqrt{\text{se}\left(\hat{\gamma}_{\text{cor}}^{(1)}\right)^2 + \text{se}\left(\hat{\gamma}_{\text{cor}}^{(2)}\right)^2}} \sim \mathcal{N}(0,1) \qquad (7)$$

Significant difference between $\widehat{\text{MPs}}$ was defined by a two-sided p-value ≤ 0.05. Of note, this z-test assumes independence between $\hat{\gamma}^{(1)}$ and $\hat{\gamma}^{(2)}$ which is not always guaranteed (i.e., when comparing P_{EM} thresholds), hence the resulting p-values may be lenient.

Omics and trait summary statistics

We used mQTL data from the GoDMC consortium (n=32,851)[6], which contains > 170,000 whole blood DNAm sites with at least one significant *cis*-mQTL (P < 1e-6, < 1 Mb from the DNAm site, n > 5000). *Cis*-eQTL data were taken from the eQTLGen consortium (n = 31,684)[8] which includes *cis*-eQTLs (< 1 Mb from gene center, 2-cohort filter) for 19,250 transcripts (16,934 with at least one significant *cis*-eQTL at FDR < 0.05 corresponding to P < 1.8e-05).

GWAS summary statistics for outcome traits came from the largest (n_{average} > 320,000), predominantly European-descent, publicly available studies, as listed in Supplementary Data 1. Thirty-seven out of the 50 traits were continuous biomarkers or continuous physical measures with the GWAS conducted on the UK Biobank[69] (http://www.nealelab.is/uk-biobank). Remaining GWAS data came mostly from case/control studies made available by the consortium of the respective disease. For binary outcome traits, log-odds ratios were used as effect sizes and results should be interpreted on the liability scale.

Prior to each mediation analysis, exposure and mediator omics, GWAS and the reference panel data were harmonized. The analysis was conducted on autosomal chromosomes, and palindromic single nucleotide variants (SNPs), as well as SNPs with an allele frequency difference > 0.05 between any pairs of datasets were removed. If allele frequencies were not reported by the GWAS summary statistics, allele frequencies from the UK Biobank were used. Z-scores of summary statistics (molecular and outcome GWAS) were standardized by the square root of the sample size to be on the same SD scale.

DNAm-to-transcript MR analysis

As follow-up analyses, we calculated MR causal effects between all available DNAm sites and transcripts in *cis* (± 500 kb) following the same procedure as in the univariable MR to obtain total effects $\hat{\theta}_T$. First, near-independent (r^2 < 0.05) and significant (P < 1e-6) exposure IVs were selected and IVs not passing the aforementioned Steiger filter were discarded. MR causal effects were then computed based on Eq. (1) for pairs with ≥3 exposure IVs.

Pearson correlation coefficient with previously reported DNAm-transcript correlations[37] was calculated on common DNAm-transcript pairs to explore agreement. DNAm probe annotations with respect to the assessed transcript were from the IlluminaHumanMethylation450kanno.ilmn12.hg19 R package (v0.6.1)[70].

Simulation studies

We conducted simulation studies to assess the robustness of our model and to identify sources of bias in the estimated MP. Simulation settings were set up post-hoc to replicate mediation results obtained for real data (Supplementary Figs. 4-5; Supplementary Table 1).

We considered an exposure with heritability h_E^2 and m_E independent IVs. Effect sizes β_i^E for m_E IVs were drawn from a normal distribution $\beta_i^E \sim \mathcal{N}(0,\sqrt{h_E^2/m_E})$ and rescaled to total h_E^2. $N_{\text{med,pot}}$ potential mediators were simulated, among which N_{med} were contributing to the indirect effect θ_M. Each mediator k associated with m_M IVs with direct effects $\beta_{\text{direct},i}^{M_k} \sim \mathcal{N}(0,\sqrt{h_{M,\text{direct},k}^2/m_M})$ rescaled to $h_{M,\text{direct},k}^2$, the direct heritability of the mediator that does not take into account the additional heritability coming through the exposure. Causal effects of the exposure on the mediator ($\alpha_{\text{EM},k}$) and of the mediator on the outcome ($\alpha_{\text{MY},k}$) for N_{med} mediators were drawn from a bivariate normal distribution $\alpha_{\text{EM},k},\alpha_{\text{MY},k} \sim \mathcal{N}(\mathbf{0},\boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma}$ the covariance matrix:

$$\boldsymbol{\Sigma} = \begin{bmatrix} \text{var}(\alpha_{\text{EM}}) & \rho \cdot \sqrt{\text{var}(\alpha_{EM}) \cdot \text{var}(\alpha_{MY})} \\ \rho \cdot \sqrt{\text{var}(\alpha_{\text{EM}}) \cdot \text{var}(\alpha_{MY})} & \text{var}(\alpha_{MY}) \end{bmatrix}$$

where ρ is the correlation between $\alpha_{\text{EM},k}$ and $\alpha_{\text{MY},k}$. For the remaining $N_{\text{med,pot}}$ - N_{med} mediators, $\alpha_{\text{EM},k}$ and $\alpha_{\text{MY},k}$ causal effects were set to zero. The vector of effect sizes $\boldsymbol{\beta}^{M_k}$ of size $m_E + N_{\text{med}} \cdot m_M$ for each mediator k was constructed to have effect sizes equalling $\beta_i^E \cdot \alpha_{\text{EM},k}$ for m_E exposure SNPs and effect sizes equalling $\beta_{\text{direct},i}^{M_k}$ for m_M mediator-associated SNPs. The effect sizes of remaining IVs associated to mediators $i \neq k$ were set to zero. Likewise, effect sizes of the $N_{\text{med}} \cdot m_M$ IVs on the exposure in the $\boldsymbol{\beta}^E$ vector were set to zero.

The indirect effect θ_M, direct effect θ_D and total effect θ_T were calculated as:

$$\theta_M = \sum_k \alpha_{\text{EM},k} \cdot \alpha_{\text{MY},k} \,;\; \theta_D = \theta_M\left(\frac{1}{\text{MP}} - 1\right) ;\; \theta_T = \theta_D + \theta_M$$

These quantities allowed to generate the outcome effect size vector $\boldsymbol{\beta}^Y$:

$$\boldsymbol{\beta}^Y = \theta_D \cdot \boldsymbol{\beta}^E + \sum_k \alpha_{\text{MY},k} \cdot \boldsymbol{\beta}^{M_k}$$

For each scenario, we simulated 500 data sets to each time get $\boldsymbol{\beta}^E$, $\boldsymbol{\beta}^{M_k}$ and $\boldsymbol{\beta}^Y$. Normally distributed noise, as a function of the sample size N, $\epsilon_i^E \sim \mathcal{N}(0,1/N_E)$, $\epsilon_i^M \sim \mathcal{N}(0,1/N_M)$ and $\epsilon_i^Y \sim \mathcal{N}(0,1/N_Y)$ was added to each simulated vector. To approximate our real data, exposure effect sizes of SNPs serving as mediator instruments were set to zero again. We then estimated for each model $\hat{\theta}_T$ and $\hat{\theta}_D$ by including mediators that satisfied P_{EM} (p-value of the causal effect from the exposure on the mediator) denoted $N_{\text{med,sig}}$. Causal effects $\hat{\theta}_D$ were regressed on $\hat{\theta}_T$ to estimate the coefficient $\hat{\gamma}$ which after accounting for regression dilution (Eq. (6)) allowed to obtain the estimated $\widehat{\text{MP}}$.

Reporting summary

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

Data availability

Methylation QTLs used in this study are from the GoDMC mQTL meta-analysis and are available on the GoDMC Consortium website (http://mqtldb.godmc.org.uk/downloads). Expression QTLs are from the eQTLGen eQTL meta-analysis and are available on the eQTLGen Consortium website (https://www.eqtlgen.org/cis-eqtls.html). The list of GWAS summary statistics used in this study is in Supplementary Data 1, all of which are publicly available. UK10K individual-level data are available upon request (https://www.uk10k.org/data_access.html). Source data are provided with this paper.

Code availability

Software to conduct univariable MR-IVW (molecular trait → outcome, molecular trait 1 → molecular trait 2) and multivariable MR-IVW (molecular trait 1 → molecular trait 2 → outcome) is available at https://github.com/masadler/smrivw(https://doi.org/10.5281/zenodo.7324709[71]). Source code (C++, released under GPL v2 license) and executable file (for Linux platforms, released under MIT license) are provided which rely on functionalities and the data management architecture of the SMR software v1.03 (https://cnsgenomics.com/software/smr[24]). The provided documentation hosted on the GitHub repository guides users in reproducing the mediation results and conducting univariable and multivariable MR on their own combinations of QTL and GWAS datasets.

References

1. Buniello, A. et al. The NHGRI-EBI GWAS catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucl. Acids Res.* **47**, D1005–D1012 (2019).
2. Tam, V. et al. Benefits and limitations of genome-wide association studies. *Nat. Rev. Genet.* **20**, 467–484 (2019).
3. Maurano, M. T. et al. Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190–1195 (2012).
4. Nicolae, D. L. et al. Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet.* **6**, e1000888 (2010).
5. Hasin, Y., Seldin, M. & Lusis, A. Multi-omics approaches to disease. *Genome Biol.* **18**, 1–15 (2017).
6. Min, J. L. et al. Genomic and phenotypic insights from an atlas of genetic effects on DNA methylation. *Nat. Genet.* **53**, 1311–1321 (2021).
7. Consortium, G. et al. Genetic effects on gene expression across human tissues. *Nature* **550**, 204 (2017).
8. Võsa, U. et al. Large-scale cis-and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. *Nat. Genet.* **53** 1300–1310 (2021).
9. Sun, B. B. et al. Genomic atlas of the human plasma proteome. *Nature* **558**, 73–79 (2018).
10. Folkersen, L. et al. Genomic and drug target evaluation of 90 cardiovascular proteins in 30,931 individuals. *Nat. Metab.* **2**, 1135–1148 (2020).
11. Ferkingstad, E. et al. Large-scale integration of the plasma proteome with genetics and disease. *Nat. Genet.* **53**,1712–1721 (2021).
12. Shin, S.-Y. et al. An atlas of genetic influences on human blood metabolites. *Nat. Genet.* **46**, 543–550 (2014).
13. Lotta, L. A. et al. A cross-platform approach identifies genetic regulators of human metabolism and health. *Nat. Genet.* **53**, 54–64 (2021).
14. Giambartolomei, C. et al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.* **10**, e1004383 (2014).
15. Hormozdiari, F. et al. Colocalization of GWAS and eQTL signals detects target genes. *Am. J. Hum. Genet.* **99**, 1245–1260 (2016).
16. Gusev, A. et al. Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.* **48**, 245–252 (2016).
17. Barbeira, A. N. et al. Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. *Nat. Commun.* **9**, 1–20 (2018).
18. Zhu, Z. et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet.* **48**, 481–487 (2016).
19. Porcu, E. et al. Mendelian randomization integrating GWAS and eQTL data reveals genetic determinants of complex and clinical traits. *Nat. Commun.* **10**, 1–12 (2019).
20. Porcu, E. et al. Differentially expressed genes reflect disease-induced rather than disease-causing changes in the transcriptome. *Nat. Commun.* **12**, 5647 (2021).
21. 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).
22. Giambartolomei, C. et al. A Bayesian framework for multiple trait colocalization from summary association statistics. *Bioinformatics* **34**, 2538–2545 (2018).
23. Gleason, K. J., Yang, F., Pierce, B. L., He, X. & Chen, L. S. Primo: integration of multiple GWAS and omics QTL summary statistics for elucidation of molecular mechanisms of trait-associated snps and detection of pleiotropy in complex traits. *Genome Biol.* **21**, 1–24 (2020).
24. Wu, Y. et al. Integrative analysis of omics summary data reveals putative mechanisms underlying complex traits. *Nat. Commun.* **9**, 1–14 (2018).
25. Hannon, E. et al. Leveraging DNA-methylation quantitative-trait loci to characterize the relationship between methylomic variation, gene expression, and complex traits. *Am. J. Hum. Genet.* **103**, 654–665 (2018).
26. Carter, A. R. et al. Mendelian randomisation for mediation analysis: current methods and challenges for implementation. *Eur. J. Epidemiol.* **36**, 465–478 (2021).
27. Sanderson, E. Multivariable Mendelian randomization and mediation. *Cold Spring Harb. Perspect. Med.* **11**, a038984 (2021).
28. Burgess, S. et al. Dissecting causal pathways using Mendelian randomization with summarized genetic data: application to age at menarche and risk of breast cancer. *Genetics* **207**, 481–487 (2017).
29. Burgess, S. & Thompson, S. G. Bias in causal estimates from Mendelian randomization studies with weak instruments. *Stat. Med.* **30**, 1312–1323 (2011).
30. Zuber, V., Colijn, J. M., Klaver, C. & Burgess, S. Selecting likely causal risk factors from high-throughput experiments using multivariable Mendelian randomization. *Nat. Commun.* **11**, 1–11 (2020).
31. Knuiman, M. W., Divitini, M. L., Buzas, J. S. & Fitzgerald, P. E. Adjustment for regression dilution in epidemiological regression analyses. *Ann. Epidemiol.* **8**, 56–63 (1998).
32. Howe, K. L. et al. Ensembl 2021. *Nucl. Acids Res.* **49**, D884–D891 (2021).
33. Sanderson, E., Spiller, W. & Bowden, J. Testing and correcting for weak and pleiotropic instruments in two-sample multivariable Mendelian randomization. *Stat. Med.* **40**, 5434–5452 (2021).
34. Hemani, G. et al. The MR-Base platform supports systematic causal inference across the human phenome. *elife* **7**, e34408 (2018).
35. Bird, A. DNA methylation patterns and epigenetic memory. *Genes Dev.* **16**, 6–21 (2002).
36. Wan, J. et al. Characterization of tissue-specific differential DNA methylation suggests distinct modes of positive and negative gene expression regulation. *BMC Genomics* **16**, 1–11 (2015).
37. Grundberg, E. et al. Global analysis of DNA methylation variation in adipose tissue from twins reveals links to disease-associated variants in distal regulatory elements. *Am. J. Hum. Genet.* **93**, 876–890 (2013).
38. Rauluseviciute, I., Drabløs, F. & Rye, M. B. DNA hypermethylation associated with upregulated gene expression in prostate cancer demonstrates the diversity of epigenetic regulation. *BMC Med. Genomics.* **13**, 1–15 (2020).
39. Ruiz-Arenas, C. et al. Identification of autosomal cis expression quantitative trait methylation (cis eQTMs) in children's blood. *eLife* **11**, e65310 (2022).
40. Lippai, R. et al. Immunomodulatory role of Parkinson's disease 7 in inflammatory bowel disease. *Sci. Rep.* **11**, 14582 (2021).

41. Di Narzo, A. F. et al. High-throughput identification of the plasma proteomic signature of inflammatory bowel disease. *J. Crohn's Colitis*. **13**, 462–471 (2019).
42. Singh, Y. et al. DJ-1 (Park7) affects the gut microbiome, metabolites and the development of innate lymphoid cells (ILCs). *Sci. Rep.* **10**, 1–19 (2020).
43. Zhang, J. et al. Deficiency in the anti-apoptotic protein DJ-1 promotes intestinal epithelial cell apoptosis and aggravates inflammatory bowel disease via p53. *J. Biol. Chem.* **295**, 4237–4251 (2020).
44. Moschen, A. R. et al. Lipocalin 2 protects from inflammation and tumorigenesis associated with gut microbiota alterations. *Cell Host Mcrobe.* **19**, 455–469 (2016).
45. Gieger, C. et al. New gene functions in megakaryopoiesis and platelet formation. *Nature* **480**, 201–208 (2011).
46. Nürnberg, S. T. et al. A GWAS sequence variant for platelet volume marks an alternative DNM3 promoter in megakaryocytes near a MEIS1 binding site. *Blood, J. Am. Soc. Hematol.* **120**, 4859–4868 (2012).
47. Watkins, N. A. et al. A HaemAtlas: characterizing gene expression in differentiated human blood cells. *Blood, J. Am. Soc. Hematol.* **113**, e1–e9 (2009).
48. Bielczyk-Maczyńska, E. et al. A loss of function screen of identified genome-wide association study loci reveals new genes controlling hematopoiesis. *PLoS Genet.* **10**, e1004450 (2014).
49. Liang, Y. et al. Demethylation of the FCER1G promoter leads to FcεRI overexpression on monocytes of patients with atopic dermatitis. *Allergy* **67**, 424–430 (2012).
50. Pairo-Castineira, E. et al. Genetic mechanisms of critical illness in Covid-19. *Nature* **591**, 92–98 (2021).
51. Initiative, C.-. H. G. et al. Mapping the human genetic architecture of COVID-19. Nature 600:472–477 (2021).
52. Smieszek, S. P. & Polymeropoulos, M. H. Loss of Function Mutations in the IFNAR2 in COVID-19 Severe Infection Susceptibility. J. Glob. Antimicrob. Resist. **26**, 239–240 (2021).
53. Zhu, H., Wang, G. & Qian, J. Transcription factors as readers and effectors of DNA methylation. *Nat. Rev. Genet.* **17**, 551–565 (2016).
54. Yin, Y. et al. Impact of cytosine methylation on DNA binding specificities of human transcription factors. *Science* **356**, eaaj2239 (2017).
55. Whalen, S., Truty, R. M. & Pollard, K. S. Enhancer–promoter interactions are encoded by complex genomic signatures on looping chromatin. *Nat. Genet.* **48**, 488–496 (2016).
56. Jjingo, D., Conley, A. B., Soojin, V. Y., Lunyak, V. V. & Jordan, I. K. On the presence and role of human gene-body DNA methylation. *Oncotarget* **3**, 462 (2012).
57. Richmond, R. C., Hemani, G., Tilling, K., Davey Smith, G. & Relton, C. Challenges and novel approaches for investigating molecular mediation. *Hum. Mol. Genet.* **25**, R149–R156 (2016).
58. Verbanck, M., Chen, C.-y, Neale, B. & Do, R. Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. *Nat. Genet.* **50**, 693–698 (2018).
59. Pellegrino-Coppola, D. et al. Correction for both common and rare cell types in blood is important to identify genes that correlate with age. *BMC Genomics.* **22**, 1–12 (2021).
60. Merkin, J., Russell, C., Chen, P. & Burge, C. B. Evolutionary dynamics of gene and isoform regulation in mammalian tissues. *Science* **338**, 1593–1599 (2012).
61. Garrido-Martín, D., Borsari, B., Calvo, M., Reverter, F. & Guigó, R. Identification and analysis of splicing quantitative trait loci across multiple tissues in the human genome. *Nat. Commun.* **12**, 1–16 (2021).
62. Zhu, Z. et al. Causal associations between risk factors and common diseases inferred from gwas summary data. *Nat. Commun.* **9**, 1–12 (2018).
63. UK10K et al.The UK10K project identifies rare variants in health and disease. Nature 526, 82 (2015).
64. Hemani, G., Tilling, K. & Davey Smith, G. Orienting the causal relationship between imprecisely measured traits using GWAS summary data. *PLoS Genet.* **13**, e1007081 (2017).
65. Lynch, M., Walsh, B. et al. Genetics and analysis of quantitative traits, vol. 1 (Sinauer Sunderland, MA, 1998).
66. Gutierrez-Arcelus, M. et al. Passive and active DNA methylation and the interplay with genetic variation in gene regulation. *eLife* **2**, e00523 (2013).
67. Sönmez Flitman, R. et al. Untargeted metabolome-and transcriptome-wide association study suggests causal genes modulating metabolite concentrations in urine. *J. Proteome Res.* **20**, 5103–5114 (2021).
68. Burgess, S., Bowden, J., Fall, T., Ingelsson, E. & Thompson, S. G. Sensitivity analyses for robust causal inference from Mendelian randomization analyses with multiple genetic variants. *Epidemiol.* **28**, 30 (2017).
69. Bycroft, C. et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
70. Hansen, K. IlluminaHumanMethylation450kanno. ilmn12. hg19: annotation for Illumina's 450k methylation arrays. *R. Package version 0. 6. 0* **10**, B9 (2016).
71. Sadler, M. Quantifying the role of transcript levels in mediating DNA methylation effects on complex traits and diseases (2022). Masadler/smrivw, https://doi.org/10.5281/zenodo.7324709.

Acknowledgements

This work was supported by the Swiss National Science Foundation (310030_189147) to Z.K. L.D. was calculated based on the UK10K data resource (EGAD00001000740, EGAD00001000741). Computations we performed on the JURA cluster of the University of Lausanne. We have used RNA-seq data from 557 CoLaus participants to compute gene-gene correlations, which were kindly made available by Sven Bergmann.

Author contributions

M.C.S., E.P. and Z.K. conceived and designed the study. M.C.S. performed statistical analyses. K.L. contributed to the statistical analyses. EP provided guidance on statistical analyses. Z.K. supervised all statistical analyses. All the authors contributed by providing advice on interpretation of results. C.A. contributed with the biological interpretation of the results. M.C.S., E.P. and Z.K. drafted the manuscript. C.A. contributed to the writing of specific sections. All authors read, approved, and provided feedback on the final 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/s41467-022-35196-3.

Correspondence and requests for materials should be addressed to Marie C. Sadler or Zoltán Kutalik.

Peer review information *Nature Communications* thanks Jean Morrison and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

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

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

Supplementary Materials for

Quantifying the role of transcript levels in mediating DNA methylation effects on complex traits and diseases

Marie C. Sadler[1,2,3,*], Chiara Auwerx[1,2,3,4], Kaido Lepik[1,2,3], Eleonora Porcu[1,2,3,4,5], Zoltán Kutalik[1,2,3,5,*]

[1] University Center for Primary Care and Public Health, Lausanne, Switzerland

[2] Department of Computational Biology, University of Lausanne, Lausanne, Switzerland

[3] Swiss Institute of Bioinformatics, Lausanne, Switzerland

[4] Center for Integrative Genomics, University of Lausanne, Lausanne, Switzerland

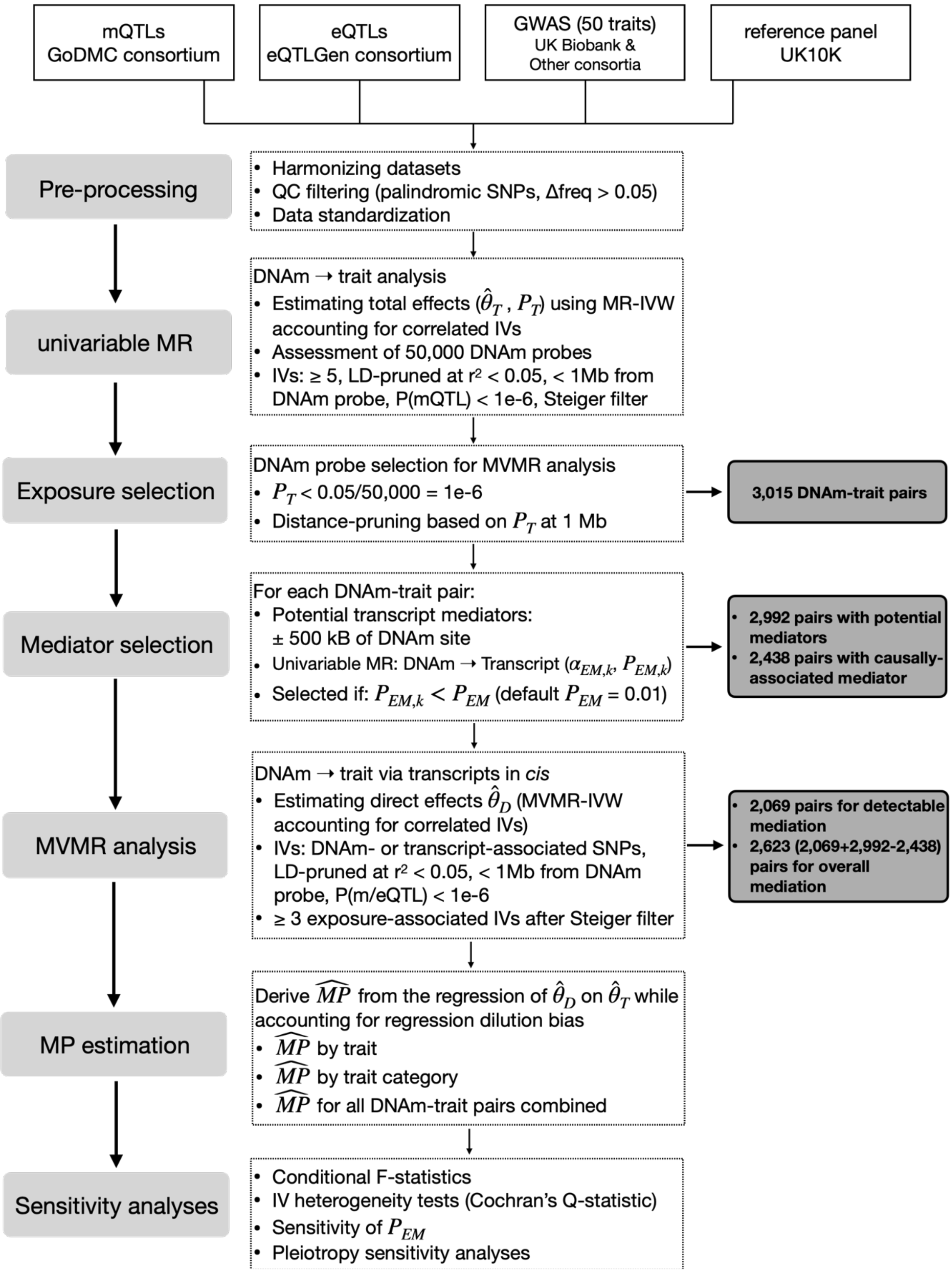
[5]Authors jointly supervised this work

*Corresponding authors: marie.sadler@unil.ch, zoltan.kutalik@unil.ch

Table of Contents

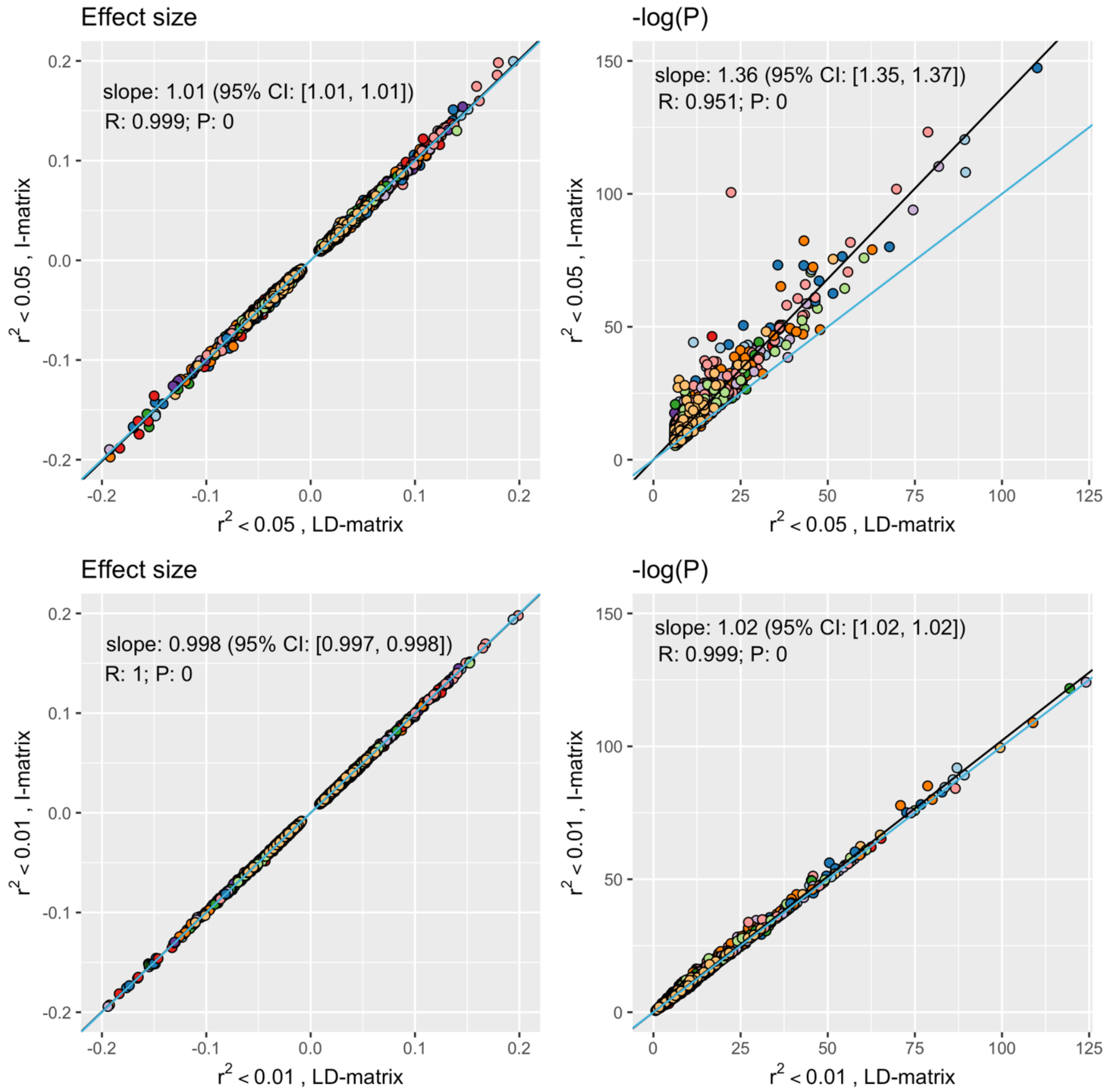
Supplementary Figures

3S-MVMR Workflow

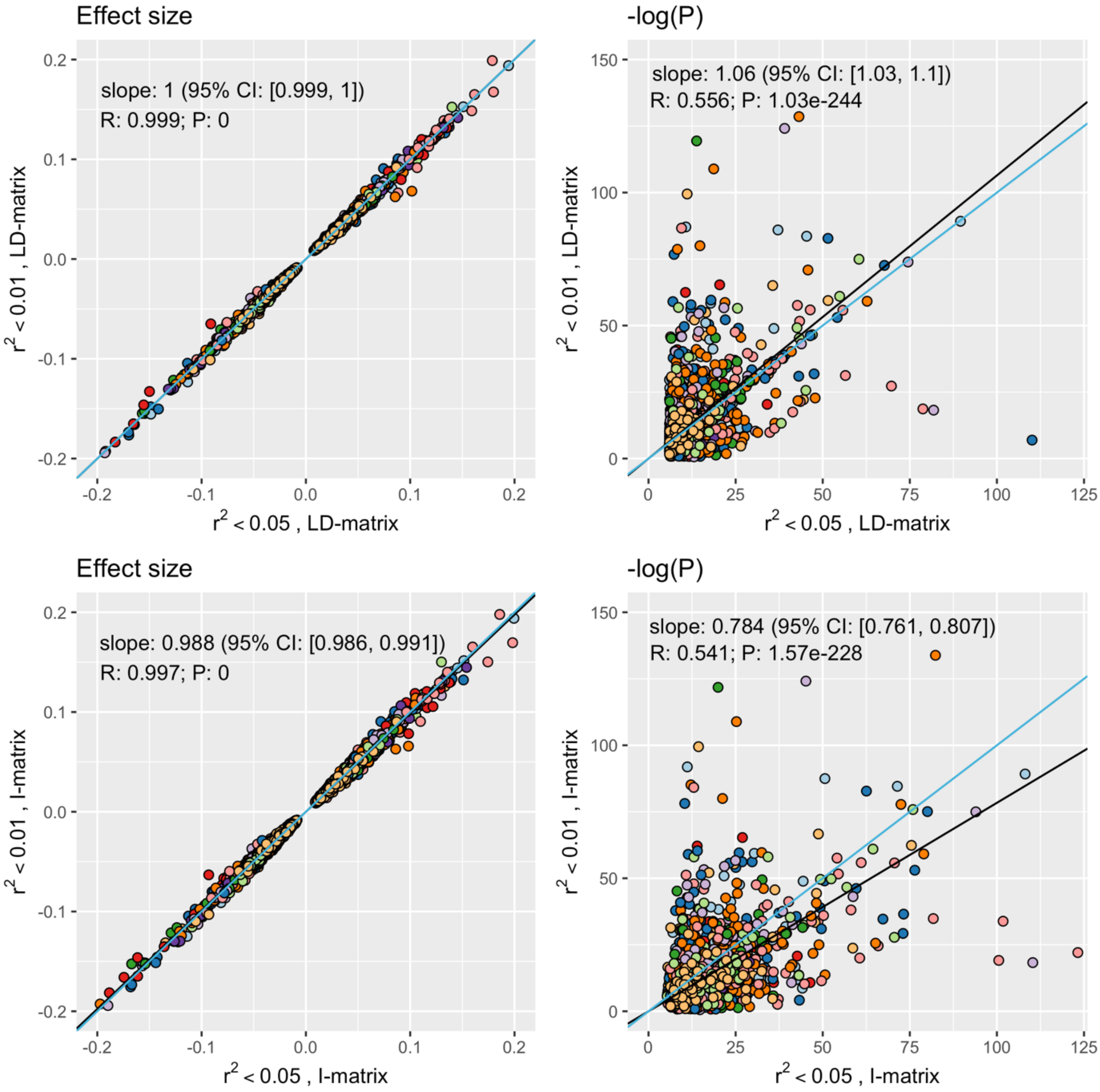

mQTLs
GoDMC consortium
eQTLs
eQTLGen consortium
GWAS (50 traits)
UK Biobank &
Other consortia
reference panel
UK10K
Pre-processing
• Harmonizing datasets
• QC filtering (palindromic SNPs, Δfreq > 0.05)
• Data standardization
univariable MR
DNAm → trait analysis
• Estimating total effects ($\hat{\theta}_T$, P_T) using MR-IVW accounting for correlated IVs
• Assessment of 50,000 DNAm probes
• IVs: ≥ 5, LD-pruned at $r^2 < 0.05$, < 1Mb from DNAm probe, P(mQTL) < 1e-6, Steiger filter
Exposure selection
DNAm probe selection for MVMR analysis
• P_T < 0.05/50,000 = 1e-6
• Distance-pruning based on P_T at 1 Mb
3,015 DNAm-trait pairs
Mediator selection
For each DNAm-trait pair:
• Potential transcript mediators: ± 500 kB of DNAm site
• Univariable MR: DNAm → Transcript ($\alpha_{EM,k}$, $P_{EM,k}$)
• Selected if: $P_{EM,k} < P_{EM}$ (default P_{EM} = 0.01)
• 2,992 pairs with potential mediators
• 2,438 pairs with causally-associated mediator
MVMR analysis
DNAm → trait via transcripts in *cis*
• Estimating direct effects $\hat{\theta}_D$ (MVMR-IVW accounting for correlated IVs)
• IVs: DNAm- or transcript-associated SNPs, LD-pruned at $r^2 < 0.05$, < 1Mb from DNAm probe, P(m/eQTL) < 1e-6
• ≥ 3 exposure-associated IVs after Steiger filter
• 2,069 pairs for detectable mediation
• 2,623 (2,069+2,992-2,438) pairs for overall mediation
MP estimation
Derive $\widehat{MP}$ from the regression of $\hat{\theta}_D$ on $\hat{\theta}_T$ while accounting for regression dilution bias
• $\widehat{MP}$ by trait
• $\widehat{MP}$ by trait category
• $\widehat{MP}$ for all DNAm-trait pairs combined
Sensitivity analyses
• Conditional F-statistics
• IV heterogeneity tests (Cochran's Q-statistic)
• Sensitivity of P_{EM}
• Pleiotropy sensitivity analyses

Supplementary Figure 1. **3S-MVMR workflow**. DNAm-to-trait mediation analysis workflow.

MR parameter sensitivity analyses

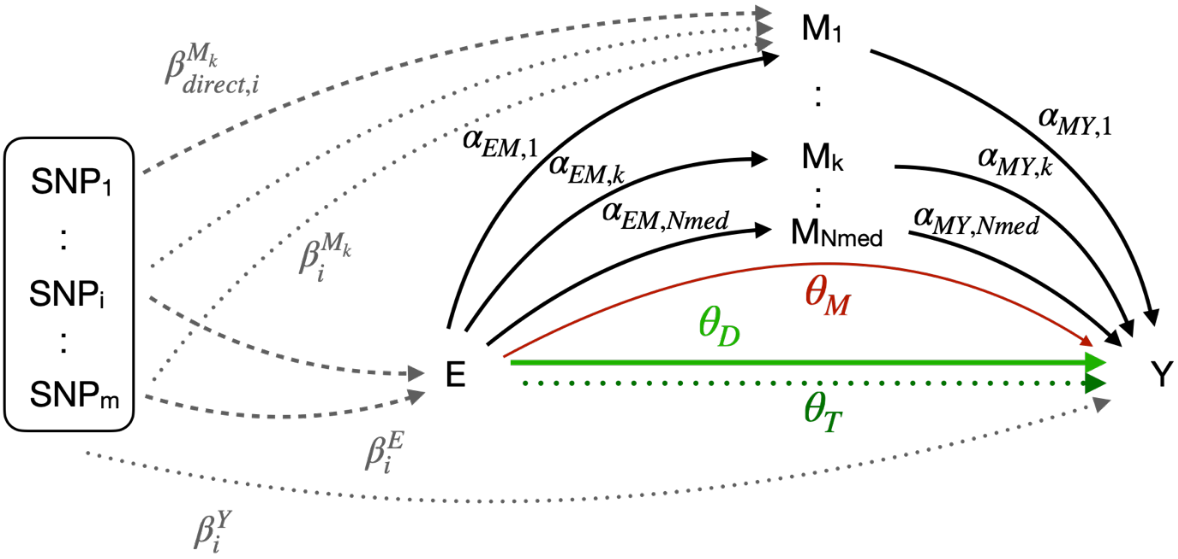


Supplementary Figure 2. **Sensitivity analyses to assess the influence of the LD-matrix**. The analyses were performed on the 3,015 DNAm-trait pairs for which there was a significant causal effect (Supplementary Fig. 1). Effect sizes and corresponding p-values were derived from IVW-MR estimates, once accounting for correlation between instruments (i.e. inclusion of LD-matrix) and once setting the LD-matrix to the identity matrix I (i.e. standard IVW estimates). The first row shows the results when setting the pruning threshold r^2 to 0.05 and the second row to 0.01. The analyses show that MR effect estimates are not affected by the LD-matrix, but if the LD-matrix was omitted p-values of causal effects were deviating towards lower values at $r^2 = 0.05$, while this was no longer the case at $r^2 = 0.01$. This indicates that at $r^2 = 0.01$, independence between IVs can be confidently claimed, while at $r^2 > 0.01$, the LD-matrix should be included to avoid false positives. Each dot represents a DNAm-trait pair colour-coded by the physiological category of the trait as defined in Supplementary Fig. 8. The slope is indicated in black with numerical values shown in the plotting area and the identity line in blue. The reported p-values (*P*) for the corresponding Pearson correlations (*R*) arise from a two-sided t-test and are shown in the plotting area.

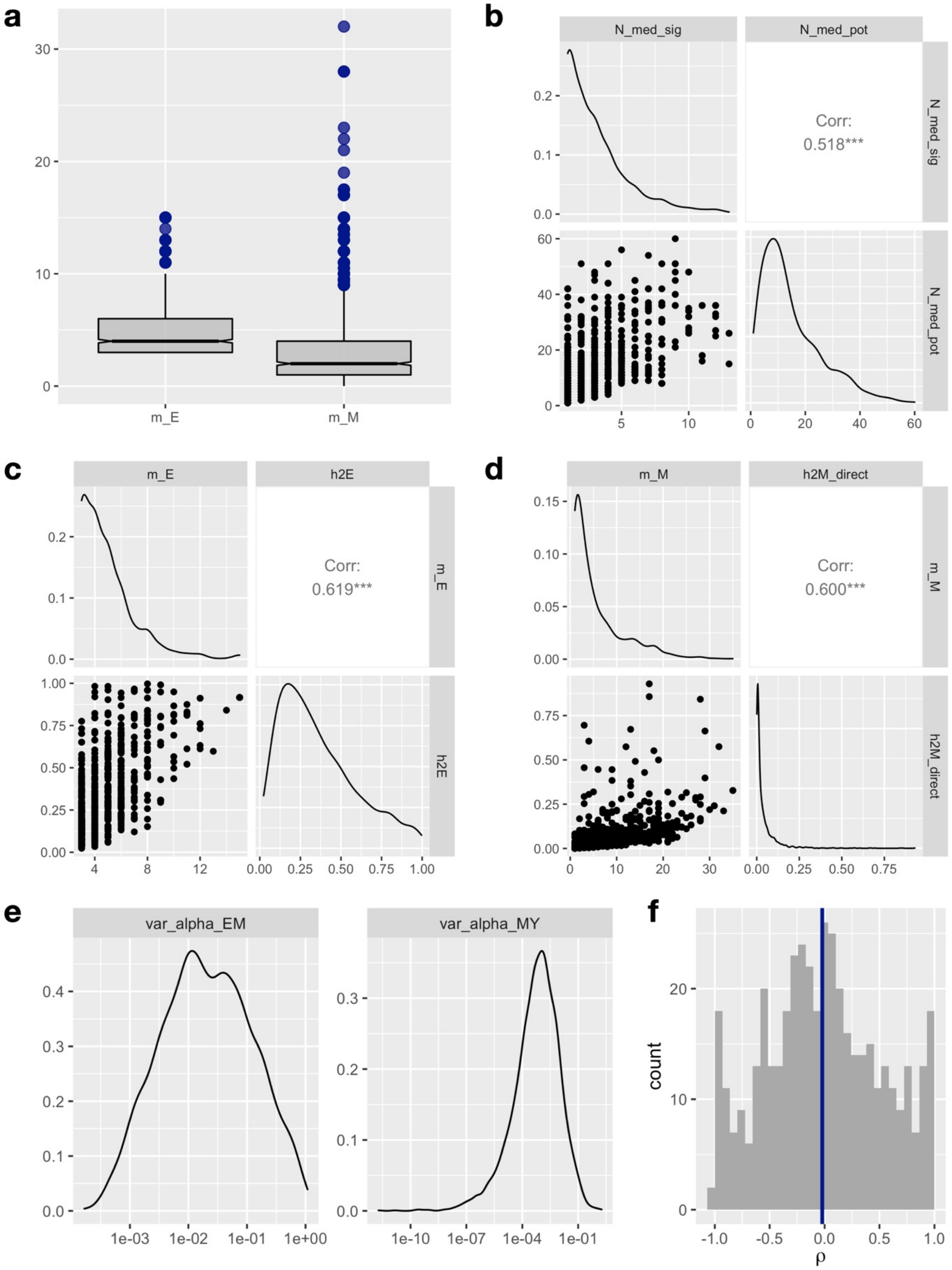


Supplementary Figure 3. Sensitivity analyses to assess the influence of the pruning threshold r^2. Again, the analyses were performed on the 3,015 DNAm-trait pairs for which there was a significant causal effect (Supplementary Fig. 1). Effect sizes and corresponding p-values were derived from IVW-MR estimates at different pruning thresholds. The first row shows the results when including the LD-matrix and the second row when setting it to the identity matrix I. The analyses show that MR effect estimates are not affected by the different pruning thresholds. However, the significance levels (p-values) can differ between both thresholds, although no overall significant difference was found as shown by the slope (provided the LD-matrix is included). When omitting the LD-matrix, lower p-values were found at r^2 = 0.05, a threshold at which, however, neglecting correlation between IVs was demonstrated to be an invalid approach (Supplementary Fig. 2). Each dot represents a DNAm-trait pair colour-coded by the physiological category of the trait as defined in Supplementary Fig. 8. The slope is indicated in black with numerical values shown in the plotting area and the identity line in blue. The reported p-values (*P*) for the corresponding Pearson correlations (*R*) arise from a two-sided t-test and are shown in the plotting area.

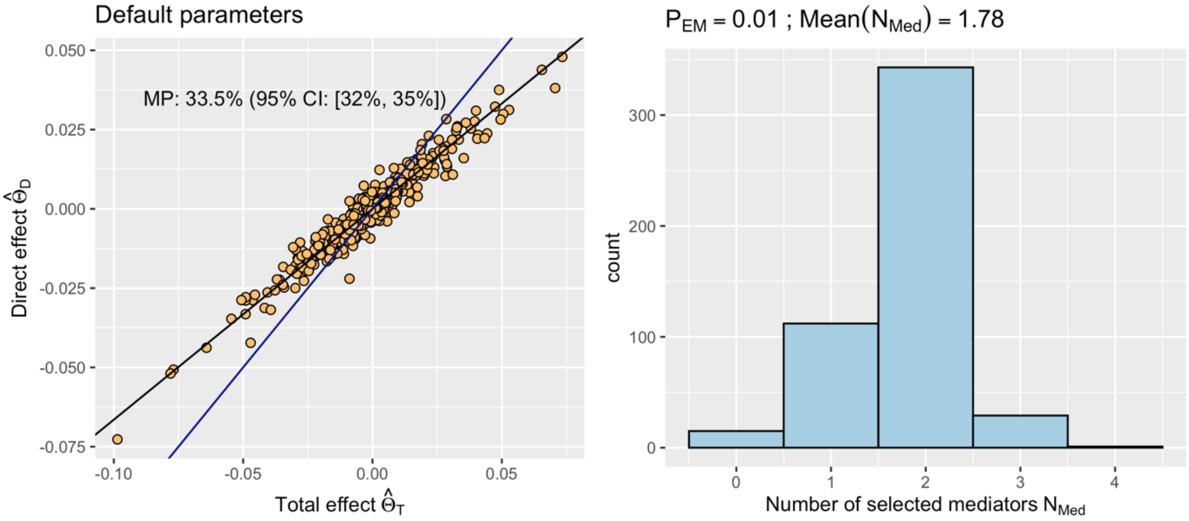
Simulation studies



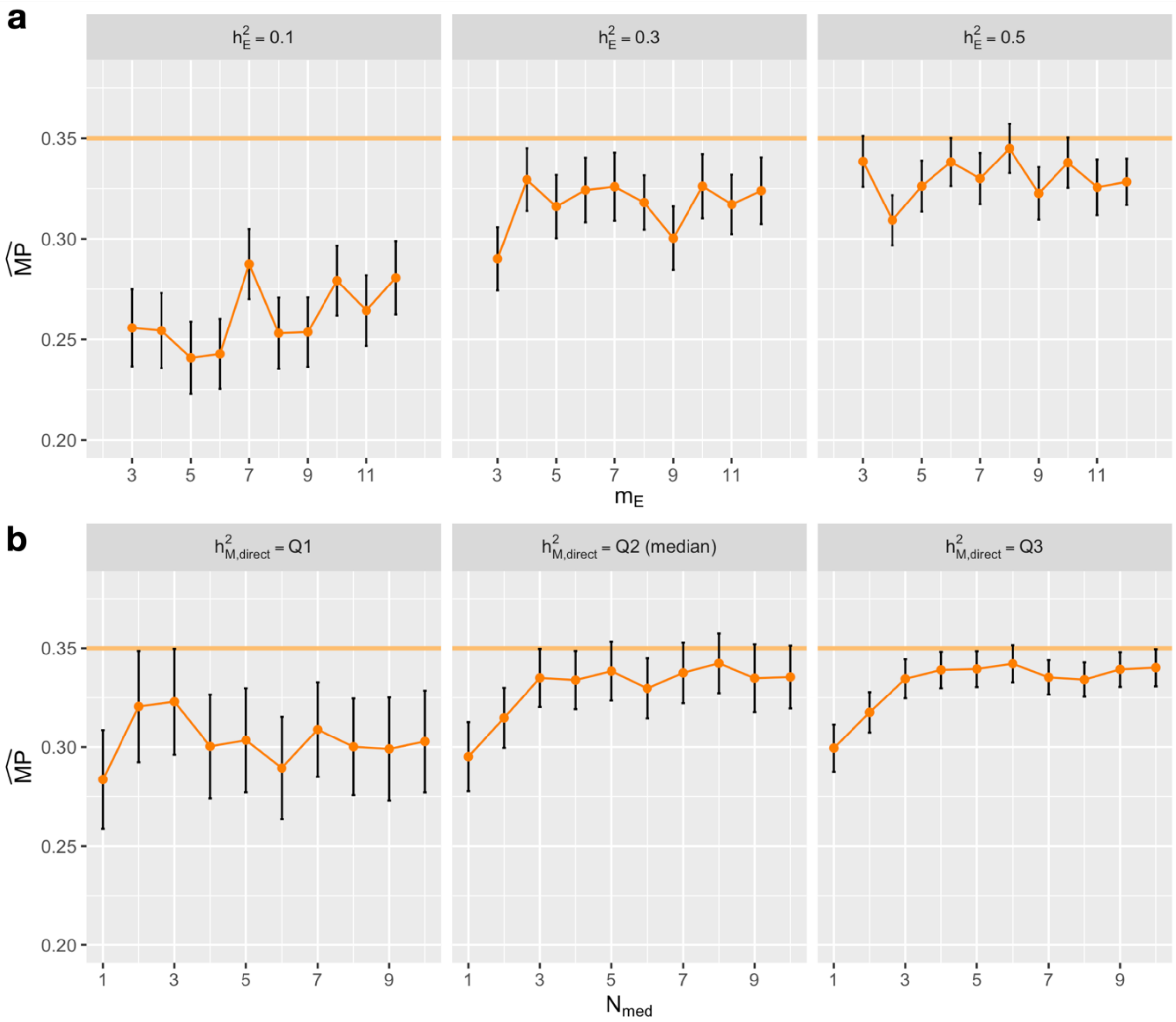
Supplementary Figure 4. **Model used in the simulation settings to estimate the total, direct and indirect causal effects ($\boldsymbol{\Theta_T}$, $\boldsymbol{\Theta_D}$ and $\boldsymbol{\Theta_M}$, respectively).** Genetic variants (SNPs) are either directly associated (dashed arrow) with the exposure *E* or mediators *M$_k$* (1 to *N*$_{\text{med}}$), or indirectly (dotted arrow) with *M$_k$* through *E*. The genetic effect sizes are denoted by β, where β^E are direct effects to *E*, β^{M_k} total effects to *M$_k$* made of the direct effects $\beta^{M_k}_{direct}$ to *M$_k$* and the indirect effects through *E*, and β^Y are total effects to the outcome *Y* through either *E* or *M*. Causal effects from *E* to *M$_k$* are denoted by $\alpha_{EM,k}$ and causal effects from *M* to *Y* by $\alpha_{MY,k}$.



Supplementary Figure 5. Distribution of the simulation parameters as observed in real data. Numerical values (interquartile ranges, mean) are shown in Supplementary Table 1 and parameter choices of the different simulation settings to explore the full range of realistic parameter estimates are summarized in Supplementary Table 2. **a** Distribution of the number of exposure-associated (m_E) and mediator-associated (m_M) independent instrumental variables (IVs) calculated on 1,836 DNAm-trait pairs. Boxes bound the 25th, 50th (median, centre), and the 75th quantile. Whiskers range from minima (Q1 – 1.5*IQR) to maxima (Q3 + 1.5*IQR) with points above or below representing potential outliers. **b** Distribution of the number of selected mediators ($N_{\text{med,sig}}$) and of the total number of potential mediators in the region ($N_{\text{med,pot}}$ with $N_{\text{med,pot}} \geq \text{N}_{\text{med,sig}}$). The Pearson correlation coefficient (Corr) is shown with the stars (***) indicating that the corresponding p-value (two-sided test-statistic) was below 2.2e-16 (exact p-value equalled 4.14e-64). **c** Distribution of the exposure heritability (h_E^2) in relationship with the number of exposure-associated IVs (m_E). Same Pearson correlation calculation as in b (exact p-value equalled 3.12e-93). **d** Distribution of the direct heritability of each mediator k ($h^2_{M,direct,k}$) - ignoring the heritability coming through the exposure - in relationship with the number of mediator-associated IVs (m_M). Same Pearson correlation calculation as in b (exact p-value equalled 5.08e-176). **e** Distribution of the variance (across all mediators) of the exposure-to-mediator causal effects (var($\alpha_{EM,k}$)) and mediator-to-outcome effects (var($\alpha_{MY,k}$)) as estimated by $\widehat{\alpha}_k^2 - \text{se}(\widehat{\alpha}_k)^2$. **f** Distribution of the correlation (ρ) between $\alpha_{EM,k}$ and $\alpha_{MY,k}$. Estimation was done by considering DNAm-trait pairs with at least 3 mediators and calculating for each pair the correlation between $\alpha_{EM,k}$ and $\alpha_{MY,k}$ that were estimated for each mediator.

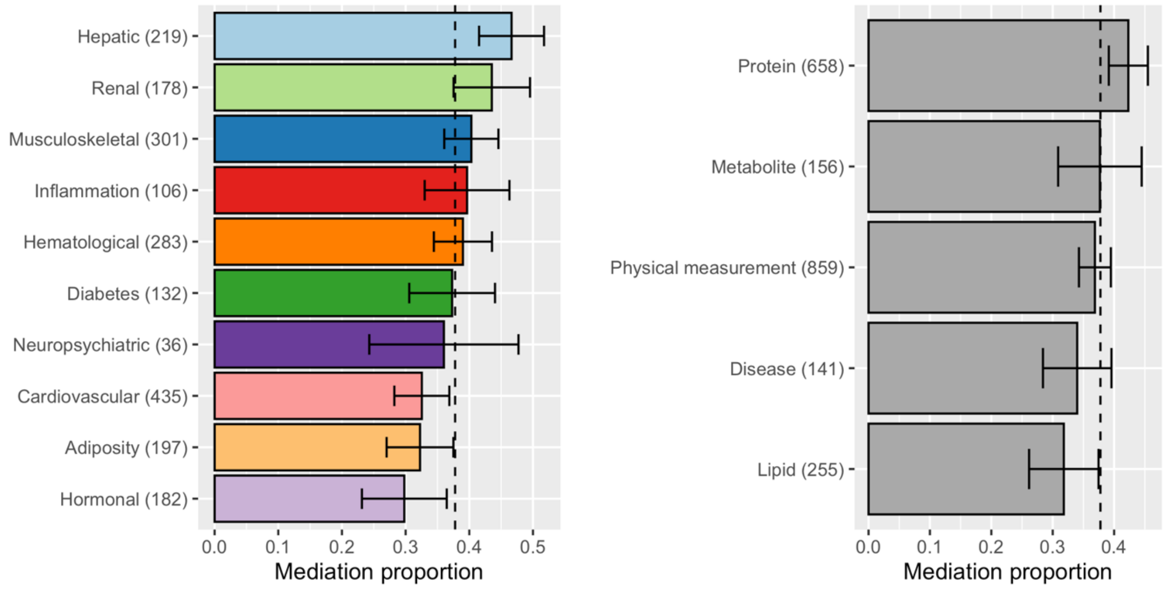


Supplementary Figure 6. Simulation results with the parameter default settings as indicated in Supplementary Table 2. 500 exposure-outcome pairs were simulated and for each a direct and total effect was estimated. The estimated mediation proportion (%) together with the 95% CI are displayed in the plot area (resulting from the regression of $\widehat{\Theta}_D$ against $\widehat{\Theta}_T$) with the corresponding slope plotted in black (blue line represents the identity line). Mediators were selected based on a p-value threshold P_{EM} and the distribution of the selected number of mediators (among a set of 12 potential mediators) is shown in the histogram. The true number of relevant mediators was 2 and the true MP was 35% (Supplementary Table 2).



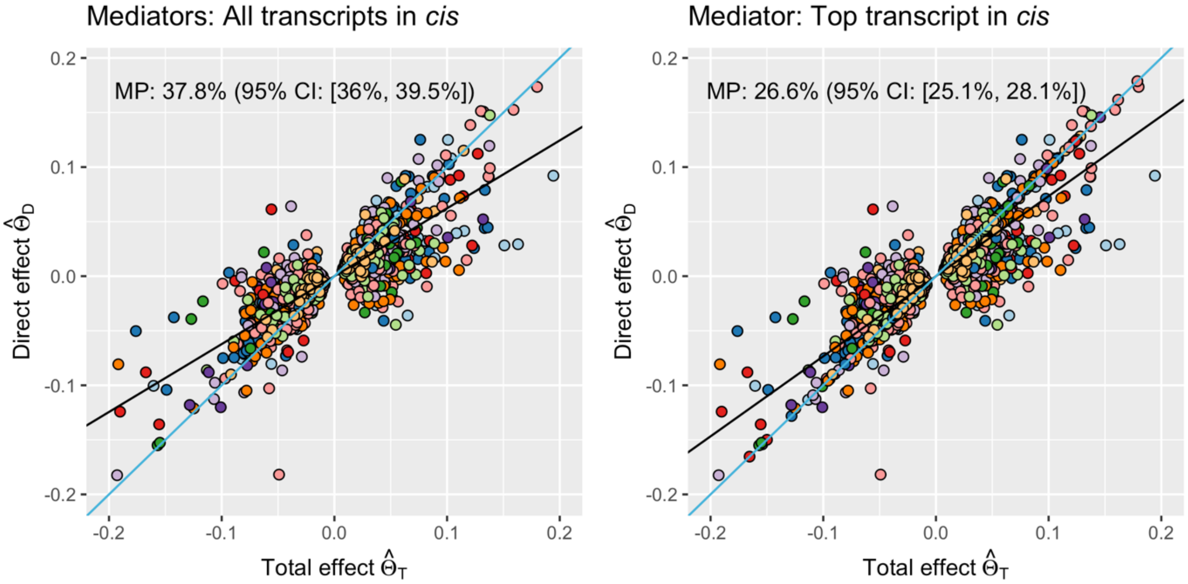
Supplementary Figure 7. Simulation results varying m_E and N_{med} (Supplementary Table 2). a The number of exposure-associated instrumental variables (IVs) m_E was changed for different exposure heritabilities h^2_E. The estimated mediation proportions were more dependent on h^2_E than on the polygenicity of the exposure. **b** Dependence of the estimated mediation proportion on the number of true mediators N_{med} (i.e., mediators contributing to the indirect effect) stratified by $h^2_{M,direct}$. Underestimations were observed for fewer mediators (1-2) and when the direct mediator heritability was low (first quartile). With fewer true mediators N_{med}, missing a relevant mediator has a greater impact on the estimated mediation proportion than if multiple N_{med} are contributing towards the mediated effect. Error bars represent 95% CI calculated on 500 simulated exposure-outcome pairs.

Mediation proportions by physiological and structural categories



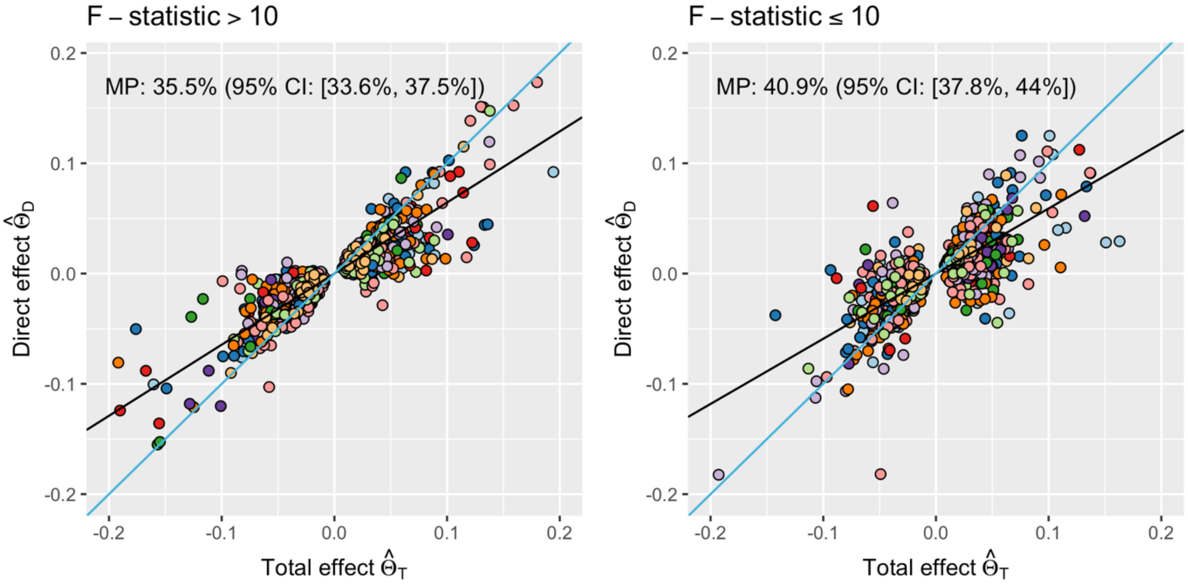
Supplementary Figure 8. Mediation proportion of traits grouped by physiological (left) and structural (right) categories. Further information about trait classification are shown in Supplementary Data 1. The vertical dotted lines denote the mean mediation proportion across all DNAm-trait pairs. 95% confidence intervals are represented by the error bars. $\widehat{\mathrm{MP}}s$ per category were derived by regressing $\hat{\theta}_D$ against $\hat{\theta}_T$. The number of DNAm-trait pairs falling into each category and on which the regression was performed is indicated in parentheses.

Mediation through the top transcript mediator

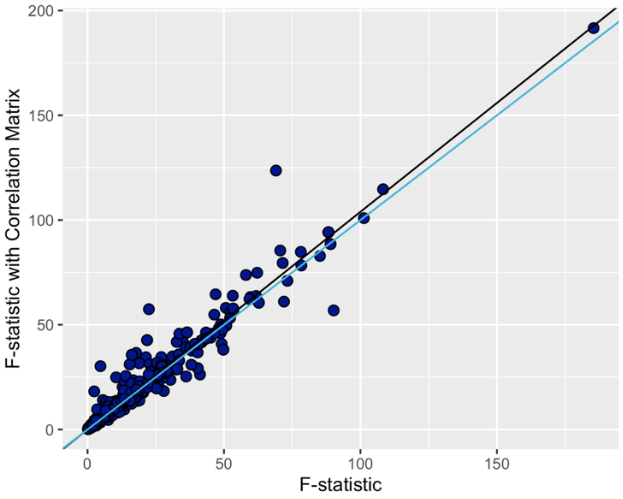


Supplementary Figure 9. Mediation through the top mediator. When restricting the mediation through the top transcript, i.e., the transcript most significantly associated to the DNAm site, the mean mediation proportion drops from 37.8% to 26.6% (evaluated are the 2,069 pairs with at least 1 causally associated transcript in *cis*). Plotted is the direct effect against the total effect together with the slope (black line). Each dot represents a DNAm-trait pair colour-coded by the physiological category of the trait as defined in Supplementary Fig. 8. The identity line is plotted in blue.

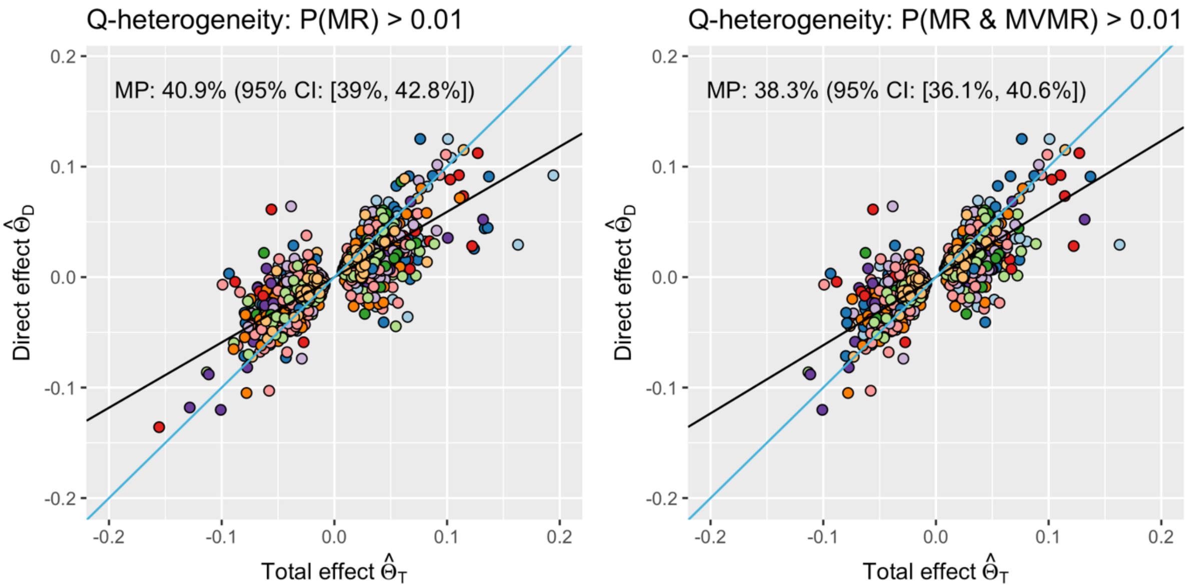
MVMR sensitivity analyses



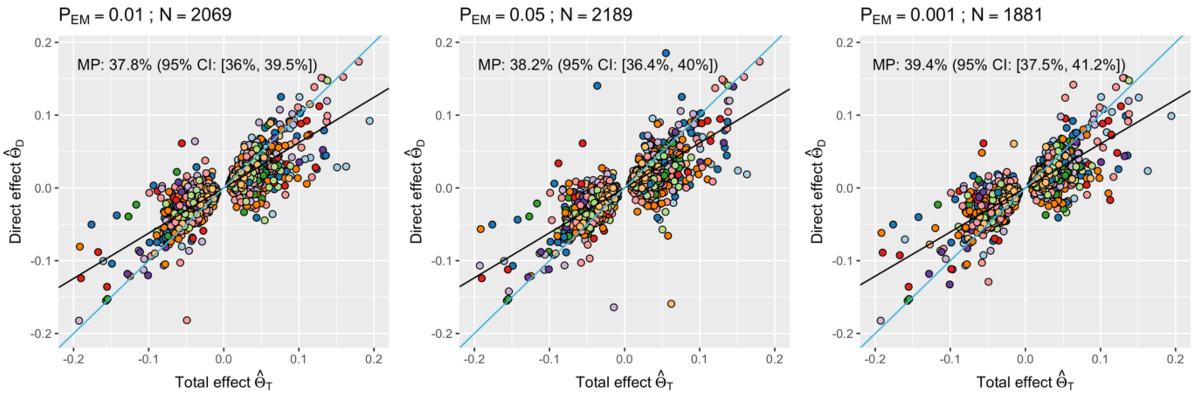
Supplementary Figure 10. MVMR sensitivity analysis to stratify DNAm-trait pairs by their conditional F-statistic. At an F-statistic below 10, the mediation analysis might suffer from weak instrument bias which can result in unreliable direct effect estimates. Among the 2,069 DNAm-trait pairs, 1,061 had an F-statistic above 10 and 1,008 below. DNAm-trait pairs are colour-coded by the physiological category of the trait as defined in Supplementary Fig. 8. The slope is plotted in black (numerical values shown in plotting area resulting from the regression of $\hat{\Theta}_D$ against $\hat{\Theta}_T$) and the identity line in blue.



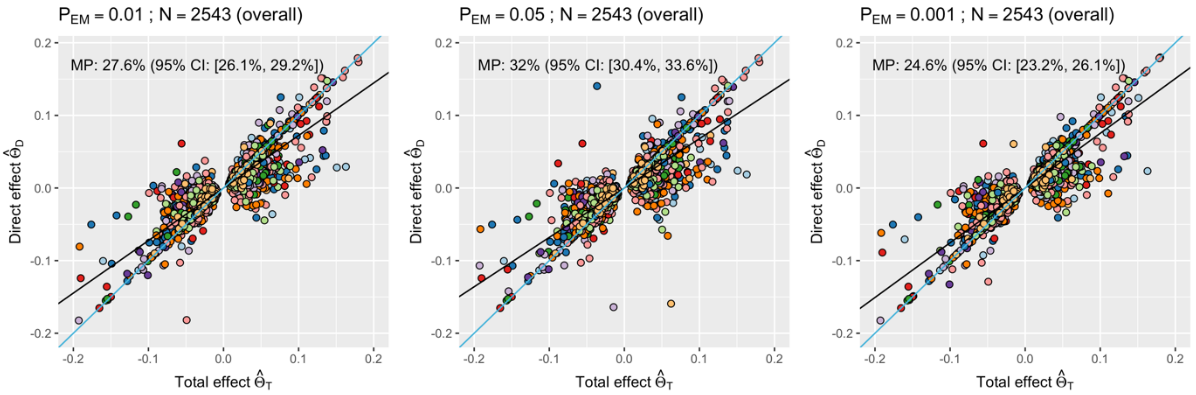
Supplementary Figure 11. Conditional F-statistics with and without the correlation matrix between mediators. Conditional F-statistics with transcript-transcript correlations were calculated for all DNAm-trait pairs with at least 2 mediators and for which at least half of them had available correlation data. In total, 1,208 pair were assessed with the mean F-statistics being 13.55 with the correlation matrix and 12.85 without. The slope is plotted in black and the identity line in blue.



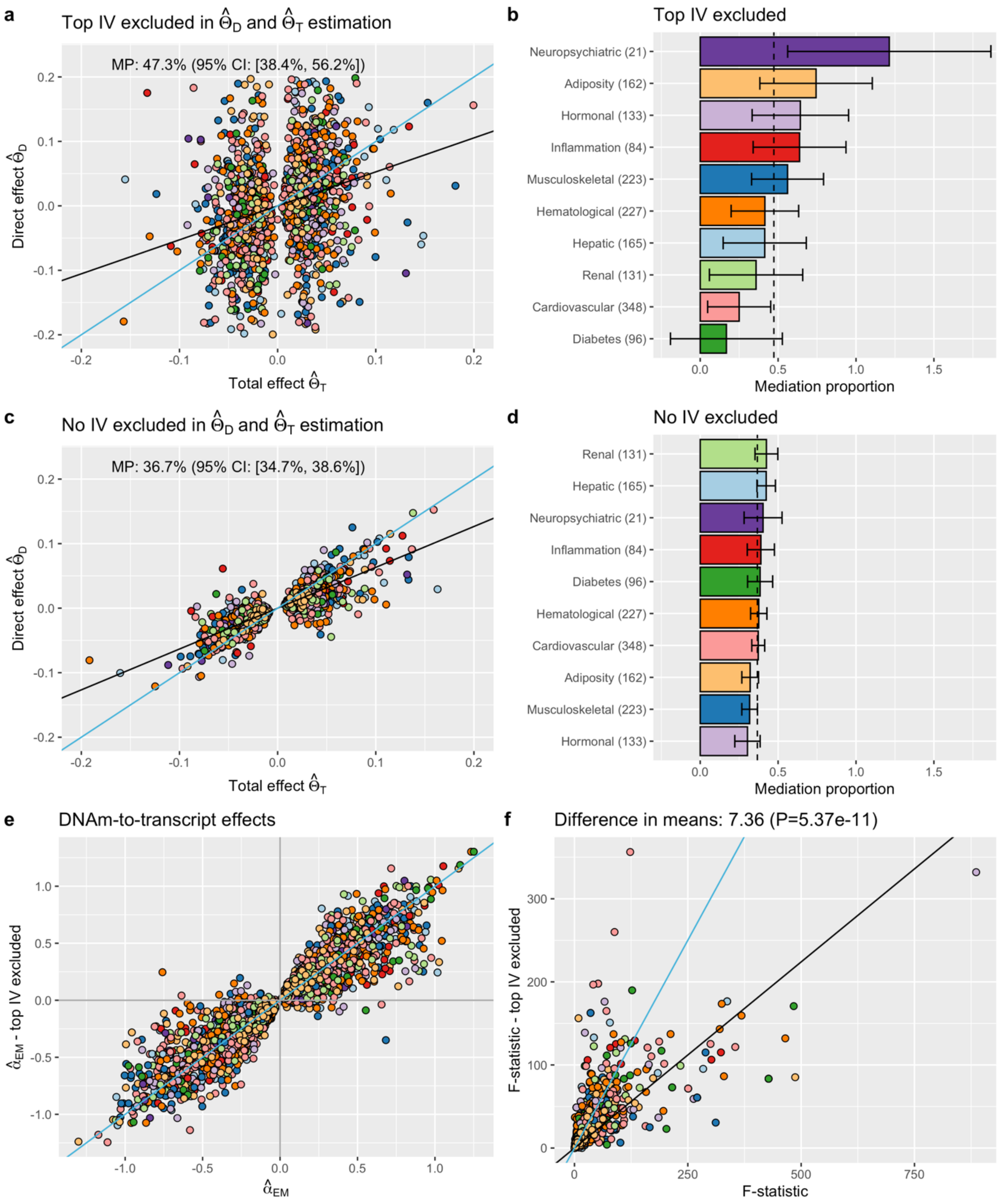
Supplementary Figure 12. MVMR sensitivity analysis to test for heterogeneity within the IV set. In the left figure, the 2,069 were filtered for those that showed no signs of heterogeneity in the univariable MR analyses (Q-heterogeneity p-value > 0.01; 1,757 pairs). In the right figure, the filtering was applied on the p-values of the Q-statistics of both the univariable and multivariable MR analyses (1,405 pairs). DNAm-trait pairs are colour-coded by the physiological category of the trait as defined in Supplementary Fig. 8. The slope is plotted in black (numerical values shown in plotting area resulting from the regression of $\widehat{\Theta}_D$ against $\widehat{\Theta}_T$) and the identity line in blue.



Supplementary Figure 13. MVMR sensitivity analysis to assess the influence of the P_{EM} thresholds to select mediators in case of "detectable mediation" analyses. Shown are the results for three thresholds (0.01, 0.05 and 0.001 from left to right). The calculation of the MP is done on DNAm-trait pairs (*N* pairs) with at least 1 transcript in the *cis* region causally associated to the DNAm site. DNAm-trait pairs are colour-coded by the physiological category of the trait as defined in Supplemetary Fig. 8. The slope is plotted in black (numerical values shown in plotting area resulting from the regression of $\widehat{\Theta}_D$ against $\widehat{\Theta}_T$) and the identity line in blue.



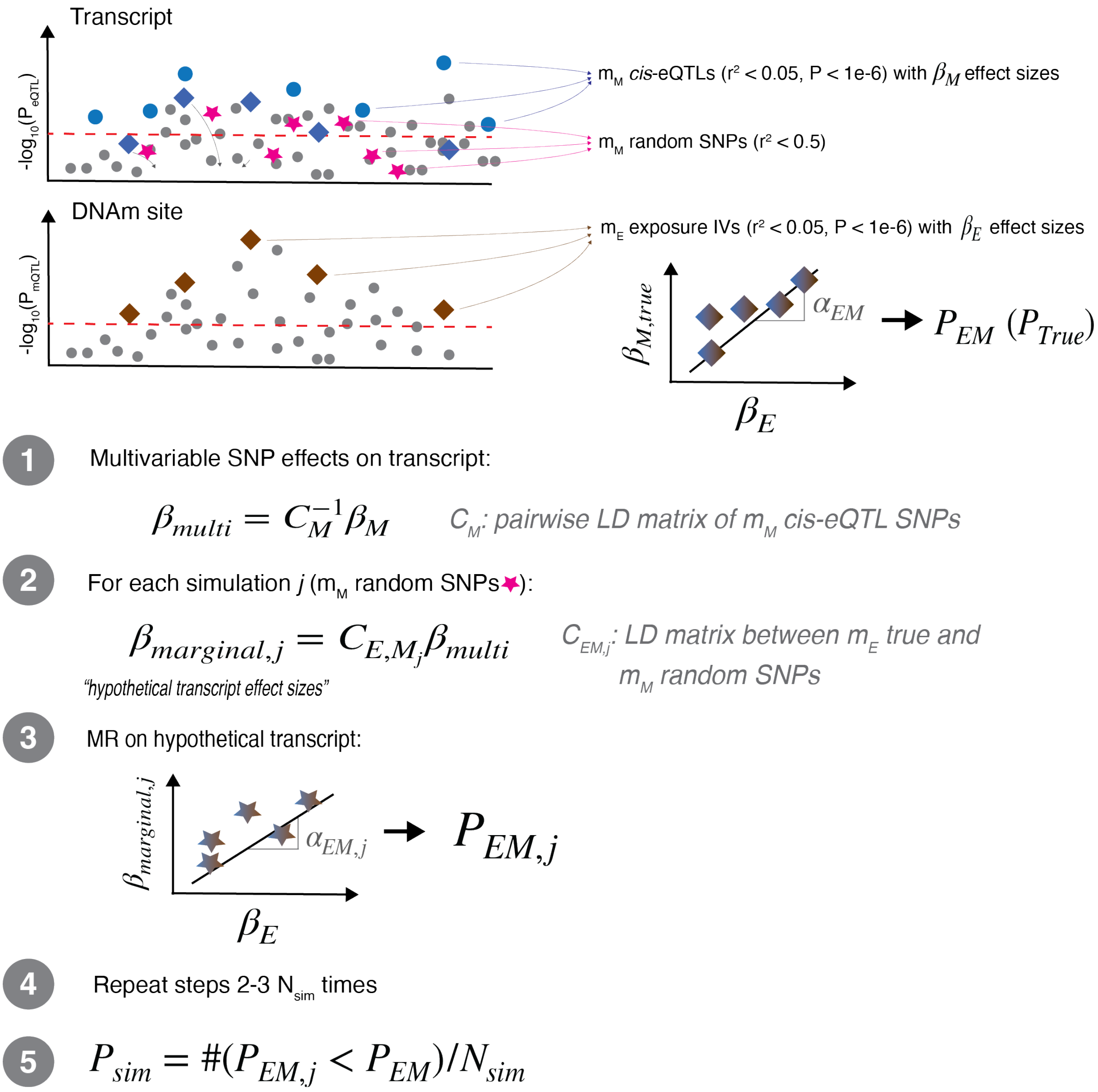
Supplementary Figure 14. MVMR sensitivity analysis to assess the influence of the P_{EM} thresholds to select mediators in case of the overall MP. Shown are the results for three thresholds (0.01, 0.05 and 0.001 from left to right). The overall MP is calculated on all DNAm-trait pairs (*N* pairs) with least 1 transcript in the *cis* region (not necessarily causally associated to the exposure) and for which a mediation analysis could be performed in all three settings (the number of IVs being the limiting factor). DNAm-trait pairs are colour-coded by the physiological category of the trait as defined in Supplementary Fig. 8. The slope is plotted in black (numerical values shown in plotting area resulting from the regression of $\widehat{\Theta}_D$ against $\widehat{\Theta}_T$) and the identity line in blue.



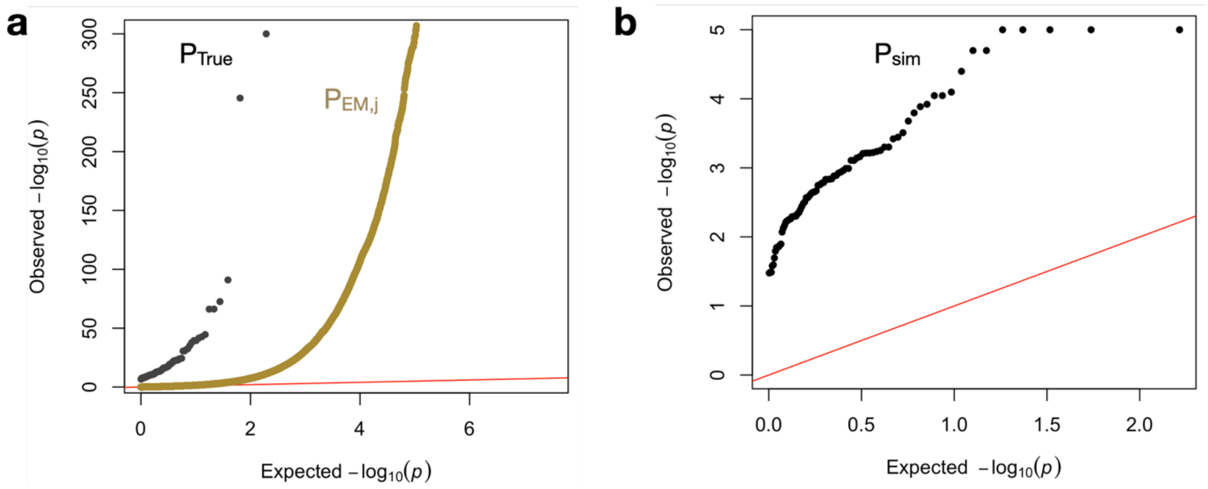
Supplementary Figure 15. MVMR sensitivity analysis excluding the top instrumental variable (pleiotropy sensitivity analysis). Mediation analyses were conducted for all DNAm-trait pairs with at least 3 exposure-associated IVs after excluding the top IV (i.e., exposure-associated IV with the lowest p-value; 1,590 DNAm-trait pairs). **a** $\widehat{\mathrm{MP}}$ and 95% CI calculated on these pairs excluding the top IV in both the total and direct effect calculation. The slope is shown by the black line and the identity line by the blue line. **b** Corresponding MPs of traits grouped by physiological categories. The vertical dotted line corresponds to the mean MP across all DNAm-trait pairs and error bars represent the 95% CI. $\widehat{\mathrm{MP}}$s per category were derived by regressing $\hat{\Theta}_D$ against $\hat{\Theta}_T$. **c**, **d** Same analysis as in **a** and **b**, respectively, but without excluding the top IV (same 1,590 pairs). **e** DNAm-to-transcript MR effects (α_{EM}) of the exposure-mediator pairs included in the mediation analyses of the 1,590 DNAm-trait pairs are shown before and after the exclusion of the top IV. **f** Conditional F-statistics calculated on the 1,590 DNAm-trait pairs before and after the exclusion of the top IV. Conditional F-statistics were on average 7.36 higher before excluding the top IV (two-sided t-test p-value = 5.37e-11) pointing out that weak instrument bias was more present in the analyses where the top IV was missing.

Overall, the analyses show that excluding the top IV results in noisier MR estimates as a consequence of weaker instruments. While the top IV is crucial in getting robust molecular MR estimates, the analyses show that the remaining IVs support same effect size magnitudes and directionalities as the top IV.

Specifically, excluding the top IV significantly increased the MP estimated over all the DNAm-trait pairs (panel **a** vs **c**, P_{diff} = 0.0228). This difference is likely due to weak instrument bias as it was not present for pairs with F>10 (845/1,590 pairs, MP = 40.9%, 95% CI: [29.3%, 52.4%] – top IV excluded; P_{diff} = 0.48). The estimated MP did not depend on the conditional F-statistic when all IVs were considered (Supplementary Fig. 10).


Transcript
$-\log_{10}(P_{eQTL})$
m_M *cis*-eQTLs ($r^2 < 0.05$, $P < 1e\text{-}6$) with β_M effect sizes
m_M random SNPs ($r^2 < 0.5$)
DNAm site
$-\log_{10}(P_{mQTL})$
m_E exposure IVs ($r^2 < 0.05$, $P < 1e\text{-}6$) with β_E effect sizes
$\beta_{M,true}$
α_{EM}
β_E
P_{EM} (P_{True})
1
Multivariable SNP effects on transcript:
$\beta_{multi} = C_M^{-1}\beta_M$
C_M: pairwise LD matrix of m_M cis-eQTL SNPs
2
For each simulation *j* (m_M random SNPs):
$\beta_{marginal,j} = C_{E,M_j}\beta_{multi}$
"hypothetical transcript effect sizes"
$C_{EM,j}$: LD matrix between m_E true and m_M random SNPs
3
MR on hypothetical transcript:
$\beta_{marginal,j}$
$\alpha_{EM,j}$
β_E
$P_{EM,j}$
4
Repeat steps 2-3 N_{sim} times
5
$P_{sim} = \#(P_{EM,j} < P_{EM})/N_{sim}$

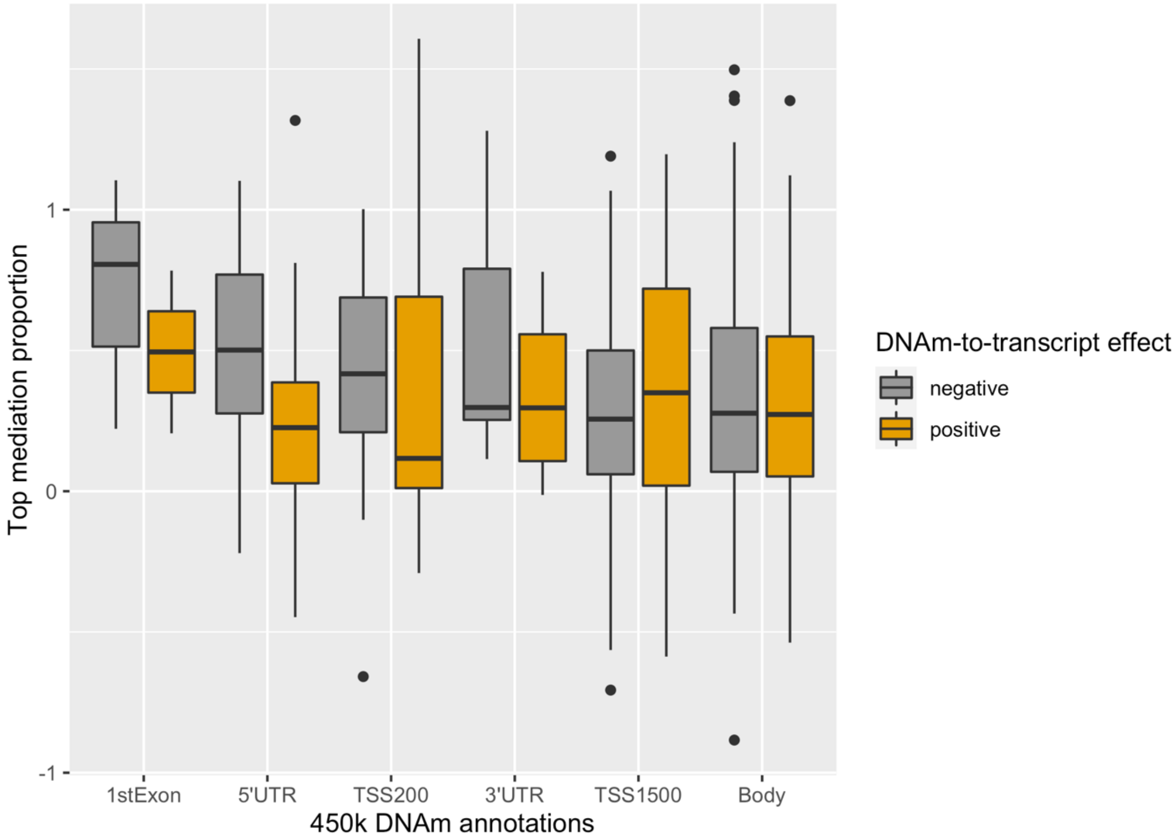
Supplementary Figure 16. Schematic illustrating the horizontal pleiotropy simulation analysis to assess the possibility of DNAm-to-transcript associations because of horizontal pleiotropy as a result of LD between mQTLs and eQTLs. First DNAm-transcript pairs with a significant MR effect at $P_{EM} < 1e\text{-}6$ are selected. Then, multivariable SNP effects on the transcript are calculated based on independent *cis*-eQTLs (step 1). In each of the following simulations, m_M random SNPs are selected for which marginal SNP-transcript effects are calculated. Note that these hypothetical transcript effect sizes have identical multivariable eQTL effect size distribution as the real transcript (step 2). Next, a univariable MR analysis on this hypothetical transcript yields $P_{EM,j}$ (step 3). Steps 2-3 are repeated N_{sim} times (step 4) which allows to calculate the simulation p-value P_{sim} (step 5).



Supplementary Figure 17. Simulation analysis to assess the possibility of DNAm-to-transcript associations due to horizontal pleiotropy. For a significant DNAm-to-transcript MR association (P_{EM}, herein called P_{True}), we performed simulation tests (N_{sim} = 100,000) by randomly selecting eQTL-SNPs with identical multivariable eQTL effects in the region (Supplementary Fig. 16). Each simulated marginal eQTL effect estimate resulted in a random DNAm-to-transcript MR estimate ($P_{\mathrm{EM,j}}$) from which we could derive the simulation p-value ($P_{\mathrm{sim}} = \#(P_{\mathrm{EM,j}} < P_{\mathrm{True}})/N_{\mathrm{sim}}$). **a** Comparison (normal QQ-plot) of the true (P_{True}) and random ($P_{\mathrm{EM,j}}$) p-values from 100 DNAm-transcript MR estimates (the transcript outcome being the true and hypothetical transcript *j* from each simulation run, respectively). **b** Normal QQ-plot of the simulation p-values P_{sim}.

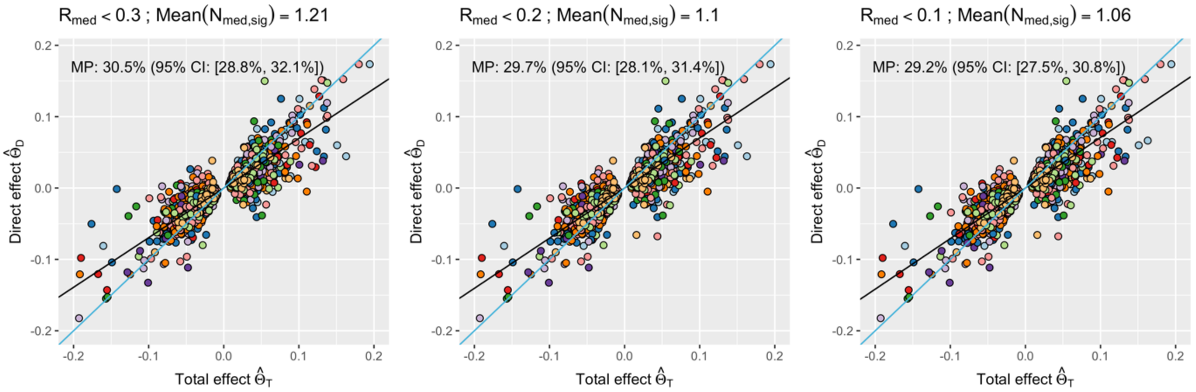
The analysis shows that while MR p-values from hypothetical transcript effects are inflated, they are much less significant than the true p-values ensuring that horizontal pleiotropy is not at the root of observed methylation-expression causal effects.

Stratification by DNAm annotations

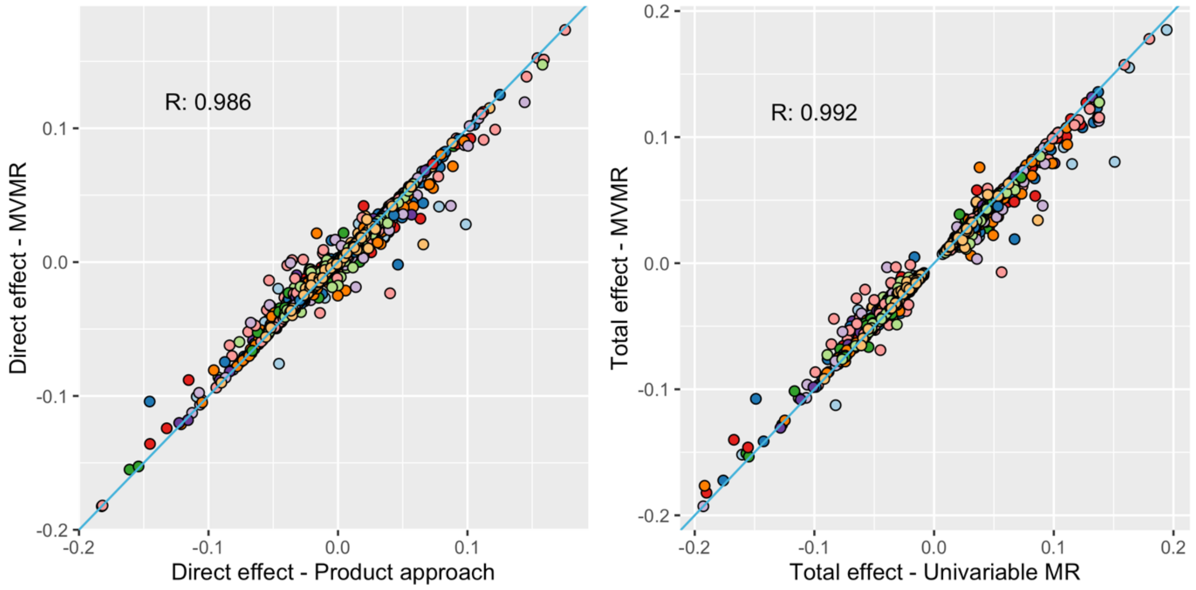


Supplementary Figure 18. Mediation proportion stratified by DNAm site location. Boxplots representing the top mediation proportion (MP_{top}) stratified by DNAm site location with respect to the top mediator and by the causal effect direction of the DNAm on the transcript level. The annotation groups are shown in decreasing order with respect to the mediation proportion (negative and positive DNAm-to-transcript effect pairs combined). Number of DNAm-trait pairs within each boxplot are as follows: 1stExon (negative: 3, positive: 2), 5'UTR (negative: 60, positive: 26), TSS200 (negative: 33, positive: 15), 3'UTR (negative: 14, positive: 12), TSS1500 (negative: 68, positive: 66), Body (negative: 172, positive: 104). Boxes bound the 25th, 50th (median, centre), and the 75th quantile. Whiskers range from minima (Q1 – 1.5*IQR) to maxima (Q3 + 1.5*IQR) with points above or below representing potential outliers. Note that annotations were not available for all DNAm sites and DNAm sites mapping to multiple annotations were omitted.

Mediation analyses with uncorrelated mediators

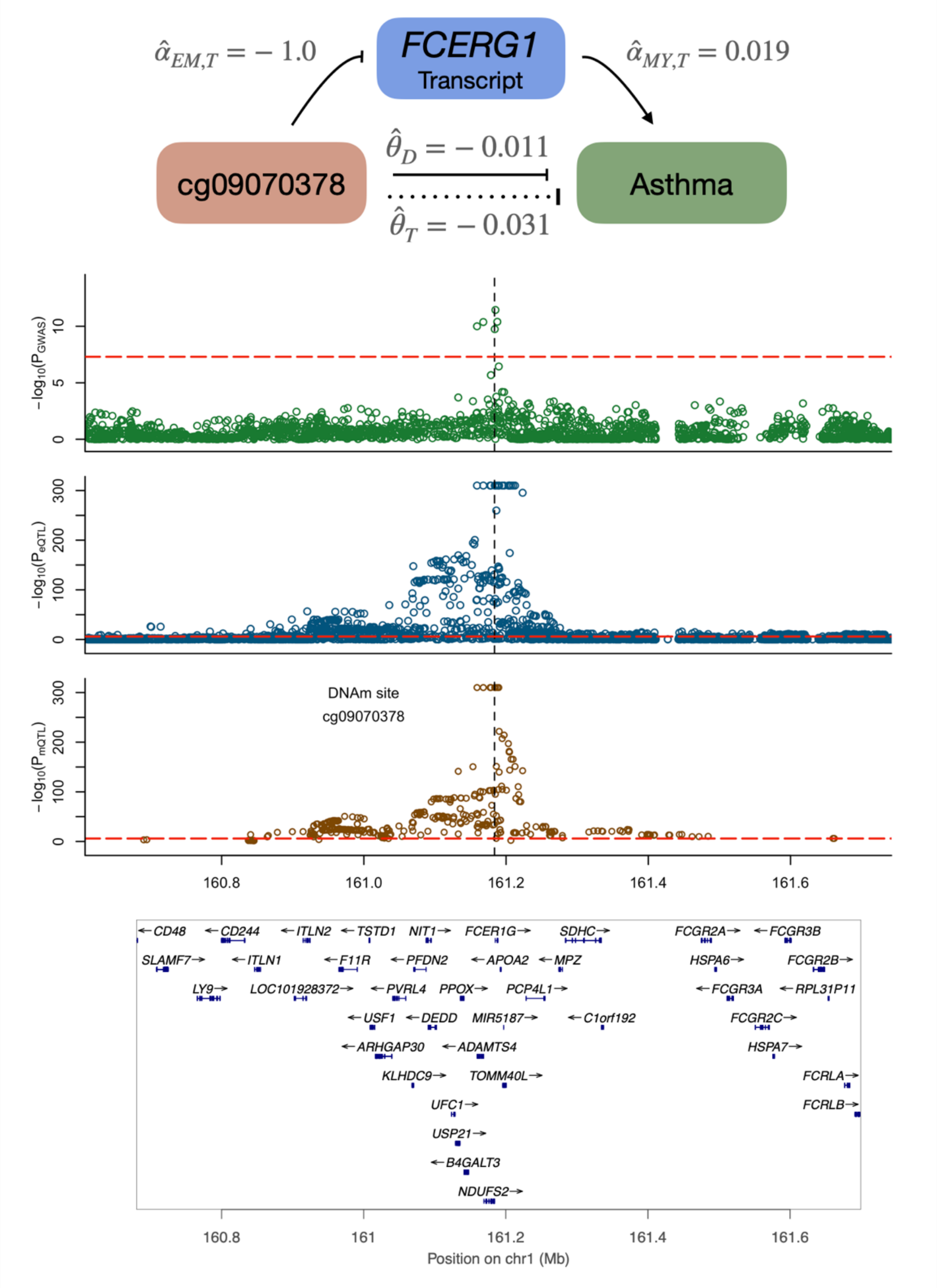


Supplementary Figure 19. Mediation analysis with uncorrelated mediators. Mediation analyses conducted with uncorrelated mediators at different R_{med} thresholds (0.3, 0.2, and 0.1 from left to right) for all 2,069 DNAm-trait pairs (colour-coded by the physiological category of the trait as defined in Supplementary Fig. 8). R_{med} is the maximum correlation between the mediators for a given exposure-outcome pair. As this threshold decreases, the average number of selected mediators ($N_{\text{med,sig}}$) decreases. The slope (black line) and the mediation proportion together with the 95% CI are displayed in the plot area (blue line represents the identity line).

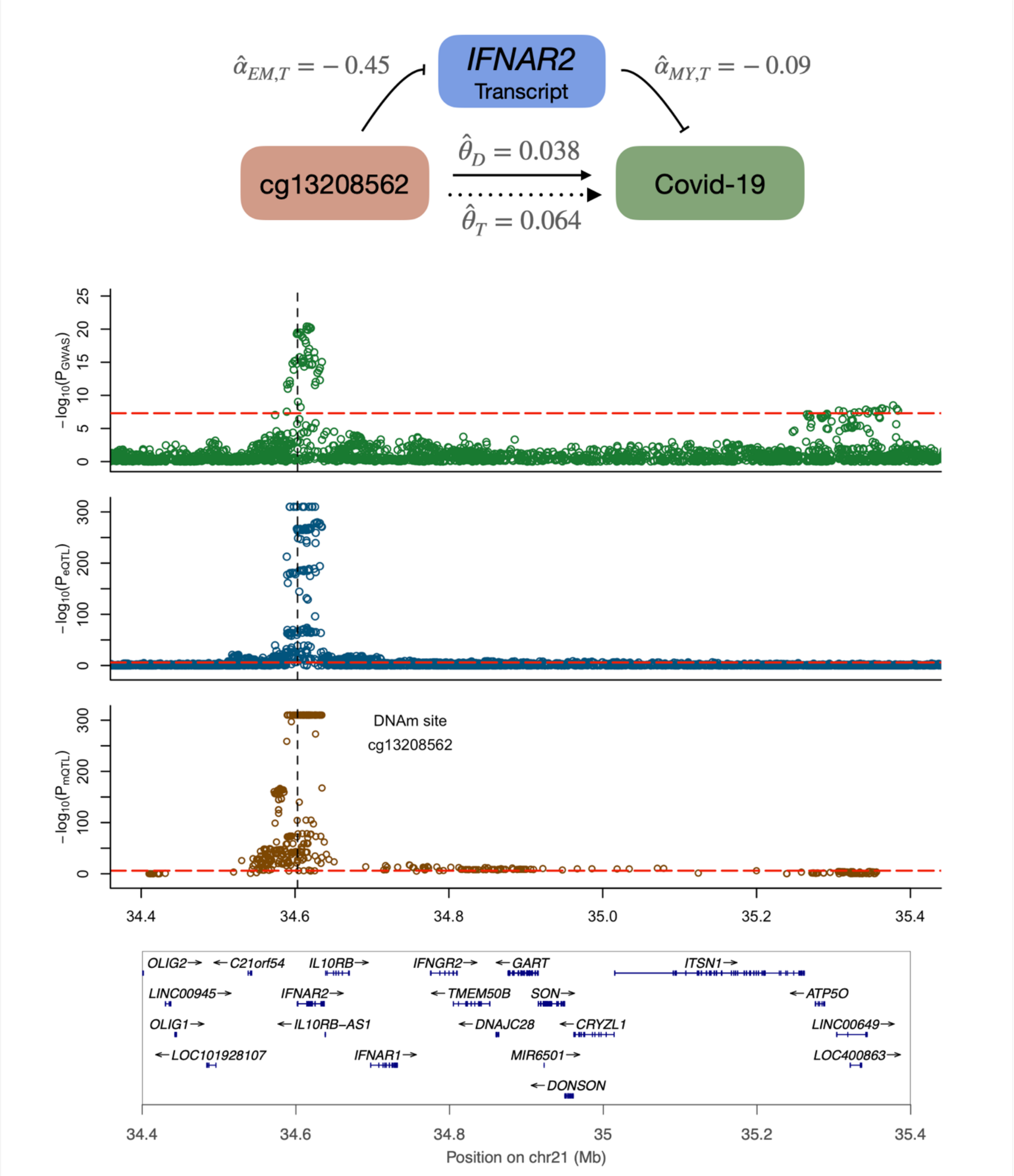


Supplementary Figure 20. **Agreement between the product of coefficients and difference in coefficients methods to estimate direct and indirect effects.** In the left panel, the agreement between direct effects estimated from the multivariable Mendelian randomization regression (MVMR, difference in coefficients methods) and direct effects from the product approach is shown. In the product approach, exposure-to-mediator effects are multiplied with mediator-to-outcome direct effects and summed up across mediators to get the indirect effect. The direct effect is then calculated by subtracting this indirect effect from the total effect. The right panel shows the agreement between total effects obtained from the univariable MR regression and total effects reconstructed by summing the direct and indirect effects derived from the MVMR regressions. In the latter, the direct effect refers to the "Direct effect – MVMR" from the left panel and the indirect effect to the one obtained in the product approach. The *R* coefficient displayed in the plot area is the Pearson correlation coefficient (identity line is plotted in blue). Results are shown for all 2,069 DNAm-trait pairs colour-coded by the physiological category of the trait as defined in Supplementary Fig. 8.

Multi-omics mechanisms of action



Supplementary Figure 21. Plausible DNAm-transcript-trait regulatory mechanism for asthma disease at the *FCERG1* locus. The top row displays a schematic of the mechanism with the calculated univariable and multivariable MR effects. The three following rows show the regional SNP associations (-log$_{10}$(p-values)) with the trait (green), transcript (blue) and DNAm probe (brown), respectively. Red dashed lines indicate the significance thresholds of the respective SNP associations and the vertical black dashed line represents the DNAm probe position. The bottom row shows the positions of the genes in the locus with their respective strand direction.



Supplementary Figure 22. Plausible DNAm-transcript-trait regulatory mechanism for Covid-19 (hospitalized vs population) at the *IFNAR2* locus. Same figure composition as Supplementary Fig. 21.

Supplementary Tables

Supplementary Table 1. Means and interquartile ranges of the simulation parameters as observed in real data. The full distribution of each parameter is shown in Supplementary Fig. 5.

	Quartile 1	Median	Mean	Quartile 3
$N_{\text{med,pot}}$	7	12	14.7	21
$N_{\text{med,sig}}$	1	2	3.3	4
m_E	3	4	5.09	6
m_M	1	3	5.65	8
h^2_E	0.179	0.319	0.403	0.539
$h^2_{M,\text{direct}}$	4.75E-03	0.0148	0.0418	0.047
var(α_{EM})	5.48E-03	0.0196	0.0789	0.0751
var(α_{MY})	1.18E-04	7.37E-04	9.52E-03	3.78E-03
ρ	-0.39	-0.0216	-0.0112	0.361

Supplementary Table 2. Values used in the different simulation settings to mimic mediation of DNAm-to-trait effects through transcript levels. Results of the default model are shown in Supplementary Fig. 6, results of varying the sample size N_{M}, the mediator selection threshold P_{EM}, and heritabilities $h^2_{\text{M,direct}}$ and h^2_{E} in Fig. 2, and the remaining simulation settings in Supplementary Fig. 7. Median parameter values are used in the default model and values comprising the interquartile range when varying the respective parameter.

<table>
<tr><th></th><th>Default model (median values)</th><th>Varying N_M</th><th>Varying P_{EM}</th><th>Varying $h^2_{M,\text{direct}}$</th><th>Varying h^2_E</th><th>Varying m_E</th><th>Varying $N_{\text{med,sig}}$</th></tr>
<tr><td>N_{med}</td><td colspan="2">12</td><td>20</td><td colspan="3">12</td><td>20</td></tr>
<tr><td>$N_{\text{med,sig}}$</td><td colspan="6">2</td><td>[1 - 10]</td></tr>
<tr><td>m_E</td><td colspan="5">4</td><td>[3 - 12]</td><td>4</td></tr>
<tr><td>m_M</td><td colspan="7">3</td></tr>
<tr><td>h^2_E</td><td colspan="4">0.319</td><td>[0.05 - 1]</td><td>0.1, 0.3, 0.5</td><td>0.319</td></tr>
<tr><td>$h^2_{M,\text{direct}}$</td><td colspan="3">0.0148</td><td>[3E-04 - 0.64]</td><td colspan="2">0.0148</td><td>4.75E-03, 0.0148, 0.047</td></tr>
<tr><td>ρ</td><td colspan="7">-0.02</td></tr>
<tr><td>P_{EM}</td><td colspan="2">0.01</td><td>[1E-06 - 1]</td><td colspan="4">0.01</td></tr>
<tr><td>N_M</td><td>30,000</td><td>[100-100,000]</td><td colspan="5">30,000</td></tr>
<tr><td>N_E</td><td colspan="7">30,000</td></tr>
<tr><td>N_Y</td><td colspan="7">300,000</td></tr>
</table>

var(α_{EM})	0.02
var(α_{MY})	1.00E-03
MP	0.35

Supplementary Table 3. Enrichment analysis of negative DNAm-to-transcript causal effects within each annotation group. The first two columns show the number of distinct DNAm-transcript pairs with negative and positive causal effects, respectively. An enrichment analysis for negative causal effects was conducted as a two-sided Fisher's test where each annotation group was tested against the remaining other groups combined. Annotation groups significantly enriched or deprived for negative causal effects (after correcting for multiple testing at $P < 0.05/6$) are highlighted in bold.

Annotation group	DNAm → Transcript negative effect	DNAm → Transcript positive effect	Proportion of negative effects	OR (negative effect enrichment)	P-value (negative effect enrichment)
1stExon	291	188	0.608	**1.33**	**2.67E-03**
3'UTR	863	828	0.510	0.89	1.63E-02
5'UTR	1773	1429	0.554	1.07	8.03E-02
Body	9675	8824	0.523	**0.87**	**2.15E-10**
TSS1500	4380	3433	0.561	**1.12**	**1.24E-05**
TSS200	1702	1284	0.570	**1.15**	**3.81E-04**

Appendix B

Multi-layered genetic approaches to identify approved drug targets

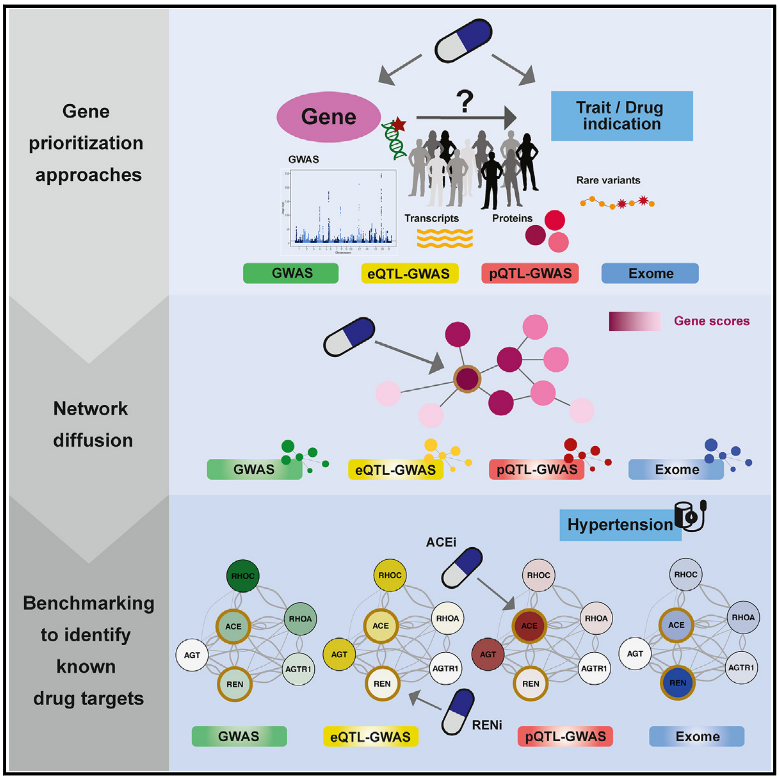
This article (Sadler *et al.*, 2023, *Cell Genomics*) is presented in Chapter 3.

Cell Genomics

Article

Multi-layered genetic approaches to identify approved drug targets

Graphical abstract



Highlights

- Gene prioritization methods based on GWAS, exome, and QTL data
- Comparison of methods for drug target identification across 30 clinical traits
- We found a 1.3- to 2.2-fold enrichment for drug targets among prioritized genes
- Network diffusion of prioritized genes significantly boosted performance

Authors

Marie C. Sadler, Chiara Auwerx, Patrick Deelen, Zoltán Kutalik

Correspondence

zoltan.kutalik@unil.ch

In brief

Sadler et al. compared and benchmarked genetically informed approaches combined with network diffusion to prioritize drug target genes. Gene prioritization methods were based on large-scale genetic studies such as genome-wide association and whole-exome studies, as well as tissue-wide and whole-blood expression and protein quantitative trait loci.

Sadler et al., 2023, Cell Genomics *3*, 100341
July 12, 2023 © 2023 The Author(s).
https://doi.org/10.1016/j.xgen.2023.100341

Cell Genomics

CellPress
OPEN ACCESS

Article

Multi-layered genetic approaches to identify approved drug targets

Marie C. Sadler,[1,2,3] Chiara Auwerx,[1,2,3,4] Patrick Deelen,[5,6] and Zoltán Kutalik[1,2,3,7,*]
[1]University Center for Primary Care and Public Health, Route de Berne 113, 1010 Lausanne, Switzerland
[2]Swiss Institute of Bioinformatics, Quartier Sorge, 1015 Lausanne, Switzerland
[3]Department of Computational Biology, University of Lausanne, 1015 Lausanne, Switzerland
[4]Center for Integrative Genomics, University of Lausanne, 1015 Lausanne, Switzerland
[5]Department of Genetics, University of Groningen, 9700 Groningen, the Netherlands
[6]Oncode Institute, 3521 Utrecht, the Netherlands
[7]Lead contact
*Correspondence: zoltan.kutalik@unil.ch

https://doi.org/10.1016/j.xgen.2023.100341

SUMMARY

Drugs targeting genes linked to disease via evidence from human genetics have increased odds of approval. Approaches to prioritize such genes include genome-wide association studies (GWASs), rare variant burden tests in exome sequencing studies (Exome), or integration of a GWAS with expression/protein quantitative trait loci (eQTL/pQTL-GWAS). Here, we compare gene-prioritization approaches on 30 clinically relevant traits and benchmark their ability to recover drug targets. Across traits, prioritized genes were enriched for drug targets with odds ratios (ORs) of 2.17, 2.04, 1.81, and 1.31 for the GWAS, eQTL-GWAS, Exome, and pQTL-GWAS methods, respectively. Adjusting for differences in testable genes and sample sizes, GWAS outperforms e/pQTL-GWAS, but not the Exome approach. Furthermore, performance increased through gene network diffusion, although the node degree, being the best predictor (OR = 8.7), revealed strong bias in literature-curated networks. In conclusion, we systematically assessed strategies to prioritize drug target genes, highlighting the promises and pitfalls of current approaches.

INTRODUCTION

Drugs whose targets have genetic support were found to be more likely to succeed in clinical trials.[1,2] Although multiple methods have been proposed to establish such genetic support, leveraging genetic data to find disease genes, and ultimately drug targets, has proven to be challenging.[3–6] The most straightforward approach maps genome-wide association study (GWAS) signals to the closest genes, with more sophisticated methods incorporating linkage disequilibrium (LD) structure and gene annotation information to compute gene scores.[7–9] Over the past decade, large-scale molecular quantitative trait loci (mQTL) datasets facilitated the discovery of disease mechanisms and the identification of potential new drug targets.[10–15] Several methods, including Mendelian randomization studies, transcriptome-wide association studies, and colocalization methods have integrated expression and protein QTL data with GWASs to pinpoint likely causal genes for complex traits and diseases.[16–22] More recently, the availability of high-throughput sequencing data enabled the discovery and analysis of rare variants and their aggregated effects to reveal gene-disease associations.[23,24] Whole-exome sequencing (WES) in the UK Biobank (UKBB) showed that genes prioritized this way are 3.6 times more likely to be targets of drugs approved by the US Food and Drug Administration (US FDA).[25]

Genes prioritized by GWASs, mQTL-GWAS integration methods, and WES burden tests may not be drug targets themselves, but may be up- or downstream of those in pharmacological pathways. Propagating gene prioritization scores on networks has proven to be a promising approach to identify known drug target genes.[26–30] Starting from seed genes (i.e., prioritized disease-associated genes), network connectivity can identify neighboring genes that strongly interact with disease genes, but lack direct genetic evidence that explains their therapeutic effect. Gene networks can be derived from literature or high-throughput experiments and thus are prone to yielding very different results when used for (seed) gene score diffusion.[31]

Here, we took a comprehensive approach to examine the contribution of each method component to the success of drug target prioritization. First, we focused on four different approaches to prioritize (seed) genes: (1) LD-aware gene score computation from the largest GWASs with full publicly available summary statistics (Pascal[9]); (2) Mendelian randomization (MR) combining tissue-wide expression QTLs and GWASs (eQTL-GWAS); (3) MR combining plasma protein QTL with GWAS (pQTL-GWAS); and (4) UKBB WES burden tests (Exome). We then used three different networks to diffuse the seed gene scores: (1) the STRING protein-protein interaction (PPI) network[32]; (2) an RNA-sequencing (RNA-seq) coexpression network[33]; and (3) the FAVA network.[34] All 12 combinations of

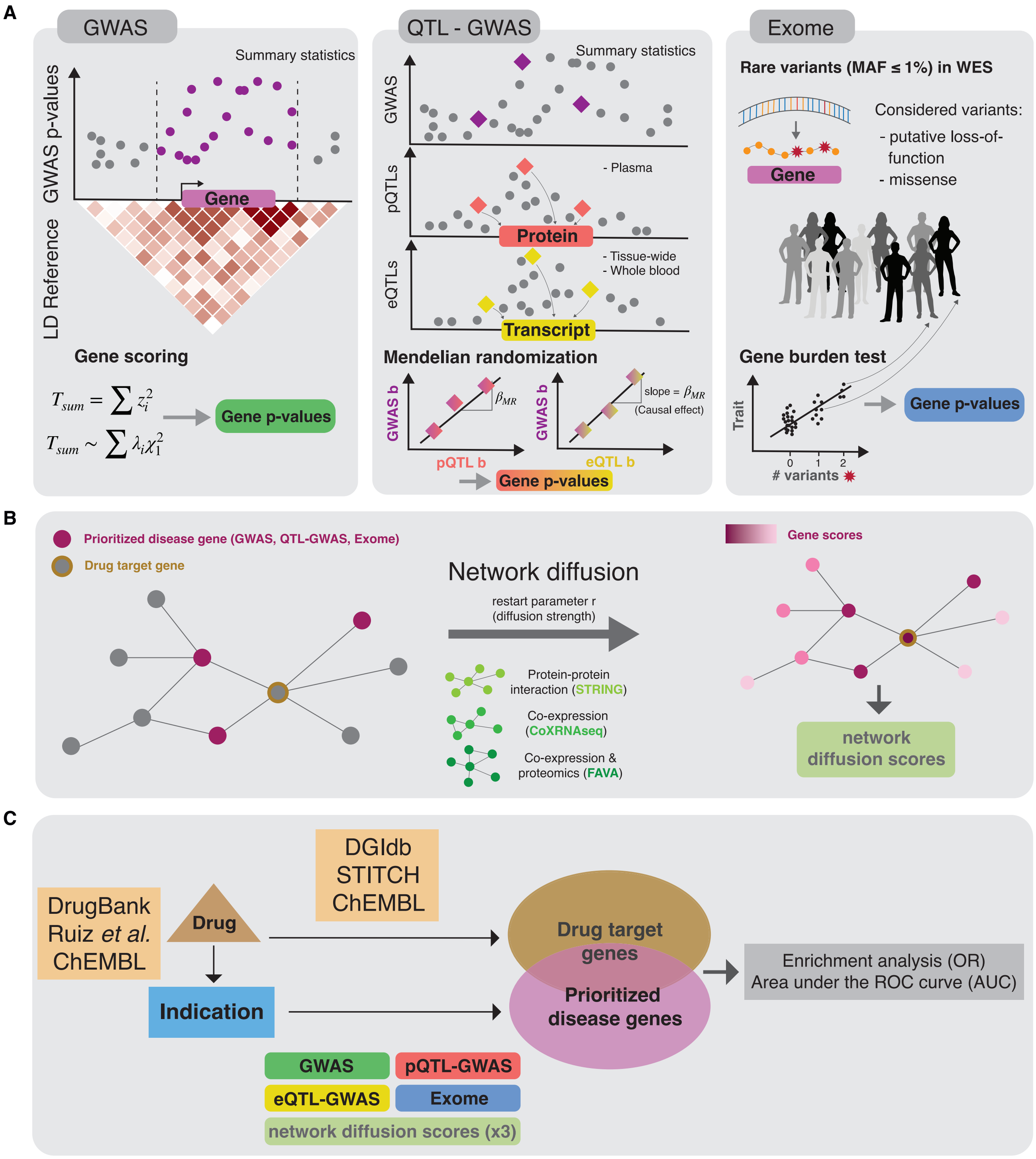


Figure 1. Overview of the analysis workflow

(A) Three different gene prioritization methods were tested in this study. The first one uses GWAS summary statistics as input (GWAS). The second combines molecular QTL and GWAS summary statistics (QTL-GWAS): either expression QTL (eQTL) or protein QTL (pQTL) data. The third leverages individual-level whole-exome sequencing (WES) data (Exome). In the GWAS method, gene p values are based on the sum of squared SNP Z scores (T_{sum}) that follows a weighted χ_1^2 distribution. The QTL-GWAS method integrates QTL and GWAS summary statistics through Mendelian randomization (MR). MR causal effect sizes (β_{MR}) are calculated from GWAS and mQTL effect sizes (GWAS b and mQTL b, respectively) and gene scores are the corresponding p values. The Exome method aggregates rare variants from WES data. Putative loss-of-function and missense variants with minor allele frequencies (MAF) below 1% are collapsed in burden tests, which results in gene p values. The different approaches were benchmarked for their ability to prioritize drug target genes.

(legend continued on next page)

the four seed-generating methods and the three networks were applied to 30 traits (Figure 1) using five different reference sets of target genes (DrugBank,[35] Ruiz et al.,[36] ChEMBL,[37] DGIdb,[38] and STITCH[39]). Overall, we provide an in-depth comparison of all combinations of these approaches, identifying their respective strengths and caveats.

RESULTS

Overview of the analysis

In this study, we calculated gene prioritization scores and tested their ability to identify drug targets across 30 traits (Figure 1). We focused on three types of method, termed GWAS, QTL-GWAS, and Exome, that allow the computation of gene scores provided genetic association data (Figure 1A).

The GWAS method takes as input GWAS summary statistics together with a matching LD reference panel. Gene p values are calculated based on the sum of squared test statistics for SNPs falling into the gene region.[9] The QTL-GWAS methods integrate GWAS summary statistics with mQTL data for the gene of interest. We calculated gene scores using (1) eQTL-GWAS data from the largest available whole-blood eQTL study (eQTLGen study, n = 31,684)[13] as well as tissue-wide eQTL data from the GTEx Consortium v.8 (n = 65–573 for 48 tissue types)[40] and (2) pQTL-GWAS data from the largest available plasma pQTL study (deCODE study, n = 35,559).[14] Integration was done by performing MR analyses using either the protein or the transcript as exposure and the GWAS trait as outcome. If not specified otherwise, the eQTL-GWAS method refers to the tissue-wide analysis in which the eQTLGen and GTEx data are combined by considering the tissue for which the MR effect was the most significant (STAR Methods). While the GWAS and QTL-GWAS methods focus on common genetic variants, the Exome method considers only rare variants from WES data with minor allele frequencies (MAFs) below 1%. Gene scores were based on gene burden tests that aggregate putative loss-of-function and missense variants, and we used the resulting p values from the WES analysis in the UKBB.[25] To allow for a fair comparison with the Exome method while also exploiting disease-specific consortium GWAS summary statistics with maximized case counts, we calculated gene prioritization scores for the GWAS and QTL-GWAS methods using both consortium GWAS and UKBB GWAS data that matched Exome sample sizes (Tables S1 and S2; STAR Methods).

Disease genes may not coincide with drug target genes, but they may be in close proximity in terms of molecular interaction (Figure 1B). Through diffusion based on random walks, we leveraged network connectivity to prioritize neighbors of disease genes, which may be drug targets. We tested this hypothesis on three different network types: the STRING PPI network, which relies on literature interactions, among other data types[32]; a gene coexpression network based on 31,499 RNA-seq samples (CoXRNAseq)[33]; and a gene coexpression network based on single-cell RNA-seq and proteomics data (FAVA).[34] Gene prioritization scores were obtained following diffusion at six different restart parameter values (r = 0, 0.2, 0.4, 0.6, 0.8, 1) (STAR Methods).

Disease drug target genes were defined using public databases. Specifically, drug-disease indications were retrieved from DrugBank,[35] Ruiz et al.,[36] and ChEMBL,[37] while drug-drug target pairs originated from DGIdb,[38] STITCH,[39] and ChEMBL.[37] Drug target enrichment analyses were calculated for the following five database combinations: DrugBank/DGIdb, DrugBank/STITCH, Ruiz/DGIdb, Ruiz/STITCH, and ChEMBL/ChEMBL.

Finally, prioritized disease genes, defined as the top 1% of genes identified through the 12 combinations of gene prioritization and network diffusion methods (5% for combinations involving the pQTL-GWAS method to account for the smaller set of testable genes), were then tested for enrichment with the five drug target genes using Fisher's exact test (Figure 1C). Background genes were defined as all genes that could be tested by the respective method, and sensitivity analyses were performed on background genes testable for all methods. Second, we calculated the area under the receiver operating characteristic curve (AUC) values, which has the advantage of not requiring any thresholds. To compute a combined enrichment score per method, we aggregated results across traits and drug databases termed overall odds ratios (ORs) or overall AUC values (STAR Methods).

Concordance of prioritized genes among gene scoring methods

We first analyzed whether genes prioritized by the GWAS, QTL-GWAS, and Exome methods were concordant (Figure 2). For each of the 30 traits, we calculated gene scores for the testable autosomal protein-coding genes (GWAS, ~19,150; eQTL-GWAS, ~12,550 (blood) and ~16,250 (tissue-wide); pQTL-GWAS, ~1,870; Exome, ~18,800). In the tissue-wide eQTL-GWAS method, the tissue with the most significant MR p value was selected. In Figure S1, we show the proportion of genes mapped to a particular tissue category. The contributions of glandular-endocrine, neural central nervous system (CNS), and whole-blood (eQTLGen) tissue categories were the highest (respective means of 15.3%, 12.8%, and 12.6% across the 30 traits; Tables S3 and S4). Although each trait had genes mapped to nearly all tissues, a few distinctive patterns could be observed: cardiac muscle tissues contributed the most to atrial fibrillation (16.4%); vascular tissues the most to coronary artery disease (16.5%), followed by diastolic (11.1%) and systolic (9.9%) blood pressure; and the neural CNS the most to schizophrenia (16.9%) and bipolar disease (16.6%).

(B) The effects of network diffusion using three different network types and different diffusion strengths (i.e., restart parameter *r*) were evaluated. Drug target genes may be prioritized only following signal propagation from neighboring disease genes.

(C) Diseases were linked to target genes through public drug databases: first, we used drug-indication information to connect the 30 traits to drugs and then leveraged drug target information to link the drugs to genes. Prioritized disease genes and corresponding diffusion scores (obtained via strategies described in A and B) were then tested for overlap with drug target genes through Fisher's exact test, resulting in odds ratios (ORs), and through area under the receiver operating characteristic curve (AUC) values.

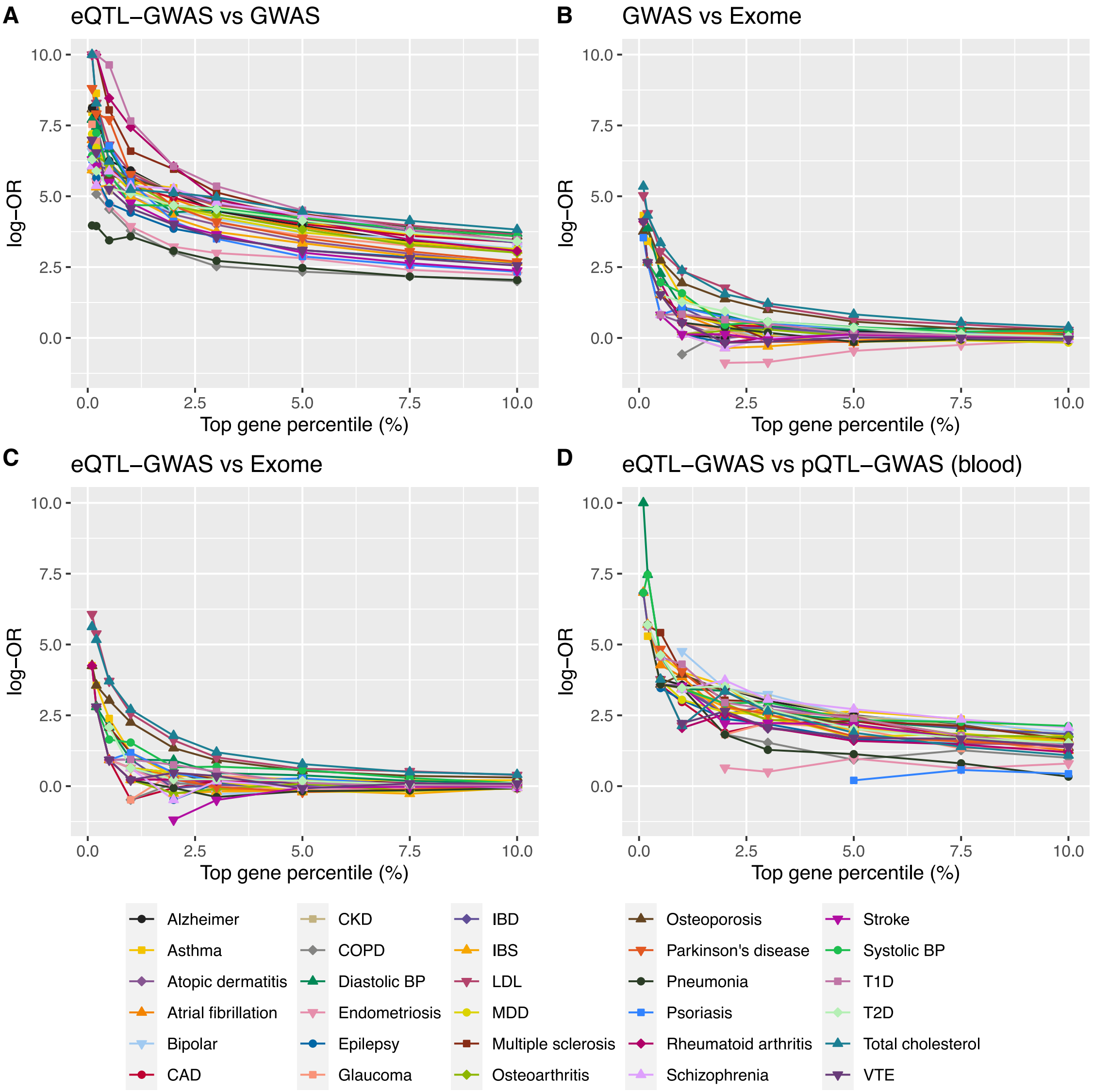


Figure 2. Evaluating the concordance of prioritized genes among gene scoring methods

(A–D) The top prioritized genes between pairs of methods were compared at different thresholds for each of the 30 traits/drug indications. The logarithm of odds ratios (log-OR) was calculated from Fisher's exact tests. Log-ORs are plotted only for percentiles at which common genes between pairs of methods were found. Comparisons were conducted on the same background genes and same data origins (i.e., on UK Biobank GWASs for comparisons with the Exome method). Tissue-wide eQTL-GWAS gene prioritizations were considered for the comparison with the GWAS and Exome methods and the blood-only eQTL-GWAS gene prioritization method for the comparison with the pQTL-GWAS method.

The concordance of prioritized genes among pairs of methods is summarized in Figure 2. For each trait, we calculated Fisher's exact tests between the top prioritized genes at thresholds ranging in the top 0.1%–10% (STAR Methods). The overlap was the highest between the GWAS and the eQTL-GWAS methods (Figure 2A). At 1%, the median OR was 212.2, which dropped to 51.0 and 22.1 at 5% and 10%, respectively. The overlap of prioritized genes was the lowest with the Exome method. The top 1% GWAS vs. Exome and eQTL-GWAS vs. Exome overlaps (based only on UKBB GWAS summary statistics), yielded median ORs of 1.7 and 1.9, respectively, which dropped to 1.0 at 10% for both methods (Figures 2B and 2C). Median ORs between eQTL-GWAS (whole blood) and pQTL-GWAS (blood plasma) were 8.5 and 4.6 at the top 5% and 10%, respectively (Figure 2D).

Enrichment of prioritized genes for drug targets

Next, we assessed the extent to which prioritized genes overlapped with drug target genes. For each trait, we conducted enrichment analyses for the GWAS, eQTL-GWAS, pQTL-GWAS, and Exome methods using our five definitions of drug target genes.

In Figure 3A, we show the resulting ORs for the DrugBank/DGIdb database combination. Across methods, genetic support for drug targets was the highest for low-density lipoprotein (LDL)

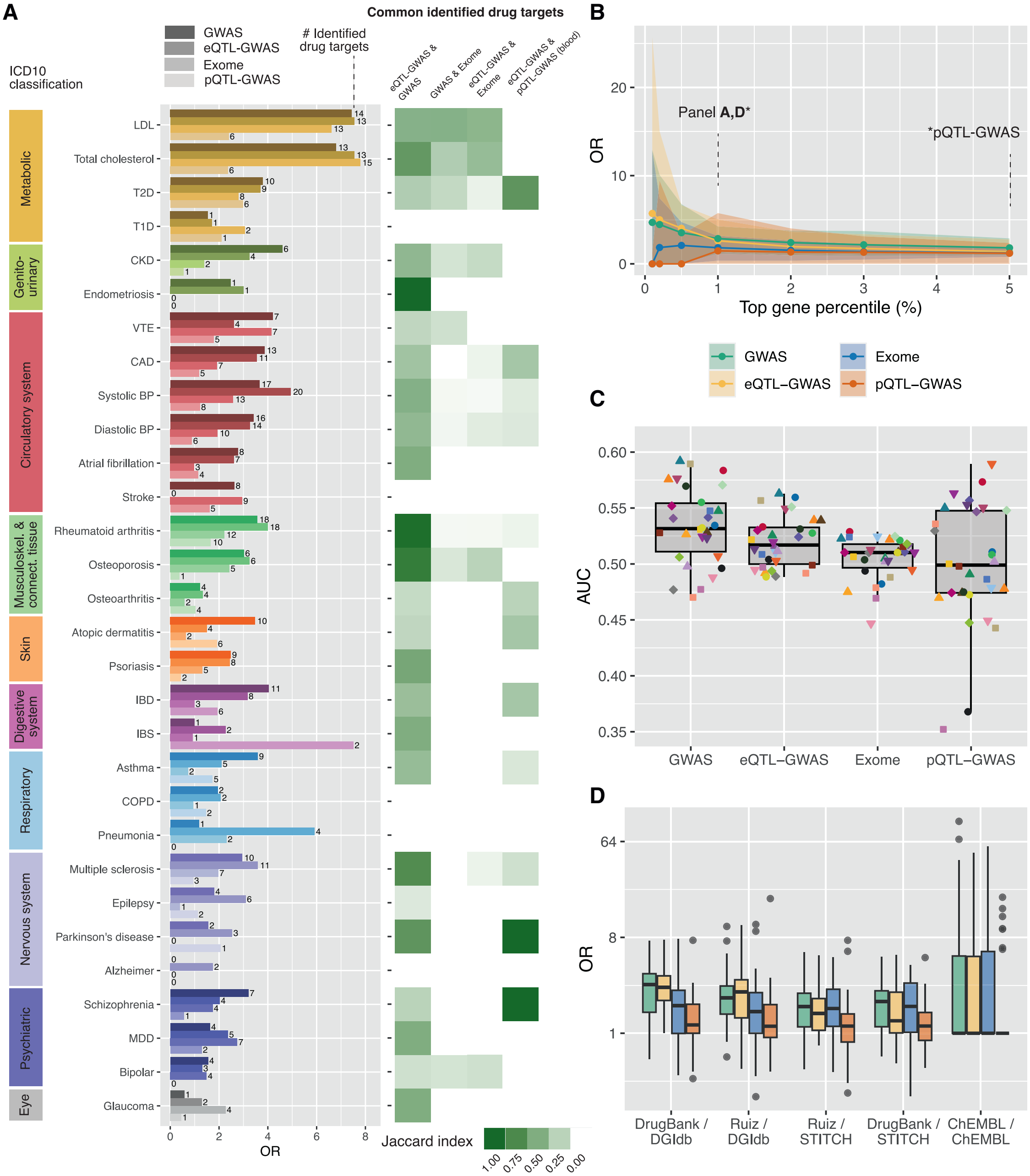


Figure 3. Enrichment of prioritized genes for drug targets

(A) Left: bar plot with odds ratios (ORs) calculated from Fisher's exact tests between drug target genes and the top 1% (5% for pQTL-GWAS) prioritized genes for the four tested methods and 30 traits, classified according to trait category. Drug target genes were defined by DrugBank/DGIdb, and only drug target genes that could be tested by the respective method were considered. The number on the right of each bar indicates the number of identified drug target genes. Right: overlap of identified drug target genes between pairs of methods quantified through the Jaccard index. The blood-only eQTL-GWAS gene prioritization method was used for the comparison with the pQTL-GWAS method. Plots using UKBB GWASs only are shown in Figure S3.

(B) ORs at different top prioritized gene percentiles for the four methods. The plotted dots correspond to the median OR across the 30 traits, and the shaded area bounds the 10% and 90% percentiles.

(C) Boxplots showing the area under the receiver operating characteristic curve (AUC) values. AUC values were calculated for each trait as indicated by the points (legend in Figure 2) and using the same background genes and drug target definitions as in (A).

(D) ORs calculated for the five drug target definitions and for all four methods (legend in B). The OR was set to 1 for traits with no identified drug target genes. In (C) and (D), the boxplots bound the 25th, 50th (median, center), and 75th quantiles. Whiskers range from minima (Q1 − (1.5 × IQR)) to maxima (Q3 + (1.5 × IQR)) with points outside representing potential outliers.

and total cholesterol (average ORs of 5.99 and 6.12, respectively). Lowest enrichment ratios were obtained for neuropsychiatric traits (average OR of 1.56) and glaucoma (average OR of 1.14). The average OR across traits was 2.48, 2.68, 1.65, and 1.26 for the GWAS, eQTL-GWAS, Exome, and pQTL-GWAS methods, respectively. We explored a range of top disease gene percentiles (0.1%–5%), and the corresponding ORs are shown in Figure 3B. Restricting disease genes to the top 0.1% for all methods increased the average ORs without changing the method ranking, with average ORs of 3.68, 4.02, 2.40, and 1.44 for the GWAS, eQTL-GWAS, Exome, and pQTL-GWAS methods, respectively. We further analyzed whether identified drug targets were the same across methods and found that prioritized drug target genes were similar between GWAS and eQTL-GWAS methods (average Jaccard index of 0.39), were less so between eQTL-GWAS and pQTL-GWAS methods (blood tissues; average Jaccard index of 0.15), and were very different from Exome identified targets (average Jaccard index of 0.06 between GWAS and Exome and between eQTL-GWAS and Exome methods). Average AUC values across traits were 53.4%, 51.9%, 50.5%, and 49.9% for the GWAS, eQTL-GWAS, Exome, and pQTL-GWAS methods (Figure 3C).

While the number of drugs reported per indication was similar across databases (average of 43.9, 41.8, and 40.4 for Ruiz et al., ChEMBL, and DrugBank, respectively), the average number of reported drug targets was much higher for Ruiz/STITCH (285), Ruiz/DGIdb (274.8), DrugBank/DGIdb (263.4), and DrugBank/STITCH (244.2) than for ChEMBL/ChEMBL (24.8; Table S6). We repeated drug target enrichment calculations for all drug database combinations (Figures 3D and S2). The average ORs for the GWAS/eQTL-GWAS methods were 2.48/2.68, 2.80/2.53, 2.18/2.12, 1.78/1.61, and 1.78/1.51 for DrugBank/DGIdb, ChEMBL/ChEMBL, Ruiz/DGIdb, Ruiz/STITCH, and DrugBank/STITCH, respectively. Overall, the variability in ORs across traits was the highest in the ChEMBL database (Figures 3D and S2), likely due to the low average number of reported drug targets, which leads to very high ORs when drug targets figured among the prioritized genes (e.g., for LDL and total cholesterol), but for many traits drug target genes were not among the prioritized genes (e.g., for type 1 diabetes, atopic dermatitis, and inflammatory bowel disease).

Since enrichment results can differ widely across traits and reference databases, we calculated overall enrichment and AUC values across traits and drug databases, including sensitivity analyses on UKBB data only, to match Exome sample sizes and common background genes (Table S8 and Figure S4; STAR Methods). The overall ORs were 2.17 (UKBB, 1.72), 2.04 (UKBB, 1.67), and 1.81 and 1.31 (UKBB, 1.30) for the GWAS, eQTL-GWAS, and Exome and pQTL-GWAS methods, respectively. There were no significant differences between these four methods in terms of enrichment OR (p_{diff} > 0.05, including in the sensitivity analyses). Overall AUCs were 54.3% (UKBB, 52.8%), 52.8% (UKBB, 51.4%), and 51.7% and 51.3% (UKBB, 50.6%) for the GWAS, eQTL-GWAS, and Exome and pQTL-GWAS methods, respectively. Judging by the AUC values, GWAS performed significantly better than eQTL-GWAS (p_{diff} = 3.1e−5) and also when considering only testable eQTL genes (p_{diff} = 2.9e−4). When excluding eQTLGen from the tissue-wide eQTL-GWAS, the performance of eQTL-GWAS dropped slightly (AUC of 52.2% compared with 52.8%; p_{diff} = 0.019). Significantly higher AUC values were obtained for GWAS compared with Exome on consortium data (p_{diff} = 2.2e−4), which was no longer the case on UKBB data (p_{diff} = 0.06). The difference between eQTL-GWAS and Exome was not significant on either dataset (p_{diff} = 0.12 and 0.77 on consortium and UKBB data, respectively). The number of testable genes was much lower for the pQTL-GWAS method (~1,870 genes). With this set of background genes, GWAS still scored a higher overall AUC (55.1%, p_{diff} = 2.1e−3). No difference was observed between the pQTL-GWAS and the tissue-wide or whole blood eQTL-GWAS methods (p_{diff} = 0.66 and 0.87, respectively).

Examples of drug target prioritization ranks

In Figure 4, we highlight drug targets and their gene prioritization ranks for a few examples (complete list in Table S9). Major anti-hypercholesterolemic drug targets *PCSK9* (evolocumab, alirocumab), *HMGCR* (statins), and *NPC1L1* (ezetimibe) were top ranked by all methods (except for no pQTLs being available for *HMGCR* and *NPC1L1*; Figure 4A). HCN4, the target of the antiarrhythmic drug dronedarone, was prioritized as a disease gene for atrial fibrillation only through the GWAS method. Although highly expressed in the atrial appendage and left ventricle of the heart, no eQTL was reported for this gene (Figure 4B). Several antiepileptic drugs target SCN1A, which was highly prioritized by the GWAS and eQTL-GWAS methods, with the strongest MR effect found in the nucleus accumbens (basal ganglia) of the brain (Figure 4C). The antiplatelet drug dipyrimadole used in the prevention and treatment of vascular diseases such as stroke and coronary artery disease is listed to target 23 genes of the *PDE* superfamily in ChEMBL. Of these, four (*PDE4D*, *PDE3A*, *PDE3B*, *PDE6B*) were ranked in the top 1% by the exome method for stroke (Figure 4D). None of the other methods prioritized any of these 23 genes. For coronary artery disease, another superfamily member (*PDE5A*) had a low ranking (<2%) by the GWAS and QTL-GWAS methods, supported by solid GWAS and e/pQTL colocalization (Figure 4E).

Heritability of drug target transcripts and proteins

Previous drug target enrichment analyses have shown that drug target genes are more likely to have lower residual variance intolerance scores (RVISs), i.e., are less tolerant to change.[1] Furthermore, limited overlap between eQTL and GWAS hits has been found, and it has been suggested that GWAS and eQTL genes are under different selective constraints.[41] Hence, under the assumption that drug target genes are more likely to be key (core) GWAS genes, we expected that drug target genes are less likely to harbor QTLs. To test this hypothesis, we assessed whether drug target transcript or protein levels are less amenable to regulation by common genomic variations, which could explain the lower than expected performance of QTL-GWAS approaches.

To this end, we compared the *cis* heritability of drug target genes vs. non-drug target genes that were measured in the respective studies (i.e., also those with no reported e/pQTLs; STAR Methods), where lower heritability would point toward a negative selection.[42] We conducted the analysis per trait and for each of the five drug target gene definitions; however, we could not observe a clear difference between *cis* heritabilities

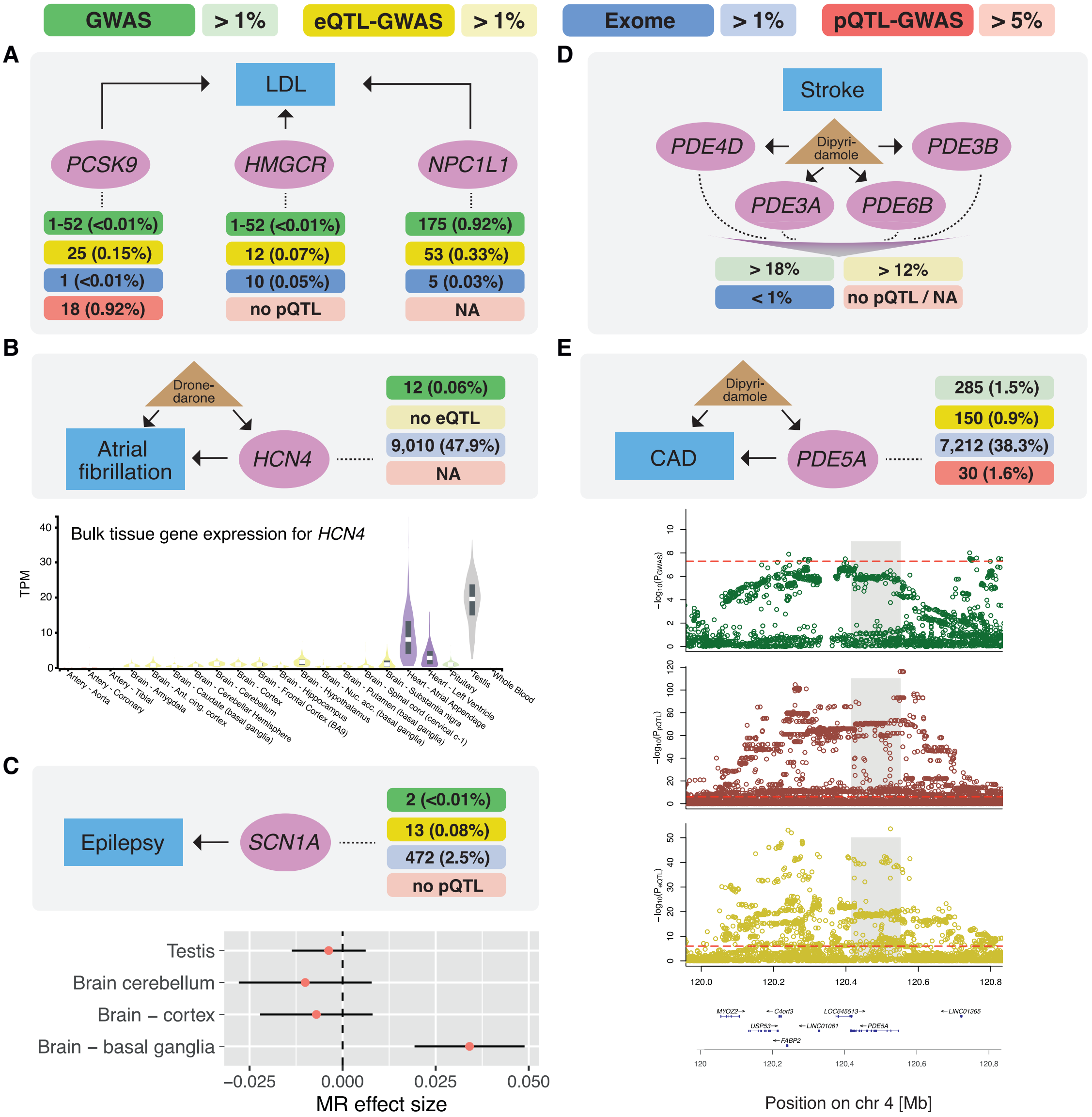


Figure 4. Examples illustrating drug target genes and their prioritization ranks

(A) Three drug target genes (*PCSK9* [evolocumab, alirocumab], *HMGCR* [statins], and *NPC1L1* [ezetimibe] shown in purple) for LDL cholesterol (blue box) and their prioritization ranks (top percentiles shown in parentheses) of each of the four methods (GWAS in green, eQTL-GWAS in yellow, Exome in blue, and pQTL-GWAS in red). Genes that were not testable by a given method are reported as NA (no e/pQTL means that the gene was measured, but had no QTL), and a range of ranks (i.e., 1–52) indicates tied p values.

(B) Top plot shows the prioritization ranks of *HCN4*, the target of the antiarrhythmic drug dronedarone. Bottom plot shows the gene expression profile of *HCN4* across GTEx tissues (TPM, transcripts per million) with "testis," "heart-atrial appendage," and "heart-left ventricle" dominating.

(C) Top plot shows the prioritization ranks of *SCN1A* (sodium voltage-gated channel alpha subunit 1), a drug target gene of several antiepileptic drugs. Bottom plot shows Mendelian randomization (MR) effects (red dots) with 95% CI (black bars) across tissues in which there was a significant eQTL.

(D) Antiplatelet drug dipyrimadole and gene prioritization ranks of its multiple drug targets (a non-exhaustive selection) of the phosphodiesterase (*PDE*) superfamily.

(E) Top plot shows the gene prioritization ranks of *PDE5A*, another reported target for dipyrimadole. Bottom plot shows the regional SNP associations ($-\log_{10}$(p)) with coronary artery disease (CAD; GWAS, green), PDE5A protein (pQTL, red), and *PDE5A* transcript (eQTL, yellow) (red dashed lines indicate the significance thresholds of the respective SNP association, and gray shading marks the position of *PDE5A*). Bottom row illustrates the position and strand direction of the genes in the locus.

of drug target and non-drug target genes (Figure S5). While this means that we cannot explain why the QTL-GWAS approach does not perform better, it may also imply that drug target genes are not necessarily typical GWAS genes or so-called core genes.

Network diffusion to prioritize drug target genes

Finally, we assessed whether network diffusion can identify drug target genes for which there is no direct genetic evidence. Gene scores from prioritization methods defined the initial distribution

$\boldsymbol{p}_0$ of the diffusion process. This process is regulated by a restart parameter r, whereby lower values result in a stronger diffusion (i.e., genes can be prioritized even when distant from initial disease genes; STAR Methods). The stationary distribution was calculated for six different restart parameters, ranging from no diffusion ($r = 1$) to complete diffusion ($r = 0$), and for each of the three networks: the STRING PPI network,[32] an RNA-seq coexpression network (CoXRNAseq),[33] and a coexpression and proteomics network (FAVA).[34] Since the set of testable proteins (∼1,870) is enriched for drug target genes (two-sided binomial test: p = 1.3e−47 for DrugBank/DGIdb; complete results in Table S15; STAR Methods), the AUC values were artificially inflated upon projecting the gene scores onto the network, and pQTL-GWAS results are hence not discussed.

Applying diffusion using the STRING network massively boosted the overlap between the diffused prioritized genes and the drug target genes (Figures 5A, 5B, S6, and S7). At no diffusion, overall AUC values across the 30 traits were 54.3%, 52.8%, and 51.7% for the GWAS, eQTL-GWAS, and Exome methods, respectively, which increased to 68.9%, 67.7%, and 66.9% at a diffusion parameter of $r = 0.6$, and further increased to 73.5%, 72.9%, and 72.3% at stronger diffusion ($r = 0.4$; Figures 5A and S6 and Table S11). A stronger enrichment of prioritized genes for drug targets upon diffusion was also observed when enrichment scores for the top 1% genes were calculated, with overall ORs of 4.63, 5.21, and 5.07 at $r = 0.4$ (Figures 5B and S7 and Table S11). On the other hand, improvements were modest when considering coexpression networks. At $r = 0.6$, overall AUC values increased to 54.9%, 54.7%, and 53.5% in the case of the CoXRNAseq network for the GWAS, eQTL-GWAS, and Exome methods, respectively. Although small, the difference was significant compared with no diffusion (p_{diff} of 5.11e−3, 4.12e−14, and 4.83e−5, respectively). In the same scenario, overall ORs at $r = 0.6$ were 2.28, 2.04, and 1.91, which were not significantly different ($p_{\text{diff}} > 0.05$) compared with no diffusion. Likewise, in the FAVA network, overall AUC values at $r = 0.6$ were 55.9%, 54.2%, and 53.6% (p_{diff} compared with no diffusion of 2.23e−5, 3.08e−3, and 7.3e−6), and ORs were 2.38, 2.02, and 1.77 ($p_{\text{diff}} > 0.05$), for GWAS, eQTL-GWAS, and Exome methods, respectively (Figures S6 and S7; Tables S11 and S12).

We further assessed which method's AUC values benefited the most from network diffusion. To allow fair comparison with the Exome methods, we used UKBB GWAS data for the GWAS and eQTL-GWAS methods. Across all diffusion parameters r, overall AUC values were significantly higher for GWAS compared with eQTL-GWAS in the STRING and FAVA network ($p_{\text{diff}} < 4.45\text{e}{-4}$), but not any different in the RNA-seq coexpression (CoXRNAseq) network ($p_{\text{diff}} > 0.05$). A nominally significant difference in favor of GWAS compared with Exome was observed only in the STRING network at r values of 0.4, 0.6, and 0.8 (p_{diff} of 0.0262, 7.36e−3, and 0.0146, respectively). No statistical differences were observed between the eQTL-GWAS and the Exome method except for a nominally significant difference in favor of eQTL-GWAS at $r = 0.2$ in the CoXRNAseq network ($p_{\text{diff}} = 0.0113$).

When investigating the network connectivity, we observed that drug target genes were significantly more likely to be hub genes, i.e., to have more connections in the network in comparison with other genes (Figures 5C and S8). This observation was particularly strong in the STRING network (mean log-degree = 13.0 vs. 12.3, $p_{\text{diff}} = 6.6\text{e}{-284}$ for DrugBank/DGIdb), but also present in the coexpression networks (Δ log-degree = 0.064, $p_{\text{diff}} = 0.011$ for CoXRNAseq; Δ log-degree = 0.3, $p_{\text{diff}} = 6.6\text{e}{-11}$ for FAVA). As a consequence, the network's node degree (a gene's number of connections to other genes adjusted by the edge weight) was found to be a good predictor of drug targets, and the best performance was found for the network degree in STRING (overall AUC = 77.6%, overall OR = 8.71). Given this bias, we generated random initial disease gene scores and determined to what extent genetically informed $\boldsymbol{p}_0$ distributions performed better compared with random $\boldsymbol{p}_0$ distributions. Although the GWAS, eQTL-GWAS, and Exome methods had significantly higher AUC values compared with random score distributions for any given r value in the STRING network ($p_{\text{diff}} < 1.62\text{e}{-7}$; Table S12), the performance of a mildly diffused ($r = 0.8$) random score (which is unaware of the target disease) performed significantly better than any disease gene prioritization method without diffusion (p_{diff} of 4.18e−6, 3.58e−10, and 2.10e−12 compared with GWAS, eQTL-GWAS, and Exome, respectively). In line with this observation, the network degree was still significantly better than gene prioritization methods at a stronger diffusion of $r = 0.2$ (p_{diff} of 8.98e−6, 9.87e−13, and 1.89e−11 compared with GWAS, eQTL-GWAS, and Exome, respectively).

Examples of prioritized genes through network diffusion

In the following, we describe several examples for which drug targets figured among the top 1% genes only after network diffusion (complete list in Table S13). Amyloid-beta precursor protein (APP) targeted by the monoclonal antibody aducanumab in the treatment of Alzheimer's disease (AD) was ranked 506 (top 2.7%) prior to and 152 (top 0.8%) after diffusion on the STRING network ($r = 0.6$; Figure 6A) based on the eQTL-GWAS method. Prioritization was largely influenced by its interacting neighbor apolipoprotein E (APOE), which was the top 5 ranked gene for AD by the eQTL-GWAS method and among the top 6 genes (tied p values) by the GWAS method. Although rare mutations in APP are a known cause of AD,[43] the Exome method did not highly prioritize this gene (>top 10%), likely because of low statistical power due to the younger and healthier nature of the UKBB cohort. Indeed, APP was among the top 1% for the GWAS method, leveraging the AD consortium data, but did not reach the top 10% when restricting the analysis to the UKBB. Tumor necrosis factor (TNF), a drug target in the treatment of inflammatory diseases such as psoriasis, was ranked 1,558th (top 8%; Exome-psoriasis) prior to and 182nd (top 0.98%; $r = 0.6$) post-propagation in the STRING network (Figure 6B). While initially the drug target F2 (coagulation factor II, thrombin) for venous thromboembolism (VTE) ranked only in the top 2%, it moved up to the top 1% regardless of the network used for diffusion at $r = 0.6$ (top 0.9%, 0.6%, and 0.7% for STRING, CoXRNAseq, and FAVA, respectively). In the STRING and CoXRNAseq networks, this boost could largely be attributed to the interacting fibrinogen genes (FGA, FGB, and FGG) that ranked in the top 0.06% (Figure 6C).

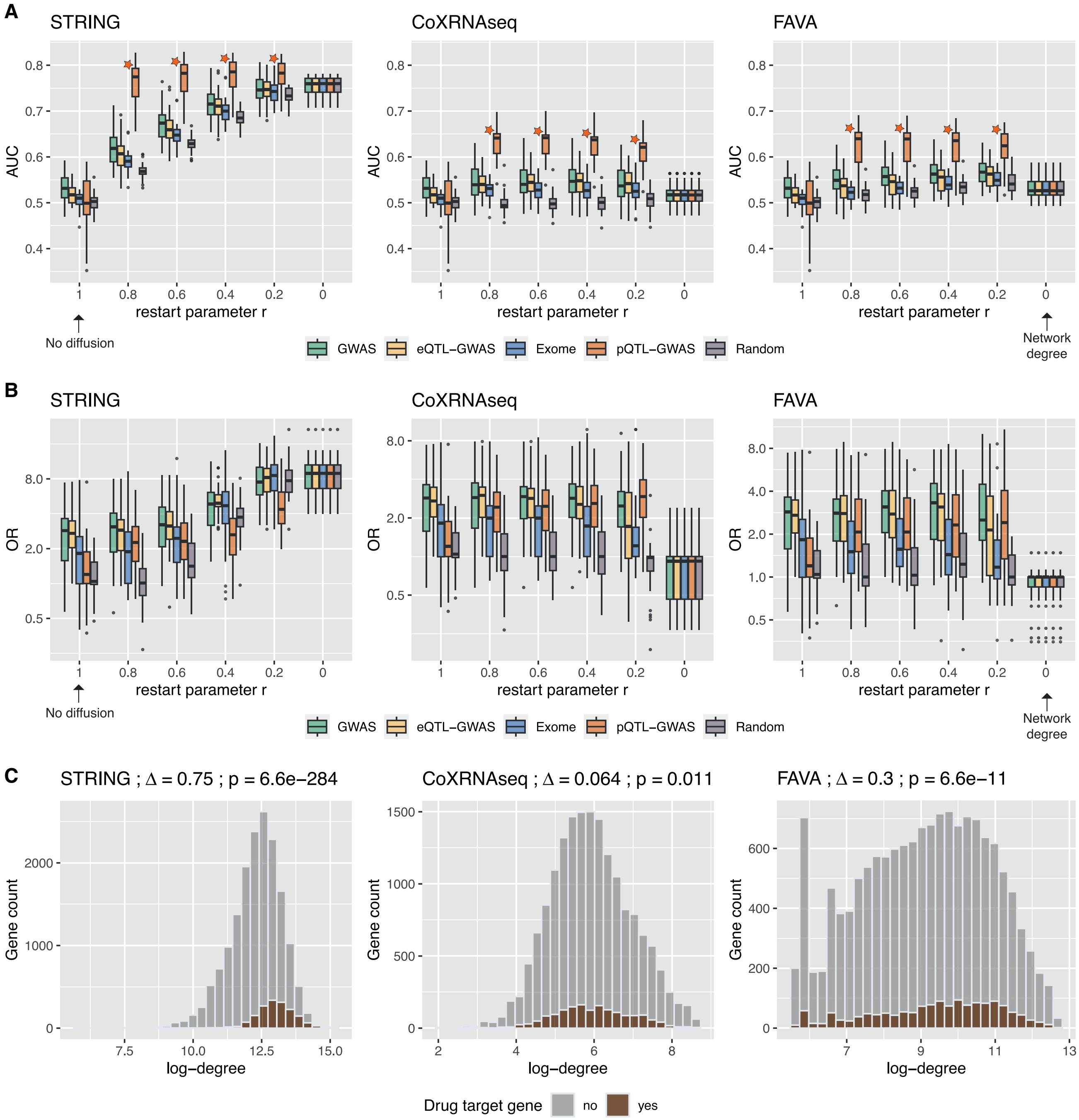


Figure 5. Effect of network diffusion to prioritize drug target genes

(A) Boxplots showing the area under the receiver operating characteristic curve (AUC) values for each network type (STRING, CoXRNAseq, and FAVA) and method at different restart parameter values *r*. AUC values were calculated for each of the 30 traits, and drug target genes were defined by DrugBank/DGIdb. At an *r* value of 1 (no network diffusion), the analysis corresponds to the results in Figure 3B, and at an *r* value of 0, the gene prioritization rank is based simply on the degree of the network nodes. At *r* < 1, the background genes are the genes reported in the respective network. The star next to the pQTL-GWAS method signals that the set of testable genes for this method is enriched for drug target genes, and therefore, higher AUC values were obtained when adding background genes with zero-valued initial scores.

(B) Odds ratios (ORs) between prioritized genes (top 1%) and drug target genes for each network type and method at different *r* values across the 30 traits (same drug target and background genes as in A). The OR was set to 1 for traits with no identified drug target genes.

(C) Histograms showing the degree distribution of drug target genes and non-drug target genes in each network. The difference in log-degree (Δ) and the p values from two-sided t tests are shown at the top. In (A) and (B), the boxplots bound the 25th, 50th (median, center), and 75th quantile. Whiskers range from minima (Q1 − (1.5 × IQR)) to maxima (Q3 + (1.5 × IQR)) with points above or below representing potential outliers.

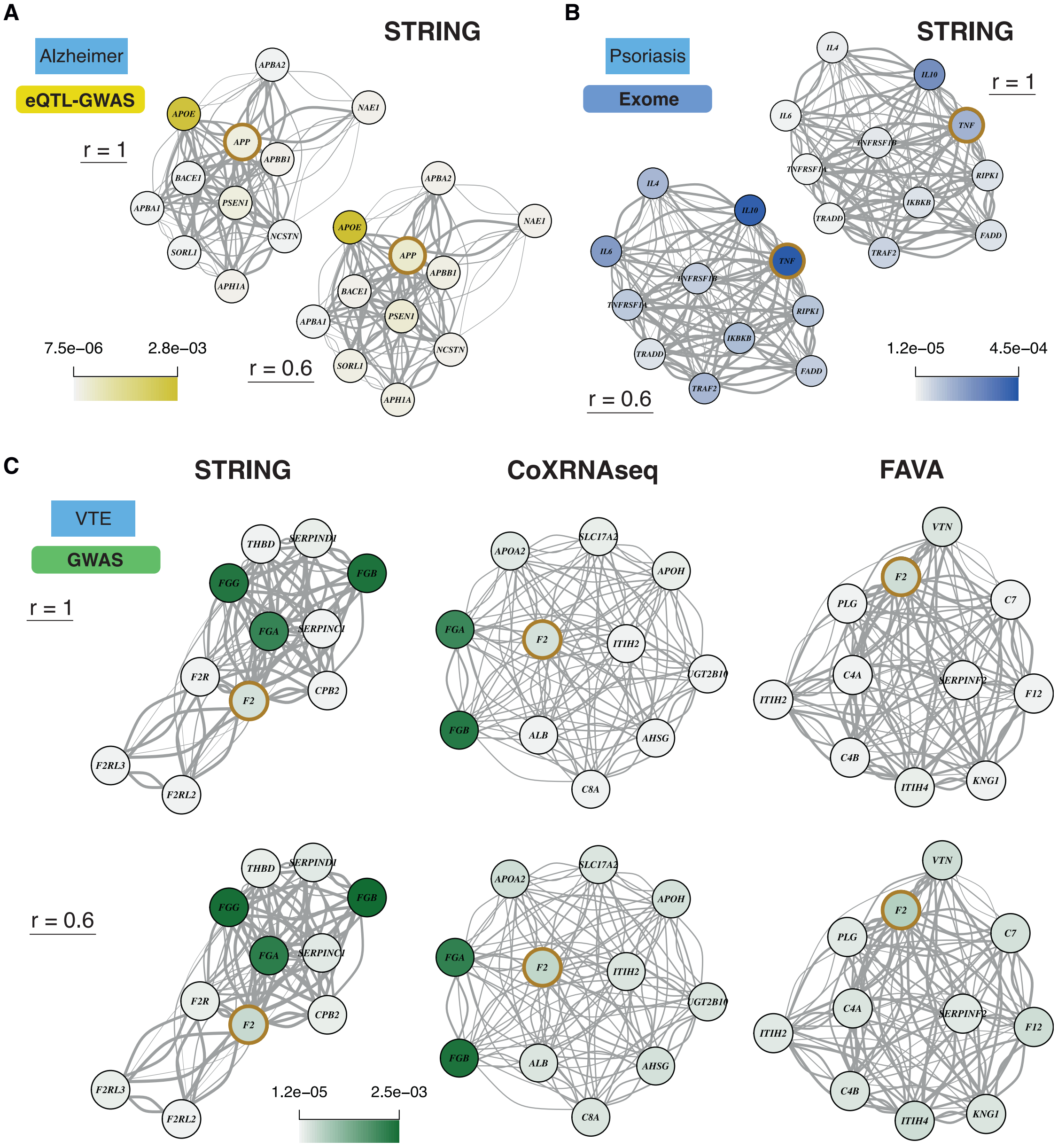


Figure 6. Examples illustrating prioritized drug target genes through network diffusion

(A) Top 10 network neighbors of drug target APP (brown circle) and their prioritization values (i.e., normalized node probabilities) by the eQTL-GWAS method for Alzheimer's disease are shown before ($r = 1$) and after diffusion ($r = 0.6$) on the STRING network.

(B) Same representation as in (A) showing Exome prioritization values for psoriasis and tumor necrosis factor (TNF) drug target.

(C) Top 10 network neighbors of drug target F2 (coagulation factor II, thrombin) in the STRING, CoXRNAseq, and FAVA networks. GWAS prioritization values for venous thromboembolism (VTE) are shown before ($r = 1$) and after diffusion ($r = 0.6$) on each network. In each network example (A–C), the drug target gene was among the top 1% prioritized genes only after diffusion at $r = 0.6$.

DISCUSSION

Summary of findings

We conducted a comprehensive benchmarking between different genetically informed approaches (GWAS, QTL-GWAS, and Exome) combined with network diffusion to prioritize drug target genes. The strength of our analysis lies in the side-by-side comparison of gene prioritization methods that individually have proven to be successful in identifying drug targets. In line with previous reports, we find a 1.3- to 2.2-fold enrichment for drug targets among (the top 1%) prioritized genes.[1,2] Recently, methods have emerged that combine multiple genetic predictors to derive an aggregate score, often using machine-learning techniques.[27,44,45] These scores have demonstrated high enrichment for drug targets but reveal little about underlying molecular mechanisms. Our aim was to disentangle the importance of the choice of the ground truth (i.e., drug target genes) and the input data (such as mQTLs, WES) in combination with different molecular networks to highlight added benefits while also exposing weaknesses compared with using GWAS data alone.

Comparison of gene prioritization methods

Adjusting for differences in background genes and data origins, GWAS yielded higher AUC than eQTL- and pQTL-GWAS, but no significant difference was found with Exome. Genes prioritized by the Exome method were different from those identified by the GWAS and QTL-GWAS methods, which was also reflected in the identified drug targets. While this could imply that rare and common variant genetic architectures are complementary, differences could also be due to power issues. Possibly, with increased sample size, the implicated genes will converge, but the extent to which they can be perturbed by regulatory vs. rare coding variants might remain different. Considering ORs, we lacked the statistical power to claim significant differences between methods, since the number of drug targets among the top 1% prioritized genes can be very low. Overall enrichment ORs for drug targets were 2.17, 2.04, 1.81, and 1.31 for the GWAS, eQTL-GWAS, Exome, and pQTL-GWAS methods, respectively. Although ORs for the pQTL-GWAS method may seem lower, it should be noted that testable proteins (i.e., proteins with pQTLs) accounted for ~10% of GWAS-testable genes. On the same background genes, ORs for the tissue-wide and blood-only eQTL-GWAS methods were 1.38 and 1.22, respectively. For the AUC metric, no significant difference between eQTL-GWAS and pQTL-GWAS was found. In the method comparisons, we considered multiple drug target gene definitions. The number of targets per drug drastically differed between ChEMBL and the DGIdb or STITCH database due to differences in their construct. Drug target genes in the ChEMBL database are manually curated and should not contain false positives, but it remains debatable whether one should consider only primary or also secondary target genes. For instance, ChEMBL lists only HMGCR as a drug target for statins, whereas the DGIdb database also includes APOA5, APOB, and APOE, among others. For this reason, we considered different databases and present enrichment results for both broad and narrow drug target definitions, as well as aggregates.

Benefits and pitfalls of network diffusion

Network diffusion was beneficial for prioritizing drug target genes with weaker genetic support. A remarkable increase in drug target identification was achieved when using the STRING PPI network. However, this improvement may be due to a circularity in the data generation process, whereby drug target genes are more researched and hence have more chance to be found to interact with other proteins, i.e., they tend to look more hub-like. Although genetically informed gene sets performed better than random ones, the genes prioritized by their node degree in the STRING network resulted in the highest AUC values overall. Thus, care has to be taken when relying on literature-derived gene-gene interactions, as aggressive diffusion will point to the same drug target genes, irrespective of the disease, due to the intrinsic bias stemming from under- and overstudied proteins. While the STRING network resource remains of immense value to identify interacting proteins, non-random missing of network edges leads to a biased network structure, which makes this resource less suitable as input for discovering new drug targets. The improvements made with coexpression networks, which do not suffer from publication/curation biases, were minor in comparison. Although significant with the AUC metric, ORs were not significantly increased with a diffusion of $r = 0.6$ compared with no diffusion for any of the methods.

Limitations of the study

Several limitations should be considered. First, we do not take into account the directionality of therapeutic and genetic effects, i.e., whether the drug is an agonist or antagonist. Although found to be less performant than GWAS, QTL-GWAS methods have the advantage of specifying directionality, as opposed to gene scores from the GWAS approach, which ignores SNP effect directions. Second, the mQTL datasets used cover only a small fraction of possible intermediate traits through which SNPs exert their disease-inducing effects.[46] Third, we focus only on common genetic variants when associating transcript and protein levels. With the advent of coupled rare variant-protein level data, either from populations enriched for rare variants or sequencing data,[14,47] more powerful QTL-GWAS methods are likely to emerge that combine mechanistic insights gained from QTL approaches while capturing rare variant associations previously missed. Fourth, drug target data are sparse, which limits the statistical power in benchmarking analyses. Given the required resources to test a drug target in clinical settings, focusing on top ranking genes is of most interest. This scenario is best described with a threshold that defines highly prioritized genes for enrichment analyses. However, ROC curves that quantify the performance at all prioritization thresholds (i.e., use all data at hand) were better powered to detect subtle differences between methods. Resulting AUC values are relatively low (51%–54%), which may be because ranks of genes with non-significant p values are likely unreliable, but these dominate most of the ROC curve. Related to this, even for low false positive rates, there is room for improvement of the gene prioritization methods. Combining prioritization methods could increase AUC values, as suggested by the distinct drug target sets identified by GWAS and Exome methods, as could the integration of additional functional genomic annotations.[27,44] Finally, our analysis compares methods using historical drug discovery data as the ground truth. These data are highly biased, with G-protein-coupled receptors being targets of a third of FDA-approved drugs.[48] Many other genes may be effective targets, but have never been tested in clinical trials. Thus, our results may not reflect how well the tested genetic approaches uncover true disease genes, but rather how well they identify targets that were historically prioritized in drug development processes. Since the emergence of robust GWASs, more and more clinical trials are motivated by genetically informed targets. Thus, drug target databases will tend to overlap better with GWAS-inspired genes, leading to artificially higher overlap.

Conclusion

To conclude, we systematically evaluated major gene prioritization approaches for their ability to identify approved drug target genes. Our analyses highlight the power of harnessing multiple data sources by capitalizing on QTLs for mechanistic insights, sequencing data for rare variant associations, GWASs when mQTL signals are missing, and network propagation to leverage gene-gene interactions.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - ○ Lead contact
  - ○ Materials availability
  - ○ Data and code availability
- METHOD DETAILS
  - ○ GWAS data
  - ○ GWAS gene scores
  - ○ Molecular QTL-GWAS gene scores
  - ○ Exome gene scores
  - ○ Drug target genes
  - ○ Transcript and protein level heritabilities
  - ○ Networks
  - ○ Network diffusion
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - ○ Concordance of gene scoring methods
  - ○ Drug target enrichment and AUC calculations
  - ○ Enrichment of proteins for drug targets

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.xgen.2023.100341.

ACKNOWLEDGMENTS

This work was supported by the Swiss National Science Foundation (310030_189147) to Z.K. This research was conducted using the UK Biobank Resource under application 16389. LD was calculated based on the UK10K data resource (EGAD00001000740, EGAD00001000741). Computations were performed on the JURA cluster at the University of Lausanne. We also would like to acknowledge the participants and investigators of the UK Biobank and FinnGen study. We thank Daniel Krefl for his help and support in implementing the PascalX software, Doug Speed for his help and support in calculating heritability estimates, and Liza Darrous for critically reading the draft.

AUTHOR CONTRIBUTIONS

M.C.S. and Z.K. conceived and designed the study. M.C.S. performed statistical analyses. P.D. provided guidance on statistical analyses. Z.K. supervised all statistical analyses. C.A. contributed to the collection and interpretation of pharmacological and biological data. All the authors contributed by providing advice on interpretation of results. M.C.S. and Z.K. drafted the manuscript. All authors read, approved, and provided feedback on the final manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: February 27, 2023
Revised: April 25, 2023
Accepted: May 16, 2023
Published: June 15, 2023

REFERENCES

1. Nelson, M.R., Tipney, H., Painter, J.L., Shen, J., Nicoletti, P., Shen, Y., Floratos, A., Sham, P.C., Li, M.J., Wang, J., et al. (2015). The support of human genetic evidence for approved drug indications. Nat. Genet. *47*, 856–860. https://doi.org/10.1038/ng.3314.
2. King, E.A., Davis, J.W., and Degner, J.F. (2019). Are drug targets with genetic support twice as likely to be approved? Revised estimates of the impact of genetic support for drug mechanisms on the probability of drug approval. PLoS Genet. *15*, e1008489. https://doi.org/10.1371/journal.pgen.1008489.
3. Edwards, S., Beesley, J., French, J., and Dunning, A. (2013). Beyond GWASs: illuminating the dark road from association to function. Am. J. Hum. Genet. *93*, 779–797. https://doi.org/10.1016/j.ajhg.2013.10.012.
4. Cao, Y., Shi, Y., Qiao, H., Yang, Y., Liu, J., Shi, Y., Lin, J., Zhu, G., and Jin, Y. (2014). GWAS and drug targets. BMC Genom. *113*, 1–9. https://doi.org/10.1186/1471-2164-15-S4-S5.
5. Tam, V., Patel, N., Turcotte, M., Bossé, Y., Paré, G., and Meyre, D. (2019). Benefits and limitations of genome-wide association studies. Nat. Rev. Genet. *20*, 467–484. https://doi.org/10.1038/s41576-019-0127-1.
6. Gazal, S., Weissbrod, O., Hormozdiari, F., Dey, K.K., Nasser, J., Jagadeesh, K.A., Weiner, D.J., Shi, H., Fulco, C.P., O'Connor, L.J., et al. (2022). Combining SNP-to-gene linking strategies to identify disease genes and assess disease omnigenicity. Nat. Genet. *54*, 827–836. https://doi.org/10.1038/s41588-022-01087-y.
7. Liu, J.Z., Mcrae, A.F., Nyholt, D.R., Medland, S.E., Wray, N.R., Brown, K.M., Hayward, N.K., Montgomery, G.W., Visscher, P.M., Martin, N.G., and Macgregor, S. (2010). A versatile gene-based test for genome-wide association studies. Am. J. Hum. Genet. *87*, 139–145.
8. de Leeuw, C.A., Mooij, J.M., Heskes, T., and Posthuma, D. (2015). MAGMA: generalized gene-set analysis of GWAS data. PLoS Comput. Biol. *11*, e1004219. https://doi.org/10.1371/journal.pcbi.1004219.
9. Lamparter, D., Marbach, D., Rueedi, R., Kutalik, Z., and Bergmann, S. (2016). Fast and rigorous computation of gene and pathway scores from SNP-based summary statistics. PLoS Comput. Biol. *12*, e1004714. https://doi.org/10.1371/journal.pcbi.1004714.
10. Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., et al. (2013). The genotype-tissue expression (GTEx) project. Nat. Genet. *45*, 580–585. https://doi.org/10.1038/ng.2653.
11. Sun, B.B., Maranville, J.C., Peters, J.E., Stacey, D., Staley, J.R., Blackshaw, J., Burgess, S., Jiang, T., Paige, E., Surendran, P., et al. (2018). Genomic atlas of the human plasma proteome. Nature *558*, 73–79. https://doi.org/10.1038/s41586-018-0175-2.
12. Folkersen, L., Gustafsson, S., Wang, Q., Hansen, D.H., Hedman, Å.K., Schork, A., Page, K., Zhernakova, D.V., Wu, Y., Peters, J., et al. (2020). Genomic and drug target evaluation of 90 cardiovascular proteins in 30,931 individuals. Nat. Metab. *2*, 1135–1148. https://doi.org/10.1038/s42255-020-00287-2.
13. Võsa, U., Claringbould, A., Westra, H.-J., Bonder, M.J., Deelen, P., Zeng, B., Kirsten, H., Saha, A., Kreuzhuber, R., Yazar, S., et al. (2021). Large-scale cis-and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. Nat. Genet. *53*, 1300–1310. https://doi.org/10.1038/s41588-021-00913-z.
14. Ferkingstad, E., Sulem, P., Atlason, B.A., Sveinbjornsson, G., Magnusson, M.I., Styrmisdottir, E.L., Gunnarsdottir, K., Helgason, A., Oddsson, A., Halldorsson, B.V., et al. (2021). Large-scale integration of the plasma proteome with genetics and disease. Nat. Genet. *53*, 1712–1721. https://doi.org/10.1038/s41588-021-00978-w.

15. Sun, B.B., Chiou, J., Traylor, M., Benner, C., Hsu, Y.-H., Richardson, T.G., Surendran, P., Mahajan, A., Robins, C., Vasquez-Grinnell, S.G., et al. (2022). Genetic regulation of the human plasma proteome in 54,306 UK biobank participants. Preprint at bioRxiv. https://doi.org/10.1101/2022.06.17.496443.

16. Giambartolomei, C., Vukcevic, D., Schadt, E.E., Franke, L., Hingorani, A.D., Wallace, C., and Plagnol, V. (2014). Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. PLoS Genet. *10*, e1004383. https://doi.org/10.1371/journal.pgen.1004383.

17. Hormozdiari, F., van de Bunt, M., Segrè, A., Li, X., Joo, J., Bilow, M., Sul, J., Sankararaman, S., Pasaniuc, B., and Eskin, E. (2016). Colocalization of GWAS and eQTL signals detects target genes. Am. J. Hum. Genet. *99*, 1245–1260. https://doi.org/10.1016/j.ajhg.2016.10.003.

18. Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B.W.J., Jansen, R., de Geus, E.J.C., Boomsma, D.I., Wright, F.A., et al. (2016). Integrative approaches for large-scale transcriptome-wide association studies. Nat. Genet. *48*, 245–252. https://doi.org/10.1038/ng.3506.

19. Barbeira, A.N., Dickinson, S.P., Bonazzola, R., Zheng, J., Wheeler, H.E., Torres, J.M., Torstenson, E.S., Shah, K.P., Garcia, T., Edwards, T.L., et al. (2018). Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. Nat. Commun. *9*, 1–20. https://doi.org/10.1038/s41467-018-03621-1.

20. Wu, Y., Zeng, J., Zhang, F., Zhu, Z., Qi, T., Zheng, Z., Lloyd-Jones, L.R., Marioni, R.E., Martin, N.G., Montgomery, G.W., et al. (2018). Integrative analysis of omics summary data reveals putative mechanisms underlying complex traits. Nat. Commun. *9*, 918–1014. https://doi.org/10.1038/s41467-018-03371-0.

21. Porcu, E., Rüeger, S., Lepik, K., Santoni, F.A., Reymond, A., and Kutalik, Z. (2019). Mendelian randomization integrating GWAS and eQTL data reveals genetic determinants of complex and clinical traits. Nat. Commun. *10*, 1–12. https://doi.org/10.1038/s41467-019-10936-0.

22. Sadler, M.C., Auwerx, C., Lepik, K., Porcu, E., and Kutalik, Z. (2022). Quantifying the role of transcript levels in mediating DNA methylation effects on complex traits and diseases. Nat. Commun. *13*, 7559. https://doi.org/10.1038/s41467-022-35196-3.

23. Cirulli, E.T., White, S., Read, R.W., Elhanan, G., Metcalf, W.J., Tanudjaja, F., Fath, D.M., Sandoval, E., Isaksson, M., Schlauch, K.A., et al. (2020). Genome-wide rare variant analysis for thousands of phenotypes in over 70,000 exomes from two cohorts. Nat. Commun. *11*, 542–610. https://doi.org/10.1038/s41467-020-14288-y.

24. Kosmicki, J.A., Horowitz, J.E., Banerjee, N., Lanche, R., Marcketta, A., Maxwell, E., Bai, X., Sun, D., Backman, J.D., Sharma, D., et al. (2021). Pan-ancestry exome-wide association analyses of COVID-19 outcomes in 586,157 individuals. Am. J. Hum. Genet. *108*, 1350–1355.

25. Backman, J.D., Li, A.H., Marcketta, A., Sun, D., Mbatchou, J., Kessler, M.D., Benner, C., Liu, D., Locke, A.E., Balasubramanian, S., et al. (2021). Exome sequencing and analysis of 454,787 UK biobank participants. Nature *599*, 628–634. https://doi.org/10.1038/s41586-021-04103-z.

26. Guney, T., Alisik, M., Alkan, A., Barábasi, A.-L., Akinci, S., Hacibekiroglu, T., Dilek, I., and Erel, O. (2016). Network-based in silico drug efficacy screening. Redox Rep. *21*, 1–5. https://doi.org/10.1038/ncomms10331.

27. Fang, H., ULTRA-DD Consortium; De Wolf, H., Knezevic, B., Burnham, K.L., Osgood, J., Lledó Lara, A., Lledó Lara, A., Kasela, S., De Cesco, S., et al. (2019). A genetics-led approach defines the drug target landscape of 30 immune-related traits. Nat. Genet. *51*, 1082–1091. https://doi.org/10.1038/s41588-019-0456-1.

28. MacNamara, A., Nakic, N., Amin Al Olama, A., Guo, C., Sieber, K.B., Hurle, M.R., and Gutteridge, A. (2020). Network and pathway expansion of genetic disease associations identifies successful drug targets. Sci. Rep. *10*, 20970. https://doi.org/10.1038/s41598-020-77847-9.

29. Han, Y., Wang, C., Klinger, K., Rajpal, D.K., and Zhu, C. (2021). An integrative network-based approach for drug target indication expansion. PLoS One *16*, e0253614. https://doi.org/10.1371/journal.pone.0253614.

30. Barrio-Hernandez, I., Schwartzentruber, J., Shrivastava, A., Del-Toro, N., Gonzalez, A., Zhang, Q., Mountjoy, E., Suveges, D., Ochoa, D., Ghoussaini, M., et al. (2023). Network expansion of genetic associations defines a pleiotropy map of human cell biology. Nat. Genet. *55*, 389–398. https://doi.org/10.1038/s41588-023-01327-9.

31. Luck, K., Kim, D.-K., Lambourne, L., Spirohn, K., Begg, B.E., Bian, W., Brignall, R., Cafarelli, T., Campos-Laborie, F.J., Charloteaux, B., et al. (2020). A reference map of the human binary protein interactome. Nature *580*, 402–408. https://doi.org/10.1038/s41586-020-2188-x.

32. Szklarczyk, D., Gable, A.L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., Doncheva, N.T., Morris, J.H., Bork, P., et al. (2019). STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res. *47*, D607–D613. https://doi.org/10.1093/nar/gky1131.

33. Deelen, P., van Dam, S., Herkert, J.C., Karjalainen, J.M., Brugge, H., Abbott, K.M., van Diemen, C.C., van der Zwaag, P.A., Gerkes, E.H., Zonneveld-Huijssoon, E., et al. (2019). Improving the diagnostic yield of exome-sequencing by predicting gene–phenotype associations using large-scale gene expression analysis. Nat. Commun. *10*, 2837. https://doi.org/10.1038/s41467-019-10649-4.

34. Koutrouli, M., Líndez, P.P., Bouwmeester, R., Martens, L., and Jensen, L.J. (2022). FAVA: high-quality functional association networks inferred from scRNA-seq and proteomics data. Preprint at bioRxiv. https://doi.org/10.1101/2022.07.06.499022.

35. Wishart, D.S., Feunang, Y.D., Guo, A.C., Lo, E.J., Marcu, A., Grant, J.R., Sajed, T., Johnson, D., Li, C., Sayeeda, Z., et al. (2018). Drugbank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res. *46*, D1074–D1082. https://doi.org/10.1093/nar/gkx1037.

36. Ruiz, C., Zitnik, M., and Leskovec, J. (2021). Identification of disease treatment mechanisms through the multiscale interactome. Nat. Commun. *12*, 1796. https://doi.org/10.1038/s41467-021-21770-8.

37. Gaulton, A., Hersey, A., Nowotka, M., Bento, A.P., Chambers, J., Mendez, D., Mutowo, P., Atkinson, F., Bellis, L.J., Cibrián-Uhalte, E., et al. (2017). The ChEMBL database in 2017. Nucleic Acids Res. *45*, D945–D954. https://doi.org/10.1093/nar/gkw1074.

38. Freshour, S.L., Kiwala, S., Cotto, K.C., Coffman, A.C., McMichael, J.F., Song, J.J., Griffith, M., Griffith, O., and Wagner, A.H. (2021). Integration of the drug–gene interaction database (DGIdb 4.0) with open crowdsource efforts. Nucleic Acids Res. *49*, D1144–D1151. https://doi.org/10.1093/nar/gkaa1084.

39. Szklarczyk, D., Santos, A., von Mering, C., Jensen, L.J., Bork, P., and Kuhn, M. (2016). Stitch 5: augmenting protein–chemical interaction networks with tissue and affinity data. Nucleic Acids Res. *44*, D380–D384. https://doi.org/10.1093/nar/gkv1277.

40. Aguet, F., Anand, S., Ardlie, K.G., Gabriel, S., Getz, G.A., Graubert, A., Hadley, K., Handsaker, R.E., Huang, K.H., Kashin, S., et al. (2020). The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science *369*, 1318–1330.

41. Mostafavi, H., Spence, J.P., Naqvi, S., and Pritchard, J.K. (2022). Limited overlap of eQTLs and GWAS hits due to systematic differences in discovery. Preprint at bioRxiv. https://doi.org/10.1101/2022.05.07.491045.

42. O'Connor, L.J., Schoech, A.P., Hormozdiari, F., Gazal, S., Patterson, N., and Price, A.L. (2019). Extreme polygenicity of complex traits is explained by negative selection. Am. J. Hum. Genet. *105*, 456–476. https://doi.org/10.1016/j.ajhg.2019.07.003.

43. O'brien, R.J., and Wong, P.C. (2011). Amyloid precursor protein processing and Alzheimer's disease. Annu. Rev. Neurosci. *34*, 185–204. https://doi.org/10.1146/annurev-neuro-061010-113613.

44. Mountjoy, E., Schmidt, E.M., Carmona, M., Schwartzentruber, J., Peat, G., Miranda, A., Fumis, L., Hayhurst, J., Buniello, A., Karim, M.A., et al. (2021). An open approach to systematically prioritize causal variants and genes at all published human GWAS trait-associated loci. Nat. Genet. *53*, 1527–1533. https://doi.org/10.1038/s41588-021-00945-5.

45. Forgetta, V., Jiang, L., Vulpescu, N.A., Hogan, M.S., Chen, S., Morris, J.A., Grinek, S., Benner, C., Jang, D.-K., Hoang, Q., et al. (2022). An effector index to predict target genes at GWAS loci. Hum. Genet. *141*, 1431–1447. https://doi.org/10.1007/s00439-022-02434-z.

46. Yao, D.W., O'Connor, L.J., Price, A.L., and Gusev, A. (2020). Quantifying genetic effects on disease mediated by assayed gene expression levels. Nat. Genet. *52*, 626–633. https://doi.org/10.1038/s41588-020-0625-2.

47. Dhindsa, R.S., Burren, O.S., Sun, B.B., Prins, B.P., Matelska, D., Wheeler, E., Mitchell, J., Oerton, E., Hristova, V.A., Smith, K.R., et al. (2022). Influences of rare protein-coding genetic variants on the human plasma proteome in 50,829 UK biobank participants. Preprint at bioRxiv. https://doi.org/10.1101/2022.10.09.511476.

48. Hauser, A.S., Chavali, S., Masuho, I., Jahn, L.J., Martemyanov, K.A., Gloriam, D.E., and Babu, M.M. (2018). Pharmacogenomics of GPCR drug targets. Cell *172*, 41–54.e19. https://doi.org/10.1016/j.cell.2017.11.033.

49. Krefl, D., and Bergmann, S. (2021). BergmannLab/PascalX: PascalX v0.0.1. https://doi.org/10.5281/zenodo.4429922.

50. Sadler, M. (2022). masadler/smrivw: v1.1, 10.5281/zenodo.7324709. https://doi.org/10.5281/zenodo.7324709.

51. Willer, C.J., Li, Y., and Abecasis, G.R. (2010). METAL: fast and efficient meta-analysis of genome-wide association scans. Bioinformatics *26*, 2190–2191. https://doi.org/10.1093/bioinformatics/btq340.

52. Csardi, G., and Nepusz, T. (2006). The igraph software package for complex network research. InterJournal, complex systems *1695*, 1–9.

53. Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., and Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinf. *12*, 77. https://doi.org/10.1186/1471-2105-12-77.

54. Durinck, S., Spellman, P.T., Birney, E., and Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nat. Protoc. *4*, 1184–1191. https://doi.org/10.1038/nprot.2009.97.

55. Speed, D., Holmes, J., and Balding, D.J. (2020). Evaluating and improving heritability models using summary statistics. Nat. Genet. *52*, 458–462. https://doi.org/10.1038/s41588-020-0600-y.

56. Sadler, M. (2023). masadler/DrugTargetMethodComparison: v1.0.1. https://doi.org/10.5281/zenodo.7857973.

57. Evangelou, E., Warren, H.R., Mosen-Ansorena, D., Mifsud, B., Pazoki, R., Gao, H., Ntritsos, G., Dimou, N., Cabrera, C.P., Karaman, I., et al. (2018). Genetic analysis of over 1 million people identifies 535 new loci associated with blood pressure traits. Nat. Genet. *50*, 1412–1425. https://doi.org/10.1038/s41588-018-0205-x.

58. Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., et al. (2018). The UK Biobank resource with deep phenotyping and genomic data. Nature *562*, 203–209. https://doi.org/10.1038/s41586-018-0579-z.

59. Wuttke, M., Li, Y., Li, M., Sieber, K.B., Feitosa, M.F., Gorski, M., Tin, A., Wang, L., Chu, A.Y., Hoppmann, A., et al. (2019). A catalog of genetic loci associated with kidney function from analyses of a million individuals. Nat. Genet. *51*, 957–972. https://doi.org/10.1038/s41588-019-0407-x.

60. Kurki, M.I., Karjalainen, J., Palta, P., Sipilä, T.P., Kristiansson, K., Donner, K., Reeve, M.P., Laivuori, H., Aavikko, M., Kaunisto, M.A., et al. (2022). FinnGen: unique genetic insights from combining isolated population and national health register data. Preprint at medRxiv. https://doi.org/10.1101/2022.03.03.22271360.

61. UK10K; Walter, K., Min, L.J., Huang, J., Crooks, L., Memari, Y., McCarthy, S., Perry, R.B.J., Xu, C.J., Futema, M., Lawson, D., et al. (2015). The UK10K project identifies rare variants in health and disease. Nature *526*, 82. https://doi.org/10.1038/nature14962.

62. Zhu, Z., Zhang, F., Hu, H., Bakshi, A., Robinson, M.R., Powell, J.E., Montgomery, G.W., Goddard, M.E., Wray, N.R., Visscher, P.M., and Yang, J. (2016). Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. Nat. Genet. *48*, 481–487. https://doi.org/10.1038/ng.3538.

63. Hemani, G., Tilling, K., and Davey Smith, G. (2017). Orienting the causal relationship between imprecisely measured traits using GWAS summary data. PLoS Genet. *13*, e1007081. https://doi.org/10.1371/journal.pgen.1007081.

64. Chambers, J., Davies, M., Gaulton, A., Hersey, A., Velankar, S., Petryszak, R., Hastings, J., Bellis, L., McGlinchey, S., and Overington, J.P. (2013). UniChem: a unified chemical structure cross-referencing and identifier tracking system. J. Cheminform. *5*, 3. https://doi.org/10.1186/1758-2946-5-3.

65. Nightingale, A., Antunes, R., Alpi, E., Bursteinas, B., Gonzales, L., Liu, W., Luo, J., Qi, G., Turner, E., and Martin, M. (2017). The Proteins API: accessing key integrated protein and genome information. Nucleic Acids Res. *45*, W539–W544. https://doi.org/10.1093/nar/gkx237.

66. Tong, H., Faloutsos, C., and Pan, J.Y. (2006). Fast random walk with restart and its applications. In Sixth International Conference on Data Mining (ICDM'06) (IEEE, IEEE), pp. 613–622. https://doi.org/10.1109/ICDM.2006.70.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
UK Biobank	UK Biobank	https://www.ukbiobank.ac.uk/
UK Biobank GWAS summary statistics	UK Biobank	http://www.nealelab.is/uk-biobank
UK Biobank GWAS summary statistics	UK Biobank	https://pan.ukbb.broadinstitute.org
FinnGen GWAS summary statistics	FinnGen	https://www.finngen.fi/en/access_results
Consortia GWAS summary statistics	Various sources	Table S1
Multiple sclerosis GWAS summary statistics	IMSGC	https://imsgc.net/?page_id=31
Whole blood expression QTLs	eQTLGen	https://www.eqtlgen.org/cis-eqtls.html
Tissue-wide expression QTLs	GTEx project	https://gtexportal.org/home/datasets
Plasma protein QTLs	deCODE study	https://www.decode.com/summarydata/
Whole exome gene burden tests	GWAS Catalog	accession IDs are in Table S2
UK10K data	UK10K	https://www.uk10k.org/data_access.html
DrugBank database	DrugBank	https://go.drugbank.com
ChEMBL database	ChEMBL	https://www.ebi.ac.uk/chembl/
DGIdb database	DGIdb	https://www.dgidb.org
STITCH database	STITCH	http://stitch.embl.de
Ruiz et al. Drug-disease links	[36]	https://doi.org/10.1038/s41467-021-21770-8
STRING network	STRING	https://string-db.org
Co-expression network	[33]	https://github.com/molgenis/systemsgenetics/wiki/Downstreamer
FAVA network	[34]	https://doi.org/10.5281/zenodo.6803472
Software and algorithms		
Main pipeline and analysis code	This paper	https://doi.org/10.5281/zenodo.7857973
PascalX	[49]	https://github.com/BergmannLab/PascalX,
SMR-IVW	[50]	https://github.com/masadler/smrivw,
METAL	[51]	https://github.com/statgen/METAL
R package igraph v1.3.5	[52]	https://igraph.org
R package pROC v1.15.3	[53]	https://doi.org/10.1186/1471-2105-12-77
biomaRt v2.50.3	[54]	https://doi.org/10.18129/B9.bioc.biomaRt
LDAK software v5.2	[55]	https://dougspeed.com

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Zoltán Kutalik (zoltan.kutalik@unil.ch).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- This paper analyzes existing, publicly available data. Accession numbers for the datasets are listed in the key resources table.
- Drug target genes and prioritized ranks are included in the supplemental material of this paper.
- All original code has been deposited at Github (https://github.com/masadler/DrugTargetMethodComparison) and archived at Zenodo (https://doi.org/10.5281/zenodo.7857973).[56]
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

METHOD DETAILS

GWAS data

We used the largest (to-date), publicly available GWAS summary statistics for each analyzed condition (Table S1). GWAS data came mostly from consortia specific to the respective disease, and were often a meta-analysis comprising the UKBB. Twenty-four out of the 30 conditions were case/control studies, the remaining 6 being continuous traits: diastolic and systolic blood pressure (DBP and SBP, respectively[57]), low-density lipoprotein and total cholesterol (LDL and TC, respectively[58]), estimated glomerular filtration rate (eGFR[59]) and heel bone mineral density ([58]) proxying chronic kidney disease (CKD) and osteoporosis, respectively. For four traits with low case count in the UK Biobank (< 20,000; chronic obstructive pulmonary disease (COPD), endometriosis, pneumonia and psoriasis) and no large-scale GWAS meta-analysis available, we performed a meta-analysis between the UK Biobank[58] and FinnGen[60] using METAL.[51]

GWAS gene scores

We used PascalX[9,49] to compute gene scores based on GWAS summary statistics. The software takes as input GWAS p values, gene annotations and LD structure. SNPs are assigned to genes and their squared z-scores are summed. This sum, under the null, was shown to follow a weighted chi-square distribution with weights being defined by the local LD structure from which gene p values can be derived.[9] We applied PascalX with default parameters (gene ± 50 kB) on protein-coding genes using the Ensembl identifiers and annotations (Ensembl GRCh37.p13 version) and the UK10K reference panel.[61] Across traits, ~19,150 protein-coding genes could be tested which were ranked by their PascalX p value.

Molecular QTL-GWAS gene scores

We integrated molecular quantitative trait loci (QTL) and GWAS summary statistics using Mendelian randomization (MR) implemented in the smr-ivw software.[22,62,50] The exposure (transcript or protein levels) and outcome disease were instrumented with independent genetic variants, also called instrumental variables (IVs; $r^2 < 0.01$) and used to calculate putative causal effect estimates of the exposure on the outcome (β_{MR}). IVs were required to be strongly associated to the exposure ($P_{QTL} < 1e\text{-}6$) and had to pass the Steiger filter ensuring no significantly stronger effect on the outcome than on the exposure.[63] We used expression QTLs (eQTLs) from the eQTLGen consortium[13] (whole blood; $n = 31{,}684$) and tissue-specific QTLs from the GTEx v8 release[40] (European ancestry; $n = 65$–573 for 48 tissue types; Table S3) to estimate causal transcript-trait effects. In the eQTLGen dataset there were ~ 12,550 protein-coding genes with at least 1 IV which increased to ~16,250 when integrating the GTEx dataset. MR results from both datasets (whole blood from eQTLGen and 49 tissues from GTEx) were aggregated by considering the MR causal effect with the lowest p value across tissues (Tables S3 and S4). Protein QTLs (pQTLs) from the deCODE study[14] (whole blood; $n = 35{,}559$) were used to estimate protein-trait causal effects with ~1,870 proteins having at least 1 IV. Prior to the analysis, e/pQTL and GWAS data were harmonized, palindromic SNPs were removed as well as SNPs with an allele frequency difference > 0.05 between datasets. All transcripts and proteins were mapped to Ensembl identifiers as provided by eQTLGen, GTEx and deCODE.

Exome gene scores

We used gene burden test results computed on WES data from the UK Biobank.[25] We extracted gene-trait associations based on putative loss of function (pLOF) and deleterious missense variants with MAF < 1% (M3.1 nomenclature in original publication) with phenotypes matching the investigated conditions as indicated in Table S2. Associations were provided for ~18,800 genes which were ranked by the association p value and retrieved by the provided Ensembl identifier.

Drug target genes

We extracted drug target genes from public resources by combining drug-indication and drug-target links from various databases. A given disease/indication was linked to a drug if the drug was indicated to be prescribed for the selected indication and subsequently, the target genes of these drugs were extracted. For drug-indication pairs we consulted DrugBank, Ruiz et al. and ChEMBL:

- DrugBank 5.0[35] (download: May 2022): DrugBank indications are manually curated from drug labels and underwent an expert review process. Drug indications have their own DrugBank condition numbers and drugs their DrugBank identifiers.
- Ruiz et al.[36]: A drug-disease dataset was created by querying multiple sources such as the Drug Repurposing Database, the Drug Repurposing Hub, and the Drug Indication Database and extracting information from drug labels, DrugBank and the American Association of Clinical Trials Database. Drug–disease pairs were filtered for FDA-approved treatment relationships. This dataset uses NLM UMLS CUIDS identifiers (National Library of Medicine - Unified Medical Language System Controlled Unique Identifier) for diseases and DrugBank identifiers for drugs.
- ChEMBL[37] (download: May 2022): ChEMBL drug indications are extracted from multiple sources including DailyMed package inserts, Anatomical Therapeutic Chemical (ATC) classification and ClinicalTrials.gov. Mapping of disease terms to Medical Subject Headings (MeSH) vocabulary and the Experimental Factor Ontology (EFO) is done through a combination of text-mining, automated mapping and manual curation/validation. Drugs are reported with ChEMBL identifiers.

The mapping of GWAS traits to the drug indication identifiers of the respective database is shown in Table S5.

Drug target genes were extracted from the DGIdb, STITCH and ChEMBL databases:

- Drug Gene Interaction database (DGIdb) 4.0[38] (release: January 2021): Aggregated drug-gene interactions from multiple sources including DrugBank, Drug Target Commons, the Therapeutic Target Database and Guide to Pharmacology. Genes were matched to Ensembl identifiers using the provided gene vocabulary file. Drugs were reported through DrugBank or ChEMBL identifiers, and mapping from ChEMBL to DrugBank identifiers was done with UniChem,[64] using PubChem IDs as intermediates..
- Search Tool for Interacting CHemicals (STITCH) 5.0[39]: Aggregated drug-protein interaction data from high-throughput experiments data, manually curated datasets and prediction methods. Only high confidence drug-protein relationships (confidence score $\geq$ 700) of the type “inhibition" and “activation" were considered. STITCH uses PubChem Chemical Identifiers (CID) for drugs and mapping to DrugBank IDs was done through the *chemical sources* file provided by STITCH. Protein Ensembl identifiers were mapped to gene Ensembl identifiers using biomaRt (GRCh37, v2.50.3)[54].
- ChEMBL[37] (download: May 2022): ChEMBL provides drug targets which have been manually curated from literature. Drug targets are identified by ChEMBL IDs with mapping to UniProt Accessions provided by ChEMBL. UniProt identifiers were then mapped to gene Ensembl identifiers through the UniProt REST API[65].

In this analysis we considered drug target genes resulting from the following combinations: DrugBank/DGIdb, DrugBank/STITCH, Ruiz/DGIdb, Ruiz/STITCH, and ChEMBL/ChEMBL. The number of drugs and drug target genes per indication is shown in Table S6.

Transcript and protein level heritabilities

Transcript and protein level *cis*-heritabilities were estimated from QTL effects using a restricted maximum likelihood method (reml) with the LDAK-thin heritability model. The LDAK heritability model assumes that the expected heritability contributed by each SNP depends on its MAF and LD. The analysis was conducted with the LDAK software (v5.2; reml method[55]) based on all SNPs in proximity of the transcript/protein ($\pm$ 500 kB) and the UK10K reference panel.[61] We set the –power to -0.25 and the –ignore-weights flag to YES to specify the LDAK-thin heritability model. The analysis was restricted to high-quality SNPs which were defined as being non-ambiguous, having a sample size > 5,000 and a MAF $\geq$ 0.01.

Protein heritabilities were based on the deCODE plasma protein dataset[14] and transcript heritabilities for whole blood on the eQTLGen dataset.[13] Of the 14,022 protein-coding transcripts in eQTLGen, reml converged for 12,218. Likewise, 3,716 of the 4,502 autosomal proteins in deCODE converged (estimated *cis*-heritabilities are in Table S10). Genes not converging were omitted in *cis*-heritability downstream analyses.

To calculate the difference in heritabilities between drug target and non-drug target genes, we considered all transcripts and proteins measured in the respective study which were classified accordingly. Per trait, the difference in heritability was then calculated through a two-sided t-test. Heritability tests were only performed for traits with at least three drug targets within the respective set of measured transcripts/proteins.

Networks

To calculate network diffusion scores, we used the following three networks:

- Search Tool for Retrieval of Interacting Genes/Proteins (STRING) v11[32]: The protein-protein (PPI) interaction network results from predictions based on genomic context information, coexpression, text-mining, experimental biochemical/genetic data and curated databases (curated pathways and protein-complex knowledge). Protein Ensembl identifiers were mapped to gene Ensembl identifiers using biomaRt (GRCh37, v2.50.3).[54] We use interaction confidence scores as edge weights.
- CoXRNAseq[33]: This network was constructed by first performing a principal component analysis on the gene coexpression correlation matrix of 31,499 RNA-seq samples. Reliable principal components were retained from which the final network was constructed via Pearson correlations. We filtered pairwise interactions to only retain those with z-scores above 4. Genes were reported with Ensembl identifiers and z-scores were used as edge weights.
- Functional Associations using Variational Autoencoders (FAVA)[34]: This network is based on single cell RNA-seq read-count data from the Human Protein Atlas and proteomics data from the PRoteomics IDEntifications (PRIDE) database. First, the high-dimensional expression data was reduced into a latent space using variational autoencoders. From this latent space, the network was derived via pairwise Pearson correlations. Each reported interaction has a score which we use as edge weight (final network reports interactions with scores above 0.15). Protein Ensembl identifiers were mapped to gene Ensembl identifiers using biomaRt.

A summary of network properties is given in Table S14. In all analyses, we use weighted networks, and we refer to weighted node degrees (i.e., sum of edge weights linking the node of interest to adjacent nodes) as node degrees.

Network diffusion

We calculated network diffusion scores based on Markov random walks. Starting from an initial node distribution $\boldsymbol{p}_0$, a stationary distribution is calculated based on network connectivity. This diffusion process depends on a restart parameter r which determines how often the random walker returns to the initial values. Analytically, the stationary distribution ($\boldsymbol{p}_\infty$) is given by:

$$\boldsymbol{p}_\infty = (\boldsymbol{I} - (1 - r)\cdot \boldsymbol{W})^{-1}\cdot \boldsymbol{p}_0 \qquad \text{(Equation 1)}$$

where $\boldsymbol{W}$ is the column-normalized weighted adjacency matrix and $\boldsymbol{I}$ the identity matrix of the same dimension as $\boldsymbol{W}$[66]. The initial node distribution $\boldsymbol{p}_0$ was determined by the squared z-scores derived from the gene p values (normalized to sum up to 1). Genes that could not be tested by a given method had their initial value set to 0. Additionally, we tested the performance of network diffusion on random initial distributions $\boldsymbol{p}_0$. For each trait, a random distribution was generated which all were different, but consistent across analyses. Resulting network diffusion scores $\boldsymbol{p}_\infty$ were ranked for AUC calculations, and the top 1% scored genes were used in the enrichment analyses.

Network manipulations, visualization and degree calculations were performed with the R igraph package v1.3.5.[52]

QUANTIFICATION AND STATISTICAL ANALYSIS

Concordance of gene scoring methods

We tested whether prioritized genes were similar or dissimilar between pairs of methods. First, only genes (based on Ensembl identifiers) that were common between the two tested methods were selected into the gene background. Then, prioritized genes were defined at different top percentile cut-offs (0.1%, 0.2%, 0.5%, 1%, 2%, 3%, 5%, 7.5%, 10%). The enrichment of prioritized genes between methods was quantified by a Fisher's exact test using common genes as background genes. When calculating median ORs, ORs of traits for which no prioritized genes overlapped at a given percentile were set to 1. Results of this analysis are presented in the result section "concordance of prioritized genes among gene scoring methods".

Drug target enrichment and AUC calculations

Enrichment for drug target genes was calculated through two-sided Fisher's exact tests. A contingency table was constructed based on testable genes (i.e., background genes), with genes categorized into prioritized (top 1% or 5% for the pQTL-GWAS) and drug target genes. In rare instances (i.e., pQTL-GWAS background genes and ChEMBL/ChEMBL drug targets) where diagonal values were 0, these were changed to 1. If no prioritized gene coincided with a drug target gene, the resulting OR was set to 1 (for visualization purposes this was not done in barplots where each trait was shown individually). AUC values and standard errors were calculated using the R package pROC v1.15.3.[53]

Log-OR and AUC values (both are denoted b_i herein) were aggregated across traits and drug databases ($m = 30\cdot 5 = 150$ observations per method) as follows:

$$\overline{b} = \frac{1}{m}\sum_{i}^{m} b_i \qquad \text{(Equation 2)}$$

with corresponding variance:

$$\mathrm{var}(\overline{b}) = 1' \cdot \boldsymbol{S} \cdot \boldsymbol{R} \cdot \boldsymbol{S}\cdot 1 \big/ m^2 \qquad \text{(Equation 3)}$$

where $\boldsymbol{S}$ is a diagonal matrix of size mxm containing standard errors of b_i and $\boldsymbol{R}$ is the correlation matrix between drug databases and traits. This matrix was derived from the Kronecker product of the drug database correlation matrix and phenotypic trait correlation matrix (Tables S6 and S7). The drug database correlation matrix was derived on the gene level (i.e., 1 if the gene was a drug target for any of the 30 traits, 0 if not) and the phenotypic trait correlation on individual-level data from the UKBB (codes in Table S1B). $\overline{b}$ was referred to as the overall AUC/ log-OR (overall OR after an exponential transformation).

To calculate the statistical difference of $\overline{b}_1$ and $\overline{b}_2$ for method 1 and 2, respectively, we derived the variance of the difference as follows:

$$\mathrm{var}(\overline{b}_1 - \overline{b}_2) = \mathrm{var}(\overline{b}_1) + \mathrm{var}(\overline{b}_2) - 2\cdot\mathrm{cov}(\overline{b}_1,\overline{b}_2) \qquad \text{(Equation 4)}$$

with $\mathrm{cov}(\overline{b}_1,\overline{b}_2) \approx r \times (1' \cdot \boldsymbol{S}_1 \cdot \boldsymbol{R} \cdot \boldsymbol{S}_2 \cdot 1 / m^2)$, where r is the empirical correlation between $\boldsymbol{b}_1$ and $\boldsymbol{b}_2$. From the resulting z-score, a two-sided p value was calculated and significance was defined at a p value below 0.05. Results of these analyses are presented in the result sections "enrichment of prioritized genes for drug targets" and "network diffusion to prioritize drug target genes".

Enrichment of proteins for drug targets

We conducted binomial tests to verify whether the set of testable (i.e., at least 1 pQTL) and measured proteins (∼1,870 and ∼ 4,450, respectively) were enriched for drug target genes. We performed the analysis on each of the five drug target definitions and proceeded as follows: 1) we extracted the number of testable/measured proteins that are drug targets (“number of successes"), 2) considering all protein-coding autosomal genes (19,430), we extracted those that are drug targets (“number of trials"), 3) we determined the proportion of testable/measured proteins among all protein-coding genes (“expected probability of success"). From these numbers, we conducted two-sided exact binomial tests (Table S15).

Cell Genomics, Volume *3*

Supplemental information

Multi-layered genetic approaches to identify approved drug targets

Marie C. Sadler, Chiara Auwerx, Patrick Deelen, and Zoltán Kutalik

Supplemental information

Multi-layered genetic approaches to identify approved drug targets

Marie C. Sadler, Chiara Auwerx, Patrick Deelen, Zoltán Kutalik

Supplemental Figures

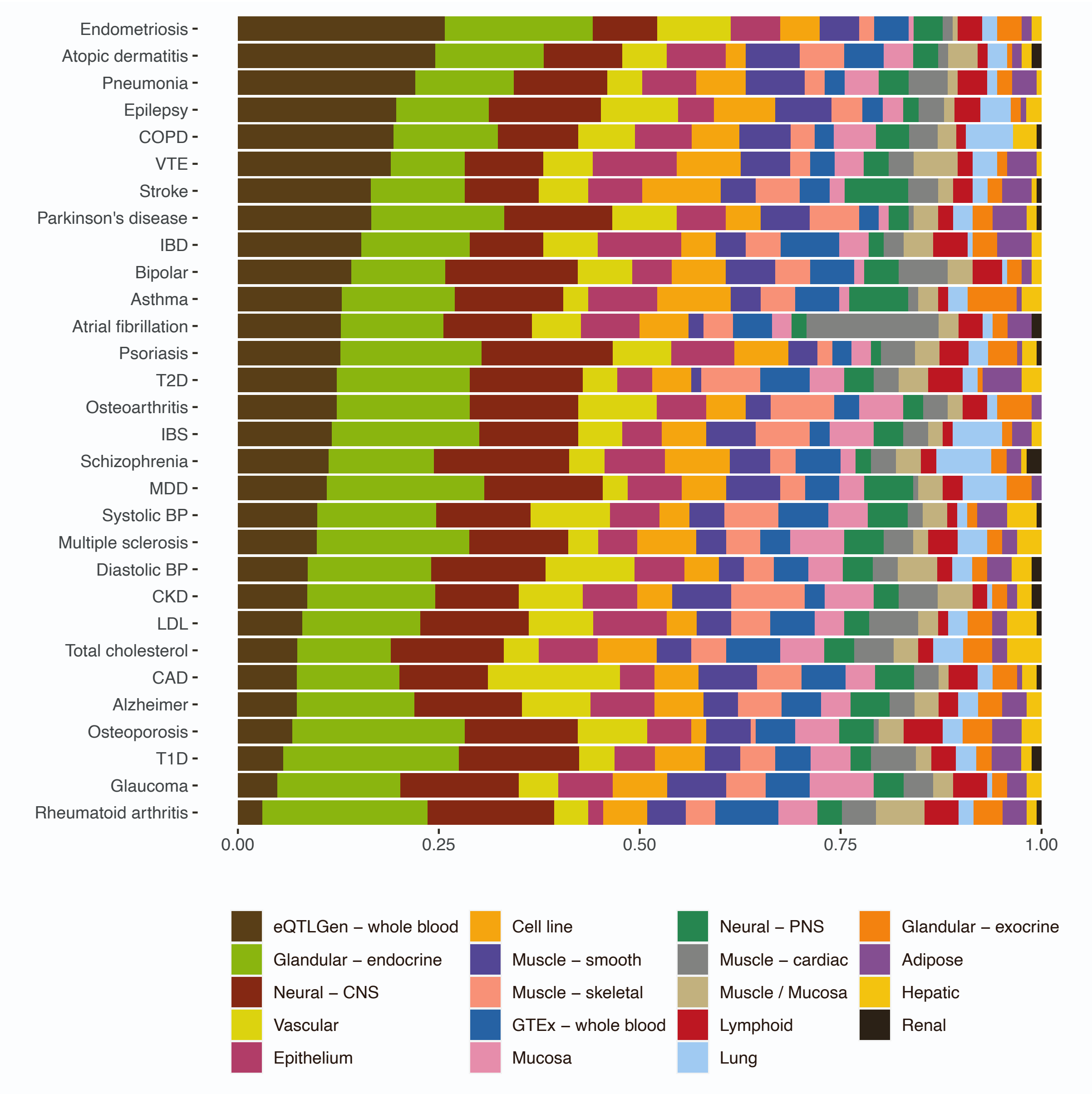


Figure S1. Gene-tissue mapping proportions.
Proportion of genes mapped to a particular tissue category in the tissue-wide expression quantitative trait locus (eQTL)-genome-wide association analysis (GWAS) analysis. For each gene, the tissue with the lowest Mendelian randomization (MR) p-value was selected. Tissue category belonging are shown in Table S4 and numerical proportion values in Table S5. This figure is related to the eQTL-GWAS method in Figures 2,3 and 5.

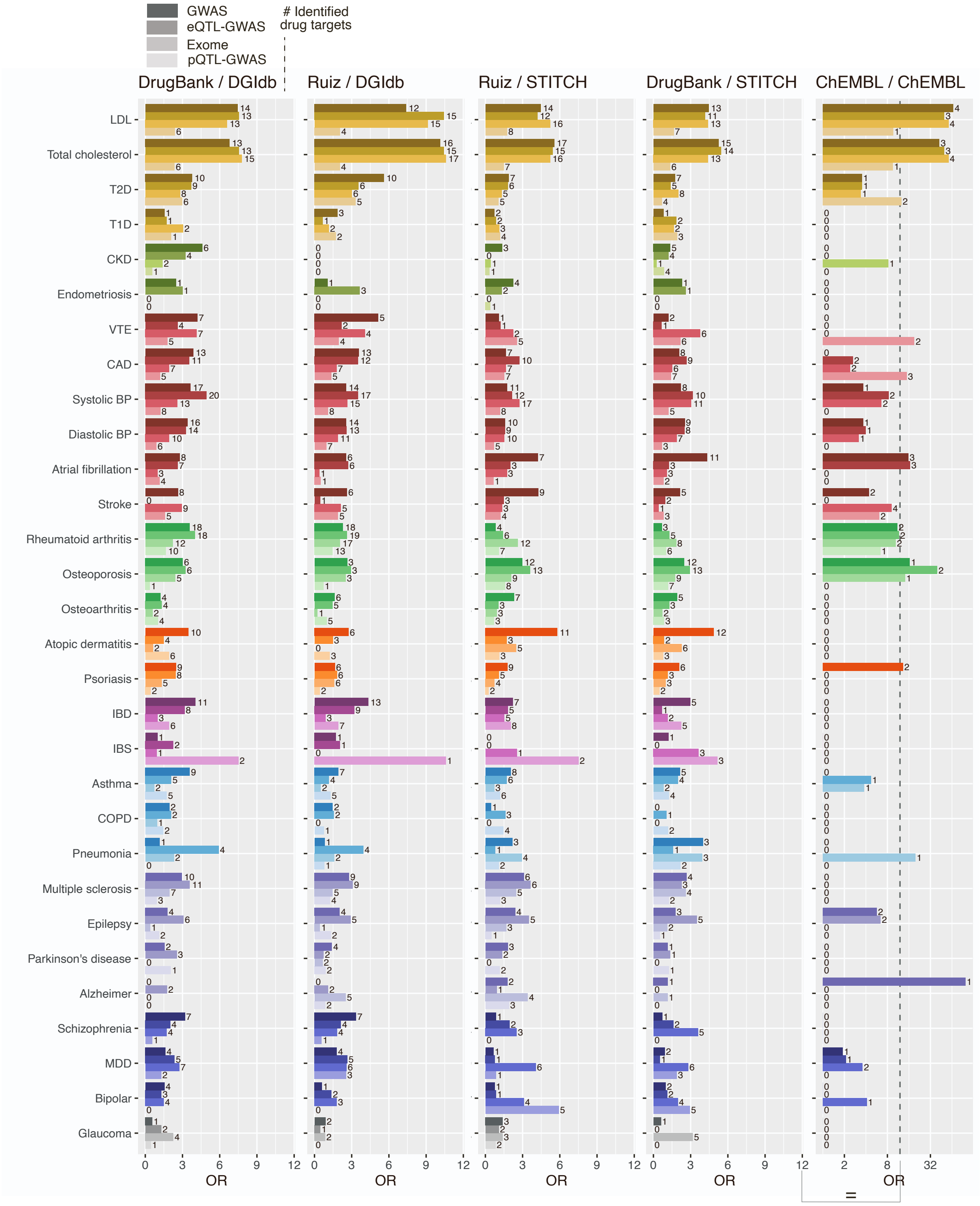


Figure S2. **Enrichment for drug target genes across drug databases.**

Barplots with odds ratios (ORs) calculated from Fisher's exact tests between drug target genes and prioritized genes for the four tested methods and thirty traits. Prioritized genes were defined as the top 1% percentile of the GWAS, eQTL-GWAS and Exome methods, and 5% of the pQTL-GWAS method. Drug target genes were defined by the drug database combinations (drug-indication and drug-target links) shown in the title of each barplot. Only drug target genes that could be tested by the respective method were considered. The number on the right of each bar indicates the number of identified drug target genes. In the barplot corresponding to the ChEMBL/ChEMBL database, the x-axis is log-transformed and therefore ORs of 0 (i.e., no identified drug target) were set to 1. This figure is related to Figure 3A which shows enrichment for the DrugBank /DGIdb combination.

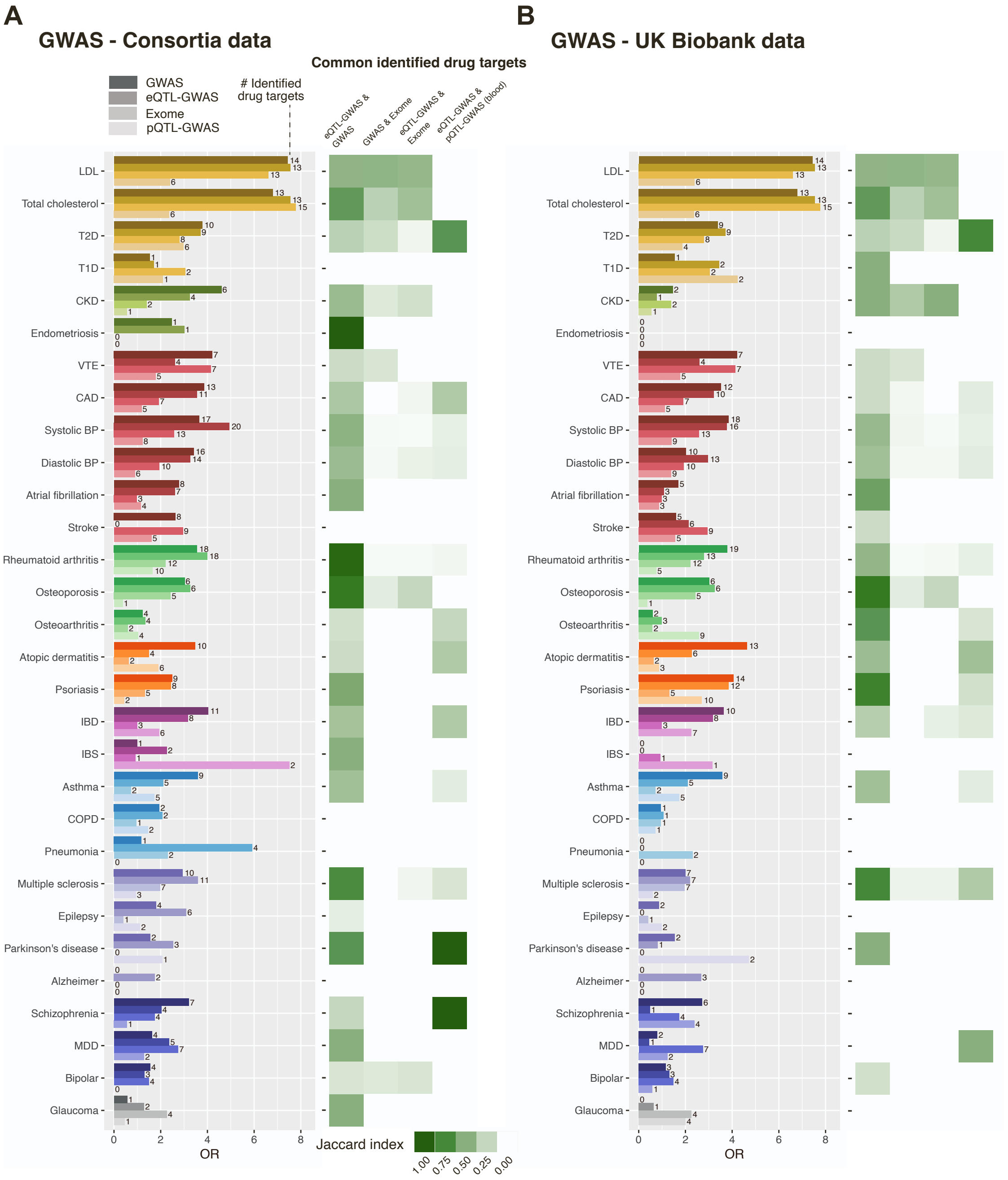


Figure S3. **Comparing consortia and UK Biobank GWAS data in drug target enrichment analyses.**

(A) Enrichment analysis using consortia GWAS summary statistics in the GWAS, eQTL-GWAS and pQTL-GWAS methods.

(B) Enrichment analysis using UKBB GWAS summary statistics in the GWAS, eQTL-GWAS and pQTL-GWAS methods. The Exome analysis is only performed on UK Biobank data. Left: Barplot with odds ratios (ORs) calculated from Fisher's exact tests between drug target genes and prioritized genes for the four tested methods and thirty traits. Prioritized genes were defined as the top 1% percentile of the GWAS, eQTL-GWAS and Exome methods, and 5% of the pQTL-GWAS method. Drug target genes were defined from the DrugBank and DGIdb databases, and only drug target genes that could be tested by the respective method were considered. The number on the right of each bar indicates the number of identified drug target genes. Right: Overlap of identified drug target genes between pairs of methods quantified through the Jaccard index. The blood-only eQTL-GWAS gene prioritization method was used for the comparison with the pQTL-GWAS method. This figure is related to Figure 3A which shows enrichment for drug targets using consortia GWAS.

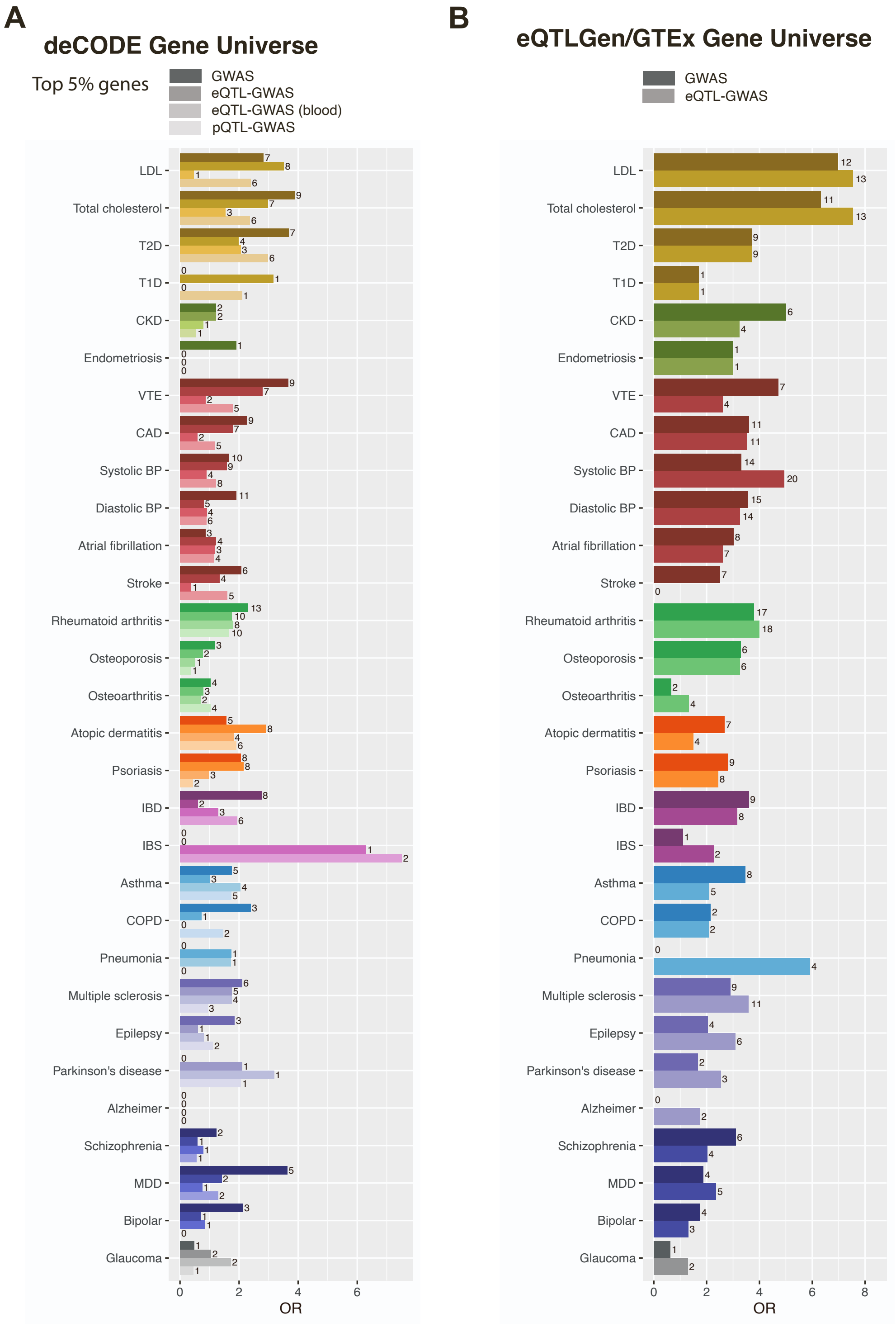


Figure S4. **Enrichment for drug target genes using the same background genes.**
(A) Enrichment analysis was performed by subsetting the gene universe of the GWAS and eQTL-GWAS (tissue-wide and whole blood only) methods to the genes available in the deCODE study (i.e., proteins used in the pQTL-GWAS analysis).
(B) Enrichment analysis was performed by subsetting the gene universe of the GWAS method to the genes available in the tissue-wide eQTL-GWAS analysis. Both plots show barplots with odds ratios (ORs) calculated from Fisher's exact tests between drug target genes and prioritized genes for the four tested methods and thirty traits. Drug target genes were defined from the DrugBank and DGIdb databases, and only drug target genes that could be tested by the respective method were considered.

The number on the right of each bar indicates the number of identified drug target genes. This figure is related to Figure 3A in which background genes were different among methods.

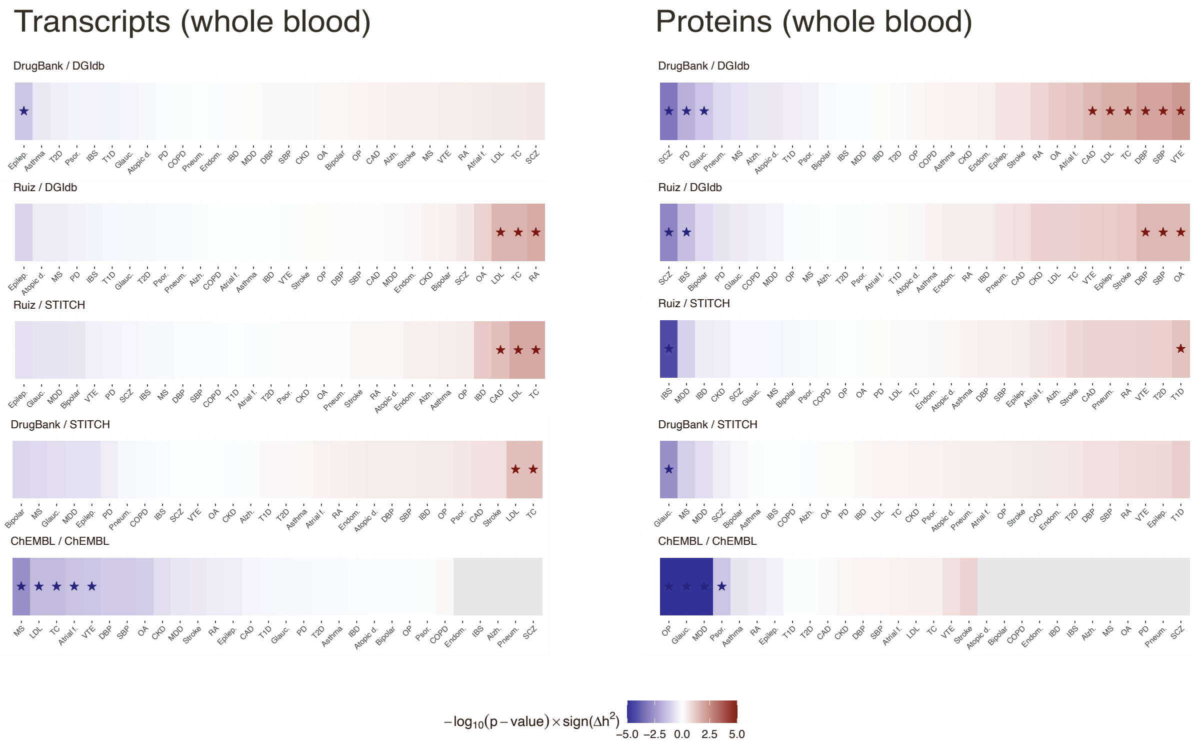


Figure S5. **Heritability of drug target genes.**

Difference in *cis*-heritability of drug target compared to non-drug target measured transcript and protein levels. For each trait, the difference in heritability was calculated through a two-sided t-test. When the difference was negative (i.e., drug target genes were less heritable), the -$\log_{10}$(p-value) is plotted in blue, otherwise in red. Traits for which the difference was nominally significant (p-value < 0.05), are indicated with a star. If less than three drug target genes could be tested for a trait, a grey box is plotted. This figure is related to the result section "Heritability of drug target transcripts and proteins".

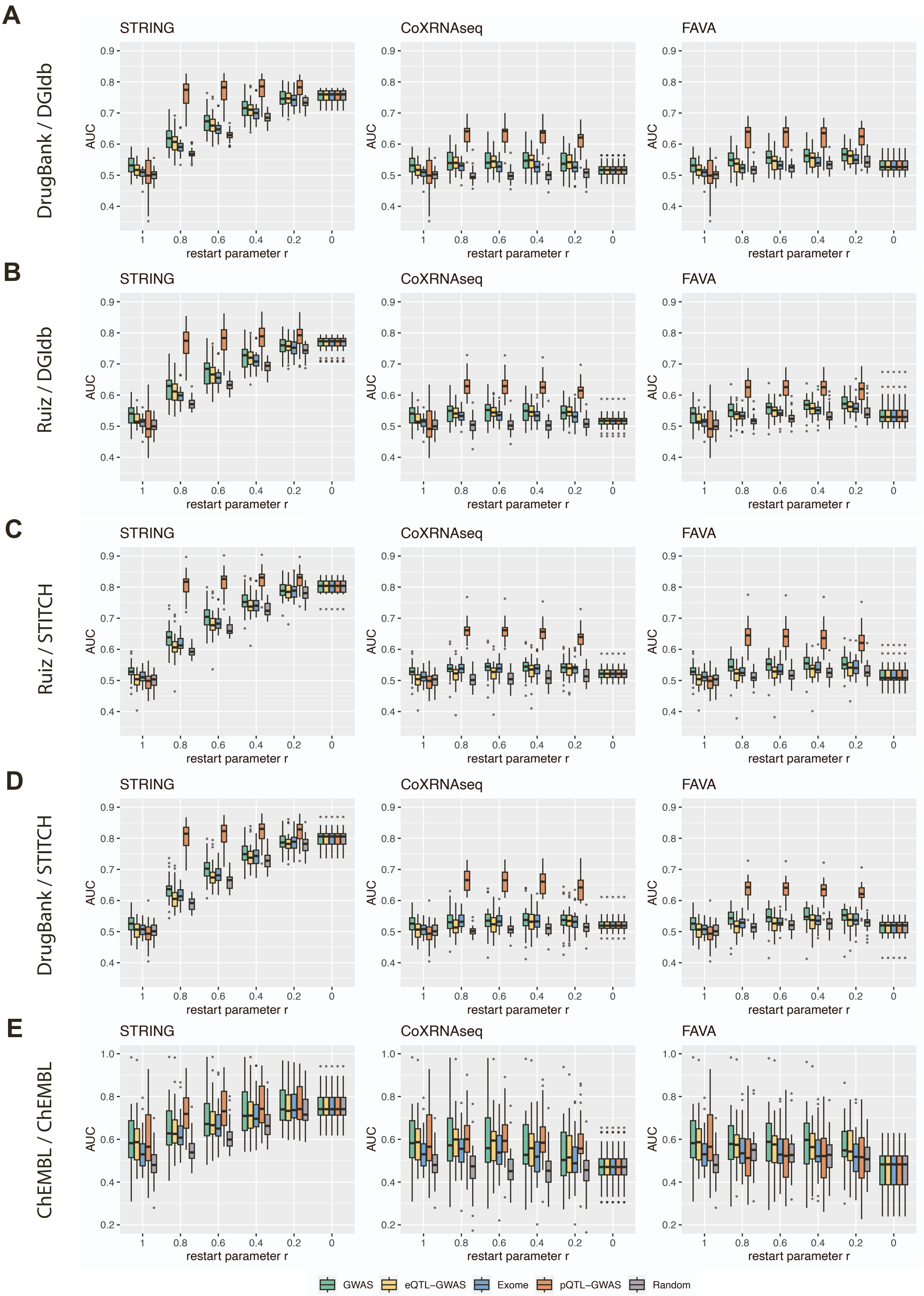


Figure S6. Effect of network diffusion to prioritize drug target genes across drug databases (AUC values).

Boxplots showing the area under the receiver operating characteristic curve (AUC) values for each network type (STRING, CoXRNAseq and FAVA) and method at different restart parameter values *r*.

AUC values were calculated for each of the thirty traits, and drug target genes were defined by the respective drug database combination (drug-indication and drug-target links, (A)-(E). The boxplots bound the 25th, 50th (median, centre), and the 75th quantile. Whiskers range from minima (Q1 – 1.5 • IQR) to maxima (Q3 + 1.5 • IQR) with points above or below representing potential outliers. This figure is related to Figure 5A which shows AUC values for the DrugBank /DGIdb combination.

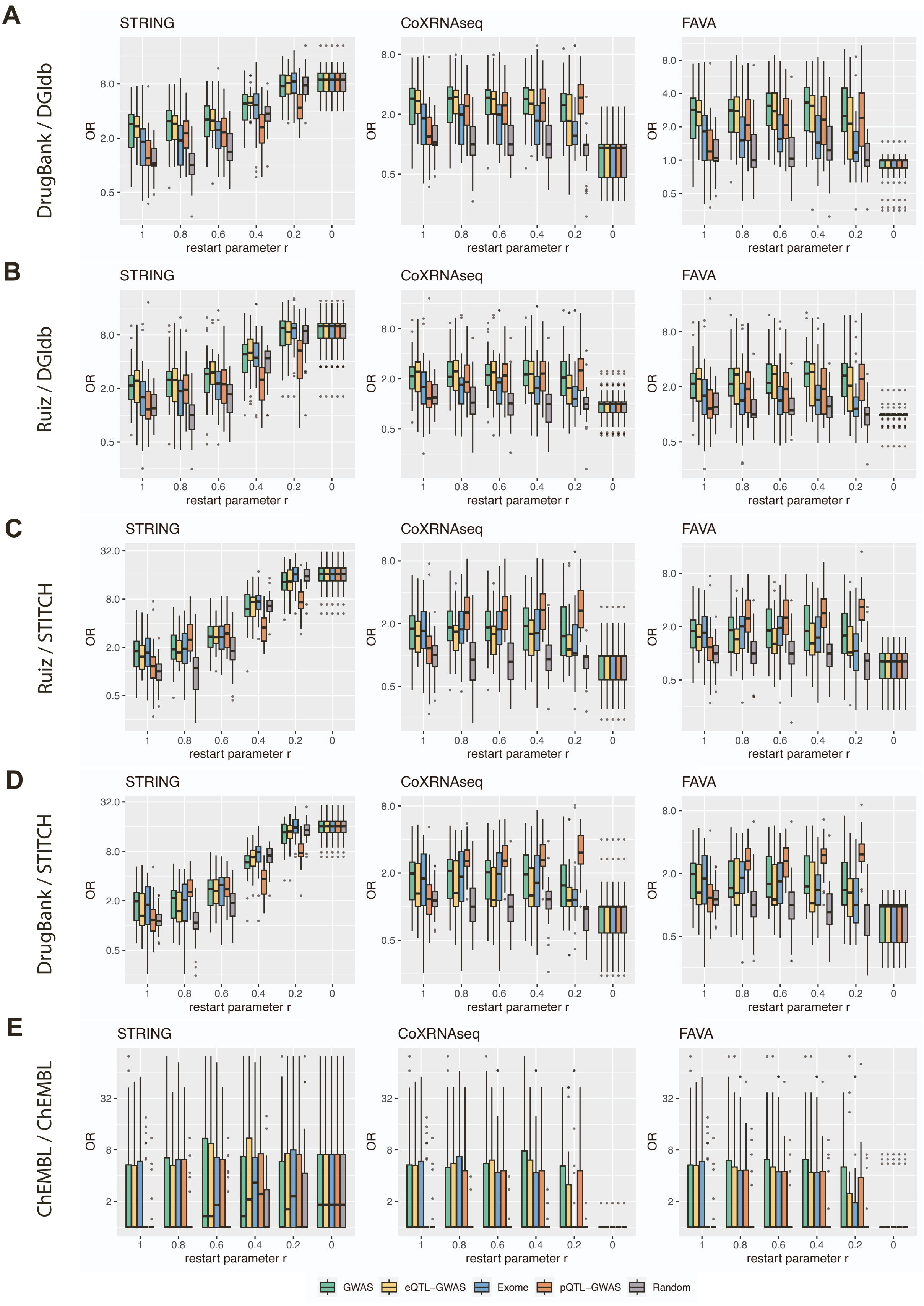


Figure S7. **Effect of network diffusion to prioritize drug target genes across drug databases (ORs).** Odds ratios (ORs) between prioritized genes (top 1%) and drug target genes for each network type (STRING, CoXRNAseq and FAVA) and method at different restart parameter values *r*. Drug target genes were defined by the respective drug database combination (drug-indication and drug-target links, (A)-(E)). The OR was set to 1 for traits with no identified drug target genes. The boxplots bound the 25th, 50th (median, centre), and the 75th quantile. Whiskers range from minima (Q1 – 1.5 • IQR) to maxima (Q3 + 1.5 • IQR) with points above or below representing potential outliers. This figure is related to Figure 5B which shows ORs for the DrugBank /DGIdb combination.

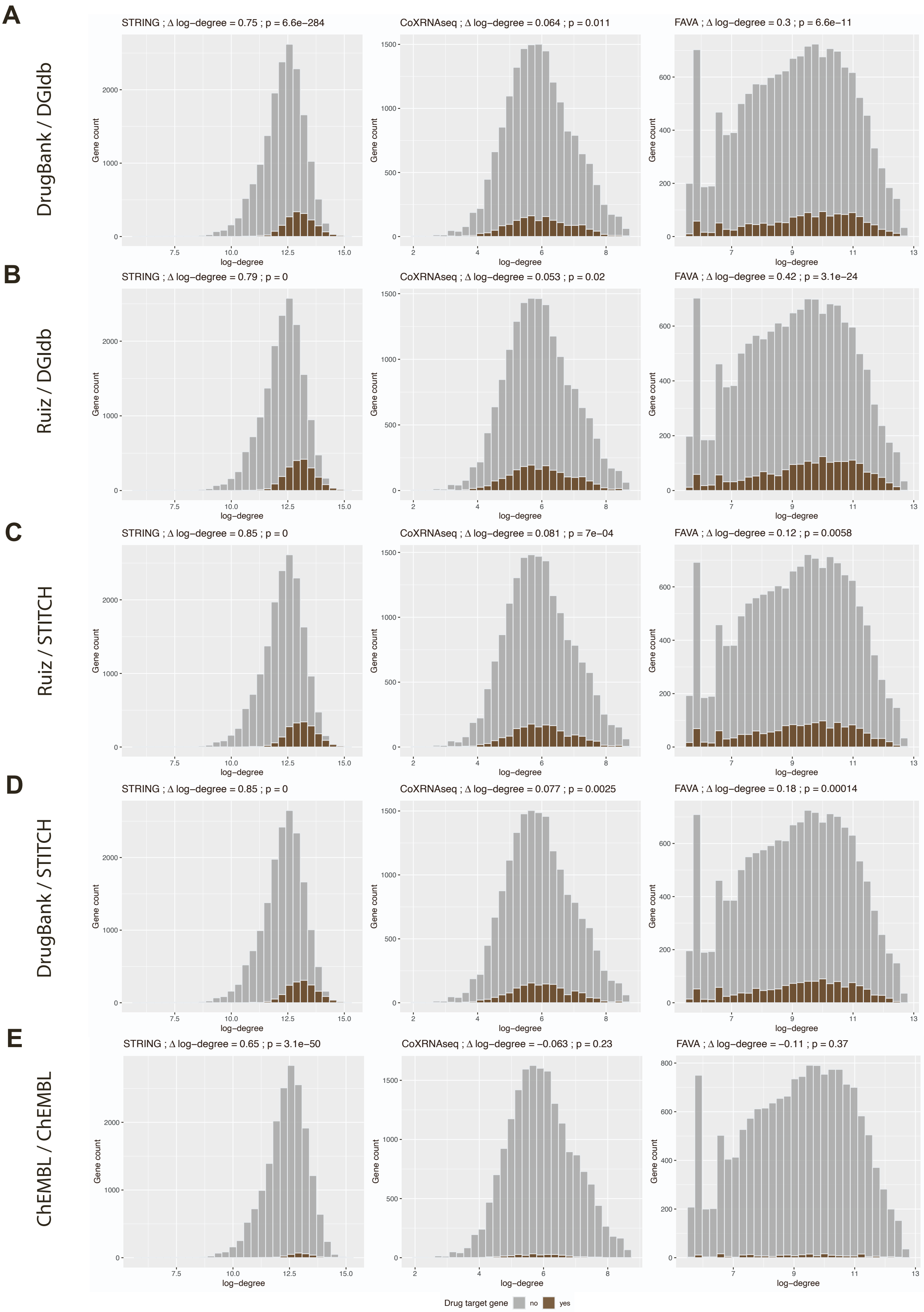


Figure S8. **Network degree distribution of drug target genes.**

Histograms showing the degree distribution of drug target genes and non-drug target genes in each network across drug databases (drug-indication and drug-target links, (A)-(E)). The difference in log-degree and the p-values from two-sided t-tests are shown in the title. This figure is related to Figure 5C which shows network degree distributions for the DrugBank /DGIdb combination.

Supplemental Tables

Table S14 . Network properties.
Network properties of the weighted networks that were analysed in this study (STRING: protein-protein interaction network, FAVA: co-expression network including proteomics, CoXRNAseq: co-expression network); related to STAR Methods section "Networks".

Nodes: number of nodes (genes) in the network.
Edges: number of total edges in the network.
Median degree: median degree in the network (i.e., weighted node degree)
Average log-degree: mean log-degree in the network
sd log-degree: standard deviation of the log-degree in the network

Network	STRING	FAVA	CoXRNAseq
Nodes	18573	15829	18695
Edges	11136598	951878	1119670
Median degree	257968.00	10672.79	356.44
Average log-degree	12.35	9.12	5.93
sd log-degree	0.96	1.73	1.02

Supplementary Table 15: Enrichment of testable and measured proteins for drug target genes.
Two-sided binomial test results to determine the enrichment of testable (~1,870, proteins that had at least 1 pQTL) and measured (~4,450) proteins for drug target genes among all protein-coding genes; related to STAR Methods section "Enrichment of proteins for drug targets".

Drug database	Observed proportion	Expected proportion	Pval	Set
Ruiz/DGIdb	0.2057	0.0964	2.05E-49	Testable
Ruiz/STITCH	0.2172	0.0964	2.49E-51	Testable
DrugBank/DGIdb	0.2164	0.0964	1.30E-47	Testable
DrugBank/STITCH	0.2246	0.0964	3.20E-50	Testable
ChEMBL/ChEMBL	0.1324	0.0964	5.04E-02	Testable
Ruiz/DGIdb	0.4045	0.2291	1.80E-69	Measured
Ruiz/STITCH	0.4689	0.2291	3.88E-108	Measured
DrugBank/DGIdb	0.4171	0.2291	1.15E-64	Measured
DrugBank/STITCH	0.4820	0.2291	3.76E-105	Measured
ChEMBL/ChEMBL	0.3015	0.2291	5.98E-03	Measured

Appendix C

Cardiometabolic drug response pharmacogenetics using EHRs from biobanks

This article is presented in Chapter 4.

Cardiometabolic drug response pharmacogenetics using EHRs from biobanks

Marie C. Sadler[1,2,3], Alexander Apostolov[3], Diogo M. Ribeiro[3],
Russ B. Altman[4], Zoltán Kutalik [1,2,3,*]

[1]University Center for Primary Care and Public Health, Lausanne, Switzerland

[2]Swiss Institute of Bioinformatics, Lausanne, Switzerland

[3]Department of Computational Biology, University of Lausanne, Lausanne, Switzerland

[4] Department of Bioengineering, Stanford University, Stanford, CA, USA

[*]Corresponding author: zoltan.kutalik@unil.ch

Abstract

DNA variants are known to contribute to inter-individual variability in drug response. Yet, despite the success of genome-wide association studies (GWAS) to unravel disease genetics, the genetic architecture of pharmacogenetic efficacy remains poorly understood. Here, we extract clinical and medication prescription data from electronic health records (EHRs) and conduct GWAS and rare variant burden test in the UK Biobank (discovery) and the All of Us program (replication) on ten cardiometabolic drug response outcomes including lipid response to statins, HbA1c response to metformin and blood pressure response to antihypertensives (N = 780-26,365). At genome-wide significance level, we replicate previously reported findings while also identifying *PCSK9* as a novel genetic determinant of LDL cholesterol response to statins (N = 17,063). We compare drug response genetics to disease and disease progression genetics in medication-naive individuals and find strong concordance, with 7 out of 14 signals being general prognostic and not drug-specific genetic markers. Furthermore, we demonstrate that individuals whose baseline condition is worse than expected based on their respective polygenic risk scores (PRS) are expected to have improved treatment efficacy. In summary, we highlight the value of using EHRs to study drug response and to identify clinically relevant genetic and environmental predictors that enable optimized treatment strategies.

Introduction

Genetic factors can contribute to inter-individual variability in drug response. However, despite the immense progress of genome-wide association studies (GWAS) for complex traits and diseases, progress in pharmacogenetics (PGx) to find genetic predictors of drug response is much slower. PGx GWAS represent less than 10% of all entries in the GWAS Catalog with median sample sizes of 1,220 for PGx GWAS published between 2016 and 2020 [1]. As a result of low sample size and lack of cohorts suitable for pharmacogenomic studies, relatively few robust PGx associations have been identified to date [1, 2, 3].

Several PGx GWAS consortia have formed over the years to study the genetics of drug efficacy in larger sample sizes. For instance, the Genomic Investigation of Statin Therapy (GIST) consortium has identified variants in the *LPA*, *APOE*, *SORT1*/*CELSR2*/*PSRC1* and *SLCO1B1* regions as modulators of low-density lipoprotein cholesterol (LDL-C) response to statins by combining randomized controlled trials (RCTs) and observational studies [4]. Similarly, the Metformin Genetics (MetGen) consortium has identified *SLC2A2* as influencing hemoglobin A1c (HbA1c) response to metformin [5], and more recently a meta-GWAS on HbA1c response to GLP-1 receptor agonists found variants in *ARRB1* to influence drug efficacy [6]. Furthermore, the International Consortium for Antihypertensive Pharmacogenomics Studies (ICAPS) has published multiple GWAS investigating blood pressure response to several antihypertensive drug classes (beta blockers, calcium channel blockers (CCBs), thiazide/thiazide-like diuretics, ACE-inhibitors (ACEi)/angiotensin receptor blockers (ARB)) [7, 8, 9].

Biobanks coupled with electronic health records (EHRs) that comprise medication data provide new opportunities to discover PGx associations [1, 10]. These massive datasets have already contributed to the replication of known PGx interactions as well as the discovery of new putative associations in national biobanks such as the Estonian [11] and UK Biobank (UKBB) [12, 13]. More recently, GWAS on longitudinal medication pattern extracted from the Finnish nationwide drug purchase registry in the FinnGen study identified tens of cardiometabolic risk loci specific to medication use and not associated with the underlying indication [14]. Yet, PGx biobank studies so far have either focused on known pharmacogenes and their associations with adverse drug reactions, drug dosage and drug prescribing behavior or analyzed the genetics of temporal medication use in isolation of disease phenotypes. What remains largely unexplored is the integration of longitudinal medication and phenotypic data to screen for genetic determinants of drug efficacy at a biobank scale.

Here, we extracted clinical and medication prescription data from EHRs and conducted PGx association analyses on the change in biomarkers following cardiometabolic drug therapy (Figure 1a). We assessed associations with both common and rare variants by performing GWAS and rare variant burden tests on sequencing data. Discovery analyses were conducted in the UK Biobank (UKBB) [15] and replication analyses in the All of Us (AoU) research program [16] (Figure 1b). In follow-up analyses, we compared drug response genetics to the genetics of baseline and longitudinal biomarker changes in medication-naive individuals to dissect medication- and disease-specific components (Figure 1c). Finally, we performed stratification analyses based on the polygenic risk scores (PRS) of the underlying disease and demonstrate their value in predicting drug response. In summary, we present a comprehensive resource of the genetic architecture of cardiometabolic drug response and showcase the value and challenges in analyzing EHR-coupled biobanks to study inter-individual variability in drug response.

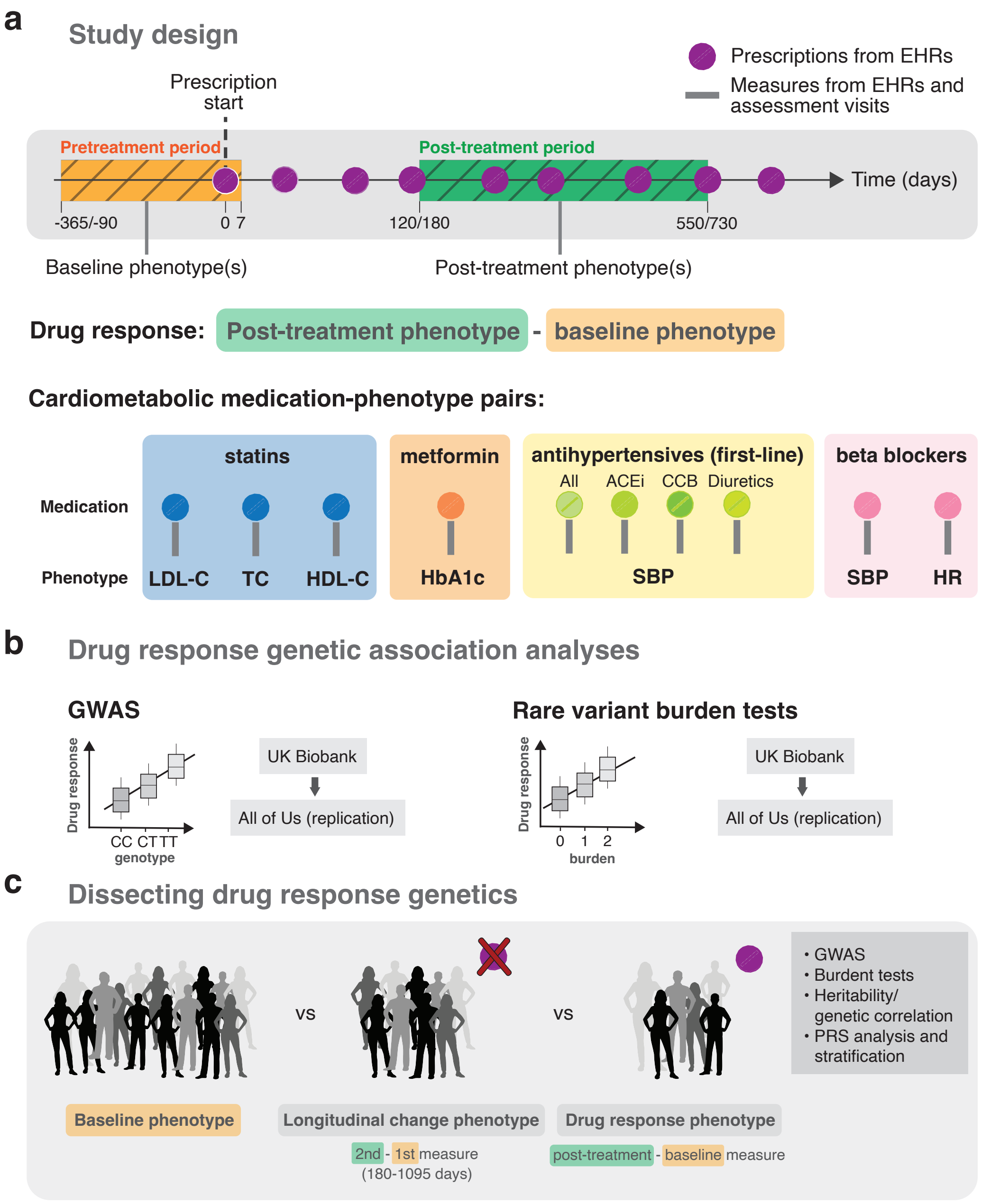


Figure 1: Study design. **a** Drug response study design using electronic health records (EHRs) from the UK and All of Us biobanks. Baseline and post-treatment phenotypes were extracted from EHRs or biobank assessment visits before and after the first recorded prescription, respectively. Different timings relative to the first prescription were tested as well as the use of single and average values over multiple baseline and post-treatment measures if available. Drug response phenotypes defined by the difference in post-treatment and baseline measures were tested for ten cardiometabolic medication-phenotype pairs. **b** Discovery genetic association analyses were conducted in the UK Biobank and replicated in the All of Us research program on common variants (GWAS analysis) and rare variants through burden tests. **c** Follow-up analyses compared the genetics of baseline, longitudinal change and drug response genetics.

Results

Overview of the analysis

In the drug response discovery analyses, we extracted longitudinal prescription and phenotypic data from the UKBB primary care data which we combined with phenotypic data from the assessment visits. We then constructed EHR-derived drug response cohorts for the following medication-phenotype pairs: statin-lipids (LDL-C, high-density lipoprotein cholesterol (HDL-C), total cholesterol (TC)), metformin-HbA1c, antihypertensive-systolic blood pressure (SBP; by antihypertensive class (ACEi, CCB, thiazide diuretics) and all classes combined), beta blocker-SBP and beta blocker-heart rate (HR). Individuals were only part of a drug response cohort, if a phenotype measurement was available before and after treatment initiation in addition to passing several other quality control (QC) steps (Method section: Study design and phenotype definitions, Figure S1, Table S3). For each drug response phenotype, we considered two filtering scenarios, a stringent and a lenient one. While more stringent QC should result in a cleaner phenotype definition, this comes at the cost of reduced sample size and thus statistical power. Given the sharp drop in sample size with more stringent criteria, the lenient filtering strategy constitutes the default setting throughout this study. In both stringent and lenient scenarios, we tested single and average baseline and post-treatment values over multiple measures with average values being the default (Figure 1).

In each drug response cohort, we first conducted GWAS to discover common genetic predictors (minor allele frequency (MAF) ≥ 0.01) of drug efficacy. In a second step, we performed genome-wide burden tests using whole exome sequencing (WES) data to assess associations with rare variants (MAF < 0.01). Replication analyses of identified PGx variants in the discovery analyses and across the literature were conducted in ~250,000 participants of the AoU research program with available whole genome sequencing data (WGS). Following genetic association studies, we compared drug response and underlying disease genetics, assessed baseline trait PRS as predictor of drug response and investigated the regression-to-the-mean phenomenon, whereby individuals converge to their genetically predicted biomarker level over time.

Drug response GWAS using EHRs from the UKBB

In the LDL-C response to statin GWAS, four loci were identified with the strongest signal being in the *APOE* region on chromosome 19 (rs7412 T>C, beta = -0.35, p-value = 1.53e-80), followed by the *SLC22A3*/*LPA* locus on chromosome 6 (rs10455872 G>A, beta = 0.15, p-value = 1.1e-21), *SORT1*/ *CELSR2*/*PSRC1* (rs7528419 G>A, beta = -0.086, p-value = 3.03e-15) and *PCSK9* locus (rs11591147 T>G, beta = -0.27, p-value = 2.2e-12) on chromosome 1 with the *SLC22A3*/*LPA* and *APOE* harbouring secondary signals (Table 1; Figure 2a; lenient filtering with average values if available, N = 17,063). TC response to statins, for which we had a larger sample size (more TC than LDL-C measures are available in the primary care data, N = 26,365) confirmed the identified loci at *PCSK9*, *SLC22A3*/*LPA*, and *APOE*. Two additional loci were found, *CETP* on chromosome 16 (rs12149545 A>G, beta = 0.05, p-value = 3.6e-10) and rs4149056 C>T in the *SLCO1B1* locus, also known as Val174Ala or SLCO1B1*5, on chromosome 12 (beta = 0.059, p-value = 2.1e-9). This SNP has previously been associated with LDL-C statin response [17] as well as clinical myopathy [18]. HDL-C response to statin GWAS identified *CETP* as single genome-wide significant locus (rs11076175 G>A, beta = -0.04, p-value = 5.4e-11) with the HDL-C component likely being responsible for this same signal in the TC response GWAS. No genome-wide significant hits were found in the HbA1c response to metformin GWAS (N = 4,124; Figure S7), SBP response to antihypertensives (N = 1,236-6,199; Figure S9) and HR/SBP response to beta blockers (N = 764-2,173; Figure S7).

The impact of single *vs* average baseline/post-treatment measures was minimal, and a difference was only observed for TC response to statin, with *SLCO1B1* reaching genome-wide significance only with average values (Figure S5; Table S5). The difference between stringent and lenient filtering was more pronounced, as sample sizes almost doubled with more lenient settings. For statins, this rise was largely due to the extended baseline period. For metformin and antihypertensives, we excluded individuals taking any related medication in the stringent filtering setting, whereas in the lenient setting, sample size largely increased by allowing metformin and antihypertensives to act as add-on therapy to sulfonylureas and other antihypertensives, respectively, if consistently taken during pre- and post-treatment periods of the studied medication (Figure S2-3). As a consequence of lower statistical power, only 9 out of the 14 signals found in the lipid-statin GWAS were detected in the stringent filtering scenarios (Figures S4-5; Table S5).

Replication analysis in the All of Us research program

We conducted replication analyses in the AoU program (v7; N $\approx$ 250,000 with available short-read WGS data). As in the UKBB, longitudinal prescription and phenotypic data were extracted from EHRs and used to construct drug response cohorts by following the same methodology as in the UKBB (Methods, Table S6). Cohort characteristics were similar as in the UKBB (Table S6, Figure S10). Mean statin starting age was 58 years compared to 61 years in the UKBB and as in the UKBB post-treatment lipid levels were on average measured a year after the first prescription. The main difference was observed in the regularity of statin prescriptions. Whereas in the UKBB, participants had on average a prescription every two months 87% of the time, this number dropped to 42% in the AoU. There were slightly less statin users as in the UKBB, but similar to the UKBB, the main reasons for being excluded in the PGx cohort were missing baseline and/or post-treatment measures in the considered time windows leaving 9,944, 6,713 and 11,120 individuals in the LDL-C, TC and HDL-C response to statins, respectively. Among the 14 signals, 7 replicated at the Bonferroni-corrected replication threshold of 0.05/14 = 0.00357 and 10 at a nominal p-value of 0.05 (all directionally concordant). Signals not replicating nominally include the secondary signal at the *SLC22A3*/*LPA* locus in the LDL-response to statin GWAS, and the *SLCO1B1*, *CETP* and secondary *APOE* (*PVR*/*CEACAM19*/*IGSF23* locus) in the TC response to statin GWAS, likely owing to the much lower sample size of TC in the AoU compared to the UKBB (N = 6,713 vs 26,365).

Table 1: Genome-wide significant loci in discovery analyses (UK Biobank) across all assessed cardiometabolic drug response traits together with replication results (All of Us).
Chr, chromosome; EAF, frequency of effect allele.

Pharmacogenetics trait	Chr	Position (GRCh37)	Lead SNP	Gene	Effect allele	Other allele	EAF (discovery)	N (discovery)	beta (discovery)	p-value (discovery)	EAF (replication)	N (replication)	beta (replication)	p-value (replication)
LDL-C response to statin	1	55505647	rs11591147	*PCSK9*	T	G	0.013	17063	-0.274	2.20E-12	0.10	9944	-0.210	3.89E-04
LDL-C response to statin	1	109817192	rs7528419	*PSRC1*, *SORT1*, *CELSR2*	G	A	0.207	17063	-0.086	3.03E-15	0.206	9941	-0.098	8.03E-12
LDL-C response to statin	6	160910282	rs9457927	*SLC22A3*, *LPA*	G	A	0.021	17063	0.211	7.44E-12	0.032	9942	0.039	2.08E-01
LDL-C response to statin	6	161010118	rs10455872	*SLC22A3*, *LPA*	G	A	0.090	17063	0.149	1.10E-21	0.050	9943	0.086	1.17E-03
LDL-C response to statin	19	45134682	rs62119267	*PVR*, *CEACAM19*, *IGSF23*	C	A	0.020	17063	-0.232	3.31E-13	0.012	9942	-0.135	9.98E-03
LDL-C response to statin	19	45412079	rs7412	*APOE*	T	C	0.059	17063	-0.354	1.53E-80	0.062	9944	-0.279	2.22E-31
TC response to statin	1	55505647	rs11591147	*PCSK9*	T	G	0.012	26365	-0.203	4.59E-10	0.009	6713	-0.156	4.82E-02
TC response to statin	6	160751531	rs9295128	*SLC22A3*, *LPA*	T	G	0.020	26365	0.177	8.31E-12	0.015	6713	0.185	2.56E-03
TC response to statin	6	161005610	rs55730499	*SLC22A3*, *LPA*	T	C	0.091	26365	0.127	1.37E-24	0.054	6709	0.069	3.49E-02
TC response to statin	12	21331549	rs4149056	*SLCO1B1*	C	T	0.150	26365	0.059	2.13E-09	0.142	6713	0.007	7.46E-01
TC response to statin	16	56993161	rs12149545	*CETP*	A	G	0.309	26365	0.048	3.60E-10	0.261	6711	0.011	5.06E-01
TC response to statin	19	45134682	rs62119267	*PVR*, *CEACAM19*, *IGSF23*	C	A	0.020	26365	-0.152	3.72E-09	0.012	6710	-0.071	2.90E-01
TC response to statin	19	45412079	rs7412	*APOE*	T	C	0.059	26365	-0.185	2.80E-35	0.058	6713	-0.122	8.15E-05
HDL-C response to statin	16	57006378	rs11076175	*CETP*	G	A	0.182	23429	-0.040	5.35E-11	0.198	11120	-0.041	2.77E-06

EHR- compare to cohort-derived PGx GWAS

From the literature we extracted genetic predictors reported for the assessed cardiometabolic medication-biomarker pairs. We adopted the criteria from Nelson *et al.*, 2016 [2] that provide a curated list up to July 2015 by querying the GWAS catalog [19]. Briefly, genetic variants were required to pass the genome-wide significance threshold of 5e-8 and show evidence of independent replication. Reported GWAS stem either from randomized controlled trials or observational studies often meta-analyzed together.

Five independent loci were reported for LDL-C response to statins of which three (*APOE*, *LPA*, and *SORT1*) and two (*APOE* and *SORT1*) passed genome-wide significance in the discovery (UKBB) and replication (AoU) cohort, respectively (Table 2). *SLCO1B1* locus was nominally significant in the UKBB (p-value = 1.21e-03) although genome-wide significant with TC as the assessed biomarker for which sample size was larger (p-value = 2.13e-09). *ABCG2* associated with LDL-C reduction following rosuvastatin therapy in the JUPITER trial [20] was found to be insignificant in the UKBB and AoU (p-values of $>$ 0.05) and did also not reach genome-wide significance in a later, larger GWAS meta-analysis of all statins combined [4]. HDL-C response GWAS to statins identified *CETP* as single genome-wide significant locus, in line with the UKBB-derived GWAS [21]. Overall, EHR-derived PGx signals on lipids agree well with those reported in cohort studies, with *PCSK9* found to be novel among the signals robustly replicating in the AoU.

GWAS of HbA1c-response to metformin identified *ATM* [22], *SLC2A2* [5] and *PRPF31* [23], none of which replicated in the EHR PGx GWAS (p-values $>$ 0.05). While this could be a power issue given the lower sample sizes (4,124 and 4,676 in the UKBB and AoU, respectively), it should also be noted that none of the studies have reported the same locus twice and the *ATM* and *SLC2A2* loci were insignificant in the ACCORD clinical trial GWAS that was conducted later (p-value $>$ 0.1) [23]. Although several loci have been found to influence blood pressure response to anti-hypertensives at a suggestive p-value threshold, no genome-wide significant hits have been reported [24].

Table 2: Genetic predictors of cardiometabolic drug response reported in the literature that were discovered or reproduced via genome-wide association study and passed a genome-wide significance threshold of 5e-8. Corresponding effect sizes and significance levels were retrieved from the EHR-derived genetic analyses in the discovery (UK Biobank) and replication (All of Us) cohort.
r^2 (%), percentage of variation explained derived from Z-score divided by the square root of the sample size (N); OR, odds ratio; UKBB, UK Biobank, AoU, All of Us; N/A, not available.

*results for LD-proxy rs45499402 (r^2 = 1)
**all statins combined
***r^2 with rs7412 = 0.64

Medication	Phenotype	Gene	SNP	r^2 (%)/OR literature	p-value literature	N literature	Reference	r^2 (%) UKBB	p-value UKBB	r^2 (%) AoU	p-value AoU
Rosuvastatin	LDL-C	*ABCG2*	rs2199936	0.96%	2.1E-12	3,523	[20]	0.0003%*,**	0.82	0.016%**	0.20
Statins	LDL-C	*APOE*	rs445925***	0.59%	8.52E-29	17,522	[4]	1.13%	6.91E-44	0.36%	2.28E-09
	LDL-C	*LPA*	rs10455872	0.54%	7.41E-44	31,056	[4]	0.54%	1.1E-21	0.11%	1.17E-03
	LDL-C	*SORT1*	rs646776	0.12%	1.05E-09	38,599	[4]	0.36%	6.88E-15	0.42%	8.44E-11
	LDL-C	*SLCO1B1*	rs2900478	0.12%	1.22E-09	24,253	[4]	0.061%	1.21E-03	0.022%	0.14
	HDL-C	*CETP*	rs247616	0.18%	8.52E-13	27,720	[21]	0.156%	1.46E-09	0.20%	1.96E-06
Metformin	*HbA1c* below 7%	*ATM*	rs11212617	1.35 (OR)	2.9E-09	3,920	[22]	0.00%	0.99	0.00%	0.96
	HbA1c	*SLC2A2*	rs8192675	0.17%	6.6E-14	10,577	[5]	0.018%	0.39	0.00%	0.89
	HbA1c	*PRPF31*	rs254271	0.26%	1.2E-08	8,273	[23]	N/A	N/A	0.027%	0.26

Rare variants have a modest impact

While common genetic variants have been assessed as predictors of drug response phenotypes in multiple studies, the impact of rare variation is less well known. Making use of sequencing data (WES and WGS in the UKBB and AoU, respectively), we conducted rare variant burden tests for all ten drug response phenotypes (Figure 1). We included missense and putative loss-of-function (LoF) variants with MAF $<$ 1% in optimal kernel association tests (SKATO) [25]. Two genes survived multiple testing correction (p-value $<$ 0.05/18,983 = 2.63e-06), namely *PCSK9* for LDL-C response (p-value = 7.55e-08) and *ABCA1* (p-value = 2.30e-07) for HDL-C response to statin. Given the genome-wide significant association of rs11591147 within the *PCSK9* locus, we conducted a sensitivity analysis to ascertain that the signal is not driven by rare variants in LD with this SNP. Conditioning on rs11591147, the association with *PCSK9* remained significant (p-value = 3.89e-08) meaning that common and rare variants have an independent effect on LDL-C response in this gene region. *ABCA1* is known to be responsible for cellular cholesterol efflux and mutations in this gene have been shown to cause familial HDL deficiency and Tangier disease, a rare and severe form of HDL deficiency [26]. Both gene associations replicated in the AoU at a p-value $<$ 0.05/2 = 0.025 (p-values of 1.21e-05 and 0.0243 for *PCSK9*-LDL-C and *ABCA1*-HDL-C, respectively).

Baseline, longitudinal change and drug response genetics are similar

As most of the identified genetic loci can also be identified in GWAS of the underlying disease, we tested whether drug response variability is primarily due to drug-gene interactions or liability to disease. First, we repeated the GWAS analysis by including the respective lipid biomarker PRS as covariate. In the LDL-C response to statin GWAS, the *SORT1*/*CELSR2*/*PSRC1* locus no longer reached genome-wide significance, and neither did the *CETP* locus in the TC and HDL-C response GWAS (Figure 2a, Table S5). In addition, effect sizes of the *APOE* locus were attenuated in the LDL-C and TC response GWAS. Disappearance and attenuation of GWAS loci suggests that drug response signals may be due to the underlying disease liability and not the treatment itself while also suggesting that disease PRS can serve as predictors of drug response (see next section).

We further conducted a control GWAS, or disease progression GWAS, where instead of before and after measures relative to medication start, we used two measures with an equivalent time interval between them in medication-naïve individuals (Methods). The *SORT1*/*CELSR2*/*PSRC1*, *CETP* and *APOE* loci (Figure S10, Table S6) all reached genome-wide significance, evidencing that these loci are not or only partly specific to the treatment intervention. The same trend held for the two rare-variant associations that were detected in statin-users. Although not surviving multiple testing correction, the burden of rare coding variants in both *PCSK9* and *ABCA1* were strongly associated with LDL-C (3.43e-06) and HDL-C (p-value = 3.91e-06) longitudinal change, respectively. Genetic associations mirrored in drug response and disease progression analyses are likely to arise due to the regression-to-the-mean effect, i.e., the convergence to genetically predicted levels to which these loci strongly contribute. When calculating the genetic correlations between the biomarker levels vs their change upon medication use, we observed indeed a strong concordance across drug-response phenotypes ranging from r_g of 0.20 to 0.80 (Figure 2b, Table S7). Genetic correlations between drug response and disease progression (r_g of 0.27 to 0.69), as well as between disease progression and baseline biomarker levels ($r_g > 0.73$) were also found to be very high (Table S7).

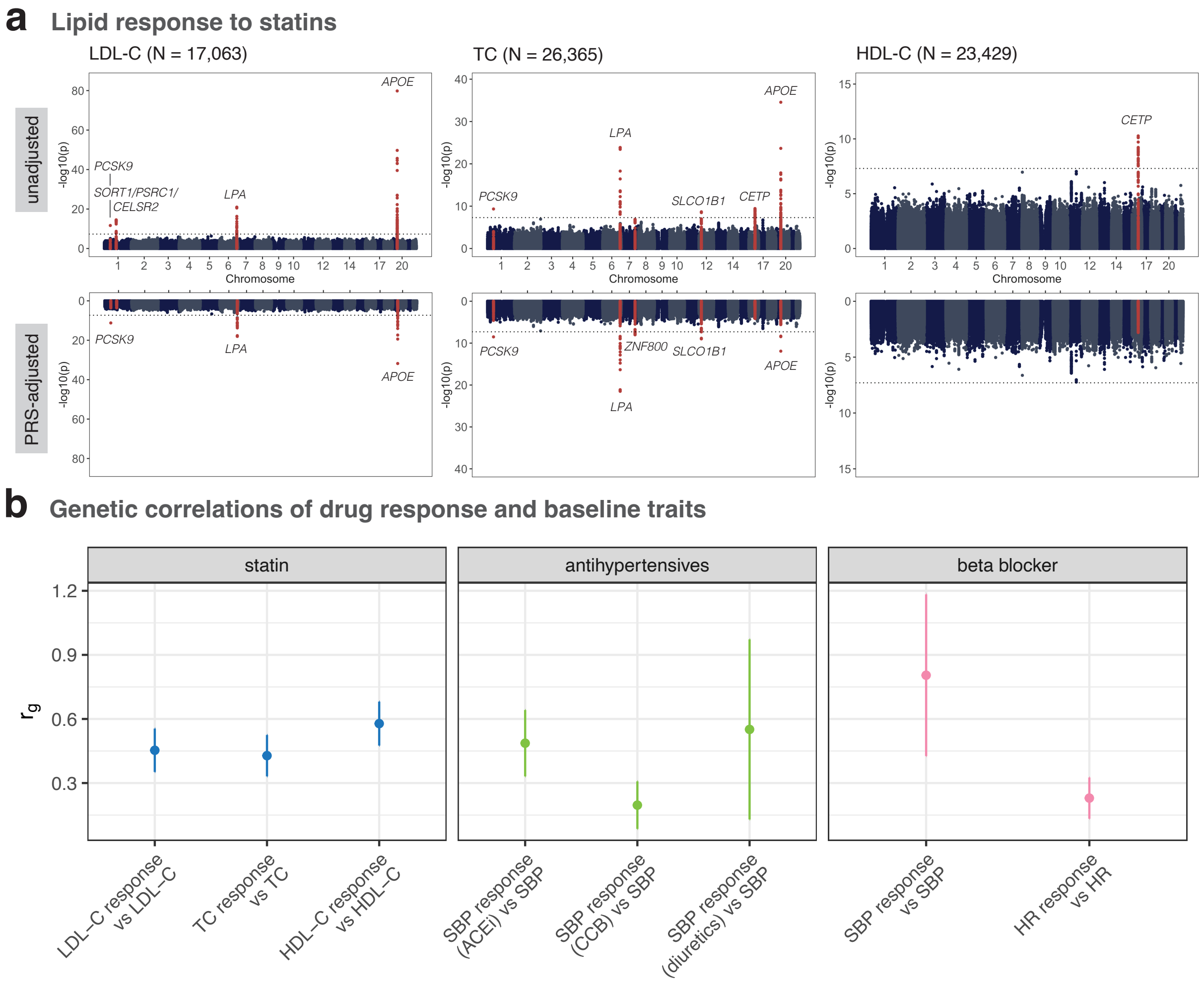


Figure 2: PGx GWAS results and genetic correlation with baseline traits. **a** Manhattan plots of lipid response to statins derived from EHRs of the UKBB (lenient filtering with average values over multiple measures). GWAS association results of the bottom row are adjusted for the polygenic risk scores (PRS) of the assessed lipid trait. Loci significant in either the adjusted or unadjusted setting are highlighted in red and annotated with the closest gene. The horizontal line denotes genome-wide significance (p-value $<$ 5e-8). **b** Genetic correlations (r_g) between PGx drug response and underlying baseline trait for cardiometabolic medication-phenotype pairs. Error bars correspond to the standard error.

In Figure 3a-d, we depict the comparison between disease, disease progression and drug response genetics at the SNP-level. The signal at the *APOE* locus was strongest for baseline LDL-C (b_{base} = 0.60, p-value $<$ 1e-300), remains for LDL-C longitudinal change (b_0 = 0.20, p-value = 1.80e-108), but with a significantly weaker effect than for the drug response (b_{drug} = 0.30, p-value = 1.60e-24; p_{diff} = 0.0074; simvastatin 40mg users who represent the largest starting statin type-dose group; Figure 3a). Conversely, the *SORT1* signal was significant in all three scenarios, but with no difference in effect sizes between longitudinal change and drug response LDL-C (p_{diff} = 0.11; Figure 3b). *SLCO1B1* is the only locus associated to neither TC baseline (p-value = 0.12) nor longitudinal change (p-value = 0.92), evidencing its sole implication in pharmacokinetics (Figure 3c). Similarly to *SORT1*, the role of *CETP* in drug response is merely driven by the regression-to-the-mean phenomenon (p_{diff} = 0.29; Figure 3d). Assessing all 14 genome-wide significant signals, we found that 7 were driven by disease liability ($p_{\text{diff}} > 0.05$; Table S9). While comparison with disease progression in controls allows to disentangle which loci are disease- and which are medication-specific, it also shows that PRS-adjustment in drug-response GWAS can already hint the difference to some extent (*PCSK9* remained significant after PRS adjustment, although analysis on the SNP-level could not find a difference between drug response and longitudinal change).

Furthermore, we disentangled whether high PRS contribute to better or worse drug response outcomes. While lipid PRS clearly determine baseline levels, they also affect lipid evolution (Figure 3e-f). Nonetheless, the effect of LDL-C PRS on drug response (b_{drug} = 0.134, p-value = 7.42e-39) was slightly stronger than on longitudinal change (b_0 = 0.131, p-value = 1.68e-257; p_{diff} = 5.5e-26) suggesting that a high PRS has an independent deteriorating effect on treatment efficacy. This effect was reversed for individuals with a high HDL-C PRS: Upon statin treatment, increase in HDL-C was not as high as for statin-free controls (p_{diff} = 0.044).

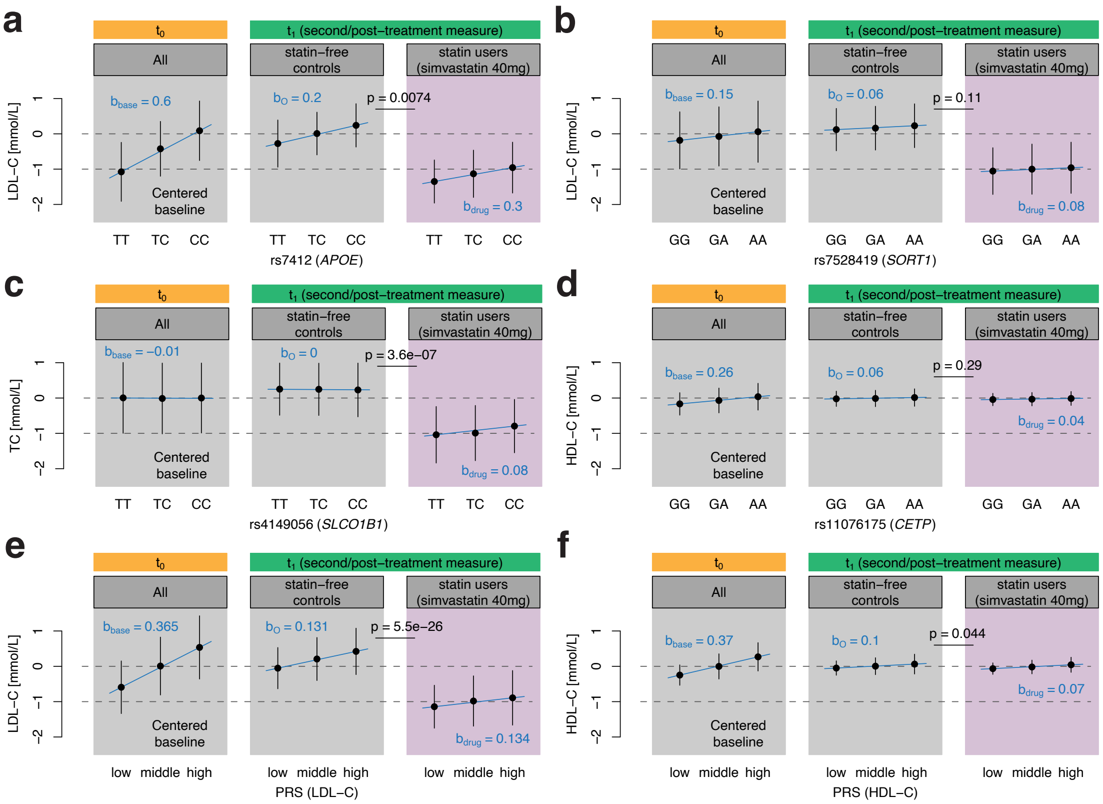


Figure 3: Comparison of baseline, longitudinal change and drug response genetics stratified by PGx variants (**a**-**d**) and lipid PRS (**e**-**f**). The baseline panel (t_0) groups statin-free controls and statin users (simvastatin 40mg corresponding to the largest starting statin type-dose group), and shows their sex and age adjusted, centered baseline level stratified by genotype/PRS. The following two panels (t_1) show for each group their follow-up measure (either second or post-treatment measure) stratified by genotype/PRS. The follow-up measure is adjusted for sex, age and baseline with adjustment performed prior statin-status stratification. Genotype/PRS association slopes with lipid levels at baseline (b_{base}), second time point in statin-free controls (b_0) and statin users (b_{drug}) were derived through regression of the standardized phenotypes (baseline/second/post-treatment measure, respectively) on the genotype dosage/PRS adjusted for sex, age and baseline (for t_1 measures). The p-value (p) shows the difference in slope between statin and non-statin users and corresponds to the significance level of the genotype dosage/PRS and drug status interaction term. Low, middle and high PRS stratified groups were defined by the lowest and highest decile cut-offs, respectively, however, in all regression analyses the continuous PRS distribution was evaluated. Dots correspond to the mean of the adjusted baseline and second/post-treatment measure in each stratified group and error bars to the standard deviation. Numeric values and number of individuals within each stratum are shown in Table S10.

Polygenic risk scores as predictors of drug response

The strong associations between drug response and PRS, as well as the high genetic correlations between drug response and underlying traits, indicates the potential of PRS for drug response prediction and patient stratification. Although PRS can serve as a predictor of drug response, starting baseline levels remain the best predictor of post-treatment levels (Figure 4a). Focusing on LDL-C response to statin, HbA1c response to metformin and SBP response to antihypertensives, we found that baseline levels explained 23.9%, 10.4%, and 12.8% of the variation, respectively. Explained variance increased to 25.8%, 10.6%, and 12.84% when integrating corresponding LDL-C, type 2 diabetes (T2D) and hypertension (HT) PRS with the p-values of the PRS, conditional on baseline, being 2.61e-93, 0.011 and 0.041, respectively. When additionally including all cohort-specific covariates (Table S3 with the exception of principal components), the explained variance of drug response increased to 30.0%, 17.2%, and 14.8%. Strong associations between drug response and PRS were also observed for the remaining medication-phenotype pairs (Table S11).

In Figure 4b, we highlight how additional stratification by drug response genetic signals can improve prediction accuracy for post-treatment LDL-C levels following statin initiation. Additional stratification by the *APOE* genotype, the top signal in the LDL-C response GWAS, increased the explained variance to 26.2% compared to 25.8% for baseline and LDL-C PRS predictors alone. It is important to note that the effect of PRS on both baseline levels and drug efficacy can lead to differing interpretation as to whether high PRS increases treatment benefits or leads to worse outcomes. LDL-C reduction is biggest for individuals with high LDL-C PRS (1SD increase in PRS leads to 0.075 mmol/L increased reduction, p-value = 2.81e-28), however, once accounting for baseline levels, LDL-C reduction is smaller in these individuals (1SD increase in PRS leads to 0.11 mmol/L less reduction, p-value = 2.61e-93; Table S11). Thus, for same starting levels a higher genetic burden decreases drug efficacy.

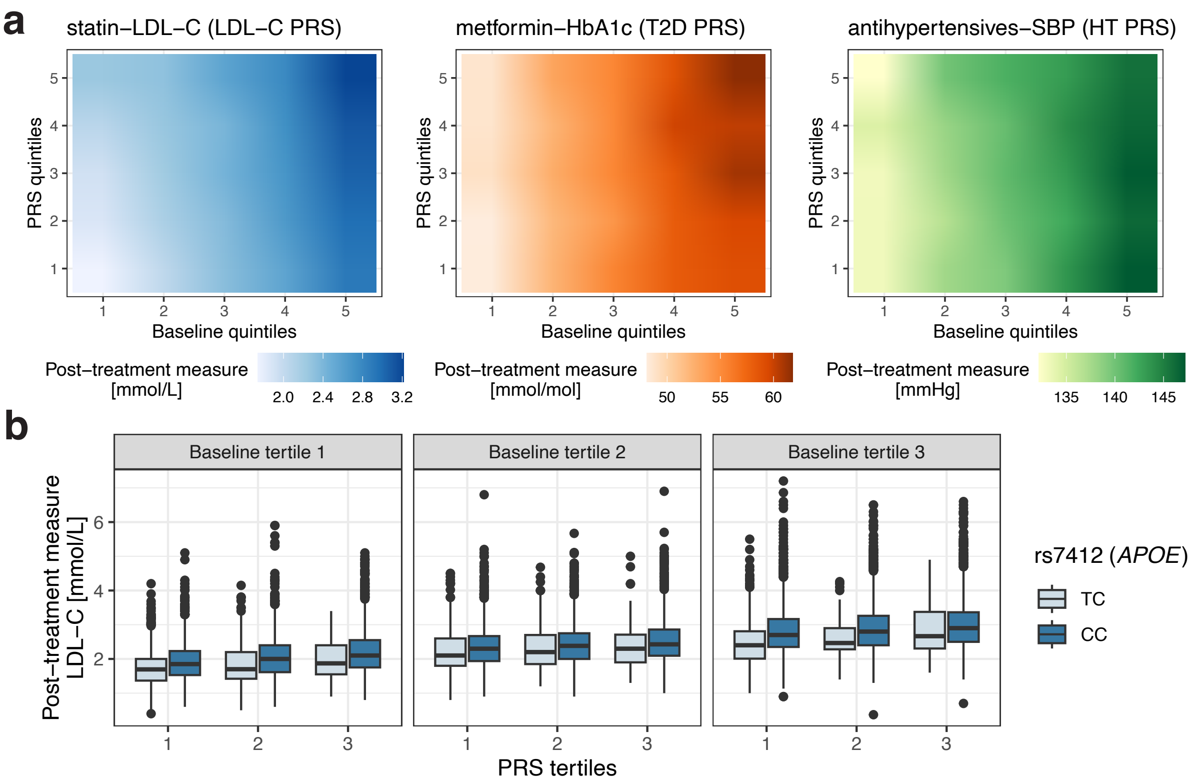


Figure 4: Drug response measures stratified by baseline and PRS. **a** Individuals stratified by 1) baseline levels (LDL-C, HbA1c and SBP, respectively) and 2) PRS (LDL-C, T2D and hypertension (HT), respectively) quintiles with each tile displaying the average post-treatment value (LDL-C, HbA1c and SBP, respectively). Number of individuals for statin-LDL, metformin-HbA1c and antihypertensives-SBP are 17,044, 4,136 and 6,226, respectively. **b** Individuals taking statins stratified by 1) LDL-C baseline levels, 2) LDL-C PRS and 3) rs7412 genotype (individuals with the TT genotype are omitted as their sample size was too low). Boxes bound the 25th, 50th (median, centre), and the 75th quantile. Whiskers range from minima (Q1 – 1.5*IQR) to maxima (Q3 + 1.5*IQR) with points above or below representing potential outliers. Numerical values and number of individuals within each stratum are shown in Table S12.

Discussion

In this study, we demonstrate the value of biobanks coupled to EHRs to study the genetics of cardiometabolic disease medications. We conducted discovery in the UKBB and replication analyses in the AoU, and assessed the impact of common and rare variations on drug efficacy. We show that signals from EHR-derived PGx GWAS are concordant with those observed in the literature and dissect medication- and disease-specific components.

Overall, we found only a few genetic variants to influence cardiometabolic drug response in line with other studies that often identified only a few or even no genome-wide significant signals [6, 8, 27]. A review on drug efficacy GWAS reported that only 15% of drugs exhibit robust gene-treatment interactions [2]. While we could identify a novel PGx lipid locus, samples sizes remain too low to have a definite answer on whether low numbers of genetic predictors are a consequence of limited statistical power or a lack of genetic influence on drug response which would be corroborated by low and often insignificant heritability estimates (Table S8). We could not find evidence for rare variants to play a major role in drug response variability and the associations we found are likely driven by disease susceptibility.

Merely assessing drug response variability in treated individuals can make it challenging to distinguish between prognostic (related to disease progression) and treatment-specific genetic markers. While RCT data with a control arm remain the gold standard to differentiate between the two, large biobank data also allow to construct (non-randomized) control groups. In several comparative analyses we show the similarity between baseline, drug response and disease progression genetics and highlight which loci exhibit pure prognostic effects. We also demonstrate that integrating PRS of the underlying disease as covariate in GWAS can at least partially correct for disease- while sparing drug-specific effects.

Even though disease liability is not treatment-specific, the predictive value of disease PRS for drug response can be clinically relevant and we found that high PRS led to lower biomarker reductions when accounting for baseline levels. Previously, several studies showed significant associations between disease PRS and drug response, although as highlighted before, the effect of PRS on both baseline and drug efficacy, and adjustments thereof, can result in opposing findings or interpretations as to whether high PRS increases treatment benefits or risk for treatment resistance. A recent study showed that sulfonylureas therapy was more effective in participants with higher T2D PRS with findings replicated in a separate cohort [28]. On the other hand, high schizophrenia PRS were found to reduce antipsychotic efficacy [29] and similarly high LDL-C and SBP PRS were associated with uncontrolled hypercholesterolaemia and hypertension, respectively [30]. Using RCT data, high coronary heart disease (CHD) genetic risk was found to associate with increased CHD risk, although the comparison between controls and treated participants revealed that relative risk reductions were higher among individuals with a high PRS, suggesting that this group benefited the most from lipid-lowering therapy [31, 32, 33, 34]. Taken together, our results and these studies seem concordant with the paradigm that a genetic burden leads to worse outcomes overall, however, treatment potential being higher in high-risk patients (i.e., higher baselines), we can observe larger relative treatment benefits in individuals with increased genetic risk. The reason for these opposing forces, we believe, is that higher baseline levels for non-genetic reasons can be easier alleviated by medication, while those with unfavourable biomarker levels due to genetic reasons are less amenable to correction via medication. An alternative explanation (which can also be an argument for simple longitudinal change, i.e., non-drug-specific change) is that individuals with temporally increased biomarker level at baseline are bound to regress back to their lifecourse mean, and even more so if their genetic risk is low.

Our study has several limitations. First, we rely on data from EHRs to derive before and after treatment biomarker levels, and thus cannot exclude the possibility that individuals were already on medication prior the first recorded prescription. Second, despite a large fraction of individuals with medication records in the biobanks, final PGx cohort sample sizes are limited by the number participants on a certain medication and further reduced due to incomplete or missing data. Of the ∼65,000 participants with a statin prescription in the UKBB, 63% could not be considered for the LDL-C response analysis because of missing baseline and/or post-treatment measures. Third, polypharmacy has only been taken into account within and not across medication groups. Even within, especially for antihypertensives where frequent changes in medication regimen occur, it can be difficult to determine appropriate filtering and covariate strategies to study individual drug classes as sample sizes are too low when restricting the analysis to individuals taking antihypertensives from a single class (i.e., stringent filtering strategy). Forth, our analysis focuses on continuous biomarkers and not clinical events. LDL-C, SBP, HR and HBA1c merely serve as surrogate end points of CHD and T2D events, and the genetic interplay with drug efficacy may be different when assessing hard clinical endpoints. Finally, we rely on observational data to draw conclusions about drug efficacy. Although, we contrast the results with control analyses on longitudinal biomarker change, control and medication groups were not defined randomly and by definition have markedly different disease profiles.

To conclude, we show that EHRs enable new opportunities to study drug response and reveal the complex contribution of genetic and environmental components to drug efficacy. While we find that the influence of common and rare genetic variants on drug response is relatively low, larger sample sizes will be needed to capture the full extent.

Methods

Study population

The UK Biobank is a prospective study of ~500,000 participants of whom 45% (N $\approx$ 230,000) are linked to the primary care data of the United Kingdom's National Health System [15]. The primary care resource contains longitudinal data of GP prescription records (datafield #42039) and GP clinical event records (datafield #42040) encoded through British National Formulary (BNF), National Health Service (NHS) dictionary of medicines and devices (DM+D), Read V2 and Clinical Terms Version 3 (CTV3) codes and are available up to 2016 or 2017 (depending on the data supplier). Analyses were conducted on individuals of white British ancestry, with no excessive number of relatives and differing reported and inferred gender, excluding participants who have withdrawn their consent (UKBB Sample-QC #531; N $\approx$ 200,000).

Study design and drug response phenotypes

We derived drug response phenotypes for the following cardiometabolic medication-phenotype pairs: statin-lipids (LDL-C, HDL-C, TC), metformin-HbA1c, antihypertensive-SBP (by antihypertensive class and all classes combined), beta blocker-SBP and beta blocker-HR. For each drug response phenotype, we considered stringent and lenient filtering scenarios which differed by regularity in prescription pattern, pre-treatment and post-treatment time windows as well as handling of treatment changes (e.g. dose change) and concomitant medication (e.g. add-on therapy). In Figure S1 and Table S3, we outline the different QC filters applied to each scenario. To further increase the number of available clinical measures, we added measures from the initial and repeated assessment visits with their respective time stamps to the pool of longitudinal data (LDL-C: #30780, HDL-C: #30760, TC: #30690, HbA1c: #30750, SBP: #4080, HR: #102). Read V2 and CTV3 codes encoding these variables in the primary care data are listed in Table S1 (see Note S1 for HbA1c unit conversion). Baseline measures were taken three months (stringent filtering) or up to a year (lenient filtering) before treatment initiation and 7 days after, either as the closest measure to treatment start or an average of all available measures during the pre-treatment period. Post-treatment period was defined as 6 months after medication start (4 months for SBP and HR as the effects of antihypertensives and beta blockers are expected to be more immediate than for statins and metformin) up to 1.5 (stringent) and 2 (lenient) years after, and either the closest measure to treatment start or an average of all available measures during the post-treatment period were taken. Consequently, we derived drug response phenotypes for four scenarios: stringent filtering-single measure, stringent filtering-average measures, lenient filtering-single measure, lenient filtering-average measures.

To determine medication regimens (medication start, treatment changes, prescription regularity), we first extracted all available prescriptions for each broader medication class (lipid-regulating, antidiabetic including insulin, and antihypertensives; BNF and Read V2 codes in Table S2). We then selected individuals with entries of the medication of interest (primary medication) and omitted individuals taking medications other than the primary medication of the same class within a year of initiating the primary medication. In the lenient filtering scenarios, we considered exceptions to this rule such as metformin being an add-on therapy to sulfonylureas, with sulfonylureas treatment being a covariate. Allowed scenarios for add-on therapy for antihypertensive are shown in Figure S2. Individuals taking primary medications in combination with a medication of the same class (e.g. statins in combination with ezetemibe) were filtered out (see Note S2). If multiple drugs corresponded to a medication class (e.g. different statin types), we included all drugs taken by at least 20 individuals. When BNF codes were truncated to miss the drug ingredient, we extracted them by matching drug names and brand names in the drug description. Likewise, dosage information was retrieved from the description using regular expressions [12].

In Table S4, we show the study characteristics of the individuals in each drug response phenotype cohort. Furthermore, flow diagrams in Figure S3 show the number of individuals after each QC step. The different QC steps were as follows: i) available baseline and post-treatment measures, ii) presence of a primary care record other than baseline/primary medication within the two years preceding medication start to avoid falsely considering a change to a new health care provider as a first prescription, iii) presence of a prescription part of the broader medication class after post-treatment measure, iv) drug change between medication start and post-treatment measure, v) regular prescriptions proxying drug adherence and vi) minimum baseline level (e.g. LDL-C $\geq$ 2.6 mmol/L). We only considered cohorts with more than 500 individuals for GWAS analyses.

GWAS

In the genetic association analyses, we define the drug response phenotype as the post-treatment measure adjusted for baseline measure and study-specific covariates including sex, age at the time of medication start, time between medication start and post-treatment measure, drug type and dose if applicable (Table S3) as follows:

$$\textit{post-treatment phenotype} = G_i + \textit{Baseline phenotype} + \textit{covariates} + \textit{PC1-20} + \epsilon \tag{1}$$

where G_i is the genotype under assessment and PC1-20 are the first 20 principal components. Using post-treatment measures adjusted for baseline levels results in the same analysis as using the difference between post-treatment and baseline measures adjusted for baseline levels [35] .

GWAS analyses were conducted using REGENIE (v3.2.6) which accounts for sample relatedness [36]. REGENIE first fits a whole-genome regression model (step 1) before testing each SNP in a leave-one-chromosome-out (LOCO) scheme (step 2). In step 1, genotyped SNPs were filtered as follows using PLINK2 [37]: minor allele frequency (MAF) $\geq$ 0.01, Hardy-Weinberg equilibrium p-value $\geq$ 1e-15, genotyping rate $\geq$ 0.99, not present in high linkage disequilibrium (LD) regions [38], not involved in inter-chromosomal LD [36] and passing LD pruning at $r^2 <$ 0.9 with a window size of 1,000 markers and a step size of 100 markers which resulted in 424,544 SNPs included in step 1. In step 2, variants imputed by the Haplotype Reference Consortium panel with a MAF $\geq$ 0.01 were tested (up to 7.5 million markers depending on phenotype sample size). Individuals with missing genetic data and/or not passing genetic QC were excluded from the analysis. Independent signals were defined as $r^2 <$ 0.001 and clumping was performed using PLINK.

Rare variant analysis

Rare variant analyses were conducted using REGENIE (v3.2.9). Phenotype definitions and step 1 whole genome regression were the same way as in the GWAS analyses. In step 2, we performed rare variant burden tests using optimal kernel association tests (SKATO) [25]. Masks were constructed from rare variants (MAF $<$ 0.01) including missense and putative LoF variants. Variant annotations and gene set definitions were derived following the original quality functionally equivalent (OQFE) protocol and provided on the UK Biobank DNAnexus research analysis platform [39]. Burden tests were then conducted on OQFE WES data [39].

Replication in the All of Us biobank

The All of Us research program is a prospective cohort recruiting up to 1 million participants [16]. Replication analyses were conducted in the current release (v7) in which genotype data were available for ~310,000 and WGS data for ~250,000 individuals. In the AoU database, the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) is used for standardized vocabularies and harmonized data representations. Medication records were retrieved based on concept id codes from the RxNorm vocabulary and phenotypes from the SNOMED vocabulary. Replication analyses were restricted to lipid response to statins and HbA1c response to metformin for which genome-wide significant signals were obtained either in the UKBB analyses or reported in the literature.

Similarly to the UKBB, we extract medication records by starting from the broader medication class (lipid modifying agents (concept id 21601853) and blood glucose lowering drugs (concept id 21600744)) which were then classified into primary medications (statins (concept id 21601855) and metformin (concept id 1503297)), combination therapies (lipid modifying agents, combinations (concept id 21601898), blood glucose lowering drugs, combination (concept id 21600765) and sulfonylureas (concept id 21600749)) and related medication from the same class. Dose information were extracted from the drug concept entries using regular expression or imputed by the median dose of the drug in question when not available. Phenotypes were extracted based on the following ancestor concept ids: LDL-C

(3028437), HDL-C (3007070), TC (3027114) and HbA1c (3004410). Only measures with available units and values in the plausible range were retained (Table S1). While lipid measures were recorded as mmol/L and primarily as mmol/mol for HbA1c in the UKBB (Note S1), units were mg/dL and % for lipids and HbA1c, respectively, which we left unconverted. Following extraction of longitudinal medication and phenotype measures, we followed the same QC steps as in the UKBB by applying the lenient filtering strategy with average baseline and post-treatment measures (Figure S1). Drug prescription regularity was found to be lower in the AoU, likely because drug prescriptions are only recorded from participating EHR sites. As a consequence, we lowered the drug regularity QC parameter and required a single prescription between medication start and post-treatment measures (QC9, Figure S1). Cohort characteristics and reason for removal are reported in Table S6 and Figure S10, respectively.

GWAS: GWAS analyses were conducted using REGENIE (v3.2.4). For step 1, we used genotyped SNPs and filtered them as follows using PLINK2: autosomal SNPs, MAF $\geq$ 0.01, Hardy-Weinberg equilibrium p-value $\geq$ 1e-15, genotyping rate $\geq$ 0.99, not present in high linkage disequilibrium (LD) regions [38] and passing LD pruning at $r^2 <$ 0.9 with a window size of 1,000 markers and a step size of 100 markers which resulted in 238,888 SNPs. The first 20 PCs were computed on the same set of SNPs using the FastPCA algorithm implemented in PLINK2 [40]. In step 2, we used WGS data from the Allele Count/Allele Frequency (ACAF) threshold callset to test associations between the genotypes of interest and drug response phenotypes.

Rare variant analysis: We conducted SKATO analyses on rare variants from the exon regions using REGENIE (v3.2.4) with step 1 being the same as in the GWAS. Variant annotations and gene set definitions were extracted from the Variant Annotation Table (VAT) provided by the AoU. Missense variants and putative LoF variants defined as stop gained, frameshift, splice donor and splice acceptor with MAF $<$ 0.01 were included in the burden tests.

PRS and genetic correlations

We extracted PRS from the UKBB that were derived from external GWAS data only ("the Standard PRS set" [41]) for the following phenotypes: LDL-C: #26250, HDL-C: #26242, HbA1c: #26238, T2D: #26285, HT: #26244. We then calculated the association between PRS and the change in biomarker level (baseline - post-treatment phenotype) either adjusted or unadjusted for baseline levels.

We calculated genetic correlations between traits using the GenomicSEM R package (v0.0.5c) [42]. Trait GWAS summary statistics were obtained from the following consortia: LDL-C, HDL-C and TC from the Global Lipids Genetics Consortium [43] (N up to 1,320,016; European ancestry), HbA1c from the UKBB (#30750, N = 344,182), SBP from a meta-analysis of the UKBB and the International Consortium of Blood Pressure [44] (N up to 757,601) and HR from the UKBB (#102, N = 340,162).

Longitudinal biomarker change GWAS

We conducted biomarker change GWAS in individuals part of the primary care data that did not have any drug prescription indicated for the investigated disease/surrogate end point (i.e., broad medication class, Table S2). All participants in this set with two available measures spaced between 6 months and 3 years which corresponds to the maximum allowed time interval between baseline and post-treatment measures were included. GWAS analyses were conducted analogous to the drug response GWAS, replacing baseline with first and post-treatment with second phenotype measure. We used the same covariates as in the corresponding drug response cohorts omitting drug-specific variables (Table S3).

Data availability

Genetic and phenotypic data from the UK Biobank and All of Us Resource are available to approved researchers.

All GWAS and rare variant burden test summary statistics will be available at the GWAS Catalog upon publication.

Code availability

GWAS calculations were performed with REGENIE (v3.2.6) which is available at `https://github.com/rgcgithub/regenie`. Genetic correlations were calculated with the GenomicSEM R package (v0.0.5c) available at `https://github.com/GenomicSEM/GenomicSEM`.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

MCS and ZK conceived and designed the study. MCS performed statistical analyses in the UK Biobank. AA conducted replication analyses in the All of Us research program under the supervision of MCS and RBA. DMR provided guidance on analyzing rare variants from sequencing data. ZK supervised all statistical analyses. All the authors contributed by providing advice on interpretation of results. MCS and ZK drafted the manuscript. All authors read, approved, and provided feedback on the final manuscript.

Acknowledgements

This work was supported by the Swiss National Science Foundation (310030_189147) to ZK. MCS would like to thank the Fulbright Program for funding her research stay at Stanford University where this research work has started under the co-supervision of RBA. This research has been conducted using the UK Biobank Resource under Application Number 16389. LD was calculated based on the UK10K data resource (EGAD00001000740, EGAD00001000741). Computations were performed on the Urblauna cluster of the University of Lausanne. We also would like to acknowledge the participants and investigators of the UK Biobank and All of Us study. We thank Greg McInnes for his help and support in analyzing medication prescription data in the UK Biobank and Ewan Pearson for his advice in defining drug response phenotypes from the UK Biobank primary care data.

References

[1] Gregory McInnes, Sook Wah Yee, Yash Pershad, and Russ B Altman. Genomewide association studies in pharmacogenomics. *Clinical Pharmacology & Therapeutics*, 110(3):637–648, 2021.

[2] Matthew R Nelson, Toby Johnson, Liling Warren, Arlene R Hughes, Stephanie L Chissoe, Chun-Fang Xu, and Dawn M Waterworth. The genetics of drug efficacy: opportunities and challenges. *Nature Reviews Genetics*, 17(4):197–206, 2016.

[3] Munir Pirmohamed. Pharmacogenomics: Current status and future perspectives. *Nature Reviews Genetics*, pages 1–13, 2023.

[4] Iris Postmus, Stella Trompet, Harshal A Deshmukh, Michael R Barnes, Xiaohui Li, Helen R Warren, Daniel I Chasman, Kaixin Zhou, Benoit J Arsenault, Louise A Donnelly, et al. Pharmacogenetic meta-analysis of genome-wide association studies of ldl cholesterol response to statins. *Nature communications*, 5(1):5068, 2014.

[5] Kaixin Zhou, Sook Wah Yee, Eric L Seiser, Nienke Van Leeuwen, Roger Tavendale, Amanda J Bennett, Christopher J Groves, Ruth L Coleman, Amber A Van Der Heijden, Joline W Beulens, et al. Variation in the glucose transporter gene slc2a2 is associated with glycemic response to metformin. *Nature genetics*, 48(9):1055–1059, 2016.

[6] Adem Y Dawed, Andrea Mari, Andrew Brown, Timothy J McDonald, Lin Li, Shuaicheng Wang, Mun-Gwan Hong, Sapna Sharma, Neil R Robertson, Anubha Mahajan, et al. Pharmacogenomics of glp-1 receptor agonists: a genome-wide analysis of observational data and large randomised controlled trials. *The Lancet Diabetes & Endocrinology*, 11(1):33–41, 2023.

[7] Erika Salvi, Zhiying Wang, Federica Rizzi, Yan Gong, Caitrin W McDonough, Sandosh Padmanabhan, Timo P Hiltunen, Chiara Lanzani, Roberta Zaninello, Martina Chittani, et al. Genome-wide and gene-based meta-analyses identify novel loci influencing blood pressure response to hydrochlorothiazide. *Hypertension*, 69(1):51–59, 2017.

[8] Sonal Singh, Helen R Warren, Timo P Hiltunen, Caitrin W McDonough, Nihal El Rouby, Erika Salvi, Zhiying Wang, Tatiana Garofalidou, Frej Fyhrquist, Kimmo K Kontula, et al. Genome-wide meta-analysis of blood pressure response to β1-blockers: results from icaps (international consortium of antihypertensive pharmacogenomics studies). *Journal of the American Heart Association*, 8(16):e013115, 2019.

[9] Caitrin W McDonough, Helen R Warren, John R Jack, Alison A Motsinger-Reif, Nicole D Armstrong, Joshua C Bis, John S House, Sonal Singh, Nihal M El Rouby, Yan Gong, et al. Adverse cardiovascular outcomes and antihypertensive treatment: A genome-wide interaction meta-analysis in the international consortium for antihypertensive pharmacogenomics studies. *Clinical Pharmacology & Therapeutics*, 110(3):723–732, 2021.

[10] Chiara Auwerx, Marie C Sadler, Alexandre Reymond, and Zoltán Kutalik. From pharmacogenetics to pharmaco-omics: Milestones and future directions. *Human Genetics and Genomics Advances*, 2022.

[11] Tõnis Tasa, Kristi Krebs, Mart Kals, Reedik Mägi, Volker M Lauschke, Toomas Haller, Tarmo Puurand, Maido Remm, Tõnu Esko, Andres Metspalu, et al. Genetic variation in the Estonian population: pharmacogenomics study of adverse drug effects using electronic health records. *European Journal of Human Genetics*, 27(3):442–454, 2019.

[12] Gregory McInnes and Russ B Altman. Drug response pharmacogenetics for 200,000 uk biobank participants. In *BIOCOMPUTING 2021: Proceedings of the Pacific Symposium*, pages 184–195. World Scientific, 2020.

[13] Mustafa Adnan Malki, Adem Y Dawed, Caroline Hayward, Alex Doney, and Ewan R Pearson. Utilizing large electronic medical record data sets to identify novel drug–gene interactions for commonly used drugs. *Clinical Pharmacology & Therapeutics*, 110(3):816–825, 2021.

[14] Tuomo Kiiskinen, Pyry Helkkula, Kristi Krebs, Juha Karjalainen, Elmo Saarentaus, Nina Mars, Arto Lehisto, Wei Zhou, Mattia Cordioli, Sakari Jukarainen, et al. Genetic predictors of lifelong medication-use patterns in cardiometabolic diseases. *Nature Medicine*, 29(1):209–218, 2023.

[15] Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O'Connell, et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203–209, 2018.

[16] The All of Us Research Program Investigators. The "All of Us" research program. *The New England Journal of Medicine*, 381(7):668–676, 2019.

[17] Jemma C Hopewell, Sarah Parish, Alison Offer, Emma Link, Robert Clarke, Mark Lathrop, Jane Armitage, Rory Collins, and MRC/BHF Heart Protection Study Collaborative Group. Impact of common genetic variation on response to simvastatin therapy among 18 705 participants in the heart protection study. *European heart journal*, 34(13):982–992, 2013.

[18] Jaideep Patel, H Robert Superko, Seth S Martin, Roger S Blumenthal, and Lisa Christopher-Stine. Genetic and immunologic susceptibility to statin-related myopathy. *Atherosclerosis*, 240(1):260–271, 2015.

[19] Elliot Sollis, Abayomi Mosaku, Ala Abid, Annalisa Buniello, Maria Cerezo, Laurent Gil, Tudor Groza, Osman Güneş, Peggy Hall, James Hayhurst, et al. The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. *Nucleic Acids Research*, 51(D1):D977–D985, 2023.

[20] Daniel I Chasman, Franco Giulianini, Jean MacFadyen, Bryan J Barratt, Fredrik Nyberg, and Paul M Ridker. Genetic determinants of statin-induced low-density lipoprotein cholesterol reduction: the Justification for the Use of Statins in Prevention: an Intervention Trial Evaluating Rosuvastatin (JUPITER) trial. *Circulation: Cardiovascular Genetics*, 5(2):257–264, 2012.

[21] Iris Postmus, Helen R Warren, Stella Trompet, Benoit J Arsenault, Christy L Avery, Joshua C Bis, Daniel I Chasman, Catherine E de Keyser, Harshal A Deshmukh, Daniel S Evans, et al. Meta-analysis of genome-wide association studies of HDL cholesterol response to statins. *Journal of medical genetics*, 53(12):835–845, 2016.

[22] Lorna W Harries, Andrew T Hattersley, Alex SF Doney, Helen Colhoun, Andrew D Morris, Calum Sutherland, D Grahame Hardie, Leena Peltonen, Mark I McCarthy, et al. Common variants near

ATM are associated with glycemic response to metformin in type 2 diabetes. *Nature genetics*, 43(2):117–120, 2011.

[23] Daniel M Rotroff, Sook Wah Yee, Kaixin Zhou, Skylar W Marvel, Hetal S Shah, John R Jack, Tammy M Havener, Monique M Hedderson, Michiaki Kubo, Mark A Herman, et al. Genetic variants in CPA6 and PRPF31 are associated with variation in response to metformin in individuals with type 2 diabetes. *Diabetes*, 67(7):1428–1440, 2018.

[24] Gustavo H Oliveira-Paula, Sherliane C Pereira, Jose E Tanus-Santos, and Riccardo Lacchini. Pharmacogenomics and hypertension: current insights. *Pharmacogenomics and personalized medicine*, pages 341–359, 2020.

[25] Seunggeun Lee, Michael C Wu, and Xihong Lin. Optimal tests for rare variant effects in sequencing association studies. *Biostatistics*, 13(4):762–775, 2012.

[26] Michel Marcil, Angela Brooks-Wilson, Susanne M Clee, Kirsten Roomp, Lin-Hua Zhang, Lu Yu, Jennifer A Collins, Marjel van Dam, Henri OF Molhuizen, Odell Loubster, et al. Mutations in the ABC 1 gene in familial HDL deficiency with defective cholesterol efflux. *The Lancet*, 354(9187):1341–1346, 1999.

[27] Cong Zhang, Konstantin Shestopaloff, Benjamin Hollis, Chun Hei Kwok, Claudia Hon, Nicole Hartmann, Chengeng Tian, Magdalena Wozniak, Luis Santos, Dominique West, et al. Response to anti-IL17 therapy in inflammatory disease is not strongly impacted by genetic background. *The American Journal of Human Genetics*, 2023.

[28] Josephine H Li, Lukasz Szczerbinski, Adem Y Dawed, Varinderpal Kaur, Jennifer N Todd, Ewan R Pearson, and Jose C Florez. A polygenic score for type 2 diabetes risk is associated with both the acute and sustained response to sulfonylureas. *Diabetes*, 70(1):293–300, 2021.

[29] Jian-Ping Zhang, Delbert Robinson, Jin Yu, Juan Gallego, W Wolfgang Fleischhacker, Rene S Kahn, Benedicto Crespo-Facorro, Javier Vazquez-Bourgon, John M Kane, Anil K Malhotra, et al. Schizophrenia polygenic risk score as a predictor of antipsychotic efficacy in first-episode psychosis. *American Journal of Psychiatry*, 176(1):21–28, 2019.

[30] Neo M Tapela, Jennifer Collister, Xiaonan Liu, Lei Clifton, Alexander Stiby, Federico Murgia, Jemma C Hopewell, and David J Hunter. Are polygenic risk scores for systolic blood pressure and ldl-cholesterol associated with treatment effectiveness, and clinical outcomes among those on treatment? *European Journal of Preventive Cardiology*, 29(6):925–937, 2022.

[31] Jessica L Mega, Nathan O Stitziel, J Gustav Smith, Daniel I Chasman, Mark J Caulfield, James J Devlin, Francesco Nordio, Craig L Hyde, Christopher P Cannon, Frank M Sacks, et al. Genetic risk, coronary heart disease events, and the clinical benefit of statin therapy: an analysis of primary and secondary prevention trials. *The Lancet*, 385(9984):2264–2271, 2015.

[32] Pradeep Natarajan, Robin Young, Nathan O Stitziel, Sandosh Padmanabhan, Usman Baber, Roxana Mehran, Samantha Sartori, Valentin Fuster, Dermot F Reilly, Adam Butterworth, et al. Polygenic risk score identifies subgroup with higher burden of atherosclerosis and greater relative benefit from statin therapy in the primary prevention setting. *Circulation*, 135(22):2091–2101, 2017.

[33] Amy Damask, P Gabriel Steg, Gregory G Schwartz, Michael Szarek, Emil Hagström, Lina Badimon, M John Chapman, Catherine Boileau, Sotirios Tsimikas, Henry N Ginsberg, et al. Patients with high genome-wide polygenic risk scores for coronary artery disease may receive greater clinical benefit from alirocumab treatment in the ODYSSEY OUTCOMES trial. *Circulation*, 141(8):624–636, 2020.

[34] Nicholas A Marston, Frederick K Kamanu, Francesco Nordio, Yared Gurmu, Carolina Roselli, Peter S Sever, Terje R Pedersen, Anthony C Keech, Huei Wang, Armando Lira Pineda, et al. Predicting benefit from evolocumab therapy in patients with atherosclerotic disease using a genetic risk score: results from the fourier trial. *Circulation*, 141(8):616–623, 2020.

[35] Lei Clifton and David A Clifton. The correlation between baseline score and post-intervention score, and its implications for statistical analysis. *Trials*, 20(1):1–6, 2019.

[36] Joelle Mbatchou, Leland Barnard, Joshua Backman, Anthony Marcketta, Jack A Kosmicki, Andrey Ziyatdinov, Christian Benner, Colm O'Dushlaine, Mathew Barber, Boris Boutkov, et al. Compu-

tationally efficient whole-genome regression for quantitative and binary traits. *Nature genetics*, 53(7):1097–1103, 2021.

[37] Christopher C Chang, Carson C Chow, Laurent CAM Tellier, Shashaank Vattikuti, Shaun M Purcell, and James J Lee. Second-generation plink: rising to the challenge of larger and richer datasets. *Gigascience*, 4(1):s13742–015, 2015.

[38] Hannah V Meyer. plinkQC: genotype quality control in genetic association studies, 2022. meyer-lab-cshl/plinkQC, `https://doi.org/10.5281/zenodo.3934294`.

[39] Olga Krasheninina, Yih-Chii Hwang, Xiaodong Bai, Aleksandra Zalcman, Evan Maxwell, Jeffrey G Reid, and William J Salerno Jr. Open-source mapping and variant calling for large-scale ngs data from original base-quality scores. *bioRxiv*, pages 2020–12, 2020.

[40] Kevin J Galinsky, Gaurav Bhatia, Po-Ru Loh, Stoyan Georgiev, Sayan Mukherjee, Nick J Patterson, and Alkes L Price. Fast principal-component analysis reveals convergent evolution of adh1b in europe and east asia. *The American Journal of Human Genetics*, 98(3):456–472, 2016.

[41] Deborah J Thompson, Daniel Wells, Saskia Selzam, Iliana Peneva, Rachel Moore, Kevin Sharp, William A Tarran, Edward J Beard, Fernando Riveros-Mckay, Carla Giner-Delgado, et al. Uk biobank release and systematic evaluation of optimised polygenic risk scores for 53 diseases and quantitative traits. *MedRxiv*, pages 2022–06, 2022.

[42] Andrew D Grotzinger, Mijke Rhemtulla, Ronald de Vlaming, Stuart J Ritchie, Travis T Mallard, W David Hill, Hill F Ip, Riccardo E Marioni, Andrew M McIntosh, Ian J Deary, et al. Genomic structural equation modelling provides insights into the multivariate genetic architecture of complex traits. *Nature human behaviour*, 3(5):513–525, 2019.

[43] Sarah E Graham, Shoa L Clarke, Kuan-Han H Wu, Stavroula Kanoni, Greg JM Zajac, Shweta Ramdas, Ida Surakka, Ioanna Ntalla, Sailaja Vedantam, Thomas W Winkler, et al. The power of genetic diversity in genome-wide association studies of lipids. *Nature*, 600(7890):675–679, 2021.

[44] Evangelos Evangelou, Helen R Warren, David Mosen-Ansorena, Borbala Mifsud, Raha Pazoki, He Gao, Georgios Ntritsos, Niki Dimou, Claudia P Cabrera, Ibrahim Karaman, et al. Genetic analysis of over 1 million people identifies 535 new loci associated with blood pressure traits. *Nature genetics*, 50(10):1412–1425, 2018.

Supplementary Materials for

Cardiometabolic drug response pharmacogenetics using EHRs from biobanks

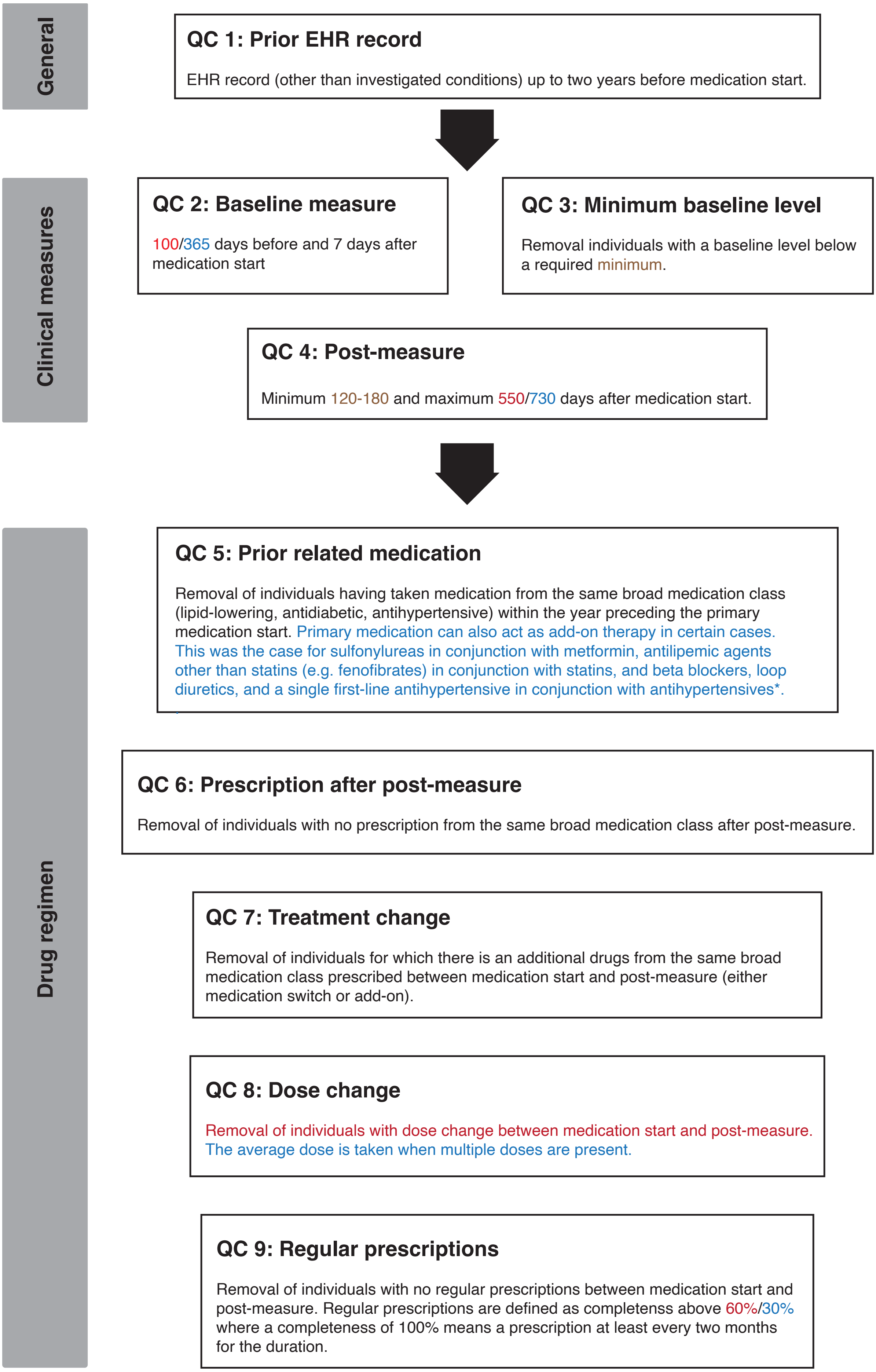
Supplementary Note 1.

HbA1c values were either DCCT (Diabetes Control and Complications Trial) aligned (codes: 42W4. And XaERp; percentage unit) or IFCC (International Federation of Clinical Chemistry and Laboratory Medicine) aligned (42W5. And XaPbt; mmol/mol unit). For consistency, we used mmol/mol units and converted DCCT units using the NGSP/IFCC equation recommended by the National Glycohemoglobin Standardization Program (NGSP) network (https://ngsp.org/ifcc.asp): NGSP = [0.09148 * IFCC] + 2.152.

Supplementary Note 2.

Medication codes can correspond to multiple active ingredients taken in combination, among which the primary medication of interest. Since we cannot disentangle the effect of the primary medication compared to a second ingredient taken in combination, we filter out individuals with prescriptions corresponding to combination therapies during the study period. For statins we eliminate combination therapies with ezetimibe and fenofibrate, for metformin, combination therapies with sitagliptin, linagliptin, saxagliptin, alogliptin, dapagliflozin, canagliflozin, empagliflozin, rosiglitazone, pioglitazone, vildagliptin and for beta blockers, combination therapies with diuretics and aspirin.

Note that this step is specific to drugs with a combined formulation and is different from the QC step where individuals taking a drug from the same medication class, but with a separate prescription code, are filtered out.



General

QC 1: Prior EHR record

EHR record (other than investigated conditions) up to two years before medication start.

Clinical measures

QC 2: Baseline measure

100/365 days before and 7 days after medication start

QC 3: Minimum baseline level

Removal individuals with a baseline level below a required minimum.

QC 4: Post-measure

Minimum 120-180 and maximum 550/730 days after medication start.

Drug regimen

QC 5: Prior related medication

Removal of individuals having taken medication from the same broad medication class (lipid-lowering, antidiabetic, antihypertensive) within the year preceding the primary medication start. Primary medication can also act as add-on therapy in certain cases. This was the case for sulfonylureas in conjunction with metformin, antilipemic agents other than statins (e.g. fenofibrates) in conjunction with statins, and beta blockers, loop diuretics, and a single first-line antihypertensive in conjunction with antihypertensives*.

QC 6: Prescription after post-measure

Removal of individuals with no prescription from the same broad medication class after post-measure.

QC 7: Treatment change

Removal of individuals for which there is an additional drugs from the same broad medication class prescribed between medication start and post-measure (either medication switch or add-on).

QC 8: Dose change

Removal of individuals with dose change between medication start and post-measure. The average dose is taken when multiple doses are present.

QC 9: Regular prescriptions

Removal of individuals with no regular prescriptions between medication start and post-measure. Regular prescriptions are defined as completenss above 60%/30% where a completeness of 100% means a prescription at least every two months for the duration.

Figure S1. Flow diagram of quality control steps. After selecting individuals taking the primary medication of interest, individuals with missing clinical measures (i.e., drug response phenotype), medication therapy changes prior posttreatment measures, irregular prescriptions, or not enrolled in the health care system prior medication start. Stringent filtering criteria are highlighted in red and lenient ones in blue. Medication/phenotype-specific criteria are highlighted in brown.

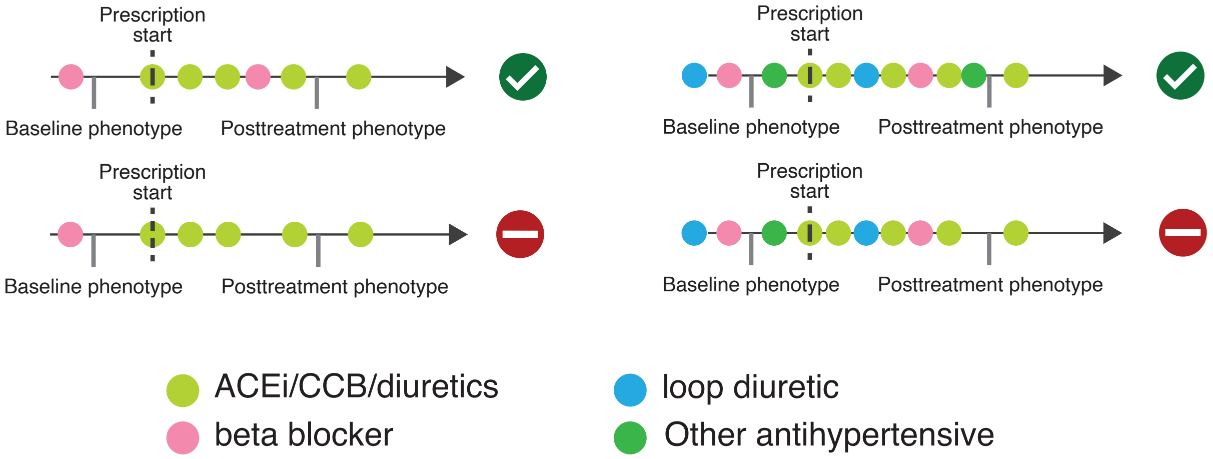
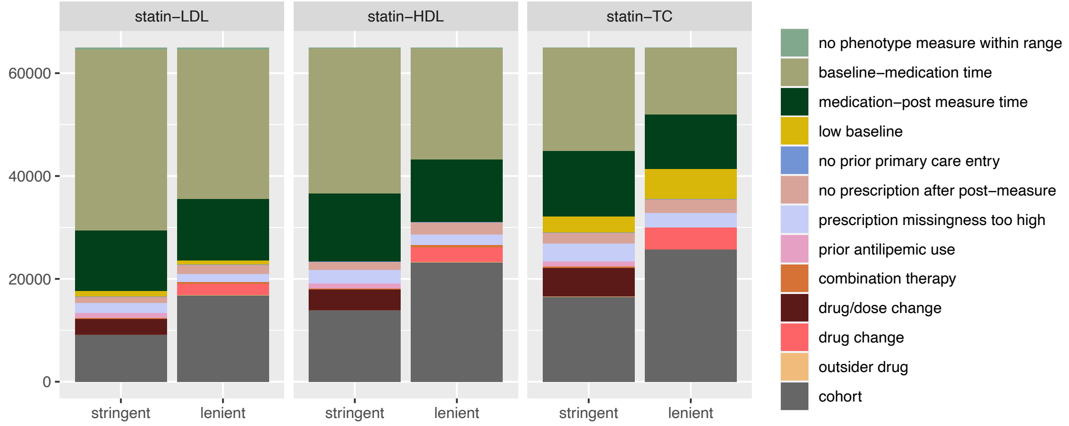
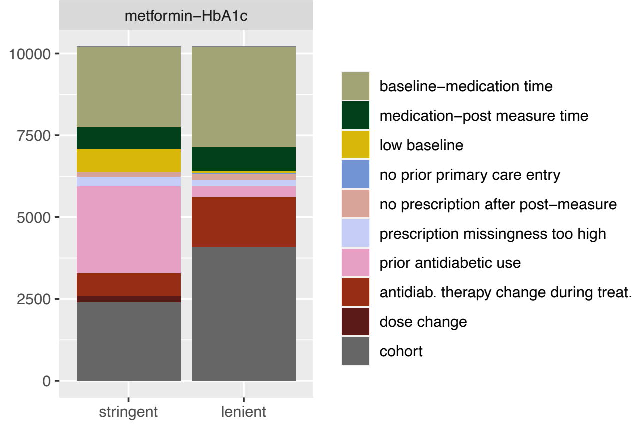
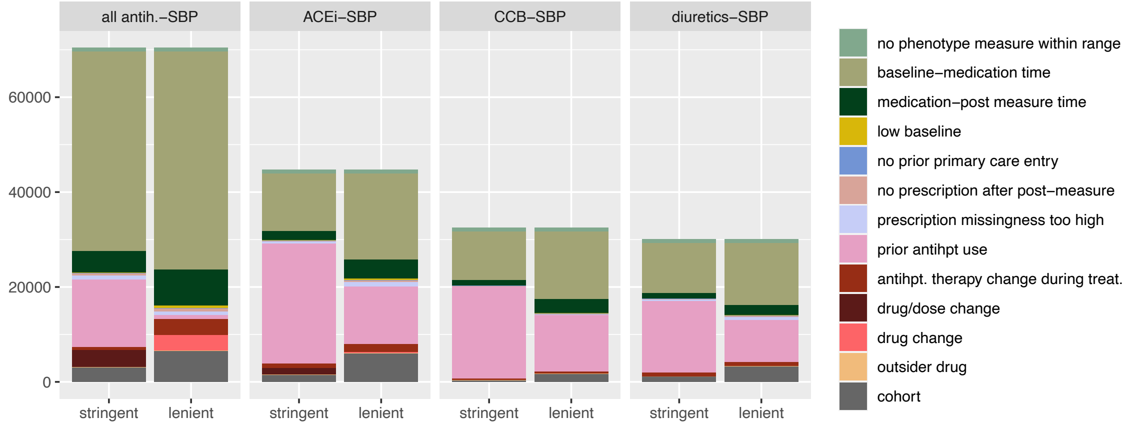


Figure S2. Add-on therapy definition. For antihypertensives, primary medication (ACEi, CCB and thiazide diuretics) could also act as add-on therapy to beta blockers, loop diuretics and a single other antihypertensive. However, medication prescribed prior primary medication start was required to be prescribed afterwards (at least until posttreatment measurement time). If the start of a concomitant medication was after the prescription start of the primary medication, this would count as “treatment change” and the individual would be removed.


statin–LDL
statin–HDL
statin–TC
60000
40000
20000
0
stringent
lenient
no phenotype measure within range
baseline–medication time
medication–post measure time
low baseline
no prior primary care entry
no prescription after post–measure
prescription missingness too high
prior antilipemic use
combination therapy
drug/dose change
drug change
outsider drug
cohort


metformin–HbA1c
10000
7500
5000
2500
0
stringent
lenient
baseline–medication time
medication–post measure time
low baseline
no prior primary care entry
no prescription after post–measure
prescription missingness too high
prior antidiabetic use
antidiab. therapy change during treat.
dose change
cohort


all antih.–SBP
ACEi–SBP
CCB–SBP
diuretics–SBP
60000
40000
20000
0
stringent
lenient
no phenotype measure within range
baseline–medication time
medication–post measure time
low baseline
no prior primary care entry
no prescription after post–measure
prescription missingness too high
prior antihpt use
antihpt. therapy change during treat.
drug/dose change
drug change
outsider drug
cohort

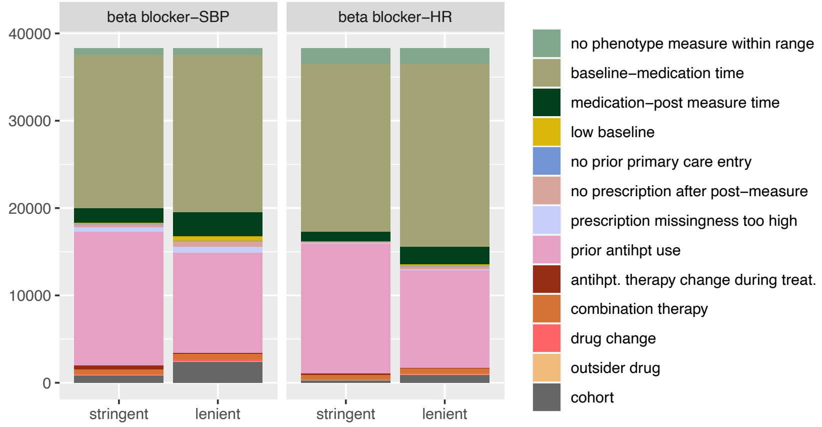

beta blocker–SBP
beta blocker–HR
40000
30000
20000
10000
0
stringent
lenient
no phenotype measure within range
baseline–medication time
medication–post measure time
low baseline
no prior primary care entry
no prescription after post–measure
prescription missingness too high
prior antihpt use
antihpt. therapy change during treat.
combination therapy
drug change
outsider drug
cohort

Figure S3. Number of individuals in each UK Biobank drug response cohort and reasons for removal (stacked barplot). The height of the bar represents the number of individuals having at least one prescription of the investigated drug. The bottom grey bar represents the number of individuals after QC steps. Note some filtering reasons are not mutually exclusive. For instance, baseline-medication time filtering was done after checking for prior related medication. Therefore, for metformin-HbA1c, it seems that more individuals were filtered out because of baseline-medication time than in the stringent scenario. However, given that individuals with previous sulfonylureas use were excluded in the stringent, but included in the lenient filtering setting, there is a larger pool of individuals for whom baseline measures are potentially missing. The same reasoning holds for antihypertensives where individuals with prior antihypertensive prescriptions were included in certain scenarios (see Figure S2) in the lenient filtering setting.

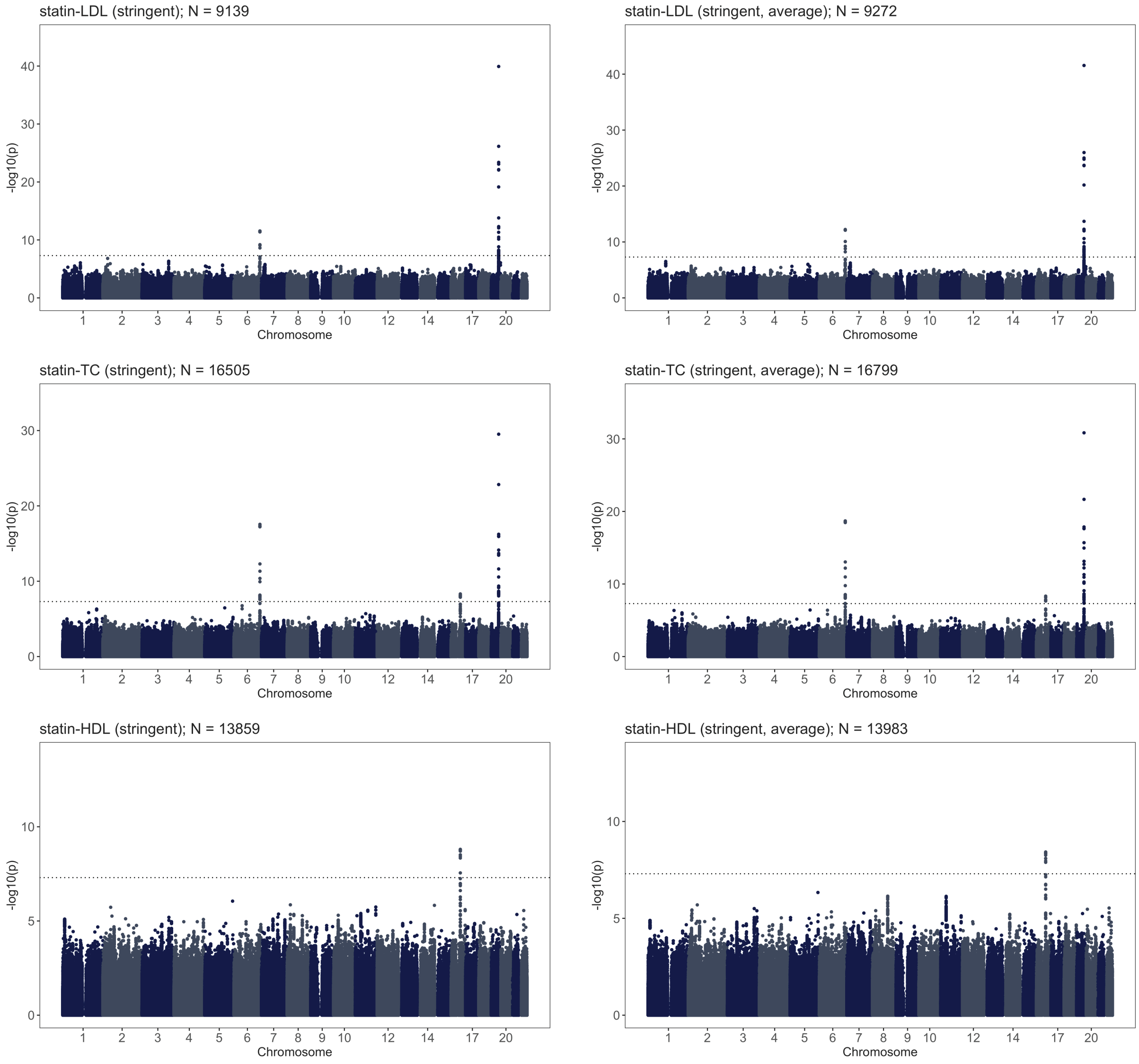


Figure S4. Lipid response to statin GWAS results in the stringent filtering scenario. Plots on the left use single baseline and posttreatment measures and plots on the right average values if available. The horizontal line denotes genome-wide significance (p-value < 5e-8).

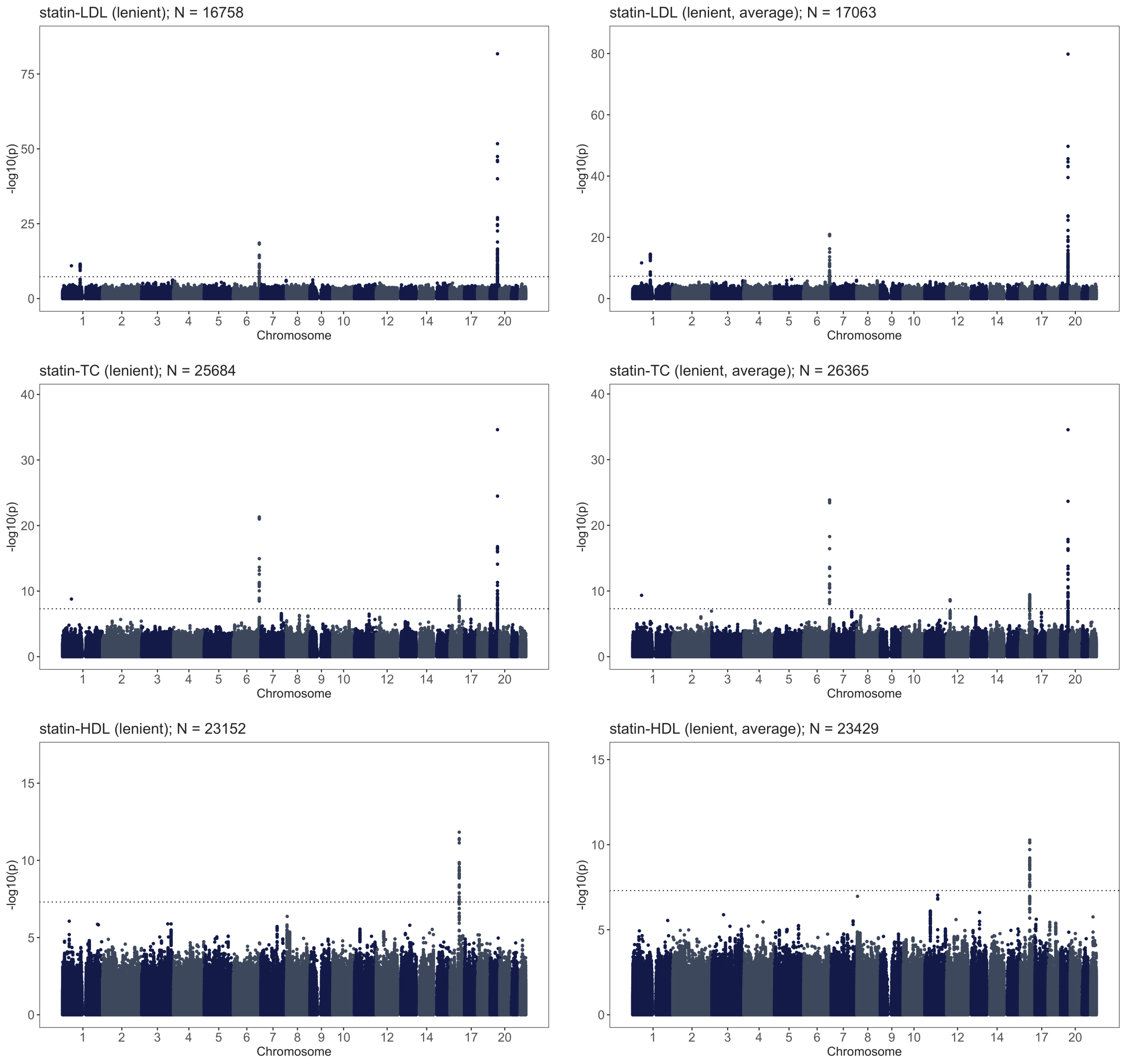


Figure S5. Lipid response to statin GWAS results in the lenient filtering scenario. Plots on the left use single baseline and posttreatment measures and plots on the right average values if available. The horizontal line denotes genome-wide significance (p-value < 5e-8).

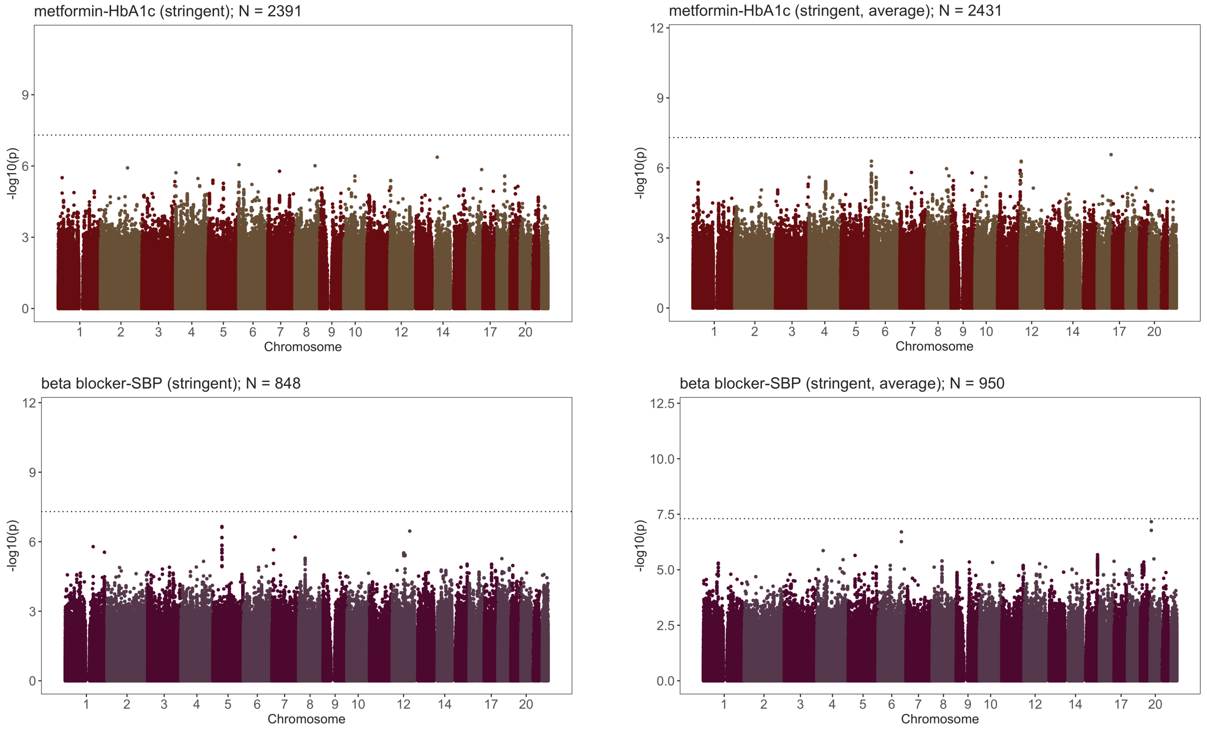


Figure S6. HbA1c response to metformin and SBP response to beta blocker GWAS results in the stringent filtering scenario. Plots on the left use single baseline and posttreatment measures and plots on the right average values if available. The horizontal line denotes genome-wide significance (p-value < 5e-8).

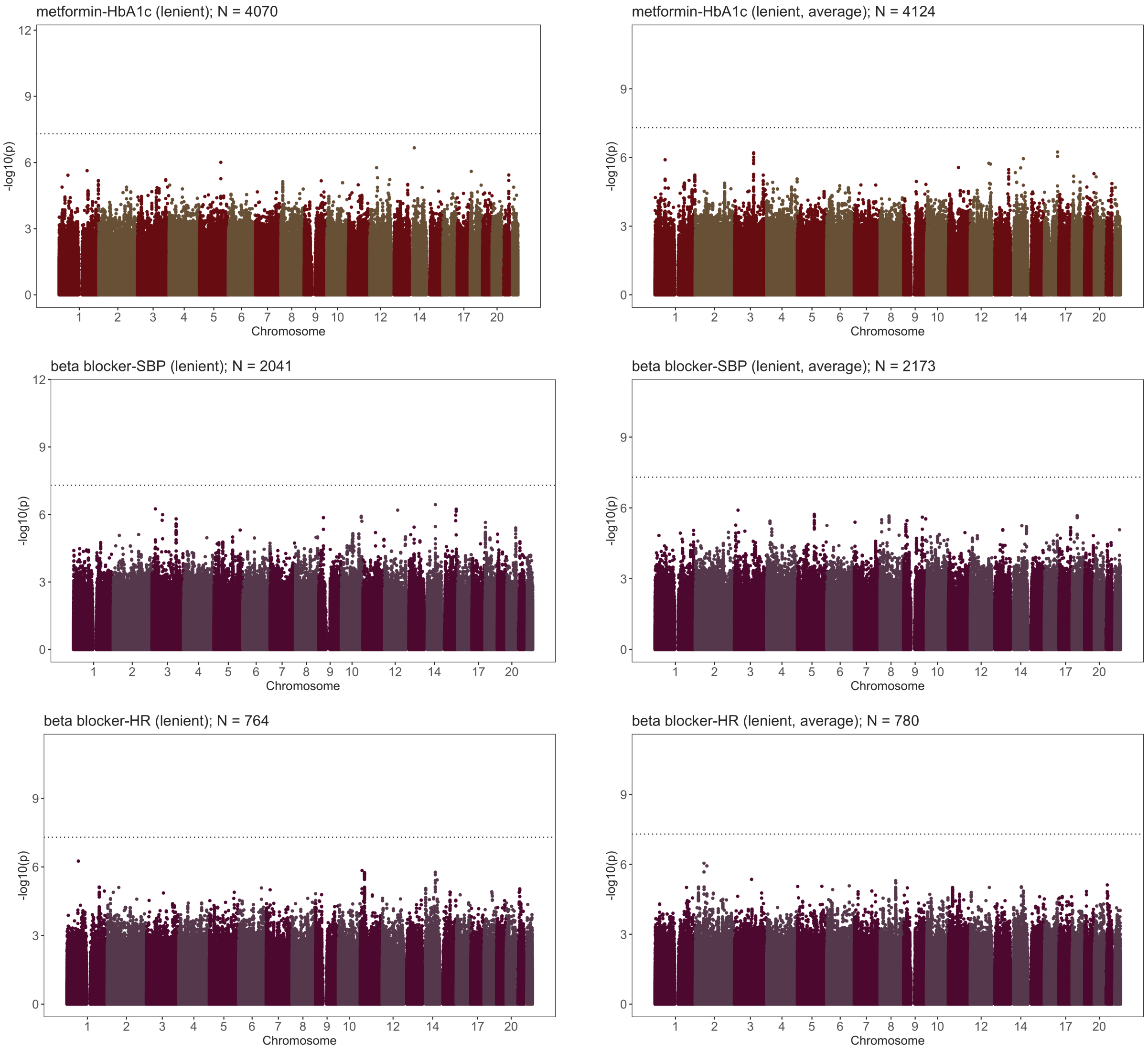


Figure S7. HbA1c response to metformin and SBP response to beta blocker GWAS results in the lenient filtering scenario. Plots on the left use single baseline and posttreatment measures and plots on the right average values if available. The horizontal line denotes genome-wide significance (p-value < 5e-8).

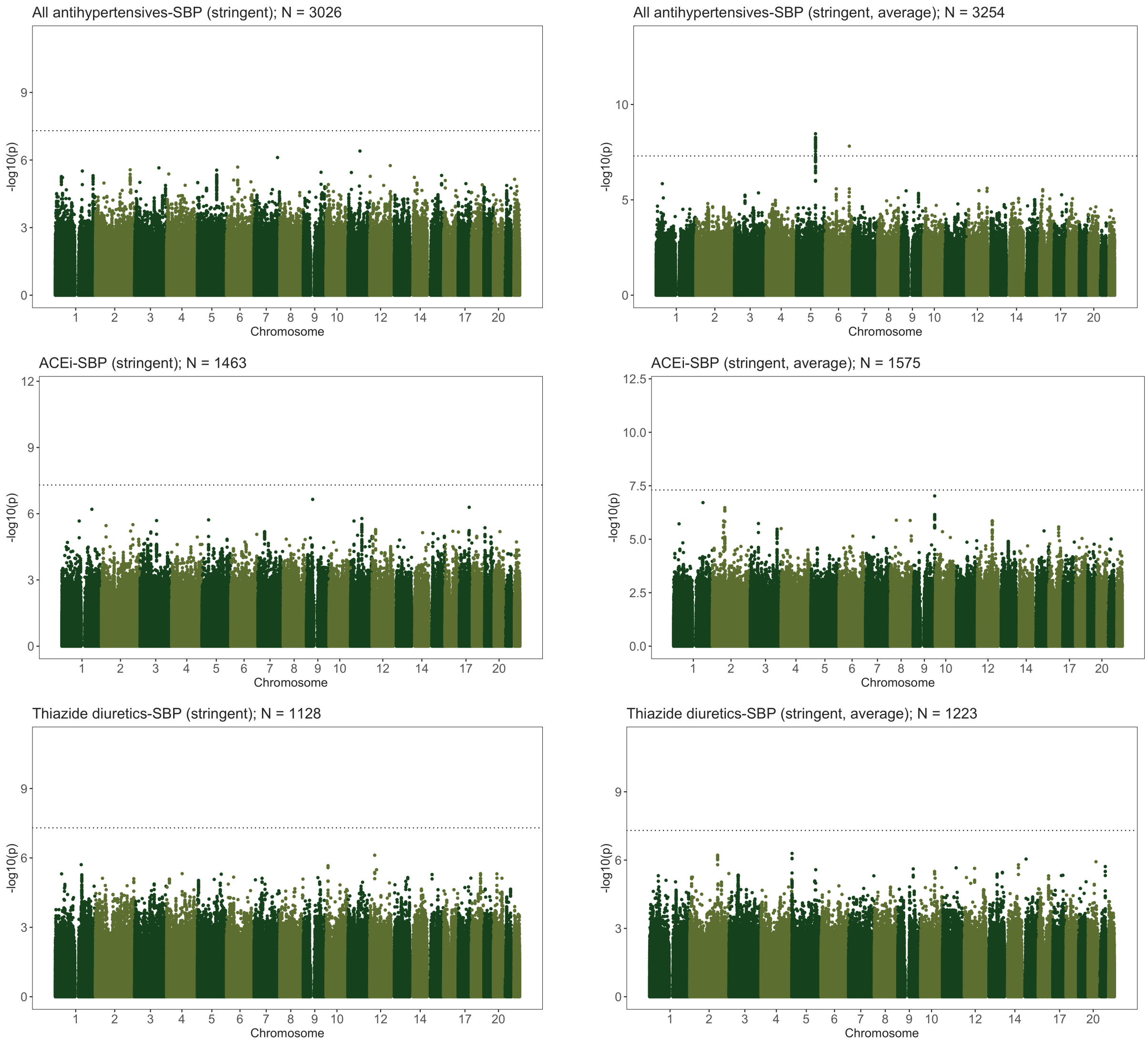


Figure S8. SBP response to antihypertensives GWAS results in the stringent filtering scenario. Plots on the left use single baseline and posttreatment measures and plots on the right average values if available. The horizontal line denotes genome-wide significance (p-value < 5e-8).

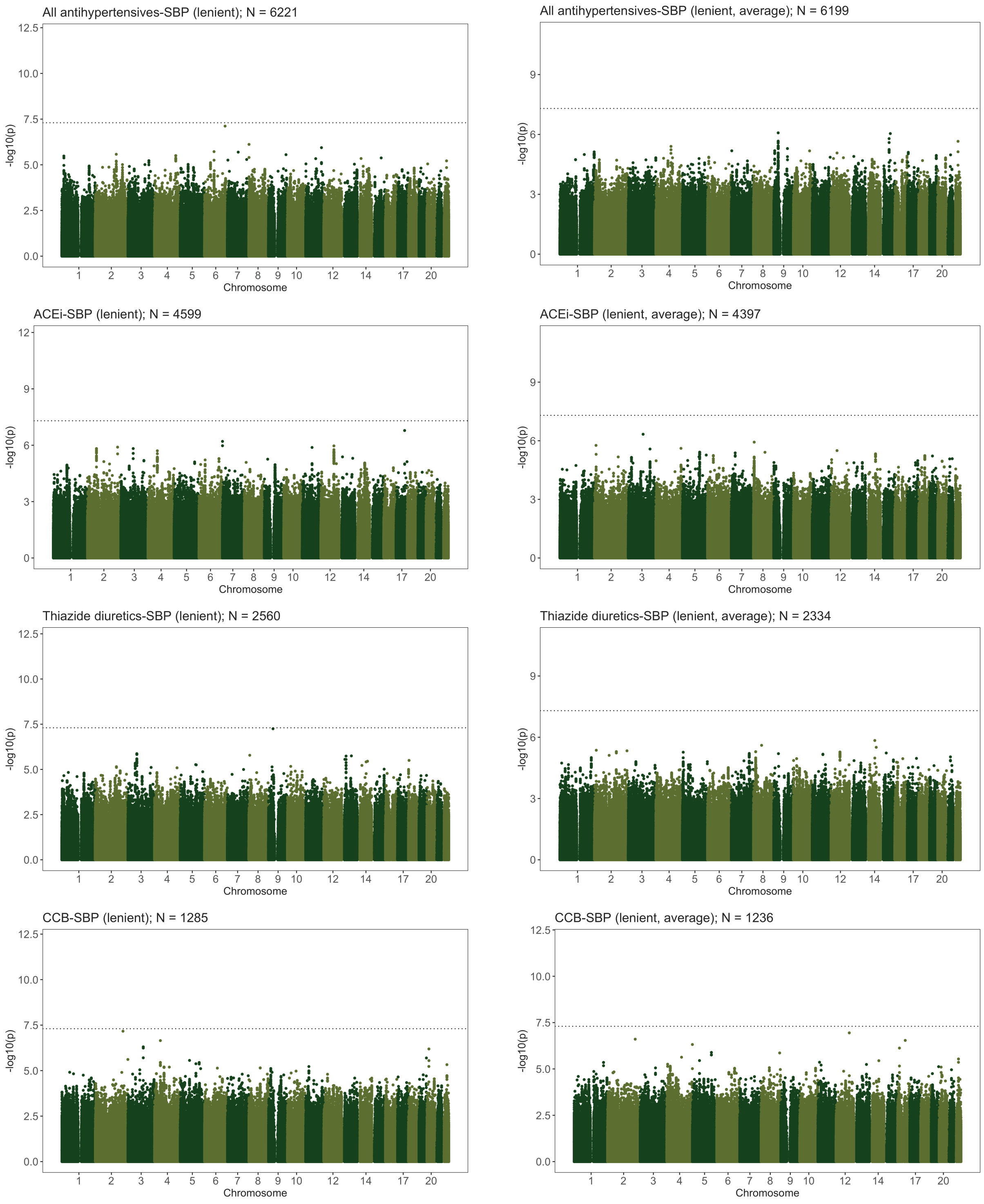


Figure S9. SBP response to antihypertensives GWAS results in the lenient filtering scenario. Plots on the left use single baseline and posttreatment measures and plots on the right average values if available. The horizontal line denotes genome-wide significance (p-value < 5e-8).

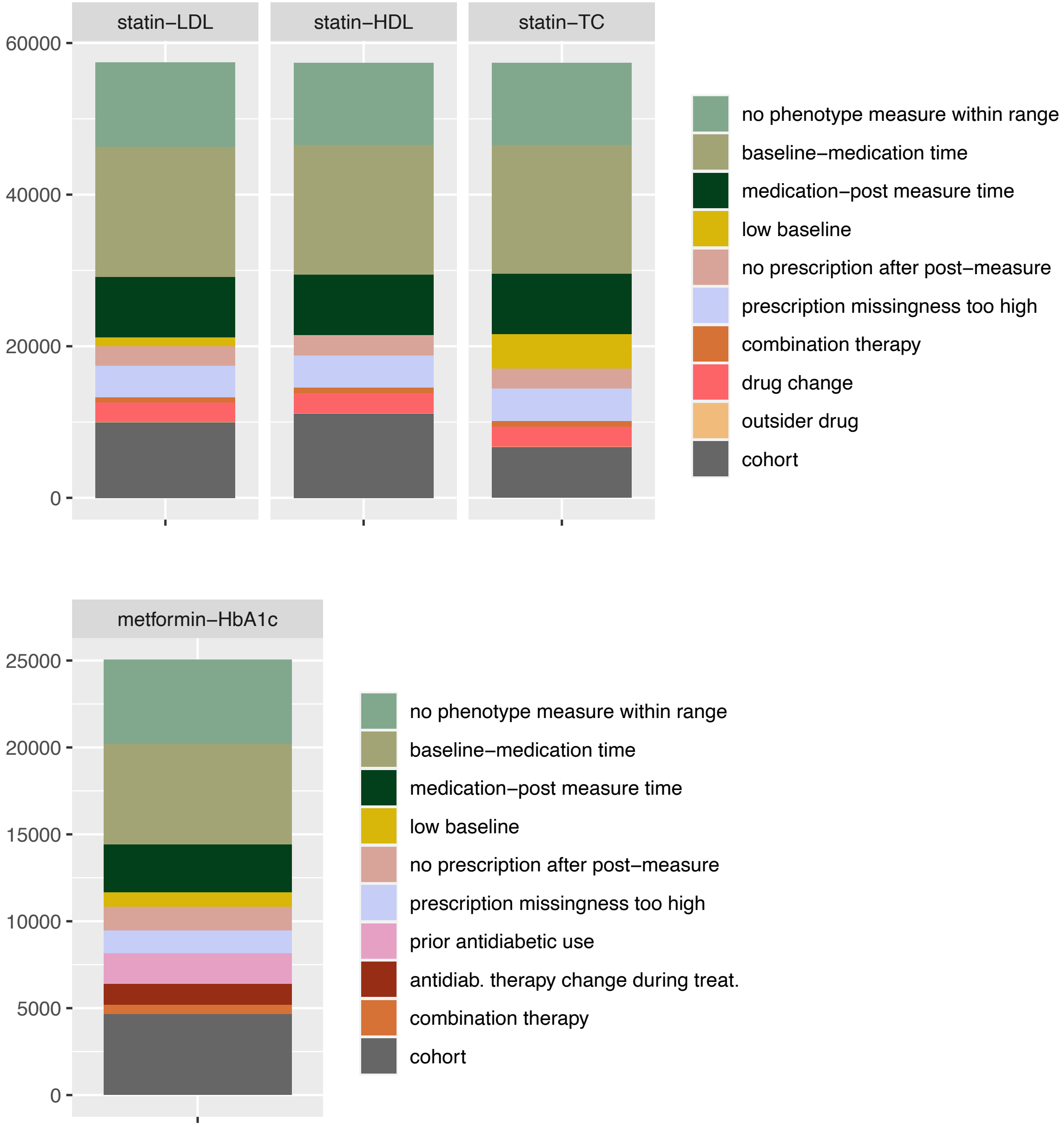


Figure S10. Number of individuals in each All of Us drug response cohort and reasons for removal (stacked barplot). The height of the bar represents the number of individuals having at least one prescription of the investigated drug. The bottom grey bar represents the number of individuals after QC steps. Note some filtering reasons are not mutually exclusive.

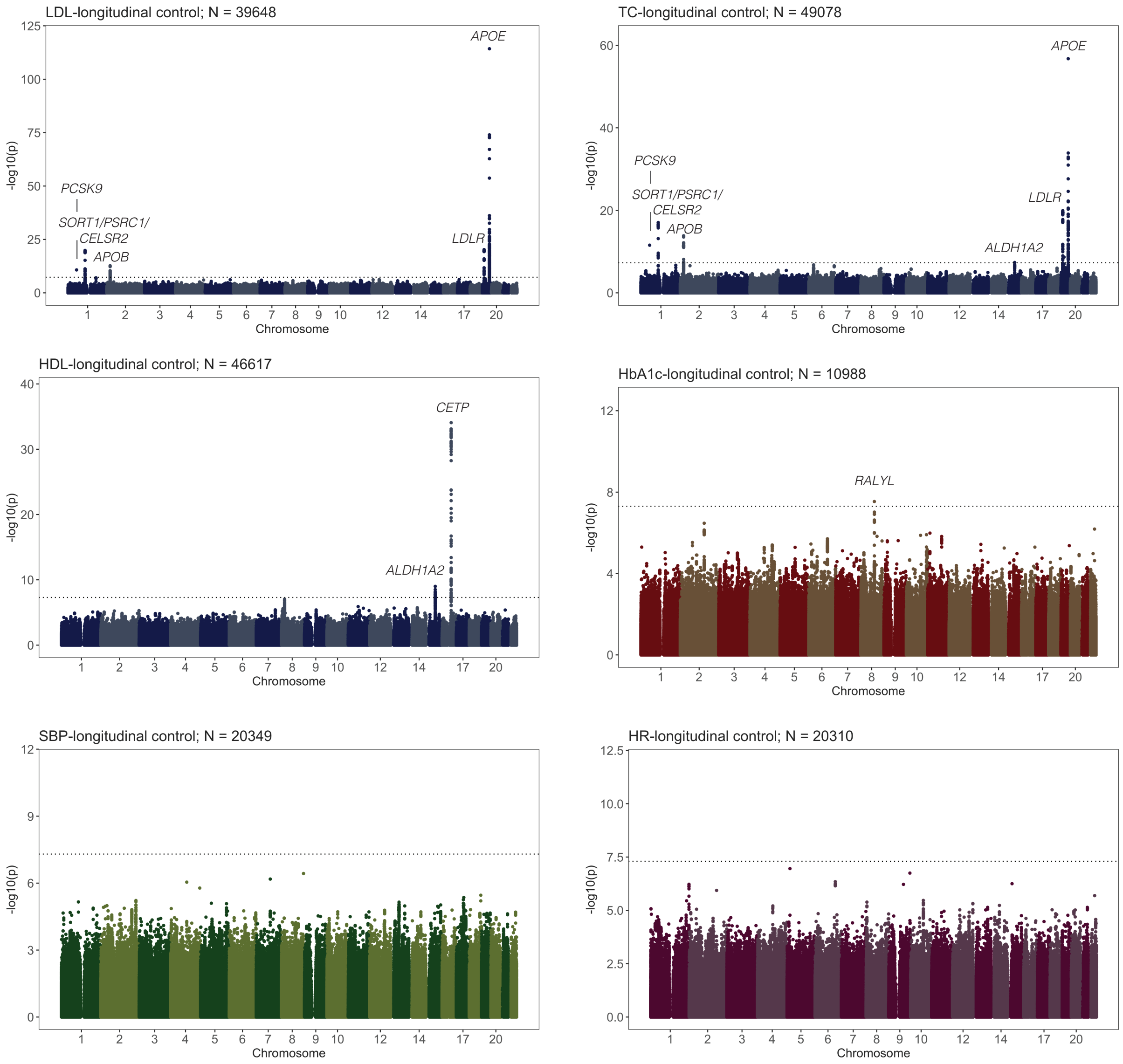


Figure S11. Longitudinal biomarker change GWAS in medication-naïve individuals. Genome-wide significant loci are annotated with the closest gene. The horizontal line denotes genome-wide significance (p-value < 5e-8).